

MecCog: A knowledge representation framework for genetic disease mechanism

Kunal Kundu^{1,2}, Lindley Darden³, John Moulton^{2,4} *

¹ Computational Biology, Bioinformatics and Genomics, Biological Sciences Graduate Program, University of Maryland, College Park, MD 20742, USA

² Institute for Bioscience and Biotechnology Research, University of Maryland, 9600 Gudelsky Drive, Rockville, MD 20850, USA

³ Department of Philosophy, University of Maryland, College Park, MD 20742, USA

⁴ Department of Cell Biology and Molecular Genetics, University of Maryland, College Park, MD 20742, USA

*Corresponding author

jmoulton@umd.edu

ABSTRACT

Motivation: Experimental findings on genetic disease mechanisms are scattered throughout the literature and represented in many ways, including unstructured text, cartoons, pathway diagrams, and network graphs. Integration and structuring of such mechanistic information will greatly enhance its utility.

Results: MecCog is a graphical framework for building integrated representations (mechanism schemas) of mechanisms by which a genetic variant causes a disease phenotype. A MecCog mechanism schema displays the propagation of system perturbations across stages of biological organization, using graphical notations to symbolize perturbed entities and activities, hyperlinked evidence tagging, a mechanism ontology, and depiction of knowledge gaps, ambiguities, and uncertainties. The web platform enables a user to construct, store, publish, browse, query, and comment on schemas. MecCog facilitates the identification of potential biomarkers, therapeutic intervention sites, and critical future experiments.

Availability and Implementation: The MecCog framework is freely available at <http://www.meccog.org>.

Contact: jmoult@umd.edu

Supplementary information: Supplementary material is available at *Bioinformatics* online.

INTRODUCTION

Findings from experimental studies of disease mechanism are reported across multiple publications in varying combinations of structured and unstructured data and many different diagrammatic representations. A number of projects have addressed different aspects of the resulting knowledge integration problem. These include building disease-specific knowledge management resources (for example alzforum.org (Kinoshita and Clark, 2007)) and ontologies (ADO (Malhotra *et al.*, 2014), PDON (Younesi *et al.*, 2015), CVDO <https://biportal.bioontology.org/ontologies/CVDO>); compiling disease etiology databases (HGMD (Stenson *et al.*, 2017), ClinVar (Landrum *et al.*, 2018), CIVIC (Griffith *et al.*, 2017), PanelApp (Martin *et al.*, 2019)); development of biomedical text mining methods (DARPA's Big Mechanism program (Cohen, 2015)); development of statistical methods for evidence integration and assessment (Konopka and Smedley, 2020); and community-driven expert systems medicine disease maps projects (Mazein *et al.*, 2018). Each of these contributes elements of a solution, but a major omission is an integrated representation of mechanism knowledge in a clear, precise, and comprehensive manner.

There have also been major technological advances in the development of tools to support mechanism descriptions, such as graphical notations (SGBN (Systems Biology Graphical Notation) (Novère *et al.*, 2009)) and languages (SBML (Systems Biology Markup Language) (Hucka *et al.*, 2018), KGML (KEGG Markup language - <https://www.genome.jp/kegg/xml/>), BCML (Biological Connection Markup Language) (Beltrame *et al.*, 2011), BioPAX (<http://www.biopax.org/>), BEL (Biological Expression Language - <https://bel.bio/>)) to encode representations; software to draw and visualize models (GO-CAM (Thomas *et al.*, 2019), PathWhiz (Pon *et al.*, 2015), Cytoscape (Shannon *et al.*, 2003)); linked data formats such as Nanopublications (Mina *et al.*, 2015) to organize provenance and metadata for scientific

assertions; and databases to store and query graph-based representations (Neo4j - <https://neo4j.com/>, (Himmelstein *et al.*, 2017)).

With the help of these tools; pathway, network, and disease map representation types have been created to describe aspects of biological system mechanism and in some cases disease mechanisms as well. For instance, KEGG (Kanehisa *et al.*, 2016), and Reactome (Fabregat *et al.*, 2017) pathways represent normal and perturbed molecular interactions that are part of cellular or metabolic processes. STRING (Szklarczyk *et al.*, 2018) and GeneMANIA (Franz *et al.*, 2018) networks represent integrated information on protein-protein interactions and associations that are part of normally functioning biological systems. Gene ontology (GO) causal activity models (Thomas *et al.*, 2019) integrate GO annotations to generate larger models of normal biological function (such as ‘pathways’) in a semantically structured manner. The Disease Maps Project (Mazein *et al.*, 2018) provides an encyclopedic description of disease-related signaling, metabolic, and gene regulatory processes. Although together these representations aptly describe the normal working of biological systems, representation of the disease related perturbations is limited. In the existing representations (such as KEGG or Reactome disease pathways), disease state perturbations and consequences are added locally to the depiction of the normal state of the biological system. Adding disease perturbation information to already complex pathway diagrams can be useful, but limits clarity. Also, uncertainties, ambiguities, and ignorance in mechanistic knowledge are not presented in most representations (with the exception of Reactome pathway diagrams, but these label uncertain reaction types only). Such knowledge gaps exist in almost all disease mechanisms (Greenberg and Amato, 2004; Kametani and Hasegawa, 2018).

These considerations led us to propose a graphical framework with an integrated representation of genetic disease mechanisms from gene to phenotype. Our design goals were that the representation framework depict mechanism components across stages of biological organization; display perturbation propagation; make use of standard biomedical ontology terms wherever possible to name the components; provide an intuitive way to visualize ignorance, uncertainties, and ambiguities; and allow tight linkage to evidence in the literature and databases. The MecCog mechanism representation framework (Darden *et al.*, 2018) incorporates all these features. The representation formalism is based on the analysis of biological mechanisms developed in the philosophy of biology (Craver and Darden, 2013): Mechanisms are characterized as entities and activities organized such that they are productive of regular changes from start or set-up to finish or termination conditions. In MecCog, a mechanism by which a genetic variant causes a disease phenotype is represented as a *mechanism schema* that displays the propagation of entity and activity perturbations across biological organizational stages (DNA→RNA→Protein→Complex→Organelle→Cell→Tissue→Organ→Phenotype) in the form of a graph (nodes are biological entities; directed edges are causal and labeled with productive activities) constructed from information in the biomedical literature in addition to established biological concepts. The schema structure uses graph properties such as branching, merging, and looping of sub-paths.

In this article, we describe the implementation of the MecCog framework as a web platform with a collaborative environment to manually construct, store, publish, browse, query, and comment on mechanism schemas for genetic diseases. The schema building tool in MecCog is supported by specially designed graphical notations, curated ontology-informed terminology for the annotation of mechanism components (entities and activities), an interactive graphical user interface (GUI) to construct the schema drawings, application programming interfaces (APIs) to fetch reference information and

scientific figures, tight integration and hyperlinking of evidence to the graphics, and a secure server to save schemas as JSON (JavaScript Object Notation) objects. The platform supports edit, version, and share operations on each schema to facilitate collaborative work. Mature schemas can be published on the platform, thereby adding to the collection of disease mechanisms available for browsing by MecCog web-site visitors. Sketchier schemas with gaps, ambiguities, and uncertainties can also be published to indicate where additional work needs to be directed.

METHODS AND RESULTS

Mechanism schema representation structure

In MecCog, a mechanism by which a genetic variant causes a disease phenotype is represented by multiple steps. Each step consists of a triplet with an input substate perturbation, a mechanism module, and an output substate perturbation (SSP-MM-SSP). A substate perturbation represents a perturbed biological entity (e.g. a DNA base change, altered stability of a protein, altered abundance of a molecular complex, altered state of a cell). A mechanism module represents the productive activity (e.g. transcription, translation, or protein-protein interaction) by which the input sub-state perturbation produces the output sub-state perturbation. The succession of overlapping SSP-MM-SSP triplets represents perturbation propagation across stages of biological organization (DNA, RNA, Protein, Complex, Organelle, Cell, Tissue, Organ, and Phenotype), and together form a mechanism schema. In MecCog, a schema is represented as a graph where the nodes are SSPs and edge labels are MMs, as illustrated in Figure 1.

