UNIVERSITY OF MARYLAND | NIST
INSTITUTE FOR BIOSCIENCE
& BIOTECHNOLOGY RESEARCH

# Knowledge Representation and Integration for Complex Trait Diseases

Lipika Ray Pal[1], Kunal Kundu[1,2], Lindley Darden[3], and John Moult[1,4*]

[1]Institute for Bioscience and Biotechnology Research, University of Maryland at College Park, Rockville, MD, USA
[2]Computational Biology, Bioinformatics and Genomics, Biological Sciences Graduate Program, University of Maryland at College Park, MD, USA
[3]Department of Philosophy, University of Maryland at College Park, College Park, MD, USA
[4]Department of Cell Biology and Molecular Genetics, University of Maryland at College Park, College Park, MD, USA
*(jmoult@umd.edu)

## ABSTRACT

Genome wide association studies (GWAS) have provided a wealth of new insight into which genetic loci are associated with disease phenotypes for many complex trait diseases. However, identification of these loci has led to limited advances in understanding the underlying disease mechanisms, and many questions remain unanswered, including which subsystems are involved, the extent and nature of epistatic effects, and possible new therapeutic strategies. Addressing these issues requires an appropriate representation and a means of integrating knowledge from many sources. We have utilized formal concepts of biological mechanism to develop a graphical framework to represent disease mechanisms. Each disease associated locus and its corresponding disease phenotype is linked through a series of substate perturbations at the DNA, RNA, protein, organelle, cellular, and other stages. Each pair of consecutive perturbed substates is connected by a mechanism module representing the activity that transforms one substate to the next. The set of intersecting mechanism chains form a mechanism graph, whose features may be used for subsystem and epistatic analysis. A simple graphical language is used to represent the chain. User-friendly tools, including pull down menus for ontology terms, allow approved contributors to build and edit chains.

## OBJECTIVE

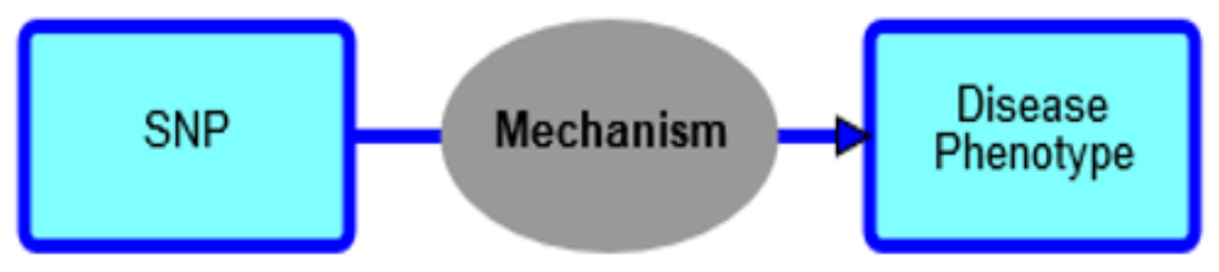

*Figure 1: Every statistical association between the presence of a genetic perturbation, a SNP, and perturbation of a disease phenotype, such as altered disease risk, implies the presence of an underlying mechanism. Without further information, that mechanism is unknown – a 'black box'. A goal of this project is to replace these black boxes with detailed mechanism descriptions.*

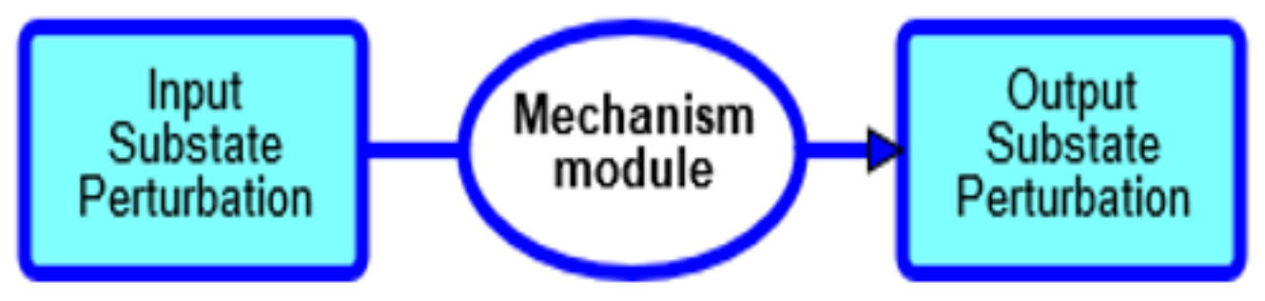

*Figure 2: Each step in a mechanism chain consists of an input substate perturbation, a mechanism module, and an output substate perturbation.*

An optimum framework for describing and analyzing complex disease mechanisms must fulfill a number of requirements:

(1) The representation of mechanisms must be clear, intuitive, and uncluttered. We use a graphical representation of disease mechanisms, at the same time providing extensive links to pathway, GWAS, expression, proteomics, literature, and other sources.

