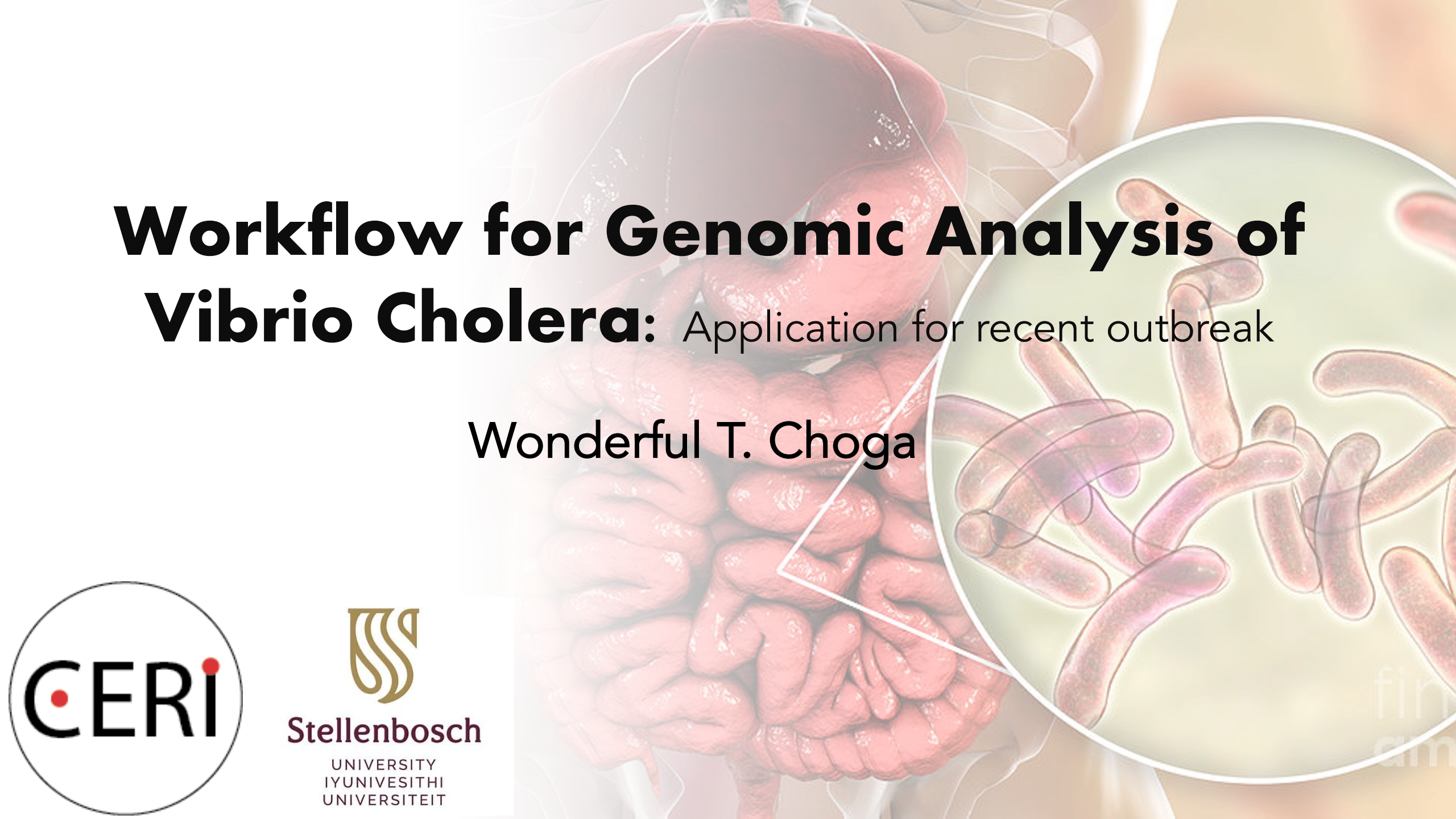


Workflow for Genomic Analysis of Vibrio Cholera: Application for recent outbreak

Wonderful T. Choga

CERI

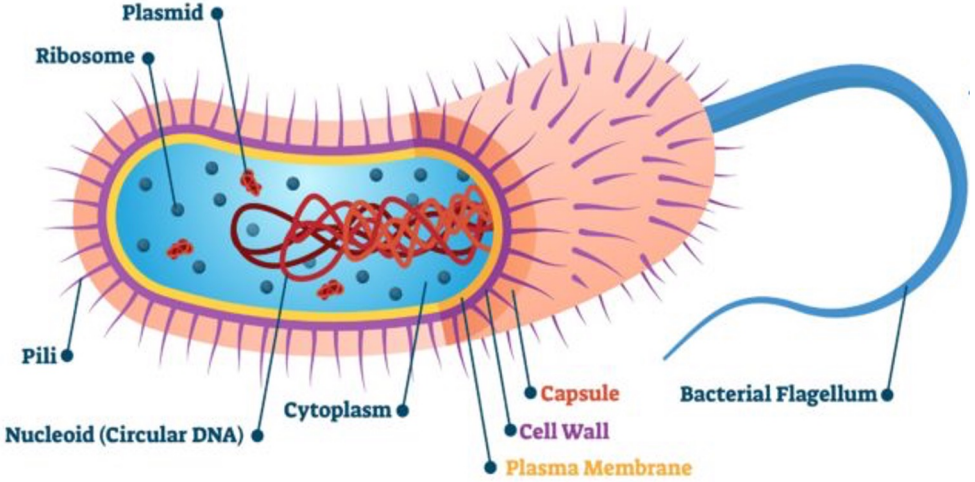
Stellenbosch
UNIVERSITY
IYUNIVESITHI
UNIVERSITEIT

