# VariablesEffectingPerformance

*David A Hughes*

*27/09/2019*

## What are the variables that are influencing the performance of our experiments??

1. First how do we define and/or measure performance ?

```
+ the number of reliable reads: reads retained after QC and mapping filtering (secondary alignments and
+ enrichment factor (EF): (reliable-on-target reads / production reads ) / (target space/genomic space)
+ library complexity (LC): # reliable reads / total number of mapped reads, including duplicates
+ capture sensitivity (CS): # of target regions covered by 1 read / total number of target regions
+ capture specificity (CSp): reliable-on-target / reliable reads
```

2. What are some of the variable that we think may influence performance?

```
+ subspecies -- a proxy for environmental condition and sample quality
+ geogrpahic sampling site
+ sample
+ total DNA concentration
+ endogenous DNA concentration
+ % endogenous DNA
+ DNA fragment size
+ the sample pool it belongs to
+ the hybridization
+ production or sequencing reads acquired for a sample/hybridization
+ pipeting volume used to make a library (????)  -- this could be something to evaluate, but perhaps tha
+ pipeting volume used to make the pool  (????)
```

## Step1) Read in the data and report the variables available in each table

```
## $`Table S1`
## [1] "Sample"                          "Site"
## [3] "Subspecies"                      "Common name"
## [5] "Total DNA Concentration (ng/ul)" "Endogenous DNA (qPCR - ng/ul)"
## [7] "% Endogenous DNA"                "Average Fragment Size"
##
## $`Table S2`
## [1] "Sample"                          "Site"
## [3] "Subspecies"                      "Common name"
## [5] "Total DNA Concentration (ng/ul)" "Endogenous DNA (qPCR - pg/ul)"
## [7] "% Endogenous DNA"
##
## $`Table S3`
## [1] "Sites"              "Median Endogenous"  "Average Endogenous"
## [4] "Min"                "Max"
##
## $`Table S4`
##  [1] "Extract ID"
##  [2] "Sequencing Batch"
##  [3] "Capture Pool"
```

```
##  [4] "Starting DNA  (ug)"
##  [5] "Production Reads"
##  [6] "Production Bases"
##  [7] "Mapped Reads"
##  [8] "Percentage Mapped Reads"
##  [9] "Unique Reads"
## [10] "Percentage Unique Reads"
## [11] "Reliable Reads"
## [12] "Percentage Reliable Reads"
## [13] "OnTarget Reliable Reads"
## [14] "OnTarget Reliable Bases"
## [15] "Percentage OnTarget Reliable Reads"
## [16] "Percentage OnTarget Reliable Bases"
## [17] "Coverage OnTarget"
## [18] "Enrichment Factor (EF)"
## [19] "Capture Specificity (CSp)"
## [20] "Library Complexity (LC)"
## [21] "Capture Sensitivity (CS) DP1"
## [22] "Capture Sensitivity (CS) DP4"
## [23] "Capture Sensitivity (CS) DP10"
## [24] "Capture Sensitivity (CS) DP50"
##
## $`Table S5`
##  [1] "Capture Pool"
##  [2] "Production Reads"
##  [3] "Mapped Reads"
##  [4] "Percentage Mapped Reads"
##  [5] "Unique Reads"
##  [6] "Percentage Unique Reads"
##  [7] "Uniq HQ Reads"
##  [8] "Percentage Unique HQ Reads"
##  [9] "OnTarget Uniq HQ Reads"
## [10] "Percentage OnTargetUnique HQ Reads"
## [11] "Average Coverage OnTarget"
## [12] "...12"
##
## $Downsampled
##  [1] "Extract ID"
##  [2] "Sequencing Batch"
##  [3] "Capture Pool"
##  [4] "Starting DNA (ug)"
##  [5] "Production Reads"
##  [6] "Production Bases"
##  [7] "Mapped Reads"
##  [8] "Percentage Mapped Reads"
##  [9] "Unique Reads"
## [10] "Percentage Unique Reads"
## [11] "Reliable Reads"
## [12] "Percentage Reliable Reads"
## [13] "OnTarget Reliable Reads"
## [14] "OnTarget Reliable Bases"
## [15] "Percentage OnTarget Reliable Reads"
## [16] "Percentage OnTarget Reliable Bases"
## [17] "Coverage OnTarget"
```

```
## [18] "Enrichment"
## [19] "Specificity"
## [20] "LC"
## [21] "DP1"
## [22] "DP4"
## [23] "DP10"
## [24] "DP20"
## [25] "DP50"
```

## Step 2) Bring the necessary data together

– Add average fragment length from Table S1 to Table S2 –

– Add the data from Table S2 to Table S4 to produce the full data set together –

– Create a Pool variable –

– Add the data from Table S2 to Table Downsampled to produce the Down-Sampled data set together –

– Create a Pool variable –

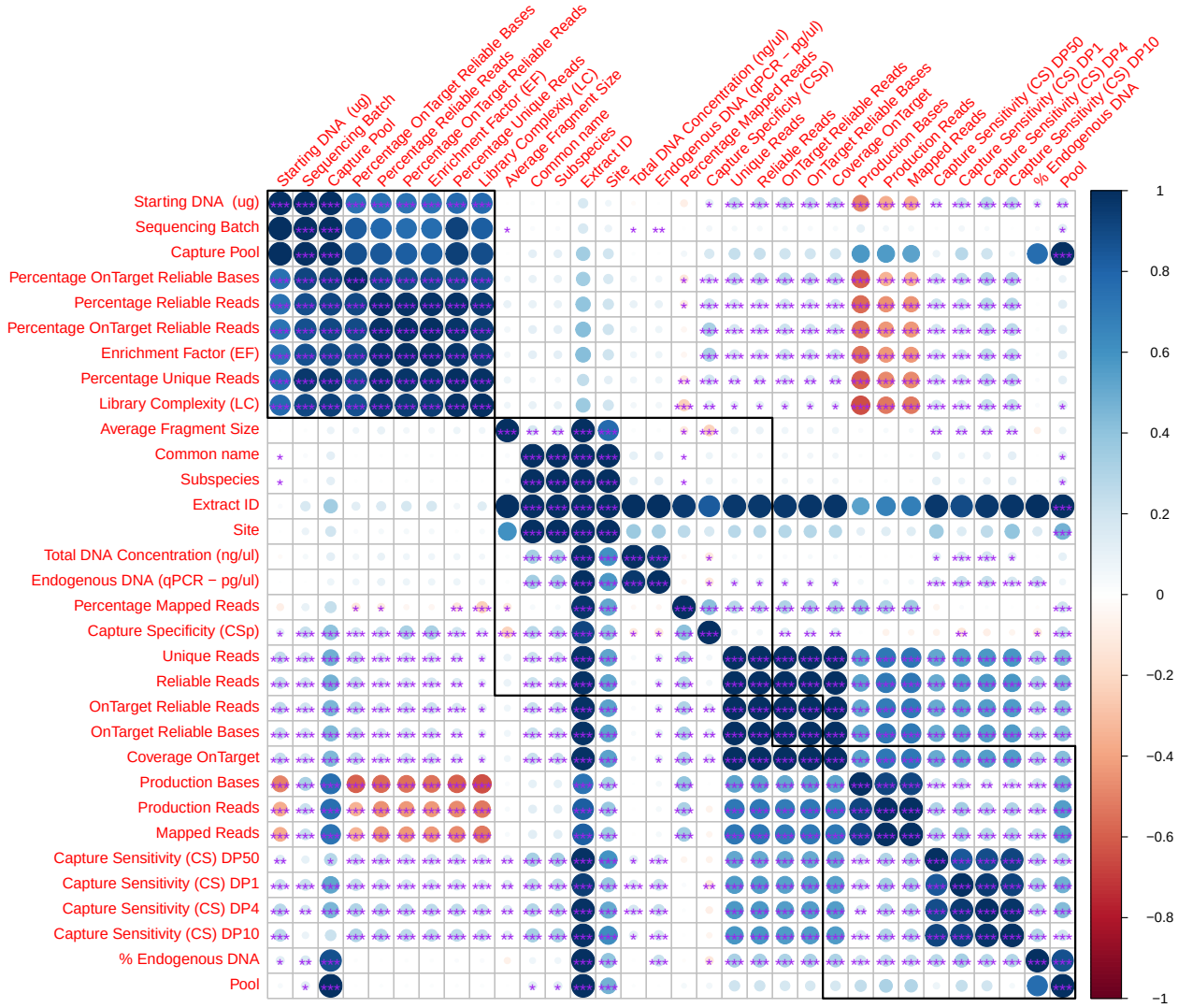
## Step 3) Correlation Analysis : everything on everything

– Using the funtions in the NILC R package, estimate a correlation matrix among all variables in the study –

– The correaltion estiamtes are based on (1) spearman's rho (N-on-N), (2) Carmer's V (F-on-F), and (3) univariate linear model (F-on-N) sqrt(R2) | sqrt(etasq)

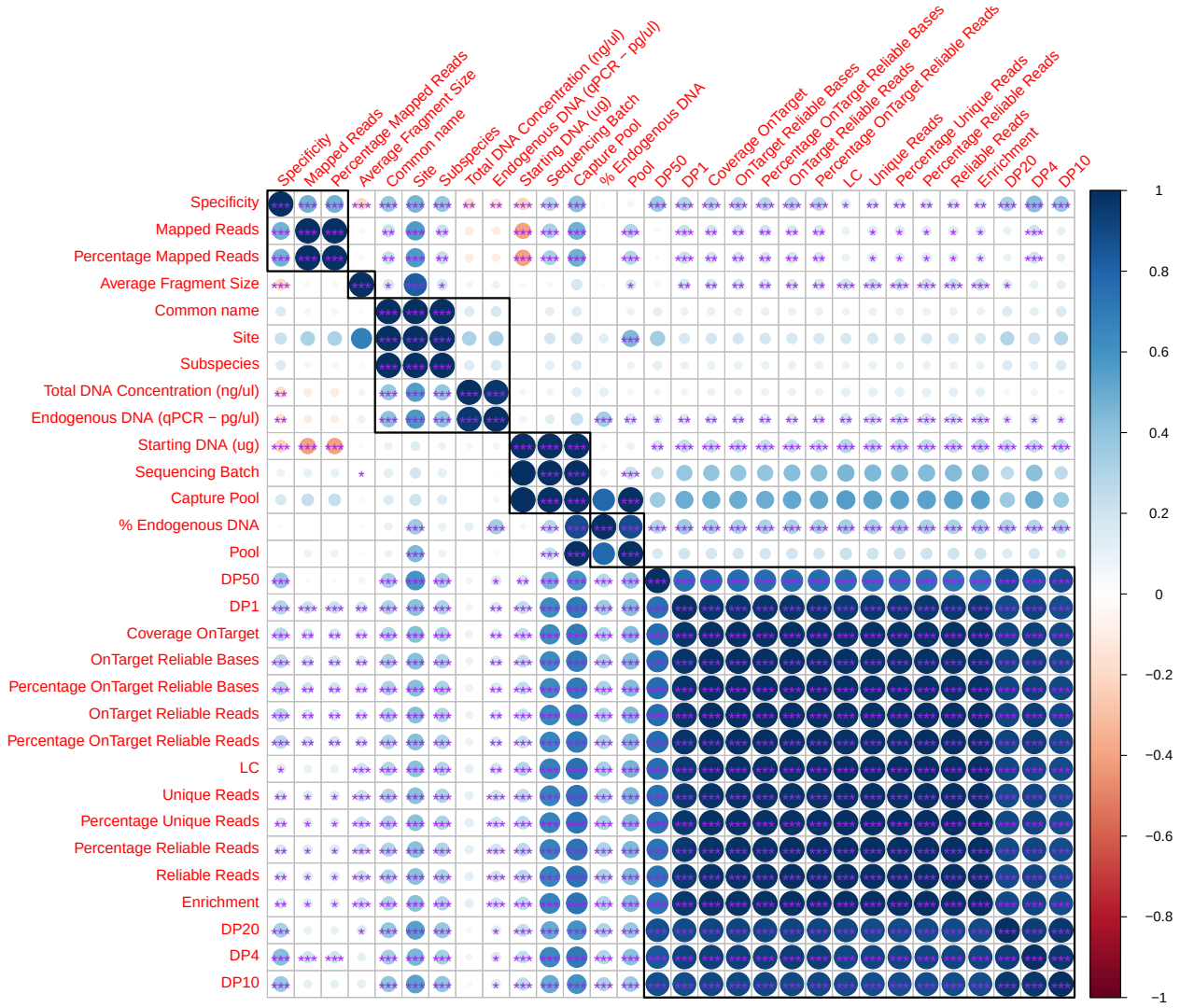
– Correlation matrix for the the *complete* data set –

```
## pdf
##   2
```



