COMMUNITY SCIENCE PROJECT WORKFLOW (with example)

Introduction

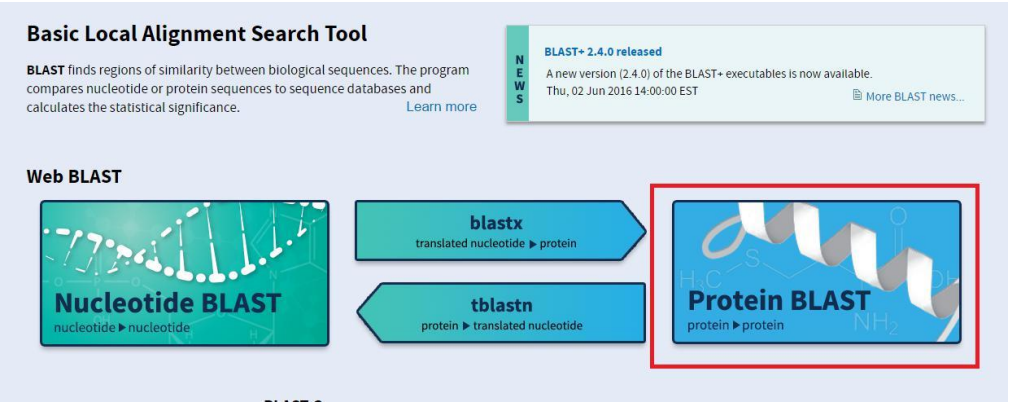
How does exchange of genetic material between bacteriophages and bacteria drive bacterial evolution? This is the question we're attempting to answer with the Community Science Project. You can help us gather data to address this question.

We will start with a phage called Fen4701_41 with Accession Number: **NC_027641**.

One of the predicted proteins in the database for Fen4704_41 is the **Major capsid protein VP1.** This phage has a small number of genes that are predicted to encode proteins, but our previous work has shown that only VP1 has homologs in bacteria, so that's the only one we'll investigate here.

```
major capsid protein VP1"
                     /protein_id="YP_009160373.1"
                     /db_xref="GI:906476351"
                     /db_xref="GeneID:25102737"
                     /translation="MGLFETNQAPRVPRAKFDLSHDRKLTMNMGTLIPVLNKEVLPGD
                     VWRCSPTVFLRFLALLAPIMHKVNVKLEVFFCAYRLVWPNFPDFVSGGESGTAAPVFP
                     TFNTTQCNVTNTTILQDGSLLDYLGFPTYPQSGTYNAAYNGNTFSILHPRVYTLIYDT
                     WYRNQNVETTVIGTTFGADGPQNAGDYPFIFSLRSRQWERDYFTSCQPSTQRGTQVVI
                     PISSTITSPSGNTNIPIVRVSGGAANTVSHAIGSKAADGSLMDQTSLAGVYLDPNGSL
                     IASNTSLTVNALRASITLQTYKERMQRGGGRYIEYLWNVFGVASDDARLQRPELIYTL
                     REPVTVSEVLTTSTGTGTGQTPPGTLAGAAMSIGSGGSGVYAAKEHGCIMALMSVIPR
                     SAYQNGIGPEFWRQVNTDFYVPDFAHLGEQTVIQQNVYYDMTNAAQASNGGVVFGYQS
                     RFAEYKYANDQVSGQFKSTLGFWHLGRQFNSTSVPALNLAFCQCVPRKDIFAVTTDVD
                     HIVVDIYWNISVLRPMPYFGTPSTLNAVS"
```

