PEPR: Pipelines for Evaluating Prokaryotic References

Nathan D. Olson · Justin M. Zook · Daniel V. Samarov · Scott A. Jackson · Marc L. Salit

Received: date / Accepted: date

Abstract The rapid adoption of microbial whole genome sequencing in public health, clinical testing, and forensic laboratories requires the use of validated measurement processes. Reference materials that are well characterized, homogeneous, and stable can be used to evaluate measurement processes and help to establish confidence in the results. Given the variety of microbial genome sequencing applications and platforms, as well as the vast microbial genomic diversity, there is a need for application-specific genomic materials for method validation. We have developed a reproducible and transparent bioinformatics tool for characterizing prokaryotic genomic materials; "PEPR", Pipelines for Evaluating Prokaryotic References. We demonstrate the tool and its output using using sequencing data while developing a Staphylococcus aureus candidate genomic reference material.

Keywords Microbiology · Whole genome sequencing · Bioinformatics

1 Introduction

Over the past decade, the availability of affordable and rapid Next-Generation Sequencing (NGS) technology has revolutionized the field of microbiology. Arguably the most discriminatory typing method available, whole genome sequencing (WGS), has been adopted not only by the research community, but also by public health laboratories, clinical testing laboratories, and the forensic community. High stakes decisions are often made based on the outcome of a WGS assay. To increase confidence in WGS assay results a critical assessment of the errors inherent to the measurement processes is required. A number of sources of error associated with the WGS measurement process have been identified, but the degree to which they can be predicted, controlled, or compensated varies significantly [14].

Well-characterized, homogeneous, and stable genomic materials can be used to evaluate methods and aid in establishing confidence in results from a measurement process. For example, we recently characterized a whole human genome reference material (National Institute of Standards and Technology, NIST; Reference Material 8398) to assess
performance of variant calling in human genomic samples [25], but no reference materials for microbial sequencing currently exist. Four microbial genomic DNA candidate reference materials are being developed at the National Institute of Standards and Technology (NIST) to meet this need. When considering the extensive genomic diversity of prokaryotic organisms as well as the rapidly evolving and diverse DNA sequencing applications, we envision the need for a wide variety of application-specific genomic materials for use in method validation and benchmarking. Currently, many laboratories and sequencing centers are using in-house materials as part of a regular method validation and quality control system. However, the degree to which these materials are characterized varies significantly, and therefore a common pipeline for characterizing prokaryotic genomic materials is needed.

PEPR, Pipelines for Evaluating Prokaryotic References, a set of reproducible and transparent bioinformatic pipelines, was developed to characterize genomic materials for use in WGS method validation. Using the pipeline increases confidence in method validation through the ability to develop better characterized control materials. PEPR characterizes prokaryotic genomic material for purity and homogeneity of the genome sequence, as well as presence of genomic material other than the material genus. The general approach to material characterization that guided the development of PEPR is the use of orthogonal sequencing methods along with technical replicates to obtain consensus values for the characterized properties. These consensus values are our best current estimates of the true values. We do not assert probabilistic estimates of confidence or confidence classification values with the sequence data, as we lack good models of biases or systematic errors of current sequencing technologies. Here we will first describe PEPR then show how PEPR was used to characterize NIST *Staphylococcus aureus* genomic DNA candidate reference reference material.

2 Methods

2.1 Pipelines for Evaluating Prokaryotic References: PEPR

PEPR consists of three bioinformatic pipelines written in python (Fig. 2.1). The three bioinformatic pipelines are genome evaluation, genome characterization, and genomic purity. A YAML file (http://yaml.org) is used to define pipeline inputs. The pipeline coordinates the execution of a number of commandline tools, logging the standard output and standard error for each executed command in time-stamped files for reference and debugging. Pipeline code is available at (https://github.com/usnistgov/pepr). To reduce the barrier for reuse, two Docker (https://www.docker.com/) containers are available with pre-installed pipeline dependencies. Docker is a lightweight virtual environment that facilitates the sharing and distribution of computing environments and can be run on any desktop, cloud, or high performance computing environment, regardless of the operating system. The *pepr* container (https://registry.hub.docker.com/u/natedolson/pepr) includes dependencies for the genome evaluation and characterization pipelines, excluding the Genome Analysis Toolkit (due to licensing restrictions). The *docker-pathoscope* container has dependencies for the genomic purity pipeline installed (https://registry.hub.docker.com/u/natedolson/docker-pathoscope/).