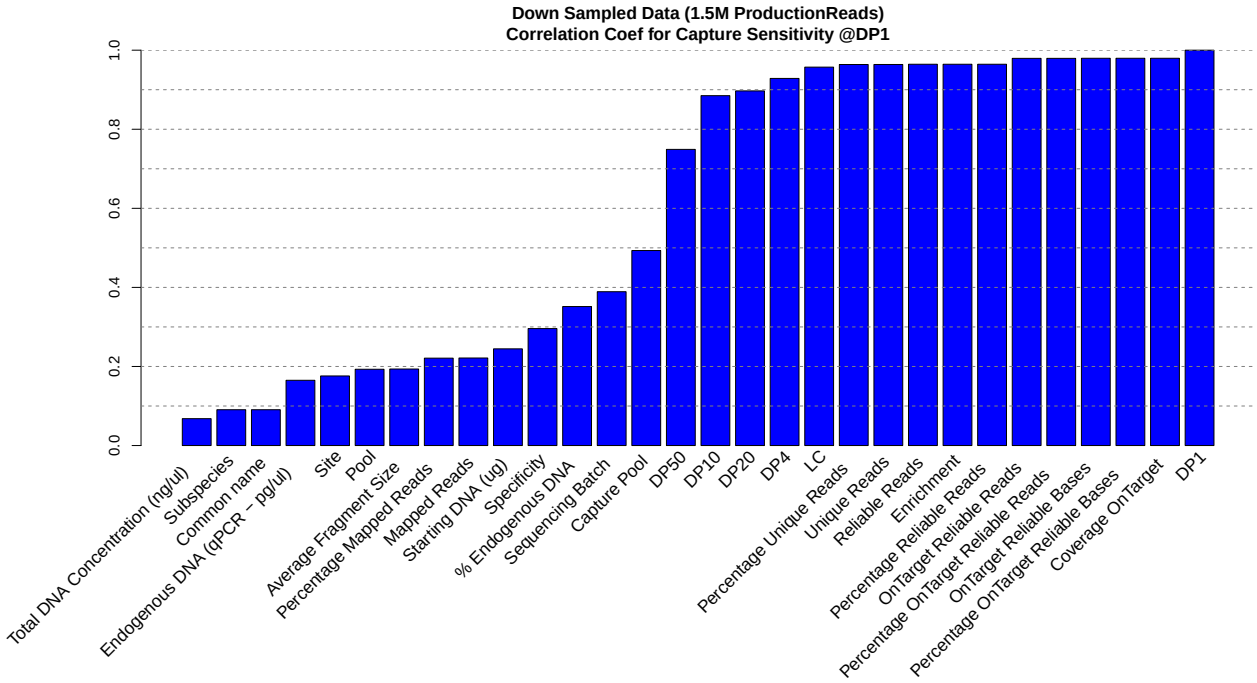

– Correlation matrix for the the *downsampled* data set –

```
## pdf
##   2
```




– **A closer look at univaraite *capture sensitivity @DP1* rho (correlation coefficents)** –

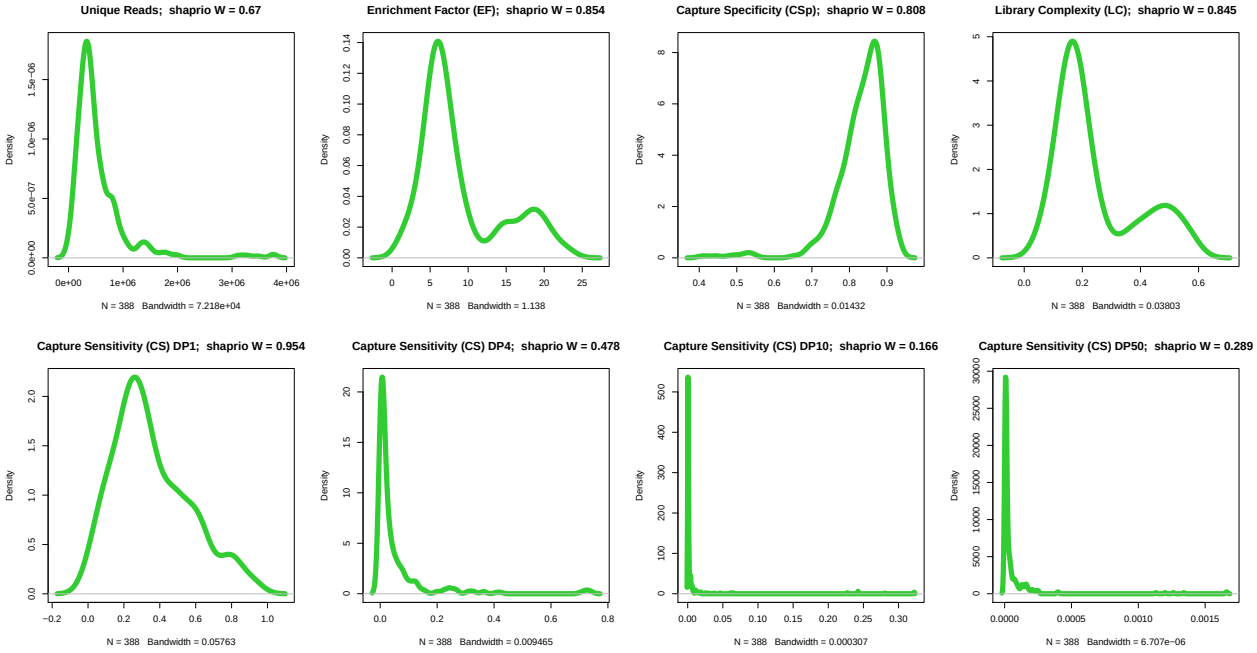
```
## pdf
##   2
```




## Step 4) Explicity Univariate linear modeling

– Distribution of summary statistics –

– Are the distributions normal?? –




– Analysis of the Complete data set –

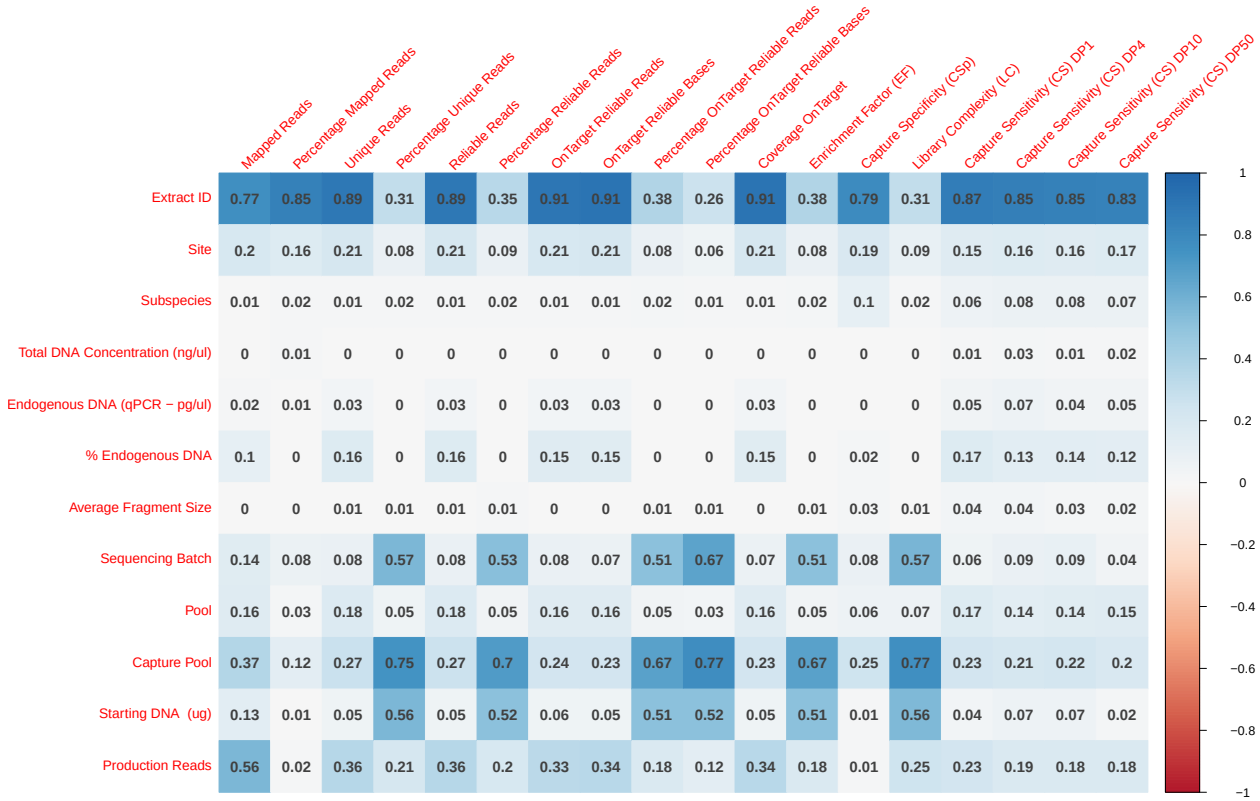
- mapped reads
- unique reads
- reliable reads

- EF : enrichment factor
- LC : library complexity
- CS : capture sensitivity
- CSp : capture specificity

– as influenced by site, DNA [concentration], %eDNA, fragment size, pool, amount of DNA in hybridaztion, hybridization, Sequencing run, production reads

```
## pdf
##   2
```

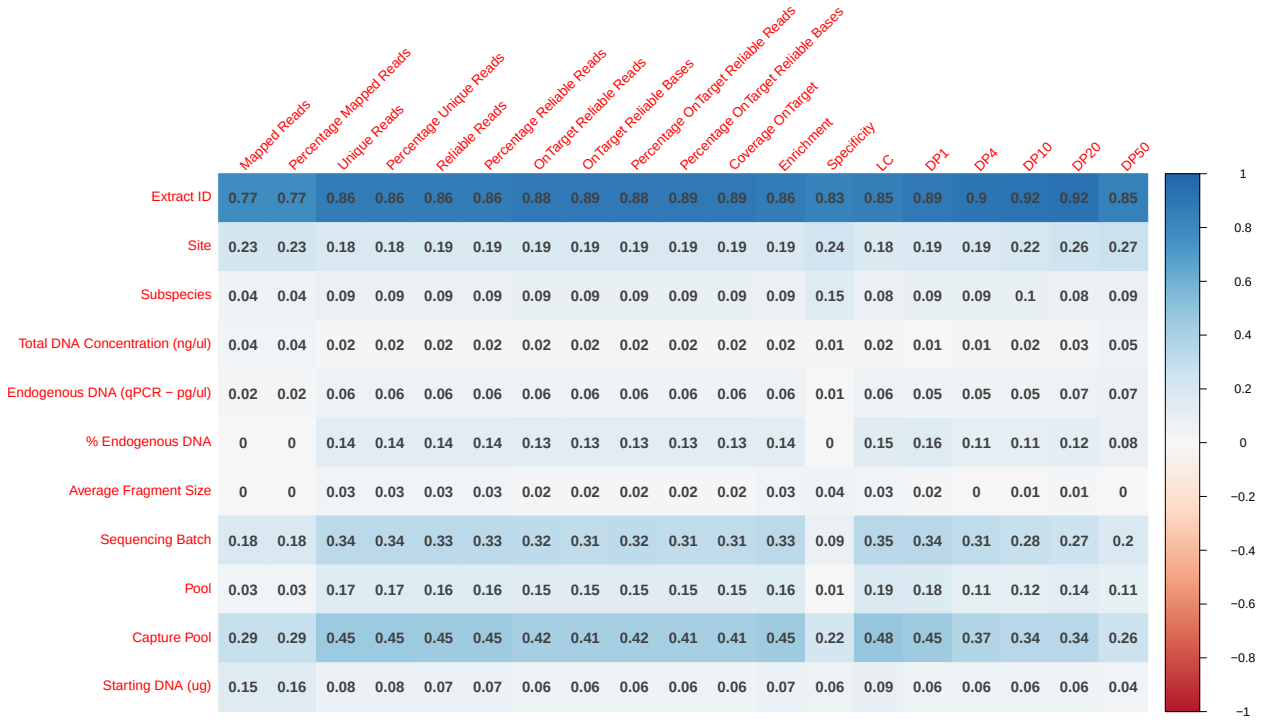
```
## pdf
##   2
```




– Analysis of the Down-sampled data set ( 1.5M Production reads samples only )–

```
## pdf
##   2
```

```
## pdf
##   2
```



