# Statistical Methods for Troubleshooting of Genomic Shotgun Data.

Eugene Goltsman, Vasanth Singan, Stephan Trong, Alla Lapidus, Kerrie Barry,
Alex Copeland

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

## INTRODUCTION

1. Abnormalities in WGS datasets present assembly problems that are expensive and time-consuming to resolve. **Cloning bias, contamination and large repeats** in genomic DNA are the three biggest obstacles requiring a great deal of human involvement and expertise.
2. Shotgun libraries and assemblies exhibit clearly recognizable patterns that have been shown to fit certain known statistical models.
3. Deviations from these models often stem from flaws and abnormalities in the input data, which may reflect problems in the protocol, chemistries, or in the DNA being sequenced.
4. Assembly behavior can be better predicted and understood by studying some of these trends. Large-scale facility-wide problems can be revealed early in the process, as well as case-specific caveats.
5. We developed several robust methods for detecting cloning bias, DNA contamination and presence of long repeats in the early stages of WGS projects.
6. These methods are based on modeling and comparative analyses of **depth of coverage distributions, progressive (iterative) assembly "kinetics",** and **GC composition distributions**
7. Ideally, the distribution of the number of alignments of a read to any other read should follow a Poisson curve. By comparing real depth distributions to the fitted Poisson function and observing the associated *Asymptotic Standard Error (ASE)* values, we can routinely reveal cloning bias and contamination in genomic libraries.
8. Progressively assembling a steadily growing dataset and looking for deviations from a pre-determined model, a process we termed Progressive Assembly Plotting (PAP), can reveal high levels of repeats.
9. Differences in GC composition between different genomes, libraries and even plates allows to identify cases of contamination by identifying bimodal patterns in the GC content distribution among the sequences.

## METHODS

To look for manifestations of cloning bias in depth distributions we randomly selected datasets from a large pool of prokaryotic genomic libraries, all subcloned and sequenced using common shotgun methods. We knew from prior experience that a portion of these libraries contained strong cloning biases, contamination or both. Libraries that didn't display such features were known to be ideal for fast and smooth gap closure and final polishing. In total, 310 libraries were included, consisting of 3 kb pUC18 and 6-8kb pMCL200 plasmid libraries, and 28-40kb pFos libraries.

Using an in-house program (estDepthFromAce) that parses and extracts per-base depth information from .ace files (Phrap assembly format) , we generated a high-resolution depth vector for each individual library. The resulting distributions were then plotted against the fitted Poisson function using Gnuplot. The goodness of fit was reflected by the value of *Asymptotic Standard Error*.

As an alternative to .ace file-derived coverage, we used Blastn to produce all possible read vs. read alignments. The frequency distribution of the hits could then be also quantitatively compared to the fitted Poisson distribution.

To model the Reads vs. Contigs dynamics for a typical single chromosome genome we used the following approximated equation:

$f(x) = x*exp((-n/L)*x)$ ($n$ = mean read length, $L$ = genome size).

To generate reads vs. contigs Progressive Assembly Plots, *in-silico* shotgun datasets were simulated from real finished genomes. Then, a series of progressive Phrap assemblies was performed, where a set number of reads would be added before each iteration. The lengths of the *in-silico* reads and clones followed normal distributions, with mean and std. deviations approximating real life cases. Real reads were not used in order to avoid any possible effects of case-specific differences in read length or quality.

GC content was plotted from the average GC% values calculated for all trimmed and vector-screened production reads. This procedure is currently a part of the standard QC protocol for JGI's Sanger-type sequences.

For assessing the effects of contamination we selected a notorious case of host contamination in Ehrlichia canis. Several datasets were available, starting from the initial, heavily contaminated set, and ending with just the reads that make it into the finished genome submission. An artificial contamination case was also made by combining reads from two distinct finished genomes.

Repeat analysis was done on Frankia Sp. (JGI) and Shewanella oneidensis (TIGR) genomes with average to high repeat content relative to other microbes. Using the Vmatch software, we identified and masked all continuous non-unique sequence fragments over 500 bp in length. After this, shotgun simulations and iterative assemblies were performed on the masked and non-masked data, as described above, and the resulting PAP curves were compared to each other and to the model.

## RESULTS AND DISCUSSION

### *Bias Detection*

Observed deviations of depth curves from Poisson distribution agree with general expectations:

♣ Libraries with relatively large observed Asymptotic Standard Error (ASE) are known to have failed at some point of the QC process, or have seriously confounded finishing, mainly by virtue of uncaptured (uncloned) gaps or insufficient clone depth. The noticeable broadening and flattening of the curve is the result of under-representation of some regions and over-sampling of others (Fig.1). It is worth noting that a normal bell-shaped distribution is still apparent in most biased cases.

