Published in final edited form as:

J Proteomics. 2014 April 4; 100: 37–43. doi:10.1016/j.jprot.2013.10.023.

***SAINTexpress*: improvements and additional features in Significance Analysis of Interactome software**

Guoci Teo¹, Guomin Liu², Jianping Zhang², Alexey I. Nesvizhskii³, Anne-Claude Gingras², and Hyungwon Choi^{4,*}

¹Department of Statistics and Applied Probability, National University of Singapore, Singapore

²Lunenfeld-Tanenbaum Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada

³Departments of Pathology and Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI, USA

⁴Saw Swee Hock School of Public Health, National University of Singapore and National University Health System, Singapore

Abstract

Significance Analysis of INTERactome (SAINT) is a statistical method for probabilistically scoring protein-protein interaction data from affinity purification-mass spectrometry (AP-MS) experiments. The utility of the software has been demonstrated in many protein-protein interaction mapping studies, yet the extensive testing also revealed some practical drawbacks. In this paper, we present a new implementation, *SAINTexpress*, with a simpler statistical model and a quicker scoring algorithm, leading to significant improvements in computational speed and sensitivity of scoring. *SAINTexpress* also incorporates external interaction data to compute a supplemental topology-based score to improve the likelihood of identifying co-purifying protein complexes in a probabilistically objective manner. Overall, these changes are expected to improve the performance and user experience of SAINT across various types of high quality datasets.

Keywords

Affinity-purification; protein-protein interaction; probabilistic scoring

INTRODUCTION

Objective probabilistic scoring of protein-protein interaction data is a crucial step in AP-MS data analysis. Significance Analysis of INTERactome (SAINT) was one of the first algorithms developed to perform such scoring, and it has been successfully adopted and

Corresponding Author: All correspondence should be addressed to Dr. Hyungwon Choi by email at hyung_won_choi@nuhs.edu.sg.

Author Contributions

The software was written by GT; the data analysis was done by GT and HC; the paper was drafted by GT and HC, and it was finalized with contributions of all authors. ACG provided test data and feedback on the software performance; GL and JZ helped with iRefIndex analysis and integration of *SAINTexpress* within ProHits; AIN integrated *SAINTexpress* into the CRAPome interface. All authors have given approval to the final version of the manuscript.

Supporting Information. This software and user manual is freely available via the Internet at <http://saintapms.sourceforge.net/>.

tested by many studies. The method was originally developed for large-scale AP-MS experiments for which control purifications were not implicitly used for scoring [1]. The method was later extended to analyze datasets with negative control purifications, which allows more robust removal of background noise interactions from the data [2, 3]. While SAINT versions 1 through 2.1 (denoted v1 - v2.1 hereafter) were exclusively written for spectral count data, more recently, the algorithm was extended to use intensity-based quantitative data (v2.2 - v2.3.4) [4].

The statistical model in SAINT postulates that a prey protein identified in an affinity-purified sample is either a true interactor of the bait protein (true interaction) or a non-specific binder or contaminant (false interaction) depending on the quantitative data. In the case of true interaction, the prey must be observed in a sufficiently high abundance, especially at a level significantly higher than the abundance observed in the negative control purifications, which provide the information for the noise observations. Likewise, in the case of false interaction, the prey in the bait purification is not significantly more abundant than in the controls. To distinguish these possibilities, SAINT derives the posterior probability of true interaction from the two counterfactual models (true/false), given the quantitative data for a prey protein in a bait purification and control purifications (when available). Ultimately the probability score of SAINT gives an intuitive representation of the confidence level of putative interactions.

