Chi Squared Tests for Heterogeneity

Rachael 'Rocky' Aikens, Voight Lab

June 20, 2017

Set Up

This analysis requires data on the counts of each 3mer, 5mer, and 7mer polymorphism type private to each of the 1,000 genomes nonadmixed continental groups (AFR, EUR, EAS, and SAS). All of this data has been preprocessed using the process_chrom_counts function in "code/data_wrangling/process_chrom_counts.R". I call these "count dataframes." An example is shown here:

```
## [1] "/Users/raikens/Documents/research/voightlab/mutation_rate/analysis/data"
```

```
##    Context  X3mer X1mer Count         Rate context_in_genome chr1 chr2
## 1 AAAAA->C AAA->C  A->C 13912 1.884656e-09          16034992 1110 1220
## 2 AAAAA->G AAA->G  A->G 22852 3.085974e-09          16034992 1824 1966
## 3 AAAAA->T AAA->T  A->T  9107 1.240046e-09          16034992  723  762
## 4 AAAAC->C AAA->C  A->C  3625 1.519035e-09           5188115  277  294
## 5 AAAAC->G AAA->G  A->G  4450 1.870248e-09           5188115  317  426
```

Pairwise Chi Squared Tests

This section details how to perform the pairwise chi squared tests from Harris 2015. These steps were mostly used for replication.

Methodology

Two R functions that I use in this analysis:

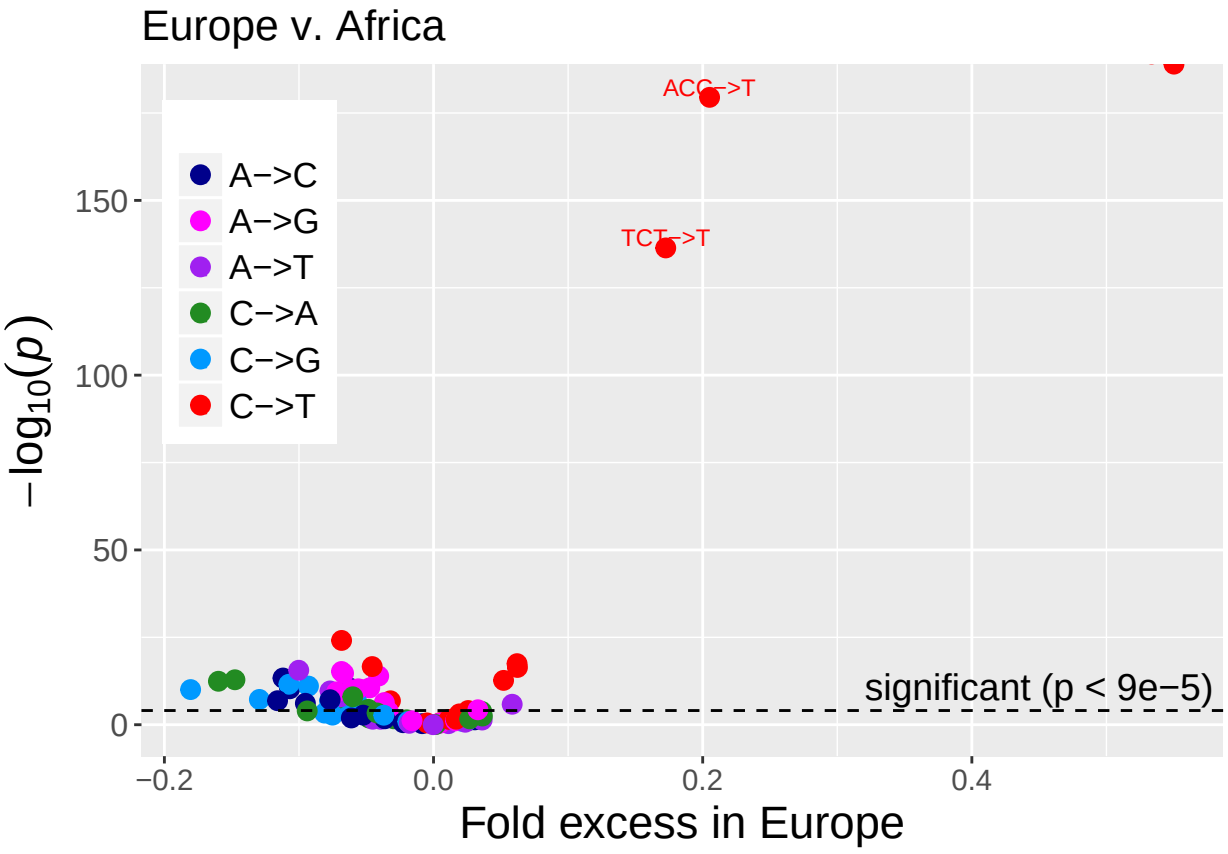
- **pairwise.chi** Given two count dataframes, output a dataframe of chi-squared test results for each context. The arguement 'filter' (set by default to be true), will output "NA" as the p-value for any test for which the chi squared assumptions may not be correct.
- **volcano.plot** Given two count dataframes, plus the output from pairwise.chi, construct a volcano plot as in Harris 2015. Also takes the arguement lab.lim, which determines the lower p-value limit for which polymorphisms types should be labeled.

The code used to definte pairwise.chi is shown below:

```
# calculates homogeneity test p values for pairwise comparisons of two dfs of counts
pairwise.chi <- function(counts.1, counts.2, filter = T){
  n.contexts = length(counts.1$Context)
  result <- data.frame(matrix(ncol=2,nrow=n.contexts))
  colnames(result) <- c("Context", "p")

  result$Context <- counts.1$Context
  sum.1 <- sum(counts.1$Count)
  sum.2 <- sum(counts.2$Count)

  for (i in 1:n.contexts){
    c.a <- c(counts.1$Count[i], counts.2$Count[i])
    c.b <- c(sum.1, sum.2) - c.a
    data <- cbind(c.a, c.b)
    warning <- is(tryCatch(chisq.test(data), warning = function(w) w), "warning")
    if (filter == T & warning){
      result$p[i] <- NA
    }
    else result$p[i] <- chisq.test(data)$p.value
  }

  return(result)
}
```

Examples

As previously mentioned, these methods are not central to my analysis and mostly important for replication. It's worthwhile to note that these functions and data replicate volcano plots like the ones from Kelley Harris's leading figure in PNAS 2015:



Fourway Tests for Homogeneity

In order to lighten the multiple hypothesis testing burden of running $\binom{4}{2} = 6$ pairwise comparisons for each possible polymorphism type, we switched to a homogeneity testing framework, which helps us rank polymorphism types based on how much they vary between populations. This is the dominant analysis technique we use to identify polymorphisms which are heterogeneous across continental groups.

Methodology

This section defines one function for calculation and two for visualization. Again, I'm hiding the code for the plotting functions in the compiled report because it's not essential to understanding.

- **fourway.chi** Given four count dataframes, output a dataframe of chi-squared test results for each context.
- **hom.test.plot** Given the output from fourway.chi, construct a volcano plot as in Harris 2015. Also takes the arguement lab.lim, which determines the lower p-value limit for which polymorphisms types should be labeled, and the boolean, NoGGA, which, when True, leaves out labels for any polymorphism with the 3mer subcontext GGA->A.
- **sigs.plot** Given the same arguments as hom.test.plot, make a plot of just the significant results.

