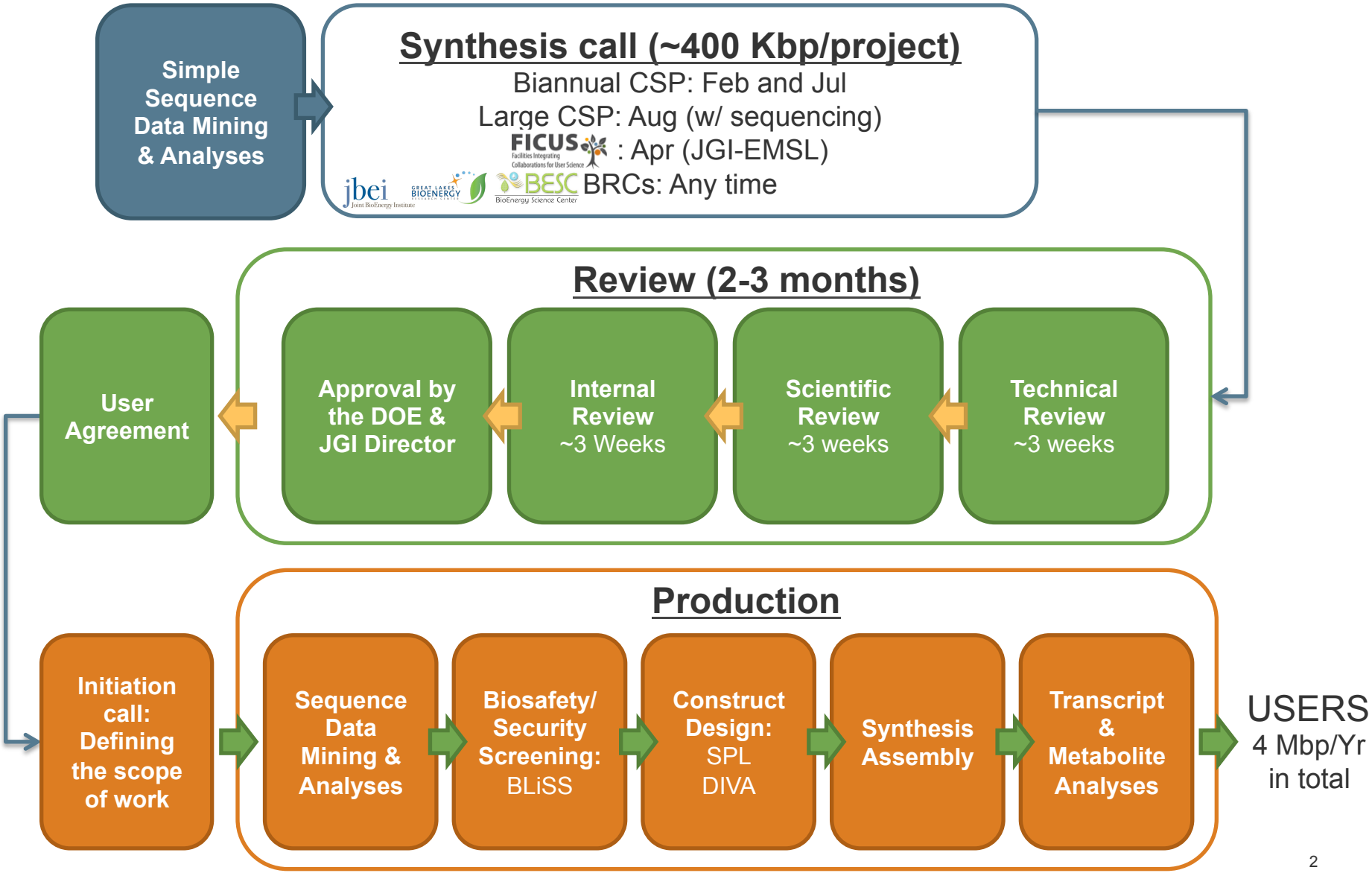
JGI JOINT GENOME INSTITUTE
UNITED STATES DEPARTMENT OF ENERGY

Design tools for synthetic biology

Nathan J. Hillson
njhillson@lbl.gov

2016 JGI User Meeting – GenTech Workshop
March 22, 2016

DNA Synthesis Science CSP Project



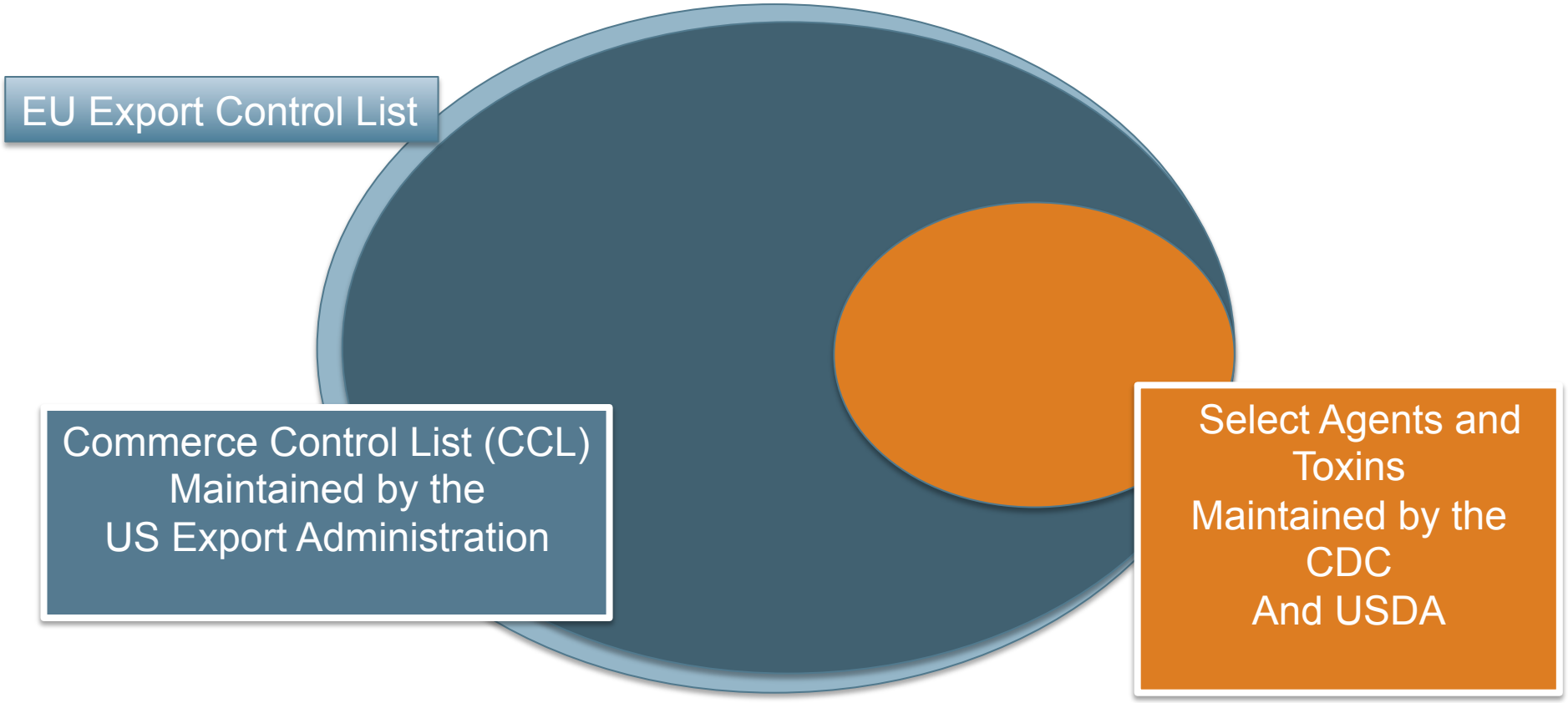
Synthetic Biology – Biosecurity Issues

```
>DNA_construct_seq
ATGAAGAGAAAGAACAAAGTACT
ATCTATCCTGCTTACCCTGTTGC
TTATAATTTCGACAACCTCAGTG
AACATGAGTTTTGCTGAAGCTAC
```

