The Influenza Research Database is a “one-stop shop”

Emma Tyrnauer & Edward Bachoura

Department of Biology & Department of Computer Science

Loyola Marymount University

October 4, 2017

https://www.fludb.org/

Outline

1. General information
2. Scientific quality of IRD
3. General utility to the scientific community
4. Summary judgment

Outline

1. General information
2. Scientific quality of IRD
3. General utility to the scientific community
4. Summary judgment

General Information (Content)

- Comprised of 3 main features:
  - Influenza virus-related data (surveillance, clinical, phenotypic, genomic, and proteomic)
  - Tools (analysis and visualization)
  - Workbench (storage)
- Includes both primary and secondary data
- Both electronic and manual in-house curation

https://doi.org/10.1093/nar/gkw857 and https://www.fludb.org/brc/home.spg?decorator=influenza

General Information (Maintenance and Funding)

- **Maintained by 29 individuals**
  - Northrop Grumman Health and Human Services, J. Craig Venter Institute, and Vecna Technologies
- **Public encouraged to submit their own material**
- **Funding from National Institute of Allergy and Infectious Diseases, National Institutes of Health, and Department of Health and Human Services**

https://doi.org/10.1093/nar/gkw857 and https://www.fludb.org/brc/home.spg?decorator=influenza

Outline

1. General information
2. Scientific quality of IRD
3. General utility to the scientific community
4. Summary judgment

Scientific Quality (Content and Usefulness)

- 2,408,618 aggregated and 1,434,077 derived records
- Focuses on avian and mammalian interactions with the influenza virus
- Useful for:
  - Researching and developing vaccines
  - Understanding diagnostics and therapeutics against influenza

https://doi.org/10.1093/nar/gkw857 and https://www.fludb.org/brc/home.spg?decorator=influenza

Scientific Quality (Content and Usefulness)

- Objective is “to provide a one-stop shop for influenza virus data and analysis tools to drive new discoveries about influenza virus transmission, virulence, host range and pathogenesis, and to develop novel strategies for diagnosis, prevention and therapeutic intervention”

https://doi.org/10.1093/nar/gkw857

Scientific Quality (Relevance and Upkeep)

- The Influenza virus is a major global public threat
- NCBI’s Influenza Virus Database is similar
- Most distant publication was in 2006
- Sequences from Genbank downloaded and curated daily
- Generated and imported data updates occur less frequently
    - May 2017- Oct 2017

https://doi.org/10.1093/nar/gkw857 and https://www.fludb.org/brc/home.spg?decorator=influenza

Outline

1. General information
2. Scientific quality of IRD
3. General utility to the scientific community
4. Summary judgment

General Utility

- Within IRD, some of the data is generated in house, but the database also regularly imports data from the following external databases:
  - NCBI - Genbank
  - NCBI - RefSeq
  - Immune Epitope Database
  - UniProt
  - RCSB Protein Data Bank
  - Catalytic Site Atlas
  - Avibase