The r code used to define fourway.chi is shown below:

```
# calculates homogeneity test p values for Fourway comparisons of counts dfs
fourway.chi <- function(AFR, EUR, EAS, SAS, filter = T){
  n.contexts = length(AFR$Context)

  # make dataframe for results
  result <- data.frame(matrix(ncol=9,nrow=n.contexts))
  colnames(result) <- c("Context", "X5mer","X3mer", "X1mer",
                        "AFR.Count", "EUR.Count", "EAS.Count", "SAS.Count", "p")
  result$Context <- AFR$Context
  result$X5mer <- AFR$X5mer # for smaller contexts, X3mer and X5mer columns do not exist,
  result$X3mer <- AFR$X3mer # and will disappear at this step
  result$X1mer <- AFR$X1mer
  result$AFR.Count <- AFR$Count; result$EUR.Count <- EUR$Count
  result$EAS.Count <- EAS$Count; result$SAS.Count <- SAS$Count

  # start setting up tables
  sums <- c(sum(AFR$Count), sum(EUR$Count), sum(EAS$Count), sum(SAS$Count))

  # set up table and run test for each context
  for (i in 1:n.contexts){
    c.a <- c(AFR$Count[i], EUR$Count[i], EAS$Count[i], SAS$Count[i])
    c.b <- sums - c.a
    data <- cbind(c.a, c.b)
    warning <- is(tryCatch(chisq.test(data), warning = function(w) w), "warning")
    if (filter == T & warning){
      result$p[i] <- NA}
    else result$p[i] <- chisq.test(data)$p.value
    }
  return(result)
}
```

3mers

Now using these functions, we can run the following tests for 3mer polymorphism types which are heterogeneous across ancestral groups. We can begin with the 3mer context paradigm, which is most commonly used in the literature in this area. The table below shows all 3mer polymorphism types, ranked according to their p value in a fourway.chi test for heterogeneity across populations.

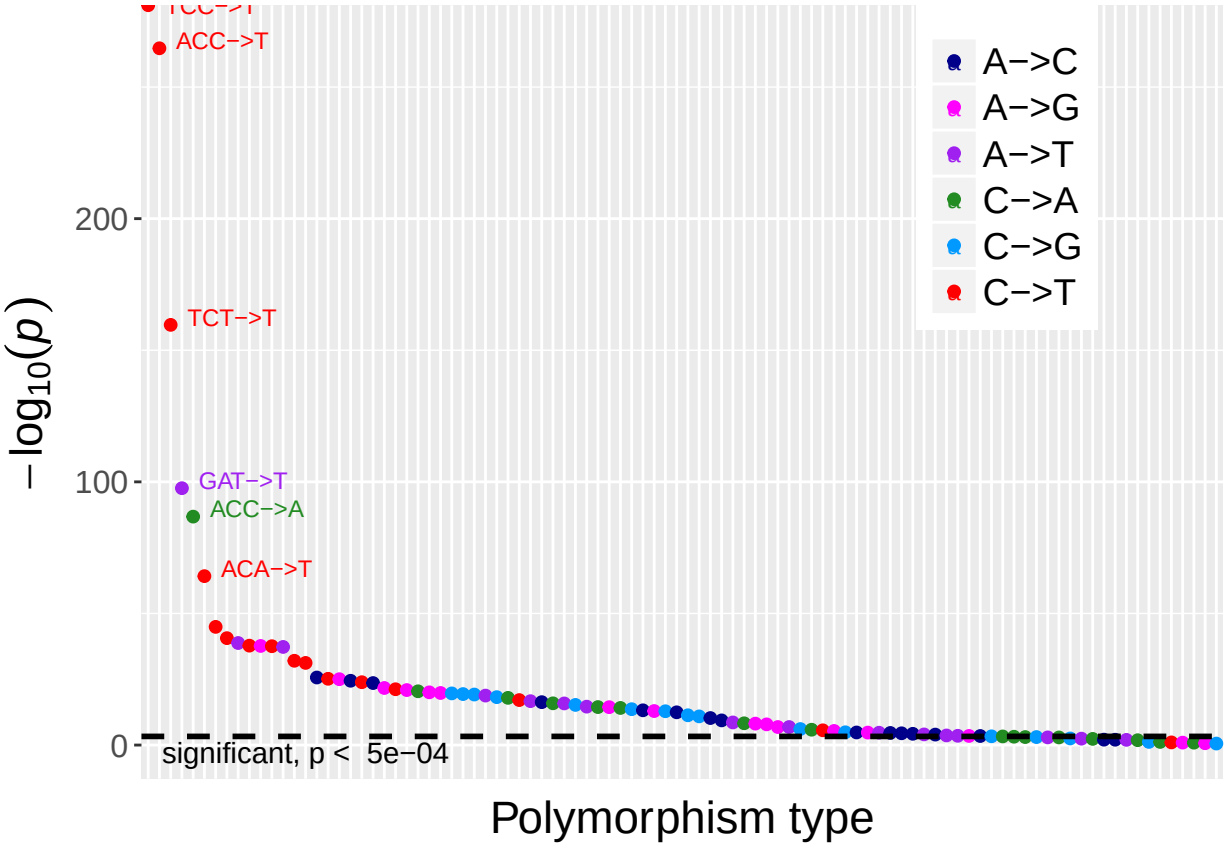


Table 1: 10 most significant 3mer polymorphisms

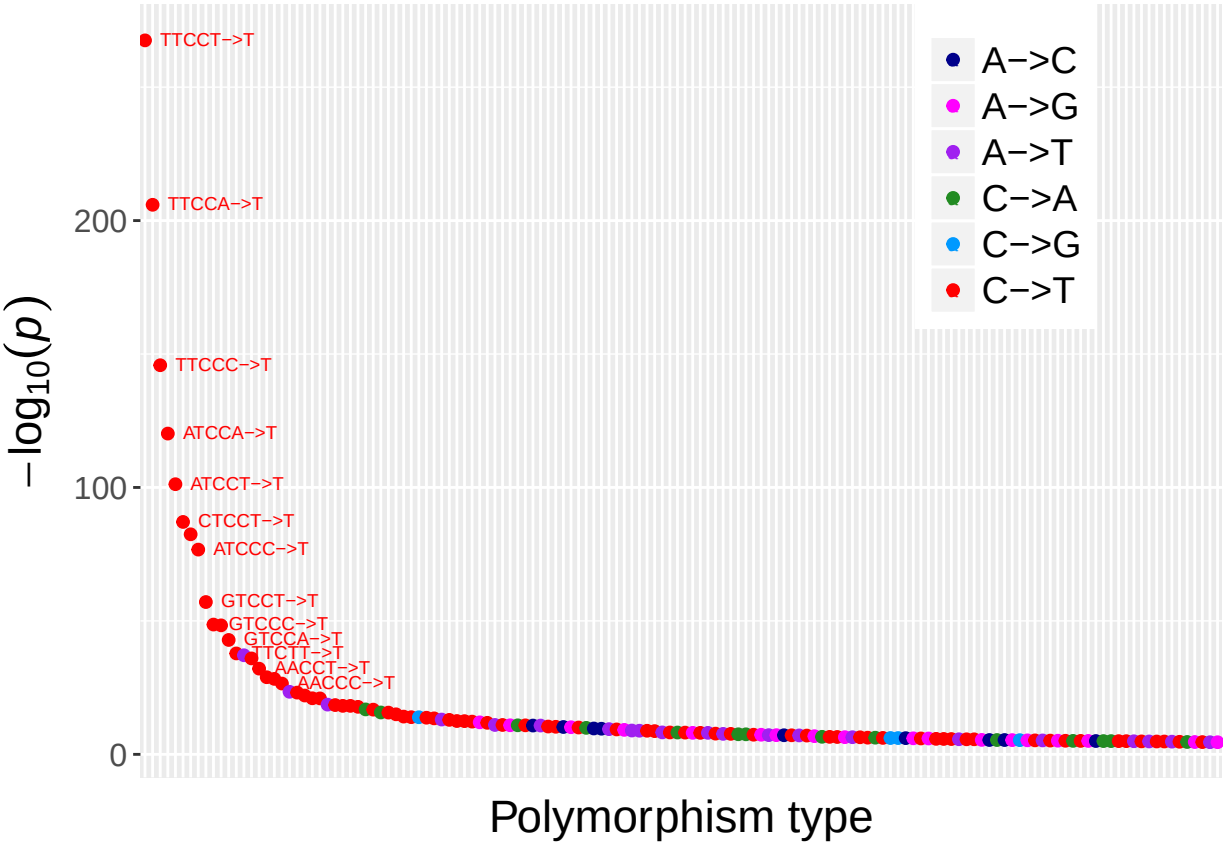
Context	X1mer	AFR.Count	EUR.Count	EAS.Count	SAS.Count	p
TCC->T	C->T	122497	34785	34204	41260	0.000000e+00
ACC->T	C->T	127253	28092	32997	38350	1.828424e-265
TCT->T	C->T	135528	29110	37044	40348	2.250941e-160
GAT->T	A->T	44242	8578	14839	14021	2.547968e-98
ACC->A	C->A	72168	13693	23008	22332	1.613236e-87
ACA->T	C->T	214890	37570	55425	58567	6.646004e-65
CCC->T	C->T	125730	24461	33148	36184	1.258539e-45
ACG->T	C->T	313789	57783	90415	91917	2.404467e-41
TAA->T	A->T	42297	6972	11040	10458	1.696467e-39
TCG->T	C->T	192028	35851	55422	57105	1.709131e-38

This plot highlights the top six contexts, which are significant at p<1e-60. They include GGA->A, ACC->T, and AGA->A, which have been previously reported as part of a European signal of C->T elevation. The next three contexts have not been noted by any previous analyses of mutation rate heterogeneity. There are 76 significant polymorphisms falling out from this analysis after Bonferroni correction.

