Title

High-resolution molecular networks from novel 'omics' approaches elucidate survival strategies in organisms from land to sea

Permalink

https://escholarship.org/uc/item/0491n31k

Author

Trigg, Shelly A.

Publication Date

2018

Supplemental Material

https://escholarship.org/uc/item/0491n31k#supplemental

UNIVERSITY OF CALIFORNIA SAN DIEGO

High-resolution molecular networks from novel ‘omics’ approaches elucidate survival strategies in organisms from land to sea

A dissertation submitted in partial satisfaction of the requirements for the degree Doctor of Philosophy

in

Biology

by

Shelly Ann Trigg

Committee in charge:

Professor Joseph Ecker, Chair
Professor Joanne Chory
Professor Theresa Gaasterland
Professor Trey Ideker
Professor Jose Pruneda-Paz

2018

The Dissertation of Shelly Ann Trigg is approved, and is acceptable in quality and form for publication on microfilm and electronically:

Chair

University of California San Diego

2018

DEDICATION

To my brother, Scott Mark Wanamaker.

To my husband Austin Engel Trigg.

To my parents Mark and Twila Wanamaker.

To my family: BFS, JAS, AC, GEP, DMT, KET, and RMT

To my undergraduate advisor JRC, and former professors.

To my students.

To the American people, that they may always have easy access to current and past publicly funded scientific research reports.

EPIGRAPH

Study the science of art.

Study the art of science.

Develop your senses – especially learn how to see.

Realize that everything connects to everything else.

Principles for the Development of a Complete Mind

TABLE OF CONTENTS

LIST OF SUPPLEMENTAL FILES

LIST OF FIGURES AND TABLES

ACKNOWLEDGEMENTS

I would first like to acknowledge Joseph Ecker for giving me the opportunity to work in his Genomic Analysis Lab over the past 6 years. He has provided a unique work environment and importantly has taught me to fearlessly think in a global scientific perspective. I thank my committee members Terry Gaasterland, Joanne Chory, Trey Ideker, and Jose Pruneda-Paz for their continued guidance and support throughout the pursuit of my doctoral degree.

Joe provided access to plethora of technological resources and a team talented researchers, which made my doctoral work possible. I would specifically like to thank lab members Renee Garza, Haili Song, Joseph Nery, Anna Bartlett, Rosa Castanon, Cesar Barragan, Carol Huang, Zhuzhu Zhang, and former lab members Mary Galli, Andrew MacWilliams, Ronan O'Malley, and Florian Jupe for acting as a strong support system inside and outside of the lab. The Salk Plant Biology department has been instrumental in facilitating discussions and stimulating new ideas. Lab administrators Kim Emerson and Nancy Benson have helped immensely in planning, logistics, and nailing deadlines. Carol, Ronan, Zhuzhu, Joe Nery, Lantian Gai, Huaming Chen, Joseph Feeney, student Joaquin Reyna, and Saket Navlahka have served as wonderful teachers and support of my learning bioinformatics without any formal computational training. Renee Garza and Andrew MacWilliams assisted in experiments, Joe Nery and Rosa Castanon performed all Illumina Sequencing runs.

This work has been funded by the NSF Graduate Research Fellowship Program and the Mary K. Chapman Foundation to Shelly Trigg, and by an NSF grant and Department of Energy grant to Joseph Ecker.

I would like to acknowledge Jens Lykke-Anderson, Suzi Harlow, Marifel Alfaro, Tom Tomp, and Cathy Pugh at the UCSD Division of Biology for their impeccable support of graduate students under all circumstances. The Salk and UCSD communities for constantly inspiring scientific discovery and thought.

I would lastly like to thank my family and friends who have unconditionally understood my sacrifice of time with them to pursue my ambitions in science.

Chapter 2, in full, is a reformatted reprint of the material as it appears in the August 2017 issue of the journal *Nature Methods*. Trigg SA, Garza RM, MacWilliams A, Nery JR, Bartlett A, Castanon R, Goubil A, Feeney J, O'Malley R, Huang SC, Zhang ZZ, Galli M, and Ecker JR. 2017. CrY2H-seq: a massively multiplexed assay for deep-coverage interactome mapping. *Nature Methods*, August 2017. The dissertation author was the primary investigator and author of this material.

Chapter 3, in full, consists of the following manuscript in preparation for submission to the journal *Scientific Reports* in the format of an article.

Shelly A. Trigg, Paul McElhany, Michael Maher, Danielle Perez, D. Shallin Busch, and Krista M. Nichols. 2018. Uncovering mechanisms of global ocean change effects on the Dungeness crab (*Cancer magister*) through metabolomics analysis. The dissertation author was the primary investigator and author of this material.

VITA

SHELLY ANN TRIGG

swanamak@ucsd.edu

EDUCATION

2018	Ph.D., Biology, University of California San Diego
2010	B.S., Biochemistry, Magna Cum Laude. Simmons College, Boston, Massachusetts
2005	High School Diploma, Bedford High School, Massachusetts

RESEARCH EXPERIENCE

2017-2018	NSF Graduate Research Internship Program Fellow, Genetics and Evolution section of the Conservation Biology Division, NOAA Northwest Fisheries Science Center, Seattle, WA
2016	Visiting Scientist, Center for Cell Circuits, Broad Institute, Cambridge, MA
2013-2017	Graduate Student Researcher, Ecker Genomic Analysis Lab, Salk Institute, CA
2011-2013	Research Assistant II, Ecker Genomic Analysis Lab, Salk Institute, CA
2010-2011	Laboratory Technician, Center for Cancer Systems Biology, Dana Farber Cancer Institute, Boston
2009-2010	Undergraduate Researcher, Roecklein-Canfield Lab, Chemistry Department, Simmons College, Boston

FELLOWSHIPS/HONORS/AWARDS

2014-2018	NSF Graduate Research Fellowship
2016	Excellence in Teaching Award. Division of Biological Sciences, University of California San Diego

2015 Mary K. Chapman Foundation Graduate Student award

2015 People’s Choice Award. Salk Institute annual retreat poster presentation

2014 Rookie of the Year Award. Association for Women in Science, San Diego. Awarded to new chapter member who demonstrated a commitment to advancing the goals of the organization by volunteering efforts, providing ideas, and promoting the organization outside of AWIS

2010 Allen Douglass Bliss Award in Chemistry for academic achievement and promise in the field. Simmons College, Boston

2009 Sigma Xi Biological Honors Society.

2008 Tri Beta Biological Honors Society.

MANUSCRIPTS IN PREPARATION

Trigg SA, McElhany P, Maher M, Perez D, Busch DS, and Nichols KM. Uncovering mechanisms of global ocean change in Dungeness crab (*Cancer magister*) through metabolomics analysis

PUBLICATIONS

Trigg SA, Garza RM, MacWilliams A, Nery JR, Bartlett A, Castanon R, Goubil A, Feeney J, O’Malley R, Huang SC, Zhang ZZ, Galli M, and Ecker JR. 2017. CrY2H-seq: a massively multiplexed assay for deep-coverage interactome mapping. Nature Methods. 14(8):819-825

Trigg SA, Garza RM, MacWilliams A, Nery JR, Bartlett A, Castanon R, Goubil A, Feeney J, O’Malley R, Huang SC, Zhang ZZ, Galli M, and Ecker JR. 2017. CrY2H-seq interactome screening. Protocol Exchange. doi:10.1038/protex.2017.058

Yang X, [15 others], Trigg SA, [20 others], and Vidal M. 2015. Widespread expansion of protein interaction capabilities by alternative splicing. Cell. 164(4):805-817

Rolland T, [53 others], Trigg SA, [14 others], and Vidal M. 2014. A proteome-scale map of the human interactome network. Cell. 159(5):1212-1226

Corominas R, [8 others], Trigg SA, [18 others], Vidal M, and Iakoucheva LM. 2014. Protein interaction network of alternatively spliced isoforms from brain links genetic risk factors for autism. Nature Communications. 5:3650

Rozenblatt-Rosen O, [36 others], Wanamaker S, [13 others], and Vidal M. 2012. Interpreting cancer genomes using systematic host network perturbations by tumour virus proteins. Nature. 487(7408):491-495

PRESENTATIONS/MEETINGS

Shelly A. Trigg, Maher M, Perez D, Busch DS, McElhany P, Nichols KM. 2018. Uncovering mechanisms of global ocean change effects on Dungeness crab through metabolomics analysis. NOAA NWFSC 6th Science Symposium. Seattle, WA

Shelly A. Trigg, Maher M, Perez D, Busch DS, McElhany P, Nichols KM. 2018. Uncovering mechanisms of global ocean change effects on Dungeness crab through metabolomics analysis. Ocean Sciences Meeting, Portland, OR

Shelly A. Trigg, Garza RM, MacWilliams A, Nery JR, Bartlett A, Castanon R, Goubil A, Feeney J, O'Malley R, Huang SC, Zhang ZZ, Galli M, and Ecker JR. 2016. ProCREate: A new assay for generating large-scale protein-protein interaction maps. HHMI Scientific Meeting, Janelia Campus, VA

Shelly A. Trigg, RM Garza, A MacWilliams, JR Nery, J Reyna, A Bartlett, R Castanon, A Goubil, J Feeney, R O'Malley, CS Huang, ZZ Zhang, M Galli, and JR Ecker (Sept 2016) ProCREate: A New Assay for Generation of Large-Scale Protein-Protein Interaction Maps, UCSD Biology annual retreat, Lake Arrowhead, CA

Shelly A. Trigg, RM Garza, A MacWilliams, JR Nery, J Reyna, A Bartlett, R Castanon, A Goubil, J Feeney, R O'Malley, CS Huang, ZZ Zhang, M Galli, and JR Ecker (Mar 2016) Next-Generation Protein Interactomes for Plant Systems Biology and Biomass Feedstocks Research, USDA-DOE Plant Feedstock Genomics for Bioenergy Meeting, Tyson's Corner, VA

Shelly A. Trigg, RM Garza, A MacWilliams, JR Nery, J Reyna, A Bartlett, R Castanon, A Goubil, J Feeney, R O'Malley, CS Huang, ZZ Zhang, M Galli, and JR Ecker (October 2015) The ProCREate interactome mapping technique reveals crosstalk and directionality between Arabidopsis regulatory networks, Salk Science at the Seaside, La Jolla, CA

Shelly A. Trigg, RM Garza, A MacWilliams, JR Nery, J Reyna, A Bartlett, R Castanon, A Goubil, J Feeney, R O'Malley, CS Huang, ZZ Zhang, M Galli, and JR Ecker (Sept 2015) ProCREate: A New Assay for Generation of Large-

Scale Protein-Protein Interaction Maps, UCSD Biology annual retreat, Lake Arrowhead, CA

Shelly A. Trigg. (Aug 2015) Finding biological relevance in large-scale protein network studies, NSF Data Science Workshop, Seattle, WA

Shelly A. Trigg, RM Garza, A MacWilliams, JR Nery, J Reyna, A Bartlett, R Castanon, A Goubil, J Feeney, R O'Malley, CS Huang, M Galli, and JR Ecker (Feb 2015) Next-Generation Protein Interactomes for Plant Systems Biology and Biomass Feedstocks Research, USDA-DOE Plant Feedstock Genomics for Bioenergy Meeting, Tyson's Corner, VA

Shelly A. Trigg, A MacWilliams, JR Nery, A Bartlett, R Castanon, A Goubil, J Feeney, R O'Malley, M Galli, and JR Ecker (Feb 2014) Next-Generation Protein Interactomes for Plant Systems Biology and Biomass Feedstocks Research, USDA-DOE Plant Feedstock Genomics for Bioenergy Meeting, Arlington, VA

Shelly A. Trigg, J Feeney, JR Nery, A Bartlett, R Castanon, A Goubil, R O'Malley, M Galli, and JR Ecker (Feb 2013) Next-Generation Protein Interactomes for Plant Systems Biology and Biomass Feedstocks Research, USDA-DOE Plant Feedstock Genomics for Bioenergy Meeting, Bethesda, MD

ACADEMIC AND COMMUNITY SERVICE

2017 Head Instructional Assistant, Eukaryotic Gene Expression lecture course. University of California San Diego.

2015-2016 Instructional Assistant, Genomics Research Initiative laboratory course, 2016; Biochemical Techniques laboratory course, 2015; Student Instructor, Quantitative Biology graduate seminar. University of California San Diego

2008-2010 Teaching Assistant, Biochemistry II, 2010; Organic Chemistry I, 2009; Organic Chemistry II, 2008; General Chemistry, 2008-2010. Simmons College, Boston

2014-2016 Academic Mentorship, Division of Biological Sciences Peer Mentor Committee Co-chair, 2015-2016; Rotation student Eileen Gonzales, 2015-2016; Rotation student Natasha Slepak, 2014-2015. University of California San Diego.

2017 National Marine Sanctuaries ocean acidification multimedia toolkit development

2012-2017	Association for Women in Science Outreach Committee member
2014-2017	Salk Institute New Frontiers in Science Education curriculum developer
2014-2017	Reuben H. Fleet Science Center #2Scientists program volunteer scientist
2014-2015	Expanding Your Horizons AWIS San Diego presenter
2014	Albert Einstein Academy Family Science Night AWIS San Diego presenter
2013	Greater San Diego Science Festival Expo Day AWIS San Diego presenter
2013-2014	Greater San Diego Science Festival AWIS San Diego poster judge
2016	Explore Salk Science Day lab tour guide and research booth expert
2014-2018	Biological Sciences annual research showcase judge. University of California San Diego

ABSTRACT OF THE DISSERTATION

High-resolution molecular networks from novel ‘omics’ approaches elucidate survival strategies in organisms from land to sea

by

Shelly Ann Trigg

Doctor of Philosophy in Biology

University of California San Diego, 2018

Professor Joseph R. Ecker, Chair

Molecular networks drive nearly all cellular processes. With the advent of omics technologies involving next-generation sequencing and mass spectrometry, we have started to uncover highly complex gene, protein, and biochemical networks that underlie survival mechanisms like growth and stress tolerance.

However, the study of cell wide of protein-protein interactions that importantly link genomic, transcriptomic, and proteomic data to cellular activity has been hindered by technical limitations, and the need to survey and track billions of possible combinatorial protein interactions. Moreover, many omics technologies have not yet been developed or applied to non-model organisms, particularly to ecologically and economically important species that may be negatively impacted by climate change. To address these shortcomings, I developed a modified yeast two-hybrid technology called CrY2H-seq that enables massively multiplexed protein interaction screening through a Cre-lox reporter and next generation sequencing, and demonstrated its applications for generating interactome resources by screening a comprehensive set of Arabidopsis transcription factors. Lastly, I applied metabolomics analysis to investigate how ocean acidification might impact the Dungeness crab. I found transcription factor families preferentially interact with others, and relationships among families supported by recent independent studies may drive mechanisms underlying reproductive development and hormone signaling. I also found that metabolomes of developing Dungeness crab show treatment specific responses to low oxygen and low pH seawater treatments. Taken all

together, the new biological insight gained from these novel omics approaches can valuably be used to inform targeted future experiments aimed at optimizing crop cultivation and predicting how economically and economically important species might response to future environmental stress.

CHAPTER 1:

NEXT-GENERATION MODIFICATIONS OF THE YEAST TWO-HYBRID SYSTEM TO ADVANCE PROTEIN-PROTEIN INTERACTION DATA COLLECTION

ABSTRACT

Knowledge of proteome-wide protein-protein interaction (PPI) networks, or interactomes, that promote robust growth or that are perturbed by environmental stress could progress strategies for improving crop cultivation and conservation management. However, current technologies for obtaining interactome data are not suitable for non-model organisms because of time, cost, and sensitivity constraints. To better resolve interactomes for both model and non-model organisms, modifications to the yeast two-hybrid system were explored. The addition of a new Cre recombinase reporter gene with mutant lox sequence-containing expression plasmids enabled one-pot screening of complex libraries, and massively paralleled sequencing to detect intracellularly-created DNA identifiers of interaction pairs. Compared to the gold standard yeast two-hybrid assay, this modified system showed improved sensitivity and reproducibility in the screening of an ORF collection previously used for assay benchmarking. This new technological development facilitates obtaining experimentally derived interactome data, with promising applications for non-model organism studies.

INTRODUCTION

Most biological processes involve an enormous number of different proteins functioning in complex and dynamic networks. Mapping these networks in the form of an interactome, a complete list of physical interactions mediated by all proteins of an organism, provides a more comprehensive perspective of the overall physical and functional cellular landscape than the traditional reductionist approach[1,2]. Interactomes show functions for uncharacterized genes, inconspicuous crosstalk between inter and intracellular networks, network evolution, and mechanisms of disease manifestation[3–6]. More specifically, interactomes reveal highly connected proteins, involved in multiple pathways, which are often the targets of pathogens[7–9]. Observing how networks become re-wired when exposed to stimuli can elucidate vulnerable network regions and how fragile or flexible certain pathways are. Understanding these system characteristics will help improve the way disease is treated, our understanding of evolution, and perhaps approaches for optimizing crop productivity[10–14].

Although interactomes reveal global and profound connections, the interactomics field has historically been hindered by insufficient technology to study the sheer volume of proteome-wide interactions. For the model plant species *Arabidopsis thaliana*, screening the estimated 35,000 distinct protein coding genes easily translates into screening over 1.2×10^9 interactions, considering isoforms and post-translational modifications. An ideal methodology would measure all interactions in real-time as they are happening in the cell, however technology has not yet evolved to simultaneously image and identify a complex pool of interacting protein pairs. The convention is typically to express proteins from cloned coding sequences in expression plasmids, and assay for one set of interactions at a time. The challenge to develop a systematic, cost-effective, time-efficient, non-targeted approach has led to two widely-used technologies for large-scale interactome mapping: high-throughput yeast two-hybrid (HT-Y2H) and high-throughput affinity purification/mass spectroscopy (HT-AP/MS) screening.

While HT-Y2H and HT-AP/MS are complementary approaches that together have led to uncovering large interactome space[15], they address different questions about protein interactions. HT-AP/MS characterizes proteins that co-purify with an affinity-tagged protein from lysate, yielding information about complexes, but not about direct interactions between two proteins. HT-Y2H assays for direct, binary protein interactions and measures phenotypic changes in yeast that result from the interaction of two proteins being expressed in each cell. More specifically, one gene is expressed as a bait protein fused to a DNA-binding domain and the other gene is expressed as a prey protein fused to an activation domain. Thus, upon protein interaction, a functional transcription factor is reconstituted and can only then drive the expression of a reporter gene.

Matrix formatting has dramatically improved the throughput of these techniques. Some of the latest HT-AP/MS technology has made it possible to screen 96 samples per hour[16], however this platform requires tens of thousands of costly assay plates to cover a proteome. HT-Y2H screening strategies include either individual pairwise screening where a single bait protein is screened against a single prey protein, or library screening where a single bait protein is screened against a pool of prey proteins. Both strategies rely on matrix format to maintain the identity of at least one of the proteins being screened. The time and cost constraints that this amounts to is exemplified in the generation of the *Arabidopsis* Interactome 1 (AI-1), where all pairwise combinations of 8,000 proteins were tested, taking upwards of 5 years and $8,000,000 to complete[6]. The 6,205 AI-1 interactions identified correspond to about only 2% of the estimated complete *Arabidopsis* interactome. The low coverage indicates that despite thorough screening, the number of screens was insufficient for this Y2H gold standard assay to achieve saturation. This problem has historically been a major source for little overlap between HT-Y2H-derived interactome datasets[17].

To cover the depth of proteome-wide interactions, requires a massively paralleled strategy similar to next-generation sequencing that can identify over a billion of different sequences in one sequencing run. Efforts to couple next-generation sequencing with yeast two-hybrid experiments have led to the development of many assay variations, for instance, Stitch-seq[18] where coding sequences are individually amplified from positive colonies and stitched together by a subsequent PCR. While Stich-seq largely increased rate of protein-protein interaction identification, the requirement of a matrix-based format for tracking interacting partners during screening persisted. Soon after, another variation was developed called QIS-seq, where one bait protein is screened against a library of prey proteins[19]. Because the identity of the one bait is known, only prey proteins need to be identified and can be pooled and sequenced *en masse*. QIS-seq solved the problem of keeping individual colonies isolated, but still required tens of thousands of assays to screen different baits. If the identity of both interacting partners could be maintained despite pooling, it would be possible to rapidly screen a proteome in Y2H. In 2007, a proof of concept rendition of this idea exploiting cre-lox unidirection recombination, called BI (binary interaction)-tag Y2H[20], showed small bait and prey cDNA libraries could be screened *en masse* by inducing a physical linkage between corresponding coding sequences of interacting proteins in cells surviving selection. However, limited overlap between traditional Y2H analysis and BI-tag Y2H hindered this concept from further development. Given the ease and cost-effectiveness of this strategy, an optimized variation of this technology could progress Systems Biology to a new level by enabling the rapid generation of interactomes.

The ability to rapidly screen whole expression clone libraries in one tube would greatly facilitate interactome data collection, and finally enable near complete interactomes to be generated[21]. Larger search spaces could be interrogated, making screens untargeted and datasets less biased. Unlimited replicate screens could easily be carried out, allowing for interactions to be screened to saturation. Next-generation sequencing for PPI detection would eliminate the variability from scoring by visual phenotyping, leading to more consistent interaction scoring and higher sensitivity for transient interactions. The ability to sensitively quantitate PPIs from linked coding sequences could even potentially measure interaction affinity. Combining existing technology for creating comprehensive full length cDNA yeast expression libraries[22,23] with cre-lox unidirectional recombination for intracellular PPI tracking would lead to a revolutionary advancement in the interactomics field, and provide a platform from which to screen cDNA library x cDNA library. This would open new doors enabling interactomes to be generated for nearly any species, and comparative interactomics where network changes could be observed between interactomes generated from different tissues ecotypes, treatments, disease states, development stages, and beyond.

Towards achieving an improved interactome mapping technology, we built a Cre recombinase reporter gene-based yeast two-hybrid approach such that interacting proteins trigger the endogenous expression of Cre recombinase. Cre activity then irreversibly recombines mutant lox sites that flank protein coding sequences on bait and prey expression plasmids, forming a physical linkage between protein coding sequences (**Fig. 1**).

RESULTS

We chose to use the yeast two-hybrid mating strains Y8800 (*MAT*a) and Y8930 (*MAT*α), and expression plasmids pDEST-AD and pDEST DB, which have all previously been described in detail[18,24]. These strains have become the gold standard strains[5,6,9,25–27] because they show a heightened sensitivity to histidine deprivation and the plasmids are low copy which protects the cells from toxic effects from overabundance of foreign protein. Additionally, they allow plasmid shuffling for identifying self-activating DB fusion proteins. We modified the plasmids to contain mutant lox sequences downstream the Gateway cloning site. Cre recombinase was integrated into the *GAL7*::LacZ reporter gene locus of Y8930 by homologous recombination, replacing the LacZ reporter.

For a proof-of-concept test, the modified Y8930 (now called CRY8930) and its mate strain Y8800 were transformed with lox-modified bait and prey plasmids that harbored the well-characterized BZIP63 and BZIP53 plant transcription factor protein-protein interaction, respectively, and subsequently screened. As a negative control, the original Y8930 strain harboring the same BZIP63-containing bait plasmid was screened with Y8800 in parallel. Y8800/CRY8930 colonies surviving *HIS3* reporter gene selection gave rise to amplicons (**Fig. 2**) containing BZIP53 and BZIP63 coding sequences linked together in a tail-to-tail manner (**Fig 1d**). The Y8800/Y8930 colonies surviving selection did not give rise to these amplicons. Unmodified plasmids in either combination of strains also did not give rise to these amplicons (**Fig 2**).

To benchmark this modified Y2H, we screened the ORF set AI-1_{REPEAT} that had previously been thoroughly screened in Y2H iterations and for which high confidence PPI data existed[6]. These ORFs were cherry picked from the AtORFeome 2.0 collection, individually Gateway cloned into pDEST AD and DB lox plasmids and transformed into yeast strains in matrix format, as described previously[24]. In the yeast two-hybrid system, false positives may arise from self-activating proteins; that is a protein capable of inducing reporter gene expression on its own[24]. To identify these prior to library pooling and *en masse* screening, DB proteins were screened 1x1 against an pDEST-AD empty plasmid strain, and clones identified as self-activating were not included in library pooling. The remaining non-self-activating clones were finally pooled from individually grown yeast clones in approximately equal amounts, and frozen library stock aliquots were prepared from the pooled libraries.

Three replicate screens of AI-1$_{\text{REPEAT}}$ were carried out as pictured in **Fig. 3**. Briefly, bait and prey libraries were mated *en masse* and selected for on synthetic complete media lacking histidine and supplemented with 1mM 3AT. AD and DB primers were used in a multi-template PCR to amplify Cre-recombined coding sequences from a bulk yeast plasmid prep. Amplicons were fragmented at 300-500bp and 3 sequencing libraries (1 for each replicate) were prepared following the Illumina Truseq library prep protocol. We named this assay CrY2H-seq, Cre reporter-mediated yeast-two hybrid using next-generation sequencing.

