Windshield splatter analysis with the Galaxy metagenomic pipeline

SERGEI KOSAKOVSKY POND[1,2,*], SAMIR WADHAWAN[3,6*], FRANCESCA CHIAROMONTE[4], GURUPRASAD ANANDA[1,3], WEN-YU CHUNG[1,3,7], JAMES TAYLOR[1,5], ANTON NEKRUTENKO[1,3] and THE GALAXY TEAM[1]

[*] These authors contributed equally to this work
[1]http://galaxyproject.org
[2]Division of Infectious Diseases, Division of Biomedical Informatics, School of Medicine University of California San Diego
[3]Huck Institute for the Life Sciences, Penn State University
[4]Department of Statistics, Penn State University
[5]Department of Biology and Mathematics & Computer Science, Emory University
Present address: [6]Department of Genetics, University of Pennsylvania Medical School
Present address: [7]Cold Spring Harbor Laboratory

A live version of this supplement with full access to analyses and workflows can be accessed via this URL:

http://usegalaxy.org/u/aun1/p/windshield-splatter

Supplemental Materials and Figures

Comparison between Galaxy pipeline and Megan

The first step of a homology-based metagenomic analysis is to contrast a collection of sequencing reads against a database whose entries are assigned to taxonomic ranks. Following the procedure of (Huson et al. 2007) we used the non-redundant protein database (NR) from the National Center for Biotechnology Information (NCBI, http://www.ncbi.nlm.nih.gov). There are several avenues for importing large sets of alignments into Galaxy. First, alignments can be generated directly within Galaxy (see the following section). Alternatively, alignments generated elsewhere (e.g., using local BLAST installations of web-based resources such as CAMERA (Seshadri et al. 2007); see below) can be uploaded in either tab-delimited or XML format. To demonstrate this functionality, we generated alignments in BLAST XML format outside of Galaxy using the BLASTx program of the BLAST package (Altschul et al. 1990) and then uploaded them into Galaxy's history. Galaxy includes a parser for XML generated by BLAST programs that produces a tab-delimited format that can be easily used in downstream analyses. Only 243 (or ~2% from 3,812,372 alignments) and 1,192 (or ~11% from 3,581,932 alignments) reads from samples 1 and 2-4, respectively, did not produce matches against the NR database. These counts were slightly higher than those reported in Huson et al. because we set the BLAST E value flag (-e) to 0.01 instead of the default value of 10 (used in (Huson et al. 2007)) removing many weakly supported alignments and significantly decreasing the size of the resultant file. Similarly to Huson and colleagues we further filtered BLAST alignments by retaining only those hits that were within 5% of the best score for every read using a combination of Galaxy tools. This significantly reduced number of hits to 54,458 and 62,647 in samples 1 and 2-4, respectively, although the number of reads producing these hits did not change (9,757 and 8,808 reads, respectively).

Because every entry within the NR database is assigned a taxonomy id, it is straightforward to create a phylogenetic profile of every read that aligns against a database sequence. Galaxy features the Fetch Taxonomic Ranks tool that quickly parses NCBI taxonomy and writes out a taxonomic string consisting of 21 taxonomic ranks from superkingdom to subspecies. Application of this tool to filtered BLAST hits produced 54,458 and 62,647 taxonomic strings for samples 1 and 2-4, respectively. Note that because the numbers of taxonomic strings greatly surpass the numbers of sequencing reads (9,757 and 8,808, respectively), each read is likely represented by multiple phylogenetic profiles. As a result all reads can be divided into two categories: diagnostic and non-specific. A diagnostic read consistently hits database sequences belonging to the same taxonomic group, while its non-specific counterpart identifies with multiple taxa. (An extreme example of a non-specific read will produce alignments with both eukaryotic and prokaryotic sequences and as a result will be useless for phylogenetic profiling of metagenomic samples). Furthermore, as biological classification is hierarchical, a read can be diagnostic at one level and non-specific at another: if a given read produces alignments with multiple database sequences yet all these sequences belong to the same genus, we consider such read diagnostic for that genus. It is easy to envision a situation when a read diagnostic to a genus will hit multiple species within a genus. For instance, a read producing 10 alignments all within the genus Drosophila may, at the species level, align with sequences from D. melanogaster and D. ananassae. Thus such a read is diagnostic at the genus level but non-specific at the species level. In addition, even when a read represents species A, it will likely also produce alignments with a closely related species B and therefore will appear non-specific at the species level. There are two ways to address this situation. First, one can tabulate a list of reads diagnostic at a predefined taxonomic level. In Galaxy this is achieved with the "Find diagnostic hits" tool (Table 1) within which the user specifies desired taxonomic ranks and the tool returns reads diagnostic for such ranks. Alternatively, one can traverse the taxonomic strings of every read by identifying and removing reads with more than one taxonomic label (see explanation of the tool's algorithm at the Galaxy web site under "Metagenomic Tools" - "Find lowest taxonomic rank"). This approach is conceptually identical to the Lowest Common Ancestor (LCA) algorithm of (Huson et al. 2007) and is implemented in Find lowest diagnostic rank tool. We used this tool here to directly compare our implementation to results produced by MEGAN software. For samples 1 and 2-4 we identified 9,380 and 7,847s reads that we were diagnostic below the Kingdom level. These numbers are slightly higher than those reported by Huson et al. One reason for this is the fact that these used a version of the NR database that is roughly two years older than the one used by us in this study. Finally, we visualized results of our analysis using the Draw phylogeny tool that renders phylogenetic trees using taxonomy datasets as input. Figure 1 shows the portion of the tree for Gammaproteobacteria in the two samples. The resulting topology and read numbers are nearly identical to those produced by MEGAN with this dataset (see Figures 3C and 3D in (Huson et al. 2007)) suggesting that our approach works correctly.