5mers

Now we move to higher levels of sequence context, which may capture more detail in how mutation rates vary. In this section, we run the same analysis as above for 5mers, identifying variable polymorphism types which may not have been highlighted at the 3mer level.

The plot below shows the homogeneity test p values for just the 142 5mers which are significant after bonferroni correction.



It is clear from this plot (and the one for 7mers) that many of the signficant polymorphisms at the 5mer and 7mer level are a part of the signal of C->T elevation that we observe at the 3mer level. This begs the question: how many significant 5mer signals are there outside of the 3mer subcontexts we have already idenified? To answer this question, I removed from the significant 5mer set all mutations whose 3mer subcontexts correspond to the European C->T elevation (GGA->A, ACC->T, AGA->A, and CCC->T), or highlighted by the three additional highly significant variable polymorphisms (ATC->A, ACC->A, ACA->T) identified in the previous section. This leaves a total of 110 new significant polymorphisms. The following table shows the most highly significant 5mers *outside* of these 3mer signals that have already been noted:

Table 2: 10 most significant 5mers not noted on a 3mer level

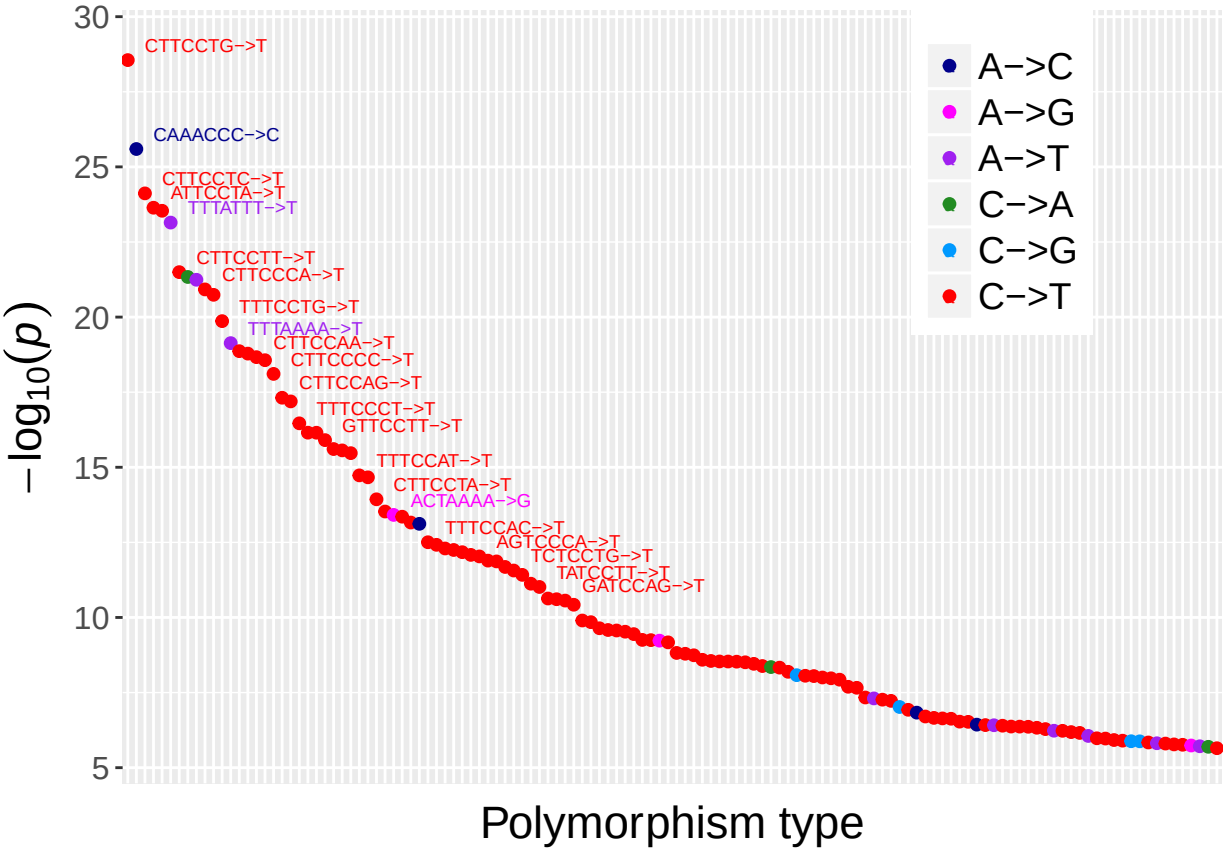
Context	AFR.Count	EUR.Count	EAS.Count	SAS.Count	p
TTCCT->T	16613	5162	4634	5754	3.009513e-268
TTCCA->T	10564	3418	2941	3774	1.205919e-206
TTCCC->T	12276	3585	3496	4217	1.587023e-146
ATCCA->T	7197	2234	1910	2489	6.414071e-121
ATCCT->T	8767	2546	2435	2951	6.029380e-102
CTCCT->T	16097	4121	4496	5113	8.310700e-88
CTCCA->T	9362	2557	2640	3235	3.553694e-83

Context	AFR.Count	EUR.Count	EAS.Count	SAS.Count	p
ATCCC->T	8058	2248	2256	2736	1.892591e-77
GTCCT->T	6567	1824	1839	2110	8.474776e-58
GTCCC->T	6151	1651	1709	2090	2.345134e-49

Note that the most highly significant new 5mer is TTAAA->T, which corresponds to the 8th most significant 3mer, TAA->T. As we will see, the 7mer TTTAAAA->T is also one of the top significantly variable 7mers.

7mers

Now we move to the same analysis at the 7mer level, beginning with a homogeneity test plot:



The plot above shows heterogeneity test p values for the 128 7mers signifcant after bonferroni correction. We can ask the same question about these results as we did with the 5mers: which of these 7mers are results that we have not previously picked out from our 3mer analysis? Filtering these signals leaves 118 significant results, the top ten of which are shown below:

Table 3: 10 most significant 7mers not noted on a 3mer level

Context	AFR.Count	EUR.Count	EAS.Count	SAS.Count	p
CTTCCTG->T	1284	430	365	482	2.783932e-29
CAAACCC->C	120	17	101	11	2.540623e-26
CTTCCTC->T	1250	410	343	380	7.617262e-25
ATTCCTA->T	620	242	171	230	2.286859e-24
CTTCCAT->T	652	257	230	235	2.887277e-24
TTTATTT->T	2578	387	743	438	7.149845e-24
CTTCCTT->T	1771	519	474	594	3.187821e-22
AAACAAA->A	2866	394	709	520	4.619710e-22
ATTAAAA->T	3545	456	791	750	5.718279e-22
CTTCCCA->T	1117	365	336	404	1.200361e-21

Summary

The following table summarizes the numbers of significant results from this section.

Context Model	Number Significant	Number New
3mer	76	–
5mer	142	110
7mer	128	118

False Discovery Rate Corrections

All of the tests in the above section use the Bonferroni Correction, which is conservative even when hypothesis tests are positively correlated (as is most-likely the case here.) However, the Bonferroni correction is often criticized as being *too* conservative. For these reasons, it may be useful to apply other significance thresholds which account for the multiple testing burden.