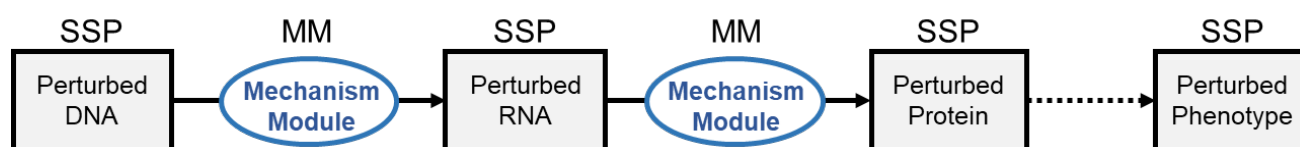


Figure 1. Principles of a mechanism schema. SSP: Substate Perturbation; MM: Mechanism Module. Each SSP represents a perturbed biological entity and each MM represents a productive activity (or a group of entities and activities) that produce an output SSP.

Evidence about SSPs and MMs is curated from the literature. Ambiguities in a mechanism and possible alternative mechanisms are represented in a schema by branching. Branch points may be labeled with the logical operators AND, OR, or AND/OR. The degree of confidence as to whether each SSP and MM is part of a schema is indicated by an evidence strength color code (red least confidence to green most confident) for the corresponding symbol. In addition to SSPs and MMs, five other types of mechanism components are defined in MecCog: 1. *Unknown mechanism module* to represent ignorance about a mechanism component; 2. *Biomarker* to represent entities correlated with a disease phenotype; 3. *Environmental factor* to represent relevant external factors; 4. *Hypothetical therapeutic intervention site*; 5. *Known therapeutic intervention site*.

MecCog platform web-architecture

Figure 2 shows the web-architecture of MecCog. On the server-side, Node.js (an open-source JavaScript runtime environment) is used as the web-server, Sails.js is used to build the model-view-controller compliant web-application, and a MySQL relational database is used to store data on users and mechanism schemas. The MySQL database is connected to the web-application by the Object-relational mapper (ORM), Waterline, in Sails and all the database transactions use REST APIs secured by CSRF (Cross-site request forgery). The front-end GUI of MecCog is implemented using HTML, CSS, and Javascript, and is made responsive by Bootstrap.js javascript library. The schema building and visualizing GUI is powered by the Rappid Diagramming Javascript library (<https://www.jointjs.com/>). Rappid also provides a feature for converting diagrams to JSON format and for communicating with the

database via AJAX requests. An open-source version of the IntenseDebate commenting system

(<https://www.intensedebate.com/>) is used to render commenting forms.

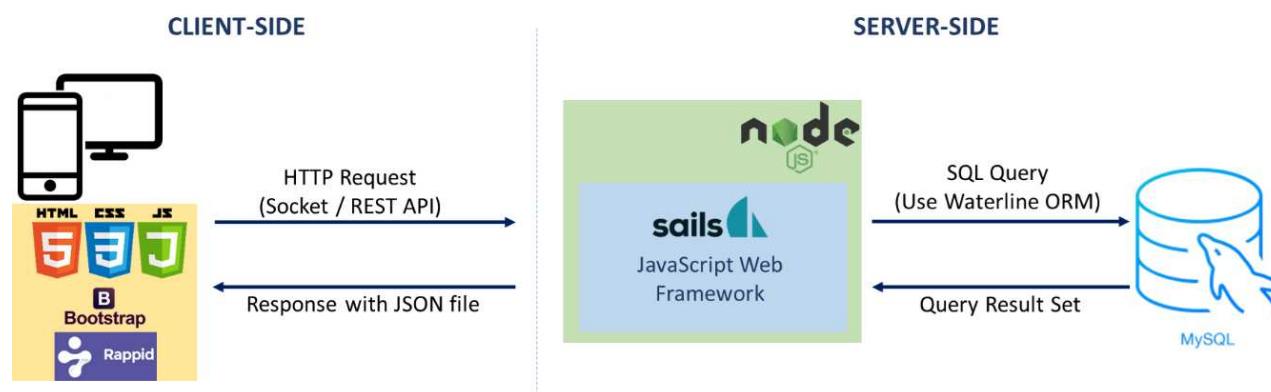


Figure 2. MecCog Web-Application Architecture. HTTP=Hypertext Transfer Protocol, REST API = Representational State Transfer Application Programming Interface, JSON=JavaScript Object Notation, SQL = Structured Query Language, ORM= Object-relational mapping.

Graphical notation of mechanism components in MecCog

Graphical notations symbolize components of a mechanism schema (Figure 3). An SSP (substate perturbation) is represented by a rectangle containing three types of information – the biological stage where the SSP occurs, the perturbation class name (curated from standard biomedical ontologies wherever possible), and the instance of that perturbation class. For example, a truncated *NOD2* protein can be represented by an SSP with *Protein* as the stage name, *Truncated Protein* as the SSP perturbation class name (from BioAssay Ontology (Visser *et al.*, 2011)), and *NOD2:1007fs* as the instance of the SSP class. A biomarker is a special case of an SSP and is represented by the same shape but with a different color. An environmental factor is represented by a cloud icon. Known and hypothetical therapeutic intervention sites are represented by pink and blue octagons respectively. A known mechanism module is represented by a clear oval displaying the MM class or instance name, such as transcription or protein folding. An unknown mechanism module is represented by a black oval.

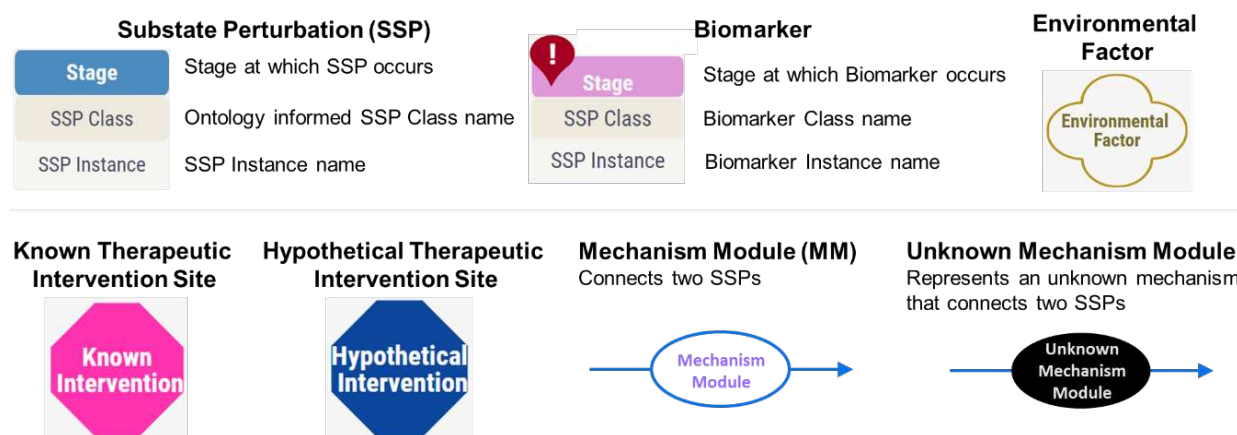


Figure 3. Graphical notations for components in a mechanism schema.

Mechanism schema meta-information and schema component annotations in MecCog

Table 1 summarizes the mechanism schema data model. Table 1A shows the meta-information of a schema. Each schema in MecCog is identified by a unique accession number automatically generated by the platform. Schema authors provide a schema name, a schema caption, genes that are part of the schema, keywords relevant to the mechanism, names of authors who constructed the schema, and the name of the curator who publishes the schema, monitors comments, and approves changes. Authors also provide a schema description with scientific background information.

Table 1B shows the mechanism component annotations. All mechanism components in a schema are annotated with a unique component identifier. Nine stage names may be assigned to an SSP component notation – DNA, RNA, Protein, Complex, Organelle, Cell, Tissue, Organ, and Phenotype. For the molecular stages (DNA, RNA Protein, and Complex), a set of stage-specific SSP perturbation class names, such as SNV, mRNA abundance, or protein stability, have been compiled. Molecular stage MM classes, such as transcription, translation, and protein folding, are also defined (Table 2). Whenever possible the SSP and MM class names are curated from existing biomedical ontologies. Currently used

ontologies are listed in Table 2. Where required, ontology terms may be prefixed with a modifier – increased, decreased, or altered. A MecCog schema builder may choose from the curated set of classes for a step, or may add new class names if needed. SSP and MM instance names are in free text. We are in the process of developing a disease mechanism ontology, based on the class names. Such an ontology is potentially useful for automatic text mining of SSP, MM, and triplet information from the literature, so speeding schema building by identifying relevant papers and sections of papers. Environmental factor names are in free text. Therapeutic intervention site components may be annotated with a potential therapy name or known drug name.

