Hindawi Publishing Corporation
Sequencing
Volume 2010, Article ID 782465, 12 pages
doi:10.1155/2010/782465

Research Article

Identification and Quantification of Genomic Repeats and Sample Contamination in Assemblies of 454 Pyrosequencing Reads

Alexander J. Nederbragt, Trine Ballestad Rounge, Kyrre L. Kausrud, and Kjetill S. Jakobsen

Department of Biology, Centre for Ecological and Evolutionary Synthesis (CEES), University of Oslo, P.O. Box 1066 Blindern, 0316 Oslo, Norway

Correspondence should be addressed to Kjetill S. Jakobsen, kjetill.jakobsen@bio.uio.no

Received 26 May 2009; Revised 28 September 2009; Accepted 5 November 2009

Academic Editor: Nick Loman

Copyright © 2010 Alexander J. Nederbragt et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Contigs assembled from 454 reads from bacterial genomes demonstrate a range of read depths, with a number of contigs having a depth that is far higher than can be expected. For reference genome sequence datasets, there exists a high correlation between the contig specific read depth and the number of copies present in the genome. We developed a sequence of applied statistical analyses, which suggest that the number of copies present can be reliably estimated based on the read depth distribution in *de novo* genome assemblies. Read depths of contigs of *de novo* cyanobacterial genome assemblies were determined, and several high read depth contigs were identified. These contigs were shown to mainly contain genes that are known to be present in multiple copies in bacterial genomes. For these assemblies, a correlation between read depth and copy number was experimentally demonstrated using real-time PCR. Copy number estimates, obtained using the statistical analysis developed in this work, are presented. Per-contig read depth analysis of assemblies based on 454 reads therefore enables *de novo* detection of genomic repeats and estimation of the copy number of these repeats. Additionally, our analysis efficiently identified contigs stemming from sample contamination, allowing for their removal from the assembly.

1. Introduction

During assembly of shotgun datasets using the 454 (Newbler) assembly program, reads stemming from regions in the genome that are repeated, that is, present in multiple copies with a high degree of similarity, "collapse" into a single contig [1, 2]. Collapsed contigs start and end with the parts of reads that extend into the collapsed alignment. Specifically, reads that align partly within and partly outside a repeated region are divided between two contigs (in contrast to some other assembly programs, e.g., phrap [3], that do not divide reads between contigs). Thus, each collapsed contig only represents a repeated part of the genome including all (parts of) reads derived from the repeated regions. Consequently, these contigs have a higher read depth, where read depth is defined as the number of bases from all the reads used to assemble the contig, divided by the contig consensus length.

Studying contigs of a shotgun assembly of a cyanobacterial genome (*Planktothrix rubescens NIVA CYA 98*), we discovered that certain genes were likely present in several copies in the genome [4]. The number of copies of these genes seemed to be correlated to the number of reads used to assemble the contigs. In this study, we investigate whether this phenomenon is a general property of 454 shotgun assemblies. We hypothesize that, due to the even distribution of 454 reads over the genome [1, 5–8], the per-contig read depth should be linearly proportional to the number of genomic copies present in the genome.

To this end, we have analyzed whole genome shotgun assemblies of sequence read datasets obtained with 454 technology, focusing on bacterial genome and BAC datasets for which no pairwise information was present. An analysis of consensus contigs based on read depth, that is, the number of sequence reads that represent (cover) a base in the consensus contig, is presented. We also present a statistical approach to *de novo* estimate the number of copies of these regions present in the genome. By taking the per-contig read depth into account, *de novo* assemblies based on 454 reads

can yield additional information on the genome of study regarding repeated sequence regions.

Additionally, we find that if we assemble genomes based on DNA samples that include some level of contamination, read depth analysis can be used to effectively identify, quantify, and remove the contaminant.

2. Methods

2.1. Sequencing Datasets. Escherichia coli str. K-12 substr. MG1655 GS FLX reads were taken from a sample run accompanying the GS FLX Data Analysis Software Manual (version 1.1.03, December 2007). The GS FLX sample run consisting of only shotgun data was used. There were 547,055 GS FLX reads, totaling ~127.1 million bases (average read length 232 bases) in the dataset. The annotated reference sequence was taken from GenBank, accession number NC_000913 [9].

The *Porphyromonas gingivalis W83* GS FLX reads were obtained from the NCBI Short Read Archive [10] under accession number SRA001027. The dataset consisted of 510,032 reads, totaling ~116.4 million bases (average read length 228 bases). Genbank accession number AE015924 [11] was used as annotated reference sequence.

All other datasets were obtained using the GS FLX (Roche/454) at the Ultra-high Throughput Sequencing Platform of the Centre for Ecological and Evolutionary Synthesis (CEES), Department of Biology, University of Oslo, except for part of the *P. rubescens* NIVA CYA98 dataset, which was sequenced at Roche, Penzburg. The standard protocol for GS FLX shotgun sequencing was followed; no paired end data was used for any of the 454 read datasets.

The sequencing of *P. rubescens* NIVA CYA98 is described elsewhere [4]. The dataset consisted of 570,912 reads, totaling 149.6 million bases (average length 262 bases). The sequencing of *A. flos-aquae strain* 10E6 is described in Stüken et al. [12]. 683,403 reads, totaling 153.4 M bases (average length 224 bases) were available for this strain.

The bacterial genomes studied had GC percentages as follows: *E. coli*: 50.8%; *P. gingivalis*: 48.4%; *P. rubescens*: 39.5%; *A. flos-aquae*: 38.2%.

The read datasets for BACs 184H23 and 114L13 were kindly provided by Unni Grimholt, CEES, Department of Biology, University of Oslo. The datasets consisted of 74,826 reads (12.1 M bases) and 55,269 reads (8.8 M bases), respectively.

2.2. Assembly, Mapping and Output Parsing. All datasets were assembled using Newbler (version 1.1.03) [13] with standard parameter settings, and the optional ace-file for small genomes as additional output file. The ace file is generated by Newbler in a format that follows the specifications of the phrap assembly program [3] for genome assembly, and contains information on all aligned reads used to make the contigs. Ace files were parsed, using a custom PERL script, to generate metrics on the assembly, that is, contig statistics, number of reads used, and so forth, and on the contigs, that is, read depth, percentage GC, and so forth. The script was further used to split the assembly into contigs based on length or read depth.

Linear regression analyses were performed and graphs were plotted using Origin 8 (OriginLab, Northampton, USA), and the statistical package *R* version 2.8.1 [14].

2.3. Statistics of Read Depth Distribution and Copy Number Estimation. We developed a method for assessing the probabilistic relationship between read depth and the number of repeats. The read depth of independent, nonrepeated sequences approaches a Poisson process, but the existence of some repeated sequences within each contig and the fact that some nucleotide sequences are read more frequently than others cause the mean: variance ratio for single-nucleotide read depths to diverge from the 1 : 1 assumption of the Poisson distribution. Hence, we use robust, non-parametric methods to avoid biases. A robust measure of read depth (λ) for contig i is assumed to be the median ($\lambda_{m,i}$), with the mean to be found between the quantiles q_l and q_u of the nucleotide read depths ($\lambda_{ql,i}, \lambda_{qu,i}$) and most likely to be within one standard deviation of the median.

From the resulting empirical distribution of $\lambda_{m,i}$'s, we pick out the tallest peak of the kernel density as representing the mean read depth of nonrepeated sequences, under the assumption that the bacterial genome is relatively sparse in repeated sequences. Treating this peak as a nonparametric density distribution, we assume that the best estimate θ of the mean read depth of nonrepeated sequences is found within the range of values covered by a proportion δ of its relative maximum probability. Furthermore, we assume that most contigs are repeated an integer number of times. This is implemented numerically, estimating θ as the value within the range defined by δ that minimizes the sum of square distances between the estimated number of repeats for each contig, $\lambda_i\theta^{-1}$, and their closest integers. Median, upper, and lower confidence intervals (CIs) for the average number of times contig i repeated in the genome is thus found as $\lambda_{m,i}\theta^{-1}$, $\lambda_{qu,i}{\theta_L}^{-1}$ and $\lambda_{ql,i}{\theta_U}^{-1}$, respectively, where θ_L and θ_U are the lower and upper limits for θ as defined by δ and the empirical kernel density. Using $\delta = 0.05$, $q_l = 0.025$, and $q_u = 0.975$ was seen to result in almost no errors in deciding between single-copy and repeated sequences in the known *E.coli* and *P. gingivalis* genomes (i.e., for a single-copy contig, the upper limit of the empirical CI not being above 1.5x, and similarly, for a repeated contig, the lower limit of the empirical CI not being below 1.5x). These settings seem to reflect a fairly good trade-off between precision and confidence. Due to varying degrees of degree of overlap between confidence intervals, while a fairly good estimate can be made, there will always be a trade-off between false positives and false negatives when distinguishing single from multiple copy numbers.

