In [2]:
```
%load_ext rpy2.ipython
# Activates R cell magic
```

```
/home/wilfried/conda/wil_works/lib/python3.7/site-packages/rpy2/robject
s/pandas2ri.py:14: FutureWarning: pandas.core.index is deprecated and w
ill be removed in a future version.  The public classes are available i
n the top-level namespace.
  from pandas.core.index import Index as PandasIndex
```

This document was generated to answer peer-reviewers. We are thankful for their construstive criticisms.

Table S8

In [55]:
```
# SNPs in Upstream inside stable G4s
infile = open('/nfs/brubeck.bx.psu.edu/scratch5/wilfried/nonB/Mutation/U
pstream1k.SFS.high.test', 'rt')

span = 0
count = 0

for line in infile:
    chrom, start, end, SFS = line.strip().split('\t')
    span += int(end) - int(start) + 1

    homo, pongo, ref, alt, MAF = SFS.split('|')

    if ref != 'NA' and homo == 'NA':
        count += 1

print(str(count)+' SNPs in '+str(span)+' bp')
```

```
8639 SNPs in 1610914 bp
```

In [52]:
```
# SNPs in Upstream inside unstable G4s

infile = open('/nfs/brubeck.bx.psu.edu/scratch5/wilfried/nonB/Mutation/Upstream1k.SFS.low.test', 'rt')

span = 0
count = 0

for line in infile:
    chrom, start, end, SFS = line.strip().split('\t')
    span += int(end) - int(start) + 1

    homo, pongo, ref, alt, MAF = SFS.split('|')

    if ref != 'NA' and homo == 'NA':
        count += 1

print(str(count)+' SNPs in '+str(span)+' bp')
```

```
2421 SNPs in 988262 bp
```

In [54]:
```
# SNPs in upstream outside of G4s

infile = open('/nfs/brubeck.bx.psu.edu/scratch5/wilfried/nonB/Mutation/Upstream1k.SFS.control', 'rt')

span = 0
count = 0

for line in infile:
    chrom, start, end, SFS, score, strand = line.strip().split('\t')
    span += int(end) - int(start) + 1

    homo, pongo, ref, alt, MAF = SFS.split('|')

    if ref != 'NA' and homo == 'NA':
        count += 1

print(str(count)+' SNPs in '+str(span)+' bp')
```

```
52209 SNPs in 51174740 bp
```

In [56]:
```
%%R
# Proportion test: Inside vs Outside G4s

fisher.test(cbind(c(8639+2421, 1610914+988262),c(52209, 51174740)))
```

```
        Fisher's Exact Test for Count Data

data:  cbind(c(8639 + 2421, 1610914 + 988262), c(52209, 51174740))
p-value < 2.2e-16
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 4.086046 4.258102
sample estimates:
odds ratio
  4.170985
```

In [123]:
```
# FNS in Upstream inside stable G4s

infile = open('/nfs/brubeck.bx.psu.edu/scratch5/wilfried/nonB/Mutation/Upstream1k.SFS.high.test', 'rt')

span = 0
count = 0

for line in infile:
    chrom, start, end, SFS = line.strip().split('\t')
    span += int(end) - int(start) + 1

    homo, pongo, ref, alt, MAF = SFS.split('|')

    if homo != 'NA' and homo != 'No':
        count += 1

print(str(count)+' FNS in '+str(span)+' bp')
```

```
36771 FNS in 1610914 bp
```

In [124]:
```
# FNS in Upstream inside unstable G4s

infile = open('/nfs/brubeck.bx.psu.edu/scratch5/wilfried/nonB/Mutation/Upstream1k.SFS.low.test', 'rt')

span = 0
count = 0

for line in infile:
    chrom, start, end, SFS = line.strip().split('\t')
    span += int(end) - int(start) + 1

    homo, pongo, ref, alt, MAF = SFS.split('|')

    if homo != 'NA' and homo != 'No':
        count += 1

print(str(count)+' FNS in '+str(span)+' bp')
```

```
18026 FNS in 988262 bp
```

In [125]:
```
# FNS in Upstream outside G4s

infile = open('/nfs/brubeck.bx.psu.edu/scratch5/wilfried/nonB/Mutation/Upstream1k.SFS.control', 'rt')

span = 0
count = 0

for line in infile:
    chrom, start, end, SFS, score, strand = line.strip().split('\t')
    span += int(end) - int(start) + 1

    homo, pongo, ref, alt, MAF = SFS.split('|')

    if homo != 'NA' and homo != 'No':
        count += 1

print(str(count)+' FNS in '+str(span)+' bp')
```

```
909807 FNS in 51174740 bp
```

In [3]:
```
%%R
# Proportion test: Inside vs Outside G4s


fisher.test(cbind(c(13572+6937, 498364+349140),c(582596, 33114500)))
```

```
        Fisher's Exact Test for Count Data

data:  cbind(c(36771 + 18026, 1610914 + 988262), c(909807, 51174740))
p-value < 2.2e-16
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.175511 1.196215
sample estimates:
odds ratio
  1.185842
```

In [126]:
```
# SNP in Enhancers inside stable G4s

infile = open('/nfs/brubeck.bx.psu.edu/scratch5/wilfried/nonB/Mutation/enhancers.SFS.high.test', 'rt')

span = 0
count = 0

for line in infile:
    chrom, start, end, SFS = line.strip().split('\t')
    span += int(end) - int(start) + 1

    homo, pongo, ref, alt, MAF = SFS.split('|')

    if ref != 'NA' and homo == 'NA':
        count += 1

print(str(count)+' SNPs in '+str(span)+' bp')
```

```
2612 SNPs in 498364 bp
```

In [127]:
```
# SNPs in Enhancers inside unstable G4s

infile = open('/nfs/brubeck.bx.psu.edu/scratch5/wilfried/nonB/Mutation/enhancers.SFS.low.test', 'rt')

span = 0
count = 0

for line in infile:
    chrom, start, end, SFS = line.strip().split('\t')
    span += int(end) - int(start) + 1

    homo, pongo, ref, alt, MAF = SFS.split('|')

    if ref != 'NA' and homo == 'NA':
        count += 1

print(str(count)+' SNPs in '+str(span)+' bp')
```

```
601 SNPs in 349140 bp
```

In [128]:
```
# SNPs in Enhancers outside G4s

infile = open('/nfs/brubeck.bx.psu.edu/scratch5/wilfried/nonB/Mutation/enhancers.SFS.control', 'rt')

span = 0
count = 0

for line in infile:
    chrom, start, end, SFS = line.strip().split('\t')
    span += int(end) - int(start) + 1

    homo, pongo, ref, alt, MAF = SFS.split('|')

    if ref != 'NA' and homo == 'NA':
        count += 1

print(str(count)+' SNPs in '+str(span)+' bp')
```

```
20514 SNPs in 33114500 bp
```

In [4]:
```
%%R
# Proportion test: Inside vs Outside G4s


fisher.test(cbind(c(2612+601, 498364+349140),c(20514, 33114500)))
```

```
        Fisher's Exact Test for Count Data

data:  cbind(c(2612 + 601, 498364 + 349140), c(20514, 33114500))
p-value < 2.2e-16
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 5.894276 6.352869
sample estimates:
odds ratio
  6.119761
```

In [61]:
```
# FNS in Enchancers inside stable G4s

infile = open('/nfs/brubeck.bx.psu.edu/scratch5/wilfried/nonB/Mutation/enhancers.SFS.high.test', 'rt')

span = 0
count = 0

for line in infile:
    chrom, start, end, SFS = line.strip().split('\t')
    span += int(end) - int(start) + 1

    homo, pongo, ref, alt, MAF = SFS.split('|')

    if homo != 'NA' and homo != 'No':
        count += 1

print(str(count)+' FNS in '+str(span)+' bp')
```

```
13572 FNS in 498364 bp
```

In [62]:
```
# FNS in Enhancers inside unstable G4s

infile = open('/nfs/brubeck.bx.psu.edu/scratch5/wilfried/nonB/Mutation/enhancers.SFS.low.test', 'rt')

span = 0
count = 0

for line in infile:
    chrom, start, end, SFS = line.strip().split('\t')
    span += int(end) - int(start) + 1

    homo, pongo, ref, alt, MAF = SFS.split('|')

    if homo != 'NA' and homo != 'No':
        count += 1

print(str(count)+' FNS in '+str(span)+' bp')
```

```
6937 FNS in 349140 bp
```

In [64]:
```
# FNS in Enhancers outside G4s

infile = open('/nfs/brubeck.bx.psu.edu/scratch5/wilfried/nonB/Mutation/enhancers.SFS.control', 'rt')

span = 0
count = 0

for line in infile:
    chrom, start, end, SFS = line.strip().split('\t')
    span += int(end) - int(start) + 1

    homo, pongo, ref, alt, MAF = SFS.split('|')

    if homo != 'NA' and homo != 'No':
        count += 1

print(str(count)+' FNS in '+str(span)+' bp')
```

```
582596 FNS in 33114500 bp
```

In [65]:
```
%%R
fisher.test(cbind(c(13572+6937, 498364+349140),c(582596, 33114500)))
```

```
        Fisher's Exact Test for Count Data

data:  cbind(c(13572 + 6937, 498364 + 349140), c(582596, 33114500))
p-value < 2.2e-16
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.356178 1.395013
sample estimates:
odds ratio
   1.37546
```

In [89]:
```
# SNPs in NCNR inside stable G4s

infile = open('/nfs/brubeck.bx.psu.edu/scratch5/wilfried/nonB/Mutation/NCNR.SFS.high.bed.test', 'rt')

span = 0
count = 0

for line in infile:
    chrom, start, end, SFS = line.strip().split('\t')[0:4]
    span += int(end) - int(start) + 1

    homo, pongo, ref, alt, MAF = SFS.split('|')

    if ref != 'NA' and homo == 'NA':
        count += 1

print(str(count)+' SNP in '+str(span)+' bp')
```

```
1244 SNP in 308300 bp
```

In [90]:
```
# SNPs in NCNR inside unstable G4s

infile = open('/nfs/brubeck.bx.psu.edu/scratch5/wilfried/nonB/Mutation/NCNR.SFS.low.bed.test', 'rt')

span = 0
count = 0

for line in infile:
    chrom, start, end, SFS = line.strip().split('\t')[0:4]
    span += int(end) - int(start) + 1

    homo, pongo, ref, alt, MAF = SFS.split('|')

    if ref != 'NA' and homo == 'NA':
        count += 1

print(str(count)+' SNP in '+str(span)+' bp')
```

```
874 SNP in 513210 bp
```

In [91]:
```
# SNPs in NCNR outside G4s

infile = open('/nfs/brubeck.bx.psu.edu/scratch5/wilfried/nonB/Mutation/NCNR.SFS.control.shuf', 'rt')

span = 0
count = 0

for line in infile:
    chrom, start, end, SFS = line.strip().split('\t')[0:4]
    span += int(end) - int(start) + 1

    homo, pongo, ref, alt, MAF = SFS.split('|')

    if ref != 'NA' and homo == 'NA':
        count += 1

print(str(count)+' SNP in '+str(span)+' bp')
```

```
14812 SNP in 20000000 bp
```

In [92]:
```
%%R
fisher.test(cbind(c(1244+874, 308300+513210),c(14812, 20000000)))
```

```
        Fisher's Exact Test for Count Data

data:  cbind(c(1244 + 874, 308300 + 513210), c(14812, 2e+07))
p-value < 2.2e-16
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 3.324672 3.644112
sample estimates:
odds ratio
  3.481184
```

In [81]:
```
# FNS in NCNR inside stable G4s

infile = open('/nfs/brubeck.bx.psu.edu/scratch5/wilfried/nonB/Mutation/NCNR.SFS.high.bed.test', 'rt')

span = 0
count = 0

for line in infile:
    chrom, start, end, SFS = line.strip().split('\t')[0:4]
    span += int(end) - int(start) + 1

    homo, pongo, ref, alt, MAF = SFS.split('|')

    if homo != 'NA' and homo != 'No':
        count += 1

print(str(count)+' FNS in '+str(span)+' bp')
```

```
8153 FNS in 308300 bp
```

In [87]:
```
# FNS in NCNR inside unstable G4s

infile = open('/nfs/brubeck.bx.psu.edu/scratch5/wilfried/nonB/Mutation/NCNR.SFS.low.bed.test', 'rt')

span = 0
count = 0

for line in infile:
    chrom, start, end, SFS = line.strip().split('\t')[0:4]
    span += int(end) - int(start) + 1

    homo, pongo, ref, alt, MAF = SFS.split('|')

    if homo != 'NA' and homo != 'No':
        count += 1

print(str(count)+' FNS in '+str(span)+' bp')
```

```
10553 FNS in 513210 bp
```

In [86]:
```
# FNS in NCNR outside G4s

infile = open('/nfs/brubeck.bx.psu.edu/scratch5/wilfried/nonB/Mutation/NCNR.SFS.control.shuf', 'rt')

span = 0
count = 0

for line in infile:
    chrom, start, end, SFS = line.strip().split('\t')[0:4]
    span += int(end) - int(start) + 1

    homo, pongo, ref, alt, MAF = SFS.split('|')

    if homo != 'NA' and homo != 'No':
        count += 1

print(str(count)+' FNS in '+str(span)+' bp')
```

```
276323 FNS in 20000000 bp
```