DNA Synthesis

The HH&S Guidance

1) Customer Screening – establish the legitimacy of customers
2) Sequence Screening – identify when “sequences of concern” are ordered
3) Follow-up screening – verify the legitimacy of the end-use



Identifying "Sequences of Concern"

JGI JOINT GENOME INSTITUTE

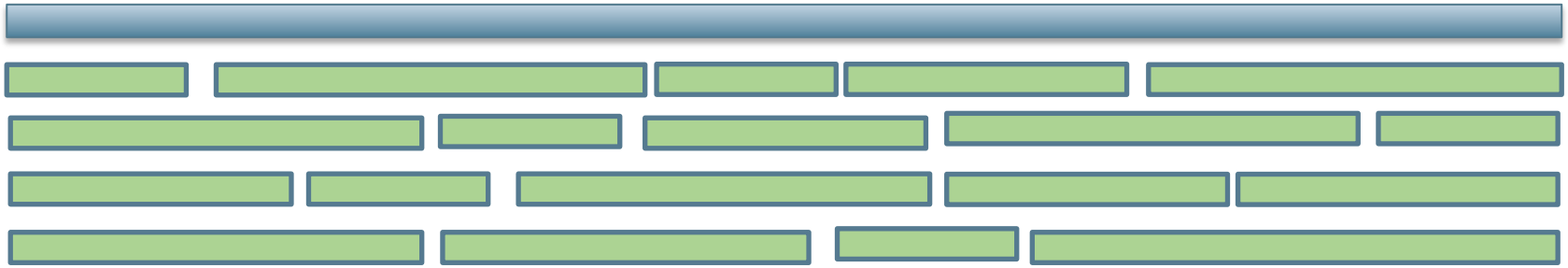
Technical recommendations for identification:

- Identify sequences that are 200 bp or longer
- Screen both DNA strands, and all six-frame translations
- Use sequence alignment methods against a public sequence database
- Use a "Best Match" approach to determine whether a query sequence is derived from or encodes a Select Agent or Toxin or, a sequence from a CCL-listed item
- Screening should identify sequences that are "unique" to Select Agents and Toxins and that "house-keeping genes" that maintain normal cellular physiology should be excluded

BLiSS Pipeline

1) BLAST the query sequence (to be synthesized) using blastn (nr) and blastx (nt)

ATGAAGAGAAAGAACCTGCTTACCC.....TGCAAAGAGTTTTGCTGAAGCTAC

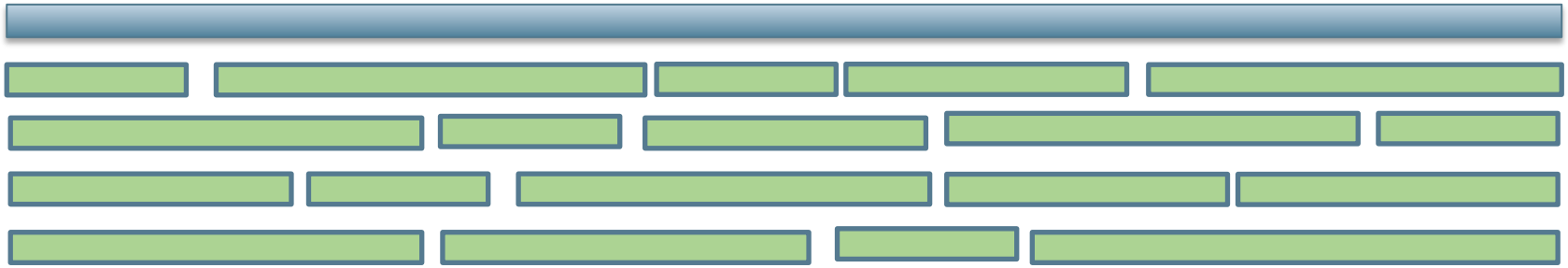


This yields a collection of alignments for each DNA strand,
and all six frame translations (amino acids)

BLiSS Pipeline

2) Discard all alignments that are under 200 nucleotides or 66 amino acids in length

