Next Generation Protein Interactomes for Plant Systems Biology and Biomass Feedstocks Research

Mary Galli,[1] Shelly Wanamaker[1] and **Joseph R. Ecker[1]***(ecker@salk.edu)

[1]Salk Institute for Biological Studies La Jolla, California

Project goals: Proteins execute most biological processes, but rarely do so alone. Therefore, large-scale efforts to interrogate the composition of the macromolecular complexes in which these proteins operate and their dynamic behavior are essential for a genome-wide understanding of cellular networks. To facilitate the identification of such interactions on a proteome-wide scale in plants, our project aims to create a high throughput protein-protein interaction system in which interacting pairs can be captured in the form of a nucleic acid readout and identified using next-generation sequencing.

Large-scale protein interaction maps are currently available for several model organisms including yeast, worm, fly, and human. In addition, we have recently reported the first large-scale plant interactome containing ~6200 novel interactions for nearly 3000 Arabidopsis proteins. While this effort resulted in the generation of the largest plant interaction dataset to date, we estimate it covers only ~2% of all protein interactions that exist in Arabidopsis. This observation is consistent with previous mapping efforts in other model organisms, which also suffered from low coverage. Therefore new technologies with the capability to significantly expand assay coverage are critical to enable comprehensive interactome mapping. This is especially pertinent for plants (in particular non-model plant systems with biomass-production potential) which contain a significantly higher number of protein coding genes relative to other eukaryotes.

We aim to develop an experimental approach that harnesses the unparalleled depth of next-generation sequencing and allows for the efficient and cost-effective interrogation of all pairwise interactions for all proteins encoded in a genome. To accomplish this goal, we have proposed to adapt the yeast 2-hybrid (Y2H) system, currently widely used in both small and large-scale protein interaction studies, to a format that is amenable with a high-throughput sequencing based output. Specifically, our strategy exploits the ability of CRE recombinase from the bacteriophage P1 to catalyze the physical linkage of two pieces of lox-containing DNA in vivo. We have introduced non-revertible lox sites into a set of Gateway-compatible Y2H vectors containing bait and prey ORF-fusion cassettes. Upon transformation into an inducible CRE-containing yeast strain, an interaction between bait and prey fusion proteins results in the reconstitution of a functional yeast GAL4 transcription factor which then drives expression of a growth-based reporter gene while simultaneously driving expression of CRE recombinase. The expression of CRE triggers the physical linkage of the bait and prey plasmids that encode the interacting proteins. Because the interacting pairs are locked together, large numbers of colonies that survive selection can be pooled *en masse* and processed as a single sample using next-generation sequencing technologies to identify the interacting pairs. A

simple shearing strategy of the pooled, fused interacting DNAs followed by ligation with Illumina adaptors simplifies the sequencing library preparation process. Our current Illumina HiSeq platform yields over 600 Gigabases of DNA sequencing output per run and thus provides sufficient coverage for the identification of millions of interaction pairs.

Following initial proof of concept experiments using a preexisting set of Arabidopsis positive and random reference set clones, we plan to re-query a subset of the Arabidopsis Interactome pairwise space. This will allow us to obtain a comparative estimate of the increase in assay coverage and quality of our next-generation interactome approach. Ultimately, such a system has the potential to move beyond traditional time-consuming pairwise tests using well-defined ORF collections and accommodate library vs. library screens using 'cDNA shatter libraries'. Flexible formats such as this will pave the way for the generation of interactome maps in organisms, such as potential biomass fuel stocks, for which no pre-existing ORF collections are available.

We gratefully acknowledge the U.S. DOE Office of Biological and Environmental Research in the U.S. DOE Office of Science for funding this project; DOE-DE-SC0007078.