Supplementary Table 1. Gamma-proteobacterial genera

Genus	# reads		B/A ratio
	Trip A	Trip B	
Acinetobacter	97	15	0.155
Aeromonas	539	21	0.039
Alcanivorax	11	1	0.091
Aliivibrio	1	1	1.000
Azotobacter	16	1	0.063
Buchnera	9	57	6.333
Candidatus	1	1	1.000
Citrobacter	668	212	0.317
Cronobacter	43	22	0.512
Dickeya	4	1	0.250
Enterobacter	4,142	5,507	1.330
Enterovibrio	3	1	0.333
Erwinia	2	240	120.000
Escherichia	811	299	0.369
Francisella	1	1	1.000
Haemophilus	3	1	0.333
Halomonas	10	4	0.400
Klebsiella	15,121	1,695	0.112
Kluyvera	14	1	0.071
Marinobacter	3	4	1.333
Pantoea	32	14	0.438
Pectobacterium	122	59	0.484
Photorhabdus	57	1	0.018
Proteus	26	1	0.038

Genus	# reads		B/A ratio
	Trip A	Trip B	
Providencia	122	3	0.025
Pseudomonas	1,616	383	0.237
Psychrobacter	3	2	0.667
Psychromonas	1	1	1.000
Raoultella	12	7	0.583
Salmonella	4,023	1,859	0.462
Serratia	3,239	29	0.009
Shewanella	29	6	0.207
Shigella	674	376	0.558
Sodalis	127	40	0.315
Stenotrophomonas	92	9	0.098
Vibrio	44	64	1.455
Wigglesworthia	1	1	1.000
Xanthomonas	20	15	0.750
Yersinia	1,257	196	0.156

Supplemental Figures

Supplementary Figure 1. Genus-level phylogenetic profile of class Gammaproteobacteria reconstructed from protein-level comparisons. The color of the branches represents the relative abundance of sequencing reads representing that branch (red = more; blue = less). Numbers within each box signify the number of sequencing reads associated with a given taxon. Branches without labels identify reads that do not identify with any ranks above genus level (in this case unidentified uncultured gammaproteobacterium)

Supplementary Figure 2. Analysis of Sargasso Sea metagenomic reads from Samples 1 (A) and 2-4 (B) as described in (Huson et al. 2007). Read length = distribution or read lengths. Alignment length = distribution of lengths of megaBLAST hits produced by aligning the reads against NT and WGS databases. Alignable fraction = distribution of proportion of each read's length covered by megaBLAST hits against NT and WGS databases. Q1, Q2, and Q3 = first, second (median), and third quartiles.

Supplementary Figure 3. Genus-level phylogenetic profile of class Gammaproteobacteria reconstructed from nucleotide-level comparisons. The color of the branches represents the relative abundance of sequencing reads representing that branch (red = more; blue = less). Numbers within each box signify the number of sequencing reads associated with a given taxon. Branches without labels identify reads that do not identify with any ranks above genus level (in this case unidentified uncultured gammaproteobacterium).

Supplementary Figure 4. Analysis of 454 read quality for trip A (A) and B (B). The distribution of base quality scores (in phred metric) for all sequencing reads in the experiment. To produce this image each read was divided into 20 equal sized segments and quality scores for all bases falling within each segment were averaged. These average quality values from all read were then used to produce the box plot.

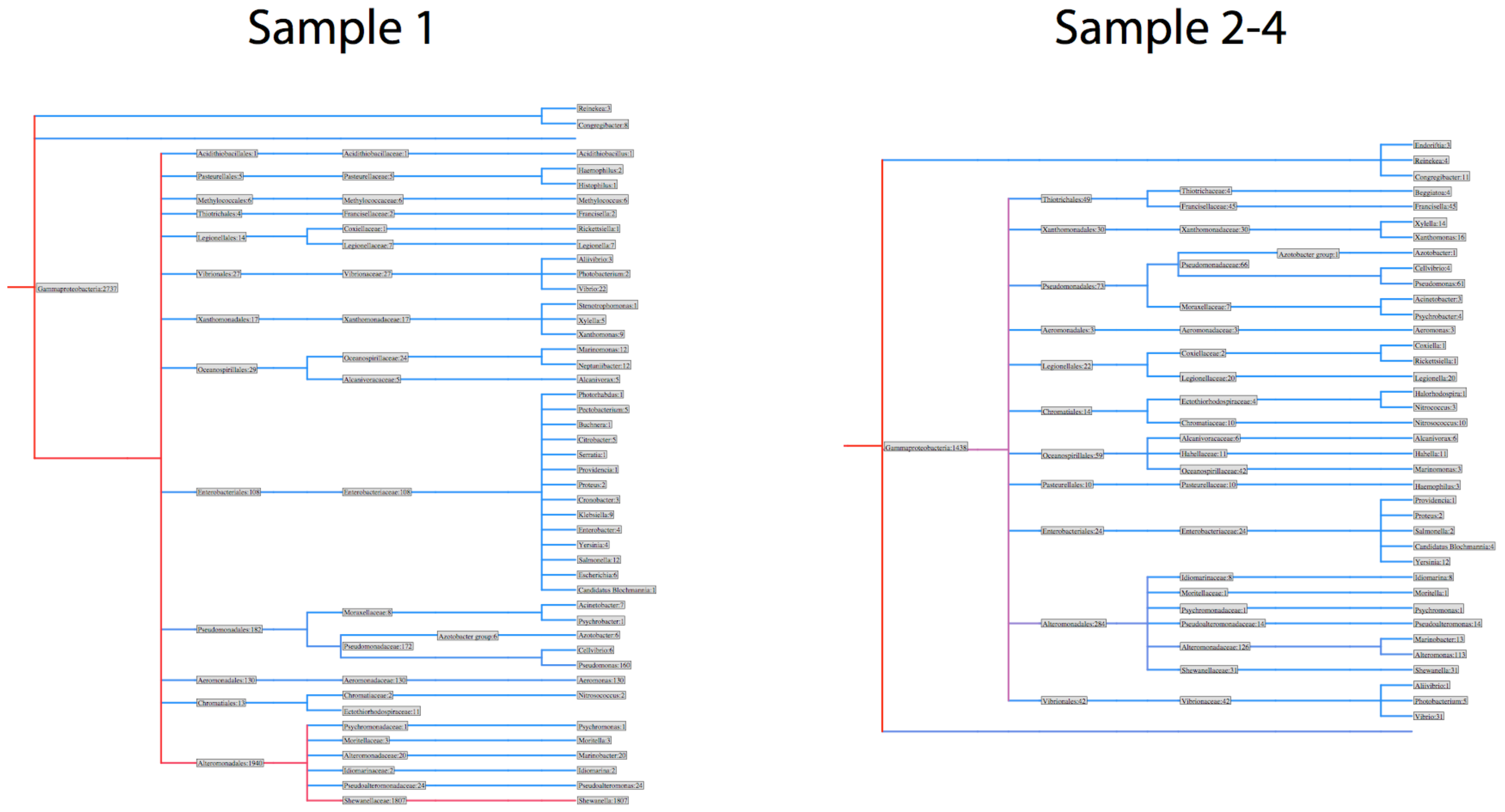
Supplementary Figure 5. Analysis of read fragmentation by low quality bases. (A). Length distribution of fragments generated by splitting the reads on any base with quality score < 20 (phred metric). (B). Length distribution of fragments generated by splitting the reads on bases with quality score < 20 that are NOT in the proximity of homopolymer runs.

