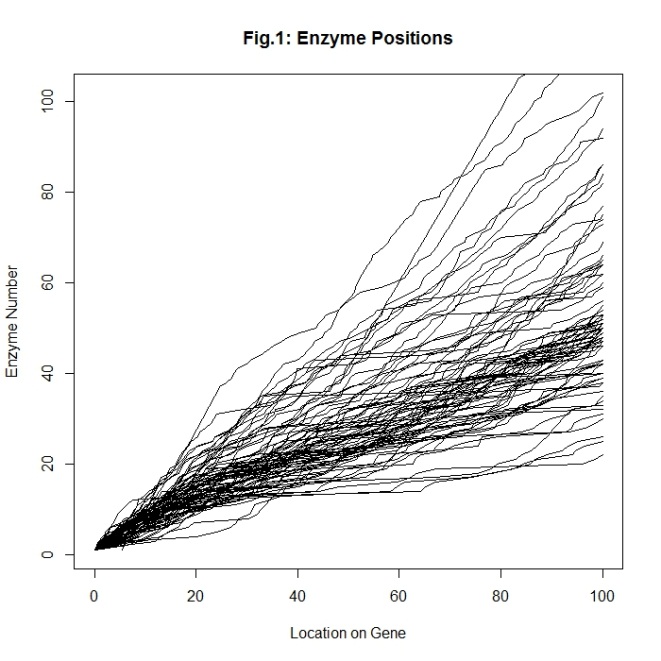
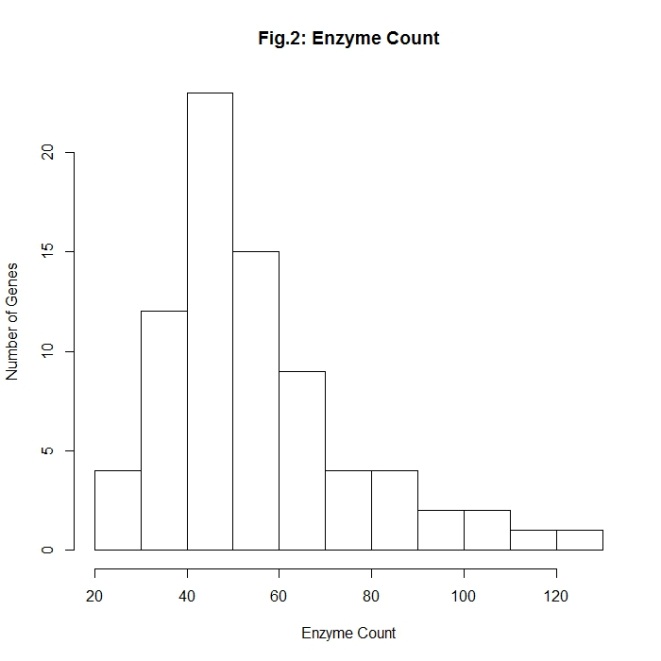
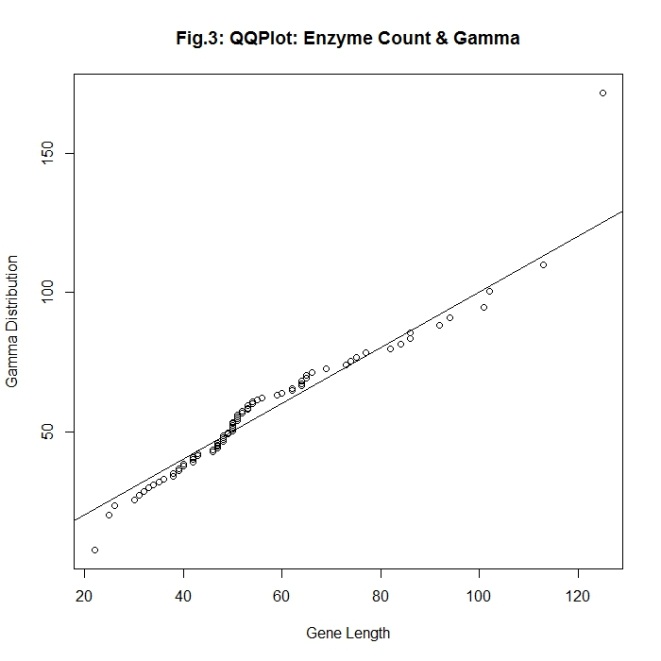
Our analysis started by looking at the raw polymerase location data along each gene. It can be seen that there is some considerable variation in the number of enzymes (Fig.1).

A histogram of the enzyme number per gene (Fig.2) shows that there is a definite trend.

The number of enzymes in the genes follows a gamma distribution, at least as a first order approximation, with shape = 7.554547 & rate = 0.1362295 as derived from the mean and variance of the enzyme counts. This can be seen from the quantile-quantile plot of the data against the gamma distribution. (Fig.3).

