Functional dissection of human mitotic proteins using CRISPR-Cas9 tiling screens

Jacob A. Herman[1,2]*, Lucas Carter[2], Sonali Arora[2], Jun Zhu[3], Sue Biggins[1], and Patrick J. Paddison[2]*

[1]Howard Hughes Medical Institute, Basic Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA

[2]Human Biology Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, USA

[3]Department of Genetics and Genomic Sciences, Icahn Institute of Genomics and Multiscale Biology, Icahn School of Medicine at Mount Sinai, New York, NY, 10029, USA

*To whom correspondence should be addressed: jherman@fredhutch.org or paddison@fredhutch.org.

Key words: CRISPR-Cas9, functional genomics, human genome, human proteome, mitosis, kinetochore, spindle assembly checkpoint, MAD1L1, Mad1.

SUMMARY

Kinetochores are large protein assemblies that mark the centromere and bind to mitotic spindle microtubules to ensure accurate chromosome segregation. Like most protein-coding genes, the full multifunctional nature of kinetochore factors remains uncharacterized due to the limited experimental tools for unbiased dissection of human protein sequences. Here, we develop a method that leverages CRISPR-Cas9 induced mutations to comprehensively identify key functional regions within protein sequences that are required for cellular outgrowth. Our analysis of 48 human kinetochore genes revealed hundreds of regions required for cell proliferation, corresponding to known and uncharacterized protein domains. We biologically validated 15 of these regions, including amino acids 387-402 of Mad1, which when deleted compromise Mad1 kinetochore localization and chromosome segregation fidelity. Altogether, we demonstrate that CRISPR-Cas9-based tiling mutagenesis enables *de novo* identification of key functional domains in protein-coding genes, which elucidates separation of function mutants and allows functional annotation across the human proteome.

INTRODUCTION

The sequencing of the human genome (Lander et al., 2001) and large-scale genome characterization projects (Consortium, 2012) have provided a reliable list of >20,000 protein coding genes (e.g., UniProt KB database (Breuza et al., 2016)). However, despite this knowledge and intense study of many genes, our understanding of protein domain functions remains largely incomplete. Current protein annotation methods mainly rely on homology-based comparative genomic approaches to define domains and infer putative gene functions. For example, the human proteome contains 5494 separate conserved protein family (Pfam) domains each with a putative function (e.g., methyltransferase-like domain) (Mistry et al., 2013). However, >3000 of these domains have unknown function (Bateman et al., 2010) and about half of proteome is entirely unannotated (El-Gebali et al., 2019; Mistry et al., 2013; Punta et al., 2012). Even within conserved domains experimental validation remains a critical need since homology-based inference is no guarantee for functional similarity. Despite this clear need, systematic approaches to validate these predictions or independently identify functional domains in the human genome remain severely limited.

One promising technology that may have utility for annotating important functional regions in the human genome is CRISPR-Cas9. In particular, the use of CRISPR-Cas9 nuclease, interference, and activation systems in combination with sgRNA tiling libraries has proven useful for identifying transcriptional regulatory elements (e.g., enhancers) across small intergenic regions (Canver et al., 2015; Canver et al., 2020; Fulco et al., 2016; Simeonov et al., 2017). For protein coding genes, a few studies have employed tiling sgRNA:Cas9 strategy to define new design rules for sgRNAs (Munoz et al., 2016; Shi et al., 2015). These studies found that sgRNA targeting functional domains and in particular Pfam domains had the most penetrant effect on

cell proliferation (Munoz et al., 2016), suggesting this approach might be useful in identifying critical sequences within protein-coding genes. In fact, analysis of editing events at genomic loci revealed that in-frame repairs were responsible for this phenomenon (Michlits et al., 2020; Shi et al., 2015). In these scenarios, after Cas9 nuclease induces a double-stranded DNA break, error prone repair creates an in-frame insertion or deletion in ~20% of cases (Chakrabarti et al., 2019; van Overbeek et al., 2016), giving rise to a random but site-specific mutation. Thus, if mutations occur in amino acids that are functionally constrained (e.g., important binding interface or enzymatic center) they will be less tolerated than mutations occurring in other genic regions, and result in more severe phenotypic consequences. Thus, when a gene is saturated with sgRNAs and their effect on cell proliferation is plotted against where they target a protein coding DNA sequence (CDS) then functional regions can be identified as phenotypic 'peaks' that are most sensitive to CRISPR-Cas9 targeting (Fig. 1A).

A key advantage of this approach is its physiological relevance. Other approaches generally rely on ectopic overexpression of mutant ORFs, while mutagenizing endogenous loci retains a gene's regulatory elements and contextualizes protein function in normal cell physiology. Thus, CRISPR-Cas9-based tiling mutagenesis is likely to reveal functional motifs independent of computational predictions and common experimental biases. Previous studies observed qualitative overlap between tiling sensitive regions and known protein domains, but these efforts were limited by interrogating a gene set lacking robust biological characterization (He et al., 2019; Munoz et al., 2016). Moreover, the ability to identify novel functional regions was suggested but lacked serious biological validation (He et al., 2019). Here we survey a set of human mitotic genes to test the hypothesis that tiling mutagenesis can validate and refine

our understanding of functional motifs within protein-coding genes, while also identifying novel ones.

Kinetochores are large multi-subunit complexes that are assembled upon centromeres to bind microtubules of the mitotic spindle and power chromosome movements that ensure their equal segregation (Fig. 1B). Accurate chromosome segregation requires that each pair of kinetochores on duplicated chromosomes bind to microtubules anchored to opposite ends of the mitotic spindle. However, at the onset of mitosis kinetochore-microtubule attachments form randomly, resulting in erroneous attachments (i.e. both kinetochores bound to microtubules anchored at the same end of the mitotic spindle) that must be corrected or result in chromosome mis-segregation (Hara and Fukagawa, 2020). Thus, the kinetochore leverages two mechanisms to ensure accurate chromosome segregation. First, an error correction pathway senses and corrects erroneous attachments, while also actively stabilizing correct ones. Second, erroneous kinetochore-microtubule attachments generate a biochemical signal that activates the Spindle Assembly Checkpoint (SAC), which prevents premature separation of duplicated chromosomes or mitotic exit until every pair of kinetochores have formed accurate microtubule attachments (London and Biggins, 2014b; Musacchio, 2015).

To accomplish these diverse tasks the kinetochore is associated with over 80 protein components that form at least seven protein sub-complexes, including the constitutive centromere associated network, chromosomal passenger, Mis12, mitotic checkpoint, Ndc80, RZZ, and Ska complexes (Hara and Fukagawa, 2020) (Figure 1; Table S1). Over three decades of study have revealed much of the underlying chemical and physical properties that enable kinetochore assembly, attachment to microtubules, and chromosome movements, yet we still do not fully understand the multifunctional nature of these factors. In fact, alternative

and moonlighting functions have recently been identified in kinetochore proteins only through targeted mutations or synthetic interactions with oncogenic signals (Ding et al., 2013; Herman et al., 2020; Nguyen et al., 2021). Thus, kinetochore-associated proteins are an ideal set of targets to determine if CRISPR-Cas9 tiling mutagenesis can resolve both known and uncharacterized protein sequences that contribute to essential functions.

Here, we performed tiling mutagenesis on 48 kinetochore-associated genes and leverage three methods to analyze these data within sequence space, ultimately identifying hundreds of essential protein regions of which ~30% appear to have never been characterized. We then performed the most robust biological validation of tiling mutagenesis to date and show that a novel domain identified in our screen contributes to the SAC by facilitating Mad1 kinetochore localization.

RESULTS

Tiling mutagenesis strategy using CRISPR-Cas9 nuclease

In tiling screens, like other CRISPR-Cas9 screens, the nuclease is targeted to a unique genomic site where it generates a blunt-ended, double-stranded break 3 bp upstream of a PAM (Jinek et al., 2012; Makarova et al., 2011). In somatic cells, this dsDNA break is commonly repaired by error-prone non-homologous end joining (NHEJ), leaving repair scars in the form of small insertion/deletion (indel) mutations (Hartlerode and Scully, 2009; Lieber et al., 2003). Because of the triplet nucleotide nature of protein reading frames, if repair is unbiased, 2/3 of these indels will trigger frameshift mutations, while 1/3 will retain frame (Fig. 1A). However, recent deep sequencing of >100 genomic loci in human cells targeted by Cas9, suggest that ~20% of mutations retain the reading frame (Chakrabarti et al., 2019). Thereby, when using a

single sgRNA targets a *population* of diploid cells, 45-64% of cells should harbor biallelic frameshift mutations, while the remaining cells will carry at least one in-frame allele (Fig. 1A). For essential genes, biallelic frameshift mutations will cause a total loss of protein function and thus these cells will deplete from the population independent of where the gene is targeted (Fig. 1A, frameshift). However, a cell gaining an in-frame mutation at a critical codon in a phenotypically constrained domain (e.g., in a catalytic domain) is more likely to die than a cell gaining an in-frame mutation in a more mutable, less essential part of the gene (Fig. 1A, in-frame mutation). Thus, when saturating the CDS of essential proteins with sgRNAs the sequences that are most depleted from the population likely target essential motifs.

Generation of a CRISPR-Cas9 tiling library targeting kinetochore-associated factors

We identified 48 kinetochore-associated factors spanning biological functions and genomic contexts to be targeted in the tiling library. This included proteins that regulate the dynamic mitotic spindle, kinetochore factors that directly bind microtubules, and proteins involved in genomic surveillance processes like error correction and the SAC (Figure 1B). We also targeted mitotic factors with 'moonlighting' interphase functions such as Survivin/BIRC5 that maintains centromere integrity during mitosis and inhibits apoptosis during interphase. We also targeted paralogous genes (CLASP1/2, MAPRE1/2/3) on the chance that cell-specific requirements for such factors may be uncovered. With genes chosen we then identified all the *unique* sgRNA targeting the CDS that (with a few exceptions) did not target other coding regions of the genome (Fig. S1A). This resulted in a library of 6500 sgRNAs with median spacing of 14 nt between cut sites within the CDS and a maximum spacing of 148 nt due to a lack of the spCas9 PAM sequence (NGG) (Figure 1C).

For tiling to identify functional motifs within a protein, there must be an unbiased distribution of in-frame indels/mutations across a protein's CDS. Using two different predictors for repair bias after CRISPR editing, we found the library on average did not contain any positional bias for sgRNA predicted to favor frameshifting edits (Fig. S1B-C) (Chakrabarti et al., 2019; Shen et al., 2018). This analysis indicated that on a per gene basis, some bias could be observed in genes with a short CDS because each bin contained only 1-2 sgRNA (Fig. S1D-E). Thus, at the sgRNA/gene coverage in our library bias is unlikely for proteins longer than 300 amino acids, while for shorter proteins the screen may under-report the boundaries or number of functional regions. This is not surprising as the concept of tiling is better suited for characterizing large multifunctional proteins.

We also included 601 non-targeting control (NTC) sgRNA sequences which cause no editing in the human genome such that positive and negative changes in representation could be related to an unperturbed rate of proliferation (Sanjana et al., 2014). Finally, to monitor screen performance we also included 3-4 sgRNAs targeting genes that have been shown in numerous screens to positively (CDKN2A, TP53, etc.) or negatively (POLR2L, HEATR1, etc.) regulate proliferation (Feldman et al., 2019; Toledo et al., 2015). This final library contained 7147 sgRNAs (Fig. 1C) which were synthesized in a pool and inserted into an 'all-in-one' single lentiviral expression vector.

Tiling proliferation screen is technically and biological reproducible, and most potent when targeting functional protein regions

To test the effect of each individual sgRNA on proliferation we infected three independent replicates of cells such that each sgRNA was incorporated into the genome of 650 cells (Fig. 1D). The population was then enriched for integrants using puromycin selection, and then an initial sample from each replicate was harvested for sequencing analysis and frozen. Replicates were sub-cultured for eight days until a final sample from each replicate was harvested (Fig. 1D). The sgRNA sequences themselves were PCR amplified from genomic DNA of each population and then identified through Illumina sequencing to determine how the representation of each sgRNA within the replicate had altered over the 12 days of proliferation. The change in normalized sequencing reads for each sgRNA were used to calculate $\log_2$(fold change) values and a Z score such that the different cell lines and experimental replicates could be directly compared (Supplementary Table 1).

To ensure that this approach could be used generally and was not driven by unique cellular or genomic contexts (copy number variations, doubling rates, etc.) we performed the tiling screen in four diverse cell types, which varied in their transformation status, degree of aneuploidy, and tissue of origin. This included common cell lines HeLa and HCT116, as well as a TERT immortalized retinal pigment epithelial cell line ($ARPE^{TERT}$) and laboratory transformed derivative with numerous genetic alterations including an ectopic copy of oncogenic HRAS ($ARPE^{RAS}$). We found that despite these unique cellular environments, on-target sgRNA (non-targeting controls excluded) affected proliferation similarly in all cell types (Fig. 2A-B). The sgRNA affected proliferation in $ARPE^{TERT}$ and $ARPE^{RAS}$ cells extremely similarly (Pearson coefficient 0.96) as expected from their shared parental cell line, yet sgRNA effects on viability were also strongly correlated (Pearson coefficients > 0.81) in cells from diverse tissue and disease types (Fig. 2A-B). These correlations were also observed when only sgRNA with the

most potent *decreases* in proliferation were analyzed (bottom quartile) (Fig. S2A-B). These results indicated that our techniques were reproducible (data span unique preparations of lentiviral particles, library amplifications, and sequencing runs), but more importantly, that our tiling library had similar *phenotypic* outcomes among diverse cells lines despite the semi-random nature of DNA-damage repair following CRISPR-Cas9 targeting.

The goal of the tiling library was to identify motifs within sequence space that contribute to the essential activity of mitotic factors. Thus, we first determined at the gene-level, which of our targets had a negative effect on cell proliferation. We first identified all sgRNA with a Z score less than -1, indicating these sgRNAs had been depleted from the population by at least one standard deviation. We then asked if more than 9.5% of all sgRNA targeting a single gene met this threshold and reported the number of cell lines for which this was true (Fig. 2C). Genes in which more than two cell lines met this threshold are colored teal and can be compared to a similar metric (dependency score) used in the DepMap platform, which at this time had performed CRISPR screens with 4-6 sgRNA targeting a single gene in more than 750 cell lines (Meyers et al., 2017; Tsherniak et al., 2017). This analysis revealed several expected outcomes, such as sgRNA targeting CLAPS1 or CLASP2 not affecting proliferation because these paralogs function redundantly for most known activities (Mimori-Kiyosue et al., 2005; Pereira et al., 2006). However, there were a few surprising findings, particularly how MAD1L1, MIS12, SKA2, and CENPM behaved differently between our screen and DepMap reports (Fig. 2C).

We found that targeting the MIS12 gene did not have a significant effect on proliferation, while a negative effect on proliferation was seen in 90% of cell lines tested by DepMap researchers (Fig. 2C). Moreover, during mitosis the Mis12 protein functions within a heterotetramer with Dsn1, Pmf1, and Nsl1 (Petrovic et al., 2016) and all three of those proteins

met the negative proliferation threshold. Looking at all the sgRNAs targeting MIS12 showed that ~80% of them performed similarly to non-targeting controls suggesting they either failed to cause editing or the repair was biased towards the wild-type sequence and thus the gene overall did not meet our threshold (Figure S2C). Interestingly, three of the four sgRNA sequences from the DepMap library were in our data and happen to be some of most active sgRNA (Sanson et al., 2018). We also observed the reverse behavior in MAD1L1. We found that all four cell lines had negative proliferation outcomes when this spindle assembly checkpoint member was targeted, yet DepMap screening suggests that <1% of cell lines are affected (Fig. 2C). We identified four of the six DepMap sgRNA sequences in our data and found that most of those sequences did not affect proliferation, yet with our increased number of sgRNA we saw many other sequences had a strong negative effect (Sanson et al., 2018)(Fig. S2C). Thus, the high-density data derived for each gene from tiling libraries can both obscure and elucidate biology seen in genome-wide approaches. We also found that targeting CENPM and SKA2 had negative proliferation outcomes, whereas DepMap data suggest no growth defects (Fig. 2C). This is because DepMap sgRNA sequences for these genes target rarely transcribed exons (Fig. S2D). Thus, because of the sheer number of sgRNA, tiling libraries can interrogate rare exons without confounding our understanding at the gene level. Altogether this indicated that tiling analysis is limited by functional biology (redundant paralogs CLASP1/2), more than biases in CRISPR technology (poor performance of MIS12 sgRNAs).

