Check for updates

RESEARCH ARTICLE

Enhancement of COPD biological networks using a web-based collaboration interface [version 1; peer review: 3 approved]

The sbv IMPROVER project team (in alphabetical order), Stéphanie Boué[1], Brett Fields[2], Julia Hoeng[1], Jennifer Park[2], Manuel C. Peitsch[1], Walter K. Schlage[1], Marja Talikka[1], The Challenge Best Performers (in alphabetical order), Ilona Binenbaum[3], Vladimir Bondarenko[4], Oleg V. Bulgakov[5], Vera Cherkasova[9], Norberto Diaz-Diaz[6], Larisa Fedorova[7], Svetlana Guryanova[8], Julia Guzova[11], Galina Igorevna Koroleva[10], Elena Kozhemyakina[11], Rahul Kumar[12], Noa Lavid[13], Qingxian Lu[14], Swapna Menon[15], Yael Ouliel[13], Samantha C. Peterson[16], Alexander Prokhorov[8], Edward Sanders[17], Sarah Schrier[18], Golan Schwaitzer Neta[13], Irina Shvydchenko[19], Aravind Tallam[20], Gema Villa-Fombuena[21], John Wu[22], Ilya Yudkevich[23], Mariya Zelikman[24]

[1]Philip Morris International R&D, Philip Morris Products S.A., Quai Jeanrenaud 5, 2000 Neuchâtel, Switzerland
[2]Selventa, One Alewife Center, Cambridge, MA, 02140, USA
[3]Systems Bioengineering Group - National Technical University of Athens, Ethniko Metsovio Politechnio, , 28is Oktovriou 42, Athina, 106 82, Greece
[4]Touro University Nevada, 874 American Pacific Drive, Henderson, NV, 89052, USA
[5]University of Pittsburgh, 4200 Fifth Ave, Pittsburgh, PA, 15260, USA
[6]Intelligent Data Analysis Group (DATAi), School of Engineering, Pablo de Olavide University, Ctra. de Utrera, km. 1 41013, Sevilla, Spain
[7]University of Toledo, 2801 W Bancroft St, Toledo, OH, 43606, USA
[8]Shemyakin & Ovchinnikov Institute of Bioorganic Chemistry, 16/10, Miklukho-Maklay str., Moscow, 117997, Russian Federation
[9]Private, Washington DC, USA
[10]USAMRIID, Attn: MCMR-UIZ-R, 1425 Porter Street, Frederick, MD, 21702-5011, USA
[11]Private, Boston, MA, USA
[12]Institute of Microbial Technology, Chandigarh, 160036, India
[13]Technion - Israel Institute of Technology, Technion City, Haifa, 3200003, Israel
[14]Louisville University, 301 E. Muhammad Ali Blvd, Louisville, KY, 40202, USA
[15]AnalyzeDat Consulting Services, Ernakulam, India
[16]Northeastern University, 360 Huntington Ave, Boston, MA, 02115, USA
[17]Edward Sanders Scientific Consulting, Rue du Clos 33, 2034 Peseux, Switzerland
[18]Massachusetts Institute of Technology, 77 Massachusetts Ave, Cambridge, MA, 02139, USA
[19]Kuban State University of Physical Education, Sport and Tourism, 161, Budennogo Str., Krasnodar City, 350015, Russian Federation
[20]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, 7, avenue des Hauts-Fourneaux, 4362 Esch sur Alzette, Luxembourg
[21]Pablo de Olavide University, Ctra. de Utrera, km. 1 41013, Sevilla, Spain
[22]Cal Biopharma, 710 Somerset Ln, Foster Cit, CA, 94404-3728, USA
[23]University of Manchester, Oxford Rd, Manchester, M13 9PL, UK
[24]University of Washington, 1959 NE Pacific Street, HSB T-466, Seattle, WA, USA

First published: 29 Jan 2015, **4**:32
https://doi.org/10.12688/f1000research.5984.1
Latest published: 20 May 2015, **4**:32

Open Peer Review

https://doi.org/10.12688/f1000research.5984.2

Abstract

The construction and application of biological network models is an approach that offers a holistic way to understand biological processes involved in disease. Chronic obstructive pulmonary disease (COPD) is a progressive inflammatory disease of the airways for which therapeutic options currently are limited after diagnosis, even in its earliest stage. COPD network models are important tools to better understand the biological components and processes underlying initial disease development. With the increasing amounts of literature that are now available, crowdsourcing approaches offer new forms of collaboration for researchers to review biological findings, which can be applied to the construction and verification of complex biological networks. We report the construction of 50 biological network models relevant to lung biology and early COPD using an integrative systems biology and collaborative crowd-verification approach. By combining traditional literature curation with a data-driven approach that predicts molecular activities from transcriptomics data, we constructed an initial COPD network model set based on a previously published non-diseased lung-relevant model set. The crowd was given the opportunity to enhance and refine the networks on a website (https://bionet.sbvimprover.com/) and to add mechanistic detail, as well as critically review existing evidence and evidence added by other users, so as to enhance the accuracy of the biological representation of the processes captured in the networks. Finally, scientists and experts in the field discussed and refined the networks during an in-person jamboree meeting. Here, we describe examples of the changes made to three of these networks: *Neutrophil Signaling*, *Macrophage Signaling*, and *Th1-Th2 Signaling*. We describe an innovative approach to biological network construction that combines literature and data mining and a crowdsourcing approach to generate a comprehensive set of COPD-relevant models that can be used to help understand the mechanisms related to lung pathobiology. Registered users of the website can freely browse and download the networks.

Keywords

COPD , Chronic Obstructive Pulmonary Disease , network model , signaling pathway , crowdsourcing , crowd verification , jamboree , online collaboration

Approval Status ✓ ✓ ✓

	1	2	3
version 2 (revision) 20 May 2015			
version 1 29 Jan 2015	✓ view	✓ view	✓ view

1. **Winston A. Hide**, Harvard School of Public Health, Boston, USA
2. **Gary D. Bader**, University of Toronto, Toronto, Canada
3. **Patrick Ruch**, University of Applied Sciences—HEG, Geneva, Switzerland

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Julia Hoeng (Julia.Hoeng@pmi.com)

Competing interests: Selventa and PMI authors performed this work under a joint research collaboration funded by PMI. The Network Verification Challenge was funded by PMI.

Grant information: The research described in this article was funded by Philip Morris International in a collaborative project with Selventa.

Copyright: © 2015 The sbv IMPROVER project team (in alphabetical order) *et al*. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. Data associated with the article are available under the terms of the Creative Commons Zero "No rights reserved" data waiver (CC0 1.0 Public domain dedication).

How to cite this article: The sbv IMPROVER project team (in alphabetical order), Boué S, Fields B *et al.* **Enhancement of COPD biological networks using a web-based collaboration interface [version 1; peer review: 3 approved]** F1000Research 2015, **4**:32 https://doi.org/10.12688/f1000research.5984.1

First published: 29 Jan 2015, **4**:32 https://doi.org/10.12688/f1000research.5984.1

Introduction

Molecular networks, such as the KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways[1,2], aid in understanding the complex interplay of signaling pathways in disease. Biological network models (hereafter referred to as networks) depict the inter-relationships between multiple signaling pathways and how their perturbations may dysregulate biological processes, eventually leading to the disease.

In previously published reports, we described the construction of a set of 90 networks that captured a large range of biological processes relevant to non-diseased lung tissue[3–7]. The generation of this set of networks relied on both manual curation of published literature and a data-driven reverse causal reasoning (RCR) methodology[8] to augment the causal biological framework underlying the network architecture (Figure 1). We used the Biological Expression Language (BEL) to represent precise biological relationships in a computable and standardized format[8]. We have built upon this approach and describe here a unique, three-phase systems biology and crowdsourcing approach to construct a comprehensive set of 50 molecular networks that describe the biological processes relevant to chronic obstructive pulmonary disease (COPD) and lung biology (Figure 2). COPD is the fourth leading cause of death worldwide and its incidence is increasing among chronic diseases in the USA[9,10]. COPD is a chronic, progressive inflammatory disease induced by cigarette smoking, inhalation of pollutants, dust, chemicals, or other foreign matter, which ultimately manifests as tissue destruction in the alveolar compartments and airflow limitation, leading to reduced oxygen exchange[11–15]. COPD affects a wide spectrum of biological processes in lung tissue, such as oxidative stress, inflammation, apoptosis, proliferation, and senescence[16,17]. Understanding the mechanisms involved in these processes is important in understanding the onset of the disease and in identifying drug targets to develop effective COPD treatments[18,19]. As recently reported by the Global Initiative for Chronic Obstructive Lung Disease (GOLD), current pharmacologic therapies cannot cure the disease but only reduce the symptoms, and the frequency and severity of exacerbations, i.e., slow down the rate of disease progression[11]; thus it appears most efficient to target the COPD-specific pathomechanisms at the earliest distinguishable state, when the extent of irreversible damage is still small, and their molecular processes are not yet convoluted with secondary processes and comorbidities, e.g., bacterial and viral

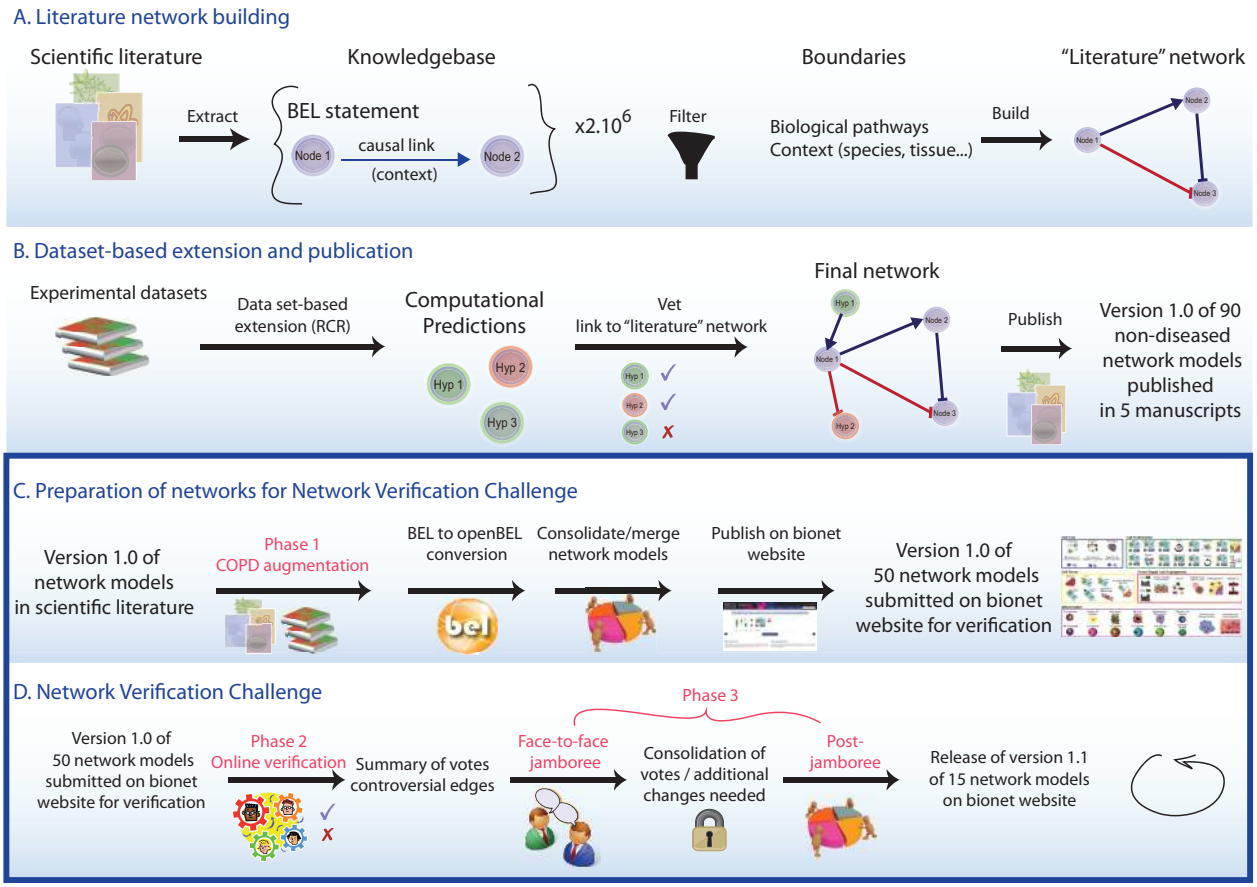

Figure 1. Network construction using a systems biology and crowdsourcing approach. Networks were constructed using published literature and data sets, and opened to the public for comment and editing in the Network Verification Challenge. The three phases of COPD network construction are shown. (**A** and **B**) Phase 1: COPD augmentation using literature and data. (**C** and **D**) Phase 2: Online verification by the public during an "open phase", and Phase 3: Face-to-face jamboree meeting where scientists and subject matter experts gathered to discuss the networks and make final decisions for the next versions.

