Widespread epistasis shapes RNA Polymerase II active site function and evolution

Authors: Bingbing Duan[1], Chenxi Qiu[2], Sing-Hoi Sze[3, 4], Craig Kaplan[1*]

Affiliations:

[1] Department of Biological Sciences, University of Pittsburgh, Pittsburgh, PA 15260.

[2] Department of Genetics, Harvard Medical School, Boston, MA 02215.

[3] Department of Computer Science and Engineering, Texas A&M University, College Station, TX 77843.

[4] Department of Biochemistry & Biophysics, Texas A&M University, College Station, TX 77843.

Contact information: craig.kaplan@pitt.edu

ABSTRACT

Multi-subunit RNA Polymerases (msRNAPs) are responsible for transcription in all kingdoms of life. At the heart of these msRNAPs is an ultra-conserved active site domain, the trigger loop (TL), coordinating transcription speed and fidelity by critical conformational changes impacting multiple steps in substrate selection, catalysis, and translocation. Previous studies have observed several different types of genetic interactions between eukaryotic RNA polymerase II (Pol II) TL residues, suggesting that the TL's function is shaped by functional interactions of residues within and around the TL. The extent of these interaction networks and how they control msRNAP function and evolution remain to be determined. Here we have dissected the Pol II TL interaction landscape by deep mutational scanning in *Saccharomyces cerevisiae* Pol II. Through analysis of over 15000 alleles, representing all single mutants, a rationally designed subset of double mutants, and evolutionarily observed TL haplotypes, we identify interaction networks controlling TL function. Substituting residues creates allele-specific networks and propagates epistatic effects across the Pol II active site. Furthermore, the interaction landscape further distinguishes alleles with similar growth phenotypes, suggesting increased resolution over the previously reported single mutant phenotypic landscape. Finally, co-evolutionary analyses reveal groups of co-evolving residues across Pol II converge onto the active site, where evolutionary constraints interface with pervasive epistasis. Our studies provide a powerful system to understand the plasticity of RNA polymerase mechanism and evolution, and provide the first example of pervasive epistatic landscape in a highly conserved and constrained domain within an essential enzyme.

INTRODUCTION

Transcription from cellular genomes is carried out by conserved multi-subunit RNA polymerases (msRNAPs) (*1-3*). Bacteria and Archaea use a single msRNAP to transcribe all genomic RNAs (*4-6*), while Eukaryotes have at least three msRNAPs (Pol I, II, and III) for different types of RNAs (*7-10*). RNA synthesis by msRNAPs occurs by iterative nucleotide addition cycles (NAC) of nucleotide selection, catalysis and polymerase translocation (*11-15*). msRNAPs active sites accomplish all three steps in the NAC using two conformationally flexible domains termed the bridge helix (BH) and the trigger loop (TL) (*5, 12, 16-19*). The multi-functional natures of the BH and TL likely underlie their striking conservation, serving as interesting models for studying the function and evolution of extremely constrained protein domains.

Nearly all catalytic cycle events are associated with the concerted conformational changes in the TL, and potentially the BH (*12-14, 19-23*).The BH is a straight helix in most msRNAP structures (*7, 8, 12, 13, 24*), but was found to be kinked in the *Thermus thermophilus* (Bacteria) RNAP structures(*5, 25*). The dynamics between the straight and kink conformations has been simulated and proposed to promote msRNAP translocation (*16, 25-27*). Even more importantly, the TL has been observed in various conformations that confer different functions. Among the observed conformations, a catalytic disfavoring “open” state facilitates translocation and a catalytic favoring “closed” conformation promotes catalysis (*12, 13, 28*). During each NAC, the TL nucleotide interaction region discriminates correct NTP over non-matched NTPs or dNTPs and initiates a TL conformational change from the open to the closed state (*11, 20, 29, 30*). The closure of the TL promotes phosphodiester bond formation (*20, 31*). Pyrophosphate release accompanies TL opening, which is proposed to support polymerase translocation to the next position downstream on the template DNA, allowing for the subsequent NAC (*32-34*)(**Fig. 1A**). Consistent with the model, mutations in the TL conferred diverse effects in every step of transcription (*13, 17, 23, 35-39*). For instance, mutations in the TL NIR impair interactions between TL and substrates, resulting in hypoactive catalysis and reduced elongation rate in vitro (Loss of function, LOF) (*17, 23, 36, 37, 40, 41*). Mutations in the TL hinge region and C-terminal portion appear to disrupt the inactive state of TL (open state) and shift the TL towards the active state (closed state), leading to hyperactive catalysis and increased elongation rate but impaired transcription fidelity (Gain of function, GOF) (*17, 28, 42*). TL conformational dynamics and functions are likely balanced by residue interactions within and around the TL (*41, 43-46*).

Intramolecular interactions in the active sites of msRNAPs control catalytic activity and underpin transcriptional fidelity. The TL is embedded in the conserved active site and interacts with other domains such as the BH, α-46 and α-47 helices, which form a five helix bundle with two TL helices enclosing a hydrophobic pocket **(Fig. 1A)** (*12, 28*). Many residue interactions observed between the TL and its proximal domains are critical for proper transcription, as catalytic activity and transcription fidelity can be altered by active site mutations within the TL (described above) and domains close to the TL in many msRNAPs (examples include BH GOF T834P and LOF T834A, funnel helix α-21 GOF S713P, and Rpb2 GOF Y769F) (*17, 23, 25, 26, 35-37*). These mutant phenotypes suggest that TL conformational dynamics and function are finely balanced and could be sensitive to allosteric effects from proximal domains (*47-50*). Understanding how “connected” the TL is to the rest of the polymerase will reveal the networks that integrate its dynamics with the rest of the enzyme and pathways for how msRNAP activity and evolution might be controlled.

Physical and functional intramolecular interactions between amino acids define the protein function and evolvability (*51-54*). Dependence of mutant phenotypes on the identities of other amino acids (epistasis) contributes to protein evolvability by providing a physical context and an evolutionary window in which some intolerable mutations may be tolerated (*55-57*). Recent studies have shown that mutations can alter the protein function, allostery, and evolvability, suggesting that even conserved residues are subject to distinct epistatic constraints dependent on context (*58-65*). In line with this prediction, distinct phenotypes for the same conserved residues have been observed in a number of proteins, including Pol I and Pol II in yeast(*63, 64, 66*). For example, the yeast Pol I TL domain is incompatible when introduced into Pol II even though about 70% of residues in the two yeast TLs are identical (*67*). The results strongly imply even functions of ultra-conserved domains are shaped by individually evolved enzymatic contexts (higher order epistasis).

Functional interactions between residues can be revealed by genetic interactions of double mutants of interacting residues (*17, 37, 68-70*). Previous studies from our lab on a small subset of site-directed substitutions have identified distinct types of Pol II double mutant interactions including suppression, enhancement, epistasis, and sign-epistasis(*17, 37, 69*). Suppression was common between LOF and GOF mutants as expected if each mutant is individually acting in the double mutant and therefore, opposing effects on activity are balanced. Similarly, synthetic sickness and lethality were commonly observed between mutants of the same (GOF or LOF) class, consistent with the combination of mutants with partial loss of TL function having greater defects when combined. However, we have also observed lack of enhancement between mutants of similar classes (epistasis), suggesting single mutants might be functioning at the same step, and in one case, sign-epistasis, where a mutant phenotype appears dependent on the identity of a residue at another position. For example, the GOF TL substitution Rpb1 F1084I was unexpectedly lethal with the LOF TL substitution Rpb1 H1085Y (instead of the predicted mutual suppression for independently (*37, 69*). This was interpreted as F1084I requiring H1085 for its GOF characteristics and becoming a LOF mutant in the presence of H1085Y (**Fig. 1B**). How representative these interactions are, and the nature of interactions across the Pol II active site requires a more systemic analysis to fully describe and understand the networks that control Pol II activity and the requirements for each mutant phenotype.

Deducing complex residue interaction networks on a large scale is challenging. To accomplish this for Pol II, we have previously established genetic phenotypes predictive of biochemical defects (*37*) and coupled this with a yeast Pol II TL deep mutational scanning system (Phenotypic landscape) (*17, 69, 71*). Here we develop experimental and analytical schemes to extend this system to a wide range of double and multiple mutants within the *S. cerevisiae* Pol II TL and between the TL and adjacent domains (Interaction landscape). By analyzing 11818 alleles including single mutants and a curated subset of double mutants, we have identified intricate intra- and inter-TL residue interactions that strongly impact TL function. Additionally, the examination of 3373 haplotypes including evolutionarily observed TL alleles and co-evolved residues revealed that TL function is heavily dependent on the msRNAP context (epistasis between TL and the rest of Pol II). These results suggest that despite being highly conserved, epistasis within msRNAPs contexts functions through derived residues and potentially reshapes functions of conserved residues. Finally, statistical coupling analyses reveals putative allosteric pathways appear to converge on the TL and may modulate active site activity upon factor

binding. Our analyses indicate TL function and evolution are dominated by widespread epistasis.

RESULTS

Systematic dissection of the Pol II active site interaction landscape. We developed an experimental and analytical framework, which we term the Pol II TL interaction landscape, to dissect residue interactions that shape Pol II TL function and evolution in *S. cerevisiae*. We designed and synthesized 15174 variants representing all possible Pol II TL single mutants, a subset of targeted double mutants, evolutionary haplotypes and potential intermediates in ten libraries (Supplemental Table 4). This approach follows our prior analysis of the TL phenotypic landscape (*17*) with modifications (see **Methods** and **Fig. S1A**). Libraries were transformed, screened under diverse conditions and phenotyped by deep sequencing (**Fig. 1C, Fig. S1A**). Growth phenotypes of mutants are calculated as the relative allele frequency shift from a control condition and normalized to the WT under the same conditions. Biological replicates indicate high reproducibility (**Fig. S1B-C**). Individual libraries were min-max normalized (*72*) to account for scaling differences between libraries (**Fig. S2A**) and same mutants present among different libraries indicate high correlation of fitness determinations in each library (**Fig. S2B-C**).

We defined a conceptual framework for evaluating genetic interactions among TL mutations. First, we assume that independence of mutant effects would result in log additive defects. This means that predicted double mutant fitness defects should be the combination of both single mutant defects, as is standardly assumed (*57, 68, 70, 73*). Deviation from log additive fitness defects represents potential genetic interactions between single mutants: either less than expected (i.e. suppression) or more than expected (i.e. synthetic sickness or lethality). Second, Pol II has two classes of active site mutants (GOF and LOF) that each confer fitness defects, and we previously observed activity additive interactions, meaning suppression between mutants of different classes (GOF+LOF) or synthetic sickness/lethality between mutants of the same class (GOF+GOF or LOF+LOF) in a set of mutants. We wished to distinguish specific epistatic interactions from activity-dependent suppression or synthetic interactions with mutant catalytic defects. For the purposes of our analysis, we defined an interaction as epistasis when we observed positive deviation in mutants of the same activity class (GOF+GOF, LOF+LOF), where we would expect synthetic sickness or lethality if mutants were functioning independently. We defined an interaction as sign epistasis for situations where we observed negative interaction for combinations between the classes (GOF+LOF), where we would expect suppression if mutants were functioning independently (**Fig. S3**).

Finally, the Pol II active site interaction landscape is based on accurate classification of mutant classes. We have previously demonstrated that mutant growth profiles across a select set of growth conditions are predictive of in vitro measured catalytic effects (Pol II TL phenotypic landscape). We extended this analysis by training multiple logistic regression models to predict phenotypic classes. We trained two models based on 65 mutants with measured in vitro catalytic defects and their conditional growth fitness to distinguish between GOF or LOF classes. Both models worked well in classifying GOF or LOF mutants (**Fig. S4A**). These two models were applied to all viable mutants (fitness score > -6.5 for control growth condition) and classified the mutants into three groups, GOF, LOF and those that did not belong to either one of the two groups ("unclassified"). To visually inspect the classification results, we applied t-SNE projection and k-means clustering for all measured mutants in all growth conditions to examine

clustering relationships to predictions from multiple logistic regression models. As shown in **Figure S4B**, we observed separated GOF and LOF clusters consistent with logistic regression classifications. With all phenotypic data, GOF and LOF mutants were further classified into different clusters, suggesting more fine-grained separation using additional phenotypes (**Fig. S4C**). In summary, we developed an experimental and analytical framework to dissect Pol II active site residue-residue interaction landscape in high throughput.

Widespread epistasis in the Pol II TL interaction landscape. To determine the TL-internal interaction networks, we rationally selected 2-4 different substitutions for each TL residue and combined them with the selected substitutions at all other TL positions. Substitutions were chosen to represent diverse phenotypes (GOF, LOF, lethal, or unclassified mutants). This curated set of 3790 double mutants represents potential interactions between any two TL residues (**Fig. 2A**). We compared the observed fitness of these double mutants with expectations from the additive model, and noticed the observed double mutant fitness deviated from the predicted fitness (r^2=0.21), which is much smaller than the r^2 (about 0.65-0.75) reported in other studies (*54, 70, 74-76*) (**Fig. 2B**), suggesting epistasis in the ultra-conserved TL domain might be more prevalent. About half of the combinations (1776/3790) matched the additive model (observed fitness ≈ expected fitness), while the rest showed positive (observed fitness > expected fitness, n=612) or negative (observed fitness < expected fitness, n=1402) interactions (**Fig. 2B**). From these positive or negative interactions, we distinguished the ratio of epistasis relative to activity-additive interactions. In all GOF/LOF combinations, we observed 43% activity-additive suppression and 41% negative interactions (sign epistasis). In all GOF/GOF or LOF/LOF combinations, activity-additive synthetic sick or lethal interactions were much more common than epistasis in combinations within the same class. We observed ~2% positive (epistasis) and 95% negative (activity-additive synthetic sick or lethal) interactions in GOF/GOF combinations, and 6% positive (epistasis) and 84% negative (synthetic sick or lethal interactions) interactions in LOF/LOF combinations (**Fig. 2C, Fig 2E and Fig S5**). Interactions were distributed throughout the TL and covered every TL residue, supporting connectivity across the TL. Observed epistasis was concentrated within the C-terminal TL helix and adjacent regions (**Fig 2D**), supporting functional dependency of TL-C terminal residues and consistent with their proposed function to collaboratively stabilize the TL open state.

Genetic interactions reveal further insight into the nature of previously lethal or unclassified individual mutants. First, most lethal mutants could be suppressed by at least one predicted GOF mutant (**Fig. 2F, Fig. S5**), suggesting that most lethal mutants likely have reduced activity (LOF) below a viable threshold, as might be predicted from greater probability of any individual mutant being a LOF than a GOF. However, two lethal mutations could be suppressed by most LOF mutations or specific other lethal mutants, but not GOF mutants, implying that their lethality resulted from being GOF (select A1076 substitutions). Second, unclassified single mutants mostly did not show widespread interactions with GOF, LOF, or lethal classes. However, a few unclassified mutants showed suppression in combination with GOF mutants, suggesting potential atypical LOF not detected by phenotypic analysis, or potential sign epistasis (**Fig. 2G, Fig. S5**).

Allele-specific interactions suggest unique properties of individual mutants with similar phenotypes. TL conformational dynamics and function are balanced by residue interactions within the TL (TL-internal interactions) and between the TL and TL-proximal domains (TL-external interactions). The properties of GOF and LOF mutants adjacent to the TL appear

similar to those inside but how they behave upon TL perturbation is not known. We analyzed the scope and nature of TL-internal and TL-external interactions by exploring interaction space of 12 previously studied GOF and LOF mutants (8 within the TL and 4 outside) each combined with all possible single TL mutants (**Fig. 3A**). These 12 mutants function as probes for the genetic interaction space of the TL and how it might be altered in allele-specific fashion by perturbation of the "probe" mutation. TL adjacent mutants showed similar scale of widespread interactions with TL substitutions as when TL-internal mutants were used as probes (**Fig. 3B, Fig. S6**). For these TL adjacent substitutions, we conclude their impact on Pol II function is of similar magnitude and connection as substitutions within the TL.

We further compared the similarity of interaction networks for substitutions with apparently similar biochemical and phenotypic defects. These analyses were designed to detect if changes to TL function might reflect simple alteration to TL dynamics, or additional alteration to folding trajectories or conformations. In the former case, mostly additive interactions might be predicted due to TL operating in the same fashion in double mutants versus single mutants, with phenotypes deriving from differences in kinetics or distributions of existing states. In the latter case where a mutation alters TL folding trajectories or changes TL conformations, it might be predicted that individual mutants that are superficially similar will show allele-specific genetic interactions reflecting epistatic changes to TL function. A subset of probe mutants showed widespread expected activity-additive suppression between GOF/LOF mutations and activity-additive synthetic lethality between same classes of substitutions (LOF/LOF or GOF/GOF). However, allele-specific epistasis and sign epistasis were also observed and were much higher for some mutants than others (**Fig. 3B-C, Fig. S6-S7**). 127/620 TL substitutions showed unique interactions with specific probe mutants; for example, some lethal substitutions could only be suppressed by Y769F, a GOF TL-proximal probe mutant in Rpb2 (**Fig. S8**). Moreover, two TL-adjacent GOF probe mutants, Rpb1 S713P (funnel α-helix 21) and the BH allele Rpb1 T834P displayed greatly distinct interaction networks despite similarly increased activities. Rpb1 S713P exhibited widespread suppression of LOF TL substitutions (96 instances) consistent with generic enhancement of activity but preservation of TL function. In contrast, Rpb1 T834P exhibited much lower suppression ability (33 instances). In addition to much lower ability to suppress, T834P showed a much greater amount of sign epistasis than Rpb1 S713P (102 instances to 38 instances) **(Fig 3C, Fig. S9A)**. These results are consistent with a model that perturbation to the BH structure is coupled to extensive changes to TL functional space and that T834P function as a GOF mutant requires most TL residues to be WT.

A similar distinction as above but between two internal TL GOF substitutions, Rpb1 E1103G and Rpb1 F1084I, was also apparent **(Fig. S9B)**. Rpb1 E1103G showed widespread suppression of LOF TL substitutions (184 instances), consistent with site-directed mutagenesis studies (*37*) **(Fig 3C, Fig. S6 and Fig S9B)**. These results suggest E1103G primarily may alter TL dynamics consistent with biochemical data that it promotes TL closure (*23*) and that it allows TL mutants primarily to maintain their effects. In contrast, Rpb1 F1084I showed more limited suppression of LOF alleles (43 instances) while showing much more widespread synthetic lethality **(Fig 3C, Fig. S6 and Fig S9B)**. These results indicate F1084I has a much greater requirement for WT residues at many TL positions to maintain its GOF characteristics. When TL function is additionally perturbed, F1084I appears to switch from a GOF to a LOF. These results imply that individual probe mutants distinctly reshape the Pol II active site, though they might share catalytic and phenotypic defects as single mutants.

