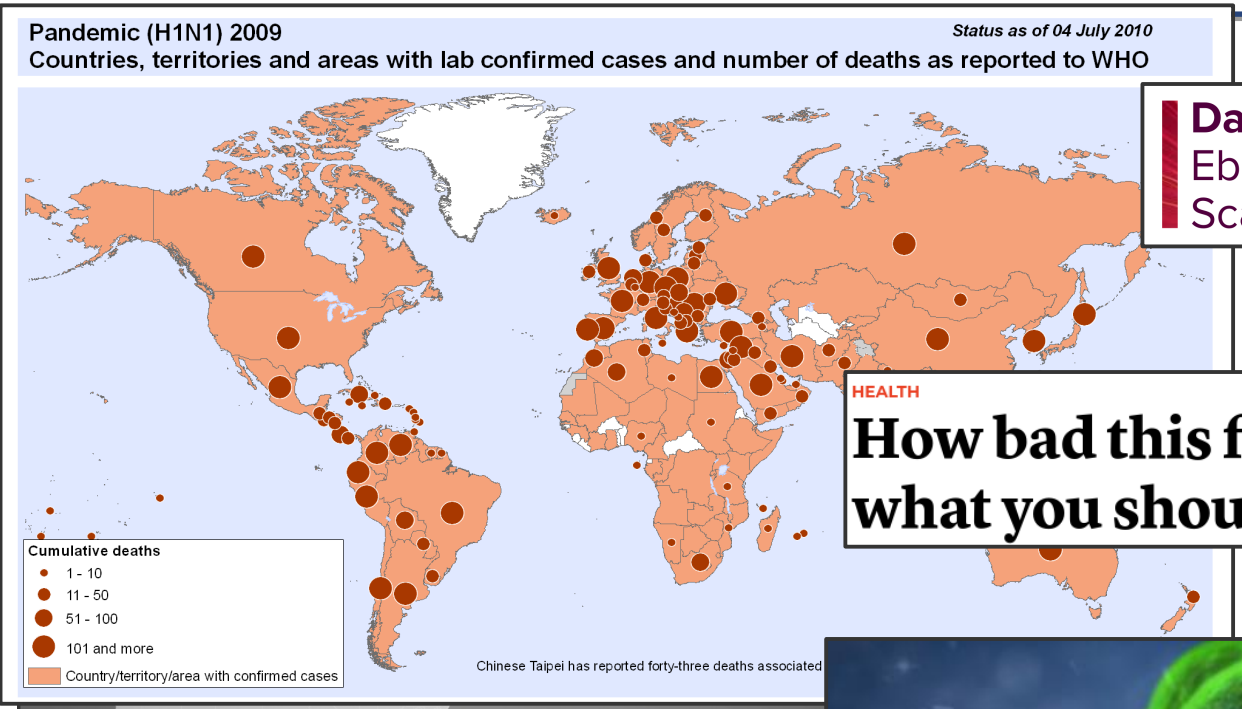
JGI JOINT GENOME INSTITUTE
A DOE OFFICE OF SCIENCE USER FACILITY

Establishing standards and best practices for describing genome sequences of uncultivated viruses (MIUViGs)

Simon Roux, DOE Joint Genome Institute, LBNL
GSC 20 – San Diego, May 4th 2018

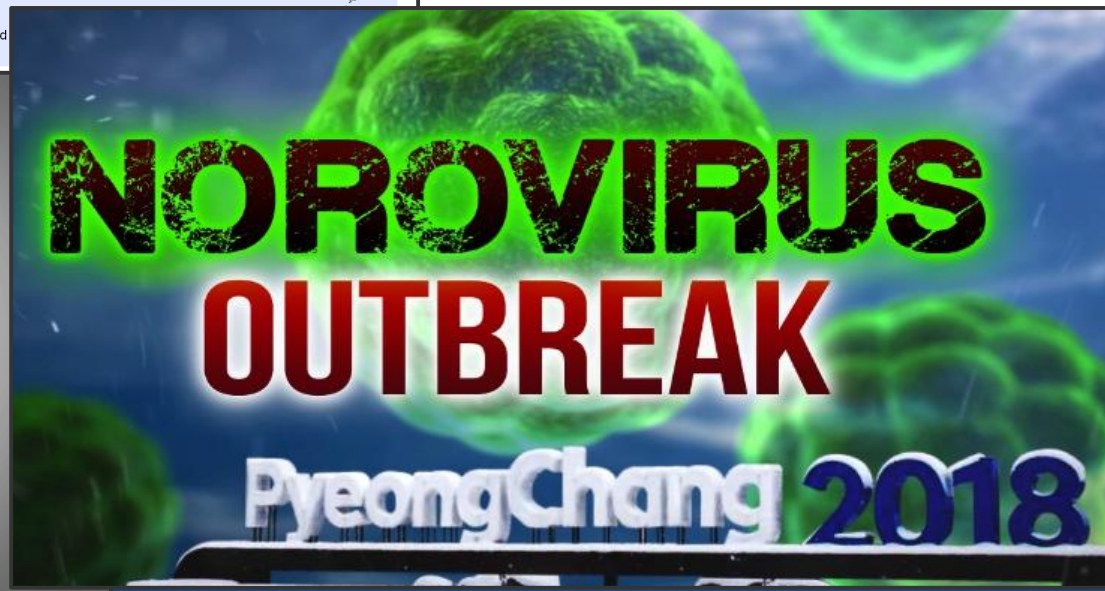
Viruses in the news



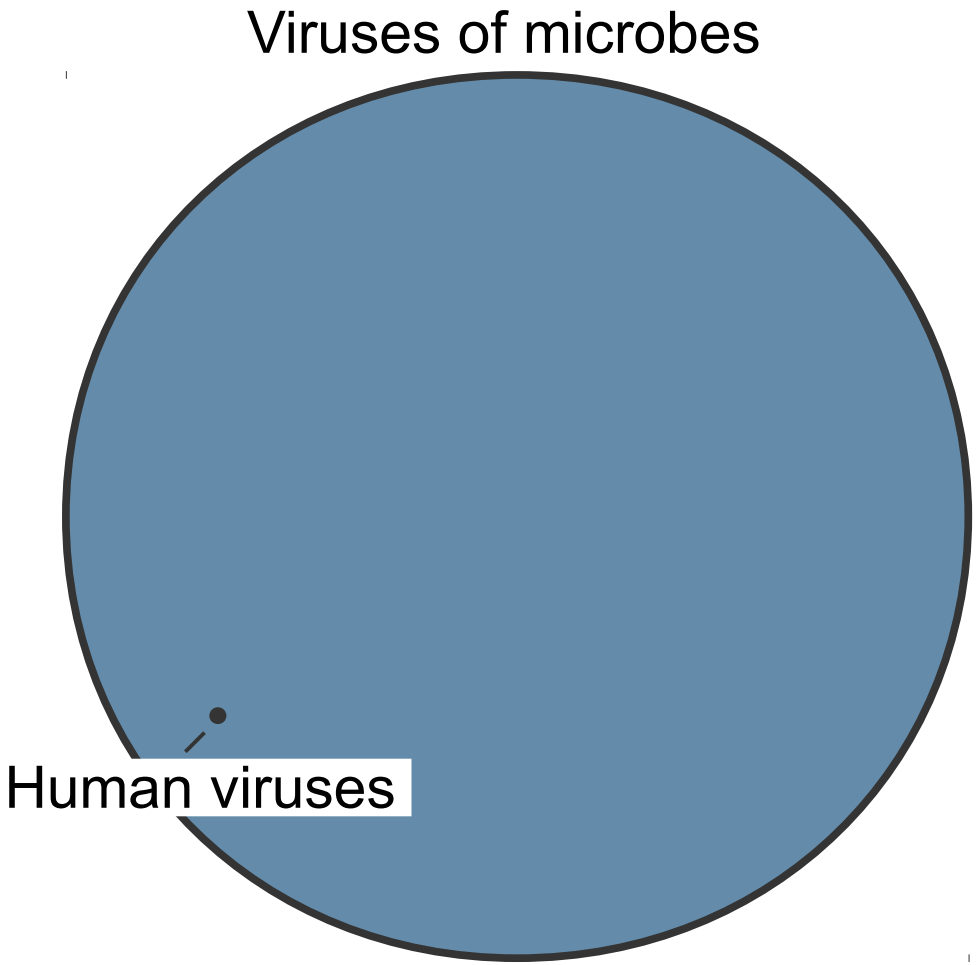
David Quammen:
Ebola to Zika and Beyond:
Scary Viruses in a Globalized World

HEALTH

How bad this flu season really is—and what you should do about it

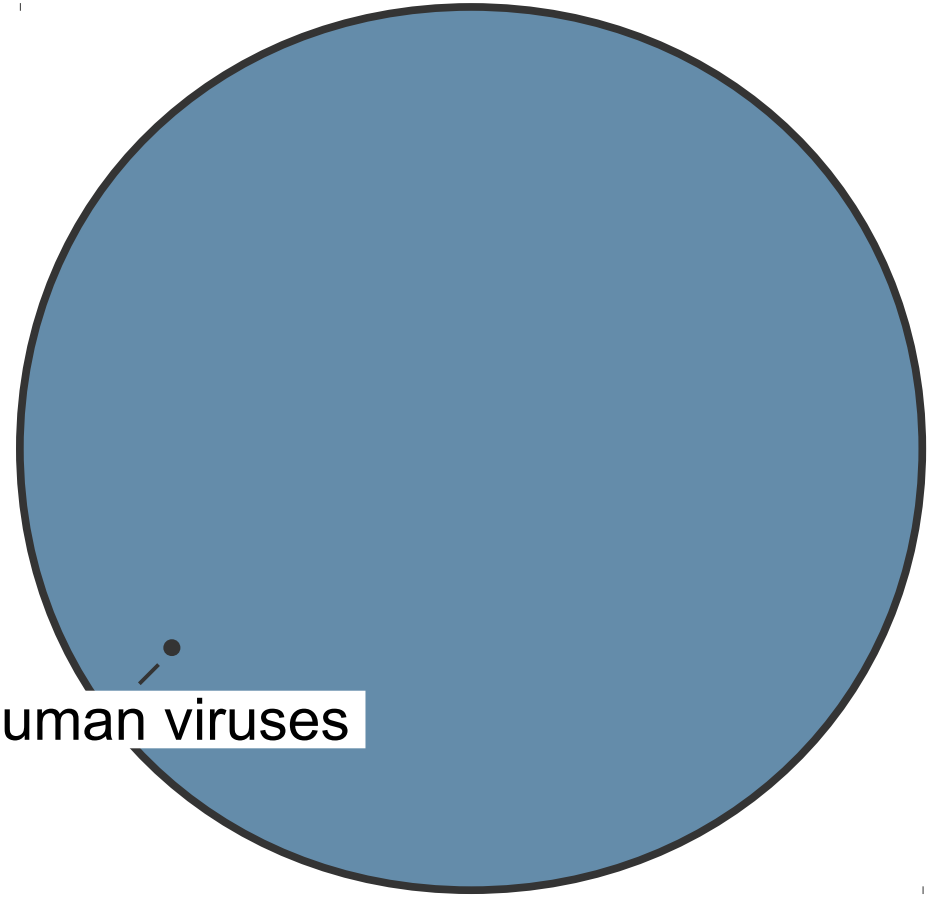


Viruses of microbes are everywhere



Viruses of microbes are everywhere

BioSciences



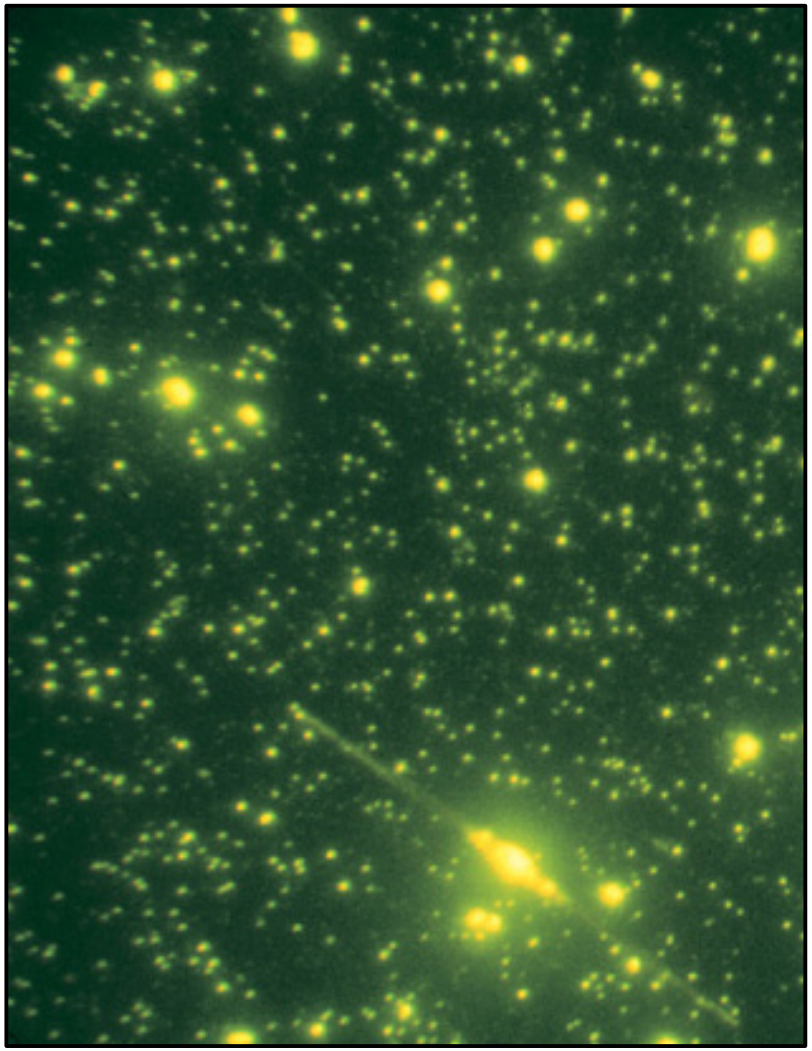
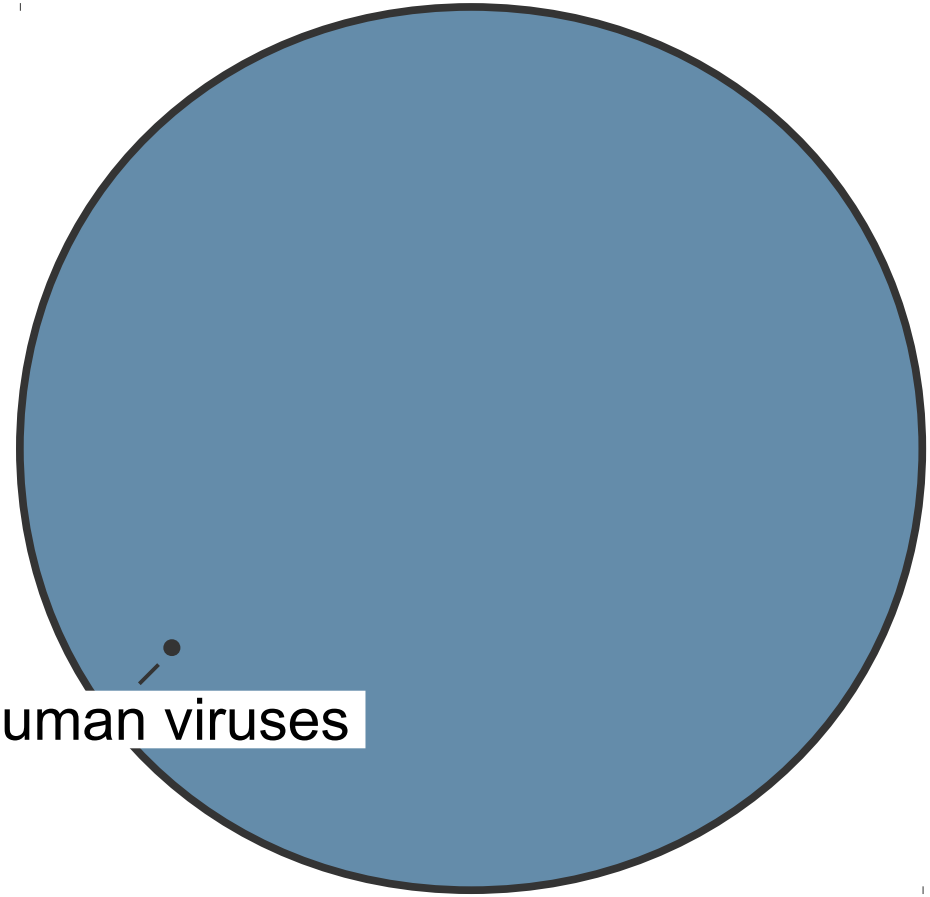


