Identifying Protein Interactions Amongst DNA Damage Repair Proteins in Yeast

by

Siyang Li

A thesis submitted in conformity with the requirements
for the degree of Master of Science

Molecular Genetics
University of Toronto

Identifying Protein Interactions Amongst DNA Damage Repair Proteins in Yeast

Siyang Li

Master of Science

Molecular Genetics
University of Toronto

2014

Abstract

Mapping protein-protein interactions (PPIs) is crucial for understanding cellular systems. PPIs can be studied with binary or co-complex methods. A major economical binary method is the yeast two-hybrid (Y2H) system. However, current Y2H is still very time-consuming for studying large protein matrices. Here, I applied a new technology developed in our lab, Barcode Fusion Genetics Yeast Two-Hybrid (BFG-Y2H), to study ~69,000 protein pairs amongst 263 yeast DNA damage repair proteins. BFG-Y2H combines Y2H, DNA-barcoding, barcode-fusion genetics and next-generation sequencing to increase the current throughput of Y2H. I demonstrated that BFG-Y2H screens are reproducible and of comparable quality to a previously published high-quality Y2H dataset. In addition, I also discovered and confirmed five novel Y2H interactions amongst the top scores with the BFG-Y2H screens when compared to all previously published literature. Thus, BFG-Y2H can be used to more efficiently capture PPIs.

Acknowledgments

Foremost, I want to express my sincerest gratitude to my supervisor, Dr. Frederick Roth, for providing patient guidance, supportive mentorship and an excellent work environment during my M.Sc. studies. In addition to teaching me the subjects of high-throughput biology, he has also indirectly taught me on how to be a generous, forgiving and kind person by setting such an excellent example. It has been a great privilege and pleasure to work under his supervision for the past two and a half years.

I want to also thank my committee members, Dr. Amy Caudy and Dr. Alex Ensminger, for the helpful guidance, valuable suggestions and continuous mental support to complete my project. Their support has been invaluable to me.

I am truly grateful for the help and support of my lab-mates and collaborators. I am indebted to Dr. Nozomu Yachie for training me as a high-throughput biologist and for setting an excellent example of a great scientist. I am extremely grateful to Dr. Evangelia Petsalakis for her continuous support on everything, especially data analysis, and her friendship, which has made the hardest times much easier. I want to thank Dr. Javier Diaz for helping me to select the genes for my project. I want to thank Dr. Atina Cote and Dr. Joe Mellor for sharing their knowledge of yeast biology and sequencing with me. I am grateful to Takafumi for all of his patient mentorship and advice on sequencing analysis and Jochen for his patience in answering my R questions. I want to thank Marta and Analyn for their technical support in the lab and Nidhi from the Vidal lab for providing clones for my project. I am thankful to the rest of the Roth lab, especially Marinella, Song and Mariana for providing warm friendships and support for the last couple of years. I would also like to thank Dr. Corey Nislow and his lab at the University of Toronto for providing guidance and support for the first year of my graduate studies.

I want to thank my friends, Gowtham, Kirill, Tetyana and Victoria for being my support system, both personally and scientifically.

Finally, and most importantly, I want to thank my parents and Walter for their endless understanding, support, patience and love during my MSc studies.

Table of Contents

List of Tables

List of Figures

1 Introduction

1.1 Mapping protein-protein interactions

Proteins mediate many functions within the cell and interactions between them regulate the systems-level behavior of cells. As such, mapping protein-protein interactions (PPIs) is crucial for understanding cellular systems of an organism of interest.

Since the release of the complete genome sequence of the model organism *Saccharomyces cerevisiae* in 1996[1], functional annotations for predicted gene products have exploded. Given that most proteins require physical interactions with other proteins to fulfill their biological role, it was proposed that functional annotations for proteins can be obtained by systematically identifying potential PPIs[2]. In an effort to accelerate functional annotations of proteins, many innovative methods for the identification of PPIs have been presented and several of these methods are currently in use in laboratories around the world[3,4] .

There are two general classes of methods for detecting PPIs: those that detect co-complexed proteins and those that detect binary interactions. An example of a co-complex method is tandem affinity purification coupled to mass spectrometry[5], which determines the protein components of a complex for a tagged protein of interest by pulling it down from a cell lysate along with proteins in the same complex and identifying these proteins with mass spectrometry. In some instances, complexes such as the ribosome have been purified and analysed directly in the mass spectrometer[6]. An example of a binary method is the yeast two-hybrid (Y2H) method (see Section **1.2** for more information) where physical binary interactions between two proteins are captured. The interaction between Ras and the protein kinase Raf was

first detected by Y2H[7] and later demonstrated in mammalian cells[8]. We now know that Ras and Raf are part of a very important signaling pathway that causes many types of cancer when misregulated[9–12], and have therefore, become the focus of cancer therapy[13–15]. Many pharmacological solutions, such as sorafenib[16], have been developed and used clinically to inhibit overly active Raf kinases. This is one of many examples where studying PPIs have not only improved our understanding of molecular biology but have also led to therapeutic outcomes.

Even though many innovative methods for the identifying PPIs have been presented, Y2H is one of the very few methods that can be practically adapted for high-throughput strategy.

1.2 Yeast two-hybrid

Yeast two-hybrid (Y2H) is a major strategy for identifying direct binary protein-protein interactions (Figure 1). In Y2H, the activation domain (AD) and DNA-binding domain (DB) of a transcription factor are separated and fused to proteins of interest, Y (also known as “prey”) and X (also known as “bait”). This results in haploid yeast cells with plasmids that carry AD-Y or DB-X fusion proteins, which are targeted to the yeast nucleus by nuclear localization signals. When haploid cells containing AD-Y and DB-X plasmids are mated together, the resulting diploid cell will contain both AD-Y and DB-X plasmids. Physical interaction between AD-Y and DB-X will reconstitute the transcription factor and allow transcription of the reporter gene. In the case of using histidine as a selectable marker, a physical interaction between AD-Y and DB-X permits the survival of the diploid cell on media without histidine.

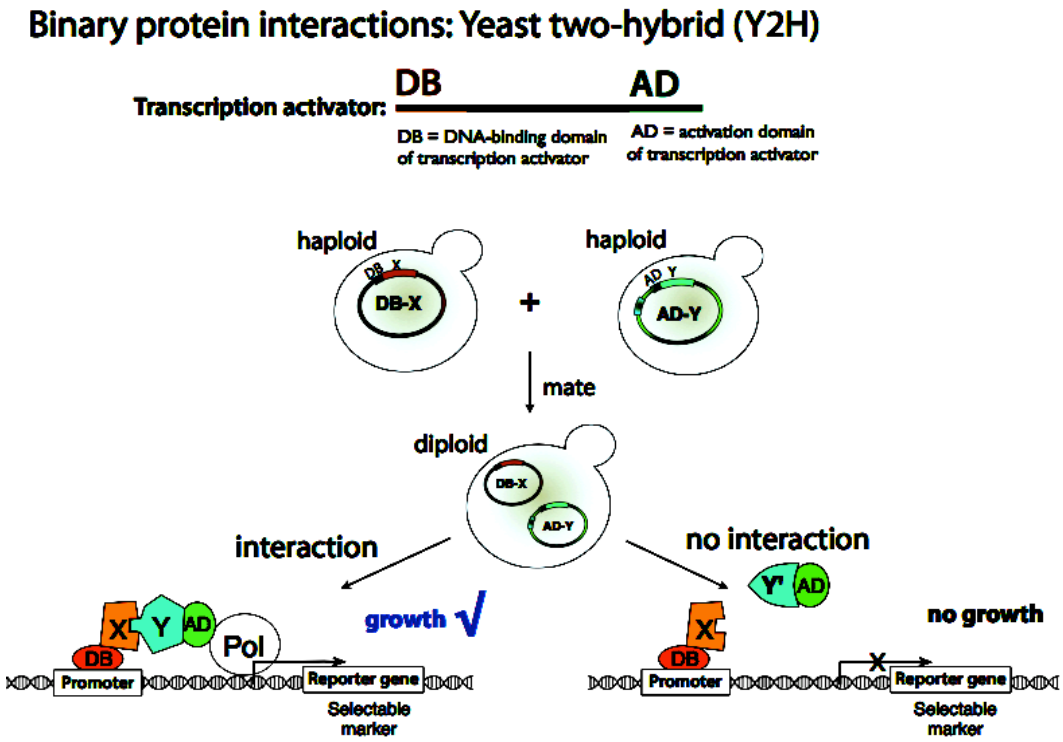


Figure 1. Yeast two-hybrid (Y2H). The activation domain (AD) and DNA-binding domain (DB) of a transcription factor are separated and fused to proteins of interest, Y and X, respectively. This results in haploid cells with plasmids carrying AD-Y or DB-X fusion proteins. When haploid cells containing AD-Y or DB-X plasmids are mated together, the resulting diploid cell will contain both AD-Y and DB-X plasmids. Physical interaction between AD-Y and DB-X will reconstitute the transcription factor and allow transcription of the reporter gene. Figure modified from Dr. N Yachie.

1.3 Limitations and advantages of the classical Y2H method

Like every method, Y2H has trade-offs. Since the fusion proteins in the classical Y2H method are all targeted to the nucleus, they must be able to fold and exist stably in order to retain their activity. Fusion proteins may also change the proteins' native conformations, which may alter their activity and/or binding. Studying the PPIs of the fusion proteins in both directions (DB-X/AD-Y and DB-Y/AD-X) would help to partially circumvent this problem. The readout of Y2H is dependent on the reporter's response to transcriptional activation; therefore, it is possible that a DB fusion protein could activate transcription on its own (autoactivation). Fortunately, this can be dealt with by implementing proper controls in the experiment[17]. However, the most important concern is the biological relevance of two positive Y2H interactors: due to both time constraints and cellular localization, two proteins that do not interact under regular cellular contexts could interact in Y2H (and vice versa). This last limitation can be mitigated by studying PPIs with complementary assays and follow-up experiments.

There are also numerous advantages for the Y2H method. Using yeast as a host can be considered as an advantage over a bacterial host, due to greater resemblance to higher eukaryote systems. Compared to classical biochemical methods, no protein purification or antibodies are needed; the cDNAs of the proteins of interest are sufficient. Additionally, transient and weak interactions can be picked up by Y2H[18], which are very important in signaling pathways. The biggest advantages are its simplicity, speed, low costs and the relative high-throughput capacity compared to other methods.

1.4 Quality-control for Y2H and recent advances

In 2009, Venkatesan *et al.*[19] developed a framework to assess the quality of Y2H high-throughput screens. Here, Y2H, amongst many other orthogonal assays, was evaluated against a set of 92 previously known and curated positive protein interactions based on 4 parameters: screening completeness, assay sensitivity, sampling sensitivity and precision. Screening completeness is the fraction of tested protein pairs out of the total possible space. Assay sensitivity is the ability to detect all possible biophysical interactions. Sampling sensitivity is the fraction of all possible protein interactions detected with one experiment; a lower sampling sensitivity means that multiple repeats need to be done in order to capture all possible interactions. Precision is the fraction of captured interactions that are true positives. Together, all of these parameters help to assess the quality of high-throughput Y2H screens. Y2H's assay sensitivity was determined to be 17%. False-positive rate (defined as number of false positives divided by the sum of true negatives and false positives) was determined by benchmarking against a random set of 188 protein pairs that have excluded all known binary interactions, and it was estimated to be <0.5%[19]. The precision (true positives divided by the sum of true positives and false positives) of the Y2H assay was shown to be 79%[19]. In another study done by Braun *et al.*[20], all of the previously mentioned parameters for Y2H have been shown to be on par with other assays used to detect binary protein interactions.

Traditionally, haploid cells carrying plasmids of fusion proteins are mated individually on plates and identification of the positive Y2H interactions is done from the coordinates of the colonies on the plates. This method is labour-intensive when surveying a large matrix because of the enormous number of plates required. In 2001, Walhout *et al.*[17] increased the throughput of Y2H by mating pools of haploid cells, instead of mating individually. Identification of the

interacting protein pairs was done by Sanger sequencing across the entire open reading frame (ORF) of the X and Y proteins. Although this new method decreased the number of plates required, it increased the cost associated with Y2H by using Sanger sequencing and still remained labour-intensive with colony-picking.

1.5 Overview of Barcode Fusion Genetics Yeast Two-Hybrid (BFG-Y2H)

Barcode Fusion Genetics Yeast Two-Hybrid (BFG-Y2H) is a technology being developed in the Roth lab (Dr. Nozomu Yachie, post-doctoral fellow). It combines Y2H, DNA-barcoding, barcode-fusion genetics and next-generation sequencing to increase the current throughput of Y2H (Figure 2). Increasing the throughput would allow us to query larger matrices as well as studying these matrices under different conditions.

Here, each plasmid (AD-X or DB-Y) has two unique DNA barcodes, which are flanked by specific recombination sites. Upon mating, a diploid cell containing both plasmids is formed. During inductive conditions for the Cre/Lox system[21], one of the two barcodes on each plasmid swaps positions with each other, resulting in fused barcodes. PCR amplification and next-generation sequencing (NGS) of the fused barcodes identify interacting protein pairs. The advantage of BFG-Y2H is that sets of both DB-X and AD-Y strains can be pooled together, allowing for *en masse* mating. Only sequencing of the fused barcode regions is needed to identify interacting pairs, thus, bypassing the current Y2H bottleneck steps of plating and colony-picking.

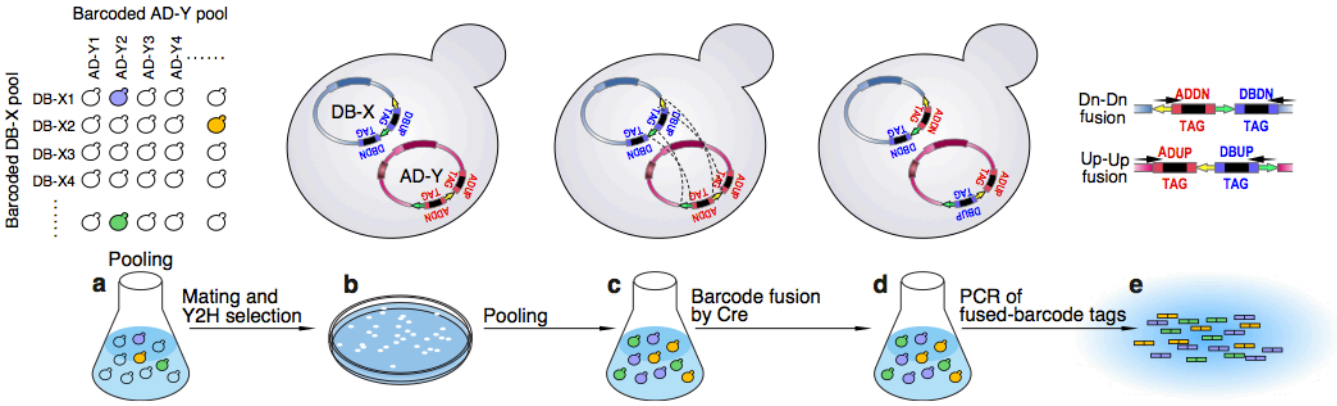


Figure 2. Barcode Fusion Genetics Y2H (BFG-Y2H). BFG-Y2H combines Y2H, DNA-barcoding, barcode-fusion genetics (BFG) and next-generation sequencing (NGS) to increase the current throughput of Y2H. Here, each plasmid (AD-X/DB-Y) has two unique DNA barcodes (either the DB TAGs or the AD TAGs in the figure); upon mating and inductive conditions for the Cre/Lox system, one of the two barcodes on each plasmid swaps positions with each other within the diploid cell (shown by dotted lines), producing fused barcodes. PCR amplification and next-generation sequencing of the fused barcodes identify interacting protein pairs. Image modified from Yachie *et al.* (unpublished).

1.6 BFG-Y2H helps to uncover novel interactions amongst DNA damage repair proteins

In this study, we applied BFG-Y2H to capture PPIs amongst 263 selected yeast DNA damage repair proteins. DNA damage repair proteins are of particular interest because they preserve genome integrity in all organisms[22]. If left unrepaired, DNA damage can have deleterious effects because it can result in mutations and chromosomal aberrations. Defects in DNA damage are known to cause increased cancer risks, developmental defects and neurodegenerative diseases[23]. Moreover, DNA damage repair proteins are nuclear, making them a great choice to use in BFG-Y2H.

I found novel Y2H interactions with BFG-Y2H. I confirmed five of these interactions to be novel when checked against literature. Additionally, I determined which positive controls should be used in the BFG-Y2H screens and I also demonstrated that the BFG-Y2H screens are highly reproducible.

2 Materials and Methods

2.1 Barcode-Fusion Genetics: background strains

Background strains, RY1010 (MAT a) and RY1030 (MAT alpha), were generated by Dr. N. Yachie (post-doctoral fellow, Roth lab). Each plasmid contains two unique barcodes (UPTAG and DNTAG), placed in tandem with *loxP* and *lox2272* sites. The Cre recombinase is controlled by the Tet-On system[24] in our background strains (Figure 3). In the presence of doxycycline, Cre is able to bind to the rtTA transcription factor and together, they are able to bind DNA at the $tetO_2$ promoter, allowing the transcription of Cre (Figure 3). The Cre recombinase allows the barcodes from the AD-X and DB-Y plasmids to switch positions, generating UPTAG-UPTAG and DNTAG-DNTAG barcode fusions. RY1010 and RY1030 are derived from Y8800 and Y8930, respectively. Y8800 and Y8930 are generated from Charlie Boone's lab by adding cycloheximide resistance to strains generated by James *et al.*[25](see Table 1 for genotypes).

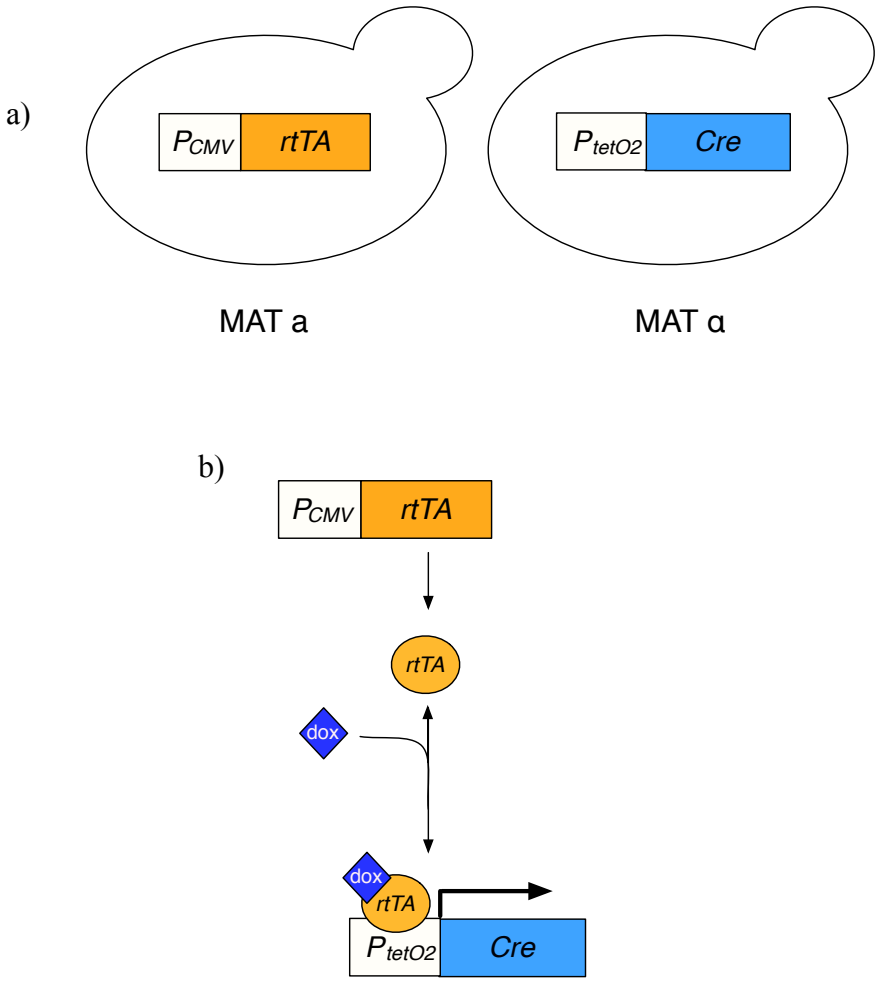


Figure 3. Activation of Cre recombinase in our background Y2H strains. A) rtTA and Cre are integrated into the genomes of the MATa (RY1010) and MATα (RY1030) strains, respectively. B) In the presence of doxycycline, rtTA is able to bind to the tetO2 promoter and activate the transcription of Cre recombinase in the diploid yeast cell.

Strain	Genotype
Y8800	*MATa leu2-3,112 trp1-901 his3-200 ade2-101 ura3-52 gal4Δ gal80Δ cyh2^R LYS2::P_{GAL1}-HIS3 P_{GAL2}-ADE2 MET2::P_{GAL7}-lacZ*
Y8930	*MATα leu2-3,112 trp1-901 his3-200 ade2-101 ura3-52 gal4Δ gal80Δ cyh2^R LYS2::P_{GAL1}-HIS3 P_{GAL2}-ADE2 MET2::P_{GAL7}-lacZ*
RY1010	*MATa leu2-3,112 trp1-901 his3-200 ade2-101 ura3-52 gal4Δ gal80Δ cyh2^R LYS2::P_{GAL1}-HIS3 P_{GAL2}-ADE2 MET2::P_{GAL7}-lacZ can1Δ::P_{CMV}-rtTA-KanMX4*
RY1030	*MATα leu2-3,112 trp1-901 his3-200 ade2-101 ura3-52 gal4Δ gal80Δ cyh2^R LYS2::P_{GAL1}-HIS3 P_{GAL2}-ADE2 MET2::P_{GAL7}-lacZ can1Δ::T_{ADH1}-P_{tetO2}-Cre-T_{CYC1}-KanMX4*

Table 1. Genotypes for the BFG-Y2H background strains. RY1010 and RY1030 are the BFG-Y2H background strains, which are derived from strains Y8800 and Y893014, respectively.

2.2 Generation of barcoded destination vectors

Barcoded destination vectors were generated (Figure 4) in collaboration with Dr. E. Petsalakis and Dr. N. Yachie (post-doctoral fellows, Roth lab).

Gibson isothermal assembly[26] was used to assemble three pieces of DNA fragments: the appropriate destination vector backbone (pDEST-AD or pDEST-DB[17], CEN-origin in yeast) and two unique randomly generated barcode fragments, creating a pool of barcoded destination vectors. This pool of barcoded destination vectors was transformed into One Shot ccdB Survival 2 T1R competent cells (Invitrogen) and spread onto 245mm x 245mm square LB+ampicillin plates. They were incubated at 37 °C for one day. Single colonies were arrayed into individual wells with the QPix robot (Genetix), so that their identities could be determined by NGS.

Row-column-plate PCR was used with NGS to determine the barcodes' sequences (Figure 5) where row, column and plate tags were added to each well through separate rounds of PCR. All resultant amplified tagged barcodes were then pooled together and NGS was used to identify the barcodes. Quality-check was also performed on *loxP, lox2272* sites to verify their sequences through sequencing. Barcodes that were successfully identified via sequencing and passed all quality-check points were used to generate barcoded expression vectors through Gateway LR reactions.

