REPUTATION-BASED COLLABORATIVE NETWORK BIOLOGY*

The sbv IMPROVER project team (in alphabetical order): JEAN BINDER[1], STEPHANIE BOUE[1], ANSELMO DI FABIO[2], R. BRETT FIELDS[3], WILLIAM HAYES[3], JULIA HOENG[1†], JENNIFER S. PARK[3], MANUEL C. PEITSCH[1]

[1] *Philip Morris International R&D, Philip Morris Products S.A., Quai Jeanrenaud 5, 2000 Neuchâtel, Switzerland*
[2] *Applied Dynamic Solutions, LLC. 220 Davidson Avenue, Suite 100, Somerset, NJ, 08873, USA*
[3] *Selventa, One Alewife Center, Cambridge, MA 02140, USA*
[†] *Corresponding author: julia.hoeng@pmi.com*

A pilot reputation-based collaborative network biology platform, Bionet, was developed for use in the sbv IMPROVER Network Verification Challenge to verify and enhance previously developed networks describing key aspects of lung biology. Bionet was successful in capturing a more comprehensive view of the biology associated with each network using the collective intelligence and knowledge of the crowd. One key learning point from the pilot was that using a standardized biological knowledge representation language such as BEL is critical to the success of a collaborative network biology platform. Overall, Bionet demonstrated that this approach to collaborative network biology is highly viable. Improving this platform for *de novo* creation of biological networks and network curation with the suggested enhancements for scalability will serve both academic and industry systems biology communities.

1. Introduction

Biological networks represent our knowledge about biological mechanisms as diagrams of nodes (e.g. molecular entities) and edges (relationships between entities). Network biology concerns itself with the building and maintenance of such networks. This requires a great deal of contextual knowledge that is generally beyond the scope of individual biologists. This makes network biology an excellent field for collaborative efforts; such efforts include WikiPathways (http://wikipathways.org), BioPax (http://biopax.org), and OpenBEL (http://openbel.org). Here we propose another model that may be more effective for collaborating in this field, namely, a reputation-based collaborative network biology platform designed to build, edit and verify networks. It can allow more scientists (e.g. subject matter experts) to bring their perspectives to bear on large representations of mechanisms. A reputation-based system can also provide self-moderation, making this a more scalable approach than an assigned moderator-managed collaborative platform. It can also incorporate peer review functions. Here, we present a prototype solution called Bionet and share the usage results from the initial Network Verification Challenge (http://bionet.sbvimprover.com) using this platform.

* This work is supported by Philip Morris International.

1.1 Network Verification Challenge (NVC)

The NVC, the third challenge of the sbv IMPROVER project[1], is an effort to validate industrial research approaches and resulting biological networks focused on lung biology and lung diseases such as chronic obstructive pulmonary disease (COPD). The NVC was supported by the creation of a platform for collaborative network biology, called Bionet (http://bionet.sbvimprover.com), to help verify and enhance the COPD biological networks. The pilot phase of the NVC consisted of a 5-month open phase during which participants could log into the website and contribute by voting on evidence and edges of fifty biological networks. The open phase was followed by a 3-day in-person Jamboree meeting where the best performers and subject matter experts in the field of lung and COPD biology were invited to discuss and agree on changes to the networks.

1.2 Large-scale Collaboration

Some crowdsourcing efforts, such as the Critical Assessment of Protein Structure Prediction initiative (CASP) [2], require intense effort and a high level of expertise while others, such as Foldit [3], Mechanical Turk (https://www.mturk.com) or Wikipedia (http://wikipedia.com) require less effort and expertise. The combined level of expertise and effort is generally inversely proportional to the number of people participating in a crowdsourcing effort. Consequently, less intense crowdsourcing efforts which offer relatively strong incentives attracted many more participants. For example, Foldit, a crowd-sourced protein folding game, attracted over 531,000 participants (http://foldit.com) while Assemblathon2, a very high effort, high expertise crowdsourcing effort recruited "only" 21 teams [4].

An important aspect of any crowdsourcing effort is to define the appropriate incentives. In the NVC, access to the resulting networks and network biology (i.e., the possibility to download the networks for further analyses/visualization) was a significant motivator along with a reputation system and associated benefits. Another benefit was an invitation based on earned reputation points to attend the Jamboree review of the resulting networks after the initial phase of network verification and enhancement.

