Comparative and Critical Analysis of the Results and Guided Discussion using Slido

September 03, 2025, Niamh Lacy-Roberts, Research Assistant at DTU, nlac@food.dtu.dk and Joana Mourão (Postdoc at DTU, PharmD, PhD joam@food.dtu.dk)

Icebreaker question: What did you have for breakfast this morning?

Outline and Learning Objectives

This session consists of the following elements:

- Participants will engage in a structured discussion of their analysis results using live Slido polls and questions.
- Topics covered include:
  - SimEx recap
  - Downloading data from SRA
  - Phylogenetic tree construction and interpretation.
  - Identification of outbreak clusters.
  - Hypotheses regarding the source of the outbreak strain.
  - Correlation with antimicrobial resistance profiles, virulence factors, biotype/serotype and genomic islands.

Comparative and Critical Analysis of Results:

- Live discussion of participant responses and aggregated poll results.
- Facilitators will highlight consistencies and discrepancies with the original study and explore different analytical paths and conclusions.

Simulated Exercise recap

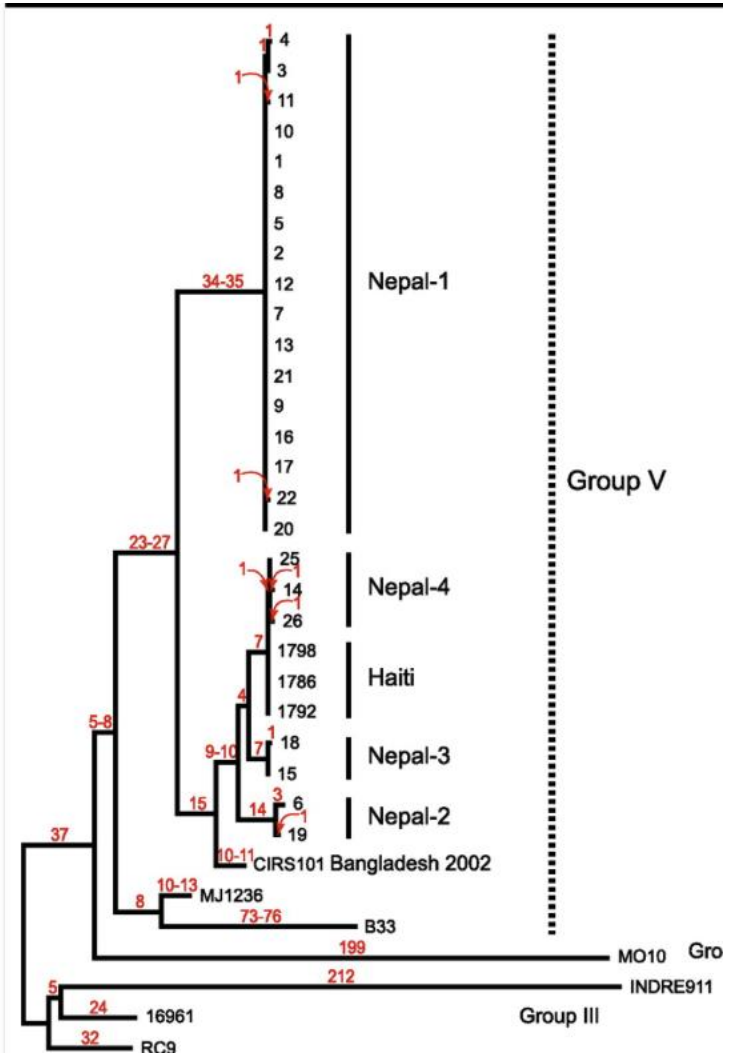
Task: Reproduce the phylogenetic analysis from Hendriksen et al. (2011) using tools of your choice.

