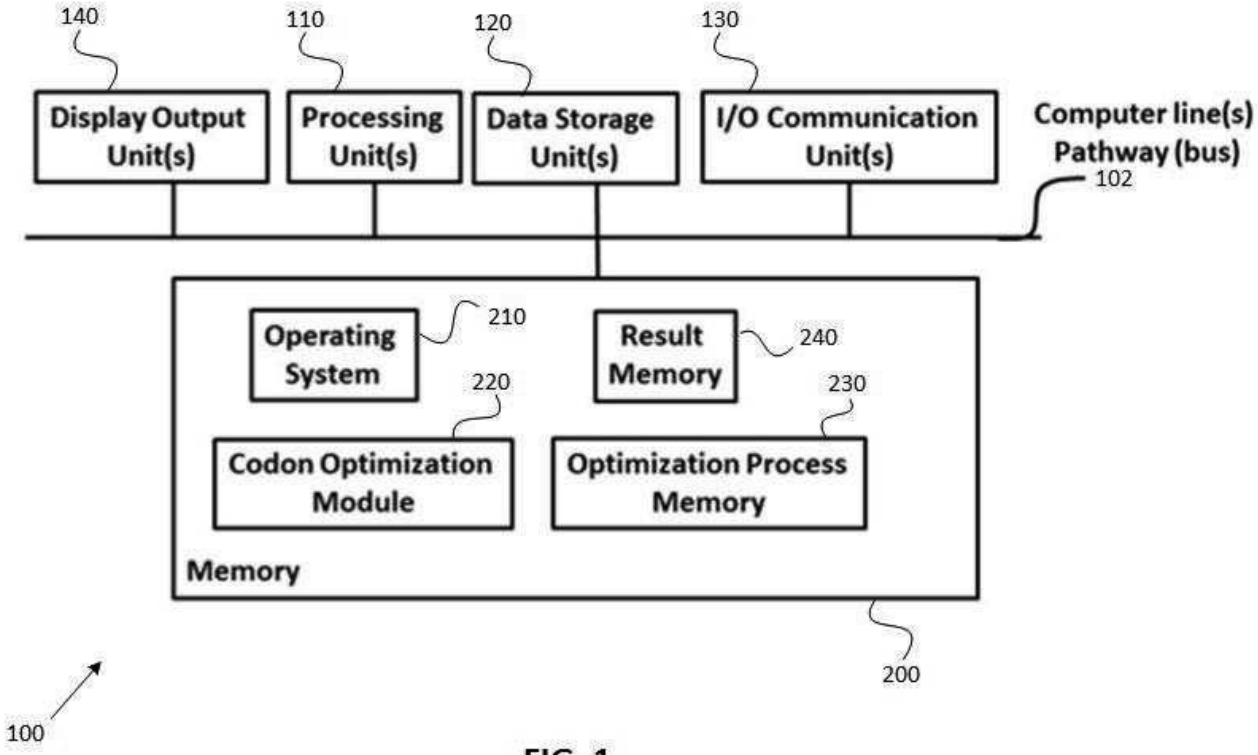

140
110
120
130
Display Output Unit(s)
Processing Unit(s)
Data Storage Unit(s)
I/O Communication Unit(s)
Computer line(s) Pathway (bus)
102
Operating System
210
Result Memory
240
220
230
Codon Optimization Module
Optimization Process Memory
Memory
200
100

FIG. 1

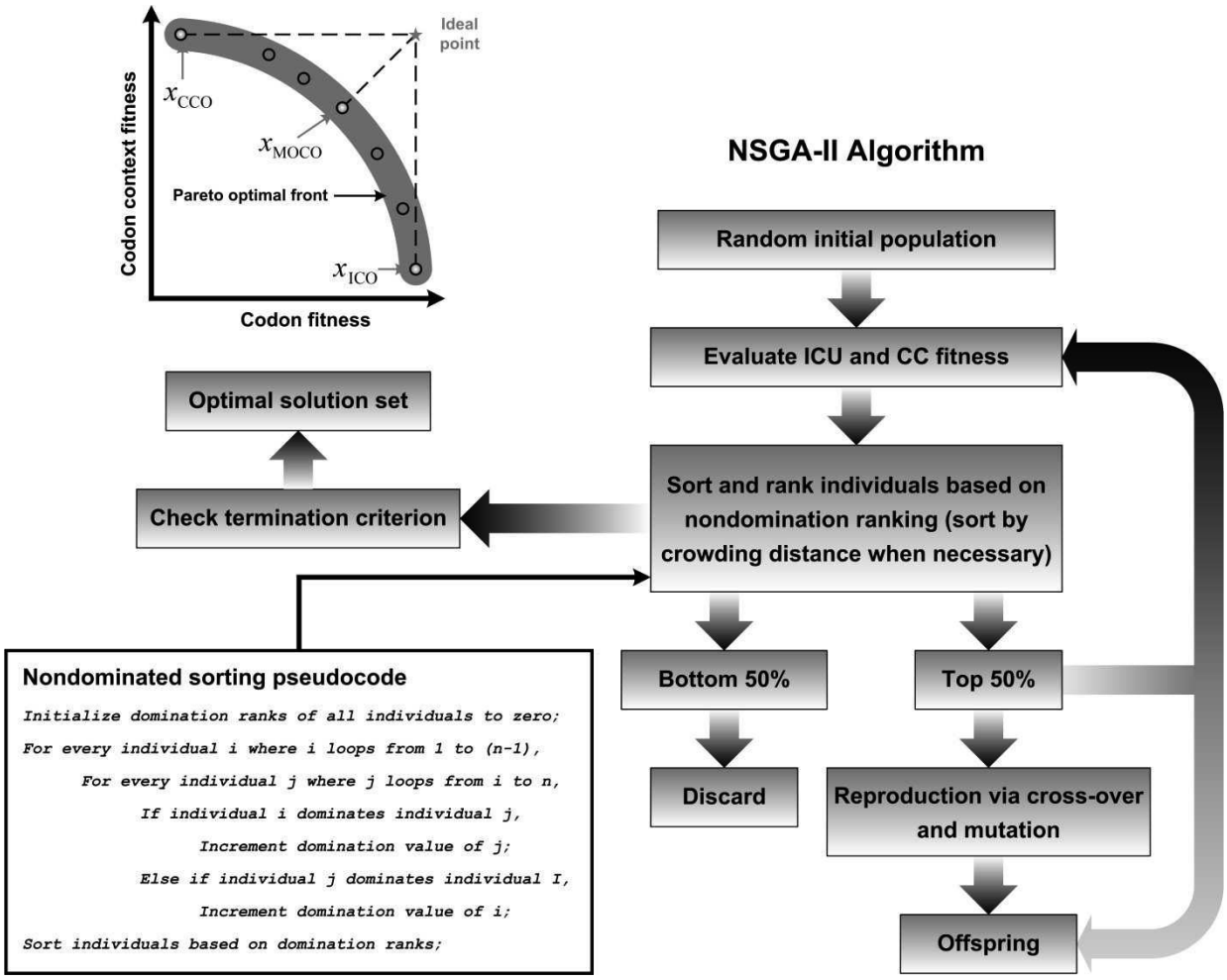

Ideal point
x_{CCO}
x_{MOCO}
Pareto optimal front
x_{ICO}
Codon context fitness
Codon fitness
NSGA-II Algorithm
Random initial population
Evaluate ICU and CC fitness
Sort and rank individuals based on nondomination ranking (sort by crowding distance when necessary)
Bottom 50%
Top 50%
Discard
Reproduction via cross-over and mutation
Offspring
Check termination criterion
Optimal solution set
Nondominated sorting pseudocode
Initialize domination ranks of all individuals to zero;
For every individual i where i loops from 1 to (n-1),
For every individual j where j loops from i to n,
If individual i dominates individual j,
Increment domination value of j;
Else if individual j dominates individual I,
Increment domination value of i;
Sort individuals based on domination ranks;

