BIOGRAPHICAL SKETCH

NAME: Haiyuan Yu

eRA COMMONS USER NAME (credential, e.g., agency login): HAIYUANYU

POSITION TITLE: Associate Professor

EDUCATION/TRAINING

INSTITUTION AND LOCATION	DEGREE	Completion Date	FIELD OF STUDY
Peking University, Beijing, P.R. China	B.S.	06/00	Biophysics and Physiology
Yale University, New Haven, CT	Ph.D.	01/06	Computational Biology and Bioinformatics
Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA	Postdoctoral	12/09	Systems Biology

A. Personal Statement

I have considerable expertise in both experimental and computational systems biology. On the computational front, I am specialized in network analysis and structural modeling. I was the first to implement a machine-learning scheme to quantitatively evaluate the degree to which TF-target relationships can be reliably transferred across species as a function of sequence similarity (***Genome Research***, 2004). On topics related to this proposal, I was among the first to perform comparative network analysis and developed a myriad of algorithms and tools to identify key nodes and edges in regulatory network using topological parameters (such as betweenness) and to investigate the effects on the whole network and individual pathways from disruptions of certain TFs (***Genome Biology***, 2017). Here at Cornell, I led my group to develop a novel proteome-scale homology modeling algorithm to construct the first 3D human interactome network to analyze and predict potential molecular mechanisms of disease mutations (***Nature Biotechnology***, 2012).

I would like to emphasize that I have extensive training and experience in high-throughput experiments. In my lab here at Cornell, we have established a fully automated pipeline using my Tecan Freedom Evo 200 bio-robot with LiHa, MCA, and RoMa arms for high-throughput Gateway cloning, minipreps, transcriptome readout and enhancer/promoter assays, and interactome screening. I led my group to finish the first binary interactome map in *S. pombe* by screening the whole proteome three times (>75 million pairs tested) using our InPOINT pipeline (***Cell***, 2016). We have also successfully set up the first massively-parallel site-directed mutagenesis pipeline, Clone-seq, and used it to generate >4,000 WT and mutant clones; we investigated molecular mechanisms for 204 disease mutations (***PLoS Genetics***, 2014). We were active members of the 1000 Genome Project Functional Interpretation Group (Dr. Gerstein is the co-Chair) to experimentally investigate functional relevance of population variants (***Science***, 2013). I believe that my expertise and experience have prepared me to lead the project in this proposal. Furthermore, with the established on-going collaborations among the Yu, Rubin, Levchenco, and Gerstein groups, we are well positioned to successfully complete the proposed project.

1. Liang, S., Tippens, N.D., Zhou, Y., Mort, M., Stenson, P.D., Cooper, D.N., and **Yu, H.,** iRegNet3D: three-dimensional integrated regulatory network for the genomic analysis of coding and non-coding disease mutations. ***Genome Biology,*** 2017. 18(1):10.
2. Vo, T., Das, J., Meyer, M., Cordero, N., Akturk, N., Wei, X., Fair, B., Degatano, A., Fragoza, R., Liu, L., Matsuyama, A., Trickey, M., Grimson, A., Yamano, H., Yoshida, M., Roth, F., Pleiss, J., Xia, Y., and **Yu, H.**, A proteome-wide fission yeast interactome reveals network evolution principles from yeasts to human. ***Cell***, 2016. 164(1):310-323.
3. Wei, X., Das, J., Fragoza, R., Liang, J., Bastos de Oliveira, F. M., Lee, H. R., Wang, X., Mort, M., Stenson, P. D., Cooper, D. N., Lipkin, S. M., Smolka, M. B., and **Yu, H.**, A massively parallel pipeline to clone DNA variants and examine molecular phenotypes of human disease mutations. ***PLoS Genetics***, 2014. 10(12):e1004819.

4. Khurana, E., Fu, Y., …, Lipkin, S. M., …, 1000 Genomes Project Consortium, Dermitzakis, E., **Yu, H.**, *Rubin, M.*, Tyler-Smith, C., and *Gerstein, M.*, Integrative annotation of variants from 1,092 humans: application to cancer genomics*.* ***Science***, 2013. 342(6154):1235587.

