scientific reports

Check for updates

OPEN

Long-read sequencing to detect full-length protein–protein interactions

Stephanie Schaefer-Ramadan[1,4], Yue Guan[1], Ayeda A. Ahmed[2], Jovana Aleksic[1], Khadija A. Elmagarmid[3], Leena F. Syed[3], Yasmin A. Mohamoud[2] & Joel A. Malek[1,2]✉

Given the increased predictions on interactome size and demand for protein function information, methods for detecting protein-protein interactions remain a significant development area. The all-vs.-all sequencing (AVA-Seq) method utilizes a convergent fusion plasmid design to make two-hybrid technology amenable to next-generation sequencing. Here, we further innovate to take advantage of synthetic DNA technologies and Oxford Nanopore Technologies long-read sequencing improvements to allow us to determine full-length protein-protein interactions. We tested 3,115 human protein-protein pairs using this approach and recovered 159 protein-protein interactions from a set of 57 full-length human proteins. Fifteen of the 159 full-length protein-protein interactions matched known human interactions. When referencing a human gold standard set of interactions, eight full-length protein-protein interactions were recovered from an expected 28 interaction pairs (28.6%), a typical recovery rate for two-hybrid technologies. The AVA-Seq method, in combination with the ease of synthetic DNA production and the MinION platform, offers a low-cost, high-throughput alternative for determining protein-protein interactions, which can be utilized in research labs at all stages.

Keywords Oxford nanopore technologies, Long-read sequencing, Protein–protein interaction, Two-hybrid, Single-molecule sequencing

Methods for protein interaction mapping that utilize the cost and speed benefits from next-generation sequencing (NGS) are limited[1–10]. In the case of two-hybrid systems, this is primarily because the information about which 'bait' and 'prey' interact are often on separate plasmids that are not physically connected.

Both bacterial and yeast two-hybrids have similar underlying components which allow for robust readout of interactions. Figure 1 (top) represents a traditional two-hybrid experiment where the two proteins being tested (bait and prey). The genes for each protein are each on different plasmids, each harboring half of an essential transcription factor. When there is an interaction between the bait and prey, the transcription factor is reconstituted, and the system can recruit RNA polymerase which facilitates the readout of the HIS3 gene (other reporter genes can be used as well but not illustrated in this figure[11]). The selection conditions for this example include a competitive inhibitor of HIS3 called 3-AT. As the concentration of 3-AT increases the interaction strength must also increase in order for the interaction to overcome the inhibition (readout in yeast and bacterial two-hybrid is colony growth). Individual colonies, which represent a bait and prey plasmid that successfully recruited RNAp, are selected, amplified, DNA extracted, and Sanger sequenced.

High-throughput protein interactome mapping methods using the two-hybrid method have utilized full-length proteins across many species. A few examples include yeast two-hybrid studies for expanding the protein interactome of human[12–14] Arabidopsis[15] SARS-CoV2-human[16] Helicobacter pylori[17,18] Caenorhabditis elegans[19] and Saccharomyces cerevisiae[20,21].

In recent years, our lab has designed a convergent fusion plasmid (pAVA), allowing for high-throughput sequencing of interacting protein fragments in a bacterial system without the need for barcodes[8]. The idea behind all-vs.-all sequencing (AVA-Seq) development was to create a high throughput way to increase the number of bait and prey pairs being tested, utilize liquid growth (rather than colony selection) and next generation sequencing (rather than Sanger sequencing). Figure 1 (bottom) highlights how AVA-Seq was able to transform the essential information from the two-plasmid system and orient it in a converged manner on a single plasmid. This means the information of which 'bait' and 'prey' interacted is retained making the method amendable to

[1]Present address: Department of Genetic Medicine, Weill Cornell Medicine in Qatar, Doha, Qatar. [2]Genomics Laboratory, Weill Cornell Medicine in Qatar, Doha, Qatar. [3]Medical Education, Weill Cornell Medicine in Qatar, Doha, Qatar. [4]Department of Human Biology, University of Wisconsin-Green Bay, Green Bay, USA. ✉email: jom2042@qatar-med.cornell.edu

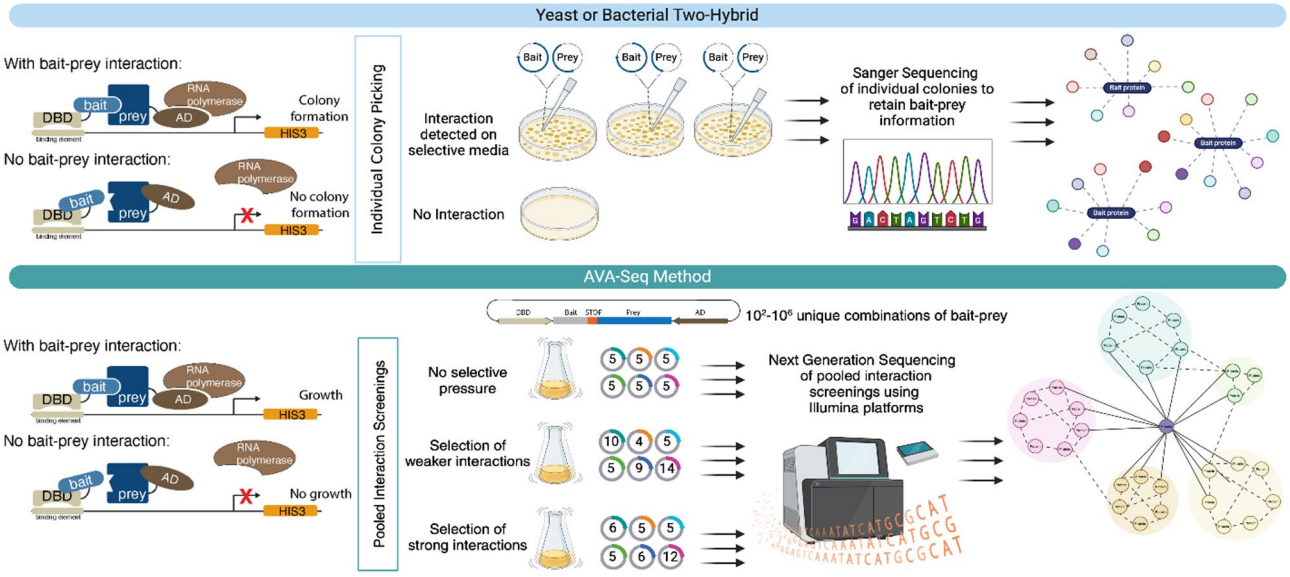

Fig. 1. Comparison of a two-hybrid method for determining protein-protein interactions and the all-vs.-all sequencing (AVA-Seq) method. Top panel: represents a traditional two-hybrid experiment where the two proteins being tested are each on different plasmids (bait and prey). Each plasmid contains half of an essential transcription factor that reconstitutes when they are within proximity of each other. When there is an interaction between the bait and prey, the transcription factor is reconstituted, and the system can recruit RNA polymerase which facilitates the readout of the HIS3 gene. The selection conditions for this example include a competitive inhibitor of HIS3 called 3-AT. As the concentration of 3-AT increases the interaction strength must also increase in order for the interaction to overcome the inhibition. The readout in yeast and bacterial two-hybrid is colony growth. Each individual colonies harbors information about which bait and prey plasmid successfully recruited RNAp. Each colony is selected, amplified, plasmid DNA is extracted, and Sanger sequenced. Bottom panel: represents the all-vs.-all sequencing (AVA-Seq) method where the readout for an interaction is growth in liquid culture. This growth (and change in growth) can be later quantified (log2FC after sequencing). Because the information for which bait and prey are on a single plasmid DNA can be extracted in bulk and next generation sequencing can be utilized.

liquid culture selection, DNA extraction in bulk and next generation sequencing. This transforms individual colony counting from traditional two-plasmid system into a quantifiable readout based on read counts. Until recently, we exclusively used gene fragments in the range of 360–450 bp (120 to 150 amino acids) for a combined size of ~800–900 bp, the amplicon size limit for most short-read technologies. While the information derived from protein fragments of this size can provide information on the broader protein interaction, it would be beneficial to develop the all-vs.-all sequencing (AVA-Seq) system to allow full-length protein testing.