The analysis method was implemented as script for the statistical package *R* [14]. The 454 AlignmentInfo.tsv generated by the Newbler assembly program was used as input for this script. This file contains, among other information, for each position in each contig the read depth of that position. The script, as well as relevant files for the *E. coli*

and *P. gingivalis* assemblies, is available from the following website: http://www.sequencing.uio.no/services/scripts/.

2.4. Real-Time PCR. Contigs with a range of read depths were selected randomly from the assemblies. PCR primers with identical Tm, generating ~200 bp PCR products were designed using Vector NTI (Invitrogen, Carlsbad, USA) and Primer3 [15]. BLAST searches against the assemblies were used to check that the primers were unique to the contig template. Average read depth at the sites of primer hybridization was calculated from the assembly alignment info files. Supplementary Table I (in Supplementary Material available online at doi:10.1155/2010/782465) gives an overview of the primers designed.

As template, the same DNA isolate, as was used for the 454 sequencing, was taken for both *P. rubescens* and *A. flos-aqua*. PCRs with four different 10-fold dilutions of the template were made to test PCR-efficiency. The PCR setup included HotGoldStar enzyme and qPCR Core kit for SYBR Green I (Eurogentec, Liege, Belgium) using a Lightcycler 480 (Roche, Basel, Switzerland). All reactions were performed in duplicate.

2.5. Taxonomic Profiling of Contigs Using MEGAN. Contigs, minimum length 500 bp for the *P. rubescens* and *A. flos-aquae,* and 100 bp for BAC assemblies, were compared to the nonredundant NCBI protein database using BLASTX [16], at the University of Oslo Bioportal (www.bioportal.uio.no). Settings were: Matrix: BLOSUM62, Gap Penalties: Existence: 11, Extension: 1, Maximum E -value 10; the alignments of the 25 best hits were recovered as a text file. This output file was imported into the program MEGAN, version 2beta9 [17]. The bit score cutoff chosen for the *P. rubescens* and *A. flos-aquae* contigs was 100, for the BAC contigs 30 (bit scores represent alignment quality). Trees were collapsed to the Family taxonomic level, with "Reads summarized" (the sum of all the reads assigned to that node and all nodes below it) shown for each node.

3. Results and Discussion

3.1. Per Contig Read Depth of 454 Read Assemblies and Genomic Copies. We have chosen two bacterial genomes for which both 454 reads and high quality annotated reference genomes were available, *Escherichia coli K-12* [9] and *Porphyromonas gingivalis W83* [11], as "model systems" for our study. The read datasets, obtained from the NCBI Short Read Archive [10], were used to explore the read distribution among contigs generated by assembling the reads.

Both GS FLX datasets were assembled, using the 454 Newbler assembler with default settings. The assembly for the *E. coli* dataset resulted in 100 contigs of at least 500 bp. For each contig, the per-contig read depth was determined. This per-contig read depth indicates, on average for all positions in a contig, how many times that base was covered by a read; this is sometimes referred to as "coverage". However, coverage is also used instead of oversampling (or redundancy), which is defined as the number of bases sequenced divided by genome size. The distribution of the per-contig read depths for the contigs of the *E. coli* assembly is shown in Figure 1(a). The read depth distribution peaks between 26 and 29x (Figure 1(a)). As the genome was sequenced to an oversampling level (or redundancy) of 27.4x (~127.1 million bases sequenced and genome size ~4.6 Mbp), contigs showing an approximate Poisson distribution around this number were expected. However, there was also a number of contigs with a higher read depth, up to 267x.

A similar per-contig read depth pattern was observed for the assembly of the *P. gingivalis* GS FLX reads, which were assembled into 124 contigs of 500 bp or more. Figure 1(b) shows that the per-contig read depth distribution for this assembly peaks around 40 to 60x, corresponding to 49.7x oversampling for this genome (~116.4 million bases, genome size ~2.3 Mbp). Again, this assembly also showed a number of contigs with higher read depths, up to 531x (Figure 1(b)).

Using the annotated, high quality reference genomes available for *E.coli K12* [9, 18] and *P. gingivalis* [11], BLASTN [16] was used to compare the contig sequences of both datasets with the corresponding reference genome (maximum E value 10^{-16}, 97% or more identity at the nucleotide level, thereby allowing for a few mismatches). Based on the BLAST results, for each contig the number of hits in the reference genome was calculated by summing the lengths of the High-scoring Segment Pairs (the portion of the contig sequence that aligned to the reference genome) for each hit, and dividing that by the contig length. The results are shown in Figures 1(c) and 1(d) for the *E. coli* and *P. gingivalis* assemblies, respectively. Linear regression analysis showed a high correlation (R-squared values 0.995 and 0.984, resp.) between the contig read depth and the number of BLAST hits in the reference genome. These results indicate that contig read depth is indeed proportional to the number of copies present in the genome with a very high correlation.

For both genomes, Sanger reads are available. We assembled these reads, using Newbler to make sure sequences from repeated regions would collapse into single contigs with higher read depth. When the per-contig read depth analysis comparison with number of blast hits to the reference genome was repeated for these assemblies, a lower correlation was found compared to the assemblies using 454 reads (R-squared values 0.897 and 0.871 for *E. coli* and *P. gingivalis*, resp.). This is a result of the less even distribution of the Sanger reads across the genome [1, 5–8]. In addition, it has been reported that especially repeated sequences show a particularly low cloning efficiency and therefore are underrepresented in Sanger read datasets [19], a fact that contributed to reduced correlation between read depth and copy number for Sanger based assemblies. Bailey et al. [20] observed a strong correlation when comparing the number of whole genome shotgun Sanger reads mapped to 5 kb segments of the human genome with known diploid copy number. However, the genomic regions studies were masked for repeat sequences in that study.

Making use of annotations of the reference genomes, we analyzed which genomic regions the high read depth contigs of the GS FLX assemblies were derived from. Supplementary