While it has been successfully applied in many studies, extensive testing has also revealed some drawbacks over the past few years. First, we discovered that the mathematical assumption of the model is not flexible enough to capture heterogeneous patterns in the quantitative data, notably when the selected bait proteins are chosen from a densely interconnected network [5]. By “densely interconnected”, we mean that some bait proteins in the target network share a large number of common interactors. When multiple baits are analyzed, the model assumes that the quantitative data of a common interactor (prey) are realizations from an identical distribution. This assumption becomes too stringent when the same prey is observed with varying quantitative levels with different baits, which occurs more often than not when many bait proteins interact with the same prey protein. For such preys, the interactions with weaker quantitative evidence tend to get penalized and assigned low scores, even if their abundance is well above that of control purifications [6]. To address this limitation, we have modified our statistical model to capture *all* the interactions with sufficient quantitative evidence, regardless of the interaction data of the same prey in other baits. While an alternative solution is to analyze each bait separately, as exemplified in the histone deacetylase (HDAC) interaction network data we analyze later [5], this requires preparation of separate input files for each bait and the model parameters may be estimated less reliably from a smaller data pool (data for each bait). The change we made in *SAINTexpress* allows fitting of one integrated model for all baits without penalizing the aforementioned cases.

Second, SAINT (v1 - v2.3.4) has used the quantitative data for each bait-prey pair to score the confidence of their interaction, without relying on any external information about the prey proteins. In some experiments, however, some prey proteins are clearly expected to co-purify (e.g. subunits of a protein complex), yet the quantitative evidence is not as convincing

for some of those preys and therefore they are assigned low scores by SAINT. As a remedy, the probability model in *SAINTexpress* incorporates this prior information regarding prey-to-prey relationship into the scoring by the Markov Random Field (MRF), which can adjust the posterior probabilities for the prey pairs that are known to be related. For example, if a previous experiment suggested that two preys are true interaction partners, a strong evidence for one of the preys in the current experiment will boost the score for the other prey in the same bait, and vice versa. The MRF model incorporates this knowledge in an objective manner, and the adjusted probability score is reported under the label of TopoAvgP, which stands for “topology-aware average probability score.”

Third, the statistical model was originally formulated as a Bayesian hierarchical model with a Markov chain Monte Carlo (MCMC) sampling procedure for nonparametric Bayes estimation, which had two practical constraints. MCMC is time consuming since it requires thousands of iterations to achieve convergence to the posterior distributions of model parameters, which can take tens of minutes in large datasets. Moreover, due to the nature of sampling-based estimation, the probabilities reported in the final output could vary depending on the seed in the random number generator. Lastly, the computational cost of the sampling-based estimation algorithm for the newly introduced MRF model was deemed prohibitive even for moderate-sized datasets. To address this challenge, we adopted the Iterated Conditional Mode (ICM) method for general MRF models [7], which generates the final output much faster than the Bayesian alternative. In this manuscript, we first explain these changes in more details and illustrate all three major changes and their impact on the analysis.

METHODS

The statistical model and the probability score in SAINT

We first review the statistical model of SAINT (as implemented in version 2.3.4). For clarity, we discuss the spectral count model with control purifications. The model for SAINT is a simple two-component mixture model

$$P(X_{ij}=x)=\prod_{i,j:j\in E}\{\pi_T P(x;\theta_i^T)+(1-\pi_T)P(x;\theta_i^F)\}\prod_{i,j:j\in C}P(x;\theta_i^F) \quad (\text{Eq. 1})$$

where θ^T and θ^F are the parameters of generalized Poisson distributions, including the level of abundance for true and false interactions, respectively. This is known as a semi-supervised mixture model in the sense that the negative distribution is estimated entirely from the data from negative control purifications. The model assumes that each interaction (bait j – prey i) has the underlying hidden states Z_{ij} , which takes on 1 if the interaction is true and 0 otherwise. Had there not been any experimental data, the prior probability of true interaction is π_T . After acquiring data, we report the posterior probability of true interaction as the final score (AvgP in the output file), which is computed as

$$P(Z_{ij}=1|X_{ij}=x)=\pi_T P(x;\theta_i^T)/\{\pi_T P(x;\theta_i^T)+(1-\pi_T)P(x;\theta_i^F)\} \quad (\text{Eq. 2})$$

by the Bayes theorem for each replicate and averaged over the replicates.

Change 1—Note that, in an experimental design where some bait proteins have more replicate experiments than others, it takes more reproducible spectral counts for the baits with more replicates to achieve the same probability score when AvgP is computed by taking the average of replicate-level probability scores. To address the unequal stringency in the AvgP score, *SAINTexpress* now offers the users an option to choose the best scoring R replicates for each interaction (the default is set to $R=100$ to use all replicates of each bait in most datasets). For example, if two replicate experiments were done for the majority of the baits while three replicates for a few selected baits, the appropriate choice of R would be 2.