A software package, *peprr*, was developed for the statistical computing language R [17] to compile the output from the genome evaluation, characterization, and genomic purity pipelines. The compiled data was formated into a series of data tables within a SQLite, *peprDB*, database to facilitate downstream analysis [23]. The package includes functions to generate a number of summary tables and figures, including those in this publication.

2.2 Genome Evaluation Pipeline

The Genome Evaluation Pipeline, is the first step in the PEPR workflow, and is used to reduce errors in the user provided genome assembly prior to characterization. The evaluation pipeline consists of three steps. Illumina sequencing data are retrieved from the Genbank Sequence Read Archive (SRA) using the sratoolkit *fastq-dump* command (http://ncbi.github.io/sra-tools/). Next, sequencing reads are mapped to the reference genome using BWA mem algorithm [7]. Finally, Pilon is used to evaluate and polish the reference assembly [22]. The corrected reference genome is then used as input for the Genome Characterization Pipeline.
2.3 Genome Characterization Pipeline

The Genome Characterization Pipeline uses replicate sequence dataset from multiple sequencing platforms to characterize the corrected reference genome produced by the Genome Evaluation Pipeline at the individual base level. Illumina data are aligned to the reference genome using the same methods as the evaluation pipeline. IonTorrent PGM data using the TMAP algorithm [3]. Additionally, if Pacific Biosciences (PacBio) sequencing data is used to generate the input reference assembly. The PacBio sequencing data is mapped to the reference genome using the BWA mem algorithm [7] and Next the sequence alignment files are processed prior to downstream analysis by marking duplicates with Picard's MarkDuplicates command (http://broadinstitute.github.io/picard) and realigning reads mapping to regions with insertions or deletions using the GenomeAnalysisToolKit [10, 2]. After refining the alignment files, base level analysis is performed using the short read sequencing data. For each platform a VCF (variant call format) file with a number of summary statistics is generated using SAMtools mpileup [8]. A base purity metric is calculated from the resulting VCF files. The base purity metric is the number of bases in reads aligned to a genome position that are in agreement with the reference base divided by the total number of reads supporting the alternate base called by SAMtools at that genome position. Homogeneity analysis, a measure of genomic content similarity between vials of the reference material, is performed by first generating a pileup file using SAMtools mpileup for each dataset then performing pairwise tumor-normal variant calling using VarScan [5]. In this work, VarScan looks specifically for differences between vials in the proportion of reads containing variants. A standard Benjamini-Hochberg procedure was used to assess
the power of the homogeneity analysis (Supplemental Material, https://github.com/DanSBS/NGSPower). Additionally, a number of summary statistics are calculated for the sequencing datasets using Picard’s Collect Multiple Metrics (http://broadinstitute.github.io/picard).

2.4 Genome Purity Pipeline

THE Genomic Purity Pipeline assesses the purity of the genomic material, in terms of the presence of DNA from sources other than the expected genus was assessed using the metagenomic taxonomic read classification algorithm PathoScope 2.0 [4]. This method uses an expectation maximization algorithm where the sequence data are first mapped to a database comprised of all sequence data in the Genbank nt database. Then, through an iterative process PathoScope re-assigns ambiguously mapped reads to a taxonomic group based on the proportion of reads mapped unambiguously to individual taxonomic groups in the database. Using short read sequencing data as input PathoScope 2.0 first filters and trims low quality reads (PathoQC), followed by mapping reads to a reference database (PathoMap - a wrapper for bowtie2 [6]), then the expectation-maximization algorithm (PathoID) is used for the taxonomic classification. The annotated Genbank nt database provided by the PathoScope developers was used as the reference database (ftp://pathoscope.bumc.bu.edu/data/nt_ti.fa.gz).

2.5 Candidate Reference Material S. aureus Sequencing Data

Sequencing data and the reference assembly for the NIST candidate reference material S. aureus was used to demonstrate how PEPR is used to characterize a genomic material. A de novo genome assembly from Pacific Biosciences (PacBio) long read sequencing data was used as input for PEPR. Prior to being used as input the assembly was validated using optical mapping data. Eight replicate vials of the candidate reference material was sequenced on the Illumina MiSeq and Ion Torrent PGM sequencing platforms (Supplemental Methods).

