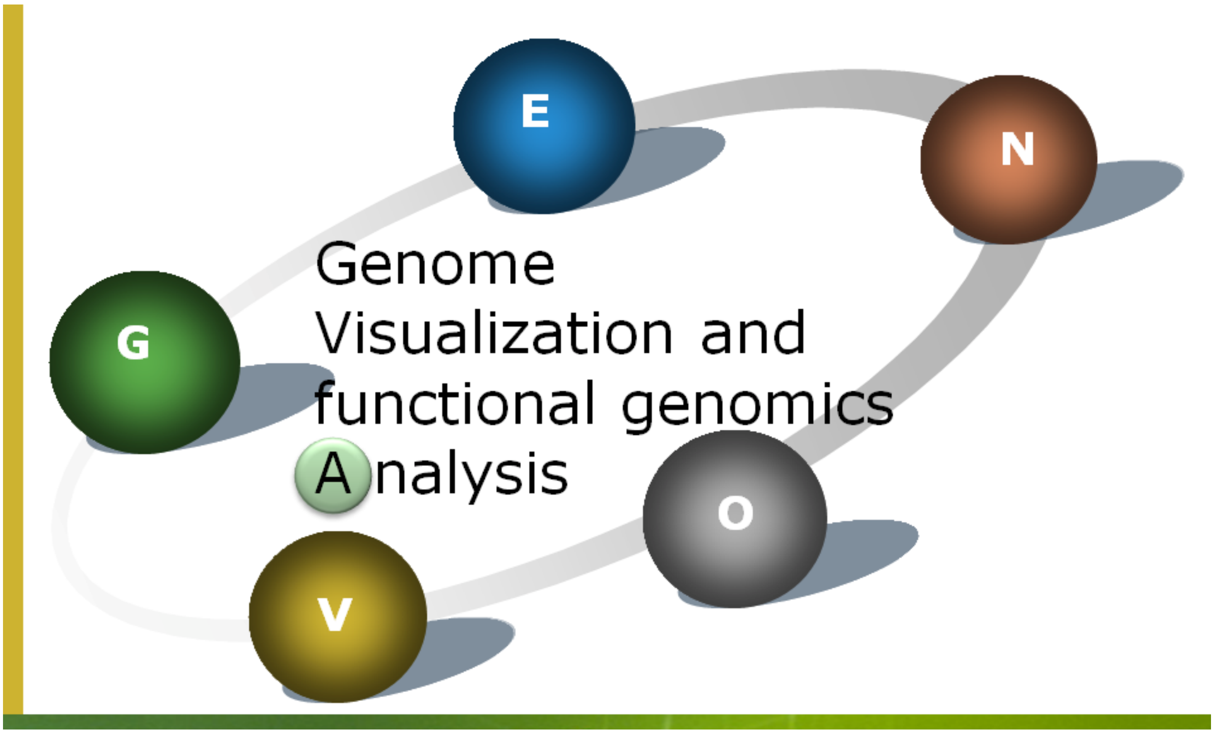
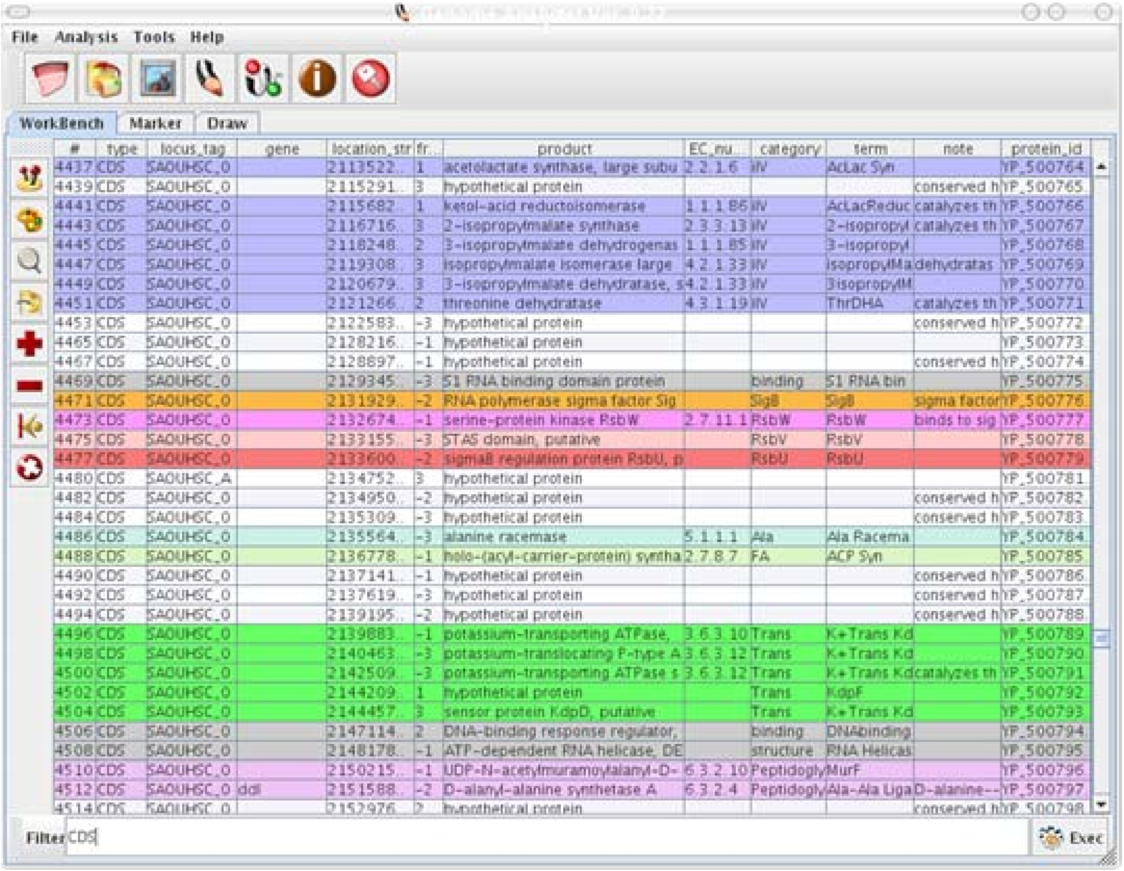


