

GenColors: Annotation and Comparative Genomics Made Easy



Marius Felder,^{1,2} Alessandro Romualdi,³ Gernot Glöckner,² Matthias Platzer² and Jürgen Sühnel¹

¹ Biocomputing and ² Genome Analysis Groups, Leibniz Institute for Age Research – Fritz Lipmann Institute: FLI,

Jena Centre for Bioinformatics, Beutenbergstr. 11, D-07745 Jena, Germany

³ SIRS-Lab GmbH, Winzerlaer Str. 2, D-07745 Jena, Germany

SGB

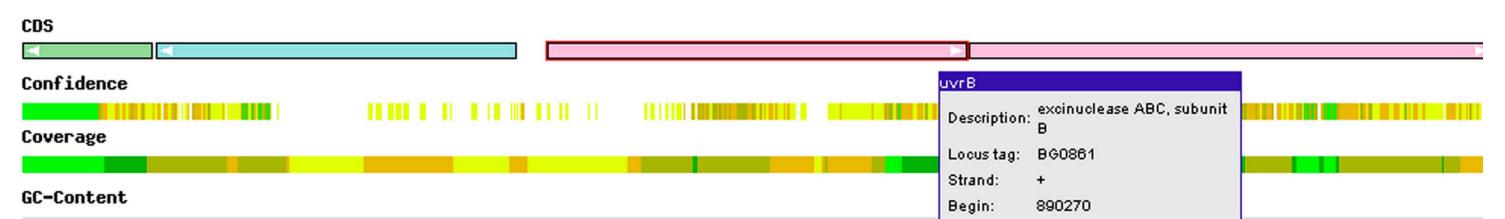
Introduction

GenColors [1,2] is a new web-based software/database system aimed at an improved and accelerated annotation of prokaryotic genomes considering information on related genomes and making extensive use of genome comparison. It offers a seamless integration of data from ongoing sequencing projects and annotated genomic sequences obtained from GenBank. With GenColors dedicated genome browsers containing a group of related genomes can be easily set up and maintained. The tool has been efficiently used for *Borrelia garinii* [3,4] and is currently applied to various ongoing genome projects on *Legionella, Pseudomonas* and *E. coli* genomes. Examples for freely accessible GenColors-based dedicated genome browsers are the Spirochetes Genome browser SGB (sgb.fli-leibniz.de), the Photogenome Browser CGB (cgb.fli-leibniz.de) and the Enterobacter Genome Browser ENGENE (engene.fli-leibniz.de). The system has now been adapted to handle also eukaryotic genomes. A first application of this feature is the ongoing annotation and analysis of two fungal species. Another GenColors-based tool is the Jena Prokaryotic Genome Viewer JPGV (jpgv.flileibniz.de). Contrary to the dedicated browsers it offers information on almost all finished bacterial genomes. As of July 10, 2008 it includes 1140 genomic elements of 293 species.

Gene information sheet

The most detailed information on a gene can be found on the Gene information sheets (Figure 4). These sheets start on top with a Gene environment graph. The DNA bases of both strands and their translation are displayed in the Basepair view.

In the central part of the sheet general gene information is provided. If the corresponding protein sequence is included in UniProt its description and all the external database links are also shown here.



The Browser

The available analysis and display options fall into three categories: General information, Search and Genome comparison.

The Spirochetes Genome Browser

Home	Genomes	Methods	Documentation	External links	Options	Contact	FLI	Jena
QuickSearch		□ Per	fect match Go				2008-05-07	14:54:31
General inform	ation			Genome comp	arison		You are logged in as ale	ex
							Logout	
Gene list				Best bid	irectional hits	(protein-based	analysis)	
Browse g	ene lists of sele	ected genomes.			a list of best bi of two genomes	directional hits b	etween the	
COG fun	ctional catego	ries						
<u> </u>	Select genes of specific COG functional categories			Gene co	re sets (proteir	is)		
across ge	enomes.				lated genes of t or user defined	wo or more geno lists.	omes, genomic	
Genome	plots							
Generate	Generate circular and linear plots.				rtnerships (DN	sis)		
						linear genomic e		
Des Condition in the second	le) linear whole	-				of potential gene on. This analysis		
	clickable linear	~		-		annot be found a	•	
	element coloure I classification.	d according to	the COG		ased bidirection			
Conomo	lists			Protein	variations (pro	tein-based ana	lysis)	
Genome Genorate	nenome lists of	selected geno	mor	[👫 List and a	analyse all best	bidirectional pro	tein hits.	