Change 2—The estimation of statistical model parameters in SAINT (up to 2.3.4) was based on the Markov chain Monte Carlo (MCMC), a sampling algorithm to draw samples from appropriate posterior distribution of each model parameter. The major drawback of MCMC is that typically tens of thousands of samples are required to obtain robust estimates and thus running the algorithm can be very time consuming. This situation was likely to be aggravated if extra sampling steps were to be added for the MRF model. Hence we removed the MCMC-based estimation and instead used the Iterated Conditional Mode [7], a fast approximation of the posterior distribution of individual protein level parameters. With this procedure, the computation time was reduced significantly and datasets of any size can now be analyzed almost instantly.

Change 3—Estimation of θ^F . A major challenge with the SAINT model is in the estimation of the parameter θ^F , which is especially problematic for the preys that appear in a small number of baits. If this parameter is estimated for each prey protein, then the model will still attempt to discriminate each observed datum into the true and false interaction distributions based on available data. If the quantitative data varies wildly across the baits, however, the same protein observed with smaller spectral count(s) in one bait will be penalized by larger spectral count(s) in another bait because the model will learn θ^F as the average of all counts surpassing those in control purifications. This was frequently observed in datasets where many bait proteins share common interaction partners and prey proteins showed a lot of variation across the baits. To address this, we avoid estimating this parameter from the data and rather set it at $\theta^F=5\theta^T$. This will require that the spectral counts of true interactions should exceed around 2.25 times of the standard deviation from the average spectral counts in the controls. This way, as long as there is convincing quantitative evidence that a prey protein can be distinguished from negative controls, *SAINTexpress* will score those interactions high without imposing an unreasonable penalty. In SAINT (v2.2.3 - v2.3.4), we have included an option called “lowMode” to address the issue, but only for the high spectral count interactions (>100 counts), and therefore the option was not effective in datasets where most spectral counts do not exceed 100 counts. In *SAINTexpress*, the simplified rule above applies across all the interactions regardless of the range in spectral counts now.

Change 4—The current scoring method in SAINT (v1 - v2.3.4) is entirely based on the quantitative data alone and we assume that nothing is known about the relationship between

preys. With the growing databases of protein-protein interaction data such as iRefIndex, it is possible to incorporate these data to improve the outcome of filtering, especially co-purification of known protein complexes. *SAINTexpress* adopts the discrete local Markov random field (MRF) model developed in the area of image segmentation [7], where any pair of two prey proteins known to be interaction partners are assigned a boosted prior probability of co-purification in the same bait. The MRF model considers the unknown status of true and false interaction as a hidden state (Z). Just as in the mixture model above, the MRF model imposes the probability distribution over the entire set of hidden states, i.e. for all observed interactions. Denoting $Z = \{Z_{ij}\}$ and the MRF parameters $\Phi = (\alpha_0, \alpha_1, \beta)$, the model can be written as

$$p(Z; \Phi) \propto \exp(\alpha_0 n_0 + \alpha_1 n_1 + \beta n_{11})$$

where n_0 and n_1 denote the number of interactions in states 0 (false interaction) and 1 (true interaction), and n_{11} is the number of linked interactions (two preys in the same bait) that are in the true interaction state in the user-provided interaction data. Because we assume $\beta > 0$, this model encourages that the user-provided interaction partners are likely to co-purify in the same bait. This model describes an approximate prior probability “field” of true interactions across the entire data, which reduces to the individual interaction level by

$$p(Z_{ij}=1; Z_{\partial ij}, \Phi) \propto \exp\{\alpha_k + \beta u_{\partial ij}(1)\}$$

where Z_{ij} denotes the hidden states of all connected neighbor of prey i in bait j , and $u_{ij}(k)$ is the number of those neighbors that are also in the same state k . Using this local model to derive the equivalent of the mixture proportion parameter π_T in Equation (1), we derive the posterior probability of true interaction for each bait-prey pair $j \sim i$. Again, this model reinforces a simple principle in scoring: if many neighbors of prey i are true interactors of bait j , prey i is also likely to be a true interactor of the same bait. This probability is reported as a supplemental score “TopoAvgP” in the new output. We remark that this new score does not necessarily replace the original probability score AvgP. If TopoAvgP is greater than AvgP for an interaction, then this means that the prey protein has at least one other protein that co-purified with the same bait and therefore the given prey protein is likely to be a functionally relevant interaction partner of the bait.

