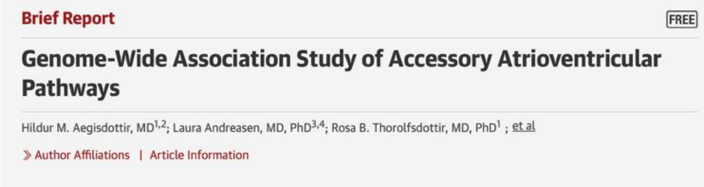
Genotype Imputation

Filling out missing data in an informed manner

Ioannis Louloudis
PhD fellow
Section for Health Data Science and AI,
Department of Public Health

UNIVERSITY OF COPENHAGEN

University of Copenhagen, Department of Public Health

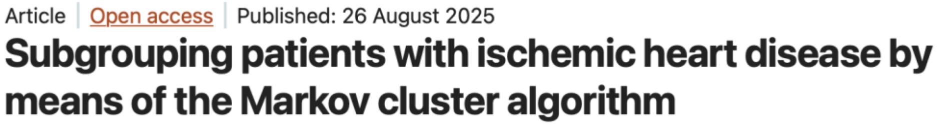


Brief Report FREE

Genome-Wide Association Study of Accessory Atrioventricular Pathways

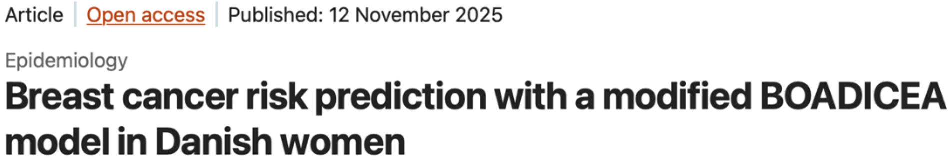
Hildur M. Aegisdottir, MD[1,2]; Laura Andreasen, MD, PhD[3,4]; Rosa B. Thorolfsdottir, MD, PhD[1] ; et al

» Author Affiliations | Article Information



Article | Open access | Published: 26 August 2025

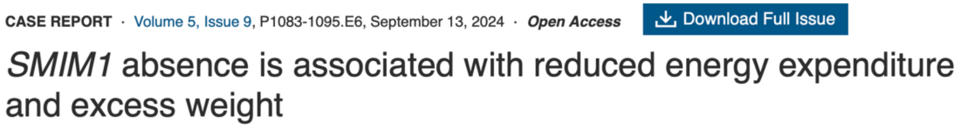
Subgrouping patients with ischemic heart disease by means of the Markov cluster algorithm



Article | Open access | Published: 12 November 2025

Epidemiology

Breast cancer risk prediction with a modified BOADICEA model in Danish women



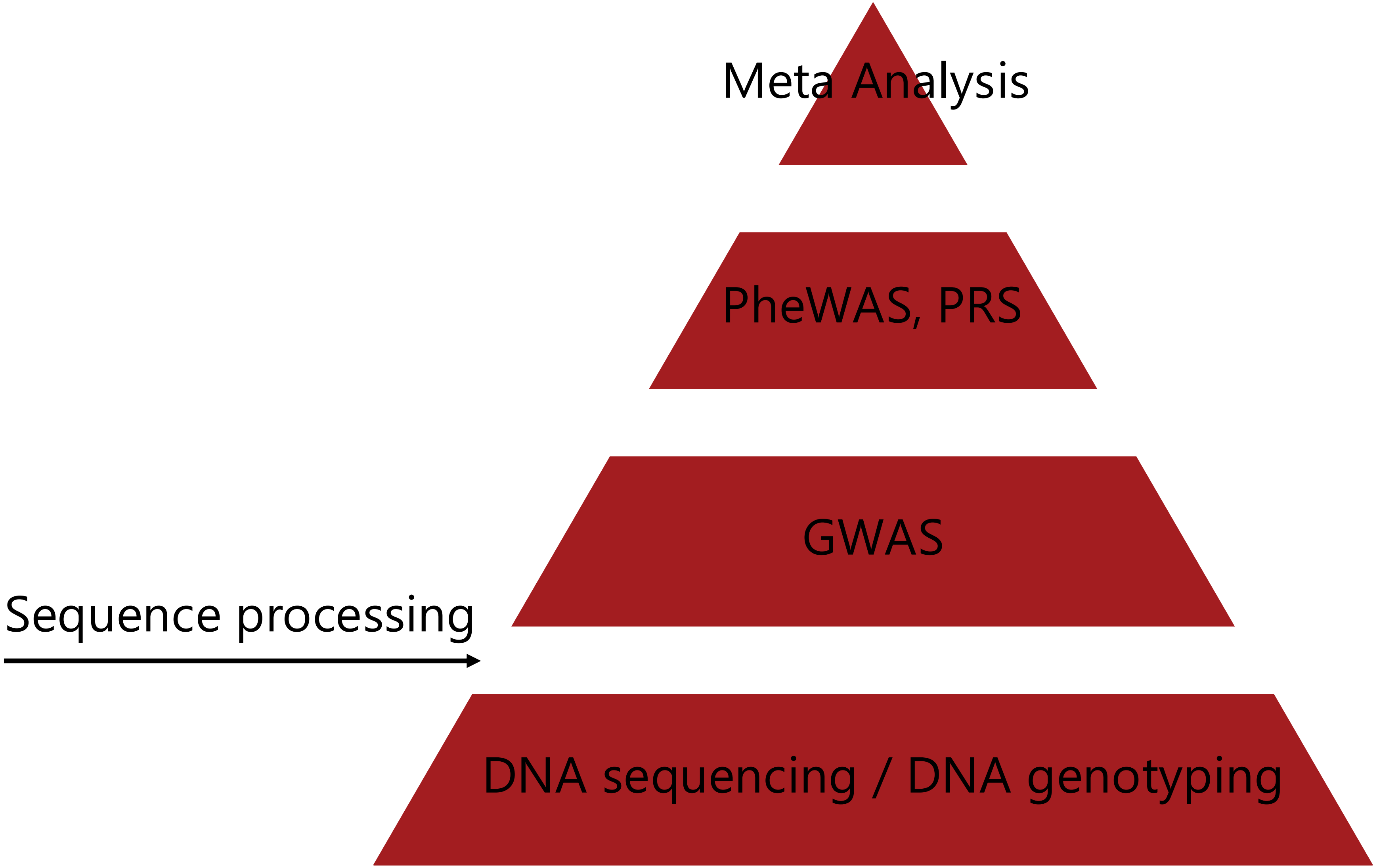
CASE REPORT · Volume 5, Issue 9, P1083-1095.E6, September 13, 2024 · *Open Access* Download Full Issue

SMIM1 absence is associated with reduced energy expenditure and excess weight

Research interests

- Drug repurposing
- Systems biology & Precision medicine
- Survival analysis modelling
- Genetics:
  - Genotype imputation
  - Polygenic Risk Score Estimation
  - Genome-Wide Association Studies
  - Phenome-Wide Association Studies
- Early-onset pancreatic cancer prediction
- Big data approach to reviewing of case reports
- Birth control side-effect identification using laboratory test values

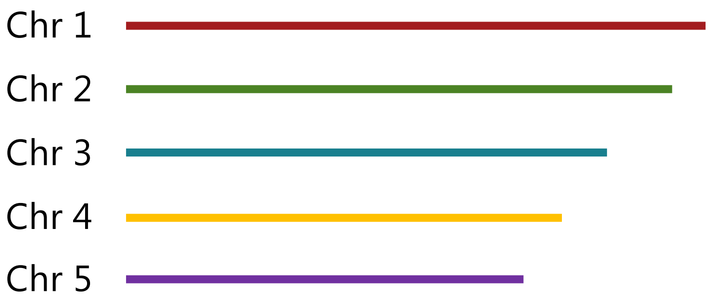
Population genetics pipeline



Sequencing vs *Genotyping*

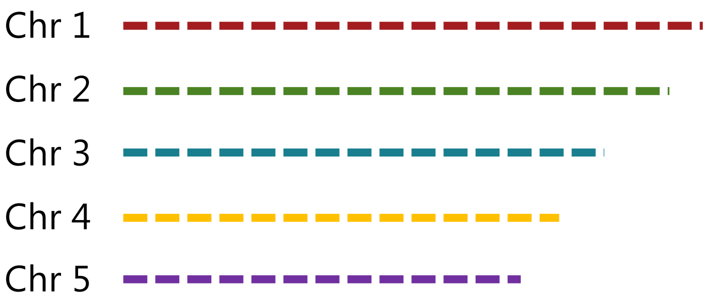
Sequencing

Reading short/long stretches of the genome.