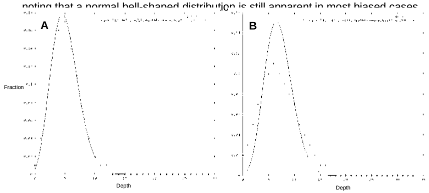


**Fig.1** Two distributions of read depths in assembled pUC18 libraries Dotted lines represent the datasets, solid lines are fitted Poisson curves. **A.** An unbiased library; ASE = +/- 0.03331; **B.** A library with confirmed excessive cloning bias; ASE = +/- 0.1924

♣ *GC-content correlation* is observed in plasmid libraries. Reads derived from large-insert fosmid libraries display a significantly better fit, but may be under-represented for purposes of this study.

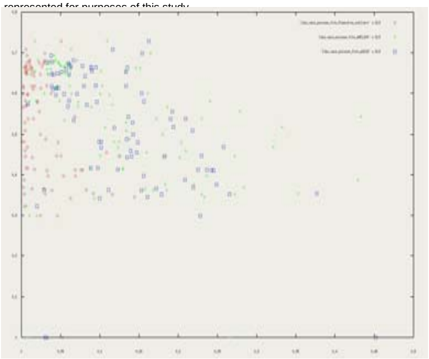


**Fig.2** Correlation of ASE scores with GC composition. In pUC18 and pMCL200 plasmid libraries, low GC datasets display higher ASE.

This depth distribution-based method could allow us to check libraries for bias routinely and reliably, with proper steps taken to either eliminate or compensate for the bias. The observed behavior should be noticeable even with only a fraction of the dataset available. This makes it possible to detect bias as soon as the reads begin coming off the sequencers.

### *Contamination Detection*

GC composition analysis provides a reliable way to identify possible extra-chromosomal contamination, vector contamination, mixed source DNA, and even large scale plate mix-ups in the lab. In a typical dataset of genomic shotgun reads, the GC content of the sequences should follow a normal distribution. Any considerable amount of contamination from a foreign source with the mean GC content distinct from the genome of interest may reveal itself as an additional peak or a "shoulder" in the GC distribution (fig.3).

♦ GC plotting of shotgun datasets that have been screened (by Blastn and Cross_match ) for known common contaminants and had them removed shows the effectiveness of the screening procedure used (fig.3).

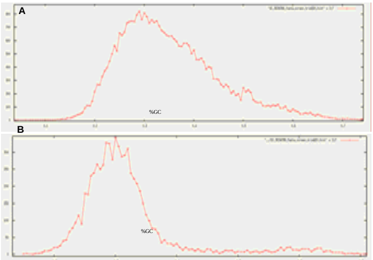


**Fig. 3** Two GC plots of *Ehrlichia chaffensis* shotgun reads. **A.** Contamination is present, as suggested by a long high GC outlier. **B.** Host contamination removed after database screening. Small amounts of high-GC reads are still present, meaning that not all of the contaminants were removed.

♦ Library-level contamination is distinguished from the genome-wide kind by plotting each library's GC composition separately. Fig. 4A shows a case of DNA-level contamination of a genome with a foreign source – all libraries have the secondary peak. In fig. 4B, wrong DNA samples were used during preparation of libraries AIGA and AIGB, while the rest was done with the correct DNA. This latter case specifically illustrates a situation when an entire library contains only the foreign reads.

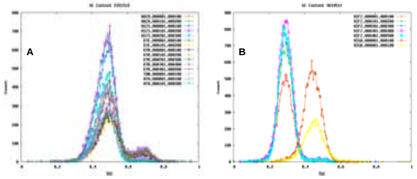


**Fig. 4** Library-specific GC plots. **A.** DNA-level contamination; secondary peak present in all libraries. **B.** Library mix-up; two of the libraries exhibit a completely distinct GC signature.

When both the target and the contaminant have similar GC content and can't be visualized with GC plotting, analyses of PAPs and depth distributions provides clues to potential contamination problems.

♦ In Fig 5, the initial PAP curve differs grossly from the model. This is caused by many additional contigs forming from the contaminant reads during assembly. After database screening (automatic QC), the deviation is still strong, indicating that contamination still remains in the dataset. After a long and tedious process of cleanup and Finishing, the reads that make it into the final submission are contamination-free and the PAP curve fits the model very well.