Some suggestions:

- Online: CSI Phylogeny, MINTyper, iTOL, MEGA, Microreact
- Command-line: Snippy, IQ-TREE / RAxML

Aim:

- Identify clusters and closely related isolates
- Determine how Haitian and Nepalese isolates group
- Estimate the likely source of the outbreak strain based on tree topology & SNP distances



Hendriksen RS, Price LB, Schupp JM,Gillece JD, Kaas RS, Engelthaler DM,Bortolaia V, Pearson T, Waters AE,Prasad Upadhyay B, Devi Shrestha S,Adhikari S, Shakya G, Keim PS,Aarestrup FM.2011.Population Genetics of Vibrio cholerae from Nepal in 2010: Evidence on the Origin of the Haitian Outbreak. mBio2:10.1128/mbio.00157-11.https://doi.org/10.1128/mbio.00157-11

Data for the SimEx

Available isolates:

- 24 Nepalese *V. cholerae*
- 3 Haitian *V. cholerae*
- 5 additional V. cholerae isolates linked to the 2010 Haitian outbreak (optional to include in your analysis)

Data format:

- Illumina sequences
- Provided as zipped FASTA files, latter FASTQ files provided

Access:

- All files are available on ScienceData via this link: https://sciencedata.dk/shared/7ba608b7d30942995cc1e94baa475102

Your homework recap

Prepare a short document including:

- Phylogeny: Screenshot of your tree with key clades labelled.
- Interpretation:
  - Identify which isolates form clades.
  - Determine where the likely source is.
  - Report how many SNP differences you observe.
- Genomics: Summary of detected AMR genes, virulence factors, and genomic elements.
- Critical reflection: Notes on any limitations or uncertainties in your analysis.

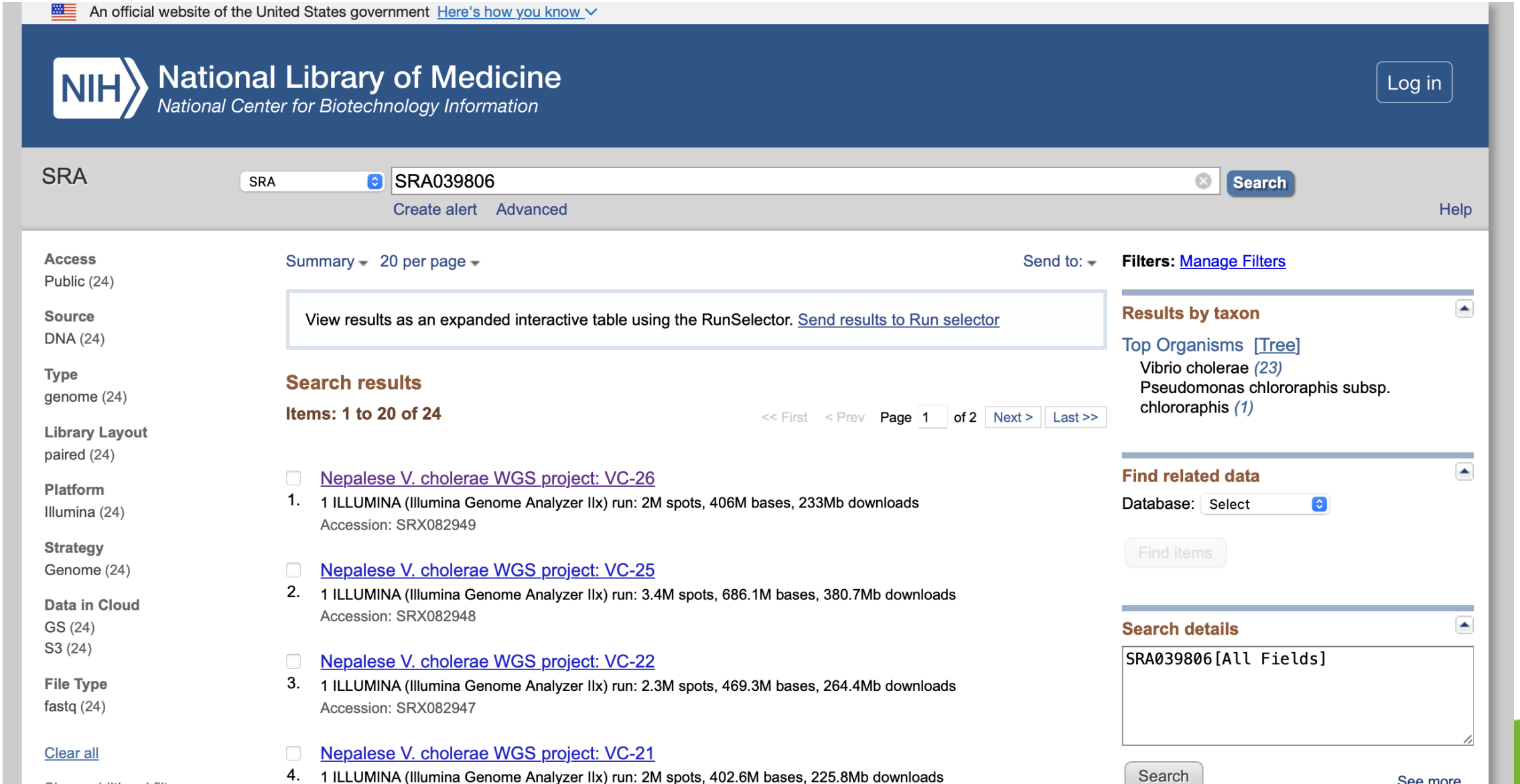
Do you attempt the exercises?