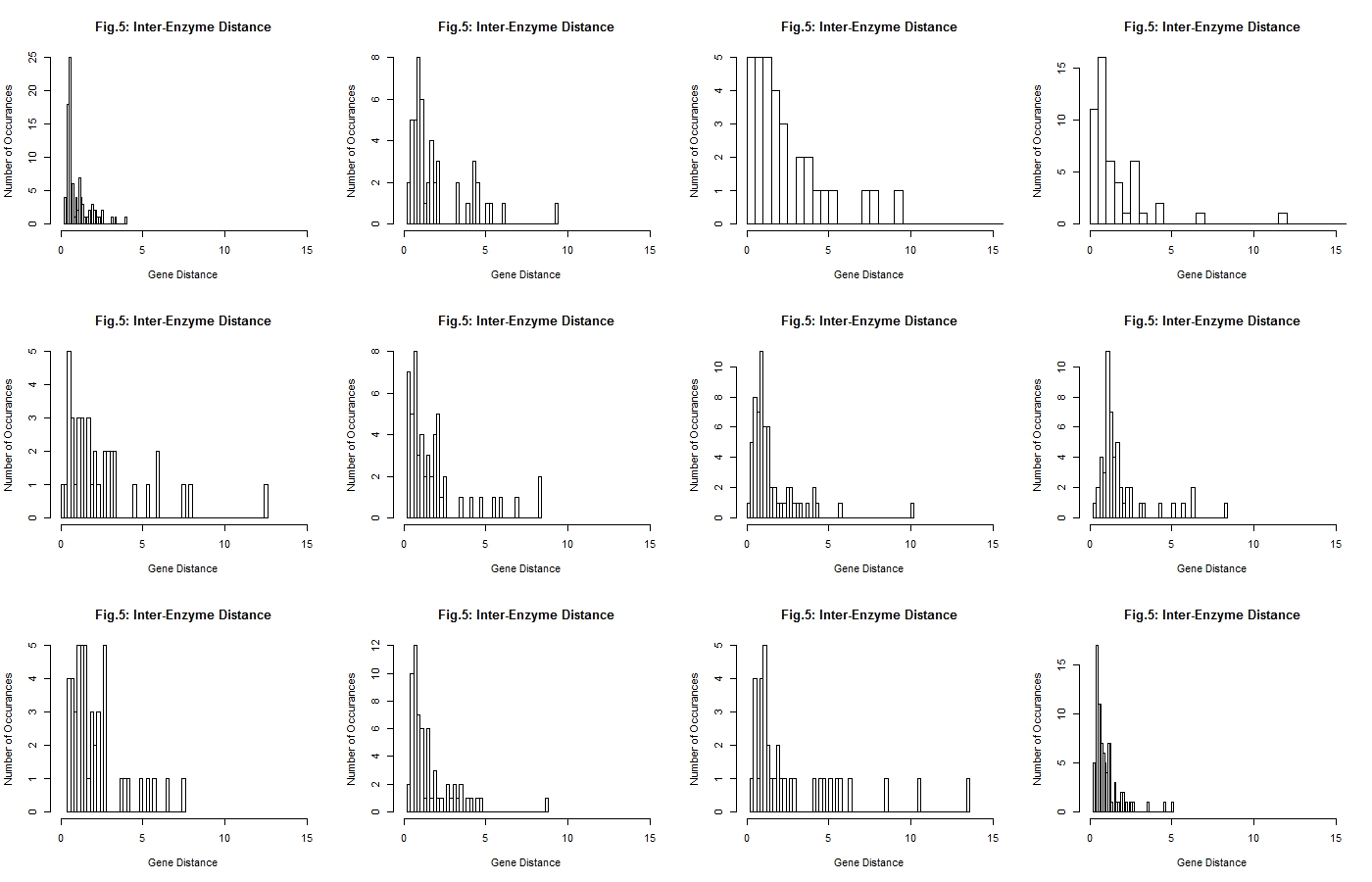


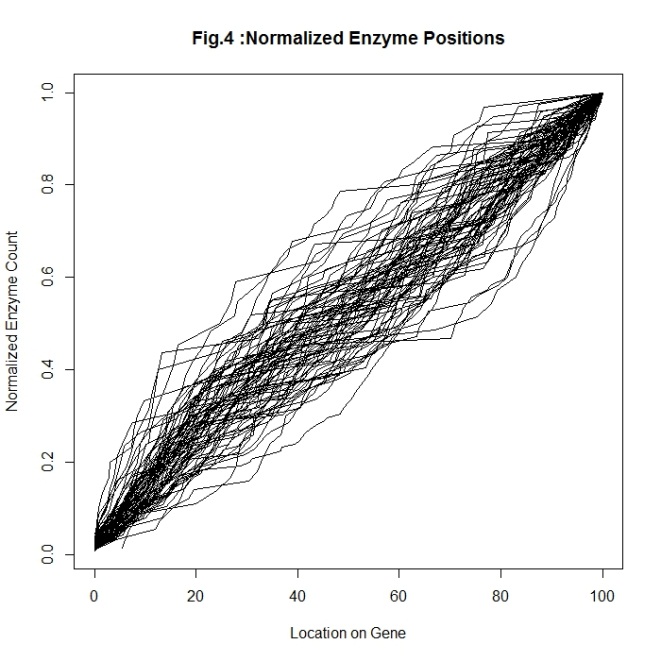




Normalizing the enzyme locations by length (Fig.4) shows that the enzyme locations along a gene are spaced in such a way there is no systematic increase or decrease in their interspacing along the gene readily apparent to the eye.

