Functional dissection of RNA polymerase active sites by deep mutational scanning

by

Bingbing Duan

B.S., Lanzhou University, Lanzhou, Gansu, China, 2014

M.S., Lanzhou University, Lanzhou, Gansu, China, 2017

Submitted to the Graduate Faculty of the

Dietrich School of Arts and Sciences in partial fulfillment

of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2024

UNIVERSITY OF PITTSBURGH

DIETRICH SCHOOL OF ARTS AND SCIENCES

This dissertation was presented

by

Bingbing Duan

It was defended on

January 9, 2024

and approved by

Karen Arndt, PhD, Professor, Department of Biological Sciences

Andrea Berman, PhD, Associate Professor, Department of Biological Sciences

Andrew VanDemark, PhD, Associate Professor, Department of Biological Sciences

Anne-Ruxandra Carvunis, PhD, Assistant Professor, Department of Computational and Systems Biology

Thesis Advisor/Dissertation Director: Craig Kaplan, PhD, Professor, Department of Biological Sciences

Functional dissection of RNA polymerase active sites by deep mutational scanning

Bingbing Duan, PhD

University of Pittsburgh, 2024

Transcription in eukaryotes is carried out by three RNA polymerases (Pol), Pol I, II, and III, which are structurally conserved though they have evolved to have their own regulation and produce different classes of transcripts. At the heart of these RNA polymerases is an ultra-conserved active site domain, the trigger loop (TL), coordinating transcription speed and fidelity by critical conformational changes impacting all three steps of nucleotide addition cycle (NAC) in transcription elongation, substrate selection, catalysis, and translocation. Previous genetic and biochemical studies have shown that substitutions of TL residues disturb its balance and then alter its function. Additionally, studies from our lab have observed different types of residue-residue interactions in Pol II TL, implying the TL's function is facilitated by residue interaction networks within and around it. Furthermore, identical mutations in a residue conserved between yeast Pol I and Pol II TLs yielded opposite biochemical phenotypes, implying even functions of conserved residues are shaped by individually evolved residue interactions in enzymatic contexts (epistasis). However, the specific mechanisms by which the TL is regulated and how it communicates with the rest of the enzyme remain unclear. Through analysis of over 15,000 alleles representing single mutants, a subset of double mutants, and evolutionarily observed TL haplotypes by deep mutational scanning, I identified intricate pairwise and higher-order epistatic interaction networks controlling TL function. Substituting residues creates allele-specific networks and propagates epistatic effects across the Pol II active site. Additionally, the interaction landscape further distinguishes alleles with similar growth phenotypes, suggesting increased resolution over the previously reported single mutant phenotypic landscape. Furthermore, we distinguished intricated layers of higher-order epistatic interaction networks within TL haplotypes and TL residues with distinct classes of epistatic patterns in affecting these higher-order interactions. Finally, co-evolutionary analyses reveal groups of co-evolving residues across Pol II converge onto the active site, where evolutionary constraints interface with pervasive epistasis. Our studies provide a powerful system to understand the plasticity of RNA polymerase mechanism and evolution and provide the first example of pervasive epistatic landscape in a highly conserved and constrained domain within an essential enzyme.

Table of Contents

List of Tables

List of Figures

Preface

I would like to express my deepest appreciation to my advisor, Dr. Craig Kaplan. Dr. Kaplan has been instrumental in my academic journey, providing guidance and support in numerous ways. He has dedicated a tremendous amount of time to help me improve in every aspect, including but not limited to communication skills, designing and interpreting experiments, writing and presentation. He always provides me invaluable insight from a higher level that largely extends my thinking and understanding towards things. In addition to his scientific guidance, Dr. Kaplan has offered important life advice, like maintain a positive mindset during stressful time. He has also shared recommendations for hiking in national parks and a lot of food, enriching my overall experience. Above all, I appreciate Dr. Kaplan's high standards and requirements, which has been challenging me to improve continuously. At this point, I realize that these challenging moments are where I experienced the most growth. I cannot express my gratitude enough for all that Dr. Kaplan has taught me. His mentorship will continue to bring positivity into my life.

I extend my deepest appreciation to my committee members, Dr. Karen Arndt, Dr. Andrea Berman, Dr. Andrew VanDemark, and Dr. Anne-Ruxandra Carvunis for all the feedback and discussions on my project, their valuable insight on helping me figuring out my career path, whether in academia or industry, and their kind support has been a source of warmth during stressful time. I would like to sincerely acknowledge Dr. VanDemark for his effort in explaining the comprehensive exam rules to me, and Dr. Carvunis for the insightful discussions on data analysis. I also would like to specifically appreciate Dr. Arndt and Dr. Berman for their continuous encouragement and support.

I also want to express my appreciation for the discussions I've had during KASHCKML lab meetings, which include the Arndt lab, the Hainer lab, and the Lee lab. Additionally, I want to acknowledge the journal club I had together with the Arndt lab. I would like to express my special thanks to Dr. Sarah Hainer for her kind support, and for sharing their Qubit with me during earlier days, as well as the discussion we had regarding postdocs and industry suitability.

I would like to thank Dr. Hays Rye and Dr. Jim Hu from Texas A&M University and their lab members for their guidance during my rotations. I would like to thank Dr. Rodolpho Aramayo (Texas A&M) for his exceptional teaching in Bioinformatics. I would also like to thank Dr. Karen Arndt and Dr. Jon Boyle for their guidance in the paper reading class. This is the first course in this PhD journey which helped me overcome my difficulties in expressing my opinion, and Dr. Arndt has always been encouraging me to communicate since then.

I am grateful to our collaborators, Dr. Sing-Hoi Sze (Texas A&M University), Dr. Steven Lockless (Texas A&M University), Dr. Bercem Dutagaci (University of California Merced) and Dr. Michael Feig (Michigan State University). Dr. Sing-Hoi Sze, in particular, for his help on data analysis.

My heartfelt thanks go out to everyone I've met in the Kaplan lab. Dr. Chenxi Qiu, Dr. Tingting Zhao, and Dr. Yunye Zhu have guided me not only in my study but also in navigating life's challenges. As senior students in the lab when I joined, they provided excellent models of what I could achieve after several years, serving as a tremendous source of motivation and inspiration. Payal Arora, I am glad I have her by my side to navigate through every step, from lab relocation, comprehensive exam to the challenges posed by the pandemic. I would like to extend my appreciation to Pratik Basnet and James Kuldell for the daily support and help. I would also like to thank the others lab member, Dr. Anandhakumar Jayamani, Evelyn Callaway, Shrabani

Basu, Joanna Yao and Lea Lubenski for their help. Special thanks to all the undergraduate students who have contributed to my project, Gabrielle Fortier, Douglas Walko, Tiffany Hom, Ethan Gretzula, Dharrnhi Purushothaman, and Nikitha Vasudevan.

I would like to thank all my friends for always being there for me. Particularly, Yongjin Liu, Xueting Tang, Zhifeng Xiao, Jirapat Thongchon, Haoyun Peng, and Jixing Lyu for their continuous help and support. Additionally, I'm grateful for the remote support provided by Panpan Zhang, Sipan Zhang, and Keli Tian.

I express my deepest gratitude to my family, and most special and important, my grandparents, for all the unconditional love and support. Without them, I would not have reached this point in my life.

1.0 Introduction

DNA-dependent RNA transcription is an essential process for all domains of life, facilitated by multi-subunit RNA polymerases (msRNAPs) (Cramer, 2002; Roeder & Rutter, 1969). Bacteria and archaea employ a single msRNAP to transcribe their entire genomes (G. Zhang et al., 1999), while in most eukaryotes, transcription is executed by three msRNAPs: Pol I, Pol II, and Pol III, each of which is specialized to a gene subset and has a varying subunit number (Cramer, Bushnell, & Kornberg, 2001; Fernandez-Tornero et al., 2013; Hoffmann et al., 2015). All msRNAPs are conserved in structure and function, especially their active centers (Werner & Grohmann, 2011). Within the active center of each msRNAP, a mobile domain, trigger loop (TL) promotes transcription in a fast but accurate manner by switching between different conformations (Kaplan, 2013; D. Wang, Bushnell, Westover, Kaplan, & Kornberg, 2006). An open question in the field is what controls the TL motion. Moreover, considering the strikingly conservation level of TLs from all msRNAPs in structure, function and mechanism, are TLs switchable among different msRNAPs? Our lab has shown several genetic residue-residue interactions within the Pol II TL (Kaplan, Jin, Zhang, & Belyanin, 2012; Qiu et al., 2016), suggesting a functional interaction network of residues controlling TL activity and potentially genetically separable steps in Pol II TL function. Additionally, identical mutations have opposite biochemical phenotypes when introduced into yeast Pol I and Pol II (Viktorovskaya et al., 2013), implying TL domain is context dependent (Qiu et al., 2016), likely because residue interactions between the TL and its specific enzymatic context diverged with the evolution of msRNAPs. The extent of these interaction networks and how they control TL function and evolution remain to be determined. This thesis comprehensively determines Pol II TL residue interaction networks and their impact on fitness and transcription related phenotypes in budding yeast *Saccharomyces cerevisiae* by an extended deep mutational scanning approach. Chapter 2 dissects TL pairwise residue interaction networks using double mutants. Chapter 3 determines higher-order interactions with many mutant combinations (haplotypes).

I will review four aspects in the introduction. Firstly, I will describe the conservation of all msRNAPs in structure and function in Archaea, Bacteria, and Eukaryotes. Secondly, using yeast Pol II as an example, I will review the current understanding of the structure, function and mechanism of the RNA polymerase active site. Thirdly, I will discuss the impacts of residue epistasis on protein function and evolution. Finally, I will review the usage of deep mutation scanning in understanding mechanisms of protein function and evolution.

1.1 Conserved multisubunit RNA polymerases execute transcription in all domains of life

In this section, I summarize the subunits, architecture and basic functions of multisubunit RNA polymerases in different domains of life. The key point is though the number of subunits are different, the architecture of all msRNAPs is conserved across all domains of life, which serves as the basis of their functional conservation (Cramer, 2019a; Kaplan, 2013; Kramm, Endesfelder, & Grohmann, 2019; Sauguet, 2019; Werner & Grohmann, 2011) (**Figure 1 and Table 1**).

The first step in gene expression, transcription, is carried out by msRNAPs. MsRNAPs from all domains of life consist of varying numbers of subunits, yet these subunits are evolutionarily related (Allison, Moyle, Shales, & Ingles, 1985; Werner & Grohmann, 2011; Wu et al., 2012). The bacterial msRNAP, as illustrated in **Figure 1A**, consists of five subunits (**Table 1**) (Vassylyev et al., 2007). Archaeal msRNAP has 11-13 subunits (**Figure 1B and Table 1**) (Korkhin et al., 2009). Five subunits of archaeal msRNAP are conserved with bacterial subunits. The additional 6 subunits are homologous to eukaryotic msRNAPs. One subunit, Rpo13, is specific to archaeal msRNAPs. Notably, Rpo13, as well as Rpo8 are found in some but not all archaeal msRNAPs. Bacteria and Archaea use one msRNAP to transcribe their entire genomes, while Eukaryotic cells have evolved at least three RNA polymerases (Pol I, II, III) for transcribing different classes of RNAs in their genomes (**Figure 1C-E**). Pol II has 12 subunits (**Figure 1C**) and transcribes protein-coding genes and many non-coding RNAs (Armache, Kettenberger, & Cramer, 2003; Bushnell & Kornberg, 2003; Cramer, 2002). Pol II is the smallest enzyme among the three eukaryotic polymerases, but it is most extensively studied due to the close connection between its transcripts and cell development (Kaplan, 2013). The largest eukaryotic Pol, Pol III, has 17 subunits (**Figure 1E**) and synthesizes specific short non-coding RNAs, including all transfer RNAs (tRNAs), ribosomal 5S rRNA, spliceosomal U6 snRNA, and some other small structured RNAs (Hoffmann et al., 2015). Pol I has 14 subunits (**Figure 1D**) and solely transcribes the large ribosomal RNA (rRNA) precursor gene, producing 35S pre-rRNA, which is processed into 28S, 18S, and 5.8S rRNA co- and post-transcriptionally. With only one target gene, which is the fewest of the three eukaryotic RNAPs, Pol I transcripts account for over 60% of total transcripts in growing cells (Khatter, Vorlander, & Muller, 2017; Sadian et al., 2019). Despite their non-overlapping targets in transcription, the three eukaryotic RNAPs have ten conserved subunits, with five subunits being shared among the three complexes (**Table 1**) (Cramer, 2002; Khatter et al., 2017; Vannini & Cramer, 2012). Plants have two additional msRNAPs which are specialized versions of Pol II, the 12 subunit Pol IV (Onodera et al., 2005) and Pol V (Wierzbicki, Haag, & Pikaard, 2008). They exhibit a noteworthy similarity with Pol II, with seven out of ten core subunits shared with Pol II (**Table 1**) (Ream et al., 2009). Pol IV and Pol V orchestrate non-coding RNA-mediated gene silencing processes in plants (Ahlquist, 2002; Y. Huang et al., 2015).

Although subunit number is different in different msRNAPs, the general structure and function of the subunits are conserved. These subunits can be separated into three groups: those forming the active site and facilitating catalysis, those involved in assembly, and those serving auxiliary functions. Subunits for catalysis and assembly, also known as the core subunits, are the minimal requirement for all functional msRNAPs like bacterial msRNAP. Bacterial RNAP's largest β' and β subunits harbor the active site responsible for nucleotide addition during transcription. Two α subunits play crucial roles in the assembly of the bacterial msRNAP complex together with the ω subunit (Werner & Grohmann, 2011; G. Zhang et al., 1999). The ω subunit is an auxiliary subunit, known to co-purify with the core subunits like β', β and α subunits (Doherty, Fogg, Wilkinson, & Lewis, 2010; Gentry & Burgess, 1990; Pero, Nelson, & Fox, 1975). Auxiliary subunits are not required for minimal msRNAP function but can be associated with DNA or transcription factor binding such as Pol II Rpb4 and Rpb7 in higher domains of life. Homologs of bacterial RNAP subunits can be found in all msRNAPs (**Table 1**). With Pol II as an example, Rpb1 and Rpb2, homologous to the bacterial β' and β subunits, form the central mass of the enzyme that is responsible for catalysis (Cramer et al., 2001; Gnatt, Cramer, Fu, Bushnell, & Kornberg, 2001; Vannini & Cramer, 2012; D. Wang et al., 2006; Werner & Grohmann, 2011). Rpb3 and Rpb11, homologs of bacterial α subunits, form a stable assembly platform with Rpb10 and Rpb12. Pol II auxiliary subunits include Rpb6, homologous to bacterial ω subunit, which also functions in assembly (Cramer, 2002; Nouraini, Archambault, & Friesen, 1996; Vannini & Cramer, 2012). The stalk subunits Rpb4 and Rpb7 (Armache et al., 2003) function in guiding away transcript from the elongating RNAP (Hirtreiter, Grohmann, & Werner, 2010) and interacting with transcription factors (Schier & Taatjes, 2020). Rpb5 helps in DNA melting (Kostrewa et al., 2009), and Rpb9 for TFIIS binding and influencing transcription fidelity (Kaster, Knippa, Kaplan, & Peterson, 2016; Walmacq et al., 2009; Ziegler, Khaperskyy, Ammerman, & Ponticelli, 2003). These 12 subunits form a “crab-claw” framework of Pol II. This architecture, similar to the bacterial msRNAP, is also shared with archaeal and other eukaryotic msRNAPs, indicating their architectures are conserved.

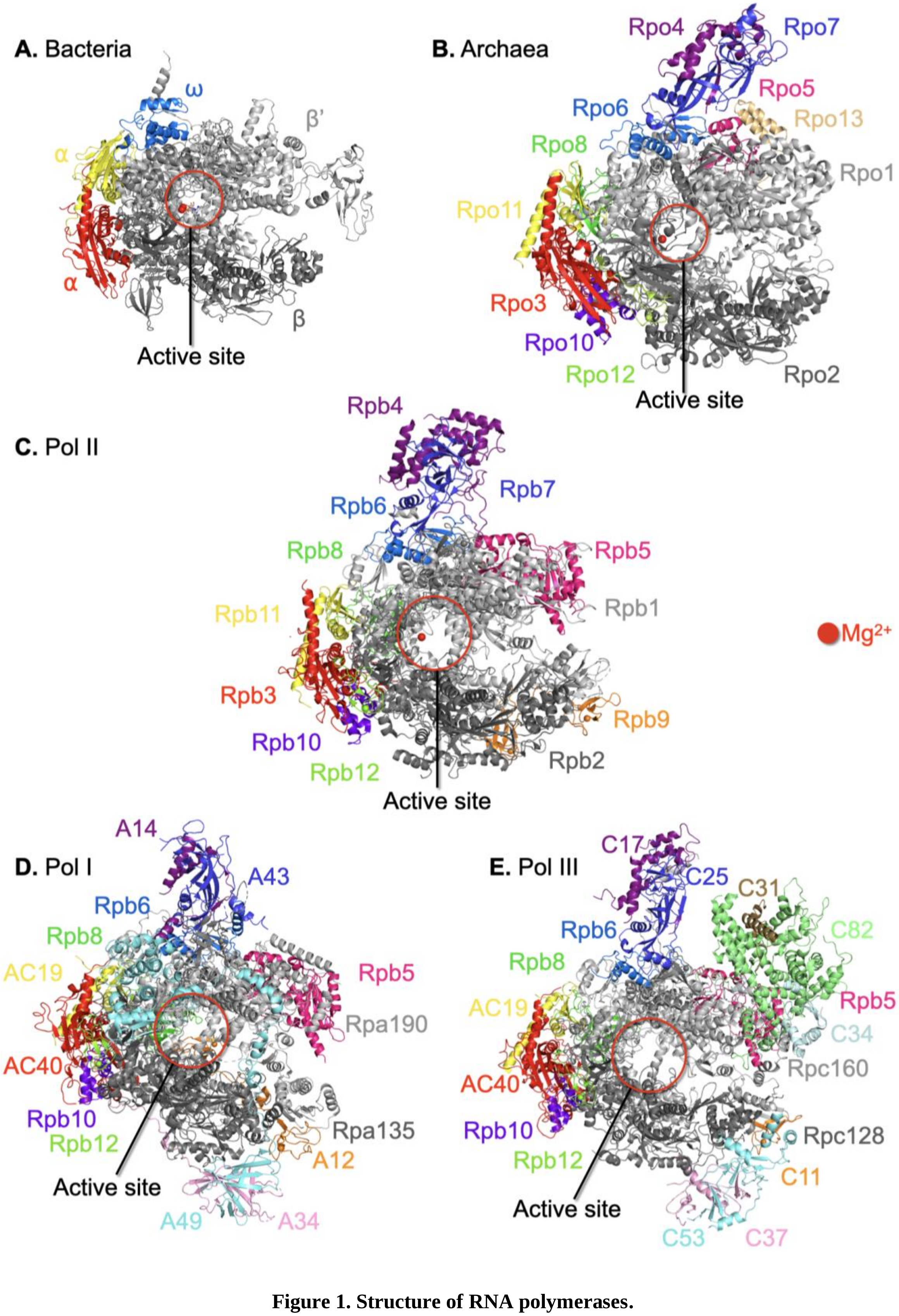


Figure 1. Structure of RNA polymerases.

(**A**). *T. thermophilus* (Bacteria) RNAP, visualized with Pymol (PDB: 2O5J) (Vassylyev et al., 2007). (**B**). *S. shibatae* (Archaea) RNAP (PDB: 2WAQ) (Korkhin et al., 2009). (**C**). *S. cerevisiae* (Eukaryote) Pol II (PDB: 5C4X) (Barnes et al., 2015). (**D**). *S. cerevisiae* (Eukaryote) Pol I (PDB: 6RWE) (Sadian et al., 2019). (**E**). *S. cerevisiae* (Eukaryote) Pol III (PDB 5FJ8) (Hoffmann et al., 2015). Conserved subunits are labeled with the same color across all RNAPs.

Table 1. Conserved RNA polymerases subunits and their functions in three domains of life

	Bacteria	Archaea	Eukaryote					Function in Pol II or Pol I/Pol III
			Pol II	Pol I	Pol III	Pol IV (Plant)	Pol V (Plant)	
Core subunits	β'	Rpo1	Rpb1	A190	C160	NRPD1	NRPE1	Catalysis
	β	Rpo2	Rpb2	A135	C128	NRPD/E2	NRPD/E2	
	α	Rpo3	Rpb3	AC40	AC40	Rpb3	Rpb3	RNAP assembly
	α	Rpo11	Rpb11	AC19	AC19	Rpb11	Rpb11	
	ω	Rpo6	Rpb6	Rpb6	Rpb6	Rpb6	Rpb6	TFIIH binding
Conserved subunits in Archaea and Eukaryote		Rpo5	Rpb5	Rpb5	Rpb5	Rpb5	NRPE5	DNA melting
		Rpo8	Rpb8	Rpb8	Rpb8	Rpb8	Rpb8	
		Rpo10	Rpb10	Rpb10	Rpb10	Rpb10	Rpb10	RNAP assembly
		Rpo12	Rpb12	Rpb12	Rpb12	Rpb12	Rpb12	
		Rpo4	Rpb4	A14	C17	NRPD/E4	NRPD/E4	“Stalk” domain
		Rpo7	Rpb7	A43	C25	NRPD7	NRPE7	
Archaea subunit		Rpo13						
Eukaryote exclusive subunit			Rpb9	A12	C11	NRPD9b	Rpb9	TFIIF binding. Elongation speed, fidelity and processivity
Exclusive subunits in eukaryotic Pol I and Pol III				A34.5	C37			Elongation speed, fidelity and processivity
				A49	C53			
					C82			

Eukaryotic Pol III-specific subunits	C34		PIC formation and DNA melting
	C31		

1.2 The structure, function, and mechanism of the Pol II active site

In this section, I'll start with a brief overview of the structure and function of the Pol II active site. Then I will describe the functions and mechanisms of two critical domains within this site, the TL and the BH, in facilitating transcription. Finally, I'll introduce how the functions of the TL and BH are regulated through intricate interaction networks.

1.2.1 The fundamental structure and function of the RNA polymerase II active site

In the center of the 'crab claw' architecture of msRNAPs is the active site, which is highly conserved in structure and function and is where RNA synthesis occurs in transcription, highly conserved in structure and function. Using *Saccharomyces cerevisiae* Pol II as an example, I will summarize its structure and function. In the transcribing Pol II, the downstream DNA extends along the major DNA-binding channel to the active center and the double strand DNA is unwound to form a transcription bubble. RNA is synthesized and remain bound by the enzyme in the form of RNA-DNA hybrid (**Figure 2**) (Cramer et al., 2001; Gnatt et al., 2001). Two critical domains in the active site, the bridge helix (BH) and the trigger loop (TL) have been proposed to dynamically facilitate all critical steps of transcription elongation (Kaplan, 2013; Vassylyev et al., 2002; D. Wang et al., 2006). Mg^{2+} present in the active site is required for catalysis (**Figure 1**) (G. Lin et al., 2023; Zaychikov et al., 1996). Other domains in the active site, including Rpb1 470-485, α-46 and α-47 helices, funnel helices α-20 and α-21, and Rpb2 link domain 757-776, participate in transcription likely through interaction with the TL and the BH (**Figure 2**) (Da et al., 2016; Kaplan, 2013; Kaster et al., 2016; Qiu et al., 2016; D. Wang et al., 2006; Weinzierl, 2010). The detailed roles of the TL and BH will be discussed in following sections.

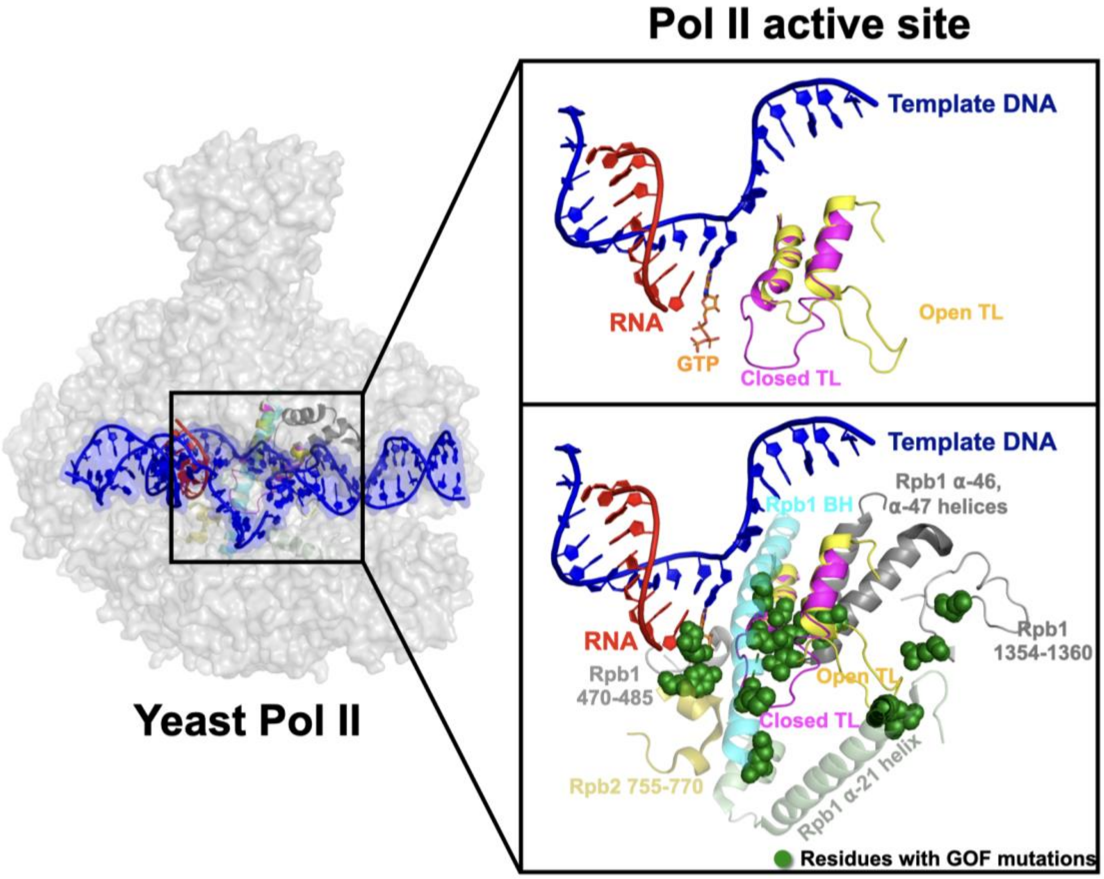


Figure 2. Pol II active site.

The Pol II active site is embedded in the center of a 12-subunit complex (left panel). Pol II functions are supported by distinct TL conformational states. An open TL (PDB: 5C4X) (Barnes et al., 2015) and closed TL (PDB: 2E2H) (D. Wang et al., 2006) conformations are shown in the right top panel. GOF mutations have been identified in the TL and its proximal domains (right bottom panel), suggesting TL mobility and function may be impacted by adjacent residues.

1.2.2 The conserved mechanism of transcription elongation

1.2.2.1 The TL and the BH facilitate fast and accurate transcription

RNA synthesis in Pol II active site is an iterative process of nucleotide addition cycle (NAC), involving nucleotide selection, catalysis and translocation (Kaplan, 2013). A strikingly conserved domain in the active site, the TL, participates in every step of NAC, promoting transcription in a fast and accurate manner (Sauguet, 2019; Scherrer, 2018; Schier & Taatjes, 2020). The elongation rate described in prior studies, depending on the types of DNA templates and the species, ranges from 1 to 5kb per minute (Mason & Struhl, 2005). The transcription error rate, influenced by error types and species, typically is around 1 in 10^6 bases (Gout et al., 2017; Gout, Thomas, Smith, Okamoto, & Lynch, 2013; Imashimizu, Oshima, Lubkowska, & Kashlev, 2013; Reid-Bayliss & Loeb, 2017).

The rapid and precise process is associated with the flexible and mobile nature of the TL, as has been observed in structural and biophysical/biochemical studies (Fouqueau, Zeller, Cheung, Cramer, & Thomm, 2013; Kaplan, 2013; Larson et al., 2012; Mazumder, Lin, Kapanidis, & Ebright, 2020; B. Wang, Predeus, Burton, & Feig, 2013; D. Wang et al., 2006). The TL residues (Rpb1 1076-1106) can be divided into three regions: an N-terminal helix (Rpb1 1076-1085) containing the nucleotide interaction region (NIR, Rpb1 1078-1085), a loop region (Rpb1 1086-1096) containing the TL tip (Rpb1 1090-1096), characterized as a random-coil region, and the C-terminal helix (Rpb1 1097-1106).

Substrate addition involves multiple TL conformations to close on/recognize matched substrates and then subsequently promote catalysis. At the beginning of each NAC, Pol II is at the post-translocation state and the TL is in a catalysis-disfavoring, “open” state, with the TL tip region (Rpb1 1090-1096) being away from the substrate site, where one template DNA base is accessible for substrate binding and the end of the growing RNA chain is positioned for substrate addition in anticipation of upcoming substrate binding. A new substrate entering the active site potentially induces a TL conformational change from the "open" state to a partially "closed" state (Cheung, Sainsbury, & Cramer, 2011). In this process, residues in the TL nucleotide interaction region (NIR, Rpb1 1078-1085) interact with the upcoming substrate and facilitate the discrimination of a correct NTP over dNTPs or non-matched NTPs. In detail, L1081 (S. cerevisiae residue numbers are used here and in the following description) forms hydrophobic interaction with the substrate (D. Wang et al., 2006). Q1078, N1082 and a non-TL Rpb1 residue N479, form an interaction network to recognize the 2'- and 3'OH of the substrate (Belogurov & Artsimovitch, 2019; Kaplan et al., 2012; Svetlov, Vassylyev, & Artsimovitch, 2004; Westover, Bushnell, & Kornberg, 2004). Matched substrate binding has been proposed to induce complete TL closure, allowing capture of the correct NTP, promoting the formation of a phosphodiester bond, and subsequent release of pyrophosphate (Fong et al., 2014; Kaplan, 2013; G. Lin et al., 2023; Malinen et al., 2012; B. Wang et al., 2013; D. Wang et al., 2006; L. Xu et al., 2014). The transition from random coil to helix upon TL closing may be promoted by substrate interactions and in turn may promote catalysis (Mejia, Nudler, & Bustamante, 2015; B. Wang et al., 2013; Windgassen et al., 2014). A TL residue, H1085, directly interacts with the β-phosphate of the substrate and has been proposed to promote catalysis by being a chemical catalyst (Castro et al., 2009; D. Wang et al., 2006) or a positional catalyst (Mishanina, Palo, Nayak, Mooney, & Landick, 2017). The release of pyrophosphate enables the TL to switch back from the "closed" to the "open" state. This transition has been proposed to support polymerase translocation to the next open nucleotide of the template DNA, enabling the subsequent NAC (Da, Wang, & Huang, 2012; B. Liu, Zuo, & Steitz, 2016; Seibold et al., 2010) (**Figure 2**). Moreover, additional TL conformations can be identified during other Pol II states such as pausing and backtracking (Cheung & Cramer, 2011; Mosaei & Zenkin, 2021; D. Wang et al., 2009; J. Zhang, Palangat, & Landick, 2010). It is worth pointing out that the proposed mechanism of TL residues participating in nucleotide addition is based on the structural observations and molecular dynamics simulations. Time resolved structural information is needed in the future to experimentally determine the exact Pol II mechanism.

The TL conformational changes have been proposed to be coupled with its adjacent domain, the highly conserved BH (Rpb1 815-848) (Ahearn, Bartolomei, West, Cisek, & Corden, 1987; Vassylyev et al., 2002). The BH was observed as a straight helix in most bacterial and all archaeal and eukaryotic msRNAP structures (Cramer et al., 2001; Gnatt et al., 2001; Kaplan, 2013; X. Liu, Bushnell, & Kornberg, 2013; D. Wang et al., 2006), but in a kinked conformation in *Thermus thermophilus* (Bacteria) RNAP structures (Kaplan & Kornberg, 2008; Kireeva et al., 2012; Vassylyev et al., 2002; G. Zhang et al., 1999), implying its potential flexibility. Though the kinked BH has never been detected in eukaryotes, the flexibility of eukaryotic BH has been supported by molecular dynamic simulations and has been proposed to promote msRNAP translocation in the NAC (Kaplan & Kornberg, 2008; Silva et al., 2014; Tan, Wiesler, Trzaska, Carney, & Weinzierl, 2008; Weinzierl, 2010). The dynamic correspondence between BH conformational changes and TL conformational changes, along with their interactions with other domains in the active site to define proper transcription, remains to be fully elucidated.

1.2.2.2 Mutations identified in TL and TL surrounding domains affect transcription

TL conformational changes are associated with catalysis during transcription. Mutations in the TL and other TL-proximal active site residues can cause specific phenotypes consistent with altered catalysis, leading to either defective elongation rate or fidelity by many msRNAPs (Kaplan et al., 2012; Kaplan & Kornberg, 2008; Kaplan, Larsson, & Kornberg, 2008; Kireeva et al., 2008;

Malagon et al., 2006; Qiu et al., 2016; Tan et al., 2008). With yeast as an example, mutations in the TL NIR disrupt interactions between the TL and substrates, leading to defective catalysis and a reduced elongation rate *in vitro* (Loss of function, LOF) (Kaplan et al., 2012; Kaplan et al., 2008; Kireeva et al., 2008; Nayak, Voss, Windgassen, Mooney, & Landick, 2013; Qiu et al., 2016; Windgassen et al., 2014). Some mutations in the TL C-terminal region break interactions stabilizing the inactive "open" state of TL and shift the TL towards the active "closed" state, leading to enhanced catalysis and increased elongation rate but compromised transcription fidelity (Gain of function, GOF) (Fouqueau et al., 2013; Kaplan et al., 2012; Kaplan et al., 2008; Kireeva et al., 2008; Malagon et al., 2006; Nayak et al., 2013; Qiu et al., 2016; Windgassen et al., 2014).

Interestingly, TL GOF and LOF mutants exhibit specific, distinct conditional growth phenotypes, as briefly summarized in **Table 2** (Aguilera, 1994; Braberg et al., 2013; Cui, Jin, Vutukuru, & Kaplan, 2016; Greger & Proudfoot, 1998; Kaplan, Holland, & Winston, 2005; Kaplan et al., 2012; Malik, Qiu, Snavely, & Kaplan, 2017; Qiu et al., 2016; Simchen, Winston, Styles, & Fink, 1984). These distinct phenotypic patterns of TL GOF or LOF mutants are consistent with their distinct defects in transcription. For example, most GOF mutants are sensitive to Mycophenolic acid (MPA) while LOF are resistant to it (Kaplan et al., 2012; Qiu et al., 2016). The MPA sensitive phenotype in transcription factors is due to altered transcription initiation at the *IMD2* promoter, which is regulated by multiple start sites and its proper expression is essential for MPA resistance (Hyle, Shaw, & Reines, 2003; Jenks, O'Rourke, & Reines, 2008; Kuehner & Brow, 2008). Pol II GOF fail to induce *IMD2* expression under MPA conditions due to abnormal transcription start site selection (Kaplan et al., 2012; Malik et al., 2017).

GOF and LOF mutations have also been found in nearly all TL-proximal domains, including the BH GOF T834P and LOF T834A, funnel helix α-21 GOF S713P, and Rpb2 GOF

Y769F (Braberg et al., 2013; Leng et al., 2020; Qiu et al., 2016). These mutations in TL-proximal domains share similar phenotypes with TL GOF or LOF mutants. The shared phenotypes suggest TL-proximal domains could participate in transcription by interacting with the TL (Braberg et al., 2013; Kaster et al., 2016; Leng et al., 2020; Taatjes, 2020). However, a key question about these residues is if and how they work with the TL and do they simply shift the balance of existing states or create an altered active site where perturbations extend to alterations across the TL (new or different conformations for example). It remains unclear whether these mutations go beyond putatively altering the balance of conformational states observed in the WT enzyme. Additionally, the specific TL residues through which they communicate to ensure proper transcription remain unidentified. The observed phenotypic patterns establish connections between mutant catalytic defects and conditional growth defects. This connection enables the prediction of catalytic defects through growth patterns without performing biochemical studies, paving the way for understanding of TL residue mechanisms in large scale. Qiu et al developed the high throughput phenotyping system with deep mutational scanning (Qiu et al., 2016), allowing these questions to be targeted, which will be discussed in Chapter 2.

Table 2. Growth phenotypes of mutants are linked with their catalytic defects

Stress conditions	Rationale	**WT** phenotype	**GOF** phenotype	**LOF** phenotype
SC-Leu+MPA	MPA depletes GTP levels in yeast. To generate GTP, *IMD2* expression is required, but its promoter has multiple transcription start site (TSS),	**Resistant** (Can use both upstream and downstream TSS)	**Sensitive** (Can only use upstream TSS)	**Resistant** (Can only use downstream TSS)

	only downstream TSS can successfully express *IMD2*.			
SC-Leu+Mn	Mn compromises transcription fidelity	**Resistant**	**Sensitive** (Fidelity has already been compromised)	**Resistant** (Fidelity is not compromised)
YPrafGal	The termination and RNA processing of *gal10Δ56* is problematic, which interfere with the initiation of the downstream *GAL7*.	**Sensitive**	Some GOF mutants are **resistant**	Most LOF mutants are **resistant**
SC-Lys	The *Ty* transposable element insertion in *lys2-128Ə* stops normal expression of *LYS2*	**Lys^-** (Lys auxotroph)	**Lys^+** (Lys prototroph) (Can utilize the normally silent promoter within the *Ty* insertion)	**Lys^-** (Lys auxotroph)
SC-Leu+Formamide	Formamide destabilizes hydrogen bonds of protein and RNA	**Resistant**	**Resistant**	**Slightly sensitive**

1.2.2.3 Residue interactions within and surround the TL control its dynamics and function

Distinct residue interactions are observed within and around the TL, suggesting that TL function is likely impacted by residue interaction networks. These networks can be categorized into four levels.

First, TL residues form interaction networks within the TL. Previous genetic studies in our lab have identified enhancement interactions (where the double mutant phenotype is worse than both singles) and lack of enhancement, epistatic interactions (where the double mutant phenotype is similar to either single mutant), indicating functional dependency of TL residues (Kaplan et al., 2012; Qiu et al., 2016; Qiu & Kaplan, 2019).

Second, residue interactions have been identified between the TL and other active site domains. Evidence from different species has shown that many active site domains (BH, funnel helix etc.) directly impact the TL function (Qiu et al., 2016; Silva et al., 2014; B. Wang et al., 2013; Weinzierl, 2010). For instance, structural observations indicate hydrophobic residues in two TL helices form a bundle with BH and two α helices (α-46 and α-47 helices) in the active site. This five-helices bundle is universally conserved in msRNAPs and is proposed to stabilize the inactive "open" state of TL (Barnes et al., 2015). Consistently, many substitutions in these hydrophobic residues result in GOF mutations (i.e. A1076, L1101) (Qiu et al., 2016).

Third, Pol II residues outside of the active site allosterically control the functions of Pol II active site, suggesting potential allosteric pathways. For example, the deletion of a small Pol II subunit, Rpb9, which does not directly contact Pol II active site, confers GOF phenotypes (Jenks et al., 2008; Kaster et al., 2016). *RPB9* mutation suppresses severe growth defects caused by a mutation in funnel helix α-21 (Kaster et al., 2016; Koyama, Ueda, Ito, & Sekimizu, 2010), suggesting a potential allosteric pathway that the funnel helix interacts with both open TL and Rpb9 to stabilize the open state TL, which has important roles in maintaining transcription fidelity.

Fourth, this kind of allosteric effect in regulating the active site may from outside of the Pol II enzyme, with transcription factors involved. For example, the Pol II elongation factor TFIIS can directly insert into the Pol II active site, functioning in rescuing arrested Pol II and also in proofreading, which requires TL being in the open state (Kettenberger, Armache, & Cramer, 2004; D. Wang et al., 2009). Notably, certain yeast transcription elongation factors, such as Spt5 (Swanson, Malone, & Winston, 1991; Winston, Chaleff, Valent, & Fink, 1984), Spt6 (Close et al., 2011; Hartzog, Wada, Handa, & Winston, 1998), Elf1 (Prather, Krogan, Emili, Greenblatt, & Winston, 2005), and Paf1C (Wade et al., 1996; Y. Xu et al., 2017), promote elongation without directly interacting with the Pol II active site in yeast. Indirect evidence from genetics studies illustrate mutants in some of these elongation factors decrease elongation rate (Archambault, Lacroute, Ruet, & Friesen, 1992; Malagon et al., 2006; Mayer et al., 2012; Prather et al., 2005; Riles, Shaw, Johnston, & Reines, 2004), implying potential communication between these elongation factors and the TL. The arising question is how these yeast transcription factors regulate the active site during elongation without direct contact with the domains. Residue interaction networks may serve as allosteric pathways for the active site being regulated. Studies on Paf1C in mammals suggest possible allosteric mechanism through BH interaction, but this specific elongation factor interaction is replaced by Rpb2-Rpb1 interactions in yeast (Chen et al., 2023; Vos, Farnung, Linden, Urlaub, & Cramer, 2020). We will design experiments to detect the potential interactions with yeast elongation factors in Chapter 4.

In summary, considering that TL is sensitive to minor alterations, TL makes multiple interactions in various states, and TL’s function is potentially controlled by residue interaction networks from the active site to transcription factors, it is crucial to comprehensively detect these residue interaction networks within and surround the TL. In Chapter 2, I will describe a high-throughput screening system to dissect these residue interactions. In Chapter 4, I will describe a potential experiment for us to probe how TL is regulated by the transcription elongation factors.

1.3 Residue epistasis impacts protein function and evolution

1.3.1 Overview of epistasis

In this dissertation, genetic interactions are used to probe mechanisms. A key concept in understanding genetic interactions is determining when the actions of two mutations are independent of each other or functionally dependent. Functional dependence can link function of residues together into a network. Such networks can underlie how proteins work. In this section I will describe the concept of epistasis, the prevalence and strength of epistasis in proteins, and the divergent residue epistasis networks in three highly related RNA polymerases.

The effect of a mutation is not only dependent on the mutation's amino acid, but also on the genetic background to which it is introduced. The interaction between the mutation and the genetic background is termed as epistasis. The concept was first used in 1909 to describe how one genetic variant could mask the effect of the other (Squire, 1909). Later, in 1919 epistasis was defined as the deviation from the expected combination of two loci's effects (Fisher, 1919). The term epistasis has been developed to various related meanings later, but the most common used one is to describe the deviation from the expected effect when combining mutations. A simple model has been developed, termed the "multiplicative" or "log additive" model, where the expected fitness effect of combined mutations should be equal to the product of constituent mutations' effects, thus predicted fitness effects are multiplicative, or the sum of the logarithm of the individual fitnesses. Deviation from the log additive model when combining mutations is considered epistasis. (Domingo, Baeza-Centurion, & Lehner, 2019; Mani, St Onge, Hartman, Giaever, & Roth, 2008; Phillips, 1998, 2008).

Epistatic interactions can be classified based on various criteria (Poelwijk, Kiviet, Weinreich, & Tans, 2007; Starr & Thornton, 2016). These include whether the outcome is better (positive epistasis) or worse (negative epistasis) than expected from the log additive model, whether the interaction is specific to a particular mutation (specific epistasis) or applicable to a range of mutations (nonspecific epistasis), the number of mutations involved (pairwise epistasis for two mutations or higher-order epistasis for many mutations), and whether the strength of mutation effect changes (epistasis, also termed as magnitude epistasis) or the direction of mutation effect changes (sign epistasis).

Epistatic interactions among residues influence protein function and evolution (Cisneros et al., 2023; Karageorgi et al., 2019; Pinney et al., 2021). To understand protein mechanisms, unravel their evolutionary history, and potentially predict the direction of future evolution by predicting the phenotype of upcoming mutations, it is crucial to comprehensively quantify the prevalence and magnitude of epistasis within proteins.

1.3.2 General epistasis prevalence and magnitude

Epistasis between genes can be employed to determine the hierarchical order of genes in a pathway (Goodwin & Ellis, 2002; Sternberg & Horvitz, 1989; Thomas, Birnby, & Vowels, 1993). This involves utilizing double mutants of genes within a pathway, especially regulatory switch pathways with two distinct outputs. If the effect of one mutant masks the impact of another, it is inferred that the first gene is epistatic to the second. The rationale behind this conclusion, depending on the pathway, could be that the first gene operates downstream of the second gene. Consequently, when the downstream gene is mutated, any effects upstream are not observable. Epistasis can also extend to functional relationships where the output is simple growth defects, represented with fitness. As noted above, mutants can show positive epistasis when double mutants have better fitness than expected from the fitnesses of single mutants. In this case, it can be interpreted that single mutants may act at a related step and that the double mutant has no further defect than each single mutant. This type of epistasis can infer mutants are acting in the same pathway or step. Along these lines, modern approaches that facilitate genome screening with single and double gene deletions or inhibitions have uncovered the degree of intergenic epistasis across various organisms including yeast. For instance, by generating about 23 million double-knockout gene combinations from almost all genes in budding yeast and screening for colony size, approximately 4% of the combinations have been identified to exhibit epistasis (Costanzo et al., 2016; Domingo et al., 2019).

Epistasis within genes has been quantified by mutating protein sequences (Aakre et al., 2015; Cisneros et al., 2023; Domingo et al., 2019; Johnson, Reddy, & Desai, 2023; Starr & Thornton, 2016). One way to quantify the prevalence of epistasis is to ask how predictable the double mutant's effect is by combining the constituent single mutants' effects using the log additive model (Araya et al., 2012; Bank, Hietpas, Jensen, & Bolon, 2015; Fowler et al., 2010; Melamed, Young, Gamble, Miller, & Fields, 2013; Olson, Wu, & Sun, 2014; Starr & Thornton, 2016). For example, in the absence of any epistasis, all observed double mutants' effects should perfectly match their predicted effects by adding the effects of involved single mutants ($R^2 = 1$ in the correlation between predicted and observed effects of the combined mutants). Conversely, in the presence of full epistasis, the effect of the combined mutant is independent to the effects of involved single mutants (R^2 is approximately 0). The R^2 observed in experiments (Araya et al., 2012; Fowler et al., 2010; Melamed et al., 2013) is around 0.65 – 0.75, indicating the effect of the combined mutants can be moderately well predicted by the involved single mutants. In addition to the prevalence, to quantify the strength of epistasis, a factor representing the deviation of the observed double mutant's effect from the prediction based on adding the single mutants' effects could be determined. For example, a thorough examination of pairwise interactions in protein G domain 1 revealed that weak epistasis (deviation factor < 2) impacted around 30% of all pairs, while approximately 5% of mutation pairs exhibited strong deviations from additivity (deviation factor > 2) (Bank et al., 2015; Olson et al., 2014; Starr & Thornton, 2016). These experimental findings indicated an intermediate level of epistasis prevalence. Notably, weak effect epistasis was more common in the analyzed proteins (observed in ~30% of all pairs with a deviation factor < 2), while strong effect epistasis was less frequent (observed in ~5% of all pairs with a deviation factor > 2). This suggests two trends of epistasis where it could be weak and prevalent, or strong and rare (Domingo et al., 2019; X. Lin et al., 2022; Starr & Thornton, 2016). However, how representative this trend is needs to be assessed, considering that proteins may have different levels of allostery and conservation. Highly conserved proteins, often subject to significant selection pressure during evolution, are likely to exhibit different levels of epistasis than previously reported proteins with lower levels of conservation.

1.3.3 The highly conserved active site of msRNAPs may have evolved various epistasis networks

Emerging mutations change the pre-existing residue epistatic interaction networks within proteins, affecting the accessibility of future mutations. The accumulation of changes in epistasis networks can result in varying mechanisms and future evolution paths of proteins, even for highly similar homologous proteins (Domingo et al., 2019; Johnson et al., 2023; Starr & Thornton, 2016). The evidence for this is distinct effects have been observed when identical residues are introduced into conserved proteins in various systems, including Pol I and Pol II in yeast (Doud, Ashenberg, & Bloom, 2015; Haddox, Dingens, Hilton, Overbaugh, & Bloom, 2018; Natarajan et al., 2013; Viktorovskaya et al., 2013). Substitution of the hyper conserved E1103 residue with a glycine (E1103G) in Pol I and Pol II result in distinct and opposite phenotypes. The Pol II *rpb1* E1103G variant has been extensively studied by biochemical, biophysical and genetic assays, demonstrating an enhanced catalysis along with compromised translocation and fidelity *in vitro*. Pol II E1103G is proposed to reduce transcription translocation by disrupting TL C-terminal interactions required to stabilize the inactive “open” state of TL. This disruption biases TL dynamics toward the active “closed” state, leading to enhanced catalysis and reduced fidelity (Dangkulwanich et al., 2013; Kaplan et al., 2008; Kireeva et al., 2008; Larson et al., 2012; Malagon et al., 2006; Viktorovskaya et al., 2013). To explain the contrasting effects of E1103G (Pol I E1224G) in the highly conserved Pol I and Pol II TLs, Viktorovskaya et al. proposed that Pol I and Pol II have different rate limiting steps based on modeling of reaction rates by Larson et al., which indicated that the Pol II E1103G alters catalysis and translocation but the major rate-limiting step appears to be catalysis (Larson et al., 2012; Viktorovskaya et al., 2013). The idea was that if the E1224G mutation altered both Pol I catalysis and translocation, but if translocation, rather than catalysis were Pol I's major rate-limiting step, then E1224G would result in a slower mutant. Dangkulwanich et al. proposed that the rate limiting step in Pol II includes both catalysis and translocation but also observed that Pol II E1103G showed increased catalysis and reduced translocation rates compared with WT (Dangkulwanich et al., 2013). Later, biochemical experiments showed a decrease in nucleotide addition rate for Pol I E1224G, accompanied by reduced elongation rate and compromised fidelity (Scull, Ingram, Lucius, & Schneider, 2019), suggesting that E1224G reduced Pol I catalysis in transcription and the catalysis is the major rate limiting step for it. This individual mutation provides a striking example for different effects of the same substitution for a conserved residue in different but highly similar proteins.