Having identified 36 genes that were required for wild-type levels of proliferation in our data, we set out to determine if any global characteristics drove the performance of the sgRNAs targeting these genes - primarily if we also observe that sgRNAs targeting functional protein motifs have the most negative effect on proliferation. After synthesizing our library, it was shown

that sgRNAs enriched for pyrimidines near their 3' end cause premature polymerase termination from the U6 promoter (Graf et al., 2019). We identified sgRNAs with these pyrimidine rich (Y rich) sequences within our own data and confirmed those findings independently. Pyrimidine rich sgRNAs on average had no effect on cell proliferation and instead behaved like a non-targeting control (Fig. 2D). We excluded these sequences from our global analysis and asked which protein or genomic features were associated with sgRNA activity in our outgrowth screen. First, we tested conventional wisdom that targeting early exons (within 5-65% of the CDS) resulted in more penetrant loss of activity due to more robust nonsense mediated decay (Doench et al., 2016; Sanson et al., 2018). We found no association between targeting early exons and sgRNA performance (Fig. 2E), instead our data were consistent with the recent suggestion that sgRNAs targeting functional protein motifs result in the most potent phenotypes because in-frame edits are not tolerated in essential domains (Michlits et al., 2020; Munoz et al., 2016; Shi et al., 2015). We found that sgRNAs targeting PFAM domains or functional regions annotated directly from literature had on average the most negative effect on proliferation (Fig. 2E). This was promising for our goal of using in-frame mutations to identify functional regions, so we next tested if sites predicted to cause in-frame or frameshifting mutations had the strongest effect on proliferation. In fact, sgRNAs that are predicted to be more likely overall (>50% of cases) or more likely than the median (>22.4% of cases) to create in-frame edits were associated with decreased proliferation (Fig. 2E). This strongly suggests that proliferation phenotypes in our screen are *not* driven by frameshift mutations. Together this global analysis is consistent with the observation that in-frame edits caused by repair after CRISPR-Cas9 nuclease activity are common and have the most potent effect on cell proliferation when they occur in an essential region of the CDS.

Multiple approaches for integrating tiling data within sequence space perform similarly to reveal known and uncharacterized protein regions at different scales

The power of the tiling library is to gain an unbiased understanding of protein function within sequence space. Thus, for each gene we can display the average Z score for all the targeting sgRNAs from the four cell lines (vertical gray bars) along the translated CDS (Fig. 3A). In KIF18A we observe sgRNAs with a strong negative effect on proliferation and sgRNAs that appear largely inactive since they behave similarly to the average of non-targeting controls (Fig. 3A). To integrate these data over the CDS we used two previously published approaches, CRISPR-SURF and ProTiler, (He et al., 2019; Hsu et al., 2018) that transform results from tiling CRISPR screens into a stepwise function. From this stepwise function ranges of nucleotides or amino acids that are negatively enriched compared to non-targeting controls or a local 'zero' are identified (colored regions within each gray bar, Supplementary Table 2). As demonstrated by KIF18A, ProTiler and SURF tended to identify many high-resolution regions (10-15 amino acids) that correlate with key functional motifs such as the nucleotide binding pocket in the kinesin motor domain. To complement these high-resolution approaches, we paired tiling data with a convex fused lasso (TiVex) to generate a more smoothed stepwise function (Parekh and Selesnick, 2015). This allowed TiVex to identify larger windows that overlapped *many* SURF or ProTiler regions and often represented discretely folded protein domains such as the KIF18A kinesin domain (Supplementary Table 2) (Fig. 3A).

This trend was also evident in an overview of regions identified by each analysis method in our 36 target genes (Fig. 2C). SURF and ProTiler identified ~500 small (10-15 amino acid) regions and TiVex identified ~150 large (50-100 amino acid) regions (Figure 3B). The resolution of each method can also be demonstrated in three dimensions by mapping SURF or TiVex

regions onto the crystal structure of target protein Bub3 bound to a fragment of Bub1 (Larsen et al., 2007; Pettersen et al., 2004) (Fig. 3C). The entire Bub3 protein folds into a WD40 structure that overall contributes to the interaction with Bub1, thus TiVex identified essentially the entire protein as important for proliferation (Fig. 3C, left, green residues). However, SURF primarily identified the two beta sheets that contain a pair of tryptophan residues that are specifically required for Bub1 binding (Fig. 3C, right, blue residues) (Larsen et al., 2007; Pettersen et al., 2004). Similarly, mapping TiVex and SURF regions onto the structure of a single domain within chTOG/CKAP5 bound to a tubulin dimer showed that TiVex identified the entire fold as important while SURF regions clustered more specifically at the binding interface (Figure S3A).

For some small proteins like Bub3 TiVex identified most of the sequence as important, which is consistent with the reality of protein function. We find that on average TiVex identifies ~55% of the protein sequence within each gene as contributing to proliferation, which is similar to the same analysis with Pfam domains or literature defined functional motifs which cover 50-60% of protein sequence (Fig. S3B). The high-resolution nature of SURF and ProTiler are also highlighted by this metric as they identify 20-30% of the overall protein sequence as important for proliferation (Fig. S3B). Thus, we propose TiVex may be better suited to identify protein regions that are *sufficient* for a specific activity, while ProTiler and SURF are likely to reveal the regions that are *required* for a specific activity.

TiVex identified protein domains of similar size to Pfam, yet unlike Pfam, TiVex was not restricted to conserved protein sequences. We calculated an average conservation score for each region identified by SURF, ProTiler, or TiVex based on the nucleotide conservation among 100 vertebrates (PhyloP) within the UCSC genome browser (Kent et al., 2002; Pollard et al., 2010). Approximately 70% of protein motifs identified in all analysis methods demonstrated

some sequence conservation among vertebrate species ($P < 0.05$), while the sequence was not constrained in the remaining ~30% of protein motifs. This suggests that CRISPR tiling screens are applicable to proteins with rapidly evolving functions and those containing disordered regions. Consistent with this interpretation, when we cross referenced regions identified by SURF, ProTiler, and TiVex with our manually curated list of functional regions identified in literature we found that 34-39% of regions identified by tiling have, to our knowledge, not yet been characterized by the field (Fig. 3B). Some of these unstudied motifs overlapped with conserved regions (Pfam domains) but many of them fell in regions predicted to be disordered, or not within either of those categories (Supplementary Table 3) (Figure 3E).

Overall, we see strong agreement between all three analysis methods. In pairwise comparisons 100% of the protein regions identified by each method overlap in 20-50% of genes (Fig. 3F). These discrepancies are primarily focused in 3-4 genes like KNTC1 and CENPF (Fig. S3C) and likely arise from how each method defines 'zero'. Computational analysis of our CRISPR tiling library targeted at mitotic factors further confirmed the early findings for this approach. We see that sgRNAs most strongly affecting proliferation are correlated with targeting functional protein regions. Moreover, the most potent sgRNAs are *not* predicted to favor frameshifting mutations, nor must they target an early exon. We successfully applied two previously published high-resolution approaches to integrate tiling data and extended this by developing a broader analysis method, TiVex. Finally, we found that all three methods identify functional motifs outside conserved regions, including in disordered regions. With these promising findings, we next performed the most robust biological validation of tiling libraries, to date.

Biological validation of essential regions identified by CRISPR tiling reveals high accuracy, and minority of domains contribute to protein stability

We validated our finding that uncharacterized regions identified by tiling were important for cell proliferation using sgRNA:Cas9 gene knockouts paired with complementation. First, we generated wildtype proteins that were resistant to Cas9 editing, and then created mutant proteins which contained small (10-40 amino acid) deletions corresponding to regions identified by SURF, ProTiler, and/or TiVex. These wildtype and mutant proteins were N-terminally tagged with a 2xFlag tag and/or EGFP to enable downstream analysis through biochemistry and cell biology, and each one named for the first residue within the small deletion (Ska3$^{\Delta 238\text{-}253}$ shortened to Ska3$^{238\Delta}$ or 238Δ). We then used the Flp Recombinase to insert the DNA coding for these exogenous proteins at a unique genomic locus within the parental cell line (Gossen and Bujard, 1992; O'Gorman et al., 1991; Taylor et al., 1998) (Fig. 4A). Cell lines encoding the wildtype or mutant proteins were then electroporated with a Cas9 in complex with 1-2 synthetic sgRNAs either targeting the endogenous protein or not targeting the genome at all (Hoellerbauer et al., 2020a; Hoellerbauer et al., 2020b). Doxycycline was either withheld or added after electroporation to test the effect of endogenous gene knockout and whether expression of the wildtype or mutant protein complemented its essential activity. We tested regions within Ska3, BuGZ, and Cenp-K that were identified by all three methods, and one additional region in Ska3 that was not identified as a control (Fig. 4B-D). In all cases, the wildtype proteins provided a significant rescue for cell proliferation following endogenous protein knockout, as did the control deletion in Ska3 (Fig. 4B-D). We observed the same behavior in Cenp-H and Sgo1 deletion mutants that were predicted by all three methods, but also found that a region identified solely by ProTiler was a false positive and was not required for proliferation in our validation study (Fig.

S4A-B). Altogether, using this complementation approach we verified that 10/11 ProTiler regions, and 10/10 TiVex overlapping SURF regions were required for cell proliferation.

To begin characterizing what function these regions played in the overall protein context we first tested their contribution to protein folding/stability at steady state. When wildtype and mutant proteins were expressed in the presence of endogenous protein, we found that 2-3 mutant proteins were present but unstable, while one mutant protein appears to accumulate well beyond wildtype levels (Fig. 4B-D, S4A-B). This indicates that tiling screens identify protein regions that regulate or drive protein folding and stability. This comprehensive analysis suggests that CRISPR-Cas9 tilling libraries are a reliable means to identify previously uncharacterized protein regions that contribute to essential activities.

Tiling MAD1L1 reveals a protein motif that contributes to its kinetochore localization

Our initial validation focused on regions predicted by all three analysis methods and thus are some of the most robust signals, so next we asked how tiling performed in more of an edge case, the MAD1L1 gene. In the case of Mad1, previous literature, SURF, ProTiler, and TiVex all agreed that the C-terminus of the protein is particularly important for its essential activity (Fig 5A). This region contributes to Mad1 forming a homodimer and then binding to multiple other kinetochore factors like Bub1 and Cdc20 (Allan et al., 2020; Brady and Hardwick, 2000; Kim et al., 2012; Lara-Gonzalez et al., 2021; London and Biggins, 2014a; Piano et al., 2021). However, in the 600 amino acids upstream of that region we saw very little agreement between SURF, ProTiler, and TiVex (Fig. 5A). Thus, we generated four deletion mutants outside the well characterized C-terminus that were identified by only SURF or SURF and TiVex (Fig. 5A). Using the same approach as before (Figure 4A) we tested the ability of mutant proteins to complement

MAD1L1 knockout. Despite targeting the gene with multiple sgRNAs and waiting 7 days to assay cell number we saw only partial defects in proliferation, which is consistent with previous findings that depleting Mad1 protein is exceedingly difficult (Allan et al., 2020; Rodriguez-Bravo et al., 2014). Despite this difficulty, we found that wildtype protein and a control deletion (Mad1$^{170\Delta}$) rescued the partial proliferation defect, while two mutants (Mad1$^{25\Delta}$ and Mad1$^{272\Delta}$) identified by SURF and TiVex did not (Fig 5B). A final mutant identified only using SURF analysis appeared to provide a rescue but could not be statistically resolved (Fig. 5B). To better interrogate these mutant proteins, we instead analyzed their mitotic behavior in the presence of endogenous protein. We validated that none of the mutations compromised protein stability (Fig 5C) and then determined if mutant proteins were able to perform an essential Mad1 activity, maintaining the spindle assembly checkpoint. We induced expression of each Mad1 protein overnight and then treated cells with the microtubule destabilizing drug nocodazole for 20 hours, which should trigger a robust SAC arrest. However, we found that fewer cells expressing Mad1$^{387\Delta}$ arrested in mitosis following this treatment, indicating this region of Mad1 contributes to SAC signaling (Fig. 5D).

Robust SAC signaling requires that Mad1 localize to the kinetochore where it serves as a scaffold to assemble the biochemical inhibitor of mitotic progression (Brady and Hardwick, 2000; De Antoni et al., 2005; Lara-Gonzalez et al., 2021; Piano et al., 2021). Using quantitative fluorescence microscopy, we found that all mutant proteins localized to unperturbed kinetochores at the same level as wildtype protein (Fig. S5); however, the SAC assay was performed in the microtubule poison nocodazole, so we next measured kinetochore localization of mutant proteins in nocodazole. Consistently, we found that the Mad1$^{387\Delta}$ protein exhibited defects in kinetochore localization under this condition (Fig. 5E). While Mad1$^{387\Delta}$ kinetochore

levels were reduced, a significant amount of protein still localized to kinetochore, indicating at least one kinetochore recruitment mechanism remained functional in this mutant.

Mad1$^{387\Delta}$ and Mad1^{R617A} contribute to kinetochore recruitment independently

This finding fits well with our current understanding of Mad1 function. In budding yeast mutating a conserved arginine within an RLK motif to alanine prevents Mad1 from binding Bub1 and localizing to the kinetochore (Brady and Hardwick, 2000); however, in other eukaryotes, including human cells, this mutation or Bub1 depletion reduces the amount of Mad1 at kinetochores but does not ablate it (Kim et al., 2012; Zhang et al., 2015). In fact, recent evidence suggests that Mad1 is initially recruited to kinetochores through this interaction with Bub1, but when kinetochores remain unattached to microtubules for long periods (such as in nocodazole) the fibrous corona expands and the RZZ complex (Rod, Zw10, Zwilch) recruits a separate population of Mad1 to kinetochores (Kim et al., 2012; Rodriguez-Rodriguez et al., 2018; Zhang et al., 2015). Thus, we hypothesized that Mad1$^{387\Delta}$ exhibited kinetochore localization defects specifically in nocodazole because this region contributes to an interaction with RZZ or other corona factors. Our hypothesis would also explain why this region was identified only by a single analysis method: in normally cycling cells the Bub1-associated Mad1 is sufficient for SAC activity, while corona-associated Mad1 is required in just a subset of an unperturbed population.

Based on this hypothesis we looked for co-conservation of RZZ proteins and our region of interest. We performed multiple sequence alignments of dozens of putative Mad1 proteins from diverse eukaryotes (van Hooff et al., 2017), which clearly shows conservation of the RLK motif but failed to identify homology near amino acids 387-402 that comprise our region of interest (Figure S6A). In fact, this region is highly divergent and full of gaps while the immediately

flanking sequences align with high occupancy and some shared identity. Even when restricting alignments to species that contain RZZ orthologs (primarily metazoan), alignments were driven by downstream sequences and many gaps were present in our region of interest (Figure S6B). We finally observed some features of conservation among coelomates when flanking Mad1 sequences were excluded from the alignment (Figure S6C). This suggests our region is evolutionarily divergent relative to other Mad1 motifs and is further evidence that CRISPR tiling screens can identify regions not easily recognized through sequence homology searches.

With unclear evolutionary evidence for co-conservation between RZZ members and the 387Δ mutation, we instead tested this experimentally. We mutated Arg 617 of the RLK motif to Ala ($Mad1^{R617A}$) alone or in combination with $Mad1^{387\Delta}$ to determine if they gave additive phenotypes (Fig. 6A). We overexpressed these mutant proteins and found that fewer cells were able to maintain a SAC arrest in mitosis when expressing $Mad1^{387\Delta}$ or $Mad1^{R617A}$ versus $Mad1^{WT}$ (Fig. 6B). When the mutations were combined, we observed a slight additive effect, but we suspect this was limited by the presence of endogenous Mad1 protein (Fig. 6B). Consistent with the loss in SAC activity observed when either $Mad1^{387\Delta}$ or $Mad1^{R617A}$ are over-expressed, we found that both mutations compromised Mad1 association with kinetochore by ~50% after one hour of nocodazole treatment (Fig. 6C). When the mutations were combined the protein virtually failed to localize to kinetochores. Consistent with previous results (Kim et al., 2012), this suggests that neither $Mad1^{387\Delta}$ nor $Mad1^{R617A}$ tend to dimerize with endogenous protein and that Mad1 residues 387-402 contribute to its kinetochore localization, independent of its interaction with Bub1. Thus, this region likely mediates or stabilizes an interaction with RZZ or another member of the fibrous corona, but this remains to be formally tested.

DISCUSSION

We have demonstrated that CRISPR-Cas9 tiling mutagenesis of endogenous protein coding sequences in the human genome can be used to functionally validate and identify critical protein regions, including conserved and divergent protein sequences. Our approach takes advantage of the naturally occurring mutagenic properties of error-prone NHEJ in human cell lines after a dsDNA break is introduced by sgRNA:Cas9 activity. Focusing on 48 kinetochore-associated genes, we were able to interrogate 36 which produced reliable results for our phenotypic readout (i.e., cell growth) (Figure 2C), and produce a unique sgRNA tiling profile for each (Figures 3-4). The negatively enriched sgRNA regions from these screens, we argue, represent differentially constrained regions of the encoded protein, which are less mutable than other regions due to the presence of an essential gene activity.

We can imagine three explanations for why these regions are sensitive to mutations: that they harbor a critical protein-protein interaction or catalytic activity; that they direct protein folding or stability (Labbadia and Morimoto, 2015); or that they contain non-coding DNA elements (e.g., transcriptional or splicing enhancers) embedded in the protein coding sequence. Our correlations between sgRNA performance and literature defined functional domains (Figs. 2-3) and our complementation studies in cells (Figs. 4, S4) are consistent with the first two explanations. In our biological interrogation of 15 domains, we identify regions that contribute to protein folding/stability and regions required for more specific activities (i.e., subcellular localization). This analysis suggests that protein folding/stability explains a minority of functional motifs as we observed this phenotype in ~2 of 15 of the mutant proteins we tested. Instead, tiling mutagenesis is likely to reveal unique protein-protein interactions or enzymatic behaviors.

