# Supplementary method and results for *P. vivax* data

## Preprocessing

### Method: Preprocessing VCF for minSNPs

VCF format describing the filtered high quality bi-allelic SNPs for all the 259 *P. vivax* isolates were obtained from: <https://www.malariagen.net/resource/24>. We then made use of vcf2phylip (<https://github.com/edgardomortiz/vcf2phylip>) (1) to convert the file into FASTA format, with the following command:

python vcf2phylip.py -i Pv\_Nature\_communications\_2018\_VQSLOD3.vcf.gz -f

And the following is used to obtain the list of reference positions for the SNPs:

bcftools query -f '%CHROM\t%POS\n' Pv\_Nature\_communications\_2018\_VQSLOD3.vcf.gz | cat -n > ref\_pos\_vivax.tsv && echo -e "snps\tchrom\treference\_pos" | cat - ref\_pos\_vivax.tsv > /tmp/out && mv /tmp/out ref\_pos\_vivax.tsv

For list of SNPs and translation to reference genome positions, see *sup\_4\_1\_ref\_pos\_vivax.tsv.*

### Method: Dealing with ambiguity codes

The initial transformed alignment contained 34.8% (183,509) non-standard bases, this includes both ambiguity codes (heterozygous calling) and failed sequencing, which would normally be ignored by MinSNPs analysis. To make use of these SNPs, we developed a protocol where the non-standard bases were substituted with the major allele (the most common allele for a SNP among the samples) at that position. Positions with >=75% samples exhibiting ambiguity codes (potential misalignments or sequencing errors) were excluded (2 such positions were found and excluded).

seed <- 2021 *#Setting a seed to ensure reproducibility*

set.seed(seed)

library("minSNPs")

vivax <- read\_fasta("./Pv\_Nature\_communications\_2018\_VQSLOD3.min0.fasta")

resolved\_vivax <- resolve\_IUPAC\_missing(vivax, log\_file = "replace2.tsv", max\_ambiguity = 0.75, replace\_method = "most\_common")

write\_fasta(resolved\_vivax, filename = "n2\_vivax.fasta")

In subsequent sections, “vivax” refers to the original alignment, where all the positions containing at least an ambiguity code are excluded from the analysis, and “N2\_vivax” refers to this new matrix, with ambiguity codes substituted.

## Derivation of SNP sets to discriminate "K2" strain with "percent" mode

We regarded 26 specimens (“K2” strain) as a model surveillance target and used percent mode in MinSNPs to identify SNP sets that distinguish all K2 specimens from all other specimens.

k2 <- c("PY0019-C", "PY0024-C", "PY0026-C", "PY0027-C", "PY0034-C",

"PY0044-C", "PY0053-C", "PY0056-C", "PY0057-C", "PY0058-C",

"PY0060-C", "PY0061-C", "PY0067-C", "PY0068-C", "PY0072-C",

"PY0073-C", "PY0074-C", "PY0075-C", "PY0101-C", "PY0105-C",

"PY0117-C", "PY0035-C", "PY0042-C", "PY0076-C", "PY0119-C", "PY0085-C")

vivax <- read\_fasta("./Pv\_Nature\_communications\_2018\_VQSLOD3.min0.fasta")

p\_vivax <- process\_allele(vivax)

k2\_snps <- find\_optimised\_snps(p\_vivax, metric = "percent", goi = k2, max\_depth = 5, number\_of\_result = 5)

output\_result(k2\_snps, view = "csv", file\_name = "k2\_snps\_vivax.tsv")

#n2\_vivax <- read\_fasta("./n2\_vivax.fasta")

#p\_n2\_vivax <- process\_allele(n2\_vivax)

#n2\_k2\_snps <- find\_optimised\_snps(p\_n2\_vivax, metric = "percent", goi = k2, max\_depth = 5, #number\_of\_result = 5)

#output\_result(n2\_k2\_snps, view = "csv", file\_name = "k2\_snps\_n2.tsv")