To the best of our knowledge, no other protein interaction method enables protein-protein interactions to be determined using a single plasmid. This unique feature of AVA-Seq represents a significant technical advantage. By incorporating both proteins (whether full-length genes or gene fragments) into a single plasmid, our method substantially improves transformation efficiency compared to approaches requiring multiple plasmids to be co-transformed into the same cell. Furthermore, AVA-Seq leverages bacteria as a host, which offers greater transformation efficiency and scalability compared to yeast-based systems commonly used in other methods.

The incorporation of full-length genes into the pAVA plasmid is an important and valuable advancement to the AVA-Seq platform. Unlike earlier studies that utilized gene fragments, full-length genes provide a more comprehensive and physiologically relevant representation of protein-protein interactions. This improvement aligns better with the intended use of human positive reference sets (PRS libraries), enabling more direct and accurate comparisons of recovered interactions. Moreover, this is the first study to integrate the Oxford Nanopore Sequencing platform into a protein-protein interaction workflow, adding a novel dimension to AVA-Seq's capabilities and expanding the technological toolkit available for interactome studies.

The decision to incorporate full-length proteins into the AVA-Seq workflow is rooted in the need to recover biologically relevant interactions under experimental selection conditions. By including full-length genes rather than fragments, this study enhances the accuracy of interaction studies and ensures that the resulting data more closely reflect native interactions as they occur in living organisms. To explore the feasibility of this approach, we leveraged synthetic biology technologies, which offer numerous advantages for full-length protein studies: 1) a customizable design to allow for the selective inclusion or exclusion of restriction sites, enabling greater flexibility in downstream workflows, 2) codon optimization for E. coli expression ensures the availability of the appropriate tRNAs, improving translation efficiency and protein yield and 3) inclusion of 5'- and 3'-adapters for directional cloning into the pAVA plasmid, streamlining the workflow. At the time of this study, the synthesis of full-length

genes up to 1.8 kbp was available with short turnaround times, making it an affordable and efficient option for small-scale projects. For research labs without access to plasmids containing the desired genes, synthetic biology offers a practical solution, enabling the rapid and scalable production of custom sequences. Furthermore, with advances in synthetic biology, gene synthesis capabilities have expanded to support sequences up to 5 kbp, significantly increasing the scope of full-length protein-protein interaction studies.

The long-read sequencing technology by Oxford Nanopore Technologies (ONT) allows for novel genome assembly[22] improving existing genomes[23,24] epigenetic studies[25] and transcriptome splicing analysis, among other applications that benefit from longer sequences. The ONT platform has also been utilized for interactions between RNA binding proteins with RNA[26] protein-DNA interactions[27–29] host-virus interactions[30] host-pathogen[31] (plant gene-for-gene network), and enhancer-promoter interactions[32]. To our knowledge, single-molecule, long-read sequencing has not yet been applied to studying protein-protein interactions.

Bacterial two-hybrids are valuable tools in the protein interaction field and are robust in determining known and novel protein-protein interactions (PPIs) across many species. Recently, full-length SARS-Covid ORFs were utilized in a bacterial two-hybrid system to determine interactions with the human ACE2 protein[33]. Their application demonstrates that the bacterial system can handle full-length protein interaction studies and could be adapted to third-generation sequencing technologies such as Oxford Nanopore.

In this work, we utilized 57 human proteins (Choi et al., 2019), which could be synthesized on a single DNA fragment (less than 1.8 kbp). The goal was to adapt the AVA-Seq method to accommodate full-length proteins, allowing a more direct comparison to traditional protein interaction methods while simultaneously using third-generation sequencing technologies. This study demonstrates the combination of AVA-Seq convergent fusion design with the long-read capabilities of ONT to provide a rapid, low-cost protein interactome for research groups of any size.

Results

Individually tested full-length pairs in pAVA (low-throughput)

We first wanted to establish if the convergent fusion design of the pAVA plasmid is usable for screening full-length protein-protein interactions, as it was initially optimized for shorter fragments. To do this, a selected number of known human interacting pairs and non-interacting pairs[34] were convergently fused and inserted into pAVA in both orientations (fused to DNA binding domain (DBD) or RNAp activation domain (AD)) individually and confirmed by Sanger sequencing. When we began this project, the current size limit for synthesized genes is 1.8 kbp though this length will likely change in the future.

We performed two initial growth assays with protein pairs in isolation. The first was on solid agar (Supplemental Table 1; scratch test; colony formation as a readout), and the second was in liquid media (Supplemental Table 1; AVA one by one; OD_{600} as a readout) in the absence or presence of 3-AT, a competitive inhibitor of the HIS3 reporter gene. Protein pairs which were tested and grew with no selective pressure are indicated with an "x" on the table (Supplemental Table 1; scratch test; column C). Pairs which did not show colony formation on the solid agar in the presence of 3-AT were marked with "-" to represent non-interacting (Supplemental Table 1; scratch test; column D and E). Pairs which had colony formation in the presence of 2 or 5 mM 3-AT are marked with an "x" and the bold font indicates a recovered interaction (Supplemental Table 1; scratch test; column D and E). We then moved the individual pairs to liquid culture by positive reference and random reference pairs with the hope to get an estimate of baseline growth (Supplemental Table 1; AVA one-by-one; column G, H and I). All the protein pairs we were able to successfully ligate and test individually in liquid culture had some level of growth in the liquid culture assay in all conditions (media supplemented with 3-AT or media lacking 3-AT). This is unsurprising as even in our first manuscript, the negative controls and empty vector gave some level of baseline growth[8]. When looking at the data, OD_{600} growth > 0.6 in 2 mM conditions generally indicates an interaction when using the non-interacting pairs or proteins paired with frameshifted fragments as a reference (Supplemental Table 1; AVA one by one). This thresholding allows us to determine a reasonable estimate of a binary protein-protein interaction.

As we incorporated full-length protein pairs into the pAVA plasmid, we wanted to determine whether the orientation of the genes played a significant role in the growth and resulting observed interaction. There is an equal chance for a given protein-protein pair to show growth when adjacent to either DBD or AD. Several known human interacting pairs were tested manually (in the low-throughput liquid growth assay; Supplemental Table 1; AVA one by one; columns G, H and I) and showed growth ($OD_{600} > 0.6$) in at least one orientation, as shown in Fig. 2. We observed a preference for some pairs in one orientation versus the other while other pairs grew equally well fused to DBD or AD. Additionally, all the known interacting pairs tested showed higher average growth in the mild 2 mM 3-AT growth conditions compared to the strong competitive inhibition of 5 mM 3-AT except for PSMD4|RAD23A (one orientation). We conclude from these low throughput growth experiments that known human protein interactions can occur in using full length genes pairs ligated into the pAVA plasmid.

Low-throughput assessment of growth time

Due to the large span of protein pair lengths being expressed on a single plasmid, we wanted to ensure the growth time was sufficient to achieve separation between the 2 mM and 5 mM conditions in all possible protein lengths. To test this, we compared 9- and 16-hour growth in 0 mM, 2 mM, and 5 mM 3-AT conditions for pairs representing different protein-pair lengths. We observed no apparent bias regarding protein length influencing how well the pair grew in liquid media. Our results indicate that a 9-hour growth time is likely sufficient and might be preferred as it gives more separation in the 2 and 5 mM samples (Supplemental Fig. 1). Furthermore, the 9-hour growth is consistent with our previous AVA-Seq studies[8,35–37].

