Interactive analysis of codon usage in prokaryotes

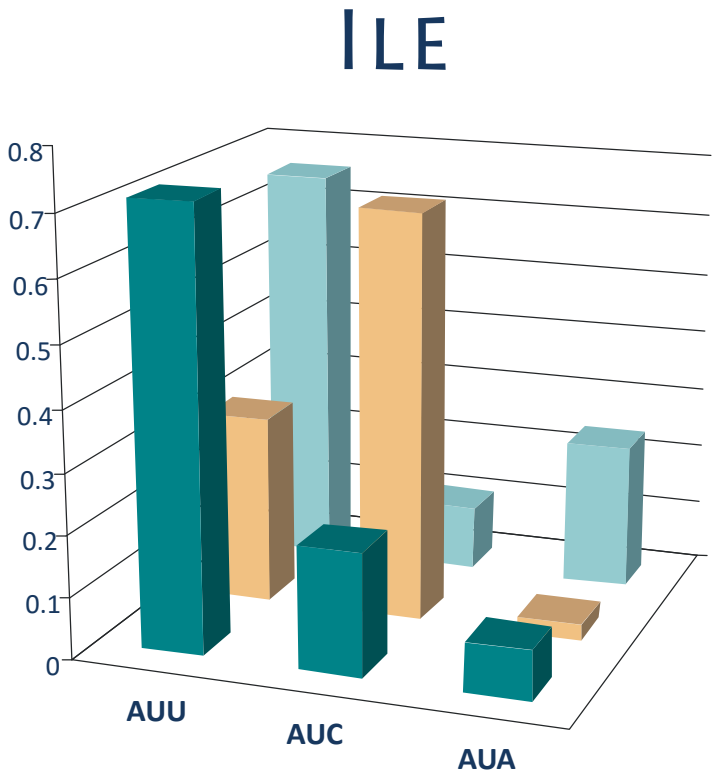
Predicting protein expressivity

Kristian Vlahoviček
Zagreb University
Croatia

Synonymous codon usage

1st position (5' end)	2nd position U	2nd position C	2nd position A	2nd position G	3rd position (3' end)
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	Stop	Stop	A
	Leu	Ser	Stop	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Related by Transitions in the 3rd Position



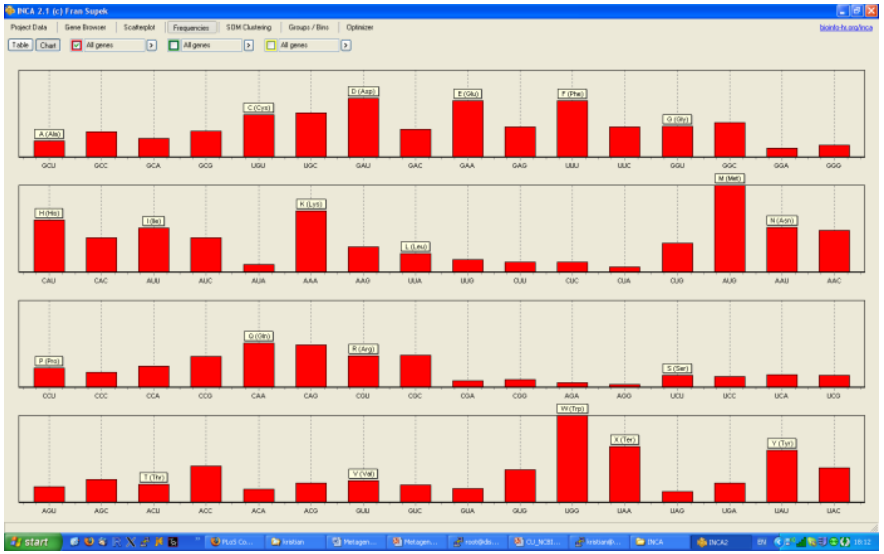
CU bias in microbial genomes

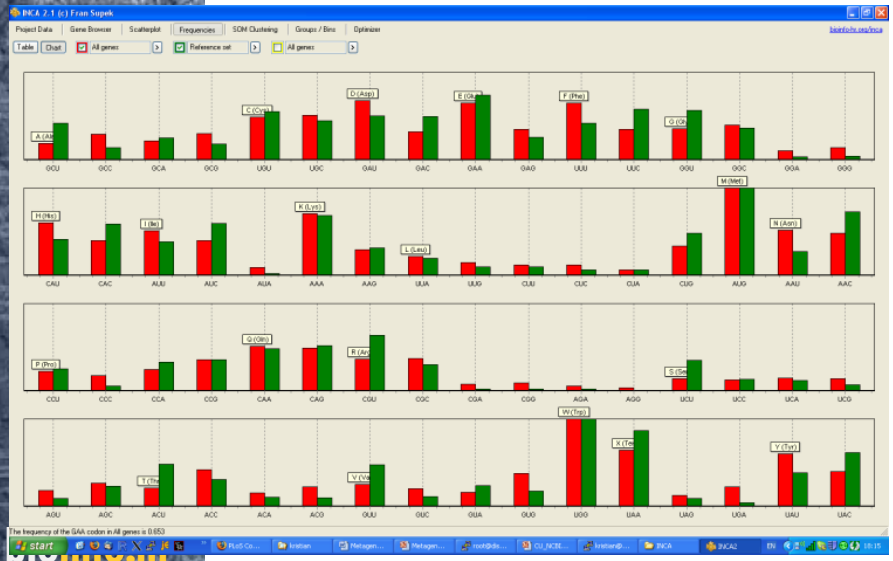
- Synonymous codons used differently
  - Between different species
    - GC content and AA composition
  - Within a single genome
    - “optimaly” encoded genes choose codons compatible to tRNA abundance and mRNA folding
      - Ribosomal proteins
      - Elongation factors
      - Chaperones
    - Background selection for “lifestyle specific” functions