1.3. Reputation System

A reputation system can be used to support self-moderation of a crowd-sourced curation system. Examples of initiatives using reputation systems are ResearchGate (https://www.researchgate.net) and StackOverflow (http://stackoverflow.com). The ResearchGate reputation score (RGScore) is used only to provide a ranking compared with other researchers, while the StackOverflow reputation score is used in crowd management of the StackOverflow question and answer crowdsourcing site. As members develop their StackOverflow reputations, more powerful moderation features are unlocked; i.e. as members 'prove' themselves and become trusted community members, they are given rights and responsibilities of managing the StackOverflow site and participant community. A side effect of the StackOverflow reputation score and associated badges of activity, such as 'Great Question' or 'Guru', is that these badges are now used as expertise credentials in the software developer community.

2. Materials and Methods

2.1. Biological Networks

In the NVC, fifty networks were made available on the Bionet website for crowd verification. These networks were based on previously constructed non-diseased networks that describe cell proliferation [5], cell stress [6], DNA damage, autophagy, cell death and senescence [7], pulmonary inflammation [8], and tissue repair and angiogenesis [9]. The networks used in the NVC were enhanced with COPD-relevant mechanisms using a literature and data approach (manuscript in preparation).

2.2. Collaborative Web Platform Functionality

Bionet gives the participants the ability to search for and navigate to a network based on various conditions: name of the network, official name and synonyms of the nodes and edges, and references supporting the edges (PubMed IDs). In the network viewer, the participant can navigate the network by nodes or edges and can use the node and edge lists to quickly view a sorted and filtered list of all elements in the network. By selecting an edge from the edge list, the participant can view a list of published evidence related to the edge, then, by selecting an evidence item, the participant can view the complete details of the evidence. This approach to network navigation can also be applied to the network visualization tool. When a participant selects a node or an edge on the network visualization page, the associated information is presented allowing participants to drill down into the evidence-level detail.

Edge creation and Evidence voting/creation are tied directly to the Reputation system. Participants gain reputation points by verifying and enhancing the networks in various ways: extend networks with new edges, provide additional evidence for edges, and/or approve/reject evidence that has been posted in support of network edges. Participants also gain points if the evidence or edge they added has been approved by other users. To vote on evidence, a participant selects the evidence and is then presented with the option of approving or rejecting the evidence. Based on the type of evidence selected, the participant is asked questions to document the rationale for the approval or rejection. After evidence is submitted, it becomes active, which allows other participants to vote on the newly added evidence. If the evidence is voted on by enough participants, and a majority of participants approve or reject it, the evidence is locked and the edge is marked approved or rejected. If no consensus is reached, the evidence is marked as ambiguous. The network visualization tool reflects the status of the edges using different colors. Moreover, users are able to visualize their own changes in the network viewer.

Participants can create a new edge by selecting a node from the network using the Biological Expression Language (BEL). BEL is a syntax that can represent biological relationships in a standardized computable format. The Bionet application provides a BEL syntax generator to help the participant create a proper BEL statement for the edge. Additionally, a tool to help create one or several BEL statements based on an excerpt from an academic paper (evidence source) is provided in Bionet. When an edge is created, the participant adds evidence to the edge and submits it to the network. After submission, all participants can see it on the network visualization page and can vote on it like any other edge in the application.

Bionet provides a community section that allows participants to see the latest network activity for all users or him/herself and possibly filtered by network. This community area is critical to participation because participants can use it to view the network of interest to them and track the actions taken by other participants. Participants can then vote on the action directly in the activity feed or go to the network to see the action in context.

Leaderboards are used on the site to help participants gauge their level of participation and reputation score in relation to their peers. Points are given for voting and evidence creation and badges are awarded for various actions. Participants can filter the leaderboards by teams, votes, evidence creation, and edge creation.

2.3. Funnel of Participation

The funnel of participation, as coined by Clow, refers to the process of gaining participants for a participatory project [10], which was applicable for this pilot phase of NVC (Figure 1). Awareness was achieved through emails to potential participants based on their research record, presentations at relevant scientific conferences, publications (both peer-reviewed and science news media channels) [11] as well as seminars at selected network biology-focused laboratories. In addition, we engaged 'NVC ambassadors' who personally called and/or emailed contacts in their scientific network to help increase awareness by notifying and teaching potential participants about the NVC. We estimated that the Bionet Awareness campaign resulted in 1,000,000 impressions on potential participants. From this, there were 1,298 unique visitors to the collaborative website. This resulted in 132 Bionet registered participants from which 26 highly active participants were selected as Best Performers for the Challenge (Figure 1).

