Characterization of Genomic and Coding Sequence Variation in the SKI2 Locus Following Whole Genome Duplication

Bioinformatic identification and characterization of coding sequence and stop codon variation using MAFFT and comparative alignment.

Maximiliam G Schylander

Bachelor Project • 15 hp

Swedish University of Agricultural Sciences, SLU

Faculty of Natural Resources and Agricultural Science

Crop Production (BSc) Program

Uppsala, 2026

Characterization of Genomic and Coding Sequence Variation in the SKI2 Locus Following Whole Genome Duplication

Identification and characterization of coding sequence and stop codon variation using MAFFT and comparative alignment.

Maximiliam G Schylander

Supervisor: Levi Yant, SLU, Department of Plant Biology
Examinator: Adrien Sicard, SLU, Department of Plant Biology

Credits: 15 hp
Level: G2E
Course title: Independent project in Biology
Course code: EX0894
Program/education: Crop Production (BSc)
Course coordinating dept: Plant Biology
Place of publication: Uppsala
Year of publication: 2026
Cover picture: DNA illustration by Geralt, from Pixabay. Licensed under the Pixabay Content License.

Series: Examensarbete / Institutionen för växtbiologi, SLU

Part number: 216

Electronic publication: https://stud.epsilon.slu.se

Keywords: Whole genome duplication (WGD), polyploidy, adaptive mutations, RNA helicase, SKI2 protein, protein structure

Preface

This thesis is a bachelor's degree project in Biology at the Swedish University of Agricultural Sciences (SLU) in Uppsala. The project corresponds to 15 HP credits and was carried out during the spring semester of 2026 as part of a Biology bachelor's degree. The topic of this study emerged from an interest in bioinformatics and genetics. The project was introduced to me by the researcher and my supervisor Levi Yant.

Polyploidy is a common phenomenon in the plant kingdom and has been shown to contribute to increased genetic diversity and evolutionary innovation. Despite its importance, many aspects of how individual genes and proteins adapt following genome duplication events remain insufficiently understood.

This study aims to explore potential signatures of adaptive change in candidate proteins in autopolyploid plants. The project focuses on identifying amino acid substitutions that may have arisen or been selected after genome duplication and examining whether these occur within conserved or functionally important protein domains.

The work presented in this thesis was carried out by Maximiliam G. Schylander.

Acknowledgements

I would like to express my gratitude to my supervisor **Levi Yant** for his guidance and support through this project. His advice on design, introduction, and scientific writing was invaluable, and his support greatly contributed to the creation of this thesis.

I would also like to thank **Sian Bray** and **Adrien Sicard**. Sian for her assistance with the structural biology aspects of this project. Her help with AlphaFold and several bioinformatics tools was extremely helpful and allowed me to better understand and analyse the structural implications of the protein mutations investigated in this study. I also thank **Adrien** for his work as my examiner and for his helpful feedback on the text.

1. Summary

This study investigated amino acid mutations in the SKI2 protein between diploid and tetraploid Mimulus plants following whole genome duplication (WGD) (See figure 1). By comparing protein sequences using multiple sequence alignment and analyzing predicted AlphaFold structures. 12 tetraploid-specific amino acid changes were found. Most changes were conservative and located in less functional regions, but three substitutions (S584Y, L918Q, and T1331M) stood out as strong candidates for potential functional changes.

Abstract

Keywords: Whole genome duplication (WGD), polyploidy, adaptive mutations, RNA helicase, SKI2 protein, protein structure

Table of Contents

Abbreviations and Definitions

AA - Amino acid
ATP - Adenosine triphosphate
DNA - Deoxyribonucleic acid
MSA - Multiple sequence alignment
RNA - Ribonucleic acid
WGD - Whole genome duplication
SKI2 - RNA helicase protein involved in RNA degradation as part of the Ski complex
RNA helicase – An enzyme that unwinds RNA molecules using ATP
Polyploidy – A Condition in which an organism contains more than two complete sets of chromosomes
Diploid – An organism with two sets of chromosomes
Tetraploid - Organism with four sets of chromosomes
Gene duplication – A process in which a gene is copied within the genome
Ortholog - Genes in different species that evolved from a common ancestral gene
Domain - Conserved region of a protein associated with a specific function
PDB - Protein Data Bank
InterPro - Database used to identify protein families and functional domains
AlphaFold - Artificial intelligence system used to predict protein structures
BLOSUM62 - Substitution matrix used to evaluate similarity between amino acid substitutions
FASTA - Text-based sequence file format used for nucleotide or protein sequences
P-loop - Phosphate-binding loop motif associated with nucleotide (e.g. ATP) binding in many ATPases and helicases
NTPase - Nucleoside triphosphatase; enzyme domain that hydrolyzes nucleoside triphosphates such as ATP
SUPV3-like helicase - Annotated RNA helicase-related domain identified in SKI2
MTR4 - Ski2-like RNA helicase associated with RNA surveillance and the exosome. Used here as a homologous reference domain annotation
Exosome – A multi-protein complex involved in RNA processing and degradation
TSV – Translated site variant
Resi - Residue
Acidic – An amino acid with a negatively charged pH
Basic – An amino acid with a positively charged pH
Polar – An amino acid interact with water or form hydrogen bonds.
Hydrophobic – An amino acid tends to avoid water.
PyMOL - Molecular visualization software to inspect and see protein structures.

2. Introduction

2.1 Evolution and genetic variation

The diversity of life on Earth is, for the most part, explained through the processes of evolution and natural selection. Charles Darwin first proposed the theory of evolution by natural selection, and since then, scientists have understood that variation among individuals within a population provides the material through which natural selection can act. Traits that improve an organism's survival rate or reproduction success tend to become more common in populations over time (Darwin,1871).

The theory of evolution by natural selection proposes that all living organisms share a common evolutionary origin and are built from similar biochemical components. Modern-day biology has demonstrated and proved that all life on earth is based on a universal genetic code. This suggests that all organisms could descend from a common ancestor. Over time, evolution has happened, and genetic variation has allowed populations to adapt to changing environments and ecological conditions (Hillis et al. 2011, pp. 2-9).

Genetic variations are thought to arise through several different mechanisms. Some of these are substitutions, genomic recombination's, and gene duplications. Among these processes, gene duplications have been recognized as an important source of evolutionary innovation. By creating additional copies of genetic material, the duplication event provides opportunities for genes to change and acquire new functions without disrupting biological processes. Mutations provide new raw material for evolution by creating variety (Hillis et al. 2011, pp, 256-257).

One of the most dramatic forms of gene duplication occurs through whole-genome duplication (WGD). This is where an organism gains an additional complete set of chromosomes. Events like this can significantly alter genomic structure and create new evolutionary opportunities by significantly increasing the amount of genetic material available for diversification and adaptation. WGD is a mutation that is dramatic in its course, where the wheat is separated from the chaff. But research on these latest genetic variations is still in its early stages, where both positive and negative outcomes still need to be researched and found (Hämälä et al. 2024, p. 1).

2.2 Whole genome duplication and polyploidy

Whole genome duplication (WGD), also known as polyploidization, is a process in which an organism creates additional complete sets of chromosomes. In diploid organisms, cells normally contain two sets of chromosomes, one inherited from each parent. But sometimes errors occur during cell division, particularly during meiosis or mitosis, and it can result in the duplication of entire chromosome sets, creating individuals with multiple full.genome copies (Hillis et al. 2011, pp. 224-234).

Polyploidy is especially common in plants and has played a significant role in plant evolution). Many plant lineages have experienced one or more WGD events during their evolutionary history. These events can generate large amounts of genetic changes, creating new opportunities for evolutionary adaptation or innovation (Heslop-Harrison et al. 2023, p. 1).

After a whole genome duplication event, organisms suddenly possess multiple copies of each gene. While some duplicated genes may eventually be lost or become nonfunctional, others can accumulate mutations and diverge in function over time and thrive. This can lead to the evolution of new gene functions, or changes in the gene regulation, increasing the genetic and functional diversity of the organism and helping it. It is an everlasting question how WGD and its ways to adapt to new environments will influence evolutionary biology (Hämälä et al., 2024, p. 6).