For all mechanism components, five types of evidence annotation are defined (Table 1C): 1. *For Evidence PubMed IDs* of papers that contain data supporting a component’s role in a mechanism; 2. *Against Evidence PubMed IDs* for papers that provide data suggesting a mechanism component is incorrect; 3. *Links to figures in PMC* that illustrate aspects of a schema by summarizing experimental results and evidence for spatial and structural features; 4. User assigned *Confidence scores* with five levels (also used to automatically encode a component’s confidence color) based on the strength of the available evidence; 5. *Evidence Comments* - brief free text comments that summarize the evidence.

Table 1. Data Model of mechanism schema and component annotations. Text in parentheses indicates the data type.

Table 1A. Mechanism schema meta-information

Accession number (<i>Automatically generated, versioned, alphanumeric</i>)
Schema Name (<i>Free text – 300 character limit</i>)
Schema Caption (<i>Free text – 500 character limit</i>)
Schema Description (<i>Free text</i>)
Genes (<i>Free text</i>)
Keywords (<i>Free text</i>)
Curators (<i>Free text</i>)

Authors (<i>Free text</i>)

Table 1B. Mechanism component annotations

Mechanism Component	Component Specific Annotation
SSP (Substate Perturbation)	Component ID (Format: SSP#)
	Stage Name (<i>Predefined list</i>)
	SSP Class Name (<i>Predefined list with the facility to add new names</i>)
	SSP Instance Name (<i>Free text</i>)
MM (Mechanism Module)	Component ID (Format: MM#)
	MM Class Name (<i>Predefined list with the facility to add new names</i>)
	MM Instance Name (<i>Free text; Optional</i>)
Biomarker	Component ID (Format: BM#)
	Stage Name (<i>Predefined list</i>)
	Biomarker Class Name (<i>Predefined list with the facility to add new names</i>)
	Biomarker Instance Name (<i>Free text</i>)
Unknown Mechanism Module	Component ID (Format: MM#)
Environmental factor	Component ID (Format: EF#)
	Factor name (<i>Free text</i>)
Known therapeutic intervention site	Component ID (Format: TT#)
	Drug or Therapy name (<i>Free text</i>)
Hypothetical therapeutic intervention site	Component ID (Format: TT#)
	Potential therapy name (<i>Free text</i>)

#' represents an integer number denoting the order of the schema component (for example SSP1, MM3, BM1, TT2)

Table 1C. Evidence annotations for the mechanism components

For Evidence PubMed IDs (<i>PMID number</i>)
Against Evidence PubMed IDs (<i>PMID number</i>)
Figure links from PubMed Central (<i>PMC figure number</i>)
Confidence score (<i>Predefined integer score range 1 to 5</i>)
Evidence comment (<i>Free text</i>)

Table 2. Curated class names for Substate Perturbations (SSP) and Mechanism Modules (MM) at the molecular stages. The class names are curated from biomedical ontologies and are prefixed with the ontology name abbreviation.

Mechanism Component	Stage Name	Number of Classes	Class Names
SSP	DNA	10	SO:SNV, NCIT:IN/DEL, NCIT:Methylation Sites, NCIT:Chromosomal Rearrangement, VARIO:Copy Number Variation, DNA Repeats, NCIT:DNA

			Structure, NCIT :Holliday Junction, DNA Supercoiling, DNA Curvature
	RNA	12	SO :SNV, NCIT :IN/DEL, VARIO :Edited RNA, Fused mRNAs, RNA Repeats, mRNA Abundance, Splicing Isoform Abundance, Edited RNA Abundance, Other RNA Abundance, EDAM :RNA Secondary Structure
	Protein	16	NCIT :Missense Mutation, NCIT :IN/DEL, BAO :Truncated Protein, NCIT :Post-Translational Modification Site(s), NCIT :Phosphorylation Site(s), Fused Proteins, Protein Sequence Repeats, PLOSTHES :Protein Abundance, Splicing Isoform Abundance, Post-Translational Modified Protein Abundance, BAO :Phosphorylated Protein Abundance, MESH :Protein Conformation, NCIT :Protein Dynamics, MESH :Protein Stability, MI :Allostery, MESH :Quaternary Protein Structure
	Complex	12	GRO :Protein-RNA Complex Abundance, CRISP :Spliceosome Abundance, DNA-RNA Complex Abundance, BIPON :RNA-RNA Complex Abundance, GO :Protein-DNA Complex Abundance, Transcription Complex Abundance, GO :Transcription Factor Complex Abundance, DNA-Scaffold Complex Abundance, DNA Replication Complex Abundance, DNA-Histone Complex Abundance, EDAM :Protein-Ligand Complex Abundance, ADMO :Protein-Protein Complex Abundance.
MM	-	24	IXNO :Cleavage Rate, NCIT :Synthesis Rate, GO :Transport Rate, CRISP :Protein Degradation, NCIT :RNA Degradation Rate, NCIT :Protein Folding Rate, NCIT :Nonsense-Mediated Decay Rate, NCIT :Transcription Rate, GO :Translation Rate, DNA Internal Interactions, RNA Internal Interactions, Protein Internal Interactions, DNA-RNA Interaction, RNA-Ligand Interaction, IOBC :RNA-RNA Interaction, GRO :DNA Protein Interaction, GO :Scaffold Protein Binding, GO :Basal Transcriptional Machinery Binding, GO :Histone Binding, NCIT :RNA-Protein Interaction, NCIT :Protein-Protein Interaction, NCIT :Ligand Binding, Le Chatelier, GO :Signaling.

SO: Sequence Ontology (Eilbeck *et al.*, 2005); *NCIT: National Cancer Institute Thesaurus* (Sioutos *et al.*, 2007); *VARIO: Variation Ontology* (Vihinen, 2014); *EDAM: EMBRACE Data and Methods* (Ison *et al.*, 2013), *MESH: Medical Subject Headings* (<https://meshb.nlm.nih.gov/>), *MI: Molecular Interactions* (<https://www.ebi.ac.uk/ols/ontologies/mi>), *CRISP: Computer Retrieval of Information on Scientific Projects Thesaurus* (<https://bioportal.bioontology.org/ontologies/CRISP>), *GRO: Gene Regulation Ontology* (<https://bioportal.bioontology.org/ontologies/GRO>), *BIPON: Bacterial interlocked Process Ontology* (Henry *et al.*, 2017), *GO: Gene Ontology* (The Gene Ontology Consortium, 2019), *ADMO: Alzheimer Disease Map Ontology* (Malhotra *et al.*, 2014), *PLOSTHES: PLOS Thesaurus* (<https://bioportal.bioontology.org/ontologies/PLOSTHES>), *BAO: BioAssay Ontology* (Visser *et al.*, 2011), *IXNO: Interaction Ontology* (<https://bioportal.bioontology.org/ontologies/IXNO>), *IOBC: Interlinking Ontology for Biological Concepts* (<https://bioportal.bioontology.org/ontologies/IOBC>).

Rules for constructing mechanism schemas in MecCog

1. Each schema begins with a genome perturbation and ends in perturbation of a disease-related phenotype such as greater risk of a disease.
2. Overall, the sequence of SSPs in a schema progresses through successive stages of biological organization, from DNA, through RNA, proteins, macromolecular complexes, organelles, cells, tissues, organs, and finally to a phenotype. There may be one or more or no SSP at any particular stage of organization and the order of the stages need not follow a prescribed order. For instance, the schema for Lynch syndrome (<http://www.meccog.org/mchain/showpubchain?accession=MS020700047.3>), where a causative mutation results in decreased DNA mismatch repair, reverts to the DNA stage after stages involving macromolecular complexes.

3. Each pair of SSPs is linked by an MM. The granularity of an MM may be a single activity (such as splicing, protein-protein interaction, ligand binding, or protein folding) or may represent telescoped combinations of entities and activities (such as protein synthesis, or cell-cell signaling). If an activity is unknown, the black oval unknown mechanism module notation is used.

4. Class names for the SSPs and MMs at the molecular stages (DNA, RNA, Protein, and Complex) can be selected from the pre-compiled list shown in Table 2. If existing names are inadequate, new names can be used. At higher organizational stages, class names are user-provided. Wherever possible these should be part of existing biomedical ontologies. The NCBO BioPortal site

(<https://bioportal.bioontology.org/>) is a source for ontology terms. Since most ontologies describe the normal state of a system, a user may select one of the in-built modifiers (increased, decreased, altered) to prefix a class name so as to represent a perturbed state.