Change 5—*SAINTexpress* also provides a few other useful summaries in the output file. First, it reports the fold change of spectral counts (or intensities) for each individual interaction, which is computed as the average spectral counts in the bait divided by the average spectral counts of the same prey in the control purifications (zero counts are replaced by 0.1). The fold change was also found to be a useful summary in the recent report describing the scoring of protein interactions using tools implemented as a part of the CRAPome resource (FC scores) [8]. This score allows the users to quickly inspect the output and rescue low probability interactions if necessary. Second, *SAINTexpress* reports the

Bayesian false discovery rate (FDR) estimates at all probability thresholds, which is computed directly from the posterior probabilities as

$$FDR(p^*) = \frac{\sum_{ij} (1 - p_{ij}) 1\{p_{ij} > p^*\}}{\sum_{ij} 1\{p_{ij} > p^*\}}$$

where $1\{A\}$ denotes the indicator function of event A . With this information, the user can determine the probability thresholds to control the FDR at the target rate.

RESULTS AND DISCUSSION

To demonstrate *SAINTexpress*, we have reanalyzed a recently published histone deacetylase (HDAC) network data [5]. The study charts the interaction map of 11 human HDACs, which are divided into four sub-classes featuring diverse functions and subcellular localizations. In this work, we benchmark *SAINTexpress* using the data for HDACs 1 through 10 only. The class IV deacetylase HDAC 11 was excluded from the analysis because considerably more interactions are captured in that bait protein potentially due to its divergent localization patterns, and hence the comparison of the two SAINT implementations would likely be dominated by the data from a single bait. We downloaded the processed spectral count data from Supplementary Tables 2 and 3 of the paper, and analyzed the same data using the most recent SAINT (v2.3.4) and *SAINTexpress*. For SAINT, we used 2,000 burn-in periods and 10,000 main iterations for the MCMC sampler, with the following three standard options: lowMode=0, minFold=1, normalize=0. The dataset consisted of 10,456 “interactions” with 2,496 prey proteins observed in at least one purification experiment across the 10 bait proteins. We remark that this analysis is different from the reported analysis in the original paper because the authors tested different options of SAINT (v2.3.2) and chose to report the SAINT analysis performed for each bait protein separately after careful examination of different outputs (note that this bait per bait SAINT analysis was intended to minimize the penalty for differential prey abundance across interconnected baits, as we have discussed above). *SAINTexpress* was applied to the same data with $R=2$ option because HDAC3-5 had 3 replicates while the other baits had 2 replicates only. The same option was applied in [5]. Control purification data was not compressed in both analyses.

Improved computation speed and sensitivity of detection

In terms of computational speed, *SAINTexpress* finished the analysis considerably faster than SAINT (v2.3.4). It took 37 minutes to run SAINT with 12,000 iterations of sampling steps, whereas *SAINTexpress* finished the analysis of the same data within merely 20 seconds. This almost instant model fit was made possible because we reduced the number of free parameters requiring statistical estimation, and especially by removing the time consuming sampling steps (Change 2).

Besides the computation time, the resulting probability scores also showed a good concordance between the two runs. Figure 1 shows the comparison of the reported probability scores (AvgP) between the two analyses. Overall, the AvgP scores reported from the two analyses were highly correlated (left panel, Pearson correlation 0.95). When the

interactions were ranked in a decreasing order of AvgP scores and benchmarked against the iRefIndex database, high confidence interactions filtered by the two analyses showed almost identical recovery of known interactions of the HDACs (right panel). Meanwhile, both analyses recovered iRefIndex interactions better than the published version of SAINT-filtered data, which was analyzed by running SAINT v2.3.2 separately on each bait. Nevertheless it should be remarked that bait-by-bait SAINT analysis in the paper was carefully chosen by the authors to address the loss of biologically relevant interactions due to the penalization of scores in densely interconnected baits in SAINT v2.3.2. This drawback seems to have been almost completely addressed in *SAINTexpress*. Overall, this comparative analysis showed that the simplified steps (Change 3) in *SAINTexpress* did not compromise the quality of filtering in exchange for the gain in speed.

