* Intro
  + We have as many genome sequenced as never before
  + Sequence genome – cure all diseases: didn’t happen
  + There are ..% coding and ..% non coding gens. For long time focus on coding (elaborate why wrong, examples)
  + Interpretation is needed (% of genes that we know functions of)
  + Among them pathogenic gens – need to find them
  + Many tools exist (examples)
  + Short on mendelian diseases
  + Lack of research on NCV causing Mendelian diseases (citations of existing – what they did[[1]](#footnote-1) (Caron et al. do XGBoost)
* Literature review
* Why it is important to use the new reference genome
* Theoretical background/ terminology
  + Mendelian desease
  + Coding/ non coding variants
  + SNP
* Data collection
  + Features
  + Databases
  + Snakemake workflow for data collection
  + Compare to previous paper
* Trained model
* Conclusion / Outlook

**Introduction**

When the Human Genome Project started in 1990the scientific world was excited about the possibilities that sequencing human genome will bring along. *Scientific code of human life is cracked by scientists* was the title of the cover article in New York Times on June 27, 2000. The expectations and hopes were boundless – beginning with*..* up to curation of countless diseases [[2]](#footnote-2).The reality though was sobering: scientists did ‘crack’ the sequence of human genome and covered almost *90%* of the genome, but this did not lead neither immediately nor many years after that to e.g. curation of the numerous diseases.

The reason for that is that simply knowing the sequence of nucleotide base pairs does not tell one anything about functionality of gens, pathogenicity of variants and many other factors that influence one’s phenotype. Cracking the genomic code to the extent that is needed for all those purposes is a much more complicated and tedious work than even the 15-year Genome Project was.

For many years biologists were focusing on particular type of genes and namely protein-coding genes. Those make out only 1% to 2% of human genome and were thought to be absolutely essential and solely responsible for any kind of disorders or traits that a living being carries. The rest of the genome was brand marked as ‘junk’ and was not studied for long time and even if this was refuted (citations of first papers) throughout many years knowledge/study/data bias towards protein-coding genes remains up to now.

Thus, the role of non-coding genes in the development/evolvement/expression of genes is highly underestimated (citation). At the moment there are only 25 non-coding variants that have been identified as causal in Mendelian diseases. At the same time approximately 88% of single nucleotide polymorphisms (SNPs) identified by genome-wide association studies (GWAS) lie in intergenetic or intronic regions (Edwards) and thus are likely to be regulatory/non-coding (which is not surprising since up to 99% of the genome is non-coding). Thus, there is a research gap in identification of pathogenic regulatory variants. This thesis aims at contributing to the research on the role regulatory genes in emerging/development of Mendelian diseases by updating a framework for identification of causal regulatory by integrating the new version of reference genome into the framework..

**Literature review**

Bed files – tab delimited files for analyses if certain positions in genes eg only exons

VCF files - tab delimited file with variants in one many samples

ToDo:

Comment Snakefile (why replace)

R to python

**Read or watch:**

<https://www.youtube.com/watch?v=hPrXcUUp70Y>

Youtube: die Merkhilfe

1. [↑](#footnote-ref-1)
2. [↑](#footnote-ref-2)