5. An evidence-based confidence score (on the scale of 1 to 5, where one indicates low confidence and five indicates high confidence) should be assigned to each SSP and MM. Evidence on which a confidence score is based should be recorded in the form of supporting/contradicting PMIDs and PMC figure URLs, together with appropriate free text commentary.

6. Two or more possible sub-paths can exist in a schema either because of ambiguity due to conflicting evidence, or alternative sub-mechanisms. Branch points should be labeled with OR, AND or AND/OR.

7. Schemas should explicitly include steps only where there is a perturbation from the normal system.

Where the function of a portion of a schema is unperturbed, for example, representing the standard activity of transcription operating on a perturbed input DNA sequence or a standard cell signaling process operates with more or less input signal, that section of the schema should be telescoped into a single mechanism module.

Steps in constructing, managing, and publishing mechanism schemas

Before beginning schema building a new user must register on the MecCog platform. A registered user may select the “Build Schema” tab to initiate building a new schema or the “My Schemas” tab to access the workspace for managing and editing their existing schemas. Figures 4A and 4B show the two interfaces used in schema construction: A. The *Initiate Mechanism Schema* form used to enter meta-information about a schema, and B. The *Schema Builder* GUI used to draw a schema. In the schema builder, mechanism components can be dragged and dropped from the mechanism component catalog panel to the drawing board panel. Clicking on a component displays five associated control icons: i) Icon to connect to other components; ii) Icon to adjust the component size; iii) Icon to clone the component; iv) Icon to show the pop-up box; and v) Icon to delete the component. Clicking on a component also renders a component-specific annotation form on the rightmost panel of the interface (labeled in Figure 4B). This form is used to enter the stage, class, and instance name of the components, prefix class names with a perturbation type if needed (increased, decreased, or altered), and record the evidence annotations (listed in the Evidence Annotations Table 1C). This is a dynamic form that automatically provides predefined possible perturbation class names for the selected stage (listed in Table 2) and creates fields for adding new PubMed IDs and PMC image URLs. The NCBI E-utilities application programming interface (API) is used on the server-side to fetch publication details for the PubMed IDs. All the evidence annotations are transferred to the current component pop-up box together with hyperlinked PMIDs and PMC image URLs. The pop-up box can be visualized by clicking the ‘*z*’ shaped control icon of the component. Confidence score values selected in the annotation form are used to automatically apply the appropriate color to the current schema component (red: score 1, orange: scores 2, 3, 4 and, green: score 5). The color of the edge connecting two components is inherited from the target component, so indicating causal confidence. Schemas are saved to the database using the *Click to*

save button. For each schema, a unique accession number is automatically generated in the database.

The accession number format has a section indicating the version (default is .1) of the schema. The schema builder GUI also has a panel of interactive buttons to undo, redo, clear page, zoom, auto-layout, export (in SVG and PNG formats), and print schema diagrams.

Figure 4C shows the view of a registered user's workspace. Each schema can be versioned, edited, shared with other MecCog users, published, or deleted, using operation-specific buttons. The workspace has three sections: i) The Unpublished Mechanism Schemas section catalogs work-in-progress schemas. ii) The Published Mechanism Schemas section catalogs published schemas. A button to remove each of these from the public collection is provided. iii) The Shared Mechanism Schemas section catalogs schemas that have been shared with the current user. For schemas with edit access privilege, the *Copy to My Space* operation is enabled, allowing the creation of a copy for the user to work on independently. All the schema accession numbers in the workspace page are hyperlinked to the schema specific landing page (described in the next section).


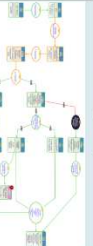

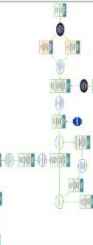
Published schemas are available for browsing via the main webpage of the MecCog site (as shown in Figure 4D) without the need for logging in. There is a search bar that allows schemas to be searched by gene name, keyword, or any component class/instance name. MYSQL's FULLTEXT indexing (<https://dev.mysql.com/doc/refman/5.6/en/innodb-fulltext-index.html>) feature is used to support the search operation. The structured organization of mechanistic knowledge in MecCog allows this search to be used to find common entities or activities and common classes used in different schemas. On the main page, schemas are presented in a masonry layout view. Each tile in the view displays the schema name, schema caption, hyperlinked accession number of the schema linking to the corresponding

Initiate Mechanism Schema

Unpublished Mechanism

Published Mechanism

Shared Mechanism

<p>NOD2 Schema for Crohn's disease</p> <p>MS031100031.9 2020-8-2 0:51</p> <p>Mechanism by which NOD2 1007fs variant causes increased Crohn's disease risk</p>  <p>The diagram illustrates the NOD2 signaling pathway. It shows the entry of a ligand (represented by a black dot) into the cell, which triggers a cascade of events. Key components include NOD2 (a green box), which is mutated (fs) in this context. The pathway involves several other proteins (blue and yellow boxes) and leads to the activation of NF-κB (a red box), which then promotes the production of inflammatory mediators (represented by red dots).</p>	<p>MSH2 Schema for Lynch syndrome</p> <p>MS020700047.2 2020-5-14 1:37</p> <p>Association of a truncating variant (rs63750245) in MSH2 gene with increased risk for Lynch syndrome.</p>  <p>The diagram shows the MSH2 protein (green box) interacting with other proteins (blue and yellow boxes) in a DNA mismatch repair pathway. A truncating variant (rs63750245) is indicated, leading to a loss of function of MSH2. This results in a failure to repair DNA mismatches, which is associated with an increased risk of Lynch syndrome.</p>	<p>CFTR F508del for Cystic Fibrosis</p> <p>MS030300006.3 2020-2-9 11:20</p> <p>The role of F508del variant of the gene CFTR contributing to the Cystic Fibrosis disease mechanism.</p>  <p>The diagram depicts the CFTR protein (green box) and its role in the cell. The F508del variant is shown, which leads to a misfolded and dysfunctional CFTR protein. This results in a failure to regulate ion transport across the cell membrane, leading to the disease mechanism of Cystic Fibrosis.</p>	<p>MS020600026.4 2019-1-21 17:09</p> <p>Relationship between Crohn's risk genetic variants and Mageritens (drug) effectiveness</p>  <p>The diagram shows a complex network of genetic variants (represented by colored dots) and their interactions with the Mageritens drug (represented by a red box). The variants are shown to influence the effectiveness of the drug in treating Crohn's disease.</p>
--	---	--	---

Mechanism Component

Drawing Board

Annotation Form

The screenshot displays the Schema Builder tool interface. On the left, a sidebar contains several categories of components: **ENVIRONMENTAL FACTOR** (with a cloud icon), **PREBIOTIC TARGET** (with a blue circle icon), **MECHANISM MODULE** (with a blue circle icon), **PROBIOTIC** (with a pink hexagon icon), and **PROBIOGENIC TARGET** (with a blue circle icon). The main workspace shows a workflow diagram with three stages: **Stage** (containing 'SSP Class' and 'SSP Instance'), **Mechanism Module** (a central blue circle), and **Protein** (containing 'Turned Protein' and 'SSP Instance'). Arrows indicate the flow from Stage to Mechanism Module and from Mechanism Module to Protein. On the right, a sidebar contains a search bar, a list of components (e.g., 'Enter MLC Component ID', 'Select a Mechanism Module', 'Select a Protein'), and a 'Save' button. The bottom of the interface shows a 'Footer' section with the text 'Page 10'.

Figure 4. Graphical User Interface (GUI) to construct, manage, and browse mechanism schemas. Figure 4A shows the form for entering meta-information on a schema. Figure 4B shows the schema builder interface, including the mechanism component catalog panel (left), the drawing board panel, and the annotation form panel (right). Figure 4C shows the user workspace interface. Figure 4D shows a portion of the MecCog main webpage, designed to facilitate browsing publicly available mechanism schemas.

schema landing page (described in the next section), and a hyperlinked schema image linked to an interactive web-based visualization of the schema (also described in the next section).