In the process of rigorously validating CRISPR-Cas9 tilling as a discovery tool, we also generated a powerful resource for the study of kinetochore genes. Despite kinetochore factors being the subject of robust and groundbreaking study for nearly four decades, our tiling screen paired with three complementary domain calling methods (i.e., SURF, ProTiler, and TiVex) identified 50-186 essential regions in 36 proteins that have not yet been studied (Fig. 3B, E). Previous efforts to dissect human kinetochore factors relied on structure or sequence homology to guide truncations or mutations, but our functional screening was not limited in this way and many uncharacterized domains of interest occurred in evolutionarily divergent or disordered sequences (Fig. 3D-E). Revealing these important regions should expedite our collective molecular understanding of kinetochore biology, as will this rich dataset of experimentally validated sgRNA sequences.

The power of CRISPR-Cas9 tilling was truly demonstrated by our in-depth analysis of a previously unstudied region in the middle of MAD1L1/Mad1. Mad1 localization to the kinetochore is dependent on distinct interactions with Bub1 and the RZZ complex (Kim et al., 2012; Rodriguez-Rodriguez et al., 2018; Zhang et al., 2015). While the interaction with Bub1 has been deeply characterized (Brady and Hardwick, 2000; Kim et al., 2012; Lara-Gonzalez et al., 2021; Piano et al., 2021), through tiling mutagenesis, we revealed an unexpected and largely evolutionarily divergent 15 amino acid sequence in the middle of the Mad1 protein (aa387-402) that contributes to its SAC activity and kinetochore localization. We found that deleting Mad1 residues 387-402 reduced its localization to the kinetochore likely by preventing an interaction with the RZZ complex, because combining the deletion with a Bub1-binding deficient mutant (R617A) virtually ablated all kinetochore localization. Our findings are consistent with several studies that indicate two populations of Mad1 exist at the kinetochore

and they each rely on distinct regulatory mechanisms (Kim et al., 2012; Rodriguez-Rodriguez et al., 2018; Zhang et al., 2019; Zhang et al., 2015). This combined evidence strongly suggests that the Mad1$^{387\Delta}$ mutant prevents an interaction between Mad1 and members of the fibrous corona like the RZZ complex. This was unexpected because a recent report found that Cyclin B1 interacts with the most N-terminal region of Mad1 and facilitates its incorporation into the fibrous corona (Allan et al., 2020). Thus, we are eager to know if the 387Δ region somehow regulate the Cyclin B1 interaction despite being physically separated by dozens of nanometers, or whether this region interacts with the corona through a distinct mechanism. While Mad1's interaction with the fibrous corona has been known for nearly a decade our unbiased tiling mutagenesis allowed us to identify a key motif within Mad1 that appears to be required for this interaction, which has been missed through traditional analysis leveraging homology searches and protein truncations.

However, there are limitations to the current tiling approach which should be considered. First, library coverage and domain resolution is partly determined by use of the "NGG" protospacer adjacent motif, required by type II CRISPR-Cas system from *Streptococcus pyogenes* (Mali et al., 2013). For example, large gaps in library coverage (148 nt maximum spacing) are due to a lack of NGG sequences (Figure 1C), which may score as false negatives. Utilizing CRISPR nucleases with a more permissive PAM sequence (e.g., xCas9 or Cas9-NG) (Hu et al., 2018; Nishimasu et al., 2018), should enhance tiling screens by allowing more uniform and closer spacing between sgRNAs. Second, in its current form, this approach can only screen genes that are singly required for a particular phenotype. Redundant genes will be missed as our analysis shows (Fig. 2C). Third, while tiling mutagenesis appears robust when assaying aneuploid cell lines, gross genetic alterations (e.g., chromosome rearrangements, gene fusions,

SNPs) may confound analysis of some genes (Munoz et al., 2016). Despite these limitations, we find that, in its current, form essential gene regions are readily identified; and there are a few if any false positives (e.g., "non-essential" regions identified as required for proliferation).

Altogether, this screening strategy is widely applicable and has benefits over other methodologies. Compared with traditional mutation screening the cost and scale of tiling libraries are magnitudes more reasonable than chemical or UV induced mutagenesis strategies in human cells. Similarly, tiling mutagenesis targets endogenous genomic loci making it a better readout of cellular activity than libraries of mutant proteins expressed with highly active promoters from ectopic loci. Tiling mutagenesis screens are also an important advance beyond computational approaches that infer function based on sequence homology because tiling annotations are derived from phenotypic outcomes, we ensure regions identified are truly important for protein function. Additionally, because sgRNA can be targeted nearly anywhere in this functional screen, important protein domains can be identified in regions resistant to homology-based analysis, namely disordered protein regions and rapidly evolving sequences.

In summary, this evidence indicates that tiling mutagenesis is a robust tool to reveal unknown high-resolution protein motifs that contribute to the essential activities of multifunctional proteins. The current data set has revealed key domains in 36 kinetochore proteins, each of which may represent an unexpected missing piece of their biological function, which would otherwise take years of lab work to identify.

METHODS

Mammalian cell culture

HeLa, ARPE^{TERT}, ARPE^{RAS}, HCT116, 293T, and HeLa FlpIn Cells (Etemad et al., 2015; Taylor et al., 1998) cells were grown in a high glucose DMEM (Thermo Fisher Scientific 11-965-

118/Gibco 11965118) supplemented with antibiotic/antimycotic (Thermo Fisher Scientific 15240062) and 10% Fetal Bovine Serum (Thermo Fisher Scientific 26140095) at 37 °C supplemented with 5% CO_2. For microscopy experiments, cells were seeded in 35mm wells containing acid washed 1.5 x 22mm square coverslips (Fisher Scientific 152222) and grown for 12-24 hours prior to transfections or immunostaining and most treatments are outlined in figures. Identity of each cell line was routinely validated by the presence of unique genetic modifications (Frt site, drug resistance genes, expression of transgenes) to ensure cross-contamination did not occur. Cell lines were also regularly screened for mycoplasma contamination using DAPI staining

To entirely depolymerize the microtubule cytoskeleton prior to immunofluorescence staining cells were treated with 10 μM nocodazole (Sigma-Aldrich M1404) for one hour. To test SAC activity cells were instead treated with 500 nM nocodazole (Sigma-Aldrich M1404) for 20 hours prior to fixation.

Library Design and cloning

All possible sgRNA target sequences within the protein coding DNA sequence (CDS) of 48 target genes were identified using the Broad Institutes GPP sgRNA Designer (since redesigned as CRISPick) by inputting gene symbols (Doench et al., 2016; Sanson et al., 2018). CRISPick output sequences for every unique spCas9 PAM (NGG) which cut within the CDS of the consensus or longest transcript for each gene. An initial library was generated from sgRNA that uniquely targeted the gene of interest and no other exonic regions within the genome. To prevent bias in the library near poly-G repeats, sgRNA were removed to ensure a minimum spacing of 5 nt between cut sites within the CDS. In regions where there were >50 nts between

sgRNA we included sequences with limited off-target sites to increase the resolution of screening. One exception being last minute additions, CENP-S and CENP-X where all unique sgRNA were included with no adjustment for minimum or maximum spacing between sgRNAs.

A pooled single stranded DNA 60-mer library containing all sgRNA sequences was synthesize by Twist Bioscences. Oligomers were designed with a universal 20 nucleotides flanking the 5' and 3' with unique sgRNA sequences in the middle 20 nucleotides. The library was PCR amplified using universal primers that annealed to the common flanking sequence and appended homologous sequences at 5' and 3' ends of the PCR product to enable Gibson assembly (New England Biolabs E2611) into pZLCv2_puro_1KF. The vector pZLCv2_puro_1KF was linearized by digestion with restriction enzyme Esp3I and both PCR product and vector were gel purified prior to assembly.

CRISPR/Cas9 Screening

Outgrowth screens were performed as previously described (Toledo et al., 2015). The library of sgRNA-containing donor plasmids, pPAX2, and pMD2.G were co-transfected into 293T cells using polyethyleneimine (PEI, Polysciences 23966-1). Virus-containing supernatant media were harvested 48 hours post transfection and passed through 0.45 µm filters, concentrated by centrifugation, and stored at -80 °C. Each cell line was infected with varying volumes of concentrated virus in the presence of polybrene (Sigma Aldrich 107689) to determine the concentration that conferred survival in puromycin to 30% of cells, representing an MOI of 0.3 where a single infection per cell is the most likely outcome. Three replicates of each cell line were infected at scale to ensure 650x representation of the library and then 24 hours later were exposed to 1 µg/mL puromycin. 72 hours after infection the puromycin containing media was

replaced with drug free media. 96 hours after infection cells were trypsinized and re-seeded to maintain 650x representation, while excess cells were harvested as an initial timepoint. Over the next eight days replicates were sub-cultured to maintain representation and eventually harvest a final population. Genomic DNA was extracted from 5 million cells (~650x representation) in the initial and final populations each using a QiaAMP DNA Blood Purification Mini Kit (Qiagen 51104) and then sgRNA sequences were amplified from each sample using a two-step PCR. For the first step, a 12 cycle PCR was performed using Phusion polymerase (New England Biolabs M0530) to amplify from all the genomic DNA extracted from the 5 million cells per sample (70-80 reactions). For the second step, an 18 cycle PCR was amplified from the pooled first step using primers coding 6 bp Illumina sequencing barcodes used for multiplexing biological samples. The final amplicon was purified from genomic DNA using a Monarch PCR and DNA Cleanup Kit (New England Biolabs T1030) and quantified with a Qubit 2.0 Fluorometer. Samples were then sequenced using an Illumina HiSeq 2500.

Computational Analysis of Tiling Data

Relative changes to the amount of sgRNA sequence detected in final versus initial samples were determined by the CRISPR-SURF package run from the command line (https://github.com/pinellolab/CRISPR-SURF) (Hsu et al., 2018). The SURF package includes 'CRISPR-SURF Count' which outputs logFC values for each sgRNA within the library. This output was used by CRISPR-SURF to deconvolve tiling data and identify the targeted genomic regions that had a negative effect on proliferation relative to non-targeting controls. This output was also the input for ProTiler (https://github.com/MDhewei/protiler) (He et al., 2019) and TiVex and was used to calculate Z scores. In a few instances data were excluded from computational

analysis. CRISPR-SURF Count did not report values for sgRNA containing a TTTT repeat due to their likelihood of causing premature transcriptional termination. No other data were removed from global lists but in the case of genes BIRC5 and KNL1 we generated the library using transcripts containing rare or mutually exclusive exons and when analyzing them at the protein level we mapped results to a more common transcript that does not contain those regions.

TiVex analysis built upon previous approaches for analyzing tiling data that used a Fused Lasso to deconvolve complex signals. The Fused Lasso optimizes the cost function $arg \min_{x}(0.5 * \|y - x\|^2 + \lambda_0\|x\| + \lambda_1\|x_i - x_{i-1}\|)$ (Hsu et al., 2018; Tibshirani and Taylor, 2011), but this was designed for sparse regulatory elements, while functional motifs in proteins are dense. If the sparsity induced penalty is reduced ($\lambda_0 = 0$), then the cost function is equivalent to identifying segmentations and not useful. To balance sparseness we used a Convex Fused Lasso (Parekh and Selesnick, 2015) to deconvolve the data. This approach optimizes the cost function, $\arg\min_{x}(0.5 * \|y - x\|^2 + \lambda_0 \sum_i f(x_i, a_0) + \lambda_1 \sum_i f(x_i - x_{i-1}, a_1))$, where $1 - a_0\lambda_0 - 4a_1\lambda_1 \geq 0$. From deconvolved data, TiVex regions were identified as negatively enriched by comparing the per gene signal to a global average of all genes.

Statistics

Outside of tiling analysis packages, GraphPad Prism version 9.1.0 was used for statistical analysis. Data normality was assessed for all experiments using the D'Agostino & Pearson test. For those with normal distributions, mean values were reported and parametric tests were used. For data sets failing the normality test, median values were reported, and non-parametric tests

were performed for comparisons. Each test (paired, multiple comparison, etc.) is specifically identified in figure legends, and generally error bars represent 95% confidence intervals.

Generation of Modified Human Cell Lines

HeLa FlpIn Trex cells encoding wild type and mutant proteins were generated as previously described (Herman et al., 2020). Briefly, HeLa FlpIn Trex cells were transfected with Flp recombinase (p0G44) and a donor plasmid encoding the protein of interest using lipofectamine 2000 (Invitrogen 11668027) according to manufacturer instructions or PEI (Polysciences 23966-1). 48 hours post transfection, media was supplemented with 500 µg/mL hygromycin (Invitrogen 10687010) and cells were negatively selected for 3 days. Expression of EGFP fusion proteins was then induced by addition of 1 µg/mL doxycycline (Sigma-Aldrich, D9891) and EGFP expressing cells were positively selected by FACS. Doubly selected polyclonal populations were frozen and stored for future experiments.

$ARPE^{TERT}$ and $ARPE^{RAS}$ were generated through single or serial retroviral transductions followed by selecting for cells expressing a single drug resistance gene as previously described (Kendall et al., 2005; Toledo et al., 2014) using methods similar to generating the lentiviral particles containing the CRISPR/Cas9 tiling library detailed above.

Nucleic Acid Reagents

Cas9 resistant and mutant alleles were generated as gBlocks (IDT) and inserted into restriction enzyme linearized pcDNA5 FRT/TO/Hygro by Gibson assembly and sequence verified.

sgRNA:Cas9 mediated gene knockout

Genes were knocked out using 1-2 synthetic sgRNAs (Snythego) in complex with spCas9 (Aldeveron) that were electroporated into cells using a nucleofector system (Lonza) according to published methods (Hoellerbauer et al., 2020a; Hoellerbauer et al., 2020b). Briefly, 120k cells were mixed with either targeting or non-targeting sgRNA:Cas9 complexes in complete SE nucleofector solution. Cell solutions were added to 16-well mini-cuvettes cells and electroporated using program CN-114. Cells were split into two well, one with doxycycline and one without, then cell numbers were assayed 5-7 days later.

Immunoblotting

Expression of Flag- and EGFP-tagged proteins was induced with media containing 1 µg/mL doxycycline (Sigma-Aldrich, D9891) 12-24 hours prior to harvesting. Cells were isolated via trypsinization and then centrifuged. Immunoblotting was performed as previously described (Herman 2020). Cells previously exposed to doxycycline to induce wildtype or mutant protein expression were harvested by trypsinization then resuspended in complete lysis buffer and frozen in liquid nitrogen. Samples were thawed and sonicated with a CL-18 microtip for 20 seconds at 50% maximum power with no pulsing three times using a Fisher Scientific FB50 sonicator. Benzonase nuclease (Millipore E1014) was added to samples and incubated at room temperature for five minutes, then samples were centrifuged at 16,100 x g at 4 °C in a tabletop centrifuge. Relative protein concentrations were determined for clarified lysates, and samples were normalized through dilution. Denatured samples were run on Tris buffered 10 or 12% polyacrylamide gels in a standard Tris-Glycine buffer. Proteins were transferred to a 0.45 µm nitrocellulose membrane (BioRad, 1620115) for two hours at 4 °C in a transfer buffer containing 20% methanol. Membranes were washed in PBS+0.05% Tween-20 (PBS-T) and blocked with

PBS-T+5% non-fat milk and incubated with primary antibodies overnight at 4 °C. Antibodies were diluted in PBS-T by the following factors or to the following concentrations: anti-GAPDH clone 6C5 (Millipore Sigma MAB374) 1 µg/mL; anti-GFP clone JL-8 (Takara 632381) 0.5 µg/mL; anti-Flag clone M2 (Sigma Aldrich F3165) 2 µg/mL. HRP conjugated anti-mouse secondary antibodies (GE Lifesciences, NA931) were diluted 1:10,000 in PBS-T and incubated on membranes for 45 minutes at room temperature. Immunoblots were developed with enhanced chemiluminescence HRP substrates SuperSignal West Dura (Thermo Scientific, 34076) using a ChemiDoc MP system (BioRad).