Figure 2

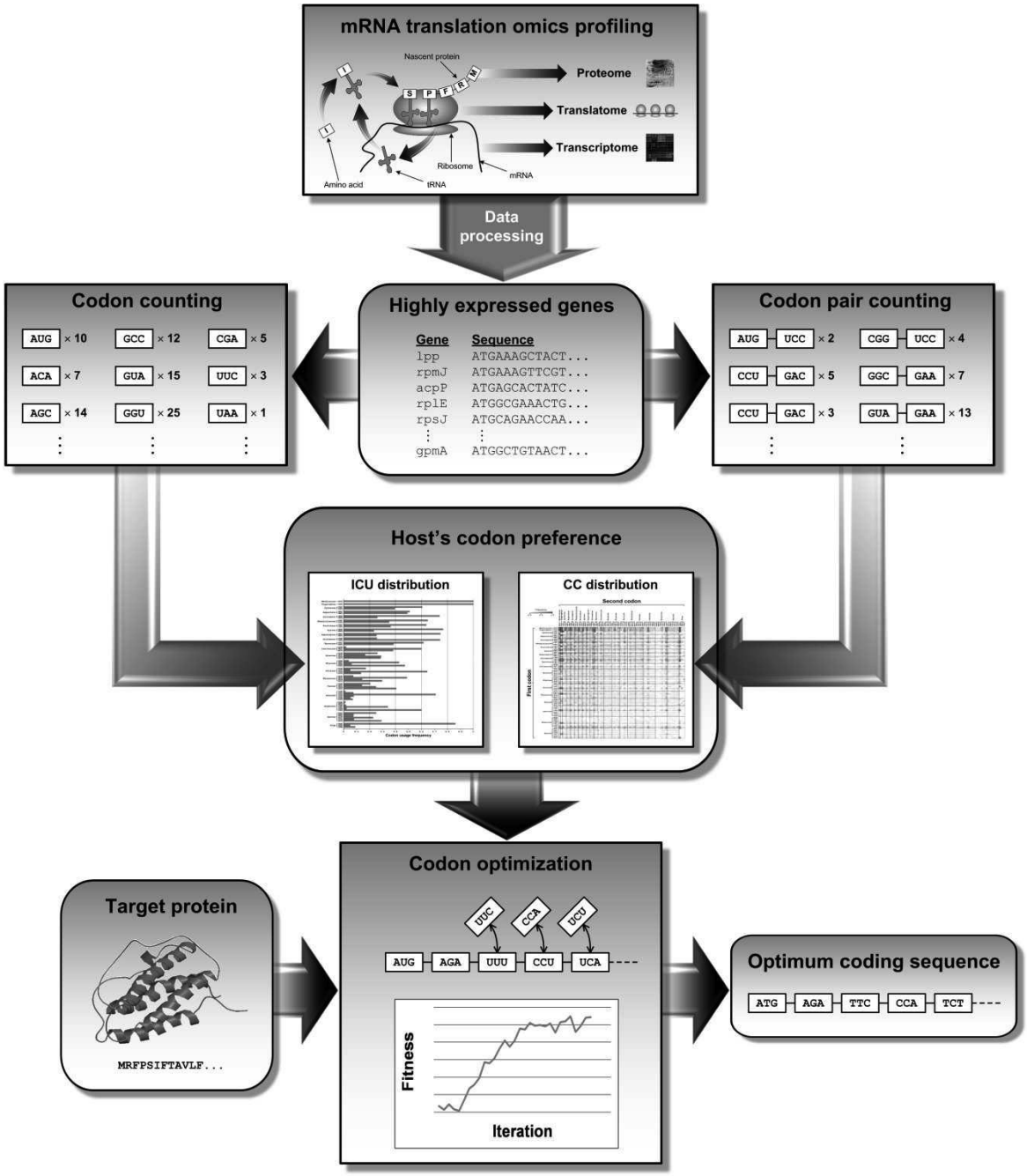

mRNA translation omics profiling
Nascent protein
Proteome
Translatome
Transcriptome
Ribosome
mRNA
tRNA
Amino acid
Data processing
Codon counting
AUG × 10
GCC × 12
CGA × 5
ACA × 7
GUA × 15
UUC × 3
AGC × 14
GGU × 25
UAA × 1
Highly expressed genes
Gene Sequence
lpp ATGAAAGCTACT...
rpmJ ATGAAAGTTCGT...
acpP ATGAGCACTATC...
rplE ATGGCGAAACTG...
rpsJ ATGCAGAACCAA...
gpmA ATGGCTGTAACT...
Codon pair counting
AUG UCC × 2
CGG UCC × 4
CCU GAC × 5
GGC GAA × 7
CCU GAC × 3
GUA GAA × 13
Host's codon preference
ICU distribution
CC distribution
Second codon
First codon
Codon optimization
UUC
CCA
UCU
AUG AGA UUU CCU UCA
Fitness
Iteration
Target protein
MRFPSIFTAVLF...
Optimum coding sequence
ATG AGA TTC CCA TCT

Figure 3

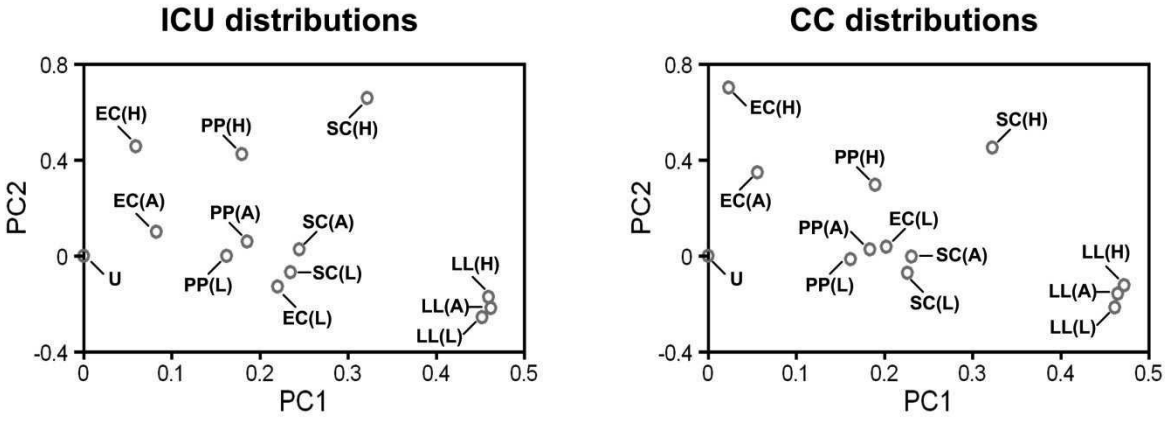
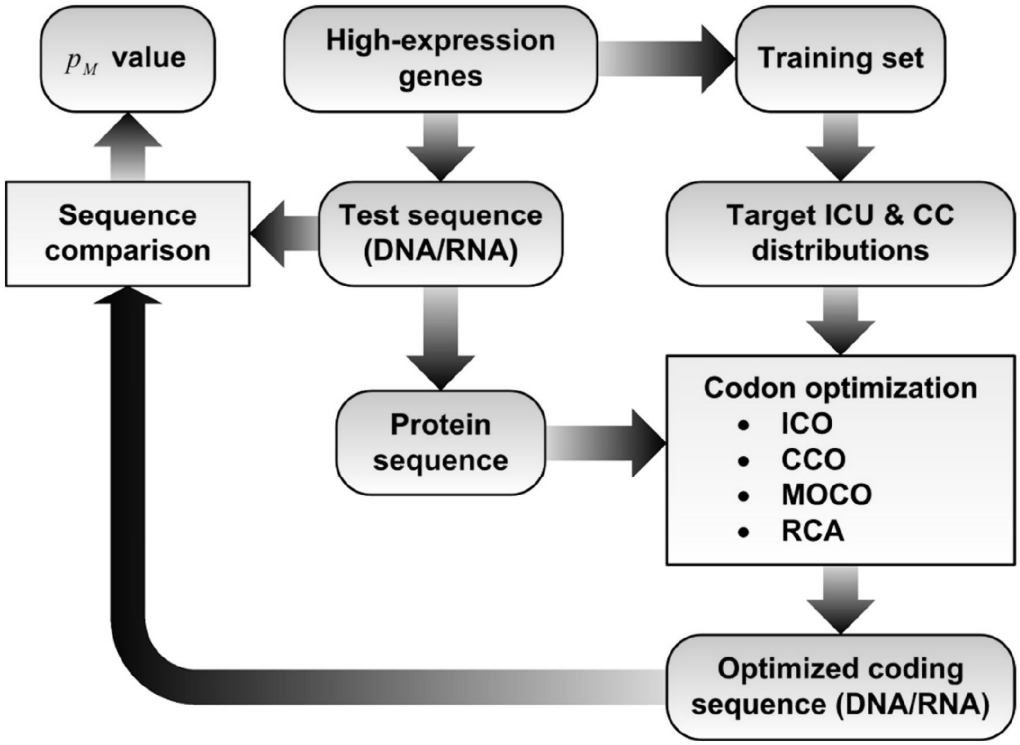

ICU distributions
CC distributions
PC1
PC2
EC(H)
PP(H)
SC(H)
EC(A)
PP(A)
SC(A)
U
PP(L)
SC(L)
EC(L)
LL(H)
LL(A)
LL(L)

Figure 4


p_M value
High-expression genes
Training set
Sequence comparison
Test sequence (DNA/RNA)
Target ICU & CC distributions
Protein sequence
Codon optimization
ICO
CCO
MOCO
RCA
Optimized coding sequence (DNA/RNA)

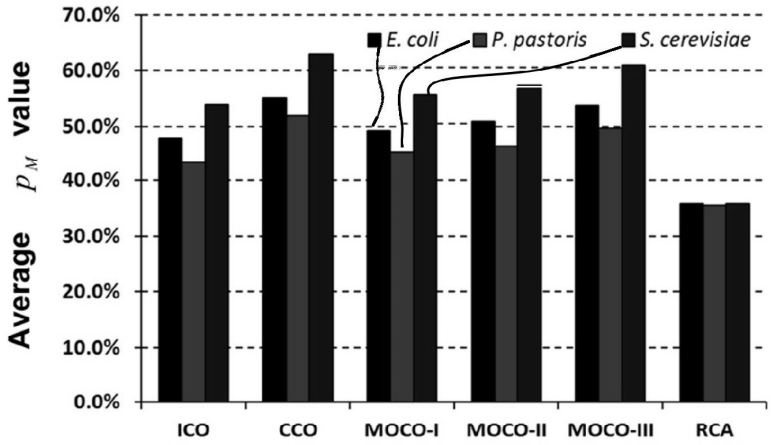
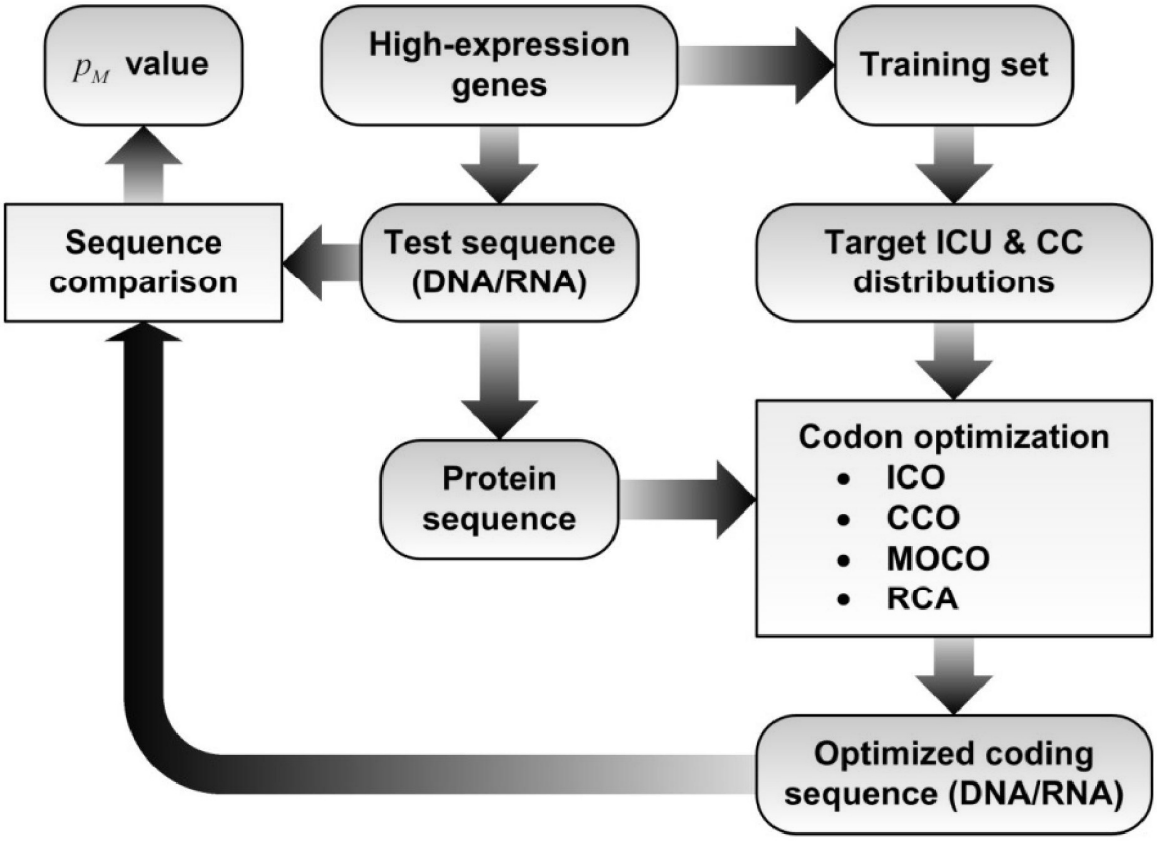