B. Positions and Honors

Positions and Employment

2006-2009	Postdoctoral Research Fellow, Department of Cancer Biology, Dana-Farber Cancer Institute and Department of Genetics, Harvard Medical School
2010-2015	Assistant Professor, Department of Biological Statistics and Computational Biology and Weill Institute for Cell and Molecular Biology, Cornell University
2015-present	Associate Professor, Department of Biological Statistics and Computational Biology and Weill Institute for Cell and Molecular Biology, Cornell University
2017-present	Director, Center for Genomic and Proteomic Technology Development (CGPT), Cornell University

Selected Awards and Honors

1999	Procter & Gamble Scholarship
2000	Award for excellent undergraduate thesis of Peking University
2000-2005	Henry L. Fan Fellowship
2013	*ad hoc* reviewer on NSF III small panel for bioinformatics (IX-BIO)
2013	*ad hoc* reviewer on NIH Program Project Grant Review Panel for National Heart, Lung, and Blood Institute (NHLBI)
2014	Cornell CALS Outstanding Accomplishments in Early Achievement Award
2014-2015	*ad hoc* reviewer on NIH Genomics, Computational Biology and Technology (GCAT) Study Section
2015	*ad hoc* reviewer on NIH Mouse Models for Translational Research (ZRG1 OTC-W-55) Study Section

C. Contribution to Science (out of a total of ***72*** publications; * co-first author, ¶ co-corresponding author)

1. ***Clone-seq: the first massively-parallel site-directed mutagenesis pipeline using next-generation sequencing.*** Due to rapid advances in next-generation sequencing technologies, tens of thousands of disease-associated mutations and millions of genomic variants have been identified in the human population. This drives an urgent need to develop high-throughput experimental methods to sift through this deluge of sequence data to quickly determine the functional relevance of each variant at the molecular level. Unfortunately, such methods do not exist currently. In my group at Cornell University, we have successfully developed Clone-seq to address this issue, which can generate >3000 mutant clones in a single lane of a 1×100 bp HiSeq run, decreasing costs by at least 10-fold. Clone-seq is entirely different from previously described random mutagenesis approaches. In Clone-Seq, each mutant clone has a separate stock. Different clones can therefore be used separately for completely different downstream assays. In conjunction with Clone-seq, we established a high-throughput comparative interactome-scanning pipeline, integrating GFP, InPOINT (PCA, LUMIER, Y2H, wNAPPA), and mass spectrometry assays to systematically examine the effect of variants on protein stability and individual interactions. We confirmed that the molecular phenotypes measured by our high-throughput GFP and InPOINT assays are biologically relevant *in vivo*. Furthermore, by comparing the molecular phenotypes, in particular the protein interaction disruption profiles, of variants to those of known disease mutations, potential candidate mutations for a variety of diseases can be identified.

a. 1000 Genomes Project Consortium, A global reference for human genetic variation*.* ***Nature***, 2015. 526(7571):68-74.

b. Wei, X., Das, J., Fragoza, R., Liang, J., Bastos de Oliveira, F. M., Lee, H. R., Wang, X., Mort, M., Stenson, P. D., Cooper, D. N., Lipkin, S. M., Smolka, M. B., and **Yu, H.**, A massively parallel pipeline to clone DNA variants and examine molecular phenotypes of human disease mutations*.* ***PLoS Genet***, 2014. 10(12):e1004819.

c. Khurana, E., Fu, Y., Colonna, V., Mu, X., Kang, H., Lappalainen, T., Sboner, A., Lochovsky, L., Chen, J., Harmanci, A., Das, J., Abyzov, A., Balasubramanian, S., Beal, K., Chakravarty, D., Challis, D., Chen, Y., Clarke, D., Clarke, L., Cunningham, F., Evani, U., Flicek, P., Fragoza, R., Garrison, E., Gibbs, R., Gümüş, Z., Herrero, J., Kitabayashi, N., Kong, Y., Lage, K., Liluashvili, V., Lipkin, S. M., MacArthur, D., Marth, G., Muzny, D., Pers, T., Ritchi, G., Rosenfeld, J., Sisu, C., Wei, X., Wilson, M., Xue, Y., Yu, F., Consortium, G.P., Dermitzakis, E., **Yu, H.**, *Rubin, M.*, Tyler-Smith, C., and *Gerstein, M.*, Integrative annotation of variants from 1,092 humans: application to cancer genomics. ***Science***, 2013. 342(6154):1235587.

d. Zhong, Q., Simonis, N., Li, Q.R., Charloteaux, B., Heuze, F., Klitgord, N., Tam, S., **Yu, H.**, Venkatesan, K., Mou, D., Swearingen, V., Yildirim, M.A., Yan, H., Dricot, A., Szeto, D., Lin, C., Hao, T., Fan, C., Milstein, S., Dupuy, D., Brasseur, R., Hill, D.E., Cusick, M.E., and Vidal, M., Edgetic perturbation models of human inherited disorders*.* ***Molecular Systems Biology***, 2009. 5:321.