Because self-activating proteins can also result from mutation or truncation during the screening process, an internal control pDEST-AD empty plasmid strain was spiked in to the library. This way, any coding sequence found fused to pDEST AD-empty plasmid could be removed computationally from the PPI data. With about 5,000,000 possible pairs maximally arising from a screening space of 1,500 bait x 2,900 prey clones, we aimed for a low 4X coverage and a high 16X coverage. The combined sequencing data from each replicate amounted to 277,000,000 reads and mapped to 36,261 protein pairs. After bioinformatically removing all pairs with self-activating proteins that showed a protein coding sequence fused to pDEST AD-empty plasmid sequence, 3,500 proteins pairs remained.

Assay reproducibility

CrY2H-seq showed increased overlap between replicate assays compared to the overlap between AI-1 replicate assay, and showed similar overlap between replicates to the overlap between intralaboratory HT-AP/MS replicate assays[28] (**Fig. 4a**). To assess reproducibility of PPIs detected by CrY2H-seq in a 1x1 pairwise traditional Y2H, a set of PPIs ranging in sequence abundance was retested and yeast colony spot growth were scored. Interactions retested positive at a higher rate when they overlapped across replicates or when they were in abundance higher than 5 reads in one replicate (**Fig. 4b**). CrY2H-seq interactions overlap with 30% $\mathrm{AI}_{\mathrm{REPEAT}}$ interactions reported in AI-1, an increase to the 9% overlap between high confidence yeast two-hybrid datasets observed in the past[29]. This suggests that one CrY2H-seq screen can more comprehensively capture high quality PPIs.

Assay sensitivity

Considering the interaction detection rate (5-10 PPIs/10,000 PPIs tested) observed across interactome datasets generated by HT-Y2H screens[6], we expected about 2,500 interactions to result from screening the $\mathrm{AI}_{\mathrm{REPEAT}}$ ORF set in CrY2H-seq. Three replicate CrY2H-seq assays identified 3,500 interactions, which is more than three times more than those identified in AI-1. Despite this increase in overall interaction detection (**Fig. 5**), CrY2H-seq does not show a plateauing pattern after three replicates, indicating the entire screen was not saturated and potentially more interactions could be detected with more screens. To more directly measure CrY2H-seq saturation, a set of PPIs reported in AI-1 and not identified in CrY2H-seq was also tested in a pairwise 1x1 Y2H test. These interactions tested positively at a rate of 70% (data not shown), which suggests that more CrY2H-seq replicate screens are necessary to see these interactions. Sub-saturation is most likely fully attributable to low mating efficiency (~2%), which limited the number of PPI combinations tested in each replicate. Because mating efficiencies have only been reported to be optimized up to 17%[30], exhaustive screening through increasing the number replicate assays and oversampling likely better strategies for ensuring all combinations are screened. However, number of CrY2H-seq replicate assays necessary to reach saturation remains to be determined.

To estimate assay sensitivity in terms of the ratio of true positives to false positives detected, a subset of well-characterized positive reference set (AtPRSv1)[6] was internally included in the screen. As another benchmark, the detection of literature-cited interactions was used to estimate a true positive detection rate. To estimate false negatives, because it is impossible to generate a set of negative interacting pairs with absolute confidence[24], a set of interactions was chosen at random from a list of all possible interactions and called random reference set. CrY2H-seq showed a higher detection rate of well-characterized interactions and a lower detection rate of random interactions compared to the AI-1 gold standard assays (**Fig. 6**). This suggests CrY2H-seq is a more sensitive assay, likely due to using sequencing for interaction detection.

DISCUSSION

The feasibility of the *en masse* recombination-based Y2H approach was tested with a thoroughly screened subset of the Ecker Lab *Arabidopsis* ORF collection, $\mathrm{AI}_{\mathrm{REPEAT}}$. After comparing detected PPIs with previously published datasets, it is apparent CrY2H-seq captures more interactions at higher sensitivity and reproducibility than current Y2H systems. Because CrY2H-seq screening was not saturated after 3 replicate screens, additional optimizations must be made to reach saturation and total number of replicate screens must be determined. Strategies for optimizing mating efficiency, including pre-growing cells in low pH media, mating in media supplemented with PEG, concentrating cells on filters for mating, and optimizing *MAT*a to *MAT*α ratio[30], could be implemented. Towards determining the number of replicate assays necessary for saturation, a saturation curve could be generated based on the decreasing rate of new interactions identified in each of the three replicate screen and based on the estimated number of PPIs CrY2H-seq can detect. Finally, an additional optimization that could reduce the sequencing space needed for identifying all PPIs in one screen could be enriching for DNA fragments containing double mutant lox sequences using a target enrichment strategy[31]. Overall, CrY2H-seq shows promise for *en masse* protein interaction screening capabilities far beyond existing technologies.

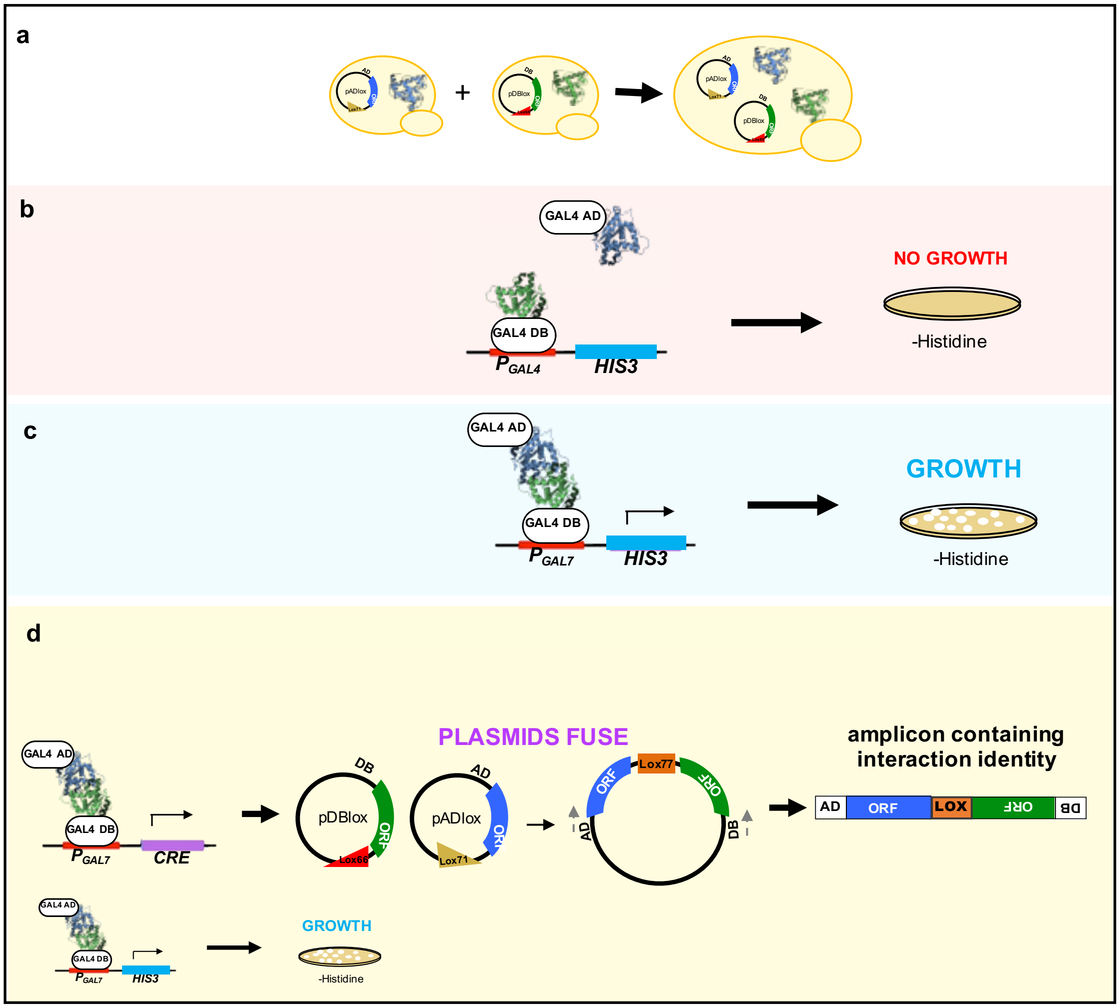

Figure 1.1. Cre reporter gene and mutant lox sequence modifications to yeast two-hybrid. (**a**) Summary of yeast two-hybrid technology. (**b**) Yeast two-hybrid outcome is no growth when no protein interaction occurs. (**c**) Yeast two-hybrid outcome is growth when a protein interaction does occur. (**d**) Cre reporter gene and mutant lox sequence modified yeast two-hybrid outcome is growth and intracellular, irreversible recombination of bait and prey plasmids. From the newly formed hybrid plasmid, activation domain (AD) and DNA binding (DB) domain specific primers (grey arrows) can the amplify coding sequences containing the protein interaction identity that can be sequenced.

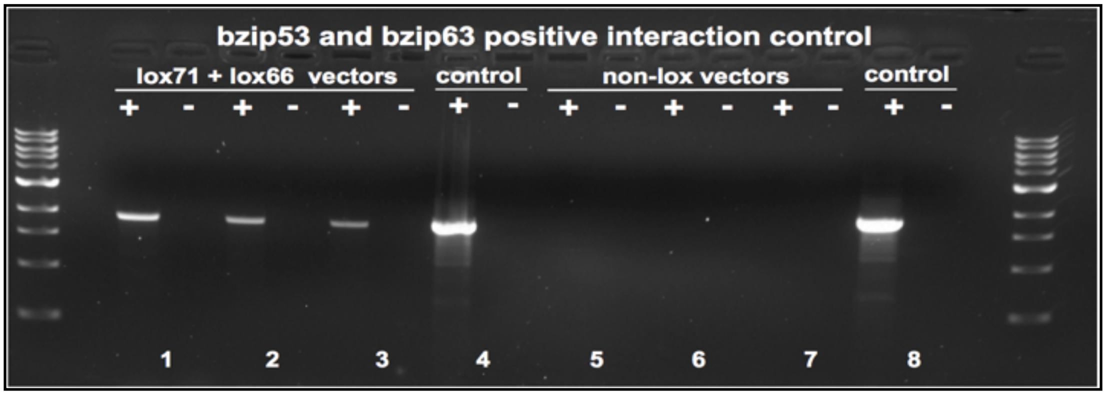

Figure 1.2. *In vivo* functionality pilot test of lox plasmids and Cre expressing yeast strain. Following interaction selection of positive control interactors BZIP53 and BZIP63 on synthetic complete media lacking histidine and supplemented with 1mM 3AT, Cre reporter activity was detected with AD and DB primers that only amplify a product in the Cre reporter yeast strain in plasmids with lox sequences. Samples 1-3 and 5-7 represent independent HIS+ colonies in + Cre and – Cre yeast strains; samples 4 and 8 are diluted *in vitro* Cre recombination reactions of lox plasmids with Cre (+) or without Cre (-).

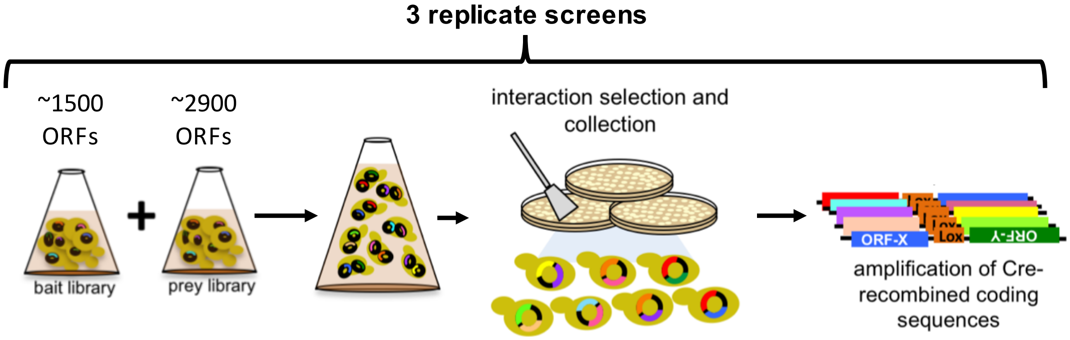
3 replicate screens
~1500 ORFs
~2900 ORFs
bait library
prey library
interaction selection and collection
ORF-X
Lox
ORF-Y
amplification of Cre-recombined coding sequences

Figure 1.3. Cre reporter modified yeast two-hybrid screening of the AI_{REPEAT} ORF collection. Complex bait and prey libraries were mated *en masse* and selected for on synthetic complete media lacking histidine and supplemented with 1mM 3AT. AD and DB primers were used in a multi-template PCR to amplify Cre-recombined coding sequences from a bulk yeast plasmid prep. Amplicons were fragmented at 300-500bp and 3 sequencing libraries (1 for each replicate) were prepared following the Illumina Truseq library prep protocol.

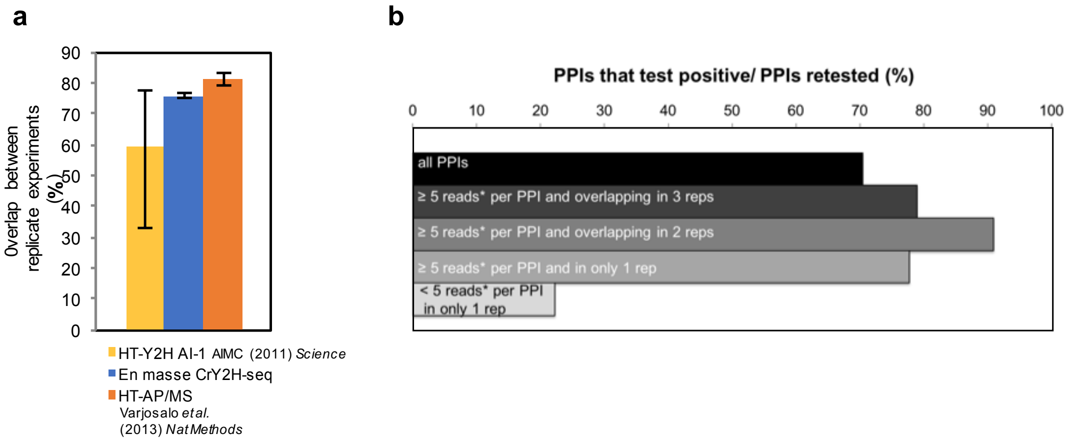

Figure 1.4. CrY2H-seq shows improved assay reproducibility. (**a**) shows increased overlap between replicate assays than AI-1 and similar overlap to interlaboratory AP/MS replicate assays. (**b**) Retest rate from 1x1 pairwise Y2H screen. Interactions identified in more than one replicate and interactions identified in one replicate with abundance higher than 5 reads retest positively at a higher rate than interactions identified in one replicate with abundance of 5 reads or less.

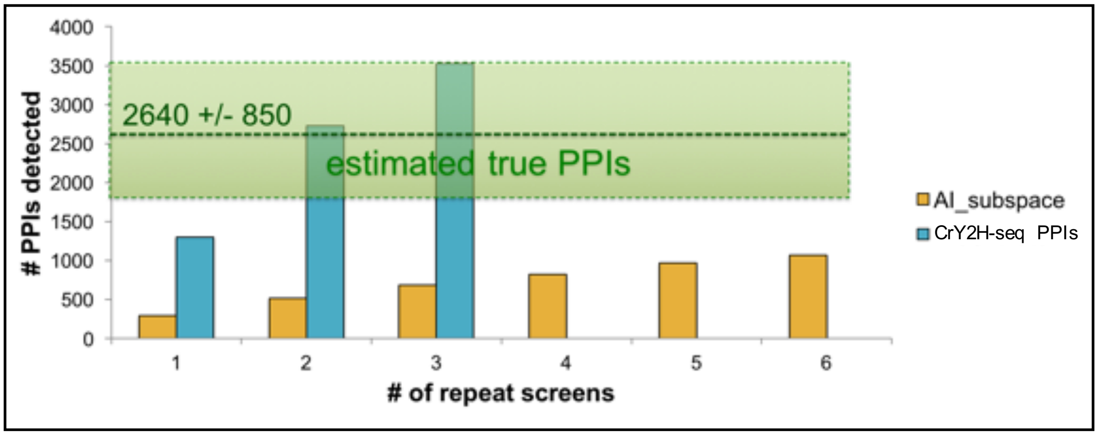

Figure 1.5. Saturation of HT-Y2H and CrY2H-seq assays compared. CrY2H-seq (teal) detects more interactions in fewer repeat screens, achieving the estimated true number of interactions after only 3 screens, but does not show a plateauing pattern that would indicate saturation has been reached. The HT-Y2H Gold Standard Assay (yellow) used in AI-1 (AIREPEAT; called AI_subspace here) does not achieve the estimated number of true interactions or screening saturation after 6 repeat screens.

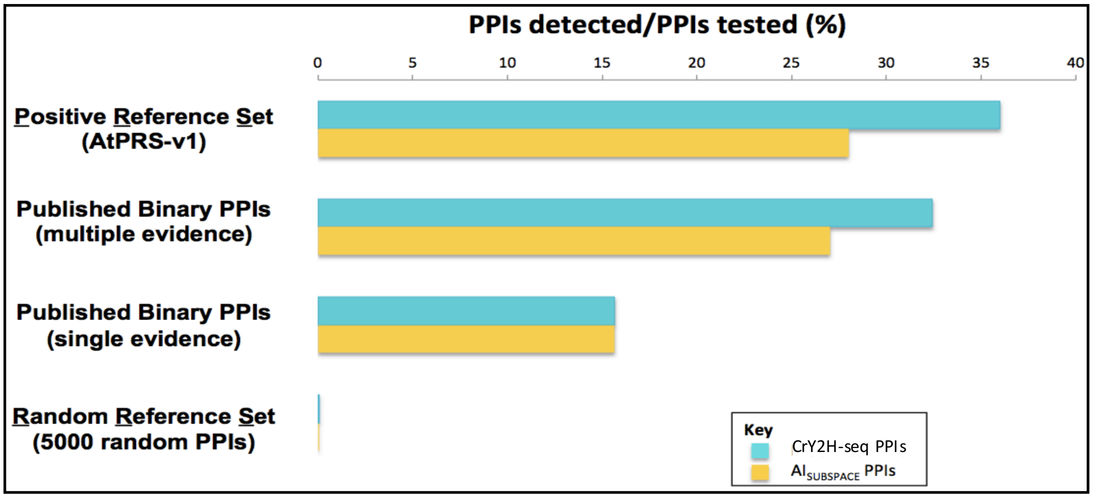

Figure 1.6. Accuracy of HT-Y2H and CrY2H-seq assays compared. CrY2H-seq (teal) shows an increased detection of published interactions than AI-1 (Gold).

REFERENCES

1. Vidal, M. Interactome modeling. *FEBS Letters* **579,** 1834–1838 (2005).

2. Ideker, T. & Krogan, N. J. Differential network biology. *Molecular Systems Biology* **8,** (2012).

3. Meneely, P. M. *Genetic Analysis: Genes, Genomes and Networks in Eukaryotes*. (Oxford University Press, 2014).

4. Klopffleisch, K., Phan, N., Augustin, K., Bayne, R. S., Booker, K. S., Botella, J. R., Carpita, N. C., Carr, T., Chen, J.-G., Cooke, T. R., Frick-Cheng, A., Friedman, E. J., Fulk, B., Hahn, M. G., Jiang, K., Jorda, L., Kruppe, L., Liu, C., Lorek, J., McCann, M. C., Molina, A., Moriyama, E. N., Mukhtar, M. S., Mudgil, Y., Pattathil, S., Schwarz, J., Seta, S., Tan, M., Temp, U., Trusov, Y., Urano, D., Welter, B., Yang, J., Panstruga, R., Uhrig, J. F. & Jones, A. M. Arabidopsis G-protein interactome reveals connections to cell wall carbohydrates and morphogenesis. *Mol. Syst. Biol.* **7,** 532–532 (2014).

5. Sahni, N., Yi, S., Taipale, M., Fuxman Bass, J. I., Coulombe-Huntington, J., Yang, F., Peng, J., Weile, J., Karras, G. I., Wang, Y., Kovács, I. A., Kamburov, A., Krykbaeva, I., Lam, M. H., Tucker, G., Khurana, V., Sharma, A., Liu, Y. Y., Yachie, N., Zhong, Q., Shen, Y., Palagi, A., San-Miguel, A., Fan, C., Balcha, D., Dricot, A., Jordan, D. M., Walsh, J. M., Shah, A. A., Yang, X., Stoyanova, A. K., Leighton, A., Calderwood, M. A., Jacob, Y., Cusick, M. E., Salehi-Ashtiani, K., Whitesell, L. J., Sunyaev, S., Berger, B., Barabási, A. L., Charloteaux, B., Hill, D. E., Hao, T., Roth, F. P., Xia, Y., Walhout, A. J. M., Lindquist, S. & Vidal, M. Widespread macromolecular interaction perturbations in human genetic disorders. *Cell* **161,** 647–660 (2015).

6. Arabidopsis Interactome Mapping Consortium. Evidence for Network Evolution in an Arabidopsis Interactome Map. *Science (80-.).* **333,** 601–607 (2011).

7. Weßling, R., Epple, P., Altmann, S., He, Y., Yang, L., Henz, S. R., McDonald, N., Wiley, K., Bader, K. C., Gläßer, C., Mukhtar, M. S., Haigis, S., Ghamsari, L., Stephens, A. E., Ecker, J. R., Vidal, M., Jones, J. D. G., Mayer, K. F. X., Ver Loren Van Themaat, E., Weigel, D., Schulze-Lefert, P., Dangl, J. L., Panstruga, R. & Braun, P. Convergent targeting of a common host protein-network by pathogen effectors from three kingdoms of life. *Cell Host Microbe* **16,** 364–375 (2014).

8. Mukhtar, M. S., Carvunis, A.-R., Dreze, M., Epple, P., Steinbrenner, J., Moore, J., Tasan, M., Galli, M., Hao, T., Nishimura, M. T., Pevzner, S.

J., Donovan, S. E., Ghamsari, L., Santhanam, B., Romero, V., Poulin, M. M., Gebreab, F., Gutierrez, B. J., Tam, S., Monachello, D., Boxem, M., Harbort, C. J., Mcdonald, N., Gai, L., Chen, H., He, Y., Vandenhaute, J., Roth, F. P., Hill, D. E., Ecker, J. R., Vidal, M., Beynon, J., Braun, P., Dangl, J. L., Union, E., Consortium, E., Vandenhaute, J., Roth, F. P., Hill, D. E., Ecker, J. R., Vidal, M. & Beynon, J. Independently Evolved Virulence Effectors Converge onto Hubs in a Plant Immune System Network. *Science (80-.).* **333,** 596–601 (2011).

9. Rozenblatt-Rosen, O., Deo, R. C., Padi, M., Adelmant, G., Calderwood, M. A., Rolland, T., Grace, M., Dricot, A., Askenazi, M., Tavares, M., Pevzner, S. J., Abderazzaq, F., Byrdsong, D., Carvunis, A. R., Chen, A. A., Cheng, J., Correll, M., Duarte, M., Fan, C., Feltkamp, M. C., Ficarro, S. B., Franchi, R., Garg, B. K., Gulbahce, N., Hao, T., Holthaus, A. M., James, R., Korkhin, A., Litovchick, L., Mar, J. C., Pak, T. R., Rabello, S., Rubio, R., Shen, Y., Singh, S., Spangle, J. M., Tasan, M., Wanamaker, S., Webber, J. T., Roecklein-Canfield, J., Johannsen, E., Barabási, A. L., Beroukhim, R., Kieff, E., Cusick, M. E., Hill, D. E., Münger, K., Marto, J. A., Quackenbush, J., Roth, F. P., Decaprio, J. A. & Vidal, M. Interpreting cancer genomes using systematic host network perturbations by tumour virus proteins. *Nature* **487,** 491–495 (2012).

10. Rusnati, M. & Presta, M. Angiogenic growth factors interactome and drug discovery: The contribution of surface plasmon resonance. *Cytokine Growth Factor Rev.* **26,** 293–310 (2015).

11. De Las Rivas, J. & Prieto, C. Protein interactions: Mapping interactome networks to support drug target discovery and selection. *Methods in Molecular Biology* **910,** 279–296 (2012).

12. Vidal, M., Cusick, M. E. & Barabási, A.-L. Interactome Networks and Human Disease. *Cell* **144,** 986–998 (2011).

13. Sharma, A., Menche, J., Chris Huang, C., Ort, T., Zhou, X., Kitsak, M., Sahni, N., Thibault, D., Voung, L., Guo, F., Ghiassian, S. D., Gulbahce, N., Baribaud, F., Tocker, J., Dobrin, R., Barnathan, E., Liu, H., Panettieri, R. A., Tantisira, K. G., Qiu, W., Raby, B. A., Silverman, E. K., Vidal, M., Weiss, S. T. & Barabási, A. L. A disease module in the interactome explains disease heterogeneity, drug response and captures novel pathways and genes in asthma. *Hum. Mol. Genet.* **24,** 3005–3020 (2014).

14. Liu, S., Liu, Y., Zhao, J., Cai, S., Qian, H., Zuo, K., Zhao, L. & Zhang, L. A computational interactome for prioritizing genes associated with complex agronomic traits in rice. *Plant J.* 177–188 (2017). doi:10.1111/tpj.13475

15. Vidal, M. & Fields, S. The yeast two-hybrid assay: Still finding connections after 25 years. *Nature Methods* **11,** 1203–1206 (2014).

