SOX2 IN FOCUS: ASSOCIATION OF SOX2 COPY NUMBER VARIATION WITH TP53 MUTATION IN TCGA PANCANCER COHORTS AND CODON OPTIMIZED DESIGN FOR DE NOVO SOX2 SYNTHESIS USING NOVEL SHINY APPLICATION

A THESIS SUBMITTED TO
THE GRADUATE SCHOOL OF ENGINEERING AND SCIENCE
OF BILKENT UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR
THE DEGREE OF
MASTER OF SCIENCE
IN
NEUROSCIENCE

BY
Siber Güneş Çelik
JANUARY 2024

SOX2 IN FOCUS: ASSOCIATION OF SOX2 COPY NUMBER VARIATION WITH TP53 MUTATION IN TCGA PANCANCER COHORTS AND CODON OPTIMIZED DESIGN FOR DE NOVO SOX2 SYNTHESIS USING NOVEL SHINY APPLICATION

By Siber Güneş Çelik

January 2024

We certify that we have read this thesis and that, in our opinion, it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

Assoc. Prof. Özlen Konu KARAKAYALI

Assoc. Prof. Dr. Urartu Özgür Şafak Şeker

Assoc. Prof. Ceren Sucularlı

Approved for the Graduate School of Engineering and Science:

Prof. Dr. Orhan Arıkan

Director of the Graduate School of Engineering and Science

Abstract

SOX2 IN FOCUS: ASSOCIATION OF SOX2 COPY NUMBER VARIATION WITH TP53 MUTATION IN TCGA PANCANCER COHORTS AND CODON OPTIMIZED DESIGN FOR DE NOVO SOX2 SYNTHESIS USING NOVEL SHINY APPLICATION

Siber Güneş Çelik

M.Sc. in Neuroscience

Advisor: Özlen KONU KARAKAYALI

January 2024

Recombinant proteins are crucial for diverse research applications such as biosensors and cancer studies. Proteins are engineered through de novo gene synthesis methods. Numerous tools and databases have emerged to facilitate the design of recombinant proteins, starting from the design of the gene sequence. De novo DNA synthesis enables the synthesis of custom-designed sequences, allowing codon optimization to enhance expression yield in heterologous systems. In cancer research, recombinant expression of proteins involved in tumorigenesis-related signaling pathways is employed for functional studies, potentially revealing new therapeutic targets. A notable example is the pivotal role of SOX2 expression in the formation of cancer stem cells (CSCs) across various cancer types. Previous studies highlight SOX2 expression functionally overlaps with TP53 expression on the PI3K/AKT signaling pathway. This association may stem from the p53-MDM2 interaction.

This thesis investigates the association between SOX2 copy number gain and TP53 mutations within TCGA PanCancer cohorts. Fisher's exact test results reveal varying association, dependent on tissue type and specific driver mutations within each cancer type. The findings suggest the potential therapeutic relevance of SOX2 in cancer research. Furthermore, the thesis employs an in-silico approach to design de novo SOX2 synthesis, utilizing a novel shiny app that integrates codon optimization and primer design functionalities. The app enables simultaneous codon optimization for multiple expression systems and offers distance analysis through hierarchical clustering. Codon optimization feature provides control over the rate of replacement value for codon substitution which validated through a case study involving human insulin.

Finally, app design set of overlapping primers with synchronized melting temperature to be used in PCR assembly for de novo SOX2 gene synthesis.

Keywords: Recombinant Protein, Cancer, TCGA-PANCAN, SOX2, TP53, Association Analysis, Codon Optimization, PCR Assembly, Primer Design

Özet

SOX2 ODAK NOKTASINDA: SOX2 GEN KOPYA SAYISI VARYOSYONUNUN TCGA PANCANCER KÜMELERİNDE TP53 MUTASYONU İLE İLİŞKİSİ VE SHINY UYGULAMASI KULLANILARAK DE NOVO SOX2 SENTEZİ İÇİN KODON OPTİMİZE TASARIM

Siber Güneş Çelik
Nörobilim Yüksek Lisans Programı
Tez Danışmanı: Özlen KONU KARAKAYALI
Ocak, 2024

Rekombinant proteinler, biyosensörler ve kanser araştırmaları gibi çeşitli araştırma alanları için önem arz eder. Bu proteinler de novo gen sentezi yöntemleriyle sentezlenebilmektedirler. Gen sekansı tasarimdan başlayarak, rekombinant proteinlerin tasarımını kolaylaştırmak üzere bir çok araç ve veritabanı ortaya çıkmıştır. De novo DNA sentezi, özel tasarım sekanslarının üretilmesini sağlayarak kodon optimizasyonuna imkan tanır, bu da heterolog sistemlerde ekspresyon verimini artırmak için kullanılabilir. Tümör oluşumuyla ilişkili sinyal yolaklarındaki hedef proteinlerin rekombinant olarak ifadesi kanser araştırmalarında potansiyel olarak yeni terapötik hedefleri ortaya çıkarabilir. Buna örnek olarak, çalışmalar SOX2 ifadesinin, çeşitli kanser türlerinde kanser kök hücrelerinin oluşumunda önemli rol oynadığını öne çıkmaktadır. SOX2 ifadesi bir başka önemli kanser geni olan TP53 ifadesi ile PI3K/AKT sinyal yolakları aracılığıyla kesişmektedir.

Bu tez, TCGA PanCancer kohortları içinde SOX2 kopya sayısı artışı ile TP53 mutasyonları arasındaki korelasyonu araştırmaktadır. Kullanılan fisher's exact testi korelasyonun, doku tipine ve her kanser türündeki belirli sürücü mutasyonlara bağlı olarak değişen bir yapıda olduğunu ortaya koymaktadır. Bulgular, SOX2 geninin kanser araştırmalarında potansiyel terapötik önemini işaret etmektedir. Ayrıca, tez, de novo SOX2 sentezi için in silico bir yaklaşım kullanarak, kodon optimizasyonu ve primer tasarım fonksiyonlarını entegre eden yeni bir shiny uygulamasını kullanmaktadır. Uygulama, birden fazla ifade sistemi için eş zamanlı kodon

optimizasyonunu sağlar ve hiyerarşik kümeleme yoluyla mesafe analizi sunar. Ek olarak kodon optimizasyon için sunulan kodon değişim oranları aplikasyon üzerinde değiştirilerek kontrol edilebilir. Bu parametrenin optimizasyon üzerindeki etkisi insan Insulin proteini kullanılarak yapılan bir analizle ortaya konmuştur. Son olarak, uygulama, SOX2 sentezini gerçekleştirebilmek adına multipleks PCR prosedüründe kullanılmak üzere çift ipliğin ayrılma ısıları senkronize edilmiş primer seti tasarlar.

Anahtar kelimeler: Rekombinant Protein, Kanser, TCGA-PANCAN, SOX2, TP53, İlişki Analizi, Kodon Optimizasyonu, Multipleks PCR, Primer Tasarım

Nothing in life is to be feared, it is only to be understood.

Now is the time to understand more,

so that we may fear less.

Marie Curie

Acknowledgements

I extend my utmost gratitude and appreciation to my supervisor, Assoc. Prof. Özlen Konu Karayalı, for her invaluable scientific guidance and unwavering encouragement that made the completion of this thesis possible. Her profound insights not only transformed my perspective but also guided me in formulating the right questions. From my undergraduate years onward, she has played a pivotal role in shaping me into a scientist, providing an environment for me to flourish. Working under her leadership has been both an extreme privilege and honor.

I would like to express my sincere thanks to the committee members, Assoc. Prof. Dr. Urartu Özgür Şafak Şeker and Assoc. Prof. Ceren Sucularlı, for dedicating their valuable time to assess and provide valuable remarks on my thesis. Additionally, I am grateful to Şükrü Atakan for his valuable insight and contributions.

Acknowledgment also goes to the members of the Konu Lab – Rana Acar, Mert Demirdizen, Abdul Moiz Aftab, Merve Vural, Seda Aktürk, Algı Demirbaş, Selvin Yıldız, Çağdaş Yanyatan, Muazzez Çelebi Çınar. I consider myself exceptionally fortunate to collaborate with such kind and intelligent individuals. Their incredible support has been instrumental in reaching this milestone, and I will forever cherish the moments spent with this remarkable team. Special thanks to our Lab alumni, Melike Tombaz, İrem Burçin Arıcı, Güneş Tok, and Rüya Tombuloğlu who were consistently there for me when needed. The time we have spent together always filled with laughter and joy.

Deepest love and gratitude are extended to my parents, Medine Güneş, and Taner Kılınç, for their unwavering love and support. The constant presence and support of my grandparents, Şuşar and Celal Kılınç, and my aunt and uncle, Tansel and Çağatay Arısoy, have contributed significantly to shaping the person I am today. Without the support of my family, this success would not have been possible I am so lucky to have them as my family, and I will always be grateful.

Lastly, I express my thanks to the Bilkent Neuroscience and Molecular Biology and Genetics departments for providing an environment and the right tools for my scientific studies. During part of my graduate studies, I was supported by the Scientific and Technological Research Council of Turkey (TUBITAK) (Project No. 1001). Special thanks are extended to Ankara University for their contributions.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Abbreviations

RSCU : Relative synonymous codon usage

CAI : Codon adaptation index

PCA : Polymerase chain assembly

PCR : Polymerase chain reaction

SGS : Simplified gene synthesis

ESC : Embryonic stem cell

CSC : Cancer stem cell

TCGA : The cancer genome atlas

EC : Esophageal carcinoma

LIHC : Liver hepatocellular carcinoma

LUAD : Lung adenocarcinoma

ESCA: Esophageal carcinoma

BC : Bladder Cancer

OV : Ovarian Cancer

THCA : Thyroid carcinoma

HEG : Highly expressed gene

PLC : Pan Lung Cancer

CNV : Copy number variation

HMG : High mobility group

DDR : DNA damage response

ESCC : Esophageal Squamous Cell Carcinoma

SNP : Single nucleotide polymorphism

BRCA: Breast invasive carcinoma

GBM: Glioblastoma multiforme

NSCLC: Non-small cell lung cancer

1 INTRODUCTION

1.1 Recombinant Protein Production

Recombinant proteins are a widely employed for numerous techniques in various research areas, involving the cloning of proteins encoded by recombinant DNA within a host system to enhance protein yield [1]. High protein yield is seen as an important goal in pharmaceutical and medical application. This is because, in many disorders stemmed from the impairment of a protein can be treated by providing the protein exogenously [2]. The most useful recombinant protein products in medical setting are hormones, enzymes, antibodies, and vaccines [3]. Numerous recombinant proteins can be offered as biopharmaceuticals for the treatment of many diseases and disorders from B12 deficiency to HIV [4]. One of the prime examples of this is recombinant insulin production. Insulin is produced in large quantities in many different expression systems, *E.coli*, yeast, mammalian, and insect cell lines being the most predominant choices [5]. Moreover, recombinant DNA technologies are used in anti-cancer treatments. Recombinantly produced immunotherapeutic can potentially kill tumor cells and stimulate immune response in a form of vaccine, immunotoxin, antibody, fusion protein, or many more [6]. Additionally, recombinant proteins find utility in biosensors and immunoassays [7], [8]. [7], [8].

Several steps are followed for recombinant protein production. The initial step involves identifying the specific gene fragment to be expressed [9]. Advances in next generation sequencing technology have provided researchers with extensive libraries of confirmed DNA sequences from many organisms, allowing for the selection of the desired fragment [10]. Often, only a distinct segment of the genome governs the

functionality under investigation, making the expression of the entire gene sequence unnecessary. For instance, in Salehi et al. 2010, the researchers successfully produced the recombinant L2 domain of the human EGFR gene. This strategic approach was chosen because the L2 domain contained the ligand-binding site. As a therapeutic approach, recombinant L2 domain is used for competitive inhibition of EGFR [13]. Such targeted expression of specific gene segments enables investigation of functional domains and potentially use them as therapeutics.

Obtaining complementary DNA (cDNA) can be achieved through amplification of genomic DNA or *de novo* oligo synthesis [11][12]. Synthesizing DNA molecule out of scratch enables gene and genome synthesis without restriction of tangible DNA. Additionally, allowing researchers to design custom sequences [13]. Mueller et al. 2008, provides an in-depth review of the historical usage of de novo synthesis of gene and genome in synthetic biology. The review highlights the method's capability of codon optimization and sequence manipulation.

De novo gene synthesis can be accomplished via several different methods[13], [14], [15]. One of the most common methods involves PCR-based assembly of short oligonucleotide sequences[12], [16], [17]. PCR-based DNA synthesis method uses short overlapping oligonucleotides (25 – 40bp) for the assembly. The overlapping regions work as a template for PCR reaction under appropriate conditions[12], [14], [17].

In 2016, Sequeira et al., recombinantly produced small animal toxins via PCR-based oligonucleotide assembly without tangible DNA [14]. Same year, another study successfully produced a synthetic bacterial genome by combining several methods

including PCR-based assembly, *in vitro* assembly, and Yeast assembly [18]. Both studies demonstrating the applications of *de novo* gene and genome assembly. Another important step in recombinant protein production is codon optimization [19]. Synonymous codons can encode the same amino acid due to the degeneracy of the genome and each organism has its unique preference on which codon to use [20]. Numerous studies have been investigated the mechanism behind the codon usage bias and many of them linked codon usage bias to translational selection, mutational bias, and mutational drift [21], [22], [23]. Different metrics that quantify codon usage bias have been developed to assess the codon usage patterns for many organisms. Codon Bias Database (CBDB) provides codon bias information for over 300 bacterial genomes [20]. Previous studies showed that codon optimization successfully increased the yield of the protein [14].

To summarize, recombinant protein production is a critical technique employed in various research areas including pharmaceuticals, medical applications, and cancer research [5], [6], [8], [24], [25], [26]. The advancements in recombinant DNA technologies allows researchers to engineer the protein of interest for various purposes [13], [27], [28]. Today's technology allows the synthesis of gene and genomes via different chemical and enzymatic reactions [14], [18], [29]. PCR-based assembly methods are predominantly used for de novo oligo synthesis [12], [14], [17]. Finally, de novo DNA synthesis allows researcher to codon optimize the gene sequence according to the host organism's codon bias to enhance protein expression [30]. [18], [30][24]However, there is still a need to combine multiple aspects of gene fragment design to synthesize as described in the following sections.

1.1.1 Codon Bias and Codon Optimization

Due to the degeneracy of the genome, multiple codons can be used interchangeably to encode the same amino acid [23]. Each organism has its own unique preference on which synonymous codon to use, therefore, there is a codon usage bias in the genome [20]. The early studies on evolution of proteins in molecular level hypothesized that most evolutionary changes in proteins likely stemmed from neutral mutations and genetic drift [31]. This hypothesis lead researches to postulate that genome composition is not random and codon usage is biased towards tRNA pool of the organism [32], [33]. The first sequencing of *E.coli* genome partially confirmed the postulated idea [34], [35].

Advancements in sequencing technology led to many more studies conducted to investigate codon usage bias of prokaryotic and eukaryotic genomes [20], [36], [37]. One striking finding from sequencing analysis was that *Escherichia coli* rarely uses synonymous arginine codon, AGA in its genome. However, AGA is the most used synonymous arginine codon in humans. Study reviles tRNA that encodes AGA codon is in very low abundance in *E.coli* which leads slow or mistranslation of the protein [23], [34], [35]. This implicates that translation efficiency can be increased in recombinant proteins by substitution of synonymous codon based on codon usage bias of the host organism. Thus, many researchers practice “codon optimization” method to increase recombinant protein yield [14].

The mechanism behind the codon usage bias is extensively studied [32], [33], [34], [35], [36], [37]. Recent paper review and discuss the findings. Paper states that codon usage bias varies not only between species but also between genes. Codon usage bias

mainly occurs through translational selection, mutational bias, and mutational drift in various organisms. However, other factors such as GC content, mRNA folding, tRNA abundance and more contribute to shape codon usage bias, thus factors shaping the codon usage bias remains complex and multidimensional [23], [32].

The mechanism behind the codon usage bias is investigated from tRNA's point of view in Rocha et al., 2004. Paper states that translation is an energetically costly process within cells, and translational efficiency is under selective pressure. The diffusion of tRNA to the ribosome is a rate limiting step during the elongation cycle of the polypeptide chain. Consequently, codons preferentially recruit the most abundant tRNAs, particularly in highly expressed genes (HEGs). The evolutionary force causing this bias in the codon composition is identified as translational selection [23]. However, studies with higher eukaryotes showed that translational selection is not driving force in every organism. For example, human genome composition is solely relying on GC content and isochore composition while fly is partially affected by translational selection [37]. These observations were explained by mutational selection and mutational drift theory which are the evolutionary forces that ensure the selection of the fittest codon in a population while random mutations are occurring in the genome [23], [38].

1.1.1.1 Quantification of Codon Bias

Quantification of codon usage bias can be achieved by calculating codon usage frequencies within the genome [20]. Different computational methods were developed with the focus on driving factors of codon usage bias [23], [32], [33], [39].

One or combinations of tRNA abundance, codon composition, mRNA stability, and other novel parameters were considered to develop codon usage bias metrics. These metrics have been used in the development of a technique known as codon optimization in many tools [21], [22], [40]. The most used metrics are explained below.

The codon adaptation index (CAI) model calculates relative frequency for each codon by using reference set of highly expressed genes (HEGs) in the genome [39], [41]. Each gene gets a score based on codon composition and relative adaptiveness index. The CAI metric is created by taking geometric mean of relative synonymous codon usage (RSCU) among synonymous codon within the reference gene set [42]. Index ranges from 0 to 1, 1 representing the most preferred synonymous codon [20], [39], [41]. Numerous tools have been proposed by either using or extending CAI metric. One example is CAIcal (http://genomes.urv.es/CAIcal) which is an online server that provides easy to access CAI calculation for given set of reference genes [42]. Similarly, VectorBuilder (https://en.vectorbuilder.com) uses CAI as an index for codon optimization [43]. For explanatory figure please refer to **Appendix Figure 2**.

The tRNA adaptation index is designed based on tRNA abundance theory which states that codon with higher abundance of anti-codon and tRNA is more preferred for translational efficiency [23]. The tRNA adaptation index calculates codon usage based on distribution of tRNA gene copy numbers and anticodon types [20], [23], [40]. As an example, Optimizer (http://genomes.urv.es/) online codon optimization tool uses tRNA based model to codon optimize bacterial genome. Tool combines CAI data with tRNA copy number information to codon optimize the sequence [22].

However, this index may not be useful for every organism since tRNA theory in not applicable to some higher eukaryotes as it was discussed previously [31], [32], [44]. Moreover, Reis et al., 2004 discussed that tRNA distribution may vary depending on the growth rate or tissue type which limits the reproducibility of the results. Furthermore, novel algorithms have been developed to predict codon usage bias in each genome. One study from 2022 introduces a novel algorithm that utilizes mRNA stability prediction model and substitute the synonymous codons accordingly. Paper states that mRNA stability correlates with codon composition through method called codon optimality and propose a novel codon optimization method based on this method [21].

1.1.1.2 Tools to Make Codon Optimization Possible

To streamline the codon optimization process, numerous tools have been developed which I have summarized below (**Table *1-1***). Each optimizer employs specific codon optimization metrics to enhance codon usage. Only some organisms are supported by each optimizer. Each tool is summarized below.

ExpOptimizer utilizes CAI metric to codon optimize DNA/RNA or protein sequence for wide range of prokaryotic and eukaryotic expression systems. Tools allows users to exclude restriction sites from the sequence [45].

VectorBuilder provides a codon optimization tool for various prokaryotic and eukaryotic expression system using CAI metrics. Tool additionally optimizes GC% in the sequence. Similar ExpOptimizer, restriction sites can be excluded in VectorBuilder [43].

GenSmart is an online codon optimization tool developed by GeneScript (http://genescript.com) and uses novel population immune algorithm to codon optimize given DNA or protein sequence. GenSmart tool allows dual codon optimization for two expression system simultaneously [46].

Optimizer tool uses combination of CAI tRNA-based codon usage index to codon optimize the sequence for bacterial expression system. It takes both DNA and protein sequences as an input [22].

IDT codon optimization tool developed by Integrated DNA Technologies Inc (www.idtdna.com). IDT tool utilizes novel algorithms to lower complexity and minimize secondary structure while codon optimizing DNA or protein sequences. Similar with previous tool, IDT tool supports wide range of host organisms [19].

iCodon codon optimization tool uses novel algorithm based on mRNA stability predicter model for codon optimization. Tool currently supports four vertebrates: mouse, human, frog, and fish [21].

Table 1-1. The summary of existing codon optimization tools available in the literature.

Optimization Tool	Codon Optimization Index	Supported Organisms
ExpOptimizer [45]	Codon Usage Bias Index	Prokaryotic and Mammalian Expression Systems
VectorBuilder [43]	CAI	Prokaryotic and Mammalian Expression Systems
GenSmart ™ [46]	Population Immune Algorithm	Prokaryotic and Mammalian Expression Systems
Optimizer [22]	tRNA-scan Software, CAI	Bacteria
IDT [19]	Novel Algorithm	Prokaryotic and Mammalian Expression Systems
iCodon [21]	mRNA Stability Prediction Model	Human - Mouse – Xenopus - Zebrafish

1.1.2 De novo Gene Synthesis via PCR Assembly

De novo gene synthesis is a recombinant DNA technology that involves the artificial construction of a gene or a genome from scratch using a combination of chemical and enzymatic methods [13], [14], [18], [29]. It is particularly useful when tangible DNA is unavailable, and it allows for custom design of gene sequences, additionally enables codon optimization [14], [17]. This technique has gained popularity in the fields of synthetic biology and biotechnology, gradually replacing classical cloning methods due to its ligase free, rapid, easy nature [12], [17], [47].

PCR-based protocols are commonly employed for de novo gene synthesis, including self-priming PCR, PCR assembly, and Polymerase Chain Assembly (PCA) [13], [48], [49]. PCR assembly-based techniques are advantageous as they do not require ligation and are independent of enzymatic ligases and utilizes small sets of short overlapping oligo sequences as the starting material for constructing larger synthetic fragments. [12], [13], [47], [49]. For explanatory figure please refer to **Appendix Figure 1**.