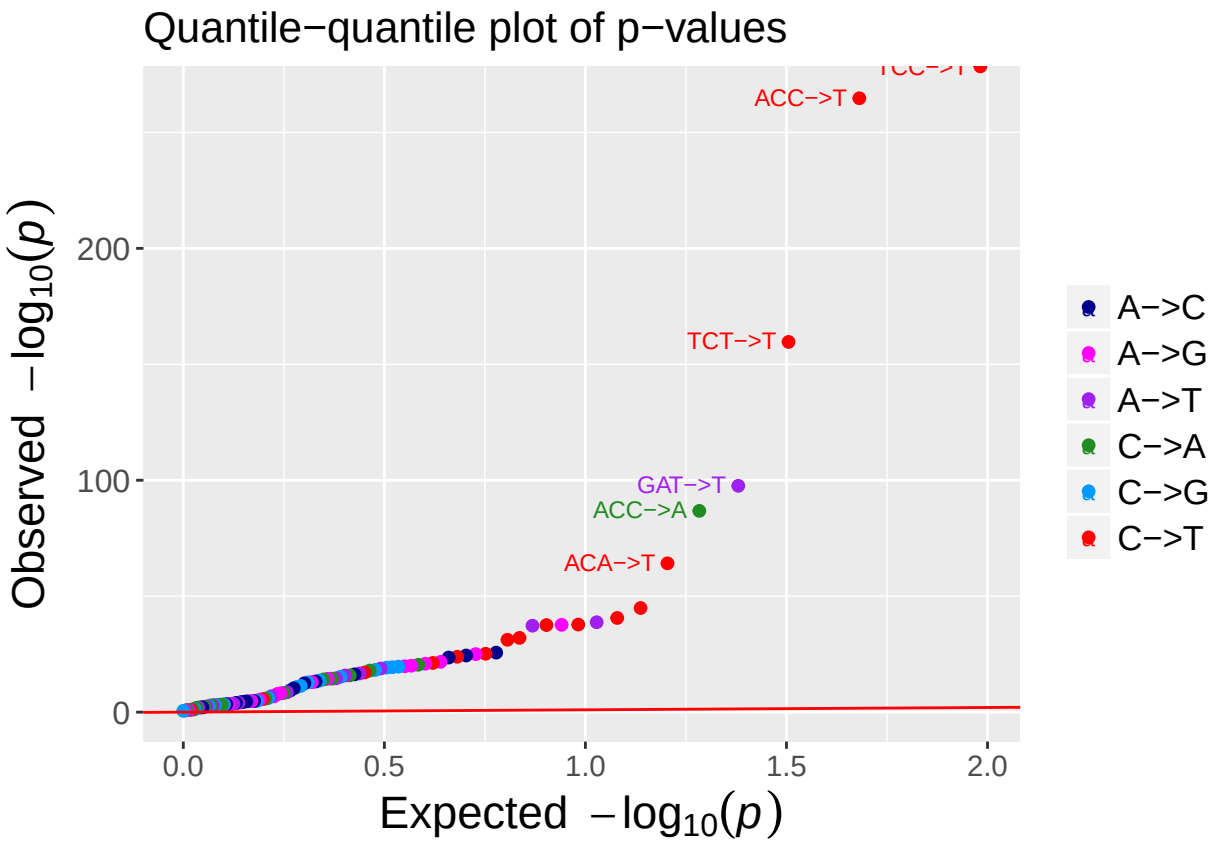
Methodology

Initially, I tried to use the qvalue package to perform false discovery rate analysis. However, this package proved difficult to use, since our p-values from our homogeneity tests don't follow a uniform [0,1] distribution (they range from 0-0.45). Instead, I decided to use the built-in R function, p.adjust(), which uses Benjamini-Hoochberg-Yekutieli. These methods should be acceptable even when the p-values are positively correlated. The following function, **fdr**, performs simple fdr analysis on an output dataframe from a chi-squared function.

```
fdr <- function(p.data){
  p.data <- p.data[complete.cases(p.data),]

  # This uses Benjamini-Hochberg-Yekutieli fdr
  p.data$fdr <- p.adjust(p.data$p, method = "fdr")

  # multiple hypothesis correction by holm
  p.data$holm <- p.adjust(p.data$p, method = "holm")

  alpha = 0.05/length(p.data$p)
  p.data <- p.data[complete.cases(p.data), ]

  n.sig <- c(length(p.data$p), sum(p.data$p < alpha), sum(p.data$holm< 0.05),
             sum(p.data$fdr< 0.1), sum(p.data$fdr< 0.05),
             sum(p.data$fdr< 0.01), sum(p.data$fdr< 0.001))

  names(n.sig) <- c("Total tests","Bonferroni", "Holm",
                    "FDR<0.1", "FDR<0.05",
                    "FDR<0.01", "FDR<0.001")

  return(list(n.sig, p.data))
}
```

I am additionally defining the funtion **qq.labels**, which takes in a p-value dataframe, a lab.lim, a title (default = "Quantile-quantile plot of p-values"), and the NoGGA arguement and returns a qq plot of all contexts, color-coded and labeled. In the following section, I will construct qq plots and run fdr analysis for each of the 3mer, 5mer, and 7mer models.

3mers

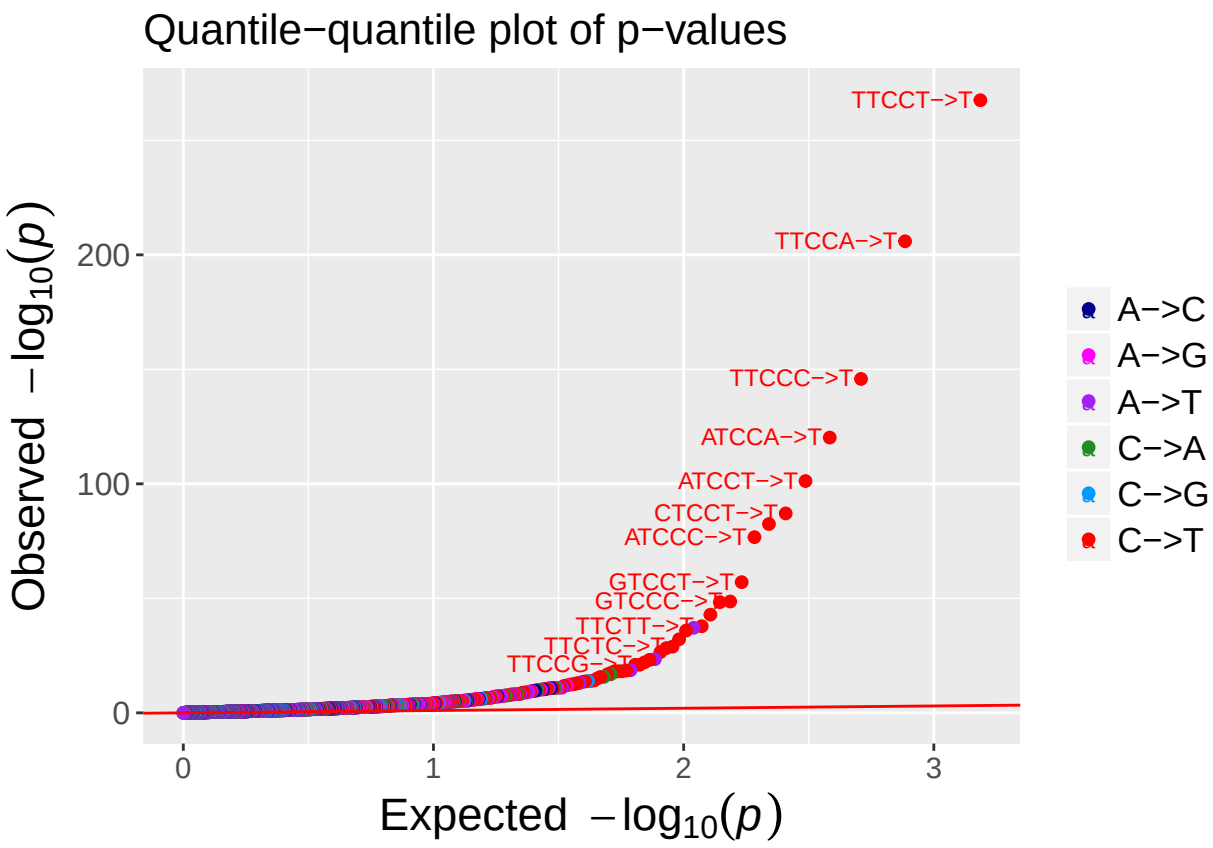



```
## [[1]]
## Total tests  Bonferroni        Holm     FDR<0.1    FDR<0.05    FDR<0.01
##          96          76          84          91          89          86
##   FDR<0.001
##          78
```

The qq plots shown above display relatively the same information as the p-value plots by context. However, it is worth noting that the observed p values, even at the lower end, are above expected p-value quantiles. This may suggest that in fact, every context is significant so that the null distribution of p values does not hold. More realistically, this appears to be an artifact of the fact that hypothesis tests set up as above are actually positively correlated (that is, a small p-value in one test probably increases the likelihood of a small p-value in another test).