Genotyping

Identification of bases at specific loci of the genome.



Sequencing vs *Genotyping*

Sequencing

Reading short/long stretches of the genome.

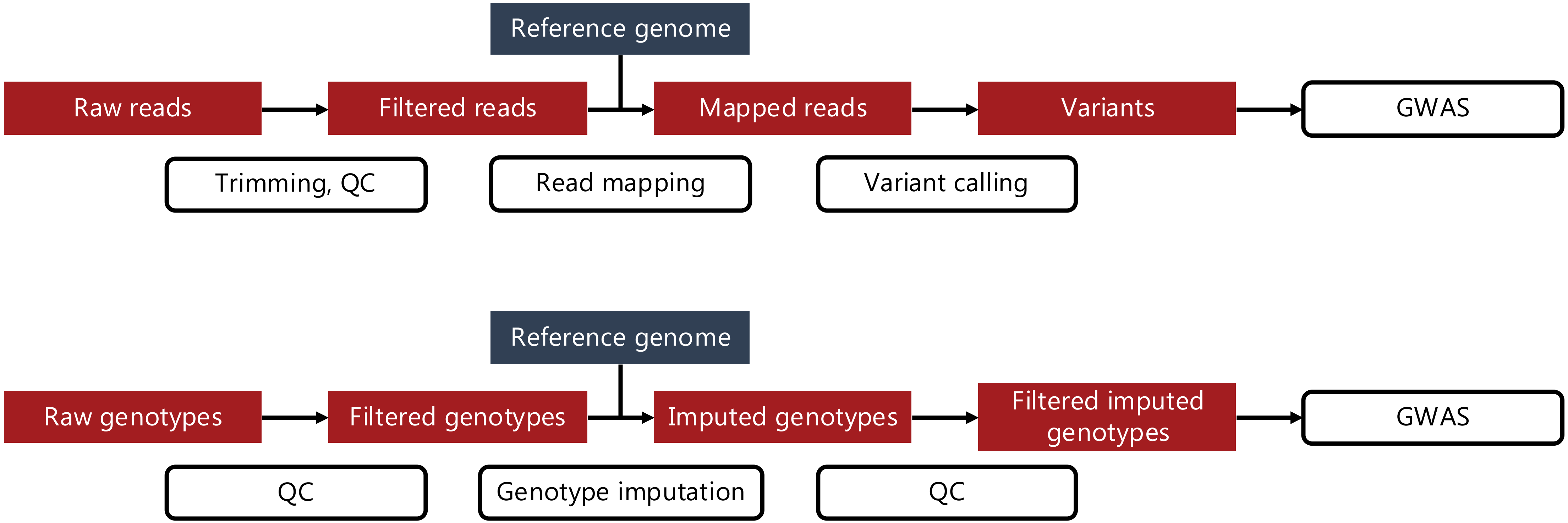
- 🎯 Low - High
- 🕓 Long reading and processing times
- 💵 Medium - High

Genotyping

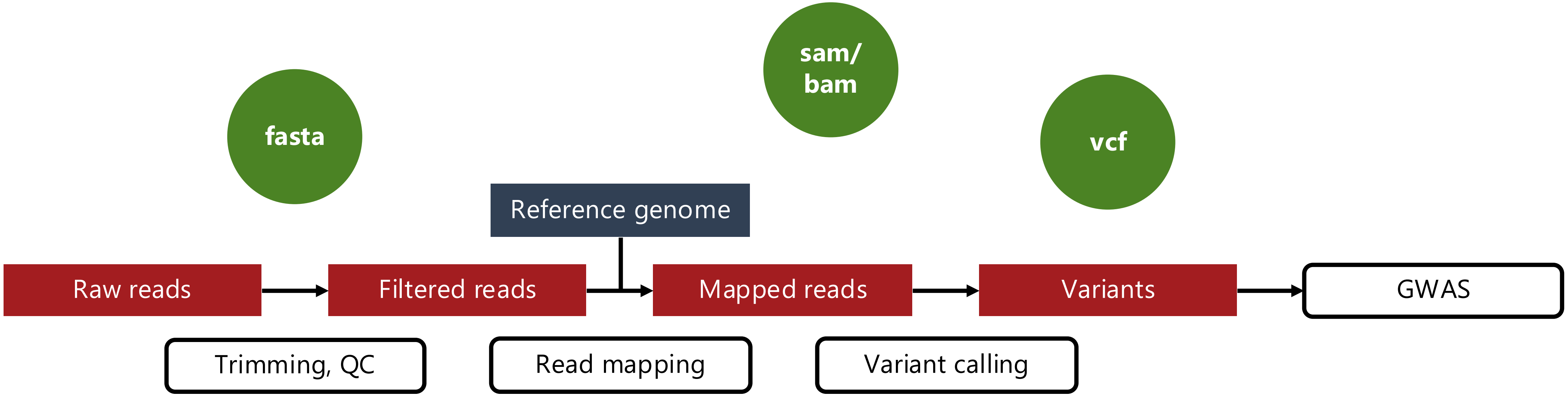
Identification of bases at specific loci of the genome.

- 🎯 Very High
- 🕓 Short time
- 💵 Very low

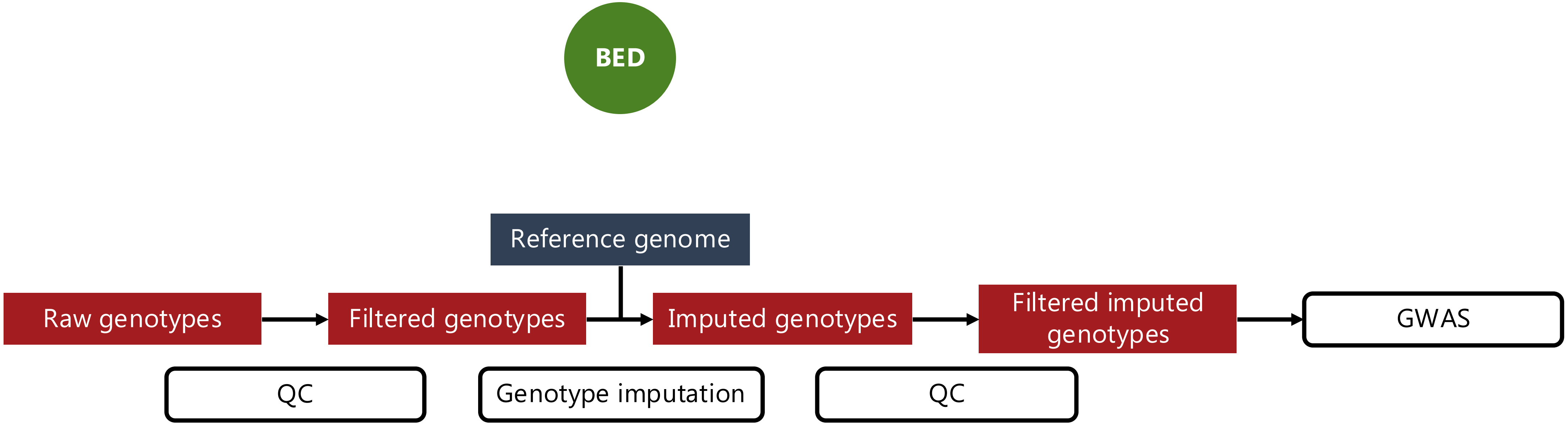
Overview of a typical population genomics pipeline



