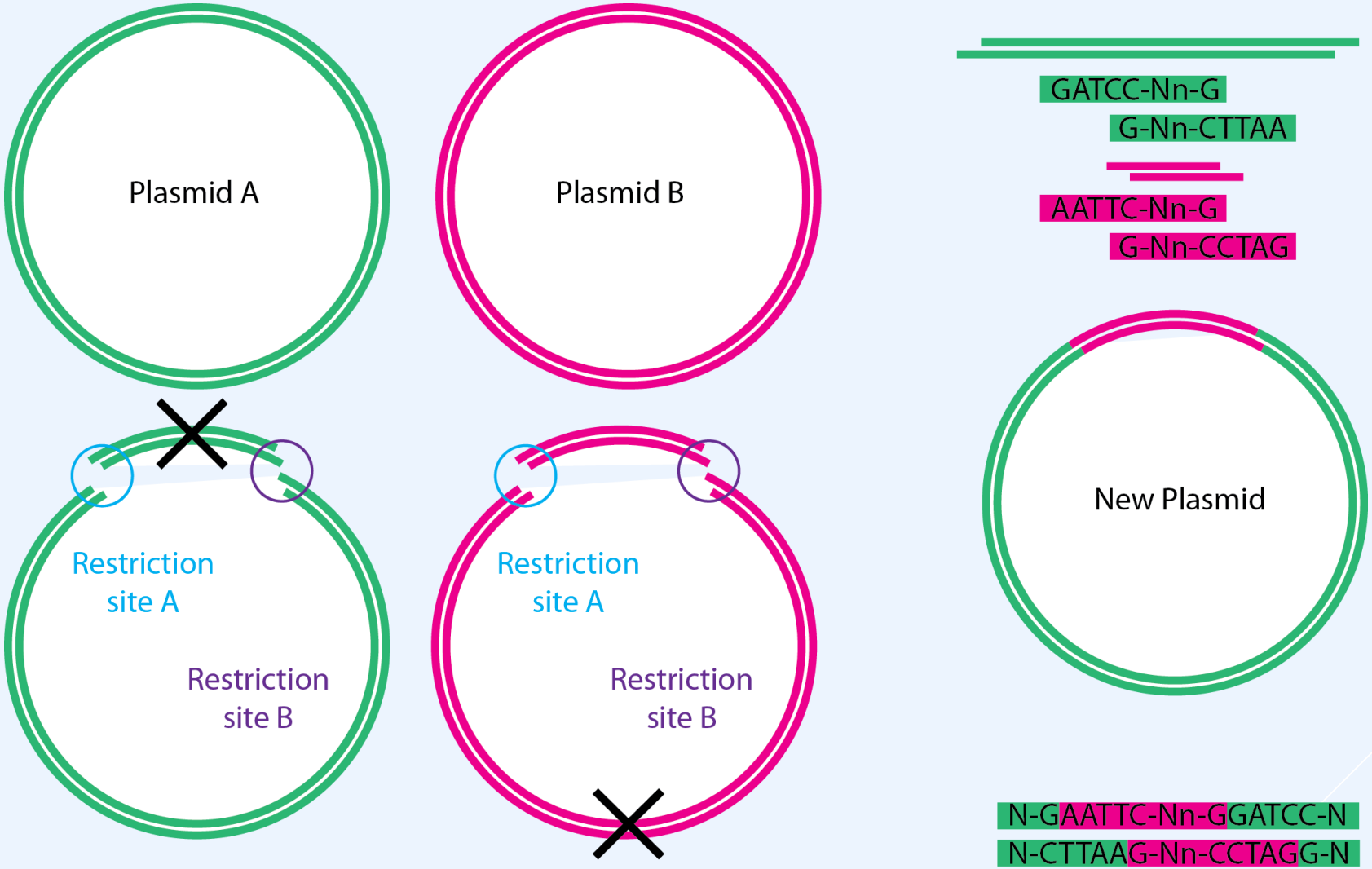
SYNTHETIC BIOLOGY PRACTICAL

Biological Databases and tools for handling DNA sequences

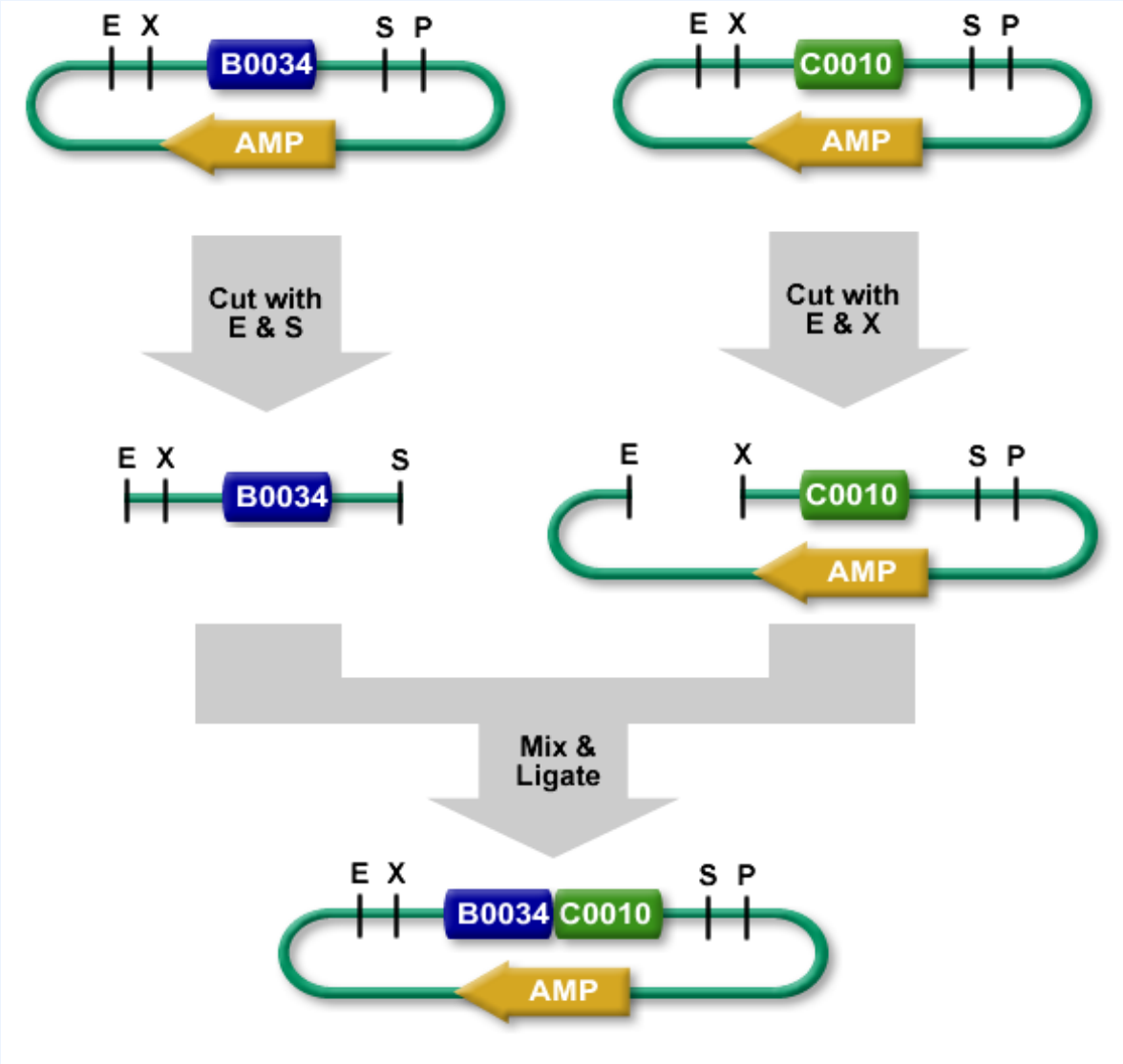
BACKGROUND

- Your boss has asked you to express the gene mecA in E. coli so that you can characterize the protein and told you that you can get a plasmid from a post doc in another lab.
- The post doc gives you a tube with the plasmid and a fasta file with the nucleotide sequence.

