NextGen

Genome-Centric Multimodal
Data Integration in Personalised
Cardiovascular Medicine.

Deliverable D3.5
Synthetic datasets for testing and piloting-1

Grant Agreement Number: 101136962

Funded by the European Union

UK Research and Innovation

The British associated partners of NextGen were funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant agreements No. 10098097, No. 10104323]

Project funded by

Schweizerische Eidgenossenschaft
Confédération suisse
Confederazione Svizzera
Confederaziun svizra

Swiss Confederation

Federal Department of Economic Affairs, Education and Research EAER
State Secretariat for Education, Research and Innovation SERI

NextGen	
Project full title	Next Generation Tools for Genome-Centric Multimodal Data Integration In Personalised Cardiovascular Medicine
Call identifier	HORIZON-HLTH-2023-TOOL-05-04
Type of action	RIA
Start date	01/01/2024
End date	31/12/2027
Grant agreement no	101136962

Funding of associated partners
The Swiss associated partners of the NextGen project were funded by the Swiss State Secretariat for Education, Research and Innovation (SERI). The British associated partners of NextGen were funded by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant agreements No 10098097, No 10104323]

D3.5 – Synthetic datasets for testing and piloting-1	
Author(s)	Marco Scutari
Editor	Francesca Mangili
Participating partners	SUPSI
Version	1.0
Status	Final
Deliverable date	M12
Dissemination Level	PU - Public
Official date	2024-12-03
Actual date	2024-12-03

Disclaimer

The NEXTGEN consortium consists of the following partners:

No	PARTNER ORGANISATION NAME	ABBREVIATION	COUNTRY
1	UNIVERSITAIR MEDISCH CENTRUM UTRECHT	UMCU	NL
2	HIRO MICRODATACENTERS B.V.	HIRO	NL
3	EURECOM GIE	EURE	FR
4	JOHANN WOLFGANG GOETHE-UNIVERSITAET FRANKFURT AM MAIN	GUF	DE
5	KAROLINSKA INSTITUTET	KI	SE
6	HUS-YHTYMA	HUS	FI
7	UNIVERSITY OF VIRGINIA	UVA	US
8	KLINIKUM RECHTS DER ISAR DER TECHNISCHEN UNIVERSITAT MUNCHEN	TUM-Med	DE
9	HL7 INTERNATIONAL FOUNDATION	HL7	BE
10	MYDATA GLOBAL RY	MYDTA	FI
11	DATAPOWER SRL	DPOW	IT
12	SOCIETE EUROPEENNE DE CARDIOLOGIE	ESC	FR
13	WELLSPAN HEALTH	WSPAN	US
14	LIKE HEALTHCARE RESEARCH GMBH	LIKE	DE
15	NEBS SRL	NEBS	BE
16	THE HUMAN COLOSSUS FOUNDATION	HCF	CH
17	SCUOLA UNIVERSITARIA PROFESSIONALE DELLA SVIZZERA ITALIANA	SUPSI	CH
18	DRUG INFORMATION ASSOCIATION	DIA	CH
19	DPO ASSOCIATES SARL	DPOA	CH
20	QUEEN MARY UNIVERSITY OF LONDON	QMUL	UK
21	EARLHAM INSTITUTE	ERLH	UK

Document Revision History

DATE	VERSION	DESCRIPTION	CONTRIBUTIONS
03/12/2024	1.0	Complete draft.	SUPSI

Authors

AUTHOR/EDITOR	ORGANISATION
Marco Scutari	SUPSI
Francesca Mangili	SUPSI

Reviewers

REVIEWER	ORGANISATION

List of terms and abbreviations

ABBREVIATION	DESCRIPTION
AI	Artificial intelligence
EDHS	European Health DataSpace
FL	Federated (machine) learning
GWAS	Genome-wide association study
KPI	Key Performance Indicator
ML	Machine learning
MMIO	Multi-model integration object
PGS	Polygenic (risk) score
VCF	Variant call format
WP	Work Package

Table of contents

1 Summary

This document reviews the generation of synthetic data (task 3.5) and its use for piloting federated genome-wide association studies (GWAS; October 2024, task 5.4). We produced low-fidelity sequence (human DNA) data to demonstrate a proof of concept whose purpose was to show the feasibility of performing a federated GWAS to find genes associated with a binary disease indicator. Given the performance and functional limitations of the software we adapted for the demo, it was impractical to work with high-fidelity or real-world data. Using real-world data would also require additional administrative burden without providing materially different insights.