Image: Fuhrman Lab

Viruses of microbes are everywhere

BioSciences



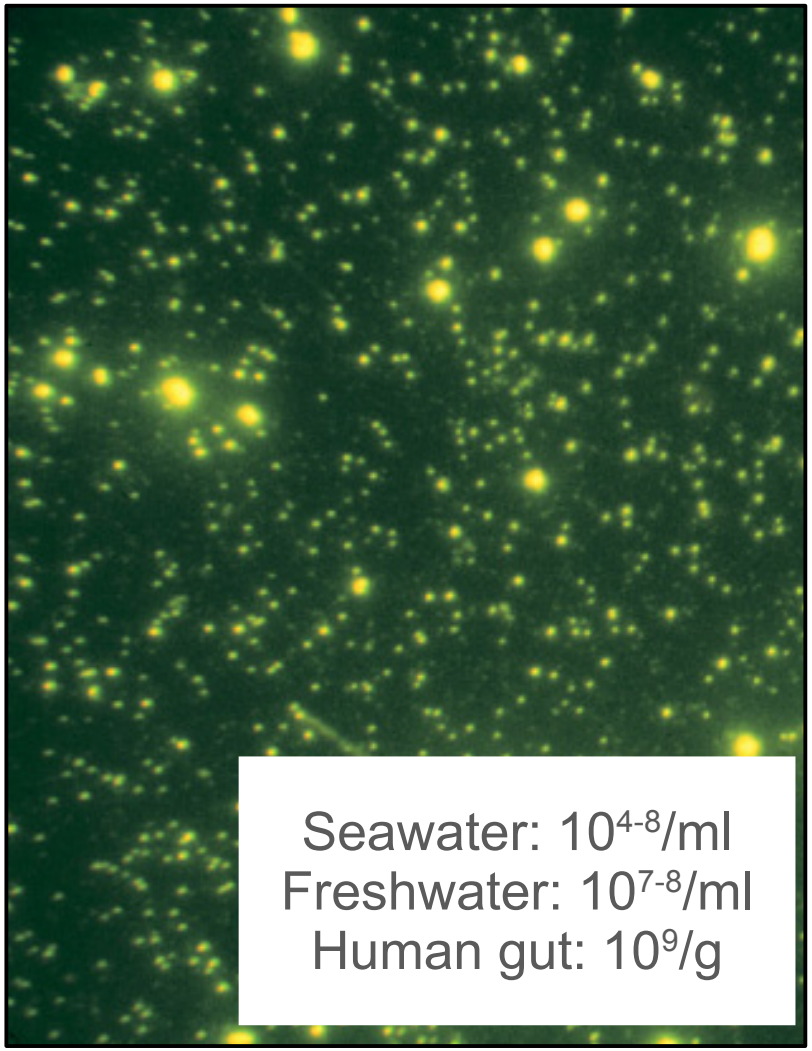
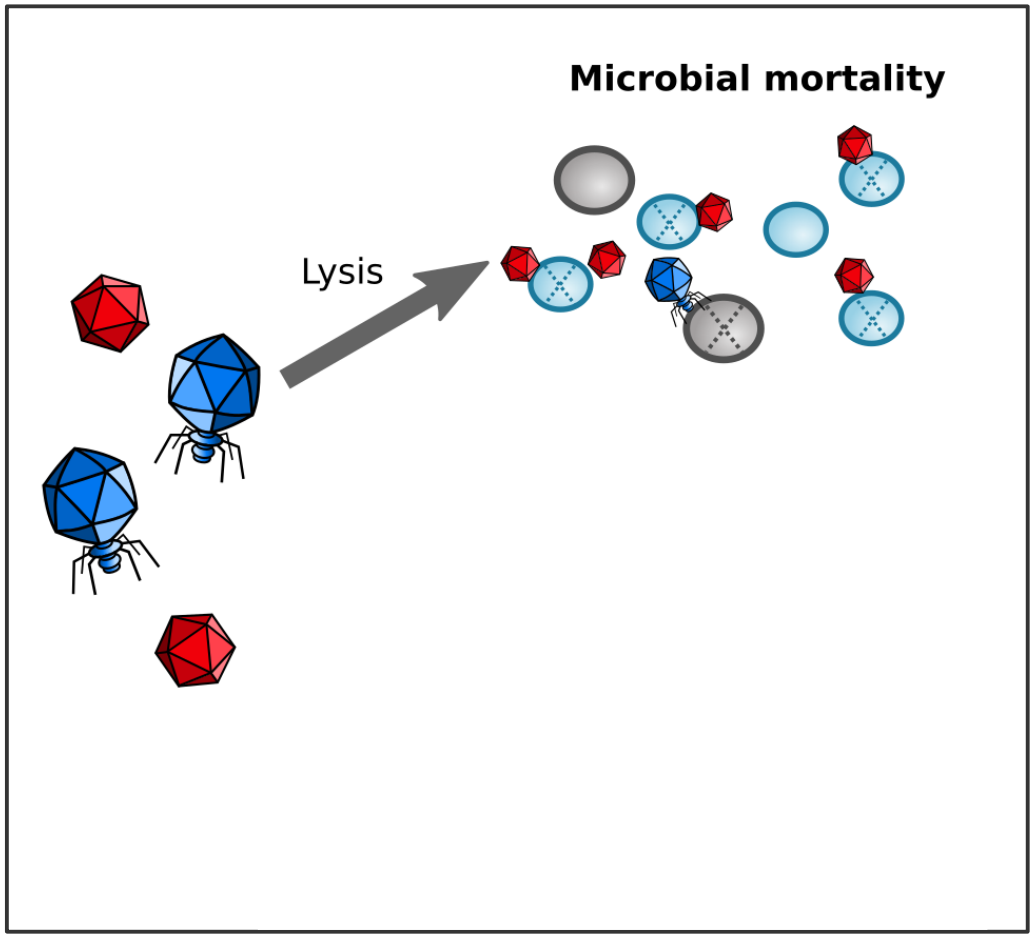


Image: Fuhrman Lab

Viruses of microbes are everywhere



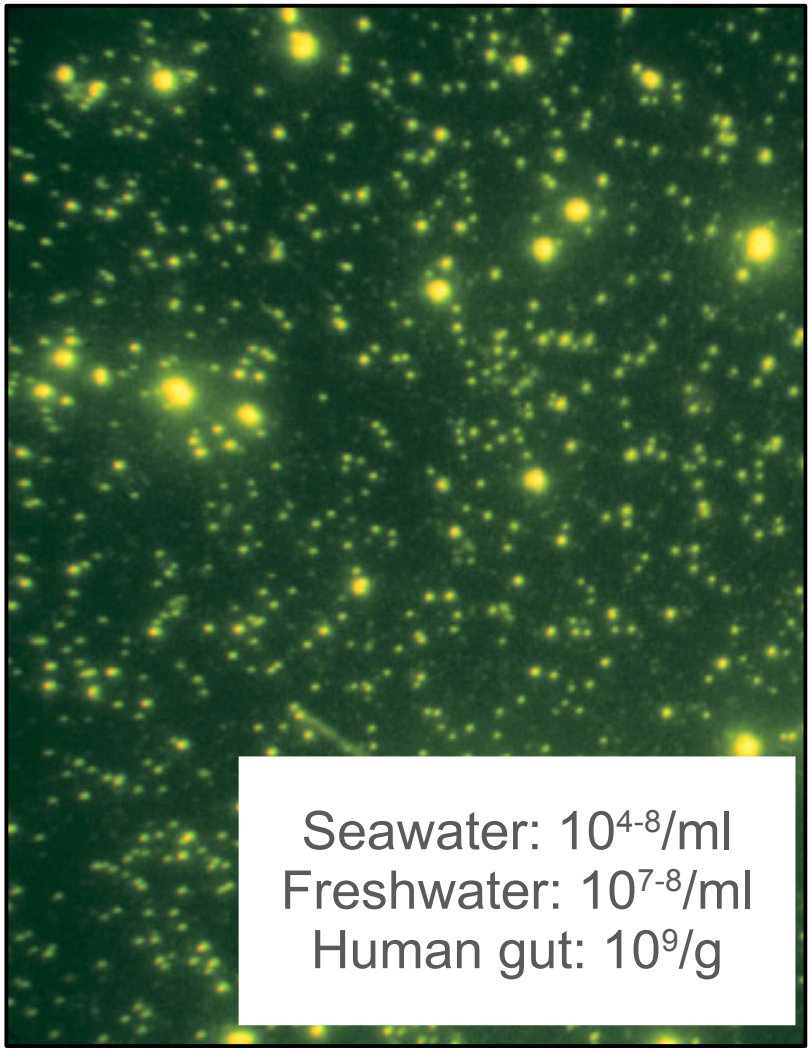
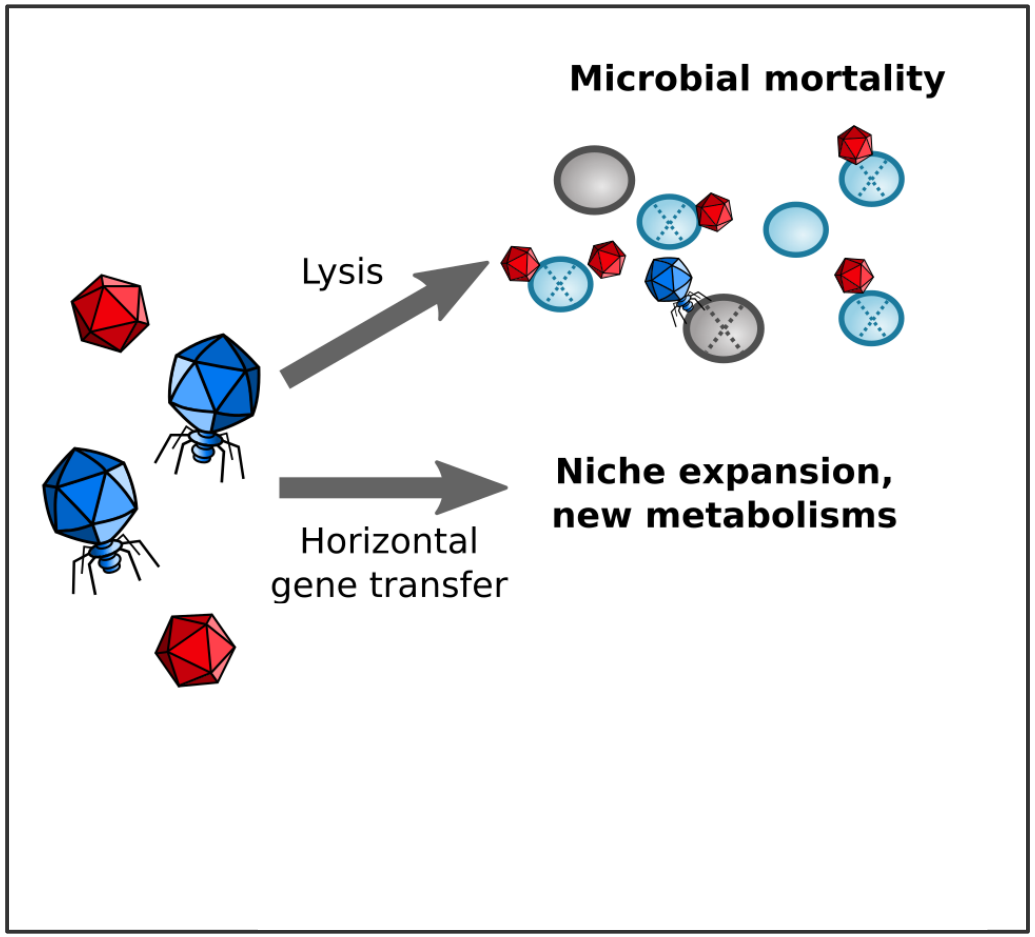


Image: Fuhrman Lab

Viruses of microbes are everywhere

BioSciences



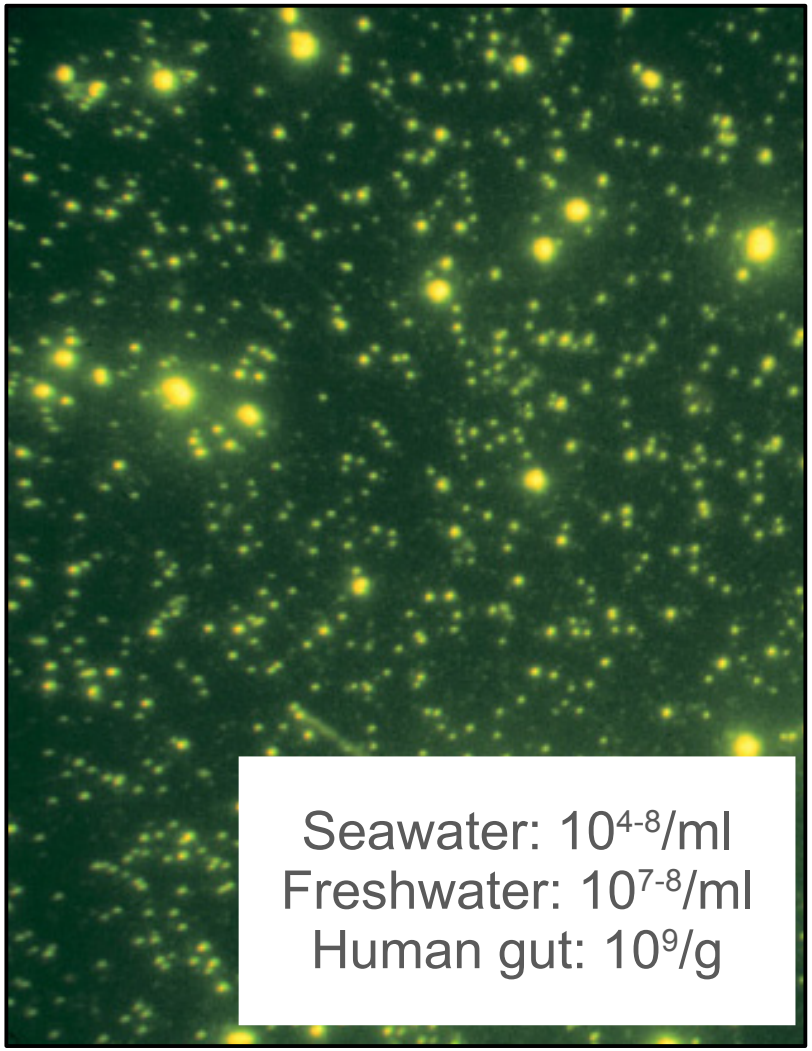
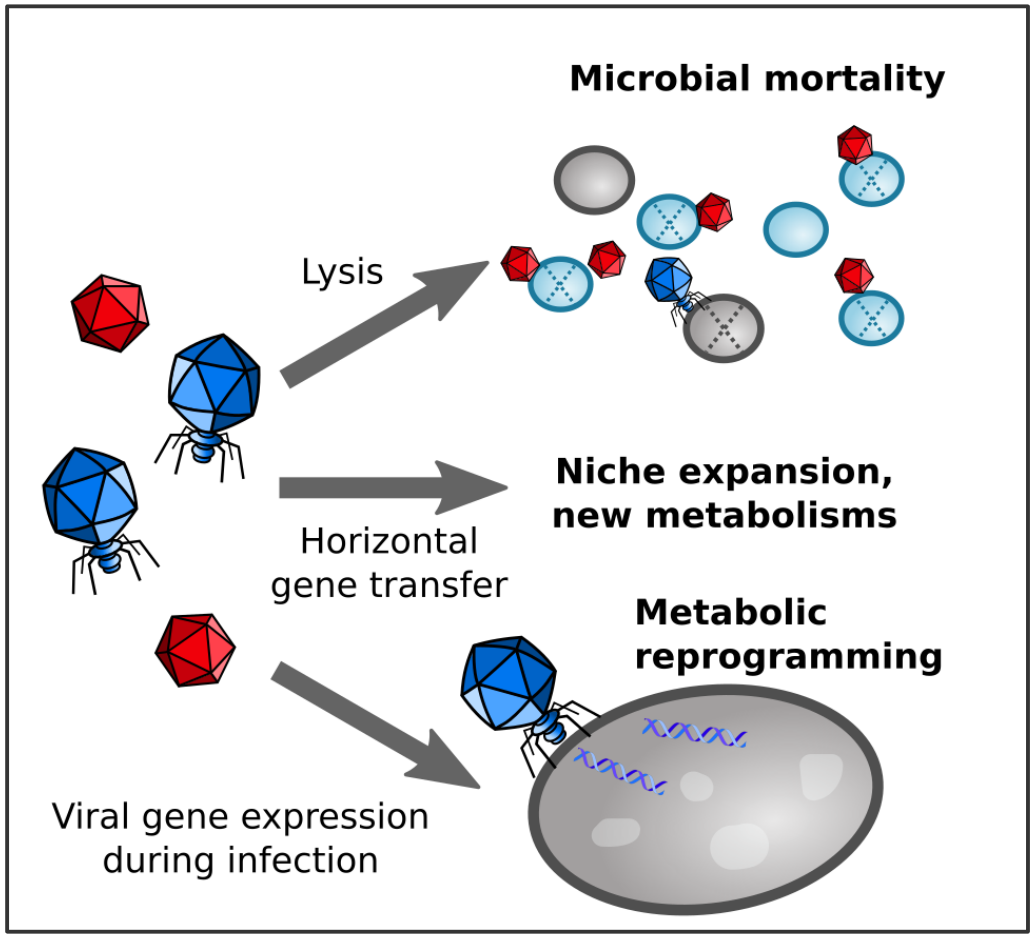


Image: Fuhrman Lab

Viruses of microbes are everywhere



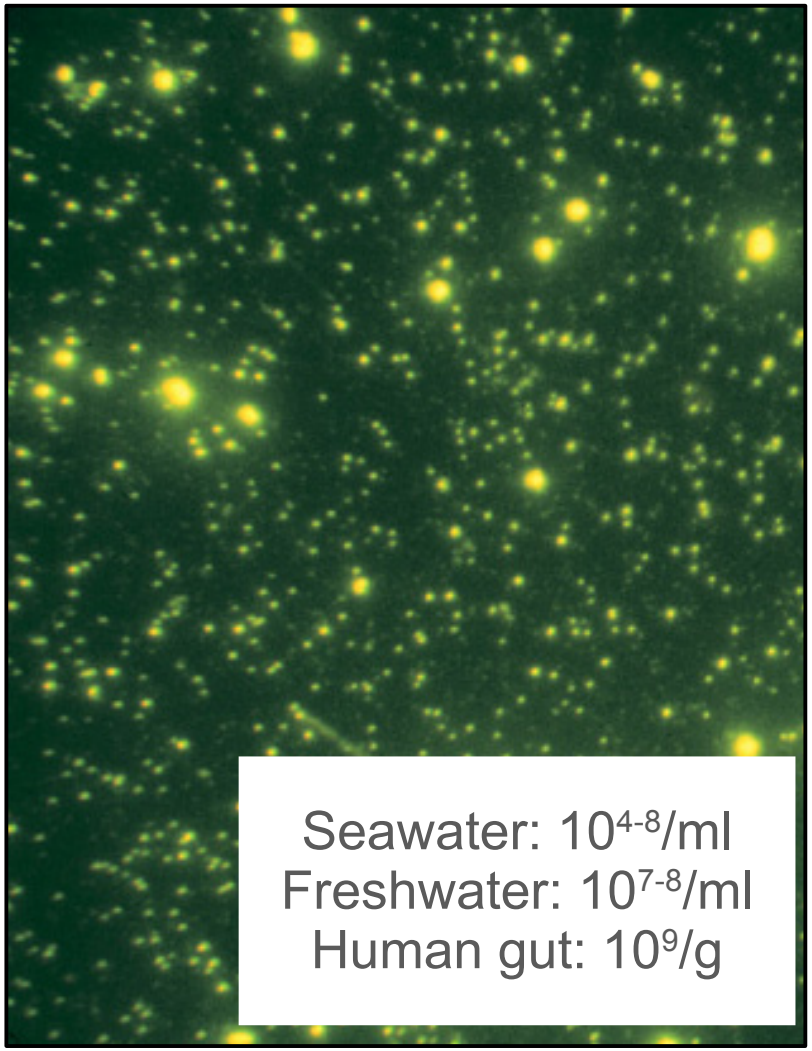
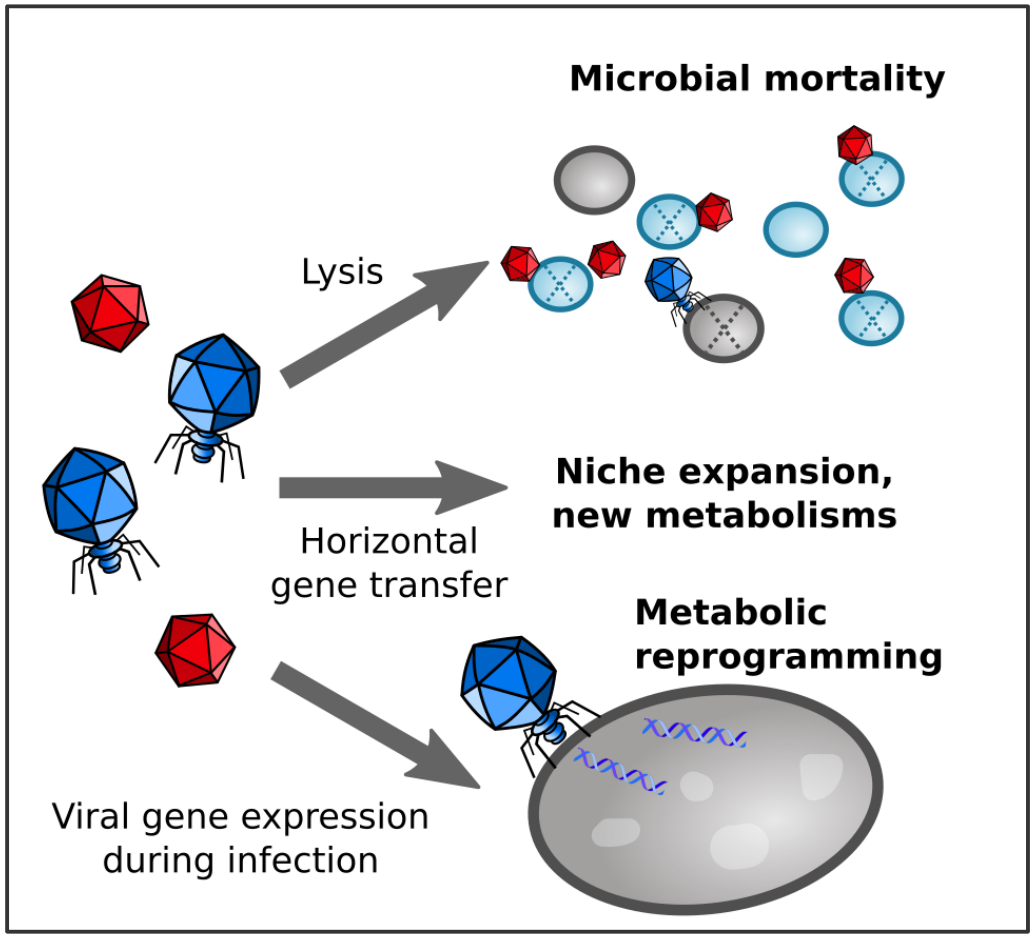