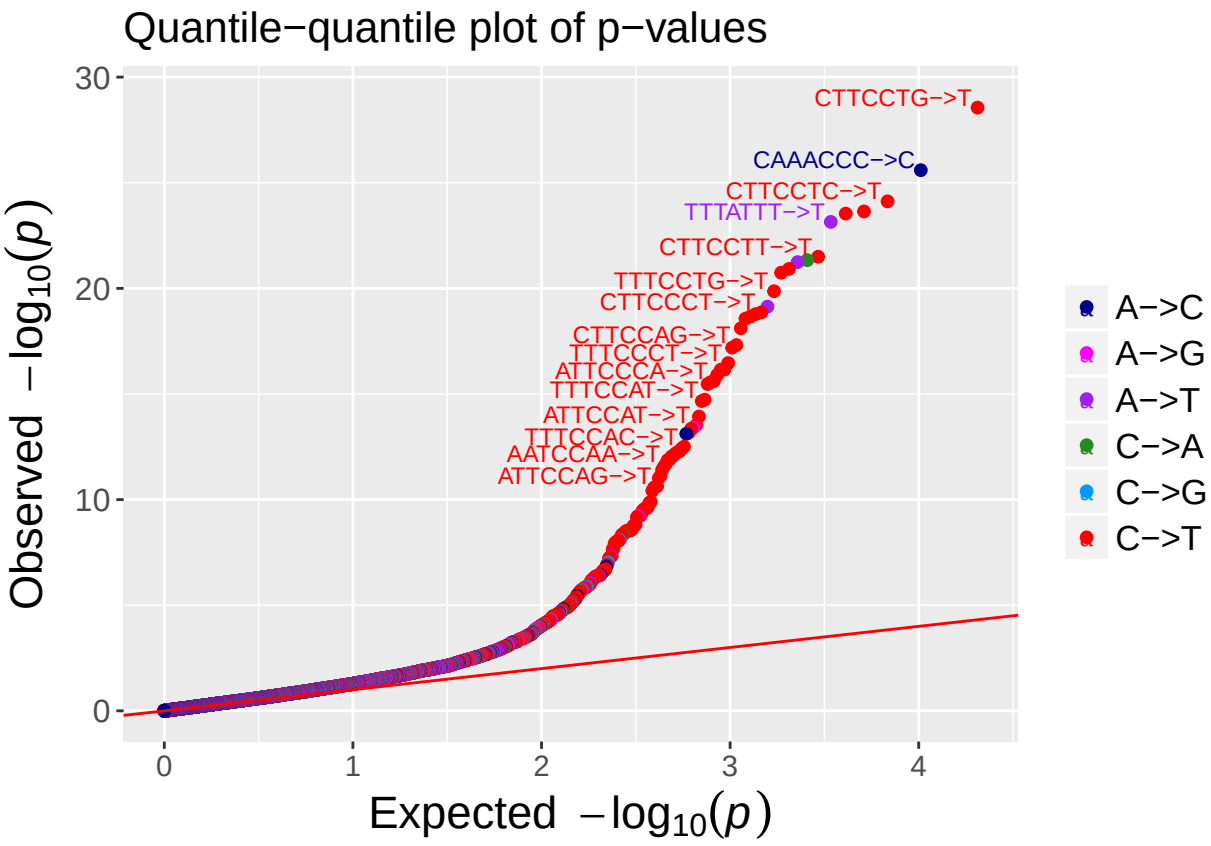
One possible solution to this problem would be to simulate null-distributed datasets to approximate an emprical distribution for expected p-value.

5mers




```
## [[1]]
## Total tests  Bonferroni        Holm     FDR<0.1    FDR<0.05    FDR<0.01
##        1535         142         143         488         393         256
##   FDR<0.001
##         157
```

7mers




```
## [[1]]
## Total tests  Bonferroni        Holm     FDR<0.1    FDR<0.05    FDR<0.01
##       20513         128         128         392         303         208
##   FDR<0.001
##         140
```

Ordered p-values

Methodology

The following function, **ordered.p**, returns a p-value dataframe with p calculated based on the methods from Harris and Pritchard, 2017. This method is proven to give less-significant results, but helps partially combat the problem of positive correlation between p values using our original methods.

```
ordered.p <- function(pdata){
  #preprocess data to order and remove nas
  pdata <- pdata[complete.cases(pdata$p),]
  myorder <- order(pdata$p)
  n.muts <- length(pdata$p)

  p.ordered <- rep(0, n.muts)

  #set largest p-value
  j <- myorder[n.muts]
  p.ordered[j] <- pdata$p[j]

  #initialize not mutated counts based on this lowest p-value mutation
  not.mut <- c(pdata$AFR.Count[j], pdata$EUR.Count[j],
               pdata$EAS.Count[j], pdata$SAS.Count[j])

  for (i in n.muts:1){
    j <- myorder[i]
    mut <-c(pdata$AFR.Count[j], pdata$EUR.Count[j],
            pdata$EAS.Count[j], pdata$SAS.Count[j])
    data <- cbind(mut, not.mut)
    p.ordered[j] <- chisq.test(data)$p.value

    #add these mutations to the not.mutated counts for future tests
    not.mut <- not.mut + mut
  }
  pdata$p <- p.ordered
  return(pdata)
}
```

In the following sections, I will repeat all of the above analyses for fourway homogeneity test in terms of ordered p value.

