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, USA

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

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

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

Contact information: craig.kaplan@pitt.edu

ABSTRACT

Multi-subunit RNA Polymerases (msRNAPs) are responsible for transcription in all kingdoms of life. These enzymes rely on dynamic, highly conserved active site domains such as the so-called “trigger loop” (TL) to accomplish steps in the transcription cycle. Mutations in the RNA polymerase II (Pol II) TL confer a spectrum of biochemical and genetic phenotypes that suggest two main classes, which decrease or increase catalysis or other nucleotide addition cycle (NAC) events. The Pol II active site relies on networks of residue interactions to function, and mutations likely perturb these networks in ways that may alter mechanisms. We have undertaken a structural genetics approach to reveal residue interactions within and surrounding the Pol II TL – determining its “interaction landscape” – by deep mutational scanning in *Saccharomyces cerevisiae* Pol II. This analysis reveals connections between TL residues and surrounding domains, demonstrating that TL function is tightly coupled to its specific enzyme context.

INTRODUCTION

Transcription from cellular genomes is performed by conserved multi-subunit RNA polymerases (msRNAPs)[1-3]. While bacteria and archaea use a single msRNAP to transcribe all RNAs[4-6], eukaryotes have at least three msRNAPs (Pol I, II, and III) for different RNA types[7-10]. RNA synthesis by msRNAPs occurs by iterative nucleotide addition cycles (NAC) of nucleotide selection, catalysis and polymerase translocation[11-15]. The msRNAPs active sites accomplish these NAC steps using two conformationally flexible domains, the bridge helix (BH) and 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.

Most catalytic cycle events involve concerted conformational changes in the TL, and potentially the BH[12-14,19-23]. The extent to which these conformations are supported by, or communicate with, the surrounding enzymatic context can be examined by structural or computational approaches, but functional importance needs to be determined experimentally. The BH is straight in most msRNAP structures[7,8,12,13,24], but appears kinked in *Thermus thermophilus* (bacteria) RNAP[5,25]. Simulations suggest these dynamics may promote msRNAP translocation[16,25-28]. The TL adopts various conformations to support its multiple functions. During each NAC, the TL nucleotide interaction region (NIR) discriminates correct NTPs from mismatched ones, initiating a shift from a catalytic-disfavoring "open" state to a catalytic-favoring "closed" state[11,20,29,30]. This closure promotes catalysis[20,31], and subsequent TL reopening is proposed to support polymerase translocation to the next template DNA position, allowing for the next NAC[32-34](**Fig. 1a**). Disrupting the TL with mutations affects transcription at all stages[13,17,23,35-39]. For instance, NIR mutations impair TL-substrate interactions, resulting in reduced catalysis and elongation rate in vitro (Loss of function, LOF)[17,23,36,37,40,41]. Mutations in the TL hinge and C-terminal regions appear to shift the TL toward the active state, enhancing catalysis and elongation rate but impairing transcription fidelity (Gain of function, GOF)[17,42,43]. Mutations may also affect other transcription processes such as proofreading, pausing and termination[12,43-46], as TL confirmations are associated with these activities in addition to catalysis and translocation.

All TL conformational dynamics and functions are likely balanced by residue interactions within and around the TL[41,45,47-49]. Embedded in the conserved active site, the TL interacts with surrounding domains such as the BH and α46-47 helices, forming a five-helix bundle enclosing a hydrophobic pocket that supports these dynamics **(Fig. 1a)**[12,42]. Mutations in TL adjacent domains (e.g., BH GOF T834P and LOF T834A, funnel helix α-21 GOF S713P, and Rpb2 GOF Y769F)[17,23,25,26,35-37] **(Fig. 1a)**, and in a Pol II subunit which does not directly contact Pol II active site, such as Rpb9[50], also alter catalytic activity and fidelity, showing phenotypes seemingly identical to TL mutants, indicating that TL dynamics and function are finely balanced and could be sensitive to allosteric effects[50-53]. Determining the extent and nature of the pathways connecting the TL to the rest of the polymerase will reveal the functional networks that shape its dynamics and the potential pathways through which Pol II has evolved or may be controlled by elongation factors.

Physical and functional intramolecular interactions between amino acids define protein function and evolvability[54-57]. Genetic interactions where one mutation modulates or alters the phenotype of another are commonly referred to as epistasis. Interactions can be of different types such as suppressive, masking, or synthetic interactions. In this study, we use "epistasis" specifically to

describe cases where mutant phenotypes allele-specifically depend 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[58-60]. Recent studies have shown how mutations can alter protein function, allostery, and evolvability, suggesting that even conserved residues are subject to unique, context dependent epistatic constraints[61-68]. Consistent with this, distinct phenotypes for the same conserved residues have been observed in various proteins, including yeast Pol I and Pol II[66,67,69]. For example, the TL domain of yeast Pol I is incompatible with Pol II, despite sharing 70% of sequence identity, while the similarly distinct Pol III TL maintains some function[70]. It is not known generally if highly conserved TL sequences are compatible in a heterologous msRNAP context. Answering this question would reveal how tightly integrated the TL is by coevolution with its embedded context, or the extent to which its functions are self-contained due to its very high conservation.

Functional interactions between residues can be revealed by genetic interactions of double mutants[17,37,71-73]. A systematic analysis is needed to fully describe and understand the requirements for each mutant's phenotype, and the nature of the interactions that control Pol II activity. Previous studies from our lab on a small subset of site-directed substitutions identified distinct types of Pol II double mutant interactions, including suppression, synthetic sickness and lethality (activity-additive), less-than-additive interactions (sometimes referred to as epistasis), and sign-epistasis[17,37,72]. Suppression interactions are common between LOF and GOF mutants, as expected if each mutant is individually acting in the double mutant, balancing opposing effects on activity. Similarly, synthetic lethality or sickness were common between mutants of the same class (GOF or LOF), consistent with greater defects arising from combining partial losses of TL function. However as just noted, we also observed mutant combinations where the double mutant was not sicker than the single mutants for mutants of the same class, suggesting these mutants might act at the same step. Additionally, we had observed a potential case of sign-epistasis, where a mutant's phenotype class appeared dependent on the identity of another residue. For example, the GOF TL substitution Rpb1 F1084I was unexpectedly lethal with the LOF TL substitution Rpb1 H1085Y, contrary to the predicted mutual suppression for a GOF/LOF combination[37,72]. This suggested that F1084I required H1085 for its GOF characteristics and became a LOF mutant in the presence of H1085Y (**Fig. 1b**).

Deducing large-scale residue interaction networks is challenging. To accomplish this for Pol II, we developed experimental and analytical schemes to extend the previously established "Pol II phenotypic landscape"[17,37,72,74] to 15,174 double and multiple mutants within the *S. cerevisiae* Pol II TL and between the TL and adjacent domains (a "Pol II Interaction landscape"). Our analyses indicate that TL function and evolution are dominated by widespread epistasis. The TL faces extensive constraints despite its high conservation, arising from residues within the TL and its interacting domains, resulting in highly similar TLs being incompatible inside yeast Pol II to a large degree. Additionally, individual mutants with similar biochemical and genetic phenotypes have quite distinct genetic interaction networks, revealing differential functional requirements for their abilities to confer altered Pol II activities. Some mutants minimally change the genetic behavior of most TL substitutions when combined, suggesting that TL conformations or interactions remain largely unchanged, with altered TL dynamics hypothesized to be the sole difference. In contrast, other alleles greatly perturb the genetic behavior of many TL substitutions, indicating these substitutions change the active site landscape in ways that would otherwise be undetectable. Together, these results suggest small changes in the Pol II active site can dramatically change its constraints and therefore, its potential evolutionary paths.

Furthermore, our findings highlight potential communication networks through evolutionary coupling of Pol II residues within the active site, which may underlie allosteric communication paths to the active site for elongation factors.

RESULTS

A system for dissection of the Pol II active site genetic interaction landscape. We developed an experimental and analytical framework, termed 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 15,174 variants representing evolutionary haplotypes, all possible Pol II TL single mutants and a subset of targeted double mutants in ten libraries (Supplemental Table 1). This approach follows our prior analysis of the TL phenotypic landscape[17] with modifications (**Methods** and **Extended Data Fig. 1a**). Libraries were transformed, screened under selected conditions (**Methods**), and sequenced (**Fig. 1c, Extended Data Fig. 1a**). Mutant growth phenotypes are calculated based on relative allele frequency shifts compared to a control condition and normalized to the WT under identical conditions. Biological replicates indicate high reproducibility (**Extended Data Fig. 1b-c**). Individual libraries were min-max normalized[75] to account for scaling differences between libraries (**Fig. S1a**) and the same mutants present among different libraries indicate high correlation of fitness determinations in each library (**Fig. S1b-c**).

We defined a conceptual framework for evaluating genetic interactions among TL mutations. First, we assume that independent mutant effects combine in log-additive manner. This means that predicted double mutant fitness defects should be the combination of both single mutant defects, as is standardly assumed[60,71,73,76]. Deviations from log-additive expectations represent genetic interactions. To quantify any potential deviation, we compared the observed fitness of a double mutant measured in our high-throughput experiment to its expected fitness, calculated as the addition of constituent single substitutions' fitnesses, which are expressed as a log value. We term these differences as a "deviation score" (Deviation score = observed fitness – expected fitness; **Extended Data Fig. 2**). We consider a deviation score between -1 and 1 to be consistent with additive behavior, meaning independent contributions of single mutants to the double mutant fitness. A deviation score > 1 is taken as a positive genetic interaction (e.g., suppression), while a score < -1 indicates a negative genetic interaction (e.g., synthetic sickness or lethality).

Importantly, our phenotyping analyses determine both general mutant fitness and predict their class of in vitro biochemical defect. Our previous studies showed that phenotypic profiles from conditional growth assays, reflecting Pol II-activity dependent phenotypes in vivo, correlate with the measured biochemical activity in vitro[17,37,77]. For example, catalytically hyperactive GOF mutants have a phenotypic profile based on sensitivity to mycophenolic acid (MPA) and weak suppression of two genetic reporters (*gal10Δ56* and *lys2-128∂*), while catalytically defective LOF mutants are MPA resistant, show strong suppression of *gal10Δ56* and fail to suppress *lys2-128∂* (more fully described in Methods). Using the established correlation between growth profiles (MPA and media used to detect phenotypes of mutants on *gal10Δ56* and *lys2-128∂* backgrounds) and known/predicted in vitro catalytic effects (GOF or LOF) of 65 TL alleles[17,37,53,78], we trained two multiple logistic regression models to distinguish Pol II mutant classes, and applied them to all viable mutants (growth fitness score > -6.5) of designed variants, classifying them into three groups, GOF (22.96%), LOF (28.11%) and unclassified (mutants not belonging to either class, 48.93%) (**Extended Data Fig. 3a,** Methods). To visually

179 inspect the classifications, we applied t-SNE projection and k-means clustering for all measured
180 mutants across all growth conditions. We observed separated GOF and LOF clusters consistent
181 with logistic regression classifications (**Extended Data Fig. 3b-c**). Notably, mutants with similar
182 defects may be further distinguished in different clusters, suggesting their similar phenotypic
183 defects may be related to different transcription defects (**Extended Data Fig. 3c**). By classifying
184 Pol II active site mutants as "GOF" or "LOF" based on predicted in vitro biochemical
185 phenotypes, we establish baseline expectations for their genetic interactions. We previously
186 observed activity-additive interactions, meaning suppression between mutants of different
187 classes (GOF+LOF) or synthetic sickness/lethality within the same class (GOF+GOF or
188 LOF+LOF). Together with the calculated deviation scores, we can distinguish specific epistatic
189 interactions from activity-dependent effects. We classified an interaction as epistasis when we
190 observed positive fitness deviation in mutants of the same activity class (GOF+GOF,
191 LOF+LOF), where we would otherwise expect synthetic sickness or lethality if mutants were
192 functioning independently. Conversely, we defined sign epistasis for situations where negative
193 deviations in combinations between the classes (GOF+LOF), where suppression would be
194 expected for independent mutants (**Fig. 1d, Extended Data Fig. 2**).

195 We compared interactions observed in sequencing to growth assays for 59 previously examined
196 double mutants[17,37], with 48 matching our observations. The other 11 were categorized as ultra-
197 sick/lethal and therefore inaccessible to sequencing analysis (**Supplemental Table 2**). We
198 constructed 50 additional mutants to assess interactions by patch assay and spot assay, and
199 these aligned with our high-throughput results (**Extended Data Fig. 4**).

200 **Widespread incompatibility of conserved TL sequences with yeast Pol II suggests the TL**
201 **is highly constrained by its coevolved environment**. We previously found that identical
202 mutations in a residue conserved between the Pol I and Pol II TLs yielded different biochemical
203 phenotypes[70,79]. Additionally, the yeast Pol I TL was incompatible within the yeast Pol II enzyme,
204 implying that TL function in that case depended on its enzymatic context[70,79]. To determine the
205 scope of TL-Pol II incompatibility, we designed a library containing evolutionary TL variants from
206 bacterial, archaeal, and eukaryotic msRNAPs and determined their compatibility in the yeast Pol
207 II context (**Fig. 2a**). Eukaryotic TL alleles were more compatible than those from archaea or
208 bacteria, with Pol II alleles being the most compatible (**Fig. 2b, Extended Data Fig. 5a-b**),
209 consistent with evolutionary distance. Growth fitness in the Pol II background showed a slight
210 negative correlation with the number of TL substitutions for most TLs but not those from bacteria
211 (**Extended Data Fig. 5c**), likely because of the bacterial TLs being largely incompatible with Pol
212 II. Despite this, some archaeal TLs provide yeast Pol II viability, yet several Pol II TLs were
213 defective or lethal. These results suggest widespread coevolution between TL sequences and
214 that the enzymatic background, beyond ultra-conserved positions, shapes TL function in
215 individual enzymes. We have explored the nature of this incompatibility and complex intra-TL
216 genetic interactions further in a separate study[80].

217 **Allele-specific interactions suggest unique properties of individual mutants with similar**
218 **phenotypes**. To determine the potential residue interactions surrounding the TL in the yeast Pol
219 II environment, we analyzed 12 previously studied GOF and LOF mutants (eight within the TL
220 and four outside), each combined with >600 perturbations in the form of all single substitution
221 TL mutants (**Fig. 3a**). These 12 mutants serve as probes for the genetic interaction space of the
222 TL, revealing how it might be altered in allele-specific fashion by each "probe" mutation.
223 Importantly, TL conformational dynamics and function are balanced by internal residue

interactions within the TL and by TL-external interactions between the TL and TL-proximal domains. While GOF and LOF mutants in TL-adjacent domains appear similar to GOF and LOF TL mutants, their underlying mechanisms or specific residue dependencies are not known. We therefore compared interaction networks among probe mutants with similar biochemical and phenotypic defects to distinguish if changes to TL function might reflect simple alterations to TL dynamics, or additional alterations to folding trajectories or conformations.

We observed distinct interaction networks among probe mutants, even for those with similar apparent biochemical and growth defects (**Fig. 3b-d**), indicating that mutants differentially impact the Pol II active site, potentially reflective of different constraints and distinct properties of individual mutants, as discussed below. A simple expectation is that mutants with altered TL dynamics yet maintenance of conformational states should show additive interactions, because the effects of substitutions on function would be maintained. In contrast, substitutions altering TL conformations/folding or propagating changes across the active site, may be subject to allele-specific interactions *i.e.* prominent epistasis. We observed that most probe mutants were placed on continuum of effects based on the balance between additive (independent) interactions and epistatic (dependent or allele-specific) interactions in double mutants, which we interpret as the degree to which they propagate their effects across the active site (**Fig. 3b-c, Extended Data Fig. 6, and Fig. S2**). Probe mutants appeared separable based on the extents of their additive or epistatic interactions, where greater numbers of independent interactions (additive) would indicate fewer effects on other residues' functions. In contrast, an increase in dependent interactions (sign epistasis) would indicate greater changes in other mutant's characters and presumably residue function. For example, Y769F, a GOF TL-proximal mutant in Rpb2 link domain (**Fig. 1a**), uniquely suppresses some TL lethal substitutions among all GOF probe mutants (**Extended Data Fig. 7**), indicating specific functional interactions between Y769F and these substitutions. Moreover, two TL-adjacent GOF mutants, Rpb1 S713P (funnel α-helix 21) and T834P (BH) displayed largely distinct interaction networks despite similarly increased biochemical activity and genetic phenotypes as single mutants[17] **(Fig 3c, Extended Data Fig. 8a)**. S713P broadly suppressed LOF TL substitutions (96 instances), consistent with generic enhancement and preservation of TL dynamics or residue function. In contrast, T834P exhibited sparser suppression (33 instances) but much greater sign epistasis than S713P (102 versus 38 instances), especially with P1099, where all substitutions showed sign epistasis with T834P but not S713P (**Extended Data Fig. 8a**). These results suggest that T834P's GOF function requires most TL residues to be WT, is especially reliant on proline at 1099, and that T834P alteration to the BH extensively changes the TL's functional space, likely through folding or altered conformations.