Recent advancements have led to the development of a simplified gene synthesis (SGS) protocols. The SGS protocol utilizes oligonucleotides with a total length of 40 nucleotides (nt) and a 20 nt overlapping region. These oligonucleotide fragments are assembled using a unique PCR-assembly reaction [13], [49].

The design of overlapping primers is a crucial step in PCR assembly as it effects the specificity and efficiency of the PCR reaction. The parameters such as melting temperature (Tm), C-G content, and primer length must be optimized to ensure successful PCR reaction Furthermore, the melting temperatures should be synchronized within the primer set to facilitate annealing of the overlapping regions [12], [13], [47], [50]. [48]As an example, Sequeira A, et al. 2016 *de novo* synthesizes 96 small genes, including venom peptides and animal toxins by using PCR assembly method. In this study Maximum length of the primers were 60nt and the length of the overlapping region ranged from 15-25 bp. In another study, PCR assembly method is used to synthesize genes encoding EPCR-1 and pUC19 β-lactamase. Similarly, overlapping length is set to 20 bp in the study [17].

1.1.2.1 Tools That Design Overlapping Primers for PCR Assembly

Automated tools that can generate overlapping primers for PCR assembly are highly desirable. Existing tools for overlapping primer design are listed in ***Table 1-2***. The table states whether the tool has melting temperature, CG% and primer length control over the design. Additionally, required inputs and programming language that the tool has been written can be found in the table.

The PCR Suite tool is developed in as an extension of a R package Primer3. Tool CGI is written in Perl and provides design for overlapping primers for PCR assembly. The PCR Suite allows optimization of melting temperature, GC content, and dimerization parameters for primer design [51].

FASTPCR is Java based software that can design PCR primers and probes for oligonucleotide assembly studies such as PCA or Gibson Assembly. FASTPCR provides melting temperature and GC content control over primer design. Additionally, restriction enzyme recognition sites can be found and created within the given coding sequence [52].

STITCHER 2.0 tool designs overlapping primers for PCR based assembly method. Tool provides extensive control over the parameters such as dimerization, self-complementation, GC content, and melting temperature [50].

Primirize is an online tool developed and funded by Stanford University. Primirize tool offers graphic representation of each designed primer over reference sequence. Additionally, tool curates a protocol based on designed primers inside the application, thus providing a unique feature within similar tools [53].

Table 1-2 The summary of existing overlapping primer tools.

Design Tool	Programming Language	Input	Tm	GC %	Length
The PCR Suite [51]	R and Perl programming Language	sequence terminator promoter		√	√
FASTPCR [52]	Java	sequence		√	√
STITCHER 2.0 [50]	JavaScript, CSS and HTML5	sequence	√	√	√
Primirize [53]	Python	sequence	√		√

1.2 Use of Recombinant Proteins in Cancer Research

Cancer is a condition mainly characterized by uncontrolled cell invasion in certain type of tissue. This condition is one of the most common causes of death in humans [54]. Besides traditional methods, new therapies have emerged for cancer treatment known as immunotherapies [55], [56], [57]. This therapy uses recombinant DNA products to either kill cancer cells or stimulate immune cells. Protein based therapeutics such as vaccines, immunotoxins, and fusion proteins are commonly used in cancer treatment [6], [55], [56], [57].

As an example, recombinant Apo2L/TRAIL (Apo2 ligand or tumor necrosis factor (TNF)-related apoptosis-inducing ligand) protein is discovered to be effective

against cancer cells since it induces apoptosis to cancer cell without damaging the normal tissue [58]. More recent study recombinantly expresses TRAIL protein in *E.coli*. In this study, authors try and increase the anti-tumor effect of recombinant TRAIL protein by combining sorafenib, tumor penetrating peptides iRDG, and recombinant TRAIL protein in cell treatment [59]. Similarly, L2 domain of human EGFR protein is recombinantly produced in *E.coli* to be used for competitive inhibition of EGFR or screening for specific ligands [25]. A novel recombinant fusion protein is developed to target HER2/neu expression in breast cancer cells. The fusion construct is generated via 2-step PCR reaction and overlapping primers were utilized [24].

The different versions of the proteins such as isoforms of mutated proteins play key role in many human cancers thus, they are extensively studied [60], [61], [62], [63], [64], [65], [66], [67]. As an example to this, one study designs a chimeric p53 protein, M3-p53-R12. Transduction of this protein induces cell cycle arrest and apoptosis in human leukemia cells [63]. Additionally, overexpression of certain genes such as SOX2 is linked with poor disease diagnosis [68], [69], [70], [71] thus, indicating the need for more functional research to either target or use recombinant products to manipulate the disease prognosis.

Two of the most important cancer genes, is a tumor suppressor/oncogene TP53 and stem cell marker and oncogene SOX2, which are discussed in more detail below. In the sections below I summarize the literature on how recombinant technologies have been used to clone and expressed these two proteins in different models and which techniques they have applied and what they have found. In addition, I also

make a point showing that these two commonly altered proteins can be studied in vitro in combination to advance our understanding of cancer mechanisms.

1.2.1 TP53

The TP53 gene which encodes p53 and situated on chromosome 17 p13, is recognized as a tumor suppressor gene [72]. The gene codes for a nuclear phosphoprotein weighing 53 kD, and it plays a distinctive role in preventing the development of tumors [73]. The wildtype p53 gene is a critical tumor suppressor that plays a pivotal role in overseeing numerous cellular processes, such as DNA repair, programmed cell death (apoptosis), halting the cell cycle, inducing cellular aging (senescence), and regulating metabolism [65], [72].

Mutations in the p53 gene, particularly those occurring in the DNA binding region, are widespread in various cancer types. These mutations have distinct effects on tumor development, influencing initiation, promotion, aggressiveness, and metastasis in diverse ways. Mutated p53 proteins lose their ability to suppress tumors but often acquire new functions that promote the growth and cell survival through affecting downstream signaling pathways [66], [74], [75].

1.2.1.1 Recombinant p53 in Cancer Research

Mutations in TP53 gene are the most observed mutations in numerous human cancers and recognized as a driver event for many tumor types [72], [76], [77]. TP53 mutation can lead to loss of tumor-suppressor function which is related to tumorigenesis [72], [76]. Furthermore, mutated (mut) p53 can work differently than wild type (wt) p53 in

the cell causing dysregulation of downstream signaling [64], [65], [78], [79]. Mutation status of TP53 gene with respect to other gene activities is extensively studied with both functional and computation methods [80], [81], [82], [83]. In 2008, one study expressed recombinant wtp53 and gonadotropin-releasing hormone-p53 (GnRH-p53) fusion protein to rescue p53-defficient tumor cells. Cell treatment resulted in low level of cell proliferation and increase in apoptosis [84]. As it is main characteristics for many tumor types, both p53-deficient and p53 knockout (KO) cells are utilized in different cancer research to imitate p53 deficient tumor cells such as breast cancer or lung cancer cells [71], [81], [85]. Conversely, overexpression of p53 via recombinant adenovirus-p53 therapy shows promising effect in cervical cancer treatment [86]. These examples and many more [82], [87], further confirms the important role of recombinant p53 molecules in cancer research.

1.2.2 SOX2

SOX2, alternatively referred to as the sex-determining region Y (SRY)-box 2 is located on chromosome 3q26.3-q27 and holds an important position as a fundamental member among the core transcription factors associated with pluripotency [88]. It promotes and sustains pluripotency in embryonic stem cells (ESCs), as well as in lineage-committed progenitors and tissue stem cells, mainly of epithelial or neural lineage[88]. Additionally, SOX2 acts as a potent oncogene in various cancer types, regulating cancer stem cells (CSCs) and other cancer hallmarks [69], [89], [90], [91], [92]. The oncogene role of SOX2 contributes to cancer-related processes[93].

Furthermore, the functionality of the protein is influenced by various secondary modifications, which can vary depending on the tissue and species[94].

1.2.2.1 The Key role of SOX2 in Various Cancer Types

The involvement of SOX2 protein in cancer is multifaceted and context dependent. It is overexpressed in various cancers like lung, breast, and esophageal cancer, correlating with poor disease prognosis[95], [96], [97]. SOX2 promotes cancer cell proliferation, migration, invasion, and likely aids in CSC maintenance [90], [91], [92], [93], [98]. However, it may also exhibit tumor-suppressive effects in certain contexts, influenced by cellular factors and coexisting genetic changes[93].

Cancer stem cells are found within tumors, and they possess the ability of self-renewal and differentiation, driving tumor growth, and contributing to cancer progression and relapse. If CSCs are not targeted during the cancer treatment, those cells with seeding capacity may cause CSCs-mediated clinical relapse in the future. [99]. High levels of SOX2 are associated with stem cell-like traits in multiple tumor types. Studies have revealed that SOX2-expressing cells can drive cancer malignancy by initiating and propagating tumor growth and giving rise to differentiated cell progenies [69], [89]. SOX2 is considered a CSC marker in lung cancer [100], thyroid cancer [101], bladder cancers [98], ovarian cancers [92], and high-grade gliomas [102], and others.

SOX2 plays a significant role in therapeutic resistance and cancer relapse. It facilitates pro-survival and anti-apoptotic signaling in various cancers, contributing to drug resistance. In lung cancer, SOX2 expression leads to resistance against drugs

like erlotinib, cisplatin, and paclitaxel by suppressing pro-apoptotic genes (BIM and BMF) and activating oncogenic EGFR and BCL2L1[103], [104]. In ovarian cancer, SOX2+ cells induce therapy resistance against staurosporine, carboplatin, cisplatin, and paclitaxel by promoting CSCs and upregulating the anti-apoptotic factor BCL2 while reducing expression of pro-apoptotic proteins PUMA and NOXA [92].

Another study, Hagerstrand et al. 2011 identifies SOX2-dependent subset of tumor- and sphere-forming glioblastoma cells. This subset of cells exhibits a distinct profile of tyrosine kinase inhibitor sensitivity. The study found that SOX2 is highly expressed in neuroglial tumors. Additionally, SOX2 is expressed in glioblastoma with other pluripotency genes . Silencing of SOX2 in glioblastoma can increase cell proliferation and tumorigenicity. The paper also provides gene expression analyses of high-grade glioma cultures and the clinical significance of the two types. The study concludes that SOX2 plays a crucial role in the development and progression of glioblastoma and may be a potential target for the development of better glioblastoma therapies [102].

1.2.2.2 Recombinant SOX2 in Cancer Research

Recent studies indicate that SOX2 is a key molecular regulator for CSCs in lung cancer [100], ovarian cancer, bladder cancer [98], and colorectal cancer [105]. CSCs are linked with asymmetric cell division and chemotherapy resistance in many cancers such as lung [106] and gastric cancer [107]. This makes SOX2 gene a potential target of immunotherapy [93], [104].

As an example, one study utilizes recombinant SOX2 protein to screen B-cell epitopes as a biomarker in glioblastoma, prostate cancer, and lung cancer [108]. Interestingly, not only SOX2 itself but also, a genetic element called SOX2OT is proven to modulate SOX2 expression [88]. Similarly, another study uses synthetic SOX2OT cDNA to create a vector that overexpress long non-coding RNA (lncRNA) SOX2OT. Sox2OT overexpression significantly reduced the cell proliferation rate [109]. As a different approach, Polakova et al., uses recombinant Sox2 proteins to develop antitumor DNA vaccine against Sox2 transcription factor. These examples further emphases the various applications of recombinant Sox2 protein in cancer early prognosis and treatment [91].

1.2.2.3 The Association Between SOX2 and TP53 Expression

The SOX protein family commonly defined with its conserved high mobility group (HMG) which is a DNA-binding element. This feature makes SOX proteins transcriptional regulators [69], [93]. SOX2 protein is the most explored member of the SOX family. This protein imposes many chemical modifications including methylation, ubiquitination, acetylation and so on [26], [69], [70]. Due to its DNA-

binding ability alongside with its role on self-renewal and stemness, SOX2 exerts various functions in ontogenesis, reprogramming, and cancer [68], [69], [89]. The research on SOX2-driven biology is extensive since it interacts with a lot of cancer pathways [71], [75], [88], [101]. The subcellular localization and regulation of SOX2 is mainly mediate by a specific stem-cell branch of PI3K/AKT signaling pathway [75]. Interestingly, the opposing branch of this pathway involves the tumor suppressor gene p53. Upon DNA damage,the interaction between p53 and MDM2 molecules weakens. This disruption causes accumulation of p53 thus effecting SOX2 expression via PI3K/AKT pathway. This interplay signifies that SOX2 and p53 functionally overlaps on PI3K/AKT signaling pathway and system gets dysregulated upon DNA damage [65], [75], [79], [88], [104], [110].

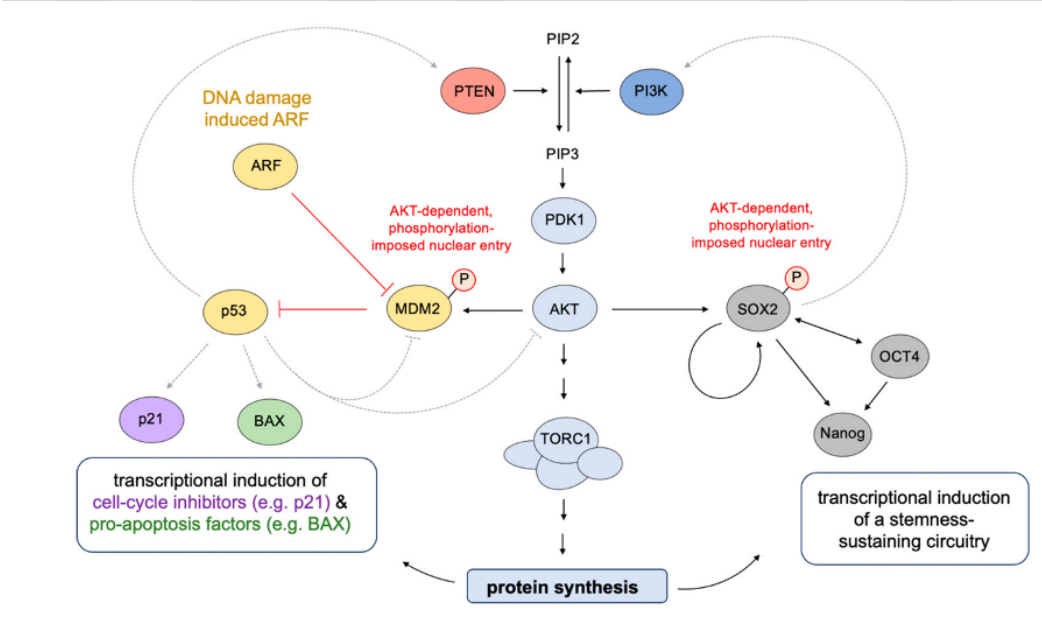


Figure 1.1 Schematic overview of PI3K/AKT signaling pathway. (Adopted from Schaefar et al., 2020)

Numerous studies delve into the potential correlation between SOX2 and p53 expression across different cancer types. An early investigation by Erdem et al. explored this connection in non-small cell lung cancer. The focus centered on the relationship between TP53 mutations and SOX2 copy number alterations. By analyzing gene expression data and conducting qPCR, the study revealed that TP53 mutations correlated with an elevated likelihood of acquiring SOX2 copy number alterations ($p = 0.017$). Notably, these alterations were more frequent in tumor tissues (34%) than adjacent non-tumorous tissues (3%).

In 2017, Lui et al. examined 227 esophageal tissue samples from 70 patients with esophageal squamous cell carcinoma (ESCC). Using next-generation sequencing, they identified shared mutations in TP53 and CDKN2A, as well as copy number alterations in specific regions: 11q (CCND1), 3q (SOX2), 2q (NFE2L2), and 9p (CDKN2A), present in more than 10% of paired IEN and ESCC tissues.

The subsequent year, Wang et al. investigated the same cancer type, ESCC. Their focus was on the prognostic value of SOX2, Cyclin D1, p53, and ki-67 in cancer patients. Through immunohistochemical analysis, all samples displayed elevated expression of the investigated genes. Subsequent statistical analysis indicated that high SOX2 expression correlated significantly with N stage ($p=0.034$), differentiation ($p=0.003$), and ki-67 expression ($p=0.001$). Conversely, heightened Cyclin D1 expression was linked to increased p53 levels ($p=0.015$). Intriguingly, patients with high SOX2 expression exhibited notably improved survival compared to those with lower expression ($p=0.021$).

Conversely, positive correlations between SOX2 and p53 expression were documented for hepatocellular carcinoma and ovarian carcinoma in subsequent years. These findings underscore the intricate behavior of SOX2 within distinct cancer contexts [67], [92], [111], [112], [113].

The studies mentioned above, along with others [89], [99], provide insight intothe correlation between SOX2 and TP53. Statistical analysis underscores that a certain level of correlation exists between SOX2 and TP53. The nature of this connection varies depending on tissue types and the cancer environment. Nonetheless, the precise mechanisms underlying this phenomenon and the link between TP53 mutations and SOX2 copy number alterations across numerous cancer types remain subjects of ongoing investigation. It is important to identify such pairs of oncogenes (mutant TP53/amplified SOX2) could clarify the role of combinatorial effects of multiple cancer inducing proteins. Use of recombinant proteins of SOX2 either through codon optimized high production of clones in cancer cells or delivery of such proteins [70], [82], [84], [87], [91], [114] could pave the way for treatment options. Use of streamlined algorithms to produce such combinations could also be beneficial for researchers to go ahead with their studies.

1.3 Online Platforms and Statistical Tools to Interrogate Copy Number Variation and Mutation Status of Cancer Genes

1.3.1 TCGA and cBioPortal

The Cancer Genome Atlas (TCGA) is a pioneering project focused on analyzing and interpreting molecular profiles of clinical tumors from various types and subtypes. Using advanced technologies like exome sequencing, copy number variation analysis, DNA methylation profiling, mRNA and microRNA expression analysis, and transcript splice variation assessment, TCGA aims to gain crucial insights into cancer's genomic complexities. Alongside genomic data, TCGA includes whole-genome sequencing and Reverse Phase Protein Arrays (RPPAs) for additional information[115]. TCGA, established under the National Institute of Health (NIH), offers openly available molecular and clinical information for over a 30 types of human cancers, encompassing comprehensive data types such as exome variant analysis, single nucleotide polymorphisms (SNPs), DNA methylation patterns, mRNA transcriptomes, microRNA (miRNA) data, and proteomes. The repository covers various sample types, including primary solid tumors, recurrent solid tumors, metastatic tumors, solid tissue and more [116].

The cBio Cancer Genomics Portal is an open-source tool that designed to explore cancer genomics data sets. It provides access to data from over 5,000 tumor samples derived from 20 cancer studies. The platform enables researchers to easily access and integrate cancer genomic data through a user-friendly interface. It features visualization options and statistical analysis. This portal effectively bridges the gap between genomic data and cancer researchers, offering them access to molecular

profiles and clinical attributes from large-scale cancer genomics projects. In conjunction with existing tools like the TCGA, the cBio portal uniquely focuses on analysis of specific genomic events in different cohorts and provides interactive network analysis [117].

1.3.2 Statistical Correlation Tests

Researchers can compare proportions of a categorical outcome among different independent groups using various statistical tests like the chi-squared test, Fisher's exact test, and z-test. These tests assess independence between two variables when dealing with independent and uncorrelated groups. The chi-squared test approximates results for large samples, while Fisher's exact test is ideal for smaller sample sizes. These statistical methods are essential for analyzing categorical data, and the choice of test depends on the specific research data and sample sizes under investigation [118].

The Fisher exact test is a nonparametric statistical method used for categorical data, serving as an alternative to the chi-square test in situations where the latter is not applicable, particularly with small sample sizes. It is employed to compare two data sets that generate a contingency table and assess the association or contingency between the two criteria. This test is particularly useful when both data sets involve a binary "yes/no" response, such as the presence or absence of tumor cells, or positive/negative blood cultures [119].

2 AIMS AND RATIONALE

The rapid progress in recombinant DNA technology have revolutionized recombinant protein production by allowing codon optimization and de novo gene synthesis. Recombinant proteins are used as immunotoxins, fusion proteins, and many more for treatment of various human cancers [2].

Traditional cloning methods are progressively being replaced by de novo gene synthesis, frequently employing PCR-based approaches with overlapping primer sets. Codon optimization holds paramount importance in *de novo* gene synthesis, especially for heterologous gene expression across various host organisms. Existing codon optimization tools often lack the capacity to simultaneously optimize and compare multiple host organisms, highlighting the need for a more comprehensive and scalable solution [13], [32]. Interestingly, there is currently no available tool that integrates both codon optimization and overlapping primer design, streamlining in silico design for de novo gene synthesis.

SOX2, a pivotal transcription factor, assumes a central role in pluripotency by influencing embryonic and tissue stem cells, while also serving as a potent oncogene across diverse cancers. Its functions extend to cellular reprogramming and oncogenesis [89].The importance of SOX2 protein in neural development and stemness creates a requirement for recombinant SOX2 production for various applications, including early embryonic development and patient-specific therapies [70].

Marked overexpression of SOX2 in various cancers underscores its significance in driving cancer cell behaviors, sustaining cancer stem cells (CSCs), and

contributing to drug resistance and relapse [93]. SOX2 expression is mediated through PI3K/AKT signaling pathway. Through this pathway, SOX2 display molecular and functional overlap with tumor suppressor p53. In 2016, Erdem et al. established a significant correlation between SOX2 and TP53 genes in non-small cell lung cancer [75], [88]. Subsequent studies have explored similar correlations across different cancer types. Investigations indicated that the behavior of SOX2 is complex, and tissue type dependent. Even though studies bring a several perspectives in the field, a comprehensive analysis across numerous cancer cohorts remains to explored [71], [83], [120].