2 Background

Generating artificial data is an essential capability in statistical genetics. The ability to produce data sets with the desired sample sizes, genotyping densities, genetic architectures and population structures makes it possible to develop, test and benchmark new statistical and machine learning models effectively and efficiently. In addition, synthetic data can be generated not to pose privacy risks and allows for reproducibility of results and comparisons on shared, public benchmarks. For this reason, there has been much research on how to generate artificial genomic data, from Markov Chain Monte Carlo (MCMC) [1] to deep learning [2] methods.

HAPGEN2 [1] is a simulation software that generates artificial data sets for case-control studies. It extends the earlier Li-Stephens [3] algorithm for modelling linkage disequilibrium using hidden Markov models and posterior sampling to generate single-nucleotide variants (SNV) with realistic linkage disequilibrium patterns. Despite its age, it is still the de facto standard in cutting-edge research in polygenic prediction [4-6], fine mapping [7-8] and missing data imputation [9]. SNVs are variations in the genome sequence spanning a single nucleotide, which can be one of adenine, cytosine, thymine and guanine (A, C, T, G). Typically, they will appear in two possible states (alleles) in a population.

HAPGEN2 is described in the Model Card in the Appendix.

3 Synthetic Data Generation

HAPGEN2 requires three files as inputs:

- a haplotype file containing the alleles for all SNVs from a population panel;
- a legend file for the SNV markers containing their ID, position, and the labels of the major and minor alleles;
- a file containing the fine-scale mapping of the SNVs to establish the recombination rate (so, the linkage disequilibrium) across the relevant regions.

The overwhelming majority of the literature uses the panels from the Thousand Genomes (1KG Phase 3) [10] and HapMap 3 [11] projects. We used the latter for the federated GWAS demo, generating the SNVs from chromosome 18 for 1000 individuals without specific risk factors. The generated genotypes and phenotypes can be easily

converted to BED/BIM/FAM or PED/MAP files, one of the standard formats used by PLINK2 [12] and other software for genome-wide association studies and genomic prediction.

HAPGEN2 has additional arguments to control the number of cases and controls, the number and position of pathogenic SNVs and their relative risk. However, how the effect sizes of the pathogenic SNV are allocated is very limited and can only generate a single binary phenotype (that is, a 0/1 disease indicator) at a time. To better control the distribution of the SNV effects, the liability and the heritability, we followed the approach from [13]. After producing the genotypes with HAPGEN2, we generated an underlying quantitative trait with the desired liabilities, controlling its heritability and then converted it into a binary phenotype. The conversions is done by thresholding the liabilities with the approriate quantile to achieve the desired prevalence in the sample. The same approach can be used to create quantitative phenotypes.

Both the panel and the generated synthetic data are described in the Data Cards in the Appendix. The code is included in the Appendix as well.

4 References

1. Su Z, Marchini J, Donnelly P. HAPGEN2: simulation of multiple disease SNPs. Bioinformatics. 2011 Aug 15;27(16):2304-5. https://mathgen.stats.ox.ac.uk/genetics_software/hapgen/hapgen2.html

2. Yelmen B, Jay F. An overview of deep generative models in functional and evolutionary genomics. Annual Review of Biomedical Data Science. 2023 Aug 10;6(1):173-8.

3. Li N, Stephens M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics. 2003 Dec 1;165(4):2213-33.

4. Ruan Y, Lin YF, Feng YC, Chen CY, Lam M, Guo Z, He L, Sawa A, Martin AR, Qin S. Improving polygenic prediction in ancestrally diverse populations. Nature genetics. 2022 May;54(5):573-80.