والمتراجعة والمتراجع	
Overview	
887770	89476
1	90424
Frame 1	Gly Ala Lys Leu Lys Leu Leu Tyr Phe Tyr Phe Phe Leu Gln Phe Leu Leu Asn Leu Phe Leu Phe Leu Pro Asn
Frame 2	Val Pro Asn Sto Asn Phe Phe Ile Phe Ile Phe Ser Cys Asn Phe Phe Sto Ile Tyr Phe Tyr Gln Ile
Frame 3	Cys Gln Ile Glu Thr Ser Leu Phe Leu Phe Phe Leu Ala Ile Ser Phe Glu Ser Ile Ser Ile Ser Thr Lys
DNA 5'-3'	G G T G C C A A A T T G A A A C T T C T T A T T T T A T T T T T
DNO 01 E1	888322 C C A C G G T T T A A C T T T G A A G A A A T A A A A A A A A A
DNA 3'-5'	
Frame 6	Ala Leu Asn Phe Ser Arg <mark>Sto</mark> Lys <mark>Sto</mark> Lys Lys Lys Cys Asn Arg Lys Phe Arg Asn Arg Gly Phe
Frame 5	His Trp Ile Ser Val Glu Lys Asn Lys Asn Lys Arg Ala Ile Glu Lys Ser Asp Ile Glu Val Leu Asn
Frame 4	Thr Gly Phe Gln Phe Lys Lys Ile Lys Ile Lys Glu Gln Leu Lys Lys Gln Ile Sto Lys Sto Trp Ile
Confidence	67 67 55 57 51 59 52 74 76 78 77 89 75 63 67 61 79 84 91 93 91 99 99 99 99 99 99 99 99 91 72 65 72 53 52 57 42 46 60 59 79 69 56 63 63 59 58 57 56 53 60 55 64 65 59 81 57 64 73 74 49 52 74 81 84 89 99
Coverage	4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4

Figure 4: Gene environment graph and basepair view

Genome comparison

Genome comparison tools constitute the major part of the GenColors system. Many of them are based on BBHs. These are defined as best BLAST hits between all protein sequences of two genomic elements that have at least 30% sequence identity and where the length of the matching region spans at least 30% of the query length.

BBH lists can further facilitate annotation as the description of a related gene can easily be transferred to the gene to be annotated by simply clicking on a transfer button.

Protein-coding genes in:			BBH in:								
Borrelia garinii			Borrelia I	Borrelia burgdorferi							
Name	Locus	GenBank description	View		Transfer		Locus GenBank description				
	tag		alignment	name	description		tag				
NA	BGB01	conserved hypothetical protein				88801	conserved hypothetical protein				
NA	BGB02	conserved hypothetical protein	=			88802	B. burgdorferi predicted coding region BBB02				

Figure 5: List of best bidirectional hits

Get information on genes of a genome acquired by horizontal gene transfer.

Generate genome lists of selected genomes.

Sequence retrieval

Retrieve genomic sequences or intergenic regions of selected genomic elements.

Search

Advanced search

Search for organisms, genes, COGs and external database information (short output with gene name, locus tag, description, location, strand information. Comprehensive output with additional information on best-bidirectional protein hits in all SGB genomes and UniProt and TrEMBL hits). one or more genomes or genomic elements.