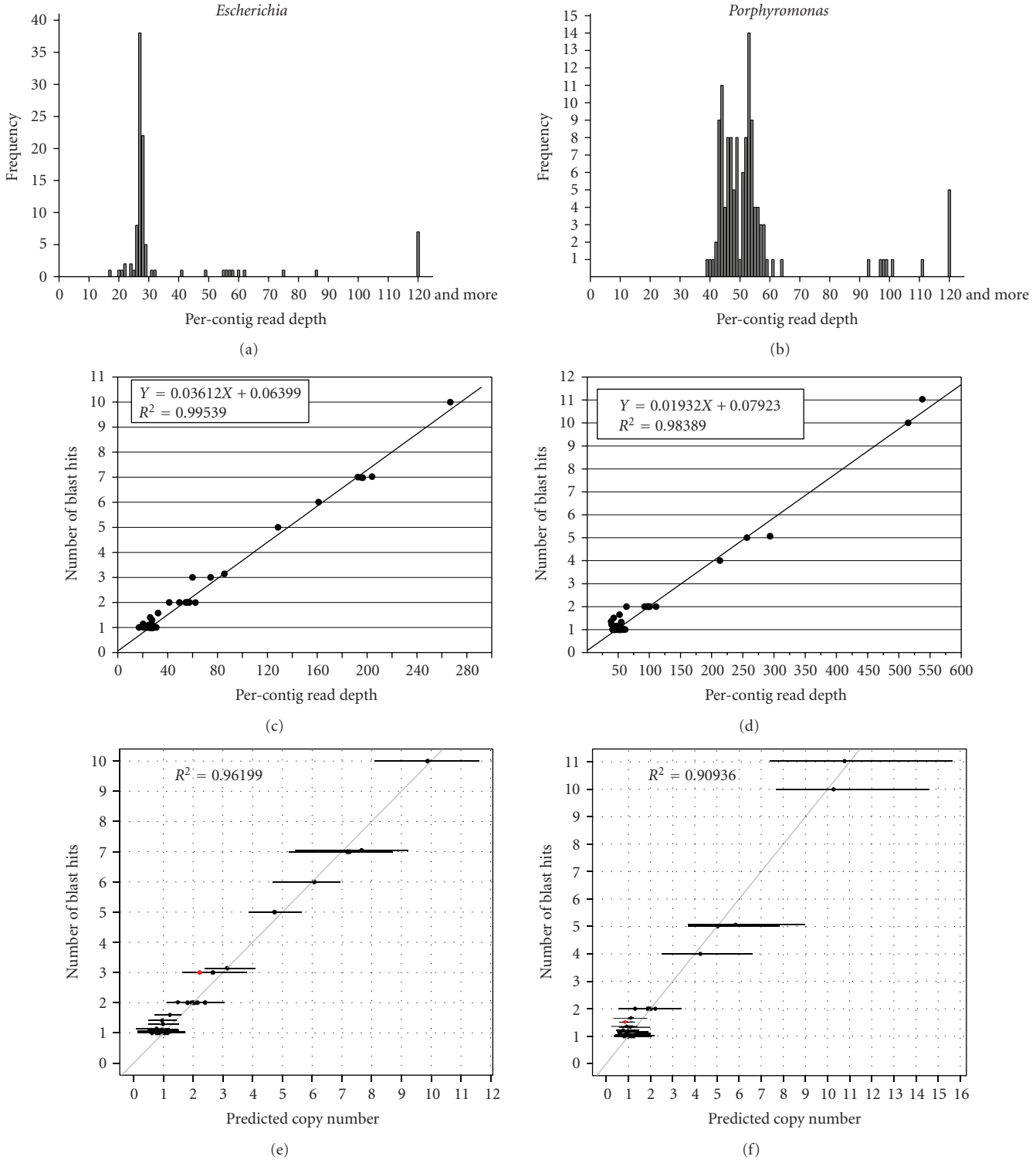


Figure 1: *Correlation between per-contig read depth and genomic copy number*. (a) and (b), per-contig read depth frequency distributions for the 454 assemblies for *E. coli K12* and *P. gingivalis*, respectively. (c) and (d): Scatter plot showing the number of BLASTX hits against the reference genome versus the per-contig read depth for each contig for the two assemblies. Linear regression curves are shown. Inset: the result of linear regression analysis. (e) and (f): comparison of the number of BLASTX hits against the reference genome versus the predicted per-contig copy number based on the statistical approach described in the paper. For each contig, the predicted copy number is indicated with the upper and lower confidence intervals. The 1 : 1 line is shown in grey. Copy number estimates with confidence intervals that do not include the known copy number (number of blast hits) are shown in red. Inset: the result of linear regression analysis.

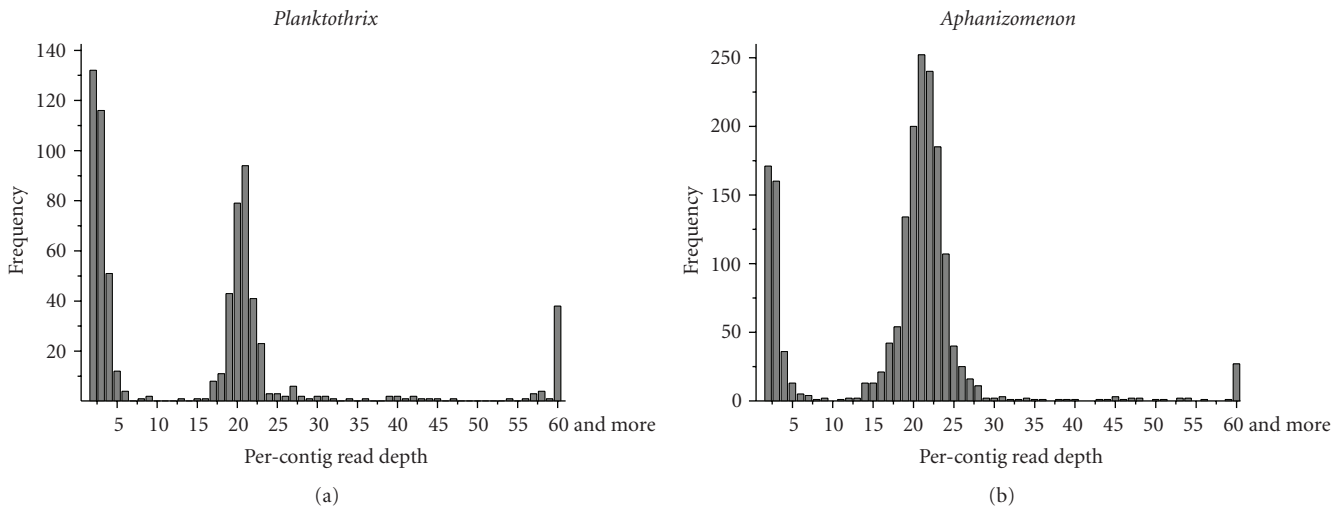


Figure 2: *Per-contig read depth distribution for the cyanobacterial de novo genome assemblies*. Per-contig read depth frequency distributions for the *P. rubescens* NIVA CYA98 (a) and *A. flos-aquae* 10E6 (b) assemblies.

Table II reports the number of BLAST hits for each contig and the annotations in the regions of the BLAST hits. Many of the regions with multiple copies were annotated as transposons or transposon-related coding sequences. The 5S, 16S, and 23S ribosomal RNA genes were also among the high read depth contigs. Transposases and ribosomal RNA regions are known to be present in several copies in bacterial genomes [21, 22].

Next, we tested whether, for a given contig with high read depth, the number of copies present in the genome could be estimated by taking into account the per-contig read depth of all contigs. This would be useful in a situation where no reference genome is available, such as in *de novo* genome sequencing. To this end, we developed a method for estimating the number of copies present in the genome by assessing the probabilistic relationship between read depth and the number of genomic copies (see Methods). Given truly random sampling of the genome, read depth would follow a Poisson distribution [23], but significant deviations from Poisson assumptions found in real data caused us to employ a more robust methodology.

For both the *E. coli and P. gingivalis* assemblies, we estimated the average oversampling level and the number of copies of each contig present in the genome, together with the upper and lower confidence intervals. The predicted copy numbers were compared to the number of blast hits for each contig for both assemblies, see Figures 1(e) and 1(f), respectively. The copy number estimates and their confidence intervals are also shown in Supplementary Table II. The results show an excellent match between predicted copy number and number of blast hits. Variance and thus confidence intervals increased with mean copy number, which is to be expected in an approximated Poisson process. The method correctly predicted that a contig came from a repeated region (copy number of at least 1.5x) of the genome for 17 out of the 18 *E. coli*, and 11 out of 12 *P. gingivalis* contigs, respectively. Therefore, both datasets showed one false negative contig. No single-copy contigs (number of blast hits below 1.5x) had a copy number estimate above 1.5x, therefore there were no false positives. The method therefore had a 100% specificity for both genomes, and a sensitivity of 94% for the *E. coli*, and 92% for the *P. gingivalis* genomes, respectively. For contigs from repeated regions of the genomes, the actual copy number fell outside the confidence intervals for only 3% of the *E. coli* contigs, and none of the *P. gingivalis* contigs, respectively, see Supplementary Table II. It cannot be excluded that this is the result of either a misassembly in the reference genome, a misassembly by Newbler, or an actual copy number variation event in the strain sequenced.

The excellent correspondence between the "known" (based on blast hits) and predicted genomic copy number indicates that the method presented here represents a robust and flexible way to estimate the number of copies for contigs when bacterial genomes or other low-redundancy sequences are sequenced *de novo*.