Overview of a typical population genomics pipeline

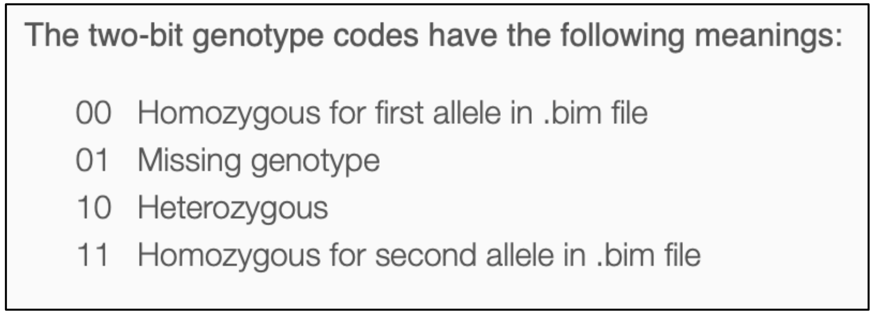


Overview of a typical population genomics pipeline



Plink file formats BED/BIM/FAM

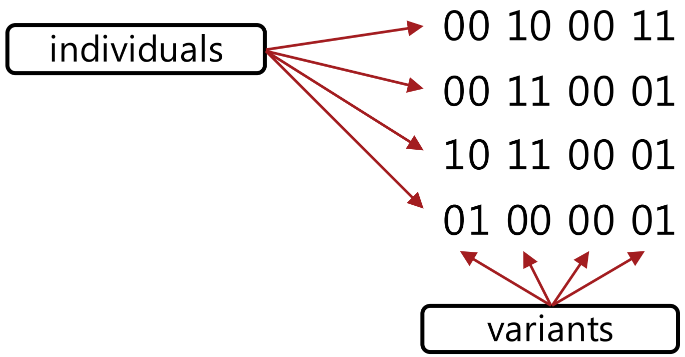
- **BED**
  - Binary file
  - Cannot be (easily) viewed



The two-bit genotype codes have the following meanings:

00	Homozygous for first allele in .bim file
01	Missing genotype
10	Heterozygous
11	Homozygous for second allele in .bim file

- For example:

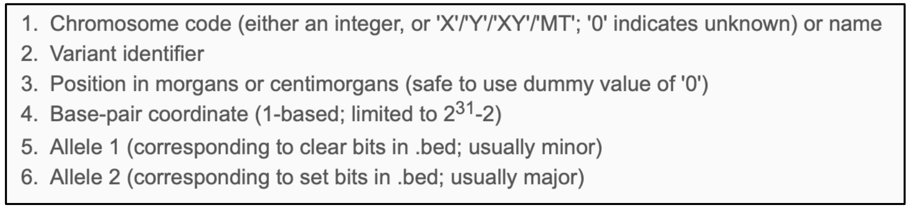


individuals

```
00 10 00 11
00 11 00 01
10 11 00 01
01 00 00 01
```

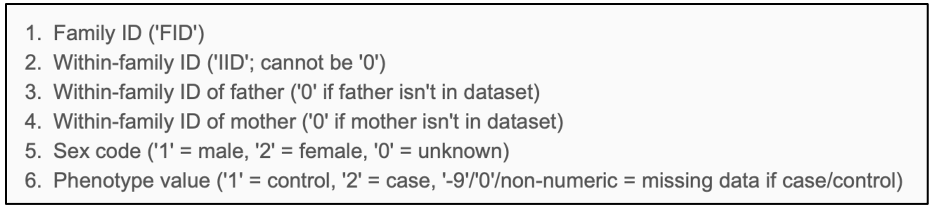
variants

- **BIM**
  - A text file, tab-separated
  - Variant information



1. Chromosome code (either an integer, or 'X'/'Y'/'XY'/'MT'; '0' indicates unknown) or name
2. Variant identifier
3. Position in morgans or centimorgans (safe to use dummy value of '0')
4. Base-pair coordinate (1-based; limited to 2^{31}-2)
5. Allele 1 (corresponding to clear bits in .bed; usually minor)
6. Allele 2 (corresponding to set bits in .bed; usually major)

- **FAM**
  - A text file, tab-separated
  - Individual (patient) information

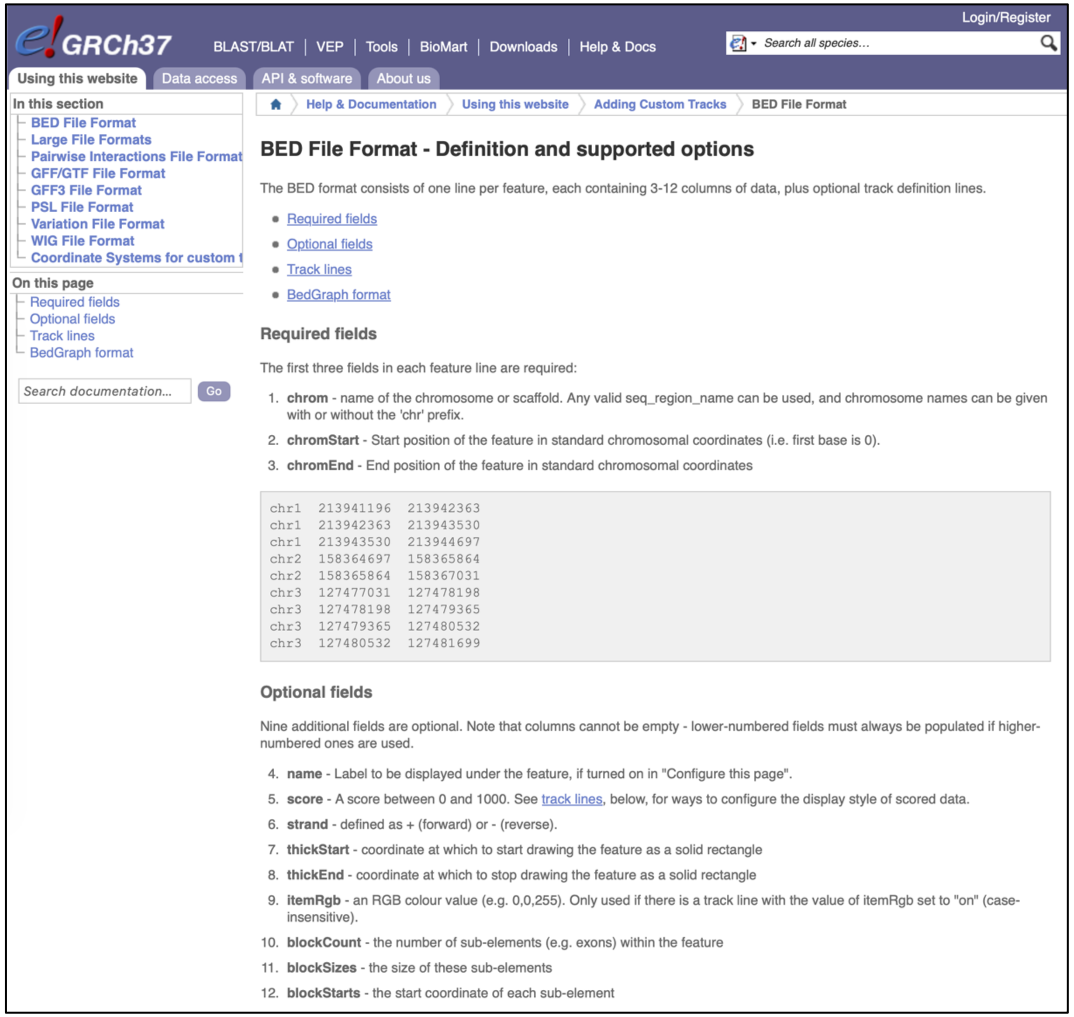


1. Family ID ('FID')
2. Within-family ID ('IID'; cannot be '0')
3. Within-family ID of father ('0' if father isn't in dataset)
4. Within-family ID of mother ('0' if mother isn't in dataset)
5. Sex code ('1' = male, '2' = female, '0' = unknown)
6. Phenotype value ('1' = control, '2' = case, '-9'/'0'/non-numeric = missing data if case/control)

https://www.cog-genomics.org/plink/1.9/

Beware!