Immunofluorescent Staining

Upon completion of experimental manipulations, cells grown on coverslips were immediately chemically crosslinked for 15 minutes with 4% PFA diluted from a 16% stock solution (Electron Microscopy Sciences, 15710) with 1x PHEM (60 mM PIPES, 25 mM HEPES, 5 mM EGTA, 8 mM $MgSO_4$) or 1x PHEM+0.5% TritonX100. Coverslips were washed with 1x PHEM+0.5% TritonX100 for 5 minutes, then washed 3 more times with 1x PHEM + 0.1% TritonX100 over 10 minutes. Cells were blocked for 1-2 hours at room temperature in 20% goat serum in 1x PHEM. Anti-centromere protein antibody or ACA (Antibodies Inc. 15-235) was diluted in 20% goat serum at a 1:600 dilution factor. Coverslips were incubated overnight at 4°C in the primary antibody, then washed four times with 1x PHEM + 0.1% TritonX100 over 10 minutes. Goat anti-human secondary antibodies conjugated to AlexaFluor 568 or 647 (Invitrogen) were diluted at 1:300 in 20% boiled goat serum. Coverslips were washed four times with 1x PHEM + 0.1% TritonX100 over 10 minutes, then stained for one minute with 30 ng/mL 4′,6-diamidino-2-phenylindole (DAPI, Invitrogen, D1306) in 1x PHEM. Coverslips were washed two times with 1x PHEM, then

immersed in mounting media (90% glycerol, 20 mM Tris [pH= 8.0], 0.5% w/v N-propyl gallate) on microscope slides and sealed with nail polish.

Microscopy and Image Analysis

Fixed cell images were acquired on either a Deltavision Elite or Deltavision Ultra deconvolution high-resolution microscope, both equipped with a 60x/1.42 PlanApo N oil-immersion objective (Olympus). Slides imaged on the Elite were collected with a Photometrics HQ2 CCD 12-bit camera, while those imaged on the Ultra were equipped with a 16-bit sCMOS detector. On both microscopes, cells were imaged in Z-stacks through the entire cell using 0.2 μm steps. All images were deconvolved using standard settings. Z projections of the maximum signal in the ACA or EGFP channel were exported as TIFFs for analysis by Cell Profiler 4.0.7 (29969450). ACA images were used to identify regions of interest after using a global threshold to remove background signal and distinguishing clumped objects using signal intensity. These signal intensity within these regions was quantified from the EGFP images, and then for background correction the regions were expanded by one pixel along the circumference and signal intensity was again quantified on the EGFP channel. Background intensity was found by subtracting the intensity of the original region from the 1-pixel expanded region. The background intensity per pixel was quantified by dividing the background intensity by the difference in area between two regions. This was then multiplied by area of the original object and subtracted from the intensity of the original object. The mean value per image was then determined and displayed in figures. Representative images displayed from these experiments are projections of the maximum pixel intensity across all Z images. Photoshop was used to crop, make equivalent, linear adjustments to brightness and contrast, and overlay images from different channels.

Computational Analysis of Protein Sequences

Databases of Conserved and Disordered Proteins

PhyloP scores for nucleotides within the genomic regions corresponding to essential regions identified by tilling were downloaded manually from UCSC genome browser and mean values calculated. Disordered regions were identified using the D2P2 database (Oates et al., 2013) based on the agreement of >75% of disorder prediction algorithms. Our comprehensive list of Pfam domains were identified using gene IDs.

Multiple Sequence Alignments

Eukaryotic Mad1 orthologs were identified previously (van Hooff et al., 2017) and a multiple sequence alignment of the entire proteins was generated with ClustalOmega (Sievers et al., 2011) default parameters and displayed in JalView 1.8 (Waterhouse et al., 2009).

FIGURE LENGENDS

Figure 1. Design and execution of CRISPR/Cas9 tiling screen. (A) Schematic of CRISPR-Cas9 tiling screen and phenotypic readout. The percent of population containing each genotype is calculated assuming unbiased repair (in-frame repairs occur at 33% of edits) or based on deep sequencing (in-frame repairs occur at 20% of edits)(Chakrabarti et al., 2019). (B) Cartoon representation of kinetochore protein complexes assembled on duplicated chromosomes where they contribute to microtubule binding while also sensing and communicating attachment state. (C) Characteristics of tiling library targeting mitotic factors. (D) Schematic of proliferation-based screening approach using lentiviral particles to insert both spCas9 and sgRNA sequences into genomic DNA. This allows next generation sequencing of sgRNA sequences themselves to serve as an indirect readout of their effect on cell growth.

Figure 2. CRISPR/Cas9 tiling screen is technically and biologically reproducible, and sgRNA effects on proliferation are associated with targeting functional protein domains. (A) Each sgRNA's average Z score from three replicates of $ARPE^{TERT}$, $ARPE^{RAS}$, and HCT116 cells are plotted relative to each sgRNA's average Z score from three replicates of HeLa cells. These data exclude non-targeting controls, and the dashed line (y = x) is plotted for reference. (B) Correlation matrix and heatmap for all targeting sgRNA in each cell line with Pearson correlation coefficients displayed. (C). Table of 48 genes tiled in the screen, "Tilling" column reports the number of cell lines in which >9.5% of sgRNA targeting that gene had a Z score < -1.0, indicating that the gene overall was important for cell proliferation. DepMap column reports the number of cell lines screened with a 'dependency score' < -0.5 at the time of writing. (D) Average Z score for non-targeting controls in all four cell lines and targeting sgRNAs that

contain the recently reported pyrimidine rich (Y rich) sequence at their 3' end (Graf et al., 2019). Box plots show median, quartiles, and range of all sgRNAs (E) Targeting sgRNA, minus those containing the Y rich sequence were binned based on which genomic or protein features they targeted. 'In-frame edits' were predicted using Indelphi (Shen et al., 2018) and binned into groups above and below the median likelihood of in-frame edits (22.4%) or above and below the overall likelihood of in-frame edits (50%). Box plots show median, quartiles, and range of the average Z score for each sgRNAs among four cell lines. Mann Whitney tests were used to determine P-values.

Figure 3. CRISPR/Cas9 tiling data analyzed at high or low resolution identifies previously known and uncharacterized functional regions including those in evolutionarily divergent and disordered proteins. (A) Median Z score from four cell lines for each sgRNA is mapped to where it targets the CDS of KIF18A (gray vertical bars with 95% confidence interval). Amino acid regions identified by SURF, ProTiler or TiVex as important for cell proliferation are shown as colored in each track corresponding to coordinates at bottom. Kif18A functional regions identified in literature are shown in a similar manner. Gray horizonal line shows the average value of all non-targeting controls representing normal levels of proliferation. (B) Table describing the characteristics of essential regions identified by analyzing tilling data with SURF, ProTiler, or TiVex in the 36 target genes previously identified (Fig. 2C). (C) Crystal structures of budding yeast Bub3 bound to a peptide of yeast Bub1 (PDB: 2I3S) are colored based on the homologous regions identified by TiVex (green) and SURF (blue) as important for proliferation (Larsen et al., 2007). Two tryptophan residues absolutely required for the protein-protein interaction are shown in blowups. (D) PhyloP values (based on the alignment of 100 vertebrate orthologs) for each

nucleotide within SURF, ProTiler, or TiVex regions were averaged. Bars represent the median, each dot is a region, and 'conserved' corresponds to a value of $P < 0.05$. (E) Regions identified by SURF, ProTiler, and TiVex, as required for proliferation were grouped by their overlap with conserved PFAM domains (El-Gebali et al., 2019; Mistry et al., 2013), functional regions identified in the literature, or predicted disordered domains (Oates et al., 2013). (F) Pairwise analysis of overlap between regions identified by each method. Graph displays the number of SURF regions within a single gene (left two columns, blue) that overlaps by at least 1 amino acid with ProTiler or TiVex.

Figure 4. Functional validation and characterization of 11 high-resolution regions within 5 genes identified by tiling (A) Schematic of generating cell lines in which expression of wildtype or mutant proteins is induced by doxycycline and the CRISPR/Cas9-based complementation approach used for functional validation. (B-D) Tiling profile, validation of proliferation phenotype, and assay of protein stability for (B) SKA3/Ska3, (C) ZNF207/BuGZ, and (D) CENPK/Cenp-K. Tiling profiles are the same as Fig. 3 while also showing regions that were deleted. Cell proliferation was assayed as the cell number after knockout of endogenous protein in the presence (green) or absence (blue) of doxycycline in each cell line. The overlap of deletion and analysis methods shown as P = ProTiler, S = CRISPR-SURF, and T = TiVex. Cell numbers were normalized to the same cell line electroporated with a non-targeting control. Each dot is a biological replicate with bars showing median values and 95% confidence intervals. Mann Whitney tests were used to determine P-values. Steady state protein levels of wildtype and mutant proteins was assayed by immunoblot in the presence of endogenous protein, using GAPDH as a loading control and exposed for a shorter interval. (More examples in Fig. S4).

Figure 5. Tiling MAD1L1/Mad1 reveals a region contributing to prolonged activation of the SAC and kinetochore localization. (A) Tiling profile for MAD1L1/Mad1, displayed the same as Fig. 4. (B) Normalized cell number after knockout of endogenous protein in the presence (green) or absence (blue) of doxycycline in each cell line. Each dot is a biological replicate and Mann Whitney tests were used to determine P-values. (C) Steady state protein levels of wildtype and mutant proteins were assayed by immunoblot in the presence of endogenous protein, using GAPDH as a loading control and exposed for a shorter interval. (D) The ability wildtype or mutant proteins to maintain a prolonged SAC arrest in the presence of endogenous Mad1 was assayed by treating cells with nocodazole for 20 hours and determining the percent of cells in mitosis based on chromatin morphology. Data in 'No Dox' are determined from all five cell lines not exposed to doxycycline. Each dot represents a biological replicate and Dunn's multiple comparisons test was used to determine P-values. (E) Kinetochore association of EGFP-Mad1 wildtype and mutant fusion proteins was determined by the EGFP overlap with anti-centromere antibody (ACA) in the presence of endogenous Mad1 and nocodazole. Representative images on left with quantifications on right. Each dot represents the average kinetochore signal from a single cell, cells from three biological replicates are colored differently. Dunn's multiple comparisons test was used to determine P-values. Scale bars are 1 μm, all averages and error bars in figure are median values and 95% confidence intervals.

Figure 6. Mad1$^{387\Delta}$ contributes to kinetochore localization independent of RLK motif that mediates Bub1 interaction. (A) Schematic of Mad1$^{387\Delta}$ deletion mutant (removal of amino acids 387-402) and Mad1^{R617A} point mutant. (B) The ability wildtype or mutant proteins to maintain a

prolonged SAC arrest in the presence of endogenous Mad1 was assayed by treating cells with nocodazole for 20 hours and determining the percent of cells in mitosis based on chromatin morphology. Data in 'No Dox' are determined from all four cell lines not exposed to doxycycline. Each dot represents a biological replicate and Dunn's multiple comparisons test was used to determine P-values. (C) Kinetochore association of EGFP-Mad1 wildtype and mutant fusion proteins was determined by the EGFP overlap with anti-centromere antibody (ACA) in the presence of endogenous Mad1 and nocodazole. Representative images on left with quantifications on right. Each dot represents the average kinetochore signal from a single cell, cells from three biological replicates are colored differently. Dunn's multiple comparisons test was used to determine P-values. Scale bars are 1 μm, all averages and error bars in figure are median values and 95% confidence intervals.

Figure S1. Tiling library design strategy benefits from a lack of positional repair bias. (A) Scheme for choosing sgRNA included in the library. (B-C) All sgRNAs in the library were binned based on the position they target (% of CDS). (B) For each bin, in each gene, the percent of sgRNA predicted to cause 1nt indels based on sequence constraints (Chakrabarti et al., 2019) are plotted. Violin plots show median (solid lines), quartiles (dotted lines) and range of bins for all genes. (C) For each bin, in each gene, the mean percent likelihood of forming a frameshift mutation based on sequence constraints (Shen et al., 2018) are plotted. Violin plots show median (solid lines), quartiles (dotted lines) and range. (D-E) The standard deviation across each bin in (B) and (C) for each gene was calculated and is plotted as a single dot. Deviations are plotted relative to the CDS length (D) or the total number of sgRNAs targeting the gene (E). Data are fitted with a single-phase exponential decay curve.

Figure S2. Tiling library is reproducible and differences from genome-wide data have biological causes. (A) Each sgRNA's average Z score from three replicates of $ARPE^{TERT}$, $ARPE^{RAS}$, and HCT116 cells are plotted relative to each sgRNA's average Z score from three replicates of HeLa cells. These data exclude non-targeting controls and only show sgRNA in the bottom quartile of Z-scores from HeLa cells. The dashed line (y = x) is plotted for reference. (B) Correlation matrix and heatmap for the bottom quartile of targeting sgRNA in each cell line with Pearson correlation coefficients displayed. (C) Z scores for each sgRNA targeting genes where disagreement between tiling and DepMap data disagree (MIS12 or MAD1L1), with sgRNA sequences present in both libraries colored teal. (D) Numerous transcript models of CENPM and SKA2 (right) along with their relative expression in cultured cells from public RNA-seq datasets (right). The location targeted by each sgRNA in the DepMap (Avana) library shown as teal bars indicating they target rarely used exons.

Figure S3. More evidence of SURF, ProTiler, and TiVex resolution. (A) (top) Tilling profile for CKAP5/chTOG as shown in Fig. 3. (bottom) Crystal structure of a single budding yeast CKAP domain, TOG2 (~aa250-500) bound to a tubulin dimer (PDB: 4U3J) are colored based on the homologous regions identified by TiVex (blue) and SURF (purple) as important for proliferation (Ayaz et al., 2014). Four residues absolutely required for the protein-protein interaction are shown in blowups. (B) Percent of each protein sequence identified by the three analysis methods or Pfam or within literature. Average values reported in Fig. 3B. (C) The number of regions identified per gene by each analysis method. Specific genes showing key differences between SURF and ProTiler are colored uniquely in all three columns.

Figure S4. Functional validation and characterization of 11 high-resolution regions within 5 genes identified by tiling, continued. (A-B) Tiling profile, validation of proliferation phenotype, and assay of protein stability for (A) CENPH/Cenp-H and (B) SGO1/Sgo. Tiling profiles are the same as Fig. 3 while also showing regions that were deleted. Cell proliferation was assayed as the cell number after knockout of endogenous protein in the presence (green) or absence (blue) of doxycycline in each cell line. The overlap of deletion and analysis methods shown as P = ProTiler, S = CRISPR-SURF, and T = TiVex. Cell numbers were normalized to the same cell line electroporated with a non-targeting control. Each dot is a biological replicate with bars showing median values and 95% confidence intervals. Mann Whitney tests were used to determine P-values. Steady state protein levels of wildtype and mutant proteins was assayed by immunoblot in the presence of endogenous protein, using GAPDH as a loading control and exposed for a shorter interval.

Figure S5. Mutations in Mad1 do not affect localization to kinetochores immediately following nuclear envelope breakdown. Kinetochore association of EGFP-Mad1 wildtype and mutant fusion proteins was determined by the EGFP overlap with anti-centromere antibody (ACA) in the presence of endogenous Mad1 immediately after nuclear envelope breakdown. Representative images on left with quantifications on right. Each dot represents the average kinetochore signal from a single cell, cells from three biological replicates are colored differently. Dunn's multiple comparisons test was used to determine P-values. Scale bars are 1 μm, all averages and error bars in figure are median values and 95% confidence intervals.

Figure S6. Mad1 region of interest (aa387-402) is divergent among most eukaryotes but shows some conservation in coelomates. (A) Previously identified (van Hooff et al., 2017) putative Mad1 orthologs were aligned using ClustalO. Multiple alignments are driven by other regions of the protein and 387-402 has many gaps. (B) Multiple sequence alignments were restricted to Mad1 orthologs from species that also encode members of the RZZ complex (primarily metazoan). (C) Multiple sequence alignments for proteins in (B) but restricting alignment to just the amino acids homologous to human Mad1 387-402. The alignment was further restricted to just coelomate species to identify some aspect of conservation (right).

ACKNOWLEDGEMENTS

We thank members of the Biggins and Paddison Labs for helpful discussions, and Bruce Clurman for providing HCT116 cells and Jennifer DeLuca for plasmids encoding SGO1 and MAD1L1. This work was supported by the: American Cancer Society (ACS-RSG-14-056-01) (PP); NIH (R01CA190957; R01NS119650, P30CA15704) (PP); NIH (R01GM064386) (SB); and Robert J. Kleberg, Jr. and Helen C. Kleberg Foundation (PP). SB is an investigator of the Howard Hughes Medical Institute.

AUTHORS' CONTRIBUTIONS

Conceptualization, JH, SB, and PP; *Methodology*, JH, JZ, SB, PP; *Validation*, JH, LC; *Investigation*, JH ; *Formal Analysis*, JH, SA, JZ, and PP; *Writing - Original Draft*, JH, JZ, SB, and PP; *Writing - Review & Editing*, JH, SB, and PP; *Visualization*, JH, SA, JZ, PP; *Supervision*, JH, SB, and PP; *Funding Acquisition*, SB and PP.

COMPETING INTERESTS

Authors declare no competing interests.

REFERENCES

Allan, L.A., Camacho Reis, M., Ciossani, G., Huis In 't Veld, P.J., Wohlgemuth, S., Kops, G.J., Musacchio, A., and Saurin, A.T. (2020). Cyclin B1 scaffolds MAD1 at the kinetochore corona to activate the mitotic checkpoint. EMBO J *39*, e103180.

Ayaz, P., Munyoki, S., Geyer, E.A., Piedra, F.A., Vu, E.S., Bromberg, R., Otwinowski, Z., Grishin, N.V., Brautigam, C.A., and Rice, L.M. (2014). A tethered delivery mechanism explains the catalytic action of a microtubule polymerase. eLife *3*, e03069.

