Genome Annotation with RAST and Artemis

Jeffrey Long

UCSC 2010

Bioinformatics

Outline

- Methods of genome annotation
  - UCSC Genome Browser, Archaeal Browser
    - Patricia Chen
  - RAST, MGRAST
- Tools for browsing annotation
  - UCSC Genome Browser, Archaeal Browser
  - RAST, MGRAST
  - Artemis Comparison Tool (ACT)
    - Artemis, DNAPlotter, WebACT, BamView

BMC Genomics

BioMed Central

Database

Open Access

The RAST Server: Rapid Annotations using Subsystems Technology

Ramy K Aziz[8,9], Daniela Bartels[3], Aaron A Best[7], Matthew DeJongh[7], Terrence Disz[2,3], Robert A Edwards[1,2], Kevin Formsma[7], Svetlana Gerdes[1], Elizabeth M Glass[2], Michael Kubal[3], Folker Meyer[2,3], Gary J Olsen[4,2], Robert Olson[2,3], Andrei L Osterman[1,5], Ross A Overbeek*[1], Leslie K McNeil[6], Daniel Paarmann[3], Tobias Paczian[3], Bruce Parrello[1], Gordon D Pusch[1,3], Claudia Reich[6], Rick Stevens[2,3], Olga Vassieva[1], Veronika Vonstein[1], Andreas Wilke[3] and Olga Zagnitko[1]

Address: [1]Fellowship for Interpretation of Genomes, Burr Ridge, IL 60527, USA, [2]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439, USA, [3]Computation Institute, University of Chicago, Chicago, IL 60637, USA, [4]Department of Microbiology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA, [5]The Burnham Institute, San Diego, CA 92037, USA, [6]National Center for Supercomputing Applications, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA, [7]Hope College, Holland, MI 49423, USA, [8]University of Tennessee, Health Science Center, Memphis, TN 38136, USA and [9]Department of Microbiology and Immunology, Cairo University, Cairo, Egypt

Email: Ramy K Aziz - ramy.aziz@gmail.com; Daniela Bartels - bartels@mcs.anl.gov; Aaron A Best - Best@hope.edu; Matthew DeJongh - dejongh@hope.edu; Terrence Disz - disz@mcs.anl.gov; Robert A Edwards - RobE@theFIG.info; Kevin Formsma - kevin.formsma@hope.edu; Svetlana Gerdes - Sveta@theFIG.info; Elizabeth M Glass - marland@mcs.anl.gov; Michael Kubal - mkubal@mcs.anl.gov; Folker Meyer - folker@mcs.anl.gov; Gary J Olsen - gary@life.uiuc.edu; Robert Olson - olson@mcs.anl.gov; Andrei L Osterman - osterman@burnham.org; Ross A Overbeek* - Ross@theFIG.info; Leslie K McNeil - lkmcneil@ncsa.uiuc.edu; Daniel Paarmann - paarmann@mcs.anl.gov; Tobias Paczian - paczian@mcs.anl.gov; Bruce Parrello - drake@mkrules.net; Gordon D Pusch - gdpusch@xnet.com; Claudia Reich - creich@ncsa.uiuc.edu; Rick Stevens - stevens@anl.gov; Olga Vassieva - OlgaV@theFIG.info; Veronika Vonstein - Veronika@theFIG.info; Andreas Wilke - wilke@mcs.anl.gov; Olga Zagnitko - OlgaZ@theFIG.info

* Corresponding author

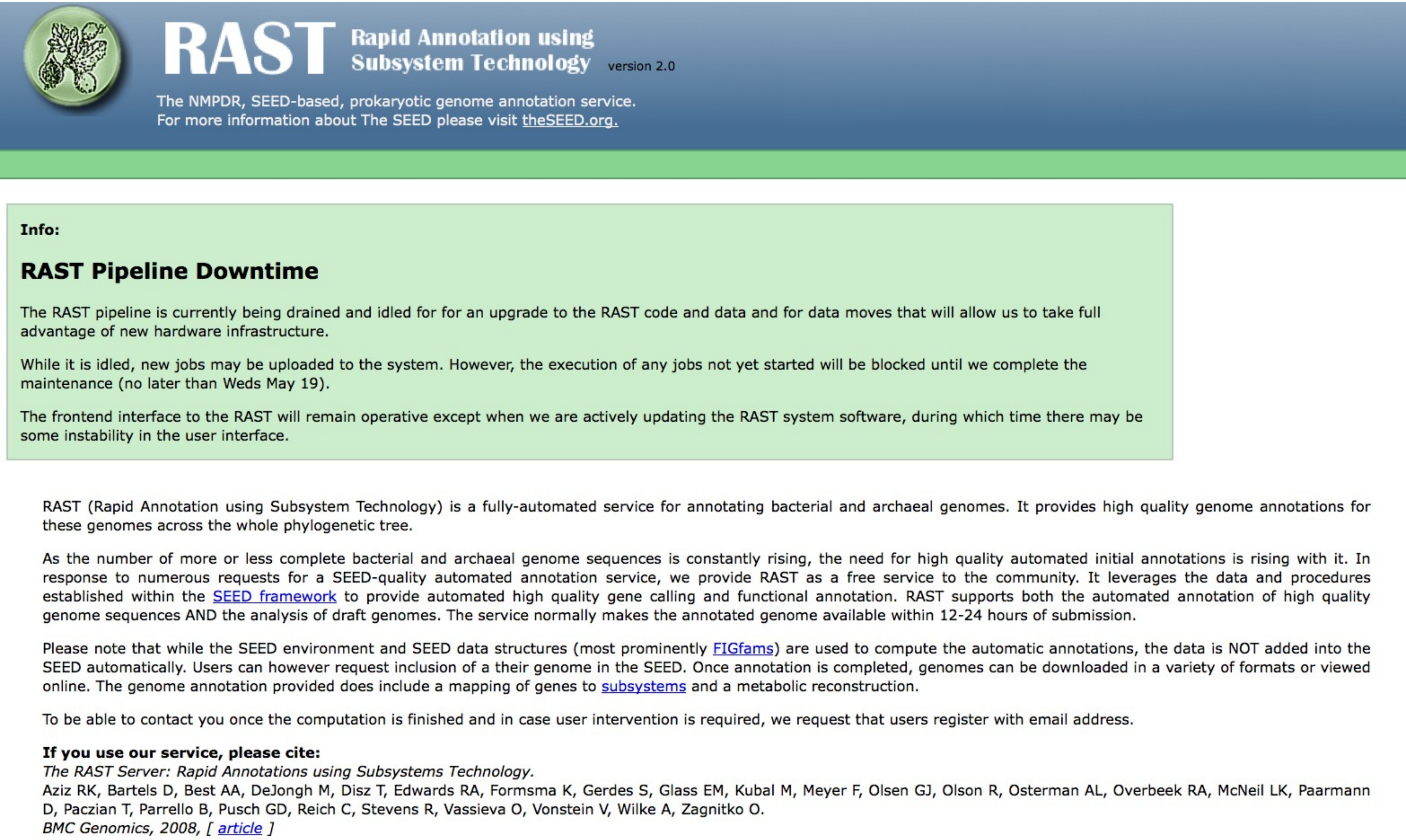
Published: 8 February 2008

BMC Genomics 2008, 9:75 doi:10.1186/1471-2164-9-75

Received: 12 September 2007

Accepted: 8 February 2008

This article is available from: http://www.biomedcentral.com/1471-2164/9/75



RAST Rapid Annotation using Subsystem Technology version 2.0

The NMPDR, SEED-based, prokaryotic genome annotation service.
For more information about The SEED please visit theSEED.org.

