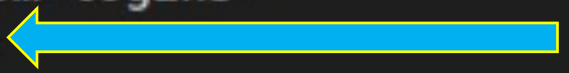
Topic

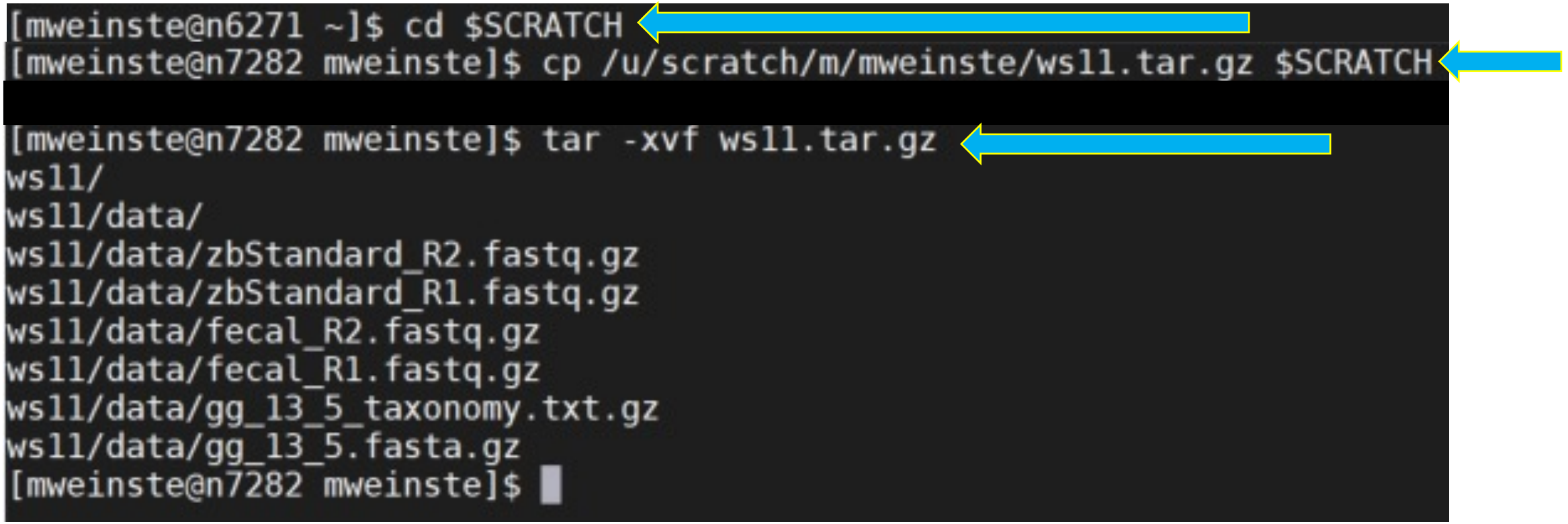
GETTING OUR SAMPLE DATA

Getting Started

```
Processes running on the login nodes which seriously degrade others'
use of the system may be terminated without warning.  Use qrsh to obtain
an interactive shell on a compute node for CPU or I/O intensive tasks.

The following news items are currently posted:

   IDRE Workshops and Training Sessions
   News Archive On Web Site

Enter shownews to read the full text of a news item.
[mweinste@login4 ~]$ qrsh -l h_data=2G,h_rt=10:00:00 -pe shared 8
Last login: Tue May 12 14:42:44 2020 from login3
[mweinste@n6271 ~]$ module load R/3.6.1

The 'gcc/4.9.3' module is being loaded

        These modules were already loaded: ATS intel/18.0.4

        Unloading the conflicting module 'intel/18.0.4'
```



Getting a Started



Install BioConductor

```
[mweinste@n7361 ~]$ module load R/3.6.1

The 'gcc/4.9.3' module is being loaded

        These modules were already loaded: ATS intel/18.0.4

        Unloading the conflicting module 'intel/18.0.4'

[mweinste@n7361 ~]$ R

R version 3.6.1 (2019-07-05) -- "Action of the Toes"
Copyright (C) 2019 The R Foundation for Statistical Computing
Platform: x86_64-pc-linux-gnu (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

  Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> install.packages("BiocManager")
Installing package into '/u/home/m/mweinste/R/x86_64-pc-linux-gnu-library/3.6'
(as 'lib' is unspecified)
```

Answer “yes” to questions about installation location if asked

Install DADA2

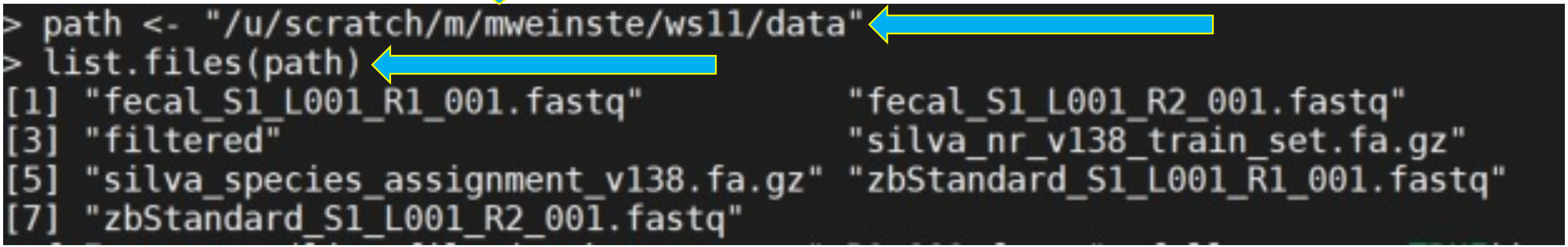
```
55: USA (MI 2) [https]
56: USA (OH) [https]
57: USA (OR) [https]
58: USA (TN) [https]
59: USA (TX 1) [https]
60: Uruguay [https]
61: (other mirrors)

Selection: 57
trying URL 'https://ftp.osuosl.org/pub/cran/src/contrib/BiocManager_1.30.10.tar.gz'
Content type 'application/x-gzip' length 40205 bytes (39 KB)
==================================================
downloaded 39 KB

* installing *source* package 'BiocManager' ...
** package 'BiocManager' successfully unpacked and MD5 sums checked
** using staged installation
** R
** inst
** byte-compile and prepare package for lazy loading
** help
*** installing help indices
** building package indices
** installing vignettes
** testing if installed package can be loaded from temporary location
** testing if installed package can be loaded from final location
** testing if installed package keeps a record of temporary installation path
* DONE (BiocManager)

The downloaded source packages are in
        '/work/tmp/RtmpqzTNfa/downloaded_packages'
> BiocManager::install("dada2", version = "3.10")
Bioconductor version 3.10 (BiocManager 1.30.10), R 3.6.1 (2019-07-05)
Installing package(s) 'BiocVersion', 'dada2'
also installing the dependencies 'ps', 'processx', 'callr', 'prettyunits', 'desc', '
```

