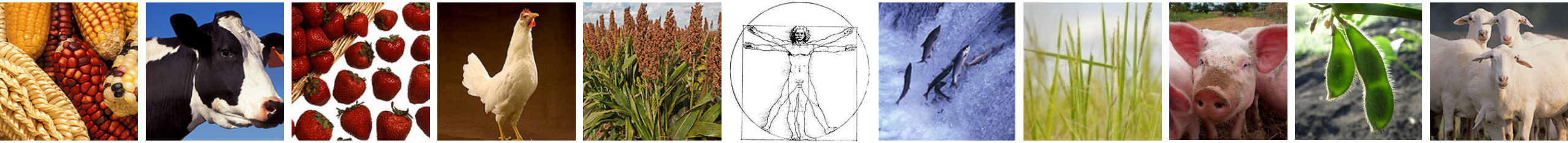


Cool New Stuff!

NCBI Resources for Phyletically-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

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

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!

Identical Protein Groups

Send to:

Argonaute family protein

GenPept FASTA Graphics BLAST

Name: Argonaute family protein
RefSeq Selected Product: NP_850110.1, 997 amino acids
Taxonomic Group: eudicots
Assembly Accessions: 2
Protein Accessions: 6
CDS Regions: 6
Total Rows: 7

Source All	CDS Region in Nucleotide	Protein	Name	Organism	Strain	Assembly
RefSeq	NC_003071.7 11871488-11876712 (+)	NP_850110.1	Argonaute family protein	Arabidopsis thaliana		GCF_000001735.3
RefSeq	NM_179779.3 111-3104 (+)	NP_850110.1	Argonaute family protein	Arabidopsis thaliana		
Swiss-Prot	N/A	Q9SJK3.2	Protein argonaute 5	Arabidopsis thaliana		
INSDC	CP002685.1 11871488-11876712 (+)	AEC08055.1	Argonaute family protein	Arabidopsis thaliana		GCA_000001735.1
PAT	FB823418.1 1-2994 (+)	CAW52555.1	unnamed protein product	Arabidopsis thaliana		
PAT	HB693477.1 1-2994 (+)	CBC23730.1	unnamed protein product	Arabidopsis thaliana		
PAT	HB840074.1 1-2994 (+)	CBD33103.1	unnamed protein product	Arabidopsis thaliana		

Search the Argonaute protein family with BLAST!

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

```
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  NVMKLLVKNYKDSHLGGKS---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 ssciname"   # 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
```

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.
```

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
```

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
```

BLAST against Ascomycota sequences!

```
blastp
    -db nr
        -query NP_850110.fsa
            -outfmt "7 qaccver saccver pident length evalue ssciname”   # 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
```

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     ODQ75158.1      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     ORY81937.1      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  31.188  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 ssciname”
                -out NP_850110.noGreenPlants.tab
                    -task blastp-fast
                        -num_threads 8
                            -evalue 0.05
                                -negative_taxidlist Viridiplantae.taxids          # exclude green plants
```

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
```

```
                                        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
```

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
```

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.

U.S. National Library of Medicine

National Center for Biotechnology Information

NCBI Resources How To

Protein

Protein (Shiga toxin AND Escherichia coli) AND "Escherichia coli"[porgn:__txid562]

Create alert Advanced

Species

Bacteria (2,900,149)

Customize ...

Source databases

Summary 20 per page Sort by Default order Send to:

See the results of this search (528 items) in our new Identical Protein Groups database.

U.S. National Library of Medicine

National Center for Biotechnology Information

NCBI Resources How To

Identical Protein Groups

Identical Protein | shiga toxin

Create alert Advanced

Source databases
INSDC (779)
PDB (2)
RefSeq (372)
UniProtKB / Swiss-Prot (5)
PIR (3)
PRF (5)

Division
Prokaryotes (900)

Protein count bins
1 (501)
2 (158)
3-25 (217)
26-50 (12)
51-100 (15)
>100 (17)
>1000 (2)

Protein count
Custom range...

Molecular weight
Custom range...

Sequence length
Custom range...

Creation date
Custom range...

Clear all

Show additional filters

Summary 20 per page Send to:

Search results

Items: 1 to 20 of 920

<< First < Prev Page 1 of 46 Next > Last >>

1. **Shiga toxin** Stx2 subunit B
   89 aa protein
   853 sequences, including WP_000738068.1
   FASTA BLAST

2. **Shiga toxin** Stx2 subunit B
   89 aa protein
   582 sequences, including WP_000738080.1
   FASTA BLAST

3. **Shiga toxin** Stx2 subunit A
   319 aa protein
   743 sequences, including WP_000649753.1
   FASTA BLAST

4. **Shiga toxin** A subunit
   143 aa protein (enterobacteria)
   1609 sequences, including WP_000227044.1
   FASTA BLAST

5. **Shiga toxin** A subunit
   149 aa protein (enterobacteria)
   2 sequences, including WP_126534086.1
   FASTA BLAST

6. **Shiga toxin** Stx2 subunit B
   92 aa protein (enterobacteria)

Why Use Identical Protein Groups?

NCBI Resources How To

Protein

Protein (Shiga toxin AND Escherichia coli) AND "Escherichia coli"[porgn:__txid562]

Create alert Advanced

Species
Bacteria (2,900,149)
Customize ...

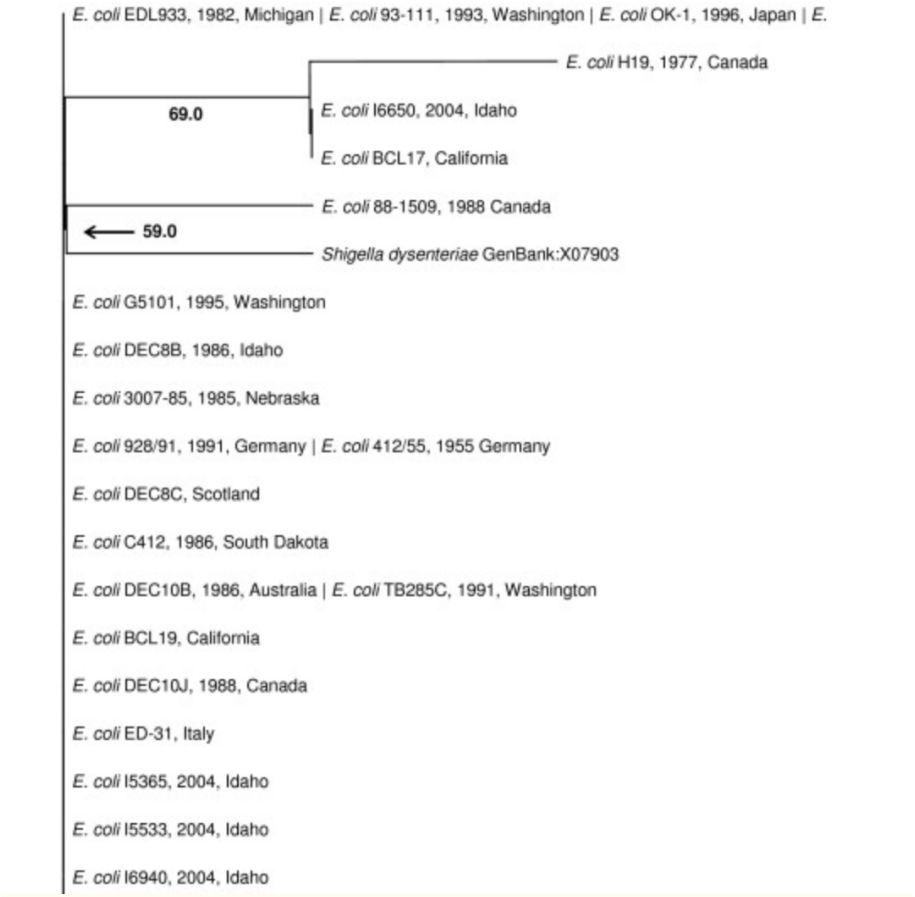
Source databases
PDB (105)
RefSeq (5,819)

Summary 20 per page Sort by Default order Send to:

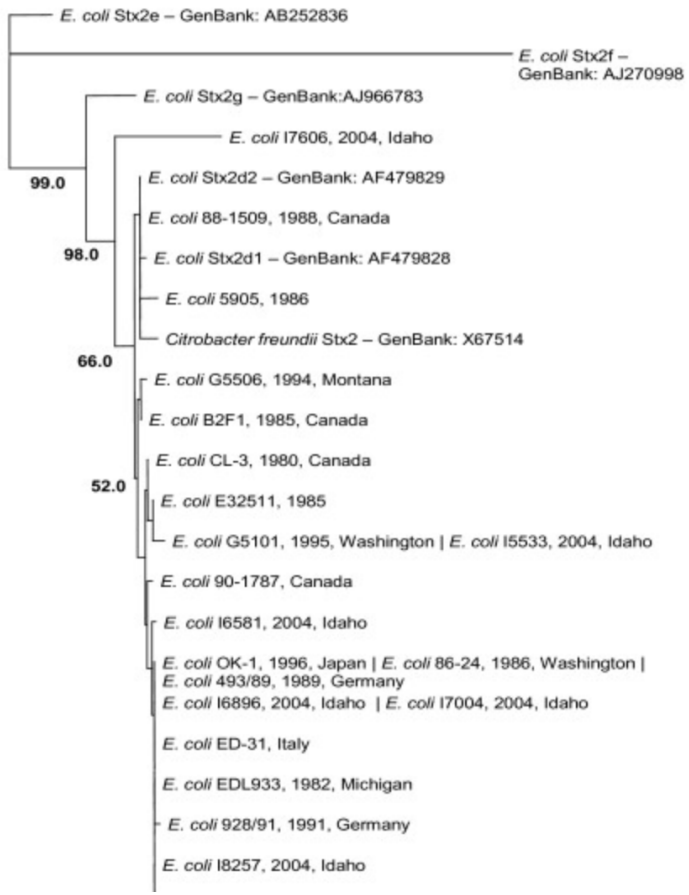
See the results of this search (528 items) in our new Identical Protein Groups database.