Therefore,

- Thesis aims to develop a novel shiny app application that integrates codon optimization and overlapping primer design capabilities. By incorporating multiple bacterial and yeast genomes into the optimization process, the app streamlines the gene synthesis workflow, significantly reducing time and effort for researchers in the field.
    - Simultaneous codon optimization for multiple host genomes
    - Generating dendrogram for evolutionary distance analysis for optimized sequences
    - Providing interactive alignment graph for input and optimized sequences.
    - Designs set of forward and reverse overlapping primers for PCR-based *de novo* gene synthesis, e.g., SOX2.

- Furthermore, thesis aims to investigate the role of the SOX2 protein in various TCGA cancer cohorts and investigate its potential association with the tumor suppressor TP53 through Fisher's exact association test. Through this investigation, the thesis seeks to contribute to a deeper understanding of the molecular mechanisms underlying cancer development and progression, paving the way for targeted interventions and personalized medicine approaches.
- Finally, thesis aims to generate an in-silico design analysis for *de novo* SOX2 gene synthesis by using the Shiny application to be used in various biological applications by significantly reducing time and effort.

3 MATERIALS AND METHODS

3.1 Shiny Application Design and Workflow

Novel shiny application is developed using R programming language and it utilizes the shiny packages. The app offers two main functions: codon optimization and overlapping primer design. For codon optimization, the app employs the GeneGA package [121], which provides R-based genetic algorithms for optimizing gene expression based on mRNA stability and codon usage bias. The package includes codon usage bias metrics, specifically the CAI metrics, for nearly 200 bacterial genomes and a yeast genome. These metrics are available in the wSet dataframe, computed from highly expressed genes predicted using the HEG-DB database.

Within the app, users input the sequence to be optimized and select their desired host organisms. The rate of replacement is adjustable to set the scale for codon optimization. The app displays each optimized sequence in a table and provides a copy button for users to easily copy and use the selected sequences in the primer design tab.

As an additional feature, the app aligns the optimized sequences via msa package [122]. The alignment object is converted into DNAbins and hierarchically clustered to generate a dendrogram. This dendrogram provides insights into the evolutionary distance between the input sequence and the selected host organisms. Moreover, an interactive alignment graph is displayed via msa package [122].

The second tab of the app focuses on primer design to be used in PCR-based gene synthesis protocols [12]. The overlapping primer design tab utilizes a novel algorithm to design a set of overlapping primers for a given sequence. Users can input

parameters such as total primer length, and desired melting temperature. The algorithm generates the primers based on these parameters, and the resulting set is displayed in a table. The table includes the primer sequences, melting temperature, and CG% information. Users can adjust the parameters to achieve optimal results.

Table 3-1. List of R packages used in the association analyses and Shiny app

Names	**Reference**
DT	Xie et al., 2022
coRdon	Elek A et al., 2023
ggmsa	Yu G., 2022
seqmagic	Yu G., 2023
cowplot	Claus O et al., 2020
dplyr	Wickham H et al., 2023
formattable	Kun Ren et al., 2021
GeneGA	Li Z et al., 2015
DECIPHER	Wright Es., 2016
RColorBrewer	Erick Neuwirth., 2022
shiny	Chang et al., 2022
shinythemes	Chang W., 2021
shinydashboard	Chang & Borges, 2021
shinyWidgets	Perrier et al., 2022
rclipboard	Sebastien Bihorel., 2022
plotly	Sievert C., 2020
ggplot2	Wickham H, 2016

kmer	Wilkonson S., 2018
dendextend	Galili T., 2015
ape	Paradis E et al., 2019
msa	Bodenhofer U et al., 2015
msaR	Rauscher B et al., 2021
Biostrings	Pages H et al., 2023
ggstatsplot	Patil I., 2021

3.1.1 Codon Optimization Tab

The Codon Optimization tab in the novel shiny app provides codon optimization for 197 bacterial genomes using the GeneGA package [121]. The required inputs for this tab are the sequence of interest, selection of host organisms, and the selection of a scale value as a fraction between 1 and 0. This tab is divided into three sub-tabs. The first sub-tab displays the codon optimized sequences for multiple organisms simultaneously. The second sub-tab provides a dendrogram for input and optimized sequences for evolutionary distance analysis. Finally in the third sub-tab, an interactive alignment graph is generated to show even the single base differences between the input sequence and the optimized sequences based on selected host organisms.

3.1.1.1 Multiple Codon Optimization

The GeneGA package includes two built-in functions for codon optimization: GeneFoldGA and GeneCodon. In our app, we currently offer codon optimization based solely on codon usage bias using the GeneCodon function.

The GeneCodon function replaces less frequently used codons with the most preferred synonymous codons, optimizing the sequence based on codon usage bias. The rate of replacement can be adjusted using the scale parameter within the function. By default, it is set to 0.5, meaning that each codon will be sampled from its 50% less frequently used synonymous codons. This means that, with the increase in the scale, the optimized sequence is going to be more and more like the input sequence. This concludes that the evolutionary distance is disproportional to the rate of replacement value.

3.1.1.2 Evolutionary Distance Analysis

Evolutionary distance analysis serves as an effective method for comparing multiple DNA sequences. Utilizing the Shiny app, the distance analysis can be conducted for both input and codon-optimized sequences, leading to the generation of dendrograms through hierarchical clustering.

The analysis process involves first alignment of the sequences via msa package, then, converting the sequences into DNA bins, which is achieved using the Biostrings package. Subsequently, the distance is calculated in an alignment-free manner, enabling a robust comparison of the sequences. Finally, the distance matrix is clustered using the built-in hclust function, and the results are visually presented in the form of a dendrogram graph.

3.1.1.3 Alignment Graph

The codon optimization tab offers a final and essential feature: the interactive alignment graph. This functionality is specifically designed to enable users to observe even single nucleotide differences between the input and codon-optimized sequences.

By giving full access to the sequences, researchers can easily identify potential promoter or enzyme sites before proceeding with PCR and cloning experiments. This feature utilizes the msa and msaR packages to effectively align and display the sequences in an interactive table.

The msa package seamlessly integrates various command-line algorithms, such as ClustalW, ClustalOmega, and MUSCLE, for multiple sequence alignment [122]. This ensures accurate and comprehensive alignment of the sequences.

Within the app, the alignment graph enables users with the ability to filter, edit, and export the alignment as necessary, providing a highly flexible and user-friendly tool for sequence analysis and preparation.

3.1.2 Overlapping Primer Design Tab

De novo gene synthesis is commonly accomplished through PCR-based methods, which involve the use of a set of overlapping forward and reverse primers. These overlapping regions are designed to anneal under proper PCR conditions.

For successful PCR assembly, several parameters need to be optimized, including the total primer length, the melting temperature, and the length of the overlapping regions. Shiny tool utilizes a novel R function that generates a series of overlapping primers with synchronized melting temperatures for a given DNA sequence.

This recursive function takes the DNA sequence as a string array and sub-sets it based on the desired total primer length, the length of the overlapping region, and the specified melting temperature for the overlapping region. By doing so, the function generates forward and reverse primers with melting temperatures similar to the previous primer in each cycle.

The user provides the total primer length (l) as input, and the function considers a range of (l-5, l+5) to accommodate slight variations. The melting temperature (t) is also user-defined, and the function ensures that the melting temperatures of each overlapping region fall within a range of (t-2, t+2). Finally, the length of the overlapping region is fixed, ranging between 15 to 25 nucleotides.

The primer design function outputs a dataframe that includes the primer sequence, directionality (reverse or forward), melting temperature, and the CG percentage of the overlapping regions.

Additionally, the primer design function utilizes the TmCalculator package to calculate the melting temperature and the CG percentage. This package originally extended from Bio.SeqUtils.MeltingTemp of python. It uses The Wallace rule which is a primary method to calculate the melting temperature (Tm) of primers with 14nt to 25 nt length [16].

3.2 The Association Analysis Between SOX2 Copy Number Variation and TP53 Mutation in Various TCGA Cancer Cohorts

3.2.1 Obtaining Data via cBioPortal

The data for analyzing the association between SOX2 copy number variation and TP53 mutation in various cancer cohorts have been collected using cBioPortal. This platform offers a gene-specific search option, allowing us to select and query cohorts for SOX2 and TP53 proteins.

After selecting the appropriate cohorts, the data was queried and plotted using the Plots tab, which provides visualizations for different data types of the genes, including clinical attributes, mutations, copy number alterations, and more. To facilitate further analysis, both the plots and the curated data used for generating them can be downloaded by users. This functionality is accessible through the download button located at the top right corner of each plot.

In line with the hypothesis of this thesis, we focused on plotting the SOX2 copy number variation data against TP53 mutation data for the selected cancer cohorts. Subsequently, the curated data was downloaded for in-depth analysis.

Table 3-2. The list of cancer cohorts that have been selected for the analysis through cBioPortal

Cancer Type	**Institute/Publication**	**Sample Size**
Esophageal Carcinoma	TCGA, PanCancer Atlas	559
Glioblastoma	TCGA, Cell	577
Liver Hepatocellular Carcinoma	TCGA, PanCancer Atlas	372
Lung Adenocarcinoma	TCGA, PanCancer Atlas	503
Pan-Lung Cancer	TCGA, Nat Genet	1144
Thyroid Carcinoma	TCGA, PanCancer Atlas	500
Breast Invasive Carcinoma	TCGA, PanCancer Atlas	1084
Ovarian Serous Cytadenocarcinoma	TCGA, PanCancer Atlas	585
Brain Lower Grade Glioma	TCGA, PanCancer Atlas	514

3.2.2 Fisher's Exact Test in R programming Environment

To investigate the potential association between SOX2 copy number variations and TP53 mutations, a Fisher's Exact test was conducted using the R programming environment. The data was obtained from cBioPortal and consists of three columns. The first column contains patient IDs, the second column indicates the putative copy number alterations of SOX2 (including amplification, deep deletion, diploid, gain, and shallow deletion), and the last column indicates the TP53 mutations for each corresponding patient (including inframe, missense, multiple, no mutation, not profiled, splice, and truncated).

Next, a contingency table was generated to explore the hypothesized association between SOX2 gain/amplification and TP53 mutation. The values in the contingency table represent the number of patients where both SOX2 gain/amplification and TP53 mutation are observed, the number of patients where neither of these conditions are observed, and two other scenarios where either one of these conditions is present.

The objective of this analysis is to examine whether there is a significant association between SOX2 gain/amplification and TP53 mutation, and to determine if these two genetic alterations occur together more frequently than expected by chance.

3.3 Recombinant SOX2 DESIGN

In silico design is generated for recombinant SOX2 protein via novel shiny app. Literature search indicates that Sox2 is recombinantly expressed in numerous organisms. *Escherichia coli* is most commonly used bacterial host organism for recombinant protein production since it is extensively studied and known [70], [91],

[114]. Beside E.coli, *Bacillus subtillis*, *Pseudomonas fluorescens*, yeast, insect and mammalian cells are used as a host organism [5], [123]. For this analysis, the organisms found in the GeneGA package dataset is curated based on the usage from the literature and following organisms were used for codon optimization: *Saccharomyces cerevisiae*, *Bacillus subtilis*, *Corynebacterium glutamicum* ATCC, *Lactococcus lactis*, *Pseudomonas fluorescens.* Additionally different *Escherichia coli* strains were used as the following: as K12, CFT073, UTI98, W3110, 536.

The parameters for overlapping primer design are set based on the existing studies that perform PCR assembly with overlapping primer set [12], [14], [17]. The total primer length set to 60 bp and the overlapping region ranges between 15 to 25 bp. The melting temperature set to accommodate 57°C.

4 RESULTS

4.1 SOX2 Copy Number Variation TP53 Mutation Analysis

The TCGA cancer cohorts used in the analysis are listed in ***Table 3-2***. The data has been analyzed using Fisher's exact test. SOX2 gain/amplification and TP53 mutation have been treated as two categorical variables. Each cohort dataset is curated, and contingency tables are generated based on these variables. A p-value of less than 0.005 was considered statistically significant in this analysis.

Results indicate that the hypothesized association between SOX2 and TP53 is observed in the following cancer cohorts: Liver Hepatocellular Carcinoma, Lung Adenocarcinoma, Breast Invasive Carcinoma, and Thyroid Carcinoma. For more detailed analysis, including cohort information and statistical results, please refer to Appendix.

4.1.1 Liver Hepatocellular Carcinoma

The Liver Hepatocellular Carcinoma cohort comprises genomic information from 372 samples. According to the cBioPortal analysis, TP53 stands out as the most frequently mutated gene in this cohort, accounting for a ratio of 30.1%. The cohort was tested for hypothesized association. The corresponding contingency table and relative frequency table can be found in ***Table 4-1*** and ***Table 4-2***, respectively. The results of the Fisher's exact test demonstrate a significant association between SOX2 gain/amplification and TP53 mutation within this cohort, yielding a p-value of 0.0001631 (**Figure 4.1**). The two-way contingency graph and Fisher's exact test outcomes are plotted using the R

programming environment. Additionally, a pie chart is generated to assess the ratio of TP53 mutation types in the presence of SOX2 gain and amplification.

The analysis reveals that missense and truncating mutations constitute the two most frequent mutation types, accounting for a percentage of 46.88% each. Additionally, splice mutations are observed in the cohort, making up 6.25% of the cases.

Table 4-1. Liver Hepatocellular Carcinoma (TCGA) contingency table.

	TP53 Mutation (-)	**TP53 Mutation (+)**
SOX2 CNV (-)	183	28
SOX2 CNV (+)	66	32

Table 4-2. Liver Hepatocellular Carcinoma (TCGA) relative frequency table.

	TP53 Mutation (-)	**TP53 Mutation (+)**
SOX2 CNV (-)	0.5922330	0.09061489
SOX2 CNV (+)	0.2135922	0.10355987

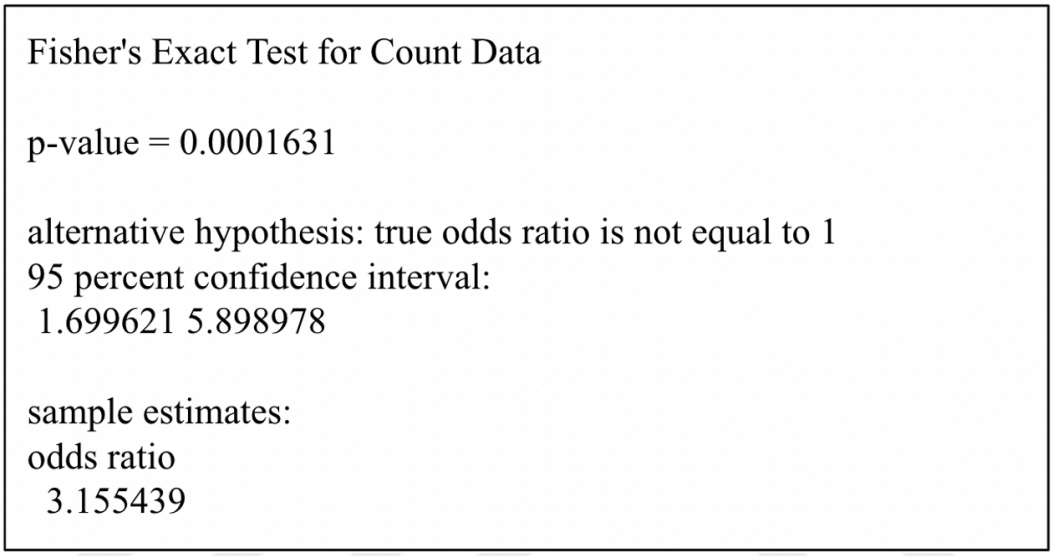


Figure 4.1 Fisher's exact test results for Liver Hepatocellular Carcinoma cohort

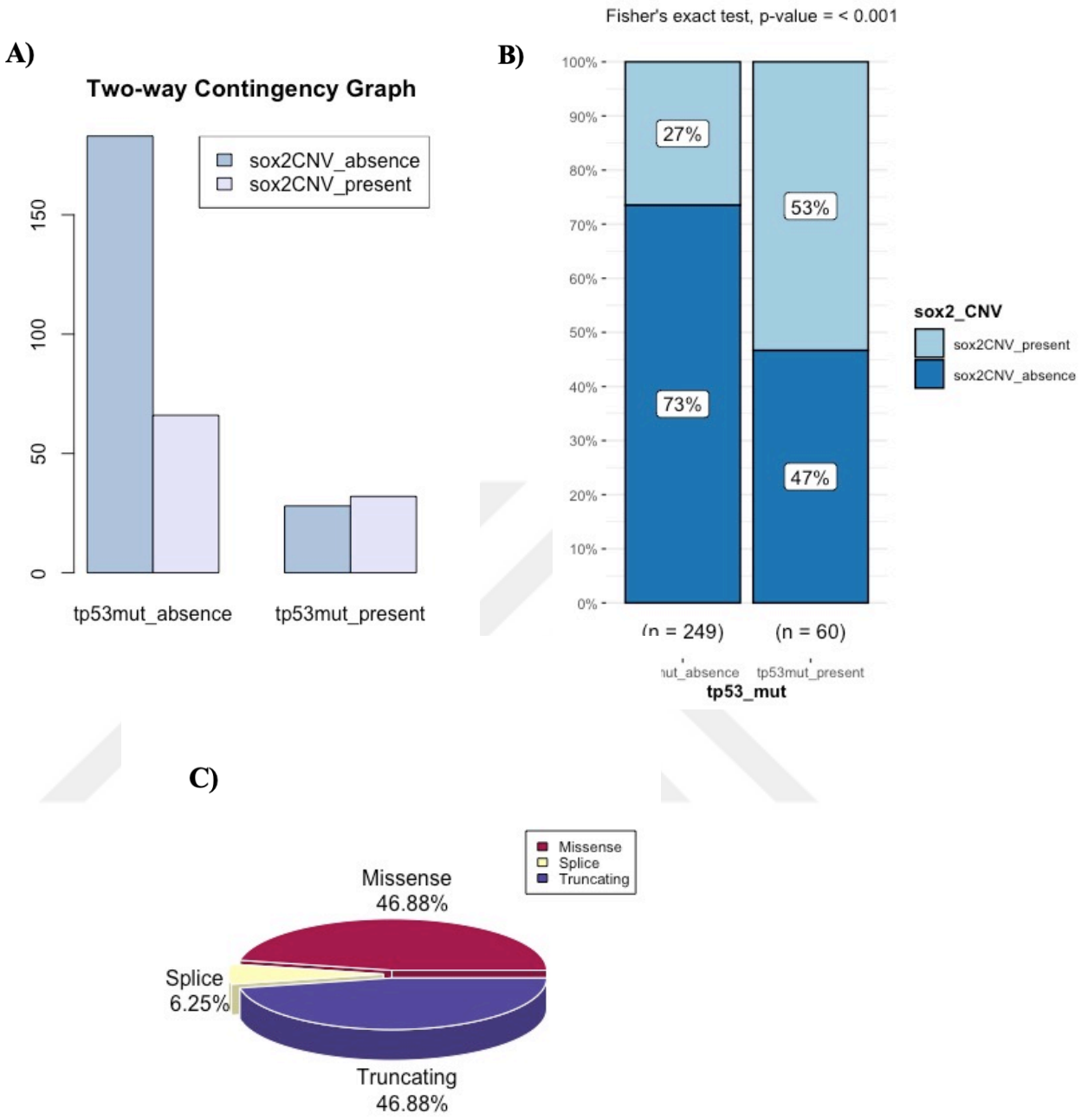


Figure 4.2 The bar plot of two-way contingency graph (A) and the bar plot of Fisher's exact test results (B) for Liver Hepatocellular Carcinoma cohort. (C) Pie chart shows the percentage of observed TP53 mutation types.

4.1.2 Lung Adenocarcinoma

The TCGA Lung Adenocarcinoma cohort comprises 503 samples. Within this cohort, TP53 is the most frequently mutated gene, exhibiting a frequency of 52.1%. The most common copy number alterations are identified in the CDKN2A and CDKN2B genes, with ratios of 16.8% and 16.4%, respectively. The Fisher's exact test revealed that there is a significant association between SOX2 gain/amplification and TP53 mutation in this cancer cohort ($p = 3.166e-09$).

Table 4-3 Lung Adenocarcinoma (TCGA) contingency table.

	TP53 Mutation (-)	**TP53 Mutation (+)**
SOX2 CNV (-)	133	35
SOX2 CNV (+)	106	107

Table 4-4 Lung Adenocarcinoma (TCGA) relative frequency table.

	TP53 Mutation (-)	**TP53 Mutation (+)**
SOX2 CNV (-)	0.3490814	0.09186352
SOX2 CNV (+)	0.2782152	0.28083990

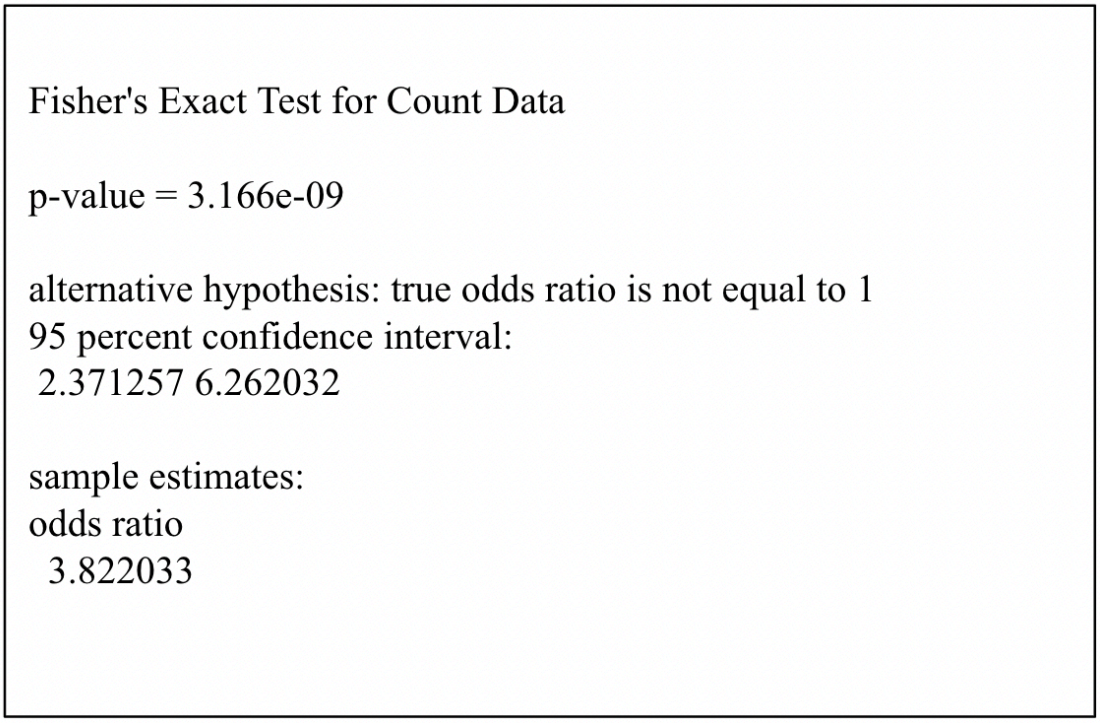


Figure 4.3 Fisher's exact test results for LUAD (TCGA)

When considering the presence of SOX2 gain and amplification, the observed mutation types for TP53 include missense, splice, multiple, truncating, and inframe mutations. Among these, the most frequently observed mutation type in this cohort is the missense mutation, accounting for 65.42% of cases. Truncating mutations come in second place, comprising 19.63% of occurrences. Following them, splice, multiple, and inframe mutations are observed with ratios of 8.41%, 4.67%, and 1.87%, respectively.