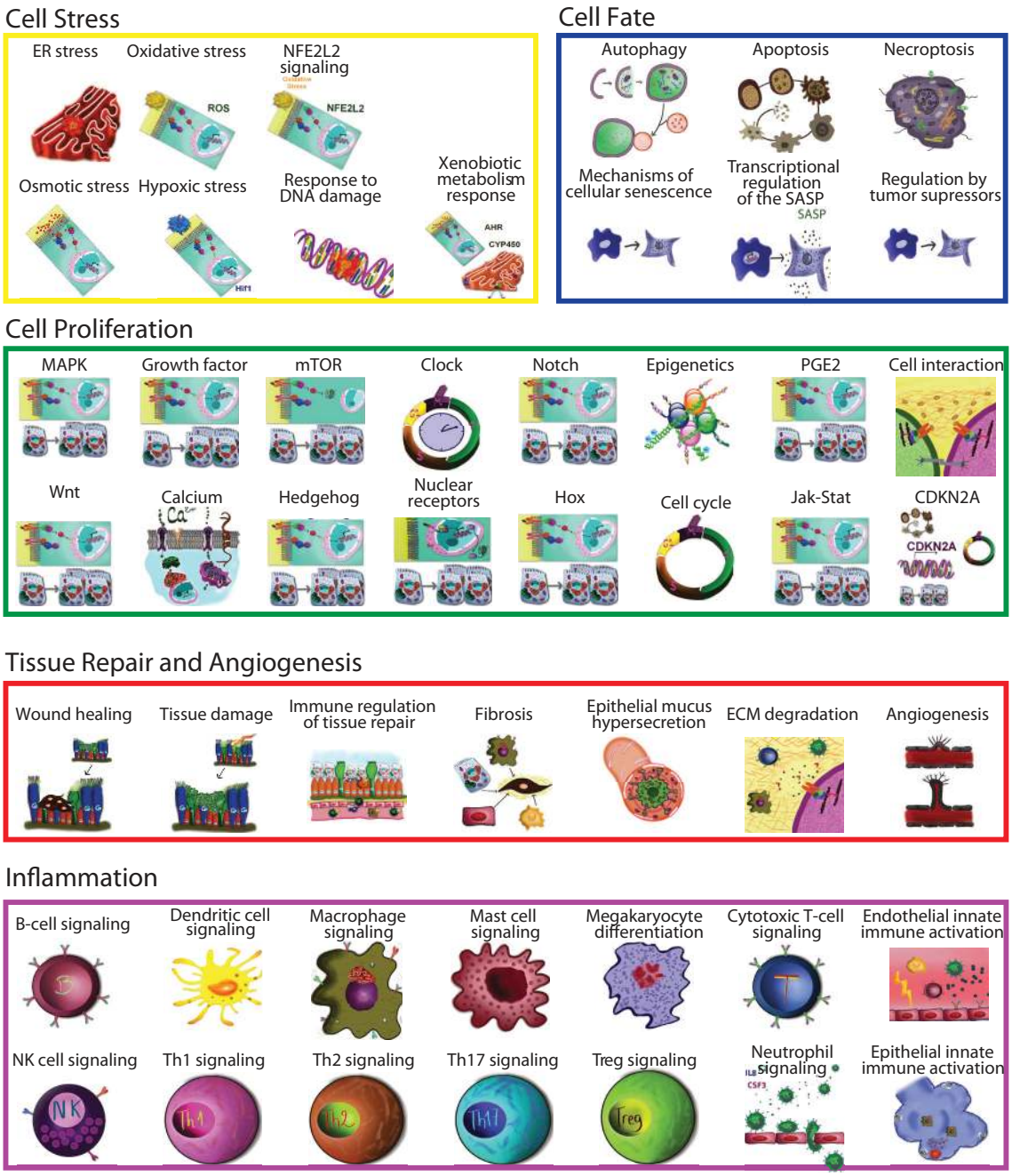

Figure 2. Fifty networks available during the network verification challenge and their associated biological processes.

infections, as they occur during the exacerbations typical for later stages of COPD. Since smoking cessation/replacement appears to be the most efficient therapy in smoking-related COPD[11], the models of early onset COPD can also be expected to be valuable tools for the development and testing of reduced risk products that may prevent COPD progression in a comparable manner as cessation does.

The networks reported here were created first from a literature scaffold and expanded via data enhancement using RCR (Phase 1), then they were made available online to the entire scientific community for critical review during the Network Verification Challenge (NVC) "Open Phase" (Phase 2) under the umbrella of the systems biology verification (sbv) IMPROVER project[20] (Figure 1). Finally, a prioritized subset of 15 of these networks was discussed during an in-person jamboree meeting where the crowd-submitted revisions were reviewed and decisions to improve the networks were finalized (Phase 3). The final versions of the networks are available at https://bionet.sbvimprover.com for the public to view, and for registered users in the NVC to continue to discuss.

A variety of COPD networks have been created by various research groups, including networks focused on muscle to study skeletal

muscle abnormalities[21], networks to compare COPD and asthma[22], and a knowledge management framework to integrate COPD clinical and experimental data[23]. To our knowledge, this is the first set of crowd-verified networks available to the broader scientific community as a unified collection on a freely accessible web-based platform. Ultimately, this interface will allow for continuous input and improvement in the networks, leading to better understanding, diagnosis, and treatment of COPD.

Methods

Results

Data File

2 Datasets

http://dx.doi.org/10.6084/m9.figshare.1284583

Phase 1: COPD enhancements using data and literature

Ninety non-diseased lung networks published previously in the areas of cell proliferation, cell stress, inflammation, DNA damage, cell death, tissue repair, and angiogenesis were used as the initial scaffolds for COPD enhancement during Phase 1[3–7]. Biological pathways implicated in COPD disease pathophysiology, including B-cell and T-cell activation, airway remodeling, extracellular matrix (ECM) degradation, efferocytosis, mucus hypersecretion, and emphysema were all captured within the modified network models. In total, 200 new nodes and 487 new edges were added: 415 of the edges were added to incorporate COPD mechanisms implicated in the literature, and 72 edges were added to incorporate 100 mechanisms predicted from COPD data by RCR to be relevant to COPD (Figure 3). Because the models were built to represent COPD in humans, human evidence was preferred and made up the majority of the networks (74%).

During Phase 1, the networks with the most significant number of COPD enhancements in terms of percentage of the network with new nodes were the *Mucus Hypersecretion* (44%), *Th2 Signaling* (37%), *Macrophage Activation* (28%), *Fibrosis* (25%), *Autophagy* (11%), and *Apoptosis* (5%) networks. Networks that were not enhanced with COPD-specific mechanisms from the literature or RCR included the *DNA Damage* and *Notch Signaling* networks. Although both these networks relevant to the development of COPD, they were not augmented beyond the original, non-diseased network scaffolds, because no studies on the differences in signaling between non-diseased and diseased states were available.

Phase 2: Networks enhanced with lung- and COPD-relevant mechanisms by the crowd during the open phase

Prior to deploying the COPD-enhanced biological networks on the NVC website for verification by the scientific community, the set of 90 networks was agglomerated by the model-building expert team to yield a more concise set of 50 networks that combined and standardized related/complementary cellular pathways (See Methods for details). For example, a new "Th1 Signaling" network model was created by merging three of the original networks that were relevant to the functional biology present in T-helper 1 cell populations: *Th1 Differentiation*, *Th1 Response*, and *T-cell Recruitment and Activation*. For a list of the original models that correspond to the agglomerated models and a description of the new models, see Dataset.

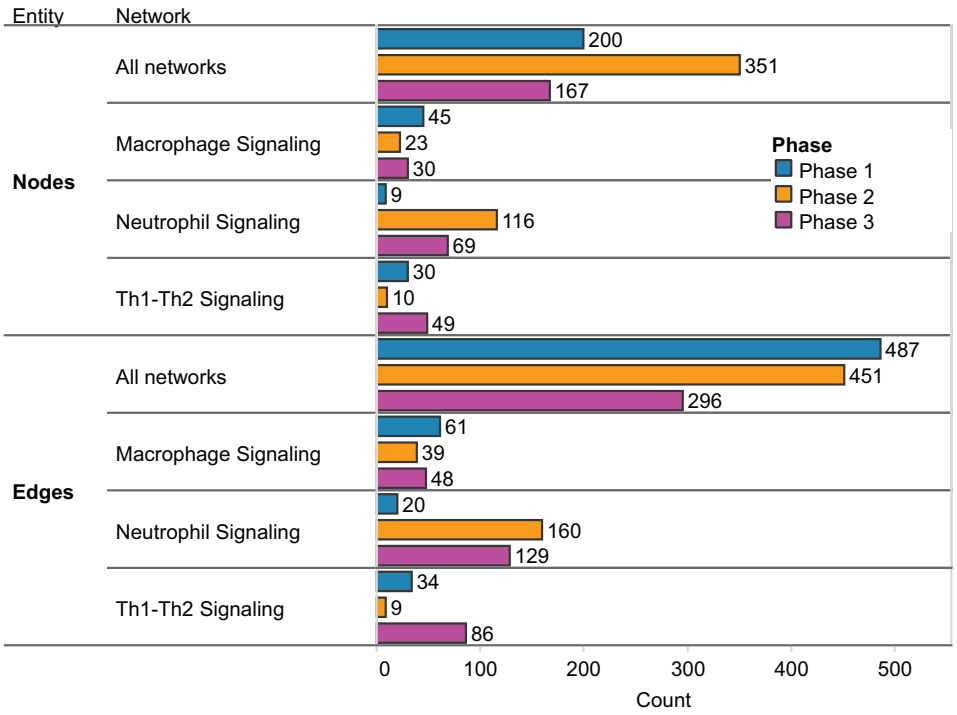

Figure 3. Nodes and edges added in each phase of COPD network construction. Summary of nodes and edges added to all networks and to three example networks in each phase. **A**) Nodes added in each phase. **B**) Edges added in each phase.

During Phase 2, a global community of scientists participated in the NVC by contributing their expertise to one or several of the network models. Scientists could contribute by verifying existing evidence for network edges using a system that allowed users to vote on evidence to indicate agreement or disagreement with its appropriateness within the network structure and boundary conditions. Participants were also encouraged to add new mechanistic biology in the form of network edges. In total, the 50 network models received 2456 evidence votes, 1795 of which supported the confirmation of evidence and 661 that favored the rejection of evidence (see Dataset). The *Neutrophil Signaling* network model received the largest share of voting activity, with 241 total votes or approximately 10% of all votes cast. Other network models that received large shares of the votes included the *Macrophage Signaling* (180 votes) and *Th1 and Th2 Signaling* network models (105 votes) (see Dataset). In addition to verifying existing literature evidence supporting edges in the network models, NVC participants could add novel biological information in the form of new literature evidence (for an existing edge) or contribute new network edges to incorporate new biological components into the network structure. In this way, the community of participants collectively contributed a significant amount of new information into the networks; among the 50 network models, a total of 885 new pieces of evidence, 351 new nodes, and 451 new edges were added (Figure 3).

Phase 3: Jamboree discussion and final decisions for next version networks

Following Phase 2, a jamboree (Phase 3) was organized for a group of invited participants to discuss the network enhancements submitted by the crowd. To represent the crowd community, the top 20 active performers who created the most pieces of evidence and submitted at least 20 votes during the NVC were invited to an in-person jamboree to discuss network refinements as a group. Additional subject matter experts in the network biology, COPD, lung biology, and biological processes represented by the networks were invited to participate in the discussions and contribute their expert feedback independent from the network-building experts. Among the 50 network models evaluated during the online NVC, 15 were prioritized and selected for discussion during Phase 3 based on the level of crowd-sourced activity and their importance in COPD onset as considered by the network-building experts (see Dataset). The goal of Phase 3 was to provide an additional layer of "verification" for the online enhancements and to provide holistic comments on the network models at the molecular/biological entity level. In doing so, the three network models that had received the largest amounts of crowd activity (*Neutrophil Signaling*, *Macrophage Signaling*, and *Th1 Signaling*) also underwent significant additional enhancements to improve granularity with respect to COPD onset and pathogenesis. In total, 167 nodes and 296 edges were added among all the network models reviewed during the jamboree sessions, and the three inflammatory networks received 89% of the nodes and 89% of the edges (148 nodes and 263 edges) (Figure 3). Many of these changes came from the identification of missing mechanistic details of processes that occur in COPD (e.g. chemotaxis mechanisms in the *Macrophage Signaling* network model described in the examples in the "Macrophage signaling" section below).

In addition to adding mechanistic details of processes that occur in COPD, enhancements were incorporated to improve the granularity and connectivity within the network structures. In several instances, the improvements involved the creation of more detailed linear pathways connecting biological components. In one example, in the *Apoptosis* network model, the original network pathway indicated that the X-ray repair complementing defective repair in Chinese hamster cells 6 (XRCC6) protein decreased the process of apoptosis[24]. During the Phase 3 discussions, additional literature evidence provided a more detailed mechanistic understanding of this phenomenon: XRCC6 was reported to decrease the activity of the BCL2-associated X protein (BAX) protein, which is known to increase mitochondrial permeability and therefore promote apoptosis (Figure 4A). The overall effect of the negative regulation of BAX by XRCC6 was therefore a decrease in apoptotic cell death[25]. By improving the granularity of this pathway in the *Apoptosis* network, a more comprehensive representation was achieved for components that are related to critical cellular processes mediating disease onset.

A similar improvement was incorporated into the *Mechanisms of Cellular Senescence* network model: the original network pathway indicated that the chemical acrolein (a common component of cigarette smoke) increased cell senescence[26]. During Phase 3 discussions, the pathway connecting these two components was expanded using additional literature evidence. In several studies, acrolein was found to decrease the activity of sirtuin 1 (SIRT1), which is a known negative regulator of the forkhead box O3 (FOXO3) transcription factor, and FOXO3 activity is known to promote cellular senescence (Figure 4B)[26–28]. Therefore, the overall observed effect was acrolein acting to potentiate cellular senescence in exposed cells, which is a well-characterized mechanism of action for this toxic chemical. Again, the generation of more comprehensive network models of biological processes in close proximity to disease onset allowed for a greater mechanistic understanding of how environmental factors can contribute to COPD development.

Exemplary outcomes of the three-phase COPD network building process

Th1 and Th2 signaling

As part of the pulmonary inflammatory process network building[6], five networks (*T-cell activation and recruitment, Th1 differentiation, Th2 differentiation, Th1 Response, Th2 response*) were built to describe Th1 and Th2 signaling in the non-disease lung context. As described previously, during the preparation phase to NVC, two networks were built around the Th1 and Th2 cells.

Phase 1: COPD augmentation of T-helper cell networks

Mechanisms that describe T-cell activation and recruitment induced by neutrophils, macrophages, and dendritic cells were added to the T-cell networks during Phase 1. These immune cells secrete various chemokines that were reported to recruit T-cell populations (i.e. CD8+ cytotoxic T-cells) to injured tissue in an acute inflammatory state[29]. Alveolar macrophages secrete interleukin 15 (IL15), which is capable of activating both the interleukin 2 (IL2) and IL15 receptors on T-cells and acts as a potent inducer of cell migration to the lung. Dendritic cells within the lung play an important role in this process by secreting chemokine (C-C motif) ligand 3 (CCL3) in response to cigarette smoke, which helps recruit CD8+ T-cells to the lung[29]. Chemokine (C-C motif) receptor 5 (CCR5) is the receptor for CCL3 and its presence in the lung has been shown to correlate