‘Measuring’ Codon Usage

Take an ORF,
count 64 frequencies,
one for each codon

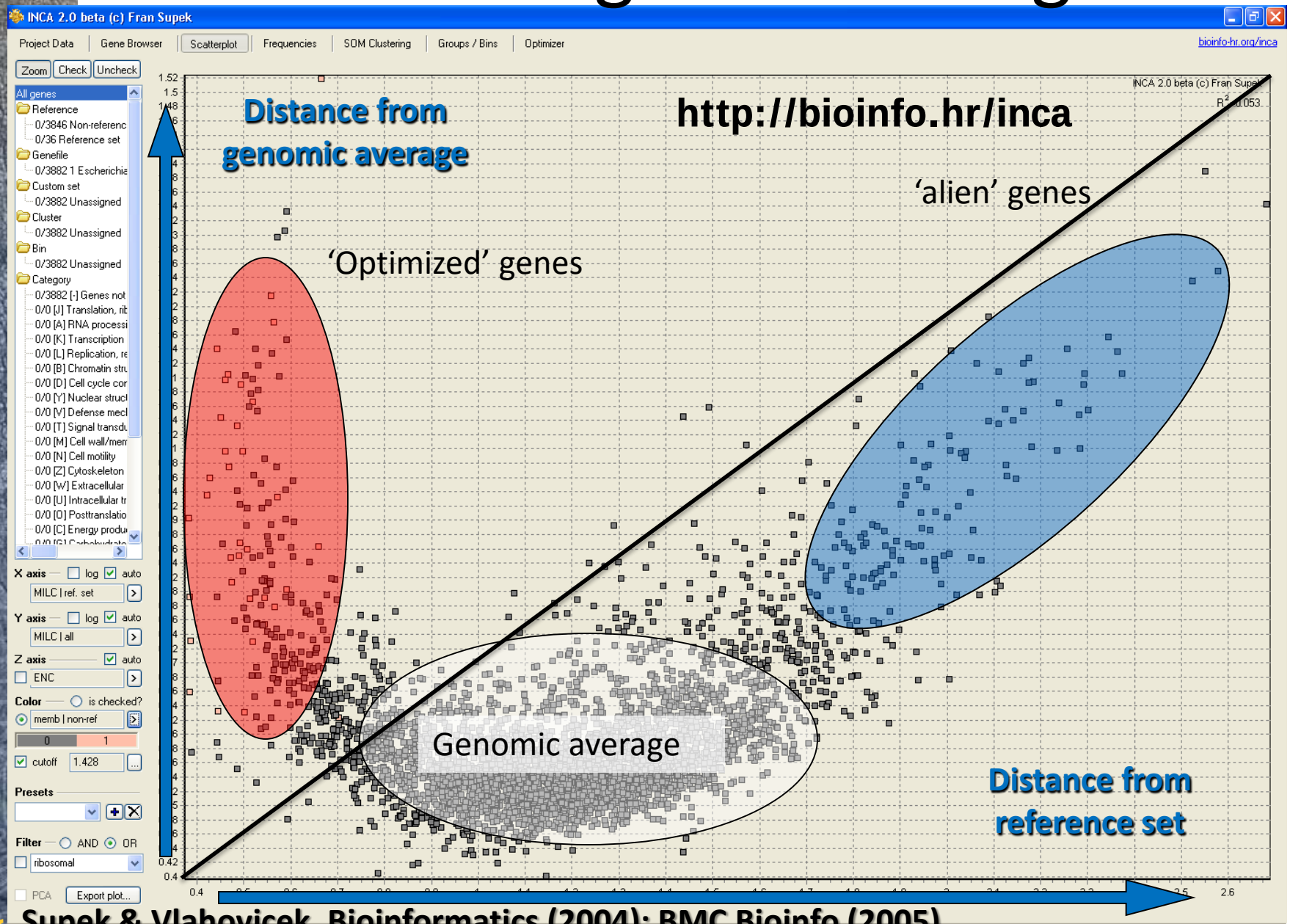
Single sequence
Whole genome
Reference sequence set





Compare CU distributions
Calculate ‘distance’

Measuring Codon Usage



Predicted expressivity

- Distance to genomic mean vs. distance to the reference set
  - ‘good’ codons ensure optimal expression rates
    - Synechocystis sp.: photosynthesis genes
    - M. janaschii: methanogenesis genes
    - D. radiodurans: membrane and detox proteins

Intermission: for the R-savy

- https://github.com/BioinfoHR/coRdon

- Work in progress
- Analysis of large-scale data
- Loads collections of .fasta files
  - Annotated or not
- Calculates codon frequencies
- Calculates distances
- Good for metagenomic data analysis