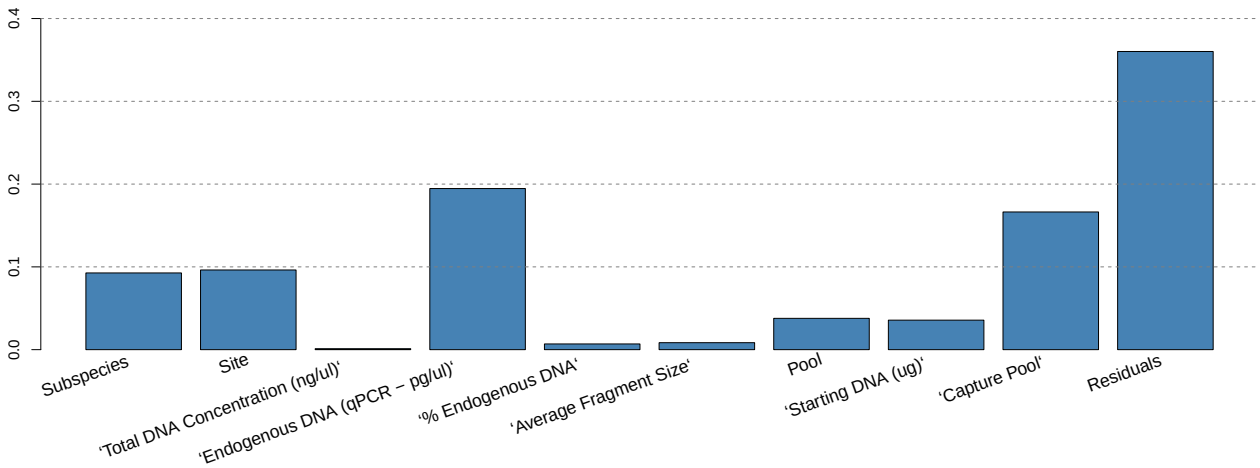

## Step 5) Multivariate Modeling of CS @ DP1

– Down Sampled Data –

– model : CS@DP1 ~ subspecies + site + [tDNA] + [eDNA] + %eDNA + fragmentsize + pool + amount_DNA-in-hyb + hybridization + seqbatch + error

```
w = which(downdata$`Production Reads` < 1500000) ## 274 samples left
#####
d = rntransform( unlist( downdata[-w, "DP1"] ) )
## sub setted model
# fit0 = lm( d ~ `Subspecies` + `Site` +
#               `Total DNA Concentration (ng/ul)` +
#               #`Endogenous DNA (qPCR - pg/ul)` +
#               `% Endogenous DNA` +
#               `Average Fragment Size` +
#               `Pool` +
#               `Starting DNA (ug)` +
#               `Capture Pool` +
#               `Sequencing Batch` , data = downdata[-w,])
#

## full model !!!
fit = lm( d ~ `Subspecies` + `Site` +
              `Total DNA Concentration (ng/ul)` +
              `Endogenous DNA (qPCR - pg/ul)` +
              `% Endogenous DNA` +
              `Average Fragment Size` +
              `Pool` +
             `Starting DNA (ug)` +
              `Capture Pool` +
```

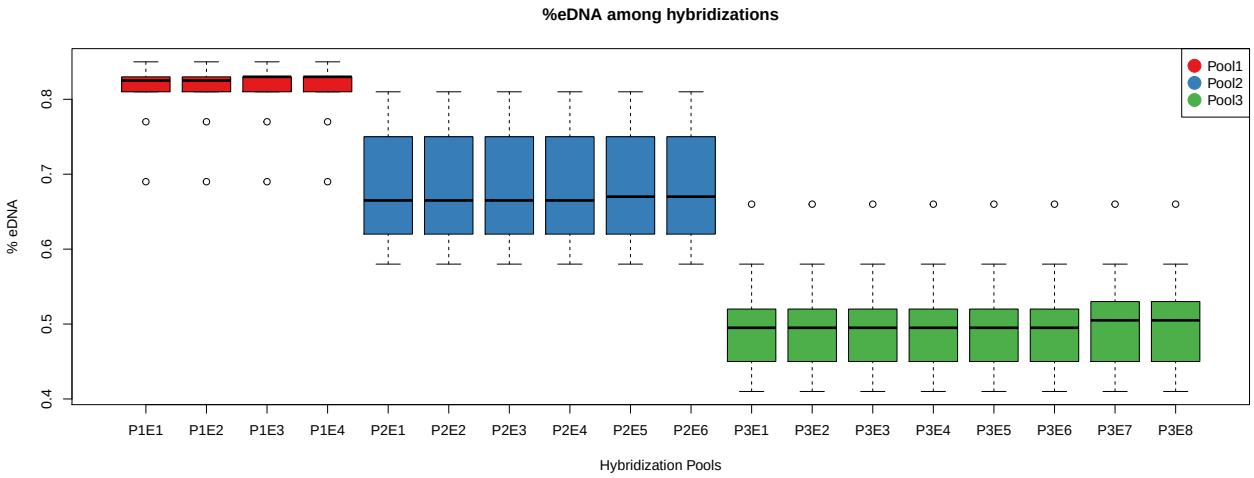
```
            `Sequencing Batch` , data = downdata[-w,])
# anova(fit0, fit)
###################

a = anova(fit)
eta2 = a[, 2]/sum(a[,2]); names(eta2) = rownames(a)
eta2 = data.frame(labels = rownames(a), eta2 = eta2)
###
par(mar = c(6.5, 5,3,3))
moose_barplot(eta2, "eta2", eta2$labels, 20, pylim = c(0,0.45))
```




– how is subspecies, site, tDNA, eDNA, %eDNA, frag-length partitioned among –> pools, hybridizations, and seq batches?

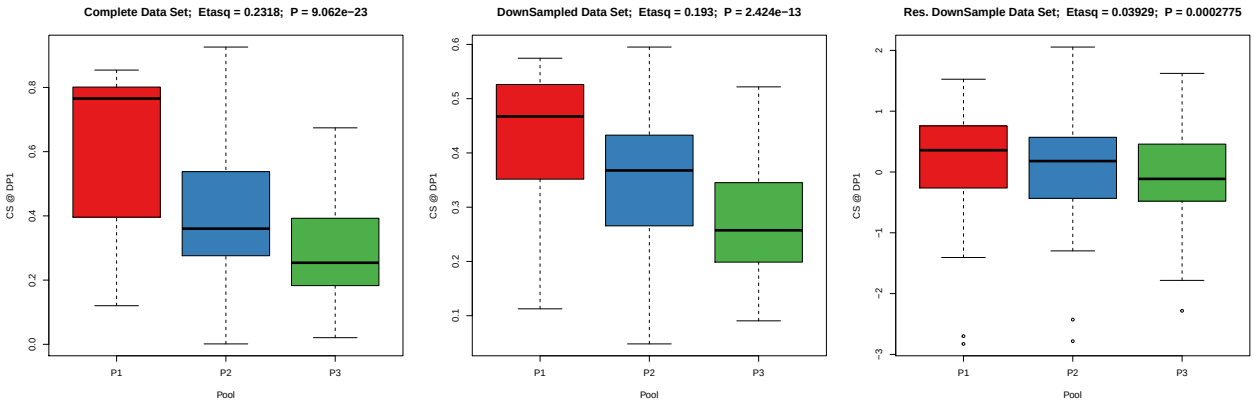
- subsepcies is randomly distributed among pools
- site is NOT randomly distributed among pools (rho = 0.44; r2 = 0.19, p = 7.546206e-12 )
- tDNA is randomly distributed among pools
- eDNA is NOT randomly distributed among pools (p = 0.007)
- %eDNA is NOT randomly distributed among pools; as designed (!) (rho = 0.88; r2 = 0.78; p = 5.0e-90)
- fragment length is NOT randomly distributed among pools (rho = 0.1608075; r2 = 0.02585905; p = 0.02872461)

```
x = RColorBrewer::brewer.pal(3, "Set1")
pcol = c( rep(x[1], 4), rep(x[2], 6), rep(x[3], 8) )
boxplot(  unlist(downdata$`% Endogenous DNA`) ~ unlist(downdata$`Capture Pool`), col = pcol,
          ylab = "% eDNA", xlab = "Hybridization Pools", main = "%eDNA among hybridizations")
legend("topright", legend = c("Pool1","Pool2","Pool3"), col = x, pch = 19, pt.cex = 2)
```




– Pool (n = 10, 20, 30) on Capture Sensitivity –

```
## varaince explained
w = which(downdata$`Production Reads` < 1500000)


###
pcol = RColorBrewer::brewer.pal(3, "Set1")
par(mfrow = c(1,3))
###
fit = lm( unlist( wdata$`Capture Sensitivity (CS) DP1`) ~ unlist(wdata$Pool) )
a = anova(fit)
etasq = signif( a[1,2]/sum(a[,2]) , d = 4)
pval = signif( a[1,5] , d = 4 )
boxplot( unlist( wdata$`Capture Sensitivity (CS) DP1`) ~ unlist(wdata$Pool), col = pcol,
          ylab = "CS @ DP1", xlab = "Pool",
          main = paste0( "Complete Data Set;  Etasq = ", etasq, ";  P = ", pval  ) )
#####
fit =  lm( unlist( downdata$`DP1`[-w]) ~ unlist(downdata$Pool[-w]) )
a = anova(fit)
etasq = signif( a[1,2]/sum(a[,2]) , d = 4)
pval = signif( a[1,5] , d = 4 )
boxplot( unlist( downdata$`DP1`[-w]) ~ unlist(downdata$Pool[-w]), col = pcol,
         ylab = "CS @ DP1", xlab = "Pool",
          main = paste0( "DownSampled Data Set;  Etasq = ", etasq, ";  P = ", pval  ))

###
dp = rntransform( downdata$DP1[-w] )
fit0 =  lm( dp ~  Site + `% Endogenous DNA` + `Average Fragment Size` + Pool, data = downdata[-w, ] )
###
res0 =  residuals( lm( dp ~  Site + `% Endogenous DNA` + `Average Fragment Size` , data = downdata[-w,
#res1 =  residuals( lm( dp ~ Site + `% Endogenous DNA` + `Average Fragment Size` + `Starting DNA (ug)`,
#fit =  lm( res0 ~ Pool, data = downdata[-w, ] )
a = anova(fit0)
etasq = signif( a[4,2]/sum(a[,2]) , d = 4)
pval = signif( a[4,5] , d = 4 )
boxplot( res0 ~ unlist(downdata$Pool[-w]), col = pcol,
         ylab = "CS @ DP1", xlab = "Pool",
          main = paste0( "Res. DownSample Data Set;  Etasq = ", etasq, ";  P = ", pval  ))
```