Image: Fuhrman Lab

Challenges and unknowns



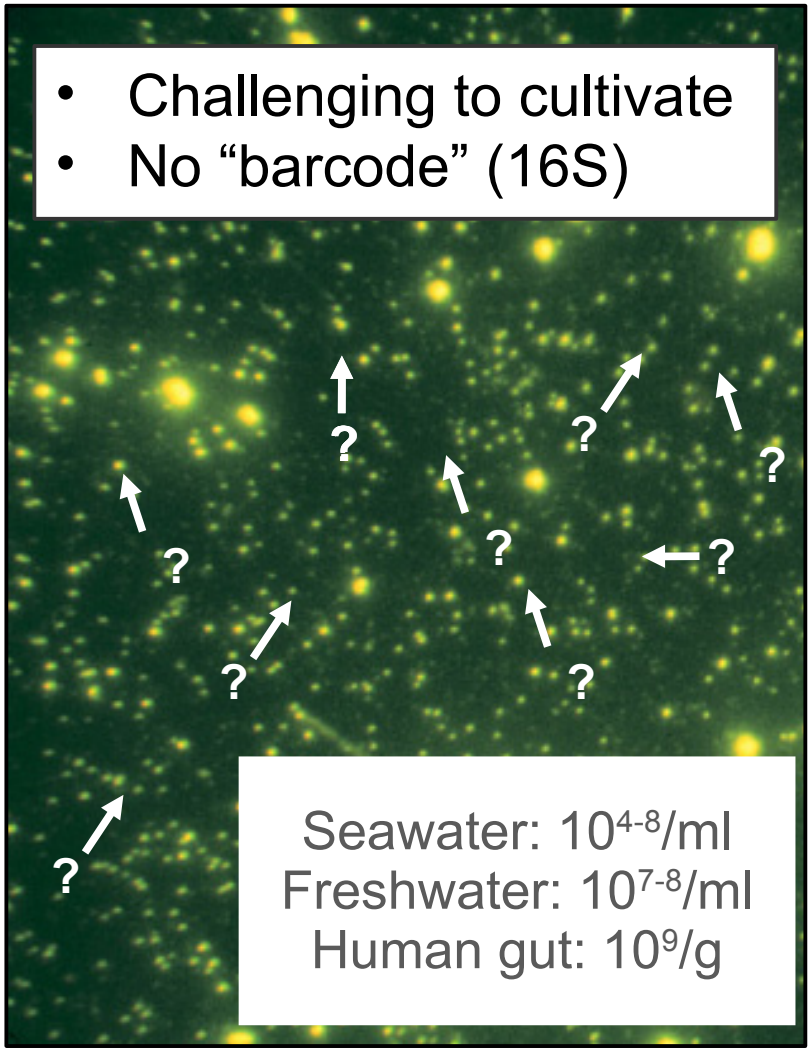
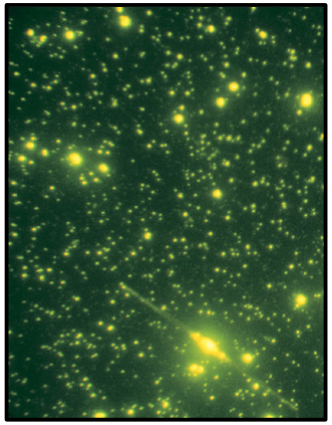
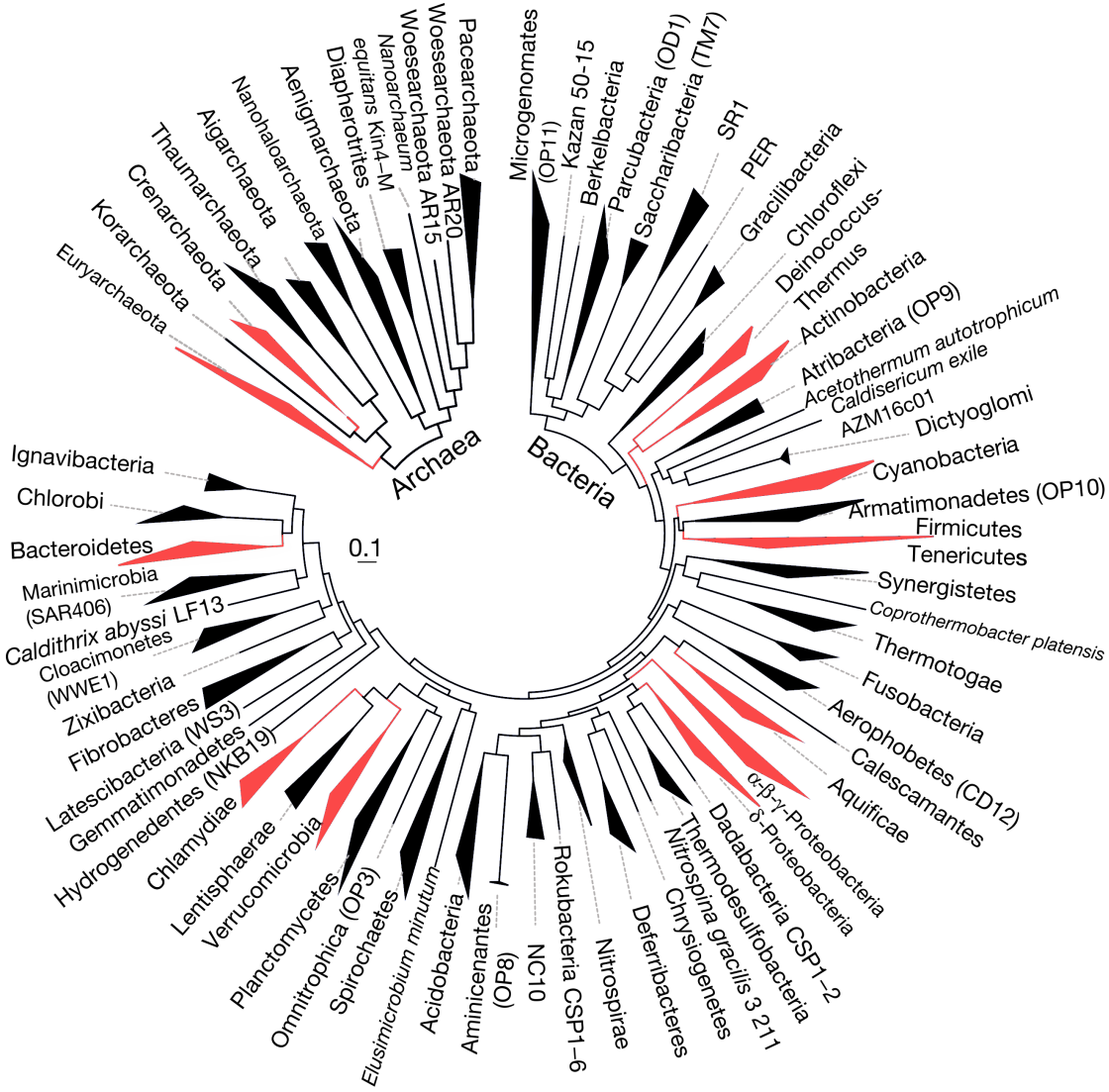


Image: Fuhrman Lab

Challenges of viral ecology

- **Low coverage of viral diversity**

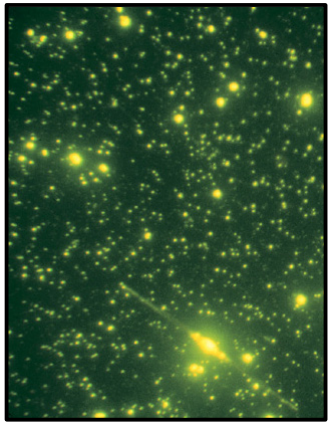


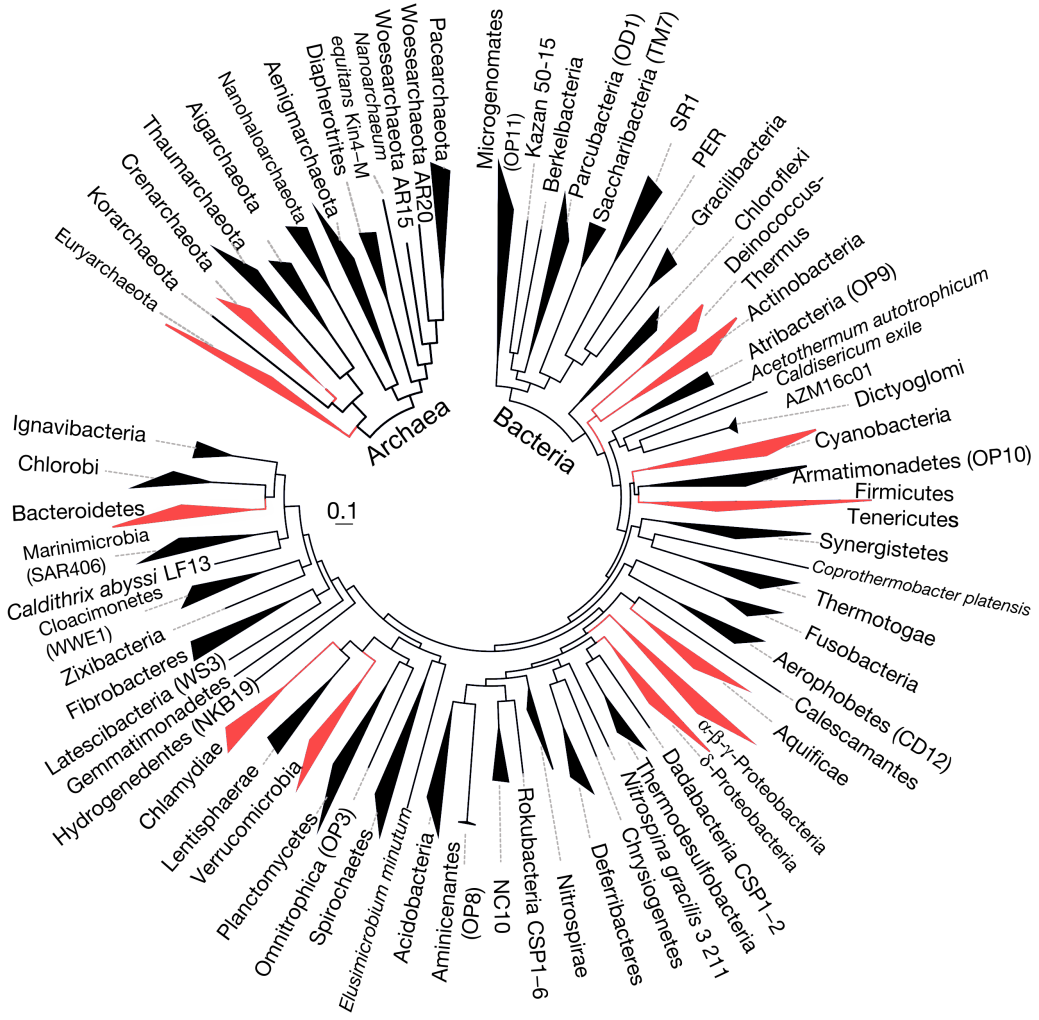


Adapted from
Paez-Espino *et al.*, 2016

Challenges of viral ecology

- Low coverage of viral diversity
- Genomes from uncultivated viruses

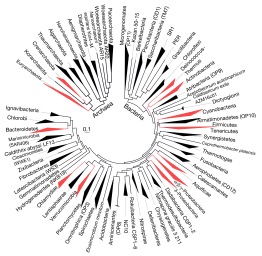


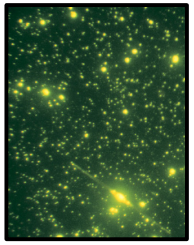


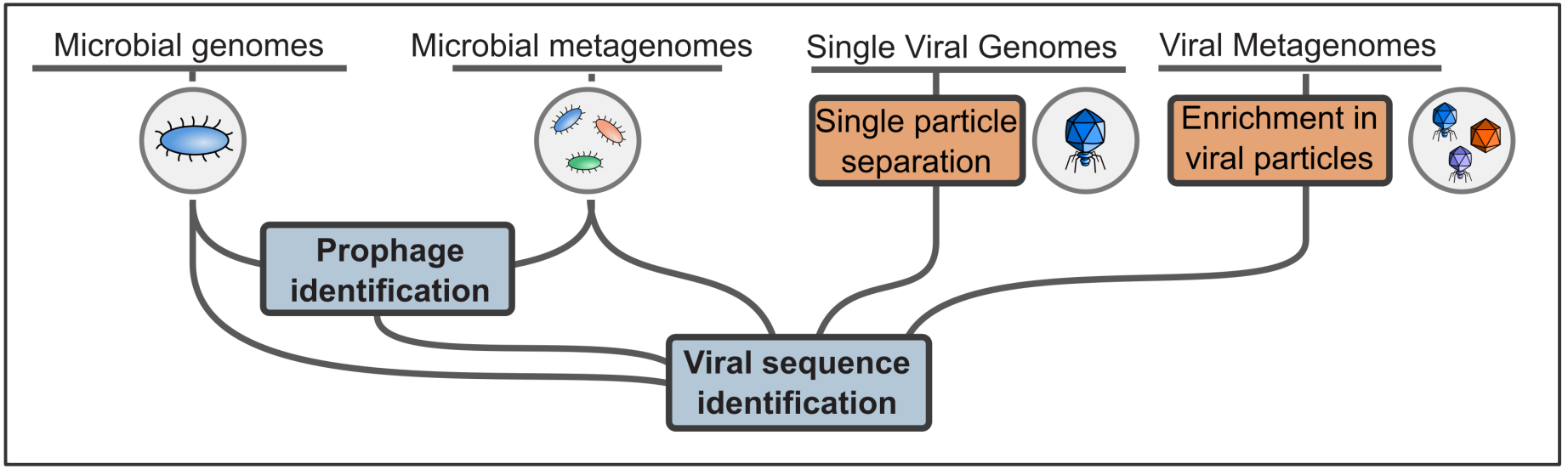
Adapted from
Paez-Espino *et al.*, 2016

Mapping the viral sequence space

- **Low coverage of viral diversity**
- **Genomes from uncultivated viruses**
  - Leverage all types of datasets