At a reasonably high probability threshold 0.8 (reported FDR 5.4%), *SAINTexpress* and SAINT reported 639 and 697 interactions with an overlap of 584 interactions (>90%). In terms of the interactions that are unique to each analysis, the 55 interactions with AvgP 0.8 only in *SAINTexpress* were mostly penalized in SAINT (v2.3.4) because of the interactions involving the same prey proteins observed with extremely high spectral counts (tens or hundreds) in other baits – which is precisely the simplified model *SAINTexpress* was designed to address (see Supplementary Table 1). Interestingly, these included the interactions of the REST co-repressor RCOR1 and RCOR2 with Class I HDAC3, nuclear receptor co-repressors NCOR1 and NCOR2 with Class IIa HDACs 4,5,7, and F-box-like/WD repeat-containing protein TBL1XR1 in HDACs 1,4,5, all of which were integrated into the HDAC interactome as high confidence interactions in the original study after acquiring the data from analysis with various options in the original study. By contrast, those 113 interactions with AvgP 0.8 only in SAINT (v2.3.4) were mostly borderline cases where the spectral counts in the baits were about a few fold above the average spectral counts in the negative control purifications, or those cases involving the prey proteins with highly heterogeneous spectral counts in the control purifications (Supplementary Table 1). While SAINT (v2.3.4) assigned high scores for these interactions, they did not pass the same threshold in *SAINTexpress* potentially due to the five-fold mean model introduced in (Change 3).

Incorporating external interaction data

To show the impact of incorporating the external data sources on the confidence scores (Change 4), we first mapped the prey proteins to two external sources to derive the prey-to-prey relationship. We first extracted all prey-to-prey interactions from the most recent release (version 11) of the human iRefIndex database to incorporate known interaction data into scoring using the MRF model. We limited the search to ~21,000 interactions where both interaction partners are annotated as human proteins in iRefIndex. This mapping resulted in 11,760 pairs among 1,100 of 2,496 prey proteins, implying that the current iRefIndex (version 11) covers on average 9 known interactions for each prey protein in this dataset (~21 for those 1,100 preys). Alternatively, we also regarded the co-membership of prey-prey pairs in common Biological Process terms in the Gene Ontology (GO) database as a potential source of adjusting the prior confidence of true interaction, and thus repeated the same exercise. To prevent large GO terms from forming spurious groups of preys, we

restricted our mapping to those terms consisting of 20 or less members in the GO database for specificity. This mapping resulted in 12,826 pairs among 1,592 prey proteins, where each prey protein had on average 5 co-members to common GO terms (8 for those 1,592 preys).

We used this prey-prey interaction map as the additional input file to *SAINTexpress*, which were used to aid computation of the topology-assisted probability score called TopoAvgP. This new score incorporates the prey-prey interaction information as follows. Suppose that preys A and B are known interaction partners and they were captured in a bait purification along with an independent prey C (not known to interact with neither A nor B). If prey A has a strong quantitative evidence, and preys B and C have a weak evidence, then the probability score will be boosted for B, but not C. The degree of boosting is governed by an objective statistical model (MRF model), which will be learned in each and every dataset. The two panels of Figure 2 show the comparison between TopoAvgP and AvgP scores in *SAINTexpress*. It is easy to see that the data from the iRefIndex database had a minimal impact on the scoring, whereas the co-membership to biological processes (of size 20 or less) changed the score for many bait-prey interactions. While the AvgP score derived solely from the AP-MS data remains the main score, the new TopoAvgP score highlights the interactions that did not score well individually, but are of potential interest for further follow-up due to their previously reported relevance with other real interactions in the same bait. These cases tend to occur when the prey protein was identified at low abundance on its own, but some of its co-members were discovered as high-confidence interaction partners of the same bait protein.