Outline

- ✓ **Brief Background & Genomic Organization (Cholera)**
  (Discovery, Epidemiology, Virology, Immunopathogenesis)

- ✓ **From Lab-work to Genomic analysis.**
  - **PreQC :** fastqc & multiqc
  - **Fastp :** Alignment
  - **Snippy :** Assembly, .vcf ,
  - **FastBAPs :** Assembly, .vcf ,
  - **Gubbins :** Assembly, .vcf ,(remove recombinant sites)
  - PI Sites : romove conserved regions.
  - Dated ML-tree analysis

- ✓ **Summary and Wayforward.**

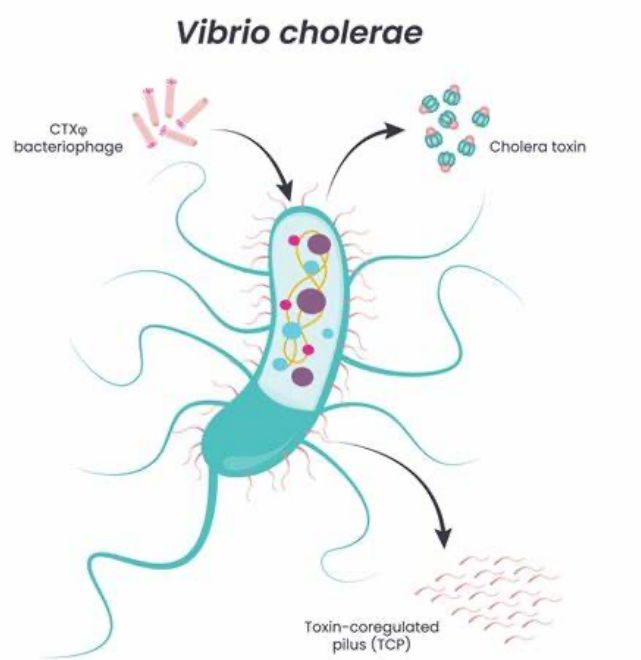


Vibrio cholerae

a) a bacterium autochthonous to the aquatic environment

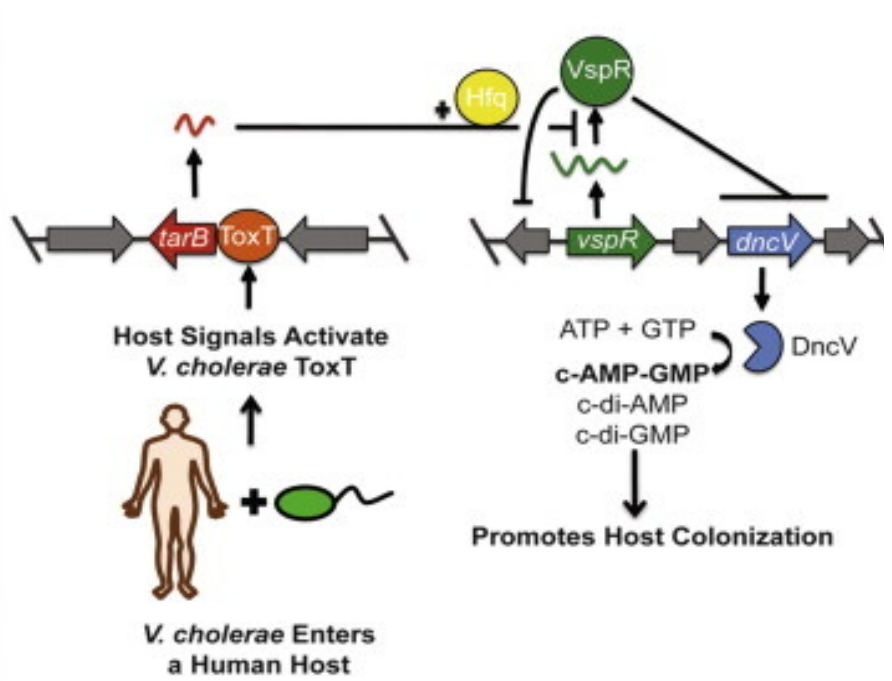
b) introduced into the human intestine through contaminated water or food

c) is the etiological agent of the acute secretory diarrheal disease described as cholera.



