CTB 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 23, 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. The necessary adaptation can occur on different levels of the organisms, starting from the adaptation of protein function to modifications in cell functions and whole tissue properties. In all of these cases the fundamental process driving adaptation is the modification of the genetic material: DNA (Griffith, 1928[1]).

Modifications on DNA level are called mutations and can be distributed throughout the genome. Here we will constrict our analysis to mutations which occur in coding regions of the genome although they can be distributed also in intergenic regions. A simple explanation is based on the outcome of mutations. Mutations in the coding regions of the genome can have a direct impact on the amino acid sequence and therefore influence the protein functionality, which is described in the following. But mutations that occur in intergenic regions do have indirect impact on many processes in the cell by mainly influencing the gene regulation. In our further analysis we are not interested in these regulatory changes rather we are interested in how changes in the coding regions affect adaptation.

Mutations can be classified into four different groups. The first group of mutations to consider is the point mutation since it is concerning just one single base pair (bp). Since these mutations only change a single point in the DNA sequence they will directly result in changes to the transcribed mRNA. Another class of mutations are insertions. Here a new sequence is inserted into the original DNA sequence which results in its elongation. Insertions can have a variety of lengths from one single base pair to a few hundred base pairs and even longer sequences. The contrasting class of insertions are deletions, which lead to a reduction of a number of base pairs. They can likewise result in a variety of lengths of the mutated DNA strand. The fourth and 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.

The impact of mutations on the biochemical processes inside a cell and an organisms phenotype is determined by its influence on protein creation. In this process the DNA first gets transcribed into mRNA, a step that is not affected by the mutations. The mRNA is translated into proteins in the following step. Here the mutations have a direct impact as each triplet base pair (codon) gets translated into a specific amino acid, which can change due to the mutation in the DNA as postulated by Gamow[2].The specific mapping between amino acids and these codons was deciphered by Nirenberg et al. [3]. Figure 1.1 a) shows how each triplet codon is translated into a specific amino acid or indicates a stop codon, providing the universal genetic code that links RNA sequences and proteins. The fact that only 20 amino acids are matched with $4^3 = 64$ unique codons led Lagerkvist [4] to the insight that this code is degenerated. This means that a single codon is not directly linked to a single amino acid, but that multiple codons are translated to the same amino acid.

Based on this translation table we can now understand that mutations can have different effects on the resulting protein. In most cases, insertions or deletions result in a shift of the reading frame and therefore change the sequence of amino acids in a wide range extending the region of the mutation. While point mutations only influence a single codon, they can still have different effects on the translated amino acid and therefore on the functionality of the resulting protein. We distinguish three possible outcomes of such a mutation on protein level: (i) A synonymous mutation, where the protein is not changed at all. Here a base pair changes but the translated codons provoke the same amino acid. An example for a synonymous mutation is provided in Figure 1.1 b). (ii) Non-synonymous mutations resulting in the exchange of the amino acid type as shown in Figure 1.1 c). Predicting the severity of non-synonymous