The evolutionary importance of gene duplication was first emphasized by Susumu Ohno, who proposed that duplicated genes provide the raw material for evolutionary novelty (Ohno, 1970). According to this idea, while one copy of a duplicated gene maintains its original function the additional copy may accumulate mutations that can lead to new biological functions without compromising essential cellular processes (Ohno, 1970).

Research has shown that whole-genome duplication events are associated with increased diversification in flowering plants. Polyploidy may allow organisms to tolerate environmental stress and adapt to new ecological niches by increasing genetic variation and providing additional opportunities for evolutionary change (Soltis et al., 2009).

Despite the potential advantages associated with polyploidy, the sudden duplication of the entire genome can also create challenges for cellular processes. The presence of multiple gene copies may disrupt gene regulation, protein interactions, and genome stability. Understanding how organisms adapt to these

genomic changes remains an important and ongoing question in evolutionary biology.

2.3 SKI2 and Ski2-like RNA helicases

SKI2 is a member of the Ski2-like family of ATP-dependent RNA helicases whose purpose, among other things, is to identify RNA that is not working as it should; its main purpose is to clean and unwind threads in the RNA that have been "entangled" and need to be untangled or cleaned. It is a group of proteins involved in RNA surveillance, RNA processing, and RNA degradation. In many organisms, Ski2-like helicases function together with the RNA exosome and play important roles in directing RNA substrates toward degradation. Structural and functional studies of Ski2-like helicases, including SKI2 and it's homolog MTR4 has shown that these proteins contain a conserved helicase-associated regions involved in ATP binding, RNA binding, and ATP dependent RNA translocation (Johnson & Jackson, 2013, Quinton et al., 2021). The functional part of Ski2-like helicases is formed by two RecA domains. These domains are responsible for ATP binding and ATP movement along the RNA. Outside the core there is an arch and helical region that helps guide RNA, support the structure, and affect how the protein changes shape during RNA handling. This makes Ski2 helicases useful for studying whether amino acid substitutions happen in parts of the protein that may affect function (Halbach, F. 2012). The study by Quinton et al (2021) identified proteins necessary for cells that have gone through WGD which are vital for cell survival. This result opens up for studies about new cancer treatments. Ski2-like helicases are therefore highly suitable candidates for comparative analyses of amino acid substitutions in relation to protein function.

SKI2 was selected as a candidate protein in this study since previous work from the Yant group laboratory identified SKI2 as a strong candidate gene, showing peaks in amino acid divergence between diploid and tetraploids in the species *Mimulus Guttatus*. Because SKI2 contains multiple conserved domains associated with helicase activity and RNA interaction, it provides a useful system for investigating whether whole genome duplication may be associated with functionally meaningful protein divergence. By examining amino acid substitutions in relation to conserved domains and known functional regions, it is possible to identify candidate mutations that may be associated with adaptive or functionally relevant divergence following whole genome duplication. This means in plain English that researchers are looking for positive selections as the outcome from whole genome duplications. A positive selection can be that the cell finds a new way to function which is more adapted to the polyploid state, and it is important to understand why the change is to the advantage for the cell. The

method is used to map out what the change looks like and also why the adaptation is to an advantage (Yant et al., 2013)

2.4 Aim and research

The aim of this study is to investigate adaptive change in candidate proteins following whole genome duplication in autopolyploid plants. By comparing protein sequences between diploids, tetraploids and well documented species this study seeks to identify amino acid substitutions that may have appeared after genome duplication and make hypotheses if these changes occur in conserved or functionally important protein regions.

2.5 Research questions

This study addresses the following research questions:

1. What are the non-synonymous changes in the coding region?
2. Do tetraploidspecific mutations occur within conserved or functionally important regions of the proteins?
3. Could these amino acid substitutions potentially influence protein structure or function?

3. Materials and Methods

3.1 Sequence collection

Protein sequences for the candidate gene SKI2 were given by the Yant laboratory at the Swedish University of Agricultural Sciences (SLU). The dataset included sequences extracted from both diploid and tetraploid Mimulus to identify mutations that could be associated with whole genome duplication.

To give an evolutionary context and see sequence conservation in homologous genes of SKI2, protein sequences from multiple model organisms were included. The model organisms are well documented. These model organism species were: *Arabidopsis thaliana, Homo sapiens, Mus musculus, Rattus norvegicus, Saccharomyces cerevisiae,* and *Caenorhabditis elegans* taken from PDB 2026/03/12.

Well-studied organisms that are distantly related were included, which helped to identify conserved regions in the protein. These regions areas served as references

to assess whether amino acid changes occurred in important regions. Literature and PDB databases were used to study these positions. By comparing known functions, a hypothesis could be formed about how changes in Mimulus affect protein structure. Kögel et al. (2023) was used as a reference to identify different regions in the alignment and to compare functionally important sites. By using PDB, it is possible to visualise where in the 3D physical structure the amino acid replacements are placed (Berman et al., 2000).

3.2 Multiple sequence alignment

Protein sequences were aligned using the MAFFT multiple sequence alignment algorithm, which was run in a Linux (Ubuntu 22.04 LTS) environment. The alignments were then imported into Geneious (version 2026.0.2) and Jalview (version 2.11.5.1) to help visualize and inspect. These tools were used to examine sequence conservation and to compare amino acid differences (Katoh & Standley, 2013).

The alignments were manually inspected to ensure that conserved regions and functional domains, were aligned. Special attention was given to positions where amino acid changes differed between diploid and tetraploid Mimulus sequences. Mutations occurring within known functional regions of the protein were considered of a higher priority of interest. PDB was used for structural reference, while published literature was used to interpret functional importance.

By comparing the locations of observed amino acid mutations with functional domains it was possible to prioritize mutations in regions likely to influence protein structure and/or function. These positions were analysed as candidate mutations potentially associated with WGD.

3.3 Domain annotation

Functional domains within the SKI2 protein were identified using domain prediction tools and previously published literature on RNA helicases. Domain functions were obtained from databases such as InterPro (Paysan-Lafosse et al., 2023) and Protein Data Bank (PDB) resources describing the protein's functions in detail in other more well documented species.

Particular attention was given to conserved domains characteristic of Ski2like RNA helicases, like the DExH helicase core, the arch domain, and the Cterminal MTR4like region which are known to play roles in ATP binding, RNA binding, and helicase activity.

3.4 Identification of amino acid substitutions

Amino acid substitutions were identified through analysis of the aligned sequences. Positions were examined to detect mutations present in tetraploid Mimulus sequences but not in the diploid reference sequence.

Candidate changes in amino acids were annotated and evaluated based on their location within the protein sequence. Priority focus was given to mutations occurring within conserved regions or known functional domains as these changes may have potential functional significance.

3.5 Analysis of Amino Acid Substitutions and Structural Impact

To identify mutations with a high potential for functional impact, amino acid changes in the SKI2 protein were first evaluated based on their physicochemical difference. Changes were classified as either conservative or non-conservative using amino acid property classes (e.g., acidic, basic, polar, and hydrophobic) and established metrics such as the BLOSUM62 matrix (See Figure 2) and the Genetic code table (See Figure 3). This allowed for the prioritization of mutations likely to cause significant evolutionary or chemical differentiation.

3.6 Alphafold and PyMOL

To further investigate these findings, three-dimensional structures of the SKI2 protein were used. Protein homology models were created using AlphaFold version 3.0 on the Czech national HPC MetaCentrum. The reduced database was used with a model preset of monomer and a maximum template data of 2021-09-30. (Varadi et al., 2022). Two protein homology models were created using AlphaFold version 3.0 on the Czech national HPC MetaCentrum. The reduced database was used with a model preset of monomer and a maximum template data of 2021-09-30. Separate models were produced for the diploid (2n) and tetraploid (4n) Mimulus lineages. This made it possible to compare how amino acid mutations were positioned in the predicted protein structures. The models were then imported into The PyMOL Molecular Graphics System, Version 3.0 Schrödinger, LLC for visualization and inspection. In PyMOL the positions of candidate amino acid mutations were mapped onto the predicted structures and examined one by one (see table 2, figure 4, figure 5). This was done to see whether the mutations were located in exposed outer regions, structured domains, or close to known functional regions such as catalytic sites. Special attention was given to mutation that were near the helicase core, ATP-binding motifs, RNA-interaction surfaces or other conserved regions previously identified from domain

annotation and published studies of SKI2-like helicases. The PyMOL models were also used to assess if residue appeared to face outward toward the protein surface or inward toward the protein core. These observations were used to support interpretations and to help make a hypothesis of whether the mutations changed the functional part of the protein.

