Caitlin Jenster[1,3], Rick Overkleeft[1,2] and Núria Queralt-Rosinach[2]
[1] 4Medbox
[2] Biosemantics, Human Genetics, Leiden University Medical Center
[3] University of Applied science Leiden

Updating The SynthDNASim tool to create diverse DNA data sets

Introduction

In genomic research of rare diseases like Huntington's Disease (HD), it is common to have a lack of genomic data to start and perform research with. One of the reasons for this is the privacy issues surrounding the obtainment and use of genomic data. Thus, a possible solution to this is the development of a tool that can create diverse synthetic DNA data sets so the created synthetic DNA data set can be used for genomic research and make the research a lot more reproducible and faster. [1]
However, the synthetic DNA data set needs to be diverse enough to represent different populations. A single disease can have many different genetic characteristics because of differences in and between populations. Thus, factors of genetic evolution and ancestry need to be taken into account while creating a diverse DNA data set. [2]

The goal of this project is to create a diverse DNA data set including genetic evolution and ancestry factors. With HD as a use case, the data set will contain disease variants from European, African, and Middle Eastern populations. Here, we show the workflow of the diversified synthetic DNA simulator (SynthDNASim). This new tool updates the Synthetic DNA Simulator (see 'Materials and Methods' below) to help researchers create synthetic DNA representative of population diversity. The next steps are to validate the synthetic DNA data set and apply the FAIR principles to make it Findable Accessible Interoperable and Reusable for the community using semantic technologies.

Huntington's disease (HD)

HD is a rare disease that is hereditary and causes degeneration of nerve cells in the brain. Because of this, HD has an impact on the functional abilities of an individual, resulting in movement, cognitive, and mental disorders. HD is caused by an extended CAG repeat within the Huntingtin gene (HTT gene). [3] There are many different variants of HD within different populations, these variants are known as haplotypes. A haplotype is a combination of single nucleotide polymorphisms (SNP's, variants) in a gene that is found on the same chromosome. [4]

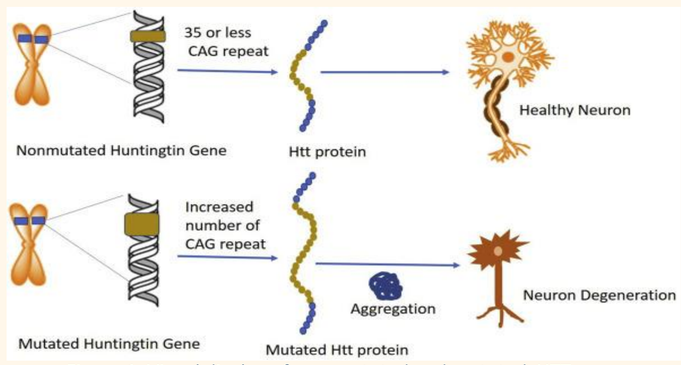
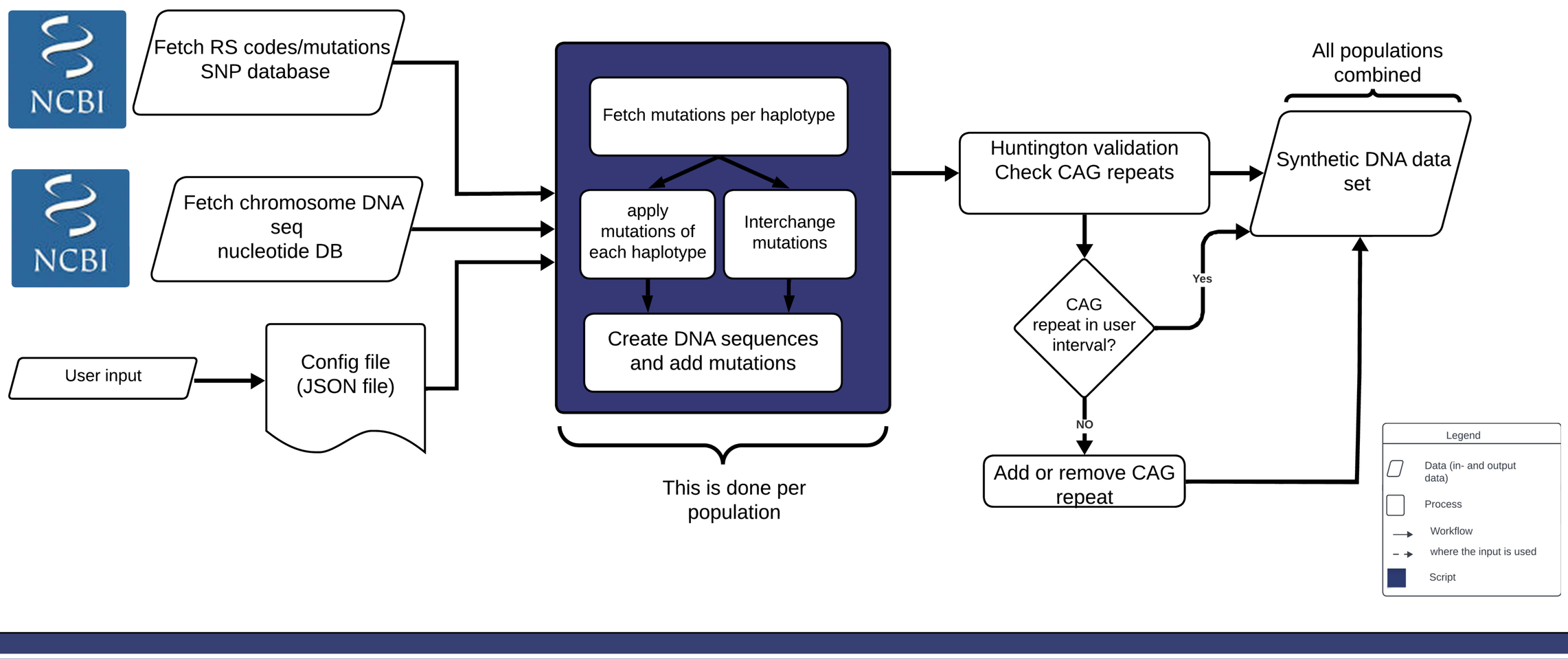