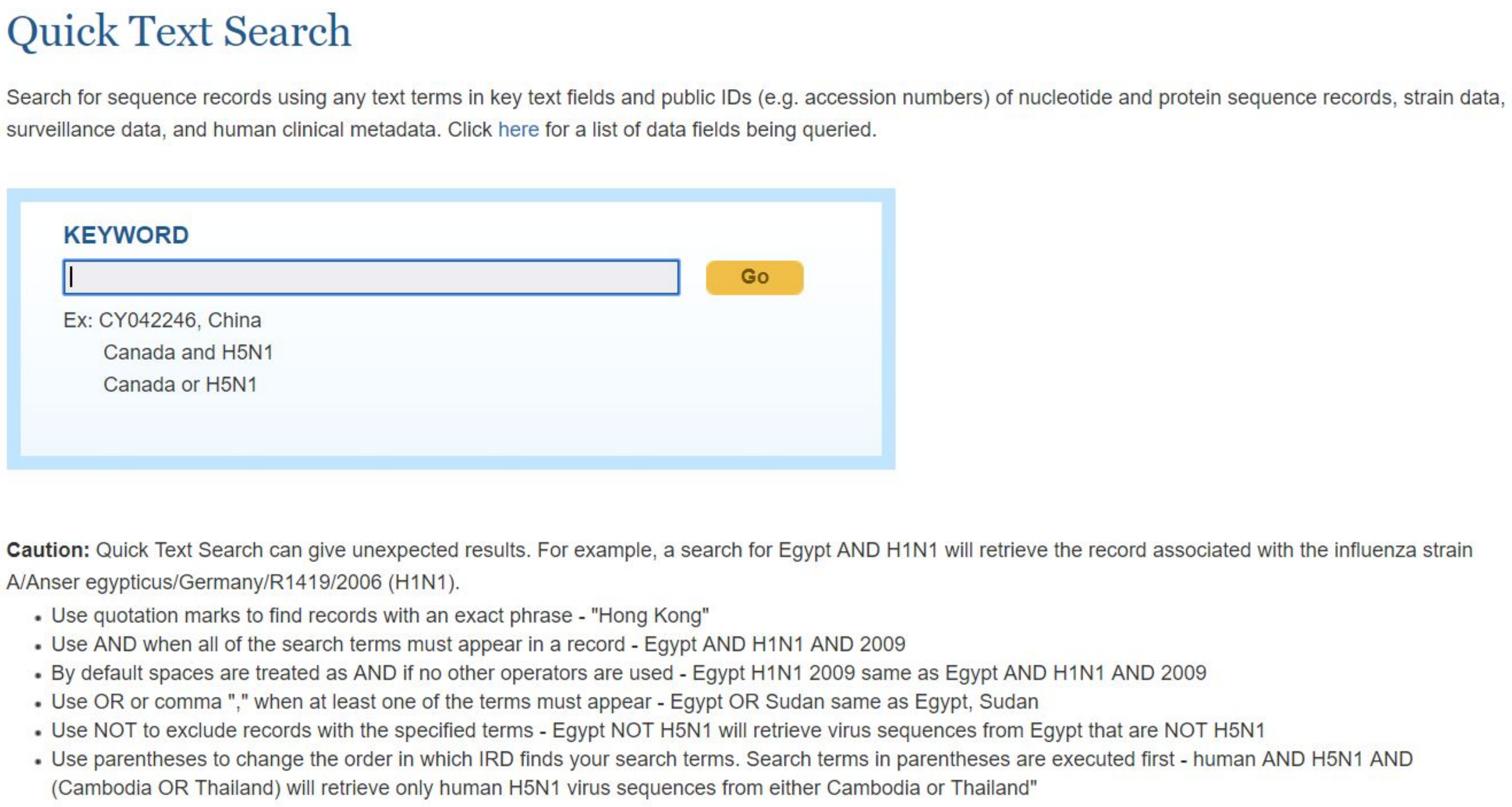
General Utility

- Multiple ways to search within their database
- Also, there are a number of tools which make it convenient to:
  - Refine and analyze the search
  - Save the data you are working on to your "workbench”

General Utility

- You can use their quick search tool to:
  - “Search for sequence records using any text terms in key text fields and public IDs (e.g. accession numbers) of nucleotide and protein sequence records, strain data, surveillance data, and human clinical metadata."
- Very useful if you know exactly what you are looking for and how to find it

General Utility



Quick Text Search

Search for sequence records using any text terms in key text fields and public IDs (e.g. accession numbers) of nucleotide and protein sequence records, strain data, surveillance data, and human clinical metadata. Click here for a list of data fields being queried.

Ex: CY042246, China
Canada and H5N1
Canada or H5N1

Caution: Quick Text Search can give unexpected results. For example, a search for Egypt AND H1N1 will retrieve the record associated with the influenza strain A/Anser egypticus/Germany/R1419/2006 (H1N1).

- Use quotation marks to find records with an exact phrase - "Hong Kong"
- Use AND when all of the search terms must appear in a record - Egypt AND H1N1 AND 2009
- By default spaces are treated as AND if no other operators are used - Egypt H1N1 2009 same as Egypt AND H1N1 AND 2009
- Use OR or comma "," when at least one of the terms must appear - Egypt OR Sudan same as Egypt, Sudan
- Use NOT to exclude records with the specified terms - Egypt NOT H5N1 will retrieve virus sequences from Egypt that are NOT H5N1
- Use parentheses to change the order in which IRD finds your search terms. Search terms in parentheses are executed first - human AND H5N1 AND (Cambodia OR Thailand) will retrieve only human H5N1 virus sequences from either Cambodia or Thailand"