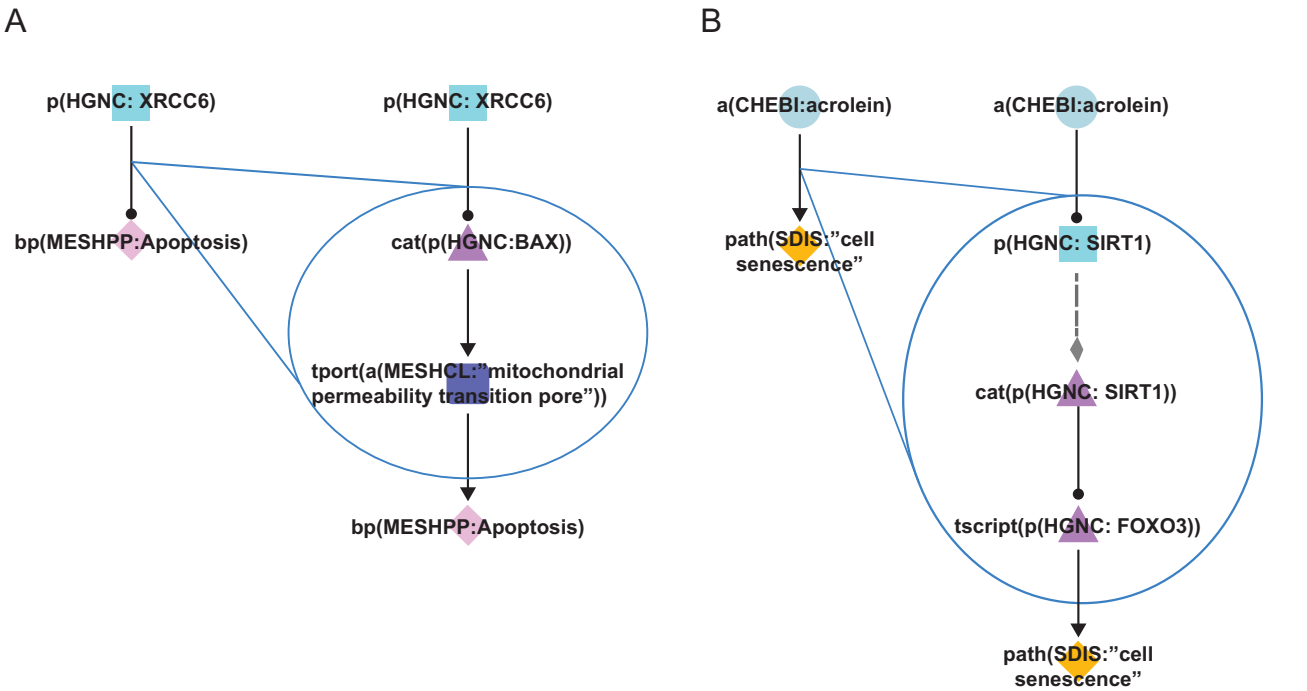

Figure 4. Improvements in the granularity of two representative network pathways. During Phase 3 of COPD network construction, improvements were made by adding mechanistic details to over-simplistic edges**. A**) In the *Apoptosis* network model, the original connection (left) simply indicated that XRCC6 decreased the process of apoptosis. The improved pathway connection (right) indicates that XRCC6 decreases the activity of BAX, which normally functions to facilitate the transport of calcium ions through the mitochondrial pores and thereby increases apoptosis. **B**) In the *Mechanisms of Cellular Senescence* network model, the original connection (left) simply indicated that acrolein increased the process of cellular senescence. The improved pathway connection (right) indicates acrolein mediates its effects on senescence via the activity of SIRT1 and the FOXO3 transcription factor. Triangle denotes activity, diamond denotes biological process or pathology, circle denotes abundance, rounded square represents transport, and square denotes protein abundance nodes. Solid edges denote causal relationships, dotted edges denote non-causal relationships such as a protein connected to its own activity.

with the severity of COPD[30]. CCL3 is one example of a node that was added during the literature-based COPD enhancement process in Phase 1 (Figure 5A). Many of the disease-relevant mechanisms identified in the literature curation phase were corroborated by mechanisms predicted from COPD-relevant data sets using RCR (see Methods), including T-cell activation mechanisms (CD28 molecule (CD28) and T cell receptor beta locus (T\RB), and chemokines and cytokines that activate and are secreted by T-cells (chemokine (C-C motif) receptor 3 (CCR3), CCR5, IL2, interleukin 4 (IL4), interleukin 6 (IL6), interleukin 10 (IL10) and interleukin 13 (IL13)). The prediction of these mechanisms in COPD data sets showed that T-cell activation and migration in response to smoke-exposed lung represents an important process in the innate immune response. In total, 30 nodes and 34 edges were added to the Th1 and Th2 networks during the internal COPD enhancement process.

Phases 2 and 3: T-cell network crowd improvements

During the open phase (Phase 2), the Th1 and Th2 networks received 105 votes from the scientific community, as well as 10 new nodes, 9 new edges, and 13 new pieces of evidence. One such addition to the *Th1 Signaling* network was the regulatory influence of early growth response 2 (EGR2) on T-cell activation; the submitted evidence demonstrated that overexpression of EGR2 promoted increased activity of the E3 ubiquitin ligase CBL-B and subsequent inhibition of T-cell activation[31] (Figure 5B).

During the Phase 3 jamboree sessions, the group decided to combine the individual *Th1 Signaling* and *Th2 Signaling* networks into a single, unified network model titled *Th1-Th2 Signaling* to better represent the interplay between the T-helper cell populations *in vivo*. It was also decided to add granularity to transcriptional pathways mediating Th1 versus Th2 cellular activation and differentiation; one example was the addition of two transcription factors, interferon regulatory factors 1 and 2 (IRF1 and IRF2), that are known to act downstream of interferon-gamma (IFNG) to suppress IL4 expression in Th2 cell populations[32]. IFNG is secreted by Th1 cells and this pathway potentiates Th1 responses while suppressing Th2 responses in the tissue. The addition of this feedback mechanism during Phase 3 contributed to a more comprehensive network describing the interactions between Th1 and Th2 cells (Figure 5C). Further network enhancements discussed in the jamboree largely emphasized the downstream effects of T-helper cells in potentiating inflammatory signaling by activating additional immune cells in a disease context. For example, secretion of IL5 activates eosinophils, whereas secretion of IL10 and IFNG activates macrophages in the diseased tissue[33–35]. This interplay between immune cell populations was incorporated into the new *Th1-Th2 Signaling* network model and better captures the signaling interconnectivity present during disease development (Figure 5D). In total, 12 new nodes and 28 new edges were added to the *Th1-Th2 Signaling* network model during the jamboree discussions, thereby creating a comprehensive

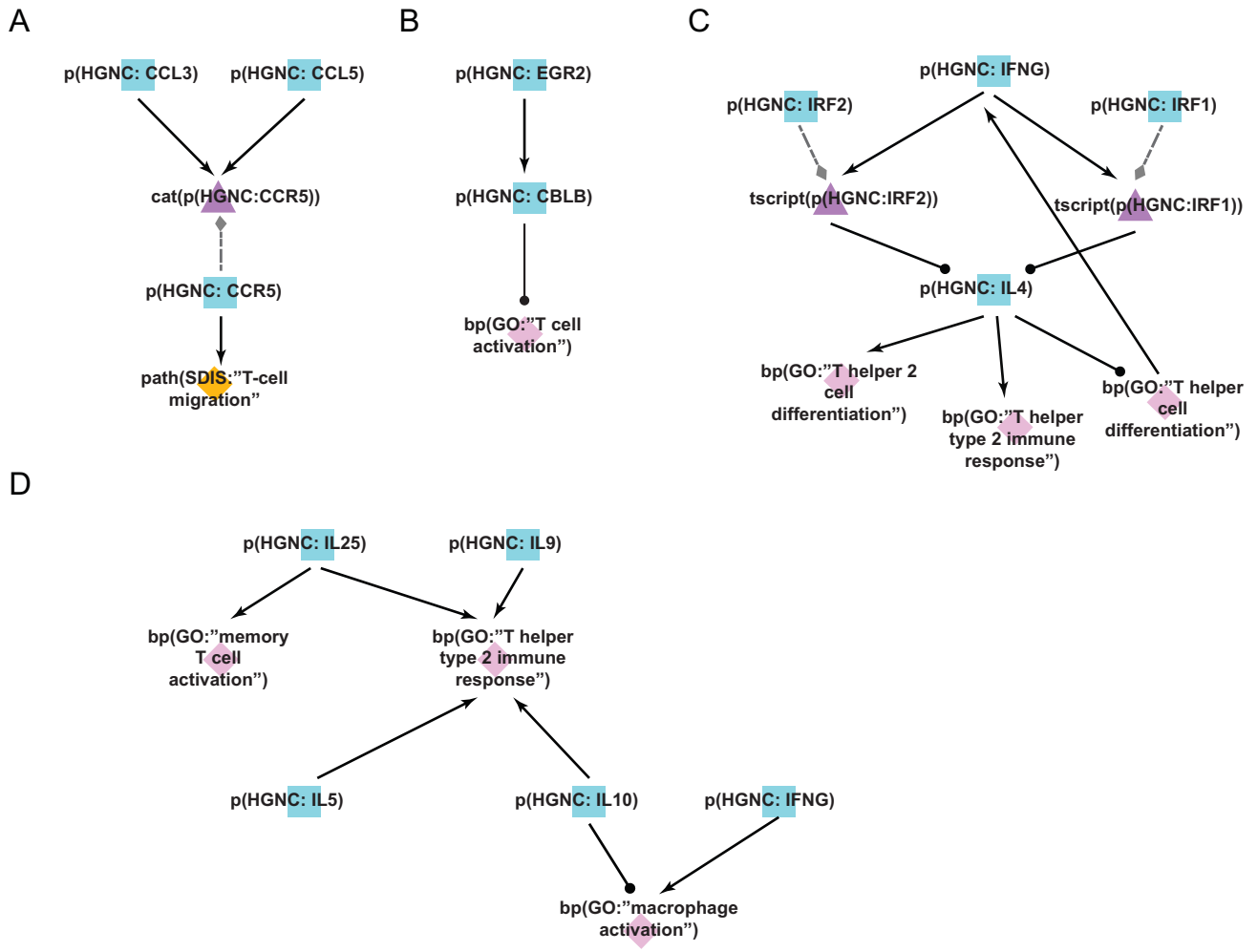

Figure 5. Enhancement of the T-cell networks during COPD network construction. A) During the literature-based COPD enhancement process in Phase 1, the protein CCL3, important for leukocyte migration and activation of T-cells, was added to the T-cell networks. **B**) During the open phase in Phase 2, the negative regulation of EGR2 on T-cell activation is a mechanistic detail that was added by the crowd. Overexpression studies demonstrated that EGR2 increased the activity of the E3 ubiquitin ligase CBL-B, which subsequently inhibited T-cell activation. **C**) During the jamboree discussions in Phase 3, the IFNG/IL-4 feedback loop mediating differentiation of Th1 vs. Th2 cellular subtypes via the activities of IRF1 and IRF2 was added to the new *Th1-2 Signaling* network model. **D**) During the jamboree discussions in Phase 3, the T-helper cell-produced chemokine effect on immune cells (e.g. IL-25 activates memory T-cells) was added to the new *Th1-2 Signaling* network. Triangle denotes activity, diamond denotes biological process or pathology, and square denotes protein abundance nodes. Solid edges denote causal relationships, dotted edges denote non-causal relationships such as a protein connected to its own activity.

biological network of T-helper cell activity and their interactions with other immune cells in the context of COPD.

Macrophage signaling

As part of the pulmonary inflammatory process network building[6], three networks (*Macrophage Differentiation, Macrophage Activation, and Macrophage-mediated Recruitment of Neutrophils*) were built to describe macrophage biology in the non-disease lung context. During the preparation phase to NVC, these three networks were merged to obtain an overall picture of macrophage biology.

Phase 1: COPD augmentation of macrophage networks

Macrophages play roles in many COPD disease processes such as clearance of apoptotic neutrophils, tissue destruction, and recruitment of other immune cells by their secretion of cytokines[36]. Macrophage signaling mechanisms were added to the network in Phase 1, with a focus on components related to efferocytosis (Figure 6A). Efferocytosis is a well-conserved mechanism for the phagocytic removal of apoptotic cells by innate immune cells, such as macrophages, and the process is critical for the resolution of inflammation via the removal of dying cells and antigenic cellular debris. Phagocytically impaired macrophages have been shown to display decreased expression of peroxisome proliferator-activated receptor gamma (PPARγ) and efferocytosis-specific bridge molecules, such as growth arrest-specific 6 (GAS6) and milk fat globule-EGF factor 8 protein (MFGE8)[37]. The number of apoptotic cells was shown to increase in COPD because of exposure of lung tissue to toxic chemicals present in cigarette smoke; for example, and their

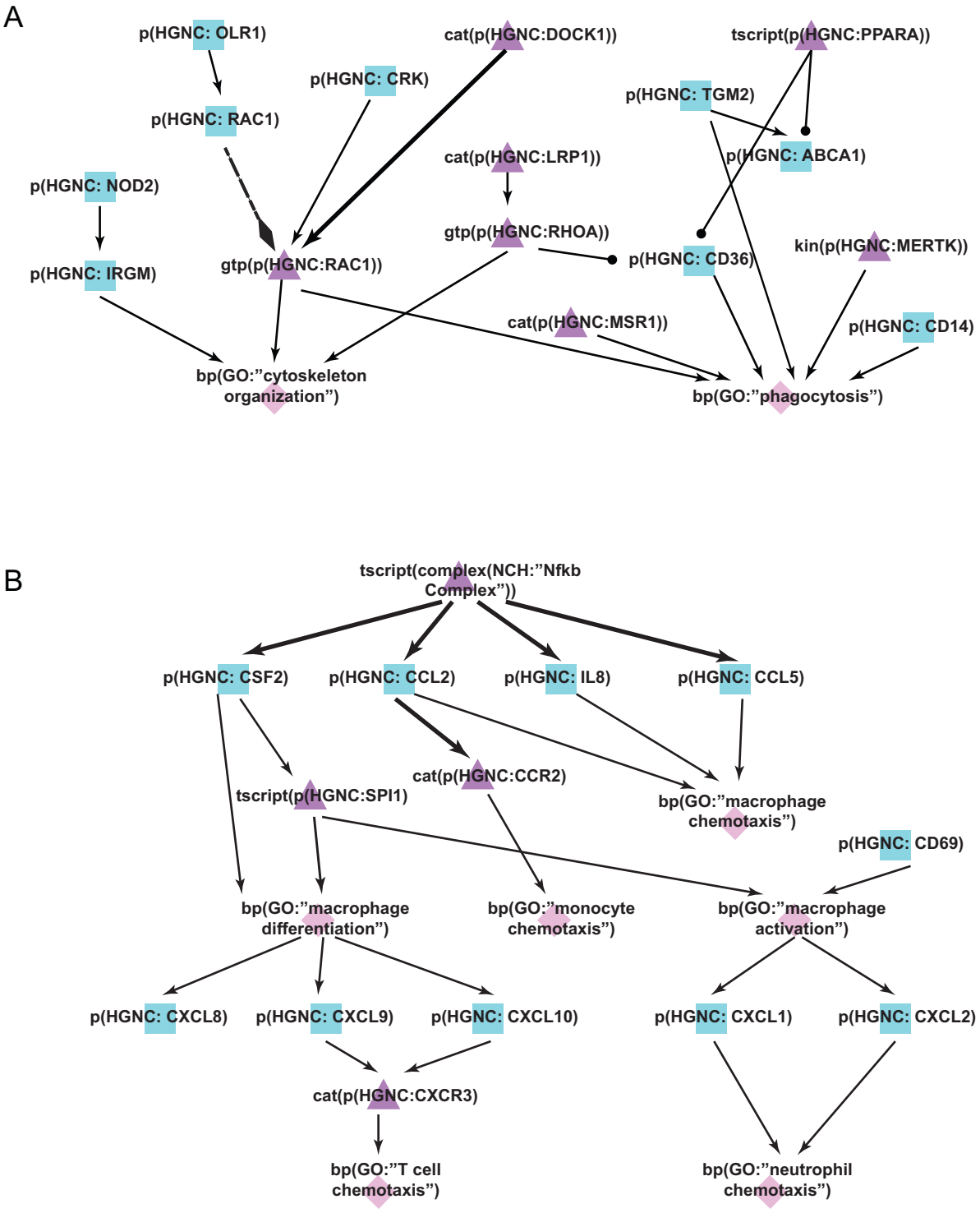

Figure 6. Enhancement of the macrophage networks during COPD network construction. A) During the literature-based COPD enhancement process in Phase 1, efferocytosis mechanisms were added to the macrophage networks to take into account its dysregulation effect in COPD. **B**) During the jamboree discussions in Phase 3, chemotaxis and differentiation mechanisms were identified and subsequently added to the latest version of the *Macrophage Signaling* network. Triangle denotes activity, diamond denotes biological process, and square denotes protein abundance nodes. Solid edges denote causal relationships, dotted edges denote non-causal relationships such as a protein connected to its own activity.

accumulation was exacerbated by the simultaneous smoke-induced impairment of the phagocytic ability of alveolar macrophages[38]. Apoptotic cells exhibit surface changes that distinguish them from viable cells, and these changes were recognized by efferocytic receptors including CD36 molecule (CD36), CD14 molecule (CD14), and Stabilin-1/2 (STAB1:STAB2)[39]. Reduced efferocytosis observed in COPD because of oxidant-driven and Rho-mediated inactivation increased the likelihood of aberrant antigen exposure from apoptotic cells, thereby perpetuating the chronic inflammatory state that is a hallmark of COPD[40–42]. In adding efferocytosis mechanisms to the macrophage network, we focused on the surface receptors and bridge proteins such as CD36 and GAS6. In total, 45 nodes and 61 new edges were added to the macrophage model during the internal COPD enhancement phase.