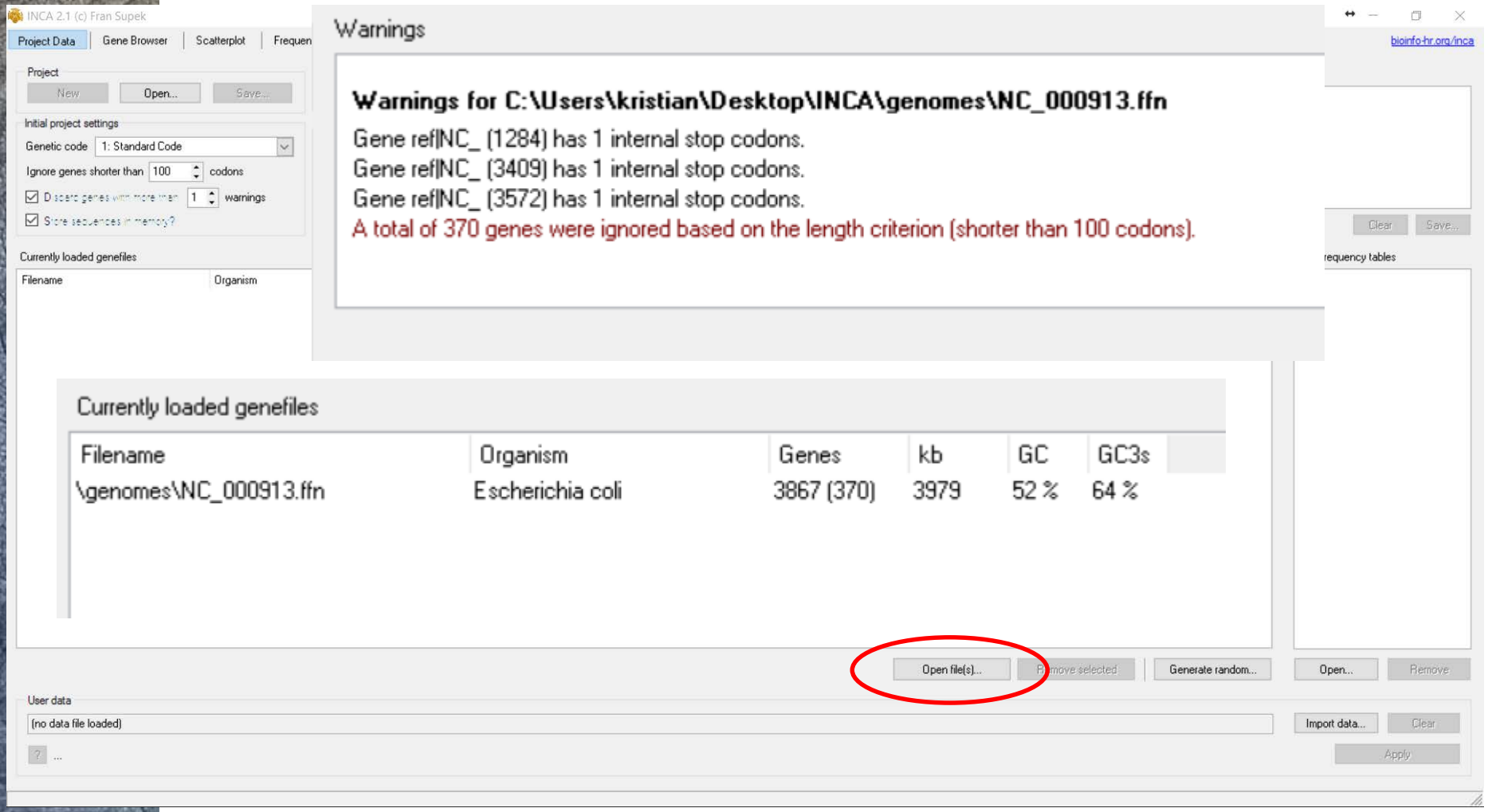
Maja Fabijanic, 2015

INCA – interactive codon usage analyzer

- Download

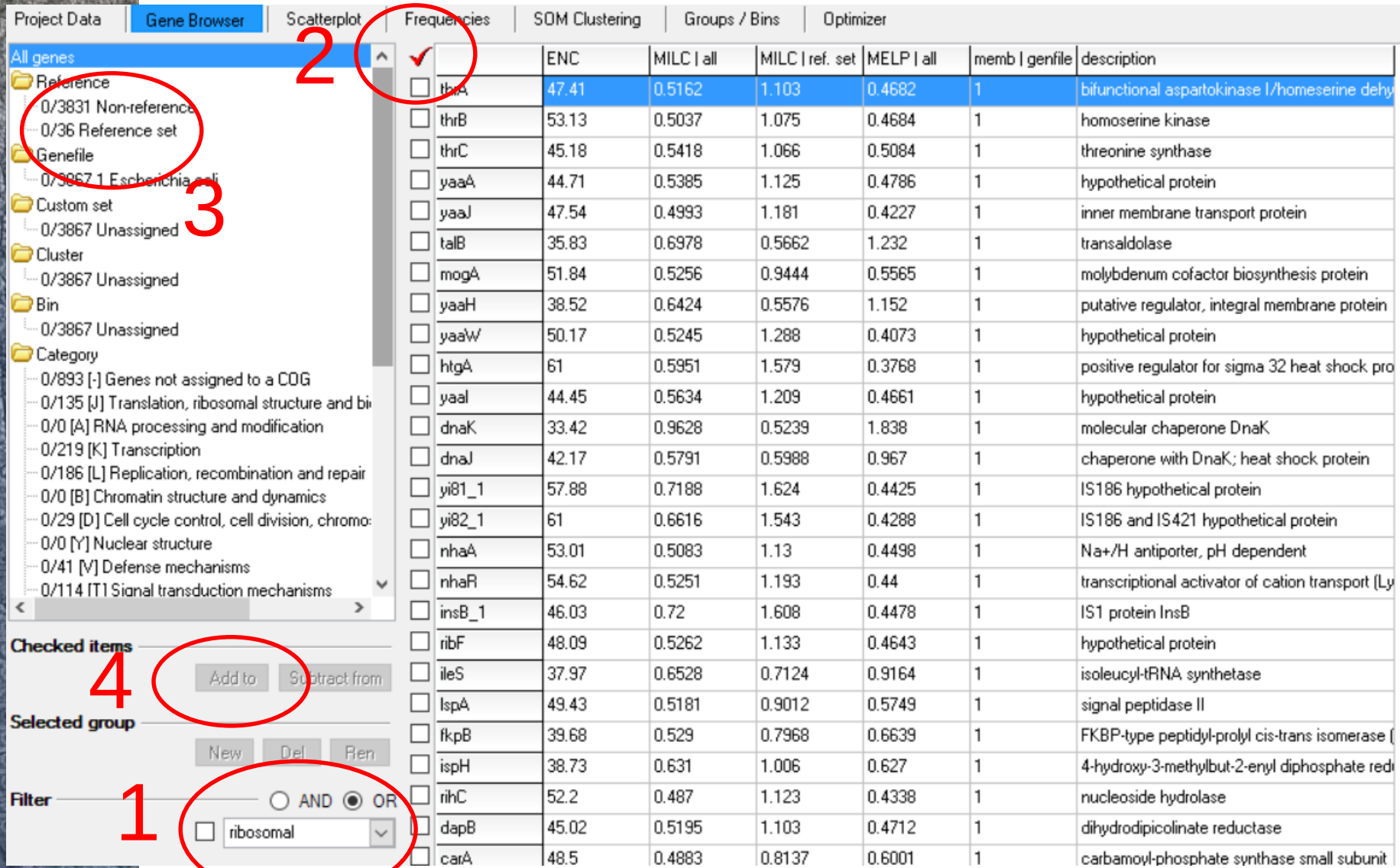
http://www.bioinfo.hr/research/inca/inca-registration/inca-download/