CLONING

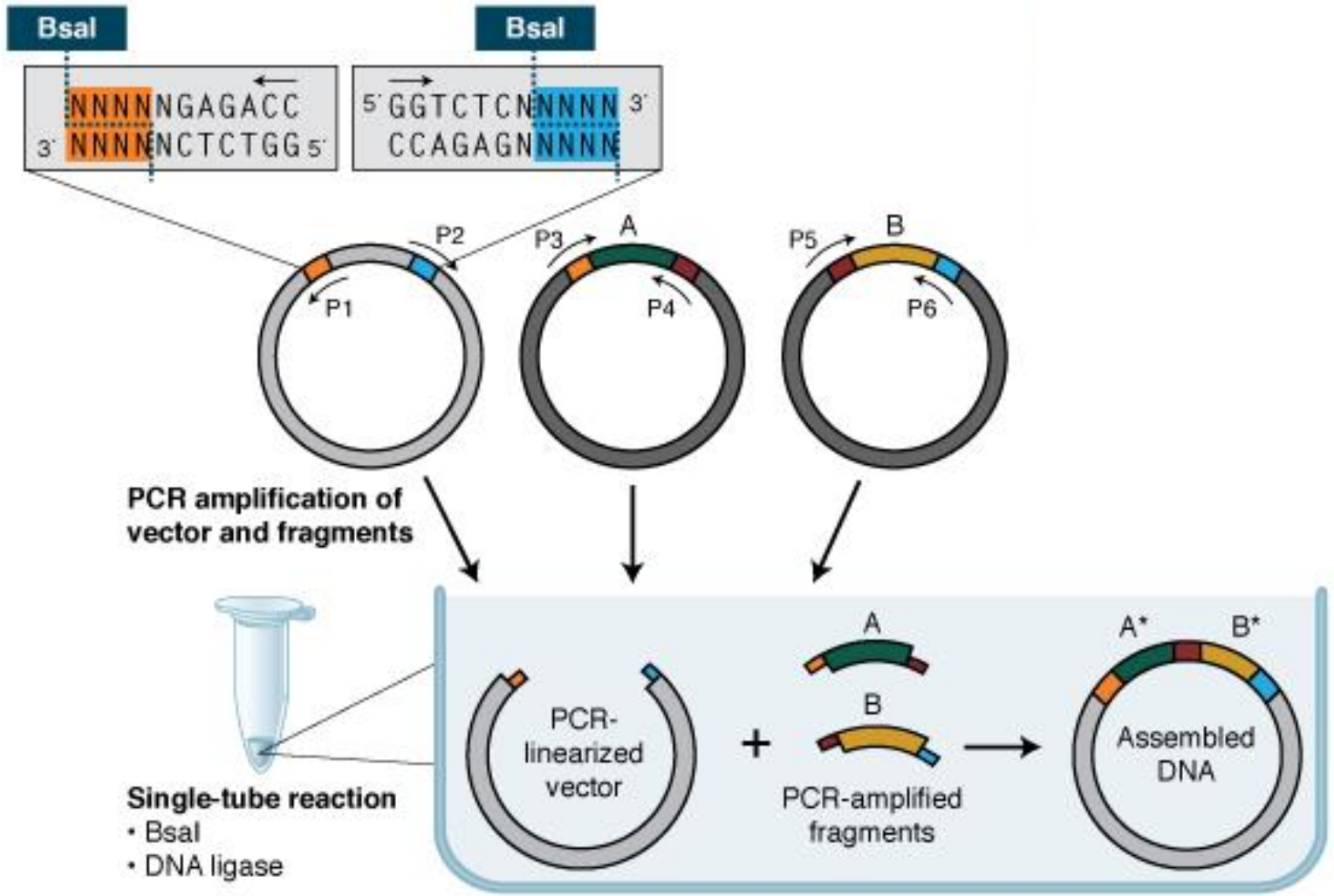

Plasmid A
Plasmid B
Restriction site A
Restriction site B
Restriction site A
Restriction site B
GATCC-Nn-G
G-Nn-CTTAA
AATTC-Nn-G
G-Nn-CCTAG
New Plasmid
N-GAATTC-Nn-GGATCC-N
N-CTTAAG-Nn-CCTAGG-N

BIOBRICK CLONING



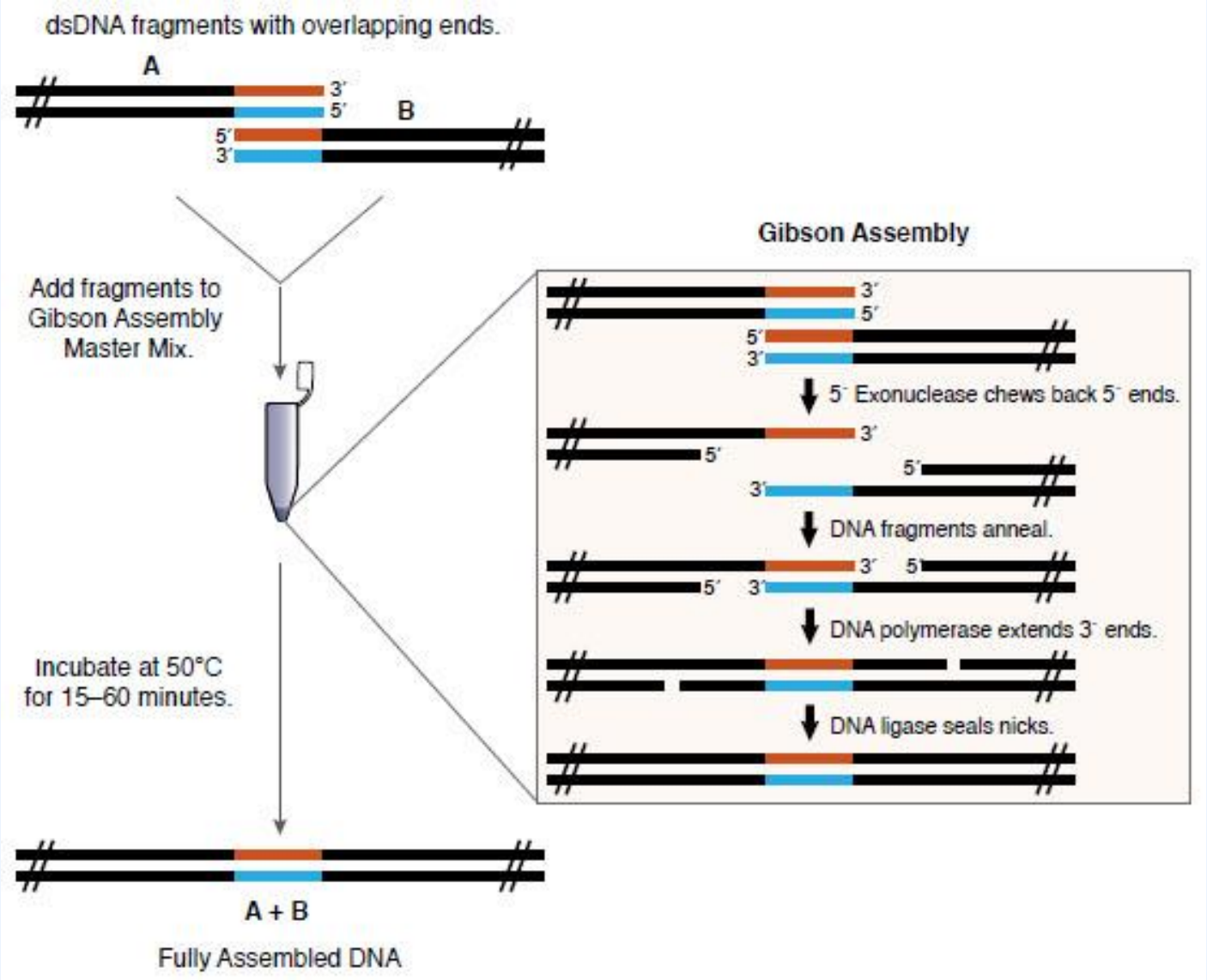
E = EcoRI
X = XbaI
S = SpeI
P = PstI

GOLDEN GATE CLONING



* While A and B insert sequences involved in 4-base overlaps are shown in separate colors for clarity, the actual assembly is seamless; 4-base overlaps are insert derived.

GIBSON ASSEMBLY



OBJECTIVES

- Reverse Engineer a plasmid
  - Identify Promoter, RBS, CDS, Terminator, Restriction sites

- Insert a new gene into plasmid
  - Use Registry of Biological Parts or biocyc to find new DNA sequences
  - Optimize expression with RBS calculator
  - Codon Optimize CDS sequence for expression in desired chassis

CLONING SOFTWARE

- Useful for Annotating DNA sequences, Primer Design, In silico Molecular biology (PCR, Restriction digestion, ligation)

