Positive selection

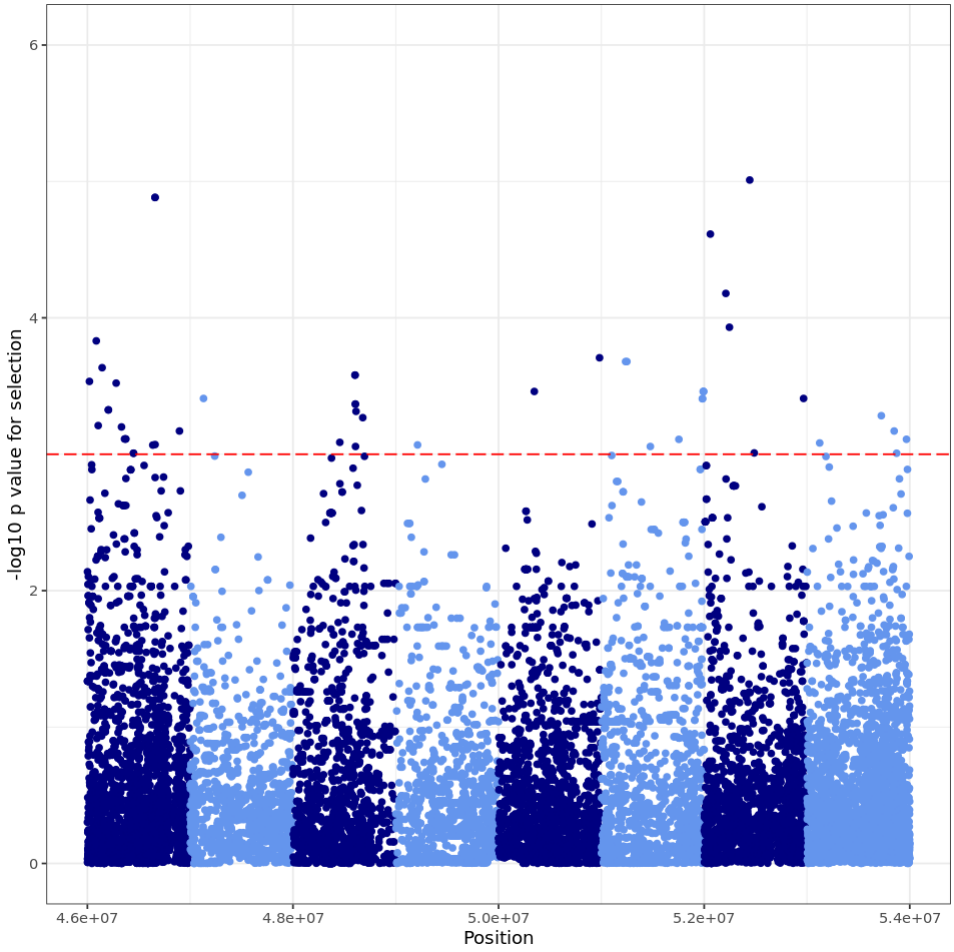
The goal is to identify positive selection in 5 African populations

In [1]:
```
library(dplyr)
library(ggplot2)
```

```
Attaching package: ‘dplyr’


The following objects are masked from ‘package:stats’:

    filter, lag


The following objects are masked from ‘package:base’:

    intersect, setdiff, setequal, union
```

In [40]:
```
relate_YRI <- read.table("relate/YRI/selection_relate_YRI.sele", header=1)

relate_YRI <- relate_YRI %>%
   mutate(in_Range = ((pos >= 4.6e+07 & pos < 4.7e+07)|
                      (pos >= 4.8e+07 & pos < 4.9e+07)|
                      (pos >= 5.0e+07 & pos < 5.1e+07)|
                      (pos >= 5.2e+07 & pos < 5.3e+07)))

ggplot(relate_YRI, aes(x=pos, y=-when_mutation_has_freq2)) +
geom_point(size=1.5, show.legend = FALSE, color=ifelse(relate_YRI$in_Range, "nav
scale_y_continuous(limits = c(0,6)) +
scale_x_continuous(limits = c(46000000, 54000000)) +
labs(x = "Position", y="-log10 p value for selection") +
geom_hline(yintercept = -log10(0.001), linetype = "longdash", color="red") +
theme_bw()
```