STEP INTO GENOVA

Part 1. Views Introduction

GENOVA provides three views for genome visualization and modification. They are the feature editor (WorkBench), the marker editor (Marker) and the visualization canvas (Draw). The following figure is a typical view of the feature editor, which is initiated automatically once the genome file (GenBank format) has been read. In this example, the *Staphylococcus aureus* RN1 genome sequence is studied in the table. Users could insert or delete nucleotide bases, genomic elements (also in batch) and feature entries of interest into the genome. Note there are two more features than in the common GenBank content, i.e., '***category***' and '***term***', where users can specify the desired tag and favorite color. Marker view is a secondary table which is designed to note the genomic rearrangement both in table and in the canvas. The last view is the canvas, users can invoke it using the "canvas" button, there GENOVA provides abundant options for obtaining a optimal visualization, zooming, range, changing terms, etc.

1. Feature Editor



#	type	locus_tag	gene	location_str	fr..	product	EC_nu..	category	term	note	protein_id
4437	CDS	SAOUHSC_0		2113522..	1	acetolactate synthase, large subu	2.2.1.6	ilv	AcLac Syn		YP_500764
4439	CDS	SAOUHSC_0		2115291..	3	hypothetical protein				conserved h	YP_500765
4441	CDS	SAOUHSC_0		2115682..	1	ketol-acid reductoisomerase	1.1.1.86	ilv	AcLacReduc	catalyzes th	YP_500766
4443	CDS	SAOUHSC_0		2116716..	3	2-isopropylmalate synthase	2.3.3.13	ilv	2-isopropyl	catalyzes th	YP_500767
4445	CDS	SAOUHSC_0		2118248..	2	3-isopropylmalate dehydrogenas	1.1.1.85	ilv	3-isopropyl		YP_500768
4447	CDS	SAOUHSC_0		2119308..	3	isopropylmalate isomerase large	4.2.1.33	ilv	isopropylMa	dehydratas	YP_500769
4449	CDS	SAOUHSC_0		2120679..	3	3-isopropylmalate dehydratase, s	4.2.1.33	ilv	3isopropylM		YP_500770
4451	CDS	SAOUHSC_0		2121266..	2	threonine dehydratase	4.3.1.19	ilv	ThrDHA	catalyzes th	YP_500771
4453	CDS	SAOUHSC_0		2122583..	-3	hypothetical protein				conserved h	YP_500772
4465	CDS	SAOUHSC_0		2128216..	-1	hypothetical protein					YP_500773
4467	CDS	SAOUHSC_0		2128897..	-1	hypothetical protein				conserved h	YP_500774
4469	CDS	SAOUHSC_0		2129345..	-3	S1 RNA binding domain protein		binding	S1 RNA bin		YP_500775
4471	CDS	SAOUHSC_0		2131929..	-2	RNA polymerase sigma factor Sig		SigB	SigB	sigma factor	YP_500776
4473	CDS	SAOUHSC_0		2132674..	-1	serine-protein kinase RsbW	2.7.11.1	RsbW	RsbW	binds to sig	YP_500777
4475	CDS	SAOUHSC_0		2133155..	-3	STAS domain, putative		RsbV	RsbV		YP_500778
4477	CDS	SAOUHSC_0		2133600..	-2	sigmaB regulation protein RsbU, p		RsbU	RsbU		YP_500779
4480	CDS	SAOUHSC_A		2134752..	3	hypothetical protein					YP_500781
4482	CDS	SAOUHSC_0		2134950..	-2	hypothetical protein				conserved h	YP_500782
4484	CDS	SAOUHSC_0		2135309..	-3	hypothetical protein				conserved h	YP_500783
4486	CDS	SAOUHSC_0		2135564..	-3	alanine racemase	5.1.1.1	Ala	Ala Racema		YP_500784
4488	CDS	SAOUHSC_0		2136778..	-1	holo-(acyl-carrier-protein) syntha	2.7.8.7	FA	ACP Syn		YP_500785
