Leiden University Medical Center

Alex Stikkelman[1], Sander van Boom[1] and Núria Queralt-Rosinach[2]
[1] 4Medbox
[2] Biosemantics, Human Genetics, Leiden University Medical Center

Creating A Synthetic Realistic Mutated Population

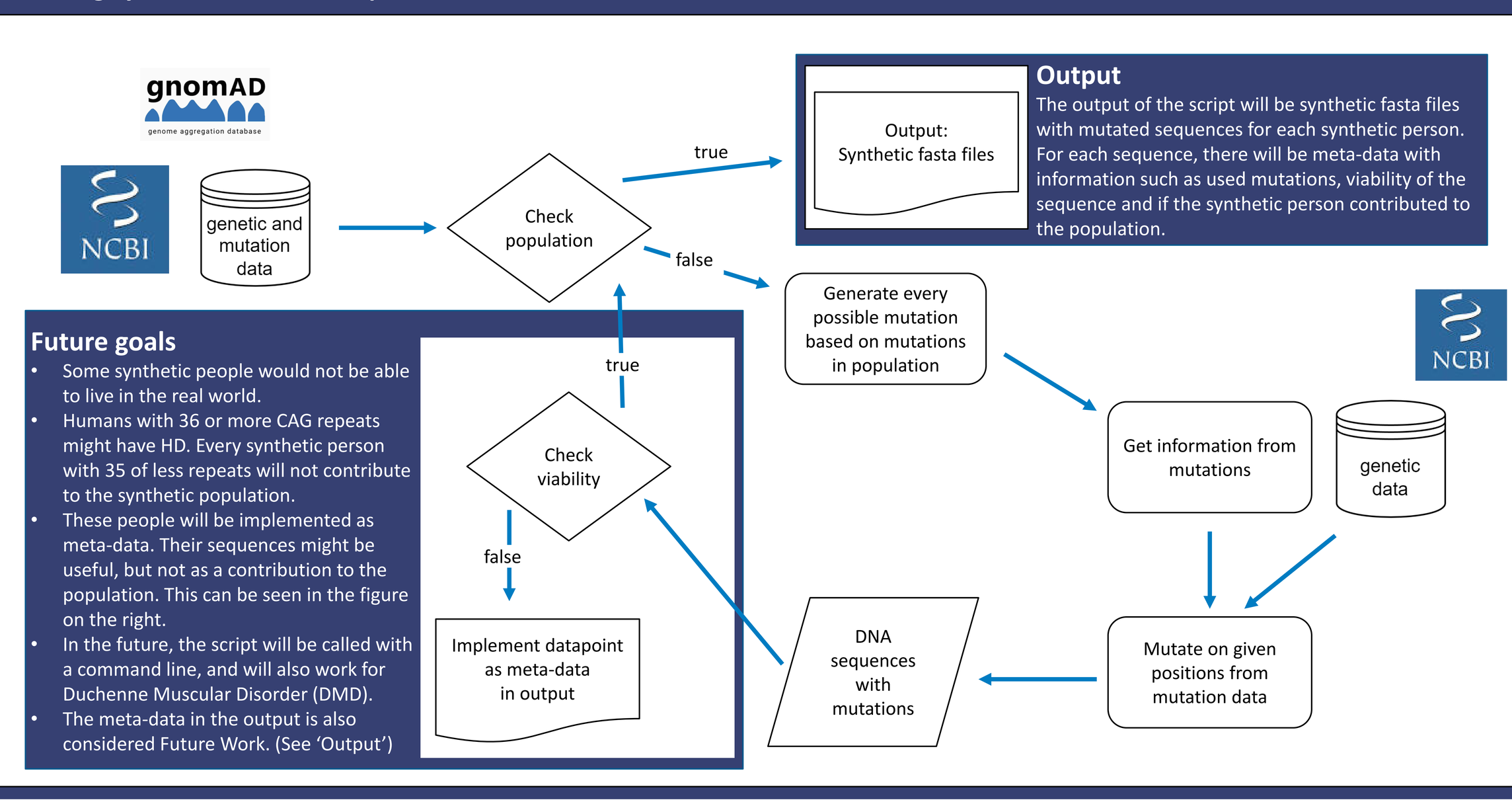
Introduction

- There is a need to have a lot of public biological research data to research and test tools on. It is not always available due to privacy concerns. Synthetic genetic data can be a solution for this problem.
- Huntington's Disease (HD) is a use-case for this problem. It is a rare neurodegenerative disease caused by an extended CAG repeat in the huntingtin (HTT) gene that manifests in young and middle-aged humans and is life-threatening. However, the script can perhaps also be used for other diseases, like Duchene Muscular Dystrophy (DMD).
- In this work, the workflow of the synthetic genetic data script will be showed with Huntington's Disease as a use-case, along side with a conclusion until now and future goals.

Synthetic research data

- There is a need for biological synthetic research data. The data can be DNA sequences, protein sequences or even synthetic hospital data, such as synthetic patient files.
- Synthetic genetic data is realistic generated data, based on real genetic data. The generated data looks like real data but is not in the genome of a real person.
- There is no privacy concern with synthetic data, because the data is derived from public available data sources like gnomAD and NCBI, and not from a real person.
- With more synthetic research data, there can be more research conducted on (rare) diseases or (rare) subtypes of diseases, even when there is not enough actual research data.

Creating synthetic mutated sequences



Output

The output of the script will be synthetic fasta files with mutated sequences for each synthetic person. For each sequence, there will be meta-data with information such as used mutations, viability of the sequence and if the synthetic person contributed to the population.

Future goals

- Some synthetic people would not be able to live in the real world.
- Humans with 36 or more CAG repeats might have HD. Every synthetic person with 35 of less repeats will not contribute to the synthetic population.
- These people will be implemented as meta-data. Their sequences might be useful, but not as a contribution to the population. This can be seen in the figure on the right.