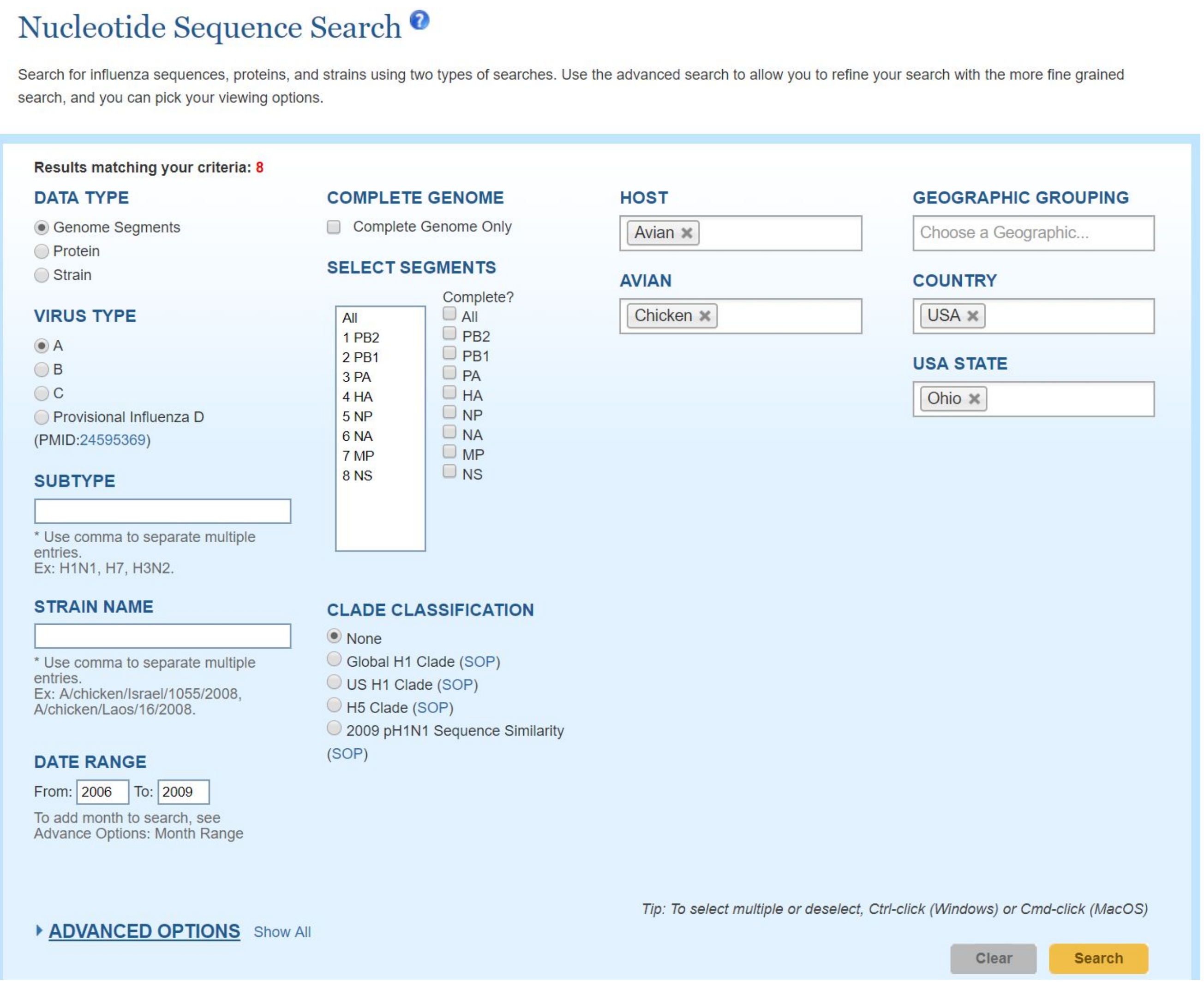
General Utility

- They also have a number of search tools that are focused on one aspect of the data.
    - Sequences & Strains
    - Animal Surveillance
    - Human Clinical Metadata
    - Host Factor Data
    - Antiviral Drugs
    - Phenotypes
    - And More!
- You would use these when you need to look through data that has a specific trait

General Utility

- For example, let’s say I want to search for Influenza Sequences or Proteins that are:
    - Type A
    - Documented from 2006 - 2009
    - Specific to Chicken
    - In Ohio
- The search would look like this:

General Utility



Nucleotide Sequence Search

Search for influenza sequences, proteins, and strains using two types of searches. Use the advanced search to allow you to refine your search with the more fine grained search, and you can pick your viewing options.

