Manual Codon Harmonizer tool (GALAXY)

The tool (codonharmonizer.systemsbiology.nl) allows for generating codon harmonized sequences, for example for synthetic gene design for heterologous protein production. The harmonization algorithm for this tool is based on the original harmonization algorithm proposed before (Angov *et al.*, 2008; Angov, Legler and Mease, 2011). If annotated genomes are available in NCBI for the original host of a gene and the expression host of interested this tool can generate harmonized genes. A detailed explanation of the algorithm and accompanying experimental results for some membrane proteins was published by us (Claassens, Siliakus *et al.*, 2017).

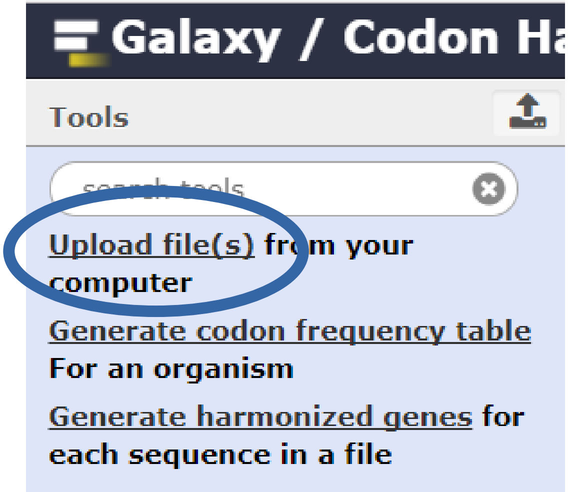
This manual aims to help users of this tool. Further questions, suggestions or experiences regarding this tool or codon harmonization can be directed to Bart Nijsse (bart.nijsse@wur.nl) or Nico Claassens (nicoclaassens@gmail.com)

Happy harmonizing!

***Please cite this tool as*:** Claassens NJ, Siliakus MF, Nijsse B, Spaans SK, Creutzburg SCA, Schaap PJ, et al. Improving heterologous membrane protein production in *Escherichia coli* by combining transcriptional tuning and codon usage algorithms. PLoS One. 2017

Harmonization in 3 steps

Step 1: Upload files

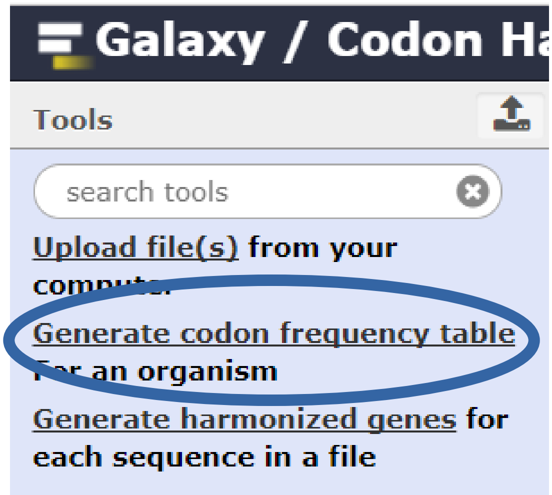

Galaxy / Codon Ha
Tools
search tools
Upload file(s) from your computer
Generate codon frequency table For an organism
Generate harmonized genes for each sequence in a file

Choose “Upload File” or Click on the green upload button. For harmonization you should upload:

- (wild-type) gene sequence(s) for genes to be harmonized (fasta format)
- Multi-fasta files (.gz) with coding sequences for the native host and expression hosts, see at the end of this manual how to obtain those easily from NCBI

You can upload files from your computer or paste a (web)link or fasta text directly file by clicking on the button: “Paste/Fetch data”.

