



**JOINT GENOME INSTITUTE**

A DOE OFFICE OF SCIENCE USER FACILITY

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

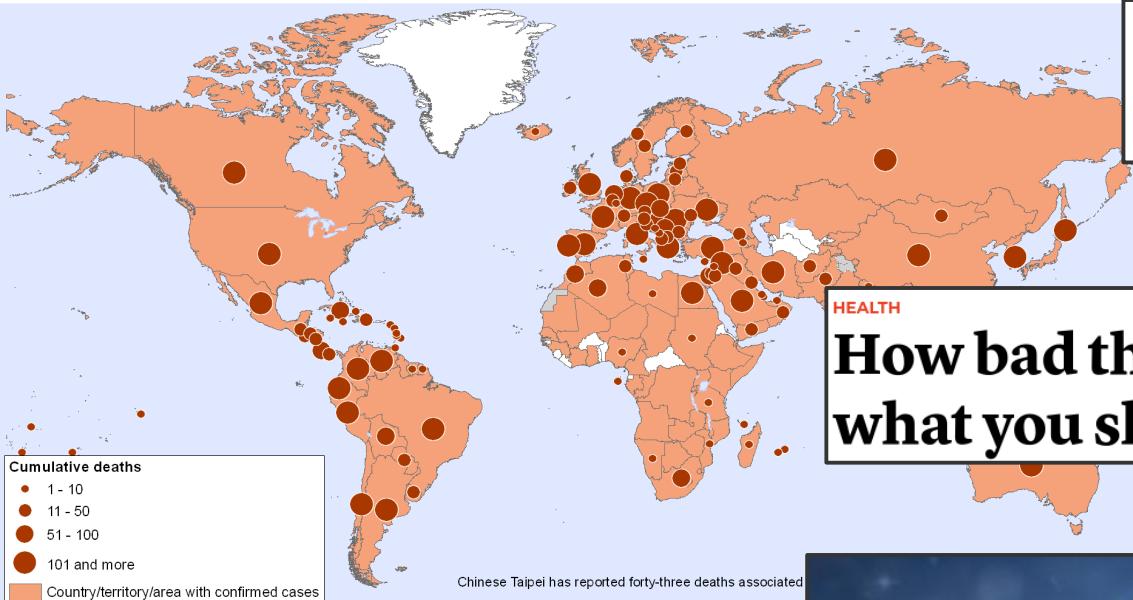
Simon Roux, DOE Joint Genome Institute, LBNL  
GSC 20 – San Diego, May 4<sup>th</sup> 2018

# Viruses in the news

Pandemic (H1N1) 2009

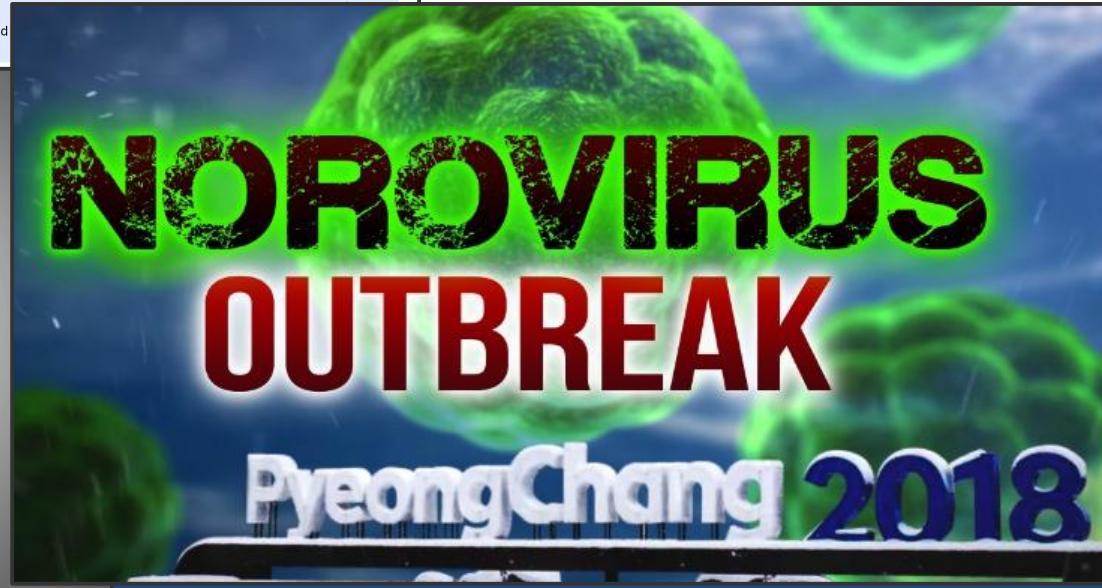
Countries, territories and areas with lab confirmed cases and number of deaths as reported to WHO

Status as of 04 July 2010



**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**



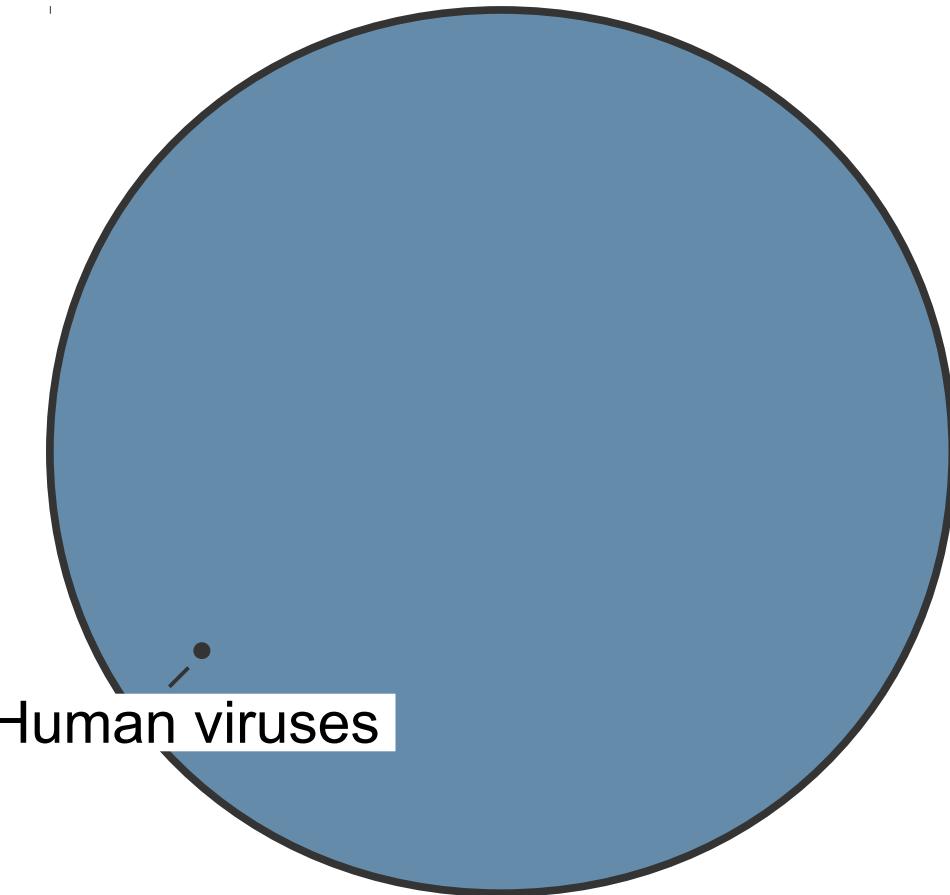
## CDC TRAVEL ALERT - ZIKA VIRUS



# Viruses of microbes are everywhere



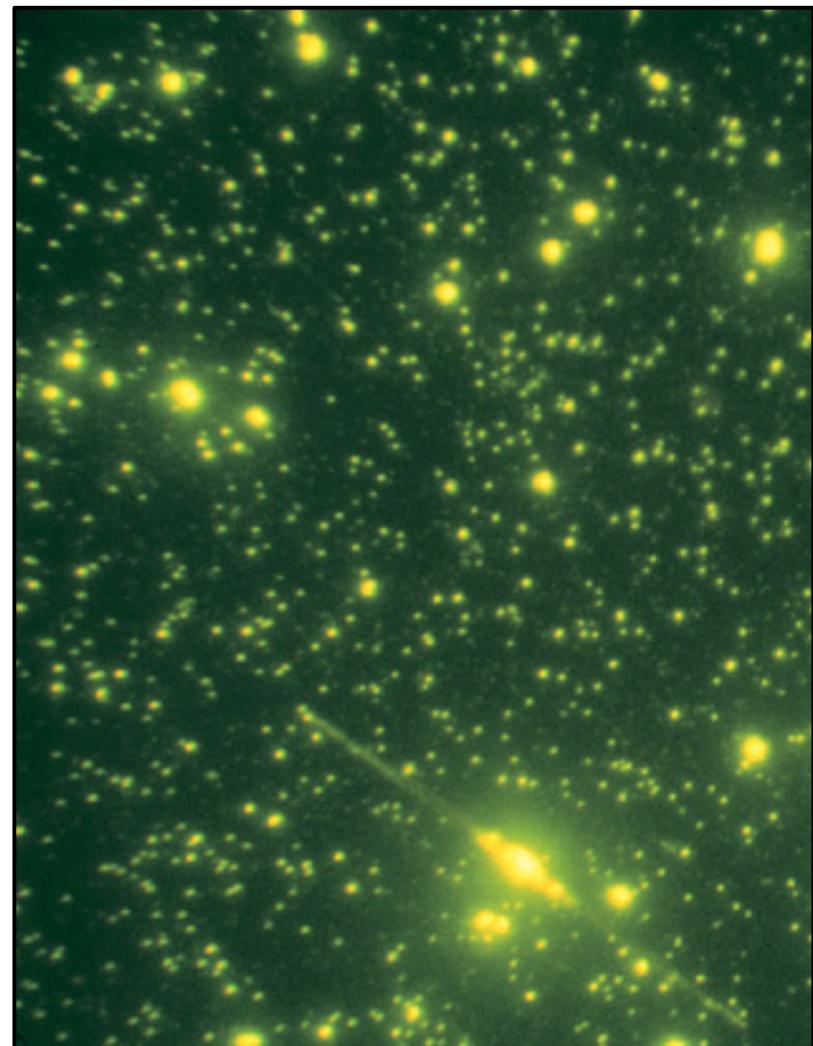
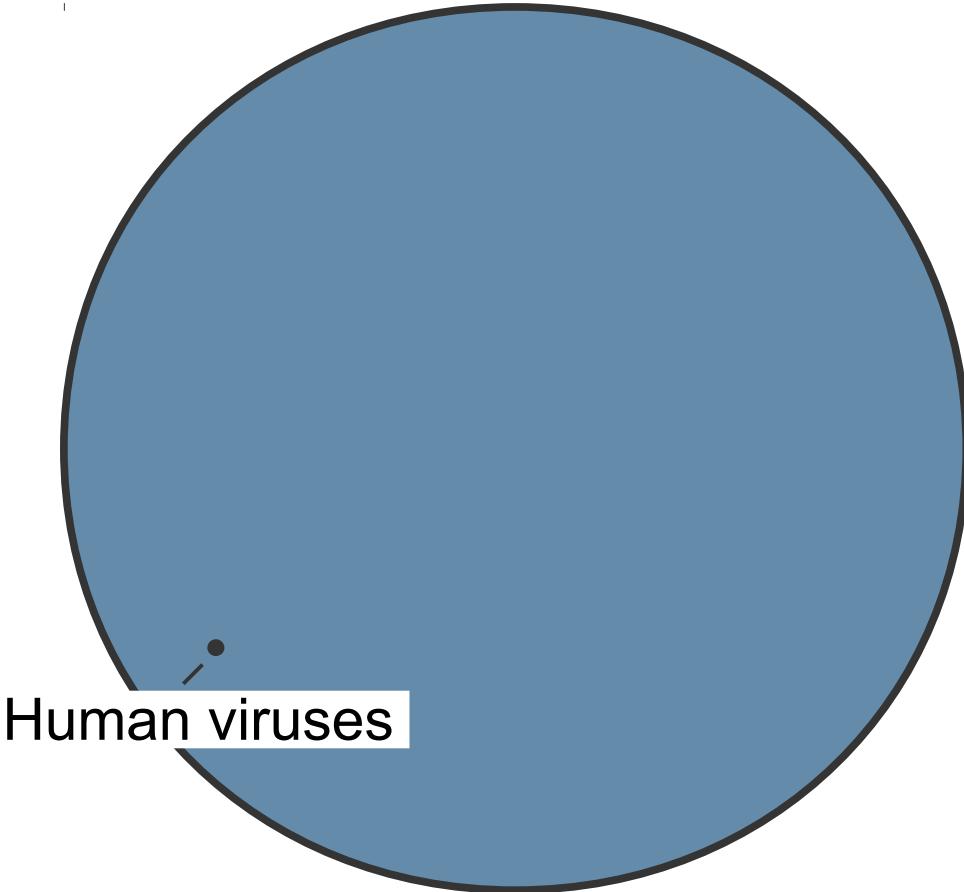
Viruses of microbes