3mers

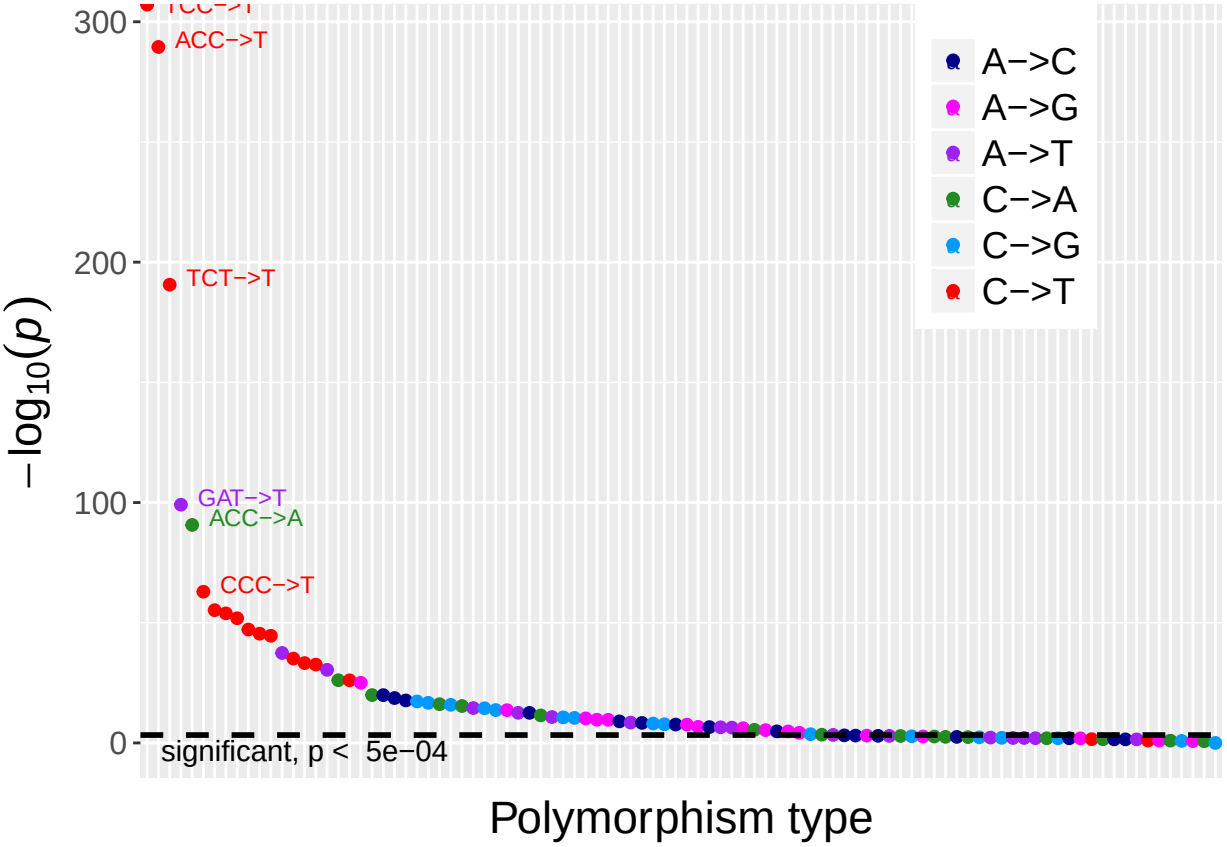
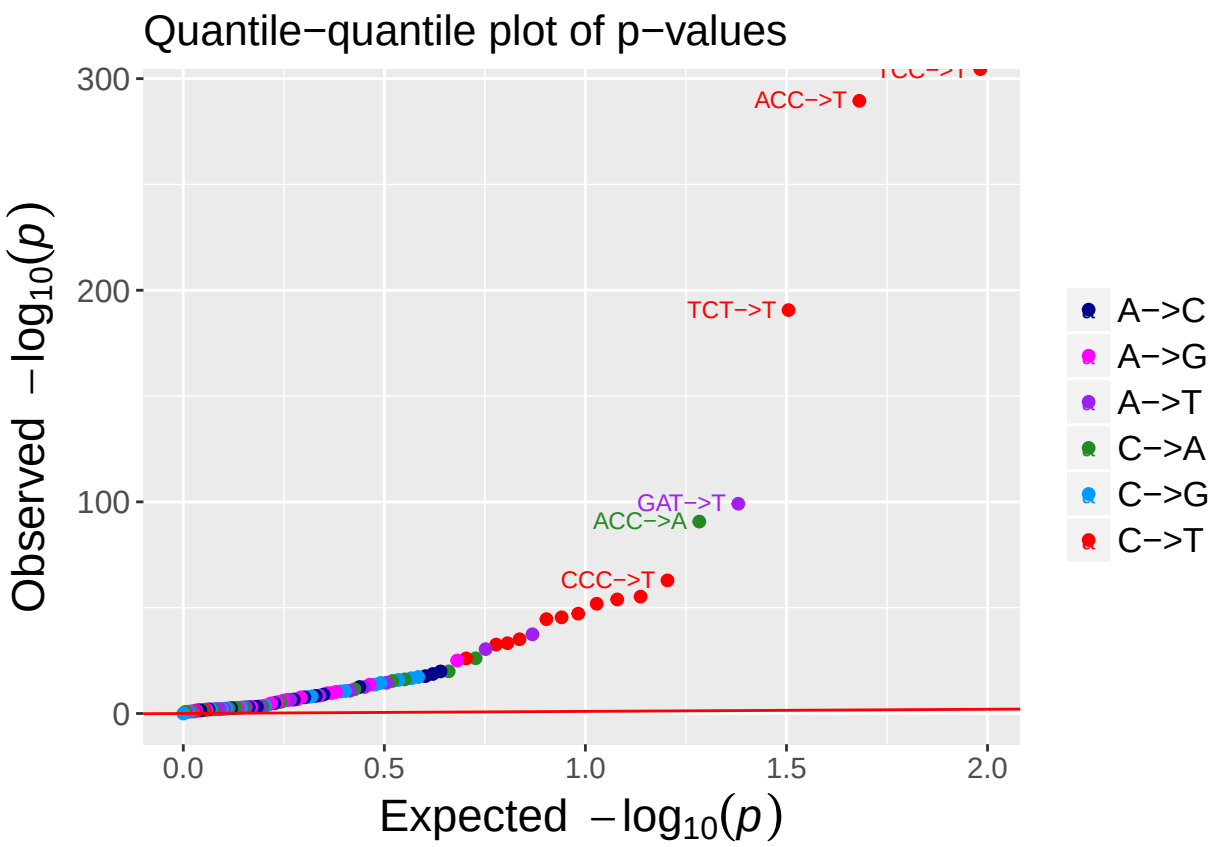


Table 5: 10 most significant 3mers using ordered p value correction

Context	X1mer	AFR.Count	EUR.Count	EAS.Count	SAS.Count	p
TCC->T	C->T	122497	34785	34204	41260	0.000000e+00
ACC->T	C->T	127253	28092	32997	38350	3.046697e-290
TCT->T	C->T	135528	29110	37044	40348	2.354790e-191
GAT->T	A->T	44242	8578	14839	14021	7.787616e-100
ACC->A	C->A	72168	13693	23008	22332	2.051216e-91
CCC->T	C->T	125730	24461	33148	36184	1.223057e-63
ACA->T	C->T	214890	37570	55425	58567	5.917902e-56
TCG->T	C->T	192028	35851	55422	57105	1.245725e-54
ACG->T	C->T	313789	57783	90415	91917	1.357347e-52
ACT->T	C->T	148193	27844	38863	41783	6.625703e-48

Notice that, as before, ATC->A and ACC->A are the 4th and 5th most significant results. Meanwhile, ACA->T, the third highly significant signal from earlier, is moved from the 6th to the 9th place in terms of significance, and no longer sticks out from the remaining mutation types as it once did. Certain mutations (for example TAA->T) have dropped in significance notably (from 8th to 17th), while C->T mutations seem to be featured much more prominently among the most significant polymorphism types.




```
## [[1]]
## Total tests  Bonferroni        Holm     FDR<0.1    FDR<0.05    FDR<0.01
##          96          62          69          89          89          77
##   FDR<0.001
##          62
```