Phases 2 and 3: Macrophage network crowd improvements

During the open phase (Phase 2), 180 total votes were cast for network evidence, with 23 new nodes and 39 new edges added by the crowd. In addition, 72 new pieces of evidence were contributed to support pre-existing edges in the network. The surfactant protein A1 (SFTPA1), which was observed to be increased in COPD[43], was added to the network. Its effect on macrophages of increasing interleukin-1 receptor-associated kinase 3 (IRAK3) and interleukin 1, beta (IL1B) were also added to the network during the open phase. Granularity enhancements around IFNG and nucleotide-binding oligomerization domain containing 2 (NOD2), both components of inflammatory signaling, were also added to augment the network models with causal relationships proximal to COPD.

During the Phase 3 jamboree discussions, several network enhancements were made in macrophage chemotaxis and differentiation (Figure 6B). Within the chemotaxis process, the nodes chemokine (C-C motif) ligand 2 (CCL2) binding to chemokine (C-C motif) receptor 2 (CCR2) and leading to macrophage chemotaxis were added. The CD69 molecule (CD69) associated with macrophage activation by cigarette smoke was also added. In addition, the effects of activated macrophages on other immune cells were expanded within the network model, including chemokine (C-X-C motif) ligand 1 (CXCL1) and chemokine (C-X-C motif) ligand 2 (CXCL2) leading to neutrophil chemotaxis, and chemokine (C-X-C motif) ligand 9 (CXCL9) and chemokine (C-X-C motif) ligand 10 (CXCL10) binding to CXCR3 and leading to T cell recruitment. In total, 30 new nodes and 48 new edges were added to the *Macrophage Signaling* network during Phase 3, thereby providing a more comprehensive network of macrophage activation and its effect on other immune cells active in COPD.

Neutrophil signaling

As part of the pulmonary inflammatory process network building[6], two networks (*Neutrophil Response* and *Neutrophil Chemotaxis*) were built to describe neutrophil biology in the non-disease lung context. During the preparation phase to NVC, these two networks were merged to constitute the *Neutrophil Signaling* network.

Phase 1: COPD augmentation of neutrophil networks

During Phase 1, the *Neutrophil Signaling* network was enhanced primarily with components related to lipid-response pathways. In response to lung damage, leukocytes and tissue-resident cells were reported to interact to generate lipid mediators that enhance the airway immune response and engage defense mechanisms[44]. Neutrophils, endothelial cells, and macrophages generate prostaglandins and leukotrienes from arachidonic acid during the initial inflammatory response, which amplifies the inflammation signals in the local area and potentiates the process of tissue destruction[45]. Subsequently, the prostaglandins PGE2 and PGD2 are generated in a cyclooxygenase-dependent way to promote synthesis of lipid mediators with anti-inflammatory activity, such as the lipoxins. Lipoxins inhibit neutrophil recruitment to inflamed sites and suppress their pro-inflammatory actions, but promote recruitment of macrophage precursors[46]. Lipoxin A4 stimulates macrophages to phagocytose apoptotic neutrophils, and resolvins and protectins, which represent another class of lipid mediators, activate anti-inflammatory pathways and stimulate clearance of inflammatory infiltrates by macrophage phagocytosis[47–49]. In total, 9 nodes and 20 edges were added to the network model including lipid mediators such as lipoxin A4, resolvin E1, and neuroprotectin D1 (Figure 7A).

Phases 2 and 3: Neutrophil network crowd improvements

The *Neutrophil Signaling* network was the network most edited by the crowd during the open phase, with the addition of 116 new nodes, 160 new edges, 181 new pieces of evidence, and 241 votes cast. The new edges described neutrophil chemotaxis including new nodes like platelet factor 4 (PF4) and protease-activated receptor 2 (F2RL1). Chemokines such as chemokine (C-X-C motif) ligand 8 (CXCL8) and chemokine (C-X-C motif) ligand 12 (CXCL12), and members of the serine/threonine kinase (AKT) family that have also been shown to induce neutrophil chemotaxis were added to the network (Figure 7B)[50].

Following the jamboree discussions, additional signaling that described cytoskeletal and adhesion mechanisms necessary for neutrophil chemotaxis, and additional neutrophil activation mechanisms, were incorporated in the new *Neutrophil Signaling* network (Figure 7B). The role of the CDC42-WASp complex in regulating neutrophil chemotaxis at the cytoskeletal level was incorporated[51], as well as other mechanisms of neutrophil chemotaxis including the role of the complement component 5 (C5) in regulating integrin, alpha M (ITGAM)[52], and the role of CCL3/CCR5 in stimulating neutrophil migration[53]. In all, 69 nodes and 129 edges were added. The new mechanisms that were incorporated into the *Neutrophil Signaling* network added significant granularity to the neutrophil chemotaxis process, which is a key driver of the inflammatory cascade that promotes the development of COPD.

Discussion

Here we report the construction of a COPD-enhanced network model set using a novel methodology that combined traditional manual literature curation and data-driven approaches with a global crowdsourcing endeavor to generate the most comprehensive representation of biological phenomenon proximal to the onset of COPD that is available to date. The three phases of network construction each contributed in different ways to building a more comprehensive network. The Phase 1 literature and data-driven enhancement of the already existing non-diseased networks resulted in the addition of

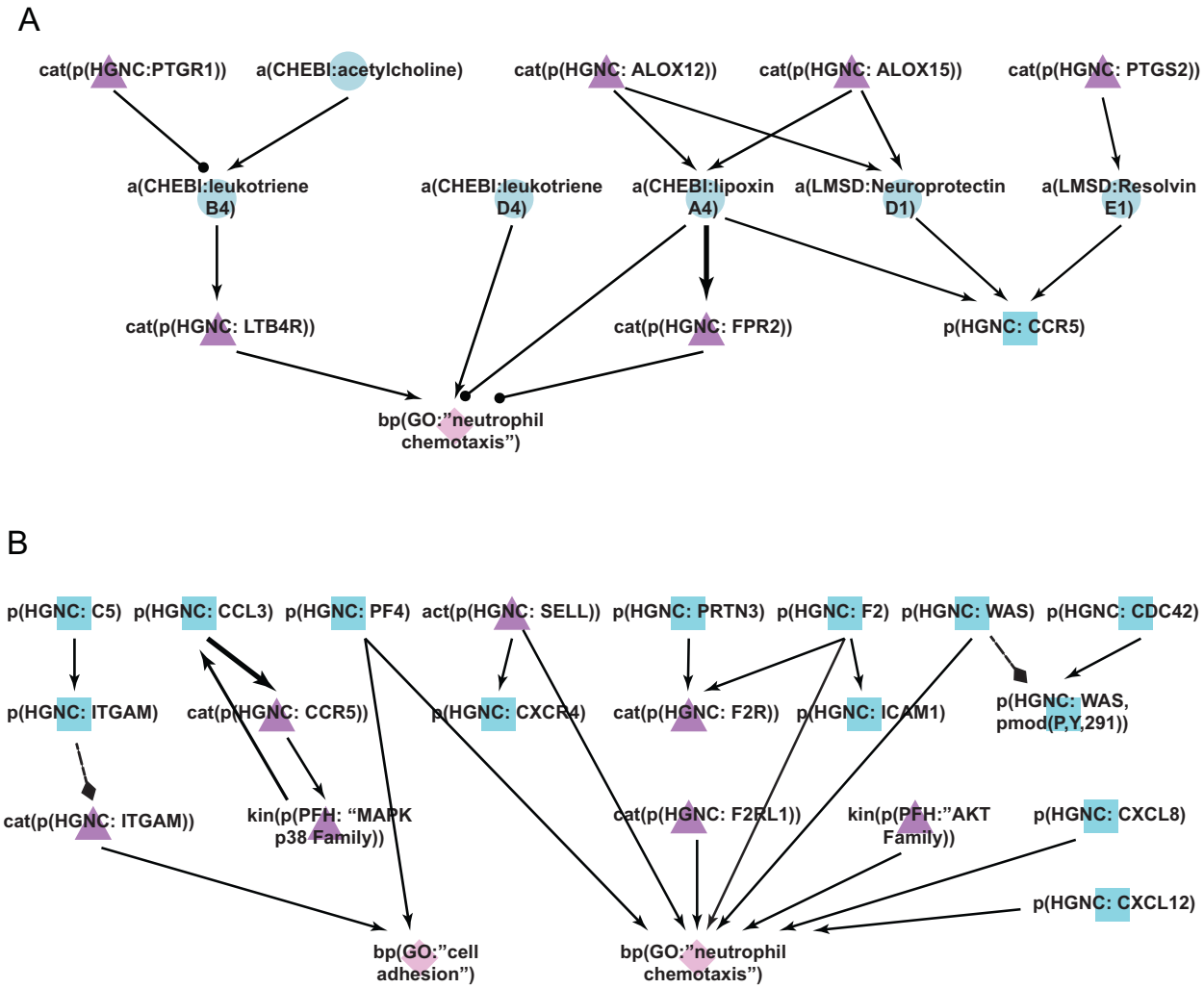

Figure 7. Enhancement of the neutrophil network during COPD network construction. A) During the literature-based COPD enhancement process in Phase 1, lipids and their effects on neutrophil chemotaxis were added to the new *Neutrophil Signaling* network. **B**) During Phases 2 and 3, neutrophil adhesion and chemotaxis mechanisms were added to the new *Neutrophil Signaling* network. Triangle denotes activity, diamond denotes biological process, circle denotes abundance, and square denotes protein abundance nodes. Solid edges denote causal relationships, dotted edges denote non-causal relationships such as a protein connected to its own activity.

COPD biomarkers and disease drivers known to be associated with COPD, while the Phase 2 crowdsourcing largely focused on contributions to cell-specific networks, and the Phase 3 jamboree discussions uncovered missing signaling processes relevant to COPD.

COPD biomarkers and processes added to non-diseased networks

During Phase 1, the non-diseased networks were expanded within the COPD context by the addition of biomarkers, disease drivers, and processes that were reported to increase in COPD, as well as mechanisms predicted in COPD data sets. Most of the edges added to the networks were lung relevant but not specifically investigated in a COPD background. Because of the limited number of mechanistic studies in COPD models that have been published, network construction was focused on adding COPD-known processes and biomarkers in tissue and experimental contexts relevant for COPD (lung, smoking) to the existing non-disease networks.

Modeling the process of efferocytosis is an example of the addition of COPD processes to the non-disease networks. The efferocytosis process of phagocytic uptake of apoptotic cells by macrophages is frequently disrupted in COPD tissue, and this disruption is thought to potentiate the chronic state of inflammation in the diseased lung[40–42]. A new network model detailing components related to efferocytosis was constructed from information available in the published literature with the majority of edges coming from general macrophage experiments. Th2 activation cascades and macrophage signaling events were also implicated generally in the context of COPD, and therefore the non-diseased network models were enhanced by the addition of these pathways from lung-relevant studies. Network

models detailing other processes not widely implicated in COPD, such as DNA damage and Notch signaling, which are more generalized conserved biological phenomenon, received very few, if any, enhancements during the COPD literature curation phase.

In addition to adding COPD processes during Phase 1, we also added COPD biomarkers and mechanisms predicted by RCR to be active in COPD data sets. Biomarkers associated with COPD included chemokines, cytokines, matrix metalloproteinases (MMPs), and other matrix degradation products. Examples of cellular mechanisms uncovered by the data-driven approach included the cytokines IL19 and IL3, as well as the serine protease inhibitor SERPINA1. IL3 is a growth-stimulating cytokine for many inflammatory cells, including macrophages, and IL19 is produced by monocytes and activates the inflammatory STAT3 pathway in several cell types. SERPINA1 is a potent elastase inhibitor, the presence of which plays a critical role in controlling the protease cascade leading to tissue destruction and emphysema. Overall, the RCR approach yielded a diverse range of biological features that were incorporated among a large percentage of the network models, thereby broadening the scope of many networks to include components with potential connections to disease that have not been investigated previously in the COPD context.

Crowdsourcing efforts focused on cell-specific networks

During the NVC, scientists from around the world browsed the publically available networks on a website, voted on and submitted new evidence, and created new nodes and edges. As may have been expected, several of the more well-studied processes in the literature (e.g. NF-kB pathways leading to inflammatory signaling) attracted a great deal of voting activity within the networks and primarily corroborated known biology. However, participants were incentivized to create new evidence to support existing edges based on the large number of points received by them for this activity. It was this aspect of the challenge that truly demonstrated the power of crowdsourcing because, in many instances, the community of users located lung-relevant and/or more recent publications to better support the existing network architecture and improve the overall relevance of the network models to COPD. With nearly 900 new pieces of evidence added by the challenge crowd, a significant overall enhancement of the networks was achieved in a relatively short time (5 months), which demonstrated the remarkable utility of harnessing knowledge from the global scientific community for a specific application. Specifically, 30% (266/885) of all the new pieces of evidence and 46% (208/451) of all the new edges that were contributed fell within three network models, namely the *Neutrophil Signaling, Macrophage Signaling*, and *Th1-Th2 Signaling* networks. These networks were edited more than other networks because of their clear boundaries, which allowed scientists to narrow their search to a particular cell type. Networks such as *Clock, Wnt*, *mTor*, and *Regulation of CDKN2A expression* were edited minimally and received more 'Down' votes than the cell-specific networks, possibly because of the more ambiguous boundaries of which cell types could be included. This observation emphasizes the need for clear boundaries in a crowdsourcing effort. In the case of general networks such as *Cell Cycle, Response to DNA Damage*, and *Oxidative Stress*, many experiments concerning these processes have been performed in cell types that were excluded in our boundaries (i.e. tumorigenic cell lines). Perhaps boundary conditions could be loosened for networks such as these if it is assumed that signaling is conserved across different cell types.