Results matching your criteria: 8

DATA TYPE

● Genome Segments
○ Protein
○ Strain

VIRUS TYPE

● A
○ B
○ C
○ Provisional Influenza D (PMID:24595369)

SUBTYPE

* Use comma to separate multiple entries.
Ex: H1N1, H7, H3N2.

STRAIN NAME

* Use comma to separate multiple entries.
Ex: A/chicken/Israel/1055/2008, A/chicken/Laos/16/2008.

DATE RANGE

From: 2006 To: 2009

To add month to search, see Advance Options: Month Range

COMPLETE GENOME

□ Complete Genome Only

SELECT SEGMENTS

	Complete?
All	□ All
1 PB2	□ PB2
2 PB1	□ PB1
3 PA	□ PA
4 HA	□ HA
5 NP	□ NP
6 NA	□ NA
7 MP	□ MP
8 NS	□ NS

CLADE CLASSIFICATION

● None
○ Global H1 Clade (SOP)
○ US H1 Clade (SOP)
○ H5 Clade (SOP)
○ 2009 pH1N1 Sequence Similarity (SOP)

HOST

Avian ✖

AVIAN

Chicken ✖

GEOGRAPHIC GROUPING

Choose a Geographic...

COUNTRY

USA ✖

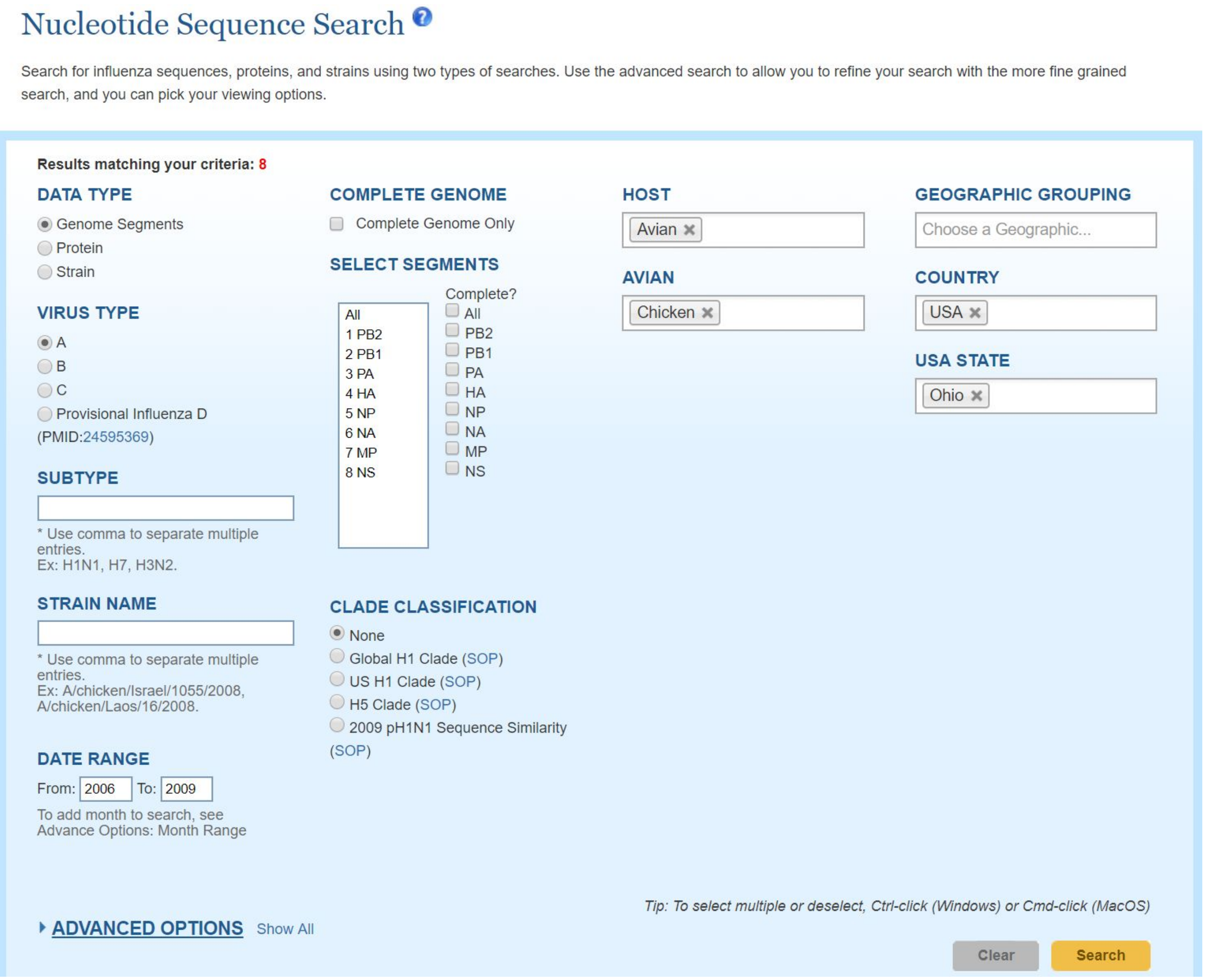
USA STATE

Ohio ✖

Tip: To select multiple or deselect, Ctrl-click (Windows) or Cmd-click (MacOS)

▸ ADVANCED OPTIONS Show All

General Utility



Nucleotide Sequence Search