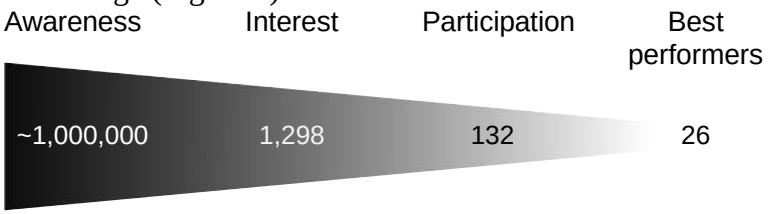

Fig. 1. Funnel of participation for the first NVC.

2.4. Evaluation of Participant Activity

User statistics from Bionet logs were analyzed to calculate a number of metrics related to user participation and network activity. A questionnaire was emailed to participants after the challenge to help understand such factors as motivation and ease of website use.

References associated with networks were counted by searching published scientific literature using Quertle (www.quertle.com) and the name of the network with the word "pathway". If the network name contained the word "signaling" it was replaced with "pathway" because this is a more specific way of describing molecular events (signaling can refer to electric signals).

2.5. Evaluation of Tissue-relevant Evidence Additions

Crowd-submitted evidence was reviewed to assess the overall degree of tissue relevancy compared with network boundaries. Evidence from primary lung tissue and lung-associated cell types (during

COPD) was deemed within the network boundaries, while additions from non-lung-relevant cell types (e.g. neural progenitor cells) and non COPD-relevant diseases (e.g. colon cancer) were rejected. The number of COPD-relevant evidence additions as a percentage of the total number submitted was taken as an overall assessment of crowd performance.

2.6. Evaluation of Quality of Participant-submitted Causal Biology

A random sampling of 100 pieces of evidence submitted by the participants was independently evaluated by two scientists with expert-level experience in these networks to assess the overall quality of contributions by manually reviewing the primary literature associated with each submission. A random number generator was used to produce 100 numbers within a 1–885 range, corresponding to the total pieces of evidence submitted among all networks. Entries were further blinded by removing all personal participant information stored with the entry (e.g. user name of submitting participant) to prevent bias during the expert evaluation. Key metrics that were evaluated during this process included, 1) relevance of evidence within the individual network, 2) relevance to COPD and/or lung biology, and 3) accuracy of capturing the biological relationship in referenced literature. Evidence meeting all three criteria was rated "Valid" and that deficient in any of the three was rated "Invalid." For the evaluation, evidence containing minor defects in BEL scripting was not rated "Invalid".

3. Results

3.1. Evaluation of Participant Activity

A global community of researchers took part in the NVC (Figure 2).

Fig. 2. NVC participant countries.

An analysis of the activity of participants in the NVC Open Phase revealed a range of participant profiles. Some researchers who registered performed only a few actions. In a follow-up questionnaire, the explanations given for low participation were mostly the lack of time and/or

specific interest in the networks. As expected, the "entry level" action of participants with low activity levels was "voting on evidence" (Figure 3, right panel). Distinctive profiles were observed among the best performers (BP) (Figure 3): some spent a lot of effort enhancing specific networks (e.g. BP2), some contributed to many networks by creating new evidence/edges (e.g. BP9), and some dedicated most time verifying (voting on) a number of networks (e.g. BP3). In general, BPs voted on more networks than the number of networks for which they created new evidence (e.g. BP2, BP3).