It is clear that epistasis in the Pol I and Pol II enzymatic backgrounds has reshaped the consequences for what effects E1103G/E1224G have. Supporting this idea, E1224G suppresses a lethal mutation that is expected to impair catalysis in Pol I, implying likely E1224G can increase catalysis in some situations just not in the otherwise WT enzyme. Additionally, the Pol I TL sequence causes lethality when used to replace Pol II TL residues (Rpb1 1076-1106) within Rpb1 while a slightly shorter Pol I TL sequence is viable when replacing Rpb1 1076-1103 though has strong growth defects. Interestingly, Pol II E1103G suppressed growth defects that the Pol I TL (1076-1103) and (1076-1106) caused in the Pol II background, suggesting that within the greater Pol II context, E1103G could behave like a GOF allele even when most of TL sequence matched Pol I (Viktorovskaya et al., 2013). The Pol I and III structures revealed after this work potentially explain this as Pol I especially has divergence in its TL-tip adjacent regions (**Figure 3**) (Engel, Sainsbury, Cheung, Kostrewa, & Cramer, 2013; Hoffmann et al., 2015; Qiu et al., 2016). The Pol I funnel helix appears to have less constraint than the Pol II or Pol III funnel helix. Moreover, unpublished data from our lab has shown that E1224K mutation could suppress a mutation with severe growth defects in Pol I, similar to the analogous mutation E1103K's effect in suppressing a slow mutant in Pol II, implying that Pol I E1224K may also increase elongation rate like Pol II E1103K does. These results indicate that TL is likely controlled by distinct residue interaction networks in enzymatic backgrounds. However, which residue interactions in the Pol II background shape the phenotypes and compatibilities of mutants remain to be elucidated. Deep mutational scanning from Qiu et al (Qiu et al., 2016) allowed for predictions and subsequent analyses looked at the overall compatibility questions, which are discussed in Chapter 3, where we detected the potential residue interactions which may cause incompatibility of TL haplotypes in yeast Pol II background.

Together, the context dependence of the analogous mutation and TL alleles, strongly imply even functions of highly conserved domains are shaped by individually evolved residue interaction networks within specific enzymatic backgrounds. Moreover, despite their high conservation, species-specific TLs may have evolved distinct, enzyme-specific mechanisms over the course of species divergence. Understanding the potential mechanism and evolution requires understanding the specialized residue interaction networks within conserved msRNAPs. Quantitative comparison of the extent of epistasis across highly related msRNAPs is needed. I will describe an experiment to compare and contrast the epistasis interaction networks in three yeast RNA polymerases in Chapter 4.

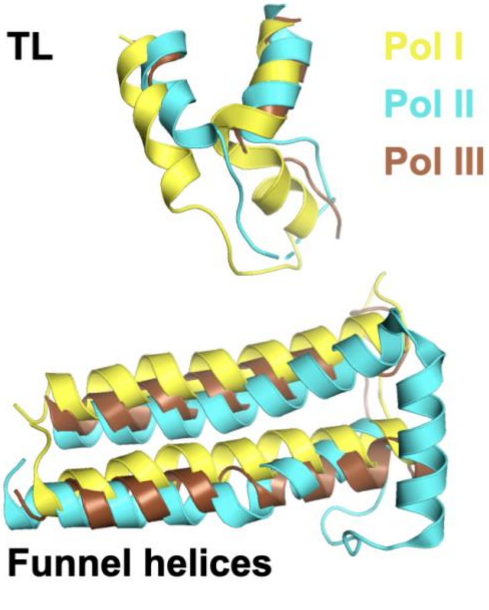


Figure 3. Different positions of funnel helices raltive to TL in three yeast RNA polymerases.

Structures of TLs and funnel helices from *S. cerevisiae* Pol I (PDB: 4C2M)(Engel et al., 2013), Pol II (PDB:5C4J)(Barnes et al., 2015), Pol III (PDB: 5fJ8)(Hoffmann et al., 2015).

1.4 Understanding protein mechanism and evolution in high throughput using deep mutational scanning

1.4.1 Determination of function from a limited numbers of substitutions may limit interpretation of mutant effects

Mutagenesis has been broadly used in understanding protein structure, function, and mechanism in various systems. However, caution is needed when interpreting protein function from limited numbers of substitutions, as potential ambiguities may arise related with the dual nature of the substitution process where the removal of the original WT residue disrupts the original interaction network, and the addition of a new residue may create new interactions. Ignoring either aspect could result in misleading conclusions. For example, E1103G was initially characterized for increasing the elongation rate in Pol II (Malagon et al., 2006). Observations from crystal structures suggested potential interactions between E1103 and T1095, leading to the hypothesis that the T1095-E1103 interaction supports the inactive “open” conformation of the TL, and loss of the interaction promotes TL closing, resulting in hyperactivity of the enzyme. This was supported by the fact that T1095G also increases elongation rate (Kireeva et al., 2008). However, further studies involving additional mutations at both T1095 and E1103 residues challenge the initial model, suggesting the interaction between them may not be critical for maintaining the TL open state (Kaplan et al., 2012). The findings include that almost all additional mutations in E1103 confer GOF phenotype, whereas no T1095 substitutions aside from T1095G exhibit such an effect (Qiu et al., 2016). These results indicate that the observed GOF phenotype of T1095G is likely attributed to the introduction of G, which promotes hyperactivity possibly through the new conformations it may allow. But the GOF phenotype of E1103G is probably due to the elimination of E at position 1103, which removes a negative role in catalysis associated with that residue.

Additionally, TL residue H1085 is proposed to promote the catalysis for its direct contact with the β-phosphate of the substrate (X. Huang et al., 2010; D. Wang et al., 2006), and H1085 is almost universally conserved in all msRNAPs (Castro et al., 2009; Palo, Zhu, Mishanina, & Landick, 2021). However, the functional mechanism of the histidine in catalysis remains unclear. The initial hypothesis is H1085 promotes catalysis by being a general acid (Castro et al., 2007; Castro et al., 2009; Lassila, Zalatan, & Herschlag, 2011), which seemed to be supported by that most substitutions (T, R, K, W, A, F, G, P, N, D, E, S) in H1085 confer severe or lethal growth defects in yeast (Kaplan et al., 2012; Kaplan et al., 2008; Palo et al., 2021; Qiu et al., 2016). However, an additional substitution L identified at H1085 argued against the initial proposal, for that H1085L is relatively healthy in yeast even though leucine does not function as proton donor like histidine (Qiu et al., 2016; Qiu & Kaplan, 2019). Studies combining genetic and biochemical data with H1085L involved have proposed a new model that histidine functions as a positional catalyst to explain why histidine could be substituted by the similar sized and shaped leucine (Mishanina et al., 2017; Palo et al., 2021). However, the introduction of L may create its specific interactions with residues in the active site. The observed slight growth defect of H1085L may not be simply supported by the similar shape and size of leucine to histidine. These observations highlight the benefit of more comprehensive analysis to identify potential residue interactions for TL residues, which is the primary focus of Chapter 2.

Deducing intricate residue interactions on a large scale requires an efficient method. The emerging technique, deep mutational scanning (DMS), which measures the effects of thousands of mutations in parallel within one experiment, has enabled us to detect potential residue interaction networks in the active site of msRNAPs.

1.4.2 General procedures of deep mutational scanning

DMS is a sophisticated method used in genetics and molecular biology to explore the relationship between genetic variations and their corresponding phenotypes. DMS was developed to comprehensively quantify the effects of variants on a large scale economically and efficiently (Fowler & Fields, 2014; Shin & Cho, 2015). Performing a DMS experiment is straightforward, typically comprises three major steps: (1) designing and generating a mutant library; (2) performing high-throughput phenotyping with selective conditions; (3) sequencing of the mutant libraries before and after selection (Matuszewski, Hildebrandt, Ghenu, Jensen, & Bank, 2016; Starita & Fields, 2015; Wei & Li, 2023). The rapid development in genetic technologies, such as gene synthesis, DNA sequencing, and high-throughput phenotyping, has significantly enhanced the effectiveness and scale of DMS experiments (Kinney & McCandlish, 2019; Weile & Roth, 2018). For instance, new methods in array-based DNA oligonucleotide synthesis have enabled creation of vast libraries of variants (Ghindilis et al., 2007; Kosuri & Church, 2014; LeProust et al., 2010), to a level of nearly a million in a single batch. Combined with deep sequencing, it enables the accurate determination of the frequency of each DNA variant (Qiu & Kaplan, 2019), which is crucial for assessing variant effects by comparing its allele frequencies before and after selections.

While the DMS strategy is conceptually straightforward, executing each step to yield reliable and meaningful data can be challenging (Qiu & Kaplan, 2019; Wei & Li, 2023). This complexity arises from managing error rates in variant sequence synthesis, PCR amplification, and establishing appropriate controls for batch-to-batch consistency and comparison in large scale experiments. Utilizing advanced methods to minimize synthesis (error correction in synthesized variant pools etc.) (Lubock, Zhang, Sidore, Church, & Kosuri, 2017) and PCR errors (emulsion PCR etc.) (Tewhey et al., 2009; Williams et al., 2006), carefully designing and incorporating meaningful controls, precisely executing procedures, and incorporation of statistical methods or frameworks (Enrich2, DiMSum etc.) (Faure, Schmiedel, Baeza-Centurion, & Lehner, 2020; Rubin et al., 2017) are essential to guarantee the results are high quality and interpretable.

1.4.3 Applications of deep mutational scanning in understanding protein mechanism and evolution

DMS has been widely applied in many studies and systems since it was initially introduced (Fowler, Araya, Gerard, & Fields, 2011; Fowler & Fields, 2014; Hietpas, Roscoe, Jiang, & Bolon, 2012) and has significantly advanced our understanding of various biological process (Wei & Li, 2023). For example, applying DMS on the SARS-Cov2 spike protein accurately identified specific mutations that became widespread during the later stage of the COVID-19 pandemic within a year of the outbreak (Starr et al., 2022; Starr et al., 2020), demonstrating the power of DMS in addressing critical problems in a relatively short timeframe. Moreover, a subset of DMS experiments aim to understand human genetic diversity, such as multiplex assays of variant effects (MAVE). Many human genetic variants with unknown impacts have been systematically classified as either benign or deleterious (Kinney & McCandlish, 2019; Starita et al., 2017; Weile & Roth, 2018). Additionally, DMS is not only limited to saturated mutagenesis of all single residues of one gene. Genetic interaction profiles between and within genes generated with double or even higher mutant combinations by DMS accurately predict protein structure (Rollins et al., 2019; Schmiedel

& Lehner, 2019) and reveal mechanisms underlying protein function and evolution (Bakerlee, Nguyen Ba, Shulgina, Rojas Echenique, & Desai, 2022; Diss & Lehner, 2018; Faure et al., 2022; X. Lin et al., 2022; Lite et al., 2020; Olson et al., 2014). Importantly, DMS can be coupled with advanced statistical methods like ancestral sequence reconstruction and machine learning or deep learning to comprehensively understand mechanisms behind protein function and evolution (Bakerlee et al., 2022; Cisneros et al., 2023; Ding et al., 2022; Domingo et al., 2019; Johnson et al., 2023; X. Lin et al., 2022). For example, with DMS and ancestral sequence reconstruction, the Thornton lab has statistically identified the decay of predictability of residue effects in long-term evolution of a protein, emphasizing the dynamic nature of residue interactions over time (Park, Metzger, & Thornton, 2022). Together, these studies demonstrate the efficacy of DMS in providing precise insights into protein structure, function, and evolution, a level of detail that is challenging to obtain through other methods.

Online platforms have been developed for compilation of source data such as the MAVEDB (Esposito et al., 2019), allowing potential future studies across different datasets to explore variant effects or residue epistatic interactions across various proteins or systems. This improvement not only emphasizes the current impact of DMS but also highlights its vast potential for future applications.

1.5 Overview of dissertation

The first step in gene expression, transcription by Pol II, requires the function of a flexible, mobile domain in the Pol II active site called the trigger loop (TL). The dynamics and function of TL are maintained by residue interactions within and surrounding it. To determine how the residue interaction networks controls TL function and evolution in large scale, we utilized deep mutational scanning together with our established genetic phenotypes predictive of biochemical defects. To detect pairwise residue interactions within the Pol II TL and between the TL and its surrounding domains in the Pol II active site, we analyzed 11,818 TL alleles including single mutants and a curated subset of double mutants (Chapter 2). To detect higher-order residue interactions within TL haplotypes, we examined 3,373 TL haplotypes including evolutionarily observed TL variants and haplotypes of all possible substitution combinations along the evolutionary path among selected TL variants (Chapter 3). Our analyses indicate TL function and evolution are shaped by widespread epistasis.

2.0 Widespread epistasis shapes RNA polymerase active site function and evolution

2.1 Introduction

Transcription from cellular genomes is carried out by conserved multi-subunit RNA polymerases (msRNAPs) (Allison et al., 1985; Cramer, 2002; Werner & Grohmann, 2011). Bacteria and Archaea use a single msRNAP to transcribe all genomic RNAs (Hirata, Klein, & Murakami, 2008; Vassylyev et al., 2002; G. Zhang et al., 1999), while Eukaryotes have at least three msRNAPs (Pol I, II, and III) for different types of RNAs (Cramer et al., 2001; Fernandez-Tornero et al., 2013; Gnatt et al., 2001; Hoffmann et al., 2015). RNA synthesis by msRNAPs occurs by iterative nucleotide addition cycles (NAC) of nucleotide selection, catalysis and polymerase translocation (Bar-Nahum et al., 2005; Dangkulwanich et al., 2013; Kaplan, 2013; Malinen et al., 2012; D. Wang et al., 2006). msRNAPs active sites accomplish all three steps in the NAC using two conformationally flexible domains termed the bridge helix (BH) and the trigger loop (TL) (Da et al., 2016; Mazumder et al., 2020; Qiu et al., 2016; Vassylyev et al., 2002; D. Wang et al., 2006; Weinzierl, 2010). The multi-functional natures of the BH and TL likely underlie their striking conservation, serving as interesting models for studying the function and evolution of extremely constrained protein domains.

Nearly all catalytic cycle events are associated with the concerted conformational changes in the TL, and potentially the BH (Dangkulwanich et al., 2013; Fouqueau et al., 2013; Kaplan, 2013; Kireeva et al., 2008; Larson et al., 2012; Mazumder et al., 2020; B. Wang et al., 2013; D. Wang et al., 2006).The BH is a straight helix in most msRNAP structures (Cramer et al., 2001; Gnatt et al., 2001; Kaplan, 2013; X. Liu et al., 2013; D. Wang et al., 2006), but was found to be kinked in the *Thermus thermophilus* (Bacteria) RNAP structures (Kaplan & Kornberg, 2008; Vassylyev et al., 2002). The dynamics between the straight and kinked conformations have been simulated and proposed to promote msRNAP translocation (Kaplan & Kornberg, 2008; Silva et al., 2014; Tan et al., 2008; Weinzierl, 2010). Even more importantly, the TL has been observed in various conformations that confer different functions. Among the observed conformations, a catalytic disfavoring “open” state facilitates translocation and a catalytic favoring “closed” conformation promotes catalysis (Barnes et al., 2015; Kaplan, 2013; D. Wang et al., 2006). During each NAC, the TL nucleotide interaction region discriminates correct NTP over non-matched NTPs or dNTPs and initiates a TL conformational change from the open to the closed state (Fong et al., 2014; Malinen et al., 2012; B. Wang et al., 2013; L. Xu et al., 2014). The closure of the TL promotes phosphodiester bond formation (Vassylyev et al., 2007; B. Wang et al., 2013). Pyrophosphate release accompanies TL opening, which is proposed to support polymerase translocation to the next position downstream on the template DNA, allowing for the subsequent NAC (Da et al., 2012; B. Liu et al., 2016; Seibold et al., 2010) (Figure 2). Consistent with the model, mutations in the TL conferred diverse effects in every step of transcription (Kaplan, 2013; Kaplan et al., 2012; Kaplan et al., 2008; Kireeva et al., 2008; Kireeva et al., 2012; Malagon et al., 2006; Matthew H. Larsona, 2012; Qiu et al., 2016). For instance, mutations in the TL NIR impair interactions between TL and substrates, resulting in hypoactive catalysis and reduced elongation rate *in vitro* (Loss of function, LOF) (Kaplan et al., 2012; Kaplan et al., 2008; Kireeva et al., 2008; Nayak et al., 2013; Qiu et al., 2016; Windgassen et al., 2014). Mutations in the TL hinge region and C-terminal portion appear to disrupt the inactive state of TL (open state) and shift the TL towards the active state (closed state), leading to hyperactive catalysis and increased elongation rate but impaired transcription fidelity (Gain of function, GOF) (Barnes et al., 2015; Cheung &

Cramer, 2011; Qiu et al., 2016). TL conformational dynamics and functions are likely balanced by residue interactions within and around the TL (Hein et al., 2014; Kettenberger et al., 2004; Lennon et al., 2012; Nayak et al., 2013; Sekine, Murayama, Svetlov, Nudler, & Yokoyama, 2015).

Intramolecular interactions in the active sites of msRNAPs control catalytic activity and underpin transcriptional fidelity. The TL is embedded in the conserved active site and interacts with other domains such as the BH, α-46 and α-47 helices, which form a five helix bundle with two TL helices enclosing a hydrophobic pocket (**Figure 2**) (Barnes et al., 2015; D. Wang et al., 2006). Many residue interactions observed between the TL and its proximal domains are critical for proper transcription, as catalytic activity and transcription fidelity can be altered by active site mutations within the TL (described above) and domains close to the TL in many msRNAPs (examples include BH GOF T834P and LOF T834A, funnel helix α-21 GOF S713P, and Rpb2 GOF Y769F) (Kaplan et al., 2012; Kaplan & Kornberg, 2008; Kaplan et al., 2008; Kireeva et al., 2008; Malagon et al., 2006; Qiu et al., 2016; Tan et al., 2008). These mutant phenotypes suggest that TL conformational dynamics and function are finely balanced and could be sensitive to allosteric effects from proximal domains (Braberg et al., 2013; Kaster et al., 2016; Leng et al., 2020; Taatjes, 2020). Understanding how "connected" the TL is to the rest of the polymerase will reveal the networks that integrate its dynamics with the rest of the enzyme and pathways for how msRNAP activity and evolution might be controlled.

Physical and functional intramolecular interactions between amino acids define the protein function and evolvability (Breen, Kemena, Vlasov, Notredame, & Kondrashov, 2012; Phillips, 1998; Starr & Thornton, 2016; Tesileanu, Colwell, & Leibler, 2015). Dependence of mutant phenotypes on the identities of other amino acids (epistasis) contributes to protein evolvability by providing a physical context and an evolutionary window in which some intolerable mutations may be tolerated (Karageorgi et al., 2019; Ortlund, Bridgham, Redinbo, & Thornton, 2007; Phillips, 2008). Recent studies have shown that mutations can alter the protein function, allostery, and evolvability, suggesting that even conserved residues are subject to distinct epistatic constraints dependent on context (Ding et al., 2022; Doud et al., 2015; Faure et al., 2022; Kondrashov, Sunyaev, & Kondrashov, 2002; Lunzer, Golding, & Dean, 2010; Natarajan et al., 2013; Park et al., 2022; Starr et al., 2022). In line with this prediction, distinct phenotypes for the same conserved residues have been observed in a number of proteins, including Pol I and Pol II in yeast (Doud et al., 2015; Haddox et al., 2018; Natarajan et al., 2013). For example, the yeast Pol I TL domain is incompatible when introduced into Pol II even though about 70% of residues in the two yeast TLs are identical (Viktorovskaya et al., 2013). The results strongly imply even functions of ultra-conserved domains are shaped by individually evolved enzymatic contexts (higher order epistasis).

Functional interactions between residues can be revealed by genetic interactions of double mutants of interacting residues (Kaplan et al., 2012; X. Lin et al., 2022; Mani et al., 2008; Qiu et al., 2016; Qiu & Kaplan, 2019). Previous studies from our lab on a small subset of site-directed substitutions have identified distinct types of Pol II double mutant interactions including suppression, enhancement, epistasis, and sign-epistasis(Kaplan et al., 2012; Qiu et al., 2016; Qiu & Kaplan, 2019). Suppression was common between LOF and GOF mutants as expected if each mutant is individually acting in the double mutant and therefore, opposing effects on activity are balanced. Similarly, synthetic sickness and lethality were commonly observed between mutants of the same (GOF or LOF) class, consistent with the combination of mutants with partial loss of TL function having greater defects when combined. However, we have also observed lack of enhancement between mutants of similar classes (epistasis), suggesting single mutants might be functioning at the same step, and in one case, sign-epistasis, where a mutant phenotype appears dependent on the identity of a residue at another position. For example, the GOF TL substitution Rpb1 F1084I was unexpectedly lethal with the LOF TL substitution Rpb1 H1085Y (instead of the predicted mutual suppression for independently (Kaplan et al., 2012; Qiu & Kaplan, 2019). This was interpreted as F1084I requiring H1085 for its GOF characteristics and becoming a LOF mutant in the presence of H1085Y (**Figure 4A**). How representative these interactions are, and the nature of interactions across the Pol II active site requires a more systemic analysis to fully describe and understand the networks that control Pol II activity and the requirements for each mutant phenotype.

Deducing complex residue interaction networks on a large scale is challenging. To accomplish this for Pol II, we have previously established genetic phenotypes predictive of biochemical defects (Kaplan et al., 2012) and coupled this with a yeast Pol II TL deep mutational scanning system (Phenotypic landscape) (Fowler & Fields, 2014; Qiu et al., 2016; Qiu & Kaplan, 2019). Here we develop experimental and analytical schemes to extend this system to a wide range of double and multiple mutants within the *S. cerevisiae* Pol II TL and between the TL and adjacent domains (Interaction landscape). By analyzing 11,818 alleles including single mutants and a curated subset of double mutants, we have identified intricate intra- and inter-TL residue interactions that strongly impact TL function. Additionally, the examination of 3,373 haplotypes including evolutionarily observed TL alleles and co-evolved residues revealed that TL function is heavily dependent on the msRNAP context (epistasis between TL and the rest of Pol II). These results suggest that despite being highly conserved, epistasis within msRNAPs contexts functions through derived residues and potentially reshapes functions of conserved residues. Finally, statistical coupling analyses reveals putative allosteric pathways appear to converge on the TL and may modulate active site activity upon factor binding. Our analyses indicate TL function and evolution are dominated by widespread epistasis.

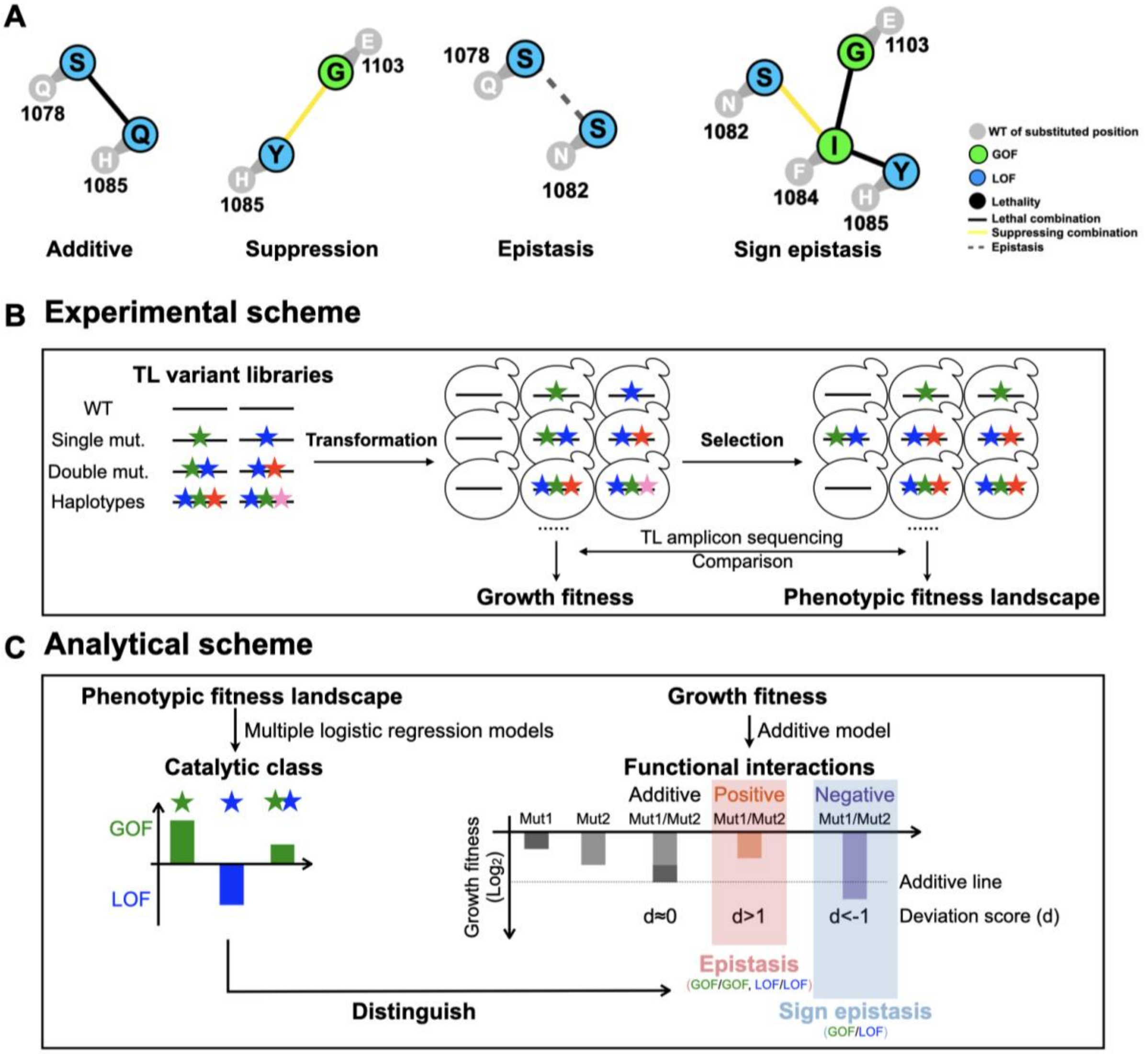


Figure 4. Schematics of the Pol II active site interaction landscape.

(**A**). Examples of inter-residue genetic interactions. WT residues are shown in grey circles with number indicating residue position in Rpb1. Mutant substitutions are shown in colored circles, with color representing mutant class. Colored lines between mutant substitutions represent types of genetic interactions. (**B**). Overview of experimental approach. We synthesized 10 libraries of TL variants represented by colored stars. Libraries were transformed into WT or mutated yeast strains. A selection assay was subsequently performed by scraping and replating the transformants onto different media for phenotyping. DNA was extracted from yeast from all conditions, and went through TL region amplification, and Illumina sequencing. Read counts for variants on general conditions were used to determine growth fitness, while read counts on other conditions were used to determine the phenotypic fitness landscape (see **2.4 Methods**). (**C**). Overview of analytical approach for determining interaction landscape. Mutant conditional growth fitnesses were calculated using allele frequencies under selective growth conditions and subjected to two logistic regression models for classification/prediction of catalytic defects. Double mutant interactions were computed using growth fitness. Classification allowed epistatic interactions to be deduced from double mutant growth fitness (see **2.4 Methods**).

2.2 Results

2.2.1 Systematic dissection of the Pol II active site interaction landscape

We developed an experimental and analytical framework, which we term the Pol II TL interaction landscape, to dissect residue interactions that shape Pol II TL function and evolution in *S. cerevisiae*. We designed and synthesized 15174 variants representing all possible Pol II TL single mutants, a subset of targeted double mutants, evolutionary haplotypes and potential intermediates in ten libraries (**Appendix Table 4**). This approach follows our prior analysis of the TL phenotypic landscape (Qiu et al., 2016) with modifications (see **2.4 Methods** and **Figure 4A**). Libraries were transformed, screened under diverse conditions and phenotyped by deep sequencing (**Figure 4B and 5**). Growth phenotypes of mutants are calculated as the relative allele frequency shift from a control condition and normalized to the WT under the same conditions. Biological replicates indicate high reproducibility (**Figure 5B-C**). Individual libraries were min-max normalized (Sergey Ioffe, 2015) to account for scaling differences between libraries (**Figure**

6A) and the same mutants present among different libraries indicate high correlation of fitness determinations in each library (**Figure 6B-C**).

We defined a conceptual framework for evaluating genetic interactions among TL mutations (**Figure 4C**). First, we assume that independence of mutant effects would result in log additive defects. This means that predicted double mutant fitness defects should be the combination of both single mutant defects, as is standardly assumed (Hill, Goddard, & Visscher, 2008; X. Lin et al., 2022; Mani et al., 2008; Phillips, 2008). Deviation from log additive fitness defects represents potential genetic interactions between single mutants: either less than expected (i.e. suppression) or more than expected (i.e. synthetic sickness or lethality). Second, Pol II has two classes of active site mutants (GOF and LOF) that each confer fitness defects, and we previously observed activity additive interactions, meaning suppression between mutants of different classes (GOF+LOF) or synthetic sickness/lethality between mutants of the same class (GOF+GOF or LOF+LOF) in a set of mutants. We wished to distinguish specific epistatic interactions from activity-dependent suppression or synthetic interactions with mutant catalytic defects. For the purposes of our analysis, we defined an interaction as epistasis when we observed positive deviation in mutants of the same activity class (GOF+GOF, LOF+LOF), where we would expect synthetic sickness or lethality if mutants were functioning independently. We defined an interaction as sign epistasis for situations where we observed negative interaction for combinations between the classes (GOF+LOF), where we would expect suppression if mutants were functioning independently (**Figure 7**).

Finally, the Pol II active site interaction landscape is based on accurate classification of mutant classes. We have previously demonstrated that mutant growth profiles across a select set of growth conditions are predictive of *in vitro* measured catalytic effects (Pol II TL phenotypic landscape). We extended this analysis by training multiple logistic regression models to predict phenotypic classes. We trained two models based on 65 mutants with measured *in vitro* catalytic defects and their conditional growth fitness to distinguish between GOF or LOF classes. Both models worked well in classifying GOF or LOF mutants (**Figure 8A**). These two models were applied to all viable mutants (fitness score > -6.5 for control growth condition) and classified the mutants into three groups, GOF, LOF and those that did not belong to either one of the two groups ("unclassified"). To visually inspect the classification results, we applied t-SNE projection and k-means clustering for all measured mutants in all growth conditions to examine clustering relationships to predictions from multiple logistic regression models. As shown in **Figure 8B**, we observed separated GOF and LOF clusters consistent with logistic regression classifications. With all phenotypic data, GOF and LOF mutants were further classified into different clusters, suggesting more fine-grained separation using additional phenotypes (**Figure 8C**). In summary, we developed an experimental and analytical framework to dissect the Pol II active site residue-residue interaction landscape in high throughput.

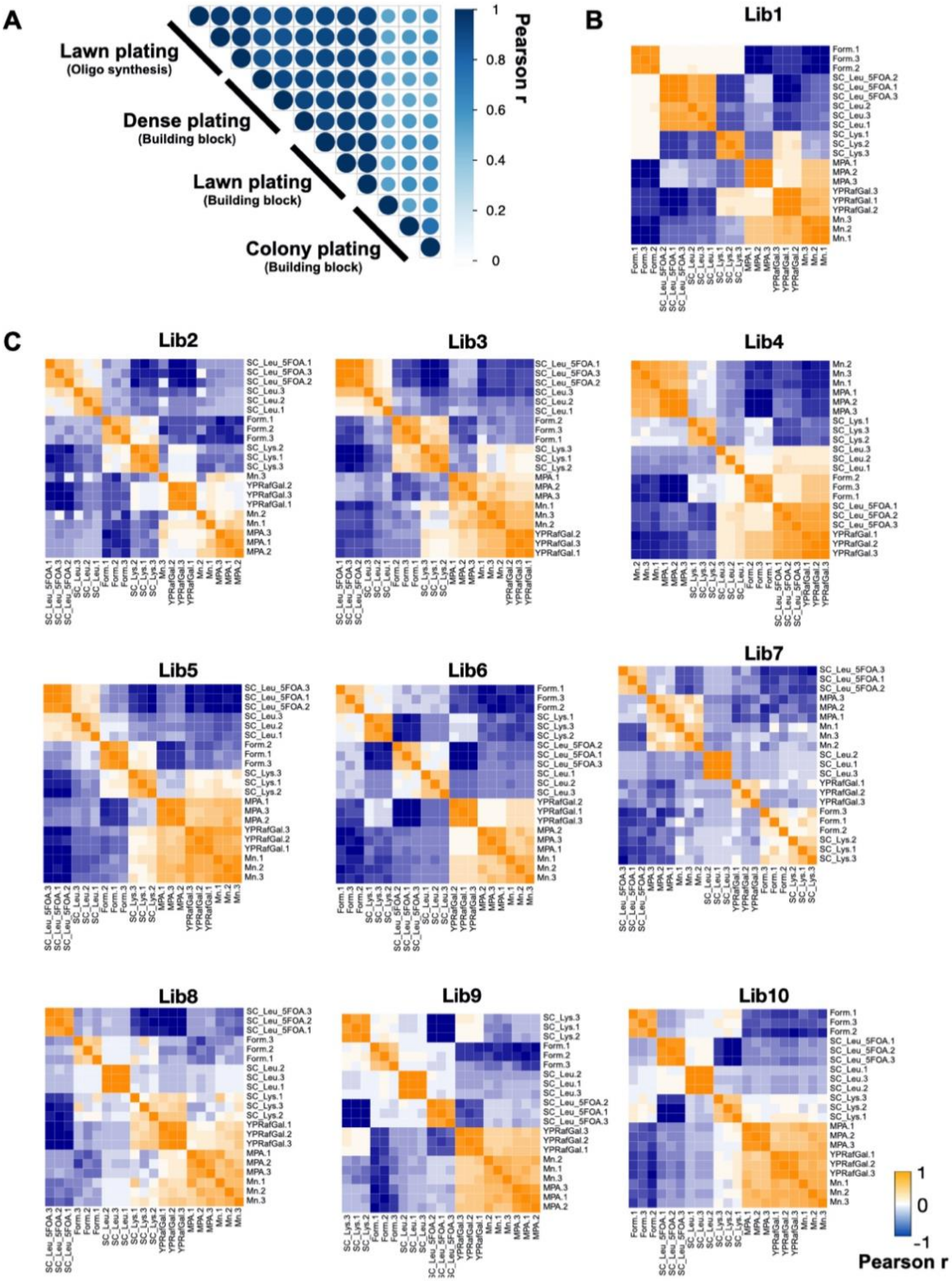


Figure 5. Pol II deep mutational scanning is highly reproducible.

(**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 (Qiu et al., 2016) plated at a lower density (colony plating) that we speculated added noise to the analysis. When plating densely (“dense” and “lawn” conditions) our new and old libraries showed highly reproducible fitness determinations for single mutants. (**B-C**). Biological replicates for each library showed high reproducibility for all conditions. Pearson correlation of each library was calculated with three replicates for viable mutant fitness on all selective conditions.

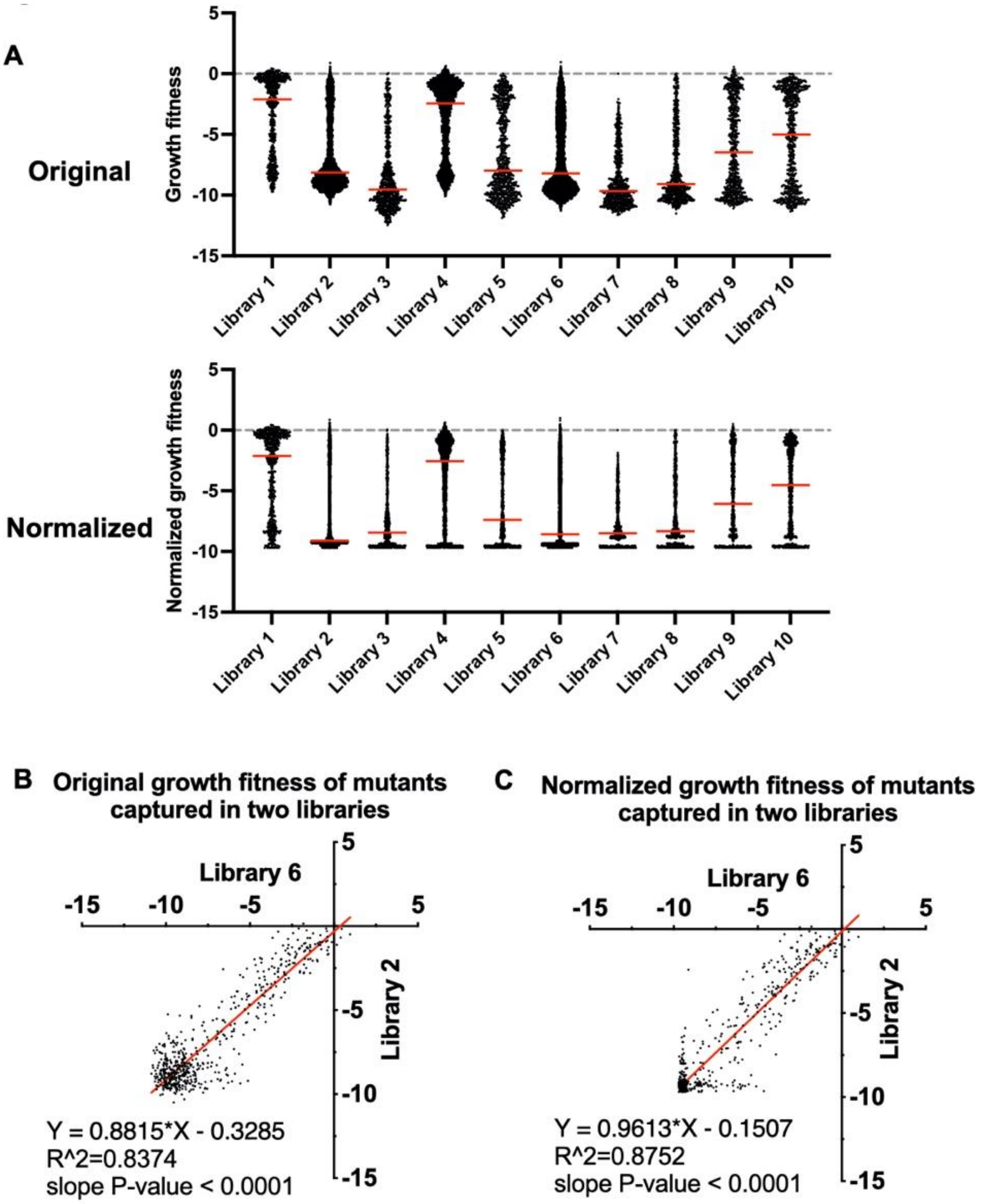


Figure 6. Min-Max normalization uniformed the fitness level of lethal mutants from different libraries without disturbing the median of mutant fitness.

(**A**). Library growth fitness distributions before and after normalization. Upper panel: The growth fitness distributions of libraries. The lowest fitness levels (fitness of lethal mutants) are different among libraries. To uniform various lowest fitness levels, we applied Min-Max normalization to minimize library effects on fitness ranges (See **2.4 Methods** for details). Lower panel: Libraries fitness distributions after normalization. The lethal mutant fitness levels of libraries were normalized to the same level while the median fitness for each library was not affected by the normalization. (**B-C**). XY-plots showing the original fitness of mutants captured in two different libraries (n=586) (**B**). These mutants present in two libraries showed improved correlation between measurements upon normalization (**C**).

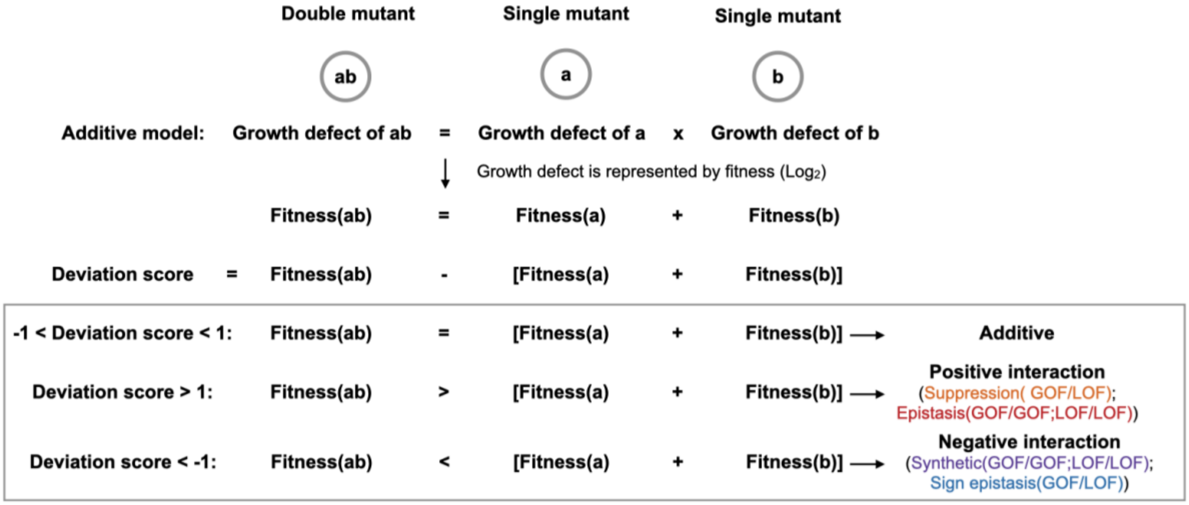


Figure 7. Detection of functional interactions by deviation score.

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

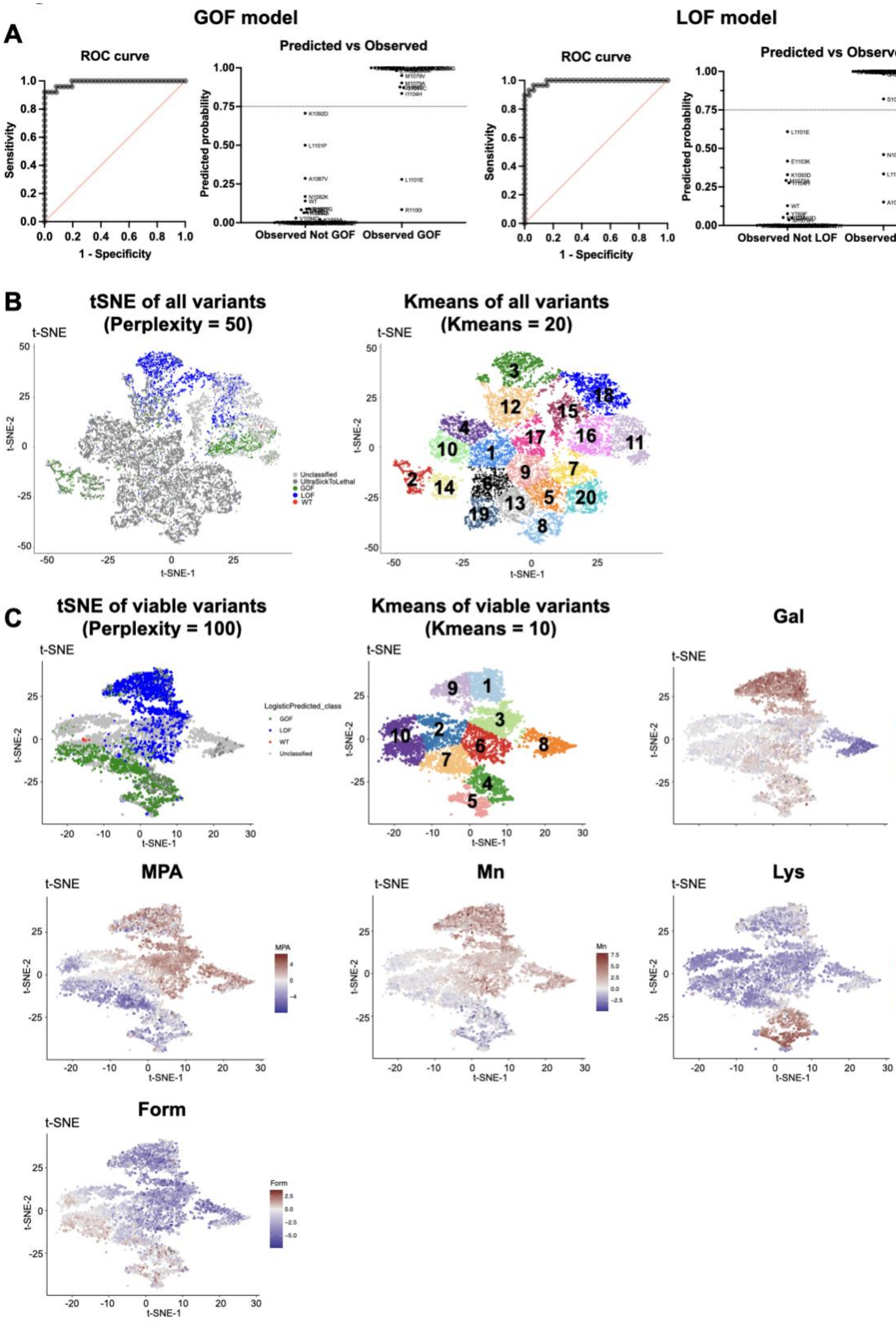


Figure 8. Classification of mutant catalytic defects with machine learning algorithms.

(**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 Appendix Table 5. (**B**). Left: t-SNE projection of all mutants (n=15174) with perplexity = 50. Right: k-means cluster of all mutants with 20 clusters. The t-SNE and k-means projections suggest GOF are in 3 clusters (cluster 2, 14, and 16), LOF are in 2 clusters (cluster 3 and 18), and unclassified mutants are in 2 clusters (11 and 15). Most ultra-sick/lethal mutants (fitness <= -6.5) are projected together into 13 clusters, likely due to significant noise from low read counts across conditions. (**C**). Feature plot of viable mutations in t-SNE and k-means projections (n=6054). Ultra-sick/lethal mutations were removed and the viable mutants were projected with t-SNE (perplexity = 100) and K-means (10 clusters). GOF were grouped into 4 clusters (4, 5, 7 and 10) and LOF were in 4 clusters (1, 3, 6, and 9). Each spot in the projection represents a mutant and it is colored based on the fitness of the mutant in selective conditions. GOF and LOF mutants in different clusters are related to various phenotype patterns. GOF clusters 7 and 10 are defined by strong MPA^S, while clusters 4 and 5 show slight MPA^S, Gal^R, Mn^S, but strong Lys$^+$. Slight FormS 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.

2.2.2 Widespread epistasis in the Pol II TL interaction landscape

To determine the TL-internal interaction networks, we rationally selected 2-4 different substitutions for each TL residue and combined them with the selected substitutions at all other TL positions. Substitutions were chosen to represent diverse phenotypes (GOF, LOF, lethal, or unclassified mutants). This curated set of 3790 double mutants represents potential interactions between any two TL residues (**Figure 9A**). We compared the observed fitness of these double mutants with expectations from the additive model, and noticed the observed double mutant fitness deviated from the predicted fitness (r^2=0.21), which is much smaller than the r^2 (about 0.65-0.75) reported in other studies (Araya et al., 2012; Fowler et al., 2010; X. Lin et al., 2022; Melamed et al., 2013; Starr & Thornton, 2016) (**Figure 9B**), suggesting epistasis in the ultra-conserved TL domain might be more prevalent. About half of the combinations (1776/3790) matched the additive model (observed fitness ≈ expected fitness), while the rest showed positive (observed fitness > expected fitness, n=612) or negative (observed fitness < expected fitness, n=1402) interactions (**Figure 9B**). From these positive or negative interactions, we distinguished the ratio of epistasis relative to activity-additive interactions. In all GOF/LOF combinations, we observed 43% activity-additive suppression and 41% negative interactions (sign epistasis). In all GOF/GOF or LOF/LOF combinations, activity-additive synthetic sick or lethal interactions were much more common than epistasis in combinations within the same class. We observed ~2% positive (epistasis) and 95% negative (activity-additive synthetic sick or lethal) interactions in GOF/GOF combinations, and 6% positive (epistasis) and 84% negative (synthetic sick or lethal interactions) interactions in LOF/LOF combinations (**Figure 9C, 9E and 10**). Interactions were distributed throughout the TL and covered every TL residue, supporting connectivity across the TL. Observed epistasis was concentrated within the C-terminal TL helix and adjacent regions (**Figure 9D**), supporting functional dependency of TL-C terminal residues and consistent with their proposed function to collaboratively stabilize the TL open state.

Genetic interactions reveal further insight into the nature of previously lethal or unclassified individual mutants. First, most lethal mutants could be suppressed by at least one predicted GOF mutant (**Figure 9F and 10**), suggesting that most lethal mutants likely have reduced activity (LOF) below a viable threshold, as might be predicted from greater probability of any individual mutant being a LOF than a GOF. However, two lethal mutations could be suppressed by most LOF mutations or specific other lethal mutants, but not GOF mutants, implying that their lethality resulted from being GOF (select A1076 substitutions). Second, unclassified single mutants mostly did not show widespread interactions with GOF, LOF, or lethal classes. However, a few unclassified mutants showed suppression in combination with GOF mutants, suggesting potential atypical LOF not detected by phenotypic analysis, or potential sign epistasis (**Figure 9G and 10**).

