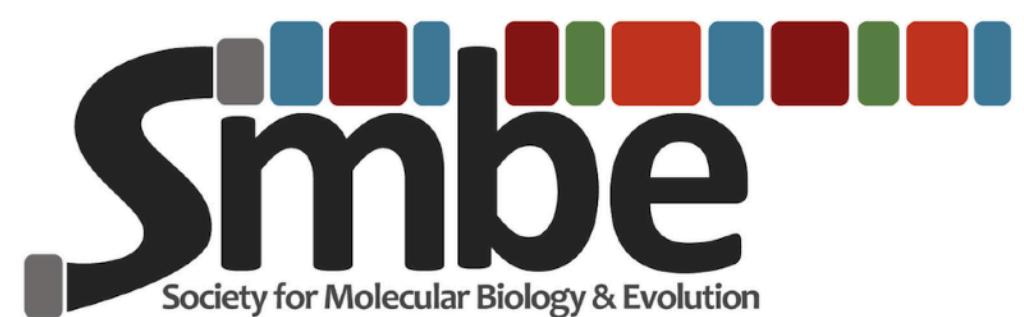


Metagenomics

Mark Stenglein



SMBE Regional Workshop
on Computational Biology
in Todos Santos, Mexico

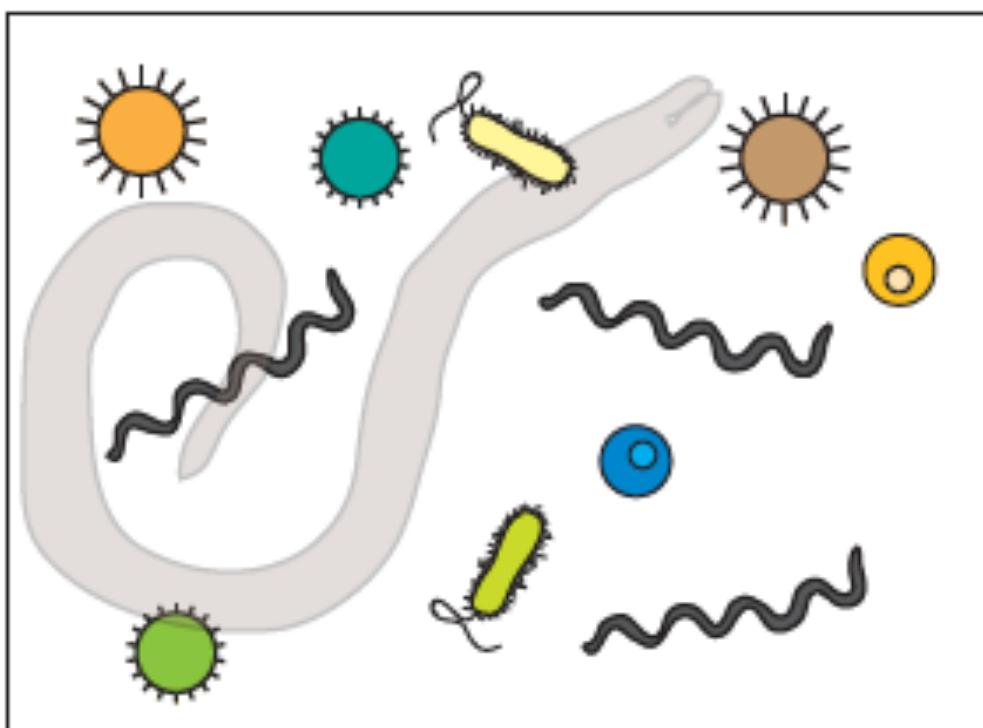


Colorado State University Todos Santos Center
April 8-12, 2019

Metagenomics is the study of >1 genome

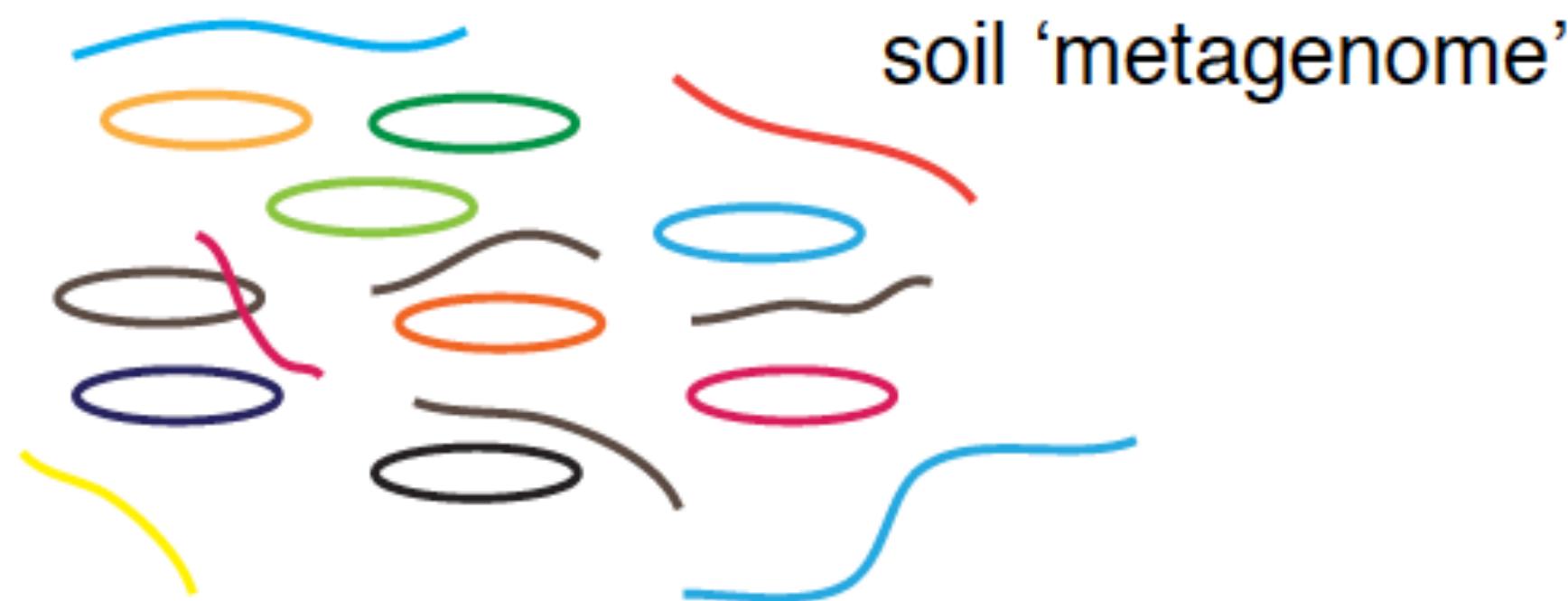
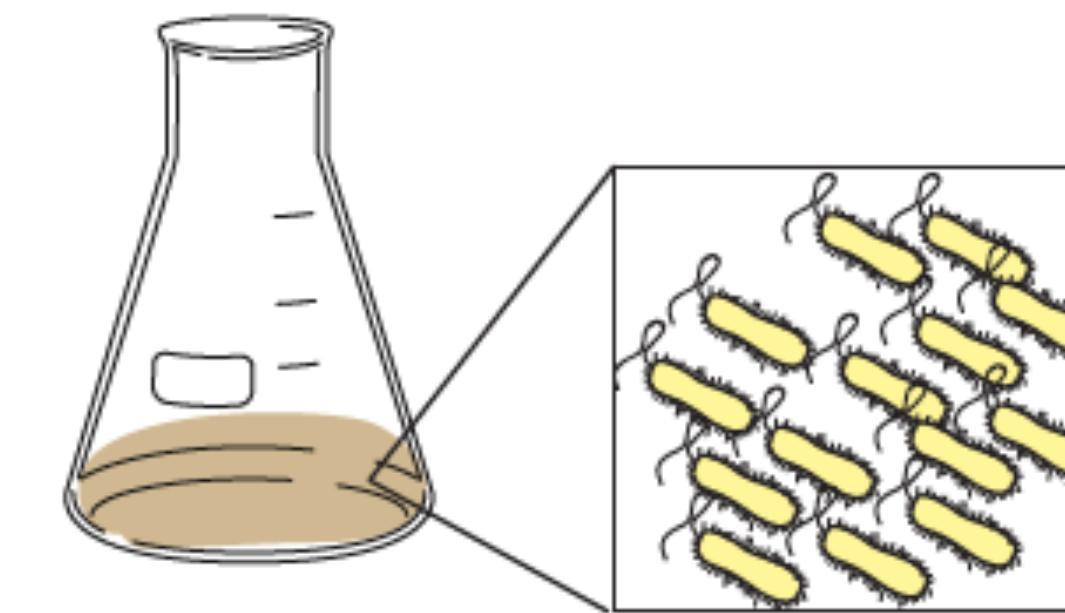
Many genomes