- In the future, the script will be called with a command line, and will also work for Duchenne Muscular Disorder (DMD).
- The meta-data in the output is also considered Future Work. (See 'Output')

Huntington's Disease (HD)

- Someone will get HD if one of their parents has the HTT gene.
- One of the characteristics of Huntington's is the CAG-repeat. If someone has 36 or more repeats, they can have HD.
- If the individual has between 27 or 35 repeats, they are not likely to develop the disease, but they could still pass the repeats on to future generations.
- Usually, the protein Huntingtin is expressed, but with HD it gains a toxic functional phenotype.
- It is unclear what Huntingtin does. For example, it helps with the antero- and retrograde exon transport and it might also be a scaffold for transport cargo and motor proteins.

Duchene Muscular Dystrophy (DMD)

- DMD is a X-linked recessive, progressive disease caused by the Dystrophin (DMD) gene.
- The DMD protein is found mainly in skeletal and heart muscles.
- The protein is in small amounts located in nerve cells in the brain.
- The protein links the cytoskeleton of each muscle cell to the muscle membrane.
- The protein might play a role in cell signaling and protects muscle cells from injury while the muscle contracts and relaxes.
- Children with DMD will have trouble with walking and breathing. Eventually, the heart and respiratory muscles will stop working. Some patients also have problems with their cognitive functions, such as delayed speech and emotional or behavioral challenges.

Where are we now

- The script works with a small starting population of less than 15 individuals. For more than 15 infividuals, multi-threading or running the script on a server might be a good solution.
- There is more to be done, such as making the script compatible for both HD and DMD and generating output data alongside meta-data.
- The script can be used for retrieving data from the gnomAD database, which can be connected to other databases.
- If time allows it, there will be a good validation method implemented. However, there is no validation method yet for synthetically created DNA sequences. The validation method needs to be researched, designed and programmed.

hogeschool Leiden
4MedBox