- Free options
  - Serial Cloner, Benchling

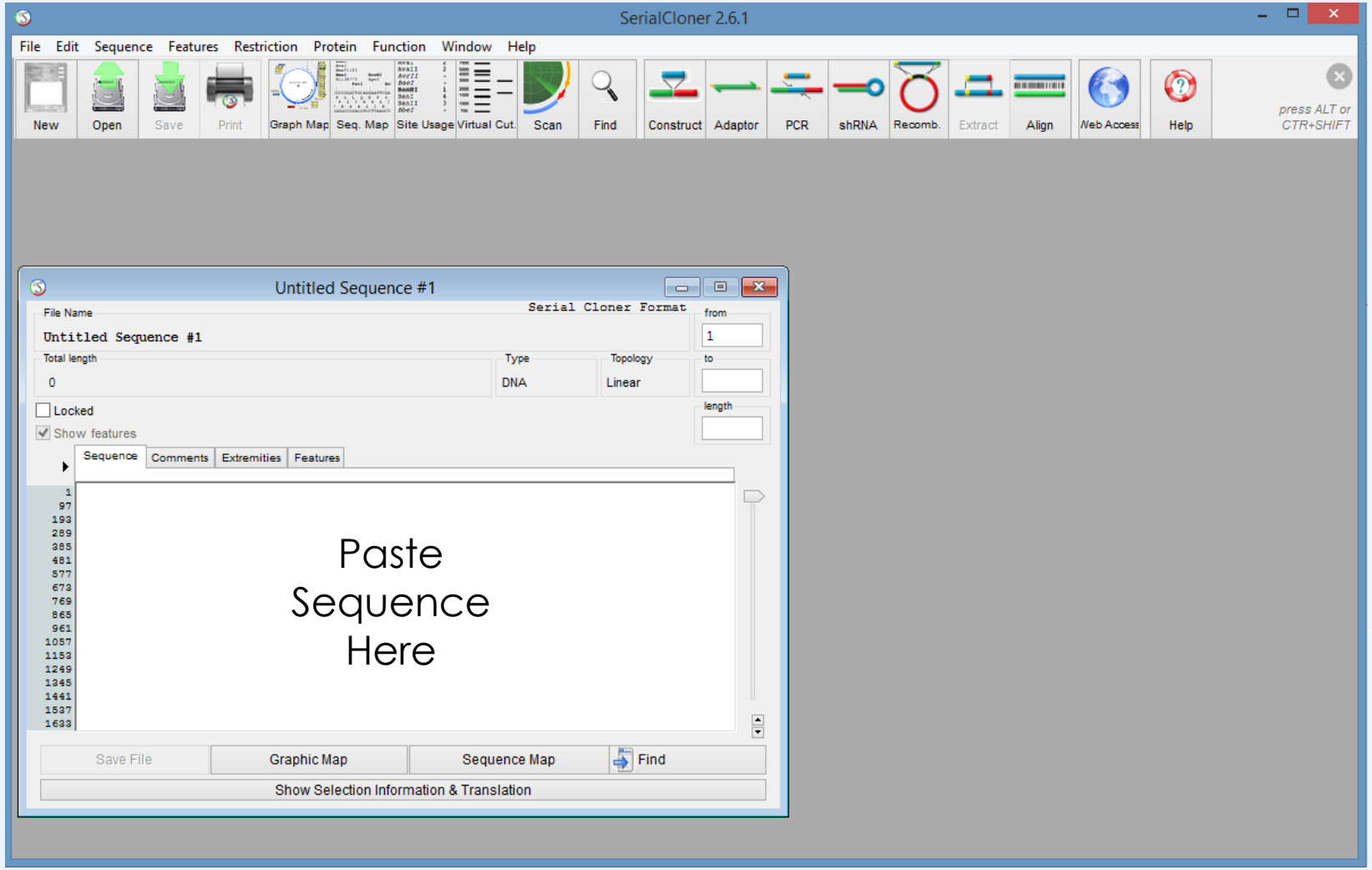
- Paid options
  - Geneious, Clone Manager, Snap Gene

NON-ANNOTATED DNA SEQUENCE

- Download sequence from https://drive.google.com/file/d/0B2ciXonm2r_jWWZld09NMnB5ZVk/view?usp=sharing

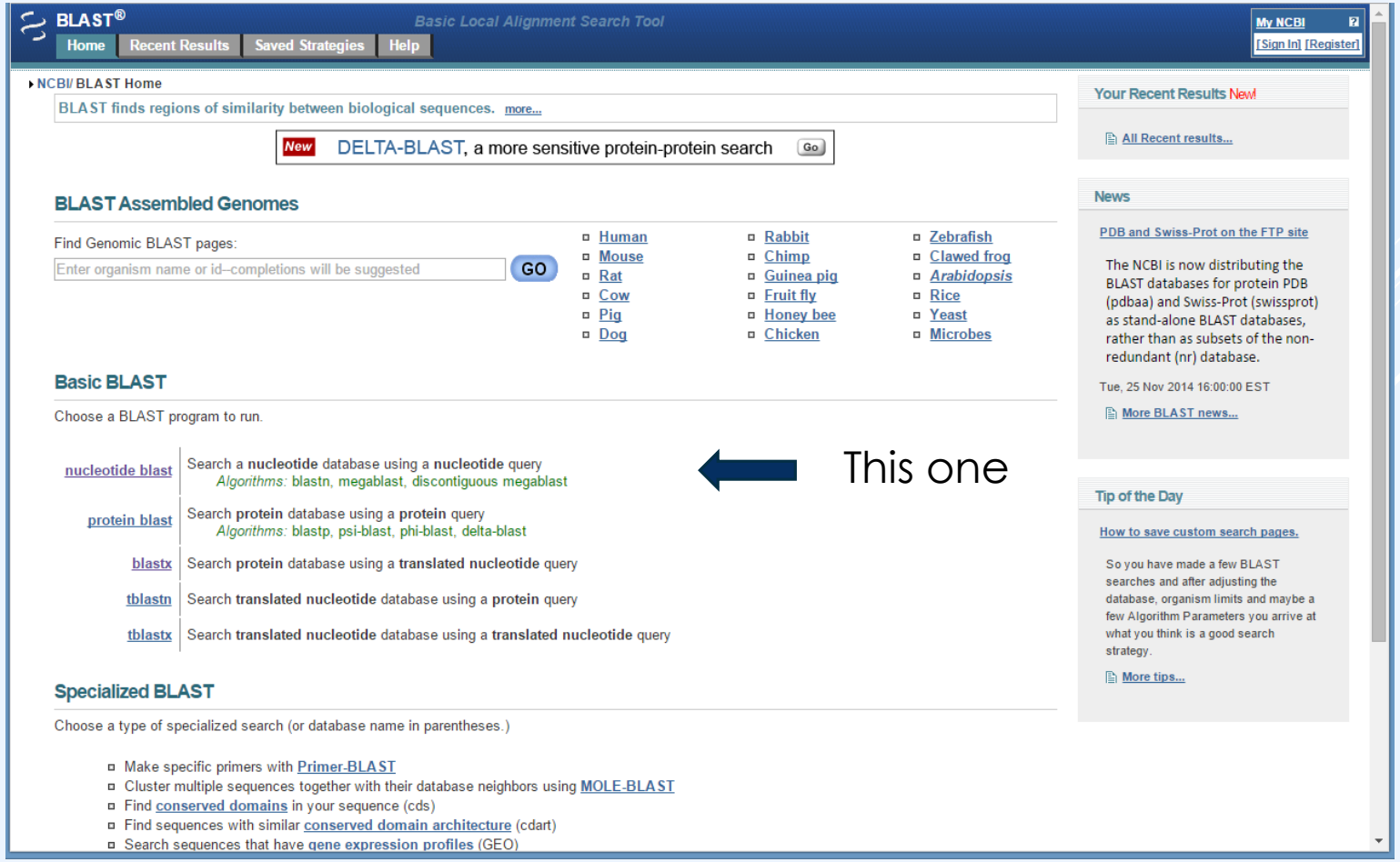
- Load sequence into SerialCloner. Annotate the features of the sequence and determine what it does.