Jamboree discussions identified missing processes relevant to COPD

The final phase of network improvements emphasized the discussion and consolidation of all submissions from the challenge crowd to synthesize more holistic changes within the set of network models. During the challenge, participants worked individually on the website adding individual edges, but did not have the ability to make major changes to the structure of the network models. The in-person jamboree discussions were therefore an opportunity to implement broader changes to better represent the biological processes as they related to COPD. These discussions were led by experts in the subject matter of the processes that the networks represented. During these sessions, missing pieces of biology and the interactions of different cell types in COPD were identified. In this manner, the jamboree was very conducive to broader network structural changes that made the set of network models more informative and representative of processes implicated in COPD and, therefore, more useful to a broader group of scientists.

Unique features of the collaborative networks

In recent years, crowdsourcing has emerged as a powerful tool to address topics related to "big data" in the domain of the life sciences, particularly in topics related to systems biology. For example, the series of DREAM challenges empowered the global scientific community to build application-specific, clinically relevant predictive biological networks using vast quantities of genomic data[54]. Similarly, the recent sbv IMPROVER challenges allowed researchers to participate in collaborative competitions to validate systems biology research, for example, by testing and validating computational approaches that are used to classify clinical samples based on transcriptional data[55–57]. In the current approach, we describe a unique paradigm for biological network construction that combines a predictive computational methodology with a large-scale crowd sourcing approach to generate very comprehensive network models describing COPD pathogenesis.

Compared with other published COPD networks, the networks described here are more comprehensive in scope, are focused on molecular pathways that can drive disease rather than on descriptions of more general clinical or physiological measures, and have been improved using crowdsourcing[21–23]. The Synergy-COPD European project is similar in its goal of creating a model of COPD for better understanding of the disease by combining information from many different sources. However, Synergy-COPD comprises seven physiological-focused mathematical networks rather than the 50 molecular networks described here, and does not currently have an intuitive web interface that allows users to freely navigate the resulting networks[23].

Compared with other more general pathway approaches such as KEGG[1], the networks we describe contain edges that have one or more detailed evidences supported by a specific literature reference and contain tissue and species-level metadata. In our approach each

of these pieces of evidence under an edge can be validated with the potential for a larger crowd with wide expertise, compared to a non-crowdsourced approach where the small group constructing the networks may not be able to sufficiently cover all the expertise necessary to verify every pathway within these networks. The BEL language syntax allows many participants to contribute by standardizing the biological representation and requiring that each node be associated with a namespace, which standardizes the representation of gene names and biological processes.

The web-based platform captures network provenance, allowing for a transparent record of what has been validated with a full revision history. The uncertainty for specific edges based on voting patterns can be demonstrated with the full voting history being captured in the network versions. By incorporating a continuous "feed" of real time enhancements submitted on the website, users are able to view the most up-to-date networks at any time; network models created using other platforms not available for crowdsourced editing remain static representations of biology and frequently do not include the most recent findings from the scientific literature. Currently networks with the most recent crowd edits can be viewed, but not downloaded. Networks with changes from the most recent Jamboree meeting are made available for download.

Another novel component of these networks is the incorporation of RCR predictions to enhance the overall biological representation within the network models. RCR analysis was performed on human COPD gene expression data sets in the public domain in order to predict potential mechanisms implicated in COPD onset and include as nodes in the networks. This unbiased approach resulted in the addition of many new nodes among the networks predicted to be active based on COPD gene expression footprints that may have less well-established or direct connections to disease etiology. As such, this important aspect of network construction potentially captures those biological components that may have "emerging" roles in disease progression. The iterative nature of the network enhancement process facilitated by the Bionet platform allows for new biology and supporting evidence to be incorporated into the networks as new findings emerge in the literature and therefore generate the most comprehensive, up-to-date COPD model sets available to the scientific community.

The enhanced crowd-verified models are publicly available on the sbv IMPROVER website (https://bionet.sbvimprover.com/) and remain open to receive further enhancements from the online community. Because the first iteration of the NVC proved the effectiveness of this approach and because the networks can continue to be reviewed by the crowd, a second iteration of the NVC (NVC2) has been started so that additional modifications and recently published literature can be incorporated. This will help to continually refine the network models and strengthen the relevance to the processes that underlie the development of COPD. The crowd verification approach continues to be refined, so, in addition to disease process-centered networks, other networks including chemical-centered networks can be built using a similar approach. These networks can aid in the development of more efficient interventions and enhance toxicological assessment of environmental exposures that may also contribute to the development of COPD.

Conclusion

Here we describe a novel approach to biological network construction and have generated a suite of COPD-relevant network models that the larger scientific community is free to edit and explore. Networks are available for download from the sbv IMPROVER website (https://bionet.sbvimprover.com/) upon registration and taking a certain number of actions as a participant (e.g., voting on an evidence). Scientists from all backgrounds are encouraged to submit additional network enhancements as participants in the NVC2[58]. By building the network model set in the BEL language format, we have generated a model framework suitable for biomarker discovery and for the interpretation of transcriptomic signatures found in human lung tissue. More generally, this large assembly of biological knowledge relevant to human lung will be of great use to both academic and industry users in promoting future research in this area of great therapeutic importance.

Methods

Phase 1: COPD enhancement using data sets and literature

Networks that described molecular mechanisms of five broad biological processes were constructed previously using a literature and data mining approach. These networks cover mechanisms of cell proliferation[5], cell stress[4], DNA damage, autophagy, cell death and senescence[3], pulmonary inflammation[6], and tissue repair and angiogenesis[7] in the non-diseased pulmonary context. To create COPD-relevant networks, these non-diseased networks were enhanced by incorporating COPD mechanisms sourced using a literature and data set approach (Figure 1) in an iterative approach, as described in detail for the non-diseased network model construction, by a team of subject matter experts in computational biology, molecular biology, inhalation toxicology, and COPD.

Boundary conditions

Because the goal of the research was to understand COPD onset, the focus of these networks was on early stage COPD mechanisms (Global Initiative for Chronic Obstructive Lung Disease (GOLD) stages I and II). When supporting literature from early COPD studies was not available, stage-independent COPD studies were used. When COPD studies were not found, the inclusion criteria were expanded to studies from non-diseased context, and mechanisms active in processes implicated in COPD were incorporated into the disease models. Literature describing the processes active in acute exacerbation in COPD patients was excluded from the supporting edges of the network models. In order to focus on the molecular mechanisms most specific to early stage COPD, we also excluded context from diseases with different pathogenesis and differential diagnosis: lung cancer and non-cancerous lung diseases, such as cystic fibrosis, acute respiratory distress syndrome, idiopathic pulmonary fibrosis, septic pneumonitis, obliterative bronchiolitis, pneumoconiosis, bronchiectasis, viral and bacterial infections, and, allergic responses/asthma, bronchitis. Animal inhalation studies with solid particles (e.g. titanium dioxide, quartz, asbestos, carbon black, and diesel exhaust) were also excluded due to their specific

mode of action. Ideally, all nodes and edges of the network model would be supported by published data from experiments conducted in the tissues and cell types found in the lung under the conditions of early COPD, e.g., airway and alveolar epithelial cells, lung fibroblasts, resident and recruited immune cells, and microvascular cells. These were prioritized but the respective cell types were also considered from other tissue origin if such lung specific context was not reported in the literature. For *in vitro*-specific exclusion criteria, tumor-derived cell lines, immortalized cell lines, neuronal cells, and cell types that are not found in the respiratory/vascular system were excluded. In some cases, we made exceptions and included non-lung cell types for canonical mechanisms for which there was additional evidence from the literature that the relationship was not tissue-specific but could also take place in the lung. Human-specific connections were prioritized, but where human data were not available, knowledge has been augmented with orthologous causal assertions derived from rat and mouse sources included after homologization in the Selventa knowledgebase where human data were not available[5].

The 90 previously published non-diseased network models used for the initial substrate included networks involved in cell proliferation[5], cell stress[4], DNA damage, apoptosis, senescence, autophagy, necroptosis (DACS)[3], pulmonary inflammation (IPN)[6], and tissue repair and angiogenesis (TRAG)[7]. The *Endothelial Shear Stress* network from the cell stress model was excluded because the focus of the COPD Network was to describe lung biology.

Literature enhancement

We conducted a broad survey of the literature to locate studies that had investigated the mechanistic biology of COPD pathogenesis and processes involved in COPD. Potential COPD biomarkers from sputum, bronchoalveolar lavage, and mouse and human blood samples, and mechanisms that regulate COPD processes were gathered from the literature and curated. Because only a small number of the studies had focused on early COPD, we expanded our searches to include stage-independent COPD studies, but excluded late-stage processes. Some processes known to be closely linked to COPD pathogenesis (e.g. B-cell activation and T-cell recruitment to lung tissue) have not been studied directly in the disease context; however, literature that detailed cell-type-specific canonical biology was sourced irrespective of the disease context.

Data enhancement

RCR was performed using Gene Expression Omnibus (GEO) COPD and emphysema data sets from lung, small airway, and alveolar macrophages of early COPD patients and healthy smokers (see Dataset)[59–63]. RCR has been used previously to predict upstream regulators from transcriptomic data[8]. Mechanisms that were predicted by RCR to be active and that were not already incorporated in the non-diseased networks were vetted on an individual basis to locate supporting literature for their potential involvement in COPD pathogenesis. Mechanisms that had not been studied directly in a COPD context were evaluated in an expanded tissue context to consider tissue deemed disease-relevant (e.g. alveolar macrophages). Mechanisms that were deemed relevant were connected in the most appropriate network based on their probable roles in COPD or lung biology.

Network agglomeration

To generate a more concise model set for presentation to the crowd during the NVC, we consolidated networks associated with related biological processes among the 90 COPD-enhanced networks. An example of this consolidation is the merging of three non-disease networks related to T-helper 1 cells (*Th1 Differentiation, Th1 Response,* and *T-cell Recruitment/Activation*) into a single new *Th1 Signaling* network. Fifty-six of the original 90 networks were combined into a concise set of 16 network models; the remaining 34 networks remained as standalone network models (see Dataset), yielding a final set of 50 models that were posted on the NVC website for review by the scientific crowd. In addition to the network agglomeration, protein, gene expression, and secretion edges were agglomerated to reduce the number of edges required for verification.

Phase 2: NVC Open Phase

The crowd verification process of improving biological networks has been published previously[20]. Briefly, the full set of 50 COPD-relevant network models was posted on the BioNet web portal[58] for a period of 20 weeks (the "Open Phase"), during which time a global community of participants were invited to submit biological improvements to the models. The improvements included submission of new evidence, additional literature publications to support existing network edges, and submission of new biological edges with supporting evidence for relationships that were not represented in a network. Users could also vote on evidence to indicate agreement or disagreement with its appropriateness within the network structure; disagreements often indicated improper tissue or experimental context for the given network. Evidence that received at least four 'Up' votes was "locked" to indicate crowd approval and evidence that received at least four 'Down' votes was "locked" to indicate rejection by the crowd. Depending on the frequency and type of submitted improvements, participants received credit points and were assigned a dynamic ranking on the community Leaderboard. For more information about the NVC challenge, see the 5-minute overview videos at https://sbvimprover.com/challenge-3/videos or the 1-hour webinars at https://sbvimprover.com/challenge-3/tutorials.

Phase 3: Jamboree meeting

When the open phase was closed, the top-ranked participants were invited to a 3-day-long in-person jamboree to discuss improvements submitted by the community and to further refine the network models. Subject matter experts in lung, COPD, and network biology, as well as experts in other related biological processes, were also invited to guide the discussions and to provide expert feedback of missing or misrepresented signaling. Scientists involved in the construction of the original non-disease networks and Phase 1-enhanced networks were present to provide feedback for the rationale behind the boundary conditions and the mechanics of network construction and BEL. During the jamboree, 15 networks were prioritized to discuss in small groups of 6–10 people focusing on one network at a time. At the end of each session, final decisions were made about

follow-up actions for each network and these actions were carried out subsequently by the scientists who constructed the original networks because of their familiarity with the mechanics of network construction and BEL.

The changes to the 15 networks that were discussed during the jamboree are posted online[58] in open-source XGMML (eXtensible Graph Markup and Modeling Language) format.

BEL: the language of the networks

The networks were built using the Biological Expression Language (BEL), which is an open source language that can represent scientific findings in the life sciences in a computable form[64]. BEL was designed to represent scientific findings by capturing causal and correlative relationships in context, where context can include information about the biological and experimental system in which the relationships were observed and the supporting publication citations. The structure of a BEL node, which includes the biological entity, the namespace or database to standardize the nomenclature of the entity, and the function that describes the type of entity (protein, chemical, biological process, family, complex, etc), is shown in Figure 8. Table 1 and Table 2 show the definition of the prefixes for BEL namespaces and functions that appear in the networks.

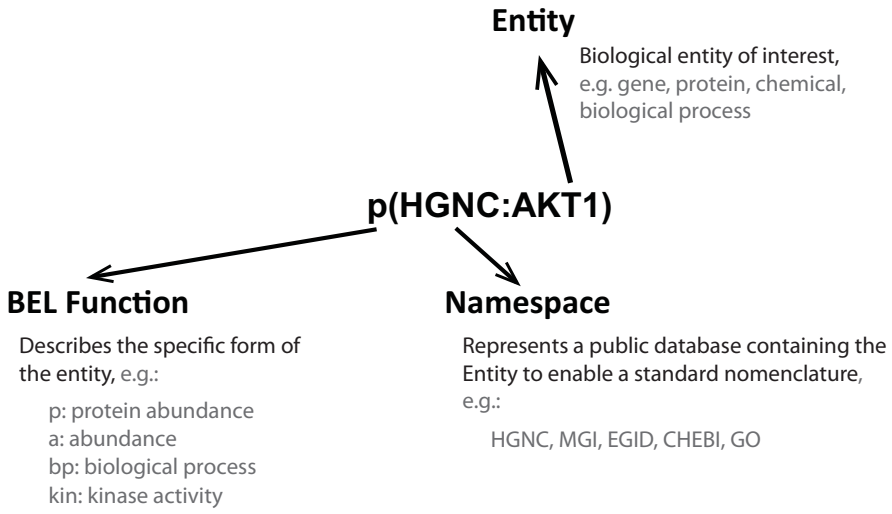

Figure 8. Structure of a BEL node. A BEL term is the standard way a node is described. It includes an entity that is described using standard nomenclature in the Namespace and the Function fields of the entity.

