Quality Control of JGI Microbial Sequencing Projects

Kerrie W. Barry, Alla L. Lapidus, Eugene V. Goltsman, Matt P. Nolan, Joel A. Martin, Mingkun Li, Paul M. Richardson, Alex C. Copeland

DOE JOINT GENOME INSTITUTE
US DEPARTMENT OF ENERGY
OFFICE OF SCIENCE

U.S. D.O.E. JOINT GENOME INSTITUTE

ABSTRACT

We provide an overview of the Joint Genome Institute's Microbial Sequencing projects with an emphasis on recent efforts to improve data quality by performing various quality control operations during sequencing and prior to beginning finishing. We apply these methods have to old projects in an attempt to correct obvious data problems and, therefore, provide a minimum standard of quality satisfied by all projects. We describe a rapid and largely automated protocol that removes low quality and likely contamination from our projects. Application of this protocol typically removes between 5 and 30 percent of the reads in a project without negatively affecting draft assemblies.

INTRODUCTION

The Joint Genome Institute (www.jgi.doe.gov) is a joint effort between staff from the national labs at Los Alamos, Livermore and Berkeley. The Production Genomics Facility (PGF) in Walnut Creek, California, a high throughput DNA sequencing facility, produces approximately 3.0 billion base pairs per month and is the largest sequencing facility of the JGI. The sequencing core facility employs around 200 people, with about 60 directly involved in production The QA group is part of the Production group and includes 5 people involved in sequence quality assurance, data analysis, and production troubleshooting.

The PGF sequences a variety of large and small genomes, which range in size from 1.0 Mb to 1.7 Gb. A list of all sequencing projects is available at http://www.jgi.doe.gov/sequencing/seqplans.html. The list of 2005 microbial projects comprises 50 projects including 6 strains of Shewanella, 7 strains of Chloroflexi, 4 strains of Rhodopseudomonas palustris, 4 organisms involved in microbial arsenic transformation, 2 species of Micromonas pusilla, 21 single microbes from 10 bacterial and 2 archael phyla, and 6 microbial communities including environmental samples from The Cedars, Iron Mountain, Obsidian Hot Springs, a PAH degrading Mycobacterial community, active methylotroph community, and viruses infecting globally distributed microalgae. A new program, the Community Sequencing Program (http://www.jgi.doe.gov/CSP/) includes 10 microbes for 2005 and over 100 proposals are now being reviewed from the 2006 RFP. Additional projects from the DOE GTL program (http://www.genomestolife.org), legacy and internal R&D projects bring the total number of microbial projects to 177.

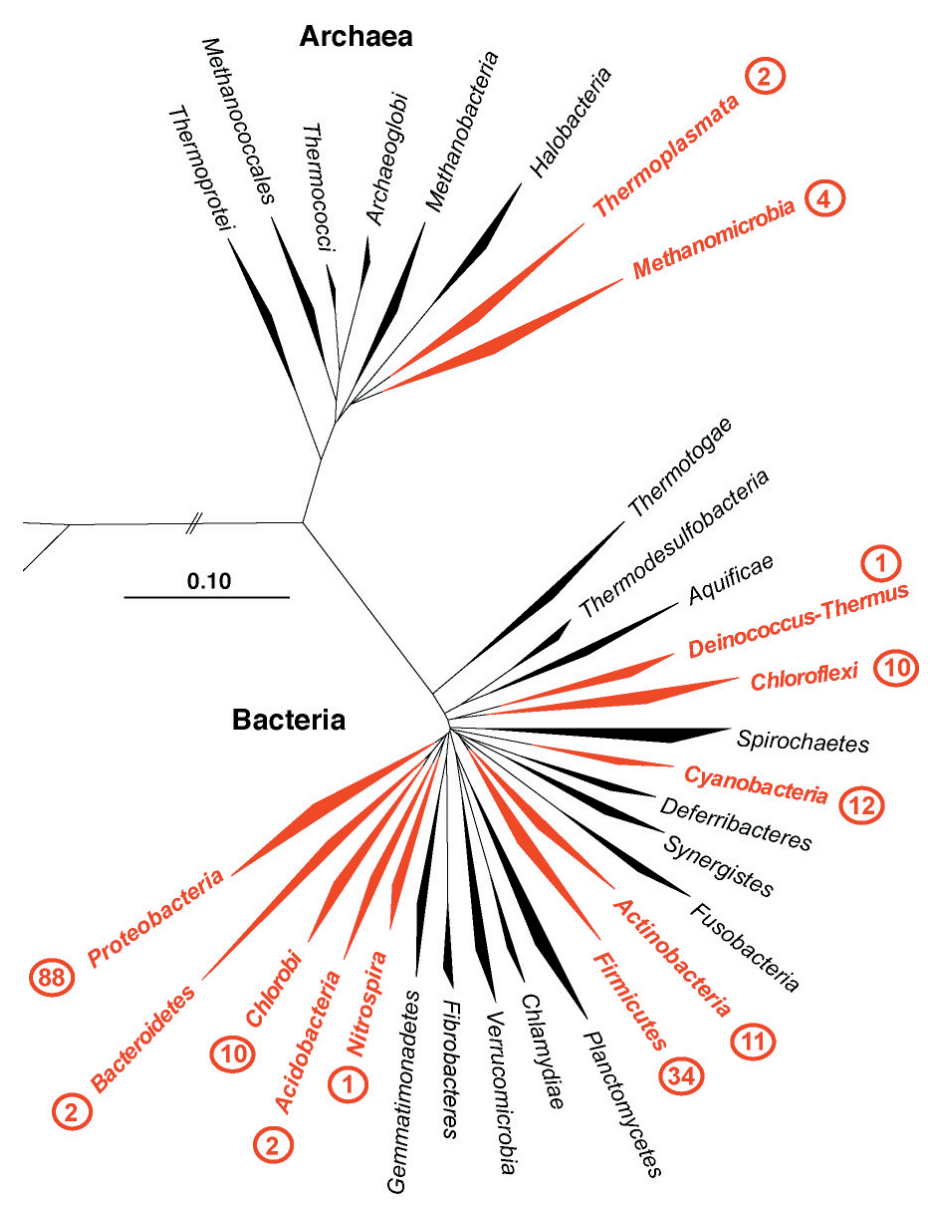


Figure 1: **Unrooted phylogram of all JGI microbes.**

The QA group works closely with finishers to insure that projects have met internal quality specifications before finishing work begins. We also work closely with production staff and collaborators to identify mix-ups and contamination and to identify bad libraries before investing significant resources into sequencing them. The goal of JGI sequencing efforts is to provide high quality sequence data to the public in a timely and cost-effective manner and, ultimately, annotated assemblies of microbial projects containing the absolute minimum of errors. The QA group has a key role in defining and maintaining project quality standards.

METHODS

See http://www.jgi.doe.gov/sequencing/protocols/prots_production.html for up to date versions of all JGI production protocols. Brief highlights of relevant aspects of the production sequencing process and subsequent quality control analysis are described below.