Please see **Appendix 7.1** (Arrangement of barcode sequences on the plasmid) for more information on the arrangement of barcode sequences, *loxP* and *lox2272* sites and **Appendix 7.2** for more information on the fused products after the induction of Cre.

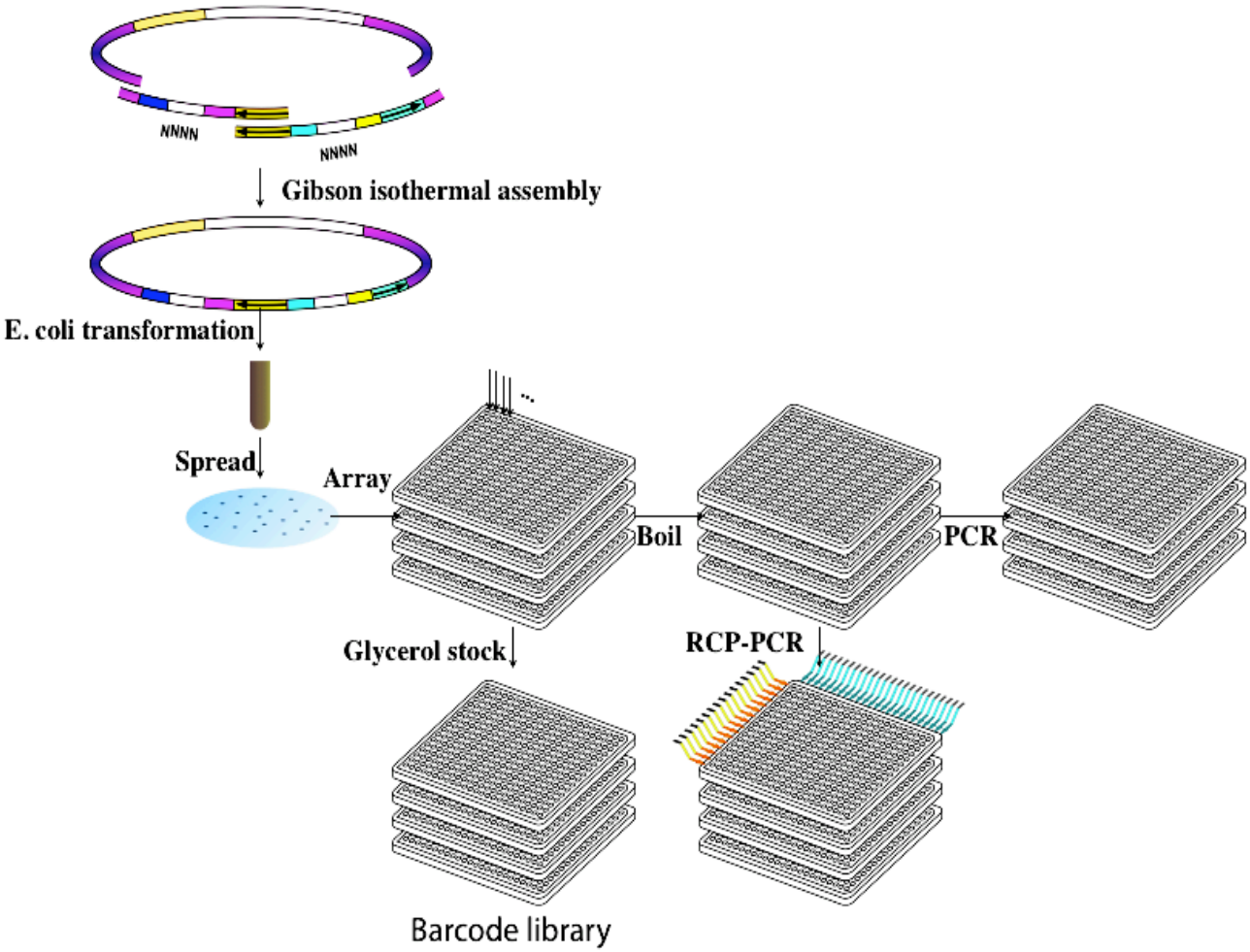


Figure 4. Generation of barcoded destination vectors. Two fragments of degenerate oligonucleotides (NNNN in the figure) and the destination vector's backbone are used to generate barcoded destination vectors. This pool of barcoded destination vectors is then transformed into E. coli cells, spread on large agar plates and single colonies are arrayed into individual wells. The identity of the barcodes in each well is identified by Row-Column-Plate PCR (RCP-PCR). Image modified Yachie *et al.* (unpublished).

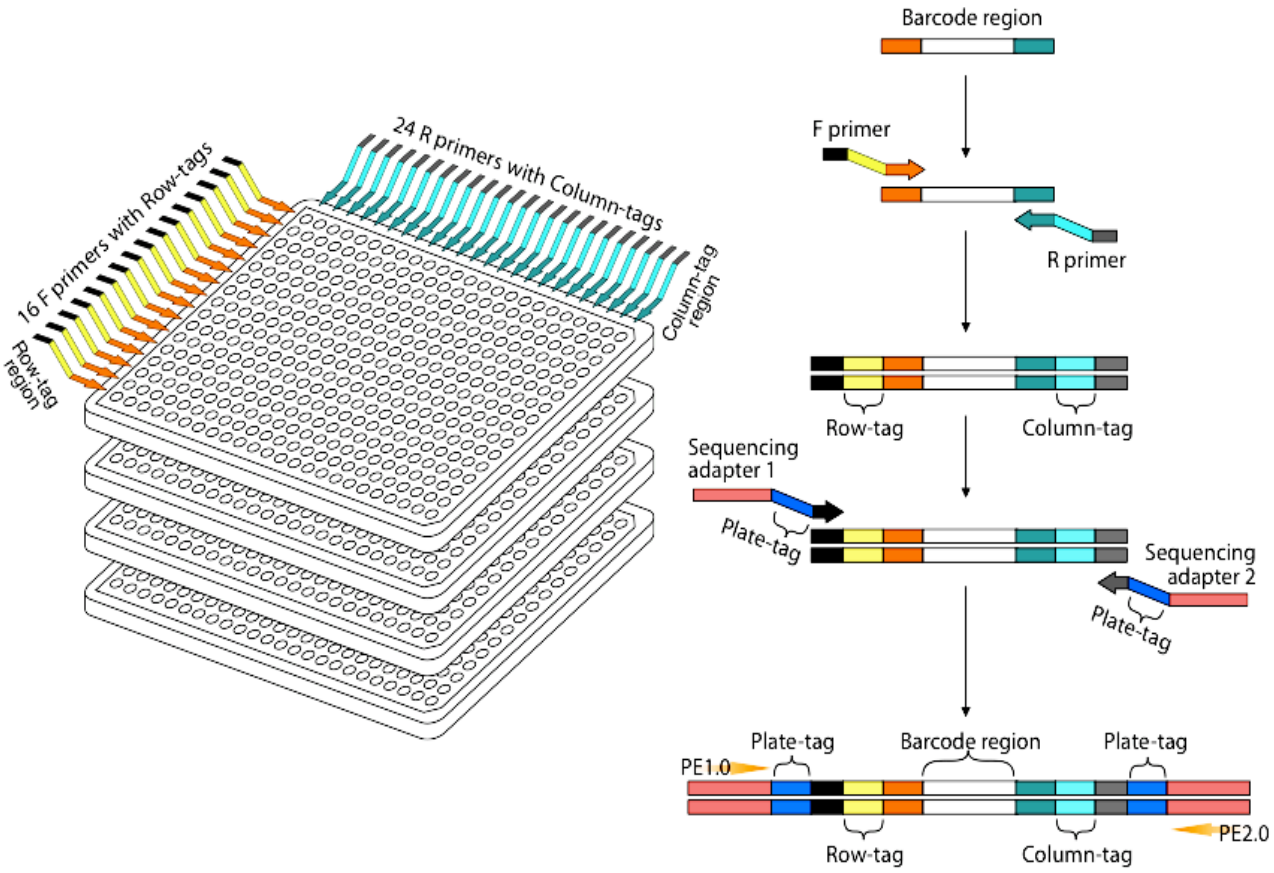


Figure 5. Row-column-plate PCR (RCP PCR). Row, column and plate tags are added to each well through separate rounds of PCR. All amplified barcodes are pooled together and NGS is then used to identify the barcodes. Image modified from Yachie *et al.* (unpublished).

2.3 Selection of DNA damage repair proteins

With the help of Dr. Javier Diaz (post-doctoral fellow, Roth lab), Dr. Atina Cote (research associate, Roth lab) and Dr. Dan Durocher (PI, Durocher lab), we selected 263 yeast DNA damage repair proteins from many resources (see Table 2): from Saccharomyces Genome Database[27] annotations (114, "sgd_annotation"), from Alvaro *et al.*[28] due to deletions increasing the levels of Rad52 foci (69, "rad52_foci"), from REPAIRtoire[29], a database with literature curated proteins that are involved in DNA repair pathways (61, "repairtoire"), manually selected (84, "core"), from FuncBase[30] predictions (6, "yeast_func") made by using GO terms from the "core" set and DNA damage repair proteins from the top 5 centroid genes from each of the 50 clusters in Costanzo *et al.*[31] (21, "survey"). Some genes fall under multiple categories, please see Table 2 for a complete list of genes.

Table 2. Complete list of DNA damage repair proteins. There are 263 DNA damage repair proteins. There are 114 proteins selected from Saccharomyces Genome Database16 annotations (sgd_annotation), 69 proteins from Alvaro et al.17 because their deletions increased the levels of Rad52 foci (rad52_foci), 61 proteins from the REPAIRtoire18 (repairtoire), 84 proteins manually selected (core), 6 proteins from FuncBase19 based on predictions (yeastfunc) and 21 proteins were selected from the top 5 genes within each of the 50 clusters in Costanzo et al.20 (survey). Some proteins fall under multiple categories.

Gene names	Annotation		Gene names	Annotation
RAD57	core/repairtoire/rad52_foci		*ULP2*	sgd_annotation
RAD54	core/repairtoire/rad52_foci		*ABF1*	sgd_annotation
RAD51	core/repairtoire/rad52_foci		*D2*	sgd_annotation
MUS81	core/repairtoire/rad52_foci		*MEK1*	sgd_annotation
MLH1	core/repairtoire/rad52_foci		*YNL194C*	sgd_annotation
POL30	repairtoire/sgd_annotation		*MGM101*	sgd_annotation
SSL1	repairtoire/sgd_annotation		*DDR48*	sgd_annotation
RAD23	repairtoire/sgd_annotation		*HIM1*	sgd_annotation
SSL2	repairtoire/sgd_annotation		*NSE1*	sgd_annotation
RAD28	repairtoire/sgd_annotation		*RAD16*	sgd_annotation
TFB1	repairtoire/sgd_annotation		*DIF1*	sgd_annotation
CDC9	repairtoire/sgd_annotation		*ALK2*	sgd_annotation
RNR4	repairtoire/sgd_annotation		*YAL044W-A*	sgd_annotation
TFB5	repairtoire/sgd_annotation		*RAD33*	sgd_annotation
TFB3	repairtoire/sgd_annotation		*YAR028W*	sgd_annotation
OGG1	repairtoire/sgd_annotation		*YGL085W*	sgd_annotation
MSH5	repairtoire/sgd_annotation		*YPR147C*	sgd_annotation
TFB2	repairtoire/sgd_annotation		*YMR099C*	sgd_annotation
RFA1	repairtoire/sgd_annotation		*YPL108W*	sgd_annotation
TFB4	repairtoire/sgd_annotation		*YMR178W*	sgd_annotation
MGT1	repairtoire/sgd_annotation		*THI4*	sgd_annotation
POL31	repairtoire/sgd_annotation		*PIN4*	sgd_annotation
IRC24	rad52_foci/sgd_annotation		*DIN7*	sgd_annotation
LCD1	yeastfunc/sgd_annotation		*NPR3*	sgd_annotation
RAD52	core/repairtoire/survey		*KRE29*	sgd_annotation
XRS2	core/repairtoire/survey		*RAD34*	sgd_annotation
HSM3	sgd_annotation/survey		*ALK1*	sgd_annotation
RRD1	sgd_annotation/survey		*SCC4*	sgd_annotation
ACK1	sgd_annotation/survey		*YMR244C-A*	sgd_annotation
FYV6	core/sgd_annotation		*SMC1*	sgd_annotation
YEN1	core/sgd_annotation		*SMC5*	sgd_annotation
IRC25	rad52_foci/survey		*IES4*	sgd_annotation
MAD2	rad52_foci/survey		*HTA1*	sgd_annotation
IRC16	rad52_foci/survey		*PRI2*	sgd_annotation
VPS72	rad52_foci/survey		*RNR2*	sgd_annotation

MAD1	rad52_foci/survey		*SMC6*	sgd_annotation
NUP133	rad52_foci/survey		*NSE5*	sgd_annotation
VPS71	rad52_foci/survey		*PAN3*	sgd_annotation
CTF19	rad52_foci/survey		*YPR022C*	sgd_annotation
IRC21	rad52_foci/survey		*YHR192W*	sgd_annotation
YKU70	core/repairtoire		*MND1*	sgd_annotation
UNG1	core/repairtoire		*PPH3*	extra_manual
MAG1	core/repairtoire		*CHL1*	extra_manual
APN2	core/repairtoire		*MMS4*	core/survey
LIF1	core/repairtoire		*SWC5*	core/survey
RAD2	core/repairtoire		*RAD17*	core/survey
RAD1	core/repairtoire		*SRS2*	core/survey
RAD50	core/repairtoire		*MRC1*	core/survey
NTG2	core/repairtoire		*RAD61*	core/survey
RAD30	core/repairtoire		*SEM1*	repairtoire
MSH6	core/repairtoire		*DPB4*	repairtoire
POL4	core/repairtoire		*SPO11*	repairtoire
MRE11	core/repairtoire		*RFC5*	repairtoire
REV3	core/repairtoire		*DPB2*	repairtoire
PHR1	core/repairtoire		*DPB3*	repairtoire
RAD10	core/repairtoire		*HNT3*	repairtoire
RAD27	core/repairtoire		*CCL1*	repairtoire
MSH2	core/repairtoire		*MRPL1*	rad52_foci
MSH4	core/repairtoire		*RTT103*	rad52_foci
MEC1	core/repairtoire		*IRC19*	rad52_foci
MPH1	core/repairtoire		*COX16*	rad52_foci
EXO1	core/repairtoire		*IRC18*	rad52_foci
PSO2	core/repairtoire		*PAC10*	rad52_foci
TDP1	core/repairtoire		*CBT1*	rad52_foci
MSH3	core/repairtoire		*IRC23*	rad52_foci
DNL4	core/repairtoire		*MDM20*	rad52_foci
RAD3	core/repairtoire		*IRC22*	rad52_foci
POL32	core/repairtoire		*IRC4*	rad52_foci
NEJ1	core/repairtoire		*IRC7*	rad52_foci
PBY1	yeastfunc/survey		*DAK2*	rad52_foci
RAD59	core/rad52_foci		*RIM9*	rad52_foci
SLX8	core/rad52_foci		*NUP60*	rad52_foci
ESC2	core/rad52_foci		*GSH2*	rad52_foci
RTT109	core/rad52_foci		*ATR1*	rad52_foci
RMI1	core/rad52_foci		*HST3*	rad52_foci
MMS1	core/rad52_foci		*PAP2*	rad52_foci
RTT107	core/rad52_foci		*TOF2*	rad52_foci
RTT101	core/rad52_foci		*BDF1*	rad52_foci
SAE2	core/rad52_foci		*ECM11*	rad52_foci
WSS1	core/rad52_foci		*BUB2*	rad52_foci
RRM3	core/rad52_foci		*HPR1*	rad52_foci
YKR075C	sgd_annotation		*YMR027W*	rad52_foci
YLR118C	sgd_annotation		*IRC3*	rad52_foci
YLR271W	sgd_annotation		*IRC8*	rad52_foci
HUG1	sgd_annotation		*IZH2*	rad52_foci
SML1	sgd_annotation		*IRC15*	rad52_foci
ADD37	sgd_annotation		*SGO1*	rad52_foci

TPP1	sgd_annotation		*LAG2*	rad52_foci
YDR262W	sgd_annotation		*BUD27*	rad52_foci
NSE3	sgd_annotation		*GDH1*	rad52_foci
HHO1	sgd_annotation		*IRC10*	rad52_foci
RCN2	sgd_annotation		*MRP17*	rad52_foci
YNL134C	sgd_annotation		*AHC1*	rad52_foci
YJR085C	sgd_annotation		*RCO1*	rad52_foci
YJR011C	sgd_annotation		*MED1*	rad52_foci
HAT1	sgd_annotation		*DDR2*	rad52_foci
IES6	sgd_annotation		*MAD3*	rad52_foci
YJL144W	sgd_annotation		*MRPS16*	rad52_foci
LSM12	sgd_annotation		*IRC6*	rad52_foci
YML131W	sgd_annotation		*YMR31*	rad52_foci
YJR096W	sgd_annotation		*PSY4*	yeastfunc
YGR126W	sgd_annotation		*YDL156W*	yeastfunc
HHT1	sgd_annotation		*PSY2*	yeastfunc
YHL018W	sgd_annotation		*HDA1*	yeastfunc
YIM1	sgd_annotation		*SHU2*	core
MMS21	sgd_annotation		*ASF1*	core
YFR017C	sgd_annotation		*CSM3*	core
RFA3	sgd_annotation		*MGS1*	core
YDL119C	sgd_annotation		*PSY3*	core
DDI3	sgd_annotation		*SLX4*	core
HHT2	sgd_annotation		*SAW1*	core
HTA2	sgd_annotation		*CHK1*	core
DDI1	sgd_annotation		*RAD55*	core
CRT10	sgd_annotation		*RPN4*	core
YBL036C	sgd_annotation		*CLA4*	core
PRI1	sgd_annotation		*DOA1*	core
RAD53	sgd_annotation		*PIF1*	core
DDI2	sgd_annotation		*NFI1*	core
MIG3	sgd_annotation		*CTF18*	core
ECO1	sgd_annotation		*MSH1*	core
BER1	sgd_annotation		*CSM2*	core
RNR3	sgd_annotation		*SLX1*	core
RAD7	sgd_annotation		*SIZ1*	core
HRR25	sgd_annotation		*RDH54*	core
LDB16	sgd_annotation		*REV1*	core
RPH1	sgd_annotation		*RAD9*	core
RNR1	sgd_annotation		*MEC3*	core
FMP41	sgd_annotation		*MLH2*	core
YOR062C	sgd_annotation		*DCC1*	core
OCA1	sgd_annotation		*SLX5*	core
ULP2	sgd_annotation		*RAD6*	core
ABF1	sgd_annotation		*RAD5*	core
D2	sgd_annotation		*TOP3*	core

2.4 Generating barcoded Y2H strains

I generated my barcoded Y2H strains by using a Gateway LR reaction[32] to a pool of three uniquely barcoded destination vectors (**Section 2.2**: Generation of barcoded destination vectors) for every individual entry clone (Figure 6). This resulted in barcoded expression vectors for both baits and preys. I obtained the entry clones from the HIP FLEXGene collection[9] (in collaboration with Dr. Marc Vidal's group). All entry clones have been previously sequence-verified by the Vidal lab. In the presence of Clonase, recombination occurs between corresponding R and L sites on the destination and entry vectors, respectively. Under the correct selection criterion, only the expression vector will remain. Since the barcodes are on the backbone of the destination vector, the expression vectors are also barcoded, resulting in a collection of barcoded expression vectors (both baits and preys) for the selected space.

I grew cells containing barcoded destination vectors from glycerol stocks in 96-deep well plates in 1 mL of TB+ampicillin media at 37 °C and 900 rpm to saturation for one day. I pooled three destination vectors (300 mL each) per well and extracted their plasmids by using NucleoSpin 96 Plasmid Core Kit (Macherey-Nagel). I also grew the entry clones from glycerol stocks in 96-deep well plates in 1 mL of TB+kanamycin media at 37 °C and 900 rpm to saturation for one day. Their plasmids were also extracted with NucleoSpin 96 Plasmid Core Kit (Macherey-Nagel). The destination vector pool (100 ng) was used with the entry plasmid (100 ng) in a Gateway LR reaction (Clonase I, Invitrogen, standard protocol). I incorporated positive (included with Invitrogen kit) and negative (no entry clone) controls for the Gateway LR reaction onto each 96-well plate. The reaction was left for one day at room temperature. The next day, I transformed the resultant of the Gateway LR reaction, the barcoded expression vectors, into 13 μL of DH5α competent cells (NEB, high efficiency) in a 96-well PCR plate. I optimized the 96-

well Gateway LR reactions such that at least 95% of the *E. coli* transformants have more than 10 colonies per 5 μL of spotted transformed cells. I assessed each experiment's cloning and transformation efficiencies by spotting 5 μL of the transformed cells onto LB+ampicillin OmniTrays. In parallel, I grew the rest of the transformed *E. coli* cells for one day in 96-deep well plates containing 1 mL of LB+ampicillin. I extracted the plasmids the following day with NucleoSpin 96 Plasmid Core Kit (Macherey-Nagel). I spot-checked the resultant expression clones on each plate with NGS to ensure correct plate orientations. I transformed these plasmids into the appropriate yeast background strains (RY1010 for the AD/preys or RY1030 for DB/baits) by using Frozen-EZ Yeast Transformation II Kit (Zymo Research).

Because generating barcoded Y2H strains is one of the most time-consuming and costly steps of BFG-Y2H, I attempted to develop a more high-throughput method by using an en masse Gateway LR reaction to increase the efficiency of generating barcoded Y2H strains. Please see **Appendix 7.3** (Using an *en masse* Gateway LR reaction to generate barcoded strains) for more details.

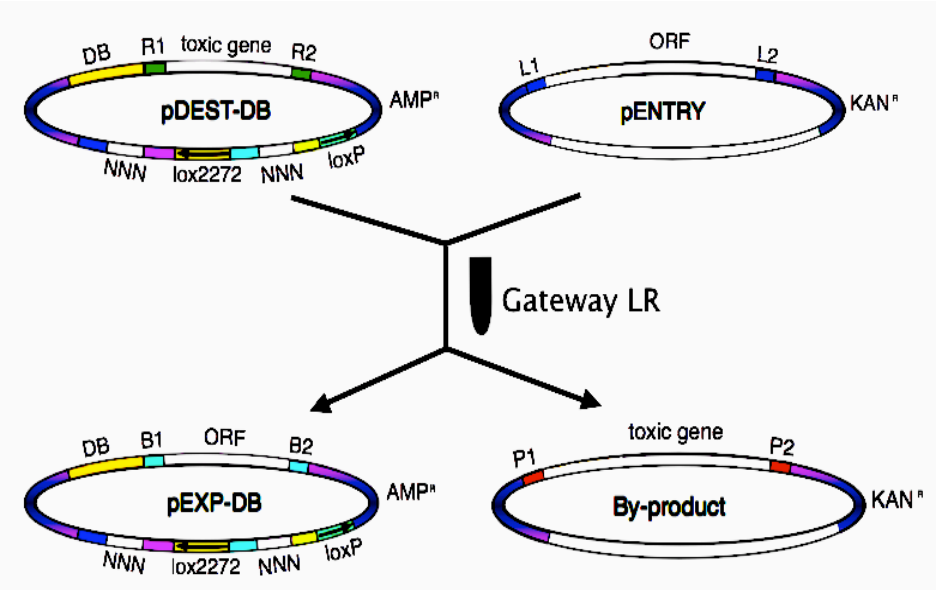


Figure 6. Gateway LR reaction between a barcoded destination vector and an entry clone. The destination vector (pDEST-DB) is barcoded (NNN) and flanked by corresponding loxP sites. In the presence of Clonase®, recombination occurs between corresponding R and L sites on the destination and entry vectors, respectively, and generates an expression product (pEXP-DB) that is barcoded and a by-product. Due to selection criterion, out of all four vectors, only the barcoded expression vector will remain.

