NCBI Resources in the Data Science Era!

Making the Transition from Sharing Data to Sharing Knowledge

Ben Busby, Ph.D.

Genomics Outreach Coordinator, Bioinformatics Training Lead

NCBI

Founder, Department of Bioinformatics and Data Science

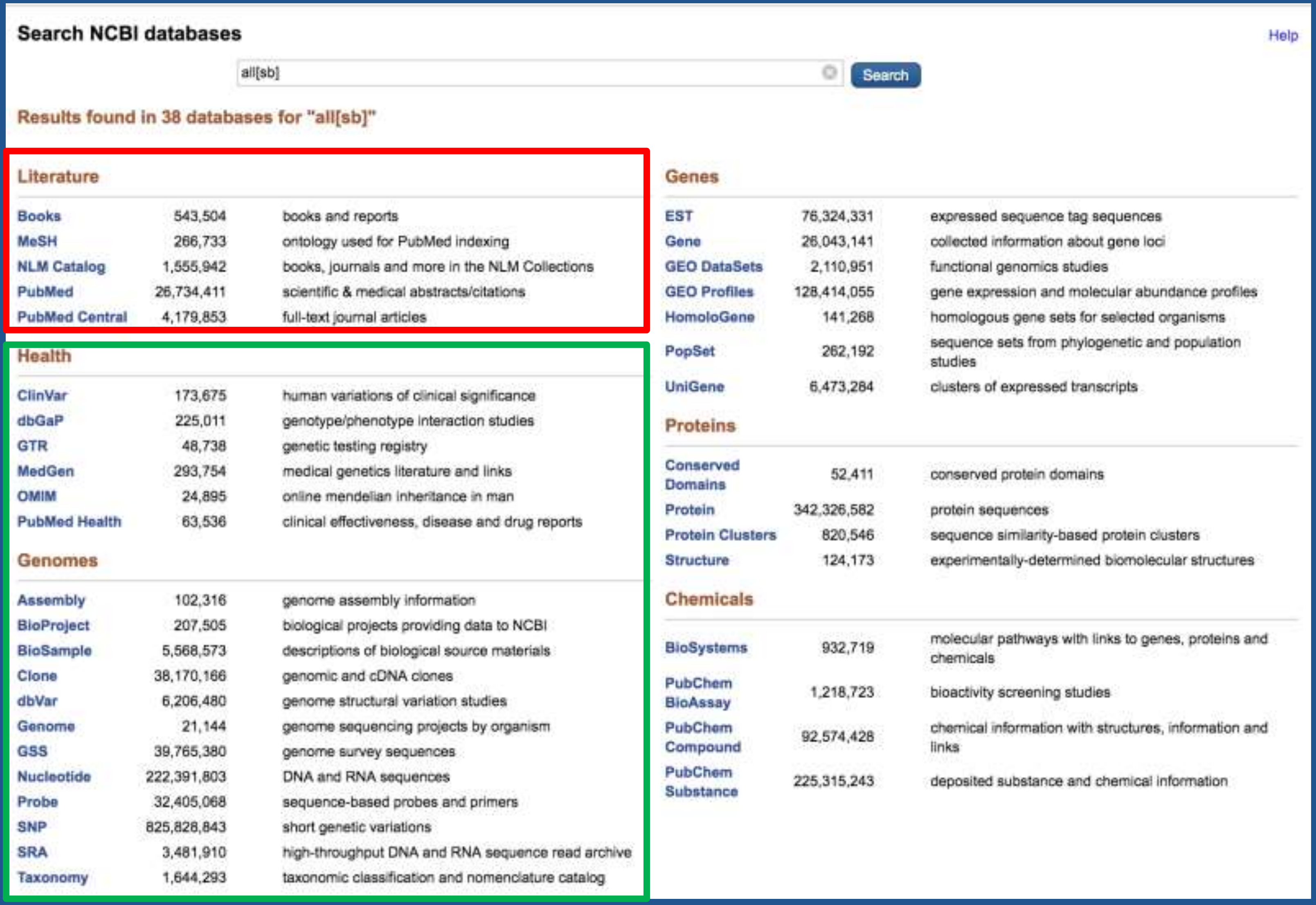
FAES

ben.busby@nih.gov

NIH U.S. National Library of Medicine

NCBI

NCBI



Search NCBI databases

Results found in 38 databases for "all[sb]"

Literature

Books	543,504	books and reports
MeSH	266,733	ontology used for PubMed indexing
NLM Catalog	1,555,942	books, journals and more in the NLM Collections
PubMed	26,734,411	scientific & medical abstracts/citations
PubMed Central	4,179,853	full-text journal articles

Health

ClinVar	173,675	human variations of clinical significance
dbGaP	225,011	genotype/phenotype interaction studies
GTR	48,738	genetic testing registry
MedGen	293,754	medical genetics literature and links
OMIM	24,895	online mendelian inheritance in man
PubMed Health	63,536	clinical effectiveness, disease and drug reports

Genomes

Assembly	102,316	genome assembly information
BioProject	207,505	biological projects providing data to NCBI
BioSample	5,568,573	descriptions of biological source materials
Clone	38,170,166	genomic and cDNA clones
dbVar	6,206,480	genome structural variation studies
Genome	21,144	genome sequencing projects by organism
GSS	39,765,380	genome survey sequences
Nucleotide	222,391,803	DNA and RNA sequences
Probe	32,405,068	sequence-based probes and primers
SNP	825,828,843	short genetic variations
SRA	3,481,910	high-throughput DNA and RNA sequence read archive
Taxonomy	1,644,293	taxonomic classification and nomenclature catalog

Genes

EST	76,324,331	expressed sequence tag sequences
Gene	26,043,141	collected information about gene loci
GEO DataSets	2,110,951	functional genomics studies
GEO Profiles	128,414,055	gene expression and molecular abundance profiles
HomoloGene	141,268	homologous gene sets for selected organisms
PopSet	262,192	sequence sets from phylogenetic and population studies
UniGene	6,473,284	clusters of expressed transcripts

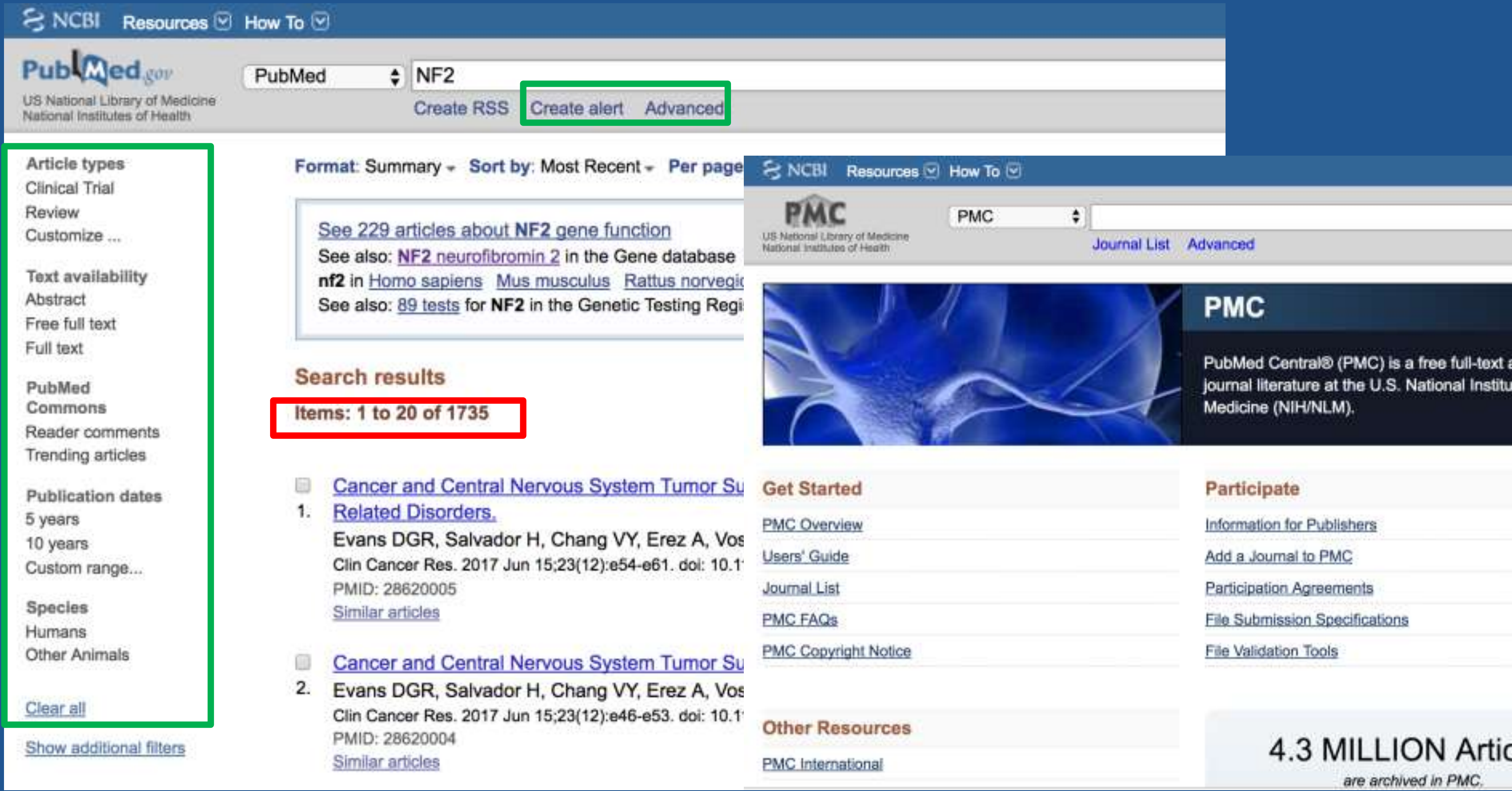
Proteins

Conserved Domains	52,411	conserved protein domains
Protein	342,326,582	protein sequences
Protein Clusters	820,546	sequence similarity-based protein clusters
Structure	124,173	experimentally-determined biomolecular structures

Chemicals

BioSystems	932,719	molecular pathways with links to genes, proteins and chemicals
PubChem BioAssay	1,218,723	bioactivity screening studies
PubChem Compound	92,574,428	chemical information with structures, information and links
PubChem Substance	225,315,243	deposited substance and chemical information

Three (and a half) ways to literature!



NF2

Article types
Clinical Trial
Review
Customize ...

Text availability
Abstract
Free full text
Full text

PubMed Commons
Reader comments
Trending articles

Publication dates
5 years
10 years
Custom range...

Species
Humans
Other Animals

Clear all

Show additional filters

Format: Summary - Sort by: Most Recent - Per page

See 229 articles about NF2 gene function
See also: NF2 neurofibromin 2 in the Gene database
nf2 in Homo sapiens Mus musculus Rattus norvegic
See also: 89 tests for NF2 in the Genetic Testing Regi

Search results

Items: 1 to 20 of 1735

1. Cancer and Central Nervous System Tumor Su Related Disorders.
Evans DGR, Salvador H, Chang VY, Erez A, Vos
Clin Cancer Res. 2017 Jun 15;23(12):e54-e61. doi: 10.1
PMID: 28620005
Similar articles

2. Cancer and Central Nervous System Tumor Su
Evans DGR, Salvador H, Chang VY, Erez A, Vos
Clin Cancer Res. 2017 Jun 15;23(12):e46-e53. doi: 10.1
PMID: 28620004
Similar articles

NCBI
Resources
How To
PMC
US National Library of Medicine
National Institutes of Health
PMC
Journal List
Advanced

PMC

PubMed Central® (PMC) is a free full-text a journal literature at the U.S. National Institu Medicine (NIH/NLM).

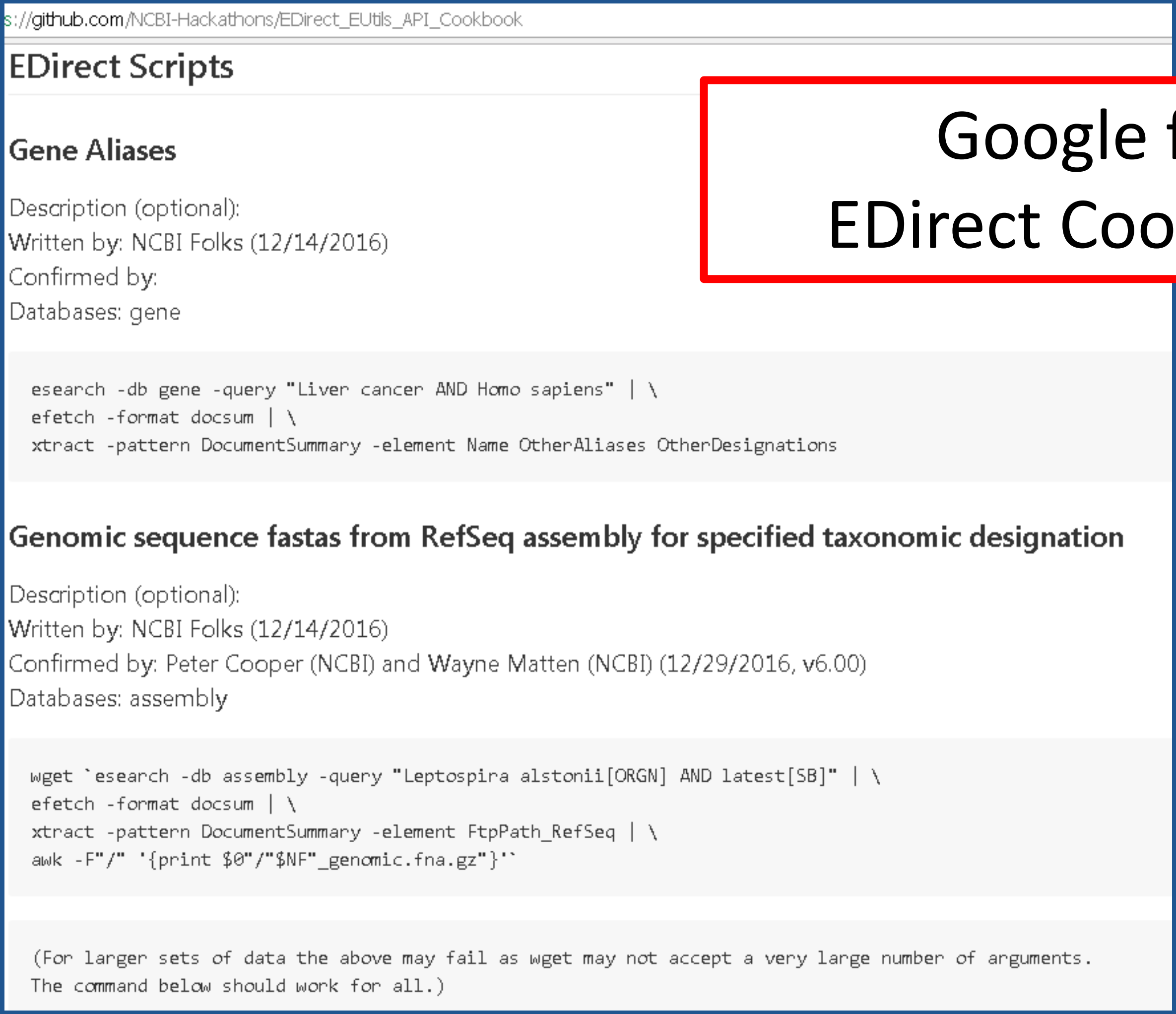
Get Started
PMC Overview
Users' Guide
Journal List
PMC FAQs
PMC Copyright Notice

Participate
Information for Publishers
Add a Journal to PMC
Participation Agreements
File Submission Specifications
File Validation Tools

Other Resources
PMC International

4.3 MILLION Artic
are archived in PMC.

EUtils (Search API) Command Line EDirect



s://github.com/NCBI-Hackathons/EDirect_EUtils_API_Cookbook

EDirect Scripts

Gene Aliases

Description (optional):
Written by: NCBI Folks (12/14/2016)
Confirmed by:
Databases: gene

```
esearch -db gene -query "Liver cancer AND Homo sapiens" | \
efetch -format docsum | \
xtract -pattern DocumentSummary -element Name OtherAliases OtherDesignations
```

Genomic sequence fastas from RefSeq assembly for specified taxonomic designation

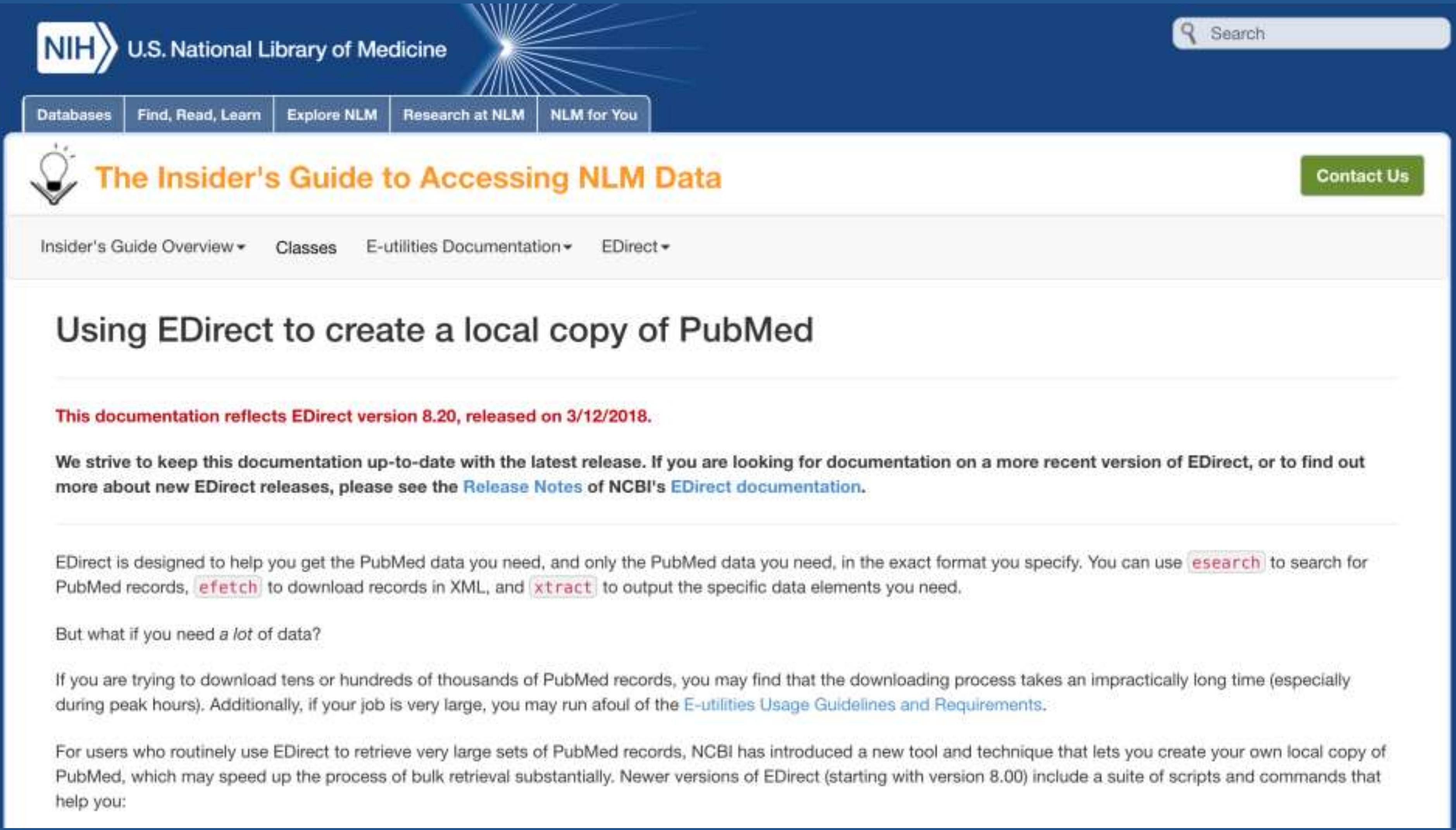
Description (optional):
Written by: NCBI Folks (12/14/2016)
Confirmed by: Peter Cooper (NCBI) and Wayne Matten (NCBI) (12/29/2016, v6.00)
Databases: assembly

```
wget `esearch -db assembly -query "Leptospira alstonii[ORGN] AND latest[SB]" | \
efetch -format docsum | \
xtract -pattern DocumentSummary -element FtpPath_RefSeq | \
awk -F"/" '{print $0"/"$NF"_genomic.fna.gz"}'`
```

```
(For larger sets of data the above may fail as wget may not accept a very large number of arguments.
The command below should work for all.)
```

Google for
EDirect Cookbook

EDirect Local Caching!



Using EDirect to create a local copy of PubMed

This documentation reflects EDirect version 8.20, released on 3/12/2018.

We strive to keep this documentation up-to-date with the latest release. If you are looking for documentation on a more recent version of EDirect, or to find out more about new EDirect releases, please see the Release Notes of NCBI's EDirect documentation.

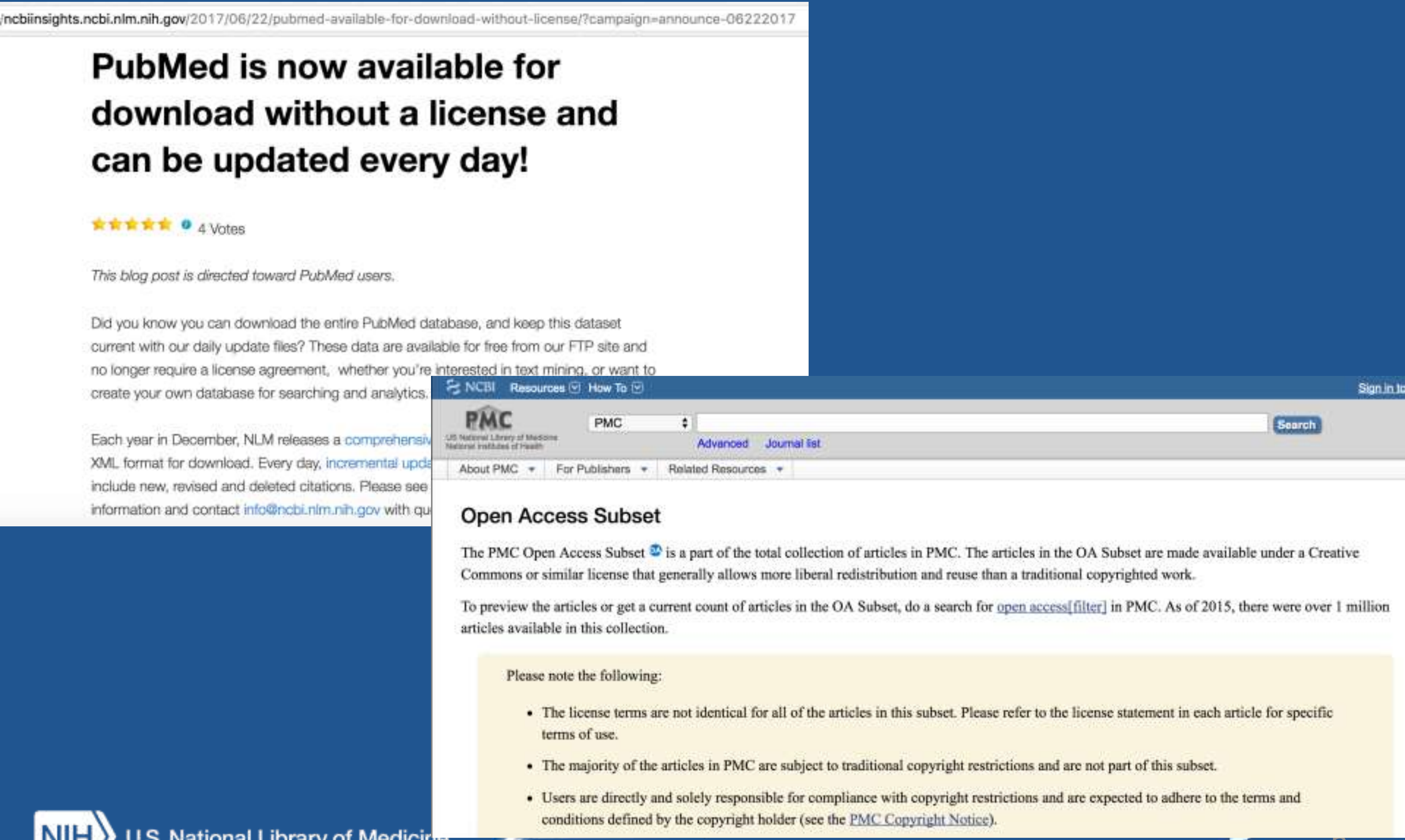
EDirect is designed to help you get the PubMed data you need, and only the PubMed data you need, in the exact format you specify. You can use `esearch` to search for PubMed records, `efetch` to download records in XML, and `xtract` to output the specific data elements you need.

But what if you need *a lot* of data?

If you are trying to download tens or hundreds of thousands of PubMed records, you may find that the downloading process takes an impractically long time (especially during peak hours). Additionally, if your job is very large, you may run afoul of the E-utilities Usage Guidelines and Requirements.

For users who routinely use EDirect to retrieve very large sets of PubMed records, NCBI has introduced a new tool and technique that lets you create your own local copy of PubMed, which may speed up the process of bulk retrieval substantially. Newer versions of EDirect (starting with version 8.00) include a suite of scripts and commands that help you:

PubMed and PMC (Open) FTP



/ncbiinsights.ncbi.nlm.nih.gov/2017/06/22/pubmed-available-for-download-without-license/?campaign=announce-06222017

PubMed is now available for download without a license and can be updated every day!

★★★★★ 4 Votes

This blog post is directed toward PubMed users.

Did you know you can download the entire PubMed database, and keep this dataset current with our daily update files? These data are available for free from our FTP site and no longer require a license agreement, whether you're interested in text mining, or want to create your own database for searching and analytics.

Each year in December, NLM releases a comprehensiv XML format for download. Every day, incremental upda include new, revised and deleted citations. Please see information and contact info@ncbi.nlm.nih.gov with qu

Open Access Subset

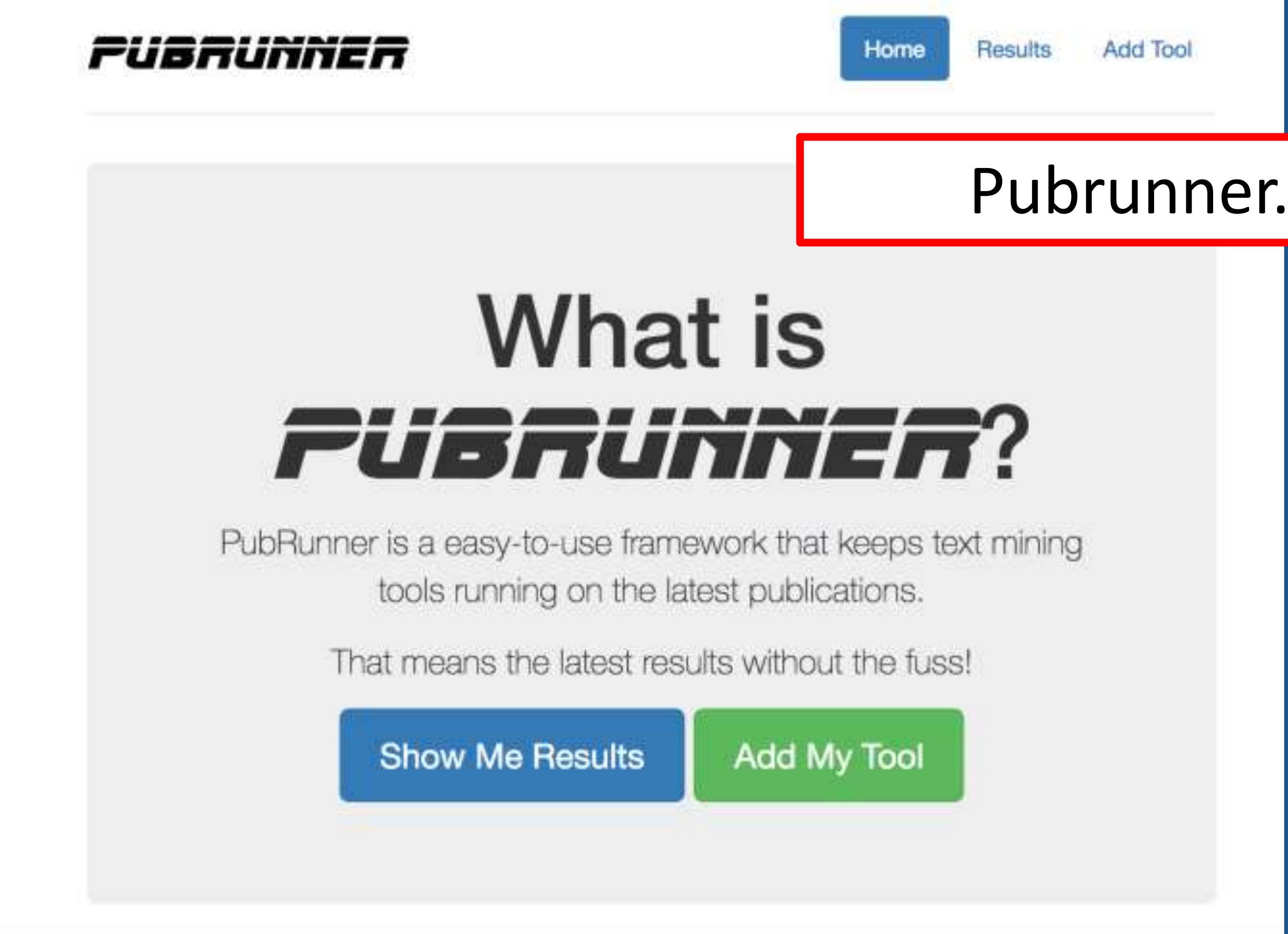
The PMC Open Access Subset is a part of the total collection of articles in PMC. The articles in the OA Subset are made available under a Creative Commons or similar license that generally allows more liberal redistribution and reuse than a traditional copyrighted work.

To preview the articles or get a current count of articles in the OA Subset, do a search for open access[filter] in PMC. As of 2015, there were over 1 million articles available in this collection.

Please note the following:

- The license terms are not identical for all of the articles in this subset. Please refer to the license statement in each article for specific terms of use.
- The majority of the articles in PMC are subject to traditional copyright restrictions and are not part of this subset.
- Users are directly and solely responsible for compliance with copyright restrictions and are expected to adhere to the terms and conditions defined by the copyright holder (see the PMC Copyright Notice).

Instead of PubMed FTP…
An automated tool (alpha)



PUBRUNNER

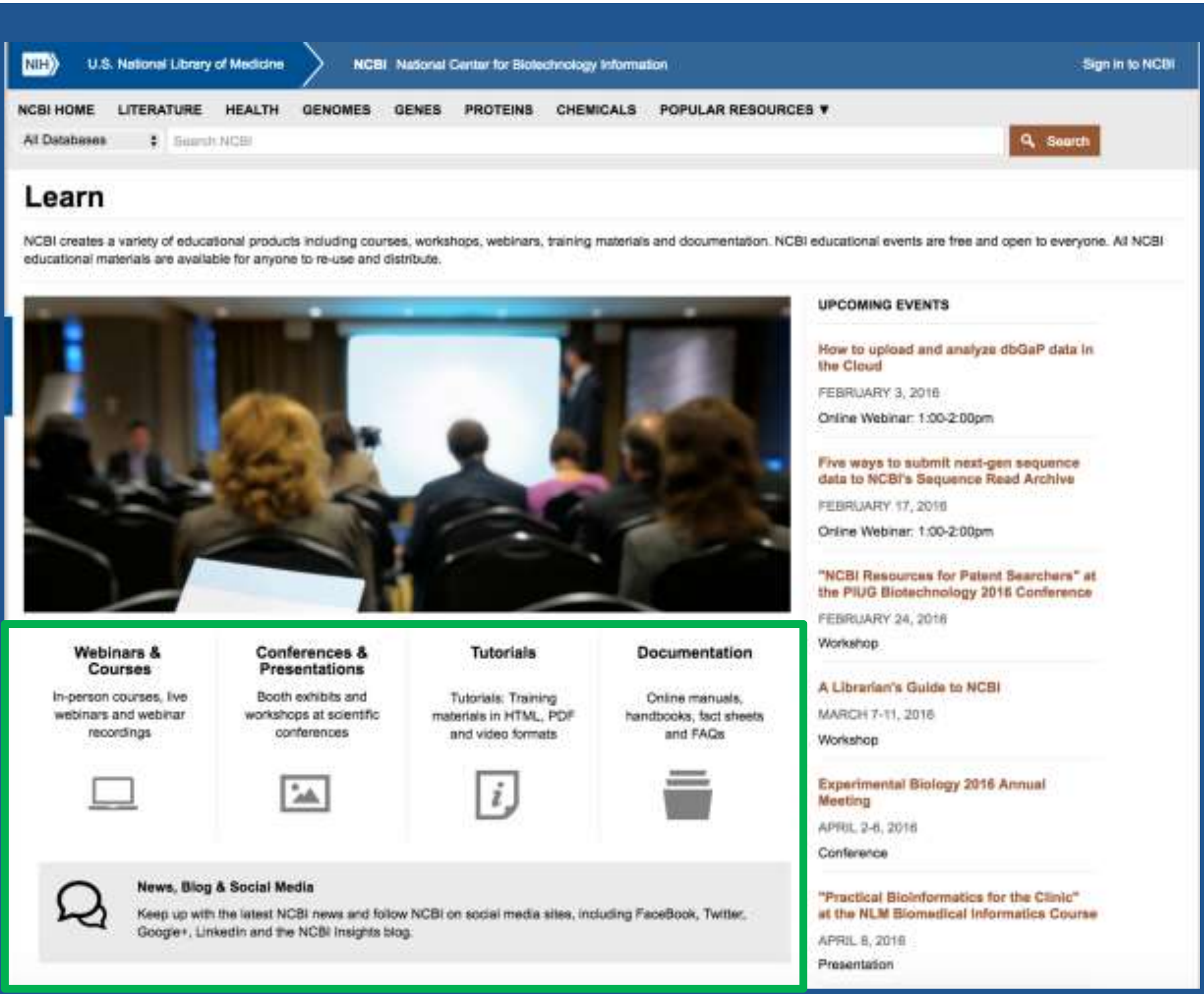
What is PUBRUNNER?

PubRunner is a easy-to-use framework that keeps text mining tools running on the latest publications.

That means the latest results without the fuss!

Pubrunner.org

For more information go to: ncbi.nlm.nih.gov/learn



NIH U.S. National Library of Medicine NCBI National Center for Biotechnology Information Sign in to NCBI

NCBI HOME LITERATURE HEALTH GENOMES GENES PROTEINS CHEMICALS POPULAR RESOURCES ▼

All Databases Search NCBI Search

Learn

NCBI creates a variety of educational products including courses, workshops, webinars, training materials and documentation. NCBI educational events are free and open to everyone. All NCBI educational materials are available for anyone to re-use and distribute.

Webinars & Courses
In-person courses, live webinars and webinar recordings

Conferences & Presentations
Booth exhibits and workshops at scientific conferences

Tutorials
Tutorials: Training materials in HTML, PDF and video formats

Documentation
Online manuals, handbooks, fact sheets and FAQs

News, Blog & Social Media
Keep up with the latest NCBI news and follow NCBI on social media sites, including FaceBook, Twitter, Google+, Linkedin and the NCBI Insights blog.

UPCOMING EVENTS

How to upload and analyze dbGaP data in the Cloud
FEBRUARY 3, 2016
Online Webinar: 1:00-2:00pm

Five ways to submit next-gen sequence data to NCBI's Sequence Read Archive
FEBRUARY 17, 2016
Online Webinar: 1:00-2:00pm

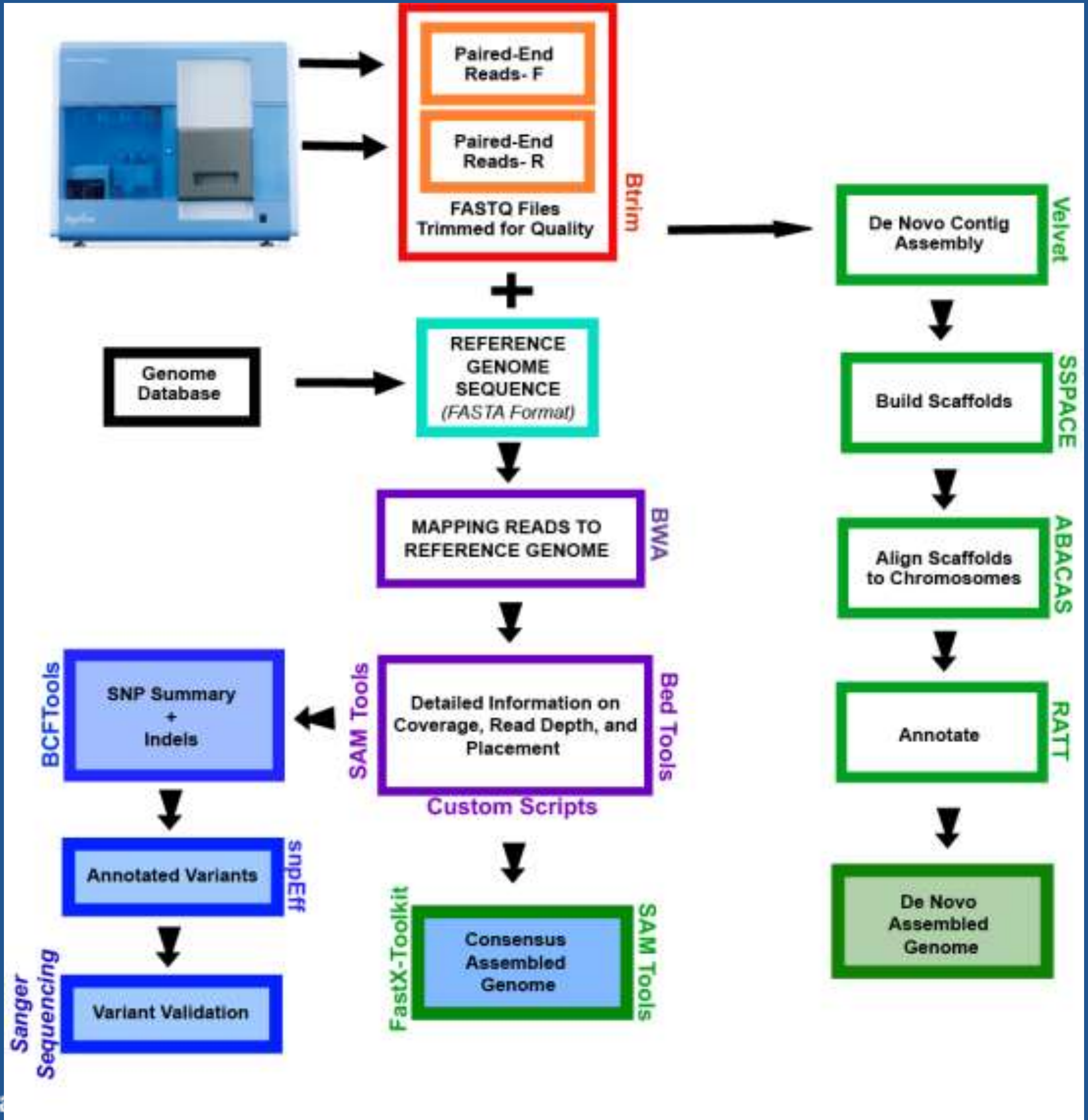
"NCBI Resources for Patent Searchers" at the PIUG Biotechnology 2016 Conference
FEBRUARY 24, 2016
Workshop

A Librarian's Guide to NCBI
MARCH 7-11, 2016
Workshop

Experimental Biology 2016 Annual Meeting
APRIL 2-6, 2016
Conference

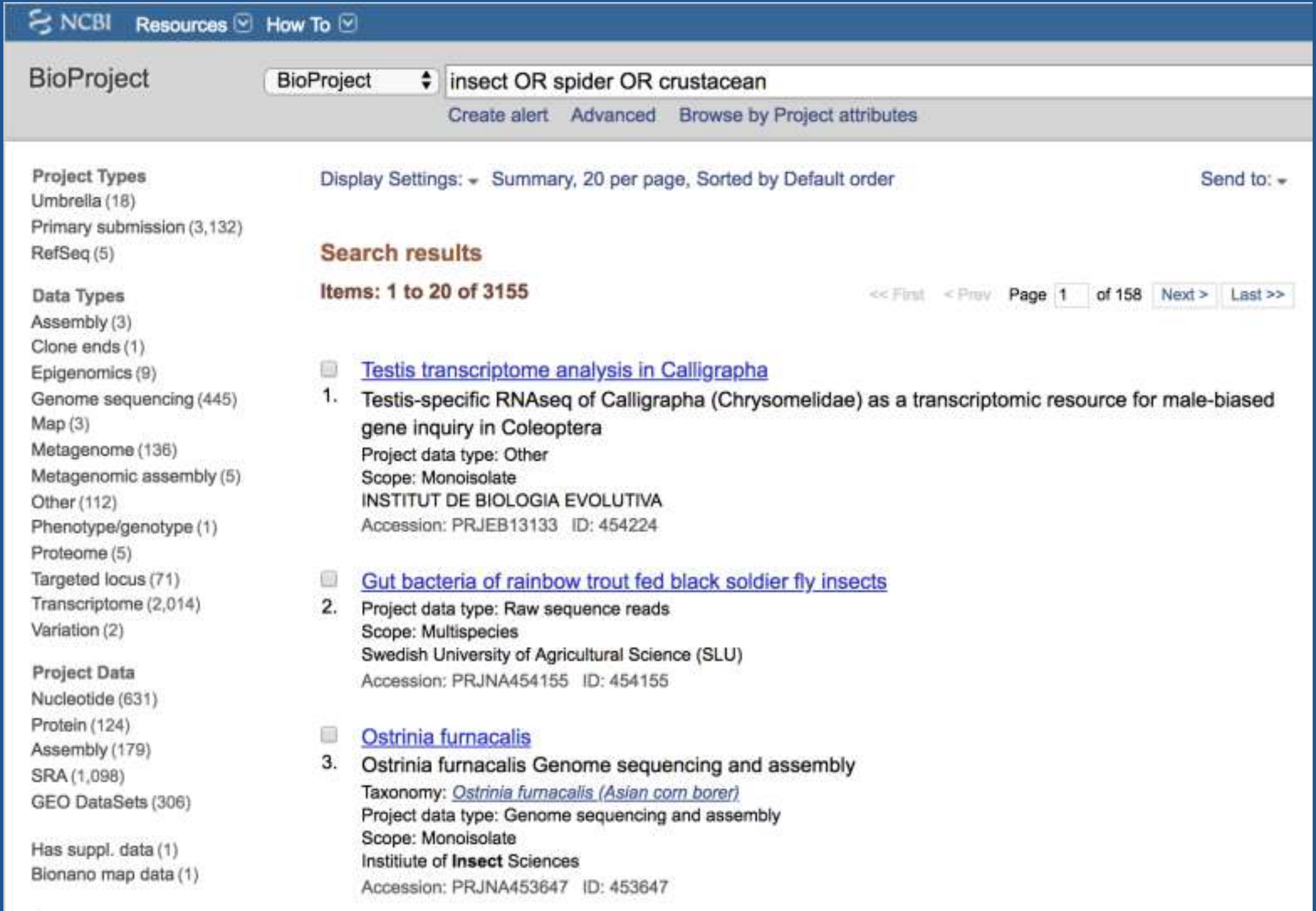
"Practical Bioinformatics for the Clinic" at the NLM Biomedical Informatics Course
APRIL 8, 2016
Presentation

NGS (c. 2013) in 90 Seconds for non-bioinformaticians



© Martine Zilversmit 2013

Cross-Data-Type Descriptors



BioProject

BioProject: insect OR spider OR crustacean

Project Types
Umbrella (18)
Primary submission (3,132)
RefSeq (5)

Data Types
Assembly (3)
Clone ends (1)
Epigenomics (9)
Genome sequencing (445)
Map (3)
Metagenome (136)
Metagenomic assembly (5)
Other (112)
Phenotype/genotype (1)
Proteome (5)
Targeted locus (71)
Transcriptome (2,014)
Variation (2)

Project Data
Nucleotide (631)
Protein (124)
Assembly (179)
SRA (1,098)
GEO DataSets (306)

Has suppl. data (1)
Bionano map data (1)

Display Settings: Summary, 20 per page, Sorted by Default order

Search results

Items: 1 to 20 of 3155

□ Testis transcriptome analysis in Calligrapha

1. Testis-specific RNAseq of Calligrapha (Chrysomelidae) as a transcriptomic resource for male-biased gene inquiry in Coleoptera
Project data type: Other
Scope: Monoisolate
INSTITUT DE BIOLOGIA EVOLUTIVA
Accession: PRJEB13133 ID: 454224

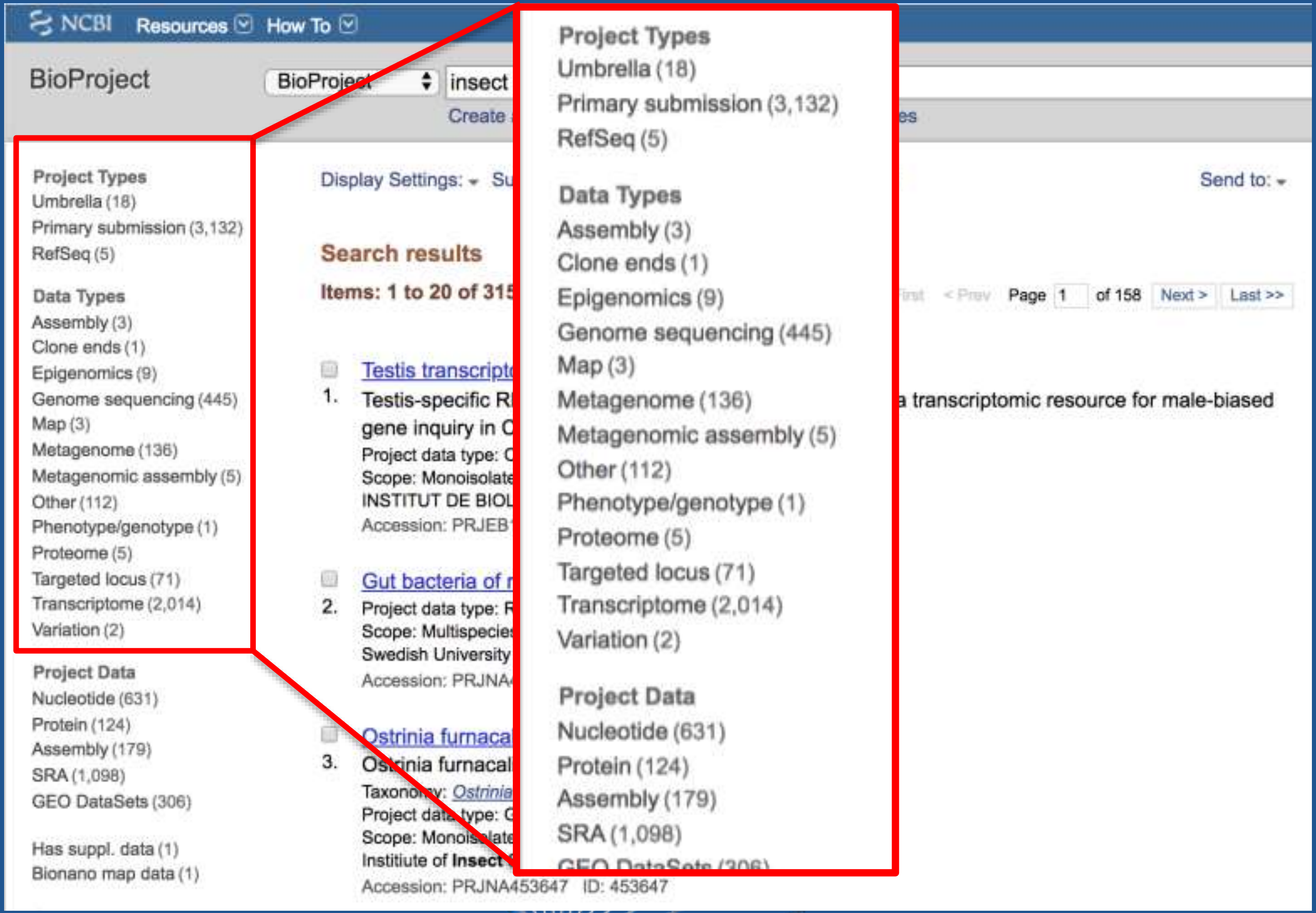
□ Gut bacteria of rainbow trout fed black soldier fly insects

2. Project data type: Raw sequence reads
Scope: Multispecies
Swedish University of Agricultural Science (SLU)
Accession: PRJNA454155 ID: 454155

□ Ostrinia furnacalis

3. Ostrinia furnacalis Genome sequencing and assembly
Taxonomy: *Ostrinia furnacalis (Asian corn borer)*
Project data type: Genome sequencing and assembly
Scope: Monoisolate
Instiitute of **Insect** Sciences
Accession: PRJNA453647 ID: 453647

Cross-Data-Type Descriptors



Project Types
Umbrella (18)
Primary submission (3,132)
RefSeq (5)