Average p_M value
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
E. coli
P. pastoris
S. cerevisiae
ICO
CCO
MOCO-I
MOCO-II
MOCO-III
RCA

FIG. 5A


p_M value
High-expression genes
Training set
Sequence comparison
Test sequence (DNA/RNA)
Target ICU & CC distributions
Protein sequence
Codon optimization
ICO
CCO
MOCO
RCA
Optimized coding sequence (DNA/RNA)

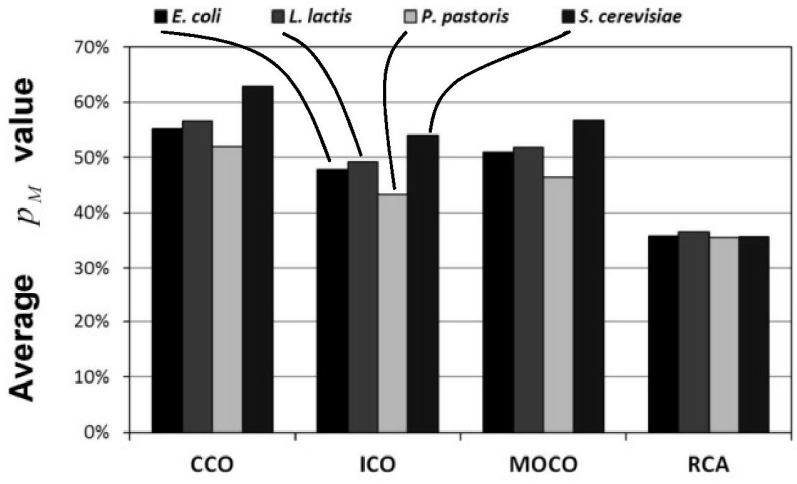

E. coli
L. lactis
P. pastoris
S. cerevisiae
Average p_M value
70%
60%
50%
40%
30%
20%
10%
0%
CCO
ICO
MOCO
RCA

FIG. 5B

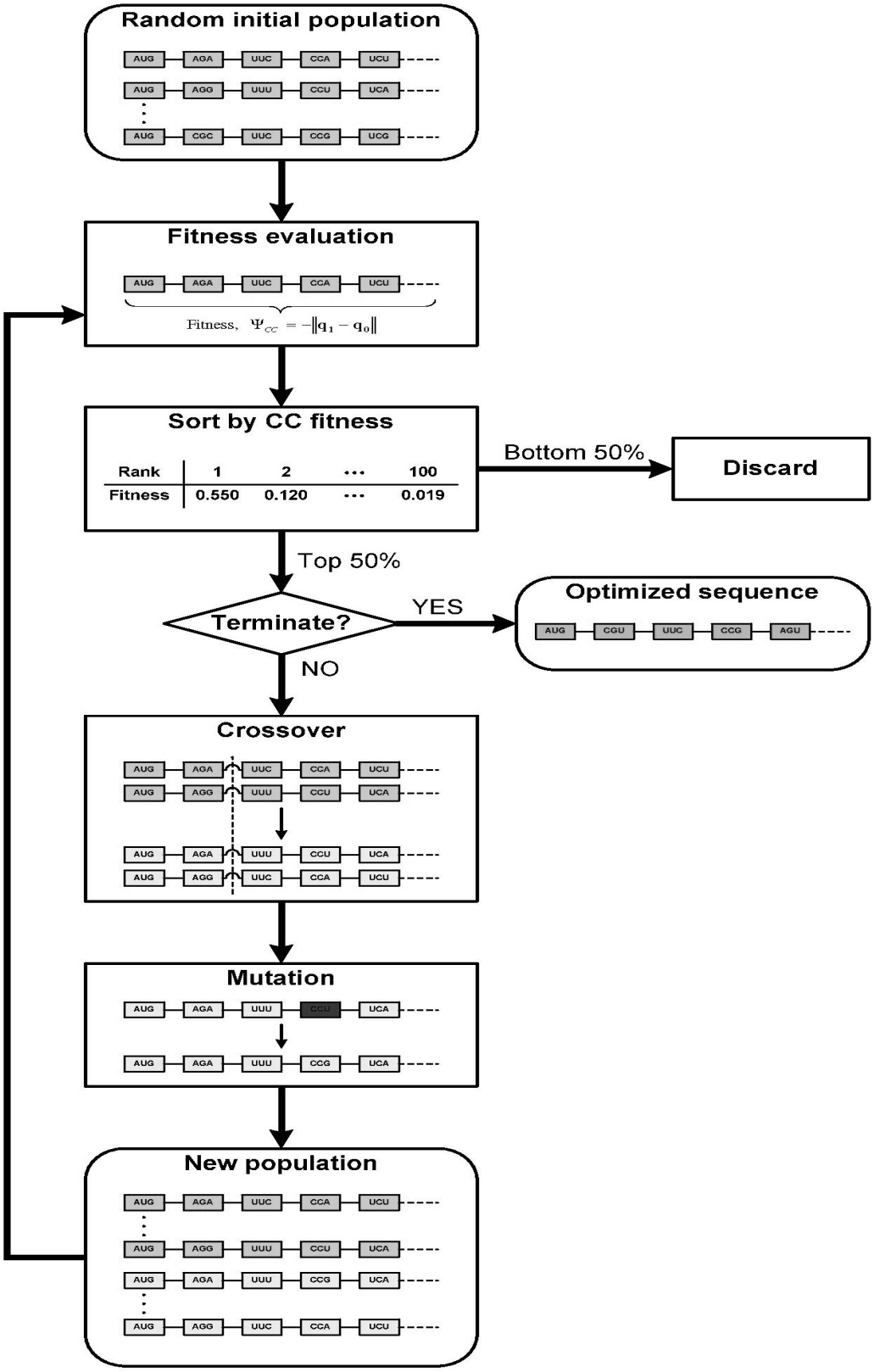

Random initial population
AUG AGA UUC CCA UCU
AUG AGG UUU CCU UCA
AUG CGC UUC CCG UCG
Fitness evaluation
AUG AGA UUC CCA UCU
Fitness, $\Psi_{CC} = -\|\mathbf{q}_1 - \mathbf{q}_0\|$
Sort by CC fitness
Rank 1 2 ... 100
Fitness 0.550 0.120 ... 0.019
Bottom 50%
Discard
Top 50%
Terminate?
YES
Optimized sequence
AUG CGU UUC CCG AGU
NO
Crossover
Mutation
New population

Figure 6A

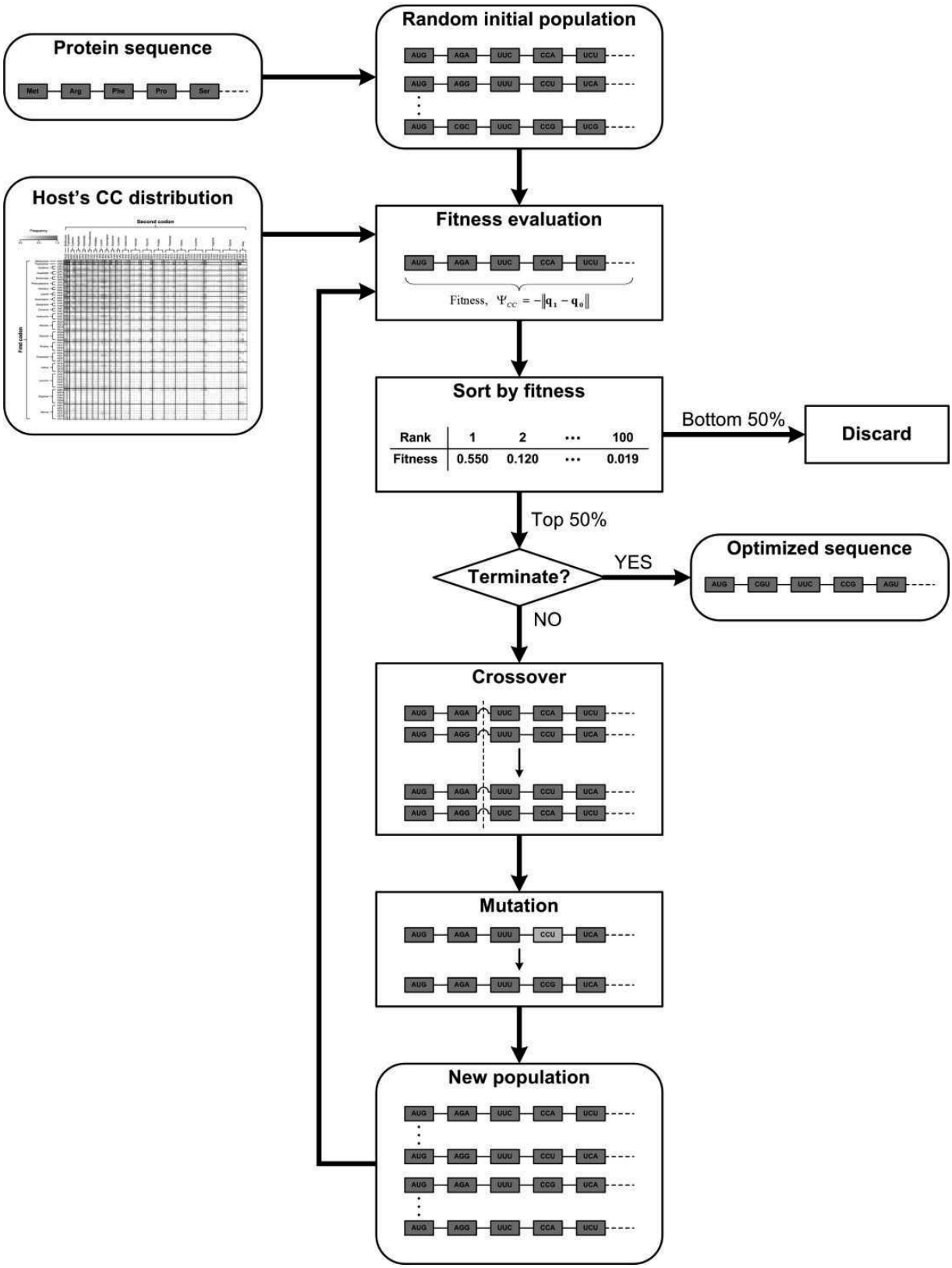

Protein sequence
Met
Arg
Phe
Pro
Ser
Random initial population
Host's CC distribution
Fitness evaluation
Fitness, $\Psi_{CC} = -\|\mathbf{q}_1 - \mathbf{q}_0\|$
Sort by fitness
Rank 1 2 ··· 100
Fitness 0.550 0.120 ··· 0.019
Bottom 50%
Discard
Top 50%
Terminate?
YES
Optimized sequence
NO
Crossover
Mutation
New population