3 Results

The output from the analysis of the S. aureus candidate reference material using PEPR is presented below. The results provide examples of the type of information, summary figures, and tables that can be generated with PEPR.

3.1 Sequencing Data Summary Statistics

Summary statistics were calculated for the number of reads, mapped read length, insert size for paired-end datasets as well as coverage for S. aureus datasets (Table 1). The MiSeq sequencing run had an average of 1.7 million paired-end reads per library with a median read length of 232 bp, whereas the PGM sequencing run produced 0.2 million reads per library on average with a median read length of 232 bp. The higher throughput and paired-end reads resulted in a higher per-library coverage for MiSeq compared to PGM (251X vs. 36X). The three PacBio datasets are technical sequencing replicates (SMRT cells) from the same sequencing library, with a median subread length 10436bp and 302X total coverage. Between the three platforms a total coverage of 4611X was obtained.

3.2 Genome Evaluation

The first step in PEPR is the Genome Evaluation Pipeline. Short read sequencing data is used to identify and correct errors in the user-provided reference genome, in this case the validated PacBio assembly. Running Pilon using the MiSeq data did not identify any assembly or base call errors.
Table 1: Summary of sequencing datasets. **Acc.** - Sequence read archive (SRA) database accessions. **Plat.** - sequencing platform: miseq: Illumina MiSeq, pgm: Ion Torrent PGM, pacbio: Pacific Biosciences RSII. **Lib.** - library replicate number for miseq and pgm, smartcell replicate for pacbio. **Reads** - number of sequencing reads in the dataset. **Length** - median read length in base pairs. **Insert** - median insert size in base pairs for paired end reads. **Cov.** - median sequence coverage across the genome.