Analyzing the inter-enzyme spacing in further depth we can plot out a distribution of the inter-enzyme spacings along a number of genes, which shows that they all exhibit a similar trend suggesting they may come from the same distribution (Fig.5)

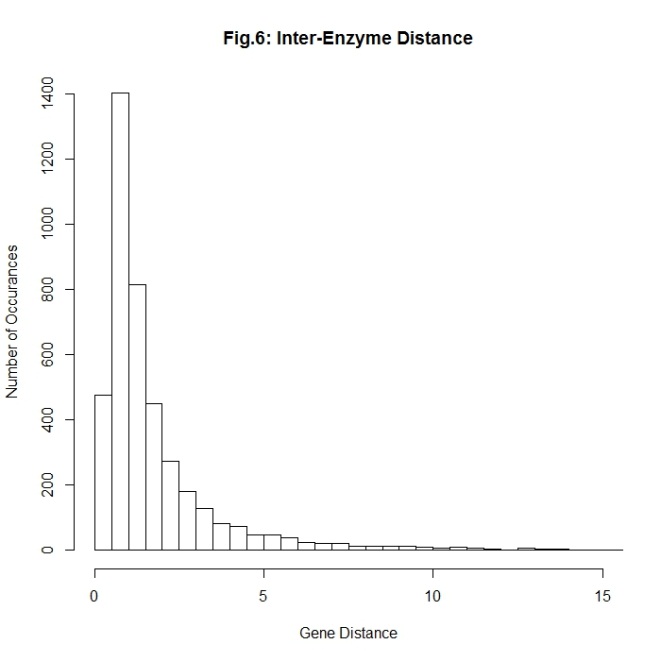


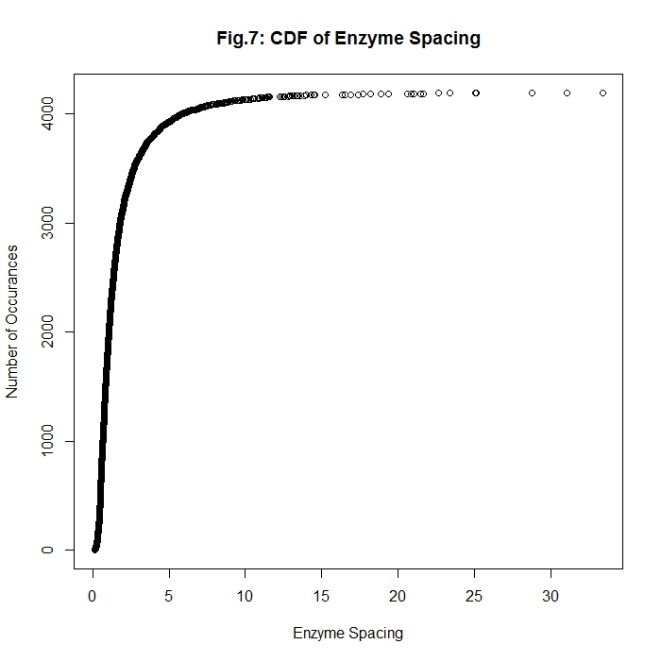


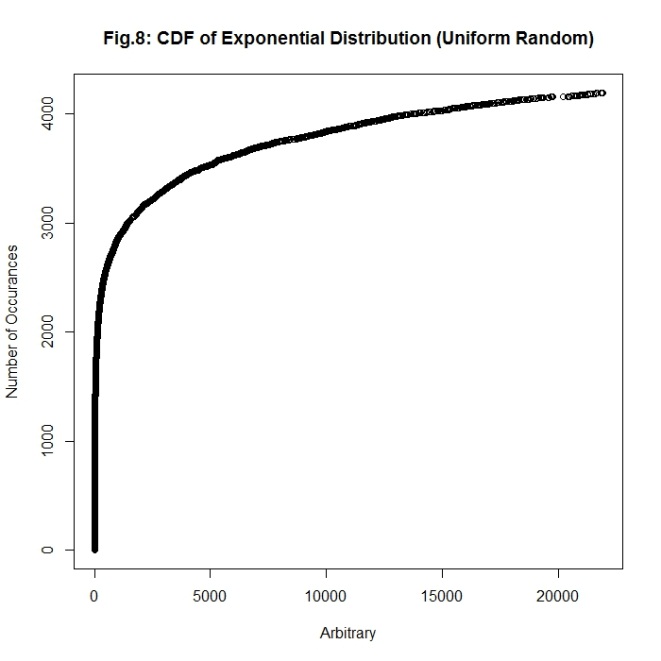
Plotting all of the inter-enzyme spacings across the 77 genes in the same distribution yields the following histogram which we would like to describe mathematically (Fig.6)

Looking at the cumulative distribution function we can see that the inter-enzyme spacing distribution (Fig.7) has a distinct shape.

