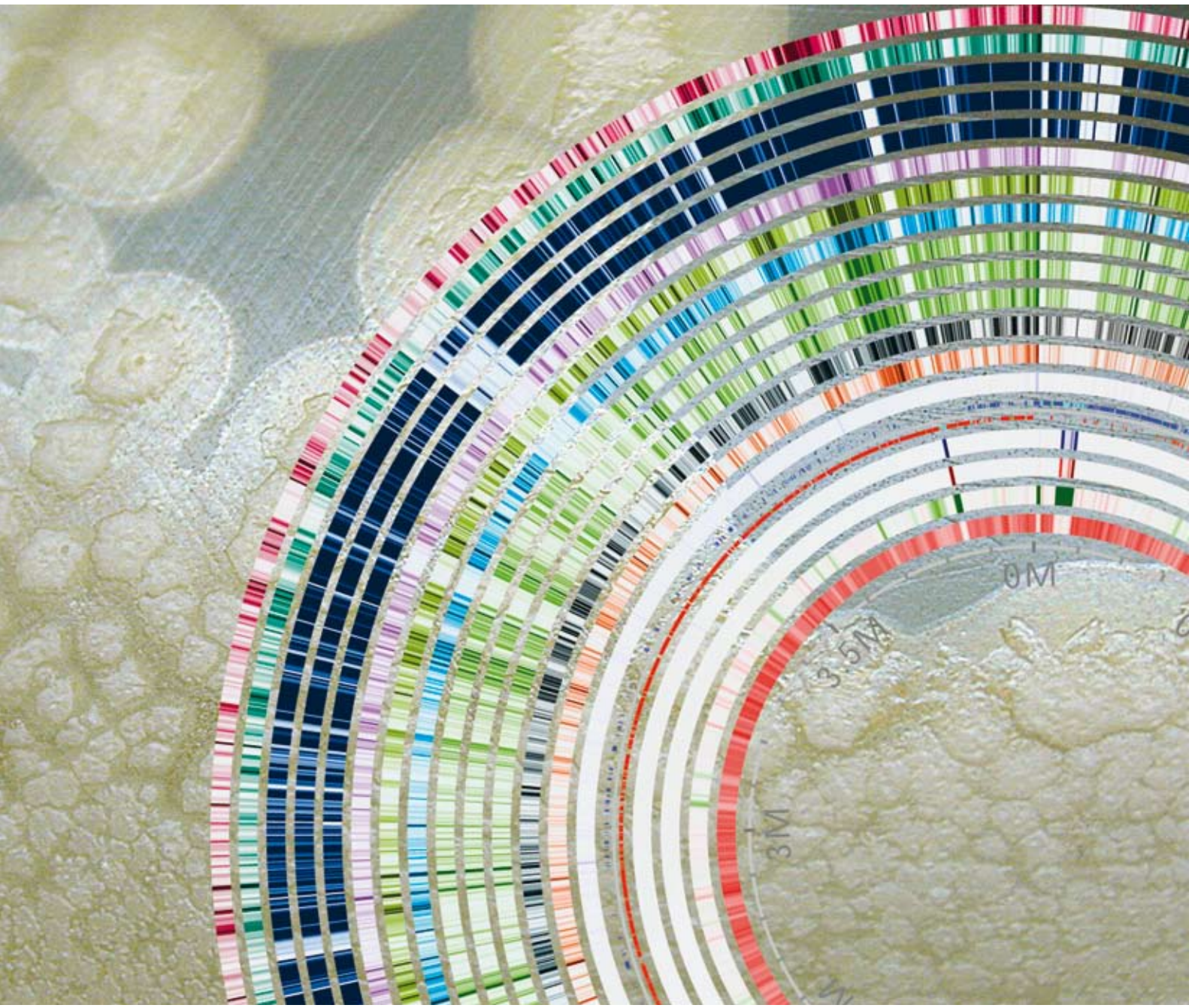
Molecular BioSystems

www.molecularbiosystems.org

Volume 4 | Number 5 | May 2008 | Pages 353–444



ISSN 1742-206X

HIGHLIGHT
Tim T. Binnewies *et al.*
The genome BLASTatlas—a GeneWiz extension for visualization of whole-genome homology

REVIEW
Eric C. Greene *et al.*
The importance of surfaces in single-molecule bioscience

RSC Publishing

1742-206X(2008)4:5;1-9

The genome BLASTatlas—a GeneWiz extension for visualization of whole-genome homology

Peter F. Hallin, Tim T. Binnewies* and David W. Ussery

DOI: 10.1039/b717118h

The development of fast and inexpensive methods for sequencing bacterial genomes has led to a wealth of data, often with many genomes being sequenced of the same species or closely related organisms. Thus, there is a need for visualization methods that will allow easy comparison of many sequenced genomes to a defined reference strain. The BLASTatlas is one such tool that is useful for mapping and visualizing whole genome homology of genes and proteins within a reference strain compared to other strains or species of one or more prokaryotic organisms. We provide examples of BLASTatlases, including the *Clostridium tetani* plasmid p88, where homologues for toxin genes can be easily visualized in other sequenced *Clostridium* genomes, and for a *Clostridium botulinum* genome, compared to 14 other *Clostridium* genomes. DNA structural information is also included in the atlas to visualize the DNA chromosomal context of regions. Additional information can be added to these plots, and as an example we have added circles showing the probability of the DNA helix opening up under superhelical tension. The tool is SOAP compliant and WSDL (web services description language) files are located on our website: (http://www.cbs.dtu.dk/ws/BLASTatlas), where programming examples are available in Perl. By providing an interoperable method to carry out whole genome visualization of homology, this service offers bioinformaticians as well as biologists an easy-to-adopt workflow that can be directly called from the programming language of the user, hence enabling automation of repeated tasks. This tool can be relevant in many pangenomic as well as in metagenomic studies, by giving a quick overview of clusters of insertion sites, genomic islands and overall homology between a reference sequence and a data set.

Center for Biological Sequence Analysis, Department of Systems Biology, The Technical University of Denmark, 2800 Lyngby, Denmark. E-mail: pfh@cbs.dtu.dk. E-mail: tim@cbs.dtu.dk. E-mail: dave@cbs.dtu.dk

Background

It has been more than 10 years since the sequencing of the first bacterial genome (ref. 1, US patent number 6,528,289), and currently sequence data are available for more than a thousand sequenced genomes.