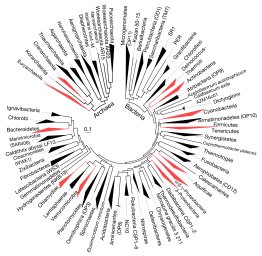


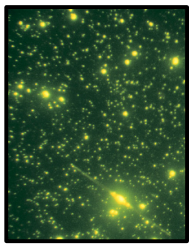


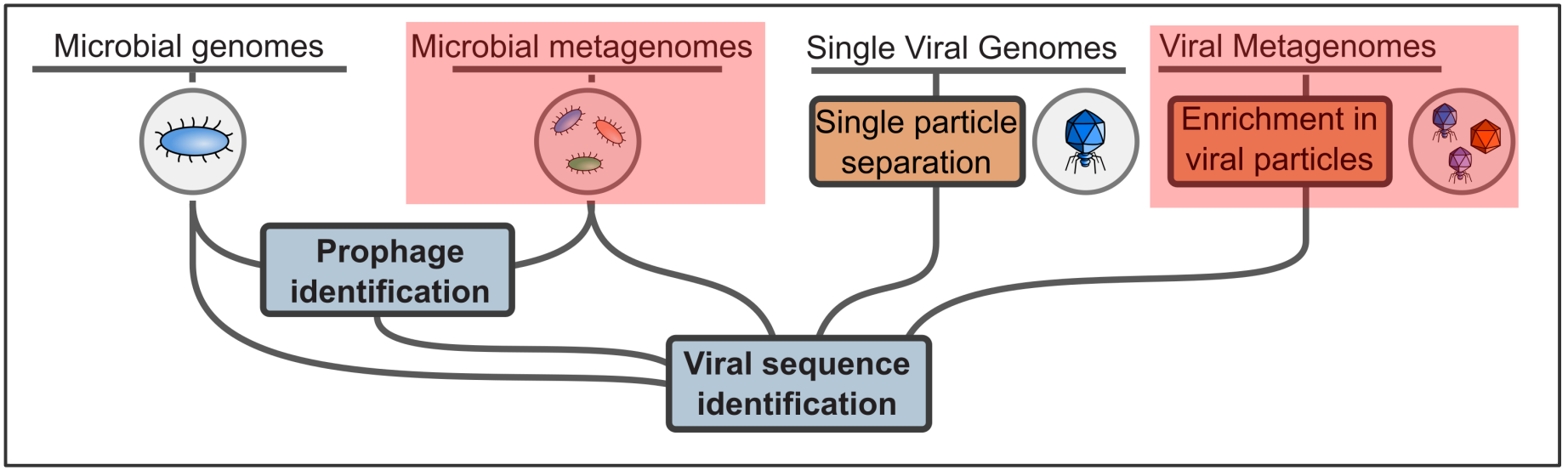


Mapping the viral sequence space

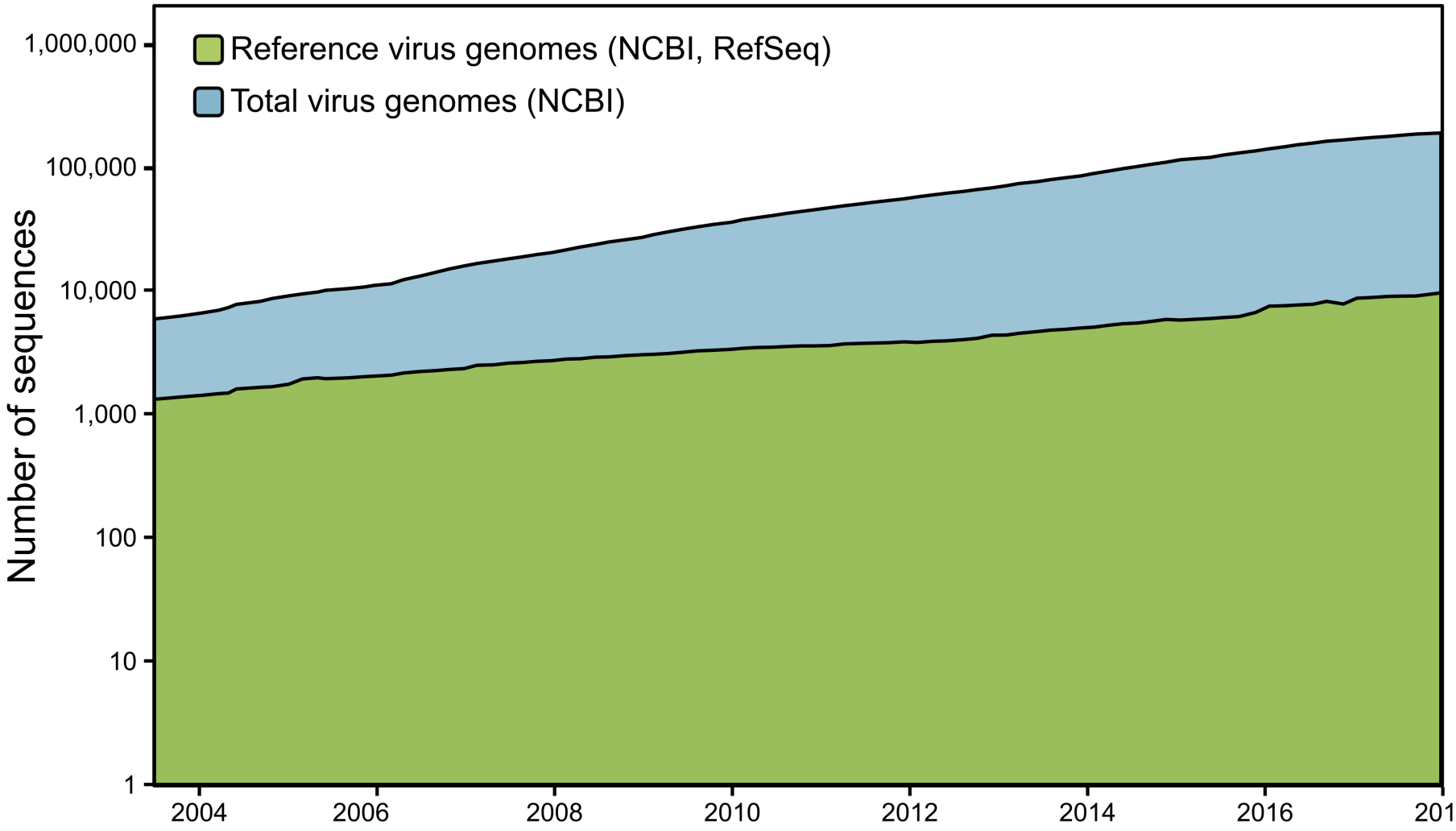
- **Low coverage of viral diversity**
- **Genomes from uncultivated viruses**
  - Leverage all types of datasets
  - Particularly useful: microbial and viral metagenomes

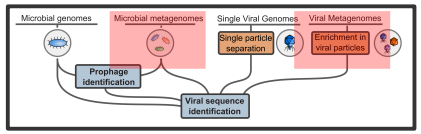




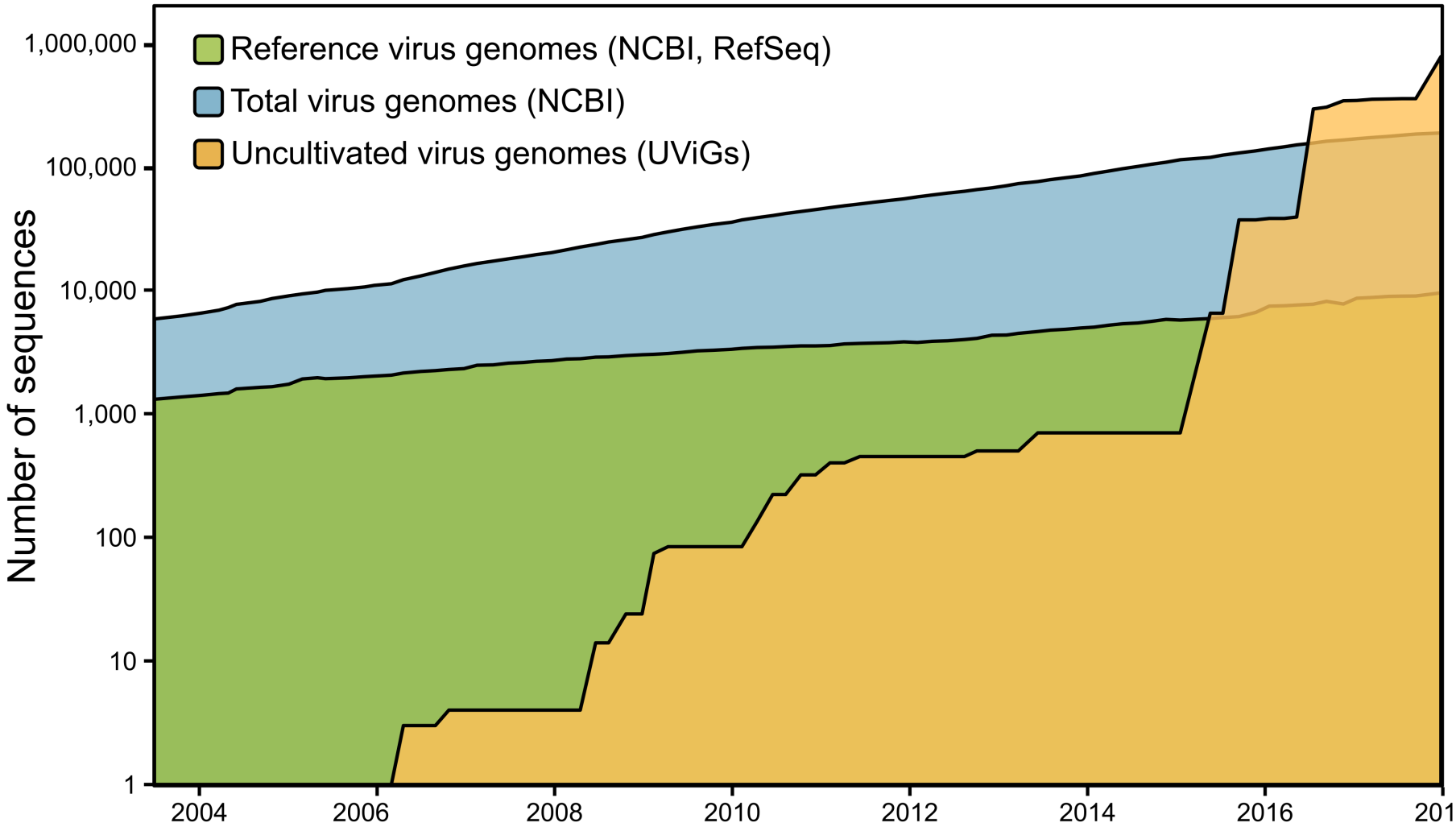


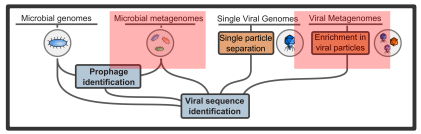
A frantic pace of viral discovery



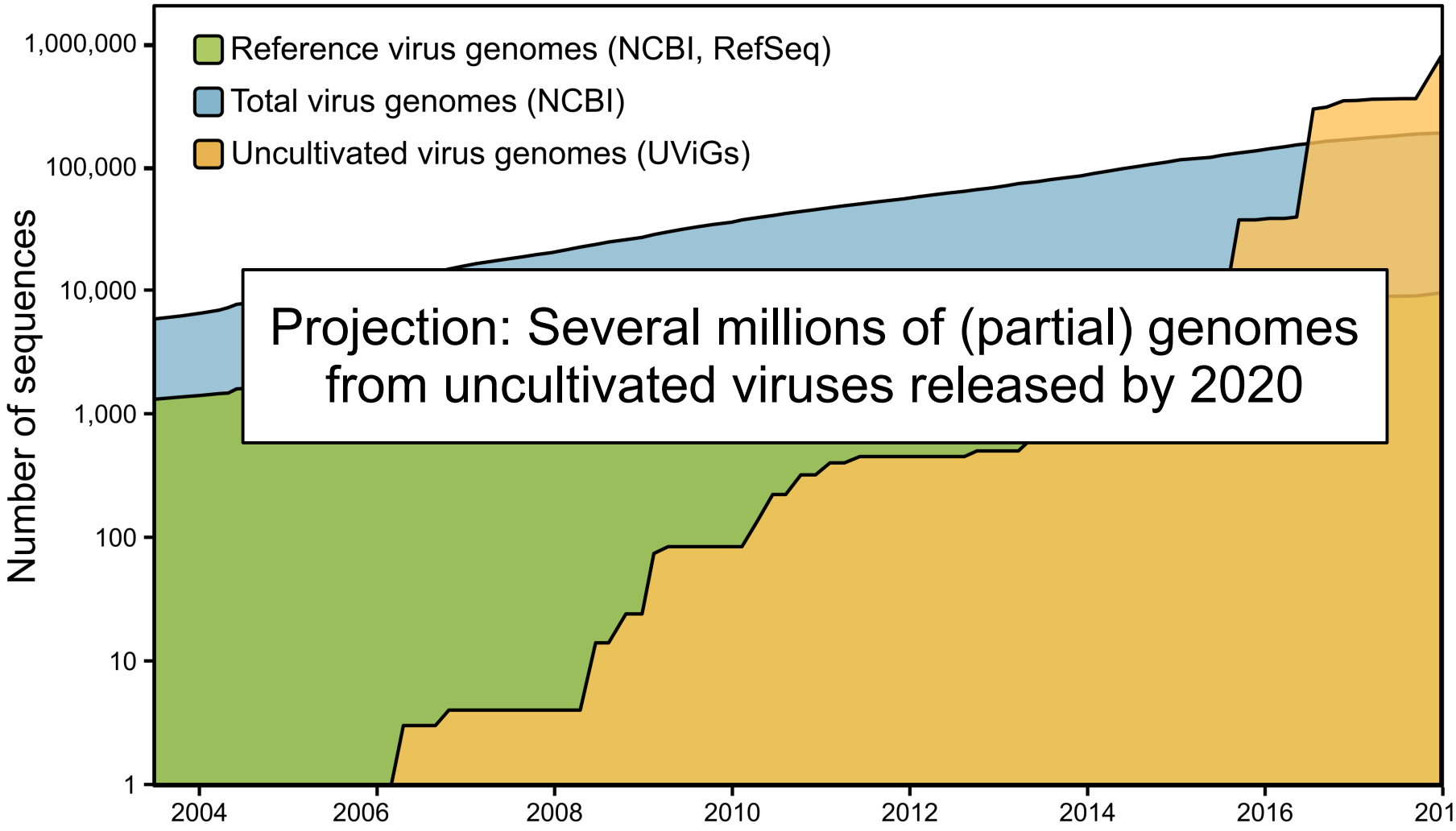


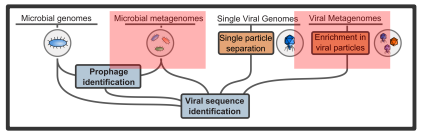
A frantic pace of viral discovery



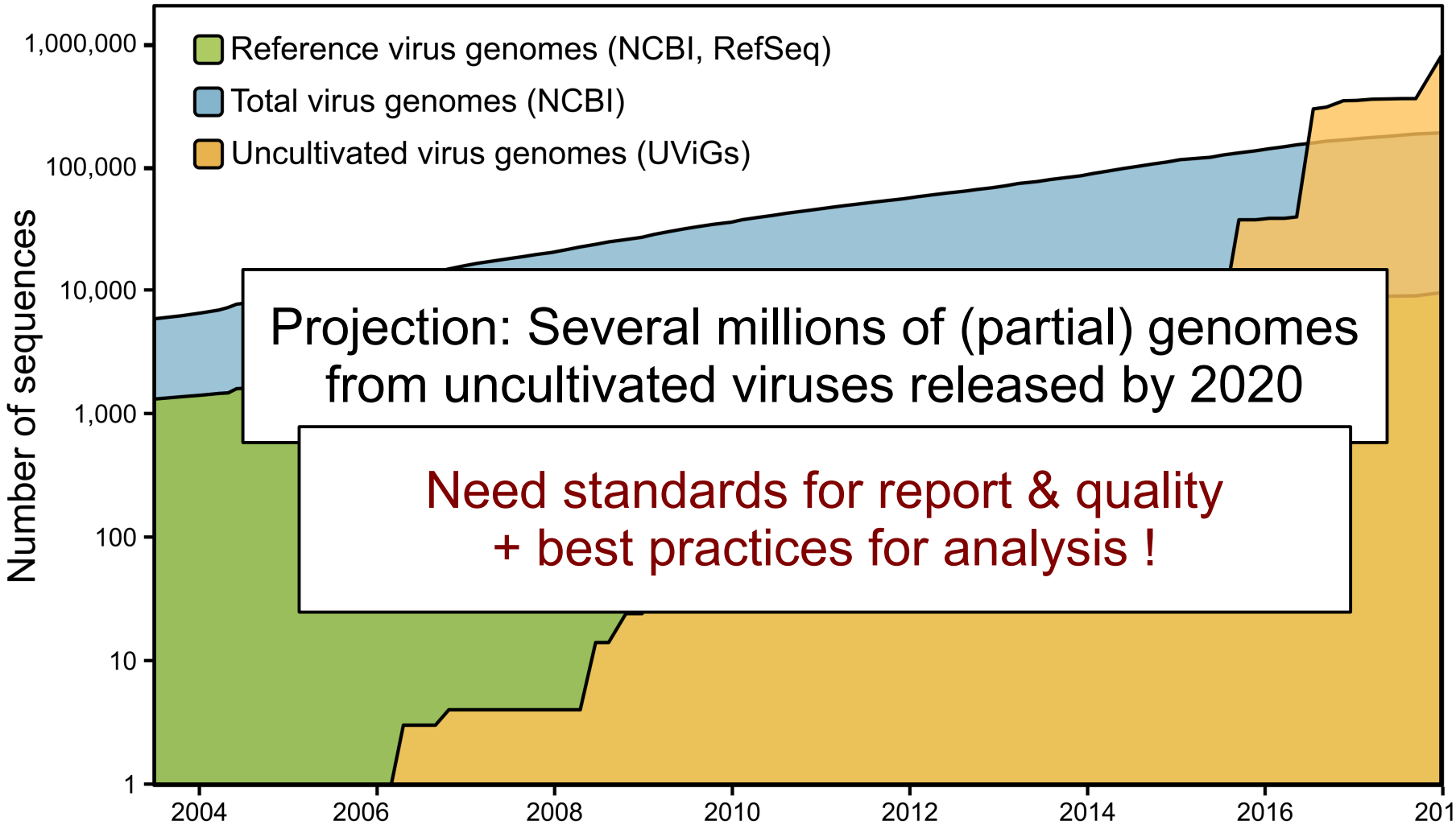


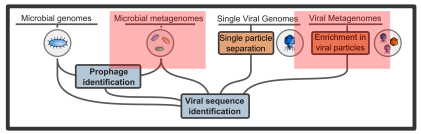
A frantic pace of viral discovery





A frantic pace of viral discovery





Built from MiMAGs & MiSAGs

nature biotechnology

PERSPECTIVE

Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea

Robert M Bowers[1], Nikos C Kyrpides[1], Ramunas Stepanauskas[2], Miranda Harmon-Smith[1], Devin Doud[1], T B K Reddy[1], Frederik Schulz[1], Jessica Jarett[1], Adam R Rivers[1,3], Emiley A Eloe-Fadrosh[1], Susannah G Tringe[1,4], Natalia N Ivanova[1], Alex Copeland[1], Alicia Clum[1], Eric D Becraft[2], Rex R Malmstrom[1], Bruce Birren[5], Mircea Podar[6], Peer Bork[7], George M Weinstock[8], George M Garrity[9], Jeremy A Dodsworth[10], Shibu Yooseph[11], Granger Sutton[12], Frank O Glöckner[13], Jack A Gilbert[14,15], William C Nelson[16], Steven J Hallam[17], Sean P Jungbluth[1,18], Thijs J G Ettema[19], Scott Tighe[20], Konstantinos T Konstantinidis[21], Wen-Tso Liu[22], Brett J Baker[23], Thomas Rattei[24], Jonathan A Eisen[25], Brian Hedlund[26,27], Katherine D McMahon[28,29], Noah Fierer[30,31], Rob Knight[32], Rob Finn[33], Guy Cochrane[33], Ilene Karsch-Mizrachi[34], Gene W Tyson[35], Christian Rinke[35] The Genome Standards Consortium[36], Alla Lapidus[37], Folker Meyer[14], Pelin Yilmaz[13], Donovan H Parks[35], A Murat Eren[38], Lynn Schriml[39], Jillian F Banfield[40], Philip Hugenholtz[35] & Tanja Woyke[1,4]

