Title: Yeast Farnesyltransferase has broad reactivity with CXXX sequences

Authors: June H. Kim[1], Emily R. Hildebrandt[1], Anushka Sarkar[1], Wayland Yeung[2], La Ryel A. Waldon[1], Natarajan Kannan[1, 2], Walter K. Schmidt[1]

[1]Department of Biochemistry and Molecular Biology, University of Georgia, Athens GA 30602

[2]Institute of Bioinformatics, University of Georgia, Athens GA 30602

Corresponding author: Walter K. Schmidt, wschmidt@uga.edu (email), 706-583-8241 (phone), 706-542-1738 (fax).

Running title: Yeast FTase acts on CXXX sequences

Keywords: protein farnesylation, farnesyltransferase, yeast genetics, substrate specificity, next-generation sequencing

Abstract (word count 144)

The basis for the current understanding of farnesyltransferase (FTase) specificity was pioneered through investigations of reporters like Ras and Ras-related proteins that possess a C-terminal CaaX motif that consists of 4 amino acid residues: Cysteine – aliphatic$_1$ – aliphatic$_2$ – variable (X). These studies led to the finding that proteins possessing the CaaX motif are subject to a 3-step post-translational modification pathway involving farnesylation, proteolysis, and carboxylmethylation. However, emerging evidence indicates that FTase can farnesylate sequences outside the canonical CaaX motif and that these motifs are only subject to farnesylation. This work investigates the scope of yeast FTase specificity through the use of genetic and next-generation sequencing methods using the reporter Ydj1, an HSP40 chaperone that only requires farnesylation for its activity. Our results support a wide array of sequence targets beyond the CaaX consensus and expand the potential target space of FTase within the yeast proteome.

Abbreviations

Farnesyltransferase (FTase), geranylgeranyltransferase (GGTase), next-generation sequencing (NGS), Ras Recruitment System (RRS), post-translational modification (PTM)

Introduction

Isoprenylation is a post-translational modification (PTM) catalyzed by several isoprenylation enzymes (i.e., FTase, GGTase-I, GGTase-II, GGTase-III). This PTM enhances protein association with cellular membranes or strengthens protein-protein interactions (Lane and Beese 2006, Leung, Baron et al. 2007, Tate, Kalesh et al. 2015, Wang and Casey 2016). Collectively, isoprenylated proteins assist in cellular signaling, cell cycle regulation, development, aging, and various other biological processes (Greenwood, Steinman et al. 2006, Palsuledesai and Distefano 2015, Jeong, Suazo et al. 2018).

Ras signaling GTPases are often cited as classical examples of isoprenylated proteins (Willumsen, Christensen et al. 1984, Wright and Philips 2006, Zhou, Wiener et al. 2016). Ras isoprenylation is catalyzed by farnesyltransferase (FTase), which acts upon a C-terminal CaaX motif consisting of 4 amino acid residues: Cysteine – aliphatic$_1$ – aliphatic$_2$ – variable (X). FTase appends the C15 isoprenoid lipid donated by farnesyl pyrophosphate to the CaaX cysteine via a thioether linkage. After farnesylation, Ras undergoes the coupled modifications of endoproteolysis to remove the -aaX portion of the motif, followed by carboxylmethylation of the farnesylated cysteine that becomes the new C-terminal residue. These PTMs regulate Ras plasma membrane localization and function, and have been extensively studied due to the significance of Ras GTPases in human disease such as cancer (Tamanoi 2011, Cox, Der et al. 2015, Hobbs, Der et al. 2016). As a result, previous Ras-based investigations have led to the general model that the three CaaX PTMs (i.e., isoprenylation, proteolysis, methylation) commonly occur across a range of proteins that are collectively referred to as CaaX proteins. CaaX

PTMs exist in all eukaryotic species, and the impact of these PTMs on CaaX protein biology are highly conserved across model systems (Omer, Kral et al. 1993, Cazzanelli, Pereira et al. 2018, Ravishankar, Hildebrandt et al. 2023). This is especially evident in the conserved similarities beween and mammalian and yeast FTase structure and function (Kohl, Diehl et al. 1991, Gomez, Goodman et al. 1993, Omer, Kral et al. 1993).

Not all farnesylated proteins undergo all three CaaX PTMs. An example is the *Saccharomyces cerevisiae* Hsp40 chaperone Ydj1 (*Sc*Ydj1) that is farnesylated on its C-terminal CASQ sequence but not subjected to endoproteolysis or carboxylmethylation. Farnesylation of Ydj1 is required for optimal yeast growth at elevated temperatures (>37 °C) and subjecting Ydj1 to all three CaaX PTMs negatively impacts this Ydj1-dependent thermotolerance growth phenotype (Caplan et al., 1992, Hildebrandt et al., 2016). We have defined this non-canonical pathway leading to a farnesylation-only PTM as the “shunt” farnesylation pathway (**Figure 1**). The observation of the shunt pathway raises the question of whether the previous use of canonically-modified CaaX protein reporters (e.g., Ras, **a**-factor), which are subject to the additional constraints of proteolysis and carboxylmethylation, has limited the breadth of sequences that can be identified as FTase substrates.

Many past investigations have attempted to probe FTase specificity. Initially, the screening of FTase substrates was explored *in vitro* on a case-by-case basis using recombinant CaaX proteins and synthetic peptides with [^{3}H]-FPP or [^{3}H]-GGPP, which was inconvenient and labor intensive for large scale studies (Moores, Schaber et al. 1991, Caplin, Hettich et al. 1994). *In vitro* investigations using peptide libraries and metabolic labeling have extended this effort but have been limited in scope and cost-

prohibitive, making it difficult to conduct systematic studies of the 8000 CXXX sequence space (Reiss, Stradley et al. 1991, Krzysiak, Scott et al. 2007, Krzysiak, Aditya et al. 2010, Wang, Dozier et al. 2014, Tate, Kalesh et al. 2015, Suazo, Schaber et al. 2016, Storck, Morales-Sanfrutos et al. 2019). While *in silico* prediction models based on structural analysis of mammalian FTase are available and have potential to define isoprenylatable sequences, these models often fail at predicting farnesylated proteins with non-canonical CaaX motifs such as *Sc*Ydj1 (CASQ), *Hs*DNAJA2 (CAHQ), *Sc*Pex19 (CKQQ), *Hs*Lkb1 (CKQQ) and *Sc*Nap1 (CKQS) and often lack orthologous *in vivo* reporter data to validate predictions (Collins, Reoma et al. 2000, Sapkota, Kieloch et al. 2001, Reid, Terry et al. 2004, Maurer-Stroh and Eisenhaber 2005, Lane and Beese 2006, Maurer-Stroh, Koranda et al. 2007, London, Lamphear et al. 2011, Berger, Yeung et al. 2022). Most of the *in vivo* reporters used to probe CaaX space have relied on CaaX protein reporters requiring 3-step modification and the superposition of 3 enzyme specificities for their activities (Boyartchuk, Ashby et al. 1997, Stein, Kubala et al. 2015). Moreover, recent reports demonstrating FTase activity on shortened and extended sequences highlight the flexibility of CaaX substrate lengths, additionally complicating the recognized model of FTase CaaX specificity (Ashok, Hildebrandt et al. 2020, Schey, Buttery et al. 2021).

The reliance of previous studies on canonical protein reporters and incomplete peptide reporter sets, in conjunction with our shunt pathway observations, led us to hypothesize that the full scope of FTase targets remains unknown. To address this gap in knowledge, we developed *Sc*Ydj1 as a genetic reporter to elucidate the specificity of the yeast FTase across all 8000 CXXX sequences. Our results indicate that yeast

FTase has a much broader target specificity than previously defined by the canonical CaaX motif.

Materials and Methods

Yeast strains: Strains used in this study are listed in **Table 1**. All plasmid-transformations into yeast were performed using a lithium acetate-based transformation procedure unless otherwise stated (Elble 1992).

Plasmid construction for Ydj1-CKQx variants and His-Nap1: Plasmids used in this study are listed in **Table 2**. Newly created plasmids needed for this study were constructed using PCR-directed recombinational cloning consistent with previously reported methods (Oldenburg, Vo et al. 1997, Hildebrandt, Cheng et al. 2016, Berger, Kim et al. 2018). For Ydj1-CKQX plasmids, pWS1132 that was linearized with *Nhe1* was co-transformed into yWS304 along with a PCR-derived DNA fragment encoding a new CaaX sequence. For the His-Nap1 plasmid (pWS1474), pRS316 that was linearized with *BamHI*, *XbaI* and *XhoI* was con-transformed into BY4741 along with a PCR derived DNA fragment encoding the *NAP1* gene. The resultant Nap1 plasmid (pWS1318) was additionally modified to encode an octa-His tag at the 5´ end of the *NAP1* gene; the tag was introduced by recombination using a PCR derived fragment and pWS1318 that had been linearized with *BsaBI*. In all cases, candidate plasmids recovered from yeast surviving appropriate genetic selection were evaluated by both restriction digest and DNA sequence analyses (GENEWIZ / Azenta Life Sciences, South Plainfield, NJ; Eurofins Genomics, Louisville, KY).

Plasmid construction for NGS library: The *YDJ1-CXXX* plasmid library (pWS1775) was designed to represent all 8000 “XXX” amino acid combinations where each amino acid is represented by a single unique codon by using trimer phosphoroamidites, aka “Trimer 20”, incorporated into the mutagenic oligos (Integrated DNA Technologies). To create the library, pWS1132 was first modified to introduce a silent unique *BsrGI* site directly 5´ to the SASQ encoding sequence using oligo oWS1330 (5´-AACTATGATTCCGATGAAGAAGAACAAGGTGGCGAAGGTGTACAATCTGCATCTCAATGATTTTC), creating pWS1728 (*CEN URA3 YDJ1-BsrGI-SASQ*). Next, PCR-derived DNA fragments were generated by pairing oWS1359 (5´-TATGATTCCGATGAAGAAGAACAAGGTGGCGAAGGTGTACAATGT-iTriMix20-iTriMix20-iTriMix20-TGATTTTCTTGATAAAAAAAGATC-3´) and oWS1308 (5´-CAGCATATAATCCCTGCTTTA-3´) and pWS1728 template using a high-fidelity Q5 polymerase (New England Biolabs, Ipswich, MA). The PCR products were purified (Omega E.Z.N.A Cycle Pure kit, Omega Bio-tek Inc., Norcross GA), digested with BsrGI and AflII, re-purified, and ligated into pWS1728 (BsrGI and AflII digested) using high concentration T4 DNA ligase (NEB M0202T). Ligations were transformed into DH5α *E. coli* prepared using Z-Competent cell kit (G-Biosciences, St. Louis, MO) and plated to achieve approximately 10^4 colonies per plate. In total, 606,800 colonies, each representing a unique clone, were pooled into a single cell pellet, followed by DNA purification (Omega E.Z.N.A Midiprep kit, Omega Bio-tek Inc., Norcross GA).

Preparation of naïve yeast library containing Ydj1-CXXX variants: The *E. coli*-derived Ydj1-CXXX plasmid library was transformed into yWS304 (*ydj1*Δ) via Frozen-EZ Yeast transformation II Kit by Zymo Research (Irvine, CA) according to manufacturer's instructions such that over 3 million individual colonies were recovered across multiple SC-Uracil plates on the same day. The colonies were collected by gently washing the plates with liquid medium, and cells concentrated by centrifugation to remove excess supernatant. The collected cells were stored at -80 °C as 100 µL aliquot stocks in 15% glycerol at a concentration of ~$2x10^9$ cells per 1 mL. For all subsequent studies, freshly thawed stock vials were used and not re-frozen for repeated use.

Thermotolerance selection of Ydj1-CXXX variants *en masse* and plasmid library recovery: An aliquot of the naïve yeast library expressing Ydj1-CXXX variants was thawed and used for each iteration of this assay. In brief, ~582,000 cells were diluted into 40 mL of room temperature SC-Uracil liquid media, and this premix was divided across 4 test tubes where sets of 2 tubes were incubated at either permissive (25 °C) or restrictive temperature (37 °C). Each test tube containing diluted culture was rapidly thermally equilibrated using an appropriate temperature water bath for 30 minutes before being placed onto a rotating wheel in an incubator at the same temperature. The cultures were incubated for approximately 24 to 48 hours until A_{600} 2.0 was reached, cells were collected by centrifugation, and plasmids were isolated from the recovered cell populations using a commercial kit (OMEGA Bio-Tek E.Z.N.A. Yeast Miniprep kit) as per manufacturer's instructions. The experiment was performed twice across two

different days, with each condition performed in duplicate, for a total of 8 replicates per temperature condition.

Next-generation sequencing and data analysis: The yeast plasmids recovered after thermoselection were subject to a shortened PCR with 15 cycles with oligonucleotide pairs oWS1408 (5´-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGCAGCATATAATCCCTGCTTTA-3´) and oWS1409 (5´-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGTCCAGGGGTGGTGCAAACTATG-3´) using a high-fidelity Q5 polymerase (New England Biolabs, Ipswich, MA) to attach overhang sequences. The PCR products were cleaned (OMEGA Bio-Tek E.Z.N.A. Cycle Pure Kit), quantified (Synergy H1 Hybrid Multi-Mode Microplate Reader), and submitted to the Georgia Genomics Bioinformatics Center (GGBC, Athens, GA) for Illumina MiSeq library assembly and sequencing (single-paired 100 bp reads starting 38 bases upstream of the CXXX region of interest).

The NGS analysis yielded over 10 million total reads, with ~96% of individual reads at or above the quality score of Q30. All 8 replicates from 25 °C yielded high-quality reads while only 7 replicates from 37 °C passed the quality cutoff. These 15 experimental samples, which represented about 3 million total reads, were carried forward for downstream analysis. In parallel, NGS was used to sequence 10 replicates of the *E. coli* plasmid library and 10 replicates of the plasmids extracted from the naïve yeast library prior to selection.

To assess the likelihood of farnesylation of Ydj1-CXXX variants, the reads within each replicate were first assessed to determine the number of occurrences for each unique CXXX sequence present. The occurrence of each CXXX sequence was summed across all replicates of the same temperature condition, then normalized to a frequency value. The latter was calculated by dividing the occurrence value for each unique CXXX by the total number of occurrences of all CXXX sequences (i.e., frequency = $count_{unique}/count_{total}$) for each of the 25 °C and 37 °C data sets. Lastly, the frequency value of each unique CXXX sequence was used to determine a unique enrichment factor (EF) score. This was calculated by dividing the frequency value of a unique CXXX sequence at restrictive temperature by that of the 25 °C condition (i.e., EF score = $frequency_{37\,°C}/frequency_{25\,°C}$). Some CXXX sequences were not recovered at 25 °C, yet present at higher temperatures (n=9; CFFM, CIFF, CILF, CIYM, CNWC, CTFA, CVFF, CVLW, CWIA). To account for such cases, a correction factor of +1 was applied across all frequency values at 25 °C so that 1 would be the denominator for CXXX sequences with 0 occurrences at 25 °C. We also applied the same correction factor of +1 to all frequency values at 37 °C for consistency.

Weblogo sequence alignments: The top 5% (n=400) sequences for both Ydj1-based and Ras-based screens were evaluated by Weblogo (http://weblogo.berkeley.edu/logo.cgi) using a custom color scheme (Crooks, Hon et al. 2004). Cys was set to blue; polar charged amino acids were set to green (Asp, Arg, Glu, His, and Lys); polar uncharged residues were set to black (Asn, Gln, Ser, Thr, and Tyr);

branched chain amino acids were set to red (Ile, Leu and Val); all other residues were set to purple (Ala, Gly, Met, Phe, Pro, and Trp).

Heatmap analysis: CXXX variants were clustered into 20-member groups having shared amino acids pairs. An average EF value was determined for these groups to allow for contextual analyses exploring the relationships between a_1 vs. a_2, a_1 vs. X, and a_2 vs. X. The averages were then analyzed with Microsoft Excel version 16.65 using the Conditional Formatting, Color Scales (Green - Yellow - Red) function to produce three distinct heatmaps (HMs). The values reported within each cell of a heatmap represent the average EF value for each respective 20-member group. The EF scores associated with RRS data (Stein et al., 2015) were evaluated in the same manner.

To determine confidence intervals, the average HM values were averaged across each individual row and column of the heatmap, and the averages were used to determine a standard deviation for either the row or column set of values. A standard deviation calculation was next used to determine a 95% confidence interval that was in turn used to establish high and low cutoffs for identifying positive selection or negative restriction. A pattern was deemed significant when 18 or more of each 20-member set (i.e., >90%) were above or below the statistical cutoff.

Likelihood of prenylation calculations for each CXXX variant: The averaged EF values from each of the 3 HMs were summed to produce a HM score. For example, to predict the likelihood of prenylation for CASQ, the averages of Cys-Ala-Ser-X (n=20 varied at the X position), Cys-Ala-X-Gln (n=20 varied at the a_2 position), and Cys-X-Ser-Gln

(n=20 varied at the a_1 position) were first calculated, then these three HM values were summed to represent the HM score of CASQ.