(2) The framework must incorporate a suitable ontology. The ontology we use integrates existing definitions with formal concepts of mechanism developed by one of us (Darden [1]). The system must be described not only graphically but also in computable and searchable form. For this purpose, we are adapting the Biological Expression Language (BEL, http://www.openbel.org/).

(3) Because of the very incomplete state of knowledge, provision must be made for directly and clearly representing ambiguities, uncertainties and knowledge gaps, and these facilities are an integral part of the mechanism language.

(4) Finally, it is not possible for any one group of investigators to construct a sufficiently complete description of a typical complex trait disease. To this end, the system is designed to manage and facilitate contributions from many participants.

## LEGEND

Symbols:

▭ : Substate perturbation

○ : Mechanism module

● : Unknown mechanism module

Information : Pop-up box for source of supporting information or any related text for any symbol

STAGE : Substate perturbations at different stages, such as DNA, RNA, Protein, Complex, Cell

## MECHANISM CHAIN FOR THE MST1 CROHN'S DISEASE LOCUS

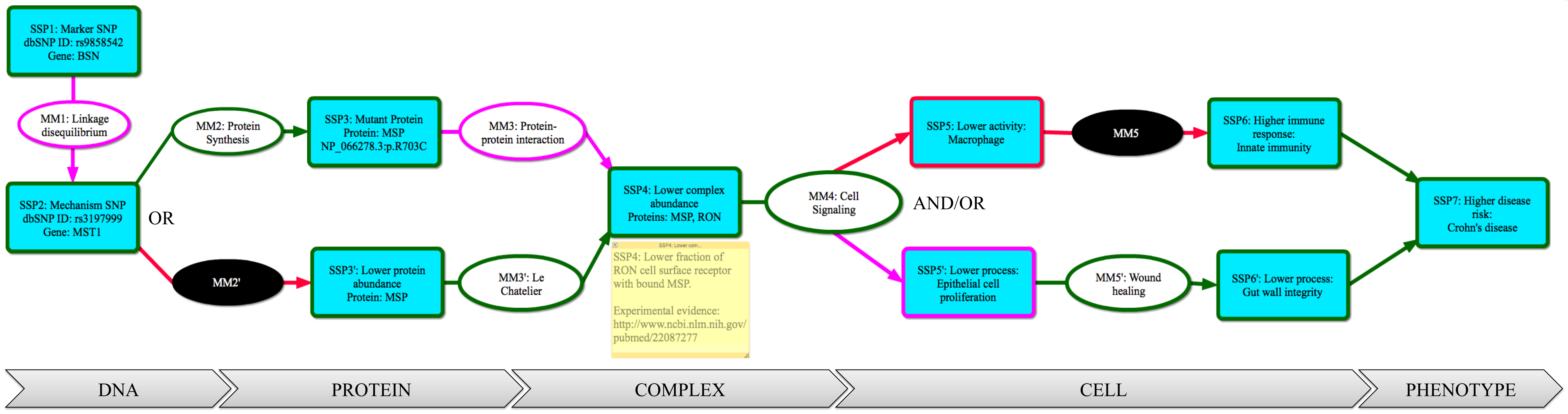
SSP1: Marker SNP
dbSNP ID: rs9858542
Gene: BSN
MM1: Linkage disequilibrium
SSP2: Mechanism SNP
dbSNP ID: rs3197999
Gene: MST1
OR
MM2: Protein Synthesis
SSP3: Mutant Protein
Protein: MSP
NP_066278.3:p.R703C
MM3: Protein-protein interaction
MM2'
SSP3': Lower protein abundance
Protein: MSP
MM3': Le Chatelier
SSP4: Lower complex abundance
Proteins: MSP, RON
SSP4: Lower fraction of RON cell surface receptor with bound MSP.
Experimental evidence: http://www.ncbi.nlm.nih.gov/pubmed/22087277
MM4: Cell Signaling
AND/OR
SSP5: Lower activity: Macrophage
MM5
SSP6: Higher immune response: Innate immunity
SSP5': Lower process: Epithelial cell proliferation
MM5': Wound healing
SSP6': Lower process: Gut wall integrity
SSP7: Higher disease risk: Crohn's disease
DNA
PROTEIN
COMPLEX
CELL
PHENOTYPE

*Figure 3: Prototype diagram of the mechanism schema for the Crohn's disease MSP (Macrophage Stimulating Protein) locus.*

## LEGEND

→ : Arrows connect one perturbed substate to the next via a mechanism module

**Confidence code:**

Arrow and symbol colors reflect confidence that a module produces the next substate

