Instructions for Obtaining a Seed Sequence for Organelle Assembly

We will be using NOVOPlasty (Dierckxsens et al., 2017) to assemble organelle genomes for our species! As we talked about at the end of Week 8 Lab, NOVOPlasty is a seed-extend based assembler. This means that the software starts with a "seed" sequence and attempts to find overlapping reads that can extend the sequence in both directions. The seed is just a string of DNA nucleotides that belong to the organelle genome of a species that is closely related to your sample. The seed sequence does not become part of the assembly, it is just used as a starting point to retrieve the first sequencing read from the raw sequencing read data set. Remember that our sequencing read files will include reads from both nuclear and organellar genomes. We don't want to include reads from the nuclear genome in our organelle assemblies!

The protocol for obtaining seed sequences will differ slightly depending on whether your species is a plant or an animal. For animals, you will only be assembling the mitochondria genome. For plants, you must assemble the chloroplast genome before you can assemble the mitochondrial genome.

Animals
Look for the sequence of a conserved mitochondrial gene in a species closely related to your species! A good place to start would be the cytochrome c oxidase subunit I (COI) gene. See the instructions below to learn how to do this.
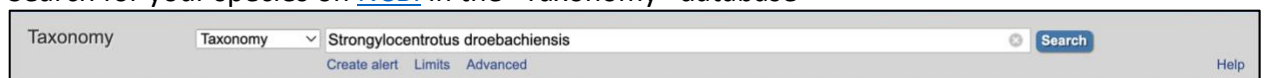
Plants
Look for a chloroplast-specific gene in a closely related species. In plants, the chloroplast and mitochondria have many similar sequences due to intergenomic transfer. It is important to find a sequence that is only found in the chloroplast, and not in the mitochondria. A good place to start would be the RuBisCO enzyme. Data on the RuBisCO enzyme for your species might be under several different synonymous names:
- Ribulose-1,5-bisphosphate carboxylase-oxygenase
- Rubisco
- RuBPCase
- RuBPco

How to Find a Seed Sequence
I will use *Strongylocentrotus droebachiensis*, the green sea urchin, to demonstrate how to follow this protocol.

1. Search for your species on NCBI in the "Taxonomy" database

2. Click on your species

3. Look at the info under "Entrez records". Entrez is a database for molecular sequence data. Is there an assembly available (a row titled "Assembly" with a number greater than zero next to it)? If so, skip to step 4. In the example below, there is not an assembly available for *Strongylocentrotus droebachiensis*.

   Look at the "Lineage" information. We are going to go up the lineage of your organism to hopefully find the most closely related species with an assembled genome. Click the last link in the lineage list (this most likely corresponds to a genus).

Strongylocentrotus droebachiensis

Taxonomy ID: 7671 (for references in articles please use NCBI:txid7671)
┌current name
│ *Strongylocentrotus droebachiensis* (Muller, 1776)
└
Genbank common name: **green sea urchin**
NCBI BLAST name: **sea urchins**
Rank: **species**
Genetic code: Translation table 1 (Standard)
Mitochondrial genetic code: Translation table 9 (Echinoderm Mitochondrial; Flatworm Mitochondrial)
Other names:
┌heterotypic synonym
│ *Strongylocentrotus droebechiensis*
└

Lineage(full)
    cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Echinodermata; Eleutherozoa; Echinozoa; Echinoidea; Euechinoidea; Echinacea;
    Camarodonta; Echinidea; Strongylocentrotidae; Strongylocentrotus

Entrez records	
Database name	Direct links
Nucleotide	509
Protein	520
Genome	1
Popset	23
PubMed Central	281
Gene	41
SRA Experiments	44
Protein Clusters	13
Identical Protein Groups	193
Bio Project	1
Bio Sample	44
Taxonomy	1

   After clicking the link, you will see all of the organisms that fall under that lineage. Click on the genus name one more time. In this case, it is *Strongylocentrotus*.

Lineage (full): cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Echinodermata; Eleutherozoa; Echinozoa; Echinoidea; Euechinoidea; Echinacea; Camarodonta; Echinidea;
Strongylocentrotidae

  ○ **Strongylocentrotus**   *Click on organism name to get more information.*
      ▪ **Strongylocentrotus droebachiensis** (green sea urchin)
      ▪ **Strongylocentrotus intermedius**
      ▪ **Strongylocentrotus pallidus** (pale sea urchin)
      ▪ **Strongylocentrotus polyacanthus**
      ▪ **Strongylocentrotus purpuratus** (purple sea urchin)
      ○ **unclassified Strongylocentrotus**
          ▪ **Strongylocentrotus sp. 17_ECHINO_001_069**
          ▪ **Strongylocentrotus sp. 17_FA_006**
          ▪ **Strongylocentrotus sp. EAC01**
          ▪ **Strongylocentrotus sp. TC-2019**

4. Look at the info under "Entrez" records. Is there an assembly available? If the answer is no, repeat steps 3 and 4 until you get to a taxonomic group that has an assembly available.