16. Frick, L. E. & LaMarr, W. A. *Using the Agilent RapidFire 365 High-Throughput Mass Spectrometry System*. (2014).

17. Huang, H., Jedynak, B. M. & Bader, J. S. Where have all the interactions gone? Estimating the coverage of two-hybrid protein interaction maps. *PLoS Comput. Biol.* **3,** 2155–2174 (2007).

18. 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. & Vidal, M. Next-generation sequencing to generate interactome datasets. *Nat. Methods* **8,** 478–480 (2011).

19. Lewis, J. D., Wan, J., Ford, R., Gong, Y., Fung, P., Nahal, H., Wang, P. W., Desveaux, D. & Guttman, D. S. Quantitative Interactor Screening with next-generation Sequencing (QIS-Seq) identifies Arabidopsis thaliana MLO2 as a target of the Pseudomonas syringae type III effector HopZ2. *BMC Genomics* **13,** 8 (2012).

20. Hastie, A. R. & Pruitt, S. C. Yeast two-hybrid interaction partner screening through in vivo Cre-mediated Binary Interaction Tag generation. *Nucleic Acids Res.* **35,** (2007).

21. Lister, R., Gregory, B. D. & Ecker, J. R. Next is now: new technologies for sequencing of genomes, transcriptomes, and beyond. *Current Opinion in Plant Biology* **12,** 107–118 (2009).

22. Cao, S., Siriwardana, C. L., Kumimoto, R. W. & Holt, B. F. Construction of high quality Gateway^{TM} entry libraries and their application to yeast two-hybrid for the monocot model plant *Brachypodium distachyon*. *BMC Biotechnol.* **11,** 53 (2011).

23. Benatuil, L., Perez, J. M., Belk, J. & Hsieh, C. M. An improved yeast transformation method for the generation of very large human antibody libraries. *Protein Eng. Des. Sel.* **23,** 155–159 (2010).

24. Dreze, M., Monachello, D., Lurin, C., Cusick, M. E., Hill, D. E., Vidal, M. & Braun, P. High-quality binary interactome mapping. *Methods Enzymol.* **470,** 281–315 (2010).

25. Rolland, T., Taşan, M., Charloteaux, B., Pevzner, S. J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R., Kamburov, A., Ghiassian, S. D., Yang, X., Ghamsari, L., Balcha, D., Begg, B. E.,

Braun, P., Brehme, M., Broly, M. P., Carvunis, A. R., Convery-Zupan, D., Corominas, R., Coulombe-Huntington, J., Dann, E., Dreze, M., Dricot, A., Fan, C., Franzosa, E., Gebreab, F., Gutierrez, B. J., Hardy, M. F., Jin, M., Kang, S., Kiros, R., Lin, G. N., Luck, K., Macwilliams, A., Menche, J., Murray, R. R., Palagi, A., Poulin, M. M., Rambout, X., Rasla, J., Reichert, P., Romero, V., Ruyssinck, E., Sahalie, J. M., Scholz, A., Shah, A. A., Sharma, A., Shen, Y., Spirohn, K., Tam, S., Tejeda, A. O., Trigg, S. A., Twizere, J. C., Vega, K., Walsh, J., Cusick, M. E., Xia, Y., Barabási, A. L., Iakoucheva, L. M., Aloy, P., De Las Rivas, J., Tavernier, J., Calderwood, M. A., Hill, D. E., Hao, T., Roth, F. P. & Vidal, M. A proteome-scale map of the human interactome network. *Cell* **159,** 1212–1226 (2014).

26. Corominas, R., Yang, X., Lin, G. N., Kang, S., Shen, Y., Ghamsari, L., Broly, M., Rodriguez, M., Tam, S., Trigg, S. A., Fan, C., Yi, S., Tasan, M., Lemmens, I., Kuang, X., Zhao, N., Malhotra, D., Michaelson, J. J., Vacic, V., Calderwood, M. A., Roth, F. P., Tavernier, J., Horvath, S., Salehi-Ashtiani, K., Korkin, D., Sebat, J., Hill, D. E., Hao, T., Vidal, M. & Iakoucheva, L. M. Protein interaction network of alternatively spliced isoforms from brain links genetic risk factors for autism. *Nat. Commun.* **5,** 3650 (2014).

27. Yang, X., Coulombe-Huntington, J., Kang, S., Sheynkman, G. M., Hao, T., Richardson, A., Sun, S., Yang, F., Shen, Y. A., Murray, R. R., Spirohn, K., Begg, B. E., Duran-Frigola, M., MacWilliams, A., Pevzner, S. J., Zhong, Q., Trigg, S. A., Tam, S., Ghamsari, L., Sahni, N., Yi, S., Rodriguez, M. D., Balcha, D., Tan, G., Costanzo, M., Andrews, B., Boone, C., Zhou, X. J., Salehi-Ashtiani, K., Charloteaux, B., Chen, A. A., Calderwood, M. A., Aloy, P., Roth, F. P., Hill, D. E., Iakoucheva, L. M., Xia, Y. & Vidal, M. Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell* **164,** 805–817 (2016).

28. Varjosalo, M., Sacco, R., Stukalov, A., Van Drogen, A., Planyavsky, M., Hauri, S., Aebersold, R., Bennett, K. L., Colinge, J., Gstaiger, M. & Superti-Furga, G. Interlaboratory reproducibility of large-scale human protein-complex analysis by standardized AP-MS. *Nat. Methods* **10,** 307–314 (2013).

29. Huang, H. & Bader, J. S. Precision and recall estimates for two-hybrid screens. *Bioinformatics* **25,** 372–378 (2009).

30. Soellick, T. R. & Uhrig, J. F. Development of an optimized interaction-mating protocol for large-scale yeast two-hybrid analyses. *Genome Biol.* **2,** RESEARCH0052 (2001).

31. Gnirke, A., Melnikov, A., Maguire, J., Rogov, P., LeProust, E. M.,

Brockman, W., Fennell, T., Giannoukos, G., Fisher, S., Russ, C., Gabriel, S., Jaffe, D. B., Lander, E. S. & Nusbaum, C. Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* **27,** 182–189 (2009).

CHAPTER 2:

CrY2H-SEQ: A MASSIVELY MULTIPLEXED ASSAY FOR DEEP-COVERAGE INTERACTOME MAPPING

CrY2H-seq: a massively-multiplexed assay for deep coverage interactome mapping

Shelly A. Trigg[1,2], Renee M. Garza[1], Andrew MacWilliams[1], Joseph R. Nery[1], Anna Bartlett[1], Rosa Castanon[1], Adeline Goubil[1,5], Joseph Feeney[1,5], Ronan O'Malley[1,3,5], Shao-shan C. Huang[1,3,4], Zhuzhu Z. Zhang[1], Mary Galli[1,5], and Joseph R. Ecker [1,3,4*]

[1]Genomic Analysis Laboratory, The Salk Institute for Biological Studies, La Jolla, California, USA.

[2]Division of Biological Sciences, University of California San Diego, La Jolla, California, USA.

[3]Genomic Analysis Laboratory, The Salk Institute for Biological Studies, La Jolla, California, USA.

[4]Howard Hughes Medical Institute, The Salk Institute for Biological Studies, La Jolla, California, USA.

[5]Present addresses: National Institute for Agricultural Research, Paris, France (A.G.); Goodreads, Amazon Inc., San Francisco, California, USA (J.F.); United States Department of Energy Joint Genome Institute, Walnut Creek, California, USA (R.O.); Waksman Institute of Microbiology, Rutgers University, Piscataway, New Jersey, USA (M.G.).

*Correspondence should be addressed to J.R.E. (ecker@salk.edu)

ABSTRACT

Broad-scale protein-protein interaction mapping is a major challenge given the cost, time, and sensitivity constraints of existing technologies. Here, we present a massively multiplexed yeast two-hybrid method, CrY2H-seq, which uses a Cre recombinase interaction reporter to intracellularly fuse the coding sequences of two interacting proteins, and next-generation DNA sequencing to identify these interactions *en masse*. We applied CrY2H-seq to investigate sparsely annotated *Arabidopsis thaliana* transcription factor interactions. By performing ten independent screens testing 36 million binary interaction combinations, and uncovering a network of 8,577 interactions among 1,453 transcription factors, we demonstrate CrY2H-seq's improved screening capacity, efficiency, and sensitivity over those of existing technologies. The deep-6coverage network resource we call AtTFIN-1 recapitulates one third of previously reported interactions derived from diverse methods, expands the number of known plant transcription factor interactions by three-fold, and reveals previously unknown family-specific interaction module associations with plant reproductive development, root architecture, and circadian coordination.

INTRODUCTION

The yeast two-hybrid (Y2H) assay is one of the most widely adopted methods for high-throughput mapping of binary protein-protein interactions. Y2H data sets[1–3] have contributed substantially to protein interaction repositories[4] and probabilistic interactome databases[5,6]. Moreover, Y2H data have revealed complexes regulating disease[7] and helped researchers identify cancer subtypes where similar network regions are affected by different somatic tumor mutations[8] and conditional subnetworks underlying different plant-pathogen infections[9]. However, broad-scale Y2H data acquisition remains constrained by cost and labor requirements of tracking interactions and the iterative screening necessary to generate complete interactome maps[10].

Advancements that leverage next-generation sequencing to identify interactions have made large-scale Y2H screening more feasible[1,11,12]. To circumvent the isolated screening of bait proteins for tracking interactions, multiplexed screening strategies that enable pools of baits to be screened against pools of preys were recently developed[12,13]. Barcode Fusion Genetics (BFG-Y2H)[12] uses intracellular DNA recombination of barcoded open reading frame (ORF) clones to identify interacting proteins, thus allowing Y2H-positive colonies to be pooled and sequenced simultaneously. However, this technology quickly becomes costly for large-scale screening on account of the isolation and sequencing requirements of each barcoded bait and prey clones before screening to make barcode-ORF associations.

To more efficiently enable iterative screening, we developed Cre-reporter-mediated yeast two-hybrid coupled with next-generation sequencing (CrY2H-seq). CrY2H-seq uses Cre recombinase as a Y2H protein-protein interaction reporter that functions intracellularly to covalently and unidirectionally link interacting bait and prey plasmids via specialized loxP sites that flank the protein-coding sequences. The linked protein-coding sequences serve as interaction-identifying DNA molecules that enable massively multiplexed screening coupled with next-generation DNA sequencing to detect protein-protein interactions.

We applied CrY2H-seq to comprehensively screen in an 'all-by-all' fashion a collection of 1,956 *Arabidopsis* transcription factors and regulators (collectively called TFs)[14]. From ten independent CrY2H-seq 'all-by-all' screens, we report a deep-coverage *Arabidopsis* transcription factor interactome composed of 8,577 binary interactions, 7,994 of which were previously unreported. After experimentally and computationally validating interactions, we identified several network modules associated with plant reproductive development, root growth, environmental regulation of circadian rhythms, and stress- and hormone-response pathway crosstalk.

RESULTS

CrY2H-seq assay development

To establish CrY2H-seq, we first generated a yeast strain, CRY8930, that carries both a Gal4-inducible *GAL7*::*CRE* expression cassette and two well-characterized *GAL1::HIS3* and *GAL2::ADE2* auxotrophic expression cassettes[1] (**Fig. 2.1a**). We then modified a widely used ARS/CEN Gateway-compatible plasmid set[1,3] to contain unidirectional lox sequences[15] flanking the 3' end of ORF inserts, such that upon Cre recombination both ORF inserts would be on the same DNA molecule in a fixed orientation (**Fig. 2.1b**). By screening yeast transformants harboring known positive and negative interaction pairs in these modified plasmids (see Online Methods), we confirmed that positive pairs induced Cre expression in addition to enabling growth selection (**Supplementary Fig. 1)**. Yeast colony PCR with Gal4-AD and Gal4-DB primers (**Fig. 2.1b** and **Supplementary Table 1**) produced amplicons only for positive pairs, which indicated that plasmids underwent Cre-recombination (**Supplementary Fig. 1b,c**). Sanger sequencing of Cre-recombination PCR products verified that a newly formed double-mutant lox site became sandwiched between the two ORF sequences, and recombination occurred in a fixed 3'-end-to-3'-end fashion **(Fig. 2.1c).** Moreover, interactions gave the same either positive or negative result regardless of whether CRY8930 or the unmodified Y8930 was used (**Supplementary Fig. 2**).

There are two main distinctions between CrY2H-seq and existing multiplexed Y2H technologies[12,13]. First, interactions detected by CrY2H-seq require the parallel activation of two reporter genes driven by distinct promoters for detection of interactions; an auxotrophic rescue reporter and *CRE*. We used *HIS3* in conjunction with *CRE* because *GAL1:HIS3* is known to be more sensitive than *ADE2* for detecting interacting proteins[16], and the use of the independent *GAL7* promoter to drive *CRE* expression reduces promoter-specific false positives[17]. Furthermore, including *CRE* as a secondary reporter gene minimizes the time and reagents required of a steroid-inducible Cre expression system[12,13]. The second distinction is that CrY2H-seq uses interacting protein coding sequences themselves to form an intracellular DNA identifier **(Fig. 2.1c)** rather than barcode identifiers[12] that could become a bottleneck in large-scale screens. These key features allowed us to circumvent current Y2H limitations and establish a general CrY2H-seq pipeline for all-by-all massively multiplexed screening (**Fig. 2.2**).

Deep interaction screening of an *Arabidopsis* transcription factor ORFeome

We loaded a set of 1,956 *Arabidopsis* TFs[14] into the CrY2H-seq pipeline and performed ten all-by-all screens with final bait and prey libraries containing 1,877 and 1,933 unique yeast clones, respectively (**Supplementary Table 2a** and Online Methods). These starting library populations showed an ORF size distribution consistent with the expected size distribution (**Supplementary Fig. 3a**), and the data showed minimal ORF size bias (**Supplementary Fig. 3b,c**). While bait proteins are typically screened for self-activation before Y2H screening, we chose to eliminate this step in order to rigorously challenge whether the assay would be able detect real interaction signal above the 'noise' from self-activator interactions. Instead, to internally control for self-activating bait proteins[18], we spiked into each screen an excess amount of a Y8800 strain harboring an empty pADlox plasmid. Libraries were mated and underwent *HIS3* reporter selection ten independent times. This deep screening tested 3.6 million potential protein combinations approximately 300 times, for an estimated total of one billion interactions surveyed (see Online Methods).

After carrying out multi-template PCR amplification on plasmid pools isolated from each screen, we randomly sheared the PCR products to ~300 bp and generated standard Illumina-based DNA sequencing libraries (**Fig. 2.2**). We then performed 100-bp paired-end Illumina sequencing, aiming for a previously established optimized coverage of 40 million reads per screen (**Supplementary Fig. 4** and Online Methods**)**. Paired-end reads were mapped and quality filtered, and fragments corresponding to Cre-recombined ORF junctions were extracted (**Supplementary Fig. 5a-e** and Online Methods). We applied a predetermined basal fragment cutoff to eliminate any putative interactors that were represented by less than three junction fragments (**Supplementary Fig. 5f** and Online Methods). The remaining interaction-identifying fragments (see Online Methods) were normalized across the ten independent screens to control for variation between sequencing runs (**Supplementary Fig. 5g** and Online Methods), and they were classified as 'normalized protein interaction fragments' (NPIFs; **Fig. 2.2**). Very minimal amplicon size bias was observed in our data set (**Supplementary Fig. 3d,e)**, although fragments mapping to homodimers were notably absent from the data, likely due to difficulty in amplification of the hairpin structure formed by fused identical ORFs, as was previously observed in small-scale experiments (**Supplementary Fig. 6**).

In total, 10.9 million NPIFs were identified from the ten CrY2H-seq screens, and these NPIFs mapped to 173,000 unique Cre-recombined ORF junctions (**Fig. 2.3a**). Among these were 299 different pDBlox ORFs fused to an empty pADlox vector, indicating that 16% of baits exhibited self-activation (**Supplementary Table 3a**). All 164,293 unique ORF combinations containing these TFs (**Supplementary Table 3b**) were excluded from the data. The remaining 1.4 million (13%) NPIFs mapped to 8,577 protein interactions, with a median of 7 NPIFs per interaction (**Fig. 2.3b**). The 8,577 interactions form the deep-coverage interactome we refer to as "*Arabidopsis thaliana* transcription factor interaction network, version 1" (AtTFIN-1) (http://signal.salk.edu/interactome/AtTFIN-1.html, **Supplementary Table 2b,c**, see Online Methods).

Validation of AtTFIN-1 Interactions

To estimate sampling sensitivity, the fraction of all identifiable interactions found in one screen[10], we simulated results for all possible orderings of replicate screens and found that one screen alone on average yielded 2012 ± 354 interactions (mean ± standard deviation). Calculating the average number of new interactions gained after each of the ten screens (**Fig. 2.3c**) revealed that, even after ten screens, saturation was not reached. We fit this data to a Michaelis-Menton modeled curve to estimate the degree of saturation and determined that, of the 15,610 ± 2,661 interactions that could have been maximally detected (**Supplementary Fig. 7** and Online Methods), we detected more than half (54.6%).

To estimate reproducibility, we retested 771 (9%) AtTFIN-1 interactions (678 of which were novel) that showed a range of NPIFs and screen occurrences (**Supplementary Table 4**) using a standard pairwise 1x1 array style Y2H screen[18] (**Supplementary Fig. 8a**). Excluding *de novo* self-activating baits identified by parallel plating on cycloheximide selection media[18], we observed an overall retest rate of 73% (422/580 novel interactions and 57/76 'known' interactions, defined below). Additionally, we observed an increased retest rate for interactions appearing in multiple screens (**Fig. 2.4a**) but a relatively similar retest rate among interactions showing different ranges of NPIFs (**Supplementary Fig. 8b**). We also tested 94 AtTFIN-1 interactions (59 of which were novel) (**Supplementary Table 5a**) using the wNAPPA assay[19] and observed that 50% of all AtTFIN-1 interactions and 25.4% of novel AtTFIN-1 interactions tested positive (**Fig. 2.4b** and **Supplementary Fig. 9**). These rates contrasted significantly with the 2.8% positive rate observed for 36 random TF interactions tested in wNAPPA.

To estimate assay sensitivity, the fraction of all detectable biophysical interactions[10], we mined both literature[3] and databases[4–6] for TF interactions that were screened in CrY2H-seq (**Supplementary Table 2b**). We refer to these mined interactions collectively as 'known' interactions. Interactions involving self-activating TFs and homodimers were excluded from this analysis. AtTFIN-1 showed the greatest overlap (52.2%) with *Arabidopsis* Interactome-1 interactions[3] and the least overlap with AraNet[6] interactions (13.4%) (**Fig. 2.4c**). We estimated a false positive rate of 0.69% ± 0.12% (mean ± standard deviation), by calculating the overlap of AtTFIN-1 interactions with ten different data sets, each composed of 8,577 randomly generated TF interactions (see Online Methods). Overall, AtTFIN-1 interactions showed significantly greater recapitulation of known interactions, including those derived from a variety of assays (**Supplementary Fig. 10a**), relative to random interactions (**Fig. 2.4c**). A precision-recall curve of these detection rates plotted as a function of the number of screen occurrences showed a large drop in precision with little gain in recall between one and two screens, which led us to classify high-confidence interactions as those identified in two or more screens (**Fig. 2.4d**).

To measure performance improvements over array-based high-throughput Y2H (HT-Y2H), we compared TF interaction detection rates between CrY2H-seq and HT-Y2H used to generate the *Arabidopsis*-Interactome-1[3]. CrY2H-seq showed a five-fold increase in general TF interaction detection relative to HT-Y2H (**Supplementary Fig. 11a**). Of the commonly screened TF interactions, CrY2H-seq showed a seven-fold increase in detection—it recovered 1,609 TF interactions, whereas HT-Y2H detected only 229 (**Supplementary Fig. 11b**). Of the commonly tested literature-curated interaction (LCI) pairs[3], CrY2H-seq recalled 33.3% while HT-Y2H recalled only 12.3% (**Supplementary Fig. 11c**). While CrY2H-seq showed a clear overall improvement to HT-Y2H, it should be noted that the *Arabidopsis* Interactome-1 was based on the union of two primary screens and was filtered by pairwise retesting, where AtTFIN-1 was based on ten primary screens that were not filtered by pairwise retesting.

To evaluate the biological relevance of AtTFIN-1 interactions, we compared expression correlations between AtTFIN-1 interactions and a random interaction data set using 6,057 different expression data sets[20]. We observed significantly higher expression correlation for transcripts encoding AtTFIN-1 interactions than for transcripts encoding random interactions (**Supplementary Fig. 12**), supporting the potential of AtTFIN-1 interactions to interact *in vivo*.

AtTFIN-1 defines expanded transcription factor modules.

We further investigated the biological significance of the 3,086 high-confidence AtTFIN-1 interactions (2,578 novel) by looking for 'preferential' intrafamily and interfamily interactions that occurred more frequently than would be expected by chance. AtTFIN-1 interactions classified by previously assigned familes[14] were compared to those in 10,000 randomly rewired degree-conserved networks (**Fig. 2.5a, Supplementary Fig. 13** and Online Methods). We observed highly significant preferential intrafamily interactions among family members known to dimerize including those in the bHLH, MADS, bZIP, NAC, WRKY, AUX-IAAs, and ARF families. We also observed highly significant preferential interfamily interactions between plant-specific families known to dimerize including growth regulating factors (GRFs) and growth regulating factor interacting factors (GIFs)[21], LUGs and YABBYs[22], and AUX-IAAs and ARFs[23]. The teosinte-branched/cycloidea/proliferating cell factor (TCP) family showed significant preference for 18 TF families (**Supplementary Fig. 13**), a broad preference consistent with previously observations of TCPs as ‘hub’ proteins[3,24].

We further examined highly significant, previously unknown preferential interfamily interactions; and we found that the preference of the ABI3-VP1/B3 family for GeBP and TRIHELIX proteins was driven by one ABI3-VP1/B3 member, AT5G60142, which showed many interactions with various TRIHELIX and GeBP members (**Fig. 2.5b**). While the GeBP and TRIHELIX members have sparse gene ontology (GO) annotations, AT5G60142 has recently been found upregulated in isolated early-stage gynoecium medial domain cells[25]. Interestingly, not only were AT5G60142 and 93% (13/14) of its TRIHELIX and GeBP interacting partners found co-expressed in this study, but five of AT5G60142’s partners (ASIL2, AT3G58630, AT1G76870, AT3G04930, and STKL1) were significantly upregulated in cells from the same distinct domain. These interactions may form part of a previously unrecognized module underlying early-stage reproductive development. We also found the preference of G2-like proteins for the GRAS family was driven by multiple phosphate response-like factors and the scarecrow-like factors (**Fig. 2.5c**). This network reveals a logical link between phosphate sensing and root development, consistent with the notion that phosphate deprivation drives altered root architecture and increased root hair density[26,27]. C2C2-CO-like TFs showed significant preferential interaction with the ‘orphans’ category of unassigned TFs (**Fig. 2.5d**). Closer examination of these interactions revealed that all proteins contained BBX domains, including the C2C2-CO-like proteins themselves. These interactions could be mediated by BBX domains as these have been shown to be crucial in mediating protein-protein interactions and transcriptional regulation[28]. Many BBX domain-containing proteins are known to have specific and sometimes opposing functions in regulating flowering, circadian clock, biotic or abiotic stress response[28]. Moreover, it was recently reported that overexpressing AtBBX32 in soybean plants increased grain yield by altering light input and expression patterns of clock genes necessary for initiation of different stages of reproductive development[29]. This AtTFIN-1 module suggests that combinatorial complexity among BBX proteins may play a role in integrating environmental signals and flowering time potentially through feedback or feed-forward loops.

Beyond the well-characterized interfamily interaction between ARFs and AUX/IAAs[23], for which we observed a significant preferential family interaction between eight ARF members and 23 AUX/IAA members, individual AUX-IAA members showed distinct interactions with other families (**Fig. 2.6**). For instance, compared with other IAAs, IAA17 heavily interacted with TCPs; this suggests that IAA17 could be the main player mediating crosstalk between auxin and TCP transcriptional regulation. IAA2, IAA10, IAA17, and IAA18 commonly interacted with methyl-CpG binding domain (MBD) proteins, which indicated the potential involvement of these IAAs in regulating DNA methylation. Particular IAAs and ARFs showed interactions with specific hormone and stress associated TFs—IAA11 with hormone or abiotic stress response factors ERF70 and DRIP2, IAA10 with defense response factors LOL2 and GEBP, and ARF18 with abscisic acid response factors VAL1 and VAL2. This suggests how these factors potentially integrate auxin response with different hormone and stress signals. This expanded ARF-AUX-IAA interactome reveals how particular TFs may play specific roles in mediating cross-talk between auxin response and other plant pathways.