2.5 BFG-Y2H screens

The BFG-Y2H screening procedure was developed and optimized by Dr. N. Yachie (Roth lab). It took me 10 days to complete the screen up to the generation of the sequencing libraries to identify the fused barcodes. A complete BFG-Y2H screen contains three conditions: with histidine (+His), without histidine (-His) and with 3-amino-1,2,4-triazole (+3AT). The +His condition is the non-selective condition, which gives us an estimate of the complexity of the screen with regard to diploid cells before being subjected to selection conditions. The –His condition is the selective condition for interacting protein pairs. The +3AT condition is a more stringent selective condition for interacting protein pairs, as 3AT is a competitive inhibitor of the *HIS3* gene product[34].

I grew cells with barcoded expression vectors from glycerol stocks in 96-deep well plates containing 1mL of SC-Leu+Ade or SC-Trp+Ade media for the DB-X and AD-Y pools, respectively. The deep well plates were grown at 30 °C and 900 rpm to saturation for two days. I pooled the DB-X and AD-Y pools separately, at equal cell densities (700 OD_{600nm} units of each) and washed twice with water. I then resuspended the cells in water and pooled them together at equal OD_{600nm} units. The cells were left for 3 hours at room temperature to increase mating efficiency[35]. I pelleted cells and removed the supernatant, and I spread the remaining cell pellets directly onto YPAD plates (technical replicates) and the plates were incubated for one day at room temperature. The next day, I scraped the cells off the plates and collected them with water. I washed the cells twice with water and I resuspended them in 500mLs of SC-Leu-Trp+His+Ade media in a 2L flask to a final 1.0 OD_{600nm}/mL. The cells were incubated for two days at 30 °C and 200 rpm to enrich for diploid cells. Afterwards, I washed the diploid cells with water and resuspended in water. I spread 200uL of cells at a density of 1.0 OD_{600nm}/mL (10^8 diploid cells)

onto 150mm agar plates. The plates were non-selective condition ("+His", SC-Leu-Trp+His+Ade and selective conditions ("-His", SC-Leu-Trp-His+Ade and "+3AT", SC-Leu-Trp-His+Ade+3AT at 1mM amino-1,2,4-triazole).

After 3 days of incubating the selection plates at 30 °C, I scraped the cells from the plates and washed them twice with water. I resuspended the cells in 5mL SC-Leu-Trp+His+Ade+Doxycycline (10μg/mL) at a final concentration of 1.0 OD_{600nm}/mL. The cells were incubated over night at 30 °C and 200 rpm for the induction of Cre. The following day, I extracted the plasmid DNA from 3 OD_{600nm} units of cells by using Charge Switch Plasmid Yeast Mini Kit (Invitrogen). Fused barcodes were amplified with primers carrying Illumina paired-end multiplexed sequencing primers. The amplified products were size-selected by E-Gel SizeSelect 2% Agarose gel (Invitrogen) and they were sequenced on the HiSeq2500 (Illumina). The barcode reads were mapped with Bowtie2[36].

2.6 Replicates for the BFG-Y2H screens

There are several ways to assess the reproducibility of BFG-Y2H screens (Figure 7) Firstly, I performed two screen replicates of the BFG-Y2H screen, which were separated at the mating stage and onwards (Figure 7, panel A). Secondly, I implemented another level of replicates at the barcoded strain level, where I created two to three different expression clones for each ORF, each represented by a different barcode. Upon mating, the resulting diploid cell will contain different combinations of the barcodes (Figure 7**,** panel B), and I consider these to be biological replicates. Lastly, there are two sets of fused barcodes within each diploid cell, UP-UP and DN-DN (Figure 7**,** panel C). Please see **Figure 8** for the visualization of the relative positions of each type of replicate within the BFG-Y2H screen.

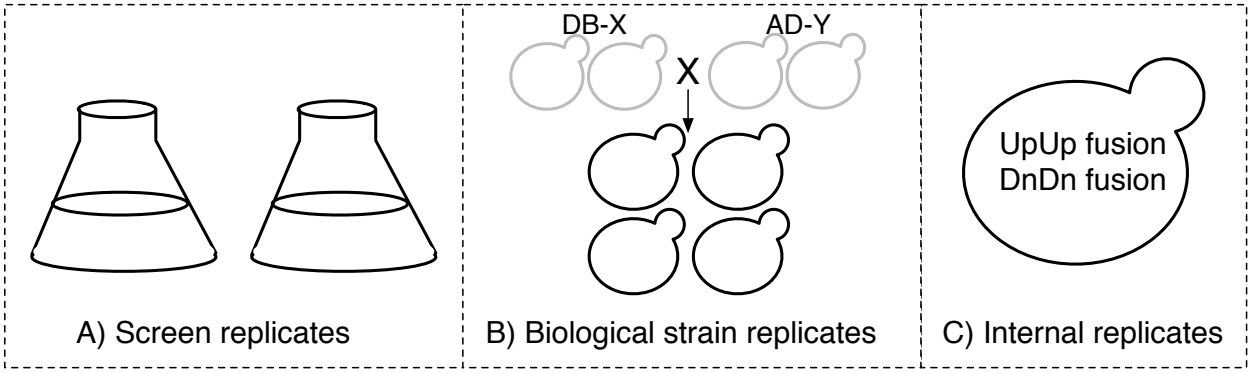


Figure 7. Assessing the reproducibility of BFG-Y2H. There are three levels of replication that I implemented during the BFG-Y2H screen. A) Screen replicates that were separated from the mating phase and onwards. B) Biological strain replicates: two to three uniquely barcoded strains for each ORF were used to mate together to generate different combinations of diploid cells. In this example, the bait and prey are represented by two different barcodes, resulting in four different combinations of the diploid cells after mating. C) Internal replicates: two pairs of fused barcodes (UP-UP and DN-DN) are within each diploid cell, serving as another way of assessing reproducibility

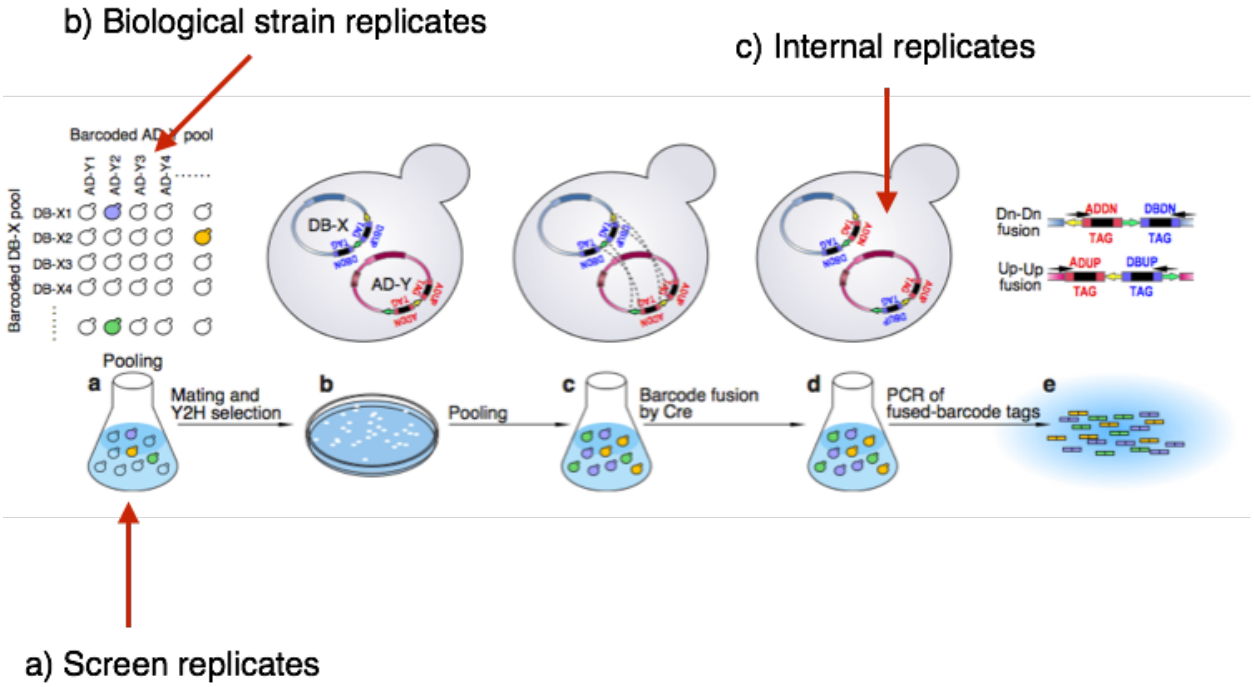


Figure 8. Replicates within the BFG-Y2H screen. Visual representation of the different kind of replicates within the BFG-Y2H screen: A) screen replicates have been separated at the mating stage, B) biological strain replicates are different diploids generated from differently barcoded haploid strains C) internal replicates of the UP-UP and DN-DN fusion barcodes within each diploid cell.

2.7 Pairwise retesting

I performed pairwise retesting for candidate interactions from the BFG-Y2H screens by following the protocol mentioned in Rual *et al*.[37] with slight modifications as follows. Four experimental repeats were performed of the mating stage and onwards.

I used AD-null cells as negative controls to determine auto-activators in the pairwise retesting experiment. The AD-null expression vectors don't have an ORF inserted (obtained from the Vidal lab). These plasmids were then transformed into RY1010. I also selected 3 pairs of previously known positive Y2H interactions as positive controls (obtained from the Vidal lab).

I grew barcoded Y2H strains from glycerol stocks in 96-well culture plates containing 150µL of SC-Leu+Ade or SC-Trp+Ade media for DB-X and AD-Y strains, respectively. The cells were grown at 30 °C at 900rpm for one day to saturation. The next day, I mated the corresponding DB-X and AD-Y cells by combining them at equal volumes (100µL:100µL). I resuspended the cells in 100µL of YPAD and repelleted them. The cells were left at room temperature for one day. In parallel, I also mated all DB-X cells with AD-null cells.

The following day, I washed the cells twice with water and resuspended them in 100µL of diploid selection media (SC-Leu-Trp+His+Ade). I enriched for diploid cells by taking 10µL of these mated samples to grow in 96-deep well plates containing 1mL of diploid selection media. The cells were incubated at 30 °C at 900rpm for two days. I washed the cells with 100µL of water twice, and resuspended the cells in 100µL of water. I diluted the cells by 100-fold, and spotted 5µL of the diluted cells onto different selection plates (+His, -His, +3AT). The plates were incubated at 30 °C for two days.

2.8 Calculation of interaction scores

For each screen, interaction scores (Δs) are calculated. The calculation can be divided into two steps: (1) the normalization of barcode counts and (2) distinguishing the real positive interactions from autoactivators. This calculation scheme was developed by Dr. N. Yachie and the implementation was carried out with help from Dr. E. Petsalakis.

The normalization of barcodes was done using the non-selective condition (+His). This is because not all protein pairs are equally represented. Some causes of this are due to, but not limited to, some ORFs' barcodes are missing due to failed cloning, toxicity effects of the ORF in the nucleus and some barcodes may not sequence as well due to inherent sequence biases.

Let ${c^+}_{ij}$ represent the barcode counts of the protein pair DB_i and AD_j in the non-selective condition (+His). Let ${f^+}_i$ be the marginal frequency of all barcodes containing DB_i, which is an entire row of the matrix:

$$ {f^+}_i = \frac{\sum_{j=1}^{n} {c^+}_{i,j}}{total^+} $$

Where j represents the row number and $total^+$ is the total count of the barcodes in the non-selective condition (+His). The frequency of the protein pair AD_i and DB_j in the +His condition is represented by ${f^+}_{ij}$:

$$ {f^+}_{ij} = {f^+}_i \times {f^+}_j $$

This is used instead of the raw barcode counts due to limited sequencing coverage across the entire matrix and sample complexity. For the selective conditions (-His or +3AT), the marginal frequency of the protein pair AD_i and DB_j is ${f^-}_{ij}$:

$$f^-_{ij} = \frac{c^-_{ij} + \alpha}{total^-}$$

The sequencing coverage for selection conditions is much better, and for those strains that survive, the raw barcode counts (c^-_{ij}) could be used to calculate the frequencies of each individual protein pair. α was set to 1 to avoid having 0 values for f^-_{ij}.

The enrichment of the positive protein interactions in the selective condition is shown with s_{ij}:

$$s_{ij} = \frac{f^-_{ij}}{f^+_{ij}}$$

Background autoactivators are eliminated by adjusting the s_{ij} scores:

If $s_{ij} - \tilde{s_i} \geq \beta_i$, then $\Delta s_{ij} = \frac{s_{ij} - \tilde{s_i}}{\beta_i}$

If $s_{ij} - \tilde{s_i} < \beta_i$, then $\Delta s_{ij} = 1$

Where $\tilde{s_i}$ is the median value of all s_i and β_i is the 75th percentile value of the positive distribution of $s_{ij} - \tilde{s_i}$. This effectively eliminates the autoactivators by taking only the top 1/8th (12.5%) of the s_i scores.

2.9 Statistical analysis and datasets used

I used R (version 3.1.0) to calculate Pearson's correlation coefficients with the 'stats' package (included in base library). I visualized interactomes with Cytoscape[38] (version 2.8.0). I used datasets from BioGRID (*Saccharomyces cerevisiae* v2.0.63).

3 Results

3.1 Reproducibility of the BFG-Y2H screens

The reproducibility of the internal replicates (two sets of barcodes within each diploid cell, Figure 7 panel C) is shown by several correlation plots (Figure 9 a-c), where the raw read counts (Up-Up fusion and Dn-Dn fusion) of fused barcodes have been transformed to a log scale. Pearson's correlation coefficients (PCC) were calculated with the raw read counts (all p-values were < 2.2e-16). The correlation value for the +His condition (0.72) is lower than the selective conditions (0.95 and 0.89 for –His and +3AT, respectively).

The +His population was used only to determine the marginal frequencies of the input barcodes; therefore, it was not necessary to obtain coverage of each individual barcode pair.

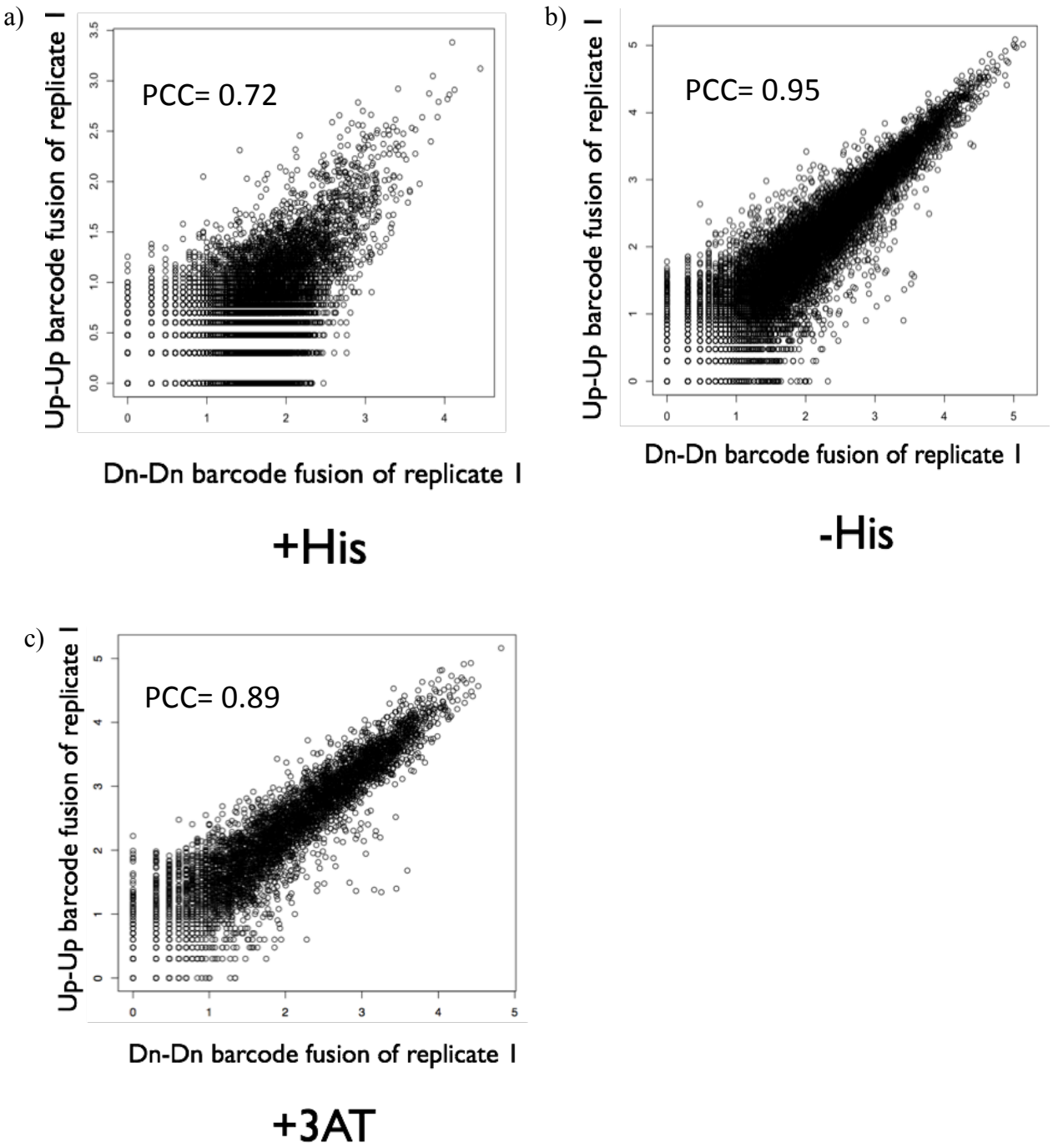


Figure 9. Reproducibility of BFG-Y2H between two internal replicates. Correlation plots for both non-selective (a) and selective conditions (b and c). The axes represent the raw read counts in log scale. The correlation is performed between the raw barcode counts. Pearson's correlation coefficients (PCC) are shown (all p-values were < 2.2e-16).

Reproducibility between biological strain replicates (barcode strain combinations, recall Figure 7 panel B) is shown in Figure 10 **a-c**. Similarly, the raw read counts of fused barcodes have been transformed to a log scale. PCCs were calculated with the raw read counts (all p-values were <2.2e16). The correlation value for the +His condition (0.55) is lower than the selective conditions (0.85 and 0.78 for –His and +3AT, respectively). There are fewer data points on these scatter plots because they only contain counts for barcode combination 1 and barcode combination 2 of replicate 1, whereas the previous plots contain all the unique barcode combinations.

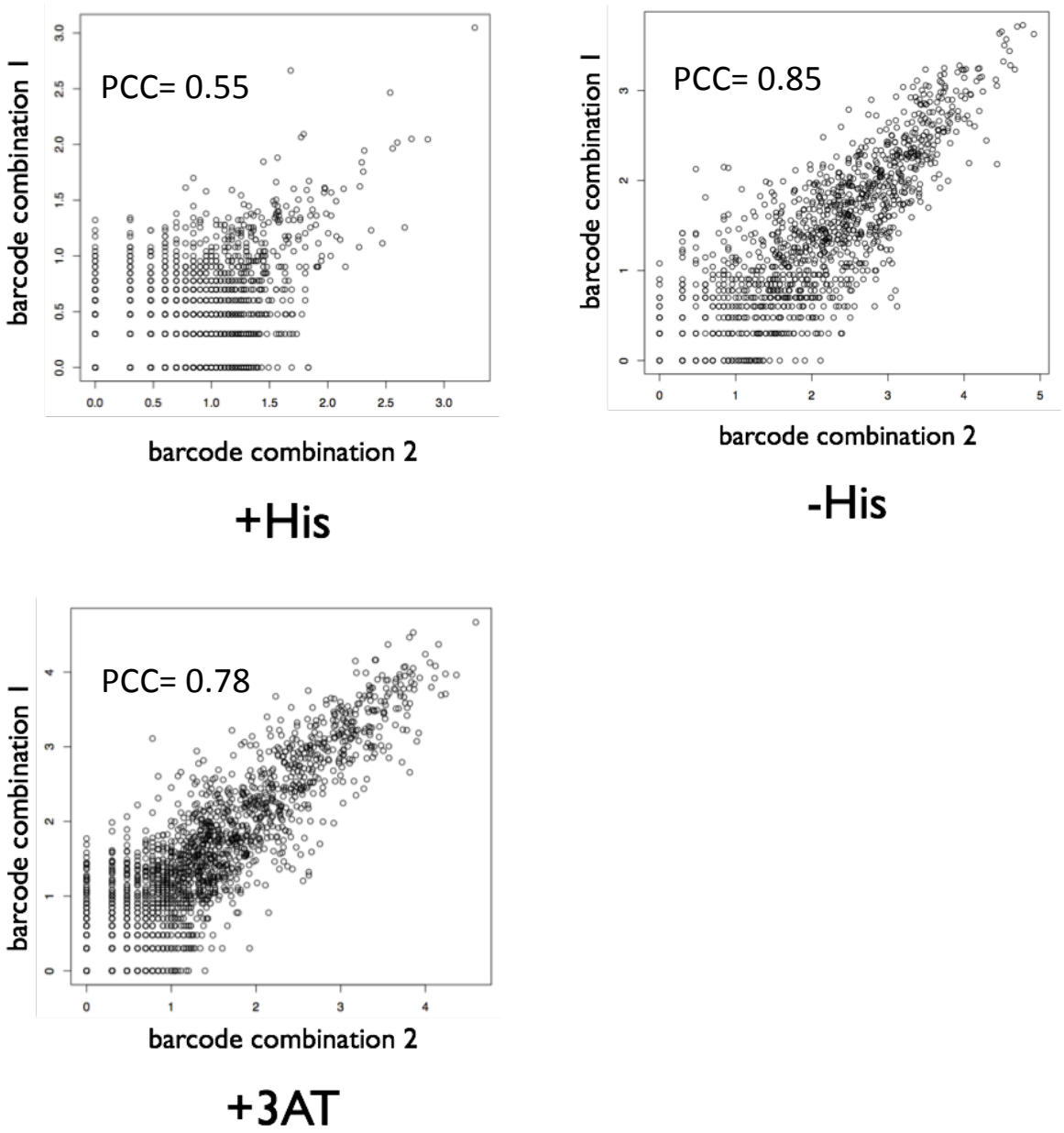


Figure 10. Reproducibility of BFG-Y2H between two biological strain replicates. These barcode combinations are diploid cells that were generated from differentially barcoded haploid cells. For all possible DB-X and AD-Y combinations, only combinations 1 and 2 are shown here (out of a total of up to 9 combinations). The axes represent the raw read counts in log scale. The correlation is performed between the raw barcode counts. Pearson's correlation coefficients (PCC) are shown (all p-values were < 2.2e-16).

