Center for Computational and Theoretical Biology

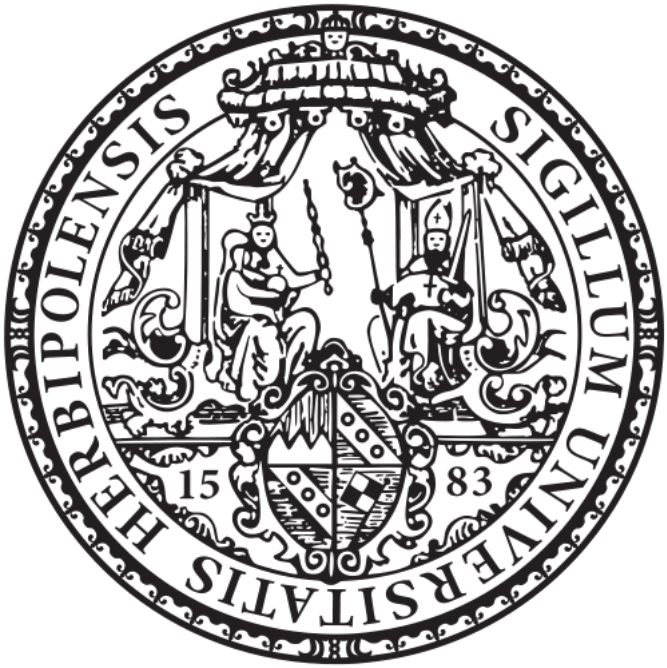
Master Thesis

Selection on loss-of-function variants

Master Biosciences
Julius-Maximilians-Universität Würzburg

submitted by Laura Steinmann
Supervisor PD Arthur Korte

March 17, 2022



Supervisor PD Arthur Korte

Second reviewer Prof. Dr. Dirk Becker

Date of Submission, office stamp

Zusammenfassung

Abstract

Contents

1. Introduction

To survive in an ecosystem organisms need to adapt to the specific abiotic and biotic factors around them. This is particularly important for plants since they can not change their habitat during their lifespan. Adaptation can operate on different levels of the organisms from the adaptation of protein function to modifications in cell functions or whole tissue functions. But the underlying fundamental process that drives adaptation is the modification of the genetic material DNA.
Modifications on DNA level are called mutations and in four different groups. In the following description we consider just mutations which occur in coding regions of the genome although they can be distributed throughout all regions of the genome. A point mutation, which is the first group of mutations and can be also called single nucleotide polymorphism (SNP), is concerning just one single base pair. These changes the DNA sequence on one single point and therefore lead to changes of the transcribed mRNA. The second class of mutations are insertions. These insert a new sequence into the former DNA sequence and thus elongate it. These insertions can have a variety of length from one single bp to a few hundred bp and even longer sequences can be added. The contrasting class of insertions are deletions, which lead to a reduction of base pairs in a variety of lengths in the former DNA strand. The last category of mutations are duplications. As implied by the name, regions of the DNA get duplicated and inserted at a different position. The regions can be copied abnormally one or even more times.
These mutations are first of all just changes in the DNA but indirectly they impact all the resulting processes. The DNA first gets transcribed into mRNA. This process is not affected by the evolved mutations. But after the transcription the mRNA gets translated into proteins. In this step the impact of the mutations appear since a triplet codon gets translated into a specific amino acid, which can change due to the mutation in the DNA as postulated by Gamow[1]. The mutations can have different effects on the resulting protein. The insertions or deletions lead mostly to a shift in the reading frame and therefore change the sequence of amino acids in a wide range. Point mutations, although they change just a single codon, can have different effects on the translated amino acid and therefore on the functionality of the resulting protein. As it was known that 20 amino acids exists and when the knowledge arises that a codon includes three base pairs the genetic code could get deciphered by Nirenberg et al. [2]. The gain of knowledge deciphering this genetic code showed that the genetic code is degenerated as identified by Lagerkvist [3], which means that not one codon is directly linked to an amino acid rather that there are multiple codons that can initiate the same amino acid, which is showed in Figure 1.1 a). For the topic of point mutations we have three possible outcomes resulting after such mutation. The first outcome is a synonymous mutation, which is the less severe one and does not change the protein at all. Here a base pair changes but the translated codons provoke the same amino acid. An schematic representation of synonymous mutations can be found in Figure 1.1 b). As shown in Figure 1.1 c) a mutation could also be a non-synonymous mutation where the modified codon leads to an exchange of the amino acid type. The prediction of how severe these non-synonymous mutations are is hard to do without further analysis. It depends on many circumstances for example if the location is a catalytic region or to which other amino acid it is exchanged. Such a mutation can be very unnoticeable or on the other hand making the protein non-functional. The last kind of point-mutations are premature stop codons. They represent a mutation of a usual codon to a stop codon as it can be seen in Figure 1.1d). This means most of the time drastic changes in the protein functionality since the resulting protein is truncated. Thereby premature stop codons represent the most interesting kind of mutations for understanding adaptation in plants since they can be recognized in the DNA level and they should change the function of the truncated protein.
Until recently in evolutionary theories the axiom exists that these described mutations occur randomly

