Short Read Alignment: extended usage

Simon Anders
Nicolas Delhomme

EMBL-EBI

What next?

NGS offers the possibility to sequence anything and aligning the reads against “reference” genome is straightforward.

But what if there is no such “reference” genome?

➔ “de novo” assembly

Assembly

- Solexa reads are too short for *de novo* assembly of large genomes.
- However, for prokaryotes and simple eukaryotes, reasonably large contigs can be assembled.
- Using paired-end reads with very large end separation is crucial.
- Most popular assembly tools:
  - Velvet (Zerbino et al.)
  - ABySS (Simpson et al.)

Data

- Yeast strain
  - 2 Solexa lanes Single-End 36 bp
    - 8 295 633 reads
  - 1 Solexa lane Paired-End 36 bp
    - 8 553 586 reads

  25 402 805 reads

“de novo” Assembler

Velvet

Zerbino, D.R. & Birney, E. Velvet: Algorithms for de novo short read assembly using de Bruijn graphs. Genome Research 18, 821(2008).

Edena

Hernandez, D. et al. De novo bacterial genome sequencing: Millions of very short reads assembled on a desktop computer. Genome Research 18, 802(2008).

EULER-SR

Chaisson, M.J., Brinza, D. & Pevzner, P.A. De novo fragment assembly with short mate-paired reads: Does the read length matter? Genome Research 19, 336(2009).

ABySS

Simpson, T.J. et al. ABySS: A parallel assembler for short read sequence data. Genome Research (early publication; 2009.) http://genome.cshlp.org/content/early/2009/02/27/gr.089532.108.abstract

SSAKE, VCAKE, ALLPATHS

Table 4. Comparison of assemblies of E. coli K12 MG1655 short read data. For each assembly, only contigs ≥100bp in length were considered. Genome coverage is based on alignments with at least 95% identity to the reference genome (see Methods).

Assembler	Contigs ≥100bp	Mean size (bp)	N50 (bp)	Largest contig (bp)	Genome coverage (%)	Number of incorrect contigs (Mean size, bp)
ABySS	233	20,258	45,362	173,852	99.44	13 (33,252)
VELVET	286	15,910	54,359	164,194	98.81	9 (52,356)
EULER-SR	216	21,074	57,497	174,041	99.76	26 (37,863)
SSAKE	931	4,906	11,450	50,668	99.99	38 (5,881)
EDENA	680	6,687	16,430	67,082	99.08	6 (13,270)