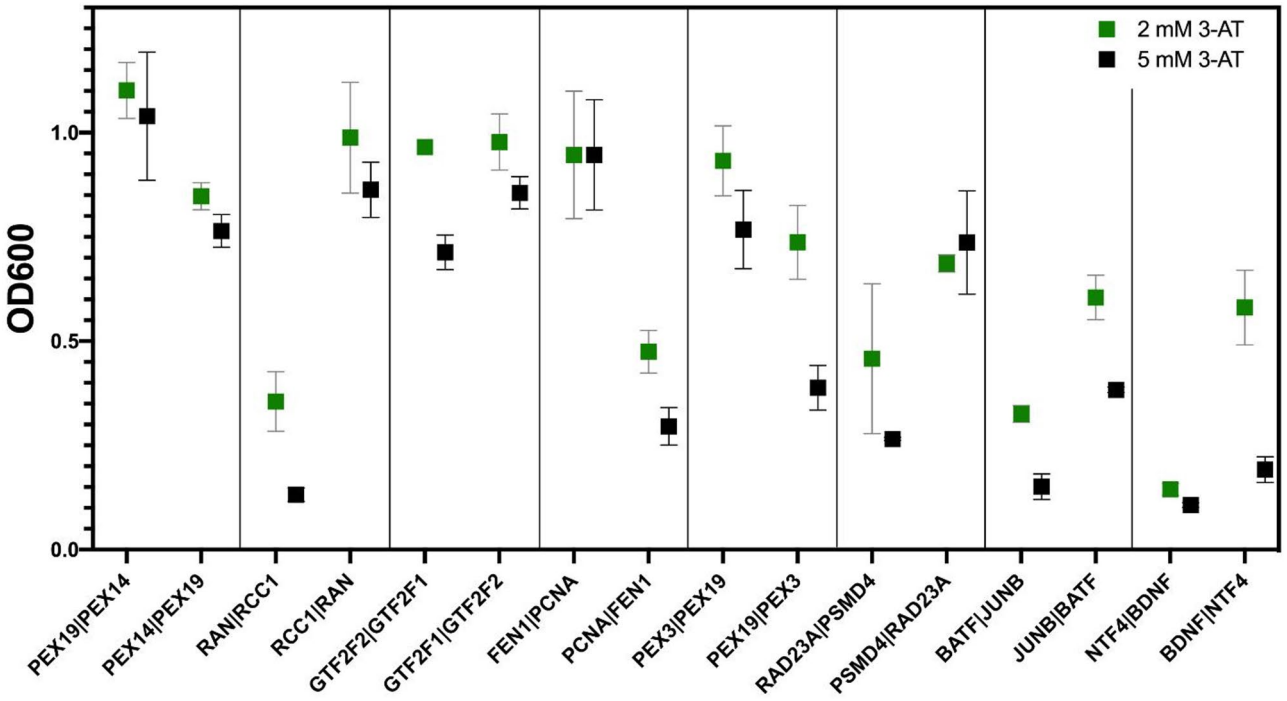

Fig. 2. Growth comparison of full-length known human interacting pairs in both orientations. Data normalized to 0 mM 3-AT 9-hours growth time. The 2 mM growth is shown in green and 5 mM growth is shown in black with error bars (SD). The gene listed first is the DBD fusion.

High-throughput AVA-Seq project design

Protein interaction screening of this dynamic range in length, where both genes are on a single plasmid, was never attempted previously; therefore, it was difficult to benchmark. Figure 3 outlines our method design. Synthesized full-length genes were separated into two sizes (< 1,299 base pairs and > 1,300 base pairs) to assist with the size selection of the NGS library. This was done for two main reasons. First, this study used a diverse range of gene lengths making pooled PCR amplification challenging to ensure a balance protein-protein pairing, and second, the first-generation ONT chemistries preferentially sequence shorter products. Meaning, a single pooled experimental design (as we initially tried; all full-length genes from this study vs. themselves) resulted in the shortest genes vs. shortest genes sequestering a majority of the sequencing pores. To overcome this the genes were split into short and longer gene pools. These two pools of genes were paired against themselves and against the other pool with the result of 3 unique experiments and all genes being paired with all others from the study. As the ONT chemistries improve, testing a true all-vs.-all on this dynamic range of gene products would be worth testing.

Coverage of interaction space (high-throughput)

AVA-Seq can simultaneously screen protein-protein pairs in two orientations, which is advantageous as some fusion pairs can show preferential interaction[34]. We obtained over 100x coverage for the gene combinations being tested. These numbers are based on the number of transformants observed as colonies on counting plates used for 3-AT screening events. When considering each orientation as a unique test pair, 3,115 protein-protein pairs were covered out of the possible test space of 3,192 (97.6%). Our requirement for 'tested' is the protein pair be in two or more experiments in at least one orientation (3,090 protein-protein pairs). Furthermore, if the first orientation is considered tested and the second orientation is only found in one experiment, it is considered tested in both orientations (25 protein-protein pairs). Figure 4 shows the tested combinations as a heat map.

If orientation is not considered (simply looking at the protein pair being covered in any orientation), 1,575 protein pairs are covered well enough to be considered tested out of a possible 1,596 protein pairs (98.7%) [total test space = (57 proteins x 56 proteins)/2; (self-pairs not considered; see below)]. Only 21 protein pairs were either never seen during sequencing or were seen in only 1 experiment (which does not meet our criteria for being 'tested') and are represented by a white box in both orientations (Fig. 4). Further method optimization such as adding an additional experiment (i.e., four experimental replicates instead of three), will likely improve the number of tested combinations.

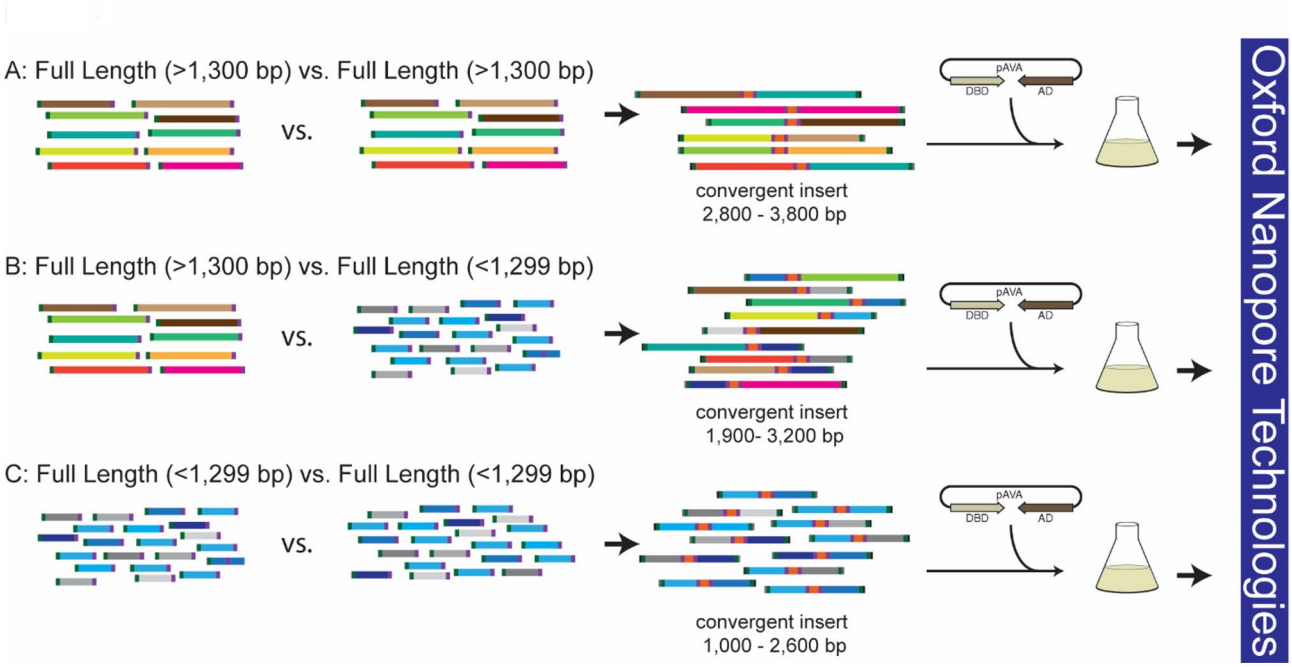

Fig. 3. AVA-Seq method experimental design using full-length proteins. (**A**) Full-length proteins that are > 1300 bp paired and screened against themselves. (**B**) Full-length proteins from human proteins > 1300 bp paired with full-length proteins < 1299 bp. (**C**) Full-length proteins < 1299 bp paired and screened against themselves. All convergent insert sizes reflect the final DNA length (including adapters) amplified for NGS. Individual protein sequences were synthesized by TWIST Biosciences and included custom adapters added to the 5' and 3' ends of the design process. All pathways utilize two sets of primers. Each set has a complementary stop insert, allowing the two fragments to be 'stitched' together (represented as orange on the convergent insert).