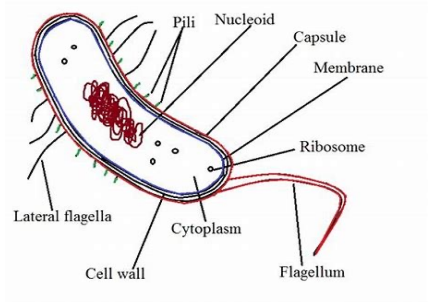
Vibrio cholerae:

- Environment : autochthonous to the aquatic environment
- Etiology : choleraic diarrheal & electrolyte imbalance

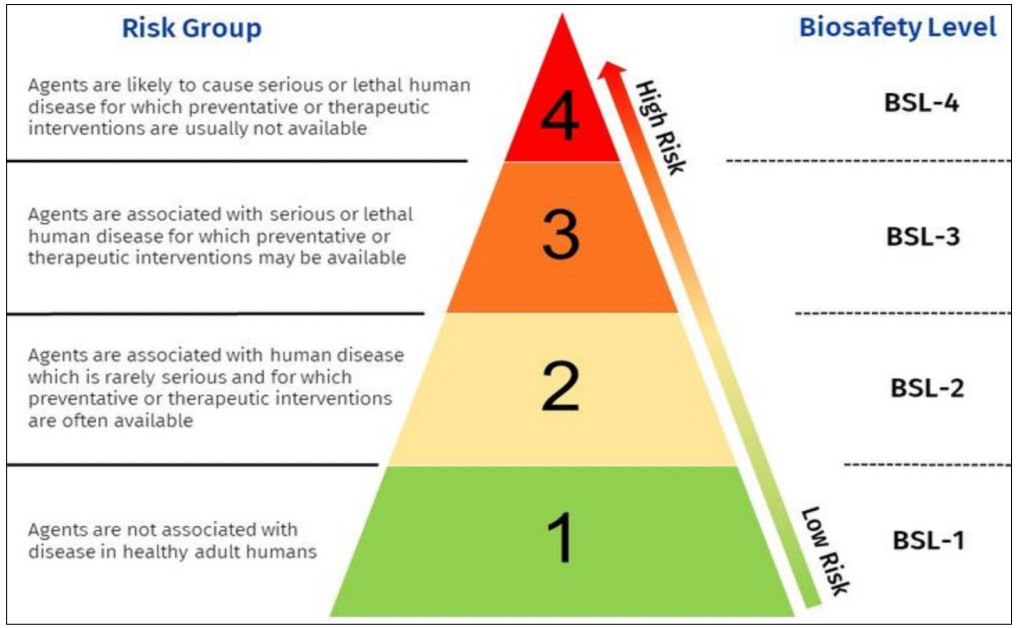


Classification of Cholera

- Species : ***V. cholerae***
- Kingdom : **Bacteria**
- Coronaviruses belong to:
  - ✓ Genus **Vibrio**
  - ✓ Family **Vibrionaceae**
  - ✓ Order **Vibrionales**
  - ✓ Class **Gammaproteobacteria**
- Phylum **Pseudomonadota**