The CDF of an exponential (Fig.8) is a close match but does not quite replicate the behavior.



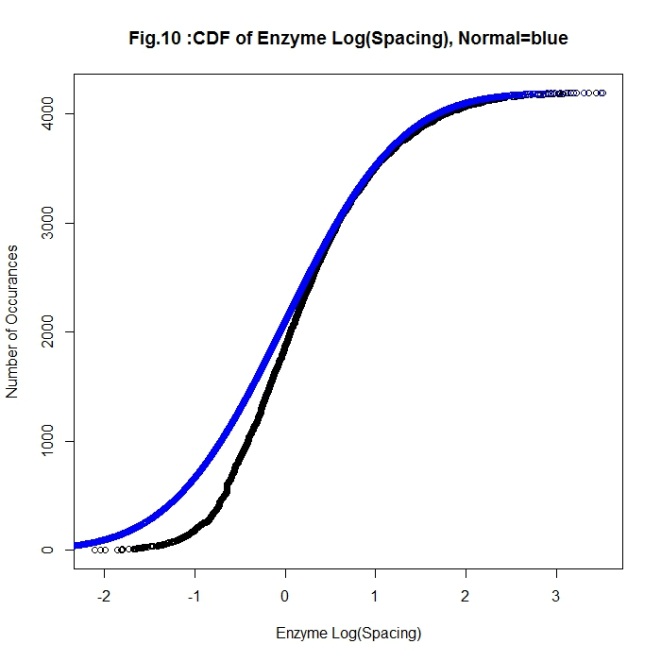
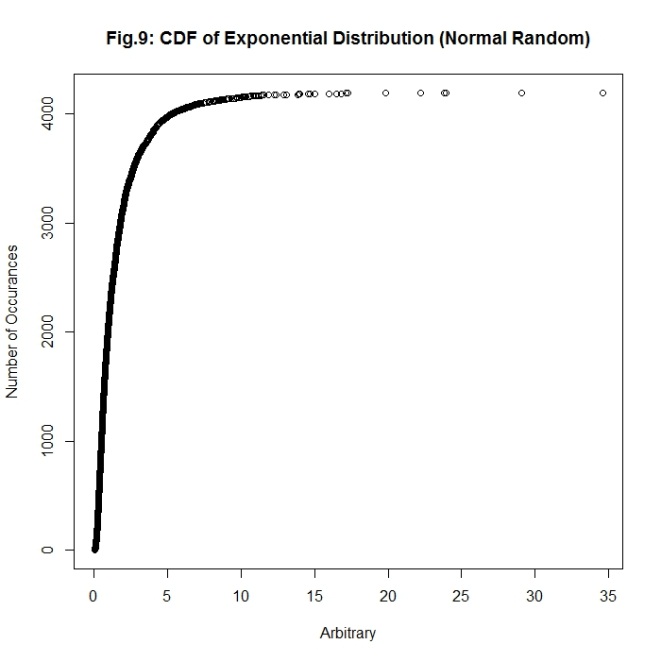


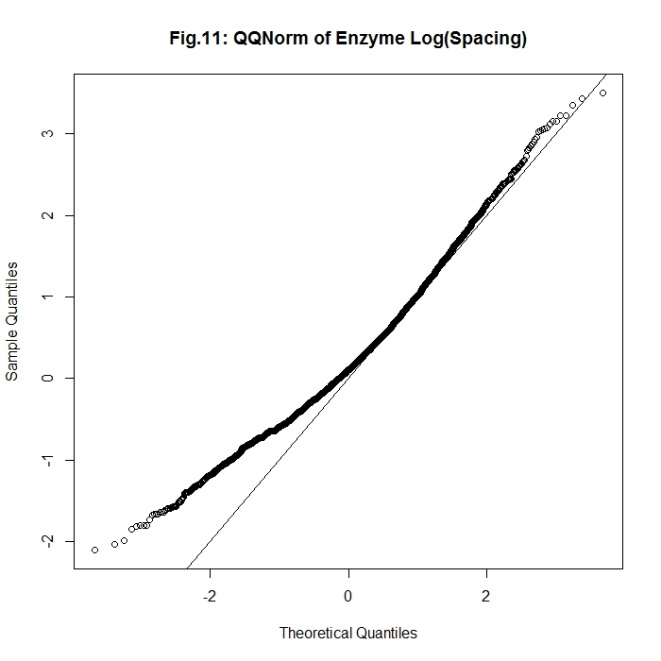


However the CDF of an exponential drawn from a normal distribution (Fig.9) is a very close match to inter-enzyme spacing.