The contrasting impact of the iRefIndex and the GO term co-membership data on the TopoAvgP score reveals an interesting insight of how the MRF model links these database interactions with the target interactome. The lack of impact by the iRefIndex data suggests either of two possibilities: the prey-to-prey interaction map in the database has a poor coverage of the proteins connected to the HDAC interactome, or all known interactions already scored considerably high before borrowing the prey-prey information. By the same reasoning, the contribution of the GO term co-membership to the TopoAvgP score indicates that the GO terms included many preys closely related to the interaction network of HDACs, resulting in boosted TopoAvgP score especially for those preys identified by low spectral counts. To see which specific GO terms contributed to boosting, we looked at the 60 interactions for which TopoAvgP score was greater than AvgP by 0.2 or more (these occurred mostly in HDAC1 and HDAC2 purifications), and collected the GO terms that were queried for co-membership. These terms were mostly the protein complexes involved in chromatin remodeling such as nucleosome disassembly, SWI/SNF complex, BAF complex, and proteasome core complex, indicating that the interactions that benefited from functionally relevant GO terms related to transcription regulation, for which histone acetyltransferases (HATs) and HDACs are obviously critical regulators [9, 10].

Utility for SAINT (v1-v2.3.4)

While the main advantage of *SAINTexpress* is the computation speed and improved sensitivity in datasets with interconnected baits, *SAINTexpress* no longer supports the

options to flexibly tune the statistical model for different kinds of datasets. We will therefore keep maintaining support for SAINT (v2.3.4) to use on datasets for which specific tailoring is needed. The statistical model in SAINTexpress is equivalent to that in SAINT with the following options: lowMode=0, minFold=1, and normalize=0. Our own experience suggests that turning off the minFold option could be beneficial when the target interactome contains many preys that appear frequently in the negative control runs such as chaperone proteins, though *SAINTexpress* was able to score most relevant interactions in an interaction network centered on HSP90 (Taipale *et al.*, *in preparation*). A recent study also showed that the normalize option is critical when heterogeneous control data were pooled from multiple sources [8], though this option is not as useful when controls and test samples are quantitatively comparable. It remains to be seen how well *SAINTexpress* handles these datasets without options through further validation in upcoming studies.

DISCUSSION AND CONCLUSION

In this work, we presented *SAINTexpress*, a new implementation of SAINT that runs much faster and addresses the vulnerability of the statistical model to the quantitative variation of a prey protein across different purification experiments. The software now provides additional probability scores that incorporate known interactions between prey proteins, in addition to the simple fold change estimate for further inspection of low scoring interactions. The new implementation no longer uses the time consuming sampling-based estimation procedure, and the output is also devoid of any random variation due to the dependence on sampling.

One of the novel features of *SAINTexpress* is the additional topology-assisted probability score (TopoAvgP), through which the user can incorporate the prior knowledge of the target interactome into the scoring step. Using an objective probability model (the MRF model), *SAINTexpress* increases the chance of identifying known protein complexes by boosting the probability score of each member of a protein complex if other members in the same complex are high confidence interactors. However, we note that large databases such as iRefIndex collate many different types of interaction data generated using different experimental techniques, each of which supports the experimental evidence of each interaction to a varying degree. While we used the entire iRefIndex database for mapping prey-prey interactions in the analysis of the HDAC dataset, users can filter the database and select a relevant subset to the experimental technique used in the given dataset. *SAINTexpress* version 3.2.0 will allow the users to provide multiple interaction files, reporting the TopoAvgP scores for each input file. This feature allows for the comparison of contribution from different types of database interactions to the scoring.

Overall, *SAINTexpress* addressed the most critical drawbacks of the existing version of SAINT and it is expected to improve the user experience significantly. The software is available for download at <http://saint-apms.sourceforge.net>, and is also available as part of the ProHits LIMS which affords AP-MS sample tracking and annotation and an easy-to-use graphical user interface [11, 12]. Lastly, *SAINTexpress* is available alongside SAINT v2.3.4 as part of the web-accessible Contaminant Repository for Affinity Purification resource (www.crapome.org) [8].

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Funding Sources

This research was supported by a National University of Singapore grant (HC), and NIH grant R01-GM-094231 (ACG and AIN).