4. Results

4.1 Amino acid substitutions identified in SKI2

A comparison of the SKI2 protein sequences from diploid and tetraploid Mimulus plants was made. 12 high frequency changes were identified in the proteins' amino acids. Some of these changes are in the functional regions of Ski. While others are in the outer areas according to the InterPro prediction (Table 1).

Multiple mutations were found within functional regions. These include the SUPV3, P-loop regions and the C-terminal MTR4 segment. Focus was placed on these changes because they occur in conserved regions. Mutations in these areas are more likely to alter how the protein functions compared to changes in less important regions.

4.2 Structural mapping of substitutions in predicted SKI2 models

In the predicted AlphaFold structure the SKI2 has the an wild shape (see figure 4 and 5) similar to other Ski2-like RNA helicase. The functional core of the protein is formed by the RecA1 and RecA2 domains. These two domains are probably responsible for ATP binding, ATP movement and RNA handling. Outside of this core domain SKI2 has an N-terminal domain, an arch-associated domain, and a helical domain. These parts help guide RNA, support the structure and affect how the protein changes shape. The model was used to see where the amino acid mutations are located. They are spread across several parts of the SKI2 protein in the N-terminal, the RecA1, RecA2 core, the arch, and the helical region. Some of these changes are conservative and located on the protein's outer surface. This means the amino acid changed keeps similar properties and is less likely to strongly change protein structure or function. Some mutations caused clear chemical changes and were in more important structures and functional parts of the protein.

The M138L mutation is in the N-terminal region. This is a conservative change between two hydrophobic amino acids (BLOSUM62 = 2). It is not expected to

significantly change the chemical properties. In the AlphaFold model residue 138 is on a small helix and is faced outward (See figure 6). Comparing with the human SKIV2L reference shows that this region is near the hSKI2 wedge segment (specifically residue W146 and G147). This means M138L is located near a regulatory region.

The E271D change is also in the N-terminal region. This is a conservative change between two acidic amino acids (BLOSUM62 = 2). In the AlphaFold model residue 271 is on a flexible loop and is faced away from the protein(See figure 7). No functional residues were found in the human SKIV2L reference that is close to this area.

In the RecA1 region of the SKI2 the I468M change was found. This is a conservative change between two hydrophobic amino acids (BLOSUM62 = 1). Even if this is a small change the AlphaFold model shows that residue 468 is in an important position on the protein structure. It is on a loop at the surface, and it is faced toward the helicase (See figure 8). In comparison with human SKIV2L shows that many important residue (such as I402 and S405) are located closely to this mutation. In humans, these residue are linked to RNA handling and the state of the protein's normal structure. This probably means that I468M is located in a functional part of the protein.

The S584Y change is in the RecA2 of the SKI2. This is a non-conservative change where a small polar residue is replaced by a larger aromatic residue (BLOSUM62 = -2). It is one of the more chemically stronger changes found in the tetraploid sequences. In the AlphaFold model, residue 584 is on a loop at the proteins surface and is faced outward (See figure 9). When compared to the human SKIV2L reference the closest functional residue is hSKI2 R483. This residue is part of a hotspot for ATPase function. Although S584Y does not directly sit on a known functional site its location could mean it occurs in a structurally and functionally important part of the RecA2.

The E591D change is also located in the RecA2 region. This is a conservative change between two acidic amino acids (BLOSUM62 = 2), meaning that it is only a small chemical effect. In the AlphaFold model residue 591 is on an exposed loop on the surface of the protein (See figure 10). It is faced outwards rather than toward the core that looks almost as it is buried. This region is near the hSKI2 R483 hotspot in the human reference. E591D does not overlap a known functional region.

In the arch region was a E897D mutation. This is a conservative change between two acidic amino acids (BLOSUM62 = 2) meaning a small chemical effect. In the

AlphaFold model residue 897 is on a helical element and is faced outward from the protein (See figure 11). When comparing this site to the human SKIV2L reference we found no functional residue in the area. The closest known functional sites are located further away.

The L918Q change is also located in the arch. This is a non-conservative change from a hydrophobic residue to a polar uncharged residue (BLOSUM62 = -2). It is one of the most chemically different changes found. In the AlphaFold model residue 918 is on a helical segment that connects the arch to the rest of the protein (See figure 12). Comparison with the human SKIV2L reference shows that the nearest functional site is R888 (equivalent to Mimulus 976). L918Q does not really overlap this site. but it clearly introduces a chemical change in the arch region.'

Several additional changes were found in the arch region forming a cluster in part a of the protein. One is V1004A it is located on a distant loop away from the center of the protein. This is a conservative change where a valine is replaced by a smaller alanine. This means only a small chemical effect. In the AlphaFold model the side chain is faced inward (See figure 13). This indicates that the residue may still be important for local structural bonding. When compared to the human reference the nearest functional site is pretty distant. No functional residues were found directly in this position.

The V1015I change is located nearby in the same arch. This is a conservative change between two hydrophobic amino acids (BLOSUM62 = 3) meaning very little chemical change. AlphaFold places residue 1015 on a loop within the protein in an accessory region. It is faced inward toward the surrounding structure and the helicase (See figure 14). Similar to V1004A no functional residue's were found in the human reference in this position.

The V1026A change is also in the arch. This is a conservative change where a valine is replaced by a smaller alanine (BLOSUM62 = 0). This means there is only a minor change in the protein's properties. In the AlphaFold model residue 1026 is part of a structure rather than a loop (see figure 15). It sits close to a nearby helical segment. Like the previous two sites no functional residue was found in the human reference close to this position.

The V1043I change is also located in the arch. This is a conservative change between two hydrophobic amino acids (BLOSUM62 = 3) meaning a very small chemical effect. In the AlphaFold model residue 1043 is positioned on a loop and is faced outward (See figure 16). As with the other sites, no functional residues were in the human reference that directly overlap this position. Together,

V1004A, V1015I, V1026A and V1043I form a cluster in the same region. They are all located in outer parts of the protein relatively distant from any mapped functional parts.

The final mutation T1331M is located in the helical domain. This is not a very conservative change from polar to a hydrophobic amino acid (BLOSUM62 = -1). This shows a clear change in the chemical properties of the site. In the AlphaFold model residue 1331 sits on a helical element and is faced outward (See figure 17). According to the human SKIV2L reference this region is near surfaces used for RNA interaction and the RNA-exit area. Specifically, it is close to residue's that correspond to positions 1308 and 1318 in the human protein. T1331M occurs in a structured part of the helical domain right next to a functionally important region.

To sum up the result of the AlphaFold mapping. The most significant chemical changes occur in the RecA2, arch, and helical domains. Other, more conservative changes form a cluster in the arch-associated accessory region. Overall, the mutations range from small changes in exposed outer areas to more important mutations in the organized parts of the protein's framework.

5. Discussion

The structural analysis shows that the changes in the tetraploid sequences do not all have an equally big effect on SKI2's function. Most mutations are conservative and located in outer regions. These changes are less likely to have an impact on SKI2's function. In comparison to more significant chemical changes that are located in better characterized regions. These sites are stronger candidates for explaining how the protein's function could have changed.

The M138L mutation is a conservative change. Based on its chemistry alone it is unlikely to have an effect on SKI2 function. The residue is faced outward meaning it does not disrupt the protein's core. But its position in the N-terminal region is close to the hSKI2 wedge segment. In humans, this segment is part of a module that helps the protein handle RNA and separate it. M138L is a lower priority candidate, and the mutation probably doesn't affect the protein's function at all. But it is possible it could affect the separation of the RNA through subtle changes in the protein's flexibility.

Even if I468M is a conservative change, its structural position says it could still be functionally relevant. The change is in the RecA1 helicase on a loop turned toward the protein's center. This means that it may affect local packing or how

different parts of the protein connect rather than being a simple change on the surface. I468M is close to several residue's (like as I402 and S405) that are known to be important in human SKIV2L. In humans these residues help the protein handle RNA and change its shape. While I468M likely does not stop the protein from working it could in some ways change how the RNA-handling region moves. This could affect how efficiently the protein uses ATP to move along the RNA strand.