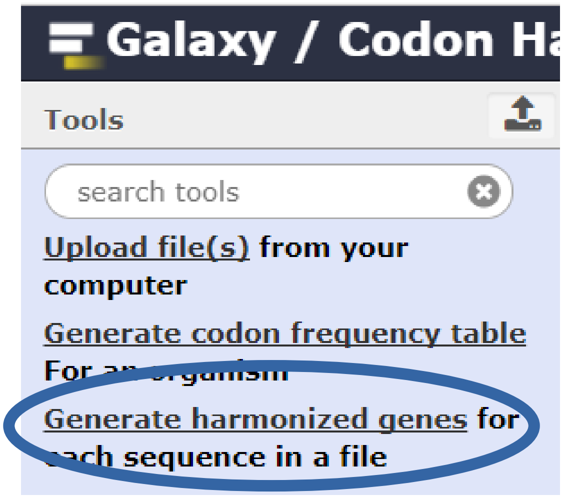
Step 2: Generate codon frequency table


Galaxy / Codon Ha
Tools
search tools
Upload file(s) from your computer
Generate codon frequency table For an organism
Generate harmonized genes for each sequence in a file

You have to use this step to convert the multi-fasta CDS files to codon frequency table files.

You can use an uploaded multi-fasta CDS file of the organism for which you can to generate the table: “Generate codon frequency table”
A name is required, you can enter e.g. the species name.

Step 3: Generate harmonized genes


Galaxy / Codon Ha
Tools
search tools
Upload file(s) from your computer
Generate codon frequency table For an organism
Generate harmonized genes for each sequence in a file

In this last step the harmonization is performed.

Input genes:
Select from the uploaded documents the gene(s) you want to harmonize (fasta file)

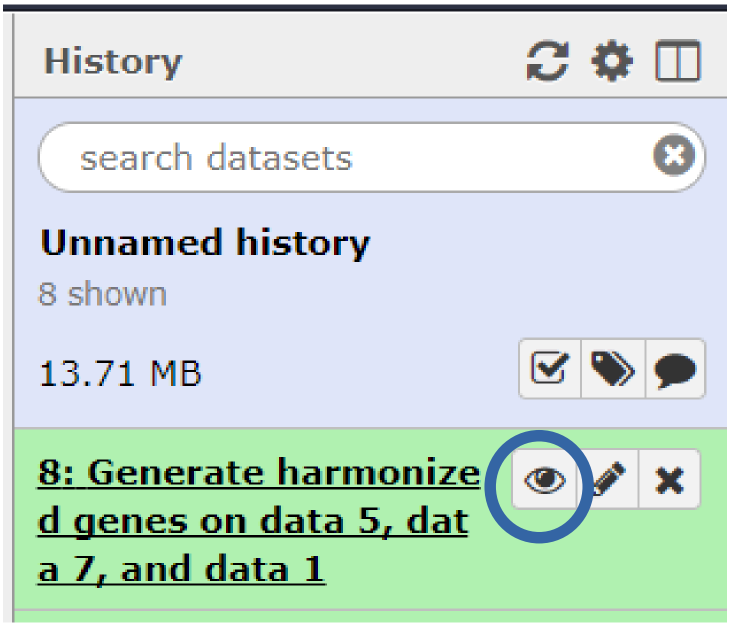
Source Frequency file:
Select the generated codon frequency table file for the source (wild-type) organisms from which the to be harmonized gene originally comes.

Targets:
Select the generated codon frequency table file for the target organism(s) (expression host) you want to harmonize your genes for. Select multiple custom frequency files by holding the **control** key.

Note: Multiple genes can be harmonized at the same time in 1 fasta file. As long they have the same "Source Organism".

Finally, press " execute" and the harmonized sequence is generated

Retrieving the harmonized gene



In the history tab on the right the harmonization output is available (green when successfully finished). Click on the 'eye logo' to obtain a zipped csv. file.

Unzip the file and you will find this file contains multiple output data and parameters including the prime output, "The harmonized gene". This gene sequence you can order for synthesis if relevant. More detailed data as codon frequency list of different gene variants and related parameters are available lower in the file. The parameters included are the Codon Adaptation Index (CAI) and Codon Harmonization Index (CHI), for explanation of these parameters see (Claassens *et al.*, 2017). Codon frequency list of the source (wild-type) gene in the wild-type host are given first, then for the source gene in the expression host, and lastly for the harmonized gene in the expression host. Below those lists of the gene codon frequencies, codon frequency tables for organisms included in the analysis are provided.