# Viruses of microbes are everywhere



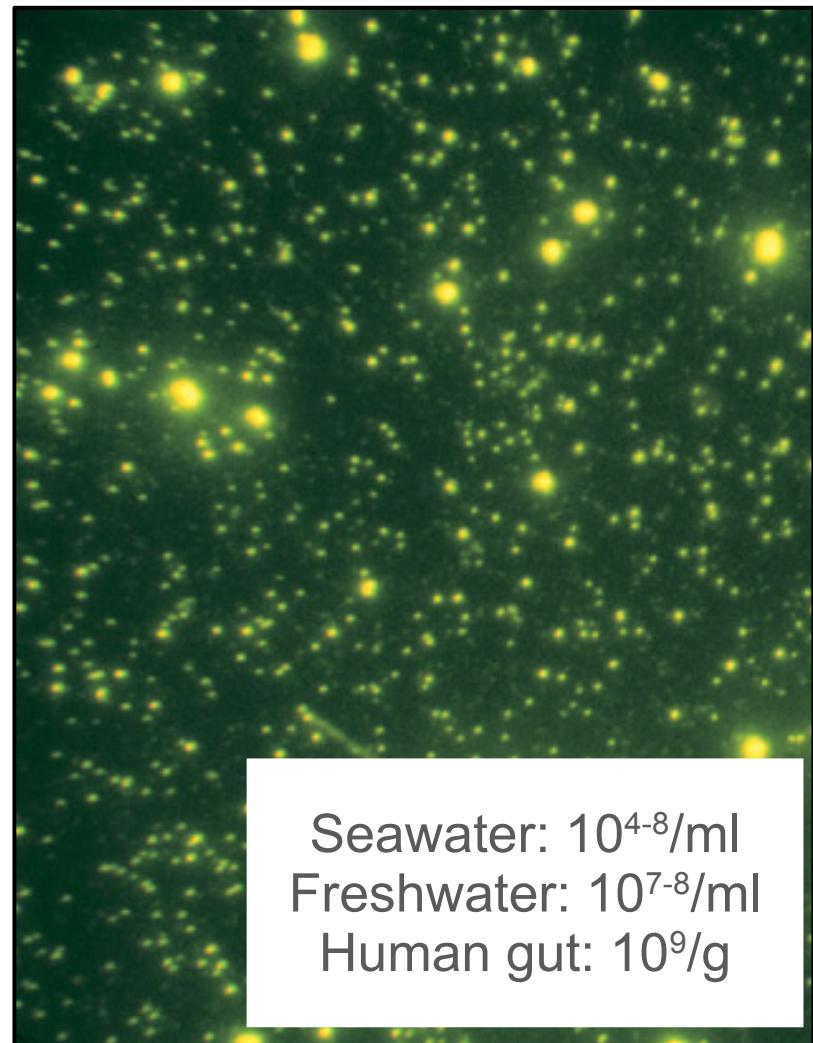
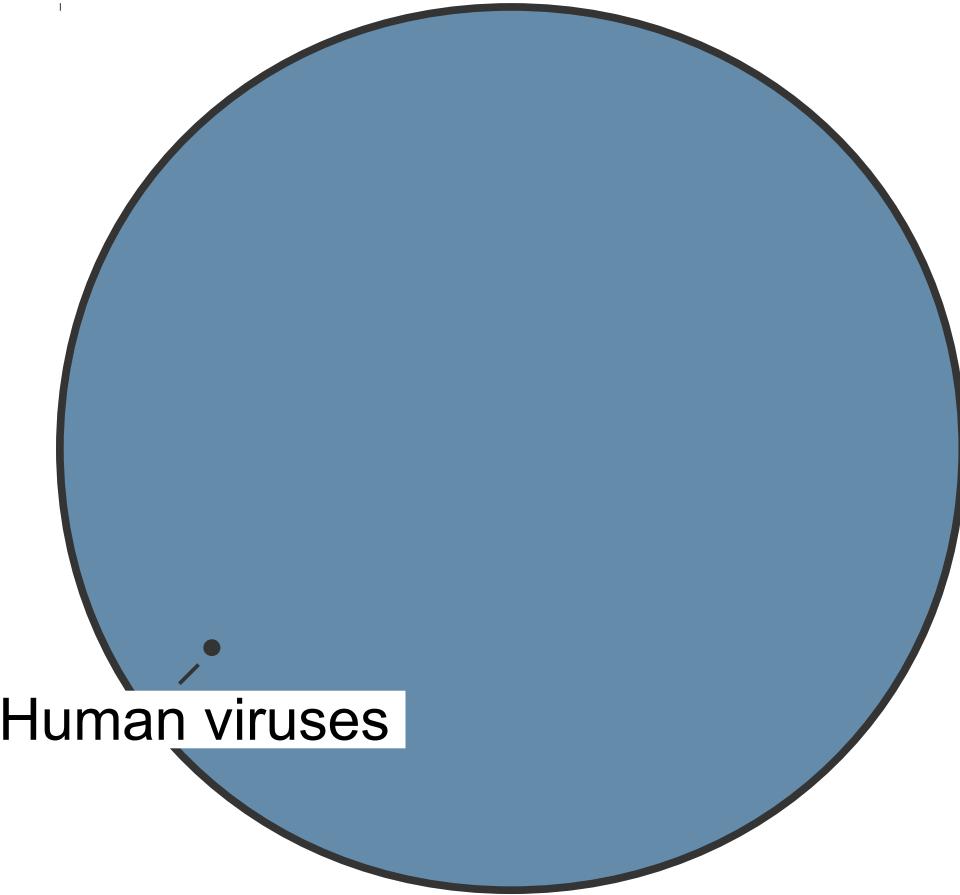
Viruses of microbes



# Viruses of microbes are everywhere

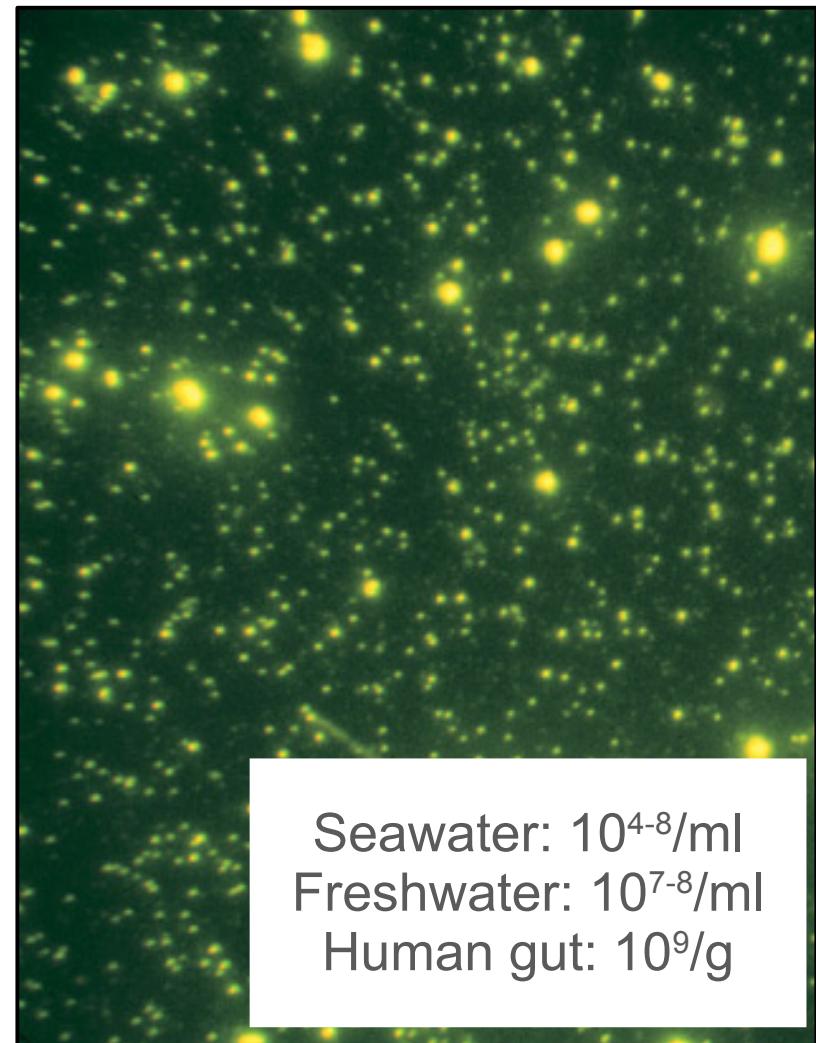
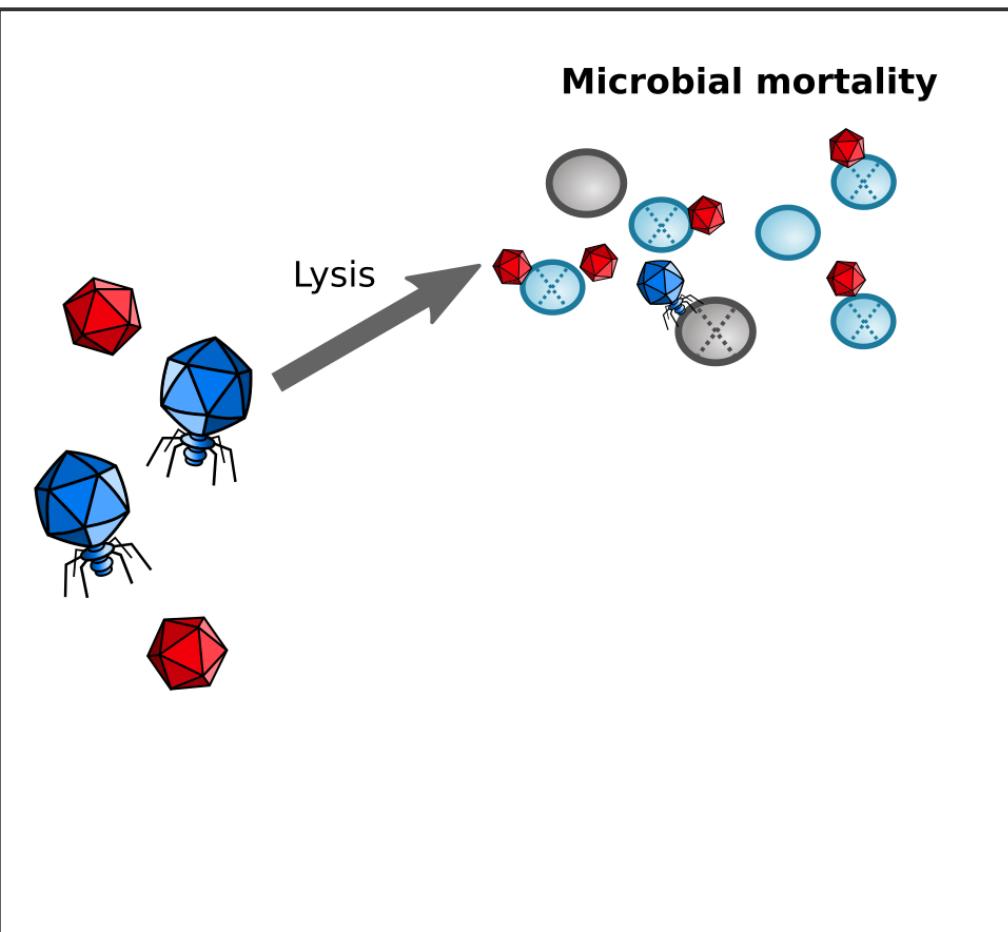


Viruses of microbes



Seawater:  $10^{4-8}/\text{ml}$   
Freshwater:  $10^{7-8}/\text{ml}$   
Human gut:  $10^9/\text{g}$

# Viruses of microbes are everywhere



Seawater:  $10^{4-8}/\text{ml}$   
Freshwater:  $10^{7-8}/\text{ml}$   
Human gut:  $10^9/\text{g}$

