Genetic Report Card – *P.vivax* User Guide

Updated May 2024

Contents

1 Introduction

The Genetic Report Card (GRC) contains malaria parasite genetic data, derived from the analysis of patient blood samples. For each sample, we have determined genotypes at 42 single nucleotide polymorphisms (SNPs) that have been chosen to analyse the diversity between parasites. Also, 10 amino acid changes putatively associated with resistance to antimalarial drugs and details of associated haplotypes are reported. The Report additionally contains data on *Plasmodium* species found in the samples and on the complexity of sample infection.

The following is a guide to interpreting the Genetic Report Card data, linking the genotyped mutations (molecular markers) to resistance to different antimalarial drugs.

The Genetic Report Card is detailed over multiple tabs:

- **GRC** - gives sample metadata details, species call, complexity of infection, sample SNP barcode and drug resistant gene haplotypes.
- **Barcode** - details the genotype result for each SNP, which is shown concatenated in GRC tab.
- **ENA** – details the sample identification numbers, both internal and external, and the accession number of each sample as seen on European Nucleotide Archive (ENA).

2 Sample Characteristics

2.1 Species Co-infections

We detect the presence of different *Plasmodium* species by sequencing two conserved regions of the parasite mitochondria. SNPs within these regions are able to differentiate the five human-infecting species of *Plasmodium* (*P. falciparum, P. vivax, P. malariae, P. ovale,* and *P. knowlesi*). We report a list of all species detected e.g. Pf/Pv. The order of the list has no biological significance.

In some cases it is possible that due to missing sequence data certain species are indistinguishable, these will be indicated on the GRC (-).

2.2 Complexity of Infection (COI)

We use the sample barcodes to infer individual sample estimates of Complexity of Infection (COI). Previously we reported COIL[1] and The real McCOIL[2] but from 2023 we only report the latter, and the COIL column does not contain data. The COI is expressed as the estimated number of genetically distinct parasites within the infection from which the sample was taken.

This is an approximation that greatly simplifies the complexities of mixed infections. However, the prevalence of highly mixed infections is a useful indicator of transmission intensity, which can be used when comparing sites or seasons. Additionally, highly polyclonal infections are sometimes favourable to exclude from certain population analyses.

2.3 Sample Barcodes

The sample barcodes are formed by concatenated genotypes at 42 SNPs across the *P. vivax* nuclear genome. The barcode is derived from a previously published list of loci from the Broad Institute [3]. These SNPs are all biallelic, i.e. only two alleles are observed; they were chosen for their usefulness in the analyses of relationships between parasites, due to their high heterozygosity, and are not associated with drug resistance.

The full list of SNPs used is given in **Appendix Table 1**. SNPs within the barcode are represented by the observed homozygous nucleotide (A, T, C and G). If the genotype is missing (could not be detected), the symbol “X” is used. The symbol “N” indicates that both alleles were observed (heterozygous call). These loci were seen to be biallelic in a global set of whole-genome sequenced parasites, however it is possible that sporadic alternate alleles do exist at low levels in populations.

Barcodes can be used for a variety of purposes, such as:

- The proportion of positions with identical alleles gives a measure of the genetic distance between two samples and can be used for generating population clusters.
- The proportion of heterozygous calls can be used to estimate the complexity of infection (COI).
- Identifying several samples with (almost) identical barcodes indicates a fast-expanding strain.
- The proportion of missingness in the barcode can give a quick indication of sample quality.

3 Drug Resistance Genes and Haplotypes

Three columns in the Genetic Report Card show the genotypes of putative drug resistance mutations. For consistency with published literature, these genotypes are reported as their amino acid translations and positions eg. A553G. For genes with multiple SNPs of interest we translate each SNP to the amino acid and concatenate these to report an amino acid haplotype eg. FSTS, which is also common in the literature. Genotypes are called using a read based approach where an allele must have 5 high quality supporting reads at a minimum proportion of 10% of all high quality reads.

For example, in the *pvdhfr-ts* gene, we genotype amino acid positions 57, 58, 61 and 117, which are implicated in resistance to antifolates (see section 3.1.1 “Pyrimethamine and Sulfadoxine Drug Resistance Mutations”). The wild-type (WT) amino acids at these positions are F, S, T and S respectively - we represent this haplotype as “FSTS”. If a parasite is reported to carry the haplotype “FRTN”, it means it is a double mutant at positions 58 (S →R), and 117 (S →N).

The following table shows the antimalarial drug and gene associated in resistance to that drug, also the amino acid positions associated with the haplotype and the wild type, non-mutated, form. Note that the association of *pvmdr1*$_{976Y}$ with chloroquine and amodiaquine resistance is somewhat controversial with limited and contradictory evidence.