In [88]:
```
%%R
fisher.test(cbind(c(8153+10553, 308300+513210),c(276323, 20000000)))
```

```
        Fisher's Exact Test for Count Data

data:  cbind(c(8153 + 10553, 308300 + 513210), c(276323, 2e+07))
p-value < 2.2e-16
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 1.623544 1.672955
sample estimates:
odds ratio
   1.64808
```

Enhancers from ENCODE

Download track from UCSC: EncodeCcreCombined

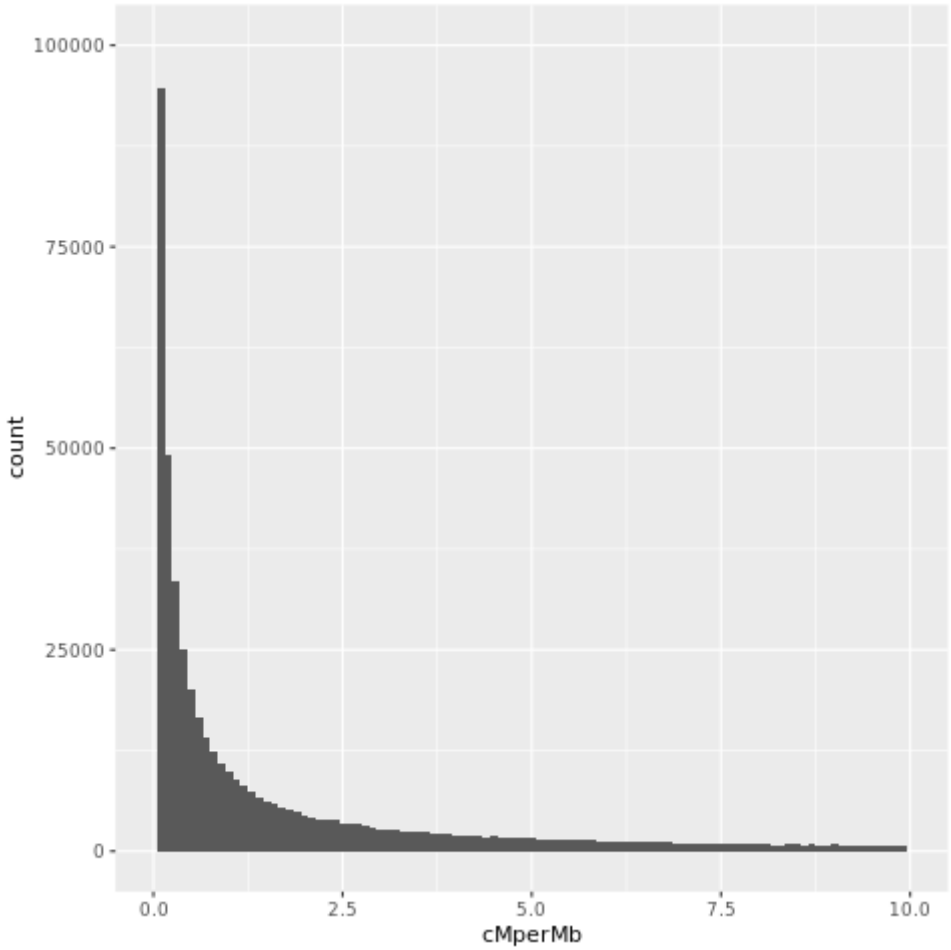
```
cut -f 1-3,13  EncodeCcreCombined > EncodeCcreCombined.bed
grep enh EncodeCcreCombined.bed > EncodeEnhancers.bed

bedtools sort -i EncodeEnhancers.bed > EncodeEnhancers.sorted.bed
liftOver EncodeEnhancers.sorted.bed hg38ToHg19.over.chain EncodeEnhancers.hg
19.bed unMapped

bedtools sort -i EncodeEnhancers.hg19.bed | grep -v chrX | grep -v chrY | gr
ep -v chrM | grep -v chrU | grep -v random > EncodeEnhancers.hg19.sorted.bed
python Locuschoice.py EncodeEnhancers.hg19.sorted.bed EncodeEnhancers.noover
lap


bedtools coverage -a EncodeEnhancers.nooverlap -b quadron+.bed | cut -f 1-3,
10 -s > EncodeEnhancers+.G4coverage
bedtools coverage -a EncodeEnhancers.nooverlap -b quadron-.bed | cut -f 1-3,
10 -s > EncodeEnhancers-.G4coverage
bedtools getfasta -fi hg19.fa -bed EncodeEnhancers.nooverlap > EncodeEnhance
rs.getfa
bedtools coverage -a quadron+.bed -b EncodeEnhancers.nooverlap  | awk '{sum+
=$11}END{print sum}' > EncodeEnhancers+.G4intersectcount
bedtools coverage -a quadron-.bed -b EncodeEnhancers.nooverlap  | awk '{sum+
=$11}END{print sum}' > EncodeEnhancers-.G4intersectcount
awk '{sum+=$3-$2+1}END{print sum}' EncodeEnhancers.nooverlap > EncodeEnhance
rs.bpcount
```

High Vs Low Recombination Rate

Find thresholds: using quantiles

In [3]:
```
%%R

library(ggplot2)
library(grid)
library(gridExtra)

data <- read.table('genetic.map', header=TRUE)
colnames(data) <- c('Chr','Begin','End','cMperMb','cM')

print(head(data))

g <- ggplot(data,aes(cMperMb)) + geom_histogram(binwidth = 0.1) +
    scale_x_continuous(limits = c(0,10)) +
    scale_y_continuous(limits = c(0,1e+05))
 #   coord_cartesian(xlim = c(0,2.5))

grid.arrange(g,nrow=1)
```

```
   Chr    Begin     End      cMperMb          cM
1 chr1 1431813 1515567 3.244924e-02 0.002717753
2 chr1 1515567 1530002 1.895966e-01 0.005454580
3 chr1 1530002 1534402 4.630789e-03 0.005474955
4 chr1 1534402 1538787 1.986456e-02 0.005562061
5 chr1 1538787 1541864 3.320831e-04 0.005563083
6 chr1 1541864 1542773 1.052292e-86 0.005563083
```



In [4]:
```
%%R

quantile(data$cMperMb)
```

```
          0%          25%          50%          75%         100% 
0.000000e+00 1.992167e-04 7.340754e-02 1.464003e+00 5.624274e+04 
```

Split genetic map in High and Low recombination

In [234]:
```
infile = open('genetic.map', 'rt')
outfile_low = open('lowrecomb.bed', 'w+')
outfile_high = open('highrecomb.bed', 'w+')

for line in infile:
    if line[0:3] == 'chr':
        array = line.strip().split('\t')
        chrom, start, end, cMperMb, cM = array

        if float(cMperMb) >= 1.464003e+00:
            outfile_low.write(chrom+'\t'+start+'\t'+end+'\n')

        elif float(cMperMb) <= 1.992167e-04:
            outfile_high.write(chrom+'\t'+start+'\t'+end+'\n')
```

```
bedtools intersect -a NCNR.nooverlap -b lowrecomb.bed -wa -wb > NCNR.lowreco
mb.intersect
bedtools intersect -a NCNR.nooverlap -b highrecomb.bed -wa -wb > NCNR.highre
comb.intersect
```

In [202]:
```
infile = open('NCNR.lowrecomb.intersect', 'rt')
outfile = open('NCNR.lowrecomb.bed', 'w+')

for line in infile:
    array = line.strip().split('\t')
    chrom, NCNRstart, NCNRend, dummy, RecombStart, RecombEnd = array

    if int(NCNRstart) < int(RecombStart):
        NCNRstart = RecombStart

    elif int(NCNRend) > int(RecombEnd):
         NCNRend = RecombEnd

    outfile.write(chrom+'\t'+NCNRstart+'\t'+NCNRend+'\n')
```

In [204]:
```
infile = open('NCNR.highrecomb.intersect', 'rt')
outfile = open('NCNR.highrecomb.bed', 'w+')

for line in infile:
    array = line.strip().split('\t')
    chrom, NCNRstart, NCNRend, dummy, RecombStart, RecombEnd = array

    if int(NCNRstart) < int(RecombStart):
        NCNRstart = RecombStart

    elif int(NCNRend) > int(RecombEnd):
         NCNRend = RecombEnd

    outfile.write(chrom+'\t'+NCNRstart+'\t'+NCNRend+'\n')
```

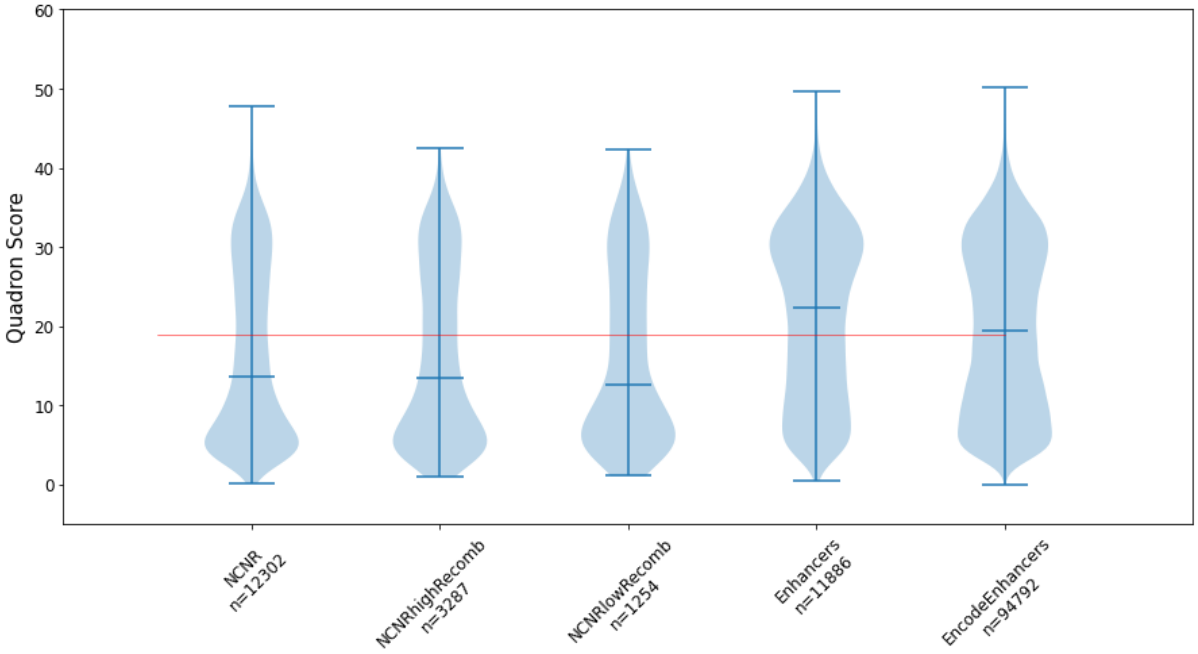
Quadron scores distributions

```
bedtools annotate -i quadron.nooverlap -files Upstream1k.nooverlap FUTR.noov
erlap Exons.nooverlap Introns.nooverlap TUTR.nooverlap Downstream1k.nooverla
p hg19RM.bed hg19Interspersed.bed RepOrigin.nooverlap eQTL.nooverlap enhance
rs.nooverlap promoters.nooverlap phastCons.nooverlap hg19CTCF.sorted.bed TAD
_boundary_regions.nooverlap recomb_hotspots.nooverlap hg19Upstream5kSorted_R
efSeq.bed hg19Downstream5kSorted_RefSeq.bed  CpGIsland.nooverlap NCNR.noover
lap eQTL.extended.nooverlap phastCons.extended.nooverlap NCNR.lowrecomb.bed
 NCNR.highrecomb.bed EncodeEnhancers.nooverlap > Annotate19Scores
```

In [235]:
```
import pandas as pd

Annotatefile = open('Annotate19Scores', 'rt')

colnames = ['chrom', 'start', 'end', 'motif', 'length', 'strand', 'scor
e', 'Upstream', 'FUTR', 'Coding', 'Introns',\
            'TUTR', 'Downstream', 'RM_no_Interspersed', 'Interspersed',
'RepOrigin', 'eQTL', 'enhancers', 'promoters',\
             'phastCons', 'CTCF', 'TAD_Boundaries', 'Recomb Hotspots', 'U
pstream5k', 'Downstream5k', 'CpG Islands',
             'NCNR', 'extended eQTL', 'extended phastCons', 'NCNRlowRecom
b', 'NCNRhighRecomb', 'EncodeEnhancers']
rows = []

for line in Annotatefile:
    array = line.strip().split('\t')
    #print(array)
    #break
    if array[6] == 'NA': # remove G4 without a score
        continue
    else:
        rows.append([array[0], int(array[1]), int(array[2]), array[3], i
nt(array[4]), array[5], float(array[6]),\
                     float(array[7]),float(array[8]),float(array[9]),floa
t(array[10]),float(array[11]),\
                     float(array[12]),float(array[13]), float(array[14]),
float(array[15]), float(array[16]),\
                     float(array[17]),float(array[18]),float(array[19]),f
loat(array[20]),float(array[21]),float(array[22]),\
                     float(array[23]), float(array[24]), float(array[25
]), float(array[26]), float(array[27]), float(array[28]),\
                     float(array[29]), float(array[30]), float(array[31
])])

dataframe_annotate = pd.DataFrame(data=rows, columns=colnames)
```