Bateman, A., Coggill, P., and Finn, R.D. (2010). DUFs: families in search of function. Acta Crystallogr Sect F Struct Biol Cryst Commun *66*, 1148-1152.

Brady, D.M., and Hardwick, K.G. (2000). Complex formation between Mad1p, Bub1p and Bub3p is crucial for spindle checkpoint function. Curr Biol *10*, 675-678.

Breuza, L., Poux, S., Estreicher, A., Famiglietti, M.L., Magrane, M., Tognolli, M., Bridge, A., Baratin, D., Redaschi, N., and UniProt, C. (2016). The UniProtKB guide to the human proteome. Database : the journal of biological databases and curation *2016*.

Canver, M.C., Smith, E.C., Sher, F., Pinello, L., Sanjana, N.E., Shalem, O., Chen, D.D., Schupp, P.G., Vinjamur, D.S., Garcia, S.P.*, et al.* (2015). BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. Nature *527*, 192-197.

Canver, M.C., Tripathi, P., Bullen, M.J., Olshansky, M., Kumar, Y., Wong, L.H., Turner, S.J., Lessard, S., Pinello, L., Orkin, S.H.*, et al.* (2020). A saturating mutagenesis CRISPR-Cas9-mediated functional genomic screen identifies cis- and trans-regulatory elements of Oct4 in murine ESCs. J Biol Chem *295*, 15797-15809.

Chakrabarti, A.M., Henser-Brownhill, T., Monserrat, J., Poetsch, A.R., Luscombe, N.M., and Scaffidi, P. (2019). Target-Specific Precision of CRISPR-Mediated Genome Editing. Mol Cell *73*, 699-713 e696.

Consortium, E.P. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57-74.

De Antoni, A., Pearson, C.G., Cimini, D., Canman, J.C., Sala, V., Nezi, L., Mapelli, M., Sironi, L., Faretta, M., Salmon, E.D.*, et al.* (2005). The Mad1/Mad2 complex as a template for Mad2 activation in the spindle assembly checkpoint. Curr Biol *15*, 214-225.

Ding, Y., Hubert, C.G., Herman, J., Corrin, P., Toledo, C.M., Skutt-Kakaria, K., Vazquez, J., Basom, R., Zhang, B., Risler, J.K.*, et al.* (2013). Cancer-Specific requirement for BUB1B/BUBR1 in human brain tumor isolates and genetically transformed cells. Cancer discovery *3*, 198-211.

Doench, J.G., Fusi, N., Sullender, M., Hegde, M., Vaimberg, E.W., Donovan, K.F., Smith, I., Tothova, Z., Wilen, C., Orchard, R.*, et al.* (2016). Optimized sgRNA design to maximize activity and minimize off-target effects of CRISPR-Cas9. Nat Biotechnol *34*, 184-191.

El-Gebali, S., Mistry, J., Bateman, A., Eddy, S.R., Luciani, A., Potter, S.C., Qureshi, M., Richardson, L.J., Salazar, G.A., Smart, A.*, et al.* (2019). The Pfam protein families database in 2019. Nucleic Acids Res *47*, D427-D432.

Etemad, B., Kuijt, T.E., and Kops, G.J. (2015). Kinetochore-microtubule attachment is sufficient to satisfy the human spindle assembly checkpoint. Nature communications *6*, 8987.

Feldman, H., Toledo, C., Arora, S., Hoellerbauer, P., Corrin, P., Carter, L., Kufeld, M., Bolouri, H., Basom, R., Delrow, J.*, et al.* (2019). Neural G0: a quiescent-like state found in neuroepithelial-derived cells and glioma. bioRxiv, 446344.

Fulco, C.P., Munschauer, M., Anyoha, R., Munson, G., Grossman, S.R., Perez, E.M., Kane, M., Cleary, B., Lander, E.S., and Engreitz, J.M. (2016). Systematic mapping of functional enhancer-promoter connections with CRISPR interference. Science *354*, 769-773.

Gossen, M., and Bujard, H. (1992). Tight control of gene expression in mammalian cells by tetracycline-responsive promoters. Proc Natl Acad Sci U S A *89*, 5547-5551.

Graf, R., Li, X., Chu, V.T., and Rajewsky, K. (2019). sgRNA Sequence Motifs Blocking Efficient CRISPR/Cas9-Mediated Gene Editing. Cell Rep *26*, 1098-1103 e1093.

Hara, M., and Fukagawa, T. (2020). Dynamics of kinetochore structure and its regulations during mitotic progression. Cellular and molecular life sciences : CMLS *77*, 2981-2995.

Hartlerode, A.J., and Scully, R. (2009). Mechanisms of double-strand break repair in somatic mammalian cells. Biochem J *423*, 157-168.

He, W., Zhang, L., Villarreal, O.D., Fu, R., Bedford, E., Dou, J., Patel, A.Y., Bedford, M.T., Shi, X., Chen, T.*, et al.* (2019). De novo identification of essential protein domains from CRISPR-Cas9 tiling-sgRNA knockout screens. Nature communications *10*, 4541.

Herman, J.A., Miller, M.P., and Biggins, S. (2020). chTOG is a conserved mitotic error correction factor. eLife *9*.

Hoellerbauer, P., Kufeld, M., Arora, S., Wu, H.J., Feldman, H.M., and Paddison, P.J. (2020a). A simple and highly efficient method for multi-allelic CRISPR-Cas9 editing in primary cell cultures. Cancer Rep (Hoboken) *3*, e1269.

Hoellerbauer, P., Kufeld, M., and Paddison, P.J. (2020b). Efficient Multi-Allelic Genome Editing of Primary Cell Cultures via CRISPR-Cas9 Ribonucleoprotein Nucleofection. Current protocols in stem cell biology *54*, e126.

Hsu, J.Y., Fulco, C.P., Cole, M.A., Canver, M.C., Pellin, D., Sher, F., Farouni, R., Clement, K., Guo, J.A., Biasco, L.*, et al.* (2018). CRISPR-SURF: discovering regulatory elements by deconvolution of CRISPR tiling screen data. Nat Methods *15*, 992-993.

Hu, J.H., Miller, S.M., Geurts, M.H., Tang, W., Chen, L., Sun, N., Zeina, C.M., Gao, X., Rees, H.A., Lin, Z.*, et al.* (2018). Evolved Cas9 variants with broad PAM compatibility and high DNA specificity. Nature *556*, 57-63.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E. (2012). A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science *337*, 816-821.

Kendall, S.D., Linardic, C.M., Adam, S.J., and Counter, C.M. (2005). A network of genetic events sufficient to convert normal human cells to a tumorigenic state. Cancer Res *65*, 9824-9828.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res *12*, 996-1006.

Kim, S., Sun, H., Tomchick, D.R., Yu, H., and Luo, X. (2012). Structure of human Mad1 C-terminal domain reveals its involvement in kinetochore targeting. Proc Natl Acad Sci U S A *109*, 6549-6554.

Labbadia, J., and Morimoto, R.I. (2015). The biology of proteostasis in aging and disease. Annual review of biochemistry *84*, 435-464.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W.*, et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Lara-Gonzalez, P., Kim, T., Oegema, K., Corbett, K., and Desai, A. (2021). A tripartite mechanism catalyzes Mad2-Cdc20 assembly at unattached kinetochores. Science *371*, 64-67.

Larsen, N.A., Al-Bassam, J., Wei, R.R., and Harrison, S.C. (2007). Structural analysis of Bub3 interactions in the mitotic spindle checkpoint. Proc Natl Acad Sci U S A *104*, 1201-1206.

Lieber, M.R., Ma, Y., Pannicke, U., and Schwarz, K. (2003). Mechanism and regulation of human non-homologous DNA end-joining. Nat Rev Mol Cell Biol *4*, 712-720.

London, N., and Biggins, S. (2014a). Mad1 kinetochore recruitment by Mps1-mediated phosphorylation of Bub1 signals the spindle checkpoint. Genes Dev *28*, 140-152.

London, N., and Biggins, S. (2014b). Signalling dynamics in the spindle checkpoint response. Nat Rev Mol Cell Biol *15*, 736-747.

Makarova, K.S., Aravind, L., Wolf, Y.I., and Koonin, E.V. (2011). Unification of Cas protein families and a simple scenario for the origin and evolution of CRISPR-Cas systems. Biology direct *6*, 38.

Mali, P., Esvelt, K.M., and Church, G.M. (2013). Cas9 as a versatile tool for engineering biology. Nat Methods *10*, 957-963.

Meyers, R.M., Bryan, J.G., McFarland, J.M., Weir, B.A., Sizemore, A.E., Xu, H., Dharia, N.V., Montgomery, P.G., Cowley, G.S., Pantel, S.*, et al.* (2017). Computational correction of copy number effect improves specificity of CRISPR-Cas9 essentiality screens in cancer cells. Nat Genet *49*, 1779-1784.

Michlits, G., Jude, J., Hinterndorfer, M., de Almeida, M., Vainorius, G., Hubmann, M., Neumann, T., Schleiffer, A., Burkard, T.R., Fellner, M.*, et al.* (2020). Multilayered VBC score predicts sgRNAs that efficiently generate loss-of-function alleles. Nat Methods *17*, 708-716.

Mimori-Kiyosue, Y., Grigoriev, I., Lansbergen, G., Sasaki, H., Matsui, C., Severin, F., Galjart, N., Grosveld, F., Vorobjev, I., Tsukita, S.*, et al.* (2005). CLASP1 and CLASP2 bind to EB1 and regulate microtubule plus-end dynamics at the cell cortex. J Cell Biol *168*, 141-153.

Mistry, J., Coggill, P., Eberhardt, R.Y., Deiana, A., Giansanti, A., Finn, R.D., Bateman, A., and Punta, M. (2013). The challenge of increasing Pfam coverage of the human proteome. Database : the journal of biological databases and curation *2013*, bat023.

Munoz, D.M., Cassiani, P.J., Li, L., Billy, E., Korn, J.M., Jones, M.D., Golji, J., Ruddy, D.A., Yu, K., McAllister, G.*, et al.* (2016). CRISPR Screens Provide a Comprehensive Assessment of Cancer Vulnerabilities but Generate False-Positive Hits for Highly Amplified Genomic Regions. Cancer discovery *6*, 900-913.

Musacchio, A. (2015). The Molecular Biology of Spindle Assembly Checkpoint Signaling Dynamics. Curr Biol *25*, R1002-1018.

Nguyen, A.L., Fadel, M.D., and Cheeseman, I.M. (2021). Differential requirements for the CENP-O complex reveal parallel PLK1 kinetochore recruitment pathways. Mol Biol Cell *32*, 712-721.

Nishimasu, H., Shi, X., Ishiguro, S., Gao, L., Hirano, S., Okazaki, S., Noda, T., Abudayyeh, O.O., Gootenberg, J.S., Mori, H.*, et al.* (2018). Engineered CRISPR-Cas9 nuclease with expanded targeting space. Science *361*, 1259-1262.

O'Gorman, S., Fox, D.T., and Wahl, G.M. (1991). Recombinase-mediated gene activation and site-specific integration in mammalian cells. Science *251*, 1351-1355.

Oates, M.E., Romero, P., Ishida, T., Ghalwash, M., Mizianty, M.J., Xue, B., Dosztanyi, Z., Uversky, V.N., Obradovic, Z., Kurgan, L.*, et al.* (2013). D(2)P(2): database of disordered protein predictions. Nucleic Acids Res *41*, D508-516.

Parekh, A., and Selesnick, I. (2015). Convex fused lasso denoising with non-convex regularization and its use for pulse detection. IEEE Xplore *10.1109/SPMB.2015.7405474*.

Pereira, A.L., Pereira, A.J., Maia, A.R., Drabek, K., Sayas, C.L., Hergert, P.J., Lince-Faria, M., Matos, I., Duque, C., Stepanova, T.*, et al.* (2006). Mammalian CLASP1 and CLASP2 cooperate to ensure mitotic fidelity by regulating spindle and kinetochore function. Mol Biol Cell *17*, 4526-4542.

Petrovic, A., Keller, J., Liu, Y., Overlack, K., John, J., Dimitrova, Y.N., Jenni, S., van Gerwen, S., Stege, P., Wohlgemuth, S.*, et al.* (2016). Structure of the MIS12 Complex and Molecular Basis of Its Interaction with CENP-C at Human Kinetochores. Cell *167*, 1028-1040 e1015.

Pettersen, E.F., Goddard, T.D., Huang, C.C., Couch, G.S., Greenblatt, D.M., Meng, E.C., and Ferrin, T.E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem *25*, 1605-1612.

Piano, V., Alex, A., Stege, P., Maffini, S., Stoppiello, G.A., Huis In 't Veld, P.J., Vetter, I.R., and Musacchio, A. (2021). CDC20 assists its catalytic incorporation in the mitotic checkpoint complex. Science *371*, 67-71.

Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res *20*, 110-121.

Punta, M., Coggill, P.C., Eberhardt, R.Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J.*, et al.* (2012). The Pfam protein families database. Nucleic Acids Res *40*, D290-301.

Rodriguez-Bravo, V., Maciejowski, J., Corona, J., Buch, H.K., Collin, P., Kanemaki, M.T., Shah, J.V., and Jallepalli, P.V. (2014). Nuclear pores protect genome integrity by assembling a premitotic and Mad1-dependent anaphase inhibitor. Cell *156*, 1017-1031.

Rodriguez-Rodriguez, J.A., Lewis, C., McKinley, K.L., Sikirzhytski, V., Corona, J., Maciejowski, J., Khodjakov, A., Cheeseman, I.M., and Jallepalli, P.V. (2018). Distinct Roles of RZZ and Bub1-KNL1 in Mitotic Checkpoint Signaling and Kinetochore Expansion. Curr Biol *28*, 3422-3429 e3425.

Sanjana, N.E., Shalem, O., and Zhang, F. (2014). Improved vectors and genome-wide libraries for CRISPR screening. Nat Methods *11*, 783-784.

Sanson, K.R., Hanna, R.E., Hegde, M., Donovan, K.F., Strand, C., Sullender, M.E., Vaimberg, E.W., Goodale, A., Root, D.E., Piccioni, F.*, et al.* (2018). Optimized libraries for CRISPR-Cas9 genetic screens with multiple modalities. Nature communications *9*, 5416.

Shen, M.W., Arbab, M., Hsu, J.Y., Worstell, D., Culbertson, S.J., Krabbe, O., Cassa, C.A., Liu, D.R., Gifford, D.K., and Sherwood, R.I. (2018). Predictable and precise template-free CRISPR editing of pathogenic variants. Nature *563*, 646-651.

Shi, J., Wang, E., Milazzo, J.P., Wang, Z., Kinney, J.B., and Vakoc, C.R. (2015). Discovery of cancer drug targets by CRISPR-Cas9 screening of protein domains. Nat Biotechnol *33*, 661-667.

Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Soding, J.*, et al.* (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular systems biology *7*, 539.

Simeonov, D.R., Gowen, B.G., Boontanrart, M., Roth, T.L., Gagnon, J.D., Mumbach, M.R., Satpathy, A.T., Lee, Y., Bray, N.L., Chan, A.Y.*, et al.* (2017). Discovery of stimulation-responsive immune enhancers with CRISPR activation. Nature *549*, 111-115.

Taylor, S.S., Ha, E., and McKeon, F. (1998). The human homologue of Bub3 is required for kinetochore localization of Bub1 and a Mad3/Bub1-related protein kinase. J Cell Biol *142*, 1-11.

Tibshirani, R.J., and Taylor, J. (2011). The solution path of the generalized lasso. Ann Statist *39*, 1335-1371.

Toledo, C.M., Ding, Y., Hoellerbauer, P., Davis, R.J., Basom, R., Girard, E.J., Lee, E., Corrin, P., Hart, T., Bolouri, H.*, et al.* (2015). Genome-wide CRISPR-Cas9 Screens Reveal Loss of Redundancy between PKMYT1 and WEE1 in Glioblastoma Stem-like Cells. Cell Rep *13*, 2425-2439.

Toledo, C.M., Herman, J.A., Olsen, J.B., Ding, Y., Corrin, P., Girard, E.J., Olson, J.M., Emili, A., DeLuca, J.G., and Paddison, P.J. (2014). BuGZ is required for Bub3 stability, Bub1 kinetochore function, and chromosome alignment. Dev Cell *28*, 282-294.

Tsherniak, A., Vazquez, F., Montgomery, P.G., Weir, B.A., Kryukov, G., Cowley, G.S., Gill, S., Harrington, W.F., Pantel, S., Krill-Burger, J.M.*, et al.* (2017). Defining a Cancer Dependency Map. Cell *170*, 564-576 e516.

van Hooff, J.J., Tromer, E., van Wijk, L.M., Snel, B., and Kops, G.J. (2017). Evolutionary dynamics of the kinetochore network in eukaryotes as revealed by comparative genomics. EMBO Rep *18*, 1559-1571.