Data Types
Assembly (3)
Clone ends (1)
Epigenomics (9)
Genome sequencing (445)
Map (3)
Metagenome (136)
Metagenomic assembly (5)
Other (112)
Phenotype/genotype (1)
Proteome (5)
Targeted locus (71)
Transcriptome (2,014)
Variation (2)

Project Data
Nucleotide (631)
Protein (124)
Assembly (179)
SRA (1,098)
GEO DataSets (306)

Has suppl. data (1)
Bionano map data (1)

Search results

Items: 1 to 20 of 315

□ Testis transcript

1. Testis-specific R
gene inquiry in C
a transcriptomic resource for male-biased
Project data type: C
Scope: Monoisolate
INSTITUT DE BIOL
Accession: PRJEB1

□ Gut bacteria of r

2. Project data type: R
Scope: Multispecies
Swedish University
Accession: PRJNA

□ Ostrinia furnaca

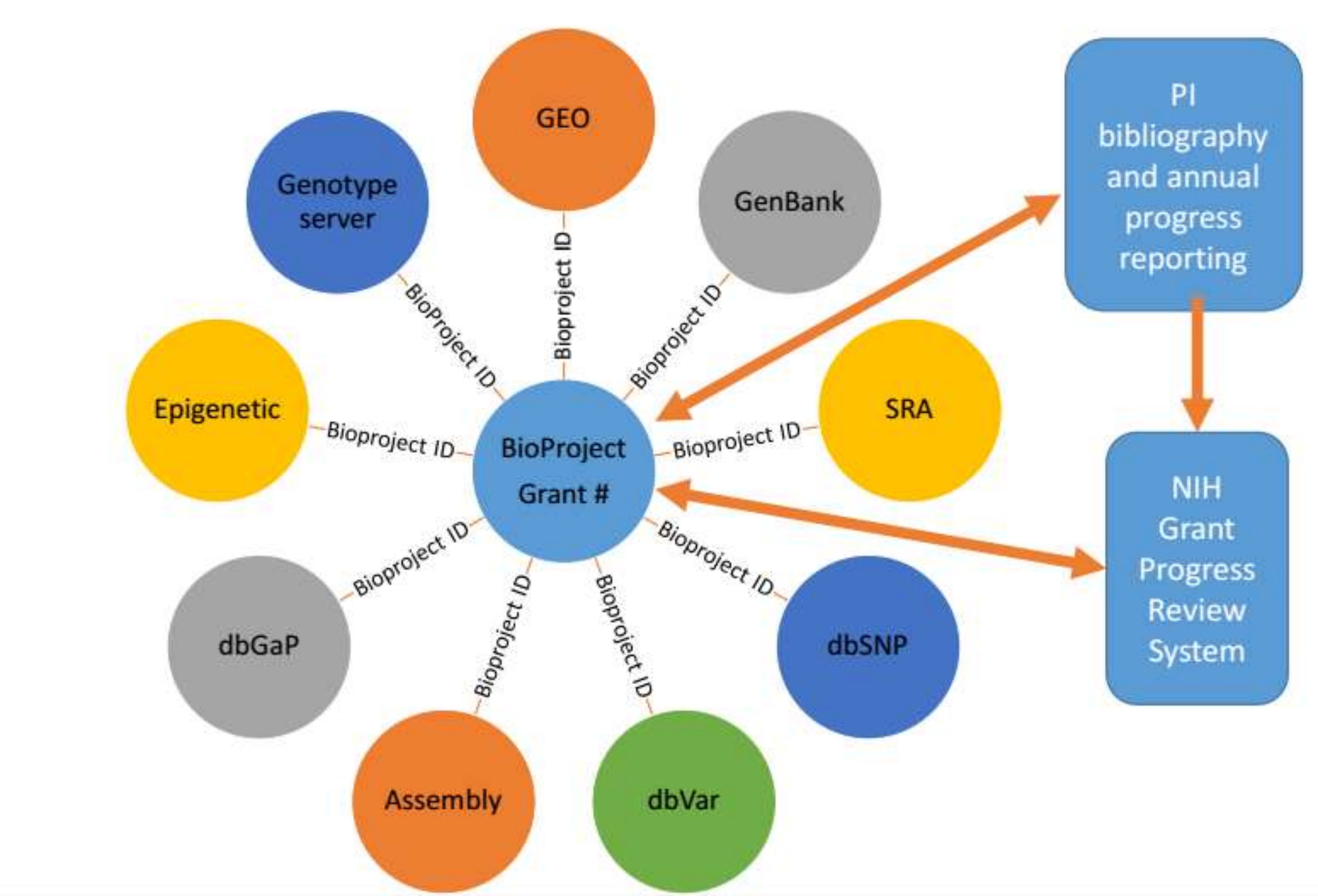
3. Ostrinia furnacal
Taxonomy: Ostrinia
Project data type: G
Scope: Monoisolate
Institute of Insect
Accession: PRJNA453647 ID: 453647

Project Types
Umbrella (18)
Primary submission (3,132)
RefSeq (5)

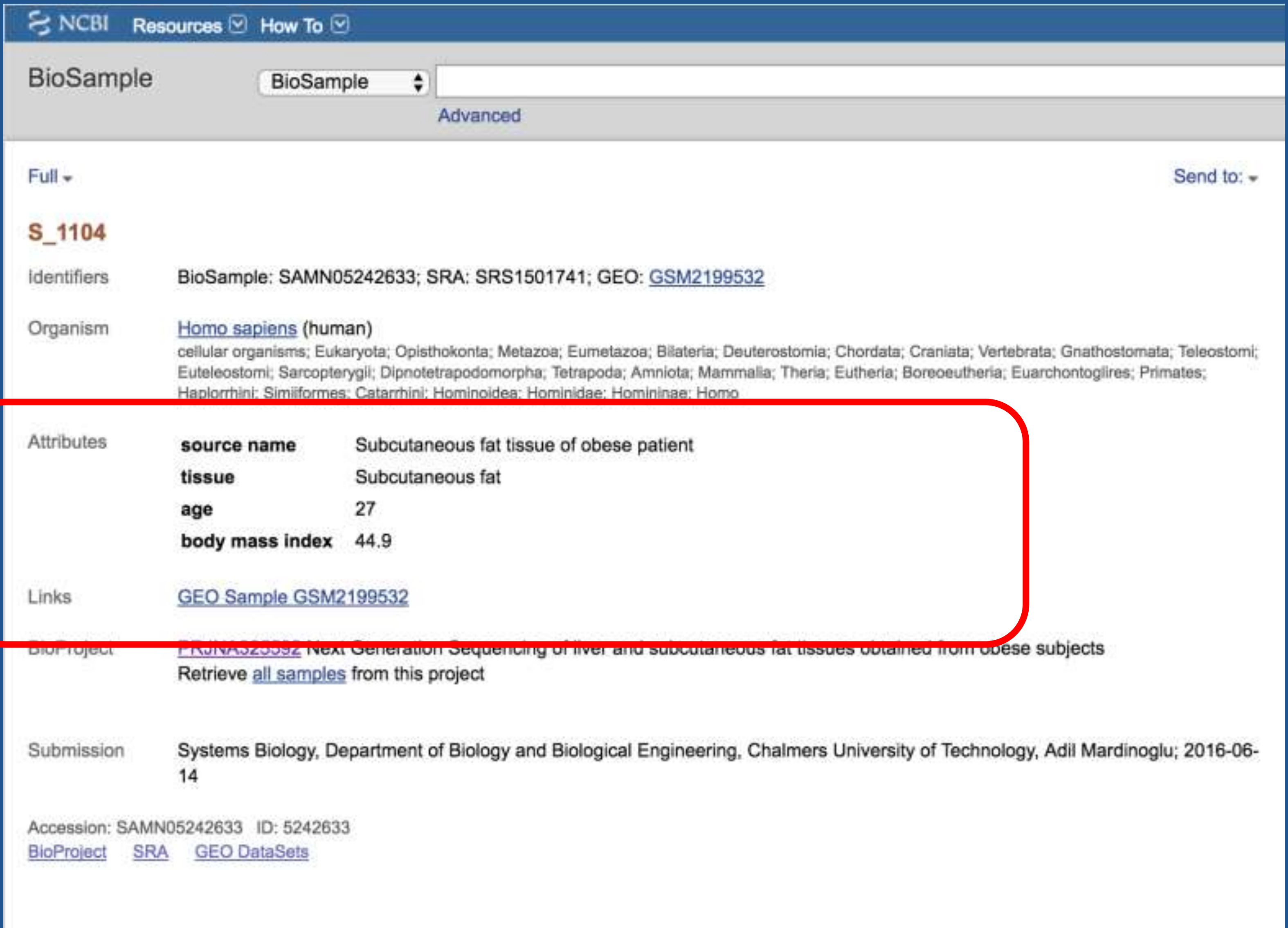
Data Types
Assembly (3)
Clone ends (1)
Epigenomics (9)
Genome sequencing (445)
Map (3)
Metagenome (136)
Metagenomic assembly (5)
Other (112)
Phenotype/genotype (1)
Proteome (5)
Targeted locus (71)
Transcriptome (2,014)
Variation (2)

Project Data
Nucleotide (631)
Protein (124)
Assembly (179)
SRA (1,098)
GEO DataSets (306)

Reporting


GEO
Genotype server
GenBank
PI bibliography and annual progress reporting
Epigenetic
BioProject Grant #
SRA
NIH Grant Progress Review System
dbGaP
dbSNP
Assembly
dbVar
BioProject ID
Bioproject ID
Bioproject ID
Bioproject ID
Bioproject ID
Bioproject ID
Bioproject ID
Bioproject ID
Bioproject ID

BioSample



S_1104

Identifiers BioSample: SAMN05242633; SRA: SRS1501741; GEO: GSM2199532

Organism Homo sapiens (human)
cellular organisms; Eukaryota; Opisthokonta; Metazoa; Eumetazoa; Bilateria; Deuterostomia; Chordata; Craniata; Vertebrata; Gnathostomata; Teleostomi; Euteleostomi; Sarcopterygii; Dipnotetrapodomorpha; Tetrapoda; Amniota; Mammalia; Theria; Eutheria; Boreoeutheria; Euarchontoglires; Primates; Haplorrhini; Simiiformes; Catarrhini; Hominoidea; Hominidae; Homininae; Homo

Attributes

source name	Subcutaneous fat tissue of obese patient
tissue	Subcutaneous fat
age	27
body mass index	44.9

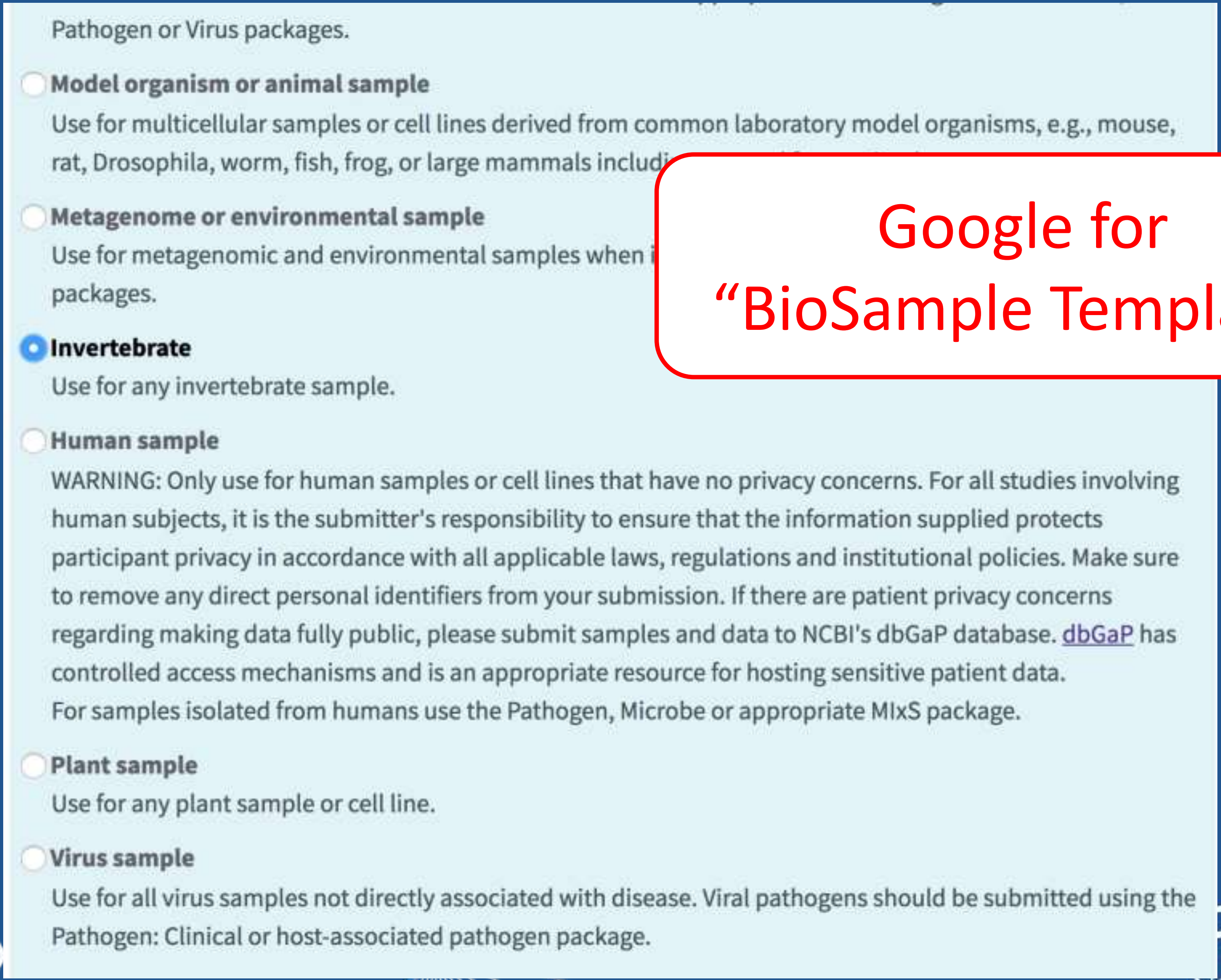
Links GEO Sample GSM2199532

BioProject PRJNA325592 Next Generation Sequencing of liver and subcutaneous fat tissues obtained from obese subjects
Retrieve all samples from this project

Submission Systems Biology, Department of Biology and Biological Engineering, Chalmers University of Technology, Adil Mardinoglu; 2016-06-14

Accession: SAMN05242633 ID: 5242633
BioProject SRA GEO DataSets

Labels that can be used!



Pathogen or Virus packages.

○ **Model organism or animal sample**

Use for multicellular samples or cell lines derived from common laboratory model organisms, e.g., mouse, rat, Drosophila, worm, fish, frog, or large mammals includi

○ **Metagenome or environmental sample**

Use for metagenomic and environmental samples when i packages.

● **Invertebrate**

Use for any invertebrate sample.

○ **Human sample**

WARNING: Only use for human samples or cell lines that have no privacy concerns. For all studies involving human subjects, it is the submitter's responsibility to ensure that the information supplied protects participant privacy in accordance with all applicable laws, regulations and institutional policies. Make sure to remove any direct personal identifiers from your submission. If there are patient privacy concerns regarding making data fully public, please submit samples and data to NCBI's dbGaP database. dbGaP has controlled access mechanisms and is an appropriate resource for hosting sensitive patient data.
For samples isolated from humans use the Pathogen, Microbe or appropriate MIxS package.

○ **Plant sample**

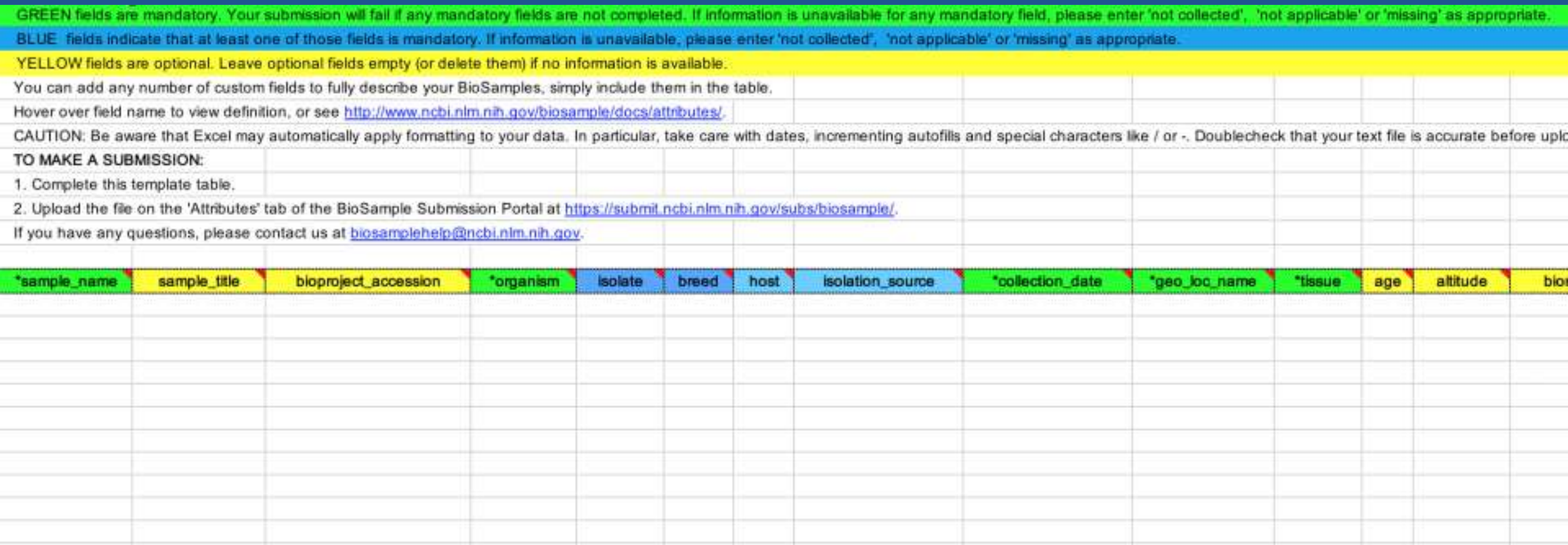
Use for any plant sample or cell line.

○ **Virus sample**

Use for all virus samples not directly associated with disease. Viral pathogens should be submitted using the Pathogen: Clinical or host-associated pathogen package.

Google for
"BioSample Template"

BioSample



GREEN fields are mandatory. Your submission will fail if any mandatory fields are not completed. If information is unavailable for any mandatory field, please enter 'not collected', 'not applicable' or 'missing' as appropriate.

BLUE fields indicate that at least one of those fields is mandatory. If infomation is unavailable, please enter 'not collected', 'not applicable' or 'missing' as appropriate.

YELLOW fields are optional. Leave optional fields empty (or delete them) if no information is available.

You can add any number of custom fields to fully describe your BioSamples, simply include them in the table.

Hover over field name to view definition, or see http://www.ncbi.nlm.nih.gov/biosample/docs/attributes/.

CAUTION: Be aware that Excel may automatically apply formatting to your data. In particular, take care with dates, incrementing autofills and special characters like / or -. Doublecheck that your text file is accurate before uplo

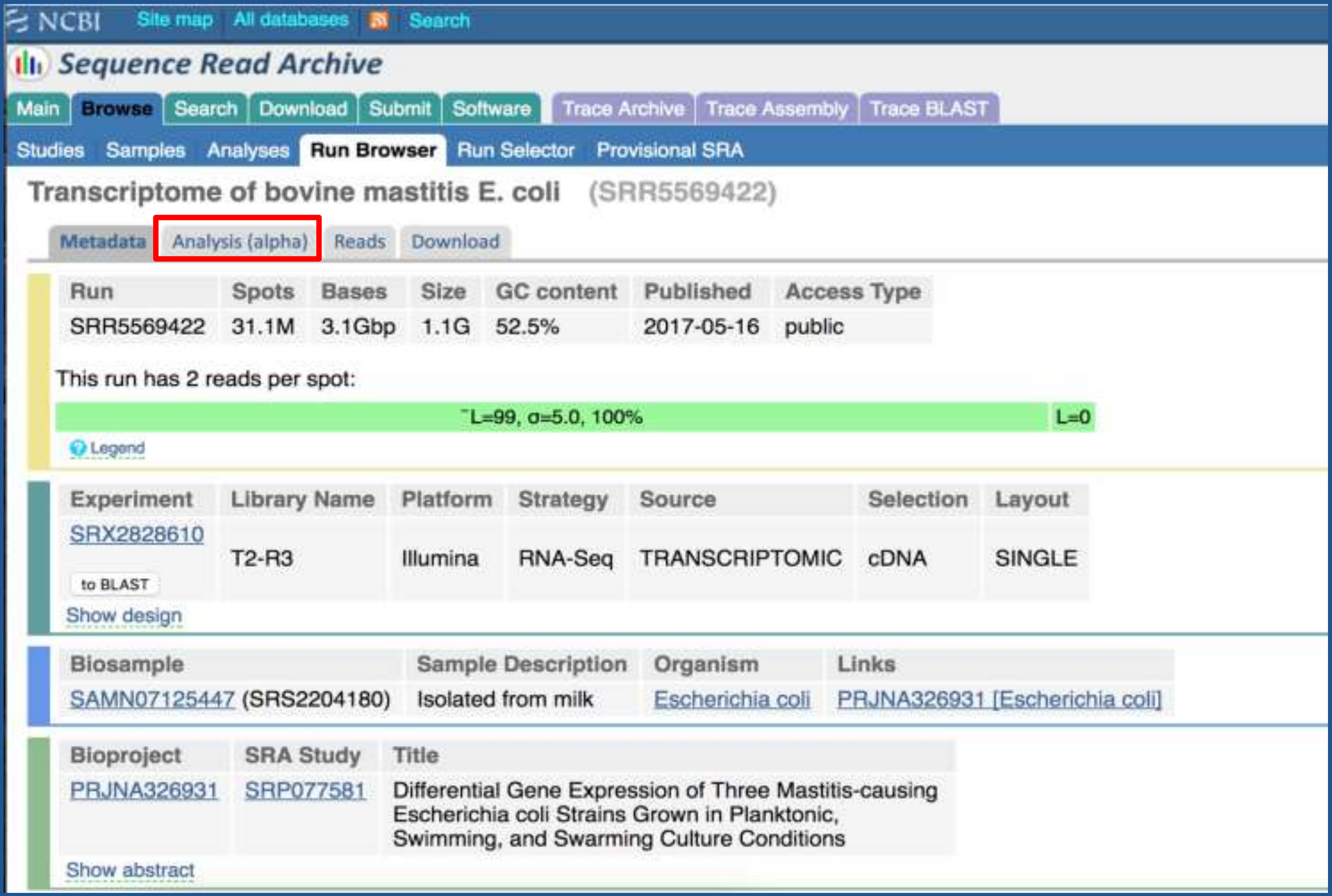
TO MAKE A SUBMISSION:

1. Complete this template table.
2. Upload the file on the 'Attributes' tab of the BioSample Submission Portal at https://submit.ncbi.nlm.nih.gov/subs/biosample/.

If you have any questions, please contact us at biosamplehelp@ncbi.nlm.nih.gov.

*sample_name	sample_title	bioproject_accession	*organism	isolate	breed	host	isolation_source	*collection_date	*geo_loc_name	*tissue	age	altitude	bio

But wait, how do I find data if the metadata is insufficient?



Transcriptome of bovine mastitis E. coli (SRR5569422)

Run	Spots	Bases	Size	GC content	Published	Access Type
SRR5569422	31.1M	3.1Gbp	1.1G	52.5%	2017-05-16	public

This run has 2 reads per spot:

¯L=99, σ=5.0, 100% L=0

Legend

Experiment	Library Name	Platform	Strategy	Source	Selection	Layout
SRX2828610 to BLAST	T2-R3	Illumina	RNA-Seq	TRANSCRIPTOMIC	cDNA	SINGLE

Show design

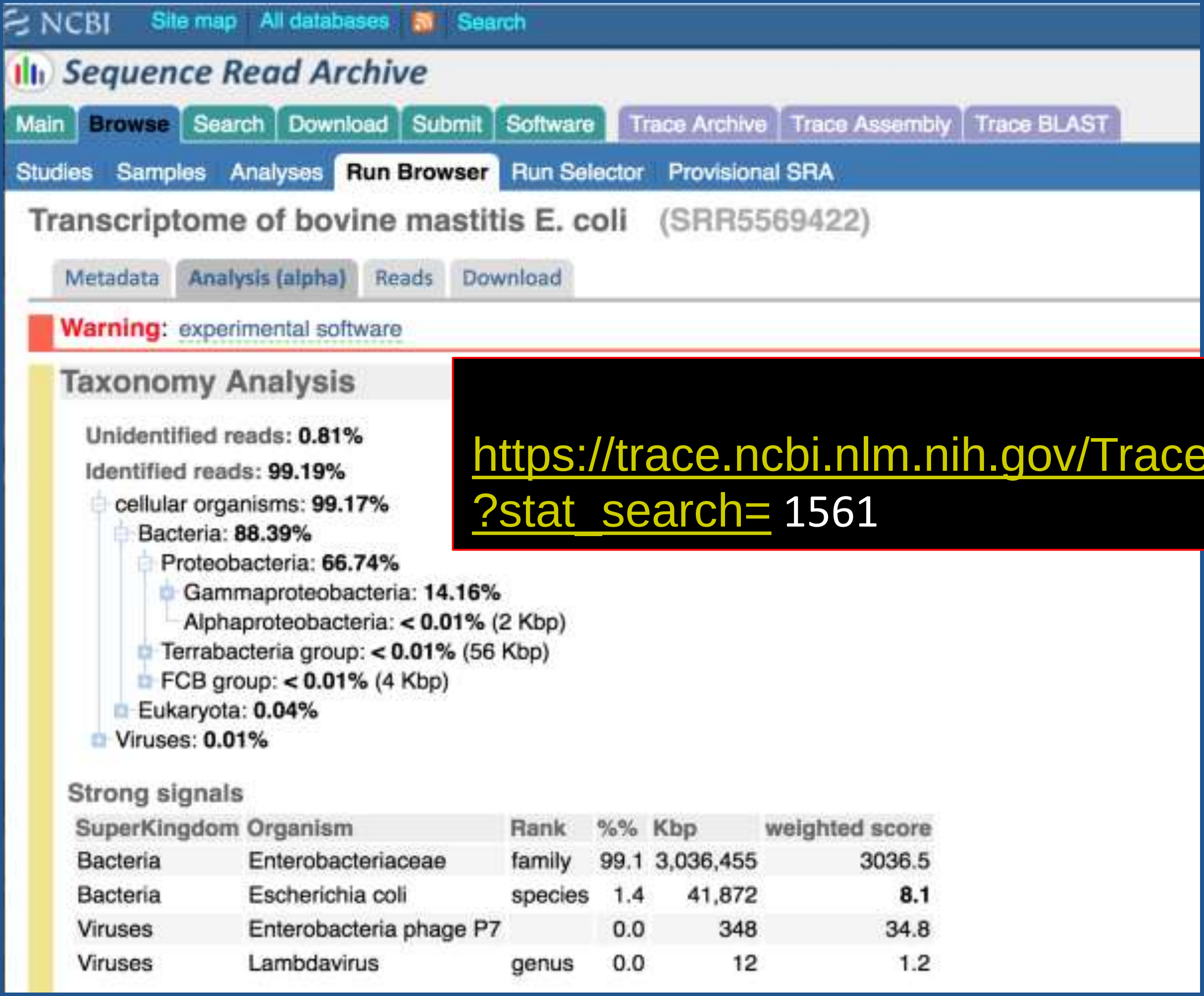
Biosample	Sample Description	Organism	Links
SAMN07125447 (SRS2204180)	Isolated from milk	Escherichia coli	PRJNA326931 [Escherichia coli]

Bioproject	SRA Study	Title
PRJNA326931	SRP077581	Differential Gene Expression of Three Mastitis-causing Escherichia coli Strains Grown in Planktonic, Swimming, and Swarming Culture Conditions

Show abstract

Now all of SRA is taxonomically indexed!

COOL THING #1 !



Sequence Read Archive

Transcriptome of bovine mastitis E. coli (SRR5569422)

Warning: experimental software

Taxonomy Analysis

Unidentified reads: **0.81%**

Identified reads: **99.19%**

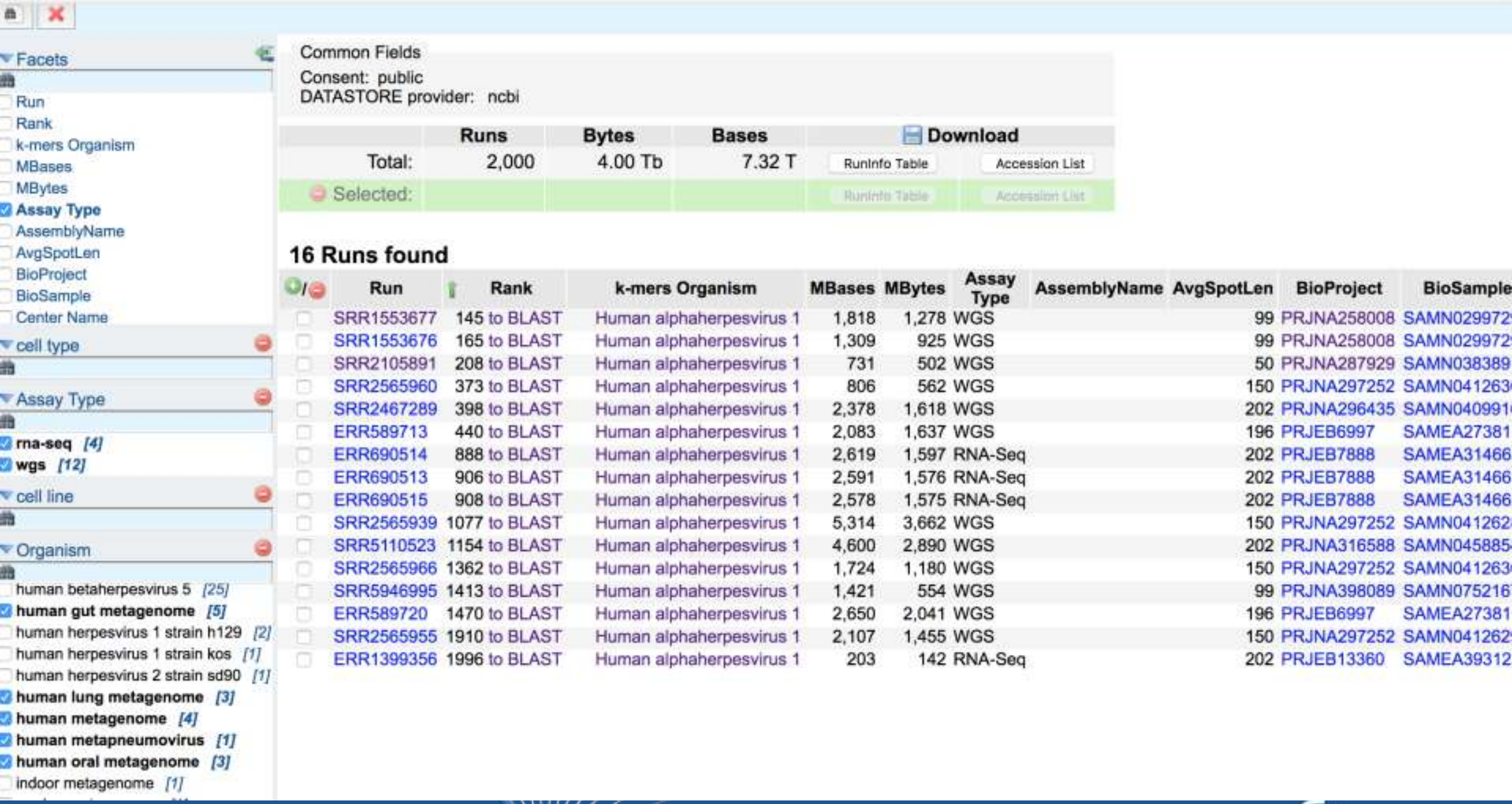
- cellular organisms: **99.17%**
  - Bacteria: **88.39%**
    - Proteobacteria: **66.74%**
      - Gammaproteobacteria: **14.16%**
      - Alphaproteobacteria: **< 0.01%** (2 Kbp)
    - Terrabacteria group: **< 0.01%** (56 Kbp)
    - FCB group: **< 0.01%** (4 Kbp)
  - Eukaryota: **0.04%**
- Viruses: **0.01%**

Strong signals

SuperKingdom	Organism	Rank	%%	Kbp	weighted score
Bacteria	Enterobacteriaceae	family	99.1	3,036,455	3036.5
Bacteria	Escherichia coli	species	1.4	41,872	**8.1**
Viruses	Enterobacteria phage P7		0.0	348	34.8
Viruses	Lambdavirus	genus	0.0	12	1.2

https://trace.ncbi.nlm.nih.gov/Traces/study/?stat_search= 1561

Now all of SRA is taxonomically indexed!



Facets
□ Run
□ Rank
□ k-mers Organism
□ MBases
□ MBytes
☑ Assay Type
□ AssemblyName
□ AvgSpotLen
□ BioProject
□ BioSample
□ Center Name
cell type
Assay Type
☑ rna-seq [4]
☑ wgs [12]
cell line
Organism
□ human betaherpesvirus 5 [25]
☑ human gut metagenome [5]
□ human herpesvirus 1 strain h129 [2]
□ human herpesvirus 1 strain kos [1]
□ human herpesvirus 2 strain sd90 [1]
☑ human lung metagenome [3]
☑ human metagenome [4]
☑ human metapneumovirus [1]
☑ human oral metagenome [3]
□ indoor metagenome [1]

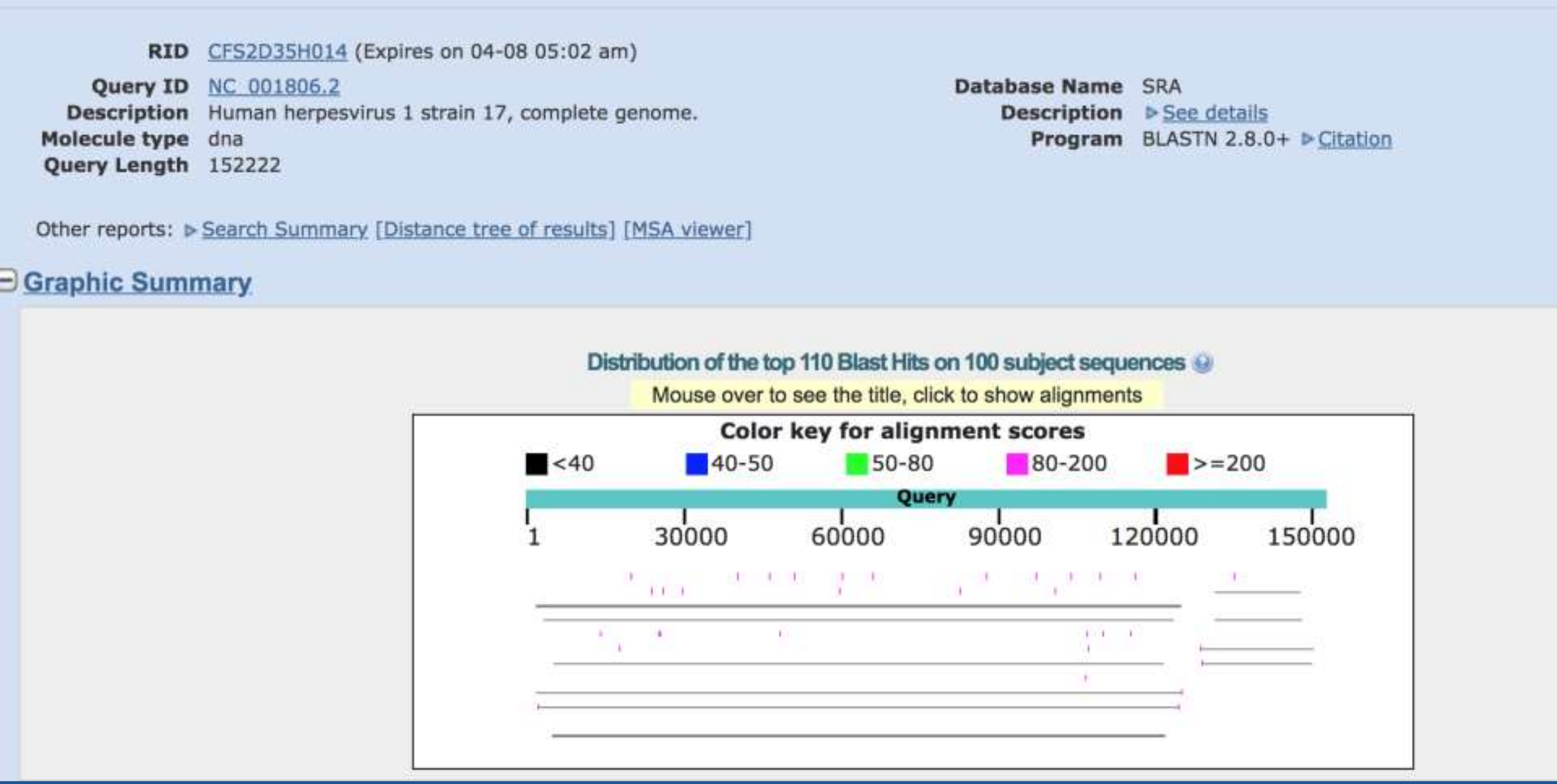
Common Fields
Consent: public
DATASTORE provider: ncbi

	Runs	Bytes	Bases	Download	
Total:	2,000	4.00 Tb	7.32 T	RunInfo Table	Accession List
Selected:				RunInfo Table	Accession List

16 Runs found

⊕/⊖	Run	Rank	k-mers Organism	MBases	MBytes	Assay Type	AssemblyName	AvgSpotLen	BioProject	BioSample
□	SRR1553677	145 to BLAST	Human alphaherpesvirus 1	1,818	1,278	WGS		99	PRJNA258008	SAMN0299972
□	SRR1553676	165 to BLAST	Human alphaherpesvirus 1	1,309	925	WGS		99	PRJNA258008	SAMN0299972
□	SRR2105891	208 to BLAST	Human alphaherpesvirus 1	731	502	WGS		50	PRJNA287929	SAMN038389
□	SRR2565960	373 to BLAST	Human alphaherpesvirus 1	806	562	WGS		150	PRJNA297252	SAMN041263
□	SRR2467289	398 to BLAST	Human alphaherpesvirus 1	2,378	1,618	WGS		202	PRJNA296435	SAMN040991
□	ERR589713	440 to BLAST	Human alphaherpesvirus 1	2,083	1,637	WGS		196	PRJEB6997	SAMEA27381
□	ERR690514	888 to BLAST	Human alphaherpesvirus 1	2,619	1,597	RNA-Seq		202	PRJEB7888	SAMEA31466
□	ERR690513	906 to BLAST	Human alphaherpesvirus 1	2,591	1,576	RNA-Seq		202	PRJEB7888	SAMEA31466
□	ERR690515	908 to BLAST	Human alphaherpesvirus 1	2,578	1,575	RNA-Seq		202	PRJEB7888	SAMEA31466
□	SRR2565939	1077 to BLAST	Human alphaherpesvirus 1	5,314	3,662	WGS		150	PRJNA297252	SAMN041262
□	SRR5110523	1154 to BLAST	Human alphaherpesvirus 1	4,600	2,890	WGS		202	PRJNA316588	SAMN045885
□	SRR2565966	1362 to BLAST	Human alphaherpesvirus 1	1,724	1,180	WGS		150	PRJNA297252	SAMN041263
□	SRR5946995	1413 to BLAST	Human alphaherpesvirus 1	1,421	554	WGS		99	PRJNA398089	SAMN075216
□	ERR589720	1470 to BLAST	Human alphaherpesvirus 1	2,650	2,041	WGS		196	PRJEB6997	SAMEA27381
□	SRR2565955	1910 to BLAST	Human alphaherpesvirus 1	2,107	1,455	WGS		150	PRJNA297252	SAMN041262
□	ERR1399356	1996 to BLAST	Human alphaherpesvirus 1	203	142	RNA-Seq		202	PRJEB13360	SAMEA39312

Now all of SRA is taxonomically indexed!



RID CFS2D35H014 (Expires on 04-08 05:02 am)

Query ID NC_001806.2
Description Human herpesvirus 1 strain 17, complete genome.
Molecule type dna
Query Length 152222

Database Name SRA
Description ▷ See details
Program BLASTN 2.8.0+ ▷ Citation

Other reports: ▷ Search Summary [Distance tree of results] [MSA viewer]

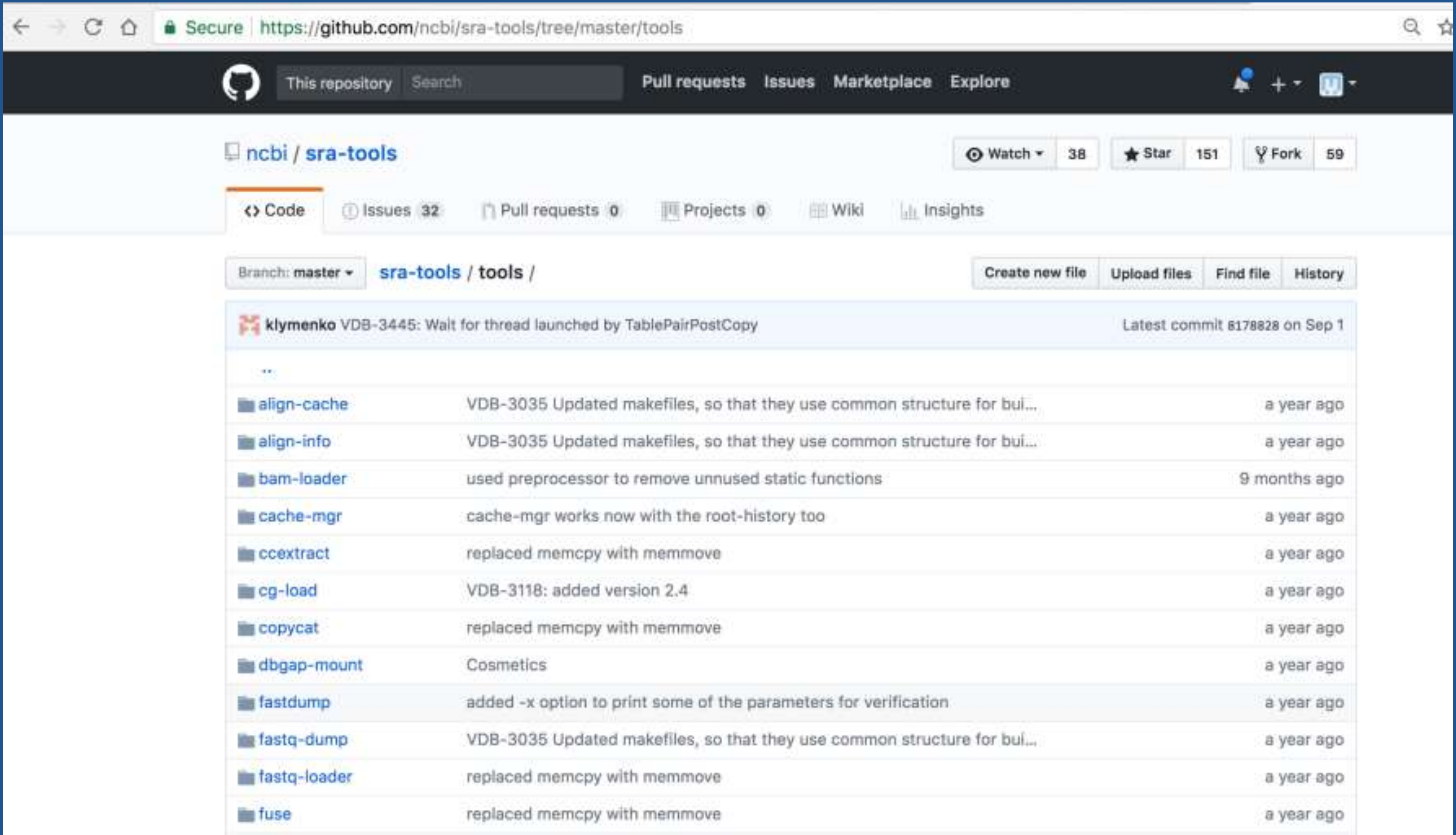
⊟ Graphic Summary

Distribution of the top 110 Blast Hits on 100 subject sequences

Mouse over to see the title, click to show alignments

But how do I extract the data?

That said, raw data is messy and hard to dump (plus its huge)



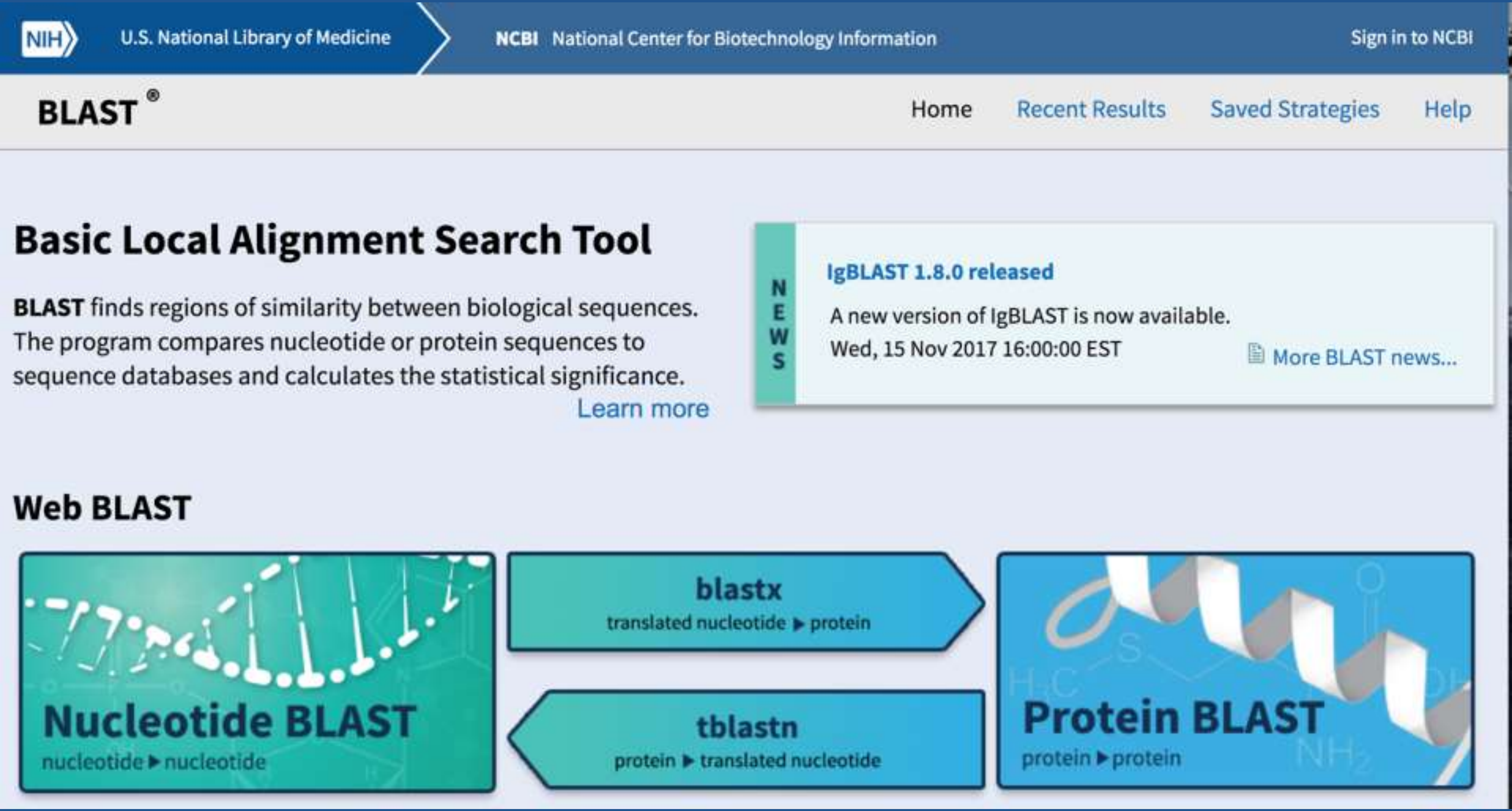
klymenko VDB-3445: Wait for thread launched by TablePairPostCopy — Latest commit 6178828 on Sep 1

..

align-cache	VDB-3035 Updated makefiles, so that they use common structure for bui…	a year ago
align-info	VDB-3035 Updated makefiles, so that they use common structure for bui…	a year ago
bam-loader	used preprocessor to remove unnused static functions	9 months ago
cache-mgr	cache-mgr works now with the root-history too	a year ago
ccextract	replaced memcpy with memmove	a year ago
cg-load	VDB-3118: added version 2.4	a year ago
copycat	replaced memcpy with memmove	a year ago
dbgap-mount	Cosmetics	a year ago
fastdump	added -x option to print some of the parameters for verification	a year ago
fastq-dump	VDB-3035 Updated makefiles, so that they use common structure for bui…	a year ago
fastq-loader	replaced memcpy with memmove	a year ago
fuse	replaced memcpy with memmove	a year ago

U.S. National Library of Medicine

So why do that?