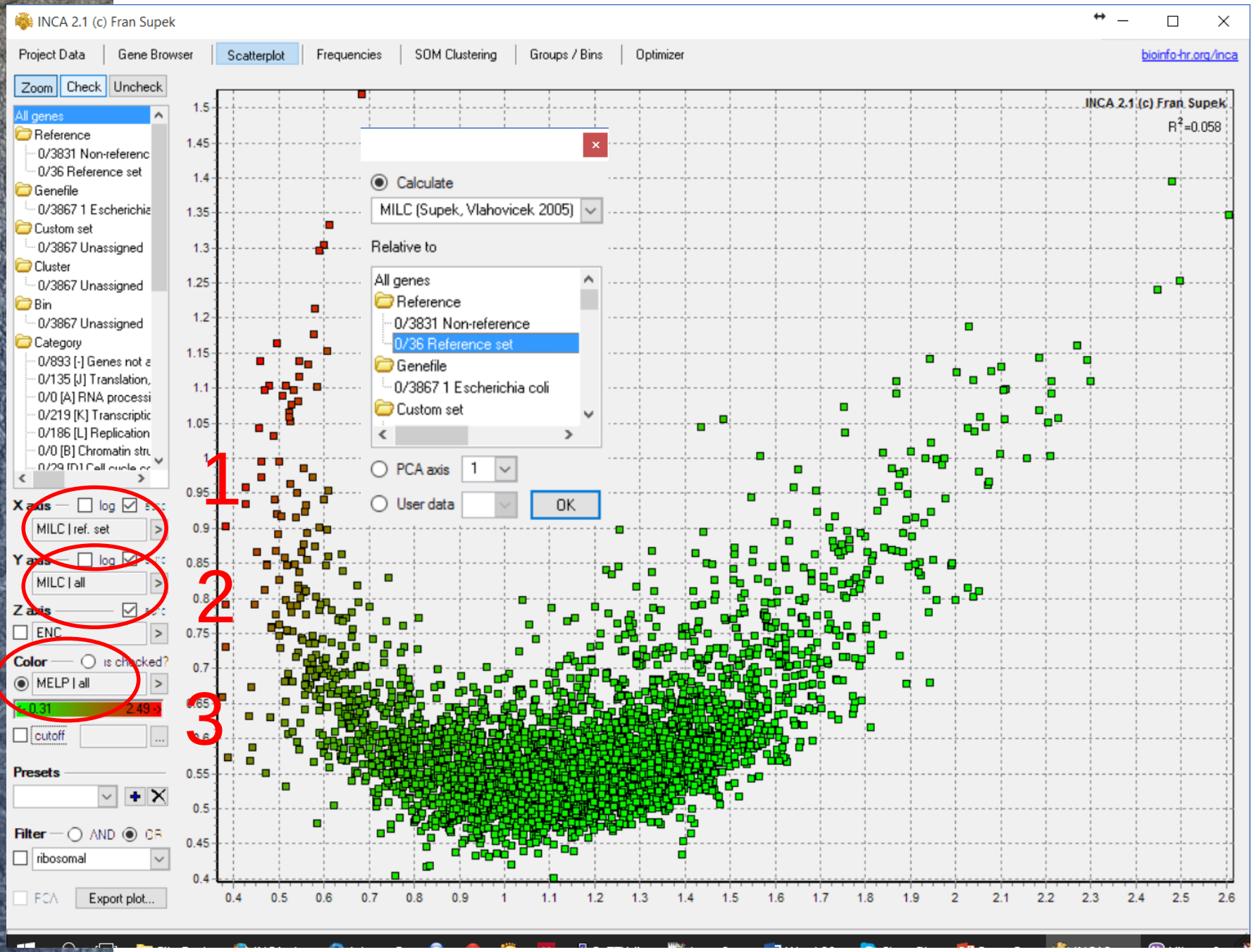
- Download version 2.1
- Unzip to Desktop
- Start INCA2
- Open file genomes/NC_000913.ffn


INCA 2.1 (c) Fran Supek
Project Data
Gene Browser
Scatterplot
Project
New
Open...
Save...
Initial project settings
Genetic code
1: Standard Code
Ignore genes shorter than
100
codons
warnings
Currently loaded genefiles
Filename
Organism
Warnings
Warnings for C:\Users\kristian\Desktop\INCA\genomes\NC_000913.ffn
Gene ref|NC_ (1284) has 1 internal stop codons.
Gene ref|NC_ (3409) has 1 internal stop codons.
Gene ref|NC_ (3572) has 1 internal stop codons.
A total of 370 genes were ignored based on the length criterion (shorter than 100 codons).
bioinfo-hr.org/inca
Clear
Save...
Currently loaded genefiles
Filename
Organism
Genes
kb
GC
GC3s
\genomes\NC_000913.ffn
Escherichia coli
3867 (370)
3979
52 %
64 %
Open file(s)...
Remove selected
Generate random...
Open...
Remove
User data
(no data file loaded)
Import data...
Clear
Apply