An even more striking example of this phenomenon can be observed by comparison of the interaction networks of two LOF substitutions at the exact same position, the ultra-conserved H1085 residue **(Fig. S9C)**. This histidine contacts incoming NTP substrates (*5, 12*), is the target for the Pol II inhibitor α-amanitin (*36*), and promotes catalysis (*12, 13*). Initial structural data and molecular dynamics simulations were interpreted as H1085 potentially functioning as a general acid for Pol II catalysis (*77-80*). Our discovery that H1085L was especially well-tolerated (*17*), and subsequent experiments from the Landick lab (*81, 82*), have led to their proposal that the TL histidine functions as a positional catalyst and a similarly sized leucine supports catalysis with relatively mild effects on biochemistry and growth. If H1085Y and L substitutions are acting on a continuum of positional catalyst activity, we might predict their interaction networks would be similar and only be distinguished by magnitude of interactions, but not identity or type of interactions. In contrast to this prediction, distinct interaction patterns were observed (**Fig. 3C**, **Fig. S6 and S9C**). Most GOF mutants were able to suppress H1085Y but not H1085L. Instead, H1085L showed synthetic lethality with most GOF mutants (putative sign epistasis). For example, almost all substitutions at E1103 showed sign epistasis with H1085L but not H1085Y (**Fig. S7B, S9C**). Distinction between H1085L and H1085Y is evident in the PCA plot of probe mutants (**Fig. 3D**). The partially unique nature of each probe mutant is also evident in the PCA plot (**Fig. 3D**). Altogether, distinguishable interaction networks of probe mutants, despite their similarity in catalytic and growth defects, even within the same residue, suggest that each mutant has ability to propagate effects across the Pol II active site. To some extent, each Pol II mutant creates a new enzyme.

Pol II TL interaction landscape reveal functional dependency of proximal residues. Several allele-specific epistatic interactions were also observed. Some of the strongest epistatic interactions were between A1076 substitutions and L1101S, which differed from all other GOF probe mutants (**Fig. 4A**), suggesting tight coupling between A1076 and L1101 for Pol II function. These two hydrophobic residues, together with other hydrophobic residues in TL proximal helices, form a five-helix bundle in the Pol II active site likely stabilizing the open TL conformation (**Fig. 4C**). Consistent with this, another pair adjacent residues, M1079 and G1097, also showed allele-specific epistasis (**Fig. 4B**).

The epistasis we identified in combinations within the same class (GOF/GOF or LOF/LOF) might also be sign epistasis (GOF suppressing GOF or LOF suppressing LOF due to a switch in residue class). We distinguished regular epistasis (lack of additivity) from sign epistasis suppression by checking conditional phenotypes predictive of biochemical defects. We reasoned that epistatic interactions would exhibit double mutant conditional phenotypes similar to single mutants while sign epistasis suppression would also exhibit suppression of conditional phenotypes. Therefore, we examined double mutants with our logistic regression models for determining phenotypic class. The majority of double mutants within each class showing positive epistasis (GOF/GOF or LOF/LOF) maintained single mutant classification. 6/10 GOF/GOF doubles showing positive epistasis were classified as GOF while 30/38 LOF/LOF doubles were classified as LOF, suggesting classic epistasis (**Fig. S10A**). In three cases of GOF/GOF combinations, all between L1101S and A1076 substitutions, the resulting double mutants were unclassified, consistent with nearly WT behavior. Here, each constituent single mutant conferred a GOF phenotype, but the double mutants show mutual suppression. This suggests tight coupling between 1101 and 1076 (see Discussion).

We also observed allele-specific interactions for predicted lethal mutants. Our threshold for lethality is likely higher than that in actuality, and very slow growing mutants may fall below our lethal threshold while still having enough data on conditional fitness assessment for logistic regression to predict mutant class. For 21 ultra sick/lethal TL substitutions predicted as GOF themselves, we observed suppression when combined with other GOF mutants (**Fig. S10B-C**). Lethal substitutions of A1076 could be suppressed by LOF probe mutants and the GOF probe L1101S, consistent with specific combinations between 1076 and 1101 showing sign-epistasis suppression or allele-specific mutual suppression. F1084R is a predicted lethal GOF but can be suppressed specifically by GOF probe Y769F. F1084 and Y769 are close to each other when the TL is in the closed, substrate bound state. Additionally, 5 ultra-sick/lethal substitutions predicted as LOF could be suppressed by a LOF allele (**Fig. S10B**). As an example, S1091G could be suppressed by almost all curated GOF mutants, yet it was also specifically suppressed by the LOF V1094D (**Fig. S10C**). S1091G and V1094D appear to compensate for each other in a allele-specific fashion. We suggest that these are the types of interactions that will allow the TL and adjacent residues to evolve and differentiate while maintaining essential functions.

We note that strong epistasis is much more prevalent in the Pol II system than in other proteins where it has been quantified (*54, 74-76, 83*) (**Fig. S10D**). We attribute this difference to the much higher rate of suppressive interactions due to Pol II mutants having opposing effects on catalysis.

TL evolution is shaped by contextual epistasis. We previously found that identical mutations in a residue conserved between the Pol I and Pol II TLs yielded different biochemical phenotypes (*67, 84*). Furthermore, the yeast Pol I TL was incompatible within the yeast Pol II enzyme, implying that TL function is shaped by the enzymatic context (*67, 84*). To determine the generality and scope of TL-Pol II incompatibility, we designed a library containing evolutionary TL variants from bacterial, archaeal, and eukaryotic msRNAPs and determined their compatibility in the yeast Pol II context (**Fig. 5A**). TL alleles of eukaryotic Pols were more compatible than those from Archaea and Bacteria, and Pol II alleles were the most compatible (**Fig. 5B**, **Fig. S11A**), consistent with evolutionary distance. The total number of TL substitutions in haplotypes were slightly negatively correlated with growth fitness in the Pol II background for Archaeal, Pol I, II and III TLs (**Fig. S11C**), though not for Bacterial TLs, likely because the bacterial TLs were almost entirely incompatible in the Pol II context (**Fig. S11C**). Conservation of TL sequence and function was high enough that some archaeal sequences could provide viability to yeast Pol II, yet at the same time a number of Pol II TLs from other species were defective if not lethal. These results suggest widespread coevolution of TL sequence outside of ultra-conserved positions shapes TL function (see Discussion).

We reasoned that evolutionarily observed lethal substitutions might be closer to functional than non-evolutionarily observed and would therefore be more likely to be suppressible by Pol II GOF alleles. To compare suppressibility between evolutionarily observed and unobserved substitutions lethal to Pol II, we extracted the highest positive deviation scores among all double mutants containing each lethal substitution. Maximum deviation scores for Pol II lethal substitutions present in TLs of existing msRNAPs were higher than for lethal substitutions that were absent, indicating the Pol II lethal mutants present in existing msRNAPs on average maintain a greater functionality and/or are suppressible by single changes (**Fig. 5C, Fig. S11B**). The TL has been estimated as providing 500-1000 fold enhancement on catalytic activity (*85-87*), while we estimate only ~10-fold effects are tolerated for yeast viability (*36*). We conclude

that lethal mutants observed as functional residues in other species are more likely to be close to the viability threshold as might result from a series of small steps to allow them to function.

TL residues co-evolve with the rest of Rpb1 through diverse pathways. Our analyses suggest that even a highly conserved domain such as the Pol II TL can be sensitive to identity of adjacent residues and that changing networks of interactions shape the Pol II active site across evolution. We employed statistical coupling analysis (SCA) to identify if there are any coevolving residue networks in Rpb1 and ask about pathways that might co-evolve the TL. SCA "Sector" analysis is especially useful for identify subgroups of coupled residues that might form allosteric communication networks (*88-90*). We extracted 410 yeast Pol II Rpb1 sequences from the recently published msRNAP large subunit multiple sequence alignment (MSA) from the Landick lab (*82*) and performed SCA (see **Methods**)(*89*). We identified 40 coevolving sectors (**Fig. S12**), and every single TL residue was found within one of the eight sectors that form generally continuous network of interactions within Rpb1(**Fig. 6**). TL residues within the TL NIR were coupled with most BH residues and the alanine-glycine linker (Rpb1 1087-1088). These residues are highly conserved (*17*), indicating this sector is driven by conservation primarily. Six of eight Rpb1 sectors containing TL residues also contained at least one BH residue, supporting functional coupling between these two domains. Coupling is not limited to residues that are close to the active site. Distal residues can potentially modulate TL function through allosteric interactions. For example, the greatest distance between a TL residue and another Rpb1 residue in the same sector is ~ 55 Å. Interestingly, the residue pair 1076-1101, for which we observed extensive epistasis, are the sole TL residues within a very large cluster containing >150 residues across Rpb1. Our epistasis studies indicate multiple allele-specific interactions between 1076 and 1101 of exactly the type that might appear as evolutionary coupling between specific substitutions at these positions. The hydrophobic TL pocket is an attractive linchpin for potential communication to the TL from throughout Pol II, and multiple sectors converge on this domain.

DISCUSSION

How individual mutants alter a protein's function is not necessarily straightforward at the mechanistic level. Amino acid substitutions both remove functionality of the WT residue but replace that functionality with something different. By altering the local environment within a protein or potentially propagating effects to distant locations through allosteric changes, each substitution potentially can be quite different. These differences may not be apparent as phenotypic outputs and phenotypic assays may not have granularity to distinguish different biophysical behaviors if they result in similar outputs. For Pol II mutants, even high-resolution phenotypic analyses, such as gene expression profiling or genetic interaction profiling between Pol II mutants and deletions in other yeast genes (*49*), suggest that LOF and GOF mutants represent a continuum of defects that match enzymatic activity in vitro. Therefore, these profiles also appear dependent on the output of Pol II activity defects and can't distinguish potential differences in underlying mechanism.

Through systematic detection of genetic interactions within the Pol II active site, we have identified functional relationships between amino acids across the TL and between TL substitutions and others. In the absence of double mutant epistasis analyses it would not be possible to differentiate similar alleles from one another. L1101S and E1103G, for example, are

two GOF alleles very close to each other in Pol II structure and confer similar phenotypic landscapes across various growth conditions. Here, we find that their distinct interactions support that substitutions at 1101 and 1103 target distinct residue networks (**Fig. 3C, Fig. 4A, and Fig. S6-S7**). 1101 functions in the five-helix bundle hydrophobic pocket while 1103 interacts and co-evolves with a number of TL external residues that together support interactions that promote the open TL conformation (**Fig. 4C**, **Fig. 6**). We also observed connections between TL C-terminal residues that suggest a limit to how disruptions to structure there can alter Pol II activity (**Fig. 2D, Fig. 6**). Helix-disrupting LOF proline substitutions in at least two TL positions showed epistasis with multiple substitutions in the back of the TL (1094-1098), suggesting that their functions require TL C-terminal helix structure and in the absence of that structure (proline disruption) effects are no longer additive.

The strongest epistatic interactions observed were between two pairs of hydrophobic residues, A1076 and L1101, and M1079 and G1097 (**Fig. 4**). Each of these contributes to the structure of a hydrophobic pocket that bundles two TL proximal helices with the BH and two others in a five-helix bundle. Supporting the dependence of these residues on each other for maintaining function, identity at these positions over evolution also shows coupling. Interestingly, these A1076 and L1101 were coupled uniquely out of TL residues with a great number of other positions in Rpb1 (**Fig. 6**).

Elongation factors bind Pol II and alter its activity, but the mechanisms by which they do so are not known (*91, 92*). We observed a high level of genetic interactions between residues outside the TL and residues within it, including allele-specific reshaping of TL mutant space upon single substitution outside the TL (**Fig. 3**). The fact that minor mutational changes outside the TL can apparently functionally perturb the TL would be consistent with the idea that minor alterations to Pol II structure upon elongation factor binding could easily propagate into the active site via the TL or the BH. As an example, human Rtf1 has been observed to project a domain into the Pol II structure adjacent to the BH (in yeast, this region is occupied instead by Rpb2 (*93*)). These contacts have been proposed to alter Pol II activity. We would propose that the paths for such alteration activity would follow the coupling sectors we have observed by SCA.

How different individual substitutions are under the surface is critical for understanding plasticity in protein mechanisms and how they might be altered by evolutionary change. A key open question in nucleic acid polymerase mechanisms is the paths for protons in the reaction (for example, deprotonation of the synthesized strand 3′-OH and protonation of pyrophosphate leaving group, for example) (e.g.(*77, 78, 80, 82, 94-96*)). For msRNAPs, the association with incoming NTP by a nearly universally conserved histidine led to the proposal that this residue might donate a proton during the reaction (*12, 79, 94*). Some substitutions at this position can provide minimal essential function (e.g. tyrosine, arginine), while others are only moderately defective (glutamine). Surprisingly, we found that H1085L was very-well tolerated for growth (*17*) and the Landick lab has proposed this substitution supports catalysis through positional but not chemical effects (*81, 82*). Our studies here were quite surprising in that they indicated that L1085 Pol II has unique behavior when perturbed by all possible TL substitutions and is entirely distinct from H1085Y (where we have direct observations of all possible intra-TL doubles) or H1085A or H1085Q (curated doubles) (**Fig. S9C, Fig. S13**). These residue specific behaviors suggest that each substitution may have different properties, and compatibility with function may not necessarily represent similar function under the surface.

Evolutionary change over time can alter protein function but it can also alter protein functional plasticity. Recent work from the Thornton lab elegantly demonstrates that phenotypes of substitutions to residues conserved over hundreds of millions of years can change over evolutionary time and can do so unpredictably and transiently during evolution (*59*). msRNAPs have structures and functions conserved over billions of years, and deep within their active sites is a mobile domain, the TL, that has large functional constraints on its sequence. The TL sequence must be able to fold into multiple states and maintain recognition of the same substrates across evolutionary space and is shows high identity even between distantly related species. Here we show that the TL, and likely the entire Pol II active site, exhibits a great amount of plasticity through non-conserved positions that are essential for compatibility of the TL and surrounding domains. Our results illustrating widespread epistasis and allele-specific effects of single and double mutants predict that comparative analyses among Pol I, II, and III will reveal widespread and enzyme-specific mechanisms due to higher order epistasis shaping function of conserved residues.

METHODS

Design and Synthesis of TL mutant libraries. We updated and extended the fitness dataset of Qiu *et al* (*17*). Using a similar methodology, but with adjusted conditions and a second-generation mutant library strategy, in order to generate a complete Pol II TL mutation-phenotype map and examine genetic interactions. Mutants were constructed by synthesis with Agilent and screened for phenotypes previously established as informative for Pol II mutant biochemical defects. Programmed oligonucleotide library pools included all 620 single TL residue substitutions and deletions for Rpb1 amino acids 1076-1106 (Library 1), 3914 pairwise double substitutions (Library 2), 4800 targeted double substitutions (Library 6), and 3373 multiple substitutions (Library 3-5), along with the WT *S. cerevisiae* Pol II TL allele at a level of ~15% of the total variants, enabling precise quantification (see Supplemental Table 4). Each synthesized region contained a mutated or WT Pol II TL sequence and two flanking regions at the 5′ and 3″ ends of the TL-encoding sequence. These flanking regions also contained designed “PCR handle” (20bp) sequences, allowing distinct subsets of oligos to be amplified from synthesized pools using selected primers for PCR, and additional flanking WT Pol II sequences allow for further extension of homology arms by PCR “sewing” (Details are in Supplemental Method 2 and 3).

Introduction of Libraries into yeast and phenotyping. Synthesized mutant pools were transformed into yeast (CKY283) along with an *RPB1*-encoding plasmid where the TL-encoding sequence was replaced with an MluI restriction site for linearization as described in Qiu et al (*17*). This strategy allows construction of *rpb1* mutant libraries by gap repair between library fragments and the linearized vector. Briefly, the synthesized oligo pools were amplified by limited cycles of emulsion PCR to limit template switching. Extension of flanking homology arms of ~200 bp were added by PCR sewing. Amplified TL sequences with extended flanking regions were co-transformed with linearized pRS315-derived *CEN LEU2* plasmid (pCK892) into CKY283, allowing gap repair via homologous flanking regions. To detect potential residue-residue interactions between the TL and TL-proximal domains including the Rpb1 Bridge Helix (BH), Funnel Helix alpha-21 and Rpb2, the Pol II TL single mutant pool (Library 1, 620 mutant alleles and 111 WT alleles) was co-transformed individually with gapped plasmids encoding an additional *rpb1* allele (Rpb1 BH T834P, T834A, or Funnel Helix alpha-21 S713P) into CKY283 respectively, or with the gapped WT *RPB1* plasmid into a strain with the genomic mutation, *rpb2*

Y769F. These co-transformations created double mutants between the TL and TL-proximal mutants. The WT allele in single mutant pool represented the single probe mutant due to substitutions outside the TL on the plasmid or in the strain background. To distinguish between a fully WT TL and a WT TL representing the TL of a mutant allele elsewhere, a WT Pol II TL allele with a silent mutant at T1083 (WT codon ACC was replaced with ACT) was co-transformed with plasmid containing gapped WT *RPB1* in a WT strain in parallel. 15% of the transformants with silent mutation were mixed with transformants of double mutants. The silent mutation allowed us to distinguish the WT and the single mutants. Each transformation was done in three biological replicates. After transformation, Leu^+ colonies were collected from SC-Leu plates by scraping into sterile water and replated on SC-Leu+5FOA to select for cells having lost the *RPB1 URA3* plasmid. 5-FOA-resistant colonies were scraped into sterile water from SC-Leu+5FOA and replated on SC-Leu, SC-Leu + 20mg/ml MPA (Fisher Scientific), SC-Leu + 15 mM Mn (Sigma), YPRaf, YPRafGal, SC-Lys, and SC-Leu + 3% Formamide (JT Baker) for phenotyping. Details of cell numbers plated on each plate and screening time of each plate are in Supplemental Table 3. Details of high efficiency transformation protocol is in supplemental method 1.

Generation of libraries for quantification by amplicon sequencing. Genomic DNA of each screened library was extracted using the Yeastar genomic DNA kit according to manufacturer's instructions (Zymo Research). To ensure adequate DNA for sequencing, the TL regions of all libraries were amplified with PCR cycles that were verified to be in the linear range by qPCR to minimize disturbance of allele distributions, and under emulsion PCR conditions (EURx Micellula DNA Emulsion & Purification (ePCR) PCR kit) to limit template switching. Details are in Supplemental Method 2 and 3. To multiplex samples, we employed a dual indexing strategy wherein 10 initial barcodes for differentiating 10 mutant libraries were added during the initial amplification using 10 pairs of custom primers. In a second amplification, 28 primers containing 28 NEB indices were used to add a second index for distinguishing conditions and replicates (NEBNext Multiplex Oligos for Illumina) (see Supplemental Table 2). As a result, a sample-specific barcodes were present for each set of variants. The indexed, pooled samples were sequenced by single end sequencing on an Illumina Next-Seq (150nt reads). On average, over 11 million reads were obtained for individual samples with high reproducibility from two rounds of sequencing. Raw sequencing data has been deposited on the NCBI SRA (Sequence Read Archive) database under BioProject PRJNA948661. Processed mutants counts and fitnesses are available through GitHub (https://github.com/Kaplan-Lab-Pitt/TLs_Screening.git).