With so many genome sequences, for several bacterial species multiple genome sequences exist; for example, at the time of writing, 10 different *Escherichia coli* genomes have been fully sequenced and published, and draft sequences for another 31 genomes are available, adding

Peter F. Hallin

Tim T. Binnewies

David W. Ussery

Peter F. Hallin was born in Odense, Denmark, and is currently a PhD student at CBS, DTU. Tim T. Binnewies grew up in Kiel, Germany, and obtained his PhD from the Technical University of Denmark, he is currently working for Roche Diagnostics AG in Switzerland. David W. Ussery was born and raised in Springdale, Arkansas. Since 1998, he has been leader for the Comparative Genomics group at CBS.

up to a total of 41 different *E. coli* genomes (according to the National Center for Biotechnology Information, NCBI Entrez, 12-Feb-2008). Table 1 lists the top 20 represented prokaryotic genera in terms of numbers of fully sequenced genomes based on recent counting in Entrez Genome Projects, although these numbers will change quickly as more genomes are being added on a regular basis. Thus, analysis of multiple genomes of the same organism (the "pangenome") is now possible, and as more metagenomic datasets are published (see for example the projects listed on the GOLD web pages[24]), there is a need for a graphical representation of how these new data compare to existing reference strains or model organisms.

We have developed a visualization method, called "BLASTatlas", for showing mapped alignments of BLAST searches of a reference sequence against one or more databases, onto the reference genome. Early implementation of a similar method[2–4] accounted for the statistical significance (*E*-value) of each hit, by color coding the expectation values [$-\log(E)$] of the alignment. This method gives a uniform color throughout the alignment (gene or protein) but shows no information about the amino acid conservation within regions of the alignment. At the level of a bacterial chromosome, this makes little difference, although when one zooms in at the level of individual genes, the older method of shading the entire gene based on the *E*-value gives no information about regions within a gene (such as functional domains) which might be strongly conserved, whilst other parts of the gene have little sequence homology within other genomes. We have refined the BLASTatlas method to map each individual amino acid residue or nucleotide back to the reference genome sequence from which the coding sequence was derived. Instead of colour-coding the significance of the entire hit, this method maps the conservation of the individual bases or amino acids. Tools such as the Artemis Comparison Tool (ACT)[5] allow detailed viewing of complete BLAST results, and this is an excellent graphical method for comparison of two genomes. ACT can also be extended to compare two genomes to a reference, placed in the middle. In contrast, the BLASTatlas method can compare many genomes to the same reference, and can provide a quick overview of chromosomal regions of gene conservation across many genomes.

As can be seen from Table 1, for many of the heavily sampled genera, there are further genome projects in the pipeline which will produce even more sequences than are currently available, and there is a need for methods for efficient comparison of these genomes, giving an overview of general trends in the data. The BLASTatlas allows the comparison of many genomes to a reference sequence. The current limit is about 60 genomes. There are two levels of comparison, the first represents a one-page map of the whole chromosome, and the second level zooming in a particular region of interest, allowing the visualization of regions of conservation within individual genes. The color-coding represents identical amino acids (or nucleic acids), based on a pairwise alignment of all protein coding regions, with the best matches for each gene in the reference genome shown. Thus, combining both levels, it is possible to get a global overview of the whole chromosome, and to then quickly identify gene conservation (or lack thereof) in regions of interest, at the level of conservation of individual amino acid residues.

Clostridium botulinum is an important human pathogen which is the causative agent of botulism, giving rise to fatal paralysis of the respiratory muscles, caused by botulinum neurotoxin (BoNT) which disrupts nerve functions. The genes encoding BoNT components are clustered on the bacterial chromosome (group I + II strains), on prophages (group III strains) or on plasmids (group IV strains). Group I strains encode type A, B and F type toxins, group II strains produce type B, E and F toxins and group III strains encode for type C and D toxins, whereas group IV strains produce type G toxin.[6] We use the BLASTatlas method to show the overall genome homology of the *C. botulinum* strain F Langeland, compared to all currently available and fully sequenced strains of the *Clostridium* genus.

Table 1 The number of species and NCBI Entrez Project IDs of the 20 most represented genera in the Entrez Genome Projects Database,[13] as accessed on 21 October 2007. The numbers in brackets show the counting of both ongoing and completed projects, whereas the first number reflects only the completed projects. Candidate genera have been excluded from this counting

Genus	Projects	Species
Streptococcus	26 [63]	8 [15]
Burkholderia	15 [55]	8 [15]
Bacillus	16 [48]	9 [16]
Clostridium	14 [43]	9 [22]
Vibrio	7 [35]	5 [14]
Mycobacterium	16 [30]	9 [14]
Salmonella	5 [30]	2 [3]
Listeria	4 [29]	3 [6]
Escherichia	10 [27]	1 [1]
Mycoplasma	13 [25]	11 [17]
Shewanella	14 [24]	10 [15]
Pseudomonas	13 [23]	7 [8]
Yersinia	9 [23]	3 [7]
Haemophilus	6 [23]	3 [4]
Staphylococcus	17 [22]	4 [5]
Synechococcus	10 [21]	2 [2]
Campylobacter	9 [20]	5 [9]
Francisella	7 [16]	1 [2]
Lactobacillus	11 [15]	10 [12]
Rickettsia	10 [15]	9 [12]

Methods

The BLASTatlas method uses all the provided annotated coding sequences (or proteins) of a reference genome, and compares each of those with one or more genomes. The total genome sequence for each organism is represented by a database and can contain any number of DNA or protein sequences. BLAST searches with a non-stringent *E*-value cut-off of 0.01 are used to identify the best alignments between the reference sequence protein and the database (genome) in question. Once identified, the single best pairwise alignment for

each of the reference sequences is obtained and included in the map.

The reference genome of a given comparison has a fixed size, whereas the sequences to be compared can be thought of as simply a ''pile of proteins'', ranging between the size from that of a small phage, to a single genome, or an entire metagenomic sample or even existing large BLAST databases, such as UniProt. It is important to emphasize that each protein in the reference genome is compared to all the proteins in the query set—regardless of orientation or location. The BLASTatlas method uses the software BLASTALL v. 2.2.11 for the search, and in BLAST terminology, the reference genome constitutes the 'query' whereas each other genome (*e.g.*, a lane or circle in the atlas) in the comparison corresponds to the 'database'. We define a lane as a visual representation of mapped database hits (individual residue matches) on to the reference genome. A lane can have a boxfilter (smoothing) applied within each of the smallest visible units of the atlas (the resolution of the graphical representation). A single BLASTatlas may contain several lanes; currently around 60 circles is the upper limit.

The input requires a file containing the genome sequence, including all annotated coding sequences (comprising protein-start, -stop and -direction) for the reference genome. The four programs 'BLASTp', 'BLASTn', 'BLASTx', and 'tBLASTn' can be used for each lane of the BLASTatlas, although of course the appropriate sequences (DNA or protein) must be provided. For example, when using ' BLASTn' or 'tBLASTn' in a lane, the required DNA sequence can be a set of open reading frames (ORFs), chromosomal contigs, entire genome sequences or even environmental (metagenomic) samples. In a pairwise fashion, the sequence of the reference is BLASTed against each database defined by the user, employing the specified BLAST algorithm.