Ydj1-CXXX thermotolerance assay: The assay was performed as previously described (Hildebrandt, Cheng et al. 2016, Berger, Kim et al. 2018, Ashok, Hildebrandt et al. 2020). In brief, yeast cultures in SC-Uracil media were incubated at 25 °C until saturation, then serially diluted 10-fold in YPD before being pinned onto YPD solid medium. Plates were incubated for 2-4 days at 25 °C, 37 °C, and 39 °C prior to results being digitally scanned face down without lids using a Cannon flat-bed scanner (300 dpi; TIFF format). Scanned images were adjusted for consistency in image rotation, contrast, and size before being copied onto Microsoft PowerPoint version 16.65 for final figure assembly. The experiment was performed twice in duplicate.

Gel-shift assay: The assay was performed as previously described (Hildebrandt, Cheng et al. 2016, Berger, Kim et al. 2018, Schey, Buttery et al. 2021). Whole cell lysates of mid log cells were prepared and separated by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE; 6% stacking with 9.5% resolving gel) then transferred onto nitrocellulose. Blots were blocked with 5% milk then sequentially incubated with rabbit anti-Ydj1 primary antibody (courtesy of A. Caplan) and HRP-conjugated goat anti-rabbit secondary antibody (Kindle biosciences, Greenwich, CT). Fluorescence was detected using the KwikQuant Western Blot Detection Kit (Kindle Biosciences) and a KwikQuant Imager and as per manufacturer's instructions. Protein bands were quantified using NIH ImageJ, and resulting values were used for calculating ratios for

prenylated and non-prenylated bands. The blots containing yeast Nap1 were treated similarly but incubated with mouse anti-HIS primary antibody (Thermo Fisher Scientific, Waltham, MA) and HRP-conjugated sheep anti-mouse secondary antibody (Kindle biosciences, Greenwich, CT).

Decision tree modeling: Sequence motif features were generated by one-hot encoding for each of the variable CaaX motif sites: a_1, a_2, and X. Additional binary features were included to describe whether the variable residue was within a given set of residues. To define these sets, we enumerated all possible amino acid combinations with a max set size of 5. In total, 65,097 features were considered. While building the decision tree classifier, entropy was used to evaluate the quality of potential splits. To determine more generalizable rules, trees were allowed a maximum depth of 3 and all nodes were required to have a minimum of 50 samples. This method was implemented using Scikit-learn (v0.22.2) using the DecisionTreeClassifier class.

Results

Many CXXX sequences sustain Ydj1-dependent growth of yeast at high temperature:

Ydj1 follows an isoprenylation-only pathway in contrast to canonical reporters (i.e., Ras GTPases, **a**-factor) that undergo additional downstream modifications (Trueblood, Boyartchuk et al. 1997, Stein, Kubala et al. 2015) (**Figure 1**). As such, the use of Ydj1 as a reporter to probe yeast FTase specificity reduces the risk of introducing additional specificity bias from CaaX proteases. To evaluate the substrate scope of FTase, we adapted methods from a previous yeast study that utilized a Ras-based

CXXX reporter (i.e., Ras Recruitment System (RRS)) along with competitive growth enrichment and next-generation sequencing (NGS) methods (Stein, Kubala et al. 2015). In our case, we investigated the ability of Ydj1-CXXX variants to sustain high temperature growth in a *ydj1Δ* genetic background (**Figure S1**). A plasmid library of Ydj1-CXXX variants was generated using single codons for each amino acid so that all 8000 CXXX sequences are represented in a relatively small library (see Material and Methods). Compared to traditional plasmid library construction relying on fully or partly degenerate oligonucleotides, a Trimer 20-based strategy was used with the aim of yielding a balanced library with respect to codon redundancy (i.e., no over-representation of amino acids with multiple codons) and no introduction of early stop codons, which effectively reduces the number of copies needed to reach statistical confidence for full coverage (Firth and Patrick 2008). While 110,500 independent clones were minimally required for statistical 100% coverage, our library contains 606,800 independent clones (i.e., 6x full coverage). By contrast, constructing a library with fully degenerate oligonucleotides to vary 3 amino acids would have required 3,200,000 independent clones for 100% coverage, leading to much higher labor and time costs for this study.

The Ydj1-CXXX plasmid library DNA was purified from *Escherichia coli* before being transformed into the *ydj1*Δ yeast strain to yield a naïve *ydj1*Δ*/YDJ1-CXXX* yeast library. Both the *E. coli* plasmid library and the naïve yeast library were confirmed by NGS to contain all potential Ydj1-CXXX variants (**File S1**). Ligation efficiency of vector alone was determined to be ~0.3% relative to vector plus insert, indicating that a small portion of the library could encode *YDJ1-SASQ* (i.e., uncut parent plasmid).

Consistently, *YDJ1-SASQ* was observed at frequencies of 0.065% and 0.071% in the *E. coli* and naïve yeast libraries, respectively. The frequencies of all CXXX sequences from the *E. coli* and naïve yeast libraries were also graphed, and no significant changes in the frequency profile between the two libraries were observed (**Figure S2**). This analysis further revealed that the libraries were not perfectly balanced such that the extremes exhibited a ~10x range in abundance.

The naïve yeast library was propagated for approximately 8 generations at permissive (25 °C) and selective (37 °C) temperatures in liquid media until the culture was well saturated (A_{600} ~2.0). The selective temperature condition was expected to enrich for farnesylated Ydj1-CXXX variants. NGS methods were then used to identify all the CXXX sequences present in each culture, from which frequencies were determined for each CXXX sequence in each temperature group. Similarities in experimental design allowed for direct comparison of Ydj1 and Ras-based data sets. In the RRS study, the likelihood of farnesylation was reported by an Enrichment Factor (RRS EF) that was defined by the frequency of a specific CXXX sequence occurring at 37 °C divided by its frequency at 25 °C (i.e., $\text{frequency}_{\text{selective}}/\text{frequency}_{\text{permissive}}$). A high RRS EF was interpreted as a high possibility of farnesylation. We performed a similar calculation for the Ydj1 data.

From our analysis, we observed that EFs for the Ydj1-based screen exhibited a narrower range (EF: 0.036 for CWWC to 13.898 for CVFF) relative to the RRS EFs for the Ras-based screen (RRS EF: 0.008 for CLRS to 70.667 for CYCM). We interpret these ranges to indicate that the majority of all CXXX sequences can support growth in the Ydj1-based screen, whereas a smaller number of sequences undergo higher

enrichment during the thermoselection process in the Ras-based screen. The comparison also revealed that the EF profiles offered two distinct sequence landscapes, especially for farnesylated sequences that are well characterized: non-canonical CASQ (*Sc*Ydj1) and CKQQ (*Hs*STK11/Lkb1); canonical CVIA (*Sc* **a**-factor) and CVLS (*Hs*H-Ras). In the Ydj1 NGS-based screen, non-canonical sequences CASQ (EF: 1.608) and CKQQ (EF: 1.627) outperformed canonical CaaX sequences such as CVIA (EF: 0.406) and CVLS (EF: 0.494) (**Figure 2A**). By contrast, in the Ras-based screen, CASQ (RRS EF: 0.399) and CKQQ (RRS EF: 0.399) were significantly less enriched while CVIA (RRS EF: 10.124) and CVLS (RRS EF: 11.804) ranked amongst the top hits (**Figure 2B**). The EFs of all CXXX sequences resulting from the Ydj1 screen are reported in **File S2**.

Our observations suggest that the Ydj1-based screen best enriches non-canonical sequences whereas the Ras-based screen best enriches canonical sequences. The top 5% (n=400) of hits from the Ydj1 NGS-based assay did not exhibit an obvious consensus sequence (**Figure 2C**); however, the same number of top hits from the Ras-based screen were enriched with aliphatic amino acids at the a_1 and a_2 positions (**Figure 2D**). In the context of the Ras reporter, an aliphatic amino acid is especially prominent at the a_2 position, which has been historically regarded as a requirement for FTase specificity. The top 5% of hits from both the Ydj1-based and Ras-based screens were also displayed as a 4D plot (**Figures 2E, F**; the size of the spot is the 4th dimension and represents relative abundance). This analysis revealed that the top hits were more widely dispersed across CXXX sequence space in the Ydj1-based data relative to the Ras-based data, and that the most abundant sequences differed

between the data sets. Analysis of the data as 3D plots, viewed along each of the 4D plot axes such that Lysine (K) is the nearest amino acid, provided additional details about the sequence space covered by each screen (**Figure S3**). In these 3D plots, the spots representing relative abundance are stacked behind each other, and the total number of spots occupying a particular node is not easily discerned. Although this arrangement makes it difficult to make detailed conclusions about the depth of coverage at each node (i.e., the number of amino acids present), it allows for clear conclusions about restrictions (i.e., amino acids that are not tolerated independent of context). From the perspective of a_1, both data sets lack several amino acids at a_2 (i.e., D, E, G, K, and R) and a single amino acid at X (i.e., R) (**Figures S3A, B**). Specific to the Ras-based data, additional amino acids were absent at a_2 (H and Y) and X (K, P, and W) or less prevalent at a_2 (i.e., A, P, Q, S, T, and W). From the perspective of a_2, no amino acids were absent at the a_1 position in both data sets, one amino acid was absent at the X position in both data sets (i.e., R), and several amino acids were absent at the X position within the Ras-based data set only (i.e., K, P, R, and W) (**Figures S3C, D**). From the perspective of X, no amino acids were absent at the a_1 position, and the amino acids that were absent at a_2 from the perspective of a_1 were again identified, with one additional amino acid being absent (i.e., S) (**Figures S3E, F**). A general interpretation of these observations is that many amino acids can be accommodated at each position of the CXXX sequence independent of whether Ydj1 or Ras is the reporter, except for charged residues that are strictly not tolerated at a_2 in either case. There also appear to be additional exceptions in the context of the Ras reporter at both

the a_2 and X position that likely reflect its need for additional modification beyond initial isoprenylation.

To further analyze our Ydj1 NGS-based data, we crosschecked CXXX sequences that were previously identified as being farnesylated through a limited scope genetic Ydj1-based Temperature Screen (YTS) (Berger, Kim et al. 2018). Mostly non-canonical sequences were identified in the earlier study (n=153). Many of these sequences exhibited high EFs in the present study (**Figure 3A**). When superposed on the Ydj1 EF profile, most YTS identified sequences were positioned in the 4th quartile (n=77) with fewer in the 3rd quartile (n=67) and fewest in the 2nd quartile (n=9). None of the YTS hits were present in the 1st quartile. By contrast, superposition of the top hits from the Ras-based screen (n=496; RRS EF > 3; the cutoff for positive hits of the Ras-based screen) on the Ydj1 EF profile revealed positioning of sequences in the 2nd, 3rd, and 4th quartiles of the EF profile (n=228, n=114, and n=153, respectively), with the fewest in the 1st quartile (n=1) (**Figure 3B**). The few canonical sequences identified by YTS (**Figure 3A**, dashed box; n=15) remained within the range displayed by RRS sequences. Together, these results are fully consistent with our Ydj1 NGS-based method being useful for enriching non-canonical sequences that are farnesylated in addition to canonical sequences observed using the Ras reporter.

<u>Non-canonical CKQX sequences are targeted by FTase</u>:

Emerging evidence indicates that some CKQX sequences are farnesylated despite the presence of non-aliphatic amino acids at a_1 and a_2 positions. CKQQ farnesylation is well documented for human STK11/Lkb1, human Nap1L1 and yeast

Pex19, whereas CKQS derived from yeast Nap1 is farnesylated in the context of the Ydj1 reporter (Collins, Reoma et al. 2000, Sapkota, Kieloch et al. 2001, Tate, Kalesh et al. 2015, Storck, Morales-Sanfrutos et al. 2019, Berger, Yeung et al. 2022). We additionally confirmed that farnesylation of CKQS occurs in the natural context of yeast Nap1 itself using the gel-shift assay, which reports on the mobility of farnesylated species that migrates faster in the case of Ydj1, and slower in the case of Nap1, than non-farnesylated species produced using an FTase-deficient strain (*ram1*Δ) (**Figure S4A**). Moreover, the ScanProsite tool (https://prosite.expasy.org/scanprosite/) identifies 5179 entries across UniProtKB/Swiss-Prot and UniProtKB/TrEMBL reference proteome sequences ending in CKQX across eukaryotes (**File S3**). The majority of these sequences (i.e., 65%) contain CKQQ or CKQS. Out of the 20 CKQX variants, most (n=17 positive hits) displayed a Ydj1 EF consistent with a high likelihood of farnesylation (**Figure 4A**). The remaining sequences (n=3) had low EFs, were well separated from the other CKQX sequences on the EF profile, and were expected to have a low likelihood of modification (i.e., negative hits). By comparison, none of the 20 CKQX sequences were predicted to be targets of FTase in the Ras-based EF profile (**Figure 4B**).

The farnesylation status of Ydj1-CaaX variants were monitored by thermotolerance and gel-shift assays (Caplan, Tsai et al. 1992, Hildebrandt, Cheng et al. 2016). The thermotolerance assay confirmed the 17 CKQX positive hits to have growth profiles similar to that of wildtype Ydj1, whereas the negative hits (n=3) had a growth profile consistent with unmodified Ydj1 (**Figure 4C, Figure S4B**). In strong agreement with the EFs and thermotolerance profiles, a prenyl-dependent gel-shift was

observed for the 17 positive Ydj1-CKQX hits and no shift observed for the three negative hits. It should be noted that CKQR displayed a doublet gel pattern in both the presence and absence of FTase. The reason for this doublet pattern is unknown, however, the pattern was identical with and without FTase, so CKQR was deemed as having 0% prenylation.

Statistical analyses reveal yeast FTase specificity:

Our evaluation of genetic and biochemical data suggested that FTase is not limited to target sequences having aliphatic amino acids at a_1 and a_2. To identify factors that could better indicate target specificity, we assessed the likelihood of prenylation for sets of CaaX motifs that shared the same context, differing only at a single amino acid position. Using heatmap analysis, we sought to identify potential patterns marked by positive selection (colored in green) or negative restriction (colored in red) across a_1, a_2, and X positions in both Ydj1-based and Ras-based results (**Figure 5**). These patterns were identified as instances where the HM values of 18 or more of each 20-member set (i.e., >90% of EF scores) were well above or below the statistical average for all values (n=400) within a heatmap (**File S4**).

For the Ydj1-based data, no positive patterns were observed for the a_1 position in combination with a_2 (**Figure 5A**), however, one restrictive pattern, D at a_1, was observed in combination with X (**Figure 5B**). Generally, a_1 appears to tolerate many different types of amino acids and does not seem to strongly influence reactivity with FTase in either a positive or negative manner. When considering the a_2 position, no positive patterns were observed, but several negative patterns were evident.

Specifically, D, K and R were restrictive in combination with either a_1 (**Figure 5A**) or X (**Figure 5C**), with E being additionally restrictive in combination with a_1. Generally, a_2 also appears to tolerate many different types of amino acids except for charged amino acids that negatively influence reactivity with FTase. For the X position, a positive pattern, Q at X, was observed in combination with a_1 (**Figure 5B**). Several negative patterns were also evident at the X position. Both P and R were restrictive in combination with either a_1 or a_2, with K being additionally restrictive in the context of a_1. Generally, X too seems to tolerate many different types of amino acids except for positively charged amino acids and structurally constrained proline (**Figure 5C**). Thus, despite repeated reports that FTase targets canonical CaaX sequences, our Ydj1-based data strongly indicates that FTase tolerates most CXXX sequences unless amino acids are present that restrict its specificity. Similar heatmap analysis of the Ras-based data also revealed fewer positive than negative patterns. The two positive patterns were I at a_2 (**Figure 5D**) and M at X (**Figure 5E**), both in combination with a_1. Most of the negative patterns observed were primarily constrained to a_2 and X positions in all contexts (**Figure 5C, F**). The a_1 position, on the other hand, displayed fewest negative patterns; only two patterns were observed at E in combination with a_2 and K in combination with X. The combined observations that both Ydj1 and Ras-based data reveal high tolerance at the a_1 position, while a_2 and X positions have more restrictions is consistent with previous reports on FTase specificity (Reid, Terry et al. 2004, Stein, Kubala et al. 2015).

To fully account for the contextual information from heatmap analysis, the three heatmap values from the Ydj1 NGS dataset were summed for each specific CaaX motif

to obtain a heatmap (HM) score. This score represents the sum of 3 averages, where each average accounts for 20 EF data points. The advantage of the HM score is that it normalizes statistical outliers that could skew results when using individual EF scores. The HM scores, when a cutoff of 3 was applied, correlated well with prior evidence of prenylation gathered from previous studies, outperforming published models such as Prenylation Prediction Suite (PrePS), RRS, and our recently published SVM algorithm for predicting farnesylation of non-canonical sequences (Maurer-Stroh and Eisenhaber 2005, Stein, Kubala et al. 2015, Berger, Yeung et al. 2022) (**Table 3**).