Items: 1 to 20 of 2900149

STX1


E. coli EDL933, 1982, Michigan | E. coli 93-111, 1993, Washington | E. coli OK-1, 1996, Japan | E.
E. coli H19, 1977, Canada
69.0
E. coli I6650, 2004, Idaho
E. coli BCL17, California
E. coli 88-1509, 1988 Canada
59.0
Shigella dysenteriae GenBank:X07903
E. coli G5101, 1995, Washington
E. coli DEC8B, 1986, Idaho
E. coli 3007-85, 1985, Nebraska
E. coli 928/91, 1991, Germany | E. coli 412/55, 1955 Germany
E. coli DEC8C, Scotland
E. coli C412, 1986, South Dakota
E. coli DEC10B, 1986, Australia | E. coli TB285C, 1991, Washington
E. coli BCL19, California
E. coli DEC10J, 1988, Canada
E. coli ED-31, Italy
E. coli I5365, 2004, Idaho
E. coli I5533, 2004, Idaho
E. coli I6940, 2004, Idaho

STX2


E. coli Stx2e – GenBank: AB252836
E. coli Stx2f – GenBank: AJ270998
E. coli Stx2g – GenBank:AJ966783
E. coli I7606, 2004, Idaho
99.0
E. coli Stx2d2 – GenBank: AF479829
E. coli 88-1509, 1988, Canada
98.0
E. coli Stx2d1 – GenBank: AF479828
E. coli 5905, 1986
Citrobacter freundii Stx2 – GenBank: X67514
66.0
E. coli G5506, 1994, Montana
E. coli B2F1, 1985, Canada
E. coli CL-3, 1980, Canada
52.0
E. coli E32511, 1985
E. coli G5101, 1995, Washington | E. coli I5533, 2004, Idaho
E. coli 90-1787, Canada
E. coli I6581, 2004, Idaho
E. coli OK-1, 1996, Japan | E. coli 86-24, 1986, Washington |
E. coli 493/89, 1989, Germany
E. coli I6896, 2004, Idaho | E. coli I7004, 2004, Idaho
E. coli ED-31, Italy
E. coli EDL933, 1982, Michigan
E. coli 928/91, 1991, Germany
E. coli I8257, 2004, Idaho

Search the Shiga toxin family protein with BLAST

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

Search the Shiga toxin family protein with BLAST

- Against the nr pr[illegible]base
- Limit the searc[illegible]o E coli

But which Shiga toxin?!

Stx1 and Stx2 Example Sequences

Shiga toxin Stx1 subunit A [Escherichia coli]

NCBI Reference Sequence: WP_123135120.1

Identical Proteins FASTA Graphics

Go to:

```
LOCUS       WP_123135120             315 aa            linear   BCT 14-NOV-2018
DEFINITION  Shiga toxin Stx1 subunit A [Escherichia coli].
ACCESSION   WP_123135120
VERSION     WP_123135120.1
KEYWORDS    RefSeq.
SOURCE      Escherichia coli
  ORGANISM  Escherichia coli
            Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales;
            Enterobacteriaceae; Escherichia.
COMMENT     REFSEQ: This record represents a single, non-redundant, protein
            sequence which may be annotated on many different RefSeq genomes
            from the same, or different, species.
            COMPLETENESS: full length.
FEATURES             Location/Qualifiers
     source          1..315
                     /organism="Escherichia coli"
                     /db_xref="taxon:562"
     gene            1..315
                     /gene="stxA1"
     Protein         1..315
                     /product="Shiga toxin Stx1 subunit A"
                     /calculated_mol_wt=34798
     Region          30..232
                     /region_name="RIP"
                     /note="Ribosome inactivating protein; pfam00161"
                     /db_xref="CDD:306635"
ORIGIN
        1 mkiiifrvlt fffvifsvnv vakeftldfs taktyvdsln virsaigtpl qtissggtsl
       61 lmidsgtgdn lfavdvrgid peegrfnnlr livernnlyv tgfvnrtnnv fyrfadfshv
      121 tfpgttavtl sgdssyttlq rvagisrmgm qinrhsltts yldlmshsgt sltqsvaram
      181 lrfvtvtaea lrfrqiqrgf rttlddlsgr syvmtaedvd ltlnwgrlss vlpdyhgqds
      241 vrvgrisfgs vnailgsval ilnchhhasr varivpdefp smcpvdrrvr githnkilwd
      301 sstlgailir raiss
//
```

Shiga toxin Stx2 subunit A [Escherichia coli]

NCBI Reference Sequence: WP_123130813.1

Identical Proteins FASTA Graphics

Go to:

```
LOCUS       WP_123130813             319 aa            linear   BCT 14-NOV-2018
DEFINITION  Shiga toxin Stx2 subunit A [Escherichia coli].
ACCESSION   WP_123130813
VERSION     WP_123130813.1
KEYWORDS    RefSeq.
SOURCE      Escherichia coli
  ORGANISM  Escherichia coli
            Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales;
            Enterobacteriaceae; Escherichia.
COMMENT     REFSEQ: This record represents a single, non-redundant, protein
            sequence which may be annotated on many different RefSeq genomes
            from the same, or different, species.
            COMPLETENESS: full length.
FEATURES             Location/Qualifiers
     source          1..319
                     /organism="Escherichia coli"
                     /db_xref="taxon:562"
     gene            1..319
                     /gene="stxA2"
     Protein         1..319
                     /product="Shiga toxin Stx2 subunit A"
                     /calculated_mol_wt=35672
     Region          30..231
                     /region_name="RIP"
                     /note="Ribosome inactivating protein; pfam00161"
                     /db_xref="CDD:306635"
ORIGIN
        1 mkcilfkwvl clllgfssvs ysreftidfs tqqsyvssln sirteistpl ehisqgttsv
       61 svinhtppgs yfavdirgld vyqarfdhlr liieqnnlyv agfvntatnt fyrfsdfthi
      121 svpdvttvsm ttdssyttlq rvaalersgm qisrhslvss ylalmefsgn tmtrdasrav
      181 lrfvtvtaea lrfrqiqref rqalsetapv ytmtpgdvdl tlnwgrisnv lpeyrgedgv
      241 rvgrisfnni sailstvavi lnchhqgars vravneesqp ecqitgdrpv ikinntlwes
      301 ntaaaflnrk sqflyttgk
//
```

Stx1 and Stx2 Example Sequences

Shiga toxin Stx1 subunit A [Escherichia coli]

NCBI Reference Sequence: WP_123135120.1

Identical Proteins FASTA Graphics

Go to:

```
LOCUS       WP_123135120             315 aa            linear   BCT 14-NOV-2018
DEFINITION  Shiga toxin Stx1 subunit A [Escherichia coli].
ACCESSION   WP_123135120
VERSION     WP_123135120.1
KEYWORDS    RefSeq.
SOURCE      Escherichia coli
  ORGANISM  Escherichia coli
            Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales;
            Enterobacteriaceae; Escherichia.
COMMENT     REFSEQ: This record represents a single, non-redundant, protein
            sequence which may be annotated on many different RefSeq genome
            from the same, or different, species.
            COMPLETENESS: full length.
FEATURES             Location/Qualifiers
     source          1..315
                     /organism="Escherichia coli"
                     /db_xref="taxon:562"
     gene            1..315
                     /gene="stxA1"
     Protein         1..315
                     /product="Shiga toxin Stx
                     /calculated_mol_wt=347
     Region          30..232
                     /region_name="RIP"
                     /note="Ribosome inactiva
                     /db_xref="CDD:306635"
ORIGIN
        1 mkiiifrvlt fffvifsvnv vakeftldfs taktyv                 issggtsl
       61 lmidsgtgdn lfavdvrgid peegrfnnlr livernn                fyrfadfshv
      121 tfpgttavtl sgdssyttlq rvagisrmgm qinrhsltt              sgt sltqsvaram
      181 lrfvtvtaea lrfrqiqrgf rttlddlsgr syvmtaedvd             wgrlss vlpdyhgqds
      241 vrvgrisfgs vnailgsval ilnchhhasr varivpdefp smcpvdrrvr githnkilwd
      301 sstlgailir raiss
//
```

Shiga toxin Stx[illegible] [Escherichia coli]

NCBI Reference S[illegible]

Identical Protei[illegible]

Go to[illegible]

```
                                        319 aa            linear   BCT 14-NOV-2018
                                        t A [Escherichia coli].

                                  coli
                                  nia coli
                                  ria; Proteobacteria; Gammaproteobacteria; Enterobacterales;
                                  terobacteriaceae; Escherichia.
                                  REFSEQ: This record represents a single, non-redundant, protein
                                  sequence which may be annotated on many different RefSeq genomes
                                  from the same, or different, species.
                                  COMPLETENESS: full length.
FEATURES             Location/Qualifiers
     source          1..319
                     /organism="Escherichia coli"
                     /db_xref="taxon:562"
     gene            1..319
                     /gene="stxA2"
     Protein         1..319
                     /product="Shiga toxin Stx2 subunit A"
                     /calculated_mol_wt=35672
     Region          30..231
                     /region_name="RIP"
                     /note="Ribosome inactivating protein; pfam00161"
                     /db_xref="CDD:306635"
ORIGIN
        1 mkcilfkwvl clllgfssvs ysreftidfs tqqsyvssln sirteistpl ehisqgttsv
       61 svinhtppgs yfavdirgld vyqarfdhlr liieqnnlyv agfvntatnt fyrfsdfthi
      121 svpdvttvsm ttdssyttlq rvaalersgm qisrhslvss ylalmefsgn tmtrdasrav
      181 lrfvtvtaea lrfrqiqref rqalsetapv ytmtpgdvdl tlnwgrisnv lpeyrgedgv
      241 rvgrisfnni sailstvavi lnchhqgars vravneesqp ecqitgdrpv ikinntlwes
      301 ntaaaflnrk sqflyttgk
//
```

Representative samples. How do I know I have the right ones?
What if they are misannotated?
Answer: BLAST ALL of them!