Column	Antimalarial	Gene	Amino Acid Positions	Wild Type
PvDHFR	Pyrimethamine	pvdhfr	57, 58, ,61, 117	F, S, T, S
PvDHPS	Sulfadoxine	pvdhps	380, 382, 383, 385, 553	E, S, A, Y, A
PvMDR-1	Amodiaquine, Chloroquine	pvmdr1	976	Y

In the haplotypes, we also use the following special characters:

- A dash ("-") indicates a missing genotype, either because the sample could not be genotyped at that position, or because the assay is still under development/refinement.
- If two alleles were detected (a heterozygous call), then both appear in brackets (e.g. [F/L]).

Details of how genetic mutations relate to resistance to specific drugs can be found in the following sections.

3.1 Resistance to Specific Antimalarials

3.1.1 Pyrimethamine and Sulfadoxine Drug Resistance Mutations

Pyrimethamine drug resistance is mediated by mutations in the bifunctional dihydrofolate reductase-thymidylate synthase (*pvdhfr*, PVP01_0526600) [4,5]. Four non-synonymous mutations at amino acid positions 57, 58, 61 and 117 have been established as being potentially important in drug resistance [4,5]. Work is ongoing to fully determine the significance of the DHFR:58 amino acid.

Sulfadoxine drug resistance is mediated by mutations in the dihydropteroate synthetase (*pvdhps*, PVP01_1429500). Two mutations at positions 383 and 553 are thought to confer some resistance [6]. We also type surrounding loci at positions 380, 382, and 385.

Pvdhps mutations are often seen in combination with *pvdhfr* mutations, since sulfadoxine and pyrimethamine are mostly used in combination (SP).

3.1.2 Amodiaquine and Chloroquine Drug Resistance Mutations

A mutation in the multidrug resistance protein (*pvmdr1*, PVP01_1010900) has been associated with parasite response to amodiaquine and chloroquine. There is limited evidence that a mutation at position 976 can mediate response to these drugs [5,7].

4 Release Change Log

- The GRC is a data output of the Amplicon toolkit bioinformatics pipeline at the Wellcome Sanger Institute, and assays are undergoing continued development and optimization.
- References provided are not the only evidence for association of molecular markers with clinical and *in vitro* phenotypes. Further research should be done before drawing conclusions on results and reporting.
- From 2023, Complexity of Infection data is reported as The real McCOIL only, the COIL column no longer contains data.

5 References

1. Galinsky K, Valim C, Salmier A, de Thoisy B, Musset L, Legrand E, et al. COIL: a methodology for evaluating malarial complexity of infection using likelihood from single nucleotide polymorphism data. Malaria Journal. 2015; 14:4.
2. Chang HH, Worby CJ, Yeka A, Nankabirwa J, Kamya MR, Staedke SG, Dorsey G, Murphy M, Neafsey DE, Jeffreys AE, Hubbart C, Rockett KA, Amato R, Kwiatkowski DP, Buckee CO, Greenhouse B. THE REAL McCOIL: A method for the concurrent estimation of the complexity of infection and SNP allele frequency for malaria parasites. PLoS Computational Biology. 2017; 13:1.
3. Baniecki ML, Faust AL, Schaffner SF, Park DJ, Galinsky K, Daniels RF, et al. Development of a single nucleotide polymorphism barcode to genotype *Plasmodium vivax* infections. PLoS Neglected Tropical Diseases. 2015; 9(3).
4. Hawkins VN, Joshi H, Rungsihirunrat K, Na-Bangchang K, & Sibley CH. Antifolates can have a role in the treatment of *Plasmodium vivax*. Trends in Parasitology. 2007; 23:213-222.
5. Marfurt J, de Monbrison F, Brega S, Barbollat L, Muller I, Sie A, et al. Molecular markers of in vivo *Plasmodium vivax* resistance to amodiaquine plus sulfadoxine-pyrimethamine: mutations in pvdhfr and pvmdr1. The Journal of Infectious Diseases. 2008; 198:409-417.

6. Korsinczky M, Fischer K, Chen N, Baker J, Rieckmann K, and Cheng Q. Sulfadoxine resistance in *Plasmodium vivax* is associated with a specific amino acid in dihydropteroate synthase at the putative sulfadoxine-binding site. Antimicrobial Agents and Chemotherapy. 2004; 48:2214-2222.
7. Suwanarusk R, Russell B, Chavchich M, Chalfein F, Kenangalem E, Kosaisavee V, et al. Chloroquine resistant *Plasmodium vivax*: in vitro characterization and association with molecular polymorphisms. PLoS One. 2007; 2(10).

6 Appendix

Table 1 – List of the 42 barcode SNPs.

The SNPs are presented in the order in which they are concatenated in the barcode. For each SNP, we show: the chromosome and position within the chromosome; the reference (PvP01 strain) and non-reference (alternative) alleles; the ID and description of the gene containing the SNP; whether the SNP is synonymous (S) or non-synonymous (N/S), intronic (int) or intergenic (-); the amino acid mutations caused by the SNP in that gene; and the coding strand for this gene (+ = sense, - = antisense).

