

Submitting Data to GenBank

Enabling others to use your hard work.

Where we are

- 13:30-14:00 – Primer Design to Amplify Microbial Genomes for Sequencing
- 14:00-14:15 – Primer Design Exercise
- 14:15-14:45 – Molecular Barcoding to Allow Multiplexed NGS
- 14:45-15:15 – Processing NGS Data – de novo and mapping assembly
- 15:15-15:30 – Break
- 15:30-15:45 – Assembly Exercise
- 15:45-16:15 – Annotation
- 16:15-16:30 – Annotation Exercise
- **16:30-17:00 – Submitting Data to GenBank**

Genbank Project Registration

- **Project Registration with BioProject**

- Need to register the project with NCBI before the project starts.
- It provides an organizational framework to access all data associated with a project
- Describes project scope, material, and objectives using a controlled vocabulary.

- **Project details**

- **Project type**: Genome Sequencing
- **Title**: Sequencing of Vaccine Preventable Disease Agents – Measles
- **Attributes**:
 - **Scope**: Multiisolate;
 - **Material**: Genome;
 - **Capture**: Whole;
- **Description**: Detailed description of the project
- **Relevance**: Medical
- **Objectives**: Sequence, Assembly, Annotation

Minimum Data Type Required for Submission

- **Here is a list of minimum metadata associated with a sample that needs to be submitted with the sequence**
- Organism name (it should be just the virus name without strain information in it.)
- Strain name
- Host from which the organism is isolated
- Collection date
- Country where isolate was collected from.
- Name, contact email and institution affiliation address of the Principal Investigator submitting the sequence.
- All these go in what the genbank calls the template file. (It can be generated using sequin)

Other Files Required for Submission

- **Sequence file in fasta format**
- **Asn file generated using Sequin OR tbl2asn**
- **If the sequence has gaps, submit an associated gapped fasta file with the estimated number of NNN denoting gaps OR with 100 NNNs in the gap region if the gap is of unknown size.**

SEQUIN

The screenshot shows a web browser window titled "Submitting Authors". The interface includes a menu bar with "File" and "Edit". Below the menu is a navigation bar with four tabs: "Submission" (highlighted with a black box), "Contact", "Authors", and "Affiliation". The main content area asks, "When may we release your sequence record?" and offers two options: "Immediately After Processing" and "Release Date:". Below this is a text input field with the placeholder text "Tentative title for manuscript (required)", which is also highlighted with a black box. At the bottom, there is a button labeled "Click here to import a template" (highlighted with a black box), and two navigation buttons: "<< Prev Form" and "Next Page >>".

Sequin is a tool from NCBI that is used to generate files in the "asn" format which is the required format for submission of sequence to Genbank

SEQUIN

Submission	Contact	Authors	Affiliation
First Name	M.I.	Last Name	Sfx
<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>

Please include country code for non-U.S. phone numbers.

Phone Fax

Email

Submission	Contact	Authors	Affiliation
First Name	M.I.	Last Name	Sfx
<input checked="" type="checkbox"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
<input checked="" type="checkbox"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
<input checked="" type="checkbox"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>

Consortium

The consortium field should be used when a consortium is responsible for the sequencing or publication of the data. Individual authors may be listed along with a consortium name.

SEQUIN

File Edit

Submission Contact Authors Affiliation

Institution

Department

Street Address

City

State/Province Zip/Postal Code

Country

[Click here to export a template](#)

- The template can be saved into a text file which can be reused; therefore not requiring to fill this multiple times.
- There is also a web version for submitting data to NCBI. Each data type has different requirements; links are in the resources.