S584Y is one of the most interesting mutations in the study. It is a large chemical change and in an interesting location in the conserved helicase. Replacing serine with tyrosine means a much larger side chain with different hydrogen bonding abilities. This makes it more likely to affect the protein structure than the smaller changes seen in the other places. The AlphaFold model shows that residue 584 sits on an outward-facing loop rand not deep inside the core but this doesn't mean that it is neutral. In Ski2-like helicases, loops can affect how the protein changes shape, how the RNA is formed or its bonding abilities. Because S584Y is located in the RecA2 close to an important site in human SKIV2L reference, it could affect how efficiently the protein moves along the RNA strand by changing the behavior of the loop or the chemistry of its surface.

By itself E591D a very strong candidate in this study. The mutation is chemically conservative keeping the same acidic and negative charge. The AlphaFold model also shows that the residue is located on an outward facing loop instead of deep inside the helicase. These features could mean it does not have a direct effect on protein stability or function. But its position in the RecA2 might still be important in a broader context. Surface loops in helicase domains can contribute to flexibility or how the surface interacts during RNA shift. E591D is best described as a little change. Its importance likely depends more on it being in the same important region as other, stronger changes such as S584Y but on its own E591D likely has no important function.

E897D is unlikely to be a functional change. The mutation is chemically conservative. The AlphaFold model shows the residue is in an outward facing position on a helical element. No important residue's from the human SKIV2L reference were found closely to this region. The most likely result is that E897D has little direct effect on SKI2's function. It could be a minor change to the protein's surface or bonding. While the arch region can be important, the structural evidence does not mean that E897D is a high-priority candidate compared to the other changes.

L918Q is one of the most interesting mutations in this study and a strong candidate. It is a significant chemical change with a location in an important part

of the arch. Replacing leucine with glutamine adds polarity to a spot that is normally hydrophobic. This change could affect how the protein folds or how the segment behaves. L918Q might have an impact on the structure. It is also in the broader arch region near the KOW fold that helps guide RNA and regulate how the protein changes shape during RNA handling. This makes this region more interesting. L918Q could affect how the arch region is positioned or how it moves during RNA handling. Because the arch helps regulate the protein, L918Q is one of the most exciting changes in our data. A possible hypothesis is that this mutation affects how the arch region is held together or how it moves rather than directly changing catalytic activity.

The mutations between positions 1004 and 1043 are best seen as a together instead of as individual mutations. In the alignment these changes are grouped together in the same arch associated region. Most of them are conservative changes and on their own they probably do not have any effect on SKI2's function. But because they close together in the same part of the protein they may still be important as a combined force of change. A possible hypothesis is that these mutations together affect how this arch-associated accessory region is held together or how it moves. This could cause small changes in local folding flexibility or the positioning of this part during RNA handling. Even if each mutation by itself makes little difference. The combined effect of several changes in the same region could create a structural change in this part of the protein.

T1331M is one of the most interesting mutations in this study. Together with L918Q it is one of the strongest candidates. It is both a clear chemical change in an important region and near a functional part of the helical domain. Replacing threonine with methionine removes a polar side chain. This adds a larger hydrophobic one. This changes the surface properties and could change how the helix interacts with other parts of the protein. The AlphaFold model shows that the residue is outward, it might still be functionally important. This is very important since it is in a region linked to RNA interactions and the RNA exit path in the human protein. While T1331M is probably not going to be a direct catalytic mutation, it could influence the helicase surface chemistry or the way the protein changes shape during its work. Because of its chemistry and its location, T1331M stands out as one of the most exciting changes in my data.

In summary, the strongest candidates in this study are S584Y, L918Q, and T1331M. If taking all amino acid substitutions in tetraploid Mimulus SKI2 it probably hasn't lost its functionality. They suggest more subtle tuning after whole-genome duplication. Most of the substitutions are conservative and are

located in exposed or outer regions of the protein. This means they are less likely to strongly affect function. A smaller number of the substitutions involve bigger chemical changes and are in more functionally important domains. The main ATP-dependent RNA helicase function of SKI2 is probably still the same but the parts of its RNA handling may have changed. This could affect RNA binding, RNA movement along the protein, unwinding or how the protein changes shape in RNA handling. This can be important after whole-genome duplication where cells have to handle changes in gene amounts and how much RNA is produced. Because SKI2 is involved in RNA surveillance and RNA degradation even small changes of its efficiency or regulation ability could help the cell maintain RNA stability when it's polyploid (Halbach, F. 2012).

5.1 Future studies

Future studies should investigate these candidate changes in wet lab experiments to test if they have functional effects on the SKI2 protein. This could include testing how the strongest candidate mutations affect RNA handling, ATP- activity, protein stability or helicase function. More structural analysis can be done by examining charge and hydrophobicity in the AlphaFold. This can give a better understanding of how the changes may change local surface chemistry, structural interactions or movement in important regions of the protein. Together, this would determine whether the predicted structural differences have real biological effects.

6. References

Berman, H. M., Henrick, K., & Nakamura, H. (2003). Announcing the worldwide Protein Data Bank. Nature Structural Biology, 10(12), 980. DOI: 10.1038/nsb1203-980.

BLOSUM62: Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. PNAS, 89(22), s. 10915-10919.

Darwin, C. (1871/1913). *On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life*. Stockholm: Albert Bonniers förlag. Available via: Litteraturbanken. https://litteraturbanken.se/forfattare/DarwinC/titlar/ArternasUppkomst/sida/1/etext (Hämtad: 2026-02-19).

Grantham distance: Grantham, R. (1974). Amino acid difference formula to help explain protein evolution. Science, 185(4154), s. 862-864.

Halbach F, Rode M, Conti E. (2012). The crystal structure of S. cerevisiae Ski2, a DExH helicase associated with the cytoplasmic functions of the exosome. The EMBO journal, doi: 10.15252/embj.2018100640

Heslop-Harrison, J. S., Schwarzacher, T. & Liu, Q. (2023). Polyploidy: its consequences and enabling role in plant diversification and evolution. Annals of Botany, 131(1), s. 1–10. https://doi.org/10.1093/aob/mcac132.

Hillis, David M., Heller, H. Craig, Hacker, Sally D., Hall, David W., Laskowski, Marta J. & Sadava, David E. (2020). Life: the science of biology. Twelfth edition. Sunderland, MA: Sinauer Associates/Macmillan

Hämälä, T., Moore, C., Cowan, L., Carlile, M., Gopaulchan, D., Brandrud, M. K., Birkeland, S., Loose, M., Kolář, F., Koch, M. A. & Yant, L. (2024). Impact of whole-genome duplications on structural variant evolution in Cochlearia. Nature Communications, 15(1), artikel nr: 5377. doi.org (https://www.nature.com/articles/s4146702449679y

Johnson, S.J. & Jackson, R.N. (2013). Ski2-like RNA helicase structures: common themes and complex assemblies. RNA Biology, 10(1), ss. 33–43. doi: 10.4161/rna.22101.

Katoh, K. & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Molecular Biology and Evolution, 30(4), s. 772–780.

Kögel, A., Keidel, A., Bonneau, F., Schäfer, I.B. och Conti, E. (2023). The human SKI complex regulates channeling of ribosome-bound RNA to the exosome via an intrinsic gatekeeping mechanism. Molecular Cell, 83(11), ss. 1836-1849.

Paysan-Lafosse, T., et al. (2023). InterPro in 2022. Nucleic Acids Research, 51(D1), s. D418–D427.

Selvitopi, R. & Ekanayake, Saliya & Guidi, Giulia & Pavlopoulos, Georgios & Azad, Ariful & Buluç, Aydin. (2020). Distributed Many-to-Many Protein Sequence Alignment using Sparse Matrices. 10.48550/arXiv.2009.14467.