Library creation & Sequencing

Our current sequencing strategy is to shotgun sequence 3kb and 8kb libraries to 8x draft coverage and to sequence 40kb fosmids to 0.5x sequence coverage. For each microbial project, three libraries are constructed: 3kb and 8kb plasmid libraries are cloned into pUC18 and pMCL200, respectively, and a 40kb fosmid library is cloned in pCC1Fos. Clones are picked into 384-well plates using Q-Pix colony pickers, DNA is amplified using RCA (7,8,9,10), sequencing is performed basically according to manufacturer's recommended protocols but at high dilution, and samples are sequenced on either ABI3730 or MB4500 platforms. The plasmid libraries are each sequenced to 4x read coverage and the fosmid library to 2x clone coverage.

Assembly

Raw trace data is automatically transferred to a holding area on a UNIX file server at the completion of each sequencing run. Traces are basecalled with phred (2,3), summary statistics are gathered for reporting, and data is moved into the appropriate project directory where it is organized by library. Raw data is immediately archived on the HPSS system at NERSC (11). Fasta files from each library are screened for the appropriate vector using parallel cross_match (4,5), then groups of 100 plates for each library are assembled using parallel phrap (4,5). These subassemblies are used for quality control purposes and to simplify the final assembly process by providing up to date, prescreened, fasta and quality files.

Library QC

7680 reads each of the 3- and 8-kb libraries for each new microbial project are prepared and sequenced as previously described. The input and assembly output is analyzed by a collection of analysis scripts that produce a collection of reports. The objectives of library QC are to determine the insertless clone rate, estimate the insert size and distribution, assess contamination levels and attempt to confirm that the correct organism has been cloned.

The fraction of insertless clones is estimated by examining the output of cross_match and identifying reads (clones) containing 80% or more vector. The insert size distribution may also be used to support the estimate by measuring the assembled distance between mate pairs, but these early low coverage assemblies are often quite biased.

Assembly QC (QD)

The core of assembly QC has been fully automated and is carried out by a single script which runs BLAST jobs, parses and filters the output, converts filtered hits into list of reads to exclude from the final assembly, produces a clean fasta file and a summary report. The details are described below.

Automated contamination identification and removal proceeds in several steps. First, MegaBLAST (6) is used to align all reads to a database of known local contaminants, which includes items such as molecular weight markers and other sequences which we have found present in many PGF projects, at fairly low stringency: -p 80 -e 1e-30. Blast hits to the contaminants database will be examined and the reads will be flagged as probable contaminants. Next, reads are aligned to a database of all sequencing vectors used at the JGI using parameters -p 98 -e 1e-30; these reads will be flagged as probable vector. Finally all reads are aligned against NCBIs 'nt' database at high stringency in order to identify any near-exact matches which can be used for flagging reads which clearly do not belong in the project. Reads are aligned using megablast with parameters -p 98, -e 1e-30 and soft masking, -F "L m" . Hits are filtered for 98% identity and minimum length of 200 bases. A list of 'gi' numbers for all of the filtered hits is constructed and used as input for 'tax_filt' (7), which separates the hits into eukaryotic and prokaryotic bins based on their NCBI taxonomy ids. As most of our projects are prokaryotic, normally we would then flag any eukaryotic hits as probable contaminants. Project or source DNA cross-contamination is identified by assessing GC plots of all reads or contig subsamples and by the 'nt' blast results.

Low quality reads have been found to negatively impact phrap assemblies so we exclude from assembly all reads having less than 100 total phred Q20 bases. A perl script calculates the total number of non-screened bases having phred score greater than 20 per read and produces a list of low quality reads.

JGI finishers have empirically determined that phrap contigs containing two or fewer reads do not in general provide useful information for finishing, since they are primarily composed of short-insert clones. Reads in such contigs are also identified and flagged for exclusion in the draft assembly.

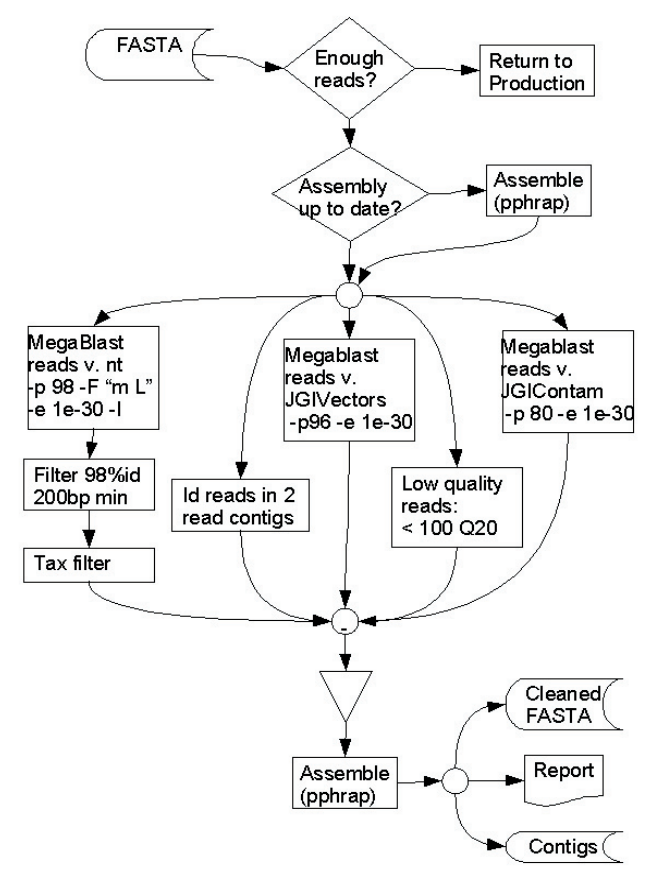
The combined list of all problematic reads identified above is removed from the project fasta and quality file and a new assembly is constructed using the filtered data set. This becomes the reference assembly for beginning the finishing process.

DISCUSSION

The first microbial genomes sequenced by the JGI were begun during the push to complete draft sequencing of Homo *sapiens*. These early attempts at microbial sequencing relied on a single 3kb library and produced draft sequence only. Some time A thorough analysis of the status of JGI's early microbial efforts demonstrated the need for more quality control, since many projects were heavily contaminated, contained abundant low quality reads, and were not finishable with the existing data. The manual and time-consuming tasks employed in cleaning up these early projects was converted into the software we used for creating high quality draft assemblies.

As previously discussed, we evaluate all new libraries for source DNA contamination, PGF process contamination, reasonable insert size distribution and we attempt to confirm the identity of the organism. Failing of a library at this step generally leads to a discussion between the project manager and the library construction group lead during which project priority, cost, DNA availability, and difficulty of remaking libraries are assessed and a decision is made whether or not to continue to sequence it. If the libraries pass, we use assembly and any other available information to refine genome size estimates. Library specific read length data and genome size estimates are combined to define the number of plates of each library type to sequence in order to achieve the project depth target. At this time, the fosmids are also sequenced and this is the information available for the QD process.