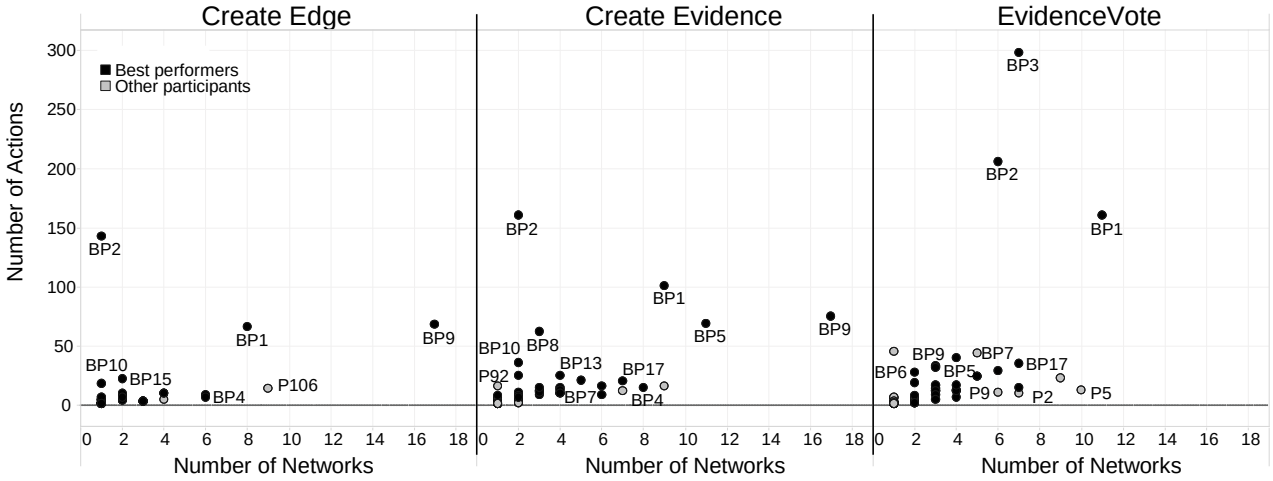

Fig. 3. Participant activity across a number of networks.

Dissecting the activity of participants per network and action type revealed interesting patterns (Figure 4): (i) cell-specific networks and widely studied biological processes (>10,000 associated references) attracted the highest number of participants, especially among those that did not have a lot of activity (non-best performers); (ii) more complex networks were approached more frequently by participants who were more experienced or who spent more time on the NVC (best performers); (iii) in most cases, the number of distinct contributors was highest for the voting actions; and (iv) of the 50 networks, nine attracted at least ten contributors.

3.2. Evaluation of species and tissue-relevant evidence additions

One of the goals of the NVC was to add relevant literature that supported edges in the networks to improve their overall relevance to human COPD. We evaluated the extent of human literature supporting the edges, as well as evidence from lung-relevant experiments added by the participants. In total, the crowd submitted 885 new pieces of evidence, the large majority of which was from human studies (65%) (Figure 5A).

There was great variability on a per-network basis in terms of quality of submissions as well as general activity, with cell-specific network additions conforming more frequently to the tissue boundary conditions. For example, the *Neutrophil Signaling* network received a preponderance of participant submissions with 179 total pieces of evidence submitted (20% of the total submitted). Among the 170 pieces of evidence with tissue metadata, 100% submissions conformed to the network boundary conditions. Similarly, in the *B-cell Signaling* network, 100% of the annotated

submissions fell within the boundary conditions of the network collection (Figure 5B). In contrast, the *Notch* network received the fewest tissue-relevant evidence additions, with only 17% of the annotated submissions falling within the boundary conditions of the network (Figure 5B).