INCA Demo cont'd

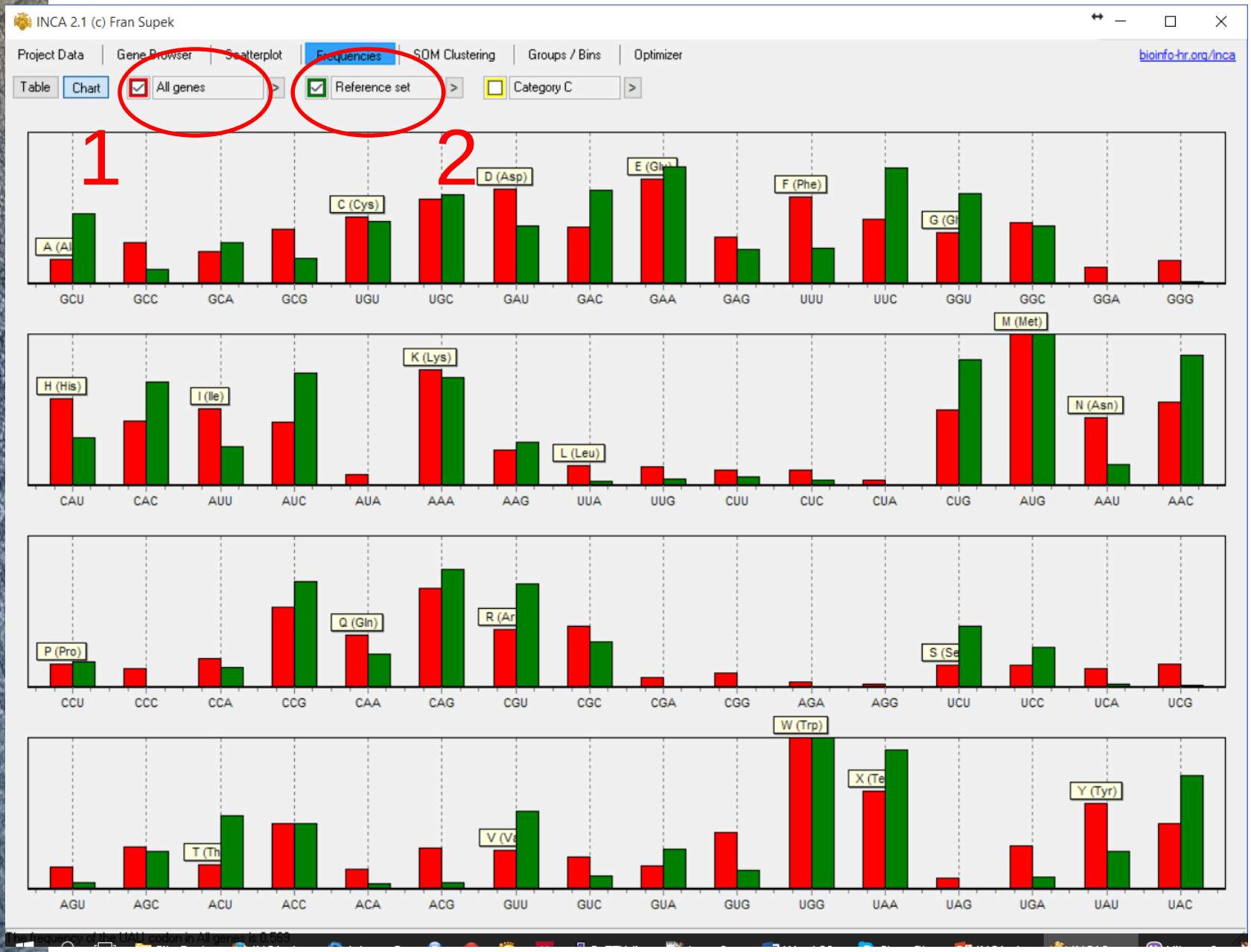
- Go to gene browser
- Filter genes by keyword 'ribosomal'
- Select all ribosomal protein genes
- Add them to reference set
- Visualize the scatterplot
  - Select different preset methods
  - Compare MILC and Karlin&Mrazek plots

