# Updating The SynthDNASim tool to create diverse synthetic DNA datasets

Caitlin Jenster[a,b,c], Rick Overkleeft[a,c] and Núria Queralt-Rosinach[c]

[a] *4MedBox Nederland B.V, Kanaalpark 157, Leiden, 2321 JW, Netherlands*
[b] *University of Applied science Leiden, Zernikedreef 11, Leiden, 2333 CK, Netherlands*
[c] *Leiden University Medical Center, Albinusdreef 2, Leiden, 2333 ZA, Netherlands* 9

**Abstract**

In biomedical research, it is common to perform numerous analyses of genomic data, for example, to understand the cause of a particular disease. Regulatory laws protect the privacy of individuals but hinder access to genomic data. One solution to this is the development of bioinformatic tools to create synthetic DNA data. One of the challenges is to capture genomic diversity representative of differences within and between populations, especially for rare genetic diseases. In this study, we present SynthDNASim, a tool for creating diverse synthetic DNA datasets. Our approach is to create diverse DNA datasets taking into account factors of genetic evolution and ancestry with Huntington's disease (HD) as a use case. In particular, with HD variants from European, African, and Middle Eastern populations. We will show our tool and future plans on applying semantic methods and tools to make SynthDNASim more FAIR (Findable Accessible Interoperable Reusable).

**Keywords**

Diverse synthetic DNA dataset, privacy, Huntington's Disease, evolution, ancestry, FAIR, semantics

## 1. Introduction

In biomedical research, it is common to perform numerous analyses of genomic data, for example, to understand the cause of a particular disease, or genetic processes, or to identify gene variants. One of the difficulties in these analyses is the collection, storage, use, and reuse of genomic data because an individual's genomic data is private. Especially if an individual has a rare disease like Huntington's disease (HD) it is theoretically possible to retrace the DNA to this individual. Thus, there are regulatory laws that protect the privacy of these individuals but hinder access to genomic data. One solution to this is the development of bioinformatic tools to create diverse synthetic DNA data so that researchers in biomedical research can create synthetic DNA data and make the research faster and more reproducible. [1] A possible issue with creating a synthetic DNA dataset is that it needs to be diverse enough to be representative of different populations. A single disease can have many different genetic characteristics because of differences in and between populations and because genetic diseases are characterized by their phenotype (symptoms). Thus, factors of genetic evolution and ancestry need to be taken into account while creating a diverse DNA dataset. [2] In this study, we present SynthDNASim, a tool for creating diverse synthetic DNA datasets. Our approach is to create diverse DNA datasets with HD as a use case. In particular, with HD variants from European, African, and Middle Eastern populations. We will show our workflow and future plans for synthetic DNA dataset validation. The FAIR principles and semantics will be applied within this project to make the tool understandable, reusable, reviewable, and open-source. HD is a rare disease that is hereditary and causes degeneration of nerve cells in the brain. Because of this, HD has a great 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]

SWAT4HCLS 2024: The 15th International Conference on Semantic Web Applications and Tools for Health Care and Life Science, February 26–29, 2024, Leiden, Netherlands

EMAIL: caitlin@4medbox.eu (A. 1); rick@4medbox.eu (A. 2); n.queralt_rosinach@lumc.nl (A. 3) ORCID: 0009-0009-7863-0962 (A. 1); 0009-0004-3529-1159 (A. 2); 0000-0003-0169-8159 (A. 3)

## 2. SynthDNASim tool

In Figure 1 we illustrate the SynthDNASim pipeline. The first step is the retrieval and pre-processing of genomic information from different data sources: National Center for Biotechnology Information (NCBI) for the SNP variants, NCBI for the sequence of Chromosome 4, and lastly the user input. Next is a sequence of steps to create the synthetic DNA sequences per population. Python is used for the user input, creating the config file (JSON file). Each sequence has its own metadata including haplotype, genetic variants, CAG repeats, gene, chromosome, etc.

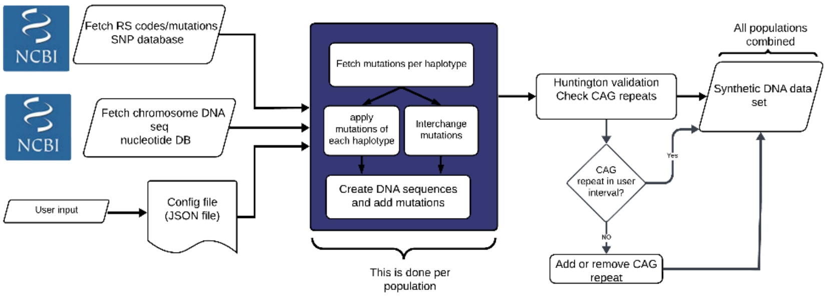
NCBI
Fetch RS codes/mutations SNP database
NCBI
Fetch chromosome DNA seq nucleotide DB
User input
Config file (JSON file)
Fetch mutations per haplotype
apply mutations of each haplotype
Interchange mutations
Create DNA sequences and add mutations
This is done per population
Huntington validation Check CAG repeats
CAG repeat in user interval?
Yes
NO
Add or remove CAG repeat
All populations combined
Synthetic DNA data set

**Figure 1:** SynthDNASim pipeline.

## 3. Future works

The remaining work of this project is to perform a validation on the generated data and to use semantic methods and tools to make the project more FAIR. One option for this is to create the metadata for the output data and the tool. For the creation of the output metadata Data Catalog Vocabulary (DCAT) can be used. [4]

## 4. Acknowledgements

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 programme under grant agreement No 847826 (Brain Involvement iN Dystrophinopathies (BIND)). We would like to thank to the EJP RD and BIND for supporting research on generating synthetic health data for rare diseases research.

## 5. References

[1] J. Walonoski, M. Kramer, J. Nichols, A. Quina, C. Moesel, D. Hall, C. Duffett, K. Dube, T. Gallagher, S. McLachlan, Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record, J. Am. Med. Inform. Assoc. 25.3 (2017) 230–238. doi:10.1093/jamia/ocx079.
[2] F. Squitieri, T. Mazza, S. Maffi, A. De Luca, Q. AlSalmi, S. AlHarasi, J. A. Collins, C. Kay, F. Baine-Savanhu, B. G. Landwhermeyer, et al., Tracing the mutated HTT and haplotype of the African ancestor who spread Huntington disease into the Middle East, Genet. Med. 22.11 (2020) 1903–1908. doi:10.1038/s41436-020-0895-1.
[3] A. B. Young, Huntingtin in health and disease, J. Clin. Investig. 111.3 (2003) 299–302. doi:10.1172/jci17742.
[4] 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/