5. Zhang H, Zhan J, Jin J, Zhang J, Lu W, Zhao R, Ahearn TU, Yu Z, O'Connell J, Jiang Y, Chen T. A new method for multiancestry polygenic prediction improves performance across diverse populations. Nature genetics. 2023 Oct;55(10):1757-68.

6. Hoggart CJ, Choi SW, García-González J, Souaiaia T, Preuss M, O'Reilly PF. BridgePRS leverages shared genetic effects across ancestries to increase polygenic risk score portability. Nature Genetics. 2024 Jan;56(1):180-6.

7. Hernández N, Soenksen J, Newcombe P, Sandhu M, Barroso I, Wallace C, Asimit JL. The flashfm approach for fine-mapping multiple quantitative traits. Nature communications. 2021 Oct 22;12(1):6147.

8. Zhou F, Soremekun O, Chikowore T, Fatumo S, Barroso I, Morris AP, Asimit JL. Leveraging information between multiple population groups and traits improves fine-mapping resolution. Nature Communications. 2023 Nov 10;14(1):7279.

9. Lam M, Awasthi S, Watson HJ, Goldstein J, Panagiotaropoulou G, Trubetskoy V, Karlsson R, Frei O, Fan CC, De Witte W, Mota NR. RICOPILI: rapid imputation for COnsortias PIpeLIne. Bioinformatics. 2020 Feb 1;36(3):930-3.

10. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, Flicek P, Gabriel SB. A global reference for human genetic variation. Nature. 2015;526(7571):68.

11. International HapMap 3 Consortium. Integrating common and rare genetic variation in diverse human populations. Nature. 2010 Sep 9;467(7311):52.

12. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. Gigascience. 2015 Dec 1;4(1):s13742-015.

13. Speed D, Cai N, Ucleb Consortium, Johnson MR, Nejentsev S, Balding DJ. Reevaluation of SNP heritability in complex human traits. Nature genetics. 2017 Jul 1;49(7):986-92.

14. Cavinato T, Rubinacci S, Malaspinas AS, Delaneau O. A resampling-based approach to share reference panels. Nature Computational Science. 2024 May 14:1-7.

15. Wharrie S, Yang Z, Raj V, Monti R, Gupta R, Wang Y, Martin A, O'Connor LJ, Kaski S, Marttinen P, Palamara PF. HAPNEST: efficient, large-scale generation and evaluation of synthetic datasets for genotypes and phenotypes. Bioinformatics. 2023 Sep 1;39(9):btad535.

16. Meyer HV, Birney E. PhenotypeSimulator: A comprehensive framework for simulating multi-trait, multi-locus genotype to phenotype relationships. Bioinformatics. 2018 Sep 1;34(17):2951-6.

5 Appendix

This appendix contains a Model Card for HAPGEN 2 and a Data Card for the syntethic data described in Section "Synthetic Data Generation".

5.1 Model Card

Model Card - HAPGEN2	
Model Details	
Developer	Zhan Su, Jonathan Marchini, Peter Donnelly
Version	2.2.0
URL	https://mathgen.stats.ox.ac.uk/genetics_software/hapgen/hapgen2.html
Licence	https://mathgen.stats.ox.ac.uk/genetics_software/hapgen/LICENCE (forbids commercial use)
Type	Generates artificial data sets for case-control studies by applying an extension of the Li-Stephens algorithm to a panel of SNV data.
Reference	[1]
Intended Use	
– Intended for use in academic research on human genetics, commercial use is forbidden. – Potentially unsuitable for inbred species, polyploid species and recombinant species. – Standard panels HapMap 3 and 1000 Genomes are recommended but no longer provided.	
Factors	
– It can only generate SNV data for the SNVs present in the input panel, within a specified range of allele frequencies and without missing data. – It provides very limited options for assigning phenotypes to the generated genotypes. – The genotypes will have the same patterns of linkage disequilibrium and average allele frequencies as the panel provided to HAPGEN. Therefore, the population the input panel comes should be chosen to reflect the target population for the specific application (CEU, YRI, CHJB, JPT, etc.).	
Training Data	
– User-provided panel of genome-wide SNV data. The panel data used to generate the data is described in the Data Card below.	