A similar distinction was observed between two internal TL GOF mutants, Rpb1 E1103G and F1084I **(Fig. 3c, Extended Data Fig. 8b)**. E1103G showed widespread suppression of LOF TL substitutions (184 instances), consistent with its proposed role in altering TL dynamics to promote TL closure[23,37] and which allows TL mutants to maintain their effects. In contrast, F1084I showed more limited suppression (43 instances) while showing widespread synthetic lethality due to predicted sign epistasis, indicating that F1084I requires many WT residues for GOF characteristics and itself may switch to a LOF upon many additional TL perturbations.

The most striking example came from two LOF substitutions at the ultra-conserved Rpb1 H1085 **(Extended Data Fig. 8c)**. This histidine, which contacts incoming NTP substrates[5,12], is the target for the Pol II inhibitor α-amanitin[36,81], and promotes catalysis[12,13]. Initial structural data and

molecular dynamics simulations suggested that H1085 functions as a general acid for Pol II catalysis[82-85]. Our discovery that H1085L was well-tolerated[17], and subsequent experiments from the Landick lab[86,87], have led to their proposal that the TL histidine functions as a positional catalyst where a similarly sized leucine substitution supports catalysis with mild effects on biochemistry and growth. If H1085Y and L substitutions represent a continuum of positional catalyst activity, then their interaction networks would be expected to be similar, and differing only in magnitude, not identity or type of interactions. Contrary to this, distinct interaction patterns were observed (**Fig. 3c, Extended Data Fig. 8c**). Most GOF mutants suppressed 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. S2b, Extended Data Fig. 8c**). The distinction between H1085L and H1085Y is further highlighted in PCA analysis (**Fig. 3d**).

Functional and epistatic interactions within the TL hydrophobic pocket tune Pol II activity. Several allele-specific epistatic interactions were observed, with some of the most extensive occurring between A1076 substitutions and L1101S (**Fig. 4a**), suggesting tight coupling between A1076 and L1101 for Pol II function. These two hydrophobic residues, together with other hydrophobic residues in TL helices, likely stabilize the open TL conformation through hydrophobic packing of the TL helices (**Fig. 4c**). Consistent with this, another pair of adjacent residues, M1079 and G1097, also showed allele-specific interactions (**Fig. 4b-c**). Additionally, three GOF/GOF combinations between L1101S and A1076 substitutions showed WT-like behavior for the double mutants based on our regression model for mutants, indicating mutual suppression due to one of the singles switching to a different class (sign epistasis), further supporting a tight functional interaction between 1101 and 1076 (**Fig. S3a**).

We also observed allele-specific interactions for predicted lethal mutants. Our lethality threshold is more conservative than the actual lethal threshold and therefore, some classified lethal but actually slow-growing mutants have enough growth in phenotypic assays for GOF or LOF class determination. Among 21 ultra-sick/lethal TL substitutions predicted as GOF, suppression by LOF mutants was expected and observed (**Fig. S3b**). For example, lethal A1076 substitutions were suppressed by LOF probe mutants, suggesting their severe defects result from strong GOF status. However, some A1076 substitutions were also suppressed by the GOF probe mutant L1101S, indicating allele-specific mutual suppression or sign-epistasis, consistent with A1076-L1101 interactions described above (**Fig. S3c**). Similarly, ultra-sick/lethal LOF substitutions were commonly suppressed by GOF mutants, but some were suppressed by specific LOF mutants (**Fig. S3b**). For example, S1091G was suppressed by most GOF mutants and specifically by the LOF V1094D (**Fig. S3c**), suggesting allele-specific compensation. Furthermore, some lethal mutants exhibited interactions with specific GOF or LOF probe mutants. For example, F1084R, was specifically suppressed by LOF probe Q1078S and GOF probe Y769F (**Fig. S3c**). Notably, F1084, Q1078S and Y769 are in close proximity in the substrate bound, catalytic-favoring TL conformation. These types of allele-specific interactions potentially 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 quantified proteins[57,88-91] (**Fig. S3d**), likely due to the high rate of suppressive interactions stemming from Pol II mutants' opposing effects on catalysis.

Mapping TL-internal residue interaction networks suggests widespread epistasis within the TL. To determine TL-internal interaction networks, we selected 2-4 substitutions for each TL residue, representing diverse phenotypes (GOF, LOF, lethal, or unclassified), and combined them with the selected substitutions across all other TL positions. This curated set of 3790 double mutants captures potential interactions between any two TL residue positions (**Fig. 5a**). Observed fitness of these double mutants was compared to predictions from the additive model, revealing widespread deviation from the model (i.e. large interactions)(r^2=0.21), much smaller than the r^2 (about 0.65-0.75) reported in other studies[57,73,88,89,91] (**Fig. 5b**), likely reflecting our selection criteria for mutations with diverse phenotypes. In general, mutations of the same class were additive or exacerbating, with some exceptions that reveal functional dependencies between residues (**Fig. 5c, 5e, Extended Data Fig. 9**). Additionally, the strongest deviations from expectations were between mutations of opposing phenotypic classes (GOF+LOF combinations). Here, commonly we observed suppression as predicted for mutations acting independently in 43% of GOF+LOF combinations but we also observed unexpected lethality or exacerbation (sign epistasis) for 41% of GOF+LOF combinations. Distinct patterns of sign epistasis for GOF/LOF combinations provide the ability to distinguish functional requirements for GOF alleles to be able to confer GOF phenotypes (discussed below). Interactions were distributed throughout the TL, covering every residue, supporting connectivity across the TL (**Fig. 5d**). Notably, epistasis clustered within the C-terminal TL helix and adjacent regions (The far-right panel of **Fig. 5d**), supporting functional dependencies of TL-C terminal residues on one another, consistent with their proposed function in collaboratively stabilizing the open TL.

Genetic interactions reveal insights into lethal or unclassified individual mutants. Most lethal mutants could be suppressed by at least one predicted GOF mutant (**Fig. 5f, Extended Data Fig. 9**), suggesting their lethality likely comes from reduced activity (LOF) below a viable threshold. However, this suppressibility also suggests that individual lethal substitution activity defects are mostly close to the viable threshold. This suggests that the contributions of most individual residues represent only a portion of total TL activity, which is expected to provide a ~1000 fold increase in catalytic activity[92-94]. Suppressibility of lethal residues by GOF alleles is also consistent with the greater probability of lethal substitutions being LOF rather than GOF. However, two lethal mutations at A1076 were suppressed by most LOF mutations or specific lethal mutants, but not GOF mutants, implying that their lethality resulted from being GOF. Additionally, unclassified single mutants generally showed limited interactions with GOF, LOF, or lethal classes, consistent with small or mostly independent effects. However, a few unclassified mutants showed suppression in combination with GOF mutants, suggesting they are either potential atypical LOF not detected by phenotypic analysis, or allele-specific interactions (**Fig. 5g, Extended Data Fig. 9**).

We reasoned that evolutionarily observed lethal substitutions would be closer to viability than those not observed and would therefore be more likely to be suppressible by Pol II GOF mutants. To determine this, we compared the suppressibility of substitutions observed in existing msRNAPs that are lethal in *S. cerevisiae* Pol II versus substitutions that are lethal but not present in extant msRNAPs by analyzing the maximum positive deviation scores from all double mutants containing individually lethal substitutions. Singly-lethal substitutions (single substitutions that are lethal on their own in yeast Pol II) identified in our msRNAP TL MSA showed higher maximum deviation scores than substitutions lethal on their own in yeast and not present in the MSA (**Fig. 6a, Extended Data Fig. 5b**). Consistently, most L1081 substitutions cannot be suppressed, whereas a L1081M, (commonly a WT residue in bacteria and archaeal

RNAPs) showed the highest suppressibility (**Fig. 6b**). These results indicate that singly lethal mutants in *S. cerevisiae* Pol II observed elsewhere msRNAP evolution, on average, maintain greater functionality and/or are more easily suppressed by single additional mutants in yeast Pol II. Interestingly, most lethal substitutions in A1087 and all lethal substitutions in G1088 cannot be suppressed (**Fig. 6b**), suggesting the strict restriction at these positions, consistent with the structural observation that A1087 and G1088 are in a small pocket between the bridge and funnel helices[12,17]. The TL has been estimated to enhance catalytic activity by 500-1000 fold[92,93,95], while yeast viability tolerates only ~10-fold effects[36]. We conclude that lethal mutants observed as functional residues in other species are likely closer to the viability threshold, as might be predicted to result from a series of small adaptive steps that allow their function.

Co-evolution of Pol II active site residues form distinct sectors within specific domains and extend to the surfaces of Rpb1 and Rpb2. Our analyses suggest that even a highly conserved domain like the Pol II TL can be tightly integrated through coevolution of less conserved residues and that substitution tolerance is shaped by widespread epistasis. Changes in residue identities alter interaction networks, shaping the Pol II active site across evolution. To identify the coevolving residue networks in Pol II active site we employed statistical coupling analysis (SCA) for residues within Rpb1 and Rpb2, the two largest subunits of Pol II, which include the entire active site. SCA is well suited for identifying coupled residue sectors, representing putative residue communication networks[96-98]. From 283 linked Rpb1/Rpb2 sequences, we identified 53 coevolving sectors (**Fig. S4**). Residues within the active site form four major sectors (**Fig. 7a-e**). First, sector 37 includes many Rpb2 link domain residues and most TL residues, including those at the NIR and some at the TL C-terminal (**Fig. 7b**). This sector tightly surrounds the substrate binding pocket (**Extended Data Fig. 10a**), befitting critical roles in catalysis. Second, sector 15 consists of four BH residues, most residues in switch 1 (part of the clamp domain), and many cleft residues, including the α46-47 loop, which accommodates incoming DNA[8,13] (**Fig. 7c**). Residues in this sector directly interact with Pol II subunits Rpb5 and Rpb6 (**Extended Data Fig. 10b**), consistent with Rpb5's role in interacting with the incoming DNA[99]. While the α46-47 loop forms a five-helix bundle with two TL helices and the BH helix[42] (**Fig. 4c**), coupling analysis indicates it is primarily evolutionarily coupled with cleft residues instead of the TL. This suggests that the α46-47 loop may influence TL function by constraining the space for TL movement, rather than directly interacting through specific residue identities. Third, sector 1 involves active site residues surrounding the Mg^{2+} at the catalytic site and most F-loop residues (**Fig. 7d**). This sector forms a potential communication path that wraps around the TL and BH, connects to the BH through two BH residues, and extends to Rpb9, itself a subunit that forms important inputs to the TL and is required for transcription fidelity[50,100,101]. This pathway likely mediates communication between Rpb9 and the Pol II active site (**Extended Data Fig. 10c**). Fourth, sector 2 contains most BH residues are coupled with the Rpb2 Fork and hybrid binding domain, forming a network linking the BH to Rpb2 residues (**Fig. 7e**). This network connects to Rpb5, which interacts with the downstream DNA, Rpb6, a shared subunit in three eukaryotic msRNAPs, Rpb9, and Rpb3 and Rpb10, which function in Pol II assembly[102] (**Extended Data Fig. 10d**), representing a residue communication path potentially tied to various transcription activities.

Notably, coupling is not restricted to residues that are within or near the active site. Some sectors extend from the active site to the surface of Rpb1/Rpb2 (**Fig. S5a-f**), suggesting that active site function, potentially through the TL, could be modulated by allosteric interactions through the binding of transcription factors. Residues at interfaces with elongation factors can

405 be coupled with those factors yet some residues at interfaces can be coupled with residues
406 distal to the interface, suggesting potential pathways for allosteric interactions. Examples of
407 these are shown for TFIIS (**Fig. S6**), LEO1 (Leo1), ELOF1 (Elf1), and SUPT5H (Spt5) (**Fig. S7**).
408 Together, our findings highlight potential communication networks through evolutionary coupling
409 that may underlie regulation paths from Pol II surface to active site.

410 **DISCUSSION**

411 How individual mutants alter a protein's function is not necessarily straightforward at the
412 mechanistic level. Amino acid substitutions both remove functionality of the WT residue but
413 replace that functionality with something different. By altering the local environment within a
414 protein or potentially propagating effects to distant locations, each substitution potentially can be
415 quite different. These differences may not be apparent as phenotypic outputs, and phenotypic
416 assays may not have granularity to distinguish different biophysical behaviors if they result in
417 similar outputs. For Pol II mutants, even high-resolution phenotypic analyses, such as gene
418 expression profiling or genetic interaction profiling between Pol II mutants and deletions in other
419 yeast genes[53], suggest that LOF and GOF mutants represent a continuum of defects that match
420 enzymatic activity in vitro. Therefore, these profiles also appear dependent on the output of Pol
421 II activity defects and can't distinguish potential differences in underlying mechanism.

422 Through systematic detection of genetic interactions within the Pol II active site, we have
423 identified functional relationships between amino acids across the TL and between TL
424 substitutions and others. In the absence of double mutant epistasis analyses it would not be
425 possible to differentiate similar alleles from one another. L1101S and E1103G, for example, are
426 two GOF alleles very close to each other in Pol II structure and confer similar phenotypic
427 landscapes across various growth conditions. Here, we find that their distinct interactions
428 support that substitutions at 1101 and 1103 target distinct residue networks (**Fig. 3c, 4a,**
429 **Extended Data Fig. 6, Fig. S2**). L1101 functions in the TL hydrophobic pocket while E1103
430 interacts and co-evolves with a number of TL external residues that together support
431 interactions that maintain the open TL conformation. We also observed connections between TL
432 C-terminal residues that suggest a limit to how disruptions to structure there can alter Pol II
433 activity (**Fig. 5d**). Helix-disrupting LOF proline substitutions in at least two TL positions showed
434 epistasis with multiple substitutions in the back of the TL (1094-1098), suggesting that their
435 functions require TL C-terminal helix structure and in the absence of that structure (proline
436 disruption) effects are no longer additive.

437 The strongest epistatic interactions were observed between two pairs of hydrophobic residues,
438 A1076 and L1101, and M1079 and G1097 (**Fig. 4**), suggesting their interdependence in
439 maintaining TL open conformation stability[42]. Considering the TL's multiple active site roles, the
440 observed interaction networks may also contribute to other NAC steps or off pathway steps (*eg.*
441 pausing or backtracking). Structural studies have proposed that these hydrophobic residues
442 form a pocket stabilizing the TL by bundling two TL helices with the BH and α46-47 loop in a
443 five-helix bundle[28,42]. However, our coupling analysis suggests that α46-47 loop is evolutionarily
444 coupled with cleft residues rather than the TL, indicating the α46-47 loop may indirectly impact
445 TL function by limiting the space available for TL movement, rather than through direct
446 interactions with TL residues (**Fig. 7c**).

447 Elongation factors bind Pol II and alter its activity, but the mechanisms by which they do so are
448 not known[103,104]. 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[105]). These contacts have been proposed to alter Pol II activity. We would propose that the paths for such alteration to activity may follow the coupling sectors we have observed by SCA (**Fig. 7**).

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[82,83,85,87,106-108]). For msRNAPs, the association with the incoming NTP by a universally conserved histidine led to the proposal that this residue might donate a proton during the reaction [12,84,106]. 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[86,87]. 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) (**Extended Data Fig. 8c, Fig. S8**). 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[62]. 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. A limitation of our study is its reliance primarily on genetic and statistical coupling analysis. Future biochemical and biophysical studies will give mechanistic insight into the activity defects underlying the functional connections we observe, for example connecting kinetic properties or structural states to specific mutants. Nonetheless, 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 shared but also enzyme-specific mechanisms due to higher order epistasis shaping the functions of conserved residues.

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.