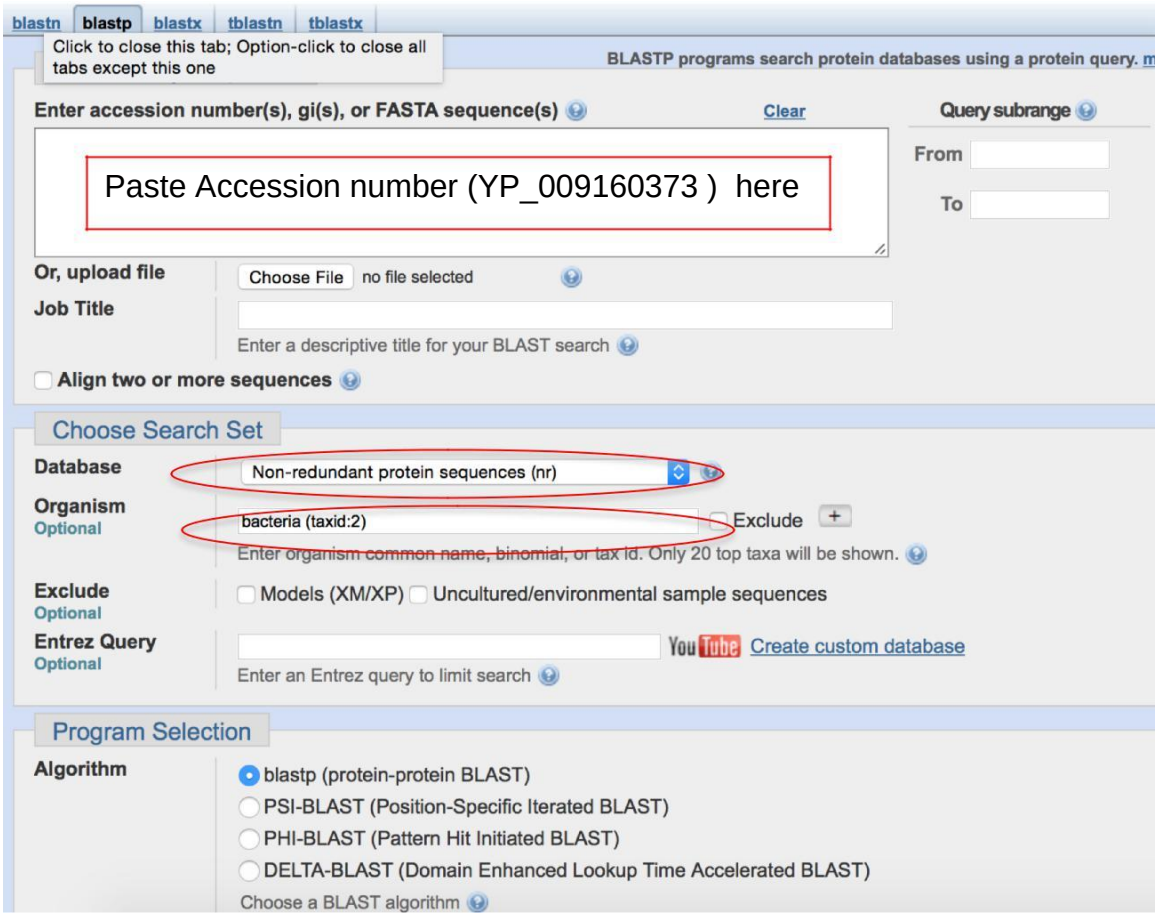
Part 1: Forward BLAST – looking for bacterial homologs of bacteriophage proteins



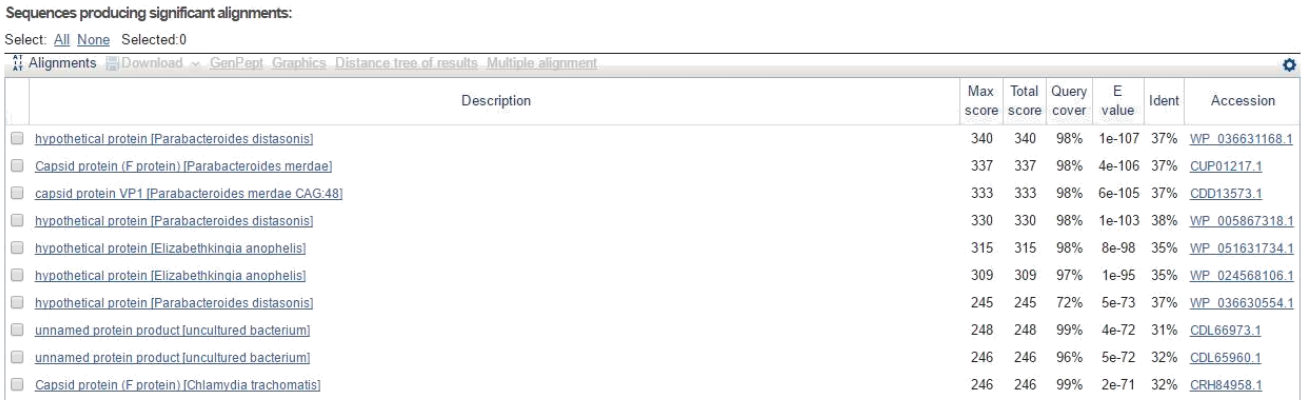
1. BLAST the Major capsid protein VP1 with Accession ID YP_009160373 in the protein database