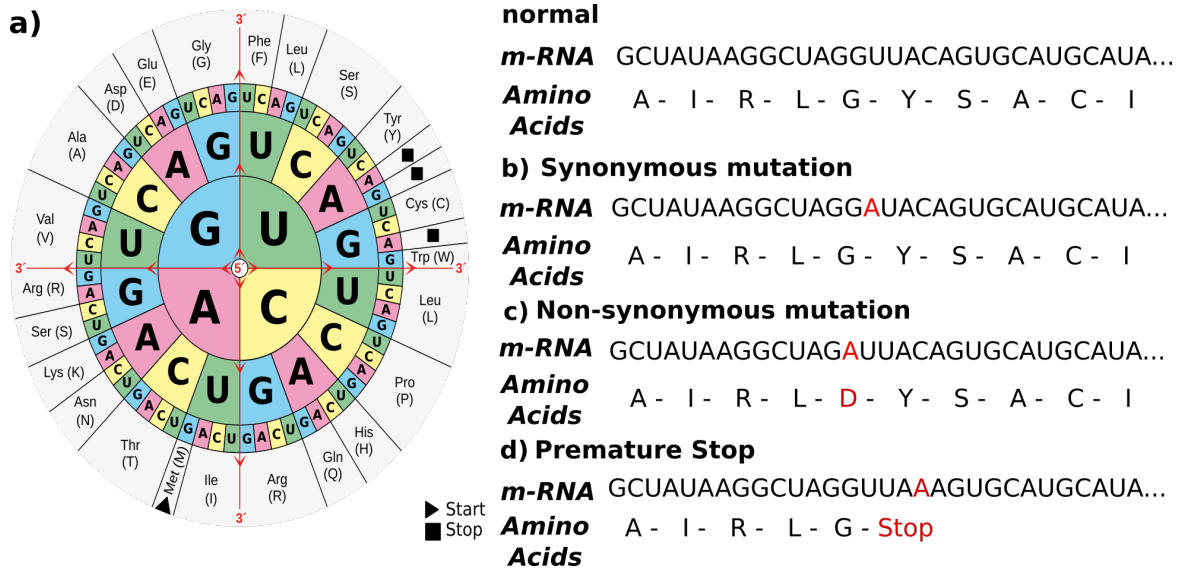


Fig. 1.1.: Schematic representation of point mutation classes
a) The genetic code (image taken from Bresch and Hausmann [4]); b) A schematic representation of a synonymous mutation, which affects the mRNA but not the amino acid sequence. c) A schematic representation of a non-synonymous mutation. The change of a single base pair in the mRNA sequence changes to an exchange of the amino acid type. d) A schematic representation of a premature stop codon. Retrieved from a exchange of a single base pair an premature stop codon appears and stops the amino acid sequence.

throughout the genome and that solely after occurrence of mutations the evolutionary selection operates and shapes the genetic pool of a population[5]. From this axiom it can be deduced that natural selection acts on this mutations by classifying fatal modifications and filter these out of the genetic pool. The time span during which this genetic pool adapts to a single mutation is depending on the severity of the modification. That also implies that mutations which have a neutral significance like synonymous mutations are not shaped by the selection force. On the other side positive impact is reinforced by the selection force so that these mutations get accumulated in the genetic pool[6].
If one now only considers the mutation type premature stop codons some hypotheses can be used to explain the effects of mutations on gene regulatory networks. Since the occurence of premature stop codons is leading directly to a loss of function in this protein one can consider that the evolutionary selection force stabilizes the gene regulatory networks by maintaining the diversion of this networks. Practically this means that if a protein is knocked out by a premature stop codons the pathways around this gene are getting more important and you expect that they are under a high conservation force by evolutionary selection and selection prevent a accumulation of premature stop codons in the rest of the gene regulatory network.
In our following project we want to study the attributes and occurence of these premature stop codons. We want to characterize how the natural selection is shaping population structures through generating loss-of-function variants and how they affect the adaptation process of organisms. We therefore follow a bioinformatical approach by applying data analysis and statistical algorithms on different genomic data and with these try to find evolutionary patterns.