Info:

RAST Pipeline Downtime

The RAST pipeline is currently being drained and idled for for an upgrade to the RAST code and data and for data moves that will allow us to take full advantage of new hardware infrastructure.

While it is idled, new jobs may be uploaded to the system. However, the execution of any jobs not yet started will be blocked until we complete the maintenance (no later than Weds May 19).

The frontend interface to the RAST will remain operative except when we are actively updating the RAST system software, during which time there may be some instability in the user interface.

RAST (Rapid Annotation using Subsystem Technology) is a fully-automated service for annotating bacterial and archaeal genomes. It provides high quality genome annotations for these genomes across the whole phylogenetic tree.

As the number of more or less complete bacterial and archaeal genome sequences is constantly rising, the need for high quality automated initial annotations is rising with it. In response to numerous requests for a SEED-quality automated annotation service, we provide RAST as a free service to the community. It leverages the data and procedures established within the SEED framework to provide automated high quality gene calling and functional annotation. RAST supports both the automated annotation of high quality genome sequences AND the analysis of draft genomes. The service normally makes the annotated genome available within 12-24 hours of submission.

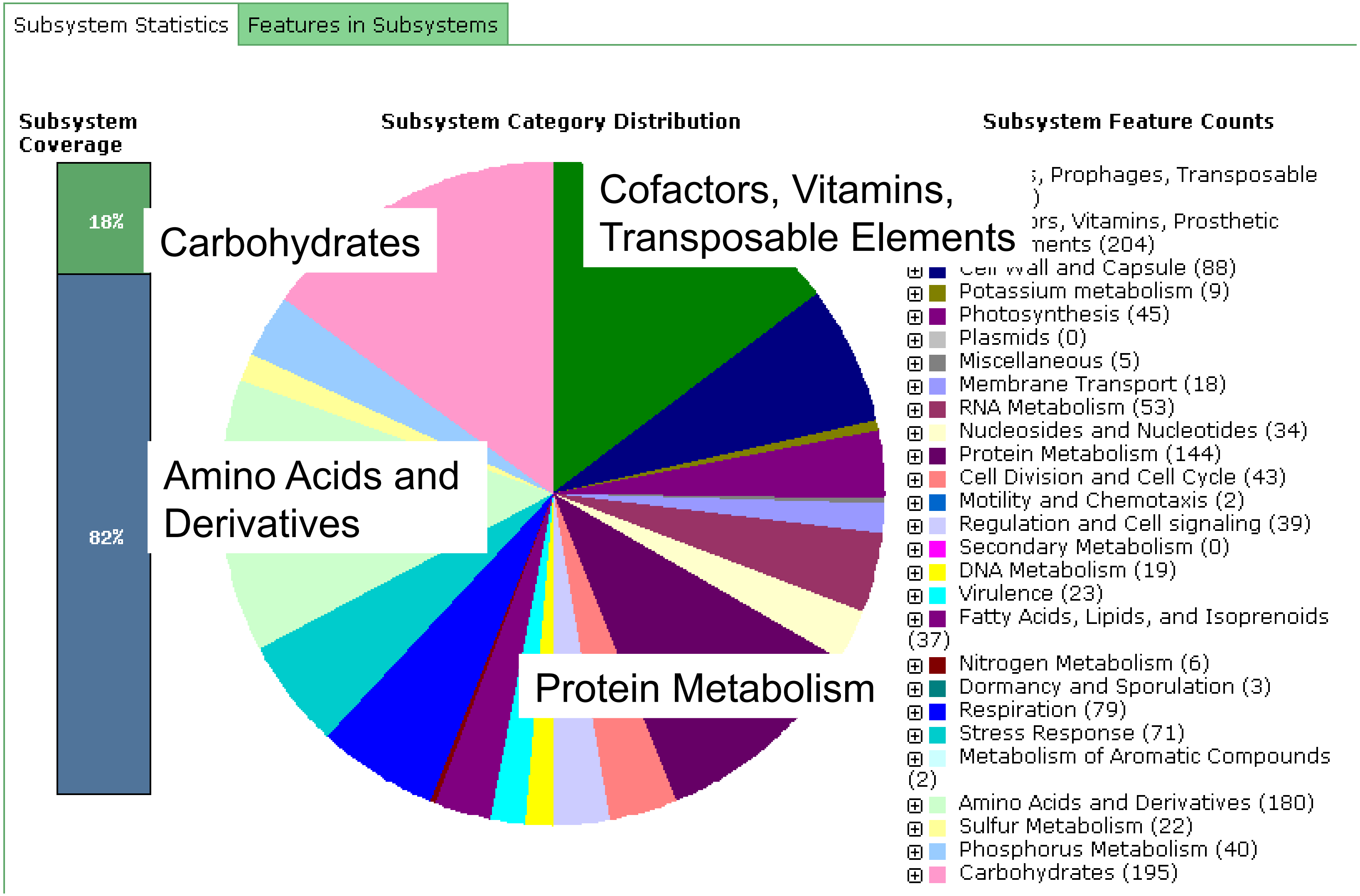
Please note that while the SEED environment and SEED data structures (most prominently FIGfams) are used to compute the automatic annotations, the data is NOT added into the SEED automatically. Users can however request inclusion of a their genome in the SEED. Once annotation is completed, genomes can be downloaded in a variety of formats or viewed online. The genome annotation provided does include a mapping of genes to subsystems and a metabolic reconstruction.

To be able to contact you once the computation is finished and in case user intervention is required, we request that users register with email address.

If you use our service, please cite:
The RAST Server: Rapid Annotations using Subsystems Technology.
Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O.
BMC Genomics, 2008, [article]