BLAST command-line parameters

```
blastp
    -db refseq-protein_v5
        -query  WP_123135120_123130813.fasta
            -outfmt "7 qaccver saccver pident length evalue ssciname”   # 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 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!

CloudBLAST and Docker Images to the Rescue!

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

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

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

Install NCBI-provided BLAST databases

The `$BLASTDB_DIR` environment variable refers to an existing, writable directory on the local host. The following command will download the `swissprot_v5` BLAST database from GCP into the `$BLASTDB_DIR` directory (notice the `-w` argument, which sets the working directory for that command):

```
docker run --rm \
  -v $BLASTDB_DIR:/blast/blastdb:rw \
  -w /blast/blastdb \
  ncbi/blast \
  update_blastdb.pl --source gcp swissprot_v5
```

Make and install my own BLAST databases

If you have your own sequence data in a file called `sequences.fsa` and want to make a BLAST database, please run the command below:

```
docker run --rm \
  -v $HOME/blastdb_custom:/blast/blastdb_custom:rw \
  -v $HOME/fasta:/blast/fasta:rw \
  -w /blast/blastdb_custom \
  ncbi/blast \
  makeblastdb -in /blast/fasta/sequences.fsa -dbtype prot -out proteins -title 'My BLASTDB title'
```

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
```

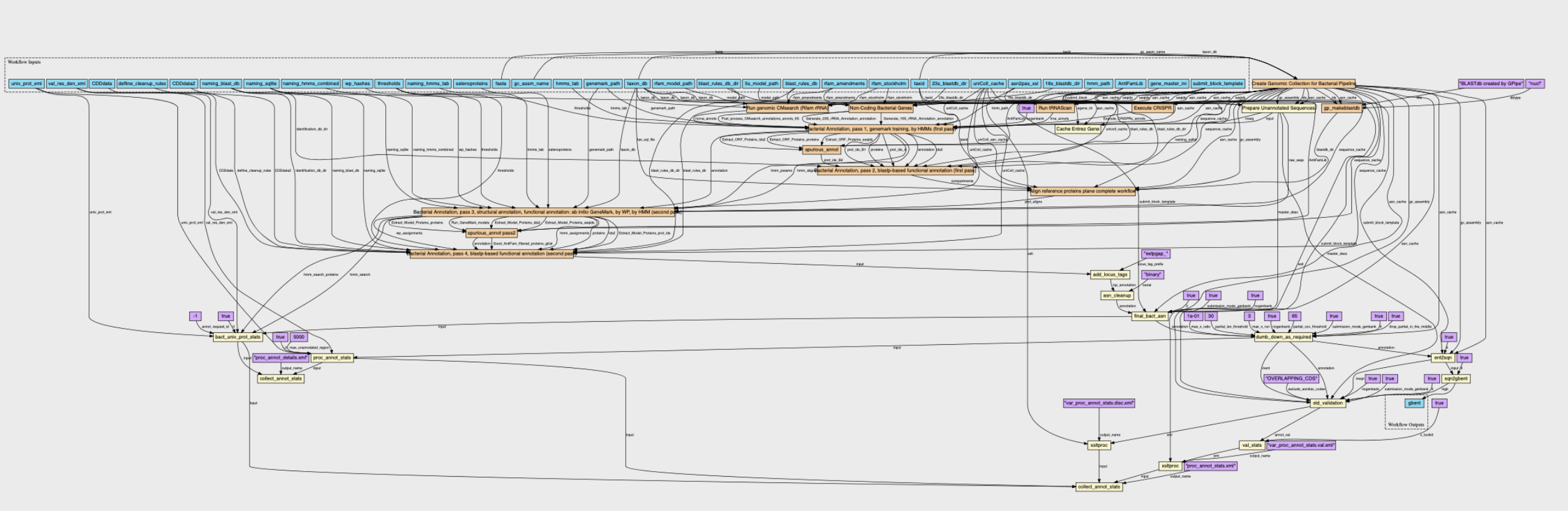
But how do I use the BLAST Docker(s)?!