- **Similar framework:**
  - List approaches
  - Standard pipeline
  - Define quality tiers

Bowers et al., 2017

Built from MiMAGs & MiSAGs

nature biotechnology

PERSPECTIVE

Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea

Robert M Bowers[1], Nikos C Kyrpides[1], Ramunas Stepanauskas[2], Miranda Harmon-Smith[1], Devin Doud[1], T B K Reddy[1], Frederik Schulz[1], Jessica Jarett[1], Adam R Rivers[1,3], Emiley A Eloe-Fadrosh[1], Susannah G Tringe[1,4], Natalia N Ivanova[1], Alex Copeland[1], Alicia Clum[1], Eric D Becraft[2], Rex R Malmstrom[1], Bruce Birren[5], Mircea Podar[6], Peer Bork[7], George M Weinstock[8], George M Garrity[9], Jeremy A Dodsworth[10], Shibu Yooseph[11], Granger Sutton[12], Frank O Glöckner[13], Jack A Gilbert[14,15], William C Nelson[16], Steven J Hallam[17], Sean P Jungbluth[1,18], Thijs J G Ettema[19], Scott Tighe[20], Konstantinos T Konstantinidis[21], Wen-Tso Liu[22], Brett J Baker[23], Thomas Rattei[24], Jonathan A Eisen[25], Brian Hedlund[26,27], Katherine D McMahon[28,29], Noah Fierer[30,31], Rob Knight[32], Rob Finn[33], Guy Cochrane[33], Ilene Karsch-Mizrachi[34], Gene W Tyson[35], Christian Rinke[35] The Genome Standards Consortium[36], Alla Lapidus[37], Folker Meyer[14], Pelin Yilmaz[13], Donovan H Parks[35], A Murat Eren[38], Lynn Schriml[39], Jillian F Banfield[40], Philip Hugenholtz[35] & Tanja Woyke[1,4]

- **Similar framework:**
  - List approaches
  - Standard pipeline
  - Define quality tiers

- **Viral-specific features:**
  - Identification of viral vs non-viral sequences
  - Taxonomic classification (“16S OTU” for viral ecology ?)
  - Host prediction

Built from MiMAGs & MiSAGs

nature biotechnology

PERSPECTIVE

Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea

Robert M Bowers[1], Nikos C Kyrpides[1], Ramunas Stepanauskas[2], Miranda Harmon-Smith[1], Devin Doud[1], T B K Reddy[1], Frederik Schulz[1], Jessica Jarett[1], Adam R Rivers[1,3], Emiley A Eloe-Fadrosh[1], Susannah G Tringe[1,4], Natalia N Ivanova[1], Alex Copeland[1], Alicia Clum[1], Eric D Becraft[2], Rex R Malmstrom[1], Bruce Birren[5], Mircea Podar[6], Peer Bork[7], George M Weinstock[8], George M Garrity[9], Jeremy A Dodsworth[10], Shibu Yooseph[11], Granger Sutton[12], Frank O Glöckner[13], Jack A Gilbert[14,15], William C Nelson[16], Steven J Hallam[17], Sean P Jungbluth[1,18], Thijs J G Ettema[19], Scott Tighe[20], Konstantinos T Konstantinidis[21], Wen-Tso Liu[22], Brett J Baker[23], Thomas Rattei[24], Jonathan A Eisen[25], Brian Hedlund[26,27], Katherine D McMahon[28,29], Noah Fierer[30,31], Rob Knight[32], Rob Finn[33], Guy Cochrane[33], Ilene Karsch-Mizrachi[34], Gene W Tyson[35], Christian Rinke[35] The Genome Standards Consortium[36], Alla Lapidus[37], Folker Meyer[14], Pelin Yilmaz[13], Donovan H Parks[35], A Murat Eren[38], Lynn Schriml[39], Jillian F Banfield[40], Philip Hugenholtz[35] & Tanja Woyke[1,4]

- **Similar framework:**
  - List approaches
  - Standard pipeline
  - Define quality tiers

- **Viral-specific features:**
  - Identification of viral vs non-viral sequences
  - Taxonomic classification (“16S OTU” for viral ecology ?)
  - Host prediction
- **Include “best practices” and benchmarks**

Built from MiMAGs & MiSAGs

nature biotechnology

PERSPECTIVE

Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea

Robert M Bowers[1], Nikos C Kyrpides[1], Ramunas Stepanauskas[2], Miranda Harmon-Smith[1], Devin Doud[1], T B K Reddy[1], Frederik Schulz[1], Jessica Jarett[1], Adam R Rivers[1,3], Emiley A Eloe-Fadrosh[1], Susannah G Tringe[1,4], Natalia N Ivanova[1], Alex Copeland[1], Alicia Clum[1], Eric D Becraft[2], Rex R Malmstrom[1], Bruce Birren[5], Mircea Podar[6], Peer Bork[7], George M Weinstock[8], George M Garrity[9], Jeremy A Dodsworth[10], Shibu Yooseph[11], Granger Sutton[12], Frank O Glöckner[13], Jack A Gilbert[14,15], William C Nelson[16], Steven J Hallam[17], Sean P Jungbluth[1,18], Thijs J G Ettema[19], Scott Tighe[20], Konstantinos T Konstantinidis[21], Wen-Tso Liu[22], Brett J Baker[23], Thomas Rattei[24], Jonathan A Eisen[25], Brian Hedlund[26,27], Katherine D McMahon[28,29], Noah Fierer[30,31], Rob Knight[32], Rob Finn[33], Guy Cochrane[33], Ilene Karsch-Mizrachi[34], Gene W Tyson[35], Christian Rinke[35] The Genome Standards Consortium[36], Alla Lapidus[37], Folker Meyer[14], Pelin Yilmaz[13], Donovan H Parks[35], A Murat Eren[38], Lynn Schriml[39], Jillian F Banfield[40], Philip Hugenholtz[35] & Tanja Woyke[1,4]

- **Similar framework:**
  - List approaches
  - Standard pipeline
  - Define quality tiers

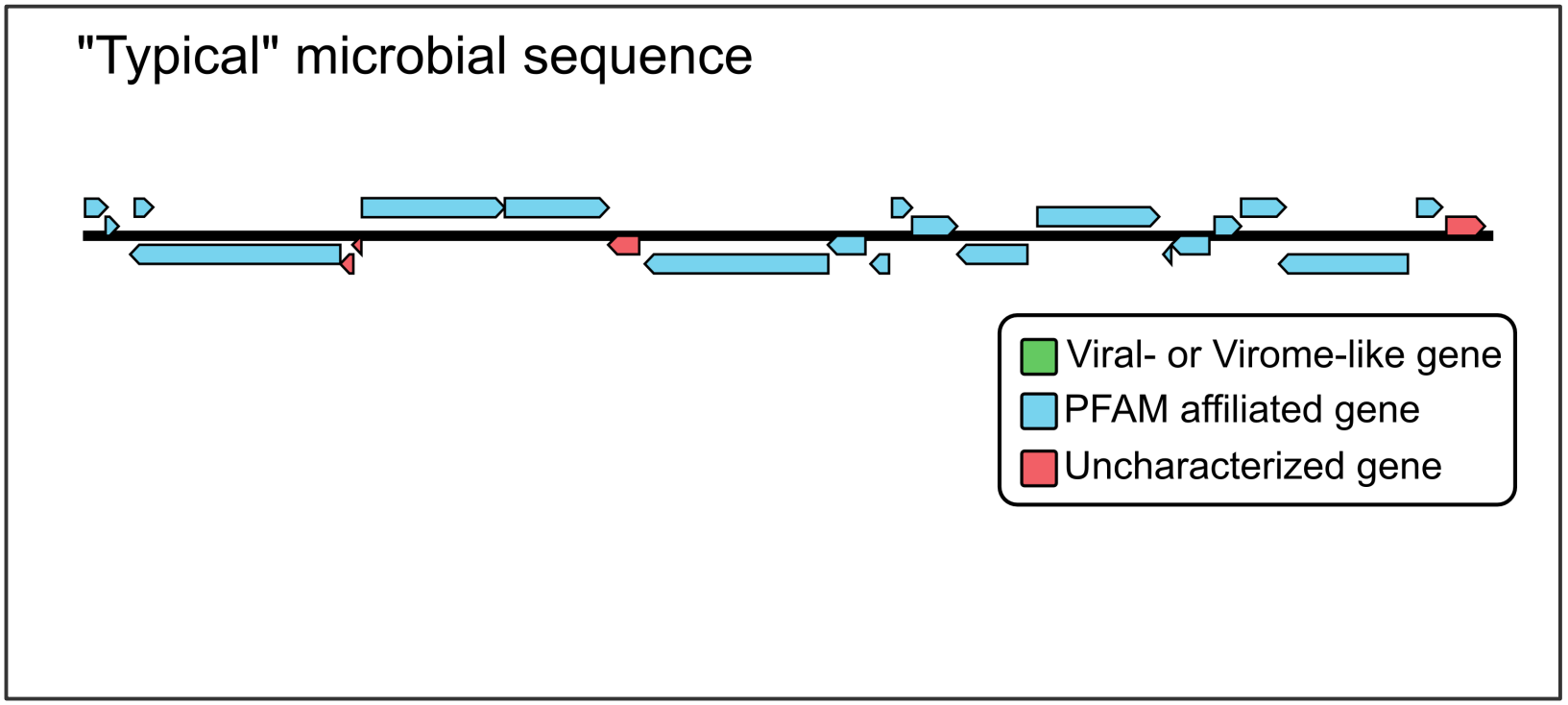
- **Viral-specific features:**
  - Identification of viral vs non-viral sequences
  - Taxonomic classification ("16S OTU" for viral ecology ?)
  - Host prediction
- **Include "best practices" and benchmarks**
- **Currently under review @ Nature Biotechnology**

Viral contigs / prophage identification

- **Established tools:**
  - Online (e.g. PHASTER, VirSorter), protocols (Earth's virome)
  - Single underlying principle:
    - "viral signature gene(s) in a viral genome-like context"

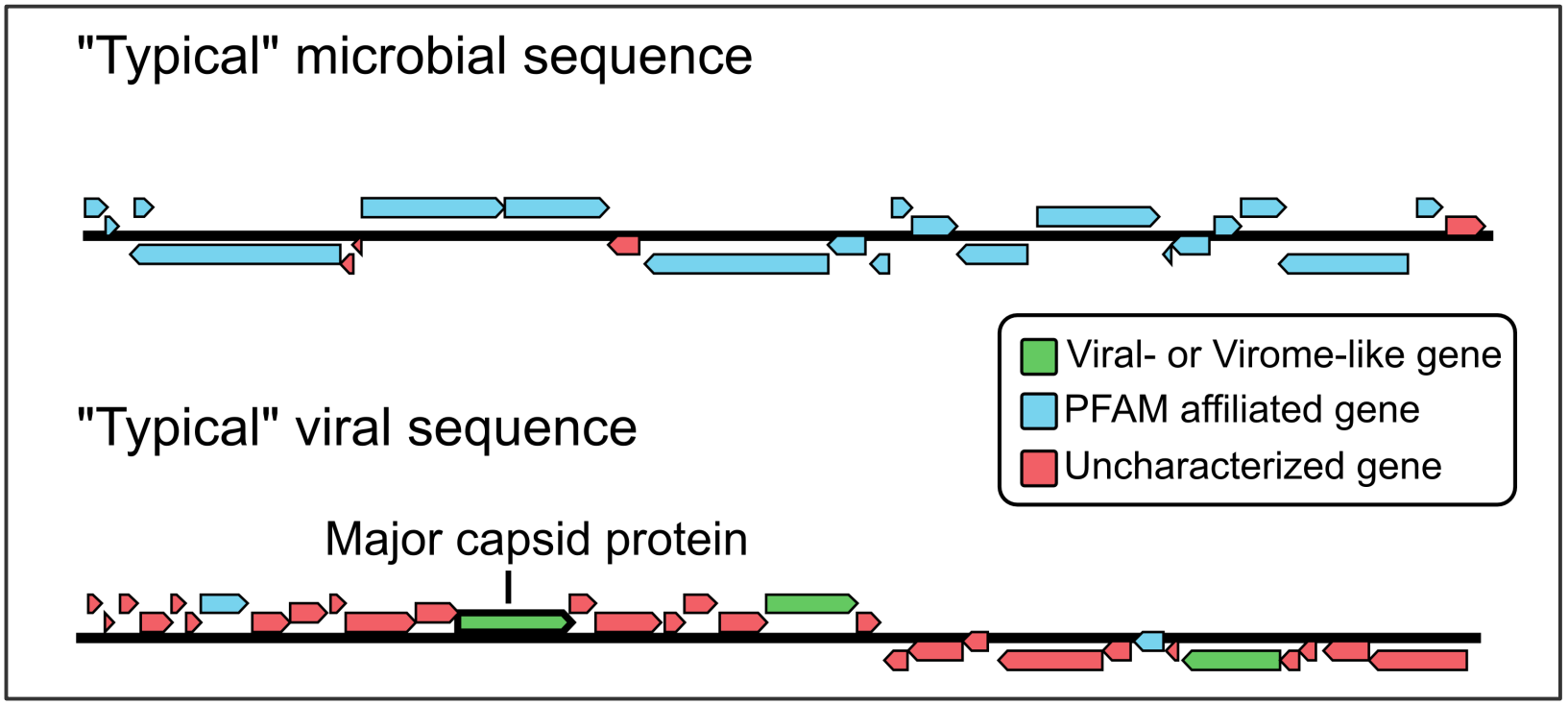
Viral contigs / prophage identification

- **Established tools:**
  - Online (e.g. PHASTER, VirSorter), protocols (Earth's virome)
  - Single underlying principle:
    - "viral signature gene(s) in a viral genome-like context"



Viral contigs / prophage identification

- **Established tools:**
  - Online (e.g. PHASTER, VirSorter), protocols (Earth's virome)
  - Single underlying principle:
    - "viral signature gene(s) in a viral genome-like context"



Viral contigs / prophage identification

- **Established tools:**
  - Online (e.g. PHASTER, VirSorter), protocols (Earth's virome)
  - Single underlying principle:
    - "viral signature gene(s) in a viral genome-like context"

- **Recent developments:**
  - K-mer approaches for viral contig identification (e.g. VirFinder, PhaMers)
  - No large-scale evaluation yet

Viral contigs / prophage identification

- **Established tools:**
  - Online (e.g. PHASTER, VirSorter), protocols (Earth's virome)
  - Single underlying principle:
    - "viral signature gene(s) in a viral genome-like context"

- **Recent developments:**
  - K-mer approaches for viral contig identification (e.g. VirFinder, PhaMers)
  - No large-scale evaluation yet

- **Requirements & recommendations:**
  - Viral contigs identification required even for viral-targeted datasets
    - methods to be reported when releasing UViGs

Viral contigs / prophage identification

- **Established tools:**
  - Online (e.g. PHASTER, VirSorter), protocols (Earth's virome)
  - Single underlying principle:
    - "viral signature gene(s) in a viral genome-like context"

- **Recent developments:**
  - K-mer approaches for viral contig identification (e.g. VirFinder, PhaMers)
  - No large-scale evaluation yet

- **Requirements & recommendations:**
  - Viral contigs identification required even for viral-targeted datasets
    - methods to be reported when releasing UViGs
  - Integrated viruses boundaries challenging to define:
    - flagged when releasing UViGs (i.e. "was part of a larger host contig")

Classification of novel UViGs

- **Taxonomic classification:**
  - Handled by ICTV

Classification of novel UViGs

- **Taxonomic classification:**
  - Handled by ICTV
  - Rules defined group by group
    - currently no universal classifier
    - still very dynamic

Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV)

Elliot J Lefkowitz ✉, Donald M Dempsey, Robert Curtis Hendrickson, Richard J Orton, Stuart G Siddell, Donald B Smith

Nucleic Acids Research, Volume 46, Issue D1, 4 January 2018, Pages D708–D717, https://doi.org/10.1093/nar/gkx932

Published: 13 October 2017 **Article history ▾**

Archives of Virology

April 2016, Volume 161, Issue 4, pp 1095–1099 | Cite as

Taxonomy of prokaryotic viruses: update from the ICTV bacterial and archaeal viruses subcommittee

Authors Authors and affiliations

Mart Krupovic, Bas E. Dutilh, Evelien M. Adriaenssens, Johannes Wittmann, Finn K. Vogensen, Mathew B. Sullivan, Janis Rumnieks, David Prangishvili, Rob Lavigne, Andrew M. Kropinski ✉, Jochen Klumpp, Annika Gillis, Francois E Rob A. Edwards, Siobain Duffy, show 4 more

Archives of Virology

April 2017, Volume 162, Issue 4, pp 1153–1157 | Cite as

Taxonomy of prokaryotic viruses: 2016 update from the ICTV bacterial and archaeal viruses subcommittee

Authors Authors and affiliations

Evelien M. Adriaenssens, Mart Krupovic, Petar Knezevic, Hans-Wolfgang Ackermann, Jakub Barylski, J. Rodney Brister, Martha R. C. Clokie, Siobain Duffy, Bas E. Dutilh, Robert A. Edwards, Francois Enault, Ho Bin Jang, Jochen Klumpp, Andrew M. Kropinski ✉, Rob Lavigne, show 8 more

Classification of novel UViGs

- **Taxonomic classification:**
  - Handled by ICTV
  - Rules defined group by group
    - currently no universal classifier
    - still very dynamic
    - e.g. currently > 200 rules for species

Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV)
Elliot J Lefkowitz ✉, Donald M Dempsey, Robert Curtis Hendrickson, Richard J Orton, Stuart G Siddell, Donald B Smith
Nucleic Acids Research, Volume 46, Issue D1, 4 January 2018, Pages D708–D717, https://doi.org/10.1093/nar/gkx932
Published: 13 October 2017 **Article history**

Virus group	Species demarcation criteria
Parvoviridae	Members of each species are antigenically distinct, as assessed by neutralization using polyclonal antisera, and natural infection is usually confined to a single host species. Generally, species are <95% related by non-structural gene DNA sequence.
P22-like viruses	Lack of DNA homology between species.
Coltivirus	RNA cross-hybridization assays: within a single species, RNA sequence that exhibit more than 74% similarity will hybridize at 36°C below the Tm of the fully base-paired duplex. Sequence analysis: Nucleotide identity of >89% in the conserved Seg12; amino acid identities of >55%, >57% and >60% respectively in VP6, VP7 and VP12 (the most variable proteins).