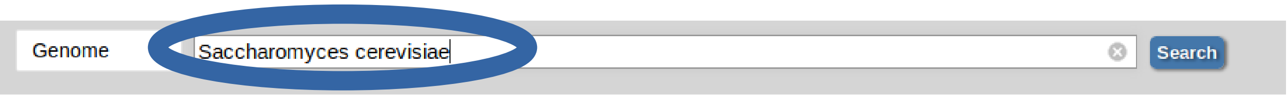
Some extra guidance on the output and usage of harmonized genes

CAI, CHI and gene codon frequency list may be relevant to inspect, also for the source/wild-type gene in the expression hosts. If the CAI value is not very low (e.g. > 0.6, but no clear threshold value known), this indicates that the wild-type contains relatively many abundant codons for the expression hosts. This CAI parameter is typically optimized in many available codon optimization algorithms. In our tool, alternatively a codon harmonization algorithm is applied, resulting a CHI value close to 0. If the CHI value is close to 0 =indicates the gene is well harmonized. If the CHI of the wild-type gene is close to this of the harmonized gene, the wild-type gene is on itself is already well harmonized for the expression host considered.

As a general warning, none of the codon adaptation algorithms available nowadays, including codon optimization and harmonization, give a guarantee for high expression. Harmonized gene variants can be a relevant variant to include in protein production screens though, but also the wild-type gene or alternative codon optimization algorithms may give good or better results.

Retrieving multi-fasta CDS files for organisms from NCBI

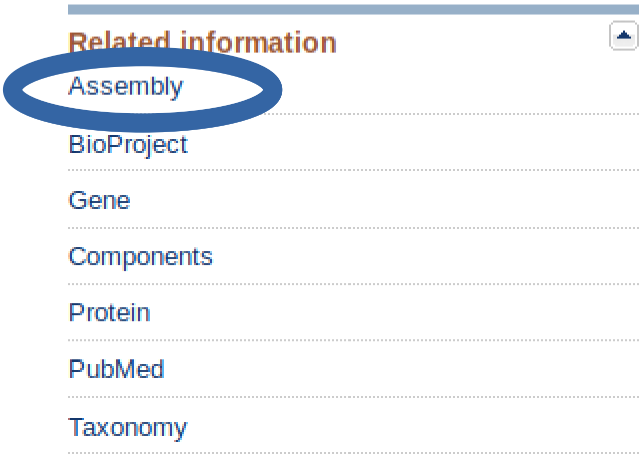
Search in NCBI genome:



Genome

Saccharomyces cerevisiae

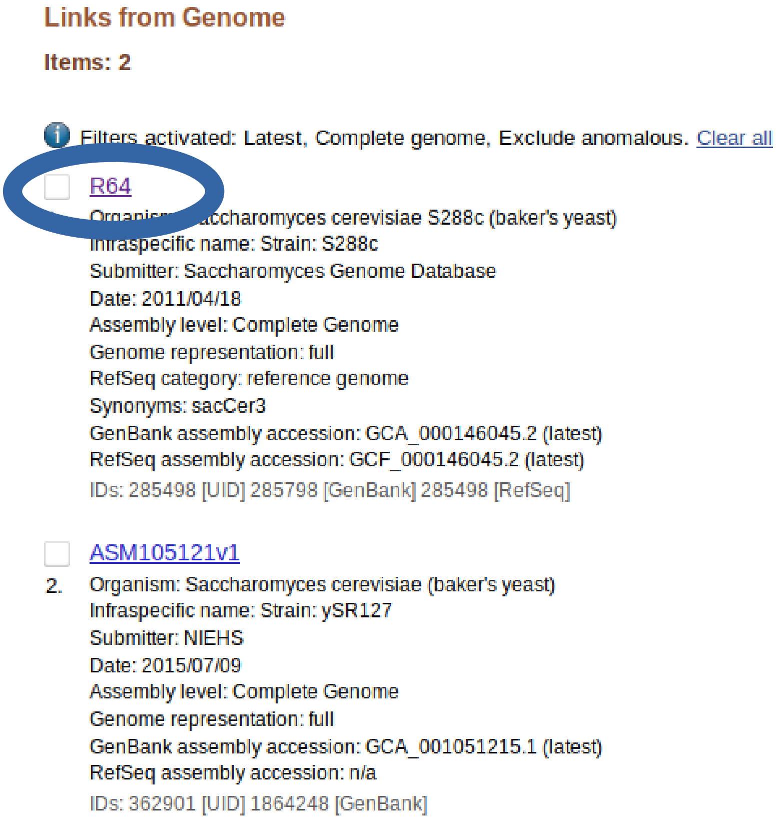
Select Assemblies:



Related information

- Assembly
- BioProject
- Gene
- Components
- Protein
- PubMed
- Taxonomy

Select the version you want:



Links from Genome

Items: 2

Filters activated: Latest, Complete genome, Exclude anomalous. Clear all

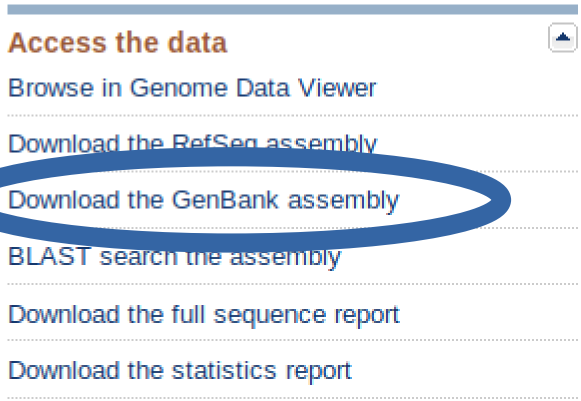
□ R64

Organism: Saccharomyces cerevisiae S288c (baker's yeast)
Infraspecific name: Strain: S288c
Submitter: Saccharomyces Genome Database
Date: 2011/04/18
Assembly level: Complete Genome
Genome representation: full
RefSeq category: reference genome
Synonyms: sacCer3
GenBank assembly accession: GCA_000146045.2 (latest)
RefSeq assembly accession: GCF_000146045.2 (latest)
IDs: 285498 [UID] 285798 [GenBank] 285498 [RefSeq]

□ ASM105121v1

2. Organism: Saccharomyces cerevisiae (baker's yeast)
Infraspecific name: Strain: ySR127
Submitter: NIEHS
Date: 2015/07/09
Assembly level: Complete Genome
Genome representation: full
GenBank assembly accession: GCA_001051215.1 (latest)
RefSeq assembly accession: n/a
IDs: 362901 [UID] 1864248 [GenBank]

Select “Download the GenBank assembly:



Access the data

Browse in Genome Data Viewer

Download the RefSeq assembly

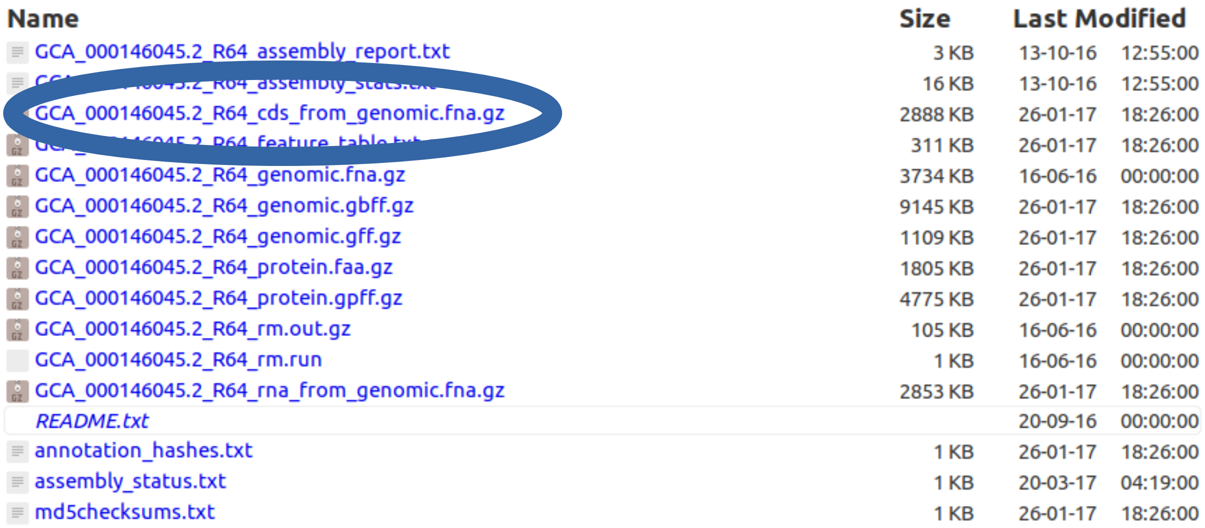
Download the GenBank assembly

BLAST search the assembly

Download the full sequence report

Download the statistics report

*Download the “*_cds_from_genomic.fna.gz” or copy the link (right click copy link location)*



Name	Size	Last Modified	
GCA_000146045.2_R64_assembly_report.txt	3 KB	13-10-16	12:55:00
GCA_000146045.2_R64_assembly_stats.txt	16 KB	13-10-16	12:55:00
GCA_000146045.2_R64_cds_from_genomic.fna.gz	2888 KB	26-01-17	18:26:00
GCA_000146045.2_R64_feature_table.txt.gz	311 KB	26-01-17	18:26:00
GCA_000146045.2_R64_genomic.fna.gz	3734 KB	16-06-16	00:00:00
GCA_000146045.2_R64_genomic.gbff.gz	9145 KB	26-01-17	18:26:00
GCA_000146045.2_R64_genomic.gff.gz	1109 KB	26-01-17	18:26:00
GCA_000146045.2_R64_protein.faa.gz	1805 KB	26-01-17	18:26:00
GCA_000146045.2_R64_protein.gpff.gz	4775 KB	26-01-17	18:26:00
GCA_000146045.2_R64_rm.out.gz	105 KB	16-06-16	00:00:00
GCA_000146045.2_R64_rm.run	1 KB	16-06-16	00:00:00
GCA_000146045.2_R64_rna_from_genomic.fna.gz	2853 KB	26-01-17	18:26:00
README.txt		20-09-16	00:00:00
annotation_hashes.txt	1 KB	26-01-17	18:26:00
assembly_status.txt	1 KB	20-03-17	04:19:00
md5checksums.txt	1 KB	26-01-17	18:26:00

References

Angov, E., Hillier, C. J., Kincaid, R. L. and Lyon, J. A. (2008) ‘Heterologous protein expression is enhanced by harmonizing the codon usage frequencies of the target gene with those of the expression host.’, *PloS one*, 3(5), p. e2189. doi: 10.1371/journal.pone.0002189.

Angov, E., Legler, P. M. and Mease, R. M. (2011) ‘Adjustment of codon usage frequencies by codon harmonization improves protein expression and folding’, *Methods Mol Biol*. 2010/12/03, 705, pp. 1–13. doi: 10.1007/978-1-61737-967-3_1.

Claassens, N. J., Siliakus, M. F., Nijsse, B., Spaans, S. K., Creutzburg, S. C. A., Schaap, P. J., Quax, T. E. F. and Oost, J. van der (2017) ‘Improving heterologous membrane protein production in *Escherichia coli* by combining transcriptional tuning and codon usage algorithms’, *PloS one*.