Search for influenza sequences, proteins, and strains using two types of searches. Use the advanced search to allow you to refine your search with the more fine grained search, and you can pick your viewing options.

Results matching your criteria: 8

DATA TYPE

● Genome Segments
○ Protein
○ Strain

VIRUS TYPE

● A
○ B
○ C
○ Provisional Influenza D (PMID:24595369)

SUBTYPE

* Use comma to separate multiple entries.
Ex: H1N1, H7, H3N2.

STRAIN NAME

* Use comma to separate multiple entries.
Ex: A/chicken/Israel/1055/2008, A/chicken/Laos/16/2008.

DATE RANGE

From: 2006 To: 2009

To add month to search, see Advance Options: Month Range

COMPLETE GENOME

□ Complete Genome Only

SELECT SEGMENTS

	Complete?
All	□ All
1 PB2	□ PB2
2 PB1	□ PB1
3 PA	□ PA
4 HA	□ HA
5 NP	□ NP
6 NA	□ NA
7 MP	□ MP
8 NS	□ NS

CLADE CLASSIFICATION

● None
○ Global H1 Clade (SOP)
○ US H1 Clade (SOP)
○ H5 Clade (SOP)
○ 2009 pH1N1 Sequence Similarity (SOP)

HOST

Avian ✖

AVIAN

Chicken ✖

GEOGRAPHIC GROUPING

Choose a Geographic...