Data cleaning and fitness calculation and normalization. Reads of mutants were sorted into appropriate libraries and conditions by detecting particular indices after sequencing. Read counts were estimated by a codon-based alignment algorithm to distinguish reads that exactly matched designated codons of mutants (*97*). To clean the data, mutant reads with coefficients of variation greater than 0.5 in the control condition (SC-Leu) were excluded from the analysis. The mutant read count was increased by 1 to calculate the allele frequency under different conditions. To measure and compare the phenotypes of all mutants, mutant phenotypic score (fitness) was calculated by allele frequency change of a mutant under selective conditions relative to the unselective condition comparing to the frequency change of WT. The formula for calculating fitness is shown below.

$$\text{Fitness (mut)} = \log [f^{\text{mut, sele}} / f^{\text{mut, unsele}} - \log [f^{\text{WT, sele}} / f^{\text{WT, unsele}}]$$

We applied min-max normalization to bring the median growth fitness of mutants measured at ten libraries to the same level for direct comparison (formula is shown below). In each library, we divided mutants into several groups based on their allele counts on the control condition. Mutants with read count differences of less than 10 are present in one group. The WT growth fitness was set as the maximum value and the minimum fitness in each group was the minimum. Min-max normalization was used to equalize the growth fitness into the same range between various groups inside each library. Additionally, we utilized min-max normalization to level the mutant fitness across all ten libraries with WT fitness as Max and minimal fitness in each library as the minimum. As a result, mutant growth fitness was scaled to one range and could be used to determine genetic interactions.

$$X' = \frac{X - Xmin}{Xmax - Xmin}$$

Determination of functional interactions. The genetic interactions between single substitutions were determined by comparing the multiple-substitution mutant normalized median fitness to the log additive of the single substitution normalized median fitness. The simplified formula is as follows:

Deviation score ($M_1M_2M_3$) = Fitness ($M_1M_2M_3$) – [Fitness (M_1) + Fitness (M_1) + Fitness (M_3)]

(1). -1 < Deviation score < 1, the interaction among the constituent single mutants is additive and mutants are acting independently.

(2). Deviation score ≥ 1, the interaction is non-additive and is positive, including suppression and epistatic interactions.

(3). Deviation score ≤ -1. the interaction is non-additive and is negative, including synthetic sick, synthetic lethal, and sign epistasis interactions.

Any mutation with fitness smaller than the lethal threshold (-6.50) was classified as an ultra-sick/ lethal mutant and its fitness was normalized to -6.50 for calculation of the deviation score. Synthetic sickness and synthetic lethality were distinguished by whether a double mutant is viable or lethal (fitness is greater than or equals to the lethal threshold -6.5) when two constituent mutations are viable. Synthetic lethality can be further classified into two types. First, additive-synthetic lethality was determined when the expected double mutant fitness calculated by additive model was lethal (expected fitness = -6.5) and the observed double mutant fitness was also lethal (fitness = -6.5) (in this case the deviation score = 0). Second, the beyond-additive synthetic lethality was determined when the expected double mutant was viable (expected fitness > -6.5) while the observed double mutant fitness was lethal (fitness = -6.5) (in this case the deviation score < 0). To separate these two situations in our figures, we labeled additive synthetic lethality as black and beyond-additive synthetic lethality as purple.

Details of formulas are in Supplemental Method 4. The codes for calculating deviation scores and generating figures are available in GitHub (https://github.com/Kaplan-Lab-Pitt/TLs_Screening.git).

Mutant classification using two multiple logistic regression models. We trained two multiple logistic regression models to distinguish GOF and LOF mutants using the phenotypic fitness on SC-Leu+MPA, SC-Lys, and YPRafGal conditions of 65 single mutants, including 25 previously identified GOF mutants, 33 LOF mutants, one WT, and six that were not GOF or LOF

mutants. Intercept, main effects, and two-way interactions were involved in defining both models. 0.75 was used as the cutoff threshold for both the GOF and LOF models.

Model for predicting the probability of a mutant being a GOF:

$$y = \frac{1}{1 + e\text{^}(1.816 + 2.542 * fMPA - 1.942 * fLys + 0.06566 * fGal - 0.5297 * fMPA * fLys - 0.08373 * fMPA * fGal + 0.02556 * fLys * fGal)}$$

Model for predicting the probability of a mutant being LOF:

$$y = \frac{1}{1 + e\text{^}(1.916 - 1.392 * fMPA - 1.328 * fLys - 0.8353 * fGal - 0.01112 * fMPA * fLys - 0.2992 * fMPA * fGal + 0.8823 * fLys * fGal)}$$

Both models showed accuracy, with the area under ROC close to one **(Fig. S3A)**. The details are provided in Supplemental Table 5.

Principal component analysis (PCA). Deviation scores of curated and probe double mutants were analyzed in PCA. The scripts using R language v4.0.3 (https://www.R-project.org/) with R packages tidyverse v1.3.1 (https://www.tidyverse.org), prompt (stats v3.6.2 (https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/prcomp)), ggplot2 v3.3.3 (https://ggplot2.tidyverse.org), dplyr v1.0.6 (https://dplyr.tidyverse.org), and missMDA v1.18 (https://dplyr.tidyverse.org), are available in GitHub (https://github.com/Kaplan-Lab-Pitt/TLs_Screening.git).

t-SNE projection. Allele frequencies for all mutants in nine conditions with three replicates were analyzed by t-SNE (Perplexity = 50) or k-means (clusters =20). Thirteen clusters with ultra-sick to lethal mutants as majority were eliminated. The remaining mutants were analyzed again with t-SNE (Perplexity = 100) and k-means (cluster =10). The scripts utilizing R language v4.0.3 (https://www.R-project.org/), along with R packages Rtsne v0.15 (https://github.com/jkrijthe/Rtsne), ggplot2 v3.3.3 (https://ggplot2.tidyverse.org), k-means (stats v3.6.2 (https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/kmeans)), are available through GitHub (https://github.com/Kaplan-Lab-Pitt/TLs_Screening.git).

Statistical coupling analysis. A published multiple sequence alignment (MSA) containing 5787 eukaryotic homologous sequences of yeast Rpb1 was used in the statistical coupling analysis (*82*). 1464 sequences were retained after sequence identity reducing to 90% with T-coffee package v12.00.7fb08c2 (*98*) through conda v4.6.14. Pol I, II, and III sequences were separated based on an ML tree constructed with FastTree 2 (*99*) and 410 Pol II Rpb1 homologous sequences were re-aligned with T-coffee, and the newly generated MSA was used for statistical coupling analysis with the python-based package pySCA v6.1 (*89*). The scripts were adapted from https://github.com/ranganathanlab/pySCA and are available via GitHub (https://github.com/Kaplan-Lab-Pitt/TLs_Screening.git).

Supplemental Tables

Supplemental Table 1. Strains and plasmids.

Supplemental Table 2. Primers.

Supplemental Table 3. Phenotyping details.

Supplemental Table 4. Library summary.

Supplemental Table 5. MLR_models_summary.

Supplemental Table 6. Kruskal-Wallis tests.

Supplemental methods

Supplemental Method 1. High efficiency large scale chemical yeast transformation protocol.

Supplemental Method 2. Emulsion PCR set up with EURx Micellula DNA Emulsion & Purification (ePCR) PCR kit.

Supplemental Method 3. Amplification/transformation/screening of mutant libraries and sequencing pool preparation.

Supplemental Method 4. Formulas of calculating functional interactions.

ACKNOWLEDGMENTS

We thank Dr. Anne-Ruxandra Carvunis (U. Pittsburgh) and Dr. Steve Lockless (Texas A&M) for discussions and advice. We thank Zhizhen Wang and Muyao Lin from the Pitt Statistical Consulting Center for their advice on checking the reproducibility of our data. We acknowledge funding from NIH R01GM097260 for initiation of this project and NIH R35GM144116 for this work. This research was supported in part by the University of Pittsburgh Center for Research Computing, RRID:SCR_022735, through the resources provided. Specifically, this work used the HTC cluster, which is supported by NIH award number S10OD028483.

REFERENCES

1. P. Cramer, Multisubunit RNA polymerases. *Curr Opin Struct Biol* **12**, 89-97 (2002).
2. F. Werner, D. Grohmann, Evolution of multisubunit RNA polymerases in the three domains of life. *Nat Rev Microbiol* **9**, 85-98 (2011).
3. L. A. Allison, M. Moyle, M. Shales, C. J. Ingles, Extensive homology among the largest subunits of eukaryotic and prokaryotic RNA polymerases. *Cell* **42**, 599-610 (1985).
4. G. Zhang, E. A. Campbell, L. Minakhin, C. Richter, K. Severinov, S. A. Darst, Crystal Structure of Thermus aquaticus Core RNA Polymerase at 3.3 Å Resolution. *Cell* **98**, 811-824 (1999).
5. D. G. Vassylyev, S. Sekine, O. Laptenko, J. Lee, M. N. Vassylyeva, S. Borukhov, S. Yokoyama, Crystal structure of a bacterial RNA polymerase holoenzyme at 2.6 A resolution. *Nature* **417**, 712-719 (2002).
6. A. Hirata, B. J. Klein, K. S. Murakami, The X-ray crystal structure of RNA polymerase from Archaea. *Nature* **451**, 851-854 (2008).
7. A. L. Gnatt, P. Cramer, J. Fu, D. A. Bushnell, R. D. Kornberg, Structural basis of transcription: an RNA polymerase II elongation complex at 3.3 A resolution. *Science* **292**, 1876-1882 (2001).
8. P. Cramer, D. A. Bushnell, R. D. Kornberg, Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. *Science* **292**, 1863-1876 (2001).
9. C. Fernandez-Tornero, M. Moreno-Morcillo, U. J. Rashid, N. M. Taylor, F. M. Ruiz, T. Gruene, P. Legrand, U. Steuerwald, C. W. Muller, Crystal structure of the 14-subunit RNA polymerase I. *Nature* **502**, 644-649 (2013).
10. N. A. Hoffmann, A. J. Jakobi, M. Moreno-Morcillo, S. Glatt, J. Kosinski, W. J. Hagen, C. Sachse, C. W. Muller, Molecular structures of unbound and transcribing RNA polymerase III. *Nature* **528**, 231-236 (2015).

11. A. M. Malinen, M. Turtola, M. Parthiban, L. Vainonen, M. S. Johnson, G. A. Belogurov, Active site opening and closure control translocation of multisubunit RNA polymerase. *Nucleic Acids Res* **40**, 7442-7451 (2012).
12. D. Wang, D. A. Bushnell, K. D. Westover, C. D. Kaplan, R. D. Kornberg, Structural basis of transcription: role of the trigger loop in substrate specificity and catalysis. *Cell* **127**, 941-954 (2006).
13. C. D. Kaplan, Basic mechanisms of RNA polymerase II activity and alteration of gene expression in Saccharomyces cerevisiae. *Biochimica Et Biophysica Acta-Gene Regulatory Mechanisms* **1829**, 39-54 (2013).
14. M. Dangkulwanich, T. Ishibashi, S. Liu, M. L. Kireeva, L. Lubkowska, M. Kashlev, C. J. Bustamante, Complete dissection of transcription elongation reveals slow translocation of RNA polymerase II in a linear ratchet mechanism. *Elife* **2**, e00971 (2013).
15. G. Bar-Nahum, V. Epshtein, A. E. Ruckenstein, R. Rafikov, A. Mustaev, E. Nudler, A ratchet mechanism of transcription elongation and its control. *Cell* **120**, 183-193 (2005).
16. R. O. Weinzierl, The nucleotide addition cycle of RNA polymerase is controlled by two molecular hinges in the Bridge Helix domain. *BMC Biol* **8**, 134 (2010).
17. C. Qiu, O. C. Erinne, J. M. Dave, P. Cui, H. Jin, N. Muthukrishnan, L. K. Tang, S. G. Babu, K. C. Lam, P. J. Vandeventer, R. Strohner, J. Van den Brulle, S. H. Sze, C. D. Kaplan, High-Resolution Phenotypic Landscape of the RNA Polymerase II Trigger Loop. *PLoS Genet* **12**, e1006321 (2016).
18. L. T. Da, F. Pardo-Avila, L. Xu, D. A. Silva, L. Zhang, X. Gao, D. Wang, X. Huang, Bridge helix bending promotes RNA polymerase II backtracking through a critical and conserved threonine residue. *Nat Commun* **7**, 11244 (2016).
19. A. Mazumder, M. Lin, A. N. Kapanidis, R. H. Ebright, Closing and opening of the RNA polymerase trigger loop. *Proc Natl Acad Sci U S A* **117**, 15642-15649 (2020).
20. B. Wang, A. V. Predeus, Z. F. Burton, M. Feig, Energetic and structural details of the trigger-loop closing transition in RNA polymerase II. *Biophys J* **105**, 767-775 (2013).
21. M. H. Larson, J. Zhou, C. D. Kaplan, M. Palangat, R. D. Kornberg, R. Landick, S. M. Block, Trigger loop dynamics mediate the balance between the transcriptional fidelity and speed of RNA polymerase II. *Proc Natl Acad Sci U S A* **109**, 6555-6560 (2012).
22. T. Fouqueau, M. E. Zeller, A. C. Cheung, P. Cramer, M. Thomm, The RNA polymerase trigger loop functions in all three phases of the transcription cycle. *Nucleic Acids Res* **41**, 7048-7059 (2013).
23. M. L. Kireeva, Y. A. Nedialkov, G. H. Cremona, Y. A. Purtov, L. Lubkowska, F. Malagon, Z. F. Burton, J. N. Strathern, M. Kashlev, Transient reversal of RNA polymerase II active site closing controls fidelity of transcription elongation. *Mol Cell* **30**, 557-566 (2008).
24. X. Liu, D. A. Bushnell, R. D. Kornberg, RNA polymerase II transcription: structure and mechanism. *Biochim Biophys Acta* **1829**, 2-8 (2013).
25. C. D. Kaplan, R. D. Kornberg, A bridge to transcription by RNA polymerase. *J Biol* **7**, 39 (2008).
26. L. Tan, S. Wiesler, D. Trzaska, H. C. Carney, R. O. Weinzierl, Bridge helix and trigger loop perturbations generate superactive RNA polymerases. *J Biol* **7**, 40 (2008).
27. D. A. Silva, D. R. Weiss, F. Pardo Avila, L. T. Da, M. Levitt, D. Wang, X. Huang, Millisecond dynamics of RNA polymerase II translocation at atomic resolution. *Proc Natl Acad Sci U S A* **111**, 7665-7670 (2014).
28. C. O. Barnes, M. Calero, I. Malik, B. W. Graham, H. Spahr, G. Lin, A. E. Cohen, I. S. Brown, Q. Zhang, F. Pullara, M. A. Trakselis, C. D. Kaplan, G. Calero, Crystal Structure of a Transcribing RNA Polymerase II Complex Reveals a Complete Transcription Bubble. *Mol Cell* **59**, 258-269 (2015).

29. N. Fong, H. Kim, Y. Zhou, X. Ji, J. Qiu, T. Saldi, K. Diener, K. Jones, X. D. Fu, D. L. Bentley, Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate. *Genes Dev* **28**, 2663-2676 (2014).
30. L. Xu, K. V. Butler, J. Chong, J. Wengel, E. T. Kool, D. Wang, Dissecting the chemical interactions and substrate structural signatures governing RNA polymerase II trigger loop closure by synthetic nucleic acid analogues. *Nucleic Acids Res* **42**, 5863-5870 (2014).
31. D. G. Vassylyev, M. N. Vassylyeva, J. Zhang, M. Palangat, I. Artsimovitch, R. Landick, Structural basis for substrate loading in bacterial RNA polymerase. *Nature* **448**, 163-168 (2007).
32. L. T. Da, D. Wang, X. Huang, Dynamics of pyrophosphate ion release and its coupled trigger loop motion from closed to open state in RNA polymerase II. *J Am Chem Soc* **134**, 2399-2406 (2012).
33. B. Liu, Y. Zuo, T. A. Steitz, Structures of E. coli sigmaS-transcription initiation complexes provide new insights into polymerase mechanism. *Proc Natl Acad Sci U S A* **113**, 4051-4056 (2016).
34. S. A. Seibold, B. N. Singh, C. Zhang, M. Kireeva, C. Domecq, A. Bouchard, A. M. Nazione, M. Feig, R. I. Cukier, B. Coulombe, M. Kashlev, M. Hampsey, Z. F. Burton, Conformational coupling, bridge helix dynamics and active site dehydration in catalysis by RNA polymerase. *Biochim Biophys Acta* **1799**, 575-587 (2010).
35. F. Malagon, M. L. Kireeva, B. K. Shafer, L. Lubkowska, M. Kashlev, J. N. Strathern, Mutations in the Saccharomyces cerevisiae RPB1 gene conferring hypersensitivity to 6-azauracil. *Genetics* **172**, 2201-2209 (2006).
36. C. D. Kaplan, K. M. Larsson, R. D. Kornberg, The RNA polymerase II trigger loop functions in substrate selection and is directly targeted by alpha-amanitin. *Mol Cell* **30**, 547-556 (2008).
37. C. D. Kaplan, H. Jin, I. L. Zhang, A. Belyanin, Dissection of Pol II trigger loop function and Pol II activity-dependent control of start site selection in vivo. *PLoS Genet* **8**, e1002627 (2012).
38. Matthew H. Larsona, Jing Zhoub,1, Craig D. Kaplanc,1, Murali Palangatd,2, Roger D. Kornberge, Robert Landickd, and Steven M. Blocka,b,f,3, Trigger loop dynamics mediate the balance between the transcriptional fidelity and speed of RNA polymerase II. (2012).
39. M. L. Kireeva, K. Opron, S. A. Seibold, C. Domecq, R. I. Cukier, B. Coulombe, M. Kashlev, Z. F. Burton, Molecular dynamics and mutational analysis of the catalytic and translocation cycle of RNA polymerase. *BMC Biophys* **5**, 11 (2012).
40. T. A. Windgassen, R. A. Mooney, D. Nayak, M. Palangat, J. Zhang, R. Landick, Trigger-helix folding pathway and SI3 mediate catalysis and hairpin-stabilized pausing by Escherichia coli RNA polymerase. *Nucleic Acids Res* **42**, 12707-12721 (2014).
41. D. Nayak, M. Voss, T. Windgassen, R. A. Mooney, R. Landick, Cys-pair reporters detect a constrained trigger loop in a paused RNA polymerase. *Mol Cell* **50**, 882-893 (2013).
42. A. C. Cheung, P. Cramer, Structural basis of RNA polymerase II backtracking, arrest and reactivation. *Nature* **471**, 249-253 (2011).
43. H. Kettenberger, K. J. Armache, P. Cramer, Complete RNA polymerase II elongation complex structure and its interactions with NTP and TFIIS. *Mol Cell* **16**, 955-965 (2004).
44. C. W. Lennon, W. Ross, S. Martin-Tumasz, I. Toulokhonov, C. E. Vrentas, S. T. Rutherford, J. H. Lee, S. E. Butcher, R. L. Gourse, Direct interactions between the coiled-coil tip of DksA and the trigger loop of RNA polymerase mediate transcriptional regulation. *Genes Dev* **26**, 2634-2646 (2012).
45. S. Sekine, Y. Murayama, V. Svetlov, E. Nudler, S. Yokoyama, The ratcheted and ratchetable structural states of RNA polymerase underlie multiple transcriptional functions. *Mol Cell* **57**, 408-421 (2015).

