New Unique Data and Analysis Tools in the Influenza Research Database (IRD)

Yun Zhang1, Catherine Macken2, Christian Zmasek1, Alexandra J. Lee, Tavis Anderson2, Amy Vincent2, David Burke1, Sam Zaremba2, Edward B. Klein1, Richard H. Scheuermann1,4,5, Craig Venter Institute, La Jolla, CA, USA; 2University of Auckland, Auckland, New Zealand; 3University of California, San Diego, CA, USA

Overview

Influenza Research Database (IRD, www.fludb.org), funded by the National Institute of Allergy and Infectious Diseases, serves as a single publicly-accessible repository of integrated datasets and analysis tools for influenza virus research.

IRD Integrates Data from External Sources and Generates Novel Data from Internal Computational Pipelines

IRD Provides Analysis and Visualization Tools
- BLAST Sequence Similarity Search
- Multiple Sequence Alignment
- Phylogenetics in Super-Computing Environment
- Sequence Variation (SNP) Analysis
- Metadata-driven Comparative Genomics Analysis
- Sequence Feature Variant Type (SFVT) Analysis
- Short Peptide Search
- PCR Primer Design
- Genome Annotation Including 5F-Annotation
- HPAI H5N1 & Swine H1 Clade Classifications
- HA Subtype Numbering Conversion
- 3D Protein Structure Visualization
- Host Factor Enrichment Analysis
- Antiviral Resistance Risk Assessment
- Batch sequence submission to GenBank

IRD Provides Personal Workbench for Data Storage & Sharing

Sequence Auto-Curation
- Flags potential sequencing artifacts
- Leverage the auto-curation results for sequence analysis
- Use edited version
- Exclude flagged sequences
- Include flagged sequences

Figure 1. The sequence auto-curation pipeline in IRD. Top: A schematic showing the curation pipeline flags potential sequencing artifacts. Middle: Sequences in IRD have been comprehensively curated with potential sequencing artifacts. Bottom: Users can choose whether and how to include problematic sequences in their analysis.

H1 & H5 Clade Classifications
- Swine H1 classification algorithm based on the USDA/OFUU swine H1 classification scheme
- H5 classification algorithm based on the CDC/WHO HPAI H5N1 classification scheme
- Annotated all IRD sequences with H1/H5 clade assignments
- H1 & H5 clade classification tools for user-provided sequences

Sequence Feature Phenotypic Variant Type
- Curated Sequence Features including phenotype markers in the CDC H5 Genetic Changes Inventory
- Computed Variant Types of each Sequence Feature
- Annotated all IRD sequences with the presence/absence of Phenotypic Variant Types (PVT)
- PVT annotation tool for user-provided sequences

Figure 2. Search sequences from avian host isolated from Asia based on H5 clades (I) from the Sequence Search page.

H5N1 Clade Classifications

Figure 3. Strain Desigs page shows that the HA sequence of the A/la/guizhi/4/2007 strain carries the I10N PVT substitution that has been shown to increase binding to alpha 2-6 receptor in the publication cited.

References

4CDC. http://www.cdc.gov/flu/pdfs/influenza/h5n1-inventory.pdf

HA Subtype Numbering Conversion
- Based on the HA subtype numbering scheme by Burke and Smith (2014)
- Automatically convert the coordinates of any HA protein sequences to coordinates of any other subtypes, in order to map functional domains or phenotype markers across subtypes
- Integrated with IRD analysis tools including Sequence Variation Analysis and metadata-driven Comparative Analysis Tool for Sequences (meta-CATS)

Custom Metadata Capturing
- New utility for capturing user-provided sequence annotations
- Analyze and visualize user-provided sequence data and metadata along with IRD data using any IRD tools

Figure 4. HA Subtype Numbering Conversion Result page showing the coordinates of user-provided HA protein sequence are converted to the coordinates of other HA subtypes.

Figure 5. A phylogenetic tree constructed from a combination of user-provided (downloaded from GISAID) and IRD sequences. Tree leaves are colored-coded by year. External nodes are colored by residue at position 10SS (node fill) and patient status (node border).

Conclusion

- Provides comprehensive enriched influenza virus sequence annotations
- Supports custom sequence annotation, analysis, and visualization

Acknowledgements

We would like to thank the primary data providers for the data that was used throughout this study. We also recognize the scientific and technical personnel responsible for supporting and developing IRD, which has been wholly supported by the NIH/NIAID (No. HHSN272201400028C).

Conflict of interest: None declared.