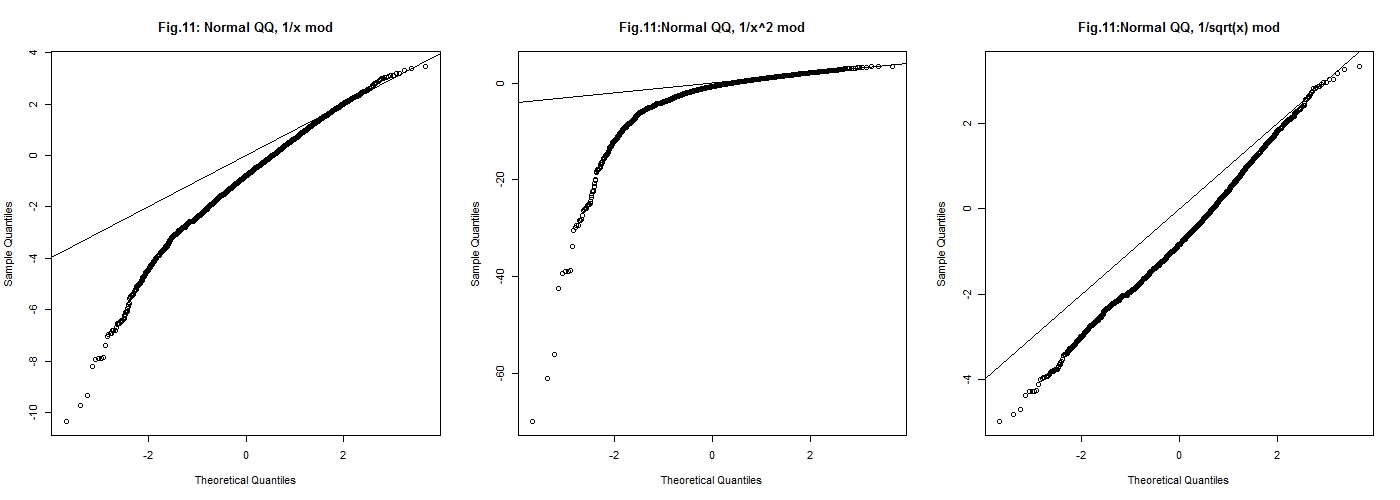
Taking the log of the inter-enzyme spacing we can see however that our distribution was not drawn from a completely normal distribution, as the CDF of the normal distribution varies at the lower extreme (Fig.10).

Plotting a QQNorm of the log of the enzyme spacings we can see once again we’re neglecting a term at the lower extreme of the distribution (Fig.11)





To correct this we need to add a term to the model. We know that the deviation happens most pronouncedly at lower values which would suggest that it is an inverse of the inter-spacing distance. Trying the inverse of some common powers : 0 , 2 , ½. We find that by adding an inverse square root we can create a linear QQNorm plot from our spacings, though one that does not fall along the expected line (Fig.11).



To tune our model we can perform a regression to fit the distribution to that of a normal

Call:

lm(formula = normVals ~ I(log(geneDiffsort)) + I(1/sqrt(geneDiffsort)))

Residuals:

Min 1Q Median 3Q Max

-0.12418 -0.01137 0.01182 0.02128 0.52959

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.234940 0.006884 179.4 <2e-16 \*\*\*

I(log(geneDiffsort)) 0.593492 0.002898 204.8 <2e-16 \*\*\*

I(1/sqrt(geneDiffsort)) -1.389865 0.006432 -216.1 <2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.04246 on 4190 degrees of freedom

Multiple R-squared: 0.9982, Adjusted R-squared: 0.9982

F-statistic: 1.156e+06 on 2 and 4190 DF, p-value: < 2.2e-16

We find that with the correct coefficients our model can be made to have a normal distribution with a coefficient of determination of R2 = 0.9982, which implies that this is a very good fit.

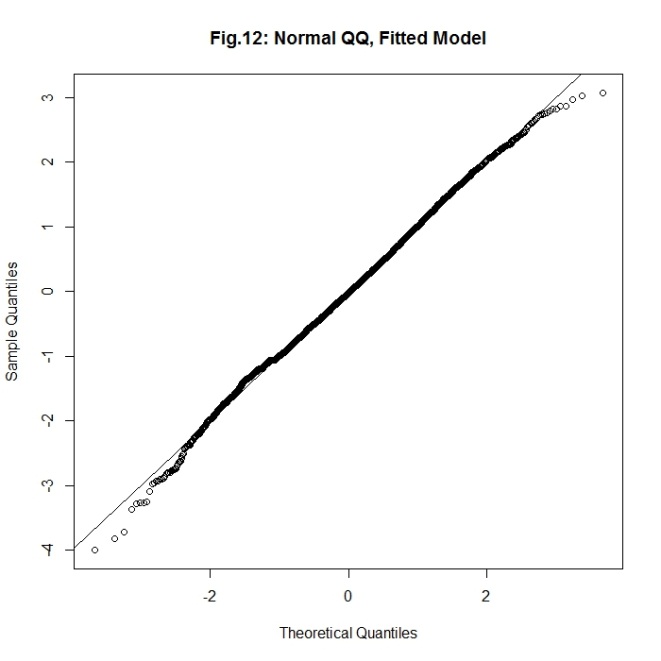
The distribution of the inter-enzyme spacings is thusly as follows:

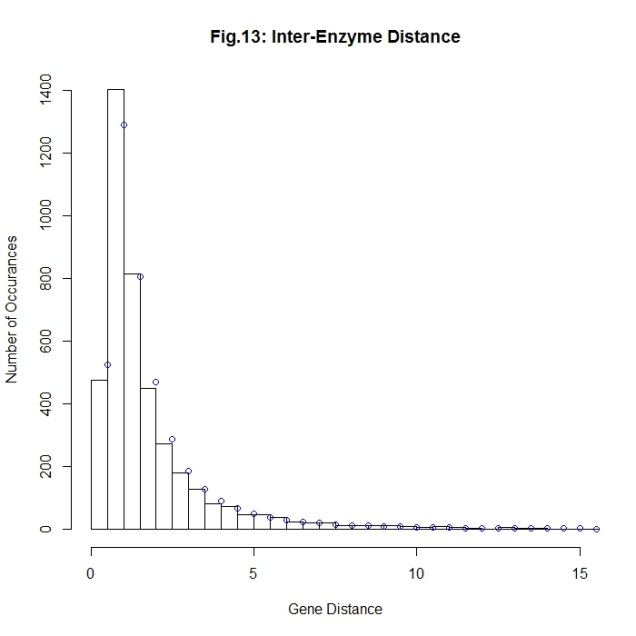
**The statistic [1.234940 + 0.593492\*log(spacing) - 1.389865/sqrt(spacing)] follows a standard normal distribution.**

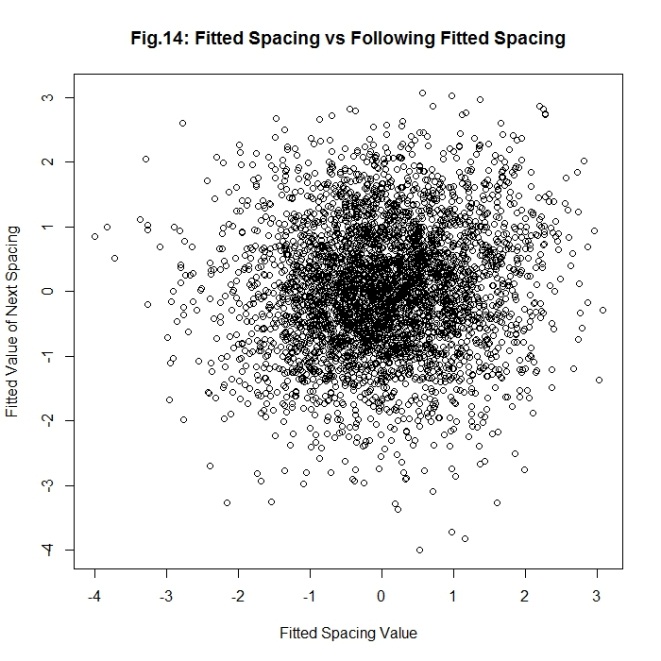
Plotting a QQNorm of the new enzyme spacing derived statistic we can see our model is now a good description of the data. (Fig.12)

The probability of the model producing a given enzyme spacing can be calculated and plotted against the histogram of the spacings (Fig.13) showing good agreement.

