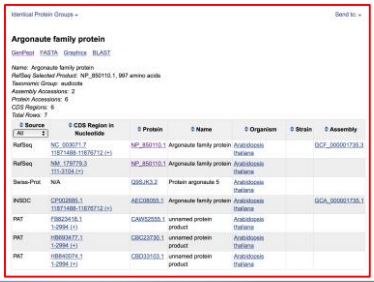
Cool New Stuff!

NCBI Resources for Phylogenetically-Defined Next Generation Analysis in and out of the Cloud

Ben Busby (ben.busby@nih.gov)
NCBI Genome Resources Workshop
PAG XXVII January 14, 2019

NIH U.S. National Library of Medicine
National Center for Biotechnology Information



Tom Madden's BLAST_v5 Talk (slightly modified)

If you'd like to see the original slides from Tom, or this slide deck, check out https://www.slideshare.net/benbusby

Search the Argonaute protein family with BLAST!

- Against the nr protein database
- Limit the search to Schizosaccharomyces pombe

BLASTDBv5: a better BLAST database!

- Limit search by taxonomy with a command-line parameter

  -taxids

- Improved performance when limiting BLAST search with accessions.
  - **Faster!**
- Retrieve sequences by taxonomy from the BLAST database.
  - Extract data!


```
Database: All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF
excluding environmental samples from WGS projects
           148,264,992 sequences; 54,322,627,498 total letters


Query= NP_850110.1 Argonaute family protein [Arabidopsis thaliana]

Length=997
                                                                      Score     E
Sequences producing significant alignments:                          (Bits)  Value

NP_587782.1  argonaute [Schizosaccharomyces pombe]                    407     2e-128


>NP_587782.1 argonaute [Schizosaccharomyces pombe]
 CAA19275.1 argonaute [Schizosaccharomyces pombe]
Length=834

 Score = 407 bits (1047),  Expect = 2e-128, Method: Compositional matrix adjust.
 Identities = 284/892 (32%), Positives = 449/892 (50%), Gaps = 91/892 (10%)

Query  136  SLPPASSKAVTFPVRPGRGTLGKKVMVRANHF-LVQVADRDLYHYDVSINPEVISKTVNR  194
            S  P+S  A+     RPG G LGK++ ++AN F ++ + +  +  Y V +       V R
Sbjct  2    SYKPSSEIAL----RPGYGGLGKQITLKANFFQIISLPNETINQYHVIVGD----GSRVPR  54

Query  195  NVMKLLVKNVKDSHLGGKS----PAYDGRKSLYTAGPLPFDSKEFVVNLAEKRADGSSGKD  251
            +L+  + +     G S     YDGR   ++ G +    + VN+      GS
Sbjct  55   KQSQLIWNSKEVKQYFGSSWMNSVYDGRSMCWSKGDIADGTIK---VNI------GSESHP  106
```

BLAST command-line parameters

```
blastp
  -db nr
    -query NP_850110.fsa
      -outfmt "7 qaccver saccver pident length evalue sscinames"  # tabular output
        -out NP_850110.4896.tab
          -task blastp-fast                    # runs faster!
            -num_threads 8                     # use 8 CPUs
              -evalue 0.05                     # limit to more significant matches
                -taxids 4896                   # limit search to Schizosaccharomyces pombe
```


Need taxids for a whole Phylum?

```
bash-4.2$ get_species_taxids.sh -n Ascomycota

Taxid: 4890
 rank: phylum
 division: ascomycetes
 scientific name: Ascomycota
 common name: ascomycetes

1 match(es) found.

bash-4.2$ get_species_taxids.sh -t 4890 > Ascomycota.taxidlist
bash-4.2$ head Ascomycota.taxidlist
4754
4896
4897
4899
4903
4909
4911
4914
```