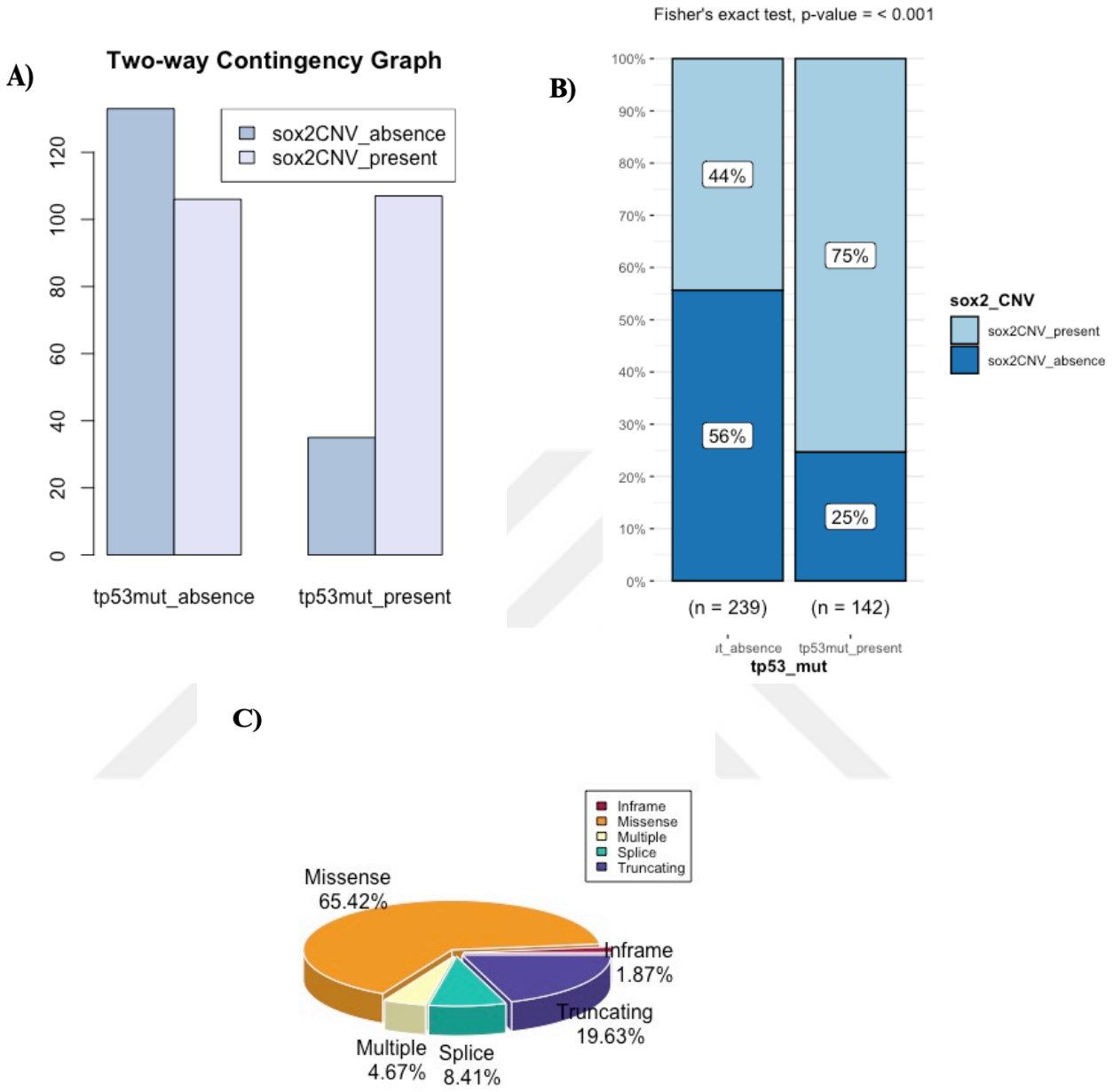


Figure 4.4 Two-way contingency graph (A) and Fisher's exact test results (B) are depicted. The pie chart represents the observed TP53 mutation types in the presence of SOX2 gain/amplification (C).

Given the notable significance of the association between SOX2 and TP53 in lung cancer, an additional cancer cohort was examined to validate this hypothesis. The second cohort was curated from the study conducted by Campbell et al. in 2016. Within this study, exome sequences and copy number profiles of 660 lung adenocarcinoma and 484 lung squamous cell carcinoma (SqCC) tumor-normal pairs were analyzed. The sample size of this cohort is 1144.

Furthermore, the Fisher's exact test provides additional confirmation of a substantial association between SOX2 and TP53 within the context of lung cancer ($p = 2.2e-16$).

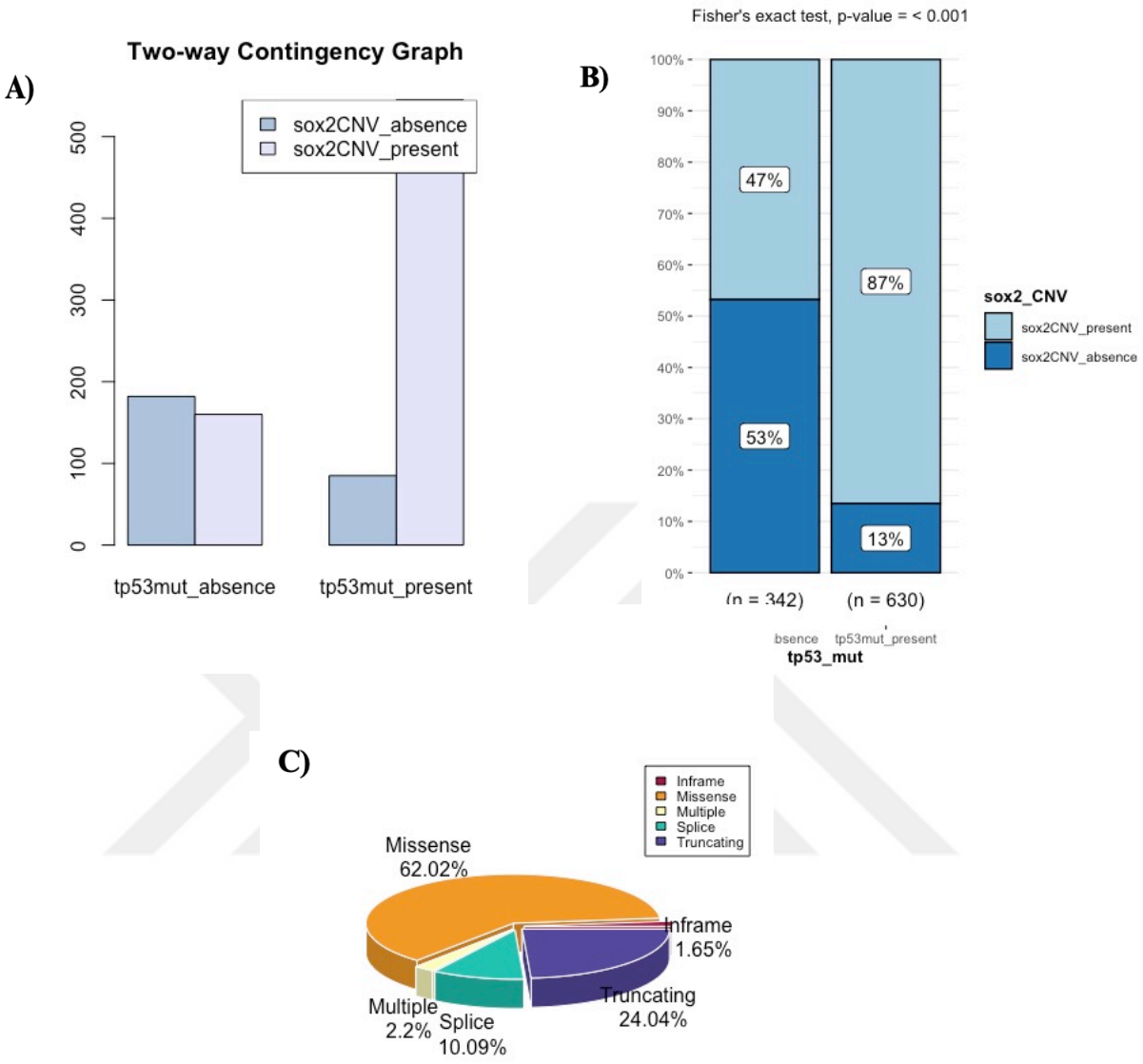


Figure 4.5 Pan Lung Cancer Cohort Analysis. Two-way contingency graph (A) and Fisher's exact test results (B) are depicted. The pie chart represents the observed TP53 mutation types in the presence of SOX2 gain/amplification.

4.1.3 Thyroid Carcinoma

The TCGA Thyroid Carcinoma cohort comprises 500 samples. Within this cohort, BRAF is the most frequently mutated gene , accounting for a ratio of 58.6%. Notably, TP53 mutations are infrequent in this cohort, constituting only 0.4% of cases. The statistical analysis reveals a significant association between TP53 mutation and SOX2 gain/amplification (p-value = 0.0009216).

Table 4-5 Thyroid Carcinoma (TCGA) contingency table.

	TP53 Mutation (-)	**TP53 Mutation (+)**
SOX2 CNV (-)	686	34
SOX2 CNV (+)	67	12

Table 4- Thyroid Carcinoma (TCGA) relative frequency table.

	TP53 Mutation (-)	**TP53 Mutation (+)**
SOX2 CNV (-)	0.85857322	0.04255319
SOX2 CNV (+)	0.08385482	0.01501877

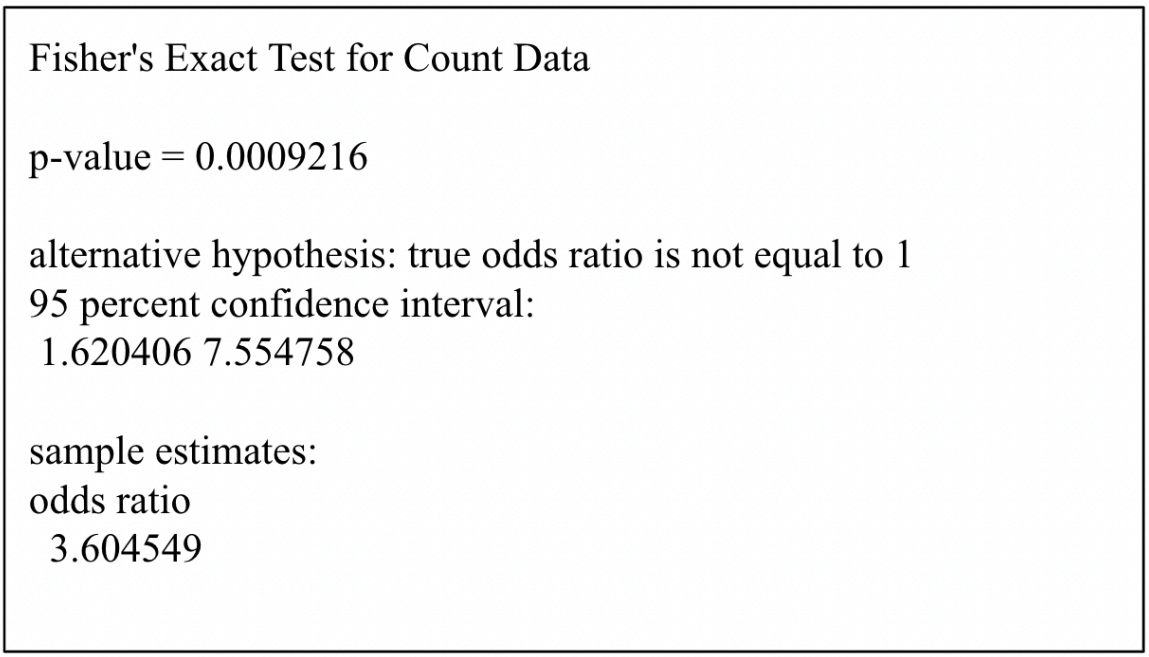


Figure 4.6 Fisher's exact test results for thyroid carcinoma cohort.

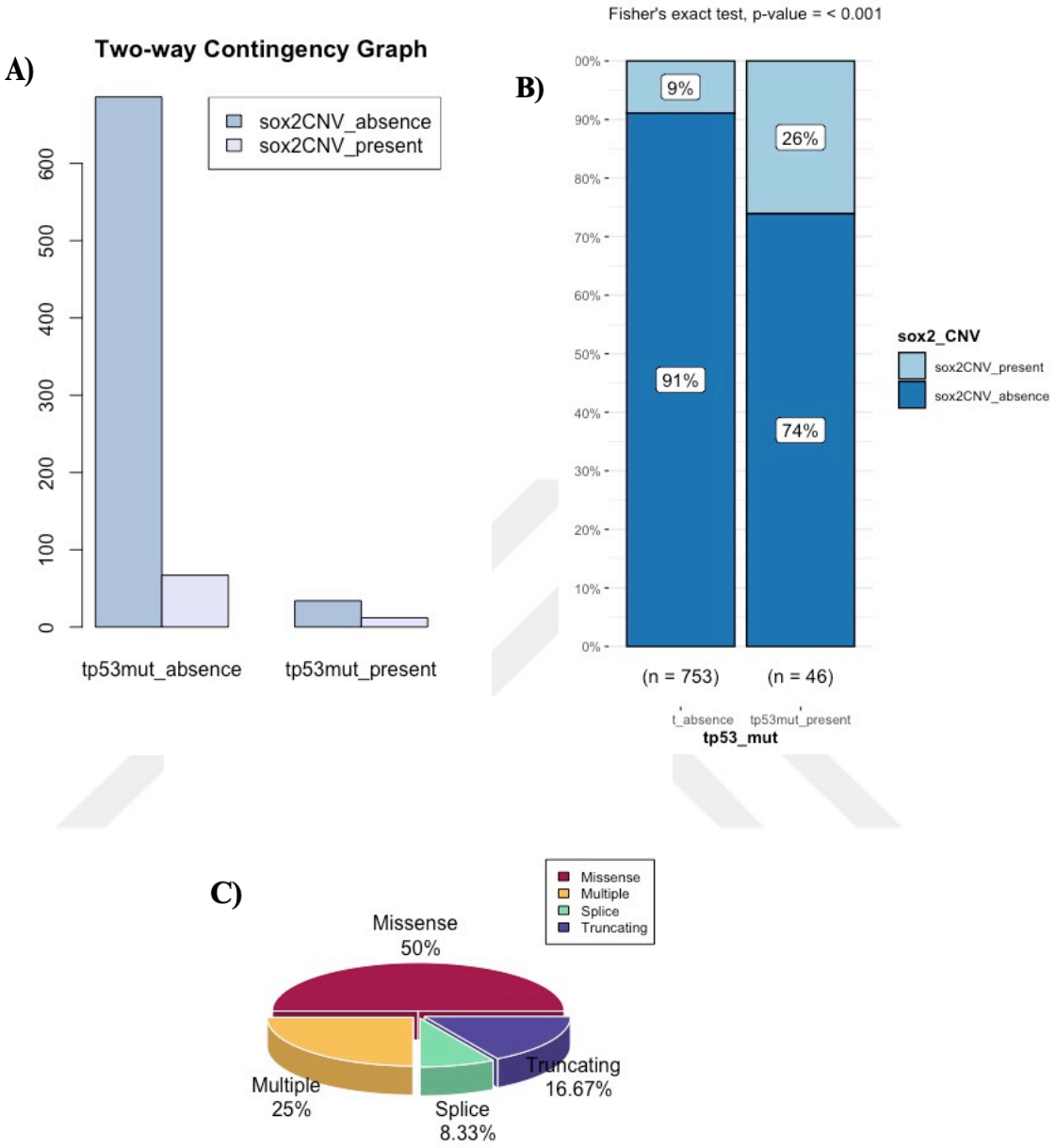


Figure 4.7 Thyroid Carcinoma Cohort Analysis. Two-way contingency graph (A) and Fisher's exact test results (B) are depicted. The pie chart represents the observed TP53 mutation types in the presence of SOX2 gain/amplification.

Consistent with the patterns observed in previous cancer cohorts, the Thyroid Carcinoma cohort similarly underscores that missense mutations represent the most common mutation type within the TP53 gene when considering the presence of SOX2 gain and amplification. Notably, approximately half of all mutations falling within this category are of the missense mutation type. (**Figure 4.7**)

4.1.4 Breast Invasive Carcinoma

The dataset for Breast Invasive Carcinoma contains genetic data from 1084 samples. According to the analysis performed using cBioPortal, the most frequently mutated genes within this cohort are PIK3CA and TP53, both accounting for a proportion of 32.6% each. The outcomes of the Fisher's exact test confirm a significant association between the presence of SOX2 gain/amplification and TP53 mutation in this cohort. The resulting p-value is 2.2e-16 (**Figure 4.8**).

Table 4-7 Breast Invasive Carcinoma (TCGA) contingency table.

	TP53 Mutation (-)	**TP53 Mutation (+)**
SOX2 CNV (-)	496	117
SOX2 CNV (+)	131	185

Table 4-8 Breast Invasive Carcinoma (TCGA) relative frequency table.

	TP53 Mutation (-)	**TP53 Mutation (+)**
SOX2 CNV (-)	0.5339074	0.1259419
SOX2 CNV (+)	0.1410118	0.1991389

```
Fisher's Exact Test for Count Data

p-value < 2.2e-16

alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 4.379576 8.183344

sample estimates:
odds ratio
  5.973056
```

Figure 4.8 Fisher's exact test results for breast invasive carcinoma cohort.

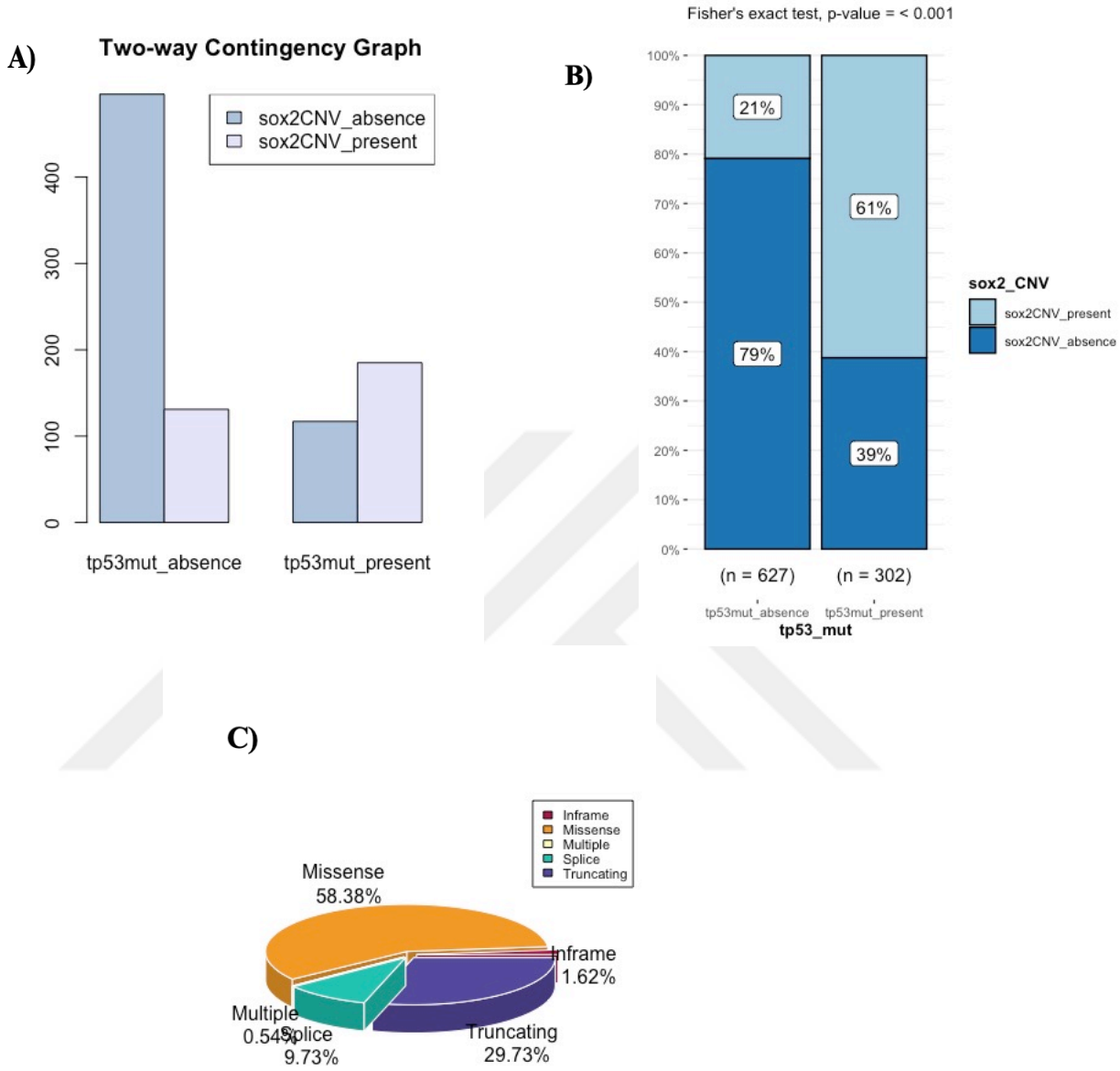


Figure 4.9 Breast Invasive Carcinoma Cohort Analysis. Two-way contingency graph (A) and Fisher's exact test results (B) are depicted. The pie chart represents the observed TP53 mutation types in the presence of SOX2 gain/amplification.