DISCUSSION

CrY2H-seq offers an untargeted, highly scalable screening approach to directly assay binary protein-protein interactions in yeast. We demonstrated that 36 million interactions could be assayed to >50% saturation with ten cost-effective and time-efficient CrY2H-seq replicate screens (**Supplementary Fig. 14**)—a scale which, to our knowledge, has not been previously achievable. Its increased interaction detection rates and significantly greater overlap with previously reported interactions (**Fig. 2.4c**, **Supplementary Fig. 10a, Supplementary Fig. 11**) suggest CrY2H-seq could increase overlap between interlaboratory Y2H screens[30] . We attribute these increases to the use of next-generation sequencing for interaction detection and to the ease of iterative screening. Moreover, the sensitivity of CrY2H-seq may even be underestimated; removal of self-activating proteins before screening could lead to the detection of missed interactions. Nonetheless, our CrY2H-seq screening was not exhaustive, nor did it completely capture all known interactions, alluding to inherent limitations of yeast two-hybrid methods, including sub-optimal protein expression levels or strain copy number in pools. CrY2H-seq could be further optimized to reduce sequencing costs by applying strategies for targeted capture of fused lox-containing DNA fragments and depletion of over-abundant DNA from sequencing libraries. Additionally, the incorporation of a unique DNA sequence into the lox region on one of the CrY2H-seq plasmids could disrupt the hairpin structure to allow the potential detection of homodimers and optimized tracking of bait/prey orientations.

The AtTFIN-1 resource generated from CrY2H-seq screening substantially expands the available interaction data among *Arabidopsis* TFs, tripling the 3,170 interactions documented in BioGRID[4]. The novel interactions we identified reveal potential involvement of poorly annotated TFs in various biological processes, including root and reproductive development, and the integration of environmental stimulus with circadian rhythms. These data can be used for future genomic analyses and data integration pipelines to further define these network modules and help identify candidate genes that could be used for crop improvement. This expanded TF network can be used to generate hypotheses regarding the specific roles of individual TFs or TF families throughout development and in response to a multitude of biotic and abiotic stressors. For instance, the activity of AtTFIN-1 interactions could be tested on different promoters to examine how interactions affect target gene expression[31]. Further understanding the roles of TF interaction partners in combinatorial gene regulation is particularly valuable for improving crop optimization strategies that currently target individual TFs[32].

Lastly, CrY2H-seq technology could be applied to Y2H assay variations. For instance, CrY2H-seq could be adapted to the split-ubiquitin system[33] or potentially even to mammalian membrane two-hybrid[34] for screening hydrophobic proteins, or to yeast one-hybrid for screening genome-wide protein-DNA interactions[35]. The ease of setting up CrY2H-seq replicate experiments permits screening on multiple media types for selection of different reporter genes, or selection on media supplemented with various hormones that may influence interactions[36]. Furthermore, while we used an array cloning strategy[18] here for mobilizing ORFs into CrY2H-seq plasmids, *en masse* cloning strategies[37,38] can be used to reduce cost and (importantly) extend the application of CrY2H-seq to cDNA library-against-cDNA library screening. This would enable comparisons of unprecedentedly large-scale interactomes derived from different ecotypes, growth conditions, or tissue types; and it would enable the identification of network differences underlying different phenotypes. Interaction maps generated by CrY2H-seq could be integrated with other ‘omics’ data to provide deeper insight into the functional relationships between genotype and phenotype, the network effects of variants, and interactome modules that certain transcriptional programs give rise to.

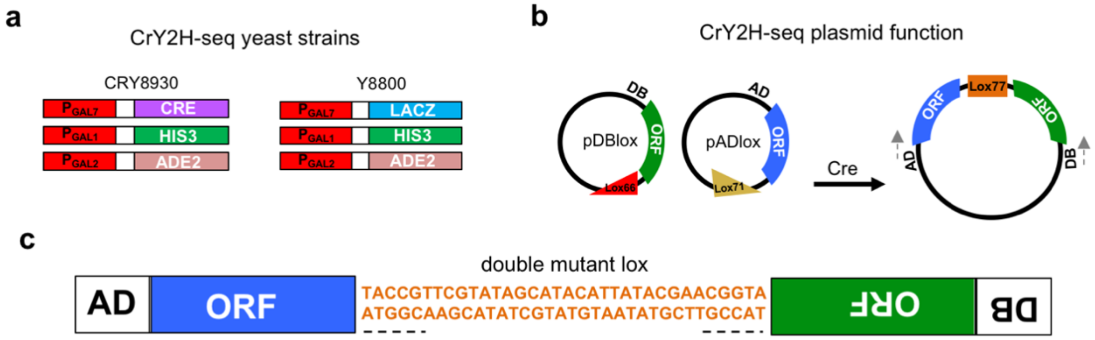

Figure 2.1. CrY2H-seq strain and plasmid design. (**a**) CrY2H-seq uses yeast strains CRY8930 and Y8800. (**b**) CrY2H-seq bait and prey plasmids pDBlox and pADlox contain mutant *lox* sites (*lox66* and *lox71*, respectively) flanking the 3' end of ORF inserts. Upon Cre/*lox*-recombination of plasmids, a fused ORF product can be recovered by PCR amplification using activation (AD) and DNA binding (DB) domain-specific primers, indicated by the grey arrows. (**c**) Representative PCR amplicon from AD and DB primers showing fused ORFs. Mutant *lox* sites are underlined.

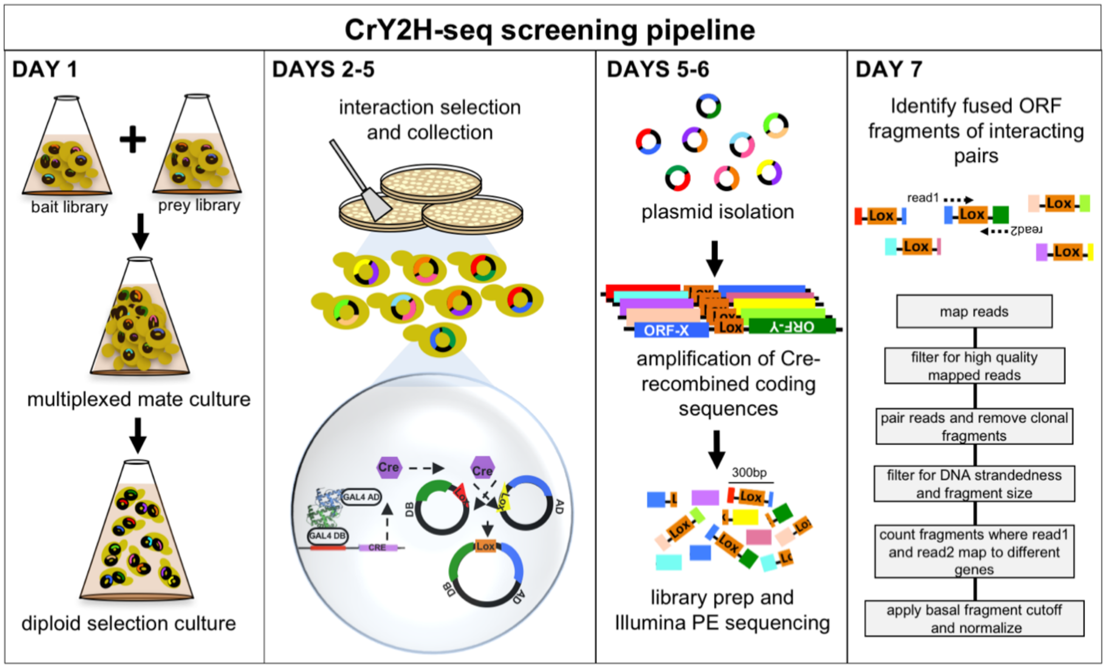
CrY2H-seq screening pipeline
DAY 1
bait library
prey library
multiplexed mate culture
diploid selection culture
DAYS 2-5
interaction selection and collection
Cre
Cre
GAL4 AD
GAL4 DB
CRE
DB
AD
Lox
DB
AD
DAYS 5-6
plasmid isolation
ORF-X
Lox
ORF-Y
amplification of Cre-recombined coding sequences
300bp
Lox
library prep and Illumina PE sequencing
DAY 7
Identify fused ORF fragments of interacting pairs
read1
read2
Lox
map reads
filter for high quality mapped reads
pair reads and remove clonal fragments
filter for DNA strandedness and fragment size
count fragments where read1 and read2 map to different genes
apply basal fragment cutoff and normalize

Figure 2.2. The CrY2H-seq screening pipeline. On day 1, archival stocks of bait and prey libraries are combined in one massively multiplexed mate culture that undergoes diploid selection overnight. On day 2, the diploid culture is plated on media to select for cells with protein-interaction-mediated Gal4 reconstitution and subsequent transcriptional activation of the *HIS3* and *CRE* reporter genes. *HIS3* expression allows cells to survive on selection media, and *CRE* expression permits unidirectional plasmid linkage, where ORF combinations corresponding to protein-protein interactions become fixed together inside cells. After 3 d of selection, surviving cells are harvested *en masse*, plasmids are purified in a single prep, and Cre-recombined ORF junctions are amplified in multitemplate PCR reactions. From these amplicons, an Illumina sequencing library is prepared and sequenced. A bioinformatics pipeline is used to identify fragments derived from Cre recombination PCR products (see **Supplementary Fig. 5** and Online Methods for more details, including those regarding media composition).

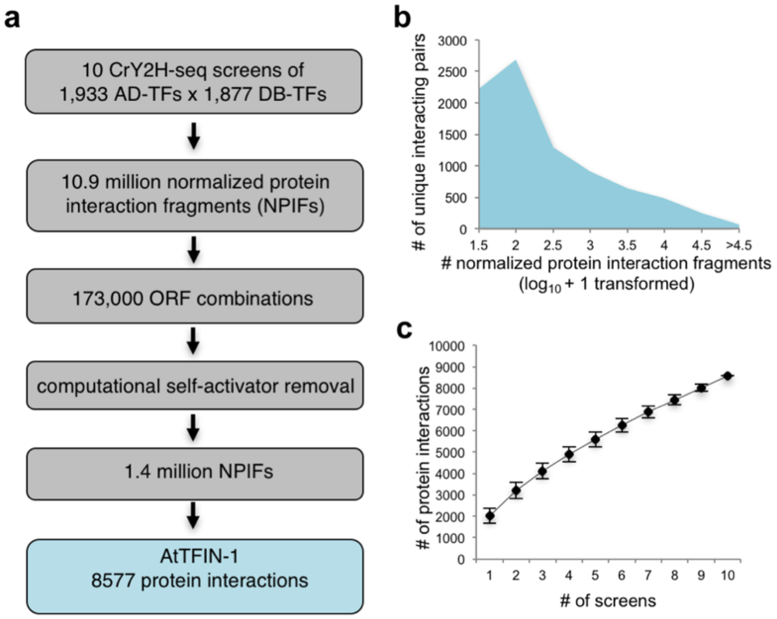

Figure 2.3. Coverage of AtTFIN-1. (**a**) Summary of TF ORFeome screening. (**b**) Cumulative coverage of unique interacting pairs detected in paired-end sequencing of all ten CrY2H-seq screens after self-activator removal. (**c**) Sampling sensitivity shown by the average number of new interactions detected after each CrY2H-seq screen considering all possible (10!) orderings of screens. Error bars, s.d.

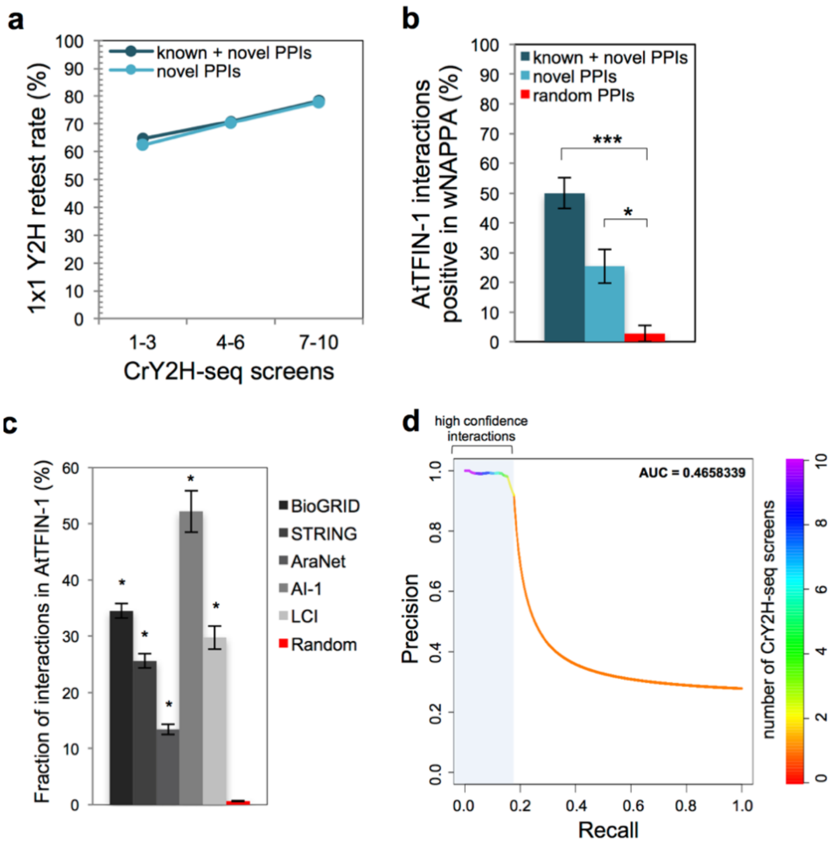

Figure 2.4. Quality of AtTFIN-1. (**a**) Fraction of AtTFIN-1 protein-protein interactions (PPIs) that were positive in 1x1 matrix style Y2H retest screen (retest rate) as a function of the number of CrY2H-seq screens that interactions were observed in. Bin sizes, 1-3: 65, 4-6: 342, and 7-10: 249. (**b**) Fraction of AtTFIN-1 PPIs that were positive in wNAPPA. Error bars, standard error of proportion. *P* values, one-sided Fisher's exact test (*** = 3.57e-08, and * = 0.002395). (**c**) Fraction of 1,368 BioGRID, 1,198 STRING, 1,355 AraNet, 182 *Arabidopsis* Interactome-1 (AI-1), 501 *Arabidopsis* Interactome literature-curated interactions (LCI), and 8,577 random interactions in AtTFIN-1. Error bars, standard error of proportion. Literature and database interactions are detected significantly more often than random interactions (*P* values, one-sided Fisher's exact test, * = 2.2e-16). (**d**) Precision-recall curve calculated using the union of known interactions as true positives and a random interaction data set as false positives plotted as a function of the number of CrY2H-seq screens in which interactions were observed. Interactions observed in two or more replicate experiments are classified as high-confidence interactions as indicated by the pale blue box.

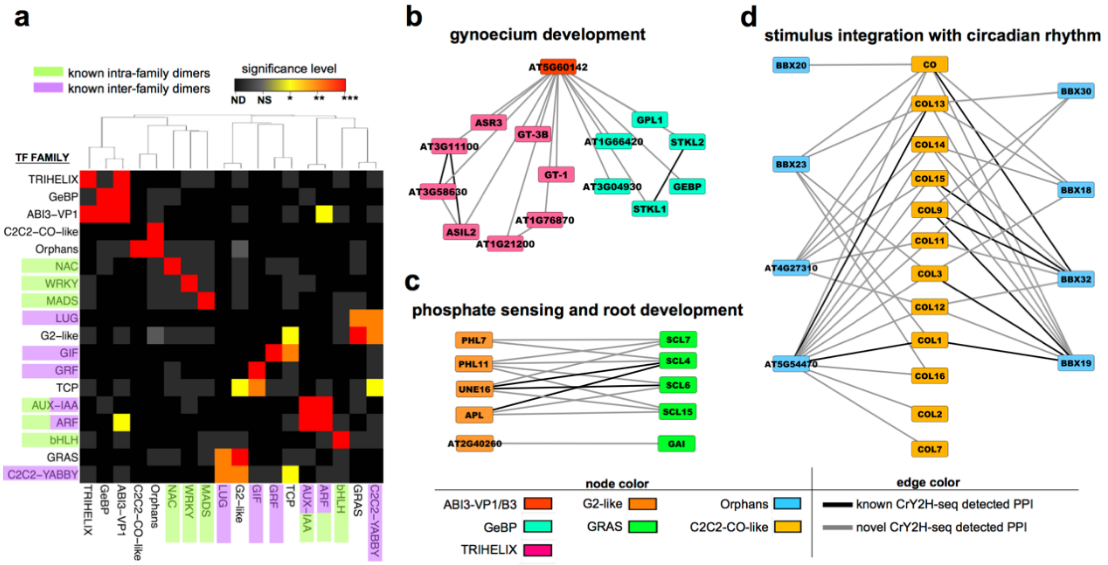

Figure 2.5. Biological functions underlying TF family interactions in AtTFIN-1. (**a**) Discrete empirical *P* values of family interactions observed more frequently in AtTFIN-1 than expected by random chance. Families are hierarchically clustered by common family interactions. Color key: ND = not detected; NS = not significant; *, *P* <0.05; **, *P* < 0.01; ***, *P* <0.001. See **Supplementary Fig. 13** for a matrix showing all TF family interactions observed. (**b**) An ABI3-VP1/B3 transcription factor preferentially interacts with many members of TRIHELIX and GeBP families, a module potentially involved in gynoecium development. (**c**) GRAS family members preferentially interact with G2-like family members providing a potential molecular link between phosphate sensing and the regulation of root development. (**d**) Preferential interaction between BBX domain-containing 'Orphans' proteins and C2C2-CO-like family members suggests a potential means by which stimulus signals are integrated with circadian rhythms.

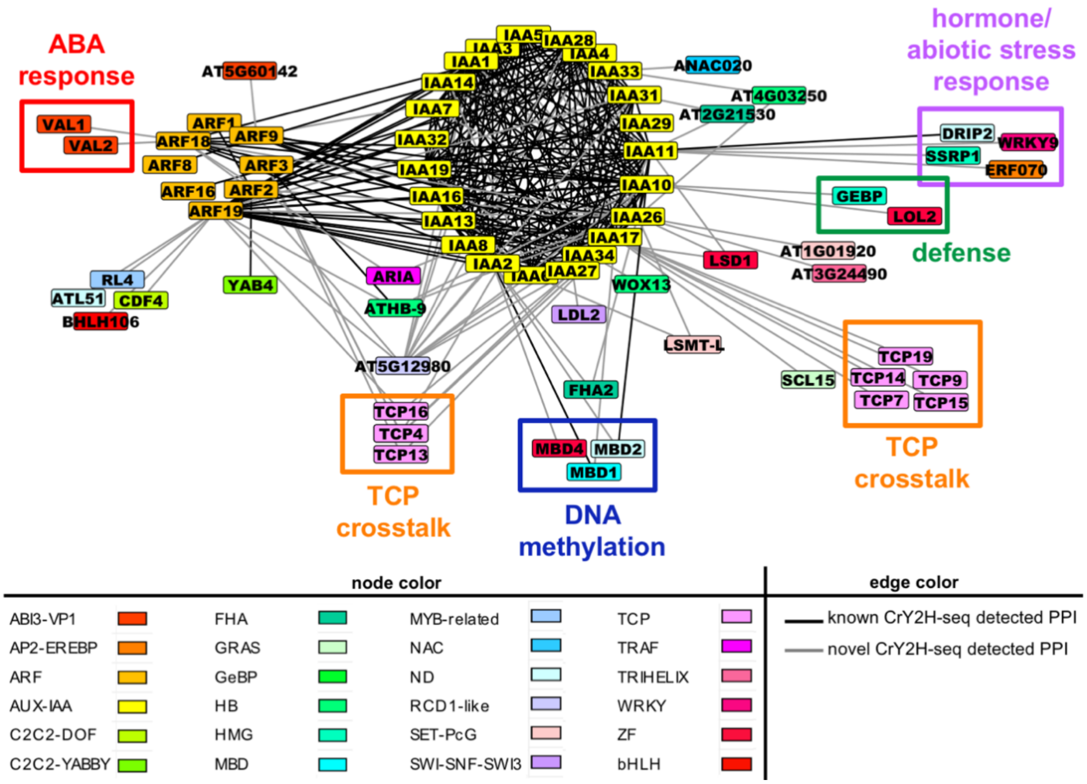

Figure 2.6. An expanded ARF-AUX-IAA transcription factor network. Distinct interactions among AUX-IAA and ARF proteins suggest certain family members have specific functions. IAA17 shows preferential enrichment for TCP family members. IAA2, IAA10, IAA17, and IAA18 commonly interact with MBD proteins. IAA11 shows distinct interactions with hormone and water stress related factors, ERF70 and DRIP2. ARF18 specifically interacts with VAL1 and VAL2 abscisic acid response factors. IAA10 interacts with LOL2 and GEBP defense response-related factors.

METHODS

A step-by-step protocol is available as a **Supplementary Protocol** (**Supplementary File 5**) and at the *Protocol Exchange*[39].

Strain and plasmid construction

Primers used to modify plasmids and the CRY8930 strain are listed in **Supplementary Table 1.** The genotype of CRY8930 is MATα *leu2-3,112 trp1-901 his3-200 ura3-52 gal4*Δ *gal80*Δ P_{GAL2}*-ADE2 LYS2::*P_{GAL1}*-HIS3 MET2::*P_{GAL7}*-CRE-HPHMX6 cyh2*^{R}*.* The genotype of Y8800 is MATa *leu2-3,112 trp1-901 his3-200 ura3-52 gal4*Δ *gal80*Δ P_{GAL2}*-ADE2 LYS2::*P_{GAL1}*-HIS3 MET2::*P_{GAL7}*-lacZ cyh2*^{R}*.* Y8800 and CRY8930 strain stocks, and pADlox and pDBlox plasmid stocks, have been deposited with the *Arabidopsis* Biological Resource Center (https://abrc.osu.edu/).

Cre reporter strain construction. The bacteriophage P1 Cre recombinase gene[40] was PCR amplified from pQTL123 GST-Cre with flanking SalI and PacI sites and ligated into SalI/PacI digested pFA6α-HPHMX6. The Cre-hygromycin resistance marker cassette was PCR amplified from the resulting plasmid and used in a homologous recombination reaction to replace the *LacZ* reporter gene within the *GAL7::LacZ* cassette integrated at the *MET2* locus of yeast strain Y8930[1]. Correct integration of *CRE* in the modified strain, referred to as CRY8930, was confirmed by sequencing of the *MET2* locus. To test *CRE* reporter gene expression, RNA was extracted from a histidine-positive diploid culture containing the known interaction pair AD-bZIP53 and DB-bZIP63[41] using the Qiagen RNeasy kit. Reverse transcription was carried out on DNAse treated RNA extract using SuperScript II (Life Technologies) followed by PCR to detect the presence of Cre cDNA (**Supplementary Fig. 1a;** primers listed in **Supplementary Table 1**).

Construction of lox *site-containing bait and prey destination vectors*. *Lox71* and *lox66* sequences[42] were inserted into the XmaI and AatII sites located downstream of the attB2 site in pDEST-AD[1] and pDEST-DB[1] respectively, using standard cloning methods. The resulting destination vectors, pADlox and pDBlox, were Sanger sequenced confirmed and transformed into One Shot ccdB Survival 2 T1R competent cells (Life Technologies). *Lox71* and *lox66* sites are modified versions of the standard *loxP* sites that display favorable forward recombination reaction equilibrium[13,15].

Pilot tests for Cre-lox recombination functionality in yeast two-hybrid. Small-scale tests were conducted to confirm the functionality of the CrY2H-seq system in yeast (**Supplementary Fig. 1b,c; Supplementary Fig. 2; and Supplementary Fig. 6**). In brief, ORFs were Gateway^{TM} cloned into pADlox and pDBlox using LR clonase and transformed into DH5α chemically competent cells. pAD-ORF-lox and pDB-ORF-lox plasmids were purified using a QIAprep Spin Miniprep kit (Qiagen) and transformed into yeast strains Y8800 and CRY8930, respectively, using a standard lithium acetate method. ORFs were also transformed into the Y8930 parental strain to serve as negative controls. Strains were mated according to published protocols[18], and grown for 3 d on interaction selection media (-Leu/-Trp/-His + 1mM 3-Amino-1,2,4-Triazole (3-

AT)). For **Supplementary Figure 1**, the known positive interaction pair bZIP53-bZIP63[41] and non-interacting pairs bZIP53-ZTL, and ZTL-bZIP63 were tested. Mated strains were also grown in parallel on diploid selection media (-Leu/-Trp). Colonies were then picked from all plates, and in the case of the non-interacting pair on interaction selection media, all background cells were scraped. Lysates were prepared as previously described[18], and PCR using AD and DB primers (**Supplementary Table 1**) was performed to detect Cre recombination products. For **Supplementary Figure 2**, before plating diploids on selection media, culture concentrations (OD_{600}) were measured on a Tecan Safire2 plate reader (**Supplementary Fig. 2b**). CRY8930/Y8800 diploids were plated adjacent to Y8930/Y8800 diploids to assess strain differences (**Supplementary Fig. 2c**). For **Supplementary Figure 6**, *HIS3*-positive colonies were picked, lysates prepared as previously described[18], and PCR using AD and DB primers (**Supplementary Table 1**) was performed to detect Cre recombination products. All PCR reactions were prepared with 1 μL of template, 0.1 μL Phusion Polymerase (NEB), 2 μL 5xGC buffer (NEB), 2 μL 5 M betaine, 200 μM each dNTP, and 0.25 μM of AD and DB primers (**Supplementary Table 1**). Reactions were run at 98°C for 2 min, 30 cycles of 98°C for 10 s, 60°C for 30 s, and 72°C for 90 s, then a final extension at 72°C for 7 min. 5 μL of each PCR reaction was run on a 1% agarose 1x TAE gel.

