Title

POLISHER: a Tool for Using Ultra Short Reads in Microbial Genome Finishing

Authors
LaButti, Kurt
Foster, Brian
Lowry, Steve
et al.

Publication Date
2008-05-26

POLISHER: a tool for using ultra short reads in microbial genome finishing

Kurt LaButti[1], Brian Foster[1], Steve Lowry[1], Stephan Trong[2], Eugene Goltsman[1], and Alla Lapidus[1]
[1]Lawrence Berkeley National Laboratory
[2] Lawrence Livermore National Laboratory
DOE Joint Genome Institute, Walnut Creek, CA 94598

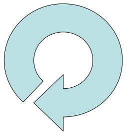
STANDARD MICROBIAL FINISHING STRATEGIES

The current strategy for finishing microbial genomes consists of repeat resolution, gap closure, and polishing. Since polishing is the most time consuming and costly stage in finishing it was targeted as an area for immediate improvement.

During the polishing phase the quality of the assembly is brought up to a predefined standard;

•all bases >= Q30
•2x coverage
•<5% 454 only

This is typically done through a cyclical process;



1. Substandard regions of the consensus tagged in ace (Figure 1)
   Low quality consensus
   <Q30Single subclone
   454 only
2. Automated oligo/plasmid template picking and ordering
   Attempts to design oligos around clustered tags
3. Sequencing
4. Data incorporation
   Requires manual assessment and intervention
   Not all are solved
5. Repeat process
   Expect 1/3 of the reactions to fail
   Typically need ~ 4 rounds

MICROBIAL FINISHING WITH THE POLISHER

In order to reduce cost, time, and increase capacity all while upholding our current finishing standard, we developed a tool that employs Illumina read data to polish substandard regions as well as fix consensus errors in our microbial projects. The tool now exists as a functional prototype that works in several phases: alignment, analysis, and polishing.

Align

The read data is aligned to a lookup table of the assembly fasta sequence using Arachne's MakeLookupTable and QueryLookupTable with the following options:

MO=10 K=12 SMITH_WAT=True MAX_ERROR_PERCENT=25 WE=10 MC=0.01

Since we are aligning to unpolished draft-like fasta we found QueryLookupTable to be the most suitable aligner at the time because of its speed and ability to align reads with a large amount of discrepancies. An alignment for each flow cell is sent off in parallel and simultaneously parsed for best hit based on percent identity. Equal scores are placed at random.

Analyze

The best hit information is parsed for Illumina coverage per consensus base. Every discrepancy (mismatch, deletion, insertion) is also tracked and this information is stored in a data structure (Figure 2.2). It then traverses the data structure and refines the information by calculating the fraction that agrees with the consensus base, and the largest fraction that disagrees. While traversing the refined data structure it looks for areas where the Illumina data suggests something is positively wrong and needs editing. These areas are kept in a list called AcefileEdits.list. An invitation to this list requires the following thresholds;

>= 10X Illumina coverage
70% of the Illumina coverage (majority discrepancy) disagrees with the consensus base

Polish

Mismatches identified in the previous step are fixed via modification of the acefile consensus base and the quality is bumped up to Q99 so they will be ignored by subsequent polishing. Our normal substandard region identification tool (tagAceforPolishing) is then run to generate a list of polishing tags (polishingTags.list). These polishing tags specify the location and type (LowQualConsensus, SingleSubclone, 454Only) of every substandard region in the acefile. Each base of every polishing tag is then interrogated to see if the Illumina data suggests it is correct or not with the following thresholds;

>= 10X Illumina coverage
70% of the Illumina coverage agrees with the consensus base

If any base in a polishing tag meets the above criteria then the polishing tag over that base is changed to solexaSupported. If the information for the base does not meet the criteria then the original tag remains and will have to be polished using traditional manual methods. The resulting modified tags are then added to the acefile and deletions suggested in the AcefileEdits.list are fixed via modification of the acefile (Figure 1).