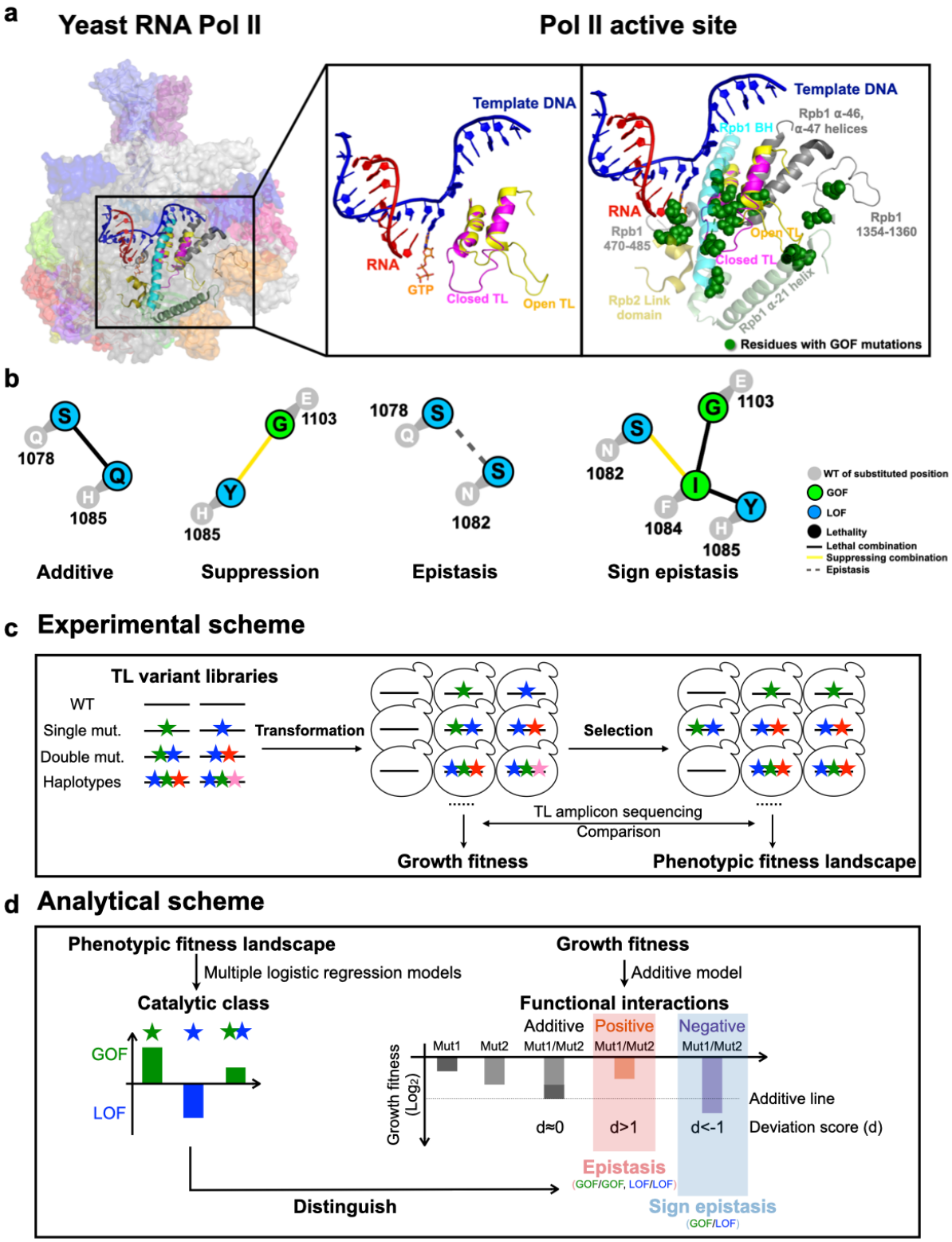
AUTHOR CONTRIBUTIONS

BD performed sequencing library construction, data analysis, and made figures, drafted and revised the manuscript. CQ designed the mutant libraries, performed screening experiments, and contributed to revising the manuscript. SHZ provided allele counts from raw sequencing data. CDK conceived the project, guided analyses and interpretation of data, provided funding, and revised the manuscript.

COMPETING INTERESTS

The authors declare no conflict of interest.

507 **FIGURES AND FIGURE LEGENDS**



508
509 **Figure 1. Schematics of the Pol II active site interaction landscape. a.** The Pol II active site
510 is embedded in the center of a 12-subunit complex (left panel). Pol II functions are supported by
511 distinct TL conformational states. An open TL (PDB: 5C4X)[42] and closed TL (PDB: 2E2H)[12]

512 conformations are shown in the middle panel. GOF mutations have been identified in the TL and
513 its proximal domains (right panel), suggesting TL mobility and function may be impacted by
514 adjacent residues. **b.** Examples of inter-residue genetic interactions. WT residues are shown in
515 grey circles with number indicating residue position in Rpb1. Mutant substitutions are shown in
516 colored circles, with color representing mutant class. Colored lines between mutant substitutions
517 represent types of genetic interactions. **c.** Overview of experimental approach. We synthesized
518 10 libraries of TL variants represented by colored stars. Libraries were transformed into WT or
519 mutated yeast strains. A selection assay was subsequently performed by scraping and replating
520 the transformants onto different media for phenotyping. DNA was extracted from yeast from all
521 conditions, and went through TL region amplification, and Illumina sequencing. Read counts for
522 variants on general conditions were used to determine growth fitness, while read counts on
523 other conditions were used to determine the phenotypic fitness landscape (see **Methods**). **d.**
524 Overview of analytical approach for determining interaction landscape. Mutant conditional
525 growth fitnesses were calculated using allele frequencies under selective growth conditions and
526 subjected to two logistic regression models for classification/prediction of catalytic defects.
527 Double mutant interactions were computed using growth fitness. Classification allowed epistatic
528 interactions to be deduced from double mutant growth fitness (see **Methods**).

529

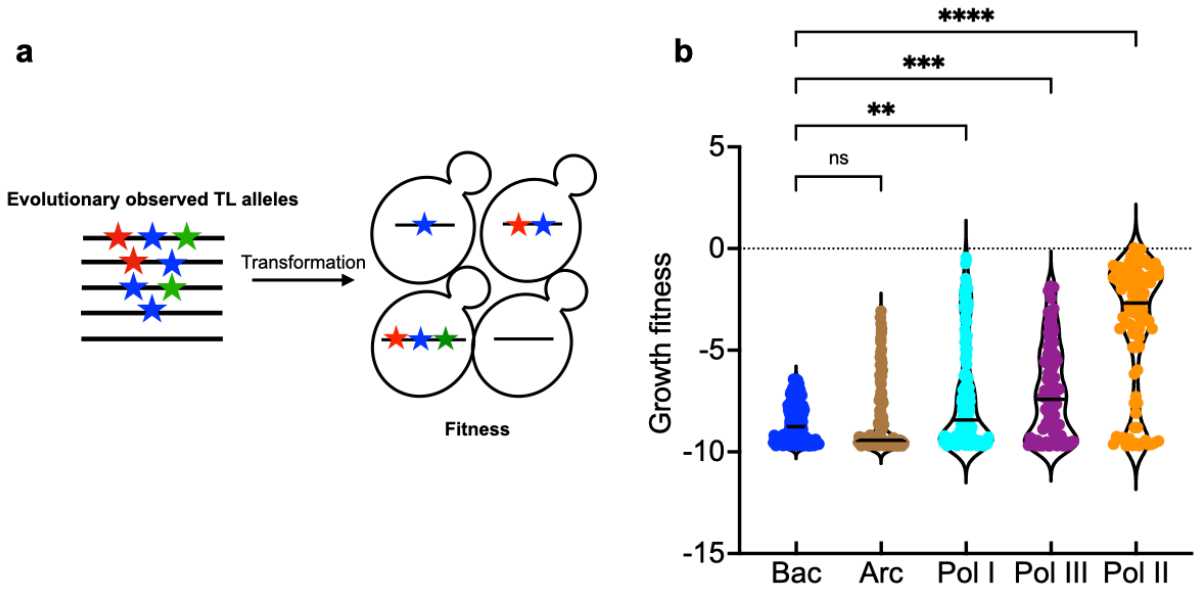
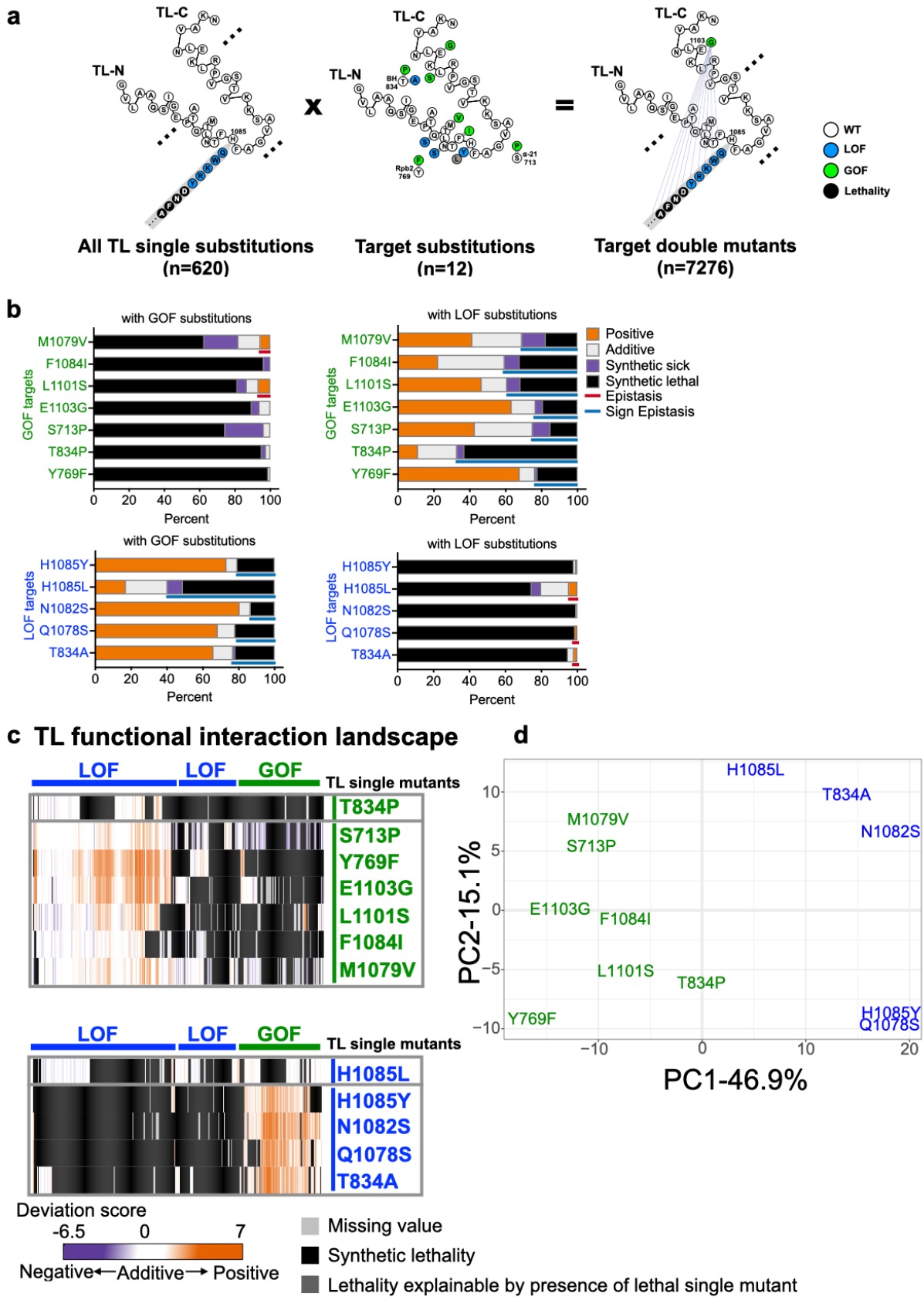


Figure 2. Contextual epistasis shapes TL evolution. a. Schematic for the TL evolutionary haplotypes library. We selected 662 TL haplotypes representing TL alleles from bacterial, archaeal and the three conserved eukaryotic msRNAPs. These TL alleles were transformed into yeast and were phenotyped under 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.



538
539 **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**).

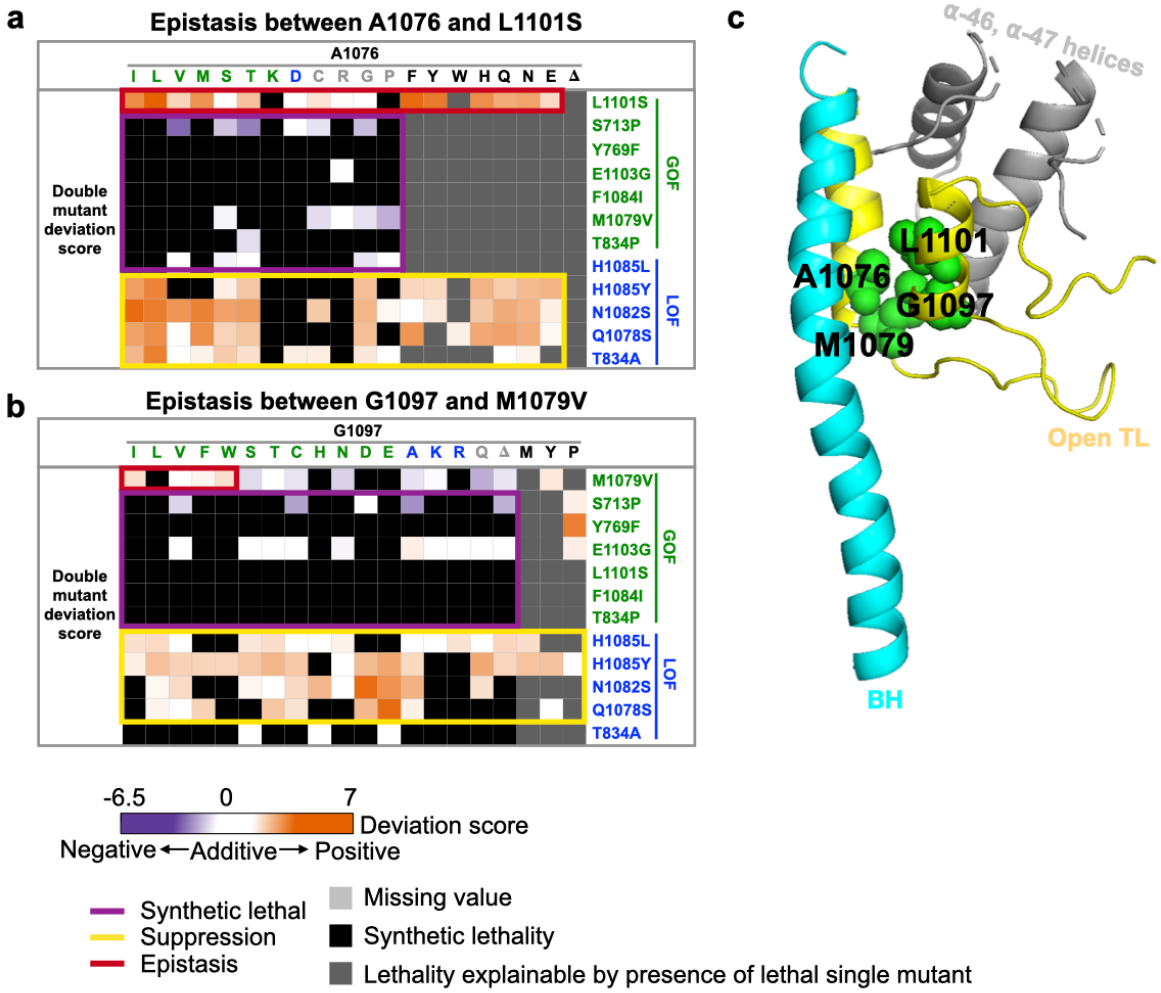


Figure 4. Pol II TL interaction landscape reveals functional dependency of proximal residues. a-b. Specific epistatic interactions observed between hydrophobic residues A1076 and L1101 (A), and M1079 and G1097 are shown as heatmaps (B). The x-axis of both heatmaps are 20 substitutions ordered by predicted 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)[42].

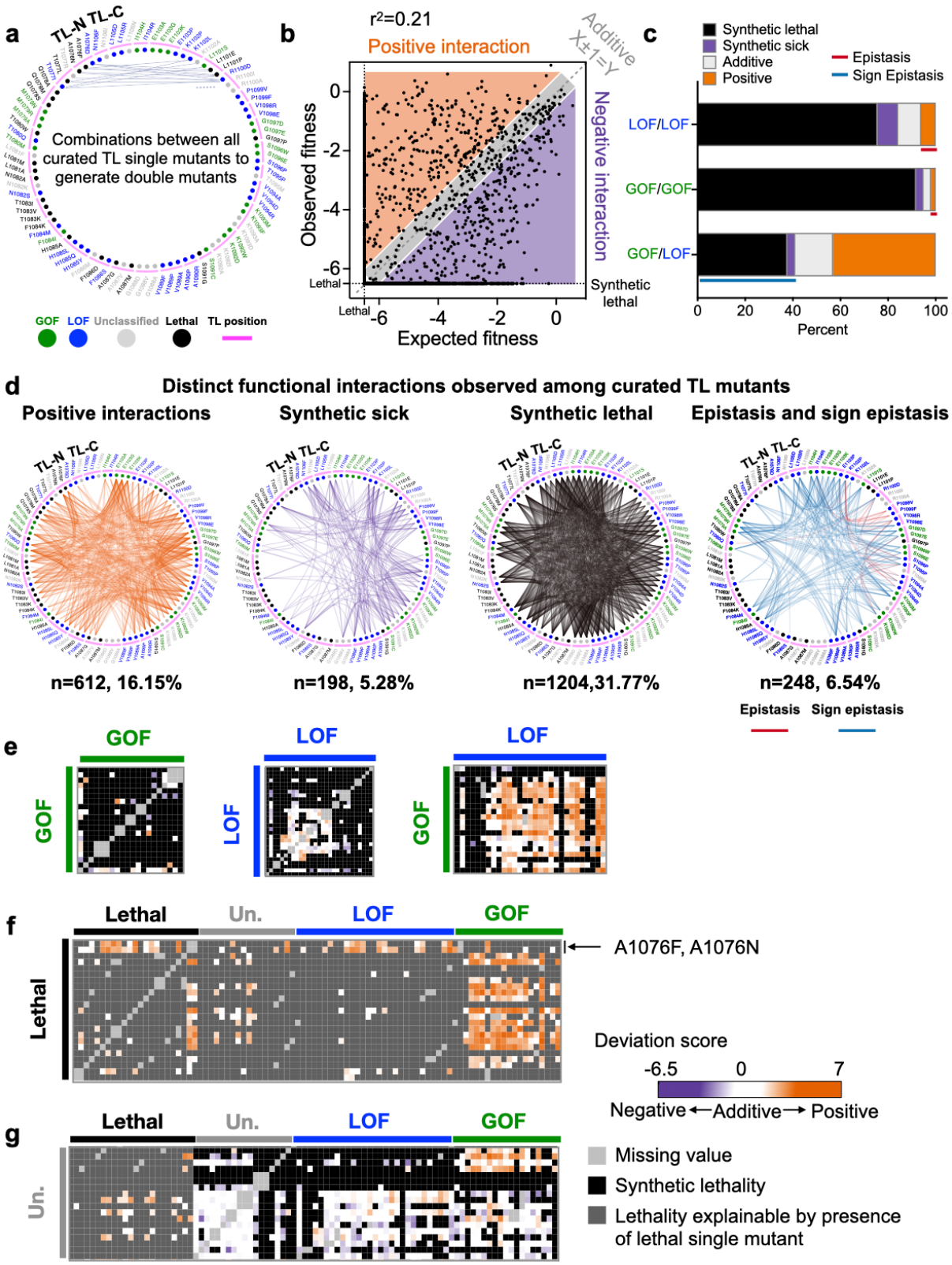


Figure 5. Widespread epistasis in the Pol II TL interaction landscape. a. Design of the pairwise double mutant library. We curated 2-4 substitutions for each TL residue (in total 90 substitutions, n(GOF) = 18, n(LOF) = 30, n(Unclassified) = 19, n(Lethal) = 23) and combined

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 a network format for clarity. In these networks, the TL structure is simplified to a circle, with the TL-N terminal and C-terminal of the TL connected. WT TL residues are represented by magenta arches, and selected substitutions are labeled beside these arches. The color of the label represents the phenotype of each substitution. A colored line connects two substitutions if there is interaction between them, with the line color representing the interaction type. For example, in the “Epistasis and sign epistasis” network, we selected three substitutions at K1102: K1102P, K1102L and K1102A. They show epistasis with V1094D and R, T1095P, S1096P, V1098R and P1099F and V. These interactions are labeled with red lines. The observed epistasis, indicated by red lines, accumulated at the TL-C terminal, indicating a functional dependency in this region. **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**.