4490	CDS	SAOUHSC_0		2137141..	-1	hypothetical protein				conserved h	YP_500786
4492	CDS	SAOUHSC_0		2137619..	-3	hypothetical protein				conserved h	YP_500787
4494	CDS	SAOUHSC_0		2139195..	-2	hypothetical protein				conserved h	YP_500788
4496	CDS	SAOUHSC_0		2139883..	-1	potassium-transporting ATPase,	3.6.3.10	Trans	K+Trans Kd		YP_500789
4498	CDS	SAOUHSC_0		2140463..	-3	potassium-translocating P-type A	3.6.3.12	Trans	K+Trans Kd		YP_500790
4500	CDS	SAOUHSC_0		2142509..	-3	potassium-transporting ATPase s	3.6.3.12	Trans	K+Trans Kd	catalyzes th	YP_500791
4502	CDS	SAOUHSC_0		2144209..	1	hypothetical protein		Trans	KdpF		YP_500792
4504	CDS	SAOUHSC_0		2144457..	3	sensor protein KdpD, putative		Trans	K+Trans Kd		YP_500793
4506	CDS	SAOUHSC_0		2147114..	2	DNA-binding response regulator,		binding	DNAbinding		YP_500794
4508	CDS	SAOUHSC_0		2148178..	-1	ATP-dependent RNA helicase, DE		structure	RNA Helicas		YP_500795
4510	CDS	SAOUHSC_0		2150215..	-1	UDP-N-acetylmuramoylalanyl-D-	6.3.2.10	Peptidogly	MurF		YP_500796
4512	CDS	SAOUHSC_0	ddl	2151588..	-2	D-alanyl-alanine synthetase A	6.3.2.4	Peptidogly	Ala-Ala Liga	D-alanine--	YP_500797
4514	CDS	SAOUHSC_0		2152976..	2	hypothetical protein				conserved h	YP_500798

Filter CDS

Fig. T1 Feature Editor – WorkBench. Features can be managed in the table, sorting and changing the values. All the cells except the first column '#' can be modified. The leading novel genome is generated and synchronized in the canvas. Operations can be carried using the buttons in the toolbar, the function are listed as below:

- Major buttons of the horizontal toolbar

Load a sequence file of GenBank format

Save as a sequence file of GenBank

Save the genomic map with notes and markers into a picture file

Plot a genomic map

Content Analysis

License information

Closing the workspace

- Editor buttons in the vertical bar

Column configuration of the feature table

Categorization palette

Search the keywords in the specified column

Set the values in a certain column of selected rows

Add a feature entry into the table

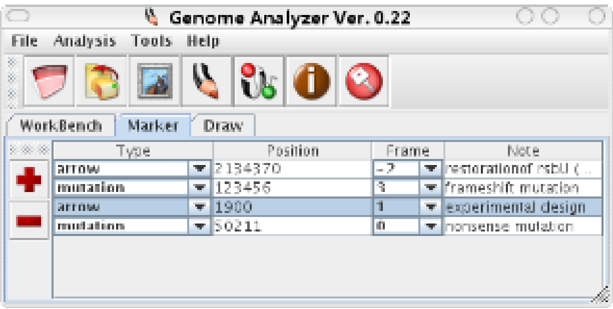
Remove selected feature entries from the table, the corresponding nucleotide sequence can be also removed out of the genome after confirmation

Insert nucleotide bases into the genome

Delete nucleotide bases of desired range from the genome

Moreover, there is a “Filter” row below the editor, which enables the user to pick and show only feature entries according to keywords, e.g. “CDS” as filter will hide all entries which contain no “CDS”. Users can easily type the category name in the “category” column. The system will give a random color, which always can be changed in the palette. All entries assigned in a category will take the color specified previously. Genes involved in the same island or in the same operon are given with preference (default setting) the same or similar color in order to distinguish these genes from others, e.g., in the table, light purple denotes all genes belong to the isoleucine-valine operon.