Reproducibility between screen replicates is shown in Figure 11 a-c. Similarly, the raw read counts of fused barcodes have been transformed to a log scale. PCCs were calculated with the raw read counts (all p-values were < 2.2e-16). The correlation value for the +His condition (0.40) is lower than the selective conditions (0.85 and 0.80 for –His and +3AT, respectively).

Reproducibility of the positive interaction candidates from the screen replicates was also assessed in Figure 12. A cutoff of 1.5 (which yields a 50% pairwise retesting rate) for normalized scores was used for calling interactions. There was an overlap of 61 proteins between the two screen replicates out of a total of 183 interactions. The additional screen increased the number of positive protein interaction candidates by approximately 50%.

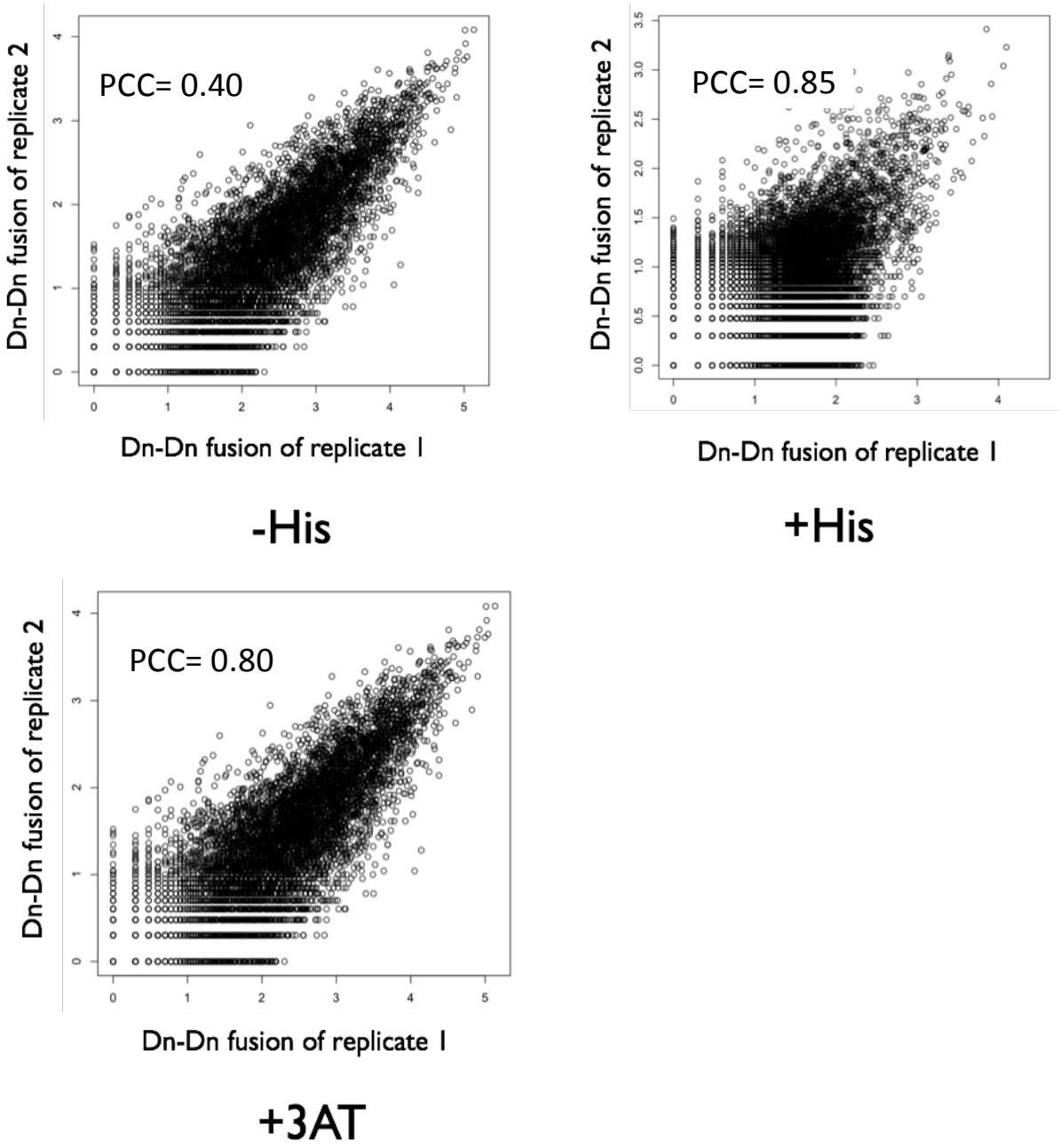


Figure 11. Reproducibility of BFG-Y2H between two screen replicates. Replicate 1 and 2 have been separated before the mating stage and onwards. The axes represent the raw read counts in log scale. The correlation is performed between the raw barcode counts. Pearson's correlation coefficients (PCC) are shown (all p-values were < 2.2e-16).

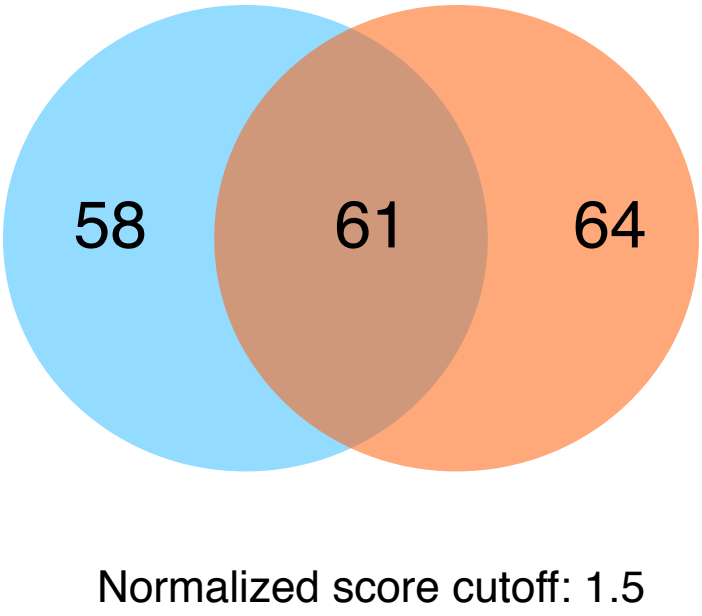


Figure 12. Reproducibility of BFG-Y2H between two screen replicates. This Venn diagram shows the overlapping positive interaction candidates between two BFG-Y2H screens. A cutoff of 1.5 for the normalized score was used to determine positive interaction candidates.

3.2 Positive controls in the BFG-Y2H screens

I included positive Y2H controls (denoted as "Positive Reference Set" or PRS from hereon) from the Vidal lab (in collaboration with Dr. Nidhi Sahni, Vidal lab) within my BFG-Y2H screens. The Vidal lab has previously used pairwise Y2H to test ~70,000 human protein pairs that were curated by the literature as positive binary interactions. This resulted in a list of ~700 positive Y2H interactions that pairwise retested as positives in the Y2H assay for more than two times (unpublished data from Dr. Nidhi Sahni, post-doctoral fellow, Vidal lab). In collaboration with Dr. N. Yachie and Dr. E. Petsalakis (Roth lab), 34 positive Y2H interacting pairs were randomly selected amongst the list of positive Y2H interacting pairs provided by the Vidal lab. These 34 positive Y2H interacting pairs were classified into two levels by the Vidal lab. Level 1 contains interacting pairs are reported as bidirectional by the Vidal lab, meaning that they are Y2H positives in both DB-X/AD-Y and DB-Y/AD-X configurations. Level 2 interaction pairs are reported as positives only in the DB-X/AD-Y direction. 15 of the 34 positive interactions were level 1 and 19 were level 2 interactions.

The positives are positioned along the diagonal of the matrices. The ideal results would be to see the diagonal lit up under selective Y2H conditions (Figure 13). The Vidal lab has not performed pairwise testing for off-diagonal protein pairs; therefore, this space is less known and not necessarily negative for Y2H protein interactions.

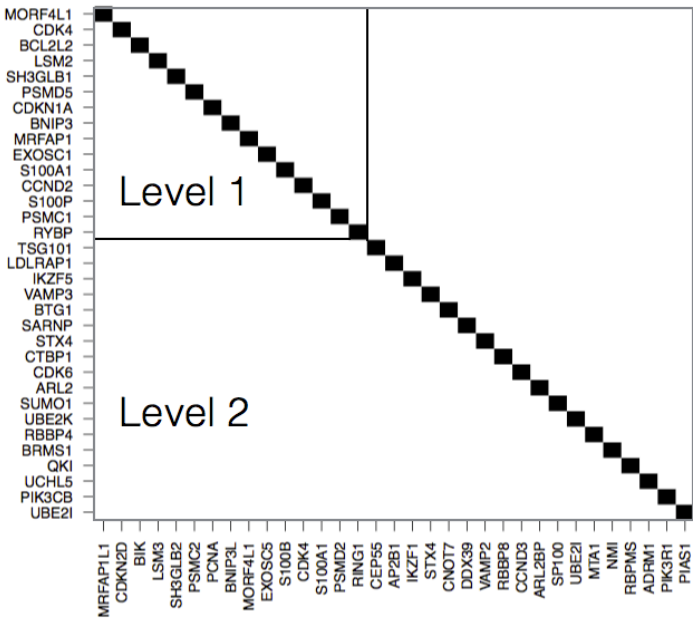


Figure 13. Expected positive interactions for BFG-Y2H. Protein pairing information was obtained from the Vidal lab. There are a total of 34 expected Y2H positive protein interactions (black) along the diagonal divided into two levels. Level 1 (15/34) interacting pairs are reported as bidirectional by the Vidal lab, meaning that they are Y2H positives in both DB-X/AD-Y and DB-Y/AD-X configurations. Level 2 (19/34) interaction pairs are reported as positives only in the DB-X/AD-Y direction. The off-diagonal space hasn't been subjected to pairwise retesting with Y2H, therefore, this space is unknown and not necessarily negative for Y2H.

As discussed in section 2.8 (Calculation of interaction scores), the +His condition (Figure 15) serves as a background level to normalize of the barcode counts for both the –His (Figure 16) and +3AT (Figure 17) conditions such that positive interactions can be quantitatively identified. Spotty sequencing coverage for strains at the +His level is not considered to be important because marginal frequencies (**Section 2.8** Calculation of interaction scores) are used for downstream normalization (please see Figure 19 for the entire matrix of marginal frequencies). Figure 15 shows the raw barcode counts for the –His selective condition before normalizing against the +His matrix and within each row. All values have been transformed to a log10 scale.

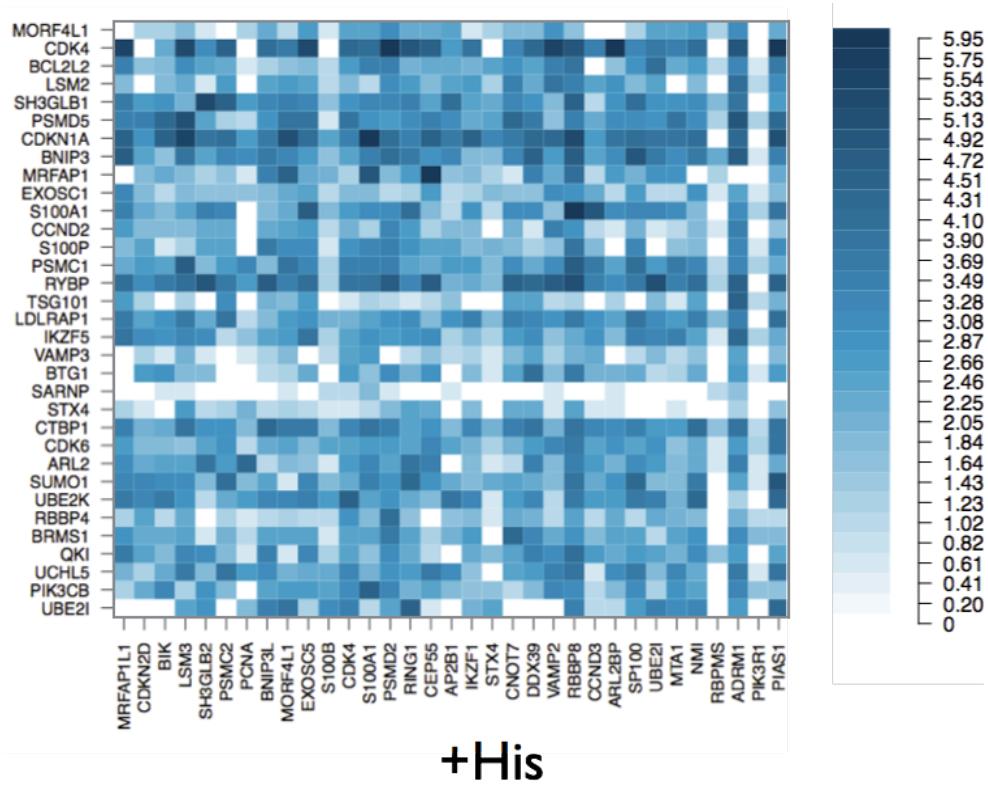
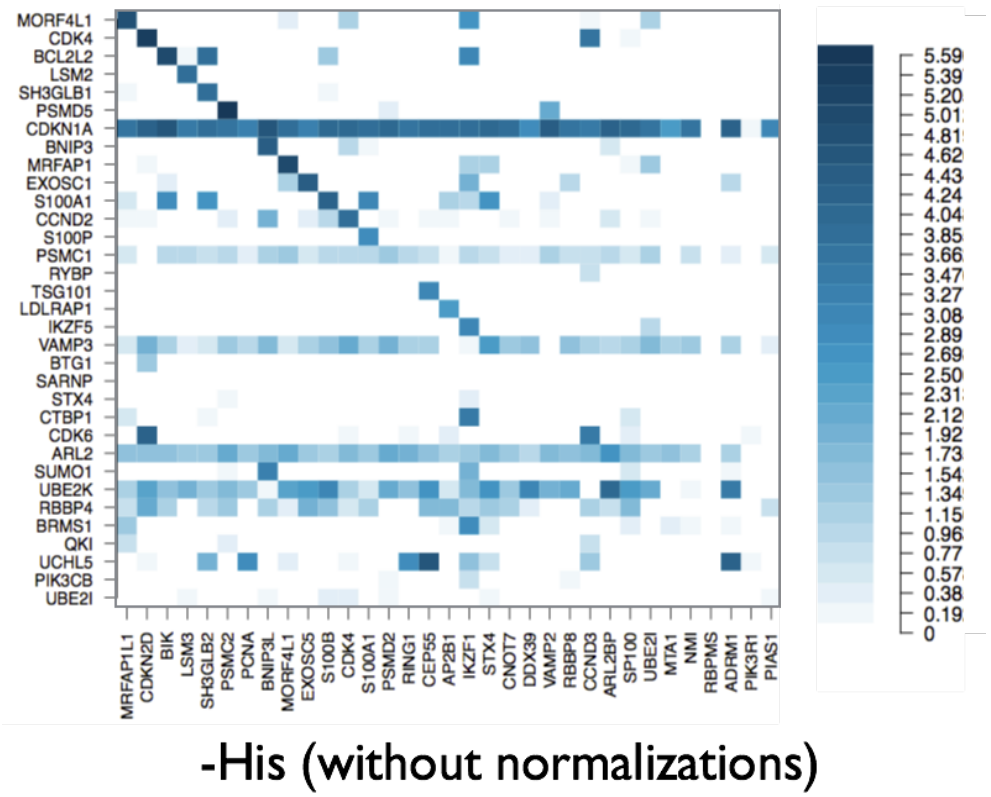
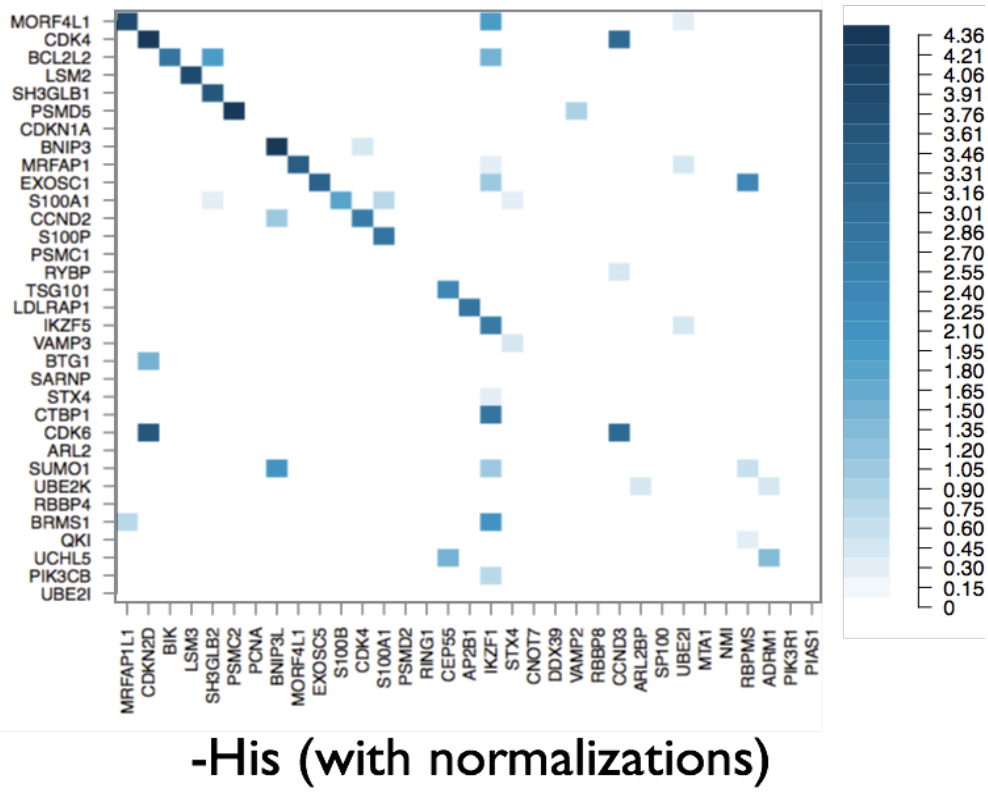


Figure 14. Positive controls of the BFG-Y2H screen in the presence of histidine (+His). This is a non-selective condition, giving a pessimistic view (due to limited sequencing coverage) of the complexity of the initial pool of diploid strains, as well as providing reference values of marginal abundance for normalization of the scores in the selective conditions downstream. The values in the heatmap are the raw barcode counts transformed to a log10 scale. The DB-X (prey) proteins are on the Y-axis and the AD-Y (bait) proteins are on the X-axis. Dark blue represents an abundance of barcode counts, whereas white represents an absence of barcode counts.



-His (without normalizations)

Figure 15. Raw barcode counts of positive controls of the BFG-Y2H screen without histidine supplements. This is a selective condition whereby positive protein interactions and autoactivators are captured. The values in the heatmap are raw barcode counts transformed to a log10 scale. These values have not been normalized against the +His matrix, nor normalized within each row. The DB-X (prey) proteins are on the Y-axis and the AD-Y (bait) proteins are on the X-axis. Dark blue represents an abundance of barcode counts, whereas white represents an absence of barcode counts.



-His (with normalizations)

Figure 16. Normalized scores of positive controls of the BFG-Y2H screen without histidine. This is a selective condition whereby positive protein interactions and *de novo* autoactivators captured. The values in the heatmap have been normalized against the +His matrix and also normalized within each row. They have also been transformed to a log10 scale. The DB-X (prey) proteins are on the Y-axis and the AD-Y (bait) proteins are on the X-axis. Dark blue represents high normalized scores, hence, positive protein interaction candidates. White represents low normalized scores, hence, no protein interaction candidates.

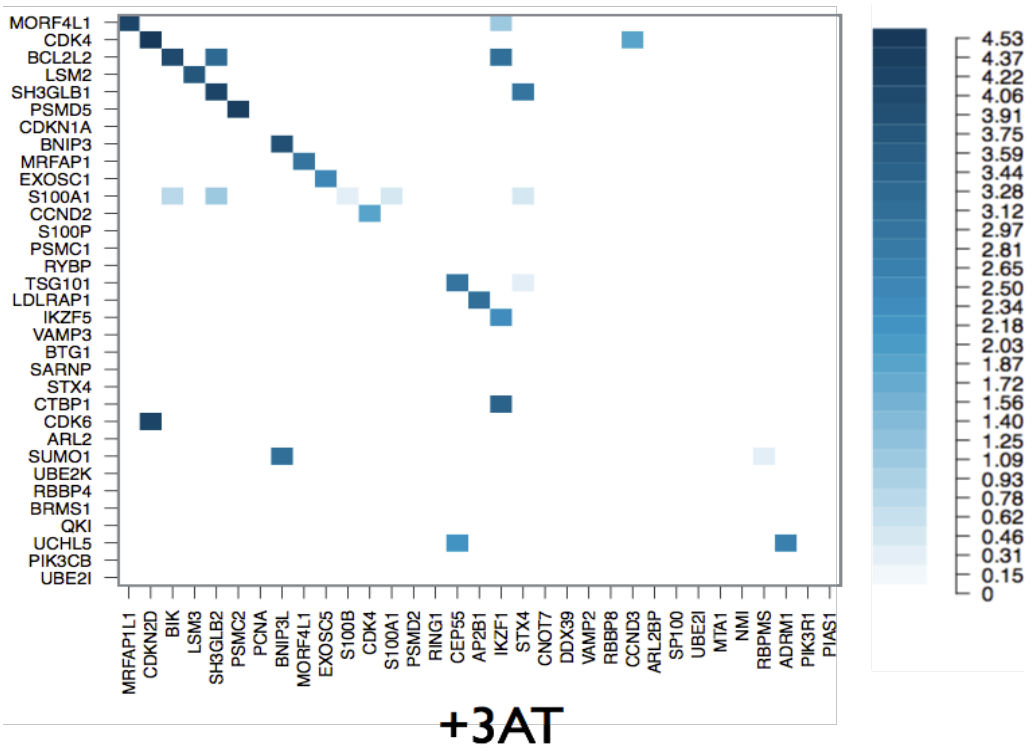


Figure 17. Normalized scores of positive controls of the BFG-Y2H screen without histidine and in the presence of 3-amino-1,2,4-triazole (+3AT). This is a selective condition that captures positive protein interactions and de novo autoactivators and it is more stringent than +His. The values in the heatmap been normalized using the marginal barcode abundance values in the +His matrix and also normalized within each row. They have also been transformed to a log10 scale. The DB-X (prey) proteins are on the Y-axis and the AD-Y (bait) proteins are on the X-axis. Dark blue represents high normalized scores, hence, positive protein interaction candidates. White represents low normalized scores, hence, no protein interaction candidates.

3.3 Untangling positive controls

To address the question of why the BFG-Y2H screens failed to pick up 20 out of 34 (58.8%) of the PRS interactions, I conducted several pairwise-retesting experiments, where each pair was tested one-on-one per **Section 2.7** (Pairwise retesting**).** A smaller matrix of the original –His matrix for the positive controls was generated for positive controls that tested positive in pairwise-retesting in the –His condition (Figure 18), where all protein pairs that tested positive in the pairwise-retesting experiment are along the diagonal.

Only 17 protein pairs out of the initial 34 PRS protein pairs retested positive, and three out of these 17 protein pairs were not detected by the BFG-Y2H screen (PSMC1 and PSMD2, BTG1 and CNOT7 and ARL2 and ARL2BP). As previously mentioned, the off-diagonal space has not been pairwise-retested, therefore, positive interactions in this space could be genuine. For example, the interaction between CDK6 and CDKN2D has been demonstrated through a crystal structure[39]. The CDK4 and CCND3 interaction has also been previously shown through Y2H screens[37] as well as affinity capture[40,41].