Lastly, we trained a predictive model using the results of the Ydj1-based NGS screen to predict whether a given CXXX sequence can be modified based on a subset of questions (decisions) and possible consequences in a decision-tree model (**Figure 6, Figure S5**). Each question pertains to the whether certain amino acids are present in the a_1, a_2, and X positions. The decision tree was created by systematically identifying the best set of questions for predicting whether a CXXX sequence can be modified. This allowed us to identify a general set of rules which govern FTase substrate specificity. Based on our model, the biggest determinant of FTase activity is the restriction of D, E, K, or R at the a_2 position, followed by the restriction of K, P, or R at the X position. However, amino acids such as I, L, M and V at the a_2 position seem to neutralize the negative effects of K, P, or R at the X position. Our findings repeatedly raise the likelihood that restrictions, rather than tolerance, are applied to a few impermissible amino acids at a_2 and X positions of potential substrates modified by FTase. It should be noted that although this simple flowchart is useful in understanding general rules of

yeast FTase specificity, HM scores provide a more individualized prediction of the prenylation status of each CXXX sequence.

Discussion

In this study, the use of Ydj1 as a reporter instead of Ras allowed for the identification of yeast FTase targets that extend beyond the canonical CaaX consensus sequence. Notably, this study corroborates previously published observations that non-canonical sequences can be targets of FTase. Restrictions by a small number of amino acids, rather than adherence to the canonical CaaX consensus, seem to guide FTase specificity. We believe that these findings can be used to improve tools for predicting farnesylation status that will outperform current prediction methods.

Structural analysis of mammalian FTase has previously shown that the aaX portion of the protein substrate makes extensive Van der Waals contact with the adjacent isoprenoid group, which is hypothesized to modulate product release (Long, Casey et al. 2002, Reid, Terry et al. 2004). However, these interactions are not predicted when non-aliphatic residues are at a_1 and a_2. Our finding that FTase can indeed tolerate non-aliphatic residues at a_1 and a_2 raises the possibility that the catalytic site of FTase may display a greater degree of structural flexibility beyond the conformations that could be sampled from mammalian FTase in complex with canonical substrates. Our findings also lead us to predict that binding of charged amino acids (D, E, K, R) may be unstable in the a_2 binding pocket. Likewise, certain amino acids (K, P, R) at the X position may be unable to coordinate with the product binding pocket of the FTase. These situations may result in the formation of a farnesylated peptide that is

unable to be released from the enzyme, thus inhibiting FTase from turning over such substrates efficiently. These possibilities could be resolved by future structural studies involving non-canonical CXXX peptide substrates in complex with either yeast or mammalian FTase, which are structurally and functionally homologous (Kohl, Diehl et al. 1991, Gomez, Goodman et al. 1993, Omer, Kral et al. 1993).

Our study was focused on the recognition of CXXX sequences with respect to FTase specificity. It has been observed, however, that upstream sequences can impact FTase specificity (Cox, Graham et al. 1993). The impact of upstream sequences has been explored by prediction algorithms, most notably by PrePS (Maurer-Stroh and Eisenhaber 2005). It is critical to acknowledge that the Ydj1-CXXX NGS screen recovered sequences that support high temperature growth in the context of Ydj1. Thus, any effect of upstream sequence has not been explored in this work. Additionally, it is possible that some Ydj1-CaaX variants affect growth in ways unrelated to prenylation status, which could impact some of our results. Still, our YDJ1 HM scoring system outperformed predictions made by PrePS, RRS, and SVM machine learning algorithm (Maurer-Stroh and Eisenhaber 2005, Stein, Kubala et al. 2015, Berger, Yeung et al. 2022) (**Table 3**). Combining PrePS predictions that consider upstream sequence in addition to the observations presented in this work is expected to provide the most robust prediction for farnesylation of individual protein targets. As more of the prenylation predictions are validated using additional techniques in the future, it will be possible to determine which prediction methods are most reliable.

Many proteins undergo a process known as isoprenylation. Historically, FTase is recognized as targeting the CaaX sequence but emerging evidence, including that

reported here, has indicated that FTase substrates are not limited to the canonical CaaX sequence. Our methods, which primarily focused on yeast FTase, provides insights into the broader specificity of this process which likely extends to human and other FTase enzymes. The discovery of farnesylated non-canonical sequences such as CKQX associated with large families of proteins like Stk11/Lkb1 and Nap1, which lack a canonical CaaX motif, opens a new understanding of post-translational isoprenylation and the potential for future discoveries in this field. Our results also demonstrate that the target specificity of FTase is mainly due to restrictions at the a_2 and X positions, rather than selection towards a canonical CaaX sequence. As Ras and other canonical reporters rely on multiple modifications for function, previous studies using these proteins may have reported on sequences that are suited for the combined specificities of FTase and downstream enzymes (i.e., CaaX proteases *Sc*Ste24 or *Hs*Rce1) instead of the specificity of FTase alone. Our study, in combination with data from previous studies, provides new ways to advance the understanding of the specificities of each enzymatic step associated with the post-translational isoprenylation pathway, laying out a crucial step toward identifying its cellular targets and the extent to which they are modified.

Data availability

Yeast strains and plasmids are available upon request. All relevant datasets for this study are included in the supplemental files of the manuscript. **File S1** contains frequency of CXXX sequences observed in naive libraries. **File S2** contains Enrichment Factor results of Ydj1-based NGS. **File S3** contains CKQX hits identified on UniProt.

File S4 contains heatmap analysis. The data derived through the Ras-Recruitment System was previously published and reanalyzed here (Stein, Kubala et al. 2015).

Supporting information

None

Acknowledgement

We thank Dr. James Hougland (Syracuse University) for early discussions about this work, Dr. Avrom Caplan (City College of New York) for anti-Ydj1 antibody, and Somdut Roy (Georgia Institute of Technology) for the raw Python 3 scripts for the 4D graphical analysis. Lastly, we thank Schmidt lab members for constructive feedback and technical assistance.

Funding and additional information

This research was supported by Public Health Service grant GM132606 from the National Institute of General Medical Sciences (WKS, NK).

Conflict of interest

The authors have declared no competing interests exist.

References

Ashok, S., E. R. Hildebrandt, C. S. Ruiz, D. S. Hardgrove, D. W. Coreno, W. K. Schmidt and J. L. Hougland (2020). "Protein Farnesyltransferase Catalyzes Unanticipated Farnesylation and Geranylgeranylation of Shortened Target Sequences." Biochemistry 59(11): 1149-1162.

Berger, B. M., J. H. Kim, E. R. Hildebrandt, I. C. Davis, M. C. Morgan, J. L. Hougland and W. K. Schmidt (2018). "Protein Isoprenylation in Yeast Targets COOH-Terminal Sequences Not Adhering to the CaaX Consensus." Genetics 210(4): 1301-1316.

Berger, B. M., W. Yeung, A. Goyal, Z. Zhou, E. R. Hildebrandt, N. Kannan and W. K. Schmidt (2022). "Functional classification and validation of yeast prenylation motifs using machine learning and genetic reporters." PLoS One 17(6): e0270128.

Boyartchuk, V. L., M. N. Ashby and J. Rine (1997). "Modulation of Ras and a-factor function by carboxyl-terminal proteolysis." Science 275(5307): 1796-1800.

Brachmann, C. B., A. Davies, G. J. Cost, E. Caputo, J. Li, P. Hieter and J. D. Boeke (1998). "Designer deletion strains derived from Saccharomyces cerevisiae S288C: a useful set of strains and plasmids for PCR-mediated gene disruption and other applications." Yeast 14(2): 115-132.

Caplan, A. J., J. Tsai, P. J. Casey and M. G. Douglas (1992). "Farnesylation of YDJ1p is required for function at elevated growth temperatures in Saccharomyces cerevisiae." J Biol Chem 267(26): 18890-18895.

Caplin, B. E., L. A. Hettich and M. S. Marshall (1994). "Substrate characterization of the Saccharomyces cerevisiae protein farnesyltransferase and type-I protein geranylgeranyltransferase." Biochim Biophys Acta 1205(1): 39-48.

Cazzanelli, G., F. Pereira, S. Alves, R. Francisco, L. Azevedo, P. Dias Carvalho, A. Almeida, M. Corte-Real, M. J. Oliveira, C. Lucas, M. J. Sousa and A. Preto (2018). "The Yeast Saccharomyces cerevisiae as a Model for Understanding RAS Proteins and their Role in Human Tumorigenesis." Cells 7(2).

Collins, S. P., J. L. Reoma, D. M. Gamm and M. D. Uhler (2000). "LKB1, a novel serine/threonine protein kinase and potential tumour suppressor, is phosphorylated by cAMP-dependent protein kinase (PKA) and prenylated in vivo." Biochem J 345 Pt 3(Pt 3): 673-680.

Cox, A. D., C. J. Der and M. R. Philips (2015). "Targeting RAS Membrane Association: Back to the Future for Anti-RAS Drug Discovery?" Clin Cancer Res 21(8): 1819-1827.

Cox, A. D., S. M. Graham, P. A. Solski, J. E. Buss and C. J. Der (1993). "The carboxyl-terminal CXXX sequence of Gi alpha, but not Rab5 or Rab11, supports Ras processing and transforming activity." J Biol Chem 268(16): 11548-11552.

Crooks, G. E., G. Hon, J. M. Chandonia and S. E. Brenner (2004). "WebLogo: a sequence logo generator." Genome Res 14(6): 1188-1190.

Elble, R. (1992). "A simple and efficient procedure for transformation of yeasts." Biotechniques 13(1): 18-20.

Firth, A. E. and W. M. Patrick (2008). "GLUE-IT and PEDEL-AA: new programmes for analyzing protein diversity in randomized libraries." Nucleic Acids Res 36(Web Server issue): W281-285.

Giaever, G., A. M. Chu, L. Ni, C. Connelly, L. Riles, S. Veronneau, S. Dow, A. Lucau-Danila, K. Anderson, B. Andre, A. P. Arkin, A. Astromoff, M. El-Bakkoury, R. Bangham, R. Benito, S. Brachat, S. Campanaro, M. Curtiss, K. Davis, A. Deutschbauer, K. D. Entian, P. Flaherty, F. Foury, D. J. Garfinkel, M. Gerstein, D. Gotte, U. Guldener, J. H. Hegemann, S. Hempel, Z. Herman, D. F. Jaramillo, D. E. Kelly, S. L. Kelly, P. Kotter, D. LaBonte, D. C. Lamb, N. Lan, H. Liang, H. Liao, L. Liu, C. Luo, M. Lussier, R. Mao, P. Menard, S. L. Ooi, J. L. Revuelta, C. J. Roberts, M. Rose, P. Ross-Macdonald, B. Scherens, G. Schimmack, B. Shafer, D. D. Shoemaker, S. Sookhai-Mahadeo, R. K. Storms, J. N. Strathern, G. Valle, M. Voet, G. Volckaert, C. Y. Wang, T. R. Ward, J. Wilhelmy, E. A. Winzeler, Y. Yang, G. Yen, E. Youngman, K. Yu, H. Bussey, J. D. Boeke, M. Snyder, P. Philippsen, R. W. Davis and M. Johnston (2002). "Functional profiling of the Saccharomyces cerevisiae genome." Nature 418(6896): 387-391.

Gomez, R., L. E. Goodman, S. K. Tripathy, E. O'Rourke, V. Manne and F. Tamanoi (1993). "Purified yeast protein farnesyltransferase is structurally and functionally similar to its mammalian counterpart." Biochem J 289 (Pt 1)(Pt 1): 25-31.

Greenwood, J., L. Steinman and S. S. Zamvil (2006). "Statin therapy and autoimmune disease: from protein prenylation to immunomodulation." Nat Rev Immunol 6(5): 358-370.

Hildebrandt, E. R., M. Cheng, P. Zhao, J. H. Kim, L. Wells and W. K. Schmidt (2016). "A shunt pathway limits the CaaX processing of Hsp40 Ydj1p and regulates Ydj1p-dependent phenotypes." Elife 5.

Hobbs, G. A., C. J. Der and K. L. Rossman (2016). "RAS isoforms and mutations in cancer at a glance." J Cell Sci 129(7): 1287-1292.

Jeong, A., K. F. Suazo, W. G. Wood, M. D. Distefano and L. Li (2018). "Isoprenoids and protein prenylation: implications in the pathogenesis and therapeutic intervention of Alzheimer's disease." Crit Rev Biochem Mol Biol 53(3): 279-310.

Kohl, N. E., R. E. Diehl, M. D. Schaber, E. Rands, D. D. Soderman, B. He, S. L. Moores, D. L. Pompliano, S. Ferro-Novick, S. Powers and et al. (1991). "Structural homology among mammalian and Saccharomyces cerevisiae isoprenyl-protein transferases." J Biol Chem 266(28): 18884-18888.

Krzysiak, A. J., A. V. Aditya, J. L. Hougland, C. A. Fierke and R. A. Gibbs (2010). "Synthesis and screening of a CaaL peptide library versus FTase reveals a surprising number of substrates." Bioorg Med Chem Lett 20(2): 767-770.

Krzysiak, A. J., S. A. Scott, K. A. Hicks, C. A. Fierke and R. A. Gibbs (2007). "Evaluation of protein farnesyltransferase substrate specificity using synthetic peptide libraries." Bioorg Med Chem Lett 17(20): 5548-5551.

Lane, K. T. and L. S. Beese (2006). "Thematic review series: lipid posttranslational modifications. Structural biology of protein farnesyltransferase and geranylgeranyltransferase type I." J Lipid Res 47(4): 681-699.

Leung, K. F., R. Baron, B. R. Ali, A. I. Magee and M. C. Seabra (2007). "Rab GTPases containing a CAAX motif are processed post-geranylgeranylation by proteolysis and methylation." J Biol Chem 282(2): 1487-1497.

London, N., C. L. Lamphear, J. L. Hougland, C. A. Fierke and O. Schueler-Furman (2011). "Identification of a novel class of farnesylation targets by structure-based modeling of binding specificity." PLoS Comput Biol 7(10): e1002170.

Long, S. B., P. J. Casey and L. S. Beese (2002). "Reaction path of protein farnesyltransferase at atomic resolution." Nature 419(6907): 645-650.

Maurer-Stroh, S. and F. Eisenhaber (2005). "Refinement and prediction of protein prenylation motifs." Genome Biol 6(6): R55.

Maurer-Stroh, S., M. Koranda, W. Benetka, G. Schneider, F. L. Sirota and F. Eisenhaber (2007). "Towards complete sets of farnesylated and geranylgeranylated proteins." PLoS Comput Biol 3(4): e66.

Moores, S. L., M. D. Schaber, S. D. Mosser, E. Rands, M. B. O'Hara, V. M. Garsky, M. S. Marshall, D. L. Pompliano and J. B. Gibbs (1991). "Sequence dependence of protein isoprenylation." J Biol Chem 266(22): 14603-14610.

Oldenburg, K. R., K. T. Vo, S. Michaelis and C. Paddon (1997). "Recombination-mediated PCR-directed plasmid construction in vivo in yeast." Nucleic Acids Res 25(2): 451-452.

Omer, C. A., A. M. Kral, R. E. Diehl, G. C. Prendergast, S. Powers, C. M. Allen, J. B. Gibbs and N. E. Kohl (1993). "Characterization of recombinant human farnesyl-protein transferase: cloning, expression, farnesyl diphosphate binding, and functional homology with yeast prenyl-protein transferases." Biochemistry 32(19): 5167-5176.

Palsuledesai, C. C. and M. D. Distefano (2015). "Protein prenylation: enzymes, therapeutics, and biotechnology applications." ACS Chem Biol 10(1): 51-62.

Ravishankar, R., E. R. Hildebrandt, G. Greenway, N. Asad, S. Gore, T. M. Dore and W. K. Schmidt (2023). "Specific Disruption of Ras2 CAAX Proteolysis Alters Its Localization and Function." Microbiol Spectr: e0269222.

Reid, T. S., K. L. Terry, P. J. Casey and L. S. Beese (2004). "Crystallographic analysis of CaaX prenyltransferases complexed with substrates defines rules of protein substrate selectivity." J Mol Biol 343(2): 417-433.

Reiss, Y., S. J. Stradley, L. M. Gierasch, M. S. Brown and J. L. Goldstein (1991). "Sequence requirement for peptide recognition by rat brain p21ras protein farnesyltransferase." Proc Natl Acad Sci U S A 88(3): 732-736.

Sapkota, G. P., A. Kieloch, J. M. Lizcano, S. Lain, J. S. Arthur, M. R. Williams, N. Morrice, M. Deak and D. R. Alessi (2001). "Phosphorylation of the protein kinase mutated in Peutz-Jeghers cancer syndrome, LKB1/STK11, at Ser431 by p90(RSK) and cAMP-dependent protein kinase, but not its farnesylation at Cys(433), is essential for LKB1 to suppress cell vrowth." J Biol Chem 276(22): 19469-19482.

Schey, G. L., P. H. Buttery, E. R. Hildebrandt, S. X. Novak, W. K. Schmidt, J. L. Hougland and M. D. Distefano (2021). "MALDI-MS Analysis of Peptide Libraries Expands the Scope of Substrates for Farnesyltransferase." Int J Mol Sci 22(21).

Sikorski, R. S. and P. Hieter (1989). "A system of shuttle vectors and yeast host strains designed for efficient manipulation of DNA in Saccharomyces cerevisiae." Genetics 122(1): 19-27.

Stein, V., M. H. Kubala, J. Steen, S. M. Grimmond and K. Alexandrov (2015). "Towards the systematic mapping and engineering of the protein prenylation machinery in Saccharomyces cerevisiae." PLoS One 10(3): e0120716.