soil community

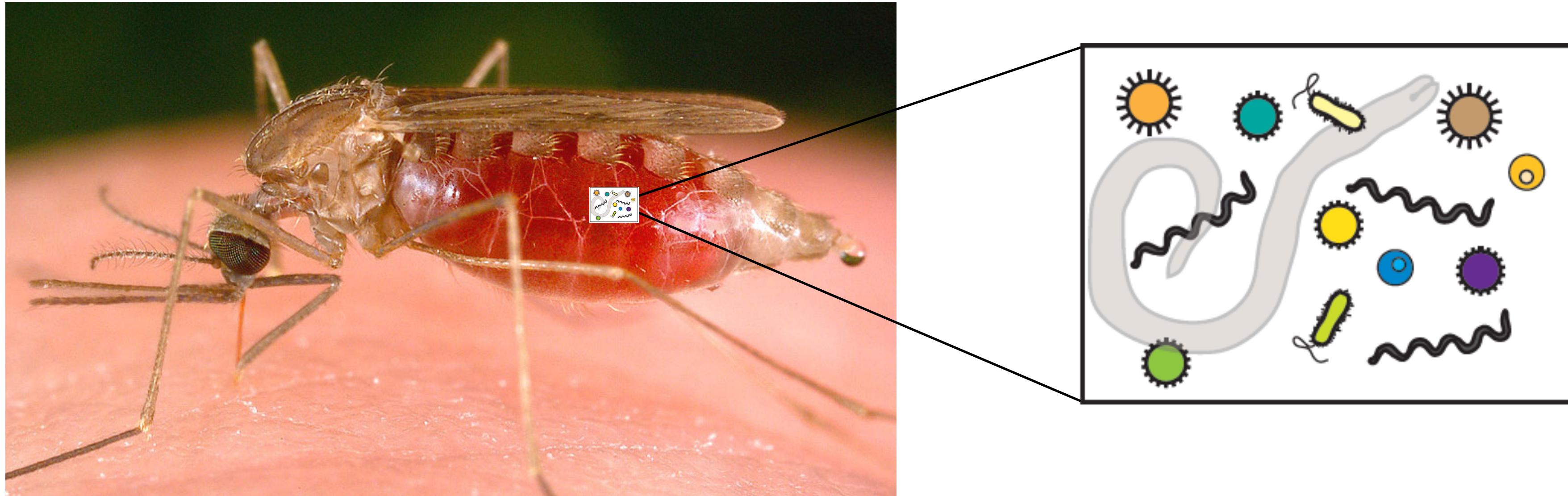


1 genome

bacterial isolate



If you sequence total nucleic acid from an intact multicellular organism,
you are doing metagenomics



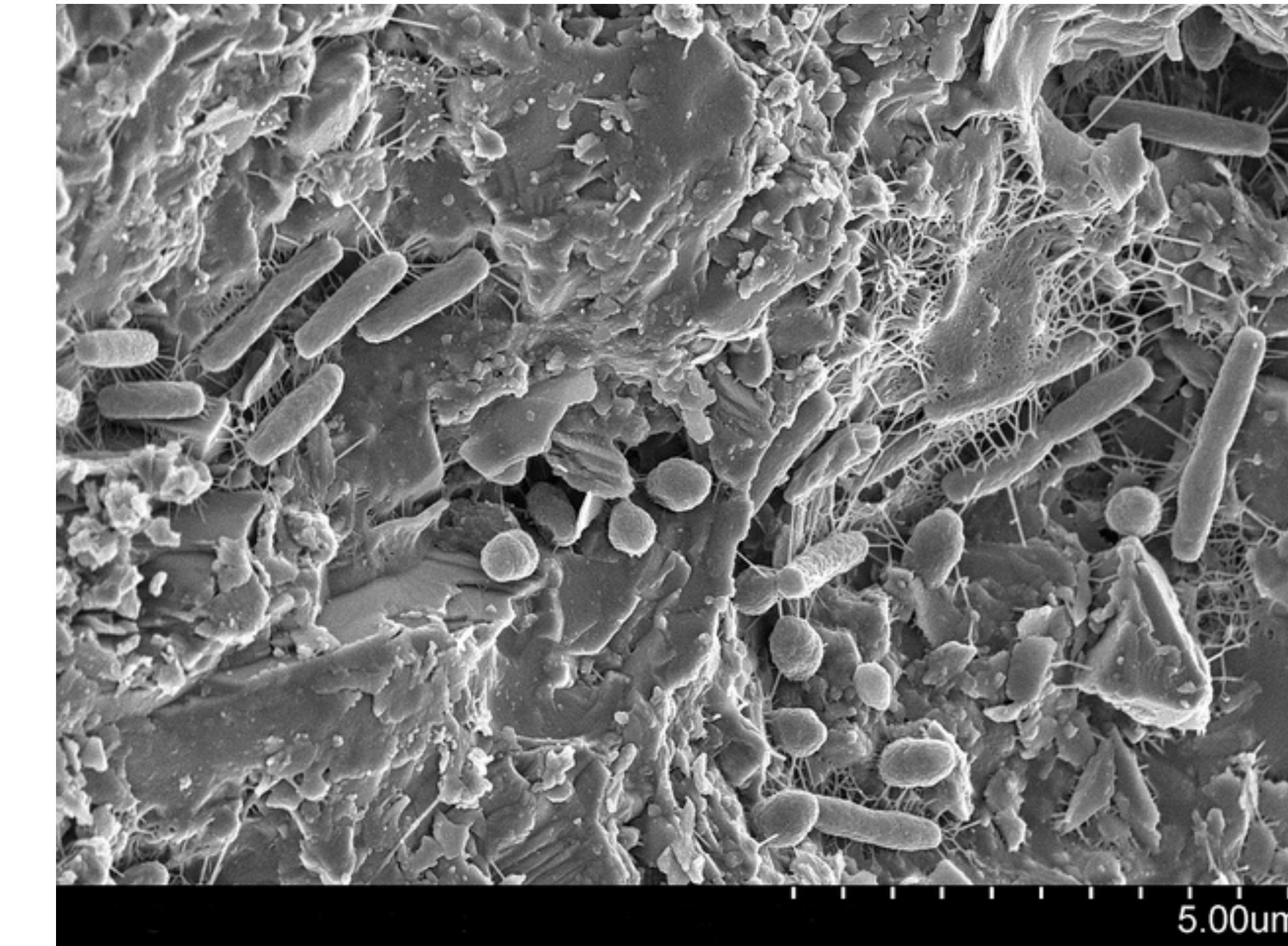
Metagenomics emerged in response to the observation that most micro-organisms can't be cultured



Morphological diversity typical of microorganisms cultured from soil on a broad spectrum medium, tryptic soy agar.