- There is another BED file format in Bioinformatics.



BED File Format - Definition and supported options

The BED format consists of one line per feature, each containing 3-12 columns of data, plus optional track definition lines.

- Required fields
- Optional fields
- Track lines
- BedGraph format

Required fields

The first three fields in each feature line are required:

1. **chrom** - name of the chromosome or scaffold. Any valid seq_region_name can be used, and chromosome names can be given with or without the 'chr' prefix.
2. **chromStart** - Start position of the feature in standard chromosomal coordinates (i.e. first base is 0).
3. **chromEnd** - End position of the feature in standard chromosomal coordinates

```
chr1  213941196  213942363
chr1  213942363  213943530
chr1  213943530  213944697
chr2  158364697  158365864
chr2  158365864  158367031
chr3  127477031  127478198
chr3  127478198  127479365
chr3  127479365  127480532
chr3  127480532  127481699
```

Optional fields

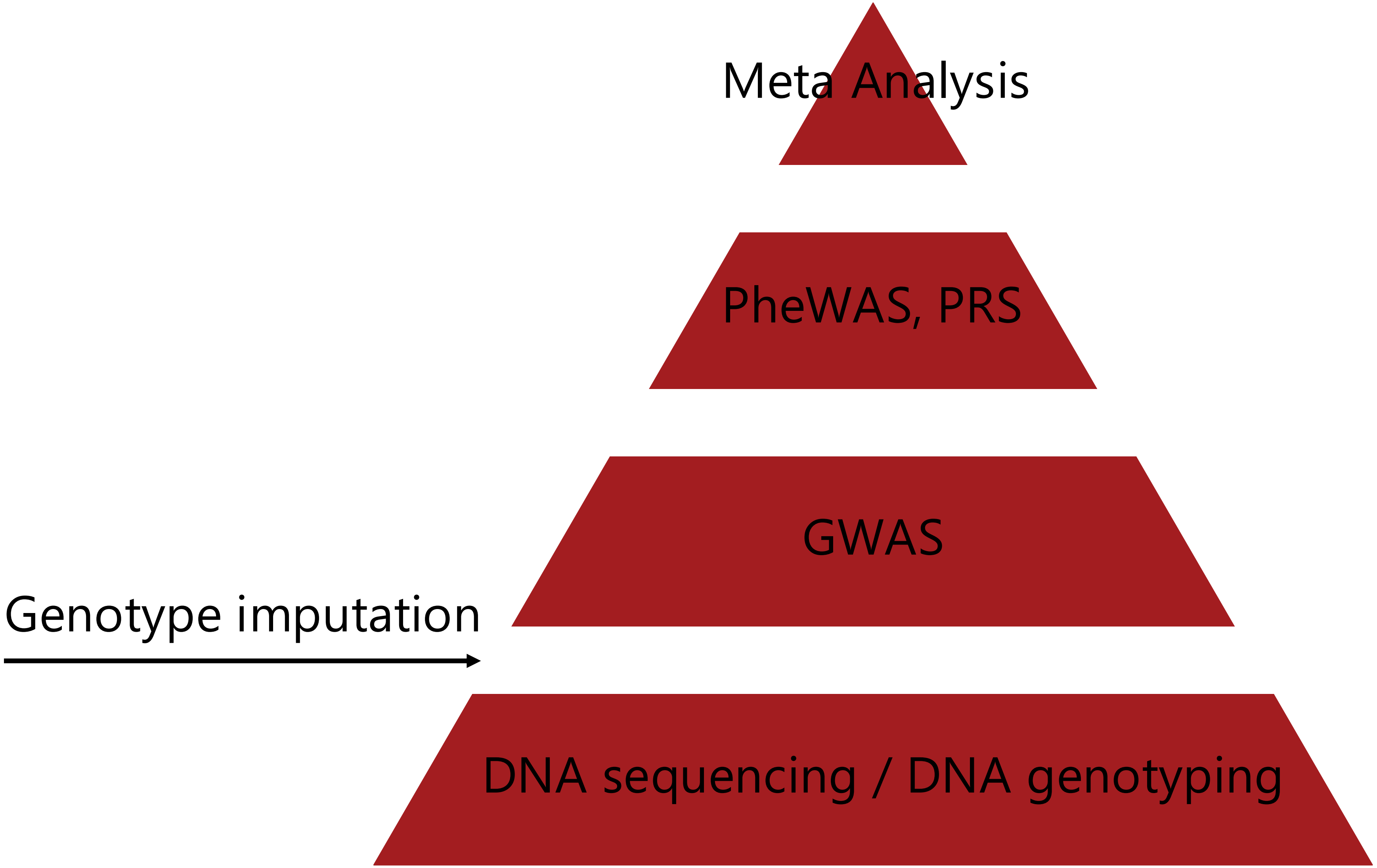
Nine additional fields are optional. Note that columns cannot be empty - lower-numbered fields must always be populated if higher-numbered ones are used.

4. **name** - Label to be displayed under the feature, if turned on in "Configure this page".
5. **score** - A score between 0 and 1000. See track lines, below, for ways to configure the display style of scored data.
6. **strand** - defined as + (forward) or - (reverse).
7. **thickStart** - coordinate at which to start drawing the feature as a solid rectangle
8. **thickEnd** - coordinate at which to stop drawing the feature as a solid rectangle
9. **itemRgb** - an RGB colour value (e.g. 0,0,255). Only used if there is a track line with the value of itemRgb set to "on" (case-insensitive).
10. **blockCount** - the number of sub-elements (e.g. exons) within the feature
11. **blockSizes** - the size of these sub-elements
12. **blockStarts** - the start coordinate of each sub-element