Additional pairwise-retesting experiments were performed to check the effect of different methods of clone generation on the Y2H results. I compared the pairwise-retesting experimental results from two different methods of generating clones: Gateway cloning and in-yeast assembly[42] and discovered that Gateway cloning performed better than in-yeast assembly, however, the two were not statistically different (see **Appendix 7.4**).

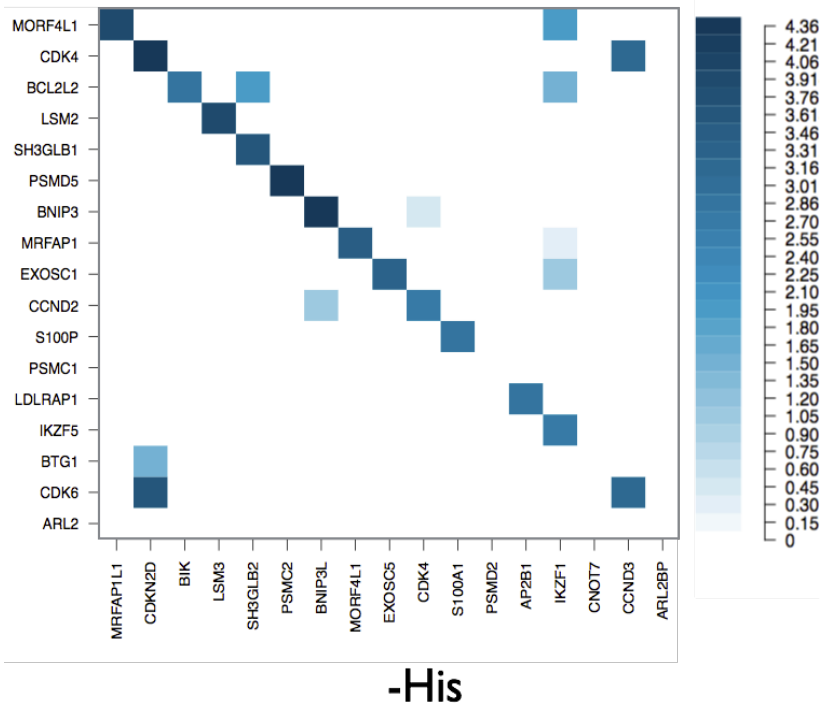


Figure 18. Readjusted positive controls for the BFG-Y2H screens without histidine. Pairs along the diagonal for the matrix are all protein pairs that tested positive in pairwise retesting. This figure is a smaller matrix of Figure 14 and contains only pairs that retested positive. The values in the heatmap have been normalized using the +His matrix and also normalized within each row. They have also been transformed to a log10 scale. The DB-X (prey) proteins are on the Y-axis and the AD-Y (bait) proteins are on the X-axis. Dark blue represents high normalized scores, hence, positive protein interaction candidates. White represents low normalized scores, hence, no protein interaction candidates.

3.4 BFG-Y2H screens

In order to uncover novel interactions amongst DNA damage repair proteins, I completed the BFG-Y2H screens with the DNA damage repair proteins and the PRS in all three conditions: marginal frequencies of +His (Figure 19), -His (Figure 21) and +3AT (Figure 22). The raw barcode counts in the +His condition ranged from 0 to 5.95 on a log10 scale (or 0 to 890,000). The distributions of the marginal frequencies of the DB-X and AD-Y pools in the +His are shown in Figure 20 (see **Section 2.8** (Calculation of interaction scores) for more details for calculating the marginal frequencies). The majority of the marginal frequencies range between 10^{-2} and 10^{-3}. Please see Figure 23 for the overlap of positive interactions between –His condition and +3AT condition.

The expected interactome (BioGRID Y2H datasets) from all of the DNA damage repair proteins present in the screen are shown in Figure 24.

. Please see **Appendix 7.5** for the top 200 interactions and their corresponding normalized scores.

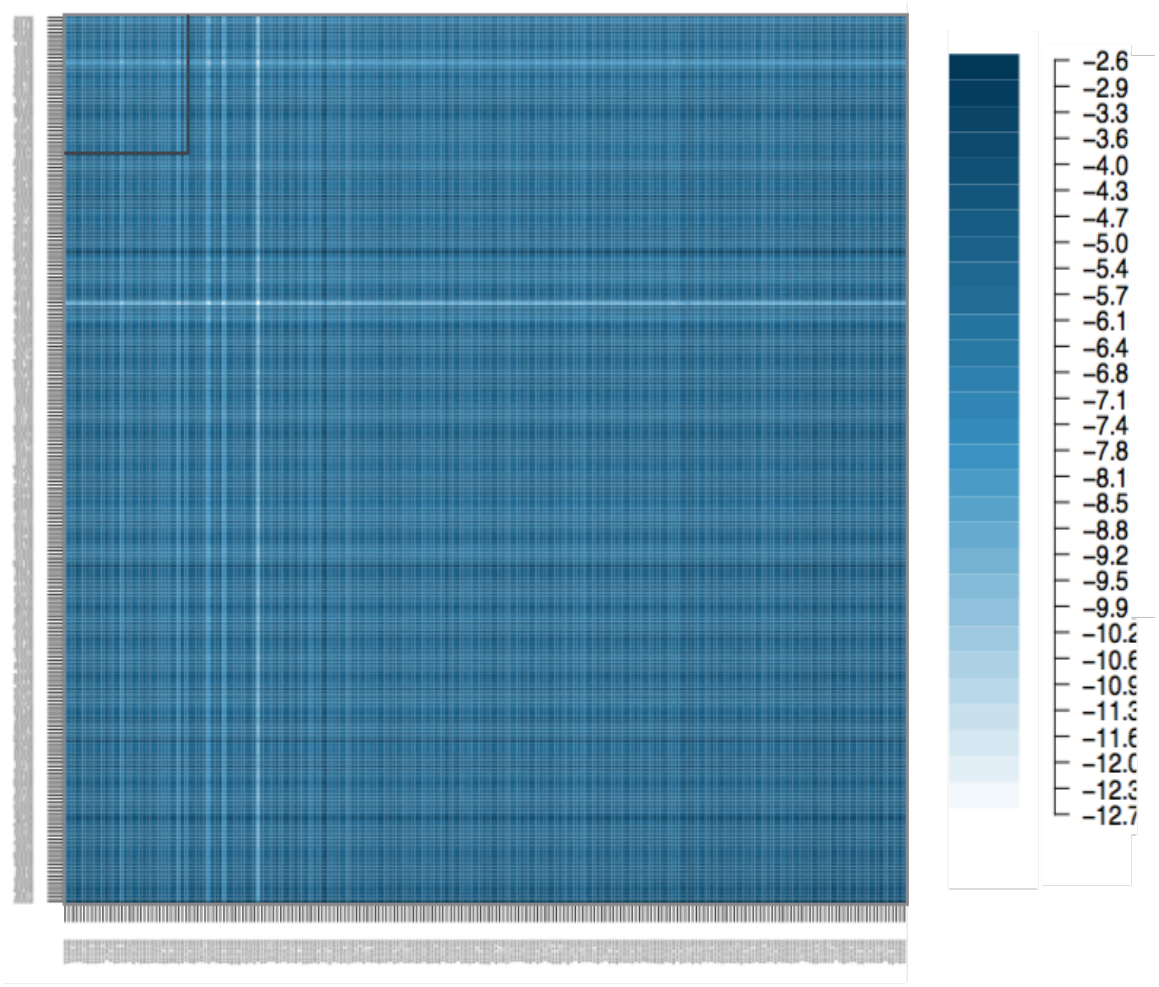


Figure 19. Marginal counts of the BFG-Y2H screen without histidine supplements. This is a non-selective condition, giving an overall view of the complexity of the diploid strains, as well as providing the background values for normalization of the scores in the selective conditions downstream. The values in the heatmap are the marginal frequencies transformed to a log10 scale. The PRS controls are within the square in the upper-left corner. The DB-X (prey) proteins are on the Y-axis and the AD-Y (bait) proteins are on the X-axis. Dark blue represents high marginal frequencies, whereas white represents low marginal frequencies.

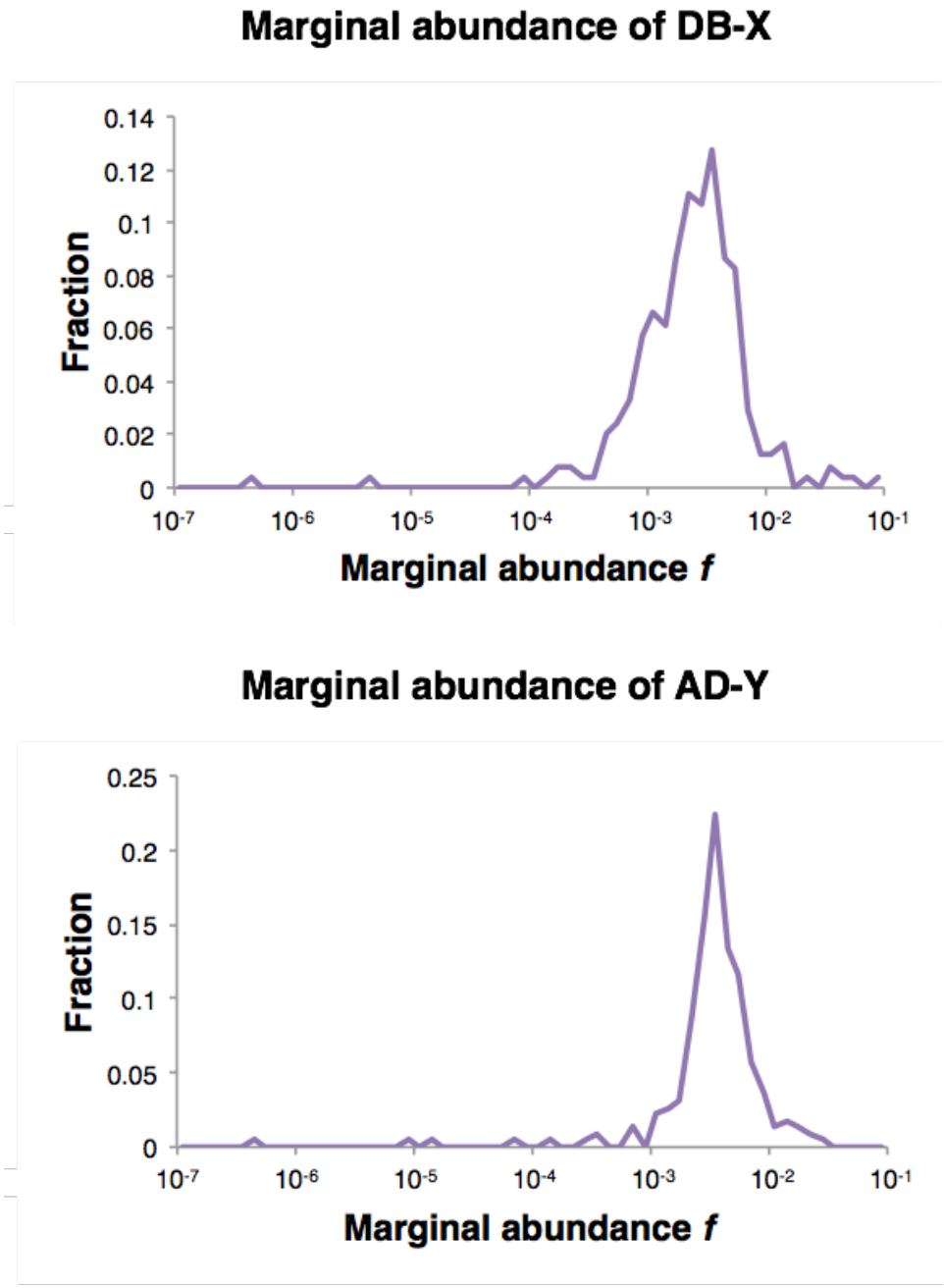


Figure 20. Distribution of marginal abundance of DB-X and AD-Y in +His. The marginal abundance (also known as marginal frequencies) for both DB-X and AD-Y in +His are calculated from **Section 2.8** (Calculation of interaction scores).

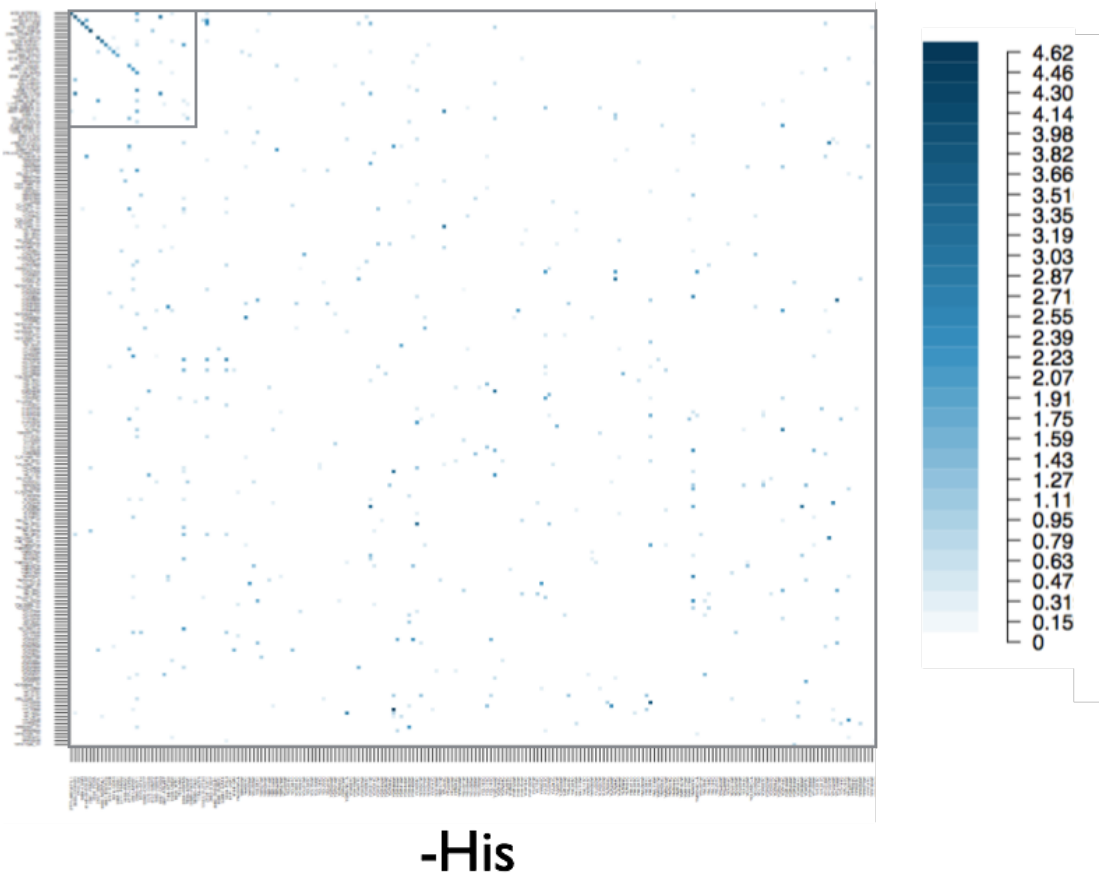


Figure 21. Normalized scores of the BFG-Y2H screen without histidine supplements. This is a selective condition whereby positive protein interactions and *de novo* autoactivators are captured. The values in the heatmap been normalized against the +His matrix and also normalized within each row. They have also been transformed to a log10 scale. The PRS controls are within the square in the upper-left corner. The DB-X (prey) proteins are on the Y-axis and the AD-Y (bait) proteins are on the X-axis. Dark blue represents high normalized scores, hence positive protein interaction candidates. White represents low normalized scores, hence, no protein interaction candidates.

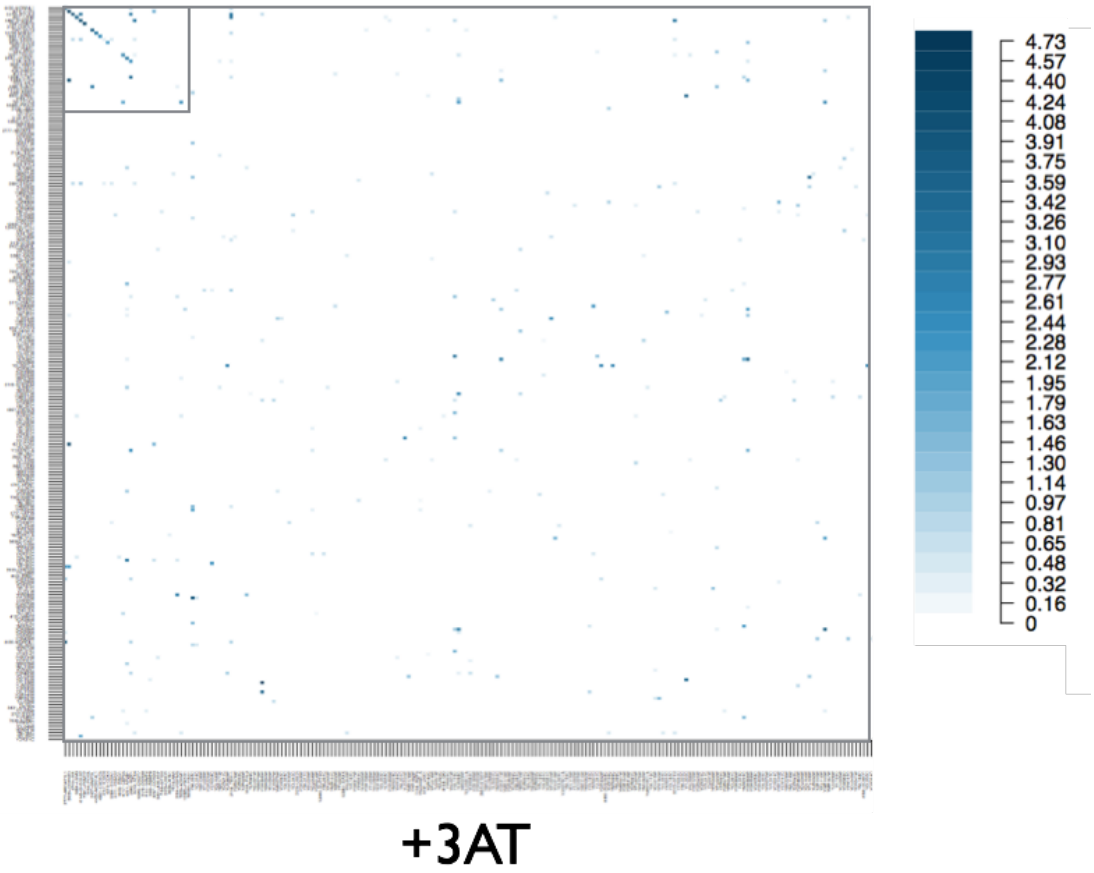


Figure 22. Normalized scores of the BFG-Y2H screen without histidine supplements and in the presence of 3-amino-1,2,4-triazole (+3AT). This is a selective condition that captures positive protein interactions and it is more stringent than –His, due to the addition of a chemical inhibitor of histidine synthesis, 3AT. The values in the heatmap been normalized against the +His matrix and also normalized within each row. They have also been transformed to a log10 scale. The PRS controls are within the square in the upper-left corner. The DB-X (prey) proteins are on the Y-axis and the AD-Y (bait) proteins are on the X-axis. Dark blue represents high normalized scores, hence positive protein interaction candidates. White represents low normalized scores, hence, no protein interaction candidates.

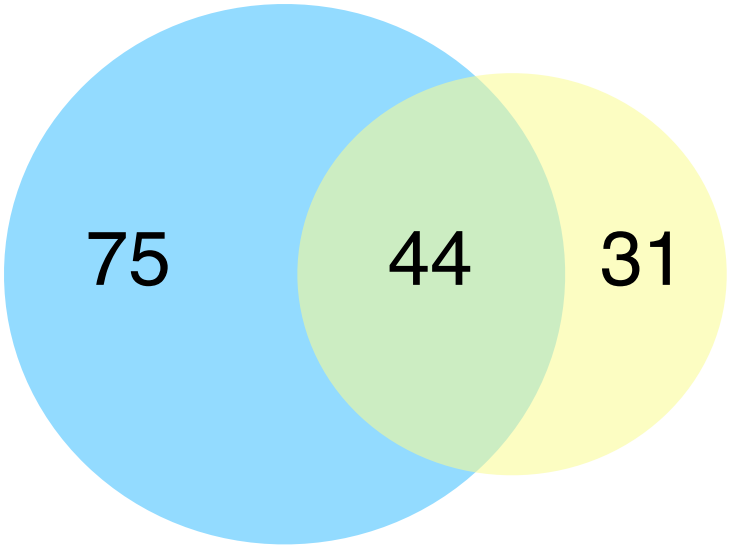


Figure 23. Overlap between –His and +3AT conditions. The blue circle represents the positive interactions found in the -His screen above a normalized score of 1.5. The yellow circle represents the positive interactions found in the +3AT screen above a normalized score of 1.5. The two screens have an overlapping 44 interactions.

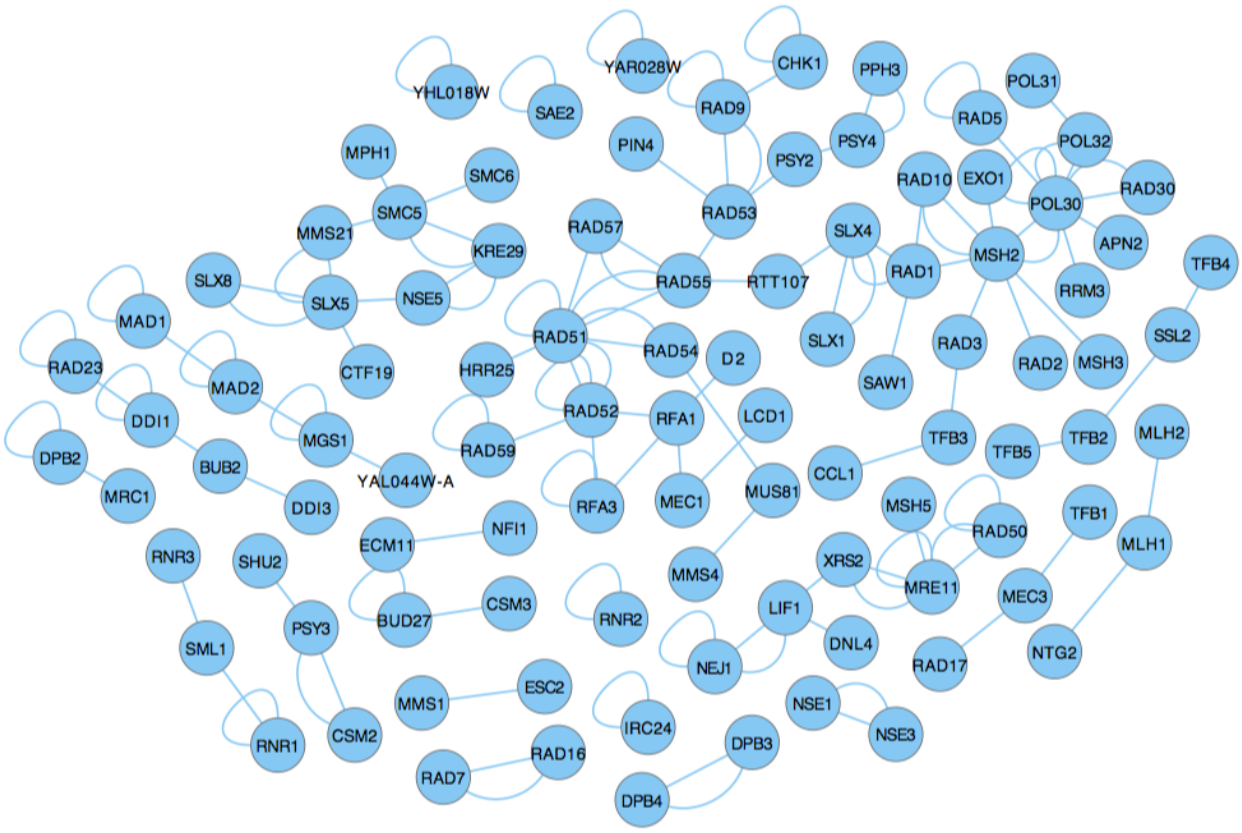


Figure 24. The expected interactome. The expected interactome from BioGRID Y2H interactions. The DNA damage repair proteins that have not been barcoded successfully have been removed from this interactome.