Image from RAST website: http://rast.nmpdr.org/

RAST Annotation



Carbohydrates

Cofactors, Vitamins, Transposable Elements

Amino Acids and Derivatives

Protein Metabolism

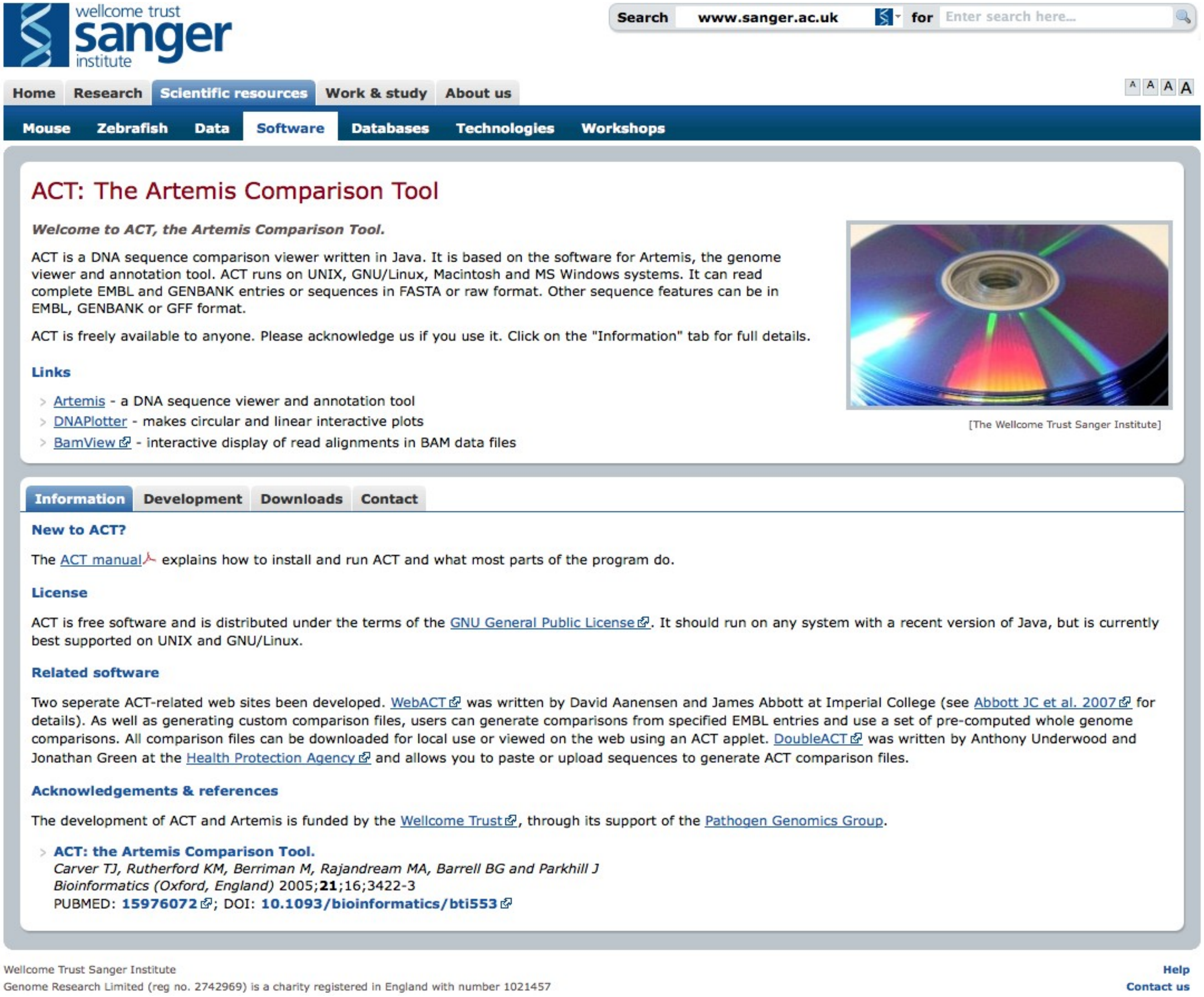
Subsystem Feature Counts

s, Prophages, Transposable
)
ors, Vitamins, Prosthetic
ments (204)
⊞ Cell Wall and Capsule (88)
⊞ Potassium metabolism (9)
⊞ Photosynthesis (45)
⊞ Plasmids (0)
⊞ Miscellaneous (5)
⊞ Membrane Transport (18)
⊞ RNA Metabolism (53)
⊞ Nucleosides and Nucleotides (34)
⊞ Protein Metabolism (144)
⊞ Cell Division and Cell Cycle (43)
⊞ Motility and Chemotaxis (2)
⊞ Regulation and Cell signaling (39)
⊞ Secondary Metabolism (0)
⊞ DNA Metabolism (19)
⊞ Virulence (23)
⊞ Fatty Acids, Lipids, and Isoprenoids (37)
⊞ Nitrogen Metabolism (6)
⊞ Dormancy and Sporulation (3)
⊞ Respiration (79)
⊞ Stress Response (71)
⊞ Metabolism of Aromatic Compounds (2)
⊞ Amino Acids and Derivatives (180)
⊞ Sulfur Metabolism (22)
⊞ Phosphorus Metabolism (40)
⊞ Carbohydrates (195)

Image from RAST website: http://rast.nmpdr.org/

The Artemis Comparison Tool (ACT)

- Artemis
- DNA Plotter
- WebACT



ACT: The Artemis Comparison Tool

Welcome to ACT, the Artemis Comparison Tool.

ACT is a DNA sequence comparison viewer written in Java. It is based on the software for Artemis, the genome viewer and annotation tool. ACT runs on UNIX, GNU/Linux, Macintosh and MS Windows systems. It can read complete EMBL and GENBANK entries or sequences in FASTA or raw format. Other sequence features can be in EMBL, GENBANK or GFF format.

ACT is freely available to anyone. Please acknowledge us if you use it. Click on the "Information" tab for full details.

Links

- Artemis - a DNA sequence viewer and annotation tool
- DNAPlotter - makes circular and linear interactive plots
- BamView - interactive display of read alignments in BAM data files

New to ACT?

The ACT manual explains how to install and run ACT and what most parts of the program do.

License

ACT is free software and is distributed under the terms of the GNU General Public License. It should run on any system with a recent version of Java, but is currently best supported on UNIX and GNU/Linux.

Related software

Two seperate ACT-related web sites been developed. WebACT was written by David Aanensen and James Abbott at Imperial College (see Abbott JC et al. 2007 for details). As well as generating custom comparison files, users can generate comparisons from specified EMBL entries and use a set of pre-computed whole genome comparisons. All comparison files can be downloaded for local use or viewed on the web using an ACT applet. DoubleACT was written by Anthony Underwood and Jonathan Green at the Health Protection Agency and allows you to paste or upload sequences to generate ACT comparison files.

