Pf3kv5_with_PNG

Somya Mehra

19/02/2019

Data Preprocessing

Variants have been processed through a pipeline following GATK best practices. Indels have been filtered out; only SNPs remain. Variants lying in subtelomeric hypervariable, subtelomeric repeat, centromeric or internal hypervariable regions have been removed. Functional annotations have been made using snpEff.

We begin by extracting functional annotations for our variants

```
metadata <- read.delim("pf3k_release_5_png_metadata_collated.txt",
                       header=TRUE, stringsAsFactors=FALSE, na.strings=c(""))

pf3kv5 <- seqOpen("pf3kv5_with_PNG.pass.snps.snpeff_ann.core.gds", readonly = FALSE)

# extract variant positions
var_annotations <- data.frame(chr = seqGetData(pf3kv5, "chromosome"),
                              pos = seqGetData(pf3kv5, "position"),
                              variant_id = seqGetData(pf3kv5, "variant.id"),
                              ref = as.character(ref(pf3kv5)))
nvar <- nrow(var_annotations)

# extract functional annotations -> multiple rows (one for each alternate allele)
# for multiallelic sites
snpeff_ann <- seqGetData(pf3kv5, "annotation/info/ANN")

func_annotations <- var_annotations[rep(1:nvar, times=snpeff_ann$length),] %>%
  mutate(snp_ann=snpeff_ann$data) %>%
  tidyr::separate(snp_ann, into=c("alt", "annotation", "impact", "gene_name", "gene_id",
                           "feature_type", "feature_id", "transcript_biotype", "rank",
                           "HGVS.c", "HGVS.p", "cDNA", "CDS", "AA"), sep="\\|", extra="drop") %>%
  select(-c("gene_name", "feature_id", "rank")) %>% distinct

readr::write_rds(func_annotations, "pf3kv5_functional_annotations.rds")
```

Our variant set contains 1485431 SNPs, segregated by chromosome as follows

SNP Counts by Chromosome

chr	count
Pf3D7_01_v3	32990
Pf3D7_02_v3	50648
Pf3D7_03_v3	61581
Pf3D7_04_v3	65612
Pf3D7_05_v3	87161
Pf3D7_06_v3	76660
Pf3D7_07_v3	81313
Pf3D7_08_v3	83673
Pf3D7_09_v3	102224
Pf3D7_10_v3	106900
Pf3D7_11_v3	137242
Pf3D7_12_v3	138788

chr	count
Pf3D7_13_v3	206208
Pf3D7_14_v3	251190
Pf3D7_API_v3	2581
Pf_M76611	660

We have permitted multiallelic sites. The number of alternate alleles per site is summarised below

Alt alleles per site

count	sites
1	1415550
2	67843
3	2038

We now consider metadata (country, region, site and sequencing facility) for each sample. Samples without assigned countries are lab-strains (either genetic crosses or simulated COI), which we exclude from further analysis. We define the following regions:

- West Africa: Senegal, The Gambia, Guinea, Mali, Ghana, Nigeria
- Central Africa: DR of the Congo, Malawi
- South East Asia: Bangladesh, Cambodia, Vietnam, Thailand, Myanmar, Laos
- PNG: Papua New Guinea

```r
west_africa <- c("TheGambia", "Ghana", "Guinea", "Mali", "Nigeria", "Senegal")
central_africa <- c("DRoftheCongo", "Malawi")
se_asia <- c("Bangladesh", "Cambodia", "Laos", "Myanmar", "Thailand", "Vietnam")

metadata$region <- NA
metadata$region[metadata$country %in% west_africa] <- "WestAfrica"
metadata$region[metadata$country %in% central_africa] <- "CentralAfrica"
metadata$region[metadata$country %in% se_asia] <- "SEAsia"
metadata$region[metadata$country=="PNG"] <- "PNG"

metadata$sequencing <- "Sanger"
metadata[metadata$country=="Senegal" & !is.na(metadata$country),]$sequencing <- "BROAD"
metadata[substr(metadata$sample, 1, 3)=="PNG" & (!is.na(metadata$country)),]$sequencing <- "BROAD"


field_isolates <- metadata %>% filter(!is.na(country))
```

We have 2661 field isolates in total, segregating by region/country/site/sequencing facility as follows

Sample Counts by Site

region	country	site	count
CentralAfrica	DRoftheCongo	Kinshasa	113
CentralAfrica	Malawi	Chikwawa	317
CentralAfrica	Malawi	Zomba	52
PNG	PNG	Alotau	30
PNG	PNG	Madang	67
PNG	PNG	Maprik	52
SEAsia	Bangladesh	Ramu	50
SEAsia	Cambodia	Pailin	84

region	country	site	count
SEAsia	Cambodia	PreahVihear	104
SEAsia	Cambodia	Pursat	235
SEAsia	Cambodia	Ratanakiri	147
SEAsia	Laos	Attapeu	85
SEAsia	Myanmar	BagoDivision	60
SEAsia	Thailand	MaeSot	106
SEAsia	Thailand	Ranong	20
SEAsia	Thailand	Sisakhet	22
SEAsia	Vietnam	BuDang	1
SEAsia	Vietnam	BuGiaMap	64
SEAsia	Vietnam	PhuocLong	32
WestAfrica	Ghana	Kassena	549
WestAfrica	Ghana	Kintampo	68
WestAfrica	Guinea	Nzerekore	100
WestAfrica	Mali	Bandiagara	9
WestAfrica	Mali	Faladje	36
WestAfrica	Mali	Kolle	51
WestAfrica	Nigeria	Ilorin	5
WestAfrica	Senegal	Thies	133
WestAfrica	Senegal	Velingara	4
WestAfrica	TheGambia	GM_Coastal	65

Sample Counts by Sequencing Facility

sequencing	count
BROAD	151
Sanger	2510

We annotate our GDS file with metadata for each sample (region, country, site, sequencing facility)

```
metadata <- metadata %>% arrange(sample)
# if(cnt.gdsn(index.gdsn(pf3kv5, "sample.annotation")) != 4) {
#  add.gdsn(index.gdsn(pf3kv5, "sample.annotation"),
#           name = "region", val = metadata$region)
#  add.gdsn(index.gdsn(pf3kv5, "sample.annotation"),
#           name = "country",val = metadata$country)
#  add.gdsn(index.gdsn(pf3kv5, "sample.annotation"),
#           name = "site", val = metadata$site)
#  add.gdsn(index.gdsn(pf3kv5, "sample.annotation"),
#           name = "sequencing_facility", val = metadata$sequencing)
# }
```

Variant and Sample Filtration

We restrict our analysis to field isolates and variants on the 14 nuclear chromosomes only

```
nuclear <- var_annotations %>% subset(chr!="Pf3D7_API_v3" & chr!="Pf_M76611")
seqSetFilter(pf3kv5, variant.id=nuclear$variant_id,
             sample.id=field_isolates$sample)
```

```
## # of selected samples: 2,661
## # of selected variants: 1,482,190
```

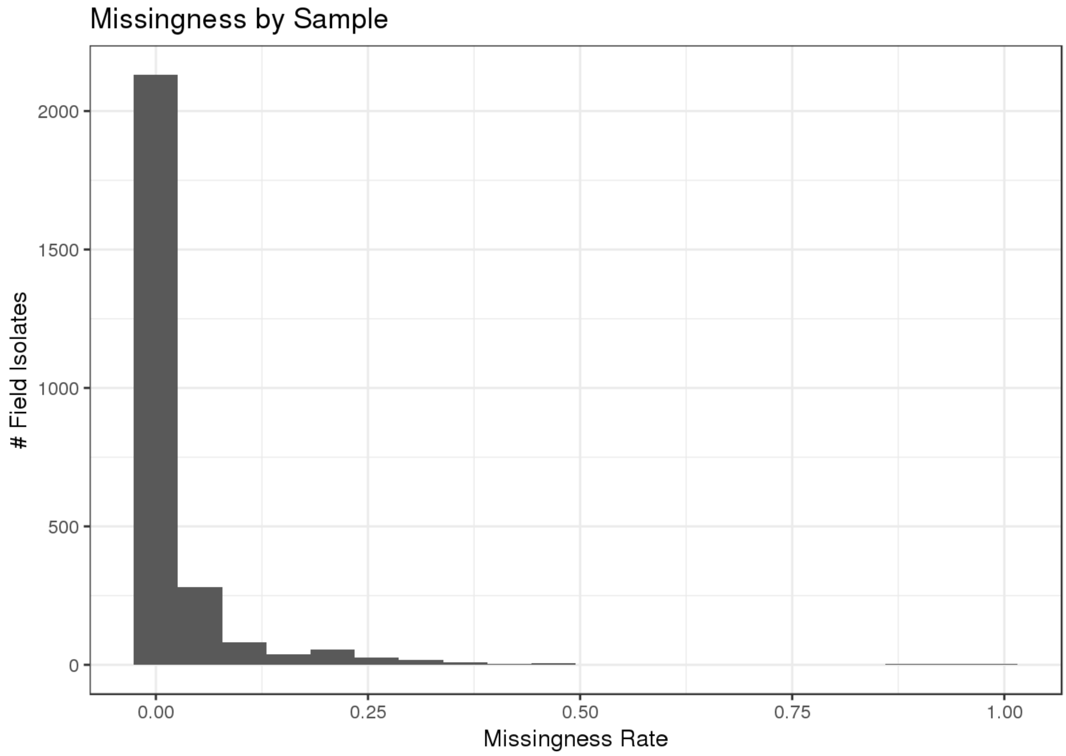
We consider the missingness per sample, and remove isolates with missingness rate >0.1

```
missing_by_sample <- data.frame(sample=field_isolates$sample, missing=seqMissing(pf3kv5, per.variant = FALSE
))
keep_samples <- as.character(missing_by_sample$sample[missing_by_sample$missing<=0.1])
field_isolates <- field_isolates %>% subset(sample %in% keep_samples)
seqResetFilter(pf3kv5)
```

```
## # of selected samples: 2,789
## # of selected variants: 1,485,431
```

```
seqSetFilter(pf3kv5, variant.id=nuclear$variant_id,
             sample.id=keep_samples)
```

```
## # of selected samples: 2,463
## # of selected variants: 1,482,190
```



We have 2463 field isolates remaining after filtration, stratified geographically as follows

Sample Counts by Site

region	country	site	count
CentralAfrica	DRoftheCongo	Kinshasa	100
CentralAfrica	Malawi	Chikwawa	314
CentralAfrica	Malawi	Zomba	50

region	country	site	count
PNG	PNG	Alotau	29
PNG	PNG	Madang	52
PNG	PNG	Maprik	50
SEAsia	Bangladesh	Ramu	45
SEAsia	Cambodia	Pailin	84
SEAsia	Cambodia	PreahVihear	104
SEAsia	Cambodia	Pursat	231
SEAsia	Cambodia	Ratanakiri	146
SEAsia	Laos	Attapeu	85
SEAsia	Myanmar	BagoDivision	60
SEAsia	Thailand	MaeSot	104
SEAsia	Thailand	Ranong	20
SEAsia	Thailand	Sisakhet	21
SEAsia	Vietnam	BuDang	1
SEAsia	Vietnam	BuGiaMap	64
SEAsia	Vietnam	PhuocLong	32
WestAfrica	Ghana	Kassena	520
WestAfrica	Ghana	Kintampo	63
WestAfrica	Guinea	Nzerekore	100
WestAfrica	Mali	Bandiagara	8
WestAfrica	Mali	Faladje	27
WestAfrica	Mali	Kolle	34
WestAfrica	Nigeria	Ilorin	5
WestAfrica	Senegal	Thies	52
WestAfrica	Senegal	Velingara	4
WestAfrica	TheGambia	GM_Coastal	58