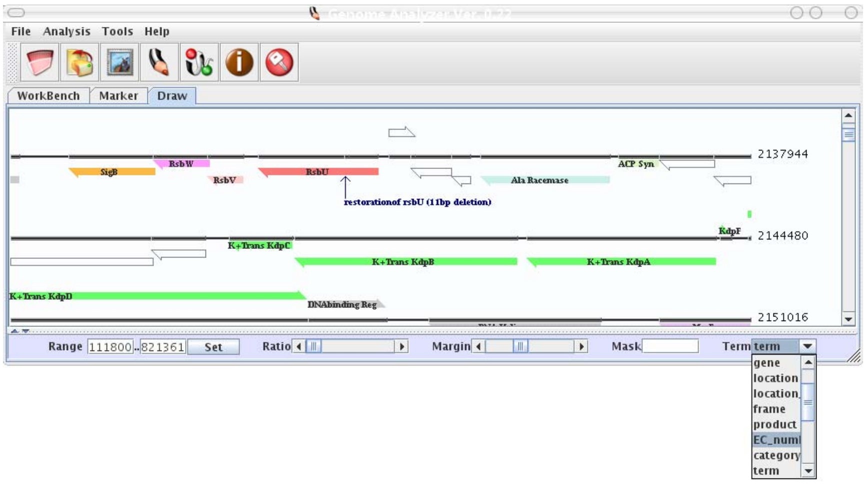
2. Marker Editor (Marker)



Type	Position	Frame	Note
arrow	2134370	-2	restorationof rsbU (...
mutation	123456	3	frameshift mutation
arrow	1900	1	experimental design
mutation	50211	0	nonsense mutation

Fig. T2 Marker Editor. Markers can be inserted, edited and removed in this view. In the "Type" column, users can specify the favorite symbol (mutation is the "X" symbol), the corresponding note information will be painted as well in the canvas. Note that the reading frame (+3,+2,+1,0,-1,-2,-3) should also be assigned according to the place users would like to insert or plot additional markers. The left control panel enables to insert the marker or remove the selected one.

3. Visualization Canvas (Draw)



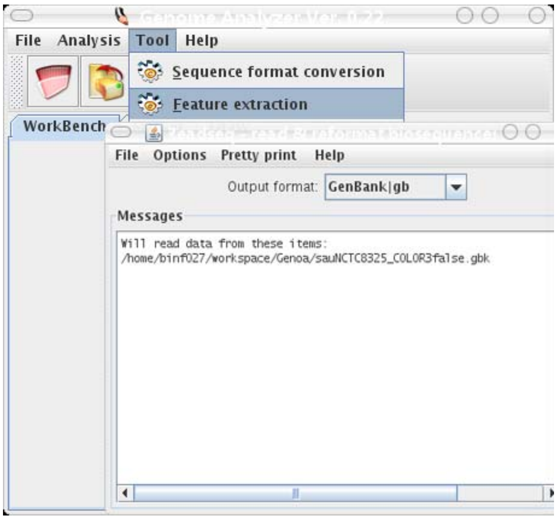
Range 111800..821361 Set

Term term

Fig. T3 Visualization canvas. The control panel below the canvas provides the possibility to set the visualization range ("Set" button), modify the zoom factor ("Ratio" slider), increase or decrease the line margin ("Margin" slider) , specify the term visualized and used for gene labeling ("Term"), e.g., *locus_tag, gene, protein_id, EC_number*, etc. There is another function named "Mask", which plays a crucial function to hide redundant symbols in the terms, in particular when the "locus_tag" is chosen as term: Similar leading words, e.g., the prefix of "SAOUHSC_" can be hidden to prevent cluttering of the figure using the mask.

4. Tools

1)

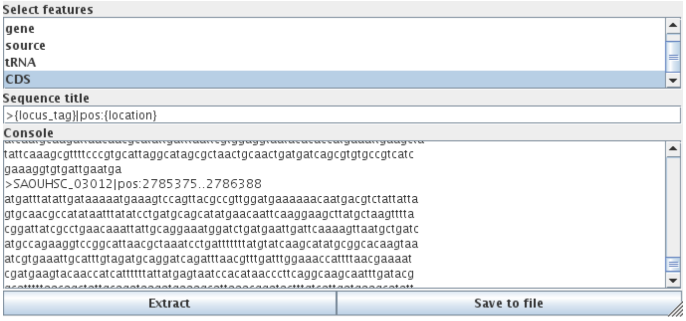


Output format: GenBank|gb

Messages

```
Will read data from these items:
/home/binf027/workspace/Genoa/sauNCTC8325_COLOR3false.gbk
```