Interpretation of BLAST alignments

For each of the sequences defined in the reference, only the best hit in each database is stored. For these hits, the alignments are mapped on to the reference genome. When aligning two DNA sequences, the map shows one of four possible states for each position: *match*, *mismatch*, *gap in query* (reference genome), and *gap in database* (lane). Only the match contributes to the overall score with a value of 1, whereas mismatches and gaps in the database get a score value of zero. When aligning two protein sequences, an additional state is introduced for conservative mismatches, indicating that two amino acids have similar physical–chemical properties; such a state will receive a score of 0.5. Match and gap states of protein alignments are defined similar to those of the DNA alignments. The occurrence of gaps in the reference sequence do not get a corresponding coordinate and are therefore ignored (see Fig. 1). In the BLASTatlas context, a map is an array of match scores. The array has the same length as the reference genome, with each position along the gene having a value of 0, 0.5 or 1: It should be noted that intergenic regions (and ncRNAs, including tRNAs and rRNAs) have values of 0, because BLASTatlases only compare protein encoding genes. We use this as a control, checking to make sure that the rRNA operons are visualized as ''gaps'' throughout all the lanes, for example. For each database defined, there will be a corresponding BLAST map within the atlas (see Fig. 2). Each database entry of the BLAST searches must contain a legend text for the lane, a colour code range and a scaling method. For the colours, an upper and lower colour is required, whereas the middle colour is usually grey; all colours are defined in RGB integers ranging from 0 to 10. The scale can be either fixed, such as ranging from 0 to 1, or scaled using any number of standard deviations around the average.

DNA properties

The BLASTatlas method allows users to add structural as well as base composition information to the atlas by using the 'DNAparameters' element in the request. These properties can be for example DNA structural properties,[7] such as intrinsic curvature,[8] global or local repeats[9] or other measures of base composition.[10] A list of possible different properties currently pre-computed can be obtained *via* the online documentation and type declarations of the web services description. The DNA property lanes are usually added near the center (or at the lowest part when seen from the outermost circle) of the atlas.

Custom properties

In addition to the standard DNA properties and BLAST maps, the web service provides a method for adding individual customer data for example gene expression values to the atlas, using the 'customMap' element in the request. Data must be provided in the form of comma separated strings, with each position in the list corresponding to the genomic position. When defining custom data lanes, the colour ranges, scaling method, and legend text must be provided.

Visualization

Details such as the atlas title and the geometry (linear or circle representation) are necessary for the final visualization. Once the BLAST searches are carried out and remapped to the reference

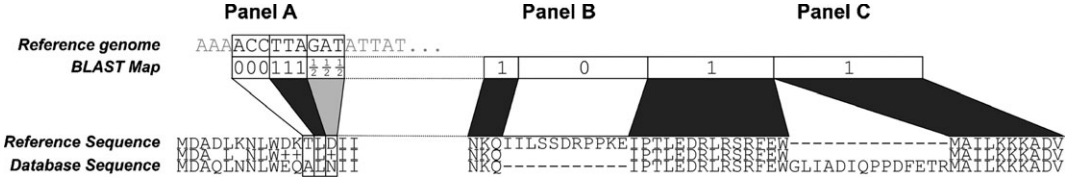


Fig. 1 Mapping of protein–protein alignment to DNA. **Panel A**: mismatches and perfect matches are assigned a score of 0 and 1, respectively. Conservative mismatches are assigned a score of 0.5. In the case of DNA alignment, only scores of 0 and 1 are possible. **Panel B**: gaps in the database sequence will be rendered as being non-conserved areas (filled with zeros). **Panel C**: gaps in the reference sequence will be neglected, since they have no corresponding region in the reference genome into which they can be mapped.

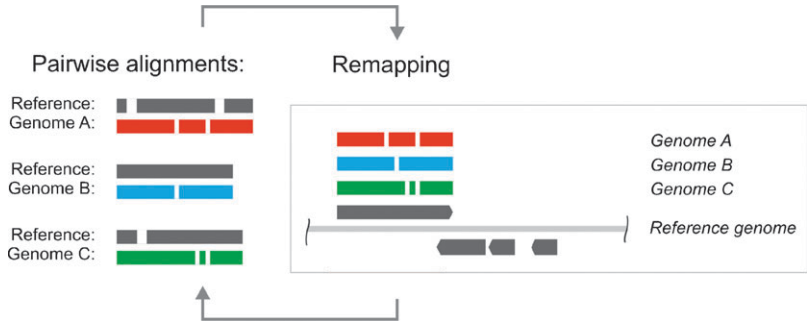


Fig. 2 Genes (or segments) from each genome are compared with a reference gene, as shown in the left panel; a pairwise comparison is made using one of the BLAST algorithms. On the right is shown the "remapping", or the representation of each of the BLAST runs on the left, mapped onto the chromosomal sequence. Note that gaps in the reference gene (grey) are not included in the colored maps of the atlas.

genome and custom data and DNA properties are collected, an XML configuration file is composed which contains all these data and the layout of the atlas. This file is then sent to the GeneWiz[7] software which produces a PostScript document, it then is base64 encoded to allow transport *via* XML. This part of the process takes place on the server and requires no user-interaction. An example atlas of a plasmid is shown in Fig. 3, and will be discussed in more detail below.

Web services implementation

A WSDL (web services description language) file is written which describes the operations (runAtlas, pollQueue, fetchAtlasResult) and the input requirements for them. The file can be downloaded. All input/output objects are defined in a separated XSD file (XML schema definition) within the WSDL file, which comprises information and type restrictions applicable in the request. This serves as documentation of the objects as well as a way to validate a request before it is submitted. Unfortunately, the validation supports only Perl modules for now that is not optimal yet, whereas this option is well implemented in tools like soapUI (http://www.soapui.org/). It should be stressed that users should, until better validation support can be implemented, be careful to correctly format the input parameters before sending the request.