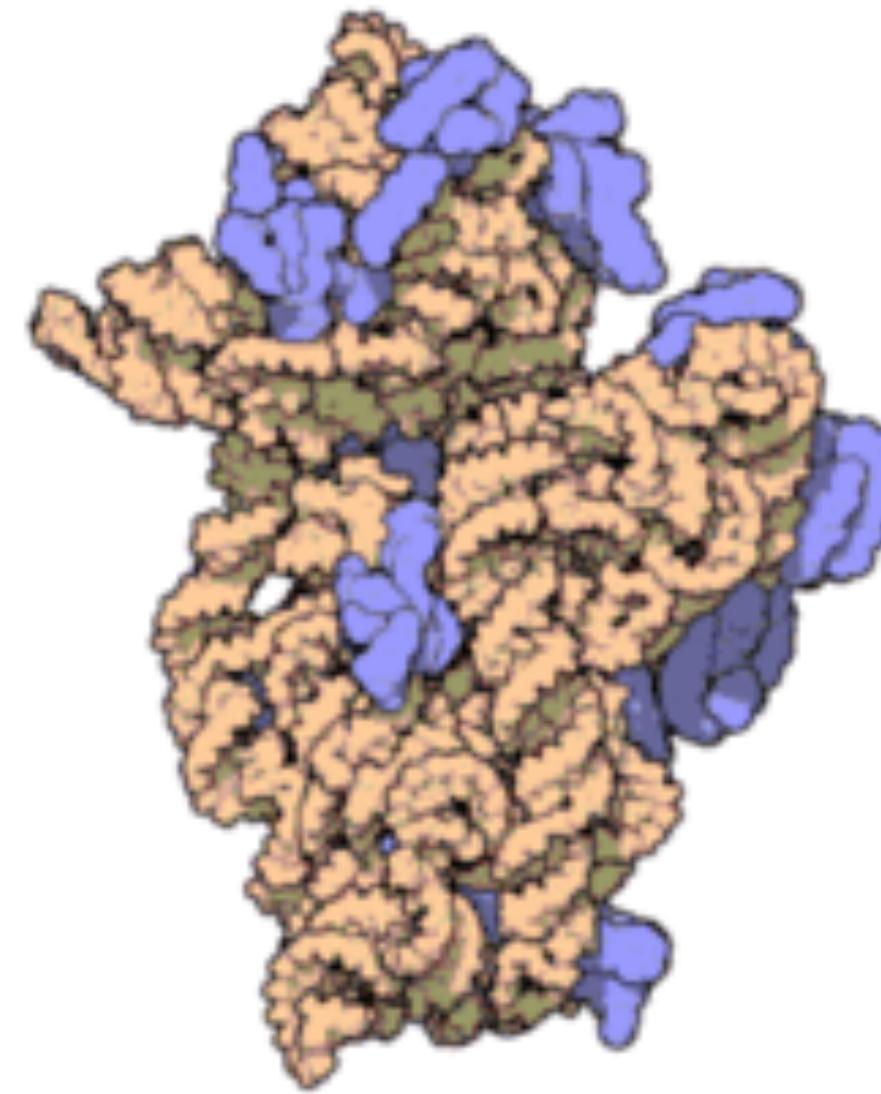
Handelsman et al (1998) Chem & Biol

Estimated: 10^8 bacteria per gram of soil of
6000-8000 different species
Only ~1% culturable (?)



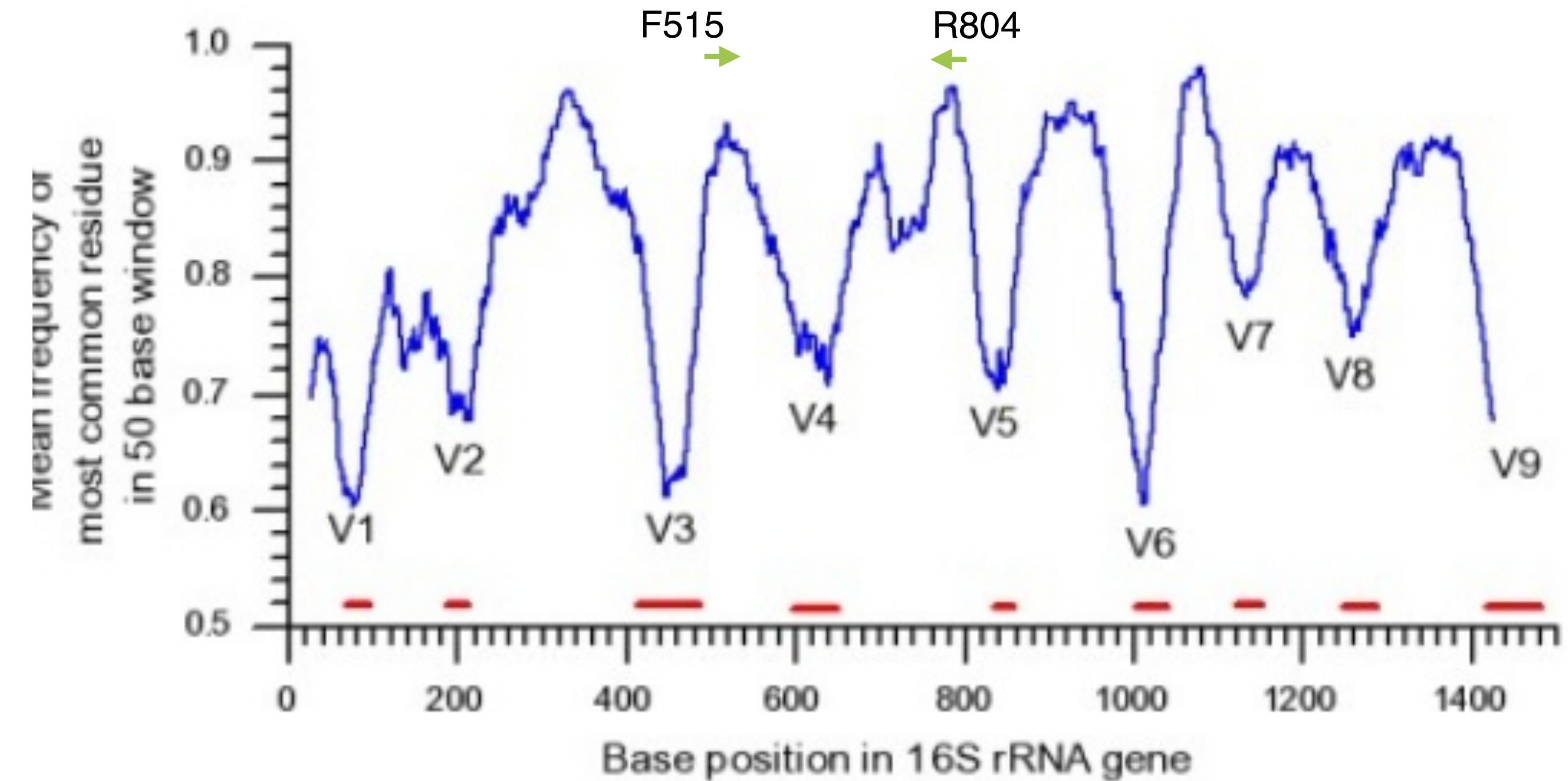
EM: Kim Lewis, Northeastern Univ.

PCR using primers targeting conserved regions of the 16S rRNA gene and sequencing enabled genotyping of bacteria and archaea without having to culture them