Exercise Break

Exercise 1: Set-up & Data exploration

Population genetics pipeline



Meta Analysis

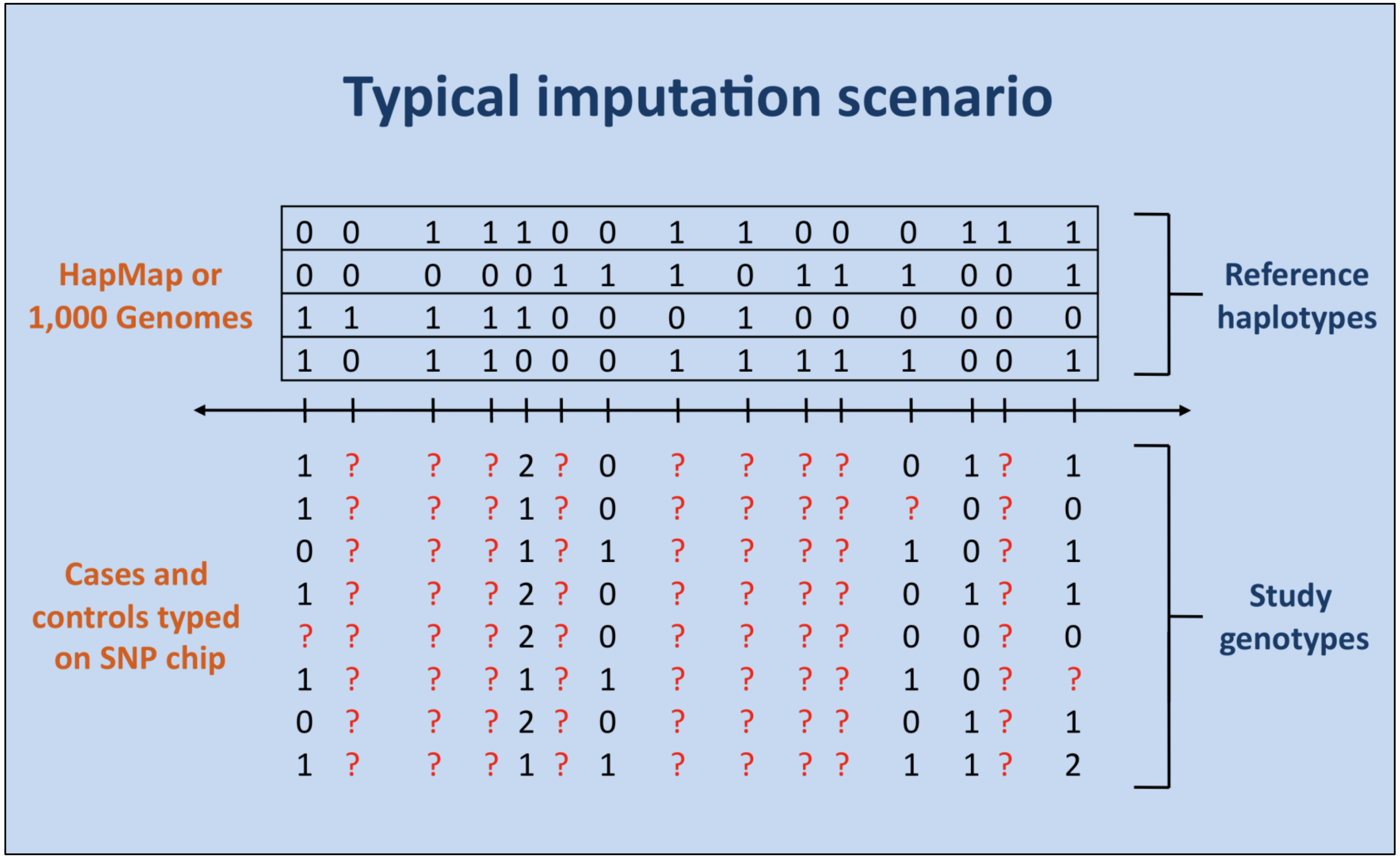
PheWAS, PRS

GWAS

Genotype imputation

DNA sequencing / DNA genotyping

Genotype Imputation



Typical imputation scenario

HapMap or 1,000 Genomes

Reference haplotypes

0	0	1	1	1	0	0	1	1	0	0	0	1	1	1
0	0	0	0	0	1	1	1	0	1	1	1	0	0	1
1	1	1	1	1	0	0	0	1	0	0	0	0	0	0
1	0	1	1	0	0	0	1	1	1	1	1	0	0	1

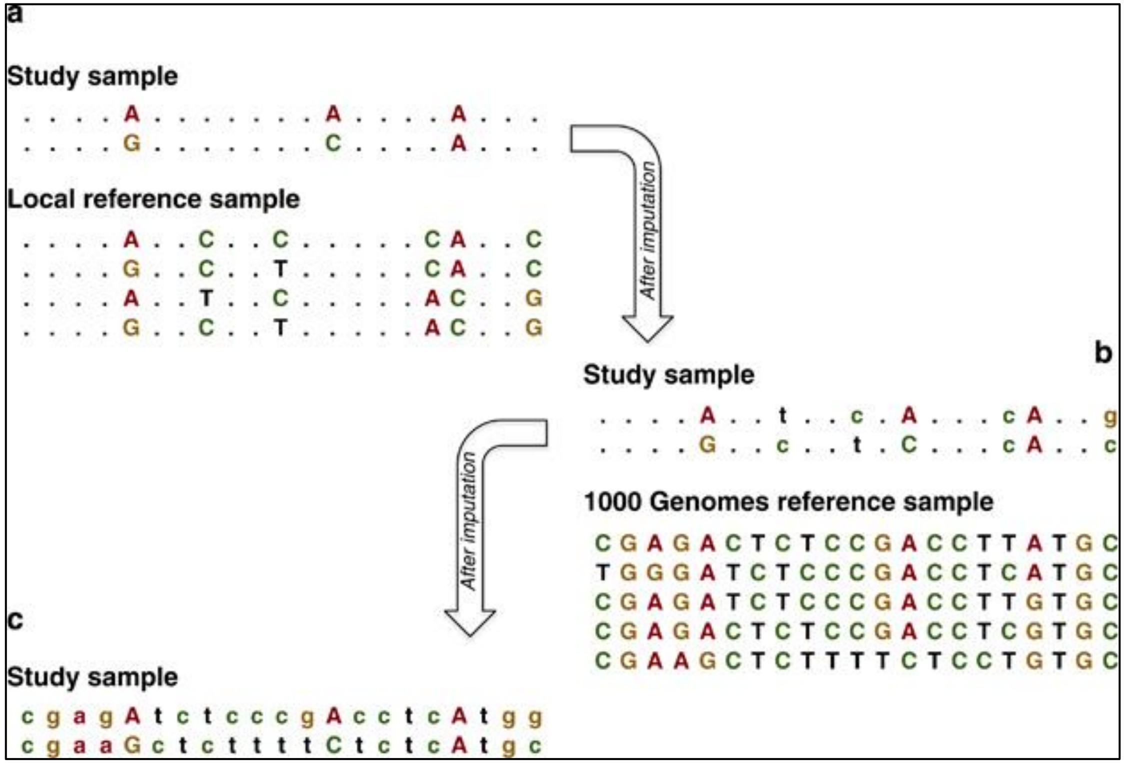
Cases and controls typed on SNP chip

Study genotypes

1	?	?	?	2	?	0	?	?	?	?	0	1	?	1
1	?	?	?	1	?	0	?	?	?	?	?	0	?	0
0	?	?	?	1	?	1	?	?	?	?	1	0	?	1
1	?	?	?	2	?	0	?	?	?	?	0	1	?	1
?	?	?	?	2	?	0	?	?	?	?	0	0	?	0
1	?	?	?	1	?	1	?	?	?	?	1	0	?	?
0	?	?	?	2	?	0	?	?	?	?	0	1	?	1
1	?	?	?	1	?	1	?	?	?	?	1	1	?	2

https://mathgen.stats.ox.ac.uk/impute/impute_v2.html

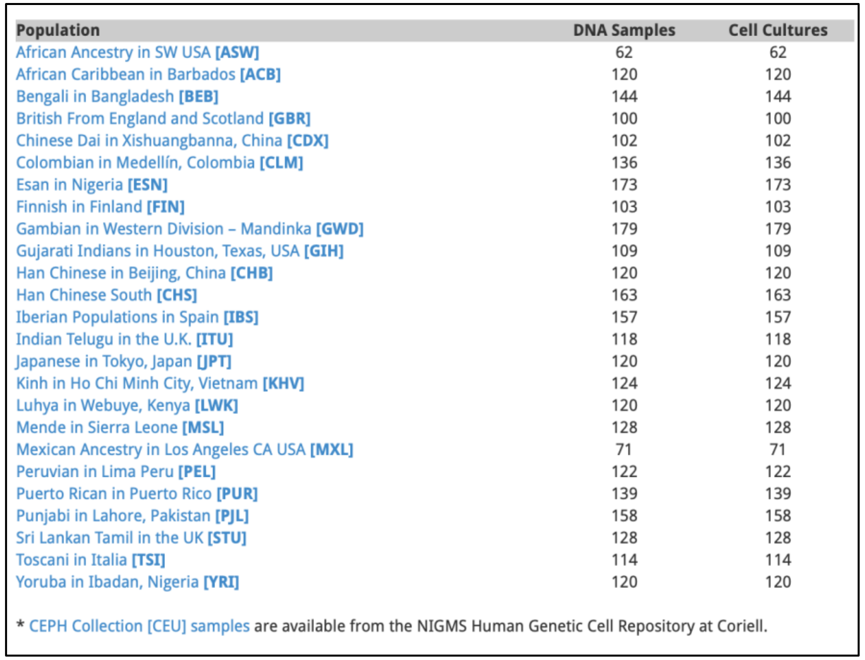
Genotype Imputation


a
Study sample
. . . . A . . . . A . . . . A . . .
. . . . G . . . . C . . . . A . . .
Local reference sample
. . . . A . . C . . C . . . . . C A . . C
. . . . G . . C . . T . . . . . C A . . C
. . . . A . . T . . C . . . . . A C . . G
. . . . G . . C . . T . . . . . A C . . G
After imputation
b
Study sample
. . . . A . . t . . c . A . . . . c A . . g
. . . . G . . c . . t . C . . . . c A . . c
1000 Genomes reference sample
C G A G A C T C T C C G A C C T T A T G C
T G G G A T C T C C C G A C C T C A T G C
C G A G A T C T C C C G A C C T T G T G C
C G A G A C T C T C C G A C C T C G T G C
C G A A G C T C T T T T C T C C T G T G C
After imputation
c
Study sample
c g a g A t c t c c c g A c c t c A t g g
c g a a G c t c t t t t C t c t c A t g c