Soltis, D.E., Albert, V.A., Leebens-Mack, J., Bell, C.D., Paterson, A.H., Zheng, C., Sankoff, D., Depamphilis, C.W., Wall, P.K. och Soltis, P.S. (2009). Polyploidy and angiosperm diversification. American Journal of Botany, 96(1), ss. 336–348. Tillgänglig via: Wiley Online Library (Hämtad 2026-02-19)

Ohno, S. (1970). Evolution by Gene Duplication [Elektronisk resurs]. Springer Berlin Heidelberg

Quinton, R.J., DiDomizio, A., Vittoria, M.A., Kotov, A.S., Konotop, M.U., Passalacqua, A.H., Helland, J., Olsen, R.R., et al. (2021). Whole-genome doubling confers unique genetic vulnerabilities on tumour cells. Nature, 590(7846), ss. 492–497. doi: 10.1038/s41586-020-03133-3

Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yushchenko, G., Huo, L., Salami, C., Nayak, D.K., Bujotzek, R., Lavigne, G., Kara, J.S., Senf, A., Akdel, M., Hegedűs, T., Gábor, I., Áron, S., Lowy, D.A., Pfeiffer, B., ... & Velankar, S. (2022). AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Research, 50(D1), s. D439–D444.

Yant, L., Hollister, J.D., Wright, K.M., Arnold, B.J., Higgins, J.D., Franklin, F.C. och Bomblies, K. (2013). Meiotic Adaptation to Genome Duplication in Arabidopsis arenosa. Current Biology, 23(21), pp. 2151–2156. doi: 10.1016/j.cub.2013.08.059.

7. Popular summary

Whole genome duplication (WGD) is when an organism gains one or more of a full set of chromosomes. This is a common event in plants and can create both opportunities but also disadvantages. This project looked at the protein SKI2 in *Mimulus Guttatus* plants to see if they had changed after going though WGD. The SKI2 protein that was looked at is involved in RNA degradation. Meaning its function is to remove faulty RNA.

Most of the amino acid changes in the tetraploid SKI2 were small and probably doesn't affect protein function. A few of the mutations stood out as stronger candidates. These were S584Y, L918Q, and T1331M. These may have caused subtle changes in how the protein functions. To summarize the results suggest that SKI2 has probably kept its main function after WGD but might have gone through small changes that could be important to adapting to a polyploid state.

8. Figures

Table 1

Identified amino acid mutation changes between diploid and tetraploid SKI2 protein sequences. Summarizes amino acid position, residue change, chemical change, BLOSUM62 score, and predicted domain location.

Position	Diploid AA	Tetraploid AA	Hydrophobic / Polarity Change	BLOSUM62 Score	Zone / Domain	Priority
138	M	L	Hydrophobic → Hydrophobic (no major polarity change)	2	Outside annotated domain	Low
271	E	D	Acidic polar (negative) → Acidic polar (negative)	2	ATPdependent RNA helicase SUPV3like	Medium

468	I	M	Hydrophobic → Hydrophobic (no major polarity change)	1	ATPdependent RNA helicase SUPV3like / DEADDEAHbox helicase domain	Medium
584	S	Y	Small polar uncharged → Bulky aromatic, weakly polar	-2	ATPdependent RNA helicase SUPV3like / Ploop containing nucleoside triphosphate hydrolase	Very High
591	E	D	Acidic polar (negative) → Acidic polar (negative)	2	ATPdependent RNA helicase SUPV3like / Ploop containing nucleoside triphosphate hydrolase	Medium
897	E	D	Acidic polar (negative) → Acidic polar (negative)	2	ATPdependent RNA helicase SUPV3like	Medium
918	L	Q	Hydrophobic → Polar uncharged	-2	ATPdependent RNA helicase SUPV3like / near Exosome RNA helicase MTR4like, betabarrel domain	High
1004	V	A	Hydrophobic → Hydrophob	0	ATPdependent RNA helicase SUPV3like / near Exosome	Medium

			ic (smaller side chain)		RNA helicase MTR4like, betabarrel domain	
1015	V	I	Hydrophobic → Hydrophobic (no major polarity change)	3	ATPdependent RNA helicase SUPV3like / near Exosome RNA helicase MTR4like, betabarrel domain	Low
1026	V	A	Hydrophobic → Hydrophobic (smaller side chain)	0	ATPdependent RNA helicase SUPV3like / near Exosome RNA helicase MTR4like, betabarrel domain	Medium
1043	V	I	Hydrophobic → Hydrophobic (no major polarity change)	3	ATPdependent RNA helicase SUPV3like / near Exosome RNA helicase MTR4like, betabarrel domain	Low
1331	T	M	Polar uncharged → Hydrophobic	1	ATPdependent RNA helicase SUPV3like / ATPdependent RNA helicase Ski2/MTR4, Cterminal region	High

Table 2

Colors used in the predicted SKI2 3D models. Categorizes color to structural regions, mapped important residues, and tetraploid-specific substitutions in the AlphaFold models.

Color	Region / domain
Pink	Arch
Blue	RecA1 domain
Purple	RecA2 domain
Light green	Helical domain
Grey	N-terminal region
Yellow	Functionally mapped reference residues/highlighted important sites
Red	Tetraploid-specific mutations analyzed in this project

Figure 1
Picture of Mimulus Guttatus.
Source - Yant, Levi (2025)

BLOSUM62 matrix

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
Ala	A	4																			
Arg	R	-1	5																		
Asn	N	-2	0	6																	
Asp	D	-2	-2	1	6																
Cys	C	0	-3	-3	-3	9															
Gln	Q	-1	1	0	0	-3	5														
Glu	E	-1	0	0	2	-4	2	5													
Gly	G	0	-2	0	-1	-3	-2	-2	6												
His	H	-2	0	1	-1	-3	0	0	-2	8											
Ile	I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4										
Leu	L	-1	-2	-3	-4	-1	-2	-3	-4	-3	2	4									
Lys	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-2	5								
Met	M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	2	-1	5							
Phe	F	-2	-3	-3	-3	-2	-3	-3	-3	-1	0	0	-3	0	6						
Pro	P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7					
Ser	S	1	-1	1	0	-1	0	0	0	-1	-2	-2	0	-1	-2	-1	4				
Thr	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	1	5			
Trp	W	-3	-3	-4	-4	-2	-2	-3	-2	-2	-3	-2	-3	-1	1	-4	-3	-2	11		
Tyr	Y	-2	-2	-2	-3	-2	-1	-2	-3	2	-1	-1	-2	-1	3	-3	-2	-2	2	7	
Val	V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	1	-2	1	-1	-2	-2	0	-3	-1	4

Figure 2

BLOSUM62 substitution table used to see amino acid changes. Positive scores indicate more common or conservative substitutions while negative scores indicate less common or more chemically distinct substitutions.
Source - Guidi, Giulia (2020)