Sample Counts by Sequencing Facility

sequencing	count
BROAD	70
Sanger	2393

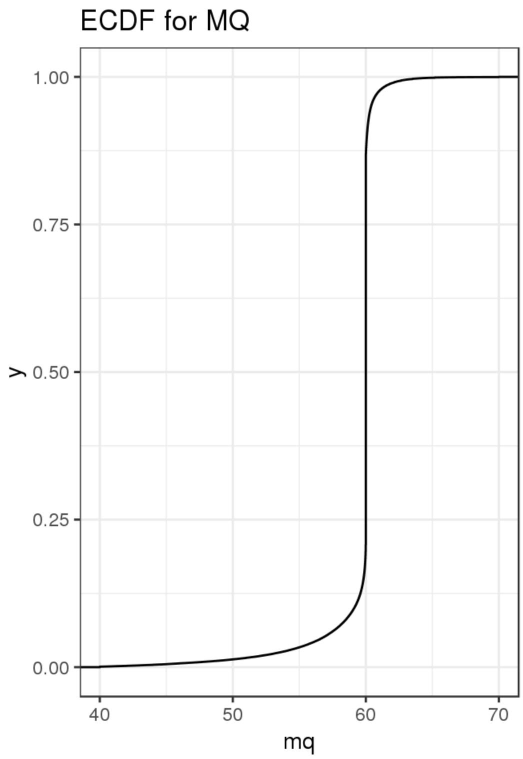
Variants have been annotated with the following quality metrics

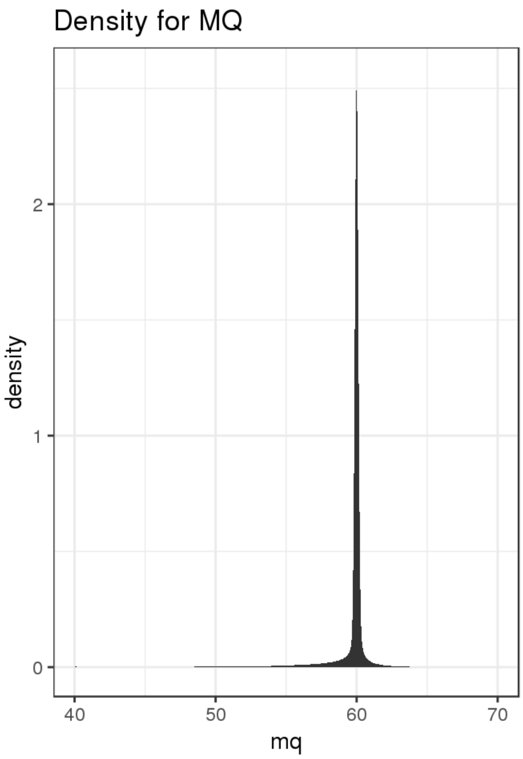
```
## Annotation, INFO variable(s):
##     AC, A, Integer, Allele count in genotypes, for each ALT allele, in the same order as listed
##     AF, A, Float, Allele Frequency, for each ALT allele, in the same order as listed
##     AN, 1, Integer, Total number of alleles in called genotypes
##     BaseQRankSum, 1, Float, Z-score from Wilcoxon rank sum test of Alt Vs. Ref base qualities
##     DP, 1, Integer, Approximate read depth; some reads may have been filtered
##     DS, 0, Flag, Were any of the samples downsampled?
##     ExcessHet, 1, Float, Phred-scaled p-value for exact test of excess heterozygosity
##     FS, 1, Float, Phred-scaled p-value using Fisher's exact test to detect strand bias
##     InbreedingCoeff, 1, Float, Inbreeding coefficient as estimated from the genotype likelihoods per-samp
le when compared against the Hardy-Weinberg expectation
##     MQ, 1, Float, RMS Mapping Quality
##     MQRankSum, 1, Float, Z-score From Wilcoxon rank sum test of Alt vs. Ref read mapping qualities
##     QD, 1, Float, Variant Confidence/Quality by Depth
##     ReadPosRankSum, 1, Float, Z-score from Wilcoxon rank sum test of Alt vs. Ref read position bias
##     SOR, 1, Float, Symmetric Odds Ratio of 2x2 contingency table to detect strand bias
##     ANN, ., String, Functional annotations: 'Allele | Annotation | Annotation_Impact | Gene_Name | Gene_I
D | Feature_Type | Feature_ID | Transcript_BioType | Rank | HGVS.c | HGVS.p | cDNA.pos / cDNA.length | CDS.
pos / CDS.length | AA.pos / AA.length | Distance | ERRORS / WARNINGS / INFO'
```

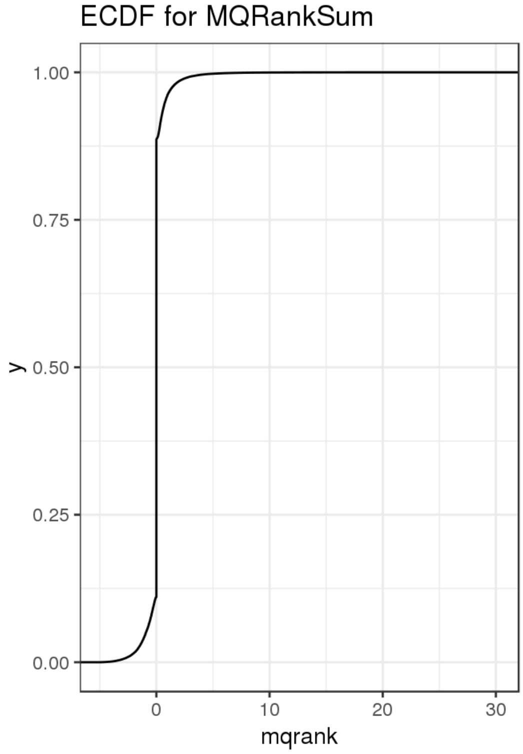
Variant quality score recalibration has been conducted, and variants failing the following filters have already been removed

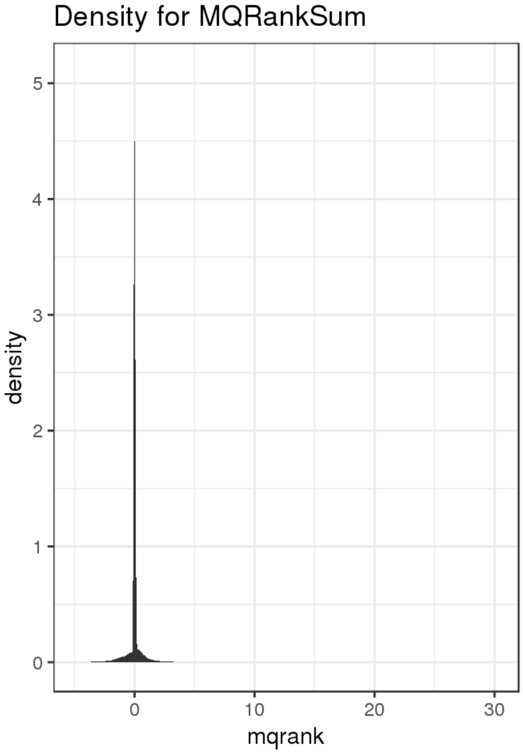
```
## Annotation, FILTER:
##     PASS, All filters passed, 1482190(100.0%)
##     LowQual, Low quality, 0(0.0%)
##     highExcessHet, Set if true: ExcessHet>54.69, 0(0.0%)
##     lowQD, Set if true: QD<2.0, 0(0.0%)
##     lowMQ, Set if true: MQ<30.0, 0(0.0%)
##     highFS, Set if true: FS>100.0, 0(0.0%)
##     highSOR, Set if true: SOR>7.5, 0(0.0%)
##     lowReadPosRankSum, Set if true: ReadPosRankSum<-5.0, 0(0.0%)
##     lowInbreedingCoeff, Set if true: InbreedingCoeff<-0.8, 0(0.0%)
##     lowMQRankSum, Set if true: MQRankSum<-5.0, 0(0.0%)
```

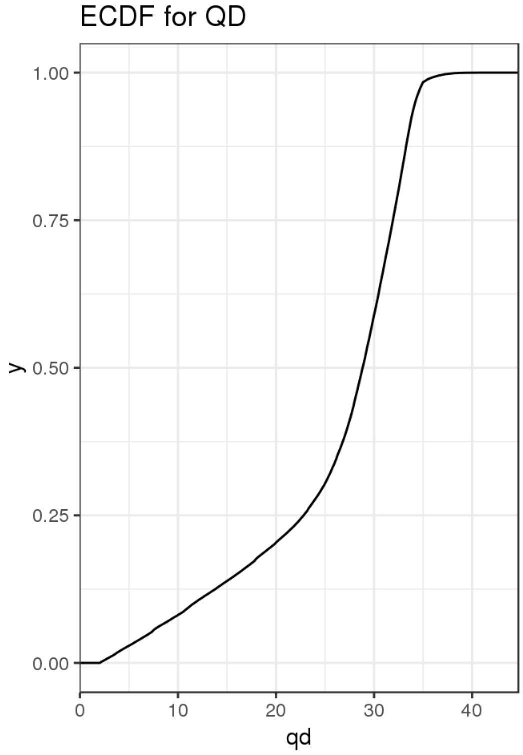
We consider various GATK metrics to see whether further hard-filtration is necessary