4.1.5 The Significance of p53 Missense Mutation

According to the results stated above, correlated results have mostly missense mutation in their p53 gene. In order to determine the statistical significance of these results, both Chi-square and Fisher's exact test was employed. A matrix and a contingency graph were created respectively to be used in statistical tests. Test results indicate that the cohorts that showed significant association between TP53 mutation and SOX2 gain and amplification did not display any association between TP53 missense mutation or in any type of p53 mutation and SOX2 gain/amplification (p-value < 0.05). Detailed test results can be found in the **Appendix Table 2**.

4.2 Shiny Application

4.2.1 Codon Optimization Tab

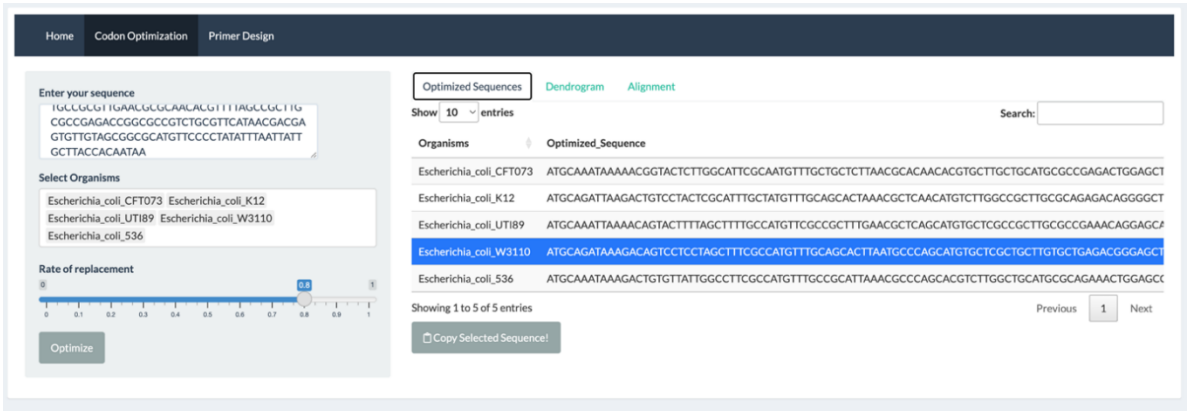
The Codon Optimization tab includes three primary functions. Each tab is accessible through a dedicated sub-tab within the application. These functions include multiple codon optimization, evolutionary distance analysis, and alignment graph (**Figure 4.8**).

On the left-hand side of the page, a sidebar panel is situated. Within this panel, users have the option to input their sequence of interest into a designated input box. Then, users can select host organisms of interest from a multiple selection box. This selection box offers a compilation of yeast and bacterial organisms drawn from the wSet dataset, a built-in dataset of the GeneGA package[121]. GeneGA package allows users to adjust the rate of codon replacement for codon optimization. This scale value can be adjusted through a slider input box on the left side of the panel.

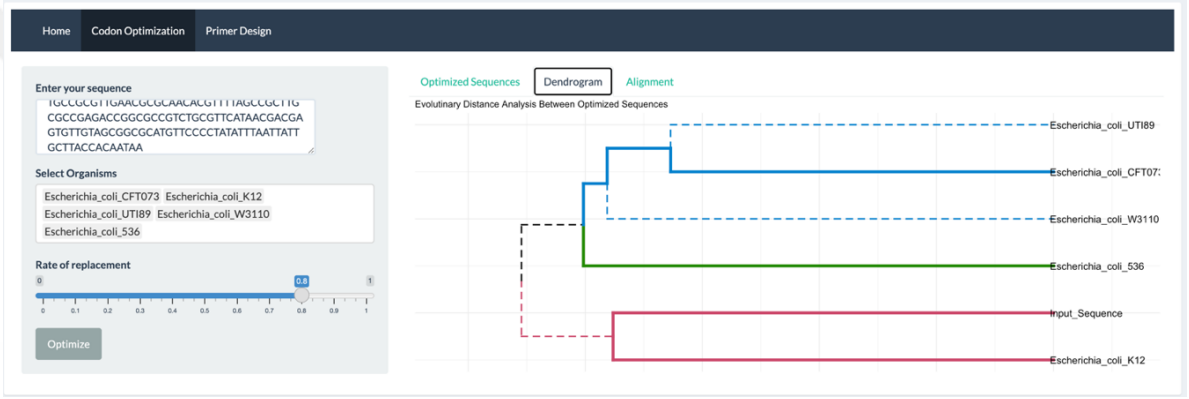
This value can be selected within the range of 0 to 1. In the "Dendrogram" sub-tab, a dendrogram plot is shown for hierarchically clsutered input and optimized sequences. This function can help researchers to compare different host organism's codon bias to assess the compability of the sequence that is going to be recombinantly expressed. Last but not least, "Alignment" sub-tab generates an interactive alignment graph for input and codon optimized sequences. This feature allows users to get base by base comparison between the sequences. Msa package allows users to edit, filter, and export the alignment graph inside the app. Additionally color scheme can be adjusted through intereactive alignment graph[122].

Upon activating the action button, all three functions provide simultaneous output analyses. The optimized sequences are presented in tabular format within the "Optimized Sequences" sub-tab. For ease of use, the selected sequence can be copied using the "Copy Selected Sequence!" button located under the tabular output.

A)



B)



C)

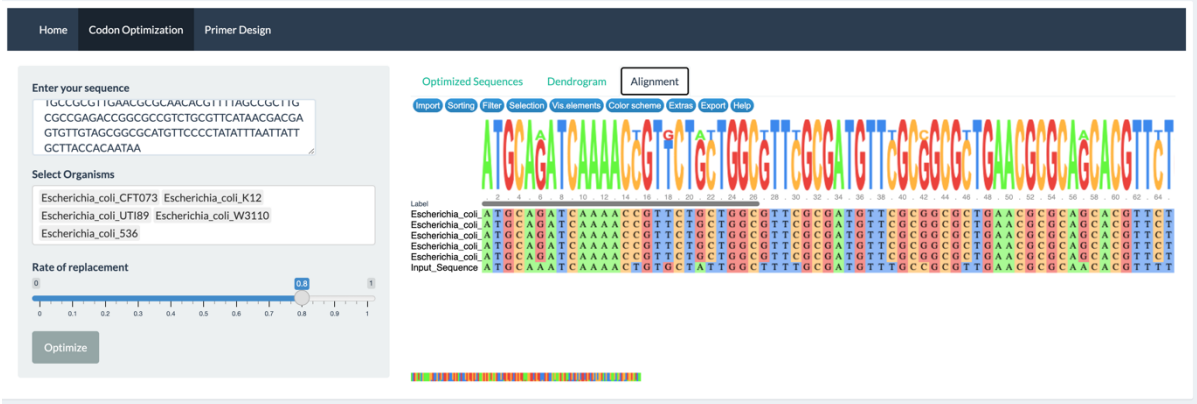


Figure 4.10 Optimized sequence subtab displays a table of codon optimized sequences (A). Dendogram subtab clusters optimzied and input sequences and display a dendrogram for evolutionary distance analysis (B). The alignment subtab displays an interactive alignment graph for optimzied and input sequences (C).

A secondary case study is conducted to test the accuracy the rate of replacement value, scale parameter, for the codon optimization step. For this test the FASTA sequence of human insulin A chain is used. Obtained sequence is codon optimized using GeneCodon function for Escherichia coli K12 strain with different rate of replacement values. The scale parameters were 0.25, 0.50, 0,75, and 1 respectively. According to the working mechanism of the rate of replacement, the distance between input sequence and optimized sequence should be disproportinate to the rate of raplecment value[121]. Optimized sequences hirerchially clustered and the results further confirmed the hypothesis. **Figure 4.11** depicts that the distance between human insulin A sequence and optimized sequence increases as the scale decreases.

Finally, “Alignment” subtab generates an interactive alignment graph for input and codon optimized sequences. This feature allows users to get base by base comparison between the sequences. The alignment grapgh can be editted and exported via msa package[122]. The maximum capacity of the codon optimization was tested and app could generate codon optimized sequences up to 35 organisms under 30 seconds.

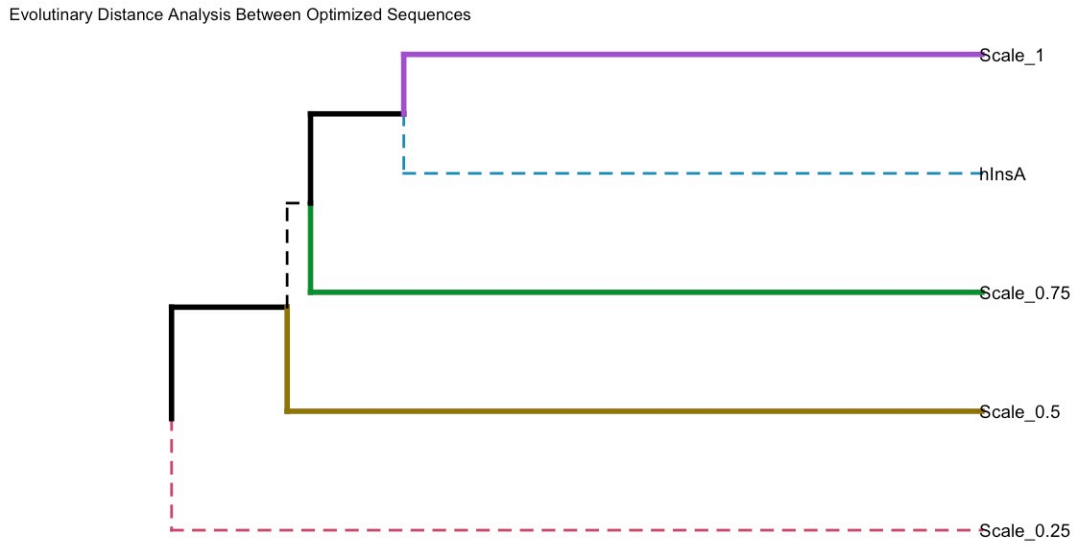


Figure 4.11 Human Insulin A chain is codon optimized for Escherichia coli K12 strain with scale of 1, 0.75, 0.5, and 0.25.

4.2.2 Primer Design Tab

The "Primer Design" tab employs a novel algorithm to generate a set of overlapping forward and reverse primers tailored for PCR-based de novo gene synthesis. For comprehensive details, please refer to **Figure 1.1**. This function's primary objective is to produce primers with consistent melting temperatures, ensuring proper functioning of PCR reactions.

On the left side of the page, a sidebar panel takes user input. Users can input their sequence of interest into a designated string input box. Additionally, the sidebar panel takes inputs for the total primer length and the desired melting temperature.

The results are presented in a tabular format. The result table provides information including primer sequences, directionality (forward [F] or reverse [R]), GC ratio, and the melting temperature of the overlapping regions (**Figure 4.12**).

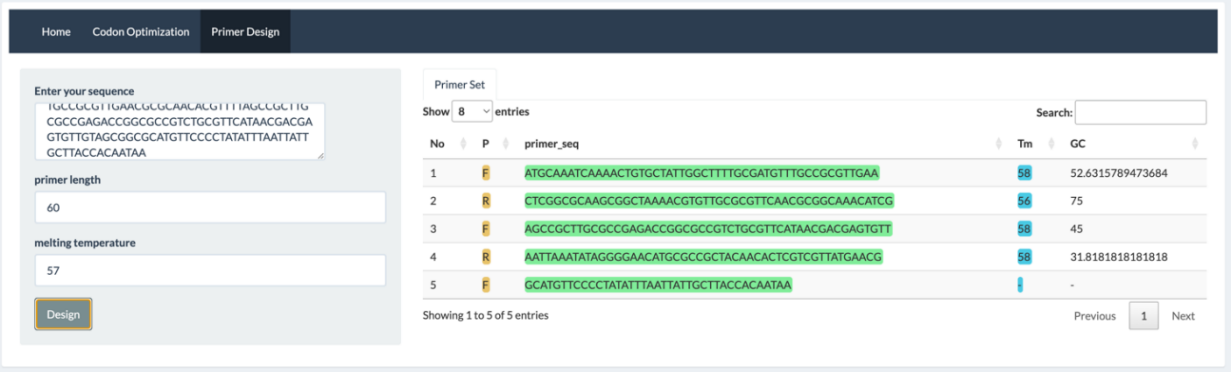


Figure 4.12 The primer design tab displays designed primers with melting temperature and GC %.

4.3 In Silico Design for Recombinant SOX2

In this analysis, FASTA sequence of Human SOX2 gene (NC_007133.7) was utilized. The sequence is obtained via NCBI database (https://www.ncbi.nlm.nih.gov). SOX2 gene is codon optimized through novel Shiny app for different bacterial and yeast host organisms. The compability of the organism is assessed in the aspect of codon bias differences through dendrogram graphs. SOX2 gene sequence codon optimized for different *Escherichia coli* strains which are labeled as K12, CFT073, UTI98, W3110, 536. Then, the most similar strain, UTI89 clustered with other commonly used host organisms: *Saccharomyces cerevisiae*, *Bacillus subtilis*, *Corynebacterium glutamicum* ATCC, *Lactococcus lactis*, *Pseudomonas fluorescens.*

Figure 4.13 Dendrogram plot for codon optimized human SOX2 for different E coli strains

Figure 4.14 Dendrogram plot for codon optimized human SOX2 for yeast and bacterial host organisms.

Based on the distance between the sequences, *Escherichia coli* UTI89 strain appears to be most compatible with original SOX2 sequence based on their codon bias information. Therefore UTI89 strain selected for the second step. Overlapping primers were designed for UTI89 codon optimized SOX2 sequence through novel Shiny app. Overlapping primers were generated by setting total primer length to 60bp and melting temperature to 57°C. Primers are listed below (***Table 4-9***).

Table 4-9 Overlapping primers for recombinant human SOX2 gene with melting temperatur values.

Dir	Primer Sequence	Tm°C
F1	ATGTACAACATGATGGAGACGGAGCTGAAGCCGCCGGGCCCGCAGCAAAC	56
R1	CGCCGCGGTGGAGTTGCCGCCGCCGCCCCCCGAAGTTTGCTGCGGGCCCG	56
F2	CAACTCCACCGCGGCGGCGGCCGGCGGCAACCAGAAAAACAGCCCGGACC	56
R2	CACCATGAAGGCATTCATGGGCCGCTTGACGCGGTCCGGGCTGTTTTTCT	56
F3	CATGAATGCCTTCATGGTGTGGTCCCGCGGGCAGCGGCGCAAGATGGCCC	56
R3	CTGATCTCCGAGTTGTGCATCTTGGGGTTCTCCTGGGCCATCTTGCGCCG	56
F4	GCACAACTCGGAGATCAGCAAGCGCCTGGGCGCCGAGTGGAAACTTTTGT	56
R4	TCGTCGATGAACGGCCGCTTCTCCGTCTCCGACAAAAGTTTCCACTCGGC	56
F5	CGGCCGTTCATCGACGAGGCTAAGCGGCTGCGAGCGCTGCACATGAAGGA	56
R5	TCCGCCGGGGCCGGTATTTATAATCCGGGTGCTCCTTCATGTGCAGCGCT	56
F6	ACCGGCCCCGGCGGAAAACCAAGACGCTCATGAAGAAGGATAAGTACACG	56
R6	GCCGCCGGGGGCCAGCAGCCCGCCGGGCAGCGTGTACTTATCCTTCTTCA	58
F7	CTGGCCCCCGGCGGCAATAGCATGGCGAGCGGGGTCGGGGTGGGCGCCGG	58
R7	AACTGTCCATGCGCTGGTTCACGCCCGCGCCCAGGCCGGCGCCCACCCCG	56

F8	ACCAGCGCATGGACAGTTACGCGCACATGAACGGCTGGAGCAACGGCAGC	58
R8	CGGGTAGCCCAGCTGGTCCTGCATCATGCTGTAGCTGCCGTTGCTCCAGC	56
F9	CCAGCTGGGCTACCCGCAGCACCCGGGCCTCAATGCGCACGGCGCAGCGC	56
R9	GGCGCTCACGTCGTAGCGGTGCATGGGCTGCATCTGCGCTGCGCCGTGCG	58
F10	GCTACGACGTGAGCGCCCTGCAGTACAACTCCATGACCAGCTCGCAGACC	56
R10	GGACATGCTGTAGGTGGGCGAGCCGTTCATGTAGGTCTGCGAGCTGGTCA	58
F11	CCCACCTACAGCATGTCCTACTCGCAGCAGGGCACCCCTGGCATGGCTCT	56
R11	GCCTCGGACTTGACCACCGAACCCATGGAGCCAAGAGCCATGCCAGGGGT	56
F12	GTGGTCAAGTCCGAGGCCAGCTCCAGCCCCCCTGTGGTTACCTCTTCCTC	58
R12	GTCCCCGGCCTGGCAGGGCGCCCTGGAGTGGGAGGAAGAGGTAACCACAG	58
F13	CTGCCAGGCCGGGGACCTCCGGGACATGATCAGCATGTATCTCCCCGGCG	56
R13	GTCTGCTGGGGGCGGCGGGTTCCGGCACCTCGGCGCCGGGGAGATACATG	58
F14	GCCGCCCCCAGCAGACTTCACATGTCCCAGCACTACCAGAGCGGCCCGGT	56
R14	CATGTGTGAGAGGGGCAGTGTGCCGTTAATGGCCGTGCCGGGCACCGGGCCGCTCTGGT	56

5 CONCLUSION, DISCUSSION, AND FUTURE PERSPECTIVES

This thesis work introduces a novel Shiny app, GeneCraft, that streamlines the *de novo* gene and oligo synthesis processes within a single tool. Despite the existence of other tools that are listed earlier in the thesis (**Table 1-1,2**), this novel app combines two crucial stages of *de novo* gene synthesis: 1) codon optimization and 2) PCR-based oligo synthesis [12], [28][13]. GeneCraft allows users to codon-optimize their sequences for nearly 200 bacterial genome and yeast, *Saccharomyces cerevisiae*. Additionally, GeneCraft facilitates the design of overlapping primers for PCR-based *de novo* gene synthesis, while ensuring synchronized melting temperatures for a successful PCR reaction.

Furthermore, the thesis investigates one of the most cloned and transfected/transduced genes, SOX2, whose amplification is found important in many cancers [69], [102]. Indeed, the significant role of SOX2 protein in various cancer types has been interrogated in combination with occurrence of mutation in TP53 [71], [120], one of the most mutated genes in TCGA cancer cohorts . TP53/p53 gene plays an important role in cancer as a tumor-suppressor [72]. This type of correlation analysis [71], [83], [124] is proven to be important in cancer research since genomic instability is a hallmark for many cancer types [125].

Lastly, the present thesis leverages the innovative Shiny app to illustrate a case study for its more in-depth analysis , while generating in silico design for recombinant human SOX2 production through de novo gene synthesis.

5.1 SOX2 CNV TP53 Mutation Analysis

The review of existing literature reveals a direct or indirect association between presence of p53 mutations and SOX2 copy number variation in the same patient across various cancer cohorts [22], [23], [38], [39]. In this thesis, a comprehensive statistical analysis of the SOX2-p53 relationship across TCGA cancer cohorts has been conducted using Fisher's exact test. The outcomes of this analysis affirm the primary findings from the literature, demonstrating varying degrees of association or no association in different cancer types.

The common challenge faced during the analysis was the inadequate sample size in certain cohorts, which has already been shown that it is limiting the statistical robustness of the conclusions [39]. Moreover, cohorts that have been curated from different populations may cause populational bias and lead to different results for different populations [126]. To provide additional insights, a detailed analysis table was included in the supplementary file.

Additionally, I analyzed whether a specific type of TP53 mutation associates with SOX2 gain or amplification more than other cancer types. First missense mutation is analyzed for cohorts that show significant association between TP53 mutation and SOX2 gain/amplification. The contingency graph is generated, and fisher's exact test is conducted. None of the cohort shows significant association between SOX2 gain/amplification and missense mutation. As a future perspective, other mutation type should be analyzed. For the fisher's exact test results please refer to **Appendix Table 2**.

The one thing to consider is that SOX2 is located on 3q chromosome alongside with PIK3CA, ACK1, PRKCI, TP63, PLD1, ECT2, and other genes. Broad 3q chromosome amplification is widely recognized as the most common chromosomal aberration found in various cancer types [120], [124][83]. Genomic copy number alterations contribute to genomic instability which lead to carcinogenesis and chemotherapeutic resistance [127], [128], [129].