A script to find taxid

Distributed with the BLAST package!

```
bash-4.2$ get_species_taxids.sh -n "Schizosaccharomyces pombe"

Taxid: 4896
 rank: species
 division: ascomycetes
 scientific name: Schizosaccharomyces pombe
 common name: fission yeast

1 match(es) found.
```


BLAST against Ascomycota sequences!

```
blastp
  -db nr
    -query NP_850110.fsa
      -outfmt "7 qaccver saccver pident length evalue sscinames"  # tabular output
        -out NP_850110.Ascomycota.tab
          -task blastp-fast                    # runs faster!
            -num_threads 8                     # use 8 CPUs
              -evalue 0.05                     # limit to more significant matches
                -taxidlist Ascomycota.taxids   # limit search to Ascomycota
```


Tabular output

```
# BLASTP 2.8.0+
# Query: NP_850110.1 Argonaute family protein [Arabidopsis thaliana]
# Database: nr
# Fields: query acc.ver, subject acc.ver, % identity, alignment length, evalue, subject sci name
# 1 hits found
NP_850110.1     NP_587782.1     31.839  892     1.57e-128       Schizosaccharomyces pombe
# BLAST processed 1 queries
```


Ascomycota results

```
# BLASTP 2.8.0+
# Query: NP_850110.1 Argonaute family protein [Arabidopsis thaliana]
# Database: nr
# Fields: query acc.ver, subject acc.ver, % identity, alignment length, evalue, subject sci name
# 501 hits found
NP_850110.1     GAO47878.1      36.139  891     1.93e-164       Saitoella complicata NRRL Y-17804
NP_850110.1     XP_019027272.1  36.550  829     1.39e-156       Saitoella complicata NRRL Y-17804
NP_850110.1     ODQ75150.1      34.220  865     1.03e-146       Lipomyces starkeyi NRRL Y-11557
NP_850110.1     CUS09702.1      31.767  894     6.68e-134       Tuber aestivum
NP_850110.1     XP_002839547.1  30.624  898     3.85e-130       Tuber melanosporum Mel28
NP_850110.1     XP_013025347.1  32.540  882     1.01e-127       Schizosaccharomyces cryophilus OY26
NP_850110.1     ORYB1937.1      32.993  882     2.93e-127       Protomyces lactucaedebilis
NP_850110.1     NP_587782.1     31.839  892     1.45e-125       Schizosaccharomyces pombe
NP_850110.1     XP_013019123.1  33.333  852     5.49e-125       Schizosaccharomyces octosporus yFS286
NP_850110.1     EPS44580.1      30.710  902     7.32e-125       Dactylellina haptotyla CBS 200.50
NP_850110.1     CUS14138.1      33.684  855     1.87e-124       Tuber aestivum
NP_850110.1     XP_011120241.1  31.180  901     2.47e-124       Arthrobotrys oligospora ATCC 24927
NP_850110.1     EWC46593.1      31.823  905     1.41e-122       Drechslerella stenobrocha 248
NP_850110.1     EWC45585.1      31.042  931     3.45e-120       Drechslerella stenobrocha 248
NP_850110.1     KFZ00212.1      30.444  969     1.22e-116       Pseudogymnoascus pannorum VKM F-4518 (FW-2643)
NP_850110.1     XP_002173826.1  31.818  858     4.95e-116       Schizosaccharomyces japonicus yFS275
NP_850110.1     XP_011120792.1  31.973  882     8.11e-116       Arthrobotrys oligospora ATCC 24927
NP_850110.1     KFZ13012.1      30.341  969     9.18e-116       Pseudogymnoascus pannorum VKM F-4520 (FW-2644)
NP_850110.1     XP_011119032.1  31.585  877     2.56e-114       Arthrobotrys oligospora ATCC 24927
NP_850110.1     KFY31959.1      30.714  967     1.34e-112       Pseudogymnoascus pannorum VKM F-4281 (FW-2241)
NP_850110.1     EPS43673.1      29.814  966     1.75e-111       Dactylellina haptotyla CBS 200.50
```

