**Connecting proteins: shareable tools for reproducible interaction mapping**

<table>
<thead>
<tr>
<th><strong>Journal:</strong></th>
<th><em>Biochemistry and Cell Biology</em></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Manuscript ID:</strong></td>
<td>bcb-2019-0285.R1</td>
</tr>
<tr>
<td><strong>Manuscript Type:</strong></td>
<td>Invited Symposia</td>
</tr>
<tr>
<td><strong>Date Submitted by the Author:</strong></td>
<td>26-Sep-2019</td>
</tr>
<tr>
<td><strong>Complete List of Authors:</strong></td>
<td>Gingras, Anne-Claude; Samuel Lunenfeld Research Institute,</td>
</tr>
<tr>
<td><strong>Keyword:</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Is the invited manuscript for consideration in a Special Issue?</strong></td>
<td>Not applicable (regular submission)</td>
</tr>
</tbody>
</table>
Connecting proteins: shareable tools for reproducible interaction mapping

Anne-Claude Gingras
Lunenfeld-Tanenbaum Research Institute
Sinai Health System
Toronto
gingras@lunenfeld.ca

A summary of the Jeanne Manery Fisher Memorial Lecture presented at the CSMB meeting in Montreal, June 2019

Proteins do not work alone: they associate with one another, for example to form molecular machines such as the ribosome, or to relay signals transmitted notably through kinase cascades. Fundamentally, interactions within molecular machines and across signalling pathways are different in their stability and affinity, and structures such as the ribosome are characterized by high-affinity interactions that tend to bury large interfaces (Figure 1). These interactions are stable, which enables biochemical purification of these complexes and the characterization of their constituents. By contrast, many of the contacts between signalling molecules involve lower affinity interactions between domains and short linear motifs, and the regulation of many of these interactions through post-translational modifications. Techniques that have been used successfully for the isolation of stable complexes may not always work for capturing these weaker interactions, especially if the purification approach involves several steps that result in exchange with the environment.

In the early 2000s, when I began my postdoctoral fellowship in the Aebersold lab in Seattle, I established in mammalian cells, a version of the then-popular “Tandem Affinity Purification”, or TAP-tagging. TAP-tagging consists of purification and re-purification of a bait and its binding partners, followed by their identification by mass spectrometry. This approach had already led Anne-Claude Gavin and others to systematically define stable complexes in yeast (Gavin et al. 2002), and indeed it enabled me to uncover new and evolutionarily conserved stable protein phosphatase complexes (Gingras et al. 2005), on which basis I started my own research group in Toronto at the end of 2005. The beauty of the approach was that the samples that were recovered were very clean, with distinct bands detectable on protein gels. Data analysis of the mass spectrometry results was therefore relatively simple, and many (and often most) of the proteins detected were true binding partners of the bait protein. Yet, what was missing from the complexes that I was identifying were the lower affinity interactors, which in the case of phosphatases would have included regulators and substrates.

What became clear is that to identify these lower affinity interactions we needed to revert to “dirtier” purifications, that would maintain the interactions through purification. However, this presented a challenge at the time, as there were no reliable computational methods to discriminate between true interactors and background contaminants. We, therefore, used to maintain our own “contaminant” lists, with proteins on this list systematically subtracted from
the list of detected proteins, or we would use several arbitrary cutoffs to define which proteins would make it to our final lists. Because my group was at the time running reasonably small projects, this was still doable, and we systematically cloned the interactors we were recovering to perform reciprocal affinity purification/mass spectrometry (AP-MS) experiments and functional validation. Here, I need to give kudos to my lab members at the time, particularly to my first two PhD students, Ginny Chen and Michelle Kean, and to my lab manager (and right arm), Marilyn Goudreault, whose scientific rigor and dedication enabled our early successes (Chen and Gingras 2007; Chen et al. 2008; Goudreault et al. 2009; Kean et al. 2011; Kean et al. 2012; Nakada et al. 2008).