NIH
U.S. National Library of Medicine
NCBI National Center for Biotechnology Information
Sign in to NCBI
BLAST ®
Home
Recent Results
Saved Strategies
Help

Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.

Learn more

NEWS

IgBLAST 1.8.0 released

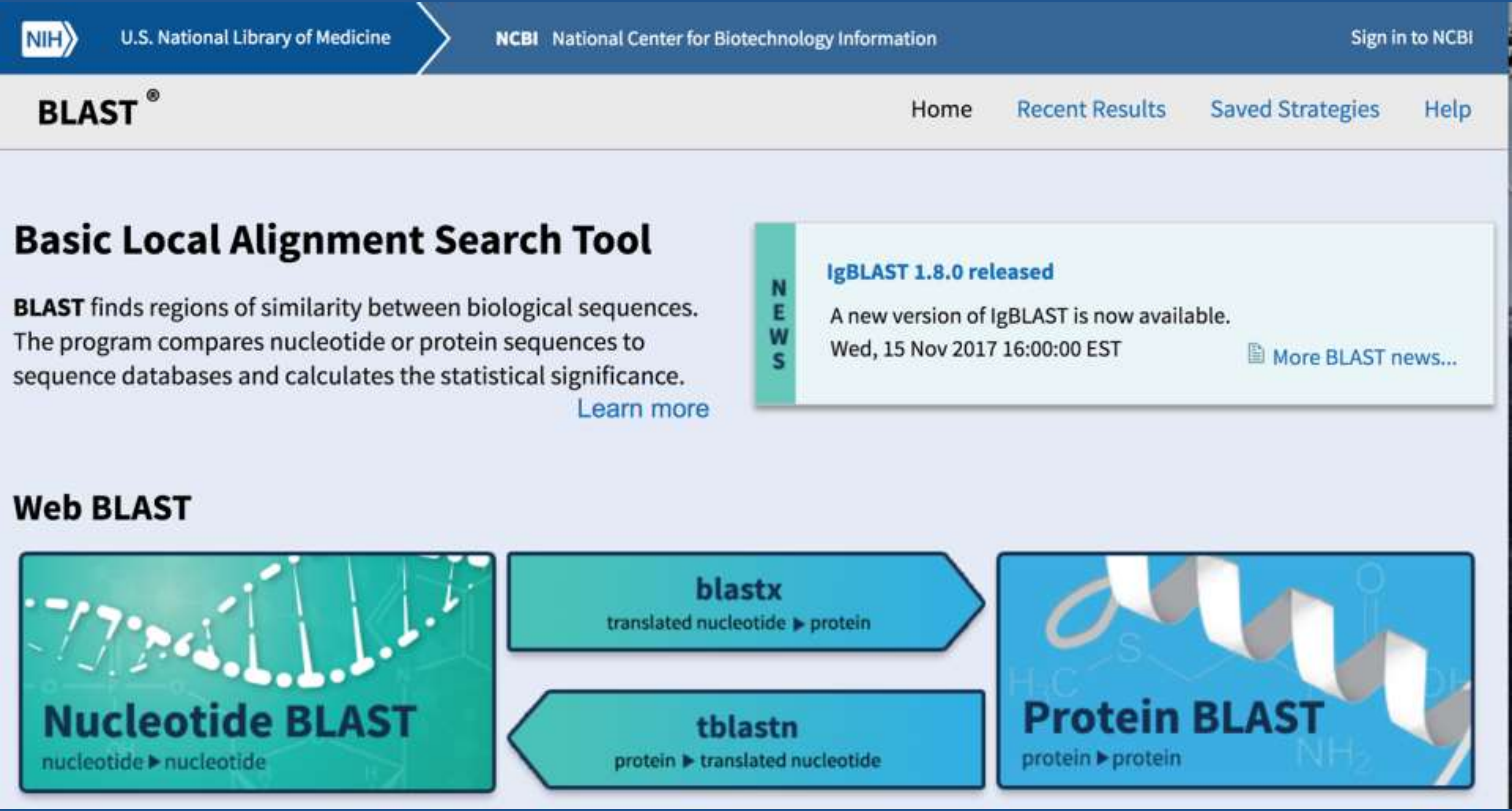
A new version of IgBLAST is now available.
Wed, 15 Nov 2017 16:00:00 EST

More BLAST news...

Web BLAST

Nucleotide BLAST
nucleotide ▶ nucleotide
blastx
translated nucleotide ▶ protein
tblastn
protein ▶ translated nucleotide
Protein BLAST
protein ▶ protein

Wait, what, that's too simple/primitive?



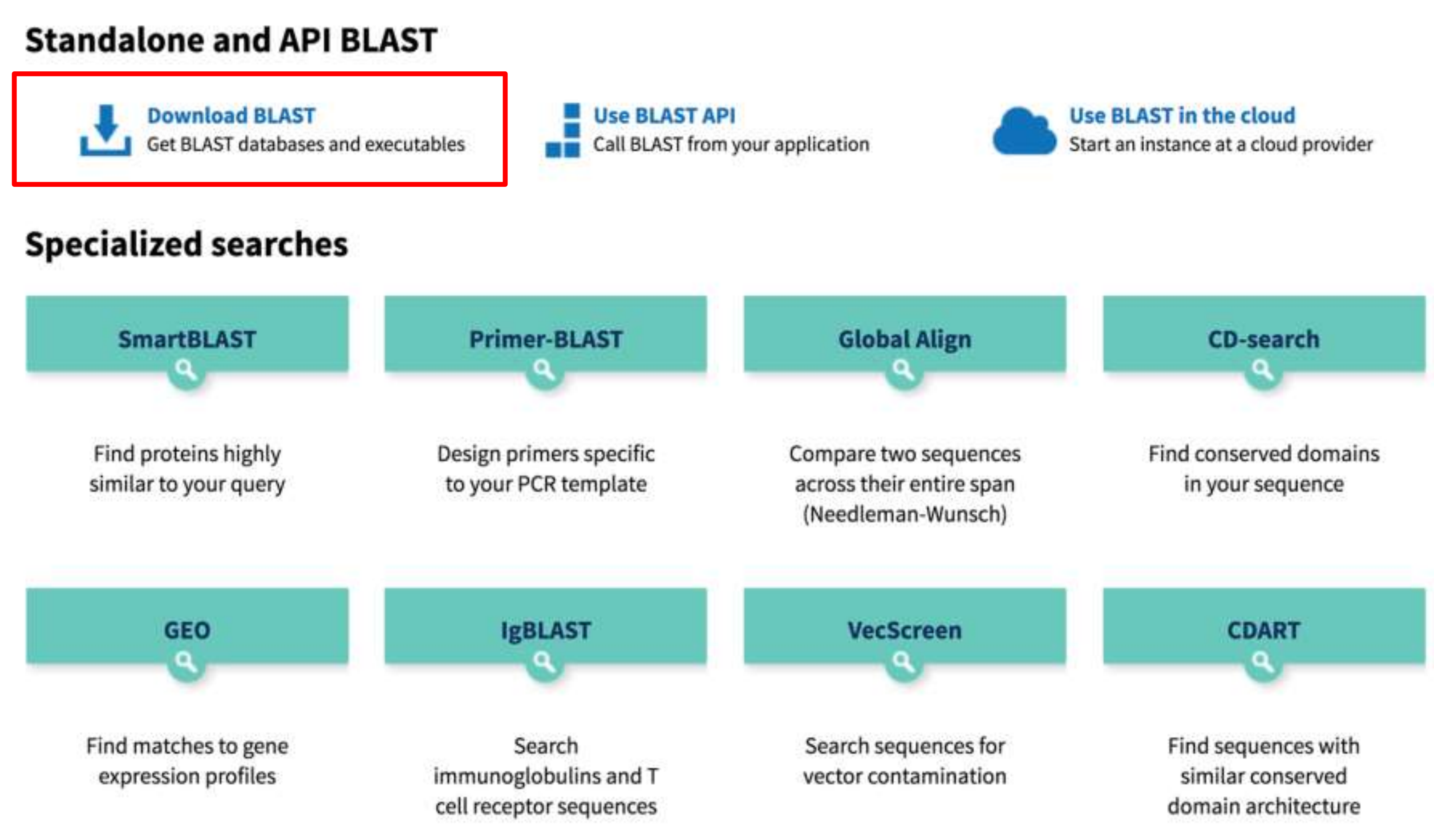
Basic Local Alignment Search Tool

BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance.

IgBLAST 1.8.0 released

A new version of IgBLAST is now available.
Wed, 15 Nov 2017 16:00:00 EST

Still too simple



Standalone and API BLAST

Download BLAST
Get BLAST databases and executables

Use BLAST API
Call BLAST from your application

Use BLAST in the cloud
Start an instance at a cloud provider

Specialized searches

SmartBLAST
Find proteins highly similar to your query

Primer-BLAST
Design primers specific to your PCR template

Global Align
Compare two sequences across their entire span (Needleman-Wunsch)

CD-search
Find conserved domains in your sequence

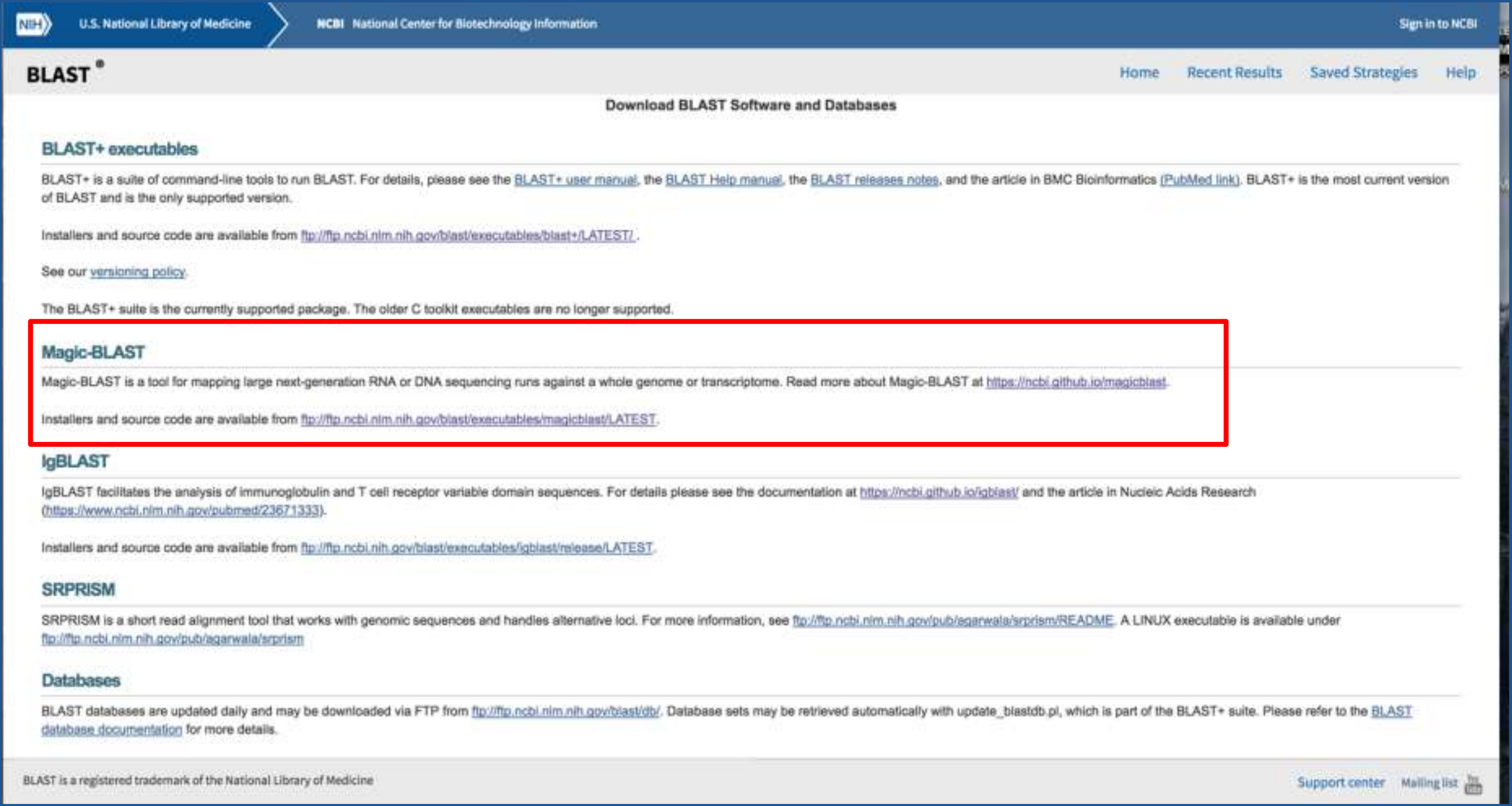
GEO
Find matches to gene expression profiles

IgBLAST
Search immunoglobulins and T cell receptor sequences

VecScreen
Search sequences for vector contamination

CDART
Find sequences with similar conserved domain architecture

MAGIC!



NIH
U.S. National Library of Medicine
NCBI National Center for Biotechnology Information
Sign in to NCBI
BLAST®
Home
Recent Results
Saved Strategies
Help

Download BLAST Software and Databases

BLAST+ executables

BLAST+ is a suite of command-line tools to run BLAST. For details, please see the BLAST+ user manual, the BLAST Help manual, the BLAST releases notes, and the article in BMC Bioinformatics (PubMed link). BLAST+ is the most current version of BLAST and is the only supported version.

Installers and source code are available from ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/ .

See our versioning policy.

The BLAST+ suite is the currently supported package. The older C toolkit executables are no longer supported.

Magic-BLAST

Magic-BLAST is a tool for mapping large next-generation RNA or DNA sequencing runs against a whole genome or transcriptome. Read more about Magic-BLAST at https://ncbi.github.io/magicblast.

Installers and source code are available from ftp://ftp.ncbi.nlm.nih.gov/blast/executables/magicblast/LATEST.

IgBLAST

IgBLAST facilitates the analysis of immunoglobulin and T cell receptor variable domain sequences. For details please see the documentation at https://ncbi.github.io/igblast/ and the article in Nucleic Acids Research (https://www.ncbi.nlm.nih.gov/pubmed/23671333).

Installers and source code are available from ftp://ftp.ncbi.nih.gov/blast/executables/igblast/release/LATEST.

SRPRISM

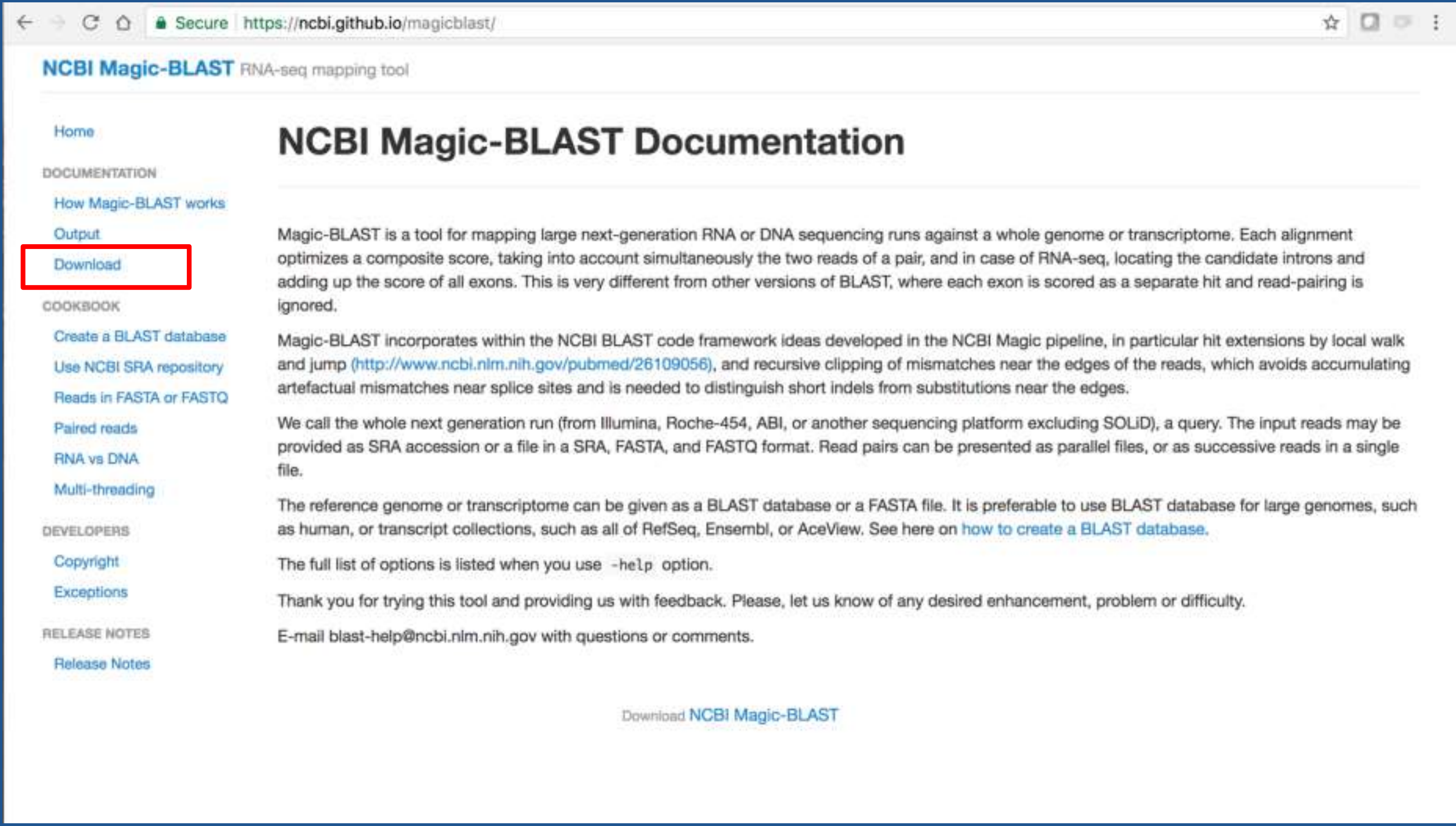
SRPRISM is a short read alignment tool that works with genomic sequences and handles alternative loci. For more information, see ftp://ftp.ncbi.nlm.nih.gov/pub/agarwala/srprism/README. A LINUX executable is available under ftp://ftp.ncbi.nlm.nih.gov/pub/agarwala/srprism

Databases

BLAST databases are updated daily and may be downloaded via FTP from ftp://ftp.ncbi.nlm.nih.gov/blast/db/. Database sets may be retrieved automatically with update_blastdb.pl, which is part of the BLAST+ suite. Please refer to the BLAST database documentation for more details.

BLAST is a registered trademark of the National Library of Medicine

Just download a binary…



NCBI Magic-BLAST RNA-seq mapping tool

NCBI Magic-BLAST Documentation

Magic-BLAST is a tool for mapping large next-generation RNA or DNA sequencing runs against a whole genome or transcriptome. Each alignment optimizes a composite score, taking into account simultaneously the two reads of a pair, and in case of RNA-seq, locating the candidate introns and adding up the score of all exons. This is very different from other versions of BLAST, where each exon is scored as a separate hit and read-pairing is ignored.

Magic-BLAST incorporates within the NCBI BLAST code framework ideas developed in the NCBI Magic pipeline, in particular hit extensions by local walk and jump (http://www.ncbi.nlm.nih.gov/pubmed/26109056), and recursive clipping of mismatches near the edges of the reads, which avoids accumulating artefactual mismatches near splice sites and is needed to distinguish short indels from substitutions near the edges.

We call the whole next generation run (from Illumina, Roche-454, ABI, or another sequencing platform excluding SOLiD), a query. The input reads may be provided as SRA accession or a file in a SRA, FASTA, and FASTQ format. Read pairs can be presented as parallel files, or as successive reads in a single file.

The reference genome or transcriptome can be given as a BLAST database or a FASTA file. It is preferable to use BLAST database for large genomes, such as human, or transcript collections, such as all of RefSeq, Ensembl, or AceView. See here on how to create a BLAST database.

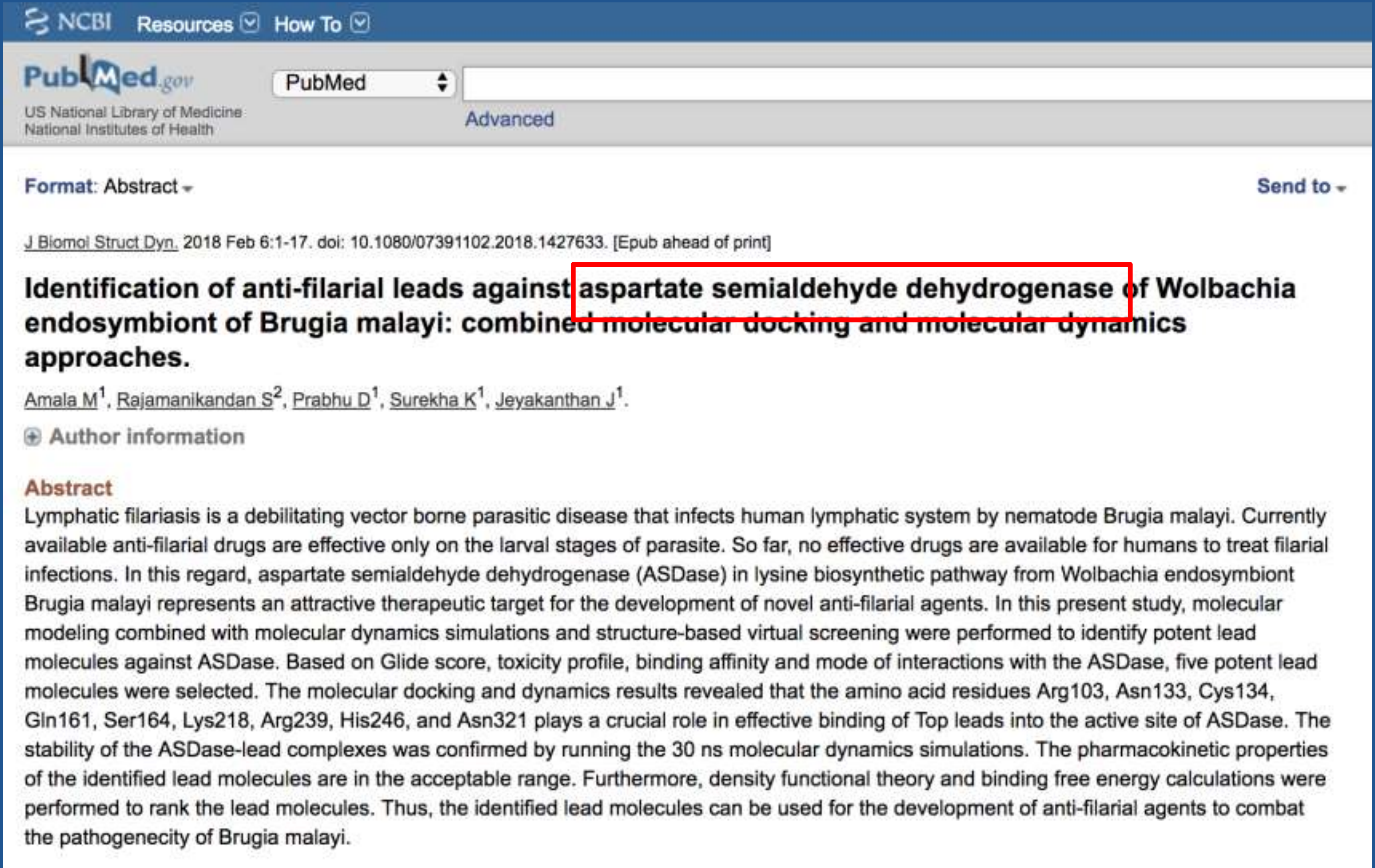
The full list of options is listed when you use `-help` option.

Thank you for trying this tool and providing us with feedback. Please, let us know of any desired enhancement, problem or difficulty.

E-mail blast-help@ncbi.nlm.nih.gov with questions or comments.

Download NCBI Magic-BLAST

Find something to BLAST into:



J Biomol Struct Dyn. 2018 Feb 6:1-17. doi: 10.1080/07391102.2018.1427633. [Epub ahead of print]

Identification of anti-filarial leads against aspartate semialdehyde dehydrogenase of Wolbachia endosymbiont of Brugia malayi: combined molecular docking and molecular dynamics approaches.

Amala M[1], Rajamanikandan S[2], Prabhu D[1], Surekha K[1], Jeyakanthan J[1].

⊕ Author information

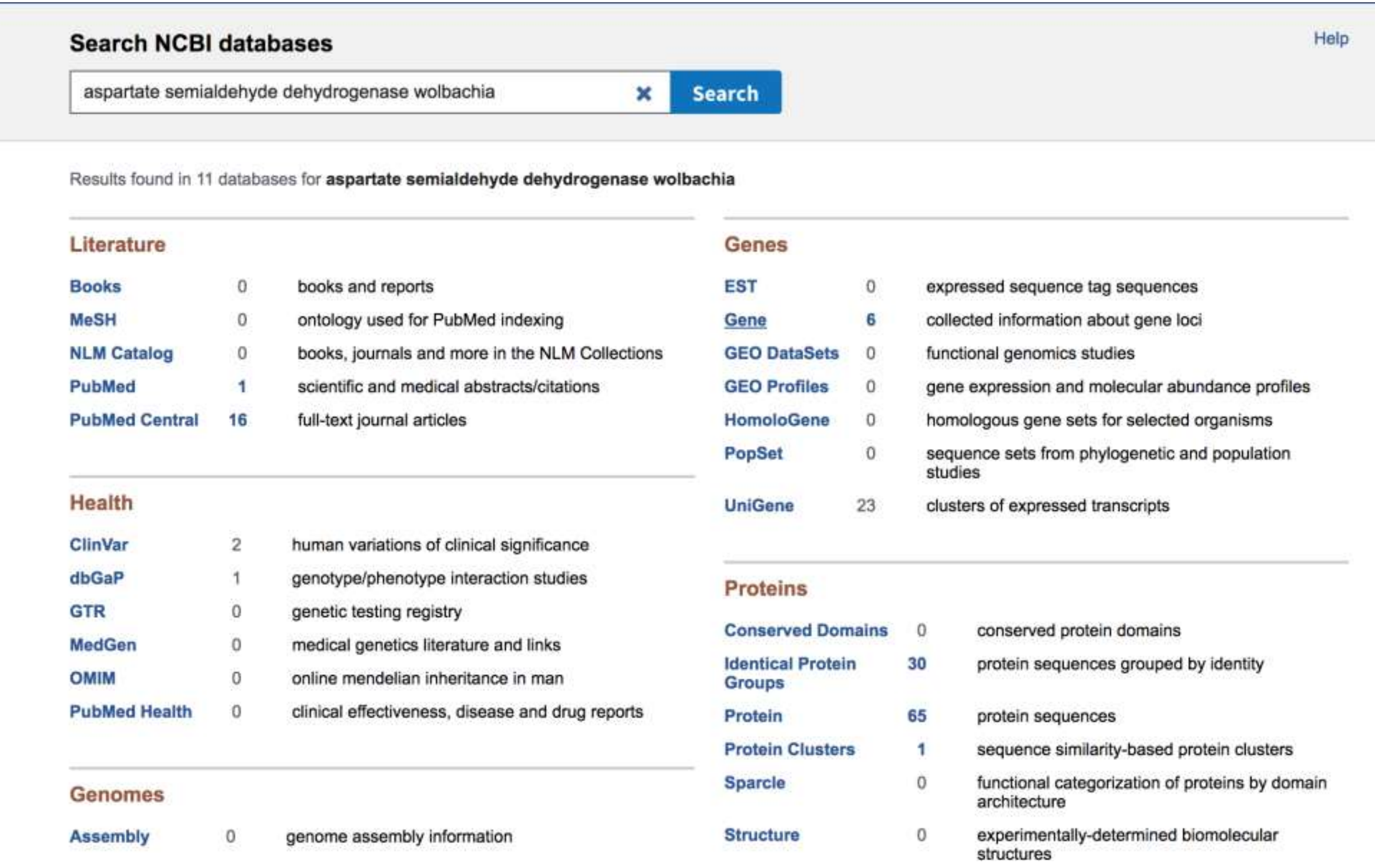
Abstract

Lymphatic filariasis is a debilitating vector borne parasitic disease that infects human lymphatic system by nematode Brugia malayi. Currently available anti-filarial drugs are effective only on the larval stages of parasite. So far, no effective drugs are available for humans to treat filarial infections. In this regard, aspartate semialdehyde dehydrogenase (ASDase) in lysine biosynthetic pathway from Wolbachia endosymbiont Brugia malayi represents an attractive therapeutic target for the development of novel anti-filarial agents. In this present study, molecular modeling combined with molecular dynamics simulations and structure-based virtual screening were performed to identify potent lead molecules against ASDase. Based on Glide score, toxicity profile, binding affinity and mode of interactions with the ASDase, five potent lead molecules were selected. The molecular docking and dynamics results revealed that the amino acid residues Arg103, Asn133, Cys134, Gln161, Ser164, Lys218, Arg239, His246, and Asn321 plays a crucial role in effective binding of Top leads into the active site of ASDase. The stability of the ASDase-lead complexes was confirmed by running the 30 ns molecular dynamics simulations. The pharmacokinetic properties of the identified lead molecules are in the acceptable range. Furthermore, density functional theory and binding free energy calculations were performed to rank the lead molecules. Thus, the identified lead molecules can be used for the development of anti-filarial agents to combat the pathogenecity of Brugia malayi.

This is where I go really fast.

Please check out slideshare to grab the details.

https://www.slideshare.net/benbusby



Search NCBI databases

aspartate semialdehyde dehydrogenase wolbachia

Results found in 11 databases for **aspartate semialdehyde dehydrogenase wolbachia**

Literature

Books	0	books and reports
MeSH	0	ontology used for PubMed indexing
NLM Catalog	0	books, journals and more in the NLM Collections
PubMed	1	scientific and medical abstracts/citations
PubMed Central	16	full-text journal articles

Health

ClinVar	2	human variations of clinical significance
dbGaP	1	genotype/phenotype interaction studies
GTR	0	genetic testing registry
MedGen	0	medical genetics literature and links
OMIM	0	online mendelian inheritance in man
PubMed Health	0	clinical effectiveness, disease and drug reports

Genomes

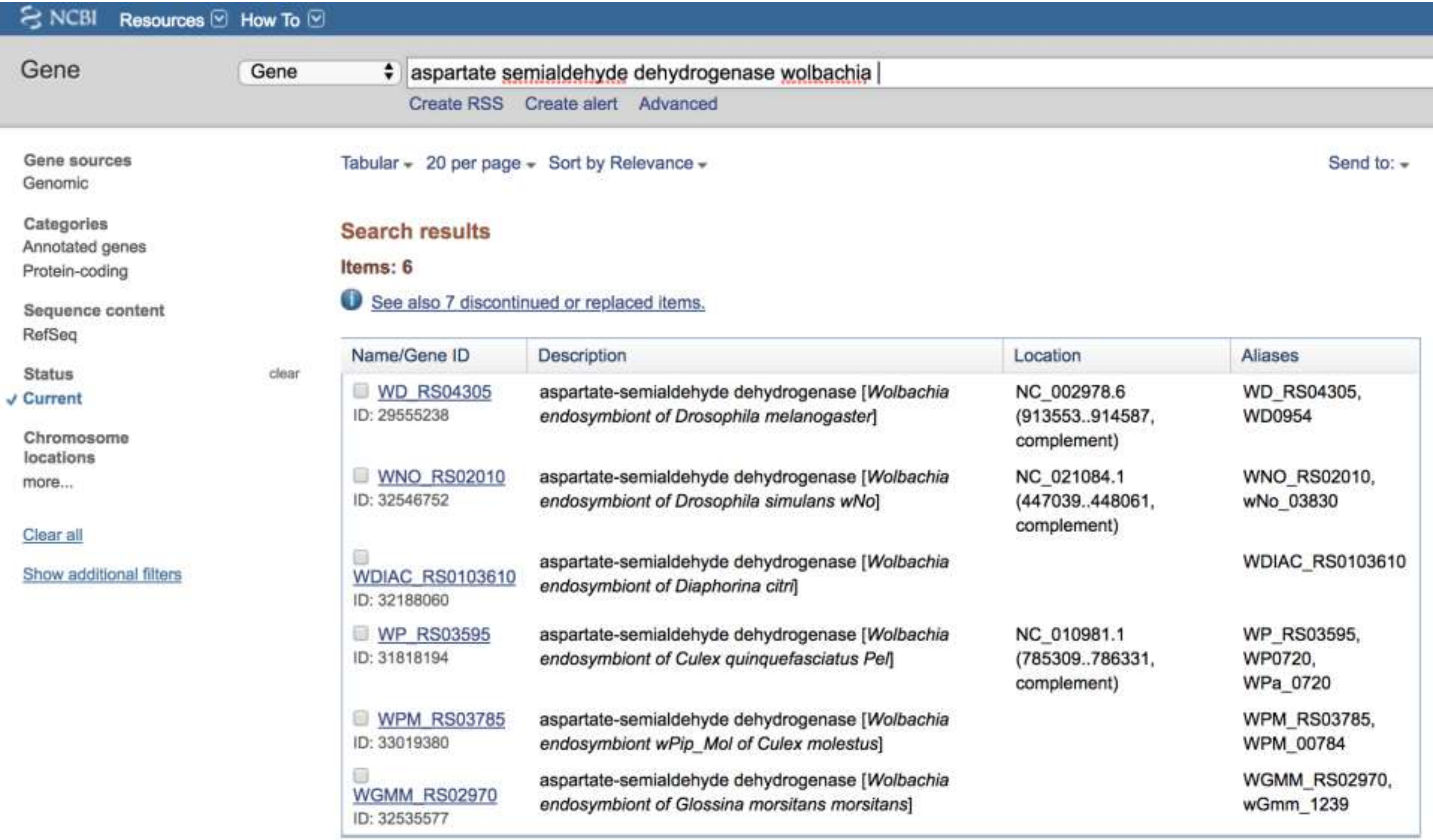
Assembly	0	genome assembly information

Genes

EST	0	expressed sequence tag sequences
Gene	6	collected information about gene loci
GEO DataSets	0	functional genomics studies
GEO Profiles	0	gene expression and molecular abundance profiles
HomoloGene	0	homologous gene sets for selected organisms
PopSet	0	sequence sets from phylogenetic and population studies
UniGene	23	clusters of expressed transcripts

Proteins

Conserved Domains	0	conserved protein domains
Identical Protein Groups	30	protein sequences grouped by identity
Protein	65	protein sequences
Protein Clusters	1	sequence similarity-based protein clusters
Sparcle	0	functional categorization of proteins by domain architecture
Structure	0	experimentally-determined biomolecular structures



Gene

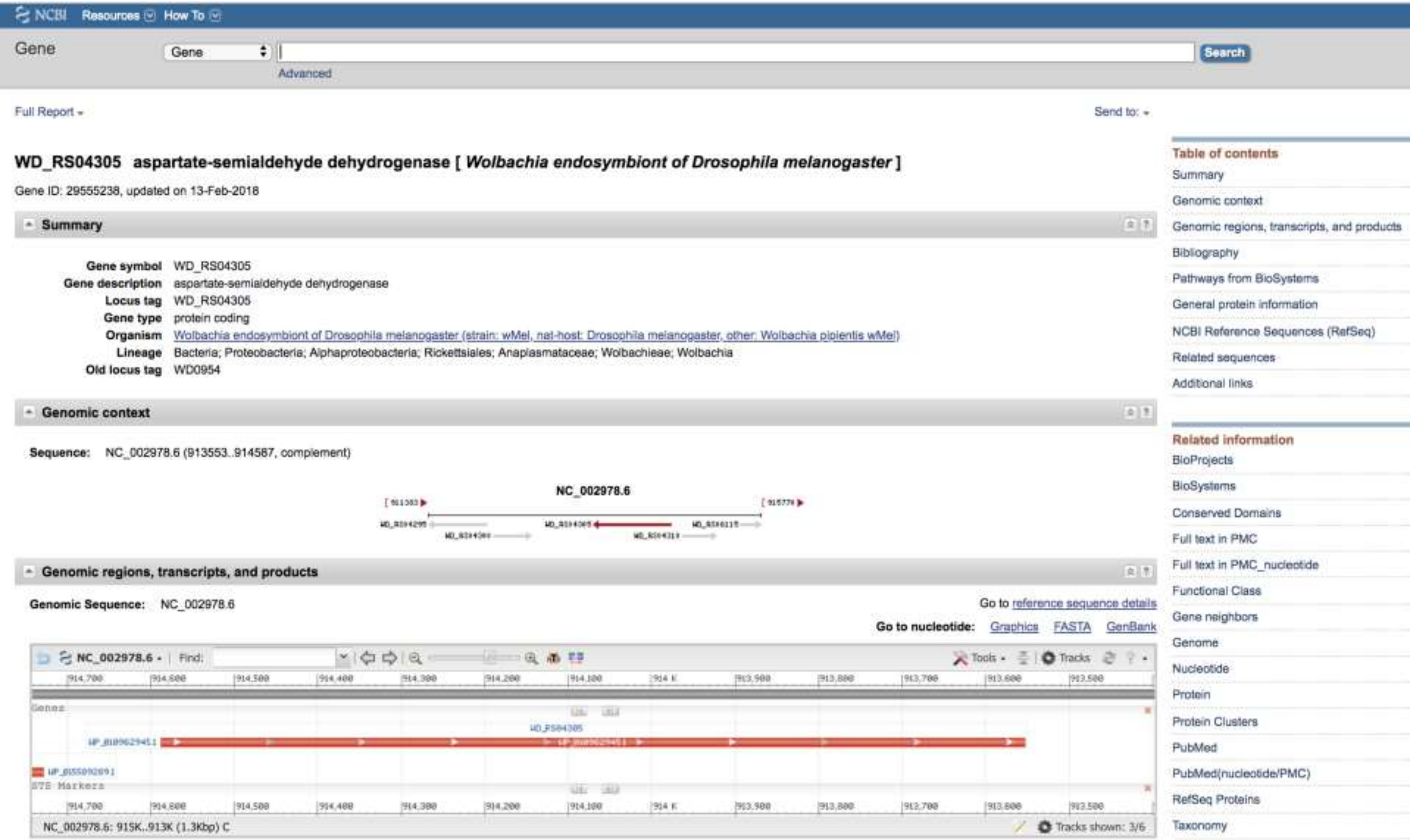
aspartate semialdehyde dehydrogenase wolbachia

Search results

Items: 6

See also 7 discontinued or replaced items.

Name/Gene ID	Description	Location	Aliases
□ WD_RS04305 ID: 29555238	aspartate-semialdehyde dehydrogenase [*Wolbachia endosymbiont of Drosophila melanogaster*]	NC_002978.6 (913553..914587, complement)	WD_RS04305, WD0954
□ WNO_RS02010 ID: 32546752	aspartate-semialdehyde dehydrogenase [*Wolbachia endosymbiont of Drosophila simulans wNo*]	NC_021084.1 (447039..448061, complement)	WNO_RS02010, wNo_03830
□ WDIAC_RS0103610 ID: 32188060	aspartate-semialdehyde dehydrogenase [*Wolbachia endosymbiont of Diaphorina citri*]		WDIAC_RS0103610
□ WP_RS03595 ID: 31818194	aspartate-semialdehyde dehydrogenase [*Wolbachia endosymbiont of Culex quinquefasciatus Pel*]	NC_010981.1 (785309..786331, complement)	WP_RS03595, WP0720, WPa_0720
□ WPM_RS03785 ID: 33019380	aspartate-semialdehyde dehydrogenase [*Wolbachia endosymbiont wPip_Mol of Culex molestus*]		WPM_RS03785, WPM_00784
□ WGMM_RS02970 ID: 32535577	aspartate-semialdehyde dehydrogenase [*Wolbachia endosymbiont of Glossina morsitans morsitans*]		WGMM_RS02970, wGmm_1239



WD_RS04305 aspartate-semialdehyde dehydrogenase [*Wolbachia endosymbiont of Drosophila melanogaster*]

Gene ID: 29555238, updated on 13-Feb-2018

Summary

Gene symbol	WD_RS04305
Gene description	aspartate-semialdehyde dehydrogenase
Locus tag	WD_RS04305
Gene type	protein coding
Organism	Wolbachia endosymbiont of Drosophila melanogaster (strain: wMel, nat-host: Drosophila melanogaster, other: Wolbachia pipientis wMel)
Lineage	Bacteria; Proteobacteria; Alphaproteobacteria; Rickettsiales; Anaplasmataceae; Wolbachieae; Wolbachia
Old locus tag	WD0954

Genomic context

Sequence: NC_002978.6 (913553..914587, complement)

Genomic regions, transcripts, and products

Genomic Sequence: NC_002978.6

Go to reference sequence details

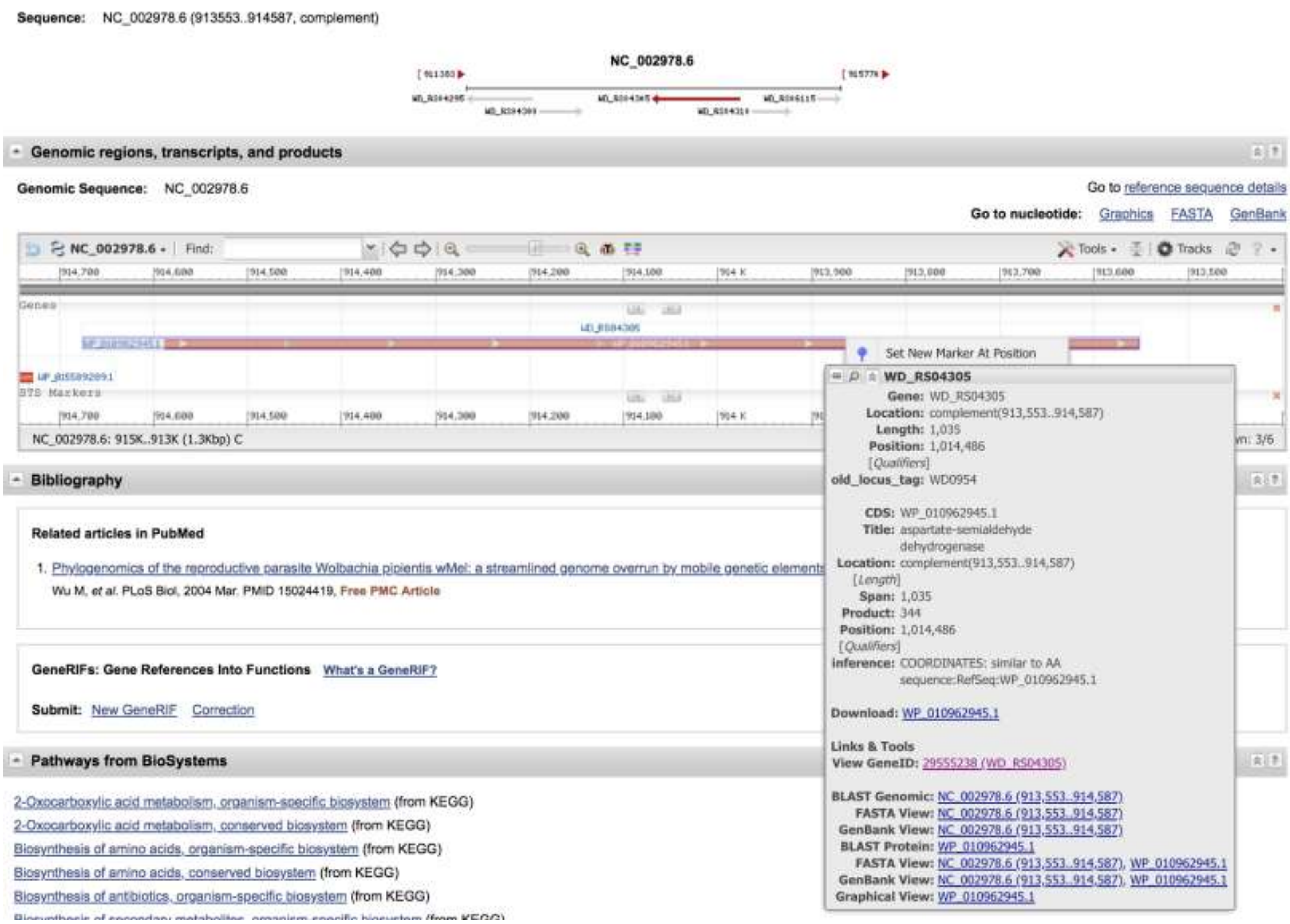
Go to nucleotide: Graphics FASTA GenBank

NC_002978.6: 915K..913K (1.3Kbp) C

Tracks shown: 3/6

Table of contents
- Summary
- Genomic context
- Genomic regions, transcripts, and products
- Bibliography
- Pathways from BioSystems
- General protein information
- NCBI Reference Sequences (RefSeq)
- Related sequences
- Additional links

Related information
- BioProjects
- BioSystems
- Conserved Domains
- Full text in PMC
- Full text in PMC_nucleotide
- Functional Class
- Gene neighbors
- Genome
- Nucleotide
- Protein
- Protein Clusters
- PubMed
- PubMed(nucleotide/PMC)
- RefSeq Proteins
- Taxonomy



Sequence: NC_002978.6 (913553..914587, complement)

Genomic regions, transcripts, and products

Genomic Sequence: NC_002978.6

Go to reference sequence details

Go to nucleotide: Graphics FASTA GenBank

NC_002978.6 ▾

NC_002978.6: 915K..913K (1.3Kbp) C

WD_RS04305

Gene: WD_RS04305
Location: complement(913,553..914,587)
Length: 1,035
Position: 1,014,486
[*Qualifiers*]
old_locus_tag: WD0954

CDS: WP_010962945.1
Title: aspartate-semialdehyde dehydrogenase
Location: complement(913,553..914,587)
[*Length*]
Span: 1,035
Product: 344
Position: 1,014,486
[*Qualifiers*]
inference: COORDINATES: similar to AA sequence:RefSeq:WP_010962945.1

Download: WP_010962945.1

Links & Tools
View GeneID: 29555238 (WD_RS04305)

BLAST Genomic: NC_002978.6 (913,553..914,587)
FASTA View: NC_002978.6 (913,553..914,587)
GenBank View: NC_002978.6 (913,553..914,587)
BLAST Protein: WP_010962945.1
FASTA View: NC_002978.6 (913,553..914,587), WP_010962945.1
GenBank View: NC_002978.6 (913,553..914,587), WP_010962945.1
Graphical View: WP_010962945.1

Bibliography

Related articles in PubMed

1. Phylogenomics of the reproductive parasite Wolbachia pipientis wMel: a streamlined genome overrun by mobile genetic elements
   Wu M, *et al.* PLoS Biol, 2004 Mar. PMID 15024419, **Free PMC Article**

GeneRIFs: Gene References Into Functions What's a GeneRIF?

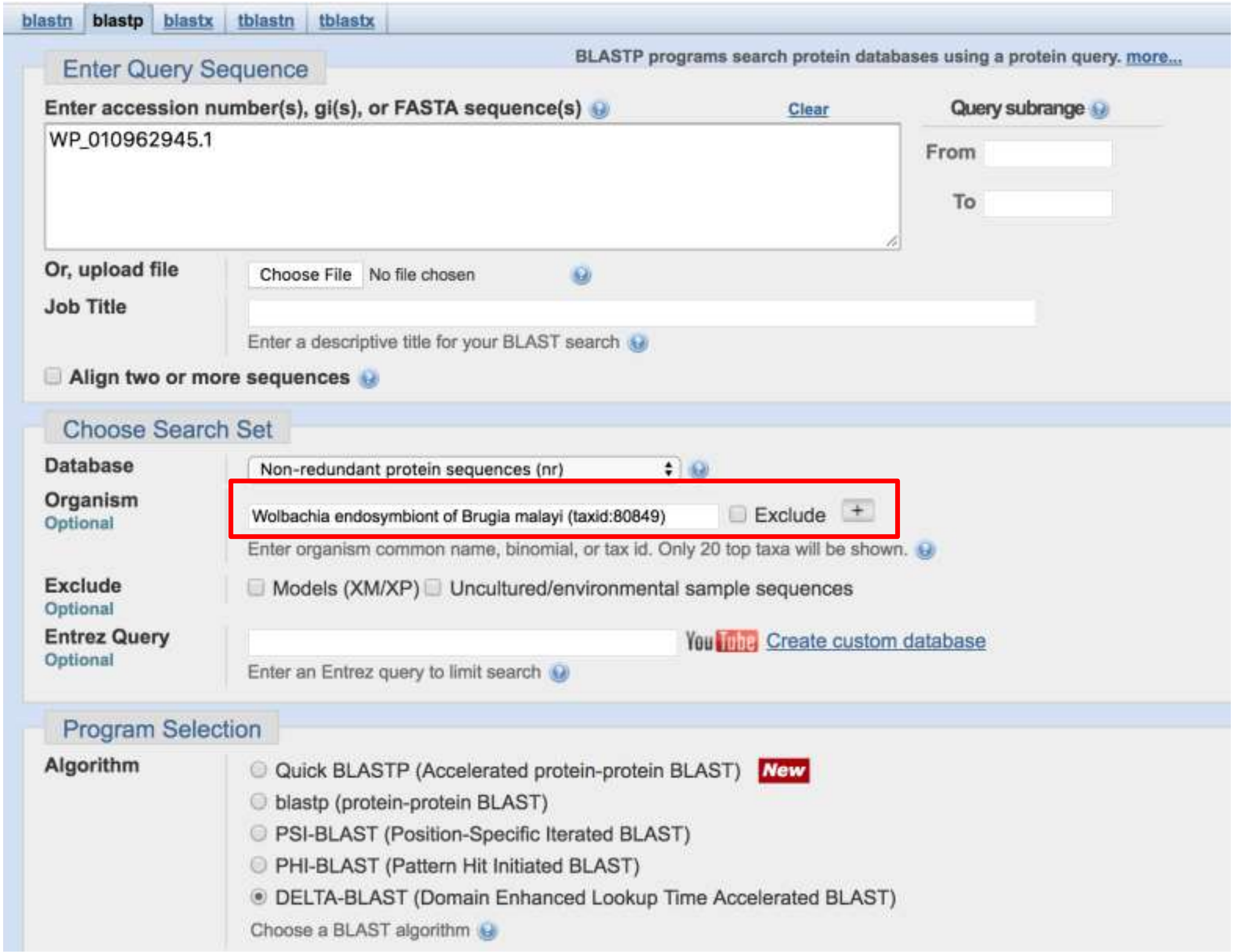
Submit: New GeneRIF Correction

Pathways from BioSystems

2-Oxocarboxylic acid metabolism, organism-specific biosystem (from KEGG)
2-Oxocarboxylic acid metabolism, conserved biosystem (from KEGG)
Biosynthesis of amino acids, organism-specific biosystem (from KEGG)
Biosynthesis of amino acids, conserved biosystem (from KEGG)
Biosynthesis of antibiotics, organism-specific biosystem (from KEGG)
Biosynthesis of secondary metabolites, organism-specific biosystem (from KEGG)

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

NCBI



BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) — Clear

WP_010962945.1

Query subrange

From

To

Or, upload file — Choose File No file chosen

Job Title

Enter a descriptive title for your BLAST search

□ Align two or more sequences

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism Optional: Wolbachia endosymbiont of Brugia malayi (taxid:80849) □ Exclude +

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown.