Once the project fasta file contains the requested amount of sequence (see Figure 2) we assemble it using pphrap. Because an in depth QC of each microbial library has already taken place, we are able to do an almost completely automated QC of the final assembly, which we call 'QD', with a turn around time of one day. The finishing group receives the resulting collection of read lists excluded in the draft assembly, the reports detailing what was removed, and the quality and assembly metrics.



FASTA → Enough reads? → Return to Production

Assembly up to date? → Assemble (pphrap)

MegaBlast reads v. nt -p 98 -F "m L" -e 1e-30 -I → Filter 98%id 200bp min → Tax filter

Id reads in 2 read contigs

Megablast reads v. JGIVectors -p96 -e 1e-30

Low quality reads: < 100 Q20

Megablast reads v. JGIContam -p 80 -e 1e-30

Assemble (pphrap) → Cleaned FASTA, Report, Contigs

Figure 2: Flow Chart of the QD process undergone by all JGI Microbes

A closer look at the draft QD reports reveals that between 5-30% of the reads for the JGI microbes are usually removed before being passed to finishing. In Figure 3, the percentage of reads removed for each category, and the percentage of reads in the final assembly are shown. Each multi-colored line represents one microbe. In general the majority of reads removed in a project are removed because of low quality, which is depicted by the sage-green colored bars. The two Ehrlichia microbes were very heavily contaminated with Canis familiaris, which we couldn't completely remove until after the WGS reads from the dog genome project were submitted to NCBI .

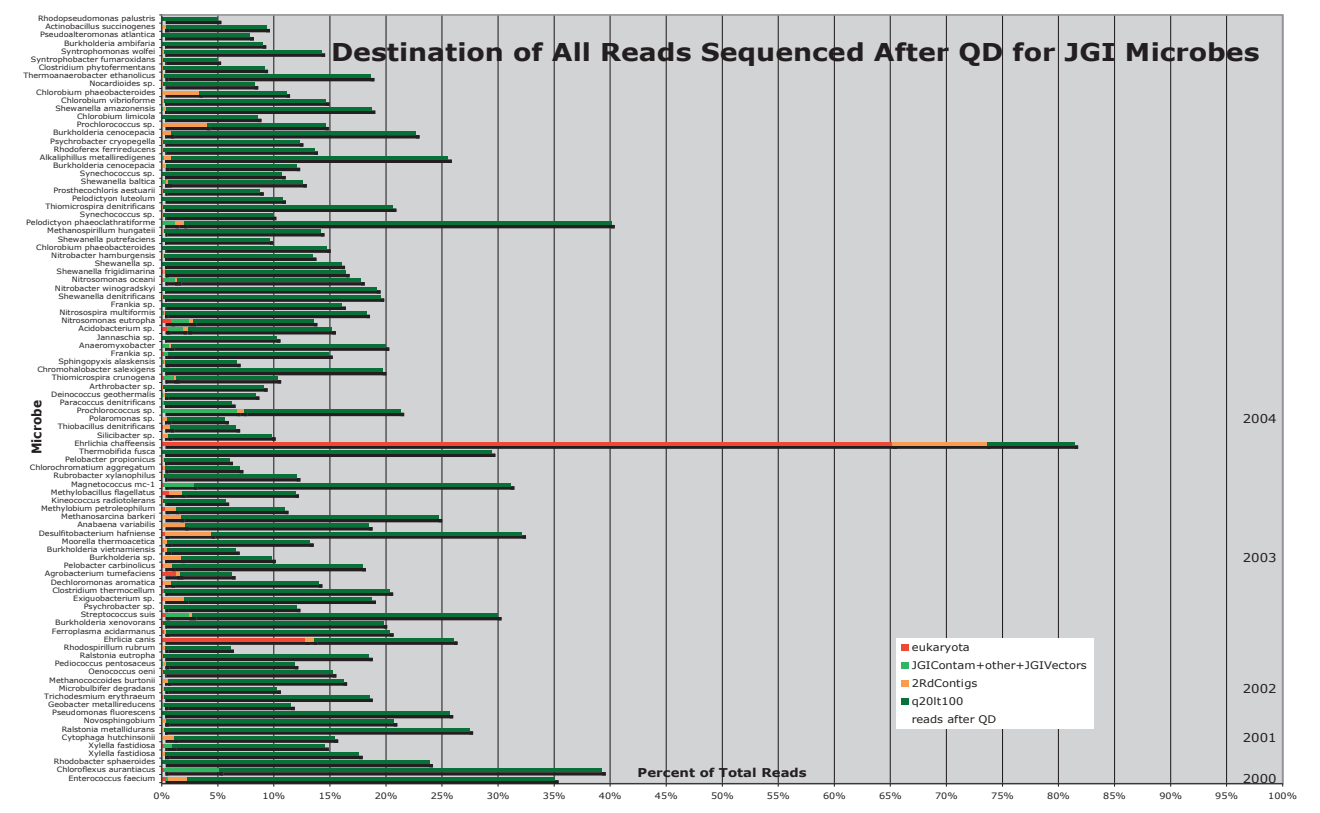
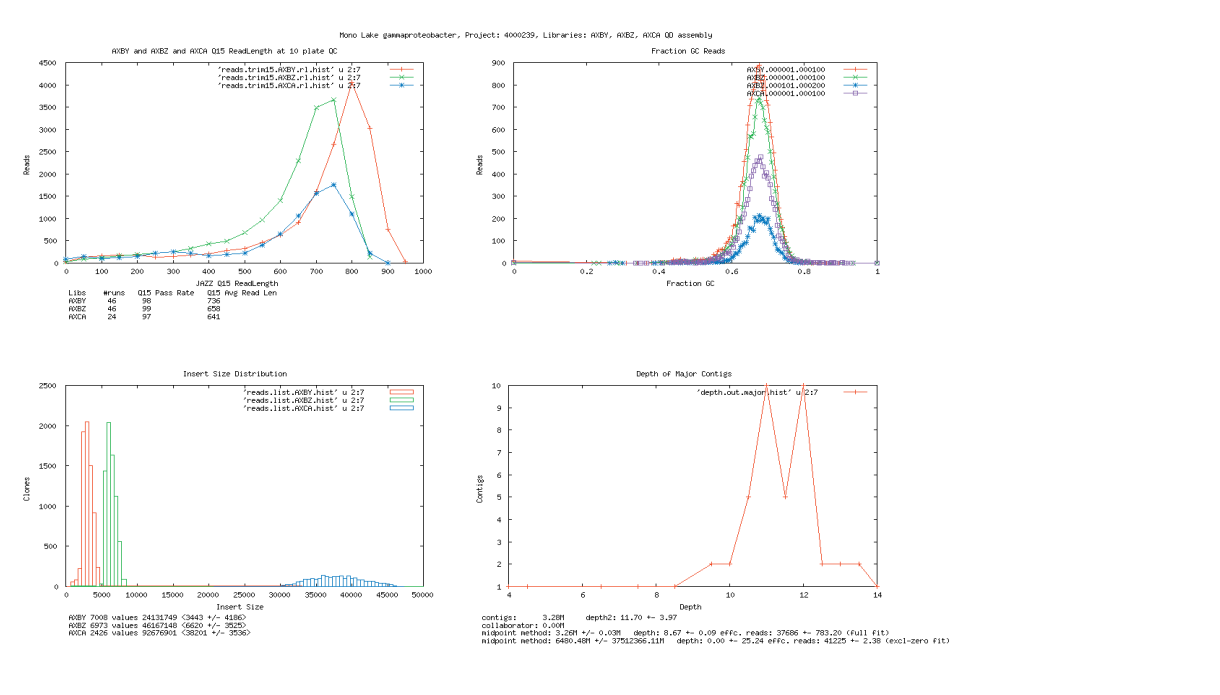