Figure 6B

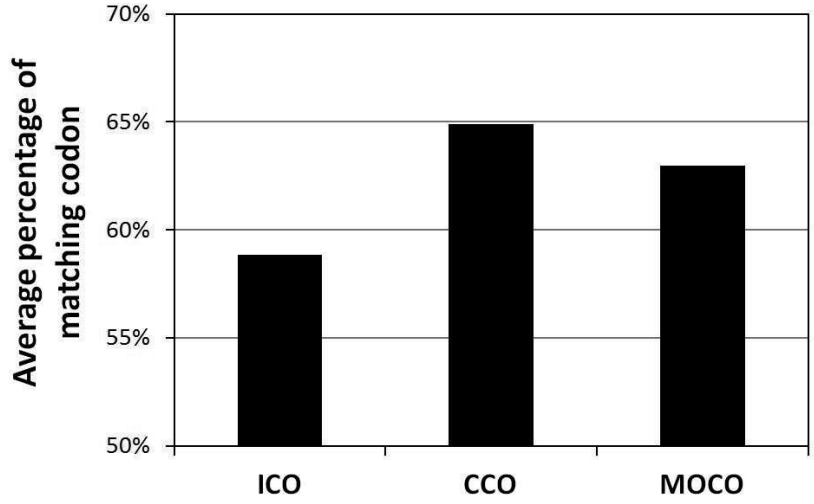

70%
65%
60%
55%
50%
Average percentage of matching codon
ICO
CCO
MOCO

Figure 7

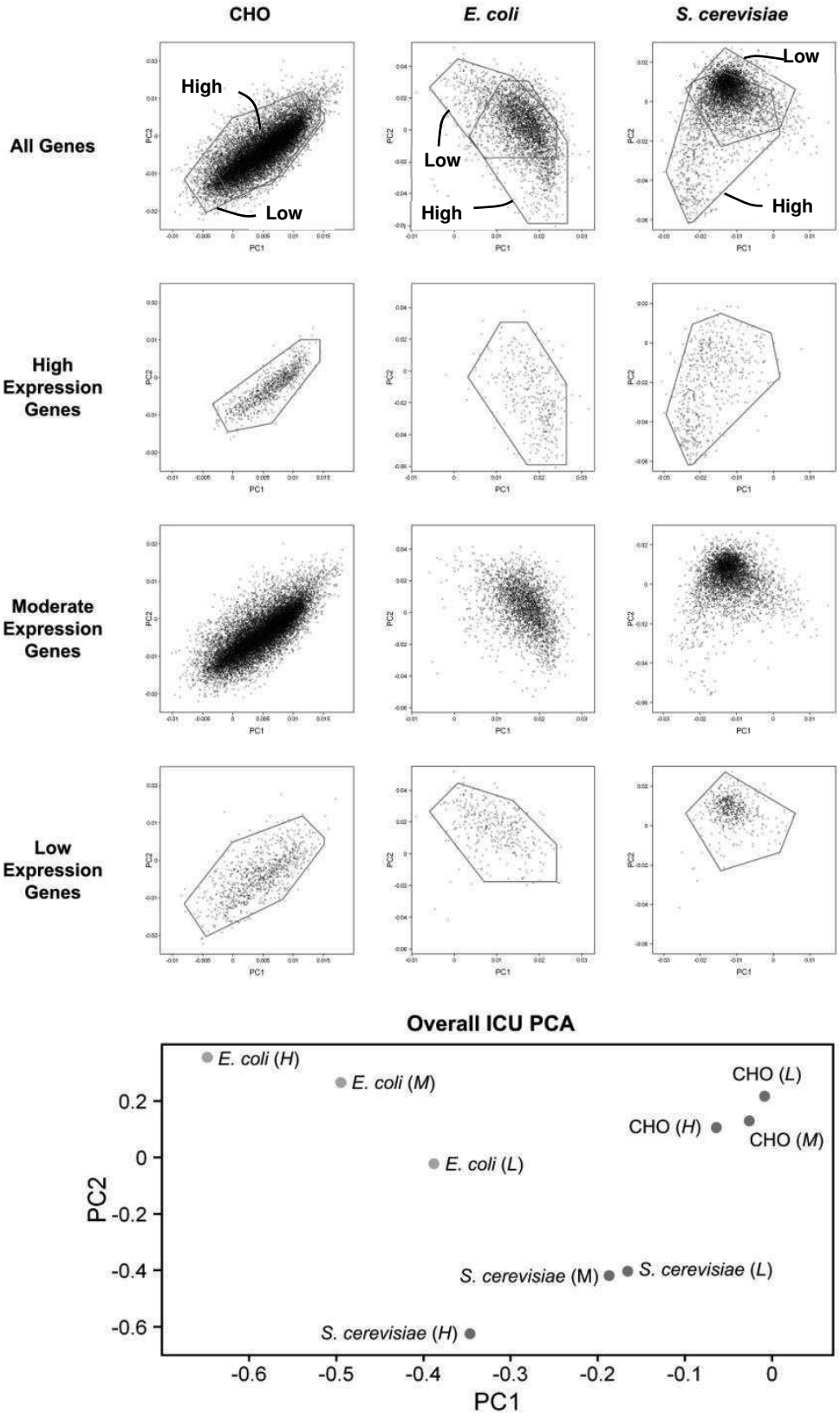

CHO
E. coli
S. cerevisiae
All Genes
High Expression Genes
Moderate Expression Genes
Low Expression Genes
High
Low
PC1
PC2
Overall ICU PCA
E. coli (H)
E. coli (M)
E. coli (L)
CHO (L)
CHO (H)
CHO (M)
S. cerevisiae (M)
S. cerevisiae (L)
S. cerevisiae (H)
0.2
0
-0.2
-0.4
-0.6
-0.6
-0.5
-0.4
-0.3
-0.2
-0.1
0

FIG. 8

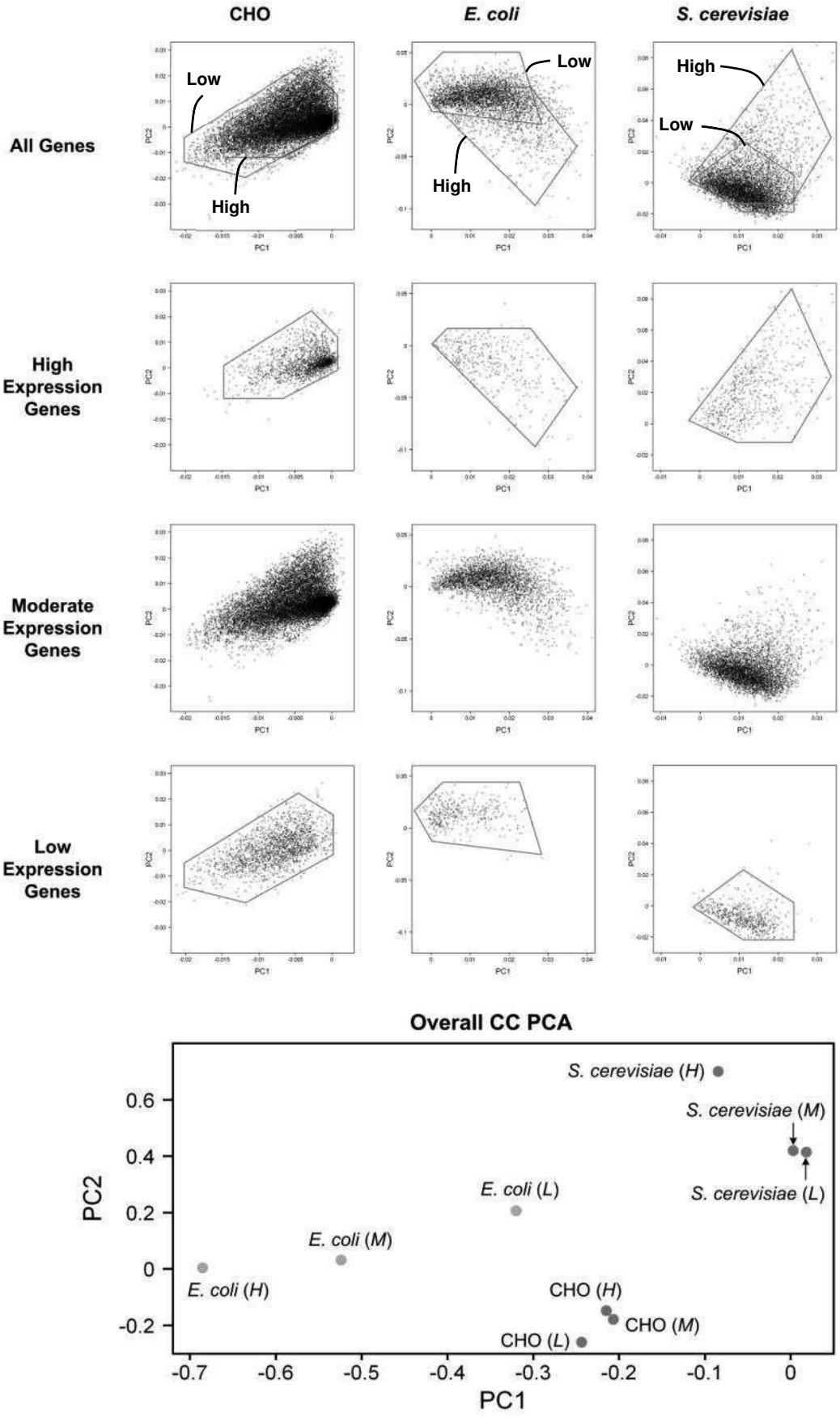

CHO
E. coli
S. cerevisiae
All Genes
High Expression Genes
Moderate Expression Genes
Low Expression Genes
Low
High
PC1
PC2
Overall CC PCA
S. cerevisiae (H)
S. cerevisiae (M)
S. cerevisiae (L)
E. coli (L)
E. coli (M)
E. coli (H)
CHO (H)
CHO (M)
CHO (L)
0.6
0.4
0.2
0
-0.2
-0.7
-0.6
-0.5
-0.4
-0.3
-0.2
-0.1
0

