Since 2008 it is possible to read the sequences of DNA (genome) and RNA (transcriptome) with a sequencing machines generally described as Next Generation Sequencing (NGS). Previously it was possible on a much smaller scale with the Sanger method used since 1977. This method was used to decipher the first complete human genome (Human Genome Project) in the year 2000. In contrast to Sanger, the NGS approach generates millions of short (currently ca 150bp bp="nucleotide base pairs") sequences, called "reads". Those short reads need to be then analysed with the specific bioinformatics methods. For example assembled in bigger sequences or aligned to the known genomes.

Nowadays, a single most advanced commercially available sequencer (Illumina 4000) can generate ca 0.5T (T="tera"=1e12) of nucleotide base pairs in a single day of run. This makes possible reading ca. 12 full human genomes in 3.5-day long run of this machine [1].

On top of the sequence analysis there are many types of statistical analysis, which converts the information about the DNA and RNA sequences into the knowledge useful for life science or medicine researchers. NGS together with the analysis techniques, including the big data processing
is becoming the basis for the new area genomic personalised medicine It is also called also "precision medicine" as in the US president Obama speech from 30 January 2015, which announced the Precision Medicine Initiative. "The Precision Medicine Initiative will leverage advances in genomics, emerging methods for managing and analysing large data sets while protecting privacy, and health information technology to accelerate biomedical discoveries" [2].

The importance of this initiative and genome-based medical approaches in general is summarised in the rest of the White House statement:

"Precision medicine is an innovative approach to disease prevention and treatment that takes into account individual differences in people’s genes, environments, and lifestyles. Precision medicine gives clinicians tools to better understand the complex mechanisms underlying
a patient’s health, disease, or condition, and to better predict which treatments will be most effective. Advances in precision medicine have already led to powerful new discoveries and several new treatments that are tailored to specific characteristics of individuals, such as a person’s genetic makeup, or the genetic profile of an individual’s tumor. This is leading to a transformation in the way we can treat diseases such as cancer. Patients with breast, lung, and colorectal cancers, as well as melanomas and leukemias, for instance, routinely undergo molecular testing as part of patient care, enabling physicians to select treatments that improve chances of survival and reduce exposure to adverse effects."

Currently the price of sequencing is falling: 1000 dollar genome is reality since ca end of 2014, while the quality is growing due to more stable and tested laboratory protocols. The bottleneck of NGS is currently the process of getting the useful biological knowledge from the big sequencing datasets. The answer for the needs of fine-grained big data analysis in high throughput genomics is the development of data science and cloud computing. During the last years, the software development and data processing paradigms called map-reduce frameworks make possible to distribute the processing among the virtual cluster of machines and assure the scalability of results to the size of data. In last 2 years also the machine learning solutions have become scalable, with allows them to be run on hundreds and thousands of computers in parallel if needed.
This opens way to developing new, specialized machine learning approaches that will be able to solve new classes of NGS big data analysis problems.

So far typical techniques for NGS data analysis come so far mostly from the area of statistics. In the gEMiLia project I would like to combine my experience in genomic data analysis that I achieved during my PhD work with the modern approach to machine learning that is the specialty of Department of Information Systems where I work since 01.01.2004.

NGS and cloud computing have been among the 12 "disruptive technologies" in the McKinsey report published in 2013 [3]. It identifies 12 technologies that could drive truly massive economic transformations and disruptions in the coming years. My project is expected to combine them together and to add the extra factor, which is the use of solution from machine learning - currently quite established, but still vibrant research and technology area. Such combination is really novel and not investigated much yet in the global scientific community, as the knowledge of genome bioinformatics, cloud computing and machine learning is rare. I believe that it is possible to achieve a lot by getting a synergy effect of them, supported of course by my hard daily research work.

My plan is to focus on three data analysis cases: quality of NGS data, DNA variant analysis and RNA analysis, which have technologically a lot in common. The results will have mostly the form of novel algorithmic solutions that can be run on many computers in the cloud in parallel. This in consequence will allow to solve new classes of data analysis projects in molecular biology and personalised medicine.

1. Illumina 4000 specifications. http://www.illumina.com/systems/hiseq-3000-4000/specifications.html
2. FACT SHEET: President Obama’s Precision Medicine Initiative. Illumina 4000 specifications.
http://www.illumina.com/systems/hiseq-3000 4000/specifications.html
3. Disruptive technologies: Advances that will transform life, business, and the global economy.
http://www.mckinsey.com/insights/business_technology/disruptive_technologies