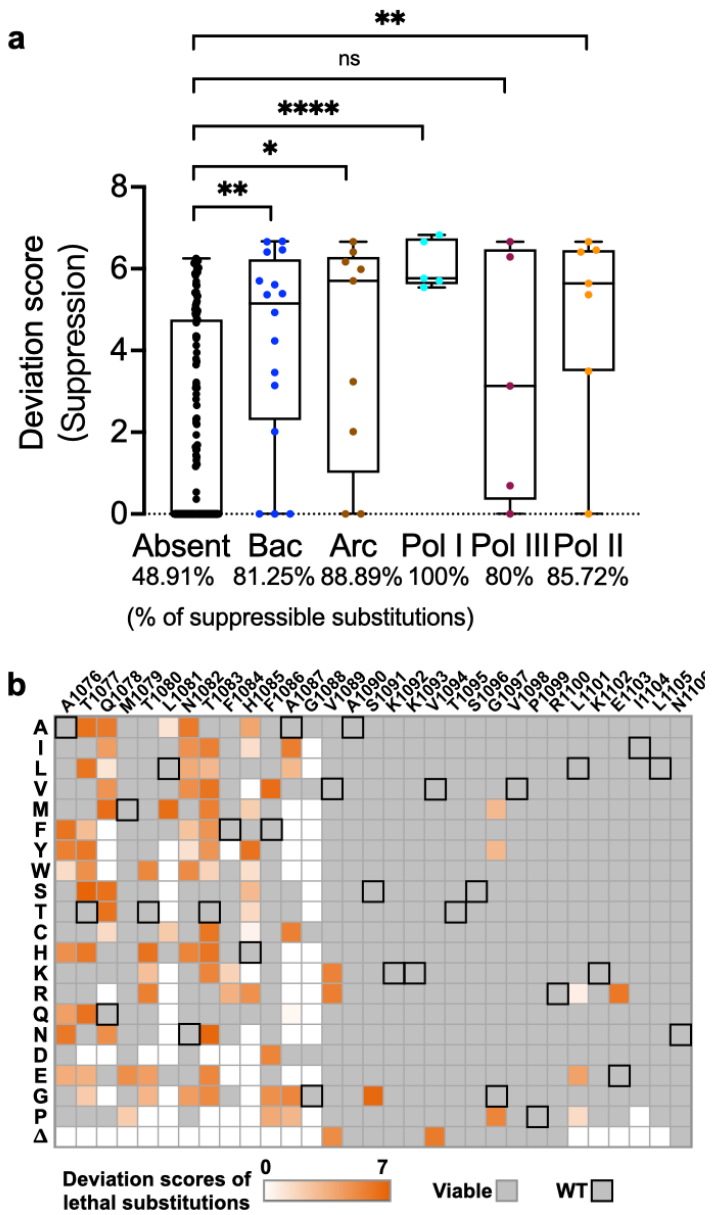


Figure 6. Contextual epistasis shapes TL evolution. a. A comparison of the maximum deviation score of each TL lethal single substitution that is present in any evolutionary TL haplotypes from bacterial, archaeal or eukaryotic Pols versus those that have not been observed in any species. The evolutionary 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 total suppressible lethal single mutants for each group is labeled at the bottom of the plot. Boxes are: center line, median; box limits, second and third quartiles; whiskers, maximum and minimum points. Statistical comparison was done with the Mann-Whitney test and the significant levels ($P < 0.05$) are shown in the figure. **b.** A heatmap displays the highest deviation scores of lethal single substitutions. The x-axis represents the wild-type residues and positions of the yeast Pol II TL, while the y-axis shows 20 substitutions. Viable

substitutions are marked with gray boxes, and wild-type amino acids are indicated with gray boxes with black edge. The highest deviation scores of lethal substitutions are shown in colored boxes, ranging from white to orange, reflecting the strength of suppression.

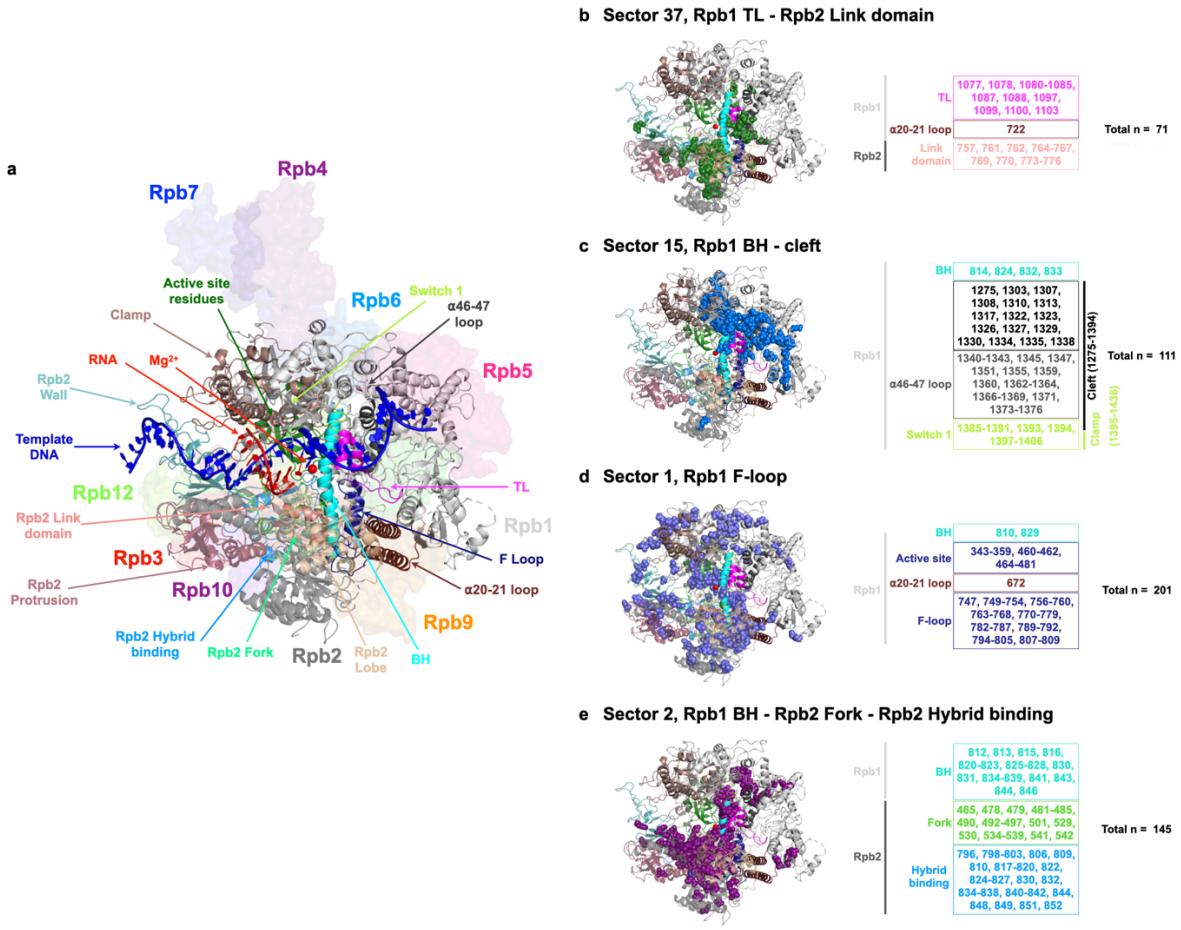
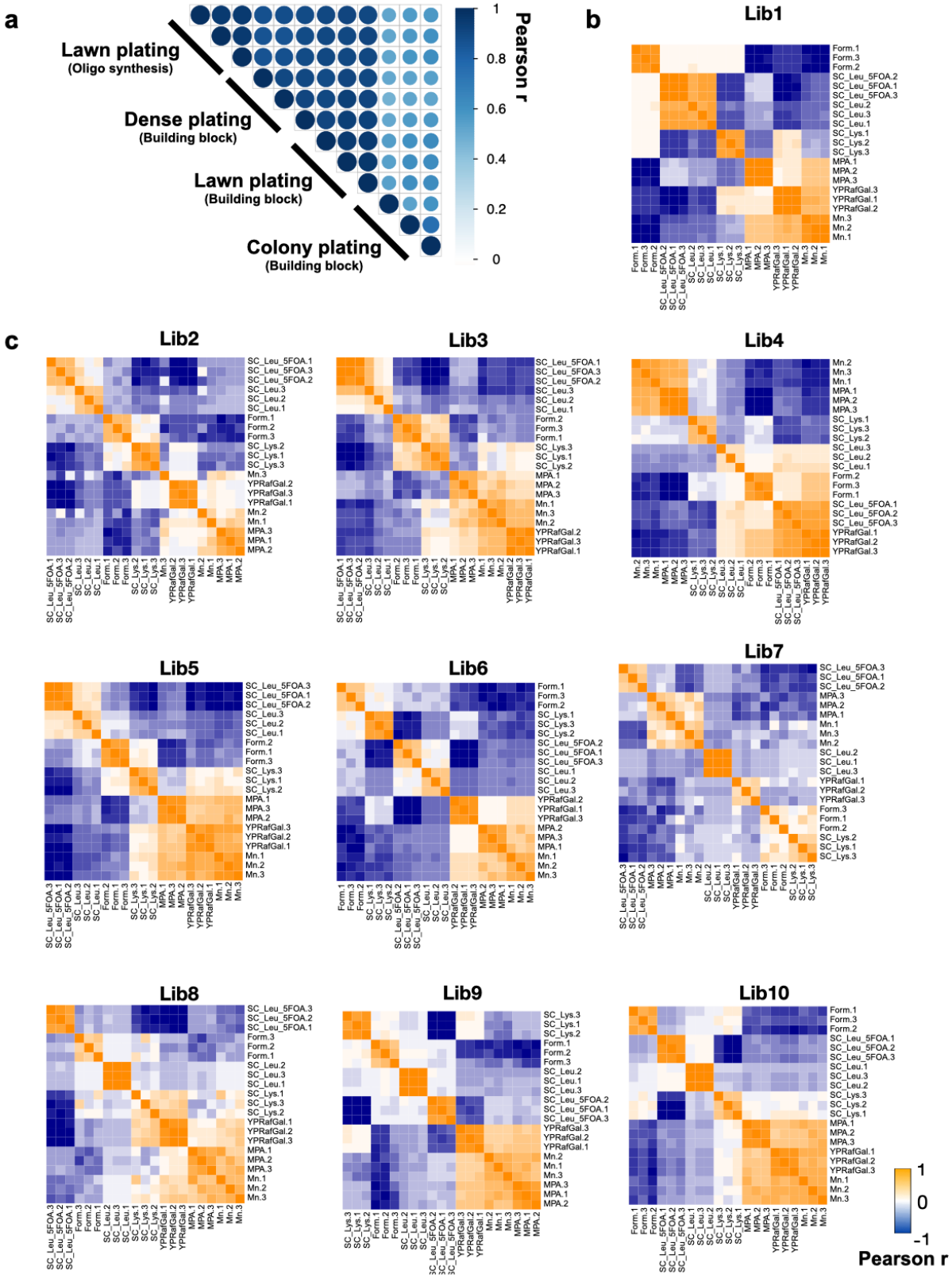
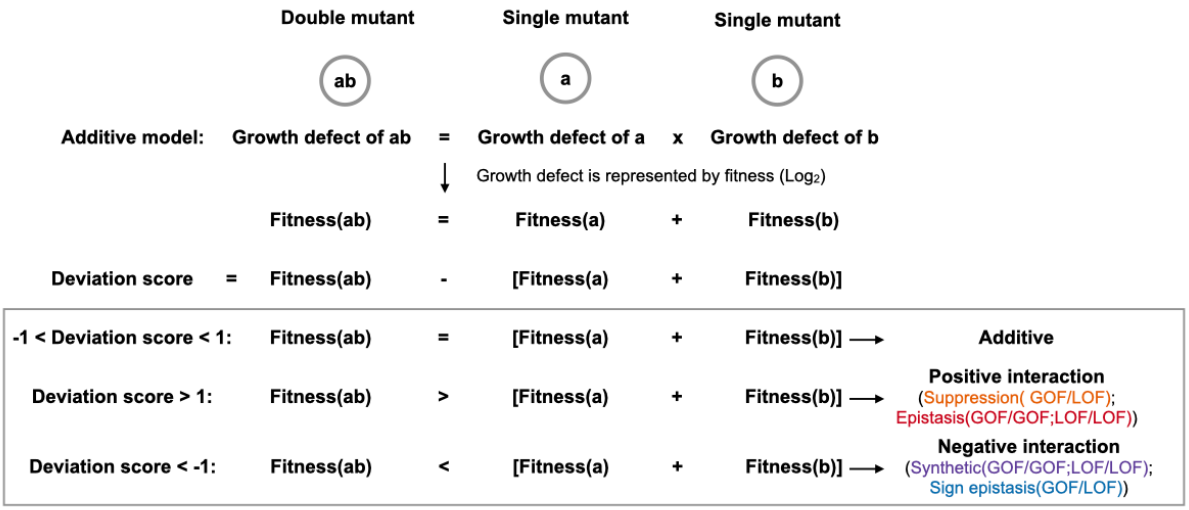


Figure 7. Ultra-conserved TL co-evolves with Pol II residues through diverse pathways. a. Structure representations of Pol II active site. Pol II subunits are showed with surface in different colors**.** The active site domains are in cartoon format. Rpb1 active site residues 346-375 and 436-508. Rpb1 TL, 1076-1106. Rpb1 BH, 810-846. Rpb1 α20-21 loop, 672-737. Rpb1 F loop, 747-810. Rpb1 α46-47 loop, 1339-1378. Rpb1 switch 1, 1385-1407. Rpb1 clamp, 1-95, 235-346, and 1395-1436. Rpb2 link domain, 757-776. Rpb2 Fork, 465-547. Rpb2 hybrid binding, 750-852. Rpb2 wall, 852-973. Rpb2 protrusion, 44-218. Rpb2 lobe, 218-405. **b-e**. Four major coevolution sectors of Pol II active site residues were identified through statistical coupling analysis. The residues within each active site domain and the total number of residues in each sector are labeled. These sectors are mapped on the yeast Pol II Rpb1/Rpb2 structure (PDB: 5C4X)[42].