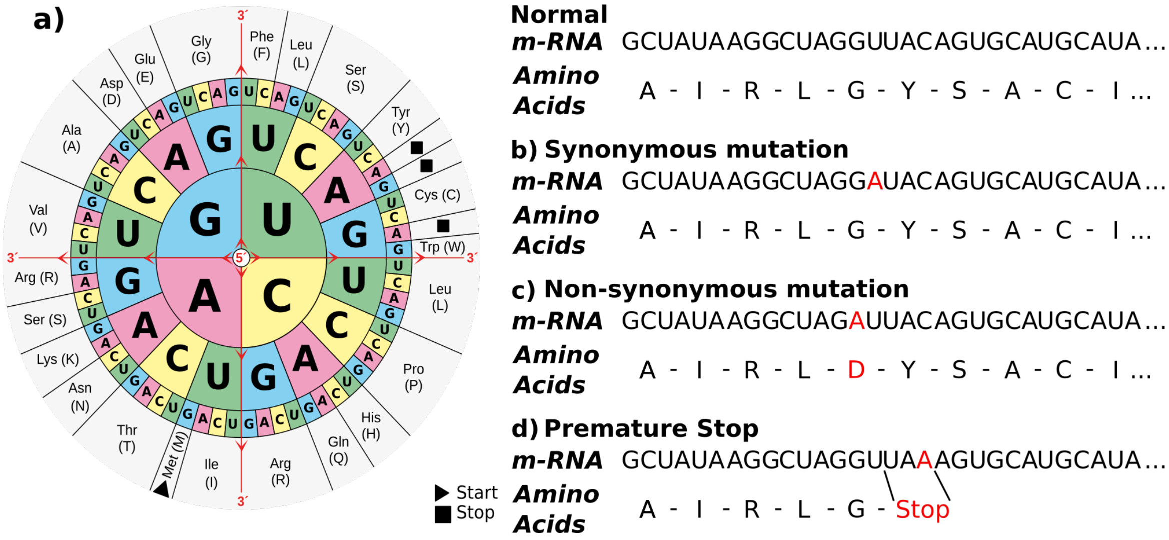
a)
Gly (G)
Phe (F)
Leu (L)
Ser (S)
Glu (E)
Asp (D)
Ala (A)
Tyr (Y)
Cys (C)
Trp (W)
Val (V)
Arg (R)
Leu (L)
Ser (S)
Lys (K)
Asn (N)
Pro (P)
His (H)
Thr (T)
Gln (Q)
Met (M)
Ile (I)
Arg (R)
3´
5´
▶ Start
■ Stop
Normal
m-RNA GCUAUAAGGCUAGGUUACAGUGCAUGCAUA...
Amino Acids A - I - R - L - G - Y - S - A - C - I ...
b) Synonymous mutation
m-RNA GCUAUAAGGCUAGGAUACAGUGCAUGCAUA...
Amino Acids A - I - R - L - G - Y - S - A - C - I ...
c) Non-synonymous mutation
m-RNA GCUAUAAGGCUAGAUUACAGUGCAUGCAUA...
Amino Acids A - I - R - L - D - Y - S - A - C - I ...
d) Premature Stop
m-RNA GCUAUAAGGCUAGGUUAAAGUGCAUGCAUA...
Amino Acids A - I - R - L - G - Stop

Fig. 1.1: Schematic representation of point mutation classes
a) The genetic code (image taken from Bresch and Hausmann [5]); b) An examplary representation of a synonymous mutation, which affects the mRNA but not the amino acid sequence. c) Example of a non-synonymous mutation. The change of a single base pair in the mRNA sequence leads to an exchange of the amino acid type. d) Example of a premature stop codon mutation. Induced by the exchange of a single base pair a premature stop codon appears and stops the amino acid sequence.

mutations is hard to do without further analysis since it depends on many circumstances like the location of the changed amino acid with respect to the catalytic region of the protein or the specific amino acids involved in the exchange. Therefore these mutations vary in their significance between almost unnoticeable and making the protein non-functional. (iii) Premature stop codons: This kind of mutation changes a usual codon to a stop codon as presented in Figure 1.1 d). This will almost always result in drastic changes to the protein functionality since the resulting protein is truncated. Due to this they represent one of the most interesting kind of mutations for understanding adaptation in plants. Premature stop codons can be recognized on the DNA level and will directly impact the function of the truncated protein in almost every case.

Driven by our interest in adaptation over an evolutionary timescale we study mutations on a population scale in this thesis. Point mutations which occur on population scale are called single-nucleotide polymorphisms (SNPs). Each SNP defines a difference at a specific position in the genome that appears in the population. This way we can distinguish individuals which have the reference allele at this position (no point mutation) from individuals which have the alternative allele (a change of the base pair). These SNPs are distributed throughout the whole genome and can be used to analyse patterns of mutations statistically and correlate them to phenotypic effects.

Until recently, evolutionary theory was working with the assumption that all types of mutations referenced above occur randomly throughout the genome initially and that evolutionary selection only influences their presence in a populations genetic pool after their occurrence (Futuyma 1986[6]). From this starting point one can deduce the natural selection removes fatal modifications from the genetic pool and generically influences the occurrence mutations depending on their severity. This also implies that mutations with neutral significance like synonymous mutations are not shaped by the selection force. On the other side positive impact is reinforced by the selection force, leading to an accumulation of these mutations in the genetic pool (Darwin 1909 [7]).

By summarizing these described processes we can draw conclusions of expected adaptation processes in organisms. Combining that mutations act as a fundamental force of evolution and that natural selection is shaping the population we can get an understanding of how adaptation shapes an individual organisms. Restricting our view to mutations of premature stop codon type and considering the adaptation of an organisms to abiotic and biotic factors as a complex gene-regulatory networks that is regulated, we can hypothesize the effects of mutations on these networks. Since the occurrence of premature stop codons is directly resulting in a loss of function of the affected protein one can assume that evolutionary selection force stabilizes the surrounding gene regulatory network by maintaining the diversion of this networks. Practically this means that if a protein is knocked out by a premature stop codon the pathways surrounding this damaged one are getting more important and you expect them to be under a high conservation force by evolutionary selection. The selection should prevent a accumulation of premature stop codons in the rest of the gene regulatory network to conserve the original functionality of a network as best as possible.