FIG. 9

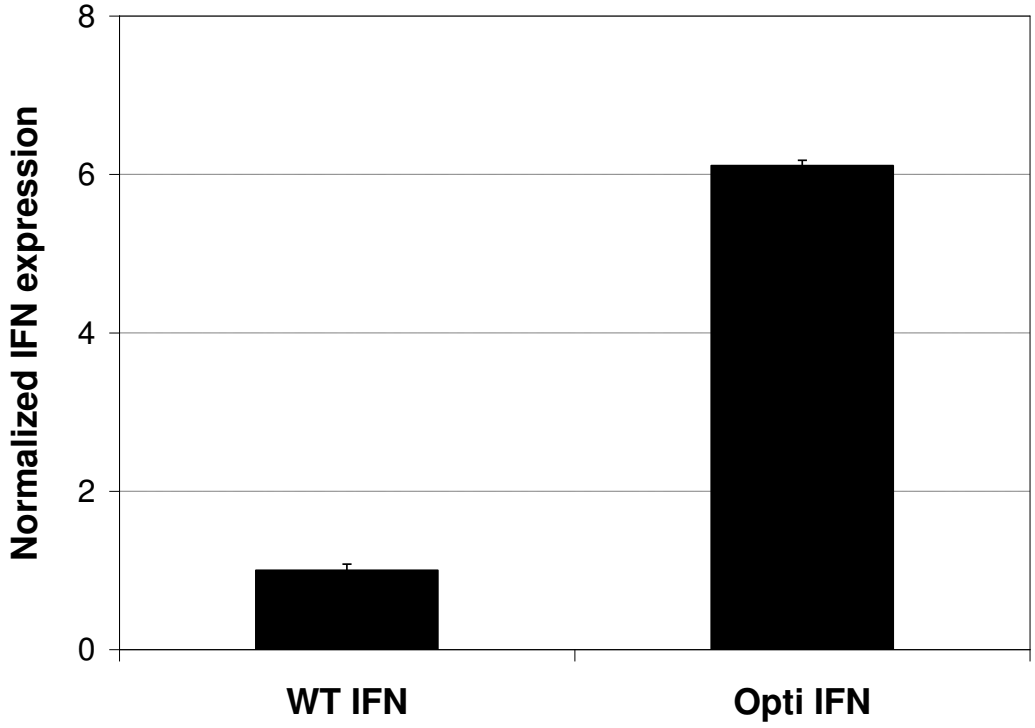
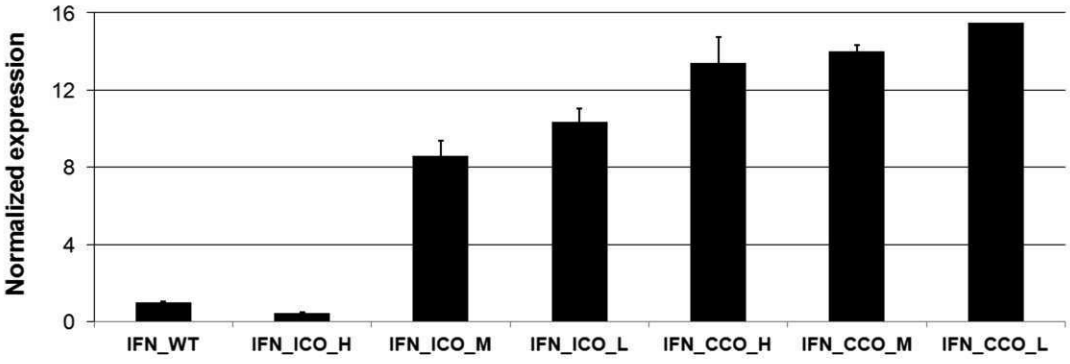


Figure 10


Normalized expression
16
12
8
4
0
IFN_WT
IFN_ICO_H
IFN_ICO_M
IFN_ICO_L
IFN_CCO_H
IFN_CCO_M
IFN_CCO_L

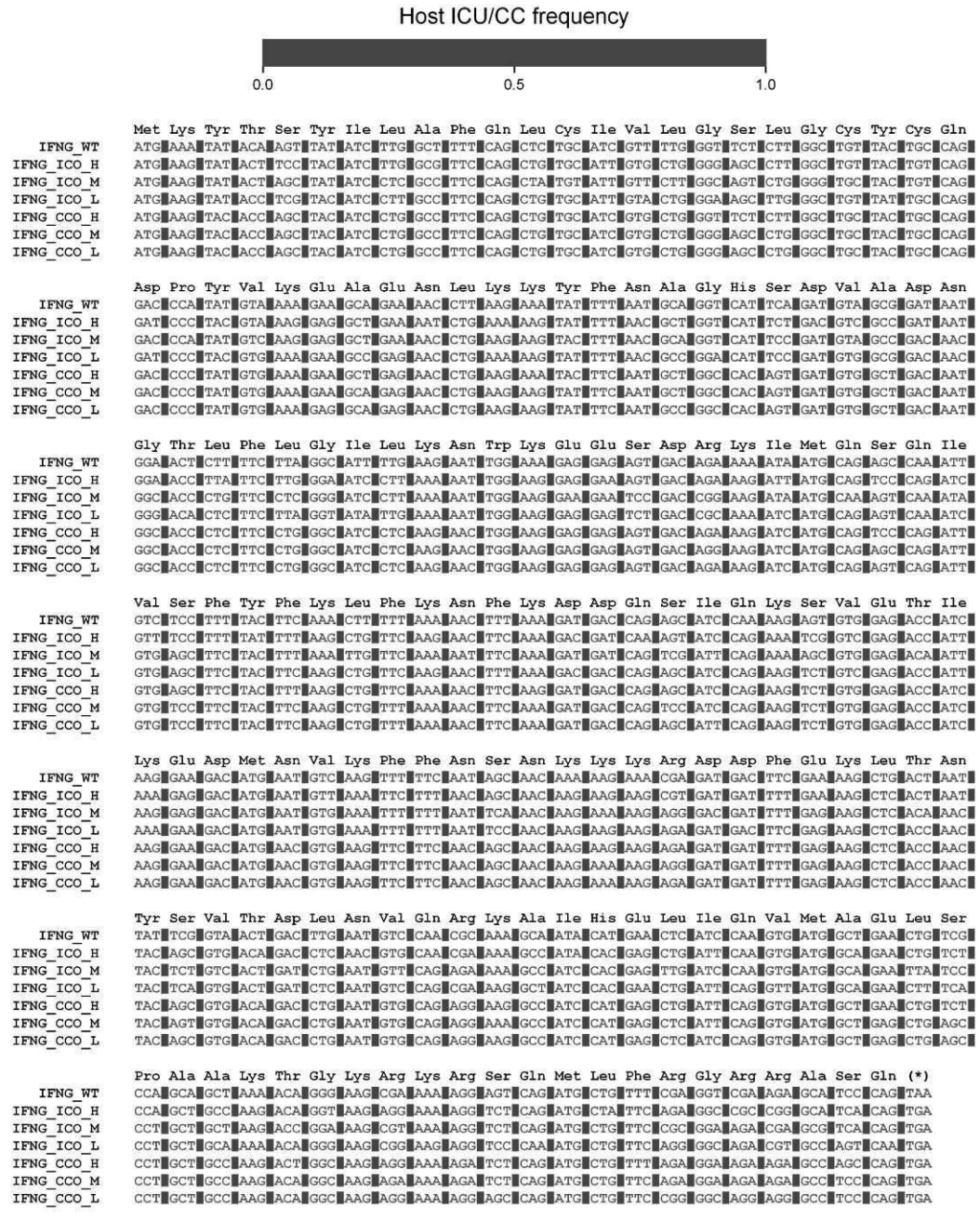
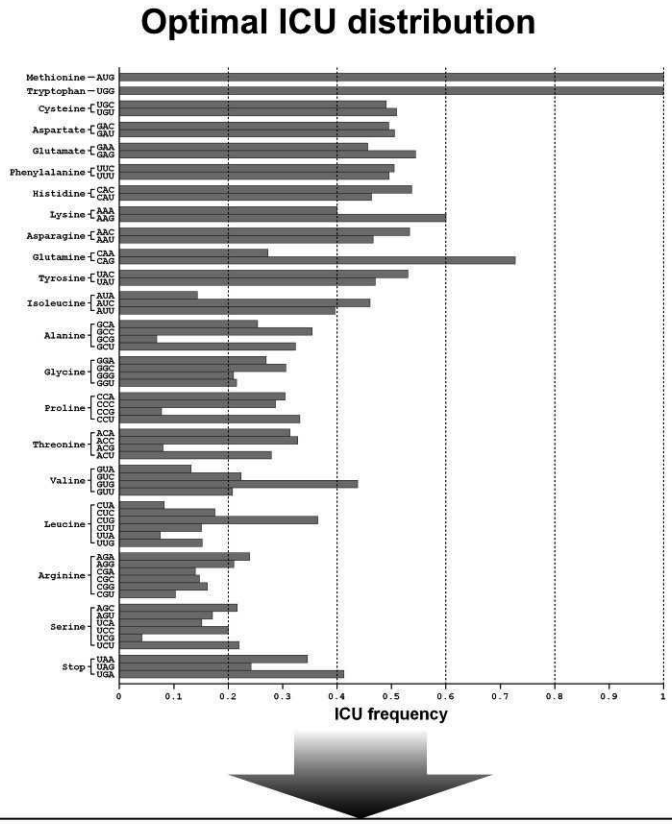