COUNTRY

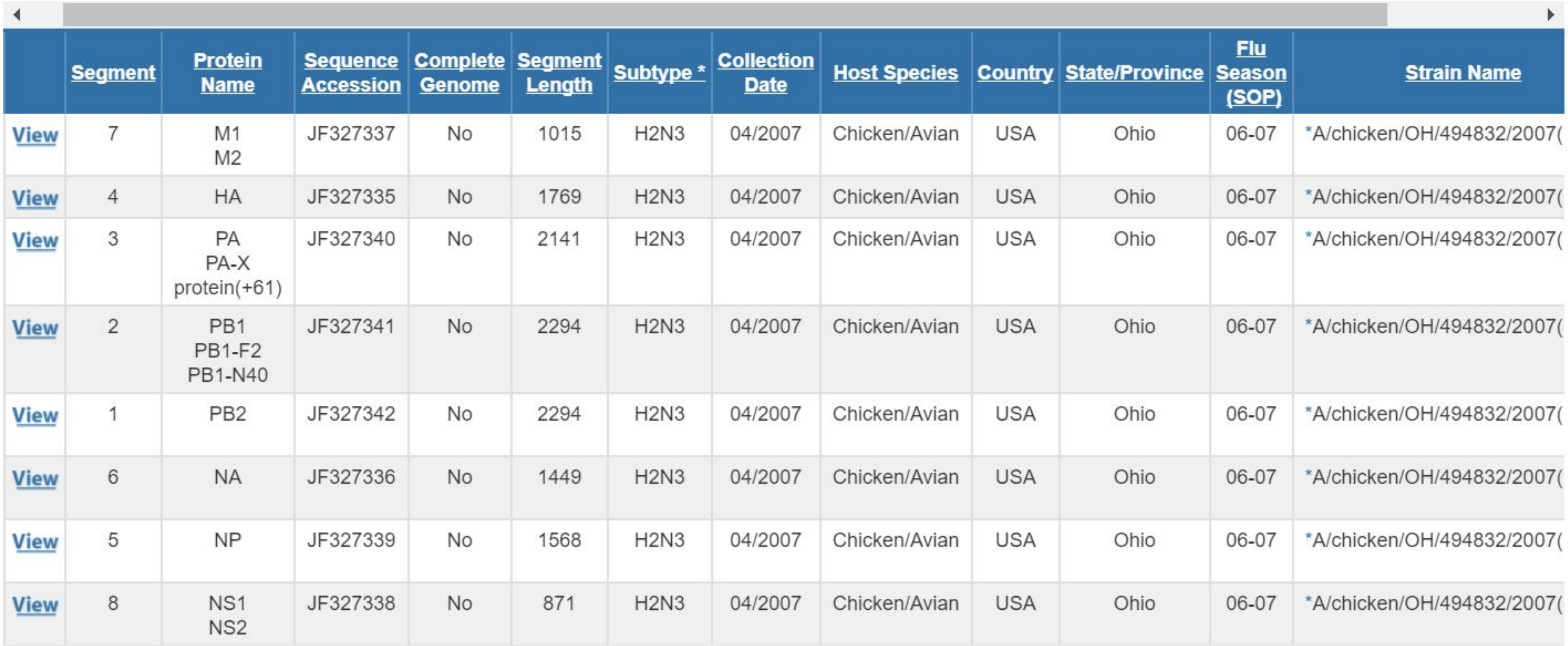
USA ✖

USA STATE

Ohio ✖

Tip: To select multiple or deselect, Ctrl-click (Windows) or Cmd-click (MacOS)

Which would, in turn, return these results:



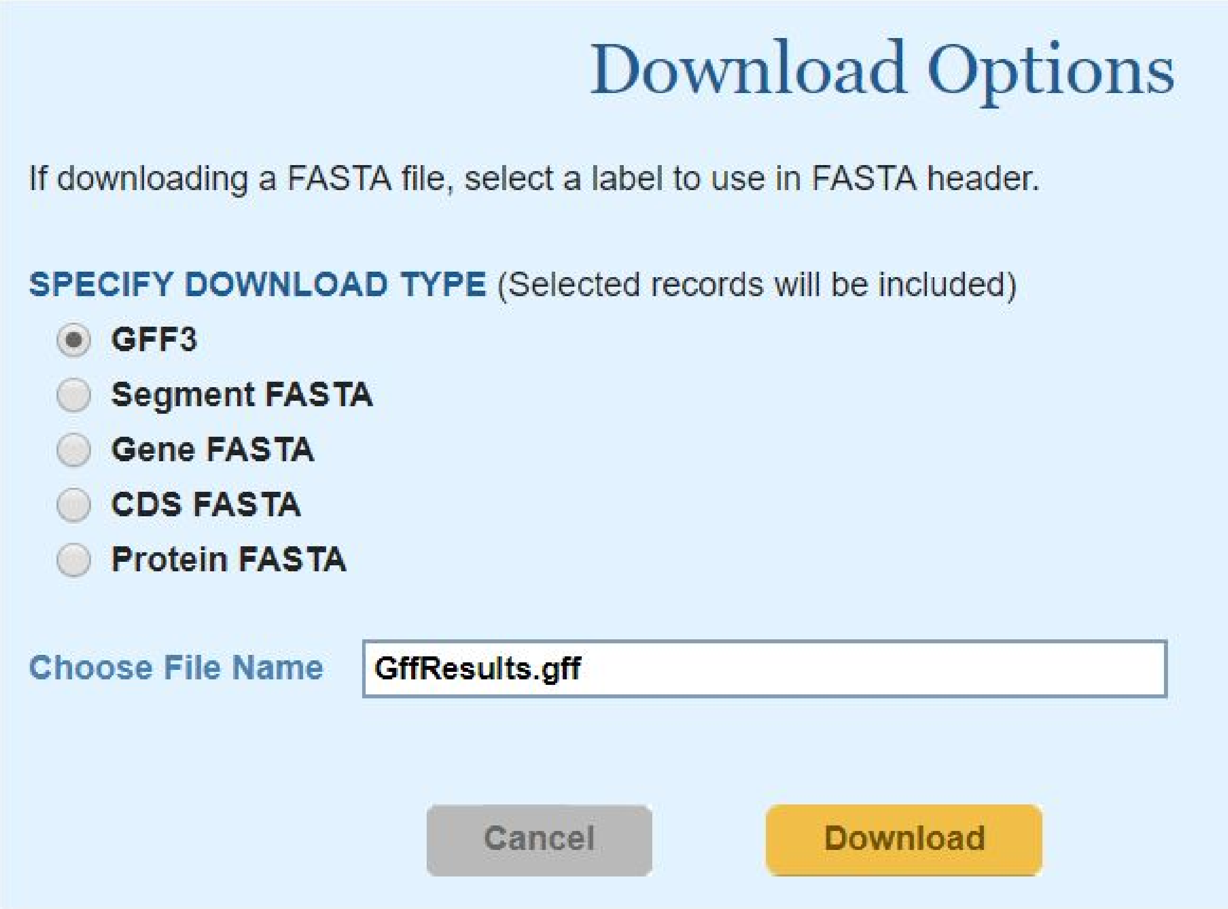
	Segment	Protein Name	Sequence Accession	Complete Genome	Segment Length	Subtype *	Collection Date	Host Species	Country	State/Province	Flu Season (SOP)	Strain Name
View	7	M1 M2	JF327337	No	1015	H2N3	04/2007	Chicken/Avian	USA	Ohio	06-07	*A/chicken/OH/494832/2007(
View	4	HA	JF327335	No	1769	H2N3	04/2007	Chicken/Avian	USA	Ohio	06-07	*A/chicken/OH/494832/2007(
View	3	PA PA-X protein(+61)	JF327340	No	2141	H2N3	04/2007	Chicken/Avian	USA	Ohio	06-07	*A/chicken/OH/494832/2007(
View	2	PB1 PB1-F2 PB1-N40	JF327341	No	2294	H2N3	04/2007	Chicken/Avian	USA	Ohio	06-07	*A/chicken/OH/494832/2007(
View	1	PB2	JF327342	No	2294	H2N3	04/2007	Chicken/Avian	USA	Ohio	06-07	*A/chicken/OH/494832/2007(
View	6	NA	JF327336	No	1449	H2N3	04/2007	Chicken/Avian	USA	Ohio	06-07	*A/chicken/OH/494832/2007(
View	5	NP	JF327339	No	1568	H2N3	04/2007	Chicken/Avian	USA	Ohio	06-07	*A/chicken/OH/494832/2007(
View	8	NS1 NS2	JF327338	No	871	H2N3	04/2007	Chicken/Avian	USA	Ohio	06-07	*A/chicken/OH/494832/2007(

Then, you can click view on the data-piece of interest:

General Utility

- The search tools offer many helpful ways to access their data and no matter which tool you use, the data that they return is exactly what you expect
- Once you are on the desired page, IRD has made it easy to download that data for your own use
- All of the datatype options that they have for download are standard formats, and are widely used in bioinformatics

General Utility



Download Options

If downloading a FASTA file, select a label to use in FASTA header.

SPECIFY DOWNLOAD TYPE (Selected records will be included)