Exclude Optional: □ Models (XM/XP) □ Uncultured/environmental sample sequences

Entrez Query Optional: YouTube Create custom database

Enter an Entrez query to limit search

Program Selection

Algorithm

○ Quick BLASTP (Accelerated protein-protein BLAST) New

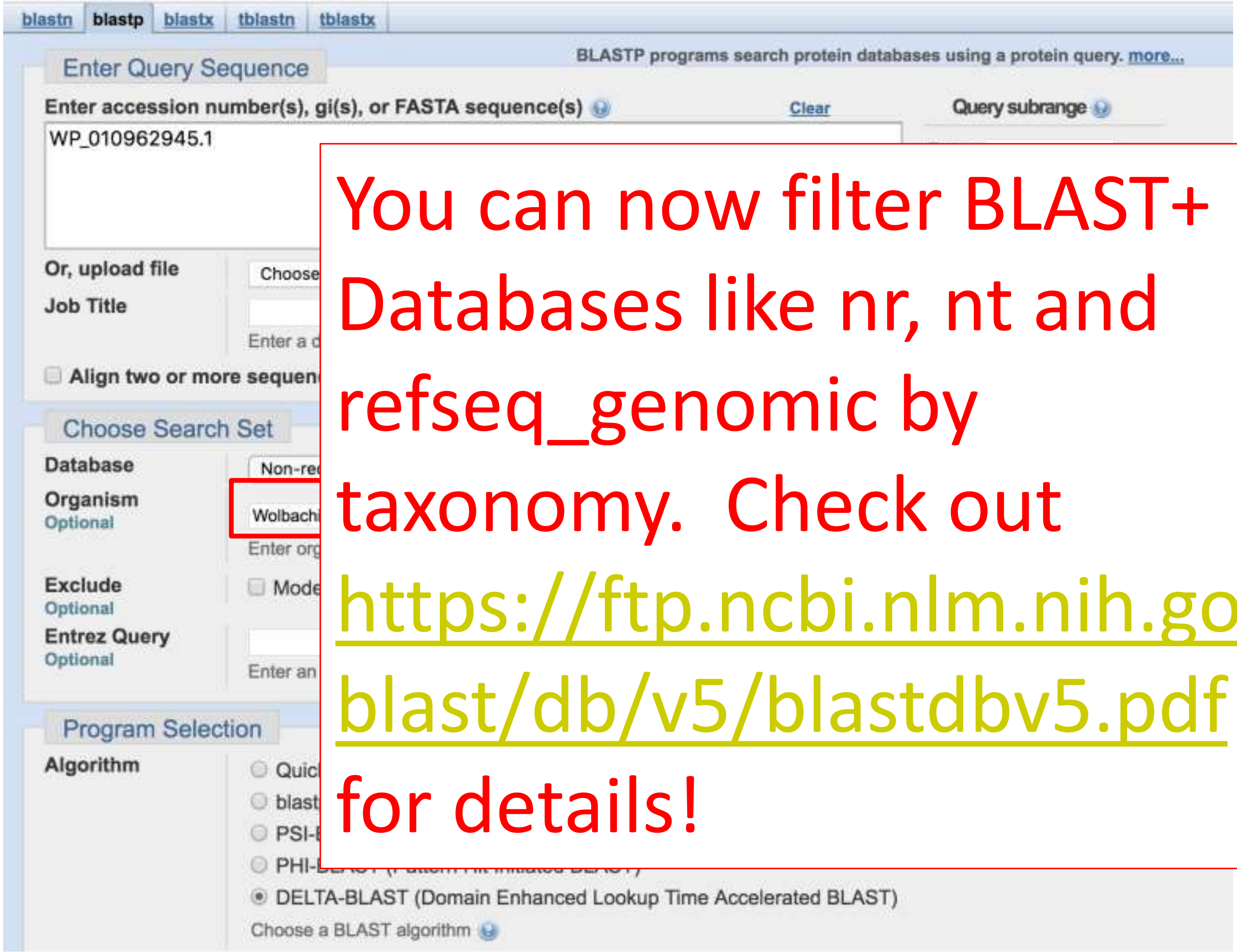
○ blastp (protein-protein BLAST)

○ PSI-BLAST (Position-Specific Iterated BLAST)

○ PHI-BLAST (Pattern Hit Initiated BLAST)

● DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm



blastn | **blastp** | blastx | tblastn | tblastx

BLASTP programs search protein databases using a protein query. more...

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

Query subrange

WP_010962945.1

Or, upload file Choose

Job Title

Enter a d

□ Align two or more sequen

Choose Search Set

Database Non-red

Organism Optional Wolbachi

Enter org

Exclude Optional □ Mode

Entrez Query Optional

Enter an

Program Selection

Algorithm

○ Quic

○ blast

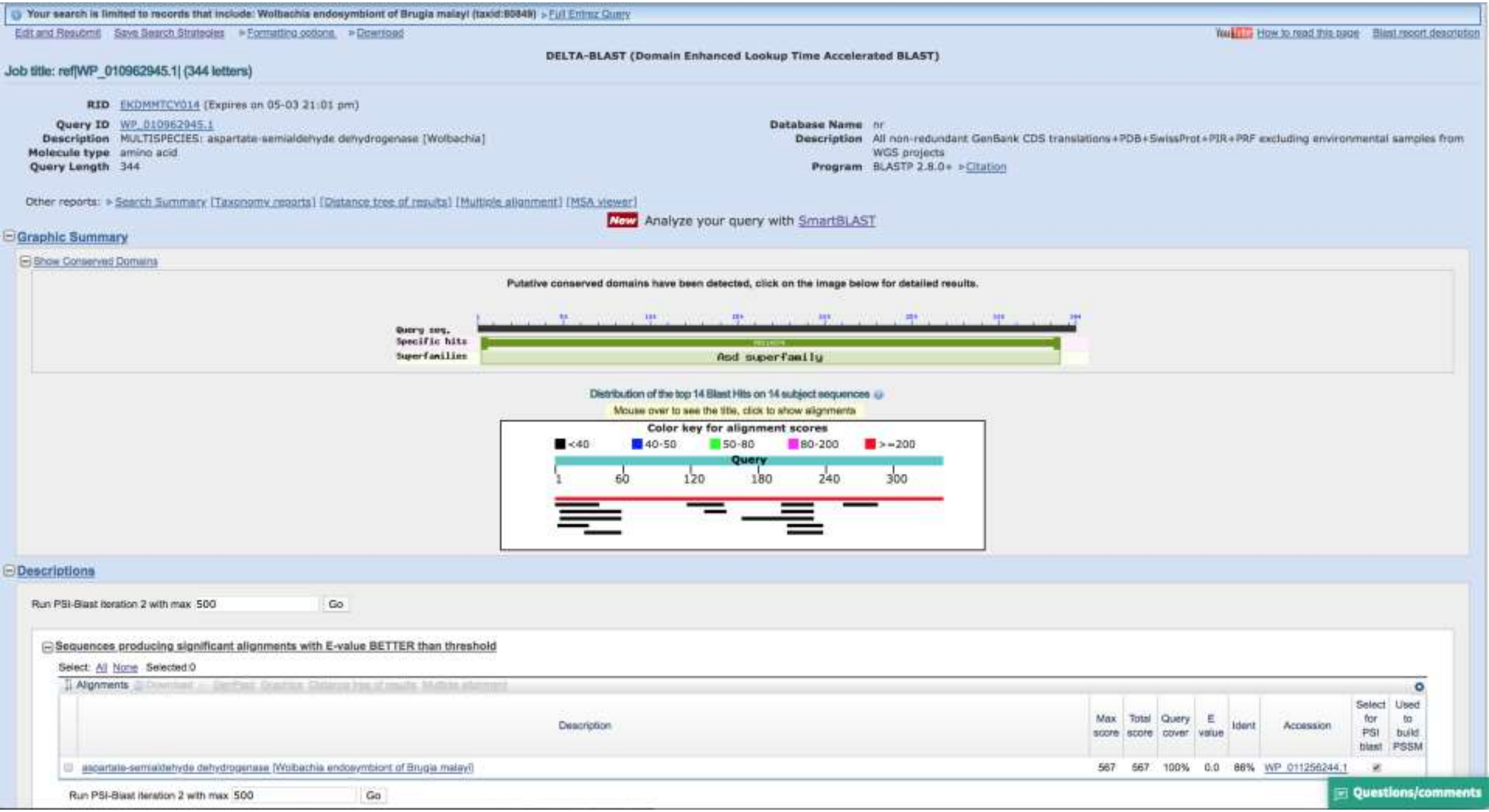
○ PSI-

○ PHI-BLAST (Pattern Hit Initiated BLAST)

● DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Choose a BLAST algorithm

You can now filter BLAST+ Databases like nr, nt and refseq_genomic by taxonomy. Check out https://ftp.ncbi.nlm.nih.gov/blast/db/v5/blastdbv5.pdf for details!



Your search is limited to records that include: Wolbachia endosymbiont of Brugia malayi (taxid:80849) > Full Entrez Query

Edit and Resubmit Save Search Strategies > Formatting options > Download How to read this page Blast report description

DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Job title: ref|WP_010962945.1| (344 letters)

RID	EKDMMTCY014 (Expires on 05-03 21:01 pm)
Query ID	WP_010962945.1
Description	MULTISPECIES: aspartate-semialdehyde dehydrogenase [Wolbachia]
Molecule type	amino acid
Query Length	344

Database Name	nr
Description	All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Program	BLASTP 2.8.0+ > Citation

Other reports: > Search Summary [Taxonomy reports] [Distance tree of results] [Multiple alignment] [MSA viewer]

New Analyze your query with SmartBLAST

⊟ Graphic Summary

⊟ Show Conserved Domains

Putative conserved domains have been detected, click on the image below for detailed results.

Distribution of the top 14 Blast Hits on 14 subject sequences

Mouse over to see the title, click to show alignments

⊟ Descriptions

Run PSI-Blast iteration 2 with max 500 Go

⊟ Sequences producing significant alignments with E-value BETTER than threshold

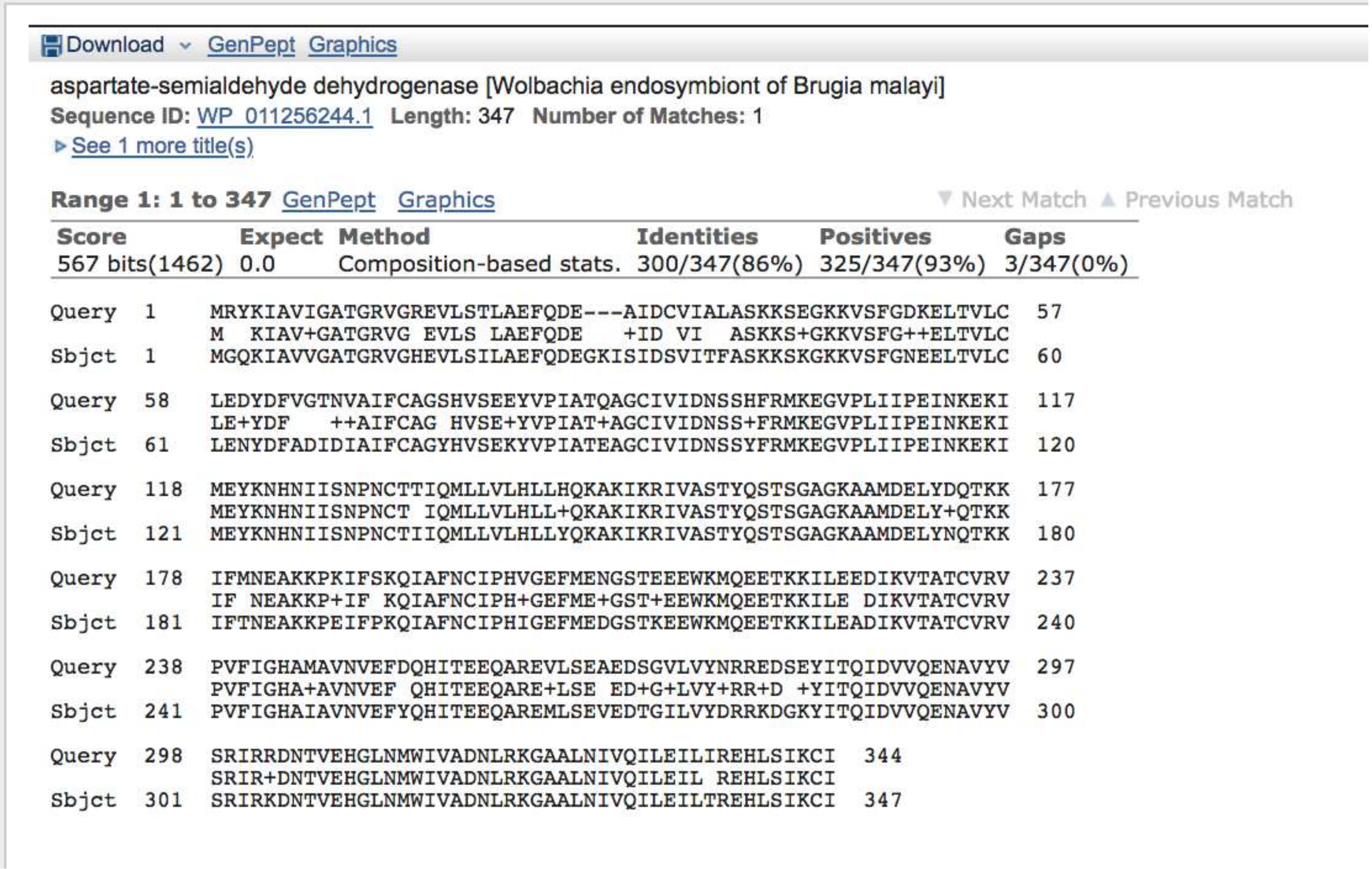
Select: All None Selected:0

Alignments Download — GenPept Graphics Distance tree of results Multiple alignment

	Description	Max score	Total score	Query cover	E value	Ident	Accession	Select for PSI blast	Used to build PSSM
□	aspartate-semialdehyde dehydrogenase [Wolbachia endosymbiont of Brugia malayi]	567	567	100%	0.0	88%	WP_011256244.1	✓	

Run PSI-Blast iteration 2 with max 500 Go

Questions/comments



aspartate-semialdehyde dehydrogenase [Wolbachia endosymbiont of Brugia malayi]

Sequence ID: WP_011256244.1 **Length:** 347 **Number of Matches:** 1

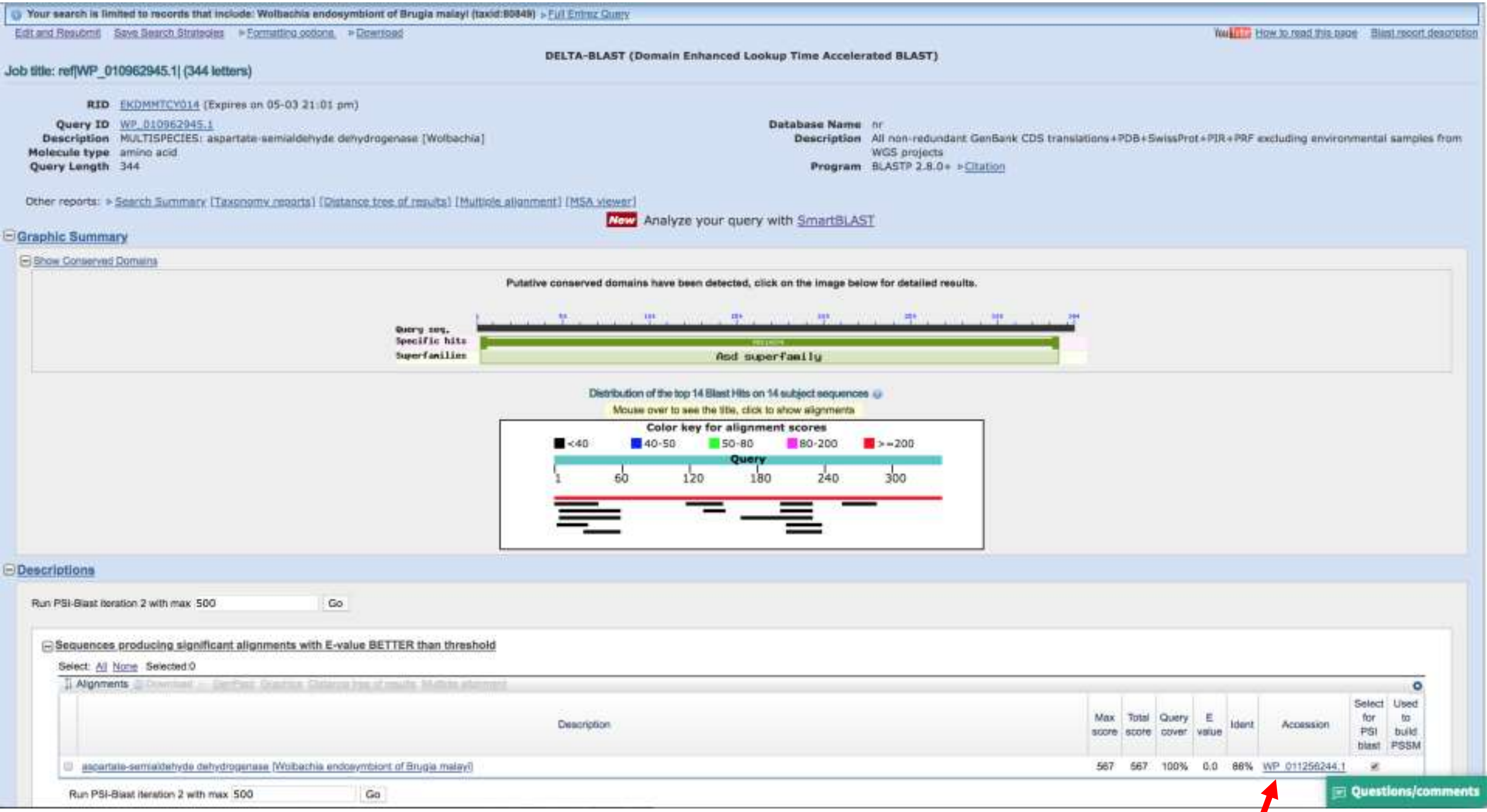
▶ See 1 more title(s)

Range 1: 1 to 347 GenPept Graphics

Score	Expect	Method	Identities	Positives	Gaps
567 bits(1462)	0.0	Composition-based stats.	300/347(86%)	325/347(93%)	3/347(0%)

```
Query  1    MRYKIAVIGATGRVGREVLSTLAEFQDE---AIDCVIALASKKSEGKKVSFGDKELTVLC  57
            M  KIAV+GATGRVG EVLS LAEFQDE   +ID VI  ASKKS+GKKVSFG++ELTVLC
Sbjct  1    MGQKIAVVGATGRVGHEVLSILAEFQDEGKISIDSVITFASKKSKGKKVSFGNEELTVLC  60

Query  58   LEDYDFVGTNVAIFCAGSHVSEEYVPIATQAGCIVIDNSSHFRMKEGVPLIIPEINKEKI  117
            LE+YDF   ++AIFCAG HVSE+YVPIAT+AGCIVIDNSS+FRMKEGVPLIIPEINKEKI
Sbjct  61   LENYDFADIDIAIFCAGYHVSEKYVPIATEAGCIVIDNSSYFRMKEGVPLIIPEINKEKI  120

Query  118  MEYKNHNIISNPNCTTIQMLLVLHLLHQKAKIKRIVASTYQSTSGAGKAAMDELYDQTKK  177
            MEYKNHNIISNPNCT IQMLLVLHLL+QKAKIKRIVASTYQSTSGAGKAAMDELY+QTKK
Sbjct  121  MEYKNHNIISNPNCTIIQMLLVLHLLYQKAKIKRIVASTYQSTSGAGKAAMDELYNQTKK  180

Query  178  IFMNEAKKPKIFSKQIAFNCIPHVGEFMENGSTEEEWKMQEETKKILEEDIKVTATCVRV  237
            IF NEAKKP+IF KQIAFNCIPH+GEFME+GST+EEWKMQEETKKILE DIKVTATCVRV
Sbjct  181  IFTNEAKKPEIFPKQIAFNCIPHIGEFMEDGSTKEEWKMQEETKKILEADIKVTATCVRV  240

Query  238  PVFIGHAMAVNVEFDQHITEEQAREVLSEAEDSGVLVYNRREDSEYITQIDVVQENAVYV  297
            PVFIGHA+AVNVEF QHITEEQARE+LSE ED+G+LVY+RR+D +YITQIDVVQENAVYV
Sbjct  241  PVFIGHAIAVNVEFYQHITEEQAREMLSEVEDTGILVYDRRKDGKYITQIDVVQENAVYV  300

Query  298  SRIRRDNTVEHGLNMWIVADNLRKGAALNIVQILEILIREHLSIKCI  344
            SRIR+DNTVEHGLNMWIVADNLRKGAALNIVQILEIL REHLSIKCI
Sbjct  301  SRIRKDNTVEHGLNMWIVADNLRKGAALNIVQILEILTREHLSIKCI  347
```

The aspartate-semialdehyde dehydrogenases are not dramatically different in Drosophila melanogaster wolbachia and Brugia malayi wolbachia



Your search is limited to records that include: Wolbachia endosymbiont of Brugia malayi (taxid:80849) > Full Entrez Query

Edit and Resubmit   Save Search Strategies   > Formatting options   > Download   How to read this page   Blast report description

DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)

Job title: ref|WP_010962945.1| (344 letters)

RID EKDMMTCY014 (Expires on 05-03 21:01 pm)

Query ID WP_010962945.1
Description MULTISPECIES: aspartate-semialdehyde dehydrogenase [Wolbachia]
Molecule type amino acid
Query Length 344

Database Name nr
Description All non-redundant GenBank CDS translations+PDB+SwissProt+PIR+PRF excluding environmental samples from WGS projects
Program BLASTP 2.8.0+ > Citation

Other reports: > Search Summary [Taxonomy reports] [Distance tree of results] [Multiple alignment] [MSA viewer]

New Analyze your query with SmartBLAST

⊟ Graphic Summary

⊟ Show Conserved Domains

Putative conserved domains have been detected, click on the image below for detailed results.

Distribution of the top 14 Blast Hits on 14 subject sequences

Mouse over to see the title, click to show alignments

⊟ Descriptions

Run PSI-Blast iteration 2 with max 500 Go

⊟ Sequences producing significant alignments with E-value BETTER than threshold

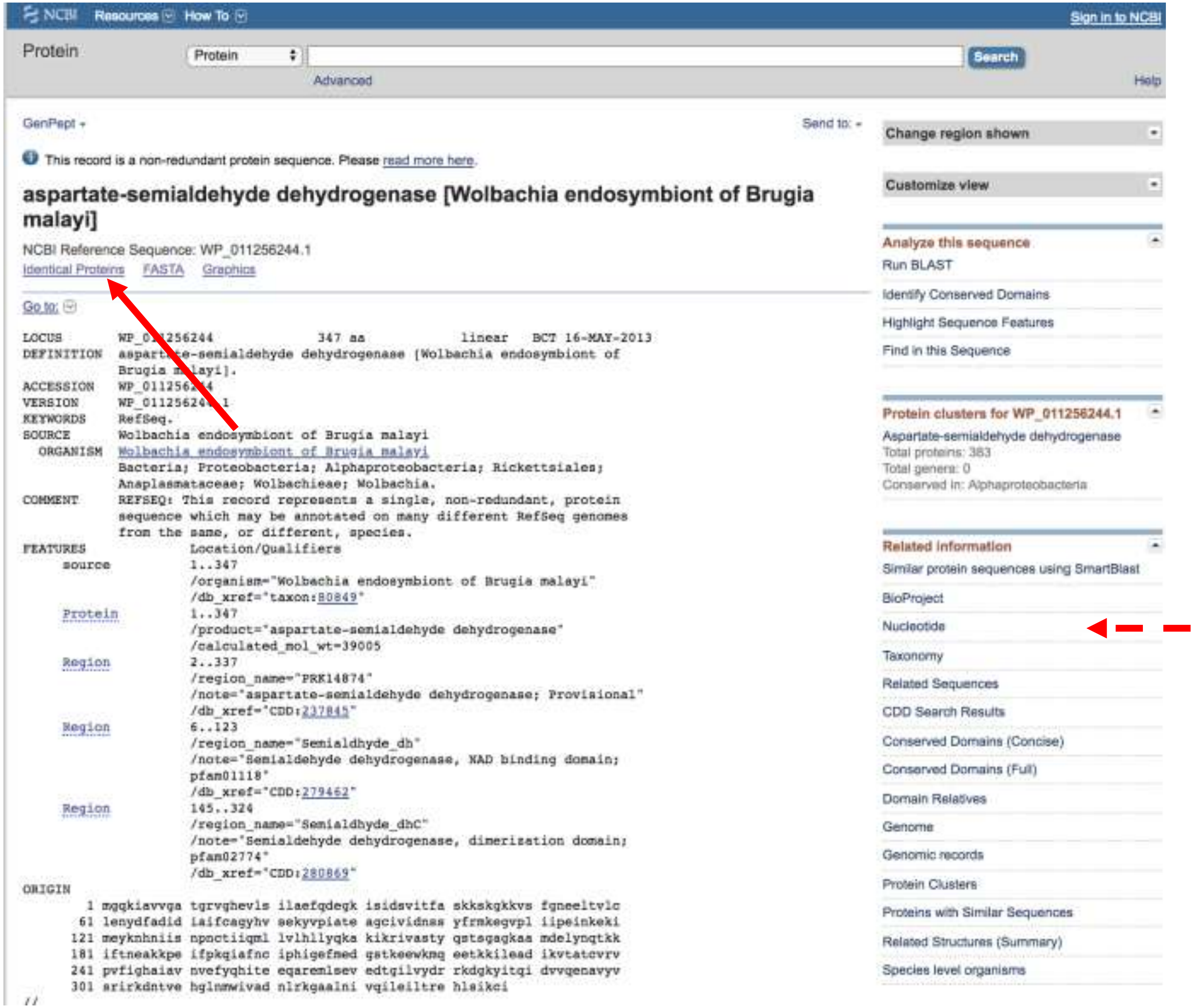
Select: All None Selected:0

Alignments

	Description	Max score	Total score	Query cover	E value	Ident	Accession	Select for PSI blast	Used to build PSSM
□	aspartate-semialdehyde dehydrogenase [Wolbachia endosymbiont of Brugia malayi]	567	567	100%	0.0	88%	WP_011256244.1	✓	

Run PSI-Blast iteration 2 with max 500 Go

Questions/comments



This record is a non-redundant protein sequence. Please read more here.

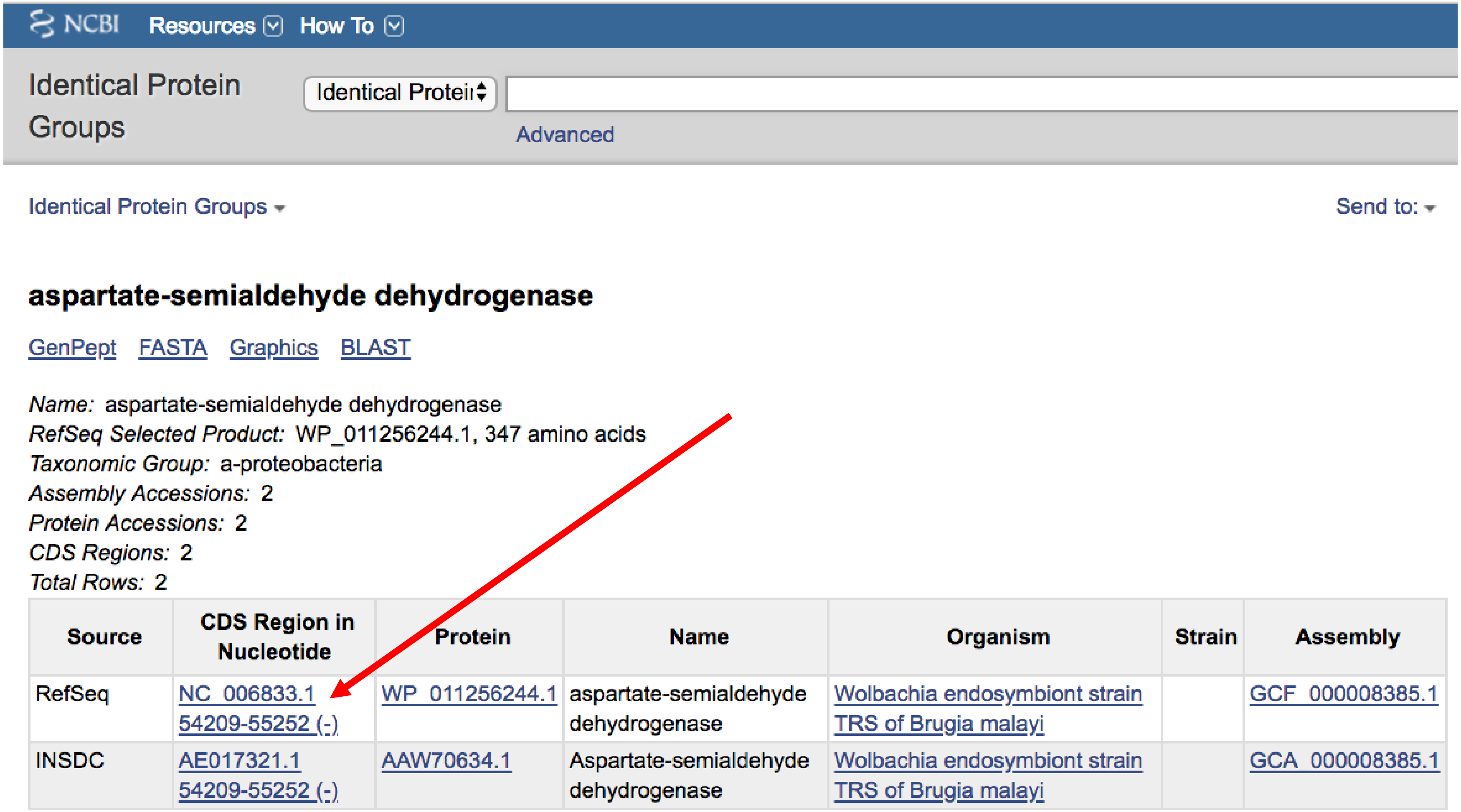
aspartate-semialdehyde dehydrogenase [Wolbachia endosymbiont of Brugia malayi]

NCBI Reference Sequence: WP_011256244.1

```
LOCUS       WP_011256244             347 aa            linear   BCT 16-MAY-2013
DEFINITION  aspartate-semialdehyde dehydrogenase [Wolbachia endosymbiont of
            Brugia malayi].
ACCESSION   WP_011256244
VERSION     WP_011256244.1
KEYWORDS    RefSeq.
SOURCE      Wolbachia endosymbiont of Brugia malayi
  ORGANISM  Wolbachia endosymbiont of Brugia malayi
            Bacteria; Proteobacteria; Alphaproteobacteria; Rickettsiales;
            Anaplasmataceae; Wolbachieae; Wolbachia.
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.
FEATURES             Location/Qualifiers
     source          1..347
                     /organism="Wolbachia endosymbiont of Brugia malayi"
                     /db_xref="taxon:80849"
     Protein         1..347
                     /product="aspartate-semialdehyde dehydrogenase"
                     /calculated_mol_wt=39005
     Region          2..337
                     /region_name="PRK14874"
                     /note="aspartate-semialdehyde dehydrogenase; Provisional"
                     /db_xref="CDD:237845"
     Region          6..123
                     /region_name="Semialdhyde_dh"
                     /note="Semialdehyde dehydrogenase, NAD binding domain;
                     pfam01118"
                     /db_xref="CDD:279462"
     Region          145..324
                     /region_name="Semialdhyde_dhC"
                     /note="Semialdehyde dehydrogenase, dimerization domain;
                     pfam02774"
                     /db_xref="CDD:280869"
ORIGIN
        1 mgqkiavvga tgrvghevls ilaefqdegk isidsvitfa skkskgkkvs fgneeltvlc
       61 lenydfadid laifcagyhv sekyvpiate agcividnas yfrmkegvpl iipeinkeki
      121 meyknhniis npnctiiqml lvlhllyqka kikrivasty qstsgagkaa mdelynqtkk
      181 iftneakkpe ifpkqiafnc iphigefmed gstkeewkmq eetkkilead ikvtatcvrv
      241 pvfighaiav nvefyqhite eqaremlsev edtgilvydr rkdgkyitqi dvvqenavyv
      301 arirkdntve hglnmwivad nlrkgaalni vqileiltre hlsikci
//
```

Protein clusters for WP_011256244.1

Aspartate-semialdehyde dehydrogenase
Total proteins: 383
Total genera: 0
Conserved in: Alphaproteobacteria

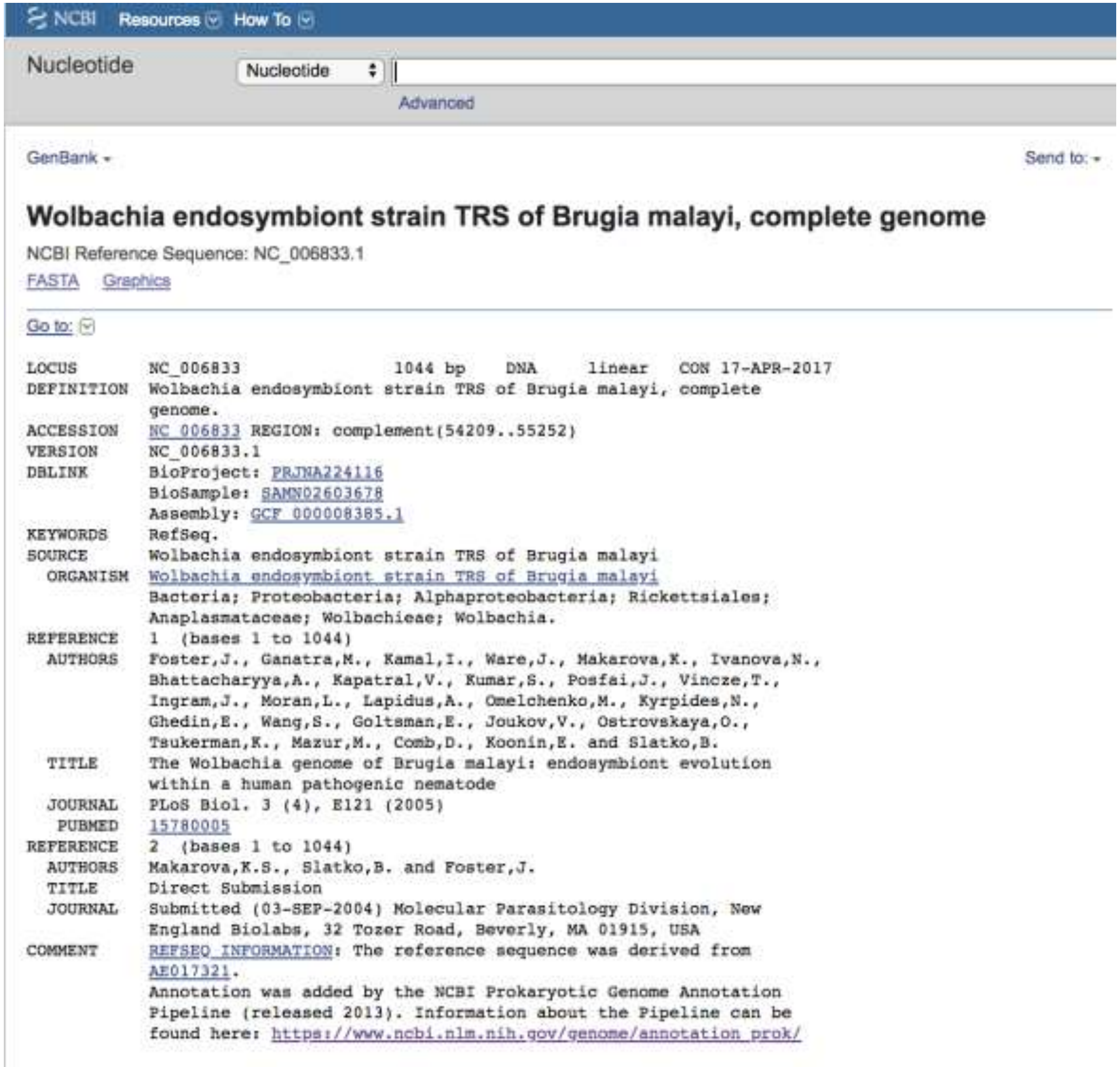


aspartate-semialdehyde dehydrogenase

GenPept FASTA Graphics BLAST

Name: aspartate-semialdehyde dehydrogenase
RefSeq Selected Product: WP_011256244.1, 347 amino acids
Taxonomic Group: a-proteobacteria
Assembly Accessions: 2
Protein Accessions: 2
CDS Regions: 2
Total Rows: 2

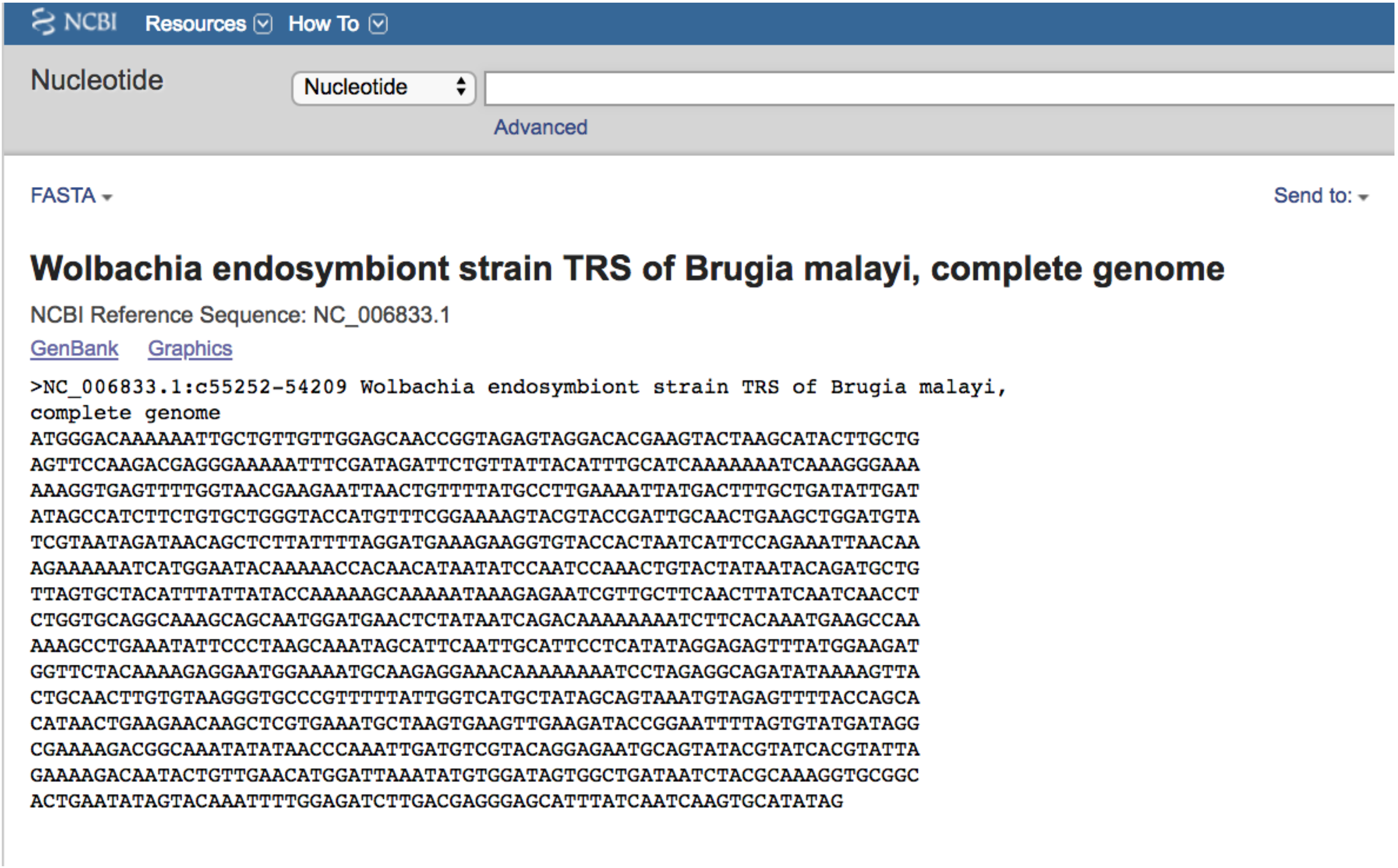
Source	CDS Region in Nucleotide	Protein	Name	Organism	Strain	Assembly
RefSeq	NC_006833.1 54209-55252 (-)	WP_011256244.1	aspartate-semialdehyde dehydrogenase	Wolbachia endosymbiont strain TRS of Brugia malayi		GCF_000008385.1
INSDC	AE017321.1 54209-55252 (-)	AAW70634.1	Aspartate-semialdehyde dehydrogenase	Wolbachia endosymbiont strain TRS of Brugia malayi		GCA_000008385.1



Wolbachia endosymbiont strain TRS of Brugia malayi, complete genome

NCBI Reference Sequence: NC_006833.1

```
LOCUS       NC_006833               1044 bp    DNA     linear   CON 17-APR-2017
DEFINITION  Wolbachia endosymbiont strain TRS of Brugia malayi, complete
            genome.
ACCESSION   NC_006833 REGION: complement(54209..55252)
VERSION     NC_006833.1
DBLINK      BioProject: PRJNA224116
            BioSample: SAMN02603678
            Assembly: GCF_000008385.1
KEYWORDS    RefSeq.
SOURCE      Wolbachia endosymbiont strain TRS of Brugia malayi
  ORGANISM  Wolbachia endosymbiont strain TRS of Brugia malayi
            Bacteria; Proteobacteria; Alphaproteobacteria; Rickettsiales;
            Anaplasmataceae; Wolbachieae; Wolbachia.
REFERENCE   1  (bases 1 to 1044)
  AUTHORS   Foster,J., Ganatra,M., Kamal,I., Ware,J., Makarova,K., Ivanova,N.,
            Bhattacharyya,A., Kapatral,V., Kumar,S., Posfai,J., Vincze,T.,
            Ingram,J., Moran,L., Lapidus,A., Omelchenko,M., Kyrpides,N.,
            Ghedin,E., Wang,S., Goltsman,E., Joukov,V., Ostrovskaya,O.,
            Tsukerman,K., Mazur,M., Comb,D., Koonin,E. and Slatko,B.
  TITLE     The Wolbachia genome of Brugia malayi: endosymbiont evolution
            within a human pathogenic nematode
  JOURNAL   PLoS Biol. 3 (4), E121 (2005)
   PUBMED   15780005
REFERENCE   2  (bases 1 to 1044)
  AUTHORS   Makarova,K.S., Slatko,B. and Foster,J.
  TITLE     Direct Submission
  JOURNAL   Submitted (03-SEP-2004) Molecular Parasitology Division, New
            England Biolabs, 32 Tozer Road, Beverly, MA 01915, USA
COMMENT     REFSEQ INFORMATION: The reference sequence was derived from
            AE017321.
            Annotation was added by the NCBI Prokaryotic Genome Annotation
            Pipeline (released 2013). Information about the Pipeline can be
            found here: https://www.ncbi.nlm.nih.gov/genome/annotation_prok/
```

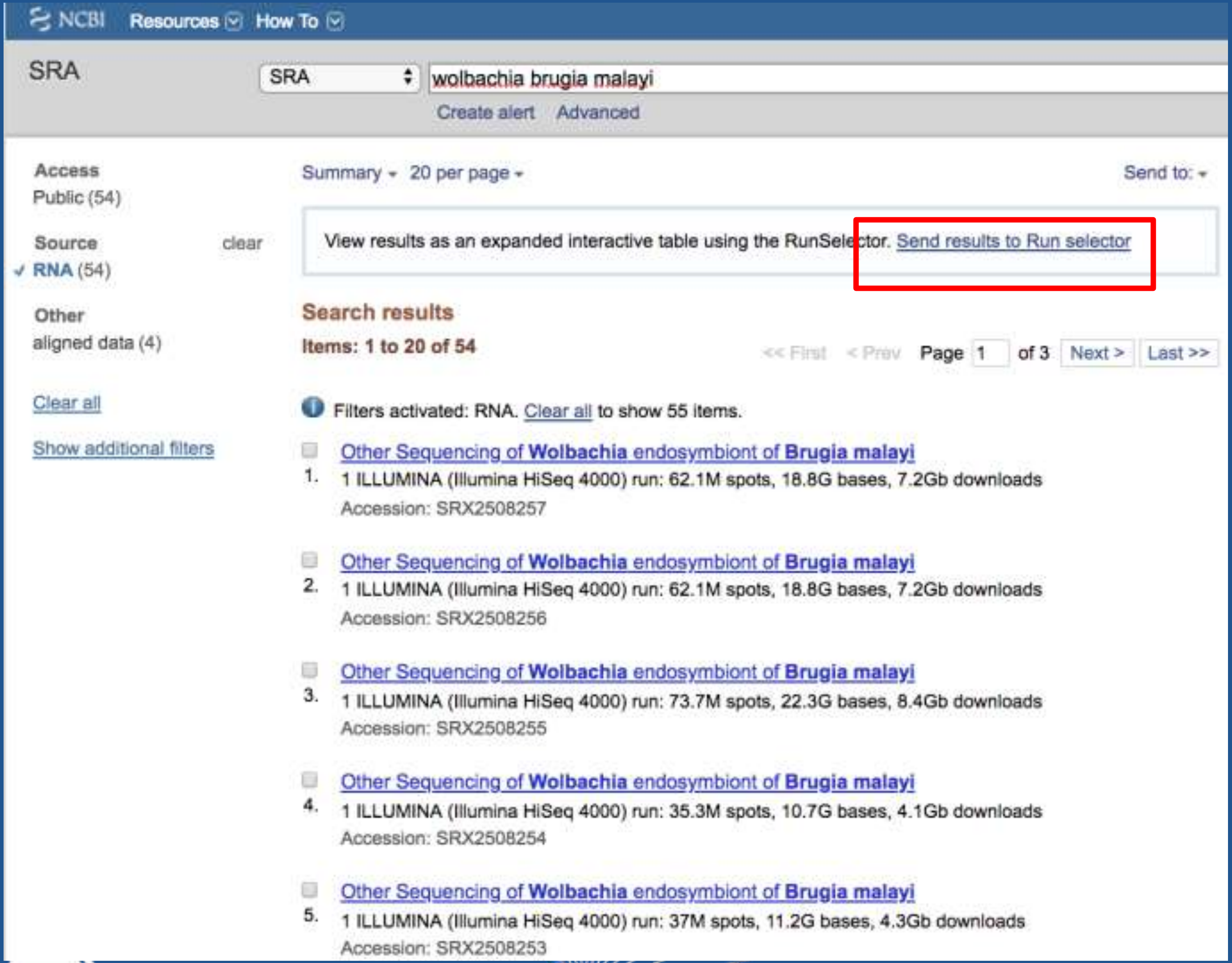


Wolbachia endosymbiont strain TRS of Brugia malayi, complete genome

NCBI Reference Sequence: NC_006833.1

```
>NC_006833.1:c55252-54209 Wolbachia endosymbiont strain TRS of Brugia malayi,
complete genome
ATGGGACAAAAAATTGCTGTTGTTGGAGCAACCGGTAGAGTAGGACACGAAGTACTAAGCATACTTGCTG
AGTTCCAAGACGAGGGAAAAATTTCGATAGATTCTGTTATTACATTTGCATCAAAAAAATCAAAGGGAAA
AAAGGTGAGTTTTGGTAACGAAGAATTAACTGTTTTATGCCTTGAAAATTATGACTTTGCTGATATTGAT
ATAGCCATCTTCTGTGCTGGGTACCATGTTTCGGAAAAGTACGTACCGATTGCAACTGAAGCTGGATGTA
TCGTAATAGATAACAGCTCTTATTTTAGGATGAAAGAAGGTGTACCACTAATCATTCCAGAAATTAACAA
AGAAAAAATCATGGAATACAAAAACCACAACATAATATCCAATCCAAACTGTACTATAATACAGATGCTG
TTAGTGCTACATTTATTATACCAAAAAGCAAAAATAAAGAGAATCGTTGCTTCAACTTATCAATCAACCT
CTGGTGCAGGCAAAGCAGCAATGGATGAACTCTATAATCAGACAAAAAAAATCTTCACAAATGAAGCCAA
AAAGCCTGAAATATTCCCTAAGCAAATAGCATTCAATTGCATTCCTCATATAGGAGAGTTTATGGAAGAT
GGTTCTACAAAAGAGGAATGGAAAATGCAAGAGGAAACAAAAAAAATCCTAGAGGCAGATATAAAAGTTA
CTGCAACTTGTGTAAGGGTGCCCGTTTTTATTGGTCATGCTATAGCAGTAAATGTAGAGTTTTACCAGCA
CATAACTGAAGAACAAGCTCGTGAAATGCTAAGTGAAGTTGAAGATACCGGAATTTTAGTGTATGATAGG
CGAAAAGACGGCAAATATATAACCCAAATTGATGTCGTACAGGAGAATGCAGTATACGTATCACGTATTA
GAAAAGACAATACTGTTGAACATGGATTAAATATGTGGATAGTGGCTGATAATCTACGCAAAGGTGCGGC
ACTGAATATAGTACAAATTTTGGAGATCTTGACGAGGGAGCATTTATCAATCAAGTGCATATAG
```

So here's some data…



wolbachia brugia malayi

Access
Public (54)

Source clear
✓ RNA (54)

Other
aligned data (4)

Clear all

Show additional filters

View results as an expanded interactive table using the RunSelector. Send results to Run selector

Search results

Items: 1 to 20 of 54

<< First < Prev Page 1 of 3 Next > Last >>

Filters activated: RNA. Clear all to show 55 items.

