Documentation for Haplo2D6, version 1.1

Maria Carolina Silva de Barros Puça, Diego Mariano, Raquel Cardoso de Melo Minardi, Taís Nóbrega de Sousa[1]

April 2026

[1]Address for correspondence: René Rachou Institute, Oswaldo Cruz Foundation (FIOCRUZ), Belo Horizonte, email: haplo2d6@gmail.com, http://bioinfo.dcc.ufmg.br/Haplo2D6

How to Use Haplo2D6 to Translate CYP2D6 Genotype to Phenotype

Haplo2D6 is a powerful tool designed to predict the star alleles and enzyme activity of CYP2D6, making the analysis of genotype data more straightforward. The software is intended for researchers and clinicians who need to streamline the process of determining CYP2D6 metabolizer status based on genetic data.

1. Input Requirements

To use Haplo2D6, you need to provide:

1.1 Genotype Data: the input file supplied by the user specifies the number of individuals analyzed, the number and type of loci (biallelic or multi-allelic locus), and the genotypes for each individual.

Examples of input files containing genotype data of 12 loci on the *CYP2D6* gene for 20 individuals in a spreadsheet format*:

20																								
12																								
P 2617 4300 4300 5222 5908 6047 6750 6816 7051 7189 7384 8381																								
SSSSSSSSSSS																								
1	C	C	C	C	C	T	C	C	A	A	G	G	T	T	C	C	C	C	G	G	G	G	G	C
2	C	G	C	C	C	C	C	C	A	A	G	G	T	T	C	C	C	T	G	G	G	G	G	C
3	C	C	C	C	C	T	C	C	A	A	G	A	T	T	C	C	C	C	G	G	G	G	G	C
4	C	G	C	C	C	C	C	T	A	A	G	G	T	T	C	C	T	T	G	G	G	G	C	C
5	C	C	C	C	C	C	T	T	A	A	G	G	T	T	C	C	T	T	G	G	G	G	C	C
6	C	C	C	C	C	T	C	C	A	A	G	A	T	T	C	C	C	T	G	A	G	G	C	C
7	C	G	C	C	C	C	C	C	A	C	G	G	T	T	C	C	C	T	G	G	G	G	G	C
8	C	C	C	C	C	C	C	C	A	A	G	G	T	T	C	C	C	C	G	G	G	G	G	G
9	C	G	C	T	C	C	C	T	A	A	G	G	T	T	C	C	T	T	G	G	G	G	C	C
10	G	G	C	C	C	C	C	C	A	A	G	G	T	T	C	C	T	T	G	G	G	G	C	C

20												
12												
P 2617 4300 4300 5222 5908 6047 6750 6816 7051 7189 7384 8381												
SSSSSSSSSSS												
1	CC	CC	CT	CC	AA	GG	TT	CC	CC	GG	GG	GC
2	CG	CC	CC	CC	AA	GG	TT	CC	CT	GG	GG	GC
3	CC	CC	CT	CC	AA	GA	TT	CC	CC	GG	GG	GC
4	CG	CC	CC	CT	AA	GG	TT	CC	TT	GG	GG	CC
5	CC	CC	CC	TT	AA	GG	TT	CC	TT	GG	GG	CC
6	CC	CC	CT	CC	AA	GA	TT	CC	CT	GA	GG	CC
7	CG	CC	CC	CC	AC	GG	TT	CC	CT	GG	GG	GC
8	CC	CC	CC	CC	AA	GG	TT	CC	CC	GG	GG	GG
9	CG	CT	CC	CT	AA	GG	TT	CC	TT	GG	GG	CC
10	GG	CC	CC	CC	AA	GG	TT	CC	TT	GG	GG	CC

Line 1 – Number of individuals analyzed

Line 2 – Number of loci/sites

Line 3 – P (upper case) indicates the position of locus i, relative to some arbitrary reference point; the loci must be in their physical order along the chromosome (ie, the positions must be increasing). Note. Position at NG_008376.3 (CYP2D6 RefSeqGene; reverse relative to chromosome).

Line 4 – Locus type: S for biallelic locus and M for multi-allelic locus.

* Haplo2D6 also supports the Phase default format (file.inp). For more information, see Phase documentation (https://stephenslab.uchicago.edu/phase/download.html).

Handling Insertions and Deletions (indels)

PHASE accepts genotype input in the form of discrete allele values (e.g., nucleotides such as A, C, T, G). Therefore, insertions and deletions (indels) cannot be directly represented as structural events in the input file. To include indels in the analysis, users must encode them as biallelic markers using a consistent symbolic representation. This involves assigning arbitrary nucleotide codes to represent the presence and absence of the polymorphism.