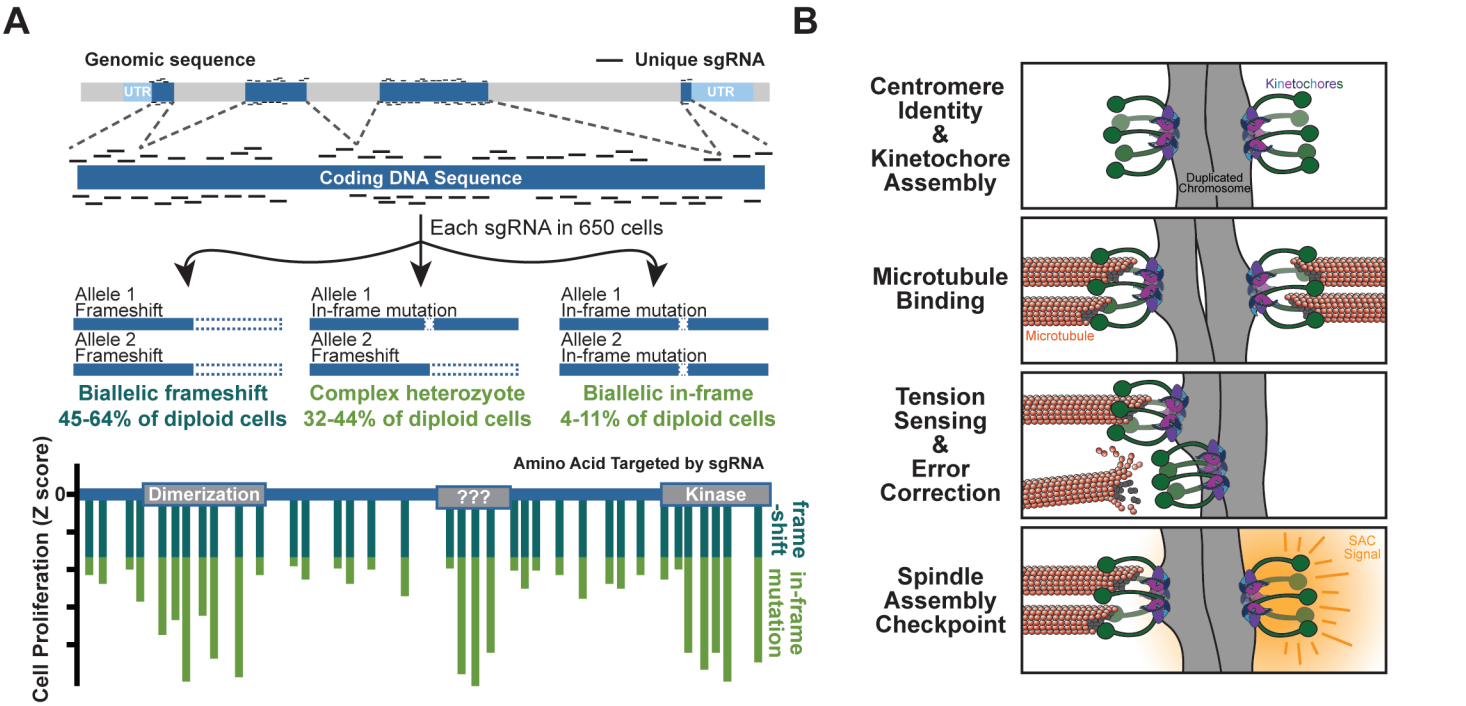
van Overbeek, M., Capurso, D., Carter, M.M., Thompson, M.S., Frias, E., Russ, C., Reece-Hoyes, J.S., Nye, C., Gradia, S., Vidal, B.*, et al.* (2016). DNA Repair Profiling Reveals Nonrandom Outcomes at Cas9-Mediated Breaks. Mol Cell *63*, 633-646.

Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M., and Barton, G.J. (2009). Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics *25*, 1189-1191.

Zhang, G., Kruse, T., Guasch Boldu, C., Garvanska, D.H., Coscia, F., Mann, M., Barisic, M., and Nilsson, J. (2019). Efficient mitotic checkpoint signaling depends on integrated activities of Bub1 and the RZZ complex. EMBO J *38*.

Zhang, G., Lischetti, T., Hayward, D.G., and Nilsson, J. (2015). Distinct domains in Bub1 localize RZZ and BubR1 to kinetochores to regulate the checkpoint. Nature communications *6*, 7162.

Figure 1



C

Library Targets	
Genes	48
sgRNA	6500
Mean Spacing	18 nt
Median Spacing	14 nt
Minimum Spacing	2 nt
Maximum Spacing	148 nt

Library Controls	
Non Targeting	601
Negative sgRNAs	18
Positive sgRNAs	28
Total sgRNA	7147

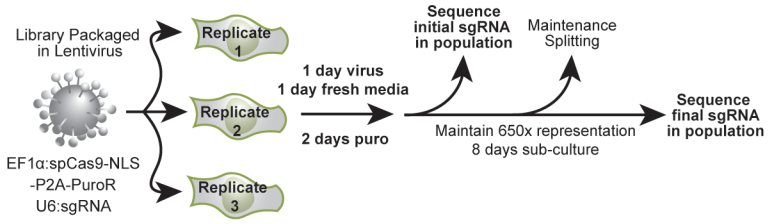
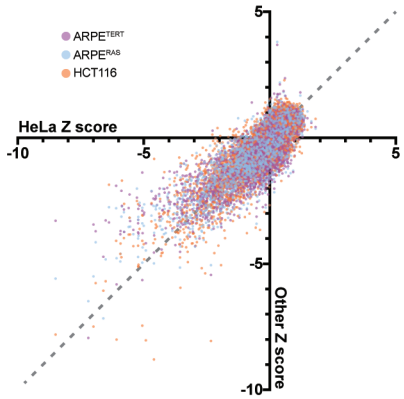
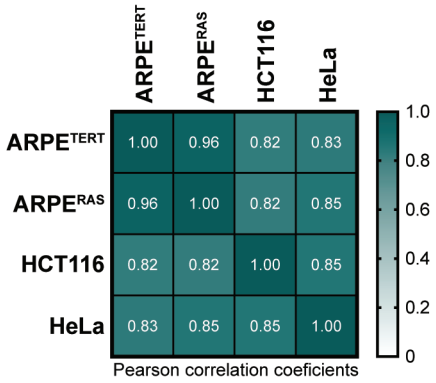


Figure 2

A



B



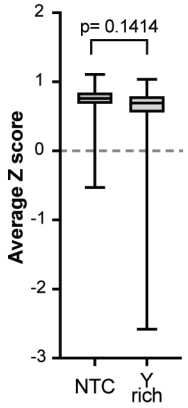
C

Gene	Tiling	DepMap	Gene	Tiling	DepMap
APC	0/4	347/808	KIF2C	2/4	313/808
AURKB	4/4	795/796	KNL1	4/4	643/796
BIRC5	4/4	808/808	KNTC1	2/4	244/808
BUB1	4/4	720/796	MAD1L1	4/4	2/808
BUB1B	4/4	805/808	MAD2L1	4/4	738/808
BUB3	4/4	808/808	MAPRE1	1/4	203/808
CENPC	4/4	665/796	MAPRE2	0/4	0/808
CENPE	3/4	630/808	MAPRE3	0/4	0/808
CENPF	2/4	120/808	MIS12	0/4	738/808
CENPH	4/4	736/799	NDC80	4/4	807/808
CENPI	4/4	712/793	NSL1	4/4	557/796
CENPK	4/4	794/796	NUF2	4/4	805/808
CENPM	3/4	0/796	PLK1	4/4	796/796
CENPS	1/4	27/808	PMF1	4/4	785/796
CENPT	4/4	773/808	SGO1	4/4	775/808
CENPW	4/4	796/796	SKA1	4/4	520/808
CENPX	2/4	291/808	SKA2	4/4	11/808
CKAP5	4/4	806/808	SKA3	4/4	731/808
CLASP1	0/4	N/A	SPDL1	4/4	780/808
CLASP2	0/4	N/A	TTK	3/4	786/808
CLIP1	0/4	0/808	ZNF207	4/4	791/808
DSN1	4/4	781/808	ZW10	1/4	567/808
KIF18A	2/4	757/808	ZWILCH	2/4	13/808
KIF2A	0/4	12/808	ZWINT	1/4	461/808

Tiling = number of cell lines where more than 9.5% of sgRNA have Z score < -1.0

DepMap = number of cell lines where probability of dependency score is > 0.5

D



E

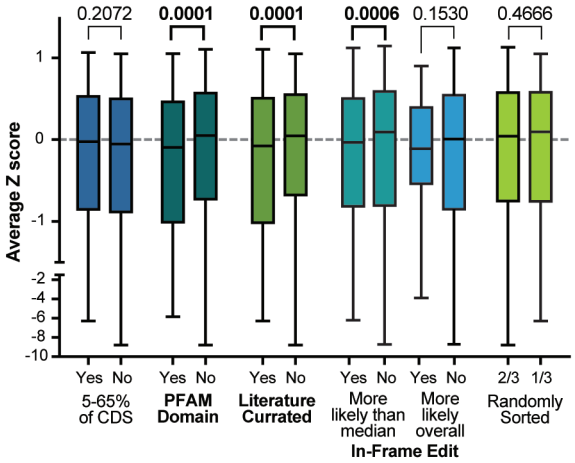
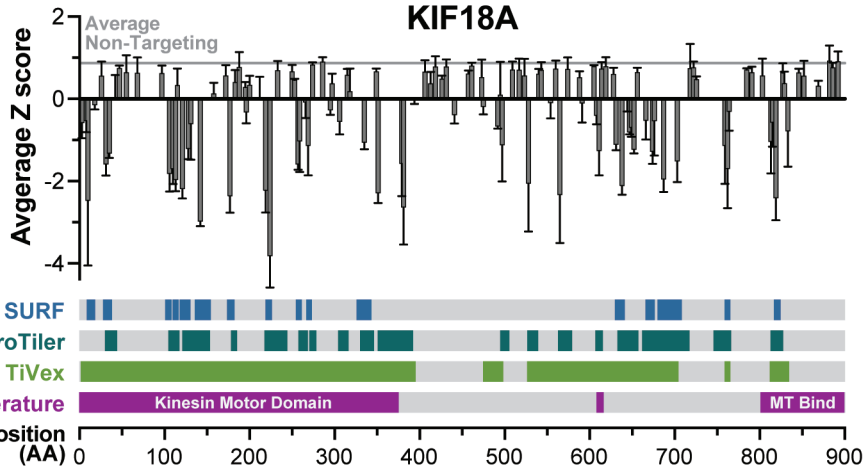


Figure 3

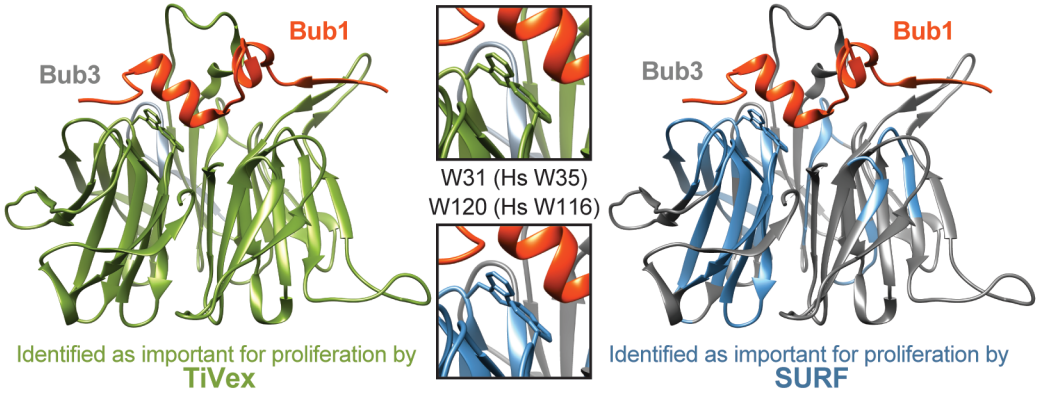
A



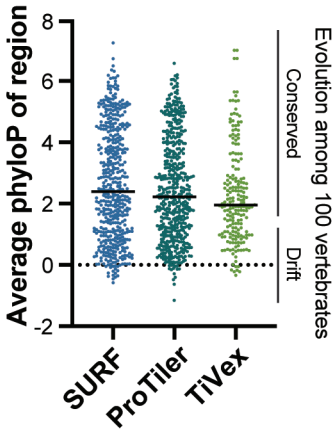
B

	SURF	ProTiler	TiVex
Genes	36/36	35/36	36/36
Regions	474	502	147
Average regions/gene	13.3	13.9	4.1
Average % protein covered	21%	26%	52%
Region Length (min / max)	4 / 51 aa	4 / 96 aa	3 / 562 aa
Region Length (mean)	9.9 aa	16.4 aa	99.6 aa
Region Length (median)	8 aa	13 aa	49 aa
Novel	186(39%)	186(37%)	50(34%)