1. □ Other Sequencing of **Wolbachia** endosymbiont of **Brugia malayi**
   1 ILLUMINA (Illumina HiSeq 4000) run: 62.1M spots, 18.8G bases, 7.2Gb downloads
   Accession: SRX2508257

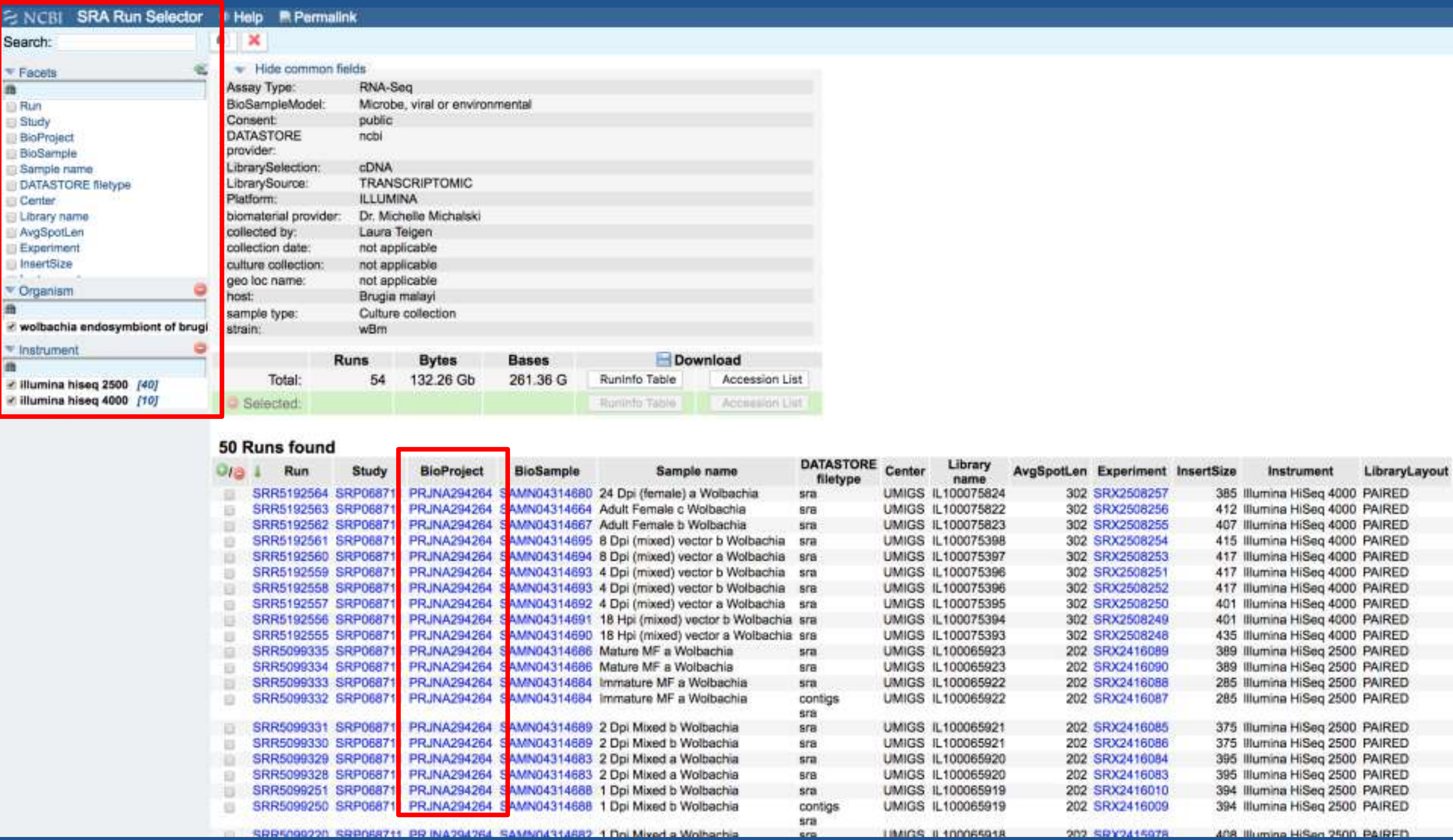
2. □ Other Sequencing of **Wolbachia** endosymbiont of **Brugia malayi**
   1 ILLUMINA (Illumina HiSeq 4000) run: 62.1M spots, 18.8G bases, 7.2Gb downloads
   Accession: SRX2508256

3. □ Other Sequencing of **Wolbachia** endosymbiont of **Brugia malayi**
   1 ILLUMINA (Illumina HiSeq 4000) run: 73.7M spots, 22.3G bases, 8.4Gb downloads
   Accession: SRX2508255

4. □ Other Sequencing of **Wolbachia** endosymbiont of **Brugia malayi**
   1 ILLUMINA (Illumina HiSeq 4000) run: 35.3M spots, 10.7G bases, 4.1Gb downloads
   Accession: SRX2508254

5. □ Other Sequencing of **Wolbachia** endosymbiont of **Brugia malayi**
   1 ILLUMINA (Illumina HiSeq 4000) run: 37M spots, 11.2G bases, 4.3Gb downloads
   Accession: SRX2508253

…and the metadata isn't horrible!



NCBI SRA Run Selector | Help | Permalink

Search: ✕

▼ Facets

- □ Run
- □ Study
- □ BioProject
- □ BioSample
- □ Sample name
- □ DATASTORE filetype
- □ Center
- □ Library name
- □ AvgSpotLen
- □ Experiment
- □ InsertSize

▼ Organism

- ☑ wolbachia endosymbiont of brugi

▼ Instrument

- ☑ illumina hiseq 2500 [40]
- ☑ illumina hiseq 4000 [10]

▼ Hide common fields

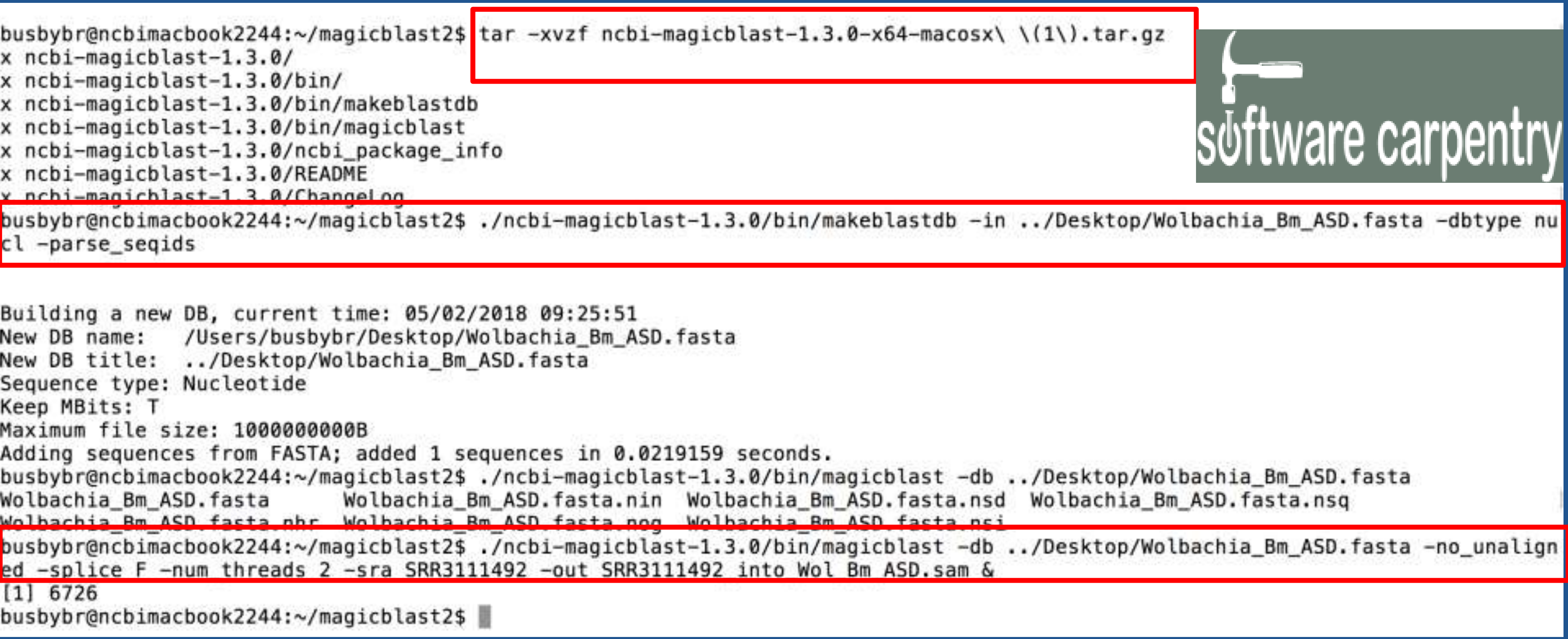
Assay Type:	RNA-Seq
BioSampleModel:	Microbe, viral or environmental
Consent:	public
DATASTORE provider:	ncbi
LibrarySelection:	cDNA
LibrarySource:	TRANSCRIPTOMIC
Platform:	ILLUMINA
biomaterial provider:	Dr. Michelle Michalski
collected by:	Laura Teigen
collection date:	not applicable
culture collection:	not applicable
geo loc name:	not applicable
host:	Brugia malayi
sample type:	Culture collection
strain:	wBm

	Runs	Bytes	Bases	Download	
Total:	54	132.26 Gb	261.36 G	RunInfo Table	Accession List
Selected:				RunInfo Table	Accession List

50 Runs found

	Run	Study	BioProject	BioSample	Sample name	DATASTORE filetype	Center	Library name	AvgSpotLen	Experiment	InsertSize	Instrument	LibraryLayout
□	SRR5192564	SRP06871	PRJNA294264	SAMN04314680	24 Dpi (female) a Wolbachia	sra	UMIGS	IL100075824	302	SRX2508257	385	Illumina HiSeq 4000	PAIRED
□	SRR5192563	SRP06871	PRJNA294264	SAMN04314664	Adult Female c Wolbachia	sra	UMIGS	IL100075822	302	SRX2508256	412	Illumina HiSeq 4000	PAIRED
□	SRR5192562	SRP06871	PRJNA294264	SAMN04314667	Adult Female b Wolbachia	sra	UMIGS	IL100075823	302	SRX2508255	407	Illumina HiSeq 4000	PAIRED
□	SRR5192561	SRP06871	PRJNA294264	SAMN04314695	8 Dpi (mixed) vector b Wolbachia	sra	UMIGS	IL100075398	302	SRX2508254	415	Illumina HiSeq 4000	PAIRED
□	SRR5192560	SRP06871	PRJNA294264	SAMN04314694	8 Dpi (mixed) vector a Wolbachia	sra	UMIGS	IL100075397	302	SRX2508253	417	Illumina HiSeq 4000	PAIRED
□	SRR5192559	SRP06871	PRJNA294264	SAMN04314693	4 Dpi (mixed) vector b Wolbachia	sra	UMIGS	IL100075396	302	SRX2508251	417	Illumina HiSeq 4000	PAIRED
□	SRR5192558	SRP06871	PRJNA294264	SAMN04314693	4 Dpi (mixed) vector b Wolbachia	sra	UMIGS	IL100075396	302	SRX2508252	417	Illumina HiSeq 4000	PAIRED
□	SRR5192557	SRP06871	PRJNA294264	SAMN04314692	4 Dpi (mixed) vector a Wolbachia	sra	UMIGS	IL100075395	302	SRX2508250	401	Illumina HiSeq 4000	PAIRED
□	SRR5192556	SRP06871	PRJNA294264	SAMN04314691	18 Hpi (mixed) vector b Wolbachia	sra	UMIGS	IL100075394	302	SRX2508249	401	Illumina HiSeq 4000	PAIRED
□	SRR5192555	SRP06871	PRJNA294264	SAMN04314690	18 Hpi (mixed) vector a Wolbachia	sra	UMIGS	IL100075393	302	SRX2508248	435	Illumina HiSeq 4000	PAIRED
□	SRR5099335	SRP06871	PRJNA294264	SAMN04314686	Mature MF a Wolbachia	sra	UMIGS	IL100065923	202	SRX2416089	389	Illumina HiSeq 2500	PAIRED
□	SRR5099334	SRP06871	PRJNA294264	SAMN04314686	Mature MF a Wolbachia	sra	UMIGS	IL100065923	202	SRX2416090	389	Illumina HiSeq 2500	PAIRED
□	SRR5099333	SRP06871	PRJNA294264	SAMN04314684	Immature MF a Wolbachia	sra	UMIGS	IL100065922	202	SRX2416088	285	Illumina HiSeq 2500	PAIRED
□	SRR5099332	SRP06871	PRJNA294264	SAMN04314684	Immature MF a Wolbachia	contigs sra	UMIGS	IL100065922	202	SRX2416087	285	Illumina HiSeq 2500	PAIRED
□	SRR5099331	SRP06871	PRJNA294264	SAMN04314689	2 Dpi Mixed b Wolbachia	sra	UMIGS	IL100065921	202	SRX2416085	375	Illumina HiSeq 2500	PAIRED
□	SRR5099330	SRP06871	PRJNA294264	SAMN04314689	2 Dpi Mixed b Wolbachia	sra	UMIGS	IL100065921	202	SRX2416086	375	Illumina HiSeq 2500	PAIRED
□	SRR5099329	SRP06871	PRJNA294264	SAMN04314683	2 Dpi Mixed a Wolbachia	sra	UMIGS	IL100065920	202	SRX2416084	395	Illumina HiSeq 2500	PAIRED
□	SRR5099328	SRP06871	PRJNA294264	SAMN04314683	2 Dpi Mixed a Wolbachia	sra	UMIGS	IL100065920	202	SRX2416083	395	Illumina HiSeq 2500	PAIRED
□	SRR5099251	SRP06871	PRJNA294264	SAMN04314688	1 Dpi Mixed b Wolbachia	sra	UMIGS	IL100065919	202	SRX2416010	394	Illumina HiSeq 2500	PAIRED
□	SRR5099250	SRP06871	PRJNA294264	SAMN04314688	1 Dpi Mixed b Wolbachia	contigs sra	UMIGS	IL100065919	202	SRX2416009	394	Illumina HiSeq 2500	PAIRED
□	SRR5099220	SRP068711	PRJNA294264	SAMN04314682	1 Dpi Mixed a Wolbachia	sra	UMIGS	IL100065918	202	SRX2415978	408	Illumina HiSeq 2500	PAIRED

Make a BLAST db, and go!



```
busbybr@ncbimacbook2244:~/magicblast2$ tar -xvzf ncbi-magicblast-1.3.0-x64-macosx\ \(1\).tar.gz
x ncbi-magicblast-1.3.0/
x ncbi-magicblast-1.3.0/bin/
x ncbi-magicblast-1.3.0/bin/makeblastdb
x ncbi-magicblast-1.3.0/bin/magicblast
x ncbi-magicblast-1.3.0/ncbi_package_info
x ncbi-magicblast-1.3.0/README
x ncbi-magicblast-1.3.0/ChangeLog
busbybr@ncbimacbook2244:~/magicblast2$ ./ncbi-magicblast-1.3.0/bin/makeblastdb -in ../Desktop/Wolbachia_Bm_ASD.fasta -dbtype nu
cl -parse_seqids


Building a new DB, current time: 05/02/2018 09:25:51
New DB name:   /Users/busbybr/Desktop/Wolbachia_Bm_ASD.fasta
New DB title:  ../Desktop/Wolbachia_Bm_ASD.fasta
Sequence type: Nucleotide
Keep MBits: T
Maximum file size: 1000000000B
Adding sequences from FASTA; added 1 sequences in 0.0219159 seconds.
busbybr@ncbimacbook2244:~/magicblast2$ ./ncbi-magicblast-1.3.0/bin/magicblast -db ../Desktop/Wolbachia_Bm_ASD.fasta
Wolbachia_Bm_ASD.fasta      Wolbachia_Bm_ASD.fasta.nin  Wolbachia_Bm_ASD.fasta.nsd  Wolbachia_Bm_ASD.fasta.nsq
Wolbachia_Bm_ASD.fasta.nhr  Wolbachia_Bm_ASD.fasta.nog  Wolbachia_Bm_ASD.fasta.nsi
busbybr@ncbimacbook2244:~/magicblast2$ ./ncbi-magicblast-1.3.0/bin/magicblast -db ../Desktop/Wolbachia_Bm_ASD.fasta -no_unalign
ed -splice F -num_threads 2 -sra SRR3111492 -out SRR3111492_into_Wol_Bm_ASD.sam &
[1] 6726
busbybr@ncbimacbook2244:~/magicblast2$ 
```

software carpentry

NIH U.S. National Library of Medicine

NCBI

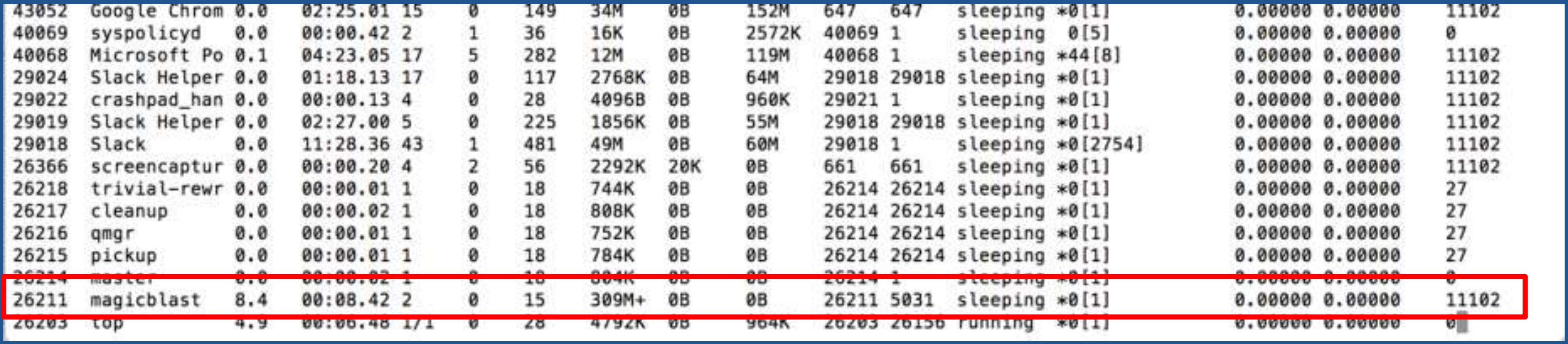
Make a BLAST db, and go!

```
tar -xvzf ncbi-magicblast...

makeblastdb -dbtype nucl -in <fasta> -parse_seqids

magicblast -db <fasta> -sra SRR... -splice F -no_unaligned
```

It runs pretty quick…

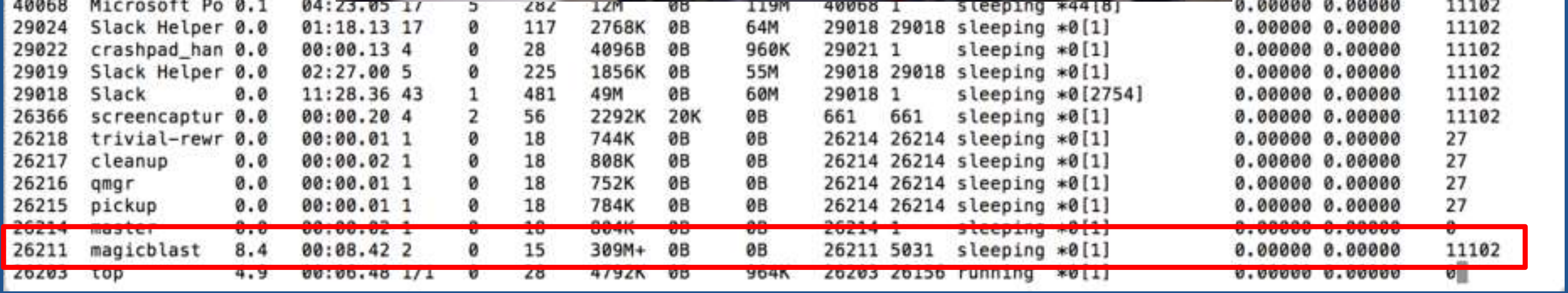
magicblast -db <fasta> -sra SRR… -splice F -no_unaligned

-num_threads X



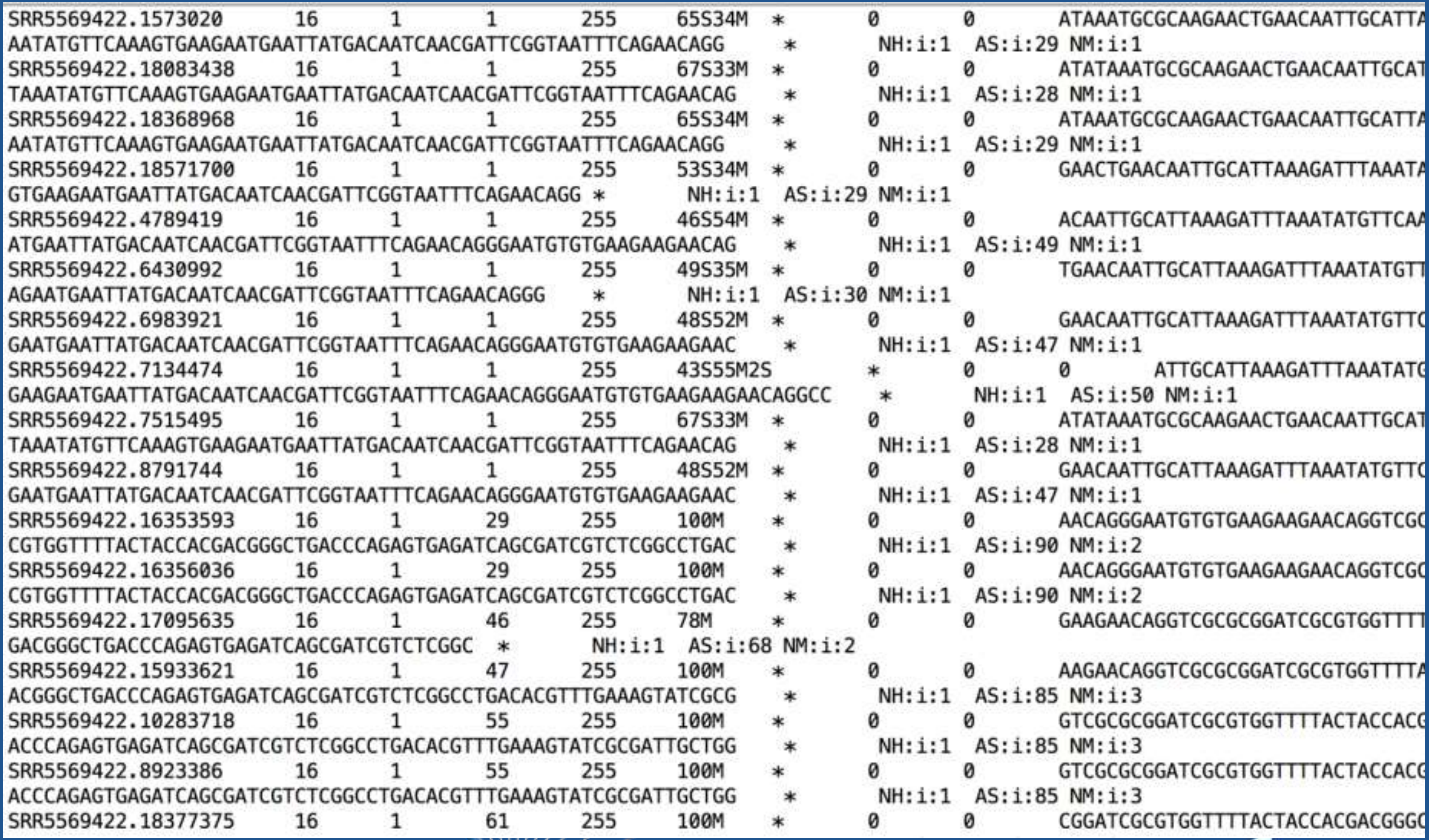
```
43052  Google Chrom 0.0  02:25.01 15   0   149  34M    0B   152M   647   647   sleeping *0[1]        0.00000 0.00000   11102
40069  syspolicyd   0.0  00:00.42 2    1   36   16K    0B   2572K  40069 1     sleeping  0[5]        0.00000 0.00000   0
40068  Microsoft Po 0.1  04:23.05 17   5   282  12M    0B   119M   40068 1     sleeping *44[8]       0.00000 0.00000   11102
29024  Slack Helper 0.0  01:18.13 17   0   117  2768K  0B   64M    29018 29018 sleeping *0[1]        0.00000 0.00000   11102
29022  crashpad_han 0.0  00:00.13 4    0   28   4096B  0B   960K   29021 1     sleeping *0[1]        0.00000 0.00000   11102
29019  Slack Helper 0.0  02:27.00 5    0   225  1856K  0B   55M    29018 29018 sleeping *0[1]        0.00000 0.00000   11102
29018  Slack        0.0  11:28.36 43   1   481  49M    0B   60M    29018 1     sleeping *0[2754]     0.00000 0.00000   11102
26366  screencaptur 0.0  00:00.20 4    2   56   2292K  20K  0B     661   661   sleeping *0[1]        0.00000 0.00000   11102
26218  trivial-rewr 0.0  00:00.01 1    0   18   744K   0B   0B     26214 26214 sleeping *0[1]        0.00000 0.00000   27
26217  cleanup      0.0  00:00.02 1    0   18   808K   0B   0B     26214 26214 sleeping *0[1]        0.00000 0.00000   27
26216  qmgr         0.0  00:00.01 1    0   18   752K   0B   0B     26214 26214 sleeping *0[1]        0.00000 0.00000   27
26215  pickup       0.0  00:00.01 1    0   18   784K   0B   0B     26214 26214 sleeping *0[1]        0.00000 0.00000   27
26214  master       0.0  00:00.02 1    0   18   804K   0B   0B     26214 1     sleeping *0[1]        0.00000 0.00000   0
26211  magicblast   8.4  00:08.42 2    0   15   309M+  0B   0B     26211 5031  sleeping *0[1]        0.00000 0.00000   11102
26203  top          4.9  00:06.48 1/1  0   28   4792K  0B   964K   26203 26156 running  *0[1]        0.00000 0.00000   0
```

Even on a plane!



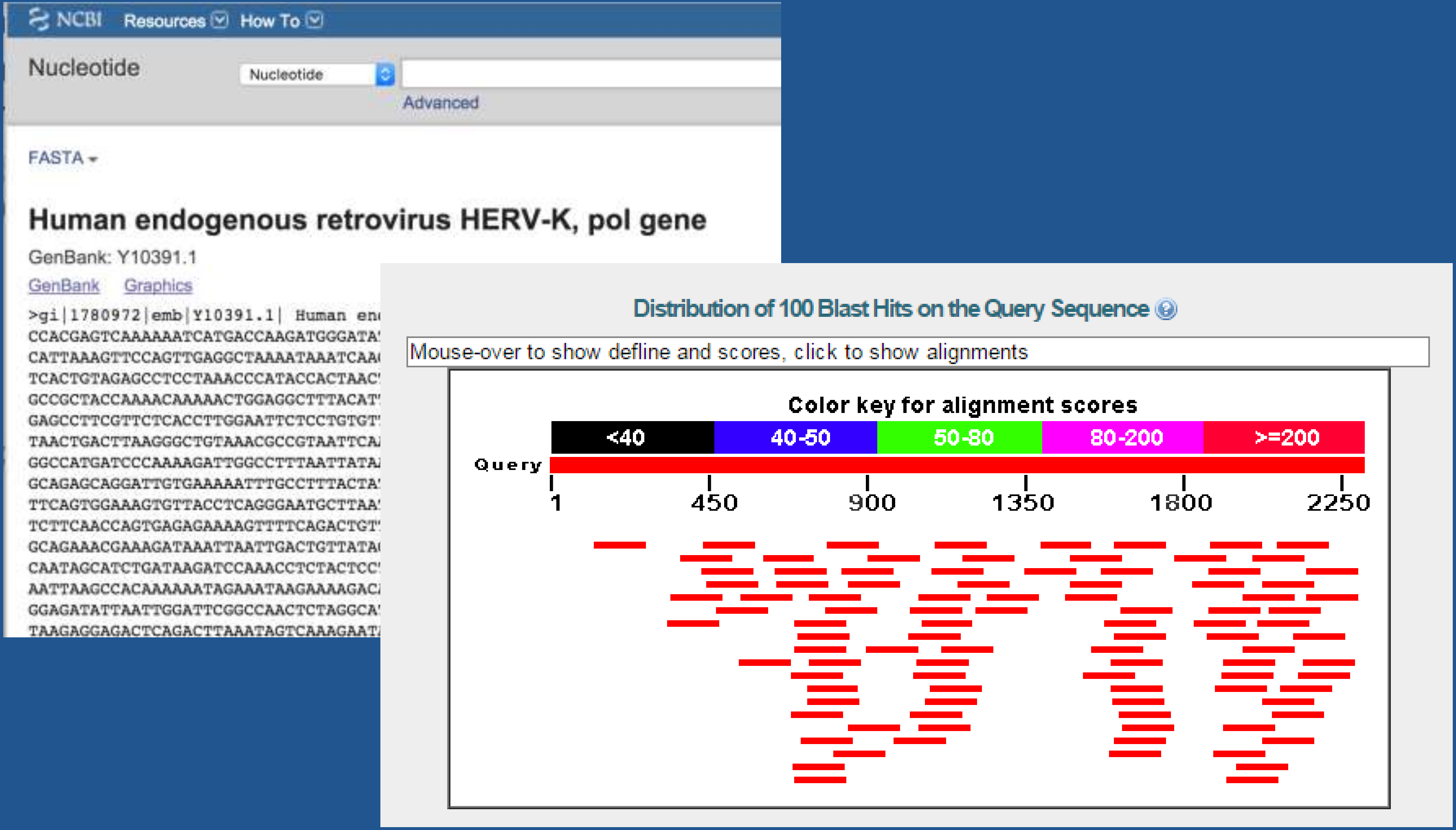
```
43052  Google Chrom 0.0   02:                                                                   0.00000 0.00000    11102
40069  syspolicyd   0.0   00:                                                                   0.00000 0.00000    0
40068  Microsoft Po 0.1   04:23.05 17      5     282    12M    0B     119M   40068 1     sleeping *44[8]        0.00000 0.00000    11102
29024  Slack Helper 0.0   01:18.13 17      0     117    2768K  0B     64M    29018 29018 sleeping *0[1]         0.00000 0.00000    11102
29022  crashpad_han 0.0   00:00.13 4       0     28     4096B  0B     960K   29021 1     sleeping *0[1]         0.00000 0.00000    11102
29019  Slack Helper 0.0   02:27.00 5       0     225    1856K  0B     55M    29018 29018 sleeping *0[1]         0.00000 0.00000    11102
29018  Slack        0.0   11:28.36 43      1     481    49M    0B     60M    29018 1     sleeping *0[2754]      0.00000 0.00000    11102
26366  screencaptur 0.0   00:00.20 4       2     56     2292K  20K    0B     661   661   sleeping *0[1]         0.00000 0.00000    11102
26218  trivial-rewr 0.0   00:00.01 1       0     18     744K   0B     0B     26214 26214 sleeping *0[1]         0.00000 0.00000    27
26217  cleanup      0.0   00:00.02 1       0     18     808K   0B     0B     26214 26214 sleeping *0[1]         0.00000 0.00000    27
26216  qmgr         0.0   00:00.01 1       0     18     752K   0B     0B     26214 26214 sleeping *0[1]         0.00000 0.00000    27
26215  pickup       0.0   00:00.01 1       0     18     784K   0B     0B     26214 26214 sleeping *0[1]         0.00000 0.00000    27
26214  master       0.0   00:00.02 1       0     18     804K   0B     0B     26214 1     sleeping *0[1]         0.00000 0.00000    0
26211  magicblast   8.4   00:08.42 2       0     15     309M+  0B     0B     26211 5031  sleeping *0[1]         0.00000 0.00000    11102
26203  top          4.9   00:06.48 1/1     0     28     4792K  0B     964K   26203 26156 running  *0[1]         0.00000 0.00000    0
```

Bam! (well, .sam)



```
SRR5569422.1573020      16      1       1       255     65S34M  *       0       0       ATAAATGCGCAAGAACTGAACAATTGCATTA
AATATGTTCAAAGTGAAGAATGAATTATGACAATCAACGATTCGGTAATTTCAGAACAGG    *       NH:i:1  AS:i:29 NM:i:1
SRR5569422.18083438     16      1       1       255     67S33M  *       0       0       ATATAAATGCGCAAGAACTGAACAATTGCAT
TAAATATGTTCAAAGTGAAGAATGAATTATGACAATCAACGATTCGGTAATTTCAGAACAG   *       NH:i:1  AS:i:28 NM:i:1
SRR5569422.18368968     16      1       1       255     65S34M  *       0       0       ATAAATGCGCAAGAACTGAACAATTGCATTA
AATATGTTCAAAGTGAAGAATGAATTATGACAATCAACGATTCGGTAATTTCAGAACAGG    *       NH:i:1  AS:i:29 NM:i:1
SRR5569422.18571700     16      1       1       255     53S34M  *       0       0       GAACTGAACAATTGCATTAAAGATTTAAATA
GTGAAGAATGAATTATGACAATCAACGATTCGGTAATTTCAGAACAGG *       NH:i:1  AS:i:29 NM:i:1
SRR5569422.4789419      16      1       1       255     46S54M  *       0       0       ACAATTGCATTAAAGATTTAAATATGTTCAA
ATGAATTATGACAATCAACGATTCGGTAATTTCAGAACAGGGAATGTGTGAAGAAGAACAG   *       NH:i:1  AS:i:49 NM:i:1
SRR5569422.6430992      16      1       1       255     49S35M  *       0       0       TGAACAATTGCATTAAAGATTTAAATATGTT
AGAATGAATTATGACAATCAACGATTCGGTAATTTCAGAACAGGG   *       NH:i:1  AS:i:30 NM:i:1
SRR5569422.6983921      16      1       1       255     48S52M  *       0       0       GAACAATTGCATTAAAGATTTAAATATGTTC
GAATGAATTATGACAATCAACGATTCGGTAATTTCAGAACAGGGAATGTGTGAAGAAGAAC   *       NH:i:1  AS:i:47 NM:i:1
SRR5569422.7134474      16      1       1       255     43S55M2S        *       0       0       ATTGCATTAAAGATTTAAATATG
GAAGAATGAATTATGACAATCAACGATTCGGTAATTTCAGAACAGGGAATGTGTGAAGAAGAACAGGCC   *       NH:i:1  AS:i:50 NM:i:1
SRR5569422.7515495      16      1       1       255     67S33M  *       0       0       ATATAAATGCGCAAGAACTGAACAATTGCAT
TAAATATGTTCAAAGTGAAGAATGAATTATGACAATCAACGATTCGGTAATTTCAGAACAG   *       NH:i:1  AS:i:28 NM:i:1
SRR5569422.8791744      16      1       1       255     48S52M  *       0       0       GAACAATTGCATTAAAGATTTAAATATGTTC
GAATGAATTATGACAATCAACGATTCGGTAATTTCAGAACAGGGAATGTGTGAAGAAGAAC   *       NH:i:1  AS:i:47 NM:i:1
SRR5569422.16353593     16      1       29      255     100M    *       0       0       AACAGGGAATGTGTGAAGAAGAACAGGTCGC
CGTGGTTTTACTACCACGACGGGCTGACCCAGAGTGAGATCAGCGATCGTCTCGGCCTGAC   *       NH:i:1  AS:i:90 NM:i:2
SRR5569422.16356036     16      1       29      255     100M    *       0       0       AACAGGGAATGTGTGAAGAAGAACAGGTCGC
CGTGGTTTTACTACCACGACGGGCTGACCCAGAGTGAGATCAGCGATCGTCTCGGCCTGAC   *       NH:i:1  AS:i:90 NM:i:2
SRR5569422.17095635     16      1       46      255     78M     *       0       0       GAAGAACAGGTCGCGCGGATCGCGTGGTTTT
GACGGGCTGACCCAGAGTGAGATCAGCGATCGTCTCGGC *       NH:i:1  AS:i:68 NM:i:2
SRR5569422.15933621     16      1       47      255     100M    *       0       0       AAGAACAGGTCGCGCGGATCGCGTGGTTTTA
ACGGGCTGACCCAGAGTGAGATCAGCGATCGTCTCGGCCTGACACGTTTGAAAGTATCGCG   *       NH:i:1  AS:i:85 NM:i:3
SRR5569422.10283718     16      1       55      255     100M    *       0       0       GTCGCGCGGATCGCGTGGTTTTACTACCACG
ACCCAGAGTGAGATCAGCGATCGTCTCGGCCTGACACGTTTGAAAGTATCGCGATTGCTGG   *       NH:i:1  AS:i:85 NM:i:3
SRR5569422.8923386      16      1       55      255     100M    *       0       0       GTCGCGCGGATCGCGTGGTTTTACTACCACG
ACCCAGAGTGAGATCAGCGATCGTCTCGGCCTGACACGTTTGAAAGTATCGCGATTGCTGG   *       NH:i:1  AS:i:85 NM:i:3
SRR5569422.18377375     16      1       61      255     100M    *       0       0       CGGATCGCGTGGTTTTACTACCACGACGGGC
```

Similar searches on the web!



Human endogenous retrovirus HERV-K, pol gene

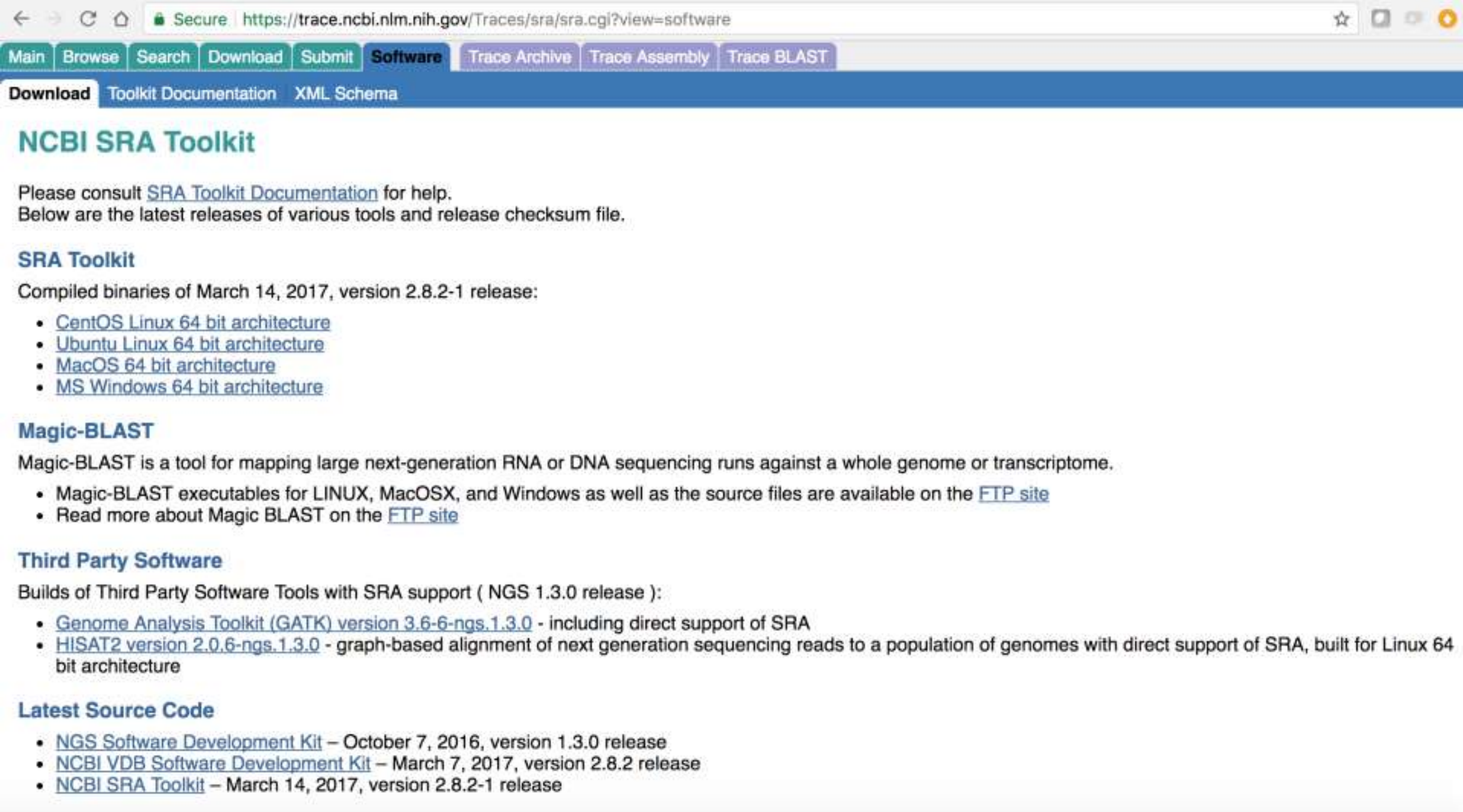
GenBank: Y10391.1

```
>gi|1780972|emb|Y10391.1| Human en
CCACGAGTCAAAAAATCATGACCAAGATGGGATA
CATTAAAGTTCCAGTTGAGGCTAAAATAAATCAA
TCACTGTAGAGCCTCCTAAACCCATACCACTAAC
GCCGCTACCAAAACAAAAACTGGAGGCTTTACAT
GAGCCTTCGTTCTCACCTTGGAATTCTCCTGTGT
TAACTGACTTAAGGGCTGTAAACGCCGTAATTCA
GGCCATGATCCCAAAAGATTGGCCTTTAATTATA
GCAGAGCAGGATTGTGAAAAATTTGCCTTTACTA
TTCAGTGGAAAGTGTTACCTCAGGGAATGCTTAA
TCTTCAACCAGTGAGAGAAAAGTTTTCAGACTGT
GCAGAAACGAAAGATAAATTAATTGACTGTTATA
CAATAGCATCTGATAAGATCCAAACCTCTACTCC
AATTAAGCCACAAAAAATAGAAATAAGAAAAGAC
GGAGATATTAATTGGATTCGGCCAACTCTAGGCA
TAAGAGGAGACTCAGACTTAAATAGTCAAAGAAT
```

Distribution of 100 Blast Hits on the Query Sequence

Mouse-over to show defline and scores, click to show alignments

Works with some other software!



NCBI SRA Toolkit

Please consult SRA Toolkit Documentation for help.
Below are the latest releases of various tools and release checksum file.

SRA Toolkit

Compiled binaries of March 14, 2017, version 2.8.2-1 release:

- CentOS Linux 64 bit architecture
- Ubuntu Linux 64 bit architecture
- MacOS 64 bit architecture
- MS Windows 64 bit architecture

Magic-BLAST

Magic-BLAST is a tool for mapping large next-generation RNA or DNA sequencing runs against a whole genome or transcriptome.