Supplementary Figure 6. Distribution of alignment length, alignment identity, and alignable fraction for 454 reads for trip A (A) and B (B). Q1, Q2, and Q3 = first, second (median), and third quartiles.

Supplementary Figure 7. Genus-level phylogenetic profile of class Gammaproteobacteria obtained by comparing trip A (A) and trip B (B) reads against NT and WGS databases.

Supplementary Figure 8. Example of using Galaxy to process alignment results generated within CAMERA system. Using the "Export" drop down of CAMERA interface we downloaded results in BLAST XML format (A). These data are them uploaded into Galaxy and processed using its XML-parser (B). Errors (red text in A) were resulting from some reads being too short to be used for computing the Altschul-Karlin statistics used in megaBLAST.

Figure S1



Sample 1

Sample 2-4

Figure S2

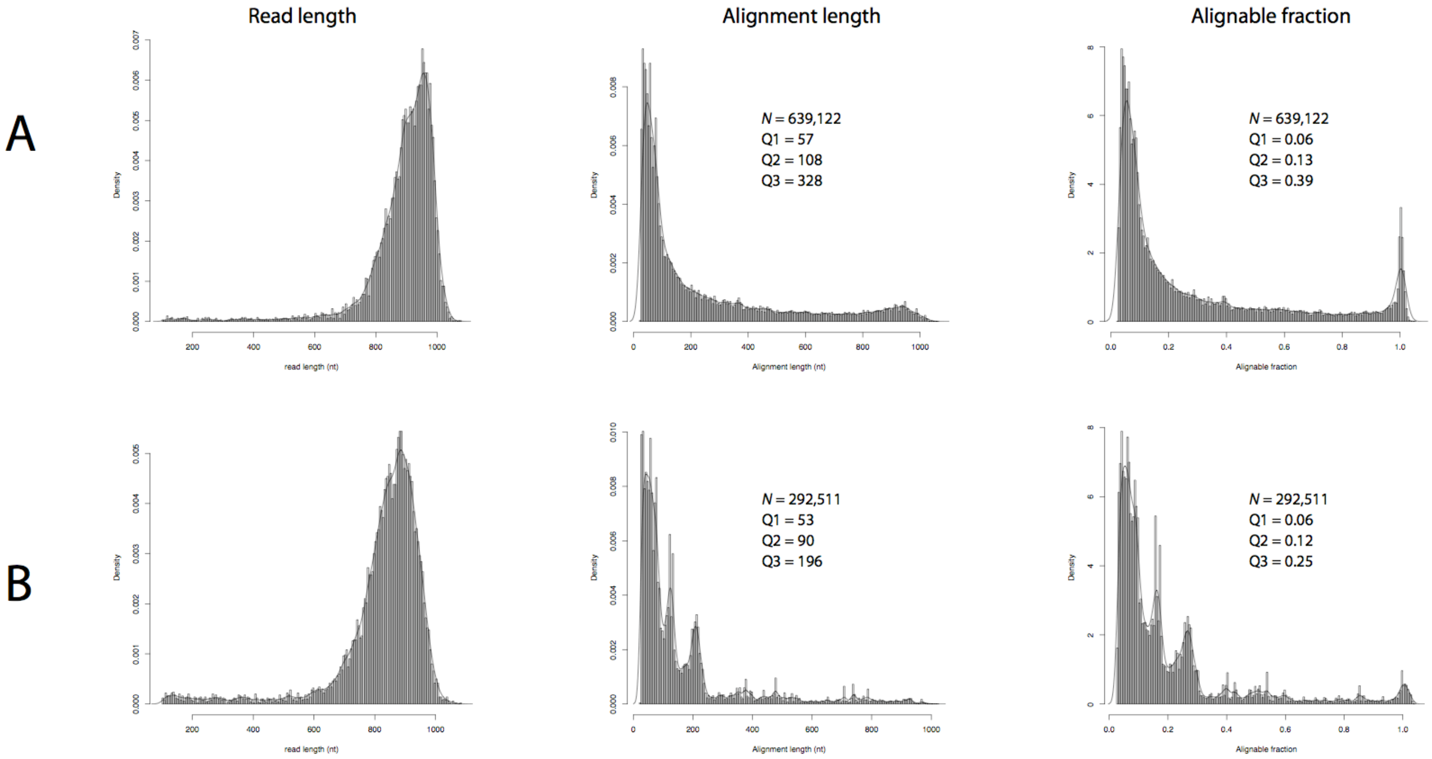
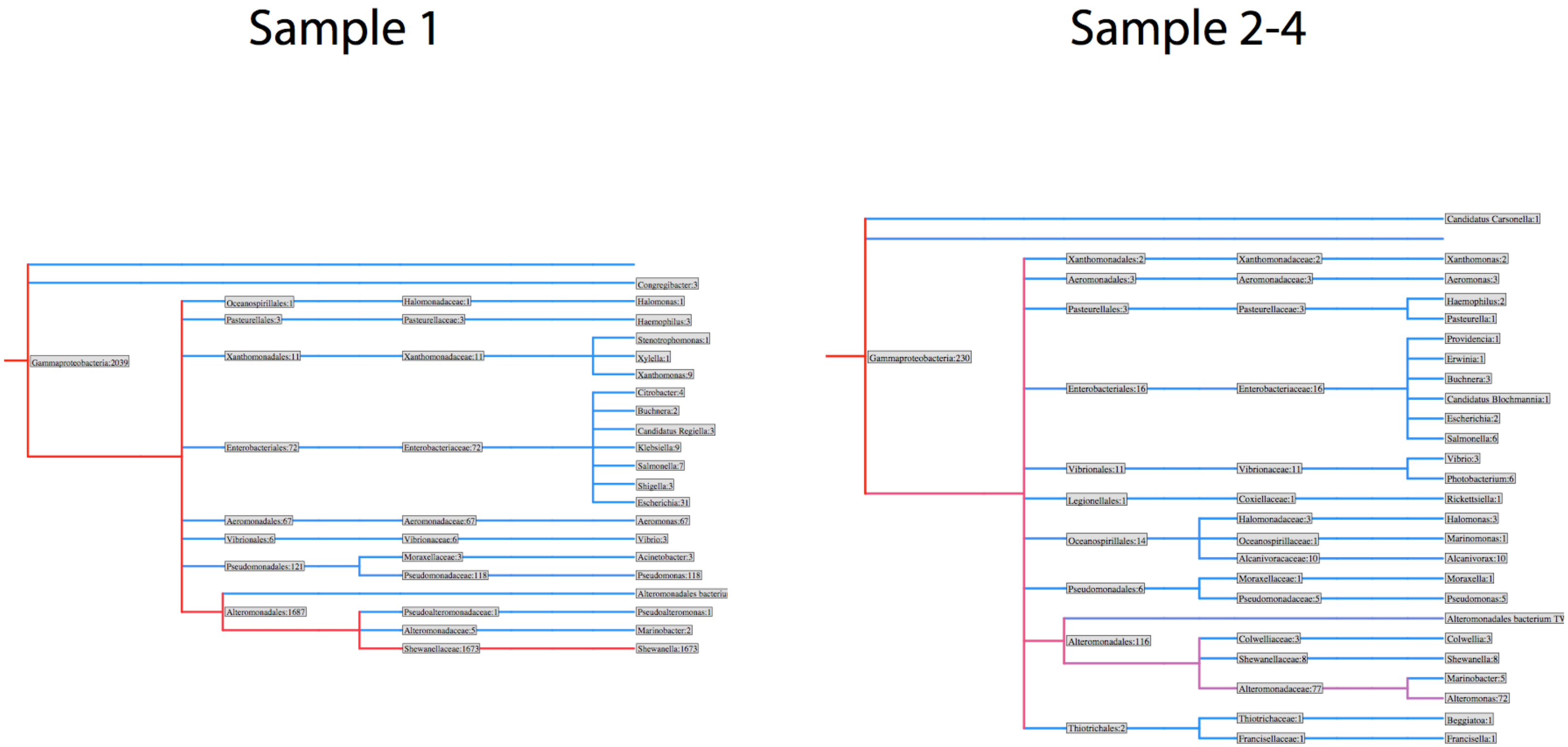