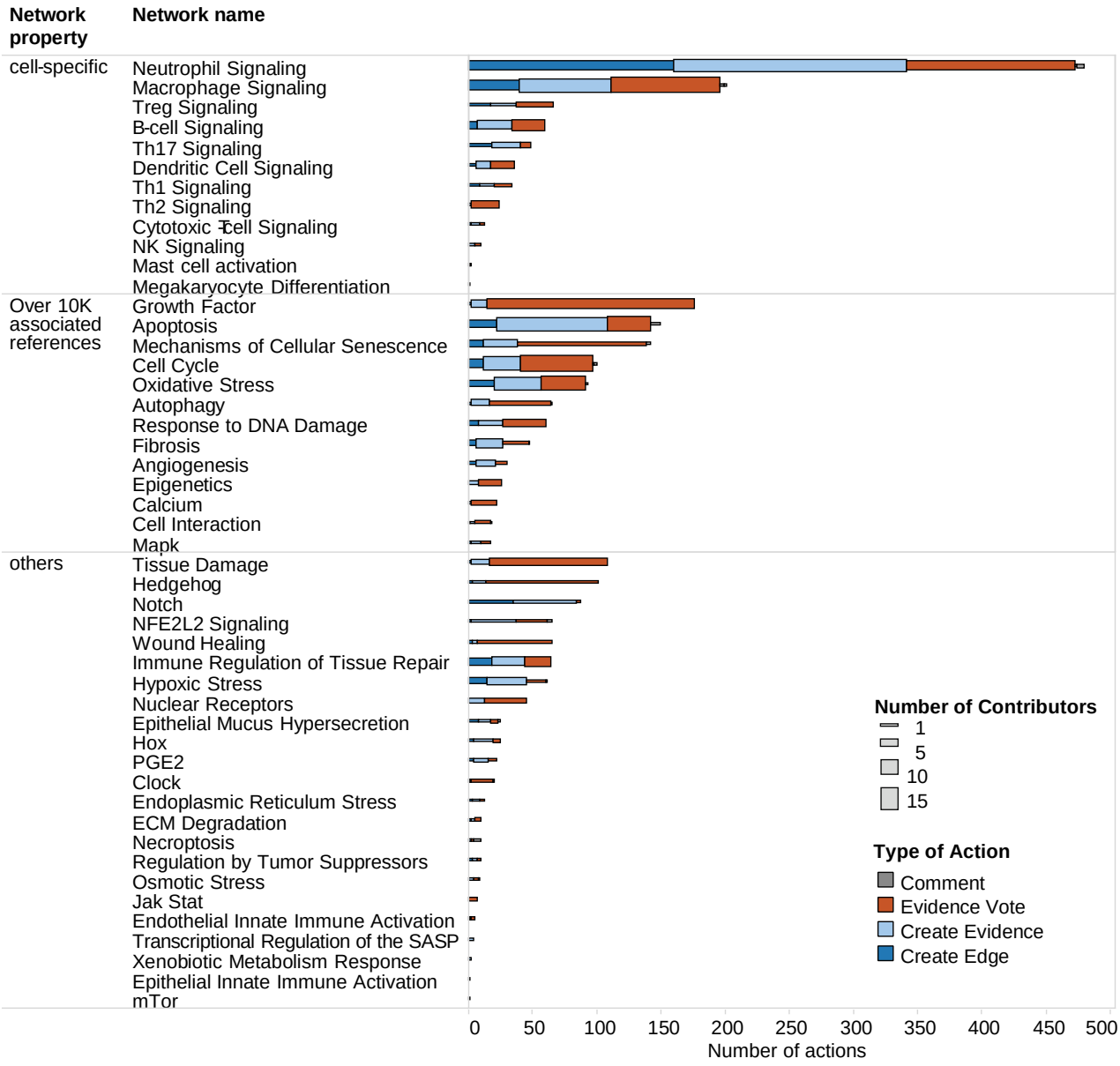

Fig. 4. Participant activity by network.

3.3. Evaluation of quality of participant-submitted causal biology

We assessed the overall quality of the 885 total pieces of evidence submitted by the NVC community using a randomized, independent review process. On average, the quality analysis resulted in a validity rate of 77%, indicating that the majority of additions enhanced the biological

foundation supporting the network connectivity. The majority of evidence deemed invalid was outside the tissue boundaries of the network. However, because the tissue boundaries may not have been obvious to all participants, to evaluate valid biological representation, we calculated the percentage validity again after removing boundaries as a criterion. The average validity rate after disregarding tissue boundaries increased to 88%. The review by one of the scientists yielded 88% valid contributions while the second independent review by another scientist yielded 85% valid contributions, although the same pieces of evidence were not always judged as valid. An assessment of inter-reviewer comparability revealed 83% agreement between the two evaluations. Among the entries where there was disagreement, most cases were caused by subjective interpretation of the primary literature and, by extension, its representation in BEL. Employing a more robust statistical measure of inter-reviewer comparability, Cohen's kappa, produced a coefficient of 0.26, illustrating the subjective nature of assessing the quality of biological submissions, and the importance for several scientists to review the same evidences, as made possible in this initiative.

