Phylogenomic and Protein Domain Architecture-based Analysis and Orthology Classification of *Herpesviridae* Proteins

Christian M. Zmasek, Ph.D.
JCVI
1. Goals

2. Background:
   1. Herpesviridae
   2. Orthology & protein function
   3. Protein domain architecture

3. Phylogenetic examples from human herpes virus genomes:
   1. DNA polymerase
   2. US22 protein family

4. Strict ortholog groups (definition)

5. Introduction to novel naming convention based on strict ortholog groups

6. Herpesviridae core strict ortholog groups examples

7. Availability of resulting data & software
Goals

• A systematic description and **classification** of the entire set of human *Herpesviridae* proteins

• Besides classifying proteins based on their **evolutionary history**, we equally consider their **domain architectures** for classification

• Goal is to classify proteins in groups (which we call "strict groups of orthologs") in which all proteins exhibit the same domain architecture and are being orthologous towards each other

• Furthermore, we attempt to provide and informative name for each these groups

• For example a "_aBG" suffix would indicate that proteins of this group are found some (but not all) human *Alphaherpesvirinae* species, and all human *Beta- and Gammaherpesvirinae* species
Herpesviridae

• Large and diverse order of dsDNA viruses that infect humans and wide range of other animals
• Human diseases caused by herpes viruses range from **blisters** to **cancer**
• **Nine** distinct viruses are known to infect humans.
• Mammalian *Herpesviridae* have been divided into **three** subfamilies:
  • **Alphaherpesvirinae**
  • **Betaherpesvirinae**
  • **Gammaherpesvirinae**
• Current research indicates that **Betaherpesvirinae** and **Gammaherpesvirinae** are more closely related to each other than to **Alphaherpesvirinae**
• In contrast to most other viruses, Herpesviridae have a long evolutionary history: the split into **Alpha-**, **Beta-**, and **Gammaherpesvirinae** occurred approximately **180 million years ago** to **220 million years ago**
Orthologs, Paralogs, Xenologs

- Homologs are defined as sequences (proteins, genes) which share a common ancestor (Fitch, 1966)
- Homologous sequences can be divided into orthologs, paralogs and xenologs
  - **Orthologs**: diverged by a speciation event (their last common ancestor on a phylogenetic tree corresponds to a speciation event)
  - **Paralogs**: diverged by a duplication event (their last common ancestor corresponds to a duplication)
  - **Xenologs**: are related to each other by horizontal gene transfer (via retroviruses, for example)
Orthology vs Function – Why are we interested in Orthologs?

- Orthologous sequences *tend* to have more similar "functions" than paralogs ("The Ortholog Conjecture")
- **Yet:** Orthologs are mathematically defined, whereas there is no definition of sequence "function" (i.e. it is a subjective term, thus the quotes)
Protein Domain Architecture

- Protein domain architecture: ordered sequence of protein domains in a protein
- Here we use domains forms the Pfam database (https://pfam.xfam.org)
- Examples (including notation) from human herpes virus genomes:
  - Uracil-DNA glycosidase (1 domain): UDG
  - Terminase (2 domains): DNA_pack_N—DNA_pack_C
  - DNA polymerase (2 domains): DNA_pol_B_exo1—DNA_pol_B
We define "Strict Ortholog Groups" (SOGs) as groups of proteins in which:

- all protein members are orthologous to each other (related by speciation events)
- and exhibit the same domain architecture
Examples:

• _A present in all human *Alphaherpesvirinae* species
• _a present in some human *Alphaherpesvirinae* species
• _aBG present in some human *Alphaherpesvirinae* species and all human *Beta*- and *Gammaherpesvirinae* species
• _ABG.B at least one domain of this architecture is present in all human *Alpha*, *Beta*- and *Gammaherpesvirinae*, yet the entire architecture is only present in (all) *Betaherpesvirinae*
Examples (3 out of 23):

- **Uracil-DNA glycosidase\_ABG** (UL2 ORF59 UL114 U81 BKRF3 ORF46): UDG

- **Helicase-primase ATPase subunit\_ABG** (UL5 ORF55 UL105 U77 BBLF4 ORF44): Herpes\_Helicase

- **Terminase\_ABG** (UL15 ORF42 UL89 U66/U60 LMP2 ORF29): DNA\_pack\_N—DNA\_pack\_C
Core SOGs: "ABG, domain rearrangements"

Examples (2 out of 8):

- **Glycoprotein B_ABG.AbG** (UL27 ORF31 U39 BALF4 ORF8): Glycoprotein_B
- **Glycoprotein B_ABG.b** (UL55): HCMVantigenic_N—Glycoprotein_B

- **DNA polymerase_ABG.a** (UL30): DNA_pol_Bexo1—DNA_pol_B—DNAPolymera_Pol
- **DNA polymerase_ABG.aBG** (ORF28 UL54 U38 BALF5 ORF9): DNA_pol_Bexo1—DNA_pol_B
Core SOGs: "ABG, non-homologous replacement"

Examples (2 out of 9):

Same Pfam clan ("DNA_clamp"), very distantly related:
• **DNA polymerase processivity subunit_A** (UL42 ORF16): Herpes_UL42—Herpes_UL42
• **DNA polymerase processivity subunit_B** (UL44 U27): Herpes_PAP
• **DNA polymerase processivity subunit_G** (BMRF1 ORF59): Herpes_DNAp_acc

Unrelated (?):
• **Glycoprotein L_A.a** (ORF60): Herpes_UL1
• **Glycoprotein L_A.S** (UL1 ): Herpes_UL1--GlyL_C
• **Glycoprotein L_B** (UL115 U82) : Cytomega_gL
• **Glycoprotein L_G** (BKRF2 ORF47): Phage_glycop_gL
Availability

• **Publication**: in review at "Virology"

• **Data**: Currently SOG analysis results are available Virus Pathogen Resource (ViPR) as searchable sequence annotations at [https://www.viprbrc.org](https://www.viprbrc.org) for the following virus families:
  • Herpesviridae
  • Poxviridae
  • Coronaviridae

• All **software** to perform DAIO analyses is freely available at: [https://sites.google.com/site/cmzrmasek/home/software/forester/daio](https://sites.google.com/site/cmzrmasek/home/software/forester/daio)
Acknowledgements

• Richard Scheuermann, J. Craig Venter Institute
• David Knipe, Department of Microbiology and Immunobiology, Harvard Medical School
• Philip Pellett, Department of Microbiology, Immunology and Biochemistry, Wayne State University School of Medicine
• Yun Zhang, J. Craig Venter Institute
• Entire ViPR/IRD development team at Northrop Grumman
• We thank the primary data providers for sharing their data in public archives, including ViPR and UniProtKB. This work was funded by the National Institute of Allergy and Infectious Diseases (NIH/DHHS) under contract no. HHSN272201400028C to RHS