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
```

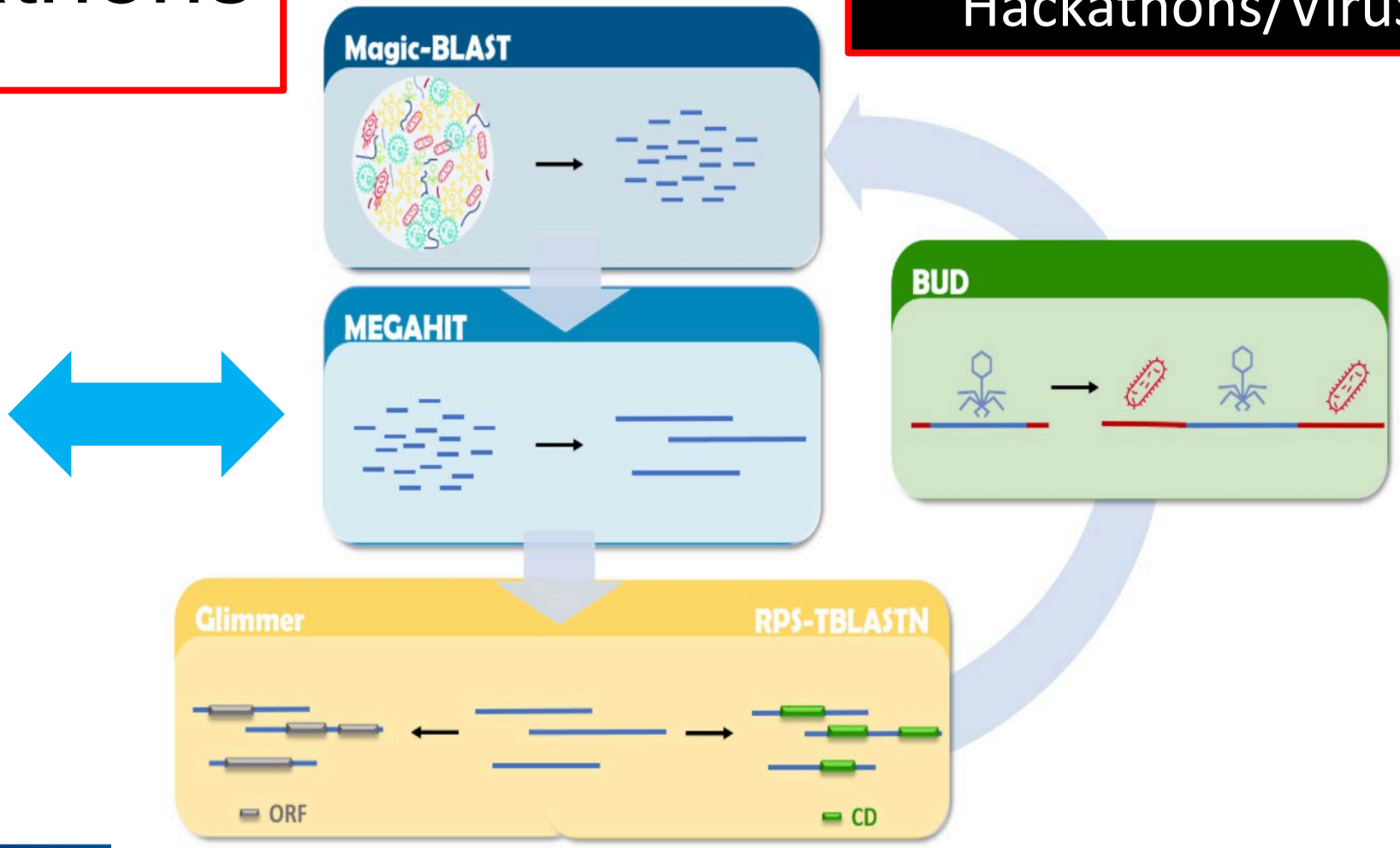
Bonus: Prokaryotic Genome Annotation Pipeline Wrapped with Docker and CWL

https://github.com/ncbi/pgap



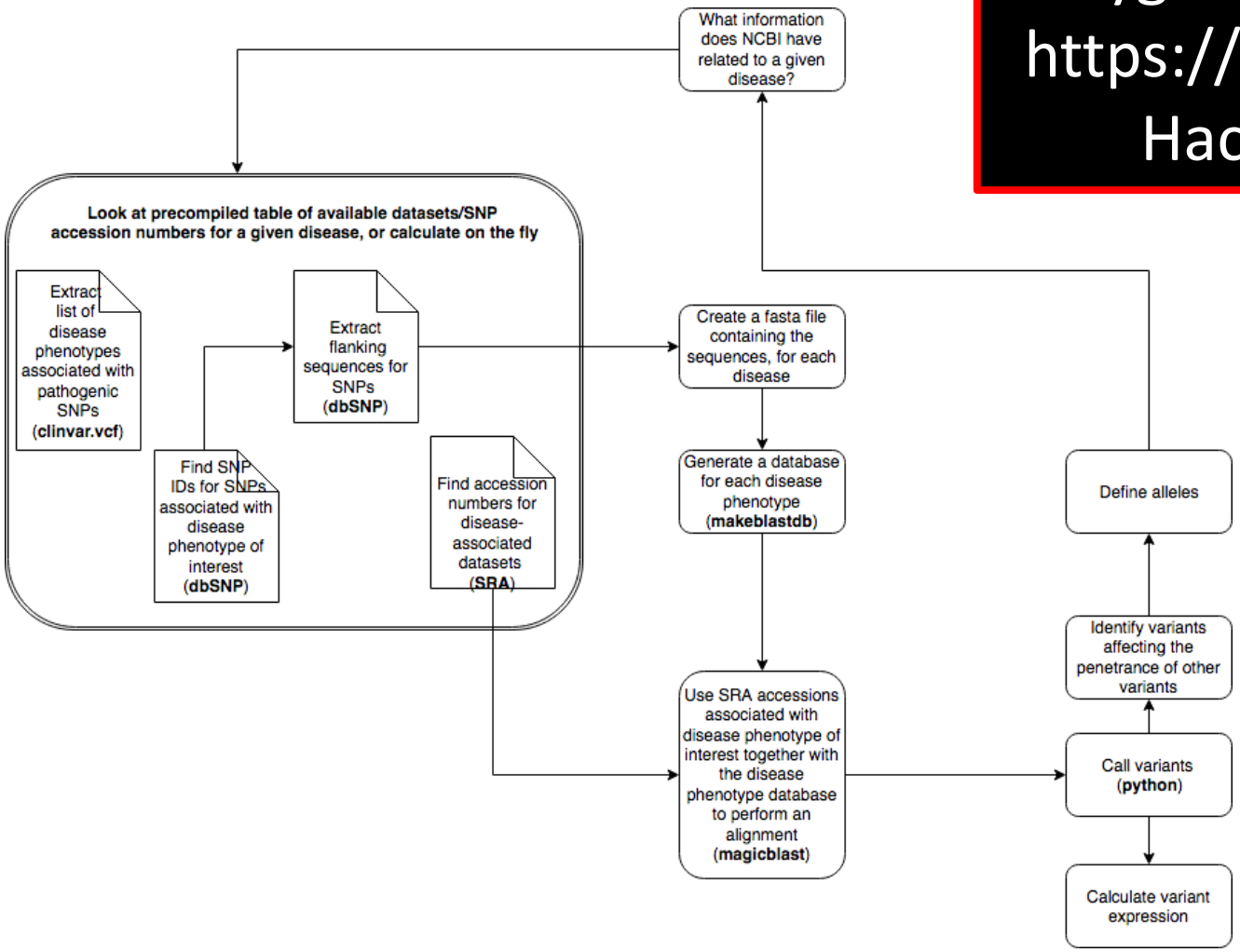
Hackathons

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



Variant calling on the fly!

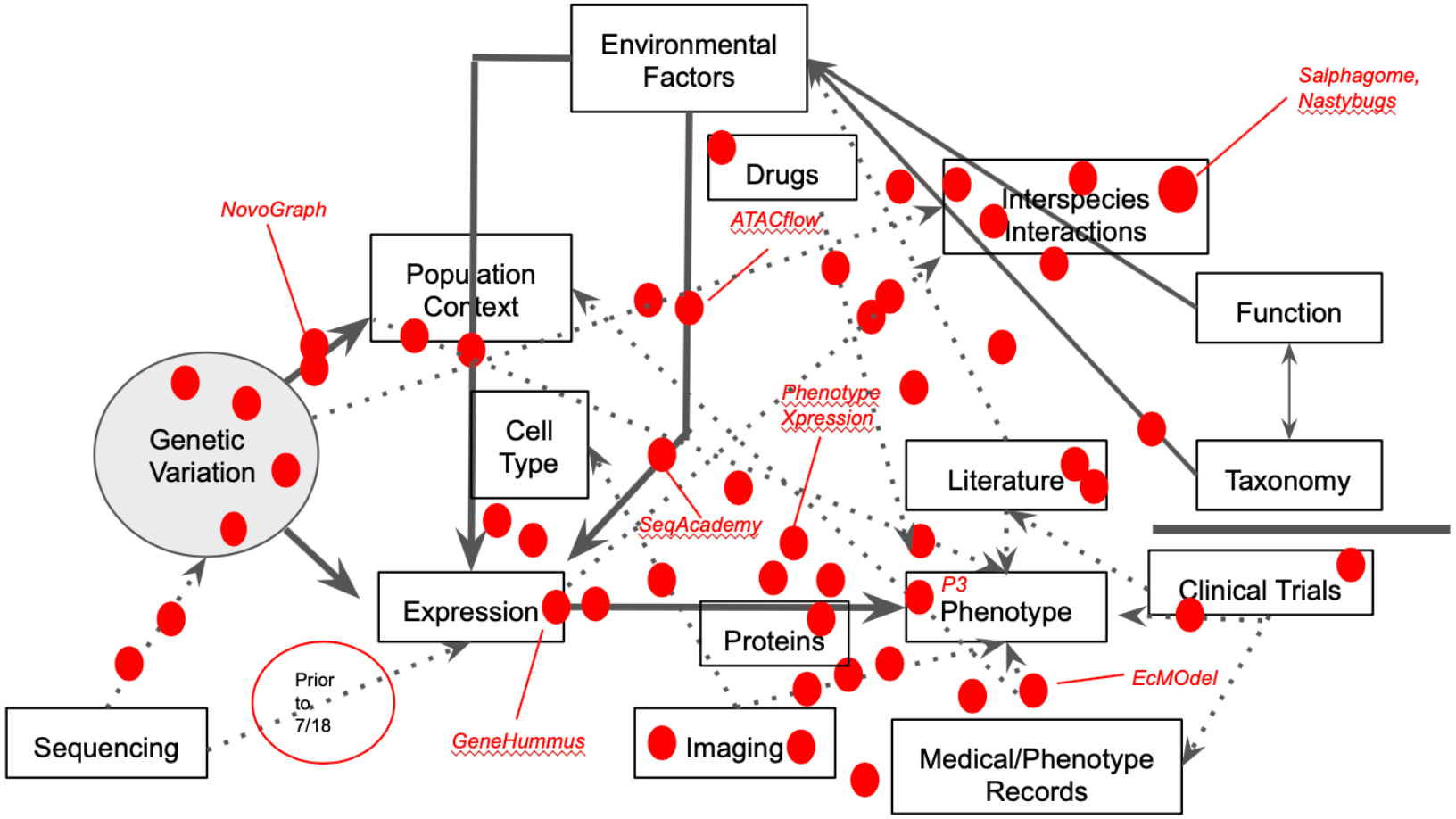
Polygenic SNP Search Tool
https://github.com/NCBI-Hackathons/PSST



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

Human, Plant and Animal Genomes

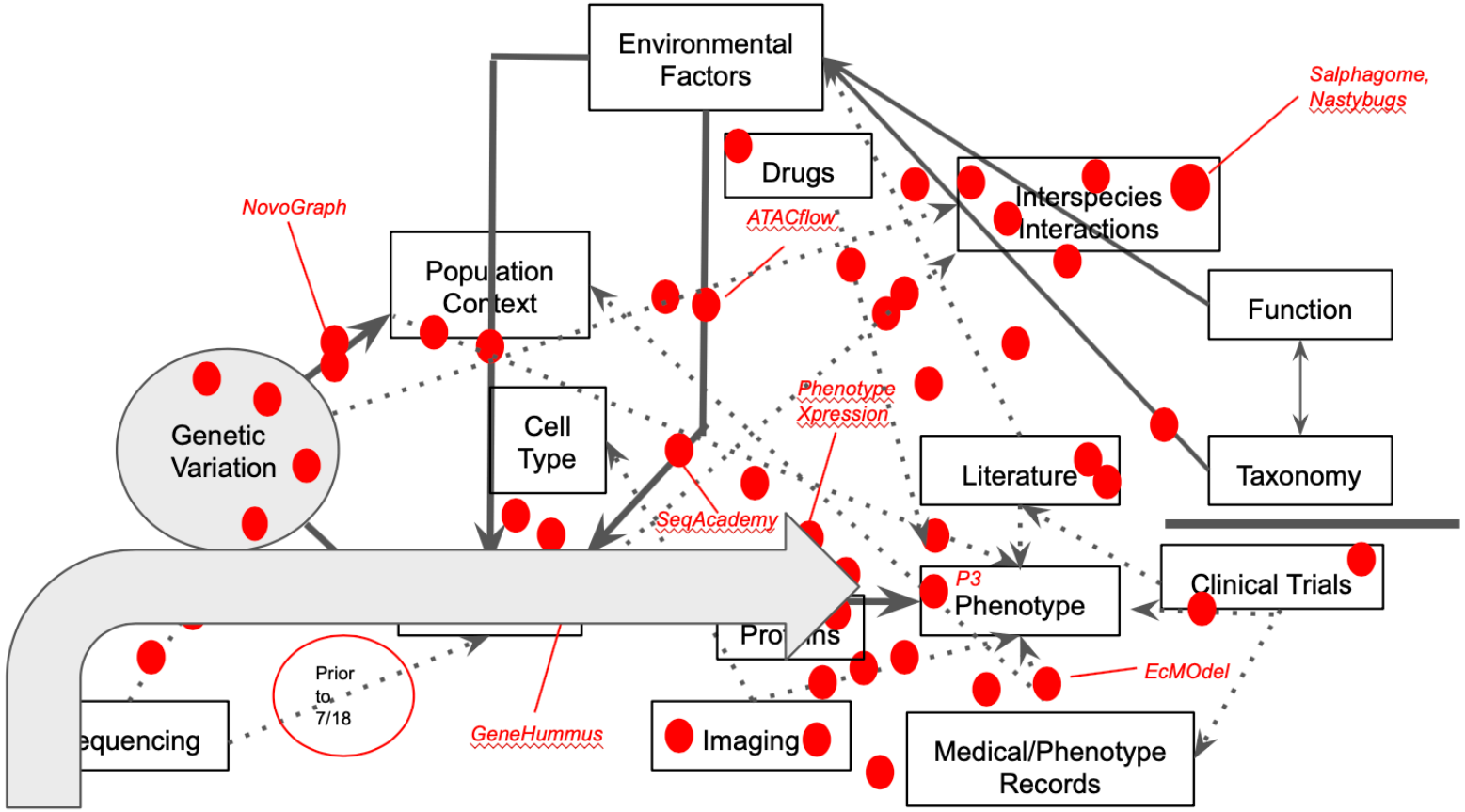
Bacterial, Viral, Fungus and Other Genomes



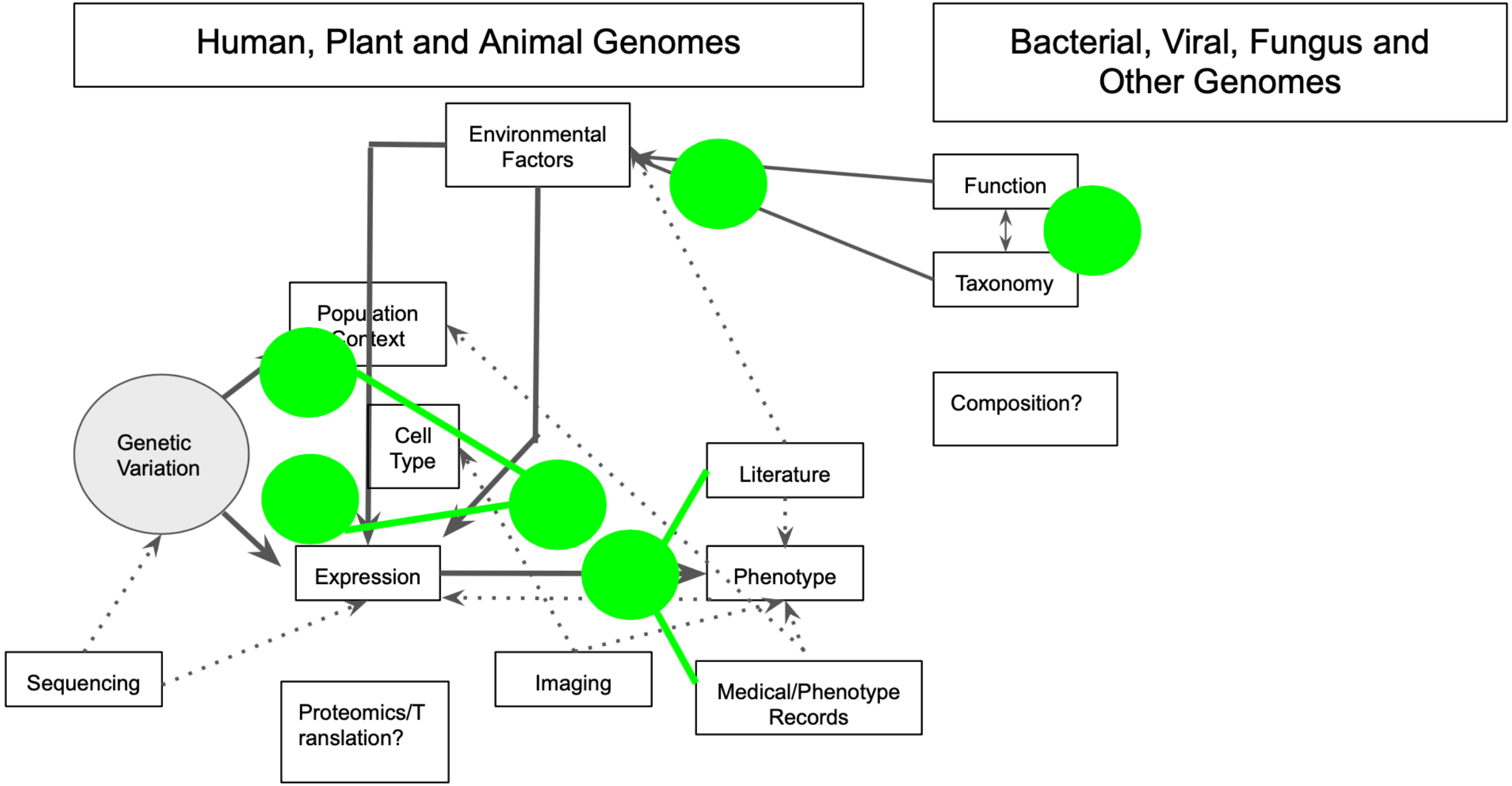
Previous Hackathons (commits 7/1/2018 - 1/1/19 only)

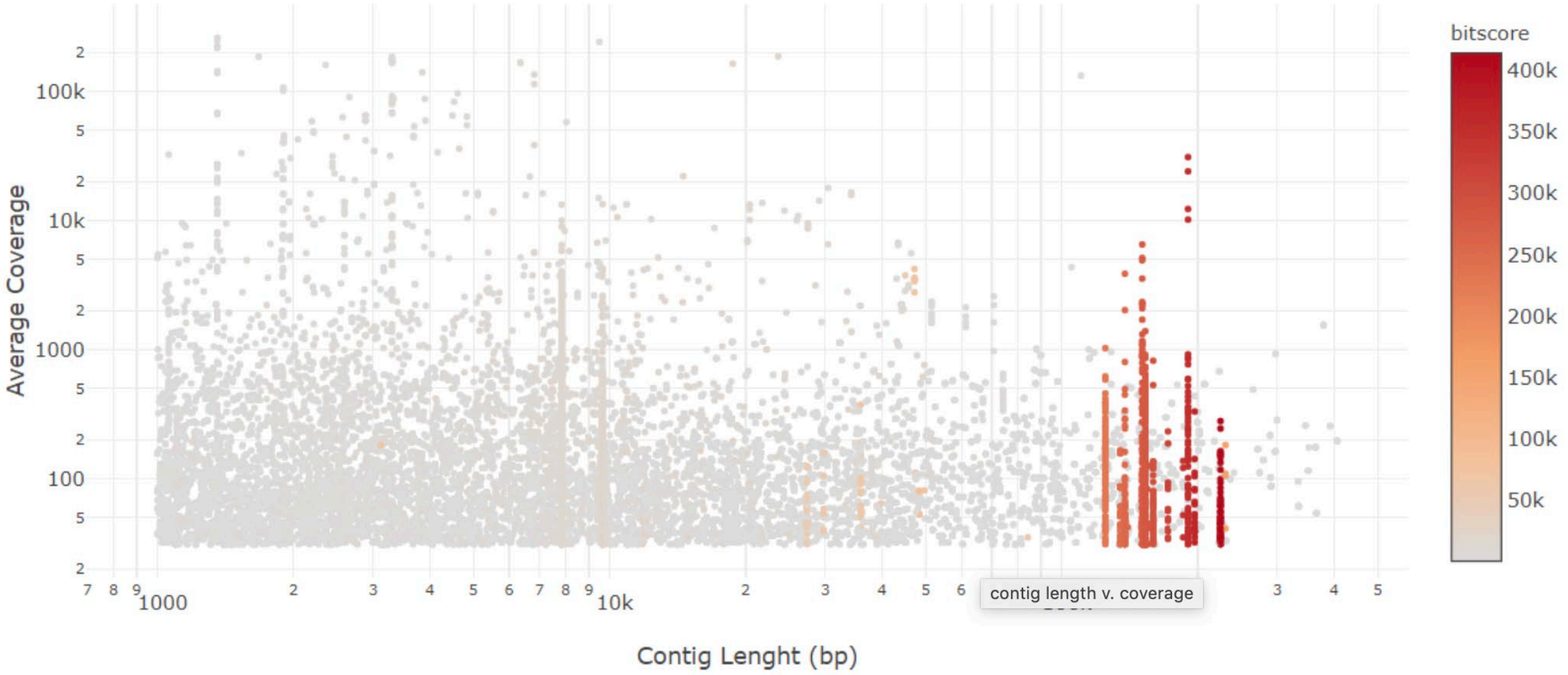
Human, Plant and Animal Genomes

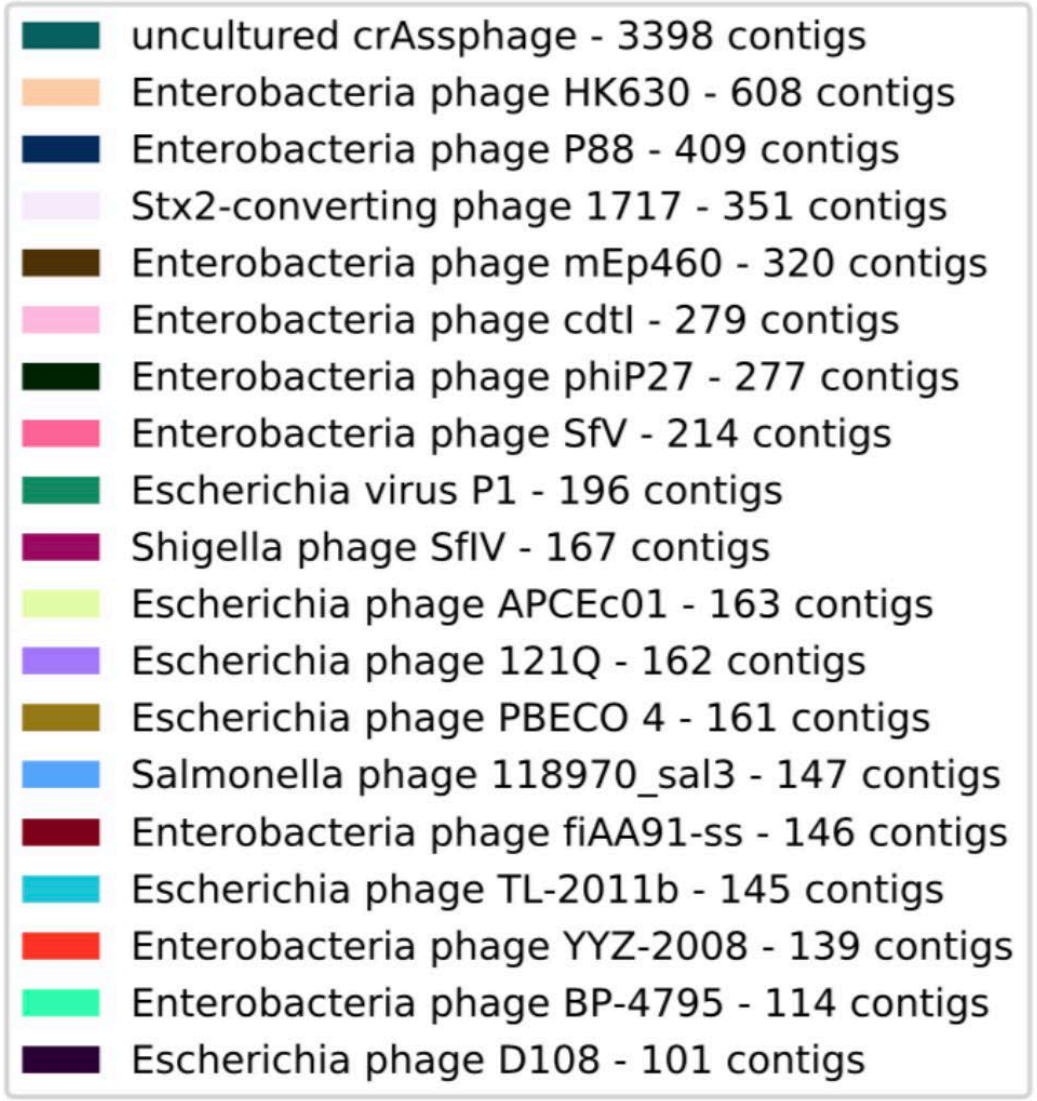
Bacterial, Viral, Fungus and Other Genomes

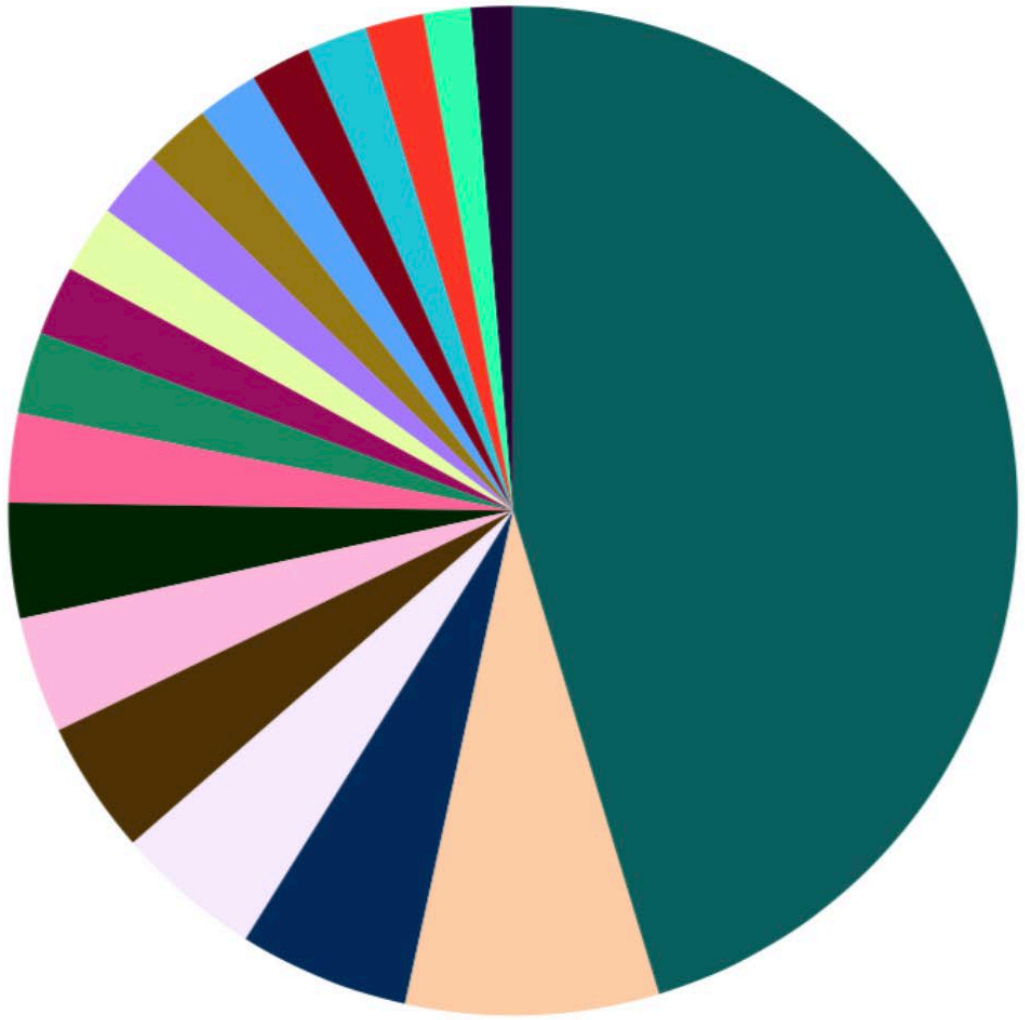


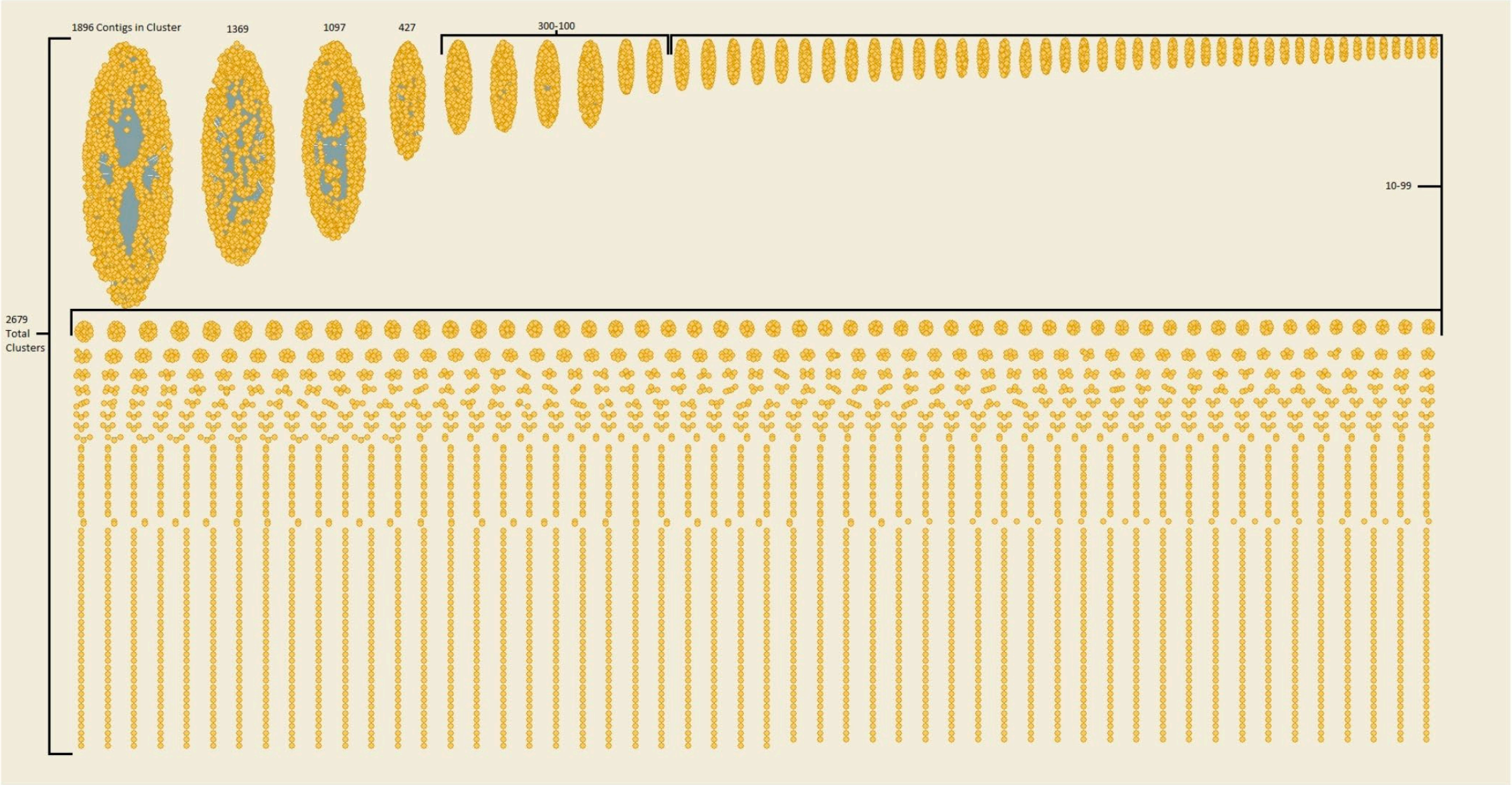
"STRIDES" Hackathons!



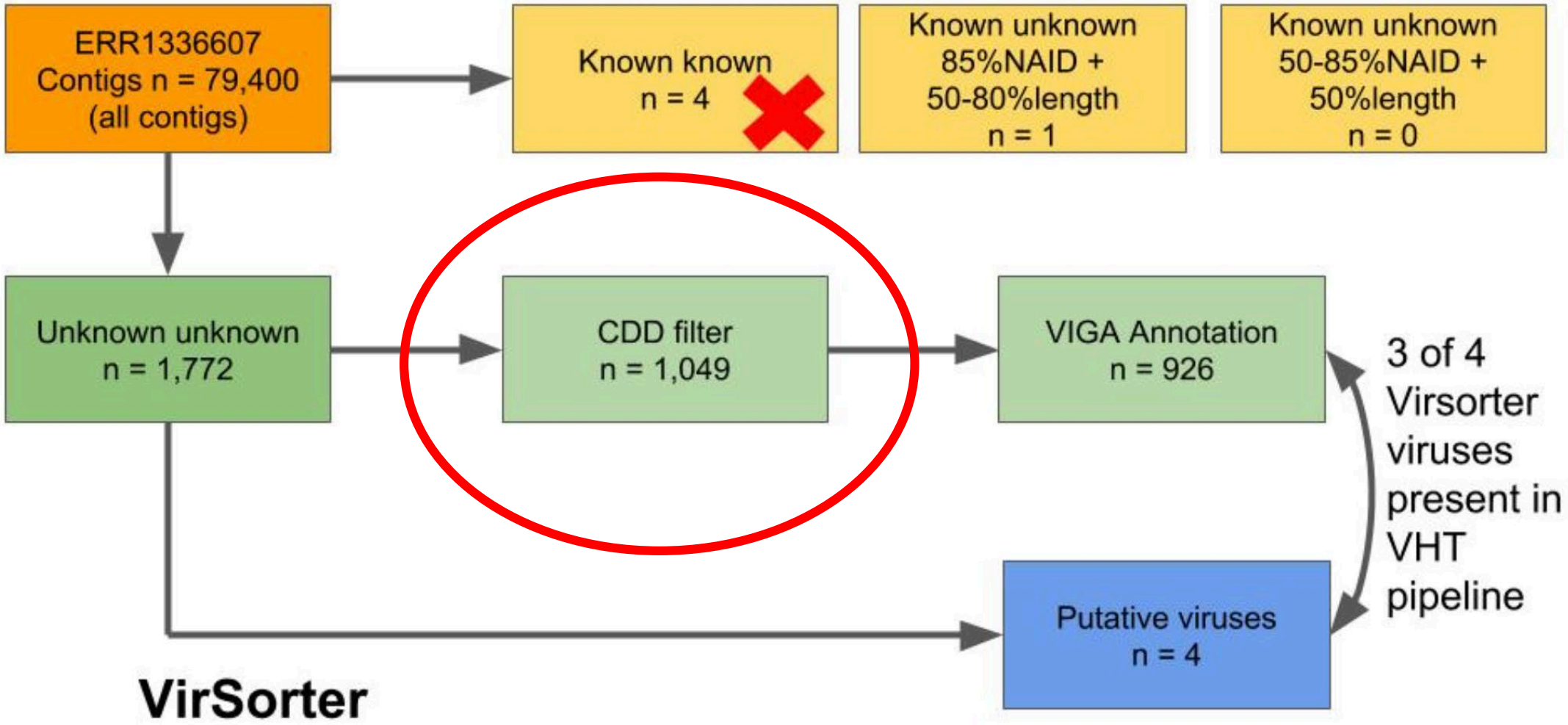

bitscore
400k
350k
300k
250k
200k
150k
100k
50k
Average Coverage
2
100k
5
2
10k
5
2
1000
5
2
100
5
2
7 8 9 1000 2 3 4 5 6 7 8 9 10k 2 3 4 5 6
3 4 5
contig length v. coverage
Contig Lenght (bp)


uncultured crAssphage - 3398 contigs
Enterobacteria phage HK630 - 608 contigs
Enterobacteria phage P88 - 409 contigs