…it's gonna take a while

```
** R
** data
*** moving datasets to lazyload DB
** inst
** byte-compile and prepare package for lazy loading
** help
*** installing help indices
** building package indices
** installing vignettes
** testing if installed package can be loaded from temporary location
** checking absolute paths in shared objects and dynamic libraries
** testing if installed package can be loaded from final location
** testing if installed package keeps a record of temporary installation path
* DONE (dada2)

The downloaded source packages are in
        '/work/tmp/RtmpqzTNfa/downloaded_packages'
Installation path not writeable, unable to update packages: backports, boot,
  class, cli, digest, glue, jsonlite, KernSmooth, lattice, MASS, Matrix, mgcv,
  nlme, nnet, pillar, Rcpp, repr, rlang, spatial, survival, uuid, vctrs
>
```

Set “path” to Where the Sequences Are

Use your own scratch folder here



Have R Find and Sort Your Sequencing Files

```
> fnFs <- sort(list.files(path, pattern="_R1_001.fastq", full.names = TRUE))
> fnRs <- sort(list.files(path, pattern="_R2_001.fastq", full.names = TRUE))
> sample.names <- sapply(strsplit(basename(fnFs), "_"), `[`, 1)
> fnFs
[1] "/u/scratch/m/mweinste/ws11/data/fecal_S1_L001_R1_001.fastq"
[2] "/u/scratch/m/mweinste/ws11/data/zbStandard_S1_L001_R1_001.fastq"
> fnRs
[1] "/u/scratch/m/mweinste/ws11/data/fecal_S1_L001_R2_001.fastq"
[2] "/u/scratch/m/mweinste/ws11/data/zbStandard_S1_L001_R2_001.fastq"
> sample.names
[1] "fecal"      "zbStandard"
```

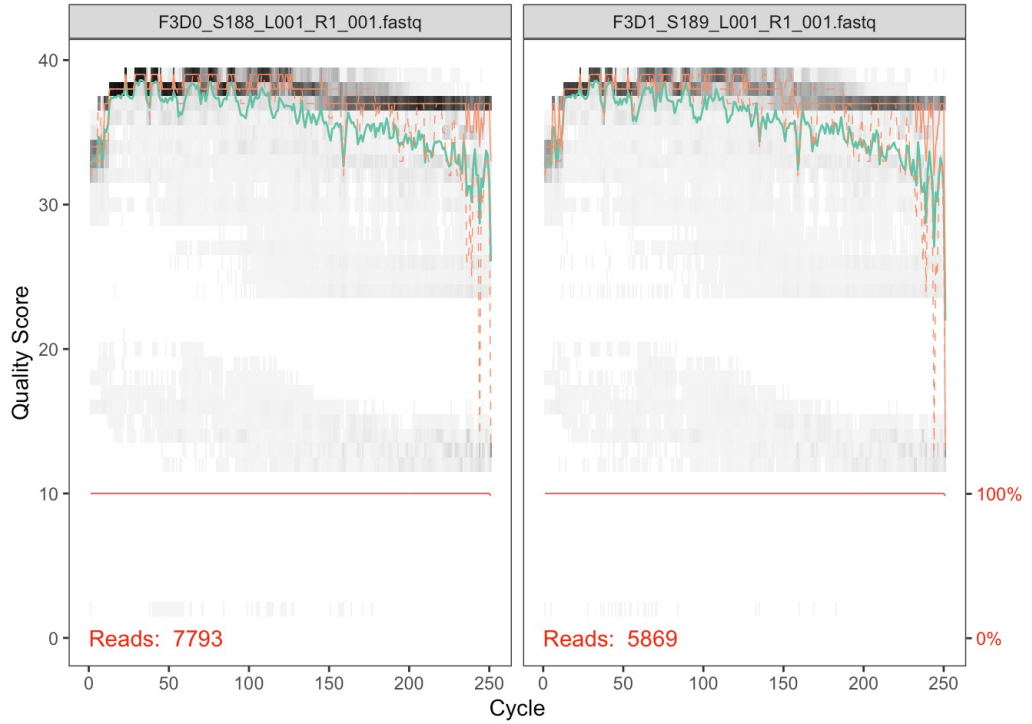
The “Official” Method

Inspect read quality profiles

We start by visualizing the quality profiles of the forward reads:

```
plotQualityProfile(fnFs[1:2])
```

```
## Scale for 'y' is already present. Adding another scale for 'y', which
## will replace the existing scale.
```



Open a New Terminal Tab/Window

```
The following news items are currently posted:

   IDRE Workshops and Training Sessions
   News Archive On Web Site

Enter shownews to read the full text of a news item.
[mweinste@login3 ~]$ qrsh -l h_data=16G,h_rt=10:00:00
Last login: Mon May 11 21:59:41 2020 from login4
[mweinste@n7282 ~]$ cd $SCRATCH
[mweinste@n7282 mweinste]$ cd ws11
[mweinste@n7282 ws11]$ ls
data  figaro
[mweinste@n7282 ws11]$ module load python/3.6.1

The 'gcc/4.9.3' module is being loaded

        These modules were already loaded: ATS intel/18.0.4

[mweinste@n7282 ws11]$ python3 figaro/figaro.py --ampliconLength 470 --forwardPrimerLength 16 --reversePrimerLength 24 --inputDirectory data/
```