2. ***The first comprehensive high-quality 3D protein interactome network for human disease proteins through an innovative proteome-scale homology-modeling approach.*** The interaction interfaces of all interactions in this 3D network have been determined at the atomic resolution. Using this network, I confirmed for the first time that alterations of specific protein interactions play a critical role in the pathogenesis of most disease genes at the genomic level. I discovered that mutations on different interaction interfaces of the same protein are significantly more likely to cause different disorders, the complete opposite of the widely accepted "guilt-by-association" principle. I found that although recessive mutations on the interaction interface of two interacting proteins tend to cause the same disease, the "guilt-by-association" principle does not apply to dominant mutations. I showed that a significant portion of truncating alleles, even those close to the N-terminus, can generate functional protein products, indicating that the common practice of considering these as "knock-out" alleles is flawed and may lead to incorrect conclusions for many truncating alleles.

a. Wang, X., Wei, X., Thijssen, B., Das, J., *Lipkin, S.M.* and **Yu, H.**, Three-dimensional reconstruction of protein networks provides insight into human genetic disease*.* ***Nature Biotechnology***, 2012. 30(2):159-64. *(Highlighted as "Research Highlights" in Nature Methods, 2012, 9: 220-1, as "Technology Feature" in Nature, 2012, 484: 271-5, and and as "Feature" in Nature, 2013, 494:416-9; Evaluated as "Exceptional" by "FACULTY of 1000")*

b. Guo, Y., Wei, X., Das, J., Grimson, A., Lipkin, S.M., Clark, A.G. and **Yu, H.**, Dissecting Disease Inheritance Modes in a Three-Dimensional Protein Network Challenges the "Guilt-by-Association" Principle*.* ***American Journal of Human Genetics***, 2013. 93(1):78-89.

c. Meyer, M. J., Das, J., Wang, X. and **Yu, H.**, INstruct: a database of high-quality 3D structurally resolved protein interactome networks*.* ***Bioinformatics***, 2013. 29(12):1577-9.

d. Das, J., Lee, H. R., Sagar, A., Fragoza, R., Liang, J., Wei, X., Wang, X., Mort, M., Stenson, P. D., Cooper, D. N., and **Yu, H.**, Elucidating common structural features of human pathogenic variations using large-scale atomic-resolution protein networks*.* ***Human Mutation***, 2014. 35(5):585-93.

3. ***Extensive experience in regulatory network analysis and machine learning algorithms.*** It has been argued that the number of interaction partners (i.e., degree) of proteins positively correlates with essentiality. I showed that this is not true. In regulatory networks, gene essentiality instead correlates with the betweenness of TFs. *I was among the first to perform comparative network analysis.* I was the first to implement a machine-learning scheme to quantitatively evaluate the degree to which TF-target relationships can be reliably transferred across species as a function of sequence similarity. I found that TFs tend not to co-express with their targets, but rather have time-delayed relationships. I found that the regulatory networks in both yeast and *E. coli* have the same hierarchical structure, with master regulators on top. This structure resembles social hierarchies, with many common characteristics optimized for efficiency. I developed three different machine-learning algorithms to successfully predict protein-protein and protein-DNA interactions on a genomic scale. More recently, we performed integrated network analysis that reveals distinct regulatory roles of TFs and microRNAs. We have established the first comprehensive high-resolution integrated 3D regulatory network (iRegNet3D) in the form of a web tool, where we resolve the interfaces of all known transcription factor (TF)-TF, TF-DNA and chromatin-chromatin interactions, for the analysis of both coding and non-coding disease-associated mutations to obtain mechanistic insights into their functional impact. Using iRegNet3D, we find that disease-associated mutations may perturb the regulatory network through diverse mechanisms including chromatin looping. iRegNet3D promises to be an indispensable tool in large-scale sequencing and disease-association studies.

a. Liang, S., Tippens, N.D., Zhou, Y., Mort, M., Stenson, P.D., Cooper, D.N., and **Yu, H.,** iRegNet3D: three-dimensional integrated regulatory network for the genomic analysis of coding and non-coding disease mutations. ***Genome Biology,*** 2017. 18(1):10.