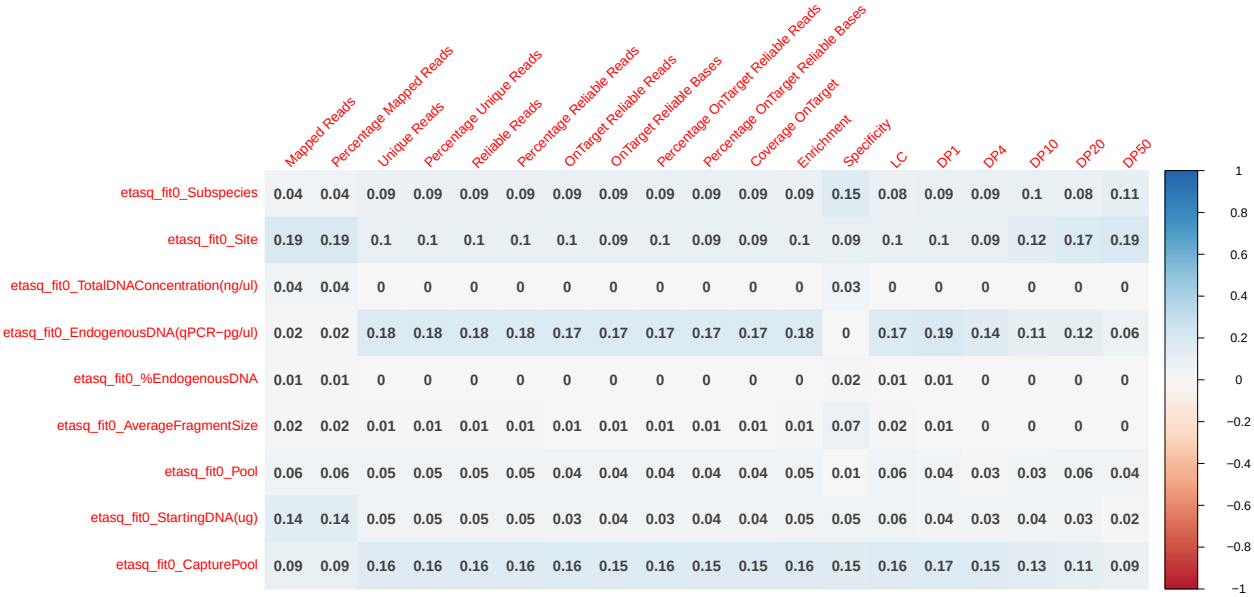

## Step 5) Multivariate Modeling

– of the complete data set –

– of the Downsampled data set limited to only those samples with 1.5M production reads –

–correlation plot the multivariate fit data –

```
## pdf
##   2
```

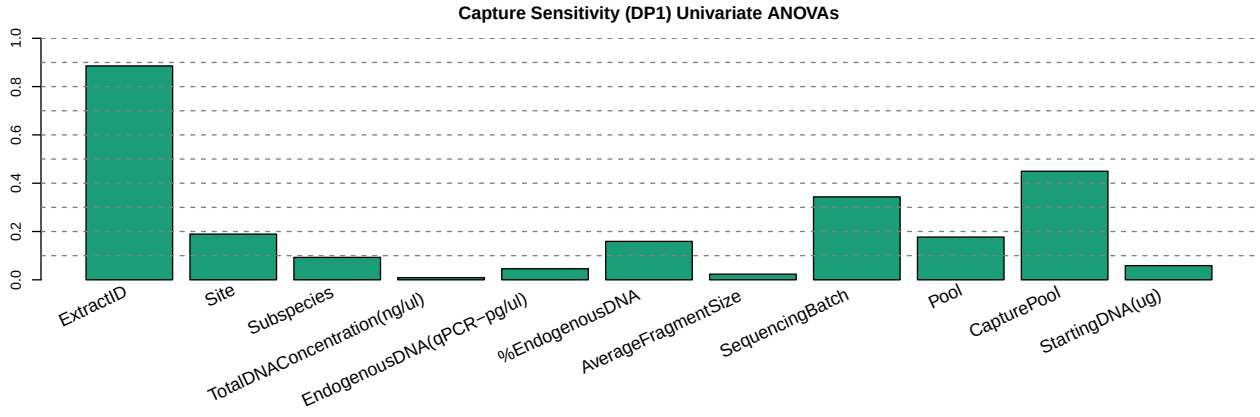


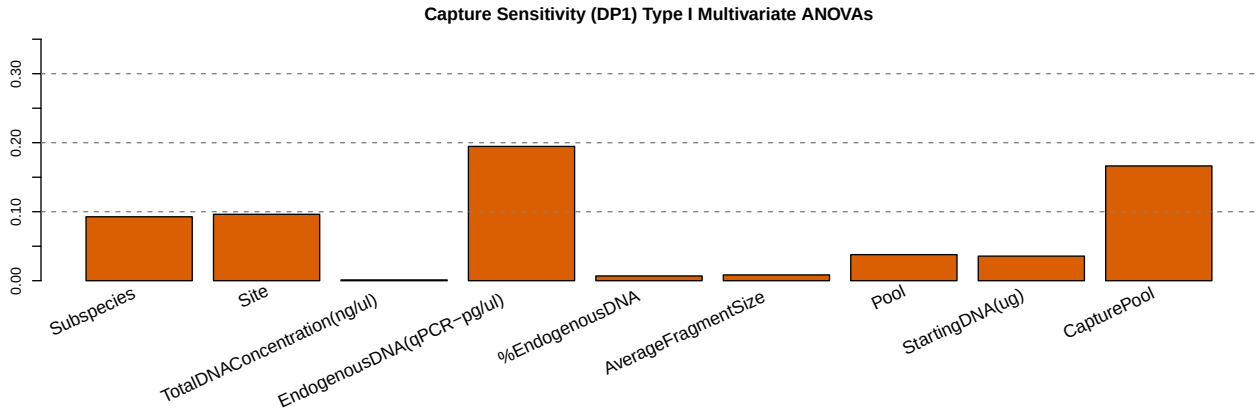

```
sort( apply(testmat, 1, mean) , decreasing = TRUE)
```

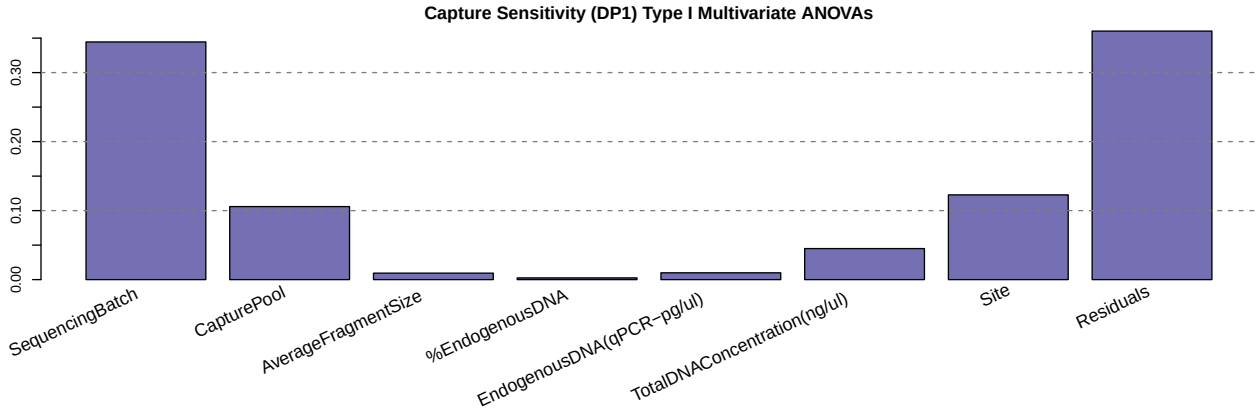
```
##                   etasq_fit0_CapturePool
##                              0.143042005
##    etasq_fit0_EndogenousDNA(qPCR-pg/ul)
##                              0.135723641
##                         etasq_fit0_Site
##                              0.116541128
##                     etasq_fit0_Subspecies
##                              0.088969515
##                etasq_fit0_StartingDNA(ug)
##                              0.049481209
```

```
##                          etasq_fit0_Pool
##                                0.043665194
##          etasq_fit0_AverageFragmentSize
##                                0.011863518
## etasq_fit0_TotalDNAConcentration(ng/ul)
##                                0.006215363
##                etasq_fit0_%EndogenousDNA
##                                0.004412848
```

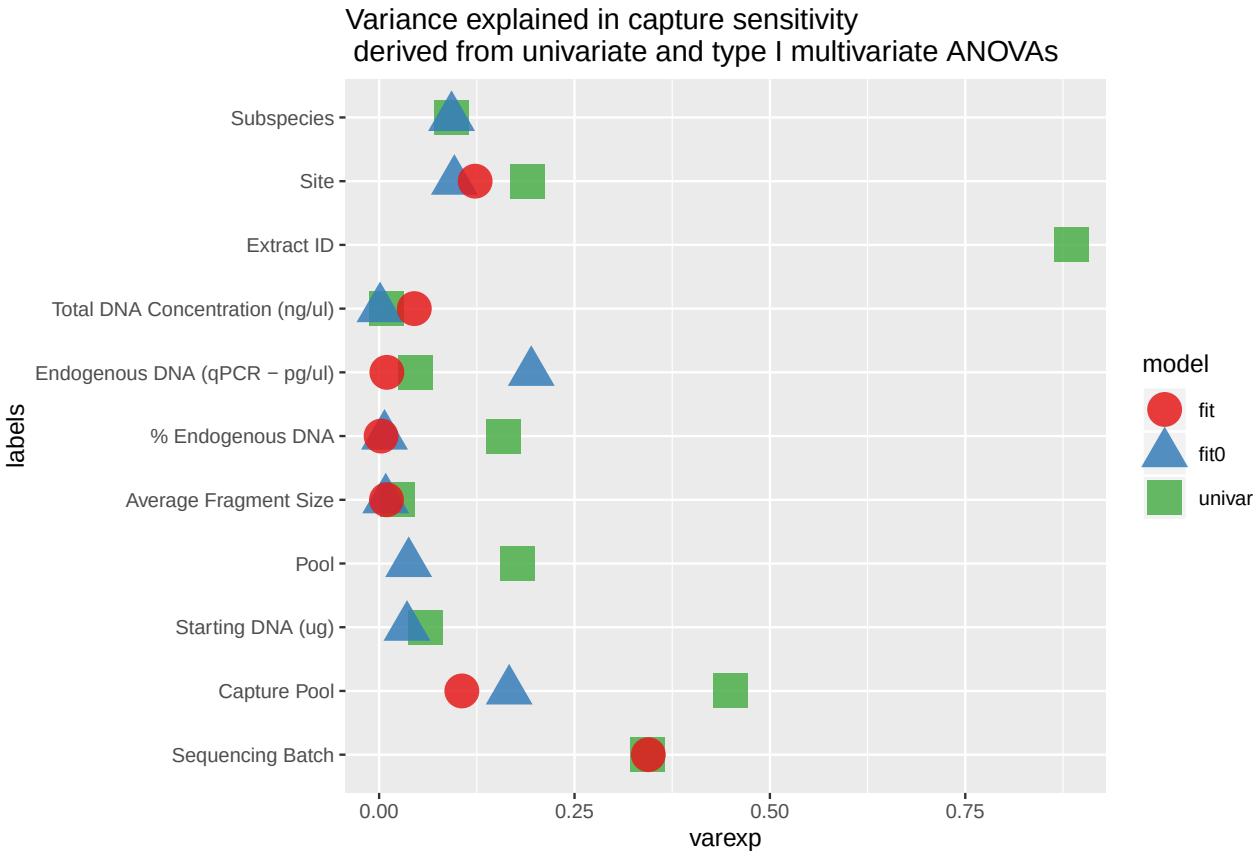
– barplots of the DownSampled Variance Explained for CS@DP1 –




```
## pdf
##   2
```