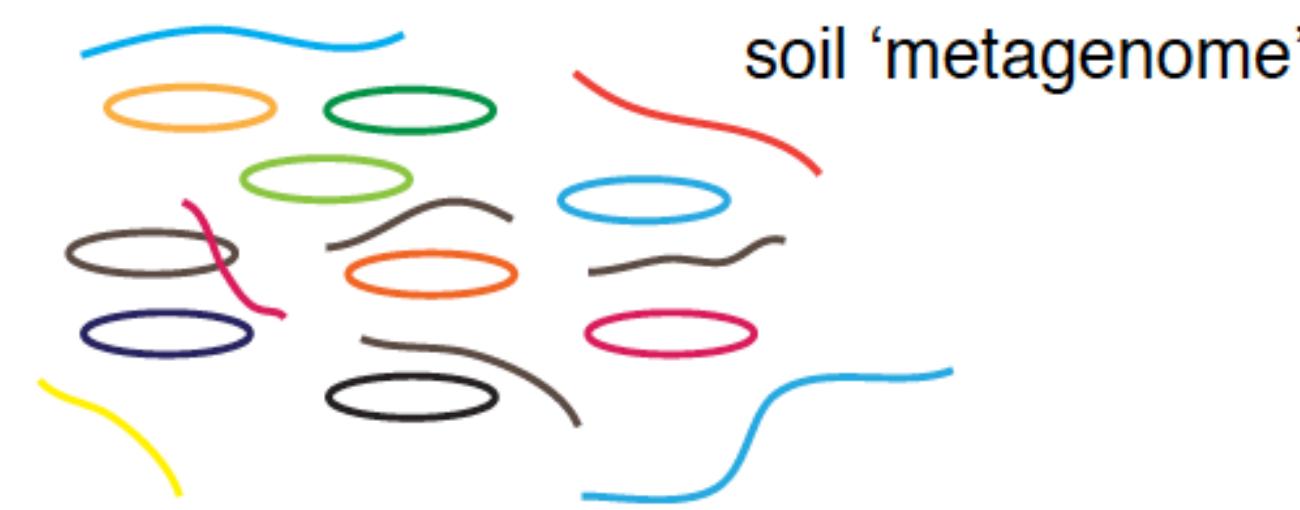
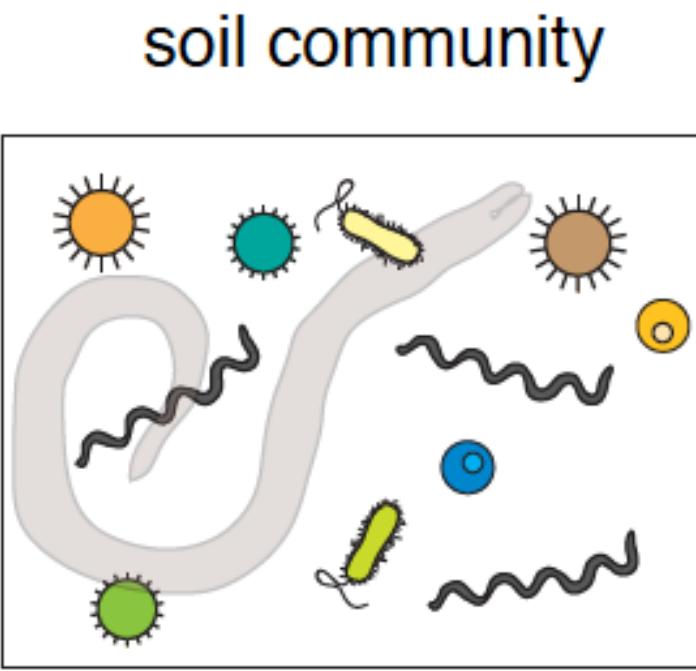
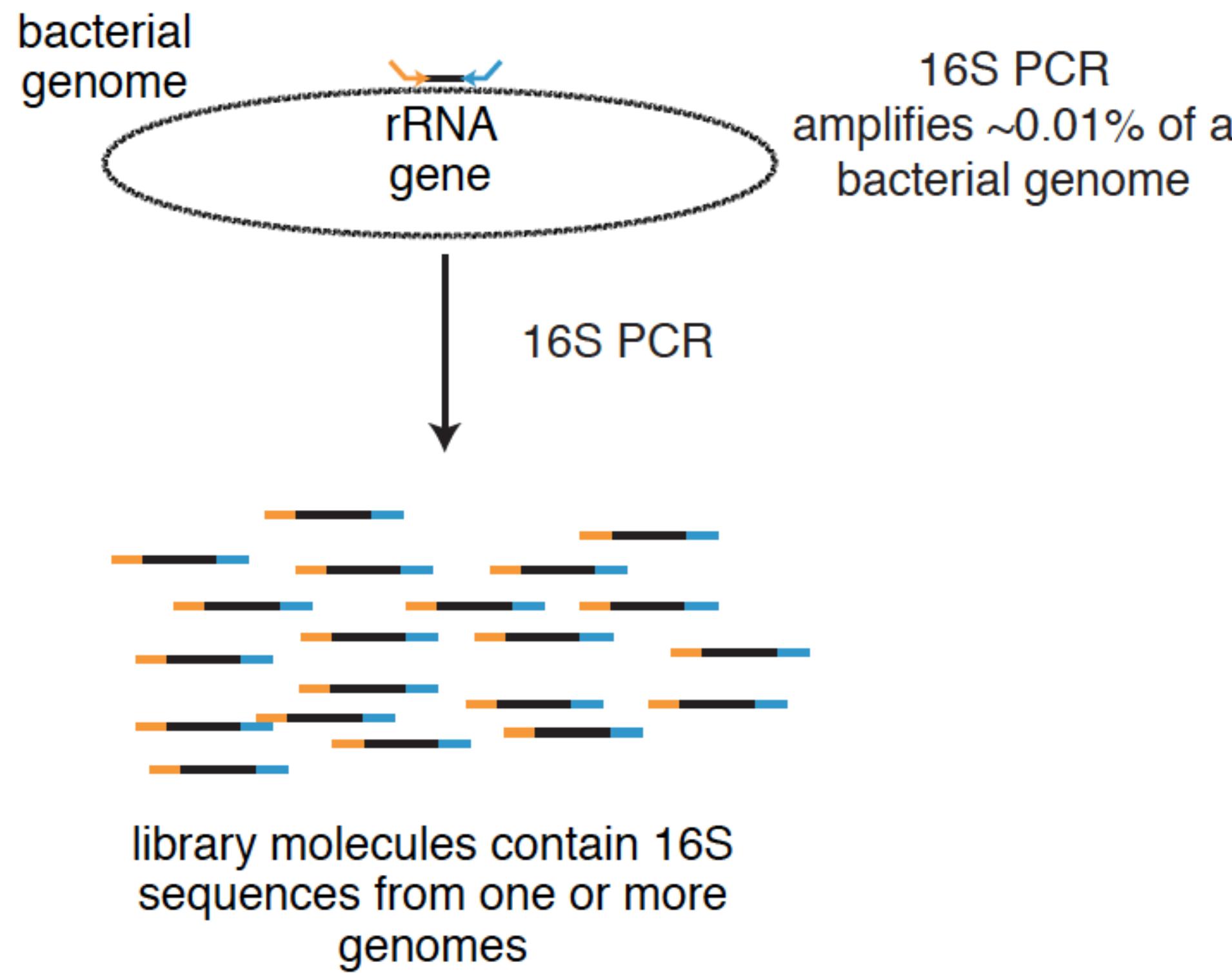
bacterial 30S ribosomal subunit
16S rRNA is in orange
(purple: ribosomal proteins)

image: wikipedia



— hypervariable

16S sequencing vs. shotgun metagenomics



- Only bacteria and archaea surveyed
- Deeper sampling of bacterial diversity per \$
- Relatively easy to make libraries and interpret results
- Appropriate if all you care about is microbial diversity / ecology

- All organisms studied*
- Decreased sampling depth per \$
- Enables analysis of other genomic features of organisms, e.g. antimicrobial resistant genes
- Analysis is significantly more difficult

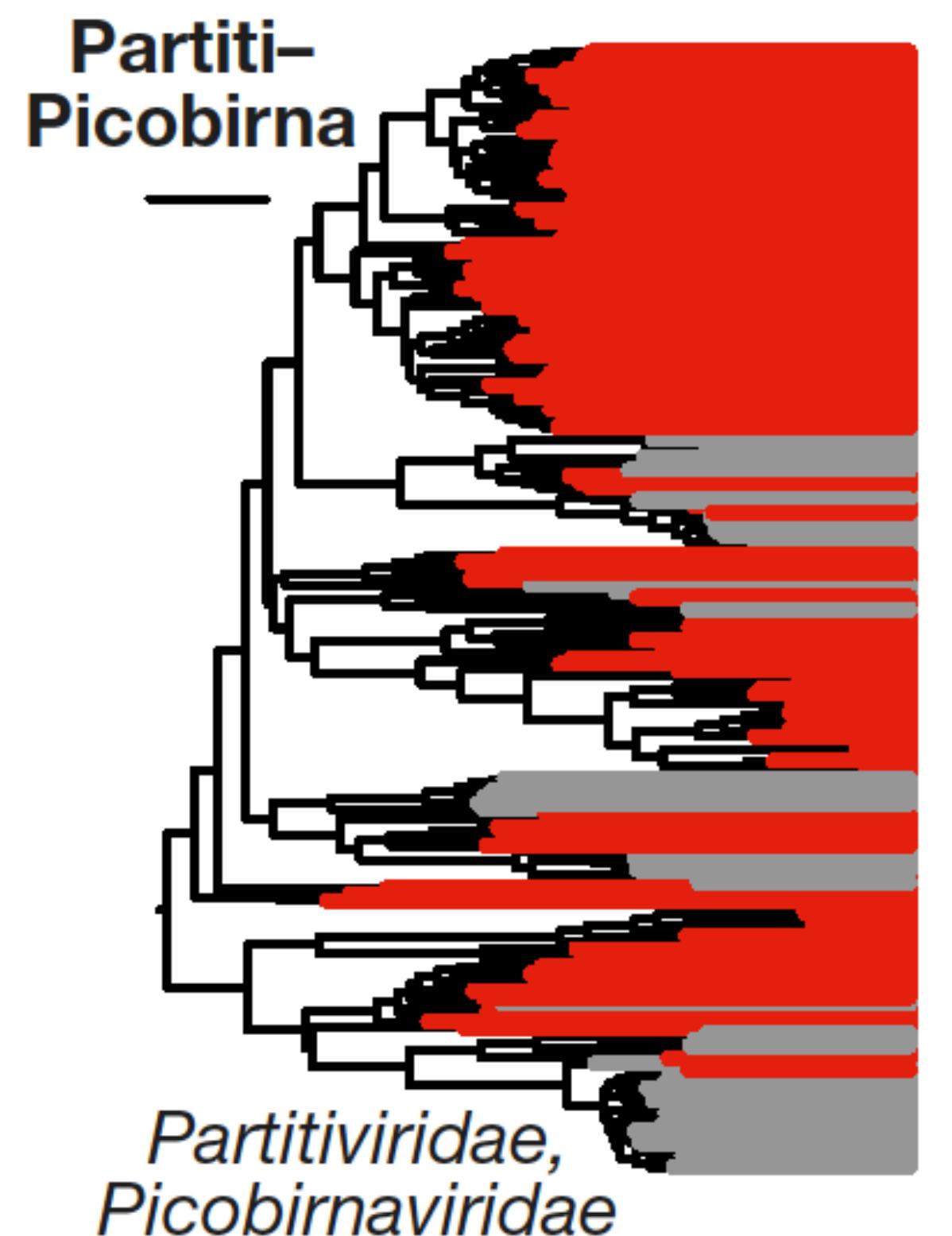
Metagenomics is leading to an explosion in discovery of new genome sequences