3.2. De Novo Assemblies of Cyanobacterial Genomes. A major focus of our research is how different variants of large nonribosomal peptide synthetases (NRPS) and polyketide synthetase gene (PKS) clusters [24, 25] evolve in the context of cyanobacterial subpopulation differentiation [26, 27]. To this end, the genomes of the cyanobacterial strains *Planktothrix rubescens* NIVA CYA98 and *Aphanizomenon flos-aquae* 10E6 were studied by whole genome shotgun 454 sequencing and the NRPS gene clusters present in both genomes were identified in the assemblies [4, 12]. In order to determine the repeat content of these genomes, the methods described above were applied to these genomes.

The *P. rubescens* NIVA CYA98 shotgun 454 dataset was assembled, using the 454 Newbler assembler with default settings, into 703 contigs of at least 500 bp [4]. For this assembly, the distribution of the per-contig read depths is shown in Figure 2(a). The distribution showed a large

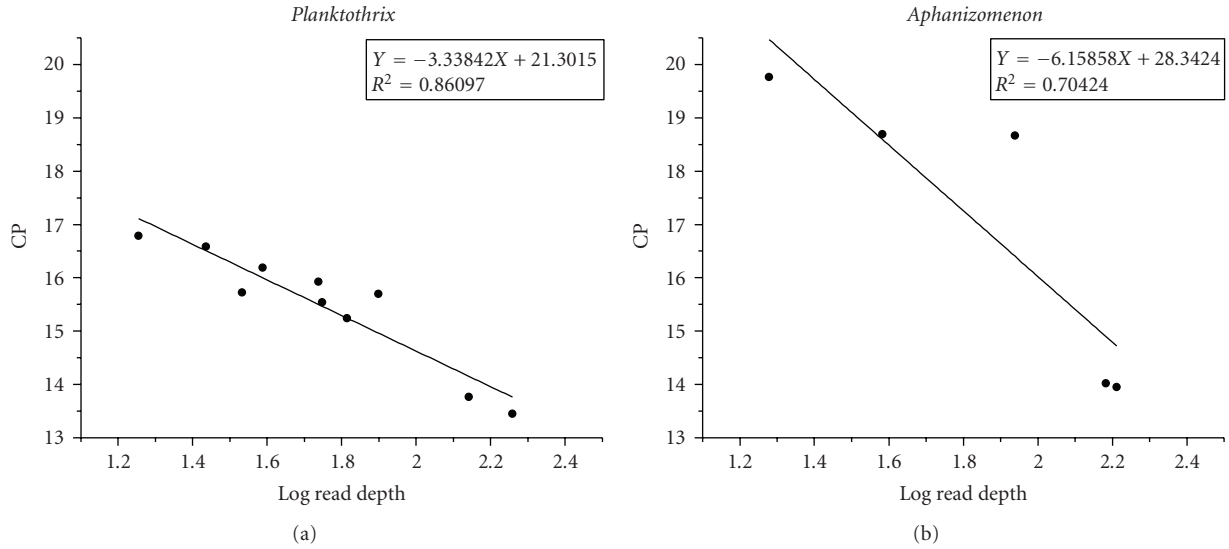


Figure 3: *Correlation between read depth and abundance as determined by real-time PCR*. Scatter plots of CP values obtained with real-time PCR, versus the logarithm of the local read depth of the region of the PCR primers. (a) *P. rubescens* assembly (b) *A. flos-aquae* assembly. Linear regression curves are shown with the regression result in the inset.

peak around a read depth between 2 and 5x and a second peak between 19 and 23x. However, there were also a number of contigs with a higher read depth, up to 260x. We also assembled the reads derived from sequencing a shotgun genomic DNA library from *A. flos-aquae* 10E6 [12]. The resulting 1811 contigs showed a similar read depth distribution as the *P. rubescens* assembly (Figure 2(b)) with a peak in the distribution from 2 to 5x, and a second peak from 17 to 24x. The highest read depth for this assembly was 301x.

We hypothesized that contigs with a read depth below 10x were derived from contaminating genomic DNA derived from other bacteria present in the culture, this will be discussed in the next section. Analogously to the *E. coli* and *P. gingivalis* assemblies, we hypothesized that the peak in the read depth distribution around 20x represented contigs that had a single copy in the genome, while the remaining, higher read depth contigs were derived from reads stemming from genomic regions present in multiple copies.

To test this latter hypothesis experimentally, real-time PCR primer sets were designed for both cyanobacterial genome assemblies for a selection of contigs with varying degrees of read depth (see Methods). For each species, real-time PCR experiments were performed on the same template for all primer sets. Next, the resulting real-time PCR CP values (CP: Crossing Point, the cycle at which the fluorescence of a sample rises above the background fluorescence; CP values are expected to drop logarithmically with the amount of starting material) were plotted against the log of the local average read depth in the contig at the region of the PCR primers. The results showed high correlation between the CP value and the log of the read depth for both assemblies, with *R*-squared values of 0.861 and 0.704, respectively (Figure 3). This further confirms our assumption that read depth of contigs assembled by Newbler is proportional to the number of copies present in the genome being sequenced.

In order to estimate the genomic copy number of the high read depth contigs for the cyanobacterial assemblies, we used the statistical method developed in this work described above, that is, the empirical, nonparametric approach. We excluded contigs with a read depth below 10x for this analysis. The results are presented in Supplementary Figure 1. For the *P. rubescens* assembly, 65 out of the 385 contigs (16.9%) were determined to be from repeated regions (estimated copy number of at least 1.5x); 16 (4.2%) contigs had an estimated copy number over 5x, with a maximum of 12.4x. These repeated contigs contained a total of 216,819 bp (3.8% of the total assembly length). Of the 1419 *A. flos-aquae* contigs, 55 (3.9%) showed an estimated copy number over 1.5x, 11 (0.8%) over 5x with a maximum of 14.2x. A total of 102,628 bp (1.8% of the total assembly length) was contained in these contigs. Confidence intervals for these estimates are larger than for the *E. coli* and *P. gingivalis* assemblies, most likely caused by the fact that the repeated contigs for these assemblies were longer, and thus showed more repeated sequences within each contig or partial contigs being repeated in other places. Table 1 shows the copy number estimates for contigs with an estimated copy number of at least 5. The table also shows the result of comparisons of these contigs against Genbank using BLASTX. For each contig, BLASTX hits are presented in the table. Many high read depth contigs had BLAST hits in transposases, including the four contigs with the highest read depths in both assemblies. Finally, The table also shows the contigs determined by BLASTN searches to contain the 16S and 23S ribosomal RNAs. Based on the copy number estimates, both the *P. rubescens* and *A. flos-aquae* genomes contain most likely four copies of these rRNAs. Supplementary Table III shows the same results as Table 1, but including all contigs

Table 1: *Annotation of the high read depths contigs for P. rubescens and A. flos-aquae assemblies*. For each contig with an estimated copy number of at least 5x, the length, read depth, and estimate of copy number ("Est. copy number") with upper and lower Confidence interval Limits (CL) are shown. In addition, BLASTX results are shown (maximum E value 10^{-16}). When a contig had hits in multiple different regions, these are separated by a comma. The species to which the BLAST hit belongs is shown in between square brackets.

(a) *P. rubescens*

Contig	Length (bp)	Read depth	Est. copy number	Lower CL	Upper CL	Features
13664	1309	259.7	12.4	9.1	19.1	transposase, IS4 family protein (Nostoc punctiforme PCC 73102)
13972	937	190.3	9.2	6.5	15.1	transposase, IS4 family protein (Cyanothece sp. PCC 8802)
13823	1190	180.2	8.9	6.3	12.9	transposase (Microcystis aeruginosa NIES-843)
13688	1109	176.3	8.8	5.8	12.9	transposase (Trichodesmium erythraeum IMS101)
136	3509	173.3	8.6	5.6	13.7	No hits
13792	9424	173.1	8.7	5.3	13.3	hypothetical protein Npun_R2618 [Nostoc punctiforme PCC 73102], DnaB domain protein helicase domain protein (Cyanothece sp. PCC 7822)
13610	852	163.2	8.0	5.2	12.6	No hits
13711	1051	145.0	7.2	5.3	10.2	No hits
13735	2163	144.0	7.2	4.0	11.5	conserved hypothetical protein (Cyanothece sp. PCC 7425)
13901	611	140.9	7.1	4.7	9.8	transposase, IS605 OrfB family (Cyanothece sp. PCC 8801)
13846	902	132.8	7.2	2.5	9.8	hypothetical protein L8106_22791 (Lyngbya sp. PCC 8106)
14014	770	123.8	6.0	4.3	9.8	Histone-like DNA-binding protein (Lyngbya sp. PCC 8106)
13843	1712	111.4	5.6	3.5	8.5	RNA-directed DNA polymerase (Microcystis aeruginosa NIES-843)
13469	669	104.2	5.2	3.8	7.4	No hits
13858	641	99.5	5.1	3.0	7.2	hypothetical protein L8106_22631 (Lyngbya sp. PCC 8106)
13921	2057	99.3	5.2	1.7	9.4	transposase (Microcystis aeruginosa NIES-843)
13462	1575	76.9	3.7	2.5	6.0	16S rRNA
13463	2891	75.4	3.7	2.3	6.0	23S rRNA