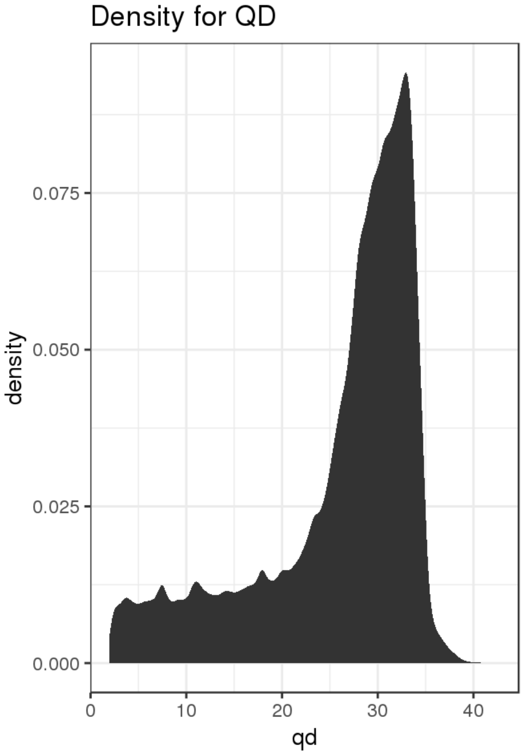


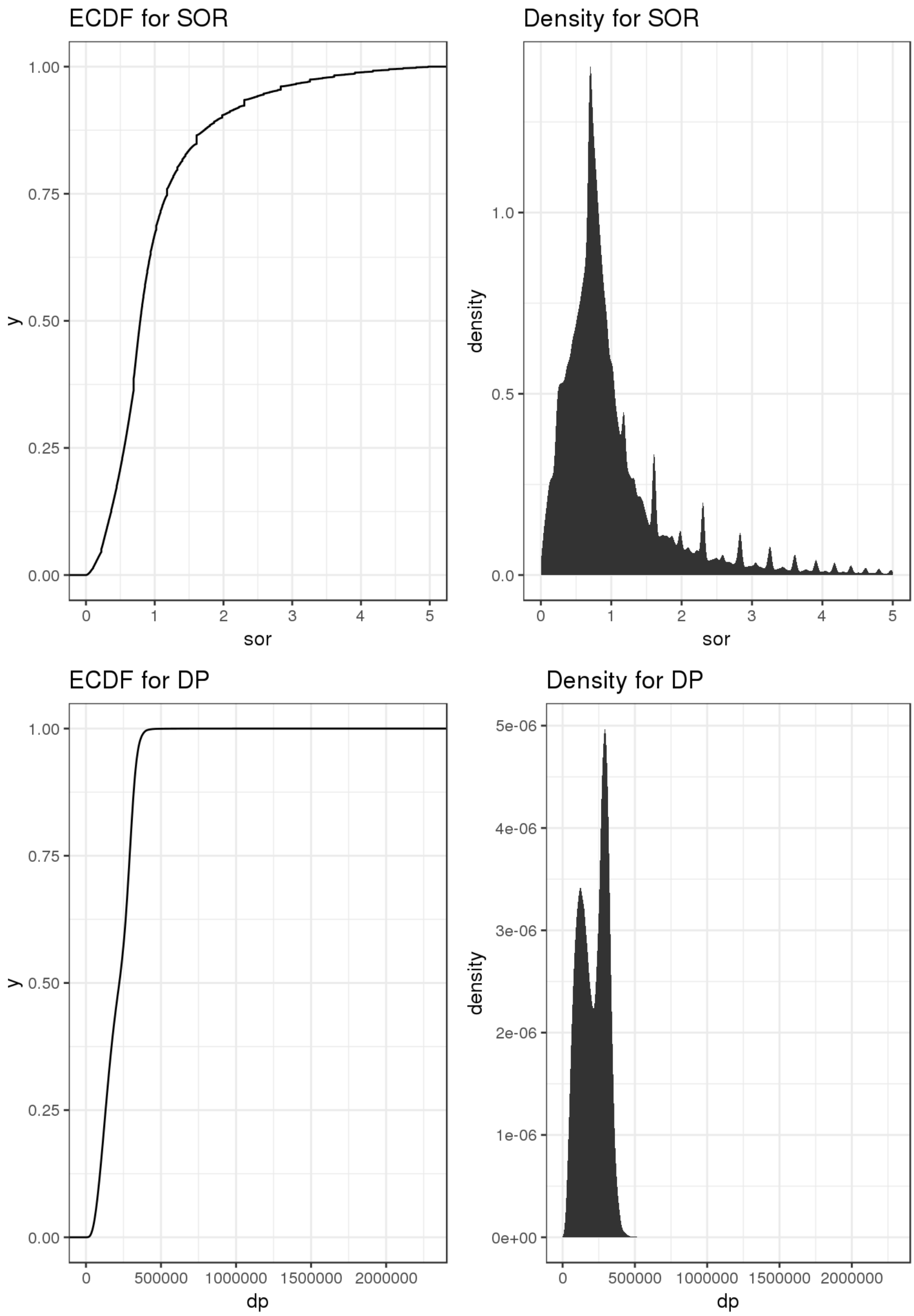



ECDF for MQRankSum
y
1.00
0.75
0.50
0.25
0.00
0
10
20
30
mqrank


Density for MQRankSum
density
5
4
3
2
1
0
0
10
20
30
mqrank


ECDF for QD
y
1.00
0.75
0.50
0.25
0.00
0
10
20
30
40
qd


Density for QD
density
0.075
0.050
0.025
0.000
0
10
20
30
40
qd

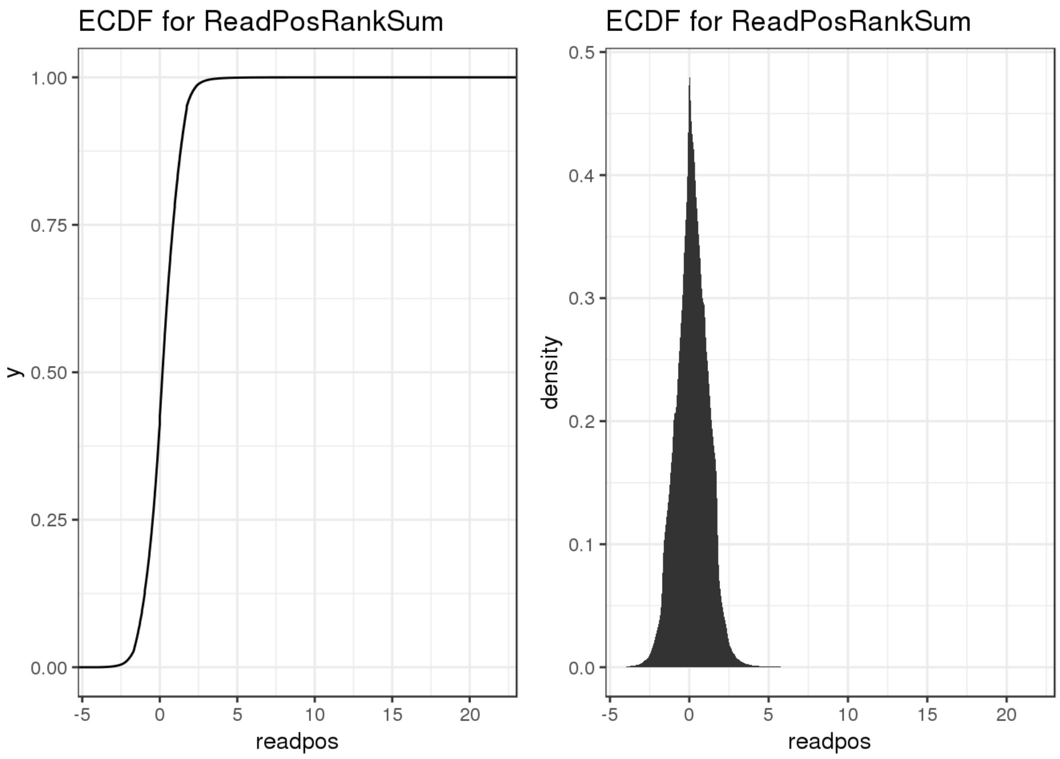


ECDF for SOR

Density for SOR

ECDF for DP

Density for DP



We apply the following hard-filters:

- QD > 20
- MQ > 50
- MQRankSum > -2 (if annotated)
- SOR < 1
- -4 < ReadPosRankSum < 4 (if annotated)

Note that there are 140 variants without MQ, 465554 without MQRankSum, 469376 without ReadPosRankSum and 676 without QD annotations.

```
hard_filtered <- gatk_metrics %>%
  subset(qd>20 & sor<1 & mq>50 &
           (mqrank>-2 | is.nan(mqrank)) &
           ((readpos>-4 & readpos<4) | is.nan(readpos)))
```

770591 SNPs are retained after hard-filtering by GATK metrics. We now consider the missingness per site, and remove sites with missingness rate >0.1.

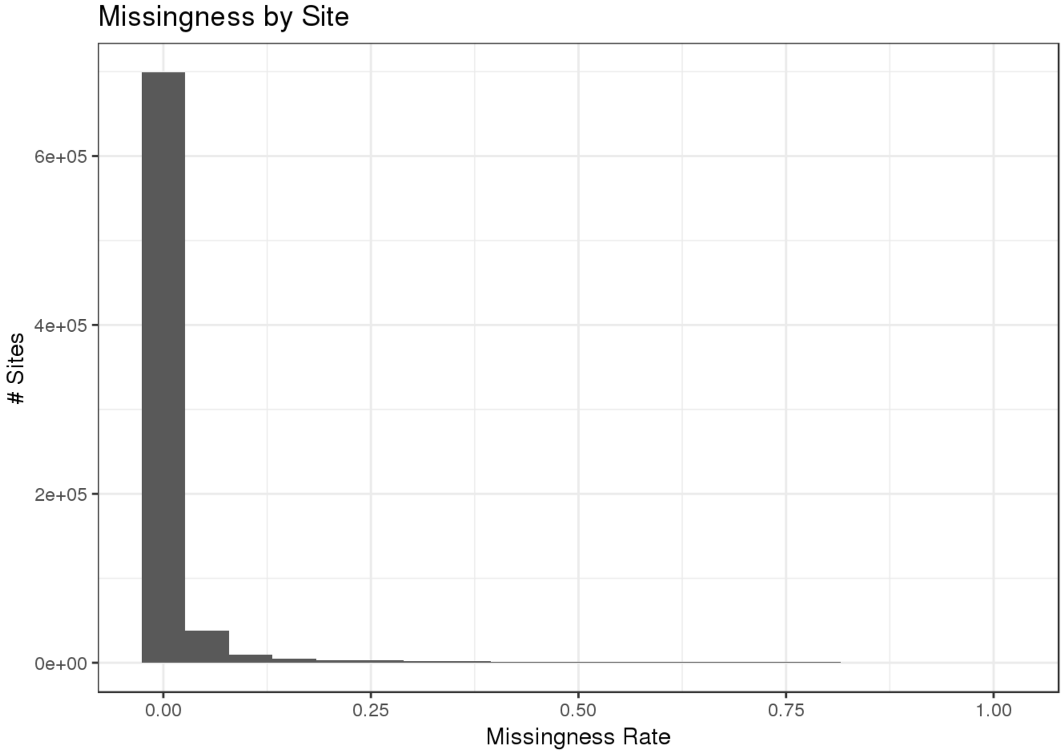
```
seqResetFilter(pf3kv5)
```

```
## # of selected samples: 2,789
## # of selected variants: 1,485,431
```

```
seqSetFilter(pf3kv5, variant.id=hard_filtered$variant.id,
             sample.id=keep_samples)
```

```
## # of selected samples: 2,463
## # of selected variants: 770,591
```

```
missing_per_snp <- data.frame(var=hard_filtered$variant.id, missing=seqMissing(pf3kv5, per.variant = TRUE))
keep_low_missing_var <- missing_per_snp$var[missing_per_snp$missing<=0.1]
hard_filtered <- hard_filtered %>% subset(variant.id %in% keep_low_missing_var)
```



742365 SNPs are retained after hard filtering by GATK metrics and missingness.

```
seqResetFilter(pf3kv5)
```

```
## # of selected samples: 2,789
## # of selected variants: 1,485,431
```

```
seqSetFilter(pf3kv5, variant.id=hard_filtered$variant.id,
             sample.id=keep_samples)
```

```
## # of selected samples: 2,463
## # of selected variants: 742,365
```

```
func_annotations %>% subset(variant_id %in% hard_filtered$variant.id) %>% readr::write_rds("pf3kv5_functiona
l_annotations_filtered.rds")
```