Redefining the invertebrate RNA virosphere

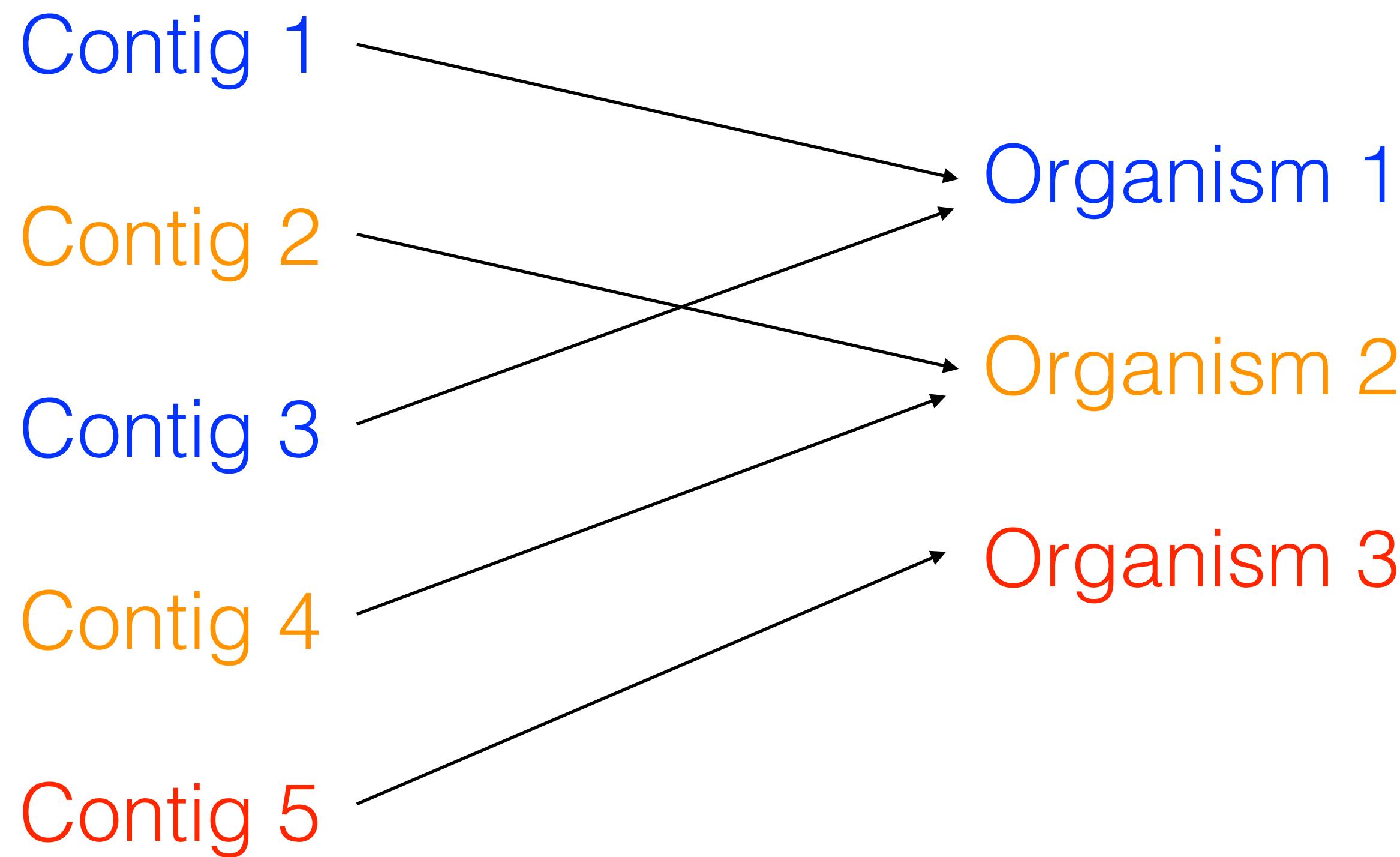
Mang Shi^{1,2*}, Xian-Dan Lin^{3*}, Jun-Hua Tian^{4*}, Liang-Jun Chen^{1*}, Xiao Chen^{5*}, Ci-Xiu Li^{1*}, Xin-Cheng Qin¹, Jun Li⁶, Jian-Ping Cao⁷, John-Sebastian Eden², Jan Buchmann², Wen Wang¹, Jianguo Xu¹, Edward C. Holmes^{1,2} & Yong-Zhen Zhang¹

Current knowledge of RNA virus biodiversity is both biased and fragmentary, reflecting a focus on culturable or disease-causing agents. Here we profile the transcriptomes of over 220 invertebrate species sampled across nine animal phyla and report the discovery of 1,445 RNA viruses, including some that are sufficiently divergent to comprise new families. The identified viruses fill major gaps in the RNA virus phylogeny and reveal an evolutionary history that is characterized by both host switching and co-divergence. The invertebrate virome also reveals remarkable genomic flexibility that includes frequent recombination, lateral gene transfer among viruses and hosts, gene gain and loss, and complex genomic rearrangements. Together, these data present a view of the RNA virosphere that is more phylogenetically and genetically diverse than that depicted in current classification schemes and provide a more solid foundation for studies in virus ecology and evolution.

Nature 540, 539–543 (22 December 2016) |



The goal of metagenomic classification software is to map sequence information to taxonomic information.



Yesterday, we did this
by BLASTing
several contigs on the NCBI
website. This is not a practical
approach for many contigs.

Nucleotide-level similarity identifies closely related organisms (blastn-like)
Protein-level similarity discovers ‘new’ organisms (blastx-like)

A nice review of metagenomic classifiers

OXFORD

Briefings in Bioinformatics, 2017, 1–15

doi: 10.1093/bib/bbx120
Paper

A review of methods and databases for metagenomic classification and assembly

Florian P. Breitwieser, Jennifer Lu and Steven L. Salzberg

Corresponding author: Steven L. Salzberg, Center for Computational Biology, Johns Hopkins University, 1900 E. Monument St., Baltimore, MD, 21205, USA.
E-mail: salzberg@jhu.edu