# Viruses of microbes are everywhere

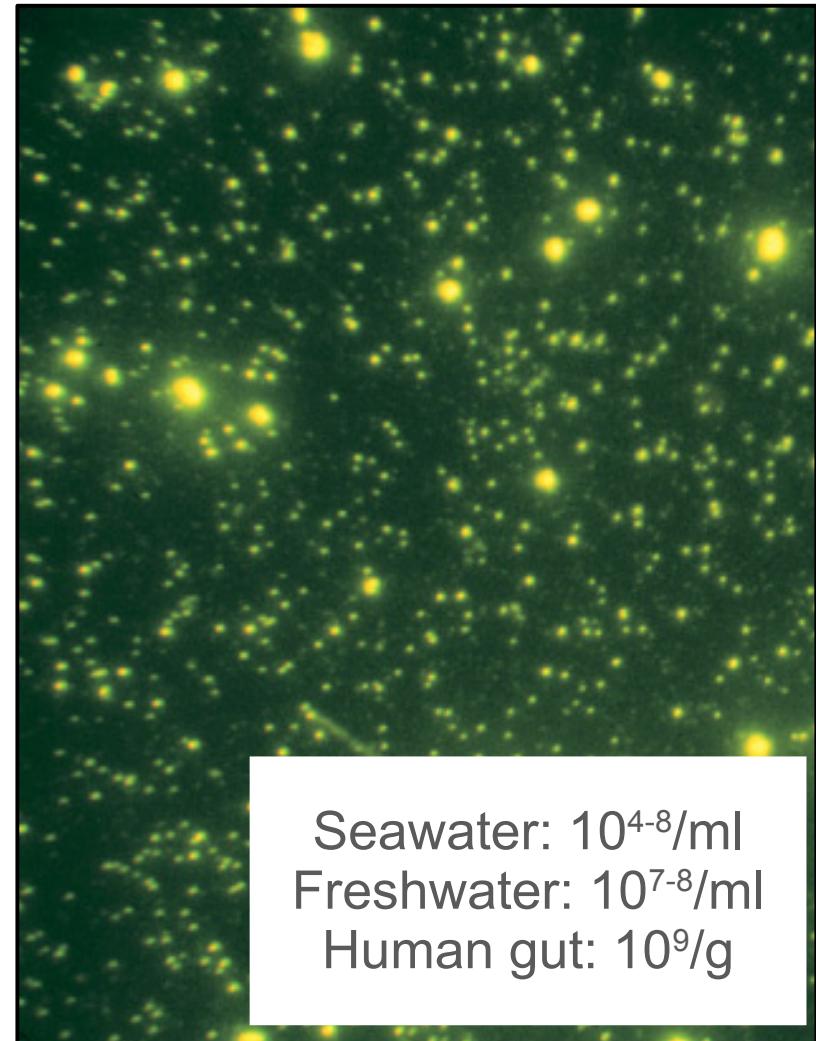
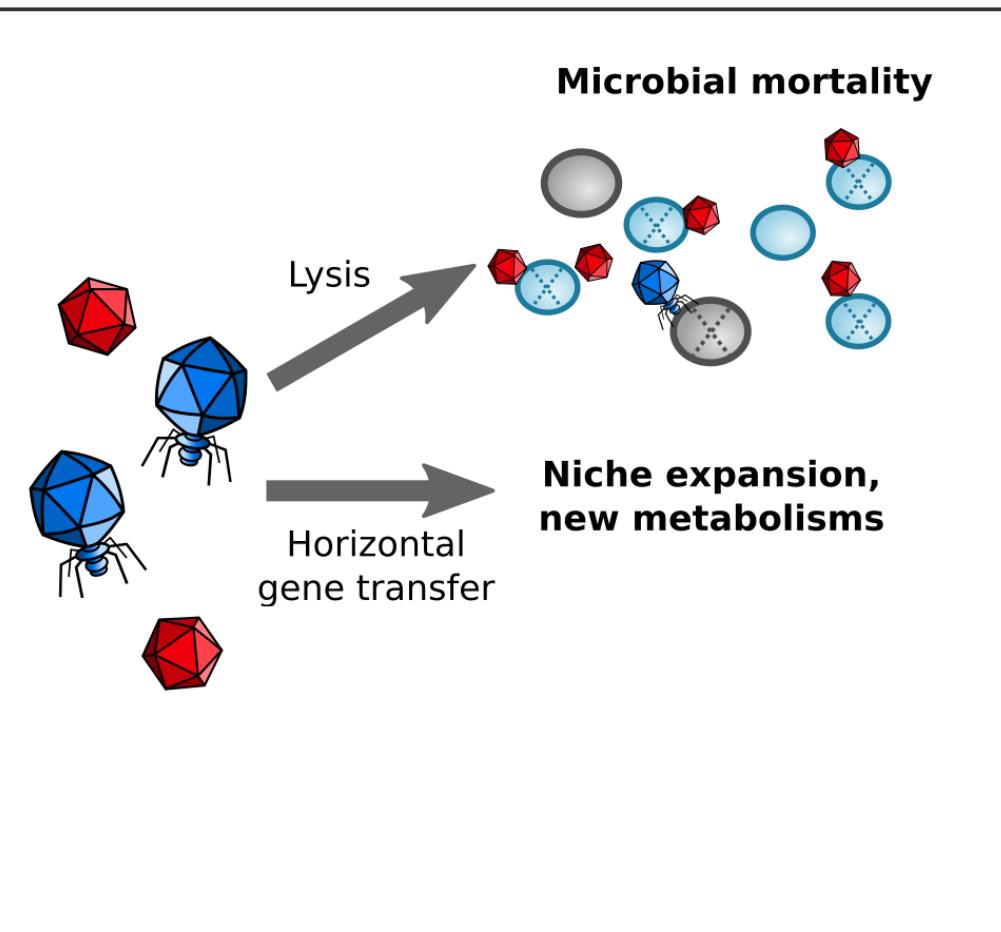
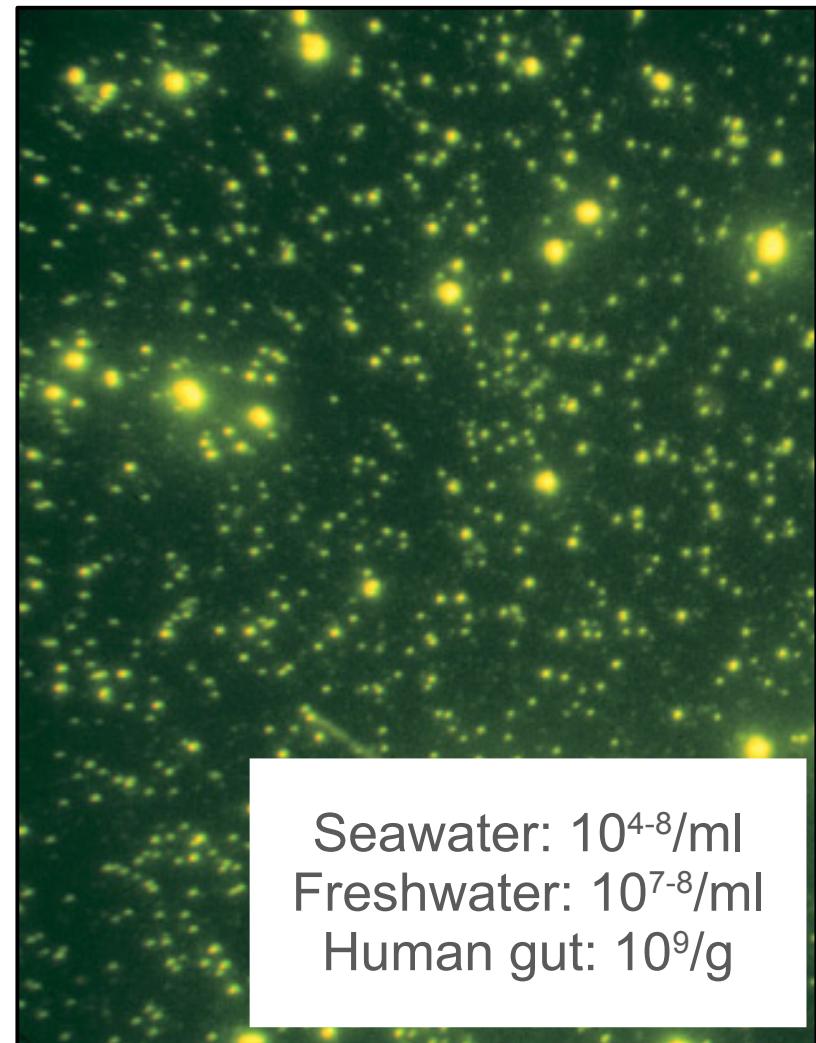
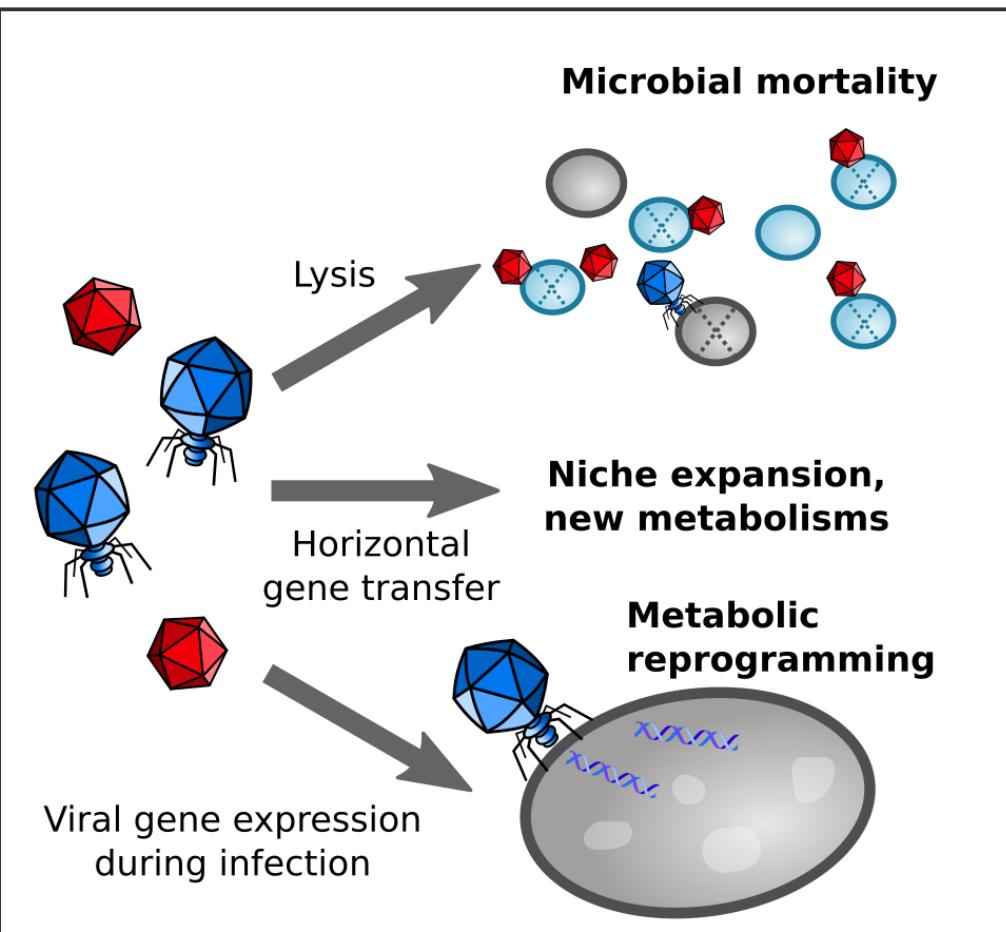
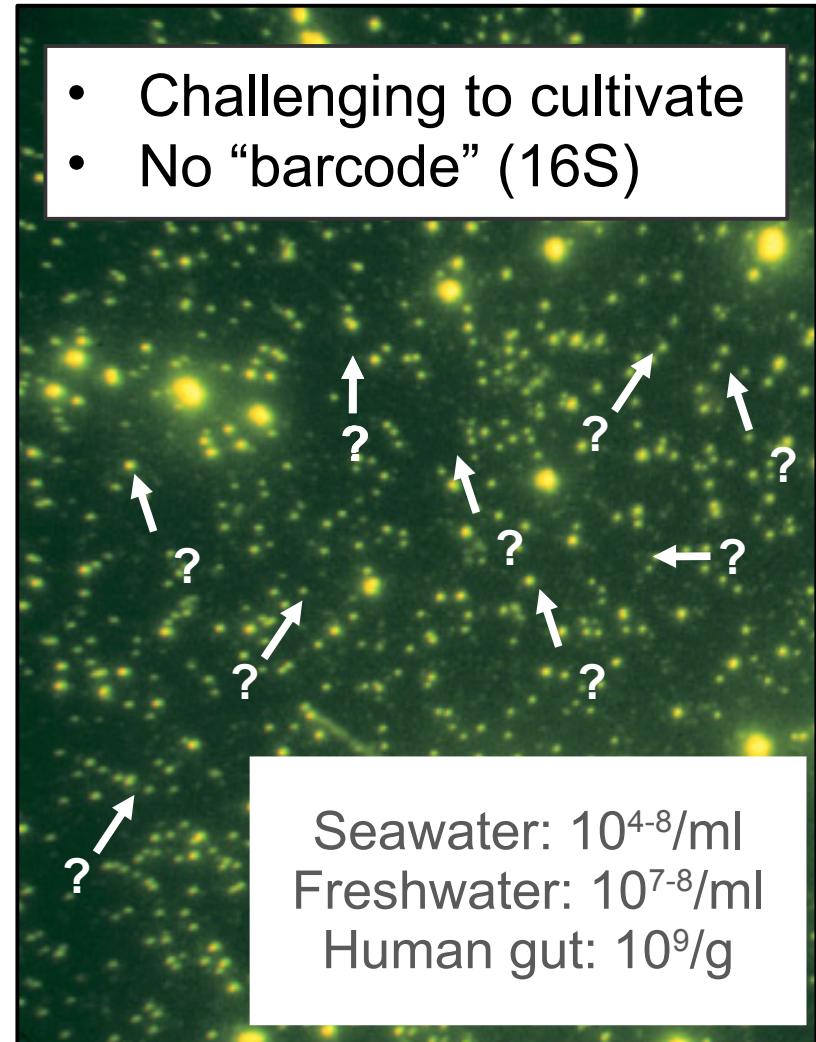
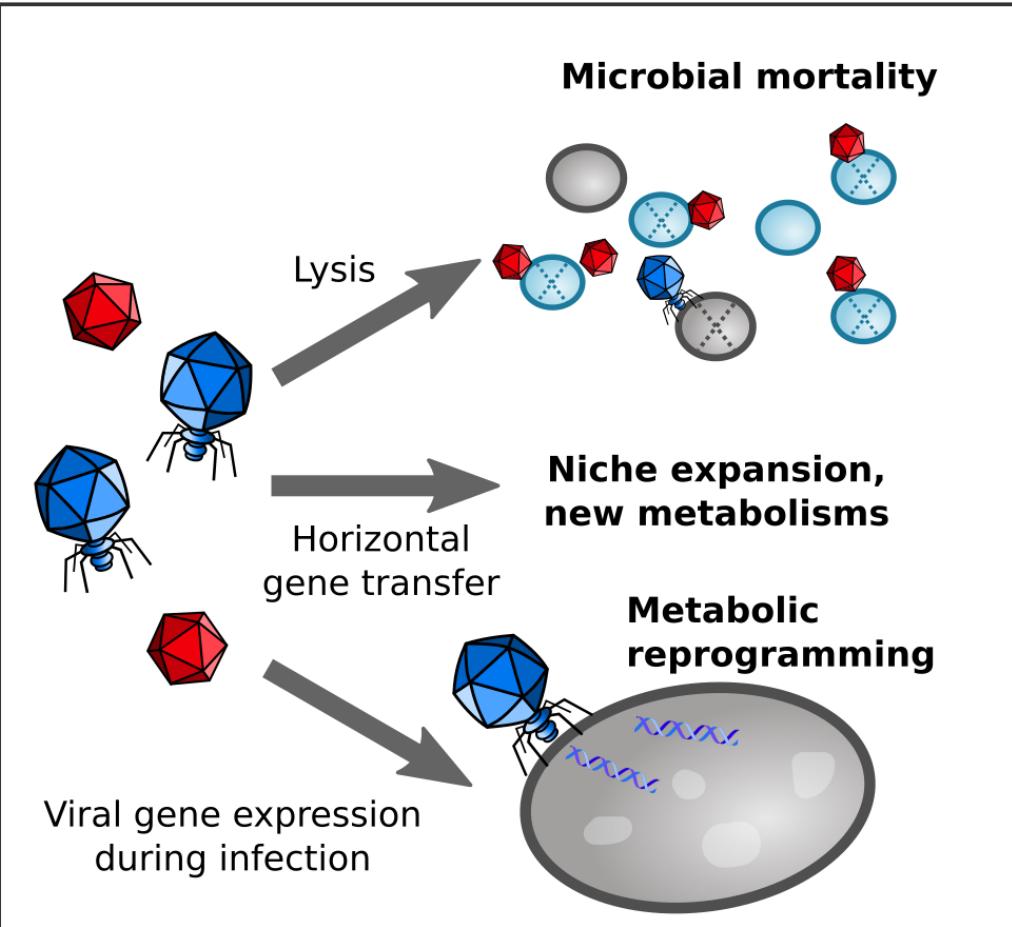


Image: Fuhrman Lab

# Viruses of microbes are everywhere

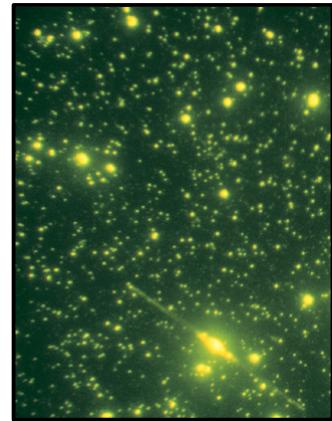
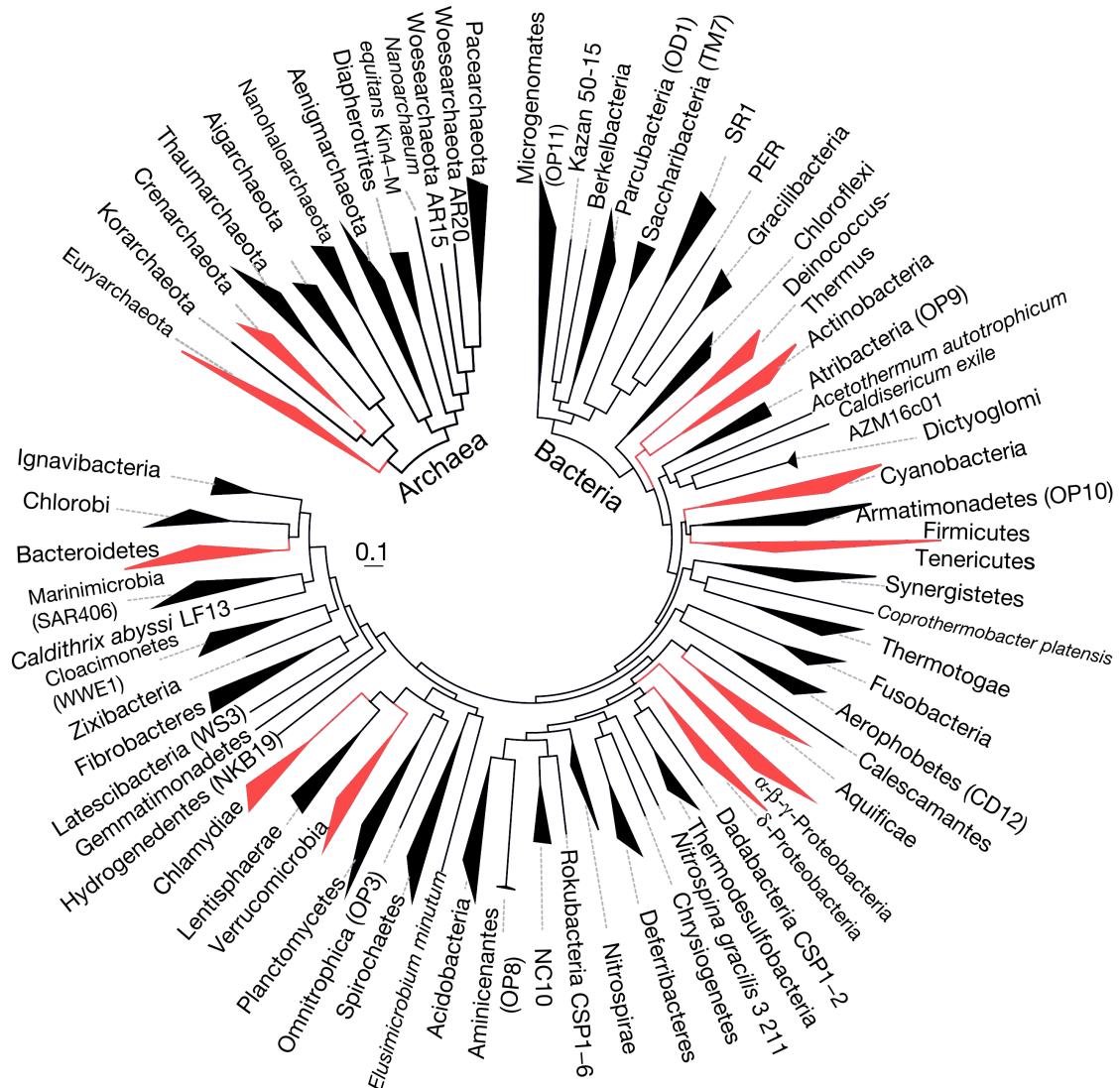