For example, a deletion or insertion at a given position can be encoded as:

- **C** = reference allele (absence of indel)
- **G** = variant allele (presence of indel)

Thus, genotypes such as:

- CC → homozygous reference
- CG → heterozygous
- GG → homozygous variant

Alternatively, any consistent pair of symbols (e.g., A/T, C/G) may be used, provided that the encoding is applied consistently across all samples and matches the allele reference file.

Important: The same encoding scheme must be used in both the genotype file and the allele reference file to ensure correct haplotype reconstruction and allele assignment.

1.2 Allele References: A reference file that maps genetic variants to star alleles and their respective functions.

Example of input file (comma separated) contains the variants that comprise each star allele, allele function assignments (normal function, no function, or decreased), and the allele activity value. Source material can be found on PharmGKB (https://www.pharmgkb.org/page/cyp2d6RefMaterials) and PharmVar websites (https://www.pharmvar.org/gene/CYP2D6). The positions follow the same order as in the previous table.

CCCCAGTCCGGG,Normal,*1,1
GCCCAGTCTGGC,Normal,*2,1
CCCCAGACCGGG,No function,*3,0
CCTCAATCCGGC,No function,*4,0
CCCCCGTCCGGG,No function,*6,0
CCCCAGTACGGG,Decreased,*9,0.25
CCTCAGTCCGGC,Decreased,*10,0.25
CCCTAGTCTGGC,Decreased,*17,0.5
GTCCAGTCTGGC,Normal,*35,1
CCCCAGTCTAGC,Decreased,*41,0.25

Recommended Reference Panel

To facilitate use and reduce the need for manual curation, Haplo2D6 provides a pre-populated reference file based on the **AMP Tier 1 CYP2D6 allele recommendations**, which represent the minimum set of alleles recommended for clinical and research genotyping. For more details, see Pratt et al. 2021 (doi:10.1016/j.jmoldx.2021.05.013).

The Tier 1 alleles included are: *CYP2D6 *2, *3, *4, *5, *6, *9, *10, *17, *29,* and **41*. These alleles are considered essential for CYP2D6 analysis due to their frequency and functional relevance across populations. The reference file has been validated to ensure accurate haplotype assignment and phenotype prediction and can be used as a template for adding or removing variants as needed.

AMP Tier 1 CYP2D6 allele Reference*:

GTCCACTCGCCC,Normal,*1,1
GTCGACTCACCG,Normal,*2,1
GTCCACACGCCC,No Function,*3,0
ATCGATTCGCCG,No Function,*4,0
GTCCCCTCGCCC,No Function,*6,0
GTCCACTAGCCC,Decreased,*9,0.25
ATCGACTCGCCG,Decreased,*10,0.25
GACGACTCACCG,Decreased,*17,0.5
GTTGACTCACTG,Decreased,*29,0.5
GTCGACTCATCG,Decreased,*41,0.5

*Reference variant positions: rs1065852, rs28371706, rs61736512, rs1058164, rs5030655, rs3892097, rs35742686, rs5030656, rs16947, rs28371725, rs59421388, rs1135840

2. Analyzing Your Data

The process of analyzing your data with Haplo2D6 involves several steps:

2.1 Data Preparation: Before running Haplo2D6, ensure that your genotype data is formatted correctly and that you have the necessary allele reference file.

2.2 Haplotype Reconstruction: Haplo2D6 integrates PHASE (version 2.1.1), a Bayesian statistical method for haplotype reconstruction (Stephens, Smith, and Donnelly 2001, Stephens and Donnelly 2003), to reconstruct haplotypes from population data. This step is crucial for accurately determining which alleles are present in each individual. Phase is an open-source software available under the following license: Phase License.

2.3 Star Allele and Enzyme Activity Prediction: Once the haplotypes are reconstructed, Haplo2D6 predicts the star alleles and corresponding enzyme activities. The output includes the predicted star alleles, allele activity value, and an associated phenotype (e.g., normal metabolizer).

Example of the output file

Show 10 entries Search:

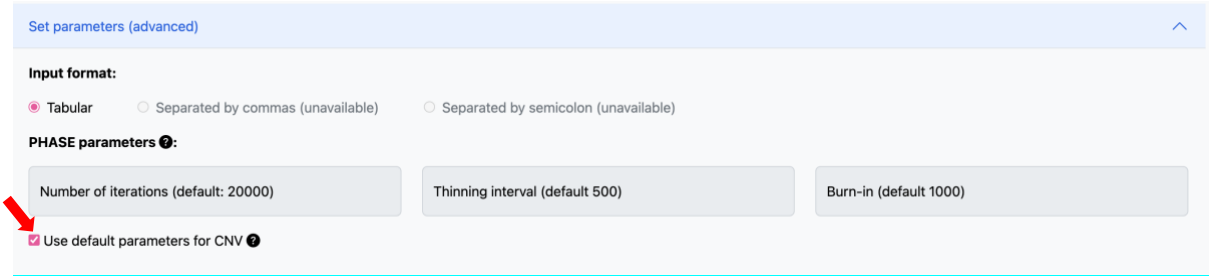
#	ID	Haplotype #1	Allele Functional #1	Allele #1	Activity Value #1	Haplotype #2	Allele Functional #2	Allele #2	Activity Value #2	Activity Score	Phenotype	CNV	Diplotype
0	1	CCCCAGTCCGGG	Normal	*1	1.0	CCTCAGTCCGGC	Decreased	*10	0.25	1.25	gNM	2	*1/*10
1	10	GCCCAGTCTGGC	Normal	*2	1.0	-	No function	*5	0	1.0	gIM	1	*2/*5
2	11	GTCCAGTCTGGC	Normal	*35	1.0	GTCCAGTCTGGC	Normal	*35xN	1.0	3	gUM	3	*35/*35xN
3	12	CCCCAGTCCGGG	Normal	*1	1.0	CCCCCGTCCGGG	No function	*6	0.0	1.0	gIM	2	*1/*6
4	13	CCCCAGTCTAGC	Decreased	*41	0.25	CCCCAGTCTAGC	Decreased	*41	0.25	0.5	gIM	2	*41/*41
5	14	CCCCAGTCCGGG	Normal	*1	1.0	CCCCAGACCGGG	No function	*3	0.0	1.0	gIM	2	*1/*3
6	15	GCCCAGTCTGGC	Normal	*2	1.0	GCCCAGTCTGGC	Normal	*2	1.0	2.0	gNM	2	*2/*2
7	16	CCTCAGTCCGGC	Decreased	*10	0.25	GCCCAGTCTGGC	Normal	*2	1.0	1.25	gNM	2	*10/*2
8	17	CCCCAGTCCGGG	Normal	*1	1.0	GTCCAGTCTGGC	Normal	*35	1.0	2.0	gNM	2	*1/*35
9	18	GCCCAGTCTGGC	Normal	*2	1.0	-	No function	*5	0	1.0	gIM	1	*2/*5

Showing 1 to 10 of 20 entries Previous 1 2 Next

2.4 Copy Number Variation (CNV) Consideration: Haplo2D6 supports the integration of user-provided copy number variation (CNV) information during the analysis. CNV data can be specified through the **"Set parameters (advanced)"** option in the interface.

By default, Haplo2D6 assumes two gene copies per individual. If CNV information is available, users can provide it in a simple tabular format using either comma- or tab-separated values (e.g., ID, CNV).

CNV information can be provided through the "Set parameters (advanced)" option in the interface.



The tool automatically incorporates CNV information into haplotype and phenotype prediction as follows:

- For a copy number of one (CNV = 1), if both inferred alleles are identical, Haplo2D6 assigns the **5* allele to represent gene deletion (e.g., **1/*1* is adjusted to **1/*5*). The activity score is then recalculated accordingly, which may result in a change in the predicted metabolizer phenotype

- For copy numbers greater than two (gene duplication or multiplication), Haplo2D6 incorporates the total number of copies into the activity score calculation. When identical alleles are present, the duplicated allele can be inferred. When different alleles are present, it is not possible to determine which allele is duplicated, and the result is reported as **indeterminate**.

This automated approach ensures consistent and reproducible CNV-aware phenotype prediction without requiring manual adjustment.

3. Additional Resources and Recommendations

- **Minimum Sets of Alleles for Pharmacogenomic Testing**: For guidance on which alleles to test, refer to this resource: PharmGKB Alleles to Test.
- **Gene-specific Information Tables**: These tables provide definitions for each star allele, their function (e.g., normal, no function, decreased function), and how these alleles combine to form metabolizer phenotypes. Access this resource here: PharmGKB CYP2D6 Reference Materials.
- **Copy Number Variation and Phenotype Translation Examples**: You can find examples and additional guidance on how to adjust phenotypes based on genotype data and CNVs in this template: CYP2D6 Genotyping Method and Data Template.

References

Pratt VM, Cavallari LH, Tredici AL Del *et al.* Recommendations for Clinical CYP2D6 Genotyping Allele Selection. *The Journal of Molecular Diagnostics* 2021;23(9):1047–64. https://doi.org/10.1016/j.jmoldx.2021.05.013.

Stephens M, Donnelly P. A Comparison of Bayesian Methods for Haplotype Reconstruction from Population Genotype Data. *The American Journal of Human Genetics* 2003;73(5):1162–9. https://doi.org/10.1086/379378.

Stephens M, Smith NJ, Donnelly P. A New Statistical Method for Haplotype Reconstruction from Population Data. *The American Journal of Human Genetics* 2001;68(4):978–89. https://doi.org/10.1086/319501.