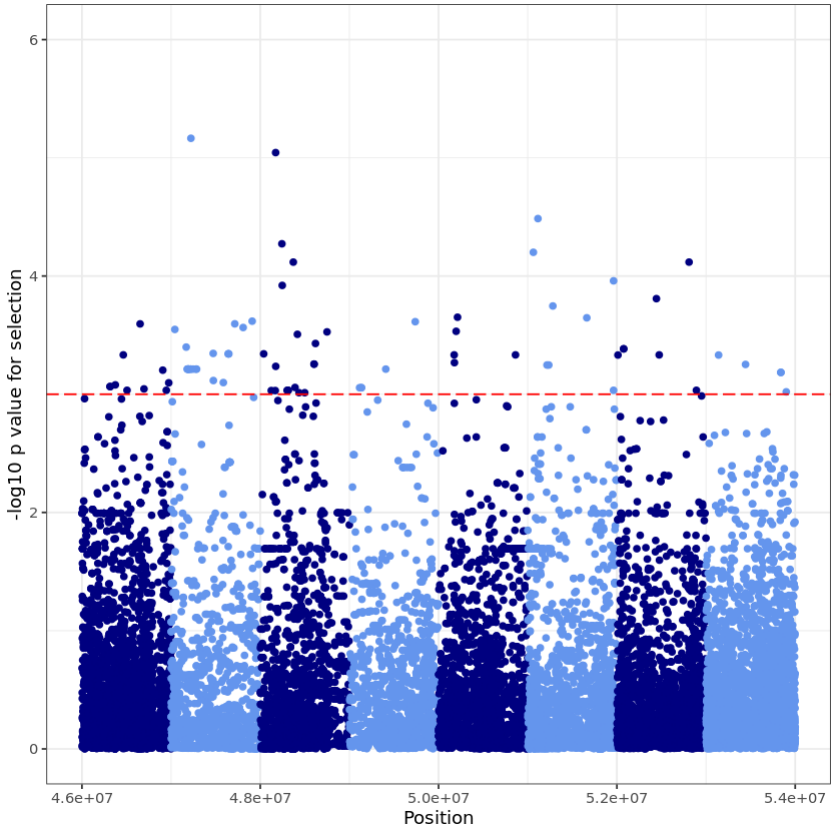
```
Warning message:
“Removed 9461 rows containing missing values (`geom_point()`).”
```



In [3]:

```
relate_LWK <- read.table("relate/LWK/selection_relate_LWK.sele", header=1)

relate_LWK <- relate_LWK %>%
   mutate(in_Range = ((pos >= 4.6e+07 & pos < 4.7e+07)|
                      (pos >= 4.8e+07 & pos < 4.9e+07)|
                      (pos >= 5.0e+07 & pos < 5.1e+07)|
                      (pos >= 5.2e+07 & pos < 5.3e+07)))

ggplot(relate_LWK, aes(x=pos, y=-when_mutation_has_freq2)) +
geom_point(size=1.5, show.legend = FALSE, color=ifelse(relate_LWK$in_Range, "nav
scale_y_continuous(limits = c(0,6)) +
scale_x_continuous(limits = c(46000000, 54000000)) +
labs(x = "Position", y="-log10 p value for selection") +
geom_hline(yintercept = -log10(0.001), linetype = "longdash", color="red") +
theme_bw()
```

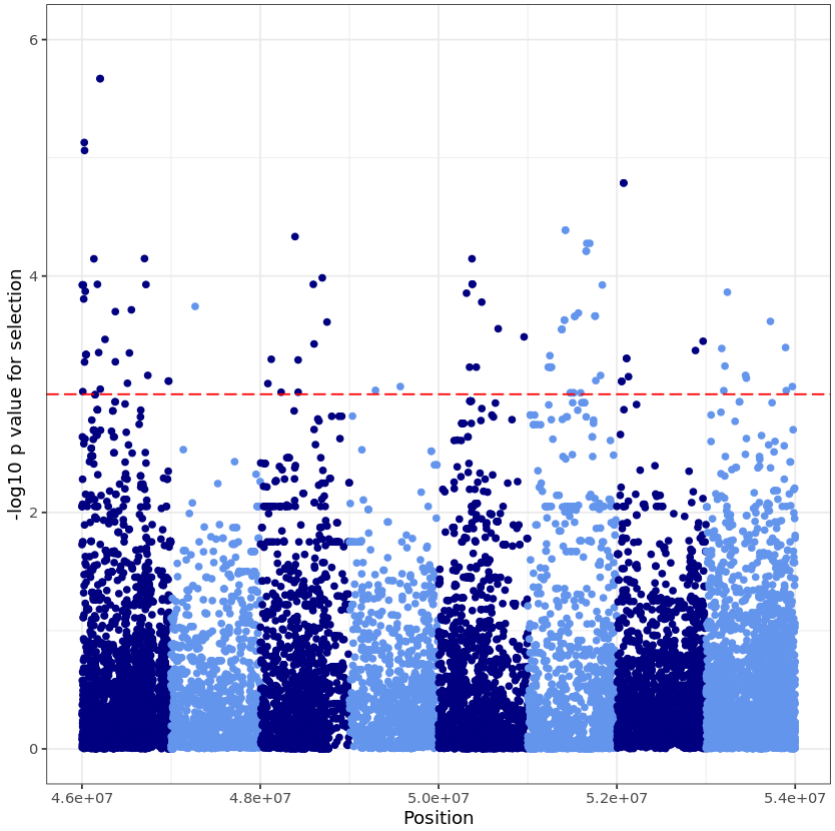
```
Warning message:
“Removed 11983 rows containing missing values (`geom_point()`).”
```