Classification of novel UViGs

- **Taxonomic classification:**
  - Handled by ICTV
  - Rules defined group by group
    - currently no universal classifier
    - still very dynamic
    - e.g. currently > 200 rules for species

Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV)

Elliot J Lefkowitz ✉, Donald M Dempsey, Robert Curtis Hendrickson, Richard J Orton, Stuart G Siddell, Donald B Smith

Nucleic Acids Research, Volume 46, Issue D1, 4 January 2018, Pages D708–D717, https://doi.org/10.1093/nar/gkx932

Published: 13 October 2017 **Article history**

- **Operational "species rank" grouping for ecological studies (vOTU)**
  - Consensus: pairwise genome comparison (nucleotide level)
  - Cutoffs on identity % and alignment fraction
  - Benchmark from NCBI RefSeq and IMG/VR → 95% ANI – 85% AF

Classification of novel UViGs

- **Taxonomic classification:**
  - Handled by ICTV
  - Rules defined group by group
    - currently no universal classifier
    - still very dynamic
    - e.g. currently > 200 rules for species

Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV)

Elliot J Lefkowitz ✉, Donald M Dempsey, Robert Curtis Hendrickson, Richard J Orton, Stuart G Siddell, Donald B Smith

Nucleic Acids Research, Volume 46, Issue D1, 4 January 2018, Pages D708–D717, https://doi.org/10.1093/nar/gkx932

Published: 13 October 2017 **Article history**

- **Operational "species rank" grouping for ecological studies (vOTU)**
  - Consensus: pairwise genome comparison (nucleotide level)
  - Cutoffs on identity % and alignment fraction
  - Benchmark from NCBI RefSeq and IMG/VR → 95% ANI – 85% AF

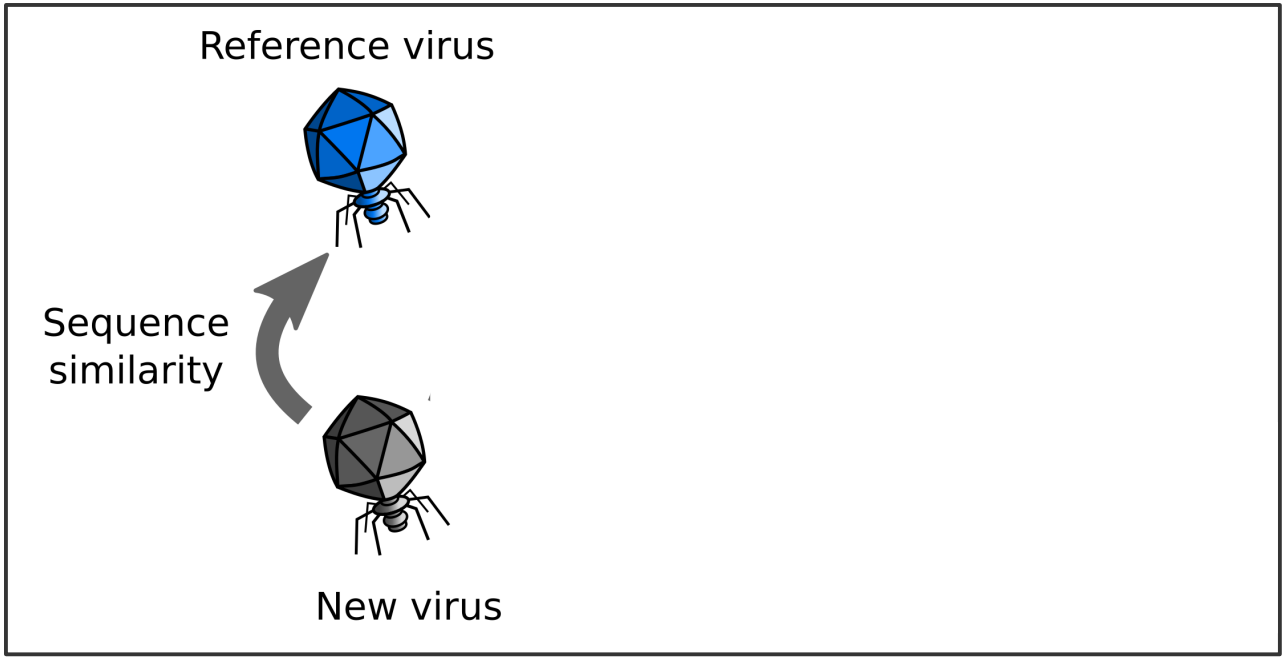
- **Recommendations**
  - For ecological studies: vOTU from pairwise ANI/AF
  - For "true" taxonomic classification: group-by-group basis (not automated)

In silico host assignment

- **4 main types of approaches**

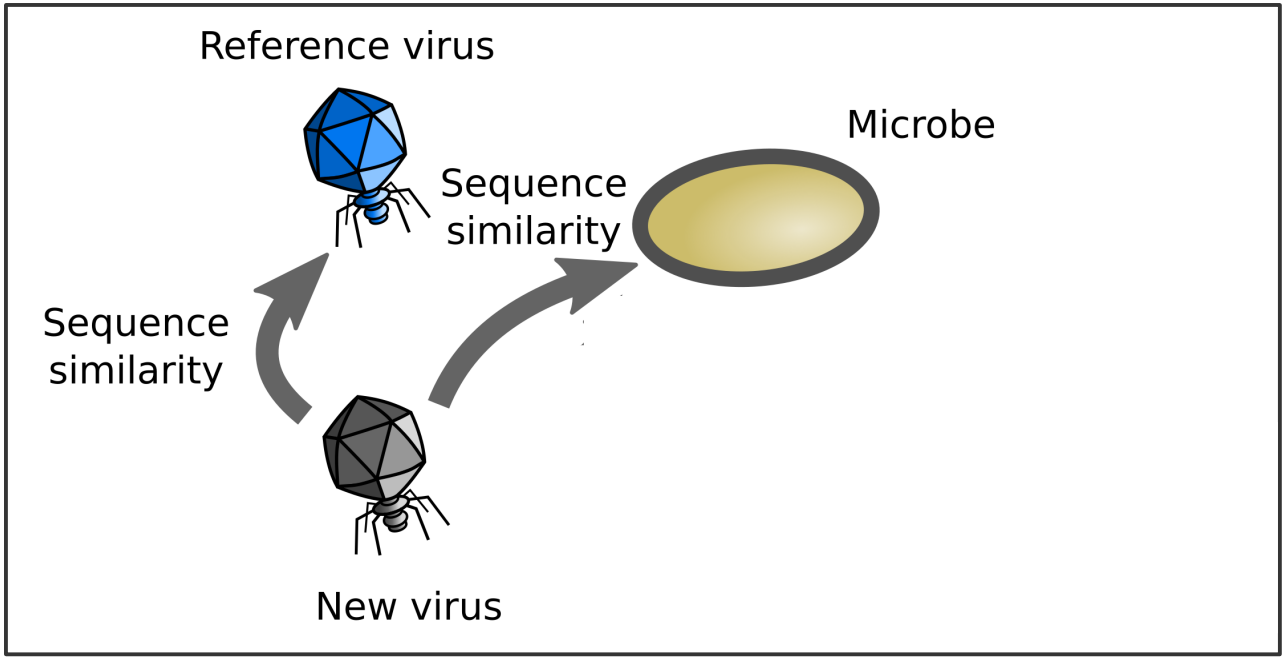
In silico host assignment

- **4 main types of approaches**
  - Sequence homology with isolate viruses



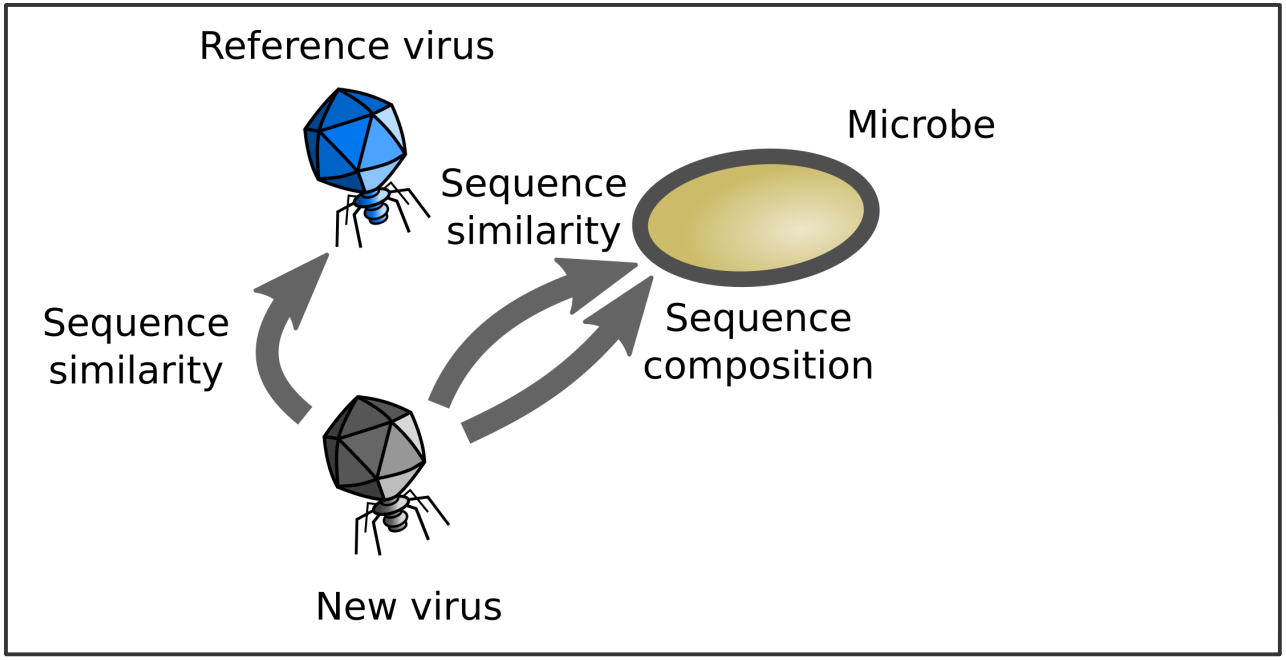
In silico host assignment

- **4 main types of approaches**
  - Sequence homology with isolate viruses
  - Sequence homology with host genome



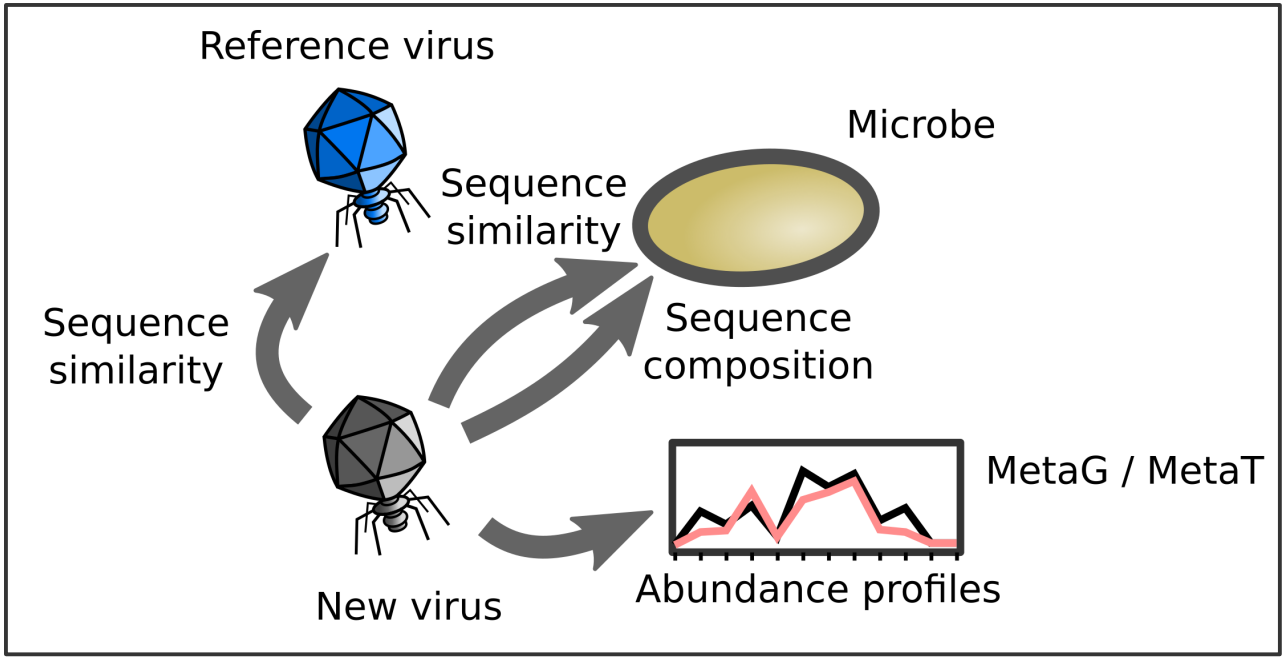
In silico host assignment

- **4 main types of approaches**
  - Sequence homology with isolate viruses
  - Sequence homology with host genome
  - Sequence composition with host genome



In silico host assignment

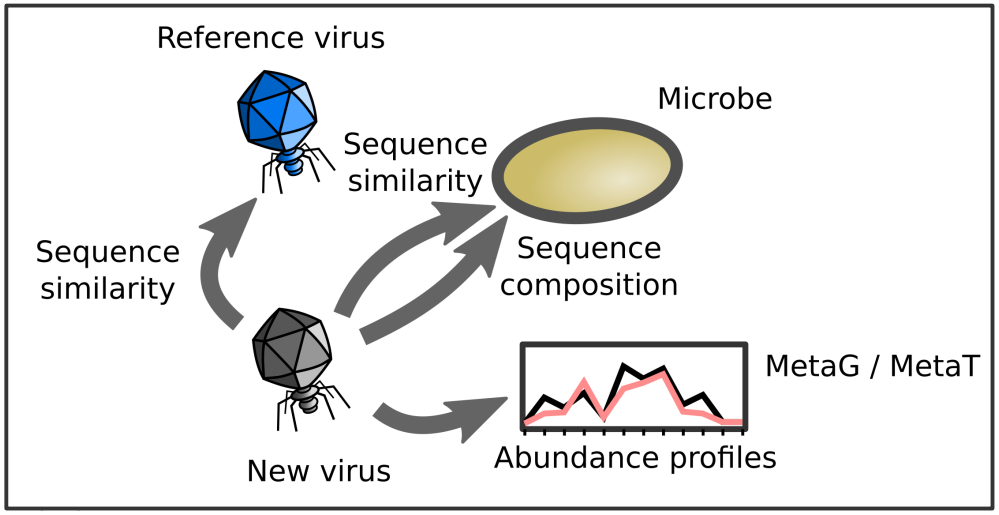
- **4 main types of approaches**
  - Sequence homology with isolate viruses
  - Sequence homology with host genome
  - Sequence composition with host genome
  - Abundance profiles

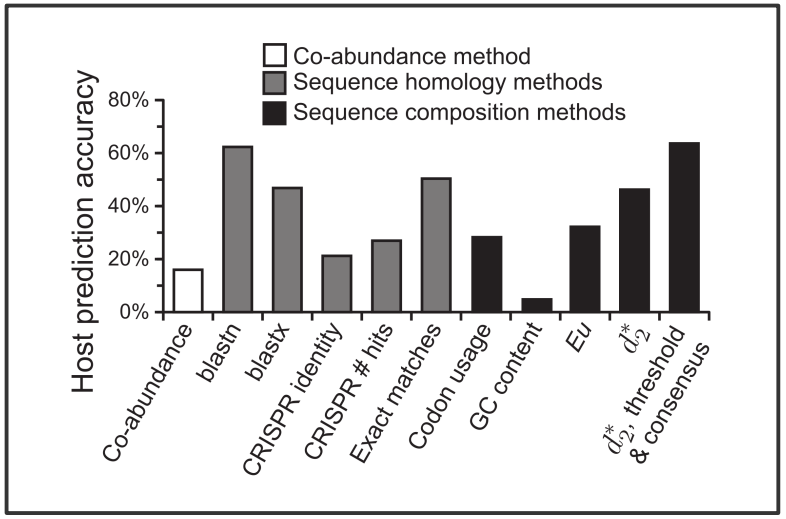


In silico host assignment

- **4 main types of approaches**
  - Sequence homology with isolate viruses
  - Sequence homology with host genome
  - Sequence composition with host genome
  - Abundance profiles

- **Challenges & limitations**
  - Limited accuracy and resolution (reviewed in Edwards et al., 2016)
  - No information on host range, infection rate and efficiency





UViG Quality estimation

- **MIMAG / MISAG:**
  - Completeness (single-copy marker genes)
  - Contamination (single-copy marker genes)
  - Presence & number of rRNA / tRNA genes

UViG Quality estimation

- **MIMAG / MISAG:**
  - Completeness (single-copy marker genes)
  - Contamination (single-copy marker genes)
  - Presence & number of rRNA / tRNA genes

UViG Quality estimation

- **MIMAG / MISAG:**
  - Completeness (single-copy marker genes)
  - Contamination (single-copy marker genes)
  - Presence & number of rRNA / tRNA genes

- **Contamination of UViG**
  - Only possible for specific group, e.g. giant viruses
  - Mentioned as a possibility but not required

UViG Quality estimation

- **MIMAG / MISAG:**
  - Completeness (single-copy marker genes)
  - Contamination (single-copy marker genes)
  - Presence & number of rRNA / tRNA genes

- **Contamination of UViG**
  - Only possible for specific group, e.g. giant viruses
  - Mentioned as a possibility but not required

- **Main UViG metrics / requirement: completeness**
  - identification of complete genomes (circular contigs)