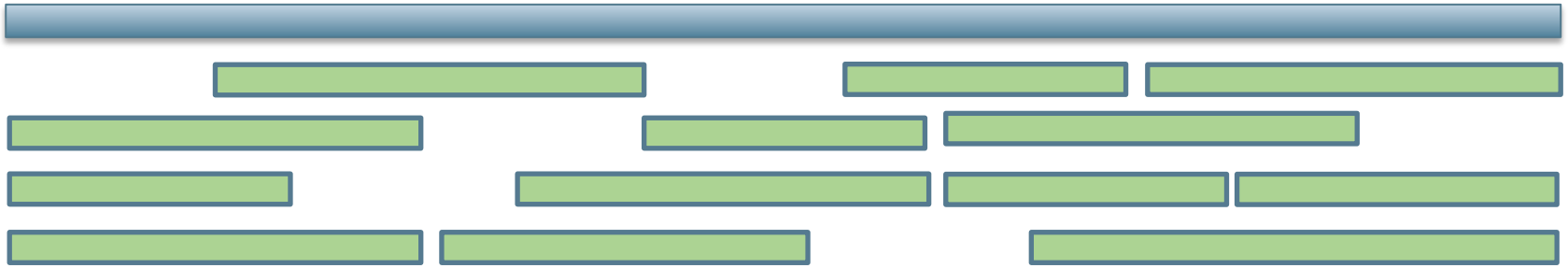
ATGAAGAGAAAGAACCTGCTTACCC.....TGCAAAGAGTTTTGCTGAAGCTAC



BLiSS Pipeline

3) Screen remaining alignments against the blacklist database

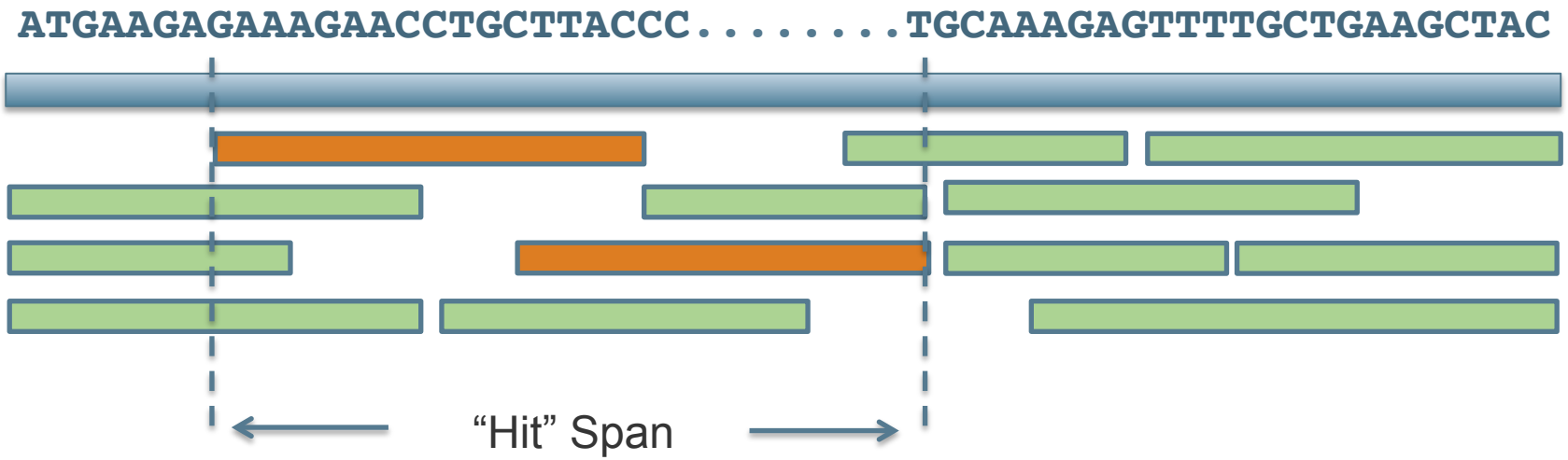
ATGAAGAGAAAGAACCTGCTTACCC.....TGCAAAGAGTTTTGCTGAAGCTAC



Matches to sequences of concern are identified

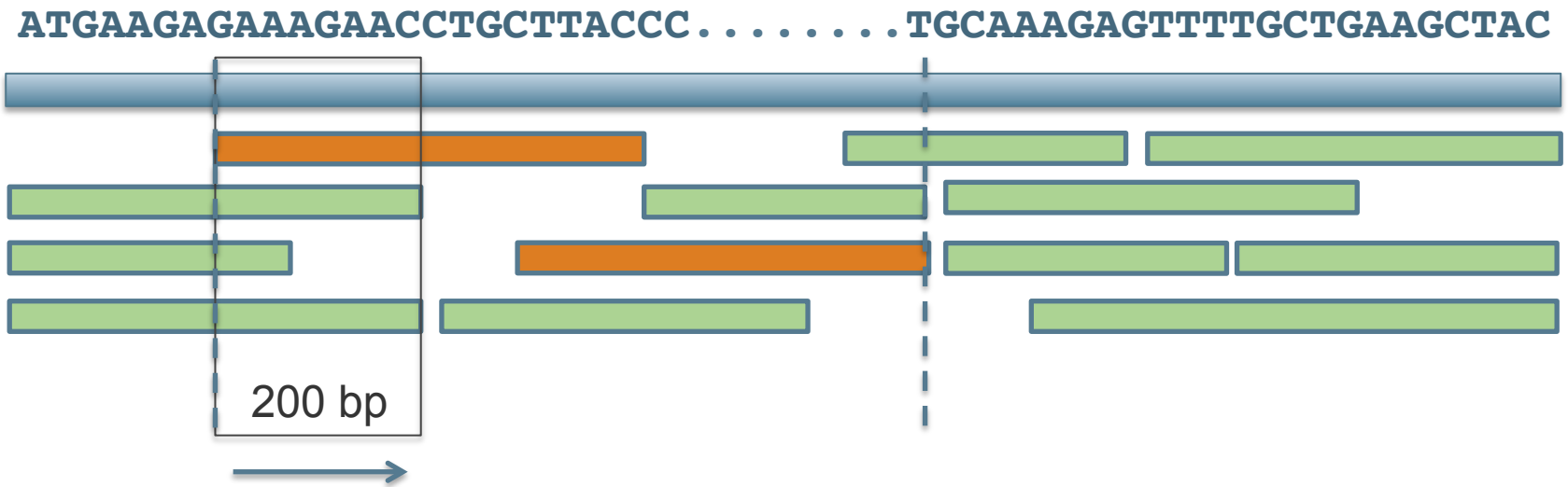
BLiSS Pipeline

4) Calculate the span(s) of the query sequence corresponding to these “hits”



BLiSS Pipeline

5) For each 200 bp window along the hit span, align with all overlapping alignments



For each 200 bp window along the hit span, BLAST the window query sequence against each alignment that overlaps the window

6) This gives us a % identity for each local alignment for the window, which we use to assign “Best Matches” for each window.