- Magic-BLAST executables for LINUX, MacOSX, and Windows as well as the source files are available on the FTP site
- Read more about Magic BLAST on the FTP site

Third Party Software

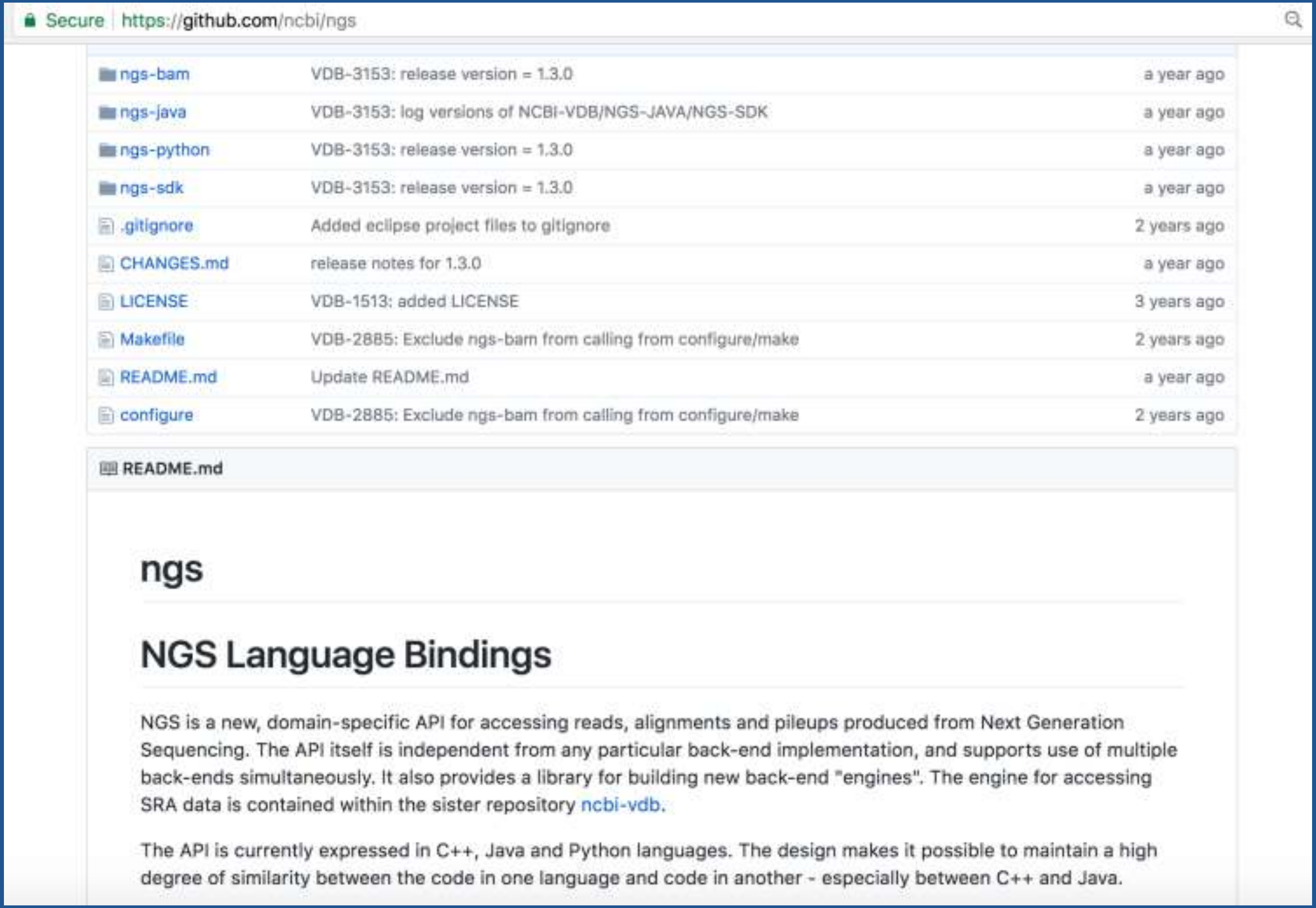
Builds of Third Party Software Tools with SRA support (NGS 1.3.0 release):

- Genome Analysis Toolkit (GATK) version 3.6-6-ngs.1.3.0 - including direct support of SRA
- HISAT2 version 2.0.6-ngs.1.3.0 - graph-based alignment of next generation sequencing reads to a population of genomes with direct support of SRA, built for Linux 64 bit architecture

Latest Source Code

- NGS Software Development Kit – October 7, 2016, version 1.3.0 release
- NCBI VDB Software Development Kit – March 7, 2017, version 2.8.2 release
- NCBI SRA Toolkit – March 14, 2017, version 2.8.2-1 release

This is how…



Secure https://github.com/ncbi/ngs

ngs-bam	VDB-3153: release version = 1.3.0	a year ago
ngs-java	VDB-3153: log versions of NCBI-VDB/NGS-JAVA/NGS-SDK	a year ago
ngs-python	VDB-3153: release version = 1.3.0	a year ago
ngs-sdk	VDB-3153: release version = 1.3.0	a year ago
.gitignore	Added eclipse project files to gitignore	2 years ago
CHANGES.md	release notes for 1.3.0	a year ago
LICENSE	VDB-1513: added LICENSE	3 years ago
Makefile	VDB-2885: Exclude ngs-bam from calling from configure/make	2 years ago
README.md	Update README.md	a year ago
configure	VDB-2885: Exclude ngs-bam from calling from configure/make	2 years ago

README.md

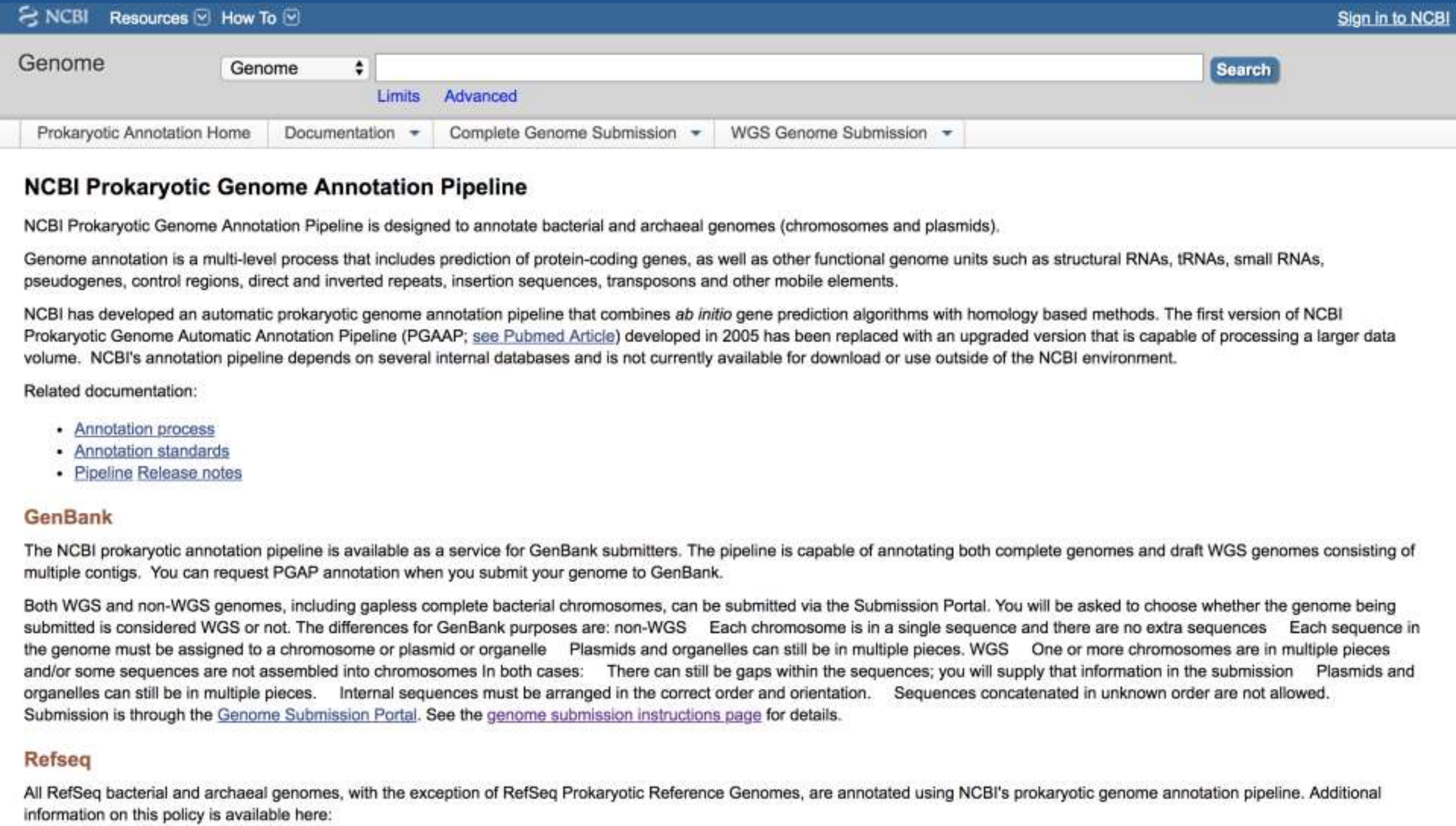
ngs

NGS Language Bindings

NGS is a new, domain-specific API for accessing reads, alignments and pileups produced from Next Generation Sequencing. The API itself is independent from any particular back-end implementation, and supports use of multiple back-ends simultaneously. It also provides a library for building new back-end "engines". The engine for accessing SRA data is contained within the sister repository ncbi-vdb.

The API is currently expressed in C++, Java and Python languages. The design makes it possible to maintain a high degree of similarity between the code in one language and code in another - especially between C++ and Java.

Prokaryotic Genome Annotation



NCBI Prokaryotic Genome Annotation Pipeline

NCBI Prokaryotic Genome Annotation Pipeline is designed to annotate bacterial and archaeal genomes (chromosomes and plasmids).

Genome annotation is a multi-level process that includes prediction of protein-coding genes, as well as other functional genome units such as structural RNAs, tRNAs, small RNAs, pseudogenes, control regions, direct and inverted repeats, insertion sequences, transposons and other mobile elements.

NCBI has developed an automatic prokaryotic genome annotation pipeline that combines *ab initio* gene prediction algorithms with homology based methods. The first version of NCBI Prokaryotic Genome Automatic Annotation Pipeline (PGAAP; see Pubmed Article) developed in 2005 has been replaced with an upgraded version that is capable of processing a larger data volume. NCBI's annotation pipeline depends on several internal databases and is not currently available for download or use outside of the NCBI environment.

Related documentation:

- Annotation process
- Annotation standards
- Pipeline Release notes

GenBank

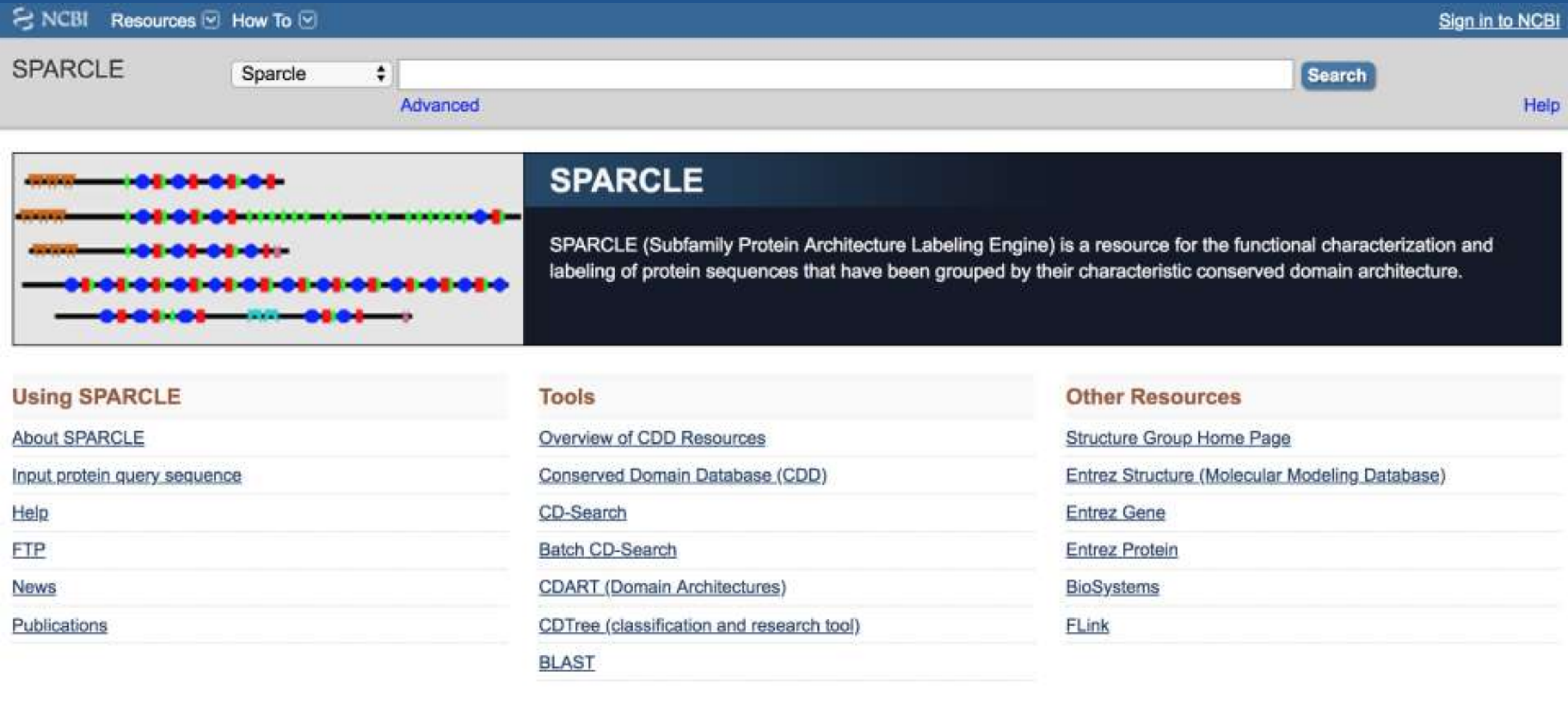
The NCBI prokaryotic annotation pipeline is available as a service for GenBank submitters. The pipeline is capable of annotating both complete genomes and draft WGS genomes consisting of multiple contigs. You can request PGAP annotation when you submit your genome to GenBank.

Both WGS and non-WGS genomes, including gapless complete bacterial chromosomes, can be submitted via the Submission Portal. You will be asked to choose whether the genome being submitted is considered WGS or not. The differences for GenBank purposes are: non-WGS Each chromosome is in a single sequence and there are no extra sequences Each sequence in the genome must be assigned to a chromosome or plasmid or organelle Plasmids and organelles can still be in multiple pieces. WGS One or more chromosomes are in multiple pieces and/or some sequences are not assembled into chromosomes In both cases: There can still be gaps within the sequences; you will supply that information in the submission Plasmids and organelles can still be in multiple pieces. Internal sequences must be arranged in the correct order and orientation. Sequences concatenated in unknown order are not allowed. Submission is through the Genome Submission Portal. See the genome submission instructions page for details.

Refseq

All RefSeq bacterial and archaeal genomes, with the exception of RefSeq Prokaryotic Reference Genomes, are annotated using NCBI's prokaryotic genome annotation pipeline. Additional information on this policy is available here:

Prokaryotic Genome Annotation



SPARCLE

SPARCLE (Subfamily Protein Architecture Labeling Engine) is a resource for the functional characterization and labeling of protein sequences that have been grouped by their characteristic conserved domain architecture.

Using SPARCLE

- About SPARCLE
- Input protein query sequence
- Help
- FTP
- News
- Publications

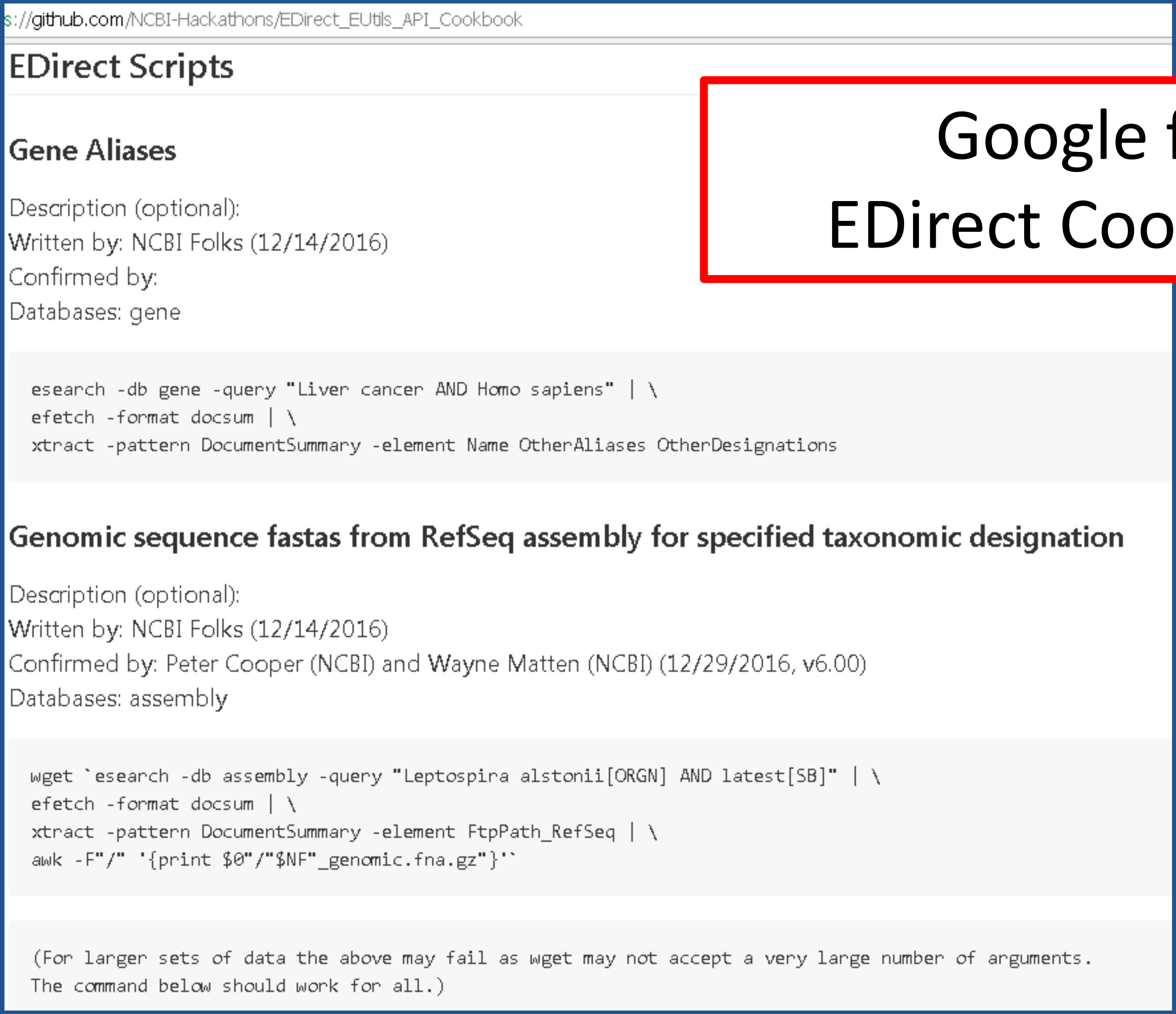
Tools

- Overview of CDD Resources
- Conserved Domain Database (CDD)
- CD-Search
- Batch CD-Search
- CDART (Domain Architectures)
- CDTree (classification and research tool)
- BLAST

Other Resources

- Structure Group Home Page
- Entrez Structure (Molecular Modeling Database)
- Entrez Gene
- Entrez Protein
- BioSystems
- FLink

EUtils (Search API) Command Line EDirect



s://github.com/NCBI-Hackathons/EDirect_EUtils_API_Cookbook

EDirect Scripts

Gene Aliases

Description (optional):
Written by: NCBI Folks (12/14/2016)
Confirmed by:
Databases: gene

```
esearch -db gene -query "Liver cancer AND Homo sapiens" | \
efetch -format docsum | \
xtract -pattern DocumentSummary -element Name OtherAliases OtherDesignations
```

Genomic sequence fastas from RefSeq assembly for specified taxonomic designation

Description (optional):
Written by: NCBI Folks (12/14/2016)
Confirmed by: Peter Cooper (NCBI) and Wayne Matten (NCBI) (12/29/2016, v6.00)
Databases: assembly

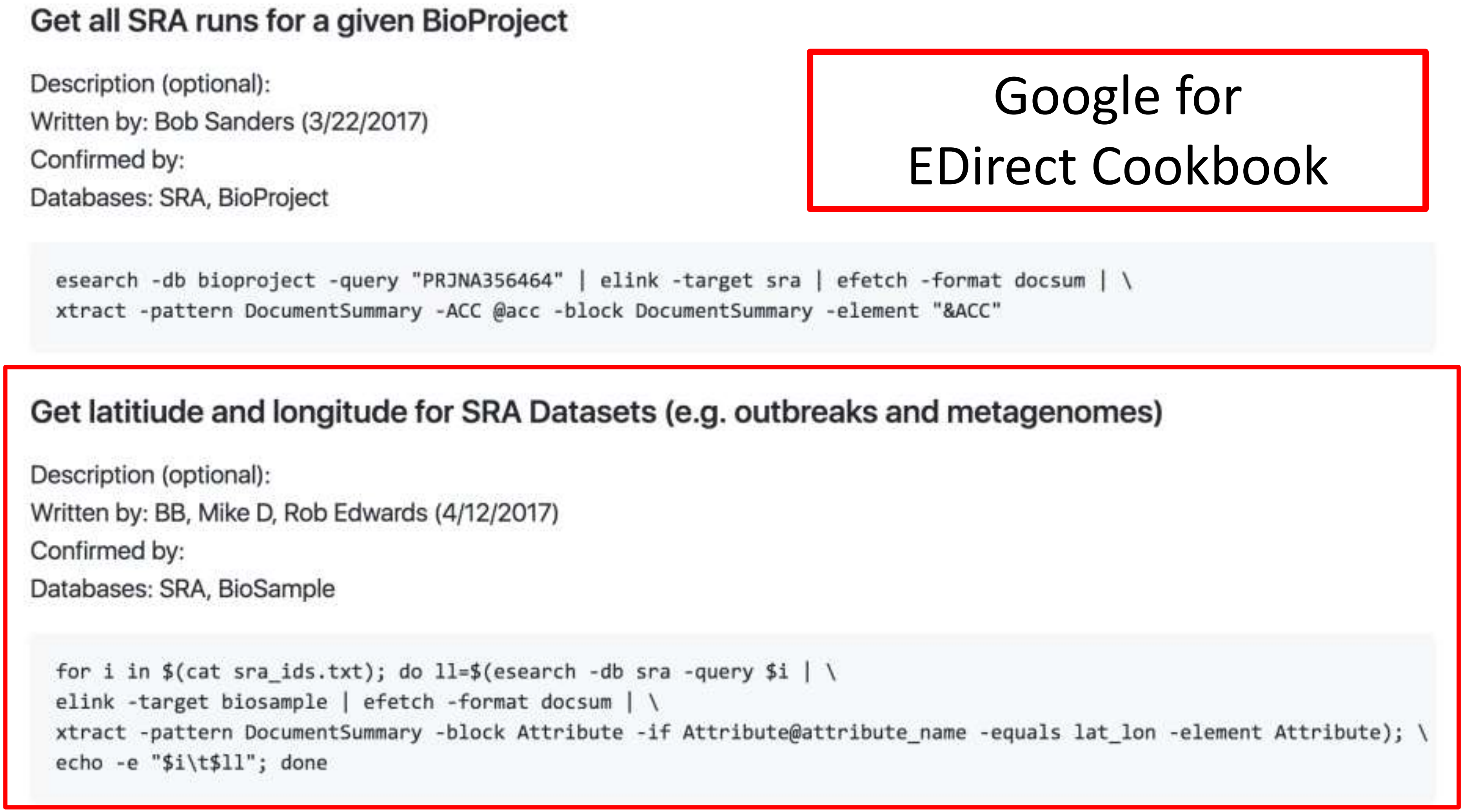
```
wget `esearch -db assembly -query "Leptospira alstonii[ORGN] AND latest[SB]" | \
efetch -format docsum | \
xtract -pattern DocumentSummary -element FtpPath_RefSeq | \
awk -F"/" '{print $0"/"$NF"_genomic.fna.gz"}'`
```

```
(For larger sets of data the above may fail as wget may not accept a very large number of arguments.
The command below should work for all.)
```

Google for
EDirect Cookbook

EUtils (Search API) Command Line EDirect

Google for
EDirect Cookbook



Get all SRA runs for a given BioProject

Description (optional):
Written by: Bob Sanders (3/22/2017)
Confirmed by:
Databases: SRA, BioProject

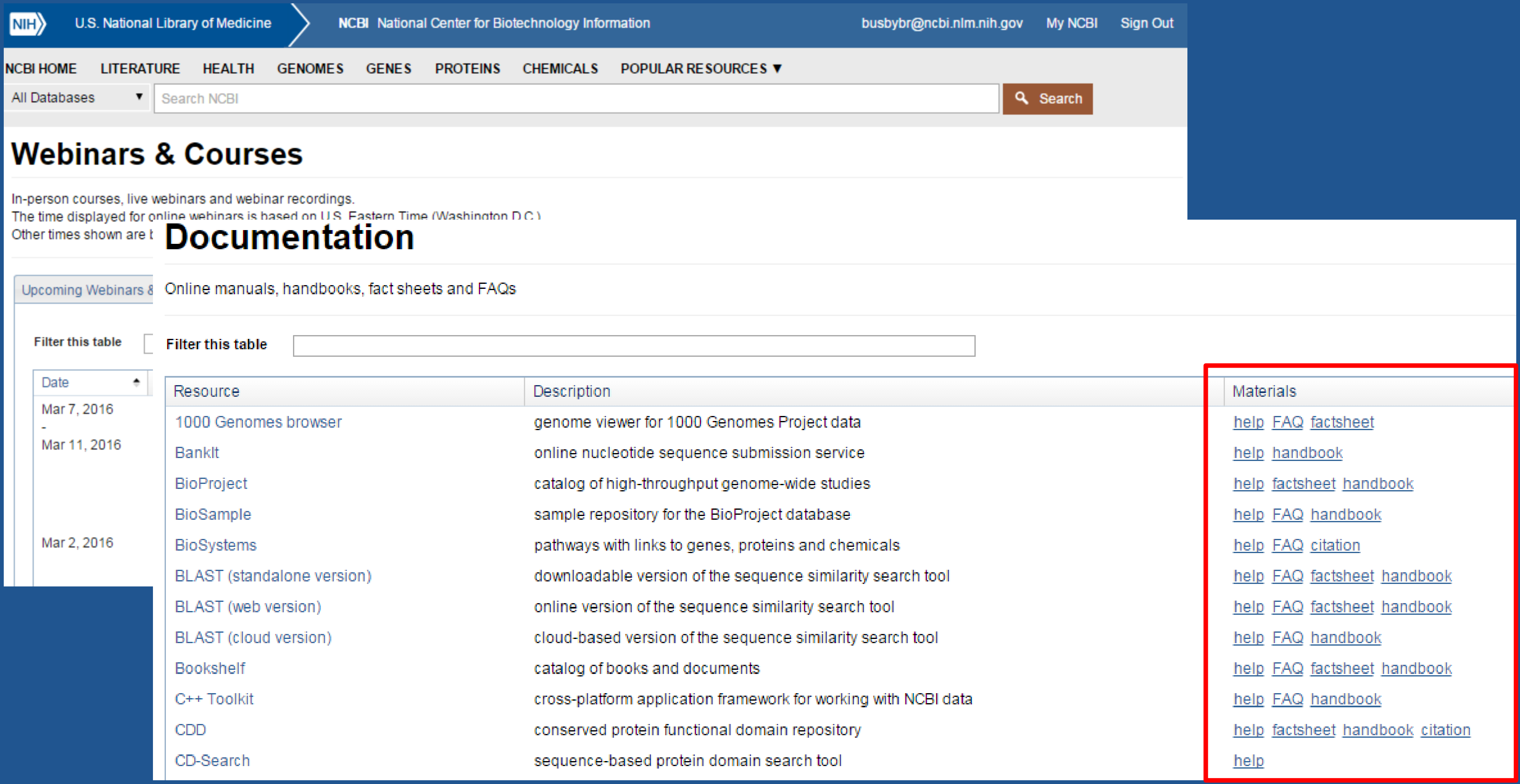
```
esearch -db bioproject -query "PRJNA356464" | elink -target sra | efetch -format docsum | \
xtract -pattern DocumentSummary -ACC @acc -block DocumentSummary -element "&ACC"
```

Get latitiude and longitude for SRA Datasets (e.g. outbreaks and metagenomes)

Description (optional):
Written by: BB, Mike D, Rob Edwards (4/12/2017)
Confirmed by:
Databases: SRA, BioSample

```
for i in $(cat sra_ids.txt); do ll=$(esearch -db sra -query $i | \
elink -target biosample | efetch -format docsum | \
xtract -pattern DocumentSummary -block Attribute -if Attribute@attribute_name -equals lat_lon -element Attribute); \
echo -e "$i\t$ll"; done
```

More Information…



Webinars & Courses

In-person courses, live webinars and webinar recordings.
The time displayed for online webinars is based on U.S. Eastern Time (Washington D.C.)
Other times shown are t

Date
Mar 7, 2016 - Mar 11, 2016
Mar 2, 2016

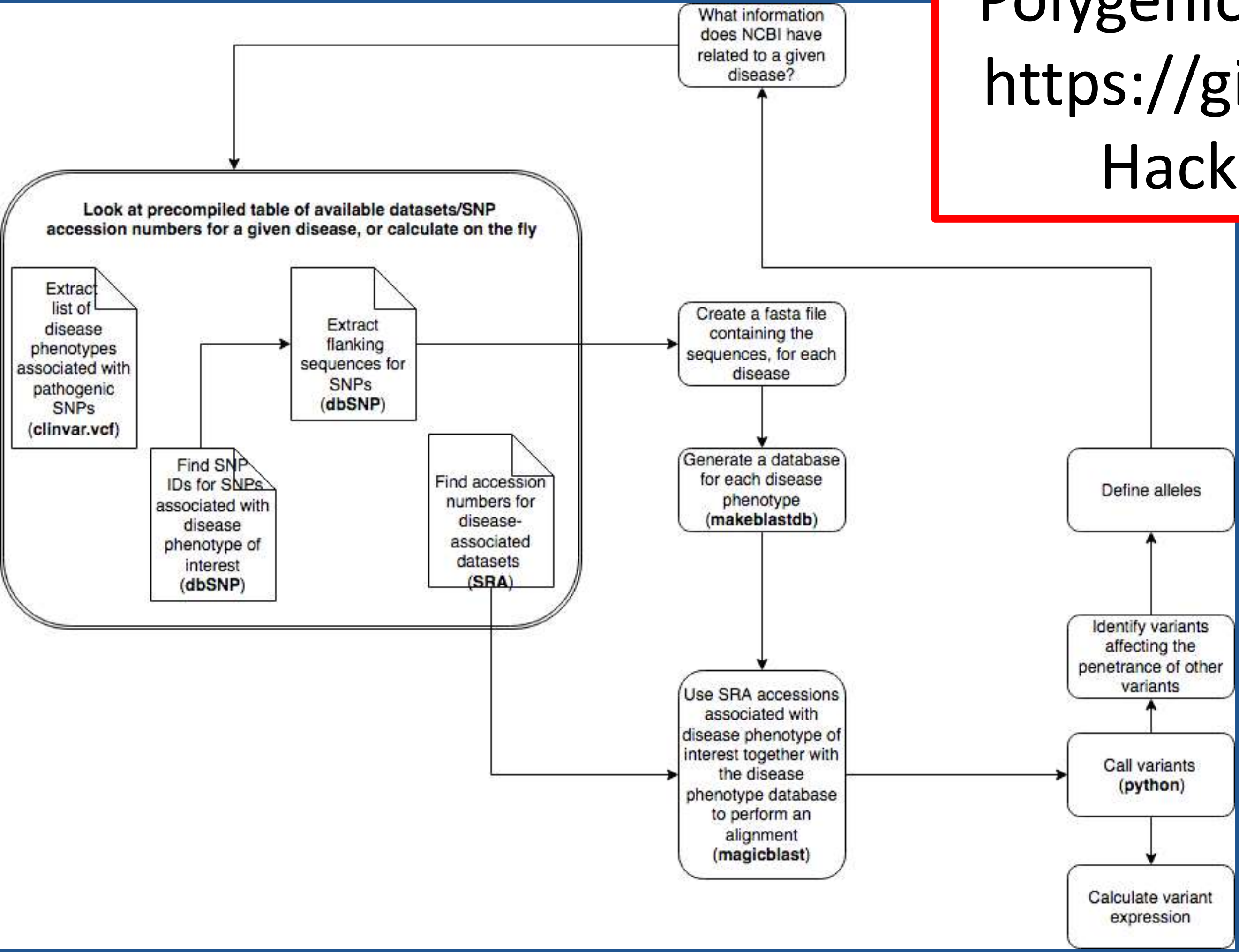
Documentation

Online manuals, handbooks, fact sheets and FAQs

Resource	Description	Materials
1000 Genomes browser	genome viewer for 1000 Genomes Project data	help FAQ factsheet
BankIt	online nucleotide sequence submission service	help handbook
BioProject	catalog of high-throughput genome-wide studies	help factsheet handbook
BioSample	sample repository for the BioProject database	help FAQ handbook
BioSystems	pathways with links to genes, proteins and chemicals	help FAQ citation
BLAST (standalone version)	downloadable version of the sequence similarity search tool	help FAQ factsheet handbook
BLAST (web version)	online version of the sequence similarity search tool	help FAQ factsheet handbook
BLAST (cloud version)	cloud-based version of the sequence similarity search tool	help FAQ handbook
Bookshelf	catalog of books and documents	help FAQ factsheet handbook
C++ Toolkit	cross-platform application framework for working with NCBI data	help FAQ handbook
CDD	conserved protein functional domain repository	help factsheet handbook citation
CD-Search	sequence-based protein domain search tool	help

SNP Calling on the Fly!

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



What information does NCBI have related to a given disease?

Look at precompiled table of available datasets/SNP accession numbers for a given disease, or calculate on the fly

Extract list of disease phenotypes associated with pathogenic SNPs (**clinvar.vcf**)

Extract flanking sequences for SNPs (**dbSNP**)

Find SNP IDs for SNPs associated with disease phenotype of interest (**dbSNP**)

Find accession numbers for disease-associated datasets (**SRA**)

Create a fasta file containing the sequences, for each disease

Generate a database for each disease phenotype (**makeblastdb**)

Use SRA accessions associated with disease phenotype of interest together with the disease phenotype database to perform an alignment (**magicblast**)

Call variants (**python**)

Calculate variant expression

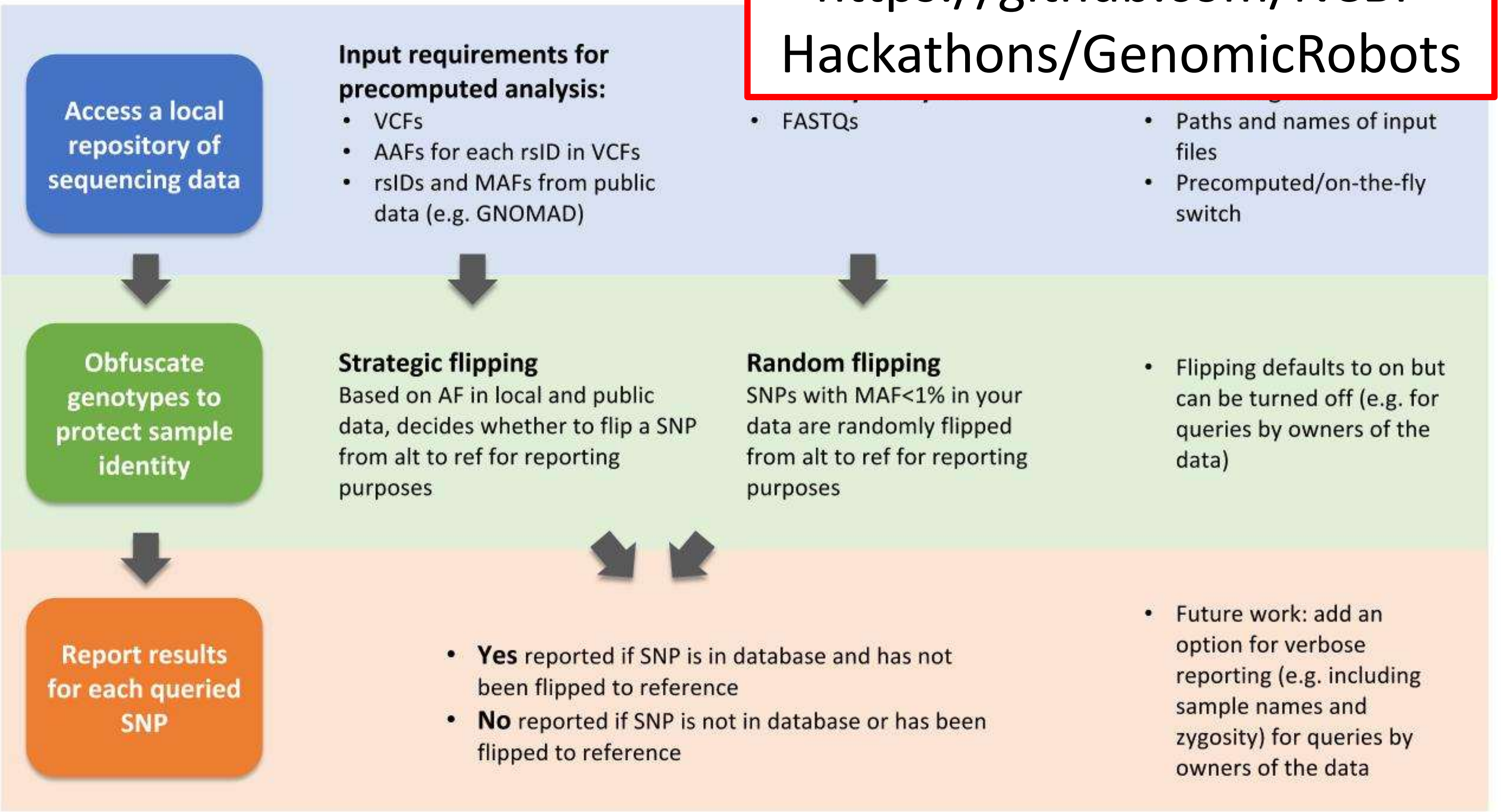
Identify variants affecting the penetrance of other variants

Define alleles

U.S. National Library of Medicine

NCBI

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



Access a local repository of sequencing data

Input requirements for precomputed analysis:

- VCFs
- AAFs for each rsID in VCFs
- rsIDs and MAFs from public data (e.g. GNOMAD)

- FASTQs

- Paths and names of input files
- Precomputed/on-the-fly switch

Obfuscate genotypes to protect sample identity

Strategic flipping
Based on AF in local and public data, decides whether to flip a SNP from alt to ref for reporting purposes

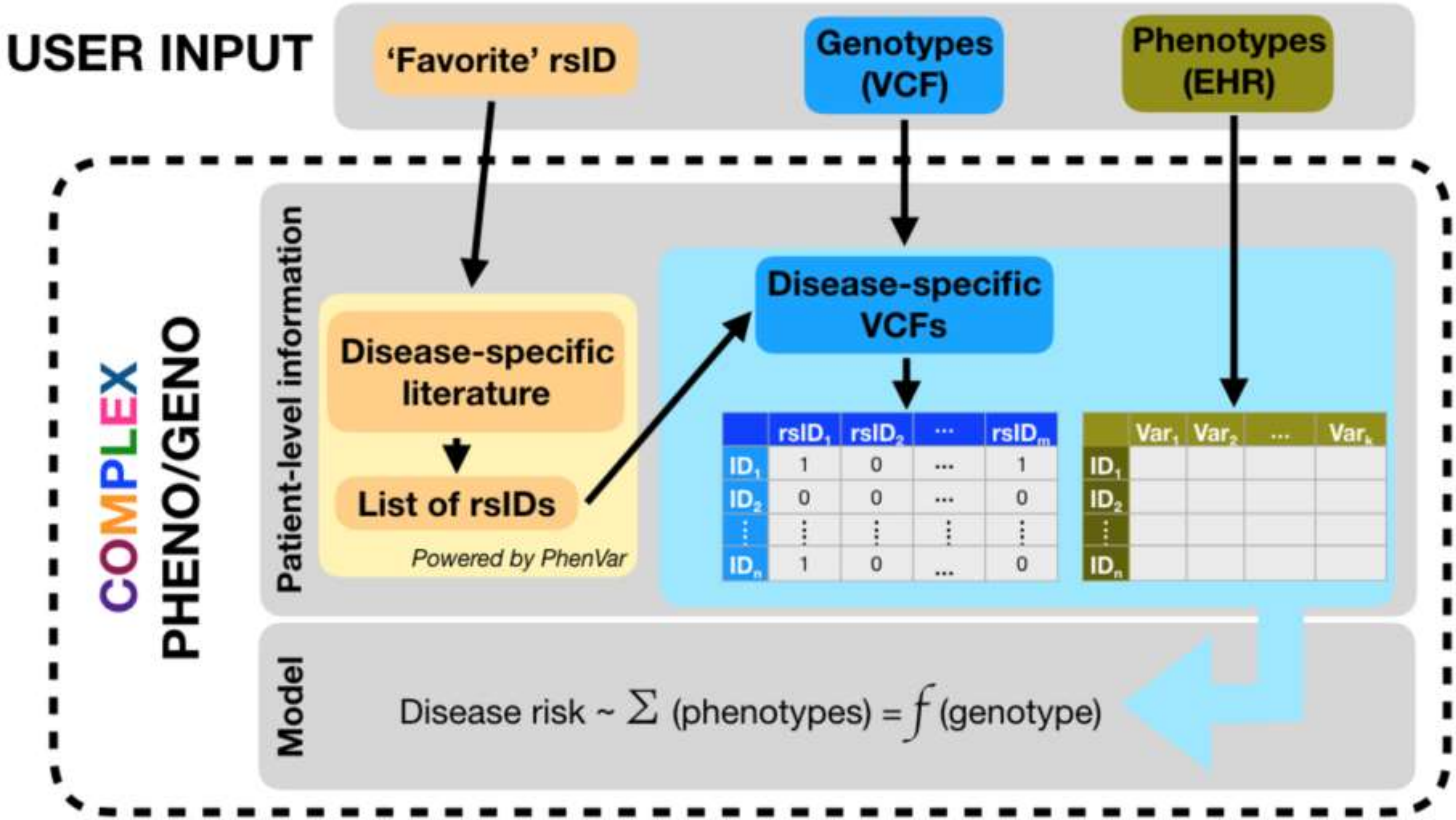
Random flipping
SNPs with MAF<1% in your data are randomly flipped from alt to ref for reporting purposes

- Flipping defaults to on but can be turned off (e.g. for queries by owners of the data)

Report results for each queried SNP

- **Yes** reported if SNP is in database and has not been flipped to reference
- **No** reported if SNP is not in database or has been flipped to reference

- Future work: add an option for verbose reporting (e.g. including sample names and zygosity) for queries by owners of the data



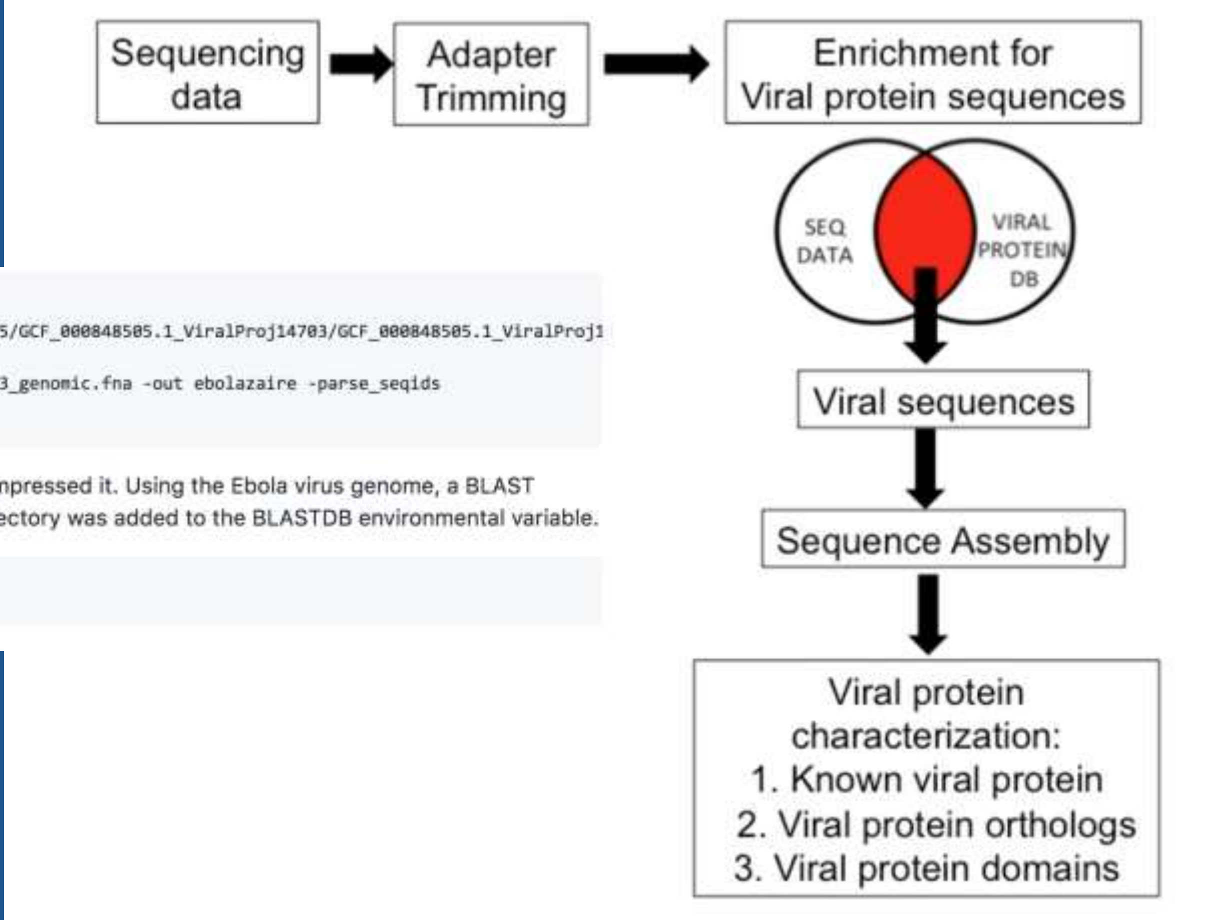
	$rsID_1$	$rsID_2$	...	$rsID_m$
ID_1	1	0	...	1
ID_2	0	0	...	0
⋮	⋮	⋮	⋮	⋮
ID_n	1	0	...	0

	Var_1	Var_2	...	Var_k
ID_1				
ID_2				
⋮				
ID_n				

Disease risk ~ Σ (phenotypes) = f (genotype)

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

Novel Virus Discovery!