Transcription factor library construction for CrY2H-seq screening

All cloning and transformations were carried out according to published protocols[18]. Briefly, 1,956 entry clones from an *Arabidopsis* transcription factor ORF collection[14] were individually Gateway^{TM} cloned in 96-well format into both pADlox and pDBlox vectors using LR clonase and transformed into chemically competent DH5α-T1^{R} cells. Transformants were selected in Terrific Broth containing carbenicillin, and plasmid DNA was extracted and purified using QiaPrep 96 turbo kits (Qiagen). Next, pADlox TF plasmids and pDBlox TF plasmids were individually transformed into the yeast strains Y8800 and CRY8930, respectively, using a 96-well lithium acetate transformation protocol[18] as follows. Plasmid DNA and yeast competent cells were combined, 96-well plates were incubated at 42°C for 1 h, cells were centrifuged, washed, spotted on SC –Trp (pADlox clones) or SC –Leu (pDBlox clones), and grown at 30°C for 3 d. Colonies were then picked and inoculated into liquid SC –Trp or –Leu, and cultures were grown for 3 d at 30°C at 200 r.p.m. to reach saturation. Equal volumes of cells from individual TF clones were pooled to make the CrY2H-seq libraries for mating. Aliquots of 1 mL containing ~3 OD_{600} were mixed with 500 μL of 50% glycerol and stored at –80°C. Additionally, 96-well glycerol stocks of individual TF clones were also made for archival storage purposes.

Characterizing starting bait and prey libraries. Plasmid DNA was purified from a 1 mL aliquot of each library, from which ORF DNA was PCR amplified with either AD or DB primer and a primer that anneals to a common sequence downstream ORF inserts **(Supplementary Table 1**). An Illumina sequencing library was then prepared from each starting library by fragmenting ORF amplicons to 300 bp with a Covaris S2 sonicator, end-repairing fragments with the End-It DNA End-Repair Kit (Epicentre-Illumina), A-tailing repaired fragments with Klenow 3'-5'exo- (NEB), and ligating Illumina Truseq adapters to fragments using T4 ligase (NEB) overnight at 16°C. The adapter-ligated libraries were then run on a 2% agarose gel, and a 400-600 bp region was excised and purified using a QIAquick gel extraction kit (Qiagen). Purified DNA was then amplified with Phusion Polymerase supplemented with 1 M betaine and Illumina Truseq primers for three cycles using Illumina-recommended conditions. A final purification with SeraMag Speedbeads (GE; 2% v/v SeraMag Speedbeads, 18% w/v PEG-8000, 1 M NaCl, 10 mM Tris HCl, 1 mM EDTA) at a 1:1 bead-to-DNA ratio was performed to remove unincorporated Truseq primers; and libraries were sequenced on an Illumina paired-end 200 cycle Rapid Run on an Illumina HiSeq 2500 platform. Each library was sequenced to ~1000x coverage (bait library, 3.7 M reads; prey library, 2.3 M reads; equivalent to 1.7% of a Rapid Run flowcell). Reads were analyzed following the next-generation sequencing analysis pipeline detailed below with the following difference: paired reads for which each of the mates aligned to the same ORF and showed different strand orientation underwent a size filter that required that the difference of the start position of one read and the end position of the read pair fall within the expected library size of 400-600bp. After this filtering, ORF-mapped fragments were totaled and libraries were further characterized by plotting the size distribution and representation of detected ORFs (**Supplementary Fig. 3a-c**). A total of 1,933 and 1,877 unique AD and DB clones respectively were identified, giving rise to ~3.6 million possible combinations.

CrY2H-seq screening of transcription factor libraries

Each replicate screen consisted of mating ~20 OD_{600} of each TF clone library (pADlox in Y8800 and pDBlox in CRY8930). Based on cell titers of 2 x 10^7 cells/OD that we observed for each library, we estimated that each replicate screen would test the ~3.6 million possible protein combinations at ten-fold excess, assuming a 10% mating efficiency.

Frozen aliquots of the 1,933 TF pADlox library and the 1,877 TF pDBlox library were thawed, separately inoculated into 200 mL of YEPD media, and grown for 1 h at 30°C and 150 r.p.m. before mating. Cell concentrations were measured and libraries were combined such that each replicate screen contained ~20 OD_{600} of each CrY2H-seq library. To internally test for self-activating proteins, a pADlox empty plasmid in the Y8800 strain was spiked into each replicate mating batch in at least three-fold excess of the average individual clone population (~2 x 10^5 cells/clone). For each replicate, mating in liquid YEPD was carried out at 30°C for 4.5 h with shaking at 50 r.p.m. Subsequently, a 10 μL aliquot of the mated culture was diluted and plated on -Leu, -Trp, and -Leu/-Trp media to determine mating efficiency, which was on average 6% with ~1.25 x 10^8 diploids formed per screen. Assuming all combinations of proteins were equally represented among the diploid population, we estimate that each possible combination was sampled ~34x in each screen (1.25 x 10^8 diploids/3.63x 10^6 total protein combinations).

The remainder of the mated cultures were washed with 1x SC and individually resuspended in 100 mL 1x SC –Leu/-Trp supplemented with 125 μg/mL hygromycin to enrich for diploids and reduce background growth. These cultures were grown at 30°C overnight shaking at 150 r.p.m. Diploid cells for each screen were then collected, washed with 1x SC, and resuspended in water at 1 OD_{600} per mL. Cells were plated at roughly 0.5 OD per plate on 1x SC –Leu –Trp –His +1mM 3-AT plates (~48 plates per screen) and grown for 3 d at 30°C to select for interactors. 48 plates, each containing more than 10,000 colony forming units, were individually scraped into 48 wells of a 96-well deep-well plate. Cells were heated at 75°C for 20 min to inactivate Cre recombinase. Cells were next treated with 300 μL zymolyase buffer (0.1 M sodium phosphate buffer pH 7.4, 1% betamercaptethanol, and 2.5 mg/mL Zymolyase 20T (US Biological), and 100 μg/mL RNase A (Qiagen) and incubated at 37°C for 1 h at 50 r.p.m. Zymolyase-treated cell suspensions were split into two wells of a 96-well deepwell plate, and plasmid DNA was prepared following the QiaPrep 96 turbo miniprep kit protocol and recommendations for purifying low-copy plasmids. DNA concentrations were measured using the dsDNA Quantifluor System (Promega) and ~5-10 ng from each well was used to PCR amplify Cre-recombined ORF pairs using Phusion Polymerase (NEB), 1x GC buffer (NEB), 1 M betaine, 200 μM each dNTP, and 0.25 μM of AD and DB primers (**Supplemental Table 1**). Reactions were run at 98°C for 2 min, 21 cycles of

98°C for 10 s, 65°C for 30 s, and 72°C for 90 s, then a final extension at 72°C for 7 min. 5 μL of each PCR reaction was run on a 1% agarose gel and showed a DNA smear corresponding to the size range expected for Cre-recombined products (~1 kb to > 4 kb). Amplicons from each PCR reaction were pooled, isopropanol precipitated, and purified with SeraMag Speedbeads (GE; 2% v/v SeraMag Speedbeads, 18% w/v PEG-8000, 1 M NaCl, 10 mM Tris HCl, 1 mM EDTA) at a 1:1 bead-to-DNA ratio to remove primers, typically yielding ~2 μg of DNA. Illumina sequencing libraries were then prepared following the exact same steps as previously mentioned for the starting bait and prey libraries.

Pilot sequencing test to determine optimal sequencing depth. The same sequencing library from one CrY2H-seq screen was sequenced to a read depth of 20 million (20M) and 80 million (80M) reads. We observed that interactions with at least three distinct identifying fragments in 20M showed an expected increase in coverage of about 4x at 80M, while those with less than 3 fragments in 20M were not consistently reproducible (**Supplementary Fig. 4)**. We therefore established a cutoff requiring at least three fragments for a PPI to be included in a screen data set. Moreover, since deeper sequencing predominantly revealed PPIs represented by less than three fragments (i.e., below our cutoff), we concluded that 20 million reads was sufficient and aimed for 40 million reads per screen library.

Sequencing of CrY2H-seq screen libraries. Libraries were sequenced with an Illumina paired-end 200 cycle Rapid Run on an Illumina HiSeq 2500 platform. The total paired reads obtained from sequencing was 583 M, equivalent to the output of 1.65 Rapid Run flowcells.

Next-generation sequence analysis of CrY2H-seq screen libraries

Reads were mapped using Bowtie2-2.0.2[43] local alignment with default settings to a custom genome composed of *Arabidopsis* TF-coding sequences from TAIR10, the *Saccharomyces cerevisiae* genome, Gal4 AD and Gal4 DB domain sequences, and the empty CrY2H-seq plasmid sequences (**Supplementary Fig. 5a**). A quality filter was applied that required reads to map with at least 30 matching bases, allowing a maximum of two mismatches, two insertions or deletions, and two bases of trimming from the beginning of the read (**Supplementary Fig. 5b**). Reads were then joined with their corresponding read pairs and included in the next analysis step only if both reads passed the first filter and mapped to *Arabidopsis* TF ORF sequences. Clonal fragments were removed from read pairs if both reads in a fragment contained the same start positions. Paired reads for which each of the mates aligned to a different ORF and showed the same strand orientation (Cre recombination occurs such that ORFs on pADlox and pDBlox plasmids become inverted in a 3'-to-3' orientation; **Supplementary Fig. 5c**) were included in further analysis. Fragments were further subjected to a size filter that required that the sum of the lengths of each read (start position of each read to the end of each ORF) and the *lox* region conformed to the expected library size of 400-600bp (**Supplementary Fig. 5c**). Remaining fragments that mapped to Cre-recombined ORF junctions were totaled (**Supplementary Fig. 5e)**. Each screen had on average ~1.4 million fragments corresponding to ORF junction sites and ~16 million fragments mapping to gene bodies. Remaining data mapped to priming site region ORF junctions or did not align. Analysis scripts can be found in **Supplementary Software**. After applying the basal fragment cutoff mentioned above to all data sets (**Supplementary Fig. 5f**), fragments were normalized by the median filtered fragments as follows. A scale factor for each replicate data set was determined by dividing the filtered protein interaction fragments by the median filtered protein interaction fragments. The number of fragments per protein pair was multiplied by this scale factor and rounded down to the nearest integer to normalize protein interaction fragments (**Supplementary Fig. 5g)**.

Identification and removal of self-activating bait proteins. Any TF found to be linked with an empty pADlox plasmid by the mapping pipeline was labeled self-activating and was not included in AtTFIN-1. A list of proteins identified as self-activating can be found in **Supplementary Table 3.**

Bait and prey orientation analysis of AtTFIN-1 interaction fragments. As the double-mutant *lox* sequence from Cre-recombined plasmids is not a full palindrome, the middle region can be used to determine bait and prey orientations of interacting proteins (**Supplementary Fig. 15a**). An analysis script was written to assess the bases at this middle region for fragments where at least one read mapped to one ORF and 15 base pairs into the *lox77* sequence (**Supplementary Software**). It should be noted that the region of the read being mapped to *lox77* was within the last 10bp of the read where sequencing quality is known to be low on account of the nature of sequencing by synthesis. Of fragments mapping to non-self-activating PPIs, 5.5% (9,662,266/14,588,892) could identify bait and prey orientations of 49.71% of (4,264/8,577) AtTFIN-1 pairs (**Supplementary Fig. 15a,b; Supplementary Table 2c**). We acknowledge that this is a partial analysis and more data would be needed to confirm the bait and prey orientations for all pairs in AtTFIN-1.

Estimating CrY2H-seq screen saturation

To estimate CrY2H-seq screening saturation (the number of interactions detected out of the number of interactions CrY2H-seq could detect for this ORF collection), we simulated results for all possible orderings (10!) for the ten replicate screens. We calculated the average number and s.d. of interactions detected at each step, considering all possible orderings (**Fig. 2.3c**). We built a model based on the average new interaction detection rate after each replicate and fit it to a Michaelis-Menten curve to predict the number of interactions detectable by CrY2H-seq after any number of screens (**Supplementary Software** and **Supplementary Fig. 7**).

Yeast two-hybrid retest

A set of 950 interaction pairs that showed a range of screen occurrences and NPIFs was selected for use in a retest assay carried out using standard 1 x 1 array-style HT-Y2H methods. Clones corresponding to interaction pairs were cherry picked from pAD-lox and pDB-lox plasmid stock plates and freshly transformed into yeast strains Y8800 and CRY8930 as described above. 771 yeast transformant pairs were recovered that could be screened in both bait and prey orientations (**Supplementary Table 4)**. This ensured that both orientations in which the interaction could have been initially detected were accounted for. A Y2H screening pipeline was followed as described previously[18], including inoculation of individual AD and DB yeast cultures, 1 x 1 mating onto YEPD medium, replica-plating onto selective SC –Leu, -Trp for diploid selection, and replica-plating onto selective SC –Leu, -Trp, -His +1mM 3-AT plates and SC -Leu, -His +1mM 3-AT plates containing 1mg/L cycloheximide. Cycloheximide containing plates select for cells that do not have the AD plasmid due to plasmid shuffling and can identify spontaneous self-activators[18]. After replica-plating onto SC –Leu, -Trp, -His +1mM 3-AT, plates were incubated at 30°C overnight, then replica-cleaned by placing each plate on a piece of velvet stretched over a replica-plating block and pressing evenly to remove excess yeast cells. Plates were incubated an additional 3 d at 30°C and phenotypes were independently scored by two researchers (for representative colonies and scoring, refer to **Supplementary Fig. 8a**). Only pairs scored as positive for *HIS3* reporter gene activation and negative for growth on cycloheximide by both researchers were considered positive interactions in the retest assay. 115 pairs (~15%) activated the *HIS3* reporter gene and showed growth on cycloheximide. These interactions were scored as self-activating and not included in subsequent analysis of the retest data set.

wNAPPA assay

TFs corresponding to 59 novel interactions that showed a range of screen occurrences and NPIFs were selected for validation in the wNAPPA assay. Additionally, 35 previously reported protein interactions that were present in At-TFIN-1 and 36 random interactions not present in AtTFIN-1 were also processed in parallel. Clones were cherry picked from TF entry clone stock plates and recombined into pIX-GST and pIX-HA destination vectors[3] using LR clonase. Reactions were transformed into DH5α-T1^R, and plasmid DNA was purified using QiaPrep 96 Turbo kits. Plasmid DNA was measured using the Quantifluor dsDNA System and a Tecan SafireII plate reader. DNA was concentrated to roughly 250 ng/μL and 1 μg of each plasmid was combined for use *in vitro* transcription-translation reactions as follows. Bait and prey proteins were co-expressed using the TNT SP6 Coupled Wheat Germ Extract System (Promega) following manufacturer recommendations. Protein expression reactions were then added to anti-GST antibody-coated detection plates (GE Healthcare, cat. no. 27-4592-01) and incubated at 15°C for 2 h. Wells were washed and blocked with 1x PBS with 0.1% Tween and 5% nonfat dry milk (PBS/T/NFM) for 1 h at room temperature, then incubated with mouse anti-HA monoclonal antibody (Covance, cat. no. MMS-101R) diluted 1:5,000 in PBS/T/NFM for 1 h at room temperature. Antibody was washed from wells with PBS/T/NFM with three quick washes followed by three longer washes, each with a five-min room temperature incubation period with gentle rotation. Wells were then incubated with anti-mouse HRP-coupled secondary antibody (GE

Healthcare, cat. no. NA931) diluted 1:2,000 in PBS/T/NFM for 1 h at room temperature. Secondary antibody was washed from the wells with PBS/T with three quick washes followed by three 5-min washes. Wells were rinsed twice with 1x PBS before adding Supersignal ELISA Femto substrate (Pierce), and then incubated for 1.5 min at room temperature with gentle shaking. Luminescence (RLU) was measured using a Tecan SafireII plate reader. Interactions were tested in both vector combinations and observed *z*-scores are listed in **Supplementary Table 5a**.

To control for plate-to-plate variation, a set of 16 pairs previously used for normalization[3] (**Supplementary Table 5b)** was included on each plate. Plate normalization and scoring were done according to previously described methods[3]. Briefly, for each plate the normalization pair average and s.d. were calculated after subtracting the average blank (empty pIX GST and empty HA plasmid mix) and taking the $\log_2$ RLU value. A *z*-score for each well was then calculated by first subtracting the normalization pair average from the RLU value and then dividing by the normalization pair s.d. To determine the recall rates, the maximum *z*-score of the two orientations tested for each pair was considered, and a scoring threshold was determined by maximizing for the number of positively scoring known interactions and minimizing for the number of positively scoring random interactions (**Supplementary Fig. 9**). A scoring threshold of 1.6 was selected based on these criteria.

Literature, database, and randomly generated data comparison with AtTFIN-1

Literature and database interaction data files were downloaded from links listed in **Supplementary Table 6**, and all interactions between TFs screened in CrY2H-seq were compiled. Interactions from different sources showed some overlap, but also many unique interactions (**Supplementary Fig. 10b**). For this reason, comparisons were made between AtTFIN-1 and individual data sets (**Fig. 2.4c**). Only high-confidence STRING and AraNet interactions with scores above 900 and 4.5 were used. To generate random TF interactions, a list of all possible combinations was first generated. From this list, 8,577 interactions were selected randomly using the script in **Supplementary Software**. This step was done a total of ten times to produce ten random interaction data sets. From each of these data sets, we excluded homodimers and interactions with TFs detected as self-activating in the CrY2H-seq screens. Comparisons between AtTFIN-1 and each list were performed and the average overlap was reported (**Fig. 2.4c**). **Supplementary Figure 10b** was generated using the web interface provided by VIB/University of Ghent Bioinformatics and Evolutionary Genomics Division, Belgium (http://bioinformatics.psb.ugent.be/webtools/Venn/). The precision-recall curve (**Fig. 2.4d**) was generated using the R package PRROC[44].

Preferential family-specific interaction analysis

The R package igraph[45] was used to generate randomly rewired interactions from a list of high-confidence AtTFIN-1 interactions using the rewire function with degree conservation. The gene IDs in the subsequent list of random interactions were converted into family names, sorted and family interactions were counted. This was done 10,000 times. The high-confidence AtTFIN-1 interactions were similarly converted to family names, and family interactions were counted. The AtTFIN-1 family-interaction observations were then compared to the 10,000 random observations and *P* values were calculated based on where the AtTFIN-1 family interaction observation occurred in the empirical distribution of all observations for each family interaction. Heatmaps (**Fig. 2.5** and **Supplementary Fig. 13**) were generated using the R package, Heatmap3[46]. Interaction networks (**Fig. 2.5** and **2.6)** were generated using Cytoscape[47].

Cost and time comparisons to existing HT-Y2H methods

Traditional Y2H and BFG-Y2H cost approximations (**Supplementary Fig. 14**) are based on appendix figure S4 in Yachie *et al.*[12]. Costs for traditional Y2H were calculated on a per-plate basis assuming minipools of 50 preys and assuming the recovery of 500 and 10,000 positive interactions from 1,000,000, and 900,000,000 PPIs screened, respectively. CrY2H-seq sequencing costs are estimated from 1 Illumina HiSeq Rapid PE Sequencing Run (cluster kit and 200 cycle kit) costing $3,126, and yielding on average 350,000,000 reads.

Statistics

Exact *n* values are reported in main text and legends for **Figure 2.4a-c**, and **Supplementary Figure 8b, 11, and 12**. For **Figure 2.4b,c** and **Supplementary Figure 12**, a one-sided Fisher's exact test was done to compare the detection rates of known and novel interactions to random interactions. For **Figure 2.5a and Supplementary Figure 13**, empirical *P* values were calculated by ranking the observed family-interaction frequency among frequencies generated from 10,000 different degree-conserved network rewirings.

Materials availability

CrY2H-seq plasmids and yeast strains are available through the *Arabidopsis* Biological Resources Center, https://abrc.osu.edu/ (stock numbers CD3-2420, CD3-2421, CD5-239, and CD5-240).

Code availability

Code generated for analysis during this study is available as **Supplementary Files 1-4**.

Data availability

Protein interaction data from this study are included in **Supplementary Tables 2-5** and have been submitted to the IMEx (http://www.imexconsortium.org) consortium through IntAct accession number

IM-25740; and interactome visualization can be found at http://signal.salk.edu/interactome/AtTFIN-1.html. Raw read data files and alignment indexes have been submitted to the Sequence Read Archive (https://www.ncbi.nlm.nih.gov/sra) through SRA accession number SRX2825531. Source data files for the figures in the study are available from the authors upon request.

AUTHOR CONTRIBUTIONS

J.R.E. conceived the project. S.A.T., R.M.G., R.O., M.G., and J.R.E. designed and/or advised research. S.A.T., R.M.G., A.M., J.R.N., A.B., R.C., A.G., and M.G. performed experiments. S.A.T. established bioinformatics pipelines and performed computational analysis with contributions from J.F., R.O., S.C.H., and Z.Z.Z. S.A.T., M.G., and J.R.E. prepared the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

ACKNOWLEDGEMENTS

This material is based upon work supported by US Department of Energy grant DOE-DE SC0007078 (to J.R.E.), and National Science Foundation grants IOS-1650227 (to J.R.E.), IOS1456950 and IOS1546873 (to M.G.), and the Graduate Research Fellowship Program under grant number DGE-1650112 (to S.A.T.). J.R.E. is an Investigator of the Howard Hughes Medical Institute. S.A.T.

is supported in part by the Mary K. Chapman Foundation. We thank M. Hofree, B. Haas, S. Navlakha, A.R. Carvunis, H. Carter, and T. Ideker for network analysis advice. S. Heinz, J. Chory, J. Law, L. Song, H. Chen, Y. He, M. Hariharan, B. Kellman, J. Reyna, L. Gai, and V. Lundblad lab members for advice and discussion; J. Pruneda-Paz (UCSD, California) for the TF ORF collection; H. Yu (Cornell University, New York) for pDEST-AD and pDEST-DB plasmids; and D. Hill (CCSB DFCI, Massachusetts) for Y8930 and Y8800 yeast strains.

Chapter 2, in full, is a reformatted reprint of the material as it appears in the August 2017 issue of the journal *Nature Methods*. Trigg SA, Garza RM, MacWilliams A, Nery JR, Bartlett A, Castanon R, Goubil A, Feeney J, O'Malley R, Huang SC, Zhang ZZ, Galli M, and Ecker JR. 2017. CrY2H-seq: a massively multiplexed assay for deep-coverage interactome mapping. *Nature Methods*, August 2017. The dissertation author was the primary investigator and author of this material.

REFERENCES

1. 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. & Vidal, M. Next-generation sequencing to generate interactome datasets. *Nat. Methods* **8,** 478–480 (2011).

2. Rolland, T., Taşan, M., Charloteaux, B., Pevzner, S. J., Zhong, Q., Sahni, N., Yi, S., Lemmens, I., Fontanillo, C., Mosca, R., Kamburov, A., Ghiassian, S. D., Yang, X., Ghamsari, L., Balcha, D., Begg, B. E., Braun, P., Brehme, M., Broly, M. P., Carvunis, A. R., Convery-Zupan, D., Corominas, R., Coulombe-Huntington, J., Dann, E., Dreze, M., Dricot, A., Fan, C., Franzosa, E., Gebreab, F., Gutierrez, B. J., Hardy,

M. F., Jin, M., Kang, S., Kiros, R., Lin, G. N., Luck, K., Macwilliams, A., Menche, J., Murray, R. R., Palagi, A., Poulin, M. M., Rambout, X., Rasla, J., Reichert, P., Romero, V., Ruyssinck, E., Sahalie, J. M., Scholz, A., Shah, A. A., Sharma, A., Shen, Y., Spirohn, K., Tam, S., Tejeda, A. O., Trigg, S. A., Twizere, J. C., Vega, K., Walsh, J., Cusick, M. E., Xia, Y., Barabási, A. L., Iakoucheva, L. M., Aloy, P., De Las Rivas, J., Tavernier, J., Calderwood, M. A., Hill, D. E., Hao, T., Roth, F. P. & Vidal, M. A proteome-scale map of the human interactome network. *Cell* **159,** 1212–1226 (2014).

3. Arabidopsis Interactome Mapping Consortium. Evidence for Network Evolution in an Arabidopsis Interactome Map. *Science (80-.).* **333,** 601–607 (2011).

4. Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A. & Tyers, M. BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* **34,** D535-9 (2006).

5. Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K. P., Kuhn, M., Bork, P., Jensen, L. J. & Von Mering, C. STRING v10: Protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43,** D447–D452 (2015).

6. Lee, T., Yang, S., Kim, E., Ko, Y., Hwang, S., Shin, J., Shim, J. E., Shim, H., Kim, H., Kim, C. & Lee, I. AraNet v2: An improved database of co-functional gene networks for the study of Arabidopsis thaliana and 27 other nonmodel plant species. *Nucleic Acids Res.* **43,** D996–D1002 (2015).

7. Wang, X., Wei, X., Thijssen, B., Das, J., Lipkin, S. M. & Yu, H. Three-dimensional reconstruction of protein networks provides insight into human genetic disease. *Nat. Biotechnol.* **30,** 159–64 (2012).

8. Hofree, M., Shen, J. P., Carter, H., Gross, A. & Ideker, T. Network-based stratification of tumor mutations. *Nat. Methods* **10,** 1108–1115 (2013).