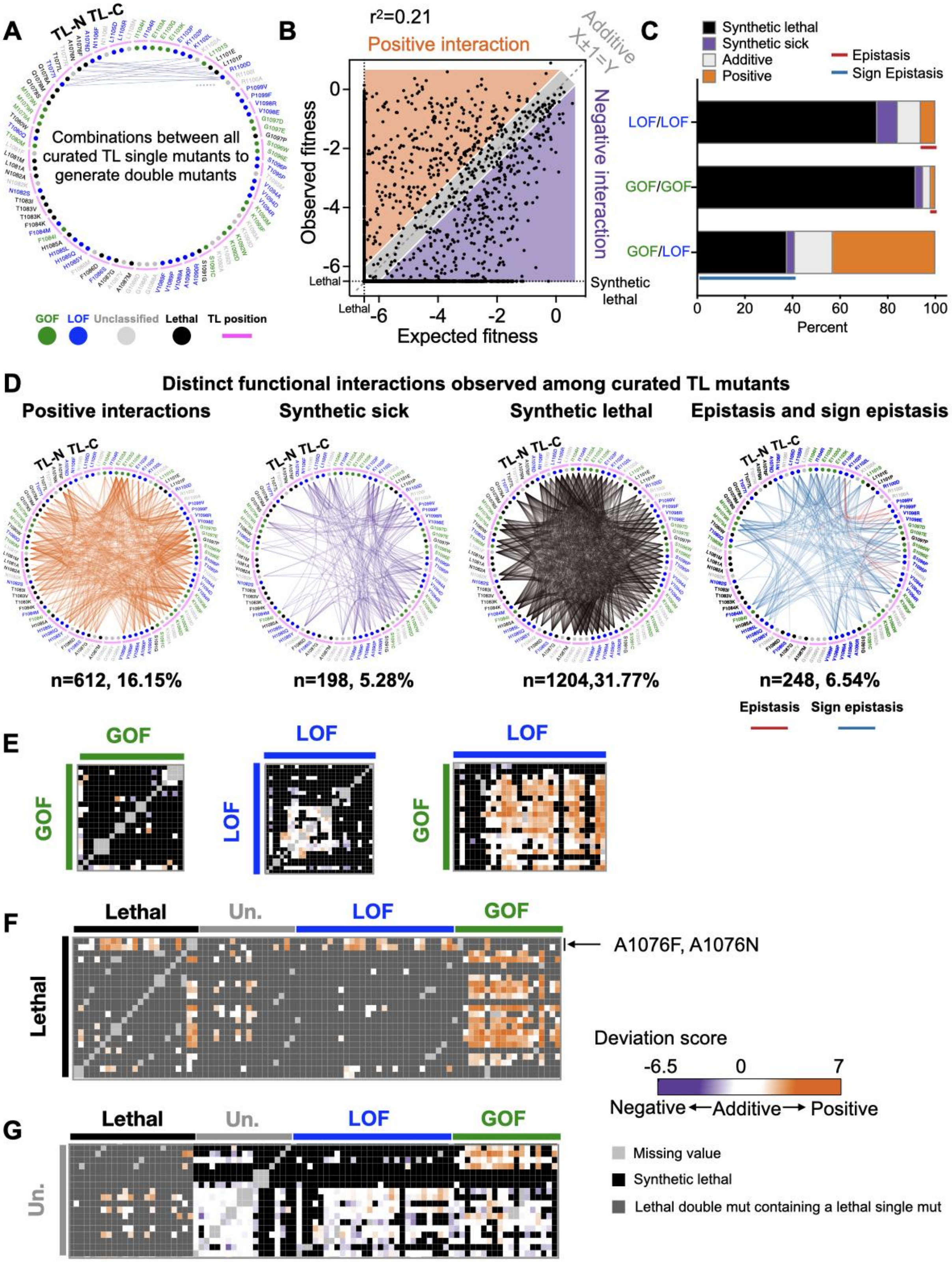


Figure 9. Widespread epistasis in the Pol II TL interaction landscape.

(**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 ± 1 =Y is indicated with dashed line. Simple linear regression was performed, and the best fit equation is $Y = 0.52X -2.55$, $r^2 = 0.21$, $P < 0.0001$. (**C**). Percent of interactions observed from each combination group. N (LOF/LOF) = 412. N (GOF/GOF) = 156. N (GOF/LOF) = 534. Epistasis and sign epistasis are indicated with colored lines. (**D**). Various groups of interactions are displayed in network format. (**E-G**). The intra-TL functional interaction heatmaps of various combinations. Double mutant deviation scores are shown in the heatmap. Annotations at the top and left indicate the curated single mutants and their predicted phenotypic classes from multiple logistic regression modeling. GOF/GOF, LOF/LOF, and GOF/LOF combinations are shown in **E**. Combinations with lethal single substitutions are in **F**. Combinations with unclassified mutants are in **G**.

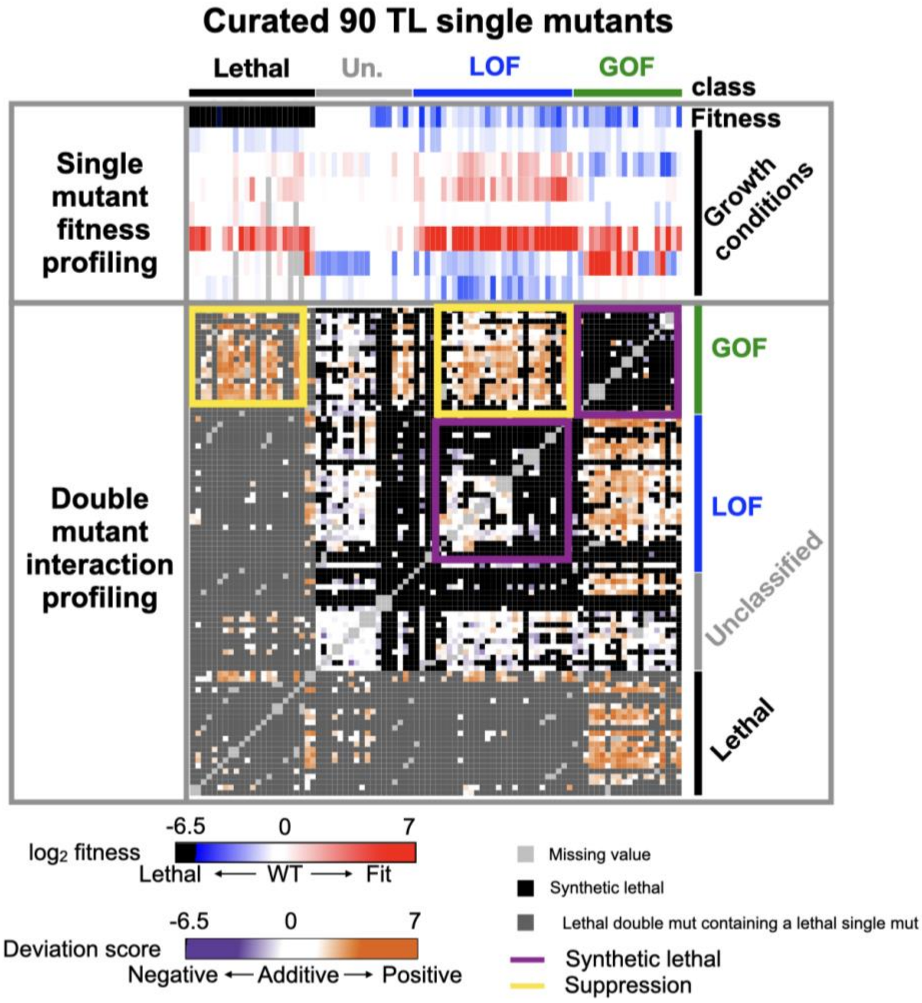


Figure 10. The intra-TL functional interaction landscape.

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

2.2.3 Allele-specific interactions suggest unique properties of individual mutants with similar phenotypes

TL conformational dynamics and function are balanced by residue interactions within the TL (TL-internal interactions) and between the TL and TL-proximal domains (TL-external interactions). The properties of GOF and LOF mutants adjacent to the TL appear similar to those inside but how they behave upon TL perturbation is not known. We analyzed the scope and nature of TL-internal and TL-external interactions by exploring interaction space of 12 previously studied GOF and LOF mutants (8 within the TL and 4 outside) each combined with all possible single TL mutants (**Figure 11A**). These 12 mutants function as probes for the genetic interaction space of the TL and how it might be altered in allele-specific fashion by perturbation of the "probe" mutation. TL adjacent mutants showed similar scale of widespread interactions with TL substitutions as when TL-internal mutants were used as probes (**Figure 11B and 12**). For these TL adjacent substitutions, we conclude their impact on Pol II function is of similar magnitude and connection as substitutions within the TL.

We further compared the similarity of interaction networks for substitutions with apparently similar biochemical and phenotypic defects. These analyses were designed to detect if changes to TL function might reflect simple alterations to TL dynamics, or additional alteration to folding trajectories or conformations. In the former case, mostly additive interactions might be predicted due to TL operating in the same fashion in double mutants versus single mutants, with phenotypes deriving from differences in kinetics or distributions of existing states. In the latter case where a mutation alters TL folding trajectories or changes TL conformations, it might be predicted that individual mutants that are superficially similar will show allele-specific genetic interactions reflecting epistatic changes to TL function. A subset of probe mutants showed widespread expected activity-additive suppression between GOF/LOF mutations and activity-additive synthetic lethality between same classes of substitutions (LOF/LOF or GOF/GOF). However, allele-specific epistasis and sign epistasis were also observed and were much higher for some mutants than others (**Figure 11B-C, 12 and 13**). 127/620 TL substitutions showed unique interactions with specific probe mutants; for example, some lethal substitutions could only be suppressed by Y769F, a GOF TL-proximal probe mutant in Rpb2 (**Figure 14**). Moreover, two TL-adjacent GOF probe mutants, Rpb1 S713P (funnel α-helix 21) and the BH allele Rpb1 T834P displayed greatly distinct interaction networks despite similarly increased activities. Rpb1 S713P exhibited widespread suppression of LOF TL substitutions (96 instances) consistent with generic enhancement of activity but preservation of TL function. In contrast, Rpb1 T834P exhibited much lower suppression ability (33 instances). In addition to much lower ability to suppress, T834P showed a much greater amount of sign epistasis than Rpb1 S713P (102 instances to 38 instances) (**Figure 11C and 15A**). These results are consistent with a model that perturbation to the BH structure is coupled to extensive changes to TL functional space and that T834P function as a GOF mutant requires most TL residues to be WT.

A similar distinction as above but between two internal TL GOF substitutions, Rpb1 E1103G and Rpb1 F1084I, was also apparent (**Figure 15B**). Rpb1 E1103G showed widespread suppression of LOF TL substitutions (184 instances), consistent with site-directed mutagenesis studies (Kaplan et al., 2012) (**Figure 11C, 12 and 15B**). These results suggest E1103G primarily may alter TL dynamics consistent with biochemical data that it promotes TL closure (Kireeva et al., 2008) and that it allows TL mutants primarily to maintain their effects. In contrast, Rpb1 F1084I showed more limited suppression of LOF alleles (43 instances) while showing much more widespread synthetic lethality (**Figure 11C, 12 and 15B**). These results indicate F1084I has a much greater requirement for WT residues at many TL positions to maintain its GOF characteristics. When TL function is additionally perturbed, F1084I appears to switch from a GOF to a LOF. These results imply that individual probe mutants distinctly reshape the Pol II active site, though they might share catalytic and phenotypic defects as single mutants.

An even more striking example of this phenomenon can be observed by comparison of the interaction networks of two LOF substitutions at the exact same position, the ultra-conserved H1085 residue (**Figure 15C**). This histidine contacts incoming NTP substrates (Vassylyev et al., 2002; D. Wang et al., 2006), is the target for the Pol II inhibitor α-amanitin (Kaplan et al., 2008), and promotes catalysis (Kaplan, 2013; D. Wang et al., 2006). Initial structural data and molecular dynamics simulations were interpreted as H1085 potentially functioning as a general acid for Pol II catalysis (Carvalho, Fernandes, & Ramos, 2011; Castro et al., 2009; X. Huang et al., 2010; Unarta, Goonetilleke, Wang, & Huang, 2023). Our discovery that H1085L was especially well-tolerated (Qiu et al., 2016), and subsequent experiments from the Landick lab (Mishanina et al., 2017; Palo et al., 2021), have led to their proposal that the TL histidine functions as a positional catalyst and a similarly sized leucine supports catalysis with relatively mild effects on biochemistry and growth. If H1085Y and L substitutions are acting on a continuum of positional catalyst activity, we might predict their interaction networks would be similar and only be distinguished by magnitude of interactions, but not identity or type of interactions. In contrast to this prediction, distinct interaction patterns were observed (**Figure 11C, 12 and 15C**). Most GOF mutants were able to suppress H1085Y but not H1085L. Instead, H1085L showed synthetic lethality with most GOF mutants (putative sign epistasis). For example, almost all substitutions at E1103 showed sign epistasis with H1085L but not H1085Y (**Figure 13B and 15C**). Distinction between H1085L and H1085Y is evident in the PCA plot of probe mutants (**Figure 11D**). The partially unique nature of each probe mutant is also evident in the PCA plot (**Figure 11D**). Altogether, distinguishable interaction networks of probe mutants, despite their similarity in catalytic and growth defects, even within the same residue, suggest that each mutant has ability to propagate effects across the Pol II active site. To some extent, each Pol II mutant creates a new enzyme.

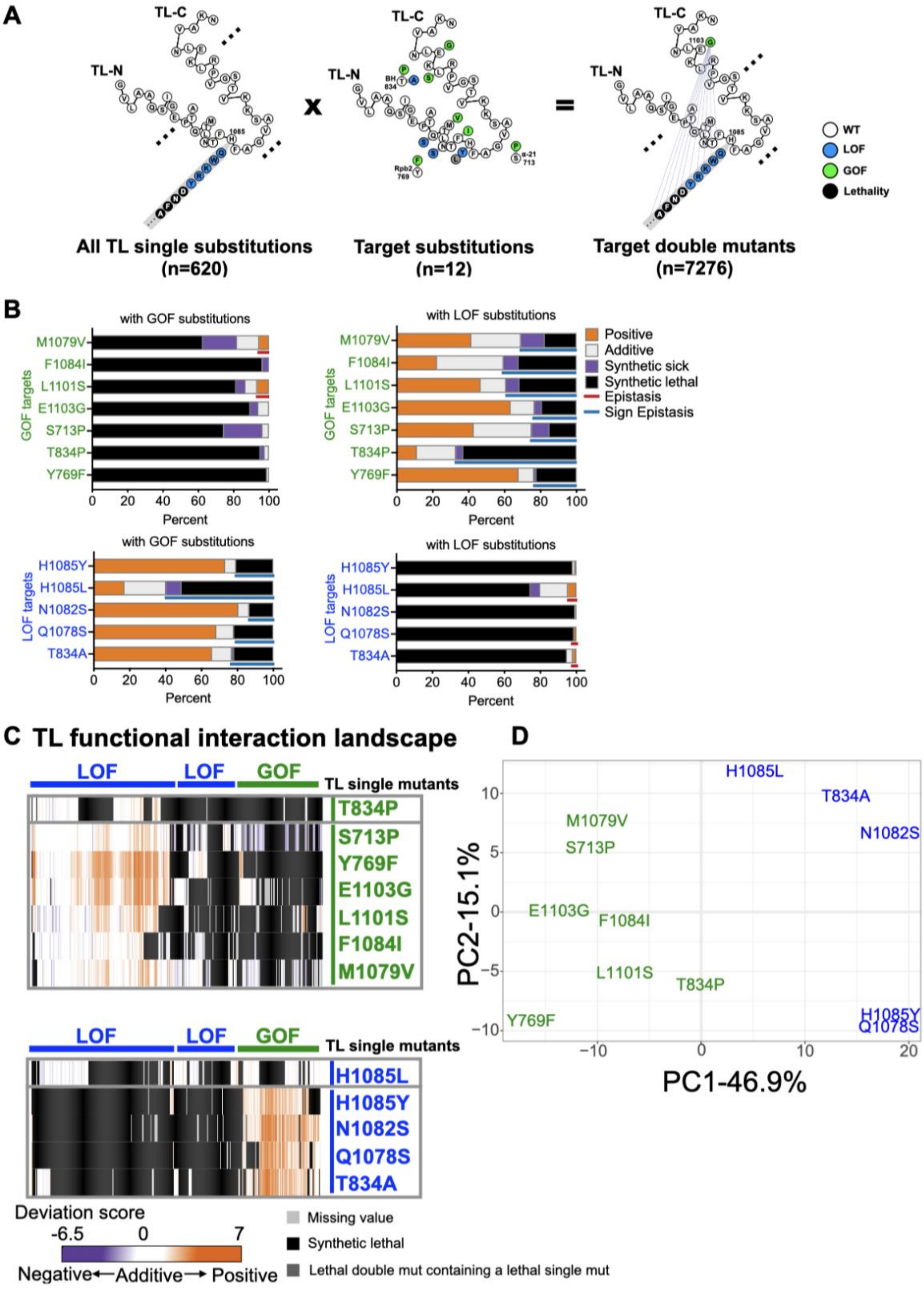


Figure 11. Pol II TL interaction landscape distinguishes mutants with similar phenotypes.

(**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 **2.4 Methods**).

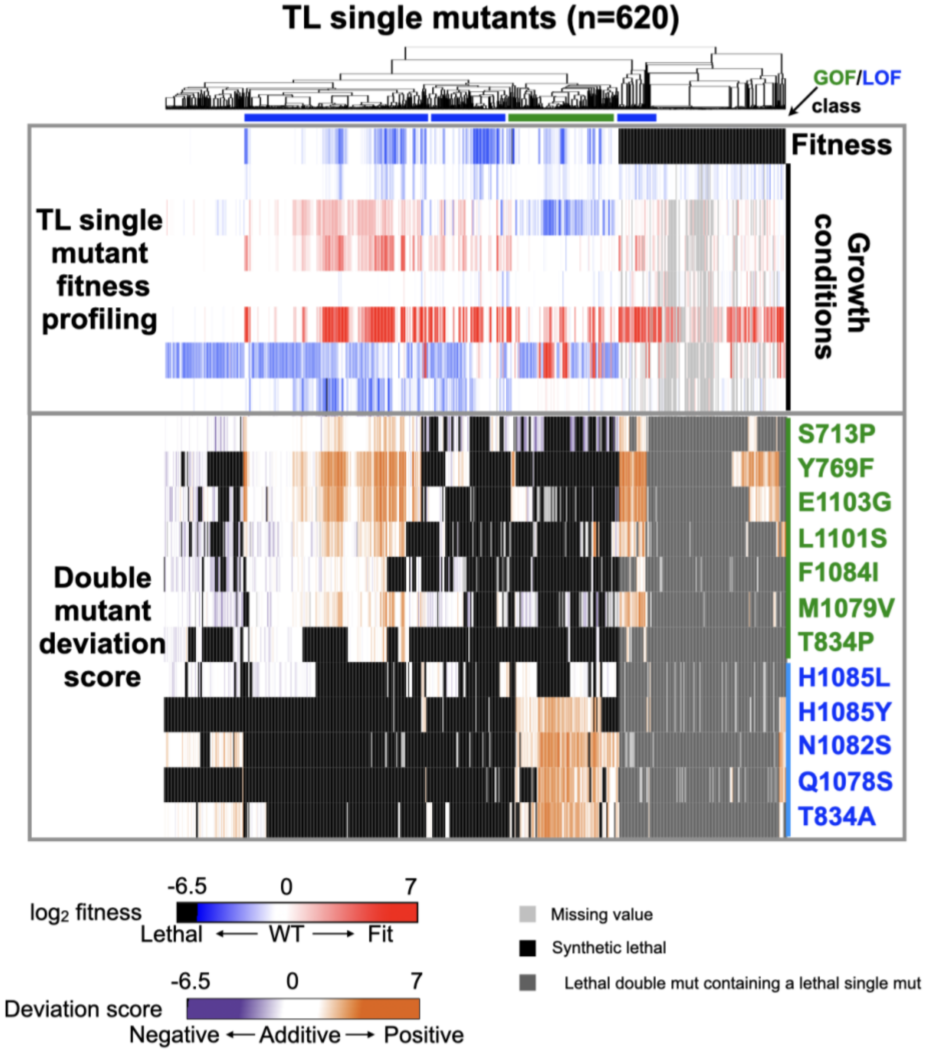


Figure 12. The functional interaction landscape of probe mutants.

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

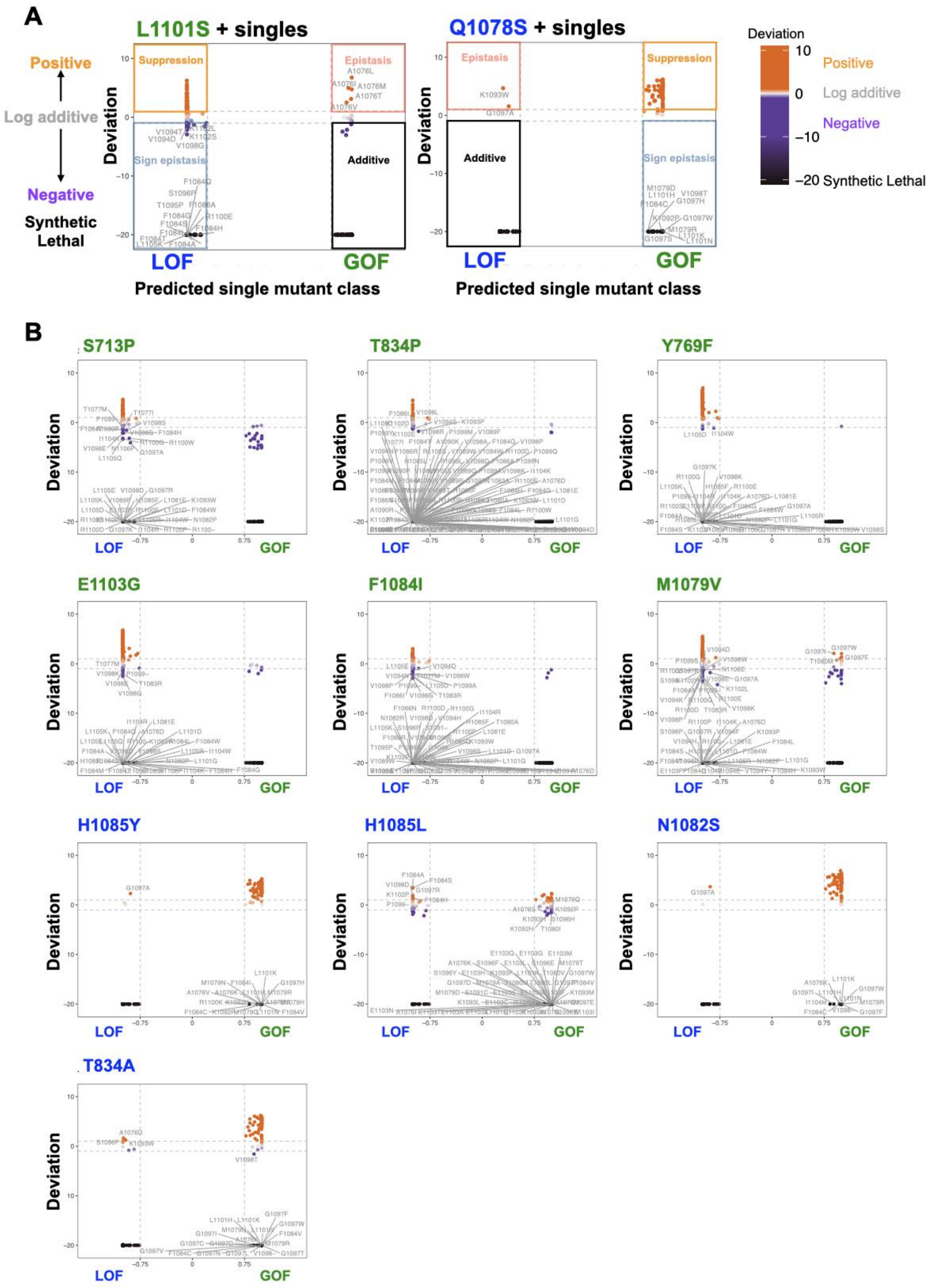


Figure 13. Identifying TL substitutions that interact with the probe mutants.

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

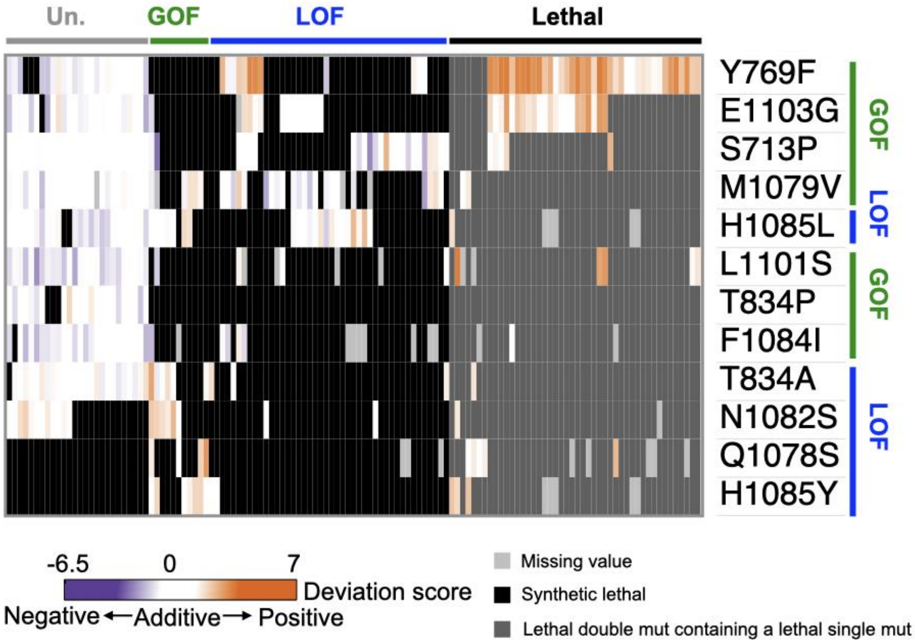


Figure 14. Allele-specific interactions.

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

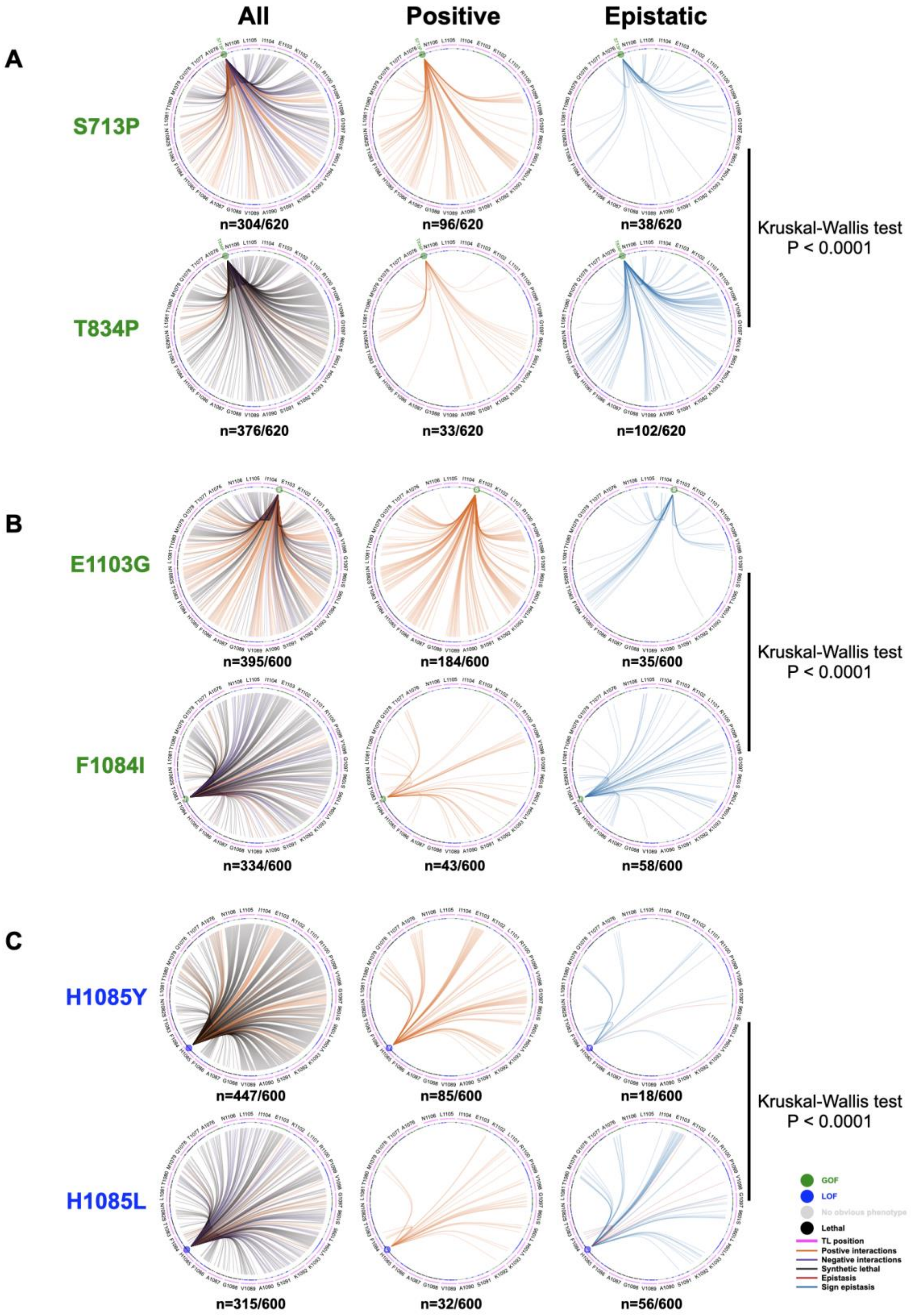


Figure 15. Interaction networks of selected probe mutants.

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

2.2.4 Pol II TL interaction landscape reveals functional dependency of proximal residues

Several allele-specific epistatic interactions were also observed. Some of the strongest epistatic interactions were between A1076 substitutions and L1101S, which differed from all other GOF probe mutants (**Figure 16A**), suggesting tight coupling between A1076 and L1101 for Pol II function. These two hydrophobic residues, together with other hydrophobic residues in TL proximal helices, form a five-helix bundle in the Pol II active site likely stabilizing the open TL conformation (**Figure 16C**). Consistent with this, another pair of adjacent residues, M1079 and G1097, also showed allele-specific epistasis (**Figure 16B-C**).

The epistasis we identified in combinations within the same class (GOF/GOF or LOF/LOF) might also be sign epistasis (GOF suppressing GOF or LOF suppressing LOF due to a switch in residue class). We distinguished regular epistasis (lack of additivity) from sign epistasis suppression by checking conditional phenotypes predictive of biochemical defects. We reasoned that epistatic interactions would exhibit double mutant conditional phenotypes similar to single mutants while sign epistasis suppression would also exhibit suppression of conditional phenotypes. Therefore, we examined double mutants with our logistic regression models for determining phenotypic class. The majority of double mutants within each class showing positive epistasis (GOF/GOF or LOF/LOF) maintained single mutant classification. 6/10 GOF/GOF doubles showing positive epistasis were classified as GOF while 30/38 LOF/LOF doubles were classified as LOF, suggesting classic epistasis (**Figure 17A**). In three cases of GOF/GOF combinations, all between L1101S and A1076 substitutions, the resulting double mutants were unclassified, consistent with nearly WT behavior. Here, each constituent single mutant conferred a GOF phenotype, but the double mutants show mutual suppression. This suggests tight coupling between 1101 and 1076 (see **2.3 Discussion**).

We also observed allele-specific interactions for predicted lethal mutants. Our threshold for lethality is likely higher than that in actuality, and very slow growing mutants may fall below our lethal threshold while still having enough data on conditional fitness assessment for logistic regression to predict mutant class. For 21 ultra sick/lethal TL substitutions predicted as GOF themselves, we observed suppression when combined with other GOF mutants (**Figure 17B-C**). Lethal substitutions of A1076 could be suppressed by LOF probe mutants and the GOF probe L1101S, consistent with specific combinations between 1076 and 1101 showing sign-epistasis suppression or allele-specific mutual suppression. F1084R is a predicted lethal GOF but can be suppressed specifically by GOF probe Y769F. F1084 and Y769 are close to each other when the TL is in the closed, substrate bound state. Additionally, ultra-sick/lethal substitutions predicted as LOF could be suppressed by a LOF allele (**Figure 17B**). As an example, S1091G could be suppressed by almost all curated GOF mutants, yet it was also specifically suppressed by the LOF V1094D (**Figure 17C**). S1091G and V1094D appear to compensate for each other in a allele-specific fashion. We suggest that these are the types of interactions that will allow the TL and adjacent residues to evolve and differentiate while maintaining essential functions.

We note that strong epistasis is much more prevalent in the Pol II system than in other proteins where it has been quantified (Araya et al., 2012; Fowler et al., 2010; Harms & Thornton, 2010; Melamed et al., 2013; Starr & Thornton, 2016) (**Figure 17D**). We attribute this difference to the much higher rate of suppressive interactions due to Pol II mutants having opposing effects on catalysis.

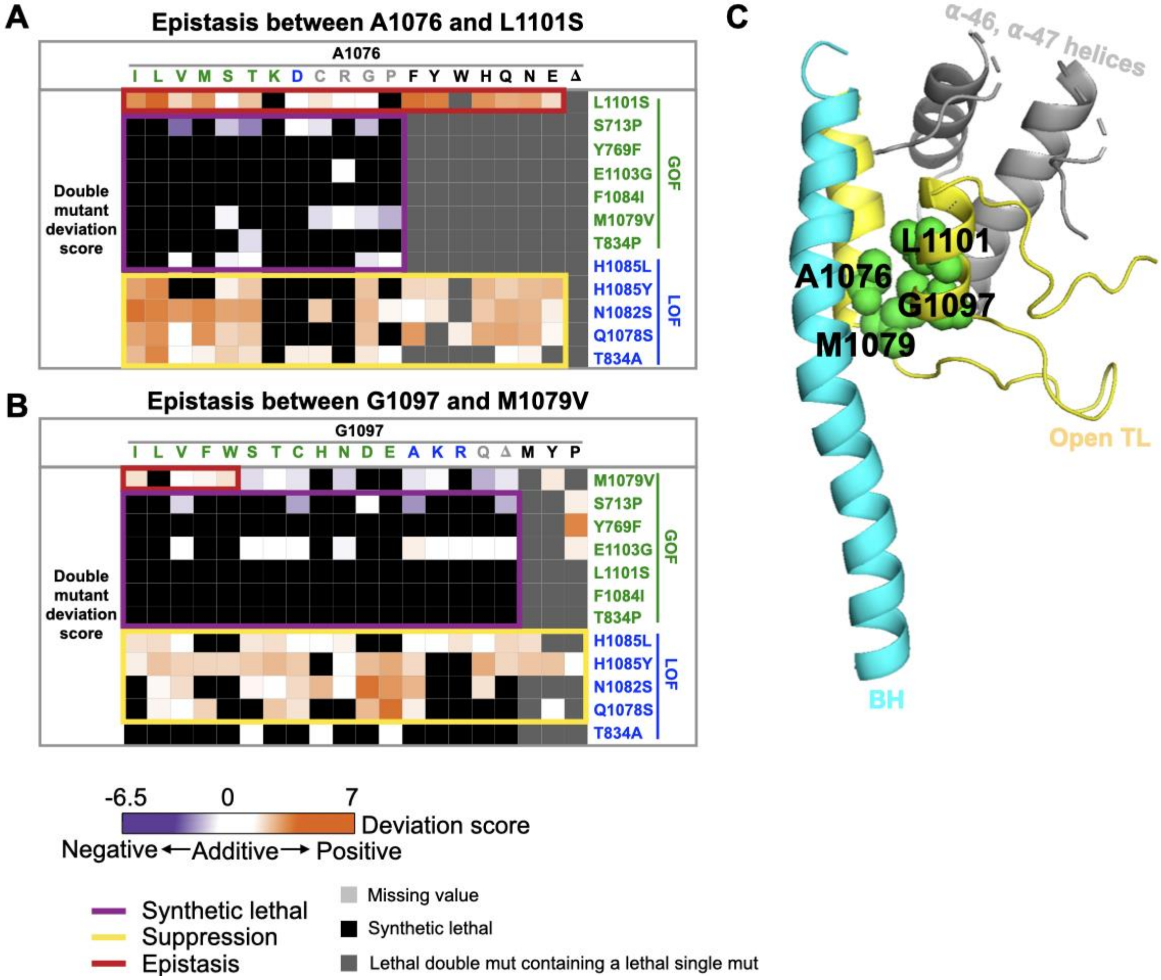
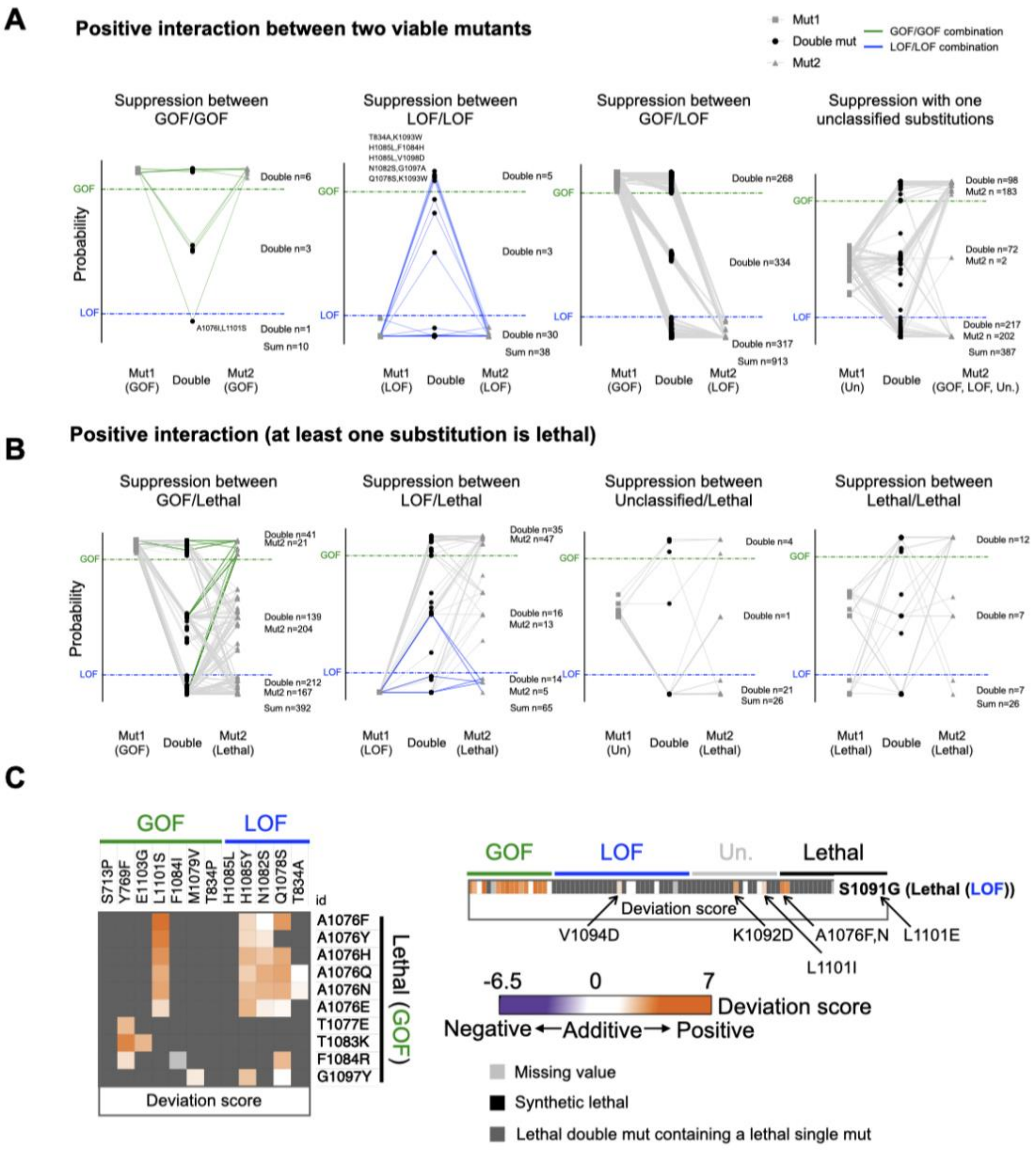


Figure 16. Pol II TL interaction landscape reveal functional dependency of proximal residues.

A-B. Specific epistatic interactions observed between hydrophobic residues A1076 and L1101 (A), and M1079 and G1097 are shown as heatmaps (B). The x-axis of both heatmaps are 20 substitutions ordered by predicted phenotypic classes, and the color of substitution represents the phenotypic class of the substitution. GOF substitution is in green, LOF is in blue, unclassified is in gray, and lethal (fitness < -6.5) is in black. **C.** The epistatic interactions we identified between A1076 and L1101, together with M1079 and G1097 are shown on the five-helix bundle of Pol II active site (PDB:5C4X)(Barnes et al., 2015).



D

Ratio of observed strong or weak epistasis

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

Figure 17. Discrimination of regular epistasis from sign epistasis.

(**A**). The phenotypic classes of double mutants consist of two viable single substitutions with positive interactions. Four plots show four kinds of combinations respectively. For each plot, the predicted GOF or LOF probabilities of a double mutant and two constituent single mutants are shown in Y-axis. The double mutant and two constituent single mutants are shown in X-axis in the order of the first constituent mutant (Mut1), the double mutant (Double), the second constituent mutant (Mut2). The double mutant and two constituent mutants are connected with lines for each pair of combinations. The numbers of double mutants belonging to GOF (top), unclassified (middle), or LOF (bottom) are labeled. GOF and LOF probability threshold are labeled with dashed lines. (**B**). The phenotypic classes of double mutants consist of one viable and one lethal single substitutions, or two lethal single substitutions with positive interactions. The arrangements of plots are similar to A. (**C**). The heatmaps of lethal GOF substitutions suppressed by GOF targets (left) and lethal LOF substitutions suppressed by LOF targets (right). (**D**). The fraction of strong and weak interactions we observed in double mutants compared with the ratio reported in other studies (Araya et al., 2012; Fowler et al., 2010; Harms & Thornton, 2010; Melamed et al., 2013; Starr & Thornton, 2016)**.**

2.2.5 TL evolution is shaped by contextual epistasis

We previously found that identical mutations in a residue conserved between the Pol I and Pol II TLs yielded different biochemical phenotypes (Scull et al., 2019; Viktorovskaya et al., 2013). Furthermore, the yeast Pol I TL was incompatible within the yeast Pol II enzyme, implying that TL function is shaped by the enzymatic context (Scull et al., 2019; Viktorovskaya et al., 2013). To determine the generality and scope of TL-Pol II incompatibility, we designed a library containing evolutionary TL variants from bacterial, archaeal, and eukaryotic msRNAPs and determined their compatibility in the yeast Pol II context **(Figure 18A**). TL alleles of eukaryotic Pols were more compatible than those from Archaea and Bacteria, and Pol II alleles were the most compatible (**Figure 18B and 19A-B**), consistent with evolutionary distance. The total number of TL substitutions in haplotypes were slightly negatively correlated with growth fitness in the Pol II background for Archaeal, Pol I, II and III TLs (**Figure 19C**), though not for Bacterial TLs, likely because the bacterial TLs were almost entirely incompatible in the Pol II context (**Figure 19C**). Conservation of TL sequence and function was high enough that some archaeal sequences could provide viability to yeast Pol II, yet at the same time a number of Pol II TLs from other species were defective if not lethal. These results suggest widespread coevolution of TL sequence outside of ultra-conserved positions shapes TL function (see **2.3 Discussion**).

We reasoned that evolutionarily observed lethal substitutions might be closer to functional than non-evolutionarily observed and would therefore be more likely to be suppressible by Pol II GOF alleles. To compare suppressibility between evolutionarily observed and unobserved substitutions lethal to Pol II, we extracted the highest positive deviation scores among all double mutants containing each lethal substitution. Maximum deviation scores for Pol II lethal substitutions present in TLs of existing msRNAPs were higher than for lethal substitutions that were absent, indicating the Pol II lethal mutants present in existing msRNAPs on average maintain a greater functionality and/or are suppressible by single changes (**Figure 18C and 19B**). The TL has been estimated as providing 500-1000 fold enhancement on catalytic activity (Toulokhonov, Zhang, Palangat, & Landick, 2007; W. Wang, Walmacq, Chong, Kashlev, & Wang, 2018; Yuzenkova et al., 2010), while we estimate only ~10-fold effects are tolerated for yeast viability (Kaplan et al., 2008). We conclude that lethal mutants observed as functional residues in other species are more likely to be close to the viability threshold as might result from a series of small steps to allow them to function.

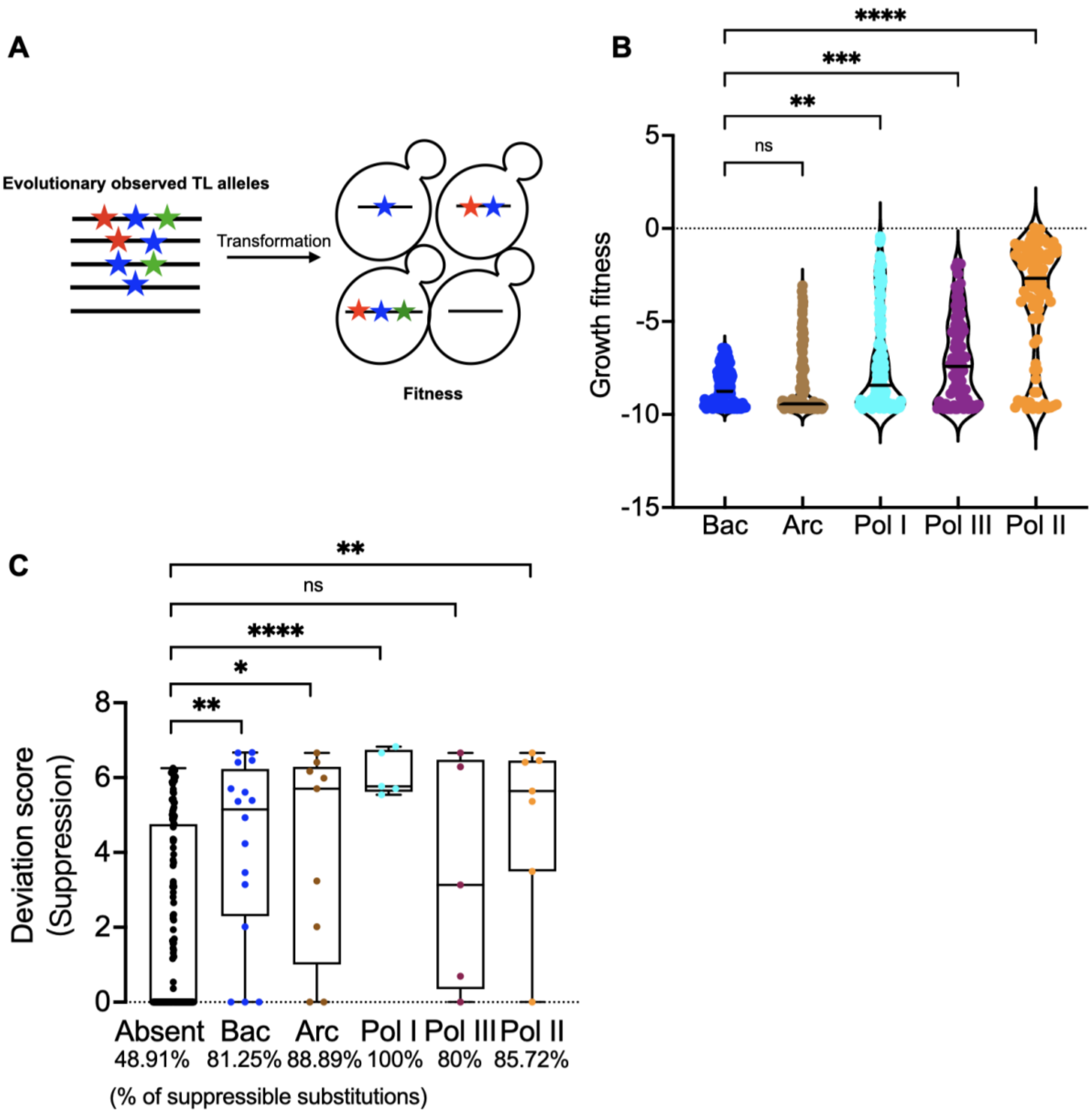


Figure 18. Contextual epistasis shapes TL evolution.