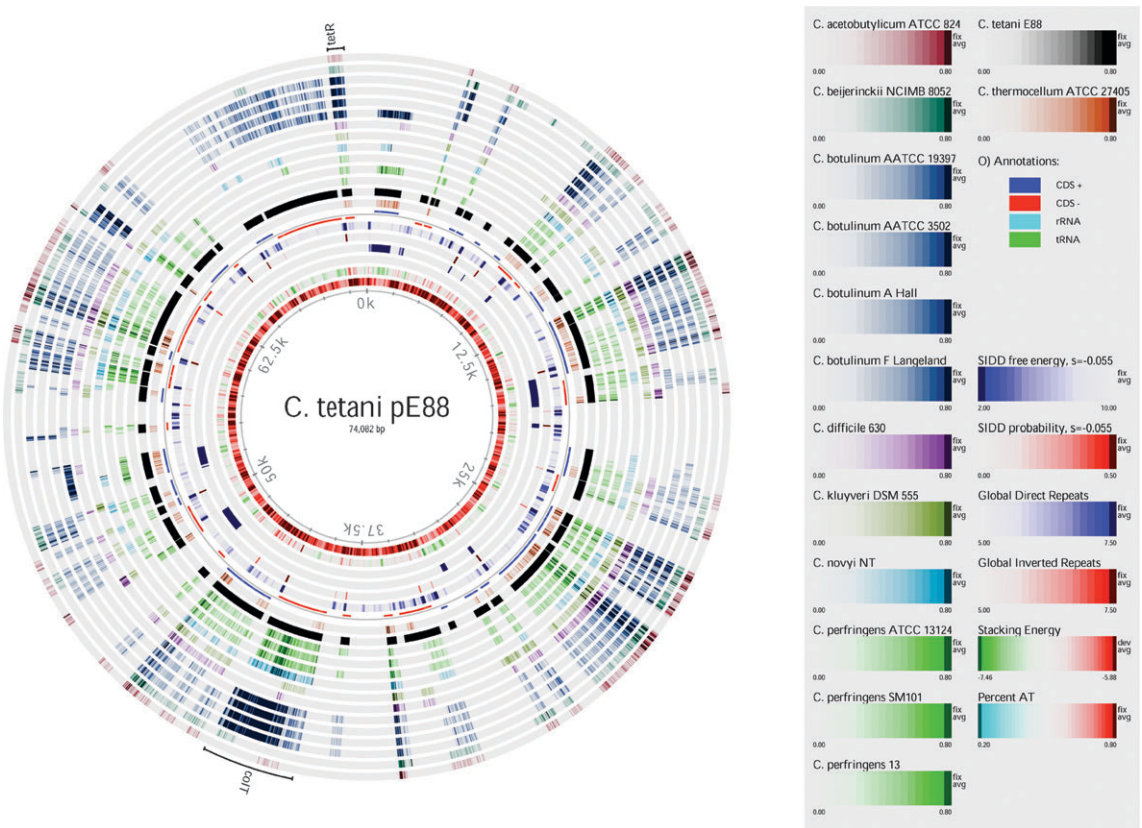


Fig. 3 BLASTatlas of pE88—a small plasmid of *Clostridium tetani* strain E88, GenBank accession number AF528097. DNA parameters percent AT, GC skew, global direct repeats, and global inverted repeats are included in the inner most lanes. BLAST lanes of all complete genome sequences of the *Clostridium* genomes (see Table 1), including plasmids are included in the outer most lanes. As examples of custom lanes, the free energy (*G*, blue kcal mol^{-1}) and the probability (*P*, red) measures of stress induced DNA duplex destabilization (SIDD) sites are included in the lanes between the DNA properties and the BLAST lanes.[23] SIDD calculations were obtained from the SIDDbase WebService (http://www.cbs.dtu.dk/ws/SIDDbase). The request XML used to construct this plot can be downloaded from the example section of the service homepage, http://www.cbs.dtu.dk/ws/BLASTatlas. As expected, there is full homology of all coding regions between the plasmids and all replicons of *C. tetani* E88 (black lane just outside of the annotations); however there appears to be limited conservation of these pE88 genes throughout the genomes for other *Clostridium* strains.

Table 2 A list of all strains and their accession numbers used in this comparison. Each row represents the NCBI Entrez sequencing project. The number of base pairs and protein coding genes are those derived as the sum within each project. *C. botulinum* str. F Langeland is that used as reference of the comparison

Species	Segments	Size	Proteins
C. acetobutylicum ATCC 824[14]	Entrez Project 77: **Chromosome: AE001437**, Plasmid pSOL1: AE001438	4.132.880	3.848
C. beijerinckii NCIMB 8052 (unpublished)	Entrez Project 12637: **Chromosome: CP000721**	6.000.632	5.020
C. botulinum A str. ATCC 19397 (unpublished)	Entrez Project 19517: **Chromosome: CP000726**	3.863.450	3.552
C. botulinum A str. ATCC 3502[6]	Entrez Project 193: **Chromosome: AM412317**, Plasmid pBOT3502: AM412318	3.903.260	3.671
C. botulinum A str. Hall (unpublished)	Entrez Project 19521: **Chromosome: CP000727**	3.760.560	3.407
***C. botulinum* F str. (unpublished)**	**Entrez Project 19519:** Chromosome: CP000728**, Plasmid pCLI: CP000729**	**4.012.918**	**3.659**
C. difficile 630[15]	Entrez Project 78: **Chromosome: AM180355**, Plasmid pCD630: AM180356	4.298.133	3.787
C. kluyveri DSM 555 (unpublished)	Entrez Project 19065: **Chromosome: CP000673**, Plasmid pCKL555A: CP000674	4.023.800	3.913
C. novyi NT[16]	Entrez Project 16820: **Chromosome: CP000382**	2.547.720	2.325
C. perfringens ATCC 13124[25]	Entrez Project 304: **Chromosome: CP000246**	3.256.683	2.876
C. perfringens SM101[17]	Entrez Project 12521: **Chromosome: CP000312**, Plasmid 1: CP000313, Plasmid 2: CP000314, Viral segment phage phiSM101: CP000315	2.960.088	2.631
C. perfringens str. 13[18]	Entrez Project 79: **Chromosome: BA000016**, Plasmid pCP13: AP003515,	3.085.740	2.723
C. tetani E88[19]	Entrez Project 81: **Chromosome: AE015927**, Plasmid pE88: AF528097	2.873.333	2.432
C. thermocellum ATCC 27405 (unpublished)	Entrez Project 314: **Chromosome: CP000568**	3.843.301	3.191
Clostridium phage[20]	**Phage c-st**: AP008983	185.683	198

Web services workflow

A workflow was written in Perl (v5.8.7), employing SOAP:Lite (v0.69) which reads the FASTA files of the database strains listed in Table 3 and produces a BLASTatlas using the *C. botulinum* strain F Langeland as reference. The script uses the online web service (see Fig. 4). The BLASTatlas figure produced by this workflow is seen in Fig. 5.

Results

Fig. 3 represents a BLASTatlas for plasmid pE88 from *Clostridium tetani* strain E88. The homology for genes in the plasmid to other sequenced genomes is shown in the circles, additional "custom lanes" represent chromosomal regions predicted to open under superhelical stress. The chromosomal location of the genes encoding colT and tetR are labelled in the figure. Notice that these two proteins contain regions of homology that are found in most of the *Clostridium* proteomes searched. Since the *C. tetani* plasmid is included in the genome sequence (black circle in the figure), all the genes are found in this genome (solid black), and most of the other *Clostridium* proteomes contain some weak homology but in general lack most of the plasmid-encoded genes. Thus, this is a quick overview of gene conservation of a plasmid compared to many sequenced genomes of the same genera.