Improving accuracy of rare variant imputation with a two-step imputation approach

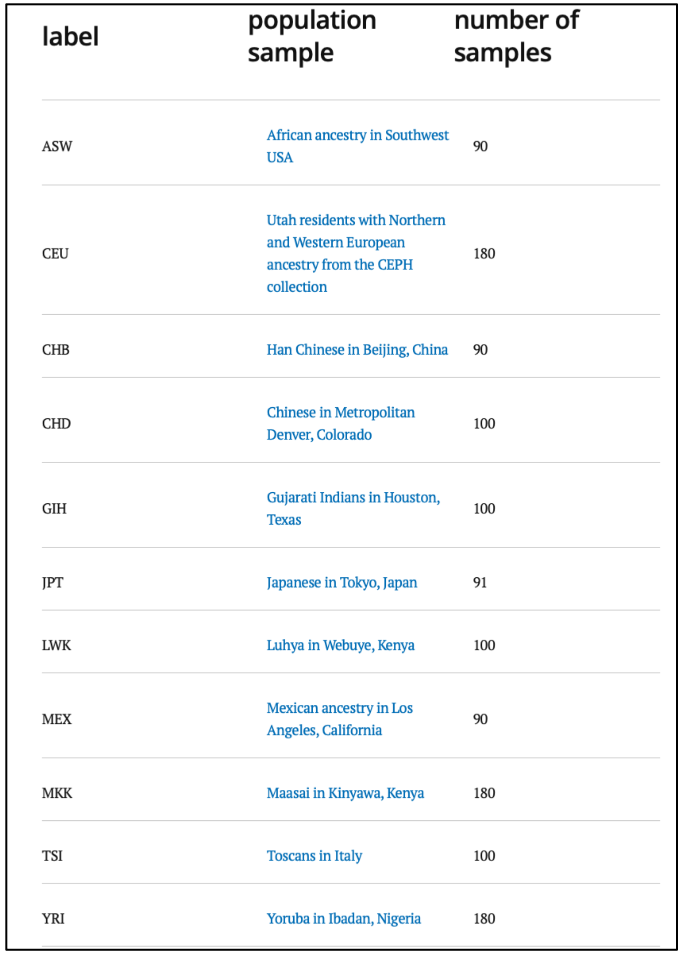
Reference panel

- Keyword: *Ancestry*
- Big (and thus generic) reference panels:
  - HapMap3
  - 1000Genomes



Population	DNA Samples	Cell Cultures
African Ancestry in SW USA **[ASW]**	62	62
African Caribbean in Barbados **[ACB]**	120	120
Bengali in Bangladesh **[BEB]**	144	144
British From England and Scotland **[GBR]**	100	100
Chinese Dai in Xishuangbanna, China **[CDX]**	102	102
Colombian in Medellín, Colombia **[CLM]**	136	136
Esan in Nigeria **[ESN]**	173	173
Finnish in Finland **[FIN]**	103	103
Gambian in Western Division – Mandinka **[GWD]**	179	179
Gujarati Indians in Houston, Texas, USA **[GIH]**	109	109
Han Chinese in Beijing, China **[CHB]**	120	120
Han Chinese South **[CHS]**	163	163
Iberian Populations in Spain **[IBS]**	157	157
Indian Telugu in the U.K. **[ITU]**	118	118
Japanese in Tokyo, Japan **[JPT]**	120	120
Kinh in Ho Chi Minh City, Vietnam **[KHV]**	124	124
Luhya in Webuye, Kenya **[LWK]**	120	120
Mende in Sierra Leone **[MSL]**	128	128
Mexican Ancestry in Los Angeles CA USA **[MXL]**	71	71
Peruvian in Lima Peru **[PEL]**	122	122
Puerto Rican in Puerto Rico **[PUR]**	139	139
Punjabi in Lahore, Pakistan **[PJL]**	158	158
Sri Lankan Tamil in the UK **[STU]**	128	128
Toscani in Italia **[TSI]**	114	114
Yoruba in Ibadan, Nigeria **[YRI]**	120	120

* CEPH Collection [CEU] samples are available from the NIGMS Human Genetic Cell Repository at Coriell.



label	population sample	number of samples
ASW	African ancestry in Southwest USA	90
CEU	Utah residents with Northern and Western European ancestry from the CEPH collection	180
CHB	Han Chinese in Beijing, China	90
CHD	Chinese in Metropolitan Denver, Colorado	100
GIH	Gujarati Indians in Houston, Texas	100
JPT	Japanese in Tokyo, Japan	91
LWK	Luhya in Webuye, Kenya	100
MEX	Mexican ancestry in Los Angeles, California	90
MKK	Maasai in Kinyawa, Kenya	180
TSI	Toscans in Italy	100
YRI	Yoruba in Ibadan, Nigeria	180

Practical applications

When working with genotype data you should:

1. Correct the input strands
2. Sort BCF
3. **QC SNPs based on: MAF, GENO**
4. **QC Individuals: MIND**
5. Chromosome split
6. Remove duplicate variants

"

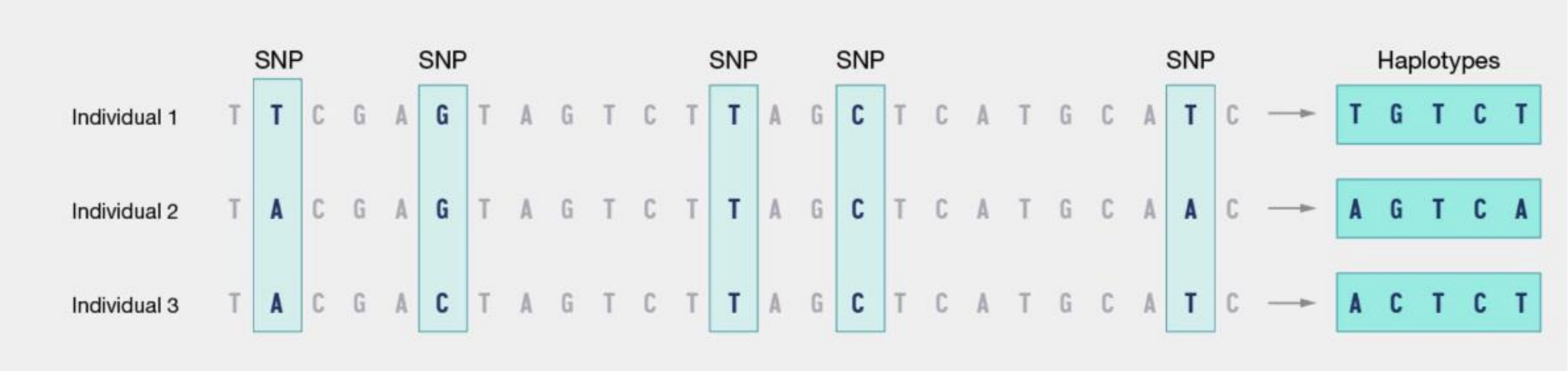
Exercise Break

Exercise 2: Data Preparation

Phasing

- Haplotype identification in the genotype data

- ShapeIt5


SNP
SNP
SNP
SNP
SNP
Haplotypes
Individual 1 T T C G A G T A G T C T T A G C T C A T G C A T C → T G T C T
Individual 2 T A C G A G T A G T C T T A G C T C A T G C A A C → A G T C A
Individual 3 T A C G A C T A G T C T T A G C T C A T G C A T C → A C T C T