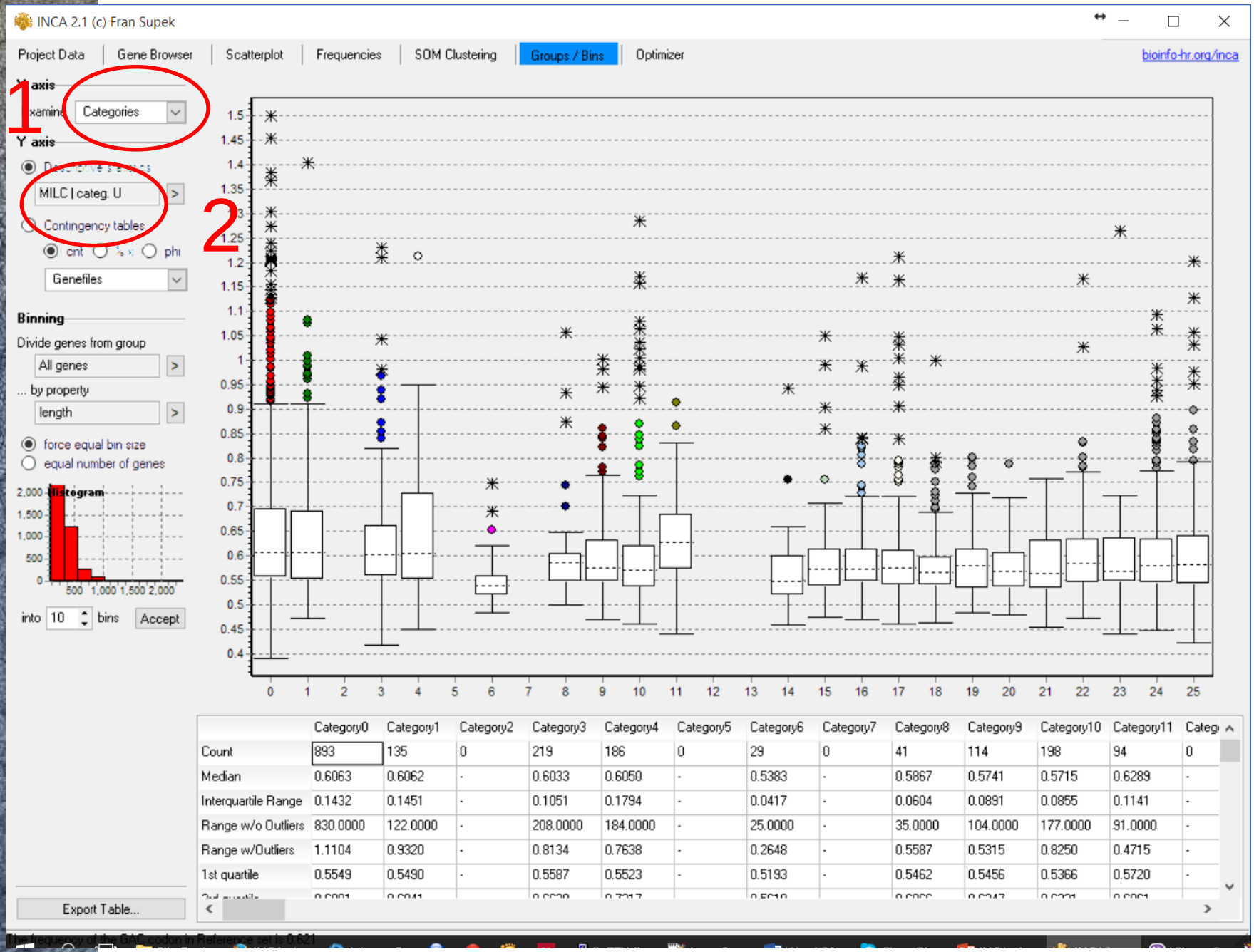

Project Data
Gene Browser
Scatterplot
Frequencies
SOM Clustering
Groups / Bins
Optimizer
All genes
Reference
0/3831 Non-reference
0/36 Reference set
Genefile
Custom set
0/3867 Unassigned
Cluster
0/3867 Unassigned
Bin
0/3867 Unassigned
Category
0/893 [-] Genes not assigned to a COG
0/135 [J] Translation, ribosomal structure and bi
0/0 [A] RNA processing and modification
0/219 [K] Transcription
0/186 [L] Replication, recombination and repair
0/0 [B] Chromatin structure and dynamics
0/29 [D] Cell cycle control, cell division, chromo
0/0 [Y] Nuclear structure
0/41 [V] Defense mechanisms
0/114 [T] Signal transduction mechanisms
Checked items
Add to
Subtract from
Selected group
New
Del
Ren
Filter
AND
OR
ribosomal
1
2
3
4
ENC
MILC | all
MILC | ref. set
MELP | all
memb | genfile
description
thrA 47.41 0.5162 1.103 0.4682 1 bifunctional aspartokinase I/homeserine dehy
thrB 53.13 0.5037 1.075 0.4684 1 homoserine kinase
thrC 45.18 0.5418 1.066 0.5084 1 threonine synthase
yaaA 44.71 0.5385 1.125 0.4786 1 hypothetical protein
yaaJ 47.54 0.4993 1.181 0.4227 1 inner membrane transport protein
talB 35.83 0.6978 0.5662 1.232 1 transaldolase
mogA 51.84 0.5256 0.9444 0.5565 1 molybdenum cofactor biosynthesis protein
yaaH 38.52 0.6424 0.5576 1.152 1 putative regulator, integral membrane protein
yaaW 50.17 0.5245 1.288 0.4073 1 hypothetical protein
htgA 61 0.5951 1.579 0.3768 1 positive regulator for sigma 32 heat shock pro
yaaI 44.45 0.5634 1.209 0.4661 1 hypothetical protein
dnaK 33.42 0.9628 0.5239 1.838 1 molecular chaperone DnaK
dnaJ 42.17 0.5791 0.5988 0.967 1 chaperone with DnaK; heat shock protein
yi81_1 57.88 0.7188 1.624 0.4425 1 IS186 hypothetical protein
yi82_1 61 0.6616 1.543 0.4288 1 IS186 and IS421 hypothetical protein
nhaA 53.01 0.5083 1.13 0.4498 1 Na+/H antiporter, pH dependent
nhaR 54.62 0.5251 1.193 0.44 1 transcriptional activator of cation transport (Ly
insB_1 46.03 0.72 1.608 0.4478 1 IS1 protein InsB
ribF 48.09 0.5262 1.133 0.4643 1 hypothetical protein
ileS 37.97 0.6528 0.7124 0.9164 1 isoleucyl-tRNA synthetase
lspA 49.43 0.5181 0.9012 0.5749 1 signal peptidase II
fkpB 39.68 0.529 0.7968 0.6639 1 FKBP-type peptidyl-prolyl cis-trans isomerase
ispH 38.73 0.631 1.006 0.627 1 4-hydroxy-3-methylbut-2-enyl diphosphate red
rihC 52.2 0.487 1.123 0.4338 1 nucleoside hydrolase
dapB 45.02 0.5195 1.103 0.4712 1 dihydrodipicolinate reductase
carA 48.5 0.4883 0.8137 0.6001 1 carbamoyl-phosphate synthase small subunit


INCA 2.1 (c) Fran Supek
Project Data
Gene Browser
Scatterplot
Frequencies
SOM Clustering
Groups / Bins
Optimizer
bioinfo-hr.org/inca
Zoom
Check
Uncheck
All genes
Reference
0/3831 Non-referenc
0/36 Reference set
Genefile
0/3867 1 Escherichia
Custom set
0/3867 Unassigned
Cluster
0/3867 Unassigned
Bin
0/3867 Unassigned
Category
0/893 [-] Genes not a
0/135 [J] Translation,
0/0 [A] RNA processi
0/219 [K] Transcriptic
0/186 [L] Replication
0/0 [B] Chromatin stru
X axis
log
MILC | ref. set
Y axis
log
MILC | all
Z axis
ENC
Color
is checked?
MELP | all
0.31
2.49
cutoff
Presets
Filter
AND
OR
ribosomal
PCA
Export plot...
INCA 2.1 (c) Fran Supek
R2=0.058
Calculate
MILC (Supek, Vlahovicek 2005)
Relative to
All genes
Reference
0/3831 Non-reference
0/36 Reference set
Genefile
0/3867 1 Escherichia coli
Custom set
PCA axis
1
User data
OK
1
2
3