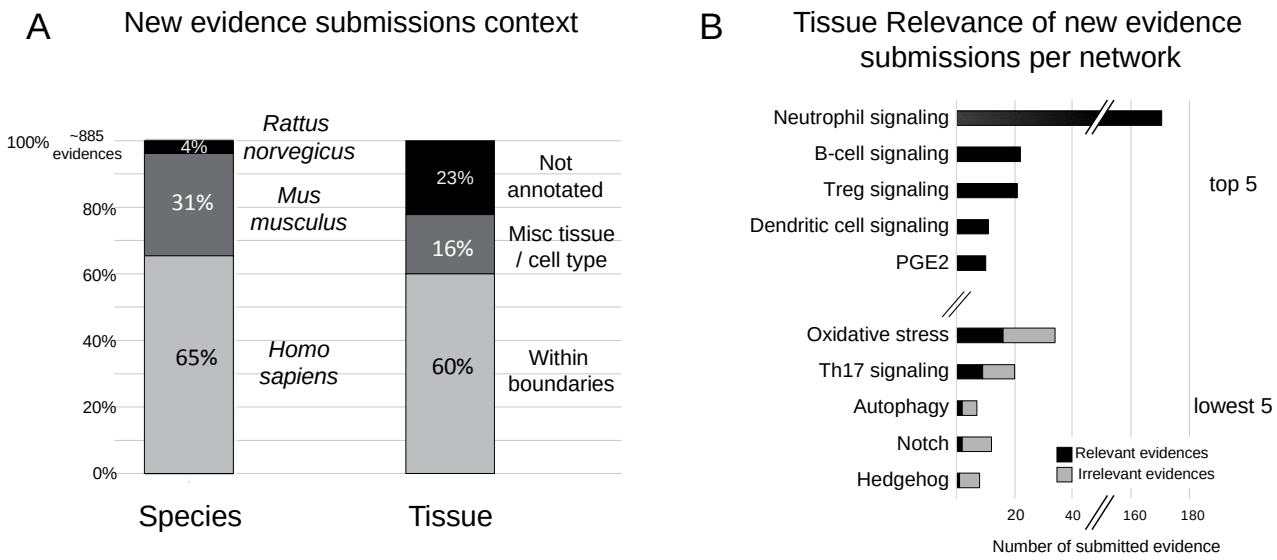

Fig. 5. Relevance of new evidence submissions overall for context (A) and for the top 5 and lowest 5 networks (B).

4. Discussion

4.1. NVC Contributions and Participant Activity

4.1.1 Overall Number of Contributions

In total, 2,456 votes were cast and 885 pieces of evidence were created (including 351 new network nodes) by a relatively small number of participants (~80). Although these numbers are small when compared with the total amount of evidence in the networks (over 180,000 pieces), the actions that were measured took place over a 5-month period in a pilot project while the existence of the NVC was still being disseminated to the scientific community. To verify 50 networks containing thousands of biological connections was overwhelming for the modest number of participants. For

future projects, one approach may be to restrict the number of networks being evaluated to help concentrate participant attention to particular areas. In addition, pilot crowdsourcing ventures with limited adoption can focus more on the creation versus the verification process for enhancing networks. Creation of new evidence and edges could add useful biology even with a small number of participants, while to reach a crowd consensus, verification/voting requires a larger number of participants to ensure a representative sample.

4.1.2. Focused Participant Activity

Overall, the participants in the NVC worked on a small number of networks (1–5) according to their scientific expertise. Participants tended to work on well-studied networks with canonical pathways (*Cell Cycle* and *Apoptosis*) that were reported in the literature, or on cell-specific networks (*Macrophage* and *Neutrophil Signaling*) for which it was straightforward to identify the relevant literature. Networks of high interest and for which there is a lot of information may be more conducive to a crowd-verification approach. However, it is the less-studied networks that might benefit more from crowd review. In the future, the NVC could be restricted to these networks to concentrate attention and effort.

Overall, participants voted rather than created new evidence, with the best performers voting on more networks and creating evidence for fewer networks. This is likely because it is easier to vote (i.e., assess the scientific validity of existing evidence) on diverse topics, whereas it requires significantly more expertise to enhance specific networks by adding new scientific evidence to existing network edges (i.e., identify and extract additional knowledge from the scientific literature) or to create new edges. Indeed, in the participant survey, most researchers scored voting to be "very easy" and rated adding new evidence to be "easy" (data not shown).