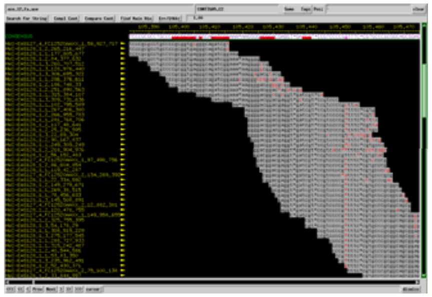


Figure 2. Mosaik visualization

Consed view of Illumina read data aligned to the draft consensus from Figure 1. Acefile and alignment were created with Mosaik for visual verification of the Polisher's performance. Note the agreement in read data below the polishing tags as well as under the extra T in the consensus (pink tag).

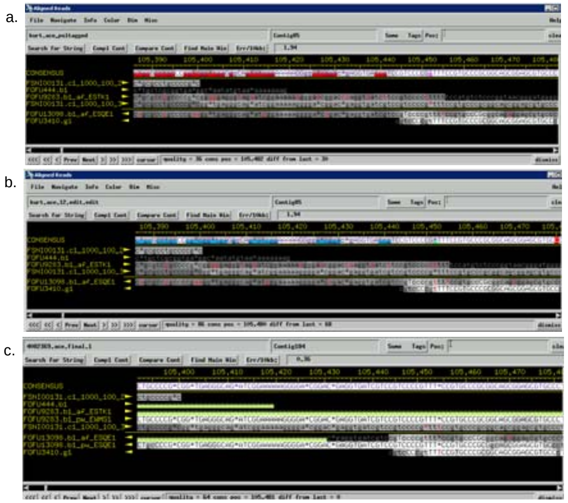
ABSTRACT

Polishing is one of the major steps of genome finishing at the JGI (Joint Genome Institute). Along with repeat resolution and gap closure, it is required to produce a fully sequenced high quality genome. Polishing consists of consensus error correction and quality improvement such that the resulting consensus meets a pre-defined standard. This has traditionally been done through targeted Sanger clone based sequencing, making it the most time consuming and resource intensive stage in finishing. Our pilot experiments, conducted using Illumina data produced by the JGI for several microbes, demonstrated that aligning ultra short reads against unpolished contigs help correct a significant amount of consensus errors and greatly reduce the amount of quality improvement necessary to produce a finished genome. A prototype tool named the "Polisher" was developed in order to automate this process. It facilitates polishing and error correction of a subject assembly (acefile), typically a draft or closed assembly, using Illumina read data.

The Polisher corrects consensus errors and supports correctly called bases that would normally be targeted for polishing due to their below standard quality by analyzing Illumina alignment data and automatically modifying the consensus. The Illumina read data in Fasta format is aligned to the subject sequence and simultaneously parsed for the best hit based on percent identity. The best hit alignment results are then used to determine coverage and discrepancy information per base in the subject. A list of errors is generated where there is overwhelming evidence that the consensus base is wrong and needs to be changed. Such areas represent mismatches, deletions, and insertion. Mismatches and deletions are automatically corrected in the acefile while insertions, for now, remain as tags for manual inspection. In addition to the previously described error correction, other areas of the genome that would normally be targeted for polishing are inspected. These areas currently represent low quality, single subclone, and 454 only regions and exist as tags in the acefile. If there is overwhelming evidence that a particular base targeted for traditional polishing is correct, then that base is termed Supported and retagged. If there is not enough evidence for support, then the original polishing tag remains for traditional manual polishing.

Figure 1. Draft consensus: pre and post polishing

(a) Portion of a draft assembly project as viewed in the assembly editor Consed. The consensus is composed of a single 454 shred and several low quality sanger reads. The consensus regions that are <Q30 have been tagged with LowQualityConsensus tags (red) in the first panel. **(b)** The second panel shows the same region after polishing with the Polisher. The Illumina data aligned to this region suggests with a high amount of certainty that the consensus bases that were tagged for polishing were correct and therefore retagged as solexaSupported (blue). Also note the green solexaCorrected tag where the polisher deleted a base. **(c)** The third panel shows the same region finished using traditional methods (sanger sequencing off of plasmid templates). It required two rounds of polishing as well as manual editing.