9. Jiang, Z., Dong, X. & Zhang, Z. Network-Based Comparative Analysis of Arabidopsis Immune Responses to Golovinomyces orontii and Botrytis cinerea Infections. *Scientific reports* **6,** 19149 (2016).

10. 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., Barabási, A.-L. & Vidal, M. An empirical framework for binary interactome mapping. *Nat. Methods* **6,** 83–90 (2009).

11. Weimann, M., Grossmann, A., Woodsmith, J., Özkan, Z., Birth, P., Meierhofer, D., Benlasfer, N., Valovka, T., Timmermann, B., Wanker, E. E., Sauer, S. & Stelzl, U. A Y2H-seq approach defines the human protein methyltransferase interactome. *Nat. Methods* **10,** 339–42 (2013).

12. Yachie, N., Petsalaki, E., Mellor, J. C., Weile, J., Jacob, Y., Verby, M., Ozturk, S. B., Li, S., Cote, A. G., Mosca, R., Knapp, J. J., Ko, M., Yu, A., Gebbia, M., Sahni, N., Yi, S., Tyagi, T., Sheykhkarimli, D., Roth, J. F., Wong, C., Musa, L., Snider, J., Liu, Y.-C., Yu, H., Braun, P., Stagljar, I., Hao, T., Calderwood, M. A., Pelletier, L., Aloy, P., Hill, D. E., Vidal, M. & Roth, F. P. Pooled-matrix protein interaction screens using Barcode Fusion Genetics. *Mol. Syst. Biol.* **12,** 863 (2016).

13. Hastie, A. R. & Pruitt, S. C. Yeast two-hybrid interaction partner screening through in vivo Cre-mediated Binary Interaction Tag generation. *Nucleic Acids Res.* **35,** (2007).

14. Pruneda-Paz, J. L., Breton, G., Nagel, D. H., Kang, S. E., Bonaldi, K., Doherty, C. J., Ravelo, S., Galli, M., Ecker, J. R. & Kay, S. A. A Genome-Scale Resource for the Functional Characterization of Arabidopsis Transcription Factors. *Cell Rep.* **8,** 622–632 (2014).

15. Oberdoerffer, P., Otipoby, K. L., Maruyama, M. & Rajewsky, K. Unidirectional Cre-mediated genetic inversion in mice using the mutant loxP pair lox66/lox71. *Nucleic Acids Res.* **31,** e140 (2003).

16. Stynen, B., Tournu, H., Tavernier, J. & Van Dijck, P. Diversity in genetic in vivo methods for protein-protein interaction studies: from the yeast two-hybrid system to the mammalian split-luciferase system. *Microbiol. Mol. Biol. Rev.* **76,** 331–82 (2012).

17. Serebriiskii, Ilya G. and Golemis, E. A. *Two-hybrid systems — methods and protocols. Methods in Molecular Biology* **177,** (Humana Press, 2001).

18. Dreze, M., Monachello, D., Lurin, C., Cusick, M. E., Hill, D. E., Vidal, M. & Braun, P. High-quality binary interactome mapping. *Methods Enzymol.* **470,** 281–315 (2010).

19. 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. & Vidal, M. An experimentally derived confidence score for binary protein-protein interactions. *Nat. Methods* **6,** 91–97 (2009).

20. He, F., Yoo, S., Wang, D., Kumari, S., Gerstein, M., Ware, D. & Maslov, S. Large-scale atlas of microarray data reveals the distinct expression landscape of different tissues in Arabidopsis. *Plant J.* **86,** 472–80 (2016).

21. Debernardi, J. M., Mecchia, M. A., Vercruyssen, L., Smaczniak, C., Kaufmann, K., Inze, D., Rodriguez, R. E. & Palatnik, J. F. Post-transcriptional control of GRF transcription factors by microRNA miR396 and GIF co-activator affects leaf size and longevity. *Plant J.* **79,** 413–426 (2014).

22. Cho, W. K., Lian, S., Kim, S. M., Seo, B. Y., Jung, J. K. & Kim, K. H. Time-course RNA-Seq analysis reveals transcriptional changes in rice plants triggered by rice stripe virus infection. *PLoS One* **10,** (2015).

23. Guilfoyle, T. J. The PB1 domain in auxin response factor and Aux/IAA proteins: a versatile protein interaction module in the auxin response. *Plant Cell* **27,** 33–43 (2015).

24. Mukhtar, M. S., Carvunis, A.-R., Dreze, M., Epple, P., Steinbrenner, J., Moore, J., Tasan, M., Galli, M., Hao, T., Nishimura, M. T., Pevzner, S. J., Donovan, S. E., Ghamsari, L., Santhanam, B., Romero, V., Poulin, M. M., Gebreab, F., Gutierrez, B. J., Tam, S., Monachello, D., Boxem, M., Harbort, C. J., Mcdonald, N., Gai, L., Chen, H., He, Y., Vandenhaute, J., Roth, F. P., Hill, D. E., Ecker, J. R., Vidal, M., Beynon, J., Braun, P., Dangl, J. L., Union, E., Consortium, E., Vandenhaute, J., Roth, F. P., Hill, D. E., Ecker, J. R., Vidal, M. & Beynon, J. Independently Evolved Virulence Effectors Converge onto Hubs in a Plant Immune System Network. *Science (80-.).* **333,** 596–601 (2011).

25. Villarino, G. H., Hu, Q., Manrique, S., Flores-Vergara, M. A., Sehra, B., Robles, L., Brumos, J., Stepanova, A. N., Colombo, L., Sundberg, E., Heber, S. & Franks, R. G. Temporal and spatial domain-specific transcriptomic analysis of a vital reproductive meristem in Arabidopsis thaliana. *Plant Physiol.* **171,** 42–61 (2016).

26. Madmon, O., Mazuz, M., Kumari, P., Dam, A., Ion, A., Mayzlish-Gati, E., Belausov, E., Wininger, S., Abu-Abied, M., McErlean, C. S. P., Bromhead, L. J., Perl-Treves, R., Prandi, C., Kapulnik, Y. & Koltai, H. Expression of MAX2 under SCARECROW promoter enhances the strigolactone/MAX2 dependent response of Arabidopsis roots to low-

phosphate conditions. *Planta* 1–9 (2016). doi:10.1007/s00425-016-2477-7

27. Sun, L., Song, L., Zhang, Y., Zheng, Z. & Liu, D. Arabidopsis PHL2 and PHR1 Act Redundantly as the Key Components of the Central Regulatory System Controlling Transcriptional Responses to Phosphate Starvation. *Plant Physiol.* **170,** 499–514 (2016).

28. Gangappa, S. N. & Botto, J. F. The BBX family of plant transcription factors. *Trends in Plant Science* **19,** 460–471 (2014).

29. Preuss, S. B., Meister, R., Xu, Q., Urwin, C. P., Tripodi, F. A., Screen, S. E., Anil, V. S., Zhu, S., Morrell, J. A., Liu, G., Ratcliffe, O. J., Reuber, T. L., Khanna, R., Goldman, B. S., Bell, E., Ziegler, T. E., McClerren, A. L., Ruff, T. G. & Petracek, M. E. Expression of the Arabidopsis thaliana BBX32 gene in soybean increases grain yield. *PLoS One* **7,** (2012).

30. Huang, H. & Bader, J. S. Precision and recall estimates for two-hybrid screens. *Bioinformatics* **25,** 372–378 (2009).

31. Llorca, C. M., Berendzen, K. W., Malik, W. A., Mahn, S., Piepho, H. P. & Zentgraf, U. The elucidation of the interactome of 16 arabidopsis bZIP factors reveals three independent functional networks. *PLoS One* **10,** (2015).

32. Century, K., Reuber, T. L. & Ratcliffe, O. J. Regulating the regulators: the future prospects for transcription-factor-based agricultural biotechnology products. *Plant Physiol.* **147,** 20–9 (2008).

33. Snider, J., Kittanakom, S., Curak, J. & Stagljar, I. Split-ubiquitin based membrane yeast two-hybrid (MYTH) system: a powerful tool for identifying protein-protein interactions. *J. Vis. Exp.* e1698 (2010). doi:10.3791/1698

34. Petschnigg, J., Groisman, B., Kotlyar, M., Taipale, M., Zheng, Y., Kurat, C. F., Sayad, A., Sierra, J. R., Usaj, M. M., Snider, J., Nachman, A., Krykbaeva, I., Tsao, M.-S., Moffat, J., Pawson, T., Lindquist, S., Jurisica, I. & Stagljar, I. The mammalian-membrane two-hybrid assay (MaMTH) for probing membrane-protein interactions in human cells. *Nat. Methods* **11,** 585–592 (2014).

35. Wilson, T. E., Fahrner, T. J., Johnston, M. & Milbrandt, J. Identification of the DNA binding site for NGFI-B by genetic selection in yeast. *Science* **252,** 1296–1300 (1991).

36. Lumba, S., Toh, S., Handfield, L. F., Swan, M., Liu, R., Youn, J. Y.,

Cutler, S. R., Subramaniam, R., Provart, N., Moses, A., Desveaux, D. & McCourt, P. A mesoscale abscisic acid hormone interactome reveals a dynamic signaling landscape in arabidopsis. *Dev. Cell* **29,** 360–372 (2014).

37. Cao, S., Siriwardana, C. L., Kumimoto, R. W. & Holt, B. F. Construction of high quality Gateway^{TM} entry libraries and their application to yeast two-hybrid for the monocot model plant *Brachypodium distachyon*. *BMC Biotechnol.* **11,** 53 (2011).

38. Benatuil, L., Perez, J. M., Belk, J. & Hsieh, C. M. An improved yeast transformation method for the generation of very large human antibody libraries. *Protein Eng. Des. Sel.* **23,** 155–159 (2010).

39. Trigg, S.A., Garza, R.M., MacWilliams, A., Nery, J.R., Bartlett, A., Castanon, R., Goubil, A., Feeney, J., O'Malley, R., Huang, S.C., Zhang, Z.Z., Galli, M., and Ecker, J. R. CrY2H-seq interactome screening. *Protoc. Exch.* (2017). doi:10.1038/protex.2017.058

40. Abremski, K. & Hoess, R. Bacteriophage P1 site-specific recombination. Purification and properties of the Cre recombinase protein. *J. Biol. Chem.* **259,** 1509–1514 (1984).

41. Ehlert, A., Weltmeier, F., Wang, X., Mayer, C. S., Smeekens, S., Vicente-Carbajosa, J. & Dröge-Laser, W. Two-hybrid protein-protein interaction analysis in Arabidopsis protoplasts: Establishment of a heterodimerization map of group C and group S bZIP transcription factors. *Plant J.* **46,** 890–900 (2006).

42. Albert, H., Dale, E. C., Lee, E. & Ow, D. W. Site-specific integration of DNA into wild-type and mutant lox sites placed in the plant genome. *Plant J.* **7,** 649–659 (1995).

43. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat Methods* **9,** 357–359 (2012).

44. Keilwagen, J., Grosse, I., Grau, J., Zucchini, W., Baldi, P., Brunak, S., Chauvin, Y., Andersen, C., Nielsen, H., Stehman, S., Fawcett, T., Lobo, J., Jiménez-Valverde, A., Real, R., Bleakley, K., Biau, G., Vert, J., Zheng, S., Yuille, A., Tu, Z., Sonnenburg, S., Schweikert, G., Philips, P., Behr, J., Ratsch, G., Alexiou, P., Maragkakis, M., Papadopoulos, G., Reczko, M., Hatzigeorgiou, A., Poulter, G., Rubin, D., Altman, R., Seoighe, C., Ohsaki, M., Abe, H., Tsumoto, S., Yokoi, H., Yamaguchi, T., Prechelt, L., Malpohl, G., Philippsen, M., Kekäläinen, J., Järvelin, K., Martin, D., Fowlkes, C., Malik, J., Grau, J., Posch, S., Grosse, I., Keilwagen, J., Lockhart, D., Winzeler, E., Yeo, G., Burge, C., Brodersen,

K., Ong, C., Stephan, K., Buhmann, J., Weirauch, M., Cote, A., Norel, R., Annala, M., Zhao, Y., Grau, J., Keilwagen, J., Gohr, A., Haldemann, B. & Posch, S. Area under Precision-Recall Curves for Weighted and Unweighted Data. *PLoS One* **9,** e92209 (2014).

45. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal Complex Syst.* 1695 (2006). doi:10.1109/ICCSN.2010.34

46. Zhao, S., Guo, Y., Sheng, Q. & Shyr, Y. Heatmap3: an improved heatmap package with more powerful and convenient features. *BMC Bioinformatics* **15,** P16 (2014).

47. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. & Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13,** 2498–504 (2003).

CHAPTER 3:
UNCOVERING MECHANISMS OF GLOBAL OCEAN CHANGE EFFECTS ON DUNGENESS CRAB (*Cancer Magister*) THROUGH METABOLOMICS ANALYSIS

Uncovering mechanisms of global ocean change effects on the Dungeness crab (*Cancer magister*) through metabolomics analysis

Shelly A. Trigg[1]*, Paul McElhany[2], Michael Maher[2], Danielle Perez[2], D. Shallin Busch[2,3], and Krista M. Nichols[2]

[1]Division of Biological Sciences, Cell and Developmental Biology Section, University of California San Diego, La Jolla, California USA

[2]Conservation Biology Division, Northwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, Seattle, Washington, USA

[3]Ocean Acidification Program, Office of Oceanic and Atmospheric Research and Northwest Fisheries Science Center, National Marine Fisheries Service, National Oceanic and Atmospheric Administration, Seattle, Washington, USA

*Correspondence should be addressed to S.A.T. (swanamak@ucsd.edu)

ABSTRACT

The Dungeness crab is an economically and ecologically important species distributed along the North American Pacific coast. It is not currently known how sensitive Dungeness crab will be to the ocean chemistry conditions predicted with global climate change. To investigate this, we used untargeted metabolomics approaches to characterize metabolite and lipid profiles in individual Dungeness crab juveniles reared in treatments that mimicked current and projected future pH and dissolved oxygen conditions. Despite limited metabolome annotation and high variation within treatment groups, we identified 94 metabolites and 127 lipids that respond in a treatment-specific manner. We show that exposure to low dissolved oxygen leads to overall greater fluctuations in known compound abundances than exposure to low pH, and that low pH affects fewer known compounds. We found through pathway analysis that juveniles may be generally responding to low dissolved oxygen through evolutionarily conserved processes including downregulating glutathione biosynthesis and upregulating glycogen storage, and may be generally responding to low pH by increasing ATP production. Most interestingly, we found that under combined low pH and low dissolved oxygen stress juveniles respond most similarly to the low oxygen single stress treatment, indicating low dissolved oxygen more significantly affects the physiology of juvenile crabs than pH. Our study elucidates metabolic flexibility and adaptations that expand our overall understanding of how the species might respond to future ocean conditions.

INTRO

The continued increase in anthropogenic carbon dioxide emissions is leading to ocean acidification, with the ocean absorbing an average of 25% of human-caused emissions annually[1]. If atmospheric carbon dioxide concentration continues to rise at the current rate, the pH of oceans is predicted to fall 0.3-0.4 units by the end of the century[2,3]. This pH drop could exacerbate conditions in the U.S. Pacific Northwest, where the ocean pH is lower than that of the global ocean due to natural oceanographic processes including regional upwelling, and could pose a greater challenge to the marine inhabitants already coping with this lower pH. An additional and compounding factor to ocean acidification is ocean deoxygenation, which co-varies with pH and temperature. Given that global ocean temperatures will rise with global warming from continued greenhouse gas emissions, hypoxic zones are expected to increase in duration, intensity, and frequency[4]. It is not certain how future ocean acidification and deoxygenation environmental stress might affect important Pacific Northwest fisheries like the Dungeness crab fishery, which is the most lucrative and valued at more than $200 million annually[5].

Ocean acidification is predicted to have negative indirect effects on Dungeness crab through loss of prey directly affected by ocean acidification[6], but knowledge of how ocean acidification might directly impact Dungeness crab is limited. It has been postulated that Dungeness crab may exhibit limited tolerance for acid-base disturbances given past observations of Dungeness crab as weak osmoregulators[7–10] and that acid-base balance and osmoregulation are tightly coupled in decapod crustaceans[11]. However, it was shown that following a brief two-week exposure to a future-predicted seawater pH of 7.4, adult Dungeness crab are able to acclimate by increasing hemolymph ion levels (bicarbonate, calcium, chloride, sulfate, and sodium), and by decreasing both oxygen consumption and nitrogen excretion[12]. One the other hand, Dungeness crab larvae have shown reduced survival and development rate in response to a 45-day exposure to the future-predicted seawater pH of 7.5 and 7.1[13]. Moreover, adult Dungeness crab have shown behavioral responses to declining oxygen conditions (21-1.5 kPa pO_2 over a 5-hour period), including reduced feeding and a preference for the area with the highest pO_2 level when placed in a seawater oxygen gradient (2.5-10.5 kPa pO_2 for 1 hour)[14]. Dungeness crab also have shown physiological responses to declining oxygen conditions (18-3 kPa pO_2 over a 6-hour period), including redistributing hemolymph to high-energy-demand tissues[15]. Despite these clear biological responses to low pH and oxygen, the biochemical mechanisms underlying Dungeness crab response to combined pH and oxygen stress have not yet been defined.

To gain broad insight into the biochemical processes underlying the physiological status of organisms, it is now possible to survey metabolomes of nearly any species with the advancement of high-throughput metabolite profiling, also known as metabolomics. Exploratory untargeted metabolomics approaches can offer unbiased analyses of the composition of all detectable metabolites for the rapid and quantitative detection of stress responses, which often leads to the development of targeted approaches and identification of stress-indicating biomarkers[16,17]. Functional analyses using pathway inference can subsequently be performed by integrating metabolomics data with databases and other “-omics” datasets using bioinformatics tools to ultimately establish causal networks between different experimental conditions and outcomes. Untargeted metabolomics in the context of understanding ocean acidification has recently been applied to reef-building coral, and revealed metabolite profiles that were predictive of primary production activity and molecules that could be used as potential biomarkers of ocean acidification[18].

To better understand which biochemical pathways might be altered in the response of Dungeness crab to ocean acidification, we applied untargeted metabolomics and lipidomics to individual juvenile crabs exposed to current pH (7.8) and future pH (7.4) conditions for an average of 32 days. Because dissolved carbon dioxide and dissolved oxygen (DO) concentrations tend to co-vary in ocean habitiats[4], we included both ambient oxygen (8.5 mg/L or 80% O_2 saturation) and low oxygen (3 mg/L or 30% O_2 saturation) treatments with our pH conditions in a factorial design to understand how pH, oxygen, and/or their interaction might influence metabolite abundances.

RESULTS

Untargeted lipidomics and metabolomics were carried out on 60 individual juvenile crabs exposed to ocean acidification treatments (**Table 3.1**) through the entire duration of their first juvenile instar until 2 days after molting to their second juvenile instar. From the lipidomics and metabolomics profiles generated by the West Coast Metabolomics Core using in-house open-access bioinformatics pipelines[19,20] (see **Methods** for additional details), a total of 3113 lipids and 651 general metabolites were detected of which 88% (3320/3764 total detected compounds) had spectra that did not match LipidBlast[21] or BinBase[22] records and were classified as unknown compounds (**Supplementary Table 1 and 2**). Eighteen lipids were detected in fewer than 50% of individual profiles and were therefore excluded from all downstream analyses. The MS identification techniques found 29/281 metabolite classes among the 246 compounds detected and listed in Human Metabolome Database (HMDB)[23] and 13/77 lipid classes among 195 compounds detected and listed in LIPID MAPS Structural Database[24] (**Supplementary Fig. 1**). Among the largest classes represented in all identified compounds were LIPID MAPS classes glycerophosphocholines, triradylglycerols, and fatty acids, and HMDB classes organic acids, carboxylic acids, and organic oxygen compounds.

In general, individual metabolite and lipid profiles showed high variation regardless of treatment group. Clustering the relative abundances of all compounds across treatment groups revealed no obvious patterns (**Supplementary Fig. 2**). Examining the distributions of coefficients of variation for individual compounds across treatment groups showed collective compound variation was generally not treatment dependent, and that individual crabs elicit diverse and dynamic responses regardless of treatment (**Supplementary Fig. 3**). To identify important features, we applied both univariate and multivariate statistical methods to combine the strengths of different methods[25]. First, to assess individual effects as well as interaction effects from pH and DO treatments on individual compounds, we used a two-way analysis of variance (ANOVA) on the raw abundance data (**Supplementary Table 1 and 2**). Prior to performing ANOVA, we verified compounds showed homogeneity of variances across treatment groups (>96% compounds showed a Levene test P value > 0.05), and that metabolite and lipid abundances were mostly (on average 69% compounds showed a Shapiro-Wilks test P value > 0.05) normally distributed within treatment groups (**Supplementary Table 3**). Although about one-third of compounds violated the ANOVA normality assumption, ANOVA can be considered robust to violations of this assumption when datasets have more than 10 samples per treatment group[26,27]. Of all compounds analyzed, 56/651 metabolites (including 24/160 known metabolites) and 98/3095 lipids (including 7/284 known lipids) showed overall model significance at $P < 0.1$ and at least one model term (pH, DO, and/or pH x DO interaction) significant at $P < 0.05$ (**Supplementary Fig. 4 and 5**). However, when a Benjamini-Hochberg P value correction for multiple testing was applied to overall model P values (**Supplementary Table 4**), no metabolite or lipid showed significance failing to pass its respective 10% false discovery rate Benjamini-Hochberg P value threshold of 1×10^{-4} and less than 3×10^{-5}.

To identify discriminatory features and model the relationship between metabolite profiles and exposure treatments using a multivariate linear regression approach, we used partial least square discriminatory analysis (PLS-

DA). For facilitating comparisons of metabolite composition among treatment groups, compound abundances were centered around the mean and scaled by the reference group (ambient pH:ambient DO) compound standard deviations prior to multivariate analysis[28]. The PLS-DA model for both the metabolite and lipid data showed partial separation of low pH-treated samples from ambient pH-treated samples in the first and second components, accounting for a total of 31% and 15% of the variation in the metabolite and lipid data, respectively (**Fig. 3.1a-b, Supplementary Table 5**). Additionally, the PLS-DA model for the metabolite data showed a separation of low DO-treated samples in the second and third components, explaining an additional 6.8% of variation, but the PLS-DA model for the lipid data did not show this (**Supplementary Fig. 6**). Because the overall PLS-DA model had weak predictive power (**Supplementary Fig. 8**), we used importance thresholds defined by points of diminishing returns in PLS-DA component loadings plots (**Supplementary Fig 7.** and **Methods**), and identified a total of 45 metabolites (including 14/160 known metabolites) and 18 unknown lipids as important discriminatory features.

Finally, we used a multivariate nonlinear-based supervised random forest classification, using the metabolomics and lipidomic data to predict treatment classes (**Supplementary Fig. 9a-b**). A total of 9 metabolites (including 4/160 known metabolites) and 15 lipids (including 1/284 known lipids) were considered important features based on their mean decrease in prediction accuracy (**Supplementary Fig. 9c-d**, **Supplementary Table 6**, and **Methods**). While no compounds were commonly identified by all 3 statistical methods, 7 metabolites were commonly identified by ANOVA and PLS-DA, 8 metabolites were in common between ANOVA and random forest, 1 metabolite overlapped between random forest and PLS-DA, and 4 lipids overlapped between ANOVA and random forest (**Fig. 3.2a-b**). Because these three statistical methods are fundamentally different approaches, we did not expect a high degree of overlap in important features that each method identified. We did expect to capture a comprehensive set of important features that no one method was capable of capturing on its own, and our set of important features identified by combining all three statistical methods was consistent with that.

Heatmaps of the statistically selected 94 metabolites (including 35 known) and 127 lipids (including 7 known), shows several compounds respond to low pH and oxygen stress treatments by commonly increasing or decreasing abundance relative to the ambient treatment (**Fig. 3.3a-b**). For known compounds, this is also summarized by violin plots of compound abundances in **Figure 3.4a-c**. For example, 5-methoxytryptamine, butyrolactam, cysteine, cysteine, homoserine, pipecolinic acid, piperidone, ribose, and xanthine commonly showed a decrease in abundance in treatments with low DO relative to treatments with ambient DO, where glutamic acid, maltose, and maltotriose commonly showed an increase in abundance in treatments with low DO relative to treatments with ambient DO (**Figure 3.4a**). In general, metabolites tend to decrease in abundance in low DO treatments relative to ambient DO treatments exemplified by the mostly blue color of compound abundance averages for the low DO treatment group (green) in **Figure 3.3a.** Specific to the low pH treatment, more metabolites show average abundances similar to ambient treatment signifying pH has a less dramatic effect on metabolites compared to low DO treatment (**Fig. 3.3a-b and Fig. 3.4a-b**). This is also apparent in the combined low pH, low DO treatment, where metabolites and lipids show average abundances similar to the low DO treatment signifying low DO has a more dominant effect on compounds than low pH (**Fig. 3.3a-b**).

To explore the physiological relevance of treatment responsive compounds, we performed biochemical pathway analysis on all known compounds. While compounds classified as sugars and fatty acids were affected by the low pH and DO treatments, most of the significantly affected compounds were part of the amino acids class, indicating amino acid metabolism was most significantly altered by the treatments. Abbreviated biochemical pathway networks focusing on affected amino acids are shown in **Figure 3.5**, and the complete biochemical pathway networks affected by low pH, low DO, and low pH and DO treatments can be found in **Supplementary Figures 10-12.** The trends in amino acid abundances resulting from either low pH, low DO, or the combined low pH and DO treatment (**Fig. 3.5a-c**), suggest different amino acid metabolic pathways are affected in response to each factor or combination of factors.