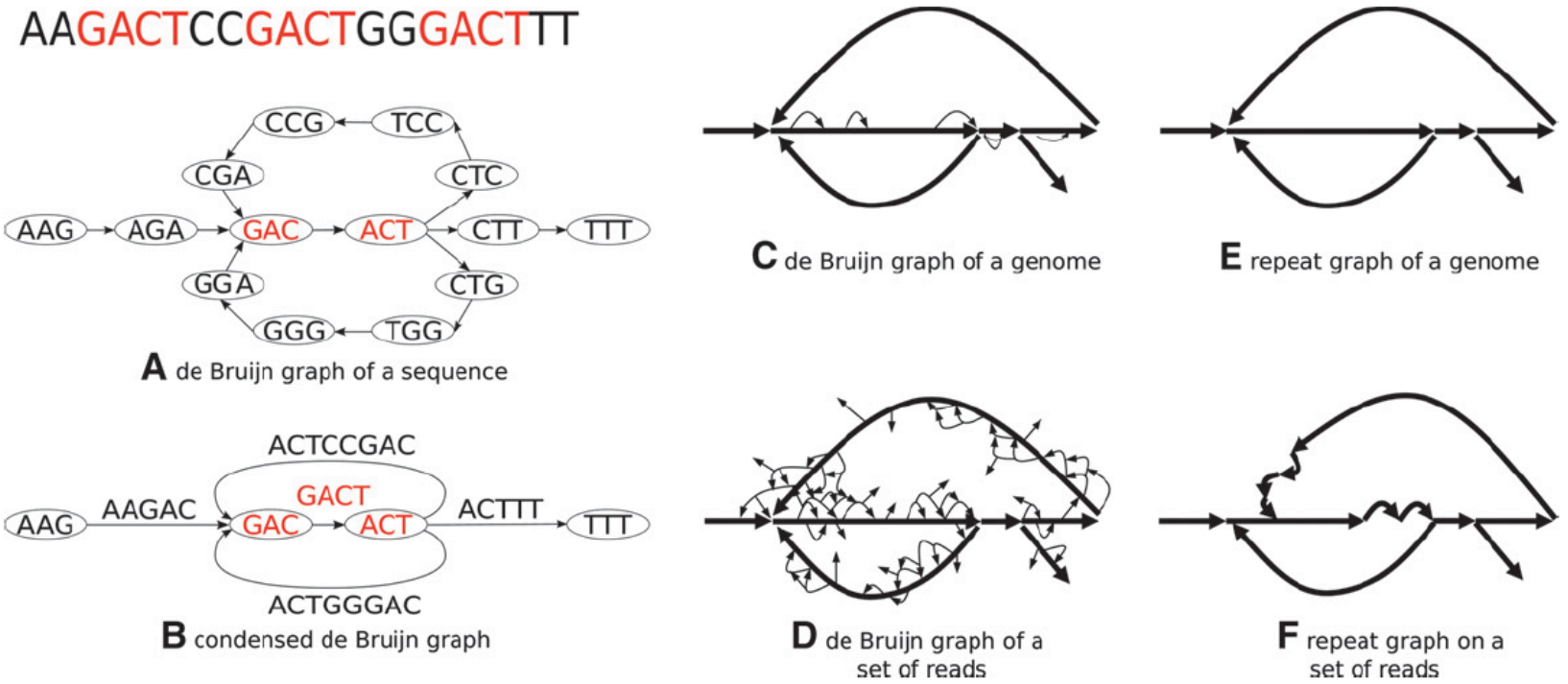
Principles

"Traditional" de novo assemblers (Sanger seq.) follow the "overlap-layout-consensus".

Most of the Short Read assemblers (but Edena) use an alternative "Eulerian" approach

Model the assembly problem as the search of an eulerian path* in a "de Bruijn" graph.

* an Eulerian path is a path in a graph which visits each edge exactly once


AAGACTCCGACTGGGACTTT
CCG
TCC
CGA
CTC
AAG
AGA
GAC
ACT
CTT
TTT
GGA
CTG
GGG
TGG
A de Bruijn graph of a sequence
ACTCCGAC
GACT
AAG
AAGAC
GAC
ACT
ACTTT
TTT
ACTGGGAC
B condensed de Bruijn graph
C de Bruijn graph of a genome
E repeat graph of a genome
D de Bruijn graph of a set of reads
F repeat graph on a set of reads

EMBL-EBI

Velvet

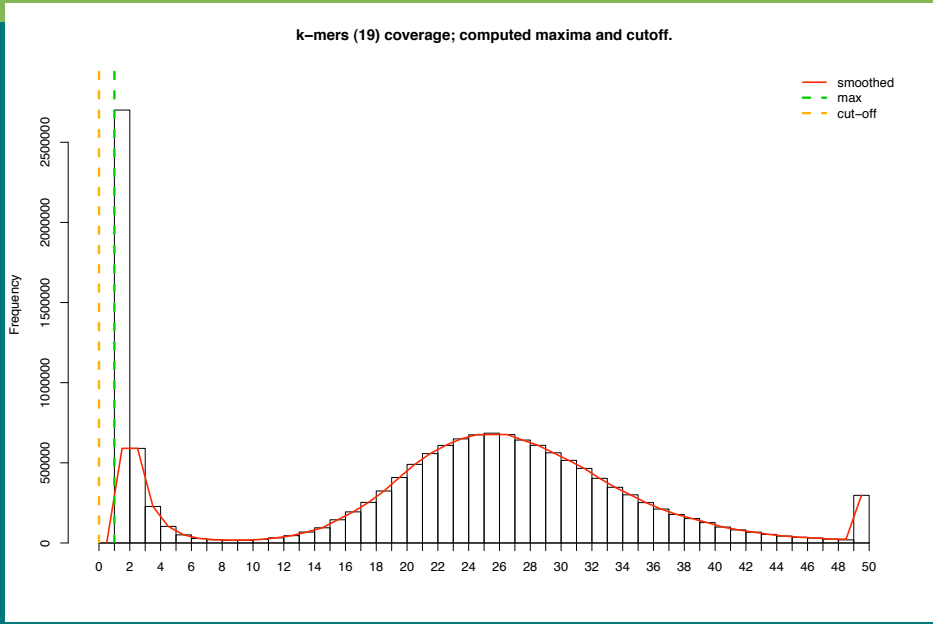
The most user friendly, more flexible and most documented one