Table 1. BEL functions.

Prefix	Function
a	abundance
bp	biological process
cat	catalytic activity
sec	cell secretion
surf	cell surface expression
chap	chaperone activity
complex	complex abundance
composite	composite abundance
deg	degradation
fus	fusion
g	gene abundance
gtp	GTP bound activity
kin	kinase activity
m	microRNA abundance
act	molecular activity
path	pathology
pep	peptidase activity
phos	phosphatase activity
p	protein abundance
pmod	protein modification
rxn	reaction
ribo	ribosylation activity
r	RNA abundance
sub	substitution
tscript	transcriptional activity
tloc	translocation
tport	transport activity
trunc	truncation

Table 2. BEL namespaces.

Prefix	Namespace
EGID	Entrez Gene Identifiers
HGNC	HGNC Approved Gene Symbols
MGI	MGI Approved Gene Symbols
RGD	RGD Approved Gene Symbols
SPAC	Swiss-Prot Proteins (Accession Numbers)
SP	Swiss-Prot (Entry Names)
HGU95AV2	Affymetrix GeneChip Human Genome U95Av2
HGU133AB	Affymetrix GeneChip Human Genome U133AB
HGU133P2	Affymetrix GeneChip Human Genome U133Plus2
MGU74ABC	Affymetrix GeneChip Mouse Genome U74ABC
MG430AB	Affymetrix GeneChip Mouse Expression Set 430
MG4302	Affymetrix GeneChip Mouse Genome 430 2.0
MG430A2	Affymetrix GeneChip Mouse Genome 430A 2.0
RG230AB	Affymetrix GeneChip Rat Expression Set 230AB
RG2302	Affymetrix GeneChip Rat Genome 230 2.0
CHEBIID	Chemicals of Biological Interest (Identifiers)
CHEBI	Chemicals of Biological Interest (Names)
LMSD*	LIPID MAPS Structure Database (Names)
GOAC	GO Biological Processes (Accession Numbers)
GO	GO Biological Processes (Names)
MESHPP	MeSH Phenomena and Processes (Names)
MESHD	MeSH Diseases (Names)
MESHCL	MeSH Cell Locations (Names)
GOCCACC	GO Cellular Component (Accession Numbers)
GOCCTERM	GO Cellular Component (Terms)
PFH	Named Human Protein Families
NCH	Named Human Complexes
PFM	Named Mouse Protein Families
NCM	Named Mouse Complexes
PFR	Named Rat Protein Families
NCR	Named Rat Complexes
SCHEM	Selventa Legacy Chemical Names
SDIS	Selventa Legacy Disease Names

*Unofficial BEL namespace to be formalized in BEL 2.0

Data availability

Up-to-date networks including all users' activity can be browsed freely on the Bionet website (https://bionet.sbvimprover.com/). Permanent URLs to each network are listed in the associated Data Set (Original networks, NVC networks and their descriptions). Networks can be downloaded by logged in users who had a few actions on the site as XGMML file for offline use in the version that started a verification phase, i.e. after review and QC by experts. The 15 networks discussed in the jamboree are available in a post-jamboree version. Moreover, different versions of the networks are available to browse and download in diverse formats from the CBN database available at causalbionet.com.

Data availability

Figshare: Original networks, NVC networks and COPD data sets used in: Enhancement of COPD biological networks using a web-based collaboration interface http://dx.doi.org/10.6084/m9.figshare.1284583[65]

Author contributions

WKS, MT, SB, JP, and BF constructed and reviewed the networks. JH and MCP conceived the idea and managed the project. JP and BF wrote the manuscript text. JP, BF, and SB developed the figures. The Challenge Best Performers: IB, VB, OVB, VC, NDD, LF, SG, JG, GIK, EK, RK, NL, QL, SM, YO, SCP, APM ES, SS, GSN, IS, AT, GVF, JW, IY, and MZ participated in the network verification process and in the review of the manuscript. All authors read and approved the final manuscript.

Competing interests

Selventa and PMI authors performed this work under a joint research collaboration funded by PMI. The Network Verification Challenge was funded by PMI.

Grant information

The research described in this article was funded by Philip Morris International in a collaborative project with Selventa.

Acknowledgements

The authors thank IBM for their help in organizing the Network Verification Challenge and jamboree, and Michael Maria and Jean Binder for their help in project management and preparation of this manuscript. The project team expresses their gratitude to the subject matter experts and moderators who actively participated in the jamboree: Maria Laura Belladonna, Michael Borchers, Maciej Cabanski, Natalia Boukharov, Stephan Gebel, Ignacio Gonzalez Suarez, Daniele Guardavaccaro, Anita Iskandar, Ulrike Kogel, Katica Jankovic, David Kling, Sophia Kossida, Hector de Leon, Karsta Luettich, Yukiko Matsuoka, Dragana Mitic Potkrajac, Michael Peck and Carine Poussin.

References

1. Kanehisa M, Goto S: **KEGG: kyoto encyclopedia of genes and genomes.** *Nucleic Acids Res.* 2000; **28**(1): 27–30.
PubMed Abstract | Publisher Full Text | Free Full Text
2. Kanehisa M, Goto S, Sato Y, *et al.*: **Data, information, knowledge and principle: back to metabolism in KEGG.** *Nucleic Acids Res.* 2014; **42**(Database issue): D199–205.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Gebel S, Lichtner RB, Frushour B, *et al.*: **Construction of a computable network model for DNA damage, autophagy, cell death, and senescence.** *Bioinform Biol Insights.* 2013; **7**: 97–117.
PubMed Abstract | Publisher Full Text | Free Full Text
4. Schlage WK, Westra JW, Gebel S, *et al.*: **A computable cellular stress network model for non-diseased pulmonary and cardiovascular tissue.** *BMC Syst Biol.* 2011; **5**: 168.
PubMed Abstract | Publisher Full Text | Free Full Text
5. Westra JW, Schlage WK, Frushour BP, *et al.*: **Construction of a computable cell proliferation network focused on non-diseased lung cells.** *BMC Syst Biol.* 2011; **5**: 105.
PubMed Abstract | Publisher Full Text | Free Full Text
6. Westra JW, Schlage WK, Hengstermann A, *et al.*: **A modular cell-type focused inflammatory process network model for non-diseased pulmonary tissue.** *Bioinform Biol Insights.* 2013; **7**: 167–192.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Park JS, Schlage WK, Frushour BP, *et al.*: **Construction of a Computable Network Model of Tissue Repair and Angiogenesis in the Lung.** *J Clinic Toxicol.* 2013; **S12**: 002.
Publisher Full Text
8. Catlett NL, Bargnesi AJ, Ungerer S, *et al.*: **Reverse causal reasoning: applying qualitative causal knowledge to the interpretation of high-throughput data.** *BMC Bioinformatics.* 2013; **14**: 340.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Lopez AD, Murray CC: **The global burden of disease, 1990–2020.** *Nat Med.* 1998; **4**(11): 1241–1243.
PubMed Abstract | Publisher Full Text
10. Ojo O, Lagan AL, Rajendran V, *et al.*: **Pathological changes in the COPD lung mesenchyme - Novel lessons learned from *in vitro* and *in vivo* studies.** *Pulm Pharmacol Ther.* 2014; **29**(2): 121–8.
PubMed Abstract | Publisher Full Text
11. From the **Global Strategy for the Diagnosis, Management and Prevention of COPD, Global Initiative for Chronic Obstructive Lung Disease (GOLD)**. 2014.
Reference Source.
12. Chapman RS, He X, Blair AE, *et al.*: **Improvement in household stoves and risk of chronic obstructive pulmonary disease in Xuanwei, China: retrospective cohort study.** *BMJ.* 2005; **331**(7524): 1050.
PubMed Abstract | Publisher Full Text | Free Full Text
13. Ekici A, Ekici M, Kurtipek E, *et al.*: **Obstructive airway diseases in women exposed to biomass smoke.** *Environ Res.* 2005; **99**(1): 93–98.
PubMed Abstract | Publisher Full Text
14. Hnizdo E, Sullivan PA, Bang KM, *et al.*: **Airflow obstruction attributable to work in industry and occupation among U.S. race/ethnic groups: a study of NHANES III data.** *Am J Ind Med.* 2004; **46**(2): 126–135.
PubMed Abstract | Publisher Full Text
15. Winchester JW: **Regional anomalies in chronic obstructive pulmonary disease; comparison with acid air pollution particulate characteristics.** *Arch Environ Contam Toxicol.* 1989; **18**(1–2): 291–306.
PubMed Abstract | Publisher Full Text
16. Fischer BM, Pavlisko E, Voynow JA: **Pathogenic triad in COPD: oxidative stress, protease-antiprotease imbalance, and inflammation.** *Int J Chron Obstruct Pulmon Dis.* 2011; **6**: 413–421.
PubMed Abstract | Publisher Full Text | Free Full Text
17. Calverley PM, Walker P: **Chronic obstructive pulmonary disease.** *Lancet.* 2003; **362**(9389): 1053–1061.
PubMed Abstract | Publisher Full Text
18. Adcock IM, Caramori G, Barnes PJ: **Chronic obstructive pulmonary disease and lung cancer: new molecular insights.** *Respiration; international review of thoracic diseases.* 2011; **81**(4): 265–284.
PubMed Abstract | Publisher Full Text
19. Barnes PJ: **Chronic obstructive pulmonary disease * 12: New treatments for COPD.** *Thorax.* 2003; **58**(9): 803–808.
PubMed Abstract | Publisher Full Text | Free Full Text
20. Ansari S, Binder J, Boue S, *et al.*: **On Crowd-verification of Biological Networks.** *Bioinform Biol Insights.* 2013; **7**: 307–325.
PubMed Abstract | Publisher Full Text | Free Full Text
21. Turan N, Kalko S, Stincone A, *et al.*: **A systems biology approach identifies molecular networks defining skeletal muscle abnormalities in chronic obstructive pulmonary disease.** *PLoS Comput Biol.* 2011; **7**(9): e1002129.
PubMed Abstract | Publisher Full Text | Free Full Text
22. Kaneko Y, Yatagai Y, Yamada H, *et al.*: **The search for common pathways underlying asthma and COPD.** *Int J Chron Obstruct Pulmon Dis.* 2013; **8**: 65–78.
PubMed Abstract | Publisher Full Text | Free Full Text
23. Maier D, Kalus W, Wolff M, *et al.*: **Knowledge management for systems biology a general and visually driven framework applied to translational medicine.** *BMC Syst Biol.* 2011; **5**: 38.
PubMed Abstract | Publisher Full Text | Free Full Text
24. Cohen HY, Lavu S, Bitterman KJ, *et al.*: **Acetylation of the C terminus of Ku70 by CBP and PCAF controls Bax-mediated apoptosis.** *Mol Cell.* 2004; **13**(5): 627–638.
PubMed Abstract | Publisher Full Text
25. Cohen HY, Miller C, Bitterman KJ, *et al.*: **Calorie restriction promotes mammalian cell survival by inducing the SIRT1 deacetylase.** *Science.* 2004; **305**(5682): 390–392.
PubMed Abstract | Publisher Full Text
26. Luo C, Li Y, Yang L, *et al.*: **A cigarette component acrolein induces accelerated senescence in human diploid fibroblast IMR-90 cells.** *Biogerontology.* 2013; **14**(5): 503–511.
PubMed Abstract | Publisher Full Text
27. Motta MC, Divecha N, Lemieux M, *et al.*: **Mammalian SIRT1 represses forkhead transcription factors.** *Cell.* 2004; **116**(4): 551–563.
PubMed Abstract | Publisher Full Text
28. Yao H, Chung S, Hwang JW, *et al.*: **SIRT1 protects against emphysema via FOXO3-mediated reduction of premature senescence in mice.** *J Clin Invest.* 2012; **122**(6): 2032–2045.
PubMed Abstract | Publisher Full Text | Free Full Text
29. Mortaz E, Kraneveld AD, Smit JJ, *et al.*: **Effect of cigarette smoke extract on dendritic cells and their impact on T-cell proliferation.** *PLoS One.* 2009; **4**(3): e4946.
PubMed Abstract | Publisher Full Text | Free Full Text
30. Freeman CM, Curtis JL, Chensue SW: **CC chemokine receptor 5 and CXC chemokine receptor 6 expression by lung CD8+ cells correlates with chronic obstructive pulmonary disease severity.** *Am J Pathol.* 2007; **171**(3): 767–776.
PubMed Abstract | Publisher Full Text | Free Full Text
31. Safford M, Collins S, Lutz MA, *et al.*: **Egr-2 and Egr-3 are negative regulators of T cell activation.** *Nat Immunol.* 2005; **6**(5): 472–480.
PubMed Abstract | Publisher Full Text
32. Elser B, Lohoff M, Kock S, *et al.*: **IFN-gamma represses IL-4 expression via IRF-1 and IRF-2.** *Immunity.* 2002; **17**(6): 703–712.
PubMed Abstract | Publisher Full Text
33. Han ST, Mosher DF: **IL-5 induces suspended eosinophils to undergo unique global reorganization associated with priming.** *Am J Respir Cell Mol Biol.* 2014; **50**(3): 654–664.
PubMed Abstract | Publisher Full Text | Free Full Text
34. Gemelli C, Zanocco Marani T, Bicciato S, *et al.*: **MafB is a downstream target of the IL-10/STAT3 signaling pathway, involved in the regulation of macrophage de-activation.** *Biochim Biophys Acta.* 2014; **1843**(5): 955–964.
PubMed Abstract | Publisher Full Text
35. Ma B, Kang MJ, Lee CG, *et al.*: **Role of CCR5 in IFN-gamma-induced and cigarette smoke-induced emphysema.** *J Clin Invest.* 2005; **115**(12): 3460–3472.
PubMed Abstract | Publisher Full Text | Free Full Text
36. Shapiro SD: **The macrophage in chronic obstructive pulmonary disease.** *Am J Respir Crit Care Med.* 1999; **160**(5 Pt 2): S29–32.
PubMed Abstract | Publisher Full Text
37. Wang X, Bu HF, Zhong W, *et al.*: **MFG-E8 and HMGB1 are involved in the mechanism underlying alcohol-induced impairment of macrophage efferocytosis.** *Mol Med.* 2013; **19**: 170–182.
PubMed Abstract | Publisher Full Text | Free Full Text
38. Brusselle GG, Joos GF, Bracke KR: **New insights into the immunology of chronic obstructive pulmonary disease.** *Lancet.* 2011; **378**(9795): 1015–1026.
PubMed Abstract | Publisher Full Text
39. Korns D, Frasch SC, Fernandez-Boyanapalli R, *et al.*: **Modulation of macrophage efferocytosis in inflammation.** *Front Immunol.* 2011; **2**: 57.
PubMed Abstract | Publisher Full Text | Free Full Text
40. Cosio MG, Saetta M, Agusti A: **Immunologic aspects of chronic obstructive pulmonary disease.** *N Engl J Med.* 2009; **360**(23): 2445–2454.
PubMed Abstract | Publisher Full Text
41. Hodge S, Hodge G, Ahern J, *et al.*: **Smoking alters alveolar macrophage recognition and phagocytic ability: implications in chronic obstructive pulmonary disease.** *Am J Respir Cell Mol Biol.* 2007; **37**(6): 748–755.
PubMed Abstract | Publisher Full Text
42. Richens TR, Linderman DJ, Horstmann SA, *et al.*: **Cigarette smoke impairs clearance of apoptotic cells through oxidant-dependent activation of RhoA.** *Am J Respir Crit Care Med.* 2009; **179**(11): 1011–1021.
PubMed Abstract | Publisher Full Text | Free Full Text
43. Ishikawa N, Hattori N, Tanaka S, *et al.*: **Levels of surfactant proteins A and D and KL-6 are elevated in the induced sputum of chronic obstructive pulmonary disease patients: a sequential sputum analysis.** *Respiration; international review of thoracic diseases.* 2011; **82**(1): 10–18.
PubMed Abstract | Publisher Full Text
44. Herold S, Mayer K, Lohmeyer J: **Acute lung injury: how macrophages orchestrate resolution of inflammation and tissue repair.** *Front Immunol.* 2011; **2**: 65.
PubMed Abstract | Publisher Full Text | Free Full Text
45. Funk CD: **Prostaglandins and leukotrienes: advances in eicosanoid biology.**