Stx2-converting phage 1717 - 351 contigs
Enterobacteria phage mEp460 - 320 contigs
Enterobacteria phage cdtI - 279 contigs
Enterobacteria phage phiP27 - 277 contigs
Enterobacteria phage SfV - 214 contigs
Escherichia virus P1 - 196 contigs
Shigella phage SfIV - 167 contigs
Escherichia phage APCEc01 - 163 contigs
Escherichia phage 121Q - 162 contigs
Escherichia phage PBECO 4 - 161 contigs
Salmonella phage 118970_sal3 - 147 contigs
Enterobacteria phage fiAA91-ss - 146 contigs
Escherichia phage TL-2011b - 145 contigs
Enterobacteria phage YYZ-2008 - 139 contigs
Enterobacteria phage BP-4795 - 114 contigs
Escherichia phage D108 - 101 contigs

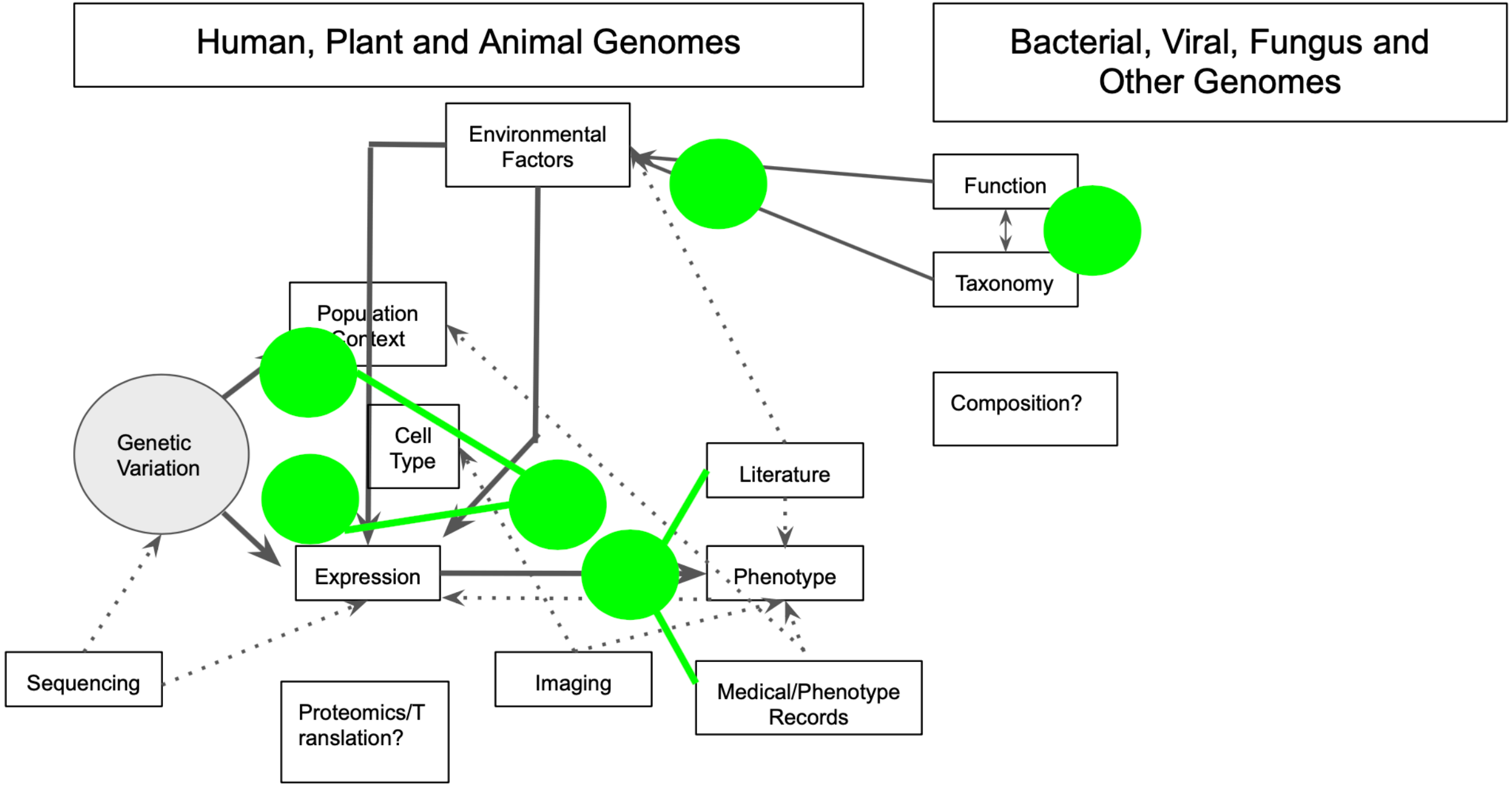



1896 Contigs in Cluster
1369
1097
427
300-100
10-99
2679
Total
Clusters

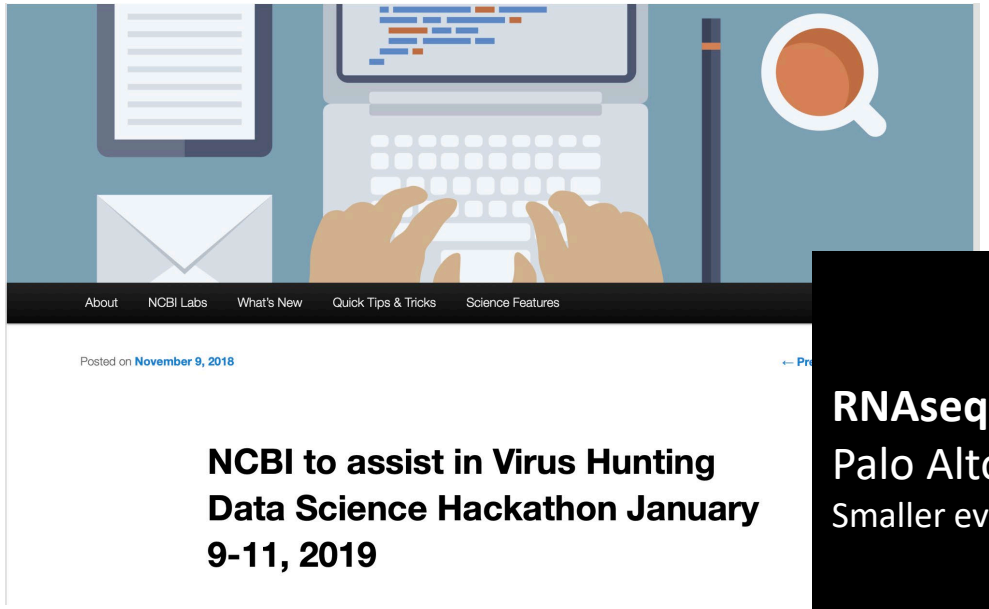
Example SRA dataset



"STRIDES" Hackathons!



Data Analysis Hackathons!



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

Creating a Community

BioHackathons | Selected Hackathon Products | About | Contact | Help | Resources | Community Resources | Join the mailing list!

Upcoming Hackathons

- ADVERTISE YOUR HACKATHON ON THIS PAGE!, ASAP
- Virus Hunting Hackathon -- San Diego (NCBI) -- APPLY NOW!, January 9-11
- Undiagnosed Disease Hackathon (SVAI), TBD
- Seattle at Fred Hutch (NCBI) -- Applications Open, APPLY NOW!, February 4-6, 2019