SERIAL CLONER

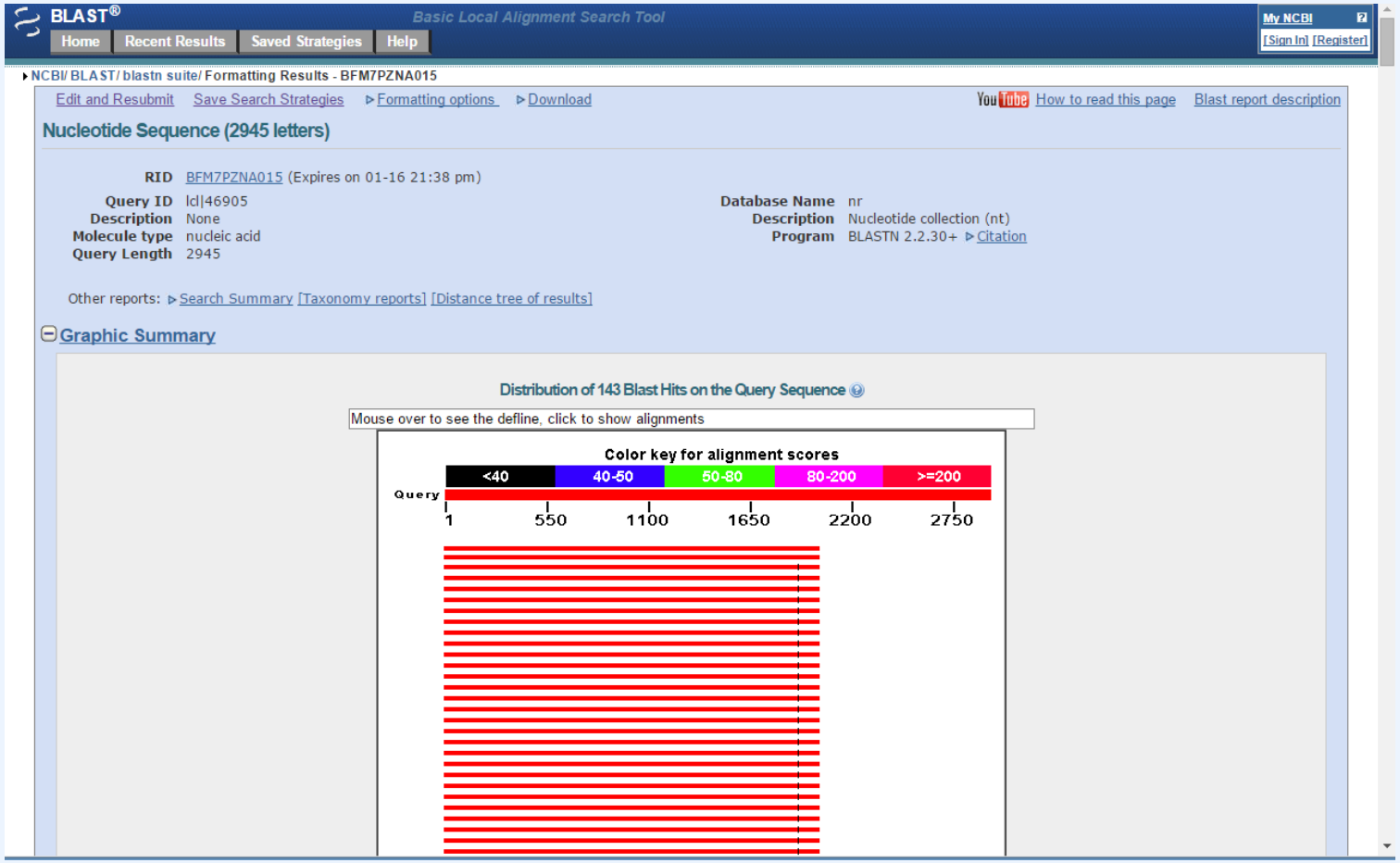


NCBI - BLAST

- First we identify the backbone
  - Go to the NCBI – Blast website (http://blast.ncbi.nlm.nih.gov/Blast.cgi)
  - NCBI Blast will search a Nucleotide or Protein database for matches to a given query sequence, or can align two sequences.
  - Click Nucleotide Blast
  - Paste DNA sequence and hit BLAST



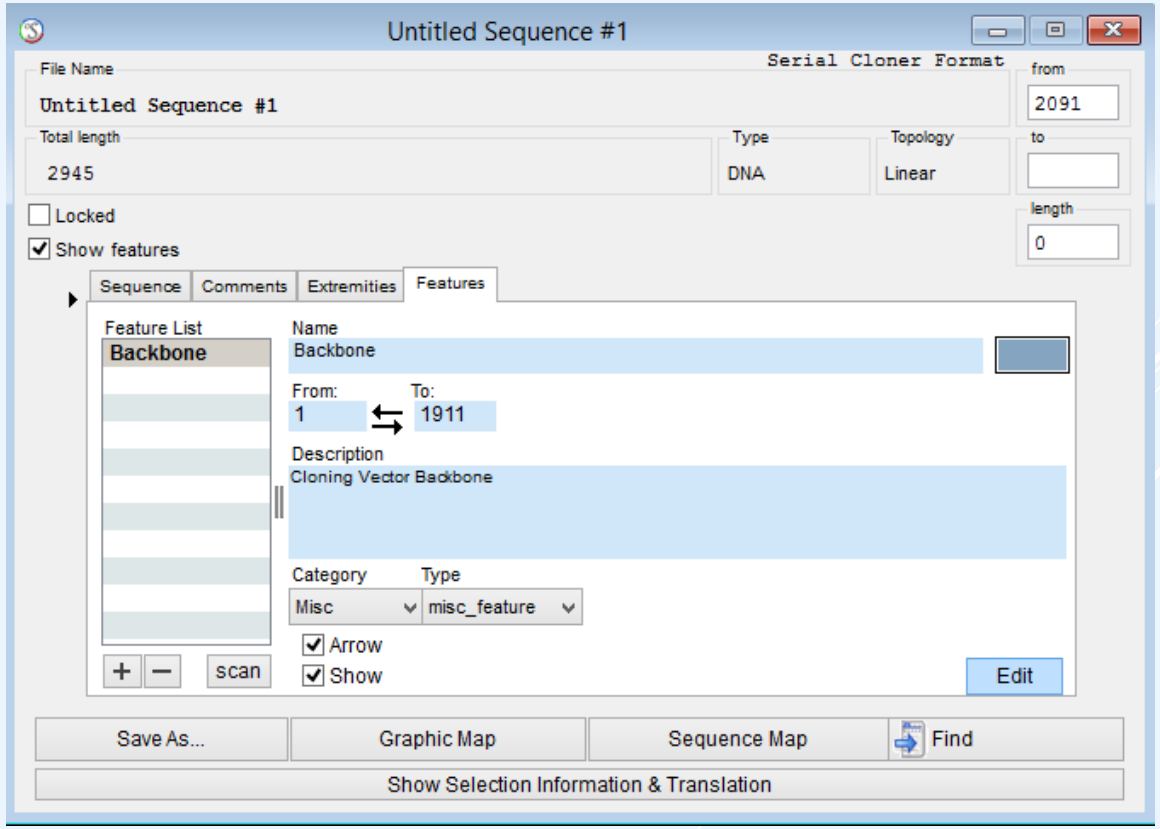
IDENTIFYING THE VECTOR BACKBONE



- The red bars show us regions with high alignment scores for sequences in the nucleotide database. A list with all those sequences and alignments is below. As many of the hits for this region are for different cloning vectors, it is likely this is a common vector back bone.

ANNOTATING THE BACKBONE

- In Serial Cloner
  - Click the Features tab of your DNA sequence
  - Click the + sign below the features list to add a new feature
  - Name the feature and annotate where it covers your DNA sequence using the alignment from NCBI – Blast.

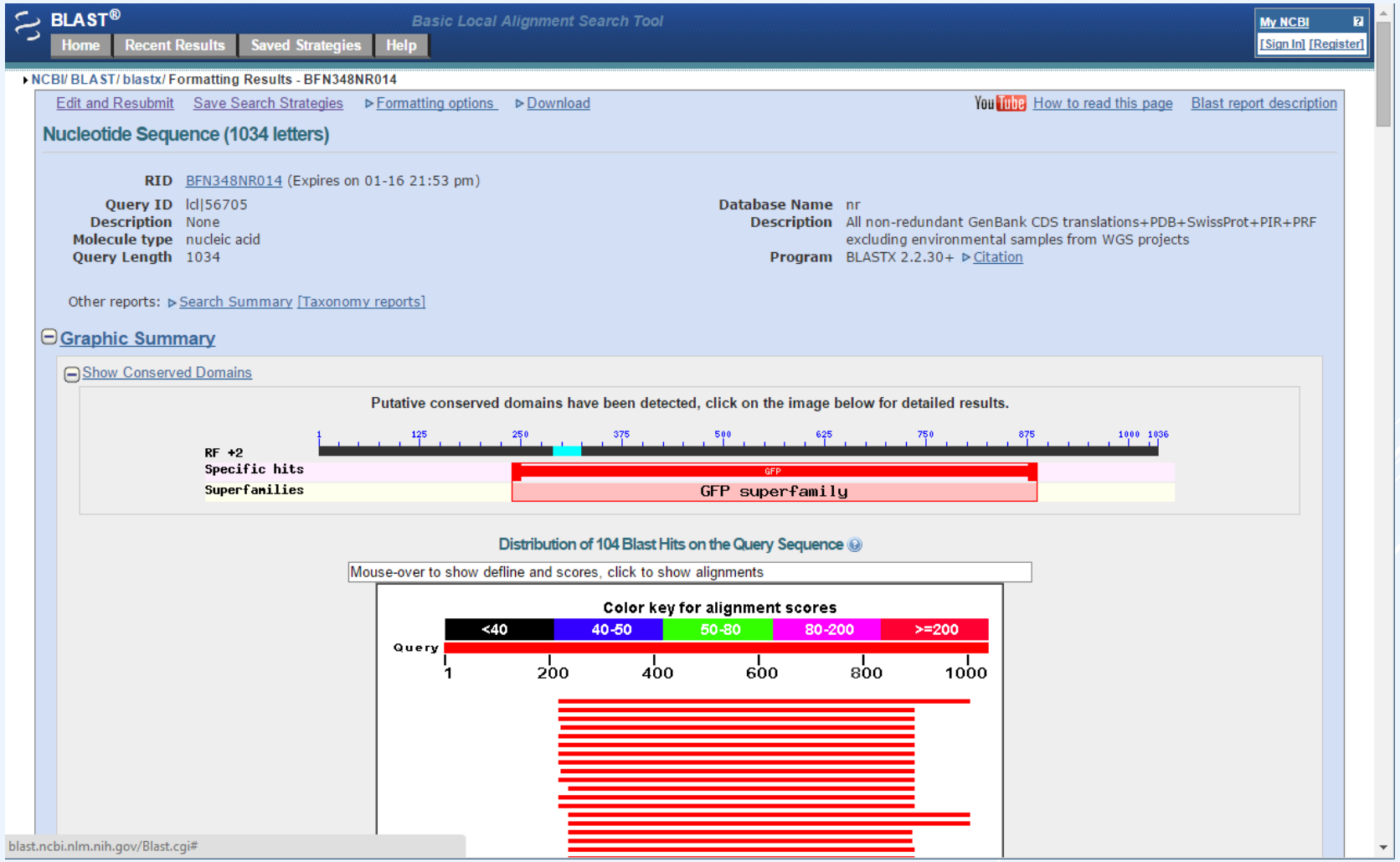


ANNOTATING THE CDS

- Some software will predict ORF within a sequence for you, however you can also find the CDS with NCBI – Blast
- Select the remaining DNA sequence that was **not** part of the backbone
- Go to the NCBI – Blast website and this time select blastx
- BlastX will attempt to translate a nucleotide sequence to match to a protein database. This is useful due to the degeneracy in the genetic code

BLASTX

- Blastx will give alignments similar to blastn, but will also tell you what protein super families it finds in a given sequence
- Blastx tells you that the sequence relates to the GFP family of proteins



BLASTX ALIGNMENTS

- Scroll down to look at the specific alignments. The Blast results show that this is monomeric RFP. Use this alignment to annotate your sequence in Serial Cloner.