2. Use <blastp>, setting the parameters to search the <nr database> but only against <bacteria>. This way your top hits will be bacteria, not other bacteriophages.

   Use the default settings for blastp.



3. The results of this BLAST will take you to the “Hits” page. Record the **top 10 hits** in an Excel spreadsheet **(provided on the website)** to enter into the Community Science Project database. Make sure to record the hits based on the following cut-off values:



Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

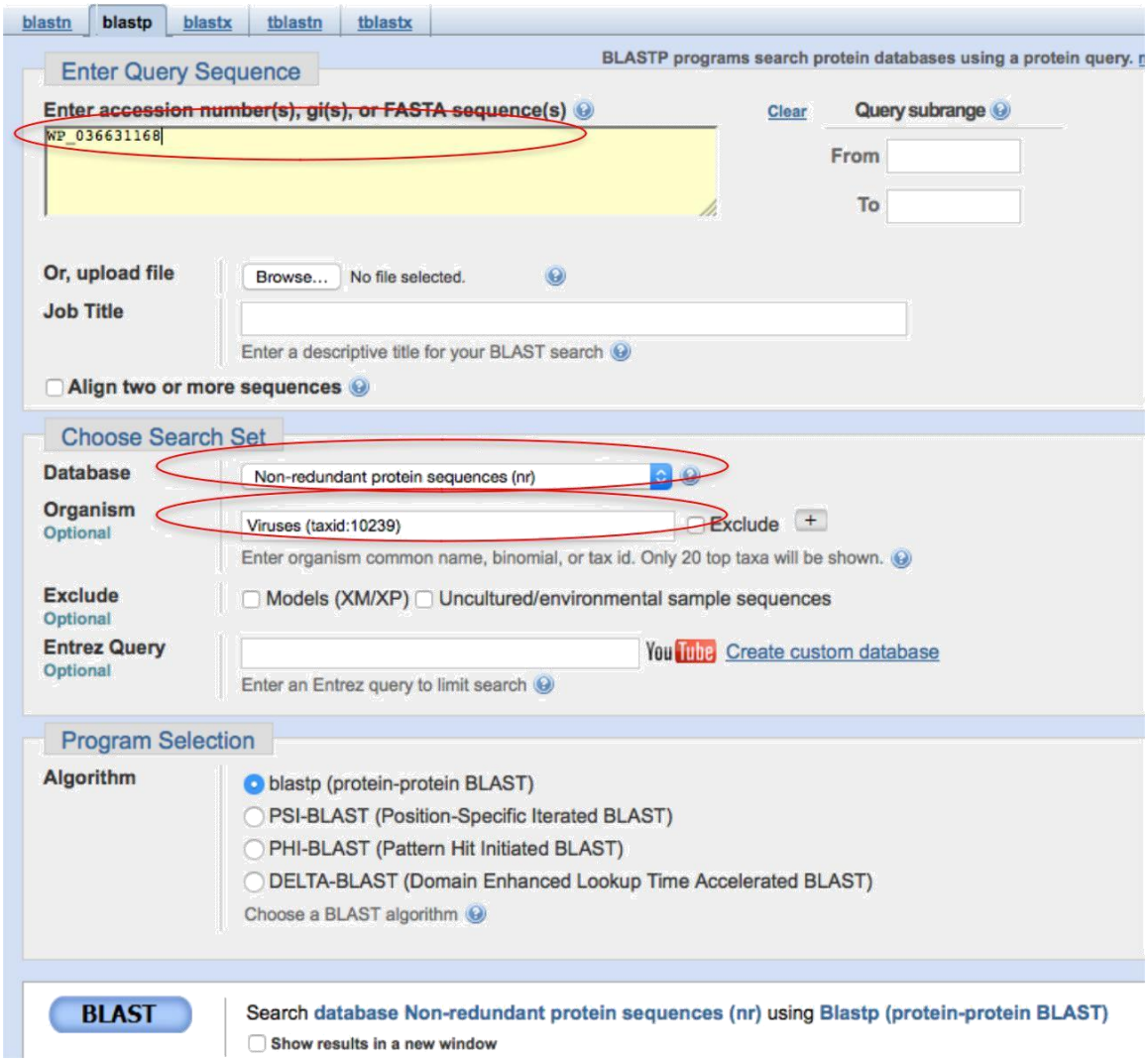
Description	Max score	Total score	Query cover	E value	Ident	Accession
hypothetical protein [Parabacteroides distasonis]	340	340	98%	1e-107	37%	WP_036631168.1
Capsid protein (F protein) [Parabacteroides merdae]	337	337	98%	4e-106	37%	CUP01217.1
capsid protein VP1 [Parabacteroides merdae CAG:48]	333	333	98%	6e-105	37%	CDD13573.1
hypothetical protein [Parabacteroides distasonis]	330	330	98%	1e-103	38%	WP_005867318.1
hypothetical protein [Elizabethkingia anophelis]	315	315	98%	8e-98	35%	WP_051631734.1
hypothetical protein [Elizabethkingia anophelis]	309	309	97%	1e-95	35%	WP_024568106.1
hypothetical protein [Parabacteroides distasonis]	245	245	72%	5e-73	37%	WP_036630554.1
unnamed protein product [uncultured bacterium]	248	248	99%	4e-72	31%	CDL66973.1
unnamed protein product [uncultured bacterium]	246	246	96%	5e-72	32%	CDL65960.1
Capsid protein (F protein) [Chlamydia trachomatis]	246	246	99%	2e-71	32%	CRH84958.1