Downloading data from SRA

Data availability

accurate loading onto the flow cell, the same quantification method was used to quantify the final pools. The pooled, paired-end libraries were sequenced on the Illumina GAIIx to a read length of at least 76 base pairs. The average genome coverage for these 24 isolates was greater than 100× with a minimum of 75×. Over 97.6 of the genomes were at 10× cover or better. The Illumina genome sequencing data were deposited in the Short Read Archive at the National Center for Biotechnology Information (NCBI) under the accession no. SRA039806.1. The three Haitian genome sequences generated by the CDC were obtained from NCBI under the following accession numbers: strain 1786, SRX031665 (Illumina) and SRX031636 (454); strain 1792, SRX032204 (Illumina) and SRX032203 (454); and strain 1798, SRX032202 (Illumina) and SRX032201 (454).

SRA – NCBI


An official website of the United States government Here's how you know
National Library of Medicine
National Center for Biotechnology Information
Log in
SRA
SRA
SRA039806
Search
Create alert
Advanced
Help
Access
Public (24)
Source
DNA (24)
Type
genome (24)
Library Layout
paired (24)
Platform
Illumina (24)
Strategy
Genome (24)
Data in Cloud
GS (24)
S3 (24)
File Type
fastq (24)
Clear all
Summary
20 per page
Send to:
Filters: Manage Filters
View results as an expanded interactive table using the RunSelector. Send results to Run selector
Search results
Items: 1 to 20 of 24
<< First
< Prev
Page 1 of 2
Next >
Last >>
1. Nepalese V. cholerae WGS project: VC-26
1 ILLUMINA (Illumina Genome Analyzer IIx) run: 2M spots, 406M bases, 233Mb downloads
Accession: SRX082949
2. Nepalese V. cholerae WGS project: VC-25
1 ILLUMINA (Illumina Genome Analyzer IIx) run: 3.4M spots, 686.1M bases, 380.7Mb downloads
Accession: SRX082948
3. Nepalese V. cholerae WGS project: VC-22
1 ILLUMINA (Illumina Genome Analyzer IIx) run: 2.3M spots, 469.3M bases, 264.4Mb downloads
Accession: SRX082947
4. Nepalese V. cholerae WGS project: VC-21
1 ILLUMINA (Illumina Genome Analyzer IIx) run: 2M spots, 402.6M bases, 225.8Mb downloads
Results by taxon
Top Organisms [Tree]
Vibrio cholerae (23)
Pseudomonas chlororaphis subsp. chlororaphis (1)
Find related data
Database: Select
Find items
Search details
SRA039806[All Fields]
Search
See more