- NIH Bethesda Campus (NCBI) -- Team Proposals Now Open!, February 20-22, 2019
- IronHack + RNAseq at USF (NCBI) -- Team Proposals Now Open!, February 25-27, 2019
- Research Triangle at UNC (NCBI) -- Plus RNAseq analyzeathon phase 2!, March 11-13, 2019
- Pangenome and Haplotype Hackathon at UCSC! (NCBI), March 25-27, 2019
- BioIT World FAIR Data Hackathon, April 15-16, 2019
- Houston at BCM pre-ASHG (NCBI), October 12-14, 2019

https://biohackathons.github.io

Previous Hackathons

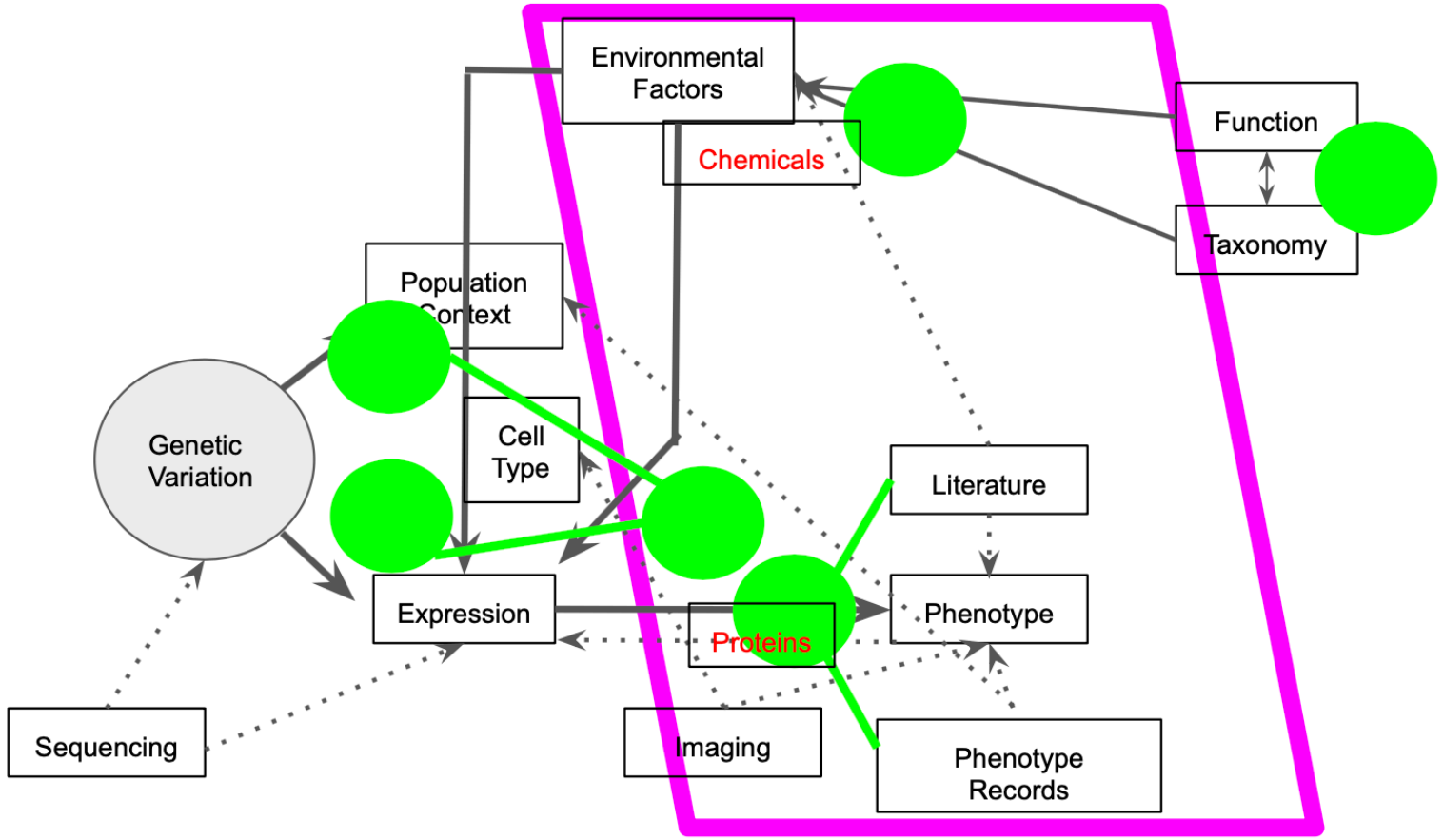
- Dallas at UTSW, pre-SC (NCBI Community Partner), November 9-10
- Cold Spring Harbor Biological Data Science (NCBI), November 10-12

MACHINE LEARNING
PROJECT MANAGEMENT
PRODUCT MANAGEMENT

Future Hackathons, Pipelining and Community Tools!

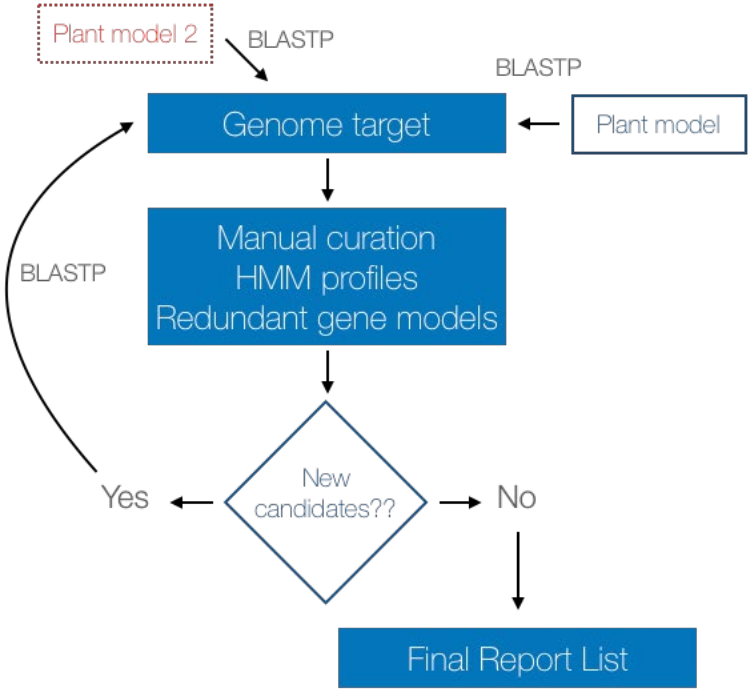
Human, Plant and Animal Genomes

Bacterial, Viral, Fungus and Other Genomes



GeneHummus

Automated Gene Family Discovery!



Schema for **exhaustive BLAST** search

Die et al. BMC Genomics (2018) 19:301
https://doi.org/10.1186/s12864-018-4695-9

BMC Genomics

RESEARCH ARTICLE Open Access

Genome-wide identification of the auxin response factor gene family in Cicer arietinum

Jose V. Die*, Juan Gil and Teresa Millan

Abstract