Genome stability is essential for cellular function, thus, controlled by the mitotic checkpoint, DNA damage checkpoint and repair machinery [130]. Genomic instability is defined as high rate of acquiring genome alteration during cell cycle process and excepted as a hallmark of human cancer. In hereditary cancers, genomic instability occurs via mutated DNA repair genes. These genes increase the spontaneous mutation rate in the precancerous tissue and contributes to genomic instability. In non-hereditary cancer on the other hand, different mechanisms take place[125]. Next-generation DNA sequencing studies have shown that mutation in DNA repair genes is less frequent in non-hereditary (sporadic) cancers. Thus, a new hypothesis called “oncogene-induced DNA replication stress model” gained popularity in recent years[128]. This hypothesis frequently includes important genes for DNA damage response such as ataxia telangiectasia mutated (ATM) and tumor suppressor gene TP53. TP53, which encodes p53 is a tumor suppressor gene and found amongst the most frequently mutated gene in numerous cancer types [77], [125], [128]. According to the hypothesis, oncogene induces DNA replication stress and double strange breakage (DSB). These effects cause activation of stress signal molecules such as ataxia telangiectasia mutated (ATM). ATM activation disrupts the

interaction between p53 and E3-type ubiquitin ligase MDM2 which is an oncogene negatively regulates p53. The dysregulation in p53-MDM2 pathway is frequently seen in various human cancers [60], [77], [128], [129]. Due to its essential role in various cellular process, the components of p53-MDM2 pathway are often targeted for cancer research [60]. As an example, p53-MDM2 pathway functionally overlap with stem-cell markers such as SOX2 on the PI3K/AKT signaling pathway [75], [88] which regulates stemness, cell proliferation, and survival [131]. Therefore, the mechanism behind the PI3K/AKT signaling pathway can provide insight for new therapeutic approaches for human cancer [132].

As an example, Erdem et al., 2016 investigates whether TP53 mutations affect SOX2 copy number alterations in early-stage of non-small cell lung cancer (NSCLC) patients. Results indicate that there is significant correlation between TP3 mutations and increased rate of SOX2 copy number alterations ($p = 0.017$). Furthermore, study silenced TP53 expression via siRNA treatment and this resulted in decrease in SOX2 and has-miR-145 expression in lung adenocarcinoma cells [71] . As has-miR-145 was previously identified as regulator for SOX2 expression which promotes further investigation to understand the mechanism behind the TP53 and SOX2 association and identify potential targets of the system .

Another study from 2019 investigates the association between p53 protein phosphorylated at serine 20 (p-p53(Ser20)) and SOX2, CD133, Notch1 expression in ovarian cancer. Results show that there is a positive correlation between p53 and p-p53(Ser20), p53 and SOX2, p-p53(Ser20) and CD133, lastly, p-p53(Ser20) and Notch1 expressions. Additionally, study found that p-p53(Ser20)/SOX2 expression

correlated with advance stage of cancer [67]. As it is reviewed in Carr et al., 2016, , DNA damage induced by the IR- and UV light causes post-translational modifications on p53 protein most commonly phosphorylation of serines (6, 9, 15, 20, 33, 37, 46), and threonines (18 and 81) residues [79], [133]. Any of this modification can disrupt the interaction between p53 and MDM2, however more recent studies established that MDM2 complex interacts with p53 via three critical residues which are Phe19, Trp23, and Leu26 [61]. This information can partially explain the lack of positive correlation between p-p53(Ser20) and SOX2 expression in ovarian cancer study as it is described above. More functional studies of the protein are needed for better understanding of the mechanism.

This finding from the literature with the findings of this thesis work support the hypothesis that there is a association between mutated p53 and SOX2 amplifications in various human cancers stemmed from dysregulations in involved pathways such as PI3K/AKT signaling pathway. This further highlights the need for more research to understand and potentially manipulate the targets of these pathways for new therapeutic approaches.

5.1.1 Liver Hepatocellular Carcinoma

Statistical test reviled that there is significant association between TP53 mutation and SOX2 gain/amplification in Liver Hepatocellular Carcinoma (LHCC) cohort ($p = 0.0002$). Similar study was conducted by Osman et al., 2020. In this study authors investigated the expression patterns of SOX2, SOX9, β-catenin, and p53 in hepatocellular carcinoma (HCC) tissue. Results show that there is a significant

association between high expression of SOX2, SOX9 and increased β-catenin and p53 expression [134]. This results partially supports the thesis results by indicating certain connection between p53 and SOX2 in this specific cohort. Additionally, previous studies established that both p53 and SOX2 plays role in HCC through different signaling pathways [111], [113]. Moreover, the importance of the interaction between p53-MDM2 for human HCC is reviewed in Azer et al., 2019. The p53-MDM2 complex functionally overlaps with SOX2 on the PI3K/AKT pathway which is another signaling pathway that its dysregulation is linked with HCC formation [110], [135].

5.1.2 Lung Adenocarcinoma

The association between TP53 mutation and SOX2 copy number alteration is previously established by Erdem et al., 2016 in non-small cell lung cancers. SOX2 is seen as a biomarker for both lung adenocarcinoma (LUAD) and non-small cell lung cancer (NSCLC) [71], however the expression rate, prognostic value, and clinical associations of SOX2 varies in different lung cancer studies [100], [106], [136], [137]. SOX2 expression associated with poor prognosis in LUAD and NSCLC patients [136], [137]. Whereas other studies suggested that SOX2 expression correlates with overall favorable impact on survival of NSCLC patients [138], [139]. Therefore, the association between SOX2 and prognostic outcome remains controversial for NSCLC. Moreover, p53-MDM2 interaction is proven to be important therapeutic target in lung cancer [140]. Studies suggest that both MDM2 and p53 work as biomarker for lung cancer and can be targeted for therapeutic

approach [141]. Statistical results of the thesis show significant association between TP53 mutation and SOX2 gain/amplification ($p < 0.0001$) for Lung Adenocarcinoma cohort. Findings of the literature and the findings of this thesis work reflects the complex nature of SOX2 and p53 in lung cancer with undeniable impact on key mechanisms leading to tumorigenesis.

5.1.3 Thyroid Carcinoma

The thesis work reviles significant association between high SOX2 expression and TP53 mutation in thyroid carcinomas. Somatic mutation profiling and whole exome sequencing analysis have conducted for thyroid carcinoma cohorts previously [142]. Key genes and pathways involved in thyroid carcinoma have been discussed and reviewed in Niciporuka et al., 2021. Literature suggests that the role of p53 in thyroid carcinoma has not been established and remains controversial [142], [143]. Similarly, the role of SOX2 was investigated for thyroid carcinoma [101], [144], [145]. Studies concluded that overexpression of SOX2 contributes to the aggressiveness of the tumor through PI3K/AKT signaling pathway [144], [145]. Additionally, SOX2 is found related to CSC related chemotherapy resistance for thyroid carcinoma [101].

The literature remains to be uncertain about the association of SOX2 and p53 for this cancer type. However, like this thesis, one study from 2011 investigates expression of BRAF, SOX2, and p53 in anaplastic thyroid carcinoma. Results indicates that SOX2 expression was higher in carcinomas with p53 expression ($p = 0.06$) and authors concluded that p53 and SOX2 expression can promote tumor progression for thyroid carcinoma [145].

5.1.4 Breast Invasive Carcinoma

Both TP53 and SOX2 plays role in progression of breast cancer [62], [69], [74], [78], [90], [96]. Studies indicates p53 mutation can induce stem-cell like stage to tumor and correlates with poor disease prognosis in breast cancer patients [78]. Different isoforms of p53 were studied to investigates their association with tumor progress, malignancy, and poor prognosis in breast cancer [62], [74], [78]. One study from 2021 shows that Δ133p53β isoform increases the expression of pluripotency markers such as *SOX2*, *NANOG* and *OCT4.* Same isoform is found to be correlated with poor prognosis and chemotherapy resistance [62]. Similarly, another p53 isoform, Δ40p53, found to be co-localize with stem-cell markers Sox2, Oct4, and Nanog in MCF-7 cell line which is breast cancer cell line. Additionally, Δ40p53 isoform correlates with high expression of Sox2, tumor growth and it can promote chemoresistance in breast cancer cells [74]. In 2020, Xiao et al. discovered that overexpression of SOX2 increases migration and blood-brain barrier (BBB) permeability, thus promotes brain metastasis of breast cancer. Functional and bioinformatical analysis reviled that high level of SOX2 expression leads to brain metastasis via upregulating FSCN1 and HBEGF genes in breast cancer cells [96]. With a similar pattern with p53 mutations, drug resistance is also induced by overexpression of SOX2 gene. Previous analysis indicates that overexpression of SOX2 gene increases the resistance to tamoxifen in MCF-7 cell line whereas, silencing of SOX2 resulted in opposite effect [69].

The findings describe above indicates that there is certain association between mutated p53 and SOX2 expression in breast cancer, however the exact mechanism remains to be uncertain.

5.1.5 Cancer Cohorts That Exhibit No Association

The other cohorts that were analyzed did not yield significant association between TP53 mutation and SOX2 gain/amplification. These cohorts are esophageal carcinoma, glioblastoma, brain lower grade glioma, ovarian serous cytadenocarcinoma. The SOX2 gene is a key regulator for cancer stemness and tumor formation in many cancers including brain, ovarian, and esophageal cancers [67], [83], [92], [97], [102], [120], [146].

SOX2 regulates and promotes sphere formation in glioblastoma. One study on grade II to IV astrocytoma shows increased expression of SOX2, SOX4, and ID4, in the presence of mutated p53 is only detected in grade II astrocytoma [120]. Similarly, SOX2 overexpression is correlated with poor prognosis in ovarian and esophageal cancers. [92], [97] However, no direct evidence is found regarding the possible positive association between SOX2 gain and TP53 mutation in any of the cancer cohorts that listed above, thus supporting the finding of the thesis.

5.2 Novel Shiny App for *De novo* Gene Synthesis

This thesis work introduces a novel Shiny application called GeneCraft, that combines codon optimization and overlapping primer design for PCR assembly under one tool. Both techniques have proven to be useful in the field of synthetic biology [14], [20], [24]. Previous studies have shown that codon optimization of the sequence is promisingly increasing the protein yield for recombinant proteins [12], [17]. Additionally, *de novo* gene synthesis allows researchers to custom design their sequences for purposes of codon optimization, variant or mutation analysis, or functional studies of the protein[2], [8], [24], [147].

Existing tools for these functions were listed in the thesis previously. None of them provides two different functions under one tool (***Table 1-1***). This makes GeneCraft a novel online software that provides simultaneous codon optimization for up to 10 host organisms and generates the hierarchical relationships between the optimized sequences from different organisms with respect to the query as a dendrogram. This feature of the GeneCraft app can help researchers to compare different organism's codon bias easily.

Keep in mind that, distance analysis based on codon bias constitutes the decision of host organism only partially. Scientists should take other aspects related to nature of the protein into account before choosing the host organism. Protein structure, solubility of the protein, post translational modifications should be considered before choosing the host organism for efficient protein yield [9].

Moreover, GeneCraft allows users to optimize the scale parameter for codon optimization. Scale parameters allow users to codon optimize the sequence in

different rate of replacement. Liu et al., 2022 states that extreme codon optimization of the sequences may resulted in decrease level of protein expression in host organism due to mRNA structural changes which is an undesirable effect [148]. Scale parameter helps researchers to fine tune their codon optimization process. Case study conducted with human insulin shows the effect of scale function on the evolutionary distance between the original and optimized sequences.

Existing tools for codon optimization are listed in **Table 1-1**. Amongst these tools only GeneSmart [46] provides simultaneous codon optimization for two host organisms. Existing tools contains codon adaptation index for more extensive list of host organisms then GeneCraft. This is because, GeneCraft app is restricted to the CAI of organisms curated from GeneGA package. As a feature perspective, existing dataset can be manually edited with new index calculations as described and reviewed in the introduction part of this thesis.

While ExpOptimizer and VectorBuilder tools offer option to exclude restriction sited from the sequence [43], [45], IDT tool only identify, and list restriction sites found in the sequence [19]. Currently GeneCraft app lacks restriction enzyme exclusion feature. However, what sets GeneCraft apart from other tools is that none of existing tools offers distance analysis or alignment graph for the codon optimized sequences .

Due its rapid cost effective and reproducible nature, PCR Assembly is one of the most common methods that is used for de novo gene synthesis [12], [14], [17]. The mechanism behind the assembly is described in the introduction part of the thesis. One of the most important aspects is that the melting temperature should be

synchronized among the primers for a successful PCR reaction. Therefore, designing the overlapping primers can be tedious work especially with longer sequences. Existing tools such as Primirize optimally work under 500nt bases [53]. GeneCraft app can work with longer sequences even over 2000 bp in length since it is a recursive function. On the other hand, Primirize and other tools such as Stitcher [50] accommodate CG control over primer design while GeneCraft currently lacks this control but compensates by providing CG% information as an output in the app. Another important aspect of the gene synthesis is the cost. Recent review shows synthesis of gene-length DNA sequences between 200-2000 bp costs from $0.10 to $0.30. Each primer synthesis cost differently based on the reagent that is used and the purification methods Click or tap here to enter text.. As a future perspective the app can calculate the cost value for primers. The cost value can be used to optimize the parameters to design primers with the lowest possible cost which can contribute to the one of the most important considerations of the gene synthesis which is a cost efficiency.

5.3 In Silico Design for Recombinant SOX2

SOX2 gene is essential transcription factor for stem cell maintenance and neural development for mammals[68], [93], [146]. Previously observed that SOX2 gene works as oncogene for different cancer types[92], [95], [96], [98], [146]. Due to its diverse effect on critical biological mechanisms, SOX2 is subjected to be recombinantly expressed for different research settings.

The first use of area is iPSC generation. In this type of studies, Sox2 is recombinantly expressed alongside with Oct4, cMyc, and other reprogramming proteins to induce pluripotency in other living cells [26], [70], [147]. SOX2 is widely expressed in many cancer stem cells (CSCs) for example in lung cancer [95], [104], esophageal squamous cell carcinoma [146], glioblastoma [102], and others [92], [98]. CSCs are resistant to the conventional cancer treatments like chemotherapy and radiotherapy due to self-renewal and self-seeding capacity [91], [92], [93]. It is previously proven that SOX2 expression increases the CSC markers and shows enhanced apoptosis resistance in numerous cancer types [91], [104], [146].

Due to these abovementioned facts, Sox2 transcription factor is a potential target for alternative cancer therapeutic approaches. In one study, antitumor DNA vaccine has been generated against Sox2 transcription factor. Vector construct was designed, and the sequence was codon optimized for bacterial host before cloning. Interestingly, codon optimization did not yield higher protein in this specific research [91]. In 2017, another study selectively targeted SOX2 by generating an artificial transcription factor (ATF). Engineered ATF/SOX2 system expressed in living cell and decreased the SOX2 expression in mRNA level [146].

This thesis work offers in silico design for recombinant human SOX2 protein using GeneCraft app. The two-step analysis, first, has provided comparisons between different host organisms based on their codon usage bias. Literature indicates recombinant SOX2 has been expressed predominantly in bacterial genome, *Escherichia coli* [26], [70], [114]. However, in some studies yeast and insect cells were utilized in its overexpression and production [26], [70], [91].

Additionally, a literature search conducted to identify most commonly used bacterial host organisms for recombinant protein production. The most commonly used host organisms from the literature is curated and compared with the list of organisms found in GeneGA package codon usage dataset [121]. Organisms overlapped in two list were used for *in silico* design of human SOX2 through GeneCraft app. . Phylogenic analysis revealed the evolutionary distance between different optimized sequences. Based on the analysis, B. subtilis and S. cerevisia were the two most closely associated one with human codon usage and hence may require less number of modificiations in the sequence.

Secondly, GeneCraft has designed a set of overlapping primers from an optimized sequence to be used in PCR-assembly with pre-set melting temperatures, with a mean of 56.7 and standard deviation of 0.9. Primer design is conducted for the SOX2 sequence with same parameters in other online tools for comparison. The Primirize tool works best under 500bp and cannot function over 1000bp. Therefore, the design for the 954nt long SOX2 sequence takes relatively longer in the application. The melting temperatures of the designed primers in Primirize are recorded with a mean of 61.63 and a standard deviation of 5.3, which is a much higher variance than the GeneCraft app. However, the Primirize app also identifies promoter sites in the sequence, whereas the GeneCraft app currently lacks this function. Both GeneCraft and Stitcher app produce a similar number of primers for the SOX2 sequence, 28 and 30 respectively. Stitcher designs primers with a mean of 58.67 and a standard deviation of 1.9, which is lower than Primirize but still higher than GeneCraft app. The high variance may results from the wide range of controls that Stitcher offers on

the primer design parameters such as CG content, self-complementarity, and dimerization. These controls can contribute to the high variance in the melting temperatures by limiting the primer design process. Controls that described previously open rooms for improvement in the GeneCraft app.

This has shown that the GeneCraft app developed in this thesis can provide an effective yet easy and fast initial design for de novo gene synthesis of any gene, as exemplified by SOX2.

References

[1] O. Khow and S. Suntrarachun, “Strategies for production of active eukaryotic proteins in bacterial expression system,” *Asian Pac J Trop Biomed*, vol. 2, no. 2, pp. 159–162, 2012, doi: 10.1016/S2221-1691(11)60213-X.

[2] V. Gupta, M. Sengupta, J. Prakash, and B. C. Tripathy, “Production of Recombinant Pharmaceutical Proteins,” in *Basic and Applied Aspects of Biotechnology*, Springer Singapore, 2017, pp. 77–101. Doi: 10.1007/978-981-10-0875-7_4.

[3] D. S. W. Ow, M. K. Oh, C. J. Chiang, and Y. P. Chao, “Editorial: Technological Advances Improving Recombinant Protein Production in Bacteria,” *Frontiers in Microbiology*, vol. 12. Frontiers Media S.A., Jul. 26, 2021. Doi: 10.3389/fmicb.2021.729472.

[4] B. Owczarek, A. Gerszberg, and K. Hnatuszko-Konka, “A Brief Reminder of Systems of Production and Chromatography-Based Recovery of Recombinant Protein Biopharmaceuticals,” *BioMed Research International*, vol. 2019. Hindawi Limited, 2019. Doi: 10.1155/2019/4216060.

[5] N. A. Baeshen *et al.*, “Cell factories for insulin production,” 2014. [Online]. Available: http://www.microbialcellfactories.com/content/13/1/141

[6] M. C. S. Pranchevicius and T. R. Vieira, “Production of recombinant immunotherapeutics for anticancer treatment: the role of bioengineering.,” *Bioengineered*, vol. 4, no. 5. Pp. 305–312, 2013. Doi: 10.4161/bioe.24666.

[7] H. H. Nguyen, S. H. Lee, U. J. Lee, C. D. Fermin, and M. Kim, “Immobilized enzymes in biosensor applications,” *Materials*, vol. 12, no. 1. MDPI AG, Jan. 02, 2019. Doi: 10.3390/ma12010121.

[8] X. Zeng, Z. Shen, and R. Mernaugh, “Recombinant antibodies and their use in biosensors,” *Analytical and Bioanalytical Chemistry*, vol. 402, no. 10. Pp. 3027–3038, Apr. 2012. Doi: 10.1007/s00216-011-5569-z.

[9] P. T. Wingfield, “Overview of the purification of recombinant proteins,” *Curr Protoc Protein Sci*, vol. 2015, pp. 6.1.1-6.1.35, 2015, doi: 10.1002/0471140864.ps0601s80.

[10] D. Qin, “Next-generation sequencing and its clinical application,” *Cancer Biol Med*, vol. 16, no. 1, pp. 4–10, 2019, doi: 10.20892/j.issn.2095-3941.2018.0055.

[11] D. G. Gibson *et al.*, “Creation of a Bacterial Cell Controlled by a Chemically Synthesized Genome,” 2023.

[12] J. Cherry, B. W. Nieuwenhuijsen, E. J. Kaftan, J. D. Kennedy, and P. K. Chanda, “A modified method for PCR-directed gene synthesis from large number of overlapping oligodeoxyribonucleotides.,” *J Biochem Biophys Methods*, vol. 70, no. 6, pp. 820–822, 2008, doi: 10.1016/j.jprot.2007.12.009.

[13] S. Mueller, J. R. Coleman, and E. Wimmer, “Putting Synthesis into Biology: A Viral View of Genetic Engineering through De Novo Gene and Genome Synthesis,” *Chemistry and Biology*, vol. 16, no. 3. Pp. 337–347, Mar. 27, 2009. Doi: 10.1016/j.chembiol.2009.03.002.

[14] A. F. Sequeira, J. L. A. Brás, C. I. P. D. Guerreiro, R. Vincentelli, and C. M. G. A. Fontes, “Development of a gene synthesis platform for the efficient large scale production of small genes encoding animal toxins,” *BMC Biotechnol*, vol. 16, no. 1, Dec. 2016, doi: 10.1186/s12896-016-0316-3.

[15] H. Ashrafi *et al.*, “De novo assembly of the pepper transcriptome (Capsicum annuum): A benchmark for in silico discovery of SNPs, SSRs and candidate genes,” *BMC Genomics*, vol. 13, no. 1, Oct. 2012, doi: 10.1186/1471-2164-13-571.

[16] S. L. Thein, J. R. Lynch, D. J. Weatherall, and R. B. Wallace, “DIRECT DETECTION OF HAEMOGLOBIN E WITH SYNTHETIC OLIGONUCLEOTIDES,” *The Lancet*, vol. 327, no. 8472. P. 93, Jan. 11, 1986. Doi: 10.1016/S0140-6736(86)90739-7.

[17] J. R. TerMaat, E. Pienaar, S. E. Whitney, T. G. Mamedov, and A. Subramanian, “Gene synthesis by integrated polymerase chain assembly and PCR amplification using a high-speed thermocycler,” *J Microbiol Methods*, vol. 79, no. 3, pp. 295–300, Dec. 2009, doi: 10.1016/j.mimet.2009.09.015.

[18] C. A. Hutchison *et al.*, “Design and synthesis of a minimal bacterial genome,” *Science (1979)*, vol. 351, no. 6280, Mar. 2016, doi: 10.1126/science.aad6253.

[19] R. A. Hughes Editor, “Synthetic DNA Methods and Protocols Methods in Molecular Biology 1472.” [Online]. Available: http://www.springer.com/series/7651

[20] A. Hilterbrand, J. Saelens, and C. Putonti, “CBDB: The codon bias database,” *BMC Bioinformatics*, vol. 13, no. 1, Apr. 2012, doi: 10.1186/1471-2105-13-62.

[21] M. Diez, S. G. Medina-Muñoz, L. A. Castellano, G. da Silva Pescador, Q. Wu, and A. A. Bazzini, “iCodon customizes gene expression based on the codon composition,” *Sci Rep*, vol. 12, no. 1, Dec. 2022, doi: 10.1038/s41598-022-15526-7.