Figure 3: Chart of the reads removed from the assembly during the QD process. The number of reads removed from each of 4 categories: low quality reads are having less than 100 phred q20 or better bases, 2 reads contigs, reads that hit the JGI contaminants database and reads that have significant sequence similarity hits to eukaryota in nt.

Detailed technical reports generated by the QC and QD processes are used by the QA group. Because many JGI staff are interested in a short, graphical summaries, the QA group provides a web-based summary report with 4 of the plots resulting from QC and/or QD. The first Quadrant shows the trimmed (Jazz Q15 trim) read length distribution of the libraries is shown. We expect the peak for the plasmid libraries to be over 700 base pairs, and for the fosmid libraries over 500 base pairs after vector screening and Jazz (8) Q15 quality trimming. Below this plot is a summary of the number of plates run, the average read length and fail rate for each library. In Quadrant II the GC plot of the libraries is shown. The third quadrant shows the insert size distribution of each library in a project. Ideally the insert size distribution should be within +/- 10% of the mean. Quadrant IV shows the depth of the major contigs in the assembly. Generally the QD assembly is 8-10X with a single peak. In the web reports, all of the graphs are image maps, and clicking anywhere inside the plots will bring up web pages containing the supporting data tables.



Libs	#runs	Q15 Pass Rate	Q15 Avg Read Len
AXBY	46	98	736
AXBZ	46	99	658
AXCA	24	97	641

```
AXBY 7008 values 24131749 (3443 +/- 4186)
AXBZ 6973 values 46167148 (6620 +/- 3525)
AXCA 2426 values 92676901 (38201 +/- 3536)
```

```
contigs:       3.26M     depth2: 11.70 +- 3.97
collaborator: 0.00M
midpoint method: 3.26M +/- 0.03M   depth: 8.67 +- 0.09 effc. reads: 37686 +- 783.20 (full fit)
midpoint method: 6480.48M +/- 37512366.11M   depth: 0.00 +- 25.24 effc. reads: 41225 +- 2.38 (excl-zero fit)
```

Figure 4: **Graphical digest of the QC Report**

Four major project categories of general interest to the JGI staff are displayed graphically and available on-line for the QC and QD assemblies of all microbial projects.

GC Profiling

We make frequent use of GC content analysis to identify possible contamination and mixed source DNA. It is important to recognize and remove contamination from a project for better efficiency when finishing (Figure 10). The basic protocol is to quality trim the reads and then calculate GC content of the trimmed output. Fasta is trimmed using sequence-trimming software developed for the Jazz assembler to avoid biasing GC estimates with low quality data. The trimmed output represents the longest possible sequence in which the average quality value in all 11bp windows is Q15. Nucleotide content of the trimmed fasta is extracted with a perl script and data is binned either by library or by plate and library.

Plate level binning appears to provide the most selectivity but it obscures problems due to contaminated source DNA. Read level binning provides insight into possible source DNA problems but variance is higher. In Figure 5 (a) the read level GC plot is wide spread and has a large variance. Using only this GC plot, it is not easy to determine whether contamination exists in the project since there is no clear peak separation, though the peak does tail to the right somewhat. In the plate level GC plot (Figure 5 (b)), a second, smaller peak is more evident. Further investigation identifies one of the project libraries (AIAW) as having a higher, broader GC distribution than the other two (Figure 5 (c)). The AIAW library is contaminated with a low level of a higher GC content genome, such as Escherichia *coli* (50% GC). Due to sequence similarity between prokaryotes the QD process does not attempt to remove the small percentage of E. *coli* contamination often detected in the fosmid libraries.

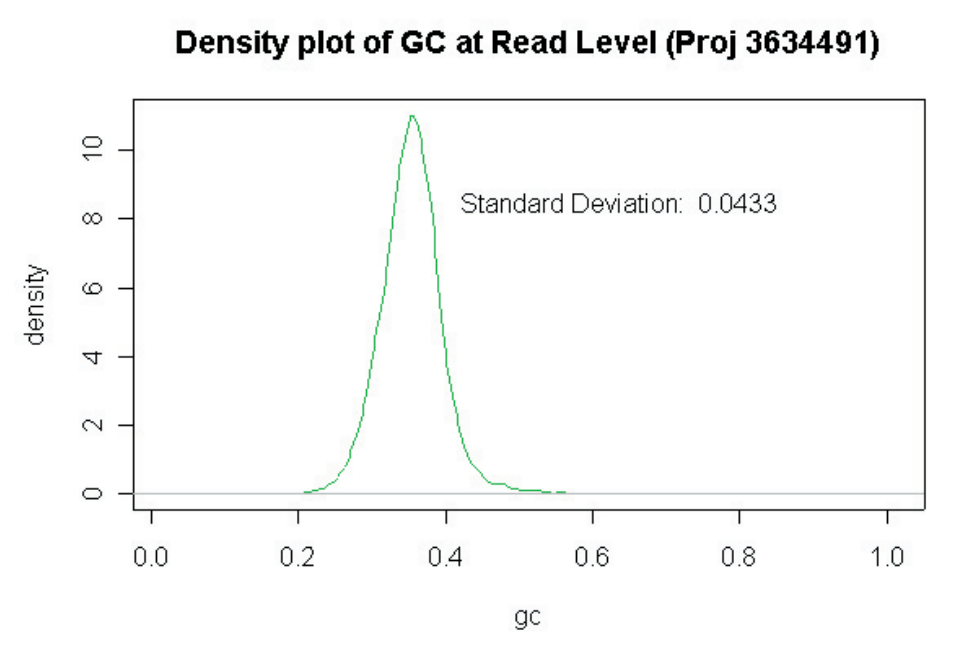


Figure 5 (a): **Density plot of the GC content by read level binning.**
The read level GC plot for Clostridium *phytofermentans* is broad and has a large variance.