Storck, E. M., J. Morales-Sanfrutos, R. A. Serwa, N. Panyain, T. Lanyon-Hogg, T. Tolmachova, L. N. Ventimiglia, J. Martin-Serrano, M. C. Seabra, B. Wojciak-Stothard and E. W. Tate (2019). "Dual chemical probes enable quantitative system-wide analysis of protein prenylation and prenylation dynamics." Nat Chem 11(6): 552-561.

Suazo, K. F., C. Schaber, C. C. Palsuledesai, A. R. Odom John and M. D. Distefano (2016). "Global proteomic analysis of prenylated proteins in Plasmodium falciparum using an alkyne-modified isoprenoid analogue." Sci Rep 6: 38615.

Tamanoi, F. (2011). "Ras signaling in yeast." Genes Cancer 2(3): 210-215.

Tate, E. W., K. A. Kalesh, T. Lanyon-Hogg, E. M. Storck and E. Thinon (2015). "Global profiling of protein lipidation using chemical proteomic technologies." Curr Opin Chem Biol 24: 48-57.

Trueblood, C. E., V. L. Boyartchuk and J. Rine (1997). "Substrate specificity determinants in the farnesyltransferase beta-subunit." Proc Natl Acad Sci U S A 94(20): 10774-10779.

Wang, M. and P. J. Casey (2016). "Protein prenylation: unique fats make their mark on biology." Nat Rev Mol Cell Biol 17(2): 110-122.

Wang, Y. C., J. K. Dozier, L. S. Beese and M. D. Distefano (2014). "Rapid analysis of protein farnesyltransferase substrate specificity using peptide libraries and isoprenoid diphosphate analogues." ACS Chem Biol 9(8): 1726-1735.

Willumsen, B. M., A. Christensen, N. L. Hubbert, A. G. Papageorge and D. R. Lowy (1984). "The p21 ras C-terminus is required for transformation and membrane association." Nature 310(5978): 583-586.

Wright, L. P. and M. R. Philips (2006). "Thematic review series: lipid posttranslational modifications. CAAX modification and membrane targeting of Ras." J Lipid Res 47(5): 883-891.

Zhou, M., H. Wiener, W. Su, Y. Zhou, C. Liot, I. Ahearn, J. F. Hancock and M. R. Philips (2016). "VPS35 binds farnesylated N-Ras in the cytosol to regulate N-Ras trafficking." J Cell Biol 214(4): 445-458.

Tables

Table 1. *Yeast strains used in this study.*

Identifier	Genotype	Reference
BY4741	*MAT***a** *his3Δ1 leu2Δ0 met15Δ0 ura3Δ0*	(Brachmann, Davies et al. 1998)
yWS304	*MAT***a** *his3Δ1 leu2Δ0 met15Δ0 ura3Δ0 ydj1Δ::KAN*^{R}	(Giaever, Chu et al. 2002)
yWS1632	*MAT***a** *his3Δ1 leu2Δ0 met15Δ0 ura3Δ0 ram1::KAN*^{R}	(Giaever, Chu et al. 2002)
yWS2542	*MAT***a** *his3Δ1 leu2Δ0 met15Δ0 ura3Δ0 ydj1Δ::NAT*^{R} *ram1Δ::KAN*^{R}	(Berger, Kim et al. 2018)
yWS2544	*MAT***a** *his3Δ1 leu2Δ0 met15Δ0 ura3Δ0 ydj1Δ::NAT*^{R}	(Berger, Kim et al. 2018)

Table 2. *Plasmids used in this study.*

Identifier	Genotype	Reference
pRS316	*CEN URA3*	(Sikorski and Hieter 1989)
pWS942	*CEN URA3 YDJ1*	(Hildebrandt, Cheng et al. 2016)
pWS1132	*CEN URA3 YDJ1-SASQ*	(Hildebrandt, Cheng et al. 2016)
pWS1286	*CEN URA3 YDJ1-CVIA*	(Hildebrandt, Cheng et al. 2016)
pWS1318	*CEN URA3 NAP1*	This study
pWS1407	*CEN URA3 YDJ1-CKQQ*	This study
pWS1411	*CEN URA3 YDJ1-CKQS*	(Berger, Yeung et al. 2022)
pWS1474	*CEN URA3 His8-NAP1*	This study
pWS1563	*CEN URA3 YDJ1-SKQQ*	This study
pWS1728	*CEN URA3 YDJ1-BsrGI-SASQ*	This study
pWS2083	*CEN URA3 YDJ1-CKQA*	This study
pWS2017	*CEN URA3 YDJ1-CKQC*	This study
pWS2018	*CEN URA3 YDJ1-CKQD*	This study
pWS2019	*CEN URA3 YDJ1-CKQE*	This study
pWS2020	*CEN URA3 YDJ1-CKQF*	This study
pWS2021	*CEN URA3 YDJ1-CKQG*	(Berger, Yeung et al. 2022)
pWS2022	*CEN URA3 YDJ1-CKQH*	(Berger, Yeung et al. 2022)
pWS2023	*CEN URA3 YDJ1-CKQI*	This study
pWS2024	*CEN URA3 YDJ1-CKQK*	This study
pWS2025	*CEN URA3 YDJ1-CKQL*	(Berger, Yeung et al. 2022)

pWS2026	*CEN URA3 YDJ1-CKQM*	This study
pWS2027	*CEN URA3 YDJ1-CKQN*	This study
pWS2028	*CEN URA3 YDJ1-CKQP*	This study
pWS2029	*CEN URA3 YDJ1-CKQR*	This study
pWS2030	*CEN URA3 YDJ1-CKQT*	This study
pWS2031	*CEN URA3 YDJ1-CKQV*	This study
pWS2033	*CEN URA3 YDJ1-CKQW*	This study

Table 3. *A comprehensive overview of in-house prediction versus published models.*

SGD Protein	Motif (n=48)	Observed[1]	%[1]	PrePS[2]	SVM[3]	RRS EF[4]	YDJ1 HM
Ras2	CIIS	+	100	1.281	1	9.011500	4.07
Hmg1	CIKS	-	41	-3.588	0	0.034717	1.93
Rho2	CIIL	+	100	0.817	1	10.313800	3.65
Ssp2	CIDL	-	0	-4.893	0	0.254070	1.82
Skt5, MiY1	CVIM	+	100	2.200	1	3.371400	3.66
Tbs1	CVKM	+	100	-2.909	0	0.066541	2.47
YDL022C-A	CSII	+	100	0.987	1	11.350000	3.34
YBR096W	CSEI	-	0	-5.759	0	0.149720	1.88
YMR265C	CSNA	+	100	-4.848	0	0.039925	3.84
Pet18	CYNA	+	100	-5.843	0	0.079849	4.95
Lih1	CSGL	-	61	-4.577	0	0.041301	2.79
Cup1	CSGK	-	59	-6.461	0	0.099811	1.75
Nap1	CKQS	+	100	-0.975	1	0.066541	4.97
Cst26	CFIF	+	100	-1.924	1	3.194000	4.81
YIL134C-A	CAPY	+	100	-2.164	1	0.199620	4.73
Atr1	CTVA	+	100	0.273	1	7.775800	4.55
Las21	CALD	+	100	-2.887	1	0.028518	4.20
YDL009C	CAVS	+	100	0.237	1	6.607500	4.69
Sua5	CIQF	+	100	-0.601	1	0.199620	5.49
NA	CAAQ	+	100	-3.280	1	0.399250	4.82
NA	CAHQ	+	100	-3.599	1	0.399250	4.74
NA	CASA	+	100	-2.693	1	0.243980	3.91
NA	CKQH	+	70	-2.442	1	0.399250	4.37
NA	CNLI	+	95	-3.804	1	0.133080	3.84
NA	CSFL	+	68	-4.425	1	0.108890	4.28
NA	CVAA	+	100	-2.789	1	0.079849	4.81
NA	CVFM	+	100	-2.939	1	2.395500	4.76
NA	CKQG	+	73	-1.876	0	0.099811	4.06

NA	CKQL	+	50	-1.680	0	0.057035	3.84
NA	CQTS	+	100	-0.922	0	0.057035	5.44
NA	CQSQ	+	100	-1.624	1	0.159700	5.27
NA	CKQA	+	95	-1.149	1	0.133080	4.57
NA	CKQC	+	100	-0.745	1	0.099811	4.12
NA	CKQD	+	67	-4.103	0	0.066541	3.51
NA	CKQE	-	19	-3.745	0	0.066541	3.36
NA	CKQF	-	28	-2.345	0	0.057035	4.62
NA	CKQH	+	73	-2.442	1	0.399250	4.37
NA	CKQI	+	70	-1.700	1	0.099811	3.90
NA	CKQK	-	0	-4.010	0	0.036295	2.22
NA	CKQM	+	100	-1.045	1	0.133080	4.50
NA	CKQN	+	88	-5.046	0	0.199620	4.48
NA	CKQP	-	13	-2.887	0	0.399250	1.97
NA	CKQQ	+	100	-0.747	1	0.399250	5.35
NA	CKQR	-	0	-4.533	0	0.399250	1.77
NA	CKQT	+	96	-1.874	1	0.079849	4.68
NA	CKQV	+	100	-1.744	1	0.133080	4.50
NA	CKQW	+	71	-5.184	0	0.199620	3.57
NA	CKQY	+	100	-3.326	0	0.399250	4.50
			# predicted correctly	30	38	17	45
			# false positive	0	0	0	2
			# false negative	18	10	31	1
			% predicted correctly	62.5%	79.2%	35.4%	93.8%
			% false positive	0.0%	0.0%	0.0%	5.1%
			% false negative	64.3%	50.0%	75.6%	11.1%

[1] Prenylation status and % gel-shift observed in context of Ydj1

[2] Values reported from (Maurer-Stroh and Eisenhaber 2005); positive cutoff > -2.

[3] Values reported from (Berger, Yeung et al. 2022); positive cutoff > 0.

[4] Values reported from (Stein, Kubala et al. 2015); positive cutoff > 3.

Figure legends

Figure 1. *The post-translational isoprenylation pathway.* Ydj1 and Lkb1/STK11 have non-canonical CaaX motifs that follow an isoprenylation-only pathway (i.e., shunted). Ras and **a**-factor have canonical CaaX motifs and are subject to the superposition of 3 modifications.

Figure S1. *Experimental scheme for genetic screening of Ydj1-CXXX variants assessed via next-generation sequencing (NGS).* A plasmid library encoding all 8000 Ydj1-CXXX variants was generated via ligation-based cloning method using PCR fragments derived using Trimer20 oligonucleotides (Integrated DNA Technologies, Newark, NJ). The library DNA was purified from *E. coli*, and plasmids were transformed into the *ydj1Δ* yeast strain. This strain cannot grow at temperatures >37 °C unless complemented by farnesylated Ydj1-CXXX variant. Populations of *ydj1Δ/YDJ1-CXXX* transformants were incubated at permissive (25 °C) and selective (37 °C) temperatures, and plasmids recovered were subject to NGS analysis. Schematic illustration created with Microsoft PowerPoint and BioRender.com.

Figure S2. *Frequencies of CXXX sequences in naïve yeast library versus E. coli plasmid library.* CXXX frequencies from the naïve yeast and *E. coli* plasmid libraries were plotted against each other. No significant change in the frequency profile of individual sequences was observed between the two libraries. Libraries were not perfectly balanced, however, across all CXXX sequences; the extremes exhibited a ~10x range in relative abundance.

Figure 2. *Comparison of results from the Ydj1-CXXX NGS screen and previously published Ras-CXXX NGS screen.* Enrichment profiles of (**A**) Ydj1-based and (**B**) Ras-based screens offer two distinct sequence landscapes. Annotated controls: non-canonical *Sc*Ydj1 (CASQ) and *Hs*Lkb1/STK11 (CKQQ); canonical *Sc* **a**-factor (CVIA) and *Hs*H-Ras (CVLS). Weblogo frequency representation of top 5% (n=400) of hits for (**C**) Ydj1-based and (**D**) Ras-based screens. A 4D representation of the space occupied by top 5% of hits recovered from (**E**) Ydj1-based and (**F**) Ras-based screens. The size of the dot is the 4th dimension and corresponds to the enrichment factor of the CXXX motif within the respective screen.

Figure S3. *3D analysis of top 5% of sequences from Ydj1-based and Ras-based data sets.* The data described in **Figures 2E, F** was plotted in 3D from the perspective of each axis of the 4D plot such that Lysine (K) is the nearest amino acid. The size of the spot is the 3rd dimension and represents relative abundance. The perspectives are along the a_1 axis (**A**, **B**), the a_2 axis (**C**, **D**), and X axis (**E**, **F**) for the Ydj1-based data (**A**, **C**, **E**) and the Ras-based data (**B**, **D**, **F**).

Figure 3. *Relative performance of previously identified farnesylated CXXX sequences.* Sequences previously identified in a (**A**) Ydj1-based genetic selection or a (**B**) Ras-based competitive growth assay were superimposed on the Ydj1 EF profile. Each sequence is represented as a point with an associated score where the second Y-axis refers to the scoring range used in these previous studies (YTS T-score - Ydj1-based

Temperature Screen Thermotolerance Score; RRS EF – Ras Recruitment System Enrichment Factor). The limited number of canonical sequences that were identified by YTS are identified by a dashed box (n=15).

Figure 4. *Evaluation of CKQX variants as FTase substrates.* Enrichment profiles of CKQX sequences in (**A**) Ydj1-based and (**B**) Ras-based screens. **C**) Thermotolerance and gel-shift assays of Ydj1-CKQX variants provide supporting evidence towards the prenylation of non-canonical CKQX sequences by yeast FTase. Thermotolerance assays were performed by culturing yeast to the same density, applying a 10x serial dilution to the cultures, spotting dilution sets onto YPD rich media, and incubating at indicated temperatures. Notably, plate-based selection requires a slightly higher temperature to mimic the growth profile observed on liquid-based selection. The strains used were transformants of yWS304 containing the indicated Ydj1-CXXX variant. Gel shift assays were performed for the indicated Ydj1-CXXX variants in the presence and absence of FTase activity. Total yeast extracts were analyzed by SDS-PAGE and immunoblotting. The strains used were yWS2544 (+FTase) and yWS2542 (*ram1*Δ; -FTase. Abbreviations: Seq = Sequence, EF = Enrichment Factor, % = percent prenylation.

Figure S4. *Thermotolerance assay of control sequences.* **A**) Gel-shift assays were performed for Ydj1 (CASQ) and His8-Nap1 (CKQS) in the presence and absence of FTase activity. Total yeast extracts were analyzed by SDS-PAGE and immunoblotting. The strains used were BY4741 (+FTase) and yWS1632 (*ram1*Δ; -FTase). **B**)

Thermotolerance assays were performed with the indicated Ydj1-CXXX variants as described for **Figure 4**. Abbreviations: Seq = Sequence, EF = Enrichment Factor, % = percent prenylation, N/A = not applicable.

Figure 5. *Heatmap analysis reveals contextual determinants of FTase specificity.* Ydj1-based comparisons of **A**) a_1 vs. a_2, **B**) a_1 vs. X, and **C**) a_2 vs. X positions, and Ras-based comparisons of **D**) a_1 vs. a_2, **E**) a_1 vs. X, and **F**) a_2 vs. X positions. Values were generated by averaging the 20 EFs of each CXXX combinations. Color scheme: red (low likelihood of farnesylation) to green (high likelihood of farnesylation).

Figure 6. *A simplified decision tree for predicting yeast FTase substrates.* The output of Ydj1 EF scores was fit into a decision tree model considering each of the variable CaaX motif sites: a_1, a_2, and X. Our results strongly suggest that a major determinant of FTase activity depends on a select few discriminatory amino acids at a_2 (D, E, K, R) and X (K, P, R) positions, with the potential for certain allowable amino acids at a_2 (I, L, M, V) that can mitigate discrimination caused at the X position. For the full decision tree, see **Figure S5**.

Figure S5. *Full decision tree model. S*equence motif features were generated by one-hot encoding each of the variable CaaX motif sites. Additional binary features were included that describe whether the variable residue was within a given set of residues. To define these rules, all possible residue combinations up to 5 amino acids were considered, yielding a total of 65,097 features. While building the decision tree

classifier, entropy was used to evaluate the quality of potential splits. Trees were allowed a maximum depth of 3 in order to determine more generalizable rules and all nodes were required to have a minimum of 50 samples. Samples refer to the number of CXXX variants from the previous node that fall within the category. Value refers to the distribution of CXXX variants that fall within specified EF ranges (0-1, 1-2, 2-3, 3-4, >4).