SRA – NCBI

Found 24 Items

Search within results

Clear

	Run	BioSample	Bases	Bytes	Experiment	Library Name	Organism	create_date	Sample Name
1	SRR308665	SAMN00632094	818.62 M	453.41 Mb	SRX082797	vc1	Vibrio cholerae	2014-05-30 22:17:00Z	Nepalese V. cholerae WGS project
2	SRR308690	SAMN00632095	742.88 M	420.26 Mb	SRX082833	VC2	Vibrio cholerae	2014-05-30 22:18:00Z	Nepalese V. cholerae WGS project: VC-2
3	SRR308691	SAMN00632097	759.82 M	436.55 Mb	SRX082834	VC3	Pseudomonas chlororaphis subsp. chlororaphis	2014-05-30 22:22:00Z	Nepalese V. cholerae WGS project: VC-3
4	SRR308692	SAMN00632098	621.91 M	356.30 Mb	SRX082835	VC4	Vibrio cholerae	2014-05-30 22:18:00Z	Nepalese V. cholerae WGS project: VC-4
5	SRR308693	SAMN00632099	645.69 M	365.36 Mb	SRX082836	VC5	Vibrio cholerae	2014-05-30 22:18:00Z	Nepalese V. cholerae WGS project: VC-5
6	SRR308703	SAMN00632100	725.47 M	416.19 Mb	SRX082837	VC6	Vibrio cholerae	2014-05-30 22:18:00Z	Nepalese V. cholerae WGS project: VC-6
7	SRR308704	SAMN00632101	676.67 M	382.74 Mb	SRX082838	VC7	Vibrio cholerae	2014-05-30 22:18:00Z	Nepalese V. cholerae WGS project: VC-7
8	SRR308705	SAMN00632102	792.35 M	444.87 Mb	SRX082839	VC8	Vibrio cholerae	2014-05-30 22:23:00Z	Nepalese V. cholerae WGS project: VC-8
9	SRR308706	SAMN00632103	718.68 M	401.41 Mb	SRX082840	VC9	Vibrio cholerae	2014-05-30 22:18:00Z	Nepalese V. cholerae WGS project: VC-9
10	SRR308707	SAMN00632104	615.77 M	353.56 Mb	SRX082841	VC10	Vibrio cholerae	2014-05-30 22:20:00Z	Nepalese V. cholerae WGS project: VC-10
11	SRR308708	SAMN00632105	545.73 M	311.76 Mb	SRX082842	VC11	Vibrio cholerae	2014-05-30 22:18:00Z	Nepalese V. cholerae WGS project: VC-11
12	SRR308709	SAMN00632106	558.69 M	319.65 Mb	SRX082843	VC12	Vibrio cholerae	2014-05-30 22:18:00Z	Nepalese V. cholerae WGS project: VC-12
13	SRR308713	SAMN00632107	438.37 M	243.09 Mb	SRX082938	VC13	Vibrio cholerae	2014-05-30 22:18:00Z	Nepalese V. cholerae WGS project: VC-13
14	SRR308715	SAMN00632108	392.78 M	219.44 Mb	SRX082939	VC14	Vibrio cholerae	2014-05-30 22:18:00Z	Nepalese V. cholerae WGS project: VC-14
15	SRR308716	SAMN00632109	587.68 M	325.54 Mb	SRX082940	VC15	Vibrio cholerae	2014-05-30 22:18:00Z	Nepalese V. cholerae WGS project: VC-15
16	SRR308717	SAMN00632110	573.61 M	312.86 Mb	SRX082941	VC16	Vibrio cholerae	2014-05-30 22:18:00Z	Nepalese V. cholerae WGS project: VC-16
17	SRR308720	SAMN00632111	374.76 M	207.80 Mb	SRX082942	VC17	Vibrio cholerae	2014-05-30 22:18:00Z	Nepalese V. cholerae WGS project: VC-17
18	SRR308721	SAMN00632112	427.70 M	240.25 Mb	SRX082943	VC18	Vibrio cholerae	2014-05-30 22:18:00Z	Nepalese V. cholerae WGS project: VC-18

Downloading files from SRA – web?