Imputation software

- Beagle
- Minimac 4
- Impute 5

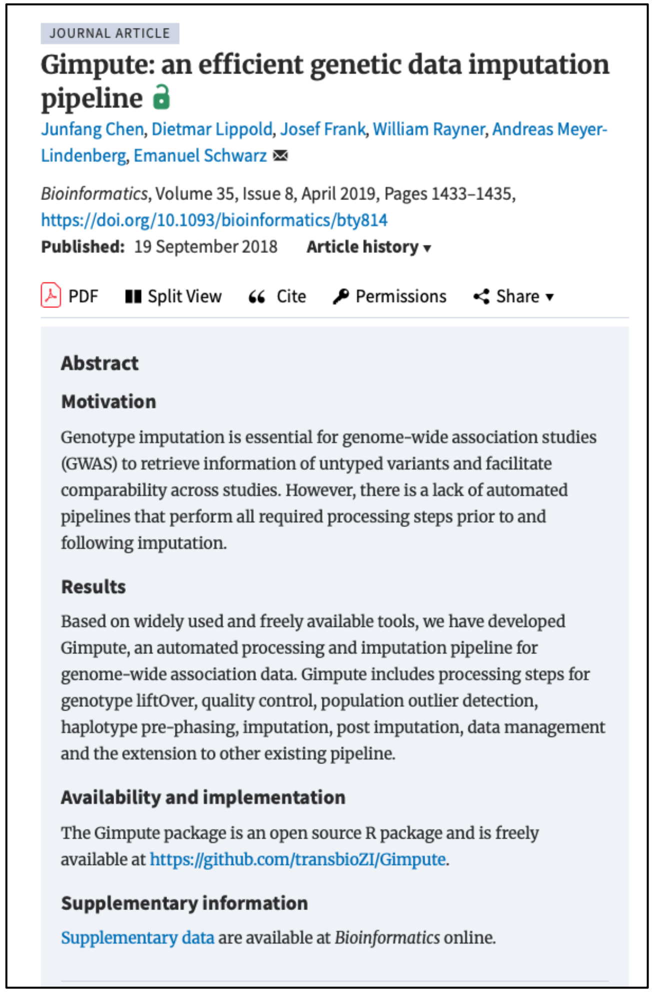
"

Exercise Break

Exercise 3: Data Preparation

Post-Imputation QC

- Info Score
  - Range [0, 1]
  - Higher is better



JOURNAL ARTICLE

Gimpute: an efficient genetic data imputation pipeline

Junfang Chen, Dietmar Lippold, Josef Frank, William Rayner, Andreas Meyer-Lindenberg, Emanuel Schwarz ✉

Bioinformatics, Volume 35, Issue 8, April 2019, Pages 1433–1435, https://doi.org/10.1093/bioinformatics/bty814

Published: 19 September 2018 **Article history** ▾

Abstract

Motivation

Genotype imputation is essential for genome-wide association studies (GWAS) to retrieve information of untyped variants and facilitate comparability across studies. However, there is a lack of automated pipelines that perform all required processing steps prior to and following imputation.

Results

Based on widely used and freely available tools, we have developed Gimpute, an automated processing and imputation pipeline for genome-wide association data. Gimpute includes processing steps for genotype liftOver, quality control, population outlier detection, haplotype pre-phasing, imputation, post imputation, data management and the extension to other existing pipeline.

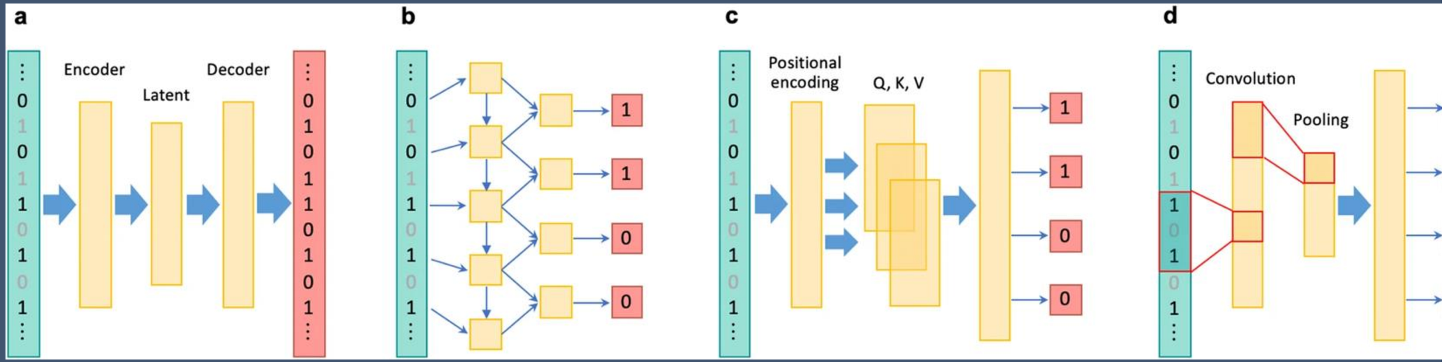
Availability and implementation

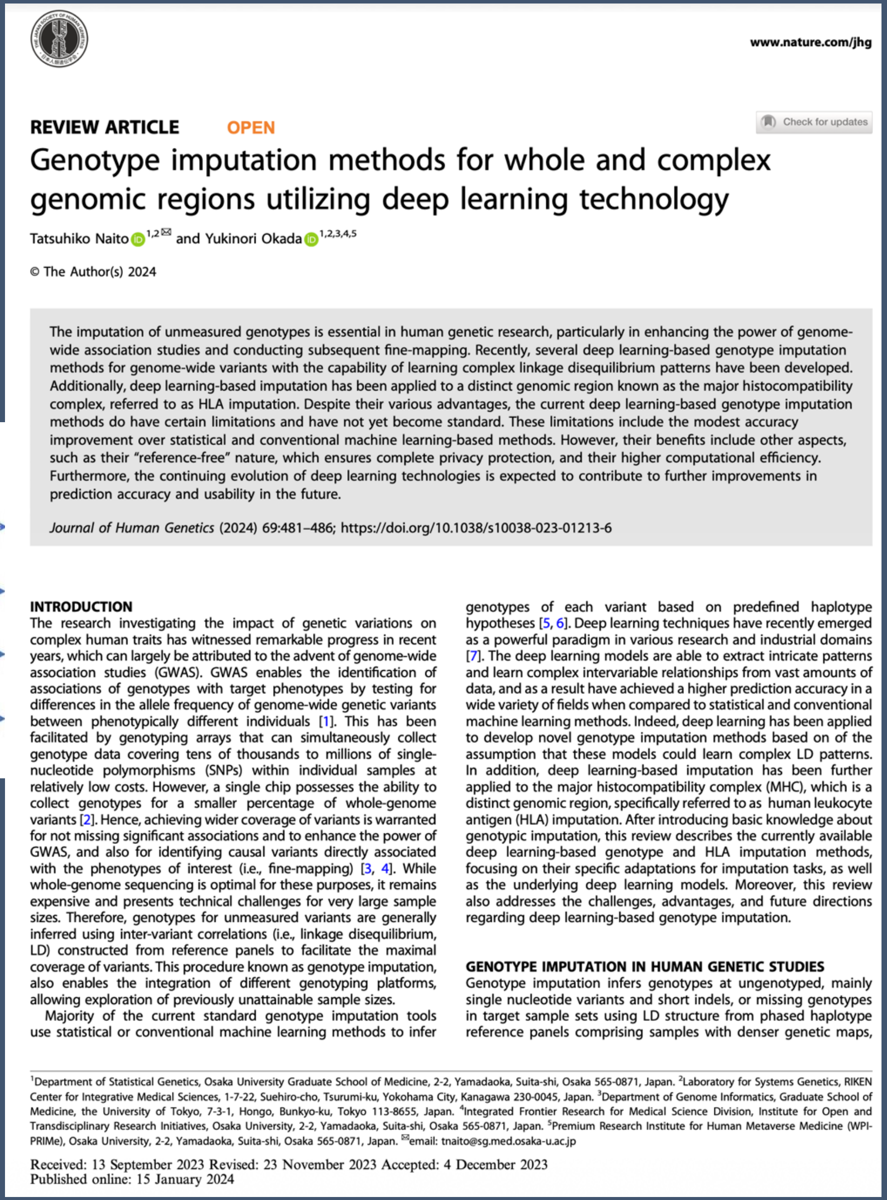
The Gimpute package is an open source R package and is freely available at https://github.com/transbioZI/Gimpute.