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

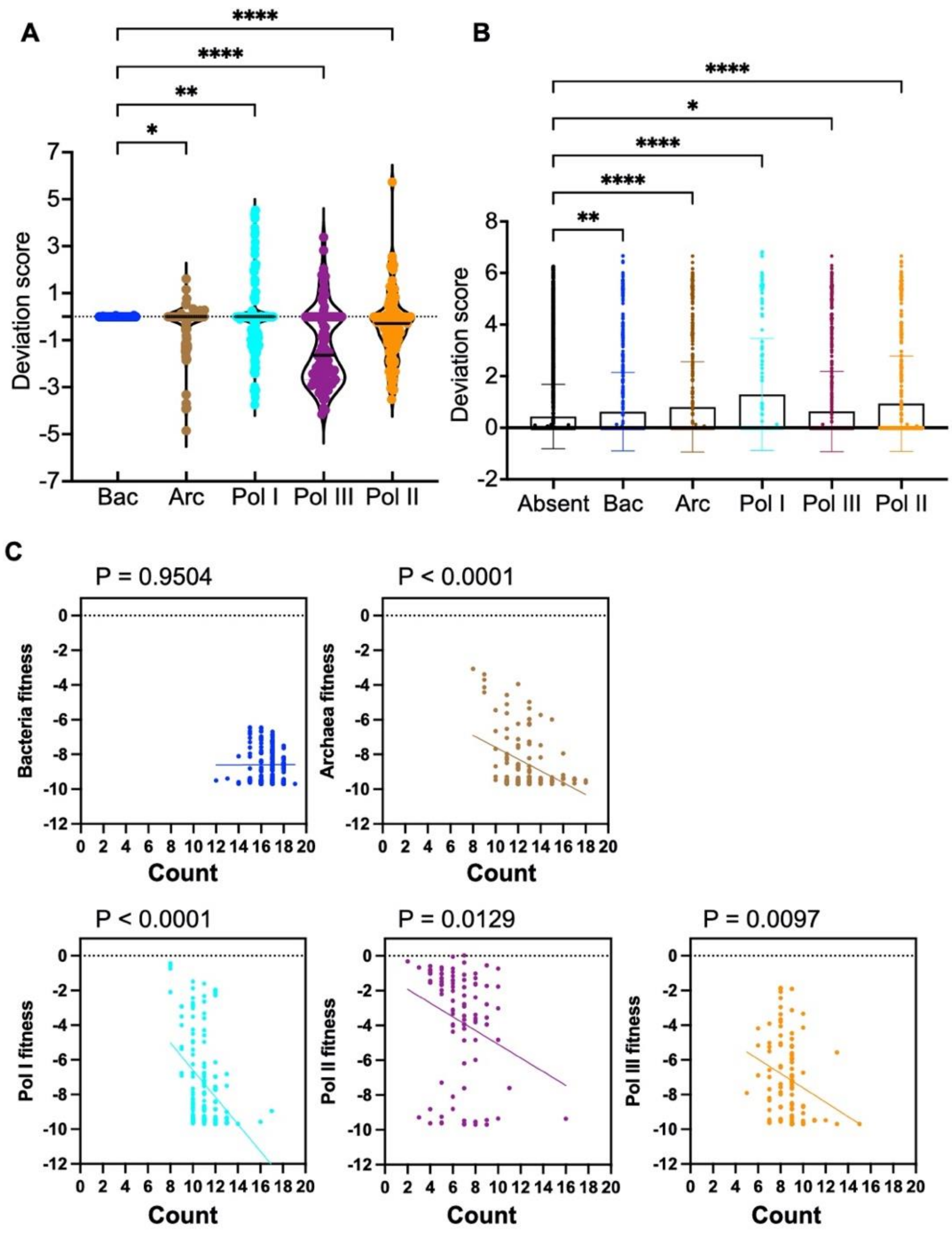


Figure 19. Contextual epistasis affects fitness of TL haplotypes.

(**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.

2.2.6 TL residues co-evolve with the rest of Rpb1 through diverse pathways

Our analyses suggest that even a highly conserved domain such as the Pol II TL can be sensitive to identity of adjacent residues and that changing networks of interactions shape the Pol II active site across evolution. We employed statistical coupling analysis (SCA) to identify if there are any coevolving residue networks in Rpb1 and ask about pathways that might co-evolve the TL. SCA “Sector” analysis is especially useful for identifying subgroups of coupled residues that might form allosteric communication networks (Halabi, Rivoire, Leibler, & Ranganathan, 2009; Rivoire, Reynolds, & Ranganathan, 2016; Salinas & Ranganathan, 2018). We extracted 410 yeast Pol II Rpb1 sequences from the recently published msRNAP large subunit multiple sequence alignment (MSA) from the Landick lab (Palo et al., 2021) and performed SCA (see **2.4 Methods**)(Rivoire et al., 2016). We identified 40 coevolving sectors (**Figure 20**), and every single TL residue was found within one of the eight sectors that form generally continuous network of interactions within Rpb1(**Figure 21**). TL residues within the TL NIR were coupled with most BH residues and the alanine-glycine linker (Rpb1 1087-1088). These residues are highly conserved (Qiu et al., 2016), indicating this sector is driven by conservation primarily. Six of eight Rpb1 sectors containing TL residues also contained at least one BH residue, supporting functional coupling between these two domains. Coupling is not limited to residues that are close to the active site. Distal residues can potentially modulate TL function through allosteric interactions. For example, the greatest distance between a TL residue and another Rpb1 residue in the same sector is ~ 55 Å. Interestingly, the residue pair 1076-1101, for which we observed extensive epistasis, are the sole TL residues within a very large cluster containing >150 residues across Rpb1. Our epistasis studies indicate multiple allele-specific interactions between 1076 and 1101 of exactly the type that might appear as evolutionary coupling between specific substitutions at these positions. The hydrophobic TL pocket is an attractive linchpin for potential communication to the TL from throughout Pol II, and multiple sectors converge on this domain.

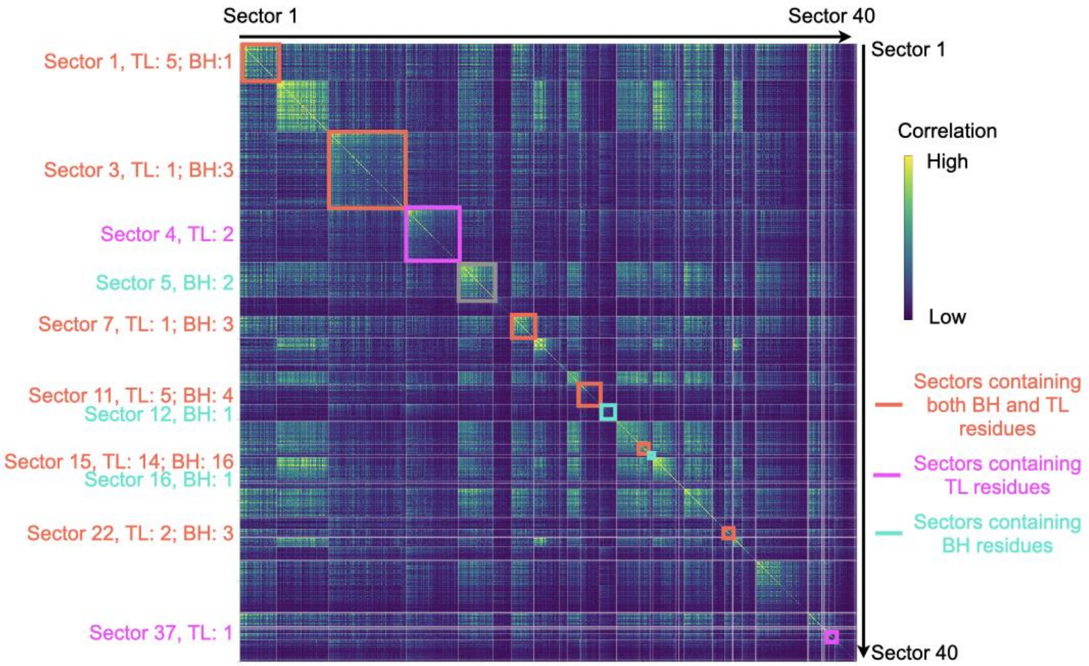


Figure 20. Rpb1 coevolutionary residue networks identified by Statistical Coupling Analysis (SCA). 40 significant and independent sectors are shown in a heatmap with correlation score calculated from the statistical coupling analysis. Sectors containing TL and BH residues are labeled. Numbers of TL and BH residues contained in each sector are labeled on the left of the heatmap. Statistical coupling analysis was applied to a published Multiple Sequence Alignment (MSA) of Rpb1 homologs (n= 410) (Palo et al., 2021). Details are in **2.4 Methods**.

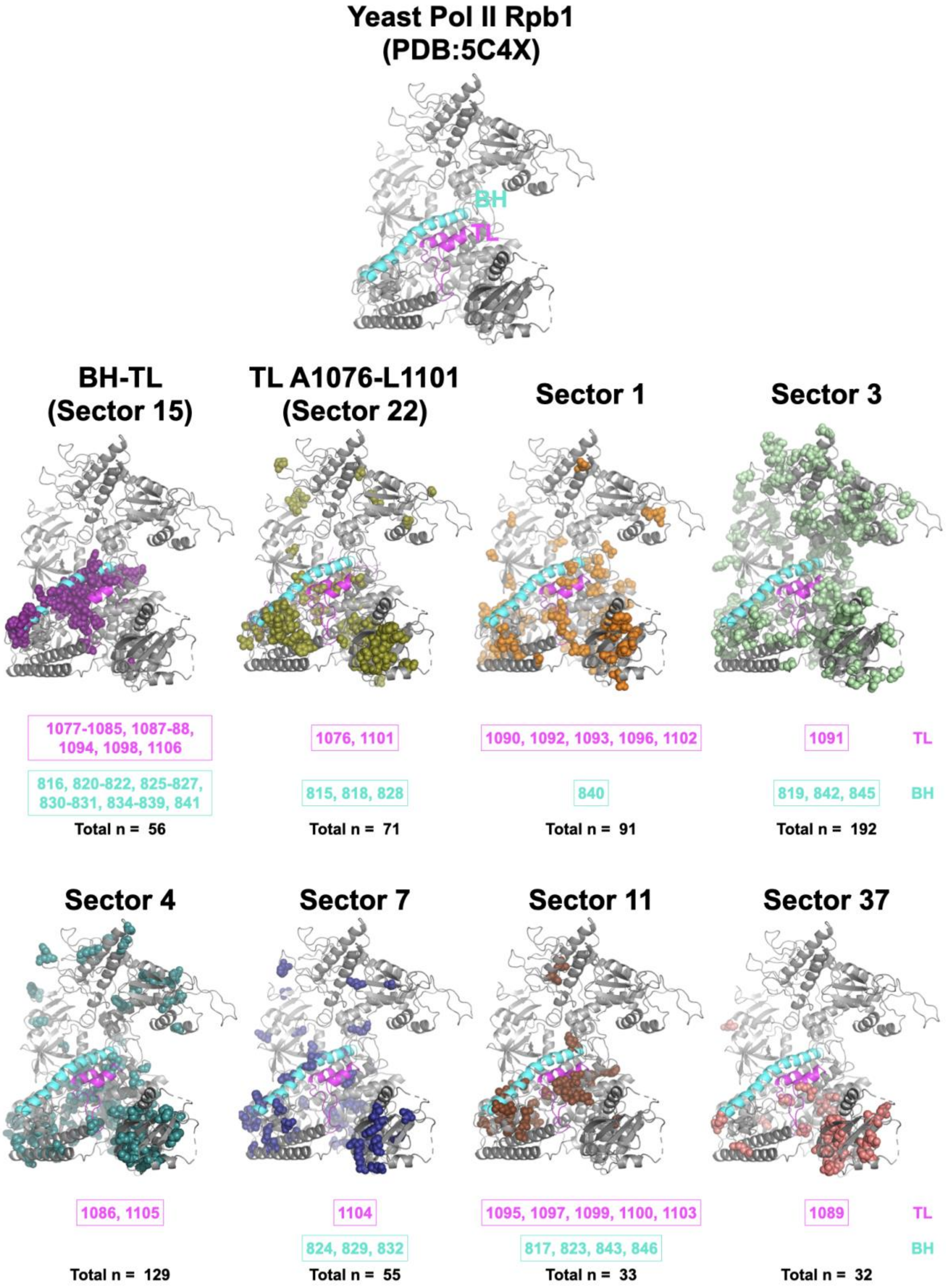


Figure 21. Ultra-conserved TL co-evolves with Pol II residues through diverse pathways.

The eight coevolution sectors containing any TL residues that were identified from statistical coupling analysis are shown on the yeast Pol II Rpb1 structure (PDB:5C4X)(Barnes et al., 2015). The TL is labeled with magenta and the BH is labeled with cyan. The TL and BH residues in each sector are labeled at the bottom of each sector. The total number of residues within each sector is also shown. The details of the statistical coupling analysis are in **2.4 Methods**.

2.3 Discussion

How individual mutants alter a protein's function is not necessarily straightforward at the mechanistic level. Amino acid substitutions both remove functionality of the WT residue but replace that functionality with something different. By altering the local environment within a protein or potentially propagating effects to distant locations through allosteric changes, each substitution potentially can be quite different. These differences may not be apparent as phenotypic outputs and phenotypic assays may not have granularity to distinguish different biophysical behaviors if they result in similar outputs. For Pol II mutants, even high-resolution phenotypic analyses, such as gene expression profiling or genetic interaction profiling between Pol II mutants and deletions in other yeast genes (Braberg et al., 2013), suggest that LOF and GOF mutants represent a continuum of defects that match enzymatic activity *in vitro*. Therefore, these profiles also appear dependent on the output of Pol II activity defects and can't distinguish potential differences in underlying mechanism.

Through systematic detection of genetic interactions within the Pol II active site, we have identified functional relationships between amino acids across the TL and between TL substitutions and others. In the absence of double mutant epistasis analyses it would not be possible to differentiate similar alleles from one another. L1101S and E1103G, for example, are two GOF alleles very close to each other in Pol II structure and confer similar phenotypic landscapes across various growth conditions. Here, we find that their distinct interactions support that substitutions at 1101 and 1103 target distinct residue networks (**Figure 11C, 12, 13, and 16A**). 1101 functions in the five-helix bundle hydrophobic pocket while 1103 interacts and co-evolves with a number of TL external residues that together support interactions that promote the open TL conformation (**Figure 16C and 21**). We also observed connections between TL C-terminal residues that suggest a limit to how disruptions to structure there can alter Pol II activity (**Figure 9D and 21**). Helix-disrupting LOF proline substitutions in at least two TL positions showed epistasis with multiple substitutions in the back of the TL (1094-1098), suggesting that their functions require TL C-terminal helix structure and in the absence of that structure (proline disruption) effects are no longer additive.

The strongest epistatic interactions observed were between two pairs of hydrophobic residues, A1076 and L1101, and M1079 and G1097 (**Figure 16**). Each of these contributes to the structure of a hydrophobic pocket that bundles two TL proximal helices with the BH and two others in a five-helix bundle. Supporting the dependence of these residues on each other for maintaining function, identity at these positions over evolution also shows coupling. Interestingly, these A1076 and L1101 were coupled uniquely out of TL residues with a great number of other positions in Rpb1 (**Figure 21**).

Elongation factors bind Pol II and alter its activity, but the mechanisms by which they do so are not known (Cramer, 2019b; Schier & Taatjes, 2020). 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 (**Figure 11**). 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(Vos et al., 2020)). These contacts have been proposed to alter Pol II activity. We would propose that the paths for such alteration activity would follow the coupling sectors we have observed by SCA.

How different individual substitutions are under the surface is critical for understanding plasticity in protein mechanisms and how they might be altered by evolutionary change. A key open question in nucleic acid polymerase mechanisms is the paths for protons in the reaction (for example, deprotonation of the synthesized strand 3′-OH and protonation of pyrophosphate leaving group) (e.g.(Belogurov & Artsimovitch, 2019; Carvalho et al., 2011; Castro et al., 2007; Gregory, Gao, Cui, & Yang, 2021; X. Huang et al., 2010; Palo et al., 2021; Unarta et al., 2023)). For msRNAPs, the association with incoming NTP by a nearly universally conserved histidine led to the proposal that this residue might donate a proton during the reaction (Castro et al., 2007; Castro et al., 2009; D. Wang et al., 2006). 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 (Qiu et al., 2016) and the Landick lab has proposed this substitution supports catalysis through positional but not chemical effects (Mishanina et al., 2017; Palo et al., 2021). Our studies here were quite surprising in that they indicated that H1085L 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) (**Figure 15C and 22**). 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 (Park et al., 2022). 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 shows high identity even between distantly related species. Here we show that the TL, and likely the entire Pol II active site, exhibits a great amount of plasticity through non-conserved positions that are essential for compatibility of the TL and surrounding domains. Our results illustrating widespread epistasis and allele-specific effects of single and double mutants predict that comparative analyses among Pol I, II, and III will reveal widespread and enzyme-specific mechanisms due to higher order epistasis shaping function of conserved residues.

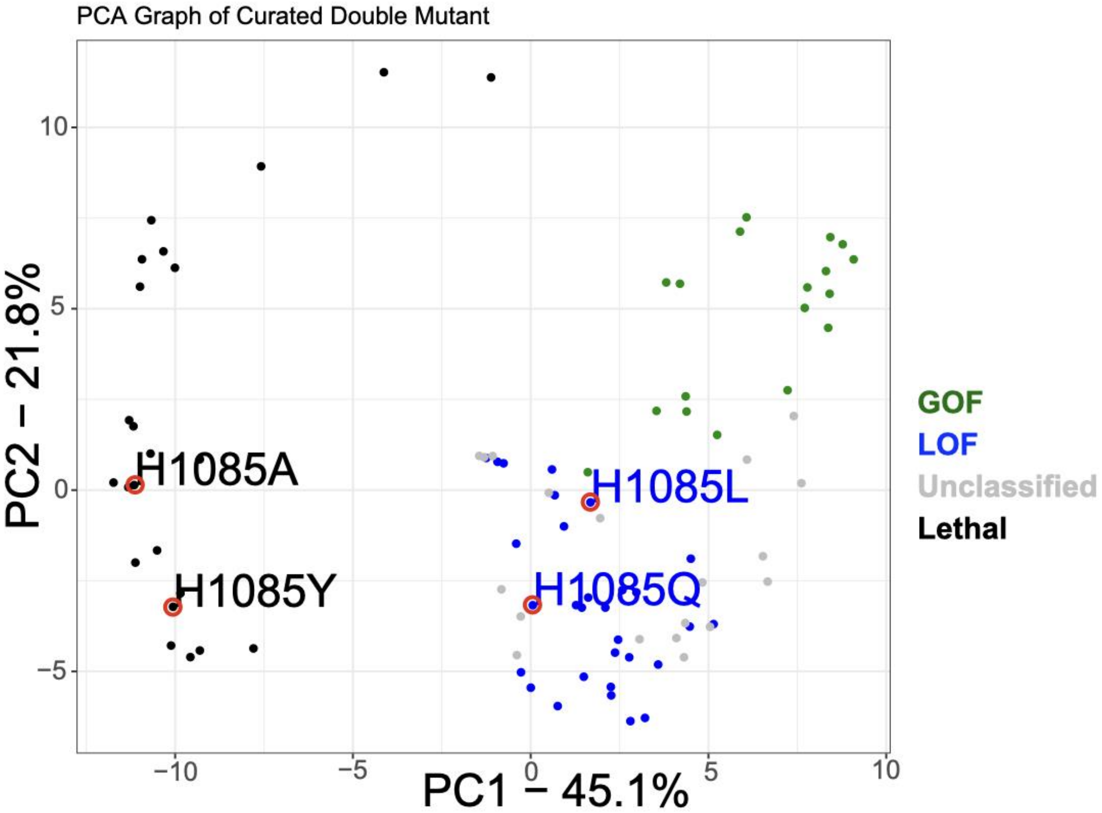


Figure 22. Four H1085 substitutions are different in some ways.

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

2.4 Methods

2.4.1 Design and Synthesis of TL mutant libraries

We updated and extended the fitness dataset of Qiu et al (Qiu et al., 2016). Using a similar methodology, but with adjusted conditions and a second-generation mutant library strategy, in order to generate a complete Pol II TL mutation-phenotype map and examine genetic interactions. Mutants were constructed by synthesis with Agilent and screened for phenotypes previously established as informative for Pol II mutant biochemical defects. Programmed oligonucleotide library pools included all 620 single TL residue substitutions and deletions for Rpb1 amino acids 1076-1106 (Library 1), 3,914 pairwise double substitutions (Library 2), 4,800 targeted double substitutions (Library 6), and 3,373 multiple substitutions (Library 3-5), along with the WT S. cerevisiae Pol II TL allele at a level of ~15% of the total variants, enabling precise quantification (see **Appendix Table 4**). Each synthesized region contained a mutated or WT Pol II TL sequence and two flanking regions at the 5′ and 3′′ ends of the TL-encoding sequence. These flanking regions also contained designed "PCR handle" (20bp) sequences, allowing distinct subsets of oligos to be amplified from synthesized pools using selected primers for PCR, and additional flanking WT Pol II sequences allow for further extension of homology arms by PCR "sewing" (Details are in **Appendix A2.2 and A2.3**).

2.4.2 Introduction of Libraries into yeast and phenotyping

Synthesized mutant pools were transformed into yeast (CKY283) along with an RPB1-encoding plasmid where the TL-encoding sequence was replaced with an MluI restriction site for linearization as described in Qiu et al (Qiu et al., 2016). This strategy allows construction of rpb1 mutant libraries by gap repair between library fragments and the linearized vector. Briefly, the synthesized oligo pools were amplified by limited cycles of emulsion PCR to limit template switching. Extension of flanking homology arms of ~200 bp were added by PCR sewing. Amplified TL sequences with extended flanking regions were co-transformed with linearized pRS315-derived CEN LEU2 plasmid (pCK892) into CKY283, allowing gap repair via homologous flanking regions. To detect potential residue-residue interactions between the TL and TL-proximal domains including the Rpb1 Bridge Helix (BH), Funnel Helix alpha-21 and Rpb2, the Pol II TL single mutant pool (Library 1, 620 mutant alleles and 111 WT alleles) was co-transformed individually with gapped plasmids encoding an additional rpb1 allele (Rpb1 BH T834P, T834A, or Funnel Helix alpha-21 S713P) into CKY283 respectively, or with the gapped WT RPB1 plasmid into a strain with the genomic mutation, rpb2 Y769F. These co-transformations created double mutants between the TL and TL-proximal mutants. The WT allele in single mutant pool represented the single probe mutant due to substitutions outside the TL on the plasmid or in the strain background. To distinguish between a fully WT TL and a WT TL representing the TL of a mutant allele elsewhere, a WT Pol II TL allele with a silent mutant at T1083 (WT codon ACC was replaced with ACT) was co-transformed with plasmid containing gapped WT RPB1 in a WT strain in parallel. 15% of the transformants with silent mutation were mixed with transformants of double mutants. The silent mutation allowed us to distinguish the WT and the single mutants. Each transformation was done in three biological replicates. After transformation, Leu+ colonies were collected from SC-Leu plates by scraping into sterile water and replated on SC-Leu+5FOA to select for cells having lost the RPB1 URA3 plasmid. 5-FOA-resistant colonies were scraped into sterile water from SC-Leu+5FOA and replated on SC-Leu, SC-Leu + 20mg/ml MPA (Fisher Scientific), SC-Leu + 15 mM Mn (Sigma), YPRaf, YPRafGal, SC-Lys, and SC-Leu + 3% Formamide (JT Baker) for phenotyping. Details of cell numbers plated on each plate and screening time of each plate are in **Appendix Table 3**. Details of high efficiency transformation protocol is in **Appendix A2.1**.

2.4.3 Generation of libraries for quantification by amplicon sequencing

Genomic DNA of each screened library was extracted using the Yeastar genomic DNA kit according to manufacturer's instructions (Zymo Research). To ensure adequate DNA for sequencing, the TL regions of all libraries were amplified with PCR cycles that were verified to be in the linear range by qPCR to minimize disturbance of allele distributions, and under emulsion PCR conditions (EURx Micellula DNA Emulsion & Purification (ePCR) PCR kit) to limit template switching. Details are in Appendix A2.2 and 2.3. To multiplex samples, we employed a dual indexing strategy wherein 10 initial barcodes for differentiating 10 mutant libraries were added during the initial amplification using 10 pairs of custom primers. In a second amplification, 28 primers containing 28 NEB indices were used to add a second index for distinguishing conditions and replicates (NEBNext Multiplex Oligos for Illumina) (see **Appendix Table 2**). As a result, a sample-specific barcodes were present for each set of variants. The indexed, pooled samples were sequenced by single end sequencing on an Illumina Next-Seq (150nt reads). On average, over 11 million reads were obtained for individual samples with high reproducibility from two rounds of sequencing. Raw sequencing data has been deposited on the NCBI SRA (Sequence Read Archive) database under BioProject PRJNA948661. Processed mutants counts and fitnesses are available through GitHub (https://github.com/Kaplan-Lab-Pitt/TLs_Screening.git).

2.4.4 Data cleaning and fitness calculation and normalization

Reads of mutants were sorted into appropriate libraries and conditions by detecting particular indices after sequencing. Read counts were estimated by a codon-based alignment algorithm to distinguish reads that exactly matched designated codons of mutants (Sing-Hoi Sze, 2018). To clean the data, mutant reads with coefficients of variation greater than 0.5 in the control condition (SC-Leu) were excluded from the analysis. The mutant read count was increased by 1 to calculate the allele frequency under different conditions. To measure and compare the phenotypes of all mutants, mutant phenotypic score (fitness) was calculated by allele frequency change of a mutant under selective conditions relative to the unselective condition comparing to the frequency change of WT. The formula for calculating fitness is shown below.

Fitness (mut) = log [$f$mut, sele / $f$mut, unsele - log [fWT, sele / fWT, unsele]

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

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

2.4.5 Determination of functional interactions

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

Deviation score (M1M2M3) = Fitness (M1M2M3) – [Fitness (M1) + Fitness (M1) + Fitness (M3)]

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

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

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

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

Details of formulas are in Appendix A2.4. The codes for calculating deviation scores and generating figures are available in GitHub (https://github.com/Kaplan-Lab-Pitt/TLs_Screening.git).

2.4.6 Computational analysis

2.4.6.1 Mutant classification using two multiple logistic regression models

We trained two multiple logistic regression models to distinguish GOF and LOF mutants using the phenotypic fitness on SC-Leu+MPA, SC-Lys, and YPRafGal conditions of 65 single mutants, including 25 previously identified GOF mutants, 33 LOF mutants, one WT, and six that were not GOF or LOF mutants. Intercept, main effects, and two-way interactions were involved in defining both models. 0.75 was used as the cutoff threshold for both the GOF and LOF models.

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

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

Model for predicting the probability of a mutant being LOF:

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

Both models showed accuracy, with the area under ROC close to one **(Figure 8A)**. The details are provided in **Appendix Table 5**.

2.4.6.2 Principal component analysis (PCA)

Deviation scores of curated and probe double mutants were analyzed in PCA. The scripts using R language v4.0.3 (https://www.R-project.org/) with R packages tidyverse v1.3.1 (https://www.tidyverse.org), prompt from R stats package v3.6.2 (https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/prcomp), ggplot2 v3.3.3 (https://ggplot2.tidyverse.org), dplyr v1.0.6 (https://dplyr.tidyverse.org), and missMDA v1.18

(https://dplyr.tidyverse.org), are available in GitHub (https://github.com/Kaplan-Lab-Pitt/TLs_Screening.git).

2.4.6.3 t-SNE projection

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

2.4.6.4 Statistical coupling analysis

A published multiple sequence alignment (MSA) containing 5787 eukaryotic homologous sequences of yeast Rpb1 was used in the statistical coupling analysis (Palo et al., 2021). 1464 sequences were retained after sequence identity reducing to 90% with T-coffee package v12.00.7fb08c2 (Notredame, Higgins, & Heringa, 2000) through conda v4.6.14. Pol I, II, and III sequences were separated based on an ML tree constructed with FastTree 2 (Price, Dehal, & Arkin, 2010) and 410 Pol II Rpb1 homologous sequences were re-aligned with T-coffee, and the newly generated MSA was used for statistical coupling analysis with the python-based package pySCA v6.1 (Rivoire et al., 2016). The scripts were adapted from https://github.com/ranganathanlab/pySCA and are available via GitHub (https://github.com/Kaplan-Lab-Pitt/TLs_Screening.git).

3.0 Higher-order epistasis within TL haplotypes

3.1 Introduction

Functional interactions among amino acid residues within a protein can be detected by genetic interactions in the form of epistasis. Epistasis can determine the phenotypic effects of mutations. Specific substitutions can constrain or relax functional constraints on the identity of residues at other positions, and impact evolutionary trajectories (Domingo et al., 2019; Johnson et al., 2023; Mani et al., 2008; Metzger, Park, Starr, & Thornton, 2023; Park et al., 2022; Phillips, 2008; Starr & Thornton, 2016). Epistasis is the modulation, positive or negative, of the effect of one mutant by the presence of other mutants. Phenotypic modulation of one mutant by another (epistasis) can be detected by a difference in the phenotype of a double mutant from that expected from the cumulative effects of the corresponding single mutants, which is the baseline expectation in the absence of epistasis (Hill et al., 2008; Mani et al., 2008; Phillips, 1998, 2008). Prior studies found many epistatic interactions at both lower-orders (i.e. among residue pairs) and higher-orders (among many residue combinations). Higher-order epistasis emerging from particular residues can reflect specific biological or physical interactions and imply crucial roles of these residues in a protein's function and evolution (Bakerlee et al., 2022; Bank et al., 2015; Ding et al., 2022; X. Lin et al., 2022; Melamed et al., 2013; Olson et al., 2014; Pokusaeva et al., 2019; Reddy & Desai, 2021). In this work, we dissect complex higher-order epistasis by deep mutational scanning within an ultra-conserved and crucial active site domain, the trigger loop (TL), of RNA polymerase II (Pol II).

Pol II transcribes eukaryotic mRNA using an iterative nucleotide addition cycle (NAC) (Bar-Nahum et al., 2005; Dangkulwanich et al., 2013; Kaplan, 2013; Malinen et al., 2012; D. Wang et al., 2006). The TL is a critical domain for the NAC, and participates in all three steps: nucleotide selection, catalysis, and likely translocation by switching between different conformations (Barnes et al., 2015; Fouqueau et al., 2013; D. Wang et al., 2006). In each NAC, a matched substrate entering the active site induces or captures a conformational change of the TL from an open, catalysis-disfavoring state to a closed, catalysis-favoring state. Thus, the TL functions in kinetic selection to distinguish matched NTPs complementary to the DNA template from mismatched ones (Barnes et al., 2015; X. Huang et al., 2010; Kaplan, 2013; Malinen et al., 2012; B. Wang et al., 2013; D. Wang et al., 2006; L. Xu et al., 2014). Full closure of the TL promotes catalysis (Vassylyev et al., 2007; B. Wang et al., 2013). Following catalysis, the TL transits from the closed to the open state, facilitating polymerase translocation for the subsequent NAC (Da et al., 2012; B. Liu et al., 2016; Seibold et al., 2010)(**Fig**ure **23A**). Additional TL confirmations have been associated with other RNA polymerase activities such as pausing and backtracking (Cheung & Cramer, 2011; Mosaei & Zenkin, 2021; D. Wang et al., 2009; J. Zhang et al., 2010). With these transitions, the mobile TL balances transcription fidelity and speed (Kaplan, 2010; Kaplan et al., 2008; Kireeva et al., 2008; Larson et al., 2012; Sydow et al., 2009; Unarta et al., 2023; D. Wang et al., 2006; Yuzenkova et al., 2010). How do residues within the TL interact with each other to ensure its proper transition? Our observations of distinct pairwise interactions suggest that TL function is maintained by residue interaction networks (Duan, Qiu, Sze, & Kaplan, 2023). For example, previous biochemical and genetics studies found mutations in the TL can modify its folding or dynamics, causing changes in catalytic activity - either an increase, leading to faster elongation rate than WT *in vitro* (gain of function, GOF) or a decrease, leading to slower elongation rate than WT *in vitro* (loss of function, LOF) (Barnes et al., 2015; Braberg et al., 2013; Kaplan et al., 2012; Kaplan et al., 2008; Kireeva et al., 2008; Nayak et al., 2013; Qiu et al., 2016; Windgassen et al., 2014). With TL double mutants including combinations between or within GOF and LOF mutations, our prior work identified distinct types of pairwise interactions resulting in either better than expected fitness from the cumulative model, i.e. suppression, commonly observed in GOF/LOF combinations, or worse than expected, i.e. synthetic sickness or lethality, normally seen in combinations of the same class (GOF/GOF or LOF/LOF). These interactions are consistent with a model that the involved single mutants independently act within double mutants, where effects are balanced in double mutants comprising constituent mutants with opposite effects and effects are enhanced in double mutants comprising constituent mutants with similar effects (Duan et al., 2023; Kaplan et al., 2012; Qiu et al., 2016; Qiu & Kaplan, 2019). However, we also observed both lack of enhancement for some "same class" combinations (epistasis), and cases where a mutant's effect appeared to be controlled by the identity of another residue (sign epistasis) (Duan et al., 2023; Kaplan et al., 2012; Qiu et al., 2016; Qiu & Kaplan, 2019).

The TL is highly conserved across evolution, supporting the idea that identities of key residues are highly constrained (Palo et al., 2021). Given this conservation, it was surprising that analogous mutations in a highly conserved residue yielded opposite effects in the conserved yeast Pol I and Pol II (Viktorovskaya et al., 2013). Given the TL's flexible and mobile character, and its high conservation of sequence and function, it was an open question of how self-contained it is functionally and if there is evolutionary coupling through non-conserved residues to maintain its ability to achieve its different conformations. Our recent studies identified widespread incompatibility between trigger loops of different species or enzymes placed into the yeast Pol II context, supporting the idea of an epistasis network from residues outside the TL is prominent in constraining TL function across evolution (Duan et al., 2023). We wished to understand the drivers of this incompatibility between conserved TLs and use diverse TL haplotypes to probe the complexity of internal TL interactions.

Recent studies have suggested that the primary contributor of epistasis to protein function is through pairwise residue interactions. Higher order epistasis is complicated and can be difficult to detect (Johnson et al., 2023; Metzger et al., 2023). We reasoned that the Pol II TL might be an excellent system to detect higher order epistasis using evolutionary haplotypes, notwithstanding constraints on function from the rest of the enzyme complex. The Pol II TL is flexible, being required to support distinct conformational states, and functionally plastic in that mutants can both increase or decrease catalysis with known suppressor relationships between these two types of mutants. We applied our TL deep mutational scanning system to detect higher order epistasis by comprehensive dissection of a set of evolutionarily derived TL sequences (haplotypes). We identified intra-TL interactions of evolutionarily observed substitutions (TL-internal epistasis) and dissected specific examples of complex interactions that would not be evident from simple analyses only considering single mutants or complete haplotypes. Our experiments suggest specific paths for TL evolution in the context of the complete enzyme may likely go through mild gain-of-function residues that allow additional changes to be tolerated.

3.2 Results

3.2.1 Systematic detection of epistatic interactions within TL haplotypes

We utilized our previously developed deep mutational scanning-based phenotyping platform for Pol II mutants in *Saccharomyces cerevisiae* (Duan et al., 2023; Qiu et al., 2016) to analyze TL-internal higher-order epistasis. The analyzed TL haplotypes included 662 natural TL variants from bacterial and archaeal msRNAPs, and eukaryotic Pol I, Pol II and Pol III (Fig**ure 23B-C**). An additional 1987 TL alleles were included, representing all possible intermediate substitution combinations for seven selected natural Pol II TL variants (Fig**ure 23D**). The seven TL variants were selected because they contain specific amino acids which exhibit phenotypes, either GOF or LOF, when introduced individually in yeast Pol II. The functional consequences of these haplotypes were measured by a set of growth phenotypes that are predictive of biochemical functions (see 3.4 Methods) (Duan et al., 2023; Kaplan et al., 2012; Qiu et al., 2016; Qiu & Kaplan, 2019). Our system allows us to profile these haplotypes and assess epistasis among multiple mutations in a highly parallel fashion.

To detect epistasis, we first fitted a log additive model for the fitness of individual mutations, which assumes all the individual mutations are independent and additive (Duan et al., 2023; Hill et al., 2008; X. Lin et al., 2022; Mani et al., 2008; Phillips, 2008) and we then computed how the observed fitness of haplotypes deviates from the predicted fitness by the log additive model. Here fitness is determined by comparing the allele frequency shifts of a mutant under selective conditions relative to control conditions, and then comparing the allele frequency change of mutants to the change observed in the WT (see **3.4 Methods**). The log of the fitness is defined as a fitness score. The log additive model assumes individual mutations are independent, so the fitness effects of individual mutations are multiplied in the double/multiple mutants (the log of the fitness scores are additive). Deviation from the log additive model is evidence for residue interactions (epistasis). Deviation from expectation could be in the form of mutual suppression (double mutant more fit than either single mutant), basic epistasis (double mutant no worse than one of the single mutants), or synthetic sickness (greater than log additive defects in the double mutant). The deviation values, termed “deviation scores”, were calculated to detect intra-TL genetic interactions within haplotypes (see 3.4 Methods).

Deviation scores were calculated in two ways. We determined a “primary deviation” score, representing total epistasis within a haplotype as the deviation of the predicted fitness from the sum of the fitness scores of individual mutants (Fig**ure 23E** and **Figure** 23). Second, to determine which residues the epistasis emerged from, we separated the haplotypes into combinations of single substitutions plus the compound mutant formed by the remaining substitutions. We therefore calculated a secondary deviation score by comparing the fitness score of the complete haplotype to the sum of the single substitution’s fitness score and the fitness score of the compound consist of the remaining substitutions (Fig**ure 23F** and **Figure 24**). The magnitude of the secondary deviation score indicates the magnitude of epistatic effects for specific substitutions when they are introduced to the compound mutant. Moreover, analyzing the secondary deviation scores of a specific substitution across various compound mutants (backgrounds) enables us to assess the breadth of its epistatic effects, which may reveal its potential to impact TL evolution.

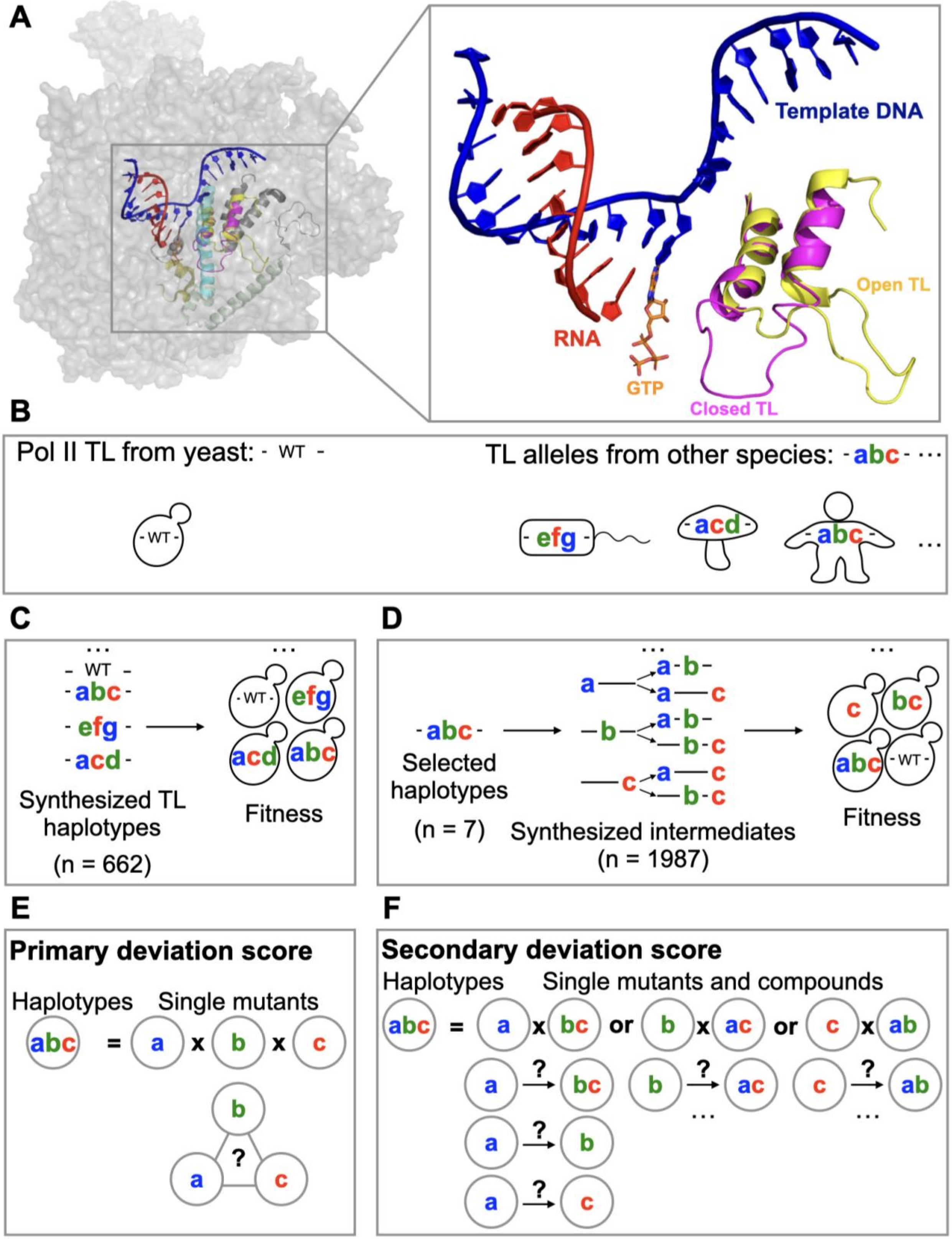
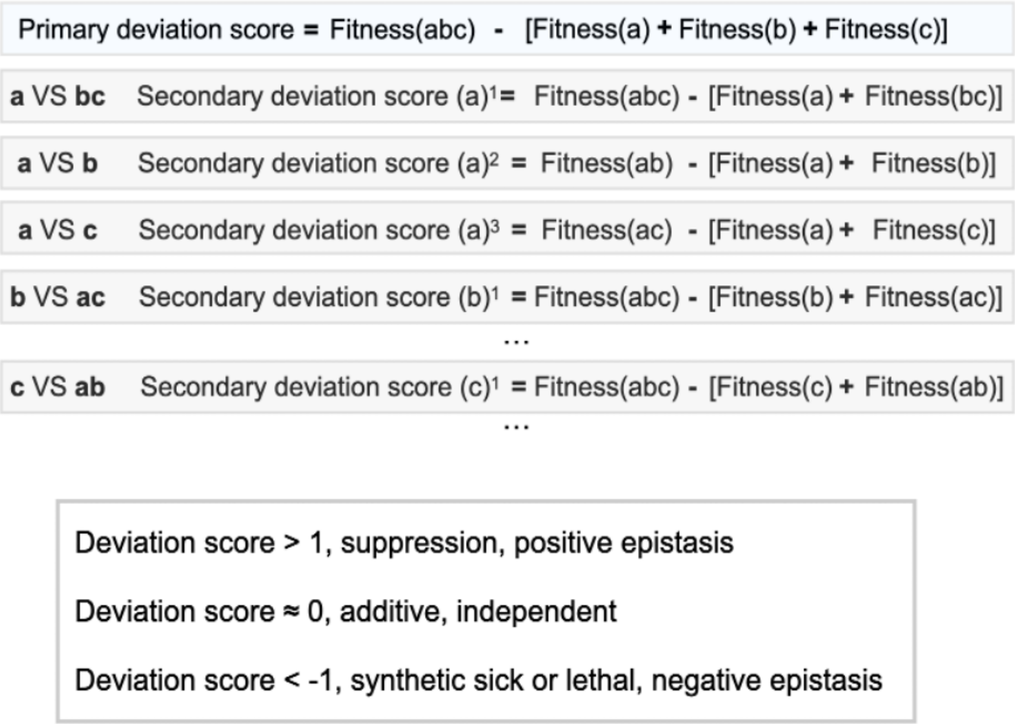


Figure 23. Systematic detection of TL-internal epistasis with natural TL alleles and intermediates.

(**A**).We selected 662 TL haplotypes representing TL alleles from bacterial, archaeal and the three conserved eukaryotic msRNAPs to detect TL-internal epistasis. (**B**). The selected TL alleles were synthesized and transformed into yeast Pol II to form chimeric Pol II enzymes. Yeast chimeric Pol II enzymes were phenotyped under selective conditions to detect growth defects, which are represented by fitness (see **3.4 Method**). (**C**). Seven Pol II TL alleles were selected to construct intermediate haplotypes representing all possible combinations of substitutions of seven TL alleles. The intermediates were transformed into yeast to measure growth defects as in (**C**). (**D-E**). Analytical scheme of primary deviation score (**D**) and secondary deviation score (**E**). Details are in **3.4 Method**.



Primary deviation score = Fitness(abc) - [Fitness(a) + Fitness(b) + Fitness(c)]

a VS **bc** Secondary deviation score $(a)^1$ = Fitness(abc) - [Fitness(a) + Fitness(bc)]

a VS **b** Secondary deviation score $(a)^2$ = Fitness(ab) - [Fitness(a) + Fitness(b)]

a VS **c** Secondary deviation score $(a)^3$ = Fitness(ac) - [Fitness(a) + Fitness(c)]

b VS **ac** Secondary deviation score $(b)^1$ = Fitness(abc) - [Fitness(b) + Fitness(ac)]

…

c VS **ab** Secondary deviation score $(c)^1$ = Fitness(abc) - [Fitness(c) + Fitness(ab)]

…

Deviation score > 1, suppression, positive epistasis

Deviation score ≈ 0, additive, independent

Deviation score < -1, synthetic sick or lethal, negative epistasis

Figure 24. Detection of higher-order interactions by primary and secondary deviation scores.

For a pseudo haplotype “abc”, the primary deviation score is calculated by comparing the observed fitness of the haplotype to the log additive of constituent substitutions’ fitness. The secondary deviation score of “a” in “abc” represents the epistatic effect of “a” on “bc” and is calculated by comparing the observed fitness score of “abc” to the additive of fitness scores of “a” and “bc”. The secondary deviation score of “a” in “ab” represents the epistatic effect of “a”on “b”. It is calculated by comparing the observed fitness score of “ab” to the additive of fitness scores of “a” and “b”. If the deviation score ≈ 0, it indicates the constituent substitutions are independent and there is no interaction among them. If the deviation score > 1, it represents positive epistasis (suppression) among the constituent substitutions. If the deviation score < -1, it represents negative epistasis (synthetic sickness or lethality) among the constituent substitutions.

3.2.2 Drivers of incompatibility for TL variants placed in the yeast Pol II context

We explored potential reasons for the widespread incompatibility between TL haplotypes from other msRNAPs and yeast Pol II (Duan et al., 2023; Qiu et al., 2016; Viktorovskaya et al., 2013). To probe this incompatibility, we grouped haplotypes based on their evolutionary source (Bacteria, Archaea and eukaryotic Pol I, Pol II and Pol III) and illustrated the fitness of all constituent single substitutions measured in yeast Pol II (Duan et al., 2023; Qiu et al., 2016), the expected fitness of the haplotypes calculated based on the log additive model, the observed fitness of the haplotypes measured from our deep mutational scanning, and the primary deviation score from comparison of the observed and expected fitness (Fig**ure 25A**-E). The incompatibility of Bacterial TLs in yeast was explained by the presence of three substitutions that are individually lethal in yeast (Fig**ure 25A**). Interestingly, among archaeal TLs examined, only about half could be explained by individual lethal substitutions while others appeared lethal due to simple additivity of mutant phenotypes. A subset however appeared to illustrate negative epistasis among constituent substitutions (Fig**ure 25B**). Fewer individually lethal single substitutions were observed in eukaryotic Pol I, Pol III and Pol II haplotypes than those in Bacteria and Archaea, while more positive or negative intra-TL interactions appeared (Fig**ure 25C**-G), likely due to eukaryotic Pols being closer to Pol II in evolution. Surprisingly, a few single substitutions from other Pol II TL haplotypes were lethal in yeast Pol II context (Fig**ure 25D**), indicating that even in highly conserved Pol II enzymes, epistasis between the TL and its Pol II context is specific and important functional constraints have evolved among close homologs.

We then asked if functional coupling of residue identities within the TL across evolution might correlate with positive interactions (Fig**ure 26A-C**). This analysis tests the hypothesis that internal coupling of residues within the TL could be important for its function even outside of their coevolved, appropriate contexts. Statistical coupling analysis is a powerful approach to identify functional interactions by statistical inference of residue identities that appear to correlate across evolution (Russ, Lowery, Mishra, Yaffe, & Ranganathan, 2005; Socolich et al., 2005). We detected 11 residues involved in coupling within the TL domain using a multiple sequence alignment containing 362 eukaryotic TL sequences (natural TL variants) (Fig**ure 26A**). To test if statistical coupling within the TL is implicated in TL function and detectable for TL haplotypes removed from their natural context, we designed two libraries to perturb residue coupling in large scale, as previously described (Russ et al., 2005; Socolich et al., 2005). One library comprised scrambled haplotypes that retained TL-internal residue couplings and conservation as determined from natural TL variants ("Monte Carlo" scrambling library) (Fig**ure 26C**). A second library comprised scrambled haplotypes but only preserved conservation but not any TL-internal residue coupling ("Random scramble" library) (See 3.4 Methods) (**Figure 26B**). Both libraries satisfy the design parameters as indicated by residue distributions (Fig**ure 26A**-C). If substantial internal TL epistasis were present within the evolutionary signal, we would predict that haplotypes in the Monte Carlo scramble library would have greater fitness on average than those from Random scramble library. We did observe more Monte Carlo scrambled haplotypes in the viable range (fitness from -6.0 to 0) than the random scramble. The difference was subtle and not significant (P-value = 0.3013) (**Figure 26D**), suggesting that without positive epistasis between TL alleles and their appropriate coevolved contexts, TL-internal epistasis is weak or undetectable based on our tested couplings.