We are grateful to the users of SAINT in the proteomics community for their helpful feedback, and to Mikko Taipale and members of the Gingras laboratory for providing unpublished test data and feedback on the scoring. We also thank Zach Wright and Dattatreya Mellacheruvu for their assistance with integration of SAINTexpress into the CRAPome interface.

ABBREVIATIONS

SAINT	significance analysis of interactome
AP-MS	affinity purification-mass spectrometry
FDR	false discovery rate
HDAC	histone deacetylase
MRF	Markov random field

References

- Breitkreutz A, Choi H, Sharom JR, Boucher L, Neduva V, Larsen B, et al. A global protein kinase and phosphatase interaction network in yeast. *Science*. 2010; 328:1043–6. [PubMed: 20489023]
- Choi H, Larsen B, Lin ZY, Breitkreutz A, Mellacheruvu D, Fermin D, et al. SAINT: probabilistic scoring of affinity purification-mass spectrometry data. *Nat Methods*. 2011; 8:70–3. [PubMed: 21131968]
- Choi H, Liu G, Mellacheruvu D, Tyers M, Gingras AC, Nesvizhskii AI. Analyzing protein-protein interactions from affinity purification-mass spectrometry data with SAINT. *Curr Protoc Bioinformatics*. 2012; Chapter 8(Unit8):15. [PubMed: 22948729]
- Choi H, Glatter T, Gstaiger M, Nesvizhskii AI. SAINT-MS1: protein-protein interaction scoring using label-free intensity data in affinity purification-mass spectrometry experiments. *J Proteome Res*. 2012; 11:2619–24. [PubMed: 22352807]
- Joshi P, Greco TM, Guise AJ, Luo Y, Yu F, Nesvizhskii AI, et al. The functional interactome landscape of the human histone deacetylase family. *Mol Syst Biol*. 2013; 9:672. [PubMed: 23752268]
- Nesvizhskii AI. Computational and informatics strategies for identification of specific protein interaction partners in affinity purification mass spectrometry experiments. *Proteomics*. 2012; 12:1639–55. [PubMed: 22611043]
- Besag J. On the Statistical-Analysis of Dirty Pictures. *J Roy Stat Soc B Met*. 1986; 48:259–302.
- Mellacheruvu D, Wright Z, Couzens AL, Lambert J-P, St-Denis N, Li T, et al. The CRAPome: a Contaminant Repository for Affinity Purification mass spectrometry data. *Nat Methods*. 2013 In press.
- Clapier CR, Cairns BR. The biology of chromatin remodeling complexes. *Annu Rev Biochem*. 2009; 78:273–304. [PubMed: 19355820]
- Muratani M, Tansey WP. How the ubiquitin-proteasome system controls transcription. *Nat Rev Mol Cell Biol*. 2003; 4:192–201. [PubMed: 12612638]

11. Liu G, Zhang J, Choi H, Lambert JP, Srikumar T, Larsen B, et al. Using ProHits to store, annotate, and analyze affinity purification-mass spectrometry (AP-MS) data. *Curr Protoc Bioinformatics*. 2012; Chapter 8(Unit8):16. [PubMed: 22948730]
12. Liu G, Zhang J, Larsen B, Stark C, Breitkreutz A, Lin ZY, et al. ProHits: integrated software for mass spectrometry-based interaction proteomics. *Nat Biotechnol*. 2010; 28:1015–7. [PubMed: 20944583]

SIGNIFICANCE

We present *SAINTexpress*, an upgraded implementation of Significance Analysis of INTeractome (SAINT) for filtering high confidence interaction data from affinity purification-mass spectrometry (AP-MS) experiments. *SAINTexpress* features faster computation and incorporation of external data sources into the scoring, improving the performance and user experience of SAINT across various types of datasets.

Figure 1.

Comparison of SAINT versus *SAINTexpress*. The left panel shows the AvgP scores computed from the two methods against each other. The right panel shows the recovery of known interactions in the iRefIndex database by the two methods.

Figure 2.

TopoAvgP versus AvgP in *SAINTexpress*. TopoAvgP scores were computed using the iRefIndex database and small GO Biological Process terms (of size <20) respectively.