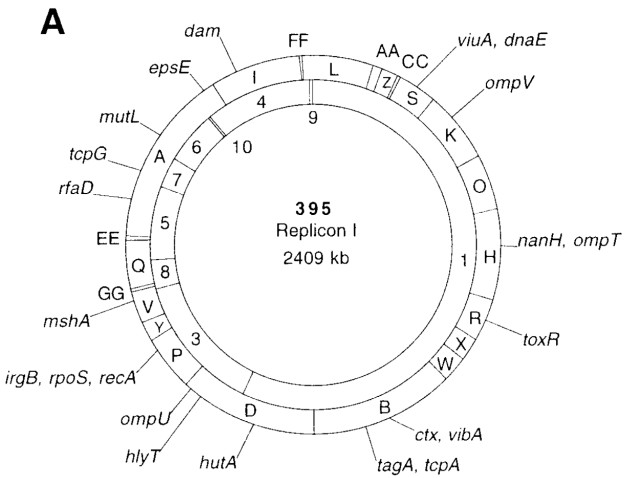
BSL 2

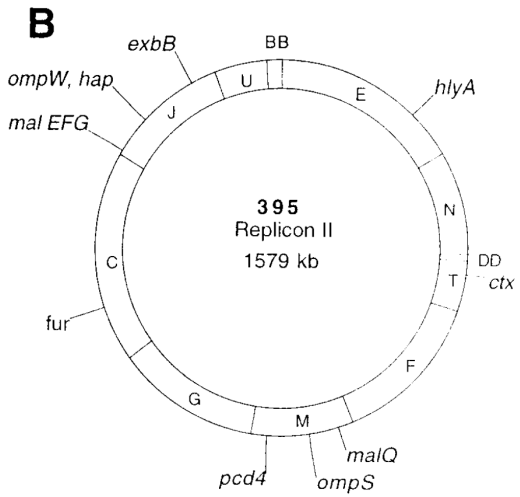


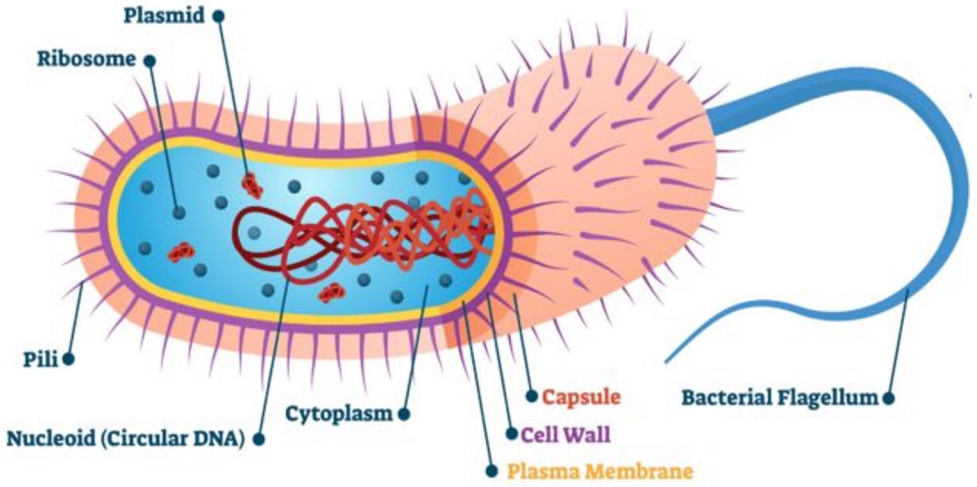
Cholera structure

- Gram –ve, facultative anaerobe
- Curved or comma-shaped bacteria
- Single polar flagellum at one pole; several pili throughout cell surface
- Two Circular DNA : 1 – produces choler toxin that causes diarrhea

  2 – does not produce directly produce toxins.







Cholera Vaccines & Treatment

Treatment

Cholera requires immediate treatment because the disease can cause death within hours.

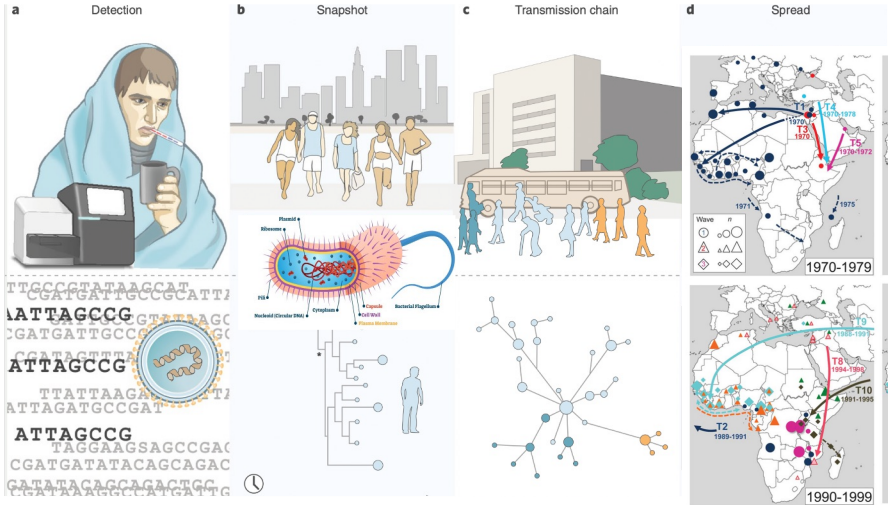
- **Rehydration.** The goal is to replace lost fluids and electrolytes using a simple rehydration solution, oral rehydration salts (ORS). The ORS solution is available as a powder that can be made with boiled or bottled water.

  Without rehydration, approximately half the people with cholera die. With treatment, fatalities drop to less than 1%.
- **Intravenous fluids.** Most people with cholera can be helped by oral rehydration alone, but severely dehydrated people might also need intravenous fluids.
- **Antibiotics.** While not a necessary part of cholera treatment, some antibiotics can reduce cholera-related diarrhea and shorten how long it lasts in severely ill people.
- **Zinc supplements.** Research has shown that zinc might decrease diarrhea and shorten how long it lasts in children with cholera.

Vaccine name (Manufacturer)	How given	Number of doses recommended	Age range	How long vaccination is effective
Vaxchora (Emergent BioSolutions)	By mouth	1 dose	2–64 years	At least 3–6 months
Dukoral (SBL Vaccines)	By mouth	2 doses, given 1–6 weeks apart (Children aged 2–5 years need 3 doses, given 1 to 6 weeks apart)	2 years and older	2 years
ShanChol ** (Sanofi Healthcare India Private Limited)	By mouth	2 doses, given at least 2 weeks apart	1 year and older	At least 3 years for 2 doses; short-term protection for 1 dose
Euvichol-Plus ** (EuBiologics)	By mouth	2 doses, given at least 2 weeks apart	1 year and older	At least 3 years for 2 doses; short-term protection for 1 dose

Epidemic Response

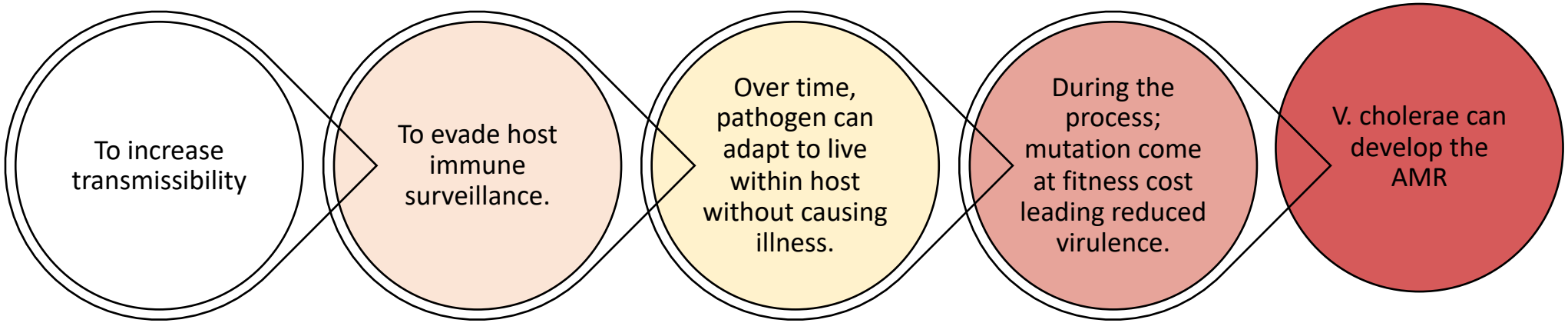
- coordinated efforts and measures aimed to control and mitigate the spread of a disease outbreak or epidemic.