3.5 Benchmarking for pairwise retesting

A threshold is needed to prioritize and select positive protein interaction candidates for pairwise retesting. This threshold should optimize the balance between false negatives and false positives, which means that a set of true interactions is needed for benchmarking. Arguably, the highest quality Y2H screen of the yeast proteome was done by Yu *et al* (2008)[43]. The background strains and the plasmids used by Yu *et al.* are also the closest to my strains used in this screen. Therefore, the overlap between the positive interactions from Yu *et al.* and the DNA-damage repair proteins in the matrix was used as a reference set of high-confidence PPIs (11 positive interactions). The true positives also include the previously mentioned PRS controls from the Vidal lab I pairwise retested to be positive interactions (17 true positive interactions, Figure 18). I plotted precision, recall and the Matthews correlation coefficient[44] (MCC). The MCC is a measurement of the quality of binary classifications. It optimizes between precision and recall, and it has a range of values from -1 (disagreement between prediction and observation) to 1 (perfect agreement between prediction and observation). The maximum precision is around 0.5. The best MCC was 0.33, corresponding to a normalized score of 3.25, which identified 30 candidate protein interactions.

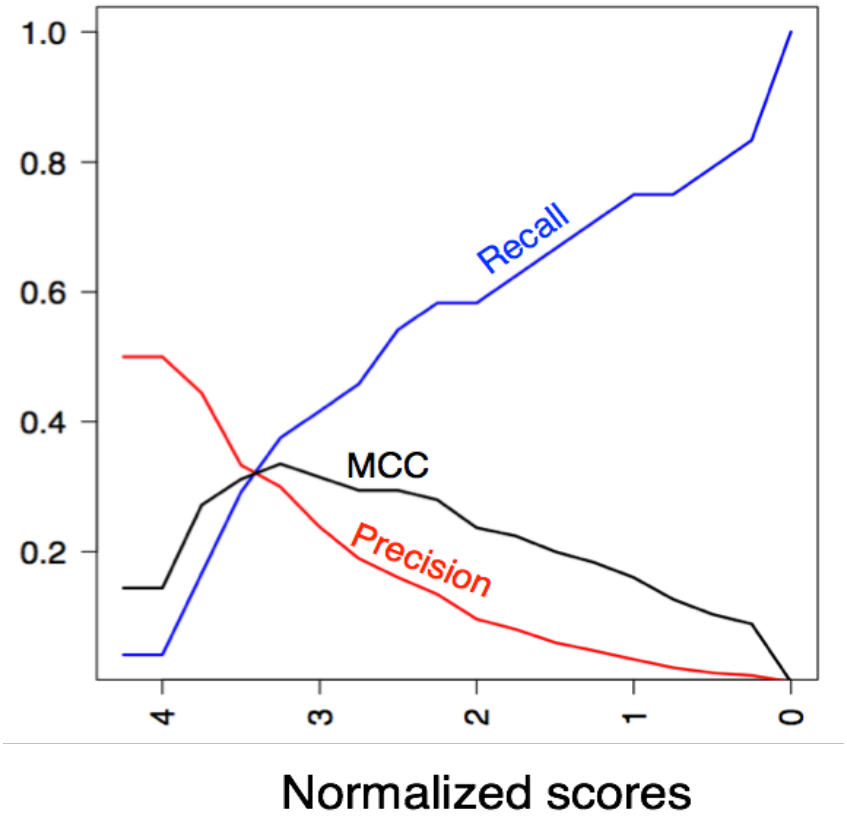


Figure 25. Prediction performance of the reported BFG-Y2H positive protein interactions for the –His condition benchmarked against Yu *et al.* [43]. The "true positive" interactions are the union of the PRS controls from the Vidal lab and the positive interactions from Yu *et al.* MCC denotes Matthew correlation coefficient, which optimizes between precision and recall.

In addition to the high-quality Y2H screen of the yeast proteome from Yu *et al* (2008)[43], many other high-throughput Y2H datasets are also included in BioGRID (such as Uetz *et al.*[45] and the "core" subset of Ito *et al.*[46], where only interactions that tested at least twice were included). Therefore, I benchmarked my dataset against all of the Y2H datasets available in BioGRID (128 positive Y2H interactions), as well as the previously pairwise retested PRS pairs from the Vidal lab (17 PRS Y2H interactions). I plotted precision, recall and the MCC (Figure 26). The maximum precision is around 0.5 as in the previous comparison (Figure 25). The maximum MCC was around 0.18, which corresponds to a normalized score of 2.5, yielding 81 candidate interactions.

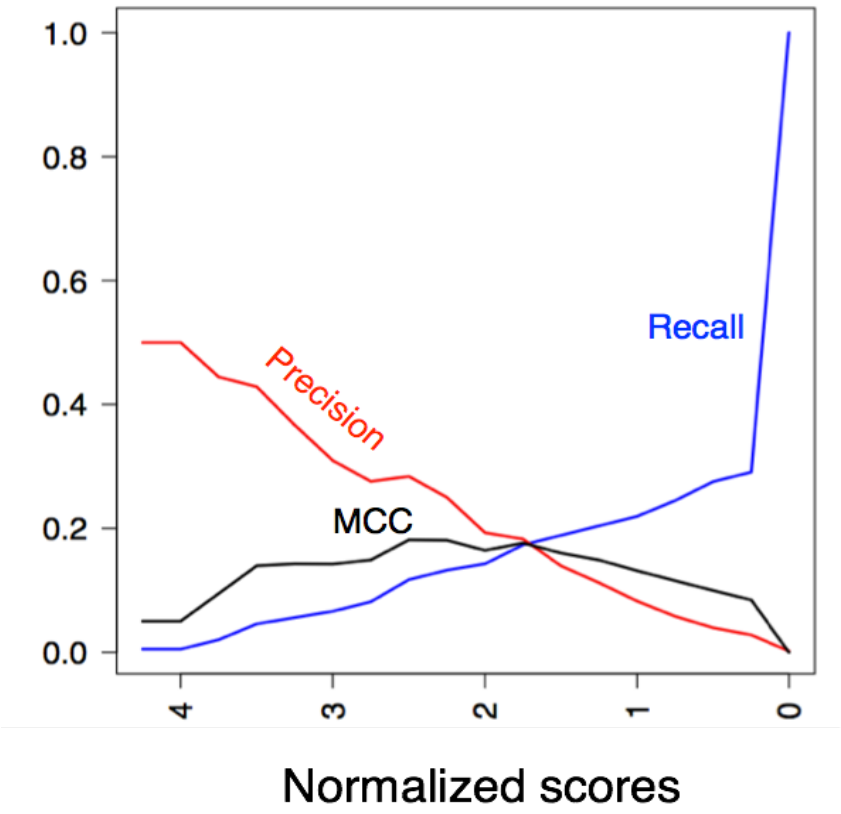


Figure 26. Prediction performance of the reported BFG-Y2H positive protein interactions for the –His condition benchmarked against all previous BioGRID Y2H data. In this case, the true positive interactions are the BioGRID Y2H data and the pairwise retested PRS interactions. MCC denotes Matthews correlation coefficient, which optimizes between precision and recall.

3.6 Pairwise retesting on BFG-Y2H candidate interactions

From the two precision and recall curves, it appears that benchmarking against the Yu et al.[43] dataset yielded better MCC and recall values. However, using the best MCC score of 0.33 as a cutoff, it would mean that we would only test the top 30 interactions from the BFG-Y2H primary hits. Instead, 52 positive interactions from the primary BFG-Y2H screen were arbitrarily selected amongst the top 185 hits and their interactions were tested with pairwise retesting (**Section 2.7** Pairwise retesting). The pairwise retesting results are shown in Figure 27; 46% (24) of the 52 candidate pairs were confirmed by pairwise retesting. Of these 24 confirmed interactions, 17 of them were novel relative to the Yu *et al.*[43] dataset, and seven were novel compared to all the Y2H data from BioGRID (Table 3). After doing a thorough search, five of these seven confirmed interactions were novel (have not been previously reported by previous literature).

I created an interactome map with interactions that were confirmed by pairwise retesting (Figure 28). The colour of the edges represent where the protein interaction was seen previously in the literature. The width of the edges represents the normalized interaction score. All novel interactions are shown in red.

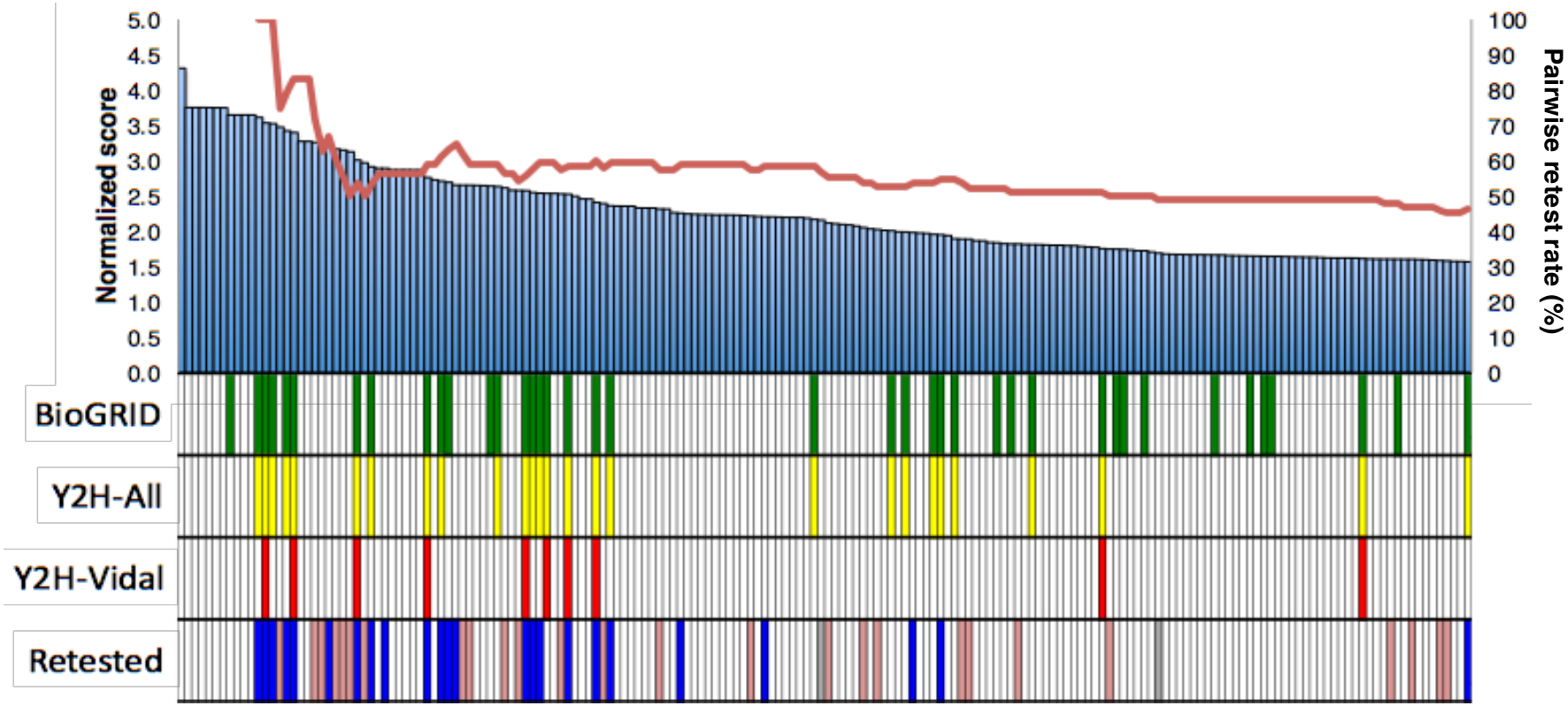


Figure 27. Pairwise retesting results of the top scores from the BFG-Y2H screens. The BFG-Y2H normalized scores for the –His condition are on the left Y-axis. The moving average of the pairwise retesting confirmation rate (%) is on the right Y-axis andis highlighted in red. Previously reported BioGRID (green), all reported Y2H interactions (yellow) and Yu et al. interactions (red) are marked. Interactions that were retested but were not positive in Y2H are in light pink, interactions that were retested and confirmed to be genuine positives are in dark blue and interactions that were retested but were found to be autoactivators are in grey.

DB-X (standard name)	DB-X (systematic name)	AD-Y (standard name)	AD-Y (systematic name)	Normalized score	Yu *et al.* novel	"All Y2H" novel	All literature novel
MEC3	YLR288C	*RAD17*	YOR368W	3.612	✔		
RAD1	YPL022W	*RAD10*	YML095C	3.533	✔		
SLX1	YBR228W	*SLX4*	YLR135W	3.523	✔		
MEC3	YLR288C	*TFB1*	YDR311W	3.421	✔		
THI4	YGR144W	*THI4*	YGR144W	3.398	✔		
MUS81	YDR386W	*YPL108W*	YPL108W	3.192		✔	
CSM2	YIL132C	*PSY3*	YLR376C	3.007	✔		
TFB1	YDR311W	*MEC3*	YLR288C	2.913	✔		
CHL1	YPL008W	*CSM3*	YMR048W	2.891		✔	✔
PSY3	YLR376C	*CSM2*	YIL132C	2.76	✔		
POL32	YJR043C	*POL31*	YJR006W	2.709	✔		
RMI1	YPL024W	*TOP3*	YLR234W	2.693		✔	
IRC15	YPL017C	*MIG3*	YER028C	2.653		✔	✔
YHL018W	YHL018W	*YHL018W*	YHL018W	2.575	✔		
POL31	YJR006W	*POL32*	YJR043C	2.548	✔		
XRS2	YDR369C	*MRE11*	YMR224C	2.539	✔		
DPB3	YBR278W	*DPB4*	YDR121W	2.524	✔		
RAD1	YPL022W	*SLX4*	YLR135W	2.407	✔		
CHK1	YBR274W	*RAD9*	YDR217C	2.361	✔		
LNP1	YHR192W	*KRE29*	YER038C	2.256		✔	✔
CRT10	YLR118C	*MIG3*	YER028C	2.204		✔	✔
YLR118C	YDR217C	*MIG3*	YBR274W	1.981		✔	✔
RAD9	YPL017C	*CHK1*	YOR368W	1.95	✔		
NEJ1	YLR265C	*LIF1*	YGL090W	1.567	✔		

Table 3. Novel Y2H positive interactions confirmed by pairwise retesting. Seventeen novel Y2H positive interactions were determined through pairwise retesting when compared to the Yu *et al.* dataset, seven of which are novel compared to all the Y2H data from BioGRID. Five were completely novel when compared to all previously reported interactions in the literature.

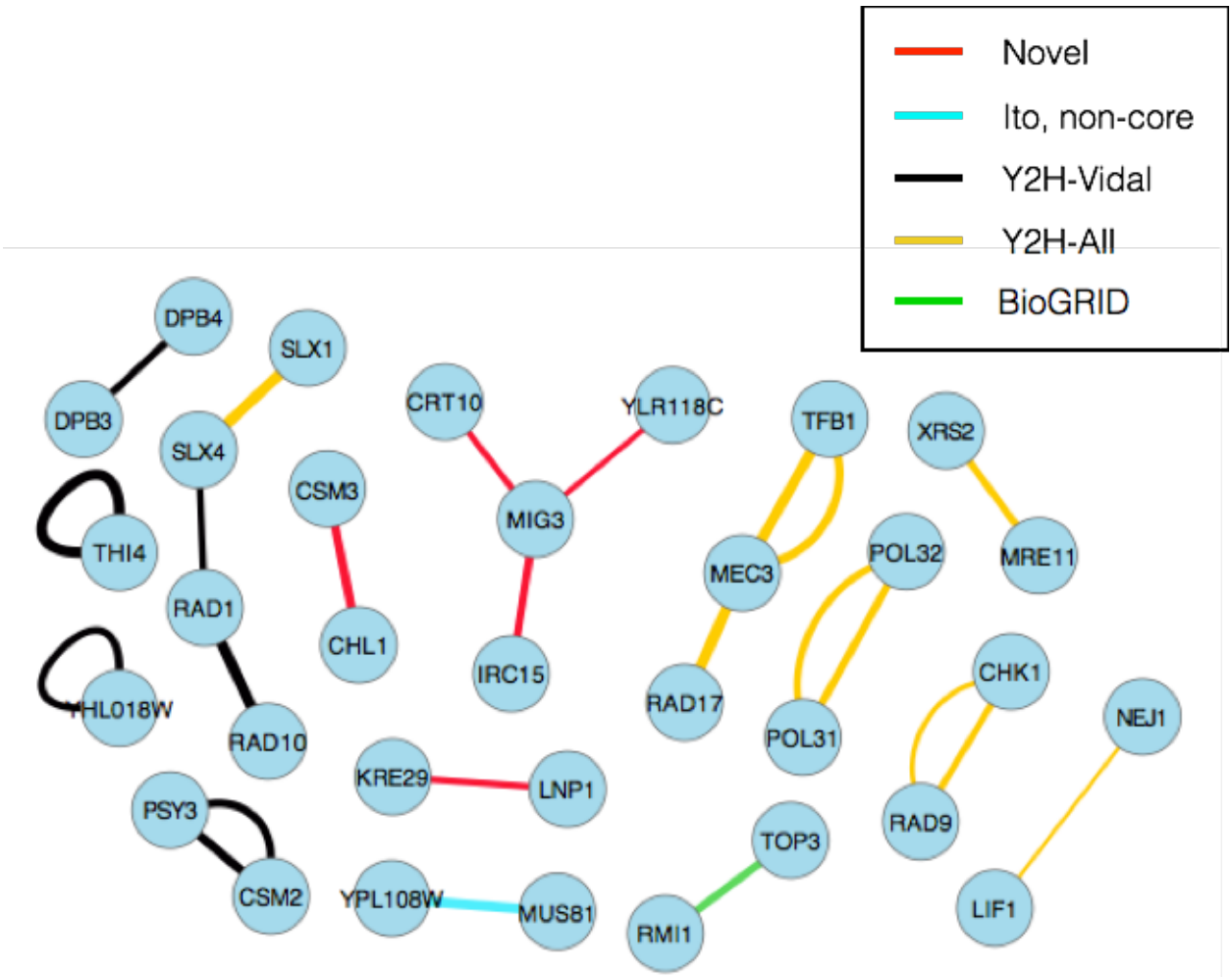


Figure 28. Interactome of the top 200 scored interactions from the BFG-Y2H screen (-His condition). All 24 Y2H protein interactions that were confirmed by pairwise retesting are shown here. Novel interactions are in red. Previously known interactions are shown as follow: Ito, non-core (teal), Y2H-Vidal (black), all Y2H (yellow), BioGRID (green). Two lines indicate that an interaction was confirmed in both DB-X/AD-Y and DB-Y/AD-X directions. The normalized scores represent the edge widths.

4 Discussion

4.1 Reproducibility of BFG-Y2H

Reproducibility was assessed in three ways (recall Figure 7): the screen replicates of the BFG-Y2H screen (separated before the mating stage and onwards), biological strain replicates (different barcode combinations), and internal barcode replicates (UP-UP and DN-DN within each diploid cell). For all three kind of replicates, their raw barcode counts were treated to generate three different sets of data: sums of raw barcode counts, maximum count for the raw barcode counts and third quartile of the raw barcode counts. The three sets of data were compared to each other, and the sums of raw barcode counts were used for normalization because they yielded better precision and recall curves. ORFs without any raw barcode counts were manually removed from the expected matrix.

Reproducibility was the highest amongst the internal replicates (PCC values of 0.72, 0.95 and 0.89 for the +His, -His and +3AT experiments, respectively). Correlation for the +His condition was lower than the selective conditions, likely because the barcodes were not sequenced to saturation given the higher complexity of this pool. The high reproducibility overall between the internal replicates indicates that barcode abundance was not greatly affected by PCR or sequencing biases on the barcodes. Reproducibility was slightly lower amongst the biological replicates (barcode strain combinations) (PCC values of 0.55, 0.85 and 0.78 for the +His, -His and +3AT experiments, respectively) compared to that of the internal replicates. This is expected because variations in mating efficiency are introduced for the barcode combinations, hence lowering the correlation between replicates compared to that of the fused barcodes within each

diploid. Similar to the internal barcode replicates, the reproducibility for the +His condition was lower than the selective conditions. Reproducibility was the lowest for screen replicates (PCC values of 0.40, 0.85, and 0.80 for the +His, -His and +3AT experiments, respectively). It is expected for these correlation values to be lower than that of the internal barcode replicates, due to variations in mating efficiency and stochastic differences rising from a population complexity bottleneck. These correlation values are very similar to the barcode combination replicates.

For all three kinds of replicates, correlation values between barcode counts were all extremely high, indicating that the BFG-Y2H screens are highly reproducible.

The reproducibility of the interactions was also examined between two BFG-Y2H screen replicates. Here, an additional screen provided a ~50% increase in the number of positive protein interaction candidates. This is because the screen is not saturated in terms of finding new interactions, in other words, due to the sampling sensitivity of the screen, multiple screens are needed to uncover all possible protein interactions. This was demonstrated by Yu *et al.*[43], where it took eight screen replicates to saturate the number of uncovered interactions. The number of new interactions uncovered after the eighth screen remained constant, which is the background noise of the screen. Yu *et al.*'s first screen uncovered approximately 65% of the total number of interactions and the second screen uncovered 78% of the total number of interactions, meaning that the second screen provided a ~20% increase in the number of positive protein interaction candidates. Because I only performed two screen replicates, it is unclear why my second screen provided an additional ~50% interactions, which is more than twice the percentage of new positive protein interaction candidates compared to Yu *et al.*[43] (~20%). Perhaps our approach will require fewer screens to reach saturation in terms of uncovering protein interactions. Alternatively, it could be due to our stochastic differences in screen replicates. Even though raw barcode counts for screen replicates have a high correlation in the –His condition, it doesn't

necessarily mean that they will have a high correlation for calling positive interactions. For example, a DB-X protein that is an autoactivator would show high correlation in –His condition for raw barcode counts between screen replicates. However, due to stochastic differences, the relative ranking of individual protein interactions could change significantly more than non-autoactivators. This is because all protein pairs containing an autoactivator have high barcode counts in the –His condition, whereas protein pairs not containing an autoactivator would have very low barcode counts, except in the case of a positive interaction. Therefore, the relative rankings are more likely to change for autoactivators than for non-autoactivators due to a smaller dynamic range. The reported screen replicates rates by Yu *et al.*[43] excluded autoactivators and this could potentially explain the differences between the percentages of additional positive interactions added by the second screen.