PERFORMANCE

The Polisher was tested on draft assemblies from 9 previously finished genomes of varying complexity and GC content to roughly gauge its performance (Figure 3). The earliest draft assembly was used for each organism because all subsequent versions contained both gap closing and polishing work incorporated. Our automated oligo picking software was run on the assembly before and after running the Polisher and the number of oligos required to polish by traditional means were compared. The results varied based on the overall status of the draft assembly and how much Illumina data available.

A second method was used to further understand how well the Polisher's performance in greater detail. In this case the reference sequence of 3 recently finished genomes was used to create modified data sets with known mutations. About 1000 random single base mutations (base change, base insertion, base deletion) were introduced into each reference fasta and the location and type were recorded. The Polisher was then run up until the Analyze function in order to create the data structure and AcefileEdits.list. The edits suggested in the AcefileEdits.list file were then compared to the known mutations and statistics were generated from the results using a confusion matrix.

a is the number of **correct** predictions that an instance is **negative**.
b is the number of **incorrect** predictions that an instance is **positive**.
c is the number of **incorrect** of predictions that an instance **negative**.
d is the number of **correct** predictions that an instance is **positive**.

		Predicted	
		-	+
Actual	-	**a**	**b**
	+	**c**	**d**

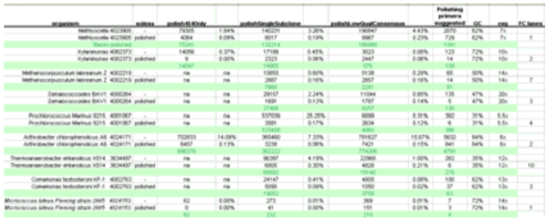
RESULTS

Employing the Polisher resulted in a dramatic reduction in the overall number of oligos the projects required for polishing (figure 3). In addition to the reduction in oligos a large amount of time was also saved since the Polisher takes on average 3 hours to run from start to finish. Comparatively each round of traditional polishing takes at least one week. This translates into an estimated 98.5% savings on traditional polishing reactions (average 81%).

The results from the confusion matrix calculations suggest that the Polisher's is better at handling mismatches than indels. Some possible reasons for this could be alignment error both in aligning the Illumina data and in aligning the polished sequence back to the reference. In addition differences in the sample DNA used to prepare the Illumina data can also result in false positives, and these can typically reflect real polymorphism in the data sets. Further investigation is required to ascertain exactly the performance differs so greatly between the mutation types.

Figure 3. Draft assembly polishing results

Table detailing polishing statistics for pre and post polished draft assemblies. The highlighted green regions specify the difference for each pre and post polished statistic. Note the reduction in oligos necessary to polish in the ninth column.



organism	solexa	polish454Only		polishSingleSubclone		polishLowQualConsensus		Polishing primers suggested	GC	cvg	FC lanes
Methylocella 4023905	-	79305	1.84%	140231	3.26%	190847	4.43%	2070	62%	7x	
Methylocella 4023905	polished	4064	0.09%	8017	0.19%	9967	0.23%	729	62%	7x	1
Bases polished		75241		132214		180880		1341			
Xylanimonas 4062373	-	14056	0.37%	17188	0.45%	3023	0.08%	123	72%	10x	
Xylanimonas 4062373	polished	9	0.00%	2323	0.06%	2447	0.06%	14	72%	10x	2
		14047		14865		576		109			
Methanocorpusculum labreanum Z 4002219	-	na	na	10855	0.60%	5138	0.29%	65	50%	14x	
Methanocorpusculum labreanum Z 4002219	polished	na	na	2887	0.16%	2857	0.16%	14	50%	14x	7
				7968		2281		51			
Dehalococcoides BAV1 4000264	-	na	na	29157	2.24%	11044	0.85%	135	47%	20x	
Dehalococcoides BAV1 4000264	polished	na	na	1691	0.13%	1787	0.14%	5	47%	20x	3
				27466		9257		130			
Prochlorococcus Marinus 9215 4001067	-	na	na	537039	25.25%	6699	0.31%	392	31%	5.5x	
Prochlorococcus Marinus 9215 4001067	polished	na	na	3581	0.17%	2634	0.12%	6	31%	5.5x	4
				533458		4065		386			
Arthrobacter chlorophenolicus A6 4024171	-	702833	14.09%	365460	7.33%	781627	15.67%	5632	64%	6x	
Arthrobacter chlorophenolicus A6 4024171	polished	6457	0.13%	3238	0.06%	7421	0.15%	841	64%	6x	2
		696376		362222		774206		4791			
Thermoanaerobacter ethanolicus X514 3634497	-	na	na	96397	4.19%	22968	1.00%	282	35%	12x	
Thermoanaerobacter ethanolicus X514 3634497	polished	na	na	6805	0.30%	4828	0.21%	6	35%	12x	10
				89592		18140		276			
Comamonas testosteroni KF-1 4002763	-	na	na	24147	0.41%	4805	0.08%	100	62%	13x	
Comamonas testosteroni KF-1 4002763	polished	na	na	5095	0.09%	1050	0.02%	37	62%	13x	3
				19052		3755		63			
Micrococcus luteus Fleming strain 2665 4024153	-	62	0.00%	273	0.01%	369	0.01%	7	72%	14x	
Micrococcus luteus Fleming strain 2665 4024153	polished	0	0.00%	41	0.00%	151	0.01%	3	72%	14x	1
		62		232		218		4			