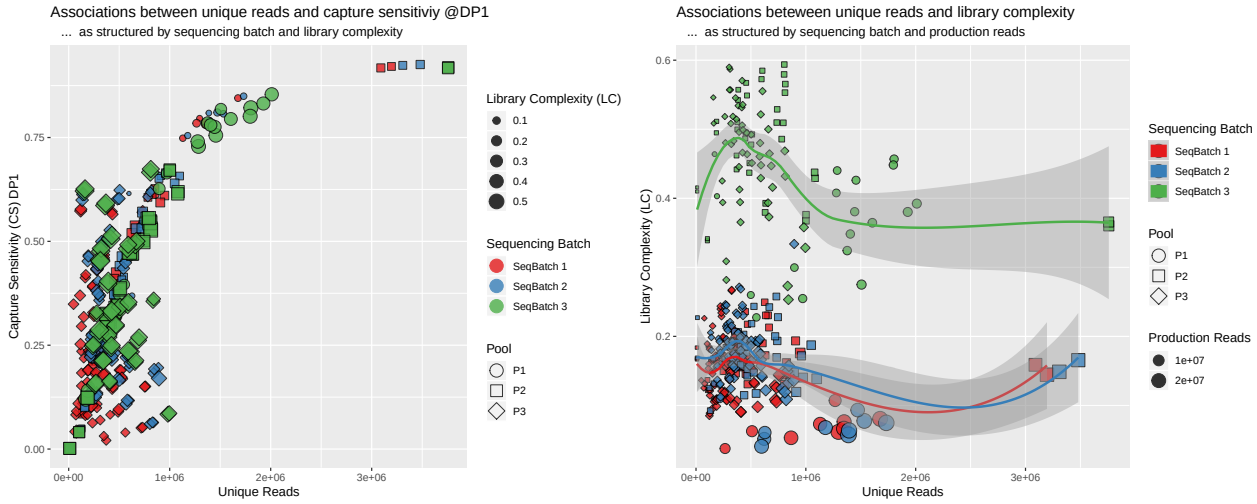



fit0 = CS@DP1 ~ ‘Subspecies‘ + ‘Site‘ + ‘Total DNA Concentration‘ + ‘% Endogenous DNA‘ + ‘Average Fragment Size‘ + ‘Pool‘ + ‘Starting DNA‘ +‘Capture Pool‘ + ‘Sequencing Batch‘

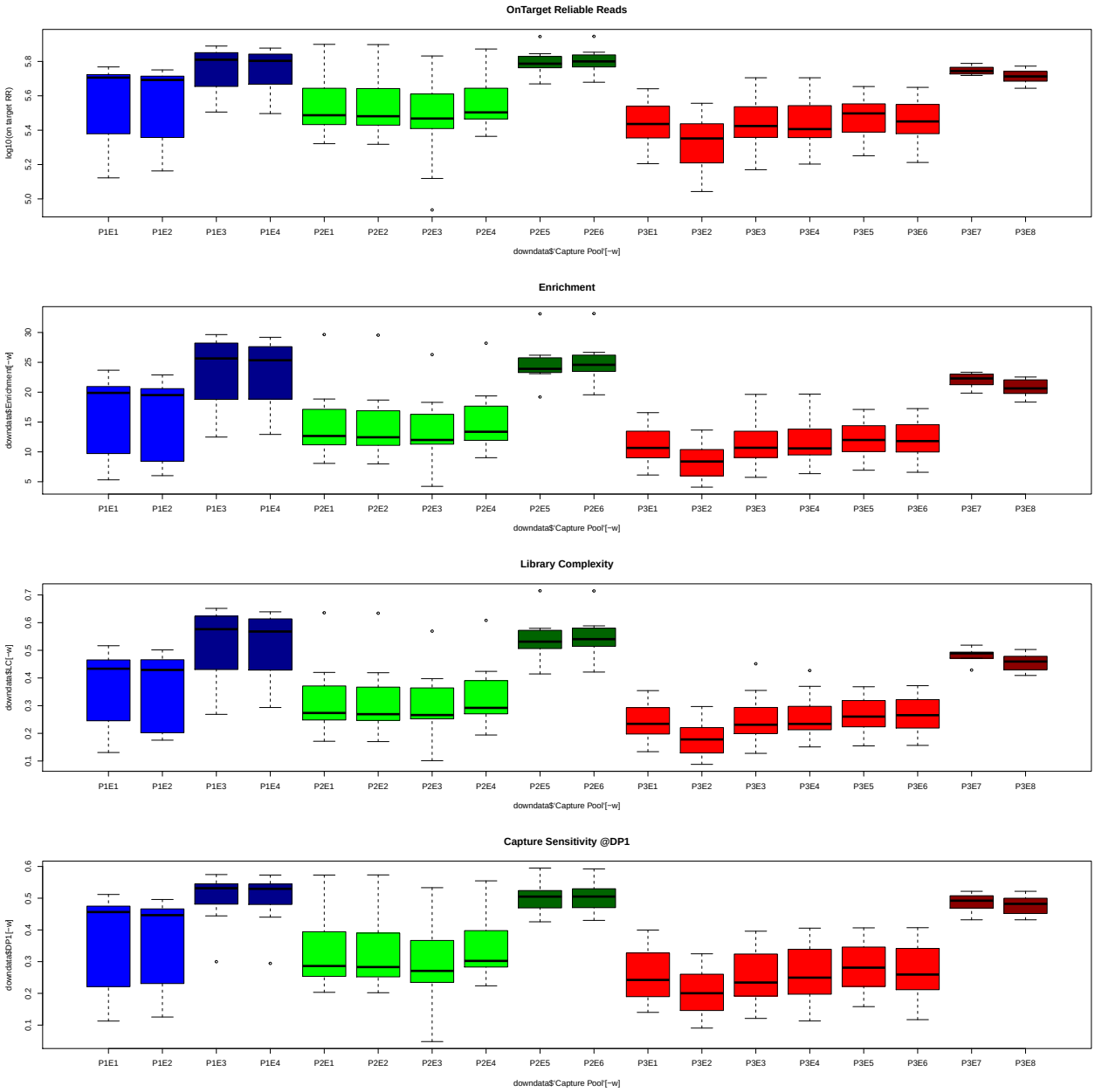
model ‘fit‘ has the order of explanatory variables reversed

## Library Complexity and CS @DP1

```
## pdf
##   2
```



## Unique Reads by hybridization




1. Samples in SeqBatch 3 all had 2ug of in the hybridization, hence the large bump in LC.
2. Drowning the data in useless sequencing also appears to have had a negative effect on LC.

What can we learn from these observations? 1. Increase the total DNA concentration in hybridization reactions (Perry paper did this) 2. Do not sequence to deeply.

## does %eDNA correlate with production reads ?

```
cp = sort( na.omit( as.character( unique(wdata$`Capture Pool`) ) ) )
x = c()
for(i in 1:length(cp)){
```

```
  w = which(wdata$`Capture Pool` == cp[i])
  a = cor.test(wdata$`% Endogenous DNA`[w], wdata$`Production Reads`[w])
  #a = cor.test(wdata$`% Endogenous DNA`[w], wdata$`OnTarget Reliable Reads`[w])
  out = c(a$estimate, a$p.value)
  x = rbind(x, out)
}
rownames(x) = cp
colnames(x) = c("rho","pval")
x
```

```
##              rho       pval
## P1E1  0.1084763 0.76548071
## P1E2  0.1519063 0.67526733
## P1E3  0.1904738 0.62351072
## P1E4  0.1897562 0.62484408
## P2E1  0.4038826 0.07738820
## P2E2  0.4603843 0.04108445
## P2E3  0.4461506 0.04863576
## P2E4  0.4525145 0.04513817
## P2E5  0.4190986 0.07408886
## P2E6  0.4193598 0.07389022
## P3E1 -0.1443537 0.44662072
## P3E2 -0.1834209 0.33194013
## P3E3 -0.1389939 0.46385179
## P3E4 -0.1720977 0.36315195
## P3E5 -0.1697775 0.36975498
## P3E6 -0.1614057 0.39415973
## P3E7 -0.1488519 0.46800527
## P3E8 -0.1502532 0.46377858
```

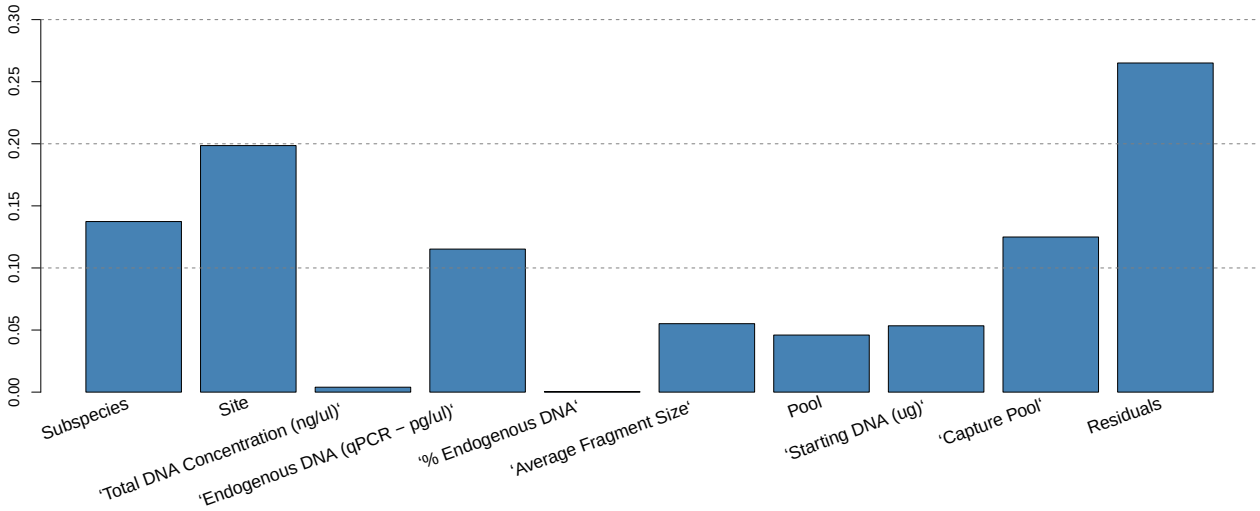
## Exclude pool 2 – Multivariate Modeling of CS @ DP1