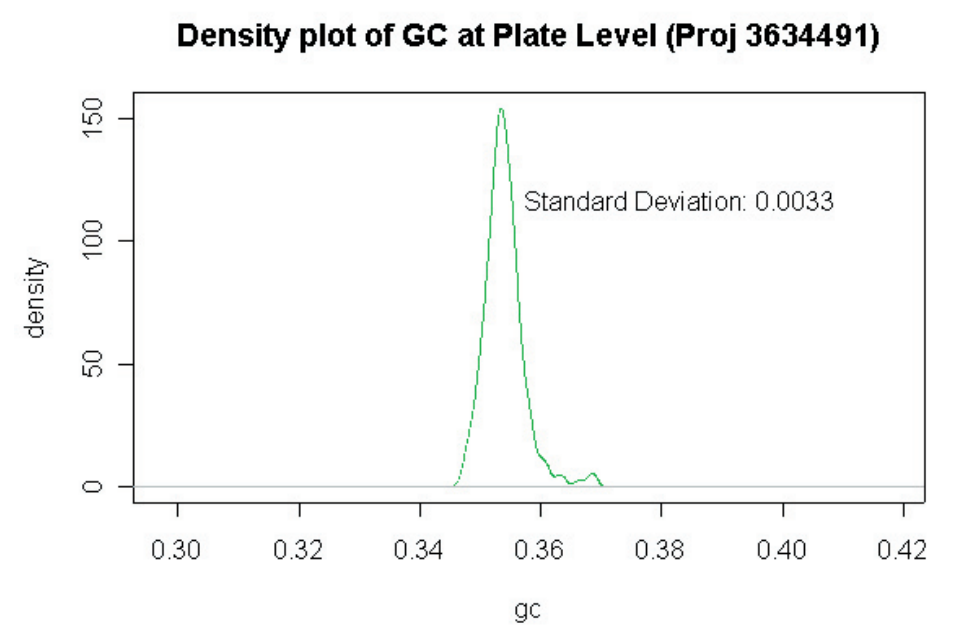
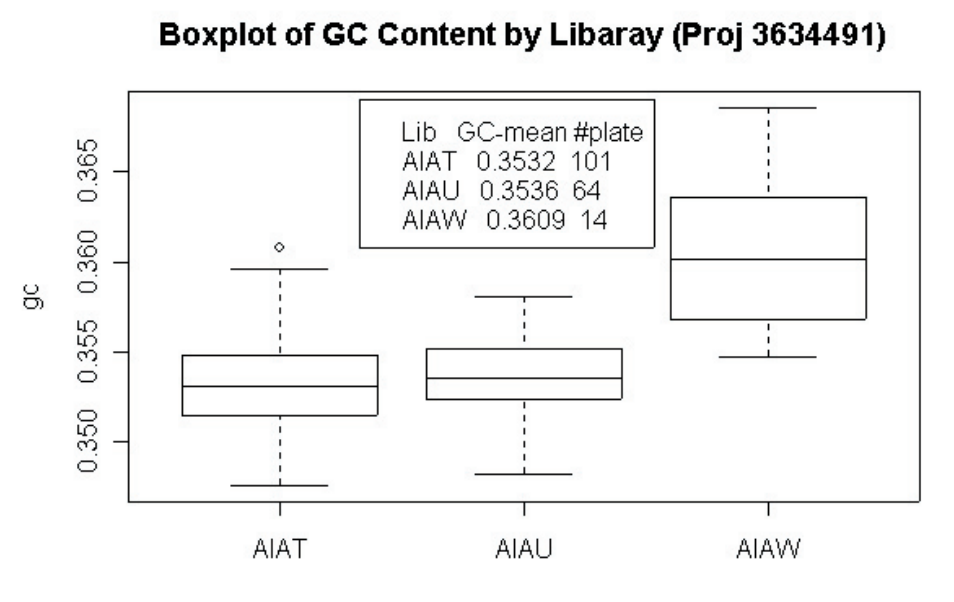


Figure 5 (b): **Density plot of the GC content by plate level binning.**
The plate level GC plot for Clostridium *phytofermentans* is narrow and has a smaller variance. In this plot, a higher GC content peak is detected at ~37% GC.



Lib	GC-mean	#plate
AIAT	0.3532	101
AIAU	0.3536	64
AIAW	0.3609	14

Figure 5 (c): **GC content boxplot of Clostridium *phytofermentans* libraries.**
The GC content of the AIAW (fosmid) library is higher and has a larger variance than that of the two plasmid libraries sequenced for this project.

In some cases GC profiling alone is sufficient to identify obvious problems. In a typical dataset of contamination-free genomic shotgun reads, the GC content of the reads should follow a normal distribution. Any considerable amount of contamination from a foreign source with a mean GC content distinct from the genome of interest may reveal itself as an additional peak in the GC distribution. The peaks can mask each other when the GC composition is similar in both sources, or when one is present in overwhelming quantities relative to the other. Figure 6 depicts a read-level GC profile for a contaminated project. In this case, the source DNA was a mixture of two Shewanella genomes. Because Shewanella are diverse, cosmopolitan microbes the GC profiles for these two strains were different enough to detect a mixed sample at the QC step by GC analysis alone.

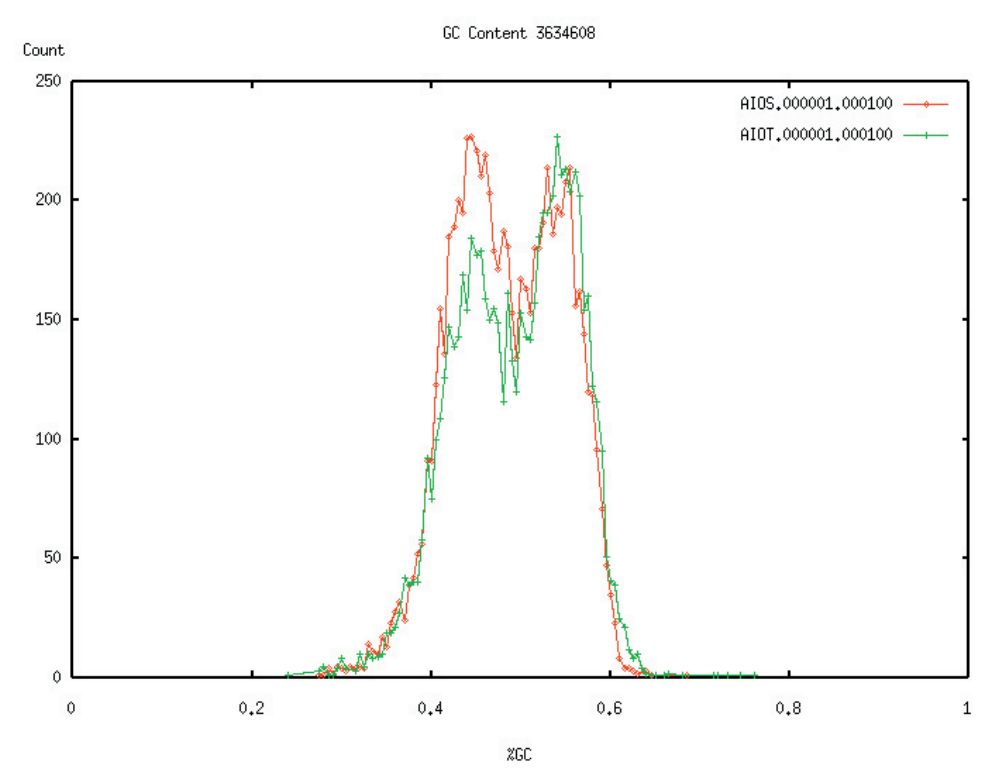


Figure 6: **GC profile of reads generated from a mixed DNA sample.**
Libraries created from a source DNA that was a mixture of two Shewanella show two distinct GC peaks.