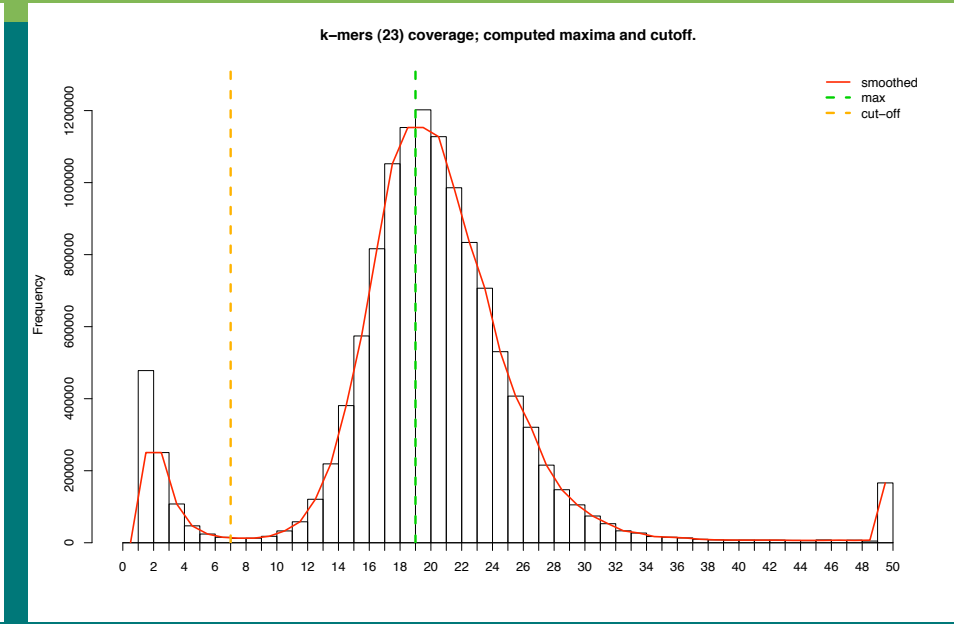
As for all “de Bruijn” based assembler, the results are highly dependant on the k-mer size

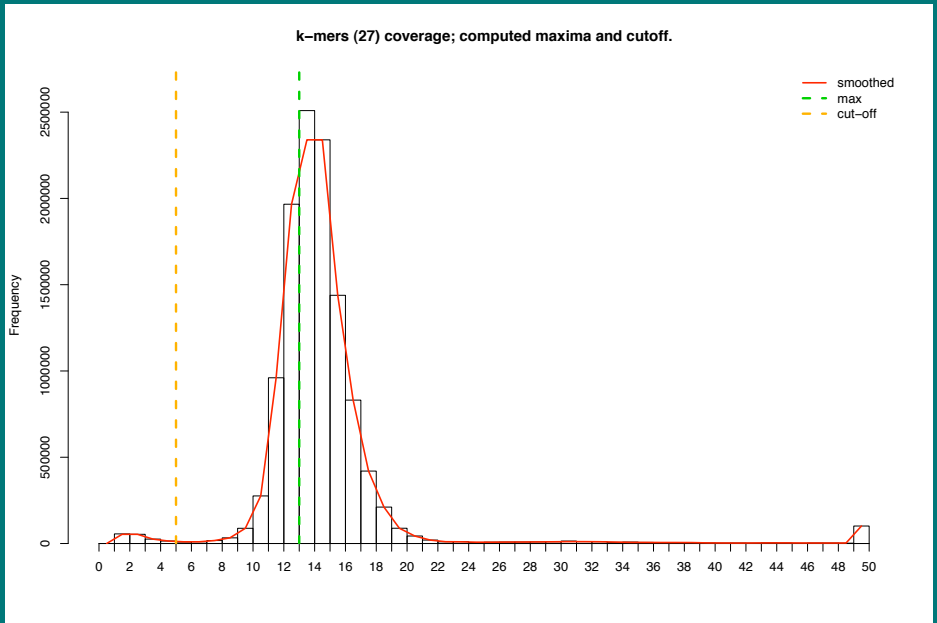
How to find the optimal k-mer?