BLAST excluding Viridiplantae

```
blastp
   -db nr
      -query NP_850110.fsa
         -outfmt "7 qaccver saccver pident length evalue sscinamе"
            -out NP_850110.noGreenPlants.tab
               -task blastp-fast
                  -num_threads 8
                     -evalue 0.05
                        -negative_taxidlist Viridiplantae.taxids          # exclude green plants
```


Everything

```
# BLASTP 2.8.0+
# Query: NP_850110.1 Argonaute family protein [Arabidopsis thaliana]
# Database: nr
# Fields: query acc.ver, subject acc.ver, % identity, alignment length, evalue, subject sci name
# 521 hits found
NP_850110.1     NP_850110.1     100.000 997     0.0     Arabidopsis thaliana
NP_850110.1     OAP08078.1      99.900  997     0.0     Arabidopsis thaliana
NP_850110.1     AAD21514.1      99.900  997     0.0     Arabidopsis thaliana
NP_850110.1     XP_002880962.1  89.634  984     0.0     Arabidopsis lyrata subsp. lyrata
NP_850110.1     XP_010473185.1  82.513  995     0.0     Camelina sativa
NP_850110.1     XP_010417939.1  81.590  994     0.0     Camelina sativa
NP_850110.1     XP_019101467.1  81.894  961     0.0     Camelina sativa
NP_850110.1     XP_010510717.1  81.894  961     0.0     Camelina sativa
NP_850110.1     XP_006296291.1  77.756  1016    0.0     Capsella rubella
NP_850110.1     XP_010435959.1  74.926  1009    0.0     Camelina sativa
NP_850110.1     XP_013692158.1  75.540  973     0.0     Brassica napus
NP_850110.1     XP_013636559.1  75.540  973     0.0     Brassica oleracea var. oleracea
NP_850110.1     XP_009103628.1  74.795  976     0.0     Brassica rapa
NP_850110.1     KFK44936.1      73.953  979     0.0     Arabis alpina
NP_850110.1     XP_013647627.1  74.693  976     0.0     Brassica napus
NP_850110.1     XP_018440495.1  79.402  869     0.0     Raphanus sativus
NP_850110.1     XP_006409803.1  72.700  989     0.0     Eutrema salsugineum
NP_850110.1     XP_010490899.1  72.819  986     0.0     Camelina sativa
NP_850110.1     XP_010521702.1  67.306  991     0.0     Cleome hassleriana
NP_850110.1     XP_021610669.1  65.119  926     0.0     Manihot esculenta
NP_850110.1     PNT55850.1      67.086  875     0.0     Populus trichocarpa
```