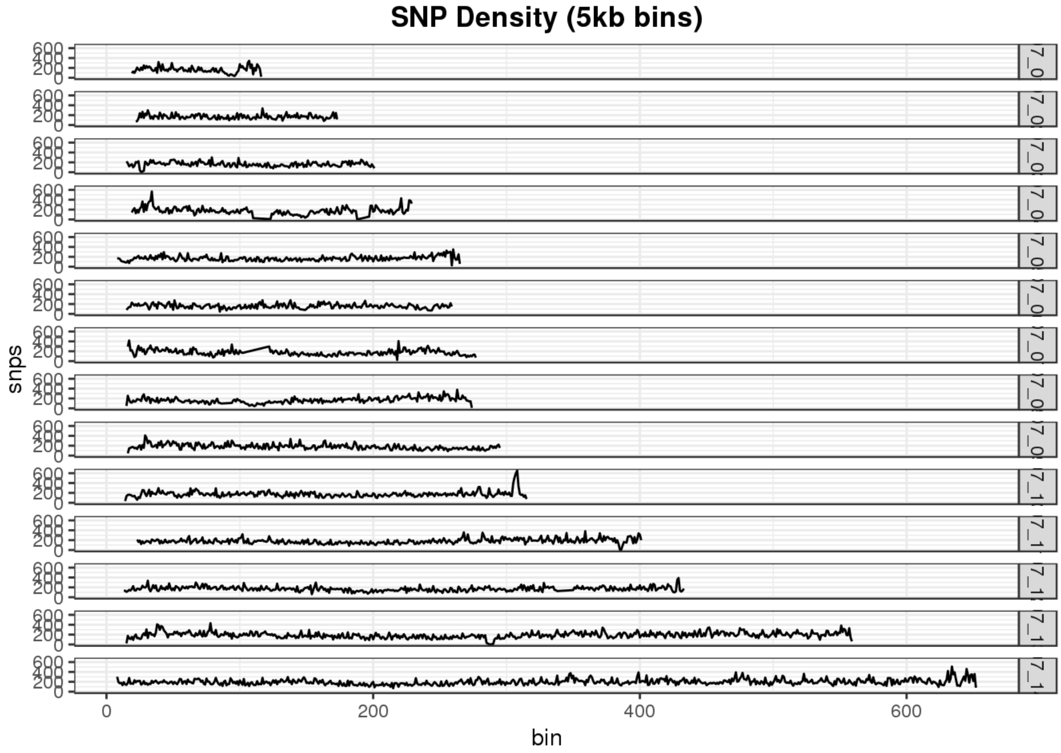
The final set of SNPs segregates across chromosomes as follows

Hard-filtered SNP counts

chr	count
Pf3D7_01_v3	16100
Pf3D7_02_v3	25432
Pf3D7_03_v3	30244
Pf3D7_04_v3	33723
Pf3D7_05_v3	43241
Pf3D7_06_v3	38174
Pf3D7_07_v3	42358
Pf3D7_08_v3	41796
Pf3D7_09_v3	50248
Pf3D7_10_v3	52970

chr	count
Pf3D7_11_v3	68240
Pf3D7_12_v3	69087
Pf3D7_13_v3	103682
Pf3D7_14_v3	127070

SNP densities across 5kb bins are shown below



A break-down of the number of alternate alleles per site is given below

Alt alleles per site

count	sites
1	697275
2	43623
3	1467

FWS and MAF Estimates

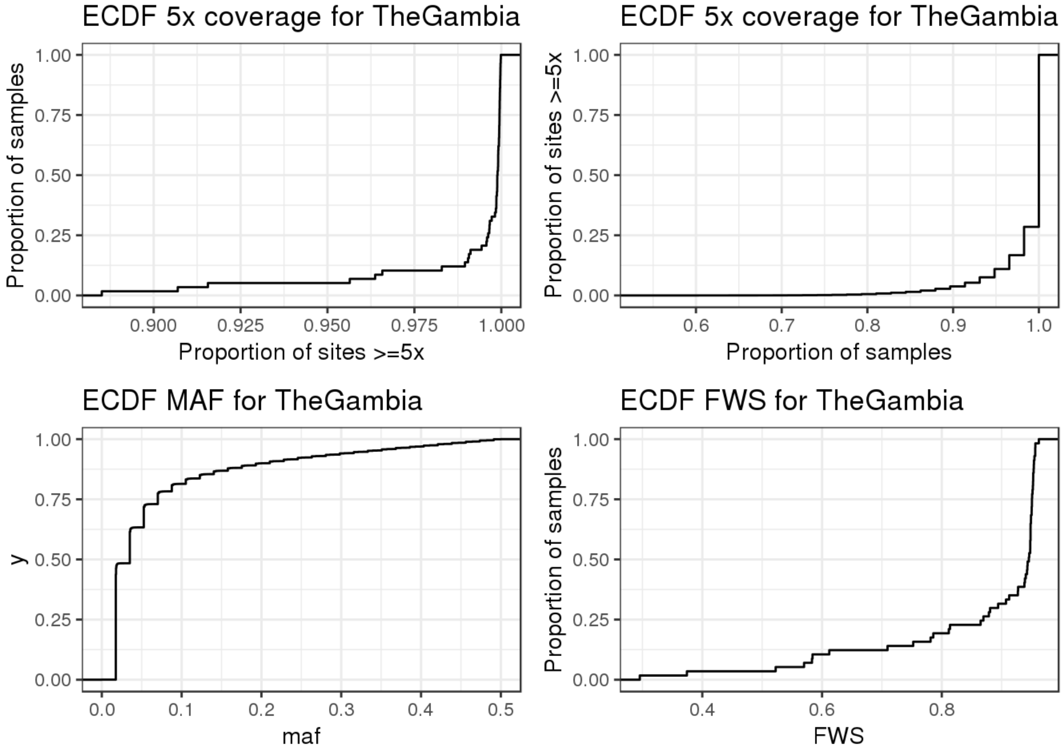
We stratify our hard-filtered data by country and sequentially apply the following filters:

1. Extract polymorphic SNPs i.e. MAF>0
2. Extract SNPs that are covered by >=5 reads in at least 90% of samples
3. Extract samples that have 90% of filtered SNPs covered by >=5 reads

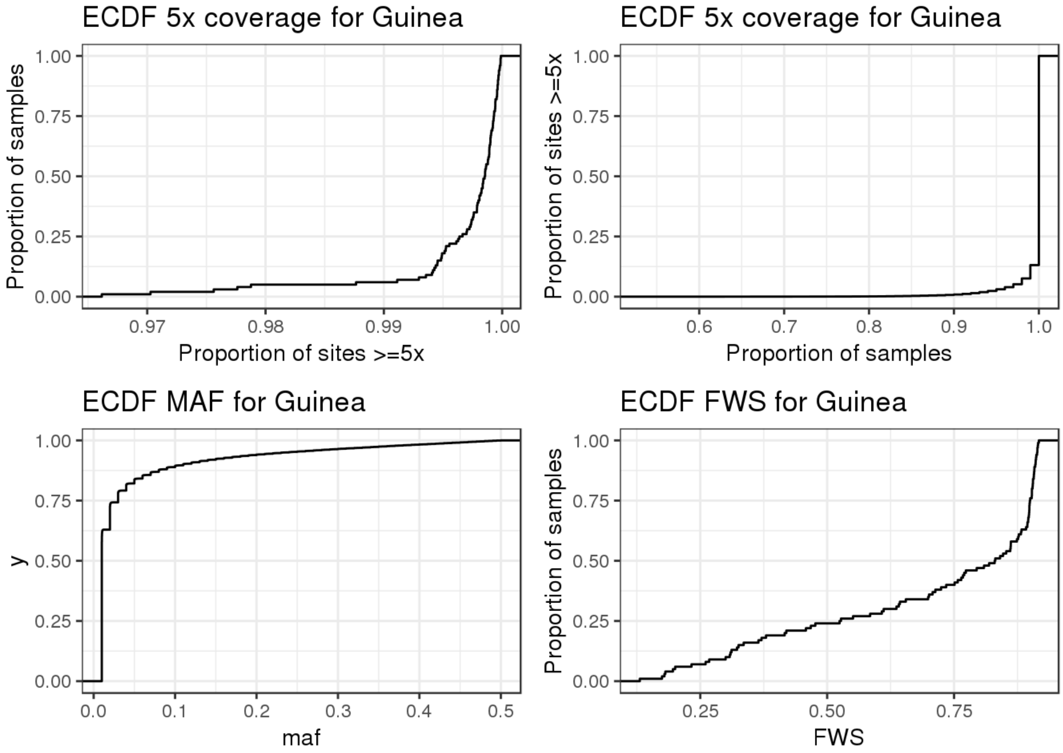
We then calculate FWS and MAF using these by-country datasets.

```
pf3kv5_stats <- list()

maf_by_site <- function(pf3kv5) {
  seqApply(pf3kv5, "genotype",
           FUN=function(x) {raf=mean(x==0L, na.rm=TRUE); return(min(raf, 1-raf))},
           as.is="double", margin="by.variant") %>% return
}

for (region in unique(field_isolates$region)) {
  pf3kv5_stats[[region]] <- list()
  countries <- unique(field_isolates$country[field_isolates$region == region])
  for (country in countries) {
    seqResetFilter(pf3kv5)
    pf3kv5_stats[[region]][[country]] <- list()
    cat(paste("## Now running", country, "estimates\n"))
    # filter samples from country of interest
    samples <- field_isolates$sample[field_isolates$country == country]
    seqSetFilter(pf3kv5, sample.id = samples, variant.id = hard_filtered$variant.id)
    # keep variants that are polymorphic in country of interest
    maf_country <- maf_by_site(pf3kv5)
    keep_polymorphic_var <- hard_filtered$variant.id[maf_country>0 & !is.na(maf_country)]
    seqSetFilter(pf3kv5, variant.id=keep_polymorphic_var)
    # keep SNPs that are covered by >=5 reads in 90% of samples
    cov_depth <- seqGetData(pf3kv5, "annotation/format/DP")$data
    var_5x_cov <- colSums(cov_depth>=5)/nrow(cov_depth)
    keep_polymorphic_5x_cov <- keep_polymorphic_var[var_5x_cov>=0.9]
    # keep samples that have 90% of filtered SNPs covered by >=5 reads
    sample_5x_cov <- rowSums(cov_depth[,var_5x_cov>=0.9]>=5)/sum(var_5x_cov>=0.9)
    keep_sample_5x <- samples[sample_5x_cov>=0.9]
    seqSetFilter(pf3kv5, variant.id=keep_polymorphic_5x_cov, sample.id=keep_sample_5x)
    # compute MAF for filtered sites + samples
    pf3kv5_stats[[region]][[country]]$maf <- data.frame(variant.id=keep_polymorphic_5x_cov,
                                                        maf=maf_by_site(pf3kv5))
    # compute FWS for filtered sites + samples
    pf3kv5_stats[[region]][[country]]$fws <- data.frame(sample.id=keep_sample_5x,
                                                        fws=getFws(pf3kv5))
    # generate ECDF curves for coverage, MAF and FWS
    maf_plot <- ggplot(pf3kv5_stats[[region]][[country]]$maf, aes(x=maf)) +
      stat_ecdf() + ggtitle(paste("ECDF MAF for", country)) + theme_bw()
    sample_cov_plot <- ggplot(data.frame(cov=sample_5x_cov), aes(x=cov)) +
      stat_ecdf() + ggtitle(paste("ECDF 5x coverage for", country)) + xlab("Proportion of sites >=5x") +
      ylab("Proportion of samples") + theme_bw()
    var_cov_plot <- ggplot(data.frame(cov=var_5x_cov), aes(x=cov)) +
      stat_ecdf() + ggtitle(paste("ECDF 5x coverage for", country)) + xlab("Proportion of samples") +
      ylab("Proportion of sites >=5x") + theme_bw()
    fws_plot <- ggplot(pf3kv5_stats[[region]][[country]]$fws, aes(x=fws)) +
      stat_ecdf() + ggtitle(paste("ECDF FWS for", country)) + xlab("FWS") +
      ylab("Proportion of samples") + theme_bw()

    print(plot_grid(sample_cov_plot, var_cov_plot, maf_plot, fws_plot))
  }
}
```