File S1. *Frequency of CXXX sequences observed in naive libraries*

File S2. *Enrichment factor results of Ydj1-based NGS*

File S3. *CKQX hits identified on UniProt*

File S4. *Heatmap analysis*

Figure 1

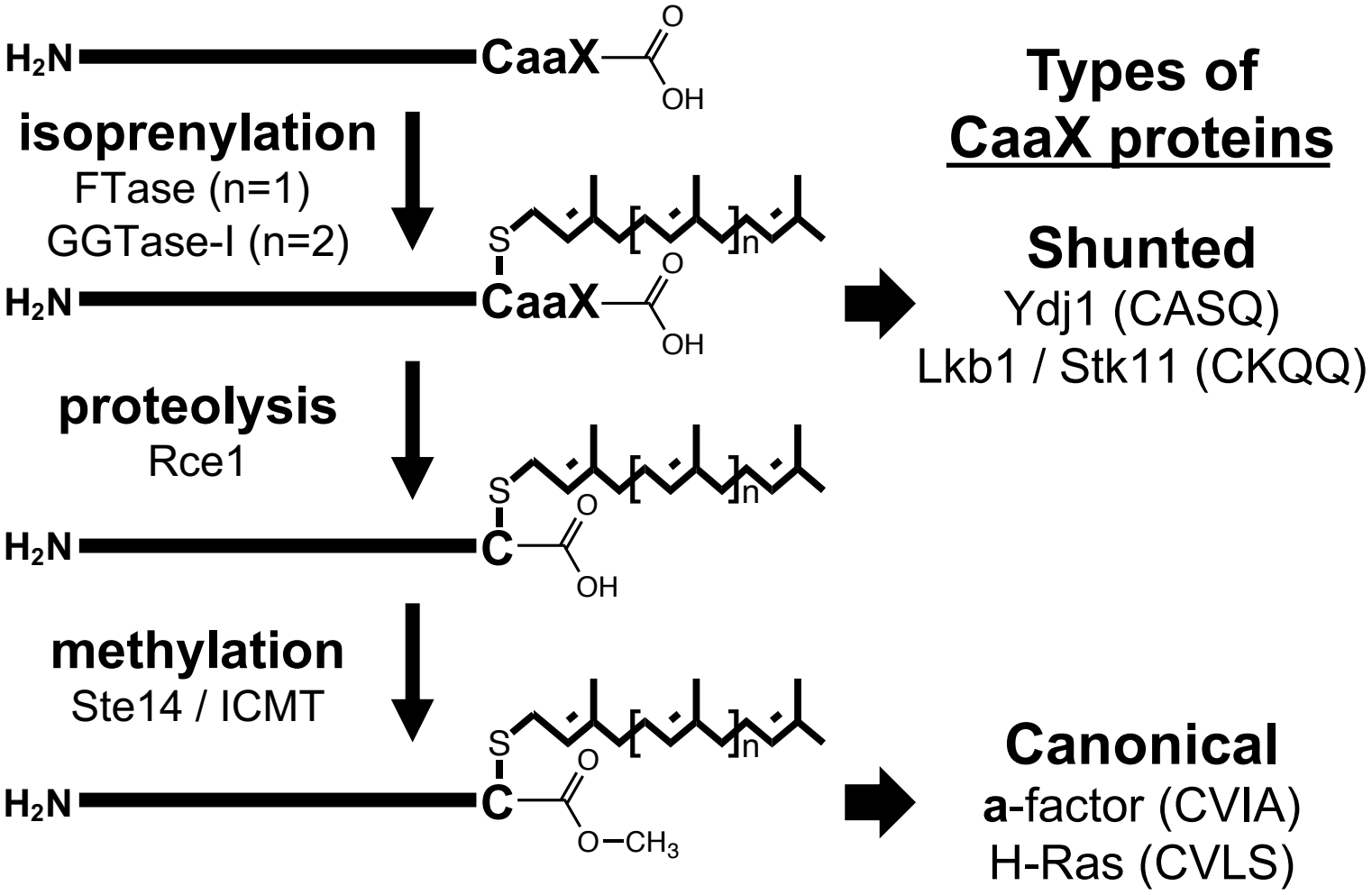

H2N
CaaX
O
OH
isoprenylation
FTase (n=1)
GGTase-I (n=2)
S
n
proteolysis
Rce1
C
methylation
Ste14 / ICMT
O−CH3
Types of
CaaX proteins
Shunted
Ydj1 (CASQ)
Lkb1 / Stk11 (CKQQ)
Canonical
a-factor (CVIA)
H-Ras (CVLS)

Figure S1

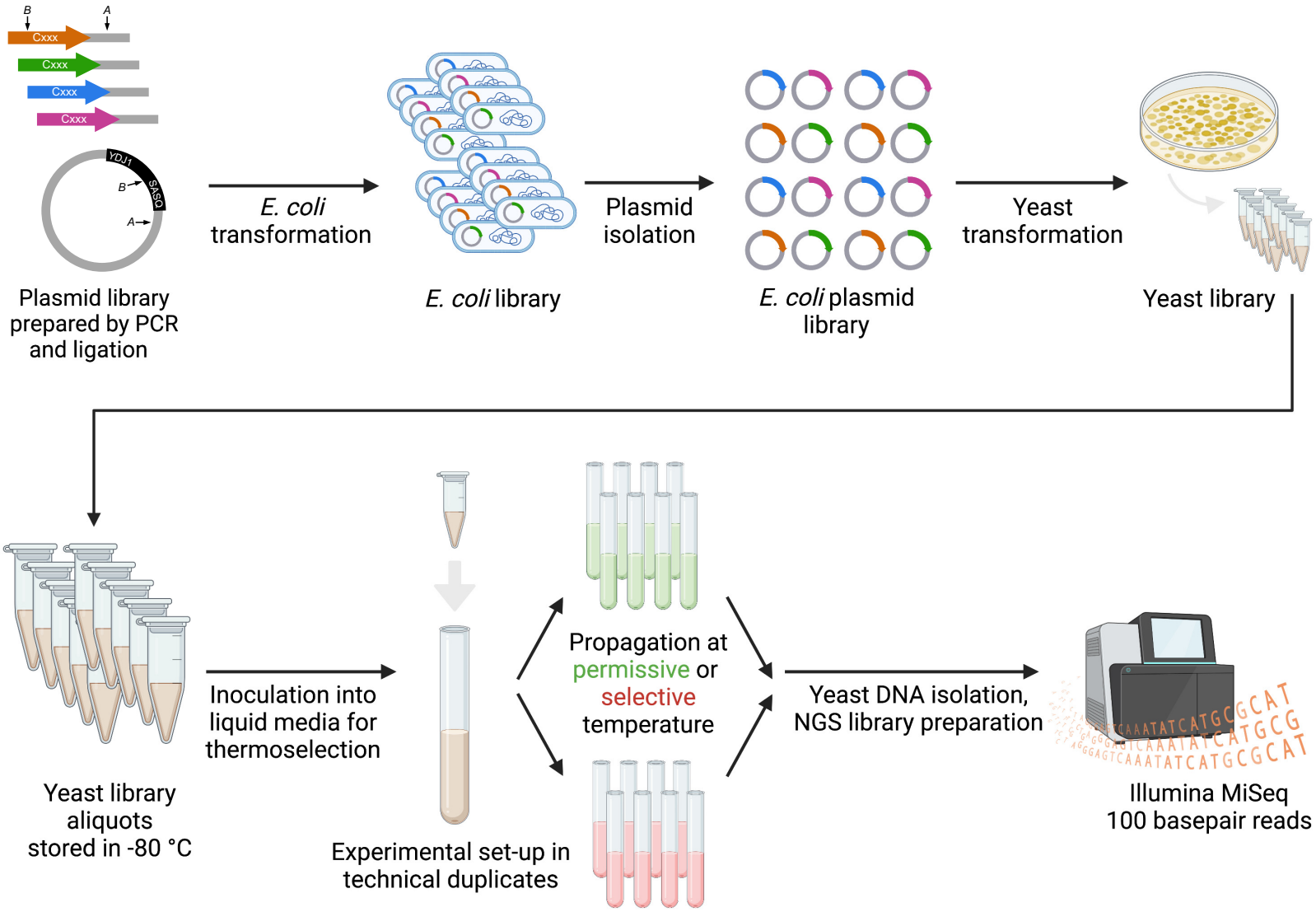

B
A
Cxxx
Cxxx
Cxxx
Cxxx
YDJ1
SASQ
B
A
Plasmid library prepared by PCR and ligation
E. coli transformation
E. coli library
Plasmid isolation
E. coli plasmid library
Yeast transformation
Yeast library
Yeast library aliquots stored in -80 °C
Inoculation into liquid media for thermoselection
Experimental set-up in technical duplicates
Propagation at permissive or selective temperature
Yeast DNA isolation, NGS library preparation
Illumina MiSeq 100 basepair reads

Figure S2

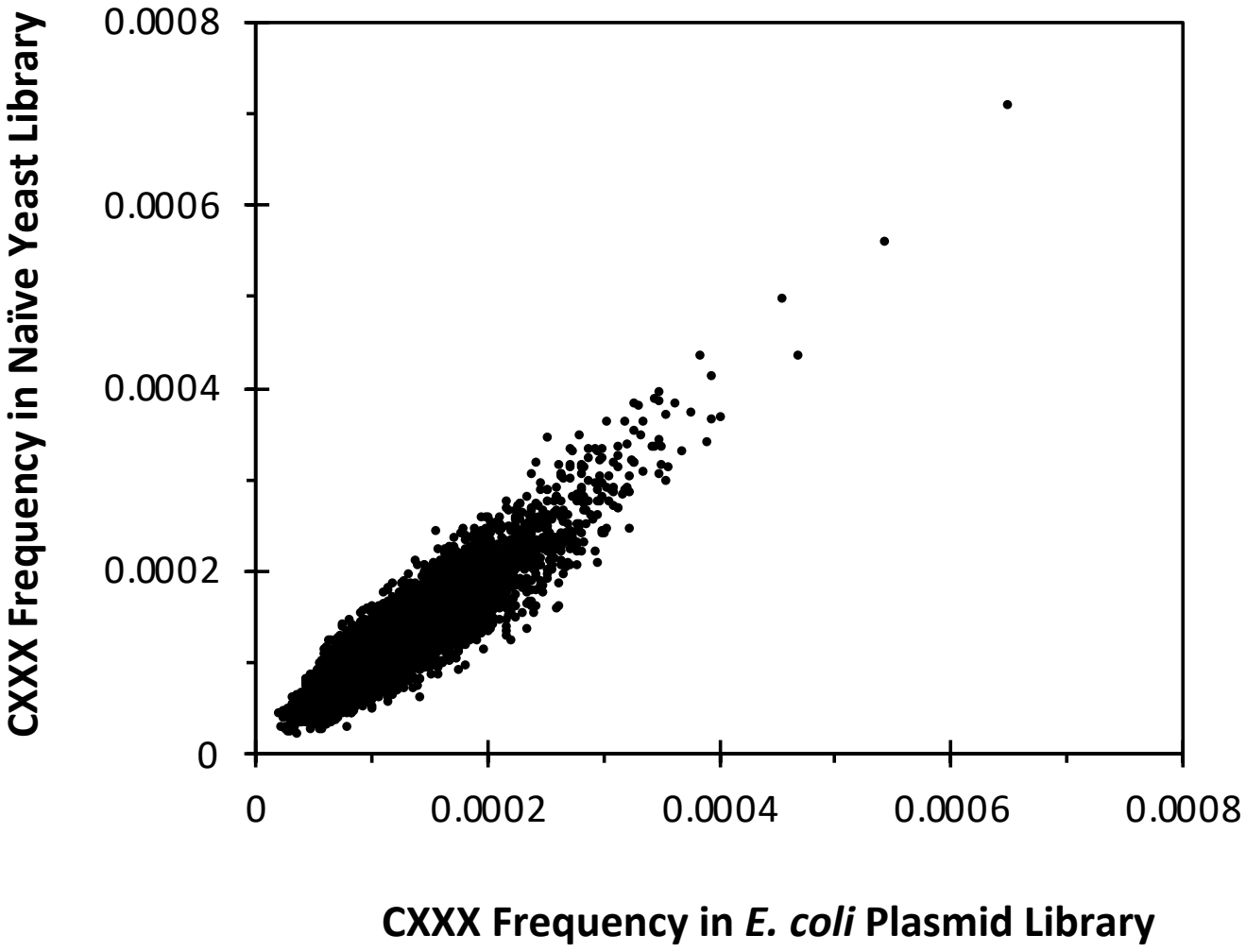

0.0008
0.0006
0.0004
0.0002
0
CXXX Frequency in Naïve Yeast Library
0
0.0002
0.0004
0.0006
0.0008
CXXX Frequency in *E. coli* Plasmid Library

Figure 2

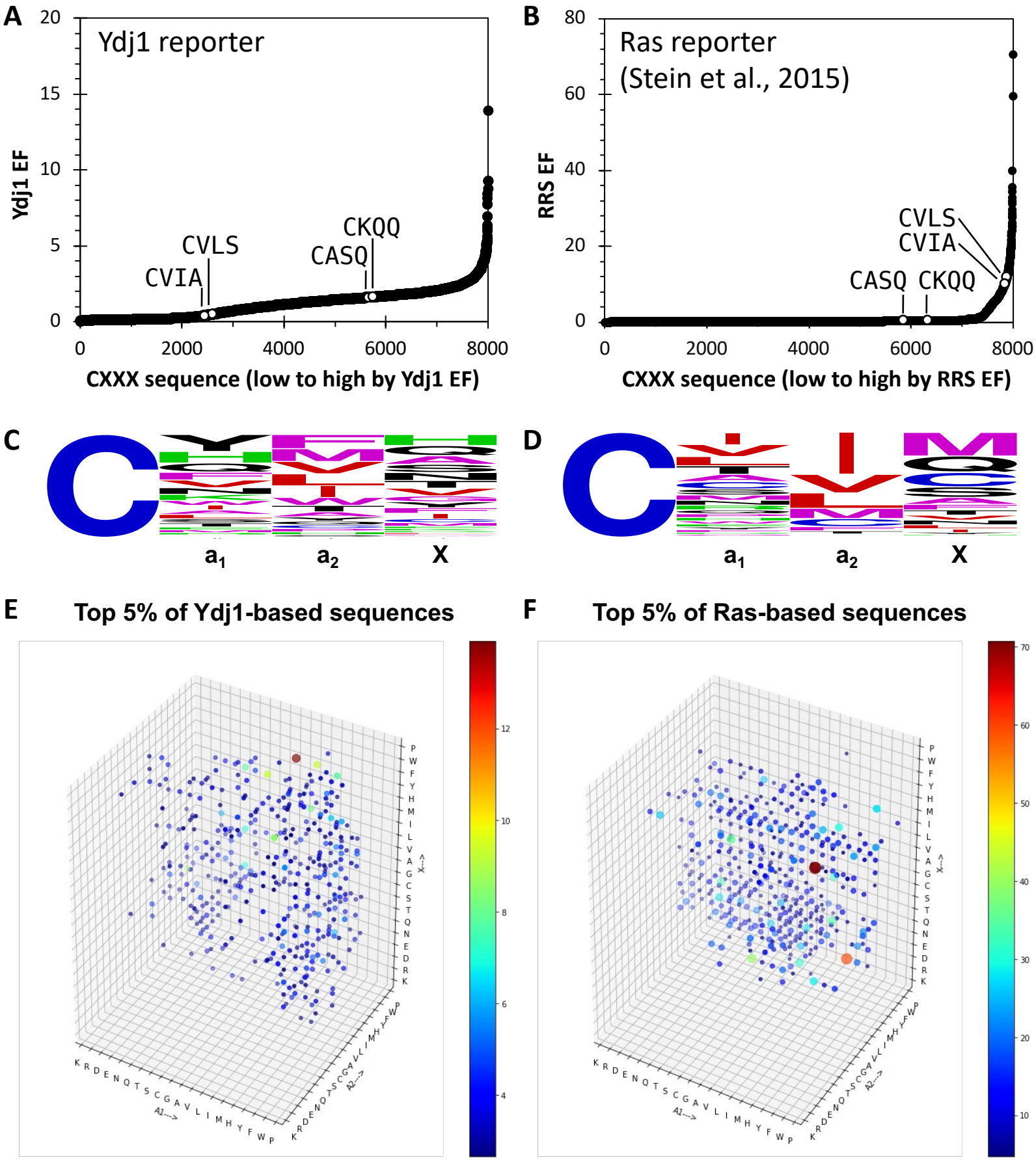

A
Ydj1 reporter
Ydj1 EF
CVIA
CVLS
CASQ
CKQQ
CXXX sequence (low to high by Ydj1 EF)
B
Ras reporter
(Stein et al., 2015)
RRS EF
CVLS
CVIA
CASQ
CKQQ
CXXX sequence (low to high by RRS EF)
C
a1
a2
X
D
a1
a2
X
E
Top 5% of Ydj1-based sequences
F
Top 5% of Ras-based sequences