The distribution of the marginal frequencies in +His condition for DB-X and AD-Y pools were mostly between 10^{-2} and 10^{-3}, which is expected for a pool of 260 ORFs ($1/260 = \sim 4 \times 10^{-3}$) assuming even distribution. The standard deviation for the distribution of the marginal frequencies for DB-X is slightly higher than that of the AD-Y pool.

4.2 Positive controls in the BFG-Y2H screens

I showed through pairwise retesting that only 17 out of the 34 PRS interactions were positive. Out of these 17 interactions, 14 were picked up by BFG-Y2H, demonstrating that BFG-Y2H works quite well, at least in the current size matrix of 280 x 280 (78400) protein pairs, or ~840 x 840 counting the multiple barcodes assigned for each ORF.

It is unclear why 17 of the 34 PRS interactions did not test positive in pairwise retesting. All of the selected 34 PRS proteins had their ORFs Sanger sequenced again in our lab to confirm

their identities. Perhaps they are due to strain differences in our Y2H background strains (recall **Section 2.1** Barcode-Fusion Genetics: background strains), or mutations on parts of the plasmids that were not sequenced (for example, a non-functional nuclear localization signal).

4.3 BFG-Y2H screen and pairwise retesting results

When benchmarked against the Yu *et al.*[43] dataset for precision and recall, the maximum precision was around 0.5. This is explained by the very small number of positive interactions previously reported by Yu *et al.*, yielding a very high "false positive" rate for my matrix. However, many of these apparent "false positive" interactions may be true and novel interactions relative to the Yu *et al.*[43] dataset and could be experimentally confirmed by orthologous assays.

To avoid being penalized for capturing novel interactions, the data was also benchmarked against all previously reported Y2H interactions from the BioGRID. Here, both recall and MCC values were much lower than when benchmarked against the Yu *et al.* dataset. The low recall values mean that my dataset performed poorly in capturing "true positives" when compared to all previously reported Y2H interactions. This could be explained by many reasons: lower quality of all of the previously reported Y2H interactions, strain differences between the background strains used in all the studies, differences in type of plasmids used and differences in selective markers used. For example, Ito *et al.*[46] used 2-micron plasmids as their expression vectors, which are high copy number vectors, increasing the level of protein expression and therefore, also possibly the detection of unspecific interactions[20].

It is difficult to judge the quality of this dataset without a perfect dataset to benchmark against. However, pairwise retesting confirmation rate was 46%, which is on par with the 50% rate that was reported by Yu *et al.*[43], indicating that the quality of the data is comparable to that

of the high-quality dataset provided by Yu *et al.*[43]. Moreover, there is an enrichment of positively pairwise retested protein pairs amongst the higher scoring pairs.

4.4 Novel interactions by pairwise retesting

Of the 52 proteins pairs that were retested by pairwise retesting, 17 novel interactions were detected when benchmarked against the high-quality dataset provided by Yu *et al.*[43] Seven interactions were novel when benchmarked against all previously reported Y2H interactions (excluding Ito *et al.* "non-core", dataset containing interactions that tested only once). Five interactions were novel when manually checked in literature.

One of the novel interactions is between Crt10 and Mig3. Dubacq *et al.*[47] found that Mig3 is a repressor of genes involved in resistance to hydroxyurea, where overexpression of Mig3 sensitizes the mutant to hydroxyurea. Hydroxyurea is a DNA damaging reagent that directly inhibits the RNR complex by quenching the tyrosyl radical found in the small subunit[48,49]. The RNR complex is known to regulate the levels of dNTPs available to the cell[50]. Interestingly, Crt10 was discovered[51] to be a negative regulator of RNR genes. Deletion of *CRT10* increases expression levels of RNR genes in the presence and absence of DNA damage and hence, increases resistance to HU. The confirmed interaction between Crt10 and Mig3 suggest that Mig3 may be interacting with Crt10 to regulate the expression level of the RNR genes together. Although there has been no previous report of an observed interaction between Mig3 and Crt10, my hypothesis is that Mig3 and Crt10 interact even in the absence of DNA damage to maintain a balanced level of the RNR complex, and therefore dNTP levels, throughout the cell cycle. This hypothesis remains to be tested.

Two additional putative interacting proteins identified in our BFG-Y2H screen, Csm3 and Chl1 are both reportedly required for sister-chromatid cohesion[52–55], in that their deletions caused defective sister-chromatid cohesion[52]. Csm3 is part of the Tof1/Csm3/Mrc1 complex that interacts with the MCM helicase complex to activate the replication fork, during the replication fork's progression, and stabilizing the replication fork when it is stalled[56]. Chl1 is suspected of being a DNA helicase and it was shown to physically interact with Eco1[54], which is responsible for promoting sister-chromatid cohesion through acetylation[57] and acts close to the replication fork[58]. My hypothesis is that Chl1 could be relaying messages between the replication fork (through the Tof1/Csm3/Mrc1 complex) and the cohesins. When the replication fork is at the cohesin ring, Chl1 could be one of the many factors that help to signal the cohesin ring to go into an "open" formation that allows the replication fork to pass through. This formation change of cohesin rings was originally proposed by Lengronne *et al.*[58]. This hypothesis remains to be tested.

The interaction between Ypl108w and Mus81 was previously found by Ito *et al.*[46], in the "non-core" dataset (tested as a positive interaction only once, therefore, it was excluded by BioGRID). Ypl108w is a protein of unknown function, though its protein expression was found to be induced by methyl methanesulfonate (a DNA damage reagent that methylates DNA)[59]. Mus81 is a protein involved in multiple pathways of DNA repair, one of which is the homologous recombination repair that can be caused by methyl methanesulfonate through DNA methylation. Although only Ito *et al.*[46] previously reported that the two proteins interact, it is possible that they indeed interact with each other and belong in the same repair pathway. Interestingly, the other direction of the interaction between the two proteins (where Ypl108w is

the bait instead of the prey) also scored very high in our BFG-Y2H screen. However, this direction was not tested.

5 Conclusions and future directions

I was able to detect 17 novel Y2H interactions when benchmarked against the high-quality dataset from Yu *et al.*[43] from pairwise retesting of 52 DNA damage repair protein pairs that were arbitrarily selected of the top 200 hits of the primary BFG-Y2H screen. When compared to all previous literature, five novel interactions were detected. In parallel, 14 out of 17 PRS controls were detected in the same BFG-Y2H screen. The raw barcode counts amongst technical replicates were highly reproducible, and the pairwise retesting confirmation rate was very similar to previously reported by Yu *et al.*[43].

Although the generation of barcoded Y2H strains was resource-intensive and time-consuming, these barcodes and barcoded strains are reusable. The BFG-Y2H screen itself is highly scalable and much more efficient compared to the traditional Y2H method. Additional optimizations may be needed for bigger matrices. Due to the scalability of BFG-Y2H, it could be used to discover conditional protein-protein interactions.

More pairwise retesting should be done to uncover more novel Y2H interactions amongst the top scores in the primary BFG-Y2H screen. Additionally, for an unbiased high-quality dataset of the yeast proteome, the BFG-Y2H screen should be expanded to include all available yeast proteins.

6 References

1. Goffeau, A. *et al.* Life with 6000 genes. *Science* **274,** 546–567 (1996).

2. Lander, E. S. The new genomics: global views of biology. *Science* **274,** 536–539 (1996).

3. Tyers, M. & Mann, M. From genomics to proteomics. *Nature* **422,** 193–197 (2003).

4. Bergg\a ard, T., Linse, S. & James, P. Methods for the detection and analysis of protein–protein interactions. *Proteomics* **7,** 2833–2842 (2007).

5. Hunt, D. F., Yates, J. R., Shabanowitz, J., Winston, S. & Hauer, C. R. Protein sequencing by tandem mass spectrometry. *Proc. Natl. Acad. Sci.* **83,** 6233–6237 (1986).

6. Sharon, M. & Robinson, C. V. The role of mass spectrometry in structure elucidation of dynamic protein complexes. *Annu Rev Biochem* **76,** 167–193 (2007).

7. Van Aelst, L., Barr, M., Marcus, S., Polverino, A. & Wigler, M. Complex formation between RAS and RAF and other protein kinases. *Proc. Natl. Acad. Sci.* **90,** 6213–6217 (1993).

8. Vojtek, A. B., Hollenberg, S. M. & Cooper, J. A. Mammalian Ras interacts directly with the serine/threonine kinase Raf. *Cell* **74,** 205–214 (1993).

9. Amado, R. G. *et al.* Wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer. *J. Clin. Oncol.* **26,** 1626–1634 (2008).

10. Cohen, Y. *et al.* BRAF mutation in papillary thyroid carcinoma. *J. Natl. Cancer Inst.* **95,** 625–627 (2003).

11. Davies, H. *et al.* Mutations of the BRAF gene in human cancer. *Nature* **417,** 949–954 (2002).

12. Wan, P. T. C. *et al.* Mechanism of Activation of the RAF-ERK Signaling Pathway by Oncogenic Mutations of B-RAF. *Cell* **116,** 855–867 (2004).

13. Poulikakos, P. I., Zhang, C., Bollag, G., Shokat, K. M. & Rosen, N. RAF inhibitors transactivate RAF dimers and ERK signalling in cells with wild-type BRAF. *Nature* **464,** 427–430 (2010).

14. Hatzivassiliou, G. *et al.* RAF inhibitors prime wild-type RAF to activate the MAPK pathway and enhance growth. *Nature* **464,** 431–435 (2010).

15. Poulikakos, P. I. *et al.* RAF inhibitor resistance is mediated by dimerization of aberrantly spliced BRAF (V600E). *Nature* **480,** 387–390 (2011).

16. Escudier, B. *et al.* Sorafenib in advanced clear-cell renal-cell carcinoma. *N. Engl. J. Med.* **356,** 125–134 (2007).

17. Walhout, A. J. & Vidal, M. High-throughput yeast two-hybrid assays for large-scale protein interaction mapping. *Methods* **24,** 297–306 (2001).

18. Chien, C.-T., Bartel, P. L., Sternglanz, R. & Fields, S. The two-hybrid system: a method to identify and clone genes for proteins that interact with a protein of interest. *Proc. Natl. Acad. Sci.* **88,** 9578–9582 (1991).

19. Venkatesan, K. *et al.* An empirical framework for binary interactome mapping. *Nat. Methods* **6,** 83–90 (2009).

20. Braun, P. *et al.* An experimentally derived confidence score for binary protein-protein interactions. *Nat. Methods* **6,** 91–97 (2009).

21. Sauer, B. Functional expression of the cre-lox site-specific recombination system in the yeast Saccharomyces cerevisiae. *Mol. Cell. Biol.* **7,** 2087–2096 (1987).

22. Wood, R. D., Mitchell, M., Sgouros, J. & Lindahl, T. Human DNA repair genes. *Science* **291,** 1284–1289 (2001).

23. Jackson, S. P. & Bartek, J. The DNA-damage response in human biology and disease. *Nature* **461,** 1071–1078 (2009).

24. Gossen, M. *et al.* Transcriptional activation by tetracyclines in mammalian cells. *Science* **268,** 1766–1769 (1995).

25. James, P., Halladay, J. & Craig, E. A. Genomic libraries and a host strain designed for highly efficient two-hybrid selection in yeast. *Genetics* **144,** 1425–1436 (1996).

26. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6,** 343–345 (2009).

27. Cherry, J. M. *et al.* Saccharomyces Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.* **40,** D700–D705 (2012).

28. Alvaro, D., Lisby, M. & Rothstein, R. Genome-wide analysis of Rad52 foci reveals diverse mechanisms impacting recombination. *PLoS Genet.* **3,** e228 (2007).

29. Milanowska, K. *et al.* REPAIRtoire—a database of DNA repair pathways. *Nucleic Acids Res.* **39,** D788–D792 (2011).

30. Beaver, J. E. *et al.* FuncBase: a resource for quantitative gene function annotation. *Bioinformatics* **26,** 1806–1807 (2010).

31. Costanzo, M. *et al.* The genetic landscape of a cell. *science* **327,** 425–431 (2010).

32. Hartley, J. L., Temple, G. F. & Brasch, M. A. DNA cloning using in vitro site-specific recombination. *Genome Res.* **10,** 1788–1795 (2000).

33. Hu, Y. *et al.* Approaching a complete repository of sequence-verified protein-encoding clones for Saccharomyces cerevisiae. *Genome Res.* **17,** 536–543 (2007).

34. Warbrick, E., Lane, D. P., Glover, D. M. & Cox, L. S. A small peptide inhibitor of DNA replication defines the site of interaction between the cyclin-dependent kinase inhibitor p21< sup> WAF1</sup> and proliferating cell nuclear antigen. *Curr. Biol.* **5,** 275–282 (1995).

35. Bickle, M. B., Dusserre, E., Moncorgé, O., Bottin, H. & Colas, P. Selection and characterization of large collections of peptide aptamers through optimized yeast two-hybrid procedures. *Nat. Protoc.* **1,** 1066–1091 (2006).

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

37. Rual, J.-F. *et al.* Towards a proteome-scale map of the human protein–protein interaction network. *Nature* **437,** 1173–1178 (2005).

38. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13,** 2498–2504 (2003).

39. Brotherton, D. H. *et al.* Crystal structure of the complex of the cyclin D-dependent kinase Cdk6 bound to the cell-cycle inhibitor p19INK4d. *Nature* **395,** 244–250 (1998).

40. Lin, J., Jinno, S. & Okayama, H. Cdk6-cyclin D3 complex evades inhibition by inhibitor proteins and uniquely controls cell's proliferation competence. *Oncogene* **20,** (2001).

41. Zhang, Q., Wang, X. & Wolgemuth, D. J. Developmentally Regulated Expression of Cyclin D3 and Its Potential in Vivo Interacting Proteins during Murine Gametogenesis 1. *Endocrinology* **140,** 2790–2800 (1999).

42. Ma, H., Kunes, S., Schatz, P. J. & Botstein, D. Plasmid construction by homologous recombination in yeast. *Gene* **58,** 201–216 (1987).

43. Yu, H. *et al.* High-Quality Binary Protein Interaction Map of the Yeast Interactome Network. *Science* **322,** 104–110 (2008).

44. Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta BBA-Protein Struct.* **405,** 442–451 (1975).

45. Uetz, P. *et al.* A comprehensive analysis of protein–protein interactions in Saccharomyces cerevisiae. *Nature* **403,** 623–627 (2000).

46. Ito, T. *et al.* A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci.* **98,** 4569–4574 (2001).

47. Dubacq, C., Chevalier, A. & Mann, C. The Protein Kinase Snf1 Is Required for Tolerance to the Ribonucleotide Reductase Inhibitor Hydroxyurea. *Mol. Cell. Biol.* **24,** 2560–2572 (2004).

48. Lassmann, G., Thelander, L. & Gräslund, A. EPR stopped-flow studies of the reaction of the tyrosyl radical of protein R2 from ribonucleotide reductase with hydroxyurea. *Biochem. Biophys. Res. Commun.* **188,** 879–887 (1992).

49. Ko\cc, A., Wheeler, L. J., Mathews, C. K. & Merrill, G. F. Hydroxyurea arrests DNA replication by a mechanism that preserves basal dNTP pools. *J. Biol. Chem.* **279,** 223–230 (2004).

50. Jordan, A. & Reichard, P. Ribonucleotide reductases. *Annu. Rev. Biochem.* **67,** 71–98 (1998).

51. Fu, Y. & Xiao, W. Identification and characterization of CRT10 as a novel regulator of Saccharomyces cerevisiae ribonucleotide reductase genes. *Nucleic Acids Res.* **34,** 1876–1883 (2006).

52. Mayer, M. L. *et al.* Identification of protein complexes required for efficient sister chromatid cohesion. *Mol. Biol. Cell* **15,** 1736–1745 (2004).

53. Petronczki, M. *et al.* Sister-chromatid cohesion mediated by the alternative RF-CCtf18/Dcc1/Ctf8, the helicase Chl1 and the polymerase-α-associated protein Ctf4 is essential for chromatid disjunction during meiosis II. *J. Cell Sci.* **117,** 3547–3559 (2004).

54. Skibbens, R. V. Chl1p, a DNA helicase-like protein in budding yeast, functions in sister-chromatid cohesion. *Genetics* **166,** 33–42 (2004).

55. Xu, H., Boone, C. & Brown, G. W. Genetic dissection of parallel sister-chromatid cohesion pathways. *Genetics* **176,** 1417–1429 (2007).

56. Nedelcheva, M. N. Uncoupling of Unwinding from DNA Synthesis Implies Regulation of MCM Helicase by Tof1/Mrc1/Csm3 Checkpoint Complex. *J. Mol. Biol.* **347,** 509–521

57. Ben-Shahar, T. R. *et al.* Eco1-dependent cohesin acetylation during establishment of sister chromatid cohesion. *Science* **321,** 563–566 (2008).

58. Lengronne, A. *et al.* Establishment of Sister Chromatid Cohesion at the< i> S. cerevisiae</i> Replication Fork. *Mol. Cell* **23,** 787–799 (2006).

59. Lee, M.-W. *et al.* Global protein expression profiling of budding yeast in response to DNA damage. *Yeast* **24,** 145–154 (2007).

60. Gillette, W. K. *et al.* Pooled ORF Expression Technology (POET) Using Proteomics to Screen Pools of Open Reading Frames for Protein Expression. *Mol. Cell. Proteomics* **4,** 1647–1652 (2005).

61. Reboul, J. *et al.* C. elegans ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nat. Genet.* **34,** 35–41 (2003).

62. Weston, A., Humphreys, G. O., Brown, M. G. & Saunders, J. R. Simultaneous transformation of Escherichia coli by pairs of compatible and incompatible plasmid DNA molecules. *Mol. Gen. Genet. MGG* **172,** 113–118 (1979).

63. Goldsmith, M., Kiss, C., Bradbury, A. R. & Tawfik, D. S. Avoiding and controlling double transformation artifacts. *Protein Eng. Des. Sel.* **20,** 315–318 (2007).

64. Velappan, N., Sblattero, D., Chasteen, L., Pavlik, P. & Bradbury, A. R. Plasmid incompatibility: more compatible than previously thought? *Protein Eng. Des. Sel.* **20,** 309–313 (2007).

7 Appendices

7.1 Arrangement of barcode sequences on the plasmid

Sequence arrangement for DB barcodes

```
CTCACTAAAGGGAACAAAAGCTGGGTACCGATAACTTCGTATAATGTATGCTATACGAAGTTATCCATACGAGCACATT
                              [------------ loxP'* ------------][----- DB-U1 --

ACGGGNNNNNNNNNNNNNNNNNNNNNNNNCTAACTCGCATACCTCTGATAACATAACTTCGTATAGGATACTTTATAC
----][------ DB-UPTAG  ------][------- DB-U2 -------][----------- lox2272 -----

GAAGTTATTGTCAGCACTCTGTCAAAATAGATCGGAAATCGATAGGTGCGTGTGAAGGNNNNNNNNNNNNNNNNNNNN
-------][---------- LINKER ----------][----- DB-D1 ------][------ DN-UPTAG  ---

NNNNCCTCAGTCGCTCAGTCAAGCCCCCTCGAGATCCGGGATCGAAGAAATGA
---][----- DB-D2 -----]
```

*loxP' is reverse complement of loxP

Sequence arrangement for AD barcodes

```
GAAGGCTTTAATTTGCAAGCTCGGGATCCGCCCTTAGAACCGAGAGTGTGNNNNNNNNNNNNNNNNNNNNNNNNCACT
                              [----- AD-U1 ------][------ AD-UPTAG  ------][---

CCGTTCGTCACTCAATAATGTCAGCACTCTGTCAAAATAGATCGGAAAATAACTTCGTATAATGTATGCTATACGAAGT
--- AD-U2 -------][---------- LINKER ----------][------------ loxP'* ----------

TATCTCCAGGGTTAGGCAGATGNNNNNNNNNNNNNNNNNNNNNNNNCAATCGCACTATCCCGCTGATAACTTCGTATA
--][----- AD-D1 -----][------ AD-DNTAG  ------][----- AD-D2 -----][------------

GGATACTTTATACGAAGTTATCCCCCTCGAGATCCGGGATCGAAGAAATGA
lox2272 ------------]
```

*loxP' is reverse complement of loxP

Please see **Appendix 7.2** for schematics of fused barcodes.

7.2 Barcode fusion products after the induction of Cre

After induction of Cre, the corresponding *loxP* and *lox2272* swap positions with one another, creating fused barcodes. Below is a schematic of what fused barcodes look like.

UP-UP fusion after induction of Cre = AD-UPTAG fused with DB-UPTAG

```
GAAGGCTTTAATTTGCAAGCTCGGGATCCGCCCTTAGAACCGAGAGTGTGNNNNNNNNNNNNNNNNNNNNNNNNCACT
                              [----- AD-U1 ------][------ AD-UPTAG  ------][---

CCGTTCGTCACTCAATAATGTCAGCACTCTGTCAAAATAGATCGGAAAATAACTTCGTATAATGTATGCTATACGAAGT
--- AD-U2 -------][---------- LINKER ----------][------------ loxP'* ----------

TATCCATACGAGCACATTACGGGNNNNNNNNNNNNNNNNNNNNNNNNCTAACTCGCATACCTCTGATAACATAACTTC
--][----- DB-U1 ------][------ DB-UPTAG  ------][------- DB-U2 -------][------

GTATAGGATACTTTATACGAAGTTATCCCCCTCGAGATCCGGGATCGAAGAAATGA
- lox2272 ---------------]
```

DN-DN fusion after induction of Cre = AD-DNTAG fused with DB-DNTAG

```
CTCACTAAAGGGAACAAAAGCTGGGTACCGATAACTTCGTATAATGTATGCTATACGAAGTTATCTCCAGGGTTAGGCA
                              [------------ loxP'* ------------][----- AD-D1 --

GATGNNNNNNNNNNNNNNNNNNNNNNNNCAATCGCACTATCCCGCTGATAACTTCGTATAGGATACTTTATACGAAGT
---][------ AD-DNTAG  ------][----- AD-D1 -----][------------- lox2272 --------

TATTGTCAGCACTCTGTCAAAATAGATCGGAAATCGATAGGTGCGTGTGAAGGNNNNNNNNNNNNNNNNNNNNNNNC
--][---------- LINKER ----------][----- DB-D1 ------][------ DN-UPTAG  ------][

CTCAGTCGCTCAGTCAAGCCCCCTCGAGATCCGGGATCGAAGAAATGA

----- DB-D2 -----]
```

7.3 Using an *en masse* Gateway LR reaction to generate barcoded strains

The most rate-limiting step in the current design of BFG-Y2H is generating all the barcoded strains. I tried to address this issue by performing Gateway LR reactions *en masse.*

Here, a pool of barcoded destination vectors (courtesy of N. Yachie) and a pool of entry clones carrying yeast ORFs would be used in one Gateway LR reaction to generate barcoded expression clones. The identity of all expression clones would be determined by NGS of multiplexed Nextera libraries generated from row, column and plate sub-pools (Figure 29**)**. This strategy has been implemented before on a pool of 688 ORFs[60], but it is unclear how many ORFs were successfully cloned through Gateway cloning because they were immediate processed downstream processes without further quantification.

