# Kmer Spectrum Primer

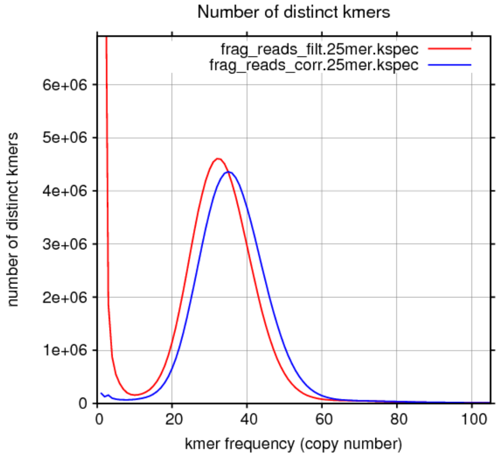
The kmer spectrum of a data set can be very helpful when diagnosing problems with an assembly or assessing data quality. ALLPATHS-LG will compute a few kmer spectra at different stages. To visualize these spectra the user can run the command `KmerSpectrumPlot.pl`, which generates and runs a `gnuplot` script, which in turn, generates three `.eps` files (and alternatively three `.gif` files).

```
KmerSpectrumPlot.pl
SPECTRA=”{*{filt,corr}*.kspec}” GIF=True
```

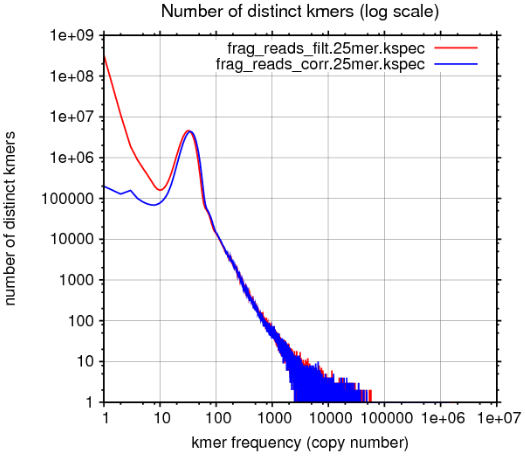
This command generates the `.eps` files:

```
kmer_spectrum.distinct.lin.lin.eps

kmer_spectrum.distinct.log.log.eps

kmer_spectrum.cumulative_frac.log.lin.eps
```

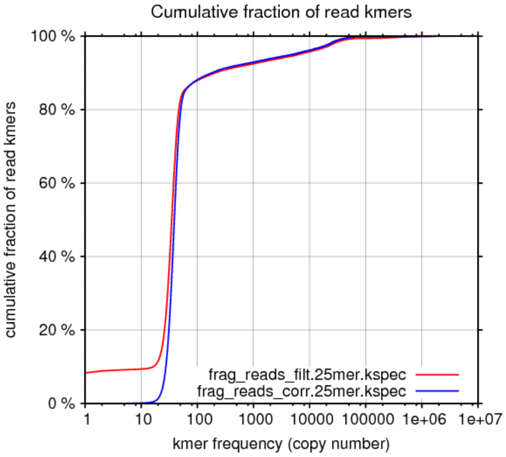
Because the option `GIF=True` is specified, corresponding `.gif` files are also generated.



The first plot depicts the kmer spectra at K=25 for the filtered fragment reads (red) and corrected fragment reads (blue). The filtered reads (red) are not error corrected, hence the strong divergence at low kmer frequency and the slight shift to lower kmer frequencies compared to the corrected reads (blue). After error correction, the divergence is gone. The closer the low frequency corrected spectrum is to zero, the better the error correction algorithm worked.



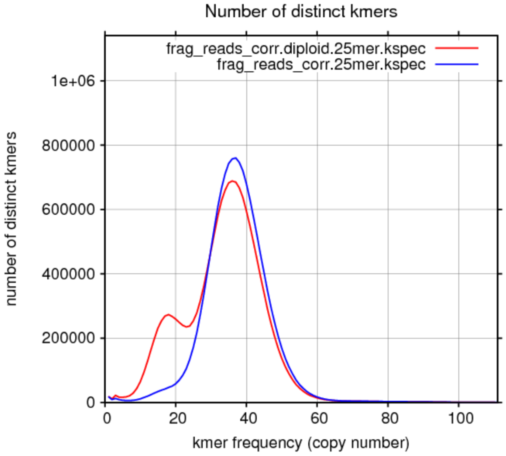
The second plot depicts the same spectra as the first except that both scales are logarithmic. One can see the dramatic effect that error correction has on the very low frequency end of the spectra. It is also clear the power law distribution at high frequencies, roughly proportional to (kmer frequency)$^{-2}$.



The third plot depicts the cumulative fraction of all **kmers in the reads** as a function of frequency. The spectrum for the filtered reads starts off at about 10%. This means that 10% of all the kmers in all the reads have very low frequencies and are most likely associated with errors. However, remember that a single base error in a read spoils K kmers, so the base error rate is not 10%. The base error rate is more like 10% / K. The corrected cumulative spectrum (blue) starts a 0%, as expected for a mostly error free data set.

At the high frequency side of the third plot reside the kmers associated with repetitiveness or copy numbers greater than one. In the case shown, the knee of the blue curve is roughly at 85%, which means that the fraction of the genome that is repetitive is about 15%.