UViG Quality estimation

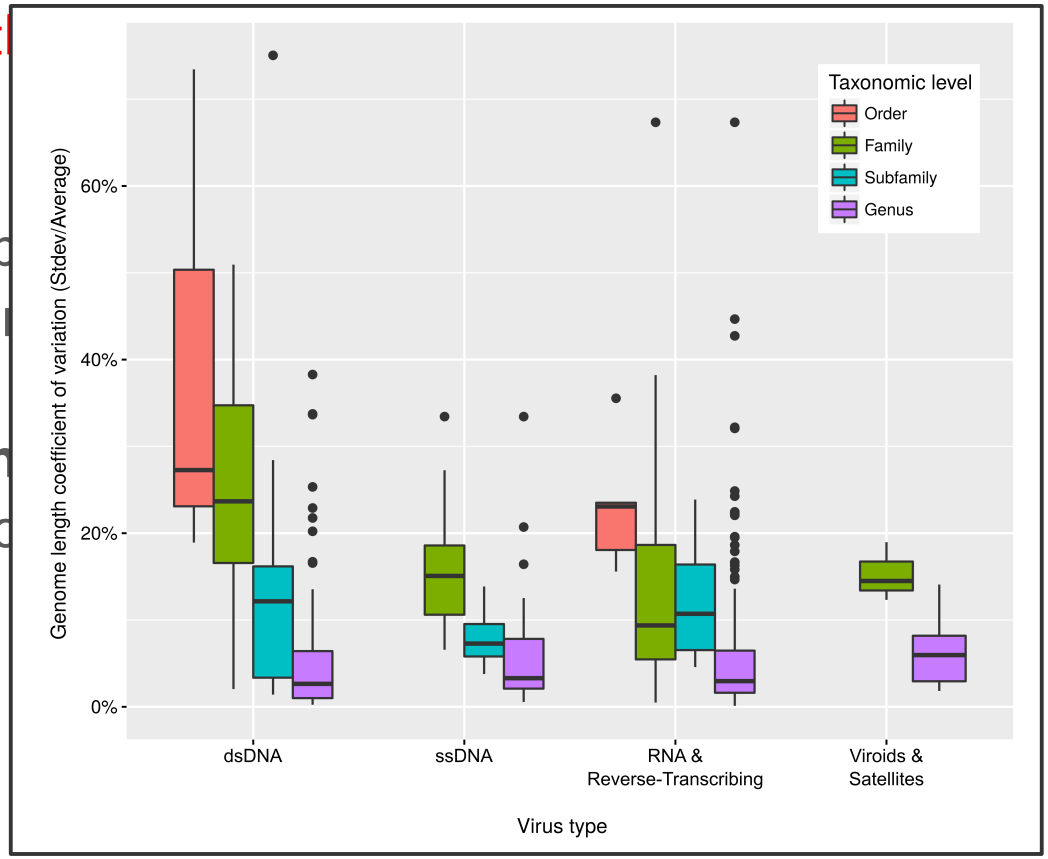
- **MIMAG / MISAG:**
  - Completeness (single-copy marker genes)
  - Contamination (single-copy marker genes)
  - Presence & number of rRNA / tRNA genes

- **Contamination of UViG**
  - Only possible for specific group, e.g. giant viruses
  - Mentioned as a possibility but not required

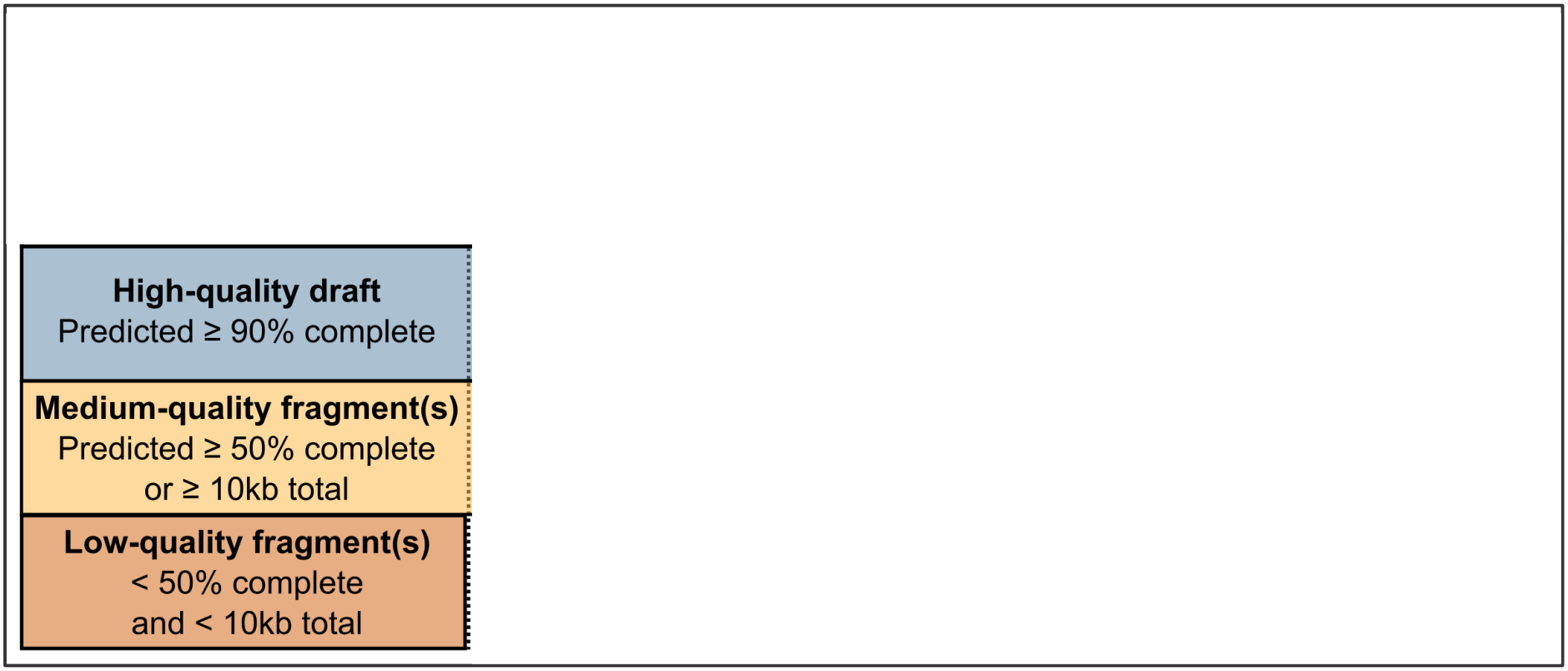
- **Main UViG metrics / requirement: completeness**
  - identification of complete genomes (circular contigs)
  - affiliation at ~ genus rank:
    expected genome size
    -

UViG Quality estimation

- **MIMAG / MISAG:**
  - Completeness (single-copy marker genes)
  - Contamination (single-copy marker genes)
  - Presence & number of rRNA / t
- **Contamination of UViG**
  - Only possible for specific group
  - Mentioned as a possibility but
- **Main UViG metrics / requiremen**
  - identification of complete genc
  - affiliation at ~ genus rank:
    - expected genome size



UViG Quality estimation

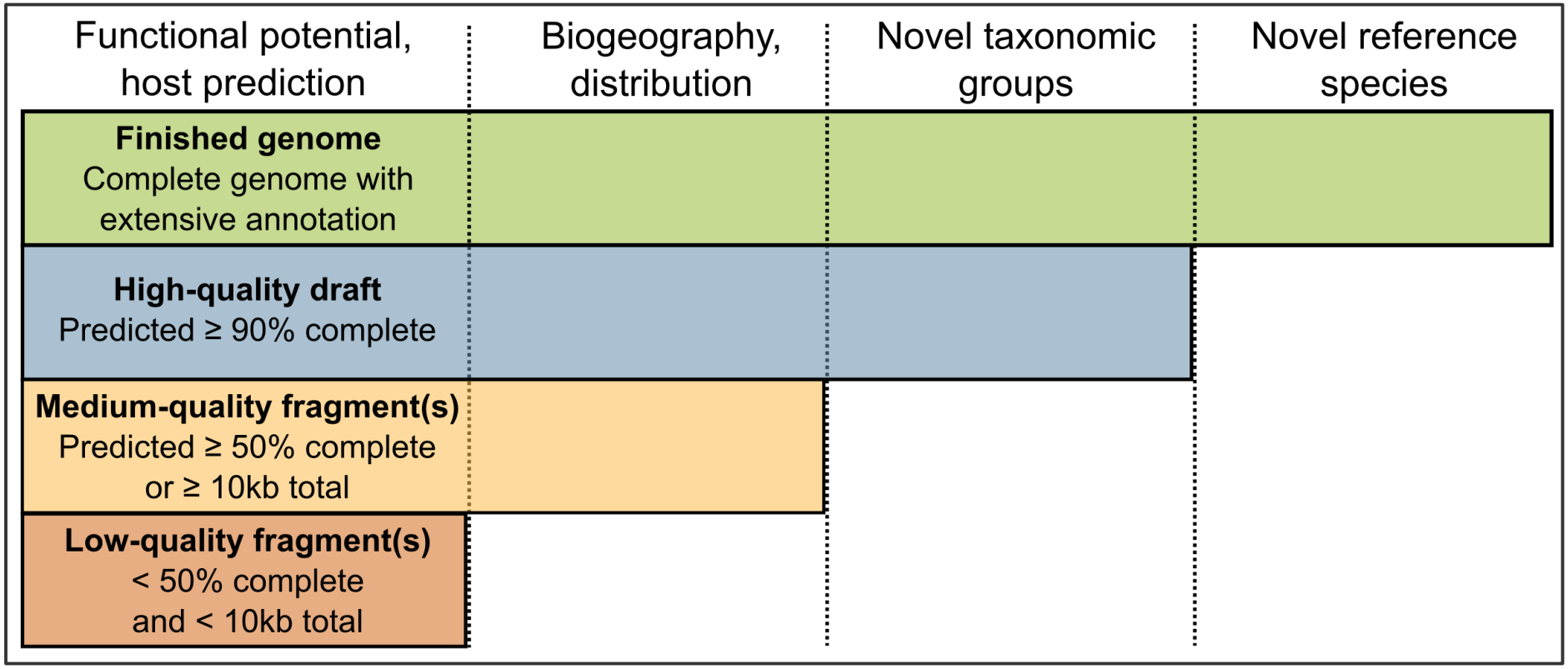


- **Classification in 4 categories**

UViG Quality estimation

- **Classification in 4 categories**

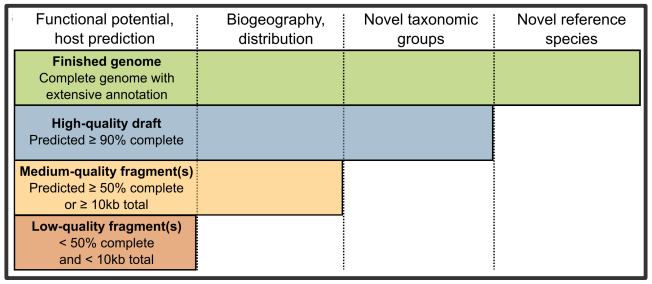
UViG Quality estimation



- **Classification in 4 categories**
  - Key information to guide genome analysis

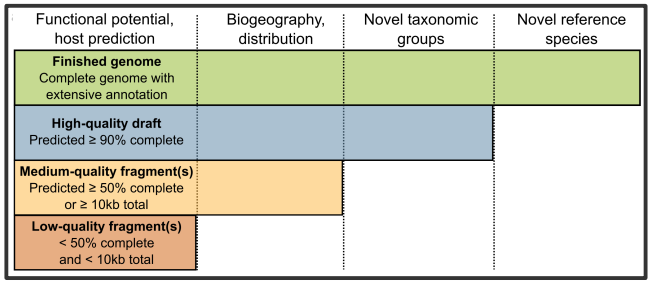
UViG Quality estimation

- **Example Global Ocean Virome**
  - 15,222 contigs (≥ 10kb)



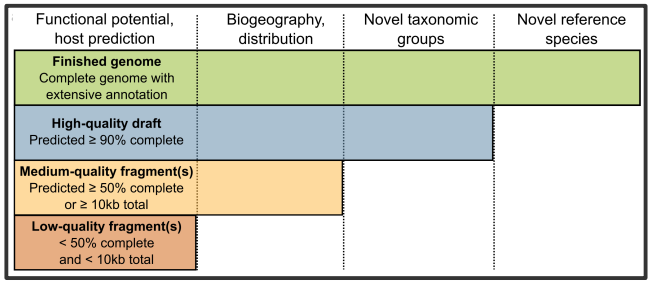
UViG Quality estimation

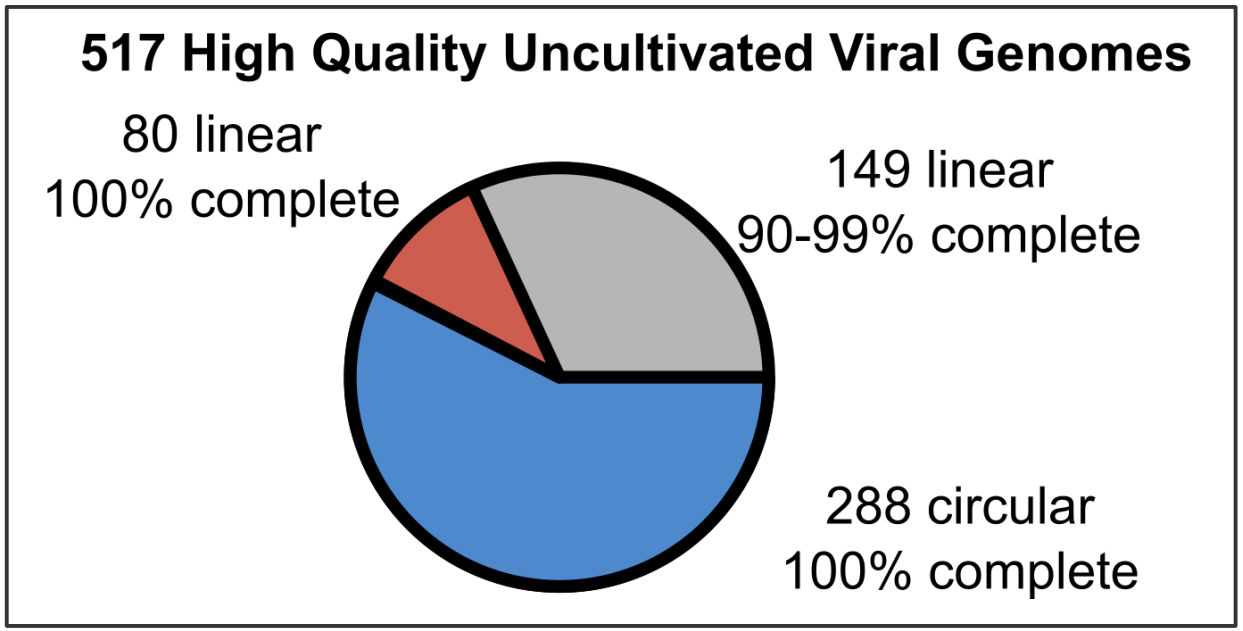
- **Example Global Ocean Virome**
  - 15,222 contigs (≥ 10kb)
  - Quality classification:
    - 0 low quality
    - 14,705 medium quality
    - 517 high quality



UViG Quality estimation

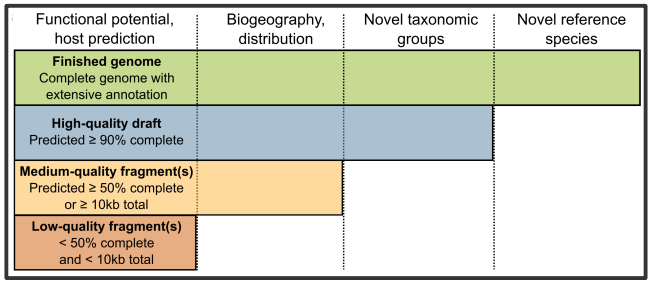
- **Example Global Ocean Virome**
  - 15,222 contigs (≥ 10kb)
  - Quality classification:
    - 0 low quality
    - 14,705 medium quality
    - 517 high quality

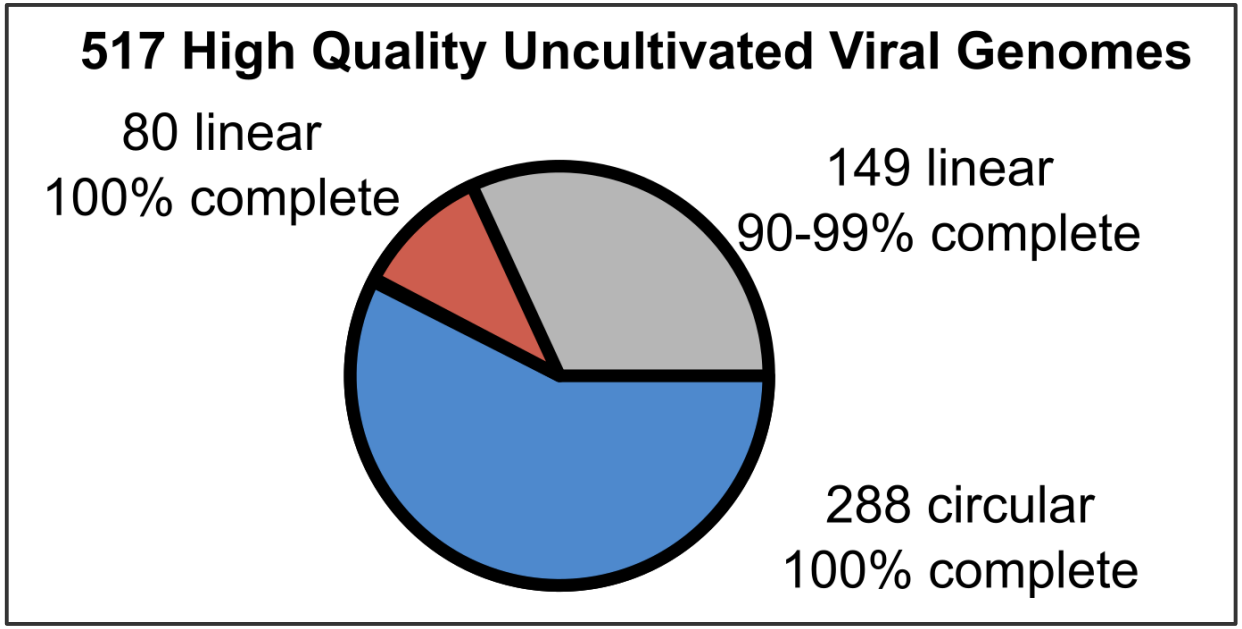




UViG Quality estimation

- **Example Global Ocean Virome**
  - 15,222 contigs (≥ 10kb)
  - Quality classification:
    - 0 low quality
    - 14,705 medium quality
    - 517 high quality
      - 3 new circular selected for manual annotation → finished genomes





UViGs: data release

- **INSDC as central repository:**
  - Member databases: NCBI, EMBL-EBI, DDBJ
  - Submission using MIUViG checklist
    and environment-relevant MIxS checklists if new data generated
  - Priority to high-quality and finished genomes

UViGs: data release

- **INSDC as central repository:**
  - Member databases: NCBI, EMBL-EBI, DDBJ
  - Submission using MIUViG checklist
    and environment-relevant MIxS checklists if new data generated
  - Priority to high-quality and finished genomes