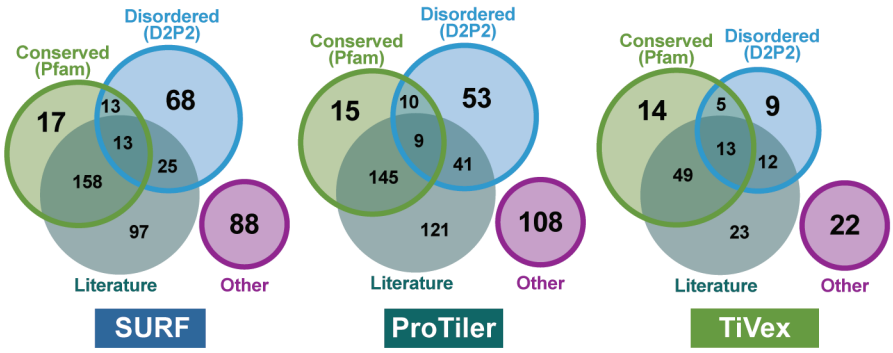
C



D



E



F

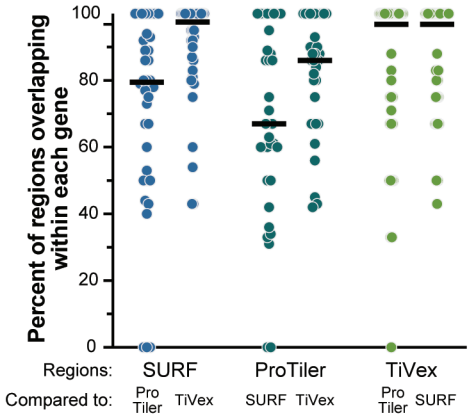


Figure 4

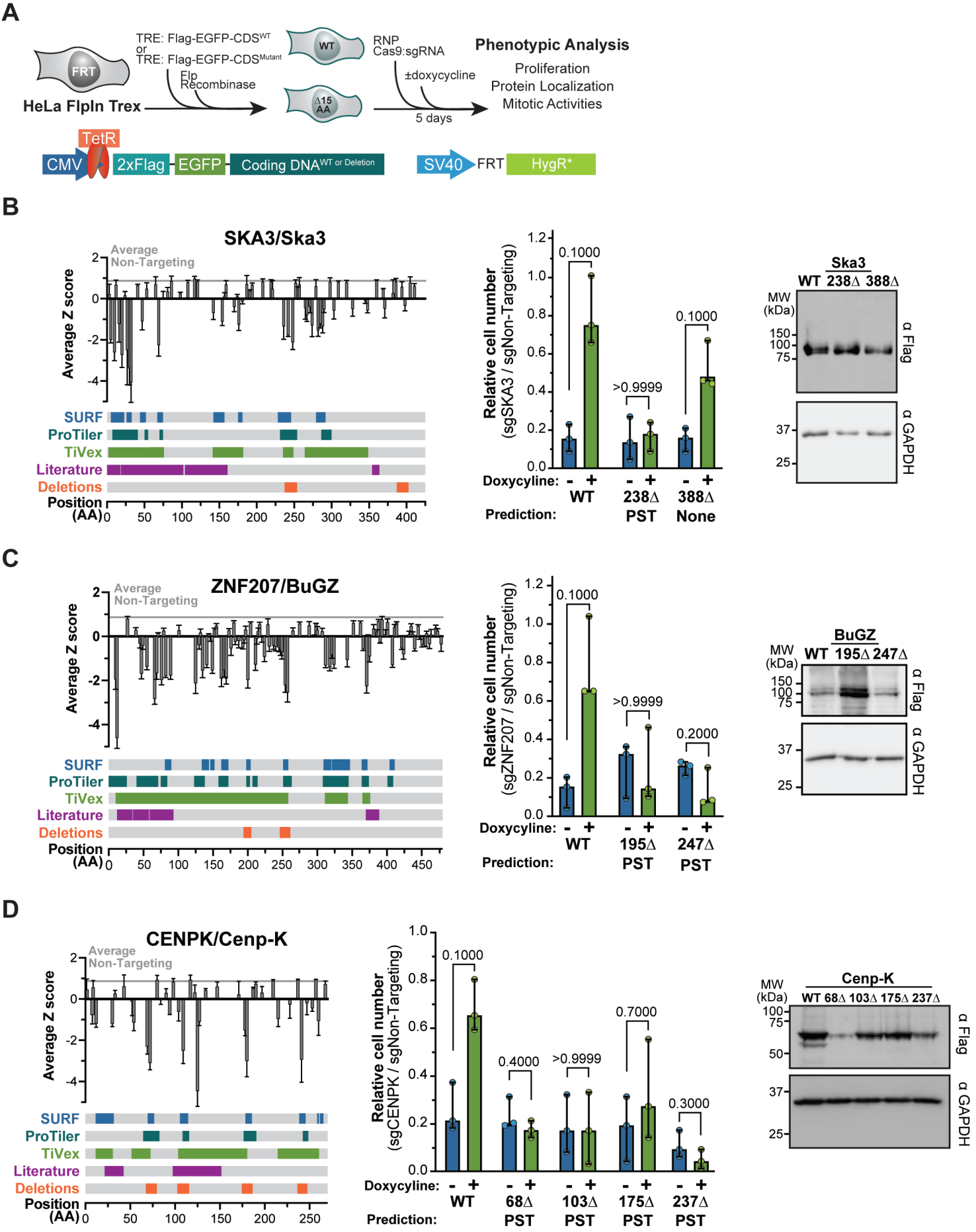

A
FRT
HeLa FlpIn Trex
TRE: Flag-EGFP-CDS^WT
or
TRE: Flag-EGFP-CDS^Mutant
Flp Recombinase
WT
Δ15 AA
RNP Cas9:sgRNA
±doxycycline
5 days
Phenotypic Analysis
Proliferation
Protein Localization
Mitotic Activities
TetR
CMV
2xFlag
EGFP
Coding DNA^WT or Deletion
SV40
FRT
HygR*
B
SKA3/Ska3
Average Non-Targeting
Average Z score
SURF
ProTiler
TiVex
Literature
Deletions
Position (AA)
Relative cell number (sgSKA3 / sgNon-Targeting)
0.1000
>0.9999
0.1000
Doxycycline: - + - + - +
WT
238Δ
388Δ
Prediction:
PST
None
Ska3
WT 238Δ 388Δ
MW (kDa)
150
100
75
α Flag
37
25
α GAPDH
C
ZNF207/BuGZ
Average Non-Targeting
Average Z score
SURF
ProTiler
TiVex
Literature
Deletions
Position (AA)
Relative cell number (sgZNF207 / sgNon-Targeting)
0.1000
>0.9999
0.2000
Doxycycline: - + - + - +
WT
195Δ
247Δ
Prediction:
PST
PST
BuGZ
MW (kDa)
WT 195Δ 247Δ
150
100
75
α Flag
37
25
α GAPDH
D
CENPK/Cenp-K
Average Non-Targeting
Average Z score
SURF
ProTiler
TiVex
Literature
Deletions
Position (AA)
Relative cell number (sgCENPK / sgNon-Targeting)
0.1000
0.4000
>0.9999
0.7000
0.3000
Doxycycline: - + - + - + - + - +
WT
68Δ
103Δ
175Δ
237Δ
Prediction:
PST
PST
PST
PST
Cenp-K
MW (kDa)
WT 68Δ 103Δ 175Δ 237Δ
100
75
50
α Flag
37
25
α GAPDH

Figure 5

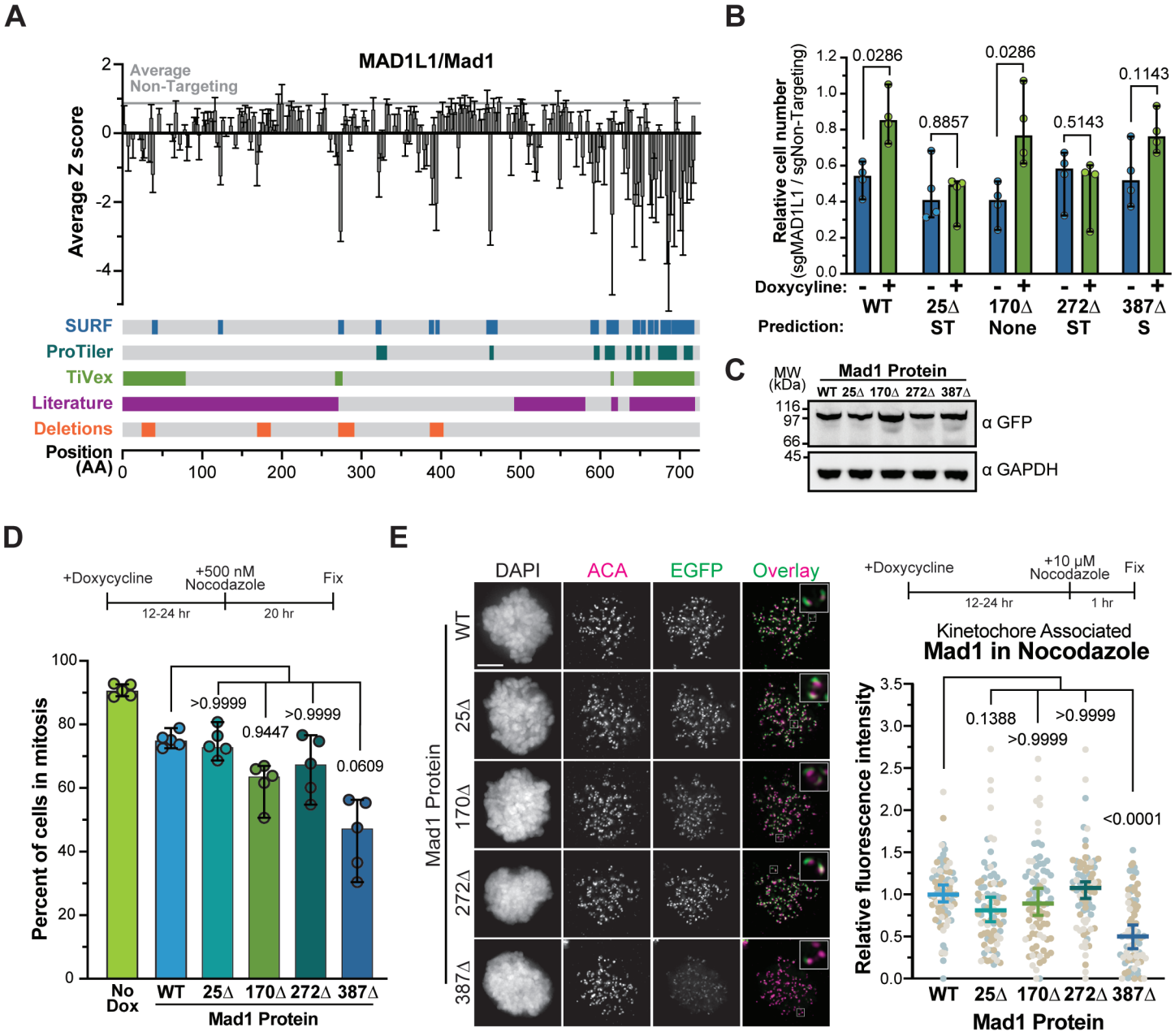
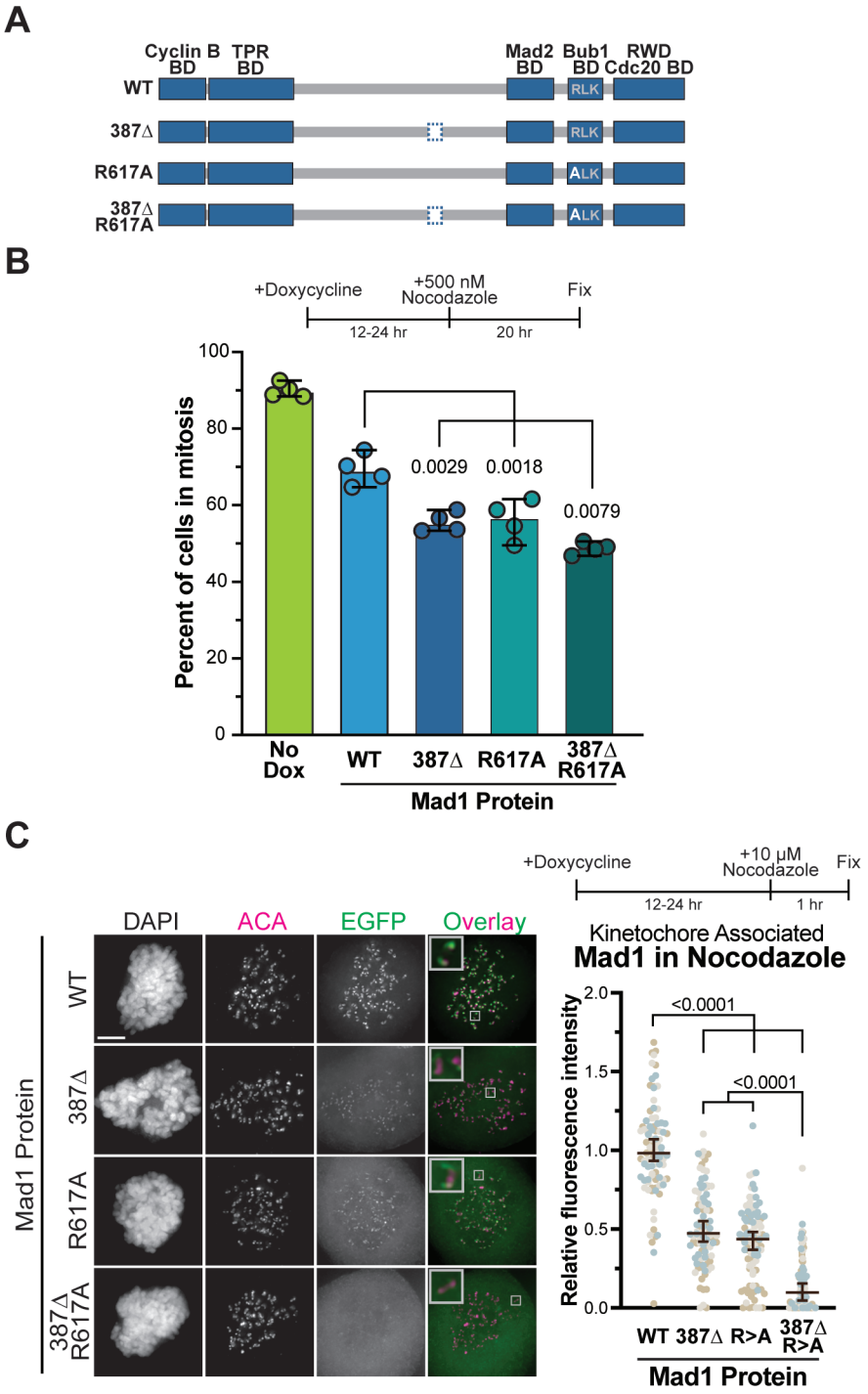

A
MAD1L1/Mad1
Average Non-Targeting
Average Z score
SURF
ProTiler
TiVex
Literature
Deletions
Position (AA)
B
Relative cell number (sgMAD1L1 / sgNon-Targeting)
0.0286
0.8857
0.0286
0.5143
0.1143
Doxycyline:
Prediction:
WT
25Δ ST
170Δ None
272Δ ST
387Δ S
C
MW (kDa)
Mad1 Protein
WT 25Δ 170Δ 272Δ 387Δ
α GFP
α GAPDH
D
+Doxycycline
+500 nM Nocodazole
Fix
12-24 hr
20 hr
Percent of cells in mitosis
>0.9999
0.9447
>0.9999
0.0609
No Dox
WT 25Δ 170Δ 272Δ 387Δ
Mad1 Protein
E
DAPI
ACA
EGFP
Overlay
Mad1 Protein
WT
25Δ
170Δ
272Δ
387Δ
+Doxycycline
+10 μM Nocodazole
Fix
12-24 hr
1 hr
Kinetochore Associated
Mad1 in Nocodazole
Relative fluorescence intensity
0.1388
>0.9999
>0.9999
<0.0001
WT 25Δ 170Δ 272Δ 387Δ
Mad1 Protein

Figure 6


A
Cyclin B BD
TPR BD
Mad2 BD
Bub1 BD
RWD Cdc20 BD
WT
387Δ
R617A
387Δ R617A
RLK
ALK
B
+Doxycycline
+500 nM Nocodazole
Fix
12-24 hr
20 hr
Percent of cells in mitosis
0.0029
0.0018
0.0079
No Dox
WT
387Δ
R617A
387Δ R617A
Mad1 Protein
C
+Doxycycline
+10 μM Nocodazole
Fix
12-24 hr
1 hr
DAPI
ACA
EGFP
Overlay
Mad1 Protein
Kinetochore Associated
Mad1 in Nocodazole
Relative fluorescence intensity
<0.0001
<0.0001
WT
387Δ
R>A
387Δ R>A
Mad1 Protein

Supplementary Figure 1

A

Broad Institute
GPP sgRNA Designer (Now CRISPick)

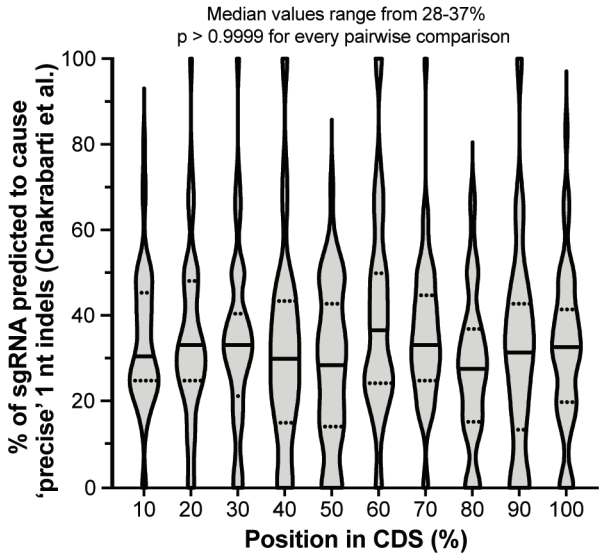
Input
Gene ID (48 mitotic regulators)

Output
All SpCas9 sgRNA targeting
Refseq transcript with longest CDS

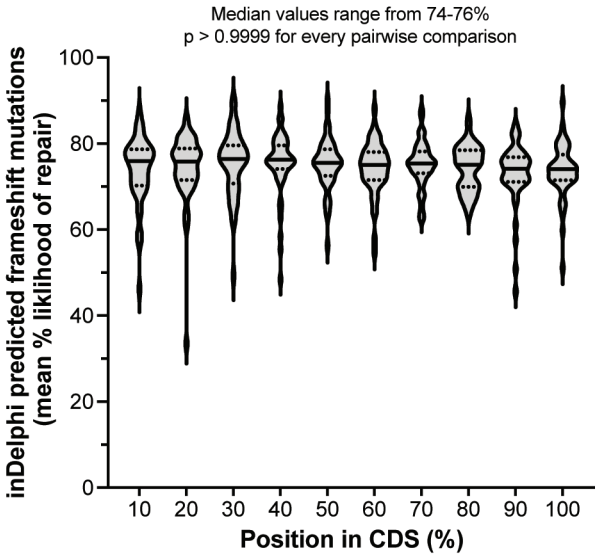
Exclude sgRNA if:
> 0 off target sites in coding regions
< 5 nt from preceeding sgRNA in CDS