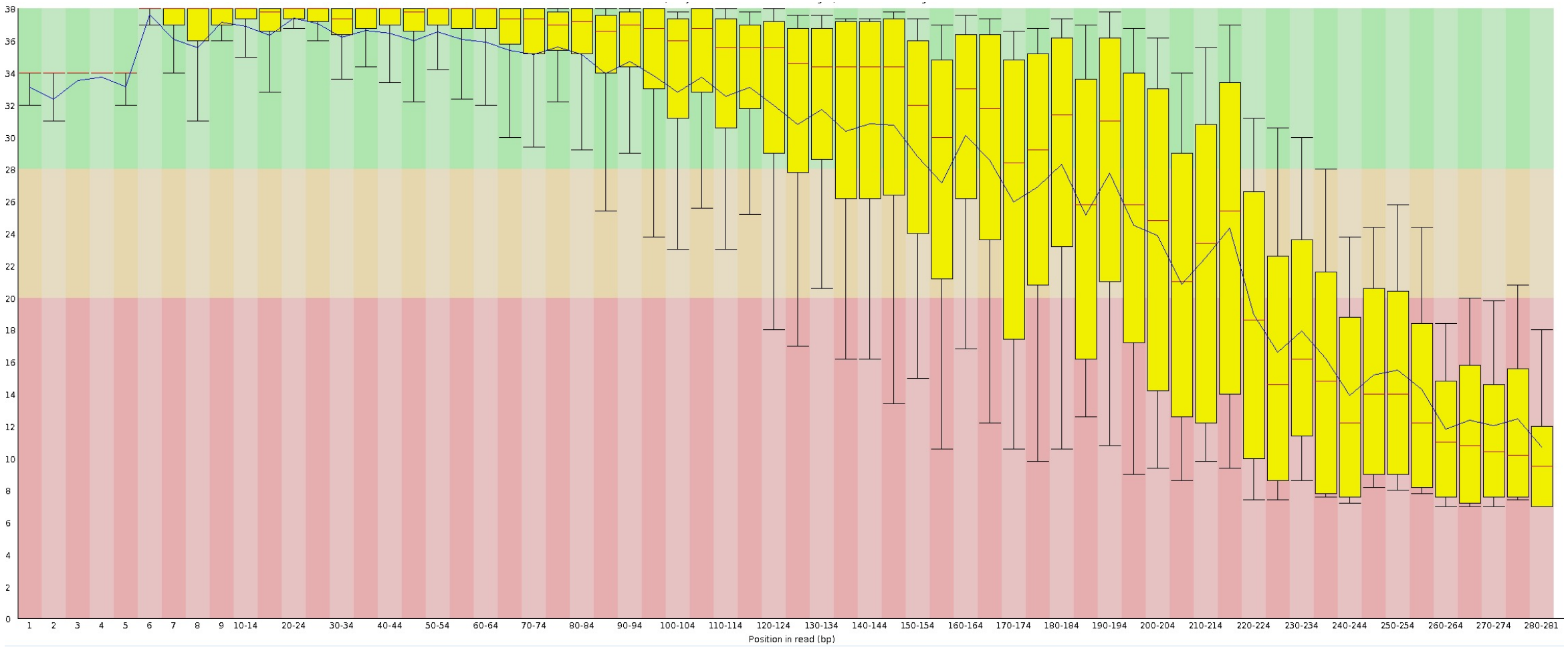
Amplicon length and primer lengths will be determined by your method of library prep and target region

Open a New Terminal Tab/Window

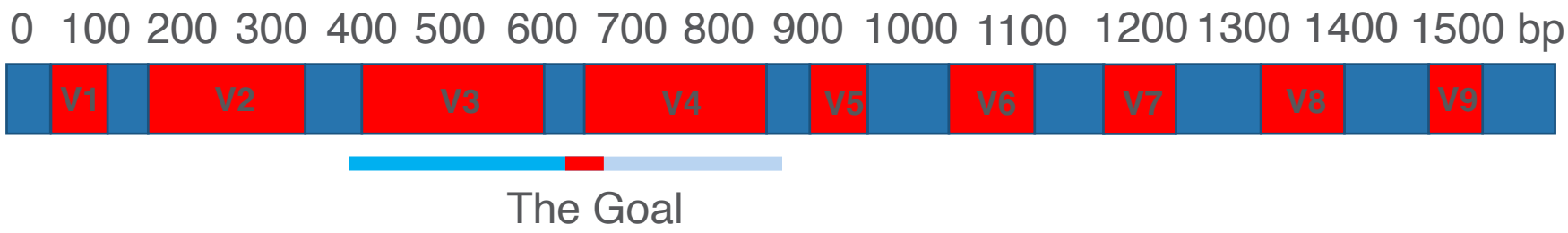
```
[mweinste@n7282 ws11]$ module load python/3.6.1

The 'gcc/4.9.3' module is being loaded

        These modules were already loaded: ATS intel/18.0.4

[mweinste@n7282 ws11]$ python3 figaro/figaro.py --ampliconLength 470 --forwardPrimerLength 16 --reversePrimerLength 24
{"trimPosition": [305, 225], "maxExpectedError": [5, 4], "readRetentionPercent": 76.9, "score": 51.896132561338405}
{"trimPosition": [304, 226], "maxExpectedError": [5, 4], "readRetentionPercent": 76.28, "score": 51.28354819103676}
{"trimPosition": [303, 227], "maxExpectedError": [5, 4], "readRetentionPercent": 75.9, "score": 50.904481622468889}
{"trimPosition": [302, 228], "maxExpectedError": [5, 5], "readRetentionPercent": 80.83, "score": 48.8339464508493}
{"trimPosition": [301, 229], "maxExpectedError": [5, 5], "readRetentionPercent": 80.68, "score": 48.6835993730207}
{"trimPosition": [300, 230], "maxExpectedError": [5, 5], "readRetentionPercent": 80.6, "score": 48.595630338120981}
{"trimPosition": [299, 231], "maxExpectedError": [5, 5], "readRetentionPercent": 80.52, "score": 48.520456799206671}
{"trimPosition": [298, 232], "maxExpectedError": [5, 5], "readRetentionPercent": 80.3, "score": 48.301333930456481}
{"trimPosition": [297, 233], "maxExpectedError": [5, 5], "readRetentionPercent": 80.25, "score": 48.253350820511191}
```

You may have to scroll back up a bit to find the first set of values.