Genomic Surveliance?

- out-beak investiagtion; to (i) intentify the genomic properties of circulating strain (e.g, AMR, escape mutations); (ii) the transmission dynamics (origins & spread).

Why do Pathogens mutate?



Genomics : *gene-omics*

Nature Reviews Genetics | Published online 4 Dec 2017; doi:10.1038/nrg.2017.

EDITORIAL | 21 January 2019

Genomics can help to monitor cholera

Sequence data from whole genomes let researchers track the spread of strains worldwide.

RESEARCH

CHOLERA

Genomic history of the seventh pandemic of cholera in Africa

François-Xavier Weill,[1,2*] Daryl Domman,[2] Elisabeth Njamkepo,[1] Cheryl Tarr,[3] Jean Rauzier,[1] Nizar Fawal,[1] Karen H. Keddy,[4,5] Henrik Salje,[6,7] Sandra Moore,[8] Asish K. Mukhopadhyay,[9] Raymond Bercion,[10,11] Francisco J. Luquero,[12] Antoinette Ngandjio,[13] Mireille Dosso,[14] Elena Monakhova,[15] Benoit Garin,[11†] Christiane Bouchier,[16] Carlo Pazzani,[17] Ankur Mutreja,[18,19] Roland Grunow,[20] Fati Sidikou,[21] Laurence Bonte,[22‡] Sébastien Breurec,[10†] Maria Damian,[23] Berthe-Marie Njanpop-Lafourcade,[24] Guillaume Sapriel,[25,26] Anne-Laure Page,[12] Monzer Hamze,[27] Myriam Henkens,[28‡] Goutam Chowdhury,[9] Martin Mengel,[24] Jean-Louis Koeck,[29§] Jean-Michel Fournier,[30] Gordon Dougan,[2,18] Patrick A. D. Grimont,[31] Julian Parkhill,[2] Kathryn E. Holt,[32] Renaud Piarroux,[8] Thandavarayan Ramamurthy,[19] Marie-Laure Quilici,[1,30||] Nicholas R. Thomson[2,33||]

PATHOGEN GENOMICS

Genomics in the time of cholera

> these findings will improve the fight against cholera

The seventh cholera pandemic, which started more than 50 years ago, has heavily affected Africa and the Americas. Now, two genomic studies published in *Science* reconstruct the history of the seventh pandemic in both continents and reveal the Asian origin of the strains responsible for it.

Cholera is caused by the bacterium *Vibrio cholerae*, which produces the cholera toxin. It causes an acute intestinal infection leading to a rapid and severe loss of bodily fluids. The seventh cholera pandemic — attributed to *V. cholerae* belonging to the O1 serogroup (or the O139 variant) and the El Tor biotype — started in Indonesia in 1961 and spread globally to South Asia in 1963, Africa in 1970, Latin America in 1991 and the Caribbean in 2010. Recently, it has been estimated that 1.4–4.0 million cases still occur every year, leading to between 21,000 and 143,000 deaths around the world, which highlights the very high burden of this pandemic.

There is ongoing debate on whether these cholera epidemics are caused by local indigenous strains or are of external origin. Although phylogenetic studies have attempted to map global dissemination, traditional pre-genomic methods such as serotyping are unsuitable for providing high-resolution phylogenetic lineages. The two new studies report the sequencing and analysis of hundreds of *V. cholerae* genomes to infer the history of epidemics spreading across Africa and the Americas.