### Results

|  |  |
| --- | --- |
| Result | Vivax |
| see: *sup\_4\_2\_k2\_snps\_vivax.txt* |
| 1 | 15847 |
| 2 | 16089 |
| 3 | 16196 |
| 4 | 16524 |
| 5 | 20645 |

We then identified all the single-member SNP sets that could completely distinguish K2 isolates from the rest and found 124 such SNPs.

excluded\_positions <- p\_vivax$ignored\_position

all\_positions <- seq\_len(length(p\_vivax$seqc[[1]]))

all\_positions <- all\_positions[! all\_positions %in% excluded\_positions]

l1\_result <- bplapply(all\_positions, cal\_met\_snp, metric = "calculate\_percent",

seqc = p\_vivax$seqc, list(goi = k2 ), BPPARAM = bp)

k2\_l1\_snps <- unlist(l1\_result)

names(k2\_l1\_snps) <- all\_positions

k2\_l1\_snps\_100 <- names(k2\_l1\_snps[k2\_l1\_snps == 1])

cat(paste(k2\_l1\_snps\_100, collapse = "\n"), file = "1\_1\_k2.tsv")

All 124 SNPs obtained from “vivax” can be found in: *sup\_4\_3\_single\_snps\_k2.txt*

The experiment was not repeated using “N2\_vivax” given the abundance of the SNPs.

## Derivation of SNP sets to discriminate Malaysian strain with "percent" mode

We then attempted to identify SNP sets that discriminate all Malaysian isolates (a superset that includes K2 specimens). Three specimens that were obtained from Malaysia were removed from the group of interest due to significant genetic divergence (on the basis of ancestry with 80% to 100% similarity to isolates found in Thailand and Indonesia) from other specimens collected in Malaysia and could be considered imported cases. To streamline the analysis, we only included one K2 specimen in the group of interest.

Malaysian\_isolates <- names(vivax)[

startsWith(names(vivax), "PY") &

(! names(vivax) %in% k2[-which(k2 == "PY0044-C")])]

removed\_Malaysian\_isolates <- names(vivax)[startsWith(names(vivax), "PY") &

(! names(vivax) %in% Malaysian\_isolates)]

vivax[removed\_Malaysian\_isolates] <- NULL

GOI <- names(vivax)[startsWith(names(vivax), "PY") &

! names(vivax) %in% c("PY0045-C", "PY0004-C", "PY0120-C")]

p\_vivax <- process\_allele(vivax)

my\_snps <- find\_optimised\_snps(seqc = p\_vivax, max\_depth = 5, accept\_multiallelic= FALSE, number\_of\_result = 5, bp = bp, metric = "percent", goi = GOI)

output\_result(my\_snps, view = "csv", file\_name = "1K2\_MY\_percent\_vivax.tsv")

## N2\_Vivax

n2\_vivax[removed\_Malaysian\_isolates] <- NULL

p\_n2\_vivax <- process\_allele(n2\_vivax)

my\_snps <- find\_optimised\_snps(seqc = p\_n2\_vivax, max\_depth = 5, accept\_multiallelic= FALSE, number\_of\_result = 5, bp = bp, metric = "percent", goi = GOI)

output\_result(my\_snps, view = "csv", file\_name = "1K2\_MY\_percent\_n2\_vivax.tsv")

### Results

|  |  |
| --- | --- |
| Results | Description |
| See: *sup\_4\_4\_1K2\_MY\_percent\_vivax.txt* | Initial analysis where positions in the alignment with at least one sample containing an ambiguity code were excluded from the MinSNPs analysis. We considered this analysis a failure, as the majority of the Malaysian specimens were not discriminated from the other specimens even when there are 5 SNPs in a set. |
| See: *sup\_4\_5\_1K2\_MY\_percent\_n2\_vivax.txt* | Subsequent analysis making use of the matrix “N2\_vivax”, which has the ambiguity codes substituted. |

### Method: Validating results obtained with data set where ambiguity codes were substituted

We created custom functions to validate the result. This result makes use of the original alignment and identifies the number of untypeable sequences. We note that in 4/5 of the SNP sets, all the group of interest are typeable.