46. P. P. Hein, K. E. Kolb, T. Windgassen, M. J. Bellecourt, S. A. Darst, R. A. Mooney, R. Landick, RNA polymerase pausing and nascent-RNA structure formation are linked through clamp-domain movement. *Nat Struct Mol Biol* **21**, 794-802 (2014).
47. A. C. S. a. D. J. Taatjes, Structure and mechanism of the RNA polymerase II transcription machinery. (2020).
48. X. Y. Leng, M. Iyanov, P. Kindgren, I. Malik, A. Thieffry, P. Brodersen, A. Sandelin, C. D. Kaplan, S. Marquardt, Organismal benefits of transcription speed control at gene boundaries. *Embo Reports* **21**, (2020).
49. H. Braberg, H. Jin, E. A. Moehle, Y. A. Chan, S. Wang, M. Shales, J. J. Benschop, J. H. Morris, C. Qiu, F. Hu, L. K. Tang, J. S. Fraser, F. C. Holstege, P. Hieter, C. Guthrie, C. D. Kaplan, N. J. Krogan, From structure to systems: high-resolution, quantitative genetic analysis of RNA polymerase II. *Cell* **154**, 775-788 (2013).
50. B. C. Kaster, K. C. Knippa, C. D. Kaplan, D. O. Peterson, RNA Polymerase II Trigger Loop Mobility: INDIRECT EFFECTS OF Rpb9. *J Biol Chem* **291**, 14883-14895 (2016).
51. T. Tesileanu, L. J. Colwell, S. Leibler, Protein sectors: statistical coupling analysis versus conservation. *PLoS Comput Biol* **11**, e1004091 (2015).
52. P. C. Phillips, The language of gene interaction. *Genetics* **149**, 1167-1171 (1998).
53. M. S. Breen, C. Kemena, P. K. Vlasov, C. Notredame, F. A. Kondrashov, Epistasis as the primary factor in molecular evolution. *Nature* **490**, 535-538 (2012).
54. T. N. Starr, J. W. Thornton, Epistasis in protein evolution. *Protein Sci* **25**, 1204-1218 (2016).
55. E. A. Ortlund, J. T. Bridgham, M. R. Redinbo, J. W. Thornton, Crystal structure of an ancient protein: evolution by conformational epistasis. *Science* **317**, 1544-1548 (2007).
56. M. Karageorgi, S. C. Groen, F. Sumbul, J. N. Pelaez, K. I. Verster, J. M. Aguilar, A. P. Hastings, S. L. Bernstein, T. Matsunaga, M. Astourian, G. Guerra, F. Rico, S. Dobler, A. A. Agrawal, N. K. Whiteman, Genome editing retraces the evolution of toxin resistance in the monarch butterfly. *Nature* **574**, 409-412 (2019).
57. P. C. Phillips, Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet* **9**, 855-867 (2008).
58. A. J. Faure, J. Domingo, J. M. Schmiedel, C. Hidalgo-Carcedo, G. Diss, B. Lehner, Mapping the energetic and allosteric landscapes of protein binding domains. *Nature* **604**, 175-183 (2022).
59. Y. Park, B. P. H. Metzger, J. W. Thornton, Epistatic drift causes gradual decay of predictability in protein evolution. *Science* **376**, 823-830 (2022).
60. D. Ding, A. G. Green, B. Wang, T. V. Lite, E. N. Weinstein, D. S. Marks, M. T. Laub, Co-evolution of interacting proteins through non-contacting and non-specific mutations. *Nat Ecol Evol* **6**, 590-603 (2022).
61. A. S. Kondrashov, S. Sunyaev, F. A. Kondrashov, Dobzhansky-Muller incompatibilities in protein evolution. *Proc Natl Acad Sci U S A* **99**, 14878-14883 (2002).
62. M. Lunzer, G. B. Golding, A. M. Dean, Pervasive cryptic epistasis in molecular evolution. *PLoS Genet* **6**, e1001162 (2010).
63. C. Natarajan, N. Inoguchi, R. E. Weber, A. Fago, H. Moriyama, J. F. Storz, Epistasis among adaptive mutations in deer mouse hemoglobin. *Science* **340**, 1324-1327 (2013).
64. M. B. Doud, O. Ashenberg, J. D. Bloom, Site-Specific Amino Acid Preferences Are Mostly Conserved in Two Closely Related Protein Homologs. *Mol Biol Evol* **32**, 2944-2960 (2015).
65. T. N. Starr, A. J. Greaney, W. W. Hannon, A. N. Loes, K. Hauser, J. R. Dillen, E. Ferri, A. G. Farrell, B. Dadonaite, M. McCallum, K. A. Matreyek, D. Corti, D. Veesler, G. Snell, J. D. Bloom, Shifting mutational constraints in the SARS-CoV-2 receptor-binding domain during viral evolution. *Science* **377**, 420-424 (2022).

66. H. K. Haddox, A. S. Dingens, S. K. Hilton, J. Overbaugh, J. D. Bloom, Mapping mutational effects along the evolutionary landscape of HIV envelope. *Elife* **7**, (2018).
67. O. V. Viktorovskaya, K. L. Engel, S. L. French, P. Cui, P. J. Vandeventer, E. M. Pavlovic, A. L. Beyer, C. D. Kaplan, D. A. Schneider, Divergent contributions of conserved active site residues to transcription by eukaryotic RNA polymerases I and II. *Cell Rep* **4**, 974-984 (2013).
68. R. Mani, R. P. St Onge, J. L. t. Hartman, G. Giaever, F. P. Roth, Defining genetic interaction. *Proc Natl Acad Sci U S A* **105**, 3461-3466 (2008).
69. C. Qiu, C. D. Kaplan, Functional assays for transcription mechanisms in high-throughput. *Methods* **159-160**, 115-123 (2019).
70. X. Lin, Y. Liu, S. Liu, X. Zhu, L. Wu, Y. Zhu, D. Zhao, X. Xu, A. Chemparathy, H. Wang, Y. Cao, M. Nakamura, J. N. Noordermeer, M. La Russa, W. H. Wong, K. Zhao, L. S. Qi, Nested epistasis enhancer networks for robust genome regulation. *Science* **377**, 1077-1085 (2022).
71. D. M. Fowler, S. Fields, Deep mutational scanning: a new style of protein science. *Nat Methods* **11**, 801-807 (2014).
72. C. S. Sergey Ioffe, Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv:1502.03167v3*, (2015).
73. W. G. Hill, M. E. Goddard, P. M. Visscher, Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet* **4**, e1000008 (2008).
74. D. Melamed, D. L. Young, C. E. Gamble, C. R. Miller, S. Fields, Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein. *RNA* **19**, 1537-1551 (2013).
75. C. L. Araya, D. M. Fowler, W. Chen, I. Muniez, J. W. Kelly, S. Fields, A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proc Natl Acad Sci U S A* **109**, 16858-16863 (2012).
76. D. M. Fowler, C. L. Araya, S. J. Fleishman, E. H. Kellogg, J. J. Stephany, D. Baker, S. Fields, High-resolution mapping of protein sequence-function relationships. *Nat Methods* **7**, 741-746 (2010).
77. A. T. Carvalho, P. A. Fernandes, M. J. Ramos, The Catalytic Mechanism of RNA Polymerase II. *J Chem Theory Comput* **7**, 1177-1188 (2011).
78. X. Huang, D. Wang, D. R. Weiss, D. A. Bushnell, R. D. Kornberg, M. Levitt, RNA polymerase II trigger loop residues stabilize and position the incoming nucleotide triphosphate in transcription. *Proc Natl Acad Sci U S A* **107**, 15745-15750 (2010).
79. C. Castro, E. D. Smidansky, J. J. Arnold, K. R. Maksimchuk, I. Moustafa, A. Uchida, M. Gotte, W. Konigsberg, C. E. Cameron, Nucleic acid polymerases use a general acid for nucleotidyl transfer. *Nat Struct Mol Biol* **16**, 212-218 (2009).
80. I. C. Unarta, E. C. Goonetilleke, D. Wang, X. Huang, Nucleotide addition and cleavage by RNA polymerase II: Coordination of two catalytic reactions using a single active site. *J Biol Chem* **299**, 102844 (2023).
81. T. V. Mishanina, M. Z. Palo, D. Nayak, R. A. Mooney, R. Landick, Trigger loop of RNA polymerase is a positional, not acid-base, catalyst for both transcription and proofreading. *Proc Natl Acad Sci U S A* **114**, E5103-E5112 (2017).
82. M. Z. Palo, J. Zhu, T. V. Mishanina, R. Landick, Conserved Trigger Loop Histidine of RNA Polymerase II Functions as a Positional Catalyst Primarily through Steric Effects. *Biochemistry* **60**, 3323-3336 (2021).
83. M. J. Harms, J. W. Thornton, Analyzing protein structure and function using ancestral gene reconstruction. *Curr Opin Struct Biol* **20**, 360-366 (2010).
84. C. E. Scull, Z. M. Ingram, A. L. Lucius, D. A. Schneider, A Novel Assay for RNA Polymerase I Transcription Elongation Sheds Light on the Evolutionary Divergence of Eukaryotic RNA Polymerases. *Biochemistry* **58**, 2116-2124 (2019).

85. I. Toulokhonov, J. Zhang, M. Palangat, R. Landick, A central role of the RNA polymerase trigger loop in active-site rearrangement during transcriptional pausing. *Mol Cell* **27**, 406-419 (2007).
86. Y. Yuzenkova, A. Bochkareva, V. R. Tadigotla, M. Roghanian, S. Zorov, K. Severinov, N. Zenkin, Stepwise mechanism for transcription fidelity. *BMC Biol* **8**, 54 (2010).
87. W. Wang, C. Walmacq, J. Chong, M. Kashlev, D. Wang, Structural basis of transcriptional stalling and bypass of abasic DNA lesion by RNA polymerase II. *Proc Natl Acad Sci U S A* **115**, E2538-E2545 (2018).
88. N. Halabi, O. Rivoire, S. Leibler, R. Ranganathan, Protein sectors: evolutionary units of three-dimensional structure. *Cell* **138**, 774-786 (2009).
89. O. Rivoire, K. A. Reynolds, R. Ranganathan, Evolution-Based Functional Decomposition of Proteins. *PLoS Comput Biol* **12**, e1004817 (2016).
90. V. H. Salinas, R. Ranganathan, Coevolution-based inference of amino acid interactions underlying protein function. *Elife* **7**, (2018).
91. P. Cramer, Organization and regulation of gene transcription. *Nature* **573**, 45-54 (2019).
92. A. C. Schier, D. J. Taatjes, Structure and mechanism of the RNA polymerase II transcription machinery. *Genes Dev* **34**, 465-488 (2020).
93. S. M. Vos, L. Farnung, A. Linden, H. Urlaub, P. Cramer, Structure of complete Pol II-DSIF-PAF-SPT6 transcription complex reveals RTF1 allosteric activation. *Nat Struct Mol Biol* **27**, 668-677 (2020).
94. C. Castro, E. Smidansky, K. R. Maksimchuk, J. J. Arnold, V. S. Korneeva, M. Gotte, W. Konigsberg, C. E. Cameron, Two proton transfers in the transition state for nucleotidyl transfer catalyzed by RNA- and DNA-dependent RNA and DNA polymerases. *Proc Natl Acad Sci U S A* **104**, 4267-4272 (2007).
95. M. T. Gregory, Y. Gao, Q. Cui, W. Yang, Multiple deprotonation paths of the nucleophile 3'-OH in the DNA synthesis reaction. *Proc Natl Acad Sci U S A* **118**, (2021).
96. G. A. Belogurov, I. Artsimovitch, The Mechanisms of Substrate Selection, Catalysis, and Translocation by the Elongating RNA Polymerase. *J Mol Biol* **431**, 3975-4006 (2019).
97. C. D. K. Sing-Hoi Sze, Codon-Based Sequence Alignment for Mutation Analysis by High-Throughput Sequencing. *2018 IEEE 8th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS)*, (2018).
98. C. Notredame, D. G. Higgins, J. Heringa, T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**, 205-217 (2000).
99. M. N. Price, P. S. Dehal, A. P. Arkin, FastTree 2--approximately maximum-likelihood trees for large alignments. *Plos One* **5**, e9490 (2010).

FIGURES/LEGENDS

Figure 1

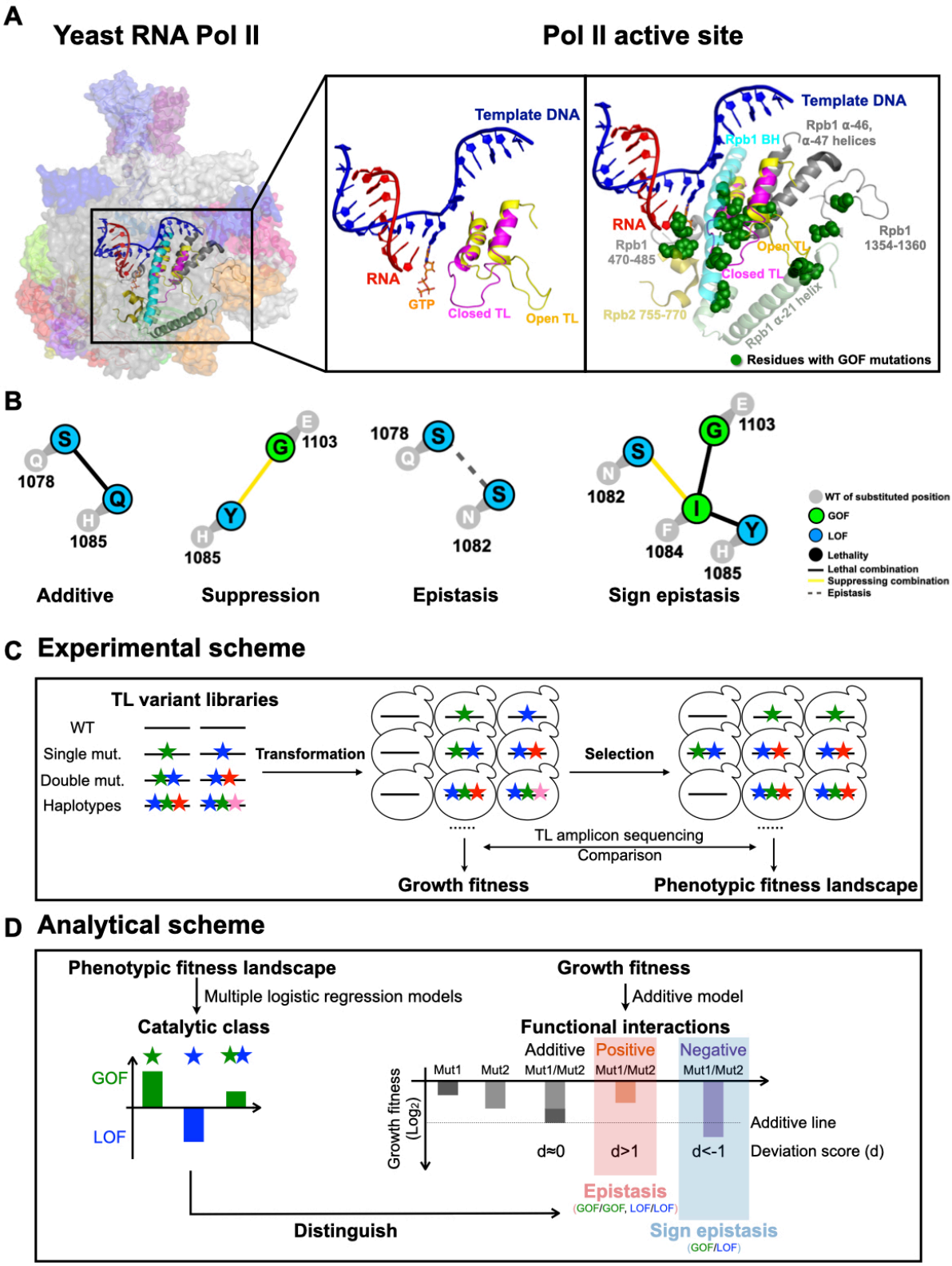


Figure 1. Schematics of the Pol II active site interaction landscape. A. The Pol II active site is embedded in the center of a 12-subunit complex (left panel). Pol II functions are supported by distinct TL conformational states. An open TL (PDB: 5C4X)(*28*) and closed TL (PDB: 2E2H)(*12*) conformations are shown in the middle panel. GOF mutations have been identified in the TL and its proximal domains (right panel), suggesting TL mobility and function may be impacted by adjacent residues. **B.** Examples of inter-residue genetic interactions. WT residues are shown in grey circles with number indicating residue position in Rpb1. Mutant substitutions are shown in colored circles, with color representing mutant class. Colored lines between mutant substitutions represent types of genetic interactions. **C.** Overview of experimental approach. We synthesized 10 libraries of TL variants represented by colored stars. Libraries were transformed into WT or mutated yeast strains. A selection assay was subsequently performed by scraping and replating the transformants onto different media for phenotyping. DNA was extracted from yeast from all conditions, and went through TL region amplification, and Illumina sequencing. Read counts for variants on general conditions were used to determine growth fitness, while read counts on other conditions were used to determine the phenotypic fitness landscape (see **Methods**). **D.** Overview of analytical approach for determining interaction landscape. Mutant conditional growth fitnesses were calculated using allele frequencies under selective growth conditions and subjected to two logistic regression models for classification/prediction of catalytic defects. Double mutant interactions were computed using growth fitness. Classification allowed epistatic interactions to be deduced from double mutant growth fitness (see **Methods**).