🟩 : Green – High confidence

🟪 : Pink – Medium confidence

🟥 : Red – Low confidence

## THE MST1 MECHANISM CHAIN

Figure 3 shows an example of one of the mechanism chains for Crohn's disease, originating in a locus containing the MST1 gene, which codes for MSP (Macrophage stimulating Protein). The mechanism begins at the perturbed DNA substate on the left, and progresses through protein, protein-protein complex, cell signalling, innate immune response and gut barrier layer stages, to disease risk. In this view, some parts of the chain, at the DNA, protein, and protein complex levels, are fully expanded, while others are partly telescoped (for example 'cell signalling', 'innate immunity') and have multiple substate and mechanism modules within them. Clicking will expand or compress chain sections, so that viewing of large and complex chains can be easily managed.

The mechanism chain starts with a marker SNP (SSP1) in gene BSN (Bassoon presynaptic cytomatrix protein), which lies in the intronic region. This gene is involved in neural development, so unlikely to be relevant for Crohn's. This SNP is in high linkage disequilibrium (MM1) with a mechanism SNP (SSP2) in gene MST1, 20Kb away from the marker SNP (Pal and Moult, [2]). Branches in the chain show possible alternative sub-mechanisms. For example, the branch beginning at the mechanism SNP reflects the facts that one study (Gorlatova et al, [3]) has reported that the SNP results in a weaker protein-protein interaction between MSP and a cell surface receptor, RON (MM2, SSP3, MM3), while another (Kauder et al, [4]) reports that the SNP affects the level of MSP in plasma (MM2', SSP3', MM3'). The two mechanism branches lead to the same outcome – a lower concentration of the MSP-RON protein complex (SSP4), so the mechanism chains converge there. A second uncertainty in this chain (the parallel paths in the right hand segment) is whether lower intra-cellular signaling (MM4) most affects macrophage activity, and consequently innate immunity (SSP5, MM5, SSP6); or whether it affects wound healing activities of epithelial cells, so primarily altering barrier integrity (SSP5', MM5', SSP6'). Black boxes (such as MM2' and MM5) represent unknown steps in the chain. Pop-up boxes, such as the one shown open for the RON-MSP complex (SSP4), provide brief text explanations, ontology terms, and GO evidence codes (Ashburner et al, [5]) indicating the source of supporting information, and links to the supporting evidence, data, and relevant literature for that chain feature.

## QUESTIONS TO COMMUNITY

To incorporate all available knowledge in just the MSP example it is evident that a wide range of expertise is needed, including:

- ◆ How a SNP affects expression of a protein?
- ◆ How such changes affect that protein's structure and function?
- ◆ How those protein changes affect interaction with a partner protein?
- ◆ How that interaction affects intracellular signalling, and in what cell types?
- ◆ If macrophage response is indeed involved, how those changes may affect other aspects of the immune response?
- ◆ If wound healing is affected, how is that altered?

Each of the 140 chains of Crohn's disease (Jostins et al, [6]) requires a similar breadth of knowledge.

!! COMMUNITY EXPERTISE REQUIRED !!

## CONCLUSION

The mechanism chain formalism advances understanding of complex trait disease in five primary ways:

(1) It clearly delineates what is known and not known about the mechanism underlying each disease association.

(2) It provides a more rigorous and objective means of defining subsystems involved in the disease.

(3) It provides a framework for identifying possible epistatic interactions, within and between subsystems.

(4) It facilitates identification of sites of potential therapeutic intervention.

(5) It provides a general framework for integrating and expanding knowledge about disease mechanisms.

Our goal is to create an **EXPERT-SOURCING INFRASTRUCTURE**, which will allow appropriate experts to construct and edit chains and to fill black boxes. For this aspect of the work we are drawing on our experience with community-wide experiments, such as CASP (predictioncenter.org), Proteopedia (proteopedia.org) and CAGI (genomeinterpretation.org).

## REFERENCES

1. Craver, CF and Darden, L, *In Search of Mechanisms: Discoveries across the Life Sciences.* Chicago, IL: University of Chicago Press (2013).
2. Pal LR & Moult J, (2015), J. Mol. Biol
3. Gorlatova N, Chao K, Pal LR, Araj RH, Galkin A, Turko I, Moult J, Herzberg O, PLoS One, 6, e27269, (2011).
4. Kauder SE, Santell L, Mai E, *et al*. PLoS One 8(12):e83958 2013.
5. Ashburner M, Ball CA *et al,* Nat Genet, 25, 25-29, (2000).
6. Jostins L, Ripke S, CA *et al*, Nature, 491, 119-124, (2012).

**Acknowledgement:**

This work was supported in part by NIH R01GM102810 and R01GM104436.