To demonstrate this for an entire bacterial genome (which is millions of bp in size, compared to a small/~75 000 bp plasmid, shown in Fig. 3), we have used the genome sequence of *C. botulinum* strain F Langeland, the largest of the *C. botulinum* genomes, to build a protein BLASTatlas of all publicly available fully sequenced *Clostridia* genomes, including all chromosomes, plasmids and phages (see Fig. 5). Each lane of the atlas corresponds to a sequencing project that contains the main chromosome plus any

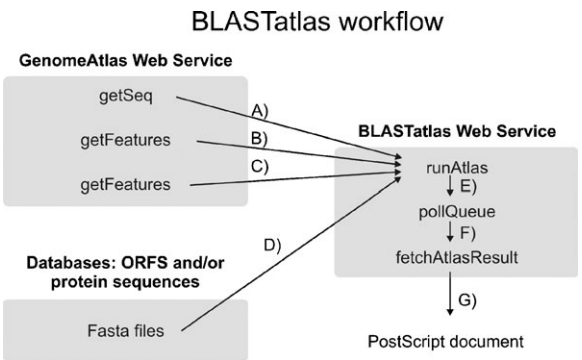


Fig. 4 Workflow description: a Perl script was written for handling the assembly of the SOAP envelope and contacting various other web services operations: (A) **obtaining genomes sequence**: using the *getSeq* operation of the GenomeAtlas Web Services (v.3.3), the genome sequence of the reference genome is obtained as one continuous string. (B) **Obtaining atlas annotations**: annotated CDS, rRNA, and tRNA features of the GenBank record of the reference genome using the *getFeatures* operation—these are the features which will be printed in a separate lane on the atlas. (C) **Obtaining ORF annotations of the reference genome**: again, using the *getFeatures* operation, all codon sequences and their translations are obtained. (D) **Obtain databases**: read FASTA files containing proteins and ORFs of the database genomes to be added as lanes. The output of A–F are assembled into a single SOAP request, including configurations of the atlas. (E) **Polling the queue**: once the job has been submitted, a 32 character hex string is returned for identifying the job, which can be used by operation *pollQueue* to see the status of the job. (F + G) **Obtaining result**: once a status "FINISHED" is obtained from *pollQueue*, the job id can submitted to *fetchResult* and the resulting PostScript image is returned.

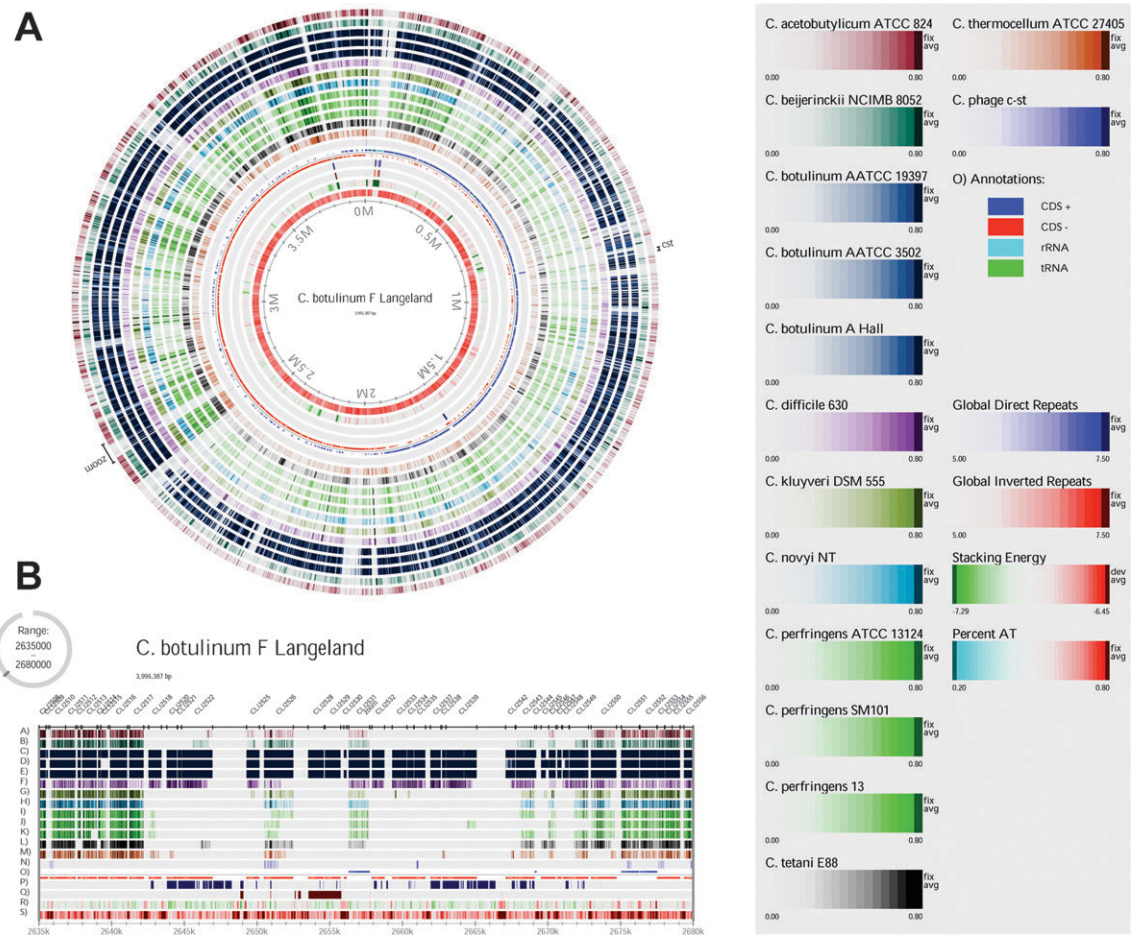


Fig. 5 BLASTatlas of *Clostridium botulinum* F strain Langeland: Lanes show genome homology of (starting from the outermost lane): *C. acetobutylicum* ATCC 824, *C. beijerinckii* NCIMB 8052, *C. botulinum* A str. ATCC 19397, *C. botulinum* A ATCC 3502, *C. botulinum* A str. Hall, *C. difficile* 630, *C. kluyveri* DSM 555, *C. novyi* NT, *C. perfringens* ATCC 13124, *C. perfringens* SM101, *C. perfringens* str. 13, *C. tetani* E88, *C. thermocellum* ATCC 27405, and *Clostridium* phage c-st genome. Inside of the annotation circle are shown global direct repeats, global inverted repeats, stacking energy, and percent AT. Blue and red annotations are coding sequences on plus and minus strand, whereas green and turquoise are rRNA and tRNA, genes respectively. The two toxin components NTNH and BoNT/A1 that are identified on phage c-st are present in the reference genome at positions 880 kb and 883 kb, respectively (marked 'cst'). The presence of the two is visible as a thin blue band on the c-st blast lane. The lower part of the figure shows a zoom of the region around 2635 kb, providing an example of a gene cluster which appears to be conserved throughout the *C. botulinum* strains and partly within the *C. difficile* 630.

phages or plasmids present in the genome. The proteins encoded by the 185 kb neurotoxin-converting bacteriophage c-st are labelled, as well as a region which is zoomed in the second panel in Fig. 5. The accession numbers, total size and total number of genes within each lane can be seen in Table 2.