In [236]:
```
NCNR = dataframe_annotate.loc[dataframe_annotate['NCNR'] == 1]['score'].
tolist()
NCNRhighRecomb = dataframe_annotate.loc[dataframe_annotate['NCNRhighReco
mb'] == 1]['score'].tolist()
NCNRlowRecomb = dataframe_annotate.loc[dataframe_annotate['NCNRlowRecom
b'] == 1]['score'].tolist()

Enhancers = dataframe_annotate.loc[dataframe_annotate['enhancers'] == 1]
\
            ['score'].tolist()
EncodeEnhancers = dataframe_annotate.loc[dataframe_annotate['EncodeEnhan
cers'] == 1]\
            ['score'].tolist()
```

In [237]:
```
# Help from https://stackoverflow.com/questions/29779079/adding-a-scatter-of-points-to-a-boxplot-using-matplotlib
import matplotlib
import matplotlib.pyplot as plt
import matplotlib.patches as mpatches
import numpy as np

# Genic elements (stranded) only



font = {'fontname':'sans'}

toplot = [NCNR, NCNRhighRecomb, NCNRlowRecomb, Enhancers, EncodeEnhancers]

fig, ax = plt.subplots(figsize=(15, 7))

sizes = []

for array in toplot:
    sizes.append(len(array))

maxsize = max(sizes)

ratiosizes = []

for size in sizes:
    ratiosizes.append(1)

violin_part1 = ax.violinplot(toplot, showmedians=True, positions=[0,2,4,6,8], widths=ratiosizes)


ax.plot([-1,8], [19,19], color='red', linewidth=0.5)

ax.set_ylabel('Quadron Score', fontsize=15)
ax.set_ylim(-5,60)
#ax.set_xlim(-1,31)
ax.set_xlim(-2,10)

#xticks = ax.set_xticks([1,5,9,13,17,21,25,29])
xticks = ax.set_xticks([0,2,4,6,8])


xticks = ax.set_xticklabels(['NCNR\nn='+str(sizes[0]),'NCNRhighRecomb\nn='+str(sizes[1]),'NCNRlowRecomb\nn='+str(sizes[2]), 'Enhancers\nn='+str(sizes[3]),\
                             "EncodeEnhancers\nn="+str(sizes[4])], rotation=45,  fontsize = 12, **font)

yticks = ax.tick_params(labelsize=12)
```


Quadron Score
60
50
40
30
20
10
0
NCNR
n=12302
NCNRhighRecomb
n=3287
NCNRlowRecomb
n=1254
Enhancers
n=11886
EncodeEnhancers
n=94792

In [238]:
```
from numpy.random import permutation
import numpy as np

def permutation_test(TestDist, ControlDist, permutations):

    mean = np.mean(TestDist)
    median = np.median(TestDist)

    deltamean = abs(np.mean(TestDist) - np.mean(ControlDist))
    deltamedian = abs(np.median(TestDist) - np.median(ControlDist))

    Pool = ControlDist + TestDist

    mockdeltameans = []
    mockdeltamedians = []
    i = 1

    while i < permutations:
        PermutedPool = np.random.permutation(Pool)
        MockTest = PermutedPool[:len(TestDist)]
        MockControl = PermutedPool[len(TestDist):]
        mockdeltamean = abs(np.mean(MockTest) - np.mean(MockControl))
        mockdeltameans.append(mockdeltamean)
        mockdeltamedian = abs(np.median(MockTest) - np.median(MockContro
l))
        mockdeltamedians.append(mockdeltamedian)
        i += 1

    deltameans = mockdeltameans + [deltamean]
    #print(len(deltameans))
    deltamedians = mockdeltamedians + [deltamedian]
    meanpvalue = sum(i >= deltamean for i in deltameans) / float(permuta
tions)# + 1)
    medianpvalue = sum(i >= deltamedian for i in deltamedians) / float(p
ermutations)# + 1)

    return meanpvalue, medianpvalue
```

In [239]:
```
import numpy as np

print('NCNRhighRecomb vs NCNRlowRecomb', np.median(NCNRhighRecomb), np.m
edian(NCNRhighRecomb) / np.median(NCNRlowRecomb), permutation_test(NCNRh
ighRecomb, NCNRlowRecomb, 1000)[1])
print('Enhancers vs EncodeEnhancers', np.median(Enhancers), np.median(En
hancers) / np.median(EncodeEnhancers), permutation_test(Enhancers, Encod
eEnhancers, 1000)[1])
```

```
NCNRhighRecomb vs NCNRlowRecomb 13.42 1.070602313522138 0.299
Enhancers vs EncodeEnhancers 22.365000000000002 1.1552169421487604 0.00
1
```

HKA

Following values were obtained with the same scripts described in the 'HKA' notebook

In [271]:
```
%%R

NCNR_fisher <- fisher.test(cbind(c(1320+1369, 8153+10553),c(32824, 276323)))
NCNR_high_fisher <- fisher.test(cbind(c(1320, 8153),c(32824, 276323)))
NCNR_low_fisher <- fisher.test(cbind(c(1369, 10553),c(32824, 276323)))
print(paste('NCNR',NCNR_fisher$estimate,NCNR_fisher$p.value))
print(paste('NCNR High',NCNR_high_fisher$estimate,NCNR_high_fisher$p.value))
print(paste('NCNR Low',NCNR_low_fisher$estimate,NCNR_low_fisher$p.value))


NCNR_highrecomb_fisher <- fisher.test(cbind(c(387+369, 2272+2869),c(6006, 51021)))
NCNR_highrecomb_high_fisher <- fisher.test(cbind(c(387, 2272),c(6006, 51021)))
NCNR_highrecomb_low_fisher <- fisher.test(cbind(c(369, 2869),c(6006, 51021)))
print(paste('NCNR highrecomb',NCNR_highrecomb_fisher$estimate,NCNR_highrecomb_fisher$p.value))
print(paste('NCNR highrecomb High',NCNR_highrecomb_high_fisher$estimate,NCNR_highrecomb_high_fisher$p.value))
print(paste('NCNR highrecomb Low',NCNR_highrecomb_low_fisher$estimate,NCNR_highrecomb_low_fisher$p.value))


NCNR_lowrecomb_fisher <- fisher.test(cbind(c(133+138, 842+1210),c(21642, 174988)))
NCNR_lowrecomb_high_fisher <- fisher.test(cbind(c(133, 842),c(21642, 174988)))
NCNR_lowrecomb_low_fisher <- fisher.test(cbind(c(138, 1210),c(21642, 174988)))
print(paste('NCNR lowrecomb',NCNR_lowrecomb_fisher$estimate,NCNR_lowrecomb_fisher$p.value))
print(paste('NCNR lowrecomb High',NCNR_lowrecomb_high_fisher$estimate,NCNR_lowrecomb_high_fisher$p.value))
print(paste('NCNR lowrecomb Low',NCNR_lowrecomb_low_fisher$estimate,NCNR_lowrecomb_low_fisher$p.value))


Enhancers_fisher <- fisher.test(cbind(c(2549+977, 13572+6937),c(74321, 582596)))
Enhancers_high_fisher <- fisher.test(cbind(c(2549, 13572),c(74321, 582596)))
Enhancers_low_fisher <- fisher.test(cbind(c(977, 6937),c(74321, 582596)))
print(paste('Enhancers', Enhancers_fisher$estimate,Enhancers_fisher$p.value))
print(paste('Enhancers High', Enhancers_high_fisher$estimate,Enhancers_high_fisher$p.value))
print(paste('Enhancers Low', Enhancers_low_fisher$estimate,Enhancers_low_fisher$p.value))


EncodeEnhancers_fisher <- fisher.test(cbind(c(15623+8728, 91161+65311),c
```

```
(827525, 6559019)))
EncodeEnhancers_high_fisher <- fisher.test(cbind(c(15623, 91161),c(82752
5, 6559019)))
EncodeEnhancers_low_fisher <- fisher.test(cbind(c(8728, 65311),c(827525,
6559019)))
print(paste('Encode Enhancers', EncodeEnhancers_fisher$estimate,EncodeEn
hancers_fisher$p.value))
print(paste('Encode Enhancers High', EncodeEnhancers_high_fisher$estimat
e,EncodeEnhancers_high_fisher$p.value))
print(paste('Encode Enhancers Low', EncodeEnhancers_low_fisher$estimate,
EncodeEnhancers_low_fisher$p.value))

EncodeEnhancers_highVslow_fisher <-fisher.test(cbind(c(15623, 91161),c(8
728, 65311)))
print(paste('Encode Enhancers High vs Low', EncodeEnhancers_highVslow_fi
sher$estimate,EncodeEnhancers_highVslow_fisher$p.value))
```

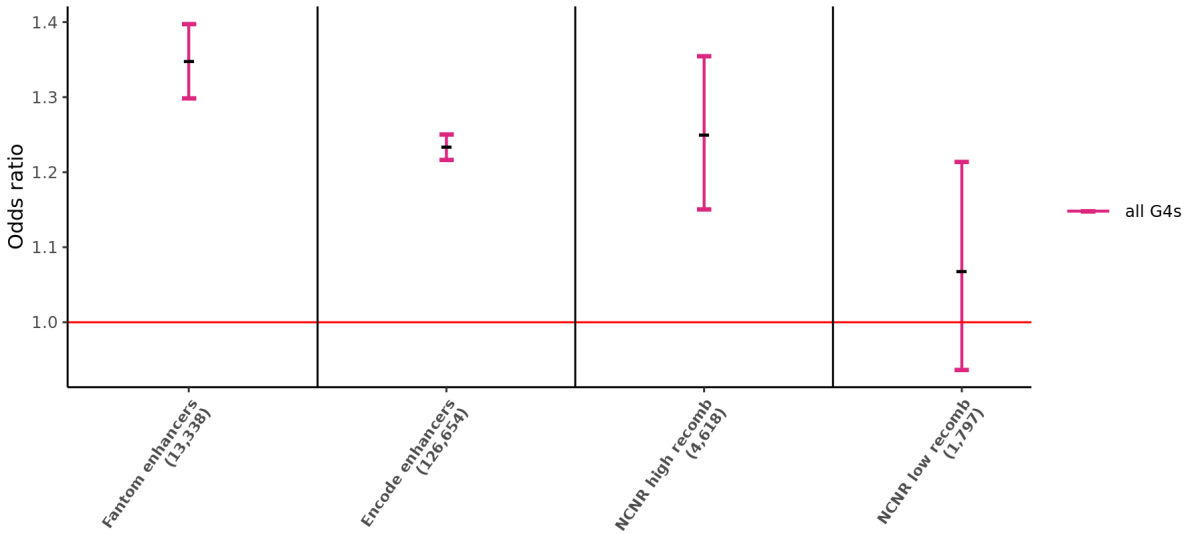
```
[1] "NCNR 1.21012485570186 2.6470537339199e-18"
[1] "NCNR High 1.36295128597438 4.24329082548456e-23"
[1] "NCNR Low 1.09210827607119 0.00287679064791438"
[1] "NCNR highrecomb 1.24921140921844 1.27587246005017e-07"
[1] "NCNR highrecomb High 1.44697714757429 3.55955462876927e-10"
[1] "NCNR highrecomb Low 1.09262425182783 0.119613216413723"
[1] "NCNR lowrecomb 1.06785283336816 0.317222327579106"
[1] "NCNR lowrecomb High 1.27717232136507 0.0103144737700312"
[1] "NCNR lowrecomb Low 0.922148949149551 0.382673392428103"
[1] "Enhancers 1.34769815601055 3.7518863981056e-54"
[1] "Enhancers High 1.47224422546273 4.03953372189528e-64"
[1] "Enhancers Low 1.10399097691151 0.00430506334346868"
[1] "Encode Enhancers 1.2334957327623 6.71747159376394e-189"
[1] "Encode Enhancers High 1.3583505435013 4.52015927230923e-251"
[1] "Encode Enhancers Low 1.05923698516702 6.10933895753114e-07"
[1] "Encode Enhancers High vs Low 1.28240671276457 1.15770126547589e-6
8"
```