In [4]:

```
relate_GWD <- read.table("relate/GWD/selection_relate_GWD.sele", header=1)

relate_GWD <- relate_GWD %>%
   mutate(in_Range = ((pos >= 4.6e+07 & pos < 4.7e+07)|
                      (pos >= 4.8e+07 & pos < 4.9e+07)|
                      (pos >= 5.0e+07 & pos < 5.1e+07)|
                      (pos >= 5.2e+07 & pos < 5.3e+07)))

ggplot(relate_GWD, aes(x=pos, y=-when_mutation_has_freq2)) +
geom_point(size=1.5, show.legend = FALSE, color=ifelse(relate_GWD$in_Range, "nav
scale_y_continuous(limits = c(0,6)) +
scale_x_continuous(limits = c(46000000, 54000000)) +
labs(x = "Position", y="-log10 p value for selection") +
geom_hline(yintercept = -log10(0.001), linetype = "longdash", color="red") +
theme_bw()
```

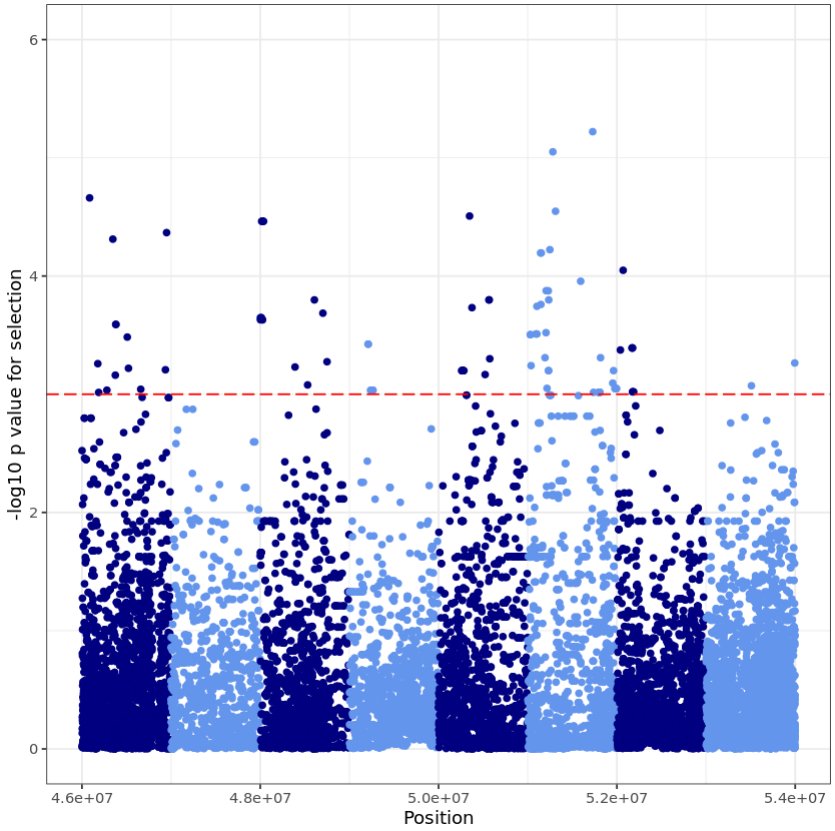
```
Warning message:
“Removed 12031 rows containing missing values (`geom_point()`).”
```