# Challenges and unknowns



# Challenges of viral ecology

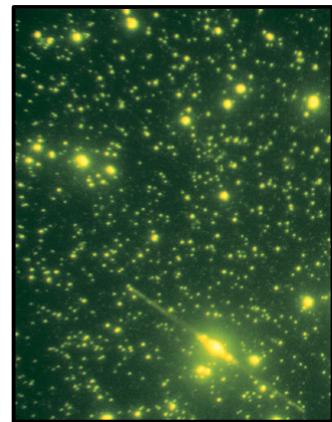
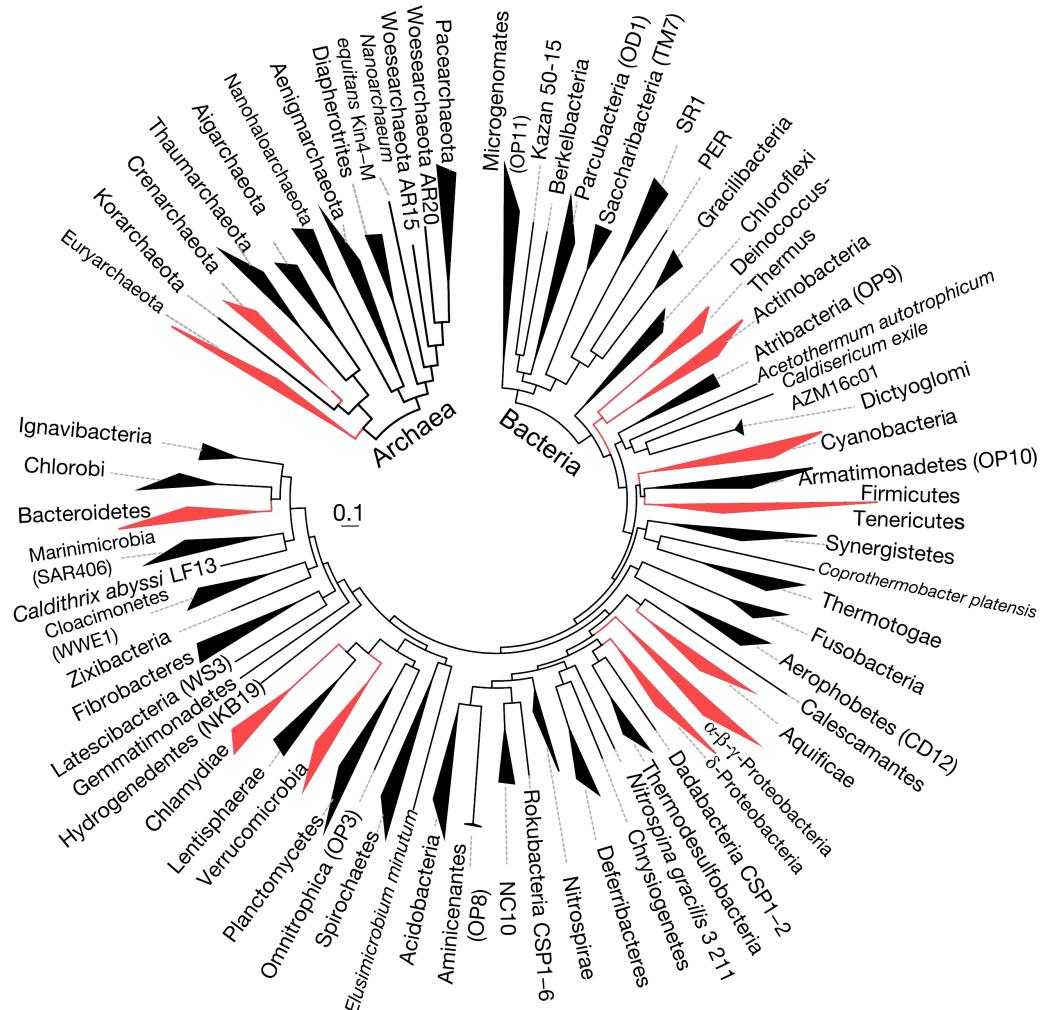
- Low coverage of viral diversity



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

# Challenges of viral ecology

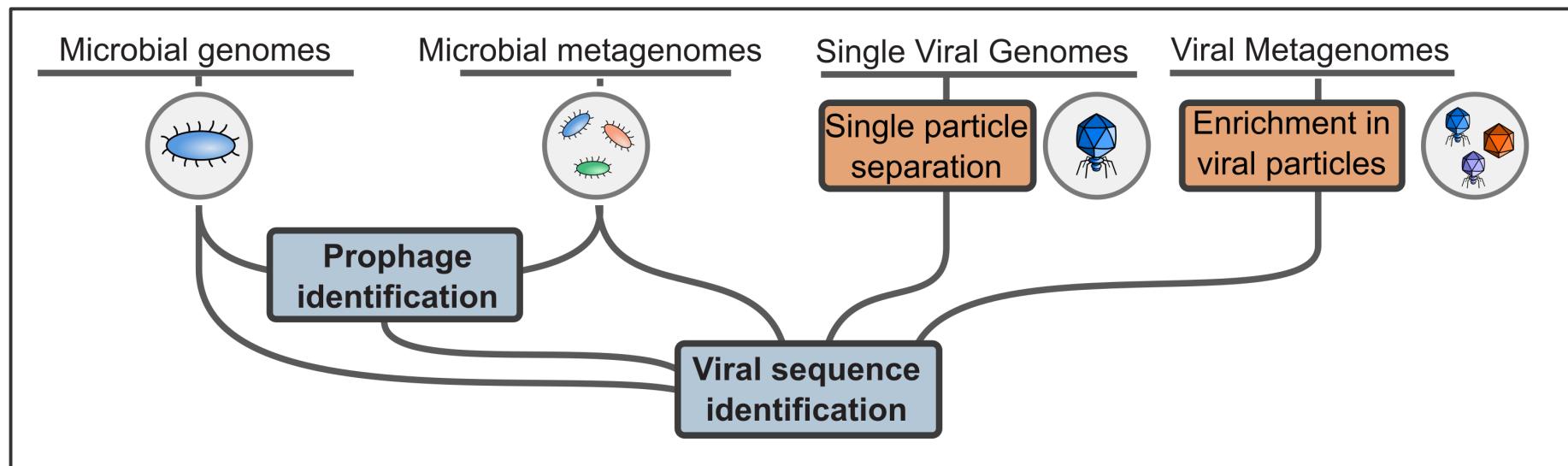
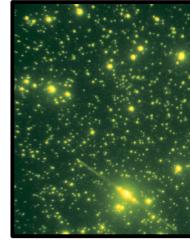
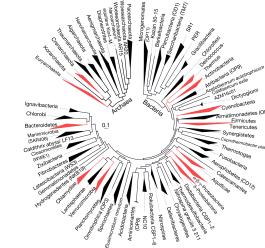
- Low coverage of viral diversity
- Genomes from uncultivated viruses



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

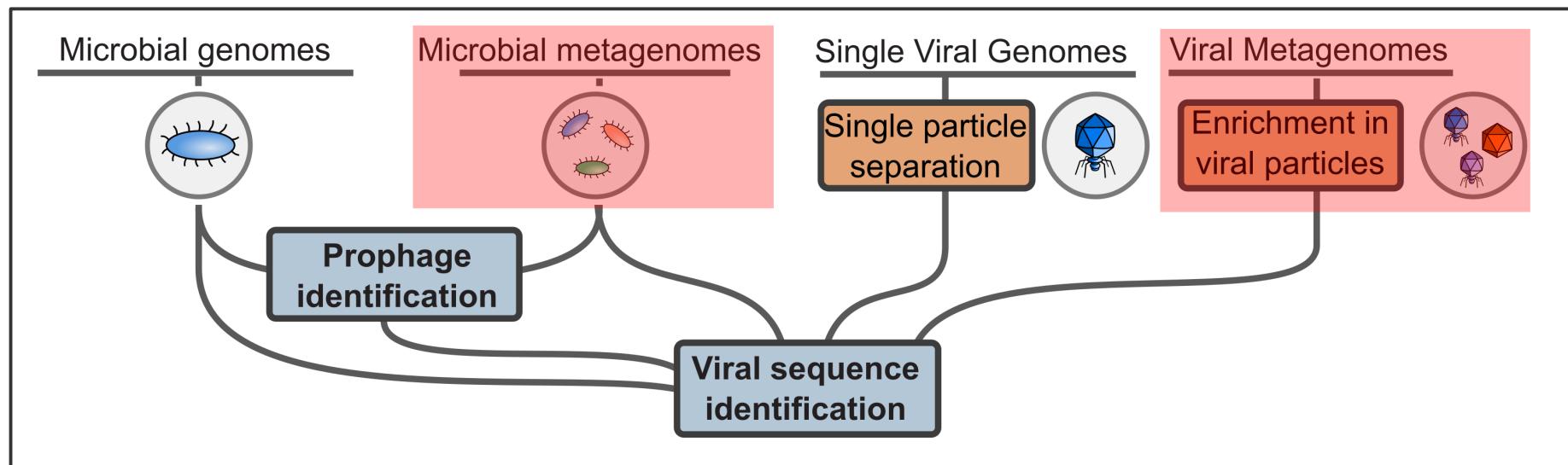
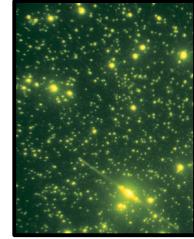
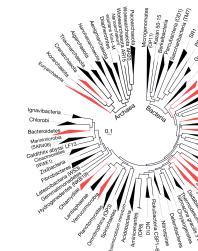
# Mapping the viral sequence space

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

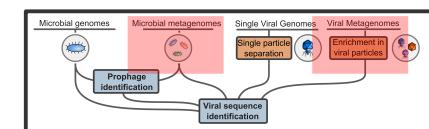
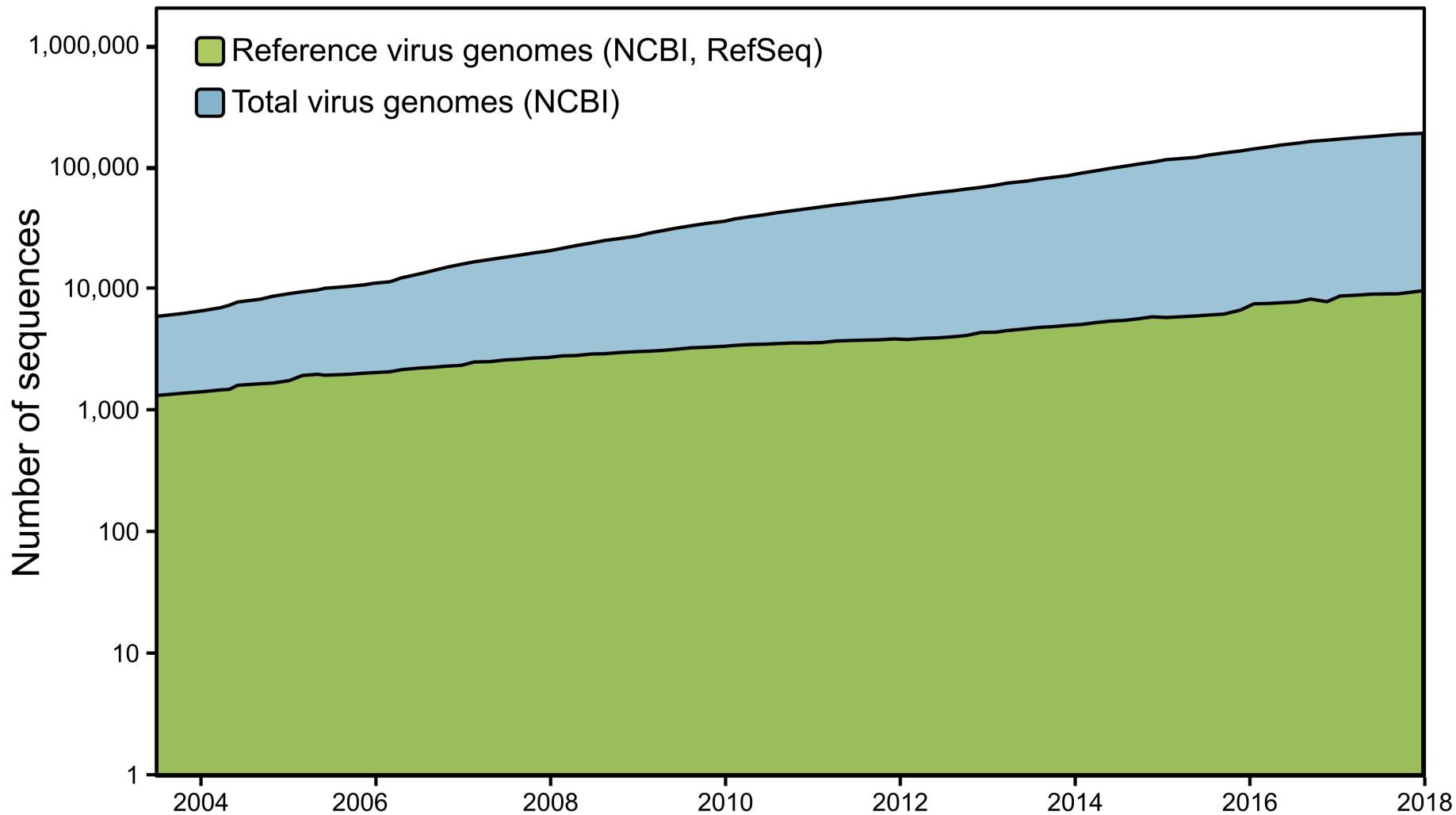