    In this case, the Entrez records indicate that there is at least one assembly available! Click on the underlined number next to "Assembly."

Strongylocentrotus

Taxonomy ID: 7664 (for references in articles please use NCBI:txid7664)
current name
Strongylocentrotus

NCBI BLAST name: **sea urchins**
Rank: **genus**
Genetic code: Translation table 1 (Standard)
Mitochondrial genetic code: Translation table 9 (Echinoderm Mitochondrial; Flatworm Mitochondrial)

Lineage(*full*)
cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Echinodermata; Eleutherozoa; Echinozoa; Echinoidea; Euechinoidea; Echinacea; Camarodonta; Echinidea; Strongylocentrotidae

External Information Resources (NCBI LinkOut)

Entrez records		
Database name	Subtree links	Direct links
Nucleotide	304,290	-
Protein	42,296	-
Structure	6	-
Genome	4	-
Popset	82	32
Conserved Domains	1	-
GEO Datasets	679	
PubMed Central	3,666	3,534
Gene	66,362	
SRA Experiments	3,332	-
Protein Clusters	13	-
Identical Protein Groups	63,460	-
Bio Project	94	1
Bio Sample	909	-
Bio Systems	269	-
Assembly	6	-
Probe	99	-
PubChem BioAssay	27	-
Taxonomy	11	1

5. A list of published genome assemblies will appear. Click on one of the listed assemblies. If many options appear, note that you can filter using options on the left side bar.

Organism group
Animals (1)
Customize ...

Status   clear
✓ Latest (1)
Latest GenBank (1)
Latest RefSeq (1)

Assembly level
Complete genome (0)
Chromosome (0)
Scaffold (1)
Contig (0)

RefSeq category
Reference (0)

Summary ▾

ⓘ Filters activated: Latest, Exclude anomalous. Clear all to show 6 items.

Spur_5.0
Organism: Strongylocentrotus purpuratus (purple sea urchin)
Sex: male
Submitter: Baylor College of Medicine
Date: 2019/09/06
Assembly level: Scaffold
Genome representation: full
RefSeq category: representative genome
GenBank assembly accession: GCA_000002235.4 (**latest**)
RefSeq assembly accession: GCF_000002235.5 (**latest**)
IDs: 4536931 [UID] 13654338 [GenBank] 13759658 [RefSeq]

Send to: ▾

Assembly Information ▲
Assembly Help
Assembly Basics
NCBI Assembly Data Model

Related Information ▲
BioProject
BioSample
Genome
Nucleotide INSDC

Shown below is the information for the latest genome assembly of the purple sea urchin (*Strongylocentrotus purpuratus*), Spur_5.0.

6. Scroll down to the bottom of the page to the section titled: "Global assembly definition"

7. Click the "non-nuclear option." For animals, look to see if there is a mitochondria option. For plants, look to see if there is a chloroplast option. Click the underlined link under "GenBank ID." If there are not any choices, go back to step 5 and choose a different assembly. If there are no other assemblies available, repeat steps 3 and 4, following the lineage further back to check for assemblies in distantly related taxonomic groups.

8. Shown below is the information for the assembled *Strongylocentrotus purpuratus* mitochondrial genome. If you scroll down, you will see all the genes in the assembly listed. For animals, try to find a COI gene. For plants, try to find a RuBisCO gene. You might have to skim through the genes or keyword search different terms. For example, the COI gene could be under "cytochrome c" or "COI."

9. Once you've found the gene you want, click on the underlined "gene" or "CDS" link. CDS stands for coding sequence. Clicking this link will show you where your sequence is in the genome assembly.

CDS	5785..7338
	/codon_start=1
	/transl_table=9
	/product="cytochrome oxidase subunit 1"
	/protein_id="CAA31153.1"
	/db_xref="GOA:P15544"
	/db_xref="InterPro:IPR000883"
	/db_xref="UniProtKB/Swiss-Prot:P15544"
	/translation="MQLSRWLFSTNHKDIGTLYLIFGAWAGMVGTAMSVIIRAELAQP
	GSLLNDDQIYNVVVTAHALVMIFFMVMPIMIGGFGNWLIPLMIGAPDMAFPRMNNMSF
	WLIPPSFILLLASAGVENGAGTGWTIYPPLSSNITHAGSSVDLAIFSLHLAGASSILG
	LINFITTIINMRTPGMSLDRLPLFVWSVFVTAFLLLLSLPVLAGAITMLLTDRNINTT
	FFDPAGGGDPILFQHLFWLFGHPEVYILILPGFGMISHVIAHYSGKREPFGYLGLVYA
	MIAIGVLGFLVWAHHMFTVGMDVDTRAYFTAATMIIAVPTGLKVFSWMAKLQGSNLQW
	SLPLLWTLGIVFLFTLGGLTGIVLANSSIDFVLHDTYYVVAHFHYVLSMGAVFAIFAG
	FTHWFPLFSGYSLHPLWGKVHFFIMFVGVNLTFFPQHFLGLAGMPRRYSDYPDAYTLW
	NTISSIGSTISVVAMLFFLFLIWEAFASQREGITPEFSHASLEWQYTSFPPSHHTFDE
	TPSTIIIVK"