Identification of auto activators and 'sticky' proteins

All experiments have internal positive and negative controls to ensure that selection conditions have worked, and sequencing was able to detect the interactions in each replicate. However, it is important to determine if any protein(s) activate the system on their own (auto-activators). When a protein interacts with one portion of the system (DNA binding domain or activation domain) we would consider it to auto-activate the system. This would only happen when it is cloned in a specific orientation allowing for that interaction to occur. Potentially the interactions detected in the other orientation would be usable. On the other hand - a protein that interacts with both parts of the system is likely "sticky" in that it isn't interacting specifically with the system components but more likely with any protein it encounters. To test for auto activators, we added frameshifted fragments that encode very short random peptides. If we observe growth levels passing criteria used to define an interaction in a single orientation (fused to DBD or AD) we would consider that protein in that orientation to be an auto activator. With these criteria, we did not identify any protein that triggered the autoactivation of the system in a single orientation.

It is readily apparent from the interaction map (Fig. 4) that several proteins have many interaction partners. Proteins such as LMNA, LMNB1, GTF2F1, IFIT1, and SLC22A15 interact with frameshifted fragments in both orientations in multiple experiments. To be cautious, our analysis annotates these proteins as 'sticky' (Supplemental Table 2). Using this approach, distinguishing between a 'sticky' protein and an auto-activator is not always possible. Furthermore, eight real interactions (as determined by APID) were annotated as 'sticky' using our criteria above with StringDB values ranging from no data ($n = 2$) to 0.999 (indicating a 99.9% chance of interaction). There are also two additional interactions with StringDB values > 0.6 (but not reported in APID) annotated as 'sticky'. Two interactions are identified as interacting protein pairs when referencing HuRI database[12] and overlap with APID (Supplemental Table 2). It, therefore, becomes a challenge to distinguish some protein-protein interactions as a 'sticky' interaction from a real interaction when there is clearly some overlap. Thus, novel interactions annotated as 'sticky' should be viewed cautiously and warrant secondary experimental validation.

Known and novel interactions recovered

In our strictest requirements, an interaction must be recovered in two or more experiments in a single fusion orientation, or in a single experiment but in both fusion orientations which are considered the equivalent of two independent experiments. With these criteria, we recovered 159 full-length protein-protein interactions. Fifteen of these interactions are known human interacting pairs and are reported in the APID[38,39].

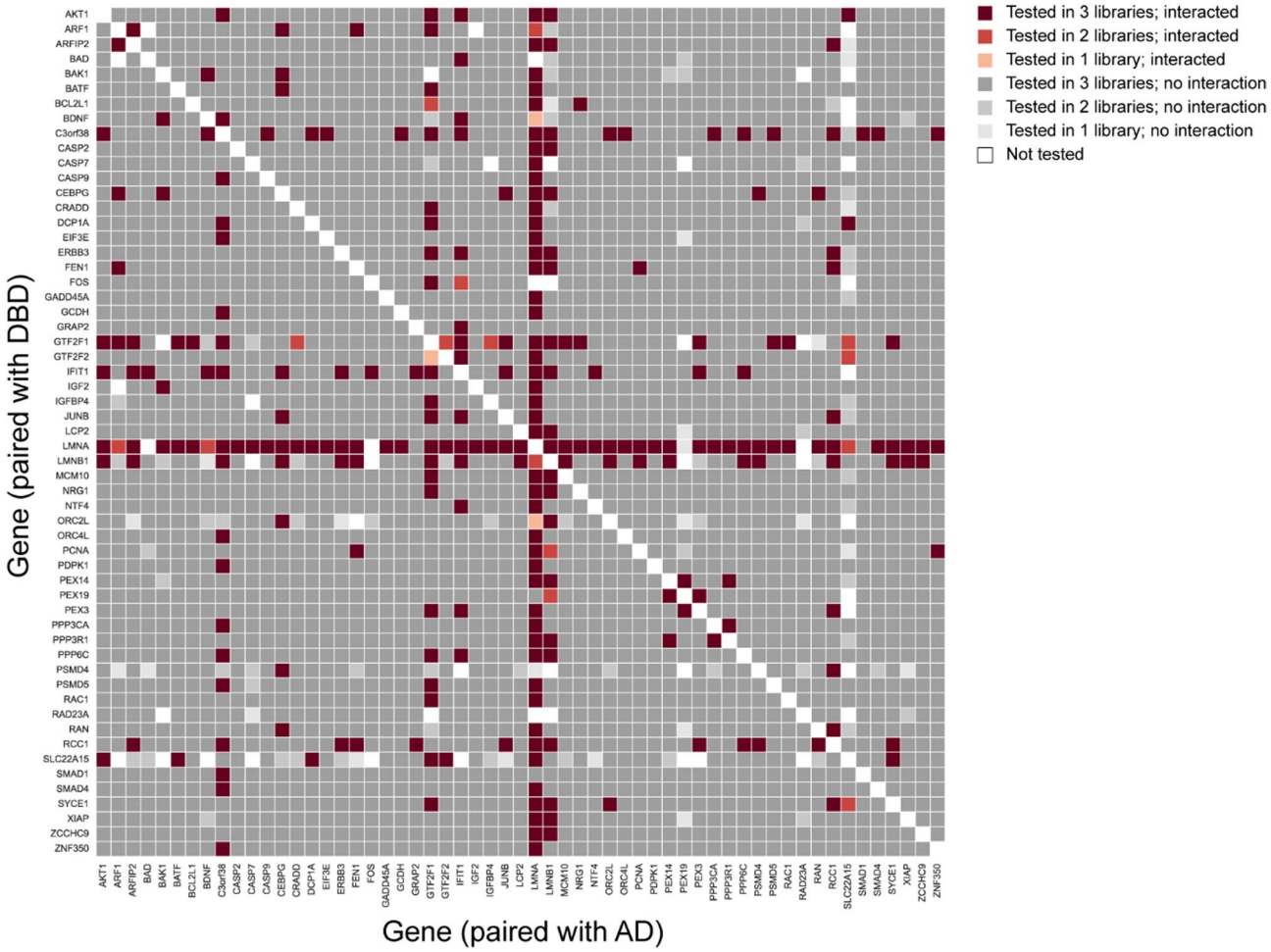

Fig. 4. Heatmap of coverage and recovered interactions. The gene pairs tested but did not interact in 3 experiments are colored dark grey and 2 experiments medium grey. The gene pairs tested and determined to interact are shaded red with a darker color indicating higher number of tested experiments. Protein pairs shown in white were not abundant enough to be considered 'tested'.

When referencing a human gold standard set of interactions, eight full-length protein-protein interactions were recovered from an expected 28 interaction pairs (28.6%, 15.3-47.1%, 95% confidence interval using sample proportions test), a typical recovery rate for two-hybrid technologies[34] and an indication of what future experiment might expect to recover.

It is important to note that the 28 tested PRS pairs from this study come from a subset of 60 PRS interactions. Similarly, the 5 tested RRS pairs came from a subset of 78 RRS interactions. To maximize the budget for this project certain proteins were selected because they participate in more than one interaction. It is possible in our selection of proteins we introduced unintentional bias. We noticed that the proteins used here have a higher average number of interactions in APID than the typical human protein that may account for the large number of interactions recovered (see below).

The remaining 144 interactions are not documented in other interaction databases and are potentially novel – several with support from StringDB[40]. Furthermore, 121 of 144 potentially novel interactions are recovered in both orientations and 2 or more experiments, indicating a robust, repeatable interaction using our method. Of the 15 APID-reported interactions we recovered, 13 were recovered in both fusion orientations and 2 or more experiments.

Five of the eight known human interacting pairs recovered were in both orientations, and all three experiments had $\log_2$FC (maximum fold change in $\log_2$) values ranging from 1.5 to 6.2. The PPP3CA|PPP3R1 interaction was one of three known interactions recovered in 5 mM 3-AT, the most stringent growth conditions. Interestingly, other methods did not recover this interaction despite the interaction domains being well documented and present[34]. Further investigation is needed as to why the AVA-Seq method recovered this interaction while no other robust method could.