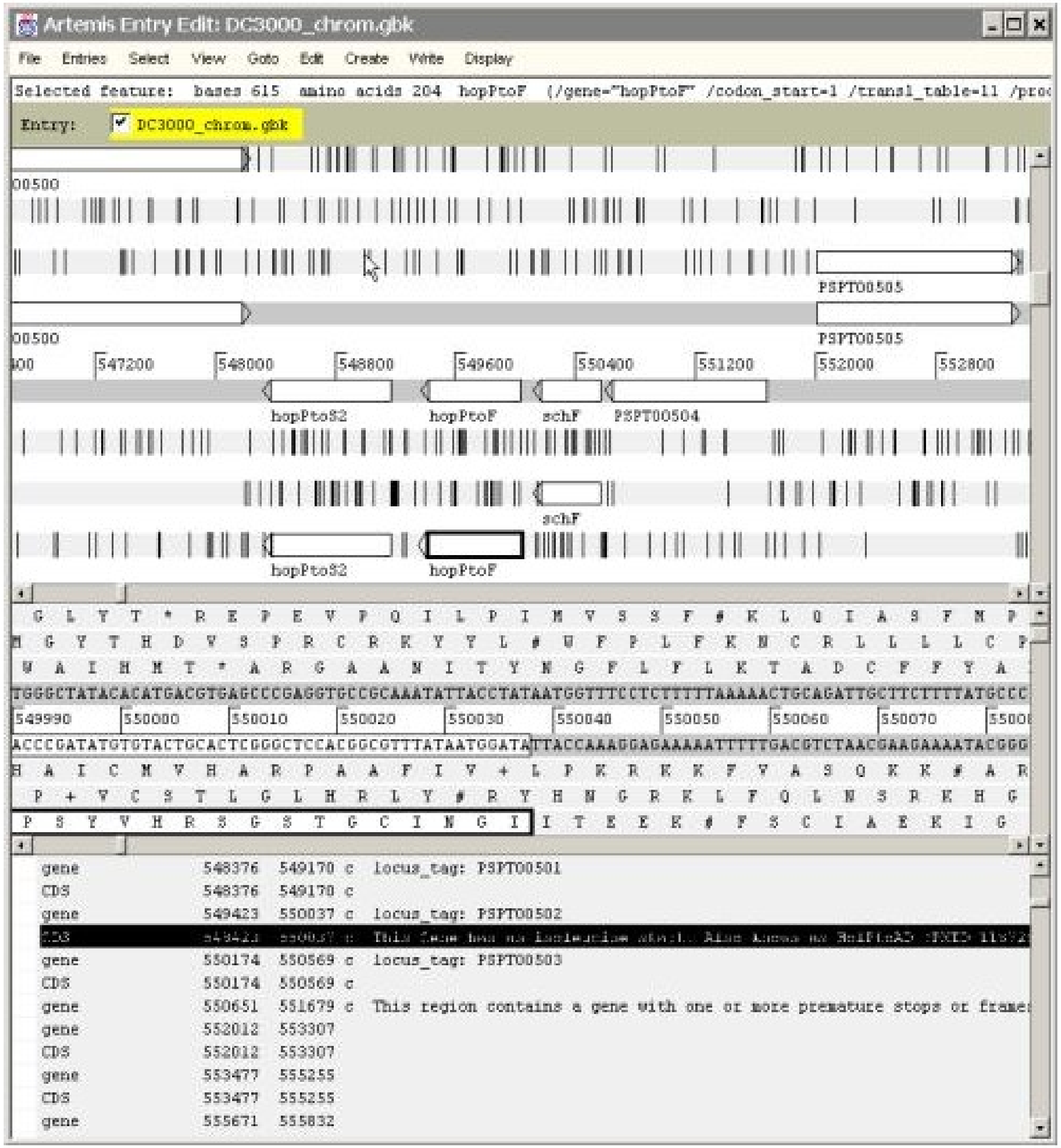
Acknowledgements & references

The development of ACT and Artemis is funded by the Wellcome Trust, through its support of the Pathogen Genomics Group.

- **ACT: the Artemis Comparison Tool.**
  Carver TJ, Rutherford KM, Berriman M, Rajandream MA, Barrell BG and Parkhill J
  Bioinformatics (Oxford, England) 2005;**21**;16;3422-3
  PUBMED: **15976072**; DOI: **10.1093/bioinformatics/bti553**

Wellcome Trust Sanger Institute
Genome Research Limited (reg no. 2742969) is a charity registered in England with number 1021457

Image from http://www.sanger.ac.uk/



```
Selected feature:  bases 615  amino acids 204  hopPtoF  (/gene="hopPtoF" /codon_start=1 /transl_table=11 /pro
```

```
 G  L  Y  T  *  R  E  P  E  V  P  Q  I  L  P  I  M  V  S  S  F  #  K  L  Q  I  A  S  F  M  P
M  G  Y  T  H  D  V  S  P  R  C  R  K  Y  Y  L  #  W  F  P  L  F  K  N  C  R  L  L  L  L  C  P
 W  A  I  H  M  T  *  A  R  G  A  A  N  I  T  Y  N  G  F  L  F  L  K  T  A  D  C  F  F  Y  A
TGGGCTATACACATGACGTGAGCCCGAGGTGCCGCAAATATTACCTATAATGGTTTCCTCTTTTTAAAAACTGCAGATTGCTTCTTTTATGCCC
549990    550000    550010    550020    550030    550040    550050    550060    550070    5500
ACCCGATATGTGTACTGCACTCGGGCTCCACGGCGTTTATAATGGATATTACCAAAGGAGAAAAATTTTTGACGTCTAACGAAGAAAATACGGG
H  A  I  C  M  V  H  A  R  P  A  A  F  I  V  +  L  P  K  R  K  K  F  V  A  S  Q  K  K  #  A  R
 P  +  V  C  S  T  L  G  L  H  R  L  Y  #  R  Y  H  N  G  R  K  L  F  Q  L  N  S  R  K  H  G
P  S  Y  V  H  R  S  G  S  T  G  C  I  N  G  I  I  T  E  E  K  #  F  S  C  I  A  E  K  I  G
```

```
gene            548376  549170 c  locus_tag: PSPTO0501
CDS             548376  549170 c
gene            549423  550037 c  locus_tag: PSPTO0502
CDS             549423  550037 c  This [illegible]
gene            550174  550569 c  locus_tag: PSPTO0503
CDS             550174  550569 c
gene            550651  551679 c  This region contains a gene with one or more premature stops or frame
gene            552012  553307
CDS             552012  553307
gene            553477  555255
CDS             553477  555255
gene            555671  555832
```

Selected feature (HopPtoF)

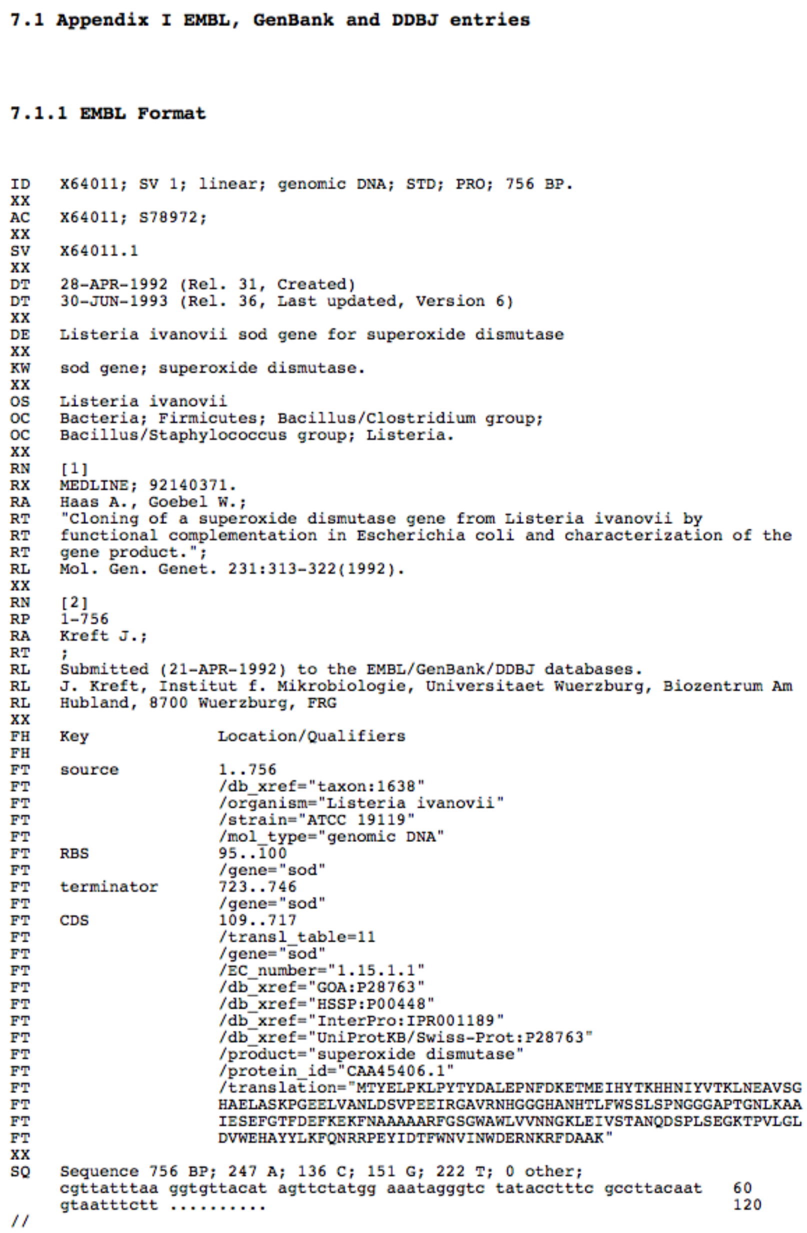
Active entry
annotated sequence of DC3000 chromosome downloaded from NCBI