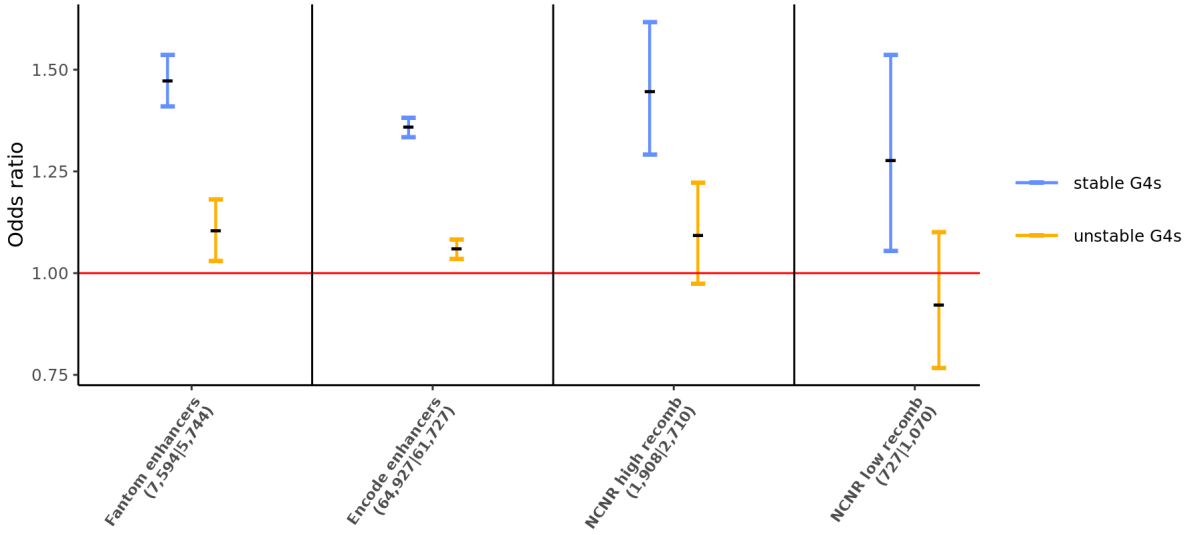
In [262]:
```
%%R
#https://stackoverflow.com/questions/14069629/how-can-i-plot-data-with-confidence-intervals

x <- c(1,2,3,4)

#print(length(x))

ORs <- c(Enhancers_fisher$estimate,
         EncodeEnhancers_fisher$estimate,
         NCNR_highrecomb_fisher$estimate,
         NCNR_lowrecomb_fisher$estimate)


Low_ORs <- c(Enhancers_low_fisher$estimate,
             EncodeEnhancers_low_fisher$estimate,
             NCNR_highrecomb_low_fisher$estimate,
             NCNR_lowrecomb_low_fisher$estimate)

High_ORs <- c(Enhancers_high_fisher$estimate,
              EncodeEnhancers_high_fisher$estimate,
              NCNR_highrecomb_high_fisher$estimate,
              NCNR_lowrecomb_high_fisher$estimate)

Ls <- c(Enhancers_fisher$conf.int[1],
         EncodeEnhancers_fisher$conf.int[1],
         NCNR_highrecomb_fisher$conf.int[1],
         NCNR_lowrecomb_fisher$conf.int[1])


Low_Ls <- c(Enhancers_low_fisher$conf.int[1],
             EncodeEnhancers_low_fisher$conf.int[1],
             NCNR_highrecomb_low_fisher$conf.int[1],
             NCNR_lowrecomb_low_fisher$conf.int[1])


High_Ls <- c(Enhancers_high_fisher$conf.int[1],
              EncodeEnhancers_high_fisher$conf.int[1],
              NCNR_highrecomb_high_fisher$conf.int[1],
              NCNR_lowrecomb_high_fisher$conf.int[1])

Us <- c(Enhancers_fisher$conf.int[2],
         EncodeEnhancers_fisher$conf.int[2],
         NCNR_highrecomb_fisher$conf.int[2],
         NCNR_lowrecomb_fisher$conf.int[2])


Low_Us <- c(Enhancers_low_fisher$conf.int[2],
             EncodeEnhancers_low_fisher$conf.int[2],
             NCNR_highrecomb_low_fisher$conf.int[2],
             NCNR_lowrecomb_low_fisher$conf.int[2])


High_Us <- c(Enhancers_high_fisher$conf.int[2],
              EncodeEnhancers_high_fisher$conf.int[2],
```

```
NCNR_highrecomb_high_fisher$conf.int[2],
NCNR_lowrecomb_high_fisher$conf.int[2])
```

In [263]:
```
%%R -w 9 -h 10 --units in -r 200



xlegend = c("Fantom enhancers\n(13,338)","Encode enhancers\n(126,654)",
"NCNR high recomb\n(4,618)","NCNR low recomb\n(1,797)")

dataf <- cbind.data.frame(xlegend,x,ORs,Low_ORs,High_ORs,Ls,Low_Ls,High_
Ls,Us,Low_Us,High_Us)
names(dataf) <- c("xlegend","x","ORs","Low_ORs","High_ORs","Ls","Low_Ls"
,"High_Ls",
                  "Us","Low_Us","High_Us")

cols <- c("all"="#DC267F",
          "high"="#648FFF",
          "low"="#FFB000")


g1 <- ggplot(dataf,) +
   geom_segment(aes(x=(x), xend=(x), y=Ls, yend=Us, colour="all"), size=
0.75) +

   geom_point(aes(x=(x), y=ORs),shape='-', size=7) +
   geom_point(aes(x=(x), y=Us, colour="all"),shape='-', size=10) +
   geom_point(aes(x=(x), y=Ls, colour="all"),shape='-', size=10) +

   geom_hline(yintercept=1, color="red") +
   geom_vline(xintercept=1.5, color="black") +
   geom_vline(xintercept=2.5, color="black") +
   geom_vline(xintercept=3.5, color="black") +
   geom_vline(xintercept=4.5, color="black") +



   theme_classic() + labs(subtitle="Classic Theme")+
   scale_x_continuous(limits = c(0.7,4.1),  breaks=c(1,2,3,4),labels=xle
gend) +
   theme(axis.text.x=element_text(angle = 55, hjust = 1,family="sans",fa
ce="bold",size=8)) +
   scale_colour_manual(name="",values=cols,labels="all G4s") +
   labs(title="", subtitle="", y="Odds ratio", x="", caption="",family=
"sans")


xlegend = c("Fantom enhancers\n(7,594|5,744)","Encode enhancers\n(64,927
|61,727)","NCNR high recomb\n(1,908|2,710)","NCNR low recomb\n(727|1,07
0)")

dataf <- cbind.data.frame(xlegend,x,ORs,Low_ORs,High_ORs,Ls,Low_Ls,High_
Ls,Us,Low_Us,High_Us)
names(dataf) <- c("xlegend","x","ORs","Low_ORs","High_ORs","Ls","Low_Ls"
,"High_Ls",
                  "Us","Low_Us","High_Us")

cols <- c("all"="#DC267F",
          "high"="#648FFF",
```

```
          "low"="#FFB000")


g2 <- ggplot(dataf,) +
   geom_segment(aes(x=(x+0.1), xend=(x+0.1), y=Low_Ls, yend=Low_Us, colo
ur="low"), size=0.75) +
   geom_segment(aes(x=(x-0.1), xend=(x-0.1), y=High_Ls, yend=High_Us, co
lour="high"), size=0.75) +

   geom_point(aes(x=(x+0.1), y=Low_ORs),shape='-',size=7) +
   geom_point(aes(x=(x-0.1), y=High_ORs),shape='-',size=7) +
   geom_point(aes(x=(x+0.1), y=Low_Us, colour="low"),shape='-', size=10)
+
   geom_point(aes(x=(x+0.1), y=Low_Ls, colour="low"),shape='-', size=10)
+
   geom_point(aes(x=(x-0.1), y=High_Us, colour="high"),shape='-', size=1
0) +
   geom_point(aes(x=(x-0.1), y=High_Ls, colour="high"),shape='-', size=1
0) +

   geom_hline(yintercept=1, color="red") +
   geom_vline(xintercept=1.5, color="black") +
   geom_vline(xintercept=2.5, color="black") +
   geom_vline(xintercept=3.5, color="black") +
   geom_vline(xintercept=4.5, color="black") +


   theme_classic() + labs(subtitle="Classic Theme")+
   scale_x_continuous(limits = c(0.7,4.1),  breaks=c(1,2,3,4),labels=xle
gend) +
   theme(axis.text.x=element_text(angle = 55, hjust = 1, face="bold",fam
ily="sans",size=8)) +
   scale_colour_manual(name="",values=cols,labels=c('stable G4s','unstab
le G4s')) +
   labs(title="", subtitle="", y="Odds ratio", x="", caption="",family=
"sans")

grid.arrange(g1,g2,nrow=2)
G <- arrangeGrob(g1,g2,nrow=2)
#ggsave(file='HKA.pdf',G, width=9, height=10, dpi=300)
```





Site Frequency Spectrums

In [248]:
```
NCNR_high_infile = open('/nfs/brubeck.bx.psu.edu/scratch5/wilfried/nonB/
Mutation/NCNR.SFS.high.bed.test')
NCNR_low_infile = open('/nfs/brubeck.bx.psu.edu/scratch5/wilfried/nonB/M
utation/NCNR.SFS.low.bed.test')
NCNR_ctrl_infile = open('/nfs/brubeck.bx.psu.edu/scratch5/wilfried/nonB/
Mutation/NCNR.SFS.control.shuf')

NCNR_high_infile.seek(0)
NCNR_low_infile.seek(0)
NCNR_ctrl_infile.seek(0)

def list_MAFs(infile):

    already_parsed = {}
    MAFs = []
    for line in infile:
        array = line.strip().split('\t')
        key = array[0]+'|'+array[1]+'|'+array[2]
        #print(key)
        #break
        SFS = array[3]
        MAF = SFS.split('|')[4]


        if MAF != 'NA':
            if key not in already_parsed:

                if float(MAF) < 1: #remove fixed
                    MAF = float(MAF)

                    if MAF > 0.5:
                        if MAF >1:
                            print('ISSUE')
                            break
                            MAF = 1
                        MAF = 1-MAF

                    if float(MAF) > 0.001792115*2: #remove single- and d
oubletons
                        MAFs.append(MAF)

        already_parsed[key] = 0

    return(MAFs)

MAFs_high = list_MAFs(NCNR_high_infile)
MAFs_low = list_MAFs(NCNR_low_infile)
MAFs_ctrl = list_MAFs(NCNR_ctrl_infile)
```

In [249]:
```
%%R -i MAFs_ctrl,MAFs_high,MAFs_low

MAFs_high <- as.numeric(as.character(MAFs_high))
MAFs_low <- as.numeric(as.character(MAFs_low))
MAFs_ctrl <- as.numeric(as.character(MAFs_ctrl))
MAFs <- c(MAFs_high,MAFs_low)

print(min(MAFs_high))
print(min(MAFs_low))
print(min(MAFs_ctrl))
```

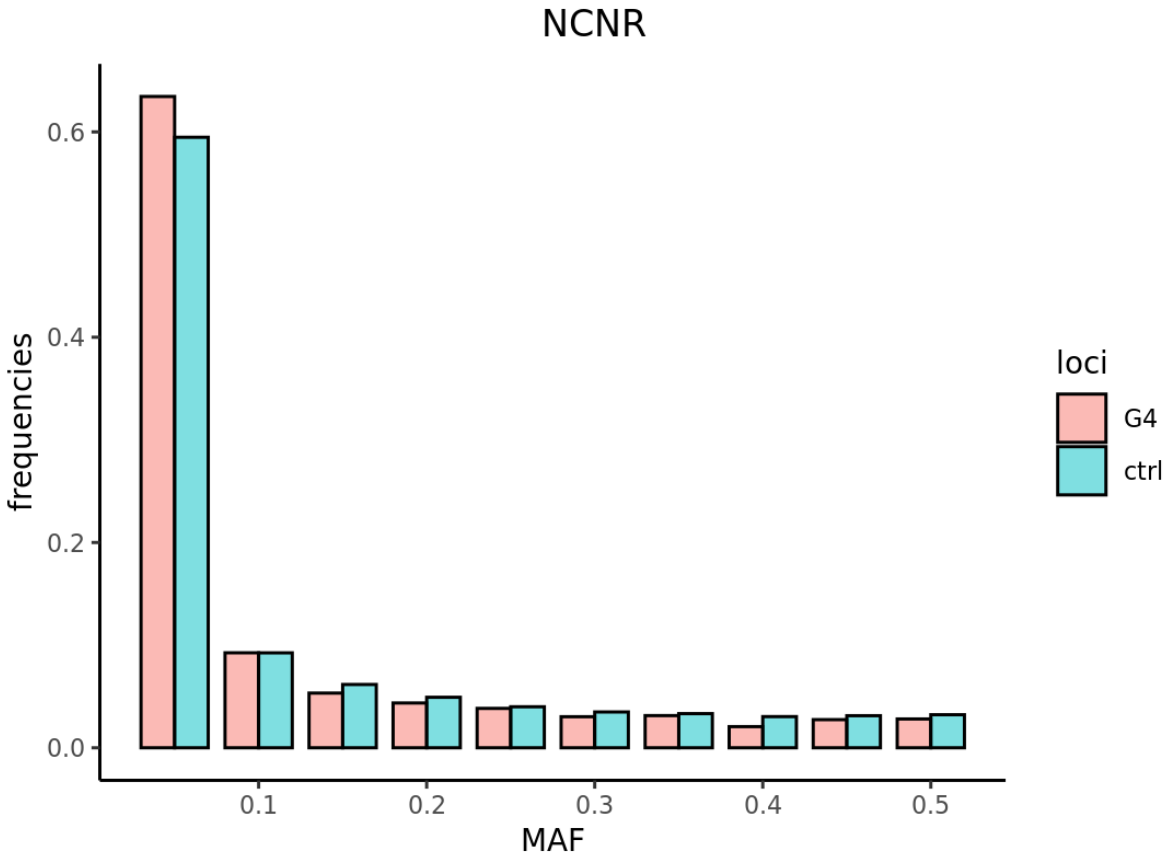
```
[1] 0.005376344
[1] 0.005376344
[1] 0.005376344
```