Figure S3

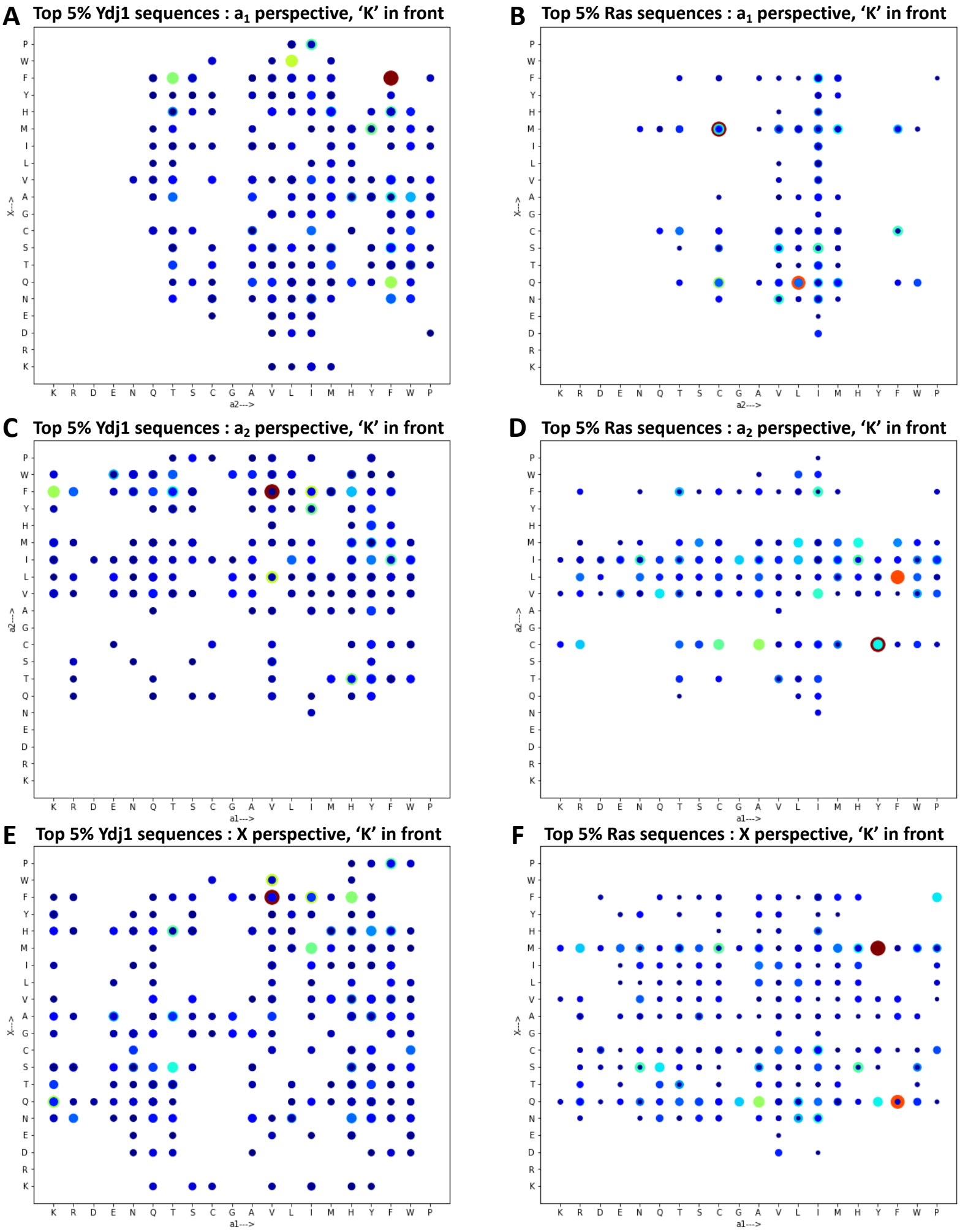

A Top 5% Ydj1 sequences : a1 perspective, 'K' in front
B Top 5% Ras sequences : a1 perspective, 'K' in front
C Top 5% Ydj1 sequences : a2 perspective, 'K' in front
D Top 5% Ras sequences : a2 perspective, 'K' in front
E Top 5% Ydj1 sequences : X perspective, 'K' in front
F Top 5% Ras sequences : X perspective, 'K' in front
X--->
a2--->
a1--->
P W F Y H M I L V A G C S T Q N E D R K
K R D E N Q T S C G A V L I M H Y F W P

Figure 3

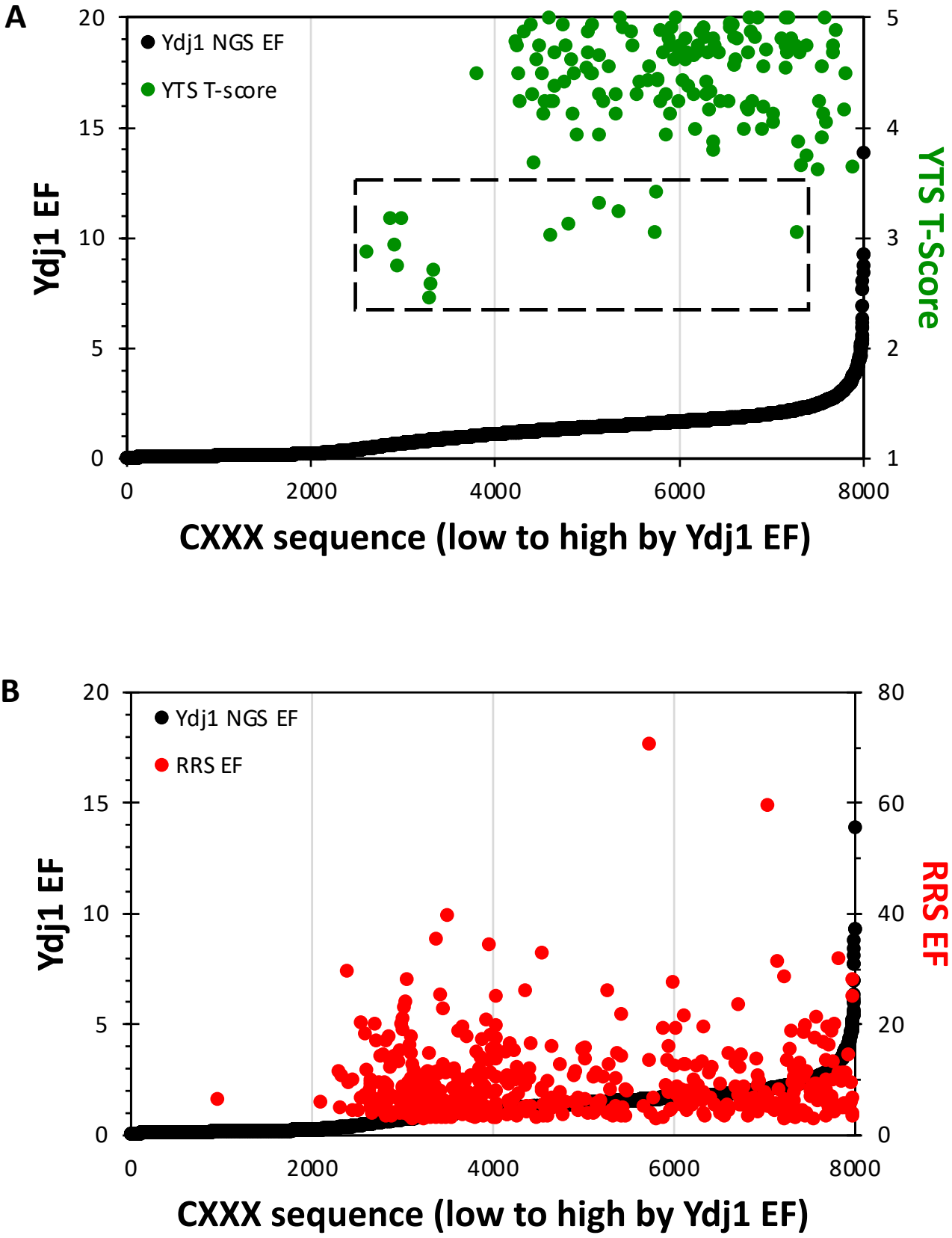
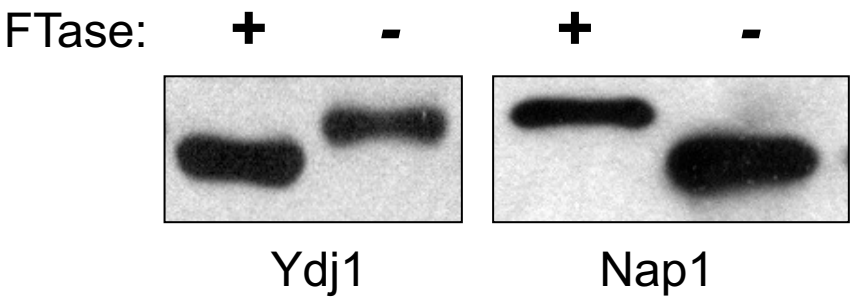

A
Ydj1 NGS EF
YTS T-score
Ydj1 EF
YTS T-Score
CXXX sequence (low to high by Ydj1 EF)
0
5
10
15
20
1
2
3
4
5
0
2000
4000
6000
8000
B
Ydj1 NGS EF
RRS EF
Ydj1 EF
RRS EF
CXXX sequence (low to high by Ydj1 EF)
0
5
10
15
20
0
20
40
60
80
0
2000
4000
6000
8000

Figure S4

A


FTase:
+
-
+
-
Ydj1
Nap1

B

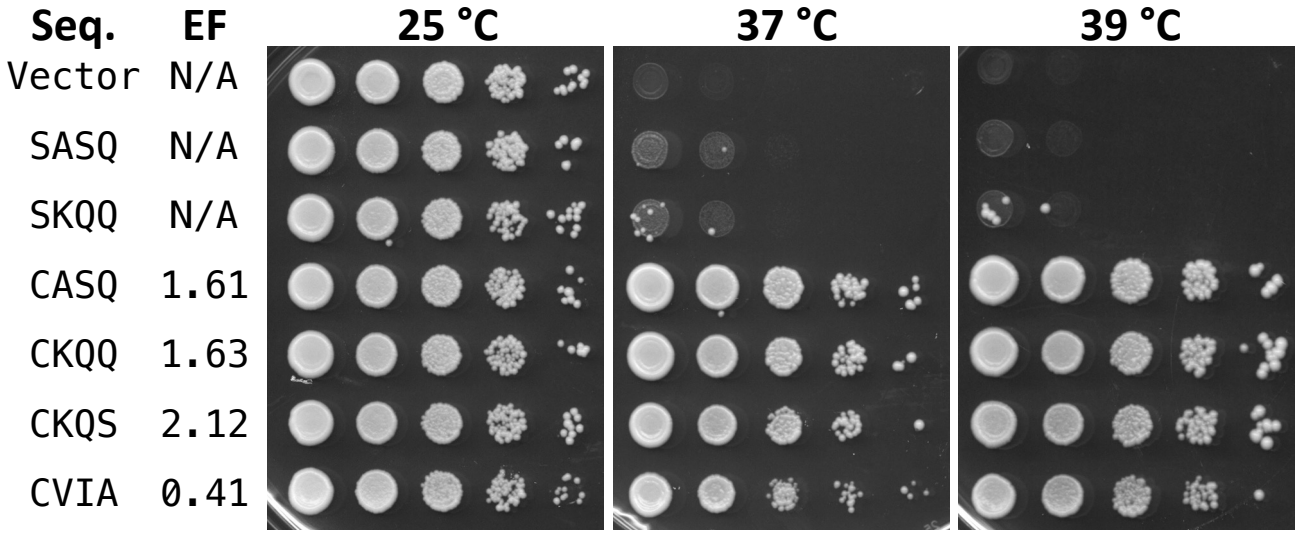
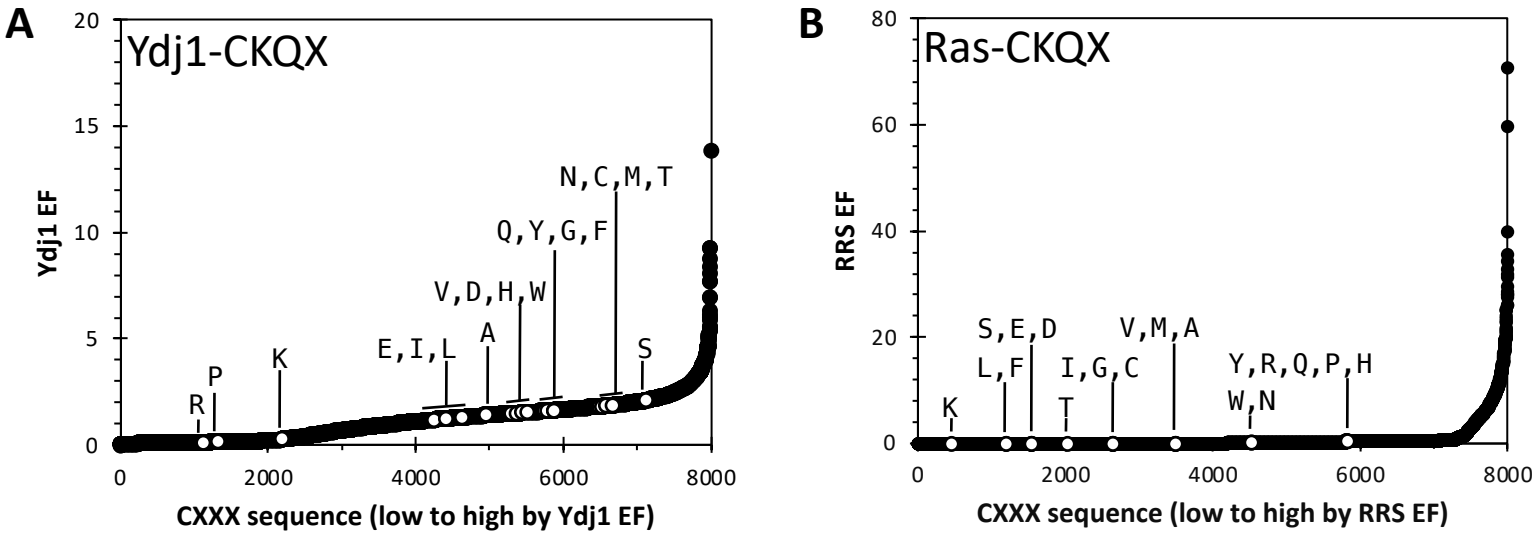
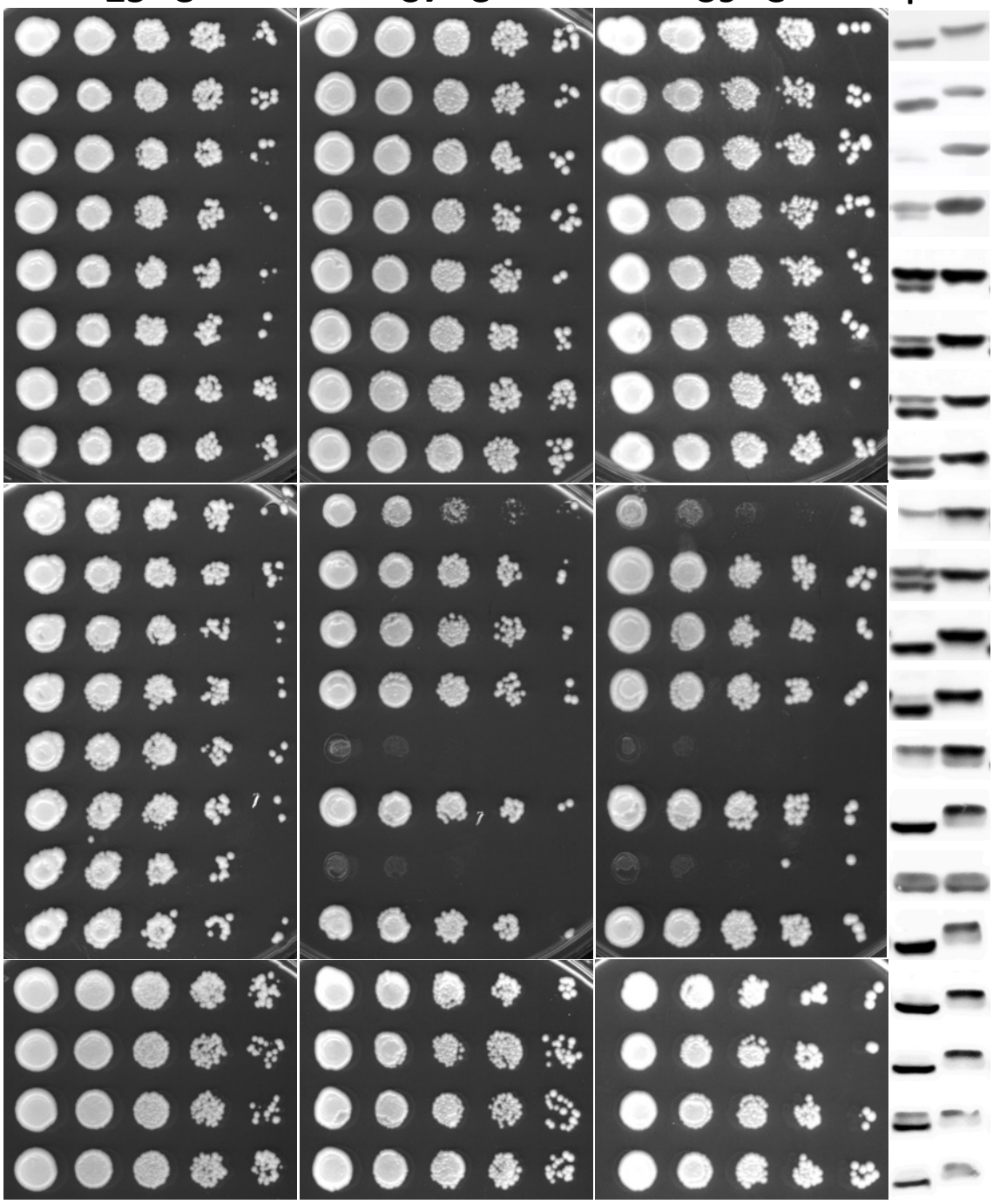