We were unable to identify full-length self-pairs confidently. This is likely due to hairpins forming during PCR amplification or NGS library creation. For this reason, we have removed all tested and interaction data for self-pairs to prevent any misinterpretation or error.

Correlation between high-throughput AVA-Seq and low-throughput AVA

A key question is whether the AVA-Seq results can correlate the log_2FC of a protein pair in read counts within a complex mixture of tested interactions to the growth of a single interaction pair, as measured by OD_{600}. With the data in Fig. 2 (a low-throughput measure of liquid growth via OD_{600}), the protein pairs can be detected as interacting when tested in isolation, possibly due to reduced competition or crowding of neighboring cells. The experiments where NGS libraries were generated were a mixed population of 1,000's of protein-protein pairs being tested. This means in each 5 mL culture tube, there could be 3,192 DNA fusions represented in millions of cells (100x coverage of possible combinations). In the presence of 3-AT, 'stronger' interactions can grow and become a higher percentage of the final DNA product – represented as more reads. Figure 5 compares the growth of individual pairs (OD_{600}) to the number of read counts (maximum log_2FC). The higher the individual growth (AVA one by one), as measured by OD_{600}, the more likely the interaction is detected in a complex mixture (full length AVA-Seq) (also see Supplemental Table 1).

Known human interactions such as PEX14|PEX19, PEX3|PEX19, RAN|RCC1 and FEN1|PCNA show strong growth in 5 mM 3-AT conditions using the low-throughput growth experiment (Supplemental Table 1; Figs. 2 and 5). However, although these pairs are present in the full-length AVA-Seq experiments which were quantified using sequencing reads (Supplemental Table 1; full length AVA-Seq), they do not show a significant increase in read counts in 5 mM 3-AT conditions indicating they do not interact in these experimental conditions. Using the low-throughput method (Supplemental Table 1; AVA one-by-one), the RAN|RCC1 interaction showed higher growth in one orientation. Still, both DBD and AD fusions had similar log_2FC in the high-throughput AVA-Seq experiment (Supplemental Table 1; Fig. 5). Competition may happen when the protein pairs are pooled together, leaving some PPIs at a disadvantage. This means we see specific pairs that grow in the low-throughput method and on solid agar (Supplemental Table 1; columns C-E and G-I) but not in the high-throughput AVA-Seq (Supplemental Table 1; columns K-M). The reason(s) remains unclear. It may be 'simply' one of the features of AVA-Seq (or pooled interaction methods), and it further reiterates that multiple methods can give different results with minimal overlap. Overall, our data suggests using sequencing read counts to quantify PPIs is appropriate, as shown previously in the short-read versions of the AVA-Seq system[8,35,37].

Discussion

The proteins in this study are a subset of the Human Positive Reference Set (hsPRS-V2), a well-established collection of reproducible protein interactions validated across various model systems. This manuscript aims to showcase the adaptation of the AVA-Seq method for full-length protein interaction studies and its innovative application of Oxford Nanopore long-read sequencing. The higher-than-expected recovery rate observed likely stems from the inherent ability of these proteins to form strong, reproducible interactions in two-hybrid and similar systems. Furthermore, the use of read counts in AVA-Seq enhances sensitivity, allowing the detection of interactions that might be missed in traditional colony-based approaches.

Experiments were repeated three times, each with three replicates across three conditions (no, mild, and strong selection). By requiring the interaction to appear in at least two independent experiments, we ensure it is repeatable in our system. While our RRS interactions may be reproducible, most would say the Log2FC values suggest interactions which are not biologically significant. Relying solely on interaction strength can exclude transient or weaker interactions, which we believe are also important. We argue that a combination of repeatability, functional analysis, and in-vivo validation is needed to identify biologically relevant interactions, not just strength. Furthermore, individuals may apply more stringent cutoffs (e.g., those yielding 0 RRS) for more conservative interpretations.

The convergent fusion design of the pAVA plasmid in combination with NGS is central to our ability to obtain quantifiable results with AVA-Seq. Others have suggested a challenge in comparing one interaction pair to another using a traditional two-plasmid B2H platform due to different expression levels of the plasmids (i.e., low copy pBT and high copy pTRG). Yet another advantage of AVA-Seq over traditional two-hybrid systems is the DNA from both genes are contained in a single DNA fragment. This means that the copies of each gene in the tested pair are equal. Therefore, there is no need to manipulate the gene transcription levels or adjust the IPTG concentrations using our AVA-Seq method[41].

There are many advantages to utilizing the Oxford Nanopore platform for studying protein-protein interactions. A top benefit is the portability of the instrumentation and low cost, allowing many more laboratories to utilize the technology. Using long-read sequencing technologies will enable researchers to screen full-length protein pairs at scale without fragmentation. Individual constructs are not required; instead, an entire pool of proteins can be created and screened all-vs.-all. Additionally, utilizing full-length genes are more likely to result in biologically relevant interactions as they consider the 3D structure, multiple binding sites, etc. With the increased availability and reduced cost of synthetic DNA, we see this as a viable option for multiple sizes of projects, significantly reducing the need to first PCR and clone genes prior to protein interaction mapping. Here, utilizing full-length synthesized genes makes the entire protein pair matrix readily attainable - allowing us to increase the search space and interaction coverage[4]. Compared to Illumina platforms, ONT has a higher error rate, potentially making distinguishing between a wildtype and a single amino acid variant sequence challenging. However, we are working to overcome this by utilizing unique barcodes for each protein or relying on individual codon optimization sequences (unpublished work).

Our initial project design aimed to conduct a true all-vs-all experiment, pairing every protein with every other protein and performing selection in 3-AT. However, during data analysis, we encountered two major obstacles that shifted the project's direction. First, LSM2 and LSM3 interacted with a disproportionately large percentage of proteins, consuming most of the sequencing reads. Additionally, shorter protein-protein pairs dominated the nanopores, preventing sufficient detection of longer protein pairs. Continuing with this design would have required significantly more sequencing reads with minimal chances of observing the longer protein