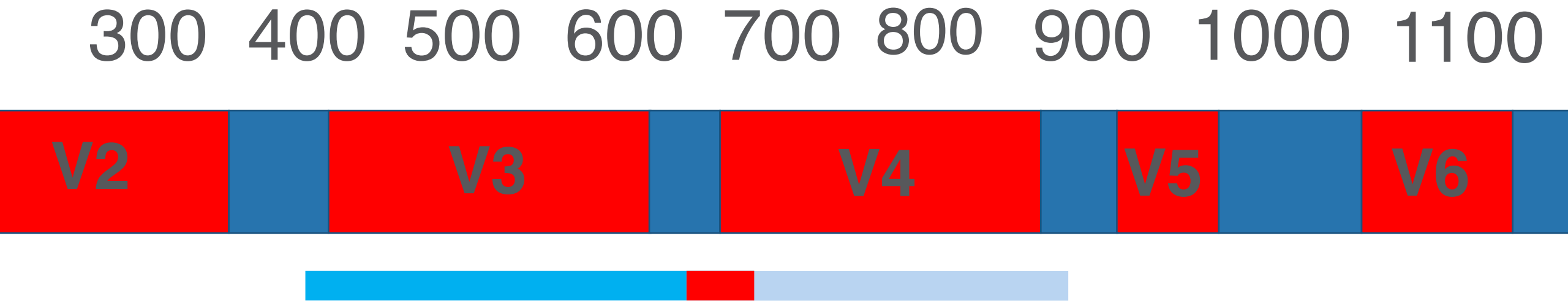
Selecting Read Trimming Parameters



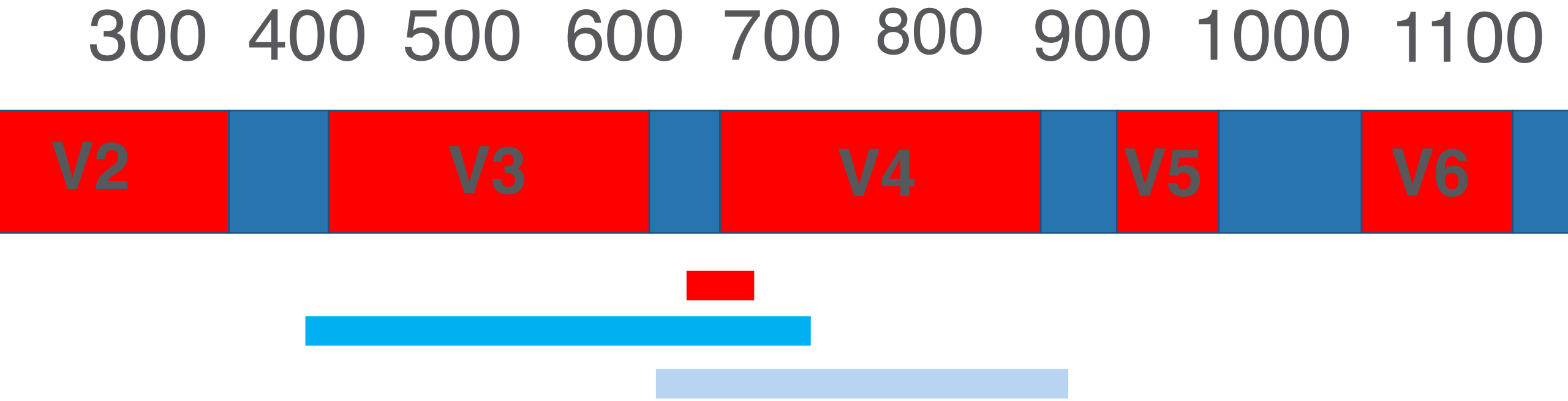
Selecting Read Trimming Parameters



Selecting Read Trimming Parameters



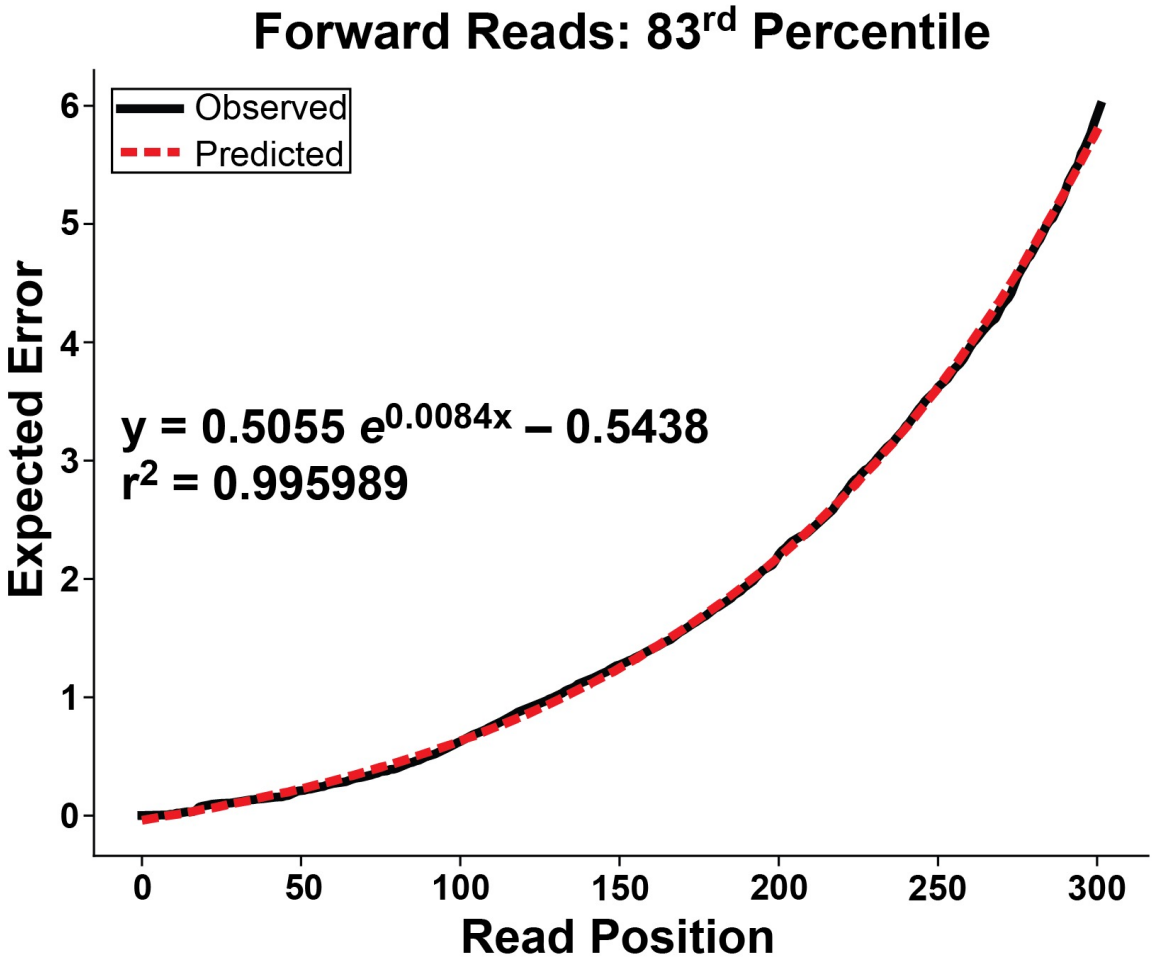
Selecting Read Trimming Parameters

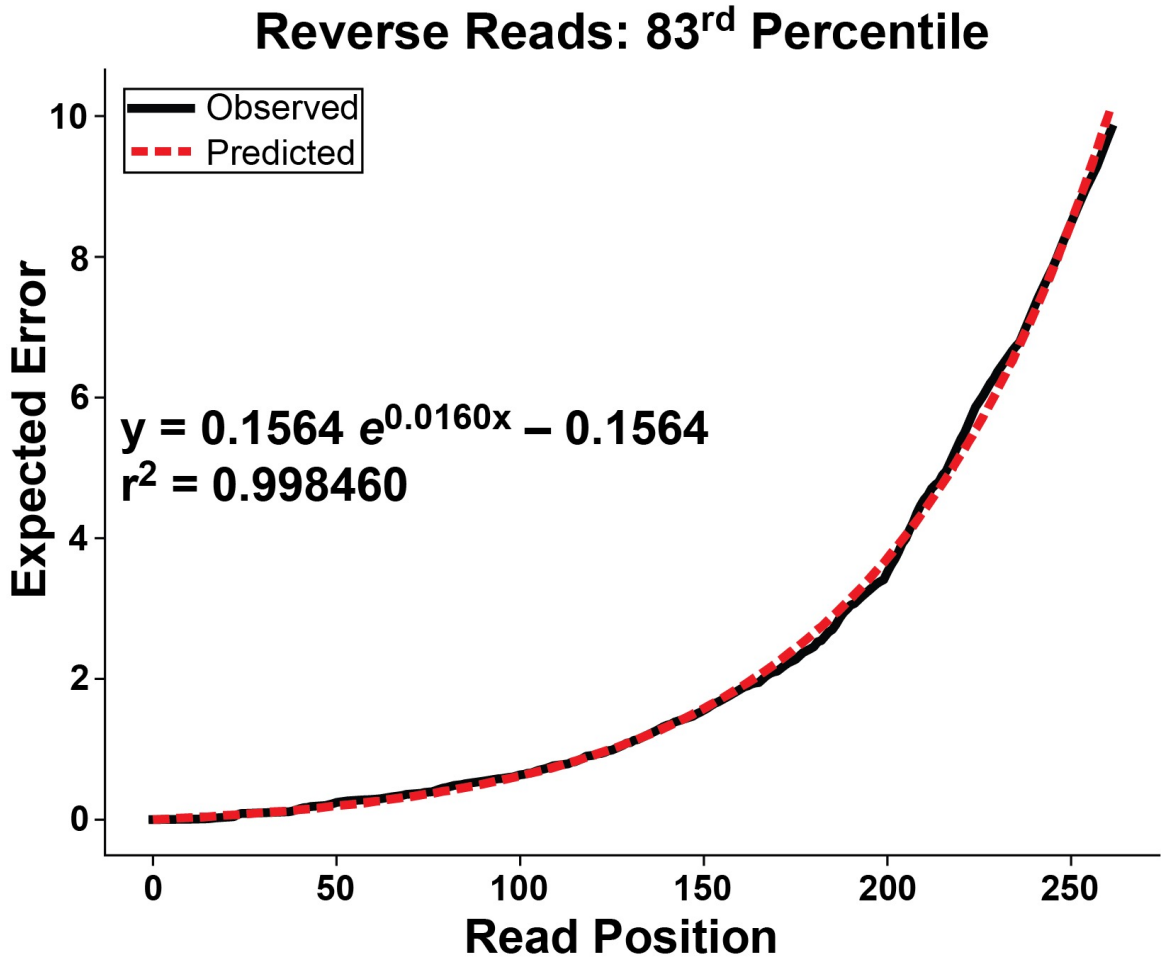


The Challenge

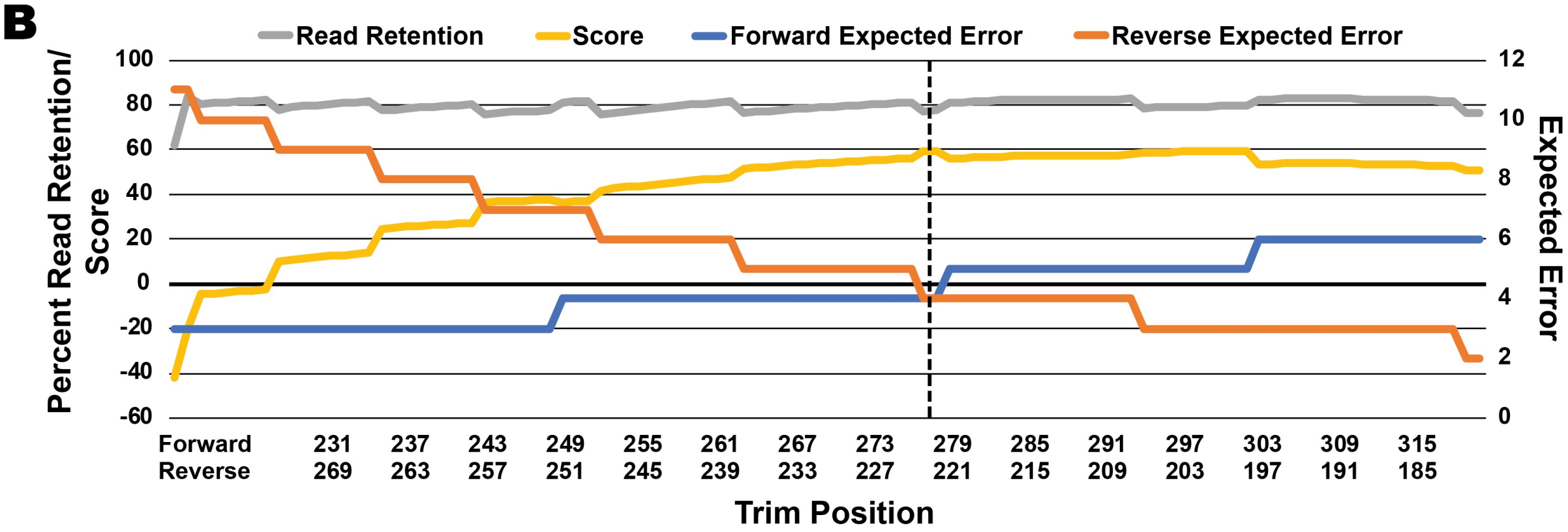
- There is some flexibility in where the overlap happens.
- Most pipelines have a technician look at the quality plots and select trimming points.
- **How do we minimize or remove human interaction in trimming parameter selection?**

Model Expected Error Accumulation





Find An Optimal Trimming Site



Return to Your Terminal Running R

Prepare To Run FilterAndTrim

```
> filtFs <- file.path(path, "filtered", paste0(sample.names, "_F_filt.fastq.gz"))
> filtRs <- file.path(path, "filtered", paste0(sample.names, "_R_filt.fastq.gz"))
> names(filtFs) <- sample.names
> names(filtRs) <- sample.names