Seq.
EF
25 °C
37 °C
39 °C
Vector N/A
SASQ N/A
SKQQ N/A
CASQ 1.61
CKQQ 1.63
CKQS 2.12
CVIA 0.41

Figure 4



C

Seq.	EF
CKQA	1.43
CKQC	1.86
CKQD	1.54
CKQE	1.22
CKQF	1.67
CKQG	1.67
CKQH	1.55
CKQI	1.27
CKQK	0.30
CKQL	1.34
CKQM	1.87
CKQN	1.85
CKQP	0.17
CKQQ	1.63
CKQR	0.16
CKQS	2.12
CKQT	1.91
CKQV	1.52
CKQW	1.57
CKQY	1.64



%
95
100
67
19
28
73
70
62
0
50
100
88
13
100
0
100
96
100
71
100

Figure 5

Ydj1 Results

RRS Results

A (a_2 × a_1)

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	1.40	1.03	0.41	0.68	1.82	1.17	1.38	1.43	0.30	1.34	1.63	1.23	1.54	1.44	0.26	1.27	1.48	1.43	1.54	1.55
C	1.09	0.92	0.48	0.82	0.86	0.36	0.84	1.09	0.35	0.94	1.23	0.77	1.19	1.17	0.21	0.90	1.22	1.09	0.46	0.88
D	0.67	0.59	0.16	0.18	1.00	0.28	0.48	1.57	0.14	1.22	1.29	0.32	0.72	0.54	0.16	0.40	0.83	1.46	0.53	0.63
E	1.14	0.86	0.19	0.33	1.57	0.73	1.21	1.88	0.17	1.76	1.61	0.86	1.05	1.09	0.19	1.00	1.28	1.88	1.53	1.28
F	1.67	1.34	0.28	0.73	1.84	0.76	1.22	1.99	0.64	2.00	1.98	1.07	1.40	1.32	0.89	1.20	1.88	1.50	1.39	1.27
G	1.00	0.97	0.14	0.29	1.24	0.64	0.75	1.35	0.15	1.32	1.21	0.56	1.01	1.00	0.17	0.84	1.19	1.47	1.16	1.08
H	1.54	1.15	0.27	0.56	2.49	0.84	1.38	1.99	0.20	1.89	2.42	1.12	1.49	1.64	0.33	1.42	2.17	2.07	1.47	1.66
I	1.65	0.77	0.37	0.43	1.91	0.89	1.51	1.26	0.27	1.73	1.67	1.19	1.61	1.54	0.16	1.26	1.16	1.50	1.19	1.86
K	1.21	0.79	0.23	0.47	1.84	0.85	1.09	1.65	0.17	1.34	2.18	1.03	1.27	1.41	0.15	1.16	1.53	1.98	1.30	1.04
L	1.43	0.67	0.32	0.49	1.41	0.76	1.30	1.75	0.28	1.50	1.84	1.11	1.45	1.32	0.16	1.15	1.13	1.83	1.51	1.50
M	1.35	1.03	0.24	0.48	1.69	0.76	1.32	1.23	0.14	1.60	1.42	0.90	1.30	1.44	0.20	1.19	1.59	1.40	1.15	1.28
N	1.20	0.85	0.23	0.41	1.83	0.95	1.27	1.83	0.16	1.40	1.63	0.93	1.29	1.32	0.17	1.32	1.36	1.97	1.90	1.28
P	1.06	0.86	0.25	0.46	1.21	0.87	1.38	1.13	0.20	1.13	1.10	1.08	1.14	1.38	0.15	1.28	1.39	1.11	1.10	1.10
Q	1.50	0.88	0.25	0.45	1.95	1.04	1.55	2.12	0.15	2.18	1.98	1.23	1.62	1.64	0.19	1.45	1.66	2.26	1.63	1.68
R	1.02	0.90	0.35	0.46	1.65	0.84	1.07	1.06	0.18	1.25	1.43	1.12	1.26	1.48	0.15	1.29	1.48	1.26	1.23	1.14
S	1.27	0.73	0.23	0.41	1.87	1.00	1.14	1.34	0.16	1.27	1.77	1.04	1.48	1.43	0.18	1.32	1.58	1.39	1.34	1.37
T	1.47	1.08	0.32	0.57	2.43	1.11	1.33	1.25	0.23	1.23	1.54	1.20	1.54	1.54	0.13	1.39	1.67	1.42	1.88	1.68
V	1.58	1.47	0.62	0.83	2.08	1.11	1.92	0.97	0.41	1.50	1.48	1.39	1.83	1.89	0.17	1.78	1.82	1.00	1.63	1.60
W	1.36	1.04	1.03	1.33	1.35	1.03	1.60	2.05	0.80	1.90	1.60	1.22	1.15	1.52	1.16	1.28	1.65	1.88	1.17	1.28
Y	2.24	1.89	0.20	0.57	1.62	0.68	1.66	2.32	0.22	2.13	2.20	1.28	1.71	1.78	0.42	1.54	2.37	1.90	1.35	1.45

B (X × a_1)

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	1.14	0.95	1.07	1.16	1.44	1.57	1.78	1.07	0.79	1.26	1.40	1.67	0.16	1.61	0.38	1.63	1.41	1.34	1.05	1.44
C	1.12	0.53	1.01	1.16	0.72	0.98	1.09	0.58	0.60	0.76	0.93	1.13	0.60	1.27	0.34	1.02	0.95	0.72	0.62	0.74
D	0.84	0.76	0.34	0.31	1.00	0.54	0.68	0.63	0.24	0.50	1.16	0.73	0.16	1.46	0.16	1.02	0.73	0.72	0.50	0.70
E	1.55	1.35	0.58	0.69	1.29	1.17	1.32	1.12	0.56	1.15	1.44	1.21	0.15	1.73	0.36	1.40	1.20	1.23	1.00	1.09
F	1.84	0.87	0.80	0.64	1.34	1.47	1.55	1.83	0.39	1.39	1.54	1.71	1.69	1.80	0.19	1.64	1.59	1.84	1.01	1.22
G	1.21	0.72	0.45	0.57	1.29	1.13	1.00	0.83	0.30	0.76	1.25	1.08	0.16	1.33	0.16	1.29	1.14	1.04	0.77	1.06
H	1.98	1.50	0.94	1.22	1.94	1.31	1.75	1.36	0.64	1.14	1.53	1.81	0.73	1.97	0.21	1.95	1.79	1.82	1.09	1.44
I	1.12	1.09	1.00	0.99	2.01	1.21	1.41	1.36	0.86	1.21	1.76	1.26	0.15	1.48	0.48	1.35	1.35	1.54	0.94	1.35
K	1.45	0.86	1.16	1.01	1.26	1.21	1.49	0.94	0.46	0.88	1.25	1.54	0.16	2.20	0.20	1.70	1.52	1.20	0.87	1.34
L	1.22	0.73	1.03	1.05	1.33	1.20	1.64	1.21	0.66	1.12	1.59	1.57	0.17	1.64	0.29	1.43	1.45	1.28	0.87	1.40
M	1.30	0.79	0.76	0.81	1.08	1.29	1.57	1.12	0.49	1.13	1.34	1.33	0.17	1.51	0.23	1.49	1.41	1.72	0.95	1.20
N	1.42	1.18	0.83	0.86	1.38	1.50	1.72	1.23	0.56	1.07	1.30	1.40	0.16	1.82	0.27	1.73	1.41	1.19	0.95	1.34
P	1.18	1.01	0.85	0.97	0.99	1.09	1.23	1.00	0.51	0.92	1.06	1.25	0.16	1.30	0.22	1.36	1.29	1.10	0.81	1.09
Q	1.60	1.39	0.99	1.02	1.59	1.49	1.88	1.40	0.66	1.26	1.58	1.90	0.16	2.06	0.35	1.80	1.85	1.79	1.22	1.43
R	1.32	0.72	1.16	1.02	1.19	1.15	1.61	0.93	0.44	0.92	1.19	1.43	0.13	1.50	0.19	1.50	1.17	1.11	0.84	1.10
S	1.27	0.80	0.86	0.88	1.37	1.45	1.84	1.05	0.60	0.99	1.15	1.36	0.16	1.53	0.26	1.52	1.31	1.40	0.99	1.52
T	1.78	0.84	1.20	1.22	1.50	1.20	2.11	1.18	0.68	1.12	1.32	1.43	0.15	1.72	0.42	1.74	1.66	1.33	1.08	1.35
V	1.51	1.59	1.29	1.22	2.23	1.23	1.78	1.54	0.94	1.34	1.43	1.35	0.14	1.53	0.61	1.19	1.37	1.51	1.67	1.62
W	1.55	0.88	1.07	1.61	1.32	1.49	1.51	1.61	0.42	1.66	1.51	1.67	1.40	1.69	0.37	1.75	1.62	1.59	1.36	1.33
Y	2.15	1.88	0.86	0.71	1.52	1.43	1.85	1.54	0.49	1.26	1.69	1.82	0.77	2.35	0.15	1.98	2.12	2.33	1.06	1.56

C (X × a_2)

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	1.72	1.40	0.78	0.68	1.82	1.29	1.55	1.86	0.20	1.45	1.67	1.58	0.30	1.82	0.16	1.79	1.69	1.79	1.32	1.99
C	0.77	0.57	1.41	1.57	0.85	1.13	1.38	0.77	0.76	1.03	1.04	1.77	0.43	1.15	0.22	1.04	1.08	0.88	0.77	1.18
D	0.23	0.34	0.17	0.16	0.36	0.25	0.72	0.24	0.17	0.24	0.90	0.20	0.18	0.96	0.17	0.32	0.26	0.25	0.21	0.25
E	0.67	0.96	0.17	0.20	0.71	0.31	0.35	0.41	0.16	0.34	1.61	0.38	0.28	1.50	0.16	1.00	0.60	0.60	0.30	0.25
F	2.26	0.85	1.40	1.56	2.27	2.05	2.81	1.51	0.82	1.42	1.25	2.23	0.38	2.40	0.41	2.69	2.23	2.19	1.31	1.64
G	1.00	0.76	0.18	0.16	1.40	0.52	1.35	0.90	0.15	0.79	1.25	0.63	0.21	1.32	0.15	1.23	0.98	1.19	1.19	1.28
H	2.08	1.52	0.64	1.04	1.51	1.24	1.61	1.62	0.30	1.20	1.90	1.52	0.28	1.75	0.16	1.72	1.72	1.92	0.74	0.93
I	1.55	1.26	1.95	2.02	1.48	1.49	2.02	0.95	2.27	1.17	1.27	2.51	0.93	1.92	1.16	1.46	1.61	1.38	1.21	1.64
K	0.23	0.24	0.15	0.17	0.27	0.20	0.18	0.26	0.14	0.23	0.63	0.25	0.24	0.77	0.14	0.31	0.26	0.31	0.18	0.19
L	1.53	1.01	1.79	1.76	1.72	1.69	2.18	1.21	1.40	1.24	1.09	1.83	0.86	2.38	0.48	1.94	1.52	1.72	1.60	1.69
M	1.96	1.17	1.71	1.57	1.71	2.06	2.86	1.55	1.21	1.62	1.38	1.94	0.42	2.29	0.44	2.16	2.06	1.79	1.34	1.97
N	1.53	1.12	0.52	0.45	1.61	1.09	1.52	0.98	0.19	0.76	1.49	1.33	0.31	1.51	0.18	1.52	1.38	1.51	0.77	0.89
P	1.92	1.00	1.32	0.44	1.51	1.91	1.35	1.70	0.16	1.48	1.81	1.57	0.31	1.83	0.16	1.88	1.88	1.91	1.16	1.76
Q	1.71	1.84	0.94	0.94	1.94	1.44	1.47	1.55	0.34	1.55	1.83	1.53	0.39	1.73	0.16	1.85	1.74	1.88	1.29	1.75
R	0.26	0.19	0.18	0.20	0.39	0.31	0.22	0.27	0.15	0.30	0.50	0.29	0.26	0.49	0.14	0.27	0.35	0.37	0.22	0.24
S	1.50	1.25	0.35	0.37	1.90	1.01	1.82	1.50	0.18	1.28	1.67	1.46	0.30	1.76	0.19	1.73	1.55	1.55	1.37	1.68
T	1.96	1.74	1.18	0.99	2.00	1.57	2.01	1.89	0.56	1.79	1.93	1.76	0.34	1.67	0.25	1.98	1.81	1.98	1.27	1.76
V	1.35	1.11	1.86	1.95	1.61	2.16	2.43	1.49	1.59	1.42	1.14	2.14	0.54	2.22	0.65	1.62	1.90	1.46	1.43	1.73
W	2.13	0.85	1.01	1.36	1.22	1.80	1.60	1.42	0.32	1.44	1.35	1.96	0.31	1.97	0.32	1.93	1.93	1.39	1.09	1.08
Y	2.21	1.26	0.53	1.52	1.51	1.58	1.37	1.45	0.20	1.12	1.69	1.75	0.30	2.06	0.15	2.08	1.77	1.74	0.90	1.43

D (a_2 × a_1)

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0.16	3.39	0.19	0.16	0.82	0.16	0.25	6.50	0.07	2.58	2.23	0.14	0.21	0.21	0.12	0.11	0.14	3.25	0.69	0.21
C	0.23	2.66	0.28	0.16	1.36	0.20	0.30	5.58	0.16	2.65	2.03	0.26	0.22	0.37	0.32	0.15	1.12	2.79	0.32	0.46
D	0.19	0.30	0.16	0.14	0.27	0.09	0.27	2.86	0.11	0.77	0.16	0.15	0.22	0.19	0.16	0.11	0.18	0.68	0.31	0.19
E	0.13	0.23	0.21	0.22	0.22	0.14	0.19	4.04	0.15	0.51	0.65	0.13	0.14	0.37	0.05	0.06	0.12	1.70	0.38	0.28
F	0.15	0.93	0.16	0.19	0.28	0.20	0.28	2.44	0.21	3.97	1.71	0.20	0.15	0.23	0.19	0.11	0.23	0.66	0.30	0.37
G	0.16	0.18	0.13	0.18	0.52	0.13	0.27	3.00	0.14	0.68	0.21	0.15	0.14	0.19	0.16	0.09	0.08	0.81	0.29	0.19
H	0.31	0.33	0.24	0.14	0.46	0.18	0.76	6.05	0.24	1.12	1.94	0.27	0.25	0.24	0.22	0.10	0.13	1.87	0.32	0.44
I	0.48	3.94	0.35	0.20	2.19	0.13	0.31	6.85	0.18	3.20	3.07	0.75	0.42	1.33	0.15	0.20	2.24	4.63	1.00	0.22
K	0.12	0.83	0.19	0.16	0.28	0.12	0.24	1.59	0.12	0.49	0.26	0.13	0.16	0.18	0.11	0.07	0.13	0.78	0.29	0.26
L	0.12	2.19	0.09	0.07	1.13	0.06	0.25	6.30	0.07	1.76	3.54	0.07	0.07	0.63	0.09	0.07	1.11	2.30	1.19	0.32
M	0.15	1.66	0.19	0.20	0.67	0.19	0.22	6.09	0.17	2.01	1.43	0.24	0.14	0.27	0.15	0.13	0.11	1.21	0.32	0.45
N	0.11	0.54	0.17	0.15	0.53	0.10	0.18	7.67	0.11	1.87	0.85	0.14	0.14	0.24	0.21	0.07	0.08	2.02	0.19	0.18
P	0.13	0.94	0.11	0.10	0.60	0.14	0.17	4.83	0.12	1.40	0.80	0.15	0.15	0.17	0.10	0.10	0.18	1.43	0.28	0.36
Q	0.13	0.38	0.18	0.20	0.46	0.13	0.34	5.67	0.20	1.34	0.21	0.17	0.35	0.34	0.17	0.10	0.22	2.63	0.32	0.31
R	0.11	1.72	0.17	0.12	0.88	0.13	0.25	4.29	0.11	2.16	0.95	0.12	0.10	0.20	0.13	0.11	0.12	1.81	0.21	0.24
S	0.09	2.05	0.09	0.11	0.41	0.08	0.14	5.15	0.09	1.88	2.02	0.11	0.13	0.27	0.09	0.16	0.10	1.87	0.21	0.16
T	0.36	1.68	0.21	0.12	1.98	0.14	0.20	5.42	0.11	2.90	2.25	0.12	0.17	0.44	0.08	0.13	0.92	4.29	0.45	0.32
V	1.26	2.25	0.14	0.08	0.77	0.10	0.33	7.51	0.12	2.94	3.56	0.16	0.24	0.36	0.08	0.23	2.26	3.24	0.35	0.22
W	0.30	1.74	0.23	0.22	0.37	0.20	0.39	3.18	0.23	1.67	1.05	0.24	0.25	0.25	0.28	0.18	0.38	2.03	0.31	0.36
Y	0.13	5.19	0.17	0.11	0.31	0.10	0.38	2.71	0.19	1.72	0.37	0.25	0.22	0.27	0.18	0.11	0.11	1.37	0.31	0.28

E (X × a_1)