Alignments of whole reference and target genomic elements (DNA-based analysis)

Analyse the DNA alignment statistics of reference and

Synteny analysis (protein-based analysis)

Gene conservation (protein-based analysis)

Codon and amino acid usages (DNA- and

on the best bidirectional protein hits.

genome and all other SGB genomes.

protein-based analysis)

Analyse syntenies between different genomes based

Analyse the gene conservation between one selected

Show or compare the codon and amino acid usages of

target genomes.

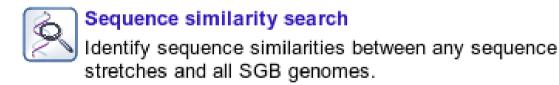


Figure 1: Methods page of the Spirochetes Genome Browser

General information

GenColors offers browsing across whole genomes (Figure 2, left), single genomic elements or according to the COG functional classification (Figure 2, right).

Borrelia garinii, PBi Generate gene lists for all or specific genomic elements or COG functional categories. Show summary table. Download as GenBank file.	Unclassified [-]	RNA processing and modification [A]	Chromatin structure and dynamics [B]		
Leptospira interrogans serovar Copenhageni str. Fiocruz	Energy	Cell cycle	Amino acid		
L1-130	production and	control, mitosis	transport and		
Generate gene lists for all or specific genomic elements or COG	conversion	and meiosis	metabolism		
	[C]	[D]	[E]		

	Query Name / Locus tag / GenBank description	Ins	s De	l Du	p Exc	Sd	Nd	s	N	ps*	pn*	ds*	dn*	ds/dn*	(Sd-Nd)/(Sd+Nd)
BB0002/: beta-N-acetylhexosaminidase, putative (1029 bp, 342 aa)	NA/BG0002: beta-N-acetylhexosaminidase, putative (1020 bp, 339 aa)	0	1	0	23	51	28	204	813	0.25	0.034	0.304	0.035	8.625	0.291
BB0003/: B. burgdorferi predicted coding region BB0003 (1365 bp, 454 aa)	NA/BG0003: hypothetical protein (1479 bp, 492 aa)	4	0	0	68	70	80	263	1098	0.267	0.073	0.33	0.077	4.262	-0.067

Figure 6: List of protein variations

Synteny groups consist of two or more neighbouring genes in one genomic element that have neighbouring BBHs in another genomic element of either the same or a different species. Neighbouring genes may be interrupted by up to five genes that have no BBH in the counterpart genomic element.

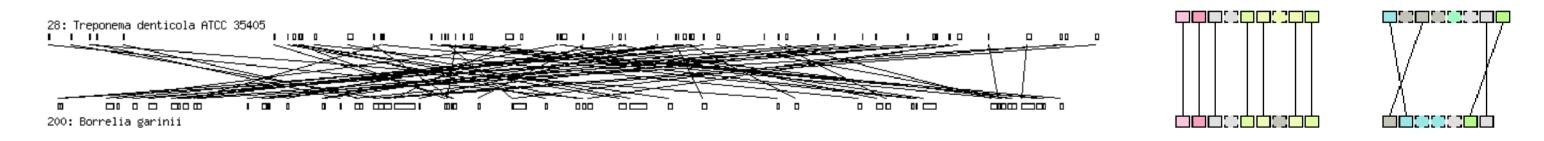
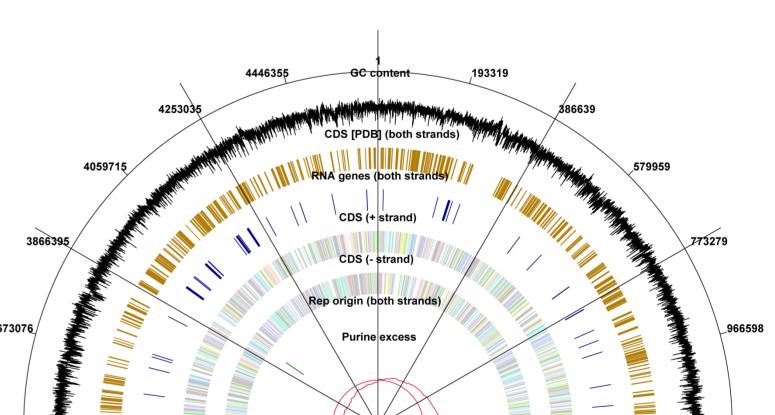


Figure 7: Overview (left) of all synteny groups identified between the chromosomes of *Treponema denticola ATCC 35405* and *Borrelia garinii*. Detailed views of two synteny groups (right) with conserved and inverted gene order.