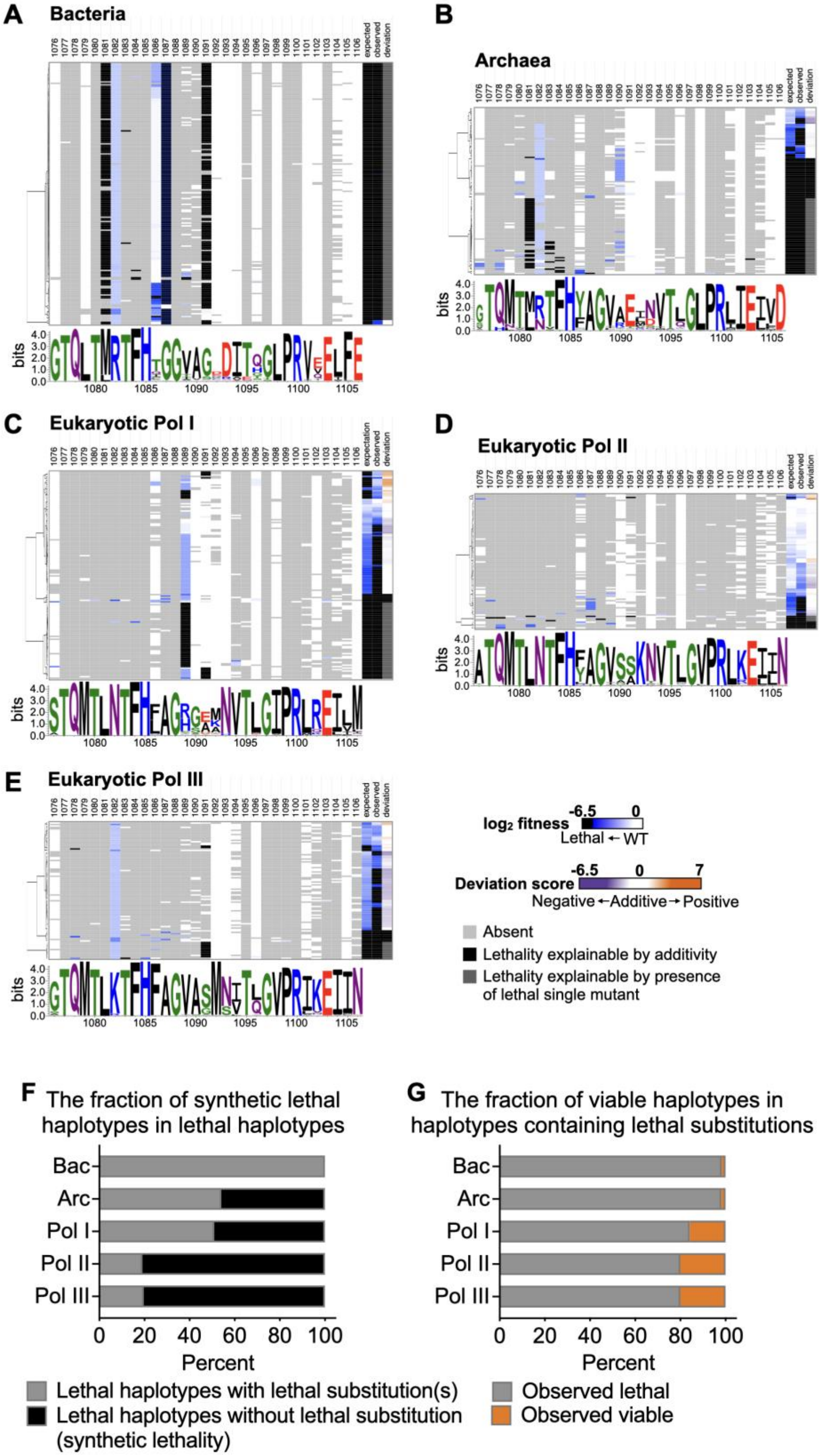
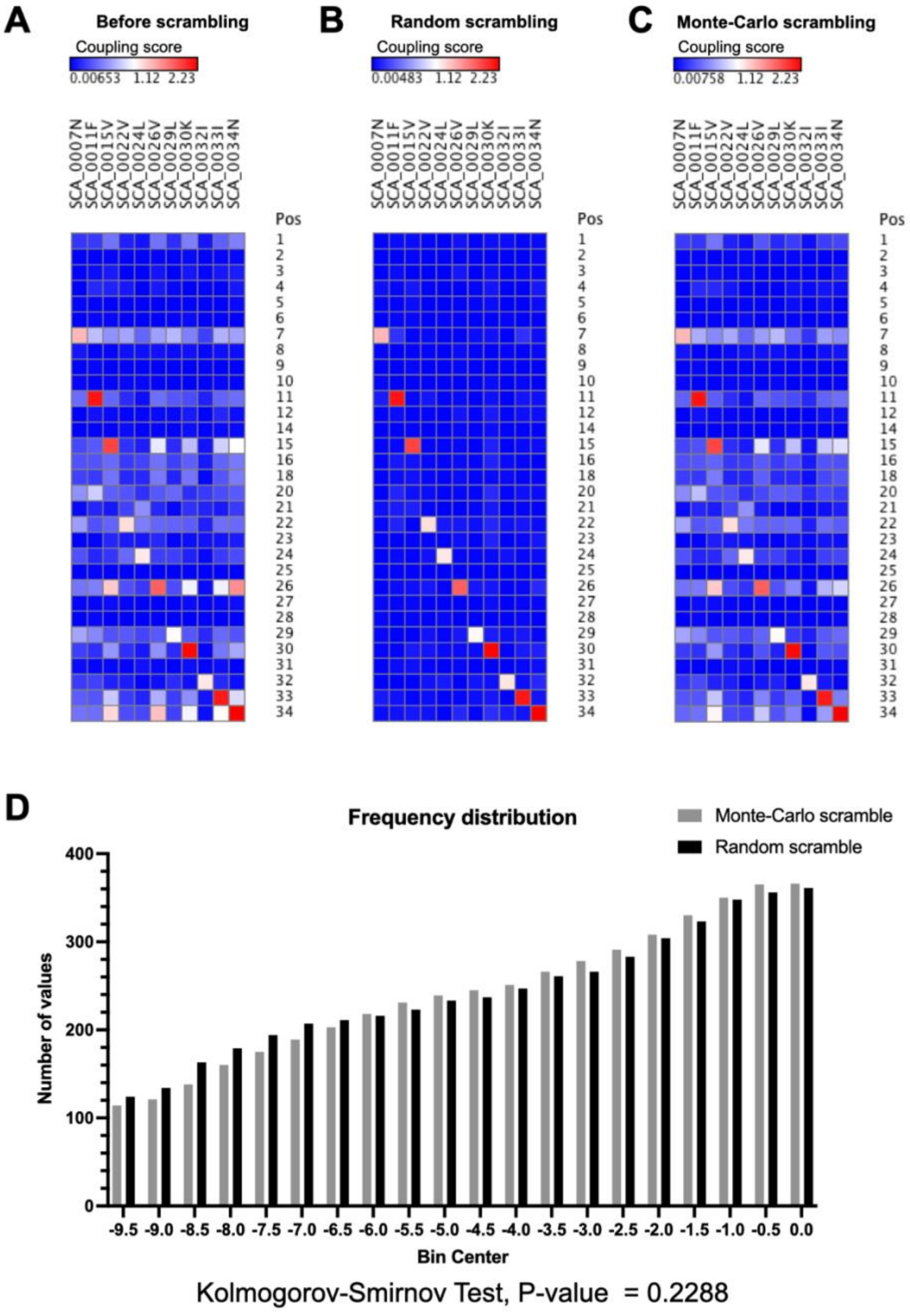


Figure 25. More TL-internal epistasis is observed with closer evolutionary distance to eukaryotic Pol II.

(**A-E**). Fitness and deviation score heatmaps of TL haplotypes in five evolutionary groups. The X-axis of each heatmap is 31 residue positions of Pol II TL (1076-1106). The Y-axis of each heatmap is the TL haplotypes belonging to each group clustered by hierarchical clustering with Euclidean distance. Each row represents one haplotype with several single substitutions. Light grey blocks in each row represent the residue from the haplotyope at the position is the same with yeast Pol II TL residue, in other words, no substitution at the position. Colored blocks represent different residues in the haplotype compared with yeast Pol II TL (substitutions). The color of the block represents growth fitness of the single substitution in yeast Pol II TL background. Expected fitness of the haplotypes were calculated from the log additive model. Observed fitness was measured in the screening experiments, and deviation scores were calculated by comparing the observed and expected scores. They are shown at the right end of each row. Sequence logos were generated with multiple sequence alignment (MSA) of the five groups individually in Weblogo 3.7.12. The labeled numbers of sequence logo represent yeast Pol II TL residue position (1076-1106). Bacteria n=465. Archaea n=426. Pol I n=605. Pol II n=405. Pol III n=444. (**F**). Lethal haplotypes contain two distinct types of lethality: one attributed to synthetic lethal interactions among substitutions, representing negative interactions, and the other arising directly from lethal substitutions. The calculated ratio specifically reflects the proportion of lethality due to negative interactions within all lethal haplotypes. (**G**). The ratio of viable haplotypes in haplotypes containing lethal substitutions, representing the ratio of positive interactions (suppression) in TL haplotypes of each group. Haplotypes containing lethal substitutions are expected to be lethal based on the additive model. If haplotypes with lethal substitutions are observed to be viable, it suggests other substitutions suppress the lethal substitution in the haplotypes, implying positive epistasis (suppression). Approximately 20% of eukaryotic TL haplotypes containing individual lethal substitutions are viable, whereas only roughly 2% of bacterial and archaeal TL haplotypes are viable, suggesting more positive interactions in eukaryotic TLs than those from Bacteria and Archaea.



Kolmogorov-Smirnov Test, P-value = 0.2288

Figure 26. The TL-internal epistasis is embedded in the Pol II enzymatic background.

(**A**). Residue coupling heatmaps of natural TL variants, Random scrambling variants and Monte Carlo scrambling variants. Residue coupling of 362 selected natural eukaryotic Pol I, II, and III TL alleles is shown in the left heatmap. (**B-C**). Heatmaps of coupled residues in Random scrambling and Monte Carlo scrambled haplotypes. The random scrambling disrupted coupled residues in selected TL variants (**B**) while the Monte Carlo scrambling haplotypes reserved coupled residues (**C**). (**D**). Cumulative fitness frequency distribution of the Random scrambling and Monte

Carlo scrambling haplotypes. The Kolmogorov-Smirnov test was used to assess the significance of the fitness distribution difference between Monte Carlo scrambling and Random scrambling haplotypes. The results indicate that they are not significantly different. Cumulative fitness frequency distribution of the Random scrambling and Monte Carlo scrambling haplotypes. The Wilcoxon matched-pairs signed rank test was used to assess the significance of the fitness distribution difference between Monte Carlo scrambling and Random scrambling haplotypes. The P-value of 0.3013 is not significant.

3.2.3 Epistasis within TL haplotypes can be attributed to residues

We next determined from which residues the epistatic effects in TL haplotypes emerge. We selected seven Pol II TL variants based on their diversity where each have 7~9 substitutions relative to the *S. cerevisiae* Pol II TL, and we constructed 1987 unique intermediate haplotypes representing 2169 combinations of substitutions from the seven variants. These seven variants cover a range of compatibility with yeast Pol II function as determined by their fitness when replacing the yeast Pol II TL (**Figure 27A-H)**. Four out of seven TL variants were not compatible in the yeast Pol II context for distinct reasons. The TL variant from *G. luxurians* was lethal presumably because it encodes G1088S, which causes lethality on its own and was similarly deleterious for all intermediate genotypes (**Figure 27A, C**). In contrast, lethalities of TL variants from *B. anathema* and *R. solani* were not explained by presence of an individual lethal substitution, and they were predicted to be viable by the summation of the fitness scores of individual substitutions. These observations suggested that the observed lethality for these two haplotypes is due to negative interactions among residues (lethality beyond additivity) (**Figure 27A, G, H**). The *S. rosetta* TL was both expected and observed to be lethal, supporting lethality from additivity of residues (**Figure 27A).** However, it's worth noting that *S. rosetta* intermediate combinations with specific substitutions exhibited fitness in the lethal range **(Figure 27F**) but were worse than predicted from the log additive model, suggesting negative interactions. In three out of seven variants, where all individual substitutions were viable and combinations between them were predicted to be viable based on the log additive model, some intermediate substitution combinations were in the severely sick or lethal range (**Figure 27B, D, E and 28**), implying negative interactions among the substitutions. Interestingly, this lethality of intermediate genotypes was suppressed by additional substitutions, resulting in final viable haplotypes and indicating positive interactions (**Figure 27B, D, E**). The fluctuations of the fitness with various residue combinations suggests complexity and potential higher order epistasis within these TL haplotypes.

To determine which specific residues contribute to the observed epistasis, we assessed deviation from expected fitness for the addition of each substitution to all intermediate haplotypes derived from subsets of substitutions of the haplotype. This analysis allows mutant effects to be compared for multiple related backgrounds (the intermediate states representing subsets of haplotype substitutions) (**Figure 23D, F**). The fitness and epistatic effects of each substitution are shown as epistasis heatmaps (**Figure 29 and 30**). For example, we observed better fitness for the *E. invadens IP1* TL haplotype than predicted from the log additive model (**Figure 30A**), suggesting positive interaction(s) within the eight substitutions comprising the haplotype. Among these eight are two with biochemically classifiable phenotypic profiles as determined from library screening of single substitution mutants on conditional media – the LOF variant V1089T and the GOF variant S1096E. Consistent with these alleles being candidates for epistatic effects, these two substitutions also have the lowest relative growth fitness of the eight *E. invadens IP1* TL substitutions (**Figure 30A-B**). S1096E had positive secondary deviation scores in most intermediate combinations, suggesting broad effects. V1089S showed positive effects but mostly in intermediates that also contained S1096E (**Figure 30B**), suggesting that V1089S's positive effect was dependent on S1096E, likely due to the suppression predicted from the combination of GOF and LOF substitutions. Interestingly, S1091E, a substitution without strong growth fitness and with no obvious phenotypes as a single substitution, showed strong positive or negative epistatic effects in most backgrounds (**Figure 30B**). These results suggest these three substitutions are responsible for intra-TL epistasis within *E. invadens IP1* TL haplotype.

To visualize these effects, we quantified the impact of constituent substitutions on epistasis within the haplotype by identifying correlations between the primary deviation score, which represents the overall epistasis of the haplotype, and the secondary deviation scores, which represents the specific epistatic effects of individual substitutions on corresponding haplotype backgrounds. Substitutions with stronger correlation (higher linear regression R^2) indicate these substitutions are the main drivers of the epistatic effects observed for the entire haplotype. Conversely, substitutions with low correlation (smaller R^2) suggest that the respective substitutions have no specific, or limited contributions to epistasis within the haplotype. To demonstrate this concept, we created a simplified simulation (**Figure 29**). Consider a TL haplotype comprising five substitutions: a, b, c, d, and e. In this scenario, the single mutant "a" is GOF, "b" and "c" are LOF, whereas "d" and "e" have no effects. We set the suppression interactions in GOF+ LOF combinations (a/b and a/c), and synthetic sick interactions in the LOF+ LOF combination b/c (**Figure 29A**). We calculated all primary and secondary deviation scores for the haplotype and all intermediate combinations. The GOF mutant "a" exhibited positive secondary deviation scores in most intermediates (**Figure 29B**). The secondary deviation scores of three single mutants with phenotypes showed strong correlation (R^2) with the primary deviation scores (**Figure 29C**), showing they are the key contributors to the observed epistatic interactions within haplotypes. The correlations observed between secondary deviation scores of “a” versus “b” and “c” indicate the secondary deviation scores of “b” and “c” are positive in most cases when “a” is present (**Figure 29D**), suggesting “b” and “c” potentially suppresses "a”. Similarly, the secondary deviation scores of “a” are consistently positive with changes in secondary deviation scores of “b” (**Figure 29E**) and “c” (**Figure 29F**), indicating “a” suppresses “b” and “c”. These observations confirmed the mutual suppression in GOF and LOF combinations. Conversely, LOF substitutions “b” and “c” showed correlated secondary deviation scores, showing their synthetic sick interactions.

To detect the key contributors to epistatic interactions observed in the *E. invadens IP1* TL haplotype, we determined the correlations between the primary and secondary deviation scores. Consistent with the epistasis landscape (**Figure 30B**), the secondary deviation scores of S1096E, V1089T, and S1091E have strong correlations ($R^2 > 0.5$) with the primary deviation score (**Figure 32A**), suggesting their substantial effect on the primary epistasis of the haplotype. ~1-3 substitutions with notable effects on haplotype epistasis were identified for all seven selected TL variants (**Figure 30B, 31, 32A, and 33**), suggesting epistasis within haplotypes is attributable to subsets of substitutions.

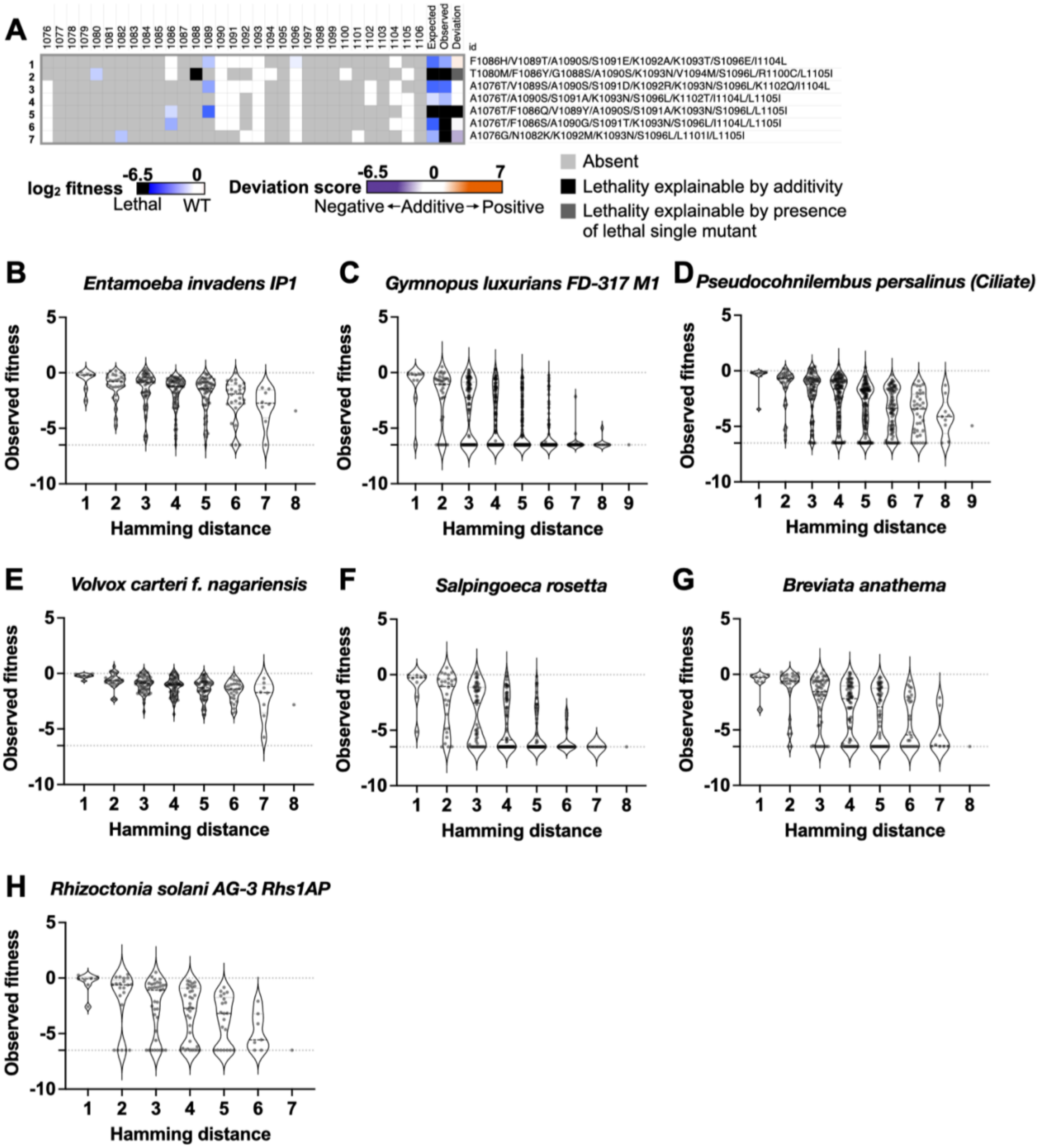


Figure 27. Interactions within TL haplotypes fluctuate with changes in substitution compositions. (**A**). Fitness of constituent single substitutions, and expected and observed fitness of the haplotypes, and deviation score are shown in heatmap for the selected seven haplotypes. The deviation shown in the heatmap is the primary deviation scores calculated by comparing observed fitness to expected fitness. (**B-H**). Distribution of growth fitness across all intermediate substitution combinations categorized by the count of substitutions (hamming distance).

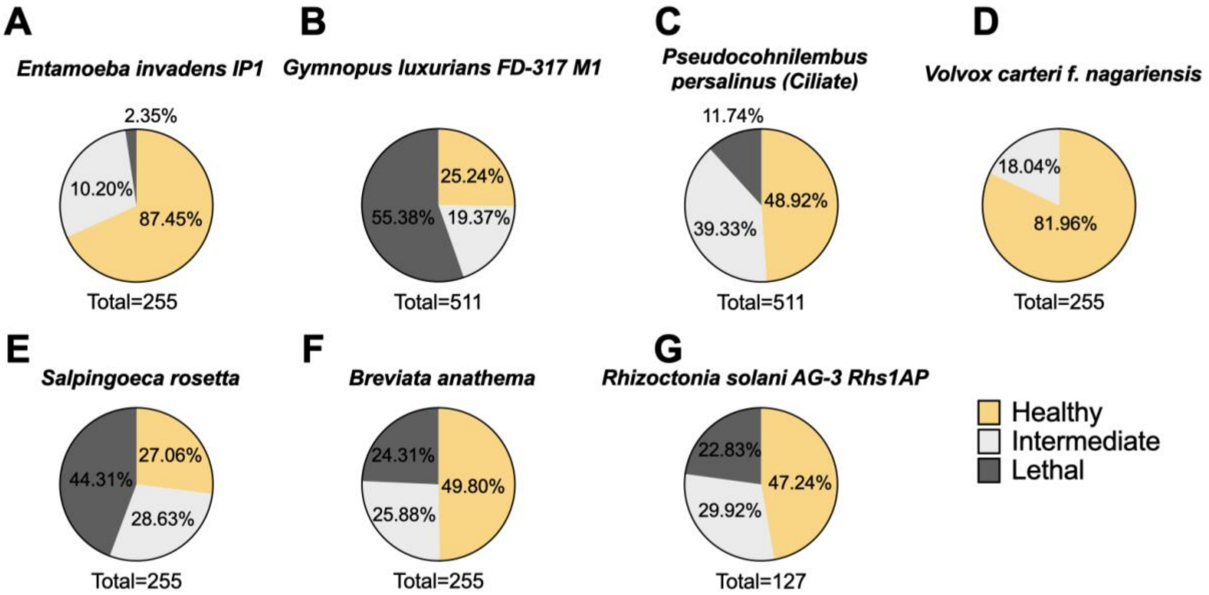


Figure 28. The percentage of fitness categories of all intermediate combinations within each haplotype.

(A-G). Healthy: Fitness > -2. Intermediate: -6.5 < Fitness ≤ -2. Lethal: Fitness ≤ -6.5**.**

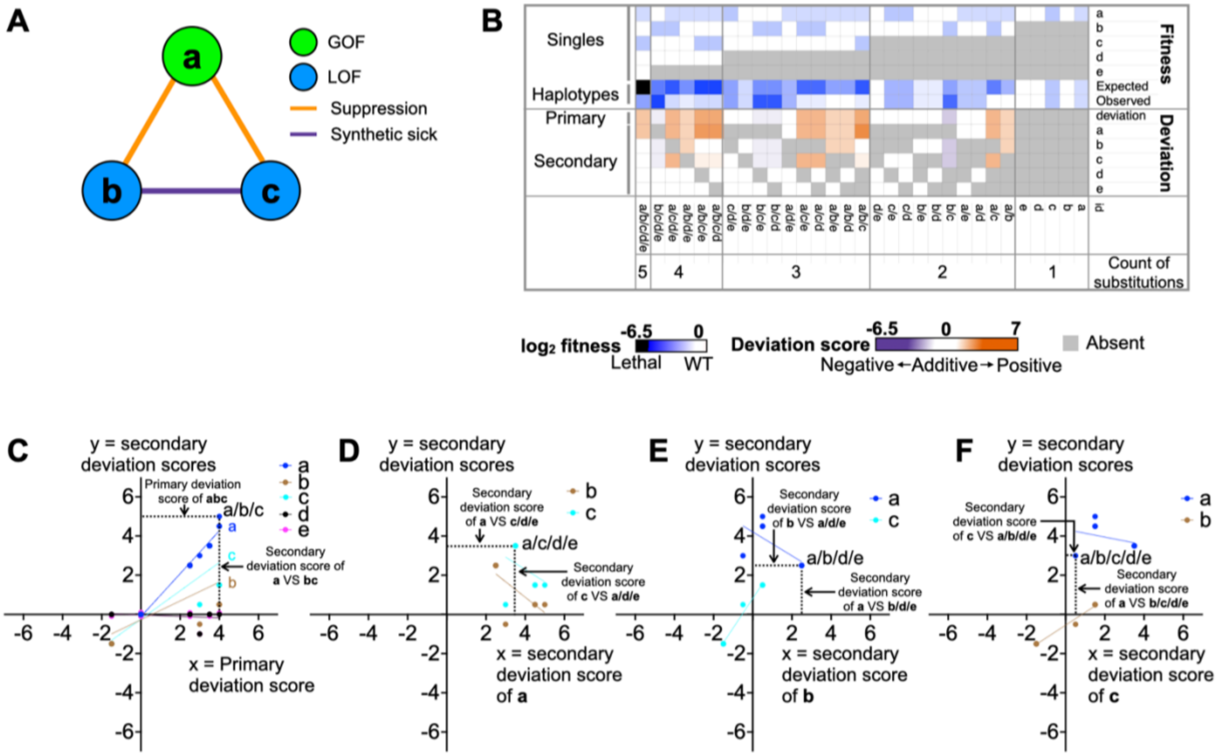


Figure 29. Example primary and secondary deviation scores for a simplified, simulated TL haplotype "abcde".

(**A**). Simulated interactions among substitutions "a", "b", and "c". We set the single mutant "a" as a GOF mutant, "b" and "c" as LOF mutants, while "e" and "f" have no effects. We also set suppression between GOF and LOF combinations (“ab” and “ac”), and synthetic sickness between LOF and LOF combinations (“bc”). (**B**). The epistasis interaction landscape of simplified TL haplotype “abcde”, illustrating the observed and expected fitness of all single substitutions and all intermediate haplotypes, as well as primary and secondary deviation scores in the simulation. (**C**). Correlations between secondary deviation scores of all five single substitutions (Y-axis) to the corresponding primary deviation scores (X-axis). To measure the correlation, linear regression was applied for secondary deviation scores of “a”, “b”, “c”, “d”, “e” to the corresponding primary deviations scores, respectively. Single substitutions with strong correlations ($R^2 > 0.5$) were labeled in the plot. Three substitutions with phenotypes show strong correlations in the simulation, “a”, “b”, and “c”. (**D-F**). The correlation between secondary deviation scores of the four substitutions (Y-axis) vs “a” (**D**), “b” (**E**), “c” (**F**) on X-axis respectively. To read the plot, with (**D**) as an example, each cyan spot represents an intermediate combination containing both “a” and “c”. The labeled spot represents a specific combination “a/c/d/e”. Its coordinate value in X-axis represents the secondary deviation score of “a” to “c/d/e”. Its coordinate value in Y-axis represents the secondary deviation score of “c” to “a/d/e”. We did a linear regression for all the cyan spots to represent the correlation between the secondary deviation scores of “c” to “a”. This correlation helps illustrate the potential interaction between “c” and “a”. The secondary deviation scores of “c” are all positive when “a” is present, meaning “c” constantly shows positive secondary score when “a” is present, indicating potential suppression of “c” to “a”.

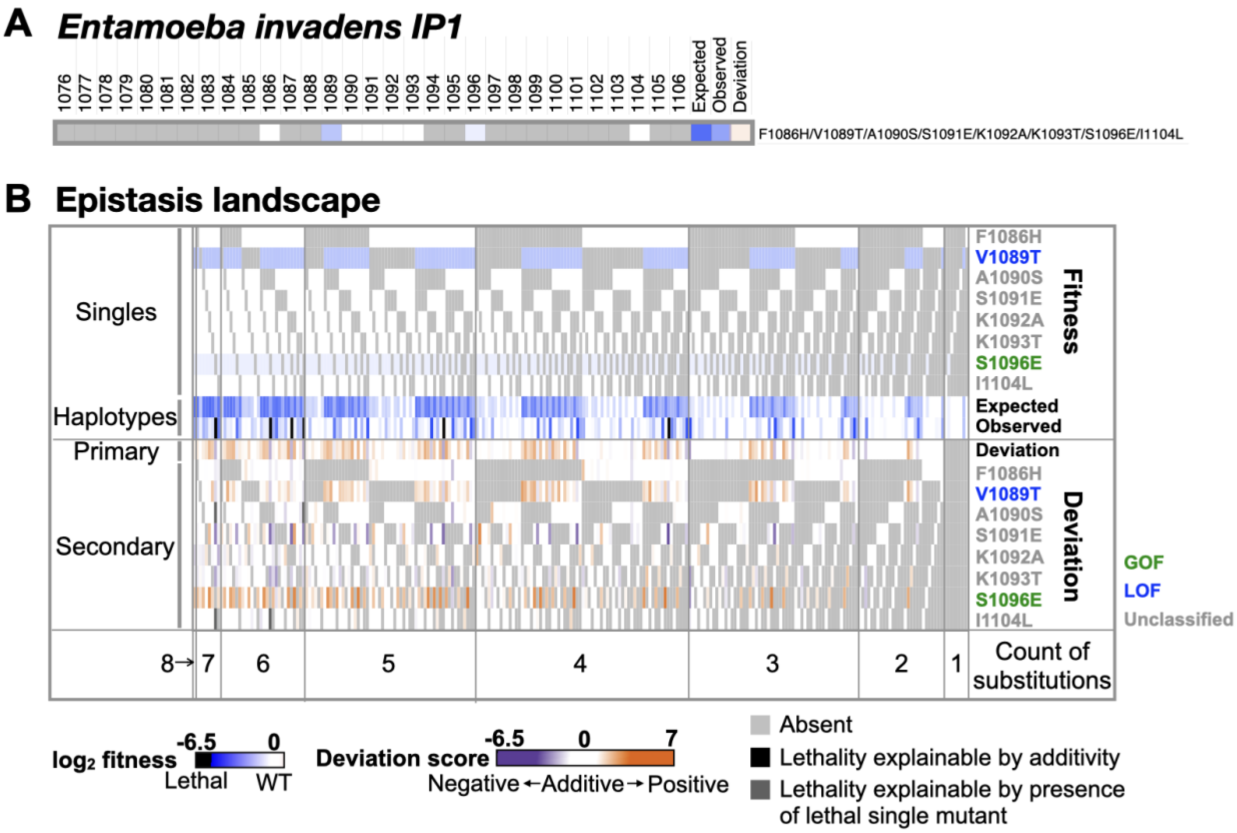


Figure 30. The epistasis landscape provides a comprehensive view of primary and secondary deviation scores, emphasizing substitutions with notable epistatic effects.

(**A**). Fitness of eight *Entamoeba invadens* IP1 TL single substitutions in yeast Pol II background, and the expected and observed fitness and the primary deviation score are shown in the heatmap. (**B**). The epistasis landscape of *E. invadens IP1* TL substitutions. The heatmap illustrates the fitness and epistasis of all unique intermediate haplotypes coming from combinations of eight substitutions. Intermediate haplotypes are grouped by counted number of substitutions from 1 to 8. The fitness values are displayed in the upper panel and the epistasis, represented by primary and secondary deviation scores is displayed in the lower panel. The colors of substitution names indicate their phenotypes, GOF is in green, LOF is in blue, unclassified is in grey.

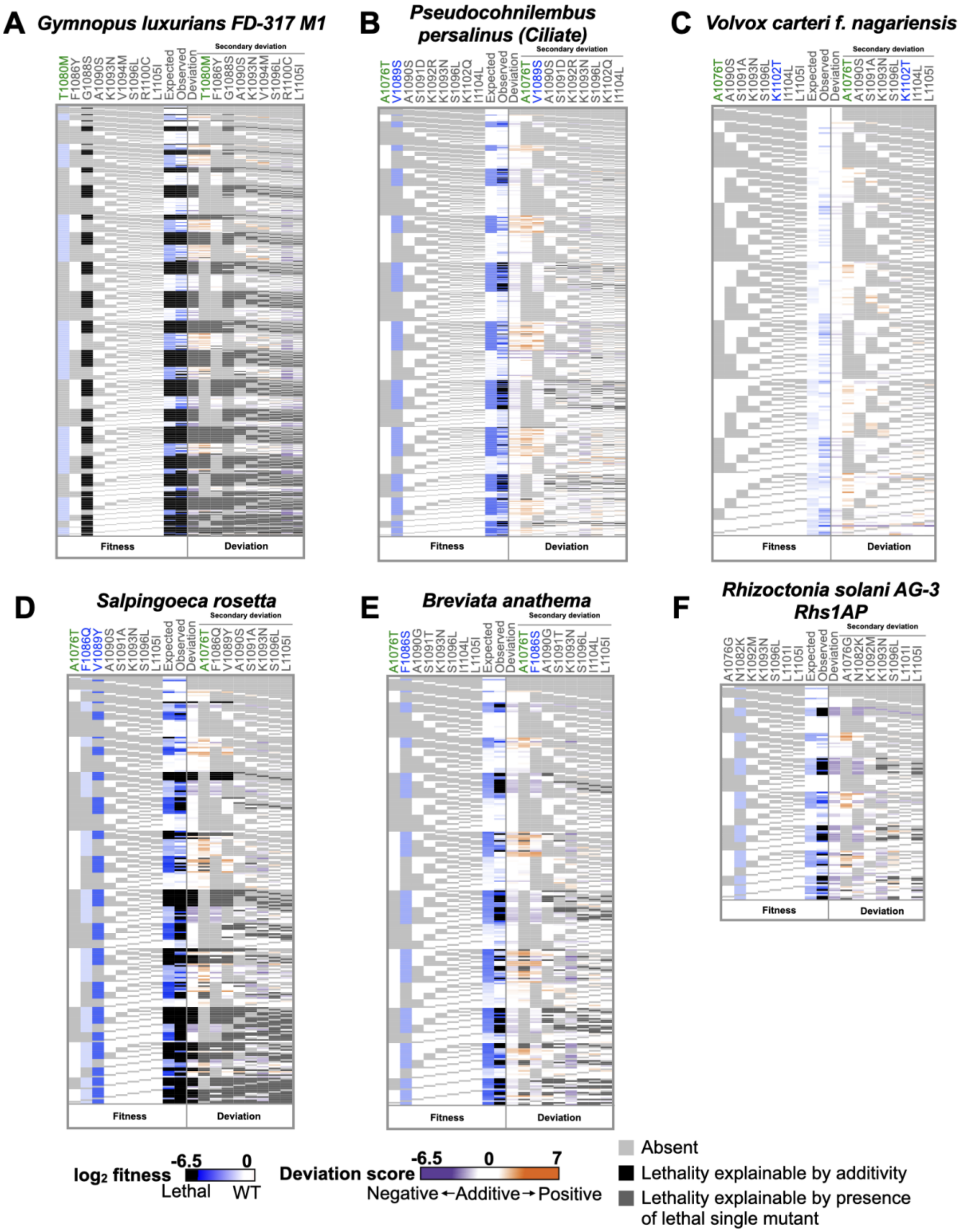


Figure 31. The epistasis landscapes of selected haplotypes.

The fitness and epistasis of all unique intermediate haplotypes from combinations of substitutions within each haplotype (**A-F**). The colors of mutants' names represent mutants' phenotypes. GOF is in green, LOF is in blue, non-classified mutants is in grey.

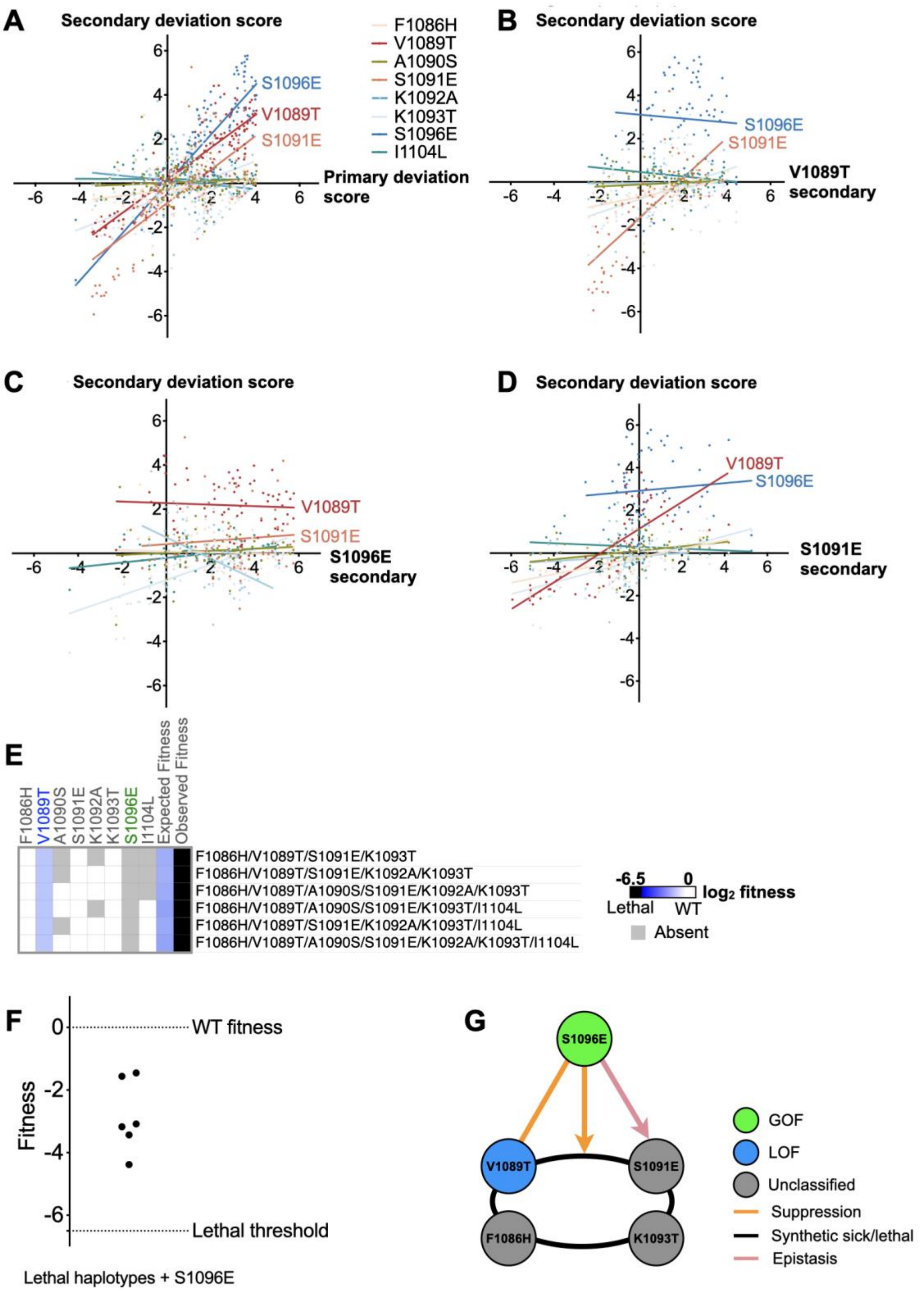


Figure 32. Correlations between deviation scores reflect specific residue interactions in *E. invadens IP1* TL substitutions.

(**A**). Correlations between secondary deviation scores of all eight substitutions (Y-axis) and the primary deviation score (X-axis). Linear regression was applied to each comparison of secondary deviation scores against primary deviation scores to check the correlation. Substitutions with an R^2 value exceeding 0.5 are annotated on the X-Y plot, indicating their substantial impact on primary epistasis of the haplotypes. (**B-D**). Correlations between secondary deviations of the other seven substitutions (Y-axis) vs V1089T (**B**), S1096E (**C**), S1091E (**D**) on X-axis respectively. (**E**). The fitness landscape of intermediate combinations with fitness in ultra-sick/lethal range. Their fitness levels are indicated with black blocks in the heatmap while the expected fitness calculated from the log additive model is in viable range (light blue blocks). The observed fitness is worse than expected, representing negative interactions. Names of substitutions are colored based on their phenotypes. GOF: green. LOF: blue. No obvious phenotype: grey. (**F**). The fitness of all ultra-sick to lethal haplotypes with S1096E incorporated is no longer in the ultra-sick/lethal range. (**G**). Scheme of specific residue interactions within substitutions of *E. invadens IP1* TL.

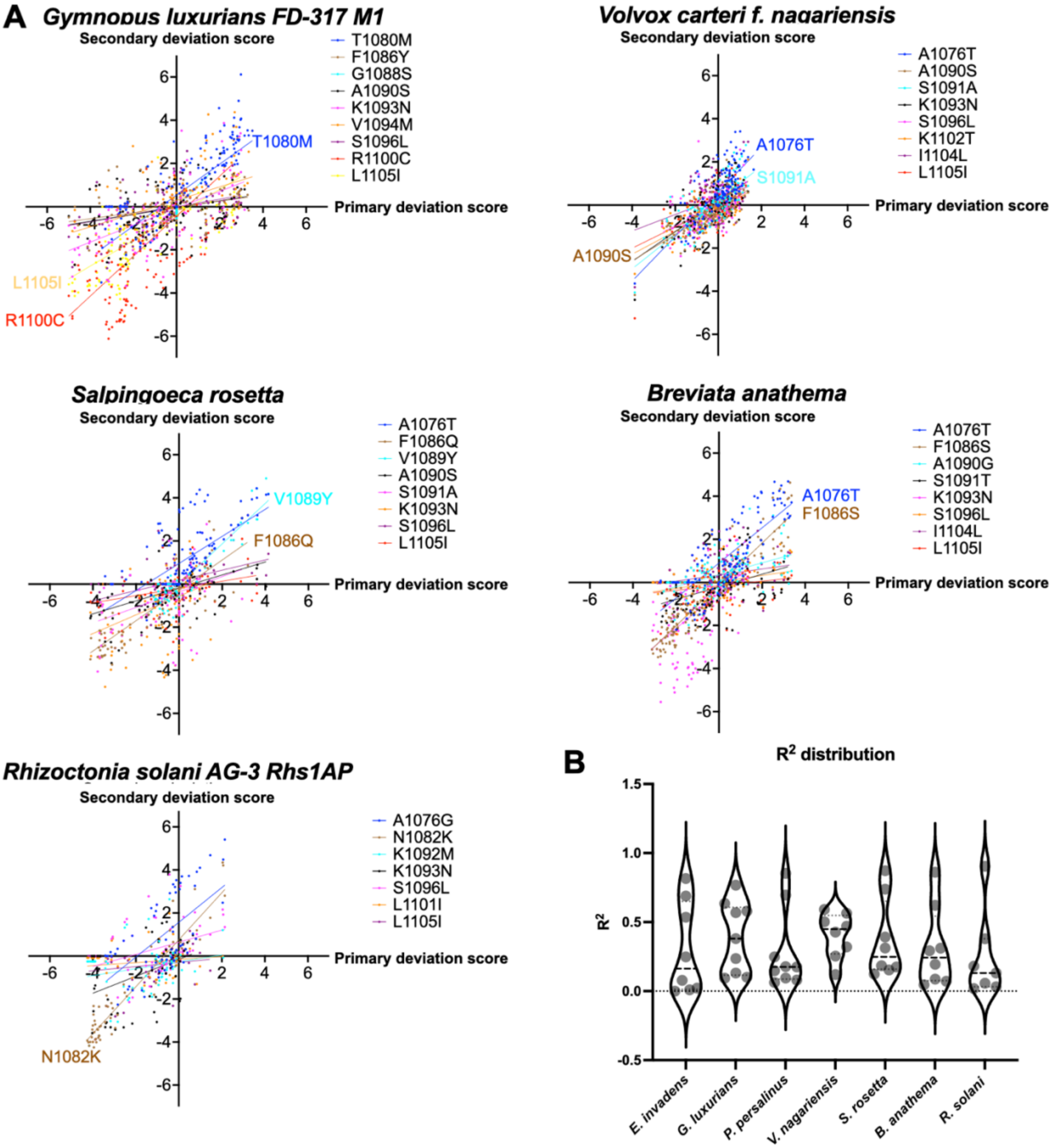


Figure 33. The epistasis within haplotypes is driven by several specific substitutions.

(**A**). The correlations between secondary deviation scores of all substitutions (Y-axis) and the primary deviation score (X-axis) for each selected haplotype. Linear regression was applied to each comparison of secondary to primary deviation scores. Substitutions with R^2 value exceeding 0.5 are annotated on the X-Y plot, indicating their substantial impact on primary epistasis of the haplotypes. (**B**). Distributions of R^2 of all selected TL haplotypes.

3.2.4 The epistasis within TL haplotypes can reflect intricate residue interactions

We dissected the interactions formed by the three substitutions S1096E, V1089T, and S1091E in the *E. invadens IP1* TL haplotype and observed an intricate, higher-order epistasis network (**Figure 32A-G**). First, we confirmed the mutual suppression between the GOF S1096E and the LOF V1089S. S1096E had a consistent positive epistatic effect in haplotypes containing V1089T **(Figure 32B**). V1089T exhibited a smaller but similar positive effect on haplotypes containing S1096E (**Figure 32C**), illustrating mutual suppression between V1089T and S1096E. These observations are consistent with the predicted suppression between GOF and LOF mutants and is illustrated by the positive values for their combinations shown in **Figure 30B**. Additionally, we observed negative interactions for S1091E when it was combined with F1086H/V1089T/K1093T, but these were dependent on the absence of S1096E. This was indicated by a low secondary deviation for S1091E when S1096E was present (**Figure 32C-D**)**.** These observations are interpreted as S1096E being epistatic to S1091E. We identified negative interactions in six intermediate combinations containing S1091E, as suggested by their fitness in the ultra-sick or lethal range beyond the log additive model (**Figure 32E**). Strikingly, F1086H, V1089T, K1093T were present in all six intermediate combinations with S1091E, pointing to negative synergistic interaction (synthetic sickness or lethality) between the four substitutions. Conversely, S1096E was absent in all six combinations (**Figure 32E**), implying the negative interaction may be suppressed by S1096E. We confirmed the suppression by the observation that the fitness of all six ultra-sick/lethal combinations with S1096E incorporated was out of the ultra-sick/lethal range (fitness > -5) (**Figure 32F**) and by spot assay (**Figure 34**). We summarize all the observed interactions in **Figure 32G**, representing the complicated higher-order epistasis network.

Substitutions of the *P. persalinus (Ciliate)* TL haplotype formed two higher-order epistasis networks. The complexity of these epistasis networks arises from the intricate layers of substitution interactions. Notably, these interactions were not immediately apparent when observing all substitutions in the entire haplotype, meaning that the predicted and expected fitnesses for the complete haplotype were similar (**Figure 35-37**). First, the haplotype contains nine single substitutions with no or slight growth fitness defects. The observed fitness of the haplotype was very similar to expected from the log additive of all individual substitutions' fitnesses, with the calculated primary deviation score close to 0. Second, when we dissected potential interactions in substitution combinations, we noticed two substitutions exhibited distinct behavious, A1076T and V1089S. Since both of these residues are classifiable as having predicted biochemical phenotypes (A1076T as a GOF and V1089S as a LOF), we anticipated suppression between them and predicted they would be drivers of epistatic interactions within intermediate combinations. Consistently, suppression was observed when A1076T and V1089S were combined (**Figure 36**), and the correlations between the secondary and primary deviation scores suggest that A1076T and V1089S strongly affect the epistasis in the haplotypes (**Figure 35B**). Additionally, two intermediate combinations, each with five substitutions, appeared to have the lowest primary and secondary deviation scores (the labeled **I** and **II in Figure 35B**), implying negative interactions (synthetic sickness/lethality) within the two combinations respectively. The combination I, A1076T/A1090S/S1091D/S1096L/I1104L, showed a negative primary deviation score (observed fitness < expected fitness) (**Figure 35C**), illustrating synthetic sickness among the five substitutions. The secondary deviation scores of five constituent substitutions were all negative (**Figure 35C**), implying the synthetic sickness could be attributed to addition of each substitution to a combination of the others. The remaining four substitutions, V1089S, K1092R, K1093N, and

K1102Q all showed positive secondary deviation scores when they are incorporated into the combination, suggesting each one of them could individually suppress the synthetic sick combination (**Figure 35C**). The observed interaction network is shown in **Figure 35D**. In contrast, sign epistasis was observed in the combination II, A1076T/A1090S/S1091D/K1092R/K1093N, where the GOF sign of A1076T appeared to change to LOF when A1090S/S1091D/K1092R/K1093N were present simultaneously with A1076T. In detail, the combination has a negative primary deviation score, representing negative interaction among the substitutions (**Figure 35E**). Surprisingly, the negative interaction only appeared when the four substitutions, A1090S, S1091D, K1092R, and K1093N were present simultaneously with A1076T (**Figure 37A**), and there were no obvious interactions among the four substitutions (**Figure 36**), suggesting the interaction between A1076T and the combination of the four substitutions A1090S/S1091D/K1092R/K1093N. We infer a change from GOF to LOF for A1076T due to the observation of lethality upon incorporation of LOF V1089S into the combination. While A1076T *suppresses* V1089S in backgrounds in nearly all synthetic sick/lethal combinations between V1089S and individual substitutions or intermediate combinations of A1090S, S1091D, K1092R, and K1093N, it *enhances* the growth defect of V1089S/A1090S/S1091D/K1092R/K1093N. This outcome aligns with the expectation of lethality in a combination with the same class substitutions (LOF V1089S and the LOF A1076T). (**Figure 37B-C**). These observations are consistent with the combination of A1090S/S1091D/K1092R/K1093N converting A1076T from a GOF to a LOF. Notably, S1096L, K1102Q, and I1104L could individually suppress A1076T/A1090S/S1091D/K1092R/K1093N when they were incorporated into the combination, as indicated by their positive secondary deviation scores (**Figure 35E**), implying each one of them could revert the sign epistasis within A1076T/A1090S/S1091D/K1092R/K1093N. V1089S is now suppressible by A1076T when either S1096L, K1102Q, or I1104L are present (**Figure 35F**). In summary, these higher order interaction networks observed in the *P. persalinus (Ciliate)* TL haplotype (**Figure 35E-F**) emphasize the complexity of higher-order epistasis, revealing layers of interactions beneath the surface.