My relatively peaceful life “in small-scale interaction proteomics” was abruptly stopped, however, when a colleague then at the Lunenfeld, Mike Tyers, asked me to help him on the analysis of a large project on the interactomes of yeast kinases and phosphatases. About 1500 pulldowns were performed in this study (almost all of them by a single technician, Zhen-Yuan Lin who is still in my group), but no systematic negative controls were embedded in the experimental design, which precluded simple solutions to discriminate between true interaction partners and contaminants. Fortunately (there is always an element of luck), I had met during my postdoctoral training Alexey Nesvizhskii, a physicist and fellow postdoc in the Aebersold lab who was then a new PI at the University of Michigan. Just as fortunately, Alexey had recruited a biostatistician, Hyungwon Choi, as a graduate student. With input from the entire team, Hyungwon implemented a non-supervised Bayesian statistical approach that enabled to score each possible pairwise interaction across very large datasets, and he created the first version of Significance Analysis of INTeractome (SAINT), which we used on the yeast kinase-phosphatase network (Breitkreutz et al. 2010). [A related approach to scoring high confidence interactions in large and disconnected datasets, CompPASS, was independently developed at around the same time by the Harper and Gygi group (Sowa et al. 2009)]. These tools used quantitative information embedded in each mass spectrometry experiment (at the simplest level, spectral counts, which are the number of times a peptide corresponding to a given protein is selected for fragmentation and identification by the mass spectrometer) to compare the recovery of a given candidate interactor (the “prey”) in a purification of a bait to that of the same prey across all baits in the dataset (Figure 2). Importantly, this statistical framework enables unbiased scoring, and selection of a cutoff could be guided by precision-recall analysis from published interactions, such as those deposited in the BioGRID database designed by Mike Tyers (Oughtred et al. 2019).

This unsupervised version of SAINT, like CompPASS, did not, however, work on small datasets, nor on highly connected datasets such as those we were generating through our iterative purifications of mammalian phosphatases. Very rapidly, Hyungwon and Alexey helped us develop a different, semi-supervised version of SAINT, which would train itself on the negative controls that were generated as part of every experiment (Choi et al. 2011; Choi et al. 2012). This turned out to be precisely what my lab needed: we could analyze small, medium or large datasets with this version of SAINT, obtaining good discrimination between true interactors and contaminants as long as the negative controls properly mimicked the behaviour of the contaminants (Figure 2). One of the first applications of this “new SAINT” was on a difficult
case, that of scoring true interactions with chaperones such as HSP90 (Skarra et al. 2011), which has enabled us to embark on this type of projects in a collaborative manner, notably with Mikko Taipale, now a new colleague in Toronto (Taipale et al. 2014). Hyungwon, who now has his own lab in Singapore, has been continuously improving SAINT, in particular to make it faster and more robust (Teo et al. 2014) and to adapt it with different types of quantitative data (Teo et al. 2016), including some tools that we helped co-develop for Data Independent Acquisition (Tsou et al. 2015; Wang et al. 2015).

What these exercises in developing scoring mechanisms and other computational tools for managing mass spectrometry data made clear was that these resources were not just needed by our own groups, but that they would benefit the entire research community. We, therefore, took an early decision to make all our computational tools freely available, and we were able to provide researchers with an efficient data management system (LIMS) for interaction proteomics, ProHits-LIMS, in another successful collaboration with Mike Tyers (Frank Liu and Jian Ping Zhang are still working on improvements in ProHits (Liu et al. 2016; Liu et al. 2012; Liu et al. 2010)). A talented postdoc (now a computational research associate in my group), James Knight, then helped us to tackle another challenge in interaction proteomics, that of data visualization. He created ProHits-viz (prohits-viz.lunenfeld.ca), a web-based tool to transform quantitative interaction lists into different simple visualizations, notably through dotplots that display at a glance the “raw” abundance values (how much prey is detected in the purification of a bait), the relative abundance value (how much prey with bait 1 versus bait 2), and the confidence of the bait-prey interaction (SAINT False Discovery Rate estimation). After an initial methods paper, James expanded his tools to include several different functions, such as correlation analysis, which is also supported by automated gene annotations within ProHits-viz (Knight et al. 2015; Knight et al. 2017). The next hurdle we were facing was the massive amount of data generated by our interaction proteomics screens: in each experiment (and particularly for the proximity-dependent biotinylation experiments discussed below), we identify as high confidence interactors hundreds of proteins. Retrieving information on these proteins through searching multiple databases became the bottleneck, and prompted James (with postdoctoral fellow Payman Samavarchi-Tehrani) to develop yet one more tool, this time a web browser plugin (Knight et al. 2019) that enables us to retrieve key and user-defined information by double-clicking on an identifier such as a gene name, directly on the webpage (e.g. NCBI PubMed or all the resources that our group developed).