Schema landing page, schema visualizer, and schema report

A schema landing page displays schema meta-information in a tabular layout (Figure 5A). A novel feature of this page is the display of references and hyper-linked PubMed IDs providing evidence for each aspect of the schema as well as PMC images selected to illustrate aspects of the mechanism. The *Schema Visualizer* button on the landing page directs a user to the GUI for interactively navigating the mechanism schema (as shown in Figure 5B). The visualizer inherits all the interactive features of the schema builder GUI (described previously). A unique feature of the visualizer is the tight integration of the graphical notations for the mechanism components and the associated evidence information (presented in the pop-up box). The pop-up box (yellow-colored box in Figure 5B) displays hyperlinked ontology sources for the SSP/MM class name (if the term is from an ontology), a brief builder-provided commentary on the evidence, and hyperlinked PMIDs and PMC figure IDs. There is a help icon (“?”) in the visualizer to display the mechanism schema key. Clicking on the Comment button in the visualizer opens a modal box to view or enter comments about the schema. The *Schema Report* button on the landing page generates a narrative report in which the meta-information, mechanism components, and evidence annotations about the schema are presented in a structured format. The schema content can be

Figure 5. *NOD2* mechanism schema entry in MecCog. This schema describes the known mechanisms by which a frameshift mutation (rs2066847) in the *NOD2* gene causes an increased risk for Crohn's disease. Figure 5A shows part of the landing page of the *NOD2* schema, displaying the meta-information in tabular format. This page includes the collection of thumbnails of PMC figures selected to illustrate aspects of the mechanism and the list of references with PubMed IDs from which evidence was derived (the list is truncated here - there are 26 references). Figure 5B shows the schema visualizer GUI used for interactive navigation of the schema. For this schema, four possible submechanisms with varying levels of evidence (indicated by the confidence colors – red=low, orange=medium, and green=high) are included. The example yellow pop-up box displays hyperlinked evidence for the associated MM. The comment button on the top can be used to open a modal box, allowing a user to view and add comments. Details of the mechanism are described in the text.

downloaded as a JSON file from the landing page using the download icon. The page also has social media share icons.

An example MecCog Schema: Known mechanisms by which a frameshift mutation in the *NOD2* gene causes an increased risk of Crohn's disease

Figure 5 shows two pages of a MecCog mechanism schema

(<http://www.meccog.org/mchain/showpubchain?accession=MS031100031.9>) describing the mechanism by which a frameshift mutation (rs2066847; NM_022162.3:c.3019dup (p.Leu1007fs)) in the *NOD2* gene causes an increased risk of Crohn's disease (CD). *NOD2* is the first gene for which variants were found to be associated with altered CD risk (Ogura *et al.*, 2001; Hugot *et al.*, 2001; Yamamoto and Ma, 2009) and the 1007fs mutation is most consistently associated with CD across multiple studies and population groups (Economou *et al.*, 2004) with a very high relative risk of 17.6 for its homozygous genotype status as compared with wild-type controls (Ogura *et al.*, 2001). Figure 5A shows the landing page displaying the meta-information of the schema and the list of references used as evidence in the schema, together with figures used to illustrate aspects of the mechanism.

Figure 5B shows the view of the NOD2 1007fs schema in the interactive visualizer. This schema was constructed using information about mechanism reported in 26 research articles. The left-most SSP (SSP1) represents the DNA stage perturbation (i.e. the single base insertion of cytosine - rs2066847 in the *NOD2* gene). The paths in the schema show how the effect of this perturbation propagates through the RNA, protein, complex, and cell stages (represented by the stage-specific SSPs and MMs) so causing the increased Crohn's disease risk phenotype (SSP12). At the RNA stage (SSP2), the rs2066847 variant causes the insertion of a cytosine after the first nucleotide of codon 1007, so introducing a premature downstream stop codon. This leads to the protein stage perturbation, a truncated NOD2 protein (SSP3) missing the last 33 amino acids of the wild-type sequence (Lécine *et al.*, 2007). Following this, the schema branches represent the multiple submechanisms by which the truncated NOD2 protein may lead to the increased Crohn's disease risk phenotype by altering the activation of the immune response (MM10) (Negroni *et al.*, 2018; Strober and Watanabe, 2011; Park *et al.*, 2007). All the branches are labeled 'AND/OR' since none has fully compelling supporting evidence. The submechanism of each branch is outlined below.

A) The top branch (SSP3→MM3 → SSP4) shows a potential alteration to NOD2-dependent regulation of Toll-like receptor (TLR) mediated NF-κB signaling that produces pro-inflammatory cytokines in response to the pathogen-associated molecular patterns (PAMPs) such as lipopolysaccharide (LPS), or muramyl dipeptide (MDP). This path is sparse and labeled medium confidence (orange) because the mechanism of interaction between NOD2 and TLRs is not known, nor is it clear how that interaction normally results in increased production of pro-inflammatory cytokines (Underhill, 2007). Different models have been proposed to describe the mechanism: synergistic production of TNF-α by NOD2 and TLR4 (Wolfert *et al.*, 2002); activation of the inflammasome by NOD2 via RICK to produce IL-1β from

pro-IL-1 β generated as the result of TLR signaling (Sarkar *et al.*, 2006); and MDP (the primary agonist for NOD2 (Grimes *et al.*, 2012)) dose-dependent TNF- α production by NOD2 and TLR2 (Borm *et al.*, 2008). Further, for none of these possibilities has the effect of the NOD2 100fs variant been investigated. These details are provided in the pop-up box for MM3.

B) The middle branch shows that truncated NOD2 protein has lost its ability to localize to the plasma membrane (MM4 \rightarrow SSP5) (Barnich *et al.*, 2005; Morosky *et al.*, 2011) where binding to incoming MDP normally produces an activated state of the protein (Al Nabhani *et al.*, 2017). In turn, activated NOD2 forms complexes with RICK and with ATG16L1 (Barnich *et al.*, 2005; Travassos *et al.*, 2010). The schema shows these effects as lower abundance of the NOD2-RICK complex (MM5 \rightarrow SSP6) (Barnich *et al.*, 2005) and the NOD2-ATG16L1 complex (MM5 \rightarrow SSP8) (Travassos *et al.*, 2010). There is no experimental evidence of the NOD2 1007fs protein's impact on complex formation. Therefore the MM5 \rightarrow SSP6 step in the schema is labeled medium confidence (orange). Following this step, the lower abundance of the NOD2-RICK complex alters downstream NF- κ B signaling (SSP6 \rightarrow MM6 \rightarrow SSP7) (Caruso *et al.*, 2014; Lécine *et al.*, 2007; Barnich *et al.*, 2005; Girardin *et al.*, 2003), resulting in lower pro-inflammatory cytokine production, so contributing to an altered activation of the immune response (MM10) (Vilela *et al.*, 2012; Park *et al.*, 2007; Strober and Watanabe, 2011; Negroni *et al.*, 2018). The perturbation of the NOD2-ATG16L1 complex affects the xenophagy process (autophagy against bacteria) (MM7) (Travassos *et al.*, 2010) so leading to an increase in the abundance of bacteria in the lamina propria (SSP9) (Sidiq *et al.*, 2016) and thereby likely contributing to a more aggressive response from other components of the immune system, as indicated by the altered activation of immune response (MM10). This sub-path is labeled high confidence (green) as its mechanism components are well understood based on the available evidence in the literature. The yellow pop-up

box for MM4 shows an example of an evidence commentary with an associated hyperlinked PMC figure and PMIDs.

C) The lower branch of the schema provides examples of the representation of a gap in knowledge and of overall low confidence. Commensal bacteria are largely prevented from penetrating the gut wall by an outer mucosal barrier and the epithelial cell layer. Paneth cells situated in the gut epithelial layer produce a range of antibiotic defensin peptides to aid in preventing commensal bacteria from traversing the mucosal layer. Some data suggest that this process is partly dependent on MDP binding to NOD2 in these cells, likely signaling that significant numbers of bacteria are getting through to the epithelial cell layer, and so triggering an increased response. Data supporting that view come from an experiment showing stimulation of NOD2 by MDP binding induces production of defensin HNP-1 (human neutrophil peptide 1) in Caco-2 cells (Yamamoto-Furusho *et al.*, 2010). It has also been shown that the NOD2 1007fs protein fails to induce the production of defensin hBD2 (human β -defensin-2) in several epithelial cell lines (Voss *et al.*, 2006). Hence the link between the presence of the 1007fs variant (SSP3) and increased defensin production (currently SSP10). But the mechanism by which MDP binding to NOD2 normally causes defensin production is unknown, hence the black oval (MM11) linking those two SSPs. There is also evidence from other studies that do not support the mechanism represented by this schema path: In two out of four CD cohort studies (Wehkamp *et al.*, 2005; Simms *et al.*, 2008; Hayashi *et al.*, 2016), and in NOD2 deficient mouse organoids (Wilson *et al.*, 2015), the association between NOD2 and defensin was not reproduced. Hence this branch is labeled low confidence (red). Further along this schema branch, the decrease in defensin production (SSP10) leads to an increased abundance of bacteria in the mucosal layer (SSP11) due to decreased bactericidal activity (MM8). In turn, this contributes to increased bacterial abundance in the lamina propria due to increased bacterial

influx from the mucosal layer (MM9) and finally leads to the altered activation of the immune response (MM10).