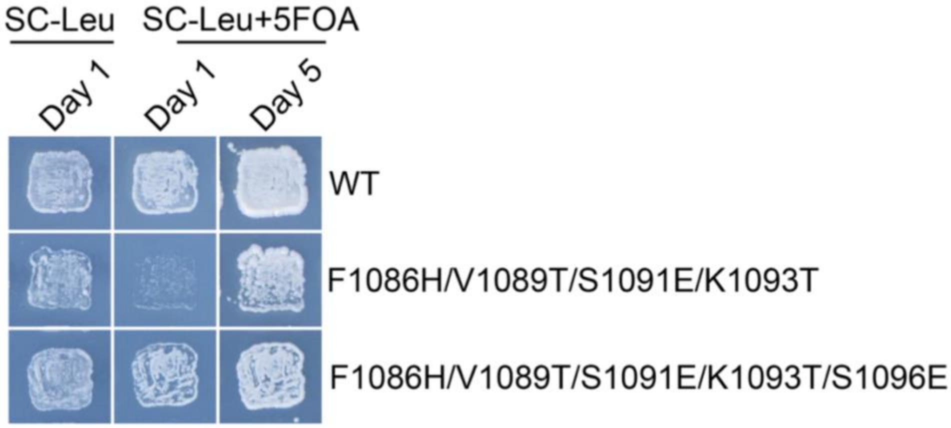

SC-Leu
SC-Leu+5FOA
Day 1
Day 1
Day 5
WT
F1086H/V1089T/S1091E/K1093T
F1086H/V1089T/S1091E/K1093T/S1096E

Figure 34. Verification of S1096E suppression on a synthetic sick intermediate haplotype F1086H/V1089T/S1091E/K1093T using patch assay.

Sequences of both haplotypes were synthesized and verified by sequencing. Growth on SC-Leu+5FOA indicates the growth defects of the haplotypes. SC-Leu is the control condition where WT *RPB1* is present.

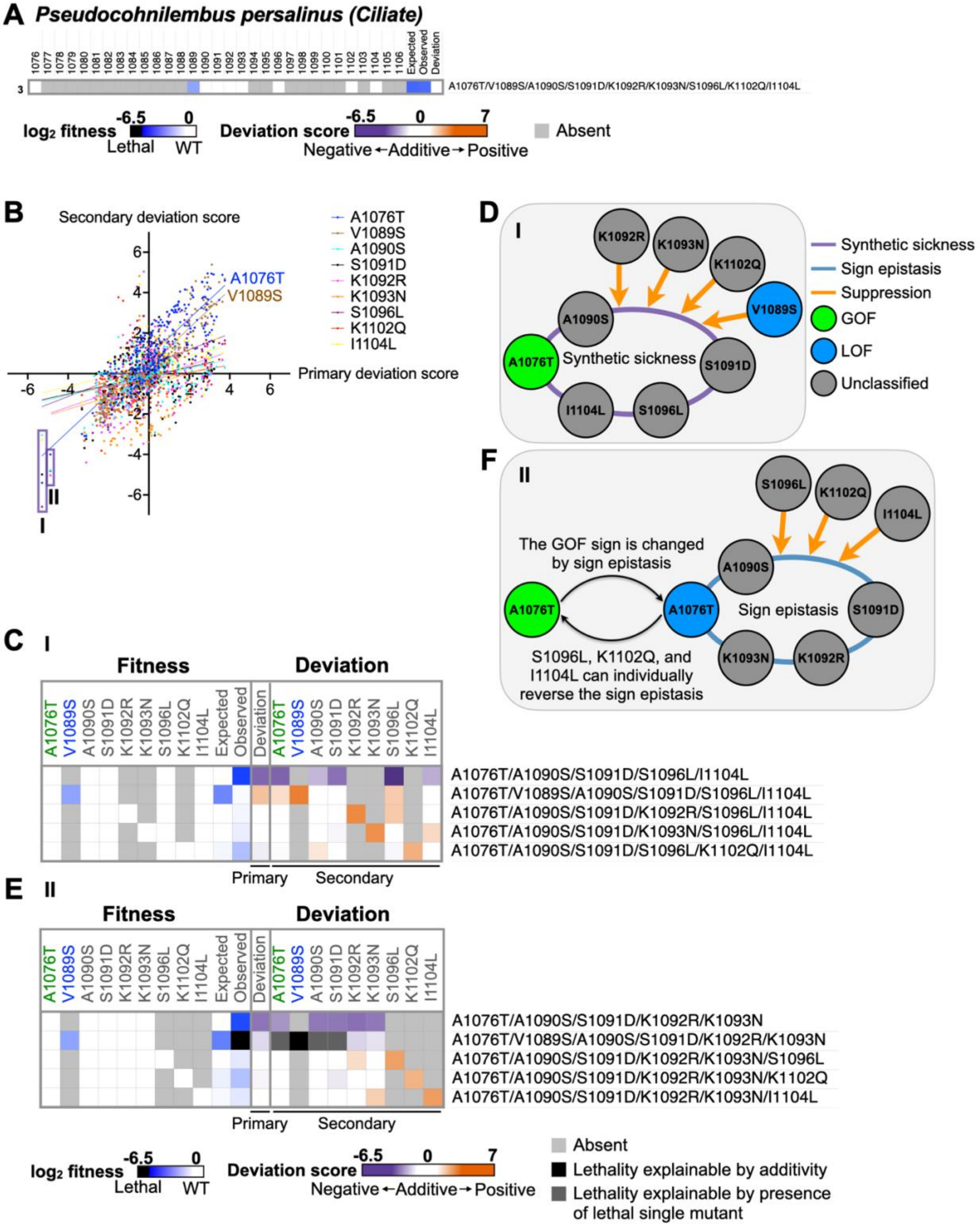


Figure 35. Intricate higher-order epistasis observed in substitutions of *P. persalinus (Ciliate)* TL haplotype.

(**A**). The heatmap displays the fitness of nine single substitutions in P. persalinus (Ciliate) TL in the yeast Pol II background, along with the epistasis between them represented by the primary deviation score. (**B**). Similar to Figure 30A, we checked correlations between secondary deviation scores (Y-axis) to the primary deviation score (X-axis) to identify substitutions with substantial impact on primary deviation scores. Simple linear regression was applied to each comparison. Substitutions with $R^2 > 0.5$ are annotated in the plot. (**C**). The fitness and deviation scores of substitution combinations related to group I are shown in the heatmap. Names of substitutions are colored based on their phenotypes. GOF: green. LOF: blue. No obvious phenotype (unclassified): grey. Each line shows the fitness and deviation scores of substitutions in a certain combination. Left, the fitness of individual substitutions, and the expected and observed fitness. Right, the primary deviations calculated by comparing observed and expected fitness and the secondary deviation scores of each constituent substitution. A1076T/A1090S/S1091D/S1096L/I1104L is in the first line. Its observed fitness is smaller than expected and when compared, resulting in a negative primary deviation score, representing a negative interaction. The secondary deviation scores of each constituent substitutions are all negative, indicating each of them showing negative interactions when adding to corresponding compounds. The following four lines represent the four combinations where V1089S, K1092R, K1093N, and K1102Q are incorporated into A1076T/A1090S/S1091D/S1096L/I1104L respectively. All observed fitness of combinations is healthier than A1076T/A1090S/S1091D/S1096L/I1104L, and the secondary deviation scores of V1089S, K1092R, K1093N and K1102Q are all positive, implying positive effect (suppression) on each combination respectively. (**D**). Scheme illustrating the substitution interaction network observed in **C**. (**E**). Similar to **C**, the fitness and deviation scores of combinations related with group II are shown in the heatmap. The first row shows the fitness and deviation scores detected within the combination A1076T/A1090S/S1091D/K1092R/K1093N. The following rows displays the corresponding fitness and deviation scores when the other four substitutions are incorporated. Notably, the effect of V1089S on A1076T/A1090S/S1091D/K1092R/K1093N cannot be determined because the observed fitness of the combination (V1089S + A1076T/A1090S/S1091D/K1092R/K1093N) is in the ultra-sick/lethal range, and its expected fitness calculated from the log additive model is also in the lethal range. However, V1089S and the compounds are sick but viable. The expected lethality is due to additivity. In this case, the secondary deviation of V1089S cannot be calculated and is represented by a black block in the heatmap. Moreover, the effect of A1076T on V1089S/A1090S/S1091D/K1092R/K1093N cannot be determined because the observed fitness falls within the lethal range. The expected fitness of (A1076T + V1089S/A1090S/S1091D/K1092R/K1093N) is also in the lethal range due to the presence of the lethal compound V1089S/A1090S/S1091D/K1092R/K1093N. The expected lethality of the combination is because it contains a lethal component. The secondary deviation score of A1076T cannot be determined either and is indicated by a dark gray block in the heatmap. (**F**). Scheme representing the substitution interaction networks observed in **E**.

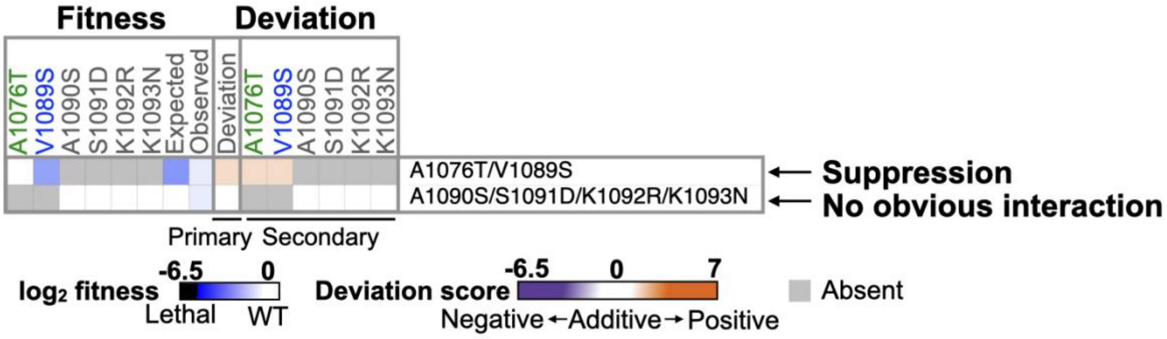


Figure 36. Heatmap displaying fitness and deviation scores in A1076T/V1089S and A1090S/S1091D/K1092R/K1093N.

Positive deviation scores in A1076T/V1089S indicate suppression, and no deviations (primary and secondary deviation scores are around zero) indicate no interactions within A1090S/S1091D/K1092R/K1093N.

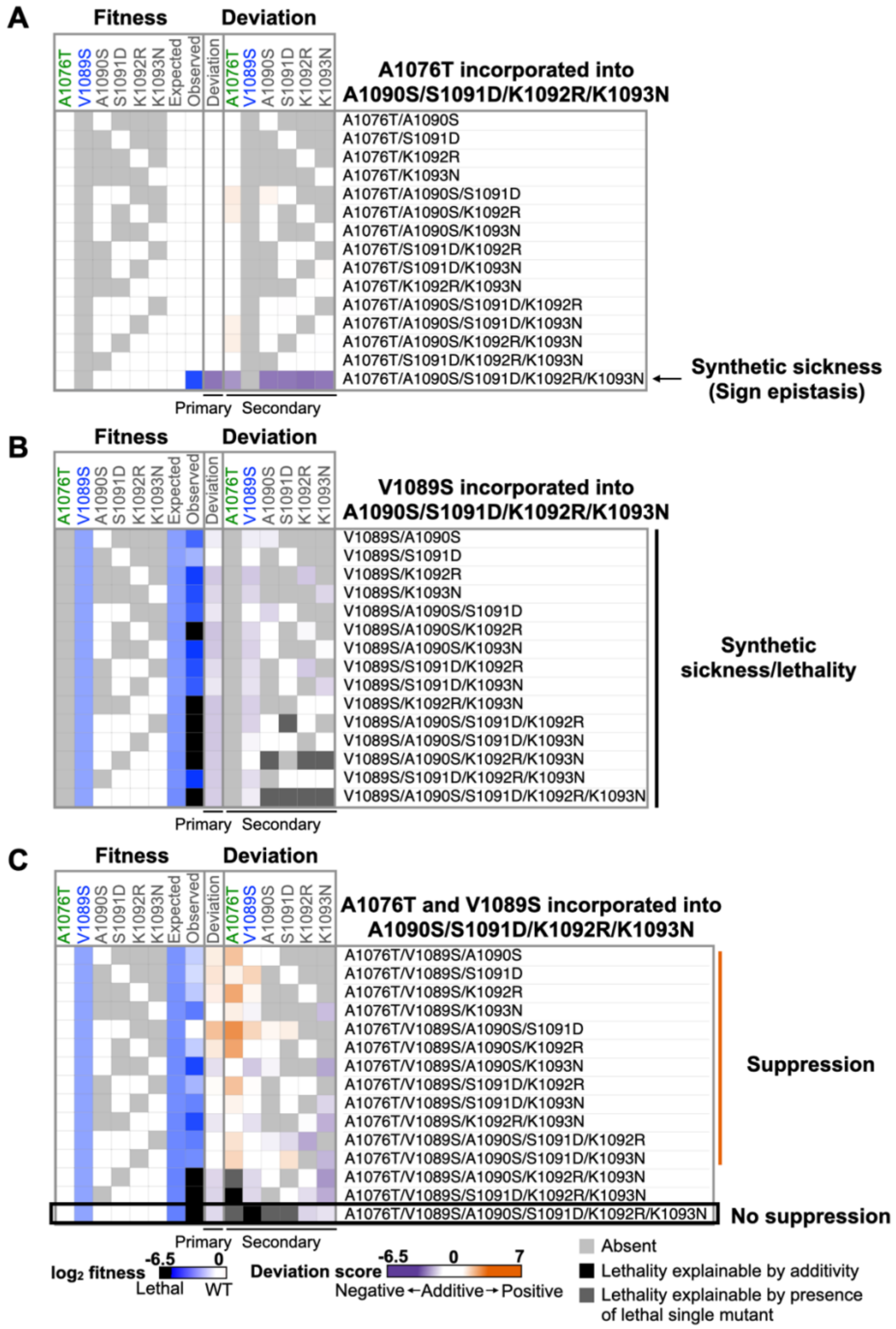


Figure 37. A1076T showed sign epistasis with A1090S/S1091D/K1092R/K1093N.

(**A**). Heatmap displays fitness and deviation scores for A1076T in all single substitutions and combinations with A1090S, S1091D, K1092R, K1093N, suggesting A1076T only showed negative interaction with A1090S/S1091D/K1092R/K1093N. (**B**). Heatmap shows fitness and deviation scores for V1089S in all single substitutions and combinations with A1090S, S1091D, K1092R, K1093N. V1089S showed negative interactions with almost all of them. (**C**). Heatmap shows fitness and deviation scores for A1076T and V1089S in all single substitutions and combinations with A1090S, S1091D, K1092R, K1093N. Positive interactions were observed in most combinations.

3.2.5 Distinct categories of residue epistasis patterns

To comprehensively compare the magnitude and consistency of epistatic effects of TL substitutions when introduced into different genetic backgrounds, we determined the distributions of secondary deviation scores for each substitution. Substitutions with strong epistatic effect (**Figure 29-30**) and primary/secondary score correlations (**Figure 32A, Figure 33, and Figure 35B**), display wide distributions of secondary deviation scores. Interestingly, while most of these impactful substitutions have classifiable phenotypes as single substitutions (GOF or LOF) (**Figure 38A-F**), some without obvious phenotypes still show wide ranges of epistatic impacts, such as S1091E (**Figure 38A**), K1093N (**Figure 38E-F**), A1076G and N1082K (**Figure 38G**). Notably, some substitutions, like A1076T, were present in more than one TL haplotype background (**Figure 38C-F**). To evaluate the overall epistatic effects of these substitutions in all tested backgrounds, we displayed the density plots of their secondary deviation scores across all non-repetitive haplotypes in which they are found (**Figure 40**) and calculated the maximum likelihood estimate (σ2) to quantify the distribution of secondary deviation effect for each substitution (Park et al., 2022) (**Figure 39**). 62.5% of substitutions exhibit mild epistatic effects (σ2 < 3) and 25% of substitutions have medium effect (3 ≤ σ2 ≤ 5). A small portion, 12.5% of substitutions, show strong epistatic effects (σ2 > 5). The epistatic effects of substitutions did not correlate with their fitness as single substitutions (Fig**ure 41A**). Moreover, the distribution of a substitution’s secondary deviation scores is expected to follow normal distribution if the substitution has no or minor epistatic effect in backgrounds to which it is introduced (Park et al., 2022). However, ~80% of TL substitutions did not have normally distributed secondary deviation scores (Fig**ure 41B**), illustrating that most TL substitutions had epistatic effects when introduced into some if not all genetic backgrounds. To investigate whether the epistatic effects of substitutions consistently remained positive (or negative) across various genetic backgrounds, we calculated the median of secondary deviation scores for substitutions respectively and plotted them against the corresponding σ2 values. Substitutions with strong epistatic effects (σ2) exhibited higher positive (or negative) median values (Fig**ure 39B**), implying that substitutions with robust epistatic effects consistently displayed either positive or negative impacts across various genetic backgrounds. We further compared the epistatic effects of substitutions in three categories (Fig**ure 39C**). Substitutions with a mild epistatic effect had little impact on the fitness of haplotypes when introduced. Substitutions with a medium epistatic effect, like A1076T and K1093N, could either enhance or reduce fitness of haplotypes. In contrast, substitutions with strong epistatic effects, such as S1096E and R1100C, were epistatic to backgrounds where they were observed, consistent with their requirement for toleration of specific substitutions in those backgrounds.

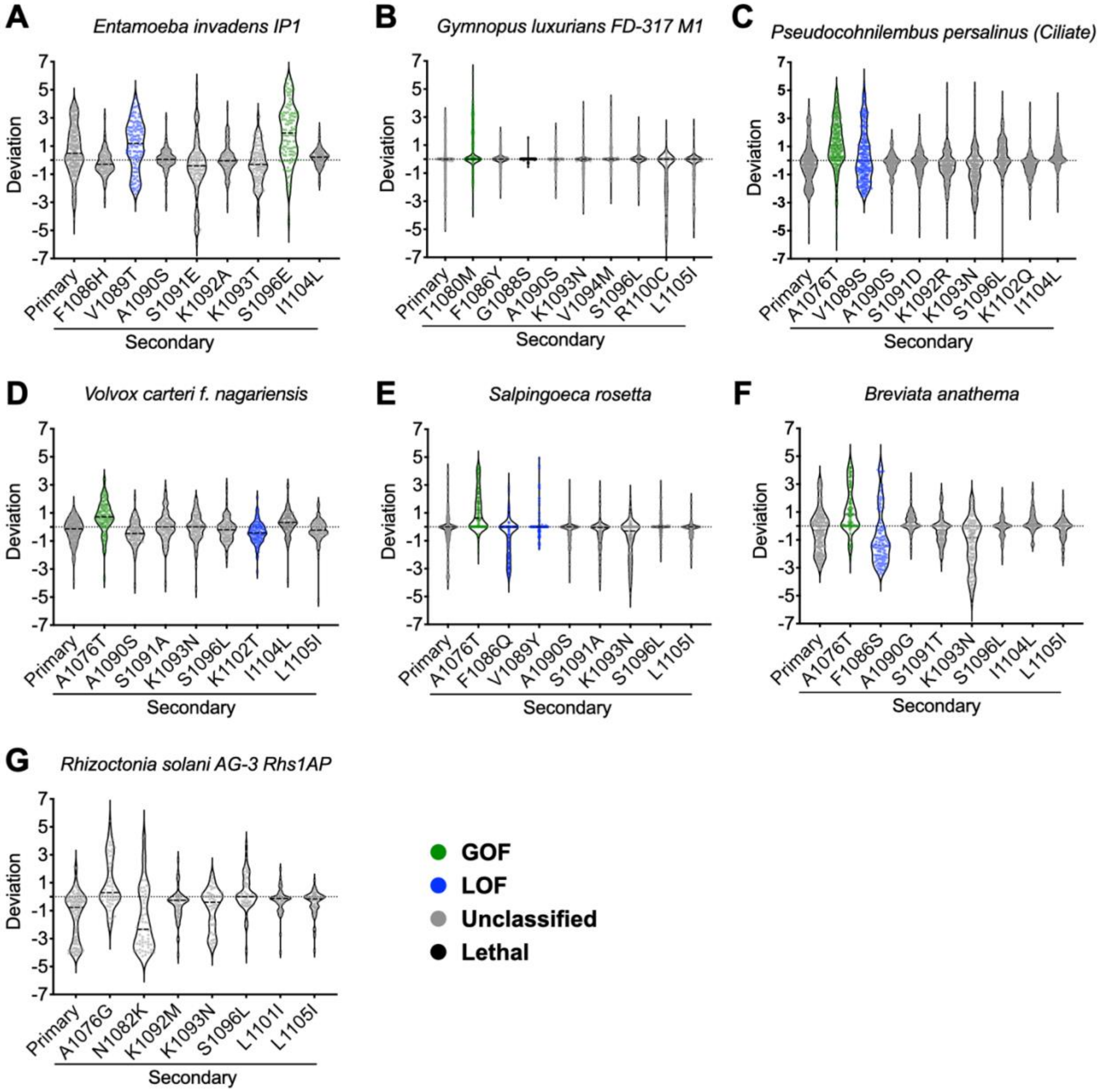


Figure 38. The distributions of primary and secondary deviation scores within each selected haplotype.

Colors of spots represent phenotypes of substitutions.

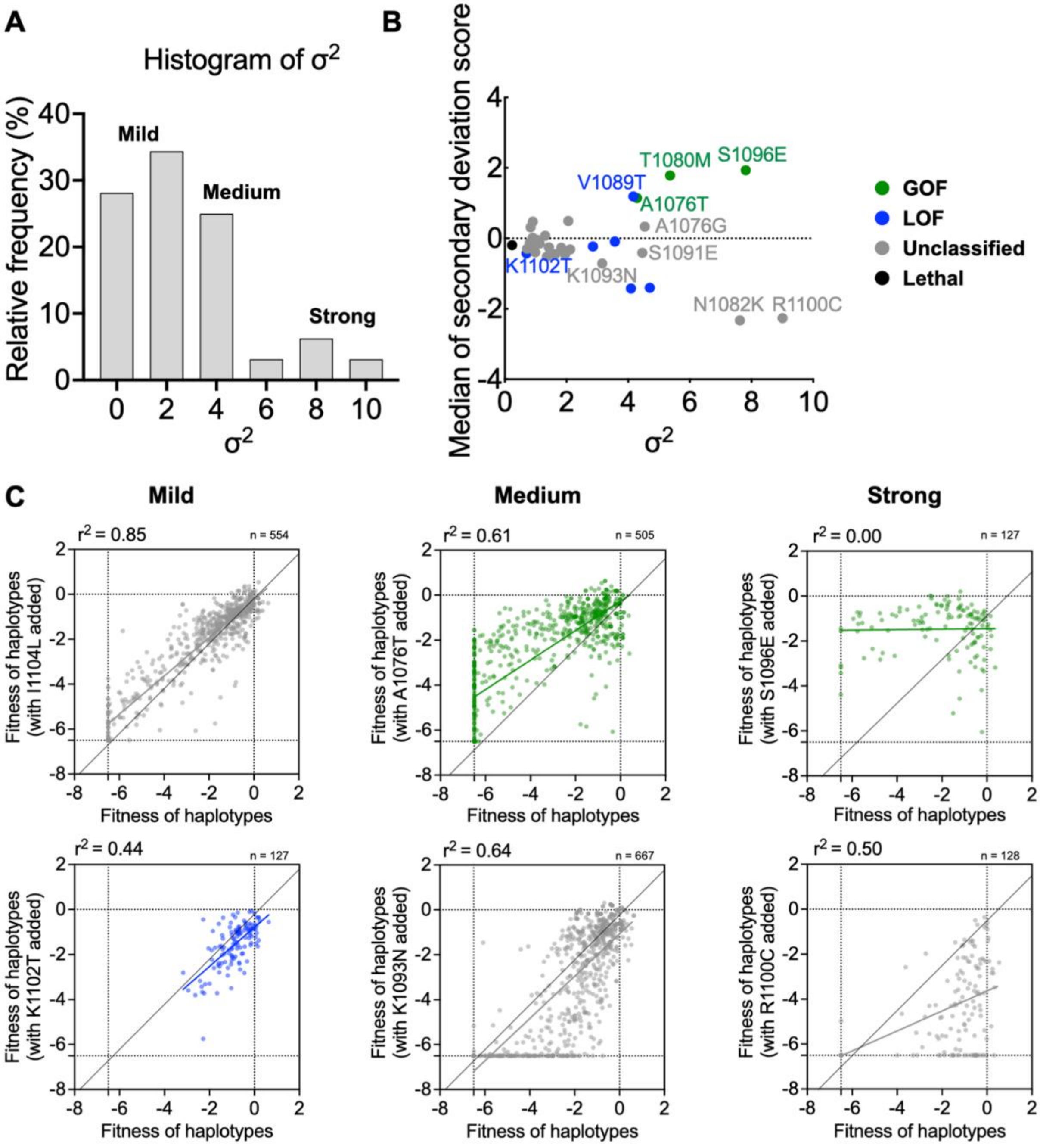


Figure 39. Different classes of epistatic effects.

(**A**). Histogram of mutants' epistatic effects, represented by their respective maximum likelihood estimate (σ2) of secondary deviation scores. Higher epistatic effect indicates greater impact of a certain substitution. (**B**). Medians of secondary deviation scores of substitutions were plotted against their corresponding σ2. Substitutions are colored based on their phenotypes. (**C**). Comparing epistatic effects of mutants in each category. Each scatter plot shows the measured fitness of haplotypes without (X-axis) versus with (Y-axis) a substitution incorporated. The colors of the plots represent the mutants' phenotypes. The colored line marks the simple linear regression of the spots, representing the observed epistatic effect of the substitution. R^2 values of the regressions are labeled in the plots. The black line indicates the additive (non-epistatic) expectation.

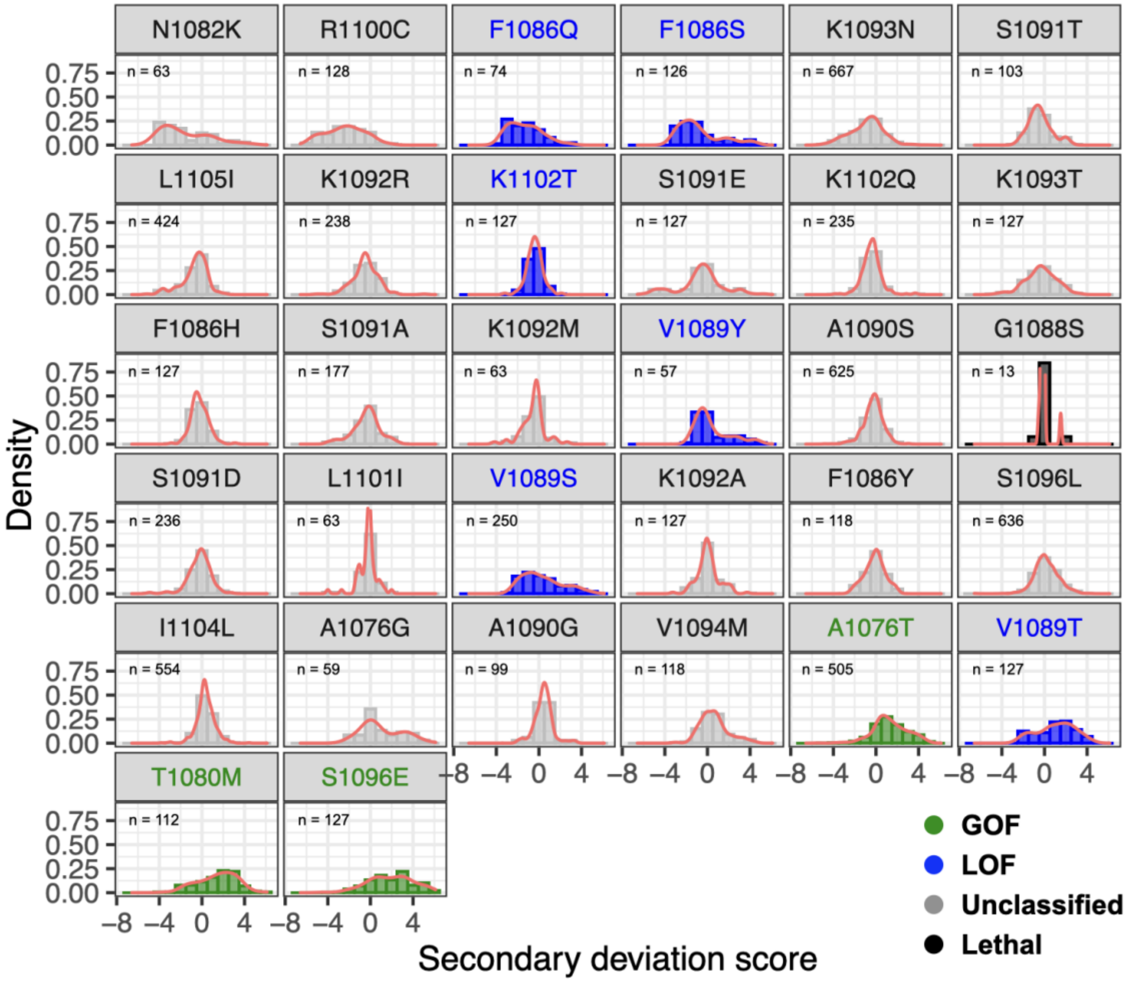


Figure 40. The density plots of secondary deviation scores of each substitution.

We made density plots to display the distribution of secondary deviation scores for each substitution. The density plots were calculated of substitutions from nonrepetitive intermediate combinations across all haplotypes. The count of nonrepetitive intermediate combinations containing the substitution is shown in the upper left of each plot.

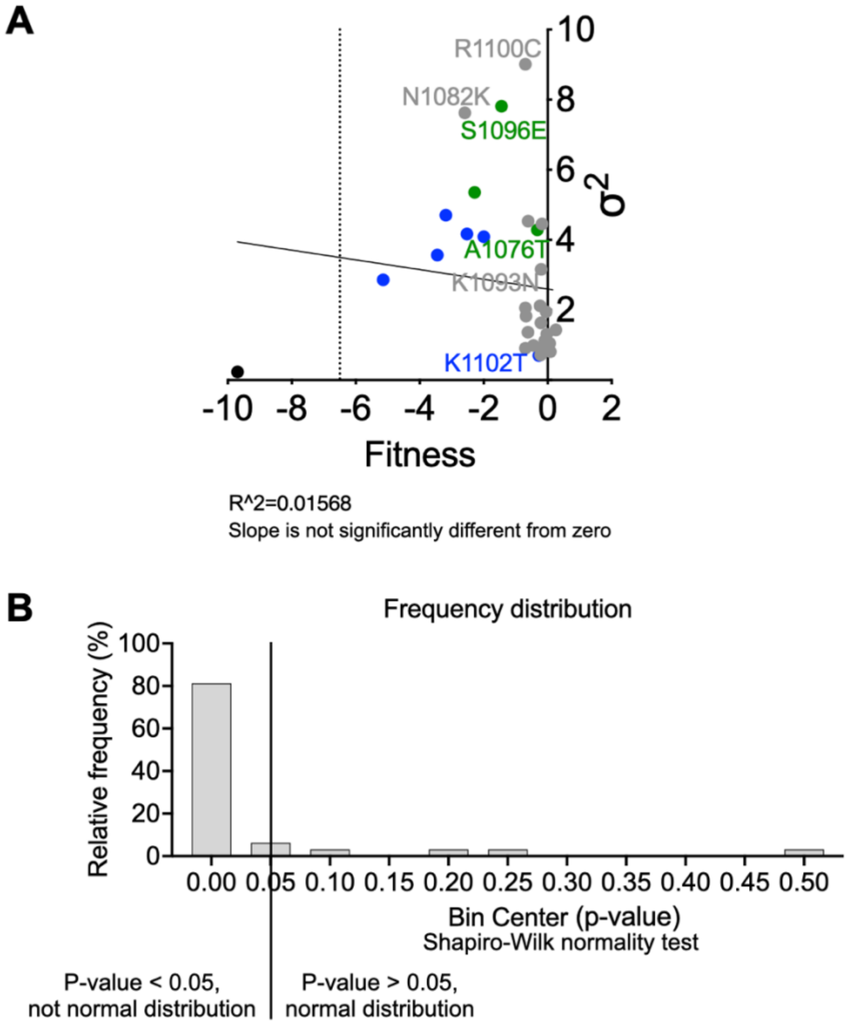


Figure 41. Slight but not significant negative correlation between σ^2 and fitness of substitutions.

(**A**). The correlation between the maximum likelihood estimate (σ2) and the fitness of each substitution. Simple linear regression was applied (Y = -0.1384*X + 2.600, R^2 = 0.01568). The slope is not significantly different from zero (P = 0.5101). (**B**). The histogram of p-values from normality test. The density plots of secondary deviation scores have been tested whether they follow normal distribution by Shapiro-Wilk normality test. The null hypothesis of the test is the distribution follows normal distribution. Substitutions with P-value > 0.05 suggest their distributions are normal and P-value < 0.05 indicate their distributions do not follow normal distribution. ~80% of substitutions do not follow normal distribution.

3.3 Discussion

RNA polymerase TL function and evolution is impacted by extensive residue interactions within and around it (Duan et al., 2023; Kaplan et al., 2012). The TL shows remarkable conservation consistent with its essential tasks for executing mechanisms in all msRNAPs (Belogurov & Artsimovitch, 2019; Cramer, 2002; Mazumder et al., 2020; Werner & Grohmann, 2011). Our observation (Duan et al., 2023) of widespread TL incompatibility in the S. cerevisiae Pol II context suggested epistasis external to the TL is the major source of epistasis for TL function (Fig**ure 26**). However, here we observed TL-internal epistatic interactions and residue coupling, although TL-internal coupling did not significantly maintain TL function when the interaction between TL and its greater evolutionary context was broken. The analyses in our recent work (Duan et al., 2023) and here indicate interactions formed by TL residues are integrated into the broader residue epistasis network within the whole enzyme (Fig**ure 25**). In summary, our results are in line with observations from other studies where the same mutation led to different phenotypes when introduced into homologous proteins (Doud et al., 2015; Haddox et al., 2018; Kondrashov et al., 2002; Lunzer et al., 2010; Natarajan et al., 2013; Park et al., 2022; Starr & Thornton, 2016; Viktorovskaya et al., 2013), highlighting the divergent nature of epistasis networks in highly conserved proteins like msRNAPs.

Even considering our above findings, our experiments here provide case studies for examples of complex interactions during evolutionary divergence. The divergence in epistasis networks of homologs accumulates through substitutions. How substitutions alter existing epistasis and shape the effects of future substitutions is critical for understanding mechanisms underlying protein functional divergence and evolution (Johnson et al., 2023; Starr & Thornton, 2016; Xie, Sun, Wang, Lehner, & Li, 2023). Using seven selected Pol II TL variants, we comprehensively determined the higher-order epistatic interactions these substitutions create. For example, S1096E exhibited strong suppression on all synthetic sick/lethal combinations of E. invadens IP1 TL haplotype substitutions (Fig**ure** 32 **and 34**), suggesting a potential path for TL evolution that would require the 1096E substitution prior to additional substitutions that lead to synthetic sickness/lethality. A caveat to this interpretation is that additional substitutions outside the TL might reshape epistasis across the entire TL. Additionally, A1076T, a substitution that consistently showed positive epistatic effects when introduced into different genetic backgrounds (**Figure 39**), was also subject to sign epistasis in certain backgrounds. Here, we would propose that additional substitutions that prevent this sign epistasis might precede appearance of a combination that would otherwise incur a fitness defect with A1076T (Fig**ure 35E**-F). A1076T-involved sign epistasis exemplifies intricate layers of substitution interactions. Emerging substitutions modify pre-existing interactions, ultimately influencing the fate of future mutations and the trajectory of protein evolution.

We further inquired whether all individual substitutions had the potential to change the pre-existing epistasis networks in the protein background they were introduced into, and to what extent. To investigate this, we analyzed the strength and consistency of the epistatic effects of substitutions across diverse genetic backgrounds. The epistatic effects of TL substitutions can be categorized into three distinct groups, with approximately 37.5% of TL substitutions consistently demonstrating medium to strong and stable epistatic effects, either positive or negative (**Figure 39A-B**), reflecting their important role in reshaping the interactions among fixed historical substitutions and influencing the phenotypes of upcoming mutations. Substitutions that consistently exhibit positive epistatic effects, such as A1076T and S1096E (**Figure 39C**), are referred to as "permissive substitutions" (Starr & Thornton, 2016). They have the potential to make certain mutations accessible that would otherwise remain inaccessible. An obvious path for evolutionary divergence in msRNAPs would be the incorporation of substitutions with mild effects on fitness but with biochemical properties of increased catalytic activity. These GOF alleles would have properties of suppression of LOF alleles that might otherwise be selected against. Conversely, substitutions with a consistent negative effect, like K1093N and R1100C, may act as "restrictive substitutions", limiting the accessibility of some mutations (Starr & Thornton, 2016). A limitation of our analyses is that our experiments tested a limited number of backgrounds. In the future, our platform has the capability to determine if the case studies presented here are rarer or the norm in the TL. In summary, our analyses provide a framework for understanding complicated epistatic interactions of substitutions and examples of how fitness landscapes of mutants are changed due to epistasis.

3.4 Methods

3.4.1 Experimental data

Experimental data were collected as described in Duan et al (Duan et al., 2023). Briefly, we synthesized TL mutant libraries including 662 natural TL variants from bacterial, archaeal, and eukaryotic RNA polymerases, 1987 TL haplotypes encompassing every possible substitution combination among seven selected Pol II TL natural variants, 724 TL haplotypes specifically designed for coupling analysis, and 620 TL single mutants serving as control to determine residue interactions. Approximately 15% wild-type *S. cerevisiae* Pol II TL allele of total variants was incorporated into each TL library for accurate quantification. All TL alleles were amplified and transformed into yeast strain CKY283 along with modified RPB1-encoding plasmid to allow construction of complete Rpb1 TL mutants within yeast through gap repair. After transformation, Leu+ colonies were collected, and re-plated into subsequent selection plates. Growth defects and phenotypes of mutants were assessed through amplicon sequencing with Illumina Next-seq (150nt reads). We constructed amplicon sequencing libraries by performing emulsion PCR on the TL region with optimized cycles to preserve the original allele frequencies. Subsequently, we employed two barcodes for multiplexing different samples. Raw sequencing data has been deposited on the NCBI SRA (Sequence Read Archive) database under BioProject PRJNA948661

3.4.2 Data cleaning and normalization

Each TL variant library was constructed and screened in three biological replicates. Read counts were estimated with a codon-based alignment algorithm (Sing-Hoi Sze, 2018) and the median counts of the three replicates was used in the analysis. Mutant with read counts coefficients of variation > 0.5 in the control condition were excluded from the analysis due to low reproducibility. We calculated fitness scores by comparing allele frequency shifts before and after selections to the shifts of wild type in the following formula. And to enable direct comparisons, we applied min-max normalization to standardize median growth fitness across all libraries. Residue with fitness equal or smaller to -6.5 is in the severe sick and lethality range and their fitness are normalized to the lethal threshold -6.5. Processed mutant count, fitness and processing codes are available through GitHub (https://github.com/Kaplan-Lab-Pitt/TLs_Screening.git).

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

3.4.3 Determination of functional interactions

Functional interactions were represented by genetic interactions calculated from the log additive model (**Figure 24**). Primary deviation score of a putative haplotype (abc) = Fitness (abc) – [Fitness (a) + Fitness (b) + Fitness (c)].

We further selected seven Pol II TL alleles and constructed all intermediate substitution combinations. For example, the putative TL haplotype has three different residues compared with wild-type *S. cerevisiae* Pol II TL allele a, b and c, we constructed all combinations of substitutions, which are a, b, c, ab, ac, bc, and abc. And we can calculate the secondary deviation scores for each substitution a, b and c. The secondary deviation score of a are calculated from all haplotypes containing a: Secondary deviation score of a_1 = Fitness (ab) – Fitness (b). Secondary deviation score of a_2 = Fitness (ac) – Fitness (c). Secondary deviation score of a_3= Fitness (abc) – Fitness (bc).

Primary and secondary deviation scores cannot be determined if the expected and observed fitness are both in the ultra-sick/lethal range (both fitness < -6.5). Haplotypes with expected fitness < -6.5 can be separated into two different situations. Firstly, all constituent substitutions are viable, but the sum of the fitness could be smaller than -6.5. This is lethality due to additivity and indicated with a dark gray block in the deviation score heatmaps. Secondly, some of the constituent substitutions or combinations are in lethal range. This lethality is because the haplotype contains at least one single substitution and indicated with a black block in the deviation heatmap.

Functional interactions are determined from the deviation scores. Positive interaction (suppression): deviation score > 1. Additive (No interaction): -1 ≤ deviation score ≤ 1. Negative interaction (synthetic sickness/lethality): deviation score < -1.

Maximum likelihood estimate (σ^2) of a certain substitution is calculated to represent its overall epistatic effect (Park et al., 2022). For a substitution with n non-repetitive secondary deviation scores, its maximum likelihood estimate: σ^2 =

$$\frac{\sum_{n}(Secondary\ deviation\ score\ of\ a\ substitution)^{2}}{n}$$

3.4.4 Coupling analysis

Coupling analysis was done as described in Russ et al (Russ et al., 2005) and Socolich et al (Socolich et al., 2005). Briefly, three independent TL haplotypes libraries were constructed to study intra-TL co-evolutionary residues. (a). Natural variants library, which consists of 362 selected natural eukaryotic Pol I, II, and III TL variants. Multiple sequence alignment (MSA) and coupling analysis based on the MSA were done for the selected TL alleles to detect the residue coupling information (**Figure 26A**). (b). Random scrambling library, where amino acids in the same position of the MSA from (a) were randomly swapped. This swapping does not disturb sequence conservation, because the composition of different amino acids in the position remains unchanged but disrupts residue coupling among different residues. Coupling analysis was done for the random scrambled MSA to confirm disrupted residue coupling (**Figure 26B**). (c). Monte Carlo scrambling library, where the amino acids in the same position were swapped and then assessed by the Monte Carlo algorithm. If the swapping perturbs residue coupling, we withdraw the swap; if not, we keep the swap. By doing this, the Monte Carlo scrambling library retains both sequence conservation and residue coupling information (**Figure 26C**). The swapping in both cases breaks the epistasis between TL residues and the Pol II context. TL haplotypes were synthesized based on three types of MSA and then screened for mutant phenotypes through phenotypic system (Duan et al., 2023; Qiu et al., 2016; Qiu & Kaplan, 2019).

4.0 Summary and future directions

4.1 Summary

This dissertation determined residue interaction networks within and around the TL, a key transcription domain in the *S. cerevisiae* Pol II active site using deep mutational scanning. In Chapter 2, we explored pairwise residue interaction networks within the TL and between the TL and other domains in Pol II active site using double mutants. In Chapter 3, we dissected higher-order residue interactions within TL haplotypes.

Pol II TL interaction landscape provides insight into the nature of lethal or previously unclassified mutants. The nature of lethal mutants is unlikely to be detected by biochemical or other methods. Using the natural suppression interaction between GOF and LOF mutants, lethal or previously unclassified mutants could be speculated. Lethal mutants caused by excessively slow transcription (LOF) are likely to be suppressed by most fast transcription mutants (GOF). Conversely, and this is relatively rare, lethal mutants due to too fast transcription (GOF) could be suppressed by most LOF. Our analysis revealed among 23 ultra-sick or lethal mutants analyzed (**Figure 9F**), 17 were suppressed by most GOF mutants, suggesting their lethality came from extreme LOF behavior like too slow catalysis. Conversely, 2 mutants were suppressed by most LOF mutants, indicating they confer phenotypes similar to GOF mutants but more severe, such as extremely fast transcription catalysis with compromised fidelity, leading to their lethality. Some unclassified mutants were suppressed by GOF, indicating atypical LOF that were not detected before or sign epistasis (**Figure 9G**). These observations suggest the double mutant interaction landscape increases the resolution beyond previously reported phenotypic analysis.

How do mutations alter protein function? Mutations alter protein function by replacing a WT amino acid with a new one. In addition to the changed biochemical features by substitution, the process leads to deeper molecular changes, where substitution disrupts the existing epistatic interaction networks involving the WT residue and introduces new interactions with the added amino acid. The resulting interaction networks vary depending on the nature of the new residue and its surrounding amino acids, illustrating a complex layer of molecular interactions beneath the surface of a simple amino acid substitution. To investigate the interaction networks that TL mutations could create, we selected 12 previously phenotyped GOF or LOF mutations, including these at the same residue at the TL or at the domains around the TL, and determined their interaction networks with the TL residues by 7,200 double mutants. We identified that none of the 12 mutations share identical interaction networks, including those that are within the same class. Strikingly, mutations at the same position, and share similar catalytic defects that both are slow, H1085L and H1085Y, have different interaction networks. For example, the growth defects of H1085Y could be suppressed by almost all 20 substitutions at E1103, whereas H1085L are synthetic sick with almost all substitutions at E1103, implying that H1085 requires the WT glutamine acid at 1103 to function properly. The change of the interaction networks may change the compatibility of future mutations, therefore shapes the potential evolution path of the TL. This emphasizes the need to consider both the biochemical nature of substitutions and the resultant changes in epistasis interaction networks when interpreting protein mechanisms through mutations.

What are the prevalence and strength of residue interactions in Pol II active site? The ultra- conserved TL domain may have experienced intense selection stress during evolution, leading to a prediction that the range and magnitude of epistasis within the TL may differ from other proteins. With about 12,000 double mutants covering interactions within the TL and between the TL and its surround domains, we determined the prevalence of epistasis by the deviation of the observed fitness of double mutants from the prediction of the log additive model, and captured notable deviations (R^2=0.21) (**Figure 9B**), which is much lower than the R^2 reported in other studies (about 0.65-0.75) (Araya et al., 2012; Fowler et al., 2010; X. Lin et al., 2022; Melamed et al., 2013; Starr & Thornton, 2016). Among these interactions, approximately 15% of observed have a strong deviation from the predicted fitness (deviation score > |2|) (**Figure 17D**), higher than the 5% in other studies. The observed high rate of strong interactions in Pol II mutants could be attributed to the much higher rate of suppressive interactions due to Pol II mutants having opposite effects on catalysis. However, we cannot rule out the possibility that in highly conserved protein like the TL, these interactions are more robust and prevalent.

Which TL residues functionally interact with each other? Our previous genetic studies have identified different types of residue interactions within the TL. To systematically understand how TL residues interacts with each other to ensure proper transcription, we designed 3790 double mutants covering double residue interactions between any two TL residues. We identified over half of the combinations exhibiting either positive or negative interactions. Among these, ~6.52% show sign epistasis or epistasis, implying functional dependence within TL residues. These interactions were distributed throughout the TL and covered every TL residue, supporting connectivity across the TL. Observed epistasis, concentrated within the C-terminal TL helix and adjacent regions, aligns with the hypothesized role of the C-terminal TL residues in maintaining the TL's open state (**Figure 9B-D**).