In [6]:

```
relate_MSL <- read.table("relate/MSL/selection_relate_MSL.sele", header=1)

relate_MSL <- relate_MSL %>%
   mutate(in_Range = ((pos >= 4.6e+07 & pos < 4.7e+07)|
                      (pos >= 4.8e+07 & pos < 4.9e+07)|
                      (pos >= 5.0e+07 & pos < 5.1e+07)|
                      (pos >= 5.2e+07 & pos < 5.3e+07)))

ggplot(relate_MSL, aes(x=pos, y=-when_mutation_has_freq2)) +
geom_point(size=1.5, show.legend = FALSE, , color=ifelse(relate_MSL$in_Range, "r
scale_y_continuous(limits = c(0,6)) +
scale_x_continuous(limits = c(46000000, 54000000)) +
labs(x = "Position", y="-log10 p value for selection") +
geom_hline(yintercept = -log10(0.001), linetype = "longdash", color="red") +
theme_bw()
```

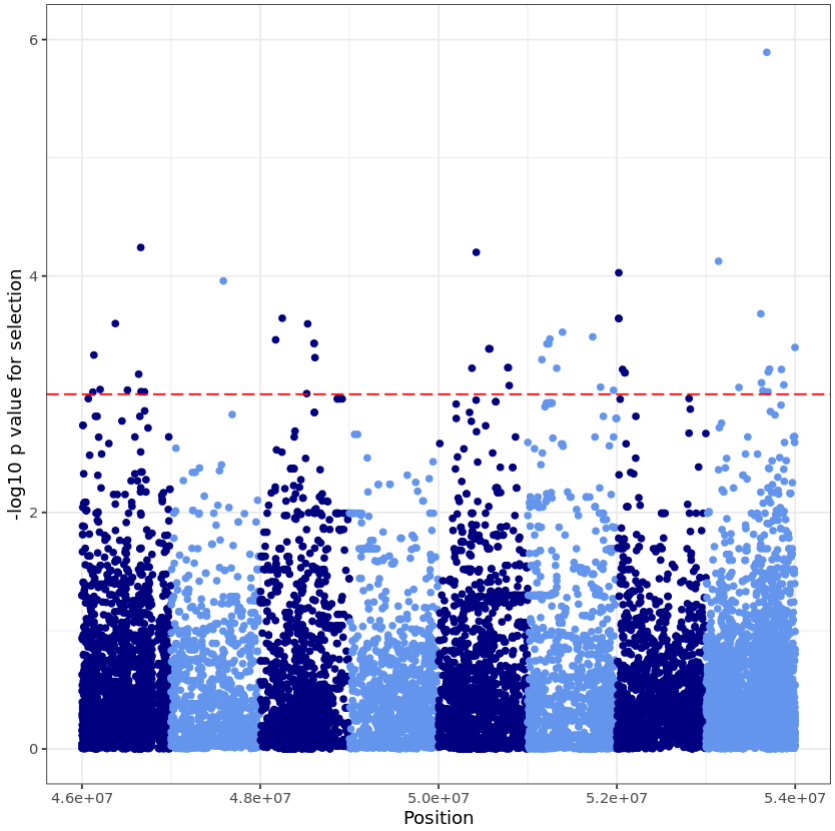
```
Warning message:
“Removed 10168 rows containing missing values (`geom_point()`).”
```



In [7]:

```
relate_ESN <- read.table("relate/ESN/selection_relate_ESN.sele", header=1)

relate_ESN <- relate_ESN %>%
   mutate(in_Range = ((pos >= 4.6e+07 & pos < 4.7e+07)|
                      (pos >= 4.8e+07 & pos < 4.9e+07)|
                      (pos >= 5.0e+07 & pos < 5.1e+07)|
                      (pos >= 5.2e+07 & pos < 5.3e+07)))

ggplot(relate_ESN, aes(x=pos, y=-when_mutation_has_freq2)) +
geom_point(size=1.5, show.legend = FALSE, , color=ifelse(relate_ESN$in_Range, "r
scale_y_continuous(limits = c(0,6)) +
scale_x_continuous(limits = c(46000000, 54000000)) +
labs(x = "Position", y="-log10 p value for selection") +
geom_hline(yintercept = -log10(0.001), linetype = "longdash", color="red") +
theme_bw()
```

```
Warning message:
“Removed 9557 rows containing missing values (`geom_point()`).”
```



In [17]:
```
relate_long <- data.frame(pos = c(relate_YRI$pos, relate_LWK$pos, relate_GWD$pos
                          log10p = c(-relate_YRI$when_mutation_has_freq2, -relate
                          in_Range = c(relate_YRI$in_Range, relate_LWK$in_Range,
                          population = c(rep("YRI", nrow(relate_YRI)), rep("LWK",
                          rs_id = c(relate_YRI$rs_id, relate_LWK$rs_id, relate_GW
```