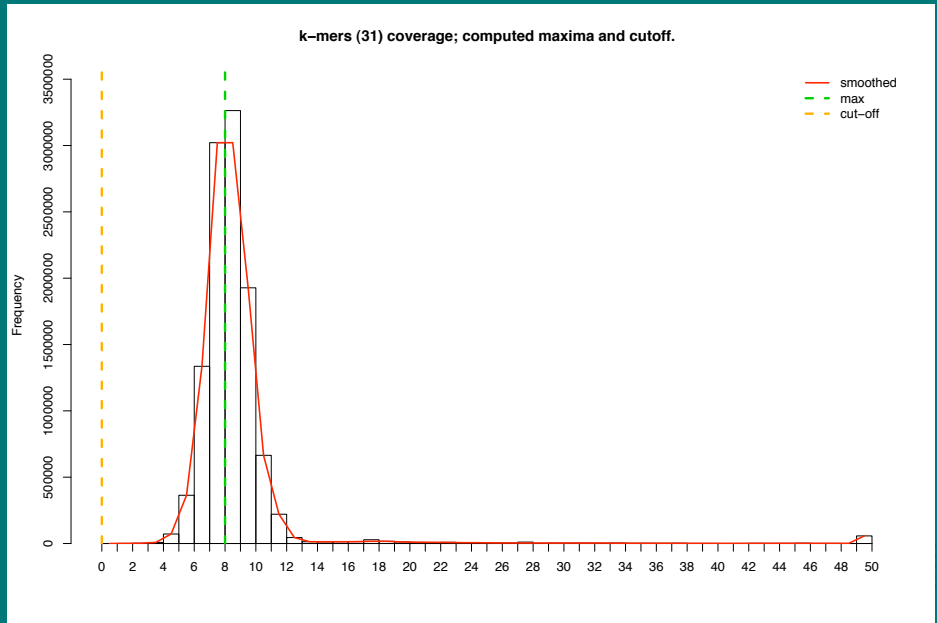
➔ run without any arguments for all sensible kmer values

How to find the cut-offs and coverage?









Finding the k-mer size

k-mer size	k-mer cov.	k-mer cutoff	N50 *	longest contig**	# nodes	# kmers
11	49	25	1	3	2075499	2069397
13	1	0	1	23	16317474	20124578
15	1	0	5	63	9428658	27135302
17	1	0	12	269	2359774	18065172
19	1	0	25	748	1002389	15203224
21	1	0	70	1626	443367	13342537
23	**19**	7	245	3488	209064	12441115
25	16	6	**800**	6586	100829	11963428
27	13	5	**2454**	**17974**	44264	11685379
29	10	3	**2903**	16285	24232	11538411
31	8	0	**799**	7039	31913	11190402

* N50 is calculated by first ordering all contigs by size and then adding the lengths (starting from the longest contig) until the summed length exceeds 50% of the total length of all contigs.

** calculated in number of k-mers

EMBL-EBI

Run with optimized parameters

k-mer size	N50	# nodes	# kmers	longest contig
21	16584	22844	11576807	72826
23	30407	16407	11583626	133739
25	32967	14099	11596286	149035
27	32569	12304	11610607	**274179**
29	23160	12399	11678034	134911
31	9763	11193	11897654	47081

1523 contigs > 100bp

Total of 11 469 222bp

92.8% coverage of the genome

ABySS

No documentation

Paired-End or Single-End mode (not combinable as in Velvet)

runs with 25, 27, 29 and 31 k-mers

- Need to filter contigs smaller than 100bp
- Need to calculate statistics (N50, longest contig)