Read length
Alignment length
Alignable fraction
A
B
Density
read length (nt)
200
400
600
800
1000
0.000
0.001
0.002
0.003
0.004
0.005
0.006
0.007
Alignment length (nt)
0
0.008
0.010
Alignable fraction
0.0
0.2
0.4
0.6
0.8
1.0
2
4
6
8
N = 639,122
Q1 = 57
Q2 = 108
Q3 = 328
N = 639,122
Q1 = 0.06
Q2 = 0.13
Q3 = 0.39
N = 292,511
Q1 = 53
Q2 = 90
Q3 = 196
N = 292,511
Q1 = 0.06
Q2 = 0.12
Q3 = 0.25

Figure S3


Sample 1

Sample 2-4

Figure S4

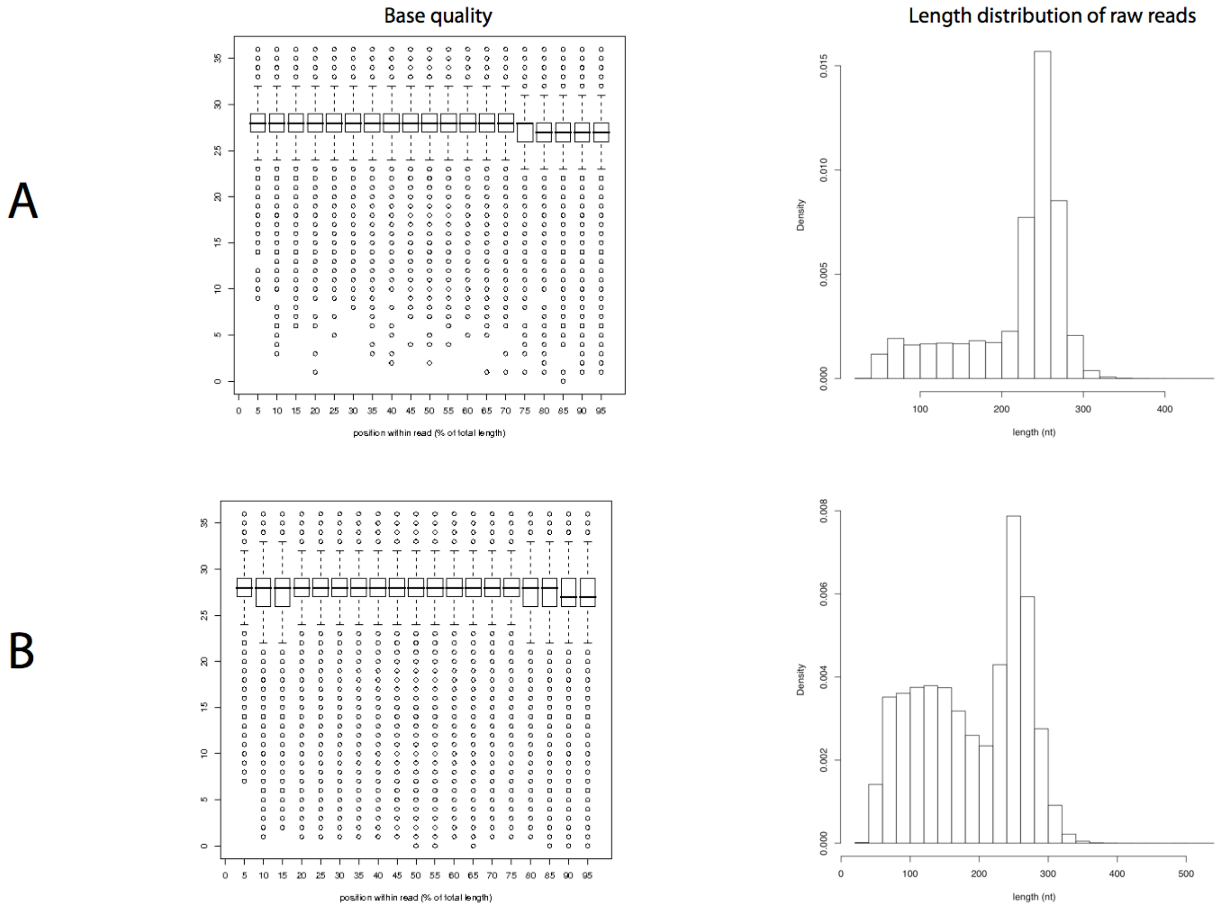

Base quality
Length distribution of raw reads
A
B
position within read (% of total length)