In this thesis we want to study the attributes and occurrence of these premature stop codons in a population of *Arabidopsis thaliana*. We 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 accomplish this by following a bioinformatical approach i.e. by applying data analysis and statistical algorithms on different genomic data in order to identify 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 the selection on loss-of-function variants. At the end of this chapter we will have a look at the *A. thaliana* dataset we use.

2.1 Computational Material and Resources

In this section I will explain the computational material. We will start with the computing infrastructure that is necessary to perform our analysis and continue with looking at the location and dependencies of my 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 the complete 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 I use mainly python but also R as an additional language to complete the workflow. They are 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 I 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 I use scipy (1.8.0). For visualizing our results I 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 I use the R language with version 4.1.3 and the vcfR library (1.12)

2.2 *A. thaliana* dataset of 1001 Genomes Project

The flowering plant *Arabidopsis thaliana* is the standard reference plant for many disciplines in biology. For example for research questions in physiological, cellular, and molecular processes it is one of the most studied model organisms (Koornneef 2010[8]). *A. thaliana* especially gain importance for studying genes and determining their function since it was the first sequence of a plant genome in 2000 (Arabidopsis Genomes Initiative [9]). With the relative small genome of 125 megabase pair that include about 25 000 genes *A. thaliana* emphasized its importance as a model organism. This reference genome is based on the col-0 ecotype of *A. thaliana* and its annotation and quality was improved with the technological progress since then.
A. thaliana is also a promoted model organism for studying natural variation since *A. thaliana* is not

known for agricultural usage or as a crop plant. This means that in *A. thaliana* we can still observe evolutionary selection and not the selection of human interests[8]. By generating and publishing genomic sequences of 1, 135 ecotypes (accessions) of *A. thaliana* in the 1001 Genomes project (1001 Genomes Consortium[10]) we have an ideal dataset to perform the analysis of adaptation in plant on population scale. These 1,135 accessions come from a worldwide hierarchical collection, which includes ecotypes from Sweden as well as the Iberian Peninsula as well as from North America and Central Asia. The accessions are a collection of naturally inbred lines that represent individuals under diverse ecological conditions. By sequencing with Illumina short read sequencing whole-genomes were systematically characterized of genome-wide polymorphisms. In the total collection 10,707,430 SNPs were identified. Based on this genomic information we can try to quantify genomic variation in a representative sample of accessions.

As an additional resource the 1001 Genomes Consortium provides 727 accessions with full-sequenced transcriptomes (Kawakatsu 2016[11]). These accessions have an average expression of 18,000 genes and allows us to combine the information of genomic and transcriptomic data. Unfortunately these 727 accessions overlap not completely with the 1, 135 accessions. If we want to look at both datasets we can just base our analysis of 665 accessions.

3 Methods

In this chapter we want to focus on the methods behind our analysis. We take a look at how the necessary statistical methods work and how we prepare the available information to extract the data we need to understand the selection in plant populations further.

3.1 Distribution of premature stop codons

3.2 Generation of a high confidential dataset

3.2.1 Classification of premature stop codons based on their gene expression differences

3.2.2 Classification of premature stop codons based on the calculation of the length of the remaining mRNA

4 Results

Since we want to study the attributes and the occurrence of premature stop codons in the 1,135 population of *A. thaliana* we try to characterize how natural selection is shaping population structures through loss-of-function variants. To answer this research question we followed an statistical workflow. We will start the project by looking first at the distribution of premature stop codons in the 1, 135 accessions of *A.thaliana* dataset and continue by generating two high confidential datasets for our further analysis. Here we will follow two separate approaches, one is based on the gene expression differences between wildtype and knock-out mutants the other one by selecting stop codons based on the calculation of the remaining length. Afterwards we want to look at the interactions between premature stop codons and finish our analysis by looking at a control group of mutations, which will be the synonymous mutations.