File Edit Search Special Projects Misc Analysis Annotate

Target Sequence

MSLSP_MSLCDC_00002.MAIN.msl.155

Done

Format

GenBank

Mode

Sequin

Style

Normal

CDS: phosphoprotein

```

LOCUS       MSLSP_MSLCDC_00002.MAIN.msl.15515900 bpss-RNA linear
20-OCT-2011
DEFINITION  Measles virus MVi/Florida.USA/19.09
ACCESSION
VERSION
KEYWORDS
SOURCE      Measles virus MVi/Florida.USA/19.09
ORGANISM    Measles virus MVi/Florida.USA/19.09
            Unclassified.
REFERENCE   1 (bases 1 to 15900)
AUTHORS     Kirkness,E.F., Halpin,R., Bera,J., Fedorova,N., Overton,L.,
            Stockwell,T., Amedeo,P., Bishop,B., Chen,H., Edworthy,P., Gupta,N.,
            Katzel,D., Li,K., Schobel,S., Shrivastava,S., Thovarai,V., Wang,S.,
            Bankamp,B., Byrd,L., Bellini,W. and Rota,P.
TITLE       Direct Submission
JOURNAL     Submitted (20-OCT-2011) J. Craig Venter Institute, 9704 Medical
            Center Drive, Rockville, MD 20850, USA

FEATURES             Location/Qualifiers
     source           1..15900
                     /organism="Measles virus MVi/Florida.USA/19.09"
                     /mol_type="genomic RNA"
                     /strain="MVi/Florida.USA/19.09"
                     /host="human"
                     /country="USA: FLORIDA"
                     /collection_date="2009"
                     /genotype="D4"
     gene             56..1744
                     /gene="N"
     CDS              108..1685
                     /gene="N"
                     /codon_start=1
                     /product="nucleocapsid protein"
                     /translation="MATLLRSLALFKRNKDKPPITSGSGGAIKIKHIIIVPIPQDSS
ITTRSRLLDRLVRLIGNPDVSGPKLTGALIGILSLFVESPQLIQRTDDPDVSIKLL
EVVQSDQSQSGLTFASRGTHMEDEADQYF SHDDPSSGDQSRSGWFENKEISDIEVQDP
EGFNMILGTLIAQIIVLLAKAVTAPDTAADSELRWIKYTKQRRVVEFRLEKWLVDV
VNRRIAEDLSLRRFMVALILDIKRTPGNKPRIAREMICDIDTYIVEAGLASFILTIKFG
IETMYPALGLHEFAGELSTLESLMNLVYQQMGETAPYVMILENSIQNKF SAGSYPLLWS
YAMGVGVLELNSMGGLNFRSYFDPAYFRLGQEMVRRSAGKVSSTLASELGITAEADAR
LVSEIAMHTTEDRI SRVAGPRQAQV SFINGDQSENELPGLGGKEDRRVKQGRGEARES
YKETGSSRASDARAHLPISTPLDVTASESGQDPQDSRRSADALLRLQAMAGILEEQ
GSDTDISRVTYNDKDLLD"

```

Genbank Format

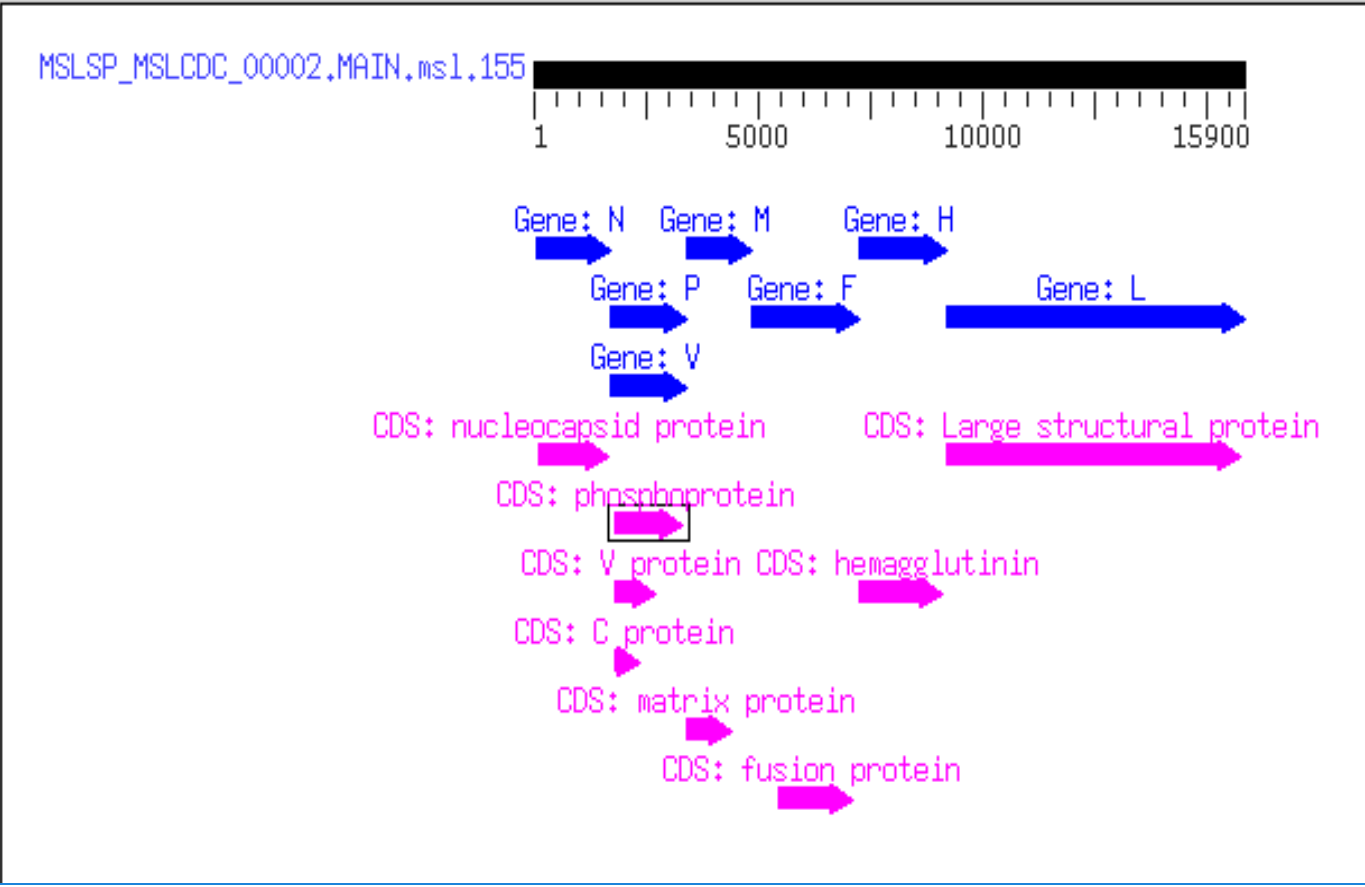
Accessions will be
assigned when
accepted by
Genbank