2. Material

In this chapter we will have a look at the necessary computing infrastructure and the code base that I developed for studying of the selection on loss-of-function variants. At the end of this chapter we will have a look at our used dataset.

2.1. Computational Material and Resources

In this section I will explain the computational material. We will start with the computing infrastructure we used for our analysis and continue with looking at the location and dependencies of our code base.

2.1.1. Computing Infrastructure

The whole analysis is calculated on the institutes own high-performance computer cluster. With access to a computer node with 80 cores and 512 GB RAM. Also not all analysis need the maximum capacity of this computer node especially working with the basic unfiltered dataset consums a lot of RAM. To recapitulate any analysis access to a high-performance computer node is therefore necessary.

2.1.2. Code Base

To study how premature stop codons get distributed in the genetic pool and shape the adaptation in plants we followed a completely bioinformatical based approach. Therefore we developed the statistical analysis and the data analysis in programming languages. To follow these ideas use mainly python but also R as an additional language to complete the workflow set up on an linux based system (Debian). If you like to recapitulate the analysis you can find the code in a github repository `https://github.com/laurasteinmann/Premature_Stop_Codons.git`.
Although at the moment there are no strict dependency on software versions I will list the used packages and there versions since there is the possibility of semantic changes in future versions. For the Python analysis part we use python (3.10) and the important packages for dealing with numerical data and big data analysis like pandas (1.4) and numpy (1.22). For calculating statistical test we use scipy (1.8.0). For visualizing our results we use the matplotlib package (3.5), the seaborn package (0.11) and the matplotlib_venn package (0.11). For dealing with genomic data and filtering it we use the R language with version 4.1.3 and the vcfR library (1.12)

2.2. Dataset of 1001 Genomes Project

3. Methods

4. Results

List of Figures

B. Bibliography

[1] Francis Crick. “Chapter 8: The genetic code”. In: *What mad pursuit: a personal view of scientific discovery. New York: Basic Books* (1988), pp. 89–101.

[2] M Nirenberg et al. “RNA codewords and protein synthesis, VII. On the general nature of the RNA code.” In: *Proceedings of the National Academy of Sciences of the United States of America* 53.5 (May 1965), pp. 1161–1168. ISSN: 0027-8424. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC301388/ (visited on 03/11/2022).

[3] U Lagerkvist. “"Two out of three": an alternative method for codon reading.” In: *Proceedings of the National Academy of Sciences* 75.4 (Apr. 1978), pp. 1759–1762. DOI: 10.1073/pnas.75.4.1759. URL: https://www.pnas.org/doi/abs/10.1073/pnas.75.4.1759 (visited on 03/16/2022).

[4] Carsten Bresch and Rudolf Hausmann. *Klassische und molekulare Genetik*. Springer-Verlag, 2013.

[5] DJ Futuyma. “Evolutionary biology, 2nd edn Sunderland”. In: *MA: Sinauer.[Google Scholar]* (1986).

[6] Charles Darwin. *The origin of species*. PF Collier & son New York, 1909.

Affirmation

I hereby confirm that my master thesis entitled Selection on loss-of-function variants
is the result of my own work./
I did not receive any support from commercial consultants. /
I have given due reference to all sources and materials used in the thesis and have listed and specified them. /
I confirm that this thesis has not yet been submitted as part of another examination process neither in identical nor in similar form. /
I agree that the thesis can be checked for plagiarism also by using a software./

Hiermit erkläre ich an Eides statt, dass ich die Masterarbeit mit dem Titel Selektion der Loss-of-Function Varianten
eigenständig und eigenhändig angefertigt habe.
Ich habe keine Unterstützung kommerzieller Berater erhalten.
Ich habe alle in der Arbeit verwendeten Quellen und Materialien ordnungsgemäß zitiert, aufgelistet und spezifiziert
Ich erkläre, dass die vorliegende Arbeit weder in gleicher noch in ähnlicher Form bereits in einem anderen Prüfungsverfahren vorgelegen hat.
Ich bestätige, dass die Thesis auch mit Hilfe einer Software auf Plagiat untersucht werden kann

Würzburg, March 17, 2022 Laura Steinmann