4.1.3. Participant Engagement

The NVC was publicized through many different avenues, including conferences, publications, emails, web searches, seminars, advertisements, and "ambassadors". NVC ambassadors tapped into personal scientific networks to promote the NVC and follow up with one-on-one educational sessions for interested scientists. This mode of promotion was found to be the most effective because of the personal nature of the contact and the opportunity for a tutorial session to ensure that potential participants understand how to use the website and create BEL statements. In fact, a participant survey showed that the majority learned about the NVC through personal contact (data not shown). Because of the success of this method, we are continuing to emphasize this personal ambassador approach among personal scientific networks to further publicize and educate researchers about the next NVC.

NVC participants were motivated by various factors, with the possibility of co-authoring an academic publication being the top motivation according to the participant survey (data not shown). Other motivations included travel and Jamboree invitation rewards, learning about the biology of the networks, the chance to download the networks to use in their research, and the challenge of the verification tasks. For continued participation these types of incentives are important to attract a community of regular participants based on periodic cycles of publications, meetings and latest version networks being released.

4.2. NVC Improved the Relevance and Comprehensiveness of the Networks

4.2.1. Evaluation of Participant Contributions

The new evidence created by participants during this short period represents an improvement in the network comprehensiveness, especially when considering the majority of evidence was concentrated among a small number of networks. The evaluation of tissue relevance for new evidence submissions was one metric used to quantify the value of crowdsourcing for improving biological networks. In this assessment, 23% of submitted evidence contained no contextual annotation, despite providing several entry fields (e.g. tissue, disease, cell type). One way to reduce non-annotated contributions in future projects could be to implement mandatory contextual fields during the submission process. Ultimately, we determined that 60% of evidence submissions conformed to the network boundary conditions as set forth in the platform user tutorials. Although entries from miscellaneous contexts are certainly a significant overall contribution, such entries could be better avoided by publishing the network boundaries more frequently throughout the platform, perhaps once again at point-of-entry for new entries to ensure participants are fully aware of the criteria prior to submission. The boundaries were heavily emphasized during scheduled webinars to promote the NVC and educate users but were less visible and detailed on the website where users probably most needed this reminder.

When the new submission context was assessed on a per-network basis, several networks received an outstanding quality of new evidence because 100% of the submissions conformed to the specified boundary conditions (Figure 5B). For example, for the *Neutrophil Signaling* network, all annotated submissions were sourced from primary literature in which the study was conducted specifically in neutrophils. In contrast, networks detailing more ubiquitous biological pathways (e.g. *Oxidative Stress* and *Notch Signaling*) often received submissions from a broader array of contexts, reflecting the abundance of primary literature among many cellular contexts, but not necessarily relevant to lung. Therefore, we concluded that networks with cell-specificity garnered submissions with better-defined contexts that were more likely to conform to the stated boundary conditions. It may be more reasonable in future challenges to loosen the boundary conditions for the more general biological pathways that are conserved across tissues.

A separate dimension related to the overall quality of crowd-submitted enhancements was assessed by an expert-level review of a random sampling of the submissions. Two independent evaluations of the same data sample revealed a "Validity" rate average of 77%. Because the tissue and disease boundary conditions of the network may not have been apparent to participants, when the validity rate was calculated without regard to these boundary conditions it came out to be 87%. This high validity rate suggested that the participants successfully entered biologically sound mechanistic statements retrieved from the literature into the website to contribute to the networks. Improving communication of the boundaries is an important lesson from this analysis.

Despite the high number of submissions deemed to be of "high quality" by the experts, the calculated Cohen's kappa statistic of 0.26 revealed a modest degree of comparability between the overall quality assignments, or 83% agreement. This statistic factors in the probability of agreement occurring by chance during a qualitative evaluation, which is heavily influenced by the fact a simple binary rating system (Valid vs. Invalid) was used during the evaluation process. Nevertheless, this

evaluation process illustrates the subjectivity inherent in assessing biological experiments. Some of the differences in the reviewer analyses were related to them not having complete information (unavailable full text) or simply reviewer error (mis-reading the paper). However, in some cases the biology was interpreted in a different way by the scientists. Owing to subjectivity of biological interpretation, in addition to the open phase during the NVC an in-person Jamboree was held to provide a forum where these subjective or controversial items could be discussed. Scientists who contributed to the networks as well as subject matter experts for the biology that each network describes participated in the Jamboree to come to a consensus for finalizing the changes to the networks. Not only was the Jamboree critical for alignment of individual controversial edges within the networks, but it provided a forum to discuss the networks more holistically and edit larger pieces of the network to improve flow, comprehensiveness, and granularity. Overall, the NVC enhancements improved the relevance and comprehensiveness of the networks.