In [29]:
```
pdf(file="plots/selectionpop.pdf", width = 8, height = 4)
#library(repr)
#options(repr.plot.width = 8, repr.plot.height = 4)
ggplot(relate_long, aes(x=pos, y=log10p)) +
   geom_point(size=0.8, color=ifelse(relate_long$in_Range, "navyblue", "cornflow
   scale_y_continuous(limits = c(0,6)) +
   scale_x_continuous(limits = c(46000000, 54000000)) +
   labs(x = "Position", y="-log10 p value", title = "Selection for populations")
   geom_hline(yintercept = -log10(0.0001), linetype = "longdash", color="red") +
   facet_wrap(~population, ncol=5) +
   theme_bw() +
   theme(axis.text.x = element_text(angle = 90)) +
   theme(axis.title = element_text(size=18), axis.text = element_text(size=12),
dev.off()
```

```
Warning message:
“Removed 53200 rows containing missing values (`geom_point()`).”
```

png: 2

In [31]:
```
above_tres <- relate_long %>%
   filter(log10p >= -log10(0.0001))
```

In [32]:
```
above_tres
```

A data.frame: 52 × 5

pos	log10p	in_Range	population	rs_id
<int>	<dbl>	<lgl>	<chr>	<chr>
46658044	4.88255	TRUE	YRI	rs2139635
46658112	4.88255	TRUE	YRI	rs2139636
52060234	4.61410	TRUE	YRI	rs79046551
52210700	4.17903	TRUE	YRI	rs352152
52443280	5.01049	TRUE	YRI	rs409803
47222433	5.16512	FALSE	LWK	rs113324123
47712083	6.35244	FALSE	LWK	rs561292917
48172687	5.04410	TRUE	LWK	rs7374376
48242938	4.27266	TRUE	LWK	rs7650437
48370580	4.11767	TRUE	LWK	rs7619865
51062970	4.20062	FALSE	LWK	rs4557193
51114357	4.48581	FALSE	LWK	rs13080170
52809525	4.11767	TRUE	LWK	rs2019065
46025992	5.12970	TRUE	GWD	rs56972507
46029398	5.06197	TRUE	GWD	rs72891712
46133965	4.14566	TRUE	GWD	rs1388602
46203832	5.66955	TRUE	GWD	rs9866593
46204315	5.66955	TRUE	GWD	rs41435749
46658737	11.22150	TRUE	GWD	rs9832679
46701599	4.14757	TRUE	GWD	rs6797188
48390016	4.33344	TRUE	GWD	rs12486944
50374568	4.14613	TRUE	GWD	rs57838764
51422188	4.38779	FALSE	GWD	rs1370124
51423110	4.38779	FALSE	GWD	rs73837442
51656370	4.21039	FALSE	GWD	rs9864693
51656376	4.21039	FALSE	GWD	rs9864443
51662407	4.27646	FALSE	GWD	rs9841077
51695865	4.27646	FALSE	GWD	rs4234644
52075944	4.78694	TRUE	GWD	rs7642787
52076019	4.78694	TRUE	GWD	rs7642881
46085854	4.66098	TRUE	MSL	rs116772011
46346640	4.31202	TRUE	MSL	rs6794300
46703270	6.11999	TRUE	MSL	rs78704534
46949379	4.36717	TRUE	MSL	rs7646906

pos	log10p	in_Range	population	rs_id
<int>	<dbl>	<lgl>	<chr>	<chr>
48016172	4.46299	TRUE	MSL	rs9862913
48022573	4.46299	TRUE	MSL	rs7426976
48029118	4.46299	TRUE	MSL	rs1013431
48037078	4.46299	TRUE	MSL	rs7433678
50347595	4.50744	TRUE	MSL	rs76579309
51144316	4.19510	FALSE	MSL	rs9865532
51151799	4.19510	FALSE	MSL	rs6772197
51248198	4.22292	FALSE	MSL	rs4619816
51281664	5.05066	FALSE	MSL	rs9809209
51311849	4.54787	FALSE	MSL	rs1480360
51728573	5.22115	FALSE	MSL	rs13090011
52070033	4.04858	TRUE	MSL	rs7643070
46659395	4.24149	TRUE	ESN	rs1474195
46659530	6.73331	TRUE	ESN	rs1474196
50422718	4.20062	TRUE	ESN	rs2236952
52021092	4.02763	TRUE	ESN	rs323894
53140630	4.12492	FALSE	ESN	rs1004808
53681521	5.89233	FALSE	ESN	rs74373443