Due to concerns about whether the efficiency of Gateway LR reaction is length-dependent[61], one Gateway LR reaction was performed with a pool of 94 entry clones (each carrying an yeast ORF) and a pool of 750 barcoded destination vectors (courtesy of N. Yachie) as a proof-of-principle experiment. Positive (provided by Invitrogen) and negative controls for the Gateway LR reaction were used. I digested 1μg of the barcoded destination vector pool with 10 units of the enzyme *SmaI* (NEB, standard protocol) for one day at 30 °C. The next day, I inactivated *SmaI* at 65 °C for 20 minutes. I used the pool of destination vectors directly in a Gateway LR reaction (Clonase II, Invitrogen) with a pool of entry clones at an equal molarity of 150ng:150ng. The reaction was left at room temperature for one day. The next day, I added 2μg of proteinase K and incubated at 37 °C for an hour to stop the reaction (as suggested per Invitrogen protocol). 100ng of the resultant expression vector pool was transformed into 50μL of DH5α competent cells (NEB, high efficiency protocol was used). I spread the transformed cells onto large 245mm x 245mm square LB+ampicillin plates and incubated them at 37 °C for one

day. I counted ~2500 colonies on the large plate. I used the QPix robot (Genetix) to select 384 single colonies from one large square plate and arrayed them into cell culture plates containing 80uL of LB+ampicillin media. These colonies were pooled such that row, column and plate sub-pools of Nextera libraries could be used to identify each plasmid in terms of ORFs and barcodes identities (Figure 29**,** panel B). The other large square plate was scraped to determine the efficiency of the *en masse* Gateway LR reaction.

For a fair comparison of the efficiency of *en masse* Gateway LR reaction, both the yeast entry clones pool and the resulting pool of expression vectors were sequenced by Nextera/Illumina sequencing. I mapped the sequencing reads to the yeast reference genome, S288C (SacCer_Apr2011/sacCer3) by using a local alignment algorithm performed with Bowtie2[36]. I wrote a customized Perl script to perform coverage analysis on each gene (normalized by ORF length). Because the overall numbers of reads were very different between the two experiments, I adjusted the coverage (under-sampled) for the experiment with more coverage (the yeast entry clones pool). The results show that the coverage of yeast ORFs after Gateway LR compared to the coverage of yeast ORFs before Gateway LR is roughly even across ORFs of lengths 200-2900bp (Figure 30**)** suggesting the value of an *en masse* Gateway LR strategy.

Given the good coverage across all ORFs post Gateway LR, the row, column and plate sub-pools of Nextera libraries were also sequenced to identify each plasmid in terms of ORFs and barcodes identities from the 96-well format. However, the initial assumption that each colony only contains one plasmid (one ORF and one set of barcodes) appeared to be false. There were significantly more ORFs than intended (numbers varied from 20-30 per each pool). This made the identification of each plasmid impossible.

These experiments demonstrated that *en masse* Gateway LR could be used to generate pools of expression vectors. However, the original scheme of identifying the ORF and barcodes for each individual well did not work due to having multiple plasmids per well. This could be explained the relatively high colony density on the original plates or the possibility that cells took in more than one plasmid in the plasmid transformation step.

Although this set of experiments did not yield individually identified barcoded expression vectors, it remains to be a very promising approach for generating pools of barcoded expression vectors. Further optimizing could be done to ensure one plasmid per colony during the plasmid transformation step in bacteria, such as spreading fewer colonies on the selection plate to ensure one colony per well. In addition, the concentration of the plasmid during the plasmid transformation step could also be lowered to decrease the chance of a bacterial cell receiving multiple plasmids, since it was previously demonstrated that there is a dose-response relationship between the number of plasmids per transformant and the amount of plasmid DNA used[62,63], which could be maintained in the same cell for many days[64].

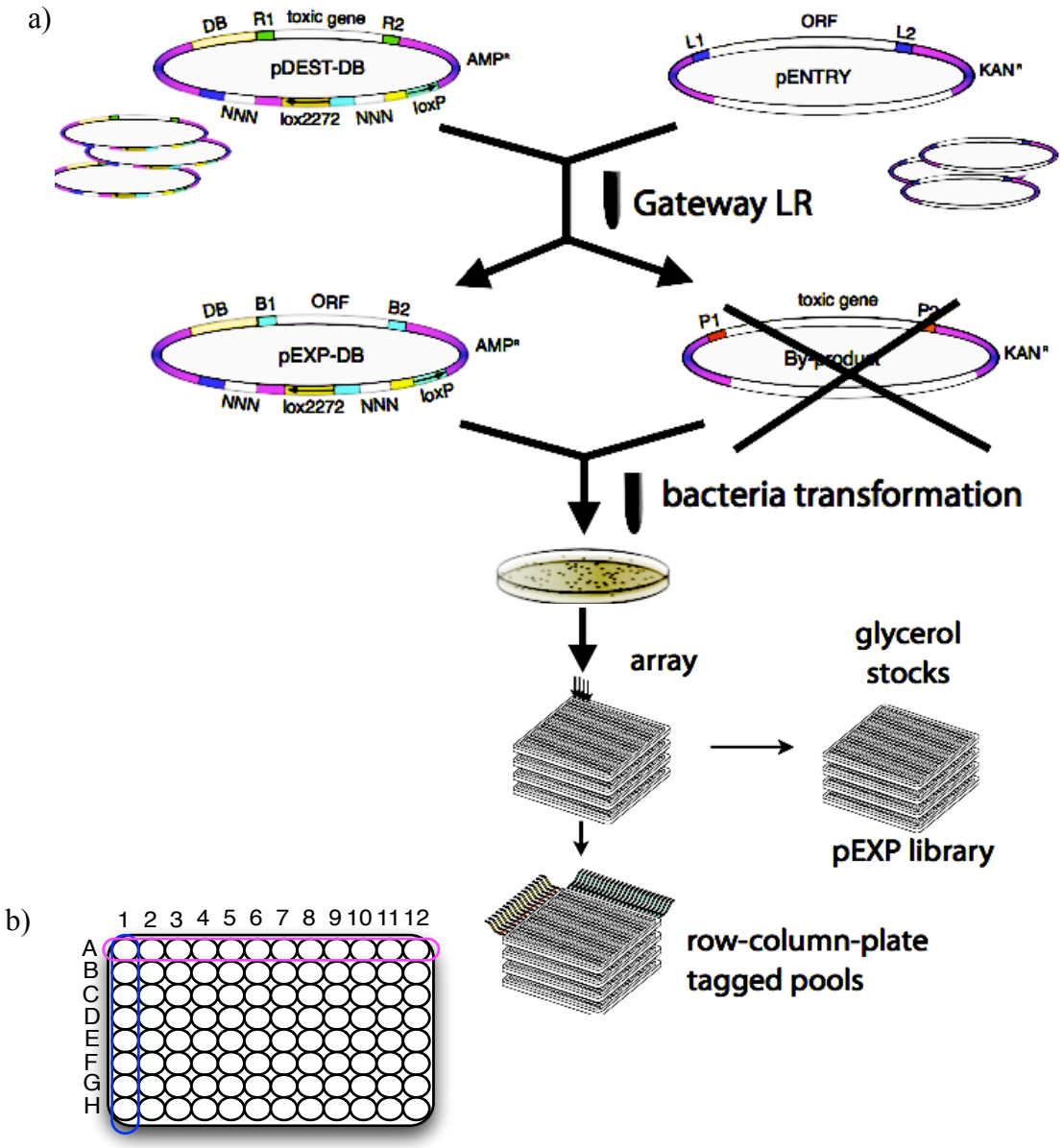


Figure 29. Overview of en masse Gateway LR. Here, a pool of barcoded destination vectors and a pool of entry vectors carrying yeast ORFs undergo one Gateway LR reaction together, generating a pool of barcoded expression vectors containing the yeast ORFs. This pool of expression vectors is then transformed into bacteria, and single colonies are picked and arrayed into plates. (b) After the plasmid DNA is extracted from each well, every column and row is pooled together. The identities of each plasmid would then be identified computationally (one common ORF shared between row A and column 1 corresponds to the original position of A1 before pooling).

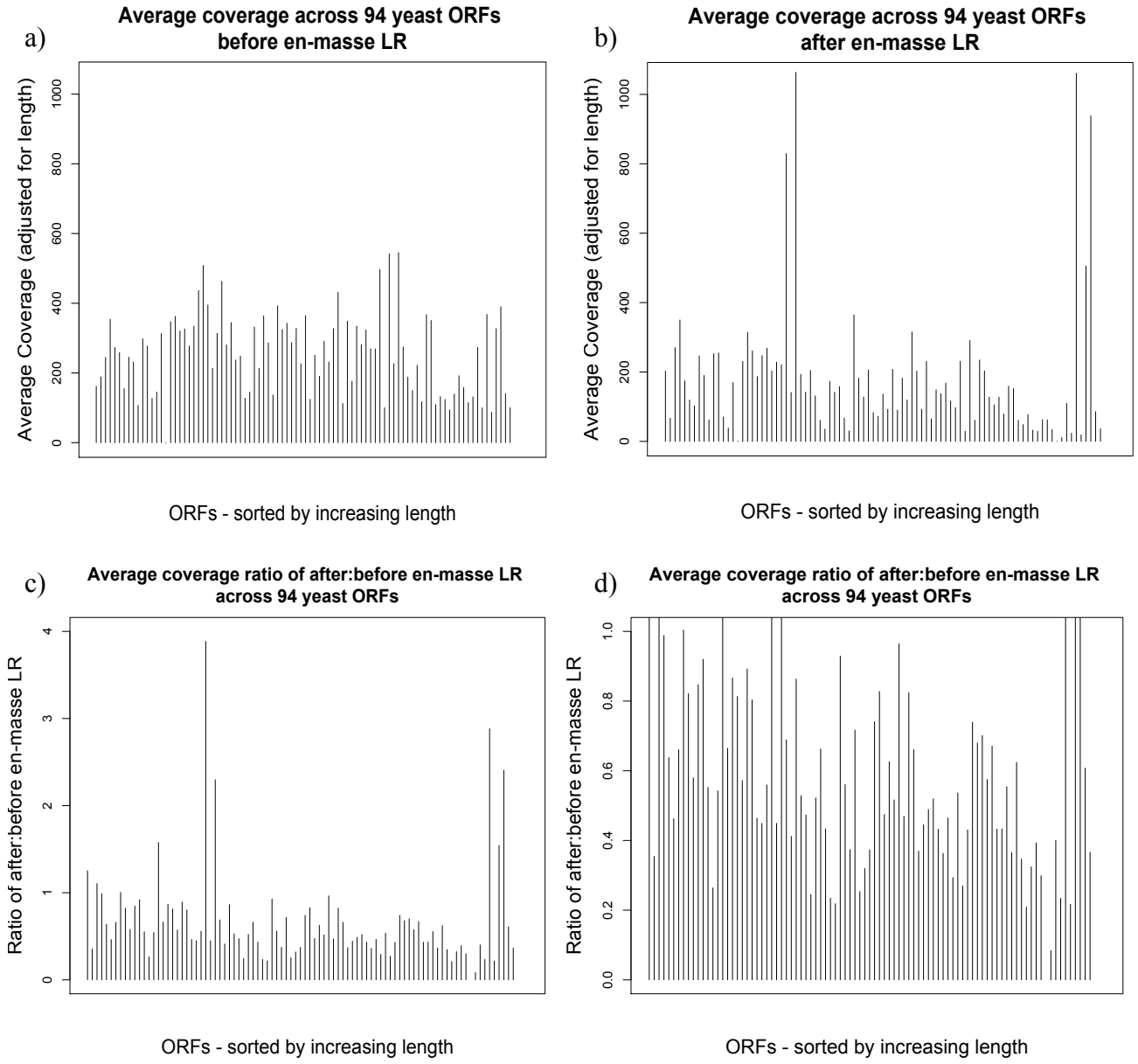


Figure 30. Results of en masse Gateway LR of a pooled barcoded destination vectors and a pool of entry vectors containing 94 different yeast ORFs of lengths between 200-2900bp. (a) Coverage of ORFs before Gateway LR. (b) Coverage of ORFs after Gateway LR (sample is adjusted such that it is at the same level of coverage as coverage before Gateway LR). (c) Coverage ratio of after:before en masse Gateway LR. (d) Same as panel (c), but scaled to 1 on the y-axis. All mapping was done with Bowtie2's 22 local alignment algorithm. Coverage was calculated with a customized Perl script and is defined as number of reads per base per ORF (adjusted for the length of ORF).

7.4 Pairwise Resting of PRS

The PRS was pairwise tested with clones generated from in-yeast-assembly[42] (IYA) (from Dr. N. Yachie, Roth lab) as well as clones that I generated from Gateway cloning.

Both sets of expression vectors have barcodes on them. The ORFs on the expression vectors used for the IYA PRS strains were verified by Sanger sequence in our lab. These expression vectors were then amplified with PCR and the resulting amplicons were used in a 4-piece IYA to generate barcoded expression vectors for the PRS. I cherry-picked the PRS entry clones from the HuORFeome v8.1 collection (previously sequenced by the Vidal lab). I then Gateway cloned them into barcoded expression vectors.

The results (Table 4) show that retesting of Gateway clones yielded slightly better results than the IYA clones. For the Gateway clones, 28 pairs of proteins retested positive out of a total of 42 pairs (66.6%) (one pair didn't grow in the +His condition after mating, probably due to a bad haploid clone). For the IYA clones, 19 (59.4%) pairs of proteins retested positive out of 32 tested pairs. If we only compare the two sets of clones for protein pairs that were tested by both methods, then we have 22 pairs of proteins retested positive out of 31 pairs of proteins for the Gateway clones (71.0%), and 18 out of 31 pairs of proteins for the IYA clones (58.0%). I did a Chi-square test with R and the p-value was 0.4259, suggesting that the two methods are not significantly different.

The differences could be explained by IYA being a more error-prone process than Gateway cloning, and therefore, a lower rate of PRS protein pairs testing as Y2H positives.

DB-X	AD-Y	Gateway	IYA
RYBP	*RING1*	1	0
SUMO1	*SP100*	1	0
VAMP3	*STX4*	1	0
S100A1	*S100B*	1	0
CDKN1A	*PCNA*	1	0
KLF3	*FHL3*	1	0
CCND2	*CDK4*	1	1
ARL2	*ARL2BP*	1	1
S100P	*S100A1*	1	1
CDK4	*CDKN2D*	1	1
MRFAP1	*MORF4L1*	1	1
PSMD5	*PSMC2*	1	1
BNIP3	*BNIP3L*	1	1
SH3GLB1	*SH3GLB2*	1	1
BCL2L2	*BIK*	1	1
LSM2	*LSM3*	1	1
MORF4L1	*MRFAP1L1*	1	1
IKZF5	*IKZF1*	1	1
LDLRAP1	*AP2B1*	1	1
CDK6	*CCND3*	1	1
TSG101	*CEP55*	1	1
TSG101	*CEP55*	1	1
UBE2K	*UBE2I*	0	0
RBBP4	*MTA1*	0	0
CTBP1	*RBBP8*	0	0
STX4	*VAMP2*	0	0
SARNP	*DDX39*	0	0
BRMS1	*NMI*	0	0
QKI	*RBPMS*	0	0
BTG1	*CNOT7*	0	1
PSMC1	*PSMD2*	0	1
EXOSC1	*EXOSC5*	No growth	1
UCHL5	*ADRM1*	1	Not tested
PIK3CB	*PIK3R1*	1	Not tested
UBE2I	*PIAS1*	1	Not tested
NAGK	*LNX1*	1	Not tested
SUPT4H1	*SUPT5H*	1	Not tested
LOC729991	*TEX11*	1	Not tested
LIN37	*TEX11*	0	Not tested
STMN2	*TEX11*	0	Not tested
DAXX	*SUMO1*	0	Not tested
RBM9	*QKI*	0	Not tested

Table 4. A comparison of pairwise retesting results for two sets of PRS generated from in-yeast-assembly and Gateway cloning. Gateway cloning and in-yeast-assembly (IYA) pairwise retesting results are shown from the –His condition. “No growth” means that that the protein pair did not grow in the +His condition. “Not tested” means that the protein pair was never tested with pairwise retesting.

7.5 Top 200 interactions in the BFG-Y2H (-His condition)

DB-X	AD-Y	Normalized score		DB-X	AD-Y	Normalized score
YPL108W	MUS81	4.308		ULP2	IRC8	2.033
HTA1	MGT1	3.752		RDH54	CSM2	2.025
HTA1	SLX8	3.752		DNL4	MND1	2.013
HTA1	YMR31	3.752		ECM11	NFI1	2.007
HTA1	MAG1	3.751		RAD2	MAD3	1.991
DDR2	MIG3	3.749		CCL1	TFB3	1.989
DDR2	MMS4	3.748		YLR118C	MIG3	1.981
HTA1	MAD1	3.647		DDI2	RAD1	1.974
HTA1	NPR3	3.647		UNG1	MUS81	1.969
HTA1	RAD10	3.647		RAD50	MRE11	1.955
HTA1	BUB2	3.647		RAD9	CHK1	1.95
MEC3	RAD17	3.612		RFA1	TFB1	1.939
RAD1	RAD10	3.533		SRS2	MRE11	1.893
SLX1	SLX4	3.523		RFA1	RAD23	1.889
RAD57	YHL018W	3.474		HSM3	NSE3	1.888
MEC3	TFB1	3.421		TFB1	RAD9	1.865
THI4	THI4	3.398		MRPS16	RAD61	1.864
MGT1	YAL044W-A	3.272		SRS2	THI4	1.84
MGT1	MEC3	3.272		LCD1	RFA1	1.837
NUP60	LIF1	3.257		ULP2	SGO1	1.823
YJR085C	SLX4	3.201		SRS2	CHL1	1.819
MUS81	YPL108W	3.192		IRC19	RAD1	1.819
MSH4	HPR1	3.164		RTT107	RAD10	1.813
YJR085C	RAD10	3.149		TFB2	TFB5	1.811
HTA2	PSY3	3.125		ULP2	RAD10	1.81
CSM2	PSY3	3.007		RAD27	THI4	1.806
DNL4	SMC6	2.968		DIN7	IRC6	1.802
TFB1	MEC3	2.913		ULP2	MUS81	1.802
ABF1	MUS81	2.891		HPR1	POL30	1.796
CHL1	CSM3	2.891		MSH2	DOA1	1.795
DDR2	RAD16	2.869		MSH2	LIF1	1.785
DDR2	HTA1	2.869		FMP41	NFI1	1.778
DDR2	MMS1	2.869		MSH2	MAD3	1.773
DDR2	YHR192W	2.868		IRC24	IRC24	1.754
DDR2	HNT3	2.867		COX16	TFB1	1.747
PSY3	CSM2	2.76		RTT109	MRE11	1.746
MLH1	UNG1	2.727		RAD2	RAD1	1.743
POL32	POL31	2.709		ECM11	WSS1	1.735
RMI1	TOP3	2.693		RAD3	RNR2	1.725
IRC15	MIG3	2.653		RAD55	SLX5	1.724
IRC15	SLX4	2.652		RAD50	PIN4	1.706
IRC15	MMS1	2.652		PSY4	MIG3	1.69
MGT1	YGL085W	2.649		YMR178W	SAE2	1.677

MGT1	RCO1	2.646		HNT3	RAD9	1.673
TOP3	RMI1	2.64		PPH3	MIG3	1.671
MSH5	MSH4	2.633		TFB2	YMR244C-A	1.669
NFI1	DPB3	2.613		YLR118C	MUS81	1.667
DNL4	LCD1	2.58		MEK1	OCA1	1.667
NTG2	RNR1	2.577		MMS1	MIG3	1.665
YHL018W	YHL018W	2.575		RFA1	RAD17	1.664
POL31	POL32	2.548		RAD27	YNL134C	1.663
XRS2	MRE11	2.539		THI4	XRS2	1.657
NSE5	KRE29	2.535		THI4	RAD17	1.654
ULP2	YLR271W	2.534		RFA1	RCO1	1.652
MGS1	MSH4	2.528		SLX8	SIZ1	1.649
DPB3	DPB4	2.524		MAD1	BDF1	1.648
MSH4	MSH2	2.495		RMI1	SAE2	1.646
HDA1	MPH1	2.458		THI4	RAD52	1.644
HDA1	YAR028W	2.456		MLH1	RAD52	1.644
RAD1	SLX4	2.407		IRC6	MIG3	1.64
RAD28	POL31	2.392		D2	IRC10	1.637
CHK1	RAD9	2.361		MLH1	PRI1	1.635
HTA1	MAD3	2.357		MLH1	TFB1	1.633
HTA1	MEC3	2.356		THI4	MLH2	1.632
HTA1	YMR244C-A	2.356		MAG1	RAD9	1.628
SMC5	RRM3	2.33		DIN7	RFC5	1.623
SMC5	APN2	2.329		PAC10	DDI1	1.621
SMC5	HIM1	2.328		RFA3	RAD2	1.621
YML131W	DDI1	2.315		MLH1	TFB5	1.621
AHC1	YPL108W	2.313		NUP133	GDH1	1.62
ADD37	RAD30	2.262		RAD10	RAD1	1.613
YHR192W	KRE29	2.256		MSH6	MMS4	1.608
TFB2	MIG3	2.245		MEC3	MRE11	1.606
CSM2	NFI1	2.24		IES6	PSY4	1.605
HDA1	MRE11	2.236		IRC15	RAD17	1.605
ULP2	MIG3	2.235		MLH1	MSH2	1.604
ULP2	ALK2	2.231		FMP41	YPL108W	1.603
HDA1	HRR25	2.231		IRC15	CHK1	1.603
ULP2	WSS1	2.23		RAD3	RAD10	1.6
ULP2	IRC15	2.228		RAD17	SLX5	1.597
MSH4	RAD30	2.224		UNG1	POL30	1.591
YJR096W	MRE11	2.215		IRC15	RIM9	1.585
SMC5	YPL108W	2.209		HHT1	MEC3	1.581
CRT10	MIG3	2.204		ACK1	WSS1	1.573
MSH2	MND1	2.203		D2	TFB2	1.572
MSH4	RPH1	2.2		MUS81	PIN4	1.567
MSH4	TFB5	2.196		NEJ1	LIF1	1.567
HTA1	MUS81	2.195		RNR1	RNR1	1.566
MSH4	DCC1	2.195		SMC5	PIN4	1.562
ABF1	POL32	2.185		SGO1	NFI1	1.562
MRE11	MRE11	2.167		THI4	YOR062C	1.559
THI4	ASF1	2.158		D2	YMR178W	1.558

IRC23	HST3	2.116		YMR178W	RAD34	1.553
DNL4	SIZ1	2.106		POL31	MAD1	1.548
YMR027W	RAD10	2.096		THI4	TFB1	1.546
NUP133	SML1	2.089		PRI2	KRE29	1.544
YLR271W	MEC3	2.073		MRPS16	YHR192W	1.542
DNL4	YLR118C	2.052		RAD27	RAD9	1.536
ULP2	IRC8	2.033		ABF1	GDH1	1.532
RDH54	CSM2	2.025		ALK2	YMR178W	1.53
DNL4	MND1	2.013		IRC16	TFB1	1.506
ECM11	NFI1	2.007		RAD23	MRE11	1.501