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
Density
0.000
0.005
0.010
0.015
0.002
0.004
0.006
0.008
length (nt)
100
200
300
400
500

Figure S5

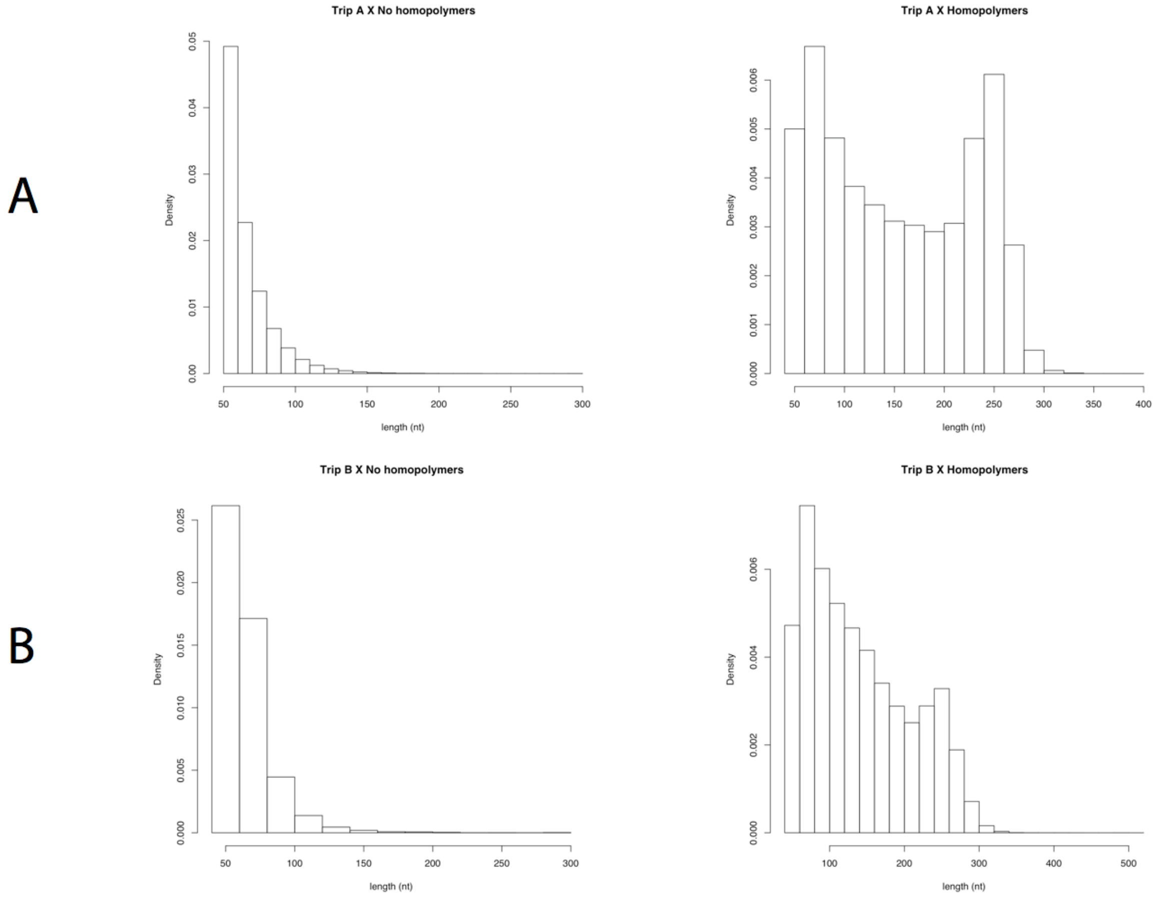

A
B
Trip A X No homopolymers
Density
0.00
0.01
0.02
0.03
0.04
0.05
50
100
150
200
250
300
length (nt)
Trip A X Homopolymers
Density
0.000
0.001
0.002
0.003
0.004
0.005
0.006
50
100
150
200
250
300
350
400
length (nt)
Trip B X No homopolymers
Density
0.000
0.005
0.010
0.015
0.020
0.025
50
100
150
200
250
300
length (nt)
Trip B X Homopolymers
Density
0.000
0.002
0.004
0.006
100
200
300
400
500
length (nt)