How do residue interactions affect TL evolution? Substitutions alter existing epistatic interactions, converting previously incompatible mutations into compatible ones, or vice versa. This alteration has profound implications on protein evolution, potentially redirecting the evolutionary path of proteins. To determine how variants of TL residues alter its potential evolutionary path, we comprehensively determined the higher-order epistatic interactions within TL haplotypes using seven selected Pol II TL variants, and all possible substitution combinations along the evolutionary path of the selected variants. We identified intricate layers of residue interactions, indicating potential TL evolutionary paths. For example, S1096E suppresses all synthetic sick/lethal mutant combinations in the *E. invadens IP1* TL variant (**Figure 32**), suggesting a potential evolutionary trajectory for the TL, where 1096E substitution may be required to appear prior to additional substitutions that lead to synthetic sickness/lethality in evolution. Moreover, The A1076T substitution, exhibiting positive epistatic effects across various genetic backgrounds, may show negative interaction (sign epistasis) in specific contexts. We hypothesize that other mutations, which counteract this sign epistasis, might occur before combinations that lead to fitness defects with A1076T (**Figure 35**). Furthermore, we classified mutations based on their epistatic effect strength into three groups. Those with medium to strong effects may have the potential to strongly influence TL variant interactions during evolution, particularly, those mutations that consistently show positive effects across various backgrounds. They enable otherwise inaccessible mutations to occur, such as the GOF mutants, which can suppress LOF mutations, preventing them to be selected against. These examples highlight the complex nature of substitution interactions and their role in shaping effects of future mutations and TL evolution pathways.

4.2 Future directions

How similar are the residue interaction networks in yeast RNA polymerases I-III? Emerging mutations modify pre-existing epistatic interaction networks within proteins and influence the accessibility of subsequent mutations. These cumulative changes in epistasis networks lead to diverse mechanisms, functions, and evolutionary direction, even among homologous proteins such as the three eukaryotic msRNAPs and their highly conserved active sites. To understand potential specialization or divergence among these homologs, we comprehensively determine the residue interaction networks for the highly related TLs in the active sites of yeast Pol I and III for comparison with our studies on yeast Pol II. Using our established high throughput platform for screening mutant phenotypes and identifying residue interactions, we plan to systematically compare and contrast the phenotypic landscapes by applying all possible 20 substitutions at each residue for Pol I and Pol III TLs, and examining the residue interaction landscapes within and around Pol I and Pol II TLs using the analogous mutations that we used to detect the interactions in Pol II TL. We have generated the designed mutant libraries, with around 8000 double mutants for both Pol I TL and Pol III TL. They are ready to be transformed into yeast for phenotyping and will be sequenced for analysis. More details are shown in Appendix B.

How are the functions of other Pol II active site domains, such as the BH, altered by mutations, and how do they interact with the TL? Similar to the TL, the BH is a conformationally flexible domain in msRNAPs' active sites that may play critical roles through TL interaction and allowing enzyme translocation (**Figure 2**). Understanding the cooperation between the BH and TL will be important to understand how they facilitate RNA polymerase activity. We plan to detect the phenotypic landscape by deep mutational scanning of all possible single substitutions in the BH (BH single mutant library). We have synthesized the BH single mutant library and finished the high-throughput phenotyping, and these samples await sequencing. More details about this are shown in **Appendix B**. In the future, we will pick mutations in the TL to combine with all BH single mutants to detect their interaction networks.

Do transcription factors allosterically affect TL function? The function of the Pol II TL is potentially regulated by transcription factors allosterically. We plan to detect potential allosteric effects in two ways. First, to understand what TL residues respond to transcription elongation factors such as TFIIS, Spt4/Spt5, Elf1 and Paf1C components, we plan to detect how the TL and BH phenotypic landscapes shift when an elongation factor is mutated or deleted. We will transform the TL and BH single mutant libraries into strains containing mutations of elongation factors in their genome and perform our high throughput screening experiment. Second, residue interaction networks may serve as allosteric pathways for the active site being regulated. Residues in these allosteric pathways may have been through co-evolution and therefore can be detected by statistically coupling analysis. We have applied the method to the largest subunit Rpb1 of Pol II and identified groups of co-evolving residues in the Rpb1 subunit that may represent pathways converge onto the active site. Because of the limit in the analysis, only one subunit of Pol II is involved. To identify the potential allosteric pathways from the whole Pol II to the active site, we plan to extend it by including all subunits of Pol II into the analysis in the future.

Does epistasis drive the evolution of RNA polymerase? RNA polymerase is essential for gene expression, regulated by internal residue interactions and interactions with regulatory proteins or molecules. Although its structure and function are conserved, RNA polymerase evolves to suit the specific needs of the host cells, such as variations in transcription elongation rate. An open question is what drives the evolution of RNA polymerases. The evolution of RNA polymerase can be influenced by various factors, including epistatic interactions. Epistasis involves residues or genes masking each other's effects, which affect protein structure stability, function and evolution of complex systems like RNA polymerases. This is because mutations in RNA polymerase’s regulatory elements can influence the function and efficiency of the entire complex, and initial mutations in RNA polymerase can influence the occurrence of subsequent mutations within the enzyme. Our results support the idea that epistasis influences the evolution of RNA polymerase. We observed widespread epistatic effects across the Pol II active site, with mutations such as A1076T and S1096E being part of complex sign epistasis networks that demonstrate intricate layers of substitution interactions. In the future, ancestral sequence reconstruction of TL sequences could provide evidence for ordering of substitutions in the TL’s evolutionary history wherein GOF mutations may build in buffering for tolerance of LOF mutations.

Appendix A Supplemental materials for Chapter 2

Appendix A.1 Supplemental tables

Appendix Table 1. Strains and plasmids

Name	**Genotype**	**Note**
pCK892	*LEU2 CEN ARS ampr ColE1* ori *rpb1* ΔTrigger Loop,T69 corrected	WT *RPB1* plasmid with TL deleted
pCK2193	*LEU2 CEN ARS ampr ColE1 ori rpb1* T834P ΔTrigger Loop	T834P plasmid with TL deleted
pCK2194	*LEU2 CEN ARS ampr ColE1 ori rpb1* T834A ΔTrigger Loop	T834A plasmid with TL deleted
pCK2198	*LEU2 CEN ARS ampr ColE1 ori rpb1* S713P ΔTrigger Loop	S713P plasmid with TL deleted
CKY283	*ura3-52 his3Δ200 leu2Δ1 or Δ0 trp1Δ63 met15Δ0 lys2-128∂ gal10Δ56 rpb1Δ::CLONATMX RPB3::TAP::KlacTRP1*	WT yeast strain
CKY3208	*ura3-52 his3Δ200 leu2Δ1* or *Δ0 trp1Δ63 met15Δ0 lys2-128∂ gal10Δ56 rpb1Δ::CLONATMX RPB3::TAP::KlacTRP1 rpb2* Y769F	*rpb2* Y769F strain

Appendix Table 2. Primers

Name	**Sequence**	**Description**	**Amplification**	**Schematic**	**Note**
CKO2751	ACACTCTTTCCCTACACGACGCTCTTCCGATCTG	TL Library amplification	TL library 1	R1-BC1-N-TL-F	Custom primer

	TCGGTAATTAGCAGCCCAATCCATTGGTG	(First amplification)			
CKO2752	ACACTCTTTCCCTACACGACGCTCTTCCGATCTAGGTCACTAGTAGCAGCCCAATCCATTGGTG	TL Library amplification (First amplification)	TL library 2	R1-BC2-NN-TL-F	Custom primer
CKO2753	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGAATCCGACACTAGCAGCCCAATCCATTGGTG	TL Library amplification (First amplification)	TL library 3	R1-BC3-NNN-TL-F	Custom primer
CKO2754	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGTACCTTGGCTCTAGCAGCCCAATCCATTGGTG	TL Library amplification (First amplification)	TL library 4	R1-BC4-NNNN-TL-F	Custom primer
CKO2755	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCATGAGGATTCGCTAGCAGCCCAATCCATTGGTG	TL Library amplification (First amplification)	TL library 5	R1-BC5-NNNNN-TL-F	Custom primer
CKO2756	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGACTGACACGAACTAGCAGCCCAATCCATTGGTG	TL Library amplification (First amplification)	TL library 6	R1-BC6-NNNNNN-TL-F	Custom primer
CKO2757	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTCAGACGAGAGTTGTTA	TL Library amplification (First amplification)	TL library 7	R1-BC7-NNNNNNN-TL-F	Custom primer

	GCAGCCCAATCCATTGGTG				
CKO2758	ACACTCTTTCCCTACACGACGCTCTTCCGATCTGATAGGCTCTCTGTGTTAGCAGCCCAATCCATTGGTG	TL Library amplification (First amplification)	TL library 8	R1-BC8-NNNNNNNN-TL-F	Custom primer
CKO2759	ACACTCTTTCCCTACACGACGCTCTTCCGATCTTGGTACAGTGTGCTCCTTAGCAGCCCAATCCATTGGTG	TL Library amplification (First amplification)	TL library 9	R1-BC9-NNNNNNNNN-TL-F	Custom primer
CKO2760	ACACTCTTTCCCTACACGACGCTCTTCCGATCTCAAGGTCTGGACAGTTATTAGCAGCCCAATCCATTGGTG	TL Library amplification (First amplification)	TL library 10	R1-BC10-NNNNNNNNNNN-TL-F	Custom primer
CKO2761	GTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTGAAGGGGTTTTCATGTTTTTGG	TL Library amplification (First amplification)	TL libraries	R2-TL-R	Custom primer
P17-B5	AGGATAGC	NEB index oligos (Second amplification)	Condition 1-1		NEB index oligos
P18-B6	CCTTCCAT	NEB index oligos (Second amplification)	Condition 1-2		NEB index oligos

P19-B7	GTCCTTGA	NEB index oligos (Second amplification)	Condition 1-3		NEB index oligos
P20-B8	TGCGTAAC	NEB index oligos (Second amplification)	Condition 2-1		NEB index oligos
P21-B9	CACAGACT	NEB index oligos (Second amplification)	Condition 2-2		NEB index oligos
P22-B10	TTACGTGC	NEB index oligos (Second amplification)	Condition 2-3		NEB index oligos
P23-B11	CCAAGGTT	NEB index oligos (Second amplification)	Condition 3-1		NEB index oligos
P24-B12	CACGCAAT	NEB index oligos (Second amplification)	Condition 3-2		NEB index oligos
P53-E5	CCGCTTAA	NEB index oligos (Second amplification)	Condition 3-3		NEB index oligos
P54-E6	TACCTGCA	NEB index oligos (Second amplification)	Condition 4-1		NEB index oligos
P55-E7	GTCGATTG	NEB index oligos (Second amplification)	Condition 4-2		NEB index oligos

P56-E8	TATGGCAC	NEB index oligos (Second amplification)	Condition 4-3		NEB index oligos
P57-E9	CTCGAACA	NEB index oligos (Second amplification)	Condition 5-1		NEB index oligos
P58-E10	CAACTCCA	NEB index oligos (Second amplification)	Condition 5-2		NEB index oligos
P59-E11	GTCATCGT	NEB index oligos (Second amplification)	Condition 5-3		NEB index oligos
P60-E12	GGACATCA	NEB index oligos (Second amplification)	Condition 6-1		NEB index oligos
P85-H1	TACTCCAG	NEB index oligos (Second amplification)	Condition 6-2		NEB index oligos
P86-H2	GGAAGAGA	NEB index oligos (Second amplification)	Condition 6-3		NEB index oligos
P87-H3	GCGTTAGA	NEB index oligos (Second amplification)	Condition 7-1		NEB index oligos
P88-H4	ATCTGACC	NEB index oligos (Second amplification)	Condition 7-2		NEB index oligos

P89-H5	AACCAGAG	NEB index oligos (Second amplification)	Condition 7-3		NEB index oligos
P90-H6	GTACCACA	NEB index oligos (Second amplification)	Condition 8-1		NEB index oligos
P91-H7	GGTATAGG	NEB index oligos (Second amplification)	Condition 8-2		NEB index oligos
P92-H8	CGAGAGAA	NEB index oligos (Second amplification)	Condition 8-3		NEB index oligos
P93-H9	CAGCATAC	NEB index oligos (Second amplification)	Condition 9-1		NEB index oligos
P94-H10	CTCGACTT	NEB index oligos (Second amplification)	Condition 9-2		NEB index oligos
P95-H11	CTTCGGTT	NEB index oligos (Second amplification)	Condition 9-3		NEB index oligos
P96-H12	CCACAACA	NEB index oligos (Second amplification)	Condition 0		NEB index oligos

Appendix Table 3. Phenotyping details

Condition	**Cells plated**	**Days for phenotyping**

SC-Leu (Pre)	30K	2 days
SC-Leu + 5FOA	400K	3 days
SC-Leu (Post)	400K	2 days
SC-Lys	1M	7 days
YPRaf	400K	6 days
YPRafGal	500K	6 days
SC-Leu + 20µg/mL MPA	1M	4 days
SC-Leu + 15mM Mn	1M	5 days
SC-Leu + 3% Formamide	500K	3 days

Appendix Table 4. Mutant counts of libraries

Index	**Handled library**	**WT**	**Allele number**
TL Lib1	Singles	111	620
TL Lib2	Pairwise Doubles	700	3914
TL Lib3	Evo present	123	662
TL Lib4	Evo Path	356	1987
TL Lib5	Coupling	130	724
TL Lib6	Target Doubles	858	4800
TL Lib7	T834P-target	111	621
TL Lib8	T834A-target	111	621
TL Lib9	Y769F-target	111	621
TL Lib10	S713P-target	111	621
Total		2722	15191

Appendix Table 5. Multiple logistic regression model for predicting GOF mutants

GOF Model				

Parameter estimates	Variable	Estimate	Standard error	95% CI (profile likelihood)
β0	Intercept	-1.816	1.549	-5.879 to 0.8657
β1	MPA	-2.542	1.294	-6.196 to -0.6109
β2	Lys	1.942	1.121	0.4256 to 5.021
β3	Gal	-0.0657	0.282	-0.7183 to 0.5200
β4	MPA : Lys	0.5297	0.2843	-0.2079 to 1.322
β5	MPA : Gal	0.08373	0.276	-0.5768 to 0.5721
β6	Lys : Gal	-0.0256	0.2288	-0.2720 to 0.5072
Odds ratios	Variable	Estimate	95% CI (profile likelihood)	
β0	Intercept	0.1626	0.002798 to 2.377	
β1	MPA	0.07874	0.002037 to 0.5429	
β2	Lys	6.974	1.530 to 151.6	
β3	Gal	0.9364	0.4876 to 1.682	
β4	MPA : Lys	1.698	0.8123 to 3.752	
β5	MPA : Gal	1.087	0.5617 to 1.772	
β6	Lys : Gal	0.9748	0.7618 to 1.661	
Area under the ROC curve				
Area	0.9889			
Std. Error	0.009628			
95% confidence interval	0.9700 to 1.000			
P value	<0.0001			
Classification table	Predicted Not GOF	Predicted GOF	Total	% Correctly classified

Observed Not GOF	36	0	36	100
Observed GOF	2	23	25	92
Total	38	23	61	96.72
Negative predictive power (%)	94.74			
Positive predictive power (%)	100			

Appendix Table 6. Multiple logistic model for predicting LOF mutants

LOF Model				
Parameter estimates	Variable	Estimate	Standard error	95% CI (profile likelihood)
β0	Intercept	-1.916	1.327	-5.389 to 0.3409
β1	MPA	1.392	1.508	-0.7599 to 5.252
β2	Spt	1.328	0.9944	-0.04761 to 4.209
β3	Gal	0.8353	0.4913	0.1660 to 2.396
β4	MPA : Spt	0.01112	0.1707	-0.3174 to 0.5524
β5	MPA : Gal	0.2992	0.3107	-0.2797 to 1.150
β6	Spt : Gal	-0.8823	0.6035	-2.785 to -0.1533
Odds ratios	Variable	Estimate	95% CI (profile likelihood)	
β0	Intercept	0.1472	0.004566 to 1.406	
β1	MPA	4.021	0.4677 to 191.0	
β2	Spt	3.774	0.9535 to 67.29	
β3	Gal	2.305	1.181 to 10.98	
β4	MPA : Spt	1.011	0.7281 to 1.737	
β5	MPA : Gal	1.349	0.7560 to 3.159	

β6	Spt : Gal	0.4138	0.06175 to 0.8579	
Area under the ROC curve				
Area	0.9914			
Std. Error	0.007495			
95% confidence interval	0.9767 to 1.000			
P value	<0.0001			
Classification table	Predicted 0	Predicted 1	Total	% Correctly classified
Observed 0	32	0	32	100
Observed 1	3	26	29	89.66
Total	35	26	61	95.08
Negative predictive power (%)	91.43			
Positive predictive power (%)	100			

Appendix Table 7. Kruskal-Wallis tests

Dunn's multiple comparisons test	**Mean rank diff.**	**Significant?**	**Summary**	**Adjusted P Value**
S713P vs. Y769F	118.1	No	ns	0.8279
S713P vs. E1103G	7.282	No	ns	>0.9999
S713P vs. L1101S	270.8	Yes	****	<0.0001
S713P vs. F1084I	383	Yes	****	<0.0001
S713P vs. M1079V	-62.19	No	ns	>0.9999
S713P vs. T834P	610.7	Yes	****	<0.0001
Y769F vs. E1103G	-110.8	No	ns	>0.9999
Y769F vs. L1101S	152.7	No	ns	0.1754

Y769F vs. F1084I	264.9	Yes	****	<0.0001
Y769F vs. M1079V	-180.3	Yes	*	0.0393
Y769F vs. T834P	492.6	Yes	****	<0.0001
E1103G vs. L1101S	263.5	Yes	***	0.0001
E1103G vs. F1084I	375.7	Yes	****	<0.0001
E1103G vs. M1079V	-69.47	No	ns	>0.9999
E1103G vs. T834P	603.4	Yes	****	<0.0001
L1101S vs. F1084I	112.2	No	ns	>0.9999
L1101S vs. M1079V	-333	Yes	****	<0.0001
L1101S vs. T834P	339.9	Yes	****	<0.0001
F1084I vs. M1079V	-445.2	Yes	****	<0.0001
F1084I vs. T834P	227.7	Yes	**	0.0017
M1079V vs. T834P	672.9	Yes	****	<0.0001
Dunn's multiple comparisons test	**Mean rank diff.**	**Significant?**	**Summary**	**Adjusted P Value**
H1085L vs. H1085Y	236.3	Yes	****	<0.0001
H1085L vs. N1082S	6.459	No	ns	>0.9999
H1085L vs. Q1078S	244.4	Yes	****	<0.0001
H1085L vs. T834A	-25.35	No	ns	>0.9999
H1085Y vs. N1082S	-229.8	Yes	****	<0.0001
H1085Y vs. Q1078S	8.097	No	ns	>0.9999
H1085Y vs. T834A	-261.6	Yes	****	<0.0001
N1082S vs. Q1078S	237.9	Yes	****	<0.0001
N1082S vs. T834A	-31.8	No	ns	>0.9999
Q1078S vs. T834A	-269.7	Yes	****	<0.0001

Appendix A.2 Adding Supplemental Documents

Appendix A.2.1 High efficiency large scale chemical yeast transformation protocol

We optimized the Gietz protocol(Gietz & Schiestl, 2007), which yields 5-15 x 105 colonies with 1μg of DNA and 10 × 108 cells. However, transformation efficiency may vary with different strains.

Day 1:

Set up 5mL cultures from a single colony in YPD.

Day 2:

1. Dilute ~400-500μL of saturated culture in 50mL of YPD (210rpm, 30 °C). The cells were grown for ~3-4 hours until the concentration of the culture reached 2×10^7 cells/mL (count the cells).

2. The cells were washed in 25 ml of sterile water and resuspended in 1 ml sterile water (volume needs to be adjusted to make the final concentration 1×10^9 cells/mL).

3. Boil ssDNA for 10mins before use, and the boiled ssDNA was kept on ice for the entire duration.

4. Aliquot of 100μL (108 cells) were centrifuged at top speed for 30s, discarded the supernatant was discarded, and the cells were resuspended in the transformation mix (240μL 50% PEG 3500, 36μL 1M LiOAc, 10μL ssDNA, 74μL DNA with water to make 100ng in total, 360μL in total).

5. The mixture was then incubated at 30 °C for 30 min.

6. Tape the tubes to a 30 °C wheel for 30 min of incubation.

7. Immediately prior to heat shock, add 36μL DMSO.

8. Heat shock was performed at 42 °C for 15 min.

9. The solution was spun down for 30 s, the liquids were removed, and re-suspend in 50µL H2O.

10. Plated all 500µL on the selection plate to achieve the highest transformation efficiency.

Appendix A.2.2 Emulsion PCR set up with EURx Micellula DNA Emulsion & Purification (ePCR) PCR kit

1. The emulsion-oil phase was prepared on ice.

Appendix Table 8. Emulsion oil phase

Oil surfactant mixture (300 µl per reaction)	(µL)
Emulsion component 1	220
Emulsion component 2	20
Emulsion component 3	60

2. Mix thoroughly by vertexing at the highest level and put in a 4 °C cold room for further use.

3. Prepare the PCR water phase on ice.

Appendix Table 9. Emulsion PCR water phase

Emulsion PCR water phase using Kaplan lab Phusion for amplification from synthesized library pool	
Components	1X
10X dNTPs(µL)	5

5X detergent free Buffer(μL)	10
H2O(μL)	32
100μM Primer F (μL)	0.5
100μM Primer R (μL)	0.5
TL library template (1ng/μL)	1
Phusion(μL)	0.5
1mg/mL BSA(μL)	0.5
Total(μL)	50

4. Create emulsion reactions by mixing the 300µl precooled oil surfactant mixture and 50µl of precooled PCR water phase with vortexing at the maximum speed for 5 min in a cold room.

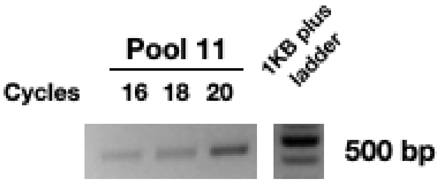
5. The solution was quickly spun down at ~1000rpm for 5 s. Dispense ~110μL aliquot into three PCR tubes.

6. PCR was performed according to the following protocol.

Note 1: Do not exceed the 95°C denaturing temperature because some buffers tend to destabilize the emulsion.

Note 2: The number of cycles was determined by Q-PCR with the same amount of template and oligos. A cycle was selected in the upper half of the linear amplification curve. Two more cycles were added to the selected cycle number because emulsion PCR tends to have a lower amplification efficiency than the standard PCR. To confirm the linearity, a further test with selected cycle, selected cycle plus two more cycles, and selected cycle plus four more cycles was applied to three emulsion PCR reactions (from the same mix). The products of the three emulsion

PCR were analyzed using agarose gel electrophoresis. The amount of PCR with the selected cycle plus two more cycles was normally between the other two cycles.



Appendix Figure 1. Example gel figure of three emulsion PCR reactions

Appendix Table 10. PCR thermal cycle

Temperature	Time	Cycle
95 °C	3min	
95 °C	15s	18 cycles
55 °C	30s	
72 °C	30s/kb	
72 °C	5 min	
12 °C	Forever	

7. Once the PCR was completed, the corresponding triplicates of each ePCR assay were pooled into a single 2 ml reaction tube. Break emulsion by adding 1.0 ml 2-butanol (or butanol). Mix by vertexing.

8. Add 400 µL of orange-colored Orange-DX buffer to the opened emulsion solution. Mix the emulsion solution with gentle agitation (e.g., on a rotator for 2 min).

9. Centrifuge for 2 min at maximum speed (e.g. 16 000 x g / approx. 14 000 rpm) for phase separation.

10. Most of the yellow-colored organic phase was removed.

11. Apply 40 µL of activation Buffer DX onto the spin column (do not spin) and keep it at room temperature until the mixture is transferred to the spin column (at least 10 min).

12. Pour the mixture (aqueous phase + interphase; max. Six hundred microliters) into a spin column/receiver tube assembly.

13. Spin down in a microcentrifuge at 12,000 rpm for 1 min, discard the flow-through.

14. Add 500 µL of Wash-DX1 buffer and spin down at 12,000 rpm (~11.000 x g) for 1 min, and discard the flow-through.

15. Add 650 µL of Wash-DX2 buffer and spin down at 12,000 rpm (~11.000 × g) for 1 min, and discard the flow-through.

16. The mixture was spun down at 12,000 rpm (~11.000 x g) for 2 min to remove traces of Wash-DX buffer.

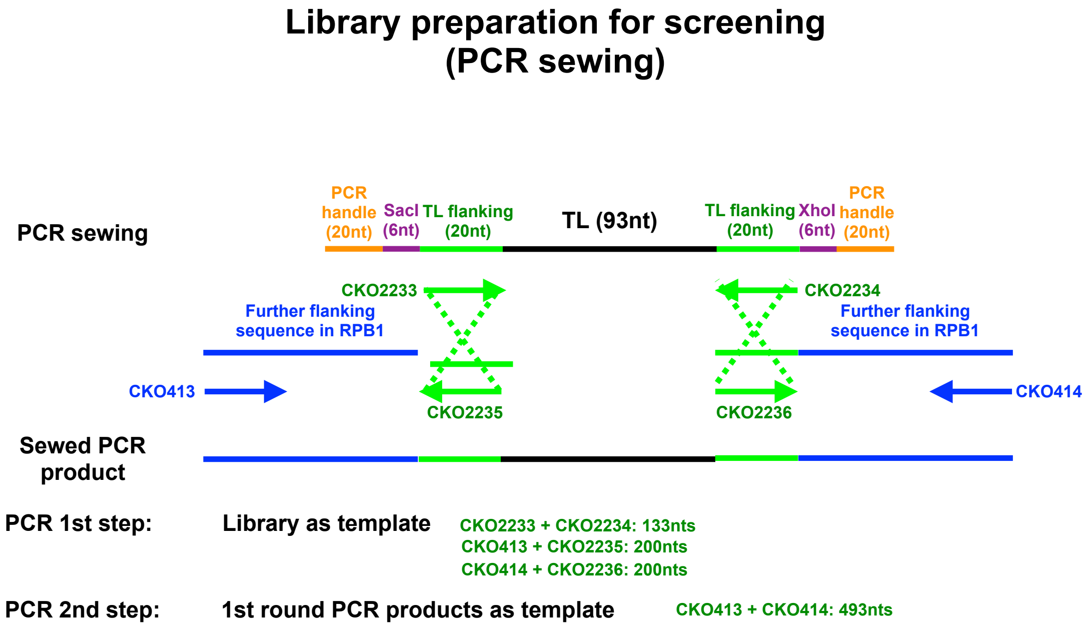
17. The spin column was placed into a new receiver tube (1.5-2 ml). Add 50-150 µl of Elution-DX buffer to elute the bound DNA.

18. Incubate the spin column/receiver tube assembly for 2 min at room temperature. Spin down at 12,000 rpm (~11.000 × g) for 1 min. The elution process was repeated.

Appendix A.2.3 Amplification of mutant libraries

1. The Agilent TL library was dissolved in 100µL of 10mM Tris buffer (pH8) as suggested by protocol. 1µL of the library was used in each 50µL PCR reaction. The number of cycles was estimated using qPCR. Two rounds of emulsion PCR were performed to obtain sufficient library products. The size of the library products was confirmed by agarose gel electrophoresis.

2. Two flanking regions were added to the TL regions amplified by PCR sequencing, as shown in the following schematic.



Appendix Figure 2. Screening library preparation by PCR sewing.

Appendix A.2.4 Transformation of mutant libraries

1. Transformation was performed using a high-efficiency large-scale chemical yeast transformation protocol (Appendix A.2.1). 100ng of MluI digested plasmid and 383ng of variants library were used in the transformation.

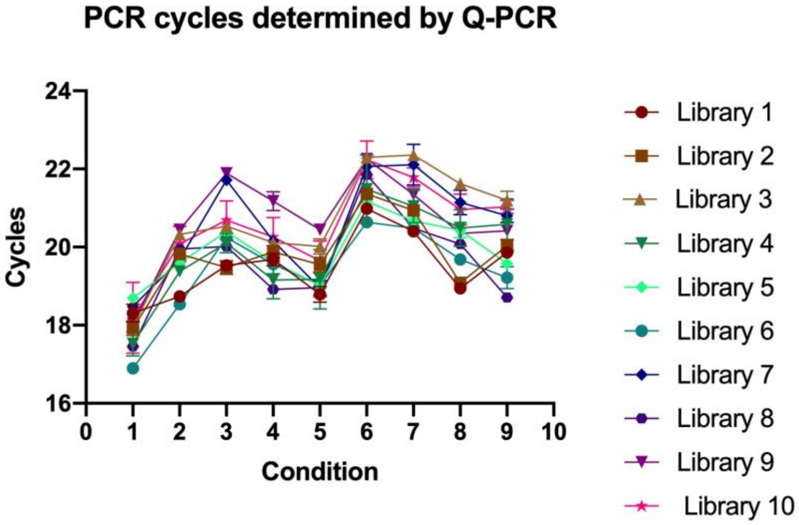
2. Screening was performed using scraping and replating. The number of cells for each condition is listed in Appendix Table 3.

Appendix A.2.5 Preparation of sequencing pool

We have 10 TL-screening mutant libraries that were tested under nine conditions; each library had three replicates, leading to 270 mutant pools (10 × 9 × 3 = 270).

1. DNA from mutant pools (n=270) was extracted using the Yeastar genomic DNA kit according to the manufacturer's instructions (Zymo Research), except that we used 1-5 x 10^8 cells.

2. To amplify the TL region from the extracted DNA and add barcodes to distinguish 10 libraries, the TL regions of 270 mutant pools were amplified by standard PCR with 10 pairs of barcoded primers (one pair of primers with a certain barcode was used for all conditions and replicates of that library). PCR cycles were determined by Q-PCR to minimize the allele frequency shifts caused by amplification. We did Q-PCR test with two or three replicates of each library to determine a cycle that was in the linear range of the amplification curve. All three replicates of each library were subjected to the same cycle, which represents the average value of the replicates in the Q-PCR test. The determined cycles are shown in the following figure.



Appendix Figure 3. Amplification cycles determined by Q-PCR.

3. PCR products from step 2, representing libraries screened under the same conditions, were pooled. We obtained 27 pools in total after a combination of 27 conditions. To limit template switching, these 27 pools were amplified using emulsion PCR technology (EURx Micellula DNA Emulsion & Purification (ePCR PCR kit). Q-PCR was performed to determine the amplification

cycle that was in the linear range for each pool. NEB barcodes containing primers were used to distinguish different conditions (NEBNext Multiplex Oligos for Illumina).

After two rounds of amplification, a sample-specific barcode sequence was added to the TL variants, and an adequate amount of TL variants was ready for sequencing. The indexed pooled samples were sequenced by single-end sequencing on an Illumina Next-seq (150nt reads). On average, over 11 million reads were obtained for individual samples with high reproducibility after two rounds of sequencing.

Appendix A.2.6 Formulas of calculating functional interactions.

For multiple mutant $M_1M_2M_3$ with count $c(M_1M_2M_3)$, we computed the ratio $rc(M_1M_2M_3) = c(M_1M_2M_3)/c(WT)$.

(1)The observed fitness of the multiple mutant $M_1M_2M_3$ is:

$$\log\,(\frac{rc\ (M1M2M3)sele}{rc(M1M2M3)unsele})$$

(2)The expected fitness of a mutant $M_1M_2M_3$ is the log additive of the constituent single mutants M_1, M_2, and M_3.

$$\log\left(\frac{rc\ (M1)sele}{rc(M1)unsele}\right)+\log\left(\frac{rc(M2)sele}{rc(M2)unsele}\right)+\log\left(\frac{rc\ (M3)sele}{rc(M3)unsele}\right)=\log\,(\frac{rc(M1)sele*rc\ (M2)sele*rc\ (M3)sele}{rc\ (M1)unele*rc\ (M2)unsele*rc\ (M3)unsele})$$

(3)We compared the fitness of $M_1M_2M_3$ with the log sum of its constituent M_1, M_2, and M_3 (compare the observed to the expected fitness), which is

$$Deviation\ score=\ \log\left(\frac{rc\ (M1M2M3)sele}{rc(M1M2M3)unsele}\right)-\log\left(\frac{rc(M1)sele*rc\ (M2)sele*rc\ (M3)sele}{rc\ (M1)unele*rc\ (M2)unsele*rc\ (M3)unsele}\right)$$

$$=\ \log\left(\frac{rc\ (M1M2M3)sele}{rc(M1)sele*rc\ (M2)sele*rc\ (M3)sele}\right)-\log\left(\frac{rc(M1M2M3)unsele}{rc\ (M1)unele*rc\ (M2)unsele*rc\ (M3)unsele}\right)$$

If

$$-1\ <\ \log\left(\frac{rc\ (M1M2M3)sele}{rc(M1)sele*rc\ (M2)sele*rc\ (M3)sele}\right)-\log\left(\frac{rc(M1M2M3)unsele}{rc\ (M1)unele*rc\ (M2)unsele*rc\ (M3)unsele}\right)<1$$

Then interaction among the constituent single mutants is additive.

If

$$\log\left(\frac{rc\ (M1M2M3)sele}{rc(M1)sele * rc\ (M2)sele * rc\ (M3)sele}\right) - \log\left(\frac{rc(M1M2M3)unsele}{rc\ (M1)unele * rc\ (M2)unsele * rc\ (M3)unsele}\right) \geq 1$$

Then the interaction is non-additive, positive interactions, including suppression and epistasis.

If

$$\log\left(\frac{rc\ (M1M2M3)sele}{rc(M1)sele * rc\ (M2)sele * rc\ (M3)sele}\right) - \log\left(\frac{rc(M1M2M3)unsele}{rc\ (M1)unele * rc\ (M2)unsele * rc\ (M3)unsele}\right) \leq -1$$

Then the interaction is non-additive, negative interactions, including synthetic sick, synthetic lethal and sign epistasis.

Appendix B Pol II BH, Pol I and Pol III TL single and double mutant libraries

Appendix B.1 Phenotyping Pol II BH single mutant library

Appendix B.1.1 Background and rationale

The BH is another highly conserved and conformationally flexible domain in the active site of msRNAPs, located adjacent to the TL (**Figure 2**). It is crucial for RNA polymerase activity through its interactions with the TL and the role in enzyme translocation. This is supported by the observation that BH mutations exhibit phenotype patterns similar to TL GOF or LOF mutants, and our double mutant interaction analysis has revealed dependent epistatic interactions between BH mutants and the TL single mutants. We aim to systematically investigate the role of the BH in RNA polymerase by examining the phenotypic landscape of BH single mutants through deep mutational scanning, focusing on two key questions**: (1) Which BH residues are particularly sensitive to mutations? (2) How are GOF and LOF mutants distributed within the BH?**

Appendix B.1.2 Experimental design

Appendix Table 11. Pol II BH single mutant library

Lib #	Content	detail	Protein	Allele amount	WT inserted	Sum
Lib13	Pol II BH singles	All possible Pol II BH single mutants	Pol II Rpb1 BH	680	108 (15%)	788

	Control	BH single mutants aliquoted from BH single mutant library	Pol II Rpb1 BH	13		13

Appendix Table 12. Pol II BH mutant screening details

Condition	**Cells plated**	**Days for phenotyping**
SC-Leu (Pre)	30K	2 days
SC-Leu + 5FOA 3days	400K	3 days
SC-Leu + 5FOA 4days	400K	4 days
SC-Leu + 5FOA 5days	400K	5 days
SC-Leu (Post)	400K	2 days
SC-Lys	1M	7 days
YPRaf	400K	6 days
YPRafGal	500K	6 days
SC-Leu + 20µg/mL MPA	1M	4 days
SC-Leu + 15mM Mn	1M	5 days
SC-Leu + 3% Formamide	500K	3 days

Appendix B.1.3 Experimental progress

Appendix Table 13. Experimental progress of BH single mutant library

Step	**Details**	**Progress**
Generating mutant library	Design	Finished
	Synthesis	Finished
	Extension/Amplification	Finished
High throughput screening	Transformation	Finished

	Screening	Finished
Deep sequencing	Construction amplicon sequencing pool	
	Deep sequencing	
	Analysis	

Appendix B.2 Detecting residue interaction networks in Pol I and Pol III active sites

Appendix B.2.1 Background and rationale

The three yeast RNA polymerases have evolved distinct residue interaction networks within their respective enzymatic backgrounds. This is evident from the fact that the analogous mutation resulted in contrasting catalytic defects in yeast Pol I and Pol II TLs, as well as the broad incompatibility of TLs from other species or polymerases when placed in the yeast Pol II background (**Figure 17-18**). To comprehensively understand the TL functions and the unique residue interaction networks in Pol I and Pol III, we aim to dissect the Pol I and Pol III TL single mutant phenotypic landscape and double mutant interaction landscape by deep mutational scanning, and compare and contrast these landscapes with those of Pol II TL. We will be focused on addressing the following questions: **(1) How similar are the Pol I and Pol III single mutant phenotypic landscapes compare to the Pol II TL landscape? (2) Can we identify distinct phenotypic patterns that differentiate Pol I and Pol III TL single mutants with GOF and LOF? (3) To what extent do the residue interaction landscapes differ among the three RNA polymerases?**

Appendix B.2.2 Experimental design

Appendix Table 14. Summary of Pol I and Pol III libraries

Lib #	**Content**	**detail**	**Protein**	**Rationale**	**Allele amount**	**WT inserted**	**Sum**
Lib11	Pol I TL singles	Pol I TL singles			620	98 (15%)	718 + 5% spike mutants
Lib14	Pol I TL Target doubles	Q1199S x singles	Rpa190 TL	Pol II TL Q1078S	600	837 (15%)	6336 + 5% spike mutants
		M1200V x singles	Rpa190 TL	Pol II TL M1079V	600		
		L1202M x singles	Rpa190 TL	Pol II TL L1081M	600		
		N1203S x singles	Rpa190 TL	Pol II TL N1082S	600		
		F1205I x singles	Rpa190 TL	Pol II TL F1084I	600		
		H1206L x singles	Rpa190 TL	Pol II TL H1085L	600		
		H1206Y x singles	Rpa190 TL	Pol II TL H1085Y	600		
		L1222S x singles	Rpa190 TL	Pol II TL L1101S	600		
		E1224G x singles	Rpa190 TL	Pol II TL E1103G	600		

		Designed single mutants	Rpa190 TL	For comparison with Pol I singles	99	99	
TL-external	Pol I TL External Target doubles	M716I x singles	Rpa135	1. M716 is the residue before Y717 (Y769 in Pol II). 2. It was identified with two parents. 3. The position we also found suppressor in Pol II.	620	15%	620 + 15% silent WT cells + 5% spike mutants
		L1484S x singles	Rpa190	1. RPA190-RPA135-RPA12.2 interacting region. 2. Multiple substitutions identified in the same position. 3. It was identified with two parents.	620	15%	620 + 15% silent WT cells + 5% spike mutants

		G1005S x singles	Rpa190 BH	1. At the BH. 2. The identical suppressor was found in Pol II	620	15%	620 +15% silent WT cells + 5% spike mutants
		F1570V x singles	Rpa190	1. We have multiple suppressors identified at the position. 2. The identical suppressor was found in Pol II	620	15%	620+15 % WT cells + 5% spike mutants
		Silent WT T1204/I1225	Rpa190	Function as WT control	15% based on cell count		
Spike mutation	Spike-in control	T1204/H1206L/I1225	Rpa190 TL	Spot assay score -3.5	5% based on cell count		
	Spike-in control	T1204/H1206Q/I1225	Rpa190 TL	Spot assay score -1			
	Spike-in control	T1204/H1206A/I1225	Rpa190 TL	Spot assay score -2.5			
	Spike-in control	T1204/H1206Y/I1225	Rpa190 TL	Observed lethal			

	Spike-in control	T1204/H1206R/I1225	Rpa190 TL	Spot assay score -2.5			
	Spike-in control	T1204/M1200H/I1225	Rpa190 TL	Very Mild defects			
	Spike-in control	T1204/V1215Q/I1225	Rpa190 TL	Very Severe defects			
Lib12	Pol III TL singles	Pol III TL singles			620	98	718 + 5% spike mutants
Lib15	Pol III TL Target doubles	Q1103S x singles	Rpo31 TL	Pol II TL Q1078S	600	837	6337 + 5% spike mutants
		M1104V x singles	Rpo31 TL	Pol II TL M1079V	600		
		L1106M x singles	Rpo31 TL	Pol II TL L1081M	600		
		K1107S x singles	Rpo31 TL	Pol II TL N1082S	600		
		F1109I x singles	Rpo31TL	Pol II TL F1084I	600		
		H1110L x singles	Rpo31 TL	Pol II TL H1085L	600		
		H1110Y x singles	Rpo31 TL	Pol II TL H1085Y	600		
		I1126S x singles	Rpo31 TL	Pol II TL L1101S	600		

		E1128G x singles	Rpo31 TL	Pol II TL E1103G	600		
		Designed single mutants	Rpo31 TL	For comparison with Pol III singles	100	100	
TL-external	Pol III TL External Target doubles	E870D x singles	Rpo31	1. At the BH. 2. Conserved position in three Pols. 3. Beside Pol II G823.	620	15%	620+15% silent WT cells + 5% spike mutants
		D1351H x singles	Rpo31	Besides Pol II H1085 suppressor Y1365C	620	15%	620 +15% silent WT cells + 5% spike mutants
		Silent WT T1108/I1129	Rpo31	Function as WT control	15% based on cell count		
Spike mutation	Spike-in control	T1108/H1110L/I1129	Rpo31 TL	Observed lethal	5% based on cell count		
	Spike-in control	T1108/H1110Q/I1129	Rpo31 TL	Spot assay score -5.5			

	Spike-in control	T1108/H1110A/I1129	Rpo31 TL	Spot assay score -6			
	Spike-in control	T1108/H1110Y/I1129	Rpo31 TL	Observed lethal			
	Spike-in control	T1108/H1110R/I1129	Rpo31 TL	Observed lethal			
	Spike-in control	T1108/R1125F/I1129	Rpo31 TL	Very Mild defects			
	Spike-in control	T1108/P1124N/I1129	Rpo31 TL	Very Mild defects			
	Spike-in control	T1108/V1119E/I1129	Rpo31 TL	Mild defects			

Appendix Table 15. Pol I and Pol III TLs mutant screening details

Condition	**Cells plated**	**Days for phenotyping**
SC-Leu (Pre)	30K	2 days
SC-Leu + 5FOA 3days	400K	3 days
SC-Leu + 5FOA 4days	400K	4 days
SC-Leu + 5FOA 5days	400K	5 days
SC-Leu (Post)	400K	2 days
SC-Leu + 0.07 µg/mL Cycloheximide	400K	Until density reaches lawn
SC-Leu + 150 mM Hydroxyurea	500K	Until density reaches lawn
SC-Leu + 15mM Mn	1M	5 days
SC-Leu + 3% Form	500K	3 days
15 ℃ (YPD)	400K	Until density reaches lawn
37 ℃ (YPD)	400K	Until density reaches lawn
30 ℃ (YPD)	400K	2 days

SC-Leu + 2ng/ml Rapamycin	500K	Until density reaches lawn
SC-Leu + 30ng/ml Rapamycin	500K	Until density reaches lawn

Appendix B.2.3 Experimental progress

Appendix Table 16. Experimental progress of Pol I and Pol III TL libraries

Step	**Details**	**Progress**
Generating mutant library	Design	Finished
	Synthesis	Finished
	Extension/Amplification	Finished
High throughput screening	Transformation	
	Screening	
Deep sequencing	Construction amplicon sequencing pool	
	Deep sequencing	
	Analysis	

Bibliography

Aakre, C. D., Herrou, J., Phung, T. N., Perchuk, B. S., Crosson, S., & Laub, M. T. (2015). Evolving new protein-protein interaction specificity through promiscuous intermediates. *Cell, 163*(3), 594-606. doi:10.1016/j.cell.2015.09.055

Aguilera, A. (1994). Formamide sensitivity: a novel conditional phenotype in yeast. *Genetics, 136*(1), 87-91. doi:10.1093/genetics/136.1.87

Ahearn, J. M., Bartolomei, M. S., West, M. L., Cisek, L. J., & Corden, J. L. (1987). Cloning and sequence analysis of the mouse genomic locus encoding the largest subunit of RNA polymerase II. *Journal of Biological Chemistry, 262*(22), 10695-10705. doi:10.1016/s0021-9258(18)61020-8

Ahlquist, P. (2002). RNA-dependent RNA polymerases, viruses, and RNA silencing. *Science, 296*(5571), 1270-1273. doi:10.1126/science.1069132

Allison, L. A., Moyle, M., Shales, M., & Ingles, C. J. (1985). Extensive homology among the largest subunits of eukaryotic and prokaryotic RNA polymerases. *Cell, 42*(2), 599-610. doi:10.1016/0092-8674(85)90117-5

Araya, C. L., Fowler, D. M., Chen, W., Muniez, I., Kelly, J. W., & Fields, S. (2012). A fundamental protein property, thermodynamic stability, revealed solely from large-scale measurements of protein function. *Proc Natl Acad Sci U S A, 109*(42), 16858-16863. doi:10.1073/pnas.1209751109

Archambault, J., Lacroute, F., Ruet, A., & Friesen, J. D. (1992). Genetic interaction between transcription elongation factor TFIIS and RNA polymerase II. *Mol Cell Biol, 12*(9), 4142-4152. doi:10.1128/mcb.12.9.4142-4152.1992

Armache, K. J., Kettenberger, H., & Cramer, P. (2003). Architecture of initiation-competent 12-subunit RNA polymerase II. *Proc Natl Acad Sci U S A, 100*(12), 6964-6968. doi:10.1073/pnas.1030608100

Bakerlee, C. W., Nguyen Ba, A. N., Shulgina, Y., Rojas Echenique, J. I., & Desai, M. M. (2022). Idiosyncratic epistasis leads to global fitness-correlated trends. *Science, 376*(6593), 630-635. doi:10.1126/science.abm4774

Bank, C., Hietpas, R. T., Jensen, J. D., & Bolon, D. N. (2015). A systematic survey of an intragenic epistatic landscape. *Mol Biol Evol, 32*(1), 229-238. doi:10.1093/molbev/msu301

Bar-Nahum, G., Epshtein, V., Ruckenstein, A. E., Rafikov, R., Mustaev, A., & Nudler, E. (2005). A ratchet mechanism of transcription elongation and its control. *Cell, 120*(2), 183-193. doi:10.1016/j.cell.2004.11.045

Barnes, C. O., Calero, M., Malik, I., Graham, B. W., Spahr, H., Lin, G., . . . Calero, G. (2015). Crystal Structure of a Transcribing RNA Polymerase II Complex Reveals a Complete Transcription Bubble. *Mol Cell, 59*(2), 258-269. doi:10.1016/j.molcel.2015.06.034

Belogurov, G. A., & Artsimovitch, I. (2019). The Mechanisms of Substrate Selection, Catalysis, and Translocation by the Elongating RNA Polymerase. *J Mol Biol, 431*(20), 3975-4006. doi:10.1016/j.jmb.2019.05.042

Braberg, H., Jin, H., Moehle, E. A., Chan, Y. A., Wang, S., Shales, M., . . . Krogan, N. J. (2013). From structure to systems: high-resolution, quantitative genetic analysis of RNA polymerase II. *Cell, 154*(4), 775-788. doi:10.1016/j.cell.2013.07.033

Breen, M. S., Kemena, C., Vlasov, P. K., Notredame, C., & Kondrashov, F. A. (2012). Epistasis as the primary factor in molecular evolution. *Nature, 490*(7421), 535-538. doi:10.1038/nature11510

Bushnell, D. A., & Kornberg, R. D. (2003). Complete, 12-subunit RNA polymerase II at 4.1-A resolution: implications for the initiation of transcription. *Proc Natl Acad Sci U S A, 100*(12), 6969-6973. doi:10.1073/pnas.1130601100

Carvalho, A. T., Fernandes, P. A., & Ramos, M. J. (2011). The Catalytic Mechanism of RNA Polymerase II. *J Chem Theory Comput, 7*(4), 1177-1188. doi:10.1021/ct100579w

Castro, C., Smidansky, E., Maksimchuk, K. R., Arnold, J. J., Korneeva, V. S., Gotte, M., . . . Cameron, C. E. (2007). Two proton transfers in the transition state for nucleotidyl transfer catalyzed by RNA- and DNA-dependent RNA and DNA polymerases. *Proc Natl Acad Sci U S A, 104*(11), 4267-4272. doi:10.1073/pnas.0608952104

Castro, C., Smidansky, E. D., Arnold, J. J., Maksimchuk, K. R., Moustafa, I., Uchida, A., . . . Cameron, C. E. (2009). Nucleic acid polymerases use a general acid for nucleotidyl transfer. *Nat Struct Mol Biol, 16*(2), 212-218. doi:10.1038/nsmb.1540

Chen, Y., Kokic, G., Dienemann, C., Dybkov, O., Urlaub, H., & Cramer, P. (2023). Structure of the transcribing RNA polymerase II-Elongin complex. *Nat Struct Mol Biol, 30*(12), 1925-1935. doi:10.1038/s41594-023-01138-w

Cheung, A. C., & Cramer, P. (2011). Structural basis of RNA polymerase II backtracking, arrest and reactivation. *Nature, 471*(7337), 249-253. doi:10.1038/nature09785

Cheung, A. C., Sainsbury, S., & Cramer, P. (2011). Structural basis of initial RNA polymerase II transcription. *EMBO J, 30*(23), 4755-4763. doi:10.1038/emboj.2011.396

Cisneros, A. F., Gagnon-Arsenault, I., Dube, A. K., Despres, P. C., Kumar, P., Lafontaine, K., . . . Landry, C. R. (2023). Epistasis between promoter activity and coding mutations shapes gene evolvability. *Sci Adv, 9*(5), eadd9109. doi:10.1126/sciadv.add9109

Close, D., Johnson, S. J., Sdano, M. A., McDonald, S. M., Robinson, H., Formosa, T., & Hill, C. P. (2011). Crystal structures of the S. cerevisiae Spt6 core and C-terminal tandem SH2 domain. *J Mol Biol, 408*(4), 697-713. doi:10.1016/j.jmb.2011.03.002