7) Each window is then assigned a “Status”

BLiSS Window Status Assignment

JGI JOINT GENOME INSTITUTE

Statuses:

Failed	Best match to an entity on the **Select Agents and Toxins List**
Controlled	Best match to an entity on the **Commerce Control List**
Passed	Not a best match to any sequence of concern

Controlled Sequences

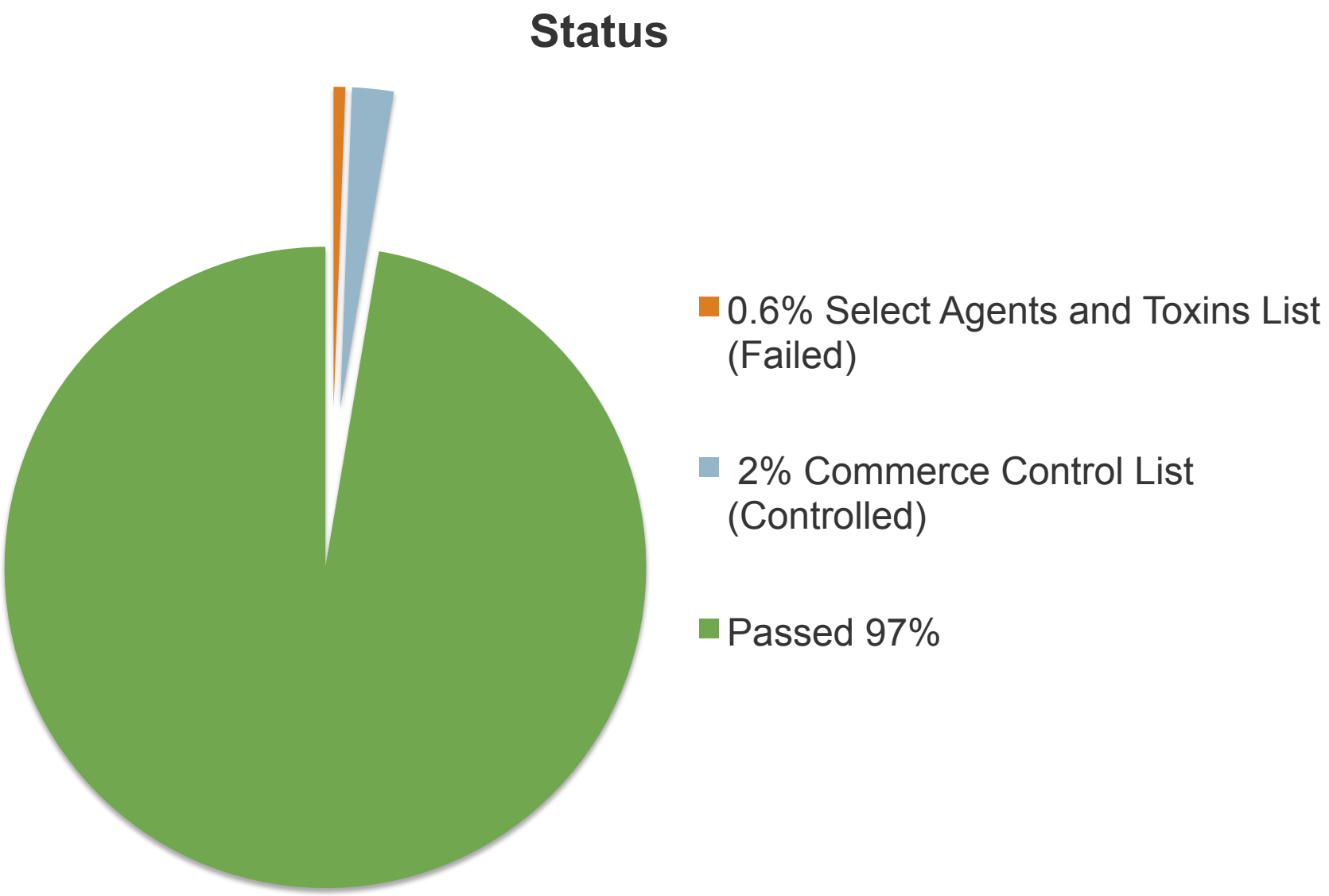
- Consult with our Export Control Department

Failed Sequences

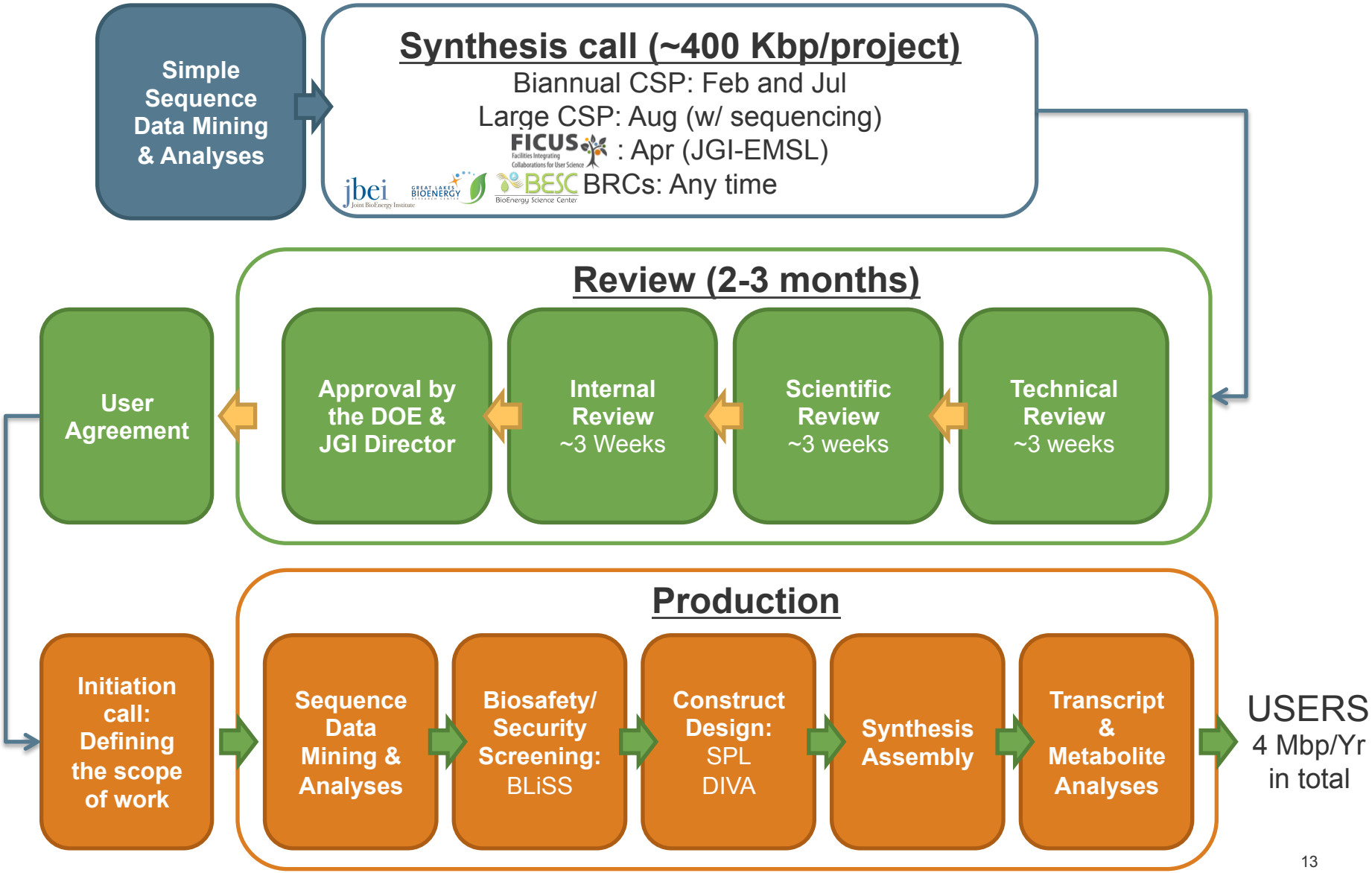
- Ask the user that submitted the sequence verify the legitimacy of the end-use
- Consult with our local FBI WMD coordinator