Nepalese V. cholerae WGS project : VC-1 (SRR308665)

Metadata | Analysis | Reads | Data access | **FASTA/FASTQ download**

Download for Experiment SRX082797

Accession	Total Bases	Spots	
		Total	Filtered
☑ SRR308665	818.6Mbases	4.1M	
☐ SRR308666	n/abases	n/a	
☐ SRR308689	n/abases	n/a	

Filter Runs

Search by sub-sequence, ✖ Filter

What can the filter be applied to?

Download

☐ Filtered ☐ Clipped FASTA or FASTQ

NCBI sra-tools and assembly

- I downloaded the accession list from the SRA Run Selector.
- I then used the SRA Toolkit https://github.com/ncbi/sra-tools to:
  - Download the data
  - convert to FASTQ with proper R1/R2 splitting
  - https://github.com/ncbi/sra-tools/wiki/HowTo:-fasterq-dump

  - Assembler: Shovill (SPAdes backend) v1.1.0
  - Polishing: Pilon (post-assembly error correction)
  - Output: draft genomes as contigs.fasta per sample
  - Renamed files

Additonal isolates- metadata

File name	Accession	Country / Region	Year	Source	Serogroup	Serotype
Vc_Mexico_environmental.fasta	ERR163271	Mexico	2010	Environmental	O1	Inaba
Vc_Mexico_clinical.fasta	ERR163268	Mexico	2010	Clinical	O1	Inaba
Vc_Dominican_Republic.fasta	SRR774788	Dominican Republic	2011	Clinical	O1	Ogawa (ctxB7)
Vc_Bangladesh_Chhatak.fasta	SRR491154	Bangladesh, Chhatak	2011	Environmental	O1	Ogawa (El Tor)
Vc_Bangladesh_Mathbaria.fasta	SRR495761	Bangladesh, Mathbaria	2011	Environmental	O1	Ogawa (El Tor)

Genomic Epidemiology of the Haitian Cholera Outbreak: a Single Introduction Followed by Rapid, Extensive, and Continued Spread Characterized the Onset of the Epidemic

Authors: Mark Eppinger, Talima Pearson, Sara S. K. Koenig, Ofori Pearson, Nathan Hicks, Sonia Agrawal, Fatemeh Sanjar, SHOW ALL (17 AUTHORS), Paul S. Keim | AUTHORS INFO & AFFILIATIONS

https://doi.org/10.1128/mbio.01721-14 · Check for updates

Eppinger, M., Pearson, T., Koenig, S. S., Pearson, O., Hicks, N., Agrawal, S., Sanjar, F., Galens, K., Daugherty, S., Crabtree, J., Hendriksen, R. S., Price, L. B., Upadhyay, B. P., Shakya, G., Fraser, C. M., Ravel, J., & Keim, P. S. (2014). Genomic epidemiology of the Haitian cholera outbreak: a single introduction followed by rapid, extensive, and continued spread characterized the onset of the epidemic. mBio, 5(6), e01721. https://doi.org/10.1128/mBio.01721-14

Integrated view of *Vibrio cholerae* in the Americas

DARYL DOMMAN, MARIE-LAURE QUILICI, MATTHEW J. DORMAN, ELISABETH NJAMKEPO, ANKUR MUTREJA, ALISON E. MATHER, GABRIELLA DELGADO, ROSARIO MORALES-ESPINOSA, PATRICK A. D. GRIMONT, [...], AND NICHOLAS R. THOMSON +10 authors Authors Info & Affiliations

SCIENCE · 10 Nov 2017 · Vol 358, Issue 6364 · pp. 789-793 · DOI: 10.1126/science.aao2136

2,712 125 CHECK ACCESS

Domman D, Quilici ML, Dorman MJ, Njamkepo E, Mutreja A, Mather AE, Delgado G, Morales-Espinosa R, Grimont PAD, Lizárraga-Partida ML, Bouchier C, Aanensen DM, Kuri-Morales P, Tarr CL, Dougan G, Parkhill J, Campos J, Cravioto A, Weill FX, Thomson NR. Integrated view of Vibrio cholerae in the Americas. Science. 2017 Nov 10;358(6364):789-793. doi: 10.1126/science.aao2136. PMID: 29123068.