[22] P. Puigbò, E. Guzmán, A. Romeu, and S. Garcia-Vallvé, “OPTIMIZER: A web server for optimizing the codon usage of DNA sequences,” *Nucleic Acids Res*, vol. 35, no. SUPPL.2, Jul. 2007, doi: 10.1093/nar/gkm219.

[23] E. P. C. Rocha, “Codon usage bias from tRNA’s point of view: Redundancy, specialization, and efficient decoding for translation optimization,” *Genome Res*, vol. 14, no. 11, pp. 2279–2286, Nov. 2004, doi: 10.1101/gr.2896904.

[24] M. G. Rosenblum, S. A. Horn, and L. H. Cheung, “A novel recombinant fusion toxin targeting HER-2/NEU-over-expressing cells and containing human tumor necrosis factor,” *Int J Cancer*, vol. 88, no. 2, pp. 267–273, 2000, doi: 10.1002/1097-0215(20001015)88:2<267::AID-IJC19>3.0.CO;2-G.

[25] E. Salehi *et al.*, “Recombinant expression and purification of L2 domain of human epidermal growth factor receptor,” *Afr J Biotechnol*, vol. 9, no. 33, pp. 5292–5296, 2010, doi: 10.5897/AJB10.216.

[26] C. Pan, W. Jia, B. Lu, and C. E. Bishop, “Expression of TAT recombinant Oct4, Sox2, Lin28, and Nanog proteins from baculovirus-infected Sf9 insect cells,” *Gene*, vol. 556, no. 2, pp. 245–248, Feb. 2015, doi: 10.1016/j.gene.2014.11.061.

[27] R. A. Hughes and A. D. Ellington, “Synthetic DNA synthesis and assembly: Putting the synthetic in synthetic biology,” *Cold Spring Harb Perspect Biol*, vol. 9, no. 1, Jan. 2017, doi: 10.1101/cshperspect.a023812.

[28] S. Kosuri and G. M. Church, “Large-scale de novo DNA synthesis: Technologies and applications,” *Nature Methods*, vol. 11, no. 5. Nature Publishing Group, pp. 499–507, 2014. Doi: 10.1038/nmeth.2918.

[29] D. G. Gibson *et al.*, “Creation of a Bacterial Cell Controlled by a Chemically Synthesized Genome,” 2023.

[30] T. Kaneko *et al.*, “Sequence Analysis of the Genome of the Unicellular Cyanobacterium Synechocystis sp. Strain PCC6803. II. Sequence Determination of the Entire Genome and Assignment of Potential Protein-coding Regions,” 1996. [Online]. Available: https://academic.oup.com/dnaresearch/article/3/3/109/342335

[31] J. L. King and T. H. Jukes, “Non-Darwinian Evolution,” 1967. Accessed: Jan. 21, 2024. [Online]. Available: https://www.science.org/doi/epdf/10.1126/science.164.3881.788?adobe_mc=MCMID%3D76837889616765450842543885410465693860%7CMCORGID%3D242B6472541199F70A4C98A6%2540AdobeOrg%7CTS%3D1705825350
[32] M. Dos Reis, R. Savva, and L. Wernisch, “Solving the riddle of codon usage preferences: a test for translational selection”, doi: 10.1093/nar/gkh834.
[33] S. T. Parvathy, V. Udayasuriyan, and · Vijaipal Bhadana, “Codon usage bias,” vol. 49, pp. 539–565, 2022, doi: 10.1007/s11033-021-06749-4.
[34] G. L. Rosano and E. A. Ceccarelli, “Rare codon content affects the solubility of recombinant proteins in a codon bias-adjusted Escherichia coli strain,” *Microb Cell Fact*, vol. 8, Jul. 2009, doi: 10.1186/1475-2859-8-41.
[35] O. G. Berg and P. J. N. Silva, “Codon bias in Escherichia coli: the influence of codon context on mutation and selection,” Oxford University Press, 1997. [Online]. Available: https://academic.oup.com/nar/article/25/7/1397/1013998
[36] D. J. Lynn, G. A. C. Singer, and D. A. Hickey, “Synonymous codon usage is subject to selection in thermophilic bacteria.” [Online]. Available: https://academic.oup.com/nar/article/30/19/4272/2376087
[37] H. Akashi, “Synonymous Codon Usage in Drosophila melanogaster: Natural Selection and Translational Accuracy,” 1994. [Online]. Available: https://academic.oup.com/genetics/article/136/3/927/6012271
[38] M. Bulmer’, “The Selection-Mutation-Drift Theory of Synonymous Codon Usage,” 1991. [Online]. Available: https://academic.oup.com/genetics/article/129/3/897/6007089
[39] P. M. Sharpl and W.-H. Li2, “Nucleic Acids Research The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications.”
[40] G. Wright, A. Rodriguez, J. Li, T. Milenkovic, S. J. Emrich, and P. L. Clark, “CHARMING: Harmonizing synonymous codon usage to replicate a desired codon usage pattern,” *Protein Science*, vol. 31, no. 1, pp. 221–231, Jan. 2022, doi: 10.1002/pro.4223.
[41] R. Jansen, H. J. Bussemaker, and M. Gerstein, “Revisiting the codon adaptation index from a whole-genome perspective: analyzing the relationship between gene expression and codon occurrence in yeast using a variety of models”, doi: 10.1093/nar/gkg306.
[42] P. Puigbò, I. G. Bravo, and S. Garcia-Vallve, “CAIcal: A combined set of tools to assess codon usage adaptation,” 2008, doi: 10.1186/1745-6150-3-38.
[43] “Codon Optimization Tool | VectorBuilder.” Accessed: Jan. 17, 2024. [Online]. Available: https://en.vectorbuilder.com/tool/codon-optimization.html
[44] P. M. Sharp1, M. Averof2, A. T. Lloyd3, G. Matassi1, and J. F. Peden1, “DNA sequence evolution: the sounds of silence,” 1995. [Online]. Available: https://royalsocietypublishing.org/
[45] “Codon Optimization (ExpOptimizer) – Online Tools.” Accessed: Jan. 17, 2024. [Online]. Available: https://www.novoprolabs.com/tools/codon-optimization#
[46] “GenSmart Codon Optimization.” Accessed: Jan. 17, 2024. [Online]. Available: https://www.genscript.com/tools/gensmart%2dcodon%2doptimization?act=register_success
[47] G. Wu, J. B. Wolf, A. F. Ibrahim, S. Vadasz, M. Gunasinghe, and S. J. Freeland, “Simplified gene synthesis: A one-step approach to PCR-based gene construction,” *J Biotechnol*, vol. 124, no. 3, pp. 496–503, Jul. 2006, doi: 10.1016/j.jbiotec.2006.01.015.

[48] A. F. Sequeira, J. L. A. Brás, C. I. P. D. Guerreiro, R. Vincentelli, and C. M. G. A. Fontes, "Development of a gene synthesis platform for the efficient large scale production of small genes encoding animal toxins," *BMC Biotechnol*, vol. 16, no. 1, Dec. 2016, doi: 10.1186/s12896-016-0316-3.

[49] A.-S. Xiong *et al.*, "Chemical gene synthesis: strategies, softwares, error corrections, and applications A.S. Xiong et al. The progress of chemical gene synthesis", doi: 10.1111/j.1574-6976.2008.00109.x.

[50] D. M. O'Halloran, I. Uriagereka-Herburger, and K. Bode, "STITCHER 2.0: Primer design for overlapping PCR applications," *Sci Rep*, vol. 7, Mar. 2017, doi: 10.1038/srep45349.

[51] M.J. van Baren and P. Heutink, "The PCR Suite," *Bioinformatics*, vol. 591, no. 3, pp. 1–20, Mar. 2004.

[52] C. Basu Editor, "PCR Primer Design Third Edition." [Online]. Available: http://www.springer.com/series/7651

[53] "Primerize: Primer Design for PCR Assembly." Accessed: Jan. 17, 2024. [Online]. Available: https://primerize.stanford.edu/design_1d/

[54] A. Upadhyay, "Cancer: An unknown territory; rethinking before going ahead," *Genes and Diseases*, vol. 8, no. 5. Chongqing University, pp. 655–661, Sep. 01, 2021. Doi: 10.1016/j.gendis.2020.09.002.

[55] M. Pourhassan-Moghaddam, B. Farhadi, K. Nejati-Koshki, and T. Moharrami, "Promoters of Cancer Genes for Recombinant Protein Expression in Human Cancer Cell Lines," *BioImpacts*, vol. 2, no. 1, pp. 55–60, 2012, doi: 10.5681/bi.2012.007.

[56] S. Khan *et al.*, "Role of Recombinant DNA Technology to Improve Life," 2016, doi: 10.1155/2016/2405954.

[57] J. R. Kintzing, M. V. Filsinger Interrante, and J. R. Cochran, "Emerging Strategies for Developing Next-Generation Protein Therapeutics for Cancer Treatment," *Trends in Pharmacological Sciences*, vol. 37, no. 12. Elsevier Ltd, pp. 993–1008, Dec. 01, 2016. Doi: 10.1016/j.tips.2016.10.005.

[58] AKIO SAIURA, MASATA KASATA, YASUNO BUHIRATA, NAGAI RYOZO, and MASATOSHI MAKUUCHI, "Differential hepatocyte toxicity of recombinant Apo2L/TRAIL versions," *Nature*, vol. 7, no. 4, pp. 383–385, 2004.

[59] R. Fadeev *et al.*, "Improved Anticancer Effect of Recombinant Protein izTRAIL Combined with Sorafenib and Peptide iRGD," *International Journal of Molecular Sciences Article*, 2019, doi: 10.3390/ijms20030525.

[60] J. J. Qin, X. Li, C. Hunt, W. Wang, H. Wang, and R. Zhang, "Natural products targeting the p53-MDM2 pathway and mutant p53: Recent advances and implications in cancer medicine," *Genes and Diseases*, vol. 5, no. 3. Chongqing University, pp. 204–219, Sep. 01, 2018. Doi: 10.1016/j.gendis.2018.07.002.

[61] P. Das, V. Satish, and K. Mattaparthi, "Computational Investigation on the p53−MDM2 Interaction Using the Potential of Mean Force Study," *Cite This: ACS Omega*, vol. 5, p. 8462, 2020, doi: 10.1021/acsomega.9b03372.

[62] S. Reinhardt *et al.*, "Citation: Cytoplasmic p53β Isoforms Are Associated with Worse Disease-Free Survival in Breast Cancer," 2022, doi: 10.3390/ijms23126670.

[63] T. C. Lu *et al.*, "Transduction of recombinant M3-p53-R12 protein enhances human leukemia cell apoptosis," *J Cancer*, vol. 7, no. 10, pp. 1360–1373, 2016, doi: 10.7150/jca.15155.

[64] N. H. Chehab, A. Malikzay, E. S. Stavridi, and T. D. Halazonetis, “Phosphorylation of Ser-20 mediates stabilization of human p53 in response to DNA damage,” Merck & Co., Inc, 1999. [Online]. Available: www.pnas.org

[65] N. H. Chehab, A. Malikzay, E. S. Stavridi, and T. D. Halazonetis, “Phosphorylation of Ser-20 mediates stabilization of human p53 in response to DNA damage,” Merck & Co., Inc, 1999. [Online]. Available: www.pnas.org

[66] N. Rivlin, R. Brosh, M. Oren, and V. Rotter, “Mutations in the p53 tumor suppressor gene: Important milestones at the various steps of tumorigenesis,” *Genes and Cancer*, vol. 2, no. 4. Pp. 466–474, Apr. 2011. Doi: 10.1177/1947601911408889.

[67] J. Bar, P. Grelewski, I. Deszcz, L. Noga, L. Hirnle, and A. Lis-Nawara, “Association between p53 protein phosphorylated at serine 20 expression and ovarian carcinoma stem cells phenotype: Correlation with clinicopathological parameters of ovarian cancer,” *Neoplasma*, vol. 66, no. 5, pp. 801–809, 2019, doi: 10.4149/neo_2018_181012N764.

[68] D. Kelberman *et al.*, “Mutations within Sox2/SOX2 are associated with abnormalities in the hypothalamo-pituitary-gonadal axis in mice and humans,” *Journal of Clinical Investigation*, vol. 116, no. 9, pp. 2442–2455, Sep. 2006, doi: 10.1172/JCI28658.

[69] E. L. Wuebben and A. Rizzino, “The dark side of SOX2: cancer-a comprehensive overview,” 2017. [Online]. Available: www.impactjournals.com/oncotarget/

[70] M. I, B. Münst, S. Mielke, and F. Edenhofer, “Cellular reprogramming employing recombinant Sox2 protein,” *Stem Cells Int*, 2012, doi: 10.1155/2012/549846.

[71] J. Samulin Erdem, V. Skaug, P. Bakke, A. Gulsvik, A. Haugen, and S. Zienolddiny, “Mutations in TP53 increase the risk of SOX2 copy number alterations and silencing of TP53 reduces SOX2 expression in non-small cell lung cancer,” *BMC Cancer*, vol. 16, no. 1, Jan. 2016, doi: 10.1186/s12885-016-2061-3.

[72] G. D’Orazi, “Recent advances in P53,” *Biomolecules*, vol. 11, no. 2. MDPI AG, pp. 1–4, Feb. 01, 2021. Doi: 10.3390/biom11020211.

[73] D. P. Guimaraes and P. Hainaut, “TP53: a key gene in human cancer,” 2002. [Online]. Available: http://www.iarc.fr/p53/

[74] L. Steffens Reinhardt, K. Groen, X. Zhang, B. C. Morten, A. Wawruszak, and K. A. Avery-Kiejda, “ARTICLE OPEN p53 isoform expression promotes a stemness phenotype and inhibits doxorubicin sensitivity in breast cancer”, doi: 10.1038/s41419-023-06031-4.

[75] T. Schaefer, R. Steiner, and C. Lengerke, “SOX2 and p53 expression control converges in PI3K/AKT signaling with versatile implications for stemness and cancer,” *International Journal of Molecular Sciences*, vol. 21, no. 14. MDPI AG, pp. 1–21, Jul. 02, 2020. Doi: 10.3390/ijms21144902.

[76] O. Hassin and M. Oren, “Drugging p53 in cancer: one protein, many targets,” *NATURE REVIEWS*, vol. 22, 2023, doi: 10.1038/s41573-022-00571-8.

[77] C. M. Eischen, “Genome stability requires p53,” *Cold Spring Harb Perspect Med*, vol. 6, no. 6, Jun. 2016, doi: 10.1101/cshperspect.a026096.

[78] H. Mizuno, B. T. Spike, G. M. Wahl, and A. J. Levine, “Inactivation of p53 in breast cancers correlates with stem cell transcriptional signatures”, doi: 10.1073/pnas.1017001108.

[79] M. I. Carr and S. N. Jones, “Regulation of the Mdm2-p53 signaling axis in the DNA damage response and tumorigenesis,” *Translational Cancer Research*, vol. 5, no. 6. AME Publishing Company, pp. 707–724, 2016. Doi: 10.21037/tcr.2016.11.75.

[80] S. Cingir Koker, E. Jahja, H. Shehwana, A. Gokce Keskus, and O. I. Konu, “Cholinergic Receptor Nicotinic Alpha 5 (CHRNA5) RNAi is associated with cell cycle inhibition, apoptosis, DNA damage response and drug sensitivity in breast cancer,” 2018, doi: 10.1371/journal.pone.0208982.
[81] P. J. Buckhaults, “TP53 knockout in MCF7 breast cancer cells induces sensitivity to fluoropyrimidine drugs,” *bioRxiv*, Feb. 2022, doi: 10.1101/2022.02.10.479997.
[82] P. Li, T. Bui, D. Gray, and H. J. Klamut, “Therapeutic potential of recombinant p53 overexpression in breast cancer cells expressing endogenous wild-type p53,” 1998.
[83] H. Wang, Y. Zhou, Q. Liu, J. Xu, and Y. Ma, “Prognostic value of SOX2, cyclin D1, P53, and ki-67 in patients with esophageal squamous cell carcinoma,” *Onco Targets Ther*, vol. 11, pp. 5171–5181, 2018, doi: 10.2147/OTT.S160066.
[84] M. Lafevre-Bernt, S. Wu, and X. Lin, “Recombinant, refolded tetrameric p53 and gonadotropin-releasing hormone-p53 slow proliferation and induce apoptosis in p53-deficient cancer cells,” *Mol Cancer Ther*, vol. 7, no. 6, pp. 1420–1429, 2008, doi: 10.1158/1535-7163.MCT-08-0078.
[85] C. Deben, V. Deschoolmeester, F. Lardon, C. Rolfo, and P. Pauwels, “TP53 and MDM2 genetic alterations in non-small cell lung cancer: Evaluating their prognostic and predictive value,” *Critical Reviews in Oncology/Hematology*, vol. 99. Elsevier Ireland Ltd, pp. 63–73, Mar. 01, 2016. Doi: 10.1016/j.critrevonc.2015.11.019.
[86] Y. Guo *et al.*, “Recombinant Human Adenovirus-p53 Therapy for the Treatment of Cervical Cancer: A Meta-Analysis,” vol. 11, 2021, doi: 10.3389/fonc.2021.748681.
[87] Q. Wang, Q. Shi, Y. Chen, G. Wang, and L. Shen, “Recombinant human p53 adenovirus injection combined with Bortezomib inhibits proliferation and promotes apoptosis in multiple myeloma,” *Leuk Res*, vol. 127, Apr. 2023, doi: 10.1016/j.leukres.2023.107041.
[88] T. Schaefer and C. Lengerke, “SOX2 protein biochemistry in stemness, reprogramming, and cancer: the PI3K/AKT/SOX2 axis and beyond,” *Oncogene*, vol. 39, no. 2. Springer Nature, pp. 278–292, Jan. 09, 2020. Doi: 10.1038/s41388-019-0997-x.
[89] W. L. Tam and H. H. Ng, “Sox2: Masterminding the Root of Cancer,” *Cancer Cell*, vol. 26, no. 1. Cell Press, pp. 3–5, Jul. 14, 2014. Doi: 10.1016/j.ccr.2014.06.024.
[90] M. Piva *et al.*, “Sox2 promotes tamoxifen resistance in breast cancer cells,” *EMBO Mol Med*, vol. 6, no. 1, pp. 66–79, Jan. 2014, doi: 10.1002/emmm.201303411.
[91] I. Polakova, M. Duskova, and M. Smahel, “Antitumor DNA vaccination against the Sox2 transcription factor,” *Int J Oncol*, vol. 45, no. 1, pp. 139–146, 2014, doi: 10.3892/ijo.2014.2402.
[92] P. M. Bareiss *et al.*, “SOX2 expression associates with stem cell state in human ovarian carcinoma,” *Cancer Res*, vol. 73, no. 17, pp. 5544–5555, Sep. 2013, doi: 10.1158/0008-5472.CAN-12-4177.
[93] M. Al Mamun, K. Mannoor, J. Cao, F. Qadri, and X. Song, “SOX2 in cancer stemness: Tumor malignancy and therapeutic potentials,” *Journal of Molecular Cell Biology*, vol. 12, no. 2. Oxford University Press, pp. 85–98, 2020. Doi: 10.1093/JMCB/MJY080.
[94] J. L. Cox, S. K. Mallanna, X. Luo, and A. Rizzino, “Sox2 uses multiple domains to associate with proteins present in Sox2-protein complexes,” *PloS One*, vol. 5, no. 11, 2010, doi: 10.1371/journal.pone.0015486.
[95] N. Karachaliou, R. Rosell, and S. Viteri, “The role of SOX2 in small cell lung cancer, lung adenocarcinoma and squamous cell carcinoma of the lung,” *Translational Lung Cancer Research*, vol. 2, no. 3. AME Publishing Company, pp. 172–179, 2013. Doi: 10.3978/j.issn.2218-6751.2013.01.01.