# Mapping the viral sequence space

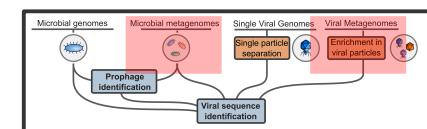
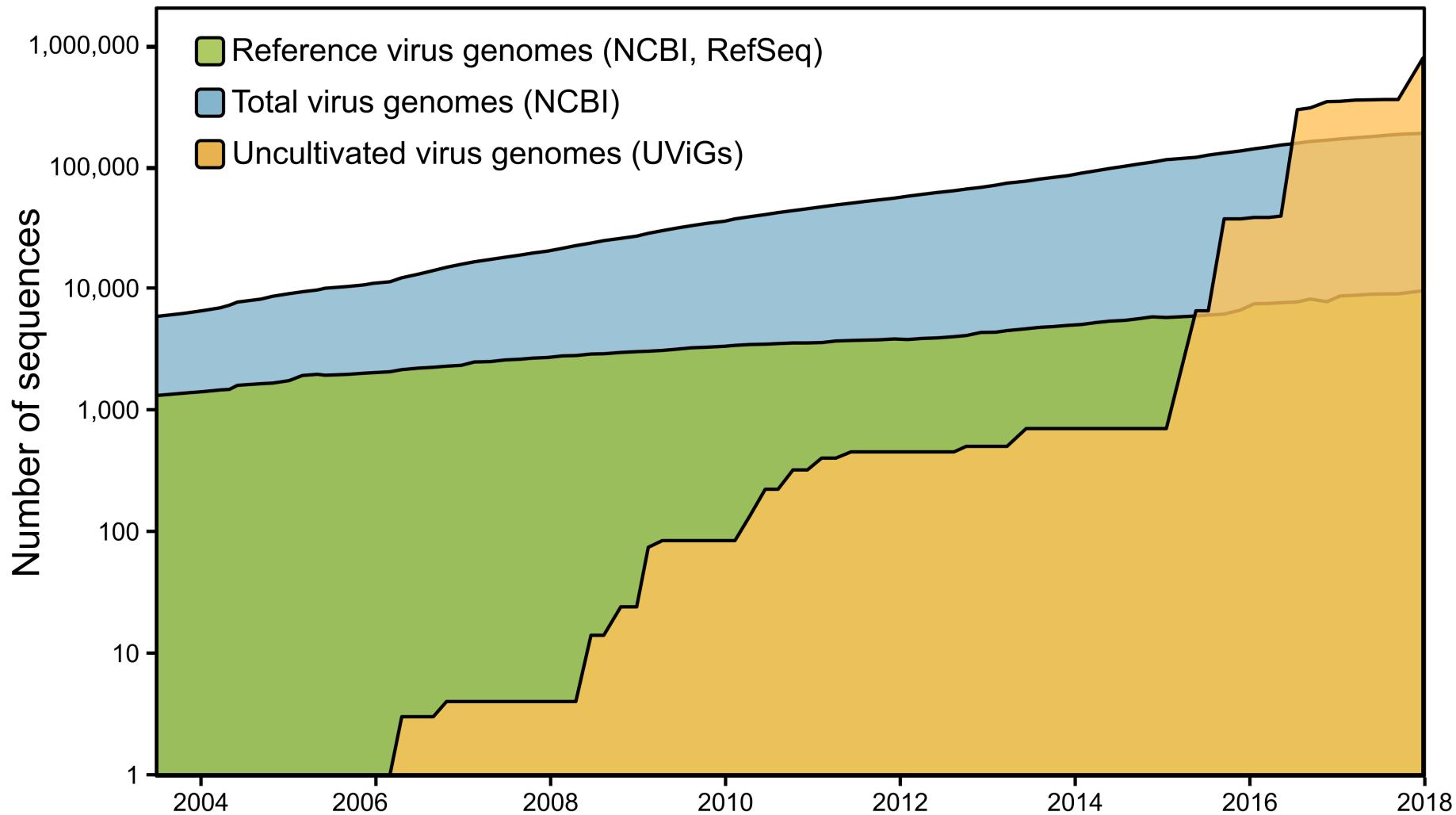
- Low coverage of viral diversity
- Genomes from uncultivated viruses
  - Leverage all types of datasets
  - Particularly useful: microbial and viral metagenomes



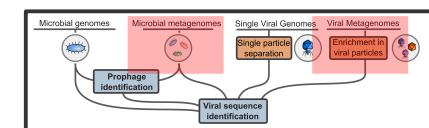
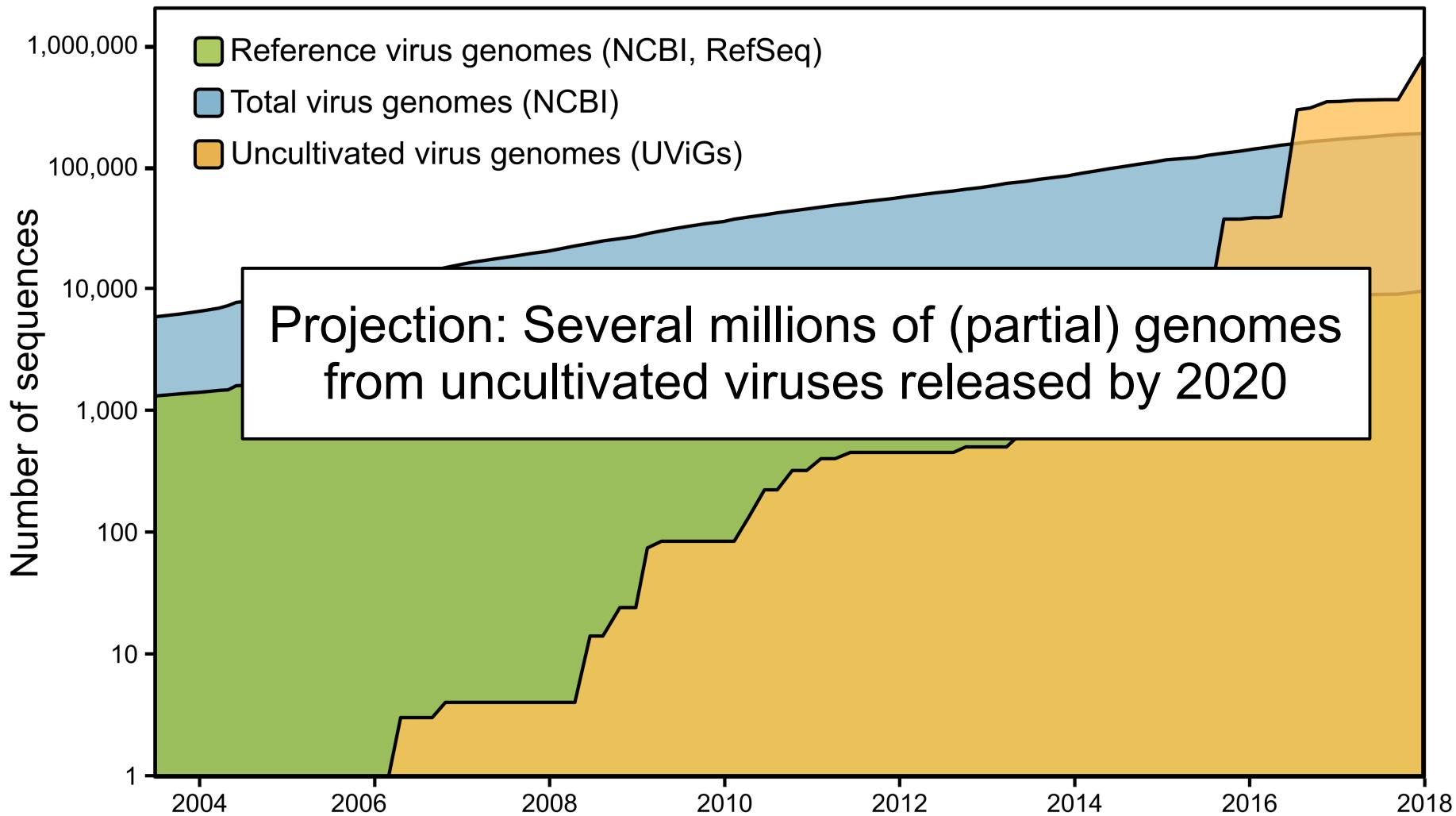
# A frantic pace of viral discovery



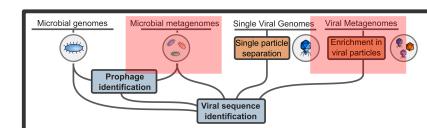
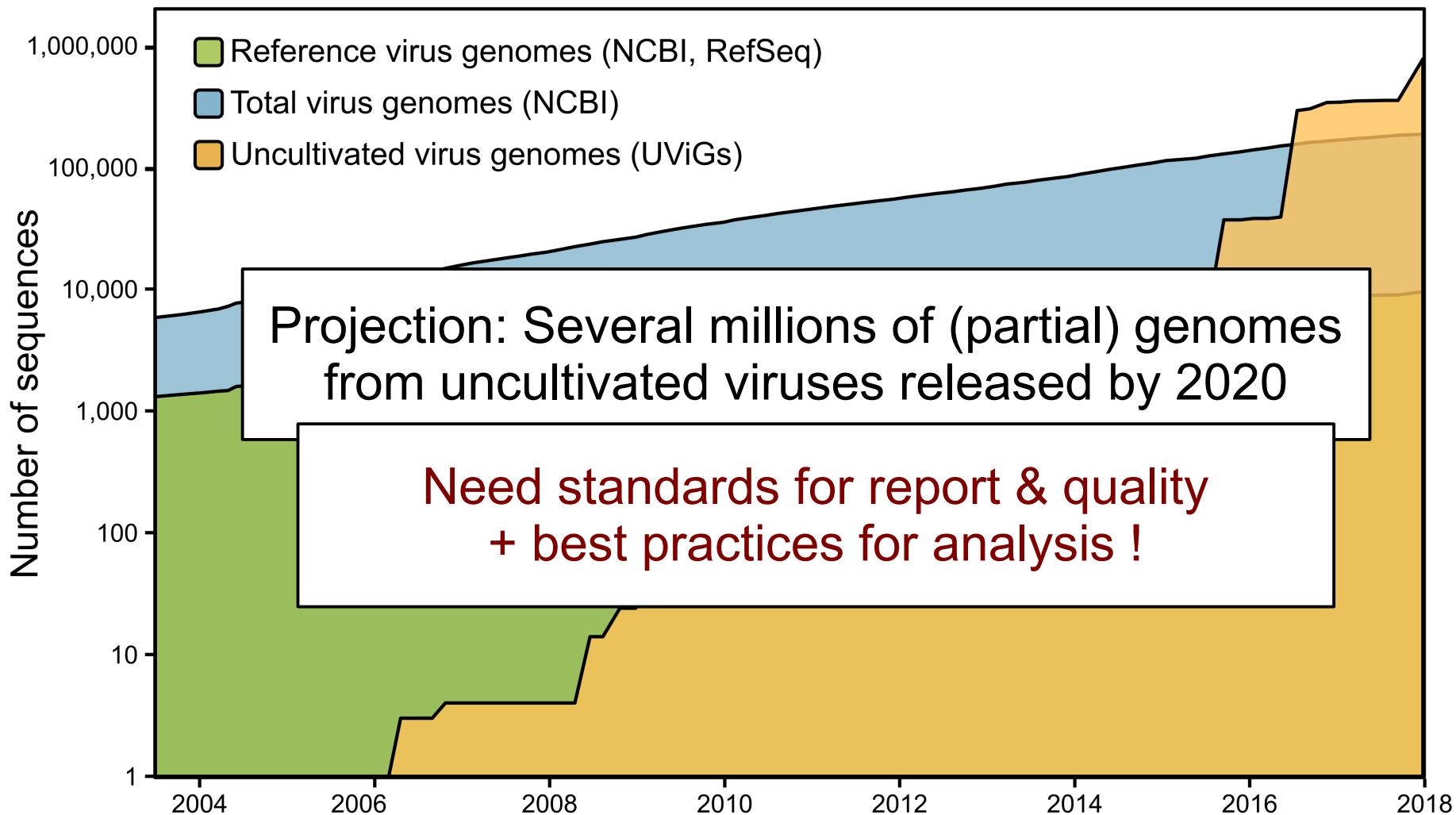
# A frantic pace of viral discovery



# A frantic pace of viral discovery



# A frantic pace of viral discovery



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

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

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

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

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

- **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

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

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

- **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**

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

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

- **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:**

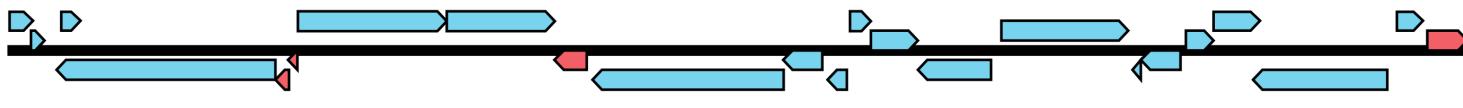
- Online (e.g. PHASTER, VirSorter), protocols (Earth's virome)
- Single underlying principle:
  - “viral signature gene(s) in a viral genome-like context”

# 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”

"Typical" microbial sequence



- Viral- or Virome-like gene
- PFAM affiliated gene
- Uncharacterized gene

# Viral contigs / prophage identification

- **Established tools:**

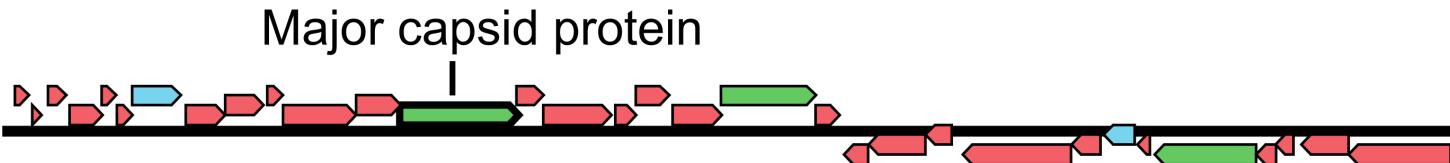
- Online (e.g. PHASTER, VirSorter), protocols (Earth's virome)
- Single underlying principle:
  - "viral signature gene(s) in a viral genome-like context"

"Typical" microbial sequence



- Viral- or Virome-like gene
- PFAM affiliated gene
- Uncharacterized gene

"Typical" viral sequence



# 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

Rob A. Edwards, Siobain Duffy, [show 4 more](#)

05/04/18

[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
- **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