b. Guo Y, Alexander K, Clark AG, Grimson A, **Yu, H.**, Integrated network analysis reveals distinct regulatory roles of transcription factors and microRNAs. ***RNA***, 2016. 22(11):1663-1672.

c. **Yu, H.**¶ and *Gerstein, M.*¶, Genomic analysis of the hierarchical structure of regulatory networks. ***Proceedings of the National Academy of Sciences of the United States of America (PNAS)***, 2006. 103(40):14724-31.

d. **Yu, H.**, Luscombe, N. M., Lu, H. X., Zhu, X., Xia, Y., Han, J. D., Bertin, N., Chung, S., Vidal, M. and *Gerstein, M.*, Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. ***Genome Research***, 2004. 14(6):1107-18.

4. ***The internally-controlled quality-assessing framework for interactome mapping.*** Combining my rigorous statistical insight and extensive experimental involvement, I developed an internally-controlled interactome mapping framework with a reference set of known positives and negatives in the relevant organism. It accounts for all potential noises in a high-throughput interactome mapping experiment and allows for quantitatively and experimentally measuring the quality of individual interactions as well as the whole dataset produced by any given technology. Applying this framework, I was able to clarify several widely-accepted misconceptions about the data quality of high-throughput technologies. For the first time, I confirmed that directy physical interactions generated by high-throughput InPOINT assays are of high-quality; Affinity-purification followed by mass spectrometry (AP-MS) methods detect co-complex membership, not binary interaction; thus AP-MS methods are not comparable, but rather complementary to InPOINT. Just like the implementation of the Phred quality score algorithm helped launch genome sequencing, this quality-assessing framework has fundamentally changed the way current and future interactome mapping projects are designed and carried out.

a. **Yu, H.**, Braun, P., Yildirim, M.A., Lemmens, I., Venkatesan, K., Sahalie, J., Hirozane-Kishikawa, T., Gebreab, F., Li, N., Simonis, N., Hao, T., Rual, J.F., Dricot, A., Vazquez, A., Murray, R.R., Simon, C., Tardivo, L., Tam, S., Svrzikapa, N., Fan, C., de Smet, A.S., Motyl, A., Hudson, M.E., Park, J., Xin, X., Cusick, M.E., Moore, T., Boone, C., Snyder, M., Roth, F.P., Barabasi, A.L., Tavernier, J., Hill, D.E., and Vidal, M., High-quality binary protein interaction map of the yeast interactome network. ***Science***, 2008. 322(5898):104-10. *(Highlighted in "Perspectives" in Science, 2008, 322: 56-7; Evaluated as "Exceptional" by "FACULTY of 1000")*

b. Cusick, M.E.*, **Yu, H.***, Smolyar, A., Venkatesan, K., Carvunis, A.R., Simonis, N., Rual, J. F., Borick, H., Braun, P., Dreze, M., Vandenhaute, J., Galli, M., Yazaki, J., Hill, D. E., Ecker, J. R., Roth, F. P., and Vidal, M., Literature-curated protein interaction datasets. ***Nature Methods***, 2009. 6(1):39-46.

c. Venkatesan, K., Rual, J. F., Vazquez, A., Stelzl, U., Lemmens, I., Hirozane-Kishikawa, T., Hao, T., Zenkner, M., Xin, X., Goh, K. I., Yildirim, M. A., Simonis, N., Heinzmann, K., Gebreab, F., Sahalie, J. M., Cevik, S., Simon, C., de Smet, A. S., Dann, E., Smolyar, A., Vinayagam, A., **Yu, H.**, Szeto, D., Borick, H., Dricot, A., Klitgord, N., Murray, R.R., Lin, C., Lalowski, M., Timm, J., Rau, K., Boone, C., Braun, P., Cusick, M.E., Roth, F.P., Hill, D.E., Tavernier, J., Wanker, E.E., Barabasi, A.L., and Vidal, M., An empirical framework for binary interactome mapping. ***Nature Methods***, 2009. 6(1):83-90.