Host ICU/CC frequency
0.0
0.5
1.0
Met Lys Tyr Thr Ser Tyr Ile Leu Ala Phe Gln Leu Cys Ile Val Leu Gly Ser Leu Gly Cys Tyr Cys Gln
IFNG_WT ATG AAA TAT ACA AGT TAT ATC TTG GCT TTT CAG CTC TGC ATC GTT TTG GGT TCT CTT GGC TGT TAC TGC CAG
IFNG_ICO_H ATG AAG TAT ACT TCC TAC ATC TTG GCG TTC CAG CTG TGC ATT GTG CTG GGG AGC CTT GGC TGT TAC TGT CAG
IFNG_ICO_M ATG AAG TAT ACT AGC TAT ATC CTC GCC TTC CAG CTA TGT ATT GTT CTT GGC AGT CTG GGG TGC TAC TGT CAG
IFNG_ICO_L ATG AAG TAT ACC TCG TAC ATC CTT GCC TTC CAG CTG TGC ATT GTA CTG GGA AGC TTG GGC TGT TAT TGC CAG
IFNG_CCO_H ATG AAG TAC ACC AGC TAC ATC CTG GCC TTC CAG CTG TGC ATC GTG CTG GGT TCT CTT GGC TGC TAC TGC CAG
IFNG_CCO_M ATG AAG TAC ACC AGC TAC ATC CTG GCC TTC CAG CTG TGC ATC GTG CTG GGG AGC CTG GGC TGC TAC TGC CAG
IFNG_CCO_L ATG AAG TAC ACC AGC TAC ATC CTG GCC TTC CAG CTG TGC ATC GTG CTG GGG AGC CTG GGC TGC TAC TGC CAG
Asp Pro Tyr Val Lys Glu Ala Glu Asn Leu Lys Lys Tyr Phe Asn Ala Gly His Ser Asp Val Ala Asp Asn
IFNG_WT GAC CCA TAT GTA AAA GAA GCA GAA AAC CTT AAG AAA TAT TTT AAT GCA GGT CAT TCA GAT GTA GCG GAT AAT
IFNG_ICO_H GAT CCC TAC GTA AAG GAG GCT GAA AAT CTG AAA AAG TAT TTT AAC GCT GGT CAT TCT GAC GTC GCC GAT AAT
IFNG_ICO_M GAC CCA TAT GTC AAG GAG GCT GAA AAC CTG AAG AAG TAC TTT AAC GCA GGT CAT TCC GAT GTA GCC GAC AAC
IFNG_ICO_L GAT CCC TAC GTG AAA GAA GCC GAG AAC CTG AAA AAG TAT TTT AAC GCC GGA CAT TCC GAT GTG GCG GAC AAC
IFNG_CCO_H GAC CCC TAT GTG AAA GAA GCT GAG AAC CTG AAG AAA TAC TTC AAT GCT GGC CAC AGT GAT GTG GCT GAC AAT
IFNG_CCO_M GAC CCC TAT GTG AAA GAA GCA GAG AAC CTG AAG AAG TAT TTC AAT GCT GGC CAC AGT GAT GTG GCT GAC AAT
IFNG_CCO_L GAC CCC TAT GTG AAA GAG GCA GAG AAC CTG AAG AAG TAT TTC AAT GCC GGC CAC AGT GAT GTG GCT GAC AAT
Gly Thr Leu Phe Leu Gly Ile Leu Lys Asn Trp Lys Glu Glu Ser Asp Arg Lys Ile Met Gln Ser Gln Ile
IFNG_WT GGA ACT CTT TTC TTA GGC ATT TTG AAG AAT TGG AAA GAG GAG AGT GAC AGA AAA ATA ATG CAG AGC CAA ATT
IFNG_ICO_H GGA ACC TTA TTC TTG GGA ATC CTT AAA AAT TGG AAG GAG GAA AGT GAC AGA AAG ATT ATG CAG TCC CAG ATC
IFNG_ICO_M GGC ACC CTG TTC CTC GGG ATC CTT AAA AAT TGG AAG GAA GAA TCC GAC CGG AAG ATA ATG CAA AGT CAA ATA
IFNG_ICO_L GGG ACA CTC TTC TTA GGT ATA TTG AAA AAT TGG AAG GAG GAG TCT GAC CGC AAA ATC ATG CAG AGT CAA ATC
IFNG_CCO_H GGC ACC CTC TTC CTG GGC ATC CTC AAG AAC TGG AAG GAG GAG AGT GAC AGA AAG ATC ATG CAG TCC CAG ATT
IFNG_CCO_M GGC ACC CTC TTC CTG GGC ATC CTC AAG AAC TGG AAG GAG GAG AGT GAC AGG AAG ATC ATG CAG AGC CAG ATT
IFNG_CCO_L GGC ACC CTC TTC CTG GGC ATC CTC AAG AAC TGG AAG GAG GAG AGT GAC AGA AAG ATC ATG CAG AGT CAG ATT
Val Ser Phe Tyr Phe Lys Leu Phe Lys Asn Phe Lys Asp Asp Gln Ser Ile Gln Lys Ser Val Glu Thr Ile
IFNG_WT GTC TCC TTT TAC TTC AAA CTT TTT AAA AAC TTT AAA GAT GAC CAG AGC ATC CAA AAG AGT GTG GAG ACC ATC
IFNG_ICO_H GTT TCC TTT TAT TTT AAG CTG TTC AAG AAC TTC AAA GAC GAT CAA AGT ATC CAG AAA TCG GTC GAG ACC ATT
IFNG_ICO_M GTG AGC TTC TAC TTT AAA TTG TTC AAA AAT TTC AAA GAT GAT CAG TCG ATT CAG AAA AGC GTG GAG ACA ATT
IFNG_ICO_L GTG AGC TTC TAC TTC AAG CTG TTC AAG AAC TTT AAA GAC GAC CAG AGC ATC CAG AAG TCT GTC GAG ACC ATT
IFNG_CCO_H GTG AGC TTC TAC TTT AAG CTG TTC AAA AAC TTC AAG GAT GAC CAG AGC ATC CAG AAG TCT GTG GAG ACC ATC
IFNG_CCO_M GTG TCC TTC TAC TTC AAG CTG TTT AAA AAC TTC AAA GAT GAC CAG TCC ATC CAG AAG TCT GTG GAG ACC ATC
IFNG_CCO_L GTG TCC TTC TAC TTC AAG CTG TTT AAA AAC TTC AAA GAT GAC CAG AGC ATT CAG AAG TCT GTG GAG ACC ATC
Lys Glu Asp Met Asn Val Lys Phe Phe Asn Ser Asn Lys Lys Lys Arg Asp Asp Phe Glu Lys Leu Thr Asn
IFNG_WT AAG GAA GAC ATG AAT GTC AAG TTT TTC AAT AGC AAC AAA AAG AAA CGA GAT GAC TTC GAA AAG CTG ACT AAT
IFNG_ICO_H AAA GAG GAC ATG AAT GTT AAA TTC TTT AAC AGC AAC AAG AAG AAG CGT GAT GAT TTT GAA AAG CTC ACT AAT
IFNG_ICO_M AAG GAG GAC ATG AAT GTG AAA TTT TTT AAT TCA AAC AAG AAA AAG AGG GAC GAT TTT GAG AAG CTC ACA AAC
IFNG_ICO_L AAA GAA GAC ATG AAT GTG AAA TTT TTT AAT TCC AAC AAG AAG AAG AGA GAT GAC TTC GAG AAG CTC ACC AAC
IFNG_CCO_H AAG GAA GAC ATG AAC GTG AAG TTC TTC AAC AGC AAC AAG AAG AAG AGA GAT GAT TTT GAG AAG CTC ACC AAC
IFNG_CCO_M AAG GAA GAC ATG AAC GTG AAG TTC TTC AAC AGC AAC AAG AAA AAG AGG GAT GAT TTT GAG AAG CTC ACC AAC
IFNG_CCO_L AAG GAA GAC ATG AAC GTG AAG TTC TTC AAC AGC AAC AAG AAA AAG AGA GAT GAT TTT GAG AAG CTC ACC AAC
Tyr Ser Val Thr Asp Leu Asn Val Gln Arg Lys Ala Ile His Glu Leu Ile Gln Val Met Ala Glu Leu Ser
IFNG_WT TAT TCG GTA ACT GAC TTG AAT GTC CAA CGC AAA GCA ATA CAT GAA CTC ATC CAA GTG ATG GCT GAA CTG TCG
IFNG_ICO_H TAC AGC GTG ACA GAC CTC AAC GTG CAA CGA AAA GCC ATA CAC GAG CTG ATT CAA GTG ATG GCA GAA CTG TCT
IFNG_ICO_M TAC TCT GTC ACT GAT CTG AAT GTT CAG AGA AAA GCC ATC CAC GAG TTG ATC CAA GTG ATG GCA GAA TTA TCC
IFNG_ICO_L TAC TCA GTG ACT GAT CTC AAT GTC CAG CGA AAG GCT ATC CAC GAA CTG ATT CAG GTT ATG GCA GAA CTT TCA
IFNG_CCO_H TAC AGC GTG ACA GAC CTG AAT GTG CAG AGG AAG GCC ATC CAT GAG CTG ATT CAG GTG ATG GCT GAA CTG TCT
IFNG_CCO_M TAC AGT GTG ACA GAC CTG AAT GTG CAG AGG AAA GCC ATC CAT GAG CTC ATT CAG GTG ATG GCT GAG CTG AGC
IFNG_CCO_L TAC AGC GTG ACA GAC CTG AAT GTG CAG AGG AAG GCC ATC CAT GAG CTC ATC CAG GTG ATG GCT GAG CTG AGC
Pro Ala Ala Lys Thr Gly Lys Arg Lys Arg Ser Gln Met Leu Phe Arg Gly Arg Arg Ala Ser Gln (*)
IFNG_WT CCA GCA GCT AAA ACA GGG AAG CGA AAA AGG AGT CAG ATG CTG TTT CGA GGT CGA AGA GCA TCC CAG TAA
IFNG_ICO_H CCA GCT GCC AAG ACA GGT AAG AGG AAA AGG TCT CAG ATG CTA TTC AGA GGC CGC CGG GCA TCA CAG TGA
IFNG_ICO_M CCT GCT GCT AAG ACC GGA AAG CGT AAA AGG TCT CAG ATG CTG TTC CGC GGA AGA CGA GCG TCA CAG TGA
IFNG_ICO_L CCT GCT GCA AAA ACA GGG AAG CGG AAG AGG TCC CAA ATG CTG TTC AGG GGC AGA CGT GCC AGT CAA TGA
IFNG_CCO_H CCT GCT GCC AAG ACT GGC AAG AGG AAA AGA TCT CAG ATG CTG TTT AGA GGA AGA AGA GCC AGC CAG TGA
IFNG_CCO_M CCT GCT GCC AAG ACA GGC AAG AGA AAA AGA TCT CAG ATG CTG TTC AGA GGA AGA AGA GCC TCC CAG TGA
IFNG_CCO_L CCT GCT GCC AAG ACA GGC AAG AGG AAA AGG AGC CAG ATG CTG TTC CGG GGC AGG AGG GCC TCC CAG TGA

Figure 11



ICU optimal IFN-γ sequence 1

Amino acid:	Met	Lys	Tyr	Ser	Tyr	...	Gly	Ser	Leu	...
Codon:	AUG	AAG	UAU	UCC	UAC	...	GGG	AGC	CUU	...
Position:	1	2	3	4	5	...	17	18	19	...