Num	Chr	Pos	Ref	Nonref	GeneId	GeneDescription	MutType	MutName	Strand
1	PvP01_01_v1	778121	T	G	PvP01_0117100	U3 small nucleolar RNA-associated protein 14, putative	S	902S	(+)
2	PvP01_10_v1	1371780	C	T	-	intergenic	-	-	-
3	PvP01_10_v1	151400	T	G	-	intergenic	-	-	-
4	PvP01_10_v1	865053	T	G	-	intergenic	-	-	-
5	PvP01_11_v1	643929	C	T	PvP01_1115100	serine/threonine protein kinase, putative	S	2588V	(-)
6	PvP01_12_v1	1220748	C	T	PvP01_1230500	conserved Plasmodium protein, unknown function	S	435R	(-)
7	PvP01_12_v1	2685195	G	A	PvP01_1264600	conserved Plasmodium protein, unknown function	S	2406H	(-)
8	PvP01_13_v1	1088863	T	C	PvP01_1325200	vacuolar protein sorting-associated protein 3, putative	S	1124R	(-)
9	PvP01_13_v1	507029	C	T	-	intergenic	-	-	-
10	PvP01_13_v1	851321	A	G	-	intergenic	-	-	-
11	PvP01_14_v1	1036358	C	T	-	intergenic	-	-	-
12	PvP01_14_v1	2310569	C	A	-	intergenic	-	-	-
13	PvP01_14_v1	439002	G	A	PvP01_1409400	conserved Plasmodium protein, unknown function	S	1365N	(-)
14	PvP01_14_v1	855554	T	C	-	intergenic	-	-	-
15	PvP01_03_v1	574236	G	A	PvP01_0313300	calcium-dependent protein kinase 4, putative	S	317R	(+)
16	PvP01_04_v1	367674	C	T	PvP01_0409000	DEAD/DEAH box helicase, putative	N	Q73P	(-)
17	PvP01_06_v1	602491	C	T	PvP01_0614200	conserved Plasmodium protein, unknown function	S	118N	(+)
18	PvP01_08_v1	1565446	C	T	PvP01_0837100	CLP1 P-loop domain-containing protein, putative	S	420Y	(+)

19	PvP01_09_v1	846288	T	C	-	intergenic	-	-	-
20	PvP01_01_v1	612295	T	C	PvP01_0113700	protein tyrosine phosphatase, putative	S	210E	(-)
21	PvP01_10_v1	1304923	T	C	PvP01_1030300	myosin B, putative	S	19N	(+)
22	PvP01_11_v1	1969060	A	G	-	intergenic	-	-	-
23	PvP01_11_v1	256884	T	C	PvP01_1106300	SNARE associated Golgi protein, putative	int	-	(+)
24	PvP01_11_v1	108200	T	C	-	intergenic	-	-	-
25	PvP01_12_v1	1139454	T	C	-	intergenic	-	-	-
26	PvP01_13_v1	1160640	C	A	-	intergenic	-	-	-
27	PvP01_14_v1	2258311	A	G	PvP01_1451700	asparagine and aspartate rich protein 1, putative	S	7095V	(+)
28	PvP01_03_v1	239662	T	C	-	intergenic	-	-	-
29	PvP01_05_v1	1409127	C	T	-	intergenic	-	-	-
30	PvP01_05_v1	163535	A	G	PvP01_0503900	conserved Plasmodium protein, unknown function	S	99T	(-)
31	PvP01_06_v1	76255	A	G	-	intergenic	-	-	-
32	PvP01_08_v1	1419576	C	T	PvP01_0833100	40s ribosomal protein S12, putative	int	-	(+)
33	PvP01_08_v1	173904	C	T	-	intergenic	-	-	-
34	PvP01_08_v1	577308	G	A	PvP01_0813500	conserved Plasmodium protein, unknown function	S	385D	(-)
35	PvP01_02_v1	527792	T	C	PvP01_0212300	conserved Plasmodium protein, unknown function	S	395N	(+)
36	PvP01_03_v1	595569	A	G	PvP01_0313800	dynein heavy chain, putative	int	-	-
37	PvP01_11_v1	1809904	G	A	PvP01_1142500	conserved Plasmodium protein, unknown function	int	-	(+)
38	PvP01_11_v1	747140	A	G	PvP01_1117500	large subunit rRNA methyltransferase, putative	int	-	(+)
39	PvP01_07_v1	625740	C	T	-	intergenic	-	-	-
40	PvP01_07_v1	1220072	C	T	PvP01_0728900	merozoite surface protein 1	S	1547V	(+)
41	PvP01_11_v1	666727	A	C	-	intergenic	-	-	-
42	PvP01_11_v1	1835972	C	T	PvP01_1143100	DNA repair protein RAD50, putative	S	918C	(+)