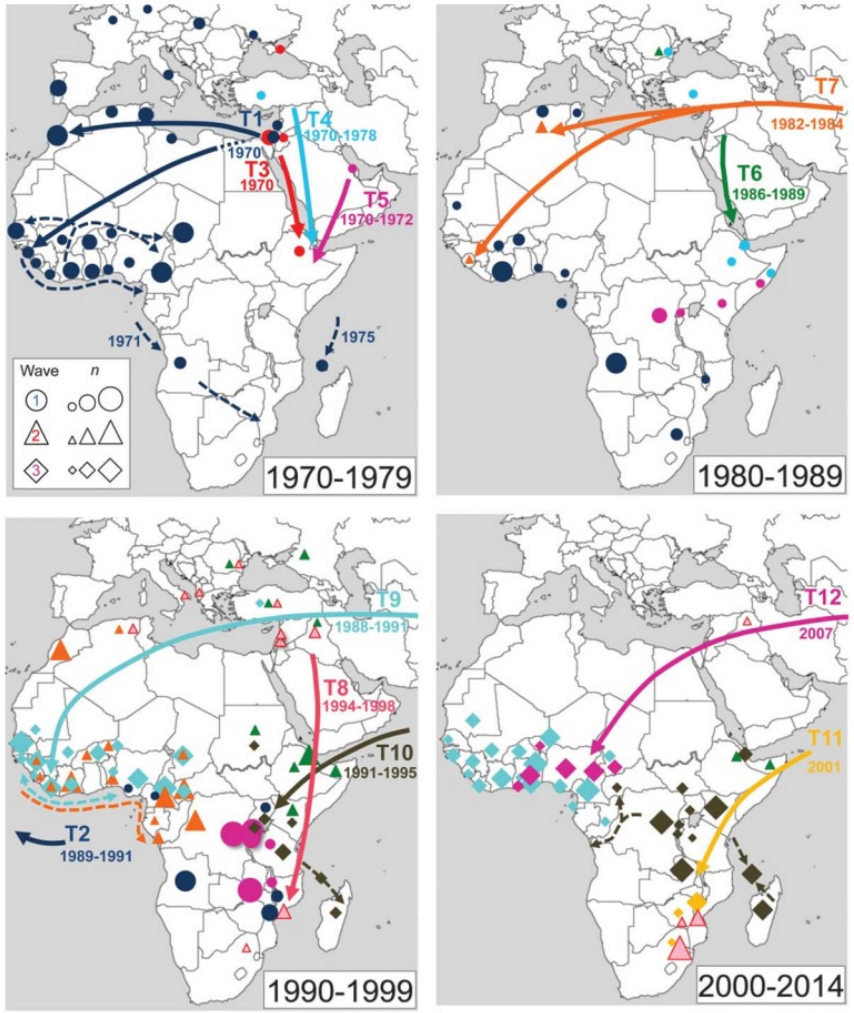
The study by Weill *et al.* analysed the genomes of 1,070 global *V. cholerae* isolates, including 651 from Africa representing a total of 45 African countries and a 49-year period. This wide temporal and geographical scope enabled the authors to reconstruct the phylogeny of the seventh pandemic in Africa and to infer the dissemination routes to, from and within the African continent. The authors determined that the different epidemics could all be traced back to a single lineage, which has been introduced at least 11 times since the first epidemic in the 1970s. The team also found that the last five introductions all originated from Asia and involved epidemics. The authors sequenced 252 isolates collected from 14 different countries between 1974 and 2014. As in the study above, this broad geographical and temporal collection enabled the authors to reconstruct the epidemic spreading and distinguish epidemic from non-epidemic lineages. Phylogenetic analysis and comparison with the data in Weill *et al.* indicated that the seventh pandemic lineages were introduced multiple times in Latin America and the Caribbean; these were globally circulating lineages of Asian or African origin. The authors were able to clearly distinguish the disease outbreaks caused by local endemic lineages from the much more severe outbreaks caused by the seventh pandemic lineages, which were introduced *de novo* through human activity.

Both studies indicate that imported strains are the major drivers of cholera epidemics in Africa and the Americas and highlight the role played by humans in spreading the disease. They also confirm previous epidemiological studies suggesting that human factors are more important in cholera dynamics than climate or other environmental factors. Importantly, these findings will improve the fight against cholera by focusing public health intervention efforts on the more dangerous global strains.

Carolina Perdigoto, Associate Editor,
Nature Communications

ORIGINAL ARTICLES Weill, F.-X. *et al.* Genomic history of the seventh pandemic of cholera in

Application on genomics to track pandemic of cholera in Africa.