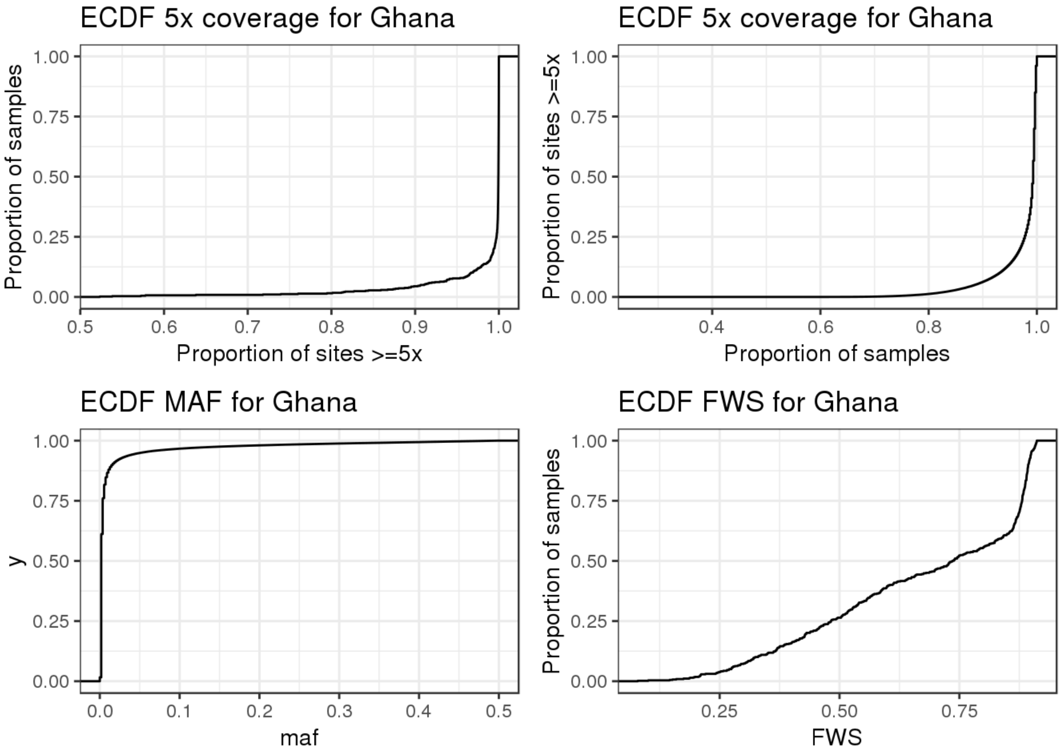
```
## # of selected samples: 2,789
## # of selected variants: 1,485,431
## ## Now running TheGambia estimates
## # of selected samples: 58
## # of selected variants: 742,365
## # of selected variants: 78,343
## # of selected samples: 57
## # of selected variants: 75,392
```



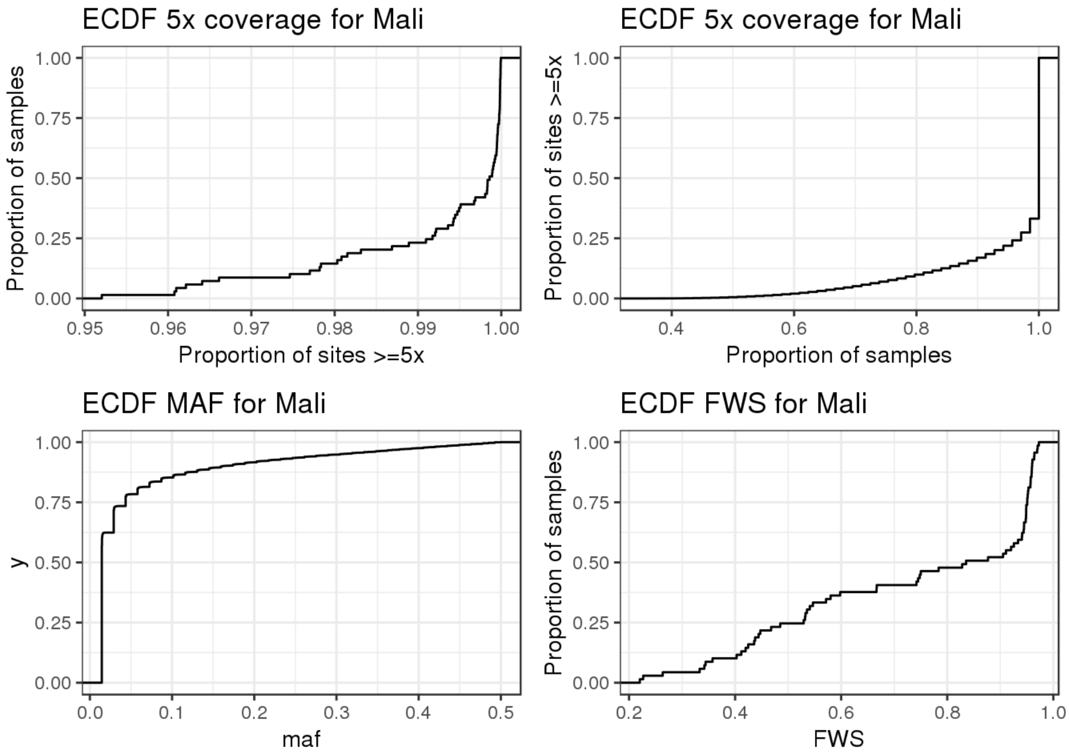
```
## # of selected samples: 2,789
## # of selected variants: 1,485,431
## ## Now running Guinea estimates
## # of selected samples: 100
## # of selected variants: 742,365
## # of selected variants: 124,771
## # of selected samples: 100
## # of selected variants: 123,879
```



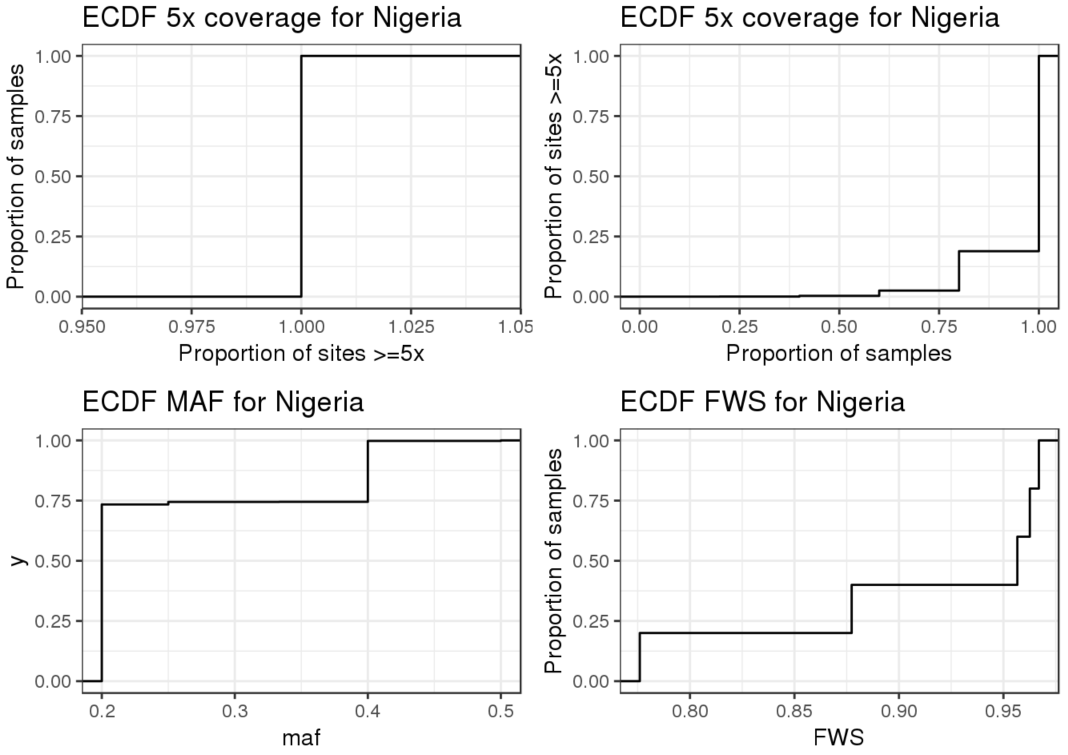
```
## # of selected samples: 2,789
## # of selected variants: 1,485,431
## ## Now running Ghana estimates
## # of selected samples: 583
## # of selected variants: 742,365
## # of selected variants: 378,843
## # of selected samples: 557
## # of selected variants: 355,378
```



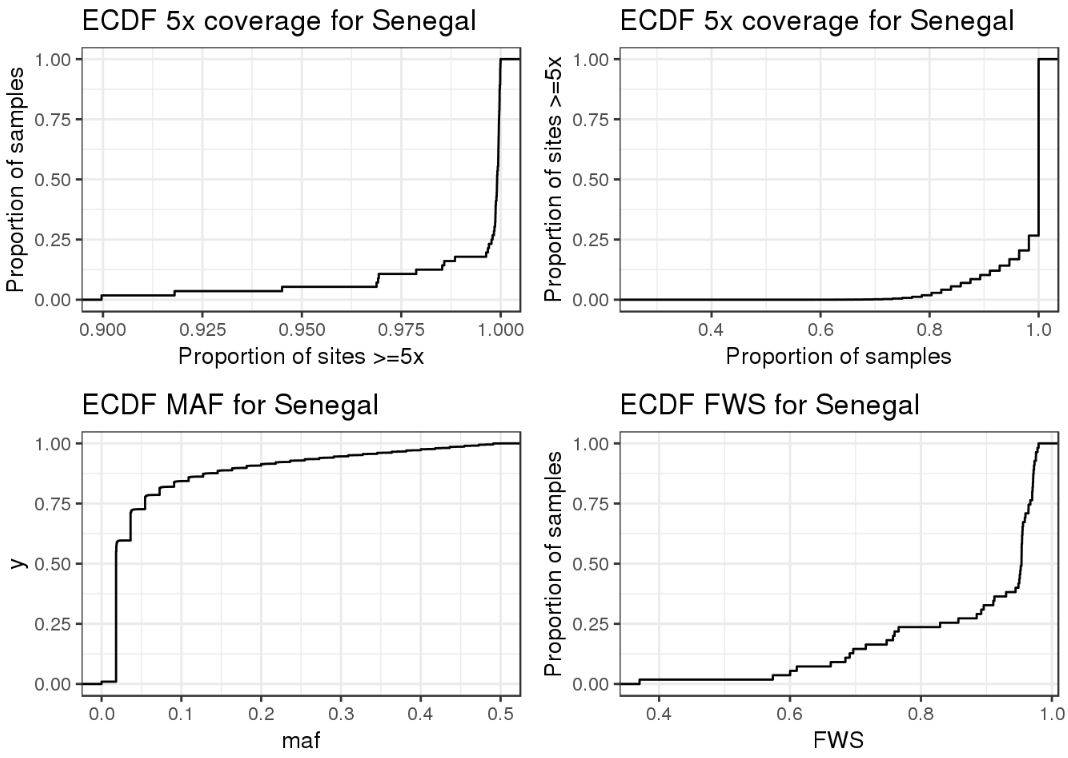
```
## # of selected samples: 2,789
## # of selected variants: 1,485,431
## ## Now running Mali estimates
## # of selected samples: 69
## # of selected variants: 742,365
## # of selected variants: 90,841
## # of selected samples: 69
## # of selected variants: 75,450
```



```
## # of selected samples: 2,789
## # of selected variants: 1,485,431
## ## Now running Nigeria estimates
## # of selected samples: 5
## # of selected variants: 742,365
## # of selected variants: 19,906
## # of selected samples: 5
## # of selected variants: 16,148
```