Find *sup\_4\_6\_custom\_metric.R* attached together

source("custom\_metric.R")

snps\_sets\_from\_n2vivax\_1k2\_MY <- list(

c(399231,352436),

c(391844,418938),

c(352436,52350,9443),

c(371027,37243,3426),

c(316619,396175,22508)

)

validations\_n2\_vivax\_1k2\_MY <- lapply(snps\_sets\_from\_n2vivax\_1k2\_MY, function(x){

return(find\_optimised\_snps(vivax, metric = "percent\_non\_standard", goi = GOI, iterate\_included = TRUE, included\_positions = x, max\_depth = 0))

})

for (result in validations\_n2\_vivax\_1k2\_MY){

output\_result(result)

cat(rep("\*", 100),"\n")

}

The result is summarised in the table below.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Vivax | | N2\_Vivax | |
|  | SNP set | % score | SNP set | % score |
| 1 | 395412, 17742, 51136, 387620, 4224 | 0.2607 | 399231, 352436 | 1 (20 untypeable) |
| 2 | 17742, 51136, 387620, 420741, 4224 | 0.2607 | 391844, 418938 | 1 (22 untypeable) |
| 3 | 51136, 146476, 387620, 264407, 4224 | 0.2512 | 352436, 52350, 9443 | 1 (52 untypeable) |
| 4 | 146476, 51994, 387620, 264407, 4224 | 0.2512 | 371027, 37243, 3426 | 1 (43 & 1 GOI untypeable) |
| 5 | 461322, 51994, 89236, 4224, 53474 | 0.2370 | 316619, 396175, 22508 | 1 (41 untypeable) |

## Derivation of SNP sets with "simpson" mode.

In this experiment, we identified five 5-member high-diversity index SNP sets.

seed <- 1633266595

set.seed(seed)

vivax <- read\_fasta("./Pv\_Nature\_communications\_2018\_VQSLOD3.min0.fasta")

## Using all the samples

p\_vivax <- process\_allele(vivax)

high\_d\_snps <- find\_optimised\_snps(seqc = p\_vivax, max\_depth = 5, number\_of\_result = 5, bp = bp, output\_progress = TRUE)

output\_result(high\_d\_snps, view = "csv", file\_name = "hd\_all\_vivax.tsv")

## Only using 80% of the samples

Malaysian\_isolates <- names(vivax)[

startsWith(names(vivax), "PY") &

(! names(vivax) %in% c("PY0045-C", "PY0004-C", "PY0120-C"))

]

Indonesian\_isolates <- names(vivax)[

startsWith(names(vivax), "PD")]

Thai\_isolates <- names(vivax)[

startsWith(names(vivax), "PJ")]

selected\_MY <- sample(Malaysian\_isolates, round(0.8\*length(Malaysian\_isolates)))

selected\_ID <- sample(Indonesian\_isolates, round(0.8\*length(Indonesian\_isolates)))

selected\_TH <- sample(Thai\_isolates, round(0.8\*length(Thai\_isolates)))

vivax <- vivax[c(selected\_MY, selected\_ID, selected\_TH)]

p\_vivax <- process\_allele(vivax)

high\_d\_snps <- find\_optimised\_snps(seqc = p\_vivax, max\_depth = 5, number\_of\_result = 5, bp = bp, output\_progress = TRUE)

output\_result(high\_d\_snps, view = "csv", file\_name = "hd\_80\_vivax.tsv")

## Using all the sample (N2\_vivax)

vivax <- read\_fasta("n2\_vivax.fasta")

p\_vivax <- process\_allele(vivax)

high\_d\_snps <- find\_optimised\_snps(seqc = p\_vivax, max\_depth = 5, number\_of\_result = 5, bp = bp, output\_progress = TRUE)

output\_result(high\_d\_snps, view = "csv", file\_name = "hd\_all\_n2.tsv")

## Using 80% of the samples (N2\_vivax)

vivax <- vivax[c(selected\_MY, selected\_ID, selected\_TH)]

p\_vivax <- process\_allele(vivax)

high\_d\_snps <- find\_optimised\_snps(seqc = p\_vivax, max\_depth = 5, number\_of\_result = 5, bp = bp, output\_progress = TRUE)

output\_result(high\_d\_snps, view = "csv", file\_name = "hd\_80\_n2.tsv")