In [250]:
```
%%R

KS <- ks.test(MAFs_high,MAFs_ctrl)
print(KS)

KS <- ks.test(MAFs_low,MAFs_ctrl)
print(KS)

KS <- ks.test(MAFs,MAFs_ctrl)
print(KS)
```

```
        Two-sample Kolmogorov-Smirnov test

data:  MAFs_high and MAFs_ctrl
D = 0.067592, p-value = 1.036e-09
alternative hypothesis: two-sided


        Two-sample Kolmogorov-Smirnov test

data:  MAFs_low and MAFs_ctrl
D = 0.031038, p-value = 0.04075
alternative hypothesis: two-sided


        Two-sample Kolmogorov-Smirnov test

data:  MAFs and MAFs_ctrl
D = 0.044978, p-value = 9.06e-08
alternative hypothesis: two-sided
```

In [251]:
```
%%R -w 5.5 -h 4 --units in -r 200

library(scales)
library(plotrix)


hist_high <- hist(MAFs, plot=FALSE, breaks = 10)$density / sum(hist(MAFs
, plot=FALSE, breaks = 10)$density)
hist_ctrl <- hist(MAFs_ctrl, plot=FALSE, breaks = 10)$density / sum(hist
(MAFs_ctrl, plot=FALSE, breaks = 10)$density)
x <- seq(0.05,0.5,0.05)

toplot <- cbind.data.frame(c(x,x),c(hist_high,hist_ctrl),c(rep('G4',leng
th(hist_high)),rep('ctrl',length(hist_ctrl))))
colnames(toplot) <- c('x','MAF', 'loci')
print(head(toplot))

p1 <- ggplot(toplot) +
    geom_bar(aes(fill = loci, x=x, y=MAF), position = "dodge", stat="ide
ntity", colour="black", width = 0.04, alpha = .5) +
    theme_classic() +
    scale_x_continuous(breaks=seq(0,0.5,0.1)) +
    labs(y="frequencies", x="MAF") +
    theme(text=element_text(size=10,family="sans"))

grid.arrange(p1,top=textGrob("NCNR",gp=gpar(fontsize=12,font=1)))
```

```
     x         MAF loci
1 0.05 0.63450998   G4
2 0.10 0.09252357   G4
3 0.15 0.05327779   G4
4 0.20 0.04363078   G4
5 0.25 0.03836878   G4
6 0.30 0.03025652   G4
```



In [252]:
```
HR_high_infile = open('/nfs/brubeck.bx.psu.edu/scratch5/wilfried/nonB/Mutation/NCNR.highrecomb.SFS.high.bed.test')
HR_low_infile = open('/nfs/brubeck.bx.psu.edu/scratch5/wilfried/nonB/Mutation/NCNR.highrecomb.SFS.control')
HR_ctrl_infile = open('/nfs/brubeck.bx.psu.edu/scratch5/wilfried/nonB/Mutation/NCNR.highrecomb.SFS.low.bed.test')

HR_MAFs_high = list_MAFs(HR_high_infile)
HR_MAFs_low = list_MAFs(HR_low_infile)
HR_MAFs_ctrl = list_MAFs(HR_ctrl_infile)
```

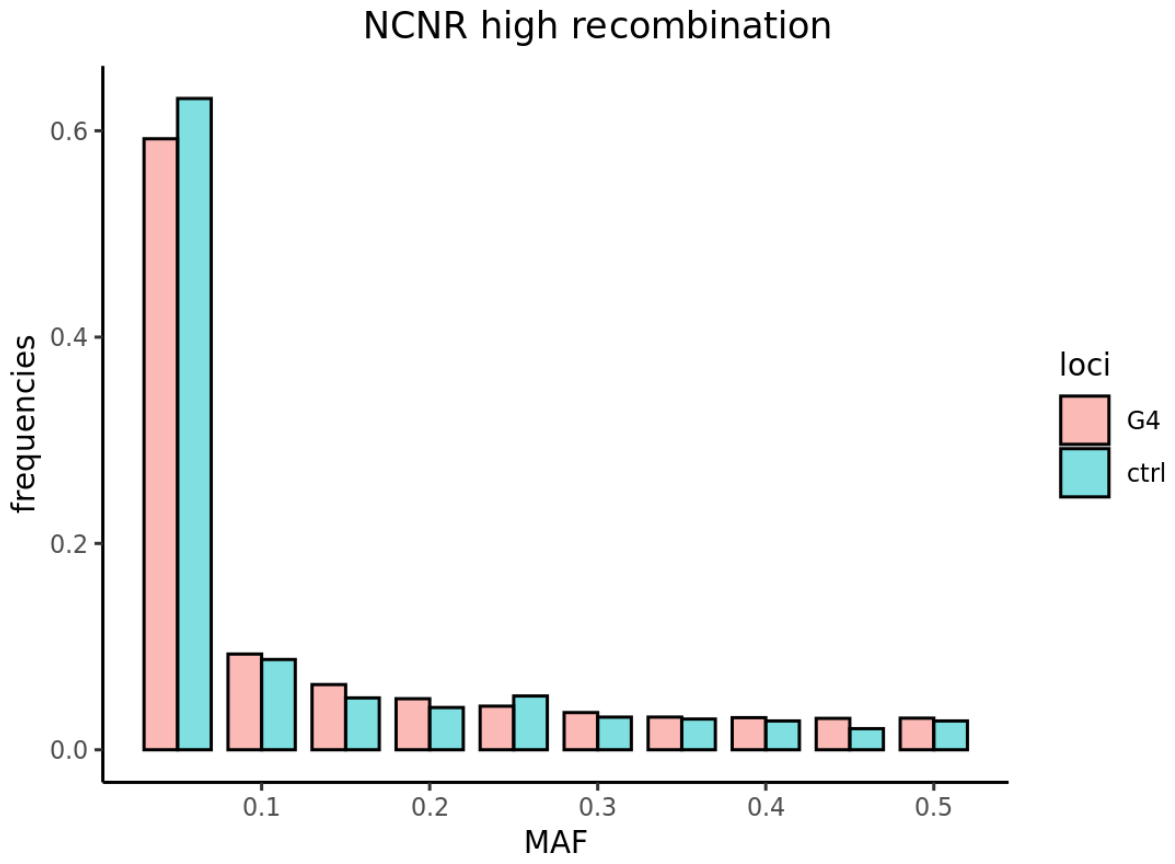
In [253]:
```
%%R -i HR_MAFs_ctrl,HR_MAFs_high,HR_MAFs_low

HR_MAFs_high <- as.numeric(as.character(HR_MAFs_high))
HR_MAFs_low <- as.numeric(as.character(HR_MAFs_low))
HR_MAFs_ctrl <- as.numeric(as.character(HR_MAFs_ctrl))
HR_MAFs <- c(HR_MAFs_high,HR_MAFs_low)
```

In [254]:
```
%%R

KS <- ks.test(HR_MAFs_high,HR_MAFs_ctrl)
print(KS)

KS <- ks.test(HR_MAFs_low,HR_MAFs_ctrl)
print(KS)

KS <- ks.test(HR_MAFs,HR_MAFs_ctrl)
print(KS)
```

```
        Two-sample Kolmogorov-Smirnov test

data:  HR_MAFs_high and HR_MAFs_ctrl
D = 0.050562, p-value = 0.4389
alternative hypothesis: two-sided


        Two-sample Kolmogorov-Smirnov test

data:  HR_MAFs_low and HR_MAFs_ctrl
D = 0.070888, p-value = 0.009198
alternative hypothesis: two-sided


        Two-sample Kolmogorov-Smirnov test

data:  HR_MAFs and HR_MAFs_ctrl
D = 0.070773, p-value = 0.00936
alternative hypothesis: two-sided
```

In [255]:
```
%%R -w 5.5 -h 4 --units in -r 200

library(scales)
library(plotrix)



hist_high <- hist(HR_MAFs, plot=FALSE, breaks = 10)$density / sum(hist(H
R_MAFs, plot=FALSE, breaks = 10)$density)
hist_ctrl <- hist(HR_MAFs_ctrl, plot=FALSE, breaks = 10)$density / sum(h
ist(HR_MAFs_ctrl, plot=FALSE, breaks = 10)$density)
x <- seq(0.05,0.5,0.05)

toplot <- cbind.data.frame(c(x,x),c(hist_high,hist_ctrl),c(rep('G4',leng
th(hist_high)),rep('ctrl',length(hist_ctrl))))
colnames(toplot) <- c('x','MAF', 'loci')
print(head(toplot))

p1 <- ggplot(toplot) +
    geom_bar(aes(fill = loci, x=x, y=MAF), position = "dodge", stat="ide
ntity", colour="black", width = 0.04, alpha = .5) +
    theme_classic() +
    scale_x_continuous(breaks=seq(0,0.5,0.1)) +
    labs(y="frequencies", x="MAF") +
    theme(text=element_text(size=10,family="sans"))

grid.arrange(p1,top=textGrob("NCNR high recombination",gp=gpar(fontsize=
12,font=1)))
```

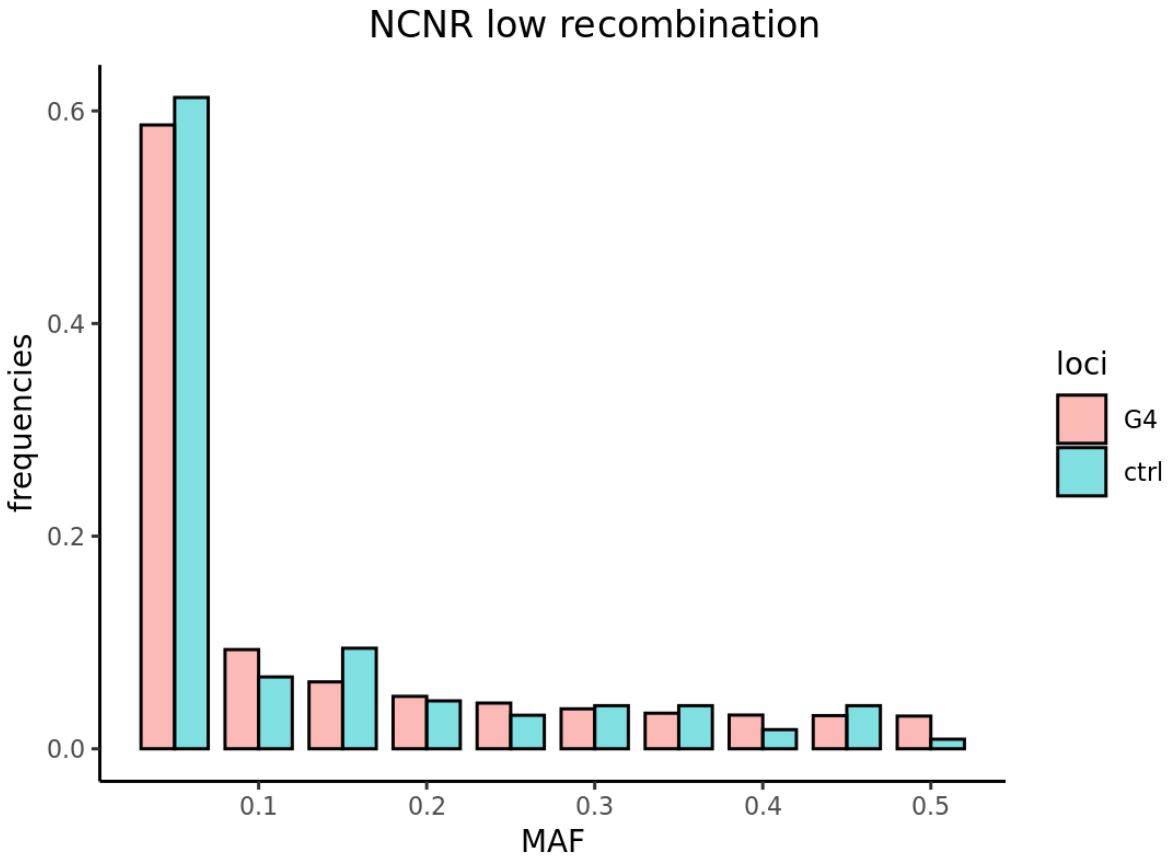
```
     x         MAF loci
1 0.05 0.59225280   G4
2 0.10 0.09278329   G4
3 0.15 0.06320602   G4
4 0.20 0.04949656   G4
5 0.25 0.04227257   G4
6 0.30 0.03607314   G4
```



In [256]:
```
LR_high_infile = open('/nfs/brubeck.bx.psu.edu/scratch5/wilfried/nonB/Mutation/NCNR.lowrecomb.SFS.high.bed.test')
LR_low_infile = open('/nfs/brubeck.bx.psu.edu/scratch5/wilfried/nonB/Mutation/NCNR.lowrecomb.SFS.control')
LR_ctrl_infile = open('/nfs/brubeck.bx.psu.edu/scratch5/wilfried/nonB/Mutation/NCNR.lowrecomb.SFS.low.bed.test')

LR_MAFs_high = list_MAFs(LR_high_infile)
LR_MAFs_low = list_MAFs(LR_low_infile)
LR_MAFs_ctrl = list_MAFs(LR_ctrl_infile)
```