Figure 2

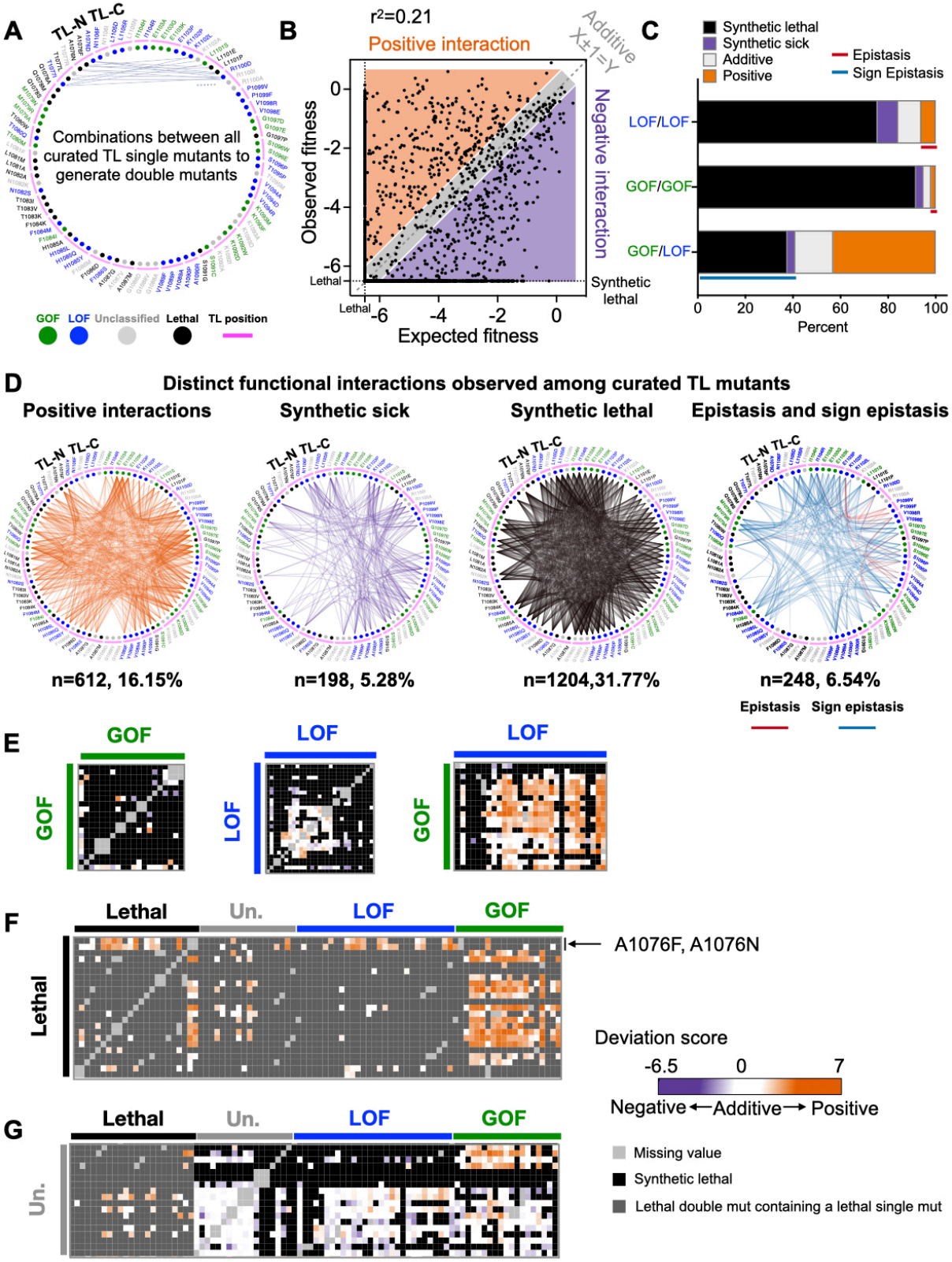
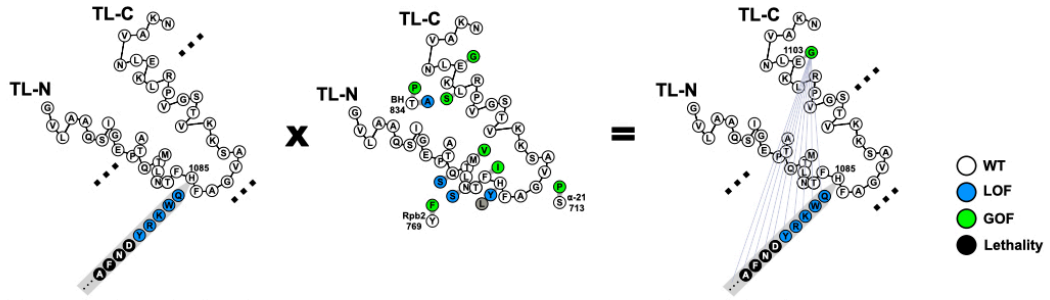

A
TL-N TL-C
Combinations between all curated TL single mutants to generate double mutants
GOF LOF Unclassified Lethal TL position
B
r²=0.21
Positive interaction
Additive X±1=Y
Negative interaction
Synthetic lethal
Observed fitness
Expected fitness
Lethal
C
Synthetic lethal
Synthetic sick
Additive
Positive
Epistasis
Sign Epistasis
LOF/LOF
GOF/GOF
GOF/LOF
Percent
D
Distinct functional interactions observed among curated TL mutants
Positive interactions
Synthetic sick
Synthetic lethal
Epistasis and sign epistasis
n=612, 16.15%
n=198, 5.28%
n=1204,31.77%
n=248, 6.54%
Epistasis
Sign epistasis
E
GOF
LOF
F
Lethal
Un.
A1076F, A1076N
Deviation score
-6.5
0
7
Negative ← Additive → Positive
G
Missing value
Synthetic lethal
Lethal double mut containing a lethal single mut

Figure 2. Widespread epistasis in the Pol II TL interaction landscape. A. Design of the pairwise double mutant library. We curated 2-4 substitutions for each TL residue (in total 90 substitutions, n(GOF) = 18, n(LOF) = 30, n(Unclassified) = 19, n(Lethal) = 23), and combined them with each other to generate double mutants. 3910 double mutants representing combinations between any two TL residues were measured and 3790 of them passed the reproducibility filter. WT TL residue positions are indicated with magenta arch. Phenotype classes of single substitutions are shown as colored circles (GOF in green, LOF in blue) while unclassified mutants are in grey and lethal mutants are in black. **B.** An xy-plot of observed double mutant growth fitness measured in our experiment (Y-axis) and expected fitness from the addition of two constituent single mutants' fitnesses (X-axis). N (positive) = 612. N (Negative) = 1402. N (Additive) = 1776. N (Sum)=3790. Lethal threshold (-6.5) is labeled with dotted lines on X and Y axis. The additive line where $X \pm 1 = Y$ is indicated with dashed line. Simple linear regression was performed, and the best fit equation is $Y = 0.52X - 2.55$, $r^2 = 0.21$, $P < 0.0001$. **C.** Percent of interactions observed from each combination group. N (LOF/LOF) = 412. N (GOF/GOF) = 156. N (GOF/LOF) = 534. Epistasis and sign epistasis are indicated with colored lines. **D.** Various groups of interactions are displayed in network format. **E-G.** The intra-TL functional interaction heatmaps of various combinations. Double mutant deviation scores are shown in the heatmap. Annotations at the top and left indicate the curated single mutants and their predicted phenotypic classes from multiple logistic regression modeling. GOF/GOF, LOF/LOF, and GOF/LOF combinations are shown in E. Combinations with lethal single substitutions are in F. Combinations with unclassified mutants are in G.

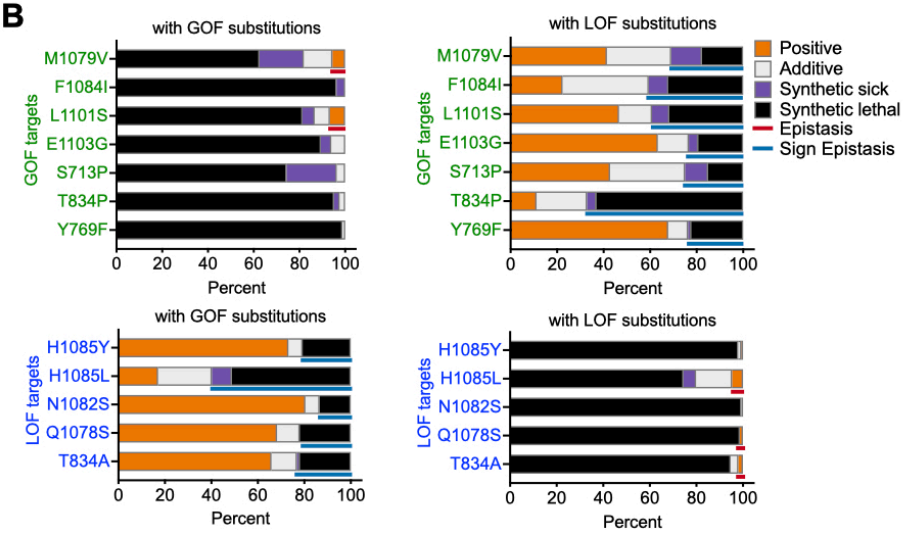
Figure 3

A

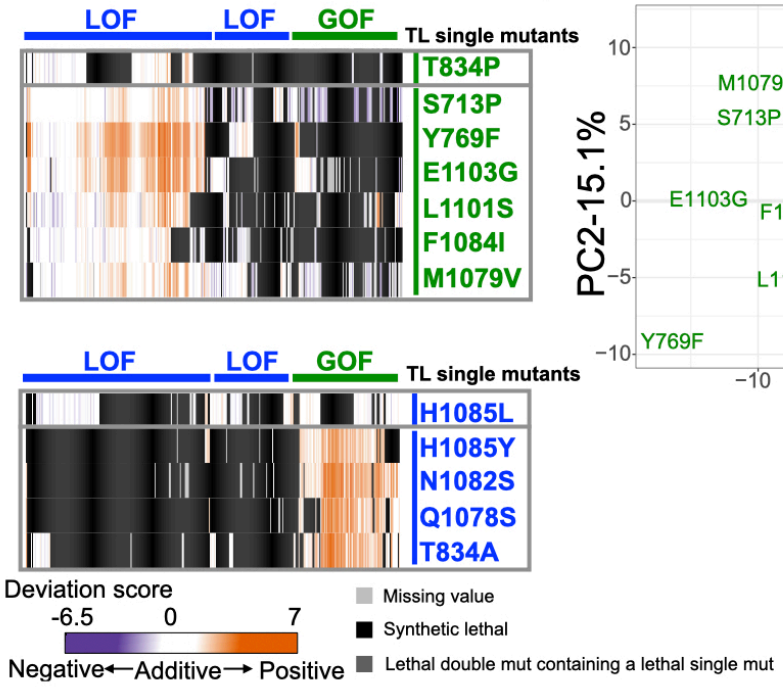


All TL single substitutions (n=620) | Target substitutions (n=12) | Target double mutants (n=7276)

B



C TL functional interaction landscape



D

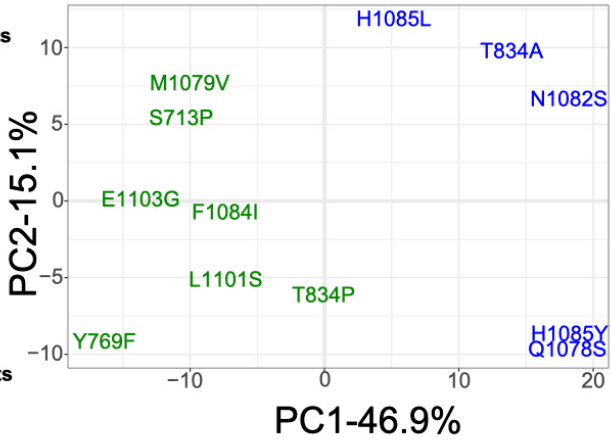


Figure 3. Pol II TL interaction landscape distinguishes mutants with similar phenotypes. **A.** Design of the targeted double mutant libraries. All possible substitutions at each TL residue (represented with a simplified format in the left panel) and twelve "probe" mutations (eight within the TL and four in TL-proximal domains) (middle panel) were combined with to generate 7280 double mutants (right panel). 7276 mutants passed the reproducibility filter and were used for interaction analyses. **B.** The percentage of functional interactions observed for each probe mutant with viable GOF or LOF TL substitutions. Epistasis and sign epistasis are labeled with colored lines. **C.** Pol II-TL functional interaction landscape with interactions represented by deviation scores. The upper panel shows interactions of GOF probe mutants in combination of viable GOF or LOF TL substitutions. The lower panel shows interactions of combinations with LOF probe mutants. **D**. Principal component analysis (PCA) of deviation scores across double mutant interactions for 12 probe mutants (see **Methods**).

Figure 4

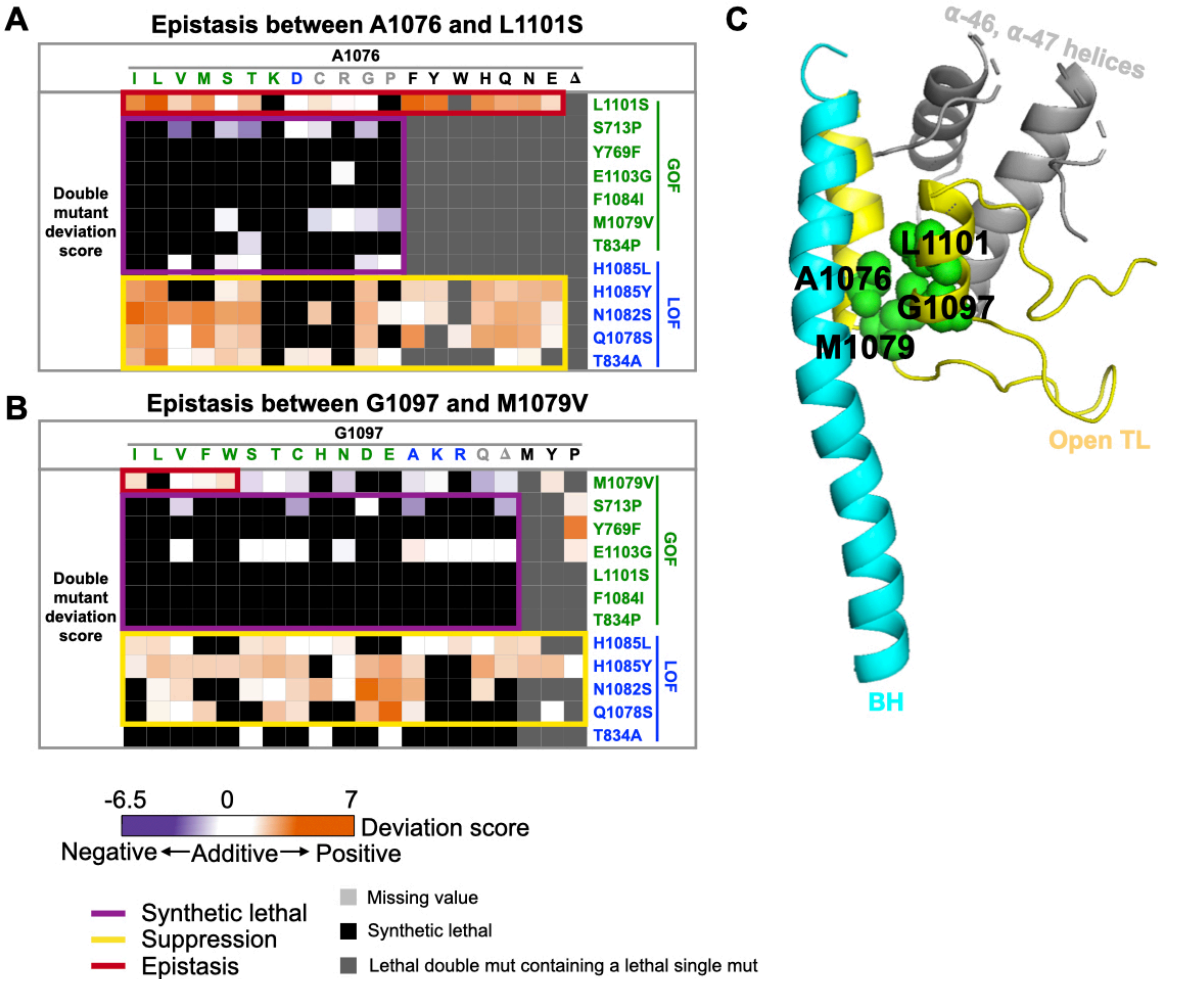


Figure 4. Pol II TL interaction landscape reveal functional dependency of proximal residues. A-B. Specific epistatic interactions observed between hydrophobic residues A1076 and L1101 (A), and M1079 and G1097 are shown as heatmaps (B). The x-axis of both heatmaps are 20 substitutions ordered by predicted phenotypic classes, and the color of substitution represents the phenotypic class of the substitution. GOF substitution is in green, LOF is in blue, unclassified is in gray, and lethal (fitness < -6.5) is in black. **C.** The epistatic interactions we identified between A1076 and L1101, together with M1079 and G1097 are shown on the five-helix bundle of Pol II active site (PDB:5C4X)(*28*).

Figure 5

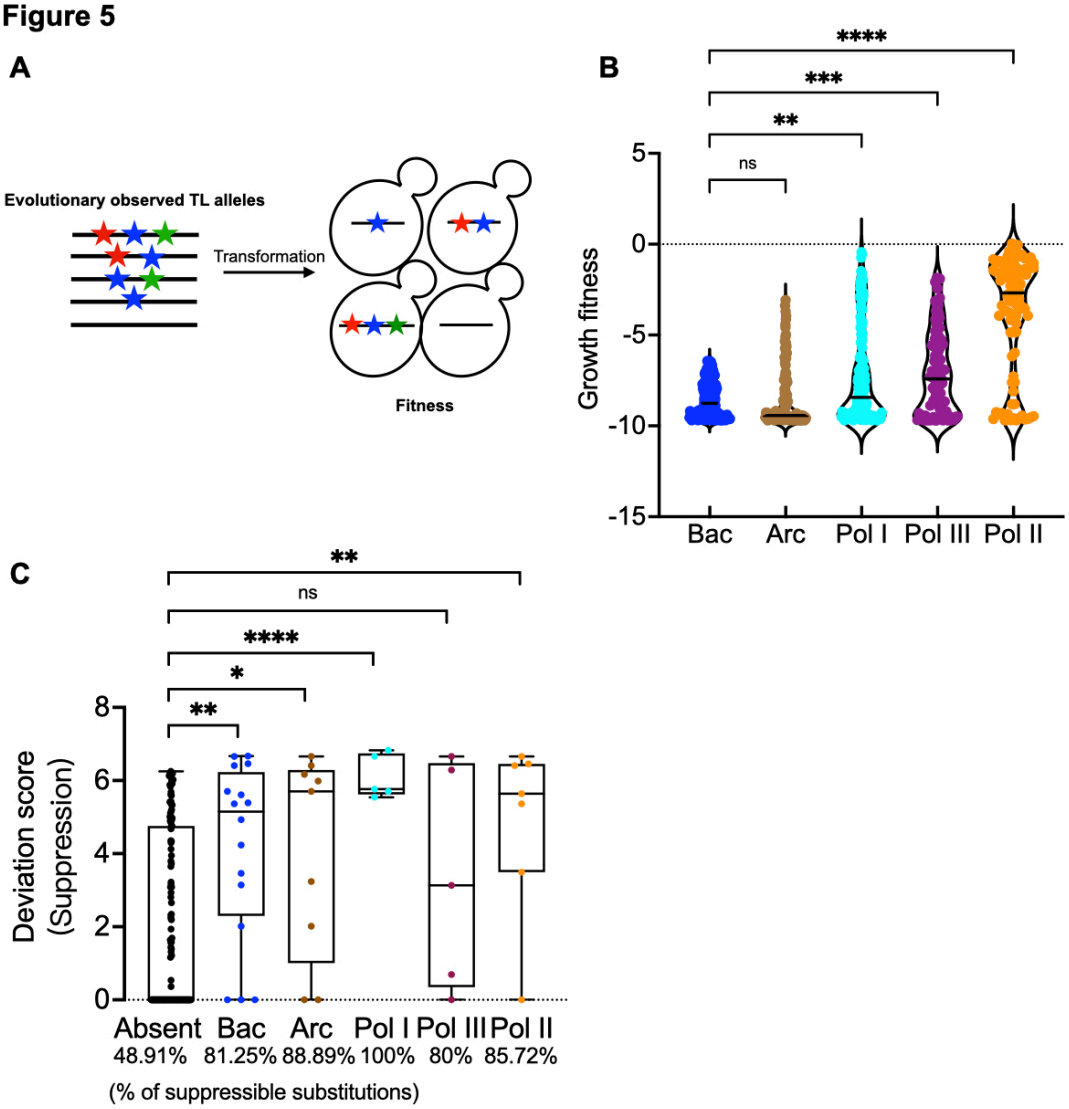


Figure 5. Contextual epistasis shapes TL evolution. A. Schematic for the TL evolutionary haplotypes library. We selected 662 TL haplotypes representing TL alleles from bacterial, archaeal and the three conserved eukaryotic msRNAPs. These TL alleles were transformed into yeast and were phenotyped under selective conditions. **B.** Fitness of evolutionarily observed TL haplotypes in the yeast Pol II background. The Pol II WT TL fitness (0) is labeled as dotted line. Kruskal-Wallis test was performed for comparison and significant levels ($P < 0.05$) were labeled. **C.** A comparison of the maximum deviation score of each TL lethal single substitution that is present in any evolutionary TL haplotypes from bacterial, archaeal or eukaryotic Pols versus those that have not been observed in any species. The evolutionary TL haplotypes were from multiple sequence alignments (MSA). 9 substitutions were found in an MSA of 542 archaeal TL sequences that are lethal when present in yeast as a single substitution. 17 were found in an MSA of 1403 bacterial TLs, 5 were found in 749 Pol I TLs, 7 were found in 499 Pol II TLs, and 5 were found in 539 Pol III TLs. Evolutionarily observed lethal substitutions were compared to those unobserved in our TL MSA. The percentage of in total suppressible lethal single mutants for each group is labeled at the bottom of the plot. Statistical comparison was done with the Mann-Whitney test and the significant levels ($P < 0.05$) are shown in the figure.