## Highly Polymorphic Genomes

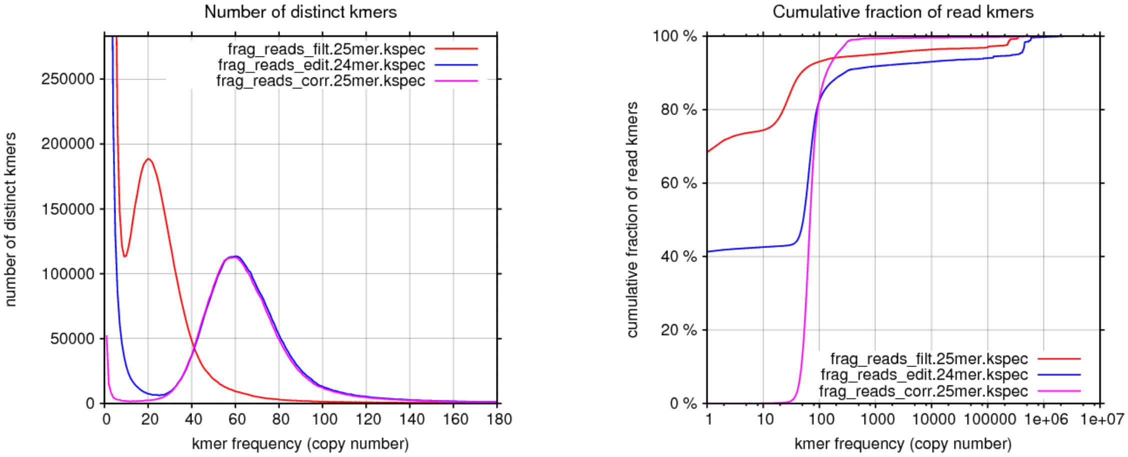


In highly polymorphic genomes the kmer spectrum can be used to estimate the polymorphic rate. In the kmer spectrum plot to the right, the red line depicts the kmer spectrum of *candida albicans* which has polymorphic rate of about 1/250. There are two peaks: the left peak corresponds to the kmers that are associated with SNPs (and other polymorphisms), while the right peak corresponds to the kmers common to both haplotypes. Counting the kmers on each peak allows for the estimation of the polymorphic rate. The blue line depicts the spectrum after ALLPATHS-LG has removed (when run with the option `HAPLOIDIFY=True`) most of the polymorphisms from the data set. Note that polymorphic rates as high as 1/100, or even higher, are not that uncommon even in large vertebrate genomes. *Candida* is definitely not an outlier.

## Poor data quality

Sometimes, the quality of sequencing data is poor. The two plots below (for *rhodobacter*) clearly show the effects of poor data quality on the kmer spectrum.

The plots show the kmer spectra at three different stages: before error correction (*filtered* reads, red), after error correction but before unique removal (*edited* reads, blue), and after error correction (*corrected* reads, magenta).

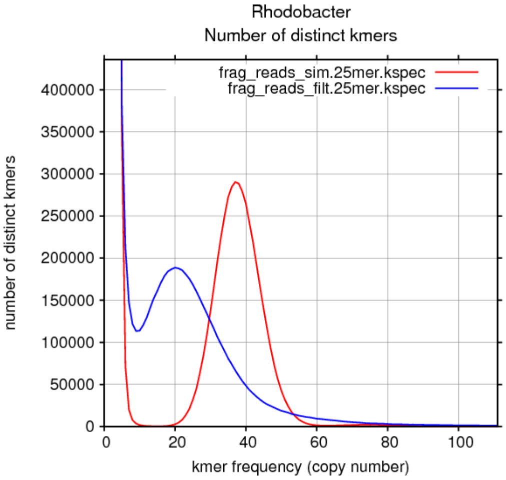


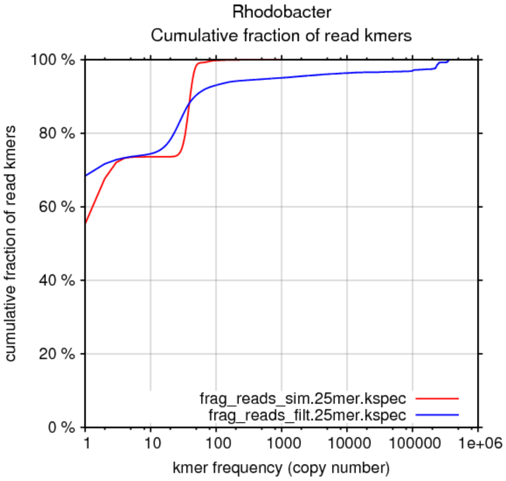
As can be seen on the cumulative plot on the right, the initial error rate is very high. More than 70% of kmers have errors (red line at low kmer frequencies) which translates to a base error rate of about 3%. After error correction (blue line) only 40% of kmers are left with errors. Finally, the magenta line depicts the cumulative number of kmers after the removal of all the reads with faulty kmers.

## High Sequencing Bias

*Rhodobacter* is known for its very high GC content which, in general, is associated with high sequencing bias.

The plots on the right compare the kmer spectra of real, highly biased, *rhodobacter* sequencing data (in blue) with zero bias, simulated data (in red) at the same coverage. The error rate of the simulated data is similar has can be assessed by the cumulative plot.


Rhodobacter
Number of distinct kmers


Rhodobacter
Cumulative fraction of read kmers

The most obvious effect is that the main peak is shifted to the left when bias is high. This means that some regions of the genome are not being covered as well as one would expect. In this case, most of the genome is being covered only half as much as expected. This can be a problem for error correction because, at low kmer frequencies, there isn't a very clear distinction between genomic and error kmers.

A more subtle effect of bias is that other regions of the genome are deeply covered. Kmers in those regions show up in the long tail of the main peak (in blue). In the cumulative plot, these presumed copy number 1 kmers are spread out through various orders of magnitude in kmer frequency which translates into the slow approach of the blue line to 100% (which is absent in the simulated case).

Fortunately (in this case), the error correction step of removing reads with faulty kmers, also reduces some of this bias. Referring back to the cumulative plot in the previous page, the magenta line (fully corrected reads) approaches 100% much faster than the line for non-corrected reads.