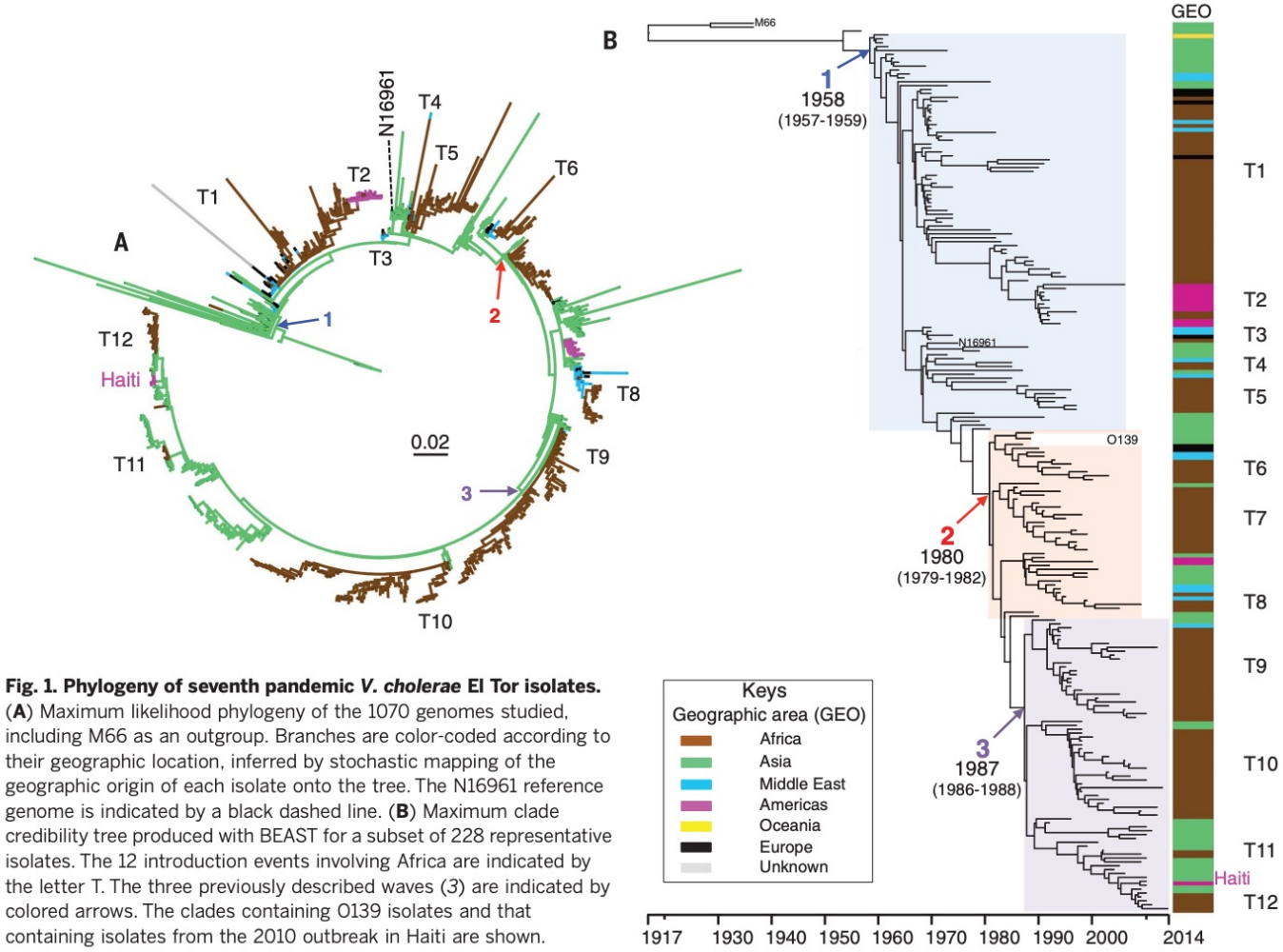


Fig. 1. Phylogeny of seventh pandemic *V. cholerae* El Tor isolates. (**A**) Maximum likelihood phylogeny of the 1070 genomes studied, including M66 as an outgroup. Branches are color-coded according to their geographic location, inferred by stochastic mapping of the geographic origin of each isolate onto the tree. The N16961 reference genome is indicated by a black dashed line. (**B**) Maximum clade credibility tree produced with BEAST for a subset of 228 representative isolates. The 12 introduction events involving Africa are indicated by the letter T. The three previously described waves (*3*) are indicated by colored arrows. The clades containing O139 isolates and that containing isolates from the 2010 outbreak in Haiti are shown.

Application continue....

RESEARCH

CHOLERA

Genomic history of the seventh pandemic of cholera in Africa

François-Xavier Weill,[1,2]* Daryl Domman,[2] Elisabeth Njamkepo,[1] Cheryl Tarr,[3] Jean Rauzier,[1] Nizar Fawal,[1] Karen H. Keddy,[4,5] Henrik Salje,[6,7] Sandra Moore,[8] Asish K. Mukhopadhyay,[9] Raymond Bercion,[10,11] Francisco J. Luquero,[12] Antoinette Ngandjio,[13] Mireille Dosso,[14] Elena Monakhova,[15] Benoit Garin,[11]† Christiane Bouchier,[16] Carlo Pazzani,[17] Ankur Mutreja,[18,19] Roland Grunow,[20] Fati Sidikou,[21] Laurence Bonte,[22]‡ Sébastien Breurec,[10]† Maria Damian,[23] Berthe-Marie Njanpop-Lafourcade,[24] Guillaume Sapriel,[25,26] Anne-Laure Page,[12] Monzer Hamze,[27] Myriam Henkens,[28]‡ Goutam Chowdhury,[9] Martin Mengel,[24] Jean-Louis Koeck,[29]§ Jean-Michel Fournier,[30] Gordon Dougan,[2,18] Patrick A. D. Grimont,[31] Julian Parkhill,[2] Kathryn E. Holt,[32] Renaud Piarroux,[8] Thandavarayan Ramamurthy,[19] Marie-Laure Quilici,[1,30]|| Nicholas R. Thomson[2,33]||