No Viridiplantae

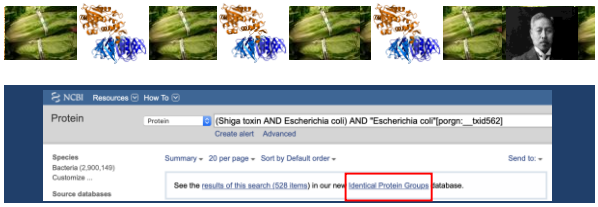
```
# BLASTP 2.8.0+
# Query: NP_850110.1 Argonaute family protein [Arabidopsis thaliana]
# Database: nr
# Fields: query acc.ver, subject acc.ver, % identity, alignment length, evalue, subject sci name
# 500 hits found
NP_850110.1     XP_021356153.1  42.081  884     0.0     Mizuhopecten yessoensis
NP_850110.1     OWF49186.1      41.968  884     0.0     Mizuhopecten yessoensis
NP_850110.1     XP_021356154.1  41.968  884     0.0     Mizuhopecten yessoensis
NP_850110.1     XP_021356147.1  41.742  884     0.0     Mizuhopecten yessoensis
NP_850110.1     XP_021965366.1  41.619  877     0.0     Folsomia candida
NP_850110.1     ODN00002.1      41.686  878     0.0     Orchesella cincta
NP_850110.1     OXA40521.1      41.733  877     0.0     Folsomia candida
NP_850110.1     XP_014666293.1  42.640  856     0.0     Priapulus caudatus
NP_850110.1     XP_002413188.1  42.272  854     0.0     Ixodes scapularis
NP_850110.1     AMN88363.1      40.765  915     0.0     Mytilus galloprovincialis
NP_850110.1     XP_014666292.1  42.740  854     0.0     Priapulus caudatus
NP_850110.1     XP_021965365.1  41.563  883     0.0     Folsomia candida
NP_850110.1     XP_021965367.1  41.563  883     0.0     Folsomia candida
NP_850110.1     XP_002045323.1  40.761  893     0.0     Drosophila sechellia
NP_850110.1     ELT99790.1      41.344  878     0.0     Capitella teleta
NP_850110.1     XP_012170889.1  40.252  954     0.0     Bombus terrestris
NP_850110.1     XP_021356148.1  41.290  884     0.0     Mizuhopecten yessoensis
NP_850110.1     XP_021356151.1  41.516  884     0.0     Mizuhopecten yessoensis
NP_850110.1     KRF99116.1      40.400  901     0.0     Drosophila willistoni
NP_850110.1     XP_015033121.1  40.400  901     0.0     Drosophila willistoni
NP_850110.1     XP_002068762.1  40.400  901     0.0     Drosophila willistoni
```


Timings

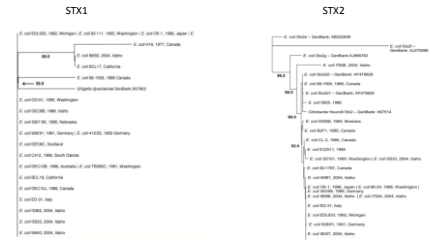
Search set	Run time (seconds)
Schizosaccharomyces pombe	8.7
Ascomycota	11.7
No Viridiplantae	101
Everything	91

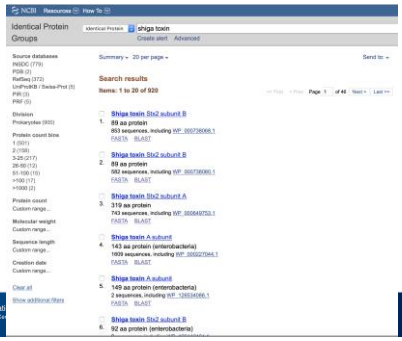
8 CPUs, best of three runs using time command.


```
                              Tax BLAST report

Query= NP_850110.1 Argonaute family protein [Arabidopsis thaliana]

Length=997

                              Organism Report

Accession                       Description                                        Score E-value

                 Mizuhopecten yessoensis (Yesso scallop) [bivalves]
 XP_021356153 protein argonaute-2-like isoform X6 [Mizuhopecten yessoensis]         668   0.0
 OWF49186     Protein argonaute-2 [Mizuhopecten yessoensis]                         666   0.0
 XP_021356154 protein argonaute-2-like isoform X7 [Mizuhopecten yessoensis]         665   0.0
 XP_021356147 protein argonaute-2-like isoform X2 [Mizuhopecten yessoensis]         665   0.0
 XP_021356148 protein argonaute-2-like isoform X3 [Mizuhopecten yessoensis]         654   0.0
 XP_021356151 protein argonaute-2-like isoform X5 [Mizuhopecten yessoensis]         654   0.0
 XP_021356158 protein argonaute-2-like isoform X11 [Mizuhopecten yessoensis]        652   0.0
 XP_021356150 protein argonaute-2-like isoform X4 [Mizuhopecten yessoensis]         652   0.0
 XP_021356157 protein argonaute-2-like isoform X10 [Mizuhopecten yessoensis]        652   0.0
 XP_021356156 protein argonaute-2-like isoform X9 [Mizuhopecten yessoensis]         652   0.0
 XP_021356155 protein argonaute-2-like isoform X8 [Mizuhopecten yessoensis]         652   0.0
 XP_021356146 protein argonaute-2-like isoform X1 [Mizuhopecten yessoensis]         652   0.0
 XP_021356149 protein argonaute-2-like isoform X1 [Mizuhopecten yessoensis]         652   0.0

                       Folsomia candida  [springtails]
 XP_021965366 protein argonaute-2-like isoform X2 [Folsomia candida]                661   0.0
 OXA40521     Protein argonaute-2 [Folsomia candida]                                660   0.0
 XP_021965365 protein argonaute-2-like isoform X1 [Folsomia candida]                657   0.0
 XP_021965367 protein argonaute-2-like isoform X3 [Folsomia candida]                657   0.0

                      Orchesella cincta  [springtails]
 ODN00002     Protein argonaute-2 [Orchesella cincta]                               660   0.0

                      Priapulus caudatus  [priapulids]
 XP_014666293 PREDICTED: protein argonaute-2-like isoform X6 [Priapulus caudatus]   659   0.0
 XP_014666292 PREDICTED: protein argonaute-2-like isoform X5 [Priapulus caudatus]   658   0.0
```