We're looking for things that are relatively closely related, therefore use **> 70% query cover** and **< 1e-50 E-value** as the cutoffs for this. If there are fewer than 10 that match these criteria, just capture these in the spreadsheet. If there are cogent reasons to change these cutoffs depending on the system you're looking at, capture this information as well.

Part 2: Reverse BLAST – looking for bacteriophage homologs of bacterial proteins

Now take the top hit from the Forward BLAST result, which in this case is:
hypothetical protein [Parabacteroides distasonis]
Accession number: WP_036631168.

1. Perform another BLAST using <blastp> but this time we will be looking for phage proteins. There is no "Bacteriophage" setting, so we'll use the <nr> database and restrict our results to the <virus> sector of the database.



2. The results of the REVERSE BLAST will take you to the “Hits” page and again you will record the top 10 hits **(spreadsheet provided on the website).** Use the same cutoffs (normally, > 70% query cover and < 1e-50 E-value) as the you did for the forward BLAST.

Part 3: Learning about the protein universe in which the protein of interest resides

Now check to see if your Top Hit from the REVERSE Blast matches the initial phage protein you had searched for.

In our search, the Reverse BLAST top hit is: major capsid protein VP1 [Parabacteroides phage YZ-2015a] with Accession Number: **YP_009218533**

This result matches with the initial phage protein we searched for -

major capsid protein VP1 [Microviridae Fen4707_41] with Accession Number: **YP_009160373**

This result conveys that the two proteins are each other's closest homologs. This isn't always the case and suggests that there are other homologs that could be explored. This information can be recorded on the Reverse Blast spreadsheet provided on the website.

Part 4: **Submitting your work to Genome Solver**

If you want to take part in the Community Science Project, the information in your two spreadsheets should be submitted to the Community Science Project pages on the Genome Solver Qubeshub website (https://qubeshub.org/community/groups/genomesolver).