Refined networks have been uploaded to Bionet so that the improvements brought by the crowd can benefit the scientific community. In particular, active participants that have earned the "Download" badge by performing a minimum number of actions may download all networks for further analysis and visualization using their favorite tools.

4.2. Vision for Biological Collaboration

4.2.1. BEL as a Universal Language for Biology

The Bionet platform is designed to require low effort but high subject matter expertise. However, the time that a potential participant needs to invest in learning BEL was a significant challenge that increased the activation effort for a participant and contributed to participant attrition. There is no standard knowledge representation that every network biologist is trained on that is comparable to the chemical reaction language for chemists. Because BEL is new to the biological academic community it has not yet achieved widespread adoption. If BEL, or some other universal network biology knowledge representation, becomes a standard representation for biological relationships, collaboration in biology, especially network biology, can become more effective. A standard representation for biology will greatly increase the viability of a network biology collaboration platform.

4.2.2. A Standard Reputation Platform for Biology

As a long-term, self-moderating platform for the creation and management of biological networks, we propose that the Bionet platform can provide great benefits to network biology. An added benefit is the potential for the reputation points and badges earned in the system based on peer-review to become an important aspect of a network biologist's *curriculum vitae*. A Reputation Score and associated Badges could become an important credentialing resource for network biologists in the research community, as has StackOverflow reputation score gained importance in the informatics community.

Our goal in creating a network biology-focused collaborative platform is, in the longer term, to provide a reputation system that supports self-management in the same manner as StackOverflow.

In the first iteration of Bionet, the reputation score only ranked one participant against the other participants and is similar to the RGScore. However, the Bionet reputation score was designed to be extended to promote high-reputation participants to moderators of the networks in the same manner as the StackOverflow reputation score. As participants gain reputation points and hit certain targets, they will be able to take on more moderation of the networks on the Bionet platform. This will allow the platform to scale to many thousands of participants and thousands of networks.

The Bionet platform can work well in both public and industry settings. The Bionet reputation-based collaborative network biology platform works well with a small number of active users. The recommended changes that will make it more scalable by automatically promoting high-reputation users to moderators are not required in an industry setting. A single platform that provides online access to networks, allows participants to edit the networks and collaborate easily with other participants regarding the networks, and provides computationally tractable networks using a common knowledge representation will be a great tool for network biology in both the academic and industry sectors.

This pilot project in large-scale collaboration for network biology has highlighted certain aspects that are required for a self-sustaining platform, including a universal biological language and a standardized and therefore valued reputation system. With the insights gained during the NVC, both the Bionet website and biological content will continue to improve and latest version networks are available on Bionet as well as on our Causal Biological Networks database (CBN, causalbionet.com). Bionet is currently open for crowd input during the ongoing second Challenge (NVC2) and associated Jamboree planned for 2015.

5. Acknowledgments

We thank IBM for their help in organizing the NVC and Jamboree, Michael Maria for help in project management, and Sam Ansari, Anita Iskandar, Robin Kleiman, Carine Poussin, Dexter Pratt, Marja Talikka, and Walter Schlage for their scientific input.

References

1. Meyer, P., *et al. Nat Biotechnol* **29**, 811-815 (2011).
2. Moult, J., Fidelis, K., Kryshtafovych, A., Schwede, T. & Tramontano, A. *Proteins: Structure, Function, and Bioinformatics* **82**, 1-6 (2014).
3. Cooper, S., *et al. Nature* **466**, 756-760 (2010).
4. Bradnam, K.R., *et al. GigaScience* **2**, 10 (2013).
5. Westra, J.W., *et al. BMC systems biology* **5**, 105 (2011).
6. Schlage, W.K., *et al. BMC systems biology* **5**, 168 (2011).
7. Gebel, S., *et al. Bioinformatics and biology insights* **7**, 97-117 (2013).
8. Westra, J.W., *et al. Bioinformatics and biology insights* **7**, 167-192 (2013).
9. Park J.S., S.W.K., Frushour B.P., Talikka M., Toedter G. et al., Vol. S12: 002 (2013).
10. Clow, D., Vol. 185-189 (ACM, Leuven, Belgium, 2013).
11. The sbv Improver project team, *et al. Bioinformatics and biology insights* **7**, 307-325 (2013).