Representation of biomarker and therapeutic intervention sites in MecCog

Figure 6A shows an example of the use of the biomarker symbol, in part of the Lynch syndrome schema (<http://www.meccog.org/mchain/showpubchain?accession=MS020700047.3>). In this schema, microsatellite instability is a diagnostic biomarker (Vilar *et al.*, 2014) for Lynch syndrome, resulting from defective base mismatch repair machinery, in turn a consequence of a mutation (rs63750245: C>T) in the *MSH2* gene. Figure 6B shows an example of a putative therapeutic intervention site in a Crohn's disease schema (<http://www.meccog.org/mchain/showpubchain?accession=MS020500019.2>) describing the mechanism by which a missense variant (rs3197999: G>A; R703C) in the *MST1* gene (coding for Macrophage Stimulating Protein, MSP) increases disease risk. The missense variant causes a lower abundance of the MSP-RON protein complex by one or both of two mechanisms: a weakened protein-protein interaction (Chao *et al.*, 2014; Gorlatova *et al.*, 2011) and reduced MSP abundance. Lower abundance of the complex is expected to result in reduced intracellular signaling affecting macrophage activation (Wang *et al.*, 2002; L. Kretschmann *et al.*, 2010; Häuser *et al.*, 2012) and/or epithelial cell survival and growth (Danilkevitch *et al.*, 2000; Neurath, 2014). An appropriate compound that bridges the structural interface between MSP and RON could restore wild-type abundance of the complex and hence signaling strength and so eliminate the downstream consequences. (Of course, many factors affect whether this is in fact an effective therapeutic strategy.)

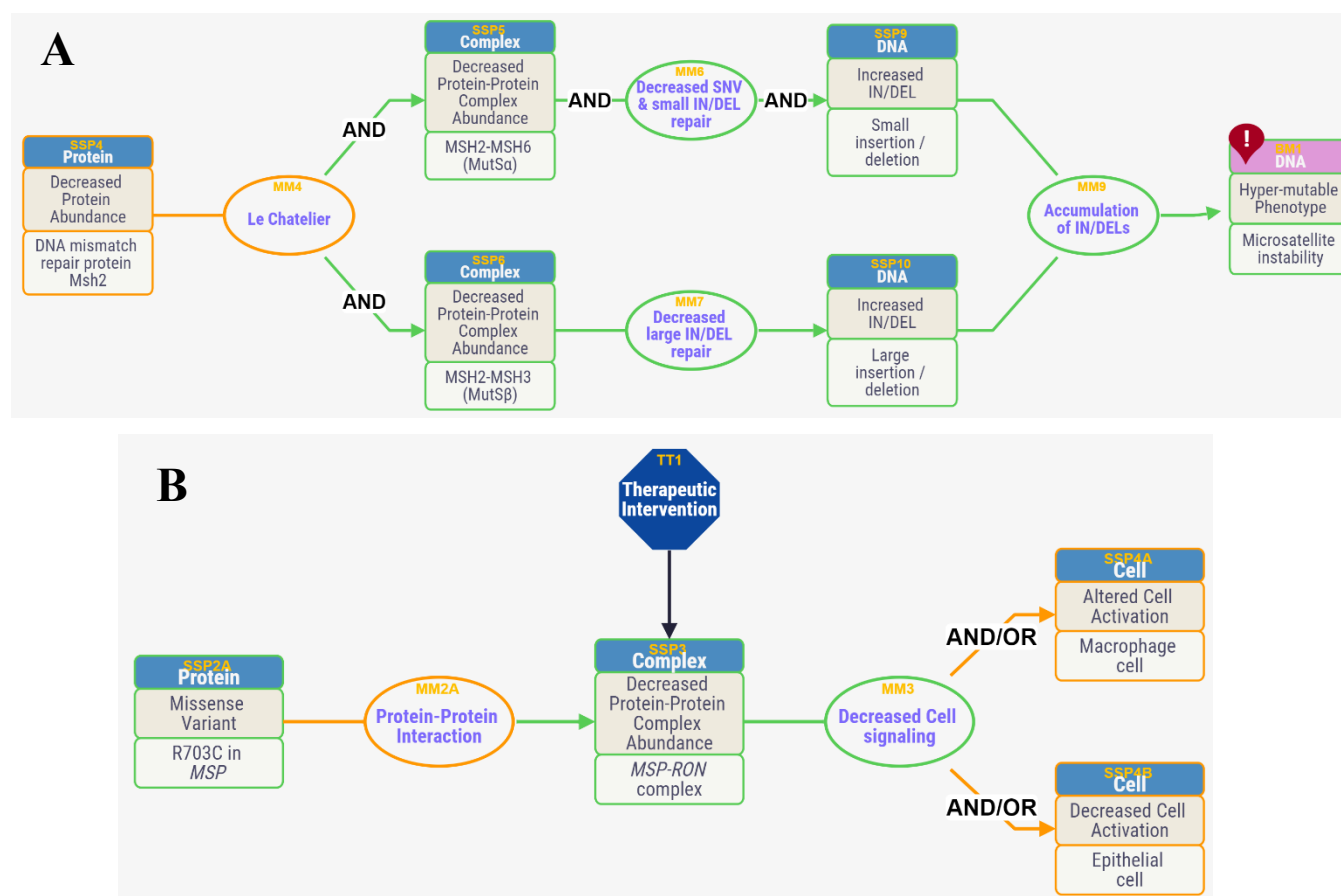


Figure 6. Biomarker and Therapeutic intervention site representation in MecCog. Figure 7A shows part of a Lynch syndrome schema (<http://www.meccog.org/mchain/showpubchain?accession=MS020700047.3>) where the presence of a nonsense mutation (rs63750245: C>T) in the MSH2 gene causes Microsatellite instability (MSI), a known biomarker (symbolized by the red location icon on the SSP) for the Lynch syndrome. Figure 7B shows part of a Crohn's disease schema (<http://www.meccog.org/mchain/showpubchain?accession=MS020500019.2>) where the decreased abundance of the MSP-RON protein complex (SSP3) is a hypothetical therapeutic intervention site, indicated by the blue octagon. In this case, an appropriate small molecule binding across the protein-protein interface might restore the wild-type abundance.

Validation of the MecCog representation framework

Eleven MecCog mechanism schemas (nine for Crohn's disease, one for cystic fibrosis, and one for Lynch syndrome) have so far been published, with additional schemas in progress on breast cancer and Alzheimer's disease. Validation and improvement of MecCog content is obtained by soliciting feedback

from specialists in the disease described in each schema. Feedback on the representation technique and platform can be provided using MecCog's *contact us* form, encouraging users to provide suggestions and report problems. MecCog has also been used as an educational tool for senior undergraduate students in a Human Genetics class at the University of Maryland, providing valuable feedback, for example, linking PMC figures to aspects of schemas.

Discussion

We have developed MecCog, a graphical knowledge representation framework, to describe genetic disease mechanisms in a structured mechanism schema format. MecCog facilitates the assembly of mechanistic information in terms of perturbation propagation across stages of biological organization, evaluation of the evidence related to that information, and identification of uncertainties, ambiguities, and ignorance. The MecCog web platform provides functionalities to create, store, browse, and search schemas. Graphical notations are annotated with ontology-informed class terms so as to consistently and intuitively represent types of mechanism components found in schemas. The schema interactive visualizer in MecCog tightly integrates the graphics, text, and hyperlinks to evidence sources.

Each schema in MecCog describes mechanisms by which a single genetic variant contributes to the increased risk for the disease phenotype. For complex trait genetic disease and cancer, multiple genetic variants contribute to disease phenotypes (Peter *et al.*, 2011; Lilyquist *et al.*, 2018). Further, contributions from variants may not be independent, as reflected by evidence of epistatic effects between pairs of variants for complex trait disease (Lin *et al.*, 2017; Li *et al.*, 2020). The MecCog formalism also supports mechanism schemas with multiple input genetic perturbations. Interactions between these inputs results in a mechanism graph. An example for Crohn's disease is a barrier integrity mechanism

graph constructed by combining schemas on loci relevant to bacterial penetration of the gut-lining mucosal layer (Figure S1). This graph incorporates a number of non-additive interactions between mucin gene variants affecting mucosal-layer integrity (*MUC1*, *MUC2*), variants affecting the unfolded protein response (*XBPI*, *ORMDL3*), and variants affecting autophagy (*NOD2*, *ATG16L1*, *LRRK2*, *IRGM*).