SNP trees

Did you manage to make a tree? E.g., SNP tree, cgMLST tree etc? If so, what tools did you use?

SNP phylogeny workflow overview

- Data → Reference → Snippy → Gubbins → IQ-TREE → iTol
- Reference genome: RefSeq GCF_000166455.1 (2010EL-1786, from Haiti 2010 outbreak)
- Hendriksen et al., used reference genome Vibrio cholerae O1 biovar El Tor strain N16961 (Bangladesh 1971).

Emerg Infect Dis. 2011 Nov;17(11):2113–2121. doi: 10.3201/eid1711.110794

Comparative Genomics of *Vibrio cholerae* from Haiti, Asia, and Africa

Aleisha R Reimer [1,2,3,4,5], Gary Van Domselaar [1,2,3,4,5], Steven Stroika [1,2,3,4,5], Matthew Walker [1,2,3,4,5], Heather Kent [1,2,3,4,5], Cheryl Tarr [1,2,3,4,5], Deborah Talkington [1,2,3,4,5], Lori Rowe [1,2,3,4,5], Melissa Olsen-Rasmussen [1,2,3,4,5], Michael Frace [1,2,3,4,5], Scott Sammons [1,2,3,4,5], Georges Anicet Dahourou [1,2,3,4,5], Jacques Boncy [1,2,3,4,5], Anthony M Smith [1,2,3,4,5], Philip Mabon [1,2,3,4,5], Aaron Petkau [1,2,3,4,5], Morag Graham [1,2,3,4,5], Matthew W Gilmour [1,2,3,4,5], Peter Gerner-Smidt [1,2,3,4,5,✉]; the V. cholerae Outbreak Genomics Task Force [1,2,3,4,5]

Author information · Copyright and License information

PMCID: PMC3310578 PMID: 22099115

Reimer, A. R., Van Domselaar, G., Stroika, S., Walker, M., Kent, H., Tarr, C., Talkington, D., Rowe, L., Olsen-Rasmussen, M., Frace, M., Sammons, S., Dahourou, G. A., Boncy, J., Smith, A. M., Mabon, P., Petkau, A., Graham, M., Gilmour, M. W., Gerner-Smidt, P., & V. cholerae Outbreak Genomics Task Force (2011). Comparative genomics of Vibrio cholerae from Haiti, Asia, and Africa. Emerging infectious diseases, 17(11), 2113–2121. https://doi.org/10.3201/eid1711.110794

Sample details

BioSample ID	SAMN00765529
Description	Pathogen: environmental/food/other sample from Vibrio cholerae O1 str. 2010EL-1786
Submitter	CDC
Strain	2010EL-1786
Source of dna	Source DNA available from Dr. Peter Gerner-Smidt, Centers for Disease Control and Prevention, 1600 Clifton Road, Atlanta GA USA
Geographic location	Haiti
Collection date	2010
Host	Homo sapiens
Note	Isolate from Haiti 2010 cholera outbreak
Serogroup	O1
Isolation source	stool sample from patient with cholera
Sample name	2010EL-1786
Collected by	Ministere de la Sante Publique et de la Population, Haiti
Culture collection	ATCC:BAA-2163

Which reference genome did you use?

Snippy

- Line up each isolate to the same reference
- Mark SNPs (single-letter differences) and indels (insertions/deletions)
- Result per isolate: SNP calls, a consensus sequence, and QC logs
- Goal: make differences comparable across samples
- https://github.com/tseemann/snippy

Core SNP phylogeny

If you call SNPs for multiple isolates from the same reference, you can produce an alignment of "core SNPs" which can be used to build a high-resolution phylogeny (ignoring possible recombination). A "core site" is a genomic position that is present in *all* the samples. A core site can have the same nucleotide in every sample ("monomorphic") or some samples can be different ("polymorphic" or "variant"). If we ignore the complications of "ins", "del" variant types, and just use variant sites, these are the "core SNP genome".