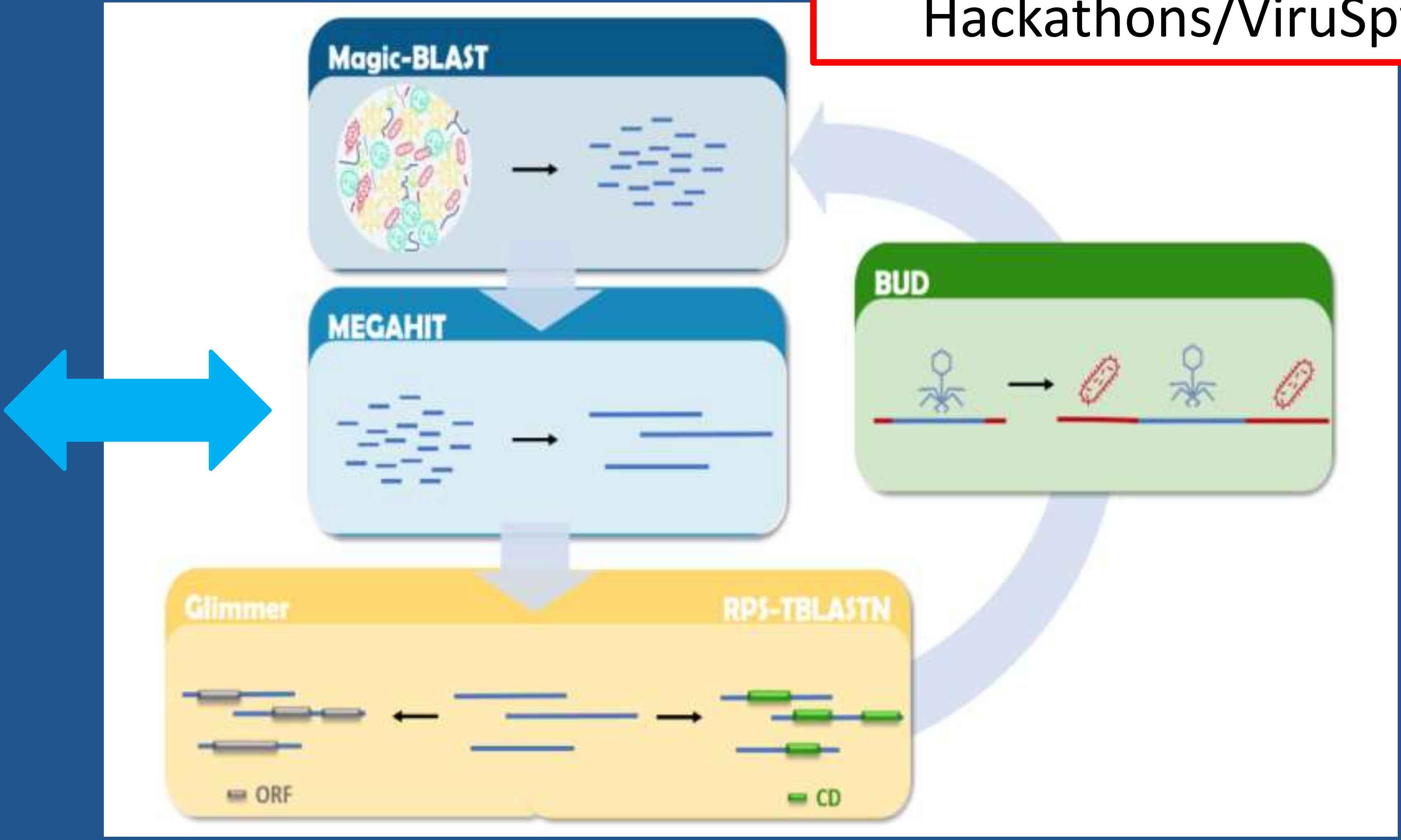


```
cd Virus_Detection_SRA/cwl/tools
wget ftp://ftp.ncbi.nlm.nih.gov/genomes/all/GCF/000/848/505/GCF_000848505.1_ViralProj14703/GCF_000848505.1_ViralProj1
gunzip GCF_000848505.1_ViralProj14703_genomic.fna.gz
makeblastdb -dbtype nucl -in GCF_000848505.1_ViralProj14703_genomic.fna -out ebolazaire -parse_seqids
export BLASTDB=$BLASTDB:`pwd`
```

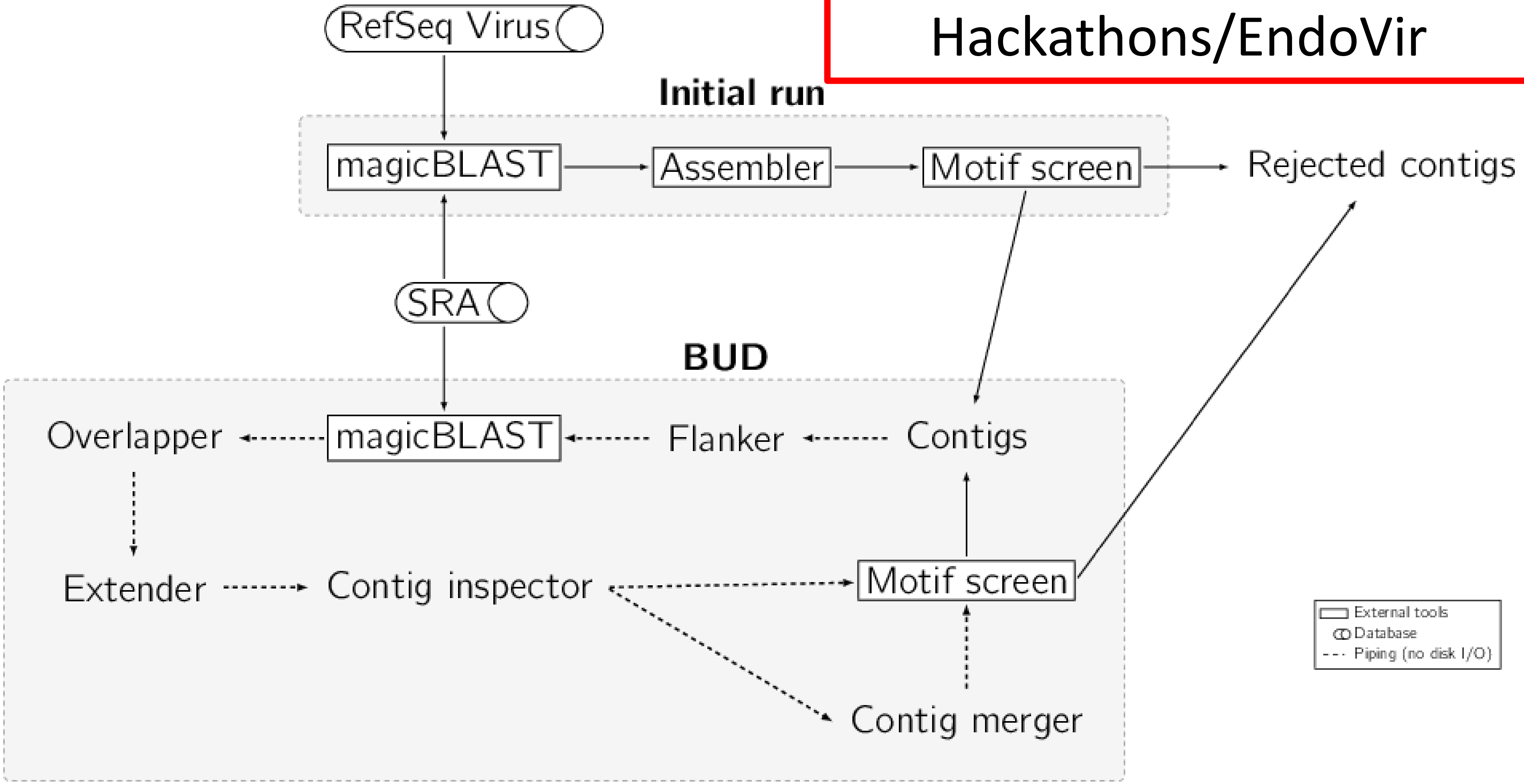
These steps downloaded the Ebola virus genome and uncompressed it. Using the Ebola virus genome, a BLAST database was created with `makeblastdb`. Then your local directory was added to the BLASTDB environmental variable.

```
sidearm.cwl sidearm.SRR1553459.ebola.yml
```

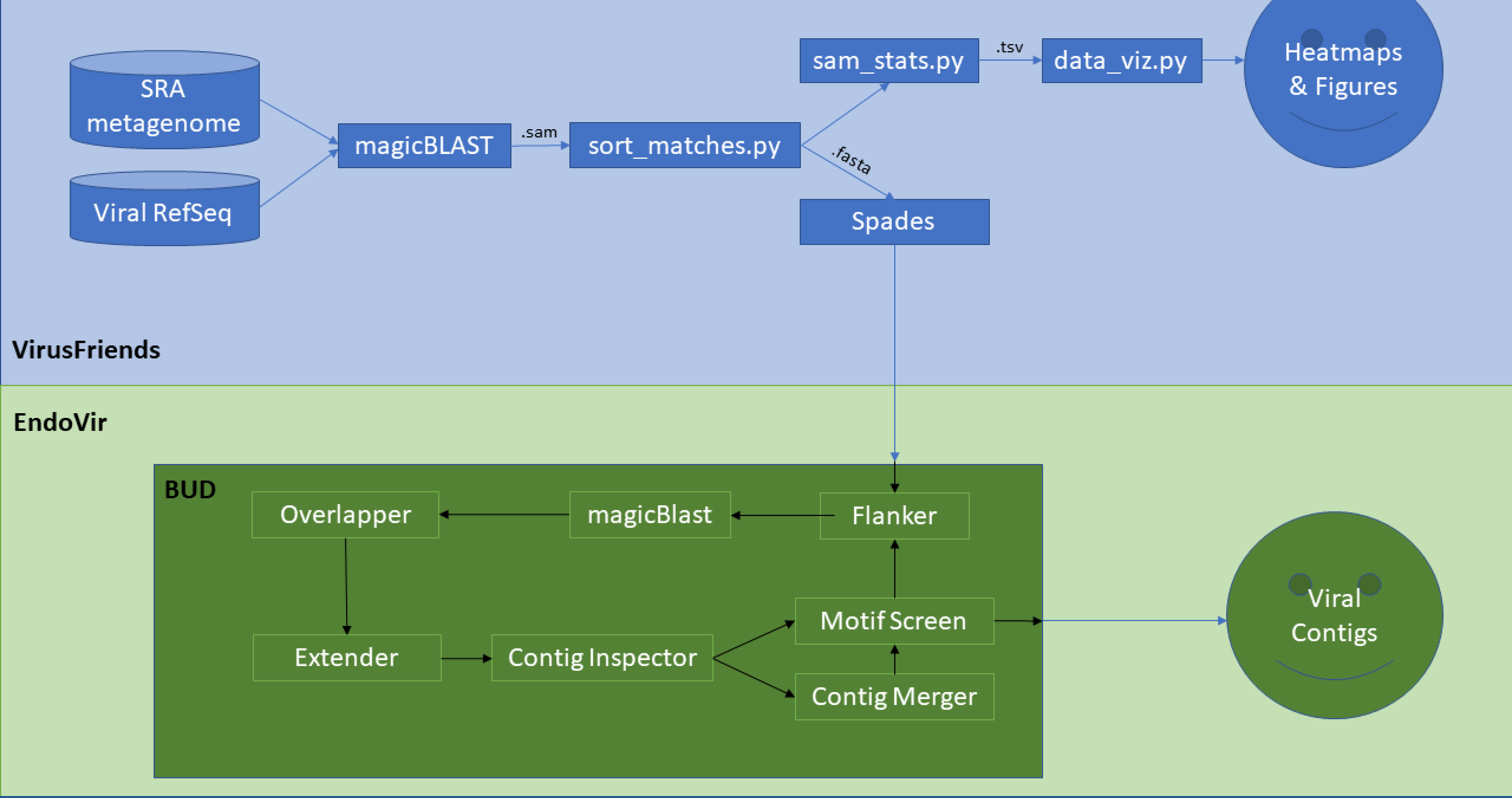
https://github.com/NCBI-Hackathons/ViruSpy

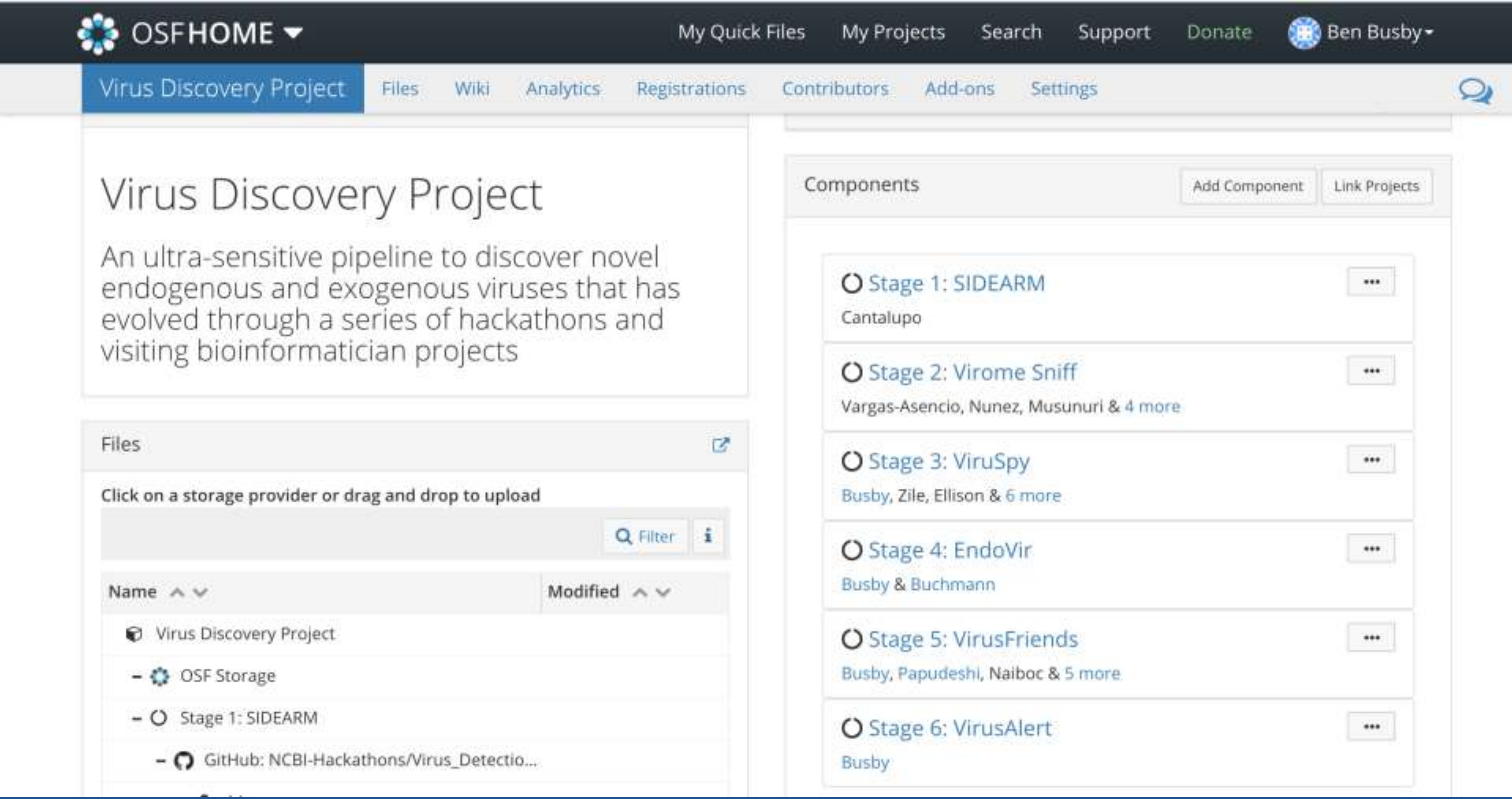


https://github.com/NCBI-Hackathons/EndoVir



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





Virus Discovery Project

An ultra-sensitive pipeline to discover novel endogenous and exogenous viruses that has evolved through a series of hackathons and visiting bioinformatician projects

Files

Click on a storage provider or drag and drop to upload

Name	Modified
Virus Discovery Project	
– OSF Storage	
– Stage 1: SIDEARM	
– GitHub: NCBI-Hackathons/Virus_Detectio...	

Components

Stage 1: SIDEARM
Cantalupo

Stage 2: Virome Sniff
Vargas-Asencio, Nunez, Musunuri & 4 more

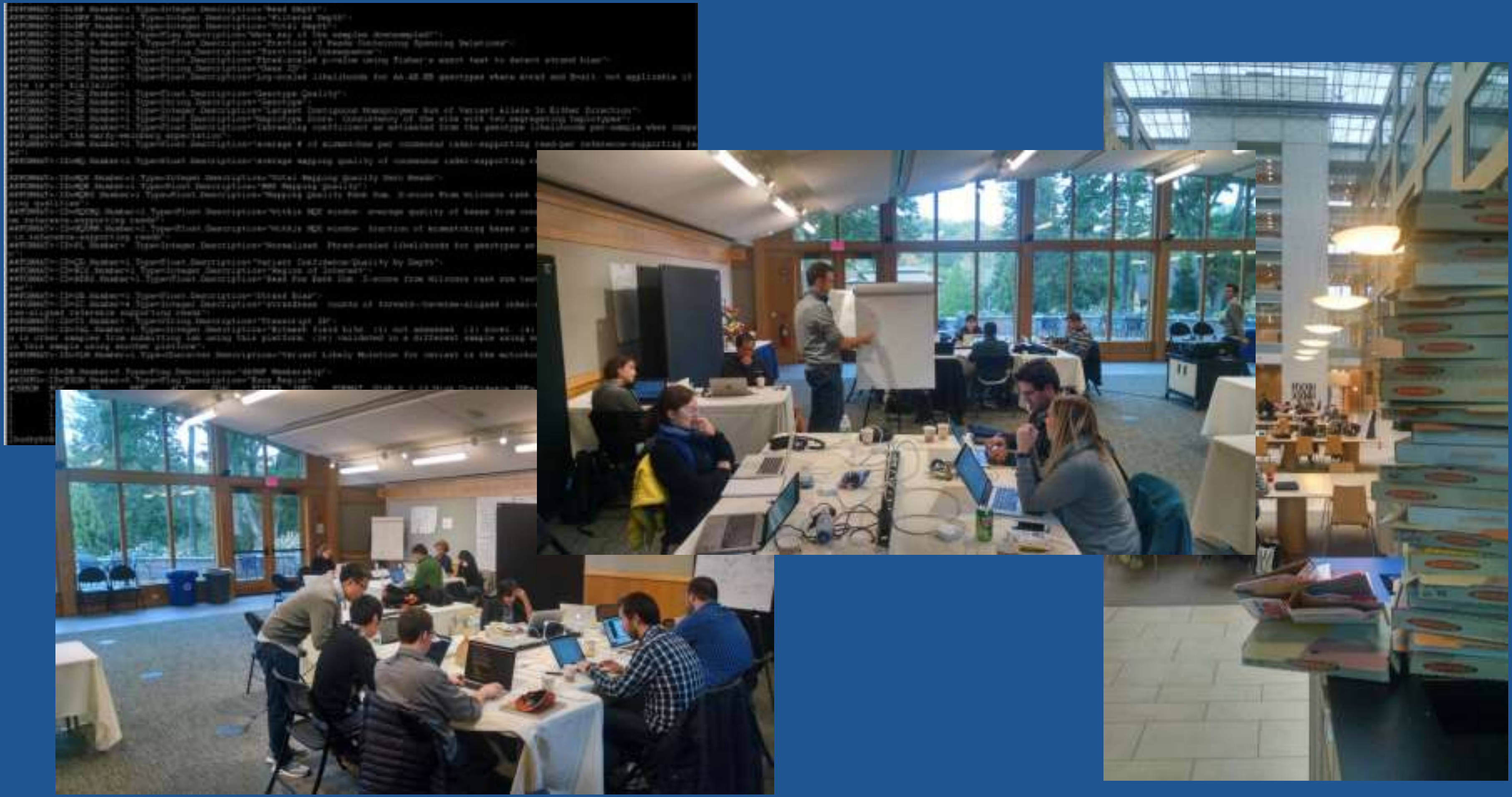
Stage 3: ViruSpy
Busby, Zile, Ellison & 6 more

Stage 4: EndoVir
Busby & Buchmann

Stage 5: VirusFriends
Busby, Papudeshi, Naiboc & 5 more

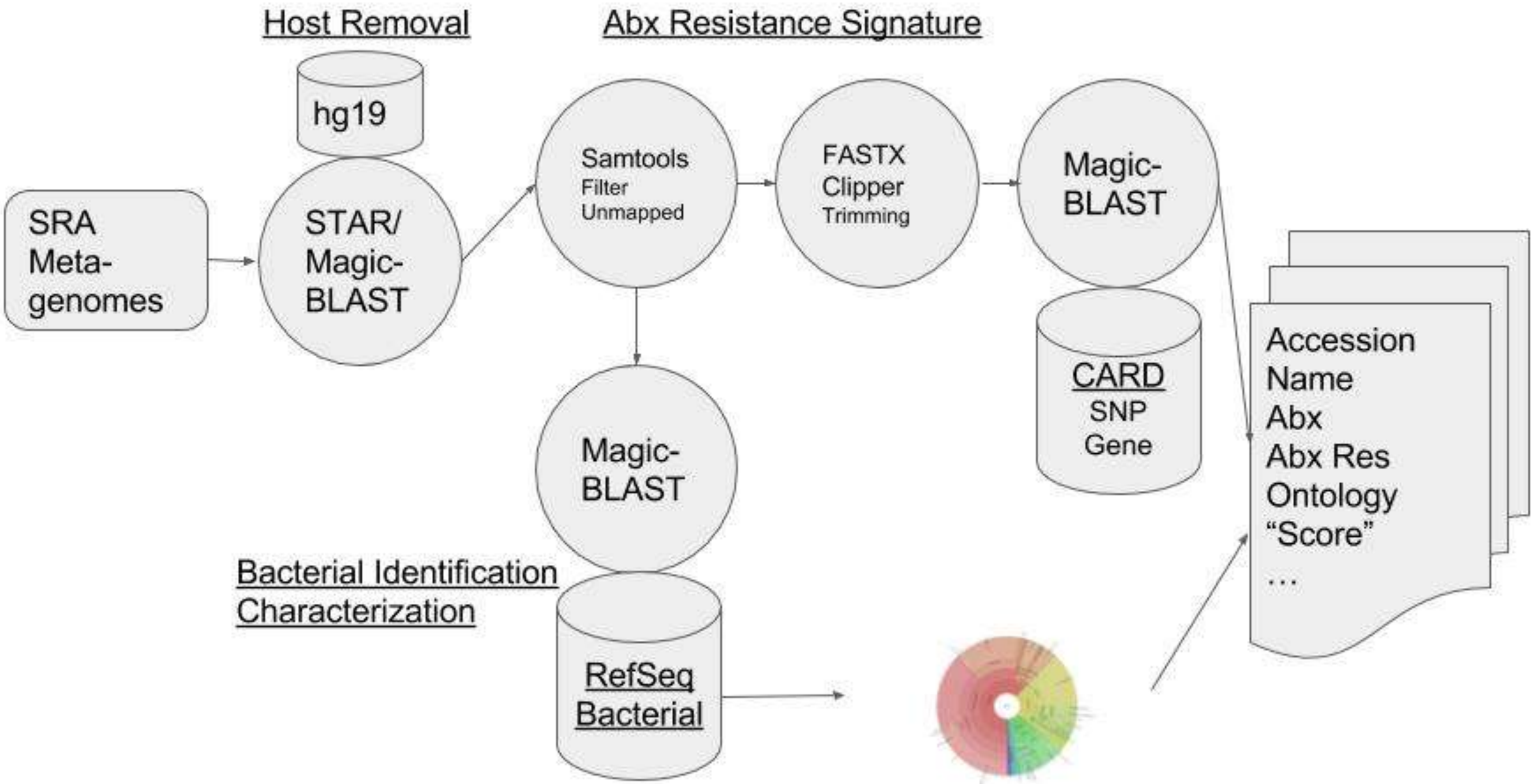
Stage 6: VirusAlert
Busby

NCBI Hackathons!



```
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=DPF,Number=1,Type=Integer,Description="Filtered Depth">
##FORMAT=<ID=DPT,Number=1,Type=Integer,Description="Total Depth">
[illegible]
##FORMAT=<ID=GQ,Number=1,Type=Float,Description="Genotype Quality">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
[illegible]
```

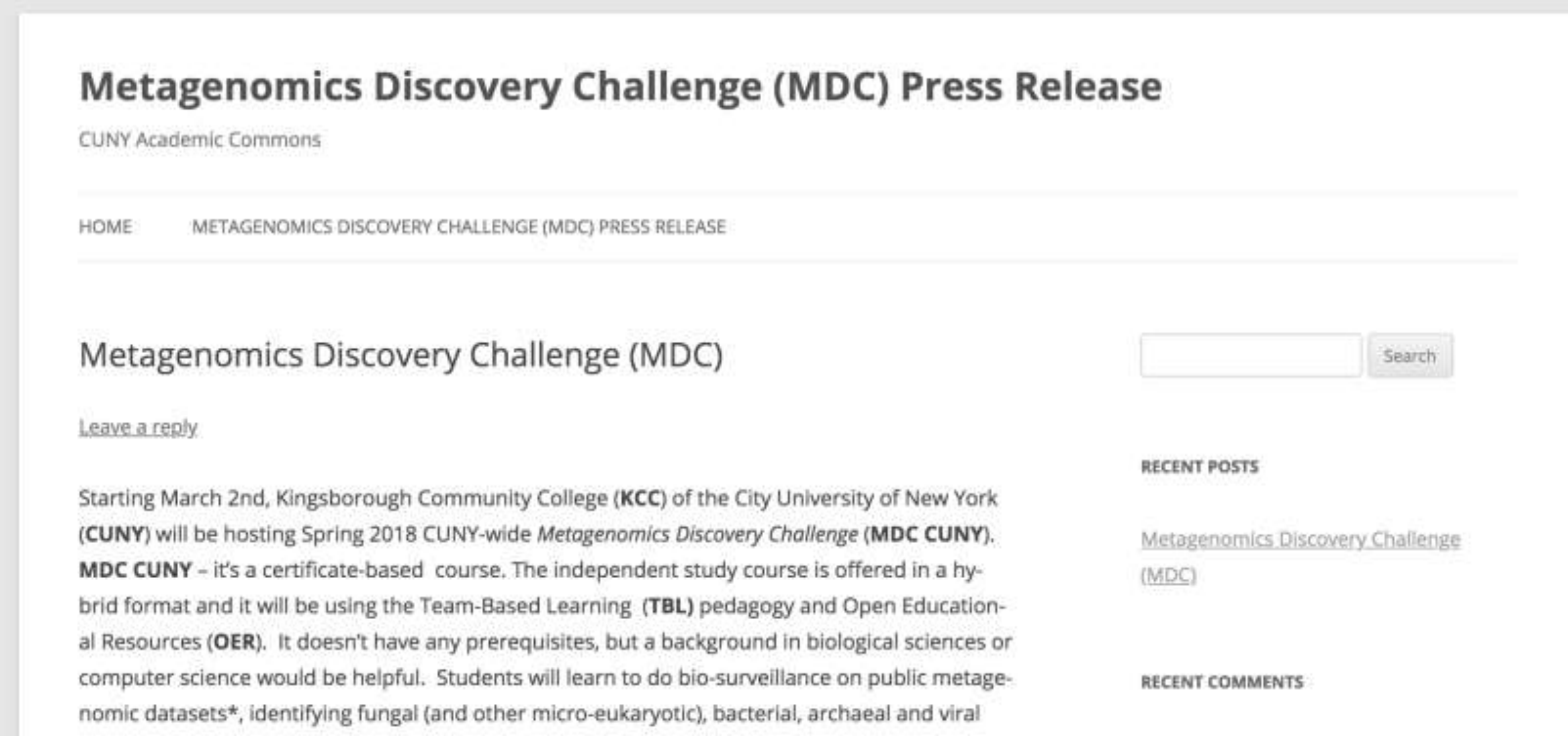
Antibiotic Resistance Detection!



Antibiotic Resistance Detection!

```
tblastn_vdb -query mdr_sequence.aa -db "SRR1427233 SRR515906" -outfmt 6 -max_target_seqs 2500 -out
sra_tblastn.tab
```

https://www.ncbi.nlm.nih.gov/core/assets/sra/files/Factsheet_SRA.pdf



Metagenomics Discovery Challenge (MDC) Press Release

CUNY Academic Commons

HOME
METAGENOMICS DISCOVERY CHALLENGE (MDC) PRESS RELEASE
Search
RECENT POSTS
Metagenomics Discovery Challenge (MDC)
RECENT COMMENTS

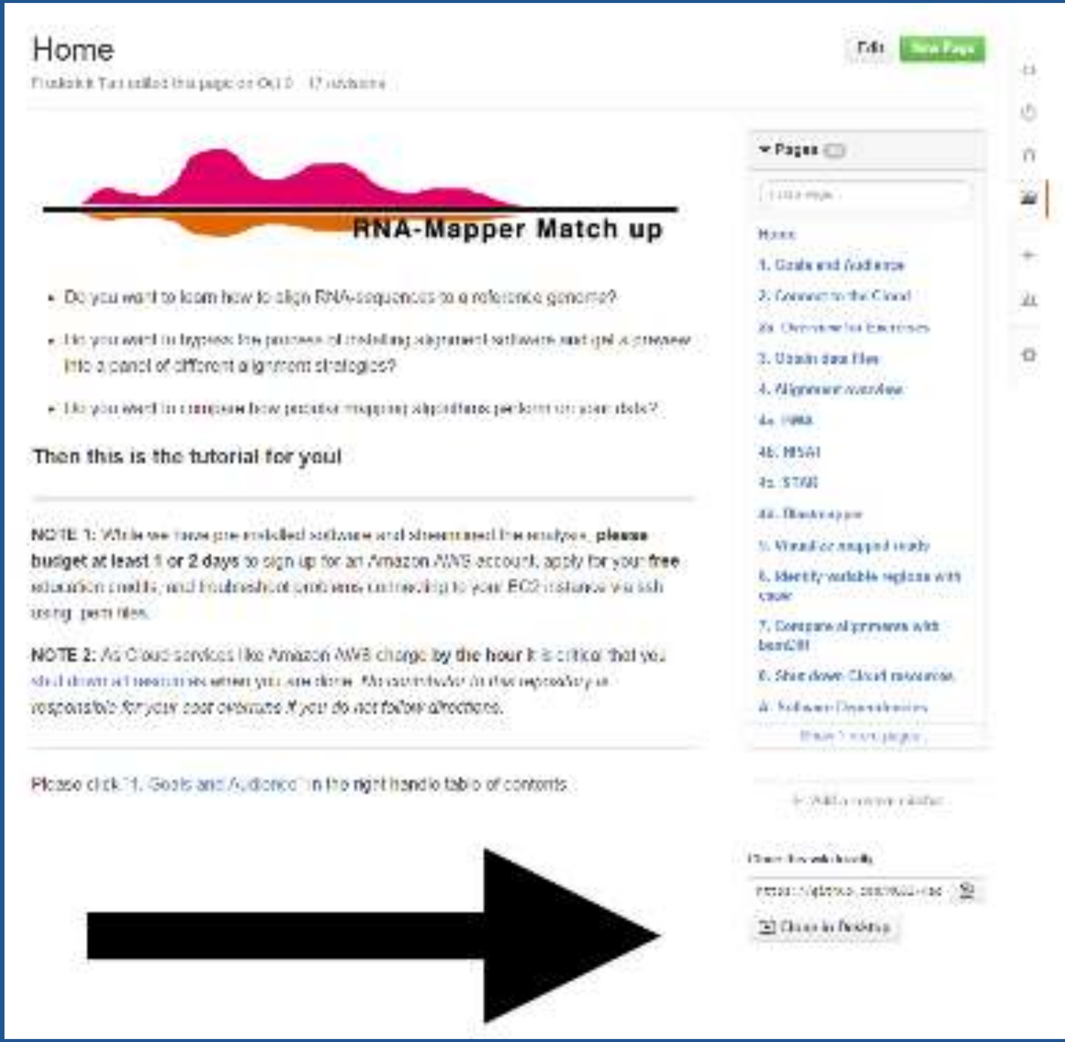
Metagenomics Discovery Challenge (MDC)

Leave a reply

Starting March 2nd, Kingsborough Community College (**KCC**) of the City University of New York (**CUNY**) will be hosting Spring 2018 CUNY-wide *Metagenomics Discovery Challenge* (**MDC CUNY**). **MDC CUNY** – it's a certificate-based course. The independent study course is offered in a hybrid format and it will be using the Team-Based Learning (**TBL**) pedagogy and Open Educational Resources (**OER**). It doesn't have any prerequisites, but a background in biological sciences or computer science would be helpful. Students will learn to do bio-surveillance on public metagenomic datasets*, identifying fungal (and other micro-eukaryotic), bacterial, archaeal and viral

An Educational Resource for RNAseq

Available to anyone!



Home

Frederick Tan edited this page on Oct 5 · 17 revisions

Edit | New Page

RNA-Mapper Match up

- Do you want to learn how to align RNA-sequences to a reference genome?
- Do you want to bypass the process of installing alignment software and get a preview into a panel of different alignment strategies?
- Do you want to compare how popular mapping algorithms perform on your data?

Then this is the tutorial for you!

NOTE 1: While we have pre-installed software and streamlined the analysis, **please budget at least 1 or 2 days** to sign up for an Amazon AWS account, apply for your **free** education credits, and troubleshoot problems connecting to your EC2 instance via ssh using .pem files.

NOTE 2: As Cloud services like Amazon AWS charge **by the hour** it is critical that you shut down all resources when you are done. *No contributor to this repository is responsible for your cost overruns if you do not follow directions.*

Please click "1. Goals and Audience" in the right hand table of contents.

▾ Pages

Find a Page…

- Home
- 1. Goals and Audience
- 2. Connect to the Cloud
- 2a. Overview for Exercises
- 3. Obtain data files
- 4. Alignment overview
- 4a. BWA
- 4b. HISAT
- 4c. STAR
- 4d. [illegible]
- 5. Visualize mapped reads
- 6. Identify variable regions with [illegible]
- 7. Compare alignments with [illegible]
- 8. Shut down Cloud resources
- A. Software Dependencies

Show 2 more pages…

+ Add a custom sidebar

Clone this wiki locally

https://github.com/NCBI-[illegible]

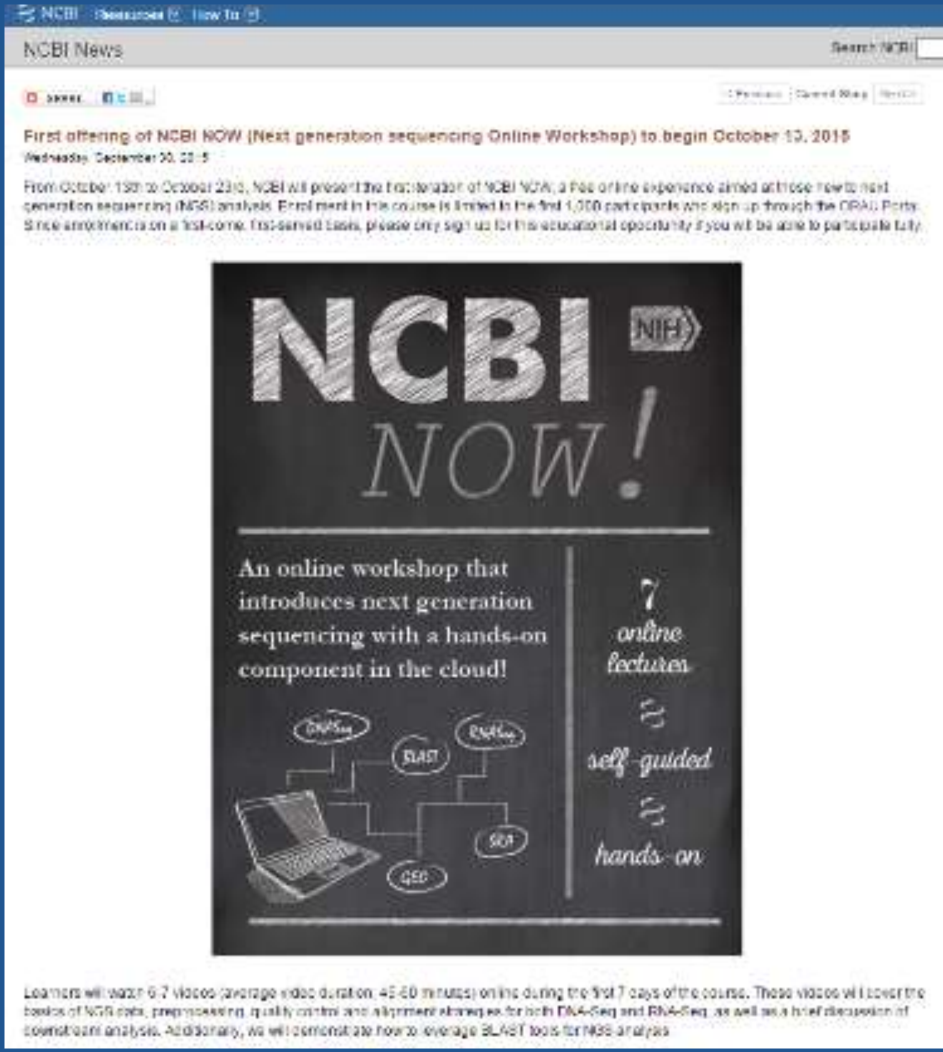
Clone in Desktop

Part of an Online Workshop

First 5 lectures

now available

on YouTube



First offering of NCBI NOW (Next generation sequencing Online Workshop) to begin October 13, 2015

Wednesday, September 30, 2015

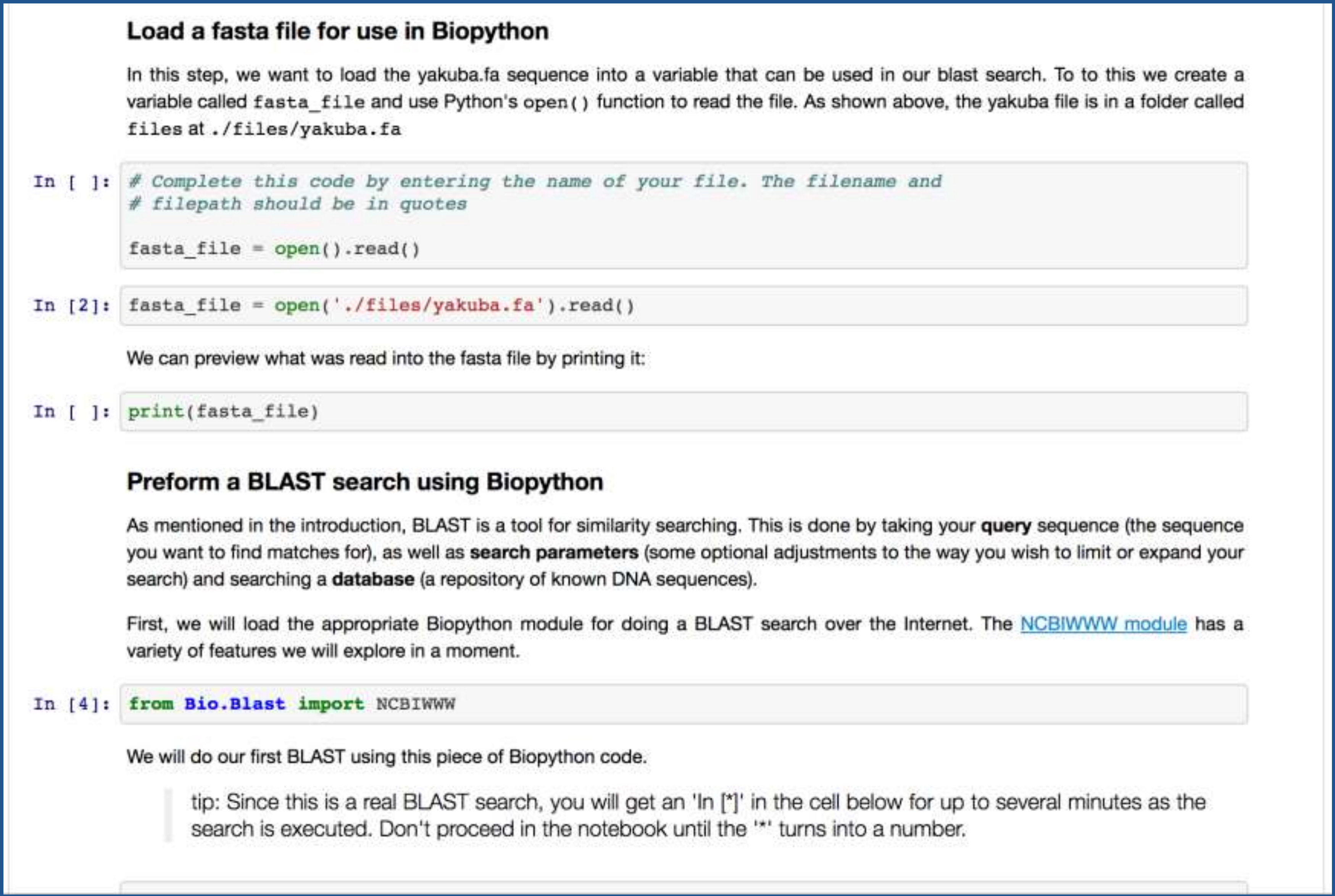
From October 13th to October 23rd, NCBI will present the first iteration of NCBI NOW, a free online experience aimed at those new to next generation sequencing (NGS) analysis. Enrollment in this course is limited to the first 1,000 participants who sign up through the ORAU Portal. Since enrollment is on a first-come, first-served basis, please only sign up for this educational opportunity if you will be able to participate fully.

NCBI NOW!

An online workshop that introduces next generation sequencing with a hands-on component in the cloud!

7 online lectures & self-guided & hands-on

Learners will watch 6-7 videos (average video duration: 45-60 minutes) online during the first 7 days of the course. These videos will cover the basics of NGS data, preprocessing, quality control and alignment strategies for both DNA-Seq and RNA-Seq, as well as a brief discussion of downstream analysis. Additionally, we will demonstrate how to leverage BLAST tools for NGS analysis.



Load a fasta file for use in Biopython

In this step, we want to load the yakuba.fa sequence into a variable that can be used in our blast search. To to this we create a variable called `fasta_file` and use Python's `open()` function to read the file. As shown above, the yakuba file is in a folder called `files` at `./files/yakuba.fa`

In [ ]:

```
# Complete this code by entering the name of your file. The filename and
# filepath should be in quotes

fasta_file = open().read()
```

In [2]:

```
fasta_file = open('./files/yakuba.fa').read()
```

We can preview what was read into the fasta file by printing it:

In [ ]:

```
print(fasta_file)
```

Preform a BLAST search using Biopython

As mentioned in the introduction, BLAST is a tool for similarity searching. This is done by taking your **query** sequence (the sequence you want to find matches for), as well as **search parameters** (some optional adjustments to the way you wish to limit or expand your search) and searching a **database** (a repository of known DNA sequences).

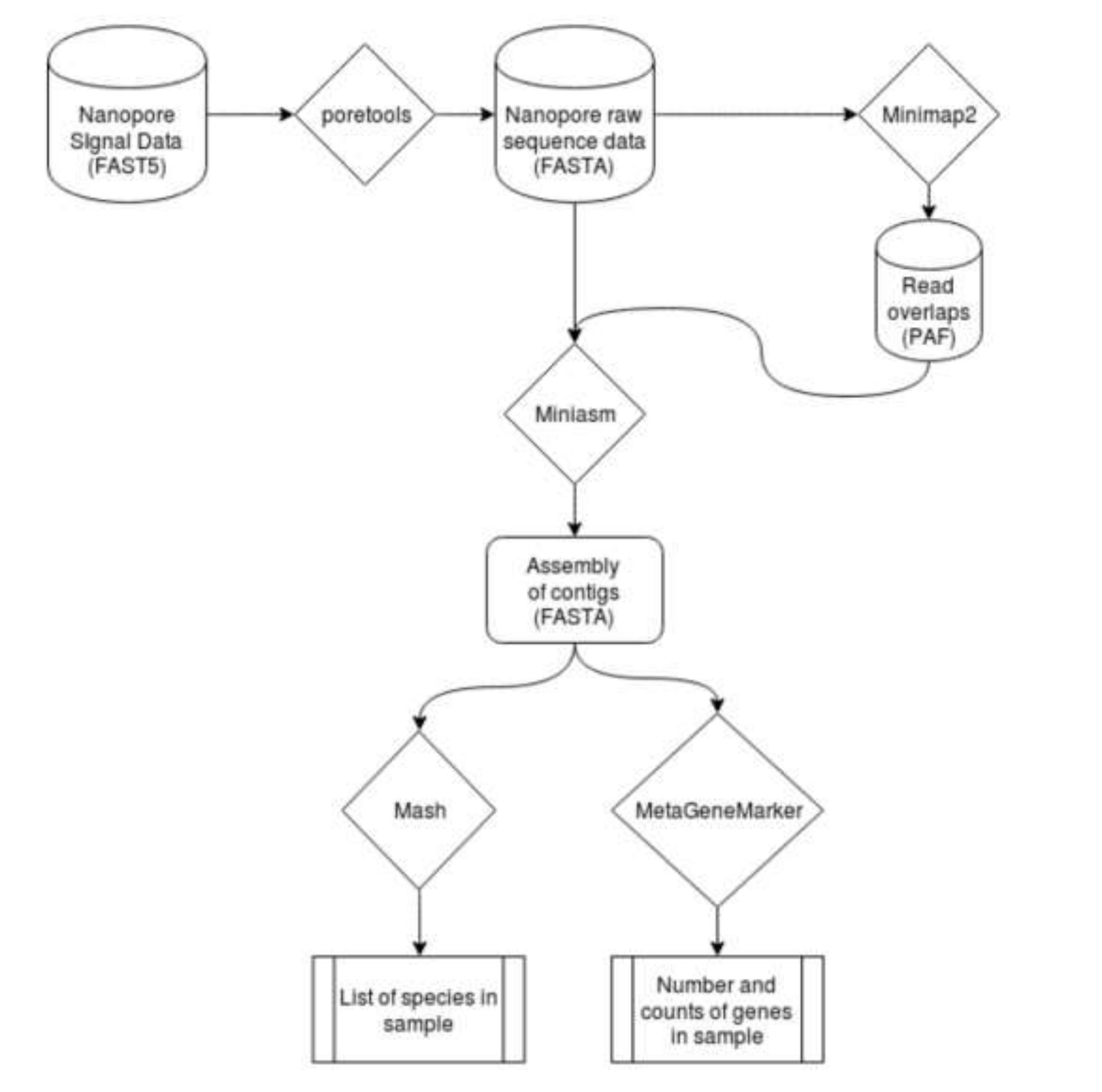
First, we will load the appropriate Biopython module for doing a BLAST search over the Internet. The NCBIWWW module has a variety of features we will explore in a moment.

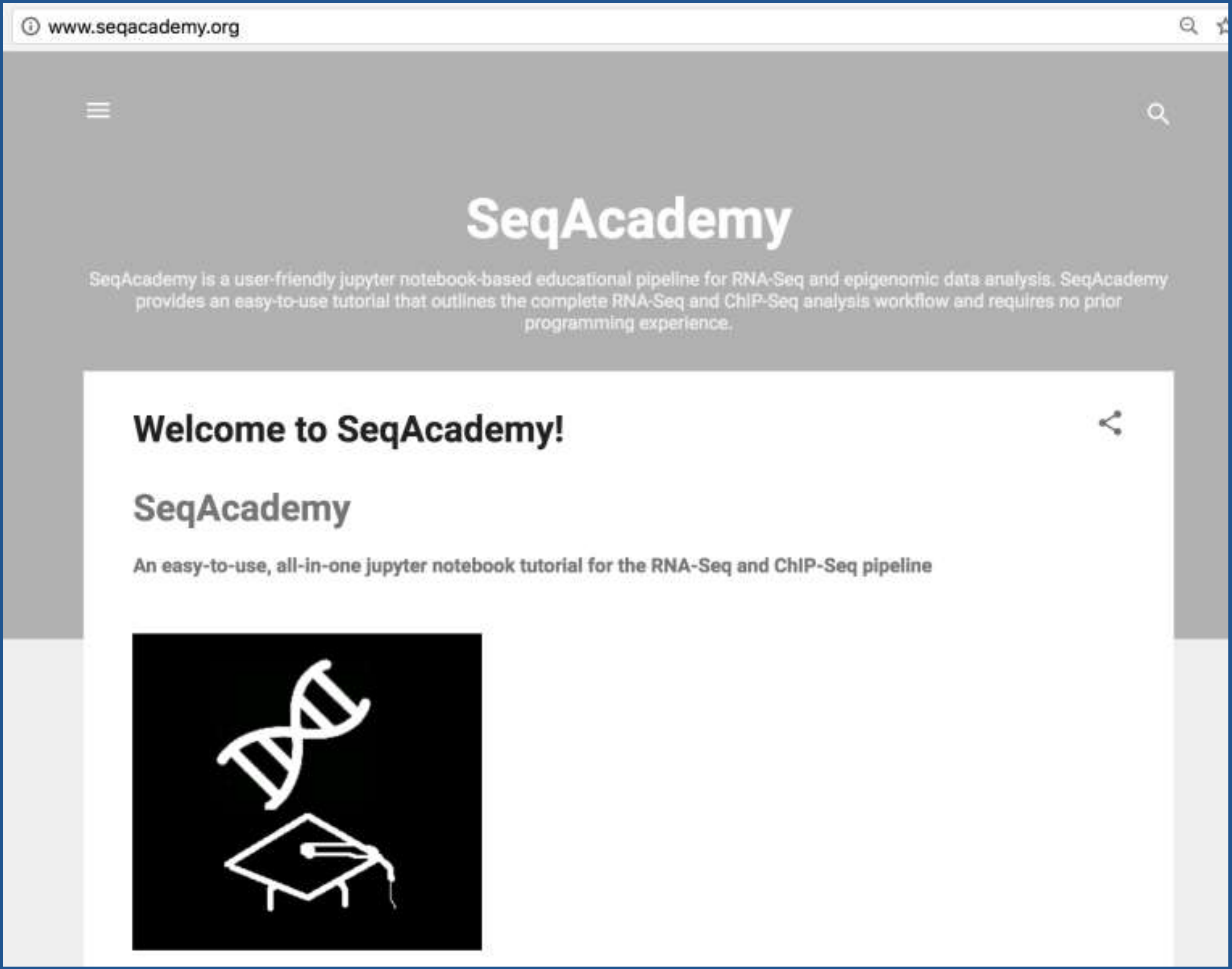
In [4]:

```
from Bio.Blast import NCBIWWW
```

We will do our first BLAST using this piece of Biopython code.

> tip: Since this is a real BLAST search, you will get an 'In [*]' in the cell below for up to several minutes as the search is executed. Don't proceed in the notebook until the '*' turns into a number.


Nanopore Signal Data (FAST5)
poretools
Nanopore raw sequence data (FASTA)
Minimap2
Read overlaps (PAF)
Miniasm
Assembly of contigs (FASTA)
Mash
MetaGeneMarker
List of species in sample
Number and counts of genes in sample



www.seqacademy.org

SeqAcademy

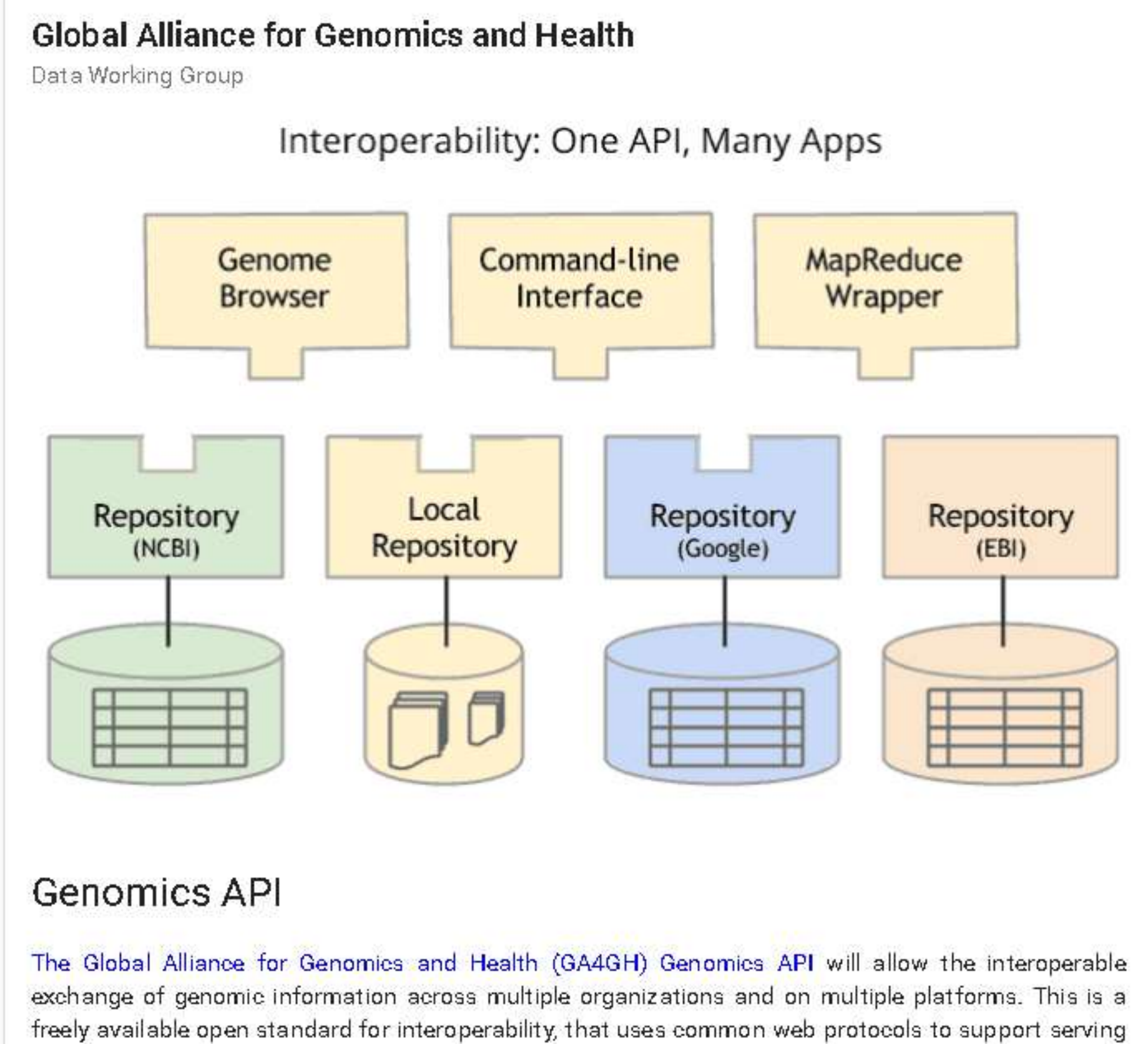
SeqAcademy is a user-friendly jupyter notebook-based educational pipeline for RNA-Seq and epigenomic data analysis. SeqAcademy provides an easy-to-use tutorial that outlines the complete RNA-Seq and ChIP-Seq analysis workflow and requires no prior programming experience.