There are several items of interest which can be seen in Fig. 5. First, the rRNA operons can be quite readily seen, near the top part of the chromosome map, labeled turquoise; these rRNA operons are more GC rich (hence less red in the inner-most lane), have direct and inverted repeats (the next two lanes), and are not shown in the proteome comparison lanes (since these genes do not encode proteins).

As expected, the circle representing the c-st phage shows little match for most of the *C. botulinum* genome, at the protein level. In general, the two other *C. botulinum* genomes (both in blue) have the highest similarity to the reference *C. botulinum* genome (also shown as a circle). In this case it is used as an internal control: all of the proteins should show a match for this lane, since the reference genome is blasted against itself. Another interesting observation is the upper-left-hand part of the genome which seems to have more homology to other *Clostridium* genomes, in particular showing many matches to the *C. perfringens* genomes (green circles), compared to the rest of the genome.

Application in metagenomics

The genera of *Prochlorococcus* belongs to the cyanobacteria and is one of the most abundant photosynthetic organisms of the ocean. It plays an important role in the planet's carbon cycle and has adapted to the various light and oxygen conditions present at the various depths.[11] As of the end of January 2008, eleven *Prochlorococcus marinus* genomes are publicly available and we have included all encoded proteins of these data with the seven metagenomic read collections from the ALOHA station near Hawaii,[12] as shown in Table 3. The strain of *P. marinus* strain MIT 9303 has the largest genome of all

Table 3 A list of all strains/sample names and their accession numbers used in the metagenomic comparison. The list is sorted by sampling depth

Source	Size	Origin	Accession/sample	Ref.	Depth
P. marinus str. MIT 9515	1 704 176 (1906 proteins)	Tropical Pacific	CP000552	Unpublished	Surface
P. marinus str. MIT 9215	1 738 790 (1983 proteins)	Equatorial Pacific	CP000825	Unpublished	Surface
P. marinus str. MED4	1 657 990 (1936 proteins)	Mediterranean Sea	BX548174	21	4 m
JGI_SMPL_HF10_10-07-02	7 482 668 (7842 contigs)	North Pacific Subtropical Gyre	—	12	10 m
P. marinus str. NATL1A	1 864 731 (2193 proteins)	North Atlantic	CP000553	Unpublished	30 m
P. marinus str. NATL2A	1 842 899 (2163 proteins)	North Atlantic	CP000095	Unpublished	30 m
P. marinus str. AS9601	1 669 886 (1921 proteins)	Arabian Sea	CP000551	Unpublished	50 m
JGI_SMPL_HF70_10-07-02	10 828 386 (10 999 contigs)	North Pacific Subtropical Gyre	—	12	70 m
P. marinus str. MIT 9211	1 688 963 (1855 proteins)	Equatorial Pacific	CP000878	21	83 m
P. marinus str. MIT 9301	1 641 879 (1907 proteins)	Sargasso Sea	CP000576	Unpublished	90 m
P. marinus str. MIT 9303	2 682 675 (2997 proteins)	Sargasso Sea	CP000554	Unpublished	100 m
P. marinus str. SS120	1 751 080 (1882 proteins)	Sargasso Sea	AE017126	22	120 m
JGI_SMPL_HF130_10-06-02	6 091 784 (6812 contigs)	North Pacific Subtropical Gyre	—	12	130 m
***P. marinus* str. MIT 9312**	**1 709 204 (1962 proteins)**	**Equatorial Pacific**	**CP000111**	**Unpublished**	**135 m**
P. marinus str. MIT MIT9313	2 410 873 (2273 proteins)	Gulf Stream	BX548175	21	135 m
JGI_SMPL_HF200_10-06-02	7 829 659 (8286 contigs)	North Pacific Subtropical Gyre	—	12	200 m
JGI_SMPL_HF500_10-06-02	8 764 642 (9027 contigs)	North Pacific Subtropical Gyre	—	12	500 m
JGI_SMPL_HF770_12-21-03	11 811 597 (11 479 contigs)	North Pacific Subtropical Gyre	—	12	770 m
JGI_SMPL_HF4000_12-21-03	11 028 821 (11 229 contigs)	North Pacific Subtropical Gyre	—	12	4000 m

currently available sequences (2.7 Mb) and was therefore used as reference in this comparison. BLAST hits between the reference and the encoded proteins of all the *P. marinus* genomes included were generated with the BLASTp algorithm, whereas hits between the reference proteins and the DNA reads of the metagenomic samples were generated using the tBLASTn algorithm. tBLASTn was used to avoid the gene prediction step of the metagenomic samples and to allow a rough estimate of the coding potential of these samples. All lanes are sorted according to the water depth at which the samples were collected (see Fig. 6). The Perl code for constructing this plot using web services is provided on the service homepage.

Discussion

The BLASTatlas method can assist biologists in finding regions along the chromosome which are conserved (or not). This information is useful for several