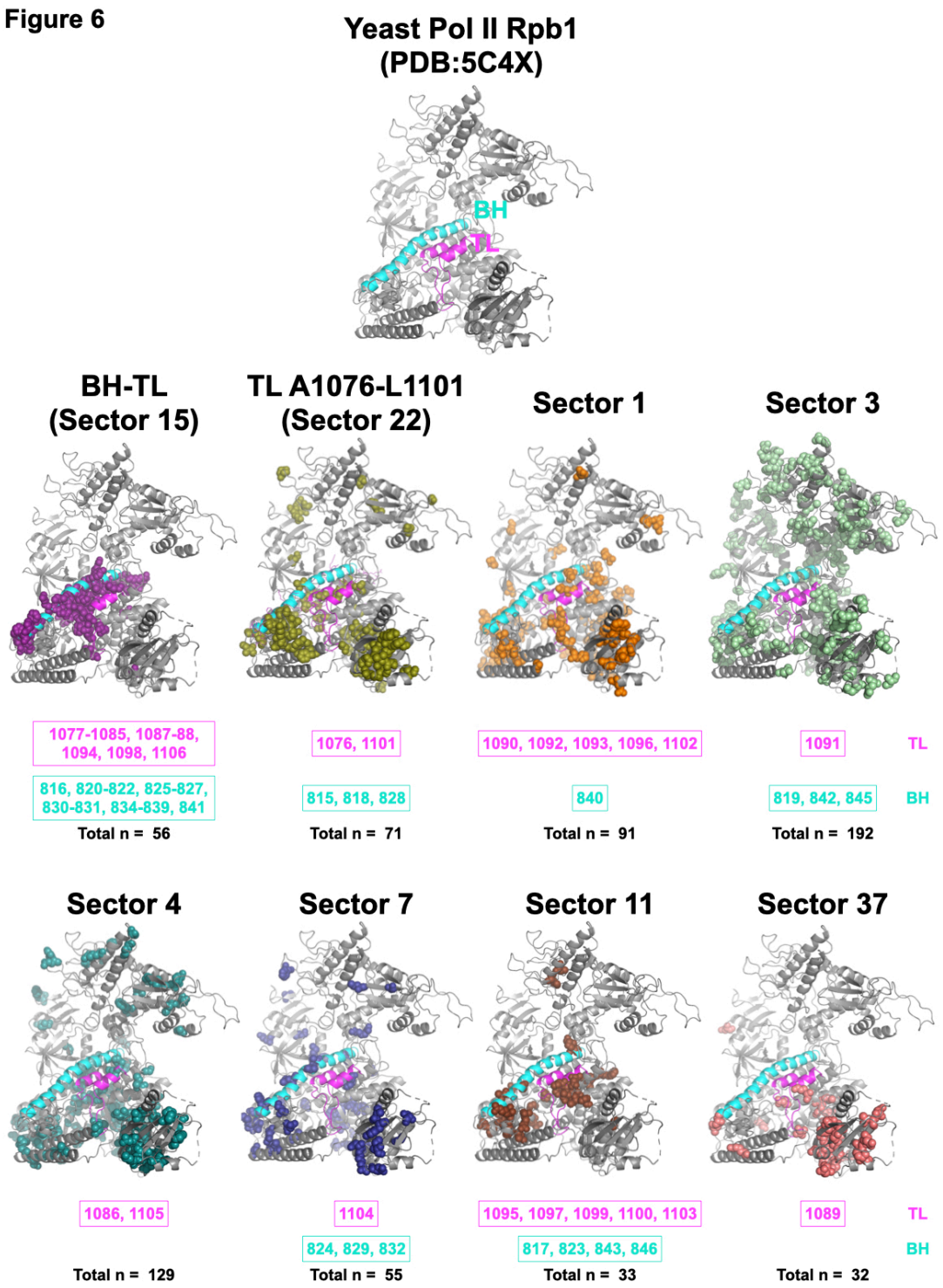


Figure 6. Ultra-conserved TL co-evolves with Pol II residues through diverse pathways. The eight coevolution sectors containing any TL residues that were identified from statistical coupling analysis are shown on the yeast Pol II Rpb1 structure (PDB:5C4X)(*28*). The TL is labeled with magenta and the BH is labeled with cyan. The TL and BH residues in each sector are labeled at the bottom of each sector. The total number of residues within each sector is also shown. The details of the statistical coupling analysis are in **Methods**.

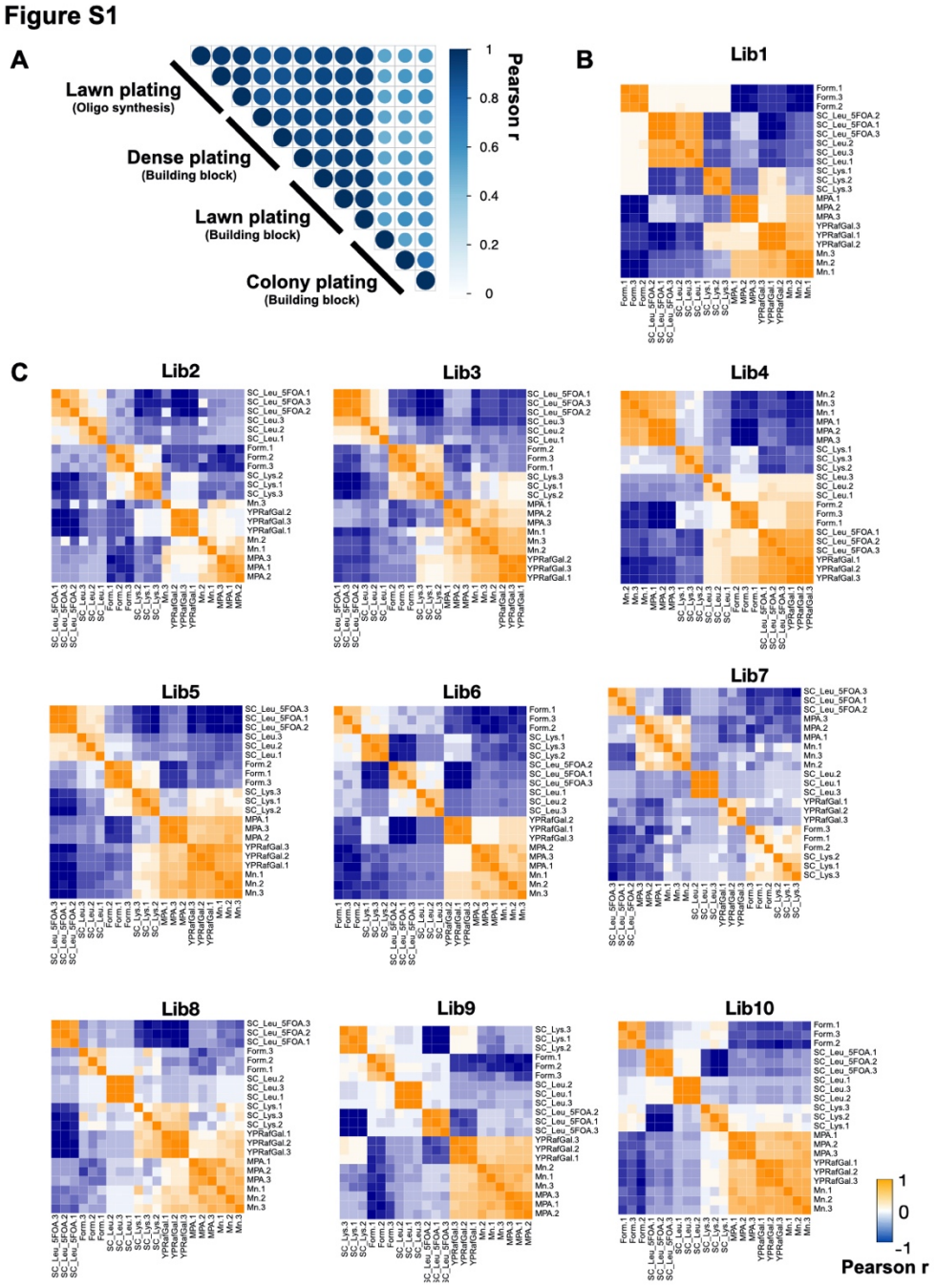


Figure S1. Pol II deep mutational scanning is highly reproducible. A. Single mutant growth fitness from mutants in libraries constructed from synthesized oligos correlated well with our previous library constructed by a random building block approach when plating conditions were the same. Qiu *et al*(*17*) plated at a lower density (colony plating) that we speculated added noise to the analysis. When plating densely ("dense" and "lawn" conditions) our new and old libraries showed highly reproducible fitness determinations for single mutants. **B-C.** Biological replicates for each library showed high reproducibility for all conditions. Pearson correlation of each library was calculated with three replicates for viable mutant fitness on all selective conditions.

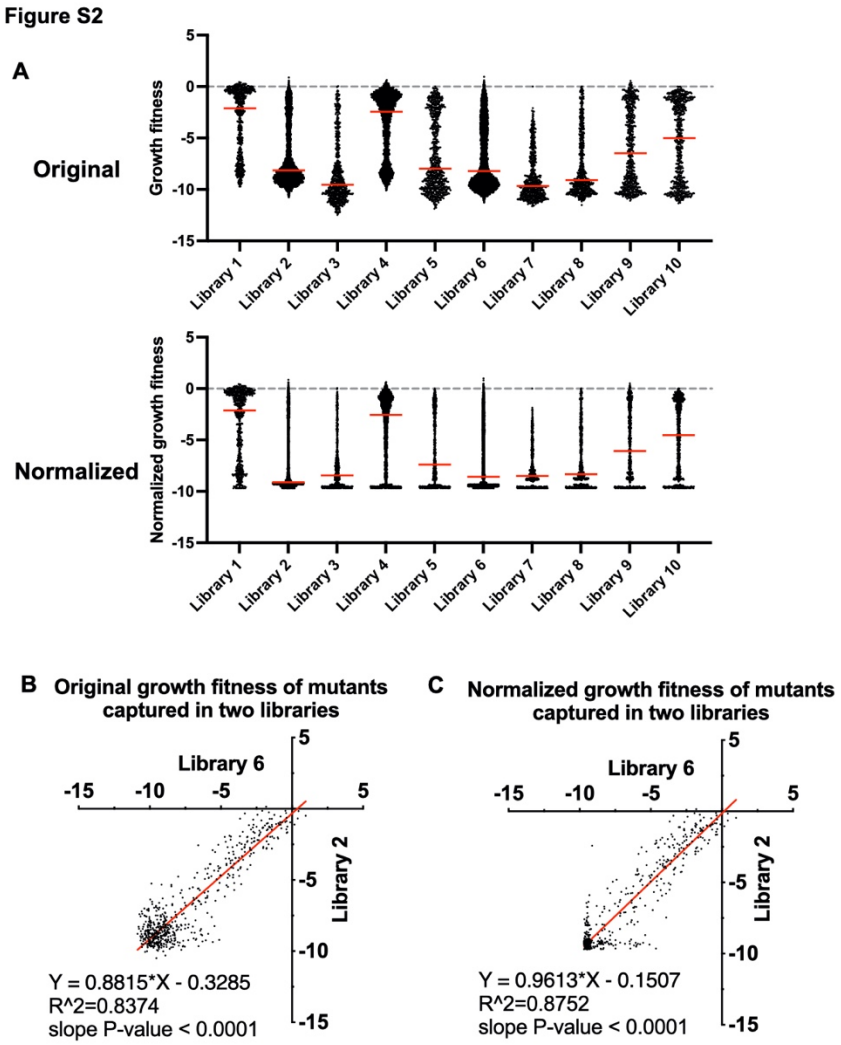


Figure S2. Min-Max normalization uniformed the fitness level of lethal mutants from different libraries without disturbing the median of mutant fitness. A. Library growth fitness distributions before and after normalization. Upper panel: The growth fitness distributions of libraries. The lowest fitness levels (fitness of lethal mutants) are different among libraries. To uniform various lowest fitness levels, we applied Min-Max normalization to minimize library effects on fitness ranges (See **Methods** for details). Lower panel: Libraries fitness distributions after normalization. The lethal mutant fitness levels of libraries were normalized to the same level while the median fitness for each library was not affected by the normalization. **B-C.** XY-plots showing the original fitness of mutants captured in two different libraries (n=586) (B). These mutants present in two libraries showed improved correlation between measurements upon normalization (C).

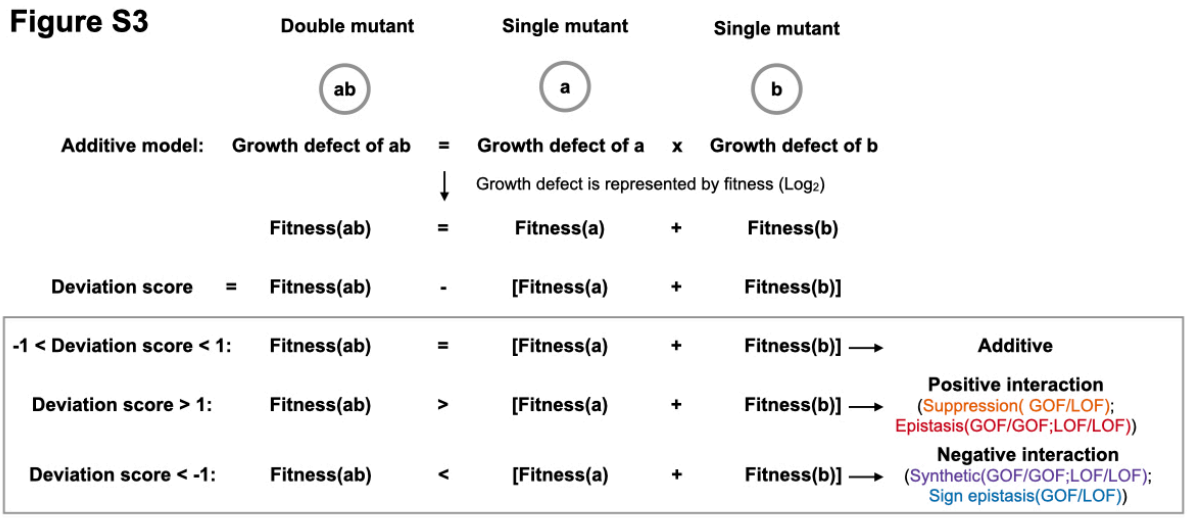


Figure S3. Detection of functional interactions by deviation score. For a pseudo double mutant *ab*, the difference between its observed fitness (*ab*) and expected fitness (*ab*) adding the fitness of two constituent single mutants (*a* and *b*) determines the type of interaction between the two mutants. Positive or negative interactions are determined if the deviation score is greater than 1 or smaller than –1. Specific epistatic interactions are further distinguished from general suppression or synthetic sick or lethal interactions using predicted mutant catalytic defect classes (GOF or LOF).

Figure S4

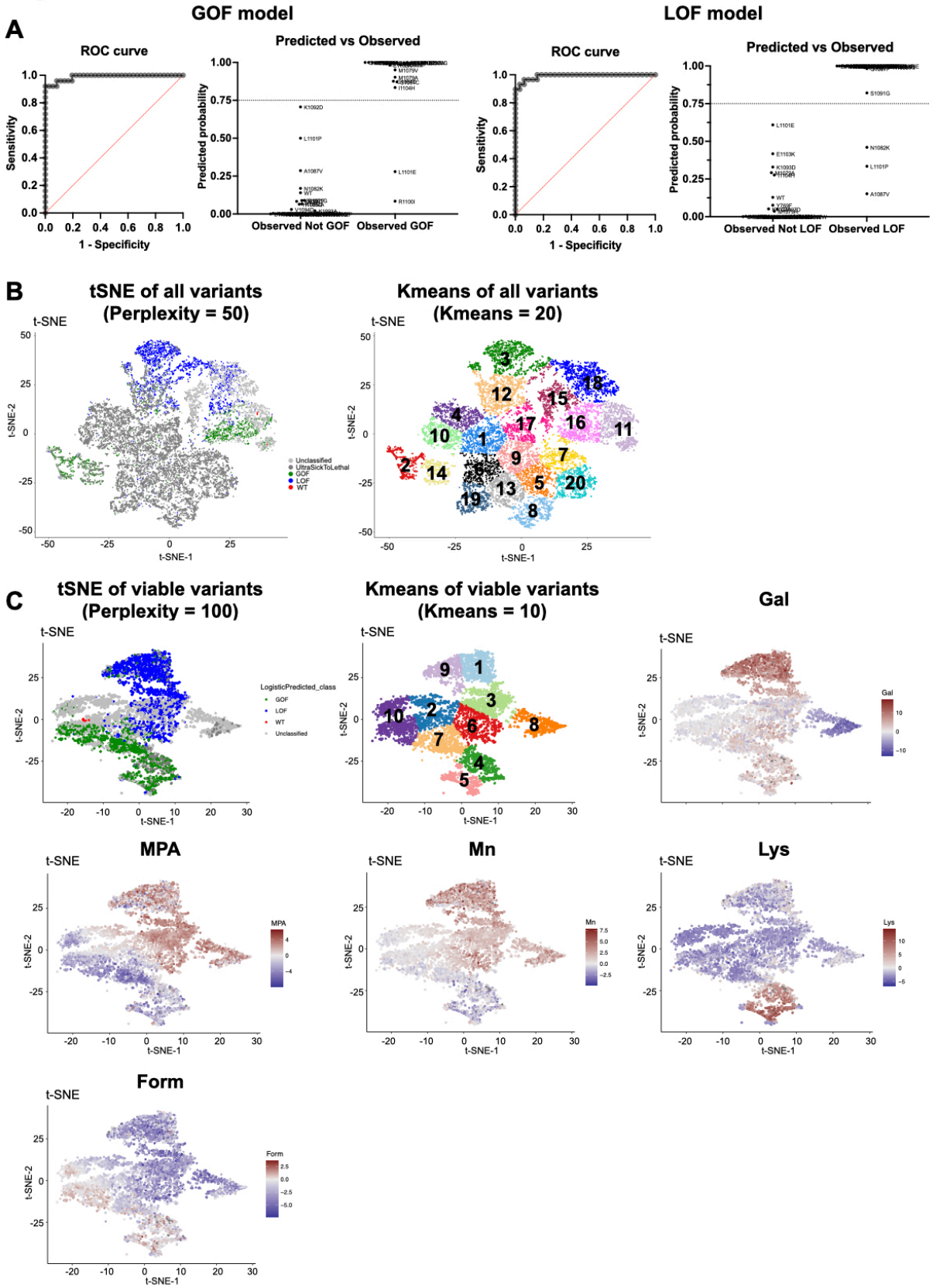

A
GOF model
LOF model
ROC curve
Predicted vs Observed
Sensitivity
1 - Specificity
Predicted probability
Observed Not GOF
Observed GOF
Observed Not LOF
Observed LOF
B
tSNE of all variants
(Perplexity = 50)
Kmeans of all variants
(Kmeans = 20)
t-SNE
t-SNE-1
t-SNE-2
Unclassified
UltraSickToLethal
GOF
LOF
WT
C
tSNE of viable variants
(Perplexity = 100)
Kmeans of viable variants
(Kmeans = 10)
Gal
MPA
Mn
Lys
Form
LogisticPredicted_class

Figure S4. Classification of mutant catalytic defects with machine learning algorithms. A. ROC curves for two multiple logistic regression models used to determine mutant catalytic class. Using 65 mutants with validated in vitro determined catalytic defects and conditional growth fitness measured in our experiment, we trained two models to classify variants as GOF or LOF. The GOF AUROC is 0.9889 ($P \leq 0.0001$), whereas the LOF ROC is 0.9914 ($P \leq 0.0001$). The predicted vs. observed graphs display the predicted probability of 65 known mutants would be GOF or LOF. The threshold we used to determine GOF or LOF mutations is shown by lines at 0.75. Details of the models are in Supplemental Table 5. **B.** Left: t-SNE projection of all mutants (n=15174) with perplexity = 50. Right: k-means cluster of all mutants with 20 clusters. The t-SNE and k-means projections suggest GOF are in 3 clusters (cluster 2, 14, and 16), LOF are in 2 clusters (cluster 3 and 18), and unclassified mutants are in 2 clusters (11 and 15). Most ultra-sick/lethal mutants (fitness <= -6.5) are projected together into 13 clusters, likely due to significant noise from low read counts across conditions. **C.** Feature plot of viable mutations in t-SNE and k-means projections (n=6054). Ultra-sick/lethal mutations were removed and the viable mutants were projected with t-SNE (perplexity = 100) and K-means (10 clusters). GOF were grouped into 4 clusters (4, 5, 7 and 10) and LOF were in 4 clusters (1, 3, 6, and 9). Each spot in the projection represents a mutant and it is colored based on the fitness of the mutant in selective conditions. GOF and LOF mutants in different clusters are related to various phenotype patterns. GOF clusters 7 and 10 are defined by strong MPA^S, while clusters 4 and 5 show slight MPA^S, Gal^R, Mn^S, but strong Lys^+. Slight $Form^S$ is a common feature across four GOF clusters. LOF clusters 3 and 6 show slight Mn^R, while clusters 1 and 9 are strongly Mn^R and Gal^R. There are three common features in all LOF clusters: MPA^R, $Form^S$, and Lys^-. Cluster 8, which mostly contain unclassified mutants, appear defined by Gal super sensitivity, indicating a potential specific defect defining this cluster.