NIH U.S. National Library of Medicine
National Center for Biotechnology Information

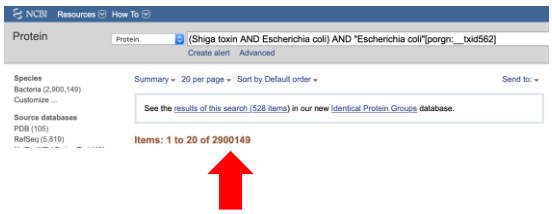




Search the Shiga toxin family protein with BLAST

- Against the nr protein database
- Limit the search to E coli

Why Use Identical Protein Groups?



Search the Shiga toxin family protein with BLAST

- Against the nr pr…base
- Limit the searc…to E coli

But *which* Shiga toxin?!

Stx1 and Stx2 Example Sequences

But what did I need to do to make my demo?

Grab sequences!

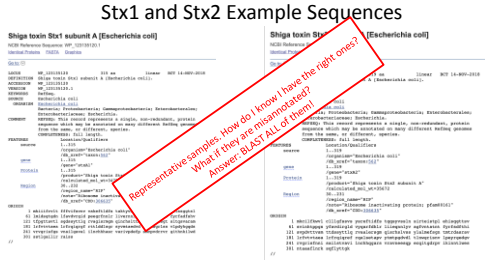
```
ben.busby@known-blast-workbench-sbx-camacho:~$ esearch -db protein -query WP_123135120 | efetch -format fasta
>WP_123135120.1 Shiga toxin Stx1 subunit A [Escherichia coli]
MKIIIFRVLTFFFVIFSVNVVAKEFTLDFSTAKTYVDSLNVIRSAIGTPLQTISSGGTSLLMIDSGTGDN
LFAVDVRGIDPEEGRFNNLRLIVERNNLYVTGFVNRTNNVFYRFADFSHVTFPGTTAVTLSGDSSYTTLQ
RVAGISRMGMQINRHSLTTSYLDLMSHSGTSLTQSVARAMLRFVTVTAEALRFRQIQRGFRTTLDDLSGR
SYVMTAEDVDLTLNWGRLSSVLPDYHGQDSVRVGRISFGSVNAILGSVALILNCHHHASRVARIVPDEFP
SMCPVDRRVRGITHNKILWDSSTLGAILIRRAISS
```

Do I need to get all of nr?!

update_blast.pl -- decompress works

But theres a better way!

Stx1 and Stx2 Example Sequences



CloudBLAST and Docker Images to the Rescue!

- http://ncbi.github.io/blast-cloud/
- https://github.com/ncbi/makeblastdb4cloud
- Also, check out magicBLAST!