2)



Select features

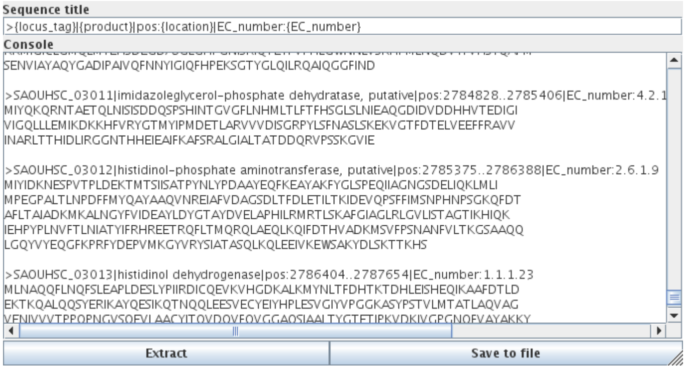
gene
source
tRNA
CDS

Sequence title: >{locus_tag}|pos:{location}

Console

```
tattcaaagcgttttcccgtgcattaggcatagcgctaactgcaactgatgatcagcgtgtgccgtcatc
gaaaggtgtgattgaatga
>SAOUHSC_03012|pos:2785375..2786388
atgatttatattgataaaaatgaaagtccagttacgccgttggatgaaaaaacaatgacgtctattatta
gtgcaacgccatataatttatatcctgatgcagcatatgaacaattcaaggaagcttatgctaagtttta
cggattatcgcctgaacaaattattgcaggaaatggatctgatgaattgattcaaaagttaatgctgatc
atgccagaaggtccggcattaacgctaaatcctgattttttatgtatcaagcatatgcggcacaagtaa
atcgtgaaattgcatttgtagatgcaggatcagatttaacgtttgatttggaaaccattttaacgaaaat
cgatgaagtacaaccatcattttttattatgagtaatccacataacccttcaggcaagcaatttgatacg
```

3)



Sequence title: >{locus_tag}|{product}|pos:{location}|EC_number:{EC_number}

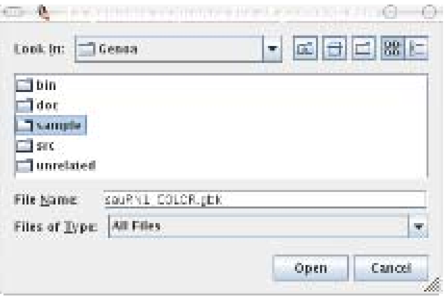
Console

```
SENVIAYAQYGADIPAIVQFNNYIGIQFHPEKSGTYGLQILRQAIQGGFIND

>SAOUHSC_03011|imidazoleglycerol-phosphate dehydratase, putative|pos:2784828..2785406|EC_number:4.2.1
MIYQKQRNTAETQLNISISDDQSPSHINTGVGFLNHMLTLFTFHSGLSLNIEAQGDIDVDDHHVTEDIGI
VIGQLLLEMIKDKKHFVRYGTMYIPMDETLARVVVDISGRPYLSFNASLSKEKVGTFDTELVEEFFRAVV
INARLTTHIDLIRGGNTHHEIEAIFKAFSRALGIALTATDDQRVPSSKGVIE

>SAOUHSC_03012|histidinol-phosphate aminotransferase, putative|pos:2785375..2786388|EC_number:2.6.1.9
MIYIDKNESPVTPLDEKTMTSIISATPYNLYPDAAYEQFKEAYAKFYGLSPEQIIAGNGSDELIQKLMLI
MPEGPALTLNPDFFMYQAYAAQVNREIAFVDAGSDLTFDLETILTKIDEVQPSFFIMSNPHNPSGKQFDT
AFLTAIADKMKALNGYFVIDEAYLDYGTAYDVELAPHILRMRTLSKAFGIAGLRLGVLISTAGTIKHIQK
IEHPYPLNVFTLNIATYIFRHREETRQFLTMQRQLAEQLKQIFDTHVADKMSVFPSNANFVLTKGSAAQQ
LGQYVYEQGFKPRFYDEPVMKGYVRYSIATASQLKQLEEIVKEWSAKYDLSKTTKHS

>SAOUHSC_03013|histidinol dehydrogenase|pos:2786404..2787654|EC_number:1.1.1.23
MLNAQQFLNQFSLEAPLDESLYPIIRDICQEVKVHGDKALKMYNLTFDHTKTDHLEISHEQIKAAFDTLD
EKTKQALQQSYERIKAYQESIKQTNQQLEESVECYEIYHPLESVGIYVPGGKASYPSTVLMTATLAQVAG
VENIVVVTPPQPNGVSQEVLAACYITQVDQVFQVGGAQSIAALTYGTETIPKVDKIVGPGNQEVAYAKKY
```

Fig. T4 Tools incorporated into the GENOVA. A tool-kit allows multiple genome sequence format conversion (1; readseq [suppl. 8]) and comprehensive feature extraction (2), as well as construction of a proteome file from GenBank format genome data (3; as long as the input file contains translated protein entries). All this can be rapidly managed in seconds.