Finding the kmer size and optimized run

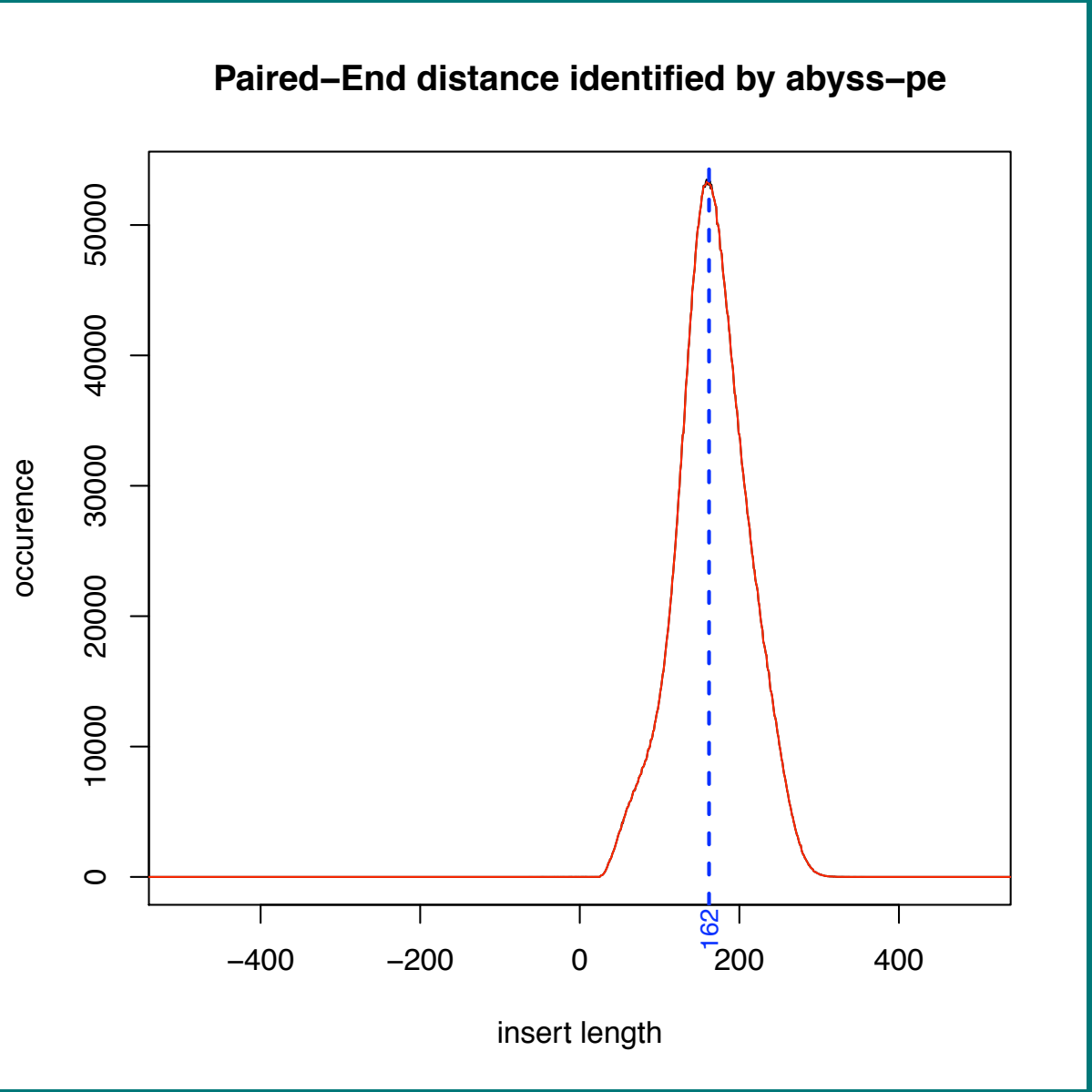
k-mer	# contig > 100	N50	longest contig	total bp	genome coverage
25	28311	5382	5270	10648862	86.12%
27	14414	2193	12218	11301463	91.40%
29	**7528**	**1036**	**18645**	11466157	92.73%
31	12343	1982	13588	**11608757**	**93.88%**

Using the Paired-end mode on the mixed single and paired end data:

contigs:	1766
longest:	79902bp,
N50:	16014bp
genome coverage:	93.4%

EMBL-EBI

Paired end distance



EMBL-EBI

minimus2 (AMOS)