<table>
<thead>
<tr>
<th>Accession</th>
<th>Plat</th>
<th>Vial</th>
<th>Lib.</th>
<th>Reads</th>
<th>Length (bp)</th>
<th>Insert (bp)</th>
<th>Cov.</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRR1979039</td>
<td>miseq</td>
<td>0</td>
<td>1</td>
<td>3305082</td>
<td>230</td>
<td>257</td>
<td>247</td>
</tr>
<tr>
<td>SRR1979040</td>
<td>miseq</td>
<td>0</td>
<td>2</td>
<td>3732088</td>
<td>216</td>
<td>233</td>
<td>263</td>
</tr>
<tr>
<td>SRR1979041</td>
<td>miseq</td>
<td>1</td>
<td>1</td>
<td>3973320</td>
<td>218</td>
<td>242</td>
<td>279</td>
</tr>
<tr>
<td>SRR1979042</td>
<td>miseq</td>
<td>1</td>
<td>2</td>
<td>3941040</td>
<td>223</td>
<td>247</td>
<td>285</td>
</tr>
<tr>
<td>SRR1979043</td>
<td>miseq</td>
<td>2</td>
<td>1</td>
<td>3442554</td>
<td>234</td>
<td>268</td>
<td>261</td>
</tr>
<tr>
<td>SRR1979070</td>
<td>miseq</td>
<td>2</td>
<td>2</td>
<td>3226726</td>
<td>232</td>
<td>268</td>
<td>240</td>
</tr>
<tr>
<td>SRR1979044</td>
<td>miseq</td>
<td>3</td>
<td>1</td>
<td>3052028</td>
<td>233</td>
<td>264</td>
<td>229</td>
</tr>
<tr>
<td>SRR1979045</td>
<td>miseq</td>
<td>3</td>
<td>2</td>
<td>4796382</td>
<td>200</td>
<td>210</td>
<td>303</td>
</tr>
<tr>
<td>SRR1979046</td>
<td>miseq</td>
<td>4</td>
<td>1</td>
<td>3338456</td>
<td>239</td>
<td>278</td>
<td>260</td>
</tr>
<tr>
<td>SRR1979047</td>
<td>miseq</td>
<td>4</td>
<td>2</td>
<td>2995090</td>
<td>237</td>
<td>277</td>
<td>231</td>
</tr>
<tr>
<td>SRR1979048</td>
<td>miseq</td>
<td>5</td>
<td>1</td>
<td>3495384</td>
<td>225</td>
<td>255</td>
<td>255</td>
</tr>
<tr>
<td>SRR1979049</td>
<td>miseq</td>
<td>5</td>
<td>2</td>
<td>3116128</td>
<td>241</td>
<td>281</td>
<td>244</td>
</tr>
<tr>
<td>SRR1979050</td>
<td>miseq</td>
<td>6</td>
<td>1</td>
<td>3129282</td>
<td>237</td>
<td>271</td>
<td>240</td>
</tr>
<tr>
<td>SRR1979060</td>
<td>miseq</td>
<td>6</td>
<td>2</td>
<td>2976312</td>
<td>242</td>
<td>280</td>
<td>233</td>
</tr>
<tr>
<td>SRR1979064</td>
<td>miseq</td>
<td>7</td>
<td>1</td>
<td>2630544</td>
<td>241</td>
<td>283</td>
<td>204</td>
</tr>
<tr>
<td>SRR1979065</td>
<td>miseq</td>
<td>7</td>
<td>2</td>
<td>3416580</td>
<td>225</td>
<td>248</td>
<td>247</td>
</tr>
<tr>
<td>SRR2002412</td>
<td>pgm</td>
<td>0</td>
<td>1</td>
<td>556903</td>
<td>231</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>SRR2002413</td>
<td>pgm</td>
<td>1</td>
<td>1</td>
<td>530117</td>
<td>224</td>
<td>38</td>
<td></td>
</tr>
<tr>
<td>SRR2002414</td>
<td>pgm</td>
<td>2</td>
<td>1</td>
<td>437527</td>
<td>231</td>
<td>33</td>
<td></td>
</tr>
<tr>
<td>SRR2002415</td>
<td>pgm</td>
<td>3</td>
<td>1</td>
<td>552692</td>
<td>232</td>
<td>42</td>
<td></td>
</tr>
<tr>
<td>SRR2002416</td>
<td>pgm</td>
<td>4</td>
<td>1</td>
<td>498479</td>
<td>232</td>
<td>37</td>
<td></td>
</tr>
<tr>
<td>SRR2002418</td>
<td>pgm</td>
<td>5</td>
<td>1</td>
<td>390070</td>
<td>235</td>
<td>30</td>
<td></td>
</tr>
<tr>
<td>SRR2002419</td>
<td>pgm</td>
<td>6</td>
<td>1</td>
<td>426196</td>
<td>232</td>
<td>32</td>
<td></td>
</tr>
<tr>
<td>SRR2002420</td>
<td>pgm</td>
<td>7</td>
<td>1</td>
<td>439119</td>
<td>238</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>SRR2056302</td>
<td>pacbio</td>
<td>9</td>
<td>1</td>
<td>163475</td>
<td>10510</td>
<td>108</td>
<td></td>
</tr>
<tr>
<td>SRR2056306</td>
<td>pacbio</td>
<td>9</td>
<td>2</td>
<td>163471</td>
<td>10436</td>
<td>103</td>
<td></td>
</tr>
<tr>
<td>SRR2056310</td>
<td>pacbio</td>
<td>9</td>
<td>3</td>
<td>163474</td>
<td>9863</td>
<td>91</td>
<td></td>
</tr>
</tbody>
</table>

3.3 Base Level Purity

A base purity metric was used to evaluate the degree to which the sequencing data supports the reference base call. We compared purity metric values between two orthogonal sequencing methods, MiSeq and PGM, for all positions in the genome (Fig. 2). Positions with purity metric values greater than 99 % were categorized as high purity and less than 99 % low purity. Out of 2909968 positions in the genome 2864925 positions had purity values greater than 99 % for both short read sequencing platforms (Table 2). Further, 2909853 and 2909965 positions had purity values greater than 99 % and 97 %, respectively, for one of the two platforms. Only 115 positions had a purity values less than 99 % for both platforms, and no positions had a purity value less than 95 % for both platforms. The positions with low purity for MiSeq were non-uniformly distributed whereas positions with low purity for PGM were uniformly distributed (Fig. 3).