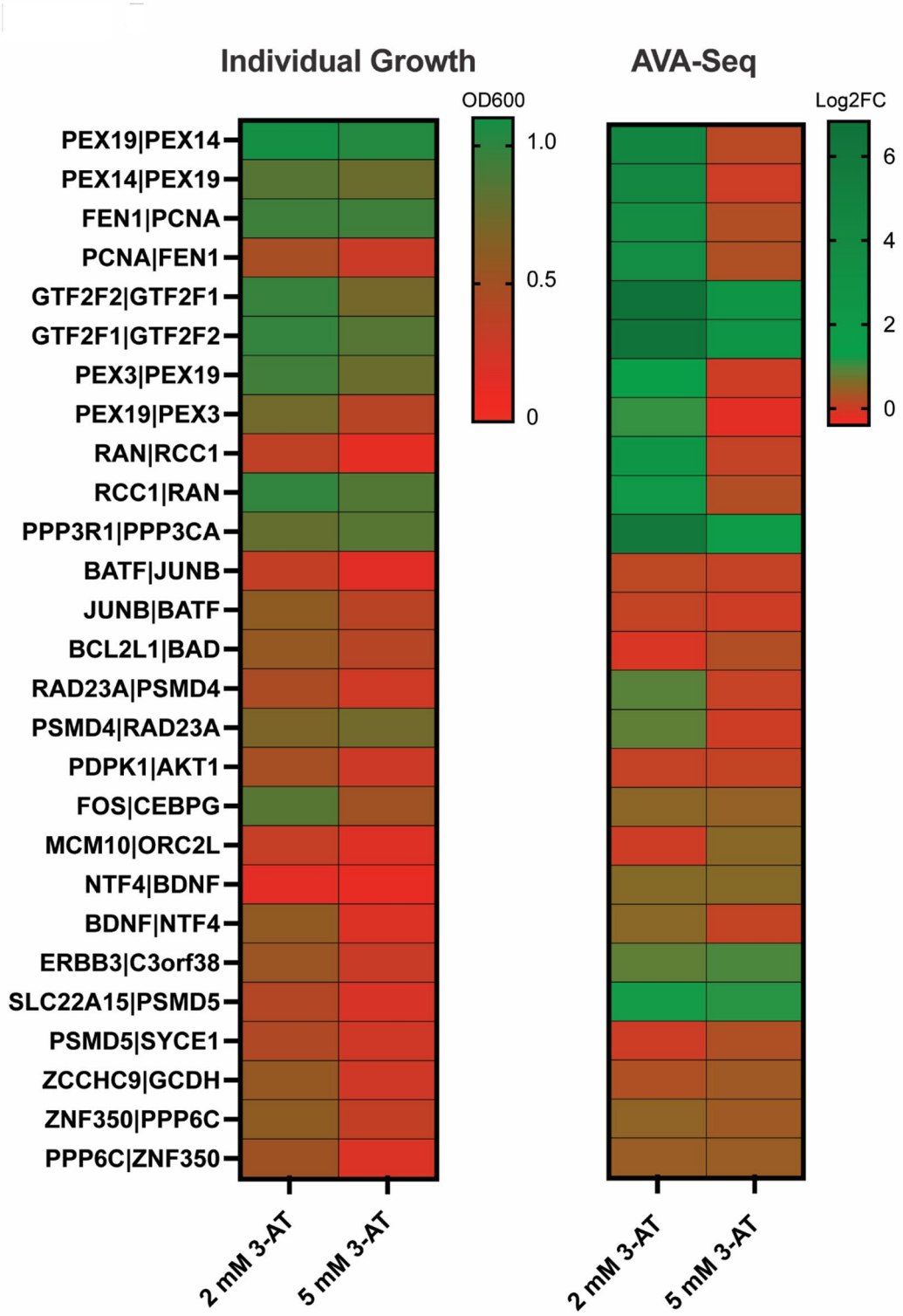

Fig. 5. Correlation of low-throughput (individual growth in liquid culture) and high-throughput all-vs.-all sequencing (AVA-Seq). Log_2FC max is shown for 2 mM and 5 mM 3-AT growth conditions. FDR < 0.1. Interactions were considered if the $log_2FC > 0.8$ and in two experiments. Known human interacting pairs from this study are listed in both orientations in the left column. Not all orientations were successfully cloned resulting in missing data for the individual growth (Supplemental Table 1). Selected individual non-interacting pairs (as reported in HsRRS-V2 [34]) ERBB3|C3orf38, SLC22A15|PSMD5, ZCCHC9|GCDH, ZNF350|PPP6C, PPP6C|ZNF350 were tested in the individual liquid growth (in those specific orientations only) and AVA-Seq high throughput. ERBB3|C3orf38 and SLC22A15|PSMD5 showed AVA-Seq growth in only 1 experiment out of three tested so they do not meet the criteria for being considered an interaction. SLC22A15 is annotated as a 'sticky' protein which may explain its growth.

interactions. To address this issue, we revised the project by dividing the proteins into two length-based pools and excluded LSM2 and LSM3, which we considered extremely "sticky." This redesign, while not perfect, allowed us to obtain meaningful interaction data.

Proteins for this full-length study were selected from the hsPRS-V2, some of which were also utilized in a fragment-based AVA-Seq project[37] allowing us to compare fragment-based and full-length versions. Differentiating between "sticky" proteins and auto-activators remains a challenge. Sticky proteins are defined as those capable of facilitating many interactions, indicating low specificity, while auto-activators primarily activate the system in only one orientation. For example, LMNA interacts with over 1,000 proteins in the APID database, compared to an average of 45 interactions per protein. Proteins tested in this study showed a mean of 170 interactions, suggesting their suitability for individual tests but highlighting limitations for large-scale all-vs-all interaction screens. Our growth time and sequencing read depth testing detected a range of interaction strengths and coverages, similar to RNA-seq, where read coverage affects gene detection.

To distinguish between "sticky" proteins and auto-activators, we analyze interactions in both orientations. Proteins interacting with frameshifted fragments paired to both RNAP and LambdaCI (two orientations) are flagged as potentially sticky, while those interacting in only one orientation are considered less likely to be sticky. The complexity arises because they are more likely to mistakenly activate the system by binding to either RNAP or LambdaCI specifically rather than binding to any protein they encounter (truly sticky). Some of this complexity arises from the nature of the AVA-Seq system itself, which tests all-by-all combinations simultaneously, unlike traditional binary yeast two-hybrid screens, where one bait is tested against many prey. This high-throughput format increases the likelihood of capturing both true and false-positive interactions, making it essential to distinguish sticky proteins from auto-activators.

We recognize that incorporating quantitative and computational approaches could enhance the distinction between sticky proteins and auto-activators. For instant, future work could involve analyzing stickiness by evaluating surface hydrophobicity or other biophysical properties associated with promiscuous interactions[42,43]. Notably, work by Ispolatov and colleagues[44] has highlighted key structural features that contribute to non-specific binding. Incorporating such insights into future AVA-Seq pipelines remains an important direction for improving interaction confidence.

We have applied the AVA-Seq protein-protein interaction method to full-length proteins using advances in synthetic biology and long-read sequencing technology. The extended dynamic range from sequencing on ONT makes applying an all-vs.-all protein interactome approach valuable – meaning the progressively weaker interactions can be detected by deeper sequencing to increase sensitivity. The AVA-Seq method takes advantage of the foundation of the B2H system and combines it with high-throughput sequencing methods. The more robust growth of a protein pair indicates a greater ability to overcome the 3-AT inhibition and correlates to a stronger protein interaction. At the same time, lack of growth in 2 mM 3-AT indicates an interaction that cannot overcome this mild inhibition. Our system challenges protein-protein interactions in 2 and 5 mM 3-AT concentrations, allowing for greater sensitivity in recovering interactions that other overly stringent methods would likely miss[9]. Furthermore, requiring that interactions must be captured in multiple experiments and multiple fusion orientations further reduces the likelihood of these interactions being false.

Expanding this project to a larger scale—such as incorporating all 60 PRS pairs and 78 RRS pairs—is of significant interest to us. The use of synthetic genes would make this feasible, but budget constraints limited the scope of the current study. Nonetheless, this work underscores the versatility of AVA-Seq, highlighting its applicability for both high-throughput research and low-throughput applications. The latter is particularly valuable in undergraduate research projects or Course-Based Undergraduate Research Experiences (CUREs), where the emphasis is on engaging students in meaningful scientific inquiry using affordable and accessible methods.

The AVA-Seq method is a powerful tool to detect protein fragment-fragment interactions[8,36,37] and full-length protein-protein interactions. Conducting full-length interaction studies with AVA-Seq allows researchers to move beyond colony formation. Utilizing NGS platforms allows for quantitative readout of PPIs as the $\log_2$FC from the high-throughput experiments correlates with the growth in individual screening using liquid culture. NGS allows for increased sensitivity in detecting weak or transient protein-protein interactions often missed with mainstream platforms. The AVA-Seq method is an important tool in the high-throughput protein interaction mapping repertoire as it recovers a unique population of interactions. Furthermore, this manuscript demonstrates a novel application of ONT in identifying full-length protein-protein interactions.

Methods

Pooling of proteins

Synthetic genes were codon optimized for E. coli and contained specific 5'- and 3'- adapter sequences to facilitate directional PCR amplification. 1 µg of lyophilized DNA were resuspended in 50 µL of ultrapure water. Genes were pooled at equal molar concentrations. Two different gene pools were made, reflecting different gene lengths. The short gene pool (< 1,299 bp) contained 35 genes and two additional frameshifted fragments. The long gene pool (genes > 1,300 bp) contained 22 genes plus two additional long frameshifted fragments. The frame-shifted fragments code for an early stop codon and help assess potential auto-activators. The protein pools were used the same day for PCR amplification and stored at -20 °C.