Combine different Velvet & Abyss assemblies

contigs:	900
longest:	274179 bp,
N50:	32569 bp
genome coverage:	95.1%

BlastN

```
>SK1.chr16
          Length = 960530

 Score = 1237 bits (624), Expect = 0.0
 Identities = 640/648 (98%)
 Strand = Plus / Plus


Query: 81     gaaacaagttattagtacaaaaatcaagctctcactacttaccagtagttattttagtag 140
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 867336 gaaacaagttattagtacaaaaatcaagctctcactacttaccagtagttattttagtag 867395


Query: 141    aatattattaactctaccgaataagtacgatatgagaaattgctgatttacttgcgtgtg 200
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 867396 aatattattaactctaccgaataagtacgatatgagaaattgctgatttacttgcgtgtg 867455


Query: 201    ttcctattattatccaggagcgcctctaatctatgcttataatttcacgttgnnnnnnnn 260
              ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 867456 ttcctattattatccaggagcgcctctaatctatgcttataatttcacgttgaaaaaaaa 867515


Query: 261    ggagtctgtataattatttcaaatccatccatttctcacggatattgtgtaccctatgag 320
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 867516 ggagtctgtataattatttcaaatccatccatttctcacggatattgtgtaccctatgag 867575


Query: 321    gtaaataattgcgctttattttctccctgtgtttgctcgctcatcttaagacgaaaaaag 380
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 867576 gtaaataattgcgctttattttctccctgtgtttgctcgctcatcttaagacgaaaaaag 867635


Query: 381    taatggaacacctaccatcaatagaagtaccacactcatgtaaataacaagcgcaagtgg 440
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 867636 taatggaacacctaccatcaatagaagtaccacactcatgtaaataacaagcgcaagtgg 867695


Query: 441    tttagtggtaaaatccaacgttgccatcgttgggcccccggttcgattccgggcttgcgc 500
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 867696 tttagtggtaaaatccaacgttgccatcgttgggcccccggttcgattccgggcttgcgc 867755


Query: 501    agattttatttttgctcccttctaaagctctggatgattcaccaacttatccaaggcaa 560
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 867756 agattttatttttgctcccttctaaagctctggatgattcaccaacttatccaaggcaa 867815
```

```
>chr16
          Length = 948062

 Score = 1189 bits (600), Expect = 0.0
 Identities = 634/648 (97%)
 Strand = Plus / Plus


Query: 81     gaaacaagttattagtacaaaaatcaagctctcactacttaccagtagttattttagtag 140
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 860025 gaaacaagttattagtacaaaaatcaagctctcactacttaccagtagttattttagtag 860084


Query: 141    aatattattaactctaccgaataagtacgatatgagaaattgctgatttacttgcgtgtg 200
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 860085 aatattattaactctaccgaataagtacgatatgagaaattgctgatttacttgtgtgtg 860144


Query: 201    ttcctattattatccaggagcgcctctaatctatgcttataatttcacgttgnnnnnnnn 260
              ||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 860145 ttcctattattatccaggagcgcctctaatctatgcttataatttcacgttgaaaaaaaa 860204


Query: 261    ggagtctgtataattatttcaaatccatccattctcacggatattgtgtaccctatgag 320
              ||||||||||||||||||||||||| |||| ||||||||||||||||||||||||||||
Sbjct: 860205 ggagtctgtataattatttcaaattcatcttttctcacggatattgtgtaccctatgag 860264


Query: 321    gtaaataattgcgctttattttctccctgtgtttgctcgctcatcttaagacgaaaaaag 380
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 860265 gtaaataattgcgctttattttctccctgtgtttgctcgctcatcttaagacgaaaaaag 860324


Query: 381    taatggaacacctaccatcaatagaagtaccacactcatgtaaataacaagcgcaagtgg 440
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 860325 taatggaacacctaccatcaatagaagtaccacactcatgtaaataacaagcgcaagtgg 860384


Query: 441    tttagtggtaaaatccaacgttgccatcgttgggcccccggttcgattccgggcttgcgc 500
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 860385 tttagtggtaaaatccaacgttgccatcgttgggcccccggttcgattccgggcttgcgc 860444


Query: 501    agattttatttttgctcccttctaaagctctggatgattcaccaacttatccaaggcaa 560
              ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 860445 agattttatttttgctcccttctaaagctctggatgattcaccaacttatccaaggcaa 860504
```

EMBL-EBI