Table 2: Number of genome positions with high and low purity, positions purity metric values higher and lower than 0.99 respectively, for the Illumina MiSeq and Ion Torrent PGM sequencing platforms.

<table>
<thead>
<tr>
<th></th>
<th>PGM-High</th>
<th>PGM-Low</th>
</tr>
</thead>
<tbody>
<tr>
<td>MiSeq-High</td>
<td>2864925</td>
<td>44534</td>
</tr>
<tr>
<td>MiSeq-Low</td>
<td>394</td>
<td>115</td>
</tr>
</tbody>
</table>
3.3.1 Base Level Homogeneity

The genomic material homogeneity was assessed through pairwise statistical analysis of the replicate MiSeq datasets using the VarScan somatic variant caller [5]. The pairwise variant analysis failed to identify any statistically significant base level differences among the replicates (Table 3).

3.4 Genomic Purity Pipeline

Short read sequencing data was used to identify the proportion of DNA in the material from an organism other than the material genus, in this case Staphylococcus, using PathoScope 2.0 [4]. Based on analysis of the MiSeq and PGM sequencing data, the reference material has minimal if any genomic contaminants (Fig. 4), with a maximum of 0.0039% reads in any dataset classified as not belonging to the genus Staphylococcus. The most abundant contaminant was Escherichia coli (Fig. 5).
Fig. 3 Distribution of genome positions by purity group. Bases with high and low purity and purity values greater than and less than 0.99 respectively for the two platforms, MiSeq - Illumina MiSeq and PGM - Ion Torrent PGM. Positions with high purity for both platforms were excluded from the figure.

4 Discussion

4.1 PEPR

Pipelines for Evaluating Prokaryotic References (PEPR) is designed to use biological and technical replicate sequencing data from orthogonal sequencing platforms to characterize the genome of a microbial material. There are two primary reasons for using replicate sequencing datasets one is to test for homogeneity within the batch of DNA being characterized the second is to minimize the impact of library specific biases. The resulting characterized genome is suitable for use in
<table>
<thead>
<tr>
<th>Position</th>
<th>Proportion of Pairs</th>
<th>Median Frequency</th>
<th>Minimum P-value</th>
<th>N Significant</th>
</tr>
</thead>
<tbody>
<tr>
<td>244332</td>
<td>0.01</td>
<td>21.31</td>
<td>0.51</td>
<td>0.00</td>
</tr>
<tr>
<td>261598</td>
<td>0.03</td>
<td>20.48</td>
<td>0.45</td>
<td>0.00</td>
</tr>
<tr>
<td>261605</td>
<td>0.08</td>
<td>25.29</td>
<td>0.15</td>
<td>0.00</td>
</tr>
<tr>
<td>261980</td>
<td>0.01</td>
<td>20.78</td>
<td>0.61</td>
<td>0.00</td>
</tr>
<tr>
<td>261986</td>
<td>0.01</td>
<td>21.54</td>
<td>0.50</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 3 Pairwise variant analysis results. Position is the position in the genome where differences in variant frequency for at least one of the 16 pairwise comparisons were reported. Proportion of pairs is the fraction of the pairwise comparisons between the 16 Illumina MiSeq datasets where a difference in the variant frequency was reported by VarScan. Median frequency is the median variant frequency for the datasets where a difference in variant frequency was reported. Minimum P-value is the lowest p-value reported by VarScan for all pairwise dataset comparisons where a difference in variant frequency was reported. N Significant is the number of datasets where a reported difference was statistically significant at the genome position.

Fig. 4 Proportion of reads from contaminant DNA. Reads categorized as genomic contaminants as classified by PathoScope as not belonging to the genus Staphylococcus.

evaluating and benchmarking whole genome sequencing methods. PEPR consists of three pipelines: genome evaluation, genome characterization, and genomic purity assessment.

The PEPR Genome Evaluation Pipeline provides an automated method for evaluating and refining a reference genome sequence. User’s provide a high quality reference genome as input to the pipeline. Pilon was chosen for the genome evaluation step as it not only assesses the accuracy of the genome but also corrects errors in the assembly. Other methods are available for evaluating reference genomes, e.g. amosValidate [15] and ALE [1], however these methods only assess assembly accuracy and do not correct misassemblies. The resulting reference assembly represents a consensus genome for the population of cells used to generate the material being characterized. The genome evaluation pipeline does not attempt to identify or characterize low frequency structural variants within the material or vial-to-vial variability of the
The reference genome, once validated, and if necessary refined by the Genome Evaluation Pipeline, is used as input for the Genome Characterization Pipeline.