In Figure 7 (a) DNA samples were swapped in the process of isolating DNA for two projects. The first library in each project was created from a separate DNA prep than subsequent libraries. When the second DNA samples were prepared for these two projects the samples got switched. Each project received one library from the correct DNA, and one library from the other project's DNA. We were able to reassign the switched libraries to the correct project as shown in Figure 7 (b).

We assign a unique color to each library in the plots below (Figures 7 (a,b)). The GC profiles indicate each library contains reads from different genomes. In the left plot of Figure 7 (a), library AIFZ is correctly assigned, while AIGA and AIGB belong to 3634478 (Syntrophomonas *wolfei* Gottingham). In the right plot of Figure 7 (a), AHYP is correctly assigned and AHYO belongs to 3634512 (Clostridium *beijerinckii)*.

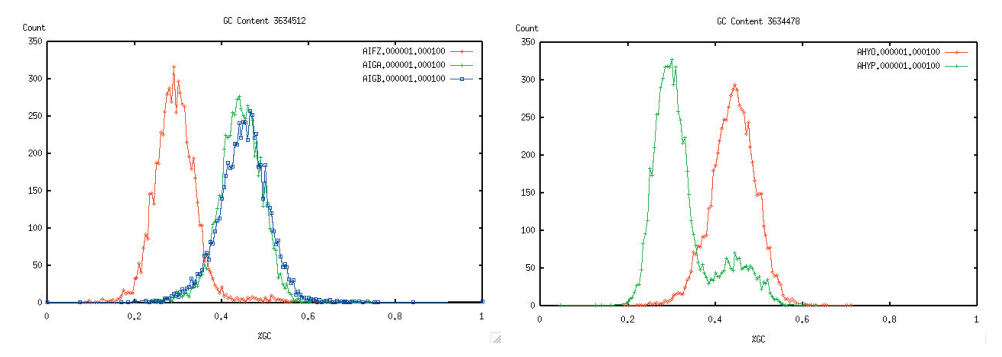


Figure 7 (a): **GC profiles of the reads for Clostridium *beijerinckii* and Syntrophomonas *wolfei* Gottingham reveal a switch during DNA isolation.**

Figure 7 (b) shows the switched libraries from 3634512 and 3634478 after reasssignment to the correct project and reassembly. The GC profile clearly shows matching GC profiles for the libraries within a project. After reassigning libraries the primary problem is solved. Note, however, another contaminant remains: library AHYP is correctly assigned to project 3634512, however, some of this library was contaminated with the source DNA for 3634478 as evidenced by the bimodal distribution of GC content.

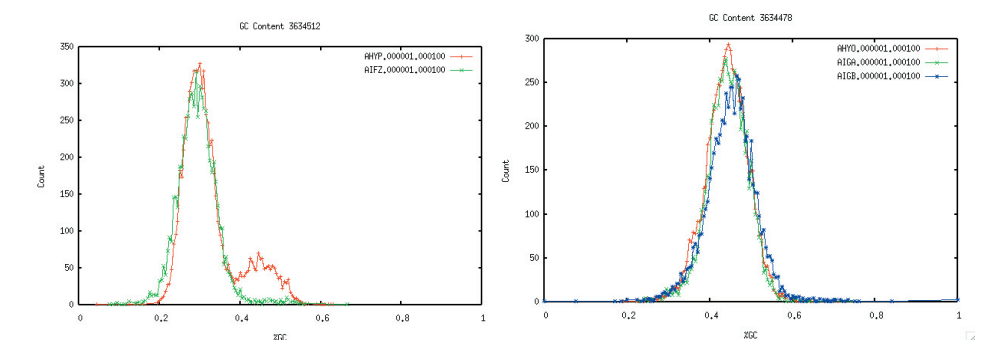


Figure 7 (b): **GC profiles of the reads for Clostridium *beijerinckii* and Syntrophomonas *wolfei* Gottingham after correctly reassigning the affected libraries.**

Assembly Dynamics

Assemblies follow a predictable pattern with respect to the number of real contigs and coverage (R. Cox *unpublished*). At 3-4X average read coverage, approximately 80% of the genome is represented by the major contigs (10+ reads, 2000+bp). As we add reads to the assembly, contig merging dominates contig creation, and the number of contigs decreases. In the progressive assembly of an ideal, uncontaminated dataset (Figure 8), the inflection point is somewhere between 3-4X read coverage. As depth increases, the number of contigs decreases as they grow and merge, incorporating smaller contigs into larger ones. Repeat content and contamination affect the location of the maxima and the convergence rate.