Supplementary information

Supplementary data are available at *Bioinformatics* online.

Deep learning imputation




www.nature.com/jhg

REVIEW ARTICLE OPEN

Check for updates

Genotype imputation methods for whole and complex genomic regions utilizing deep learning technology

Tatsuhiko Naito [1,2✉] and Yukinori Okada [1,2,3,4,5]

The imputation of unmeasured genotypes is essential in human genetic research, particularly in enhancing the power of genome-wide association studies and conducting subsequent fine-mapping. Recently, several deep learning-based genotype imputation methods for genome-wide variants with the capability of learning complex linkage disequilibrium patterns have been developed. Additionally, deep learning-based imputation has been applied to a distinct genomic region known as the major histocompatibility complex, referred to as HLA imputation. Despite their various advantages, the current deep learning-based genotype imputation methods do have certain limitations and have not yet become standard. These limitations include the modest accuracy improvement over statistical and conventional machine learning-based methods. However, their benefits include other aspects, such as their "reference-free" nature, which ensures complete privacy protection, and their higher computational efficiency. Furthermore, the continuing evolution of deep learning technologies is expected to contribute to further improvements in prediction accuracy and usability in the future.

Journal of Human Genetics (2024) 69:481–486; https://doi.org/10.1038/s10038-023-01213-6

INTRODUCTION

The research investigating the impact of genetic variations on complex human traits has witnessed remarkable progress in recent years, which can largely be attributed to the advent of genome-wide association studies (GWAS). GWAS enables the identification of associations of genotypes with target phenotypes by testing for differences in the allele frequency of genome-wide genetic variants between phenotypically different individuals [1]. This has been facilitated by genotyping arrays that can simultaneously collect genotype data covering tens of thousands to millions of single-nucleotide polymorphisms (SNPs) within individual samples at relatively low costs. However, a single chip possesses the ability to collect genotypes for a smaller percentage of whole-genome variants [2]. Hence, achieving wider coverage of variants is warranted for not missing significant associations and to enhance the power of GWAS, and also for identifying causal variants directly associated with the phenotypes of interest (i.e., fine-mapping) [3, 4]. While whole-genome sequencing is optimal for these purposes, it remains expensive and presents technical challenges for very large sample sizes. Therefore, genotypes for unmeasured variants are generally inferred using inter-variant correlations (i.e., linkage disequilibrium, LD) constructed from reference panels to facilitate the maximal coverage of variants. This procedure known as genotype imputation, also enables the integration of different genotyping platforms, allowing exploration of previously unattainable sample sizes.

Majority of the current standard genotype imputation tools use statistical or conventional machine learning methods to infer genotypes of each variant based on predefined haplotype hypotheses [5, 6]. Deep learning techniques have recently emerged as a powerful paradigm in various research and industrial domains [7]. The deep learning models are able to extract intricate patterns and learn complex intervariable relationships from vast amounts of data, and as a result have achieved a higher prediction accuracy in a wide variety of fields when compared to statistical and conventional machine learning methods. Indeed, deep learning has been applied to develop novel genotype imputation methods based on of the assumption that these models could learn complex LD patterns. In addition, deep learning-based imputation has been further applied to the major histocompatibility complex (MHC), which is a distinct genomic region, specifically referred to as human leukocyte antigen (HLA) imputation. After introducing basic knowledge about genotypic imputation, this review describes the currently available deep learning-based genotype and HLA imputation methods, focusing on their specific adaptations for imputation tasks, as well as the underlying deep learning models. Moreover, this review also addresses the challenges, advantages, and future directions regarding deep learning-based genotype imputation.

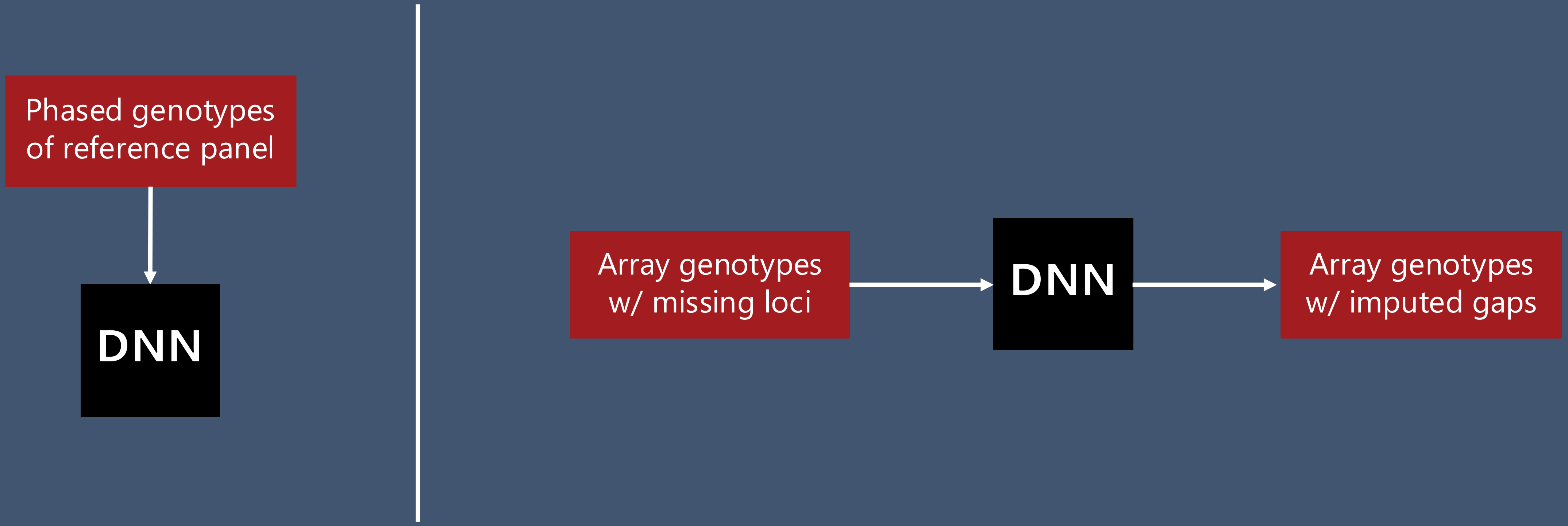
GENOTYPE IMPUTATION IN HUMAN GENETIC STUDIES

Genotype imputation infers genotypes at ungenotyped, mainly single nucleotide variants and short indels, or missing genotypes in target sample sets using LD structure from phased haplotype reference panels comprising samples with denser genetic maps,

[1] Department of Statistical Genetics, Osaka University Graduate School of Medicine, 2-2, Yamadaoka, Suita-shi, Osaka 565-0871, Japan. [2] Laboratory for Systems Genetics, RIKEN Center for Integrative Medical Sciences, 1-7-22, Suehiro-cho, Tsurumi-ku, Yokohama City, Kanagawa 230-0045, Japan. [3] Department of Genome Informatics, Graduate School of Medicine, the University of Tokyo, 7-3-1, Hongo, Bunkyo-ku, Tokyo 113-8655, Japan. [4] Integrated Frontier Research for Medical Science Division, Institute for Open and Transdisciplinary Research Initiatives, Osaka University, 2-2, Yamadaoka, Suita-shi, Osaka 565-0871, Japan. [5] Premium Research Institute for Human Metaverse Medicine (WPI-PRIMe), Osaka University, 2-2, Yamadaoka, Suita-shi, Osaka 565-0871, Japan. ✉email: tnaito@sg.med.osaka-u.ac.jp

Received: 13 September 2023 Revised: 23 November 2023 Accepted: 4 December 2023
Published online: 15 January 2024

Deep learning imputation



Phased genotypes of reference panel

DNN

Array genotypes w/ missing loci

DNN

Array genotypes w/ imputed gaps

Questions?