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 ▾

# 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
- **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)

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 ▾

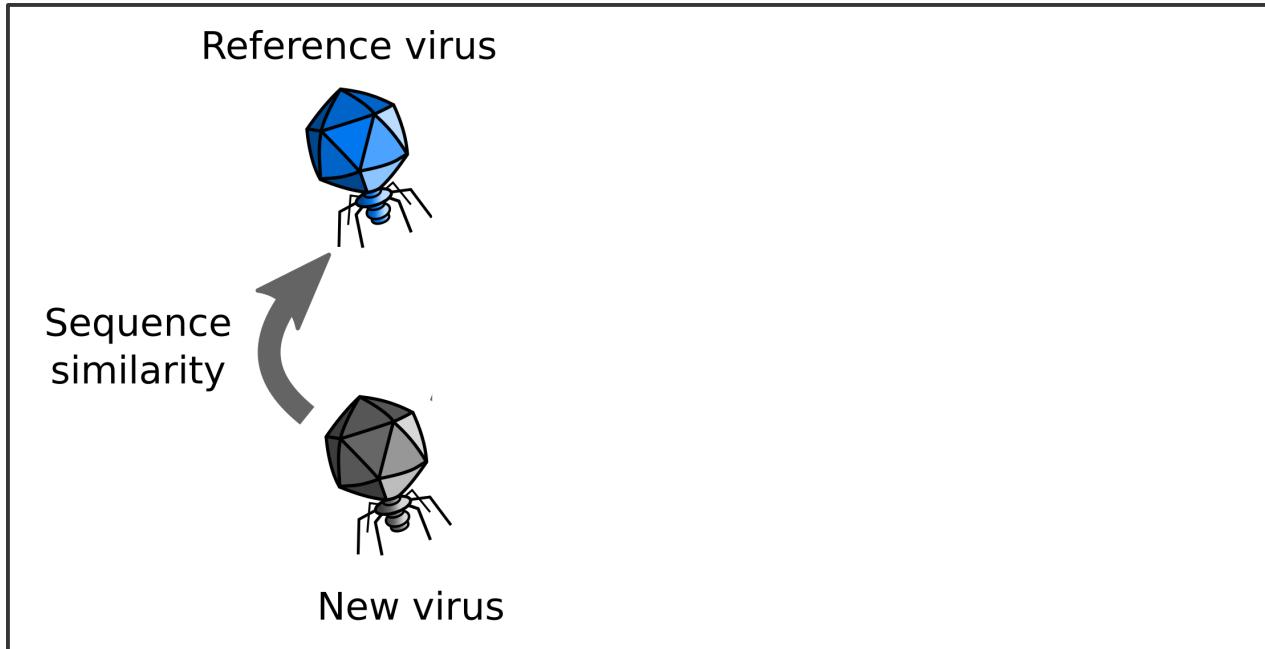
# In silico host assignment



- 4 main types of approaches

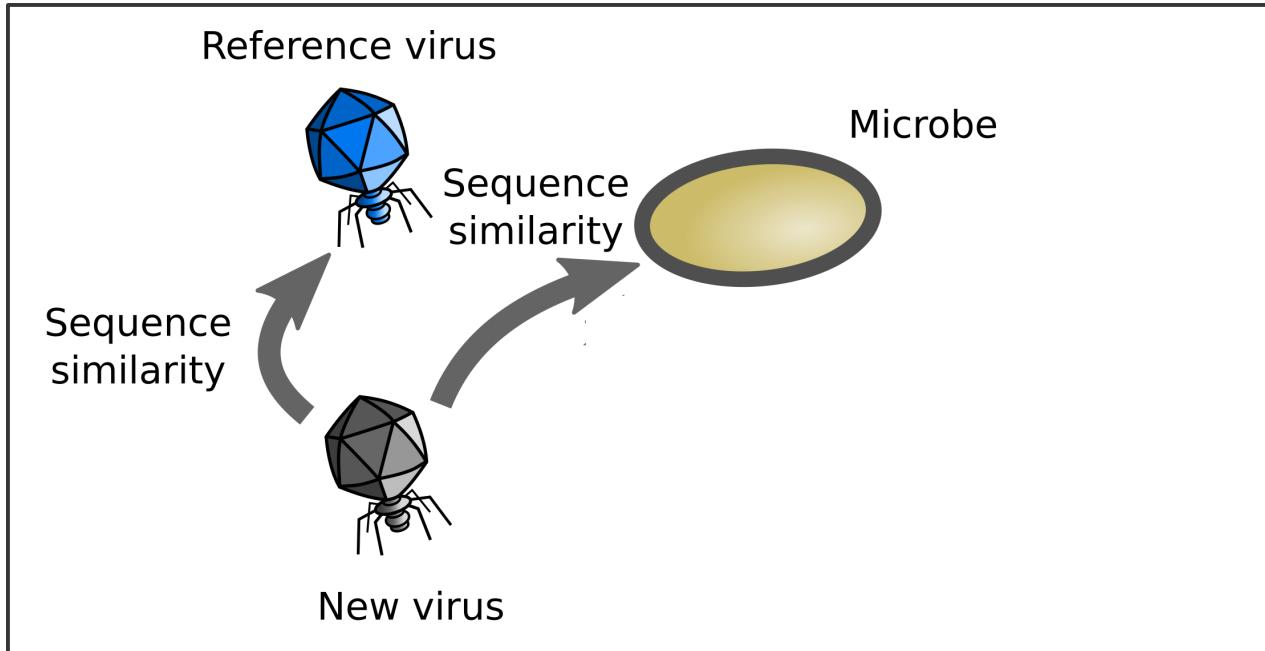
# In silico host assignment

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



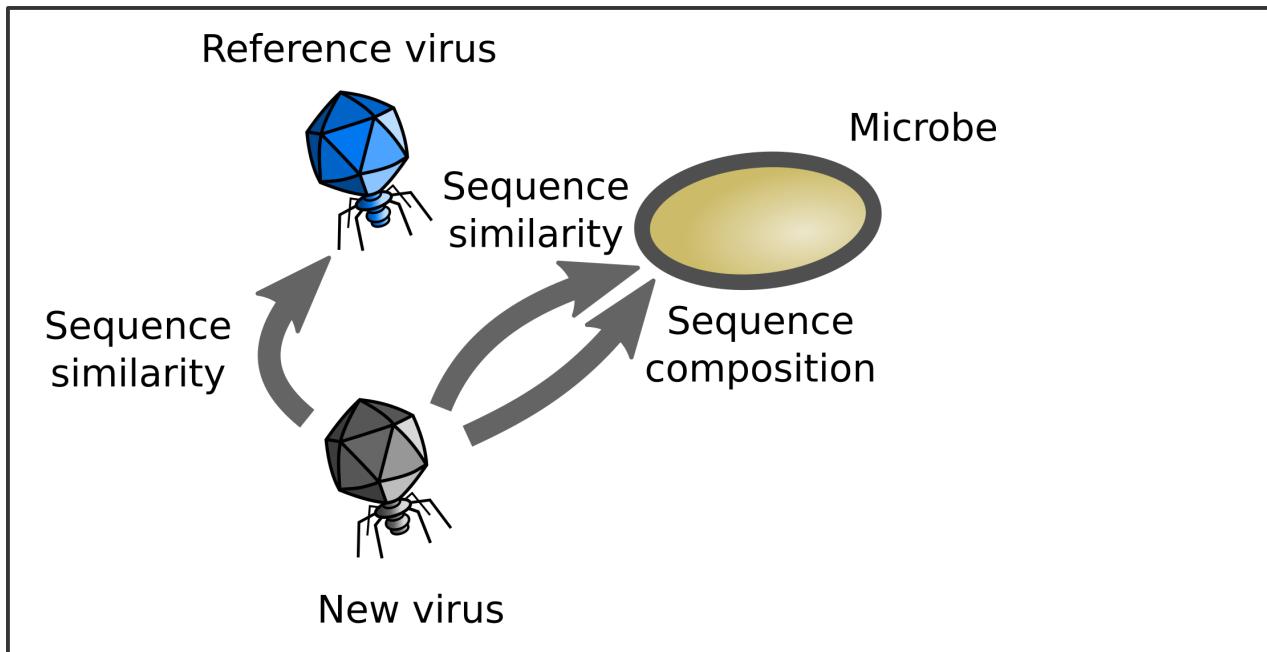
# In silico host assignment

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



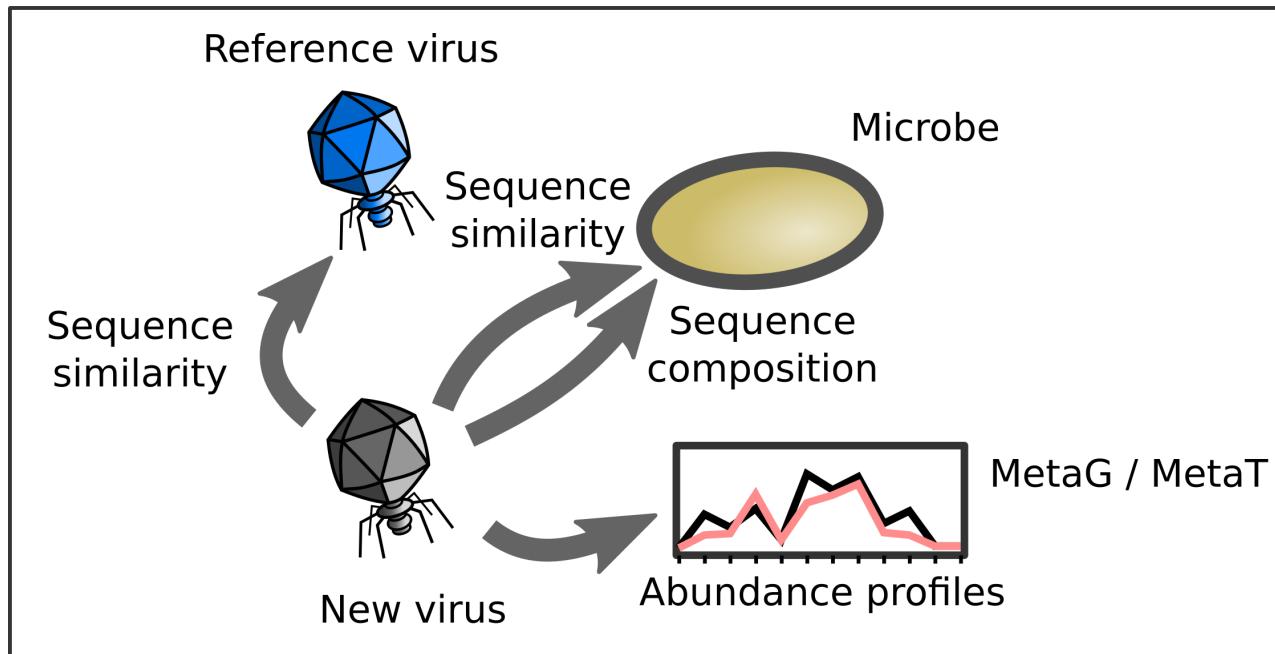
# In silico host assignment

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



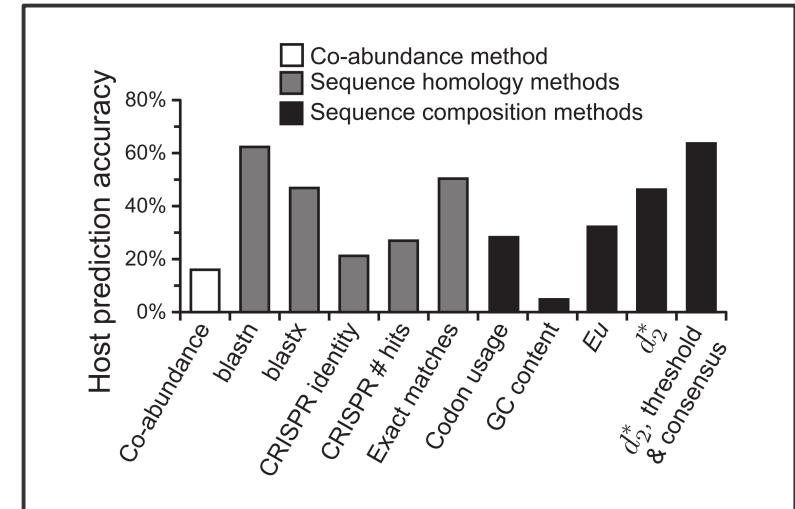
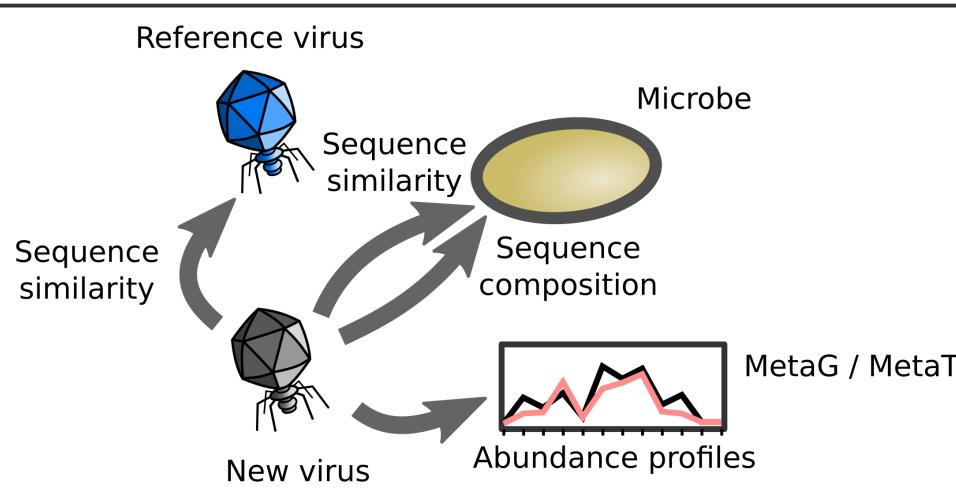
# In silico host assignment

- 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

- **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

- **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)

- **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:**

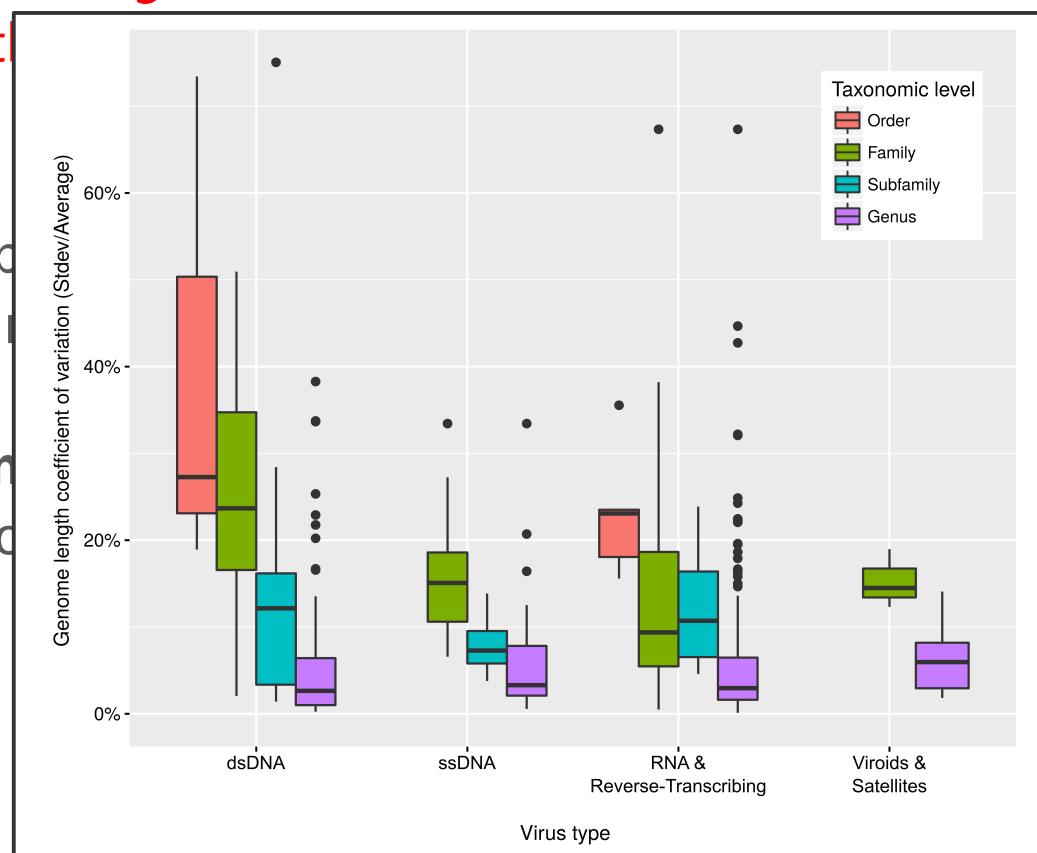
- Completeness (**single-copy marker genes**)
- Contamination (**single-copy marker genes**)
- Presence & number of rRNA / tRNA

- **Contamination of UViG**

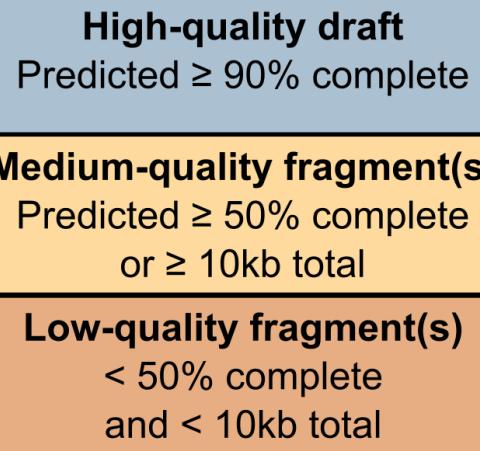
- Only possible for specific groups
- Mentioned as a possibility but not a problem