4.1 Distribution of premature stop codons in *A. thaliana*

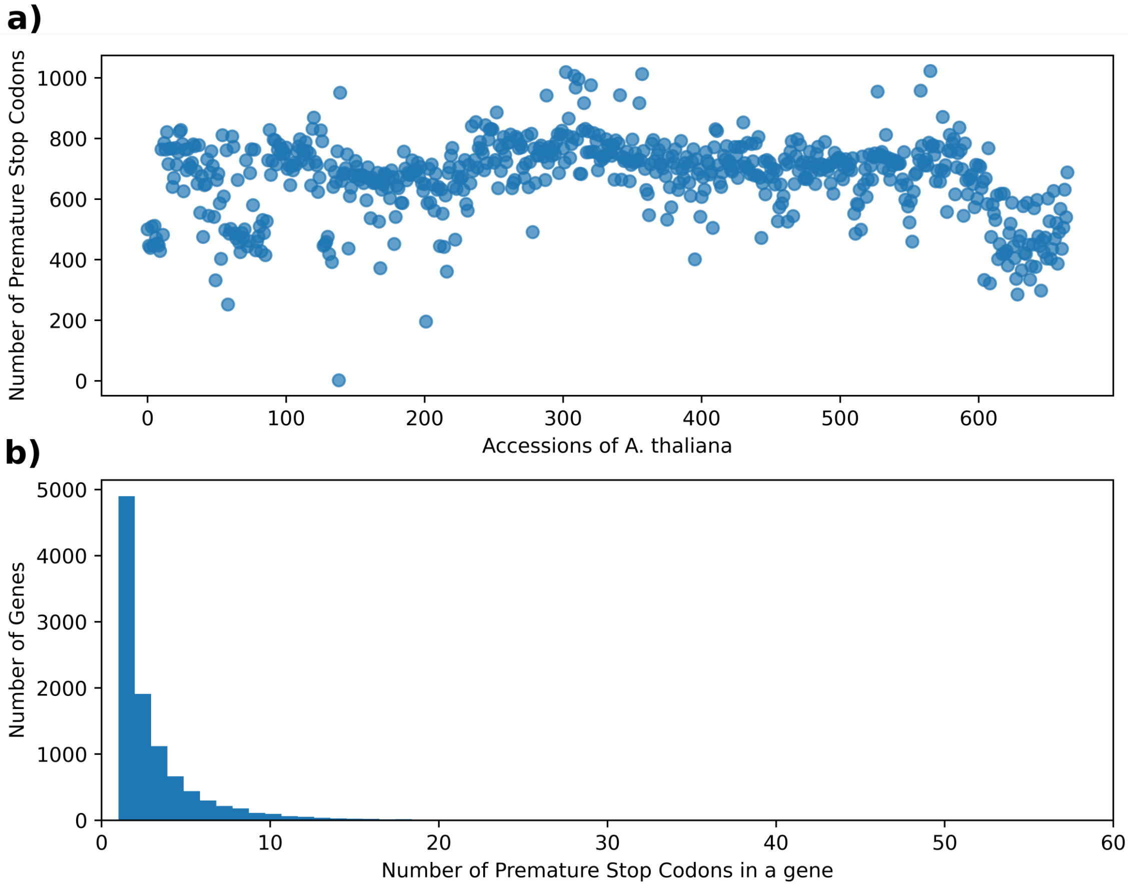

Fig. 4.1: Distribution of premature stop codons in the genome from the 665 accessions *A. thaliana* population

a) Distribution of premature stop codons along the 665 accessions. In each accessions, execept one, exist many hundreds of premature stop codons across the whole coding sequences of the genome. b) Histogram of number of stop codons in a gene. Most of the premature stop codons occur just once in each gene but there are also genes where we have multiple premature stop codons occurring.

List of Figures

List of Tables

Bibliography

[1] Fred. Griffith. “The Significance of Pneumococcal Types”. In: *The Journal of Hygiene* 27.2 (Jan. 1928), pp. 113–159. ISSN: 0022-1724. URL: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2167760/` (visited on 03/23/2022).

[2] 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.

[3] 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).

[4] 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).

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

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

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

[8] Maarten Koornneef and David Meinke. “The development of Arabidopsis as a model plant”. en. In: *The Plant Journal* 61.6 (2010), pp. 909–921. ISSN: 1365-313X. DOI: `10.1111/j.1365-313X.2009.04086.x`. URL: `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-313X.2009.04086.x` (visited on 03/22/2022).

[9] Arabidopsis Genome Initiative. “Analysis of the genome sequence of the flowering plant Arabidopsis thaliana”. eng. In: *Nature* 408.6814 (Dec. 2000), pp. 796–815. ISSN: 0028-0836. DOI: `10.1038/35048692`.

[10] 1001 Genomes Consortium. “1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana”. eng. In: *Cell* 166.2 (July 2016), pp. 481–491. ISSN: 1097-4172. DOI: `10.1016/j.cell.2016.05.063`.

[11] Taiji Kawakatsu et al. “Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions”. en. In: *Cell* 166.2 (July 2016), pp. 492–505. ISSN: 0092-8674. DOI: `10.1016/j.cell.2016.06.044`. URL: `https://www.sciencedirect.com/science/article/pii/S0092867416308522` (visited on 03/23/2022).

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 23, 2022 Laura Steinmann