Figure 4. Confusion matrix results

The following tables lists correct and incorrect predictions used to measure the performance of the Polisher. **(a)** The number of **False edits** (shouldn't have been edited but was) **Correct edits** (should have and was), **Missed edits** (should have been but wasn't), and **Non edited** (shouldn't have and wasn't) for each mutation type for each genome are detailed. **(b)** The averages of each mutation type were used to populate the confusion matrices. **(c)** Overall performance statistics were calculated for each mutation type using the confusion matrix for each mutation type.

a.

Brachybacterium faecium DSM 04810

	Mismatches	Deletions	Insertions
False edits	21	13	2
Correct edits	312	241	228
Missed edits	23	105	101
Non edited	3614637	3614643	3614626

Cryptobacterium curtum DSM 15641

	Mismatches	Deletions	Insertions
False edits	0	3	2
Correct edits	350	290	292
Missed edits	15	26	31
Non edited	1617446	1617495	1617488

Sanguibacter keddieii DSM 10542

	Mismatches	Deletions	Insertions
False edits	6	0	2
Correct edits	299	228	193
Missed edits	24	125	131
Non edited	4253061	4253031	4253060

b.

Mismatches

a	3161715	9	b
c	21	320	d

Deletions

a	3161725	2	b
c	88	238	d

Insertions

a	3461723	5	b
c	85	253	d

c.

	Mismatches	Insertions	Deletions
Accuracy	0.99999062	0.99997164	0.99997133
True positive (sensitivity)	0.93939394	0.73053279	0.74778325
False positive	0.00000285	0.00000063	0.00000169
True negative (specificity)	0.99999715	0.99999937	0.99999831
False negative	0.06060606	0.2694672	0.25221675
Precision	0.97267206	0.99165508	0.97935484

CONCLUSIONS

With the push to reduce or all together eliminate sanger data in an effort to reduce project costs in the future, the cost of finishing and polishing such projects will likely rise. A method will be needed to polish a genome in the absence of sanger templates while keeping cost in mind. Here we have presented such the Polisher which employs the use of Illumina ultra short read data to polish and correct assembly consensus. Experimental results using real and simulated data suggest by deploying the Polisher in its current state in microbial finishing we stand to reduce our finishing and overall project costs by a significant amount. As it currently functions there is room for improvement in detecting errors, however the False positive rate is relatively low and since each genome is inspected for completion at the end any small amount of errors introduced by the Polisher should be identified and fixed.

•Estimated 98.55 savings on traditional polishing reactions
•Average saving in finishing: ~25% per genome

Further experimentation on real data sets will hopefully allow us to enhance the thresholds and increase the sensitivity. In the future we also plan on integrating the Polisher into the JGI Alignment Services system. This will allow correction of insertions as well as streamline data management and speed.

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Berkeley National Laboratory under contract No. DE-AC02-05CH11231, Lawrence Livermore National Laboratory under Contract No. DE-AC52-07NA27344, and Los Alamos National Laboratory under contract No. DE-AC02-06NA25396.