Besides sharing software tools, we worked on disseminating our own proteomics data. We continue to work within the framework of existing repositories, including for mass spectrometry data deposition (our LIMS ProHits facilitates deposition of our proteomics results in the ProteomeXchange consortium (Deutsch et al. 2017) through partner MassIVE), and deposition of interaction data to BioGRID (Oughtred et al. 2019) and the IMeX consortium member IntAct (Orchard et al. 2014). Through my role as a Deputy Editor for the leading proteomics journal *Molecular and Cellular Proteomics*, I continue to advocate for sharing through these public repositories, but also realized that specialized databases are sometimes also critical. For instance, when we deposit our interaction data in databases such as BioGRID, most of the underlying quantitative information is lost, which means that as long as we deem
Draft

an interaction to be of “high-confidence”, it will be annotated the same way for stoichiometric and vastly sub-stoichiometric interactions. This limits the conclusions that can be drawn from the data. To circumvent this, we created a simple interaction repository (which we structure per project/publication) where we deposit all the interaction proteomics data in simple tabular formats that can be searched across all published datasets (we also have a “collaborative” area of the database where unpublished data can be accessed in a password-protected manner). Launched at the time of the publication of our interactome for the Hippo tumour suppressor pathway (Couzens et al. 2013), this resource (at prohits-web.lunenfeld.ca) now contains public data for multiple publications from our group and collaborators (14 independent projects are publicly available as of Aug 2019). Besides reporting the high confidence interactors for our own projects, however, we also realized that it would be crucial for the scientific community at large to have access to data that helps them scoring their own protein-protein interactions. Again in collaboration with Alexey Nesvizhskii’s lab, we created a Contaminant Repository for Affinity Purification (CRAPome.org), a resource that collates well-annotated and well-executed negative controls from the literature and enables researchers to either browse the resource, or upload their own experiments for analysis through some of the tools including SAINT (Mellacheruvu et al. 2013). Like most of the other tools that we have been creating, we have continued to improve this repository based on requests from users, and have now separated the Contaminant Repository from the analytical tools that are now within REPRINT (Resource for Evaluation of PRotein INTeractions; accessible at reprint-apms.org; Mellacheruvu et al. in prep). Our current computational pipeline is summarized in Figure 3B.

With these tools in hand – and a robust and shareable experimental pipeline (Figure 3A) – we and our collaborators have been able to explore the interactions for multiple signaling molecules. We aimed at understanding the interactome differences associated with cancer mutations and the regulation of protein-protein interactions by treatment with extracellular stimuli or pharmacological compounds (notably through the work of my ex postdoctoral fellow, JP Lambert, who currently heads his own group in Québec (Lambert et al. 2013; Lambert et al. 2019; Lambert et al. 2015; Lambert et al. 2014)). Yet, we (like everyone in the field) also noted some significant limitations with the AP-MS approach. AP-MS is defined by cell lysis and affinity purification of proteins that must both be solubilized and remain associated with their interactors throughout purification for proper detection. Poor solubility of some compartments, loss of weaker binding partners throughout even the single cell purifications, and the possibility of post-lysis artefacts are all drawbacks of standard AP-MS. Importantly, spatial resolution is also lost. Therefore, when the first in a series of methods describing biotinylation of proteins proximal to a bait in living cells was published (Roux et al. 2012), we (with Stéphane Angers and Brian Raught) were rapid in adopting the approach. BioID (as this approach is called) utilizes an abortive biotin ligase fused to a bait of interest that is expressed at near endogenous levels in living cells. The substrate (biotin), added extracellularly, is transported and then activated to a reactive intermediate by the biotin ligase; this intermediate covalently labels lysine residues, providing a handle for purification of proteins that came in proximity to the bait within living cells. Importantly, the entire computational pipeline we had developed for AP-MS could be directly repurposed for the analysis of BioID data, facilitating the adoption of this approach and its thorough benchmarking.
While multiple lab members simultaneously applied BioID to their projects, then postdoc Amber Couzens was the first in my lab to publish with the approach in the context of her Hippo signalling project (Couzens et al. 2013). Besides confirming that “insoluble” structures (as was first demonstrated by Roux et al. (Roux et al. 2012)) such as centrosomes, plasma- and endo-membranes, were amenable to BioID, Amber noted that phosphorylation-dependent and condition-specific interactions could be picked up in the absence of stimuli in BioID. This was particularly obvious for the okadaic acid-sensitive interaction between MOB1 and the Hippo kinase MST1. MST1 kinase autophosphorylates to create a docking site for the phosphopeptide-binding pocket in MOB1, a process that is opposed by an okadaic acid sensitive phosphatase (as we further structurally explored with the lab of Frank Sicheri (Couzens et al. 2017; Xiong et al. 2017; Xiong et al. 2018)). In standard AP-MS, the interaction between these proteins is only detectable following phosphatase inhibition, yet this relationship was readily picked up in BioID, without any inhibition necessary. Why is this? The long incubation times for BioID (Amber was performing a 24 h labelling) could provide an explanation: interactions between MST1 and MOB1 must constantly be cycling in cells (MST1 autophosphorylates, but at steady-state, the phosphatase “wins”). However, every time MST1 becomes phosphorylated, it has a chance to interact with the MOB1-bait become biotinylated: over 24 hours, and providing that the protein is stable, an accumulation of biotin can be detected on the pool of MST1 protein. We now refer to this phenomenon as signal amplification (see our recent review (Gingras et al. 2019)), and have since harnessed this approach for our studies of stress granules (Youn et al. 2018), which are dynamically formed membraneless organelles, while Brian Raught has applied it for the detection of substrates for E3 ubiquitin ligases (Coyaud et al. 2015).