10. On the bottom right, click the "Display: FASTA" option. This will take you to a new screen with the selected sequence in FASTA format.

Strongylocentrotus purpuratus complete mitochondrial genome

GenBank: X12631.1

GenBank   Graphics

```
>X12631.1:5785-7338 Strongylocentrotus purpuratus complete mitochondrial genome
ATGCAACTAAGACGATGATTATTTTCTACTAACCACAAGGACATCGGAACACTTTATTTAATTTTTGGGG
CCTGAGCTGGCATGGTAGGCACAGCTATGAGTGTGATTATCCGTGCCGAGTTGGCACAACCTGGTTCGCT
GCTAAAAGATGACCAGATATACAAAGTGGTCGTTACCGCACATGCGCTAGTCATGATTTTCTTCATGGTA
ATGCCAATAATGATTGGTGGATTTGGGAATTGACTCATTCCACTAATGATCGGTGCGCCAGATATGGCCT
TCCCCCGCATGAAAAATATGAGTTTTTGACTTATTCCCCCTTCTTTTATATTACTTTTAGCGTCCGCAGG
AGTAGAAAAAGGAGCAGGAACTGGCTGAACTATCTACCCTCCTCTCTCTAGTAAAATAACACACGCCGGT
AGGTCCGTTGATTTAGCAATCTTCTCCCTTCACCTGGCCGGTGCCTCTTCCATCTTGGGCCTCATTAAAT
TTATAACAACAATTATTAATATGCGGACACCGGGGATGTCTTTGGATCGTCTTCCTTTATTCGTCTGATC
CGTCTTTGTCACTGCCTTCTTGCTCCTCCTTTCTCTTCCAGTATTAGCAGGAGCAATTACAATGCTTCTC
ACAGATCGTAAAATAAACACAACTTTCTTTGACCCTGCAGGAGGGGGAGATCCAATTCTATTTCAACACC
TATTCTGGCTTTTTGGACACCCCGAGGTGTATATTCTTATCTTACCGGGATTTGGTATGATCTCACACGT
TATAGCTCACTACTCTGGTAAGCGAGAGCCTTTCGGATACCTGGGGTTGGTTTATGCCATGATTGCAATA
GGAGTTTTAGGATTCCTTGTCTGGGCCCACCATATGTTTACAGTAGGGATGGATGTTGATACACGAGCAT
ACTTCACTGCCGCCACAATGATTATTGCTGTCCCAACAGGATTAAAGGTTTTCAGATGAATGGCAAAGCT
CCAAGGGTCTAATCTACAATGAAGACTCCCTTTATTATGAACCTTGGGGATTGTATTTTTATTCACATTA
GGAGGACTCACAGGTATTGTTCTTGCCAATTCCTCCATTGACTTTGTTCTTCATGATACCTACTACGTGG
TAGCTCACTTTCACTACGTTCTTTCAATGGGGGCTGTATTTGCAATCTTCGCTGGTTTCACTCACTGGTT
TCCCCTCTTCTCTGGTTATAGCCTACACCCATTATGAGGAAAGGTTCACTTCTTCATAATGTTTGTTGGA
GTCAACTTAACCTTTTTCCCTCAACACTTCTTAGGTCTAGCCGGAATGCCACGACGGTACTCAGACTATC
CAGACGCCTATACACTTTGAAATACTATCTCCTCAATTGGATCAACCATCTCCGTAGTGGCTATGCTATT
TTTCCTCTTCTTAATCTGGGAGGCCTTCGCTTCTCAACGGGAAGGAATCACCCCAGAGTTCTCACACGCC
TCACTAGAGTGACAATACACCTCCTTTCCCCCTTCTCACCACACCTTCGATGAAACACCCTCTACCATAA
TTATTGTAAAGTAA
```

11. Copy and paste the sequence into a text editor (either Sublime or Atom, do not use TextEdit). **Make sure to include the header that beings with the ">" symbol.**

12. Save this file as seed_mito.fasta or seed_chloro.fasta. Remember where you saved it!