5mers

```
## [1] 146
## [1] 113
```

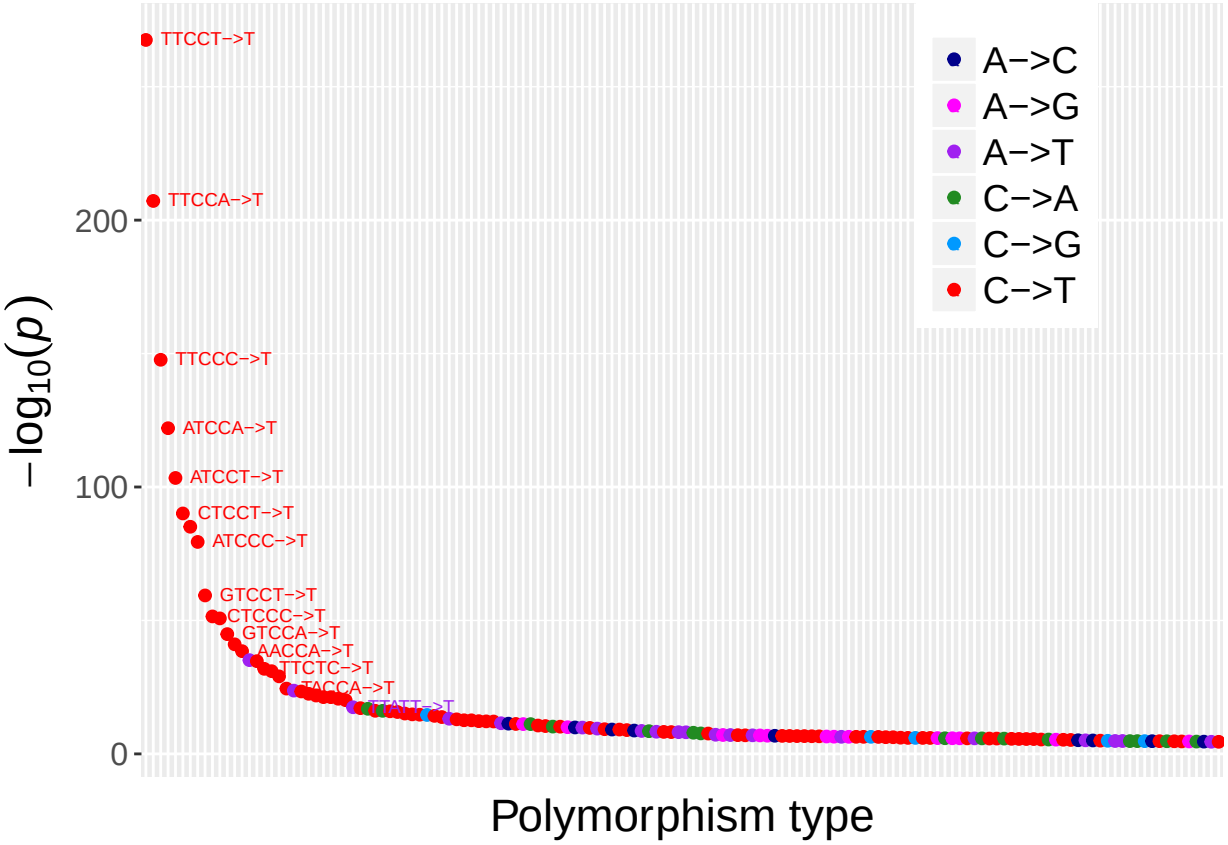
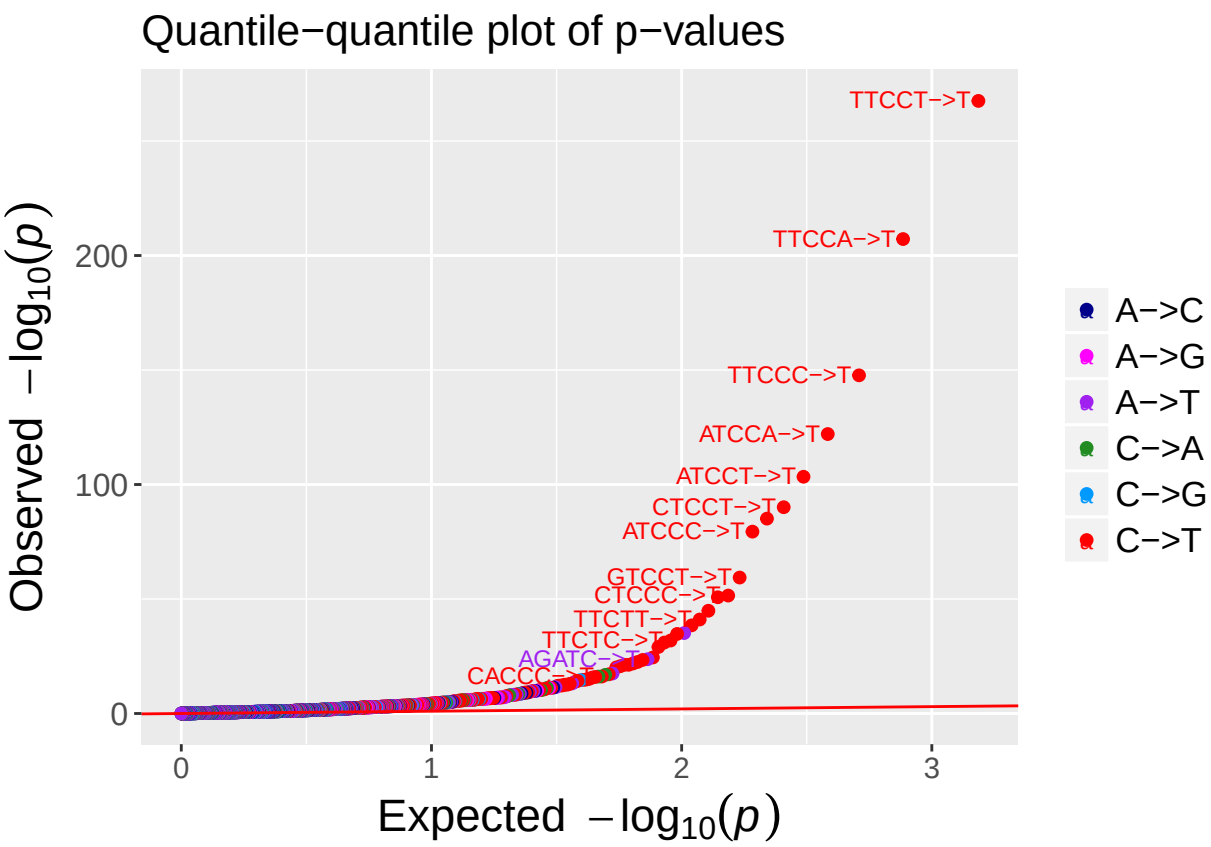


Table 6: 10 most significant new 5mers using ordered p value correction

Context	AFR.Count	EUR.Count	EAS.Count	SAS.Count	p
TTCCT->T	16613	5162	4634	5754	3.024394e-268
TTCCA->T	10564	3418	2941	3774	6.441458e-208
TTCCC->T	12276	3585	3496	4217	2.095740e-148
ATCCA->T	7197	2234	1910	2489	9.144227e-123
ATCCT->T	8767	2546	2435	2951	4.074418e-104
CTCCT->T	16097	4121	4496	5113	8.156315e-91
CTCCA->T	9362	2557	2640	3235	7.805733e-86
ATCCC->T	8058	2248	2256	2736	3.802285e-80
GTCCT->T	6567	1824	1839	2110	4.275523e-60
CTCCC->T	13171	3189	3722	4191	3.119867e-52

Notice that, as before, TTAAA->T is the most signifcant new 5mer, with a p-value several orders of magnitude smaller than the other new significant 5mers. Again, many of the top results are present in a different order than in the original test.




```
## [[1]]
## Total tests  Bonferroni        Holm     FDR<0.1    FDR<0.05    FDR<0.01
##        1535         146         148         447         370         260
##   FDR<0.001
##         166
```

7mers

```
## Warning in chisq.test(data): Chi-squared approximation may be incorrect

## Warning in chisq.test(data): Chi-squared approximation may be incorrect

## Warning in chisq.test(data): Chi-squared approximation may be incorrect

## Warning in chisq.test(data): Chi-squared approximation may be incorrect

## Warning in chisq.test(data): Chi-squared approximation may be incorrect

## Warning in chisq.test(data): Chi-squared approximation may be incorrect

## Warning in chisq.test(data): Chi-squared approximation may be incorrect

## Warning in chisq.test(data): Chi-squared approximation may be incorrect

## Warning in chisq.test(data): Chi-squared approximation may be incorrect

## Warning in chisq.test(data): Chi-squared approximation may be incorrect

## Warning in chisq.test(data): Chi-squared approximation may be incorrect

## Warning in chisq.test(data): Chi-squared approximation may be incorrect

## Warning in chisq.test(data): Chi-squared approximation may be incorrect

## Warning in chisq.test(data): Chi-squared approximation may be incorrect

## Warning in chisq.test(data): Chi-squared approximation may be incorrect

## Warning in chisq.test(data): Chi-squared approximation may be incorrect
```