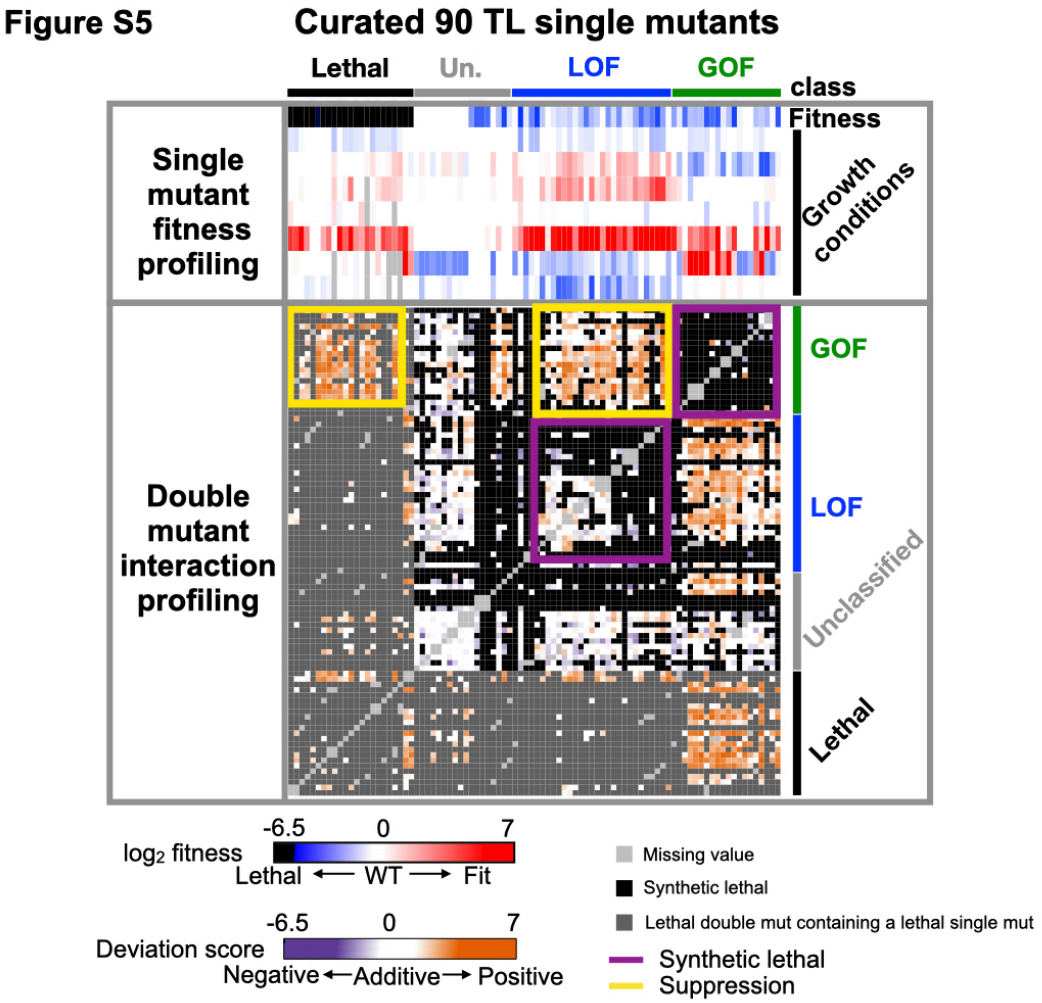


Figure S5. The intra-TL functional interaction landscape. The intra-TL functional interaction landscape is shown as a heatmap. Annotations at the top and right indicate the 90 curated single mutants and their predicted phenotypic classes from multiple logistic regression modeling. The upper part of the heatmap shows single mutant growth fitness profiling across multiple phenotypes ordered by groups predicted with logistic regression models. The lower part of the heatmap shows double mutant deviation scores where a colored block at the interaction of x and y coordinates indicates deviation score of the double mutant.

Figure S6

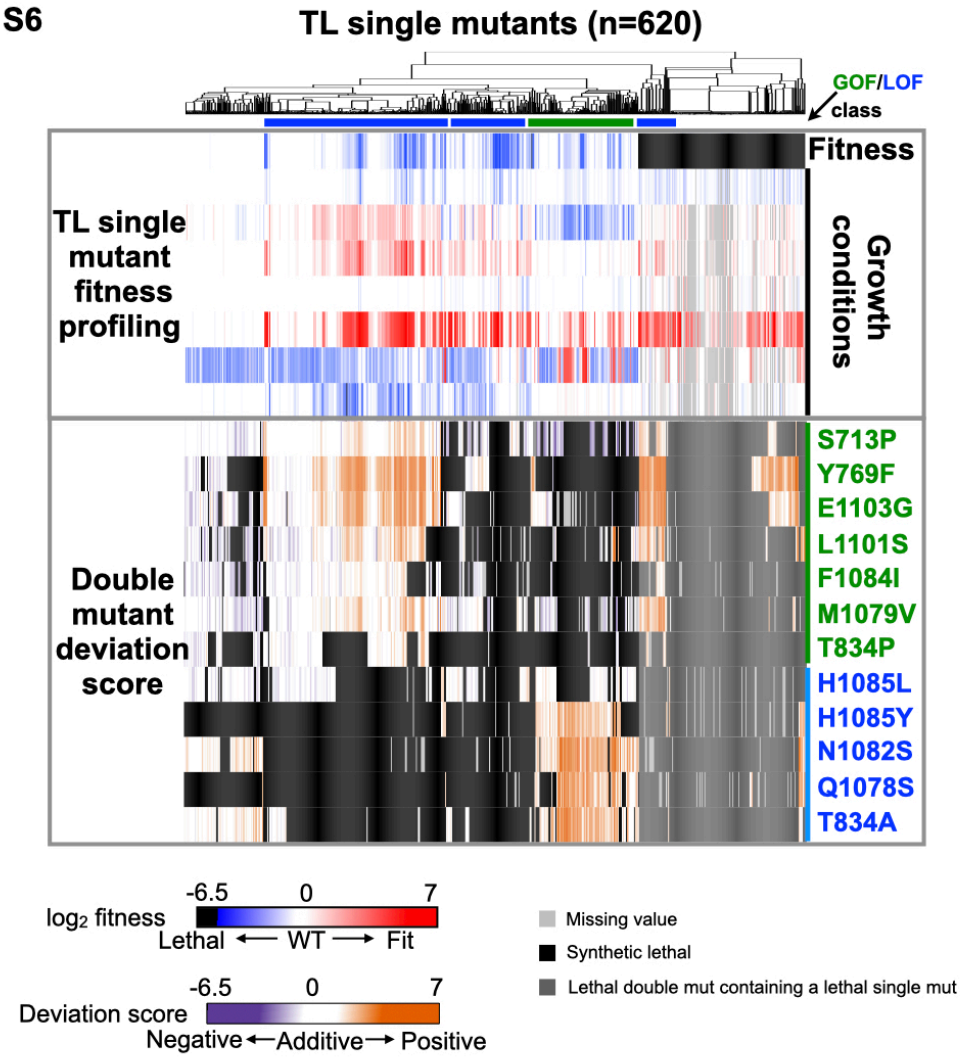


Figure S6. The functional interaction landscape of probe mutants. The functional interaction landscape is shown as a heatmap. The upper part of the heatmap shows all Pol II TL single mutant growth fitness profiling across several phenotypes and the single mutants were ordered by hierarchical clustering with Euclidean distance. The lower part of the heatmap shows double mutant deviation scores where a colored block at the interaction of x and y coordinates indicates deviation score of the double mutant.

Figure S7

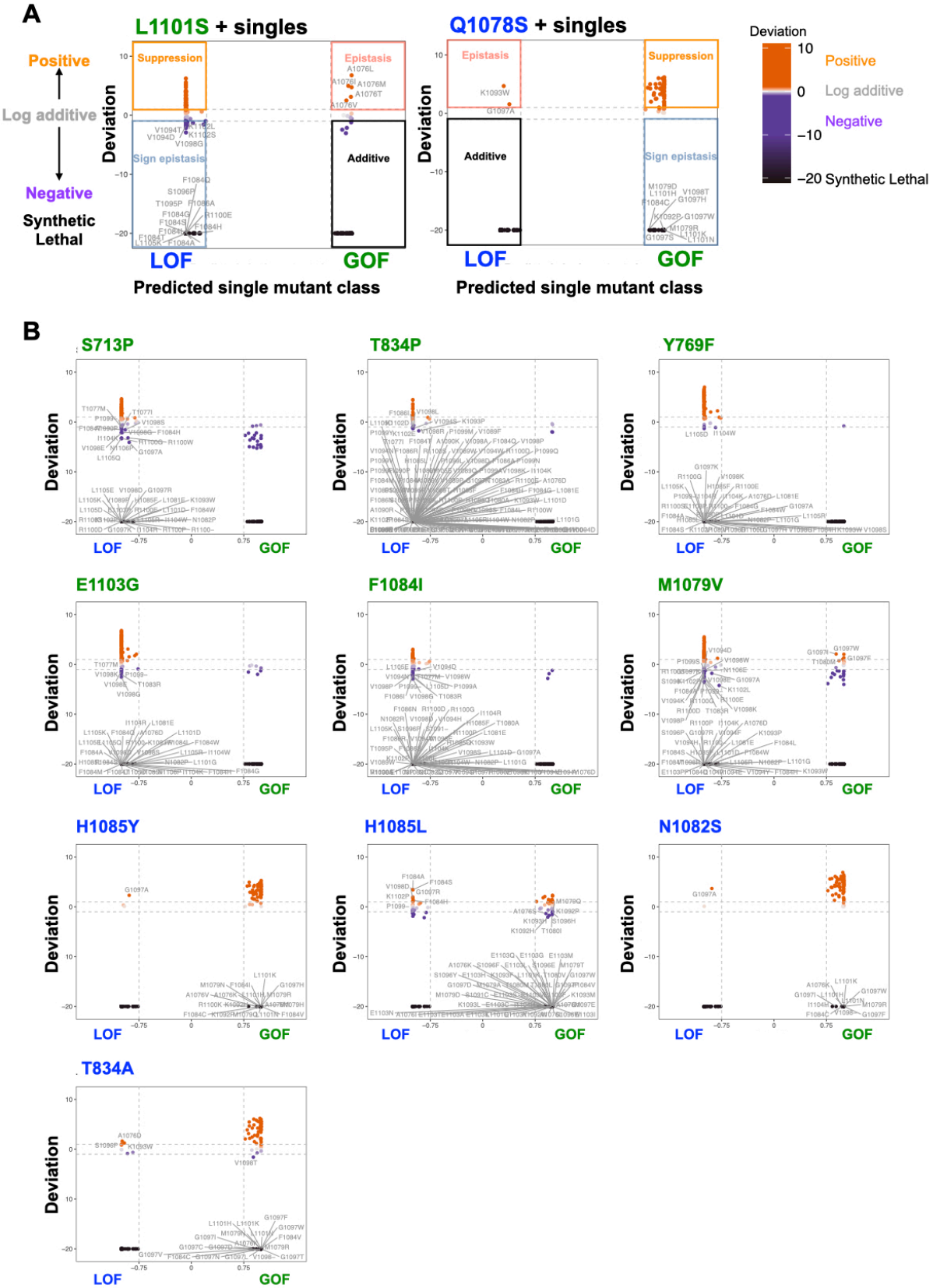

A
L1101S + singles
Q1078S + singles
Positive
Log additive
Negative
Synthetic Lethal
Deviation
Suppression
Epistasis
Sign epistasis
Additive
LOF
GOF
Predicted single mutant class
Deviation
10
0
−10
−20 Synthetic Lethal
B
S713P
T834P
Y769F
E1103G
F1084I
M1079V
H1085Y
H1085L
N1082S
T834A

Figure S7. Identifying TL substitutions that interact with the probe mutants. A. Identification of epistasis and suppression within positive interactions, and sign epistasis and synthetic sickness/lethality within negative interactions in two probe mutants, L1101S and N1082S. The deviation score of combinations (y-axis) between probe mutants and TL GOF or LOF single mutants were plotted versus the predicted probability of single mutants being GOF or LOF (x-axis). **B.** The scatter plots for distinguishing interactions of the other 10 TL probe mutants.

Figure S8

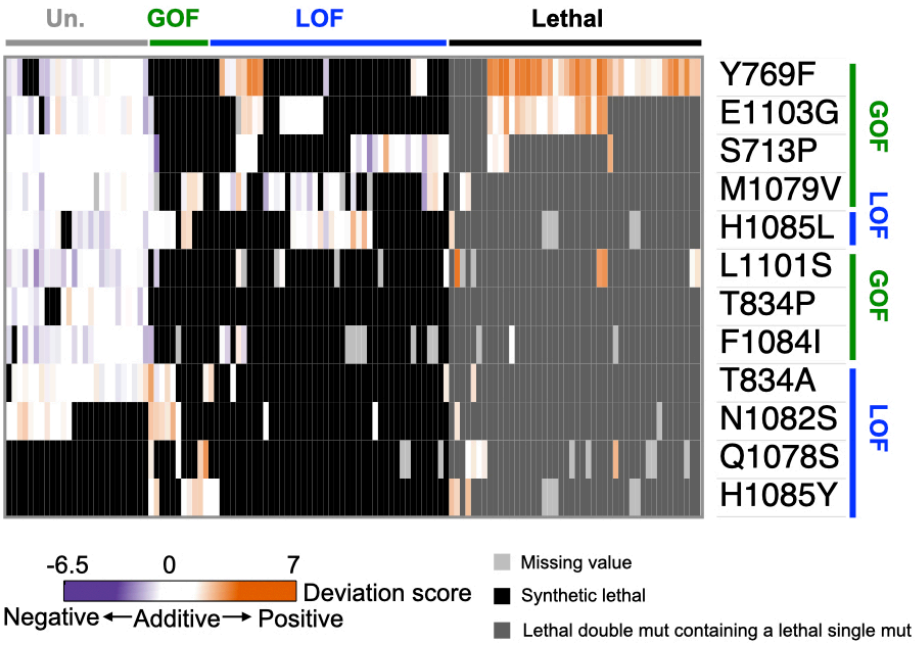


Figure S8. Allele-specific interactions. Unique interactions observed between TL substitutions and probe mutants. For each substitution, we analyzed the interquartile range (IQR) of their deviation scores with all probe mutants. Any substitution with deviation score(s) outside of the IQR were extracted and called as unique interaction(s). 127 substitutions with unique interactions with probe mutants were found out of 620 and are shown in the heatmap. The heatmap was hierarchical clustered with Euclidean distance for both rows and columns.