Failed sequences are not synthesized unless they have been cleared by the FBI

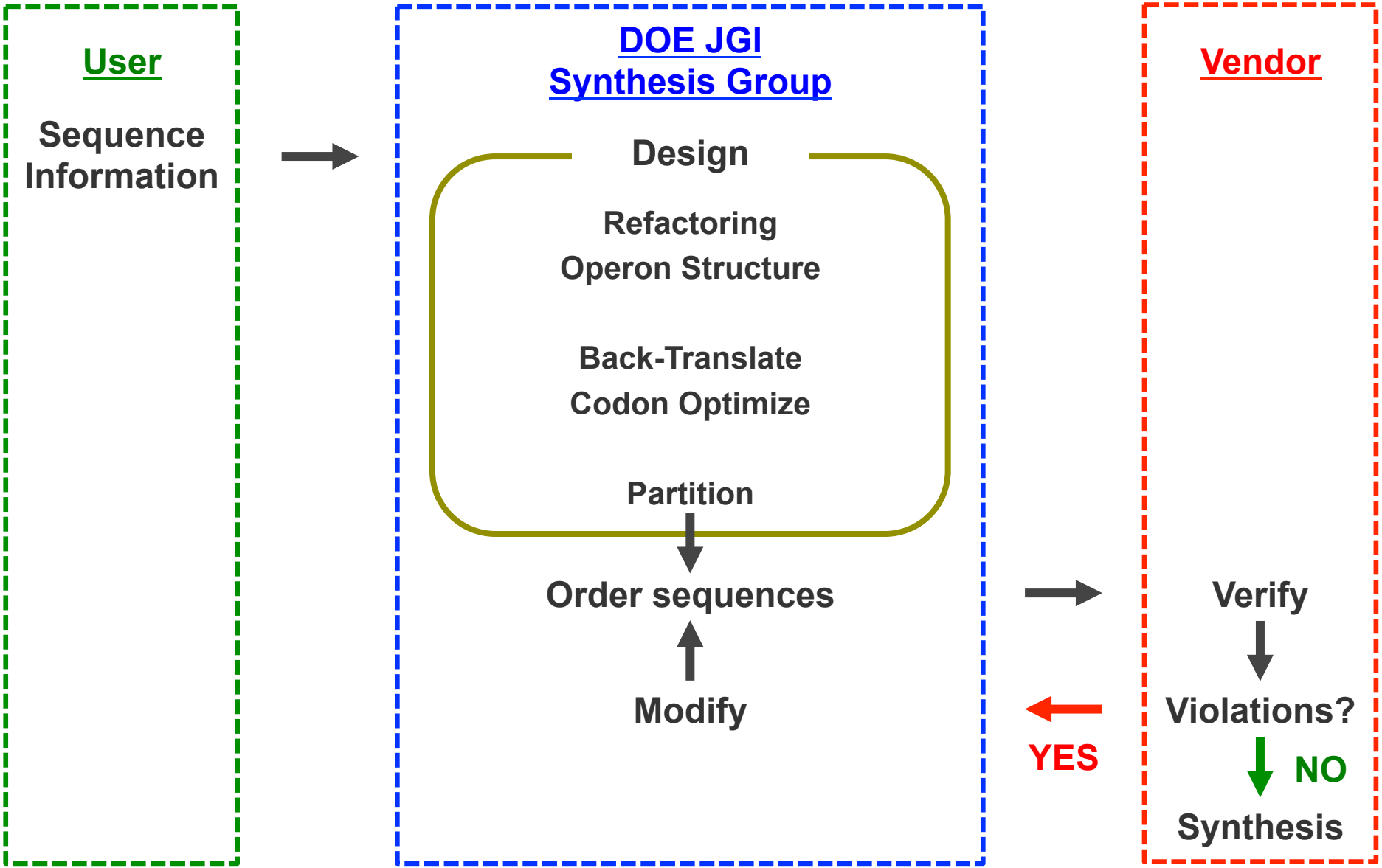
Statistics: 9192 Sequences Screened



DNA Synthesis Science CSP Project



FY 2015


User
Sequence Information
DOE JGI
Synthesis Group
Design
Refactoring
Operon Structure
Back-Translate
Codon Optimize
Partition
Order sequences
Modify
Vendor
Verify
Violations?
YES
NO
Synthesis