5.2 Data Card

Data Card – Input Panel	
Panel	HapMap 3 (release 2) haplotypes - NCBI Build 36 (dbSNP b126)
Population	CEU
Chromosome	18
Sample size	1301 individuals, 99454 SNVs
Phenotypes	None.
URL (alternative to HAPGEN's)	https://www.sanger.ac.uk/data/hapmap-3/
Licence	International HapMap Project Public Access License (copy: http://www.worldlii.org/int/other/PubRL/2003/4.html) This licence allows use of the data but not redistribution. The data must be donwloaded from the original source to reproduce the analysis described earlier in the report.
Notes	No medical or personal identifying information was obtained from individuals providing the samples. However, the samples are identified by the population from which they were collected.

Data Card – Generated Synthetic Data	
Panel	-
Population	CEU
Chromosome	18
Sample size	1000 individuals, the first 5000 SNVs in the chromosome.
Phenotypes	Binary phenotype with 1000 causative loci placed randomly in the chromosome, heritability ~60% and prevalence 25%.
URL (alternative to HAPGEN's)	https://www.sanger.ac.uk/data/hapmap-3/
Licence	HAPGEN's licence applies.

5.3 Computer Code

The following code calls HAPGEN2 to generate the SNVs.

```
#!/bin/bash -e

set -o pipefail

HAPGEN_URL=https://mathgen.stats.ox.ac.uk/genetics_software/hapgen/download/builds/x86_64/v2.2.0/
hapgen2_x86_64.tar.gz
HAPGEN=`basename $HAPGEN_URL`
OUTDIR=./output
PREFIX=hapmap
CASES=0
CONTROLS=1000

# download and extract HAPGEN 2.
wget -c $HAPGEN_URL
tar xvzf $HAPGEN hapgen2
# extract the HAPMAP panel.
gzip -d CEU.0908.chr18.hap.gz
# create a directory in which to save the intermediate files.
mkdir -p $OUTDIR
# generate the genotypes.
valgrind ./hapgen2 -h CEU.0908.chr18.hap \
   -l CEU.0908.chr18.legend \
   -m genetic_map_chr18_combined_b36.txt \
   -o $OUTDIR/$PREFIX \
   -dl 441 1 1.5 2.5 1124 0 3 4.5 1149 1 2.0 6.5 \
   -n $CONTROLS $CASES

# cut and format the data in 0/1/2 allele counts on columns
cut -d" " -f6- $OUTDIR/$PREFIX.controls.gen > $OUTDIR/formatted
split -l 500 $OUTDIR/formatted $OUTDIR/formatted2.

for FILE in `find $OUTDIR -type f -name "formatted2.*" | sort`
do
 awk '
 {
    for (i = 1; i <= NF; i++)  {
       a[NR, i] = $i;
    }
 }
 END {
    for(j = 1; j <= NF; j = j + 3) {
       str = ""
       for(i = 1; i <= NR; i++) {
          acount = a[i, j] * 0 + a[i, j + 1] * 1 + a[i, j + 2] * 2;
          str = str""acount" ";
       }
```

```
        print str;
      }
  }' $FILE > `sed -e "s/2/3/" <<< $FILE`
done

# merge the data back into a single file.
cut -d" " -f 2 $OUTDIR/$PREFIX.controls.gen | tr '\n' ' ' > $OUTDIR/$PREFIX.final.gen
echo >> $OUTDIR/$PREFIX.final.gen
paste -d" " $OUTDIR/formatted3* >> $OUTDIR/$PREFIX.final.gen
# compress and copy to the final location.
gzip -k -9 $OUTDIR/$PREFIX.final.gen
cp $OUTDIR/$PREFIX.final.gen.gz raw-$PREFIX.CEU.chr18.txt.gz
# clean up.
rm -rf output
```