Download GenPept Graphics ▼ Next ▲ Previous Descriptions

monomeric red fluorescent protein [synthetic construct]

Sequence ID: gb|AAM54544.1|AF506027_1 **Length:** 225 **Number of Matches:** 1

See 39 more title(s)

Range 1: 1 to 225 GenPept Graphics ▼ Next Match ▲ Previous Match

Score	Expect	Method	Identities	Positives	Gaps	Frame
436 bits(1122)	4e-151	Compositional matrix adjust.	225/225(100%)	225/225(100%)	0/225(0%)	+2

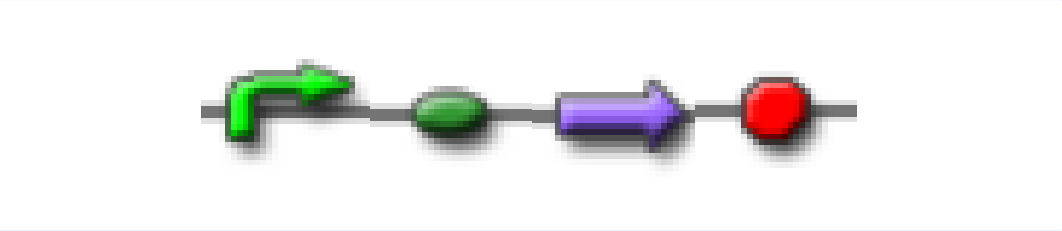
```
Query  221  MASSEDVIKEFMRFKVRMEGSVNghefeiegegegRPYEGTQTAKLKVTKGGPLPFAWDI  400
            MASSEDVIKEFMRFKVRMEGSVNGHEFEIEGEGEGRPYEGTQTAKLKVTKGGPLPFAWDI
Sbjct  1    MASSEDVIKEFMRFKVRMEGSVNGHEFEIEGEGEGRPYEGTQTAKLKVTKGGPLPFAWDI  60

Query  401  LSPQFQYGSKAYVKHPADIPDYLKLSFPEGFKWERVMNFEDGGVVTVTQDSSLQDGEFIY  580
            LSPQFQYGSKAYVKHPADIPDYLKLSFPEGFKWERVMNFEDGGVVTVTQDSSLQDGEFIY
Sbjct  61   LSPQFQYGSKAYVKHPADIPDYLKLSFPEGFKWERVMNFEDGGVVTVTQDSSLQDGEFIY  120

Query  581  KVKLRGTNFPSDGPVMQKKTMGWEASTERMYPEDGALKGEIKMRLKLKDGGHYDAEVKTT  760
            KVKLRGTNFPSDGPVMQKKTMGWEASTERMYPEDGALKGEIKMRLKLKDGGHYDAEVKTT
Sbjct  121  KVKLRGTNFPSDGPVMQKKTMGWEASTERMYPEDGALKGEIKMRLKLKDGGHYDAEVKTT  180

Query  761  YMAKKPVQLPGAYKTDIKLDITSHNEDYTIVEQYERAEGRHSTGA  895
            YMAKKPVQLPGAYKTDIKLDITSHNEDYTIVEQYERAEGRHSTGA
Sbjct  181  YMAKKPVQLPGAYKTDIKLDITSHNEDYTIVEQYERAEGRHSTGA  225
```

Related Information

Identical Proteins - Proteins identical to the subject

A TYPICAL GENETIC CASSETTE

- Most functional genetic units are made up of a Promoter, an RBS, a CDS, and finally a Terminator



Promoter RBS CDS Terminator

- This can help you look for the other parts of the cassette

IDENTIFYING THE PROMOTER AND TERMINATOR

- Promoter and terminator boundaries are often poorly characterised or annotated but there are still a few tricks to do
  - Blast the sequence upstream and downstream of the CDS to find the promoter
  - Check genbank entries of alignments to find annotations
  - Look for cloning sites/scars to determine the boundaries

Display Settings: GenBank Send:

Dual controller plasmid for Boolean Integrase Logic, complete sequence

GenBank: KC529324.1

FASTA Graphics

Go to:

```
LOCUS       KC529324                 75 bp    DNA     linear   SYN 12-MAY-2013
DEFINITION  Dual controller plasmid for Boolean Integrase Logic, complete
            sequence.
ACCESSION   KC529324 REGION: 2..76
VERSION     KC529324.1  GI:490341930
KEYWORDS    .
SOURCE      Dual controller plasmid for Boolean Integrase Logic
  ORGANISM  Dual controller plasmid for Boolean Integrase Logic
            other sequences; artificial sequences; vectors.
REFERENCE   1  (bases 1 to 75)
  AUTHORS   Bonnet,J., Yin,P., Ortiz,M.E., Subsoontorn,P. and Endy,D.
  TITLE     Amplifying genetic logic gates
  JOURNAL   Science 340 (6132), 599-603 (2013)
   PUBMED   23539178
REFERENCE   2  (bases 1 to 75)
  AUTHORS   Bonnet,J.
  TITLE     Direct Submission
  JOURNAL   Submitted (25-JAN-2013) Bioengineering, Stanford University, 473
            Via Ortega, Stanford, CA 94305, USA
FEATURES             Location/Qualifiers
     source          1..75
                     /organism="Dual controller plasmid for Boolean Integrase
                     Logic"
                     /mol_type="other DNA"
                     /db_xref="taxon:1330438"
     misc_feature    <1..21
                     /note="prefix Bba"
     regulatory      22..75
                     /regulatory_class="promoter"
                     /note="pTetO"
ORIGIN
        1 aattcgcggc cgcttctaga gtccctatca gtgatagaga ttgacatccc tatcagtgat
       61 agagatactg agcac
//
```

IDENTIFYING THE RBS

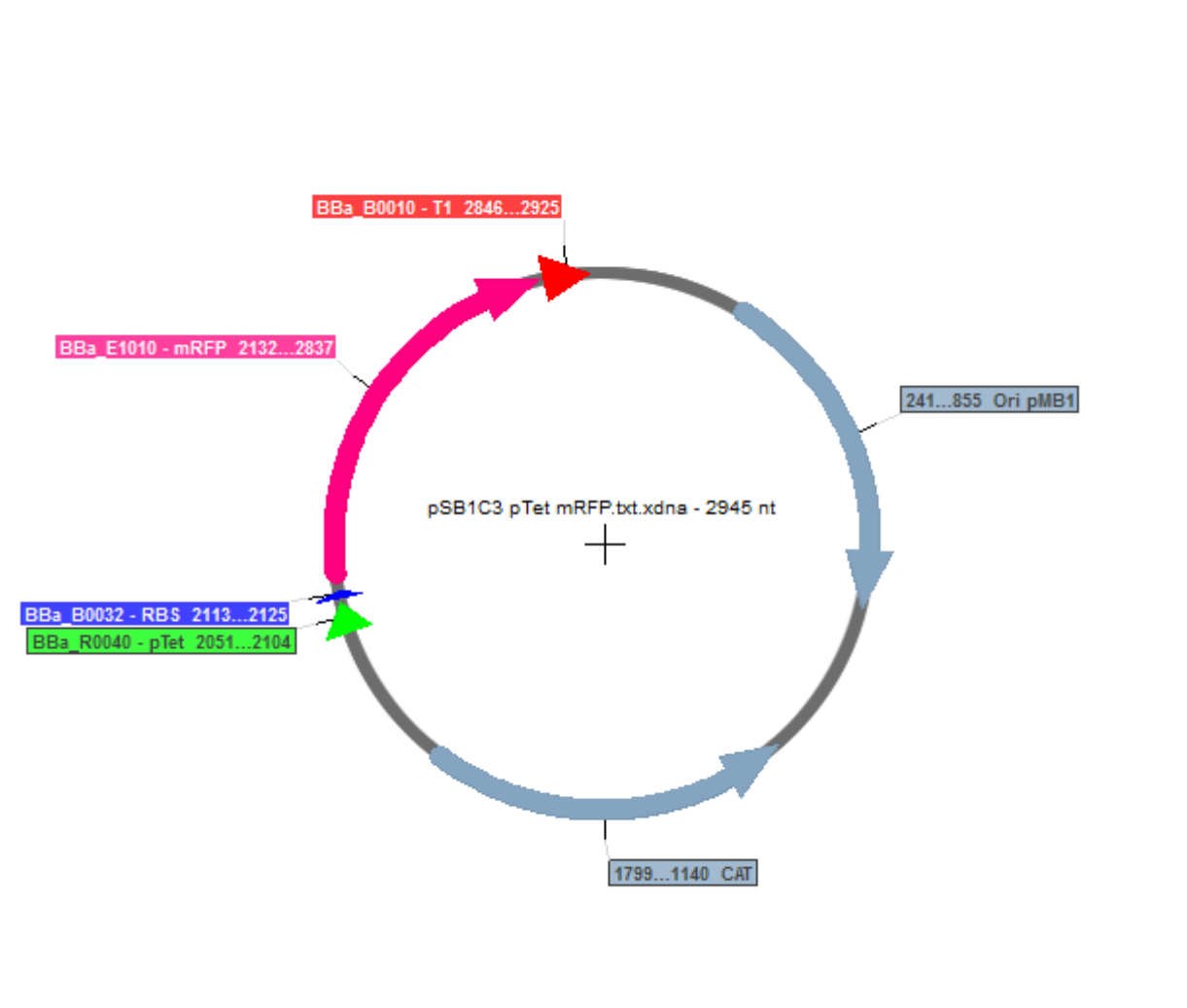
- The easiest part of the RBS to recognize is the Shine Dalgarno sequence.
  - The Consensus sequence in E. coli is AGGAGGU
  - Synthetic RBS will not necessarily follow this consensus
- If you fail to find a sequence similar to the Shine Dalgarno sequence between your promoter and CDS, search for Biobrick Scar sites or cloning sites which may indicate the RBS location.
- Compare to Synthetic RBS on parts.igem.org
- Reverse engineer mRNA with Salis RBS calculator at https://salis.psu.edu