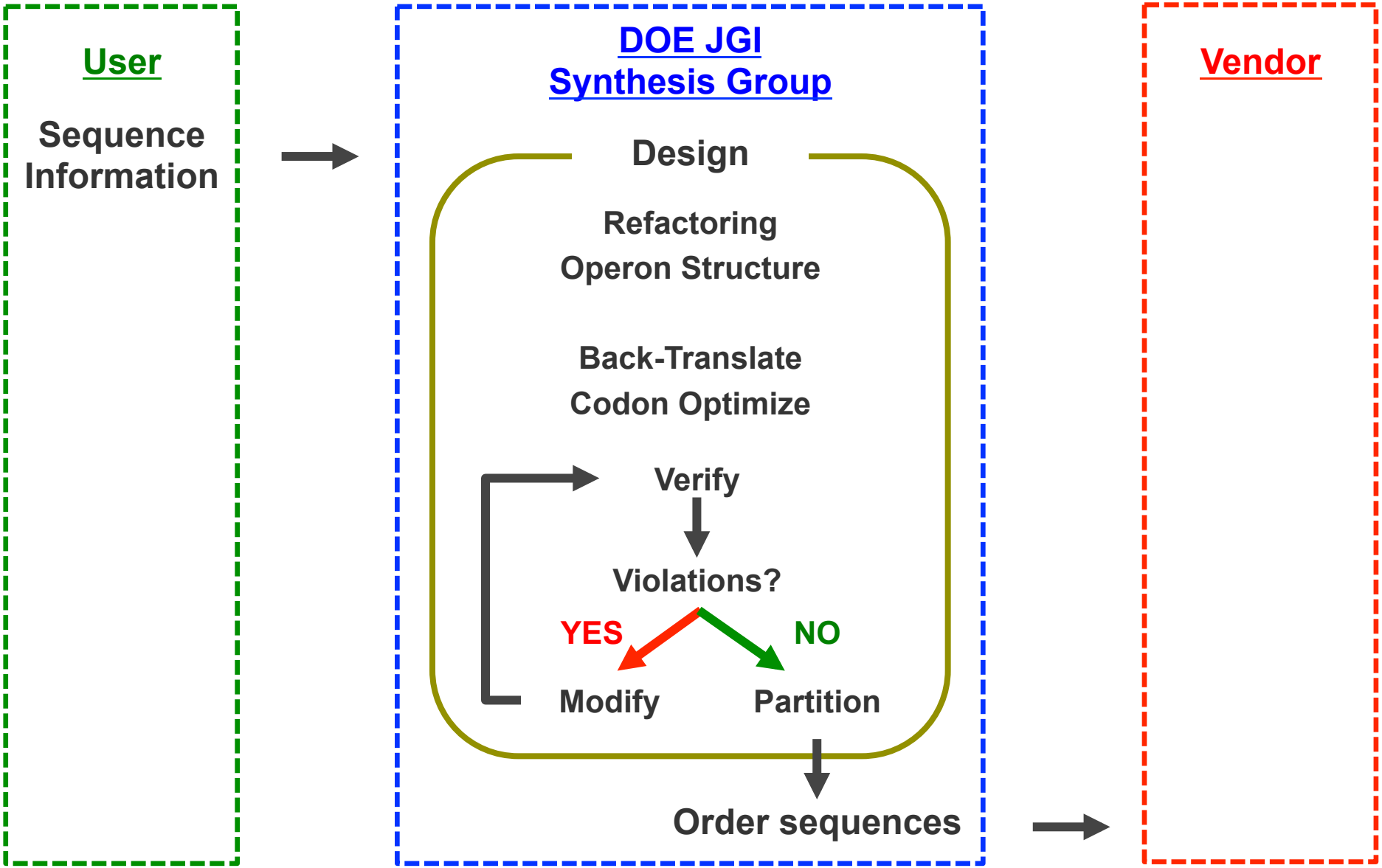
What are common Synthesis Criteria?

- **Length (bp)**
- **GC% Content**
  - Global vs. Local ("Windowed")
  - GC% Flux
- **Repeats and Repeat Coverage**
  - Global vs. Local ("Windowed")
  - Tandem vs. Interspersed
  - Direct vs. Inverted
  - Exact vs. Mutated
- **Restriction Sites**