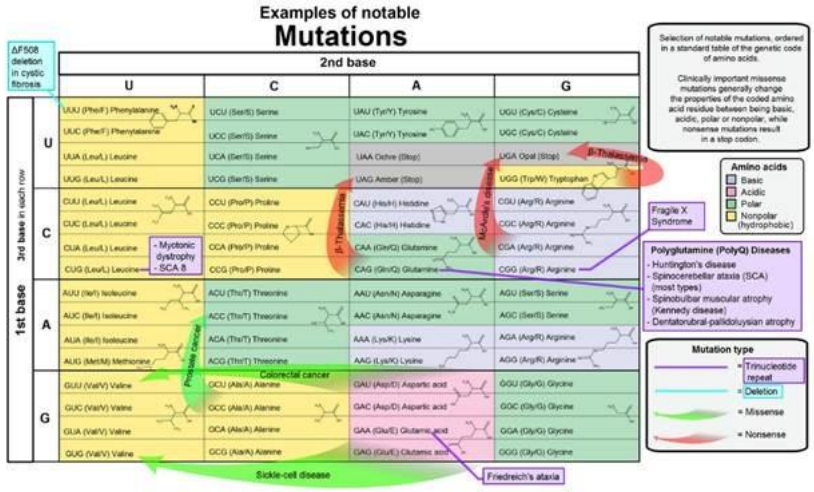
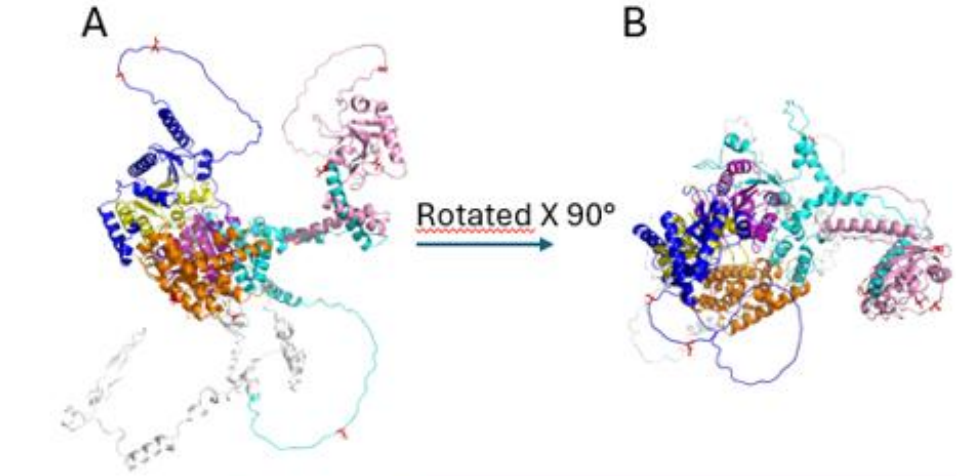
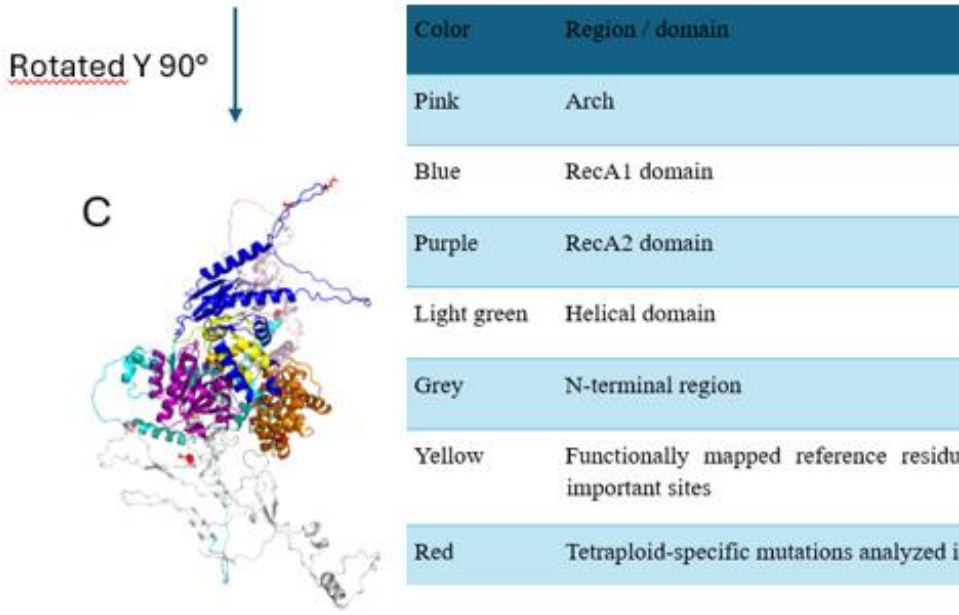


Figure 3.

Genetic code table illustrating codon-to-amino acid assignments and examples of mutation types. Amino acids are grouped according to general physicochemical properties like basic, acidic, polar, and nonpolar/hydrophobic.
Source - Häggström, Mikael (2014)

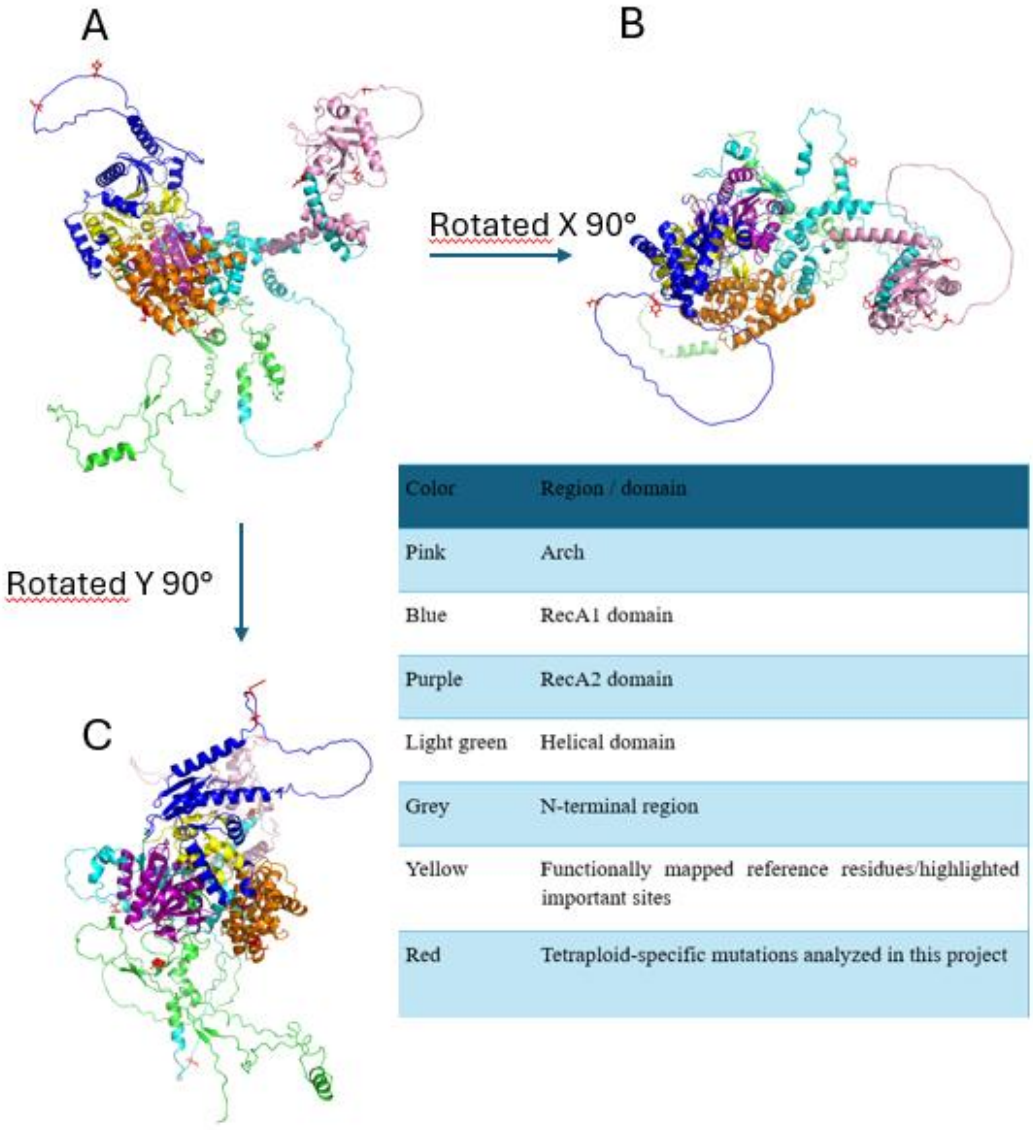




Color	Region / domain
Pink	Arch
Blue	RecA1 domain
Purple	RecA2 domain
Light green	Helical domain
Grey	N-terminal region
Yellow	Functionally mapped reference residues/highlighted important sites
Red	Tetraploid-specific mutations analyzed in this project

Figure 4 Diploid SKI2

Predicted AlphaFold structure of the Diploid SKI2 protein visualized with PyMOL. Major structural regions are colored to distinguish the N-terminal accessory domains and the central helicase core (See table 2).