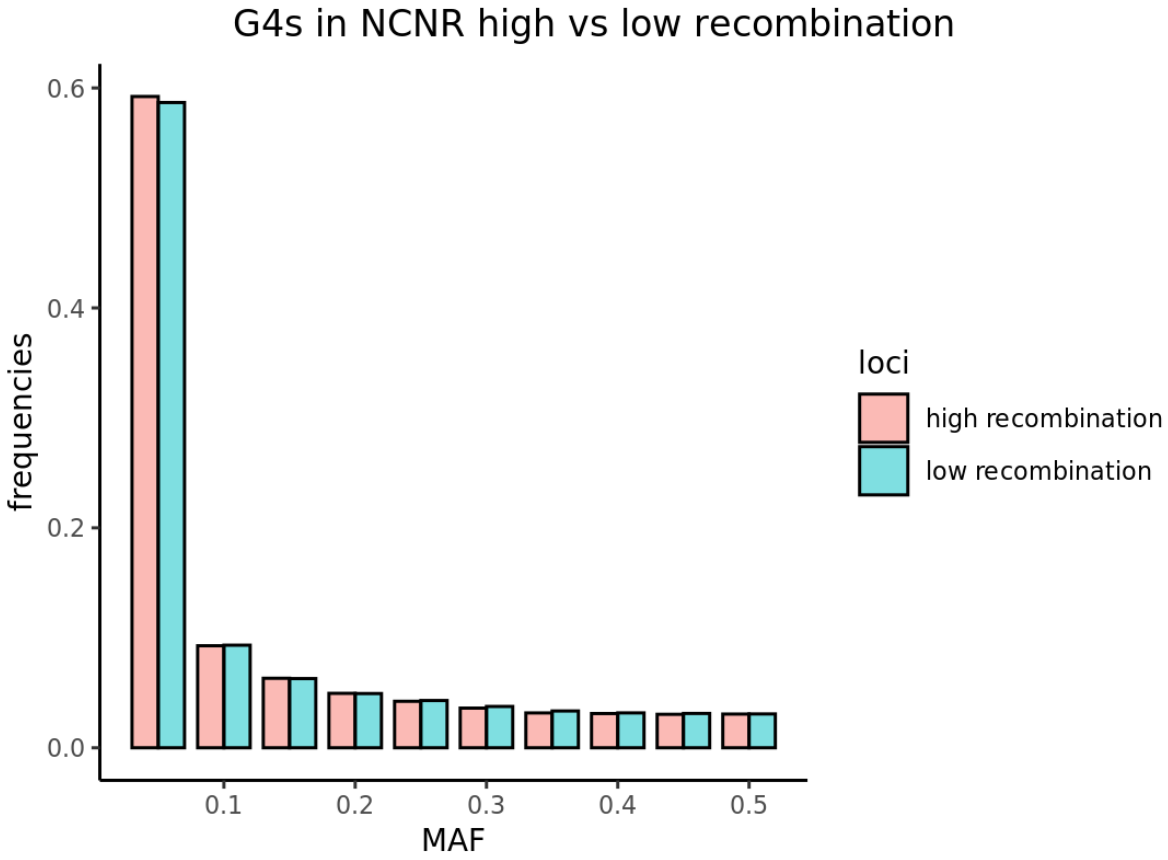
In [257]:
```
%%R -w 5.5 -h 4 --units in -r 200 -i LR_MAFs_ctrl,LR_MAFs_high,LR_MAFs_l
ow

LR_MAFs_high <- as.numeric(as.character(LR_MAFs_high))
LR_MAFs_low <- as.numeric(as.character(LR_MAFs_low))
LR_MAFs_ctrl <- as.numeric(as.character(LR_MAFs_ctrl))
LR_MAFs <- c(LR_MAFs_high,LR_MAFs_low)


KS <- ks.test(LR_MAFs_high,LR_MAFs_ctrl)
print(KS)

KS <- ks.test(LR_MAFs_low,LR_MAFs_ctrl)
print(KS)

KS <- ks.test(LR_MAFs,LR_MAFs_ctrl)
print(KS)


library(scales)
library(plotrix)


hist_high <- hist(LR_MAFs, plot=FALSE, breaks = 10)$density / sum(hist(L
R_MAFs, plot=FALSE, breaks = 10)$density)
hist_ctrl <- hist(LR_MAFs_ctrl, plot=FALSE, breaks = 10)$density / sum(h
ist(LR_MAFs_ctrl, plot=FALSE, breaks = 10)$density)
x <- seq(0.05,0.5,0.05)

toplot <- cbind.data.frame(c(x,x),c(hist_high,hist_ctrl),c(rep('G4',leng
th(hist_high)),rep('ctrl',length(hist_ctrl))))
colnames(toplot) <- c('x','MAF', 'loci')

p1 <- ggplot(toplot) +
    geom_bar(aes(fill = loci, x=x, y=MAF), position = "dodge", stat="ide
ntity", colour="black", width = 0.04, alpha = .5) +
    theme_classic() +
    scale_x_continuous(breaks=seq(0,0.5,0.1)) +
    labs(y="frequencies", x="MAF") +
    theme(text=element_text(size=10,family="sans"))

grid.arrange(p1,top=textGrob("NCNR low recombination",gp=gpar(fontsize=1
2,font=1)))
```

```
        Two-sample Kolmogorov-Smirnov test

data:  LR_MAFs_high and LR_MAFs_ctrl
D = 0.062632, p-value = 0.7237
alternative hypothesis: two-sided


        Two-sample Kolmogorov-Smirnov test

data:  LR_MAFs_low and LR_MAFs_ctrl
D = 0.044326, p-value = 0.778
alternative hypothesis: two-sided


        Two-sample Kolmogorov-Smirnov test

data:  LR_MAFs and LR_MAFs_ctrl
D = 0.043936, p-value = 0.7871
alternative hypothesis: two-sided
```



In [270]:
```
%%R -w 5.5 -h 4 --units in -r 200

KS <- ks.test(HR_MAFs,LR_MAFs)
print(KS)


library(scales)
library(plotrix)


hist_high <- hist(HR_MAFs, plot=FALSE, breaks = 10)$density / sum(hist(L
R_MAFs, plot=FALSE, breaks = 10)$density)
hist_ctrl <- hist(LR_MAFs, plot=FALSE, breaks = 10)$density / sum(hist(L
R_MAFs_ctrl, plot=FALSE, breaks = 10)$density)
x <- seq(0.05,0.5,0.05)

toplot <- cbind.data.frame(c(x,x),c(hist_high,hist_ctrl),c(rep('high rec
ombination',length(hist_high)),rep('low recombination',length(hist_ctrl
))))
colnames(toplot) <- c('x','MAF', 'loci')

p1 <- ggplot(toplot) +
    geom_bar(aes(fill = loci, x=x, y=MAF), position = "dodge", stat="ide
ntity", colour="black", width = 0.04, alpha = .5) +
    theme_classic() +
    scale_x_continuous(breaks=seq(0,0.5,0.1)) +
    labs(y="frequencies", x="MAF") +
    theme(text=element_text(size=10,family="sans"))

grid.arrange(p1,top=textGrob("G4s in NCNR high vs low recombination",gp=
gpar(fontsize=12,font=1)))
```

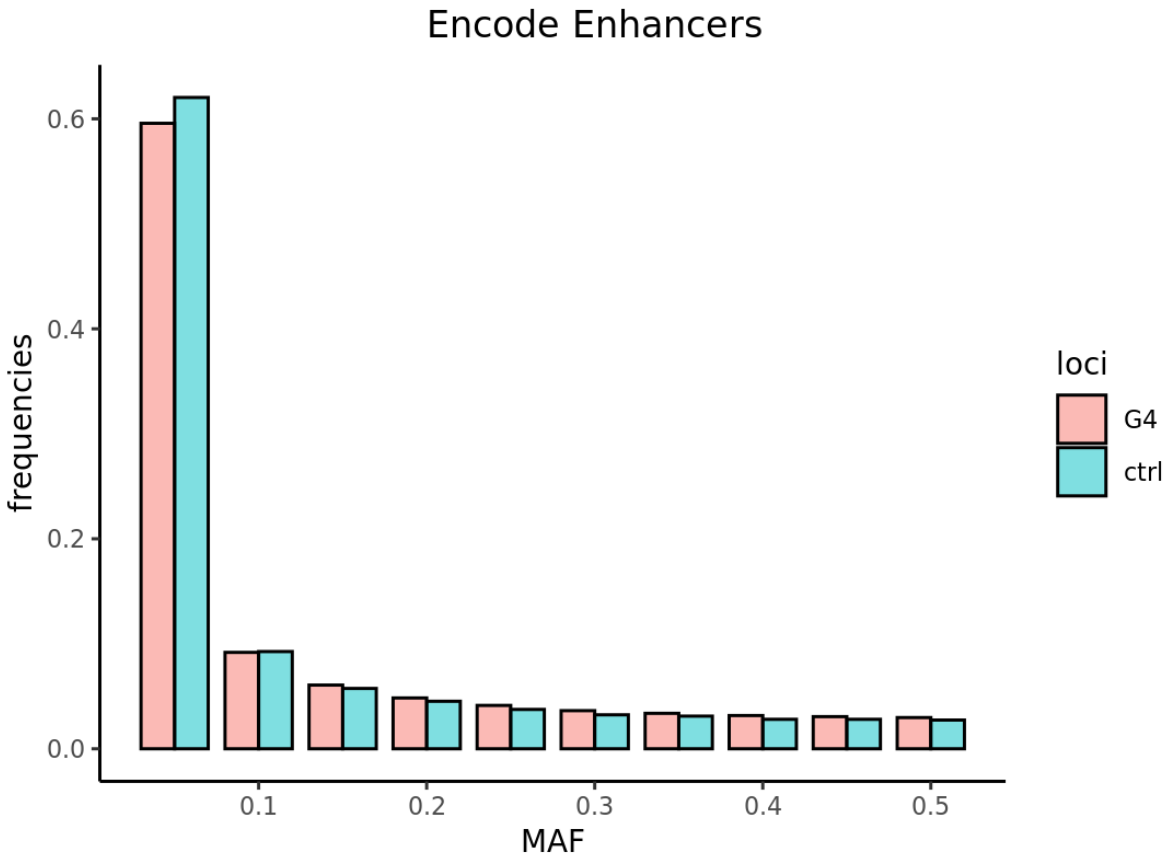
```
        Two-sample Kolmogorov-Smirnov test

data:  HR_MAFs and LR_MAFs
D = 0.0067221, p-value = 0.05468
alternative hypothesis: two-sided
```



In [258]:
```
ENCODE_high_infile = open('/nfs/brubeck.bx.psu.edu/scratch5/wilfried/nonB/Mutation/EncodeEnhancers.SFS.high.test')
ENCODE_low_infile = open('/nfs/brubeck.bx.psu.edu/scratch5/wilfried/nonB/Mutation/EncodeEnhancers.SFS.control')
ENCODE_ctrl_infile = open('/nfs/brubeck.bx.psu.edu/scratch5/wilfried/nonB/Mutation/EncodeEnhancers.SFS.low.test')

ENCODE_MAFs_high = list_MAFs(ENCODE_high_infile)
ENCODE_MAFs_low = list_MAFs(ENCODE_low_infile)
ENCODE_MAFs_ctrl = list_MAFs(ENCODE_ctrl_infile)
```

In [264]:
```
%%R -w 5.5 -h 4 --units in -r 200 -i ENCODE_MAFs_ctrl,ENCODE_MAFs_high,E
NCODE_MAFs_low

ENCODE_MAFs_high <- as.numeric(as.character(ENCODE_MAFs_high))
ENCODE_MAFs_low <- as.numeric(as.character(ENCODE_MAFs_low))
ENCODE_MAFs_ctrl <- as.numeric(as.character(ENCODE_MAFs_ctrl))
ENCODE_MAFs <- c(ENCODE_MAFs_high,ENCODE_MAFs_low)


KS <- ks.test(ENCODE_MAFs_high,ENCODE_MAFs_ctrl)
print(KS)

KS <- ks.test(ENCODE_MAFs_low,ENCODE_MAFs_ctrl)
print(KS)

KS <- ks.test(ENCODE_MAFs,ENCODE_MAFs_ctrl)
print(KS)


library(scales)
library(plotrix)


hist_high <- hist(ENCODE_MAFs, plot=FALSE, breaks = 10)$density / sum(hi
st(ENCODE_MAFs, plot=FALSE, breaks = 10)$density)
hist_ctrl <- hist(ENCODE_MAFs_ctrl, plot=FALSE, breaks = 10)$density / s
um(hist(ENCODE_MAFs_ctrl, plot=FALSE, breaks = 10)$density)
x <- seq(0.05,0.5,0.05)

toplot <- cbind.data.frame(c(x,x),c(hist_high,hist_ctrl),c(rep('G4',leng
th(hist_high)),rep('ctrl',length(hist_ctrl))))
colnames(toplot) <- c('x','MAF', 'loci')

p1 <- ggplot(toplot) +
    geom_bar(aes(fill = loci, x=x, y=MAF), position = "dodge", stat="ide
ntity", colour="black", width = 0.04, alpha = .5) +
    theme_classic() +
    scale_x_continuous(breaks=seq(0,0.5,0.1)) +
    labs(y="frequencies", x="MAF") +
    theme(text=element_text(size=10,family="sans"))

grid.arrange(p1,top=textGrob("Encode Enhancers",gp=gpar(fontsize=12,font
=1)))
```

```
        Two-sample Kolmogorov-Smirnov test

data:  ENCODE_MAFs_high and ENCODE_MAFs_ctrl
D = 0.037986, p-value = 2.542e-11
alternative hypothesis: two-sided


        Two-sample Kolmogorov-Smirnov test

data:  ENCODE_MAFs_low and ENCODE_MAFs_ctrl
D = 0.027673, p-value = 6.914e-09
alternative hypothesis: two-sided


        Two-sample Kolmogorov-Smirnov test

data:  ENCODE_MAFs and ENCODE_MAFs_ctrl
D = 0.026225, p-value = 5.013e-08
alternative hypothesis: two-sided
```