BLAST command-line parameters

```
blastp
   -db refseq-protein_v5
      -query  WP_123135120_123130813.fasta
         -outfmt "7 qaccver saccver pident length evalue sscinname"   # tabular output
            -out Shiga_Ecoli.tab
               -task blastp-fast                                      # runs faster!
                  -num_threads 32                                     # use 8 CPUs
                     -evalue 0.00005                                  # limit to more significant matches
                        -taxids   562                                 # limit search to Escherichia coli
```


But how do I use the BLAST Docker(s)?!

https://github.com/ncbi/docker/blob/master/blast/README.md

But how do I use the BLAST Docker(s)?!

https://github.com/ncbi/docker/blob/master/blast/README.md

Interactive BLAST

One can login to the container and run commands inside the container if multiple BLAST runs will be executed interactively.

```
docker run --rm -it \
  -v $BLASTDB_DIR:/blast/blastdb:ro -v $HOME/blastdb_custom:/blast/blastdb_custom:ro \
  -v $HOME/queries:/blast/queries:ro \
  -v $HOME/results:/blast/results:rw \
  ncbi/blast \
  /bin/bash
```

This will open a login shell in the container and one can run BLAST+ as if it was locally installed.

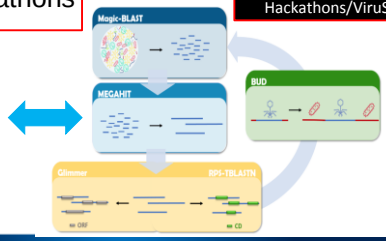
Scripted BLAST

One approach to deal with this situation is to start the blast container in detached mode and execute commands on it.

```
# Start a container named "blast" in detached mode
docker run --rm -dit --name blast \
  -v $BLASTDB_DIR:/blast/blastdb:ro -v $HOME/blastdb_custom:/blast/blastdb_custom:ro \
  -v $HOME/queries:/blast/queries:ro \
  -v $HOME/results:/blast/results:rw \
  ncbi/blast \
  sleep infinity
```


Hackathons

https://github.com/NCBI-Hackathons/ViruSpy



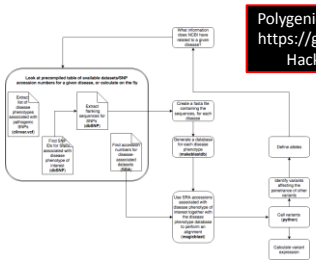
But how do I use the BLAST Docker(s)?!

https://github.com/ncbi/docker/blob/master/blast/README.md

sudo docker run --rm -v $BLASTDB_DIR:/blast/blastdb:ro -v $HOME/queries:/blast/queries:rw -w /blast/blastdb ncbi/blast blastp -db nr_v5 -query /blast/queries/WP_123135120_123130813.fasta -out /blast/queries/Shiga_Ecoli.tab -task blastp-fast -num_threads 32 -evalue .00005 -taxids 562

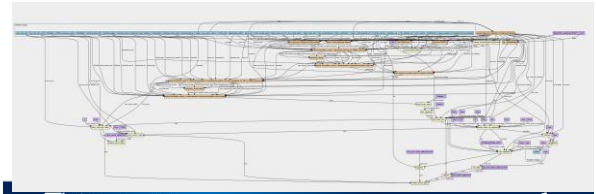
Variant calling on the fly!

Polygenic SNP Search Tool
https://github.com/NCBI-Hackathons/PSST



Bonus: Prokaryotic Genome Annotation Pipeline Wrapped with Docker and CWL

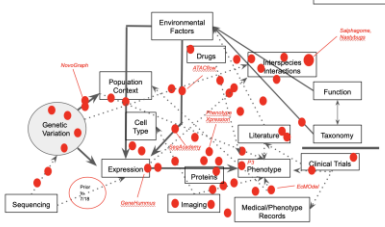
https://github.com/ncbi/pgap



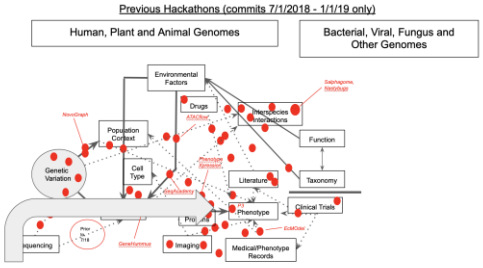
Previous Hackathons (commits 7/1/2018 - 1/1/19 only)