Name	References	URL
CaPSID	Borozan et al., 2012	https://github.com/capsid/capsid
ClassyFlu	Van der Auwera et al., 2014	http://bioinf.uni-greifswald.de/ClassyFlu/query/init
Clinical PathoScope	Byrd et al., 2014	https://sourceforge.net/p/pathoscope/wiki/clinical_pathoscope/
DUDes	Piro et al., 2016	http://sf.net/p/dudes
EnsembleAssembler	Deng et al., 2015	https://github.com/xutaodeng/EnsembleAssembler
Exhaustive Iterative Assembly (Virus Discovery Pipeline)	Schürch et al., 2014	–
FACS	Stranneheim et al., 2010	https://github.com/SciLifeLab/facs
GenSeed-HMM	Alves et al., 2016	https://sourceforge.net/projects/genseedhmm/
Giant Virus Finder	Kerepesi and Grolmusz, 2016	http://pitgroup.org/giant-virus-finder
GOTTCHA	Freitas et al., 2015	https://github.com/LANL-Bioinformatics/GOTTCHA
IMSA	Dimon et al., 2013	https://sourceforge.net/projects/arron-imsa/?source=directory
IMSA+A	Cox et al., 2017	https://github.com/JeremyCoxBML/IMSA-A
Kraken	Wood and Salzberg, 2014	https://github.com/DerrickWood/kraken
LMAT	Ames et al., 2013	https://sourceforge.net/projects/lmat/
MEGAN 4	Huson et al., 2011	http://ab.inf.uni-tuebingen.de/software/megan4/
MEGAN Community Edition	Huson et al., 2016	http://ab.inf.uni-tuebingen.de/data/software/megan6/download/welcome.html
MePIC	Takeuchi et al., 2014	https://mepic.nih.go.jp/
MetaShot	Fosso et al., 2017	https://github.com/bfoso/MetaShot
metaVIC	Modha, 2016	https://github.com/sejmodha/metaVIC
Metavir	Roux et al., 2011	http://metavir-meb.univ-bpclermont.fr/
Metavir 2	Roux et al., 2014	http://metavir-meb.univ-bpclermont.fr/
MetLab	Norling et al., 2016	https://github.com/norling/metlab
NBC	Rosen et al., 2011	http://nbc.ece.drexel.edu/
PathSeq	Kostic et al., 2011	https://www.broadinstitute.org/software/pathseq/
ProViDE	Ghosh et al., 2011	http://metagenomics.atc.tcs.com/binning/ProViDE/
QuasQ	Poh et al., 2013	http://www.statgen.nus.edu.sg/\$sim\$software/quasq.html
READSCAN	Naeem et al., 2013	http://cbrc.kaust.edu.sa/readscan/
Rega Typing Tool	Kroneman et al., 2011; Pineda-Peña et al., 2013	http://regatools.med.kuleuven.be/typing/v3/hiv/typingtool/
RIEMS	Scheuch et al., 2015	https://www.fli.de/fileadmin/FLI/IVD/Microarray-Diagnostics-RIEMS.tar.gz
RINS	Bhaduri et al., 2012	http://khavarilab.stanford.edu/tools-1/#tools
SLIM	Cotten et al., 2014	“Available upon request”
SMART	Lee et al., 2016	https://bitbucket.org/ayl/smart
SRSA	Isakov et al., 2011	“Available upon request”
SURPI	Naccache et al., 2014	https://github.com/chialab/surpi
Taxonomer	Flygare et al., 2016	https://www.taxonomer.com/
Taxy-Pro	Klingenberg et al., 2013	http://gobics.de/TaxyPro/
“Unknown pathogens from mixed clinical samples”	Gong et al., 2016	–
vFam	Skewes-Cox et al., 2014	http://derisilab.ucsf.edu/software/vFam/
VIP	Li et al., 2016	https://github.com/keylabivdc/VIP
ViralFusionSeq	Li et al., 2013	https://sourceforge.net/projects/viralfusionseq/
Virana	Schelhorn et al., 2013	https://github.com/schelhorn/virana
VirFind	Ho and Tzanetakis, 2014	http://virfind.org/
VIROME	Wommack et al., 2012	http://virome.dbi.udel.edu/app/#view=home
ViromeScan	Rampelli et al., 2016	https://sourceforge.net/projects/viromescan/
VirSorter	Roux et al., 2015	https://github.com/simroux/VirSorter
VirusFinder	Wang et al., 2013	http://bioinfo.mc.vanderbilt.edu/VirusFinder/
VirusHunter	Zhao et al., 2013	https://www.ibridgenetwork.org/#!/profiles/905559575893/innovations/103/
VirusSeeker	Zhao et al., 2017	https://wupathlabs.wustl.edu/virusseeker/
VirusSeq	Chen et al., 2013	http://odin.mdacc.tmc.edu/\$sim\$xsu1/VirusSeq.html
VirVarSeq	Verbiest et al., 2015	http://sourceforge.net/projects/virtools/?source=directory
VMGAP	Lorenzi et al., 2011	–

–, No website could be found, the workflow was unavailable.

idseq.net is a new web-based tool that does metagenomic classification
it's free, fast, and easy to use.

The screenshot shows the idseq.net website. At the top, there is a navigation bar with links to various bioinformatics tools like GenBank, blastn, blastp, and NCBI Tax. Below the navigation bar is the IDseq logo and a "Join our team" button. The main content area features a large text block about IDseq's purpose, followed by three service icons: Discover (pathogen landscape), Detect (outbreak monitoring), and Decipher (large dataset analysis). To the right, a call-to-action sidebar prompts users to learn more about IDseq, sign in, and provide their contact information for a team joining.

Apps Read Later GenBank blastn blastp tblastn NCBI Tax PubMed SRA github ViralZone Primer3 FoCo W Kuali BDSC CSU Dir Google Scholar Canvas

IDSEQ Join our team

IDseq is an unbiased global software platform that helps scientists identify pathogens in metagenomic sequencing data.

Discover
Identify the pathogen landscape

Detect
Monitor and review potential outbreaks

Decipher
Find potential infecting organisms in large datasets

Learn more about IDseq
Already have an account? [Sign in.](#)