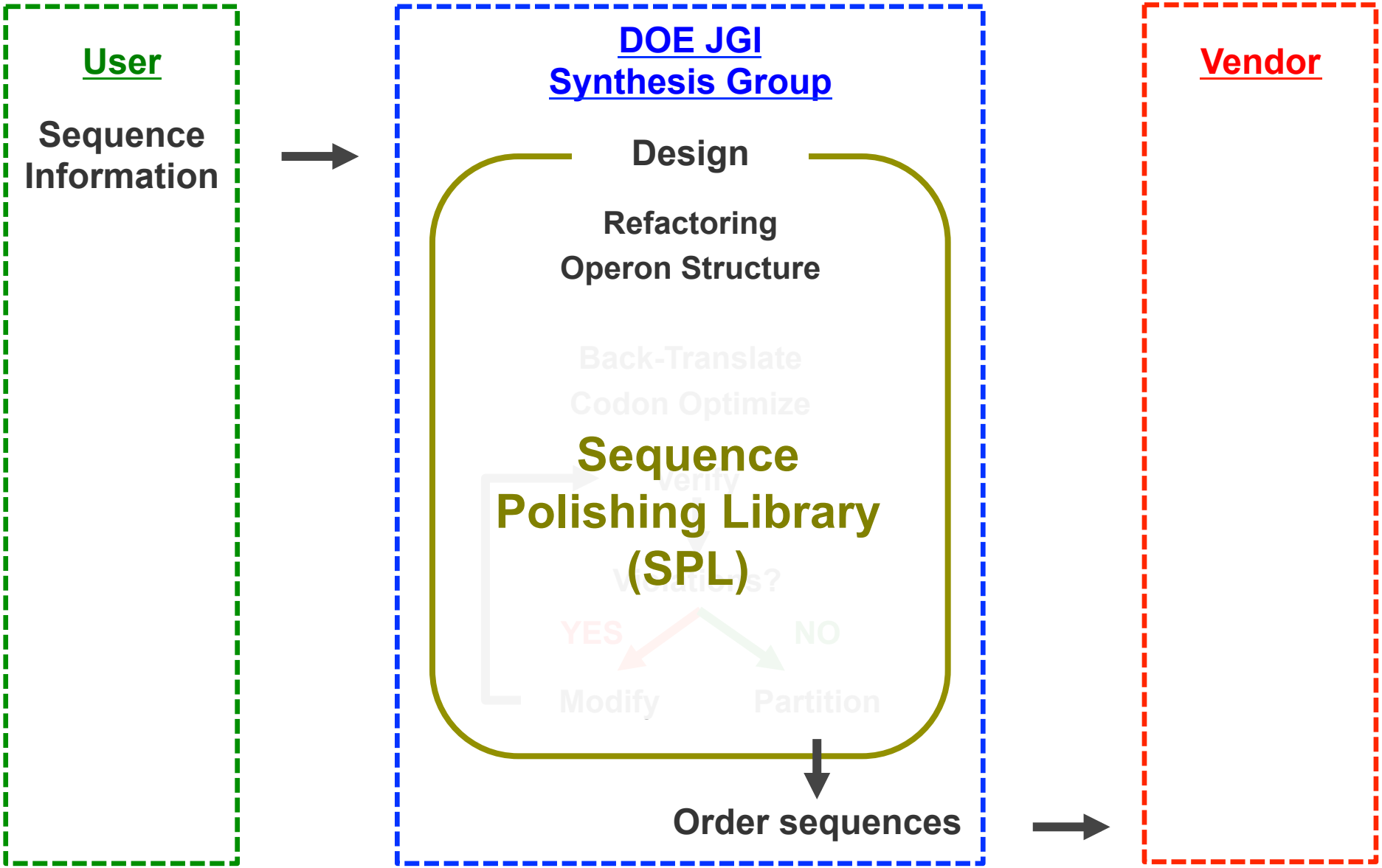
Example: **Gen9**

http://help.gen9bio.com/docs/gene-synthesis-guidelines

Improved Design for Synthesis and Assembly



Improved Design for Synthesis and Assembly



Enabling Customizable and Automated Workflows

JGI JOINT GENOME INSTITUTE

- SPL Modules

JUGGLER
- Back-Translate
- Codon Optimize

POLISHER
- Verify
- Modify

PARTITIONER
- Partition

Work in Progress:
SPL Web User-Interface

JGI JOINT GENOME INSTITUTE

JGI JOINT GENOME INSTITUTE UNITED STATES DEPARTMENT OF ENERGY | Home | Manual | Q&A Forum | Contact | Logout

Sequence Information

Enter Sequ

Select a

Choos

Enter Co

Predef

Gen9

Enter P

Pattern S

```
>AarI
CACCTG
>BsaI
GGTCTC
```

Save F

Check against synthesis constraints (e.g. GC%, repeats)?

Check against pattern sequences (e.g. restriction sites)?

Suggest Modifications?

Codon Replacement Strategy:

Random

Enter Codon Usage Table | Upload Codon Usage Table

Codon Usage Table:

```
UUU 22.2( 35846)  UCU  8.7( 14013)  UAU 16.5( 26648)  UGU  5.2(  8458)
UUC 15.9( 25565)  UCC  8.9( 14420)  UAC 12.3( 19766)  UGC  6.4( 10285)
UUA 13.8( 22316)  UCA  8.1( 13117)  UAA  2.0(  3163)  UGA  1.1(  1751)
UUG 13.0( 20904)  UCG  8.8( 14220)  UAG  0.3(   435)  UGG 15.3( 24656)

CUU 11.4( 18366)  CCU  7.2( 11657)  CAU 12.8( 20631)  CGU 20.2( 32590)
CUC 10.5( 16869)  CCC  5.6(  8961)  CAC  9.4( 15116)  CGC 20.8( 33547)
CUA  3.9(  6257)  CCA  8.4( 13507)  CAA 14.7( 23703)  CGA  3.8(  6166)
CUG 51.1( 82300)  CCG 22.4( 36178)  CAG 29.4( 47324)  CGG  6.2(  9955)

AUU 29.7( 47838)  ACU  9.1( 14639)  AAU 19.2( 30864)  AGU  9.4( 15123)
AUC 23.9( 38504)  ACC 22.8( 36724)  AAC 21.7( 34907)  AGC 16.0( 25800)
AUA  5.5(  8835)  ACA  8.1( 13030)  AAA 34.0( 54723)  AGA  2.9(  4656)
AUG 27.2( 43846)  ACG 15.0( 24122)  AAG 11.0( 17729)  AGG  1.8(  2915)

GUU 18.1( 29200)  GCU 15.4( 24855)  GAU 32.8( 52914)  GGU 24.2( 38983)
GUC 14.8( 23870)  GCC 25.2( 40571)  GAC 19.2( 30953)  GGC 28.1( 45226)
GUA 10.9( 17561)  GCA 20.7( 33343)  GAA 39.3( 63339)  GGA  8.9( 14286)
GUG 26.2( 42261)  GCG 32.3( 52091)  GAG 18.7( 30158)  GGG 11.8( 18947)
```

Save Codon Usage Table?