Color	Region / domain
Pink	Arch
Blue	RecA1 domain
Purple	RecA2 domain
Light green	Helical domain
Grey	N-terminal region
Yellow	Functionally mapped reference residues/highlighted important sites
Red	Tetraploid-specific mutations analyzed in this project

Figure 5 Tetraploid SKI2

Predicted AlphaFold structure of the tetraploid SKI2 protein visualized with PyMOL. Major structural regions are colored to distinguish the N-terminal accessory domains and the central helicase core (See table 2)

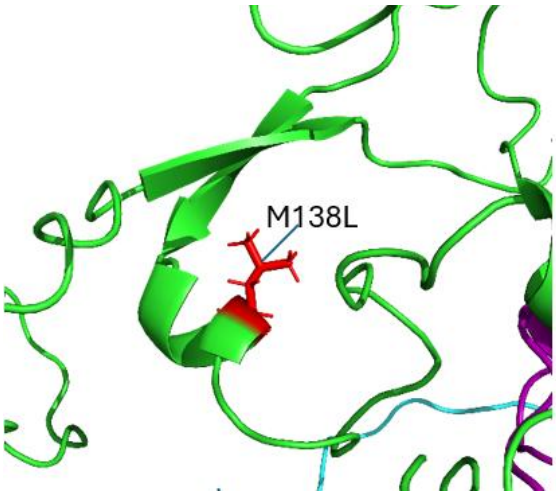


Figure 6 Residue #138

Zoom structural view of the residue 138 substitution in the SKI2 model. The residue is shown as a red stick and is located in the N-terminal region.

Figure 7 Residue #271

Zoom structural view of the residue 271 substitution in the SKI2 model. The residue is shown as a red stick and is located in a flexible loop.

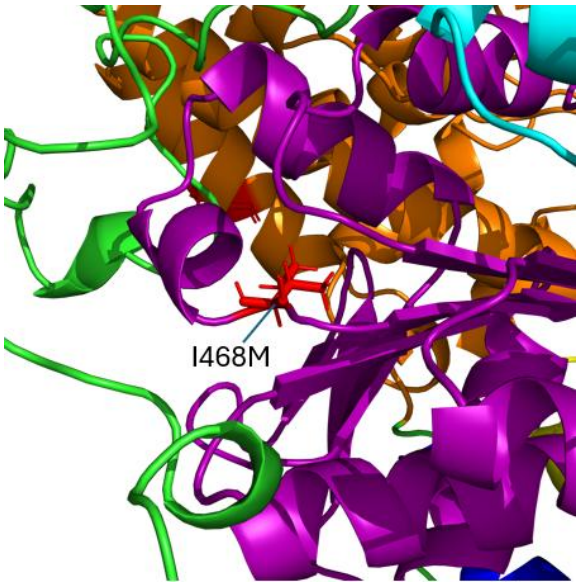


Figure 8 Residue #468

Zoom structural view of the residue 468 substitution in the SKI2 model. The residue is shown as a red stick and is located within the RecA1 region

.

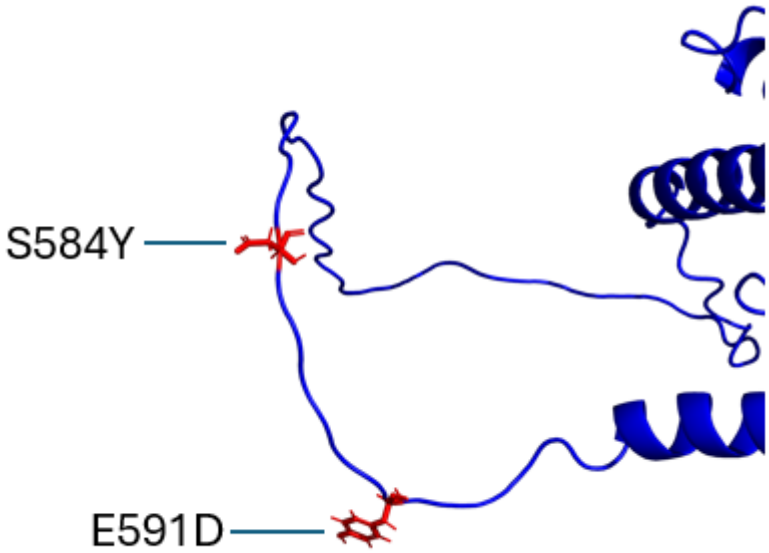


Figure 9 Residue #584

Zoom structural view of the residue 584 substitution in the SKI2 model. The residue is shown as a red stick and is located within the RecA2 region.

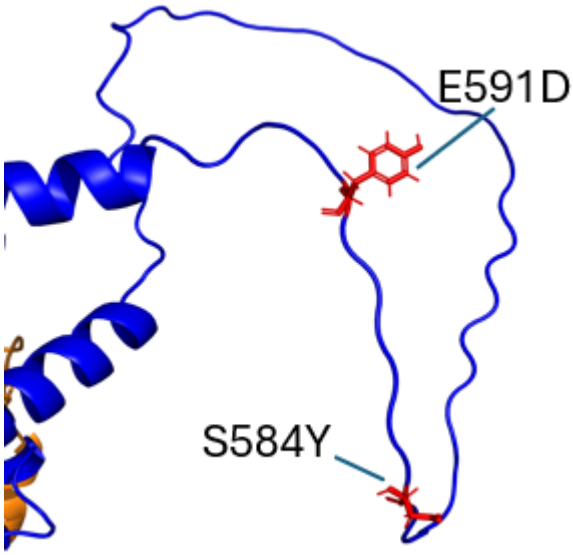


Figure 10 Residue #591

Zoom structural view of the residue 591 substitution in the SKI2 model. The residue is shown as a red stick and is located within the RecA2 region.

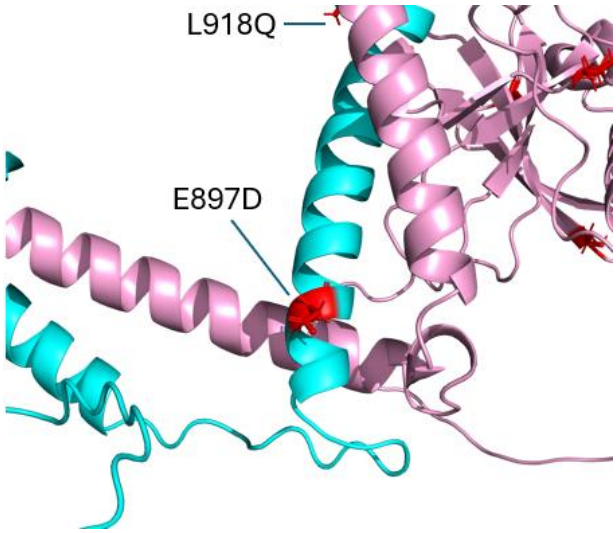


Figure 11 Residue #897

Zoom structural view of the residue 897 substitution in the SKI2 model. The residue is shown as a red stick and is located in the arch region.

.

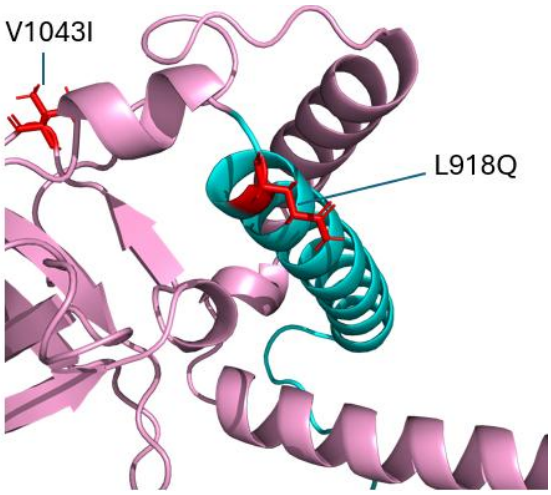


Figure 12 Residue #918

Zoom structural view of the residue 918 substitution in the SKI2 model. The residue is shown as a red stick and is located in the arch-associated region.