- **Other databases:**
  - ICTV: "coding-complete", i.e. finished & high-quality complete
  - IMG/VR: genome quality soon to be included

UViGs: data release

- **INSDC as central repository:**
  - Member databases: NCBI, EMBL-EBI, DDBJ
  - Submission using MIUViG checklist
    and environment-relevant MIxS checklists if new data generated
  - Priority to high-quality and finished genomes

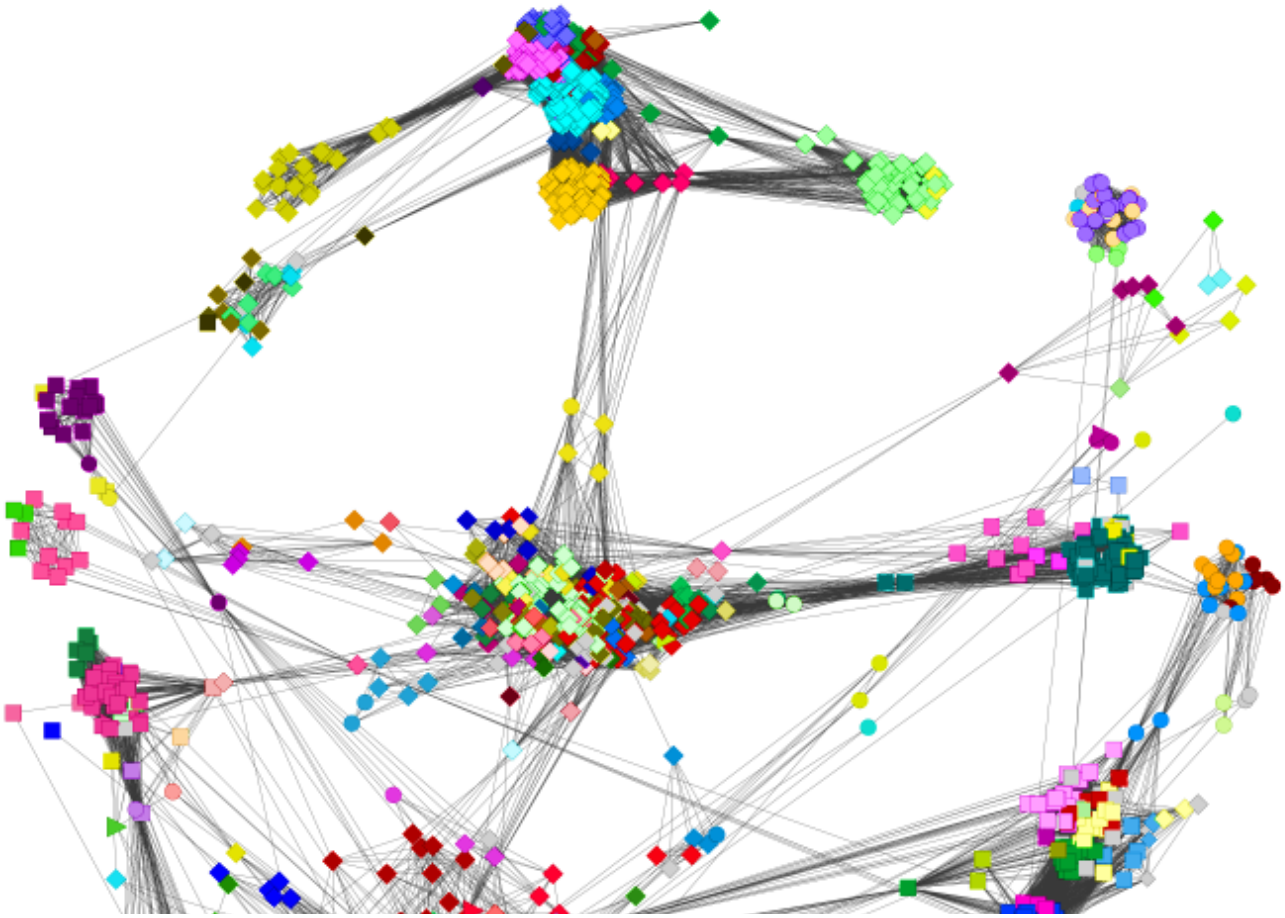
- **Other databases:**
  - ICTV: "coding-complete", i.e. finished & high-quality complete
  - IMG/VR: genome quality soon to be included

- **Active development areas**
  - "Universal" virus sequence detection tool
  - Robust genome-based taxonomy software
  - Unified, comprehensive, & annotated database of virus proteins

… All that for what ?

JGI JOINT GENOME INSTITUTE

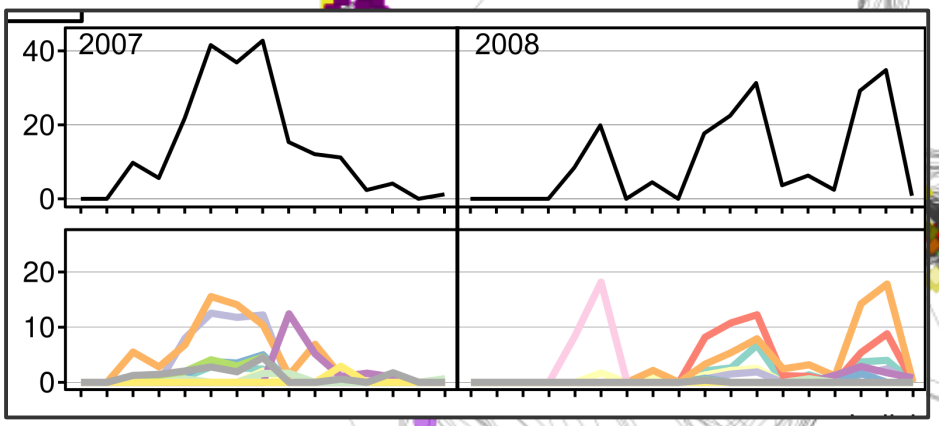
End goal: comprehensive (host-resolved) mapping of the virosphere

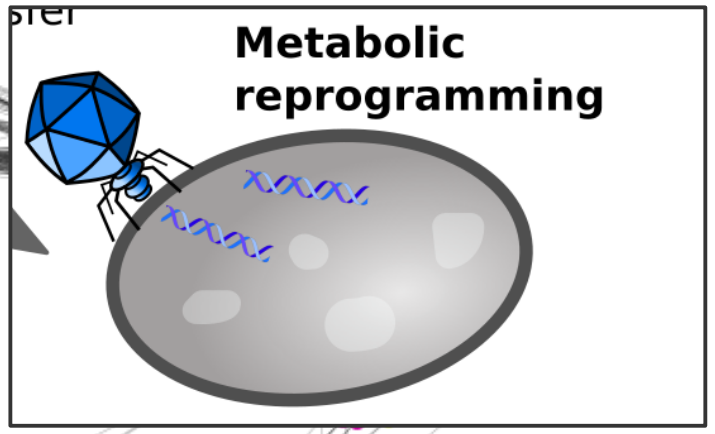


… All that for what ?

End goal: comprehensive (host-resolved) mapping of the virosphere

- **Viral impacts on ecosystems**
  - infection dynamics in nature
  - virus-aware ecosystem models





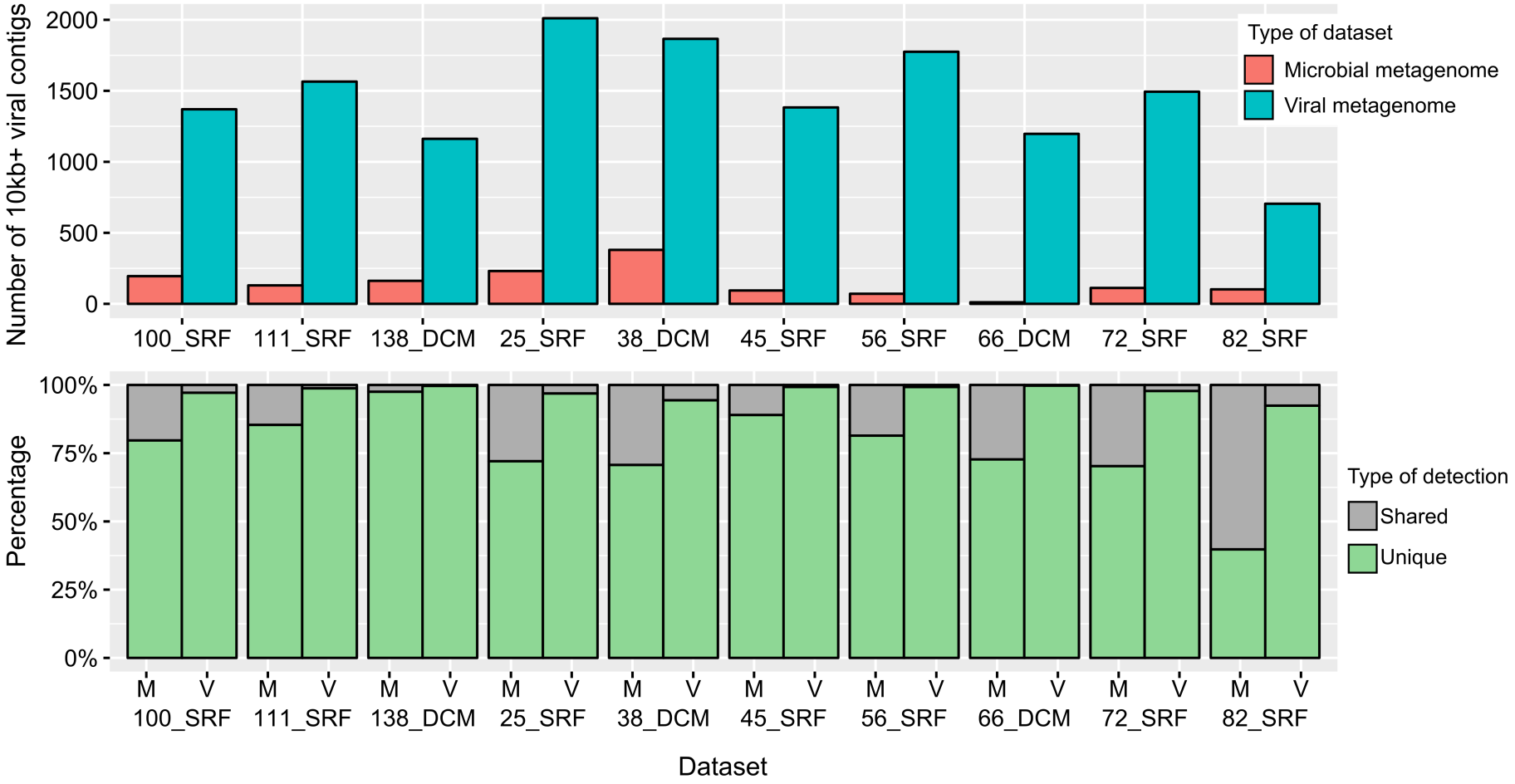
- **Microbial manipulation**
  - host cell takeover mechanisms
  - biotechnological applications

Thanks

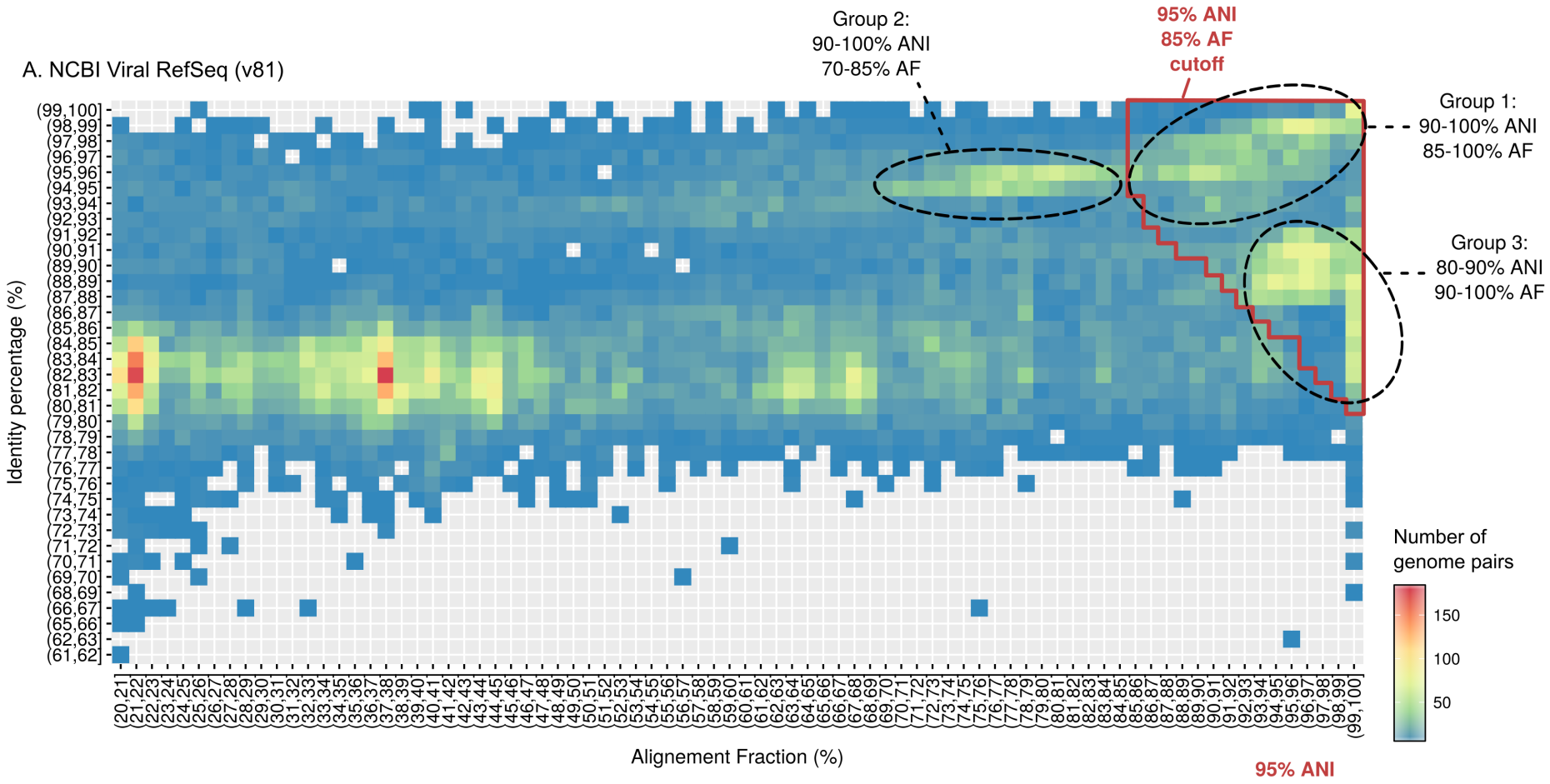
JGI
JOINT GENOME INSTITUTE

> 50 contributing labs !

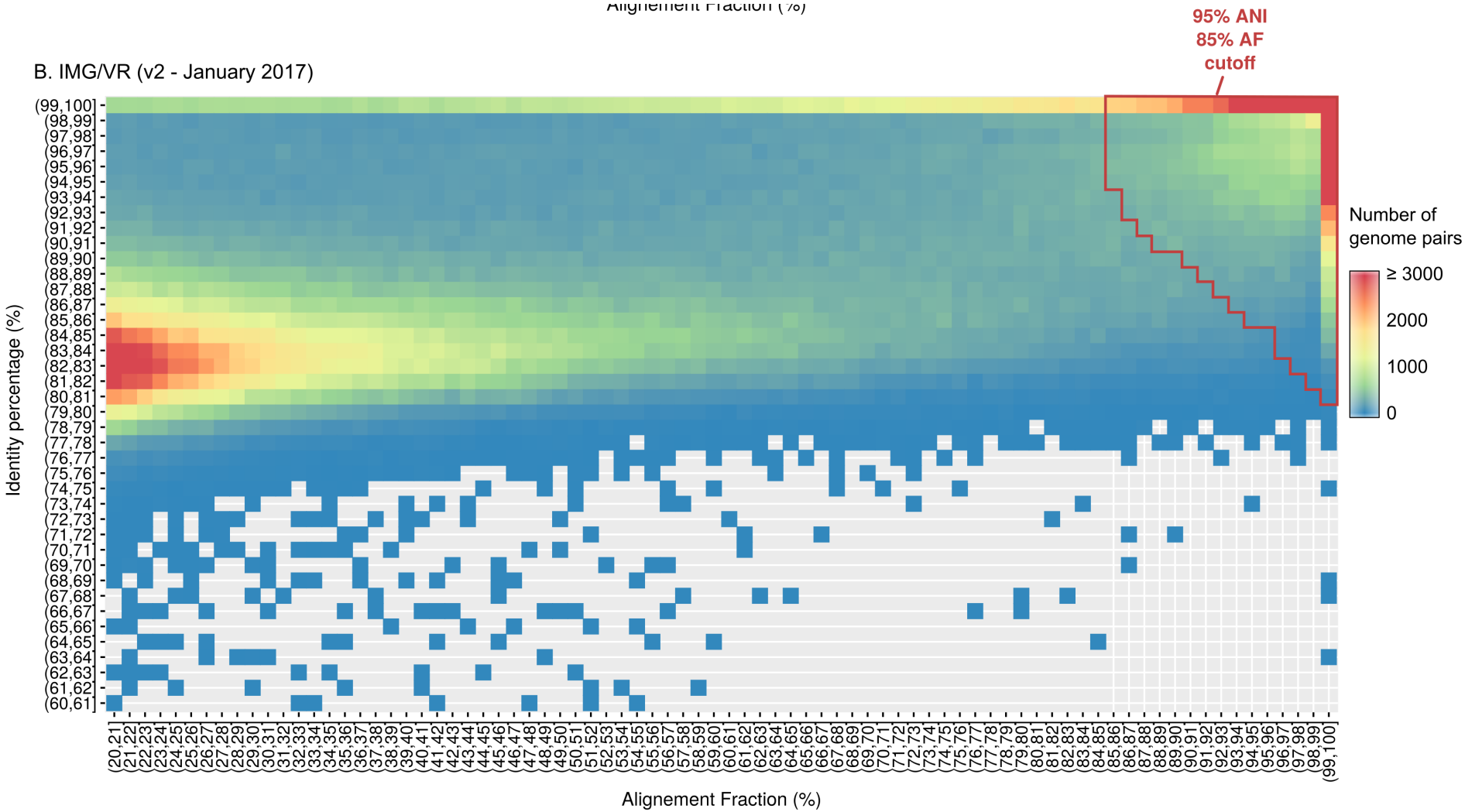
Targeted vs untargeted



vOTU threshold



vOTU threshold



vOTU threshold