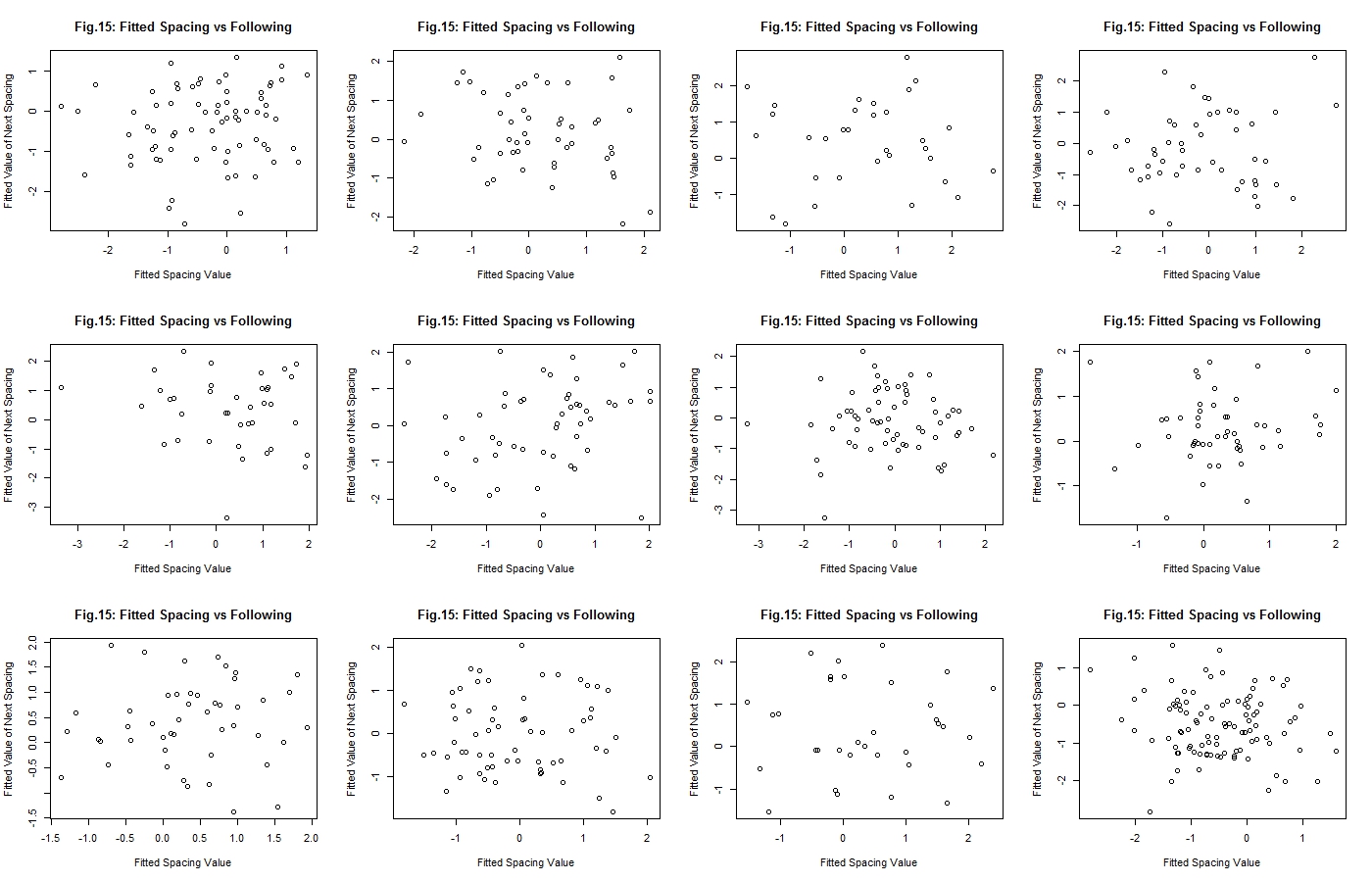
Lastly it is important to show that the inter-enzyme spacing is coming randomly from the decribed distribution and is not correlated in any way to the proceeding enzyme spacing on the gene. Plotting the statistic we arrived at for pairs of neighboring enzyme spacings in a scatter plot we find that with R2=.014 there is no correlation of one spacing value on the other. The graph is what you’d expect for paired values randomly drawn from a normal distribution, and our statistic is normally distributed. (Fig.14)



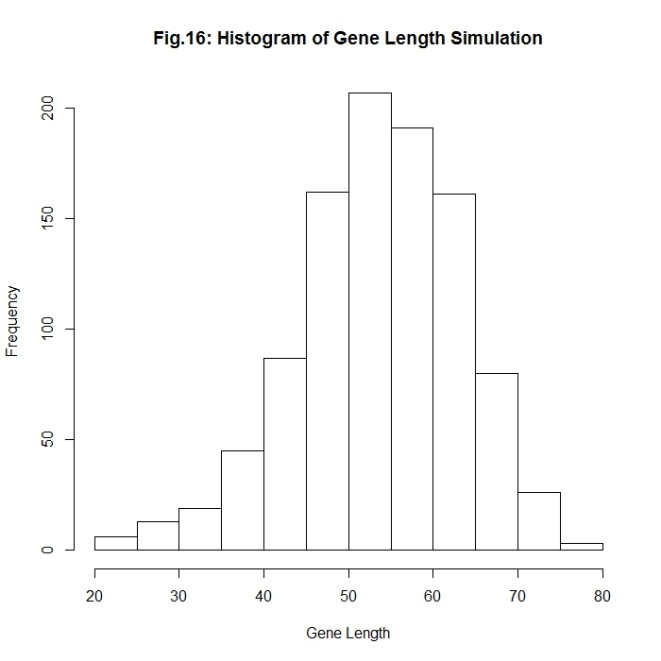




Plotting the distribution of the statistic for a number of different samples we can see that we are not missing any sort of systematic behavior by looking at the spacing values as a larger set and they all exhibit the same behavior. (Fig.15)