Specific to low DO treatment response compared to the normoxic treatment, energy conservation pathways were most apparent. The increased lysine with decreased pipecolinic acid and the piperidine derivative piperidone abundance is suggestive of downregulation of the lysine degradation pathway[29,30]. The decreased abundance of cysteine and the cysteine homodimer cystine suggests that cysteine catabolism is upregulated. This coupled with the increased abundance of glutamic acid suggests the glutathione synthesis pathway could be downregulated, as both cysteine and glutamic acid are precursor molecules in glutathione synthesis[31,32]. Glutathione itself was not detected, and can typically be challenging to detect by MS techniques in marine animals due to its reactivity[33]. The glycogen intermediates maltose and maltotriose show an increase in abundance consistent with previously observed glycogen synthesis pathway upregulation during hypoxic stress[34]. Purine and pyrimidine metabolic intermediates orotic acid, ribose, and xanthine show decreased abundance suggesting purine and pyrimidine metabolic pathways could be downregulated[35].

Specific to low pH treatment response compared to ambient pH treatment, ATP generation pathways were most apparent. The decreased abundance of aspartic acid and the subtle decreased abundance of maleic acid may suggest the citric acid cycle intermediates that they form (oxaloacetate, malic acid, and fumaric acid)[36,37] are favored. Although, oxaloacetate was not detected (likely due to the instability of the alpha keto acid compound)[38], and malic acid and fumaric acid did not show a significant difference across treatments. However, citric acid shows an increase in abundance and taken all together suggest the citric acid cycle activity could be upregulated in response to low pH[39]. An increase in glutaric acid in response to low pH suggests the catabolism of its parent molecule, glutaryl CoA, which produces ATP in addition to glutaric acid[40]. Glutaryl CoA catabolism supports citric acid cycle activity because glutaryl CoA can uncompetitively inhibit the alpha-ketoglutarate dehydrogenase complex that facilitates a rate limiting step in the citric acid cycle[41].

Among the compounds affected by low pH and low DO treatment, phosphoethanolamine and phophatidylethanolamine (PE(p-34:2) or PE(o-34:3)) show increased abundance compared to ambient treatment (**Supplementary Fig. 11)**, suggesting that low pH and DO might alter the glycerophospholipid synthesis pathway of which phosphatidulethanolamine is a product and phosphoethanolamine is a substrate intermediate in a rate limiting step[42]. Alanine shows a decrease in abundance, suggesting that alanine synthesis is decreased and that its substrate for synthesis, pyruvate, may be limited[43]. The increased abundance of cystathionine and subtle increase of methionine suggest cysteine and homoserine synthesis is likely stalled[44] in response to both low pH and low DO. This supports the decrease in homoserine abundance and upregulation of cysteine catabolism observed in low DO treatment.

DISCUSSION

We used metabolomics to explore how the Dungeness crab might respond to the simultaneous change in oxygen and carbon dioxide in predicted future climate change scenarios in the Pacific Northwest. Although the untargeted metabolomics and lipidomics approaches rapidly identified hundreds of different molecular species, the majority of compounds identified in our study, including most compounds selected by statistical methods, lack annotations in existing metabolite databases. Thus, our understanding of their role in response to low pH and low DO treatment is limited without extensive further validation. Still, through shedding light on the simultaneous activity of hundreds of known compounds, we were able to observe a dynamic range of metabolic responses among individuals within treatment groups that indicates that Dungeness crab have flexibility in how their biochemistry compensates for environmental change. While we attempted to control for variation between individuals by collecting animals from the same location within a 2-month period, we were unable to control for prior environmental exposure or genetic background of the wild-caught animals we used. Where prior studies have found high genetic variation among individuals within one sampling site[45,46], we suspect that both genetics and prior environmental exposure likely contributed to the high variation in metabolite abundances among individuals within treatment groups, which may have obscured the different treatment effects.

We applied three different statistical methods (ANOVA, PLS-DA, and random forest) for selecting important metabolites in order to combine the strengths of powerful univariate and multivariate analyses[25]. Although using univariate statistics in the strictest sense with an FDR to control for false discoveries showed no statistically significant variation for metabolites across treatments, not all compounds were identified with the same confidence level and were subject to the sensitivity of the mass spectrometry method used.

Applying an FDR correction to all detected compounds assumes that all compounds had equal chance for discovery, which can mask important biology in highly variable datasets[47]. For this reason and due to our use of wild-caught animals, we chose a more liberal focus on all compounds showing an overall ANOVA model P value < 0.1 and at least one model term P value of < 0.05, or identified as important in PLS-DA or random forest models in our pathway analysis.

While we acknowledge a level of uncertainty in our pathway analysis due to liberal feature selection criteria, the resulting proposed affected pathways are consistent with previous observations of pH and hypoxia effects on different organisms. In general, amino acid metabolism in general is a well-documented mechanism for stress tolerance[48,49]. This class of compounds contains versatile chemical structures that serve as buffering molecules, antioxidants, signaling molecules, and chemical building blocks for the synthesis of proteins important in stress response (i.e., heat shock proteins, unfolded protein response proteins, ion channels). Under low oxygen conditions, it is in the best interest of the animal to limit non-essential energy consuming pathways[50]. One way Dungeness crab may do this is through reducing the activity of evolutionarily conserved gamma-glutamyl cycle that synthesizes glutathione and consumes ATP[51,52] by limiting cysteine availability through catabolism. Interestingly, glutathione reduction in response to hypoxia has been observed in multiple mammalian cell lines[53–56]. Also under low oxygen conditions in nature, food can be scarce, and it is theorized that glycogen storage during low oxygen can help prepare cells for low nutrient conditions. In multiple mouse and human cancer cell lines[34,57,58], glycogen storage has been observed in response to hypoxia, induced by highly evolutionarily conserved hypoxia-inducible factor transcriptional signalling[34]. It seems that Dungeness crab may also adopt this strategy in combating low oxygen conditions. Under low pH, adult Dungeness crab initially develop hypercapnia which then abates over time via elevated hemolymph bicarbonate likely generated through gill restructuring and the upregulation of energy consuming ion-exchange proteins[12]. Our results indicate that this process may occur in early juveniles in low pH conditions given that we found metabolite profiles that support energy generation via increased citric acid cycle activity.

Having parsed out effects from low oxygen and low pH, we found that in combined low oxygen and low pH conditions, low oxygen has a more dramatic effect on metabolite abundance. However, the animals in this treatment group were still able to alter their metabolism to allow for upregulation of citric-acid-cycle-related metabolites and potentially cell respiration. It is not yet clear what the longer-term consequences are of these metabolic adjustments or how long these responses could be sustained. Future avenues of research to expand on these findings should include targeted metabolomics to confirm the compounds identified in this study as well as capture more antioxidants and citric acid cycle intermediates that were not detected in this study. Targeted expression profiling would also be helpful in confirming the biochemical pathway activity predictions from this study. Ultimately, longer-term exposure experiments on crabs reared over generations might best reveal the maximal duration of these metabolic responses and any long-term consequences of low pH and oxygen exposure. This exploratory metabolomics and lipidomics analysis uncovered potential biochemical pathways affected by experiments simulating ocean acidification and hypoxia, and can now serve as preliminary hypotheses for deeper investigations of how the Dungeness crab (and even other crustaceans) may tolerate global ocean change.

Table 3.1. Summary of experimental treatments.

treatment	number of crabs	time (days)	temp (◦C)	pH	DO (mg/L)
ambient pH and DO	15	31.33 ± 1.11	12	7.8	8.5
ambient pH and low DO	15	31.57 ± 0.94	12	7.8	3
low pH and ambient DO	15	31.07 ± 0.88	12	7.4	8.5
low pH and low DO	15	31.73 ± 0.96	12	7.4	3

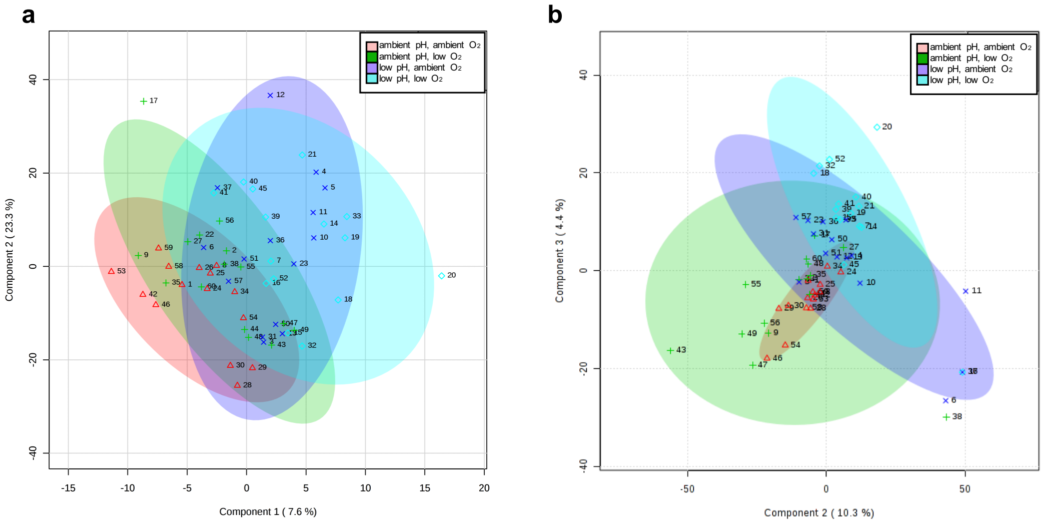

Figure 3.1. Summary of PLS-DA analysis. Partial separation of low pH treatment groups from low DO treatment groups shown by (a) PLS-DA components 1 and 2 for metabolite data and (b) PLS-DA components 2 and 3 for lipid data.

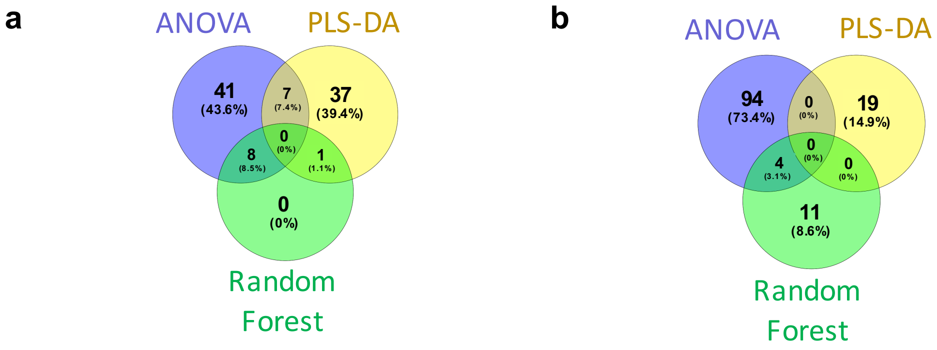

Figure 3.2. Overlap of important features identified by different statistical methods. Venn diagrams of all (**a**) metabolites and (**b**) lipids identified as significantly changed by ANOVA, PLS-DA, and/or random forest.

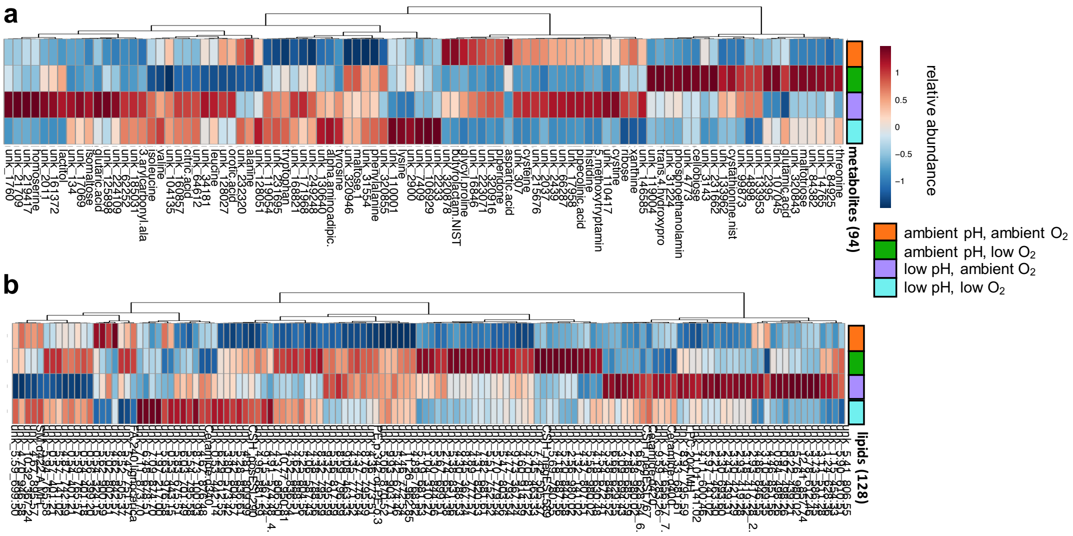

Figure 3.3. Treatment-specific effects on the relative abundance of compounds selected by multivariate and univariate statistical methods. Heatmap plots show the average compound abundance (average peak intensity) for each treatment group for the (**a**) 94 general metabolites and (**b**) 127 lipids selected by multivariate and univariate statistical methods used to evaluate treatment effects. Individual known and unknown ("unk") compound names are listed over the columns and treatment groups are listed over the rows (orange, ambient pH, ambient O_2; green, ambient pH, low O_2; purple, low pH, ambient O_2; and blue, low pH, low O_2). Compound average abundances are shown as auto-scaled within each compound ("relative abundance").

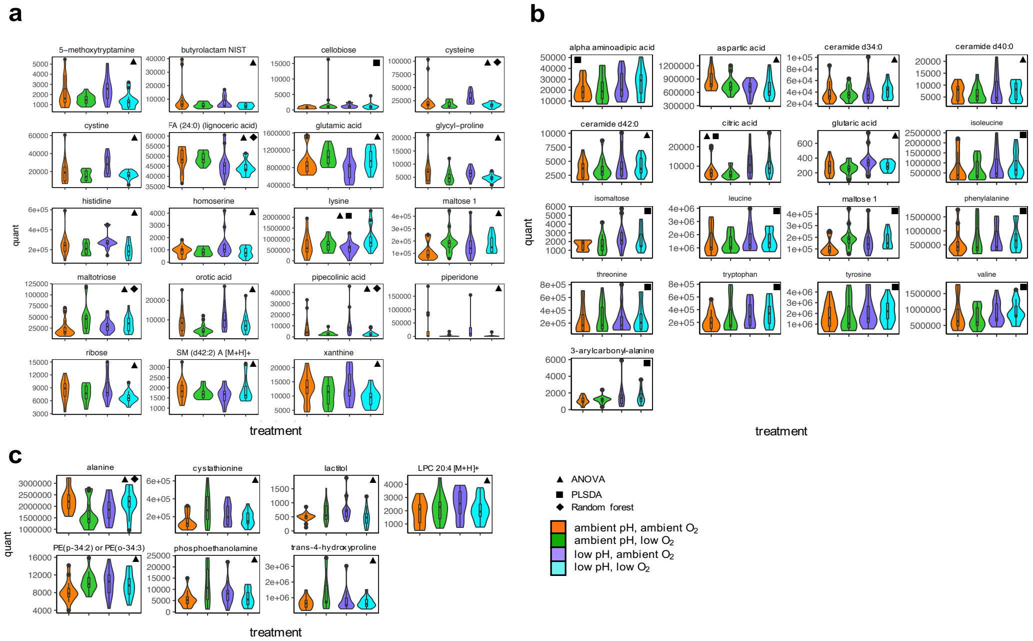

Figure 3.4. Abundance levels (peak intensities) of known compounds selected by univariate and multivariate statistical methods. Violin plots with boxplot insets show the distribution of compound abundances within each treatment group for compounds statistically showing (**a**) DO treatment effect, (**b**) pH treatment effect, and (**c**) DO:pH interaction effect. Binbase names for each known compound are listed across the top of each plot. Abundance levels (peak intensities, noted as “quant”) are listed on the y-axis while treatments are listed along the x-axis. Statistical methods that each known compound was identified by are noted by shapes in the upper corners of each plot (ANOVA, triangle; PLS-DA, square; and random forest, diamond).

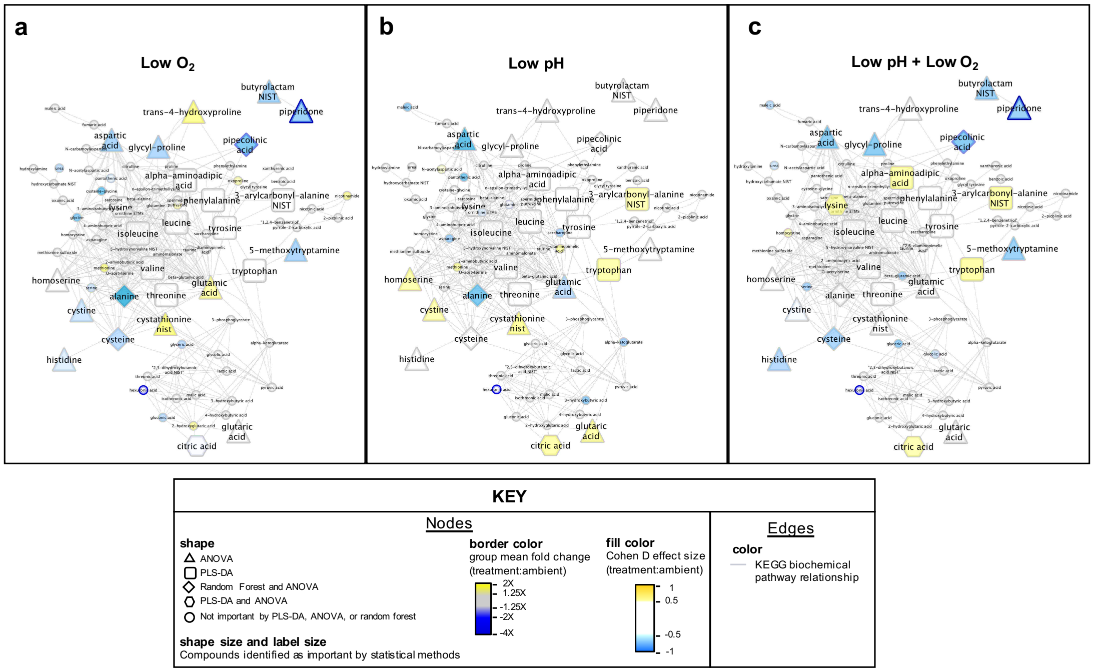

Figure 3.5. Treatment-specific differences in amino acid metabolic networks. Networks show effects from (**a**) low DO, (**b**) low pH, and (**c**) combined low pH and low DO. Compounds are clustered by chemical similarity. Node fill color is colored by Cohen D effect size comparing the treatment group to the ambient pH,ambient DO group, and node borders are colored by the treatment group mean fold change relative to the ambient pH,ambient DO group. Node shapes indicate the statistical method(s) from which the compound was classified as important. Node shape and label size are enlarged if the compound was identified as important by a statistical method. Gray edges indicate nodes sharing a KEGG biochemical pathway(s).

METHODS

Animals

Cancer magister megaolopae were collected from a single site in Puget Sound (47.950232, -122.301784) on several days between the June and September 2016 using light traps that were set overnight. The contents of each trap were immediately transferred into a 5-gallon bucket that was, within 5 minutes, pooled into a cooler with ice packs and an air bubbler. After transferring contents from 7 traps in approximately 1 hour, the cooler containing megalopae was brought within 5 minutes into the lab and megalopae were individually transferred onto the water flow-through system described below. The total time for transferring megalopae from the cooler onto the seawater flow-through system was about 2 hours.

CO_2 exposure experiments

Megalopae were held in individual 250 mL customized jars on Mobile Ocean Acidification Treatment Systems (MOATS). These systems flowed one-micron-filtered, UV-sterilized, Puget Sound seawater maintained at 12°C. Prior to flowing through jars, seawater was degassed and oxygen, nitrogen and carbon dioxide gases were resupplied to finely control dissolved gas levels. Temperature, pH, and dissolved oxygen were continuously monitored throughout the duration of the experiment by Omega thermistors, Honeywell Durafet III probes, and Vernier optical dissolved oxygen probes, respectively. The pH was additionally validated by periodic sampling of water for dissolved inorganic carbon and total alkalinity, and by bi-weekly spectrophotometric pH measurements using an Ocean Optics USB 230 2000+ Fiber Optic Spectometer with SpectraSuite software and a 5mM solution of Sigma Aldritch m-cresol purple indicator dye. MOATS chemistry parameters were automatically adjusted through a data-driven feedback system. Megalopae were fed *Artermia salina* (San Francisco Bay brand) at a target concentration of 1 nauplius per mL every 3 days. Once megalopae transitioned to stage one juveniles, they were fed small pieces of squid. Upon transitioning to stage 2 juveniles, crabs were held on the MOATS for an additional 48 hours in attempt to reduce variation due to potential stochastic physiological processes associated with molting [59,60]. Stage 2 juveniles were then immediately frozen and stored at -80°C after lightly blotting with a paper towel to remove excess seawater. To reduce variation from length of exposure, which ranged from 20-65 days, only 60 crabs with the average exposure time (30-33 days) were chosen for metabolomics analysis with 15 individuals from each treatment group.

Sample preparation

Frozen samples were sent to the West Coast Metabolomics Center, Davis, CA for sample preparation, metabolomics and lipidomics profiling. Whole animals were thawed, weighed, and derivatized as previously described[61,62] Briefly, samples were extracted at -20°C with 2 mL of degassed acetonitrile/isopropanol/water (3:3:2) solution and solvents were evaporated to complete dryness with a Labconco Centrivap cold trap concentrator. Membrane lipids and triglycerides were subsequently removed from dried samples with 50% acetonitrile, and samples were again concentrated to complete dryness. 15 mg of each sample preparation was used for metabolomic profiling. For lipidomic profiling, 15 mg of the same sample preparation was also used to which internal standards, C8-C30 fatty acid methyl esters were added. Aliquoted samples were derivatized with methoxyamine hydrochloride (Sigma-Aldrich) in pyridine (Acros Organics) and then by N-methyl-N-(trimethylsilyl) trifluoroacetamide (Sigma-Aldrich) for trimethylsilylation of acidic protons.

Metabolite and lipid data acquisition

General metabolite and lipid abundances were quantified from derivatized samples by gas-chromatography, time-of-flight mass spectrometry (GC-TOF/MS) and charged-surface, hybrid-column, electrospray-quadrupole, time-of-flight mass spectrometry (CSH-ESI QTOF MS/MS), respectively. For metabolites, an Agilent 6890 gas chromatograph (Santa Clara, CA) was used with a Leco Pagasus IV time-of-flight mass spectrometer running Leco ChromaTOF software 2.32 (St. Joseph, MI). The following temperature profile was used: 50°C to 275°C final temperature at a rate of 12°C/s and hold for 3 minutes. Injection volume was 0.5 µl with 10 µl/s injection speed on a splitless injector with a purge time of 25 seconds. Liner (Gerstel #011711-010-00) was changed after every 10 samples, (using the Maestro1 Gerstel software vs. 1.1.4.18). Before and after each injection, the 10 µL injection syringe was washed 3 times with 10 µL ethyl acetate. For gas chromatography, a 30 m long, 0.25 mm i.d. Rtx-5Sil MS column (0.25 μm 95% dimethyl 5% diphenyl polysiloxane film) with additional 10 m integrated guard column was used (Restek, Bellefonte PA). 99.9999% pure Helium with a built-in purifier (Airgas, Radnor PA) was set at a constant flow of 1 mL/minute. The oven temperature was held constant at 50°C for 1 minute and then ramped at 20°C/minute to 330°C at which it is held constant for 5 minutes. The transfer line temperature between gas chromatograph and mass spectrometer was set to 280°C. Electron-impact ionization at 70V was employed with an ion source temperature of 250°C. Acquisition rate was 17 spectra/second, with a scan mass range of 85-500 Da.

For positively charged lipids, an Agilent 6530 QTOF mass spectrometer with resolution 10,000 was used and for negatively charged lipids, an Agilent 6550 QTOF mass spectrometer with resolution 20,000 was used. Electrospray ionization was used to ionize column elutants in both positive and negative modes. Compounds were separated using a Waters Acquity ultra-high-pressure, liquid-chromatography charged surface hybrid column C18 (100 mm length x 2.1 mm internal diameter; 1.7 um particles) using the following conditions: mobile phase A (60:40 acetonitrile:water + 10 mM ammonium formiate + 0.1% formic acid, mobile phase B (90:10 isopropanol:acetonitrile + 10 mM ammonium formiate + 0.1% formic acid), 65°C column temperature, a flow rate of 0.6 mL/minute, an injection volume of 3 uL, an injection temperature of 4 C, and a gradient of 0 minutes 15%, 0-2 minutes 30%, 2-2.5 minutes 48%, 2.5-11 minutes 82%, 11-11.5 minutes 99%, 11.5-12 minutes 99%, 12-12.1

minutes 15%, and 12.1-15 minutes 15%. The capillary voltage was set to +3.5 and -3.5 kV, and the collision energy to 25 or 40 eV for positive and negative modes. Mass-to-charge ratios (m/z) were scanned from 60 to 1700 Da and spectra acquired every 2 seconds.