- **Main UViG metrics / requirements**

- identification of complete genome
- affiliation at ~ genus rank:
  - expected genome size

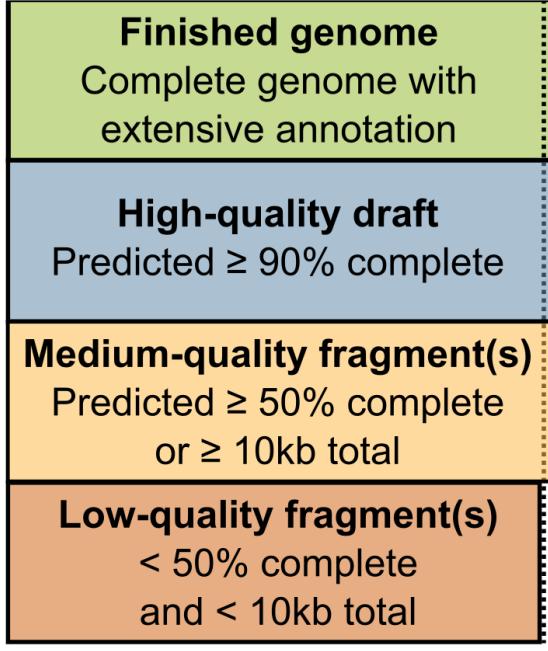


# UViG Quality estimation



- Classification in 4 categories

# UViG Quality estimation



- Classification in 4 categories

# UViG Quality estimation

Functional potential, host prediction	Biogeography, distribution	Novel taxonomic groups	Novel reference species
<b>Finished genome</b> Complete genome with extensive annotation			
<b>High-quality draft</b> Predicted $\geq 90\%$ complete			
<b>Medium-quality fragment(s)</b> Predicted $\geq 50\%$ complete or $\geq 10\text{kb}$ total			
<b>Low-quality fragment(s)</b> $< 50\%$ complete and $< 10\text{kb}$ total			

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

# UViG Quality estimation

- Example Global Ocean Virome
  - 15,222 contigs ( $\geq 10\text{kb}$ )

Functional potential, host prediction	Biogeography, distribution	Novel taxonomic groups	Novel reference species
<b>Finished genome</b> Complete genome with extensive annotation			
<b>High-quality draft</b> Predicted $\geq 90\%$ complete			
<b>Medium-quality fragment(s)</b> Predicted $\geq 50\%$ complete or $\geq 10\text{kb}$ total			
<b>Low-quality fragment(s)</b> $< 50\%$ complete and $< 10\text{kb}$ total			

# UViG Quality estimation

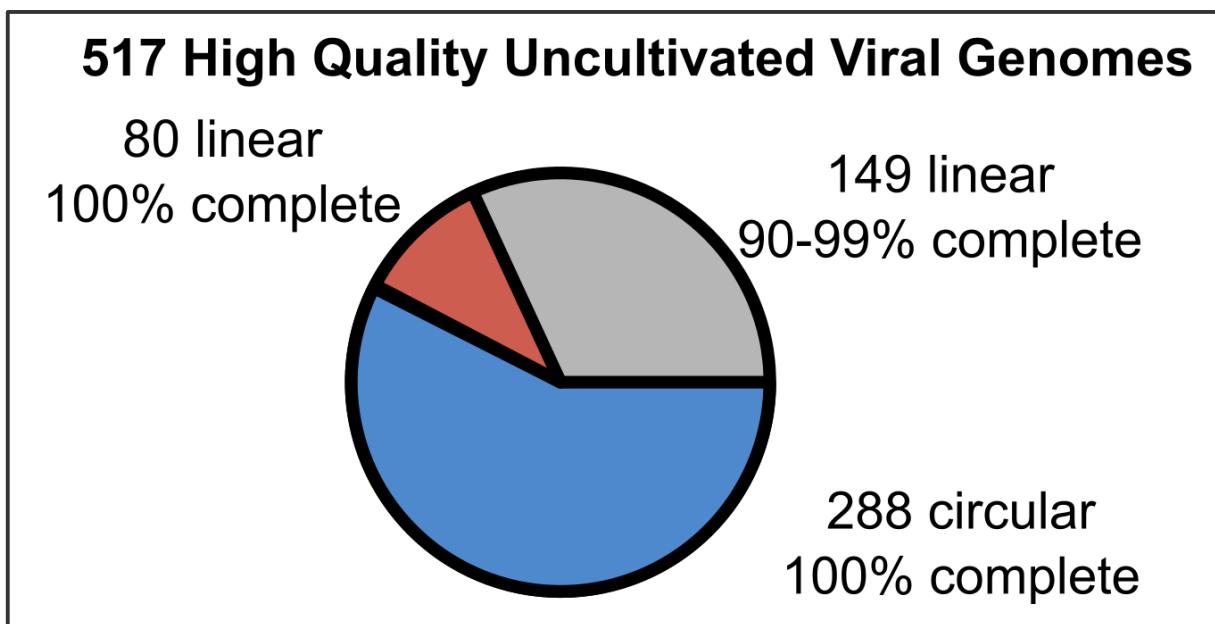
- **Example Global Ocean Virome**
  - 15,222 contigs ( $\geq 10\text{kb}$ )
  - Quality classification:
    - 0 low quality
    - 14,705 medium quality
    - 517 high quality

Functional potential, host prediction	Biogeography, distribution	Novel taxonomic groups	Novel reference species
<b>Finished genome</b> Complete genome with extensive annotation			
<b>High-quality draft</b> Predicted $\geq 90\%$ complete			
<b>Medium-quality fragment(s)</b> Predicted $\geq 50\%$ complete or $\geq 10\text{kb}$ total			
<b>Low-quality fragment(s)</b> $< 50\%$ complete and $< 10\text{kb}$ total			

# UViG Quality estimation

- **Example Global Ocean Virome**
  - 15,222 contigs ( $\geq 10\text{kb}$ )
  - Quality classification:
    - 0 low quality
    - 14,705 medium quality
    - 517 high quality

Functional potential, host prediction	Biogeography, distribution	Novel taxonomic groups	Novel reference species
<b>Finished genome</b> Complete genome with extensive annotation			
<b>High-quality draft</b> Predicted $\geq 90\%$ complete			
<b>Medium-quality fragment(s)</b> Predicted $\geq 50\%$ complete or $\geq 10\text{kb}$ total			
<b>Low-quality fragment(s)</b> $< 50\%$ complete and $< 10\text{kb}$ total			



# UViG Quality estimation

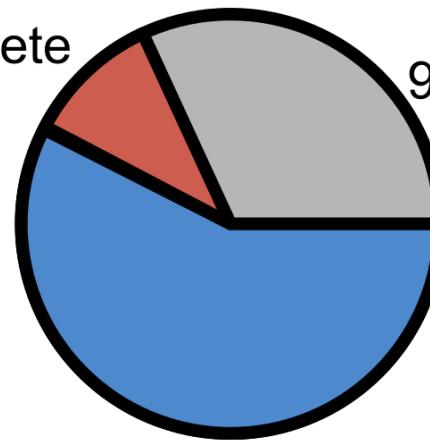
- **Example Global Ocean Virome**

- 15,222 contigs ( $\geq 10\text{kb}$ )
- Quality classification:
  - 0 low quality
  - 14,705 medium quality
  - 517 high quality
    - 3 new circular selected for manual annotation → finished genomes

Functional potential, host prediction	Biogeography, distribution	Novel taxonomic groups	Novel reference species
<b>Finished genome</b> Complete genome with extensive annotation			
<b>High-quality draft</b> Predicted $\geq 90\%$ complete			
<b>Medium-quality fragment(s)</b> Predicted $\geq 50\%$ complete or $\geq 10\text{kb}$ total			
<b>Low-quality fragment(s)</b> $< 50\%$ complete and $< 10\text{kb}$ total			

## 517 High Quality Uncultivated Viral Genomes

80 linear  
100% complete



149 linear  
90-99% complete

288 circular  
100% complete

# 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

The screenshot shows the JGI/IMG/VR interface. At the top, there's a search bar labeled "Quick Genome Search" and a "Go" button. Below the header, a navigation bar includes links for Home, Find Genomes, Find Genes, Find Functions, Compare Genomes, My IMG, Data Marts, and Help. A message box states: "The IMG/VR system (<http://nar.oxfordjournals.org/content/early/2016/10/30/nar.gkw1030>) serves as a starting point for the sequence analysis of viral fragments derived from metagenomic samples. Virus detection methods and host assignment approaches in IMG/VR are fully described in Paez-Espino et al. *Nature*, 2016 ["Uncovering Earth's virome"](#)".

The main content area displays "IMG Viral Content" with sections for "Viral Datasets", "Viral Clusters", and "With Host". Each section provides counts for Isolate Viruses (IVGs), Metagenomic Viral Contigs (mVCs), and Total Viral Datasets/Clusters/Singletons.

Below this, there's a "Ecosystems" section with a "Show Human Body Sites" button. A note states: "712115 viral scaffolds. Some projects maybe rejected via Google Maps because of bad location coordinates. Map pins represent location counts. Some pins may have multiple genomes. Map pins are grouped into clusters and clusters themselves into larger clusters."

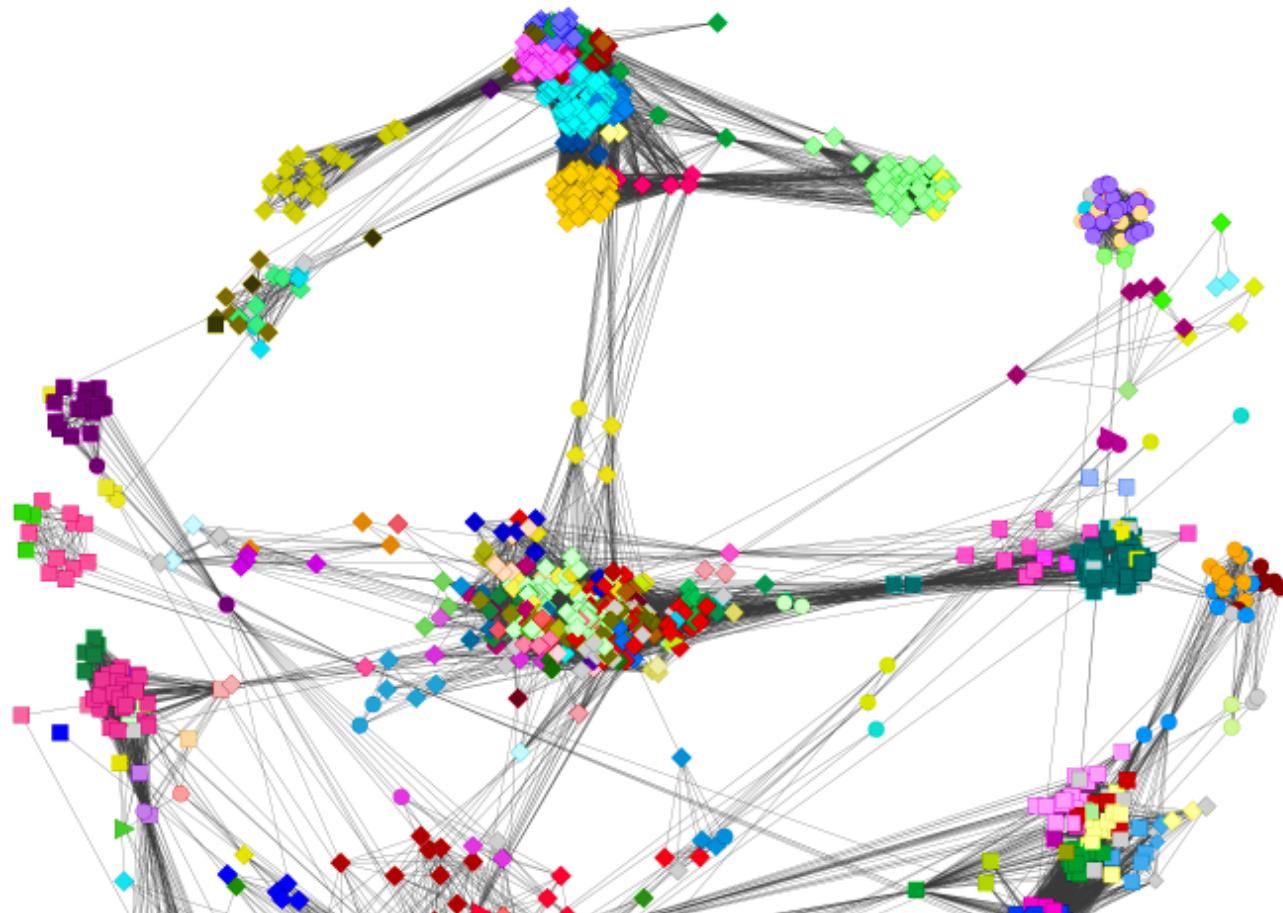
A "hint" box at the bottom left explains: "For any given genome at a location on the map, you may access the list of scaffolds that belong to a virus by clicking on a map pin and selecting the count next to the genome of interest for that location. The total count of viral scaffolds for a location is displayed in the label and tooltip of a map pin e.g. Arctic Ocean [3]."