(b) *A. flos-aquae*

Contig	Length (bp)	Read Depth	Est. copy Number	Lower CL	Upper CL	Features
13355	911	301.0	14.2	9.6	24.6	transposase (Microcystis aeruginosa NIES-843)
13273	748	257.2	12.5	6.0	25.7	transposase (Nodularia spumigena CCY9414), transposase (Nodularia spumigena CCY9414)
13683	1256	189.6	8.6	5.5	17.7	transposase (Nostoc sp. PCC 7120)
14262	560	174.4	8.2	5.5	14.7	transposase and inactivated derivatives (Syntrophus aciditrophicus)
14128	1185	165.6	8.0	5.0	13.5	unnamed protein product (Microcystis aeruginosa PCC 7806)
12918	1870	163.3	7.7	4.5	14.3	transposase (Microcystis aeruginosa NIES-843)
525	1566	160.6	7.6	4.4	13.7	hypothetical protein AM1_C0013 (Acaryochloris marina MBIC11017), hypothetical protein AM1_C0013 (Acaryochloris marina MBIC11017)
214	567	156.5	7.2	4.9	12.7	conserved hypothetical protein (Microscilla marina ATCC 23134), conserved hypothetical protein (Microscilla marina ATCC 23134)
12934	575	137.3	6.6	3.6	12.0	IS1 transposase subfamily, putative (Synechococcus sp. PCC 7335)
13740	502	115.3	5.4	3.8	9.3	transposase (Cyanothece sp. ATCC 51142)
13286	1782	109.0	6.2	0.2	14.2	No hits
13746	1585	91.0	4.3	2.3	8.1	16S
13744	2169	89.9	4.2	2.7	7.8	23S

determined to be from repeated regions (estimated copy number at least 1.5x).

The results of our analysis of *de novo* bacterial genome assemblies using the Newbler assembly program show that read depth-based copy number estimations can be used to *de novo* identify contigs containing sequences present multiple times in the genome. In addition, we show how contig specific read depth can be used to predict the number of genomic copies of these sequences in the genome, even when no paired read data is available to resolve repeats.

One rationale for sequencing the genome of *P. rubescens* was to identify all NRPS gene clusters putatively present in the genome. When we studied the read depth of the contigs containing these clusters, we concluded that two NRPS gene clusters were present in the genome in multiple copies, while the others were present as single copies [4]. Recombination occurs frequently in NRPS gene clusters (see, [25, 28]) and in cyanobacterial genomes in general [29, 30]. Identifying duplications is not only important in cyanobacterial research, but in most genome projects. The read depth based copy number estimation method presented here will allow us to assess NRPS gene cluster duplications by 454 shotgun sequencing of the relevant strains.

In principle, assessing the relation between read depth and genomic copy number should be possible for all shotgun assemblies based on reads from other next-generation sequencing technologies that do not amplify sequences by propagating in bacteria. Nevertheless, this analysis depends partly on the assembly program; not all programs treat repeats the way the Newbler program does. In addition, read coverage for Illumina/Solexa reads is likely biased towards regions with higher GC content [31, 32]. We therefore expect an analysis of contigs from assemblies using Illumina reads to show a lower correlation between read depth and genomic copy number. Several recent publications describe methods, using next-generation sequencing technology, for determining copy number variations based on read depth [6, 32–38]. These methods are all based on a well-annotated, usually human, reference genome. Our method aims to provide copy number *estimations* for *de novo* assemblies of bacterial genomes.

3.3. Detection of Contigs of Co-Cultured Bacteria in Non-Axenic Cyanobacterial Cultures. Both the *P. rubescens* and *A. flos-aquae* assemblies showed a large number of contigs with low read depths, between 2 and 5x (Figures 2(a) and 2(b)). We further examined the contigs of these assemblies by plotting the GC percentage of each contig versus the contig read depth for all contigs (Figures 4(a) and 4(b)). The figure shows the contigs with low read depth to the left of the dotted line at 10x read depth and the remaining contigs to the right of this line. For both these assemblies, the contigs with low read depths could be further divided into contigs with a low GC percentage (between 20–25% and 40–45% for *P. rubescens* and *A. flos-aquae*, resp.) and a high GC percentage (between 45–50% and 65–70%, resp.); see Figures 4(a) and 4(b).

The two groups of contigs, having low (below 10x) and high (above 10x) read depths, were separately compared to the nonredundant NCBI protein database using BLASTX [16], and the top 25 hits were recovered. This analysis was limited to the contigs of at least 500 bp because of the large number of small, low read depth contigs present in the assembly. The program MEGAN [MEtaGenomics Analysis, 17] was used to summarize the results and plot them onto the NCBI taxonomic tree (using a Lowest Common Ancestor algorithm) in a comparative way (Figures 4(c) and 4(d)). The MEGAN results indicated that the contigs with read depths of 10x and above of both the *P. rubescens* (Figure 4(c)) and *A. flos-aquae* (Figure 4(d)) assemblies were mainly cyanobacterial, with a small fraction of the *A. flos-aquae* contigs in the *Proteobacteria* part of the tree. For *P. rubescens*, the contigs with read depths below 10x fell mainly in the phylum *Proteobacteria*, with some in the genus *Bacteroides*. For *A. flos-aquae* (Figure 4(d)), the low read depth contigs mapped mainly in the order *Flavobacteriales* (genus *Bacteroides*) with the rest determined as *Proteobacteria*. For both assemblies, the contigs with low read depths and *low* GC percentage were assigned to the *Bacteroides* part of the tree, while the *Proteobacteria* subtree contained the low read depth, *high* GC percentage contigs.

The contigs with a read depth below 10x were most likely derived from co-cultured bacteria present in the sample used for DNA extraction. It is difficult to obtain axenic cyanobacterial cultures (free of other "contaminating" organisms), and therefore these cultures often contain several different co-cultured bacteria [39, 40, page 139]. The fact that contigs resulting from contaminating reads have a low read depth is a result of the low frequency of the contaminating "genomes": there are relatively few reads present from these genomes (low rate of oversampling), which are assembled in many contigs which have a low read depth. Due to the high oversampling rate of 454 sequencing (typically around 20x or more), the contaminants assemble into contigs with a significantly lower read depth than the contigs of the dominant genome in the sample. Assemblies based on Sanger reads are usually at a lower oversampling (7 to 8x), preventing the contaminants from being assembled into contigs with clearly lower read depth. In fact, in 454 assemblies of cyanobacterial genomes with a too low oversampling (e.g., below 10x), we could not detect the contaminants based on read depth alone (data not shown).

By removing the contaminating contigs (those with a read depth below 10x) from the assemblies, their statistics improved: for *P. rubescens*, the number of contigs was reduced from the original 703 to 385 (55%), and the average contig length almost doubled from 8110 to 14280, and contig N50 length went from 24899 to 25830 (the N50 length is defined such that half of the assembled bases reside in contigs having a length of at least the N50 contig length). For the *Aphanizomenon* assembly, 1419 of the 1811 (78%) contigs remained, and the average length increased from 3217 to 3920, and contig N50 length went from 5731 to 5986. Based on the number of reads present in the different contigs, we determined the level of contamination to be around 6% to 7% of the reads in the cyanobacterial samples sequenced.