METADATA

Target Sequence

Format Style Filter Scale

CDS: phosphoprotein



Graphic Format

Target Sequence

Format

CDS: Large structural protein

Feature display:

```

      10      20      30      40      50      60      70
N      1      accaaacaaa attgagtaag gatagatcaa tcaatgatca tattctagtg cacttaagat tcaagatct
N
      80      90      100     110     120     130     140
N      71      attaacaggg acaagagcag gattagggat atccagatg gccacacttt taaggagctt agcattgttc
N
nucleocapsid
nucleocapsid
      M A T L L R S L A L F
      150     160     170     180     190     200     210
N      141      aaaagaaaca aggacaaacc acccattaca tcaagatccg gtgagaccat cagaagaaac aaacacatta
N
nucleocapsid
nucleocapsid
      K R N K D K P P I T S G S G G A I R G I K H I
      220     230     240     250     260     270     280
N      211      ttatagtacc aatecttoga gattctctcaa ttaccactcg atccagacta ctgagaccgt tgcacagct
N
nucleocapsid
nucleocapsid
      I I V P I P G D S S I T T R S R L L D R L V R L
      290     300     310     320     330     340     350
N      281      aattgaaac ccgatctgga gccagcccaa actaacaggg gcactaatag gtatattate cttatttoto
N
nucleocapsid
nucleocapsid
      I G N P D V S G P K L T G A L I G I L S L F V
      360     370     380     390     400     410     420
N      351      gaatctccag gtcaattgat tcaagagatc accgatgacc ctgacottag catcaagctg ttagaagttg
N
nucleocapsid
nucleocapsid
      E S P G Q L I Q R I T D D P D V S I R L L E V
    
```

Alignment Format

Sequin is a very powerful tool for checking the accuracy of annotation

Resources for Genbank Submission

- **BioProject registration**

<http://www.ncbi.nlm.nih.gov/bioproject/>

- **Procedure for submitting data to Genbank**

<http://www.ncbi.nlm.nih.gov/guide/howto/submit-data/>

- **Bacterial Genome Submission Guide**

<http://www.ncbi.nlm.nih.gov/genbank/genomesubmit>

<https://submit.ncbi.nlm.nih.gov/>

- **Consistency Check and Validation tool for microbial genomes**

<http://www.ncbi.nlm.nih.gov/genomes/frameshifts/frameshifts.cgi>

- **Instruction for submitting Whole Genome Shotgun Sequences**

<http://www.ncbi.nlm.nih.gov/genbank/wgs>

- **Sequin**

<http://www.ncbi.nlm.nih.gov/projects/Sequin/>

You have been great for making it to the end!!

- Thanks for attending and staying!
- You won't remember everything you heard today... This workshop was to give you a “taste” of Bioinformatics for Viral Genomics.
- I like e-mail if you have questions later on tstockwell@jcvl.org
- Please consider our summer internships (applications due in March)

Questions?