Part 2. Typical steps (suggested tutorial tour)

1. Import the sequence file with the annotation into the GENOVA software using the "***open***" button. (sample sequence in the sample sub-directory)



Look In: Genoa

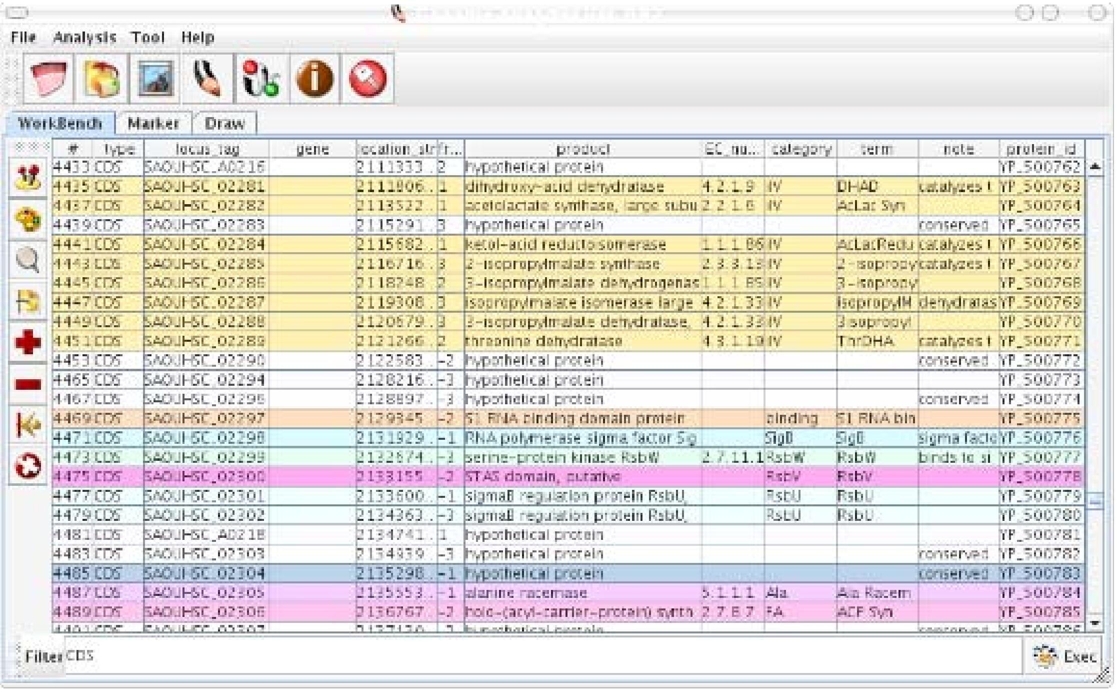
bin
doc
sample
src
unrelated

File Name: sauRN1_COLOR.gbk

Files of Type: All Files

Fig. T5 Open a GenBank file located in the sample directory.

2. Edit the features and sequences in the "feature editor" (***workbench*** tab). Firstly please type CDS in the bottom filter text-field, click the "***Exec***" button or simply press return key, so that only CDS entries are listed in the editor. Scroll them to the desired region, e.g., SAOUHSC_02300 in the following figure.