Armed with this BioID pipeline, we have been exploring different projects in increasingly ambitious or complex scales, notably the definition of the composition, specificity and overall organization of P-bodies and stress granules by Ji-Young Youn who will soon be starting her independent position in Toronto (Youn et al. 2018). An ongoing project led by PhD student Christopher Go characterized all major membrane-bound and membraneless organelles in the cells by generating 234 stable cell pools, each expressing a BioID “marker” bait. 192 of these markers yielded useable data, which were used to draft a “proximity map” of a human cell, in this case, an HEK293 cell (Go et al. submitted). Importantly, however, this project again made us revisit data analysis, visualization and data sharing, and again, computational research associate James Knight designed a user-friendly solution. Humancellmap.org (which we opened as a resource prior to submission of the manuscript) provides all data associated with this project in different searchable and navigable formats, greatly facilitating data exploration. Additionally, building on our experience with REPRINT and the CRAPome (and in fact, compatible with scoring through REPRINT), we provide the users with an easy interface for uploading their own BioID data and comparing it to the humancellmap data. This can help reveal the subcellular localization of the bait (by similarity to baits in the humancellmap) and enable identifying proteins that are specifically enriched over an organelle (or the entire humancellmap) background. The concepts behind our organelles and global static proximity interactome labs have propelled new avenues of research led by different students and postdoctoral fellows in the group, as well as by our numerous collaborators. We are exploring
alternative cellular models through the development of lentiviral delivery tools (Payman Samavarchi-Tehrani) and are applying the concept of compartment sensors to revisit the spatial aspect of cellular signalling (postdoc Geoff Hesketh). These exciting new developments (and others I don’t have the space to discuss here) are enabling us to explore the spatio-temporal aspects of subcellular regulation, and open new areas of research in health and disease.

In summary, I have had a circuitous journey through protein-protein interactions, which has taken me very far from my initial comfort zone (biochemistry and phosphatases) but has been most rewarding. The tools we are collaboratively developing, alongside protocols and reagents, are made available to the scientific community, which enable others to improve scoring and analysis of their own interaction datasets, hopefully resulting in overall more reproducible and reusable research results.

Acknowledgements
Success in science is certainly not a one-man (or woman) affair. Besides my own extremely patient academic mentors (André Darveau, Nahum Sonenberg and Ruedi Aebersold), I need to acknowledge all my colleagues throughout my career and particularly in Toronto, and the numerous collaborators that make it fun to come to work. As described in this article, the collaborations with Alexey Nesvizhskii, Hyungwon Choi and Mike Tyers have shaped the proteomics bioinformatics efforts of our group, and I have been fortunate to work with (and live with) Brian Raught for over 20 years (and 35 publications). However, it is my trainees and other employees (a.k.a. HQPs) that keep the lab alive and productive. Each and every one of them has contributed in some way to make me grow as a scientist and a mentor. Lastly, none of this would have been possible without funding, particularly from CIHR, but also from Genome Canada/Ontario Genomics, NSERC and the NIH, alongside other grants for applications of our methods to study diseases, and in particular cancer, from the CCSRI, TFRI and CRS.