Spectral data processing

Acquired metabolite data were processed using UC Davis's BinBase workflow, which performs data processing including peak detection at signal-to-noise levels of 5:1 throughout the chromatogram. Resulting apex masses are reported for use in the BinBase algorithm to facilitate metabolite identification and quantification[19]. Metabolites were identified through comparison to the BinBase database[22] and peak heights were normalized to total metabolite content[20].

For acquired lipid data, MassHunter (Qual v. B05.00) was used to find peaks from the raw data in up to 300 chromatograms. These peaks were imported into Mass Profiler Professional for alignment to determine which peaks occur in at least 30 % of the chromatograms. These peaks were then quantified with MassHunter. Resulting accurate mass data and tandem MS/MS spectra were compared to LipidBlast libraries for compound identification[21].

Statistical analyses

All statistical analyses were carried out in R, except for PLS-DA and random forest analyses which were carried out using the MetaboAnalyst web interface.

The coefficient of variation (standard deviation/mean) for the abundance of each compound was calculated for each treatment group. The standard deviation of each compound in a treatment group was calculated and then divided by the metabolite mean within the treatment group. The distribution of coefficient of variations was plotted for each treatment group to see if compound abundance variability showed a dependency on treatment (**Supplementary Fig. 3a-b**). A Kruskall-Wallis test was used to check for significant differences between treatment group coefficient of variation distributions. Heatmaps in **Figure 3.3** and **Supplementary Figure 2** were made using the heatmap function in the MetaboAnalyst web interface with Euclidean distance, ward clustering, plotting auto-scaled compound abundance averages for each treatment group. Normality and heteroscedasticity of compound abundances within treatment groups were tested using the Shapiro-Wilks test and the Levene test, respectively.

Univariate statistics

Excluding compounds that were detected in less than half of the individual crabs surveyed, two-way ANOVA was applied to each compound in the dataset. A Benjamini-Hochberg FDR correction was applied to ANOVA *P* values to correct for multiple testing, but corrected *P* values were not used in selecting important compounds. Metabolites and lipids were selected for pathway analyses if they had a raw overall model ANOVA *P* value less than 0.1 and an effect *P* value with less than 0.05 without an FDR correction applied.

Multivariate statistics

Prior to applying multivariate testing, data were normalized by mean-centering and scaling by the reference-group standard-deviation for each compound in order to make compounds more comparable to one another[28]. Normalized data was then uploaded to MetaboAnalyst with no further normalizations applied. The PLS-DA function was run with default settings and component loadings were exported from MetaboAnalyst (**Supplementary Table 5**). Loadings were plotted for PLS-DA components, and importance thresholds were placed at the point of diminishing returns in the plot curves (**Supplementary Fig. 7**). Specifically for metabolites, loadings were plotted PLS-DA components 1, 2, and 3 (**Supplementary Fig. 7a-c**) since these PLS-DA components gave the greatest separation between treatment groups (**Fig. 3.1a and Supplementary Fig. 6a**). The points of diminishing returns that importance thresholds were placed in the metabolite loadings plots were as follows: component 1, loadings threshold > 0.05; component 2, loadings threshold > 0.1; and component 3, loadings threshold > 0.05. Specifically for lipids, loadings were plotted for PLS-DA components 2 and 3 (**Supplementary Fig. 7d-e**) since these gave the best separation between treatment groups, while component 1 could only distinguish the ambient from the altered treatment groups (**Fig. 3.1b and Supplementary Fig. 6b**). The points of diminishing returns that importance thresholds were placed in the lipids loadings plots were as follows: component 2, loadings threshold > 0.05; and component 3, loadings threshold > 0.05. For the random forest analysis, the Random Forest function was run with 2000 decision trees and either 25 predictors for the metabolite dataset or 56 predictors for the lipid dataset, conforming to the default classification value being the square root of the number of variables[63]. The complete table of important variables and their mean decrease in random forest accuracy prediction exported from MetaboAnalyst (**Supplementary Table 6**). The mean decrease in prediction accuracy for each compound was ordered from largest to smallest and plotted (**Supplementary Fig. 9c-d**). Importance thresholds were drawn from the plots at the points of diminishing returns, which for metabolites was a mean decrease in accuracy > 0.001 and for lipids was a mean decrease in accuracy > 0.00047.

Pathway analysis

For visualizing the pH, DO, and pH:DO interaction effects, mean abundance fold change and Cohen D effect size values were calculated for each compound within a treatment group relative to the ambient treatment group. Cohen D effect size ($\boxed{\frac{\mu_{treatment}-\mu_{ambient}}{s_{pooled}}}$) was calculated for each compound raw abundance within a treatment group relative to the ambient treatment group using the R package effsize[64]. A table was then generated with treatment group -$\log_2$ fold change values, Cohen D effsize output from R, and corresponding -$\log_{10}$ *P* values from the ANOVA analysis for each metabolite (**Supplementary Table 7**). Pubchem and KEGG identifiers were obtained for metabolites and lipid international chemical identifier keys provided by the West Coast Metabolomics Core using Chemical Translation Service[65] and Pubchem Identifier exchange[66]. Metamapp[67] was then used to map metabolites by chemical and biochemical relationships. The generated SIF file (**Supplementary File 1**) and modified node attribute file (**Supplementary Table 7**) were imported into Cytoscape[68]. Low pH, low DO, and low pH:low DO networks were stylized using the following settings: node shapes were set according to statistical method used to select the compound (ANOVA, triangle; PLS-DA, square; random forest, diamond; both PLS-DA and ANOVA, hexagon; not selected as important by any method, circle); node border was color by the treatment group mean compound abundance fold change relative to ambient (1.25X – 2X fold change, yellow gradient; -1.25X – 1.25X fold change, grey; -1.25X – -4X fold change, blue gradient); node fill was colored by the Cohen D effect size of treatment on compound abundance relative to ambient (effect size of 0.5 – 1, yellow gradient; -0.5 – 0.5, white; and -0.5 – -1, blue gradient); node shape size and label size were set to fixed values (100 height and width with size 50 font for statistically selected compounds, and 40 height and width with size 12 font for compounds not identified by statistics (**Supplementary Fig. 10-12)**.

AUTHOR CONTRIBUTIONS

K.M.N., P.M., and D.S.B conceived the project. K.M.N., P.M., and D.S.B. advised research. M.M. and D.P. performed experiments. S.A.T. performed statistical analyses and pathway analysis. S.A.T. prepared the manuscript with edits from K.M.N., P.M, and D.S.B.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

ACKNOWLEDGEMENTS

This material is based upon work supported by the Ocean Acidification Program and Northwest Fisheries Science Center of the National Oceanic and Atmospheric Administration, and in part by the National Science Foundation Graduate Research Fellowship Program and Graduate Research Internship Program (to S.A.T.). S.A.T. is supported in part by the Mary K. Chapman Foundation. We thank the current and former members of the Ocean Acidification lab at the Northwest Fisheries Science Center Mukilteo Research Station, members of the West Coast Metabolomics Core and Fiehn lab at University of California Davis.

Chapter 3, in full, consists of the following manuscript in preparation for submission to the journal *Scientific Reports* in the format of an article.

Shelly A. Trigg, Paul McElhany, Michael Maher, Danielle Perez, D. Shallin Busch, and Krista M. Nichols. 2018. Uncovering mechanisms of global ocean change effects on the Dungeness crab (*Cancer magister*) through metabolomics analysis. The dissertation author was the primary investigator and author of this material.

REFERENCES

1. Le Quéré, C., Andrew, R. M., Canadell, J. G., Sitch, S., Ivar Korsbakken, J., Peters, G. P., Manning, A. C., Boden, T. A., Tans, P. P., Houghton, R. A., Keeling, R. F., Alin, S., Andrews, O. D., Anthoni, P., Barbero, L., Bopp, L., Chevallier, F., Chini, L. P., Ciais, P., Currie, K., Delire, C., Doney, S. C., Friedlingstein, P., Gkritzalis, T., Harris, I., Hauck, J., Haverd, V., Hoppema, M., Klein Goldewijk, K., Jain, A. K., Kato, E., Körtzinger, A., Landschützer, P., Lefèvre, N., Lenton, A., Lienert, S., Lombardozzi, D., Melton, J. R., Metzl, N., Millero, F., Monteiro, P. M. S., Munro, D. R., Nabel, J. E. M. S., Nakaoka, S. I., O'Brien, K., Olsen, A., Omar, A. M., Ono, T., Pierrot, D., Poulter, B., Rödenbeck, C., Salisbury, J., Schuster, U., Schwinger, J., Séférian, R., Skjelvan, I., Stocker, B. D., Sutton, A. J., Takahashi, T., Tian, H., Tilbrook, B., Van Der Laan-Luijkx, I. T., Van Der Werf, G. R., Viovy, N., Walker, A. P., Wiltshire, A. J. & Zaehle, S. Global Carbon Budget 2016. *Earth Syst. Sci. Data* **8,** 605–649 (2016).

2. Caldeira, K. & Wickett, M. E. Anthropogenic carbon and ocean pH. *Nature* **425,** 365 (2003).

3. Feely, R. A., Sabine, C. L., Lee, K., Berelson, W., Kleypas, J., Fabry, V. J. & Millero, F. J. Impact of anthropogenic CO2 on the CaCO3 system in the oceans. *Science (80-.).* **305,** 362–366 (2004).

4. Gobler, C. J. & Baumann, H. Hypoxia and acidification in ocean ecosystems: coupled dynamics and effects on marine life. *Biol. Lett.* **12,** (2016).

5. National Marine Fisheries Service. *Fisheries of the United States*. *Fisheries of the United States, 2016 report.* (2017). https://www.fisheries.noaa.gov/resource/document/fisheries-united-states-2016-report.

6. Marshall, K. N., Kaplan, I. C., Hodgson, E. E., Hermann, A., Busch, D. S., McElhany, P., Essington, T. E., Harvey, C. J. & Fulton, E. A. Risks of ocean acidification in the California Current food web and fisheries: ecosystem model projections. *Glob. Chang. Biol.* **23,** 1525–1539 (2017).

7. Hunter, K. C. & Rudy, P. P. Osmotic and ionic regulation in the

Dungeness crab, *Cancer magister* dana. *Comp. Biochem. Physiol. Part A* **51,** 439–447 (1975).

8. Engelhardt, F. R. & Dehnel, P. A. Ionic regulation in the Pacific edible crab, *Cancer magister* (Dana). *Can. J. Zool.* **51,** 735–743 (1973).

9. Freire, C. A., Onken, H. & McNamara, J. C. A structure-function analysis of ion transport in crustacean gills and excretory organs. *Comparative Biochemistry and Physiology - A Molecular and Integrative Physiology* **151,** 272–304 (2008).

10. Jones, L. L. Osmotic regulation in several crabs of the pacific coast of north america. *J. Cell. Comp. Physiol.* **18,** 79–92 (1941).

11. Henry, R. P. & Wheatly, M. G. Interaction of respiration, ion regulation, and acid-base balance in the everyday life of aquatic crustaceans. *Integr. Comp. Biol.* **32,** 407–416 (1992).

12. Hans, S., Fehsenfeld, S., Treberg, J. R. & Weihrauch, D. Acid-base regulation in the Dungeness crab (Metacarcinus magister). *Mar. Biol.* **161,** 1179–1193 (2014).

13. Miller, J. J., Maher, M., Bohaboy, E., Friedman, C. S. & McElhany, P. Exposure to low pH reduces survival and delays development in early life stages of Dungeness crab (*Cancer magister*). *Mar. Biol.* **163,** 118 (2016).

14. Bernatis, J. L., Gerstenberger, S. L. & McGaw, I. J. Behavioural responses of the Dungeness crab, *Cancer magister*, during feeding and digestion in hypoxic conditions. *Mar. Biol.* **150,** 941–951 (2007).

15. Airriess, C. & Mcmahon, B. Cardiovascular adaptations enhance tolerance of environmental hypoxia in the crab *Cancer magister*. *J. Exp. Biol.* **190,** (1994).

16. Turi, K. N., Romick-Rosendale, L., Ryckman, K. K. & Hartert, T. V. A review of metabolomics approaches and their application in identifying causal pathways of childhood asthma. *Journal of Allergy and Clinical Immunology* (2016). doi:10.1016/j.jaci.2017.04.021

17. Lankadurai, B. P., Nagato, E. G. & Simpson, M. J. Environmental metabolomics: an emerging approach to study organism responses to environmental stressors. *Environ. Rev.* **21,** 180–205 (2013).

18. Sogin, E. M., Putnam, H. M., Anderson, P. E. & Gates, R. D. Metabolomic signatures of increases in temperature and ocean acidification from the reef-building coral, Pocillopora damicornis.

Metabolomics **12,** (2016).

19. Fiehn, O., Wohlgemuth, G. & Scholz, M. Setup and Annotation of Metabolomic Experiments by Integrating Biological and Mass Spectrometric Metadata. in 224–239 (Springer, Berlin, Heidelberg, 2005). doi:10.1007/11530084_18

20. Fiehn, O., Wohlgemuth, G., Scholz, M., Kind, T., Lee, D. Y., Lu, Y., Moon, S. & Nikolau, B. Quality control for plant metabolomics: reporting MSI-compliant studies. *Plant J.* **53,** 691–704 (2008).

21. Kind, T., Liu, K.-H., Lee, D. Y., DeFelice, B., Meissen, J. K. & Fiehn, O. LipidBlast in silico tandem mass spectrometry database for lipid identification. *Nat. Methods* **10,** 755–8 (2013).

22. Skogerson, K., Wohlgemuth, G., Barupal, D. K. & Fiehn, O. The volatile compound BinBase mass spectral database. *BMC Bioinformatics* **12,** 321 (2011).

23. Wishart, D. S., Feunang, Y. D., Marcu, A., Guo, A. C., Liang, K., Vázquez-Fresno, R., Sajed, T., Johnson, D., Li, C., Karu, N., Sayeeda, Z., Lo, E., Assempour, N., Berjanskii, M., Singhal, S., Arndt, D., Liang, Y., Badran, H., Grant, J., Serra-Cayuela, A., Liu, Y., Mandal, R., Neveu, V., Pon, A., Knox, C., Wilson, M., Manach, C. & Scalbert, A. HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.* **46,** D608–D617 (2018).

24. Fahy, E., Sud, M., Cotter, D. & Subramaniam, S. LIPID MAPS online tools for lipid research. *Nucleic Acids Res.* **35,** W606–W612 (2007).

25. Grissa, D., Pétéra, M., Brandolini, M., Napoli, A., Comte, B. & Pujos-Guillot, E. Feature Selection Methods for Early Predictive Biomarker Discovery Using Untargeted Metabolomic Data. *Front. Mol. Biosci.* **3,** 30 (2016).

26. Witte, R. S. & Witte, J. S. *Statistics*. (J. Wiley & Sons, 2010).

27. Vinaixa, M., Samino, S., Saez, I., Duran, J., Guinovart, J. J. & Yanes, O. A Guideline to Univariate Statistical Analysis for LC/MS-Based Untargeted Metabolomics-Derived Data. *Metabolites* **2,** 775–95 (2012).

28. Timmerman, M. E., Hoefsloot, H. C. J., Smilde, A. K. & Ceulemans, E. Scaling in ANOVA-simultaneous component analysis. *Metabolomics* **11,** 1265–1276 (2015).

29. Broquist, H. P. Lysine-Pipecolic Acid Metabolic Relationships in Microbes and Mammals. *Annu. Rev. Nutr.* **11,** 435–448 (1991).

30. Cheng, J., Chen, C., Kristopher, K. W., Manna, S. K., Scerba, M., Friedman, F. K., Luecke, H., Idle, J. R. & Gonzalez, F. J. Identification of 2-piperidone as a biomarker of CYP2E1 activity through metabolomic phenotyping. *Toxicol. Sci.* **135,** 37–47 (2013).

31. Orlowski, M. & Meister, A. The gamma-glutamyl cycle: a possible transport system for amino acids. *Proc. Natl. Acad. Sci. U. S. A.* **67,** 1248–55 (1970).

32. Bodnaryk, R. P. Structure and Function of Insect Peptides. *Adv. In Insect Phys.* **13,** 69–132 (1978).

33. Mika, A., Skorkowski, E. & Stepnowski, P. The Use of Different MS Techniques to Determine Glutathione Levels in Marine Tissues. *Food Anal. Methods* **6,** 789–802 (2013).

34. Pelletier, J., Bellot, G., Gounon, P., Lacas-Gervais, S., Pouyssegur, J. & Mazure, N. M. Glycogen Synthesis is Induced in Hypoxia by the Hypoxia-Inducible Factor and Promotes Cancer Cell Survival. *Front Oncol* **2,** 18 (2012).

35. Christman, A. A. Purine and pyrimidine metabolism. *Physiol. Rev.* **32,** 303–48 (1952).

36. Winefield, C. S., Farnden, K. J. F., Reynolds, P. H. S. & Marshall, C. J. Evolutionary analysis of aspartate aminotransferases. *J. Mol. Evol.* **40,** 455–463 (1995).

37. Scher, W. & Jakoby, W. B. Maleate isomerase. *J. Biol. Chem.* **244,** 1878–1882 (1969).

38. Mamer, O., Gravel, S.-P., Choinière, L., Chénard, V., St-Pierre, J. & Avizonis, D. The complete targeted profile of the organic acid intermediates of the citric acid cycle using a single stable isotope dilution analysis, sodium borodeuteride reduction and selected ion monitoring GC/MS. *Metabolomics* **9,** 1019–1030 (2013).

39. Engelking, L. R. & Engelking, L. R. Tricarboxylic Acid (TCA) Cycle. in *Textbook of Veterinary Physiological Chemistry* 208–213 (Elsevier, 2015). doi:10.1016/B978-0-12-391909-0.50034-7

40. Menon, G. K. K., Friedman, D. L. & Stern, J. R. Enzymic synthesis of glutaryl-coenzyme A. *Biochim. Biophys. Acta* **44,** 375–377 (1960).

41. Sauer, S. W., Okun, J. G., Schwab, M. A., Crnic, L. R., Hoffmann, G. F., Goodman, S. I., Koeller, D. M. & Kölker, S. Bioenergetics in glutaryl-coenzyme A dehydrogenase deficiency: a role for glutaryl-coenzyme A.

J. Biol. Chem. **280,** 21830–6 (2005).

42. Gibellini, F. & Smith, T. K. The Kennedy pathway-De novo synthesis of phosphatidylethanolamine and phosphatidylcholine. *IUBMB Life* **62,** n/a-n/a (2010).

43. Mathews, C. K., Van Holde, K. E. (Kensal E. & Ahern, K. G. *Biochemistry*. (Benjamin Cummings, 2000).

44. Courtney-Martin, G. & Pencharz, P. B. Sulfur Amino Acids Metabolism From Protein Synthesis to Glutathione. in *The Molecular Nutrition of Amino Acids and Proteins* 265–286 (Elsevier, 2016). doi:10.1016/B978-0-12-802167-5.00019-0

45. Jackson, T. M. & O'Malley, K. G. Comparing genetic connectivity among Dungeness crab (*Cancer magister*) inhabiting Puget Sound and coastal Washington. *Mar. Biol.* **164,** (2017).

46. Jackson, T. M., Roegner, G. C. & O'Malley, K. G. Evidence for interannual variation in genetic structure of Dungeness crab (*Cancer magister*) along the California Current System. *Mol. Ecol.* (2018). doi:10.1111/mec.14443

47. Chong, E. Y., Huang, Y., Wu, H., Ghasemzadeh, N., Uppal, K., Quyyumi, A. A., Jones, D. P. & Yu, T. Local false discovery rate estimation using feature reliability in LC/MS metabolomics data. *Sci. Rep.* **5,** 17221 (2015).

48. Ding, M.-Z., Wang, X., Liu, W., Cheng, J.-S., Yang, Y. & Yuan, Y.-J. Proteomic Research Reveals the Stress Response and Detoxification of Yeast to Combined Inhibitors. *PLoS One* **7,** e43474 (2012).

49. Harding, H. P., Zhang, Y., Zeng, H., Novoa, I., Lu, P. D., Calfon, M., Sadri, N., Yun, C., Popko, B., Paules, R., Stojdl, D. F., Bell, J. C., Hettmann, T., Leiden, J. M. & Ron, D. An Integrated Stress Response Regulates Amino Acid Metabolism and Resistance to Oxidative Stress. *Mol. Cell* **11,** 619–633 (2003).

50. Murray, A. J. Metabolic adaptation of skeletal muscle to high altitude hypoxia: How new technologies could resolve the controversies. *Genome Med.* **1,** (2009).

51. Nava, G. M., Lee, D. Y., Ospina, J. H., Cai, S.-Y. & Gaskins, H. R. Genomic analyses reveal a conserved glutathione homeostasis pathway in the invertebrate chordate Ciona intestinalis. *Physiol. Genomics* **39,** 183–94 (2009).

52. Fraser, J. A., Saunders, R. D. C. & McLellan, L. I. Drosophila melanogaster glutamate-cysteine ligase activity is regulated by a modifier subunit with a mechanism of action similar to that of the mammalian form. *J. Biol. Chem.* **277,** 1158–65 (2002).

53. Mansfield, K. D., Simon, M. C. & Keith, B. Hypoxic reduction in cellular glutathione levels requires mitochondrial reactive oxygen species. *J. Appl. Physiol.* **97,** 1358–1366 (2004).

54. Cargnoni, A., Ceconi, C., Gaia, G., Agnoletti, L. & Ferrari, R. Cellular thiols redox status: A switch for NF-κB activation during myocardial post-ischaemic reperfusion. *J. Mol. Cell. Cardiol.* **34,** 997–1005 (2002).

55. Murata, Y., Ohteki, T., Koyasu, S. & Hamuro, J. IFN-gamma and pro-inflammatory cytokine production by antigen-presenting cells is dictated by intracellular thiol redox status regulated by oxygen tension. *Eur. J. Immunol.* **32,** 2866–2873 (2002).

56. Rajpurohit, R., Koch, C. J., Tao, Z., Teixeira, C. M. & Shapiro, I. M. Adaptation of chondrocytes to low oxygen tension: Relationship between hypoxia and cellular metabolism. *J. Cell. Physiol.* **168,** 424–432 (1996).

57. Pescador, N., Villar, D., Cifuentes, D., Garcia-Rocha, M., Ortiz-Barahona, A., Vazquez, S., Ordoñez, A., Cuevas, Y., Saez-Morales, D., Garcia-Bermejo, M. L., Landazuri, M. O., Guinovart, J. & del Peso, L. Hypoxia Promotes Glycogen Accumulation through Hypoxia Inducible Factor (HIF)-Mediated Induction of Glycogen Synthase 1. *PLoS One* **5,** e9644 (2010).

58. Shen, G. M., Zhang, F. L., Liu, X. L. & Zhang, J. W. Hypoxia-inducible factor 1-mediated regulation of PPP1R3C promotes glycogen accumulation in human MCF-7 cells under hypoxia. *FEBS Lett.* **584,** 4366–4372 (2010).

59. Phlippen, M. K., Webster, S. G., Chung, J. S. & Dircksen, H. Ecdysis of decapod crustaceans is associated with a dramatic release of crustacean cardioactive peptide into the haemolymph. *J. Exp. Biol.* **203,** 521–36 (2000).

60. Kuballa, A. V, Holton, T. A., Paterson, B. & Elizur, A. Moult cycle specific differential gene expression profiling of the crab Portunus pelagicus. *BMC Genomics* **12,** 147 (2011).

61. Fiehn, O. & Kind, T. Metabolite Profiling in Blood Plasma. in 3–17 (Humana Press, 2007). doi:10.1007/978-1-59745-244-1_1

62. Matyash, V., Liebisch, G., Kurzchalia, T. V., Shevchenko, A. & Schwudke, D. Lipid extraction by methyl- *tert* -butyl ether for high-throughput lipidomics. *J. Lipid Res.* **49,** 1137–1146 (2008).

63. Breiman, L. & Cutler, A. Breiman and Cutler's random forests for classification and regression. *Packag. 'randomForest'* 29 (2012). doi:10.5244/C.22.54

64. Torchiano, M. Efficient Effect Size Computation [R package effsize version 0.7.1].

65. Wohlgemuth, G., Haldiya, P. K., Willighagen, E., Kind, T. & Fiehn, O. The chemical translation service-a web-based tool to improve standardization of metabolomic reports. *Bioinformatics* **26,** 2647–2648 (2010).

66. Kim, S., Thiessen, P. A., Bolton, E. E., Chen, J., Fu, G., Gindulyte, A., Han, L., He, J., He, S., Shoemaker, B. A., Wang, J., Yu, B., Zhang, J. & Bryant, S. H. PubChem substance and compound databases. *Nucleic Acids Res.* **44,** D1202–D1213 (2016).

67. Barupal, D. K., Haldiya, P. K., Wohlgemuth, G., Kind, T., Kothari, S. L., Pinkerton, K. E. & Fiehn, O. MetaMapp: mapping and visualizing metabolomic data by integrating information from biochemical pathways and chemical and mass spectral similarity. *BMC Bioinformatics* **13,** (2012).

68. Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. & Ideker, T. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13,** 2498–504 (2003).