Genome plots

GenColors generates circular and linear genome plots (PNG, PDF, Postscript) using annotation data and calculating quantities such as GC-content, GC-skew and purine and keto excess. Different features of one or more genomes can be displayed in one representation facilitating comparative analysis.



functional categories. Show summary table.

Figure 2: Browsing options in SGB

Search

GenColors basically offers two ways of querying the underlying database. There is a QuickSearch option for text strings in gene names, descriptions or locus tags as well as an AdvancedSearch option that allows the combination of 20 different data types. These include gene identifiers/description, gene lengths, genomes or genomic elements, COG categories, PROSITE sequence patterns and the complete external database information provided by UniProt. Sequence based searches are done via BLAST.

Figure 3: AdvancedSearch in GenColors (selected categories)

Search by locus tag

Search by gene description

Search	by	gene	note

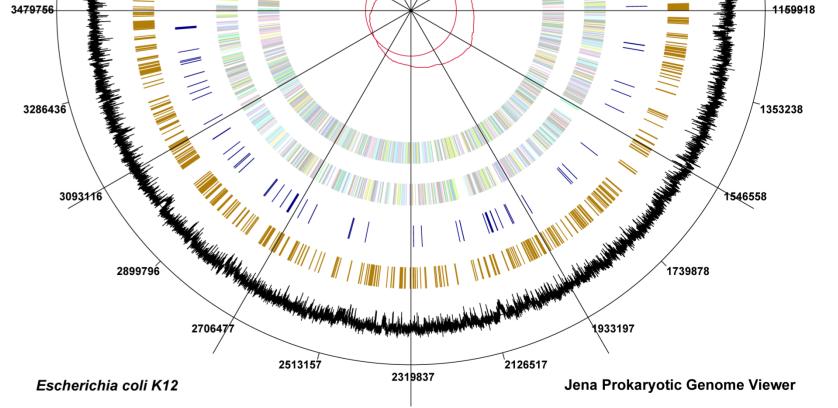
Search by gene length (please use '<', '>' or '=')

Search by protein sequence motif (Prosite format)

Select by external database identifier

Aarhus/Ghent-2DPAGE

Fig 8: Circular plot of features of the *Escherichia coli* K12 chromosome: different gene types and precomputed data (GC-content, purine excess) are displayed on different orbits. Colouring of the CDS orbits corresponds to COG functional categories.



References

[1] Romualdi, Siddiqui, Glöckner, Lehmann, Sühnel. GenColors: accelerated comparative analysis and annotation of prokaryotic genomes at various stages of completeness. Bioinformatics **2005**, 15, 3669-71.

[2] Romualdi, Felder, Rose, Gausmann, Schilhabel, Glöckner, Platzer, Sühnel. GenColors: Annotation and Comparative Genomics of Prokaryotes Made Easy. Methods Mol. Biol. 2007, 395, 75-96.

[3] Glöckner, Lehmann, Romualdi, Pradella, Schulte-Spechtel, Schilhabel, Wilske, Sühnel, Platzer. Comparative analysis of the *Borrelia garinii* genome. Nucleic Acids Res. **2004**, 32, 6038-46.

[4] Glöckner, Schulte-Spechtel, Schilhabel, Felder, Sühnel, Wilske, Platzer. Comparative genome analysis: selection pressure on the *Borrelia vls* cassettes is essential for infectivity. BMC Genomics **2006**, 7, 211.