> library(dada2)
```

This may tell you it is loading other packages (such as *Rcpp*). This is NOT a problem.

Trim and Filter

```
> out <- filterAndTrim(fnFs, filtFs, fnRs, filtRs, truncLen=c(305,225), trimLeft=c(16,24), maxN=0, maxEE=c(5,4), truncQ=2, rm.phix=TRUE, compress=TRUE, multithread=TRUE)
> head(out)
                           reads.in reads.out
fecal_S1_L001_R1_001.fastq        62708     48819
zbStandard_S1_L001_R1_001.fastq   62335     47330
```

out <- filterAndTrim(fnFs, filtFs, fnRs, filtRs, truncLen=c(305,225), trimLeft=c(16,24), maxN=0, maxEE=c(5,4), truncQ=2, rm.phix=TRUE, compress=TRUE, multithread=TRUE)

```
[mweinste@n7282 ws11]$ module load python/3.6.1

The 'gcc/4.9.3' module is being loaded

        These modules were already loaded: ATS intel/18.0.4

[mweinste@n7282 ws11]$ python3 figaro/figaro.py --ampliconLength 470 --forwardPrimerLength 16 --reversePrimerLength 24
{"trimPosition": [305, 225], "maxExpectedError": [5, 4], "readRetentionPercent": 76.9, "score": 51.896132561338405}
{"trimPosition": [304, 226], "maxExpectedError": [5, 4], "readRetentionPercent": 76.28, "score": 51.28354819103676}
{"trimPosition": [303, 227], "maxExpectedError": [5, 4], "readRetentionPercent": 75.9, "score": 50.904481622468 89}
{"trimPosition": [302, 228], "maxExpectedError": [5, 5], "readRetentionPercent": 80.83, "score": 48.8339464508493}
{"trimPosition": [301, 229], "maxExpectedError": [5, 5], "readRetentionPercent": 80.68, "score": 48.6835993730207}
{"trimPosition": [300, 230], "maxExpectedError": [5, 5], "readRetentionPercent": 80.6, "score": 48.59563033812098}
{"trimPosition": [299, 231], "maxExpectedError": [5, 5], "readRetentionPercent": 80.52, "score": 48.52045679920667}
{"trimPosition": [298, 232], "maxExpectedError": [5, 5], "readRetentionPercent": 80.3, "score": 48.30133393045648}
{"trimPosition": [297, 233], "maxExpectedError": [5, 5], "readRetentionPercent": 80.25, "score": 48.25335082051119}
```