The Genome Characterization Pipeline calculates base level statistics using replicate sequencing data from orthogonal measurement methods. Additionally, the Genome Characterization Pipeline generates summary statistics for the sequencing datasets used in the material characterization procedure as well as long read data if it was used to generate the reference assembly, if applicable. The results from the Genome Evaluation and Characterization pipeline are loaded into peprDB, a SQLite database. Loading the data into a SQLite database makes it easier to perform additional analyses of the material characterization and evaluation results.

**Fig. 5** Breakdown of contaminants by organism.
4.1.1 Process for using PEPR

If another lab wishes to characterize a new reference or quality control material, they could follow this process:

1. Ideally, the user should generate a large batch of material and aliquot it to reduce inhomogeneity.
2. Identify a high-quality genome assembly. If a good reference assembly does not exist for the sample, then long-read sequencing like PacBio may be required to generate an assembly, and ideally mapping technologies would be used to validate the assembly.
3. Short-read whole genome sequencing, preferably from two orthogonal sequencing technologies, should be generated from multiple vials of the material, ideally 2 technical replicate libraries from at least six randomly selected vials.
4. PEPR can then be run to assess base-level purity and homogeneity, genomic contaminants, and mis-assemblies. The methods presented in this work provide a straightforward pipeline that can be used by any laboratory to characterize new reference materials or in-house quality controls.

4.2 PEPR Application: Characterization of Candidate RM 8376 using PEPR

4.2.1 Preparation of Reference Assembly

PEPR requires a high quality reference assembly for input, in this work, a closed candidate reference genome was assembled from long read data generate using the Pacific Biosciences RSII platform. Optical mapping technologies and large insert mate-pair library preparation methods, including Illumina’s TruSeq Synthetic Long-Reads [9], represent orthogonal methods that can also be used to identify large misassemblies for validating reference assemblies for use in PEPR. To validate the *S. aureus* reference genome, whole genome mapping data generated with the OpGen optical mapping technology, was used to validate the genome assembly. The long DNA fragments (average size >200Mb) allow for the evaluation of large misassemblies (<3 kb) that are not easily identified using standard short read sequencing data [11]. In this work, after the initial assembly was validated for no large misassemblies, the reference assembly was characterized using PEPR.

4.2.2 Sequencing Dataset Summary

The PEPR database (peprDB) includes a number of summary statistics for the sequencing datasets processed. For sequencing datasets used to characterize the NIST candidate reference material RM8376, throughput and read lengths were as expected based on the library prep and sequencing methods used excluding the PGM datasets. For PGM the 400 bp sequencing and library preparation methods, median read lengths of 232 bp. The shorter read length is potentially due to the low GC content, which is known to challenge current sequencing technologies [16].

4.2.3 Base Level Purity

The ratio of sequencing reads supporting the reference and alternative base call was used to assess the purity of a genome position or base in the material as a whole. Through comparison of the base purity for two orthogonal sequencing methods we identified genome positions with low purity values due to platform specific systematic sequencing errors. The reference base is identified using a third orthogonal sequencing method (Pacific Biosciences RSII), which only chooses the dominant base and does not identify small impurities. Thus, a low purity (below 50 %) for one of the two short read sequencing platforms and a high purity value for the other means that two technologies (one short read and one long read) agree that the dominant base is the reference base. It is also important to acknowledge that even if the two short read sequencing platforms indicate an impurity, they are potentially susceptible to the same unknown bias. The sequencing technologies used to characterize the material are still maturing and an incomplete understanding of platform specific biases limits our ability to provide a confidence value for the base calls.
A number of base level metrics, such as strand bias, are calculated as part of the PEPR Genome Characterization Pipeline and are included in the pipeline results database. These metrics can be used to differentiate positions with low purity due to measurement error and those due to biological variability. Use of additional metrics and algorithms developed for the identification of low frequency variants, such as LoFreq [24], could help identify positions with low levels of biological variability, but are not currently implemented in PEPR.