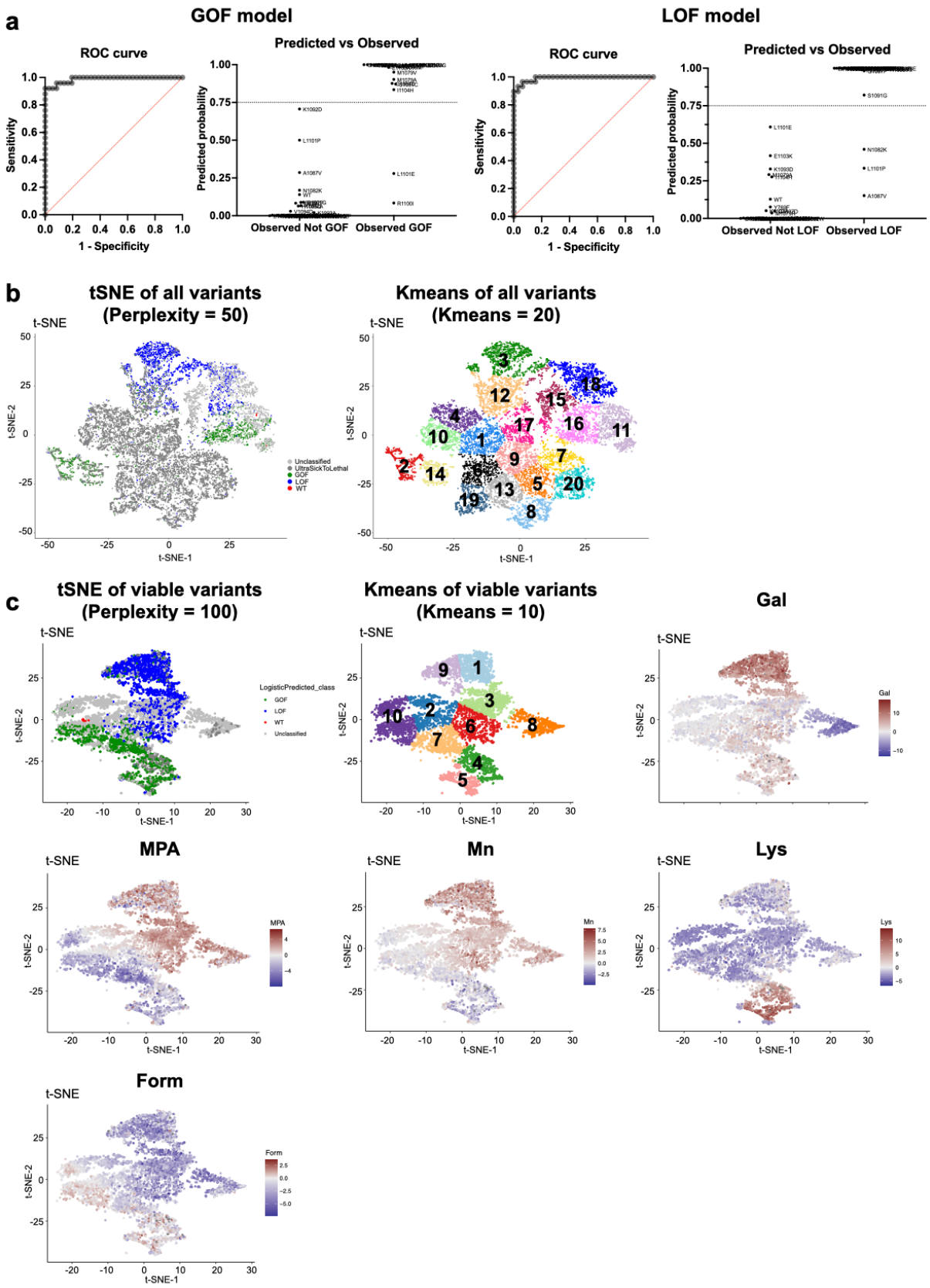


Extended Data Fig.1 Pol II deep mutational scanning is highly reproducible. a. Single mutant growth fitness from mutants in libraries constructed from synthesized oligos correlated

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.

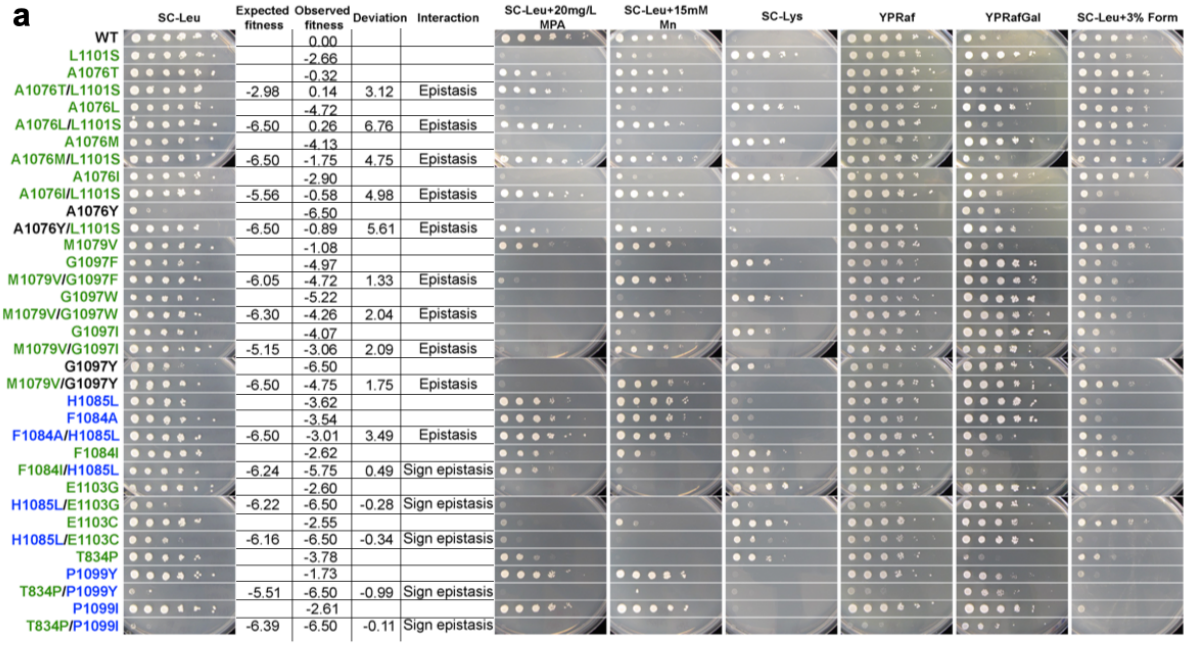


Extended Data Fig.2 Detection of functional interactions by deviation score. For a pseudo double mutant *ab*, the difference between its observed fitness (*ab*) and expected fitness (*ab*) adding the fitness of two constituent single mutants (*a* and *b*) determines the type of interaction between the 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).

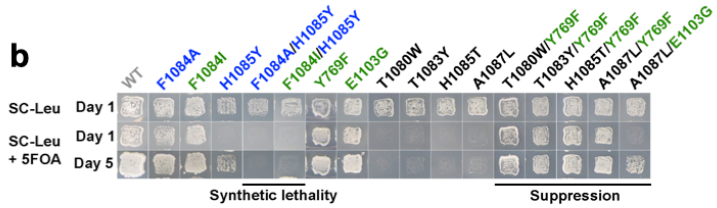


Extended Data Fig.3 Classification of mutant catalytic defects with machine learning

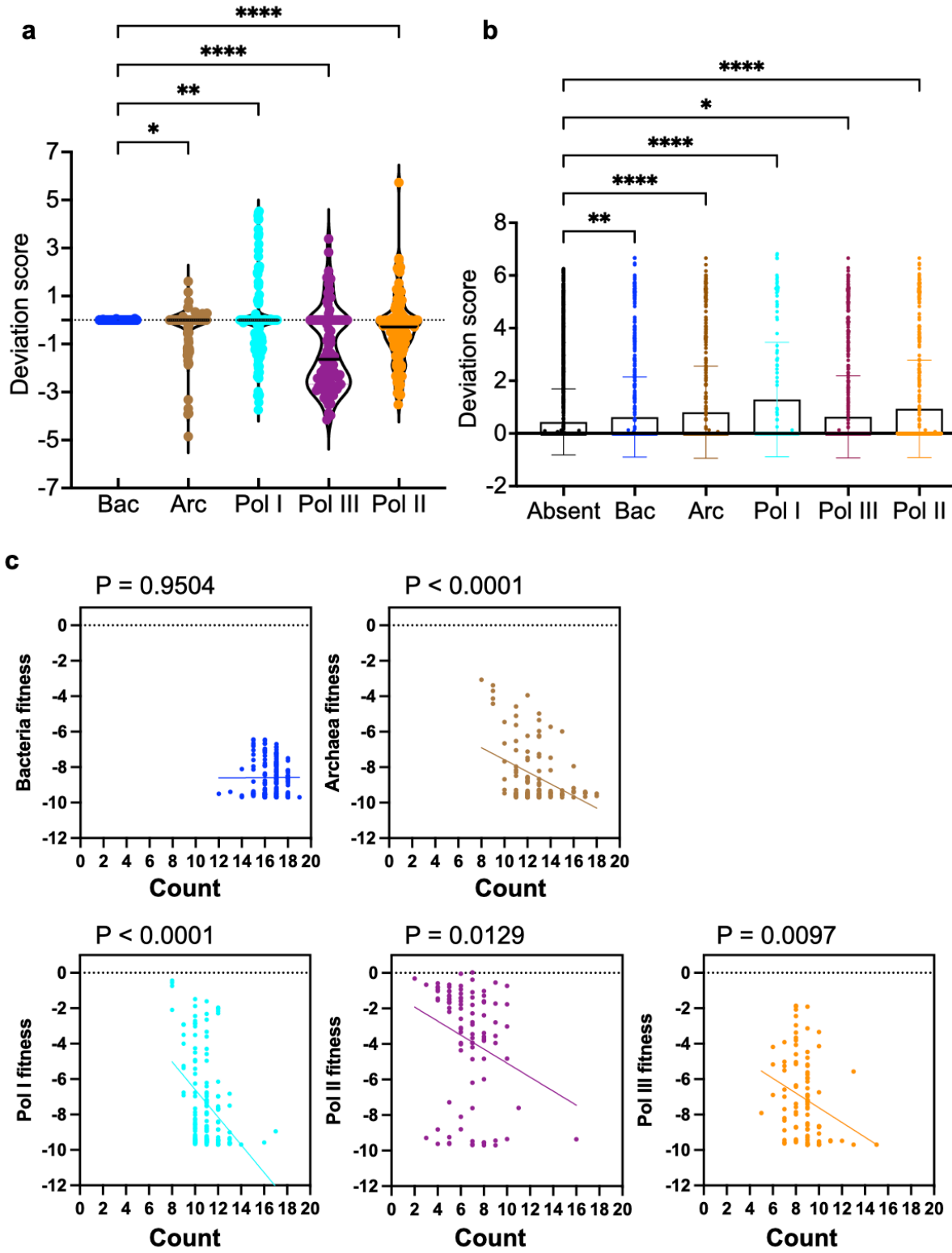
algorithms. a. ROC curves for two multiple logistic regression models used to determine mutant catalytic class. Using 65 mutants with validated in vitro determined catalytic defects and conditional growth fitness measured in our experiment, we trained two models to classify variants as GOF or LOF. The 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 6. Among the 6054 viable mutants, 1390 were classified as GOF (22.96%), 1702 as LOF (28.11%), and 2962 remained unclassified (48.93%). **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.



	Expected fitness	Observed fitness	Deviation	Interaction
WT		0.00		
L1101S		-2.66		
A1076T		-0.32		
A1076T/L1101S	-2.98	0.14	3.12	Epistasis
A1076L		-4.72		
A1076L/L1101S	-6.50	0.26	6.76	Epistasis
A1076M		-4.13		
A1076M/L1101S	-6.50	-1.75	4.75	Epistasis
A1076I		-2.90		
A1076I/L1101S	-5.56	-0.58	4.98	Epistasis
A1076Y		-6.50		
A1076Y/L1101S	-6.50	-0.89	5.61	Epistasis
M1079V		-1.08		
G1097F		-4.97		
M1079V/G1097F	-6.05	-4.72	1.33	Epistasis
G1097W		-5.22		
M1079V/G1097W	-6.30	-4.26	2.04	Epistasis
G1097I		-4.07		
M1079V/G1097I	-5.15	-3.06	2.09	Epistasis
G1097Y		-6.50		
M1079V/G1097Y	-6.50	-4.75	1.75	Epistasis
H1085L		-3.62		
F1084A		-3.54		
F1084A/H1085L	-6.50	-3.01	3.49	Epistasis
F1084I		-2.62		
F1084I/H1085L	-6.24	-5.75	0.49	Sign epistasis
E1103G		-2.60		
H1085L/E1103G	-6.22	-6.50	-0.28	Sign epistasis
E1103C		-2.55		
H1085L/E1103C	-6.16	-6.50	-0.34	Sign epistasis
T834P		-3.78		
P1099Y		-1.73		
T834P/P1099Y	-5.51	-6.50	-0.99	Sign epistasis
P1099I		-2.61		
T834P/P1099I	-6.39	-6.50	-0.11	Sign epistasis

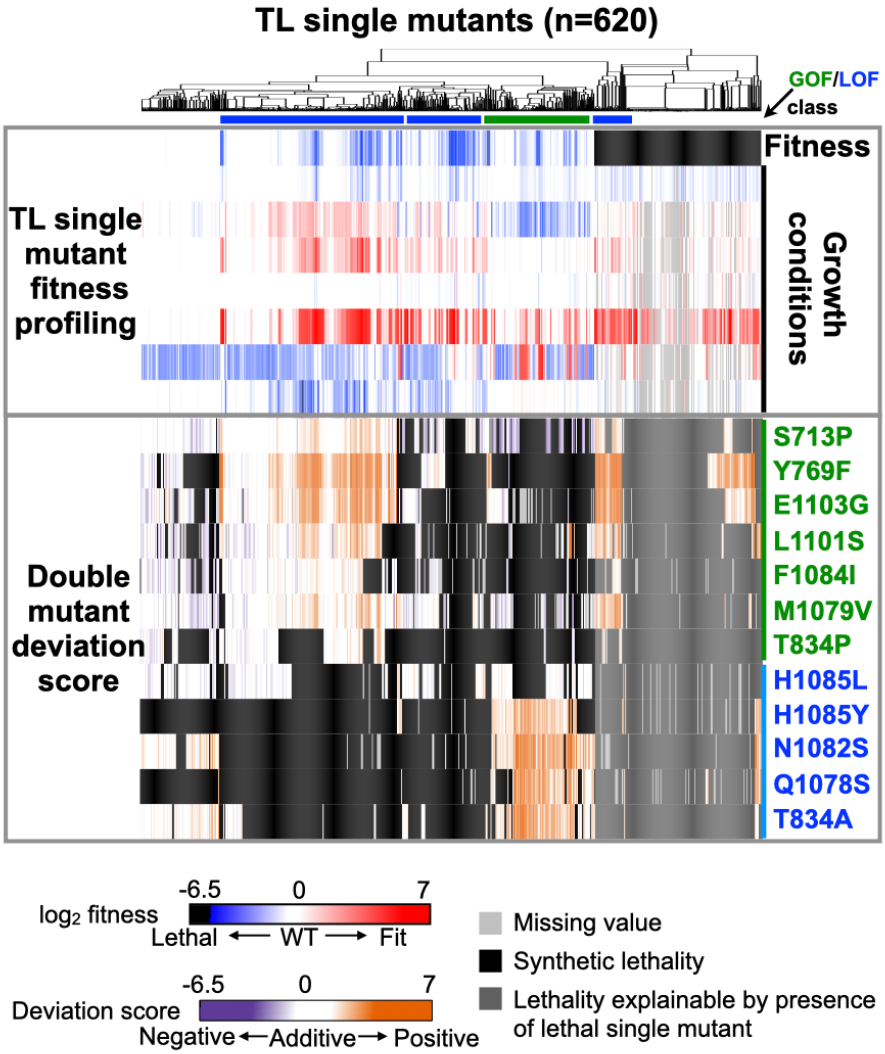


Extended Data Fig.4 Genetic interactions detected with spot assay and patch assay are consistent with high-throughput results. To verify the genetic interactions observed in high throughput experiment, we picked 50 mutants to measure the genetic interactions with spot assay (**a**) or patch assay (**b**). They are consistent. The fitness, deviation scores and corresponding genetic interactions measured in high throughput experiment are indicated in **a.** Specifically, the labeled observed fitness indicates observed general growth fitness that was measured on SC-Leu+5FOA.The names of GOF mutants are labeled with green color, LOF mutants are labeled with blue, ultra-sick to lethal mutants are labeled with black color.