Welcome to SeqAcademy!

SeqAcademy

An easy-to-use, all-in-one jupyter notebook tutorial for the RNA-Seq and ChIP-Seq pipeline

Collaboration!



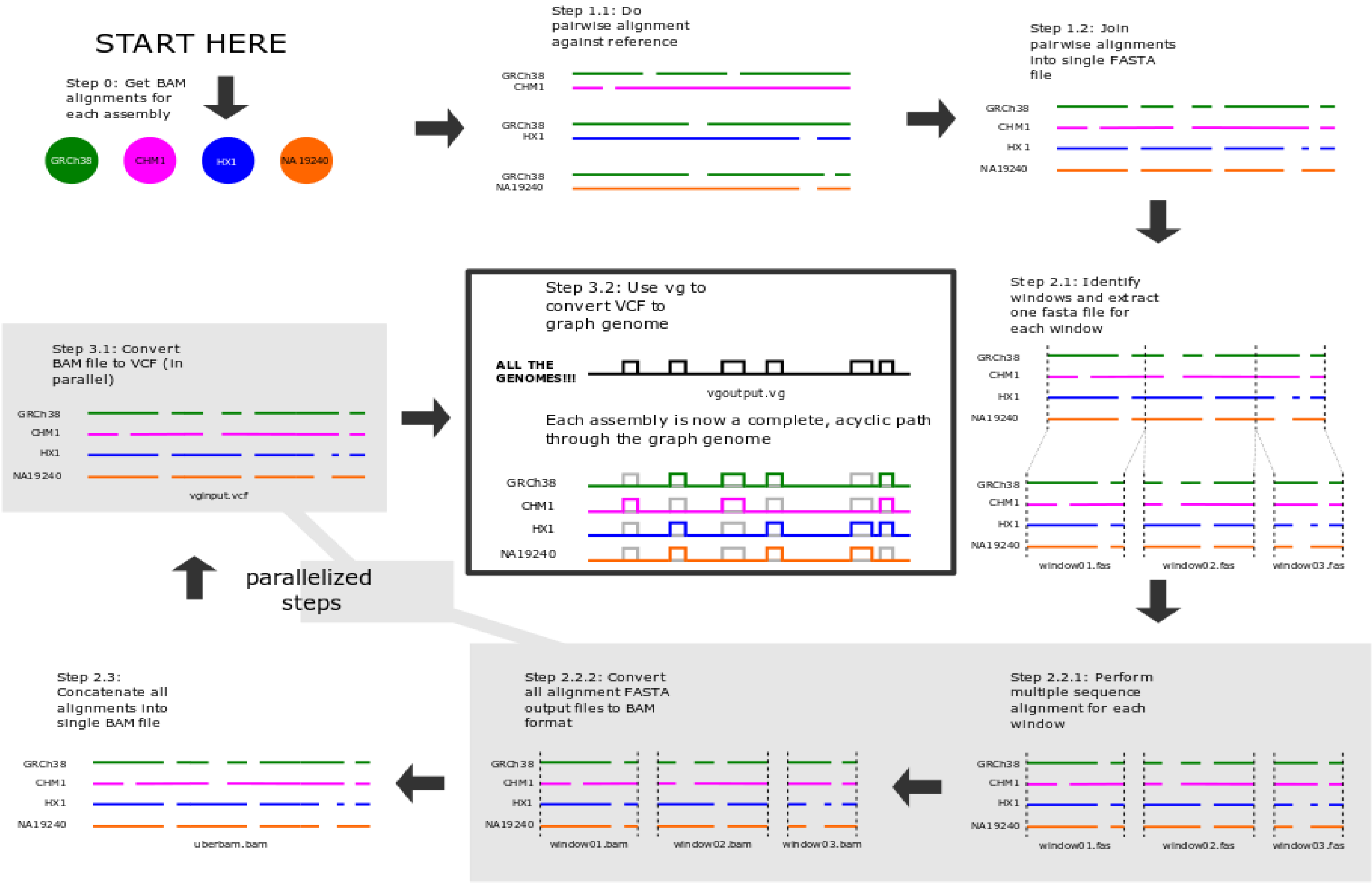
Global Alliance for Genomics and Health

Data Working Group

Genomics API

The Global Alliance for Genomics and Health (GA4GH) Genomics API will allow the interoperable exchange of genomic information across multiple organizations and on multiple platforms. This is a freely available open standard for interoperability, that uses common web protocols to support serving

Graph Genomes!!!



START HERE

Step 0: Get BAM alignments for each assembly

Step 1.1: Do pairwise alignment against reference

Step 1.2: Join pairwise alignments into single FASTA file

Step 2.1: Identify windows and extract one fasta file for each window

Step 2.2.1: Perform multiple sequence alignment for each window

Step 2.2.2: Convert all alignment FASTA output files to BAM format

Step 2.3: Concatenate all alignments into single BAM file

Step 3.1: Convert BAM file to VCF (in parallel)

Step 3.2: Use vg to convert VCF to graph genome

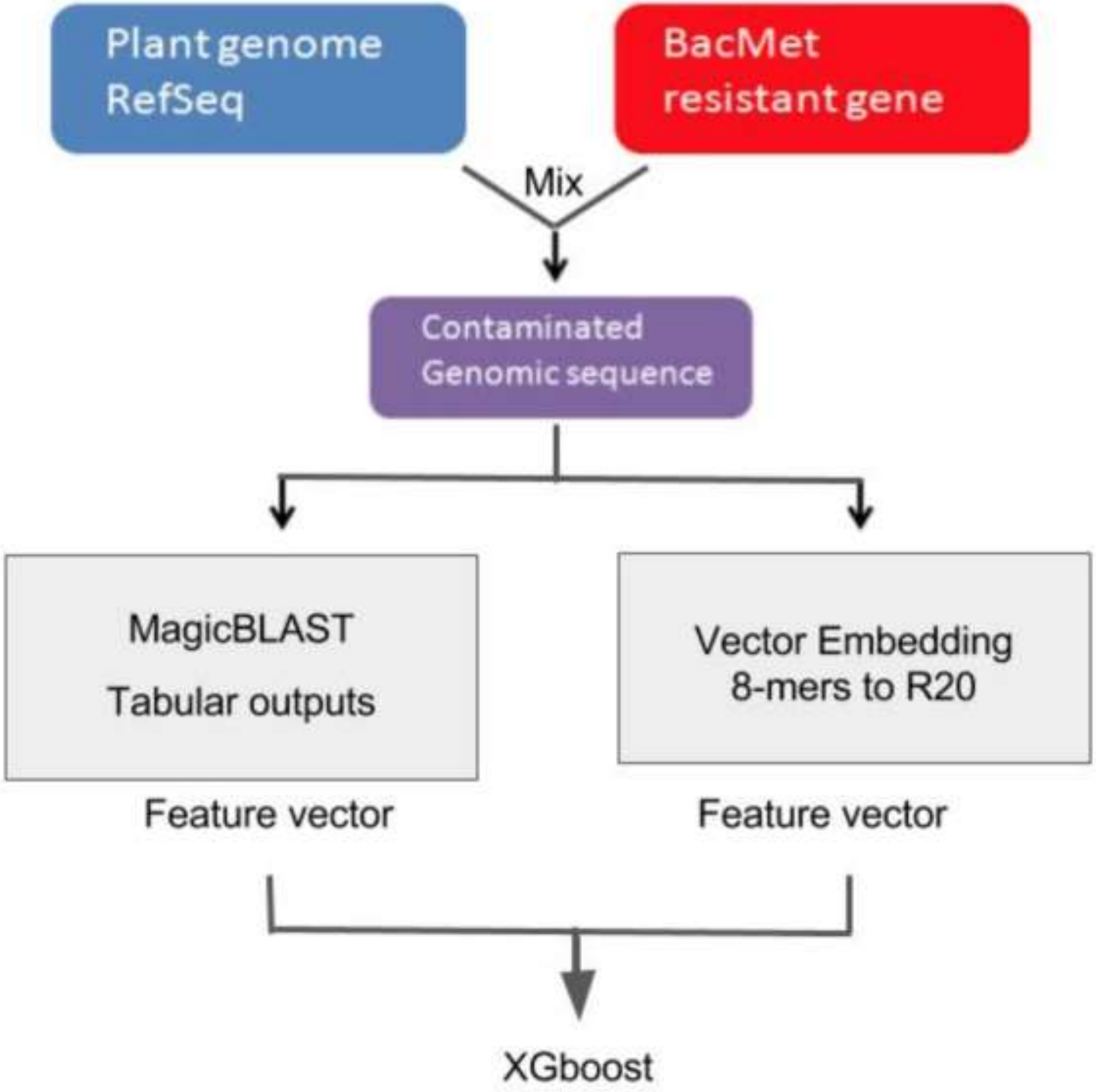
Each assembly is now a complete, acyclic path through the graph genome

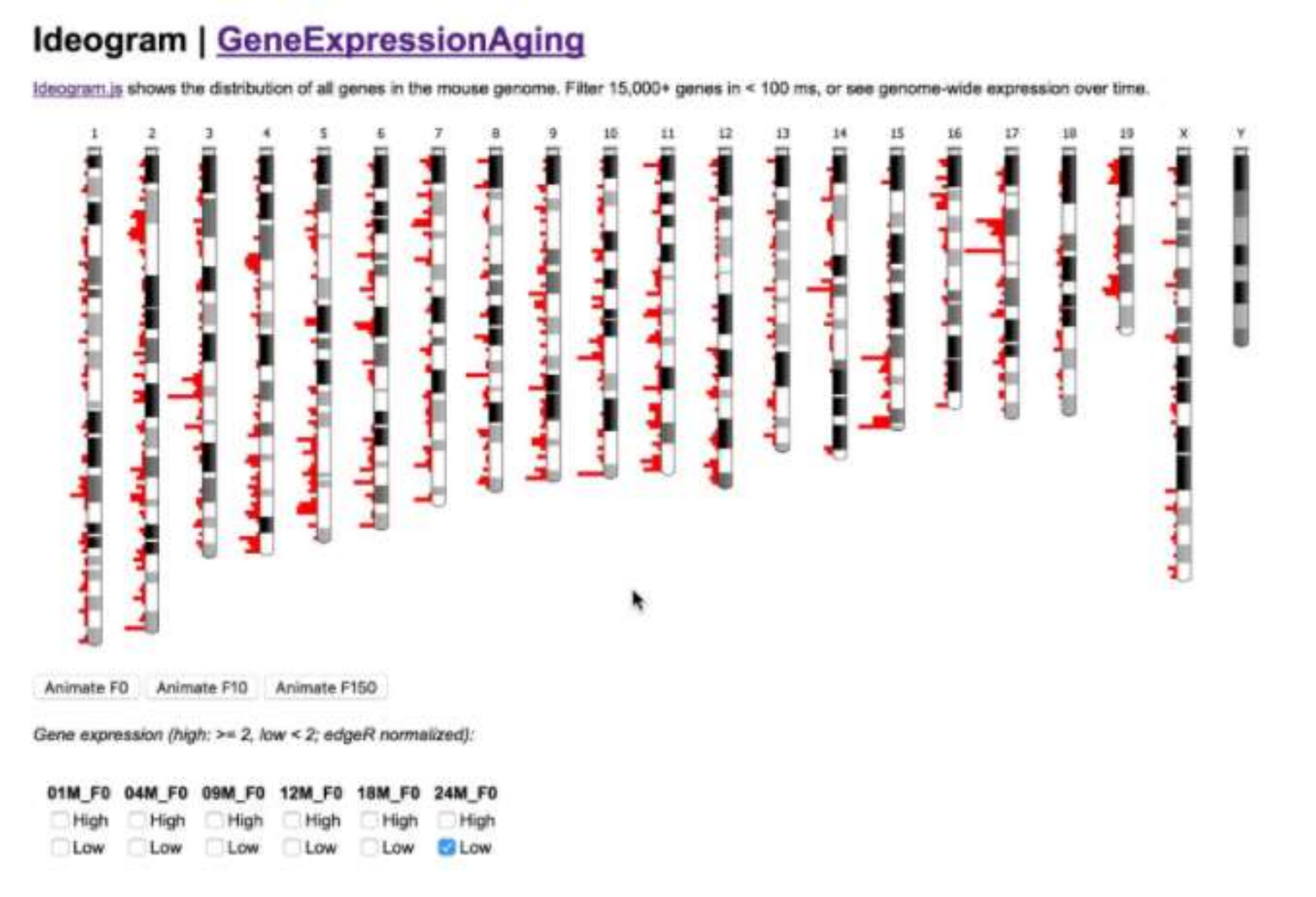
parallelized steps

Data preparation

Featurization

Classifier





Ideogram | GeneExpressionAging

Ideogram.js shows the distribution of all genes in the mouse genome. Filter 15,000+ genes in < 100 ms, or see genome-wide expression over time.

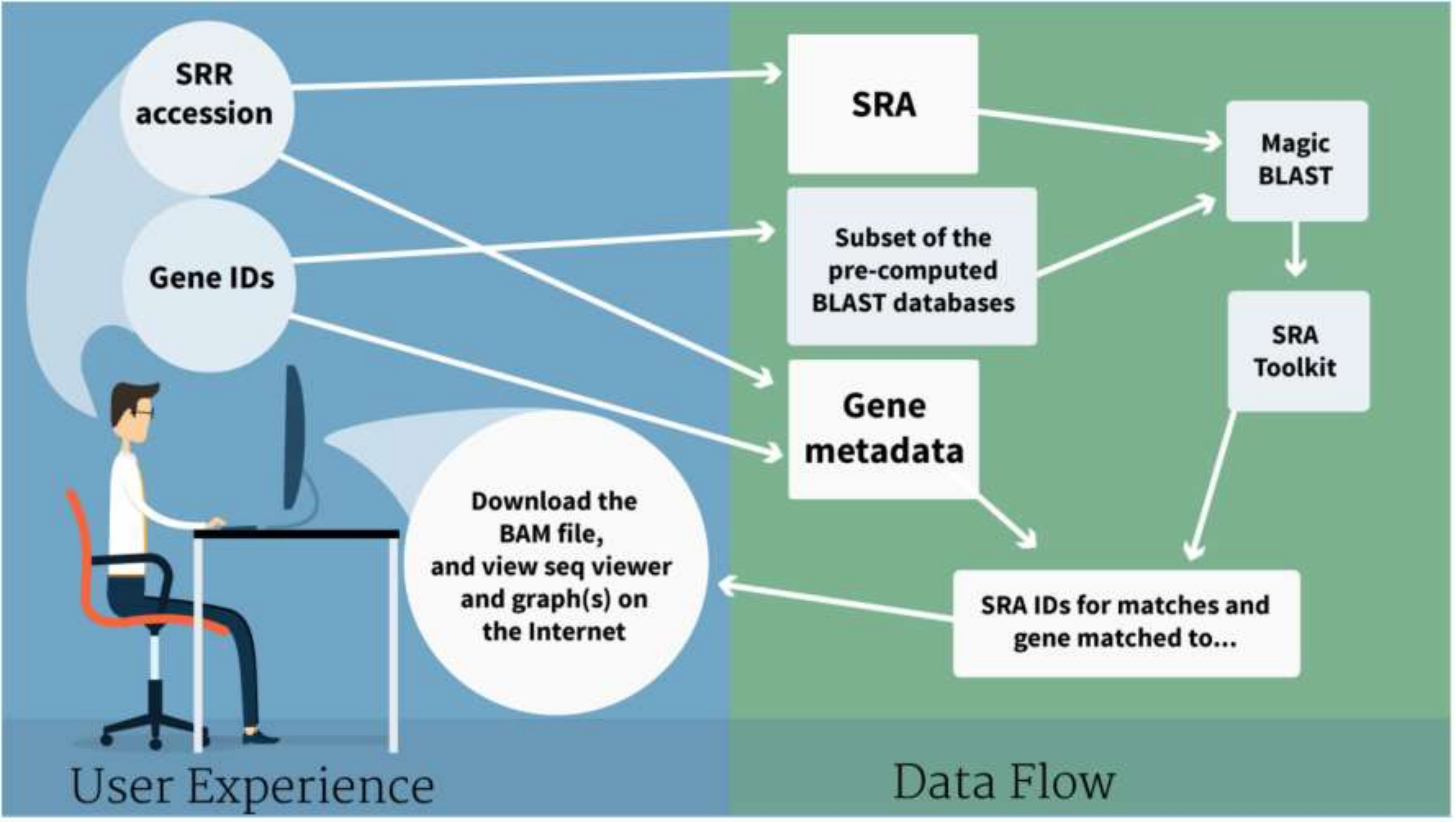
Animate F0 Animate F10 Animate F150

Gene expression (high: >= 2, low < 2; edgeR normalized):

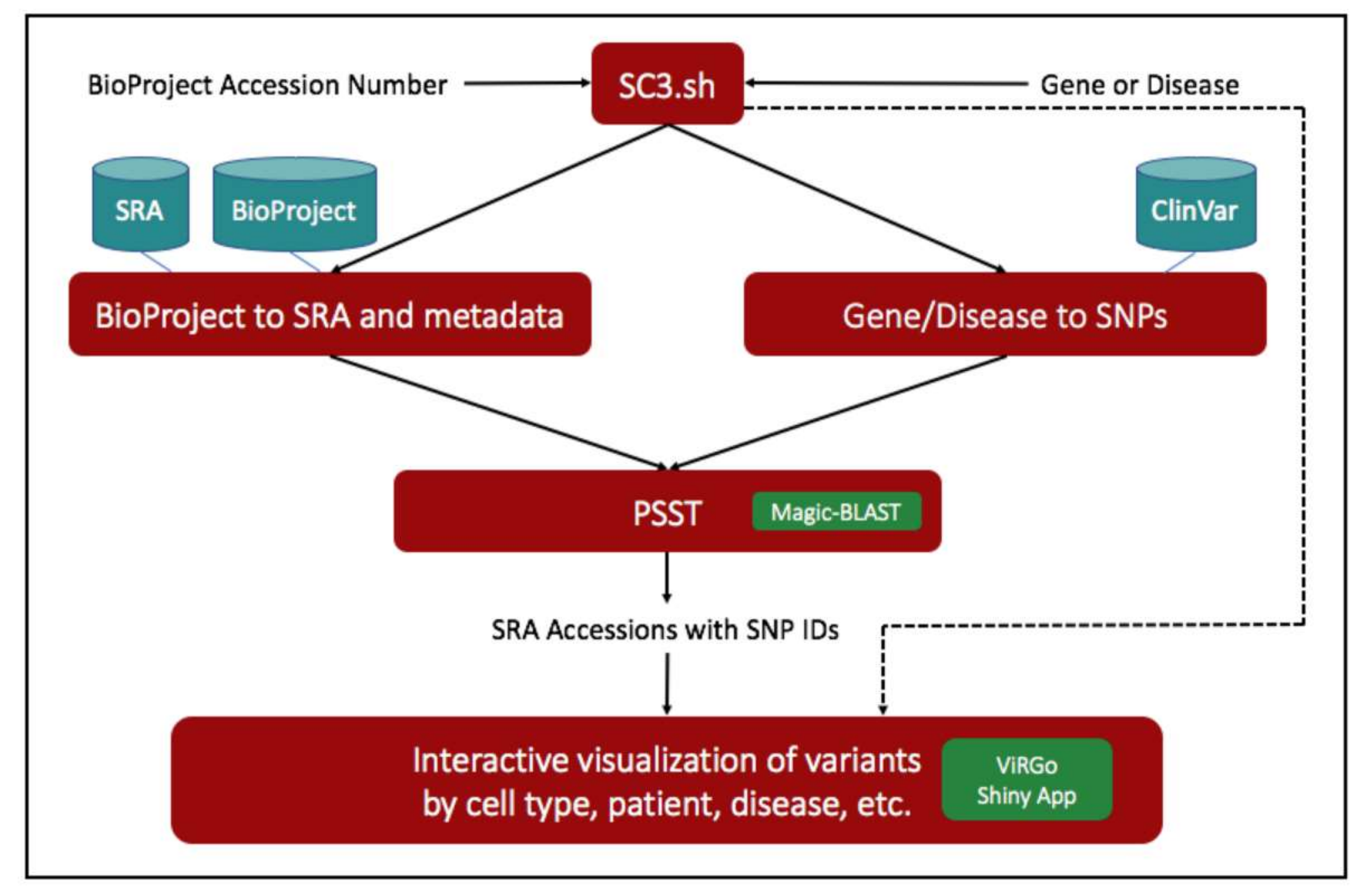
01M_F0	04M_F0	09M_F0	12M_F0	18M_F0	24M_F0
□ High	□ High	□ High	□ High	□ High	□ High
□ Low	□ Low	□ Low	□ Low	□ Low	☑ Low

https://ncbi-hackathons.github.io/GeneExpressionAging/ideogram

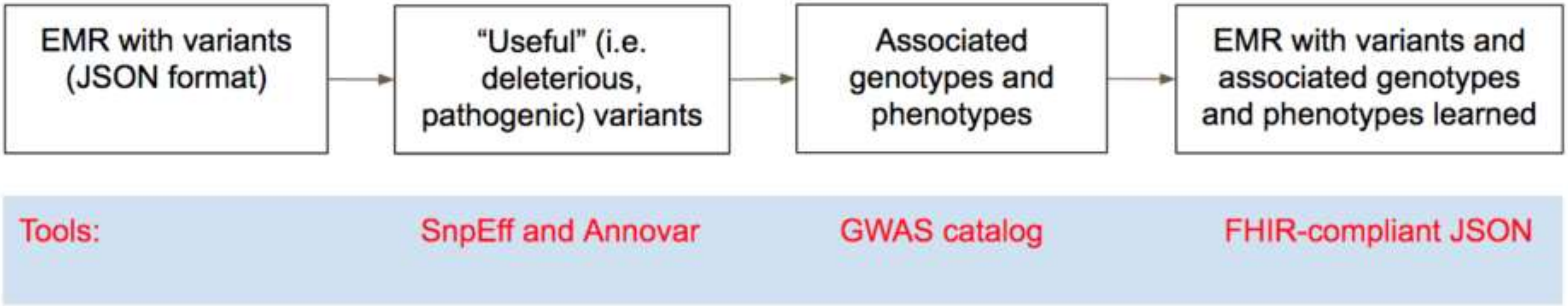
deSRA/viSRA (prototype)


SRR accession
Gene IDs
SRA
Subset of the pre-computed BLAST databases
Gene metadata
Magic BLAST
SRA Toolkit
SRA IDs for matches and gene matched to...
Download the BAM file, and view seq viewer and graph(s) on the Internet
User Experience
Data Flow

VIRGO (prototype)

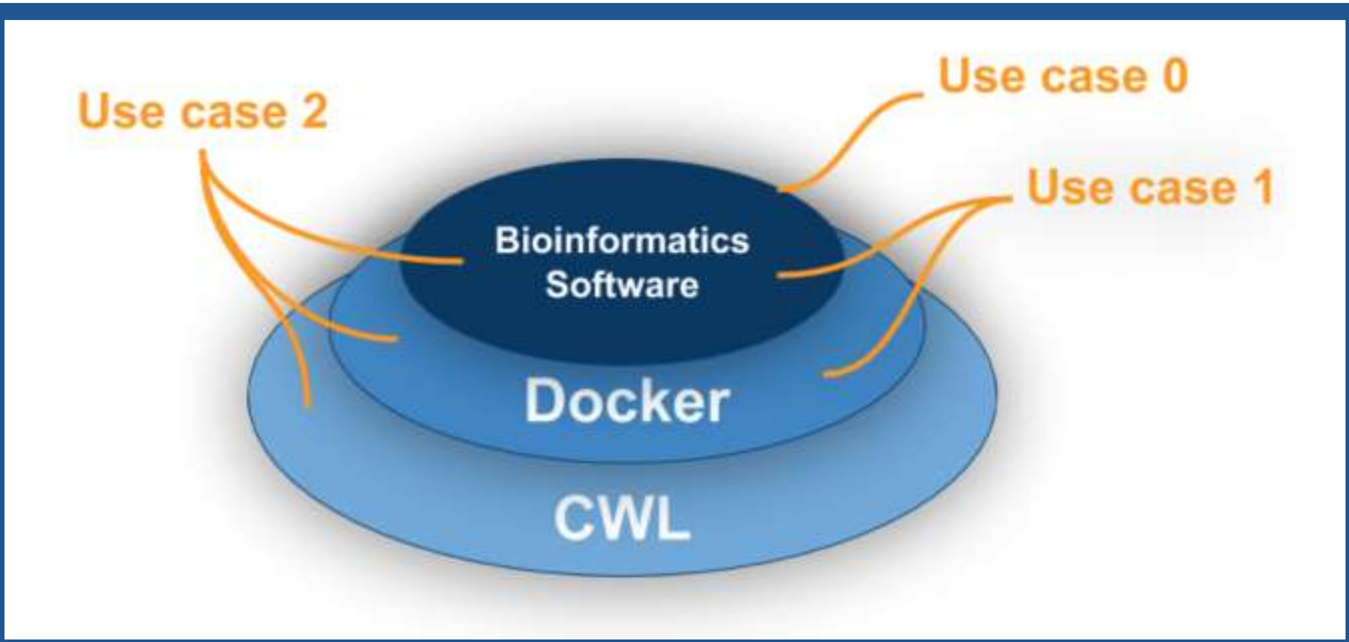

BioProject Accession Number
SC3.sh
Gene or Disease
SRA
BioProject
ClinVar
BioProject to SRA and metadata
Gene/Disease to SNPs
PSST
Magic-BLAST
SRA Accessions with SNP IDs
Interactive visualization of variants by cell type, patient, disease, etc.
ViRGo Shiny App

Translating from Bioinformatics and Clinical Informatics

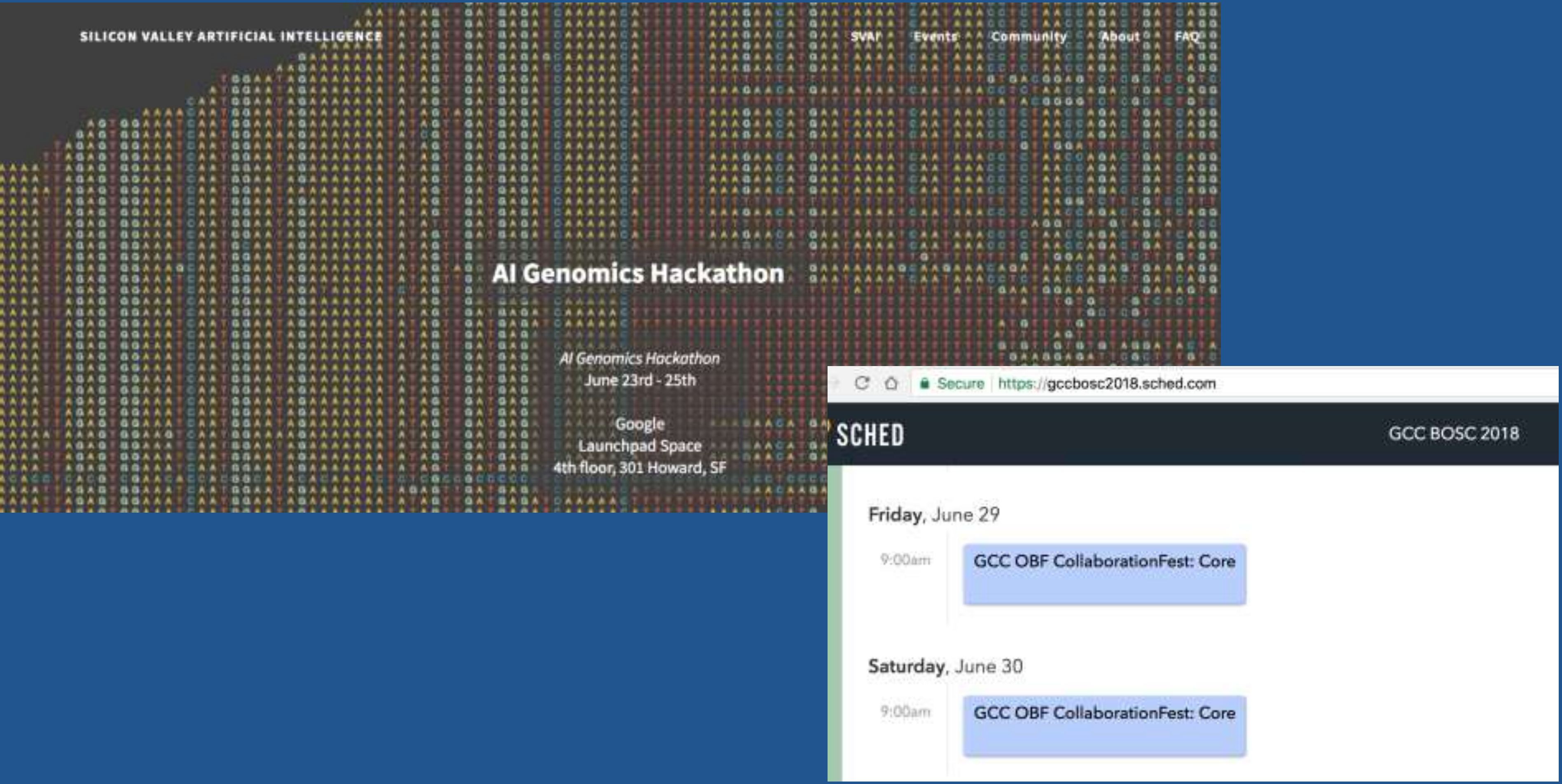


EMR with variants (JSON format) → "Useful" (i.e. deleterious, pathogenic) variants → Associated genotypes and phenotypes → EMR with variants and associated genotypes and phenotypes learned

Tools: SnpEff and Annovar | GWAS catalog | FHIR-compliant JSON



Other People's Hackathons


SILICON VALLEY ARTIFICIAL INTELLIGENCE
SVAI
Events
Community
About
FAQ

AI Genomics Hackathon

AI Genomics Hackathon
June 23rd - 25th

Google
Launchpad Space
4th floor, 301 Howard, SF

Secure
https://gccbosc2018.sched.com
SCHED

GCC BOSC 2018

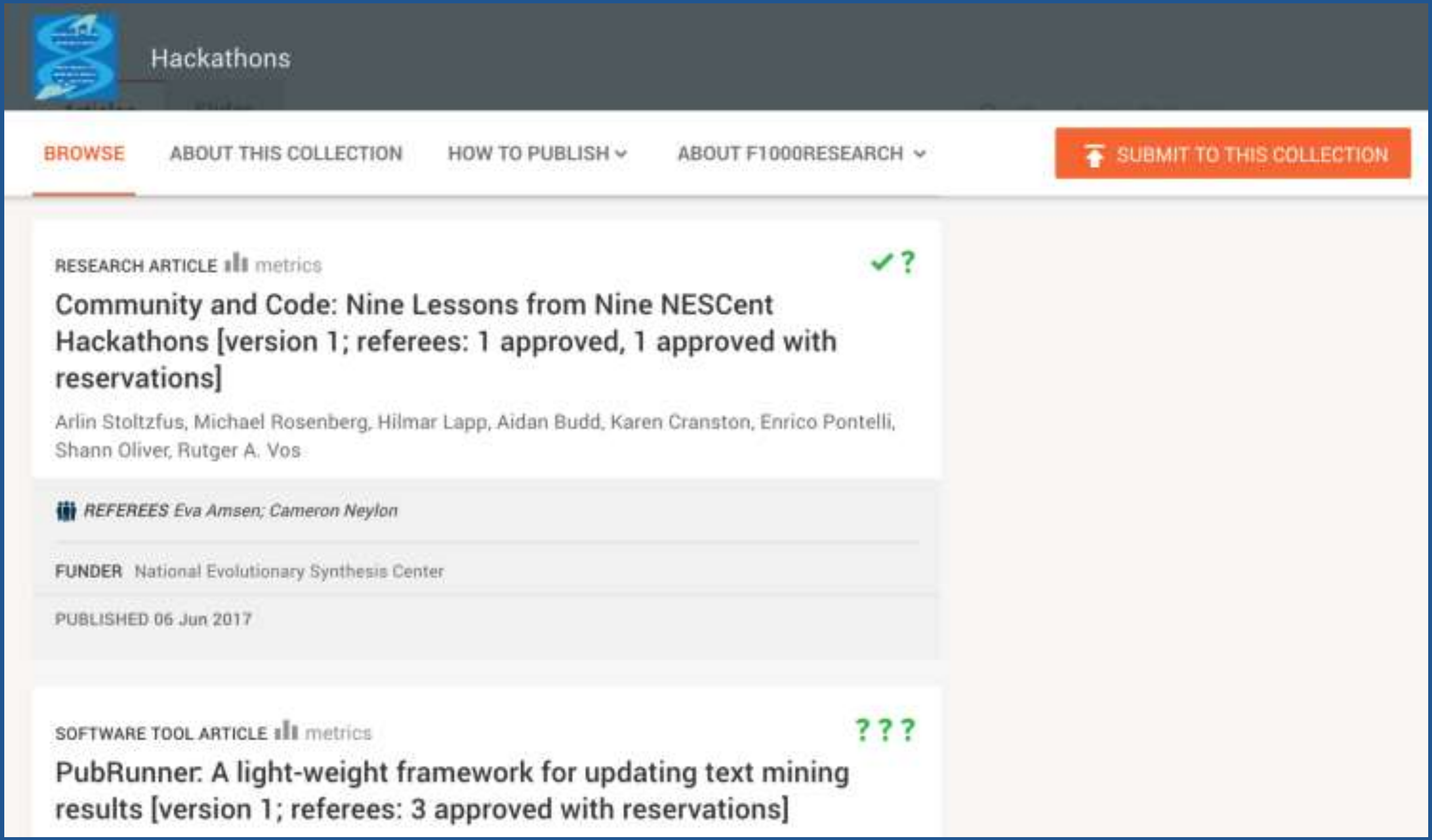
Friday, June 29

9:00am GCC OBF CollaborationFest: Core

Saturday, June 30

9:00am GCC OBF CollaborationFest: Core

Communication



RESEARCH ARTICLE metrics ✓ ?

Community and Code: Nine Lessons from Nine NESCent Hackathons [version 1; referees: 1 approved, 1 approved with reservations]

Arlin Stoltzfus, Michael Rosenberg, Hilmar Lapp, Aidan Budd, Karen Cranston, Enrico Pontelli, Shann Oliver, Rutger A. Vos

REFEREES Eva Amsen; Cameron Neylon

FUNDER National Evolutionary Synthesis Center

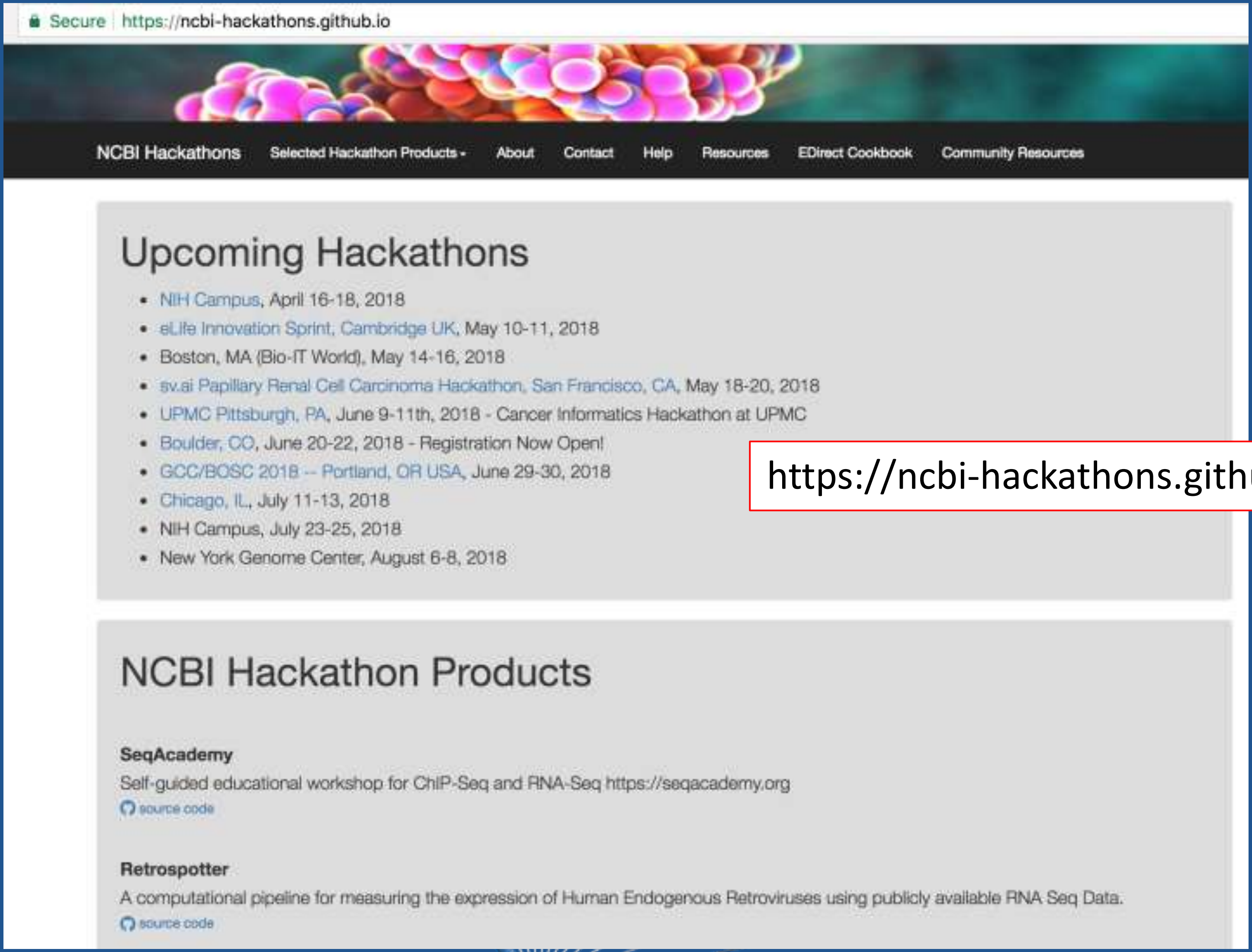
PUBLISHED 06 Jun 2017

SOFTWARE TOOL ARTICLE metrics ? ? ?

PubRunner: A light-weight framework for updating text mining results [version 1; referees: 3 approved with reservations]

@DCGenomics

Creating a Community



Upcoming Hackathons

- NIH Campus, April 16-18, 2018
- eLife Innovation Sprint, Cambridge UK, May 10-11, 2018
- Boston, MA (Bio-IT World), May 14-16, 2018
- sv.ai Papillary Renal Cell Carcinoma Hackathon, San Francisco, CA, May 18-20, 2018
- UPMC Pittsburgh, PA, June 9-11th, 2018 - Cancer Informatics Hackathon at UPMC
- Boulder, CO, June 20-22, 2018 - Registration Now Open!
- GCC/BOSC 2018 -- Portland, OR USA, June 29-30, 2018
- Chicago, IL, July 11-13, 2018
- NIH Campus, July 23-25, 2018
- New York Genome Center, August 6-8, 2018

NCBI Hackathon Products

SeqAcademy

Self-guided educational workshop for ChIP-Seq and RNA-Seq https://seqacademy.org

source code

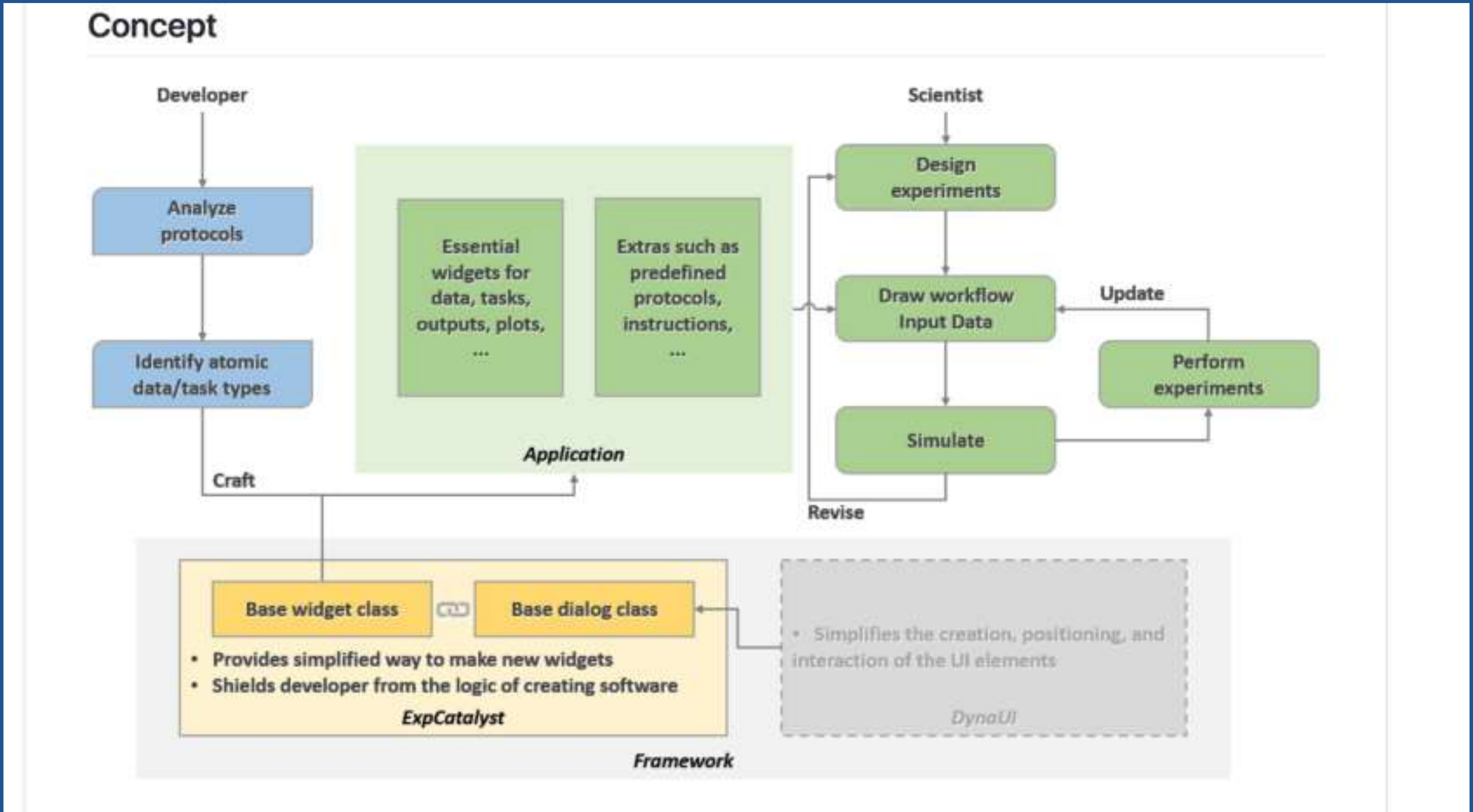
Retrospotter

A computational pipeline for measuring the expression of Human Endogenous Retroviruses using publicly available RNA Seq Data.

source code

https://ncbi-hackathons.github.io

Creating a Community



Concept

Developer

Analyze protocols

Identify atomic data/task types

Craft

Essential widgets for data, tasks, outputs, plots, ...

Extras such as predefined protocols, instructions, ...

Application

Scientist

Design experiments

Draw workflow Input Data

Simulate

Perform experiments

Update

Revise

Base widget class

Base dialog class

- Provides simplified way to make new widgets
- Shields developer from the logic of creating software

ExpCatalyst

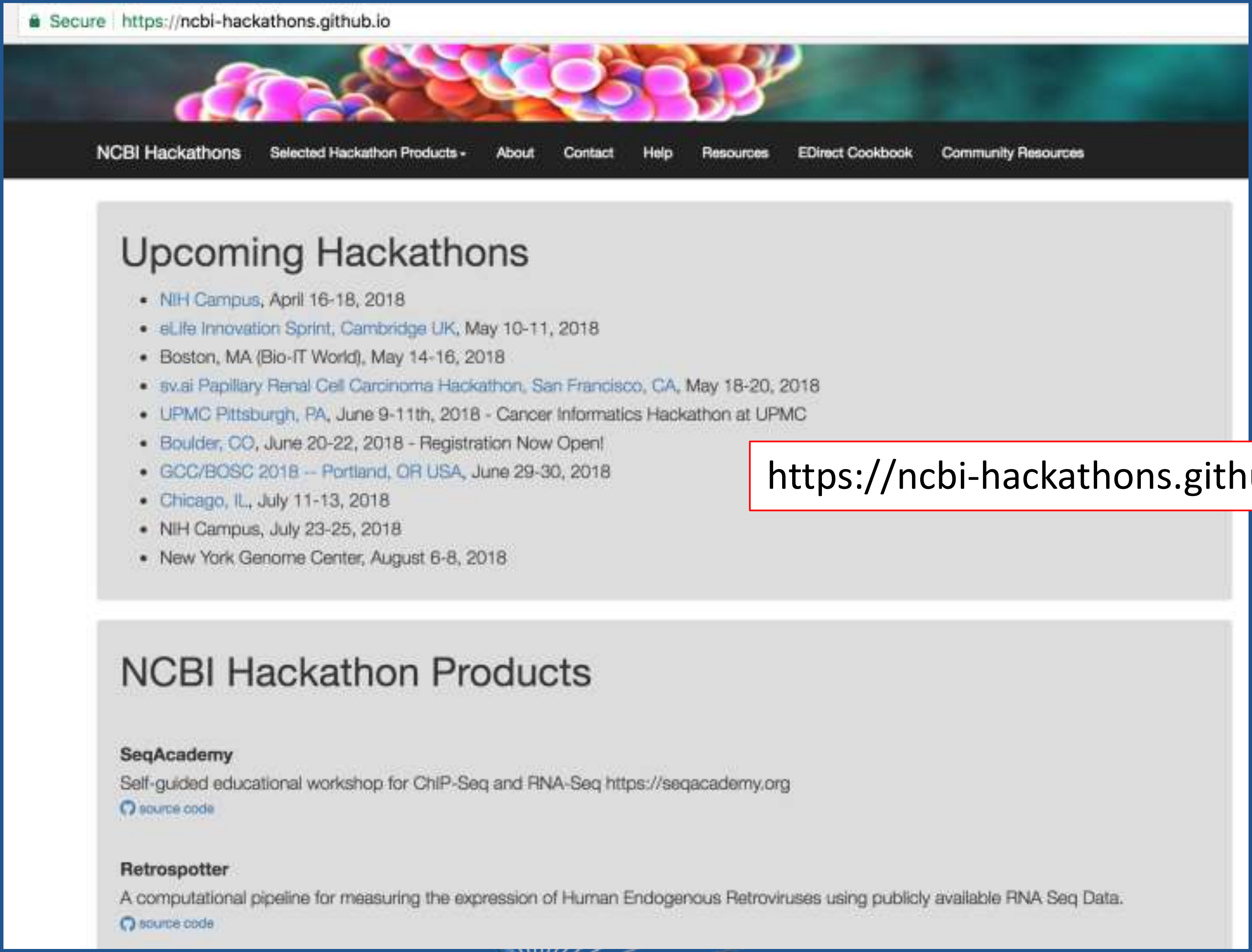
- Simplifies the creation, positioning, and interaction of the UI elements

DynaUI

Framework

Come work at NCBI for 4-6 weeks!
Email bioinformatics-training@ncbi.nlm.nih.gov
for more information!

Creating a Community



Secure | https://ncbi-hackathons.github.io

Upcoming Hackathons

- NIH Campus, April 16-18, 2018
- eLife Innovation Sprint, Cambridge UK, May 10-11, 2018
- Boston, MA (Bio-IT World), May 14-16, 2018
- sv.ai Papillary Renal Cell Carcinoma Hackathon, San Francisco, CA, May 18-20, 2018
- UPMC Pittsburgh, PA, June 9-11th, 2018 - Cancer Informatics Hackathon at UPMC
- Boulder, CO, June 20-22, 2018 - Registration Now Open!
- GCC/BOSC 2018 -- Portland, OR USA, June 29-30, 2018
- Chicago, IL, July 11-13, 2018
- NIH Campus, July 23-25, 2018
- New York Genome Center, August 6-8, 2018

NCBI Hackathon Products

SeqAcademy

Self-guided educational workshop for ChIP-Seq and RNA-Seq https://seqacademy.org

source code

Retrospotter

A computational pipeline for measuring the expression of Human Endogenous Retroviruses using publicly available RNA Seq Data.

source code

https://ncbi-hackathons.github.io