The PEPR Genome Characterization Pipeline assesses the homogeneity of a material through comparison of purity values between replicate vials. The homogeneity analysis is performed using the VarScan tumor-normal variant caller [5]. Only Illumina data was used to assess the homogeneity of the material as the higher coverage increased the statistical power of the test, and replicate libraries provide information regarding the method error rate. The Ion Torrent dataset did not include replicate libraries for the eight vials sequenced and therefore library specific sequencing errors were confounded with vial to vial variability. No statistically significant variants were identified between all pairwise comparisons indicating that the material is homogeneous. If potential inhomogeneities were found, then the PGM sequencing data could be examined for additional support for the inhomogeneities. Even without replicate libraries for the different vials it is unlikely that any library specific bias will correlate with vial-to-vial variability observed in the Illumina data by chance.

4.3 Genomic Purity Pipeline

The Genomic Purity Pipeline is used to identify DNA within the material that belongs to a genus other than the material genus. The genus level cutoff was selected based on results from a previous study characterizing the specificity of the PathoScope 2.0 classification algorithm (Olson et al. in-prep). Genomic contaminants can be from the culture itself or reagents and materials used to prepare the material or during sequencing [19, 20, 18]. Contaminants identified by the Genomic Purity Pipeline may not be present in the material. For example reagents used during library preparation may also include contaminants [21, 13, 12, 18]. Additionally, bioinformatic errors may lead to false positive contaminants, either due to errors in the database or bioinformatic errors by classification algorithm. Genomic purity analysis of the S. aureus material identified a number of candidate contaminants, the most abundant of which was E. coli. E. coli is a well documented contaminant of molecular biology reagents, and not likely a true contaminant [18]. Lower abundant contaminants maybe bioinformatic errors and not true contaminants. While, contaminants identified by the Genomic Purity Pipeline are most likely from reagents and due to bioinformatic errors, a conservative estimate of the material purity, assuming all contaminants are real, reported by the pipeline. Users will want to take into consideration the limited specificity of the taxonomic classification method. For example if the intended use of the genomic DNA is for use as part of an inclusivity exclusivity pannel, additional genomic purity assessment in addition to the PEPR Genomic Purity Pipeline is required to validate the material.

5 Conclusions

PEPR provides a framework for characterizing microbial genomic reference materials, for instance a homogenized batch of DNA from a single prokaryotic strain. The objective in developing PEPR was to provide a reproducible and transparent workflow for the characterization of prokaryotic genomic materials. The pipeline can be used to characterize reference materials as well as in-house quality control materials for which replicate sequencing datasets from multiple platforms are available. The outputs from PEPR include a corrected reference genome assembly, identified genome position with high and low purity based on biological and technical variation, base level homogeneity of the material, as well as the percentage and identity of genus level genomic contaminants. The resulting characterization values are intentionally conservative and without uncertainty or confidence estimates, as sources of bias and error associated with the measurement process are currently not fully understood. As the scientific community’s understanding of the measurement process matures new algorithms can be incorporated into the pipeline to increase the quality of material characterization process. The genomic materials characterized using PEPR will not only help increase confidence in measurement methods and
assays, but can be used to benchmark and validate sequencing platforms and bioinformatic algorithm improving our understanding of sequencing and data analysis process.

Acknowledgements The authors would like to thank Jenny McDaniel, Lindsay Vang, and David Catoe for performing the MiSeq and PGM sequencing, and Tim Muruvanda for performing the PacBio sequencing. Additionally, we would like to thank Drs. Nancy Lin and Steven Lund for conversations and suggestions while developing PEPR and drafting the manuscript. This work was supported by the Department of Homeland Security (DHS) Science and Technology Directorate under the Interagency Agreement HSHQPM-14-X-00078 with NIST and by two interagency agreements with the FDA. Opinions expressed in this paper are the authors and do not necessarily reflect the policies and views of the DHS, NIST, or affiliated venues. Certain commercial equipment, instruments, or materials are identified in this paper only to specify the experimental procedure adequately. Such identification is not intended to imply recommendation or endorsement by the NIST, nor is it intended to imply that the materials or equipment identified are necessarily the best available for the purpose. Official contribution of NIST; not subject to copyrights in USA.

References