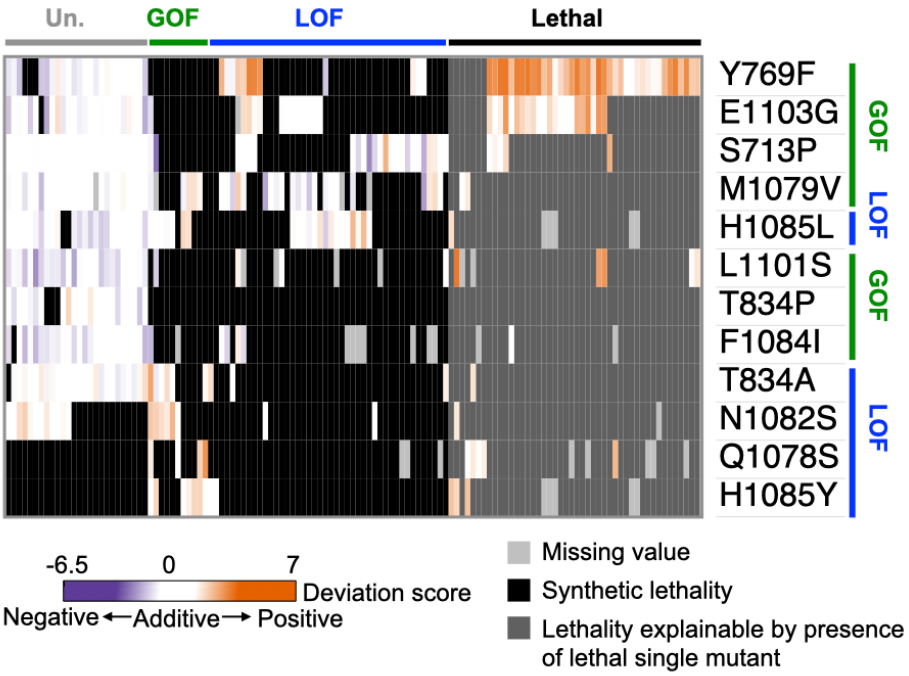


Extended Data Fig.5 Contextual epistasis affects fitness of TL haplotypes. a. Distributions of deviation scores of the TL haplotypes in each group. **b.** Comparison of the mean deviation scores of lethal single substitutions that are present in different species and those that are absent in any species. Standard deviation values are also shown in the bar plot. ANOVA

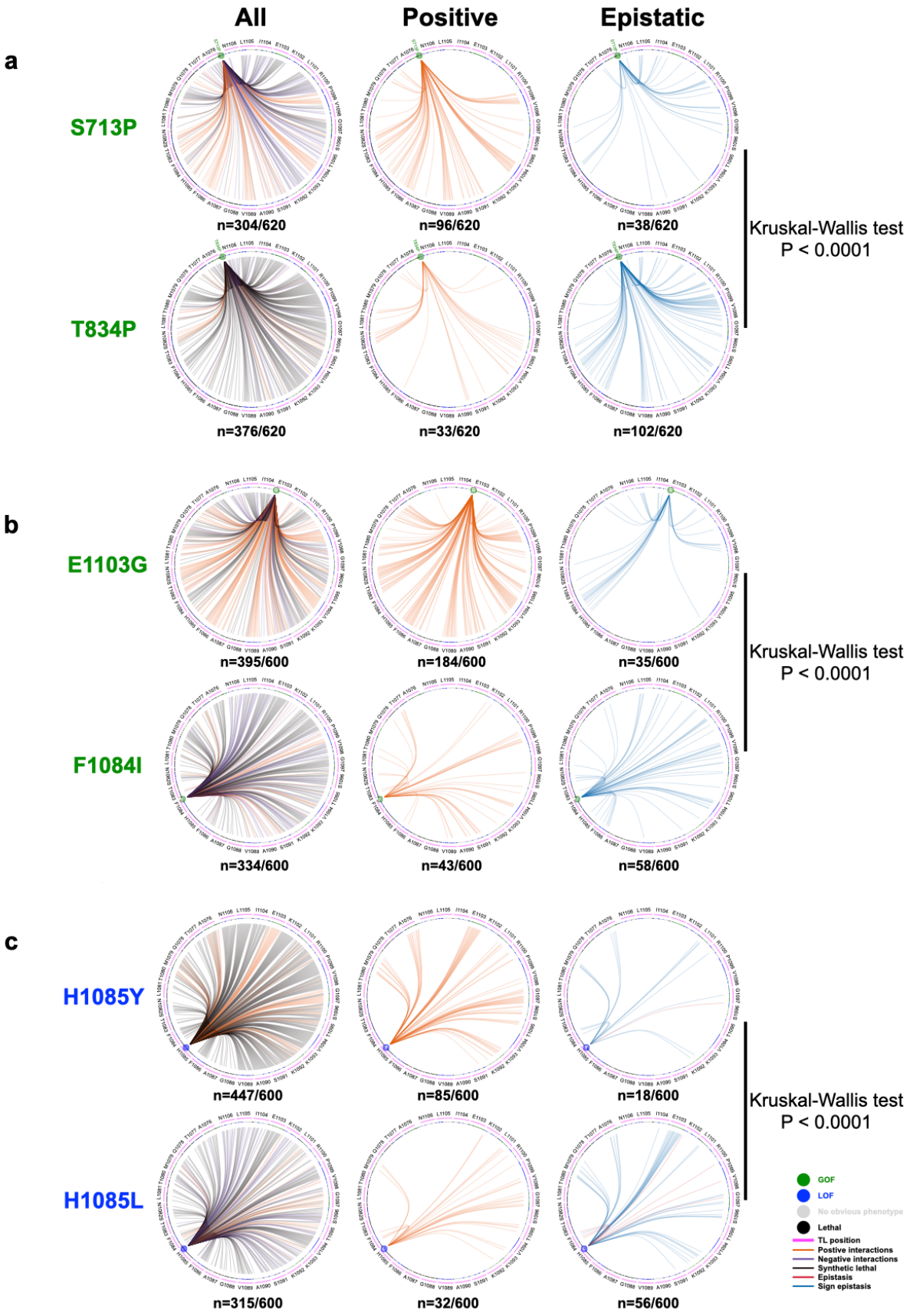
multiple comparison was applied to compare the mean deviation score of the "Absent" group to each of the other groups and significant levels ($P < 0.05$) are shown in the figure. **c.** An xy-plot of evolutionary observed TL haplotypes fitness versus the numbers of substitutions in the haplotypes. Simple 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.



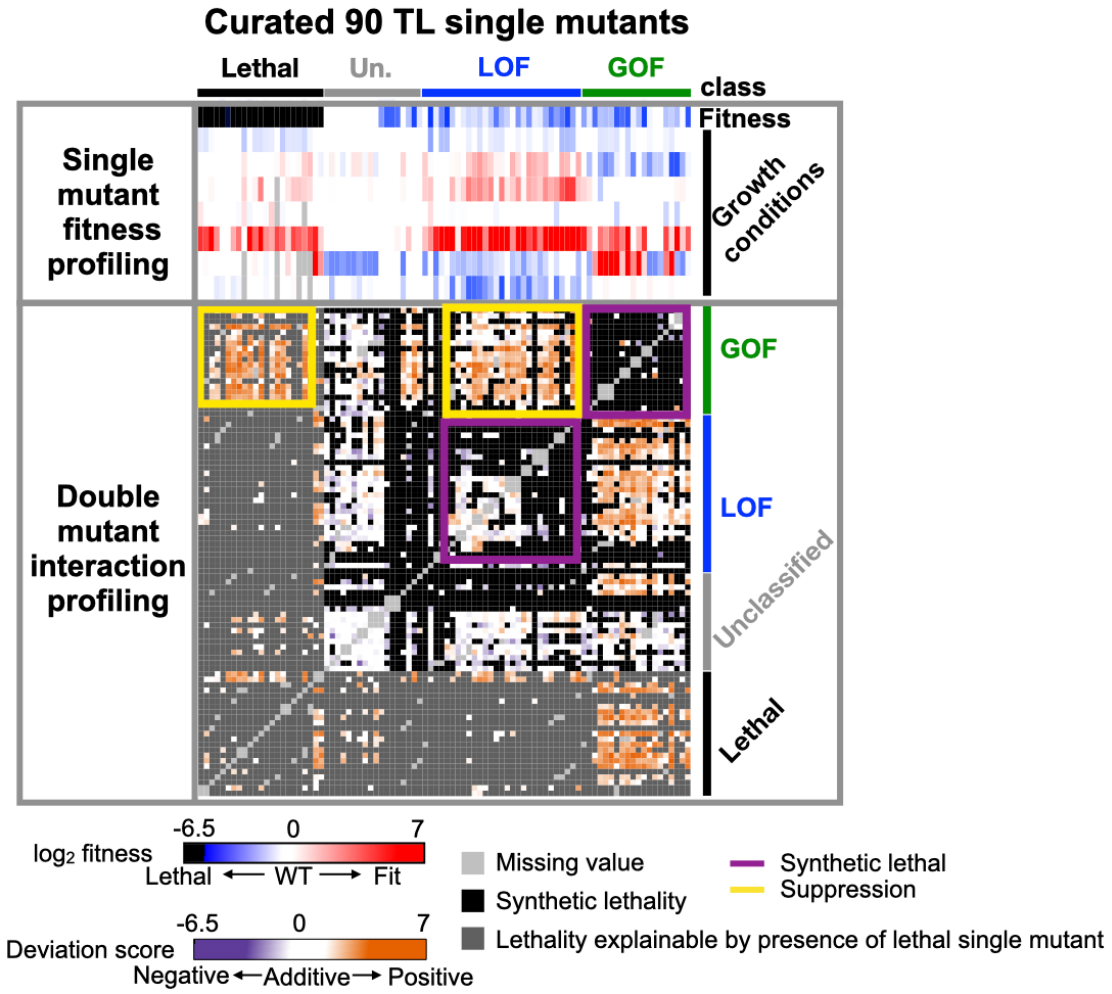
Extended Data Fig.6 The functional interaction landscape of probe mutants. The functional interaction landscape is shown as a heatmap. The upper part of the heatmap shows all Pol II TL single mutant growth fitness profiling across several phenotypes and the single mutants were ordered by hierarchical 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.



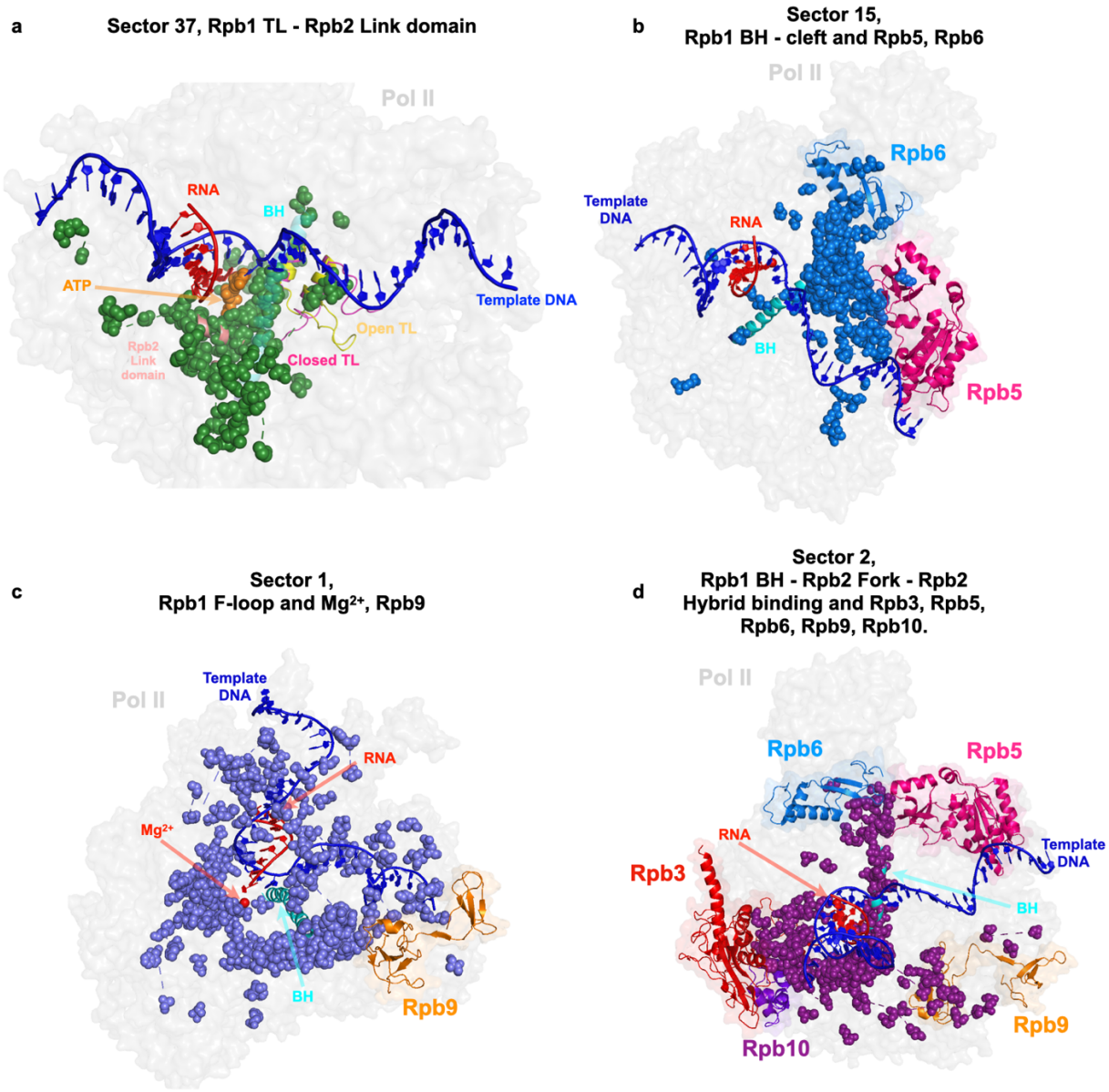
Extended Data Fig.7 Allele-specific interactions. Unique interactions observed between TL substitutions and probe mutants. For each substitution, we analyzed the interquartile range (IQR) of their deviation scores with all probe mutants. Any substitution with deviation score(s) outside of the IQR were 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.


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

703 **Extended Data Fig.8 Interaction networks of selected probe mutants.** The TL is shown in
704 circle with WT residues and positions labeled. All 20 substitutions of each TL residue are
705 represented by a magenta arc under each WT residue, with tick marks representing individual
706 substitutions at that position and are colored by mutant class. Comparison of interaction
707 networks between S713P and T834P (**a**), E1103G and F1084I (**b**) and H1085Y and H1085L (**c**)
708 showed the differences are significant (P < 0.0001). The comparisons were performed with
709 Kruskal-Wallis test with P-value correction with Dunn's multiple comparisons test (Supplemental
710 Table 7).



711
712 **Extended Data Fig.9 The intra-TL functional interaction landscape.** The intra-TL functional
713 interaction landscape is shown as a heatmap. Annotations at the top and right indicate the 90
714 curated single mutants and their predicted phenotypic classes from multiple logistic regression
715 modeling. The upper part of the heatmap shows single mutant growth fitness profiling across
716 multiple phenotypes ordered by groups predicted with logistic regression models. The lower part
717 of the heatmap shows double mutant deviation scores where a colored block at the interaction
718 of x and y coordinates indicates deviation score of the double mutant.



Extended Data Fig.10 Coevolutionary sectors of Pol II active site residues extend to the surface of Pol II, forming connections with other Pol II subunits. a. The coevolutionary sector of Rpb1 TL and Rpb2 link domain residues surrounds the substrate ATP, indicating its potential role in catalysis**.** The sector is mapped on the yeast Pol II structure with TL in open state (PDB: 5C4X) and closed state (PDB: 8U9R). **b**. Residues in sector 15 connect with Rpb5 and Rpb6. **c.** Residues on sector 1 connect with Mg^{2+} and Rpb9. **d.** Residues in sector 2 connect with Rpb5, Rpb6, Rpb3, Rpb9 and Rpb10. These sectors are mapped on the yeast Pol II structure (PDB: 5C4X)[42].

METHODS

Design and Synthesis of TL mutant libraries. We updated and extended the fitness dataset of Qiu *et al*[17]. sing a similar methodology, but with adjusted conditions and a second-generation mutant library strategy to generate a complete Pol II TL mutation-phenotype map and examine genetic interactions. Mutants were constructed by oligo synthesis (Agilent) and screened for

733 phenotypes previously established as informative for Pol II mutant biochemical defects.
734 Programmed oligonucleotide library pools included all 620 single TL residue substitutions and
735 deletions for Rpb1 amino acids 1076-1106 (Library 1), 3914 pairwise double substitutions
736 (Library 2), 4800 targeted double substitutions (Library 6), and 3373 multiple substitutions
737 (Library 3-5), along with the WT *S. cerevisiae* Pol II TL allele at a level of ~15% of the total
738 variants, enabling precise quantification (see Supplemental Table 1, 3). Each synthesized
739 region contained a mutated or WT Pol II TL sequence and two flanking regions at the 5′ and 3'
740 ends of the TL-encoding sequence. These flanking regions also contained designed "PCR
741 handle" (20bp) sequences, allowing distinct subsets of oligos to be amplified from synthesized
742 pools using selected primers for PCR, and additional flanking WT Pol II sequences allow for
743 further extension of homology arms by PCR "sewing" (Details are in Supplemental Method 3).

744 **Introduction of Libraries into yeast**. Synthesized mutant pools were transformed into yeast
745 (CKY283) along with an *RPB1*-encoding plasmid where the TL-encoding sequence was
746 replaced with an *Mlu*I restriction site for linearization as described in Qiu et al[17]. This strategy
747 allows construction of *rpb1* mutant libraries by gap repair between library fragments and the
748 linearized vector. Briefly, the synthesized oligo pools were amplified by limited cycles of
749 emulsion PCR to limit template switching. Extension of flanking homology arms of ~200 bp were
750 added by PCR sewing. Amplified TL sequences with extended flanking regions were co-
751 transformed with linearized pRS315-derived *CEN LEU2* plasmid (pCK892) into CKY283,
752 allowing gap repair via homologous flanking regions. In our experience phenotypes of integrated
753 alleles and single copy plasmids are qualitatively very similar but this has not been
754 systematically compared quantitatively. To detect potential residue-residue interactions between
755 the TL and TL-proximal domains including the Rpb1 Bridge Helix (BH), Funnel Helix alpha-21
756 and Rpb2, the Pol II TL single mutant pool (Library 1, 620 mutant alleles and 111 WT alleles)
757 was co-transformed individually with gapped plasmids encoding an additional *rpb1* allele T834P
758 (Library 7), T834A (Library 8), or Funnel Helix alpha-21 S713P (Library 10)) into CKY283
759 respectively, or with the gapped WT *RPB1* plasmid into a strain with the genomic mutation, *rpb2*
760 Y769F (Library 9). These co-transformations created double mutants between the TL and TL-
761 proximal mutants. The WT allele in single mutant pool represented the single probe mutant due
762 to substitutions outside the TL on the plasmid or in the strain background. To distinguish
763 between a fully WT TL and a WT TL representing the TL of a mutant allele elsewhere, a WT Pol
764 II TL allele with a silent mutant at T1083 (WT codon ACC was replaced with ACT) was co-
765 transformed with plasmid containing gapped WT *RPB1* in a WT strain in parallel. 15% of the
766 transformants with silent mutation were mixed with transformants of double mutants. The silent
767 mutation allowed us to distinguish the WT and the single mutants.