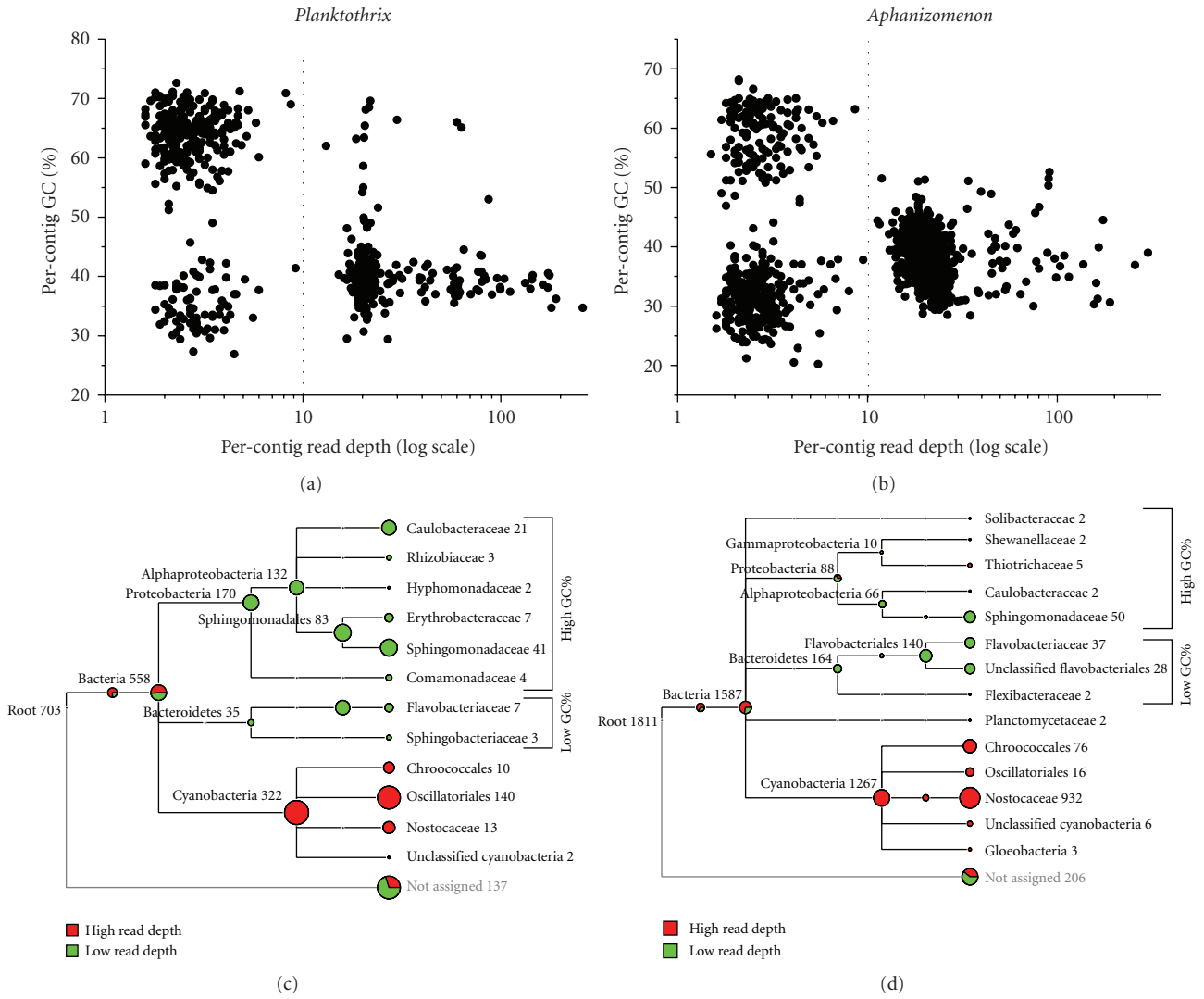


Figure 4: *Identifying contigs from contaminating bacteria in cyanobacterial genome assemblies derived from DNA from nonaxenic cultures.* (a) and (b): scatter plot showing for each contig (minimum length 500 bp), the GC percentage and read depth (log scale) for the *P. rubescens* NIVA CYA98 (a) and *A. flos-aquae* (b) assemblies. Contigs with low read depths to the left of the dotted line at 10x read depth. The low read depth contigs fall into two clusters based on GC percentage. (c) and (d): MEGAN comparisons of the low (green) and high (red) read depth contigs from the *P. rubescens* NIVA CYA98 (c) and *A. flos-aquae* (d) assemblies. Trees collapsed at the Family taxonomic level. Numbers with the taxon names are number of hits summarized to that node and all nodes below in the NCBI taxonomic tree. Circles sizes are log-scale relative to the number of hits. For the contigs with low read depths, it is indicated that if they fall into the high or low GC% cluster of contigs. "Not assigned": contigs that were not assigned to any branch of the tree due to too low bit score (cutoff at 100) or because they are the only contig that were assigned to a particular taxon.

However, these contaminating reads assembled into 45% of the contigs in the case of *P. rubescens* and 22% in the case of *A. flos-aquae*.

The contigs assembled from these contaminating reads were much shorter than the contigs derived from the remaining set of reads. This is actually a direct cause of the low frequency of the contaminating "genomes": there are relatively few reads present of these genomes (low rate of oversampling), resulting in an assembly consisting of many short contigs with a low read depth.

Our per-contig read depth analysis are thus able to distinguish contigs coming from the co-cultured bacteria from those resulting from the genome of interest. In this way, a straightforward method to obtain a "clean" assembly is to simply select the contigs with a read depth above a certain threshold.

3.4. Detection of Host Organism Genomic DNA in BAC Assemblies. We used our analysis approach based on per-contig read depth on two BAC (Bacterial Artificial Chromosome) shotgun sequencing datasets containing genomic regions from salmon. Close to 75,000 reads for BAC 184H23 were assembled into 579 contigs of at least 100 bp, and just over 55,000 reads into 471 contigs for BAC 114L13.