Background: Auxin Response Factors act as critical components of the auxin-signaling pathway by regulating the transcription of auxin-responsive genes. The release of the chickpea reference genome provides an opportunity to identify and characterize the ARF gene family in this important legume by a data mining coupled by comparative genomics approaches.

Results: We performed a comprehensive characterization and analysis of 24 ARF genes in the chickpea reference genome. Comparative phylogenetic analysis of the ARF from chickpea, *Medicago* and *Arabidopsis* suggests that recent duplications have played a very limited role in the expansion of the ARF chickpea family. Gene structure analysis based on exon-intron organization provides additional evidence to support the evolutionary relationship among the ARF members. Conserved motif analysis shows that most of the proteins fit into the canonical ARF structure model, but 9 proteins lack or have a truncated dimerization domain. The mechanisms underlying the diversification of the ARF gene family are based on duplications, variations in domain organization and alternative splicing. Concerning duplications, segmental, but not tandem duplications, have contributed to the expansion of the gene family. Moreover, the duplicated pair genes have evolved mainly under the influence of purifying selection pressure with restricted functional divergence. Expression profiles responding to various environmental stimuli show a close relationship between tissue and expression patterns. Promoter sequence analysis reveals an enrichment of several *cis*-regulatory elements related to symbiosis, and modulation of plant gene expression during the interaction with microbes.

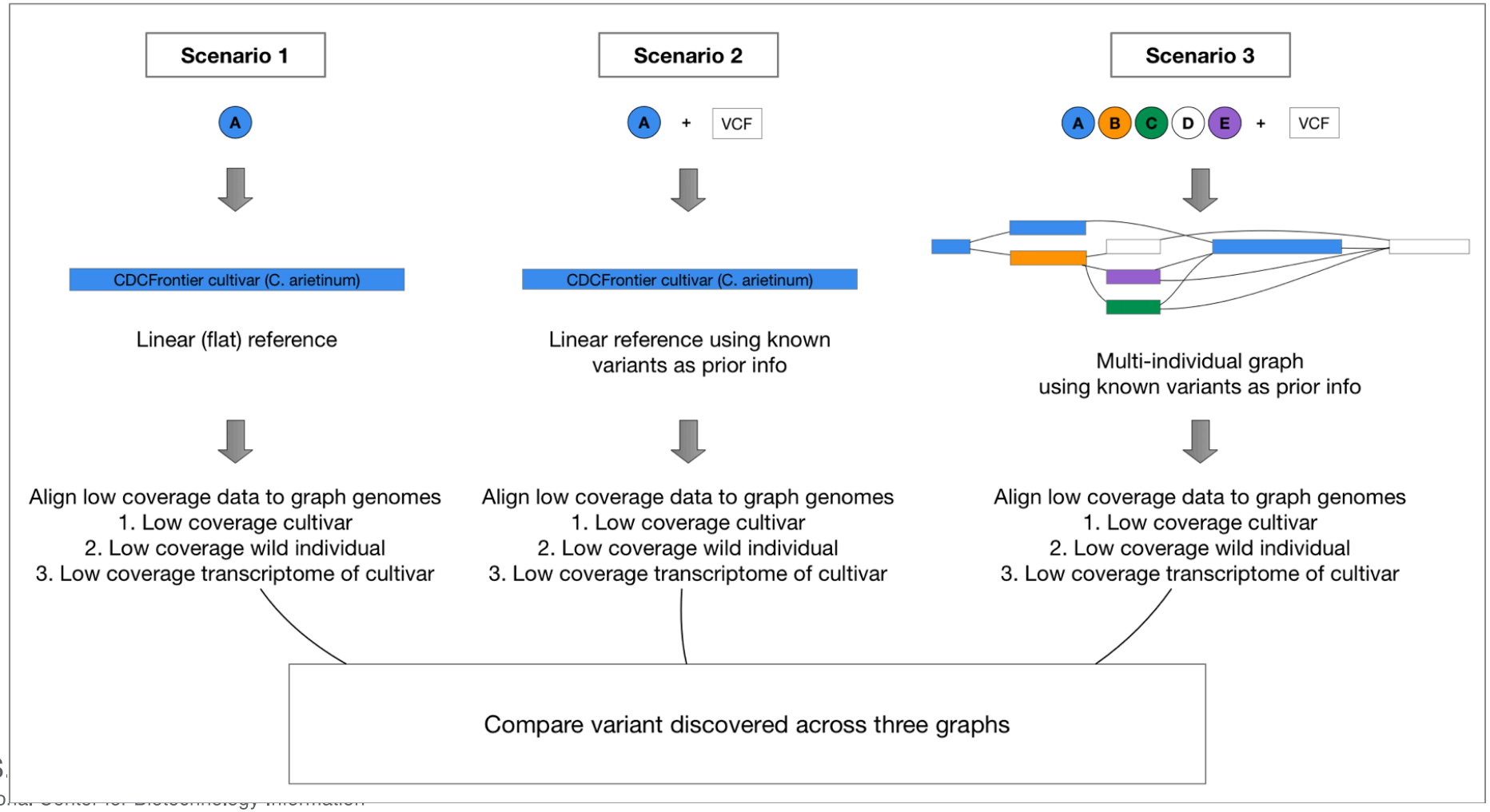
Conclusions: In conclusion, this study provides a comprehensive overview of the ARF gene family in chickpea. Globally, our data supports that auxin signaling pathway regulates a wide range of physiological processes and stress responses. Our findings could further provide new insights into the complexity of the regulation of ARF at the transcription level that may be useful to develop rational chickpea breeding strategies to improve development or stress responses. Our study also provides a foundation for comparative genomic analyses and a framework to trace the dynamic evolution of ARF genes on a large time-scale within the legume family.

Keywords: ARF, Bioinformatics, Chickpea, Gene duplication, Gene expression, RT-qPCR, Rstat

Auxin Response Factor gene family in the chickpea genome (2018)

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

Watch NCBI News for updates!

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

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