Input Requirements

- a set of Snippy folders which used the same `--ref` sequence.

Using `snippy-multi`

To simplify running a set of isolate sequences (reads or contigs) against the same reference, you can use the `snippy-multi` script. This script requires a *tab separated* input file as follows, and can handle paired-end reads, single-end reads, and assembled contigs.

```
# input.tab = ID R1 [R2]
Isolate1    /path/to/R1.fq.gz   /path/to/R2.fq.gz
Isolate1b   /path/to/R1.fastq.gz    /path/to/R2.fastq.gz
Isolate1c   /path/to/R1.fa      /path/to/R2.fa
# single end reads supported too
Isolate2    /path/to/SE.fq.gz
Isolate2b   /path/to/iontorrent.fastq
# or already assembled contigs if you don't have reads
Isolate3    /path/to/contigs.fa
Isolate3b   /path/to/reference.fna.gz
```

Snippy-core

- snippy-core keeps only sites present in all isolates → the core
- Two outputs:
  - core.full.aln = invariant + variant sites (full core)
  - core.aln = SNP-only columns (polymorphic sites)
- snippy-multi (concept): batch many isolates with one table; runs Snippy for each, then snippy-core
- https://github.com/tseemann/snippy

Then one would run this to generate the output script. The first parameter should be the `input.tab` file. The remaining parameters should be any remaining shared `snippy` parameters. The `ID` will be used for each isolate's `--outdir`.

```
% snippy-multi input.tab --ref Reference.gbk --cpus 16 > runme.sh

% less runme.sh   # check the script makes sense

% sh ./runme.sh   # leave it running over lunch
```

It will also run `snippy-core` at the end to generate the core genome SNP alignment files `core.*`.

Output Files

Extension	Description
.aln	A core SNP alignment in the `--aformat` format (default FASTA)
.full.aln	A whole genome SNP alignment (includes invariant sites)
.tab	Tab-separated columnar list of **core** SNP sites with alleles but NO annotations
.vcf	Multi-sample VCF file with genotype `GT` tags for all discovered alleles
.txt	Tab-separated columnar list of alignment/core-size statistics
.ref.fa	FASTA version/copy of the `--ref`
.self_mask.bed	BED file generated if `--mask auto` is used.

Gubbins: remove recombination before building the tree

- We don’t want SNPs introduced by recombination to drive the tree.
- After snippy-core, some SNPs may occur in dense clusters from recombination.
- Gubbins detects those recombinant regions and masks them → we keep the putative “true” SNPs from “vertical inheritance”.
- This is similar in spirit to CSI Phylogeny’s SNP pruning (e.g., pruning = 100 bp), but Gubbins is phylogeny-aware rather than a fixed-distance rule.
- https://github.com/nickjcroucher/gubbins

Did you filter for recombination? If so, what tool did you use?

IQ-TREE: Maximum-likelihood tree

- Method: Maximum-likelihood (ML) phylogeny in IQ-TREE 2
- Input: Gubbins-filtered SNP-only core alignment
- Outputs:
  - snp_tree.treefile → best ML tree
  - snp_tree.contree → bootstrap consensus (supports on nodes)
  - snp_tree.iqtree → run log & best-fit model details
- Uploaded the .treefile to iTOL
- https://github.com/iqtree/iqtree3
- Also available online! https://iqtree.github.io

Do you have a preferred tree viewing tool?

Let’s look at the ML tree in iTol ☺

Does your tree recover the key clades reported in Hendriksen et al.?

If your tree differs from the paper, what's the main difference?

Can you infer the origin of the outbreak from your tree?

Did you include any of the additional isolates in your analysis?

Did the Dominican Republic isolate cluster with the Haiti isolates?

How about the Bangladesh and Mexico isolates? Do they cluster with the Haiti isolates?

Thank you for listening! ☺

Acknowledgements

The creation of this training material was commissioned by ECDC to Institut Pasteur with the direct involvement of Niamh Lacy-Roberts, Research Assistant at DTU, nlac@food.dtu.dk