Using Simpson result obtained with all of the samples (*sup\_4\_9\_hd\_all\_n2.txt*), we also identified the number of untypeable specimens and recalculated the *D* valueby excluding the untypeable specimens when N2\_vivax was used.

source("custom\_metric.R")

snps\_sets\_from\_n2vivax\_hd <- list(

c(488401,440313,330390,246684,281026),

c(346552,77554,272561,198732,259899),

c(332306,281139,161979,507255,272561),

c(56578,320955,266789,323311,484103),

c(66048,511351,9946,427558,20105)

)

validations\_n2\_vivax\_hd <- lapply(snps\_sets\_from\_n2vivax\_hd, function(x){

return(find\_optimised\_snps(vivax, metric = "simpson\_non\_standard", iterate\_included = TRUE, included\_positions = x, max\_depth = 0))

})

for (result in validations\_n2\_vivax\_hd){

output\_result(result)

cat(rep("\*", 100),"\n")

}

### Results

|  |  |
| --- | --- |
| Result | Description |
| See: *sup\_4\_7\_hd\_all\_vivax.txt* | High-diversity index SNPs identified with all the specimens, where a position containing at least an ambiguity codes were removed. |
| See: *sup\_4\_8\_hd\_80\_vivax.txt* | High-diversity index SNPs identified with 80% of the specimens, where a position containing at least an ambiguity codes were removed. |
| See: *sup\_4\_9\_hd\_all\_n2.txt* | High-diversity indexSNPs identified with all the specimens, where ambiguity codes were transformed to major allele. |
| See: *sup\_4\_10\_hd\_80\_n2.txt* | High-diversity indexSNPs identified with 80% of the specimens, where ambiguity codes were transformed to major allele. |

While the D-index output from MinSNPs included the untypeable specimen (i.e., those with ambiguity codes but transformed to most frequent base), we recalculated the diversity-index with the original matrix and excluded all untypeable specimen (those specimens with an allelic profile containing at least an ambiguity code).

|  |  |  |
| --- | --- | --- |
| No | D-index (excluding untypeable specimen) | No. untypeable specimen |
| 1 | 0.960133071553229 | 67 |
| 2 | 0.957961647347898 | 65 |
| 3 | 0.959503039915411 | 64 |
| 4 | 0.958004960044089 | 68 |
| 5 | 0.959790587219344 | 66 |

These are visualised in a phylogenetic tree, see *sup\_4\_11\_phylogeny\_genotypes.svg* or for interactive version, visit <https://microreact.org/project/minsnps-pvivax-profiles>. The figure shows the phylogeny of the vivax isolates obtained from three different countries in South East Asia taken from original paper (2). The genotypes from SNP sets identified from all the isolates through the two different means, i.e., excluding ambiguity codes and substituting ambiguity codes are presented. For genotypes in N2\_Vivax (inner most label), white/blank label are the untypeable isolates, i.e., those with a heterozygous / ambiguity code at the selected SNP position. High-diversity index SNP sets identified from alignment with ambiguity codes removed and substituted are located (position relative to chromosome) in 340505 (chromosome (Chr) 13), 460741 (Chr 12), 854772 (Chr 10), 531315 (Chr 6), 2100572 (Chr 12), and 1269895 (Chr 14), 1240935 (Chr 13), 1812716 (Chr 11), 1717060 (Chr 9), 1141805 (Chr 10) of the PvP01\_v1 reference genome respectively.

## References

1. Ortiz EM. vcf2phylip v2.0: convert a VCF matrix into several matrix formats for phylogenetic analysis. Zenodo; 2019.

2. Auburn S, Benavente ED, Miotto O, Pearson RD, Amato R, Grigg MJ, et al. Genomic analysis of a pre-elimination Malaysian Plasmodium vivax population reveals selective pressures and changing transmission dynamics. Nature Communications. 2018;9(1):2585-.