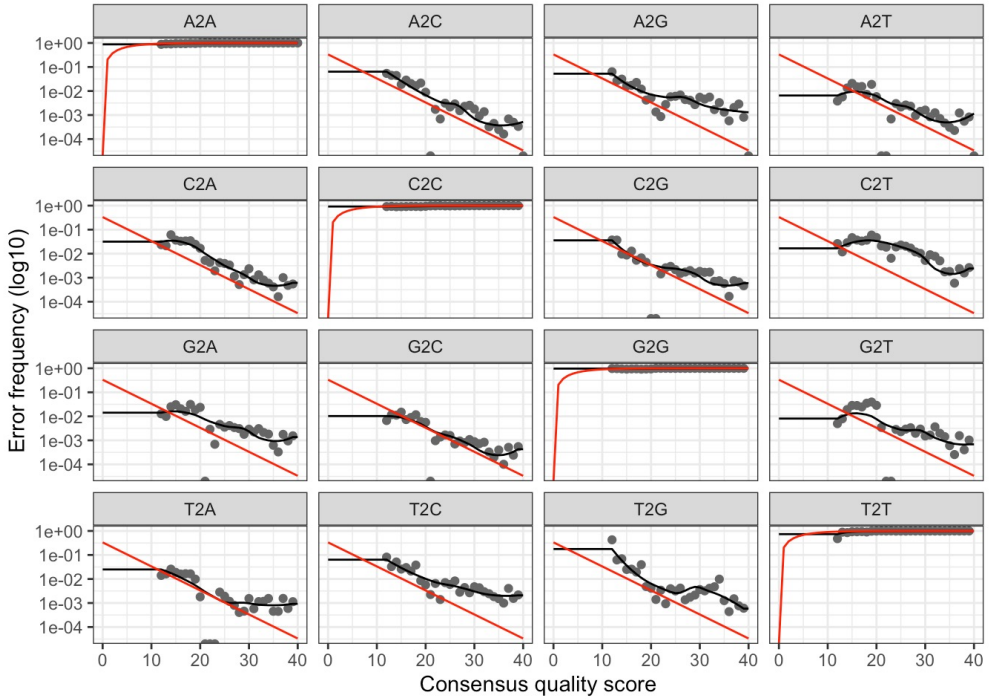
Learn Forward and Reverse Errors

```
>
> errF <- learnErrors(filtFs, multithread=TRUE)
27787061 total bases in 96149 reads from 2 samples will be used for learning the error rates.
> errR <- learnErrors(filtRs, multithread=TRUE)
19325949 total bases in 96149 reads from 2 samples will be used for learning the error rates.
>
```

Visualizing the Error Models

It is always worthwhile, as a sanity check if nothing else, to visualize the estimated error rates:

```
plotErrors(errF, nominalQ=TRUE)
```



The error rates for each possible transition (A→C, A→G, …) are shown. Points are the observed error rates for each consensus quality score. The black line shows the estimated error rates after convergence of the machine-learning algorithm. The red line shows the error rates expected under the nominal definition of the Q-score. Here the estimated error rates (black line) are a good fit to the observed rates (points), and the error rates drop with increased quality as expected. Everything looks reasonable and we proceed with confidence.

Apply the Denoising

```
> dadaFs <- dada(filtFs, err=errF, multithread=TRUE)
Sample 1 - 48819 reads in 23762 unique sequences.
Sample 2 - 47330 reads in 25257 unique sequences.
> dadaRs <- dada(filtRs, err=errR, multithread=TRUE)
Sample 1 - 48819 reads in 34623 unique sequences.
Sample 2 - 47330 reads in 36509 unique sequences.
> dadaFs[[1]]
dada-class: object describing DADA2 denoising results
192 sequence variants were inferred from 23762 input unique sequences.
Key parameters: OMEGA_A = 1e-40, OMEGA_C = 1e-40, BAND_SIZE = 16
> dadaRs[[1]]
dada-class: object describing DADA2 denoising results
95 sequence variants were inferred from 34623 input unique sequences.
Key parameters: OMEGA_A = 1e-40, OMEGA_C = 1e-40, BAND_SIZE = 16
```

Attempt to Merge Read Pairs

```
> mergers <- mergePairs(dadaFs, filtFs, dadaRs, filtRs, verbose=TRUE)
46126 paired-reads (in 154 unique pairings) successfully merged out of 47970 (in 411 pairings) input.
45087 paired-reads (in 38 unique pairings) successfully merged out of 47156 (in 124 pairings) input.
> head(mergers[[1]])


                                                    sequence
1 GTAGGGAATTTTCGTCAATGGGGGGAACCCTGAACGAGCAATGCCGCGTGAGTGAGGAAGGTCTTCGGATCGTAAAGCTCTGTTGTAAGAGAAAAACGACATTCATAGGGAATGATGAGTGAGTGATGGTATCTTACCAGAAAGTCACGGCTAACTACGTGCCAGCAGCCGCGGTAATAC
GTAGGTGGCGAGCGTTATCCGGAATGATTGGGCGTAAAGGGTGCGTAGGTGGCAGAACAAGTCTGGAGTAAAAGGTATGGGCTCAACCCGTACTGGCTCTGGAAACTGTTCAGCTAGAGAACAGAAGAGGACGGCGGAACTCCATGTGTAGCGGTAAAATGCGTAGATATATGGAAGAACAC
CGGTGGCGAAGGCGGCCGTCTGGTCTGTTGCTGACACTGAAGCACGAAAGCGTGGGGAGCAA
2     GTGAGGAATATTGGTCAATGGGCGCAGGCCTGAACCAGCCAAGTAGCGTGAAGGATGACTGCCCTATGGGTTGTAAACTTCTTTTATAAAGGAATAAAGTCGGGTATGTATACCCGTTTGCATGTACTTTATGAATAAGGATCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACG
GAGGATCCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGATGGATGTTTAAGTCAGTTGTGAAAGTTTGCGGCTCAACCGTAAAATTGCAGTTGATACTGGATATCTTGAGTGCAGTTGAGGCAGGCGGAATTCGTGGTGTAGCGGTGAAATGCTTAGATATCACGAAGAACTC
CGATTGCGAAGGCAGCCTGCTAAGCTGCAACTGACATTGAGGCTCGAAAGTGTGGGTATCAA
3     GTGAGGAATATTGGTCAATGGGCGAGAGCCTGAACCAGCCAAGTAGCGTGAAGGATGACTGCCCTATGGGTTGTAAACTTCTTTTATAAAGGAATAAAGTCGGGTATGCATACCCGTTTGCATGTACTTTATGAATAAGGATCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACG
GAGGATCCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGATGGATGTTTAAGTCAGTTGTGAAAGTTTGCGGCTCAACCGTAAAATTGCAGTTGATACTGGATATCTTGAGTGCAGTTGAGGCAGGCGGAATTCGTGGTGTAGCGGTGAAATGCTTAGATATCACGAAGAACTC
CGATTGCGAAGGCAGCCTGCTAAGCTGCAACTGACATTGAGGCTCGAAAGTGTGGGTATCAA
4                     GTGGGGAATATTGCACAATGGGGGAAACCCTGATGCAGCGACGCCGCGTGAAGGAAGAAGTATCTCGGTATGTAAACTTCTATCAGCAGGGAAGATAGTGACGGTACCTGACTAAGAAGCCCCGGCTAACTACGTGCCAGCAGCCGCGGTAATACG
TAGGGGGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGTAGACGGTGTGGCAAGTCTGATGTGAAAGGCATGGGCTCAACCTGTGGACTGCATTGGAAACTGTCATACTTGAGTGCCGGAGGGGTAAGCGGAATTCCTAGTGTAGCGGTGAAATGCGTAGATATTAGGAGGAACAC
CAGTGGCGAAGGCGGCTTACTGGACGGTAACTGACGTTGAGGCTCGAAAGCGTGGGGAGCAA
5                     GTGGGGAATATTGCACAATGGGGGAAACCCTGATGCAGCGACGCCGCGTGAGCGATGAAGTATTTCGGTATGTAAAGCTCTATCAGCAGGGAAGAAAATGACGGTACCTGACTAAGAAGCACCGGCTAAATACGTGCCAGCAGCCGCGGTAATACG
TATGGTGCAAGCGTTATCCGGATTTACTGGGTGTAAAGGGAGCGTAGACGGAGTGGCAAGTCTGATGTGAAAACCCGGGGCTCAACCCCGGGACTGCATTGGAAACTGTCAATCTAGAGTACCGGAGAGGTAAGCGGAATTCCTAGTGTAGCGGTGAAATGCGTAGATATTAGGAGGAACAC
CAGTGGCGAAGGCGGCTTACTGGACGGTAACTGACGTTGAGGCTCGAAAGCGTGGGGAGCAA
6     GTGAGGAATATTGGTCAATGGACGAGAGTCTGAACCAGCCAAGTAGCGTGAAGGATGACTGCCCTATGGGTTGTAAACTTCTTTTATACGGGAATAAAGTGAGGCACGCGTGCCTTTTTGTATGTACCGTATGAATAAGGATCGGCTAACTCCGTGCCAGCAGCCGCGGTAATACG
GAGGATCCGAGCGTTATCCGGATTTATTGGGTTTAAAGGGAGCGTAGGCGGACGCTTAAGTCAGTTGTGAAAGTTTGCGGCTCAACCGTAAAATTGCAGTTGATACTGGGTGTCTTGAGTACAGTAGAGGCAGGCGGAATTCGTGGTGTAGCGGTGAAATGCTTAGATATCACGAAGAACTC
CGATTGCGAAGGCAGCTTGCTGGACTGTAACTGACGCTGATGCTCGAAAGTGTGGGTATCAA
  abundance forward reverse nmatch nmismatch nindel prefer accept
1      4234       1       2     66         0      0      1   TRUE
2      3308       2       1     70         0      0      1   TRUE
3      3031       3       1     70         0      0      2   TRUE
4      2154       4       9     90         0      0      1   TRUE
5      1954       5       8     90         0      0      1   TRUE
6      1737       6       4     70         0      0      1   TRUE
>
```