Figure legends

Figure 1. Characteristics of different protein-protein interactions. While “molecular machines”, for example the ribosome, are characterized by high affinity and large buried surfaces, interactions in signalling pathways must be dynamically regulated, often by post-translational modifications (PTMs), and are usually weaker. They are also often mediated by the interaction of a folded domain with short linear interaction motifs (SLiMs) that can be modified post-translationally. Methods designed for the purification and analysis of molecular machines often fail at capturing weaker or more transient interactions. See (Perkins et al. 2010) for extended discussion.

Figure 2. Scoring protein-protein interactions using quantitative mass spectrometry data. In current versions of Significance Analysis of INTeractome (SAINT), quantitative information for a prey in replicates biological (BR) purifications of a bait are scored against several negative controls samples (Choi et al. 2011). When datasets are sufficiently large and the baits not
connected to one another, scoring can also be performed across the entire dataset in the absence of specific negative controls. This type of scoring was exploited in the unsupervised early version of SAINT and in the popular CompPASS tool (Breitkreutz et al. 2010; Sowa et al. 2009).

**Figure 3.** Workflow of our current FLAG AP-MS and BioID pipelines.

**References**


Nakada, S., Chen, G.I., Gingras, A.C., and Durocher, D. 2008. PP4 is a gamma H2AX phosphatase required for recovery from the DNA damage checkpoint. EMBO Rep 9, 1019-1026.


Youn, J.Y., Dunham, W.H., Hong, S.J., Knight, J.D.R., Bashkurov, M., Chen, G.I., Bagci, H., Rathod, B., MacLeod, G., Eng, S.W.M., et al. 2018. High-Density Proximity Mapping Reveals the Subcellular Organization of mRNA-Associated Granules and Bodies. Mol Cell 69, 517-532 e511.
Protein-protein interactions

Transient

PTM

Weak transient  Strong transient  Permanent

Affinity

μM  nM

Interaction mechanism:

Examples:

SLiMs  Large buried surface

Signaling pathways  Molecular machines (ribosome)
Against controls

<table>
<thead>
<tr>
<th>Baits (BR)</th>
<th>Control</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preys</td>
<td></td>
</tr>
</tbody>
</table>

“new” SAINT

Across datasets

<table>
<thead>
<tr>
<th>Baits</th>
</tr>
</thead>
<tbody>
<tr>
<td>Preys</td>
</tr>
</tbody>
</table>

SAINT 1.0; CompPASS
### A Generic experimental pipeline for FLAG AP-MS (*BioID* modifications in italics)

<table>
<thead>
<tr>
<th>Step</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cloning cDNA</td>
<td>Cloning cDNA for bait protein into a pcDNA5-based vector to express fusion (FLAG, BirA*)</td>
</tr>
<tr>
<td>Generation of cell lines</td>
<td>Transfect cDNA into 293 Flp-In T-REx cells (single integration site) and select stably-transfected cells (hygromycin)</td>
</tr>
<tr>
<td>Verification of expression</td>
<td>Test protein expression and solubility (monitor biotinylation)</td>
</tr>
<tr>
<td>Induction</td>
<td>Induce expression with tetracycline and harvest cells (add biotin for BioID)</td>
</tr>
<tr>
<td>Gentle cell lysis (harsh lysis)</td>
<td>Optional gel separation</td>
</tr>
<tr>
<td>Affinity Purification</td>
<td>On bead trypsin digest</td>
</tr>
<tr>
<td>Peptide Preparation</td>
<td>Nano-scale Reversed-phase MS</td>
</tr>
<tr>
<td>Gel-free LC-MS/MS</td>
<td></td>
</tr>
<tr>
<td>Generate cell lysate</td>
<td></td>
</tr>
<tr>
<td>Affinity Purification</td>
<td></td>
</tr>
<tr>
<td>Peptide Preparation</td>
<td></td>
</tr>
<tr>
<td>Gel-free LC-MS/MS</td>
<td></td>
</tr>
</tbody>
</table>

### B Robust bioinformatics platform

- **Data Management**
  - Mass spectrometer
  - RAW file
  - File conversion
  - Peptide/protein identification / quant
  - Search results
  - Data parsing

- **Analyst**
  - Project
  - Bait
  - Experiment
  - Sample

- **SAINT analysis**
  - Low confidence interactions and absence of interaction
  - High confidence interactions
  - Mean counts

- **Visualization (ProHits-viz)**
  - Preys
  - Preys

- **ProHits-web CRAPome / REPRINT humancellmap.org**

- **Dissemination**

---

https://mc06.manuscriptcentral.com/bcb-pubs