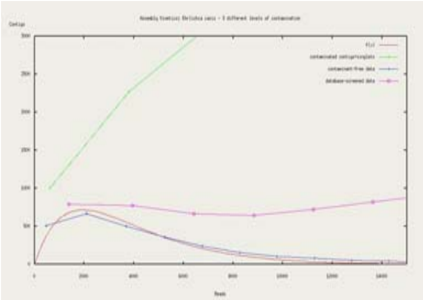
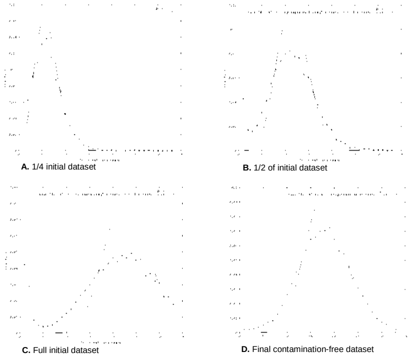


**Fig. 5** Progressive Assembly Plots of Ehrlichia canis shotgun reads - initial dataset (green), database-screened (pink), contaminant-free assembly (blue), compared to the model curve.

By using PAP analysis alone, however, it is difficult to distinguish contamination from library bias, since both result in the early leveling of the curve. The distinction is more clear in the depth distribution plot where the contaminant produces a secondary peak (reflecting different abundance levels), while cloning bias would likely cause an overall broadening of the curve.

♦ Fig 6 shows the same E.canis contamination case revealed in a series of depth plots (here in the form of frequency vs. nr.of Blast hits ), each taken at a different project stage. The over-represented low-depth outlier (i.e., contamination) is detectable even with just ¼ of the reads present.



**A.** 1/4 initial dataset

**B.** 1/2 of initial dataset

**C.** Full initial dataset

**D.** Final contamination-free dataset

**Fig. 6** Frequency plots of Ehrlichia canis shotgun reads.

### *Repeat Content Analysis*

Long repeats affect the behavior of assembly software* in a number of complex (and not so complex) ways, and thus result in definite changes in the shape of the PAP curve.

♠ Removal of long repeats (>500bp) and replacing them with random basecalls resulted in a 5% increase in the number of contigs around the peak of the PAP curve and a better fit to the model function* (Fig 7).

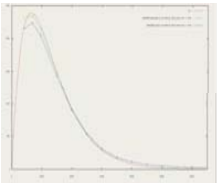


**Fig. 7.** Progressive Assembly Plots of simulated shotgun datasets generated from the published S.oneidens AEO14299 chromosome. Repeats cause the peak suppression of the of the non-masked dataset (blue), while the masked dataset (green) fits the model more closely.

* A likely explanation of this phenomenon is that in the early assembly stages, repeat copies act as depth boosters, causing small contigs to join into fewer larger ones, and do so at a faster rate than in a repeat-free dataset. As the assembly progresses with higher and higher avg. coverage, these repeats may or may not affect the number of contigs as much, depending on the repeat type and on the software design.

♠ Removal of long repeats (>500bp) from the same dataset caused a 25% increase of the mean value in the Blast-based depth distribution plot, and a clearly better Poisson fit*.

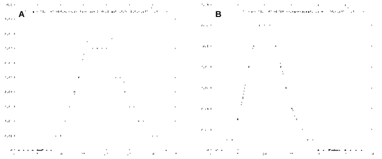


**Fig. 8** Frequency plots of simulated shotgun of S.oneidens AEO14299. **A.** Repeats present. **B.** Repeats masked

* This happens due to sequences from different repeat copies having strong pairwise hits with each other. In assemblies, this results in "over-collapsed repeats" where several repeat copies assemble together into one highly covered region. We've been able to confirm this behavior on many of our repeat-rich genomes.

Both of these distribution "fitness" approaches prove to be useful in revealing high repeat content. By running the analysis early using this method we are able to foresee potential repeat-linked assembly troubles that lie ahead. Better classification of the troublesome repeats is now needed in order to start designing better assembly strategies.

## REFERENCES:

Cox, R., Mazur, M., Goltsman, E. 2004. Statistical Mechanics of Genome Sequence Assembly. *Unpublished.*

Roach, J. C. 1998. Random subcloning, pairwise end sequencing, and the molecular evolution of the vertebrate trypsinogens. Doctoral dissertation, University of Washington.

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231 and Los Alamos National Laboratory under contract No. W-7405-ENG-36. LBNL-59860 Poster