In [ ]:

In [130]:
```
%%R


Enhancers_vs_NCNR_G4_fisher <- fisher.test(cbind(c(823+306, 13572+6937),
c(1320+1369, 8153+10553)))
print(paste('Enhancers_vs_NCNR_G4',Enhancers_vs_NCNR_G4_fisher$estimate,
Enhancers_vs_NCNR_G4_fisher$p.value))
Enhancers_vs_NCNR_Ctrl_fisher <- fisher.test(cbind(c(24791, 582596),c(32
824, 276323)))
print(paste('Enhancers_vs_NCNR_Ctrl',Enhancers_vs_NCNR_Ctrl_fisher$estim
ate,Enhancers_vs_NCNR_Ctrl_fisher$p.value))
```

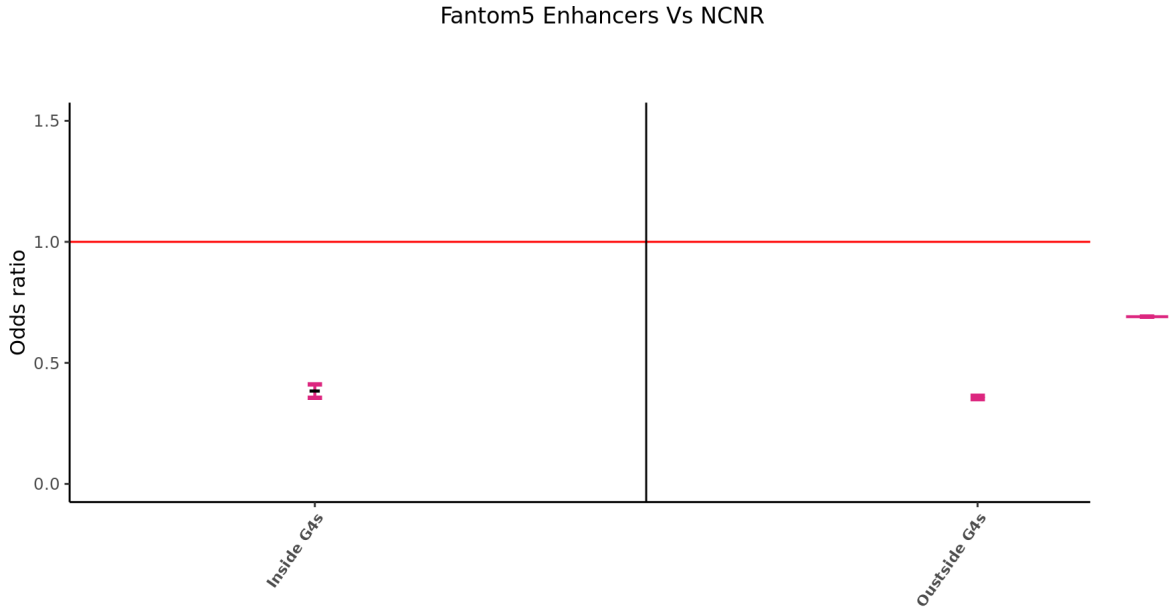
```
[1] "Enhancers_vs_NCNR_G4 0.382955035689816 2.08362232472756e-162"
[1] "Enhancers_vs_NCNR_Ctrl 0.358222031958436 0"
```

In [160]:
```
%%R
#https://stackoverflow.com/questions/14069629/how-can-i-plot-data-with-c
onfidence-intervals

x <- c(1,2)

#print(length(x))

ORs <- c(Enhancers_vs_NCNR_G4_fisher$estimate,
         Enhancers_vs_NCNR_Ctrl_fisher$estimate)

print(ORs)

Ls <- c(Enhancers_vs_NCNR_G4_fisher$conf.int[1],
        Enhancers_vs_NCNR_Ctrl_fisher$conf.int[1])

print(Ls)

Us <- c(Enhancers_vs_NCNR_G4_fisher$conf.int[2],
        Enhancers_vs_NCNR_Ctrl_fisher$conf.int[2])

print(Us)
```

```
odds ratio odds ratio
  0.382955   0.358222
[1] 0.3559354 0.3521443
[1] 0.4118281 0.3644054
```

In [173]:

```
%%R -w 9 -h 5 --units in -r 200



xlegend = c("Inside G4s","Oustside G4s")

dataf <- cbind.data.frame(xlegend,x,ORs,Ls,Us)


names(dataf) <- c("xlegend","x","ORs","Ls","Us")

cols <- c("all"="#DC267F",
          "high"="#648FFF",
          "low"="#FFB000")


g1 <- ggplot(dataf,) +
   geom_segment(aes(x=(x), xend=(x), y=Ls, yend=Us, colour="all"), size=
0.75) +

   geom_point(aes(x=(x), y=ORs),shape='-', size=7) +
   geom_point(aes(x=(x), y=Us, colour="all"),shape='-', size=10) +
   geom_point(aes(x=(x), y=Ls, colour="all"),shape='-', size=10) +

   geom_hline(yintercept=1, color="red") +
   geom_vline(xintercept=1.5, color="black") +
   geom_vline(xintercept=2.5, color="black") +
   geom_vline(xintercept=3.5, color="black") +
   geom_vline(xintercept=4.5, color="black") +
   geom_vline(xintercept=5.5, color="black") +



   theme_classic() + labs(subtitle="Classic Theme")+
   scale_x_continuous(limits = c(0.7,2.1),  breaks=c(1,2),labels=xlegend
) +
   scale_y_continuous(limits = c(0,1.5)) +
   theme(axis.text.x=element_text(angle = 55, hjust = 1,family="sans",fa
ce="bold",size=8)) +
   scale_colour_manual(name="",values=cols,labels="") +
   labs(title="", subtitle="", y="Odds ratio", x="", caption="",family=
"sans")




grid.arrange(g1,nrow=1,top=textGrob("Fantom5 Enhancers Vs NCNR",gp=gpar(
fontsize=12,font=1)))
G <- arrangeGrob(g1,g2,nrow=2)
#ggsave(file='HKA.pdf',G, width=9, height=10, dpi=300)
```

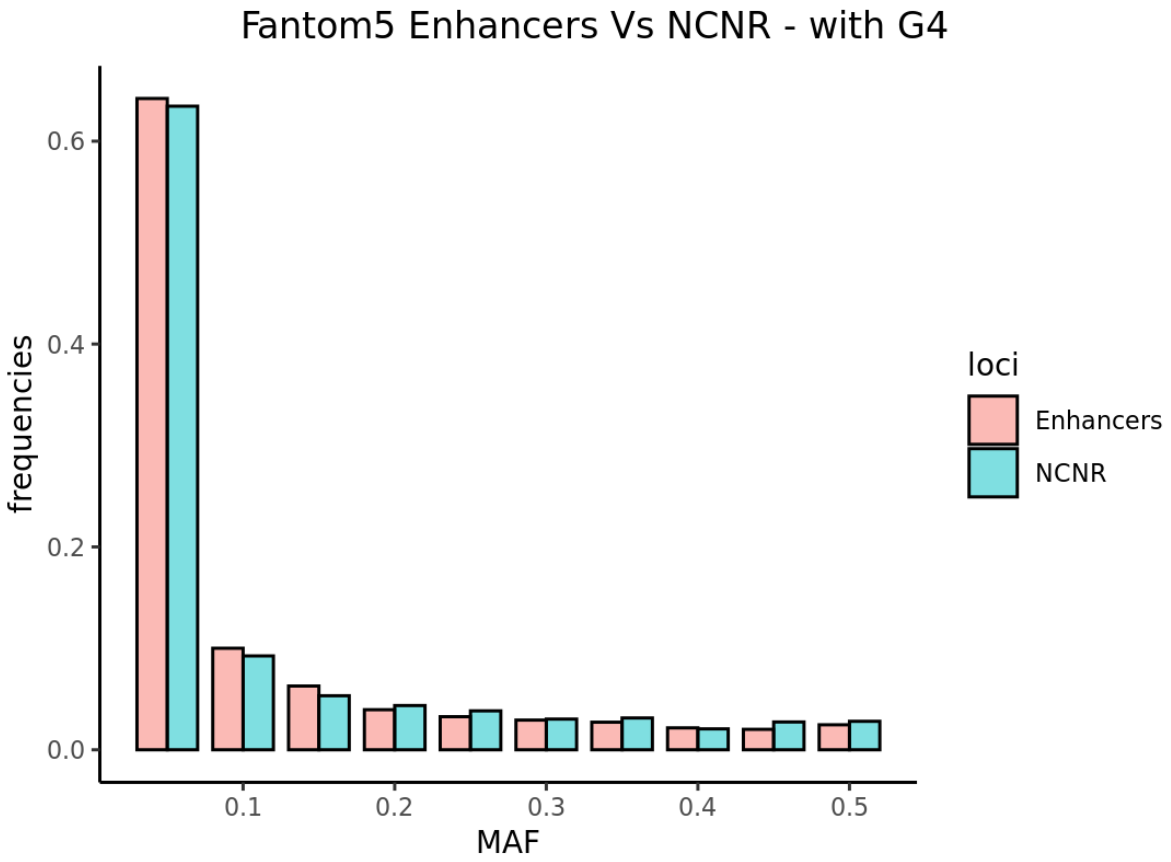


In [265]:
```
%%bash

cat /nfs/brubeck.bx.psu.edu/scratch5/wilfried/nonB/Mutation/enhancers.SFS.high.test /nfs/brubeck.bx.psu.edu/scratch5/wilfried/nonB/Mutation/enhancers.SFS.low.test > enhancers.SFS.test
cat /nfs/brubeck.bx.psu.edu/scratch5/wilfried/nonB/Mutation/NCNR.SFS.high.bed.test /nfs/brubeck.bx.psu.edu/scratch5/wilfried/nonB/Mutation/NCNR.SFS.low.bed.test > NCNR.SFS.test
```

In [266]:
```
G4_Enhancers_infile = open('enhancers.SFS.test')
G4_NCNR_infile = open('NCNR.SFS.test')

G4_Enhancers = list_MAFs(G4_Enhancers_infile)
G4_NCNR = list_MAFs(G4_NCNR_infile)
```

In [267]:
```
%%R -w 5.5 -h 4 --units in -r 200 -i G4_Enhancers,G4_NCNR

G4_Enhancers <- as.numeric(as.character(G4_Enhancers))
G4_NCNR <- as.numeric(as.character(G4_NCNR))

KS <- ks.test(G4_Enhancers,G4_NCNR)
print(KS)


library(scales)
library(plotrix)


hist_high <- hist(G4_Enhancers, plot=FALSE, breaks = 10)$density / sum(h
ist(G4_Enhancers, plot=FALSE, breaks = 10)$density)
hist_ctrl <- hist(G4_NCNR, plot=FALSE, breaks = 10)$density / sum(hist(G
4_NCNR, plot=FALSE, breaks = 10)$density)
x <- seq(0.05,0.5,0.05)

toplot <- cbind.data.frame(c(x,x),c(hist_high,hist_ctrl),c(rep('Enhancer
s',length(hist_high)),rep('NCNR',length(hist_ctrl))))
colnames(toplot) <- c('x','MAF', 'loci')

p1 <- ggplot(toplot) +
    geom_bar(aes(fill = loci, x=x, y=MAF), position = "dodge", stat="ide
ntity", colour="black", width = 0.04, alpha = .5) +
    theme_classic() +
    scale_x_continuous(breaks=seq(0,0.5,0.1)) +
    labs(y="frequencies", x="MAF") +
    theme(text=element_text(size=10,family="sans"))

grid.arrange(p1,top=textGrob("Fantom5 Enhancers Vs NCNR - with G4",gp=gp
ar(fontsize=12,font=1)))
```

```
        Two-sample Kolmogorov-Smirnov test

data:  G4_Enhancers and G4_NCNR
D = 0.026094, p-value = 0.05512
alternative hypothesis: two-sided
```