ICU optimal IFN-γ sequence 2

Amino acid:	Met	Lys	Tyr	Ser	Tyr	...	Gly	Ser	Leu	...
Codon:	AUG	AAG	UAU	AGC	UAC	...	GGG	UCC	CUU	...
Position:	1	2	3	4	5	...	17	18	19	...

Figure 12

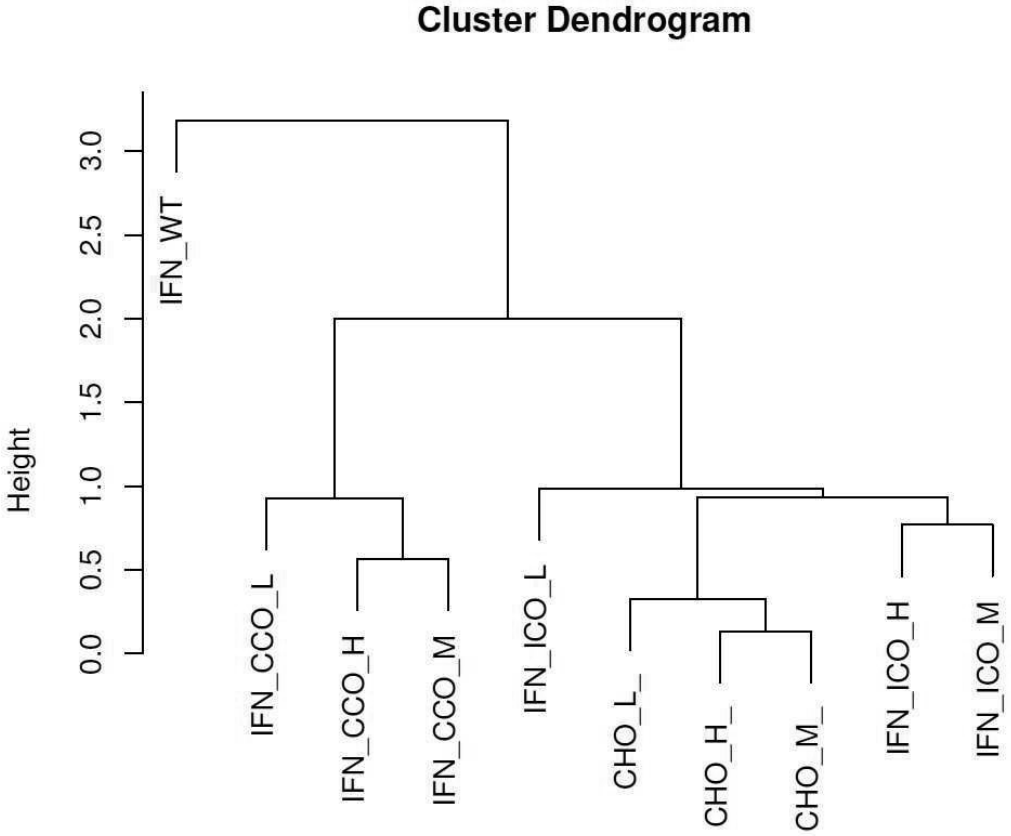

Cluster Dendrogram
Height
0.0
0.5
1.0
1.5
2.0
2.5
3.0
IFN_WT
IFN_CCO_L
IFN_CCO_H
IFN_CCO_M
IFN_ICO_L
CHO_L_
CHO_H_
CHO_M_
IFN_ICO_H
IFN_ICO_M

Figure 13

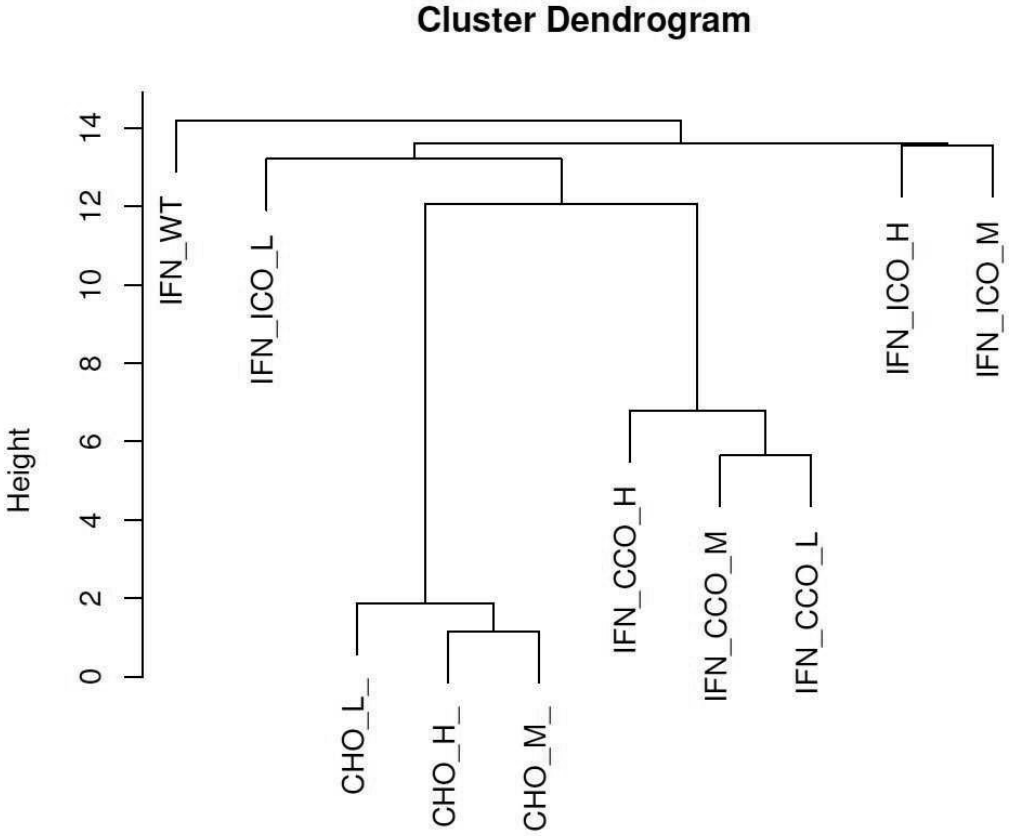

Cluster Dendrogram
Height
0
2
4
6
8
10
12
14
IFN_WT
IFN_ICO_L
IFN_ICO_H
IFN_ICO_M
IFN_CCO_H
IFN_CCO_M
IFN_CCO_L
CHO_L_
CHO_H_
CHO_M_

Figure 14

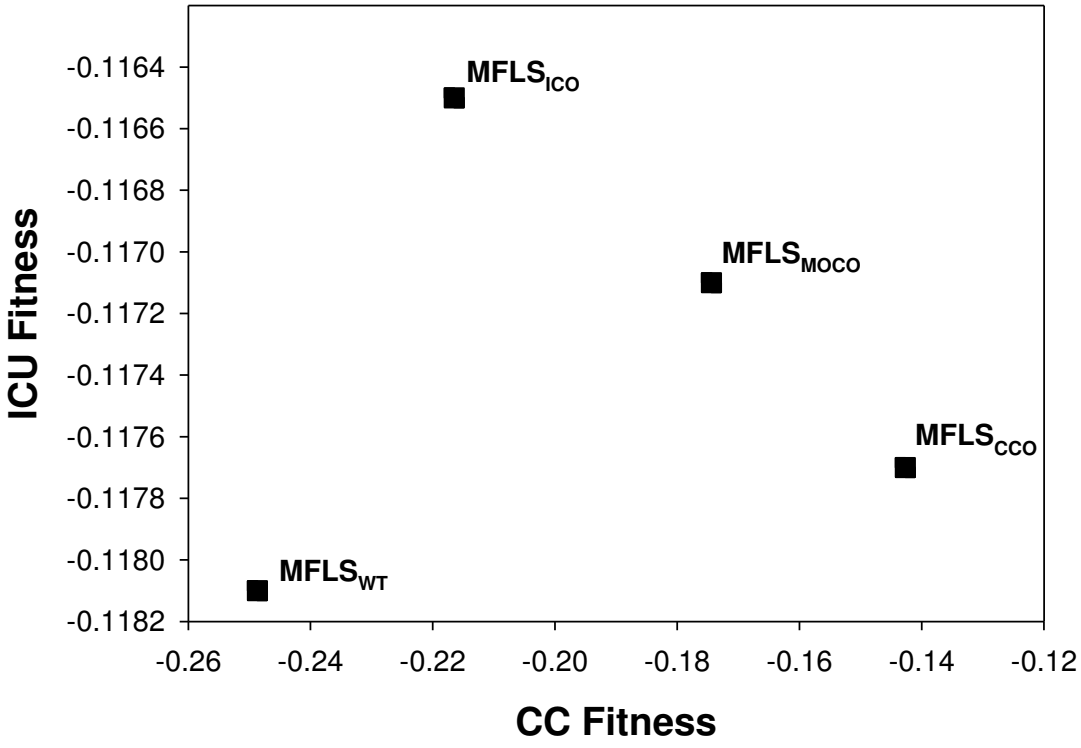

MFLS$_{ICO}$
MFLS$_{MOCO}$
MFLS$_{CCO}$
MFLS$_{WT}$
ICU Fitness
CC Fitness
-0.1164
-0.1166
-0.1168
-0.1170
-0.1172
-0.1174
-0.1176
-0.1178
-0.1180
-0.1182
-0.26
-0.24
-0.22
-0.20
-0.18
-0.16
-0.14
-0.12