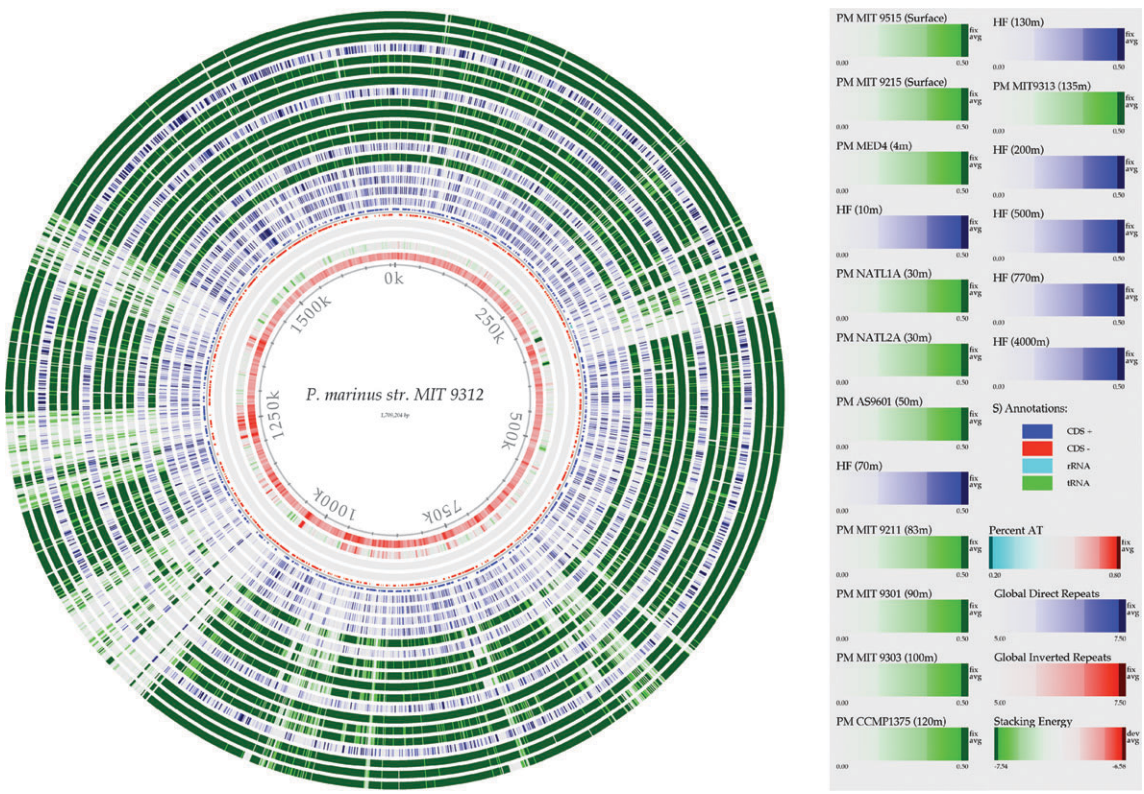


Fig. 6 BLASTatlas showing fully sequenced *Prochlorococcus* genomes (green) and the seven ALOHA metagenomic samples (blue). Outermost lanes represent samples closer to the ocean surface.

different applications, such as identifying phage insertion sites and loss of important genetic material. This method is even able to scale down to each individual nucleotide or amino acid residue. However, it is unable to deal with sequences (or parts thereof) that are not found in the reference genome. A good compromise when dealing with this issue is often to use the largest chromosome of a species as reference; in addition, it can be useful to rebuild the maps using different reference genomes. Besides this limitation, the fact that all coordinates are mapped back to the reference causes the coordinates of the database genomes to "get lost" in that only the best match is displayed, regardless of the chromosomal location in the database genomes. Other aspects of genome homology like gene synteny cannot effectively be answered by this tool. However, it is possible to use an additional circle to plot gene order conservation along the chromosome.

Currently, we see the BLASTatlas as an intermediate stage in analysis of many genomes of similar species. Soon there will be a need to compare hundreds or thousands of genome sequences, and the need for development of new methods for comparison of even larger numbers of genomes (hundreds or thousands) is ever more important.

Acknowledgements

The authors would like to thank Hans Henrik Stærfeld for assistance with server side programs and Kristoffer Rapacki for assistance on web services data types. The work was supported by a grant from the European Union through the EMBRACE network of Excellence, contract number LSHG-CT-2004-512092 and a grant from the Danish Center for Scientific Computing (DCSC).

References

1 R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick, J. McKenney, G. Sutton, W. FitzHugh, C. Fields, J. D. Gocyne, J. Scott, R. Shirley, L. I. Liu, A. Glodek, J. M. Kelley, J. F. Weidman, C. A. Phillips, T. Spriggs, E. Hedblom, M. D. Cotton, T. R. Utterback, M. C. Hanna, D. T. Nguyen, D. M. Saudek, R. C. Brandon, L. D. Fine, J. L. Fritchman, J. L. Fuhrmann, N. S. M. Geoghagen, C. L. Gnehm, L. A. McDonald, K. V. Small, C. M. Fraser, H. O. Smith and J. C. Venter, Whole-Genome Random Sequencing and Assembly of Haemophilus Influenzae Rd., *Science*, 1995, **269**(5223), 496–512.

2 L. J. Jensen, M. Skovgaard, T. Sicheritz-Ponten, M. K. Jorgensen, C. Lundegaard, C. C. Pedersen, N. Petersen and D. Ussery, Analysis of two large functionally uncharacterized regions in the Methanopyrus kandleri AV19 genome, *BMC Genomics*, 2003, **4**, 12.

3 L. J. Jensen, M. Skovgaard, T. Sicheritz-Ponten, N. T. Hansen, H. Johansson, M. K. Jørgensen, K. Kiil, P. F. Hallin and D. Ussery, Comparative genomics of four *Pseudomonas* species, in *The Pseudomonads Vol. I. Genomics, Life Style and Molecular Architecture*, ed. J. L. Ramos, Kluwer Academic/Plenum Publishers, New York, 2004, ch. 5, pp. 139–164.

4 P. F. Hallin, T. T. Binnewies and D. W. Ussery, Genome update: chromosome atlases, *Microbiology (Reading, U. K.)*, 2004, **150**, 3091–3093.

5 T. J. Carver, K. M. Rutherford, M. Berriman, M. A. Rajandream, B. G. Barrell and J. Parkhill, ACT: the Artemis Comparison Tool, *Bioinformatics*, 2005, **21**, 3422–3423.

6 M. Sebaihia, M. W. Peck, N. P. Minton, N. R. Thomson, M. T. Holden, W. J. Mitchell, A. T. Carter, S. D. Bentley, D. R. Mason, L. Crossman, C. J. Paul, A. Ivens, M. H. Wells-Bennik, I. J. Davis, A. M. Cerdeno-Tarraga, C. Churcher, M. A. Quail, T. Chillingworth, T. Feltwell, A. Fraser, I. Goodhead, Z. Hance, K. Jagels, N. Larke, M. Maddison, S. Moule, K. Mungall, H. Norbertczak, E. Rabbinowitsch, M. Sanders, M. Simmonds, B. White, S. Whithead and J. Parkhill, Genome sequence of a proteolytic (Group I) *Clostridium botulinum* strain Hall A and comparative analysis of the clostridial genomes, *Genome Res.*, 2007, **17**, 1082–1092.

7 A. G. Pedersen, L. J. Jensen, S. Brunak, H. H. Staerfeldt and D. W. Ussery, A DNA structural atlas for *Escherichia coli*, *J. Mol. Biol.*, 2000, **299**, 907–930.

8 E. S. Shpigelman, E. N. Trifonov and Bolshoy, A Curvature: software for the analysis of curved DNA, *CABIOS, Comput. Appl. Biosci.*, 1993, **9**, 435–440.

9 M. Skovgaard, L. J. Jensen, C. Friis, H. H. Stærfeldt, P. Worning, S. Brunak and D. Ussery, The Atlas Visualisation of Genome-wide Information, *Methods Microbiol.*, 2002, **33**, 49–63.

10 L. J. Jensen, C. Friis and D. W. Ussery, Three Views of Microbial Genomes, *Res. Microbiol.*, 1999, **150**, 773–777.

11 M. B. Sullivan, M. L. Coleman, P. Weigele, F. Rohwer and S. W. Chisholm, Three Prochlorococcus cyanophage Genomes: Signature Features and Ecological Interpretations, *PLoS Biol.*, 2005, **3**, e144; PMID: 15828858 [PubMed—indexed for MEDLINE].

12 E. F. DeLong, C. M. Preston, T. Mincer, V. Rich, S. J. Hallam, N.-U. Frigaard, A. Martinez, M. B. Sullivan, R. Edwards, B. R. Brito, S. W. Chisholm and D. M. Karl, Community Genomics Among Stratified Microbial Assemblages in the Ocean's Interior, *Science*, 2006, **311**(5760), 496–503.