FINDING CLONING ARTIFACTS

- A biobrick scar is the sequence actaga and results from the ligation of a XbaI cut site with a SpeI cut site.
- Finding this sequence between the promoter, RBS, CDS, and terminator is a clear indication of the boundaries of each part.
- Additionally, Common Restriction sites can indicate restriction sites used to construct the Plasmid (EcoRI, BamHI, PstI, XbaI, HindIII, KpnI, NcoI, SacI, XhoI)
- Many modern cloning techniques (golden gate, gibson assembly) are called scarless cloning techniques because they avoid these artifacts, making them more difficult to reverse engineer.

THE ANNOTATED SEQUENCE

- Serial cloner (and most forms of in silico cloning software) will allow you to visualize your annotated plasmids.
- Click on Graphic Map



CREATING A NEW VECTOR

- Modify the vector provided to express *mecA* (which confers methicillin resistence in *Staphylococcus aureus*)

  You will need the following tools:

  - Biocyc / Kegg
  - RBS calculator (https://salis.psu.edu)
  - Codon optimization (www.**jcat**.de/)

BIOCYC.ORG

BioCyc Database Collection

Pathway Tools Tutorial Date Poll

LOGIN | Why Login? | Create New Account

Enter a gene, protein, metabolite or pathway... Quick Search Gene Search

Searching *Staphylococcus aureus aureus N315* change organism database

Sites ▾ | Search ▾ | Genome ▾ | Metabolism ▾ | Analysis ▾ | SmartTables ▾ | Help ▾

BioCyc Database Collection

BioCyc is a collection of 7667 Pathway/Genome Databases (PGDBs), plus software tools for understanding their data.

Getting Started

New to BioCyc? Typical usage is:

- Select one or more databases (genomes) to search. To do so, click "change organism database" in the box in the top right of every page. By default, BioCyc searches *Escherichia coli* K-12 substr. MG1655.
- Search for a gene or pathway using the Quick Search, or see the Search menu for more options.

New User Guide >>

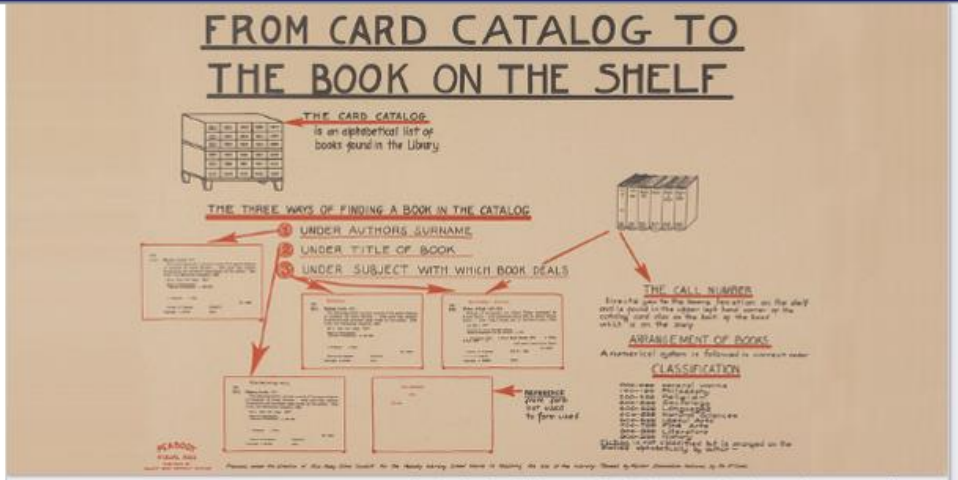


Image of Peabody Visual Aids poster by char booth. Used under creative commons license.

Search BioCyc by Locus ID

You can now enter locus IDs into the BioCyc quick search box. There is no need to first select the correct BioCyc organism database - if the organism is part of BioCyc, the unique locus ID will automatically find it. Some example locus IDs look like this: MSM_0046, MXAN_1061.

1 2 3 4 5 6 7 8 9 10 11 12

Tools

BioCyc provides tools for navigating, visualizing, and analyzing the

BioCyc Databases

The BioCyc databases are divided into three tiers, based on their quality.

BIOCYC.ORG

BioCyc Database Collection

Pathway Tools Tutorial Date Poll

LOGIN | Why Login? | Create New Account

Enter a gene, protein, metabolite or pathway... Quick Search Gene Search

Searching *Staphylococcus aureus aureus N315* change organism database

Sites | Search | Genome | Metabolism | Analysis | SmartTables | Help

gene **mecA** polypeptide **penicillin binding protein 2 prime**

Staphylococcus aureus aureus N315

log in to add to SmartTable.

Accession IDs	GJCB-41 (Saur158879Cyc) SA0038	Length	2007 bp / 668 aa
		Map Position	[45,031 <- 47,037] (1.6 centisomes, 6°) on NC_002745 chromosome View in Genome Browser

Summary | Operons | Show All

Molecular Weight of Polypeptide	76.102 kD (from nucleotide sequence)

Unification Links

Entrez	15925745
Entrez-gene	1122812
NCBI-Protein	NP_373278.1
UniProt	NP_373278.1

Report Errors or Provide Feedback
Page generated by Pathway Tools version 19.5 (software by SRI International) on Mon Jan 18, 2016, biocyc14.
saur158879cyc version 19.0.

hide

OPERATIONS

Sequences
- Protein Sequence
- Nucleotide Sequence
- Save Nucleotide Sequence to file
- Save Protein Sequence to file

Comparison Operations
- Show this gene in another database
- Change organisms/databases for comparison operations
- Search for this gene in other databases
- Show orthologs (with operon diagrams) in multiple databases
- Align in Multi-Genome Browser
- Align gene nucleotide sequence with orthologs
- Align gene product amino acid sequence with orthologs