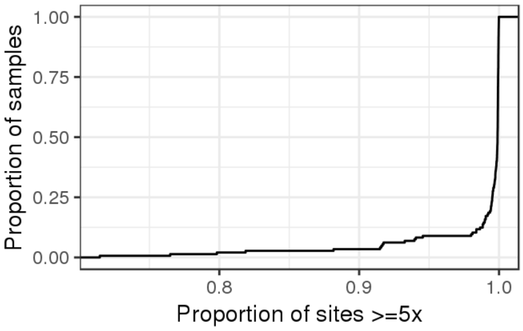


```
## # of selected samples: 2,789
## # of selected variants: 1,485,431
## ## Now running Senegal estimates
## # of selected samples: 56
## # of selected variants: 742,365
## # of selected variants: 91,147
## # of selected samples: 55
## # of selected variants: 81,792
```

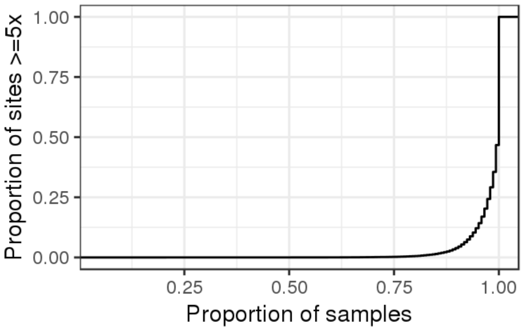


```
## # of selected samples: 2,789
## # of selected variants: 1,485,431
## ## Now running Thailand estimates
## # of selected samples: 145
## # of selected variants: 742,365
## # of selected variants: 51,518
## # of selected samples: 140
## # of selected variants: 49,506
```

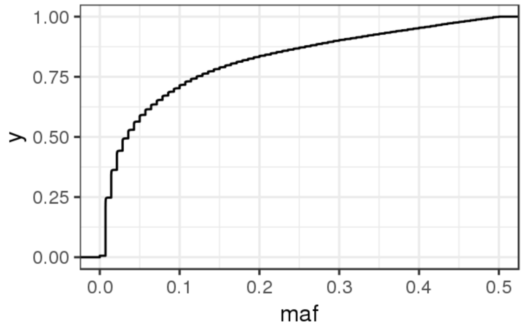
ECDF 5x coverage for Thailand



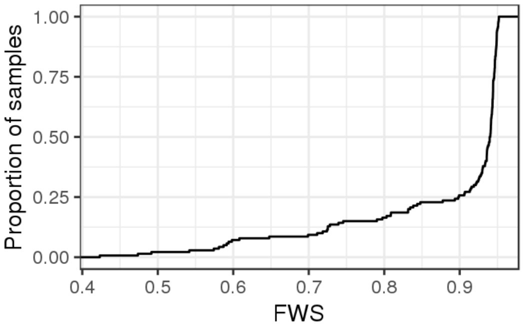
ECDF 5x coverage for Thailand



ECDF MAF for Thailand

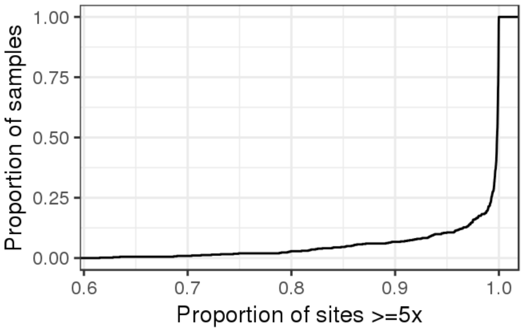


ECDF FWS for Thailand

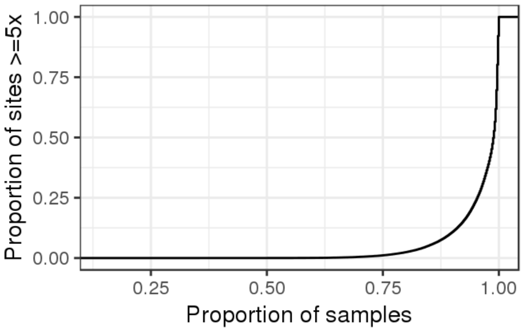


```
## # of selected samples: 2,789
## # of selected variants: 1,485,431
## ## Now running Cambodia estimates
## # of selected samples: 565
## # of selected variants: 742,365
## # of selected variants: 85,478
## # of selected samples: 527
## # of selected variants: 76,305
```