– Down Sampled Data –

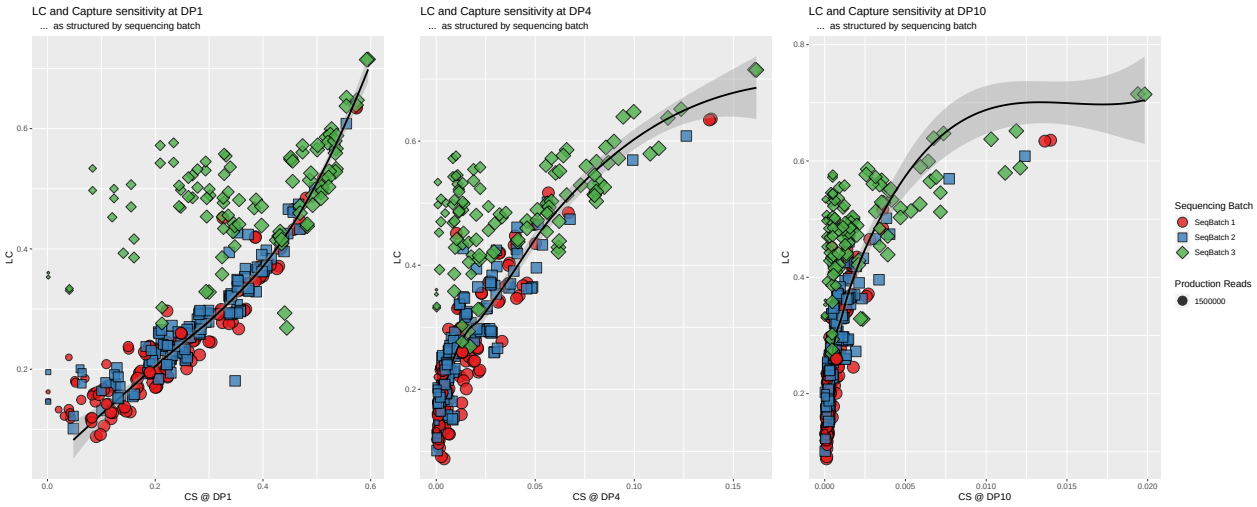
– model : CS@DP1 ~ subspecies + site + [tDNA] + [eDNA] + %eDNA + fragmentsize + pool + amount_DNA-in-hyb + hybridization + seqbatch + error

```
#w = which(downdata$`Production Reads` < 1500000  ) ## 274 samples left
w = which(downdata$`Production Reads` < 1500000 | downdata$Pool == "P2" ) ## 200 samples left
#####
d = rntransform( unlist( downdata[-w, "DP1"] ) )
## sub setted model
# fit0 = lm( d ~ `Subspecies` + `Site` +
#               `Total DNA Concentration (ng/ul)` +
#              #`Endogenous DNA (qPCR - pg/ul)` +
#               `% Endogenous DNA` +
#               `Average Fragment Size` +
#               `Pool` +
#               `Starting DNA (ug)` +
#               `Capture Pool` +
#               `Sequencing Batch` , data = downdata[-w,])
#

## full model !!!
```

```
fit = lm( d ~ `Subspecies` + `Site` +
              `Total DNA Concentration (ng/ul)` +
              `Endogenous DNA (qPCR - pg/ul)` +
              `% Endogenous DNA` +
              `Average Fragment Size` +
              `Pool` +
             `Starting DNA (ug)` +
              `Capture Pool` +
              `Sequencing Batch` , data = downdata[-w,])
# anova(fit0, fit)
###################

a = anova(fit)
eta2 = a[, 2]/sum(a[,2]); names(eta2) = rownames(a)
eta2 = data.frame(labels = rownames(a), eta2 = eta2)
###
par(mar = c(6.5, 5,3,3))
moose_barplot(eta2, "eta2", eta2$labels, 20, pylim = c(0,0.30))
```



## How are production reads influencing library complexity in the down sampled data ?

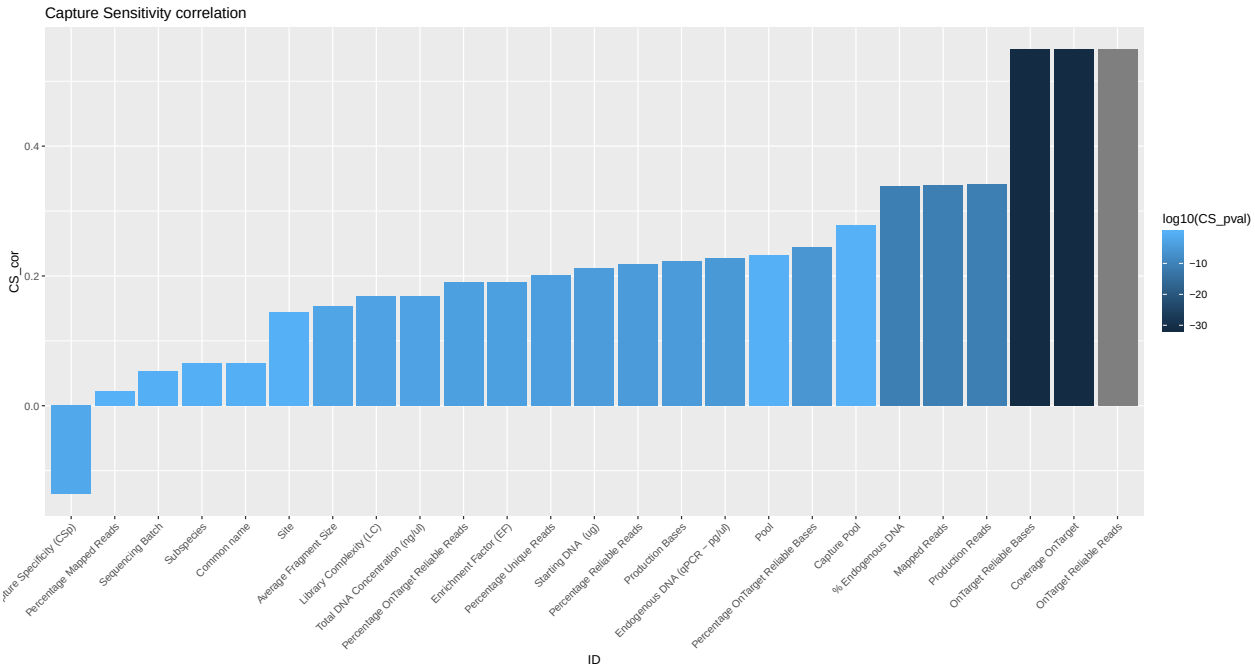



## Capture Sensitivity

It would be useful to identify a single summary statistic that summarizes what a good "performing" target capture seqeuencing experiment is. I think what we be most useful is to count the number of (population-wide) variable position that we were able to genotype (at whatever criteria uniformly executed across all samples). In the absence of this data it would seem that the best summary statistic that would predict this number is Capture Sensitivity (# of target regions covered by 1 read / total number of target regions), as having a base covered by a read provides a chance for genotyping.

Now depth in coverage gives us accuracy in genotyping heterozygosity and there is most certainly going to be some bias in capturing specific alleles, but we have some good evidence that just making hemizygous calls is informative for the gross|macro level population genetics we would like to do. However, this should eventually be quantified.

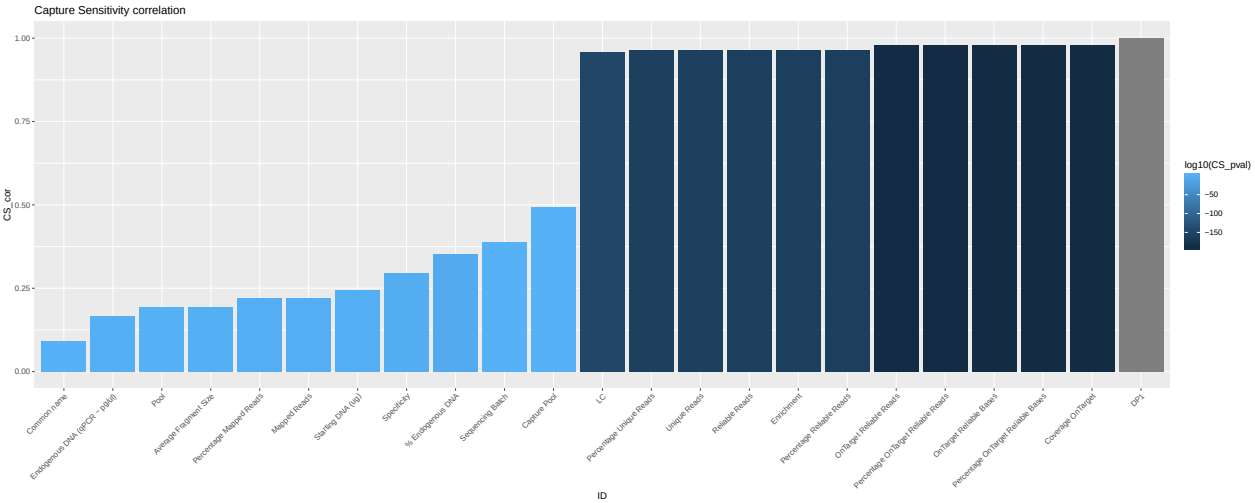
**Below we are asking in a univariate fashion how each of our variables, and sumstat, correlate with CAPTURE SENSITIVITY.**



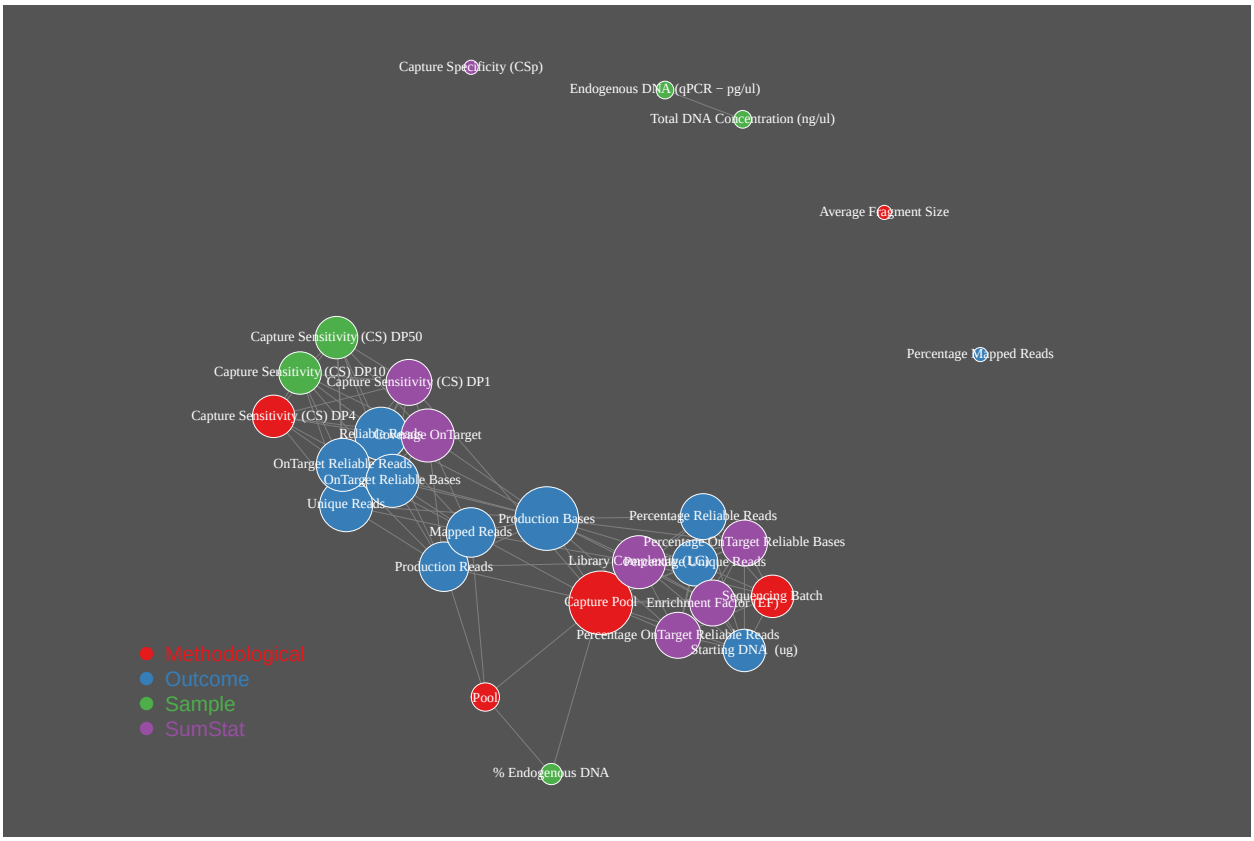

What can we learn from this analysis?