#	type	locus_tag	gene	location_str	fr...	product	EC_nu...	category	term	note	protein_id
4433	CDS	SAOUHSC_A0216		2111333..	2	hypothetical protein					YP_500762
4435	CDS	SAOUHSC_02281		2111806..	1	dihydroxy-acid dehydratase	4.2.1.9	IlV	DHAD	catalyzes t	YP_500763
4437	CDS	SAOUHSC_02282		2113522..	1	acetolactate synthase, large subu	2.2.1.6	IlV	AcLac Syn		YP_500764
4439	CDS	SAOUHSC_02283		2115291..	3	hypothetical protein				conserved	YP_500765
4441	CDS	SAOUHSC_02284		2115682..	1	ketol-acid reductoisomerase	1.1.1.86	IlV	AcLacRedu	catalyzes t	YP_500766
4443	CDS	SAOUHSC_02285		2116716..	3	2-isopropylmalate synthase	2.3.3.13	IlV	2-isopropy	catalyzes t	YP_500767
4445	CDS	SAOUHSC_02286		2118748..	2	3-isopropylmalate dehydrogenas	1.1.1.85	IlV	3-isopropy		YP_500768
4447	CDS	SAOUHSC_02287		2119308..	3	isopropylmalate isomerase large	4.2.1.33	IlV	IsopropylM	dehydratas	YP_500769
4449	CDS	SAOUHSC_02288		2120679..	3	3-isopropylmalate dehydratase,	4.2.1.33	IlV	3-isopropyl		YP_500770
4451	CDS	SAOUHSC_02289		2121266..	2	threonine dehydratase	4.3.1.19	IlV	ThrDHA	catalyzes t	YP_500771
4453	CDS	SAOUHSC_02290		2122583..	-2	hypothetical protein				conserved	YP_500772
4465	CDS	SAOUHSC_02294		2128216..	-3	hypothetical protein					YP_500773
4467	CDS	SAOUHSC_02295		2128897..	-3	hypothetical protein				conserved	YP_500774
4469	CDS	SAOUHSC_02297		2129345..	-2	S1 RNA binding domain protein		binding	S1 RNA bin		YP_500775
4471	CDS	SAOUHSC_02298		2131929..	-1	RNA polymerase sigma factor Sig		SigB	SigB	sigma facto	YP_500776
4473	CDS	SAOUHSC_02299		2132674..	-3	serine-protein kinase RsbW	2.7.11.1	RsbW	RsbW	binds to si	YP_500777
4475	CDS	SAOUHSC_02300		2133155..	-2	STAS domain, putative		RsbV	RsbV		YP_500778
4477	CDS	SAOUHSC_02301		2133600..	-1	sigmaB regulation protein RsbU,		RsbU	RsbU		YP_500779
4479	CDS	SAOUHSC_02302		2134363..	-3	sigmaB regulation protein RsbU,		RsbU	RsbU		YP_500780
4481	CDS	SAOUHSC_A0218		2134741..	1	hypothetical protein					YP_500781
4483	CDS	SAOUHSC_02303		2134939..	-3	hypothetical protein				conserved	YP_500782
4485	CDS	SAOUHSC_02304		2135298..	-1	hypothetical protein				conserved	YP_500783
4487	CDS	SAOUHSC_02305		2135553..	-1	alanine racemase	5.1.1.1	Ala	Ala Racem		YP_500784
4489	CDS	SAOUHSC_02306		2136767..	-2	holo-(acyl-carrier-protein) synth	2.7.8.7	FA	ACP Syn		YP_500785

Filter CDS

Fig. T6 Overview of operations in the genome feature editor.

3. Categorize them by writing directly into cells of the "***category***" column, e.g., *rsb_sigB*, or any other desired name (ilV, RsbV, RsbU in the sample file), system will automatically detect the repeats and assign a random color. Of course you are welcome to specify a favourite color using the "***palette***" button directly. This table is actually serving as a collector of function groups.

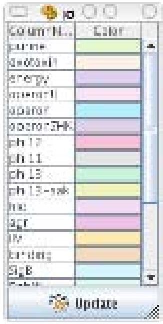
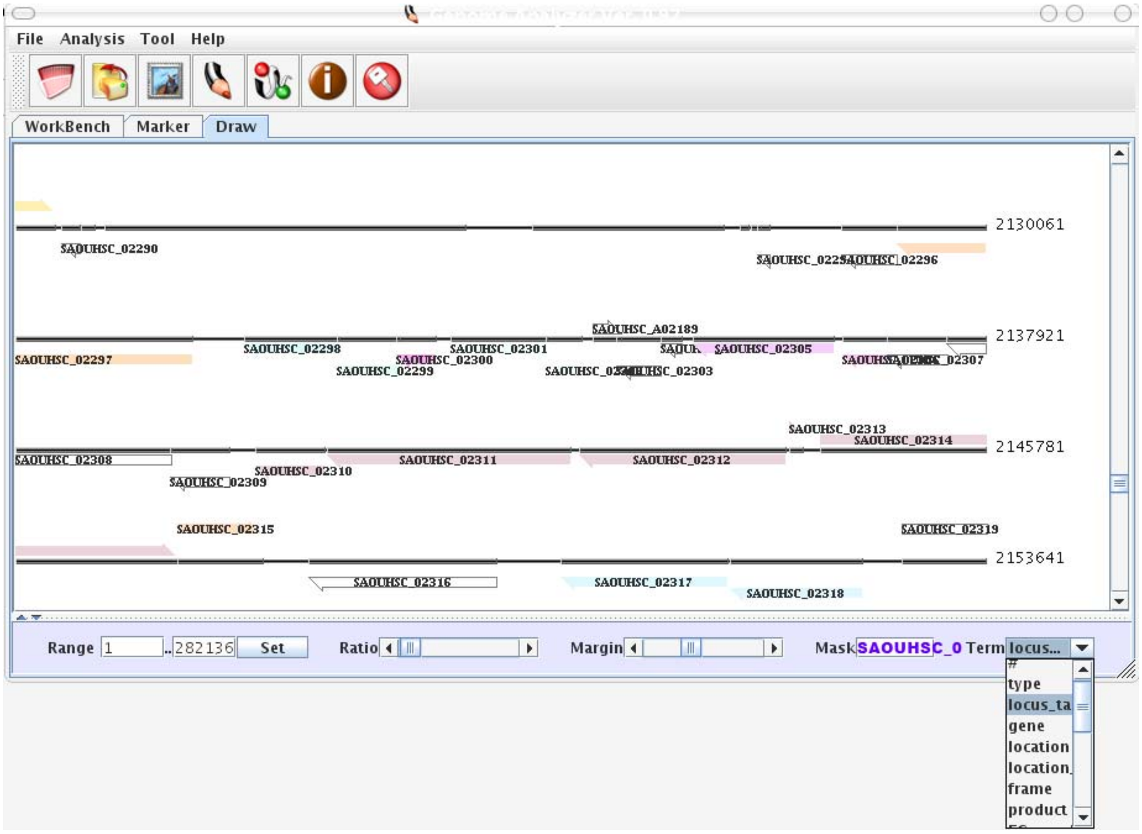