# 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 ?

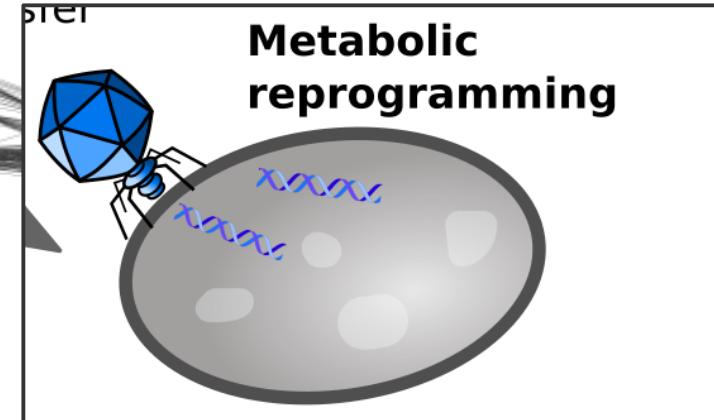
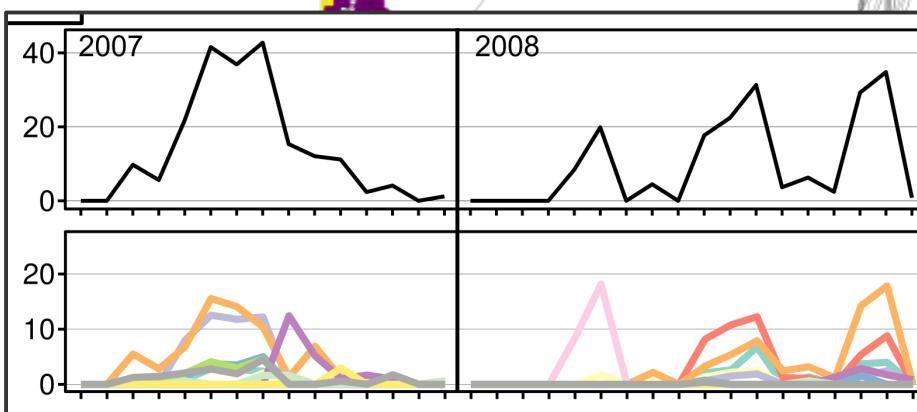
**End goal: comprehensive (host-resolved) mapping of the virosphere**



# ... 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



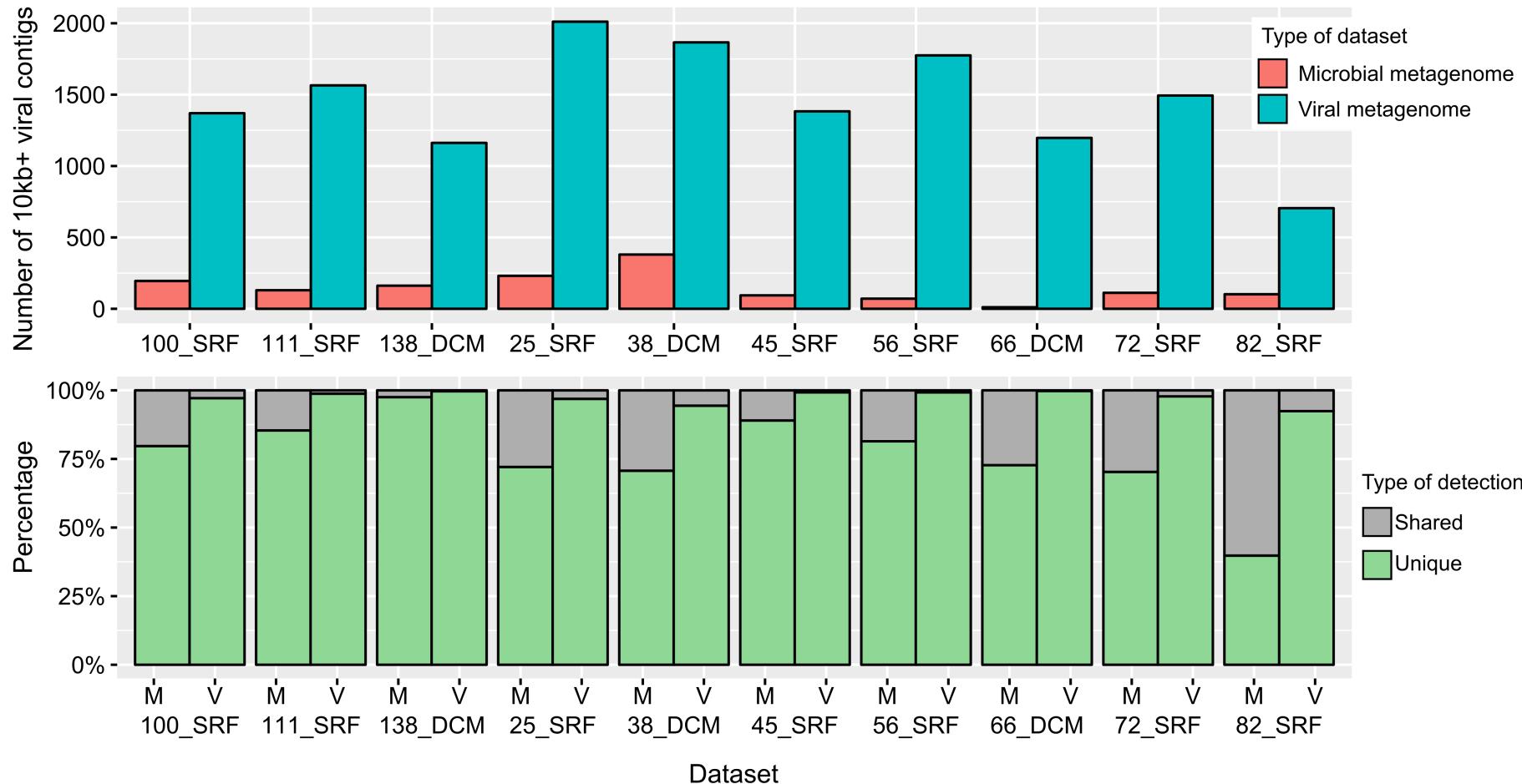
- **Microbial manipulation**
  - host cell takeover mechanisms
  - biotechnological applications

# Thanks

**> 50 contributing labs !**

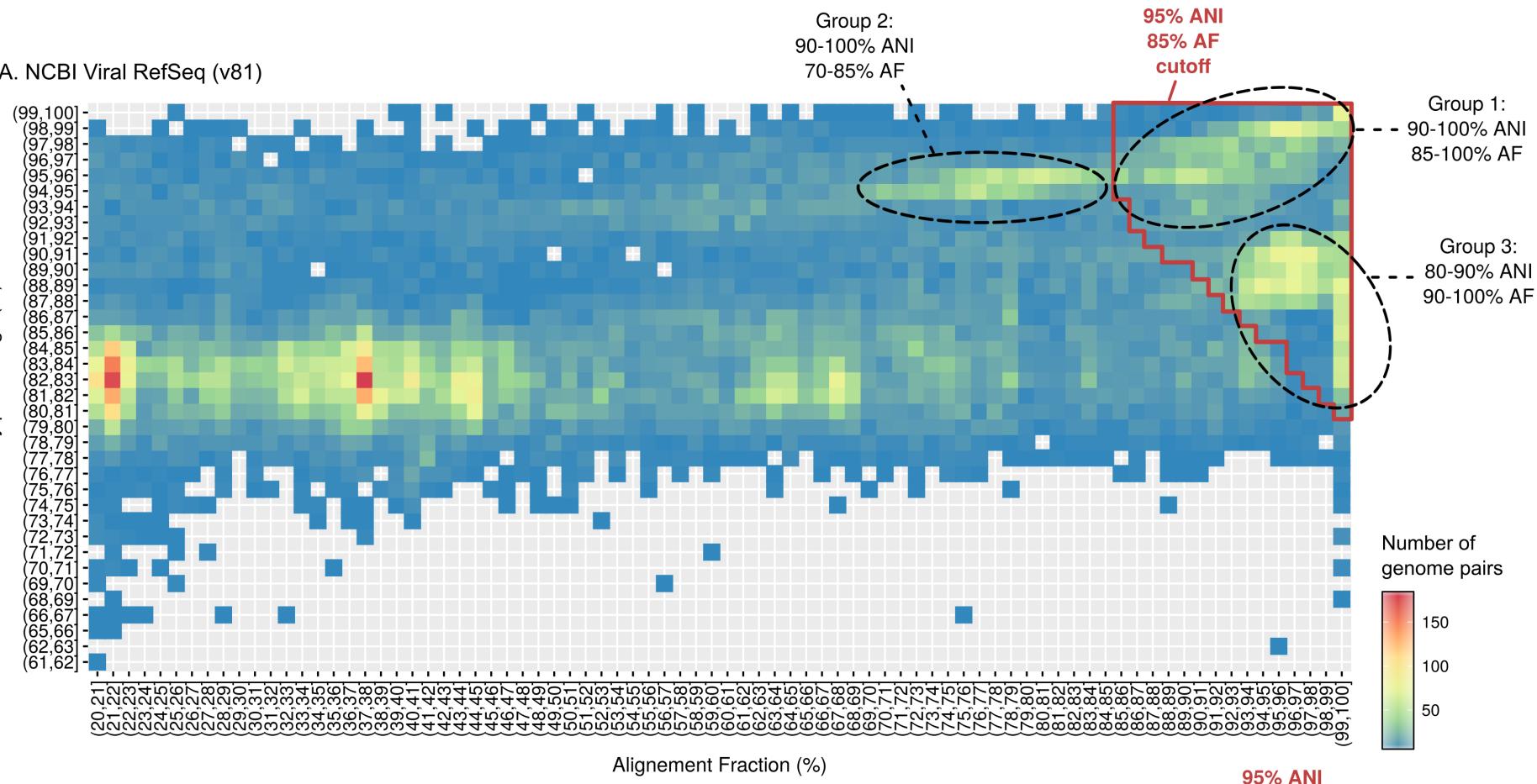


# Targeted vs untargeted



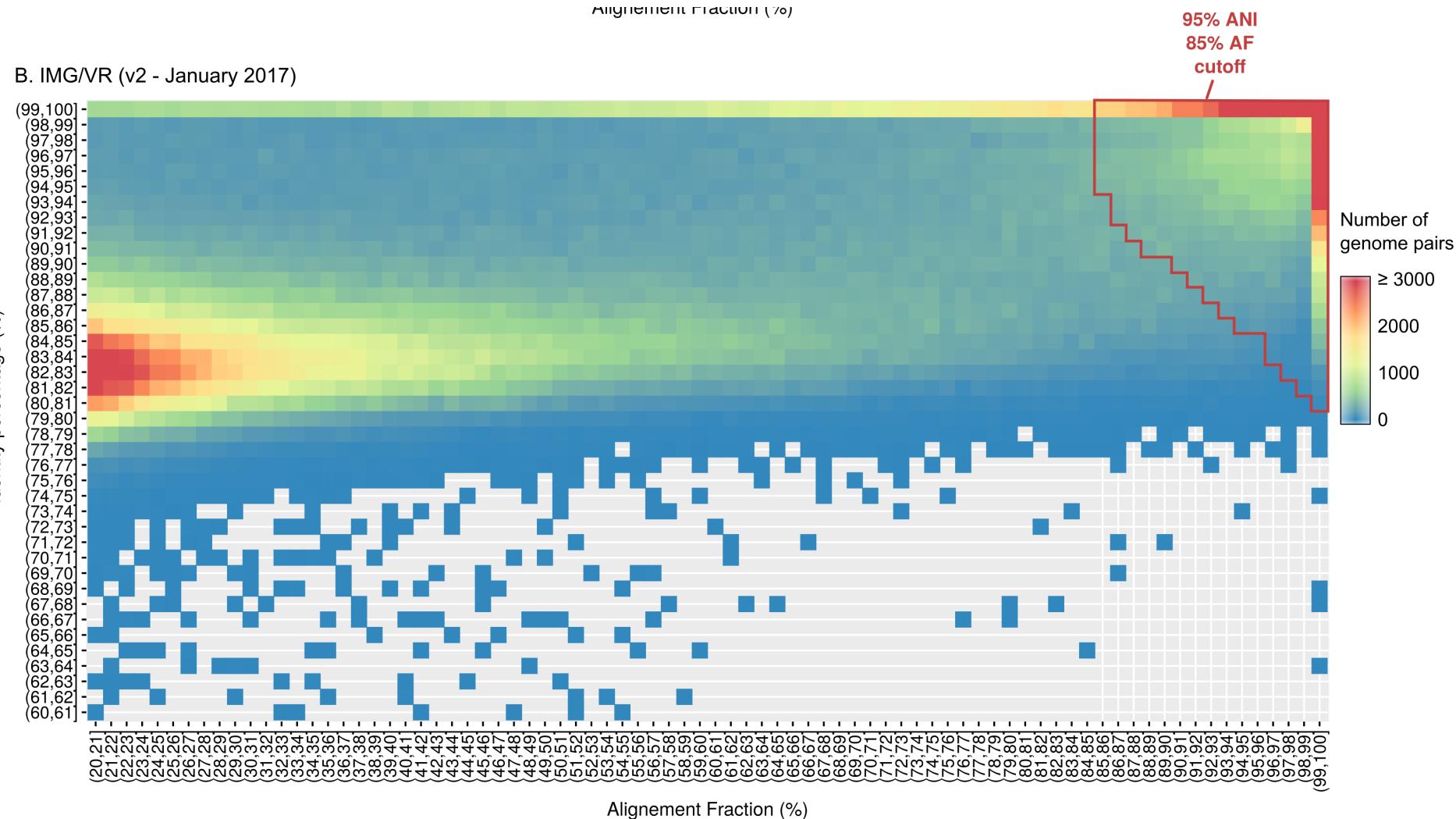
# vOTU threshold

A. NCBI Viral RefSeq (v81)



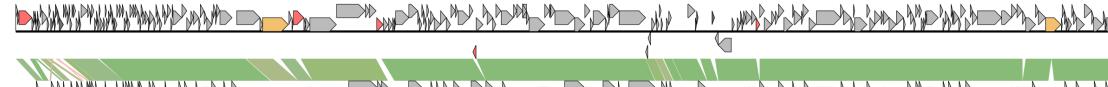
# vOTU threshold

B. IMG/VR (v2 - January 2017)



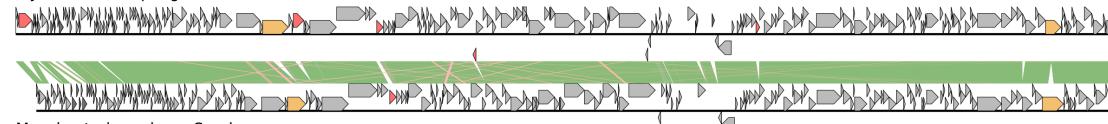
# vOTU threshold

Mycobacterium phage Gizmo      98.8% ANI, 96.6% AF, 95.5% wgANI      88% 100%



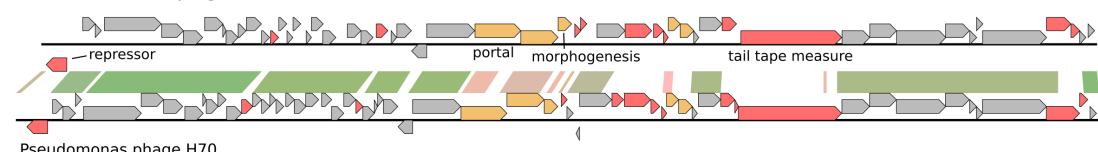
Amino acid similarity (tblastx)

Mycobacterium phage Gizmo



Mycobacterium phage Spud

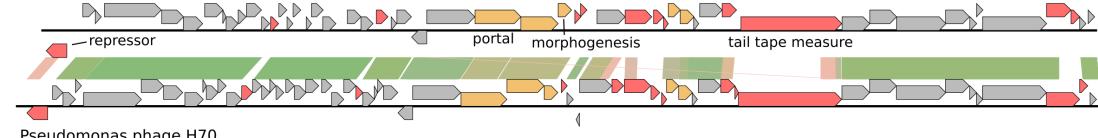
.0% ANI, 71.9% AF, 67.7% wgANI      10 Kbp      blastn



Pseudomonas phage H70

Amino acid similarity (tblastx)

Pseudomonas phage DMS3



Pseudomonas phage H70

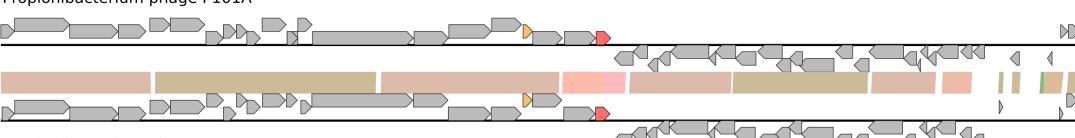
Nucleotide similarity (blastn)

Propionibacterium phage P101A

86.68% ANI, 95.48% AF, 82.7% wgANI

10 Kbp      blastn

82% 100%

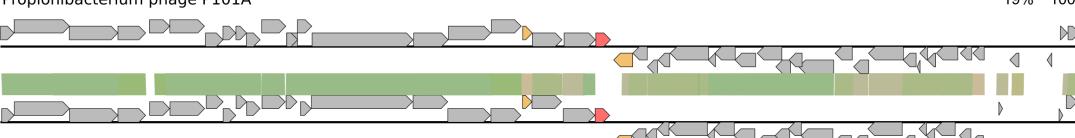


Amino acid similarity (tblastx)

Propionibacterium phage P101A

tblastx

19% 100%



Propionibacterium phage PHL010M04