Figure 15

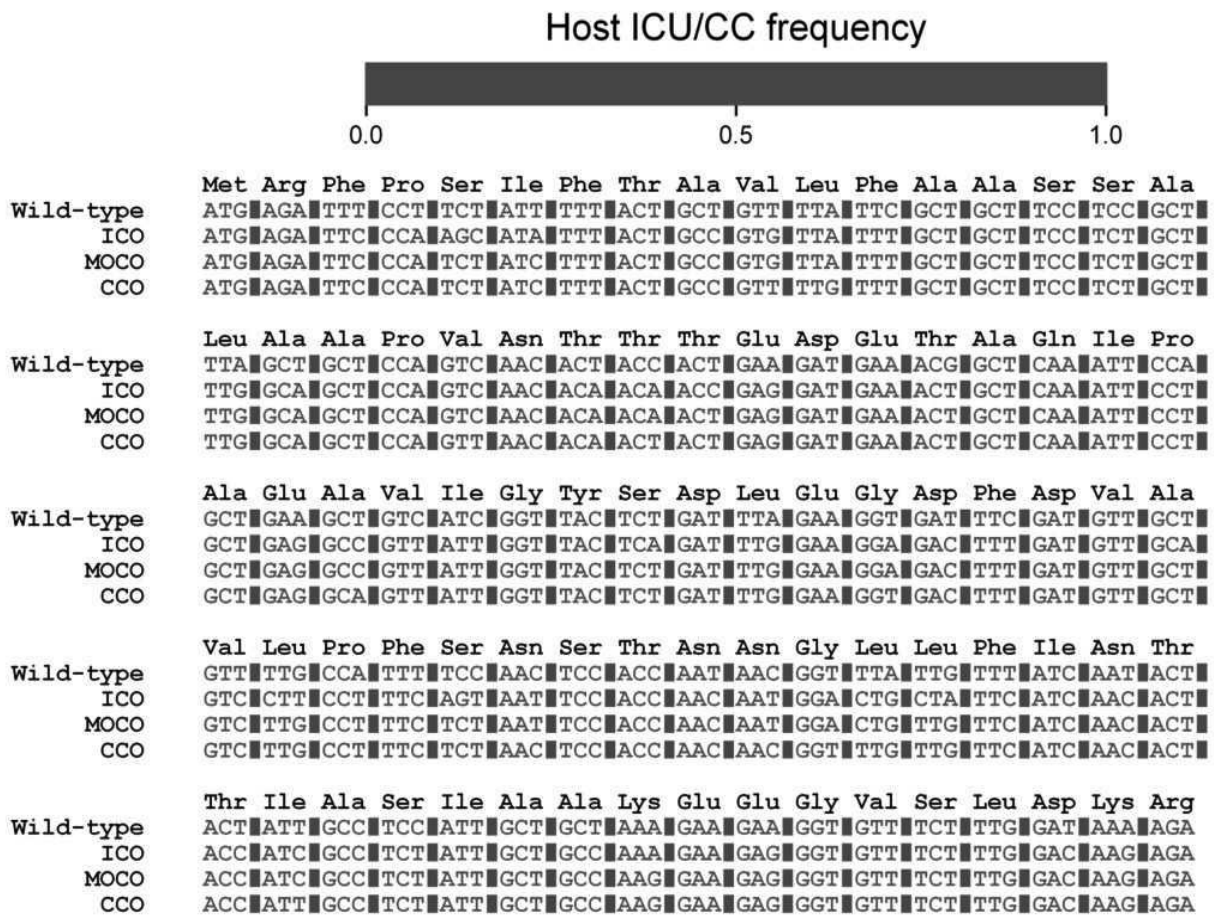

Host ICU/CC frequency
0.0
0.5
1.0
Met Arg Phe Pro Ser Ile Phe Thr Ala Val Leu Phe Ala Ala Ser Ser Ala
Wild-type ATG AGA TTT CCT TCT ATT TTT ACT GCT GTT TTA TTC GCT GCT TCC TCC GCT
ICO ATG AGA TTC CCA AGC ATA TTT ACT GCC GTG TTA TTT GCT GCT TCC TCT GCT
MOCO ATG AGA TTC CCA TCT ATC TTT ACT GCC GTG TTA TTT GCT GCT TCC TCT GCT
CCO ATG AGA TTC CCA TCT ATC TTT ACT GCC GTT TTG TTT GCT GCT TCC TCT GCT
Leu Ala Ala Pro Val Asn Thr Thr Thr Glu Asp Glu Thr Ala Gln Ile Pro
Wild-type TTA GCT GCT CCA GTC AAC ACT ACC ACT GAA GAT GAA ACG GCT CAA ATT CCA
ICO TTG GCA GCT CCA GTC AAC ACA ACA ACC GAG GAT GAA ACT GCT CAA ATT CCT
MOCO TTG GCA GCT CCA GTC AAC ACA ACA ACT GAG GAT GAA ACT GCT CAA ATT CCT
CCO TTG GCA GCT CCA GTT AAC ACA ACT ACT GAG GAT GAA ACT GCT CAA ATT CCT
Ala Glu Ala Val Ile Gly Tyr Ser Asp Leu Glu Gly Asp Phe Asp Val Ala
Wild-type GCT GAA GCT GTC ATC GGT TAC TCT GAT TTA GAA GGT GAT TTC GAT GTT GCT
ICO GCT GAG GCC GTT ATT GGT TAC TCA GAT TTG GAA GGA GAC TTT GAT GTT GCA
MOCO GCT GAG GCC GTT ATT GGT TAC TCT GAT TTG GAA GGA GAC TTT GAT GTT GCT
CCO GCT GAG GCA GTT ATT GGT TAC TCT GAT TTG GAA GGT GAC TTT GAT GTT GCT
Val Leu Pro Phe Ser Asn Ser Thr Asn Asn Gly Leu Leu Phe Ile Asn Thr
Wild-type GTT TTG CCA TTT TCC AAC TCC ACC AAT AAC GGT TTA TTG TTT ATC AAT ACT
ICO GTC CTT CCT TTC AGT AAT TCC ACC AAC AAT GGA CTG CTA TTC ATC AAC ACT
MOCO GTC TTG CCT TTC TCT AAT TCC ACC AAC AAT GGA CTG TTG TTC ATC AAC ACT
CCO GTC TTG CCT TTC TCT AAC TCC ACC AAC AAC GGT TTG TTG TTC ATC AAC ACT
Thr Ile Ala Ser Ile Ala Ala Lys Glu Glu Gly Val Ser Leu Asp Lys Arg
Wild-type ACT ATT GCC TCC ATT GCT GCT AAA GAA GAA GGT GTT TCT TTG GAT AAA AGA
ICO ACC ATC GCC TCT ATT GCT GCC AAA GAA GAG GGT GTT TCT TTG GAC AAG AGA
MOCO ACC ATC GCC TCT ATT GCT GCC AAG GAA GAG GGT GTT TCT TTG GAC AAG AGA
CCO ACC ATT GCC TCT ATT GCT GCC AAG GAA GAG GGT GTT TCT TTG GAC AAG AGA

Figure 16

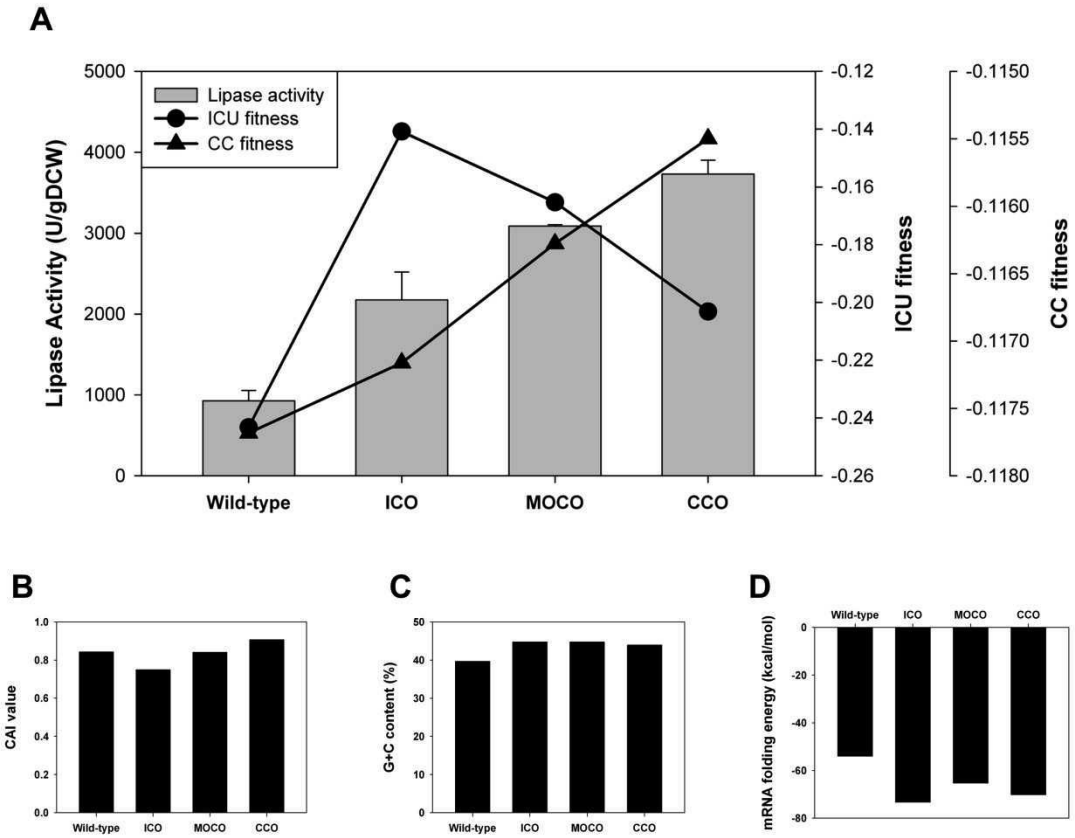

A
Lipase activity
ICU fitness
CC fitness
Lipase Activity (U/gDCW)
ICU fitness
CC fitness
Wild-type
ICO
MOCO
CCO
B
CAI value
C
G+C content (%)
D
mRNA folding energy (kcal/mol)

Figure 17