5. ***Most comprehensive binary interactome networks generated for budding yeast, fission yeast, and human.*** Two main problems with current high-throughput interaction detection technologies are low sensitivity and high false-positive rates. To solve these problems, I developed a high-throughput interaction-detection tool-kit, InPOINT, consisting of four complementary experimental assays: Protein Complementation Assay (PCA), Nucleic Acid Programmable Protein Array in Wells (wNAPPA), Y2H, and LUminescence-based Mammalian IntERactome mapping (LUMIER). The performance of these assays has been optimized using the quality-assessing framework I developed. The use of four complementary assays greatly improves the sensitivity of our experiments while essentially eliminating false-positives. Using InPOINT, I lead the effort to screened the whole budding yeast proteome three times independently to yield *a high-coverage high-quality binary interactome network in yeast consisting of 2930 interactions among 2018 proteins, the most comprehensive in any organism to date*. I developed the "Stitch-seq" technology to utilize next-generation sequencing, increasing the throughput and decreasing the cost of interactome mapping by at least 10~100-fold. *Using Stitch-seq, I generated the most comprehensive human binary interactome network at the time with 8713 interactions.* In my own lab here at Cornell, we have just generated the first fission yeast binary interactome by screening the whole proteome three times. We erformed the first systematic cross-species interactome mapping and discovered that half of the conserved interactions have undergone co-evolution and thus proteins involved in these interactions no longer interact with the orthologs of their binding partners in other species. We identified a novel stress response factor, a previously uncharacterized protein we named Snr1 in fission yeast, whose functions have diverged from its budding yeast ortholog, Ehd3.

a. Vo, T., Das, J., Meyer, M., Cordero, N., Akturk, N., Wei, X., Fair, B., Degatano, A., Fragoza, R., Liu, L., Matsuyama, A., Trickey, M., Grimson, A., Yamano, H., Yoshida, M., Roth, F., Pleiss, J., Xia, Y., and **Yu, H.**, A proteome-wide fission yeast interactome reveals network evolution principles from yeasts to human. ***Cell***, 2016. 164(1):310-323.

b. **Yu, H.**, Tardivo, L., Tam, S., Weiner, E., Gebreab, F., Fan, C., Svrzikapa, N., Hirozane-Kishikawa, T., Rietman, E., Yang, X., Sahalie, J., Salehi-Ashtiani, K., Hao, T., Cusick, M. E., Hill, D. E., Roth, F. P., Braun, P., and Vidal, M., Next-generation sequencing to generate interactome datasets. ***Nature Methods***, 2011. 8(6):478-80. *(Evaluated as "Good" by "FACULTY of 1000")*

c. Simonis, N., Rual, J. F., Carvunis, A. R., …, **Yu, H.**, …, Hill, D. E., Tavernier, J., Roth, F. P., and Vidal, M., Empirically controlled mapping of the *Caenorhabditis elegans* protein-protein interactome network. ***Nature Methods***, 2009. 6(1):47-54.

d. Braun, P., Tasan, M., Dreze, M., Barrios-Rodiles, M., Lemmens, I., **Yu, H.**, Sahalie, J. M., Murray, R. R., Roncari, L., de Smet, A. S., Venkatesan, K., Rual, J. F., Vandenhaute, J., Cusick, M. E., Pawson, T., Hill, D. E., Tavernier, J., Wrana, J. L., Roth, F. P., and Vidal, M., An experimentally derived confidence score for binary protein-protein interactions*.* ***Nature Methods***, 2009. 6(1):91-7.

Complete List of Published Work in MyBibliography:
http://www.ncbi.nlm.nih.gov/sites/myncbi/haiyuan.yu.1/bibliography/43158364/public/?sort=date&direction=descending

D. Research Support

Ongoing Research Support

1 R01 GM097358 05/01/2012 – 01/31/2017 (in no cost extension) 2.5 calendar months
NIH/NIGMS
Role: Principal Investigator
Title: Towards a comprehensive protein interactome network in *Schizosaccharomyces pombe*

1 R01 GM104424 09/01/2013 – 06/30/2017 (in no cost extension) 3 calendar month
NIH/NIGMS
Role: MPI with Steven Lipkin
Title: Using Protein Interactome Networks to Understand the Functional Role of Coding DNA Sequence Variants

1 R01 HG008126 07/01/2016 – 06/30/2019 1 calendar months
NIH/NHGRI
Role: Co-investigator (PI: *Mark Gerstein* at Yale)
Title: Prioritizing rare variants associated with cancer using non-coding annotation

1 UM1 HG009393 02/01/2017 – 01/31/2021 2 calendar months
NIH/NHGRI
Role: MPI with John Lis
Title: High-throughput functional characterization of human enhancers

Overlap

None