More tricks

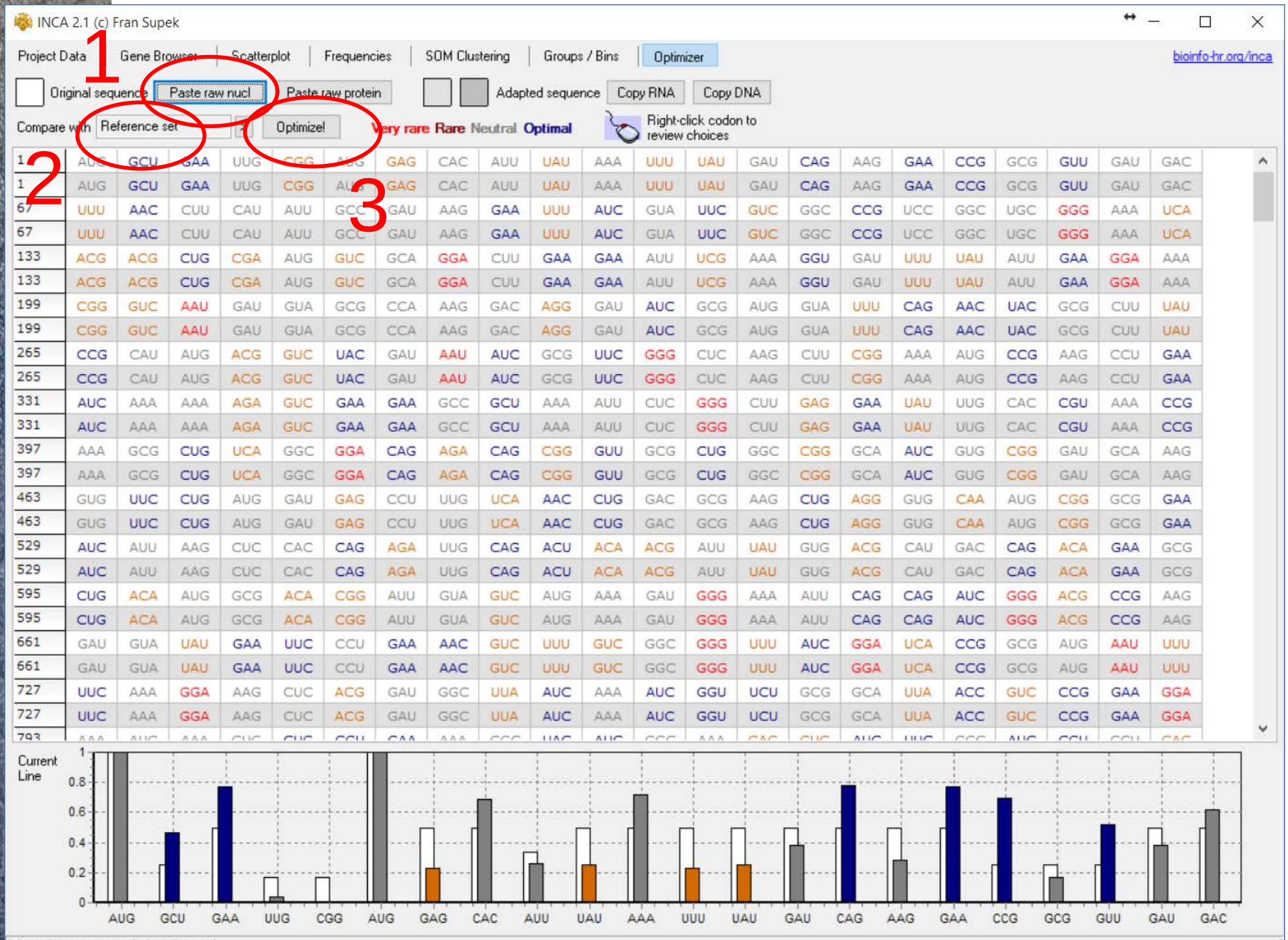
- Visualize codon usage frequencies in the reference set and the whole genome
- View expression prediction binned by COG categories
  - Go to Groups/bins
  - Select Categories on X axis
  - Select MELP on Y axis


INCA 2.1 (c) Fran Supek
Project Data
Gene Browser
Scatterplot
Frequencies
SOM Clustering
Groups / Bins
Optimizer
bioinfo-hr.org/inca
Table
Chart
All genes
Reference set
Category C
1
2
A (Al
C (Cys)
D (Asp)
E (Gl
F (Phe)
G (Gl
GCU GCC GCA GCG UGU UGC GAU GAC GAA GAG UUU UUC GGU GGC GGA GGG
H (His)
I (Ile)
K (Lys)
L (Leu)
M (Met)
N (Asn)
CAU CAC AUU AUC AUA AAA AAG UUA UUG CUU CUC CUA CUG AUG AAU AAC
P (Pro)
Q (Gln)
R (Ar
S (Se
CCU CCC CCA CCG CAA CAG CGU CGC CGA CGG AGA AGG UCU UCC UCA UCG
W (Trp)
X (Te
Y (Tyr)
T (Th
V (Va
AGU AGC ACU ACC ACA ACG GUU GUC GUA GUG UGG UAA UAG UGA UAU UAC


INCA 2.1 (c) Fran Supek
Project Data
Gene Browser
Scatterplot
Frequencies
SOM Clustering
Groups / Bins
Optimizer
bioinfo-hr.org/inca
1
X axis
examine
Categories
Y axis
MILC | categ. U
Contingency tables
cnt
phi
Genefiles
2
Binning
Divide genes from group
All genes
... by property
length
force equal bin size
equal number of genes
Histogram
into 10 bins
Accept
Export Table...
Category0 Category1 Category2 Category3 Category4 Category5 Category6 Category7 Category8 Category9 Category10 Category11
Count 893 135 0 219 186 0 29 0 41 114 198 94 0
Median 0.6063 0.6062 - 0.6033 0.6050 - 0.5383 - 0.5867 0.5741 0.5715 0.6289 -
Interquartile Range 0.1432 0.1451 - 0.1051 0.1794 - 0.0417 - 0.0604 0.0891 0.0855 0.1141 -
Range w/o Outliers 830.0000 122.0000 - 208.0000 184.0000 - 25.0000 - 35.0000 104.0000 177.0000 91.0000 -
Range w/Outliers 1.1104 0.9320 - 0.8134 0.7638 - 0.2648 - 0.5587 0.5315 0.8250 0.4715 -
1st quartile 0.5549 0.5490 - 0.5587 0.5523 - 0.5193 - 0.5462 0.5456 0.5366 0.5720 -

Are you into over-expression?