Figure S6

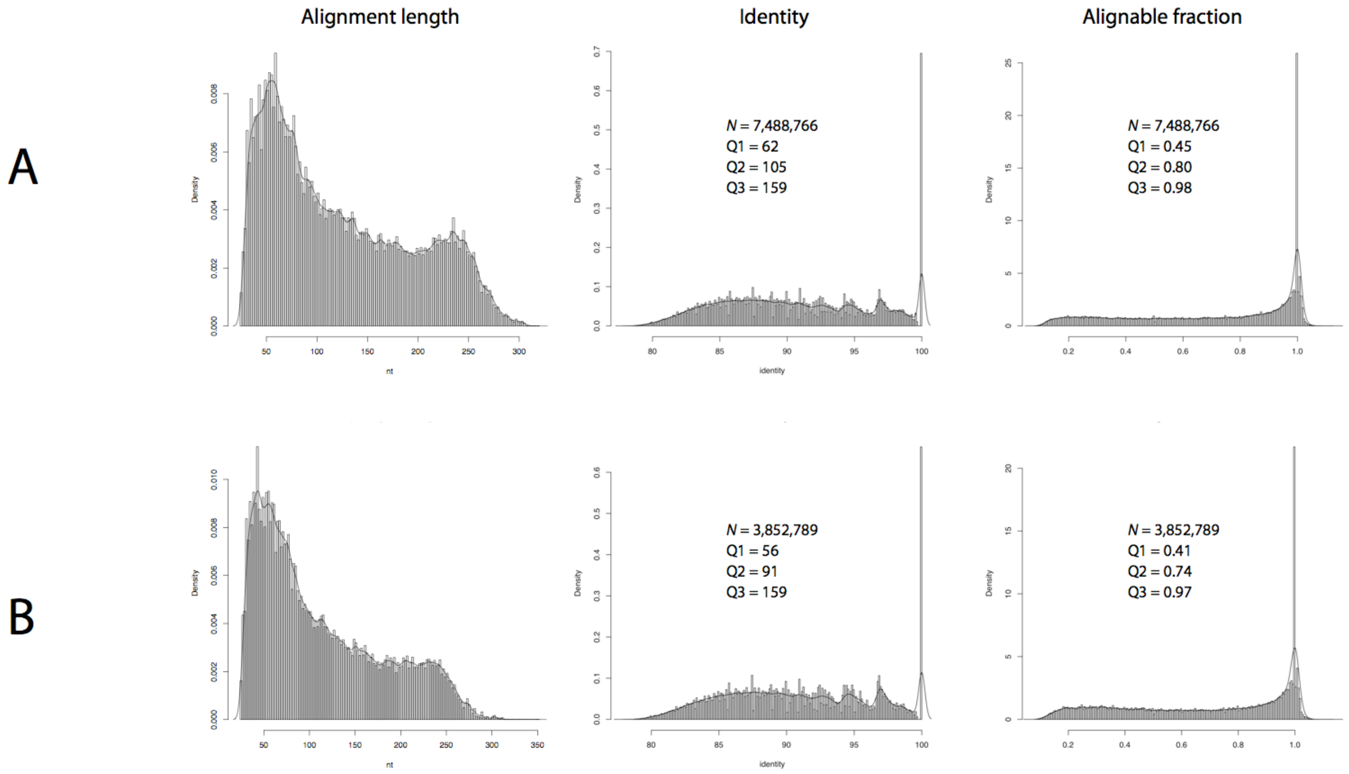

Alignment length
Identity
Alignable fraction
A
B
N = 7,488,766
Q1 = 62
Q2 = 105
Q3 = 159
N = 7,488,766
Q1 = 0.45
Q2 = 0.80
Q3 = 0.98
N = 3,852,789
Q1 = 56
Q2 = 91
Q3 = 159
N = 3,852,789
Q1 = 0.41
Q2 = 0.74
Q3 = 0.97
Density
0.000
0.002
0.004
0.006
0.008
0.010
nt
50
100
150
200
250
300
350
Density
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
identity
80
85
90
95
100
Density
0
5
10
15
20
25
0.2
0.4
0.6
0.8
1.0