1. Want to increase capture sensitivity aquire more unique reads ! + kind of obvious but great to observe and demonstrate.
2. For technical or methodological choices it would appear that + samples with more Endogenous DNA, (note that this is NOT %DNA), it is higher DNA [concentrations] perform better + captures with more DNA in the hybridization perform better

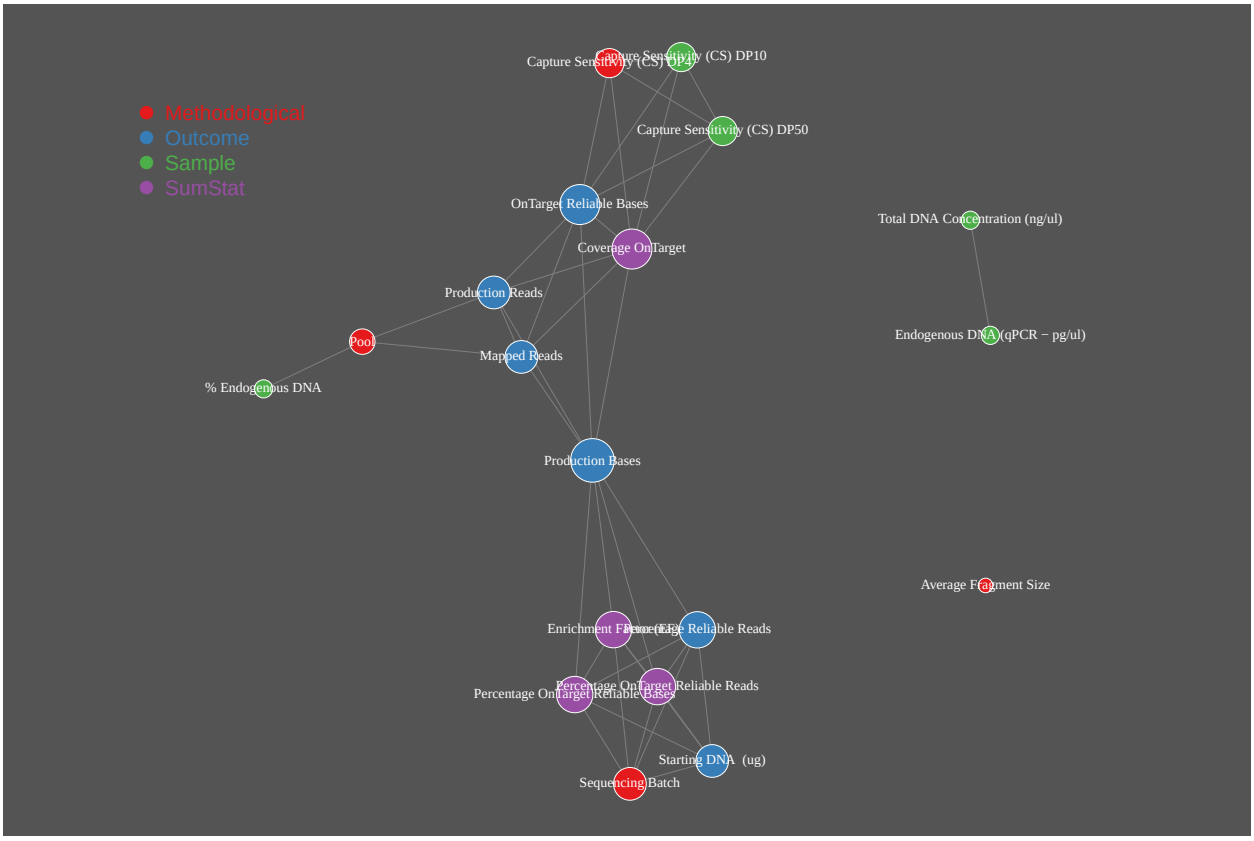
**Below we are asking in a univariate fashion how each of our variables, and sumstat, correlate with CAPTURE SENSITIVITY at a uniform production**



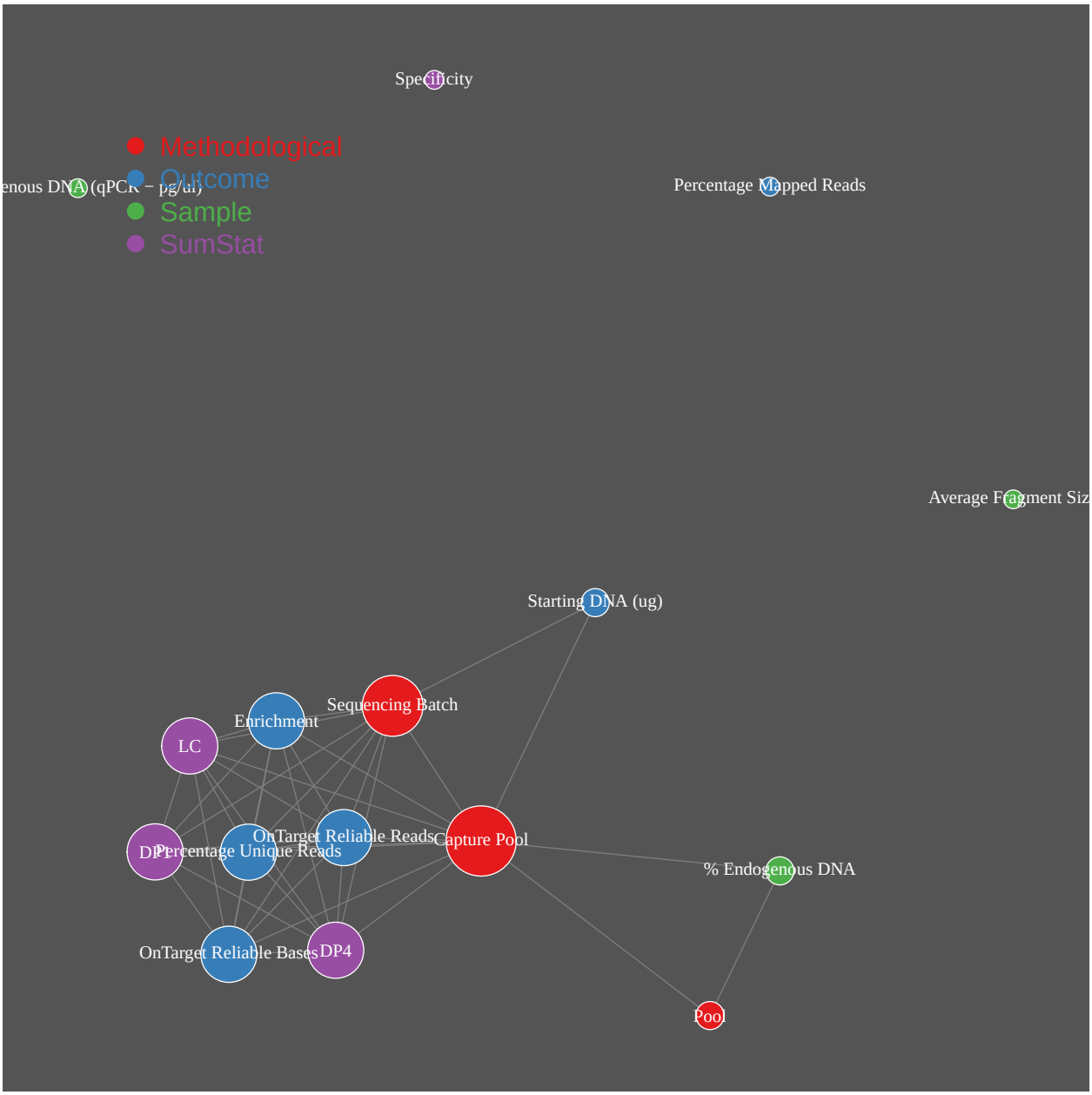
## Build a network of relationships based on correaltion estimates