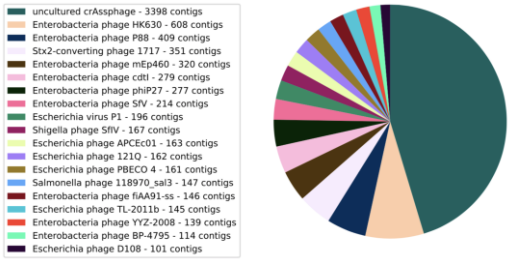
Human, Plant and Animal Genomes

Bacterial, Viral, Fungus and Other Genomes

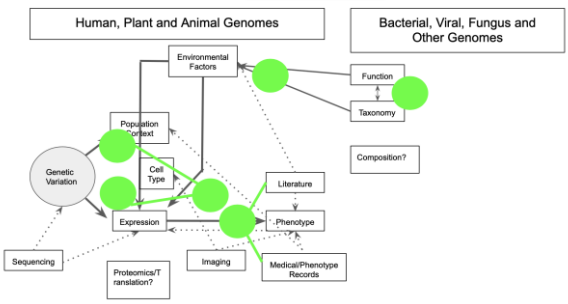


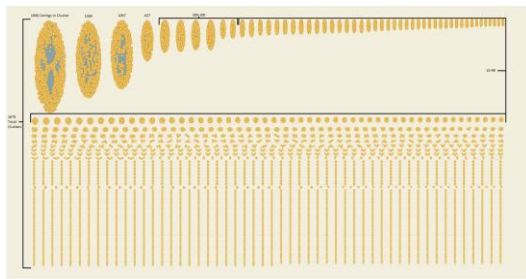
Previous Hackathons (commits 7/1/2018 - 1/1/19 only)

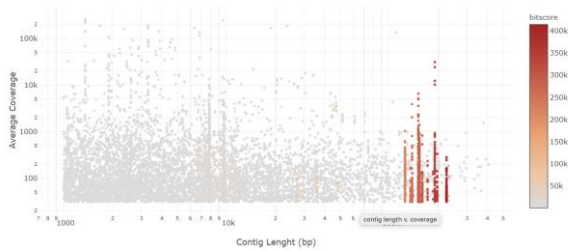




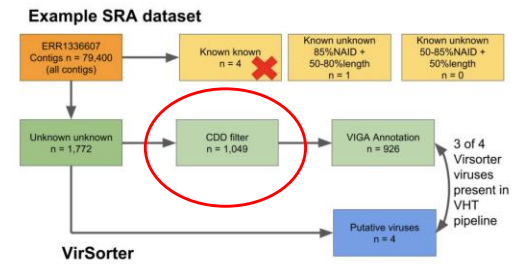
"STRIDES" Hackathons!