http://www.genome.jp/kegg/

KEGG

KEGG

Staphylococcus aureus subsp. aureus N315 (MRSA/VSSA): SA0038

Help

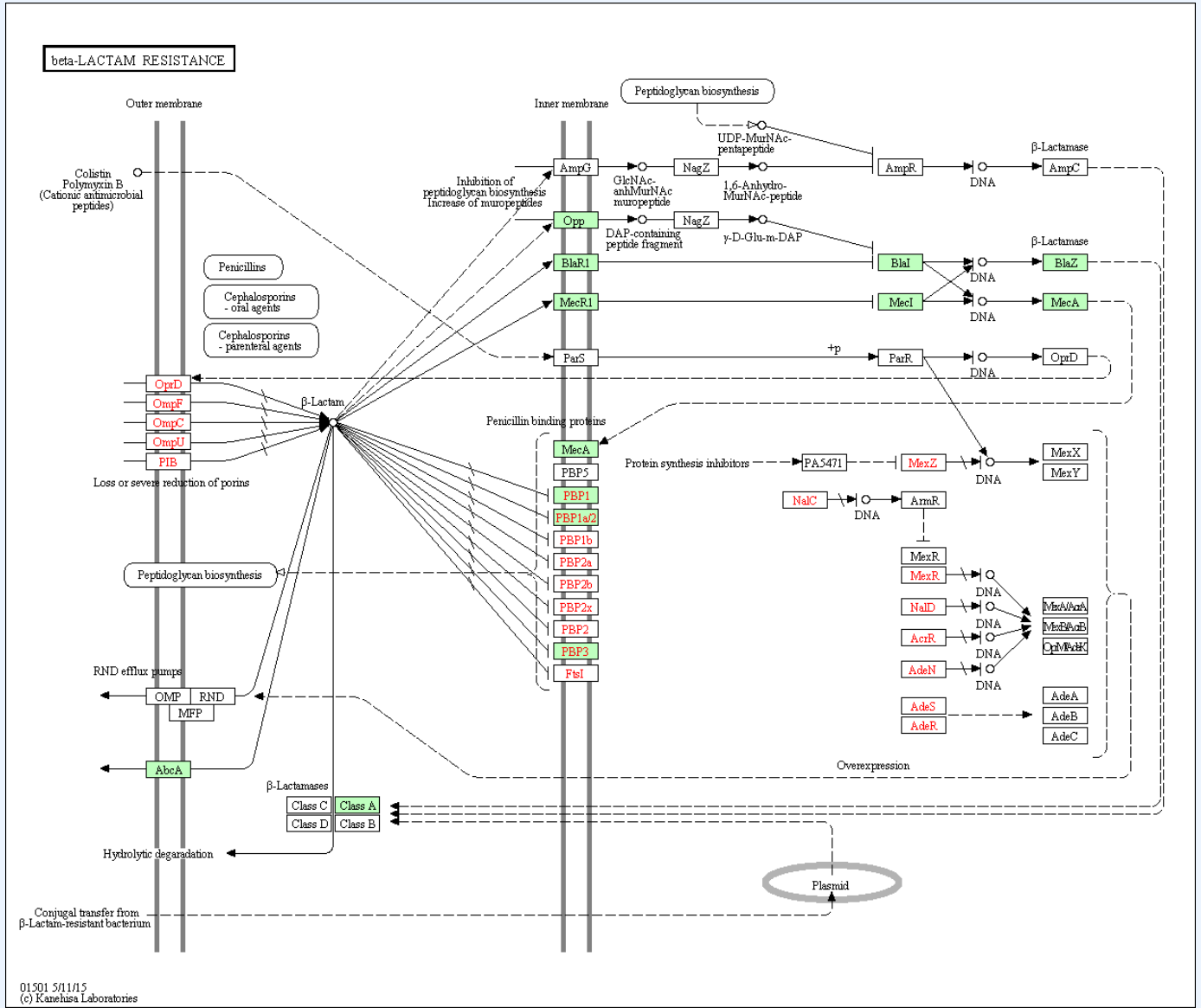
Entry	SA0038 CDS T00051
Gene name	mecA
Definition	(RefSeq) penicillin binding protein 2 prime
KO	K02545 penicillin-binding protein 2 prime
Organism	sau Staphylococcus aureus subsp. aureus N315 (MRSA/VSSA)
Pathway	sau01501 beta-Lactam resistance
Module	sau_M00625 Methicillin resistance
Brite	KEGG Orthology (KO) [BR:sau00001] Human Diseases Drug resistance 01501 beta-Lactam resistance SA0038 (mecA) BRITE hierarchy
SSDB	Ortholog Paralog Gene cluster GFIT
Motif	Pfam: Transpeptidase PBP_dimer MecA_N HK_sensor Motif
Other DBs	NCBI-GI: 15925745 NCBI-GeneID: 1122812 NITE: SA0038
Structure	PDB: 3ZG5 Thumbnail  Jmol
Position	complement(45031..47037) Genome map
AA seq	668 aa AA seq DB search MKKIKIVPLILIVVVVGFGIYFYASKDKEINNTIDAIEDKNFKQVYKDSSYISKSDNGEV EMTERPIKIYNSLGVKDINIQDRKIKKVSKNKKRVDAQYKIKTNYGNIDRNVQFNFVKED GMWKLDWDHSVIIPGMQKDQSIHIENLKSERGKILDRNNVELANTGTAYEIGIVPKNVSK KDYKAIAKELSISEDYIKQQMDQNWVQDDTFVPLKTVKKMDEYLSDFAKKFHLTTNETES

All links

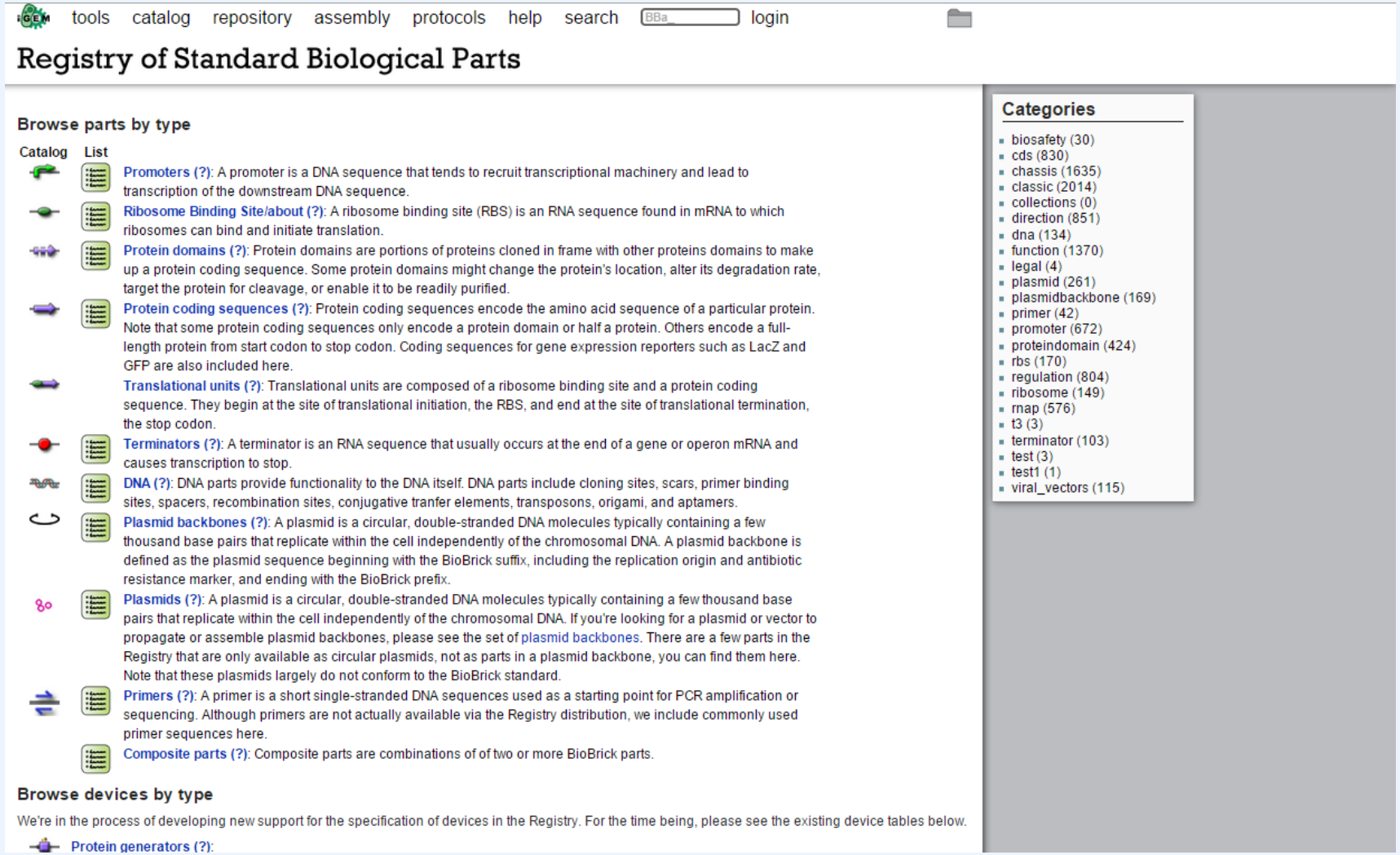
Ontology (1)
KEGG BRITE (1)
Pathway (2)
KEGG PATHWAY (1)
KEGG MODULE (1)
Genome (1)
KEGG GENOME (1)
Gene (11)
KEGG ORTHOLOGY (1)
RefGene (6)
NCBI-Gene (1)
NCBI-GI (1)
NITE (1)
OC (1)
3D Structure (1)
PDB (1)
Protein domain (4)
Pfam (4)
All databases (20)