Costanzo, M., VanderSluis, B., Koch, E. N., Baryshnikova, A., Pons, C., Tan, G., . . . Boone, C. (2016). A global genetic interaction network maps a wiring diagram of cellular function. *Science, 353*(6306). doi:10.1126/science.aaf1420

Cramer, P. (2002). Multisubunit RNA polymerases. *Curr Opin Struct Biol, 12*(1), 89-97. doi:10.1016/s0959-440x(02)00294-4

Cramer, P. (2019a). Eukaryotic Transcription Turns 50. *Cell, 179*(4), 808-812. doi:10.1016/j.cell.2019.09.018

Cramer, P. (2019b). Organization and regulation of gene transcription. *Nature, 573*(7772), 45-54. doi:10.1038/s41586-019-1517-4

Cramer, P., Bushnell, D. A., & Kornberg, R. D. (2001). Structural basis of transcription: RNA polymerase II at 2.8 angstrom resolution. *Science, 292*(5523), 1863-1876. doi:10.1126/science.1059493

Cui, P., Jin, H., Vutukuru, M. R., & Kaplan, C. D. (2016). Relationships Between RNA Polymerase II Activity and Spt Elongation Factors to Spt- Phenotype and Growth in Saccharomyces cerevisiae. *G3 (Bethesda), 6*(8), 2489-2504. doi:10.1534/g3.116.030346

Da, L. T., Pardo-Avila, F., Xu, L., Silva, D. A., Zhang, L., Gao, X., . . . Huang, X. (2016). Bridge helix bending promotes RNA polymerase II backtracking through a critical and conserved threonine residue. *Nat Commun, 7*, 11244. doi:10.1038/ncomms11244

Da, L. T., Wang, D., & Huang, X. (2012). 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*(4), 2399-2406. doi:10.1021/ja210656k

Dangkulwanich, M., Ishibashi, T., Liu, S., Kireeva, M. L., Lubkowska, L., Kashlev, M., & Bustamante, C. J. (2013). Complete dissection of transcription elongation reveals slow translocation of RNA polymerase II in a linear ratchet mechanism. *Elife, 2*, e00971. doi:10.7554/eLife.00971

Ding, D., Green, A. G., Wang, B., Lite, T. V., Weinstein, E. N., Marks, D. S., & Laub, M. T. (2022). Co-evolution of interacting proteins through non-contacting and non-specific mutations. *Nat Ecol Evol, 6*(5), 590-603. doi:10.1038/s41559-022-01688-0

Diss, G., & Lehner, B. (2018). The genetic landscape of a physical interaction. *Elife, 7*. doi:10.7554/eLife.32472

Doherty, G. P., Fogg, M. J., Wilkinson, A. J., & Lewis, P. J. (2010). Small subunits of RNA polymerase: localization, levels and implications for core enzyme composition. *Microbiology (Reading), 156*(Pt 12), 3532-3543. doi:10.1099/mic.0.041566-0

Domingo, J., Baeza-Centurion, P., & Lehner, B. (2019). The Causes and Consequences of Genetic Interactions (Epistasis). *Annu Rev Genomics Hum Genet, 20*, 433-460. doi:10.1146/annurev-genom-083118-014857

Doud, M. B., Ashenberg, O., & Bloom, J. D. (2015). Site-Specific Amino Acid Preferences Are Mostly Conserved in Two Closely Related Protein Homologs. *Mol Biol Evol, 32*(11), 2944-2960. doi:10.1093/molbev/msv167

Duan, B., Qiu, C., Sze, S. H., & Kaplan, C. (2023). Widespread epistasis shapes RNA Polymerase II active site function and evolution. *bioRxiv*. doi:10.1101/2023.02.27.530048

Engel, C., Sainsbury, S., Cheung, A. C., Kostrewa, D., & Cramer, P. (2013). RNA polymerase I structure and transcription regulation. *Nature, 502*(7473), 650-655. doi:10.1038/nature12712

Esposito, D., Weile, J., Shendure, J., Starita, L. M., Papenfuss, A. T., Roth, F. P., . . . Rubin, A. F. (2019). MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome Biol, 20*(1), 223. doi:10.1186/s13059-019-1845-6

Faure, A. J., Domingo, J., Schmiedel, J. M., Hidalgo-Carcedo, C., Diss, G., & Lehner, B. (2022). Mapping the energetic and allosteric landscapes of protein binding domains. *Nature, 604*(7904), 175-183. doi:10.1038/s41586-022-04586-4

Faure, A. J., Schmiedel, J. M., Baeza-Centurion, P., & Lehner, B. (2020). DiMSum: an error model and pipeline for analyzing deep mutational scanning data and diagnosing common experimental pathologies. *Genome Biol, 21*(1), 207. doi:10.1186/s13059-020-02091-3

Fernandez-Tornero, C., Moreno-Morcillo, M., Rashid, U. J., Taylor, N. M., Ruiz, F. M., Gruene, T., . . . Muller, C. W. (2013). Crystal structure of the 14-subunit RNA polymerase I. *Nature, 502*(7473), 644-649. doi:10.1038/nature12636

Fisher, R. A. (1919). XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Transactions of the Royal Society of Edinburgh, 52*(2), 399-433. doi:10.1017/S0080456800012163

Fong, N., Kim, H., Zhou, Y., Ji, X., Qiu, J., Saldi, T., . . . Bentley, D. L. (2014). Pre-mRNA splicing is facilitated by an optimal RNA polymerase II elongation rate. *Genes Dev, 28*(23), 2663-2676. doi:10.1101/gad.252106.114

Fouqueau, T., Zeller, M. E., Cheung, A. C., Cramer, P., & Thomm, M. (2013). The RNA polymerase trigger loop functions in all three phases of the transcription cycle. *Nucleic Acids Res, 41*(14), 7048-7059. doi:10.1093/nar/gkt433

Fowler, D. M., Araya, C. L., Fleishman, S. J., Kellogg, E. H., Stephany, J. J., Baker, D., & Fields, S. (2010). High-resolution mapping of protein sequence-function relationships. *Nat Methods, 7*(9), 741-746. doi:10.1038/nmeth.1492

Fowler, D. M., Araya, C. L., Gerard, W., & Fields, S. (2011). Enrich: software for analysis of protein function by enrichment and depletion of variants. *Bioinformatics, 27*(24), 3430-3431. doi:10.1093/bioinformatics/btr577

Fowler, D. M., & Fields, S. (2014). Deep mutational scanning: a new style of protein science. *Nat Methods, 11*(8), 801-807. doi:10.1038/nmeth.3027

Gentry, D. R., & Burgess, R. R. (1990). Overproduction and purification of the omega subunit of Escherichia coli RNA polymerase. *Protein Expr Purif, 1*(1), 81-86. doi:10.1016/1046-5928(90)90050-9

Ghindilis, A. L., Smith, M. W., Schwarzkopf, K. R., Roth, K. M., Peyvan, K., Munro, S. B., . . . McShea, A. (2007). CombiMatrix oligonucleotide arrays: genotyping and gene expression assays employing electrochemical detection. *Biosens Bioelectron, 22*(9-10), 1853-1860. doi:10.1016/j.bios.2006.06.024

Gietz, R. D., & Schiestl, R. H. (2007). High-efficiency yeast transformation using the LiAc/SS carrier DNA/PEG method. *Nat Protoc, 2*(1), 31-34. doi:10.1038/nprot.2007.13

Gnatt, A. L., Cramer, P., Fu, J., Bushnell, D. A., & Kornberg, R. D. (2001). Structural basis of transcription: an RNA polymerase II elongation complex at 3.3 A resolution. *Science, 292*(5523), 1876-1882. doi:10.1126/science.1059495

Goodwin, E. B., & Ellis, R. E. (2002). Turning clustering loops: sex determination in Caenorhabditis elegans. *Curr Biol, 12*(3), R111-120. doi:10.1016/s0960-9822(02)00675-9

Gout, J. F., Li, W., Fritsch, C., Li, A., Haroon, S., Singh, L., . . . Vermulst, M. (2017). The landscape of transcription errors in eukaryotic cells. *Sci Adv, 3*(10), e1701484. doi:10.1126/sciadv.1701484

Gout, J. F., Thomas, W. K., Smith, Z., Okamoto, K., & Lynch, M. (2013). Large-scale detection of in vivo transcription errors. *Proc Natl Acad Sci U S A, 110*(46), 18584-18589. doi:10.1073/pnas.1309843110

Greger, I. H., & Proudfoot, N. J. (1998). Poly(A) signals control both transcriptional termination and initiation between the tandem GAL10 and GAL7 genes of Saccharomyces cerevisiae. *EMBO J, 17*(16), 4771-4779. doi:10.1093/emboj/17.16.4771

Gregory, M. T., Gao, Y., Cui, Q., & Yang, W. (2021). Multiple deprotonation paths of the nucleophile 3'-OH in the DNA synthesis reaction. *Proc Natl Acad Sci U S A, 118*(23). doi:10.1073/pnas.2103990118

Haddox, H. K., Dingens, A. S., Hilton, S. K., Overbaugh, J., & Bloom, J. D. (2018). Mapping mutational effects along the evolutionary landscape of HIV envelope. *Elife, 7*. doi:10.7554/eLife.34420

Halabi, N., Rivoire, O., Leibler, S., & Ranganathan, R. (2009). Protein sectors: evolutionary units of three-dimensional structure. *Cell, 138*(4), 774-786. doi:10.1016/j.cell.2009.07.038

Harms, M. J., & Thornton, J. W. (2010). Analyzing protein structure and function using ancestral gene reconstruction. *Curr Opin Struct Biol, 20*(3), 360-366. doi:10.1016/j.sbi.2010.03.005

Hartzog, G. A., Wada, T., Handa, H., & Winston, F. (1998). Evidence that Spt4, Spt5, and Spt6 control transcription elongation by RNA polymerase II in Saccharomyces cerevisiae. *Genes Dev, 12*(3), 357-369. doi:10.1101/gad.12.3.357

Hein, P. P., Kolb, K. E., Windgassen, T., Bellecourt, M. J., Darst, S. A., Mooney, R. A., & Landick, R. (2014). RNA polymerase pausing and nascent-RNA structure formation are linked through clamp-domain movement. *Nat Struct Mol Biol, 21*(9), 794-802. doi:10.1038/nsmb.2867

Hietpas, R., Roscoe, B., Jiang, L., & Bolon, D. N. (2012). Fitness analyses of all possible point mutations for regions of genes in yeast. *Nat Protoc, 7*(7), 1382-1396. doi:10.1038/nprot.2012.069

Hill, W. G., Goddard, M. E., & Visscher, P. M. (2008). Data and theory point to mainly additive genetic variance for complex traits. *PLoS Genet, 4*(2), e1000008. doi:10.1371/journal.pgen.1000008

Hirata, A., Klein, B. J., & Murakami, K. S. (2008). The X-ray crystal structure of RNA polymerase from Archaea. *Nature, 451*(7180), 851-854. doi:10.1038/nature06530

Hirtreiter, A., Grohmann, D., & Werner, F. (2010). Molecular mechanisms of RNA polymerase--the F/E (RPB4/7) complex is required for high processivity in vitro. *Nucleic Acids Res, 38*(2), 585-596. doi:10.1093/nar/gkp928

Hoffmann, N. A., Jakobi, A. J., Moreno-Morcillo, M., Glatt, S., Kosinski, J., Hagen, W. J., . . . Muller, C. W. (2015). Molecular structures of unbound and transcribing RNA polymerase III. *Nature, 528*(7581), 231-236. doi:10.1038/nature16143

Huang, X., Wang, D., Weiss, D. R., Bushnell, D. A., Kornberg, R. D., & Levitt, M. (2010). RNA polymerase II trigger loop residues stabilize and position the incoming nucleotide triphosphate in transcription. *Proc Natl Acad Sci U S A, 107*(36), 15745-15750. doi:10.1073/pnas.1009898107

Huang, Y., Kendall, T., Forsythe, E. S., Dorantes-Acosta, A., Li, S., Caballero-Perez, J., . . . Mosher, R. A. (2015). Ancient Origin and Recent Innovations of RNA Polymerase IV and V. *Mol Biol Evol, 32*(7), 1788-1799. doi:10.1093/molbev/msv060

Hyle, J. W., Shaw, R. J., & Reines, D. (2003). Functional distinctions between IMP dehydrogenase genes in providing mycophenolate resistance and guanine prototrophy to yeast. *J Biol Chem, 278*(31), 28470-28478. doi:10.1074/jbc.M303736200

Imashimizu, M., Oshima, T., Lubkowska, L., & Kashlev, M. (2013). Direct assessment of transcription fidelity by high-resolution RNA sequencing. *Nucleic Acids Res, 41*(19), 9090-9104. doi:10.1093/nar/gkt698

Jenks, M. H., O'Rourke, T. W., & Reines, D. (2008). Properties of an intergenic terminator and start site switch that regulate IMD2 transcription in yeast. *Mol Cell Biol, 28*(12), 3883-3893. doi:10.1128/MCB.00380-08

Johnson, M. S., Reddy, G., & Desai, M. M. (2023). Epistasis and evolution: recent advances and an outlook for prediction. *BMC Biol, 21*(1), 120. doi:10.1186/s12915-023-01585-3

Kaplan, C. D. (2010). The architecture of RNA polymerase fidelity. *BMC Biol, 8*, 85. doi:10.1186/1741-7007-8-85

Kaplan, C. D. (2013). Basic mechanisms of RNA polymerase II activity and alteration of gene expression in Saccharomyces cerevisiae. *Biochimica Et Biophysica Acta-Gene Regulatory Mechanisms, 1829*(1), 39-54. doi:10.1016/j.bbagrm.2012.09.007

Kaplan, C. D., Holland, M. J., & Winston, F. (2005). Interaction between transcription elongation factors and mRNA 3'-end formation at the Saccharomyces cerevisiae GAL10-GAL7 locus. *J Biol Chem, 280*(2), 913-922. doi:10.1074/jbc.M411108200

Kaplan, C. D., Jin, H., Zhang, I. L., & Belyanin, A. (2012). Dissection of Pol II trigger loop function and Pol II activity-dependent control of start site selection in vivo. *PLoS Genet, 8*(4), e1002627. doi:10.1371/journal.pgen.1002627

Kaplan, C. D., & Kornberg, R. D. (2008). A bridge to transcription by RNA polymerase. *J Biol, 7*(10), 39. doi:10.1186/jbiol99

Kaplan, C. D., Larsson, K. M., & Kornberg, R. D. (2008). The RNA polymerase II trigger loop functions in substrate selection and is directly targeted by alpha-amanitin. *Mol Cell, 30*(5), 547-556. doi:10.1016/j.molcel.2008.04.023

Karageorgi, M., Groen, S. C., Sumbul, F., Pelaez, J. N., Verster, K. I., Aguilar, J. M., . . . Whiteman, N. K. (2019). Genome editing retraces the evolution of toxin resistance in the monarch butterfly. *Nature, 574*(7778), 409-412. doi:10.1038/s41586-019-1610-8

Kaster, B. C., Knippa, K. C., Kaplan, C. D., & Peterson, D. O. (2016). RNA Polymerase II Trigger Loop Mobility: INDIRECT EFFECTS OF Rpb9. *J Biol Chem, 291*(28), 14883-14895. doi:10.1074/jbc.M116.714394

Kettenberger, H., Armache, K. J., & Cramer, P. (2004). Complete RNA polymerase II elongation complex structure and its interactions with NTP and TFIIS. *Mol Cell, 16*(6), 955-965. doi:10.1016/j.molcel.2004.11.040

Khatter, H., Vorlander, M. K., & Muller, C. W. (2017). RNA polymerase I and III: similar yet unique. *Curr Opin Struct Biol, 47*, 88-94. doi:10.1016/j.sbi.2017.05.008

Kinney, J. B., & McCandlish, D. M. (2019). Massively Parallel Assays and Quantitative Sequence-Function Relationships. *Annu Rev Genomics Hum Genet, 20*, 99-127. doi:10.1146/annurev-genom-083118-014845

Kireeva, M. L., Nedialkov, Y. A., Cremona, G. H., Purtov, Y. A., Lubkowska, L., Malagon, F., . . . Kashlev, M. (2008). Transient reversal of RNA polymerase II active site closing controls fidelity of transcription elongation. *Mol Cell, 30*(5), 557-566. doi:10.1016/j.molcel.2008.04.017

Kireeva, M. L., Opron, K., Seibold, S. A., Domecq, C., Cukier, R. I., Coulombe, B., . . . Burton, Z. F. (2012). Molecular dynamics and mutational analysis of the catalytic and translocation cycle of RNA polymerase. *BMC Biophys, 5*, 11. doi:10.1186/2046-1682-5-11

Kondrashov, A. S., Sunyaev, S., & Kondrashov, F. A. (2002). Dobzhansky-Muller incompatibilities in protein evolution. *Proc Natl Acad Sci U S A, 99*(23), 14878-14883. doi:10.1073/pnas.232565499

Korkhin, Y., Unligil, U. M., Littlefield, O., Nelson, P. J., Stuart, D. I., Sigler, P. B., . . . Abrescia, N. G. (2009). Evolution of complex RNA polymerases: the complete archaeal RNA polymerase structure. *PLoS Biol, 7*(5), e1000102. doi:10.1371/journal.pbio.1000102

Kostrewa, D., Zeller, M. E., Armache, K. J., Seizl, M., Leike, K., Thomm, M., & Cramer, P. (2009). RNA polymerase II-TFIIB structure and mechanism of transcription initiation. *Nature, 462*(7271), 323-330. doi:10.1038/nature08548

Kosuri, S., & Church, G. M. (2014). Large-scale de novo DNA synthesis: technologies and applications. *Nat Methods, 11*(5), 499-507. doi:10.1038/nmeth.2918

Koyama, H., Ueda, T., Ito, T., & Sekimizu, K. (2010). Novel RNA polymerase II mutation suppresses transcriptional fidelity and oxidative stress sensitivity in rpb9Delta yeast. *Genes Cells, 15*(2), 151-159. doi:10.1111/j.1365-2443.2009.01372.x

Kramm, K., Endesfelder, U., & Grohmann, D. (2019). A Single-Molecule View of Archaeal Transcription. *J Mol Biol, 431*(20), 4116-4131. doi:10.1016/j.jmb.2019.06.009

Kuehner, J. N., & Brow, D. A. (2008). Regulation of a eukaryotic gene by GTP-dependent start site selection and transcription attenuation. *Mol Cell, 31*(2), 201-211. doi:10.1016/j.molcel.2008.05.018

Larson, M. H., Zhou, J., Kaplan, C. D., Palangat, M., Kornberg, R. D., Landick, R., & Block, S. M. (2012). Trigger loop dynamics mediate the balance between the transcriptional fidelity and speed of RNA polymerase II. *Proc Natl Acad Sci U S A, 109*(17), 6555-6560. doi:10.1073/pnas.1200939109

Lassila, J. K., Zalatan, J. G., & Herschlag, D. (2011). Biological phosphoryl-transfer reactions: understanding mechanism and catalysis. *Annu Rev Biochem, 80*, 669-702. doi:10.1146/annurev-biochem-060409-092741

Leng, X. Y., Iyanov, M., Kindgren, P., Malik, I., Thieffry, A., Brodersen, P., . . . Marquardt, S. (2020). Organismal benefits of transcription speed control at gene boundaries. *Embo Reports, 21*(4). doi:ARTN e49315
10.15252/embr.201949315

Lennon, C. W., Ross, W., Martin-Tumasz, S., Toulokhonov, I., Vrentas, C. E., Rutherford, S. T., . . . Gourse, R. L. (2012). Direct interactions between the coiled-coil tip of DksA and the trigger loop of RNA polymerase mediate transcriptional regulation. *Genes Dev, 26*(23), 2634-2646. doi:doi: 10.1101/gad.204693.112

LeProust, E. M., Peck, B. J., Spirin, K., McCuen, H. B., Moore, B., Namsaraev, E., & Caruthers, M. H. (2010). Synthesis of high-quality libraries of long (150mer) oligonucleotides by a novel depurination controlled process. *Nucleic Acids Res, 38*(8), 2522-2540. doi:10.1093/nar/gkq163

Lin, G., Barnes, C. O., Weiss, S., Dutagaci, B., Qiu, C., Feig, M., . . . Calero, G. (2023). Structural basis of transcription: RNA Polymerase II substrate binding and metal coordination at 3.0 A using a free-electron laser. *bioRxiv*. doi:10.1101/2023.09.22.559052

Lin, X., Liu, Y., Liu, S., Zhu, X., Wu, L., Zhu, Y., . . . Qi, L. S. (2022). Nested epistasis enhancer networks for robust genome regulation. *Science, 377*(6610), 1077-1085. doi:10.1126/science.abk3512

Lite, T. V., Grant, R. A., Nocedal, I., Littlehale, M. L., Guo, M. S., & Laub, M. T. (2020). Uncovering the basis of protein-protein interaction specificity with a combinatorially complete library. *Elife, 9*. doi:10.7554/eLife.60924

Liu, B., Zuo, Y., & Steitz, T. A. (2016). Structures of E. coli sigmaS-transcription initiation complexes provide new insights into polymerase mechanism. *Proc Natl Acad Sci U S A, 113*(15), 4051-4056. doi:10.1073/pnas.1520555113

Liu, X., Bushnell, D. A., & Kornberg, R. D. (2013). RNA polymerase II transcription: structure and mechanism. *Biochim Biophys Acta, 1829*(1), 2-8. doi:10.1016/j.bbagrm.2012.09.003

Lubock, N. B., Zhang, D., Sidore, A. M., Church, G. M., & Kosuri, S. (2017). A systematic comparison of error correction enzymes by next-generation sequencing. *Nucleic Acids Res, 45*(15), 9206-9217. doi:10.1093/nar/gkx691

Lunzer, M., Golding, G. B., & Dean, A. M. (2010). Pervasive cryptic epistasis in molecular evolution. *PLoS Genet, 6*(10), e1001162. doi:10.1371/journal.pgen.1001162

Malagon, F., Kireeva, M. L., Shafer, B. K., Lubkowska, L., Kashlev, M., & Strathern, J. N. (2006). Mutations in the Saccharomyces cerevisiae RPB1 gene conferring hypersensitivity to 6-azauracil. *Genetics, 172*(4), 2201-2209. doi:10.1534/genetics.105.052415

Malik, I., Qiu, C., Snavely, T., & Kaplan, C. D. (2017). Wide-ranging and unexpected consequences of altered Pol II catalytic activity in vivo. *Nucleic Acids Research, 45*(8), 4431-4451. doi:10.1093/nar/gkx037

Malinen, A. M., Turtola, M., Parthiban, M., Vainonen, L., Johnson, M. S., & Belogurov, G. A. (2012). Active site opening and closure control translocation of multisubunit RNA polymerase. *Nucleic Acids Res, 40*(15), 7442-7451. doi:10.1093/nar/gks383

Mani, R., St Onge, R. P., Hartman, J. L. t., Giaever, G., & Roth, F. P. (2008). Defining genetic interaction. *Proc Natl Acad Sci U S A, 105*(9), 3461-3466. doi:10.1073/pnas.0712255105

Mason, P. B., & Struhl, K. (2005). Distinction and relationship between elongation rate and processivity of RNA polymerase II in vivo. *Mol Cell, 17*(6), 831-840. doi:10.1016/j.molcel.2005.02.017

Matthew H. Larsona, Jing Zhoub,1, Craig D. Kaplanc,1, Murali Palangatd,2, Roger D. Kornberge, Robert Landickd, and Steven M. Blocka,b,f,3. (2012). Trigger loop dynamics mediate the balance between the transcriptional fidelity and speed of RNA polymerase II. Retrieved from www.pnas.org/cgi/doi/10.1073/pnas.1200939109

Matuszewski, S., Hildebrandt, M. E., Ghenu, A. H., Jensen, J. D., & Bank, C. (2016). A Statistical Guide to the Design of Deep Mutational Scanning Experiments. *Genetics, 204*(1), 77-87. doi:10.1534/genetics.116.190462

Mayer, A., Schreieck, A., Lidschreiber, M., Leike, K., Martin, D. E., & Cramer, P. (2012). The spt5 C-terminal region recruits yeast 3' RNA cleavage factor I. *Mol Cell Biol, 32*(7), 1321-1331. doi:10.1128/MCB.06310-11

Mazumder, A., Lin, M., Kapanidis, A. N., & Ebright, R. H. (2020). Closing and opening of the RNA polymerase trigger loop. *Proc Natl Acad Sci U S A, 117*(27), 15642-15649. doi:10.1073/pnas.1920427117

Mejia, Y. X., Nudler, E., & Bustamante, C. (2015). Trigger loop folding determines transcription rate of Escherichia coli's RNA polymerase. *Proc Natl Acad Sci U S A, 112*(3), 743-748. doi:10.1073/pnas.1421067112

Melamed, D., Young, D. L., Gamble, C. E., Miller, C. R., & Fields, S. (2013). Deep mutational scanning of an RRM domain of the Saccharomyces cerevisiae poly(A)-binding protein. *RNA, 19*(11), 1537-1551. doi:10.1261/rna.040709.113

Metzger, B. P. H., Park, Y., Starr, T. N., & Thornton, J. W. (2023). Epistasis facilitates functional evolution in an ancient transcription factor. *Elife*. doi:https://doi.org/10.7554/eLife.88737.1

Mishanina, T. V., Palo, M. Z., Nayak, D., Mooney, R. A., & Landick, R. (2017). Trigger loop of RNA polymerase is a positional, not acid-base, catalyst for both transcription and proofreading. *Proc Natl Acad Sci U S A, 114*(26), E5103-E5112. doi:10.1073/pnas.1702383114

Mosaei, H., & Zenkin, N. (2021). Two distinct pathways of RNA polymerase backtracking determine the requirement for the Trigger Loop during RNA hydrolysis. *Nucleic Acids Res, 49*(15), 8777-8784. doi:10.1093/nar/gkab675

Natarajan, C., Inoguchi, N., Weber, R. E., Fago, A., Moriyama, H., & Storz, J. F. (2013). Epistasis among adaptive mutations in deer mouse hemoglobin. *Science, 340*(6138), 1324-1327. doi:10.1126/science.1236862

Nayak, D., Voss, M., Windgassen, T., Mooney, R. A., & Landick, R. (2013). Cys-pair reporters detect a constrained trigger loop in a paused RNA polymerase. *Mol Cell, 50*(6), 882-893. doi:10.1016/j.molcel.2013.05.015

Notredame, C., Higgins, D. G., & Heringa, J. (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol, 302*(1), 205-217. doi:10.1006/jmbi.2000.4042

Nouraini, S., Archambault, J., & Friesen, J. D. (1996). Rpo26p, a subunit common to yeast RNA polymerases, is essential for the assembly of RNA polymerases I and II and for the stability of the largest subunits of these enzymes. *Mol Cell Biol, 16*(11), 5985-5996. doi:10.1128/MCB.16.11.5985

Olson, C. A., Wu, N. C., & Sun, R. (2014). A comprehensive biophysical description of pairwise epistasis throughout an entire protein domain. *Curr Biol, 24*(22), 2643-2651. doi:10.1016/j.cub.2014.09.072

Onodera, Y., Haag, J. R., Ream, T., Costa Nunes, P., Pontes, O., & Pikaard, C. S. (2005). Plant nuclear RNA polymerase IV mediates siRNA and DNA methylation-dependent heterochromatin formation. *Cell, 120*(5), 613-622. doi:10.1016/j.cell.2005.02.007

Ortlund, E. A., Bridgham, J. T., Redinbo, M. R., & Thornton, J. W. (2007). Crystal structure of an ancient protein: evolution by conformational epistasis. *Science, 317*(5844), 1544-1548. doi:10.1126/science.1142819

Palo, M. Z., Zhu, J., Mishanina, T. V., & Landick, R. (2021). Conserved Trigger Loop Histidine of RNA Polymerase II Functions as a Positional Catalyst Primarily through Steric Effects. *Biochemistry, 60*(44), 3323-3336. doi:10.1021/acs.biochem.1c00528

Park, Y., Metzger, B. P. H., & Thornton, J. W. (2022). Epistatic drift causes gradual decay of predictability in protein evolution. *Science, 376*(6595), 823-830. doi:10.1126/science.abn6895

Pero, J., Nelson, J., & Fox, T. D. (1975). Highly asymmetric transcription by RNA polymerase containing phage-SP01-induced polypeptides and a new host protein. *Proc Natl Acad Sci U S A, 72*(4), 1589-1593. doi:10.1073/pnas.72.4.1589

Phillips, P. C. (1998). The language of gene interaction. *Genetics, 149*(3), 1167-1171. doi:10.1093/genetics/149.3.1167

Phillips, P. C. (2008). Epistasis--the essential role of gene interactions in the structure and evolution of genetic systems. *Nat Rev Genet, 9*(11), 855-867. doi:10.1038/nrg2452

Pinney, M. M., Mokhtari, D. A., Akiva, E., Yabukarski, F., Sanchez, D. M., Liang, R., . . . Herschlag, D. (2021). Parallel molecular mechanisms for enzyme temperature adaptation. *Science, 371*(6533). doi:10.1126/science.aay2784

Poelwijk, F. J., Kiviet, D. J., Weinreich, D. M., & Tans, S. J. (2007). Empirical fitness landscapes reveal accessible evolutionary paths. *Nature, 445*(7126), 383-386. doi:10.1038/nature05451

Pokusaeva, V. O., Usmanova, D. R., Putintseva, E. V., Espinar, L., Sarkisyan, K. S., Mishin, A. S., . . . Kondrashov, F. A. (2019). An experimental assay of the interactions of amino acids from orthologous sequences shaping a complex fitness landscape. *PLoS Genet, 15*(4), e1008079. doi:10.1371/journal.pgen.1008079

Prather, D., Krogan, N. J., Emili, A., Greenblatt, J. F., & Winston, F. (2005). Identification and characterization of Elf1, a conserved transcription elongation factor in Saccharomyces cerevisiae. *Mol Cell Biol, 25*(22), 10122-10135. doi:10.1128/MCB.25.22.10122-10135.2005

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2--approximately maximum-likelihood trees for large alignments. *Plos One, 5*(3), e9490. doi:10.1371/journal.pone.0009490

Qiu, C., Erinne, O. C., Dave, J. M., Cui, P., Jin, H., Muthukrishnan, N., . . . Kaplan, C. D. (2016). High-Resolution Phenotypic Landscape of the RNA Polymerase II Trigger Loop. *PLoS Genet, 12*(11), e1006321. doi:10.1371/journal.pgen.1006321

Qiu, C., & Kaplan, C. D. (2019). Functional assays for transcription mechanisms in high-throughput. *Methods, 159-160*, 115-123. doi:10.1016/j.ymeth.2019.02.017

Ream, T. S., Haag, J. R., Wierzbicki, A. T., Nicora, C. D., Norbeck, A. D., Zhu, J. K., . . . Pikaard, C. S. (2009). Subunit compositions of the RNA-silencing enzymes Pol IV and Pol V reveal their origins as specialized forms of RNA polymerase II. *Mol Cell, 33*(2), 192-203. doi:10.1016/j.molcel.2008.12.015

Reddy, G., & Desai, M. M. (2021). Global epistasis emerges from a generic model of a complex trait. *Elife, 10*. doi:10.7554/eLife.64740

Reid-Bayliss, K. S., & Loeb, L. A. (2017). Accurate RNA consensus sequencing for high-fidelity detection of transcriptional mutagenesis-induced epimutations. *Proc Natl Acad Sci U S A, 114*(35), 9415-9420. doi:10.1073/pnas.1709166114

Riles, L., Shaw, R. J., Johnston, M., & Reines, D. (2004). Large-scale screening of yeast mutants for sensitivity to the IMP dehydrogenase inhibitor 6-azauracil. *Yeast, 21*(3), 241-248. doi:10.1002/yea.1068

Rivoire, O., Reynolds, K. A., & Ranganathan, R. (2016). Evolution-Based Functional Decomposition of Proteins. *PLoS Comput Biol, 12*(6), e1004817. doi:10.1371/journal.pcbi.1004817

Roeder, R. G., & Rutter, W. J. (1969). Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms. *Nature, 224*(5216), 234-237. doi:10.1038/224234a0

Rollins, N. J., Brock, K. P., Poelwijk, F. J., Stiffler, M. A., Gauthier, N. P., Sander, C., & Marks, D. S. (2019). Inferring protein 3D structure from deep mutation scans. *Nat Genet, 51*(7), 1170-1176. doi:10.1038/s41588-019-0432-9

Rubin, A. F., Gelman, H., Lucas, N., Bajjalieh, S. M., Papenfuss, A. T., Speed, T. P., & Fowler, D. M. (2017). A statistical framework for analyzing deep mutational scanning data. *Genome Biol, 18*(1), 150. doi:10.1186/s13059-017-1272-5

Russ, W. P., Lowery, D. M., Mishra, P., Yaffe, M. B., & Ranganathan, R. (2005). Natural-like function in artificial WW domains. *Nature, 437*(7058), 579-583. doi:10.1038/nature03990

Sadian, Y., Baudin, F., Tafur, L., Murciano, B., Wetzel, R., Weis, F., & Muller, C. W. (2019). Molecular insight into RNA polymerase I promoter recognition and promoter melting. *Nat Commun, 10*(1), 5543. doi:10.1038/s41467-019-13510-w

Salinas, V. H., & Ranganathan, R. (2018). Coevolution-based inference of amino acid interactions underlying protein function. *Elife, 7*. doi:10.7554/eLife.34300

Sauguet, L. (2019). The Extended "Two-Barrel" Polymerases Superfamily: Structure, Function and Evolution. *J Mol Biol*. doi:10.1016/j.jmb.2019.05.017

Scherrer, K. (2018). Primary transcripts: From the discovery of RNA processing to current concepts of gene expression - Review. *Exp Cell Res, 373*(1-2), 1-33. doi:10.1016/j.yexcr.2018.09.011

Schier, A. C., & Taatjes, D. J. (2020). Structure and mechanism of the RNA polymerase II transcription machinery. *Genes Dev, 34*(7-8), 465-488. doi:10.1101/gad.335679.119

Schmiedel, J. M., & Lehner, B. (2019). Determining protein structures using deep mutagenesis. *Nat Genet, 51*(7), 1177-1186. doi:10.1038/s41588-019-0431-x

Scull, C. E., Ingram, Z. M., Lucius, A. L., & Schneider, D. A. (2019). A Novel Assay for RNA Polymerase I Transcription Elongation Sheds Light on the Evolutionary Divergence of Eukaryotic RNA Polymerases. *Biochemistry, 58*(16), 2116-2124. doi:10.1021/acs.biochem.8b01256

Seibold, S. A., Singh, B. N., Zhang, C., Kireeva, M., Domecq, C., Bouchard, A., . . . Burton, Z. F. (2010). Conformational coupling, bridge helix dynamics and active site dehydration in catalysis by RNA polymerase. *Biochim Biophys Acta, 1799*(8), 575-587. doi:10.1016/j.bbagrm.2010.05.002

Sekine, S., Murayama, Y., Svetlov, V., Nudler, E., & Yokoyama, S. (2015). The ratcheted and ratchetable structural states of RNA polymerase underlie multiple transcriptional functions. *Mol Cell, 57*(3), 408-421. doi:10.1016/j.molcel.2014.12.014

Sergey Ioffe, C. S. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv:1502.03167v3*. doi:https://doi.org/10.48550/arXiv.1502.03167

Shin, H., & Cho, B. K. (2015). Rational Protein Engineering Guided by Deep Mutational Scanning. *Int J Mol Sci, 16*(9), 23094-23110. doi:10.3390/ijms160923094

Silva, D. A., Weiss, D. R., Pardo Avila, F., Da, L. T., Levitt, M., Wang, D., & Huang, X. (2014). Millisecond dynamics of RNA polymerase II translocation at atomic resolution. *Proc Natl Acad Sci U S A, 111*(21), 7665-7670. doi:10.1073/pnas.1315751111

Simchen, G., Winston, F., Styles, C. A., & Fink, G. R. (1984). Ty-mediated gene expression of the LYS2 and HIS4 genes of Saccharomyces cerevisiae is controlled by the same SPT genes. *Proc Natl Acad Sci U S A, 81*(8), 2431-2434. doi:10.1073/pnas.81.8.2431

Sing-Hoi Sze, C. D. K. (2018). *Codon-Based Sequence Alignment for Mutation Analysis by High-Throughput Sequencing*. Paper presented at the 2018 IEEE 8th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS). Conference retrieved from

Socolich, M., Lockless, S. W., Russ, W. P., Lee, H., Gardner, K. H., & Ranganathan, R. (2005). Evolutionary information for specifying a protein fold. *Nature, 437*(7058), 512-518. doi:10.1038/nature03991

Squire, J. E. (1909). Discussion on the Influence of Heredity on Disease, with special Reference to Tuberculosis, Cancer, and Diseases of the Nervous System: Introductory Address. *Proc R Soc Med, 2*(Gen Rep), 60-63. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/19973586

Starita, L. M., Ahituv, N., Dunham, M. J., Kitzman, J. O., Roth, F. P., Seelig, G., . . . Fowler, D. M. (2017). Variant Interpretation: Functional Assays to the Rescue. *Am J Hum Genet, 101*(3), 315-325. doi:10.1016/j.ajhg.2017.07.014

Starita, L. M., & Fields, S. (2015). Deep Mutational Scanning: Library Construction, Functional Selection, and High-Throughput Sequencing. *Cold Spring Harb Protoc, 2015*(8), 777-780. doi:10.1101/pdb.prot085225

Starr, T. N., Greaney, A. J., Hannon, W. W., Loes, A. N., Hauser, K., Dillen, J. R., . . . Bloom, J. D. (2022). Shifting mutational constraints in the SARS-CoV-2 receptor-binding domain during viral evolution. *Science, 377*(6604), 420-424. doi:10.1126/science.abo7896

Starr, T. N., Greaney, A. J., Hilton, S. K., Ellis, D., Crawford, K. H. D., Dingens, A. S., . . . Bloom, J. D. (2020). Deep Mutational Scanning of SARS-CoV-2 Receptor Binding Domain Reveals Constraints on Folding and ACE2 Binding. *Cell, 182*(5), 1295-1310 e1220. doi:10.1016/j.cell.2020.08.012

Starr, T. N., & Thornton, J. W. (2016). Epistasis in protein evolution. *Protein Sci, 25*(7), 1204-1218. doi:10.1002/pro.2897

Sternberg, P. W., & Horvitz, H. R. (1989). The combined action of two intercellular signaling pathways specifies three cell fates during vulval induction in C. elegans. *Cell, 58*(4), 679-693. doi:10.1016/0092-8674(89)90103-7

Svetlov, V., Vassylyev, D. G., & Artsimovitch, I. (2004). Discrimination against deoxyribonucleotide substrates by bacterial RNA polymerase. *J Biol Chem, 279*(37), 38087-38090. doi:10.1074/jbc.C400316200

Swanson, M. S., Malone, E. A., & Winston, F. (1991). SPT5, an essential gene important for normal transcription in Saccharomyces cerevisiae, encodes an acidic nuclear protein with a carboxy-terminal repeat. *Molecular and Cellular Biology, 11*(6), 3009-3019. doi:10.1128/mcb.11.6.3009

Sydow, J. F., Brueckner, F., Cheung, A. C., Damsma, G. E., Dengl, S., Lehmann, E., . . . Cramer, P. (2009). Structural basis of transcription: mismatch-specific fidelity mechanisms and paused RNA polymerase II with frayed RNA. *Mol Cell, 34*(6), 710-721. doi:10.1016/j.molcel.2009.06.002

Taatjes, A. C. S. a. D. J. (2020). Structure and mechanism of the RNA polymerase II transcription machinery. doi:10.1101/gad.335679

Tan, L., Wiesler, S., Trzaska, D., Carney, H. C., & Weinzierl, R. O. (2008). Bridge helix and trigger loop perturbations generate superactive RNA polymerases. *J Biol, 7*(10), 40. doi:10.1186/jbiol98

Tesileanu, T., Colwell, L. J., & Leibler, S. (2015). Protein sectors: statistical coupling analysis versus conservation. *PLoS Comput Biol, 11*(2), e1004091. doi:10.1371/journal.pcbi.1004091

Tewhey, R., Warner, J. B., Nakano, M., Libby, B., Medkova, M., David, P. H., . . . Frazer, K. A. (2009). Microdroplet-based PCR enrichment for large-scale targeted sequencing. *Nat Biotechnol, 27*(11), 1025-1031. doi:10.1038/nbt.1583

Thomas, J. H., Birnby, D. A., & Vowels, J. J. (1993). Evidence for parallel processing of sensory information controlling dauer formation in Caenorhabditis elegans. *Genetics, 134*(4), 1105-1117. doi:10.1093/genetics/134.4.1105

Toulokhonov, I., Zhang, J., Palangat, M., & Landick, R. (2007). A central role of the RNA polymerase trigger loop in active-site rearrangement during transcriptional pausing. *Mol Cell, 27*(3), 406-419. doi:10.1016/j.molcel.2007.06.008

Unarta, I. C., Goonetilleke, E. C., Wang, D., & Huang, X. (2023). Nucleotide addition and cleavage by RNA polymerase II: Coordination of two catalytic reactions using a single active site. *J Biol Chem, 299*(2), 102844. doi:10.1016/j.jbc.2022.102844

Vannini, A., & Cramer, P. (2012). Conservation between the RNA polymerase I, II, and III transcription initiation machineries. *Mol Cell, 45*(4), 439-446. doi:10.1016/j.molcel.2012.01.023

Vassylyev, D. G., Sekine, S., Laptenko, O., Lee, J., Vassylyeva, M. N., Borukhov, S., & Yokoyama, S. (2002). Crystal structure of a bacterial RNA polymerase holoenzyme at 2.6 A resolution. *Nature, 417*(6890), 712-719. doi:10.1038/nature752

Vassylyev, D. G., Vassylyeva, M. N., Zhang, J., Palangat, M., Artsimovitch, I., & Landick, R. (2007). Structural basis for substrate loading in bacterial RNA polymerase. *Nature, 448*(7150), 163-168. doi:10.1038/nature05931

Viktorovskaya, O. V., Engel, K. L., French, S. L., Cui, P., Vandeventer, P. J., Pavlovic, E. M., . . . Schneider, D. A. (2013). Divergent contributions of conserved active site residues to transcription by eukaryotic RNA polymerases I and II. *Cell Rep, 4*(5), 974-984. doi:10.1016/j.celrep.2013.07.044

Vos, S. M., Farnung, L., Linden, A., Urlaub, H., & Cramer, P. (2020). Structure of complete Pol II-DSIF-PAF-SPT6 transcription complex reveals RTF1 allosteric activation. *Nat Struct Mol Biol, 27*(7), 668-677. doi:10.1038/s41594-020-0437-1

Wade, P. A., Werel, W., Fentzke, R. C., Thompson, N. E., Leykam, J. F., Burgess, R. R., . . . Burton, Z. F. (1996). A novel collection of accessory factors associated with yeast RNA polymerase II. *Protein Expr Purif, 8*(1), 85-90. doi:10.1006/prep.1996.0077

Walmacq, C., Kireeva, M. L., Irvin, J., Nedialkov, Y., Lubkowska, L., Malagon, F., . . . Kashlev, M. (2009). Rpb9 subunit controls transcription fidelity by delaying NTP sequestration in RNA polymerase II. *J Biol Chem, 284*(29), 19601-19612. doi:10.1074/jbc.M109.006908

Wang, B., Predeus, A. V., Burton, Z. F., & Feig, M. (2013). Energetic and structural details of the trigger-loop closing transition in RNA polymerase II. *Biophys J, 105*(3), 767-775. doi:10.1016/j.bpj.2013.05.060

Wang, D., Bushnell, D. A., Huang, X., Westover, K. D., Levitt, M., & Kornberg, R. D. (2009). Structural basis of transcription: backtracked RNA polymerase II at 3.4 angstrom resolution. *Science, 324*(5931), 1203-1206. doi:10.1126/science.1168729

Wang, D., Bushnell, D. A., Westover, K. D., Kaplan, C. D., & Kornberg, R. D. (2006). Structural basis of transcription: role of the trigger loop in substrate specificity and catalysis. *Cell, 127*(5), 941-954. doi:10.1016/j.cell.2006.11.023

Wang, W., Walmacq, C., Chong, J., Kashlev, M., & Wang, D. (2018). Structural basis of transcriptional stalling and bypass of abasic DNA lesion by RNA polymerase II. *Proc Natl Acad Sci U S A, 115*(11), E2538-E2545. doi:10.1073/pnas.1722050115

Wei, H., & Li, X. (2023). Deep mutational scanning: A versatile tool in systematically mapping genotypes to phenotypes. *Front Genet, 14*, 1087267. doi:10.3389/fgene.2023.1087267

Weile, J., & Roth, F. P. (2018). Multiplexed assays of variant effects contribute to a growing genotype-phenotype atlas. *Hum Genet, 137*(9), 665-678. doi:10.1007/s00439-018-1916-x

Weinzierl, R. O. (2010). The nucleotide addition cycle of RNA polymerase is controlled by two molecular hinges in the Bridge Helix domain. *BMC Biol, 8*, 134. doi:10.1186/1741-7007-8-134

Werner, F., & Grohmann, D. (2011). Evolution of multisubunit RNA polymerases in the three domains of life. *Nat Rev Microbiol, 9*(2), 85-98. doi:10.1038/nrmicro2507

Westover, K. D., Bushnell, D. A., & Kornberg, R. D. (2004). Structural basis of transcription: nucleotide selection by rotation in the RNA polymerase II active center. *Cell, 119*(4), 481-489. doi:10.1016/j.cell.2004.10.016

Wierzbicki, A. T., Haag, J. R., & Pikaard, C. S. (2008). Noncoding transcription by RNA polymerase Pol IVb/Pol V mediates transcriptional silencing of overlapping and adjacent genes. *Cell, 135*(4), 635-648. doi:10.1016/j.cell.2008.09.035

Williams, R., Peisajovich, S. G., Miller, O. J., Magdassi, S., Tawfik, D. S., & Griffiths, A. D. (2006). Amplification of complex gene libraries by emulsion PCR. *Nat Methods, 3*(7), 545-550. doi:10.1038/nmeth896

Windgassen, T. A., Mooney, R. A., Nayak, D., Palangat, M., Zhang, J., & Landick, R. (2014). Trigger-helix folding pathway and SI3 mediate catalysis and hairpin-stabilized pausing by Escherichia coli RNA polymerase. *Nucleic Acids Res, 42*(20), 12707-12721. doi:10.1093/nar/gku997

Winston, F., Chaleff, D. T., Valent, B., & Fink, G. R. (1984). Mutations affecting Ty-mediated expression of the HIS4 gene of Saccharomyces cerevisiae. *Genetics, 107*(2), 179-197. doi:10.1093/genetics/107.2.179

Wu, C. C., Herzog, F., Jennebach, S., Lin, Y. C., Pai, C. Y., Aebersold, R., . . . Chen, H. T. (2012). RNA polymerase III subunit architecture and implications for open promoter complex formation. *Proc Natl Acad Sci U S A, 109*(47), 19232-19237. doi:10.1073/pnas.1211665109

Xie, X., Sun, X., Wang, Y., Lehner, B., & Li, X. (2023). Dominance vs epistasis: the biophysical origins and plasticity of genetic interactions within and between alleles. *Nat Commun, 14*(1), 5551. doi:10.1038/s41467-023-41188-8

Xu, L., Butler, K. V., Chong, J., Wengel, J., Kool, E. T., & Wang, D. (2014). Dissecting the chemical interactions and substrate structural signatures governing RNA polymerase II trigger loop closure by synthetic nucleic acid analogues. *Nucleic Acids Res, 42*(9), 5863-5870. doi:10.1093/nar/gku238

Xu, Y., Bernecky, C., Lee, C.-T., Maier, K. C., Schwalb, B., Tegunov, D., . . . Cramer, P. (2017). Architecture of the RNA polymerase II-Paf1C-TFIIS transcription elongation complex. *Nature Communications, 8*, 15741. doi:10.1038/ncomms15741
https://www.nature.com/articles/ncomms15741#supplementary-information

Yuzenkova, Y., Bochkareva, A., Tadigotla, V. R., Roghanian, M., Zorov, S., Severinov, K., & Zenkin, N. (2010). Stepwise mechanism for transcription fidelity. *BMC Biol, 8*, 54. doi:10.1186/1741-7007-8-54

Zaychikov, E., Martin, E., Denissova, L., Kozlov, M., Markovtsov, V., Kashlev, M., . . . Mustaev, A. (1996). Mapping of catalytic residues in the RNA polymerase active center. *Science, 273*(5271), 107-109. doi:10.1126/science.273.5271.107

Zhang, G., Campbell, E. A., Minakhin, L., Richter, C., Severinov, K., & Darst, S. A. (1999). Crystal Structure of Thermus aquaticus Core RNA Polymerase at 3.3 Å Resolution. *Cell, 98*(6), 811-824. doi:10.1016/s0092-8674(00)81515-9

Zhang, J., Palangat, M., & Landick, R. (2010). Role of the RNA polymerase trigger loop in catalysis and pausing. *Nat Struct Mol Biol, 17*(1), 99-104. doi:10.1038/nsmb.1732

Ziegler, L. M., Khaperskyy, D. A., Ammerman, M. L., & Ponticelli, A. S. (2003). Yeast RNA polymerase II lacking the Rpb9 subunit is impaired for interaction with transcription factor IIF. *J Biol Chem, 278*(49), 48950-48956. doi:10.1074/jbc.M309656200