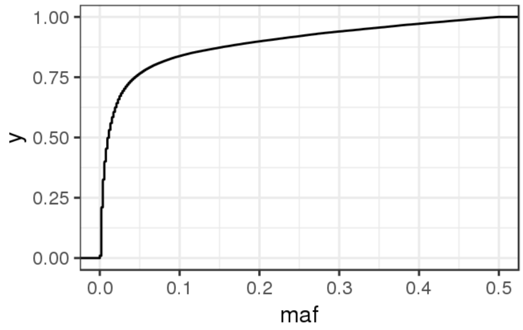
ECDF 5x coverage for Cambodia



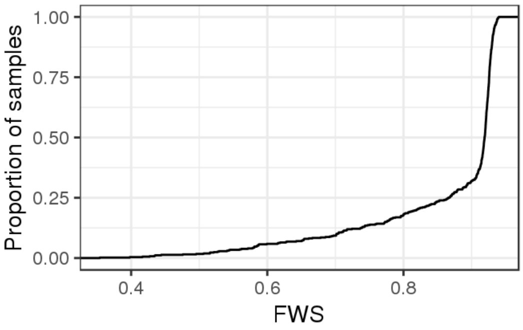
ECDF 5x coverage for Cambodia



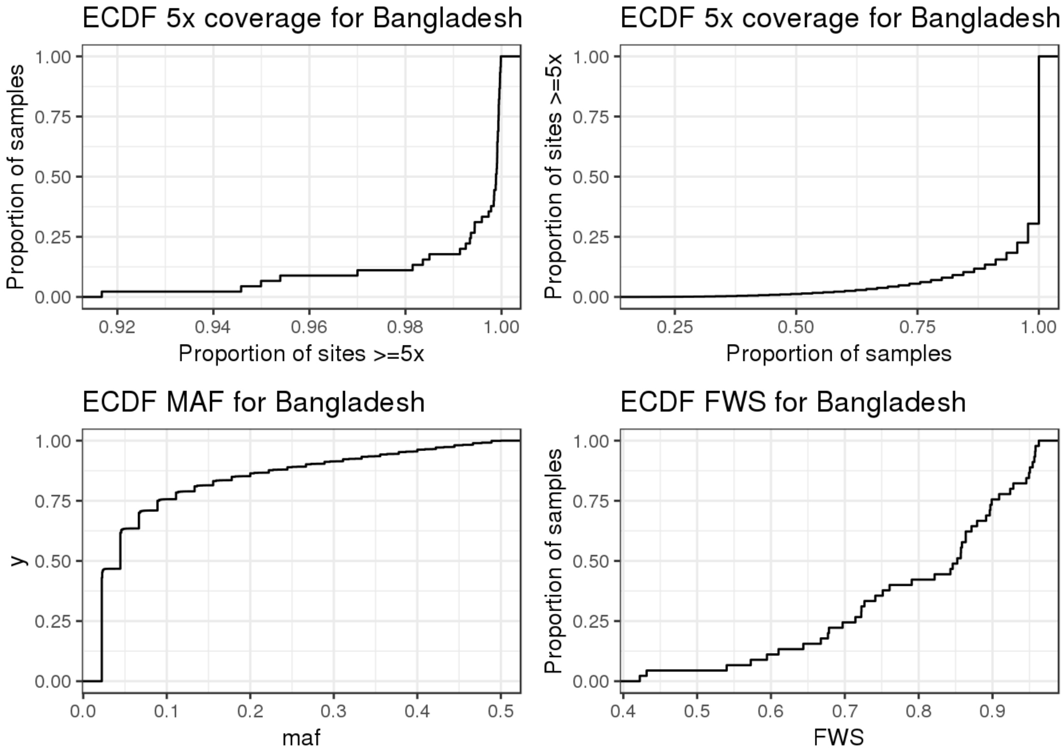
ECDF MAF for Cambodia



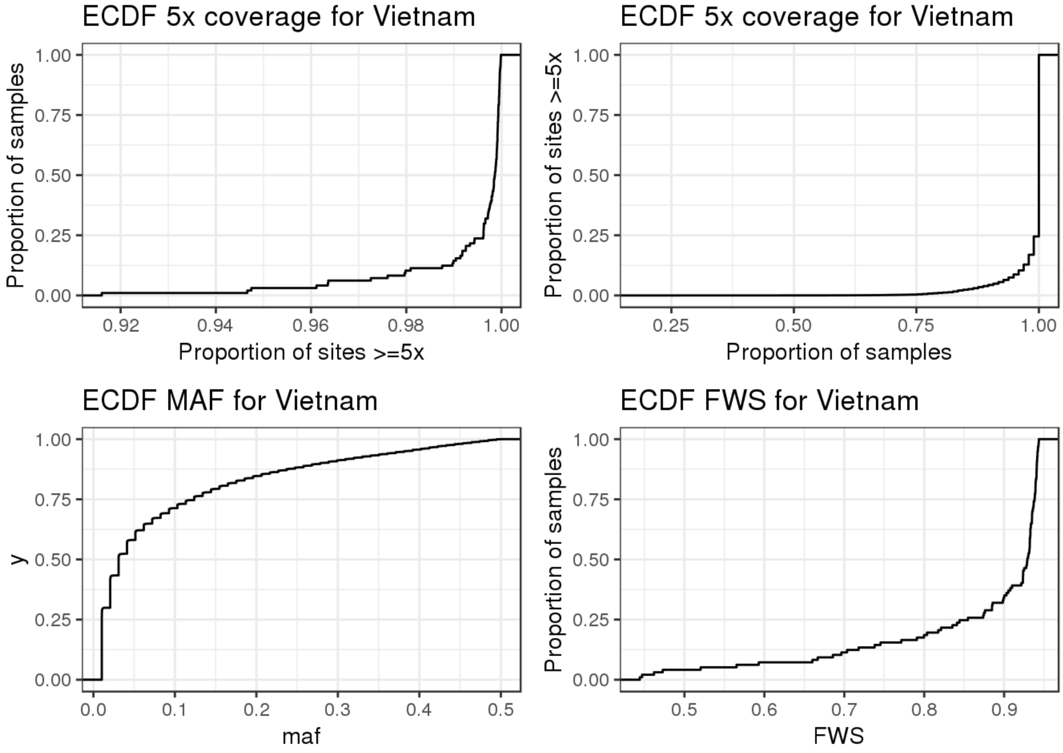
ECDF FWS for Cambodia



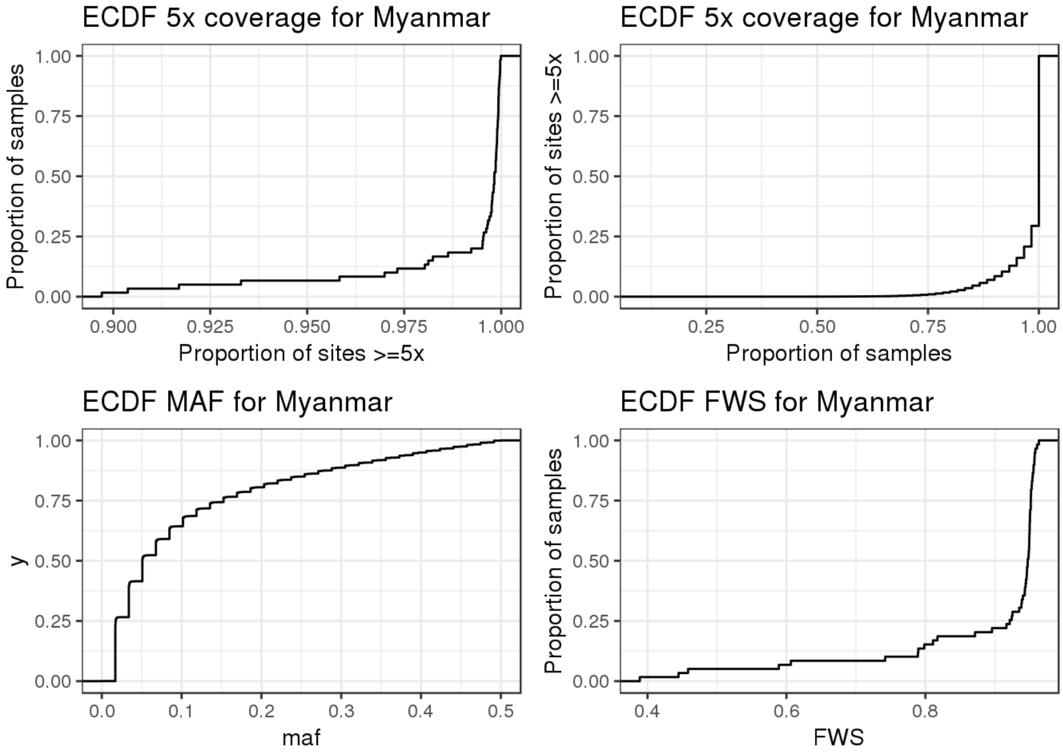
```
## # of selected samples: 2,789
## # of selected variants: 1,485,431
## ## Now running Bangladesh estimates
## # of selected samples: 45
## # of selected variants: 742,365
## # of selected variants: 57,969
## # of selected samples: 45
## # of selected variants: 50,182
```



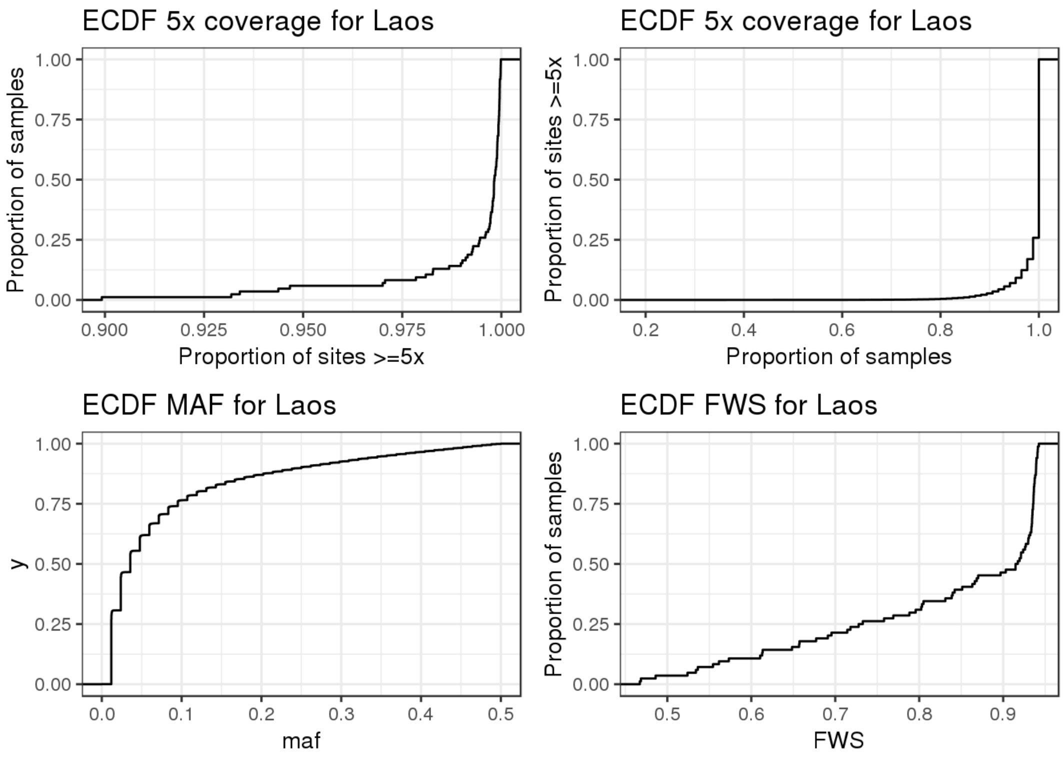
```
## # of selected samples: 2,789
## # of selected variants: 1,485,431
## ## Now running Vietnam estimates
## # of selected samples: 97
## # of selected variants: 742,365
## # of selected variants: 53,198
## # of selected samples: 97
## # of selected variants: 50,874
```



```
## # of selected samples: 2,789
## # of selected variants: 1,485,431
## ## Now running Myanmar estimates
## # of selected samples: 60
## # of selected variants: 742,365
## # of selected variants: 43,946
## # of selected samples: 59
## # of selected variants: 40,886
```



```
## # of selected samples: 2,789
## # of selected variants: 1,485,431
## ## Now running Laos estimates
## # of selected samples: 85
## # of selected variants: 742,365
## # of selected variants: 58,735
## # of selected samples: 84
## # of selected variants: 57,136
```