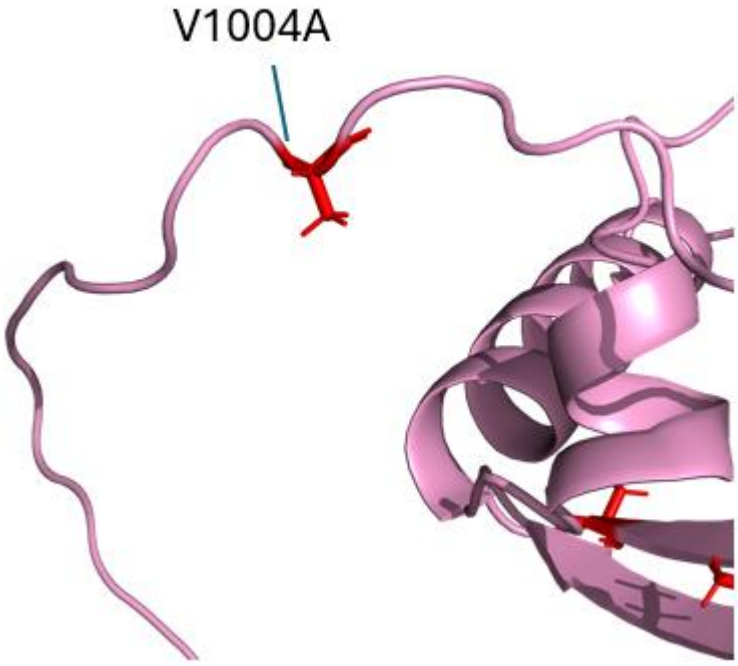


Figure 13 Residue #1004

Zoom structural view of the residue 1004 substitution in the SKI2 model. The residue is shown as a red stick and is located in the arch-associated region.

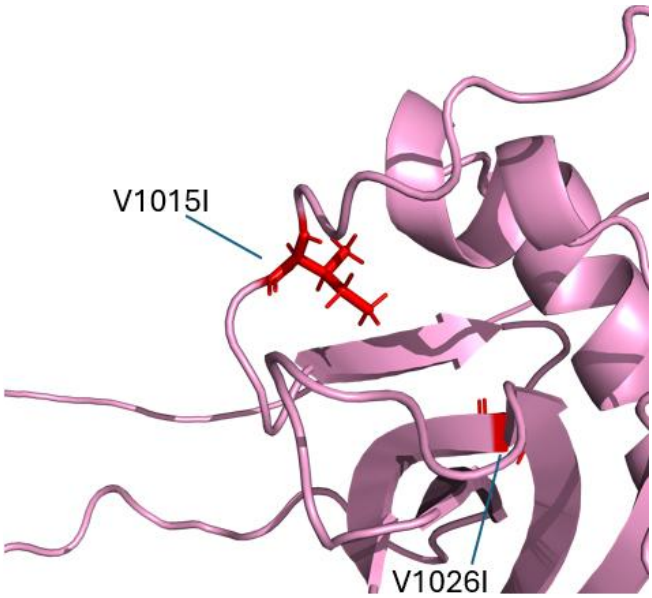


Figure 14 Residue #1015

Zoom structural view of the residue 1015 substitution in the SKI2 model. The residue is shown as a red stick and is located in the arch-associated region.

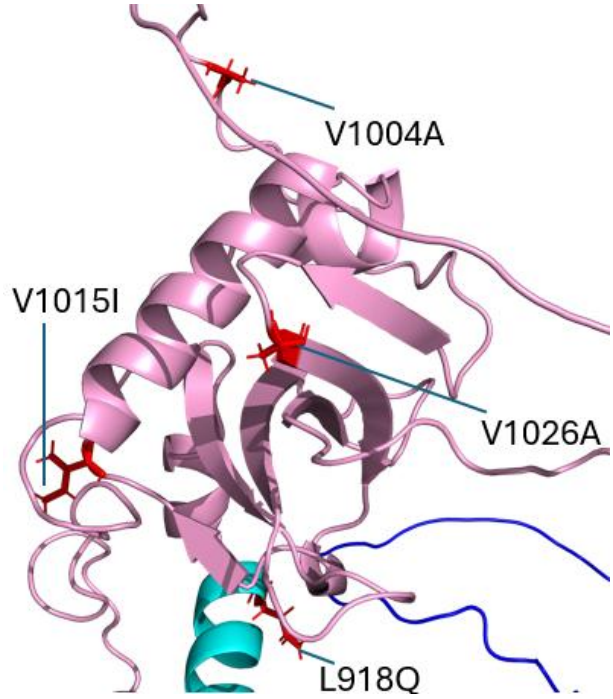


Figure 15 Residue #1026

Zoom structural view of the residue 1026 substitution in the SKI2 model. The residue is shown as a red stick and is located in the arch-associated region.

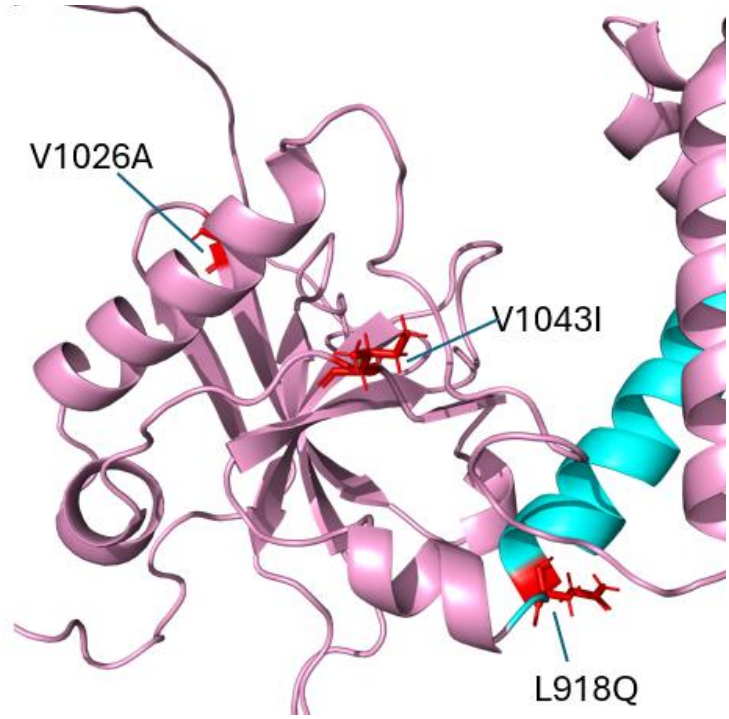


Figure 16 Residue #1043

Zoom structural view of the residue 1043 substitution in the SKI2 model. The residue is shown as a red stick and is located in the arch-associated region

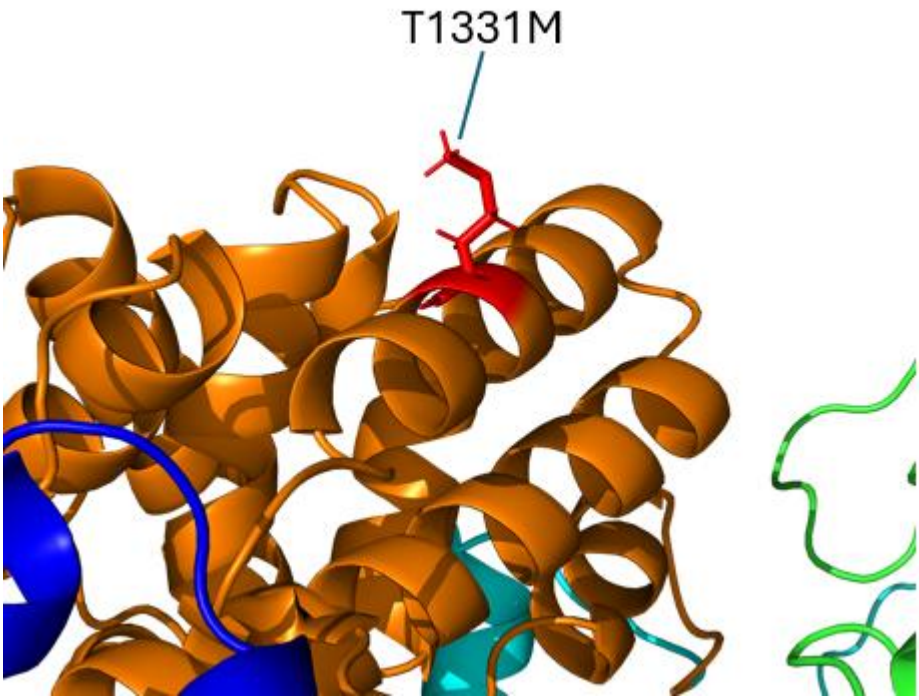


Figure 17 Residue #1331

Zoom structural view of the residue 1331 substitution in the SKI2 model. The residue is shown as a red stick and is located in the helical domain.