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0.87	1.76	0.16	0.09	1.13	0.11	0.61	1.23	0.13	1.00	4.10	0.78	0.18	4.57	0.10	2.29	0.75	0.74	0.27	0.76
C	1.07	2.52	0.23	0.17	1.22	0.29	0.63	0.77	0.20	0.64	5.07	1.30	0.25	3.30	0.36	0.93	0.73	0.93	0.26	0.72
D	0.40	1.31	0.17	0.12	0.57	0.11	0.29	0.26	0.15	0.19	1.18	0.23	0.21	0.76	0.12	0.52	0.18	0.23	0.30	0.19
E	0.59	0.92	0.17	0.16	0.48	0.10	0.23	0.46	0.17	0.36	2.84	0.15	0.19	0.81	0.23	0.60	0.46	0.28	0.20	0.49
F	0.85	0.69	0.19	0.20	0.42	0.16	0.24	0.25	0.18	0.27	1.75	0.30	0.22	4.83	0.18	0.36	0.35	0.78	0.40	0.31
G	0.43	0.70	0.12	0.12	0.30	0.19	0.19	0.20	0.28	0.17	1.41	0.17	0.21	1.89	0.11	0.33	0.14	0.17	0.29	0.29
H	0.87	0.99	0.23	0.15	0.84	0.28	0.82	1.08	0.26	0.75	3.43	0.38	0.20	1.03	0.22	2.21	0.40	0.84	0.31	0.30
I	1.58	4.75	0.48	0.25	3.13	0.81	1.74	0.66	0.12	0.92	4.37	3.15	0.27	3.48	0.16	2.77	1.40	0.99	0.28	0.53
K	0.33	0.65	0.22	0.15	0.16	0.10	0.28	0.17	0.15	0.15	1.51	0.16	0.15	0.85	0.14	0.22	0.20	0.49	0.24	0.22
L	1.24	2.80	0.11	0.10	0.63	0.39	0.44	0.66	0.13	0.99	3.59	1.68	0.12	4.91	0.11	1.72	0.63	0.87	0.16	0.16
M	0.91	1.11	0.21	0.21	1.05	0.16	0.26	0.66	0.28	0.60	4.06	0.46	0.18	2.77	0.16	0.84	0.73	0.48	0.30	0.57
N	1.06	1.93	0.17	0.17	0.53	0.10	0.21	0.90	0.12	0.44	2.49	0.46	0.16	2.09	0.12	2.07	0.38	1.19	0.18	0.82
P	0.27	2.54	0.16	0.14	1.78	0.10	0.25	0.67	0.24	0.53	3.40	0.14	0.19	0.76	0.08	0.20	0.15	0.35	0.16	0.13
Q	0.61	1.49	0.17	0.22	0.77	0.32	0.31	0.78	0.31	0.47	1.47	1.02	0.17	1.58	0.15	1.76	0.96	0.63	0.26	0.42
R	1.13	1.44	0.12	0.12	0.53	0.10	0.29	0.28	0.19	0.14	3.34	0.51	0.17	2.55	0.11	1.07	0.78	0.68	0.25	0.13
S	1.41	1.56	0.12	0.09	0.59	0.09	0.22	0.72	0.08	0.60	3.09	0.89	0.11	3.25	0.10	1.05	0.29	0.54	0.14	0.26
T	1.66	2.06	0.15	0.25	1.44	0.24	0.56	0.54	0.10	0.67	4.47	0.86	0.23	3.85	0.11	1.87	1.54	0.96	0.16	0.58
V	1.55	3.07	0.89	0.39	2.39	0.69	0.81	1.25	0.10	0.85	2.53	1.54	0.20	2.73	0.15	2.23	1.85	1.36	0.28	1.31
W	0.70	0.91	0.23	0.21	0.40	0.20	0.36	0.29	0.22	0.22	3.16	0.31	0.34	2.98	0.19	1.53	0.51	0.44	0.34	0.32
Y	0.75	1.66	0.22	0.15	0.39	0.18	0.27	0.32	0.18	0.35	5.25	0.17	0.26	1.83	0.15	0.78	0.32	0.68	0.23	0.31

F (X × a_2)

	A	C	D	E	F	G	H	I	K	L	M	N	P	Q	R	S	T	V	W	Y
A	0.17	0.21	0.16	0.13	0.76	0.12	0.19	0.17	0.17	0.08	0.70	0.19	0.13	0.89	0.09	0.13	0.09	0.10	0.17	0.15
C	0.68	1.79	0.17	0.20	2.31	0.22	0.34	0.32	0.21	0.27	12.5	0.90	0.35	9.26	0.27	1.66	0.48	0.61	0.29	0.30
D	0.29	0.25	0.20	0.13	0.28	0.13	0.35	0.16	0.14	0.14	0.16	0.18	0.19	0.23	0.10	0.07	0.09	0.11	0.22	0.24
E	0.17	0.23	0.18	0.10	0.19	0.17	0.17	0.18	0.15	0.07	0.16	0.17	0.10	0.16	0.12	0.14	0.08	0.08	0.24	0.17
F	0.38	2.47	0.20	0.18	0.33	0.18	0.31	0.33	0.24	0.19	5.93	0.24	0.18	1.93	0.11	0.20	0.23	0.24	0.34	0.31
G	0.11	0.16	0.12	0.12	0.13	0.12	0.27	0.11	0.16	0.07	0.17	0.11	0.12	0.16	0.11	0.08	0.07	0.06	0.27	0.17
H	0.57	0.33	0.20	0.22	0.23	0.27	0.49	0.26	0.26	0.34	0.35	0.14	0.30	0.51	0.16	0.16	0.14	0.27	0.30	0.22
I	7.75	8.22	1.25	0.63	7.12	1.36	2.60	7.66	0.24	6.70	9.34	6.44	0.21	8.32	0.18	9.97	7.18	8.10	0.43	4.01
K	0.11	0.17	0.11	0.14	0.19	0.10	0.21	0.17	0.16	0.10	0.14	0.16	0.14	0.18	0.08	0.08	0.09	0.11	0.25	0.21
L	2.06	5.86	0.09	0.15	0.74	0.15	0.35	0.20	0.07	0.17	9.49	1.18	0.18	12.1	0.10	3.20	0.72	0.43	0.17	0.19
M	1.24	4.26	0.24	0.18	2.08	0.14	0.66	0.36	0.14	0.33	6.64	0.61	0.26	7.37	0.32	2.10	0.72	0.59	0.27	0.78
N	0.11	0.19	0.18	0.19	0.21	0.14	0.19	0.18	0.18	0.09	0.70	0.13	0.33	0.29	0.11	0.09	0.10	0.08	0.24	0.21
P	0.17	0.23	0.20	0.13	0.57	0.12	0.19	0.16	0.14	0.06	0.25	0.15	0.19	0.43	0.08	0.12	0.10	0.13	0.23	0.19
Q	0.17	0.88	0.26	0.18	0.25	0.19	0.37	0.28	0.14	0.17	1.72	0.16	0.33	0.44	0.35	0.13	0.15	0.11	0.22	0.23
R	0.28	0.15	0.17	0.12	0.16	0.12	0.33	0.13	0.09	0.09	0.16	0.11	0.11	0.14	0.14	0.07	0.15	0.10	0.21	0.20
S	0.10	0.20	0.12	0.07	0.24	0.05	0.18	0.08	0.10	0.09	0.22	0.09	0.09	0.18	0.06	0.11	0.06	0.08	0.10	0.14
T	0.11	2.07	0.10	0.06	0.91	0.08	0.29	0.11	0.18	0.07	3.18	0.15	0.10	1.50	0.07	0.50	0.07	0.11	0.14	0.14
V	3.12	6.51	0.18	0.16	1.40	0.46	0.67	0.56	0.21	0.92	9.24	2.99	0.16	4.24	0.14	6.21	1.85	1.83	0.23	0.31
W	0.30	0.31	0.26	0.22	0.37	0.24	0.42	0.39	0.23	0.21	0.88	0.32	0.33	2.00	0.29	0.18	0.19	0.22	0.36	0.32
Y	0.41	0.35	0.16	0.13	0.29	0.35	0.41	0.33	0.37	0.15	0.56	0.25	0.22	0.47	0.15	0.15	0.20	0.23	0.32	0.33

Figure 6

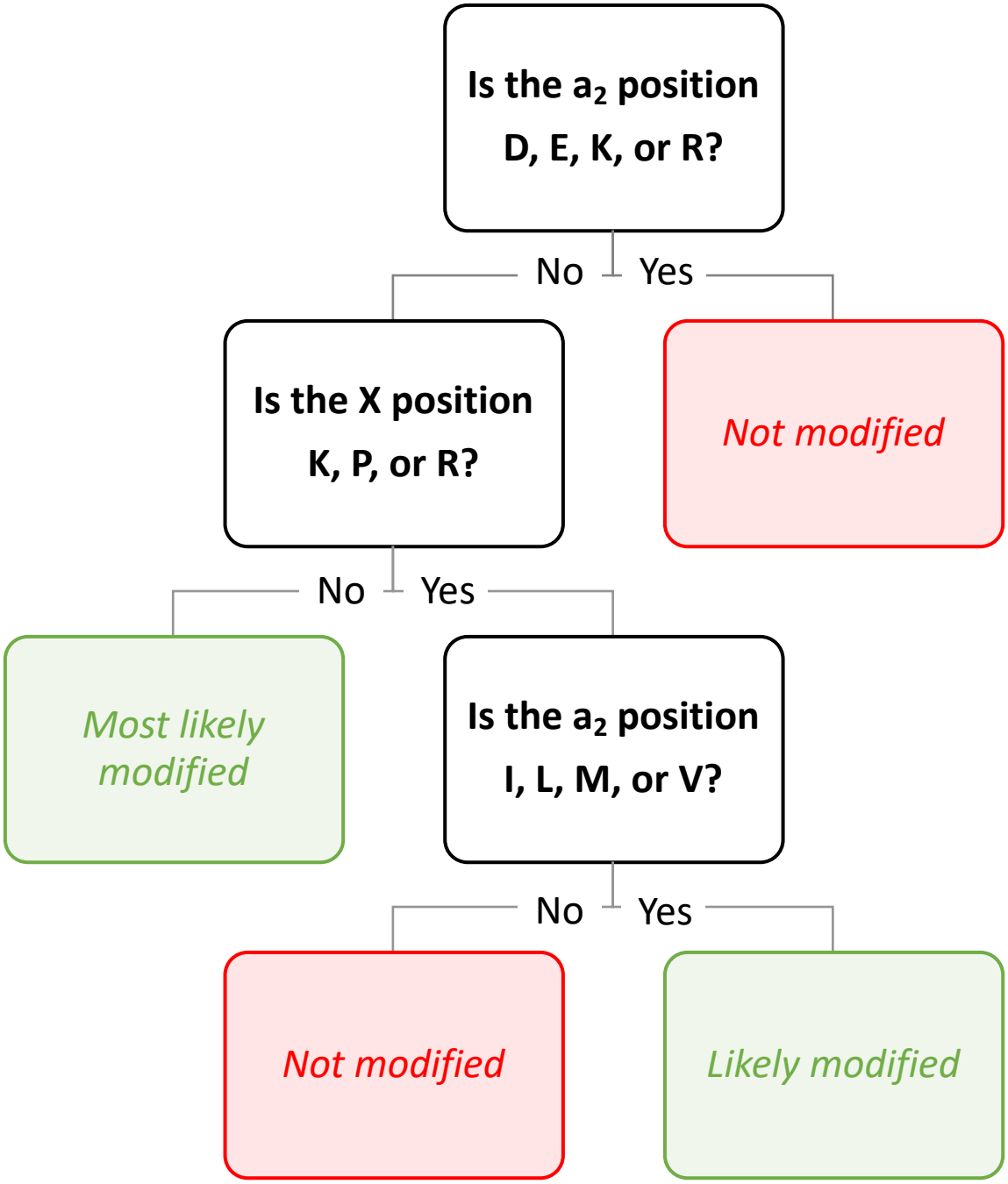
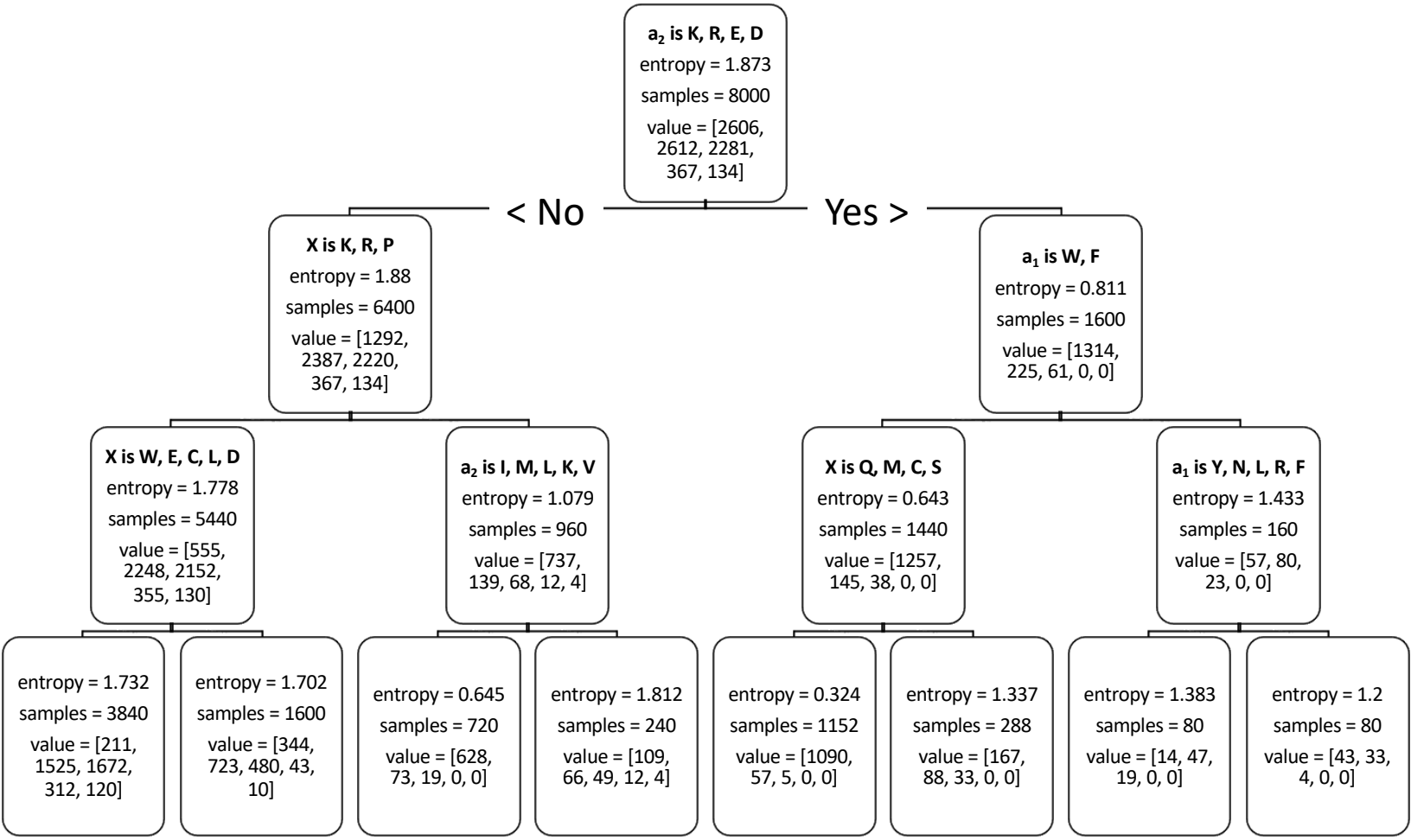

Is the a_2 position D, E, K, or R?
No
Yes
Not modified
Is the X position K, P, or R?
No
Yes
Most likely modified
Is the a_2 position I, L, M, or V?
No
Yes
Not modified
Likely modified

Figure S5


a_2 is K, R, E, D
entropy = 1.873
samples = 8000
value = [2606, 2612, 2281, 367, 134]
< No
Yes >
X is K, R, P
entropy = 1.88
samples = 6400
value = [1292, 2387, 2220, 367, 134]
a_1 is W, F
entropy = 0.811
samples = 1600
value = [1314, 225, 61, 0, 0]
X is W, E, C, L, D
entropy = 1.778
samples = 5440
value = [555, 2248, 2152, 355, 130]
a_2 is I, M, L, K, V
entropy = 1.079
samples = 960
value = [737, 139, 68, 12, 4]
X is Q, M, C, S
entropy = 0.643
samples = 1440
value = [1257, 145, 38, 0, 0]
a_1 is Y, N, L, R, F
entropy = 1.433
samples = 160
value = [57, 80, 23, 0, 0]
entropy = 1.732
samples = 3840
value = [211, 1525, 1672, 312, 120]
entropy = 1.702
samples = 1600
value = [344, 723, 480, 43, 10]
entropy = 0.645
samples = 720
value = [628, 73, 19, 0, 0]
entropy = 1.812
samples = 240
value = [109, 66, 49, 12, 4]
entropy = 0.324
samples = 1152
value = [1090, 57, 5, 0, 0]
entropy = 1.337
samples = 288
value = [167, 88, 33, 0, 0]
entropy = 1.383
samples = 80
value = [14, 47, 19, 0, 0]
entropy = 1.2
samples = 80
value = [43, 33, 4, 0, 0]