First Name Last Name

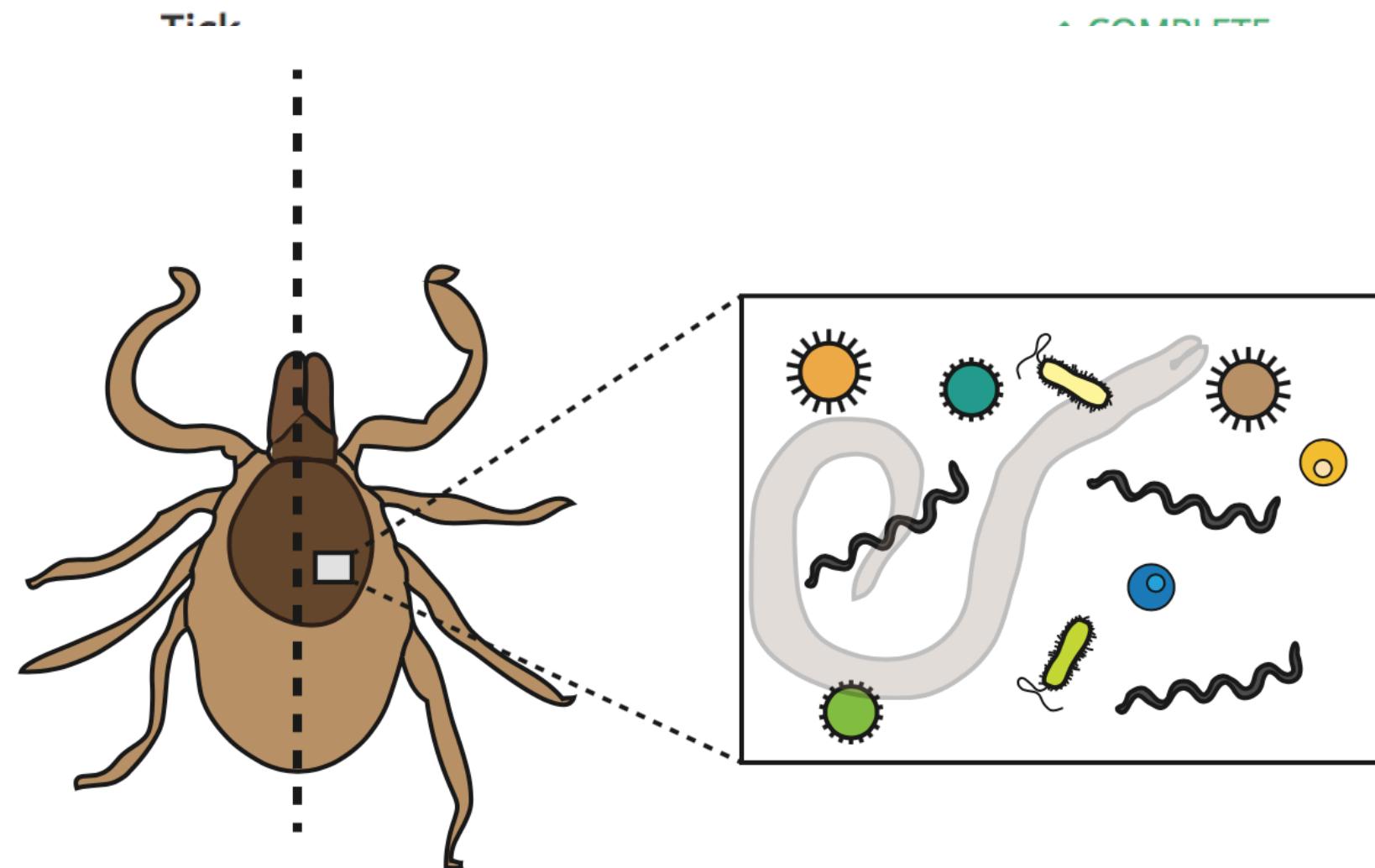
Email

Affiliated Institution or Company

How would you use IDseq? Optional

Submit

<input type="checkbox"/> Name	Total reads ▾	Passed filters ▾	Passed QC ▾	DCR ▾	Host ▾	Collection Location ▾	Status ▾
<input type="checkbox"/> Tick_16 a year ago Mark Stenglein	14,822,970	377,104 2.54%	61.82%	1.75	Tick	--	✓ COMPLETE ⌚ 50 minutes
<input type="checkbox"/> Tick_15 a year ago Mark Stenglein	8,776,796	202,590 2.31%	60.93%	1.51	Tick	--	✓ COMPLETE ⌚ 47 minutes
<input type="checkbox"/> Tick_14 a year ago Mark Stenglein	14,132,738	307,554 2.18%	63.27%	1.48	Tick	--	✓ COMPLETE ⌚ 46 minutes
<input type="checkbox"/> Tick_13 a year ago Mark Stenglein	14,892,862	356,202 2.39%	59.85%	1.58	Tick	--	✓ COMPLETE
<input type="checkbox"/> Tick_12 a year ago Mark Stenglein	13,067,958	199,586 1.53%	60.72%	1.38			
<input type="checkbox"/> Tick_11 a year ago Mark Stenglein	14,281,564	449,058 3.14%	63.74%	1.36			
<input type="checkbox"/> Tick_10 a year ago Mark Stenglein	11,685,018	200,632 1.72%	61.54%	1.29			



Stenglein_I_scap_ticks >

Tick_16 ▾

Sample Details

Share

Download ▾

Taxon name		Name Type: Scientific ▾	Background: NID Human CSF v3 ▾	Categories ▾	Threshold Filters ▾	Read Specificity: All ▾	Min Contig Size: 4 ▾
------------	--	-------------------------	--------------------------------	--------------	---------------------	-------------------------	----------------------

999 rows passing filters, out of 999 total rows.



> Taxon		Score ^	Z ^	rPM ^	r ^	%id ^	L ^	log(1/E) ^	NT NR
> Phlebovirus (3 viral species) ● 3	PATHOGENIC A	36,880,366	100.0 99.0	3,667.1 34.5	54,358 512	98.6 100.0	72.2 24.4	36 7	
> Non-genus-specific reads in family Enterobacteriaceae (1 bacterial species)		36,794,662	99.0 45.9	3,702.1 31.4	54,876 465	99.9 99.9	73.6 24.4	38 7	
> Ehrlichia (15 bacterial species) ● 15	PATHOGENIC C	22,546,226	99.0 66.6	3,838.3 2.8	56,895 42	99.1 100.0	71.3 24.6	36 6	
> Ixodes (11 eukaryotic species)		15,789,538	99.0 -0.3	2,005.1 0.1	29,722 2	98.1 100.0	69.5 25.0	34 6	
> Borrelia (12 bacterial species) ● 1		3,867,729	99.0 1.3	731.3 0.2	10,840 3	99.9 100.0	73.1 24.7	38 6	
> Non-genus-specific reads in family Boreliaceae (1 bacterial species)		1,692,640	100.0 -100.0	169.3 0.0	2,509 0	99.9 0.0	71.2 0.0	37 0	
> Non-genus-specific reads in family Anaplasmataceae (1 bacterial species)		323,922	99.0 100.0	32.6 0.1	483 2	99.9 100.0	51.9 25.0	24 6	

Stenglein_I_scap_ticks >

Tick_16 ▾

Sample Details

Share

Download ▾

[Download Report Table \(.csv\)](#)[Download Non-Host Reads \(.fasta\)](#)[Download Unmapped Reads \(.fasta\)](#)[See Results Folder](#)[Download Taxon Tree as SVG](#)[Download Taxon Tree as PNG](#)

Taxon name		Name Type: Scientific ▾	Background: NID Human CSF v3 ▾	Categories ▾	Threshold Filters ▾	Read	cifici	Score ▾	Z ▾	rPM ▾	r ▾	99.0	72.2	50
> Phlebovirus (3 viral species) ● 3		PATHOGENIC A						36,880,366	100.0 99.0	3,667.1 34.5	54,358 512	100.0	24.4	7
> Non-genus-specific reads in family Enterobacteriaceae (1 bacterial species)								36,794,662	99.0 45.9	3,702.1 31.4	54,876 465	99.9	73.6 24.4	38 7
> Ehrlichia (15 bacterial species) ● 15		PATHOGENIC C						22,546,226	99.0 66.6	3,838.3 2.8	56,895 42	99.1	71.3 24.6	36 6
> Ixodes (11 eukaryotic species)								15,789,538	99.0 -0.3	2,005.1 0.1	29,722 2	98.1	69.5 25.0	34 6
> Borrelia (12 bacterial species) ● 1								3,867,729	99.0 1.3	731.3 0.2	10,840 3	99.9	73.1 24.7	38 6
> Non-genus-specific reads in family Boreliaceae (1 bacterial species)								1,692,640	100.0 -100.0	169.3 0.0	2,509 0	99.9 0.0	71.2 0.0	37 0
> Non-genus-specific reads in family Anaplasmataceae (1 bacterial								323,922	99.0 100.0	32.6 0.1	483 2	99.9 100.0	51.9 25.0	24 6

