Genome sequencing has an unprecedented impact on our understanding of the species living on Earth. The cost of sequencing the first almost-complete human genome (completed in 2001) was about 3 billion dollars. After 20 years, we need to pay only about 1K USD. This is a huge improvement, and we can expect that this is not the end of the story.

The key players in the market are Illumina, Oxford Nanopore, and PacBio. Their sequencing instruments differ significantly. Illumina offers a high throughput sequencing and relatively short reads of high quality (100–250 base pairs). PacBio produces much longer (e.g., 15kbp) high-quality reads, but the throughput is lower. Finally, Oxford Nanopore gives even longer reads (e.g., 1Mbp), but the quality is moderate. Taking the best of all of them we can *de novo* assemble even complex genomes, like humans.

The existence of such relatively low-cost technologies resulted in the launching of several large-scale whole-genome sequencing projects. The Telomere-to-Telomere consortium published the first complete human genome. Filling all the gaps in the human genome sequence since 2001, when the first assembly was announced took 20 years, which shows how complex it was. The Human Pangenome Project (HPP) is another example. It aims at *de novo* assembling of hundreds of human genomes providing an ultimate database of genomic diversity in human populations. Until June 2022, they published almost a hundred high-quality haplotype human genomes. Such efforts should reduce the underrepresentation of many genomic sequences in the currently used reference genome. It is expected that using many genomes from various populations will remarkably improve the determination of genomic variants in personalized medicine.

There are also projects focused on collecting complete genomes of various species. One of the examples is the Vertebrate Genomes Project. Its goal is to provide high-quality reference genomes for all approx. 70,000 extant vertebrate species. The goals of the Earth BioGenome Project (EBGP) are also very ambitious. They plan to sequence all known eukaryotic species in 10 years.

One of the primary goals of genome sequencing is the identification of protein-coding genes. The first large protein databases were published in the 1990s. Nowadays, probably the most popular of them, Pfam, contains almost 20k protein families and 169M protein sequences in total. The largest family contains almost 3M sequences.

In this project, we will focus on developing new algorithms for the mentioned data. First, to address the cost of storage and transfer of huge datasets we will provide specialized compression algorithms. Second, we will develop algorithms for making use of the collected genome and protein sequences to solve selected biological problems.