Generating the full-length protein-protein fusions

DNA was synthesized (TWIST Bioscience) for 57 human genes using codon optimization for expression in E. coli, and specific 5'- and 3'-adapter sequences were added. These adapters allow for directional amplification using different overhanging primers, allowing two genes to be merged into the convergent fusion DNA strand. After assembly, the inserts have an equal chance of being fused to the DBD or the AD contained in the pAVA

plasmid. This also allows for simultaneous screening in multiple orientations. Two separate amplifications were performed using the pooled DNA to ensure high coverage of all proteins.

Ligation into pAVA plasmid and Preparation for 3-AT selection

A ligation ratio of 1:1 vector to insert was used to reduce multiple insertions. The column-cleaned inserts are ligated using T4 DNA ligase (NEB M0202) to the linearized pAVA plasmid [DPN1 (NEB R0176) and rSAP (NEB M0371) treated before use] and incubated at 16 °C overnight. The ligation product is column-cleaned (Sigma-Aldrich NA1020) and transformed using NEB Turbo Electrocompetent cells (NEB C2986; discontinued). The DNA were then extracted (Sigma-Aldrich PLN350) and transformed using the validation reporter cell line using < 5ng DNA per transformation to ensure minimal double transformants[45].

3-AT selection and controls

As previously reported, 3-AT interaction selections were performed using the transformed validation reporter cell line[37]. High-throughput AVA-Seq validation (Full length AVA-Seq): experiments were conducted in triplicate (unique transformation events). Each experiment contained three technical replicates of 0 mM, 2 mM, and 5 mM 3-AT liquid growth conditions at 37 °C for 9 h, as reported previously[8,35,37].

Low-throughput validation (AVA one-by-one): For the individual full-length known interacting protein pairs, cultures were washed four times in minimal media and diluted to a final concentration of $OD_{600} = 0.015$. Constructs were screened in a 96-deepwell plate with a volume of 1 mL in quadruplicate for all three conditions (0 mM, 2 mM, and 5 mM 3-AT) at 37 °C for 9 or 16 h. The OD_{600} of the endpoint was measured using a standard 96-well plate with 100 µL of minimal media and 40 µL of culture using the minimal media as a blank. Measurements were normalized to the final 0 mM readings.

MinION library preparation and sequencing

The DNA were extracted from each 3-AT replicate sample (Sigma-Aldrich PLN350) and quantified using QuBit HS (Invitrogen Q32854). Three 0 mM, three 2 mM and three 5 mM samples comprise one experiment. Twenty-seven libraries (9 for each replicate experiment) were prepared using the Ligation Sequencing Kit protocol for Ligation sequencing amplicons - native barcoding (SQK-LSK109 with EXP-NBD196, Oxford Nanopore Technologies). Briefly, 130 ng for 1 kb amplicon DNA was prepared for each sample. Then amplicon DNA were end-repaired using NEBNext Ultra II End-Repair/dA-tailing Module (NEB E7546). Subsequently, each sample was cleaned with AMPure XP beads (Beckman Coulter A63881). Ligation of the Native barcode was then performed by adding Blunt/TA Ligase Master Mix (NEB M0367) and a unique native barcode (Oxford Nanopore Technologies EXP-NBD196) for each end-prepped DNA. The samples were then cleaned with AMPure XP beads and quantified with QuBit HS. The library pool was prepared by combining equimolar amounts of barcoded samples into an Eppendorf tube for a final molarity of 0.4 pmoles. Finally, the adapter ligation and clean-up were performed following the manufacturer's instructions. To ensure all DNA fragments are retained, the pool was cleaned using Short Fragment Buffer (Oxford Nanopore Technologies SQK-LSK109). The resulting pool ~ 80 fmol was then loaded onto the MinION flow cell (Oxford Nanopore Technologies FLO-MIN106D) and sequenced for 72 h.

The average number of sequencing reads per protein pair is 444 reads. If we look at individual conditions, the average number of reads of in the 0 mM selection conditions (average of 3 experimental replicates) was 147 reads, 2 mM was 157 reads and 5 mM conditions had an average of 139 reads.

Consistency of the results were tested by correlation analysis of the replicate screens. We tested the log2 fold change protein pairs across the 3 independent screens (2 mM versus 0 mM) showing an R^2 correlation coefficient of about 0.8 meaning highly consistent results for the same protein pairs tested at different times (Supplementary Fig. 2).

Data analysis

FASTQ sequence files from the Oxford Nanopore sequencer contain long reads spanning the two encoded convergent proteins being tested. The FASTQ files were searched for the multiple STOP insert sequence[8] using custom Perl scripts allowing for fuzzy sequence matching. Sequences to the right and left of the STOP insert were retained and searched for sequences indicating a fusion to either the lambda cI (DNA binding domain; DBD) or RNAP (activation domain; AD) of the system. Both sequences were then searched against the sequence database for this project using BLASTN[46]. Paired sequences that exhibited high quality BLAST matches with our sequence database were retained for further analysis. Subsequently, protein pairs were compiled, and each instance of identical protein pairs observed across separate read pairs led to an increment in count. The counts for each protein pair across all nine replicates (comprising triplicates of 0 mM, 2 mM, and 5 mM 3-AT) were aggregated into a table for subsequent statistical analysis in R.

Protein pairs were removed if they had a cumulative count of less than 10 reads in the combined three baseline replicates (0 mM 3-AT). Pairs failing to meet this threshold were excluded due to challenges in statistical processing arising from low count numbers.

Using the R package edgeR[47] protein pairs exhibiting a statistically significant increase under selective conditions (2 mM or 5 mM 3-AT) compared to the background (0 mM 3-AT) were identified. Internally, edgeR performed normalization of count values to accommodate varying sequencing depths reflected in differing library sizes. A negative binomial model was applied to ascertain differential growth, utilizing Fisher's exact test for significance testing, which computed adjusted p-values or false discovery rates (FDRs) for each protein pair.

Based on our experience the $\log_2$ fold change ($\log_2$FC) threshold changes based on the project. More specifically when we use protein fragments, we can set the fold change threshold higher. We believe this is due

to several reasons. First, we can test many more combinations using the fragment-fragment screening and the fragments are also all (intentionally) the same length, so they seem to have uniform growth.

However, with the full-length proteins their lengths, and subsequent combination of lengths, vary significantly more than with the fragment approach. We still found that a 9-hour growth time was sufficient for us to get consistent change in read count (to call for interactions), but the overall $\log_2$FC was much less. Meaning that the threshold for full length protein interaction projects such as this needed to be adjusted based on the data we observed.

Further scrutiny led to the identification of protein pairs meeting specific criteria: a $\log_2$ fold change ($\log_2$FC) > 0.8 and FDR < 0.1 in the presence of 3-AT compared to 0 mM 3-AT, denoting potential interactions. That these cutoffs gave consistent and robust results was tested across multiple experiments of the same screens and volcano plots showed the same interactions passing the thresholds regularly (Supplementary Fig. 2). Additionally, more stringent filters were applied to eliminate interactions lacking robust support. For the all-versus-all analysis, reporting a PPI necessitated repeatability: at least two experiments per protein pair orientation with $\log_2$FC > 0.8 and FDR < 0.1. In instances where one orientation exhibited interaction across multiple experiments, an interaction in the other orientation (with $\log_2$FC > 0.8 and FDR < 0.1) was acknowledged, even if detected in only one experiment.

Data availability

The data underlying this article are available using BioProject ID PRJNA1095688.https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA1095688.