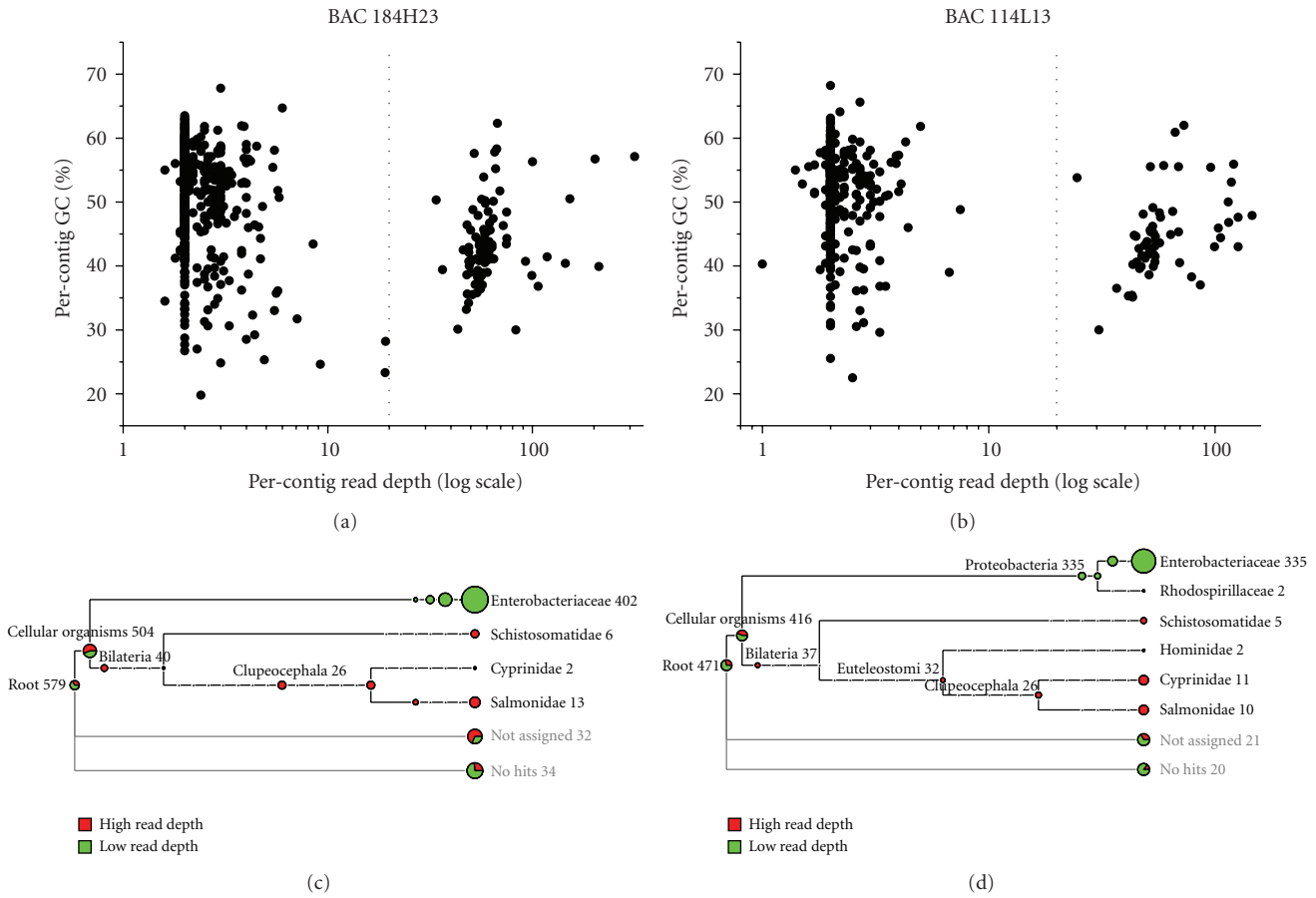


Figure 5: *Identifying bacterial host genomic DNA in BAC assemblies.* (a) and (b): scatter plot showing for each contig (minimum length 100 bp), the GC percentage and read depth (log scale) for Salmon BAC 184H23 (a) and Salmon BAC 114L13 (b) assemblies. Contigs with low read depths to the left of the dotted line at 20x read depth. (c) and (d): MEGAN comparisons of the low and high read depth contigs from the Salmon BAC 184H23 (a) and Salmon BAC 114L13 (b) assemblies. Trees collapsed at the Family level. Contigs with high read depth, in red, cluster into bony fishes (*Clupeocephala*) with a few hits classified as *Schistosomatidae* (flatworms). Numbers with the taxon names are number of hits summarized to that node and all nodes below in the NCBI taxonomic tree. Circles sizes are log-scale relative to the number of hits. "Not assigned": contigs that were not assigned to any branch of the tree due to too low bit score (cutoff at 30) or because they are the only contig that were assigned to a particular taxon.

Figures 5(a) and 5(b) show the GC percentage versus read depth plots for the assemblies of these BACs. As for the cyanobacterial datasets, two clusters of contigs could be observed for these assemblies, those with low read depths (below 20x) and those with high read depths (above 20x). Comparison of the BAC assembly contigs with the NCBI nonredundant protein database using BLASTX, and MEGAN analysis of the resulting hits (shown in Figures 5(c) and 5(d)) identified contigs with low read depths as coming from bacteria, that is, *Enterobacteriaceae* (to which *Escherichia coli* belongs). This indicated that the contigs with low read depth were derived from host genomic DNA that was copurified during BAC DNA sample preparation. The level of host contaminating reads was around 2% for the BAC datasets, however, they assembled into more than 85% of the contigs. The high read depths contigs of both BACs came exclusively from eukaryotes, mainly from *Clupeocephala*, (a group of bony fishes including salmon). A few BLASTX hits were assigned to *Schistosomatidae* (parasitic flatworms), but these contigs had both fish and *Schistosomatidae* as best BLASTX hits. The flatworm hits may result from the fact that *Schistosomatidae* worms can have fish as their host and have been shown to take up host specific genes in their genome [41].

In the case of BAC assemblies (or other sequences propagated in bacteria) where the genome sequence of the host is known, one could remove the contaminating reads with the help of read filtering against the host genome sequence, or BLAST analysis of contigs against the reference genome. We provide an alternative filtering method, based on cleaning up the assembly by using a minimum read depth cutoff.

The approach of using contig read depth selection has been described in one instance before. It was shown that for a 454 assembly of *Sulcia muelleri* [42], an insect symbiont, the genomic DNA prepared to perform 454 sequencing was contaminated with genomic DNA of the host, and of a second symbiont, *Baumannia*. Of the total of 416 contigs

assembled, a selection was made based on average depth [42, supplementary Figure S1], leading to 25 high read depth contigs, of which 23 were further assembled into the complete genome of *S. Muelleri.*

4. Conclusions

Contigs made from 454 reads using the Newbler assembler have a read depth that is linearly correlated to the number of copies present in the genome. Bacterial genome assemblies therefore show a number of contigs with higher read depths that represent repeated genomic regions (e.g., transposons, ribosomal RNA genes). We present a robust statistical method to *de novo* estimate the number of copies of these contig sequences present in the genome based on the contig specific read depth. Most importantly, analysis of contig-specific read depth will extend the amount of information that can be gained from a *de novo* 454 shotgun assembly using Newbler.

In addition, our per-contig read depth analysis is useful to detect and remove contigs assembled from reads derived from contaminating DNA, since contaminant contigs are likely to have low read depths. This is particularly useful when the sequence of the contaminating genome or genomes is not available, such as in the case of reads obtained from impure (nonaxenic) cultures. There are many other possible examples of contaminated read datasets, such as human viruses derived from cell cultures leading to human DNA contamination, and parasite DNA preparations containing host DNA. We expect contig selection based on read depth to be an essential tool for "cleaning up" contaminated 454 shotgun sequencing read assemblies in future studies.

Acknowledgments

The authors would like to thank Anke Stüken for allowing them to use the *Aphanizomenon* dataset, and Unni Grimholt for the BAC datasets. Ave Tooming-Klunderud is acknowledged for expert 454 sequencing, and Tom Kristensen and Bastiaan Star for critical reading of the manuscript. This work was supported by the Norwegian Research Council through grants from the programs AVIT, FUGE, and GenoFisk, respectively.

References

[1] T. Wicker, E. Schlagenhauf, A. Graner, T. J. Close, B. Keller, and N. Stein, "454 sequencing put to the test using the complex genome of barley," *BMC Genomics*, vol. 7, article 275, 2006.

[2] M. Pop and S. L. Salzberg, "Bioinformatics challenges of new sequencing technology," *Trends in Genetics*, vol. 24, no. 3, pp. 142–149, 2008.

[3] Phrap, http://www.phrap.org/.

[4] T. B. Rounge, T. Rohrlack, A. J. Nederbragt, T. Kristensen, and K. S. Jakobsen, "A genome-wide analysis of nonribosomal peptide synthetase gene clusters and their peptides in a *Planktothrix rubescens* strain," *BMC Genomics*, vol. 10, no. 1, article 396, 2009.

[5] 454 Case Study: Genome Coverage of *Neurospora crassa*, http://www.454.com/downloads/454_CASE_STUDY_genome_coverage.pdf.

[6] K. Swaminathan, K. Varala, and M. E. Hudson, "Global repeat discovery and estimation of genomic copy number in a large, complex genome using a high-throughput 454 sequence survey," *BMC Genomics*, vol. 8, article 132, 2007.

[7] D. A. Wheeler, M. Srinivasan, M. Egholm, et al., "The complete genome of an individual by massively parallel DNA sequencing," *Nature*, vol. 452, no. 7189, pp. 872–876, 2008.

[8] J.-M. Aury, C. Cruaud, V. Barbe, et al., "High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies," *BMC Genomics*, vol. 9, article 603, 2008.

[9] M. Riley, T. Abe, M. B. Arnaud, et al., "*Escherichia coli* K-12: a cooperatively developed annotation snapshot—2005," *Nucleic Acids Research*, vol. 34, no. 1, pp. 1–9, 2006.

[10] The NCBI Short Read Archive, http://www.ncbi.nlm.nih.gov/Traces/sra/sra.cgi.

[11] K. E. Nelson, R. D. Fleischmann, R. T. DeBoy, et al., "Complete genome sequence of the oral pathogenic Bacterium *Porphyromonas gingivalis* strain W83," *Journal of Bacteriology*, vol. 185, no. 18, pp. 5591–5601, 2003.

[12] A. Stüken, A. J. Nederbragt, and K. S. Jakobsen, "Cylindrospermopsin biosynthesis cluster in *Aphanizomenon flos-aquae*," submitted for publication.