Currently, MecCog schemas are manually constructed, relying on human understanding to extract and infer causal connections between mechanism components from literature. Given the scattered and incomplete nature of mechanistic information in literature, this process is complex and requires a combination of prior biological knowledge together with searching for and assimilating new facts and evidence from the literature. These activities are labor-intensive and work best when the schema builder is an expert on the schema subject. To achieve scale for the resource, we require an expert-crowdsourcing strategy, soliciting inputs from appropriate domain experts. The resource is structured so that experts can build schemas based on their knowledge and can also edit and comment on existing schemas. The current version of the MecCog platform supports these activities in the following ways: i) acknowledging contributors to a schema as authors and curators, ii) providing a schema specific commenting interface to solicit input, and iii) allowing versioning of schemas to update content. To implement the crowdsourcing model, we will work closely with disease-specific research communities (such as IBD Genetics, Crohn's & Colitis Foundation, and Alzforum).

An obvious question is whether mechanism schemas can be constructed automatically given the structured and unstructured data available in the biomedical domain. The structure of a mechanism schema shares features with that of knowledge graphs (KG), a knowledge representation system initiated by Google in 2012 (<https://googleblog.blogspot.com/2012/05/introducing-knowledge-graph-things->

[not.html](#)). There nodes (aka subjects) represent entities such as real-world objects, events or concepts, and edges (aka predicates) link the nodes with relationships. Because of a KG's ability to integrate and represent multi-relational databases, many biological KGs (Sosa *et al.*, 2019; Celebi *et al.*, 2019; Chen *et al.*, 2019; Himmelstein *et al.*, 2017; PDBe-KB consortium, 2020; <https://digitalinsights.qiagen.com/coronavirus-network-explorer/>) are being generated, using a combination of manual and automated mining of subject-predicate-object (SPO) triplets from biomedical literature and from bioinformatics relational databases. The elemental SSP-MM-SSP units of a MecCog schema are a subset of SPO triplets and so it is in principle possible to construct a schema by extracting appropriate triplets from a comprehensive knowledge graph. However, preliminary tests of this process suggest that current knowledge graphs do not capture a large fraction of the triplets incorporated in the corresponding mechanism schemas. There are multiple reasons for this, including the absence of biological knowledge in knowledge graphs and the absence of causal reasoning components. We envisage that in the future comprehensive and well-structured KGs will be combined with a repository of biological knowledge and reasoning machines to generate a wide variety of biological mechanisms, as well as providing evaluation of evidence strength and identifying current gaps in mechanism knowledge.

Acknowledgments

This work was supported in part by the National Institute of Health [R01GM104436 to JM]. KK's conference travel related to this research was supported in part by NSF award DGE-1632976.

We thank Rappid (<https://www.jointjs.com/>) for providing their JavaScript library under an academic license. We thank Lipika Ray, Yizhou Yin, Maya Zuhl, Christian Presley, and undergraduate students in the University of Maryland Human Genetic course for many useful suggestions and comments on

MecCog software. We thank Mark Tonelli for feedback on the MecCog framework and for reviewing the cystic fibrosis schema.

References

- Barnich,N. *et al.* (2005) Membrane recruitment of NOD2 in intestinal epithelial cells is essential for nuclear factor-KB activation in muramyl dipeptide recognition. *J. Cell Biol.*, **170**, 21–26.
- Beltrame,L. *et al.* (2011) The Biological Connection Markup Language: a SBGN-compliant format for visualization, filtering and analysis of biological pathways. **27**, 2127–2133.
- Borm,M.E.A. *et al.* (2008) The effect of NOD2 activation on TLR2-mediated cytokine responses is dependent on activation dose and NOD2 genotype. *Genes Immun.*, **9**, 274–278.
- Caruso,R. *et al.* (2014) NOD1 and NOD2: Signaling, host defense, and inflammatory disease. *Immunity*, **41**, 898–908.
- Celebi,R. *et al.* (2019) Evaluation of knowledge graph embedding approaches for drug-drug interaction prediction in realistic settings. *BMC Bioinformatics*, **20**.
- Chao,K.L. *et al.* (2014) Structural Basis for the Binding Specificity of Human Recepteur d’Origine Nantais (RON) Receptor Tyrosine Kinase to Macrophage-stimulating Protein. *J. Biol. Chem.*, **289**, 29948–29960.
- Chen,I.Y. *et al.* (2019) Robustly Extracting Medical Knowledge from EHRs: A Case Study of Learning a Health Knowledge Graph. In, *Biocomputing 2020*. WORLD SCIENTIFIC, pp. 19–30.
- Cohen,P.R. (2015) DARPA’s Big Mechanism program. *Phys. Biol.*, **12**, 045008.
- Craver,C.F. and Darden,L. (2013) In Search of Mechanisms: Discoveries across the Life Sciences University of Chicago Press, Chicago, IL.
- Danilkovitch,A. *et al.* (2000) Two Independent Signaling Pathways Mediate the Antiapoptotic Action of

- Macrophage-Stimulating Protein on Epithelial Cells. *Mol. Cell. Biol.*, **20**, 2218–2227.
- Darden,L. *et al.* (2018) Harnessing formal concepts of biological mechanism to analyze human disease. *PLoS Comput. Biol.*, **14**.
- Economou,M. *et al.* (2004) Differential effects of NOD2 variants on Crohn’s disease risk and phenotype in diverse populations: A metaanalysis. *Am. J. Gastroenterol.*, **99**, 2393–2404.
- Eilbeck,K. *et al.* (2005) The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol.*, **6**.
- Fabregat,A. *et al.* (2017) The Reactome Pathway Knowledgebase. *Nucleic Acids Res.*, **46**, 649–655.
- Franz,M. *et al.* (2018) GeneMANIA update 2018. *Web Serv. issue Publ. online*, **46**.
- Girardin,S.E. *et al.* (2003) Nod2 is a general sensor of peptidoglycan through muramyl dipeptide (MDP) detection. *J. Biol. Chem.*, **278**, 8869–8872.
- Gorlatova,N. *et al.* (2011) Protein characterization of a candidate mechanism SNP for Crohn’s disease: The macrophage stimulating protein R689C substitution. *PLoS One*, **6**.
- Greenberg,S.A. and Amato,A.A. (2004) Uncertainties in the pathogenesis of adult dermatomyositis. *Curr. Opin. Neurol.*, **17**, 359–364.
- Griffith,M. *et al.* (2017) CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat. Genet.*, **49**, 170–174.
- Grimes,C.L. *et al.* (2012) The innate immune protein Nod2 binds directly to MDP, a bacterial cell wall fragment. *J. Am. Chem. Soc.*, **134**, 13535–13537.
- Häuser,F. *et al.* (2012) Macrophage-stimulating protein polymorphism rs3197999 is associated with a gain of function: Implications for inflammatory bowel disease. *Genes Immun.*, **13**, 321–327.
- Hayashi,R. *et al.* (2016) Reduced human α -defensin 6 in noninflamed jejunal tissue of patients with Crohn’s disease. *Inflamm. Bowel Dis.*, **22**, 1119–1128.