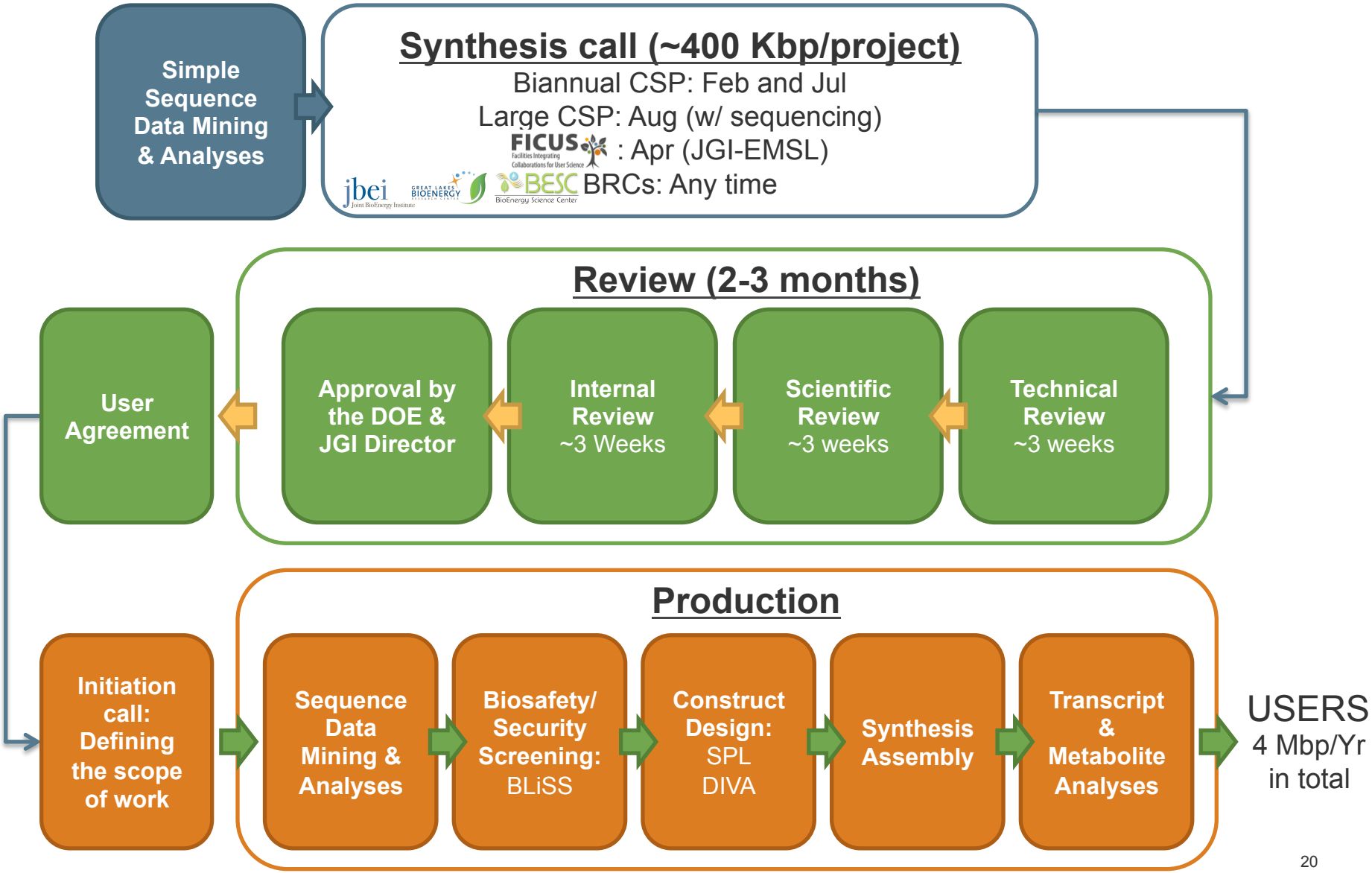
Enter codon usage table name

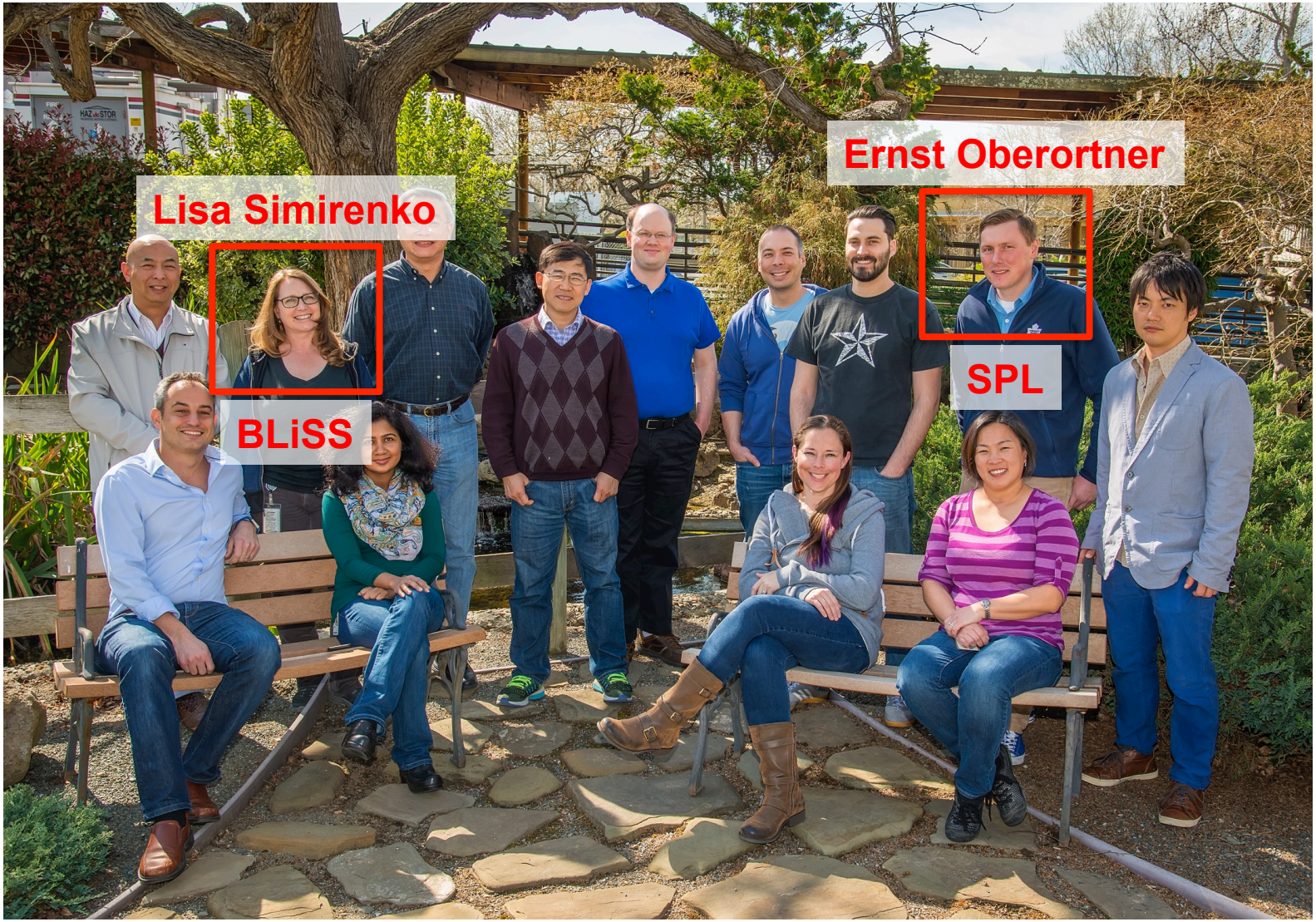
Accepted Formats:

- Codon Usage Database (CUDB)
  - Example 1
  - Example 2
- Relative Synonymous Codon Usage (RSCU)
  - Example

DOE Bioenergy Research Centers

DNA Synthesis Science CSP Project



Acknowledgements

DOE JGI DNA Synthesis Science Program