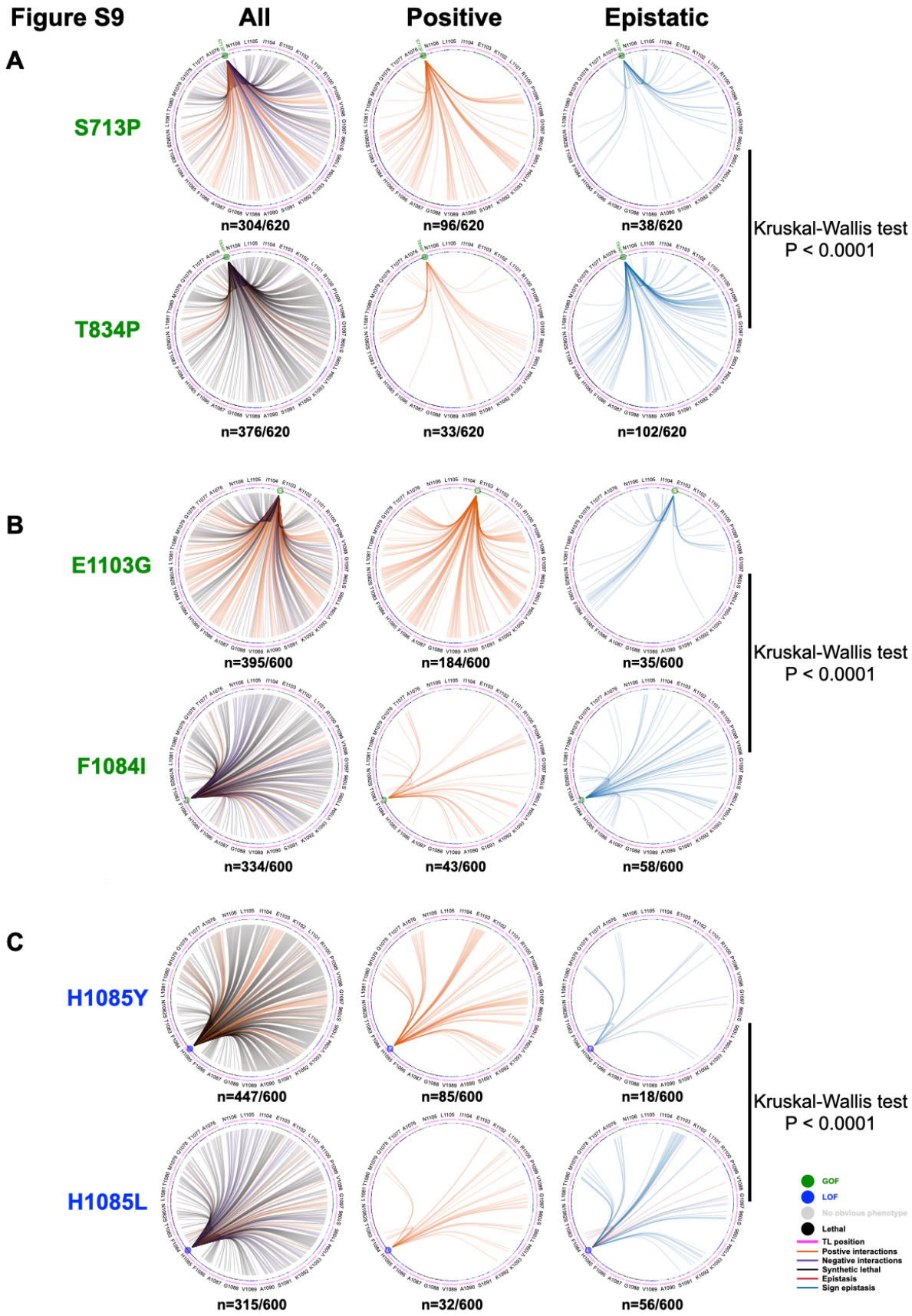
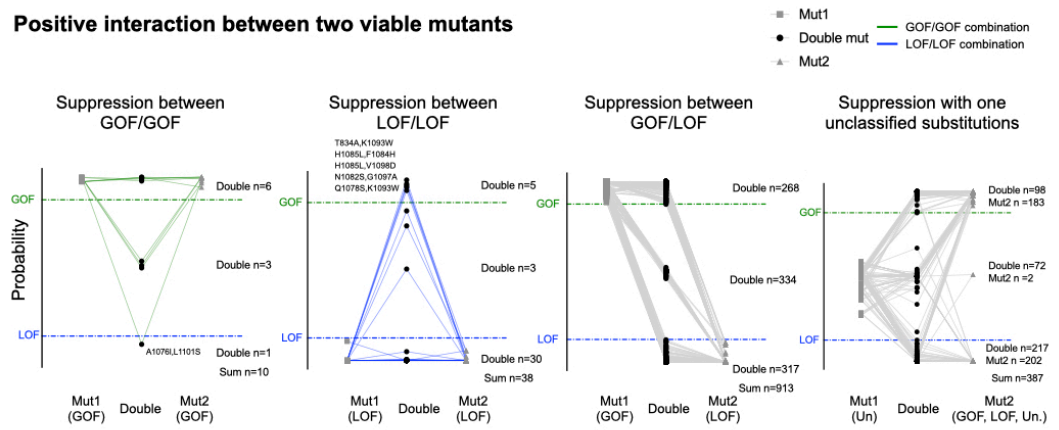


Figure S9
All
Positive
Epistatic
A
S713P
n=304/620
n=96/620
n=38/620
T834P
n=376/620
n=33/620
n=102/620
Kruskal-Wallis test
P < 0.0001
B
E1103G
n=395/600
n=184/600
n=35/600
F1084I
n=334/600
n=43/600
n=58/600
Kruskal-Wallis test
P < 0.0001
C
H1085Y
n=447/600
n=85/600
n=18/600
H1085L
n=315/600
n=32/600
n=56/600
Kruskal-Wallis test
P < 0.0001
GOF
LOF
No obvious phenotype
Lethal
TL position
Postive interactions
Negative interactions
Synthetic lethal
Epistasis
Sign epistasis

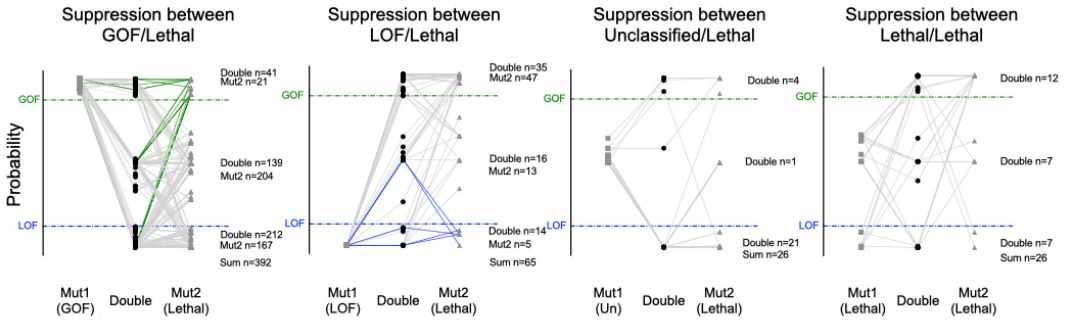
Figure S9. Interaction networks of selected probe mutants. The TL is shown in circle with WT residues and positions labeled. All 20 substitutions of each TL residue are represented by a magenta arc under each WT residue, with tick marks representing individual substitutions at that position and are colored by mutant class. Comparison of interaction networks between S713P and T834P (A), E1103G and F1084I (B) and H1085Y and H1085L (C) showed the differences are significant ($P < 0.0001$). The comparisons were performed with Kruskal-Wallis test with P-value correction with Dunn's multiple comparisons test (Supplemental Table 6).

Figure S10

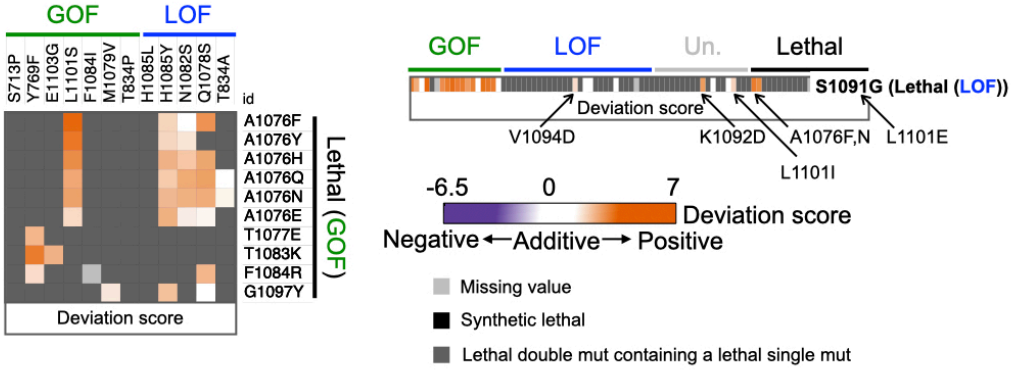
A



B



C



D

Ratio of observed strong or weak epistasis

	Strong epistasis (|d| > 2)	Weak epistasis (1< |d| < 2)
Pairwise Intra-TL doubles	0.15	0.06
Target Intra-TL doubles	0.15	0.09
Target Pol II-TL doubles	0.17	0.10
Reported	0.05	0.3

Figure S10. Discrimination of regular epistasis from sign epistasis. A. The phenotypic classes of double mutants consist of two viable single substitutions with positive interactions. Four plots show four kinds of combinations respectively. For each plot, the predicted GOF or LOF probabilities of a double mutant and two constituent single mutants are shown in Y-axis. The double mutant and two constituent single mutants are shown in X-axis in the order of the first constituent mutant (Mut1), the double mutant (Double), the second constituent mutant (Mut2). The double mutant and two constituent mutants are connected with lines for each pair of combinations. The numbers of double mutants belonging to GOF (top), unclassified (middle), or LOF (bottom) are labeled. GOF and LOF probability threshold are labeled with dashed lines. **B.** The phenotypic classes of double mutants consist of one viable and one lethal single substitutions, or two lethal single substitutions with positive interactions. The arrangements of plots are similar to A. **C.** The heatmaps of lethal GOF substitutions suppressed by GOF targets (left) and lethal LOF substitutions suppressed by LOF targets (right). **D.** The fraction of strong and weak interactions we observed in double mutants compared with the ratio reported in other studies (*54, 74-76, 83*).

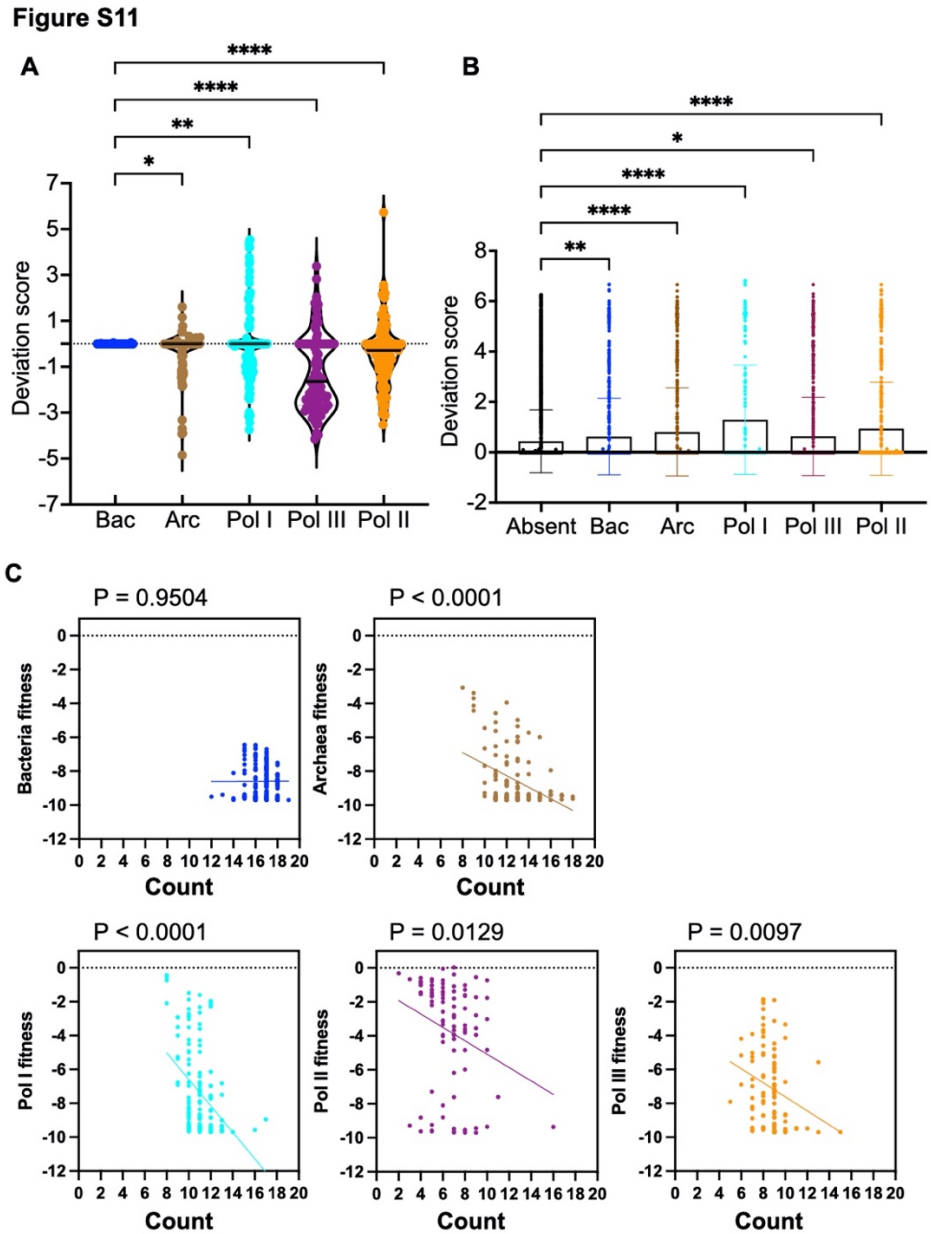


Figure S11. Contextual epistasis affects fitness of TL haplotypes. A. Distributions of deviation scores of the TL haplotypes in each group. **B.** Comparison of the mean deviation scores of lethal single substitutions that are present in different species and those that are absent in any species. Standard deviation values are also shown in the bar plot. ANOVA multiple comparison was applied to compare the mean deviation score of the "Absent" group to each of the other groups and significant levels ($P < 0.05$) are shown in the figure. **C.** An xy-plot of evolutionary observed TL haplotypes fitness versus the numbers of substitutions in the haplotypes. Simple linear regression was performed for each plot. Bacteria fitness vs count: $Y = 0.004267*X – 8.660$, $r^2=2.152e-005$. Archaea fitness vs count: $Y = -0.3406*X - 4.175$, $r^2=0.1568$. Pol I fitness vs count: $Y = -0.7818*X + 1.235$, $r^2=0.1521$. Pol II fitness vs count: $Y = -0.3943*X – 1.132$, $r^2=0.06535$. Pol III fitness vs count: $Y = -0.4148*X – 3.468$, $r^2=0.06984$. The P values of the slopes are labeled.

Figure S12

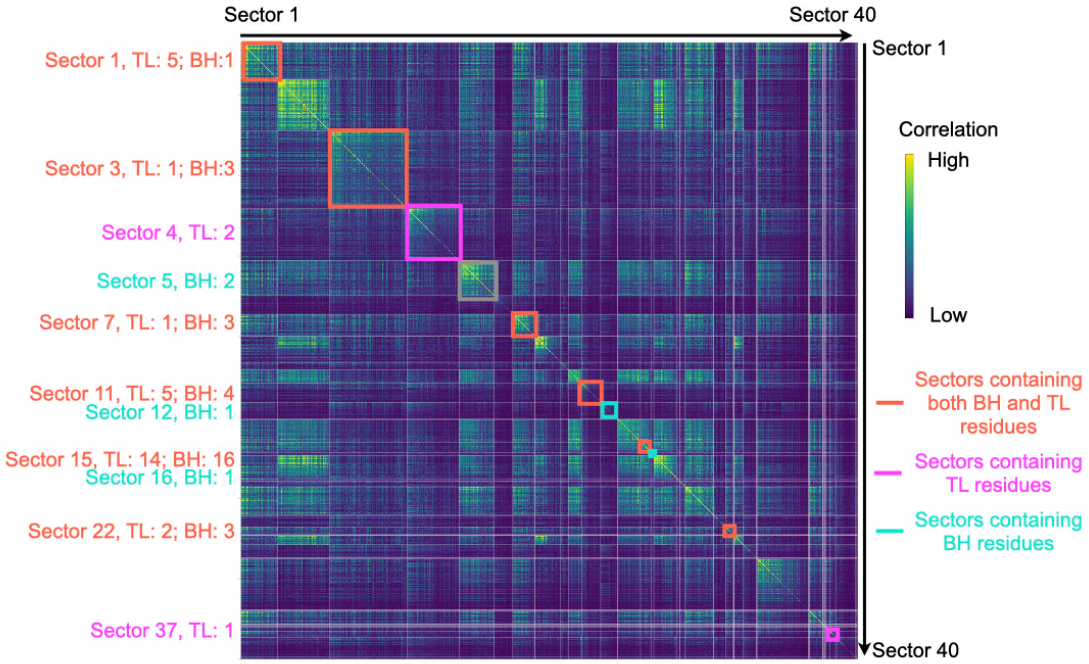


Figure S12. **Rpb1 coevolutionary residue networks identified by Statistical Coupling Analysis (SCA).** 40 significant and independent sectors are shown in a heatmap with correlation score calculated from the statistical coupling analysis. Sectors containing TL and BH residues are labeled. Numbers of TL and BH residues contained in each sector are labeled on the left of the heatmap. Statistical coupling analysis was applied to a published Multiple Sequence Alignment (MSA) of Rpb1 homologs (n= 410) (*82*). Details are in **Methods**.

Figure S13

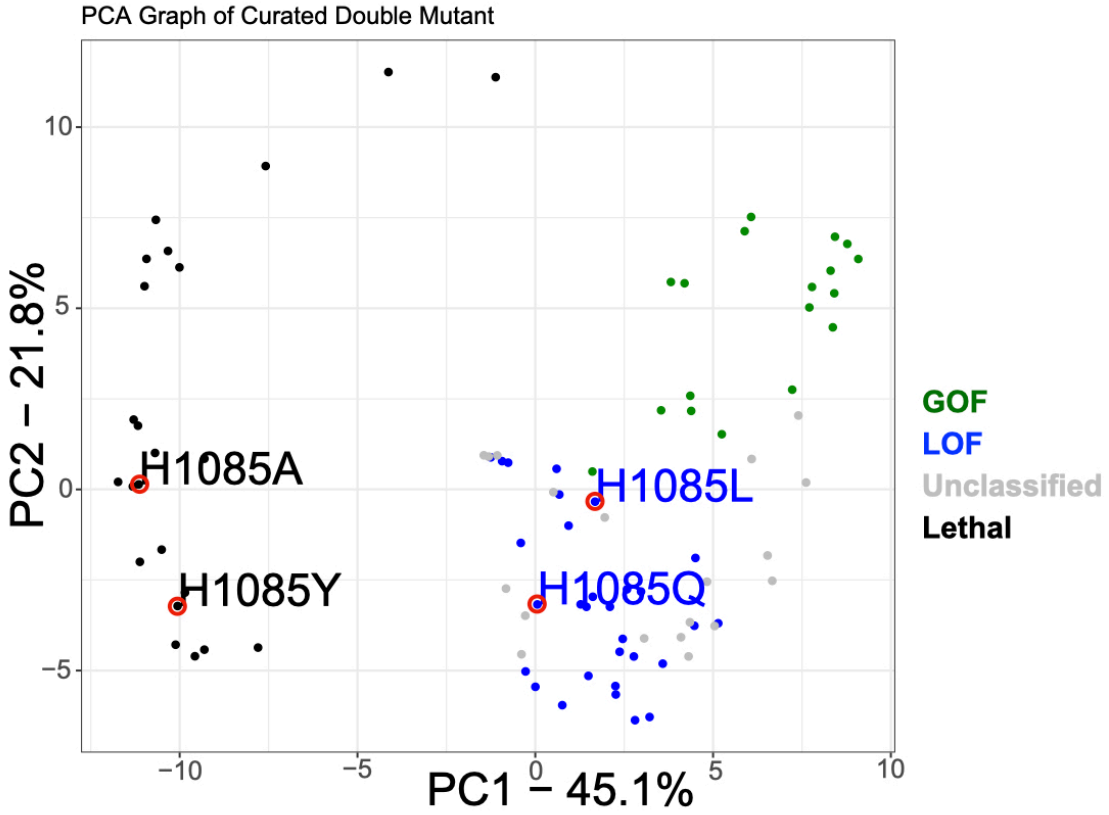


Figure S13. **Four H1085 substitutions are different in some ways. A.** Principal component analysis (PCA) with double mutant deviation scores of all curated TL single mutant substitutions, which are represented with colored dots. GOF mutants are in green, LOF mutants are in blue, unclassified mutants are in grey and lethal mutants are in black. Four H1085 substitutions are labeled and assigned with a red circle to make them visible in the plot.