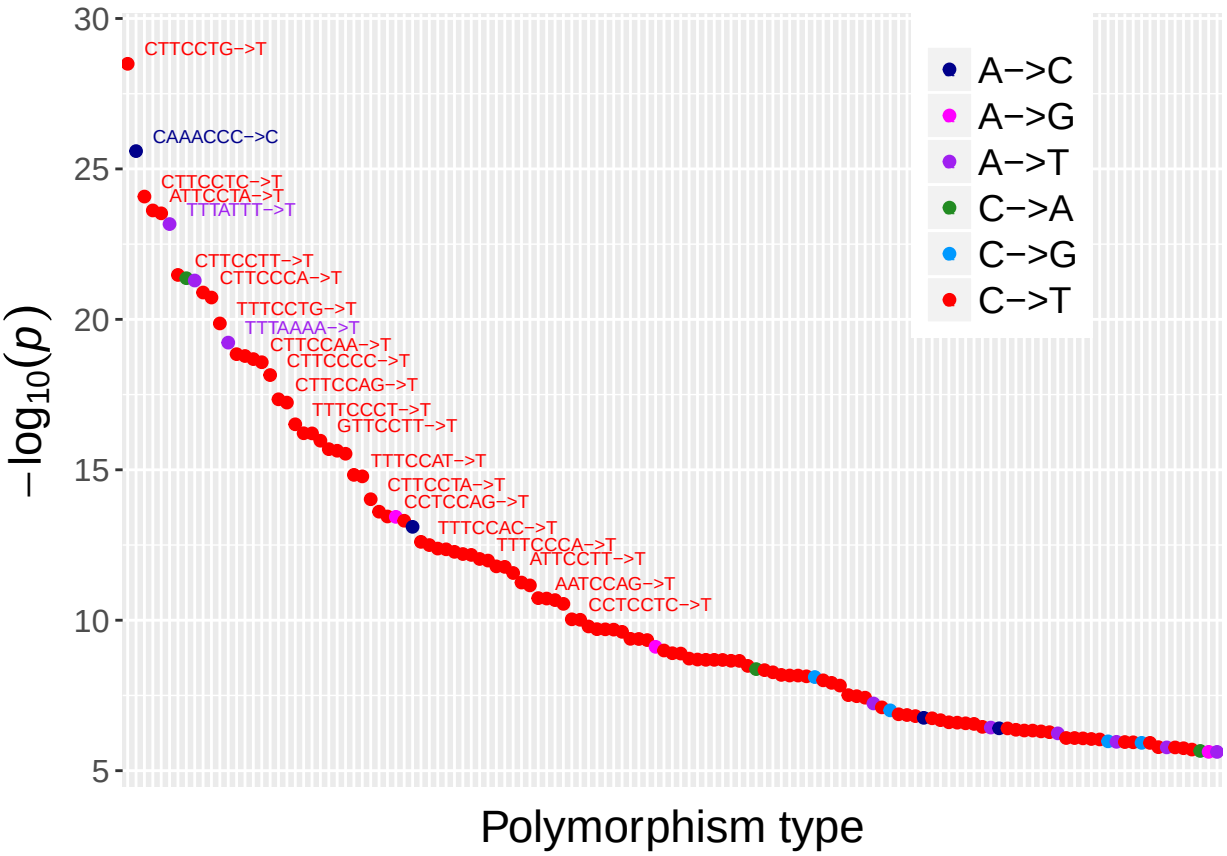
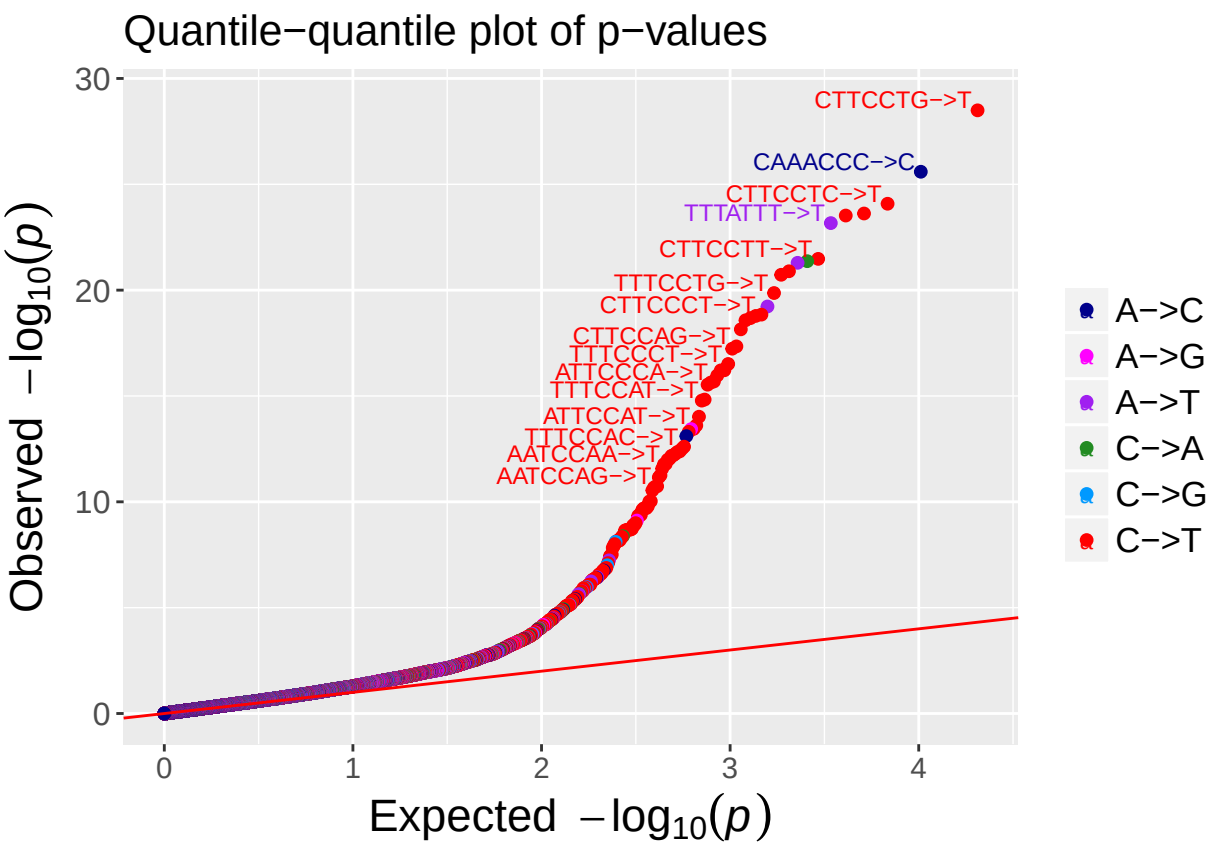


Table 7: 10 most significant new 7mers using ordered p value correction

Context	AFR.Count	EUR.Count	EAS.Count	SAS.Count	p
CTTCCTG->T	1284	430	365	482	3.208407e-29
CAAACCC->C	120	17	101	11	2.556951e-26
CTTCCTC->T	1250	410	343	380	8.284838e-25
ATTCCTA->T	620	242	171	230	2.399324e-24
CTTCCAT->T	652	257	230	235	2.992203e-24
TTTATTT->T	2578	387	743	438	6.803340e-24
CTTCCTT->T	1771	519	474	594	3.338707e-22
AAACAAA->A	2866	394	709	520	4.284298e-22
ATTAAAA->T	3545	456	791	750	5.112000e-22
CTTCCCA->T	1117	365	336	404	1.282007e-21

Notice that, for 7mers, the ordering of the top ten most significant results is entirely unchanged.




```
## [[1]]
## Total tests  Bonferroni        Holm     FDR<0.1    FDR<0.05    FDR<0.01
##       20513         131         131         412         318         213
##   FDR<0.001
##         145
```

Summary

The following table summarizes the numbers of significant results from this section.

Context Model	Number Significant	Number New
3mer	62	–
5mer	146	113
7mer	131	120

Notice that using the ordered p value calculation causes us to pick up far fewer significant 3mers, but slightly more 5mers and 7mers. Moreover, it seems that using p ordered on 3mers has a much greater effect than on 5mers and 7mers. This might be expected because, for the most significant 7mers or 5mers, our p-value calculations still include most of the data, however, for our highly significant 3mers, a much larger portion of the data is excluded.