- ● GFF3
- ○ Segment FASTA
- ○ Gene FASTA
- ○ CDS FASTA
- ○ Protein FASTA

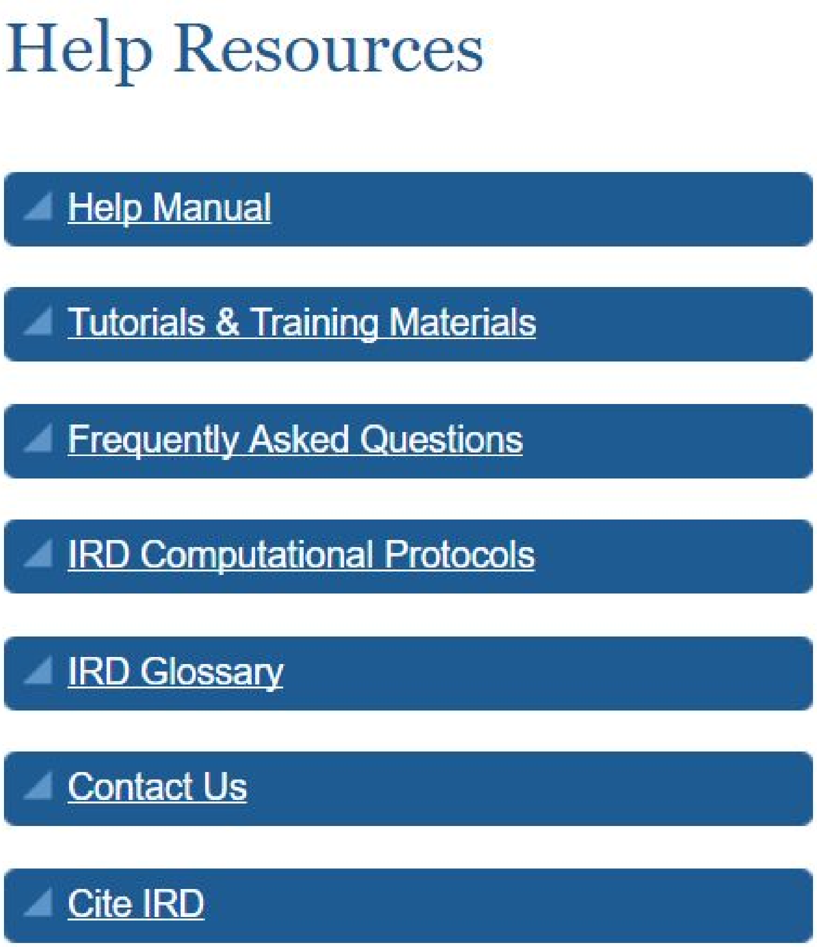
Choose File Name GffResults.gff

General Utility

- IRD is a very user-friendly database.
- They have done a good job making all of their search and download tools very obvious and easy to use.
- On that note, someone who has no biological background would have a hard time searching data, but that doesn't make the site not user-friendly.

General Utility

- And even if you can’t figure out how to use any of their tools, they have a very extensive help desk



Help Resources

- Help Manual
- Tutorials & Training Materials
- Frequently Asked Questions
- IRD Computational Protocols
- IRD Glossary
- Contact Us
- Cite IRD

General Utility

- Their help manual is very extensive with detailed written instructions on how to access/use any part of the site
- Their Tutorials and Training Materials page is also very helpful because it provides links to video instructions on how to do the most used tasks within IRD

General Utility

- After working with their site for a few hours, one might find that everything is clearly labeled and well organized
- They have five main sections right under their logo which help direct you to the part of the site that you are looking for
- The drop-down menus under each of these sections has labels that aren't confusing, and therefore take you to exactly the part of the site that you think you are going to

Outline

1. General information
2. Scientific quality of IRD
3. General utility to the scientific community
4. Summary judgment

Summary Judgment

- Influenza Research Database is good for any biologist that is either new to bioinformatics or has veteran status in the field
- It is a very well-organized site that congregates its own data with the data of most of the big name bio-databases
- The site is designed well and has clearly marked entrances and exits to every page and section which makes it good for any hobbyist or professional bioinformatician

Acknowledgments

- Dr. Dahlquist and Dr. Dionisio
- LMU Biology Department
- LMU Computer Science Department
- Class of Biological Databases

References

- LMU BioDB 2017. (2017). Week 5. Retrieved October 4, 2017, from https://xmlpipedb.cs.lmu.edu/biodb/fall2017/index.php/Week_5.
- Influenza Research Database: update 2017. Retrieved October 4, 2017, from https://www.fludb.org/brc/home.spg?decorator=influenza.
- Yun Zhang, Brian D. Aevermann, Tavis K. Anderson, David F. Burke, Gwenaelle Dauphin, Zhiping Gu, Sherry He, Sanjeev Kumar, Christopher N. Larsen, Alexandra J. Lee, Xiaomei Li, Catherine Macken, Colin Mahaffey, Brett E. Pickett, Brian Reardon, Thomas Smith, Lucy Stewart, Christian Suloway, Guangyu Sun, Lei Tong, Amy L. Vincent, Bryan Walters, Sam Zaremba, Hongtao Zhao, Liwei Zhou, Christian Zmasek, Edward B. Klem, Richard H. Scheuermann; Influenza Research Database: An integrated bioinformatics resource for influenza virus research, Nucleic Acids Research, Volume 45, Issue D1, 4 January 2017, Pages D466–D474. Retrieved October 4, 2017 from https://doi.org/10.1093/nar/gkw857.