Figure S7

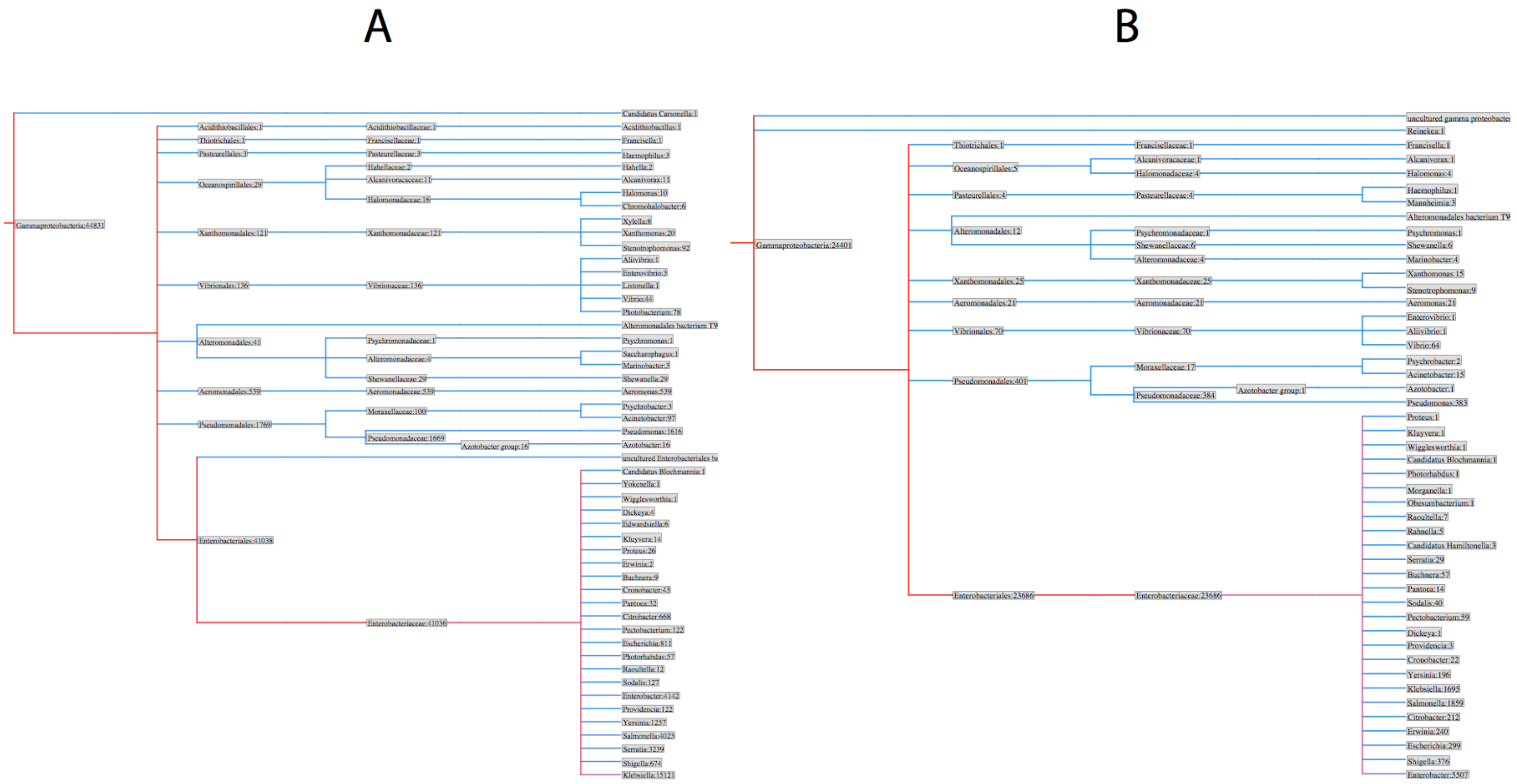
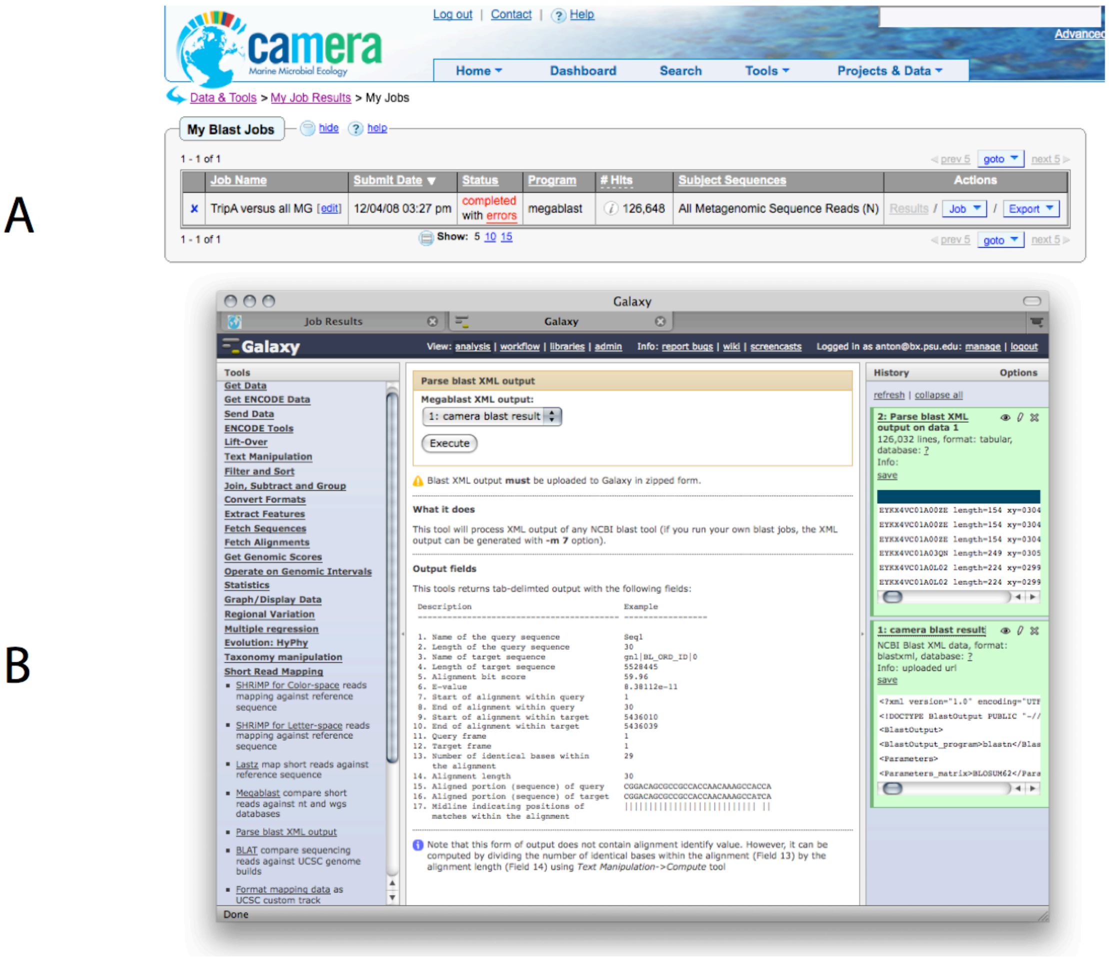