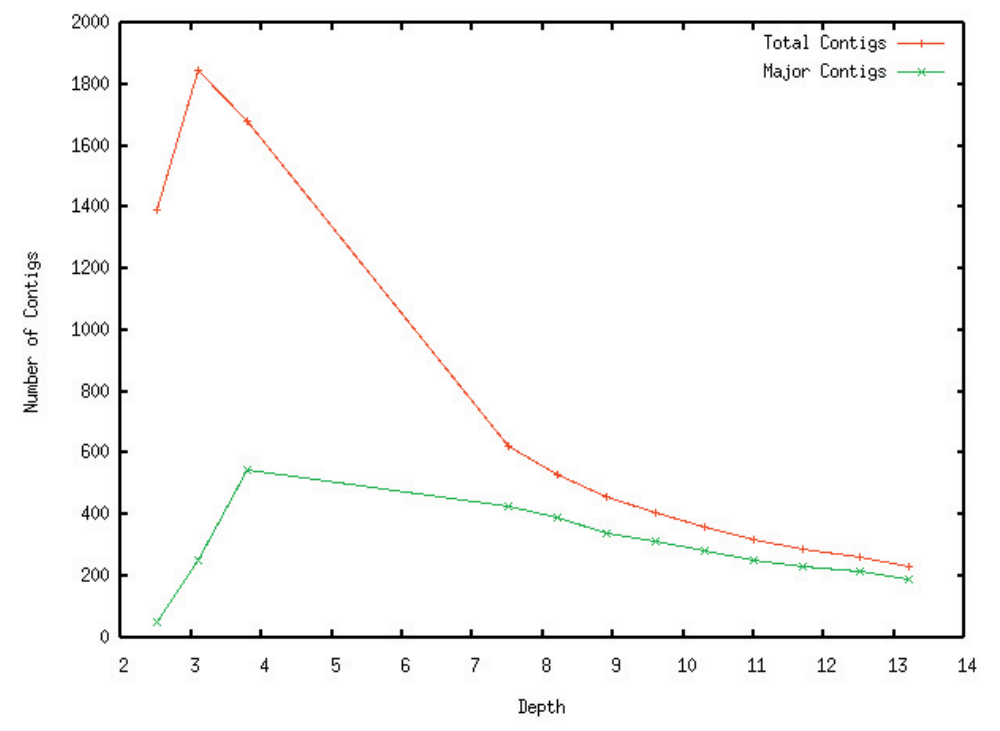


Figure 8: **Assembly Dynamics of for an uncontaminated whole genome shotgun assembly.**
In the assembly dynamics of uncontaminated, whole genome shotgun assembly the number of contigs increases as read depth increases. Between 3-4X read coverage, the assembly progresses toward fewer contigs as greater sequence depth is added.

Case Study: Ehrlichia chaffeensis

Ehrlichia *chaffeensis* was a particularly difficult project to assemble due to the high level of contamination from the host's genome. In this case study we identify the contamination using read level GC profiling, and show the detrimental affects of contamination on finishing by plotting assembly dynamics.

Source DNA for a microbial project may be contaminated by a host genome. In Figure 9, a broad tail toward higher GC content is present prior to QD.

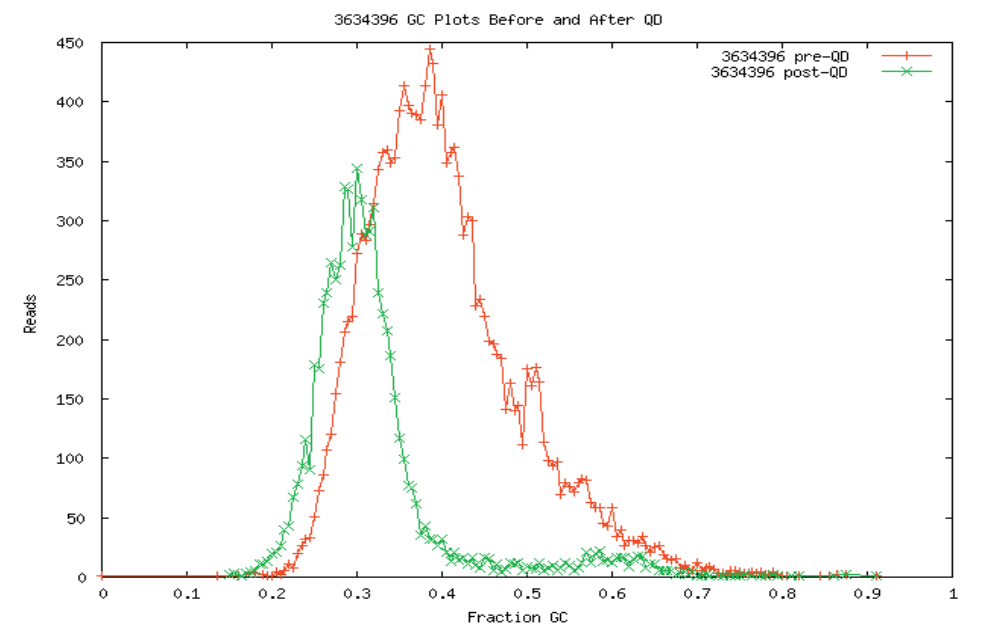


Figure 9 (a): **Host specific contamination of Ehrlichia *chaffeensis* revealed through GC distribution.**
Contamination in the GC profile of Ehrlichia *chaffeensis* is apparent prior to running QD. The GC plot is wide and has a long, broad tail toward higher GC content. **Host specific contamination of Ehrlichia *chaffeensis* is greatly reduced during QD.**

The QD process identified and removed the contaminant dog genome sequence. The low levels of contamination that remain are likely due to the high stringency of BLAST searching and hit filtering used for identifying eukaryotic contamination, as well as incomplete coverage of C. familiaris in 'nt' at the time.

Abnormal patterns in the assembly dynamics seen when running progressive assemblies on a project may indicate the presence of a contaminant. If the number of contigs continues to grow as depth increases, as in the case of Ehrlichia *chaffeensis* prior to QD (Figure 10 (a)), then the major contigs are not merging at the expected rate (see Figure 8).

The re-created progressive assembly series, using the filtered dataset, (Figure 10 (b)) has the expected shape. The contig numbers begin to decrease at 4.0X average contig read depth. This assembly behavior is especially apparent when the contaminant is from a larger genome. In this case, contaminant reads will keep forming new contigs and will skew the measured depth toward lower values, since these contigs will be small and isolated. When the contaminant is short, but over-sampled, this behavior is less apparent.

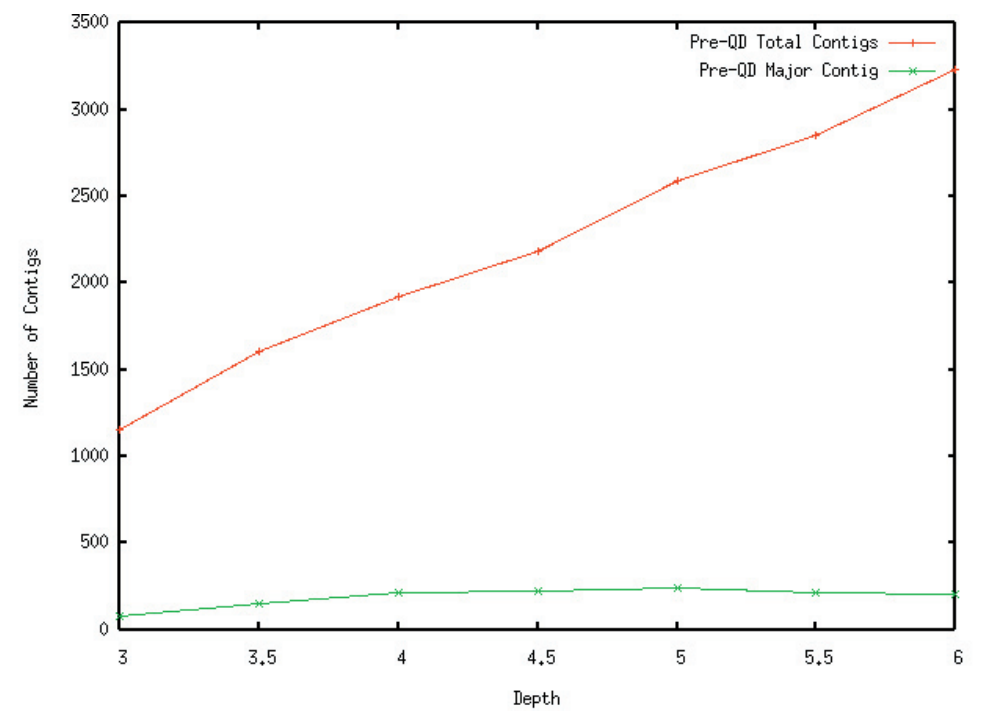


Figure 10 (a): **Assembly Dynamics of Ehrlichia *chaffeensis* prior to QD.**
As more reads are added to the pre-QD dataset of Ehrlichia *chaffeensis*, the total number of contigs continues to grow indefinitely.