768 **Phenotyping libraries with different conditions.** Each transformation was done in three
769 biological replicates. After transformation, Leu^+ colonies were collected from **SC-Leu** plates and
770 replated on **SC-Leu+5FOA** to select for cells having lost the *RPB1 URA3* plasmid and to
771 measure growth defects. 5-FOA-resistant colonies were then scraped and replated on different
772 conditions to assess functional differences between WT, GOF, and LOF mutants[17,37]. **SC-Leu +**
773 **20 µg/L MPA** (Fisher Scientific) evaluates transcription start site (TSS) shifting. MPA depletes
774 GTP levels in yeast. To generate GTP, *IMD2* expression is required, but its promoter has
775 multiple transcription start sites (TSS)[109,110]. Transcripts from the downstream TSS express
776 functional *IMD2* upon MPA treatment. WT and LOF mutants are resistant to MPA due to their
777 ability to use downstream TSSs for *IMD2* expression, while GOF mutants are sensitive as they
778 can only use upstream TSS sand fail to express functional *IMD2*[53,77]. **SC-Leu + 15 mM Mn**

779 (Sigma): Mn^{2+} compromises transcription fidelity[111] and likely also increases Pol II catalysis;
780 while WT and LOF mutants are resistant, GOF mutants are sensitive either due to decreased
781 fidelity or exacerbated catalytic increase. **YPRaf** and **YPRafGal**: the *gal10Δ56* allele
782 compromised *GAL10* 3'-end formation and polyadenylation, leading to transcription read-
783 through that inhibits the downstream *GAL7* gene [112,113]. The transcription readthrough in WT
784 *gal10Δ56* cells leads to accumulation of a toxic galactose metabolite because of defects in
785 *GAL7* expression. Some GOF and most LOF mutants show varying galactose resistance, likely
786 through increasing termination in the *GAL10-GAL7* intergenic region or otherwise reducing
787 read-through interference with *GAL7*. YPRaf serves as a control for YPRafGal, to demonstrate
788 that the defect on YPRafGal is galactose toxicity. **SC-Lys** tests lysine auxotrophy. The *lys2-*
789 *122∂* allele in the strain background has a *Ty1* transposable element insertion, which prevents
790 normal expression of *LYS2*[114]. WT and LOF mutants are Lys^-, while GOF mutants utilize a
791 normally silent promoter within the *Ty1 ∂* element insertion, becoming Lys^+. **SC-Leu + 3%**
792 **formamide** (JT Baker), formamide destabilizes protein hydrogen bonds[115], with most mutants
793 showing resistance except for slight sensitivity in LOF mutants. Details of cell numbers plated
794 on each plate and screening time of each plate are in Supplemental Table 5. Details of the high
795 efficiency transformation protocol are in Supplemental Method 1.

796 **Generation of libraries for quantification by amplicon sequencing**. Genomic DNA of each
797 screened library was extracted using the Yeastar genomic DNA kit according to manufacturer's
798 instructions (Zymo Research). To ensure adequate DNA for sequencing, the TL regions of all
799 libraries were amplified with PCR cycles that were verified to be in the linear range by qPCR to
800 minimize disturbance of allele distributions, and under emulsion PCR conditions (EURx
801 Micellula DNA Emulsion & Purification (ePCR) PCR kit) to limit template switching. Details are in
802 Supplemental Method 2 and 3. To multiplex samples, we employed a dual indexing strategy
803 wherein 10 initial barcodes for differentiating 10 mutant libraries were added during the initial
804 amplification using 10 pairs of custom primers. In a second amplification, 28 primers containing
805 28 NEB indices were used to add a second index for distinguishing conditions and replicates
806 (NEBNext Multiplex Oligos for Illumina) (see Supplemental Table 4). As a result, a sample-
807 specific barcodes were present for each set of variants. The indexed, pooled samples were
808 sequenced by single end sequencing on an Illumina Next-Seq (150nt reads). On average, over
809 11 million reads were obtained for individual samples with high reproducibility from two rounds
810 of sequencing.

811 **Data cleaning and fitness calculation and normalization**. Reads of mutants were sorted into
812 appropriate libraries and conditions by detecting particular indices after sequencing. Read
813 counts were estimated by a codon-based alignment algorithm to distinguish reads that exactly
814 matched designated codons of mutants [116]. To clean the data, mutant reads with coefficients of
815 variation greater than 0.5 in the control condition (SC-Leu) were excluded from the analysis.
816 The mutant read count was increased by 1 to calculate the allele frequency under different
817 conditions. To measure and compare the phenotypes of all mutants, mutant phenotypic score
818 (fitness) was calculated by allele frequency change of a mutant under selective conditions
819 relative to the unselective condition comparing to the frequency change of WT. The formula for
820 calculating fitness is shown below.

821 $$\text{Fitness (mut)} = \log [f^{mut,\ sele} / f^{mut,\ unsele} - \log [f^{WT,\ sele} / f^{WT,\ unsele}]$$

822 We applied min-max normalization to bring the median growth fitness of mutants measured at
823 ten libraries to the same level for direct comparison (formula is shown below). In each library,

824 we divided mutants into several groups based on their allele counts on the control condition.
825 Mutants with read count differences of less than 10 are present in one group. The WT growth
826 fitness was set as the maximum value and the minimum fitness in each group was the
827 minimum. Min-max normalization was used to equalize the growth fitness into the same range
828 between various groups inside each library. Additionally, we utilized min-max normalization to
829 level the mutant fitness across all ten libraries with WT fitness as Max and minimal fitness in
830 each library as the minimum. As a result, mutant growth fitness was scaled to one range and
831 could be used to determine genetic interactions.

832 $$X' = \frac{X - Xmin}{Xmax - Xmin}$$

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

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

838 (1). -1 < Deviation score < 1, the interaction among the constituent single mutants is additive
839 and mutants are acting independently.

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

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

844 Any mutation with fitness smaller than the lethal threshold (-6.50) was classified as an ultra-sick/
845 lethal mutant and its fitness was normalized to -6.50 for calculation of the deviation score.
846 Synthetic sickness and synthetic lethality were distinguished by whether a double mutant is
847 viable or lethal (fitness is greater than or equals to the lethal threshold -6.5) when two
848 constituent mutations are viable. Synthetic lethality can be further classified into two types.
849 First, additive-synthetic lethality was determined when the expected double mutant fitness
850 calculated by additive model was lethal (expected fitness = -6.5) and the observed double
851 mutant fitness was also lethal (fitness = -6.5) (in this case the deviation score = 0). Second, the
852 beyond-additive synthetic lethality was determined when the expected double mutant was viable
853 (expected fitness > -6.5) while the observed double mutant fitness was lethal (fitness = -6.5) (in
854 this case the deviation score < 0). To separate these two situations in our figures, we labeled
855 additive synthetic lethality as black and beyond-additive synthetic lethality as purple. Details of
856 formulas are in Supplemental Method 4.

857 **Mutant classification using two multiple logistic regression models.** We trained two
858 multiple logistic regression models to distinguish GOF and LOF mutants using phenotypic
859 fitness from SC-Leu+MPA, SC-Lys, and YPRafGal conditions of 65 single mutants. This dataset
860 included 20 mutants with transcription rates measured by biochemical experiments (13 GOF
861 and seven LOF)[17,37,53], and 45 additional mutants that were hierarchical clustered with verified
862 mutants (12 GOF, 26 LOF, 1 WT, and 6 that were not GOF or LOF mutants)[17]. Both models
863 incorporated intercept, main effects, and two-way interactions were involved in defining both
864 models, with a cutoff threshold of 0.75 applied to classify mutants in both of the GOF and LOF
865 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 **(Extended Data Fig. 2A)**. The details are provided in Supplemental Table 6.

Principal component analysis (PCA). Deviation scores of curated and probe double mutants were analyzed in PCA. The scripts used 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).

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 utilized 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)).

Statistical coupling analysis. 5787 eukaryotic homologous sequences of yeast Rpb1 from a published multiple sequence alignment (MSA) and 4000 Rpb2 sequences from NCBI nr database were cleaned and used in the statistical coupling analysis[87]. Sequence identity was reducing to 90% with T-coffee package v12.00.7fb08c2[117]. Pol I, II, and III sequences were separated based on a phylogenetic tree constructed with FastTree 2[118]. Finally, 283 matched Pol II Rpb1 and Rpb2 sequences, with each matched pair sharing the same species name, were connected with 10 Ns and aligned with T-coffee. This newly generated MSA was used for statistical coupling analysis with the python-based package pySCA v6.1[97]. The scripts were adapted from https://github.com/ranganathanlab/pySCA.

Supplemental Tables

Supplemental Table 1. Library summary

Supplemental Table 2. Double mutant interactions observed in high throughput system and growth assays in published papers

Supplemental Table 3. Strains and plasmids

Supplemental Table 4. Primers

Supplemental Table 5. Phenotyping details

Supplemental Table 6. MLR_models_summary

904 Supplemental Table 7. Kruskal-Wallis tests

905 **Supplemental methods**

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

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

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

911 Supplemental Method 4. Formulas of calculating functional interactions.

912 **DATA AVAILIABILITY**

913 Raw sequencing data has been deposited on the NCBI SRA (Sequence Read Archive)
914 database under BioProject PRJNA948661. Processed mutant count, fitness and processing
915 codes are available through GitHub (https://github.com/Kaplan-Lab-Pitt/TLs_Screening.git).
916 Strains and plasmids will be provided upon request. Source data are provided with this paper.

917 **CODE AVAILIABILITY**

918 The codes for calculating deviation scores, PCA, t-SNE projection, statistical coupling analysis,
919 and generating figures are available in GitHub (https://github.com/Kaplan-Lab-
920 Pitt/TLs_Screening.git).

921 **REFERENCES**

922 1. Cramer, P. Multisubunit RNA polymerases. *Curr Opin Struct Biol* **12**, 89-97 (2002).
923 2. Werner, F. & Grohmann, D. Evolution of multisubunit RNA polymerases in the three
924 domains of life. *Nat Rev Microbiol* **9**, 85-98 (2011).
925 3. Allison, L.A., Moyle, M., Shales, M. & Ingles, C.J. Extensive homology among the largest
926 subunits of eukaryotic and prokaryotic RNA polymerases. *Cell* **42**, 599-610 (1985).
927 4. Zhang, G. et al. Crystal Structure of Thermus aquaticus Core RNA Polymerase at 3.3 Å
928 Resolution. *Cell* **98**, 811-824 (1999).
929 5. Vassylyev, D.G. et al. Crystal structure of a bacterial RNA polymerase holoenzyme at
930 2.6 A resolution. *Nature* **417**, 712-9 (2002).
931 6. Hirata, A., Klein, B.J. & Murakami, K.S. The X-ray crystal structure of RNA polymerase
932 from Archaea. *Nature* **451**, 851-4 (2008).
933 7. Gnatt, A.L., Cramer, P., Fu, J., Bushnell, D.A. & Kornberg, R.D. Structural basis of
934 transcription: an RNA polymerase II elongation complex at 3.3 A resolution. *Science*
935 **292**, 1876-82 (2001).
936 8. Cramer, P., Bushnell, D.A. & Kornberg, R.D. Structural basis of transcription: RNA
937 polymerase II at 2.8 angstrom resolution. *Science* **292**, 1863-76 (2001).
938 9. Fernandez-Tornero, C. et al. Crystal structure of the 14-subunit RNA polymerase I.
939 *Nature* **502**, 644-9 (2013).
940 10. Hoffmann, N.A. et al. Molecular structures of unbound and transcribing RNA polymerase
941 III. *Nature* **528**, 231-6 (2015).
942 11. Malinen, A.M. et al. Active site opening and closure control translocation of multisubunit
943 RNA polymerase. *Nucleic Acids Res* **40**, 7442-51 (2012).

12. Wang, D., Bushnell, D.A., Westover, K.D., Kaplan, C.D. & Kornberg, R.D. Structural basis of transcription: role of the trigger loop in substrate specificity and catalysis. *Cell* **127**, 941-54 (2006).
13. Kaplan, C.D. 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. Dangkulwanich, M. et al. Complete dissection of transcription elongation reveals slow translocation of RNA polymerase II in a linear ratchet mechanism. *Elife* **2**, e00971 (2013).
15. Bar-Nahum, G. et al. A ratchet mechanism of transcription elongation and its control. *Cell* **120**, 183-93 (2005).
16. Weinzierl, R.O. The nucleotide addition cycle of RNA polymerase is controlled by two molecular hinges in the Bridge Helix domain. *BMC Biol* **8**, 134 (2010).
17. Qiu, C. et al. High-Resolution Phenotypic Landscape of the RNA Polymerase II Trigger Loop. *PLoS Genet* **12**, e1006321 (2016).
18. Da, L.T. et al. Bridge helix bending promotes RNA polymerase II backtracking through a critical and conserved threonine residue. *Nat Commun* **7**, 11244 (2016).
19. Mazumder, A., Lin, M., Kapanidis, A.N. & Ebright, R.H. Closing and opening of the RNA polymerase trigger loop. *Proc Natl Acad Sci U S A* **117**, 15642-15649 (2020).
20. Wang, B., Predeus, A.V., Burton, Z.F. & Feig, M. Energetic and structural details of the trigger-loop closing transition in RNA polymerase II. *Biophys J* **105**, 767-75 (2013).
21. Larson, M.H. et al. Trigger loop dynamics mediate the balance between the transcriptional fidelity and speed of RNA polymerase II. *Proc Natl Acad Sci U S A* **109**, 6555-60 (2012).
22. Fouqueau, T., Zeller, M.E., Cheung, A.C., Cramer, P. & Thomm, M. The RNA polymerase trigger loop functions in all three phases of the transcription cycle. *Nucleic Acids Res* **41**, 7048-59 (2013).
23. Kireeva, M.L. et al. Transient reversal of RNA polymerase II active site closing controls fidelity of transcription elongation. *Mol Cell* **30**, 557-66 (2008).
24. Liu, X., Bushnell, D.A. & Kornberg, R.D. RNA polymerase II transcription: structure and mechanism. *Biochim Biophys Acta* **1829**, 2-8 (2013).
25. Kaplan, C.D. & Kornberg, R.D. A bridge to transcription by RNA polymerase. *J Biol* **7**, 39 (2008).
26. Tan, L., Wiesler, S., Trzaska, D., Carney, H.C. & Weinzierl, R.O. Bridge helix and trigger loop perturbations generate superactive RNA polymerases. *J Biol* **7**, 40 (2008).
27. Silva, D.A. et al. Millisecond dynamics of RNA polymerase II translocation at atomic resolution. *Proc Natl Acad Sci U S A* **111**, 7665-70 (2014).
28. Lin, G. et al. Structural basis of transcription: RNA polymerase II substrate binding and metal coordination using a free-electron laser. *Proc Natl Acad Sci U S A* **121**, e2318527121 (2024).
29. Fong, N. et al. Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate. *Genes Dev* **28**, 2663-76 (2014).
30. Xu, L. et al. Dissecting the chemical interactions and substrate structural signatures governing RNA polymerase II trigger loop closure by synthetic nucleic acid analogues. *Nucleic Acids Res* **42**, 5863-70 (2014).
31. Vassylyev, D.G. et al. Structural basis for substrate loading in bacterial RNA polymerase. *Nature* **448**, 163-8 (2007).
32. Da, L.T., Wang, D. & Huang, X. 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**, 2399-406 (2012).