Download RDF

KEGG

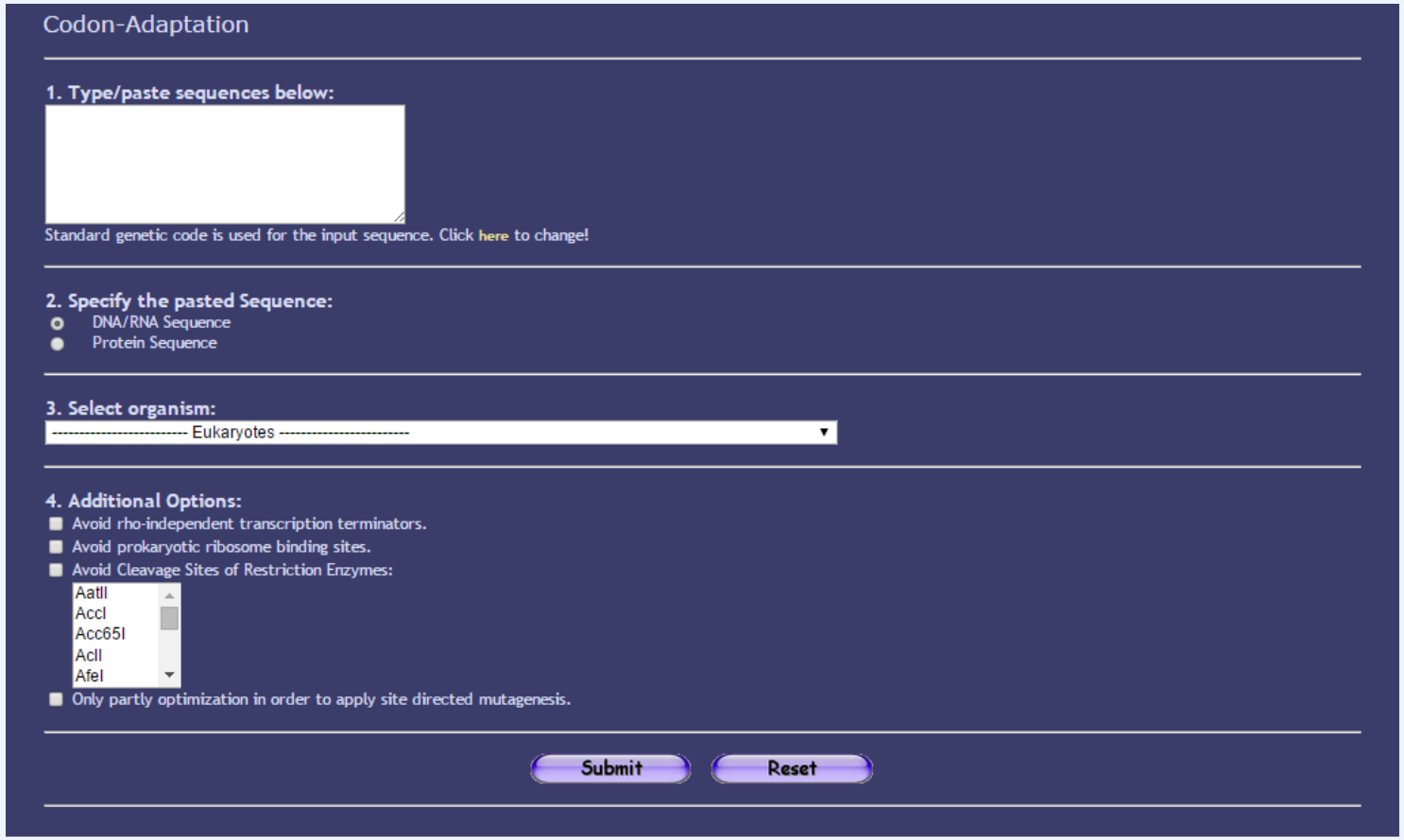

beta-LACTAM RESISTANCE
Outer membrane
Inner membrane
Peptidoglycan biosynthesis
UDP-MurNAc-pentapeptide
β-Lactamase
Colistin
Polymyxin B
(Cationic antimicrobial peptides)
Inhibition of peptidoglycan biosynthesis
Increase of muropeptides
AmpG
NagZ
AmpR
AmpC
GlcNAc-anhMurNAc muropeptide
1,6-Anhydro-MurNAc-peptide
DNA
Opp
DAP-containing peptide fragment
γ-D-Glu-m-DAP
BlaR1
BlaI
BlaZ
MecR1
MecI
MecA
Penicillins
Cephalosporins - oral agents
Cephalosporins - parenteral agents
ParS
+p
ParR
OprD
OmpF
OmpC
OmpU
PIB
β-Lactam
Loss or severe reduction of porins
Penicillin binding proteins
PBP5
PBP1
PBP1a/2
PBP1b
PBP2a
PBP2b
PBP2x
PBP2
PBP3
FtsI
Protein synthesis inhibitors
PA5471
MexZ
MexX
MexY
NalC
ArnR
MexR
NalD
AcrR
AdeN
AdeS
AdeR
AdeA
AdeB
AdeC
Overexpression
RND efflux pumps
OMP
RND
MFP
AbcA
β-Lactamases
Class C
Class A
Class D
Class B
Hydrolytic degaradation
Plasmid
Conjugal transfer from β-Lactam-resistant bacterium
01501 5/11/15
(c) Kanehisa Laboratories

THE REGISTRY OF BIOLOGICAL PARTS



- http://parts.igem.org/
- Good resource for finding DNA sequences for specific parts
- Characterization quality varies

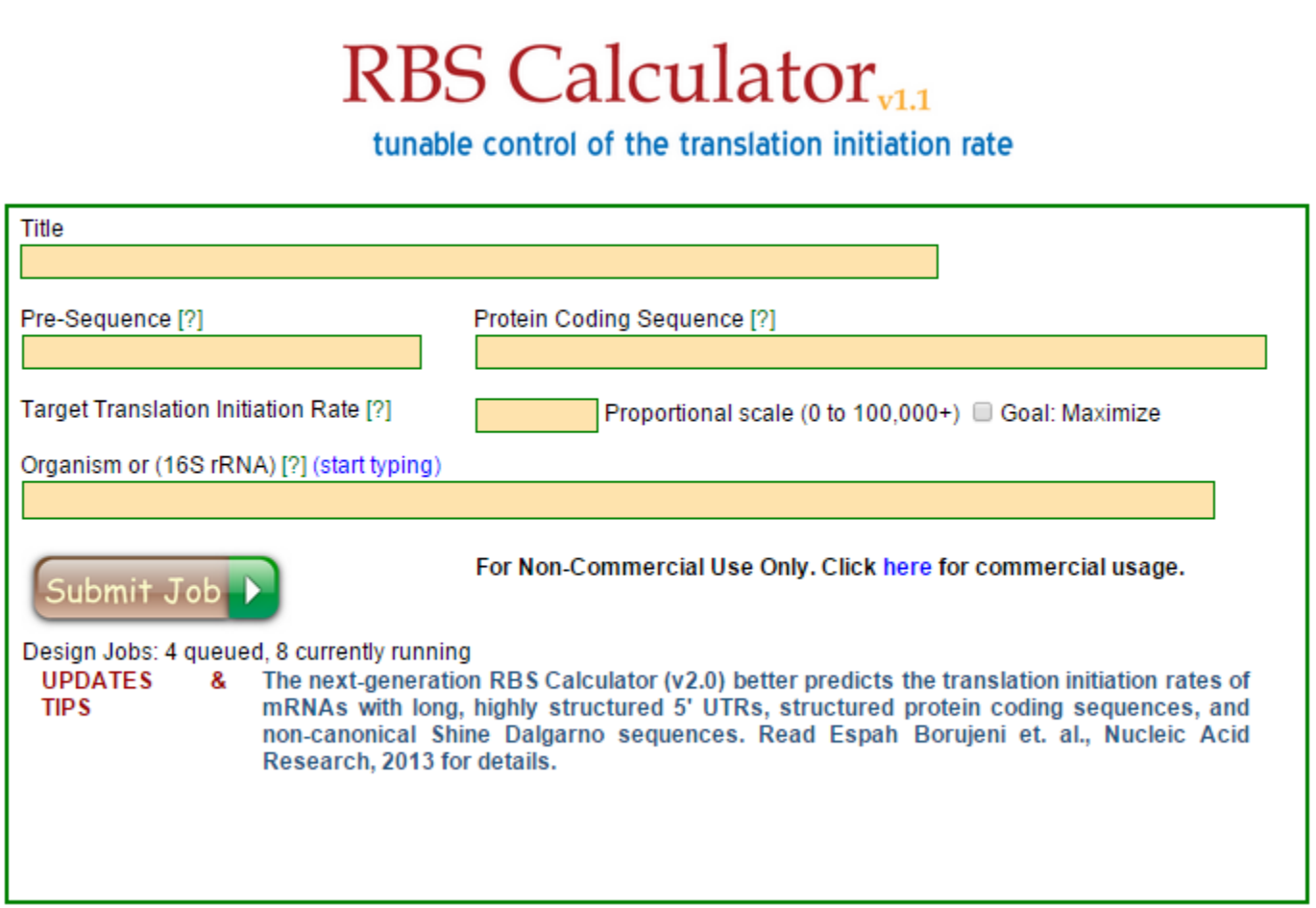
CODON OPTIMIZATION



- http://www.jcat.de/
- Alternatives from Lifetechnologies, genscript, IDT and other synthesis companies
- Optimize expression for genes when transferring between species

RBS CALCULATOR

- Translation rate depends on RBS
- RBS are not completely modular, the translation initiation rate will vary with the protein coding sequence
- Can design RBS for targeted translation initiation rate
- Can reverse engineer mRNA to determine relative translation initiation rate



FINISHING UP YOUR NEW INSERT

- Add restriction sites (for Traditional, Biobrick, or Goldengate cloning) to your sequence to prepare it for synthesis. Add a few (3-5) bases on the ends of your sequence to ensure good binding of restriction enzymes.
- **or**
- Add homology regions for gibson assembly.
- The sequence can be synthesized by DNA synthesis companies such as IDT.

PRIMERS FOR YOUR BACKBONE

Primers can be designed with primer3 (bioinfo.ut.ee/primer3/) or using cloning software. You can use the sequences below for binding to your backbone.

- pSB1C3 pTet F = TACTAGAGCCAGGCATCAA
- pSB1C3 pTet R = AGTAGTGCTCAGTATCTCTATC

Add restriction sites (for Traditional, Biobrick, or Goldengate cloning) to your primers on the 5' end to prepare them for synthesis. Add a few (3-5) bases on the 5' of your sequence to ensure good binding of restriction enzymes.

or

Add homology regions for gibson assembly.

IN SILICO CLONING

You can do in silico PCR by clicking on ‘PCR’ button in serial cloner.

Select your template (your now-annotated vector) and input the primers you designed.

You can do in silico cloning by clicking on ‘Construct’ button in serial cloner. Select your PCR product, and your new insert, and the restriction enzymes you want to use for the cloning.

Save your constructed sequence and check that it is what you expect.