Fig. T7 Categorization palette.

4. Click the "***Plot***" button to visualize the ORFs, the ***Term*** combo-box in the bottom control panel allows users to switch different tags appearing in the ORFs, e.g., *locus_tag, gene name, position, function, EC number* as well as the ***term***, which is in addition in the feature editor (column position see figure of step.2) allows any desired notes from users without changing the other original annotation. The following figure is an example when the user specify the *locus_tag* as the ORF names. Here users are suggested to remove the redundant prefix using the "***Mask***", please input SAOUHSC_0 in the text-field. Any regular expression can be used here as the mask filter. In the following steps, we suggest to choose the term in the term combo-box for the illustration. The other slider "***ratio***" is for the zoom factor, and "***margin***" is for increasing or decreasing the line margin.



Range 1 .. 282136

Mask SAOUHSC_0 Term locus...

Fig. T8 Canvas operation.

5. Add some markers in the “***marker***” view (Fig. T2) to note down the genome region as well as the interesting rearrangement events. Frame is critical for acquiring an optimal figure, since the ORFs are drawn according to their reading frame numbers (3,2,1,0,-1,-2,-3). Here 0 suggests the arrow should point to the genome scale line directly.

6. Observe them again by selecting the “***draw***” tab in the index (see Fig.2 in the GENOVA manuscript)

7. Click the “***Statistics***” button to inspect the content analysis.

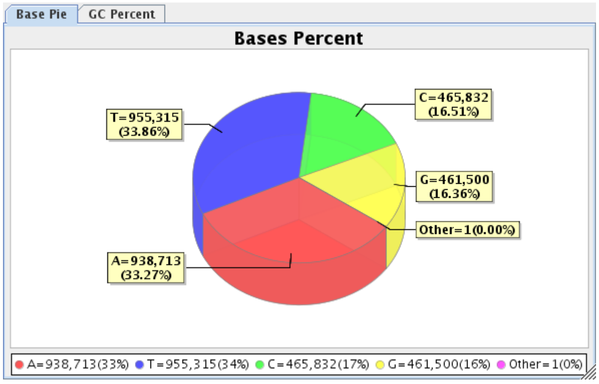


Fig. T8 Content analysis.

8. Optional: Edit the features, inserting and removing entries, adding and deleting nucleotide sequences (Fig. T1). This operation will change the genome length and all the ORFs, the software will return the result list of all the deleted entries, certain entries partially changed which required for a manual confirmation or modification.

9. Save the novel genome as a GenBank file using the “***save***” button. The categorization information will also be stored.

10. The resulting genome figure can be exported into a picture file using the “***save genome map***” button.

11. Optional: Toolkit to extract feature entries or convert the file into different formats or a proteome sequence-file (Fig. T4). The feature extraction tool allows users to select the feature type firstly in the upper list, specify the desired title-format and generate a sequence collection in fasta format. The “***save to file***” button enables to save as a permanent file in the disk.

12. Latest update, other tutorials or more sample files please visit our website located at http://genova.bioapps.biozentrum.uni-wuerzburg.de, thank you very much and we are looking forward to hearing from you.