Overview window
shows features specified in the active entry (in this case, genes and CoDing Sequences or CDS) overlaid on the two DNA strands and six translation frames

DNA view window
genes and CDSes as above, shown on the sequence level

Feature list
shows annotation record for all features in the active entry

http://pseudomonas-syringae.org/artemis_tutorial.htm



7.1 Appendix I EMBL, GenBank and DDBJ entries

7.1.1 EMBL Format

```
ID   X64011; SV 1; linear; genomic DNA; STD; PRO; 756 BP.
XX
AC   X64011; S78972;
XX
SV   X64011.1
XX
DT   28-APR-1992 (Rel. 31, Created)
DT   30-JUN-1993 (Rel. 36, Last updated, Version 6)
XX
DE   Listeria ivanovii sod gene for superoxide dismutase
XX
KW   sod gene; superoxide dismutase.
XX
OS   Listeria ivanovii
OC   Bacteria; Firmicutes; Bacillus/Clostridium group;
OC   Bacillus/Staphylococcus group; Listeria.
XX
RN   [1]
RX   MEDLINE; 92140371.
RA   Haas A., Goebel W.;
RT   "Cloning of a superoxide dismutase gene from Listeria ivanovii by
RT   functional complementation in Escherichia coli and characterization of the
RT   gene product.";
RL   Mol. Gen. Genet. 231:313-322(1992).
XX
RN   [2]
RP   1-756
RA   Kreft J.;
RT   ;
RL   Submitted (21-APR-1992) to the EMBL/GenBank/DDBJ databases.
RL   J. Kreft, Institut f. Mikrobiologie, Universitaet Wuerzburg, Biozentrum Am
RL   Hubland, 8700 Wuerzburg, FRG
XX
FH   Key             Location/Qualifiers
FH
FT   source          1..756
FT                   /db_xref="taxon:1638"
FT                   /organism="Listeria ivanovii"
FT                   /strain="ATCC 19119"
FT                   /mol_type="genomic DNA"
FT   RBS             95..100
FT                   /gene="sod"
FT   terminator      723..746
FT                   /gene="sod"
FT   CDS             109..717
FT                   /transl_table=11
FT                   /gene="sod"
FT                   /EC_number="1.15.1.1"
FT                   /db_xref="GOA:P28763"
FT                   /db_xref="HSSP:P00448"
FT                   /db_xref="InterPro:IPR001189"
FT                   /db_xref="UniProtKB/Swiss-Prot:P28763"
FT                   /product="superoxide dismutase"
FT                   /protein_id="CAA45406.1"
FT                   /translation="MTYELPKLPYTYDALEPNFDKETMEIHYTKHHNIYVTKLNEAVSG
FT                   HAELASKPGEELVANLDSVPEEIRGAVRNHGGGHANHTLFWSSLSPNGGGAPTGNLKAA
FT                   IESEFGTFDEFKEKFNAAAAARFGSGWAWLVVNNGKLEIVSTANQDSPLSEGKTPVLGL
FT                   DVWEHAYYLKFQNRRPEYIDTFWNVINWDERNKRFDAAK"
XX
SQ   Sequence 756 BP; 247 A; 136 C; 151 G; 222 T; 0 other;
     cgttatttaa ggtgttacat agttctatgg aaatagggtc tataccttc gccttacaat   60
     gtaatttctt ..........                                              120
//
```

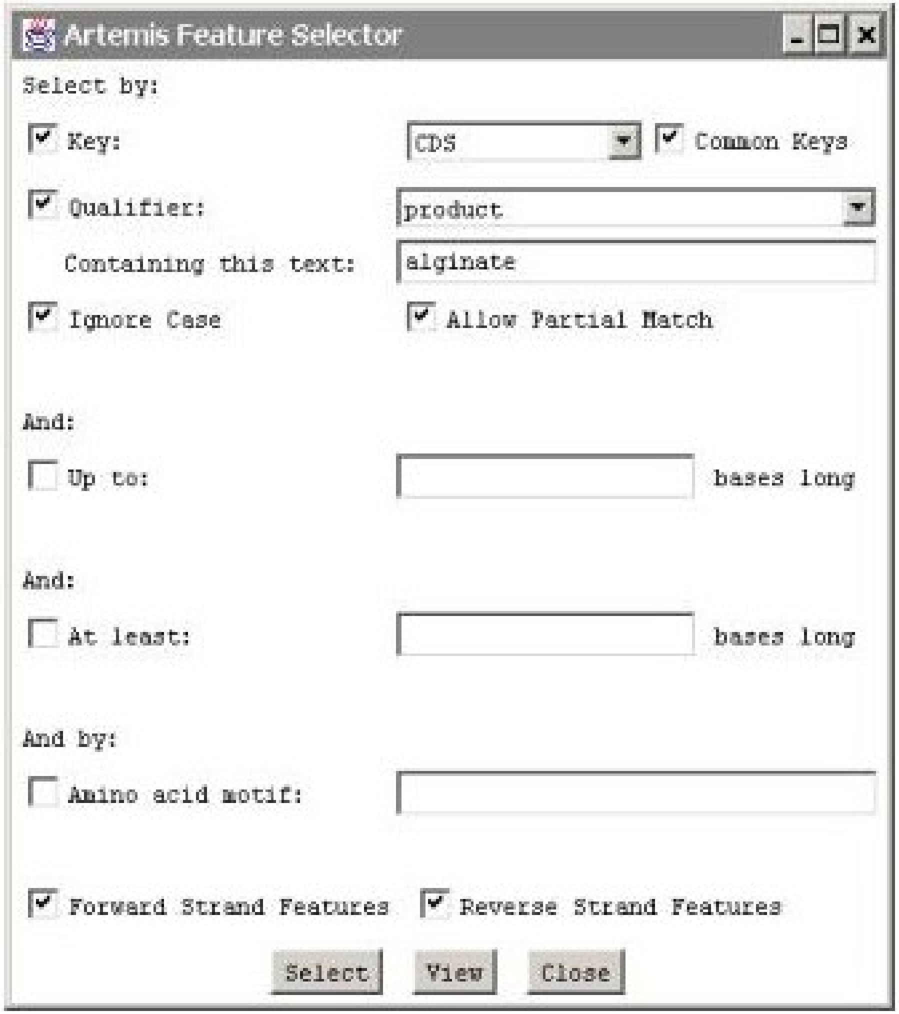
http://www.ebi.ac.uk/embl/Documentation/FT_definitions/feature_table.html

Example 1. Select coding sequences involved in alginate biosynthesis:

1. Select>"Feature selector.." (selects features based on their shared qualifiers) In this example, I am selecting coding sequences for which the "product" qualifier contains the word "alginate" (see window at right)
2. click on Select
3. click on View (brings up a window showing the list of selected features)
4. Select desired features on the list
5. Edit>"Copy selected features" (specify the entry file to which they will be copied)

Example 2. Select genes involved in alginate biosynthesis

1. Select>"Feature selector.." using the following selection terms:
   Key = gene
   Qualifier = gene
   Containing this text = alg
2. Proceed as described in Example 1.