Science. 2001; **294**(5548): 1871–1875.
PubMed Abstract | Publisher Full Text

46. Chiang N, Serhan CN, Dahlen SE, *et al.*: **The lipoxin receptor ALX: potent ligand-specific and stereoselective actions *in vivo*.** *Pharmacol Rev.* 2006; **58**(3): 463–487.
PubMed Abstract | Publisher Full Text

47. Godson C, Mitchell S, Harvey K, *et al.*: **Cutting edge: lipoxins rapidly stimulate nonphlogistic phagocytosis of apoptotic neutrophils by monocyte-derived macrophages.** *J Immunol.* 2000; **164**(4): 1663–1667.
PubMed Abstract | Publisher Full Text

48. Arita M, Ohira T, Sun YP, *et al.*: **Resolvin E1 selectively interacts with leukotriene B4 receptor BLT1 and ChemR23 to regulate inflammation.** *J Immunol.* 2007; **178**(6): 3912–3917.
PubMed Abstract | Publisher Full Text

49. Schwab JM, Chiang N, Arita M, *et al.*: **Resolvin E1 and protectin D1 activate inflammation-resolution programmes.** *Nature.* 2007; **447**(7146): 869–874.
PubMed Abstract | Publisher Full Text | Free Full Text

50. Rose JJ, Foley JF, Yi L, *et al.*: **Cholesterol is obligatory for polarization and chemotaxis but not for endocytosis and associated signaling from chemoattractant receptors in human neutrophils.** *J Biomed Sci.* 2008; **15**(4): 441–461.
PubMed Abstract | Publisher Full Text

51. Kumar S, Xu J, Perkins C, *et al.*: **Cdc42 regulates neutrophil migration via crosstalk between WASp, CD11b, and microtubules.** *Blood.* 2012; **120**(17): 3563–3574.
PubMed Abstract | Publisher Full Text | Free Full Text

52. Schmid E, Warner RL, Crouch LD, *et al.*: **Neutrophil chemotactic activity and C5a following systemic activation of complement in rats.** *Inflammation.* 1997; **21**(3): 325–333.
PubMed Abstract | Publisher Full Text

53. Ottonello L, Montecucco F, Bertolotto M, *et al.*: **CCL3 (MIP-1alpha) induces *in vitro* migration of GM-CSF-primed human neutrophils via CCR5-dependent activation of ERK 1/2.** *Cell Signal.* 2005; **17**(3): 355–363.
PubMed Abstract | Publisher Full Text

54. Prill RJ, Saez-Rodriguez J, Alexopoulos LG, *et al.*: **Crowdsourcing network inference: the DREAM predictive signaling network challenge.** *Sci Signal.* 2011; **4**(189): mr7.
PubMed Abstract | Publisher Full Text | Free Full Text

55. Meyer P, Alexopoulos LG, Bonk T, *et al.*: **Verification of systems biology research in the age of collaborative competition.** *Nat Biotechnol.* 2011; **29**(9): 811–815.
PubMed Abstract | Publisher Full Text

56. Meyer P, Hoeng J, Rice JJ, *et al.*: **Industrial methodology for process verification in research (IMPROVER): toward systems biology verification.** *Bioinformatics.* 2012; **28**(9): 1193–1201.
PubMed Abstract | Publisher Full Text | Free Full Text

57. Tarca AL, Lauria M, Unger M, *et al.*: **Strengths and limitations of microarray-based phenotype prediction: lessons learned from the IMPROVER Diagnostic Signature Challenge.** *Bioinformatics.* 2013; **29**(22): 2892–2899.
PubMed Abstract | Publisher Full Text | Free Full Text

58. [https://bionet.sbvimprover.com/].

59. Ezzie ME, Crawford M, Cho JH, *et al.*: **Gene expression networks in COPD: microRNA and mRNA regulation.** *Thorax.* 2012; **67**(2): 122–131.
PubMed Abstract | Publisher Full Text

60. Gemelli C, Orlandi C, Zanocco Marani T, *et al.*: **The vitamin D3/Hox-A10 pathway supports MafB function during the monocyte differentiation of human CD34+ hemopoietic progenitors.** *J Immunol.* 2008; **181**(8): 5660–5672.
PubMed Abstract | Publisher Full Text

61. Ammous Z, Hackett NR, Butler MW, *et al.*: **Variability in small airway epithelial gene expression among normal smokers.** *Chest.* 2008; **133**(6): 1344–1353.
PubMed Abstract | Publisher Full Text | Free Full Text

62. Shaykhiev R, Otaki F, Bonsu P, *et al.*: **Cigarette smoking reprograms apical junctional complex molecular architecture in the human airway epithelium *in vivo*.** *Cell Mol Life Sci.* 2011; **68**(5): 877–892.
PubMed Abstract | Publisher Full Text | Free Full Text

63. Shaykhiev R, Krause A, Salit J, *et al.*: **Smoking-dependent reprogramming of alveolar macrophage polarization: implication for pathogenesis of chronic obstructive pulmonary disease.** *J Immunol.* 2009; **183**(4): 2867–2883.
PubMed Abstract | Publisher Full Text | Free Full Text

64. [http://www.openbel.org/].

65. Boue S, Fields B, Hoeng J, *et al.*: **Original networks, NVC networks and COPD data sets used in: Enhancement of COPD biological networks using a web-based collaboration interface.** *Figshare.* 2014.
Data Source

Open Peer Review

Current Peer Review Status: ✔ ✔ ✔

Version 1

Reviewer Report 02 March 2015

https://doi.org/10.5256/f1000research.6402.r7527

© **2015 Ruch P.** This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

✔ **Patrick Ruch**
BiTeM group, Library and Information Sciences, University of Applied Sciences—HEG, Geneva, Switzerland

I would welcome more information about the original networks, which were generated out of text/data mining. In particular, I would like to know what keywords were used to fetch the source articles (a set of genes/gene products, e.g. AQP5, MUC5...; a list of pathologies and synonyms, e.g. COPD Chronic Obstructive Pulmonary Disorders, early-stage COPD; a list of chemical compounds...) , what search engines were used (e.g. PubMed), what collection (MEDLINE, PubMed Central full-texts). Additionally quantitative details would be welcome also: how many abstracts or full-text articles were collected first?

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 12 May 2015

Stephanie Boue, Philip Morris Products S.A., Quai Jeanrenaud 5, 2000 Neuchâtel, Switzerland

Networks were built in multiple phases, as briefly described in M&M and in figure 1. The “original” networks were built in a non-disease context as described in detail in the respective publications [1-5]. Based on relevant reviews in the area and context of interest, research papers that report the causal mechanistic relationships relevant for a specific biological process were identified and the causal relationships were extracted and added to the model. Mechanistic relationships were also added from an existing Knowledgebase, a collection of causal relationships curated over more than ten years from over 40,000 papers and over 500,000 statements.

Competing Interests: No competing interests were disclosed.

Reviewer Report 02 March 2015

https://doi.org/10.5256/f1000research.6402.r7526

© 2015 Bader G. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Gary D. Bader
The Donnelly Centre, University of Toronto, Toronto, ON, Canada

High quality curation by trained database curators is needed in our community to convert the literature to computable models, but it is difficult to imagine how manual curation will scale to handle the ever-growing data generation rate in biology. Thus, the biological research community needs to figure out how to get crowdsourcing working for everyone as a tool to improve access to computable data. This paper does a very good job of describing how a set of COPD networks were constructed and enhanced (they grew in size and level of detail) through an interesting three-phase process. However, it would be useful to better describe the utility of the resulting networks and to further analyze the crowdsourcing process itself. Addressing these points will give the work a broader impact.

Utility of networks:

It is not clear what advantages the use of causal networks brings compared to more established models in the community, such as molecular interaction networks used by many algorithms (e.g. gene function prediction, module detection, interpretation of molecular profile data, network biomarkers) or detailed biochemical pathway models (used by most textbooks and pathway databases). While many results are published in terms of causal networks (e.g. A activates B), one important issue with networks constructed by collecting these relationships is that they may be difficult to integrate across resources since they are context specific: A may activate B in the lung, but inhibit B in the heart and when these are integrated, a conflict arises. Many computational analysis methods require integration of networks from multiple sources to construct the largest available network and integrate this data with disease-specific molecular profile data (e.g. gene expression data) to gain context (as it seems is done in the RCR approach). It would be useful for the authors to further discuss the utility of context-specific causal networks for follow on discovery.

I only noticed one sentence mentioning use: “By building the network model set in the BEL language format, we have generated a model framework suitable for biomarker discovery and for the interpretation of transcriptomic signatures found in human lung tissue.” However, this sentence is not clear and doesn’t cite any prior literature. How does using the BEL format create models suitable for biomarker discovery? Can’t molecular interaction or other types of networks

also be used for biomarker discovery? What type of biomarker discovery is referred to here? How are transcriptomic signatures interpreted and analyzed?

Crowdsourcing comments:
"Networks that were not enhanced with COPD-specific mechanisms from the literature or RCR included the DNA Damage and Notch Signaling networks. Although both these networks relevant to the development of COPD, they were not augmented beyond the original, non-diseased network scaffolds, because no studies on the differences in signaling between non-diseased and diseased states were available." How do the authors know that no relevant studies were available? It seems that many papers at least have discussed links between COPD and DNA damage or Notch signaling (e.g. PMID: 19106307 published in 2009 "Down-regulation of the notch pathway in human airway epithelium in association with smoking and chronic obstructive pulmonary disease.")

"In total, 12 new nodes and 28 new edges were added to the Th1-Th2 Signaling network model during the jamboree discussions, thereby creating a comprehensive biological network of T-helper cell activity and their interactions with other immune cells in the context of COPD." How is 'comprehensive' measured? How do we know how much of the available literature was covered by the crowdsource process? That is, what is the sensitivity of the crowdsourcing process?

How many contributors were involved in enhancing each network in phase 2? Where were they from e.g. academia, industry? What incentivized them to contribute – for instance, were they COPD researchers? For the sake of research into crowdsourcing in biology, it would be very useful to provide additional analysis of the contributor community. We need to learn more about what works and what doesn't in crowdsourcing initiatives so future generations of these approaches can be improved.

The authors state "With nearly 900 new pieces of evidence added by the challenge crowd, a significant overall enhancement of the networks was achieved in a relatively short time (5 months)," How many papers (PMIDs) supported the 900 pieces of evidence?

Questions about use of BEL:
Figure 4. The shorthand BEL notation is not widely recognized as a visual format and difficult to read in general. An easy to read visualization format would make the network figures much easier to understand. Also, what do the different edge end symbols (e.g. arrow, dot, diamond) mean?

Part C of Figure 1 mentions "BEL to openBEL conversion". What's the difference between BEL and openBEL?

Other comments:
A broader review of the literature of pathway databases and crowdsourcing efforts should be included in the introduction.

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 12 May 2015

Stephanie Boue, Philip Morris Products S.A., Quai Jeanrenaud 5, 2000 Neuchâtel, Switzerland

High quality curation by trained database curators is needed in our community to convert the literature to computable models, but it is difficult to imagine how manual curation will scale to handle the ever-growing data generation rate in biology. Thus, the biological research community needs to figure out how to get crowdsourcing working for everyone as a tool to improve access to computable data. This paper does a very good job of describing how a set of COPD networks were constructed and enhanced (they grew in size and level of detail) through an interesting three-phase process. However, it would be useful to better describe the utility of the resulting networks and to further analyze the crowdsourcing process itself. Addressing these points will give the work a broader impact.

Authors' response: We have previously published several papers introducing use cases where the biological signal is interpreted in a meaningful manner using the causal network models [1-7] and have added these references to support the statement in the text. The point of the reviewer is absolutely relevant and we acknowledge that it will be of utmost importance to critically assess how the usefulness of the networks changed through each phase of the network verification project. As a first step, the previously published analyses can be repeated with the crowd-verified networks to assess the impact of network verification on data interpretation. A thorough assessment of the impact of crowd verification, requires however an extensive analysis leveraging multiple relevant datasets and to be reported thoroughly would dissolve the intended content of this manuscript that concentrates on the way networks were built and later on verified and refined through a crowdsourcing approach. We will conduct such an analysis and include the reference as soon as it will become available. We have now addressed these points in the discussion.

Utility of networks:

It is not clear what advantages the use of causal networks brings compared to more established models in the community, such as molecular interaction networks used by many algorithms (e.g. gene function prediction, module detection, interpretation of molecular profile data, network biomarkers) or detailed biochemical pathway models (used by most textbooks and pathway databases). While many results are published in terms of causal networks (e.g. A activates B), one important issue with networks constructed by collecting these relationships is that they may be difficult to integrate across resources since they are context specific: A may activate B in the lung, but inhibit B in the heart and when these are integrated, a conflict arises. Many computational analysis methods require integration of networks from multiple sources to construct the largest available network and integrate this data with disease-specific molecular profile data (e.g. gene expression data) to gain context (as it seems is done in the RCR approach). It would be useful for the authors to further discuss the utility of context-specific causal networks for follow on discovery.