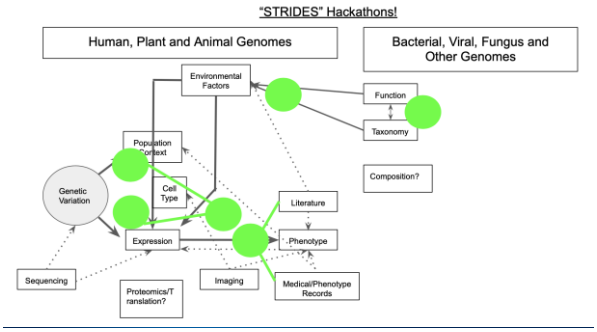


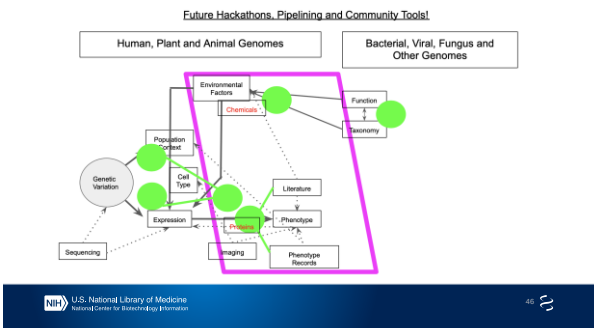




Example SRA dataset







44

Data Analysis Hackathons!

NCBI to assist in Virus Hunting Data Science Hackathon January 9-11, 2019

Planned Events

RNAseq – UNC, March 11-13 2019; Palo Alto, June 2019
Smaller events at USF and Pittsburgh

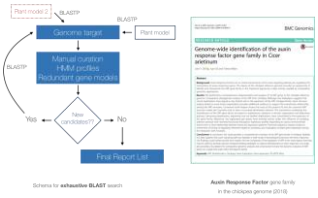
Human PanGenomics – UCSC, 25-27 March 2019;
Haplotype Annotation, Baltimore, June 2019

Variant Synthetic Dataset Generation – Boston, April 15-16, 2019

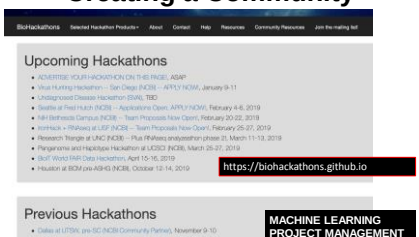
AMR and Prokaryotic Genome Annotate-athon
NIH Campus, August 2019

GeneHummus

Automated Gene Family Discovery!



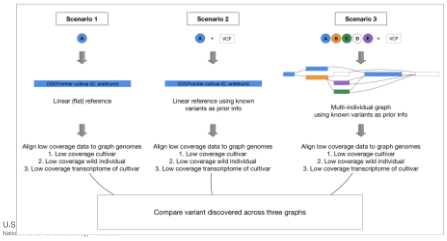
Creating a Community



MACHINE LEARNING
PROJECT MANAGEMENT
PRODUCT MANAGEMENT

HummusGraph

Because graph genomes aren't just for humans!



Thank you.

This research was supported by the Intramural Research Program of the NIH, National Library of Medicine.

Jan Buchmann
Chaochih Liu
Tom Madden
Moamen Elmassry
Kim LeBlanc
Allissa Dillman
Olaitan Awe
Anjana Raina
Peter Meric
Francoise Thibaud-Nissen
Douglas Slotta
Michael Crusoe

Hackathon Participants!

And visiting bioinformaticians!

Eric Cox
Vamsi Kodali
Brian Smith-White
Bart Trawick
Kim Pruitt

U.S. National Library of Medicine
National Center for Biotechnology Information

Watch NCBI News for updates!

http://www.ncbi.nlm.nih.gov/news/
https://www.youtube.com/user/NCBINLM

49

NCBI Genome Resources Workshop

Time	Topic
12:50 – 1:10	Submission of Genomes to GenBank *Karen Clark*
1:10 – 1:30	GEO Submissions and Usage *Steve Wilhite*
1:30 – 1:55	From Annotation to Visualization: Exploring Genes and Genomes with NCBI Tools *Eric Cox*
1:55 – 2:15	Programmatic Access to Genomic Data: E-Utilities and FTP *Vamsi K. Kodali*
2:15 – 2:35	NCBI Resources for Phyletically-Defined Next Generation Analysis in and out of the Cloud (a.k.a. Cool New Stuff!) *Ben Busby*
2:35 – 3:00	Q & A session

U.S. National Library of Medicine
National Center for Biotechnology Information

Visit NCBI Booth **223**

Contact us **info@ncbi.nlm.nih.gov**