[13] M. Margulies, M. Egholm, W. E. Altman, et al., "Genome sequencing in microfabricated high-density picolitre reactors," *Nature*, vol. 437, no. 7057, pp. 376–380, 2005.

[14] R Development Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2008.

[15] Primer3, http://primer3.sourceforge.net/.

[16] S. F. Altschul, T. L. Madden, A. A. Schäffer, et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[17] D. H. Huson, A. F. Auch, J. Qi, and S. C. Schuster, "MEGAN analysis of metagenomic data," *Genome Research*, vol. 17, no. 3, pp. 377–386, 2007.

[18] F. R. Blattner, G. Plunkett III, C. A. Bloch, et al., "The complete genome sequence of *Escherichia coli* K-12," *Science*, vol. 277, no. 5331, pp. 1453–1462, 1997.

[19] S. L. Chissoe, M. A. Marra, L. Hillier, R. Brinkman, R. K. Wilson, and R. H. Waterston, "Representation of cloned genomic sequences in two sequencing vectors: correlation of DNA sequence and subclone distribution," *Nucleic Acids Research*, vol. 25, no. 15, pp. 2960–2966, 1997.

[20] J. A. Bailey, Z. Gu, R. A. Clark, et al., "Recent segmental duplications in the human genome," *Science*, vol. 297, no. 5583, pp. 1003–1007, 2002.

[21] P. Siguier, J. Filee, and M. Chandler, "Insertion sequences in prokaryotic genomes," *Current Opinion in Microbiology*, vol. 9, no. 5, pp. 526–531, 2006.

[22] Z. M.-P. Lee, C. Bussema III, and T. M. Schmidt, "*rrn*DB: documenting the number of rRNA and tRNA genes in bacteria and archaea," *Nucleic Acids Research*, vol. 37, supplement 1, pp. D489–D493, 2009.

[23] E. S. Lander and M. S. Waterman, "Genomic mapping by fingerprinting random clones: a mathematical analysis," *Genomics*, vol. 2, no. 3, pp. 231–239, 1988.

[24] A. Tooming-Klunderud, T. Rohrlack, K. Shalchian-Tabrizi, T. Kristensen, and K. S. Jakobsen, "Structural analysis of a non-ribosomal halogenated cyclic peptide and its putative operon

from *Microcystis*: implications for evolution of cyanopeptolins," *Microbiology*, vol. 153, no. 5, pp. 1382–1393, 2007.

[25] B. Mikalsen, G. Boison, O. M. Skulberg, et al., "Natural variation in the microcystin synthetase operon mcyABC and impact on microcystin production in *Microcystis* strains," *Journal of Bacteriology*, vol. 185, no. 9, pp. 2774–2785, 2003.

[26] T. Rohrlack, B. Edvardsen, R. Skulberg, et al., "Oligopeptide chemotypes of the toxic freshwater cyanobacterium *Planktothrix* can form subpopulations with dissimilar ecological traits," *Limnology and Oceanography*, vol. 53, no. 4, pp. 1279–1293, 2008.

[27] T. B. Rounge, T. Rohrlack, B. Decenciere, B. Edvardsen, T. Kristensen, and K. S. Jakobsen, "Subpopulation differentiation associated with nonribosomal peptide synthetase gene cluster dynamics in the cyanobacterium *Planktothrix*," *the Journal of Phycology*, in press.

[28] Y. Tanabe, K. Kaya, and M. M. Watanabe, "Evidence for recombination in the microcystin synthetase (mcy) genes of toxic cyanobacteria *Microcystis* spp," *Journal of Molecular Evolution*, vol. 58, no. 6, pp. 633–641, 2004.

[29] O. Zhaxybayeva, J. P. Gogarten, R. L. Charlebois, W. F. Doolittle, and R. T. Papke, "Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events," *Genome Research*, vol. 16, no. 9, pp. 1099–1108, 2006.

[30] K. Rudi, O. M. Skulberg, and K. S. Jakobsen, "Evolution of cyanobacteria by exchange of genetic material among phyletically related strains," *Journal of Bacteriology*, vol. 180, no. 13, pp. 3453–3461, 1998.

[31] J. C. Dohm, C. Lottaz, T. Borodina, and H. Himmelbauer, "Substantial biases in ultra-short read data sets from high-throughput DNA sequencing," *Nucleic Acids Research*, vol. 36, no. 16, p. e105, 2008.

[32] D. Y. Chiang, G. Getz, D. B. Jaffe, et al., "High-resolution mapping of copy-number alterations with massively parallel sequencing," *Nature Methods*, vol. 6, no. 1, pp. 99–103, 2009.

[33] E. Arner, E. Kindlund, D. Nilsson, et al., "Database of *Trypanosoma cruzi* repeated genes: 20 000 additional gene variants," *BMC Genomics*, vol. 8, article 391, 2007.

[34] C. Alkan, J. M. Kidd, T. Marques-Bonet, et al., "Personalized copy number and segmental duplication maps using next-generation sequencing," *Nature Genetics*, vol. 41, no. 10, pp. 1061–1067, 2009.

[35] S. Yoon, Z. Xuan, V. Makarov, K. Ye, and J. Sebat, "Sensitive and accurate detection of copy number variants using read depth of coverage," *Genome Research*, vol. 19, no. 9, pp. 1586–1592, 2009.

[36] B. Daines, H. Wang, Y. Li, Y. Han, R. Gibbs, and R. Chen, "High-throughput multiplex sequencing to discover copy number variants in Drosophila," *Genetics*, vol. 182, no. 4, pp. 935–941, 2009.

[37] C. Xie and M. T. Tammi, "CNV-seq, a new method to detect copy number variation using high-throughput sequencing," *BMC Bioinformatics*, vol. 10, article 80, 2009.

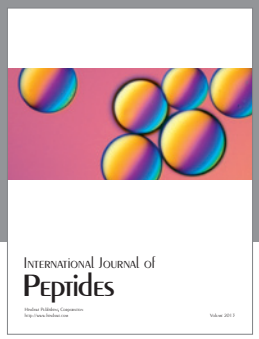
[38] J. Macas, P. Neumann, and A. Navratilova, "Repetitive DNA in the pea (*Pisum sativum* L.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and *Medicago truncatula*," *BMC Genomics*, vol. 8, article 427, 2007.

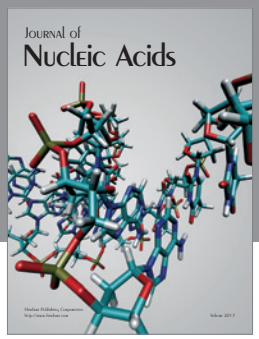
[39] M. J. Ferris and C. F. Hirsch, "Method for isolation and purification of cyanobacteria," *Applied and Environmental Microbiology*, vol. 57, no. 5, pp. 1448–1452, 1991.

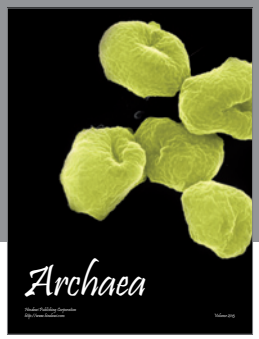
[40] H. W. Paerl, "Marine plankton," in *The Ecology of Cyanobacteria Their Diversity in Time and Space*, B. Whitton and M. Potts, Eds., pp. 121–148, Springer, New York, NY, USA, 2000.

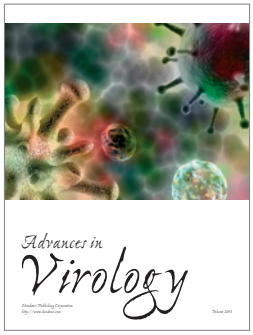
[41] F. Liu, J. Lu, W. Hu, et al., "New perspectives on host-parasite interplay by comparative transcriptomic and proteomic analyses of *Schistosoma japonicum*," *PLoS Pathogens*, vol. 2, no. 4, p. e29, 2006.

[42] J. P. McCutcheon and N. A. Moran, "Parallel genomic evolution and metabolic interdependence in an ancient symbiosis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 49, pp. 19392–19397, 2007.









Hindawi

Submit your manuscripts at
http://www.hindawi.com