Figure 1. Visual display of non-mutated and mutated HTT gene. This figure also displays the protein and the effect of the protein on neurons (source: https://pubmed.ncbi.nlm.nih.gov/33659724/).

Creating diverse DNA sequences



Materials and Methods:

- Python scripts
- Original Synthetic DNA Simulator (scan QR code)
- NCBI SNP database and nucleotide database
- The FAIR principles will be a guiding factor within this project
- Each sequence has its metadata including haplotype, genetic variants, CAG repeats, gene, chromosome, etc. ic variants, CAG repeats, gene, chromosome, etc.

Future work:

The remaining work of this project:

- Validation of the generated synthetic DNA
- Use semantic methods and tools to make the project more FAIR. Create metadata using Data Catalog Vocabulary (DCAT) and share the tool and metadata using WorkflowHub and Common Workflow Language (CWL). [5-7]

For future research:

- Determine a standard protocol to validate the created synthetic DNA data set.
- Add synthetic clinical data.

References:

1. Walonoski J, Kramer M, Nichols J, et al. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*. 2018;25(3):230-238. doi:10.1093/jamia/ocx079

2. Squitieri F, Mazza T, Maffi S, et al. Tracing the mutated HTT and haplotype of the African ancestor who spread Huntington disease into the Middle East. *Genet Med*. 2020;22(11):1903-1908. doi:10.1038/s41436-020-0895-1

3. Young AB. Huntingtin in health and disease. *J Clin Invest*. 2003;111(3):299-302. doi:10.1172/JCI200317742

4. Haplotype. Genome.gov. Accessed September 21, 2023. https://www.genome.gov/genetics-glossary/haplotype

5. Albertoni R, Browning D, Cox S, et al. Data Catalog Vocabulary (DCAT) - Version 3. W3.org. Published January 18, 2024. Accessed February 7, 2024. https://www.w3.org/TR/vocab-dcat-3/

6. Peter Amstutz, Michael R. Crusoe, Nebojša Tijanić (editors), Brad Chapman, John Chilton, Michael Heuer, Andrey Kartashov, Dan Leehr, Hervé Ménager, Maya Nedeljkovich, Matt Scales, Stian Soiland-Reyes, Luka Stojanovic (2016): Common Workflow Language, v1.0. Specification, Common Workflow Language working group. https://w3id.org/cwl/v1.0/ https://doi.org/10.6084/m9.figshare.3115156.v2

7. Carole Goble, Stian Soiland-Reyes, Finn Bacall, Stuart Owen, Alan Williams, Ignacio Eguinoa, Bert Droesbeke, Simone Leo, Luca Pireddu, Laura Rodríguez-Navas, José Mª Fernández, Salvador Capella-Gutierrez, Hervé Ménager, Björn Grüning, Beatriz Serrano-Solano, Philip Ewels, & Frederik Coppens. (2021). Implementing FAIR Digital Objects in the EOSC-Life Workflow Collaboratory. Zenodo. https://doi.org/10.5281/zenodo.4605654

Acknowledgments

We want to thank Alex Stikkelman and the 4MedBox team for all their support and help. We also want to thank Ivo Fokkema and the Biosemantics group at the LUMC for their input and help in this project. This project received funding from 4MedBox. N. Queralt-Rosinach is supported by funding from the European Union's Horizon 2020 research and innovation program under the EJP RD COFUND-EJP N° 825575 and by a grant from the European Union's Horizon 2020 research and innovation program under grant agreement No 847826 (Brain Involvement iN Dystrophinopathies (BIND)). We would like to thank the EJP RD and BIND for supporting research on generating synthetic health data for rare disease research.