Artemis Feature Selector

Select by:

☑ Key: CDS ☑ Common Keys

☑ Qualifier: product

Containing this text: alginate

☑ Ignore Case ☑ Allow Partial Match

And:

☐ Up to: bases long

And:

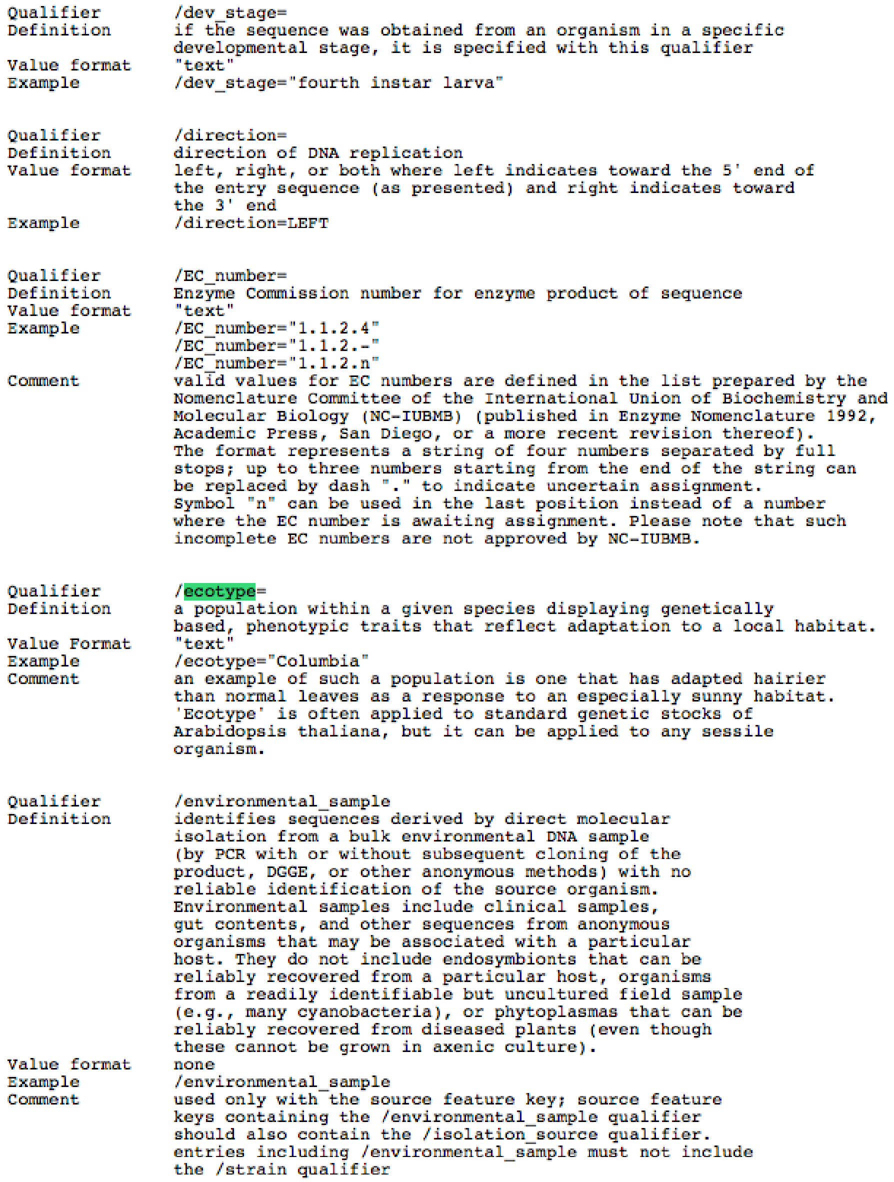
☐ At least: bases long

And by:

☐ Amino acid motif:

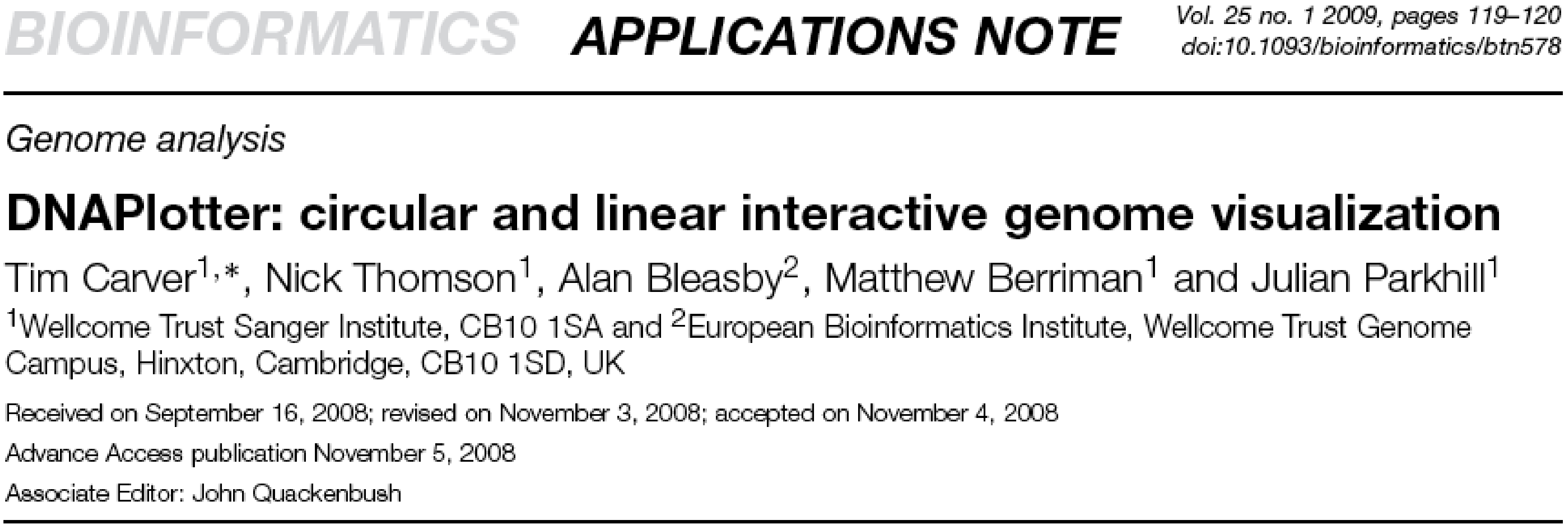
☑ Forward Strand Features ☑ Reverse Strand Features