Received: 24 June 2024; Accepted: 23 June 2025
Published online: 17 July 2025

References

1. Suter, B. et al. Next-generation sequencing for binary protein-protein interactions. *Front. Genet*. https://doi.org/10.3389/fgene.2015.00346 (2015).
2. Yu, H. et al. Next-generation sequencing to generate interactome datasets. *Nat. Methods* **8,** 478–480 (2011).
3. Yachie, N. et al. Pooled-matrix protein interaction screens using barcode fusion genetics. *Mol. Syst. Biol.* **12**, 863 (2016).
4. Weimann, M. et al. A Y2H-seq approach defines the human protein methyltransferase interactome. *Nat. Methods* **10**, 339–342 (2013).
5. Lewis, J. D. et al. Quantitative interactor screening with next-generation sequencing (QIS-Seq) identifies Arabidopsis thaliana MLO2 as a target of the Pseudomonas syringae type III effector HopZ2. *BMC Genom.* https://doi.org/10.1186/1471-2164-13-8 (2012).
6. Yang, F. et al. Development and application of a recombination-based library versus library highthroughput yeast two-hybrid (RLL-Y2H) screening system. *Nucleic Acids Res.* **46** (2018).
7. Johnson, K. L. et al. Revealing protein-protein interactions at the transcriptome scale by sequencing. *Mol. Cell* **81** (2021).
8. Andrews, S. S. et al. High-resolution protein–protein interaction mapping using all-versus-all sequencing (AVA-Seq). *J. Biol. Chem.* **294**, 11549–11558 (2019).
9. Pashkova, N. et al. DEEPN as an approach for batch processing of yeast 2-hybrid interactions. *Cell Rep.* **17** (2016).
10. Boldridge, W. C. et al. A multiplexed bacterial two-hybrid for rapid characterization of protein–protein interactions and iterative protein design. *Nat. Commun.* **14** (2023).
11. Stynen, B., Tournu, H., Tavernier, J. & Van Dijck, P. Diversity in genetic in vivo methods for Protein-Protein interaction studies: From the yeast two-hybrid system to the mammalian split-luciferase system. *Microbiol. Mol. Biol. Rev.* **76** (2012).
12. Luck, K. et al. A reference map of the human binary protein interactome. *Nature* **580**, 402–408 (2020).
13. Rolland, T. et al. A Proteome-Scale map of the human interactome network. *Cell* **159**, 1212–1226 (2014).
14. Stelzl, U. et al. A human protein-protein interaction network: A resource for annotating the proteome. *Cell* **122** (2005).
15. Dreze, M. et al. Evidence for network evolution in an Arabidopsis interactome map. *Science (1979)*. **333**, 601–607 (2011).
16. Zhou, Y. et al. A comprehensive SARS-CoV-2–human protein–protein interactome reveals COVID-19 pathobiology and potential host therapeutic targets. *Nat. Biotechnol*. **41** (2023).
17. Rain, J. C. et al. The protein–protein interaction map of Helicobacter pylori. *Nature* **409**, 211–215 (2001).
18. Hauser, R. et al. A second-generation protein-protein interaction network of Helicobacter pylori. *Mol. Cell. Proteomics*. https://doi.org/10.1074/mcp.O113.033571 (2014).
19. Li, S. et al. A map of the interactome network of the metazoan *C. elegans*. *Sci. (1979)*. **303**, 540–543 (2004).
20. Ito, T. et al. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl. Acad. Sci. U S A* **98** (2001).
21. Uetz, P. et al. A comprehensive analysis of protein–protein interactions in *Saccharomyces cerevisiae*. *Nature* **403** (2000).
22. Bonthala, B. et al. ONT-based draft genome assembly and annotation of *Alternaria atra*. *Mol. Plant-Microbe Interact*. **34** (2021).
23. Liu, X. et al. Improved whole-genome sequence of *Fusarium meridionale*, the fungal pathogen causing fusarium head blight in rice. *Mol. Plant-Microbe Interact.* **35** (2022).
24. Shi, J. et al. Improved whole-genome sequence of *Phytophthora capsici* generated by long-read sequencing. *Mol. Plant-Microbe Interact.* **34** (2021).
25. Schatz, M. C. Nanopore sequencing meets epigenetics. *Nat. Methods* **14** Preprint https://doi.org/10.1038/NMETH.4240 (2017).
26. Brannan, K. W. et al. Robust single-cell discovery of RNA targets of RNA-binding proteins and ribosomes. *Nat. Methods* **18** (2021).
27. Altemose, N. et al. DiMeLo-seq: A long-read, single-molecule method for mapping protein–DNA interactions genome wide. *Nat. Methods* **19** (2022).
28. Maslan, A. et al. Mapping protein-DNA interactions with DiMeLo-seq. *BioRxiv* 2022.07.03.498618. https://doi.org/10.1101/2022.07.03.498618 (2022).
29. Cheetham, S. W. et al. Single-molecule simultaneous profiling of DNA methylation and DNA-protein interactions with Nanopore-DamID. bioRxiv 2021.08.09.455753 https://doi.org/10.1101/2021.08.09.455753 (2022).
30. Mariani, M. et al. Higher-order chromatin structures of chromosomally integrated HHV-6A predict integration sites. *Front Cell. Infect. Microbiol* **11** (2021).
31. Sugihara, Y. et al. Disentangling the complex gene interaction networks between rice and the blast fungus identifies a new pathogen effector. *PLoS Biol.* **21**, e3001945 (2023).
32. Zuin, J. et al. Nonlinear control of transcription through enhancer–promoter interactions. *Nature* **604** (2022).
33. Springstein, B. L., Deighan, P., Grabe, G. J. & Hochschild, A. A bacterial cell-based assay to study SARS-CoV-2 protein-protein interactions. *mBio* **12** (2021).

34. Choi, S. G. et al. Maximizing binary interactome mapping with a minimal number of assays. *Nat. Commun.* **10**, 3907 (2019).
35. Al-Thani, N. M., Schaefer-Ramadan, S., Aleksic, J., Mohamoud, Y. A. & Malek, J. A. Identifying novel interactions of the colon-cancer related APC protein with Wnt-pathway nuclear transcription factors. *Cancer Cell. Int.* **22**, 376 (2022).
36. Schaefer-Ramadan, S., Aleksic, J., Al-Thani, N. M. & Malek, J. A. Novel protein contact points among TP53 and minichromosome maintenance complex proteins 2, 3, and 5. *Cancer Med.* **11**, 4989–5000 (2022).
37. Schaefer-Ramadan, S. et al. Scaling-up a fragment-based protein-protein interaction method using a human reference interaction set. *Proteins Struct. Funct. Bioinform.* **90**, 959–972 (2022).
38. Alonso-López, D. et al. APID database: Redefining protein-protein interaction experimental evidences and binary interactomes. *Database* **2019** (2019).
39. Nomizu, M., Otaka, A., Utani, A., Roller, P. P. & Yamada, Y. Assembly of synthetic laminin peptides into a triple-stranded coiled-coil structure. *J. Biol. Chem.* **269** (1994).
40. Szklarczyk, D. et al. STRING v10: Protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res.* **43** (2015).
41. Lawrie, J., Song, X., Niu, W. & Guo, J. A high throughput approach for the generation of orthogonally interacting protein pairs. *Sci. Rep.* **8**, 867 (2018).
42. Deeds, E. J., Ashenberg, O. & Shakhnovich, E. I. A simple physical model for scaling in protein-protein interaction networks. *Proc Natl. Acad. Sci. U S A* **103** (2006).
43. Zhang, J., Maslov, S. & Shakhnovich, E. I. Constraints imposed by non-functional protein-protein interactions on gene expression and proteome size. *Mol. Syst. Biol.* **4** (2008).
44. Ispolatov, I., Yuryev, A., Mazo, I. & Maslov, S. Binding properties and evolution of homodimers in protein-protein interaction networks. *Nucleic Acids Res.* **33** (2005).
45. Dove, S. L. & Hochschild, A. A. Bacterial two-hybrid system based on transcription activation. In *Protein-Protein Interactions*. Vol. 261. 231–246 (Humana Press, 2004).
46. Camacho, C. et al. BLAST+: Architecture and applications. *BMC Bioinform.* **10** (2009).
47. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. EdgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

Acknowledgements

We would like to thank the members of the Genomics Core at Weill Cornell Medicine – Qatar.

Author contributions

Conceptualization: S.S-R. and J.M., Methodology: S.S-R., J.A., A.A. and J.M., Validation: S.S-R., Y.G., J.A. and J.M. Formal analysis: J.A. and J.M., Investigation: S.S-R., Y.G. and A.A., Data curation: J.A., K.E., L.S., Y.M. and J.M. Writing-Original draft preparation: S.S-R. and J.M., Writing-Reviewing and Editing: S.S-R. and J.M., Visualization: S.S-R. and J.M., Supervision and project administration: S.S-R., Y.M. and J.M. Funding acquisition: J.M.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-08549-3.

Correspondence and requests for materials should be addressed to J.A.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.