The human genome is made of three billion letters of DNA that encode genes playing a major role in making us who we are. Therefore, the genome is often being described as "the book of life". Although the text of "the book of life" - the order of the DNA letters is known and everybody can see it and read it, in that form it does not provide any information about biological functions of its regions. Thus, for many of genes, their actual biological roles remain undiscovered. Protein-coding genes, which are genes that produce proteins in the cell are the most important ones for living organisms. It was long assumed that the genome is mainly made of protein-coding genes and that the organismal complexity depends on their number. Meaning that the more complex organism, the thicker is "the book of life". Revealing the human genome sequence was followed by huge surprise as it turned out that humans have a similar number of protein genes as worms. It was also shown that protein-coding genes constitute only 1-2% of human genome. The rest is non-protein-coding and due to its mysterious role in the cell is often called a "dark matter" of the genome. Another amazement came together with the studies investigating how "the book of life" is printed in the cell. The non-protein-coding part was proven to be pervasively printed and produce various regulatory units that often interact with protein-coding genes. The biggest and the most intriguing class of non-protein-coding elements are the long non-coding RNAs (lncRNAs). They look like protein-coding genes, but they do not produce any proteins. The genome's dark matter was largely neglected by biological studies, until recently after many lncRNAs were proven to play crucial biological roles. Some of them were also identified to be involved in the progression of various diseases, including cancer. However, until now only ~2% of human lncRNAs (out of ~19,000) have been functionally characterized, and the rest of them remains largely unexplored.

The main task is to understand which lncRNAs are functional and how those functions are kept in the genome. Achieving this goal is not trivial. Firstly, it requires knowing the location of all lncRNAs in the genome. This is challenging as lncRNAs do not provide any hint that would help to track them. Finding the protein-coding genes is much easier, as we know their final product – the protein made of amino acid letters. Therefore, it is like finding a sentence in Spanish within an English book. It first requires translating the sentence and then using it to search the book. For lncRNAs, the final product is just an RNA, so it is necessary to screen all the RNAs in the cell to find them. Moreover, they are highly tissue and cell-specific. Meaning that one lncRNA can be fully present in one cell, but only partially present or absent in another. It is like printing out the same text using all printers in the building and getting different results from each office or department. Furthermore, the presence of lncRNAs in the cell is largely dominated by the protein-coding genes that on average are much more frequently printed than the non-coding elements. Thus, cataloging lncRNAs requires deep-fishing them from the mixture of RNAs. Secondly, understanding the biological roles of lncRNAs necessitates knowing their evolutionary conservation. Evolution keeps crucial biological functions by passing them as stretches of DNA sequence from one genome to the other over millions of years. Therefore, knowing which lncRNAs are conserved across distant species could guide the researchers. However, so far the analysis of lncRNA conservation was mainly done using whole-genome comparisons, where the lncRNAs were mapped to a full genome sequence of another organism. This is not very accurate, as it only allows to detect the presence of given lncRNA in another genome, rather than finding an overlap between lncRNAs in different species. This like searching words or sentences from one chapter of a given book in the full text of another book. It allows to find matches between books, but not necessarily between chapters.

To address these limitations, firstly we are going to improve and expand zebrafish lncRNA annotation – a detailed map describing precise location of lncRNAs in the genome. We will overcome the challenges of finding lncRNAs by focusing our analysis only on the genomic fragments that potentially could contain them. This could be compared to highlighting the text in color to mark the parts of interests, rather than screening the whole text. Secondly, we will use our new lncRNA catalogue to look for the conserved sequences or their fragments in human and mouse lncRNAs. Finally, we will select evolutionary conserved candidate lncRNAs to see, whether switching them off causes any effect on zebrafish development. We think that this project will provide some important insights into understanding biological roles of lncRNAs.