13 D. L. Wheeler, T. Barrett, D. A. Benson, S. H. Bryant, K. Canese, V. Chetvernin, D. M. Church, M. DiCuccio, R. Edgar, S. Federhen, L. Y. Geer, Y. Kapustin, O. Khovayko, D. Landsman, D. J. Lipman, T. L. Madden, D. R. Maglott, J. Ostell, V. Miller, K. D. Pruitt, G. D. Schuler, E. Sequeira, S. T. Sherry, K. Sirotkin, A. Souvorov, G. Starchenko, R. L. Tatusov, T. A. Tatusova, L. Wagner and E. Yaschenko, Database Resources of the National Center for Biotechnology Information, *Nucleic Acids Res.*, 2007, **35**, D5–D12.

14 J. Nolling, G. Breton, M. V. Omelchenko, K. S. Makarova, Q. Zeng, R. Gibson, H. M. Lee, J. Dubois, D. Qiu, J. Hitti, Y. I. Wolf, R. L. Tatusov, F. Sabathe, L. Doucette-Stamm, P. Soucaille, M. J. Daly, G. N. Bennett, E. V. Koonin and D. R. Smith, Genome Sequence and Comparative Analysis of the Solvent-producing Bacterium *Clostridium acetobutylicum*, *J. Bacteriol.*, 2001, **183**, 4823–4838.

15 M. Sebaihia, B. W. Wren, P. Mullany, N. F. Fairweather, N. Minton, R. Stabler, N. R. Thomson, A. P. Roberts, A. M. Cerdeno-Tarraga, H. Wang, M. T. Holden, A. Wright, C. Churcher, M. A. Quail, S. Baker, N. Bason, K. Brooks, T. Chillingworth, A. Cronin, P. Davis, L. Dowd, A. Fraser, T. Feltwell, Z. Hance, S. Holroyd, K. Jagels, S. Moule, K. Mungall, C. Price, E. Rabbinowitsch, S. Sharp, M. Simmonds, K. Stevens, L. Unwin, S. Whithead, B. Dupuy, G. Dougan, B. Barrell and J. Parkhill, The Multidrug-resistant Human Pathogen *Clostridium difficile* has a Highly Mobile: Mosaic Genome, *Nat. Genet.*, 2006, **38**, 779–786.

16 C. Bettegowda, X. Huang, J. Lin, I. Cheong, M. Kohli, S. A. Szabo, X. Zhang, L. A. Diaz, Jr, V. E. Velculescu, G. Parmigiani, K. W. Kinzler, B. Vogelstein and S. Zhou, The Genome and Transcriptomes of the Anti-tumor Agent Clostridiumnovyi-NT, *Nat. Biotechnol.*, 2006, **24**, 1573–1580.

17 G. S. Myers, D. A. Rasko, J. K. Cheung, J. Ravel, R. Seshadri, R. T. DeBoy, Q. Ren, J. Varga, M. M. Awad, L. M. Brinkac, S. C. Daugherty, D. H. Haft, R. J. Dodson, R. Madupu, W. C. Nelson, N. J. Rosovitz, S. A. Sullivan, H. Khouri, G. I. Dimitrov, K. L. Watkins, S. Mulligan, J. Benton, D. Radune, D. J. Fisher, H. S. Atkins, T. Hiscox, B. H. Jost, S. J. Billington, J. G. Songer, B. A. McClane, R. W. Titball, J. I. Rood, S. B. Melville and I. T. Paulsen, Skewed Genomic Variability in Strains of the Toxigenic Bacterial Pathogen, *Clostridium perfringens*, *Genome Res.*, 2006, **16**, 1031–1040.

18 T. Shimizu, K. Ohtani, H. Hirakawa, K. Ohshima, A. Yamashita, T. Shiba, N. Ogasawara, M. Hattori, S. Kuhara and H. Hayashi, Complete Genome Sequence of *Clostridium perfringens*, an Anaerobic Flesh-eater, *Proc. Natl. Acad. Sci. U. S. A.*, 2002, **99**, 996–1001.

19 H. Bruggemann, S. Baumer, W. F. Fricke, A. Wiezer, H. Liesegang, I. Decker,

C. Herzberg, R. Martinez-Arias, R. Merkl, A. Henne and G. Gottschalk, The Genome Sequence of Clostridium tetani, the Causative Agent of Tetanus Disease, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 1316–1321.

20 Y. Sakaguchi, T. Hayashi, K. Kurokawa, K. Nakayama, K. Oshima, Y. Fujinaga, M. Ohnishi, E. Ohtsubo, M. Hattori and K. Oguma, The Genome Sequence of *Clostridium botulinum* Type C Neurotoxin Converting Phage and the Molecular Mechanisms of Unstable Lysogeny, *Proc. Natl. Acad. Sci. U. S. A.*, 2005, **102**, 17472–17477.

21 G. Rocap, F. W. Larimer, J. Lamerdin, S. Malfatti, P. Chain, N. A. Ahlgren, A. Arellano, M. Coleman, L. Hauser, W. R. Hess, Z. I. Johnson, M. Land, D. Lindell, A. F. Post, W. Regala, M. Shah, S. L. Shaw, C. Steglich, M. B. Sullivan, C. S. Ting, A. Tolonen, E. A. Webb, E. R. Zinser and S. W. Chisholm, Genome Divergence in Two Prochlorococcus ecotypes Reflects Oceanic Niche Differentiation, *Nature*, 2003, **424**, 1042–1047.

22 A. Dufresne, M. Salanoubat, F. Partensky, F. Artiguenave, I. M. Axmann, V. Barbe, S. Duprat, M. Y. Galperin, E. V. Koonin, F. Le Gall, K. S. Makarova, M. Ostrowski, S. Oztas, C. Robert, I. B. Rogozin, D. J. Scanlan, N. Tandeau de Marsac, J. Weissenbach, P. Wincker, Y. I. Wolf and W. R. Hess, Genome Sequence of the Cyanobacterium Prochlorococcus marinus SS120, a Nearly Minimal Oxyphototrophic Genome, *Proc. Natl. Acad. Sci. U. S. A.*, 2003, **100**, 9647–9649.

23 C. J. Benham and C. Bi, The Analysis of Stress-induced Duplex Destabilization in Long Genomic DNA Sequences, *J. Comput. Biol.*, 2004, **11**, 519–543.

24 K. Liolios, N. Tavernarakis, P. Hugenholtz and N. C. Kyrpides, The Genomes On Line Database (GOLD) v.2: a monitor of genome projects worldwide, *Nucleic Acids Res.*, 2006, **34**, D332–D334.

25 J. I. Rood and S. T. Cole, Molecular genetics and pathogenesis of *Clostridium perfringens*, *Microbiol. Rev.*, 1991, **55**, 621–648.