http://pseudomonas-syringae.org/artemis_tutorial.htm



```
Qualifier       /dev_stage=
Definition      if the sequence was obtained from an organism in a specific
                developmental stage, it is specified with this qualifier
Value format    "text"
Example         /dev_stage="fourth instar larva"


Qualifier       /direction=
Definition      direction of DNA replication
Value format    left, right, or both where left indicates toward the 5' end of
                the entry sequence (as presented) and right indicates toward
                the 3' end
Example         /direction=LEFT


Qualifier       /EC_number=
Definition      Enzyme Commission number for enzyme product of sequence
Value format    "text"
Example         /EC_number="1.1.2.4"
                /EC_number="1.1.2.-"
                /EC_number="1.1.2.n"
Comment         valid values for EC numbers are defined in the list prepared by the
                Nomenclature Committee of the International Union of Biochemistry and
                Molecular Biology (NC-IUBMB) (published in Enzyme Nomenclature 1992,
                Academic Press, San Diego, or a more recent revision thereof).
                The format represents a string of four numbers separated by full
                stops; up to three numbers starting from the end of the string can
                be replaced by dash "." to indicate uncertain assignment.
                Symbol "n" can be used in the last position instead of a number
                where the EC number is awaiting assignment. Please note that such
                incomplete EC numbers are not approved by NC-IUBMB.


Qualifier       /ecotype=
Definition      a population within a given species displaying genetically
                based, phenotypic traits that reflect adaptation to a local habitat.
Value Format    "text"
Example         /ecotype="Columbia"
Comment         an example of such a population is one that has adapted hairier
                than normal leaves as a response to an especially sunny habitat.
                'Ecotype' is often applied to standard genetic stocks of
                Arabidopsis thaliana, but it can be applied to any sessile
                organism.


Qualifier       /environmental_sample
Definition      identifies sequences derived by direct molecular
                isolation from a bulk environmental DNA sample
                (by PCR with or without subsequent cloning of the
                product, DGGE, or other anonymous methods) with no
                reliable identification of the source organism.
                Environmental samples include clinical samples,
                gut contents, and other sequences from anonymous
                organisms that may be associated with a particular
                host. They do not include endosymbionts that can be
                reliably recovered from a particular host, organisms
                from a readily identifiable but uncultured field sample
                (e.g., many cyanobacteria), or phytoplasmas that can be
                reliably recovered from diseased plants (even though
                these cannot be grown in axenic culture).
Value format    none
Example         /environmental_sample
Comment         used only with the source feature key; source feature
                keys containing the /environmental_sample qualifier
                should also contain the /isolation_source qualifier.
                entries including /environmental_sample must not include
                the /strain qualifier
```

http://www.ebi.ac.uk/embl/Documentation/FT_definitions/feature_table.html



BIOINFORMATICS APPLICATIONS NOTE

Vol. 25 no. 1 2009, pages 119–120
doi:10.1093/bioinformatics/btn578

Genome analysis

DNAPlotter: circular and linear interactive genome visualization

Tim Carver[1,*], Nick Thomson[1], Alan Bleasby[2], Matthew Berriman[1] and Julian Parkhill[1]
[1]Wellcome Trust Sanger Institute, CB10 1SA and [2]European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, UK

Received on September 16, 2008; revised on November 3, 2008; accepted on November 4, 2008
Advance Access publication November 5, 2008
Associate Editor: John Quackenbush

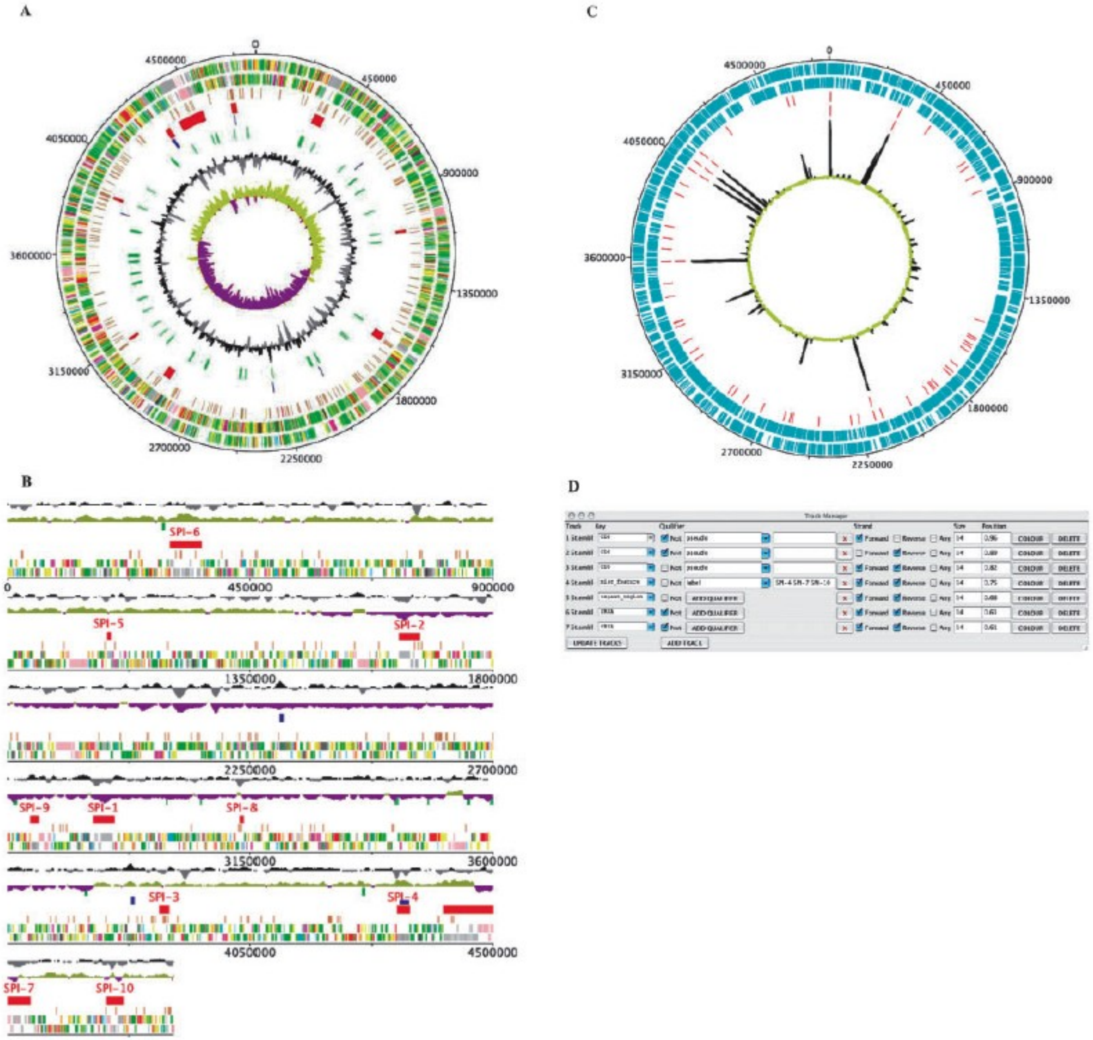
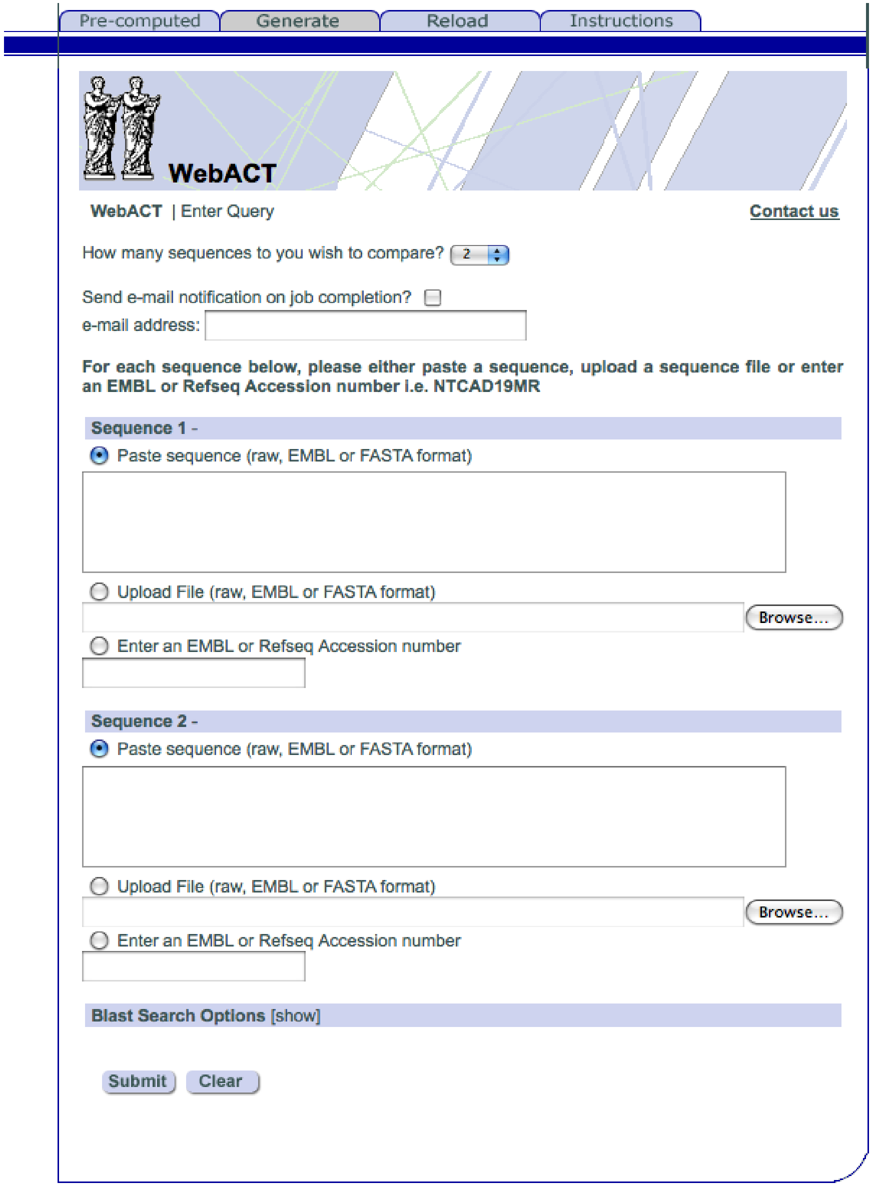