If you are interested in using idseq, please contact

Rebecca Egger <regger@chanzuckerberg.com>

Pathogen discovery using metagenomics sequencing



case and control
tissues



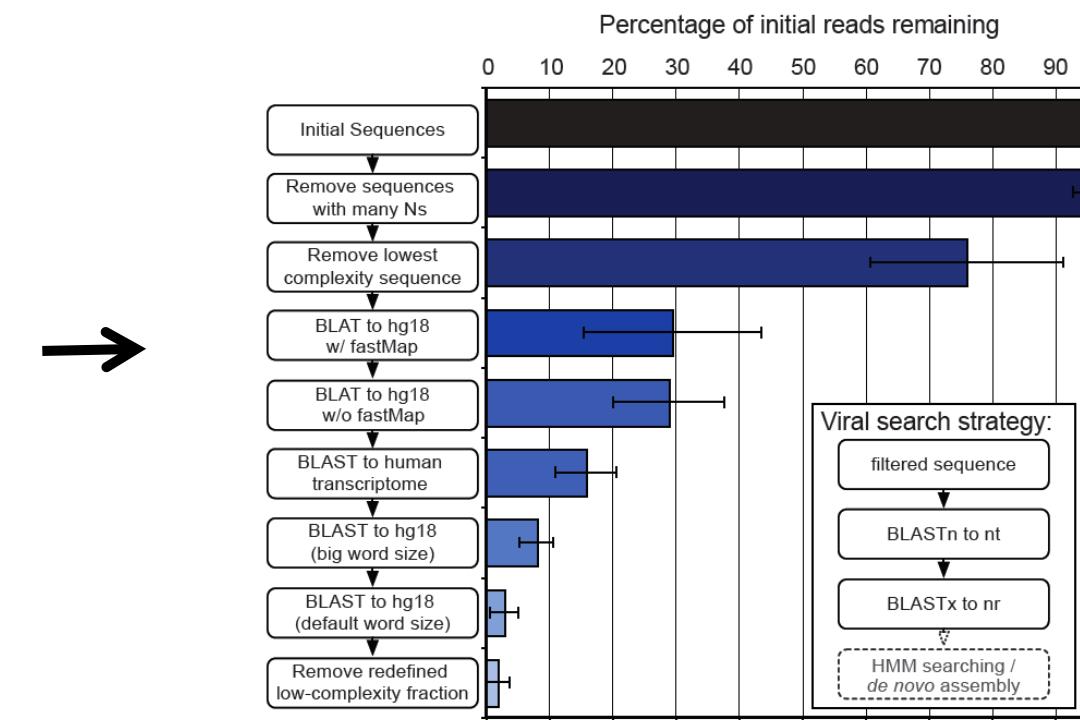
Nucleic acid



Library prep
/ barcode



Illumina
sequencing



Computational
Analysis



Follow-up

Sequencing can only give you sequences

A rabbit facility in TN experienced an outbreak of fatal gastroenteritis

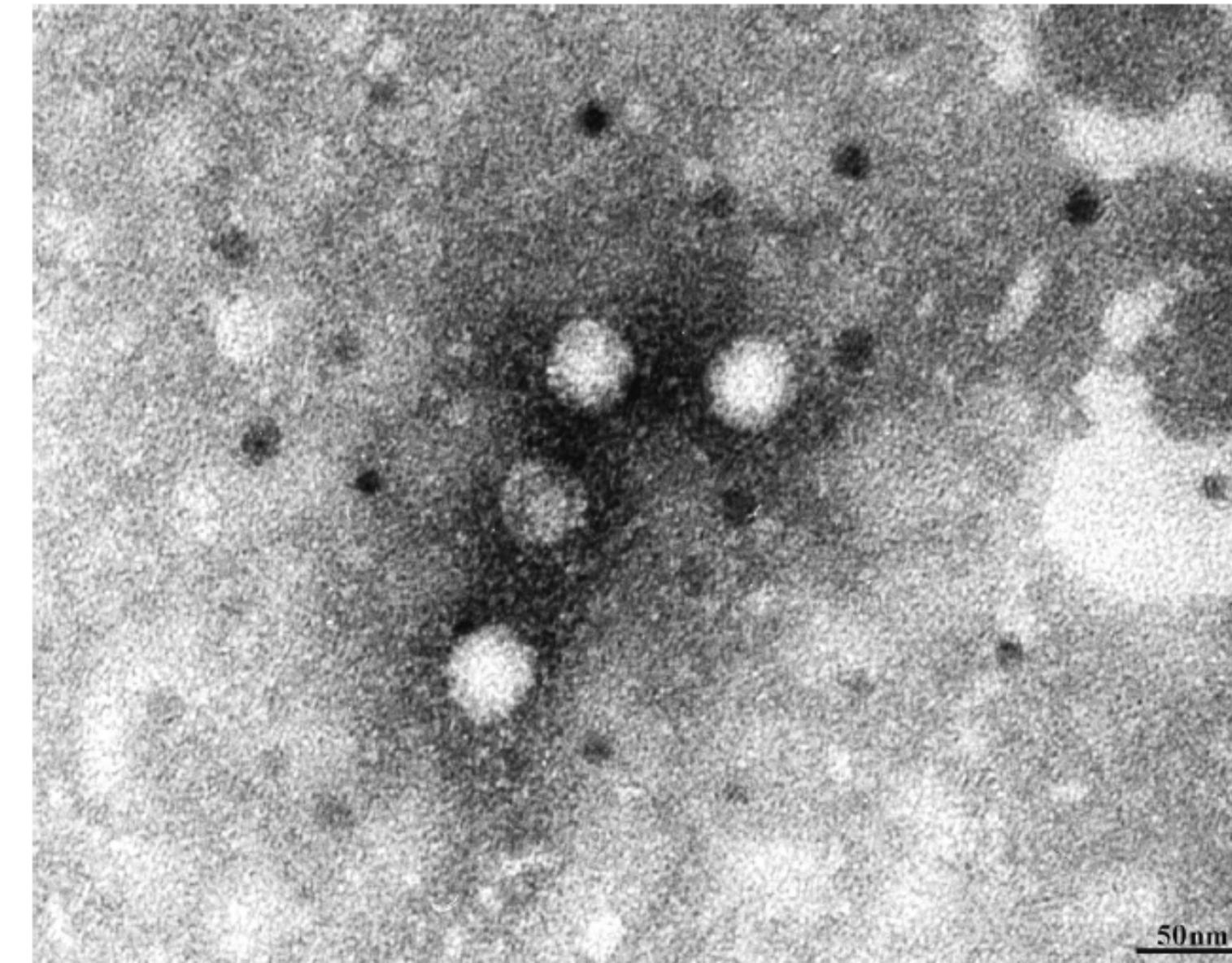


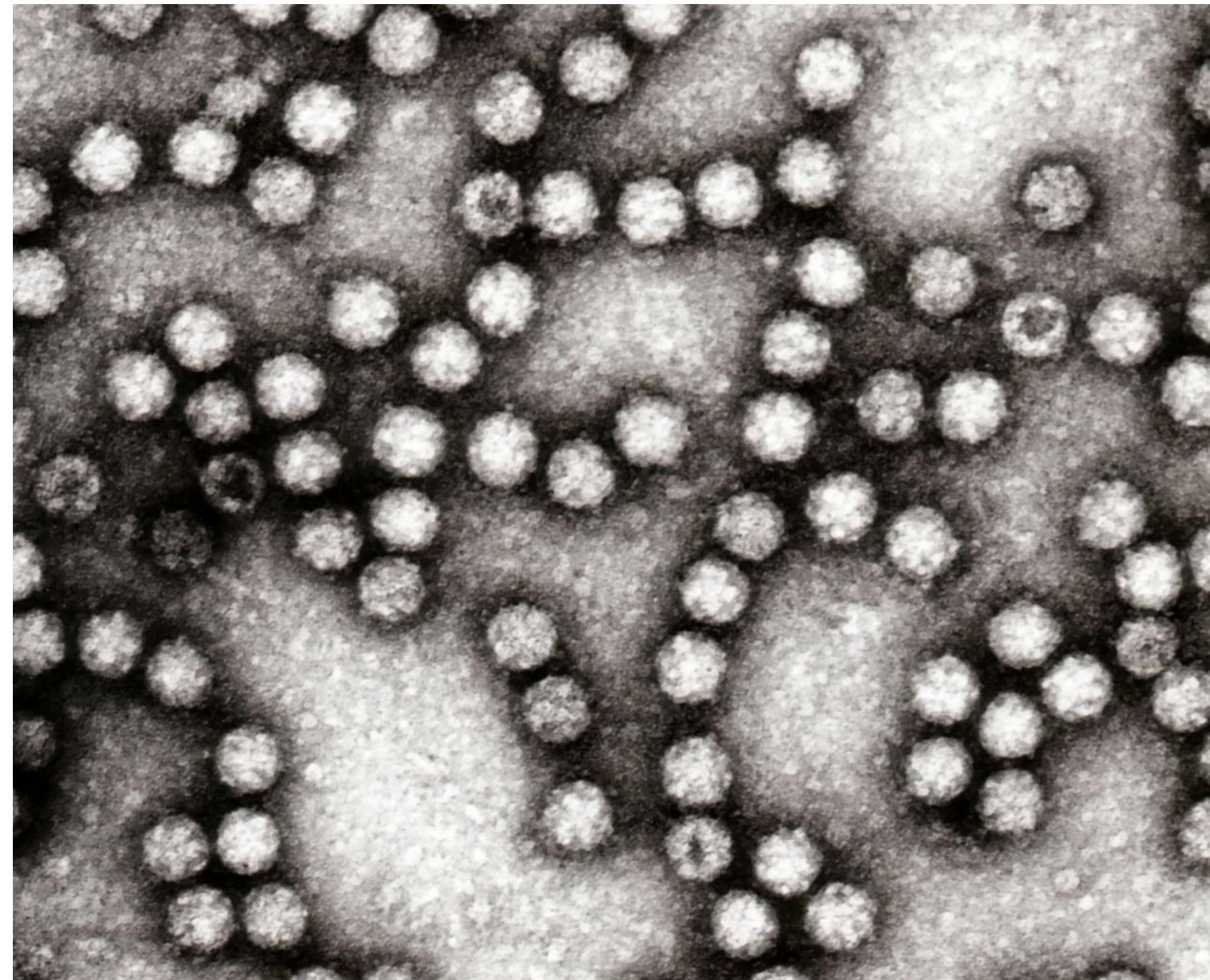
Figure 1 Electron micrograph of virus like particles in the stool of one animal (Table 1). Scale bar indicates 50 nm.

astrovirus sequences in the stool samples from sick rabbits

the virus is probably the cause of disease, but not proof

Additional experiments must be done

Astrovirus particles



JOURNAL OF CLINICAL MICROBIOLOGY, Apr. 1993, p. 955-962
0095-1137/93/040955-08\$02.00/0
Copyright © 1993, American Society for Microbiology

Vol. 31, No. 4

Characterization and Seroepidemiology of a Type 5 Astrovirus Associated with an Outbreak of Gastroenteritis in Marin County, California

KAREN MIDTHUN,^{1†*} HARRY B. GREENBERG,^{1‡} JOHN B. KURTZ,² G. WILLIAM GARY,³
FENG-YING C. LIN,⁴ AND ALBERT Z. KAPIKIAN¹

RESULTS

Volunteer study. Nineteen adult volunteers were orally administered a filtrate prepared from a 0.1% suspension of stool from one of the ill individuals in the original Marin County outbreak. None of 17 volunteers who received a 1-ml inoculum became ill. Because of this, the amount of inoculum was increased to 20 ml. Of two volunteers who received the larger inoculum, one developed a gastrointestinal illness characterized by nausea, vomiting, diarrhea, and malaise.