Include sgRNA breaking rules if:
> 50 nt from preceeding sgRNA in CDS

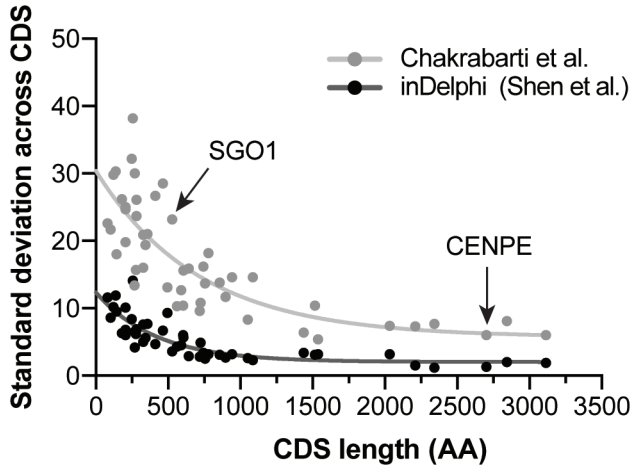
B



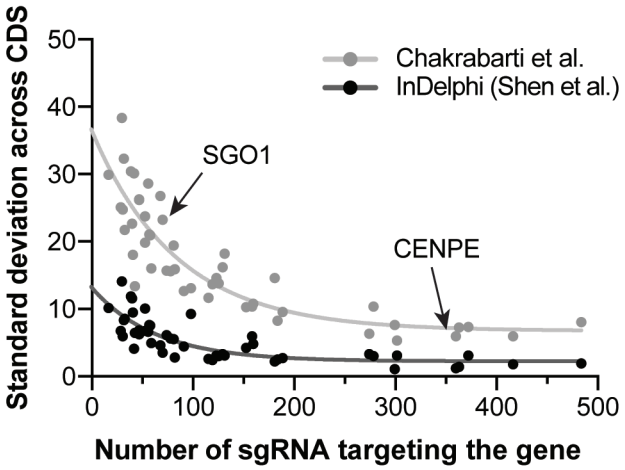
C



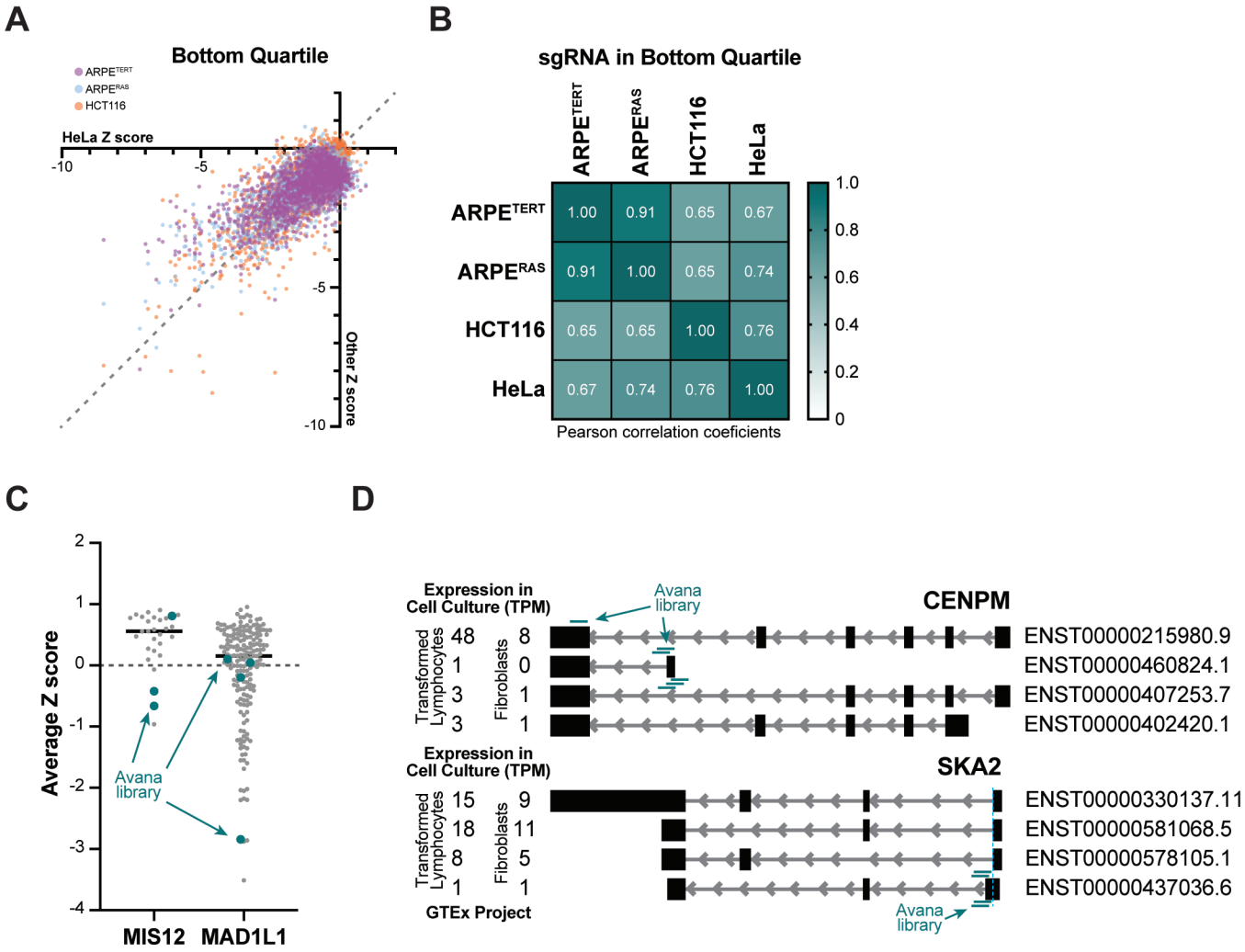
D



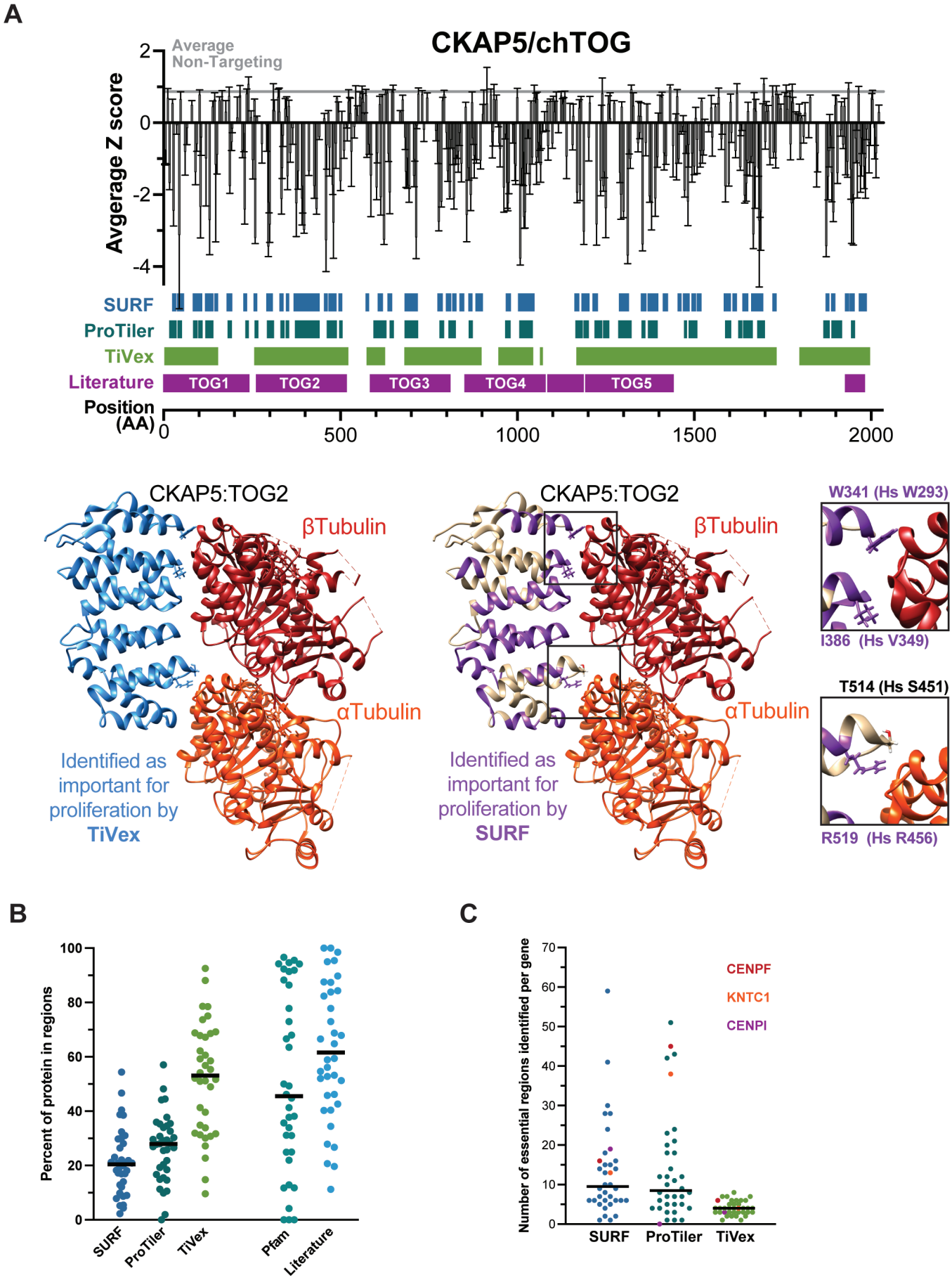
E



Supplementary Figure 2

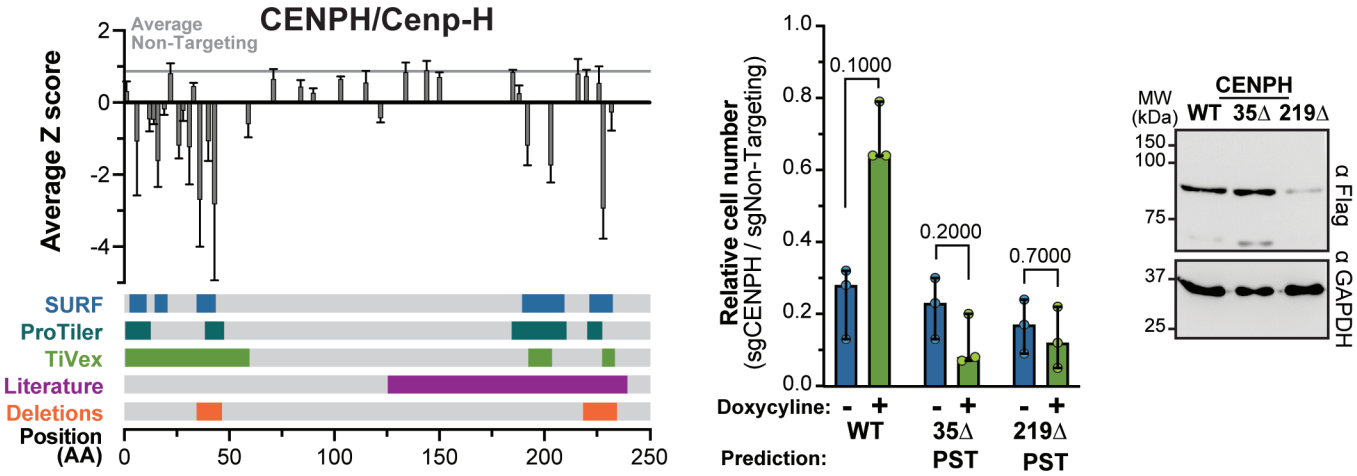


Supplementary Figure 3

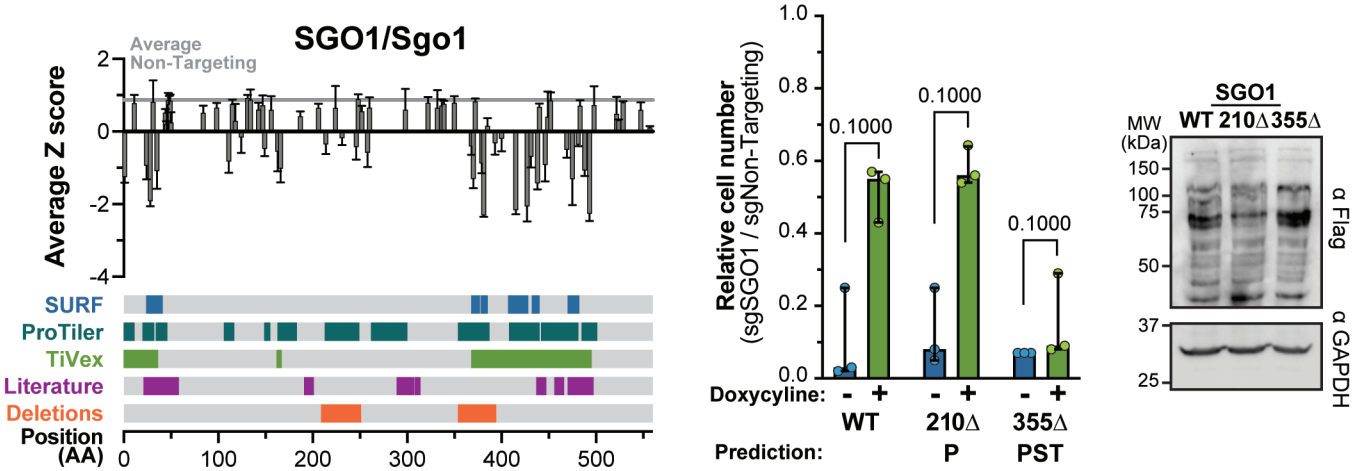


Supplementary Figure 4

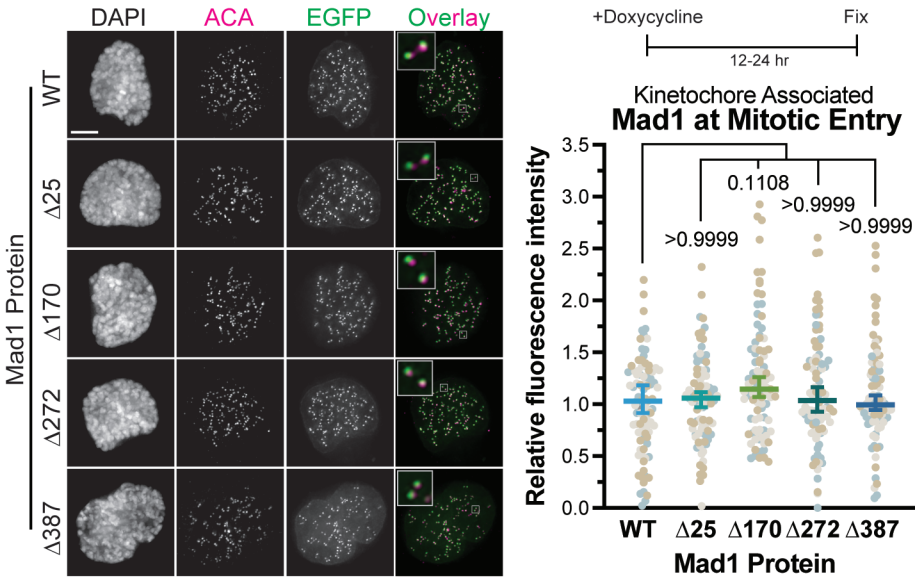
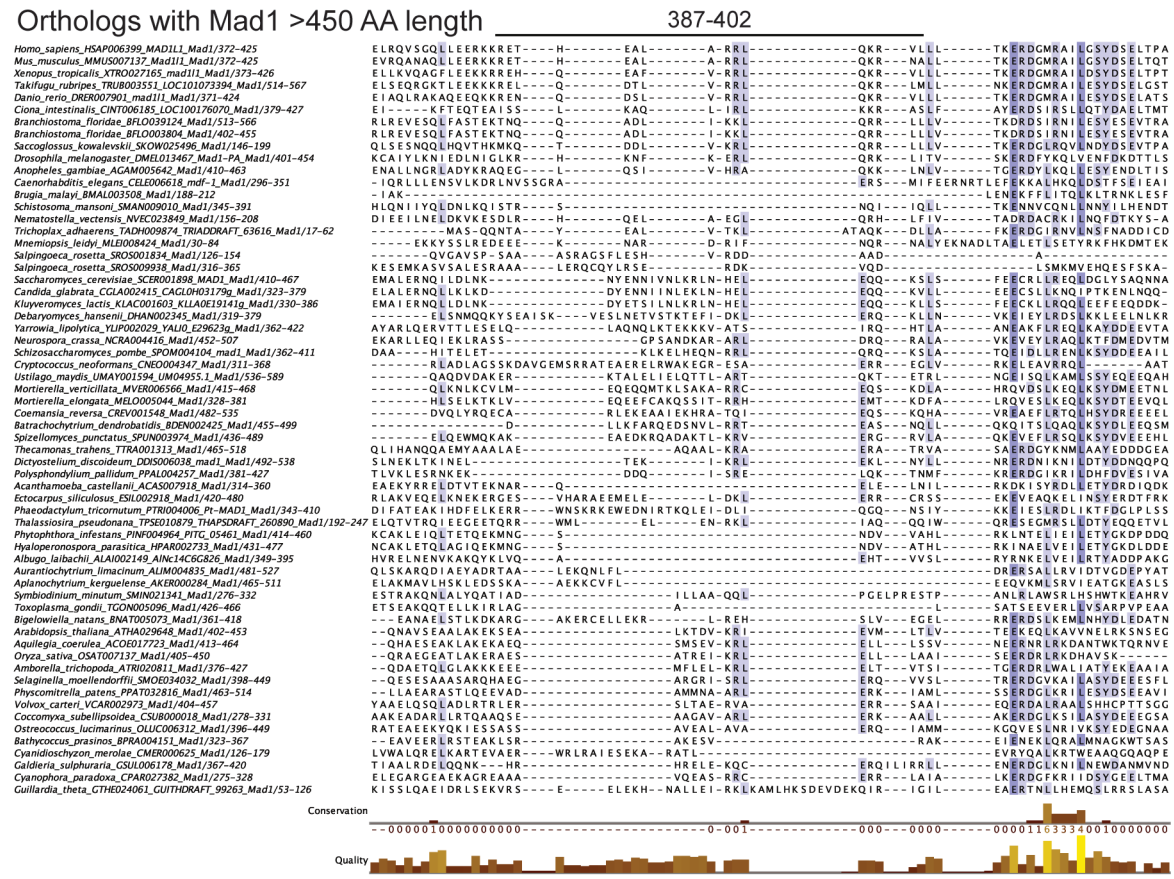
A



B



Supplementary Figure 5



Supplementary Figure 6