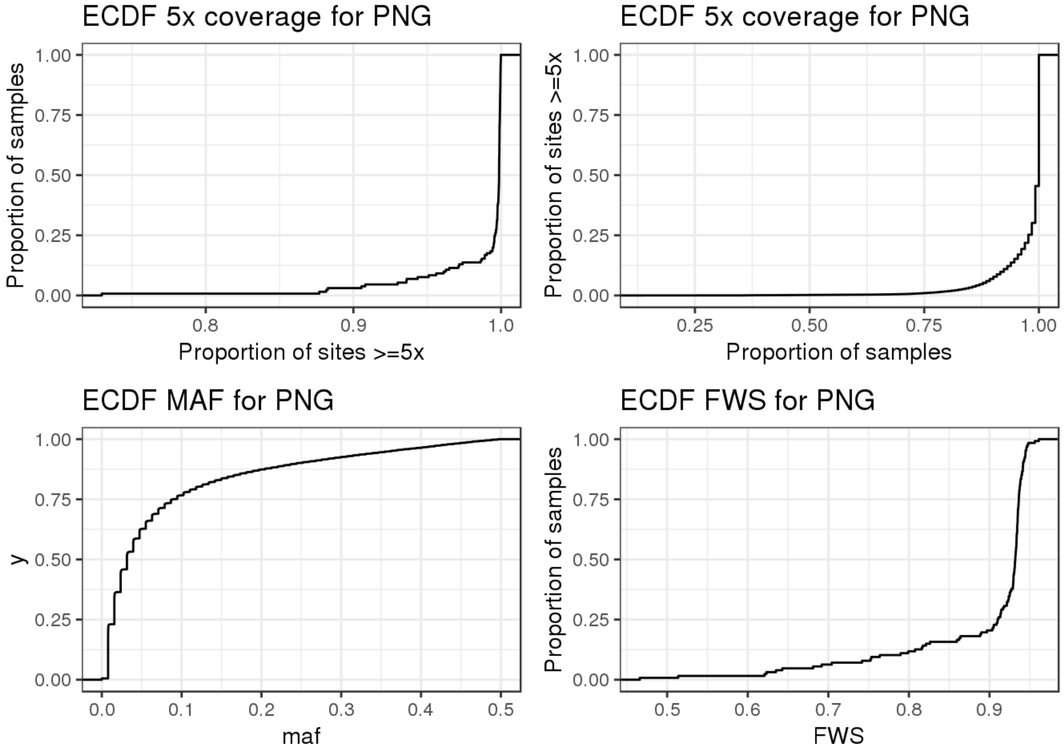
ECDF 5x coverage for Laos

ECDF 5x coverage for Laos

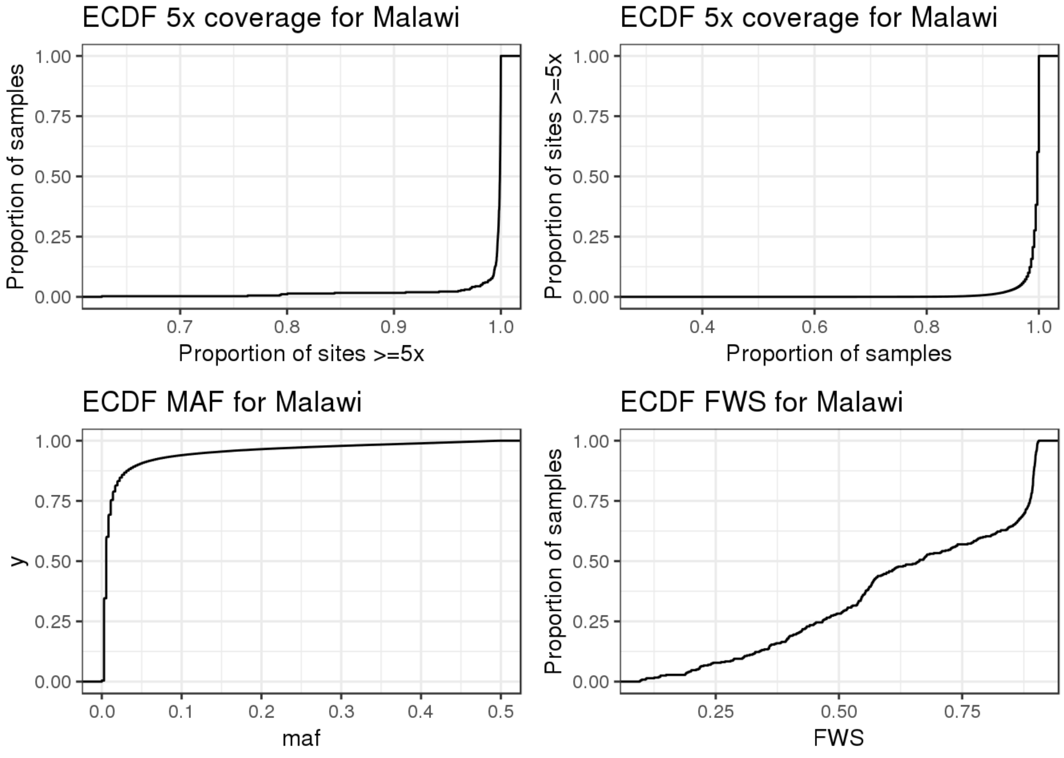
ECDF MAF for Laos

ECDF FWS for Laos

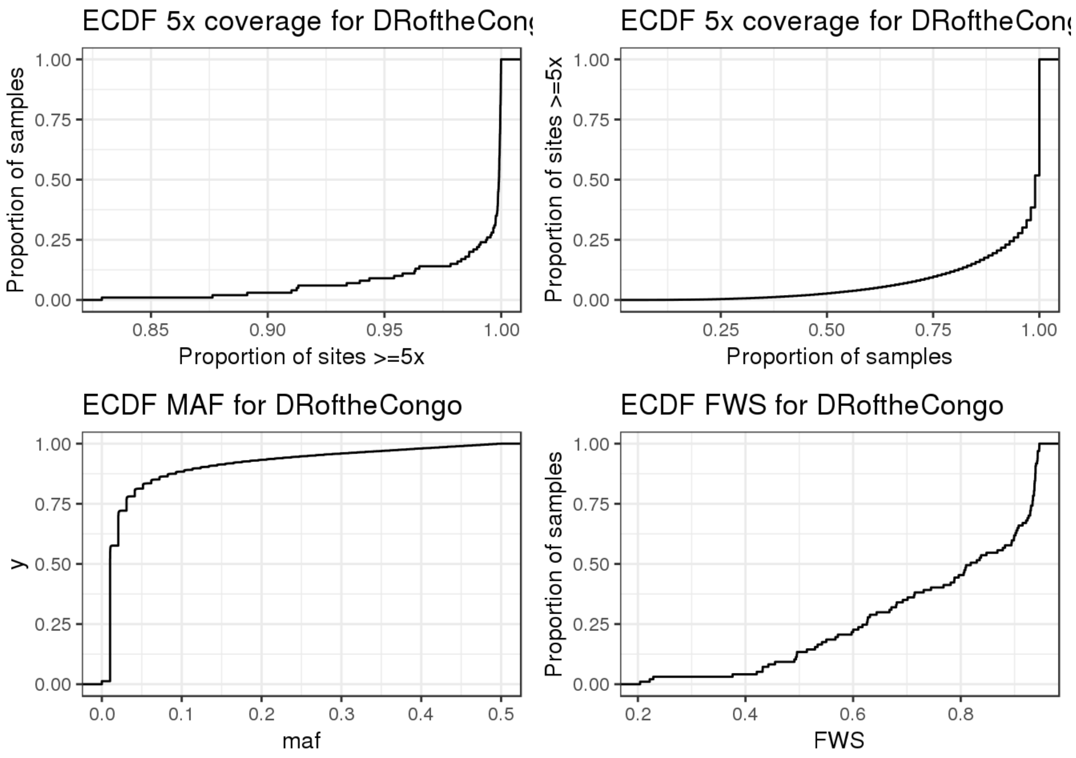
```
## # of selected samples: 2,789
## # of selected variants: 1,485,431
## ## Now running PNG estimates
## # of selected samples: 131
## # of selected variants: 742,365
## # of selected variants: 62,137
## # of selected samples: 127
## # of selected variants: 57,874
```



```
## # of selected samples: 2,789
## # of selected variants: 1,485,431
## ## Now running Malawi estimates
## # of selected samples: 364
## # of selected variants: 742,365
## # of selected variants: 207,203
## # of selected samples: 358
## # of selected variants: 205,803
```



```
## # of selected samples: 2,789
## # of selected variants: 1,485,431
## ## Now running DRoftheCongo estimates
## # of selected samples: 100
## # of selected variants: 742,365
## # of selected variants: 113,074
## # of selected samples: 97
## # of selected variants: 91,071
```



```
readr::write_rds(pf3kv5_stats, "pf3kv5_stats.rds")
```

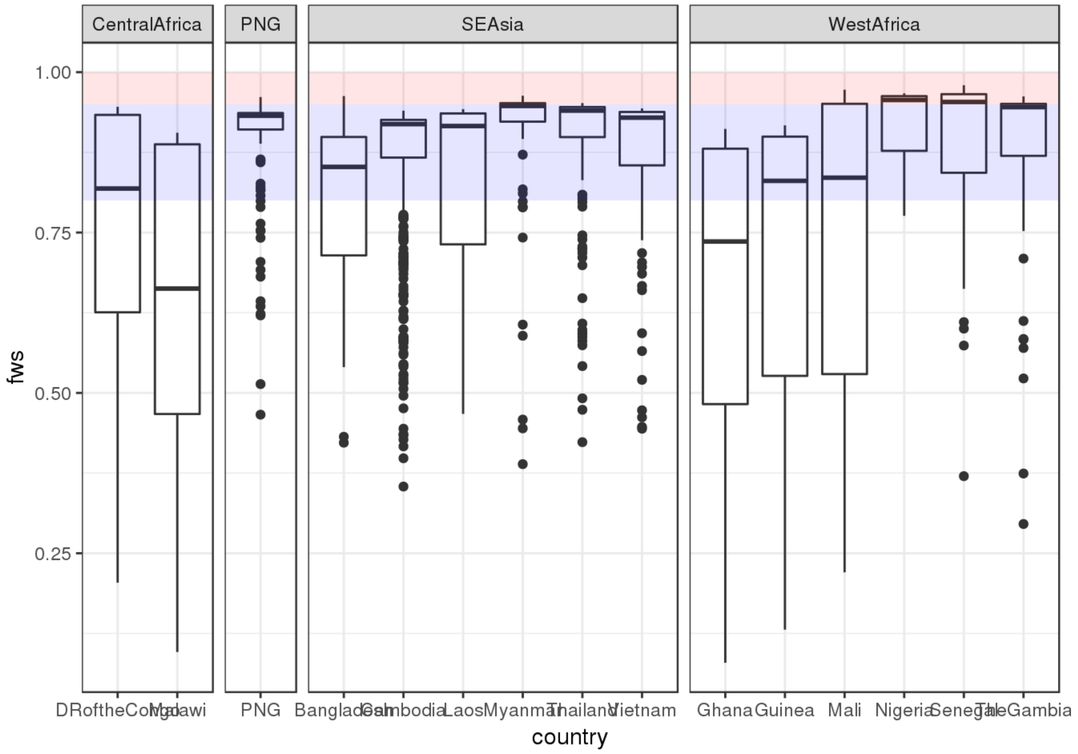
MOI Estimates

We determine MOI by considering FWS values

- MOI = 1: FWS > 0.95
- MOI = 2: 0.80 < FWS <= 0.95
- MOI > 2: FWS <= 0.80

```
fws <- data.frame(sample.id=character(0), fws=numeric(0))
for (region in names(pf3kv5_stats)) {
  for (country in names(pf3kv5_stats[[region]])) {
    fws <- pf3kv5_stats[[region]][[country]]$fws %>%
      mutate(country=country, region=region,
             moi=ifelse(fws>0.95, 1, ifelse(fws>0.8, 2, NA))) %>% rbind(fws)
  }
}
readr::write_rds(fws, "pf3kv5_fws.rds")
```

We compare the distribution of FWS values across geographic strata



A summary of MOI across our samples is given below. NA indicates indeterminate MOI > 2.

MOI by Country

moi	count
1	108
2	1391
NA	878

MOI by Country

region	country	moi	count
CentralAfrica	DRoftheCongo	2	53
CentralAfrica	DRoftheCongo	NA	44
CentralAfrica	Malawi	2	142
CentralAfrica	Malawi	NA	216
PNG	PNG	1	2
PNG	PNG	2	110
PNG	PNG	NA	15
SEAsia	Bangladesh	1	6
SEAsia	Bangladesh	2	20
SEAsia	Bangladesh	NA	19
SEAsia	Cambodia	2	433
SEAsia	Cambodia	NA	94
SEAsia	Laos	2	58

region	country	moi	count
SEAsia	Laos	NA	26
SEAsia	Myanmar	1	22
SEAsia	Myanmar	2	28
SEAsia	Myanmar	NA	9
SEAsia	Thailand	1	7
SEAsia	Thailand	2	110
SEAsia	Thailand	NA	23
SEAsia	Vietnam	2	79
SEAsia	Vietnam	NA	18
WestAfrica	Ghana	2	248
WestAfrica	Ghana	NA	309
WestAfrica	Guinea	2	53
WestAfrica	Guinea	NA	47
WestAfrica	Mali	1	18
WestAfrica	Mali	2	18
WestAfrica	Mali	NA	33
WestAfrica	Nigeria	1	3
WestAfrica	Nigeria	2	1
WestAfrica	Nigeria	NA	1
WestAfrica	Senegal	1	32
WestAfrica	Senegal	2	10
WestAfrica	Senegal	NA	13
WestAfrica	TheGambia	1	18
WestAfrica	TheGambia	2	28
WestAfrica	TheGambia	NA	11

Appendix

```
showfile.gds(closeall=TRUE)
```

```
##
FileName
## 1 /stornext/General/data/academic/lab_barry/Somya/Pf3k_PNG/report/pf3kv5_with_PNG.pass.snps.snpeff_ann.co
re.gds
##   ReadOnly  State
## 1    FALSE closed
```

```
sessionInfo()
```

```
## R version 3.5.1 (2018-07-02)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: CentOS Linux 7 (Core)
## 
## Matrix products: default
## BLAS: /stornext/System/data/apps/R/R-3.5.1/lib64/R/lib/libRblas.so
## LAPACK: /stornext/System/data/apps/R/R-3.5.1/lib64/R/lib/libRlapack.so
## 
## locale:
## [1] C
## 
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base     
## 
## other attached packages:
##  [1] bindrcpp_0.2.2    readr_1.3.1       kableExtra_1.0.1 
##  [4] knitr_1.20        cowplot_0.9.3     ggplot2_3.1.0    
##  [7] tidyr_0.8.2       moimix_0.0.1.9001 flexmix_2.3-14   
## [10] lattice_0.20-38   dplyr_0.7.8       SeqArray_1.22.3  
## [13] gdsfmt_1.18.1    
## 
## loaded via a namespace (and not attached):
##  [1] httr_1.4.0             Biobase_2.42.0         viridisLite_0.3.0     
##  [4] splines_3.5.1          assertthat_0.2.0       highr_0.7             
##  [7] stats4_3.5.1           GenomeInfoDbData_1.2.0 GWASExactHW_1.01      
## [10] yaml_2.2.0             pillar_1.3.1           backports_1.1.3       
## [13] quantreg_5.36          glue_1.3.0             digest_0.6.18         
## [16] GenomicRanges_1.34.0   RColorBrewer_1.1-2     XVector_0.22.0        
## [19] rvest_0.3.2            minqa_1.2.4            colorspace_1.4-0      
## [22] htmltools_0.3.6        Matrix_1.2-15          plyr_1.8.4            
## [25] pkgconfig_2.0.2        broom_0.5.0            SparseM_1.77          
## [28] zlibbioc_1.28.0        purrr_0.3.0            webshot_0.5.1         
## [31] scales_1.0.0           SeqVarTools_1.20.1     BiocParallel_1.16.0   
## [34] lme4_1.1-19            MatrixModels_0.4-2     tibble_2.0.1          
## [37] mgcv_1.8-27            IRanges_2.16.0         withr_2.1.2           
## [40] pan_1.6                nnet_7.3-12            BiocGenerics_0.28.0   
## [43] lazyeval_0.2.1         survival_2.43-1        magrittr_1.5          
## [46] crayon_1.3.4           mitml_0.3-6            mcmc_0.9-5            
## [49] evaluate_0.12          mice_3.3.0             nlme_3.1-137          
## [52] MASS_7.3-51.1          xml2_1.2.0             tools_3.5.1           
## [55] hms_0.4.2              stringr_1.4.0          MCMCpack_1.4-4        
## [58] S4Vectors_0.20.1       munsell_0.5.0          Biostrings_2.50.2     
## [61] compiler_3.5.1         GenomeInfoDb_1.18.1    logistf_1.23          
## [64] rlang_0.3.1            grid_3.5.1             RCurl_1.95-4.11       
## [67] nloptr_1.2.1           rstudioapi_0.9.0       labeling_0.3          
## [70] bitops_1.0-6           rmarkdown_1.11         gtable_0.2.0          
## [73] reshape2_1.4.3         R6_2.3.0               bindr_0.1.1           
## [76] jomo_2.6-5             modeltools_0.2-22      stringi_1.2.4         
## [79] parallel_3.5.1         Rcpp_1.0.0             rpart_4.1-13          
## [82] tidyselect_0.2.5       coda_0.19-2
```