The following code generates the PLINK2 files from the outputs of HAPGEN2.

```
# load phenotypes and genotypes.
pheno = readRDS("phenotypes.rds")
geno = readRDS("prepd-hapmap.rds")

geno = geno[, 10000 + 1:5000]

# FAM file.
contents = data.frame(
  FAMID = paste0("FAM", sprintf("%04d", seq(nrow(geno)))),
  WFID = rep(1, nrow(geno)),
  FA = rep(0, nrow(geno)),
  MO = rep(0, nrow(geno)),
  SEX = sample(c("F", "M"), nrow(geno), replace = TRUE),
  PHENO = as.integer(pheno$pheno)
)

contents$SEX = as.integer(factor(contents$SEX, levels = c("M", "F")))

write.table(contents, file = "generated.fam", quote = FALSE, row.names = FALSE,
  col.names = FALSE)

# PED file (FAM + calls).
for (g in seq(ncol(geno))) {

  calls = sapply(geno[, g], function(g) {

    if (g == 0)
      return(c(1, 1))
    else if (g == 2)
      return(c(2, 2))
    else
      return(sample(1:2))

  })

  contents = cbind(contents, t(calls))

}
```

```
write.table(contents, file = "generated.ped", quote = FALSE, row.names = FALSE,
  col.names = FALSE)

# MAP file.
legend = read.table("CEU.0908.chr18.legend", header = TRUE)

contents = data.frame(
  CHR = rep(18, ncol(geno)),
  ID = legend[seq(ncol(geno)), "ID"],
  MORGAN = rep(0, ncol(geno)),
  BP = legend[seq(ncol(geno)), "pos"]
)

write.table(contents, file = "generated.map", quote = FALSE, row.names = FALSE,
  col.names = FALSE)

# BIM file.
contents = cbind(contents, legend[seq(ncol(geno)), c("allele0", "allele1")])

write.table(contents, file = "generated.bim", quote = FALSE, row.names = FALSE,
  col.names = FALSE)

# finally, run: ./plink --file generated --make-bed --out binary
```

The following code generates the phenotypes from the SNVs.

```
# load the markers.
markers = readRDS("prepd-hapmap.rds")

generate.binary.phenotypes =
  function(markers, nqtl = 1000, size = 0.048, prevalence = 0.25) {

  # ... generate the noise term...
  noise = rnorm(nrow(markers), sd = 1)
  # ... sample the QTLs and the associated effect sizes...
  betas = rep(0, ncol(markers))
  betas[sample(ncol(markers), nqtl)] = rnorm(nqtl, sd = size)
  # ... and generate the continuous liabilities as intermediate phenotypes.
  liability = as.vector(as.matrix(markers) %*% betas + noise)
  # check the heritability of the phenotype.
  cat("@ estimated heritability is:",
      (var(liability) - var(noise)) / var(liability), "\n")
  # create the binary phenotypes from the liability to achieve the desired prevalence (a la LDAK).
  thr = quantile(liability, probs = 1 - prevalence)
  pheno = cut(liability, breaks = c(-Inf, thr, Inf), labels = c("CTRL", "CASE"))
  return(list(liability = liability, pheno = pheno, betas = betas))

}#GENERATE.BINARY.PHENOTYPES

pheno = generate.binary.phenotypes(markers)

saveRDS(pheno, file = "phenotypes.rds")
```

6 Conclusions and next steps

This report details how to successfully generate SNVs from a panel of sequence data using HAPGEN2, and how to produce phenotypes with set heritability, liability and prevalence from a set of disease-causing variants.

We note that more modern approaches for generating genotypes have been proposed in recent literature: RESHAPE [14] and HAPNEST [15] are two notable examples. While less explored, they may be of interest in NextGen because they are more efficient in generating large samples and can handle larger panels. Fewer options exist for generating realistic phenotypes from genotypes; the most cited example is PhenotypeSimulator [16]. HAPNEST can also generate phenotypes together with genotypes. Currently, no software or method for generating phenotypes for modalities other than numeric (say, image phenotypes) exists in the literature.