- Henry, V.J. *et al.* (2017) The bacterial interlocked process ONtology (BiPON): A systemic multi-scale unified representation of biological processes in prokaryotes. *J. Biomed. Semantics*, **8**.
- Himmelstein, D.S. *et al.* (2017) Systematic integration of biomedical knowledge prioritizes drugs for repurposing. *Elife*, **6**.
- Hucka, M. *et al.* (2018) The Systems Biology Markup Language (SBML): Language Specification for Level 3 Version 2 Core. *J. Integr. Bioinform.*, **15**.
- Hugot, J.P. *et al.* (2001) Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature*, **411**, 599–603.
- Ison, J. *et al.* (2013) EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. *Bioinformatics*, **29**, 1325–1332.
- Kametani, F. and Hasegawa, M. (2018) Reconsideration of amyloid hypothesis and tau hypothesis in Alzheimer's disease. *Front. Neurosci.*, **12**.
- Kanehisa, M. *et al.* (2016) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.*, **45**, 353–361.
- Kinoshita, J. and Clark, T. (2007) Alzforum. *Methods Mol. Biol.*, **401**, 365–381.
- Konopka, T. and Smedley, D. (2020) Incremental data integration for tracking genotype-disease associations. *PLoS Comput. Biol.*, **16**.
- L. Kretschmann, K. *et al.* (2010) The Macrophage Stimulating Protein/Ron Pathway as a Potential Therapeutic Target to Impede Multiple Mechanisms Involved in Breast Cancer Progression. *Curr. Drug Targets*, **11**, 1157–1168.
- Landrum, M.J. *et al.* (2018) ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.*, **46**.
- Lécine, P. *et al.* (2007) The NOD2-RICK complex signals from the plasma membrane. *J. Biol. Chem.*,

- 282**, 15197–15207.
- Li,Y. *et al.* (2020) Statistical and Functional Studies Identify Epistasis of Cardiovascular Risk Genomic Variants From Genome-Wide Association Studies. *J. Am. Heart Assoc.*, **9**, e014146.
- Lilyquist,J. *et al.* (2018) Common genetic variation and breast cancer Risk—Past, present, and future. *Cancer Epidemiol. Biomarkers Prev.*, **27**, 380–394.
- Lin,Z. *et al.* (2017) Genetic association and epistatic interaction of the interleukin-10 signaling pathway in pediatric inflammatory bowel disease. *World J. Gastroenterol.*, **23**, 4897–4909.
- Malhotra,A. *et al.* (2014) ADO: A disease ontology representing the domain knowledge specific to Alzheimer’s disease. *Alzheimer’s Dement.*, **10**, 238–246.
- Martin,A.R. *et al.* (2019) PanelApp crowdsources expert knowledge to establish consensus diagnostic gene panels. *Nat. Genet.*, **51**, 1560–1565.
- Mazein,A. *et al.* (2018) Systems medicine disease maps: community-driven comprehensive representation of disease mechanisms. *npj Syst. Biol. Appl.*, **4**.
- Mina,E. *et al.* (2015) Nanopublications for exposing experimental data in the life-sciences: a Huntington’s Disease case study. *J. Biomed. Semantics*, **6**, 5.
- Morosky,S.A. *et al.* (2011) Retinoic acid-induced gene-I (RIG-I) associates with nucleotide-binding oligomerization domain-2 (NOD2) to negatively regulate inflammatory signaling. *J. Biol. Chem.*, **286**, 28574–28583.
- Al Nabhani,Z. *et al.* (2017) Nod2: The intestinal gate keeper. *PLoS Pathog.*, **13**.
- Negrone,A. *et al.* (2018) NOD2 and inflammation: Current insights. *J. Inflamm. Res.*, **11**, 49–60.
- Neurath,M.F. (2014) New targets for mucosal healing and therapy in inflammatory bowel diseases. *Mucosal Immunol.*, **7**, 6–19.
- Novère,N. Le *et al.* (2009) The Systems Biology Graphical Notation. *Nat. Biotechnol.*, **27**, 735–741.

- Ogura, Y. *et al.* (2001) A frameshift mutation in NOD2 associated with susceptibility to Crohn's disease. *Nature*, **411**, 603–606.
- Park, J.-H. *et al.* (2007) RICK/RIP2 Mediates Innate Immune Responses Induced through Nod1 and Nod2 but Not TLRs. *J. Immunol.*, **178**, 2380–2386.
- PDBE-KB consortium (2020) PDBE-KB: a community-driven resource for structural and functional annotations. *Nucleic Acids Res.*, **48**.
- Peter, I. *et al.* (2011) Evaluation of 22 genetic variants with Crohn's Disease risk in the Ashkenazi Jewish population: a case-control study. *BMC Med. Genet.*, **12**.
- Pon, A. *et al.* (2015) Pathways with PathWhiz. *Nucleic Acids Res.*, **43**.
- Saqi, M. *et al.* (2019) Navigating the disease landscape: Knowledge representations for contextualizing molecular signatures. *Brief. Bioinform.*, **20**, 609–623.
- Sarkar, A. *et al.* (2006) ASC Directs NF- κ B Activation by Regulating Receptor Interacting Protein-2 (RIP2) Caspase-1 Interactions. *J. Immunol.*, **176**, 4979–4986.
- Shannon, P. *et al.* (2003) Cytoscape: A software Environment for integrated models of biomolecular interaction networks. *Genome Res.*, **13**, 2498–2504.
- Sidiq, T. *et al.* (2016) Nod2: A critical regulator of ileal microbiota and Crohn's disease. *Front. Immunol.*, **7**.
- Simms, L.A. *et al.* (2008) Reduced α -defensin expression is associated with inflammation and not NOD2 mutation status in ileal Crohn's disease. *Gut*, **57**, 903–910.
- Sioutos, N. *et al.* (2007) NCI Thesaurus: A semantic model integrating cancer-related clinical and molecular information. *J. Biomed. Inform.*, **40**, 30–43.
- Sosa, D.N. *et al.* (2019) A Literature-Based Knowledge Graph Embedding Method for Identifying Drug Repurposing Opportunities in Rare Diseases. In, *Biocomputing 2020*. WORLD SCIENTIFIC, pp.

463–474.

Stenson,P.D. *et al.* (2017) The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. *Hum. Genet.*, **136**, 665–677.

Strober,W. and Watanabe,T. (2011) NOD2, an intracellular innate immune sensor involved in host defense and Crohn’s disease. *Mucosal Immunol.*, **4**, 484–495.

Szklarczyk,D. *et al.* (2018) STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.*, **47**, 607–613.

The Gene Ontology Consortium (2019) The Gene Ontology Resource: 20 years and still GOing strong. *Nucleic Acids Res.*, **47**.

Thomas,P.D. *et al.* (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.

Travassos,L.H. *et al.* (2010) Nod1 and Nod2 direct autophagy by recruiting ATG16L1 to the plasma membrane at the site of bacterial entry. *Nat. Immunol.*, **11**, 55–62.

Underhill,D.M. (2007) Collaboration between the innate immune receptors dectin-1, TLRs, and Nods. *Immunol. Rev.*, **219**, 75–87.

Vihinen,M. (2014) Variation Ontology for annotation of variation effects and mechanisms. *Genome Res.*, **24**, 356–364.

Vilar,E. *et al.* (2014) Role of microsatellite instability-low as a diagnostic biomarker of Lynch syndrome in colorectal cancer. *Cancer Genet.*, **207**, 495–502.

Vilela,E.G. *et al.* (2012) Evaluation of inflammatory activity in Crohn’s disease and ulcerative colitis.

- World J. Gastroenterol.*, **18**, 872–881.
- Visser,U. *et al.* (2011) BioAssay Ontology (BAO): A semantic description of bioassays and high-throughput screening results. *BMC Bioinformatics*, **12**.
- Voss,E. *et al.* (2006) NOD2/CARD15 mediates induction of the antimicrobial peptide human beta-defensin-2. *J. Biol. Chem.*, **281**, 2005–2011.
- Wang,M.H. *et al.* (2002) Macrophage-stimulating protein and RON receptor tyrosine kinase: Potential regulators of macrophage inflammatory activities. *Scand. J. Immunol.*, **56**, 545–553.
- Wehkamp,J. *et al.* (2005) Reduced Paneth cell α -defensins in ileal Crohn’s disease. *Proc. Natl. Acad. Sci. U. S. A.*, **102**, 18129–18134.
- Wilson,S.S. *et al.* (2015) A small intestinal organoid model of non-invasive enteric pathogen-epithelial cell interactions. *Mucosal Immunol.*, **8**, 352–361.
- Wolfert,M.A. *et al.* (2002) The origin of the synergistic effect of muramyl dipeptide with endotoxin and peptidoglycan. *J. Biol. Chem.*, **277**, 39179–39186.
- Yamamoto-Furusho,J.K. *et al.* (2010) MDP-NOD2 stimulation induces HNP-1 secretion, which contributes to NOD2 antibacterial function. *Inflamm. Bowel Dis.*, **16**, 736–742.
- Yamamoto,S. and Ma,X. (2009) Role of Nod2 in the development of Crohn’s disease. *Microbes Infect.*, **11**, 912–918.
- Younesi,E. *et al.* (2015) PDON: Parkinson’s disease ontology for representation and modeling of the Parkinson’s disease knowledge domain. *Theor. Biol. Med. Model.*, **12**, 20.