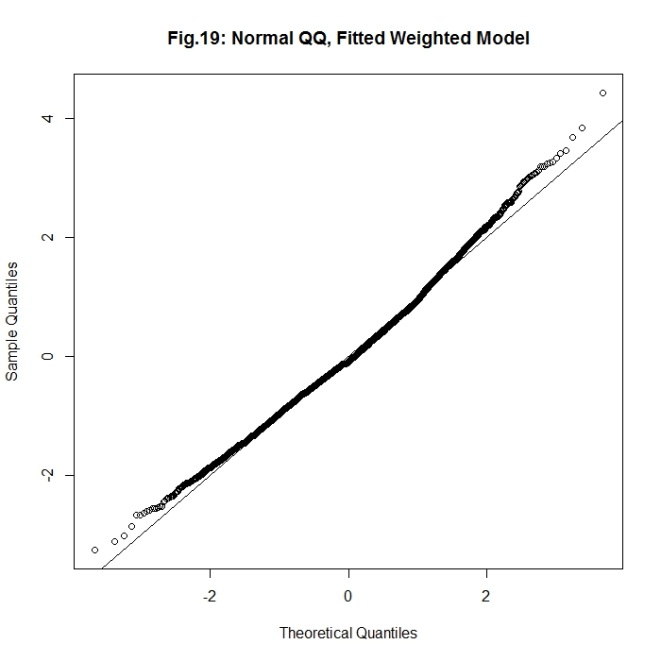
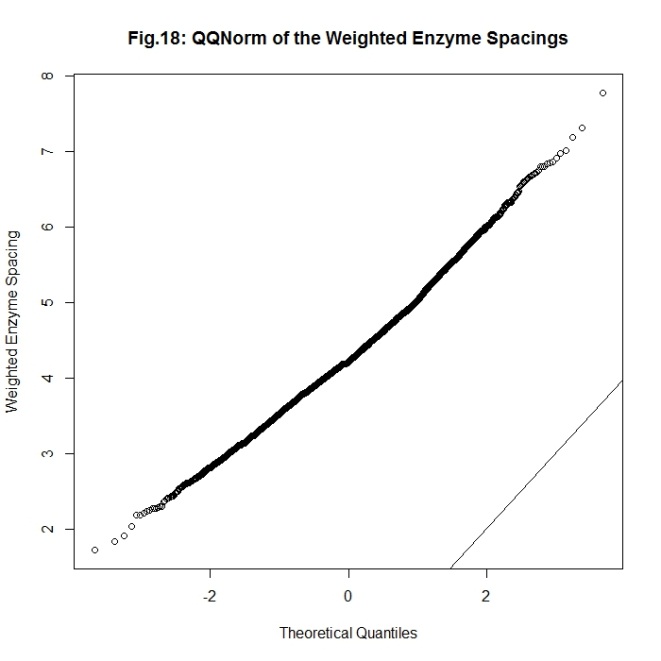
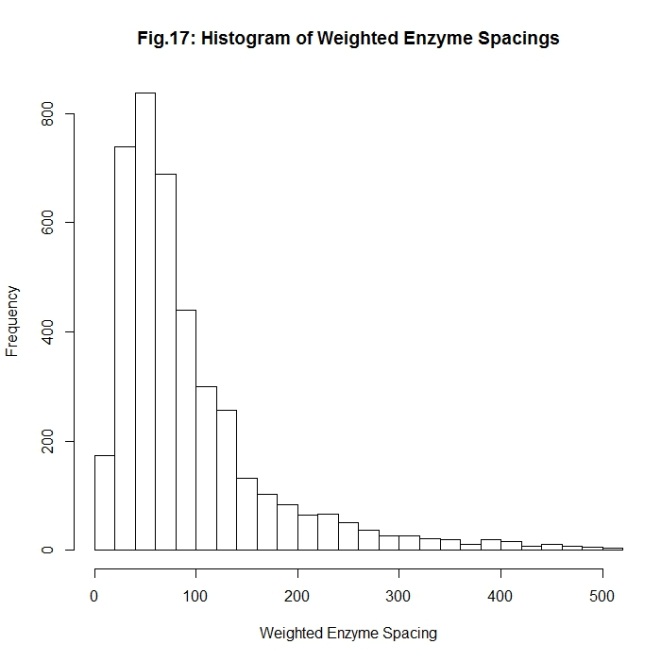
However running a simulation for the expected gene length using this distribution it can be seen that something doesn’t add up in our analysis as the simulation does not match the data (Fig. 16)



This led to the realization that the deviation from the log normal distribution found earlier simply be explaining some gene-length dependent behavior on the inter-enzyme spacing, which would follow from the percentage nature. Thusly it may be more appropriate to study the inter-enzyme spacing weighted by gene length (Fig.17)

A QQNorm plot of the log of the weighted data shows that the distribution is likely log-normal (Fig.18)

Fitting the log of the weighted data against a normal distribution it is possible to refine the log-normal fit in the same manner as before, and we find a R2=0.993 coefficient of determination(Fig.19)



lm(formula = normVals ~ I(log(geneWeightDiffListSort)))

Residuals:

Min 1Q Median 3Q Max

-0.94121 -0.04058 0.01255 0.07188 0.08923

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -5.433948 0.007171 -757.8 <2e-16 \*\*\*

I(log(geneWeightDiffListSort)) 1.268554 0.001647 770.4 <2e-16 \*\*\*

As before we can plot the probability density of our model against the histogram of the weighted data and see that it’s in good agreement (Fig.20)

And by plotting the weighted and fitted log values of the enzyme inter-spacing against the values following them we can see there is no correlation, with R2=0.0007 (Fig.21)