[96] W. Xiao *et al.*, “SOX2 Promotes Brain Metastasis of Breast Cancer by Upregulating the Expression of FSCN1 and HBEGF,” *Mol Ther Oncolytics*, vol. 17, pp. 118–129, Jun. 2020, doi: 10.1016/j.omto.2020.03.001.
[97] Y. Gen, K. Yasui, T. Nishikawa, and T. Yoshikawa, “SOX2 promotes tumor growth of esophageal squamous cell carcinoma through the AKT/mammalian target of rapamycin complex 1 signaling pathway,” *Cancer Sci*, vol. 104, no. 7, pp. 810–816, Jul. 2013, doi: 10.1111/cas.12155.
[98] F. Zhu *et al.*, “SOX2 Is a Marker for Stem-like Tumor Cells in Bladder Cancer,” *Stem Cell Reports*, vol. 9, no. 2, pp. 429–437, Aug. 2017, doi: 10.1016/j.stemcr.2017.07.004.
[99] L. V Nguyen, R. Vanner, P. Dirks, and C. J. Eaves, “Cancer stem cells: an evolving concept.,” *Nat Rev Cancer*, vol. 12, no. 2, pp. 133–43, Jan. 2012, doi: 10.1038/nrc3184.
[100] S. Wang *et al.*, “SOX2 Promotes Radioresistance in Non-small Cell Lung Cancer by Regulating Tumor Cells Dedifferentiation,” *Int J Med Sci*, vol. 2023, no. 6, pp. 781–796, 2023, doi: 10.7150/ijms.75315.
[101] V. Carina *et al.*, “Multiple Pluripotent Stem Cell Markers in Human Anaplastic Thyroid Cancer: The Putative Upstream Role of SOX2”, doi: 10.1089/thy.2012.0372.
[102] D. Hägerstrand *et al.*, “Identification of a SOX2-dependent subset of tumor-and sphere-forming glioblastoma cells with a distinct tyrosine kinase inhibitor sensitivity profile,” *Neuro Oncol*, vol. 13, no. 11, pp. 1178–1191, Nov. 2011, doi: 10.1093/neuonc/nor113.
[103] S. M. Rothenberg *et al.*, “Inhibition of mutant EGFR in lung cancer cells triggers SOX2-FOXO6-dependent survival pathways”, doi: 10.7554/eLife.06132.001.
[104] Y. T. Chou *et al.*, “The emerging role of SOX2 in cell proliferation and survival and its crosstalk with oncogenic signaling in lung cancer,” *Stem Cells*, vol. 31, no. 12, pp. 2607–2619, Dec. 2013, doi: 10.1002/stem.1518.
[105] K. Takeda *et al.*, “Sox2 is associated with cancer stem-like properties in colorectal cancer,” *SciEntific RepoRts |*, vol. 8, p. 17639, 2018, doi: 10.1038/s41598-018-36251-0.
[106] X. Wang *et al.*, “Stem Cell Factor SOX2 Confers Ferroptosis Resistance in Lung Cancer via Upregulation of SLC7A11”, doi: 10.1158/0008-5472.CAN-21-0567.
[107] T. Tian, Y. Zhang, S. Wang, J. Zhou, and S. Xu, “Sox2 enhances the tumorigenicity and chemoresistance of cancer stem-like cells derived from gastric cancer,” *J Biomed Res*, vol. 26, no. 5, pp. 336–345, 2012, doi: 10.7555/JBR.26.20120045.
[108] J. Shih *et al.*, “Dominant B-cell epitopes from cancer/stem cell antigen SOX2 recognized by serum samples from cancer patients,” 2014. [Online]. Available: www.ajcei.us/ISSN:2164-7712/AJCEI0000691
[109] K. Chen, B. Yu, and J. Liao, “LncRNA SOX2OT alleviates mesangial cell proliferation and fibrosis in diabetic nephropathy via Akt/mTOR-mediated autophagy,” *Molecular Medicine*, vol. 27, no. 1, Dec. 2021, doi: 10.1186/s10020-021-00310-6.
[110] S. A. Azer, “MDM2-P53 interactions in human hepatocellular carcinoma: What is the role of nutlins and new therapeutic options?,” *Journal of Clinical Medicine*, vol. 7, no. 4. MDPI, Apr. 01, 2018. Doi: 10.3390/jcm7040064.
[111] Z. Hosseini-Khah *et al.*, “SOX2 and Bcl-2 as a Novel Prognostic Value in Hepatocellular Carcinoma Progression,” 2021, doi: 10.3390/curroncol28040264.
[112] L. Ke *et al.*, “Somatic Mutation Profiles Revealed by Next Generation Sequencing (NGS) in 39 Chinese Hepatocellular Carcinoma Patients,” *Front Mol Biosci*, vol. 8, Jan. 2022, doi: 10.3389/fmolb.2021.800679.

[113] H. G. Woo *et al.*, "Association of TP53 mutations with stem cell-like gene expression and survival of patients with hepatocellular carcinoma," *Gastroenterology*, vol. 140, no. 3, pp. 1063-1070.e8, 2011, doi: 10.1053/j.gastro.2010.11.034.
[114] M. Thool, C. Dey, S. Bhattacharyya, S. Sudhagar, and R. P. Thummer, "Generation of a Recombinant Stem Cell-Specific Human SOX2 Protein from Escherichia coli Under Native Conditions," *Mol Biotechnol*, vol. 63, no. 4, pp. 327–338, Apr. 2021, doi: 10.1007/s12033-021-00305-y.
[115] J. N. Weinstein *et al.*, "The cancer genome atlas pan-cancer analysis project," *Nat Genet*, vol. 45, no. 10, pp. 1113–1120, Oct. 2013, doi: 10.1038/ng.2764.
[116] T. C. Silva *et al.*, "TCGA Workflow: Analyze cancer genomics and epigenomics data using Bioconductor packages," *F1000Res*, vol. 5, p. 1542, Jun. 2016, doi: 10.12688/f1000research.8923.1.
[117] E. Cerami *et al.*, "The cBio Cancer Genomics Portal: An open platform for exploring multidimensional cancer genomics data," *Cancer Discov*, vol. 2, no. 5, pp. 401–404, May 2012, doi: 10.1158/2159-8290.CD-12-0095.
[118] H.-Y. Kim, "Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test," *Restor Dent Endod*, vol. 42, no. 2, p. 152, 2017, doi: 10.5395/rde.2017.42.2.152.
[119] R. Winters, A. Winters, and R. G. Amedee, "Statistics: A Brief Overview."
[120] T. F. A. de Galatro *et al.*, "Differential Expression of ID4 and Its Association with TP53 Mutation, SOX2, SOX4 and OCT-4 Expression Levels," *PloS One*, vol. 8, no. 4, Apr. 2013, doi: 10.1371/journal.pone.0061605.
[121] Z. Li, H. Huang, and M. Z. Li, "Package 'GeneGA' Type Package Title Design gene based on both mRNA secondary structure and codon usage bias using Genetic algorithm," 2015. [Online]. Available: http://www.tbi.univie.ac.at/~ivo/RNA/
[122] U. Bodenhofer, E. Bonatesta, C. Horejš-Kainrath, and S. Hochreiter, "Msa: An R package for multiple sequence alignment," *Bioinformatics*, vol. 31, no. 24, pp. 3997–3999, Jul. 2015, doi: 10.1093/bioinformatics/btv494.
[123] M. Pouresmaeil and S. Azizi-Dargahlou, "Factors involved in heterologous expression of proteins in E. coli host," *Arch Microbiol*, vol. 205, no. 5, May 2023, doi: 10.1007/s00203-023-03541-9.
[124] P. Mendez and J. L. Ramirez, "Copy number gains of FGFR1 and 3q chromosome in squamous cell carcinoma of the lung," *Translational Lung Cancer Research*, vol. 2, no. 2. AME Publishing Company, pp. 101–111, 2013. Doi: 10.3978/j.issn.2218-6751.2013.03.05.
[125] S. Negrini, V. G. Gorgoulis, and T. D. Halazonetis, "Genomic instability an evolving hallmark of cancer," *Nature Reviews Molecular Cell Biology*, vol. 11, no. 3. Pp. 220–228, Mar. 2010. Doi: 10.1038/nrm2858.
[126] L. Ke *et al.*, "Somatic Mutation Profiles Revealed by Next Generation Sequencing (NGS) in 39 Chinese Hepatocellular Carcinoma Patients," *Front Mol Biosci*, vol. 8, Jan. 2022, doi: 10.3389/fmolb.2021.800679.
[127] H. Tanaka and T. Watanabe, "Mechanisms Underlying Recurrent Genomic Amplification in Human Cancers," *Trends in Cancer*, vol. 6, no. 6. Cell Press, pp. 462–477, Jun. 01, 2020. Doi: 10.1016/j.trecan.2020.02.019.
[128] J. J. Moon, A. Lu, and C. Moon, "Role of genomic instability in human carcinogenesis," *Experimental Biology and Medicine*, vol. 244, no. 3. SAGE Publications Inc., pp. 227–240, Mar. 01, 2019. Doi: 10.1177/1535370219826031.

[129] N. Andor, C. C. Maley, and H. P. Ji, “Genomic instability in cancer: Teetering on the limit of tolerance,” *Cancer Res*, vol. 77, no. 9, pp. 2179–2185, May 2017, doi: 10.1158/0008-5472.CAN-16-1553.
[130] Y. Y. Wei Dai, “Genomic Instability and Cancer,” *J Carcinog Mutagen*, vol. 05, no. 02, 2014, doi: 10.4172/2157-2518.1000165.
[131] R. R. Madsen, “PI3K in stemness regulation: from development to cancer,” *Biochem Soc Trans*, vol. 48, pp. 301–315, 2020, doi: 10.1042/BST20190778.
[132] N. Jiang *et al.*, “Role of PI3K/AKT pathway in cancer: the framework of malignant behavior Endometrioid type of EC EMT Epithelial-mesenchymal transition,” vol. 47, pp. 4587–4629, 2020, doi: 10.1007/s11033-020-05435-1.
[133] O. Schon, A. Friedler, M. Bycroft, S. M. V. Freund, and A. R. Fersht, “Molecular mechanism of the interaction between MDM2 and p53,” *J Mol Biol*, vol. 323, no. 3, pp. 491–501, 2002, doi: 10.1016/S0022-2836(02)00852-5.
[134] N. A. Osman, A. I. Khalil, and R. K. Yousef, “The Clinical and Prognostic Implications of Pluripotent Stem Cell Markers Expression and Their Correlation with the WNT signal pathway in Hepatocellular Carcinoma,” *Asian Pacific Journal of Cancer Prevention*, vol. 21, no. 10, pp. 2961–2970, Jan. 2020, doi: 10.31557/APJCP.2020.21.10.2961.
[135] I. Pires, P. Wongtrakoongate, Y. Zhu, and J. Li, “Recent Advances in Liver Cancer Stem Cells: Non-coding RNAs, Oncogenes and Oncoproteins,” *Oncogenes and Oncoproteins. Front. Cell Dev. Biol*, vol. 8, p. 548335, 2020, doi: 10.3389/fcell.2020.548335.
[136] J. Ying, C. Shi, C. S. Li, L. P. Hu, and W. D. Zhang, “Expression and significance of SOX2 in non-small cell lung carcinoma,” *Oncol Lett*, vol. 12, no. 5, pp. 3195–3198, Nov. 2016, doi: 10.3892/ol.2016.5065.
[137] N. Karachaliou, R. Rosell, and S. Viteri, “The role of SOX2 in small cell lung cancer, lung adenocarcinoma and squamous cell carcinoma of the lung,” *Translational Lung Cancer Research*, vol. 2, no. 3. AME Publishing Company, pp. 172–179, 2013. Doi: 10.3978/j.issn.2218-6751.2013.01.01.
[138] Y. Chen, Y. Huang, Y. Huang, J. Chen, S. Wang, and J. Zhou, “The prognostic value of SOX2 expression in non-small cell lung cancer: a meta-analysis.,” *PloS One*, vol. 8, no. 8, 2013, doi: 10.1371/journal.pone.0071140.
[139] W. Shao, H. Chen, and J. He, “The role of SOX-2 on the survival of patients with non-small cell lung cancer,” *J Thorac Dis*, vol. 7, no. 7, pp. 1113–1118, 2015, doi: 10.3978/j.issn.2072-1439.2015.07.14.
[140] A. Sinha *et al.*, “Citation: Early-Stage Lung Adenocarcinoma MDM2 Genomic Amplification Predicts Clinical Outcome and Response to Targeted Therapy,” 2022, doi: 10.3390/cancers.
[141] M. J. Sanaei, S. Razi, A. Pourbagheri-Sigaroodi, and D. Bashash, “The PI3K/Akt/mTOR pathway in lung cancer; oncogenic alterations, therapeutic opportunities, challenges, and a glance at the application of nanoparticles,” *Translational Oncology*, vol. 18. Neoplasia Press, Inc., Apr. 01, 2022. Doi: 10.1016/j.tranon.2022.101364.
[142] T. Qi *et al.*, “Somatic Mutation Profiling of Papillary Thyroid Carcinomas by Whole-exome Sequencing and Its Relationship with Clinical Characteristics,” *Int J Med Sci*, vol. 2021, no. 12, pp. 2532–2544, 2021, doi: 10.7150/ijms.50916.
[143] R. Niciporuka *et al.*, “medicina Can We Predict Differentiated Thyroid Cancer Behavior? Role of Genetic and Molecular Markers,” 2021, doi: 10.3390/medicina57101131.

[144] P. Wang *et al.*, “SRY-related HMG box-2 role in anaplastic thyroid cancer aggressiveness is related to the fibronectin 1 and PI3K/AKT pathway,” *Mol Med Rep*, vol. 21, no. 3, pp. 1201–1207, 2020, doi: 10.3892/mmr.2020.10907.

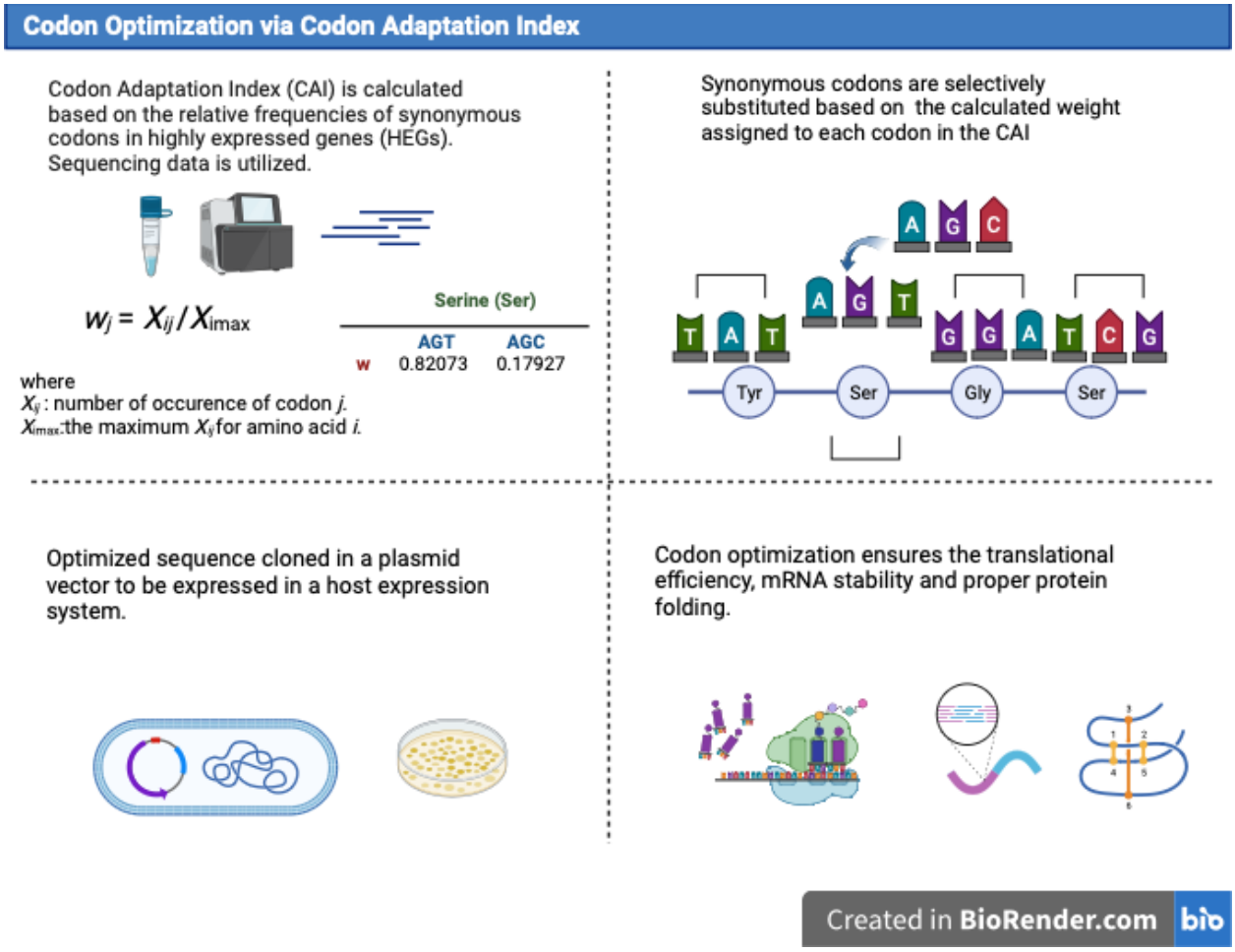
[145] G. Gauchotte *et al.*, “BRAF, p53 and SOX2 in anaplastic thyroid carcinoma: Evidence for multistep carcinogenesis,” *Pathology*, vol. 43, no. 5, pp. 447–452, 2011, doi: 10.1097/PAT.0b013e3283486178.

[146] E. Yokota *et al.*, “Targeted silencing of SOX2 by an artificial transcription factor showed antitumor effect in lung and esophageal squamous cell carcinoma,” 2017. [Online]. Available: www.impactjournals.com/oncotarget/

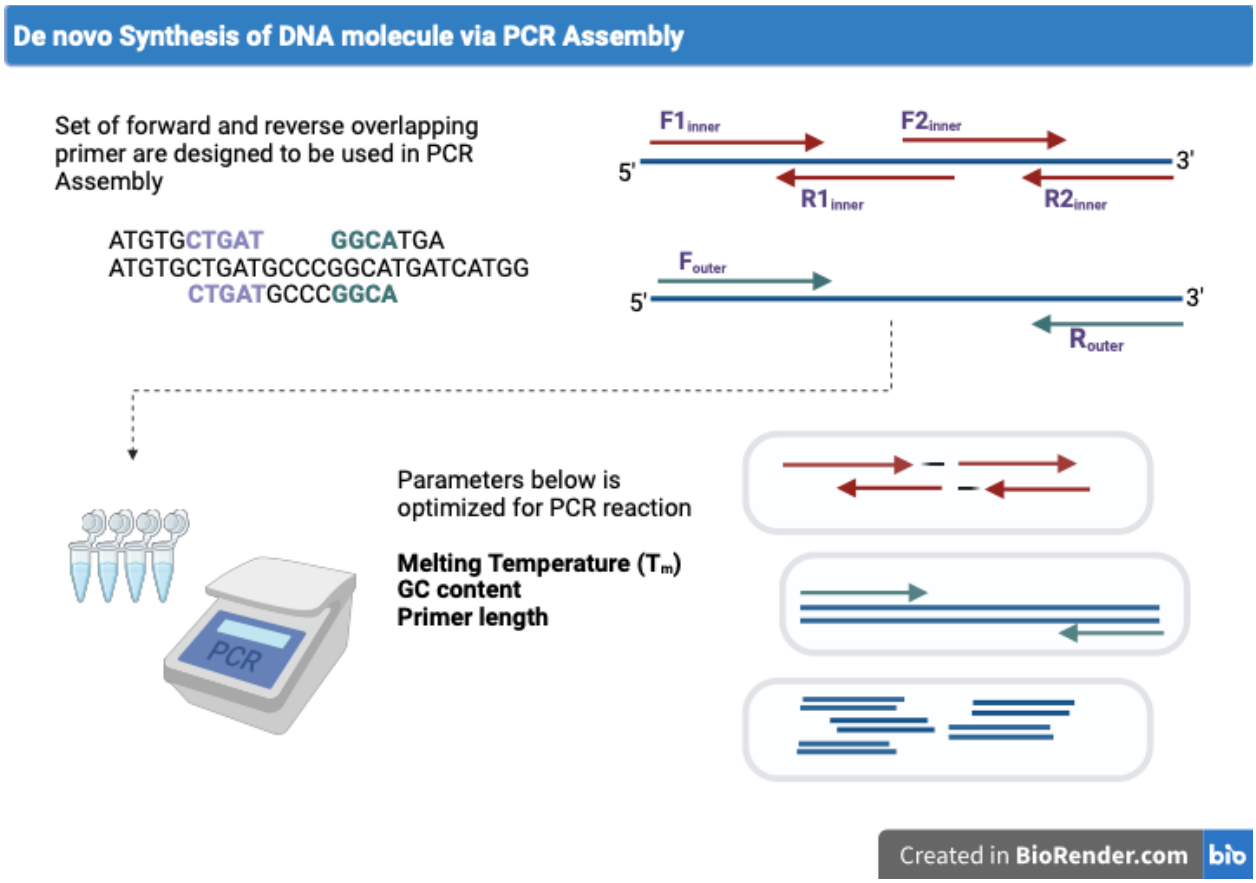
[147] G. D. Liu *et al.*, “Soluble expression of recomb inant cMyc, Klf4, Oct4, and Sox2 proteins in bacteria and transduction into living cells,” *Int J Ophthalmol*, vol. 10, no. 4, pp. 560–566, Apr. 2017, doi: 10.18240/ijo.2017.04.10.

[148] K. Liu, Y. Ouyang, R. Lin, C. Ge, and M. Zhou, “Strong negative correlation between codon usage bias and protein structural disorder impedes protein expression after codon optimization,” *J Biotechnol*, vol. 343, pp. 15–24, Jan. 2022, doi: 10.1016/j.jbiotec.2021.11.001.

Appendix



Appendix Figure 1. PCR assembly method is depicted in the figure. DNA molecules can be synthesis without tangible DNA with this application.



Appendix Figure 2. Codon adaption of the sequence based on Codon Adaptation Index (CAI) is summarized in the figure. (Created with BioRender.com)

Appendix Table 1. Detailed table for TCGA PanCancer cohorts that is analyzed for association analysis between SOX2 CNV and TP53 mutations. N-A: no association, S-A: strong association, W-A: weak association.

Cancer Type	**Sample Size**	**Association status**	**p-value**	**95% CI**	**Odds Ratio**
Esophageal Carcinoma	559	N-A	0.6595	(0.63, 2.17)	1.168
Glioblastoma	577	N-A	0.6205	(0.54, 3.67)	1.336
Liver Hepatocellular Carcinoma	372	S-A	1.6e-5	(1.7, 5.9)	3.155
Lung Adenocarcinoma	503	S-A	3.2e-6	(2.37, 6.26)	3.822
Pan Lung Cancer (TCGA, Nat Genet)	1144	S-A	2.2e-16	(5.28, 10.1)	7.276
Thyroid Carcinoma	500	S-A	9.2e-4	(1.62,7.55)	3.605
Brain Lower Grade Glioma	514	W-A	0.1861	(0.7, 5.28)	1.862
Ovarian Serous Cytadenocarcinoma	585	W-A	0.4193	(0.3, 7.9)	1.775
Breast Invasive Carcinoma	1084	S-A	2.2e-16	(4.4, 8.2)	5.973

Appendix Table 2. The results of association test for TP53 mutation status for cancer cohorts with SOX2/TP53 association. None of the cohort exhibits statistical association between TP53 mutation type and SOX2 gain/amplification.

Cancer Type	**Sample Size**	**Association status**	**p-value**	**95% CI**	**Odds Ratio**
Liver Hepatocellular Carcinoma	372	N-A	0.632	(0.7, 3.49)	1.155
Lung Adenocarcinoma	503	N-A	0.465	(0.23, 3.26)	1.867
Thyroid Carcinoma	500	N-A	0.677	(0.62, 2.25)	0.645
Breast Invasive Carcinoma	1084	N-A	0.789	(0.45, 1.89)	1.175