994 33. Liu, B., Zuo, Y. & Steitz, T.A. Structures of E. coli sigmaS-transcription initiation
995 complexes provide new insights into polymerase mechanism. *Proc Natl Acad Sci U S A*
996 **113**, 4051-6 (2016).
997 34. Seibold, S.A. et al. Conformational coupling, bridge helix dynamics and active site
998 dehydration in catalysis by RNA polymerase. *Biochim Biophys Acta* **1799**, 575-87
999 (2010).
1000 35. Malagon, F. et al. Mutations in the Saccharomyces cerevisiae RPB1 gene conferring
1001 hypersensitivity to 6-azauracil. *Genetics* **172**, 2201-9 (2006).
1002 36. Kaplan, C.D., Larsson, K.M. & Kornberg, R.D. The RNA polymerase II trigger loop
1003 functions in substrate selection and is directly targeted by alpha-amanitin. *Mol Cell* **30**,
1004 547-56 (2008).
1005 37. Kaplan, C.D., Jin, H., Zhang, I.L. & Belyanin, A. Dissection of Pol II trigger loop function
1006 and Pol II activity-dependent control of start site selection in vivo. *PLoS Genet* **8**,
1007 e1002627 (2012).
1008 38. Matthew H. Larsona, Jing Zhoub,1, Craig D. Kaplanc,1, Murali Palangatd,2, Roger D.
1009 Kornberge, Robert Landickd, and Steven M. Blocka,b,f,3. Trigger loop dynamics mediate
1010 the balance between the transcriptional fidelity and speed of RNA polymerase II. (2012).
1011 39. Kireeva, M.L. et al. Molecular dynamics and mutational analysis of the catalytic and
1012 translocation cycle of RNA polymerase. *BMC Biophys* **5**, 11 (2012).
1013 40. Windgassen, T.A. et al. Trigger-helix folding pathway and SI3 mediate catalysis and
1014 hairpin-stabilized pausing by Escherichia coli RNA polymerase. *Nucleic Acids Res* **42**,
1015 12707-21 (2014).
1016 41. Nayak, D., Voss, M., Windgassen, T., Mooney, R.A. & Landick, R. Cys-pair reporters
1017 detect a constrained trigger loop in a paused RNA polymerase. *Mol Cell* **50**, 882-93
1018 (2013).
1019 42. Barnes, C.O. et al. Crystal Structure of a Transcribing RNA Polymerase II Complex
1020 Reveals a Complete Transcription Bubble. *Mol Cell* **59**, 258-69 (2015).
1021 43. Cheung, A.C. & Cramer, P. Structural basis of RNA polymerase II backtracking, arrest
1022 and reactivation. *Nature* **471**, 249-53 (2011).
1023 44. Sydow, J.F. et al. Structural basis of transcription: mismatch-specific fidelity mechanisms
1024 and paused RNA polymerase II with frayed RNA. *Mol Cell* **34**, 710-21 (2009).
1025 45. Sekine, S., Murayama, Y., Svetlov, V., Nudler, E. & Yokoyama, S. The ratcheted and
1026 ratchetable structural states of RNA polymerase underlie multiple transcriptional
1027 functions. *Mol Cell* **57**, 408-21 (2015).
1028 46. Kuldell, J.C. & Kaplan, C.D. RNA Polymerase II Activity Control of Gene Expression and
1029 Involvement in Disease. *Journal of Molecular Biology* (2024).
1030 47. Kettenberger, H., Armache, K.J. & Cramer, P. Complete RNA polymerase II elongation
1031 complex structure and its interactions with NTP and TFIIS. *Mol Cell* **16**, 955-65 (2004).
1032 48. Lennon, C.W. et al. Direct interactions between the coiled-coil tip of DksA and the trigger
1033 loop of RNA polymerase mediate transcriptional regulation. *Genes Dev* **26**, 2634-46
1034 (2012).
1035 49. Hein, P.P. et al. RNA polymerase pausing and nascent-RNA structure formation are
1036 linked through clamp-domain movement. *Nat Struct Mol Biol* **21**, 794-802 (2014).
1037 50. Kaster, B.C., Knippa, K.C., Kaplan, C.D. & Peterson, D.O. RNA Polymerase II Trigger
1038 Loop Mobility: INDIRECT EFFECTS OF Rpb9. *J Biol Chem* **291**, 14883-95 (2016).
1039 51. Taatjes, A.C.S.a.D.J. Structure and mechanism of the RNA polymerase II transcription
1040 machinery. (2020).
1041 52. Leng, X.Y. et al. Organismal benefits of transcription speed control at gene boundaries.
1042 *Embo Reports* **21**(2020).
1043 53. Braberg, H. et al. From structure to systems: high-resolution, quantitative genetic
1044 analysis of RNA polymerase II. *Cell* **154**, 775-88 (2013).

1045 54. Tesileanu, T., Colwell, L.J. & Leibler, S. Protein sectors: statistical coupling analysis
1046 versus conservation. *PLoS Comput Biol* **11**, e1004091 (2015).
1047 55. Phillips, P.C. The language of gene interaction. *Genetics* **149**, 1167-71 (1998).
1048 56. Breen, M.S., Kemena, C., Vlasov, P.K., Notredame, C. & Kondrashov, F.A. Epistasis as
1049 the primary factor in molecular evolution. *Nature* **490**, 535-8 (2012).
1050 57. Starr, T.N. & Thornton, J.W. Epistasis in protein evolution. *Protein Sci* **25**, 1204-18
1051 (2016).
1052 58. Ortlund, E.A., Bridgham, J.T., Redinbo, M.R. & Thornton, J.W. Crystal structure of an
1053 ancient protein: evolution by conformational epistasis. *Science* **317**, 1544-8 (2007).
1054 59. Karageorgi, M. et al. Genome editing retraces the evolution of toxin resistance in the
1055 monarch butterfly. *Nature* **574**, 409-412 (2019).
1056 60. Phillips, P.C. Epistasis--the essential role of gene interactions in the structure and
1057 evolution of genetic systems. *Nat Rev Genet* **9**, 855-67 (2008).
1058 61. Faure, A.J. et al. Mapping the energetic and allosteric landscapes of protein binding
1059 domains. *Nature* **604**, 175-183 (2022).
1060 62. Park, Y., Metzger, B.P.H. & Thornton, J.W. Epistatic drift causes gradual decay of
1061 predictability in protein evolution. *Science* **376**, 823-830 (2022).
1062 63. Ding, D. et al. Co-evolution of interacting proteins through non-contacting and non-
1063 specific mutations. *Nat Ecol Evol* **6**, 590-603 (2022).
1064 64. Kondrashov, A.S., Sunyaev, S. & Kondrashov, F.A. Dobzhansky-Muller incompatibilities
1065 in protein evolution. *Proc Natl Acad Sci U S A* **99**, 14878-83 (2002).
1066 65. Lunzer, M., Golding, G.B. & Dean, A.M. Pervasive cryptic epistasis in molecular
1067 evolution. *PLoS Genet* **6**, e1001162 (2010).
1068 66. Natarajan, C. et al. Epistasis among adaptive mutations in deer mouse hemoglobin.
1069 *Science* **340**, 1324-7 (2013).
1070 67. Doud, M.B., Ashenberg, O. & Bloom, J.D. Site-Specific Amino Acid Preferences Are
1071 Mostly Conserved in Two Closely Related Protein Homologs. *Mol Biol Evol* **32**, 2944-60
1072 (2015).
1073 68. Starr, T.N. et al. Shifting mutational constraints in the SARS-CoV-2 receptor-binding
1074 domain during viral evolution. *Science* **377**, 420-424 (2022).
1075 69. Haddox, H.K., Dingens, A.S., Hilton, S.K., Overbaugh, J. & Bloom, J.D. Mapping
1076 mutational effects along the evolutionary landscape of HIV envelope. *Elife* **7**(2018).
1077 70. Viktorovskaya, O.V. et al. Divergent contributions of conserved active site residues to
1078 transcription by eukaryotic RNA polymerases I and II. *Cell Rep* **4**, 974-84 (2013).
1079 71. Mani, R., St Onge, R.P., Hartman, J.L.t., Giaever, G. & Roth, F.P. Defining genetic
1080 interaction. *Proc Natl Acad Sci U S A* **105**, 3461-6 (2008).
1081 72. Qiu, C. & Kaplan, C.D. Functional assays for transcription mechanisms in high-
1082 throughput. *Methods* **159-160**, 115-123 (2019).
1083 73. Lin, X. et al. Nested epistasis enhancer networks for robust genome regulation. *Science*
1084 **377**, 1077-1085 (2022).
1085 74. Fowler, D.M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat*
1086 *Methods* **11**, 801-7 (2014).
1087 75. Sergey Ioffe, C.S. Batch Normalization: Accelerating Deep Network Training by
1088 Reducing Internal Covariate Shift. *arXiv:1502.03167v3* (2015).
1089 76. Hill, W.G., Goddard, M.E. & Visscher, P.M. Data and theory point to mainly additive
1090 genetic variance for complex traits. *PLoS Genet* **4**, e1000008 (2008).
1091 77. Malik, I., Qiu, C., Snavely, T. & Kaplan, C.D. Wide-ranging and unexpected
1092 consequences of altered Pol II catalytic activity in vivo. *Nucleic Acids Research* **45**,
1093 4431-4451 (2017).
1094 78. (!!! INVALID CITATION !!! 17,37).

1095 79. Scull, C.E., Ingram, Z.M., Lucius, A.L. & Schneider, D.A. A Novel Assay for RNA
1096 Polymerase I Transcription Elongation Sheds Light on the Evolutionary Divergence of
1097 Eukaryotic RNA Polymerases. *Biochemistry* **58**, 2116-2124 (2019).
1098 80. Duan, B., Qiu, C., Lockless, S.W., Sze, S.H. & Kaplan, C.D. Higher-order epistasis
1099 within Pol II trigger loop haplotypes. *Genetics* **228**(2024).
1100 81. Brueckner, F. & Cramer, P. Structural basis of transcription inhibition by alpha-amanitin
1101 and implications for RNA polymerase II translocation. *Nat Struct Mol Biol* **15**, 811-8
1102 (2008).
1103 82. Carvalho, A.T., Fernandes, P.A. & Ramos, M.J. The Catalytic Mechanism of RNA
1104 Polymerase II. *J Chem Theory Comput* **7**, 1177-88 (2011).
1105 83. Huang, X. et al. RNA polymerase II trigger loop residues stabilize and position the
1106 incoming nucleotide triphosphate in transcription. *Proc Natl Acad Sci U S A* **107**, 15745-
1107 50 (2010).
1108 84. Castro, C. et al. Nucleic acid polymerases use a general acid for nucleotidyl transfer. *Nat*
1109 *Struct Mol Biol* **16**, 212-8 (2009).
1110 85. Unarta, I.C., Goonetilleke, E.C., Wang, D. & Huang, X. Nucleotide addition and cleavage
1111 by RNA polymerase II: Coordination of two catalytic reactions using a single active site.
1112 *J Biol Chem* **299**, 102844 (2023).
1113 86. Mishanina, T.V., Palo, M.Z., Nayak, D., Mooney, R.A. & Landick, R. Trigger loop of RNA
1114 polymerase is a positional, not acid-base, catalyst for both transcription and
1115 proofreading. *Proc Natl Acad Sci U S A* **114**, E5103-E5112 (2017).
1116 87. Palo, M.Z., Zhu, J., Mishanina, T.V. & Landick, R. Conserved Trigger Loop Histidine of
1117 RNA Polymerase II Functions as a Positional Catalyst Primarily through Steric Effects.
1118 *Biochemistry* **60**, 3323-3336 (2021).
1119 88. Araya, C.L. et al. A fundamental protein property, thermodynamic stability, revealed
1120 solely from large-scale measurements of protein function. *Proc Natl Acad Sci U S A* **109**,
1121 16858-63 (2012).
1122 89. Melamed, D., Young, D.L., Gamble, C.E., Miller, C.R. & Fields, S. Deep mutational
1123 scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein.
1124 *RNA* **19**, 1537-51 (2013).
1125 90. Harms, M.J. & Thornton, J.W. Analyzing protein structure and function using ancestral
1126 gene reconstruction. *Curr Opin Struct Biol* **20**, 360-6 (2010).
1127 91. Fowler, D.M. et al. High-resolution mapping of protein sequence-function relationships.
1128 *Nat Methods* **7**, 741-6 (2010).
1129 92. Toulokhonov, I., Zhang, J., Palangat, M. & Landick, R. A central role of the RNA
1130 polymerase trigger loop in active-site rearrangement during transcriptional pausing. *Mol*
1131 *Cell* **27**, 406-19 (2007).
1132 93. Yuzenkova, Y. et al. Stepwise mechanism for transcription fidelity. *BMC Biol* **8**, 54
1133 (2010).
1134 94. Diss, G. & Lehner, B. The genetic landscape of a physical interaction. *Elife* **7**(2018).
1135 95. Wang, W., Walmacq, C., Chong, J., Kashlev, M. & Wang, D. Structural basis of
1136 transcriptional stalling and bypass of abasic DNA lesion by RNA polymerase II. *Proc Natl*
1137 *Acad Sci U S A* **115**, E2538-E2545 (2018).
1138 96. Halabi, N., Rivoire, O., Leibler, S. & Ranganathan, R. Protein sectors: evolutionary units
1139 of three-dimensional structure. *Cell* **138**, 774-86 (2009).
1140 97. Rivoire, O., Reynolds, K.A. & Ranganathan, R. Evolution-Based Functional
1141 Decomposition of Proteins. *PLoS Comput Biol* **12**, e1004817 (2016).
1142 98. Salinas, V.H. & Ranganathan, R. Coevolution-based inference of amino acid interactions
1143 underlying protein function. *Elife* **7**(2018).
1144 99. Cramer, P. Structure and function of RNA polymerase II. *Adv Protein Chem* **67**, 1-42
1145 (2004).

1146 100. Walmacq, C. et al. Rpb9 subunit controls transcription fidelity by delaying NTP
1147 sequestration in RNA polymerase II. *J Biol Chem* **284**, 19601-12 (2009).
1148 101. Ziegler, L.M., Khaperskyy, D.A., Ammerman, M.L. & Ponticelli, A.S. Yeast RNA
1149 polymerase II lacking the Rpb9 subunit is impaired for interaction with transcription factor
1150 IIF. *J Biol Chem* **278**, 48950-6 (2003).
1151 102. Nouraini, S., Archambault, J. & Friesen, J.D. Rpo26p, a subunit common to yeast RNA
1152 polymerases, is essential for the assembly of RNA polymerases I and II and for the
1153 stability of the largest subunits of these enzymes. *Mol Cell Biol* **16**, 5985-96 (1996).
1154 103. Cramer, P. Organization and regulation of gene transcription. *Nature* **573**, 45-54 (2019).
1155 104. Schier, A.C. & Taatjes, D.J. Structure and mechanism of the RNA polymerase II
1156 transcription machinery. *Genes Dev* **34**, 465-488 (2020).
1157 105. Vos, S.M., Farnung, L., Linden, A., Urlaub, H. & Cramer, P. Structure of complete Pol II-
1158 DSIF-PAF-SPT6 transcription complex reveals RTF1 allosteric activation. *Nat Struct Mol*
1159 *Biol* **27**, 668-677 (2020).
1160 106. Castro, C. et al. Two proton transfers in the transition state for nucleotidyl transfer
1161 catalyzed by RNA- and DNA-dependent RNA and DNA polymerases. *Proc Natl Acad Sci*
1162 *U S A* **104**, 4267-72 (2007).
1163 107. Gregory, M.T., Gao, Y., Cui, Q. & Yang, W. Multiple deprotonation paths of the
1164 nucleophile 3'-OH in the DNA synthesis reaction. *Proc Natl Acad Sci U S A* **118**(2021).
1165 108. Belogurov, G.A. & Artsimovitch, I. The Mechanisms of Substrate Selection, Catalysis,
1166 and Translocation by the Elongating RNA Polymerase. *J Mol Biol* **431**, 3975-4006
1167 (2019).
1168 109. Jenks, M.H., O'Rourke, T.W. & Reines, D. Properties of an intergenic terminator and
1169 start site switch that regulate IMD2 transcription in yeast. *Mol Cell Biol* **28**, 3883-93
1170 (2008).
1171 110. Kuehner, J.N. & Brow, D.A. Regulation of a eukaryotic gene by GTP-dependent start site
1172 selection and transcription attenuation. *Mol Cell* **31**, 201-11 (2008).
1173 111. Qiu, C. et al. Thiolutin has complex effects in vivo but is a direct inhibitor of RNA
1174 polymerase II in vitro. *Nucleic Acids Res* **52**, 2546-2564 (2024).
1175 112. Greger, I.H. & Proudfoot, N.J. Poly(A) signals control both transcriptional termination and
1176 initiation between the tandem GAL10 and GAL7 genes of Saccharomyces cerevisiae.
1177 *Embo J* **17**, 4771-9 (1998).
1178 113. Kaplan, C.D., Holland, M.J. & Winston, F. Interaction between transcription elongation
1179 factors and mRNA 3'-end formation at the Saccharomyces cerevisiae GAL10-GAL7
1180 locus. *J Biol Chem* **280**, 913-22 (2005).
1181 114. Simchen, G., Winston, F., Styles, C.A. & Fink, G.R. Ty-mediated gene expression of the
1182 LYS2 and HIS4 genes of Saccharomyces cerevisiae is controlled by the same SPT
1183 genes. *Proc Natl Acad Sci U S A* **81**, 2431-4 (1984).
1184 115. Aguilera, A. Formamide sensitivity: a novel conditional phenotype in yeast. *Genetics*
1185 **136**, 87-91 (1994).
1186 116. Sing-Hoi Sze, C.D.K. Codon-Based Sequence Alignment for Mutation Analysis by High-
1187 Throughput Sequencing. in *2018 IEEE 8th International Conference on Computational*
1188 *Advances in Bio and Medical Sciences (ICCABS)* (2018).
1189 117. Notredame, C., Higgins, D.G. & Heringa, J. T-Coffee: A novel method for fast and
1190 accurate multiple sequence alignment. *J Mol Biol* **302**, 205-17 (2000).
1191 118. Price, M.N., Dehal, P.S. & Arkin, A.P. FastTree 2--approximately maximum-likelihood
1192 trees for large alignments. *PLoS One* **5**, e9490 (2010).

1193