- Fetch a file from http://hex.bioinfo.hr/~kristian/my_overexpression_target.txt
- Paste into optimizer
  - Notice the very rare codons!
- Optimize towards the reference set
- Optimize towards all genes
- Synthesize, clone into *E. coli*, over-express and compare ☺


INCA 2.1 (c) Fran Supek
Project Data
Gene Browser
Scatterplot
Frequencies
SOM Clustering
Groups / Bins
Optimizer
bioinfo-hr.org/inca
1
Original sequence
Paste raw nucl
Paste raw protein
Adapted sequence
Copy RNA
Copy DNA
Compare with
Reference set
2
Optimize!
Very rare Rare Neutral Optimal
Right-click codon to review choices
2
3
1 AUG GCU GAA UUG CGG AUG GAG CAC AUU UAU AAA UUU UAU GAU CAG AAG GAA CCG GCG GUU GAU GAC
1 AUG GCU GAA UUG CGG AUG GAG CAC AUU UAU AAA UUU UAU GAU CAG AAG GAA CCG GCG GUU GAU GAC
67 UUU AAC CUU CAU AUU GCC GAU AAG GAA UUU AUC GUA UUC GUC GGC CCG UCC GGC UGC GGG AAA UCA
67 UUU AAC CUU CAU AUU GCC GAU AAG GAA UUU AUC GUA UUC GUC GGC CCG UCC GGC UGC GGG AAA UCA
133 ACG ACG CUG CGA AUG GUC GCA GGA CUU GAA GAA AUU UCG AAA GGU GAU UUU UAU AUU GAA GGA AAA
133 ACG ACG CUG CGA AUG GUC GCA GGA CUU GAA GAA AUU UCG AAA GGU GAU UUU UAU AUU GAA GGA AAA
199 CGG GUC AAU GAU GUA GCG CCA AAG GAC AGG GAU AUC GCG AUG GUA UUU CAG AAC UAC GCG CUU UAU
199 CGG GUC AAU GAU GUA GCG CCA AAG GAC AGG GAU AUC GCG AUG GUA UUU CAG AAC UAC GCG CUU UAU
265 CCG CAU AUG ACG GUC UAC GAU AAU AUC GCG UUC GGG CUC AAG CUU CGG AAA AUG CCG AAG CCU GAA
265 CCG CAU AUG ACG GUC UAC GAU AAU AUC GCG UUC GGG CUC AAG CUU CGG AAA AUG CCG AAG CCU GAA
331 AUC AAA AAA AGA GUC GAA GAA GCC GCU AAA AUU CUC GGG CUU GAG GAA UAU UUG CAC CGU AAA CCG
331 AUC AAA AAA AGA GUC GAA GAA GCC GCU AAA AUU CUC GGG CUU GAG GAA UAU UUG CAC CGU AAA CCG
397 AAA GCG CUG UCA GGC GGA CAG AGA CAG CGG GUU GCG CUG GGC CGG GCA AUC GUG CGG GAU GCA AAG
397 AAA GCG CUG UCA GGC GGA CAG AGA CAG CGG GUU GCG CUG GGC CGG GCA AUC GUG CGG GAU GCA AAG
463 GUG UUC CUG AUG GAU GAG CCU UUG UCA AAC CUG GAC GCG AAG CUG AGG GUG CAA AUG CGG GCG GAA
463 GUG UUC CUG AUG GAU GAG CCU UUG UCA AAC CUG GAC GCG AAG CUG AGG GUG CAA AUG CGG GCG GAA
529 AUC AUU AAG CUC CAC CAG AGA UUG CAG ACU ACA ACG AUU UAU GUG ACG CAU GAC CAG ACA GAA GCG
529 AUC AUU AAG CUC CAC CAG AGA UUG CAG ACU ACA ACG AUU UAU GUG ACG CAU GAC CAG ACA GAA GCG
595 CUG ACA AUG GCG ACA CGG AUU GUA GUC AUG AAA GAU GGG AAA AUU CAG CAG AUC GGG ACG CCG AAG
595 CUG ACA AUG GCG ACA CGG AUU GUA GUC AUG AAA GAU GGG AAA AUU CAG CAG AUC GGG ACG CCG AAG
661 GAU GUA UAU GAA UUC CCU GAA AAC GUC UUU GUC GGC GGG UUU AUC GGA UCA CCG GCG AUG AAU UUU
661 GAU GUA UAU GAA UUC CCU GAA AAC GUC UUU GUC GGC GGG UUU AUC GGA UCA CCG GCG AUG AAU UUU
727 UUC AAA GGA AAG CUC ACG GAU GGC UUA AUC AAA AUC GGU UCU GCG GCA UUA ACC GUC CCG GAA GGA
727 UUC AAA GGA AAG CUC ACG GAU GGC UUA AUC AAA AUC GGU UCU GCG GCA UUA ACC GUC CCG GAA GGA
Current Line
1
0.8
0.6
0.4
0.2
0
AUG GCU GAA UUG CGG AUG GAG CAC AUU UAU AAA UUU UAU GAU CAG AAG GAA CCG GCG GUU GAU GAC
Adapted sequence copied to clipboard.