In [34]:
```
above_tres %>%
   filter(population == 'YRI')
```

A data.frame: 5 × 5

pos	log10p	in_Range	population	rs_id
<int>	<dbl>	<lgl>	<chr>	<chr>
46658044	4.88255	TRUE	YRI	rs2139635
46658112	4.88255	TRUE	YRI	rs2139636
52060234	4.61410	TRUE	YRI	rs79046551
52210700	4.17903	TRUE	YRI	rs352152
52443280	5.01049	TRUE	YRI	rs409803

In [35]:
```
above_tres %>%
   filter(population == 'LWK')
```

A data.frame: 8 × 5

pos	log10p	in_Range	population	rs_id
<int>	<dbl>	<lgl>	<chr>	<chr>
47222433	5.16512	FALSE	LWK	rs113324123
47712083	6.35244	FALSE	LWK	rs561292917
48172687	5.04410	TRUE	LWK	rs7374376
48242938	4.27266	TRUE	LWK	rs7650437
48370580	4.11767	TRUE	LWK	rs7619865
51062970	4.20062	FALSE	LWK	rs4557193
51114357	4.48581	FALSE	LWK	rs13080170
52809525	4.11767	TRUE	LWK	rs2019065

In [36]:
```
above_tres %>%
   filter(population == 'GWD')
```

A data.frame: 17 × 5

pos	log10p	in_Range	population	rs_id
<int>	<dbl>	<lgl>	<chr>	<chr>
46025992	5.12970	TRUE	GWD	rs56972507
46029398	5.06197	TRUE	GWD	rs72891712
46133965	4.14566	TRUE	GWD	rs1388602
46203832	5.66955	TRUE	GWD	rs9866593
46204315	5.66955	TRUE	GWD	rs41435749
46658737	11.22150	TRUE	GWD	rs9832679
46701599	4.14757	TRUE	GWD	rs6797188
48390016	4.33344	TRUE	GWD	rs12486944
50374568	4.14613	TRUE	GWD	rs57838764
51422188	4.38779	FALSE	GWD	rs1370124
51423110	4.38779	FALSE	GWD	rs73837442
51656370	4.21039	FALSE	GWD	rs9864693
51656376	4.21039	FALSE	GWD	rs9864443
51662407	4.27646	FALSE	GWD	rs9841077
51695865	4.27646	FALSE	GWD	rs4234644
52075944	4.78694	TRUE	GWD	rs7642787
52076019	4.78694	TRUE	GWD	rs7642881

In [38]:
```
above_tres %>%
   filter(population == 'MSL')
```

A data.frame: 16 × 5

pos	log10p	in_Range	population	rs_id
<int>	<dbl>	<lgl>	<chr>	<chr>
46085854	4.66098	TRUE	MSL	rs116772011
46346640	4.31202	TRUE	MSL	rs6794300
46703270	6.11999	TRUE	MSL	rs78704534
46949379	4.36717	TRUE	MSL	rs7646906
48016172	4.46299	TRUE	MSL	rs9862913
48022573	4.46299	TRUE	MSL	rs7426976
48029118	4.46299	TRUE	MSL	rs1013431
48037078	4.46299	TRUE	MSL	rs7433678
50347595	4.50744	TRUE	MSL	rs76579309
51144316	4.19510	FALSE	MSL	rs9865532
51151799	4.19510	FALSE	MSL	rs6772197
51248198	4.22292	FALSE	MSL	rs4619816
51281664	5.05066	FALSE	MSL	rs9809209
51311849	4.54787	FALSE	MSL	rs1480360
51728573	5.22115	FALSE	MSL	rs13090011
52070033	4.04858	TRUE	MSL	rs7643070

In [39]:
```
above_tres %>%
   filter(population == 'ESN')
```

A data.frame: 6 × 5

pos	log10p	in_Range	population	rs_id
<int>	<dbl>	<lgl>	<chr>	<chr>
46659395	4.24149	TRUE	ESN	rs1474195
46659530	6.73331	TRUE	ESN	rs1474196
50422718	4.20062	TRUE	ESN	rs2236952
52021092	4.02763	TRUE	ESN	rs323894
53140630	4.12492	FALSE	ESN	rs1004808
53681521	5.89233	FALSE	ESN	rs74373443