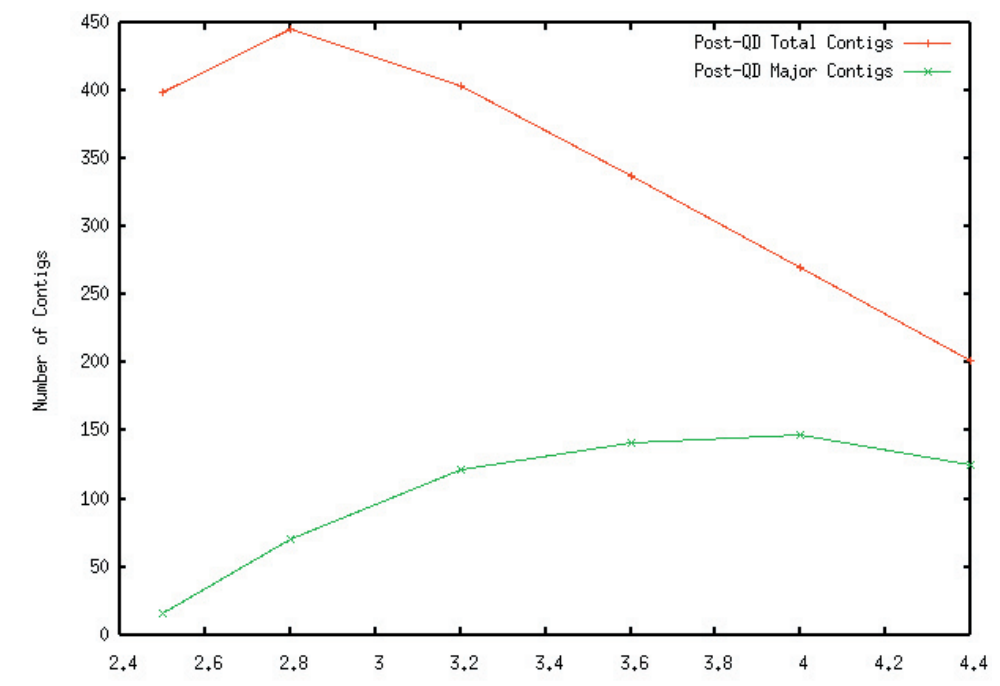


Figure 10 (b): **Assembly Dynamics of Ehrlichia *chaffeensis* after QD.**
As more reads are added to the post-QD dataset of Ehrlichia *chaffeensis*, the total number of contigs begins to decrease as smaller contigs are incorporated into larger ones.

FUTURE DIRECTIONS

Phrap makes excellent use of available data, and examination of phrap alignments shows them to include very close to the total number of bases sequenced. However, phrap does not handle low quality data well. In practice we have found that reads having less than one hundred total phred Q20 bases typically harm, more than help, microbial assemblies. Similarly, retained vector and other contamination frequently lead to assembly problems, necessitating the identification and removal of all such data prior to assembly. We have found it useful to tolerate the small risk of excluding some useful data in order to make the best possible assembly available for finishing work. These observations are empirical and in the future we would like to fully characterize read utility in assembly in order to optimal thresholds for excluding various types of problem reads.

We actively investigate other assemblers including PGA, Arachne, Jazz, AMOS and the Celera Assembler to determine if we can produce more accurate assemblies which will be simpler and less expensive to finish given the same number, or possibly fewer reads.

Our ability to identify contamination in projects continually improves. Ultimately, however, we'd like to prevent such contamination from entering the projects wherever possible. To this end, we are developing tools to aid in these investigations, including a system which fetches collections of reads based on their processing batch or location on a given instrument. The JGI Production Informatics group is rewriting our LIMS system and we anticipate that it's greatly enhanced reporting capabilities will also make these investigations simpler and more effective.

As our tools and methods continue to improve it becomes possible to automatically identify all contamination in all JGI projects and guarantee no contamination is submitted to public databases. This will also make it possible to retroactively clean up all previously submitted projects. While we are currently focused on developing and maintaining quality standards for current JGI projects, our long-term goal will be to identify and remove all contamination from public databases.

REFERENCES

1. http://www.jgi.doe.gov/sequencing/protocols/prots_production.html
2. Ewing et al.,Genome Res. 1998 Mar;8(3):175-85
3. Ewing et al.,Genome Res. 1998 Mar;8(3):186-94
4. http://www.phrap.org/phredphrapconsed.html
5. http://www.spsoft.com/products.html
6. Zhang Z., Schwartz S., Wagner L., Miller W., J.Comp.Bio. 2000 Feb, 7(1-2):203-214
7. Walker, DR. and Koonin, EV (1997) SEALS: A System for Easy Analysis of Lots of Sequences. ISMB 5:333-339.
8. Aparicio, S. et al. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science **297**, 1301-1310 (2002).
9. Dean, F. et al., Genome Research 11 1095-1099 (2001)
10. Lizardi, P. et al., Nat. Genet. 19, 225-232 (1998)
11. Estaban, J.A. et al., J. Biol. Chem. 268, 2719-2726 (1993)
12. Reagin, M.J. et al. TempliPhi: A Sequencing Template Preparation Procedure That Eliminates Overnight Cultures and DNA Purification. J Biomol. Tech. 14, 143-148 (2003.)
13. AMOS. ftp.tigr.org/software/AMOS
14. http://www.broad.mit.edu/wga
15. D.B. Jaffe, J. Butler, S. Gnerre, E. Mauceli, K. Lindblad-Toh, J.P. Mesirov, M.C. Zody, and E.S. Lander. Whole-genome sequence assembly for mammalian genomes: Arachne 2. Genome Res. 13: 91-96 (2003).
16. http://sourceforge.net/projects/wgs-assembler (CA released to public domain)
17. Myers et al.,Science. 2000 Mar 24;287(5461):2196-204.
18. **Huson et al. 17 , Bioinformatics** (Supplement 1): 132 (2001).
19. PGA (www.paracel.com - company no longer in business)
20. http://www.nersc.gov/nusers/resources/HPSS

ACKNOWLEDGEMENTS

We thank Phil Hugenholtz for the unrooted phylogram of JGI microbes.

AUSPICES

This work was performed under the auspices of the U.S. Department of Energy, Office of Biological and Environmental Research, by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, the Lawrence Berkeley National Laboratory under contract No. DE-AC03-76SF00098, and the Los Alamos National Laboratory under contract No. W-7405-ENG-36.