Authors' response: The usage of causal networks allows all applications that other network models would have, and in addition eases the biological interpretation of the results in a

mechanistic, cause and effect fashion. The new, sophisticated algorithms that have been developed to analyze molecular data using the causal network models fully exploit the specific structure of two-layer cause-and-effect network models, providing evidence that causality adds precision on top of interaction[1,2,8]. However, as the reviewer points out, causality may differ across conditions (space and time), and the usage of BEL is therefore particularly relevant, as it allows for detailed context annotation of each piece of evidence linked to a causal edge. To fully make use of this property, it is important that as much of the literature evidence are collected in a knowledgebase, which will only really be doable thanks to new text mining methods assisting the biologists with the creation of BEL evidences or via crowdsourcing efforts such as the one described here. Because it is a very large undertaking, we have so far tried to restrict the evidences to respiratory and cardiovascular context. It is not excluded, however, that as the crowd and interest for the network grows, a more comprehensive annotation of the networks are achieved, making them usable in a specific context. Furthermore, BEL is being used in both academic and industry settings and BEL converters are being developed that can translate information from other sources such as BioPAX and SBML to facilitate comprehensive aggregation of networks.

I only noticed one sentence mentioning use: "By building the network model set in the BEL language format, we have generated a model framework suitable for biomarker discovery and for the interpretation of transcriptomic signatures found in human lung tissue." However, this sentence is not clear and doesn't cite any prior literature. How does using the BEL format create models suitable for biomarker discovery? Can't molecular interaction or other types of networks also be used for biomarker discovery? What type of biomarker discovery is referred to here? How are transcriptomic signatures interpreted and analyzed?

Authors' response: We have previously published several papers introducing use cases where the biological signal is interpreted in a meaningful manner using the causal network models[1-7] and have added these references to support the statement in the text. Martin *et al.* describes the development of network signatures that identify mechanisms that may explain differential drug treatment response between individuals, demonstrating that the causal two layered networks allow analyses which go beyond what normal networks can provide, i.e. provide classification power coupled with mechanistic detail[8].

Crowdsourcing comments:
"Networks that were not enhanced with COPD-specific mechanisms from the literature or RCR included the DNA Damage and Notch Signaling networks. Although both these networks relevant to the development of COPD, they were not augmented beyond the original, non-diseased network scaffolds, because no studies on the differences in signaling between non-diseased and diseased states were available." How do the authors know that no relevant studies were available? It seems that many papers at least have discussed links between COPD and DNA damage or Notch signaling (e.g. PMID: 19106307 published in 2009 "Down-regulation of the notch pathway in human airway epithelium in association with smoking and chronic obstructive pulmonary disease.")

Authors' response: We reformulated the sentence. Although there may be papers that report on the correlation between COPD and these processes like the Notch paper you

mention, we are referring to mechanistic papers that will provide causal links within the model. For example, a paper from a NOTCH1 knockout experiment in a COPD animal model that shows a particular protein being decreased will allow us to add the causal link of NOTCH1 activity increasing that protein in the Notch signaling COPD model. These are the types of causal mechanistic papers we have searched for and have not found in the context of COPD.

"In total, 12 new nodes and 28 new edges were added to the Th1-Th2 Signaling network model during the jamboree discussions, thereby creating a comprehensive biological network of T-helper cell activity and their interactions with other immune cells in the context of COPD." How is 'comprehensive' measured? How do we know how much of the available literature was covered by the crowdsource process? That is, what is the sensitivity of the crowdsourcing process?

Authors' response: As to avoid any confusion, and because the sensitivity of crowdsourcing is not easily measurable (as it would require to assess all possible literature), we reformulated to "more comprehensive".

How many contributors were involved in enhancing each network in phase 2? Where were they from e.g. academia, industry? What incentivized them to contribute – for instance, were they COPD researchers? For the sake of research into crowdsourcing in biology, it would be very useful to provide additional analysis of the contributor community. We need to learn more about what works and what doesn't in crowdsourcing initiatives so future generations of these approaches can be improved.

Authors' response: A specific publication addresses the statistics related to participation[9]. Clearly, the most difficult part of such a crowdsourcing project is to get the right incentives for people to participate. We acknowledge that showing the usefulness of the networks and their refinements should allow for a bigger buy-in from the scientific community, and likely more participation.

The authors state "With nearly 900 new pieces of evidence added by the challenge crowd, a significant overall enhancement of the networks was achieved in a relatively short time (5 months)," How many papers (PMIDs) supported the 900 pieces of evidence?

Authors' response: 479 unique PMIDs supported the 886 new pieces of evidence. We have included this detail in the text.

Questions about use of BEL:
Figure 4. The shorthand BEL notation is not widely recognized as a visual format and difficult to read in general. An easy to read visualization format would make the network figures much easier to understand. Also, what do the different edge end symbols (e.g. arrow, dot, diamond) mean?

Authors' response: We have added a legend to the figure. Please note that the bionet website also has a legend for the network visualization part.

Part C of Figure 1 mentions "BEL to openBEL conversion". What's the difference between BEL and

openBEL?

Authors' response: BEL was a proprietary language developed by Selventa. In the interest of the growing community of researchers using BEL, an openBEL language derived from BEL has been developed and released as open source http://www.openbel.org/. One of the main differences between the two is that in the openBEL, the namespace (i.e. databases in which the biological entity is defined) is clearly stated, allowing for a better standardization of used ontologies and databases. We have added this specification in the figure legend.

Other comments:
A broader review of the literature of pathway databases and crowdsourcing efforts should be included in the introduction.

Authors' response: We have discussed the comparison of our network models with other resources in other publications[2,9,10]. We have added this statement with appropriate references in the discussion for readers, who wish to find more background information about the network models and see how they compared with other approaches to interpret data.

References

1. Hoeng J, Deehan R, Pratt D, Martin F, et al.: A network-based approach to quantifying the impact of biologically active substances. *Drug discovery today*. 2012; **17** (9-10): 413-418 PubMed Abstract | Publisher Full Text
2. Hoeng J, Talikka M, Martin F, Ansari S, et al.: Toxicopanomics: Applications of Genomics, Transcriptomics, Proteomics, and Lipidomics in Predictive Mechanistic Toxicology. *In Hayes' Principles and Methods of Toxicology, Sixth Edition*. 2014; **Edited by Hayes AW, Kruger CL** (CRC Press): 295-332 Reference Source
3. Kogel U, Schlage WK, Martin F, Xiang Y, et al.: A 28-day rat inhalation study with an integrated molecular toxicology endpoint demonstrates reduced exposure effects for a prototypic modified risk tobacco product compared with conventional cigarettes. *Food and Chemical Toxicology*. 2014; **38**: 204-217 PubMed Abstract | Publisher Full Text
4. Phillips B, Veljkovic E, Peck MJ, Buettner A, et al.: A 7-month cigarette smoke inhalation study in C57BL/6 mice demonstrates reduced lung inflammation and emphysema following smoking cessation or aerosol exposure from a prototypic modified risk tobacco product. *Food and Chemical Toxicology*. 2015; **80**: 328-345 PubMed Abstract | Publisher Full Text
5. Schlage WK, Iskandar AR, Kostadinova R, Xiang Y, et al.: In vitro systems toxicology approach to investigate the effects of repeated cigarette smoke exposure on human buccal and gingival organotypic epithelial tissue cultures. *Toxicology mechanisms and methods*. 2014; **24** (7): 470-487 PubMed Abstract | Free Full Text | Publisher Full Text
6. Talikka M, Kostadinova R, Xiang Y, Mathis C, et al.: The Response of Human Nasal and Bronchial Organotypic Tissue Cultures to Repeated Whole Cigarette Smoke Exposure. *International Journal of Toxicology*. 2014; **33** (6): 506-517 PubMed Abstract | Publisher Full Text
7. Thomson TM, Sewer A, Martin F, Belcastro V, et al.: Quantitative assessment of biological impact using transcriptomic data and mechanistic network models. *Toxicology and Applied*

Pharmacology. 2013; **272** (3): 863-878 PubMed Abstract | Publisher Full Text
8. Martin F, Sewer A, Talikka M, Xiang Y, et al.: : Quantification of biological network perturbations for mechanistic insight and diagnostics using two-layer causal models. *BMC Bioinformatics*. 2014; **15**: 238 PubMed Abstract | Free Full Text | Publisher Full Text
9. sbv IPT, Binder J, Boue S, Di Fabio A, et al.: Reputation-based collaborative network biology. *Pacific Symposium on Biocomputing*. 2015. 270-281 PubMed Abstract | Publisher Full Text
10. Boue S, Talikka M, Westra JW, Hayes W, et al.: Causal Biological Network (CBN) database: a comprehensive platform of causal biological network models focused on the pulmonary and vascular systems. *Database*. 2015; **2015** (bav030): 1-14 PubMed Abstract | Publisher Full Text

Competing Interests: No competing interests were disclosed.

Reviewer Report 06 February 2015

https://doi.org/10.5256/f1000research.6402.r7528

© 2015 Hide W. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Winston A. Hide
Department of Biostatistics, Harvard School of Public Health, Boston, MA, USA

This work uses a hybrid approach network modelling approach to incorporate predictive methodology with empirical knowledge and crowd sourcing for models of COPD pathogenesis. It is a good idea, thoroughly implemented and has produced a potentially useful set of pathways. The value of the resulting pathways is not clear as they do not have community validation, only community design. The manuscript is exhaustive in its descriptions and the process of developing the models is clear.

The work represents the first phase of understanding for knowledge driven development of network models of COPD - the process of building the models is well described and the actual outcomes of the interactions with community are informative. The question of the actual true value of the models in terms of their accuracy, adoption and accessibility is not yet convincingly addressed. That may be expected as the purpose of this work appears to be a description of the first part of the process of developing knowledge based models for a disease. The models as presented appear unvalidated and without a description of the framework for assessing the value and actioning of the networks, it is not clear how their uptake by the community will be assured.

This is a unique effort but the manuscript should make more reference to existing pathway based community annotation efforts e.g.: wikipathways and/or open science

initiatives such as those promoted by community interaction leaders such as Andrew Su. It should show how the value of this approach differs to existing efforts.

In terms of access to expertise, it is not clear how an uninvited scientist would contribute to an existing pathway model - except through the open but time-limited crowdsourcing venue. Straightforward validation of the models network is not tested in terms of their consistency or cross-valdiation within COPD high dimensional assays - where it should be possible to see evidence of enrichment for co-expression etc.

Contextual nature of networks is mentioned and attempts are made to address contextual pathway structures, but the context is not tested.

As a suggestion the authors should consider community validation

Pathway accessibility and distribution is described but it is not clear as to how these models are available in any format except web browsing. For the models to be tested by the community, value would come from making them openly available as downloadable instances in several of the most popular formats. Feedback on their accuracy could then be encouraged.

Competing Interests: I was an invitee to one of the Improver conferences to present a talk but did not contribute to the manuscript or models. I know S. Boue as a previous collaborator and colleague on a publication "ASTD: The Alternative Splicing and Transcript Diversity database.," vol. 93, no. 3, pp. 213–220, Mar. 2009."

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Author Response 12 May 2015

Stephanie Boue, Philip Morris Products S.A., Quai Jeanrenaud 5, 2000 Neuchâtel, Switzerland

This work uses a hybrid approach network modelling approach to incorporate predictive methodology with empirical knowledge and crowd sourcing for models of COPD pathogenesis. It is a good idea, thoroughly implemented and has produced a potentially useful set of pathways. The value of the resulting pathways is not clear as they do not have community validation, only community design. The manuscript is exhaustive in its descriptions and the process of developing the models is clear.

The work represents the first phase of understanding for knowledge driven development of network models of COPD - the process of building the models is well described and the actual outcomes of the interactions with community are informative. The question of the actual true value of the models in terms of their accuracy, adoption and accessibility is not yet convincingly addressed. That may be expected as the purpose of this work appears to be a description of the first part of the process of developing knowledge based models for a disease. The models as presented appear unvalidated and without a description of the framework for assessing the value and actioning of the networks, it is not clear how their uptake by the

community will be assured.

Authors' response: The point of the reviewer is absolutely relevant and we acknowledge that it will be of utmost importance to critically assess how the usefulness of the networks changed through each phase of the project. Whenever possible, orthogonal data sets were used to validate the network model during the building process. In the paper *Systematic verification of upstream regulators of a computable cellular proliferation network model on non-diseased lung cells using a dedicated dataset*, we have done just that by using a specifically designed, independent lung cell proliferation dataset to verify the correctness of the cell cycle network model [1]. The validation of all available networks requires an extensive analysis leveraging multiple relevant datasets and to be reported thoroughly would dissolve the intended content of this manuscript that concentrates on the way networks were built and later on verified and refined through a crowdsourcing approach. We will conduct such an analysis and make sure to reference it here as soon as it will be available.

This is a unique effort but the manuscript should make more reference to existing pathway based community annotation efforts e.g.: wikipathways and/or open science initiatives such as those promoted by community interaction leaders such as Andrew Su. It should show how the value of this approach differs to existing efforts.

Authors' response: We have discussed the comparison of our network models with other resources in other articles[2,3] and in a book chapter[4]. We have added this statement in the discussion for readers who wish to find more background information about the network models and see how they compared with other approaches to interpret data.

In terms of access to expertise, it is not clear how an uninvited scientist would contribute to an existing pathway model - except through the open but time-limited crowdsourcing venue.

Straightforward validation of the models network is not tested in terms of their consistency or cross-valdiation within COPD high dimensional assays - where it should be possible to see evidence of enrichment for co-expression etc.

Contextual nature of networks is mentioned and attempts are made to address contextual pathway structures, but the context is not tested.

As a suggestion the authors should consider community validation

Pathway accessibility and distribution is described but it is not clear as to how these models are available in any format except web browsing. For the models to be tested by the community, value would come from making them openly available as downloadable instances in several of the most popular formats. Feedback on their accuracy could then be encouraged.

Authors' response: The networks can be browsed on the bionet.sbvimprover.com website, including latest votes and modification. More stable versions are stored in the causalbionet.com database[2].

References

1. Belcastro V, Poussin C, Gebel S, Mathis C, et al.: Systematic verification of upstream regulators of a computable cellular proliferation network model on non-diseased lung cells using a dedicated dataset. Bioinformatics and biology insights. *Bioinformatics and biology insights*. 2013; **7**: 217-230 PubMed Abstract | Free Full Text | Publisher Full Text
2. Boue S, Talikka M, Westra JW, Hayes W, et al.: Causal Biological Network (CBN) database: a comprehensive platform of causal biological network models focused on the pulmonary and vascular systems. *Database*. 2015; **2015** (bav030): 1-14 PubMed Abstract | Publisher Full Text
3. sbv IPT, Binder J, Boue S, Di Fabio A, et al.: Reputation-based collaborative network biology. *Pacific Symposium on Biocomputing*. 2015. 270-281 PubMed Abstract | Publisher Full Text
4. Hoeng J, Talikka M, Martin F, Ansari S, et al.: Toxicopanomics: Applications of Genomics, Transcriptomics, Proteomics, and Lipidomics in Predictive Mechanistic Toxicology. *In Hayes' Principles and Methods of Toxicology, Sixth Edition*. 2014; **Edited by Hayes AW, Kruger CL** (CRC Press): 295-332 Reference Source

Competing Interests: No competing interests were disclosed.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research