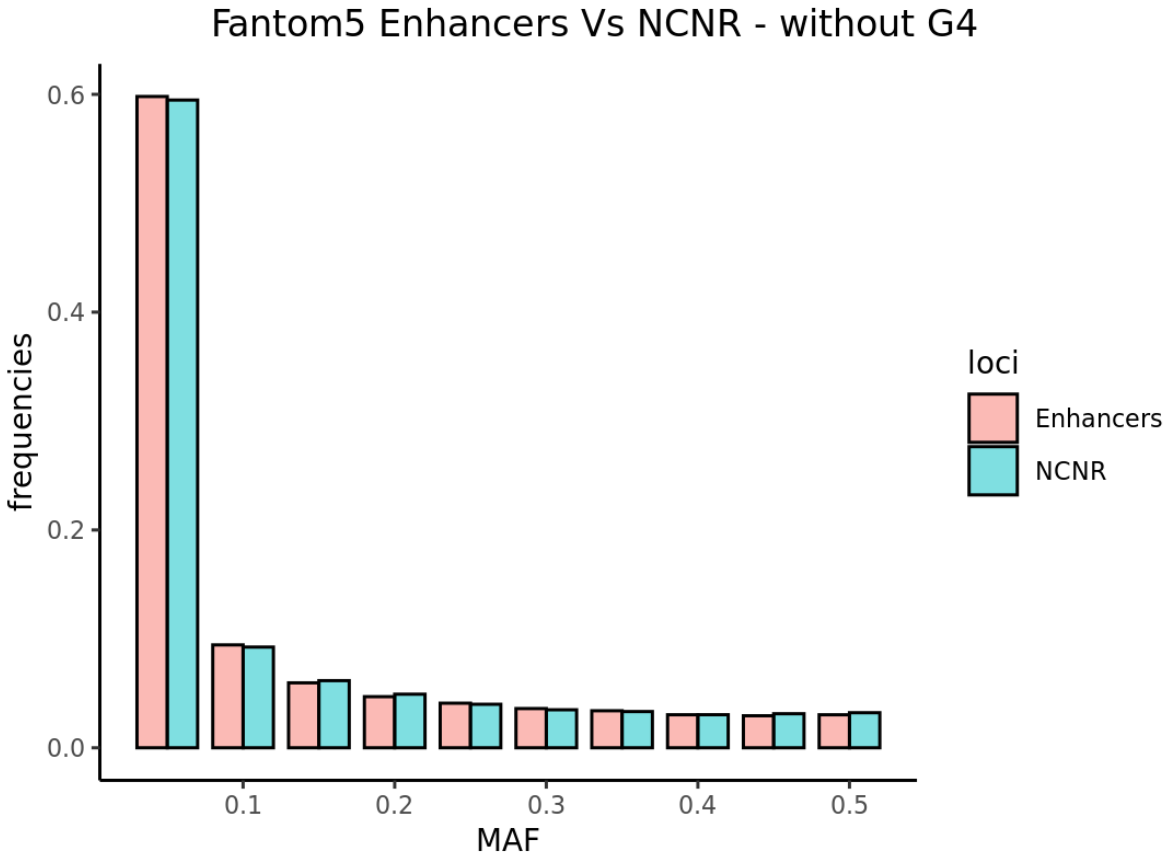


In [268]:
```
NOG4_Enhancers_infile = open('/nfs/brubeck.bx.psu.edu/scratch5/wilfried/
nonB/Mutation/enhancers.SFS.control')
NOG4_NCNR_infile = open('/nfs/brubeck.bx.psu.edu/scratch5/wilfried/nonB/
Mutation/NCNR.SFS.control.shuf')

NOG4_Enhancers = list_MAFs(NOG4_Enhancers_infile)
NOG4_NCNR = list_MAFs(NOG4_NCNR_infile)
```

In [269]:

```
%%R -w 5.5 -h 4 --units in -r 200 -i NOG4_Enhancers,NOG4_NCNR

NOG4_Enhancers <- as.numeric(as.character(NOG4_Enhancers))
NOG4_NCNR <- as.numeric(as.character(NOG4_NCNR))

KS <- ks.test(NOG4_Enhancers,NOG4_NCNR)
print(KS)


library(scales)
library(plotrix)


hist_high <- hist(NOG4_Enhancers, plot=FALSE, breaks = 10)$density / sum
(hist(NOG4_Enhancers, plot=FALSE, breaks = 10)$density)
hist_ctrl <- hist(NOG4_NCNR, plot=FALSE, breaks = 10)$density / sum(hist
(NOG4_NCNR, plot=FALSE, breaks = 10)$density)
x <- seq(0.05,0.5,0.05)

toplot <- cbind.data.frame(c(x,x),c(hist_high,hist_ctrl),c(rep('Enhancer
s',length(hist_high)),rep('NCNR',length(hist_ctrl))))
colnames(toplot) <- c('x','MAF', 'loci')

p1 <- ggplot(toplot) +
    geom_bar(aes(fill = loci, x=x, y=MAF), position = "dodge", stat="ide
ntity", colour="black", width = 0.04, alpha = .5) +
    theme_classic() +
    scale_x_continuous(breaks=seq(0,0.5,0.1)) +
    labs(y="frequencies", x="MAF") +
    theme(text=element_text(size=10,family="sans"))

grid.arrange(p1,top=textGrob("Fantom5 Enhancers Vs NCNR - without G4",gp
=gpar(fontsize=12,font=1)))
```

```
        Two-sample Kolmogorov-Smirnov test

data:  NOG4_Enhancers and NOG4_NCNR
D = 0.0082955, p-value = 0.01815
alternative hypothesis: two-sided
```



In [ ]:

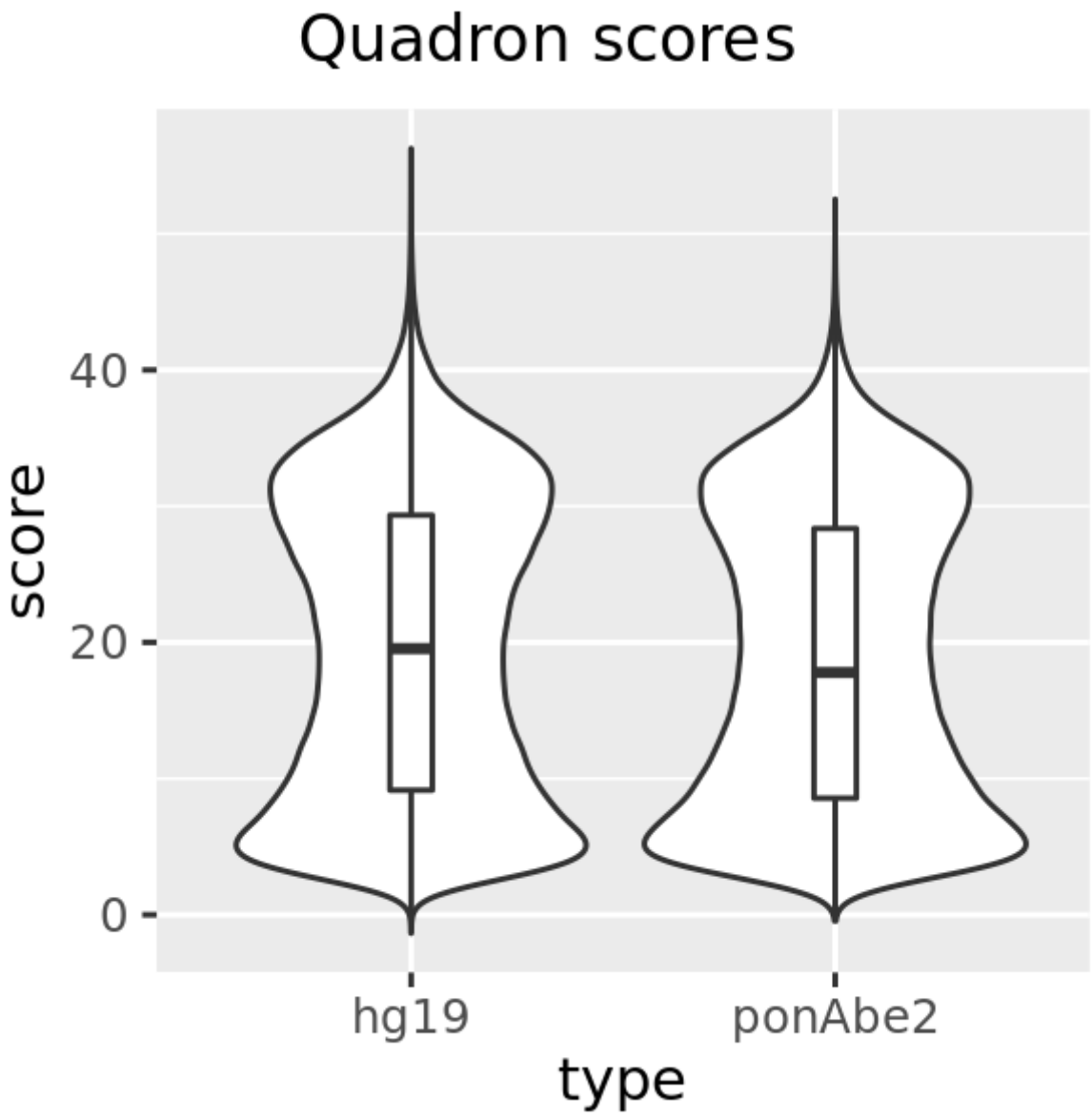
In [ ]:

In [ ]:

In [ ]:

```
python MergeAndFilter.py #Merge quadron annotations into one BED file

 liftOver ponAbe2.quadron.bed ponAbe2ToHg19.over.chain.gz ponAbe2.lifted.hg1
9.quadron.bed


cut -f 7 hs37d5.quadron.bed > hg19scores
cut -f 7 ponAbe2.lifted.hg19.quadron.bed > ponAbe2scores
```

In [188]:
```
%%R -w 3 -h 3 --units in -r 200

scores <- read.csv('hg19scores', header=FALSE)
alt_scores <- read.csv('ponAbe2scores', header=FALSE)

refs <- cbind(scores, rep('hg19', length(scores)))
alts <- cbind(alt_scores, rep('ponAbe2', length(alt_scores)))

colnames(refs) <- c('score', 'type')
colnames(alts) <- c('score', 'type')

#print(head(refs))
#print(head(alts))

df<- rbind.data.frame(refs,alts)



colnames(df) <- c('score', 'type')

#print(head(df))



library(ggplot2)
library(grid)
library(gridExtra)

p1 <- ggplot(df, aes(x=type,y=score)) + geom_violin() + geom_boxplot(wid
th=0.1)


grid.arrange(p1, ncol=1,
             top=textGrob("Quadron scores",gp=gpar(fontsize=12,family="sa
ns")))
```



In [189]:
```
%%R

t.test(scores,alt_scores)
```

```
        Welch Two Sample t-test

data:  scores and alt_scores
t = 47.597, df = 1123355, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.9184298 0.9973174
sample estimates:
mean of x mean of y
 19.55430  18.59643
```

In [216]:
```
original = open('hs37d5.quadron.bed', 'rt')
alternative = open('ponAbe2.lifted.hg19.quadron.bed', 'rt')

G4s = {}

original_count = 0

for line in original:
    original_count += 1
    array = line.strip().split('\t')
    chrom, start, end, motif, length, strand, score = array

    key = chrom+'|'+start+'|'+end

    G4s[key] = score


alternative_count = 0

Matched_G4s = {}

for line in alternative:
    alternative_count += 1
    array = line.strip().split('\t')
    chrom, start, end, motif, length, strand, score = array

    key = chrom+'|'+start+'|'+end

    if key in G4s:
        Matched_G4s[key] = G4s[key], score


print('Original G4s: '+str(original_count))
print('Alternative G4s: '+str(alternative_count))
print('Matched G4s: '+str(len(Matched_G4s)))
```

```
Original G4s: 670076
Alternative G4s: 516116
Matched G4s: 318864
```

In [223]:
```
score_increase = 0
score_decrease = 0
score_equal = 0

stable_to_unstable = 0
unstable_to_stable = 0
stay_stable = 0
stay_unstable = 0

count = 0

for G4s in Matched_G4s:

    try:
        #print(Matched_G4s[G4s])
        original_score = float(Matched_G4s[G4s][0])
        alternative_score = float(Matched_G4s[G4s][1])

    except ValueError as e:
        print('Error is: ' + str(e))
        print(Matched_G4s[G4s])

    if original_score > alternative_score:
        #print(Matched_G4s[G4s])
        score_decrease += 1

    if original_score < alternative_score:
        #print(Matched_G4s[G4s])
        score_increase += 1

    if original_score == alternative_score:
        score_equal += 1


    if original_score >= 19 and alternative_score < 19:
        stable_to_unstable += 1

    if original_score < 19 and alternative_score >= 19:
        unstable_to_stable += 1

    if original_score >= 19 and alternative_score >= 19:
        stay_stable += 1

    if original_score < 19 and alternative_score < 19:
        stay_unstable += 1

    count += 1


print("Increases: " + str(score_increase))
print("Decreases: "+ str(score_decrease))
print("Equals: "+ str(score_equal))
print("Stable to Unstable: "+ str(stable_to_unstable))
print("Unstable to Stable: "+ str(unstable_to_stable))
print("Remains Stable: "+ str(stay_stable))
```

```
print("Remains Unstable: "+ str(stay_unstable))
print(count)
```

```
Increases: 155199
Decreases: 154287
Equals: 9378
Stable to Unstable: 14723
Unstable to Stable: 13450
Remains Stable: 133144
Remains Unstable: 157547
318864
```

```
bedtools intersect -a hs37d5.quadron.bed -b ponAbe2.lifted.hg19.quadron.bed
 -f 0.9 -F 0.9 -e -wa -wb > original_alternative.intersect
```