A
Gammaproteobacteria:44831
Acidithiobacillales:1
Thiotrichales:1
Pasteurellales:3
Oceanospirillales:29
Xanthomonadales:121
Vibrionales:136
Alteromonadales:41
Aeromonadales:539
Pseudomonadales:1769
Enterobacteriales:41038
Acidithiobacillaceae:1
Francisellaceae:1
Pasteurellaceae:3
Hahellaceae:2
Alcanivoracaceae:11
Halomonadaceae:16
Xanthomonadaceae:121
Vibrionaceae:136
Psychromonadaceae:1
Alteromonadaceae:4
Shewanellaceae:29
Aeromonadaceae:539
Moraxellaceae:100
Pseudomonadaceae:1669
Azotobacter group:16
Enterobacteriaceae:41036
Candidatus Carsonella:1
Acidithiobacillus:1
Francisella:1
Haemophilus:3
Hahella:2
Alcanivorax:11
Halomonas:10
Chromohalobacter:6
Xylella:8
Xanthomonas:20
Stenotrophomonas:92
Aliivibrio:1
Enterovibrio:3
Listonella:1
Vibrio:44
Photobacterium:78
Alteromonadales bacterium TW
Psychromonas:1
Saccharophagus:1
Marinobacter:3
Shewanella:29
Aeromonas:539
Psychrobacter:3
Acinetobacter:97
Pseudomonas:1616
Azotobacter:16
uncultured Enterobacteriales ba
Candidatus Blochmannia:1
Yokenella:1
Wigglesworthia:1
Dickeya:4
Edwardsiella:6
Kluyvera:14
Proteus:26
Erwinia:2
Buchnera:9
Cronobacter:43
Pantoea:32
Citrobacter:668
Pectobacterium:122
Escherichia:811
Photorhabdus:57
Raoultella:12
Sodalis:127
Enterobacter:4142
Providencia:122
Yersinia:1257
Salmonella:4023
Serratia:3239
Shigella:674
Klebsiella:15121
B
Gammaproteobacteria:24401
Thiotrichales:1
Oceanospirillales:5
Pasteurellales:4
Alteromonadales:12
Xanthomonadales:25
Aeromonadales:21
Vibrionales:70
Pseudomonadales:401
Enterobacteriales:23686
Francisellaceae:1
Alcanivoracaceae:1
Halomonadaceae:4
Pasteurellaceae:4
Psychromonadaceae:1
Shewanellaceae:6
Alteromonadaceae:4
Xanthomonadaceae:25
Aeromonadaceae:21
Vibrionaceae:70
Moraxellaceae:17
Pseudomonadaceae:384
Azotobacter group:1
Enterobacteriaceae:23686
uncultured gamma proteobacte
Reinekea:1
Francisella:1
Alcanivorax:1
Halomonas:4
Haemophilus:1
Mannheimia:3
Alteromonadales bacterium TW
Psychromonas:1
Shewanella:6
Marinobacter:4
Xanthomonas:15
Stenotrophomonas:9
Aeromonas:21
Enterovibrio:1
Aliivibrio:1
Vibrio:64
Psychrobacter:2
Acinetobacter:15
Azotobacter:1
Pseudomonas:383
Proteus:1
Kluyvera:1
Wigglesworthia:1
Candidatus Blochmannia:1
Photorhabdus:1
Morganella:1
Obesumbacterium:1
Raoultella:7
Rahnella:5
Candidatus Hamiltonella:3
Serratia:29
Buchnera:57
Pantoea:14
Sodalis:40
Pectobacterium:59
Dickeya:1
Providencia:3
Cronobacter:22
Yersinia:196
Klebsiella:1695
Salmonella:1859
Citrobacter:212
Erwinia:240
Escherichia:299
Shigella:376
Enterobacter:5507

Figure S8



	Job Name	Submit Date ▼	Status	Program	# Hits	Subject Sequences	Actions
x	TripA versus all MG [edit]	12/04/08 03:27 pm	completed with errors	megablast	126,648	All Metagenomic Sequence Reads (N)	Results / Job / Export

1 - 1 of 1

Parse blast XML output

Megablast XML output: 1: camera blast result

Blast XML output **must** be uploaded to Galaxy in zipped form.

What it does

This tool will process XML output of any NCBI blast tool (if you run your own blast jobs, the XML output can be generated with **-m 7** option).

Output fields

This tools returns tab-delimted output with the following fields:

```
 Description                              Example
 ---------------------------------------- -----------------

 1. Name of the query sequence             Seq1
 2. Length of the query sequence           30
 3. Name of target sequence                gnl|BL_ORD_ID|0
 4. Length of target sequence              5528445
 5. Alignment bit score                    59.96
 6. E-value                                8.38112e-11
 7. Start of alignment within query        1
 8. End of alignment within query          30
 9. Start of alignment within target       5436010
10. End of alignment within target         5436039
11. Query frame                            1
12. Target frame                           1
13. Number of identical bases within       29
    the alignment
14. Alignment length                       30
15. Aligned portion (sequence) of query    CGGACAGCGCCGCCACCAACAAAGCCACCA
16. Aligned portion (sequence) of target   CGGACAGCGCCGCCACCAACAAAGCCATCA
17. Midline indicating positions of        |||||||||||||||||||||||||||| ||
    matches within the alignment
```

Note that this form of output does not contain alignment identify value. However, it can be computed by dividing the number of identical bases within the alignment (Field 13) by the alignment length (Field 14) using *Text Manipulation->Compute* tool

2: Parse blast XML output on data 1
126,032 lines, format: tabular, database: ?
Info:

```
EYKX4VC01A00ZE length=154 xy=0304
EYKX4VC01A00ZE length=154 xy=0304
EYKX4VC01A00ZE length=154 xy=0304
EYKX4VC01A03QN length=249 xy=0305
EYKX4VC01A0L02 length=224 xy=0299
EYKX4VC01A0L02 length=224 xy=0299
```

1: camera blast result
NCBI Blast XML data, format: blastxml, database: ?
Info: uploaded url

```
<?xml version="1.0" encoding="UTF
<!DOCTYPE BlastOutput PUBLIC "-//
<BlastOutput>
<BlastOutput_program>blastn</Blas
<Parameters>
<Parameters_matrix>BLOSUM62</Para
```