Fig. 1. (**A**, **B**) showing *Salmonella typhi* genome as a circular and linear plot, respectively. The tracks from the outside represent: (1) Forward CDS; (2) Reverse CDS; (3) Pseudogenes 4. Salmonella Pathogenicity Islands (red); (5) repeat regions (blue); (6) rRNA and tRNA (green); (7) %GC plot 8. GC skew [(GC)/(G+C)]. (**C**) A generated example showing a transcriptome graph (black and yellow) on a circular plot for a prokaryotic genome. The tracks from the outside represent: (1) Forward CDS; (2) Reverse CDS; (3) tRNA; (4) rRNA. (**D**) Snapshot of the track manager showing filtering criteria.

Carver et al. 2008



WebACT | Enter Query

How many sequences to you wish to compare? 2

Send e-mail notification on job completion? □

e-mail address:

For each sequence below, please either paste a sequence, upload a sequence file or enter an EMBL or Refseq Accession number i.e. NTCAD19MR

Sequence 1 -

● Paste sequence (raw, EMBL or FASTA format)

○ Upload File (raw, EMBL or FASTA format)

○ Enter an EMBL or Refseq Accession number

Sequence 2 -

● Paste sequence (raw, EMBL or FASTA format)

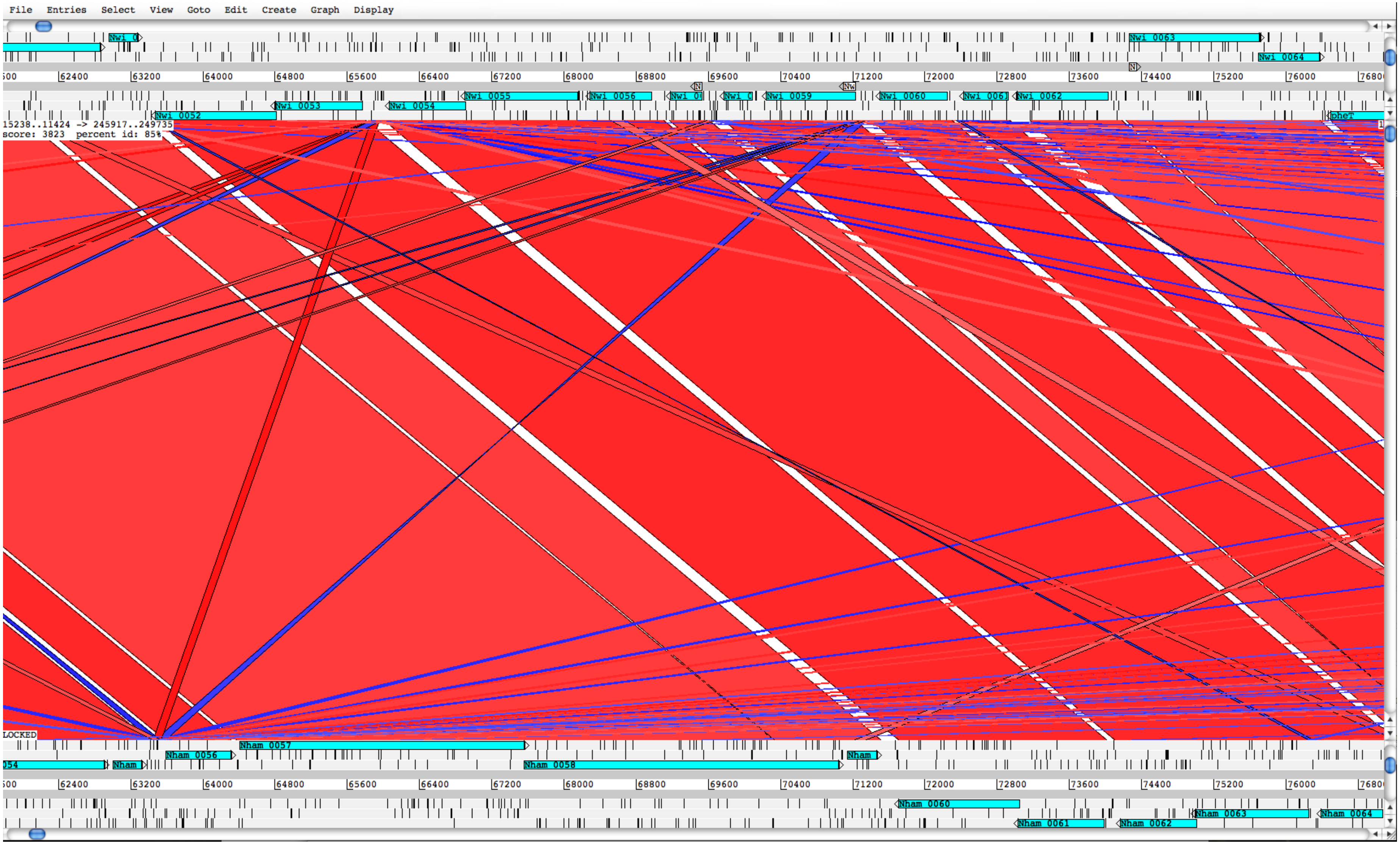
○ Upload File (raw, EMBL or FASTA format)

○ Enter an EMBL or Refseq Accession number

Blast Search Options [show]

Image from http://www.sanger.ac.uk/

WebACT Genome Comparison Visualization



```
15238..11424 -> 245917..249735
score: 3823  percent id: 85%
```

Image from http://www.sanger.ac.uk/