Inspect Our Denoised Amplicons

```
> seqtab <- makeSequenceTable(mergers)
> dim(seqtab)
[1]   2 192
> table(nchar(getSequences(seqtab)))

399 400 401 402 403 404 405 419 420 424 425 426
  4  63   4  10   7   1   1   5  45   3  42   7
>
```

Remove Chimeras and Examine Effect

```
> seqtab.nochim <- removeBimeraDenovo(seqtab, method="consensus", multithread=TRUE, verbose=TRUE)
Identified 67 bimeras out of 192 input sequences.
> dim(seqtab.nochim)
[1]   2 125
> table(nchar(getSequences(seqtab.nochim)))

399 400 401 402 403 404 405 419 420 424 425 426
  1  54   4   8   7   1   1   3  23   3  17   3
> sum(seqtab.nochim)/sum(seqtab)
[1] 0.9858463
>
```

Track Our Reads

```
> getN <- function(x) sum(getUniques(x))
> track <- cbind(out, sapply(dadaFs, getN), sapply(dadaRs, getN), sapply(mergers, getN), rowSums(seqtab.nochim))
> colnames(track) <- c("input", "filtered", "denoisedF", "denoisedR", "merged", "nonchim")
> rownames(track) <- sample.names
> head(track)
           input filtered denoisedF denoisedR merged nonchim
fecal      62708    48819     48282     48417  46126   45676
zbStandard 62335    47330     47211     47259  45087   44246
>
```

Assign Taxa

```
> taxa <- assignTaxonomy(seqtab.nochim, "/u/scratch/m/mweinste/ws11/data/silva_nr_v138_train_set.fa.gz", multithread=TRUE)
> taxa <- addSpecies(taxa, "/u/scratch/m/mweinste/ws11/data/silva_species_assignment_v138.fa.gz")
> taxa.print <- taxa
> rownames(taxa.print) <- NULL
> head(taxa.print)
     Kingdom    Phylum          Class                Order
[1,] "Bacteria" "Firmicutes"    "Bacilli"            "Bacillales"
[2,] "Bacteria" "Firmicutes"    "Bacilli"            "Lactobacillales"
[3,] "Bacteria" "Firmicutes"    "Bacilli"            "Staphylococcales"
[4,] "Bacteria" "Firmicutes"    "Bacilli"            "Lactobacillales"
[5,] "Bacteria" "Firmicutes"    "Bacilli"            "Erysipelotrichales"
[6,] "Bacteria" "Proteobacteria" "Gammaproteobacteria" "Enterobacterales"
     Family               Genus                 Species
[1,] "Bacillaceae"        "Bacillus"            NA
[2,] "Listeriaceae"       "Listeria"            NA
[3,] "Staphylococcaceae"  "Staphylococcus"      NA
[4,] "Enterococcaceae"    "Enterococcus"        NA
[5,] "Erysipelotrichaceae" "Holdemanella"       "biformis"
[6,] "Enterobacteriaceae" "Escherichia/Shigella" NA
> write.table(taxa.print, "taxaResults.txt", sep = "\t", row.names = TRUE, col.names = TRUE)
>
```

Taxa table is now in your ws11 folder