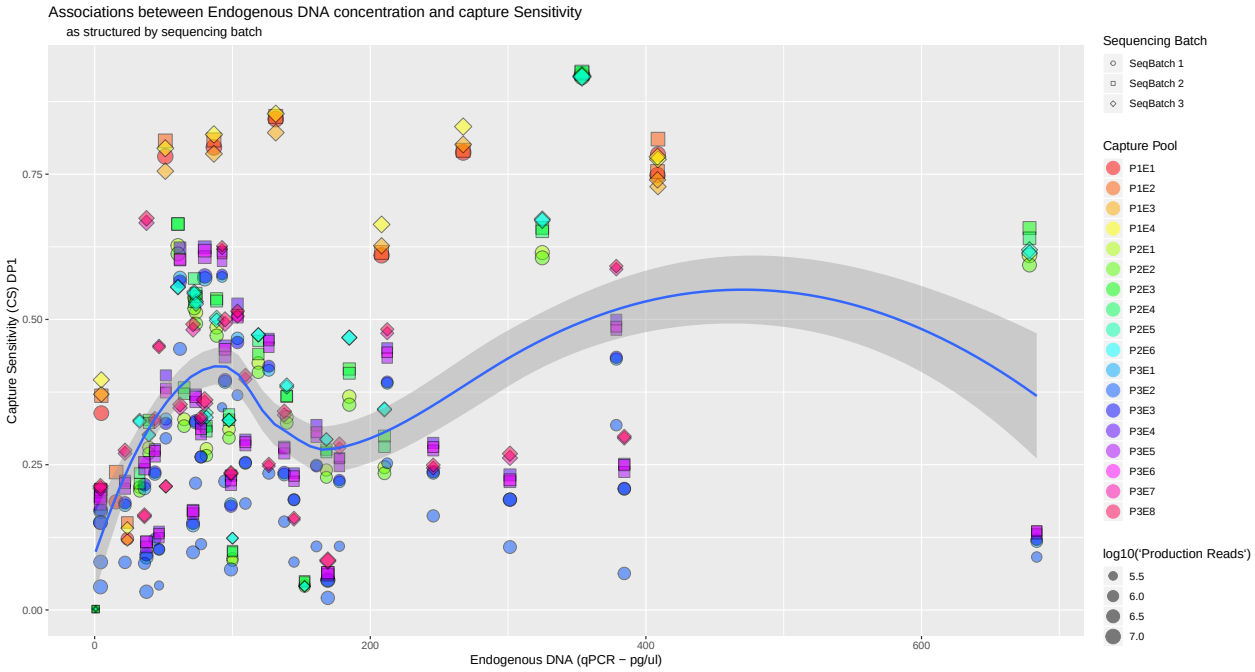
## Remove redundancies in Network

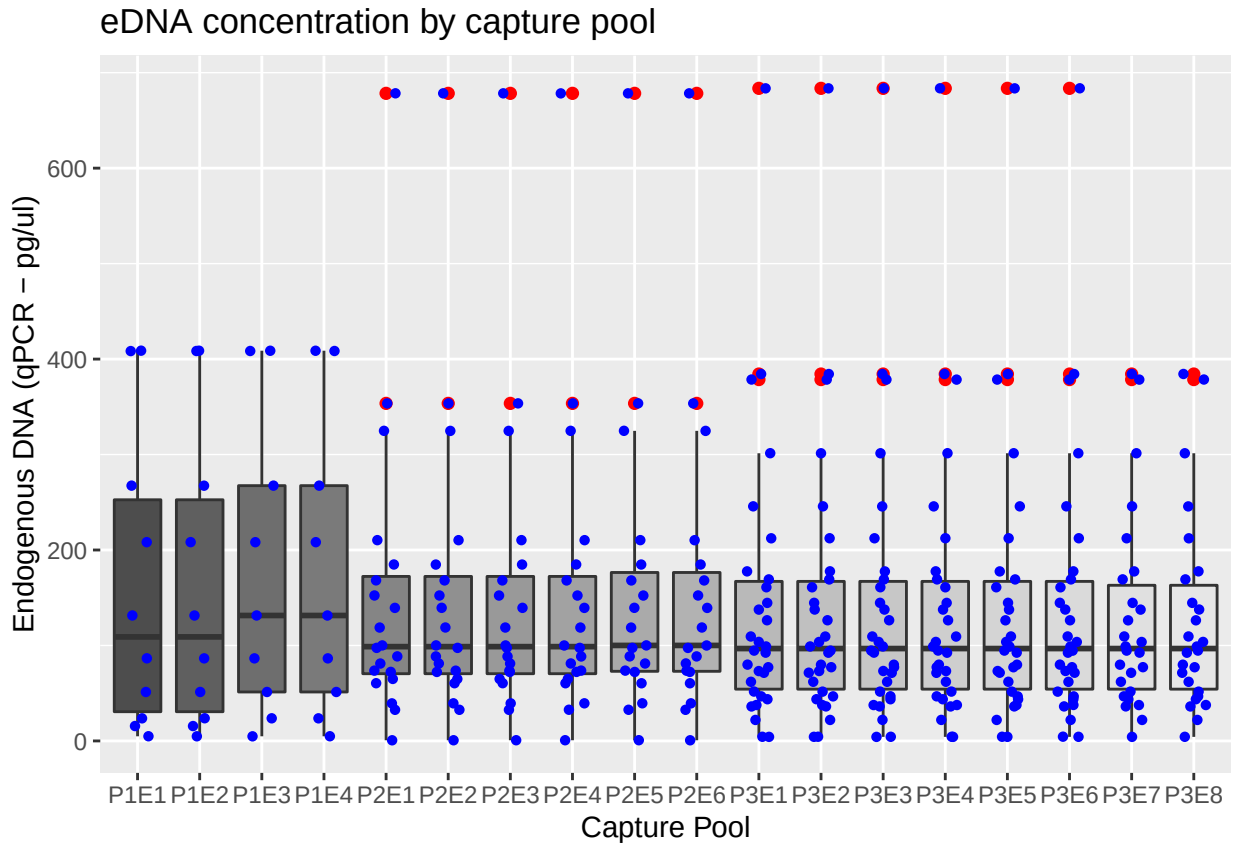


## Remove redundancies in Network and construct with down sampled data



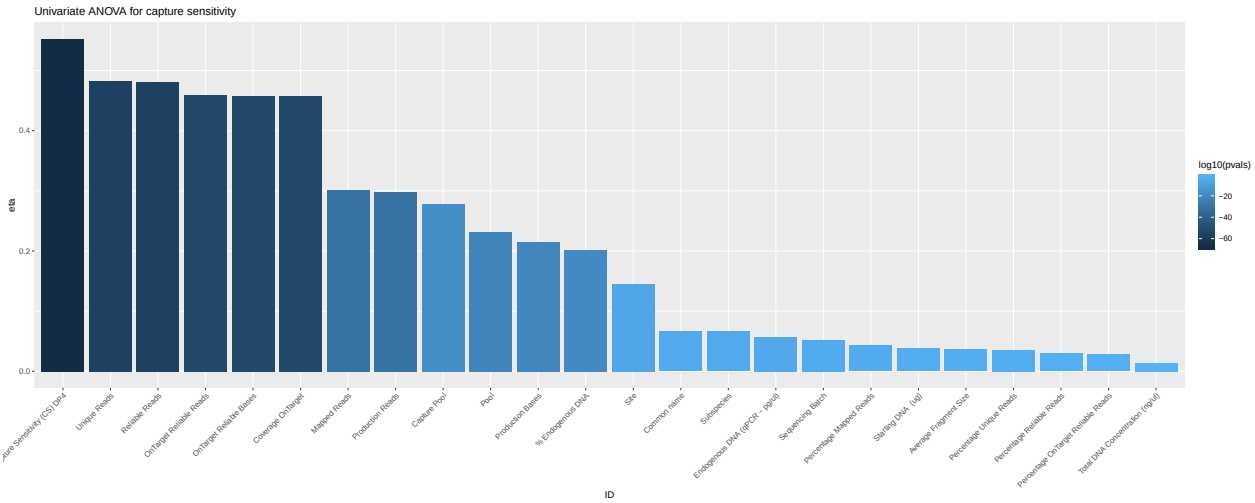
## How is the concentration of a sample influencing capture sensitivity ?



## Univariate ANOVA on Summary Statistics

```
cols2test = c(2:3, 4:23, 25, 30)
UnivarateANOVA = matrix(NA, length(cols2test), 2)
for(i in 1:length(cols2test) ){
  x = unlist(wdata[,cols2test[i] ])
  test = class( x )
  ###

    fit = lm(wdata$`Capture Sensitivity (CS) DP1` ~  x)

  ###
  a = anova(fit)
  eta = a[1,2]/sum(a[,2])
  pval = a[1, 5]
  out = c(eta, pval)
  ##
  UnivarateANOVA[i, ] = out
  }

rownames(UnivarateANOVA) = colnames(wdata)[cols2test]
colnames(UnivarateANOVA) = c("eta","pval")

## order
o = order(UnivarateANOVA[,1], decreasing = TRUE)
UnivarateANOVA = UnivarateANOVA[o,]
```




## A multivariate model to explain how sample quality, and methodological choice influences Capture Sensitivity

```
library(car)
################################
## fit a simple linear model
################################
fit = lm( `Capture Sensitivity (CS) DP1` ~ `Total DNA Concentration (ng/ul)` +
            `Endogenous DNA (qPCR - pg/ul)` +
```

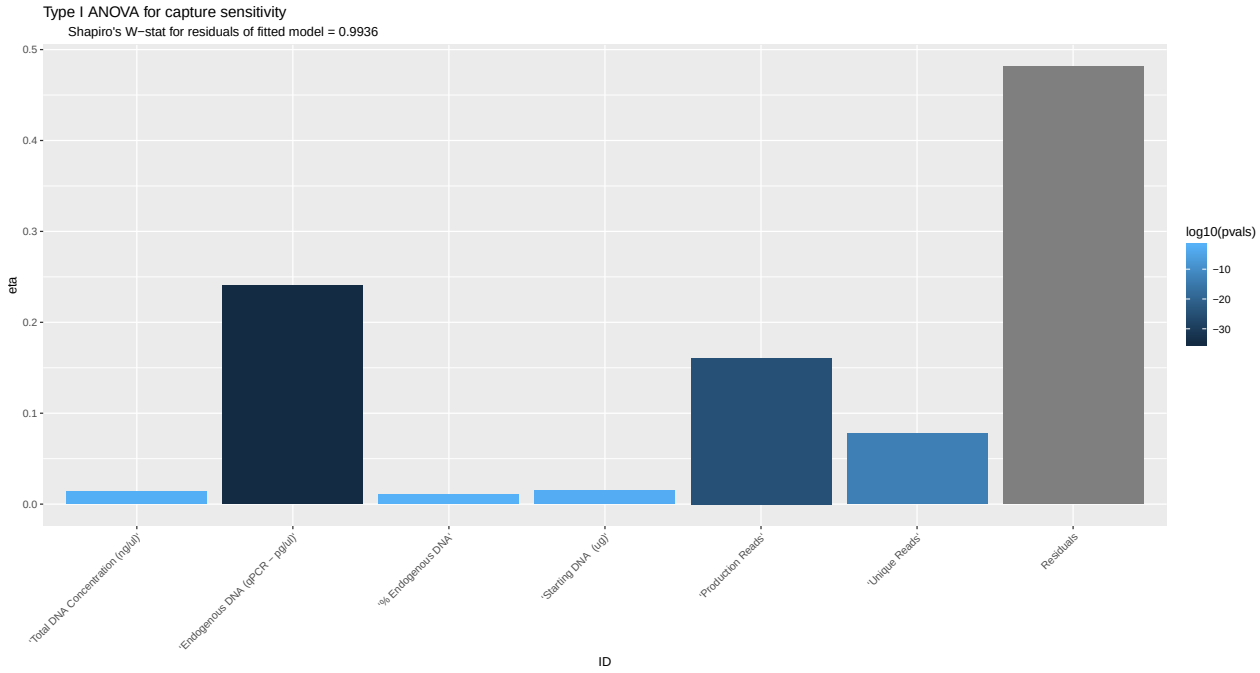
```
            `% Endogenous DNA` +
            `Capture Pool` +
            `Starting DNA  (ug)` +
            `Sequencing Batch` +
            `Production Reads` +
            `Unique Reads`
            , data = wdata )

fit = lm( `Capture Sensitivity (CS) DP1` ~ `Total DNA Concentration (ng/ul)` +
            `Endogenous DNA (qPCR - pg/ul)` +
            `% Endogenous DNA` +
            `Starting DNA  (ug)` +
            `Production Reads` +
            `Unique Reads`
            , data = wdata )


##############################
## are model residuals normal ?
##############################
W = shapiro.test(residuals(fit))

##############################
##  estiamte SS and VarExp
##  assuming an TypeI hierarchical
##  ANOVA
##############################

(a = anova(fit) )
```

```
## Analysis of Variance Table
##
## Response: Capture Sensitivity (CS) DP1
##                                   Df Sum Sq Mean Sq  F value    Pr(>F)
## `Total DNA Concentration (ng/ul)`  1 0.2370  0.2370  10.7605 0.0011324
## `Endogenous DNA (qPCR - pg/ul)`    1 4.1996  4.1996 190.6785 < 2.2e-16
## `% Endogenous DNA`                 1 0.1849  0.1849   8.3952 0.0039795
## `Starting DNA  (ug)`               1 0.2634  0.2634  11.9582 0.0006055
## `Production Reads`                 1 2.8046  2.8046 127.3397 < 2.2e-16
## `Unique Reads`                     1 1.3593  1.3593  61.7162 4.089e-14
## Residuals                        381 8.3914  0.0220
##
## `Total DNA Concentration (ng/ul)` **
## `Endogenous DNA (qPCR - pg/ul)`   ***
## `% Endogenous DNA`                **
## `Starting DNA  (ug)`              ***
## `Production Reads`                ***
## `Unique Reads`                    ***
## Residuals
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
eta = a[, 2]/ sum(a[,2])
names(eta) = rownames(a)
```

```r
summary(fit)
```

```
##
## Call:
## lm(formula = `Capture Sensitivity (CS) DP1` ~ `Total DNA Concentration (ng/ul)` +
##     `Endogenous DNA (qPCR - pg/ul)` + `% Endogenous DNA` + `Starting DNA  (ug)` +
##     `Production Reads` + `Unique Reads`, data = wdata)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.38579 -0.11332 -0.00564  0.11084  0.41269
##
## Coefficients:
##                                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     9.648e-03  5.903e-02   0.163  0.87026
## `Total DNA Concentration (ng/ul)` 3.783e-04  2.125e-03   0.178  0.85878
## `Endogenous DNA (qPCR - pg/ul)`   3.103e-05  3.760e-04   0.083  0.93427
## `% Endogenous DNA`                2.003e-01  9.798e-02   2.045  0.04158
## `Starting DNA  (ug)`              5.956e-02  2.021e-02   2.946  0.00341
## `Production Reads`                7.901e-09  2.693e-09   2.934  0.00355
## `Unique Reads`                    2.037e-07  2.593e-08   7.856 4.09e-14
##
## (Intercept)
## `Total DNA Concentration (ng/ul)`
## `Endogenous DNA (qPCR - pg/ul)`
## `% Endogenous DNA`                *
## `Starting DNA  (ug)`              **
## `Production Reads`                **
## `Unique Reads`                    ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1484 on 381 degrees of freedom
##   (12 observations deleted due to missingness)
## Multiple R-squared:  0.5188, Adjusted R-squared:  0.5113
## F-statistic: 68.47 on 6 and 381 DF,  p-value: < 2.2e-16
```


Type I ANOVA for capture sensitivity
Shapiro's W−stat for residuals of fitted model = 0.9936
eta
0.5
0.4
0.3
0.2
0.1
0.0
log10(pvals)
−10
−20
−30
'Total DNA Concentration (ng/ul)'
'Endogenous DNA (qPCR − pg/ul)'
'% Endogenous DNA'
'Starting DNA (ug)'
'Production Reads'
'Unique Reads'
Residuals
ID