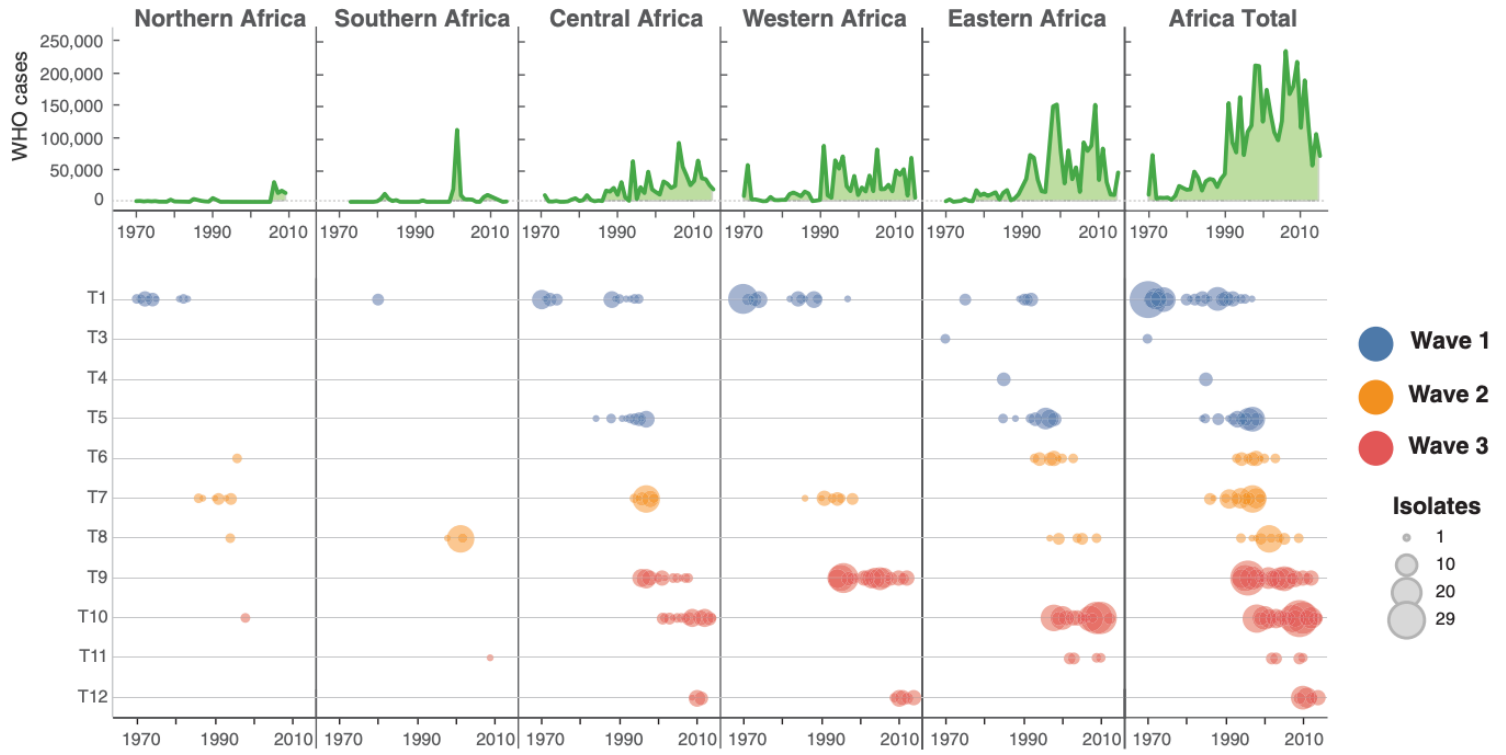


Fig. 3. Geographic and temporal distribution of seventh pandemic *V. cholerae* El Tor isolates from Africa according to their inferred introduction events (T1 to T12). The annual number of cholera cases reported to the World Health Organization (WHO) is shown in the upper panels. The United Nations subregion scheme was used for the geographic breakdown. The size of the circle scales with the number of genomes analyzed per year.

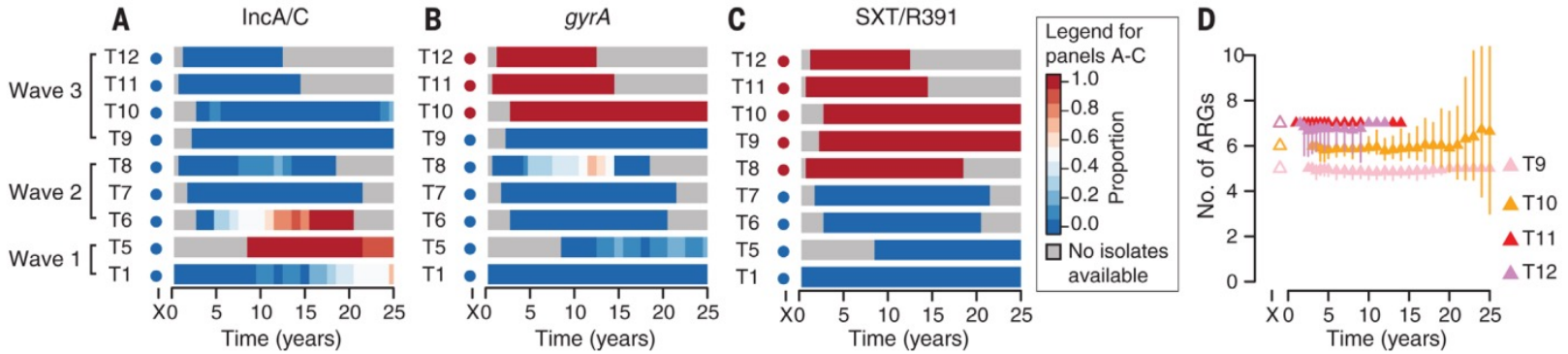


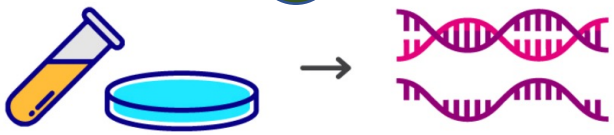
Fig. 4. Evolution of antibiotic resistance in seventh pandemic V. cholerae El Tor isolates from Africa. For each introduction into Africa (T1, T5 to T12), the proportion of genomes at different time periods following introduction that contain (A) an IncA/C plasmid, (B) a gyrA mutation, and (C) a SXT/R391 genomic island. (D) For wave 3, the mean number of antibiotic resistance genes (ARGs) per isolate at different time periods following introduction. Each value is calculated over a 10-year window

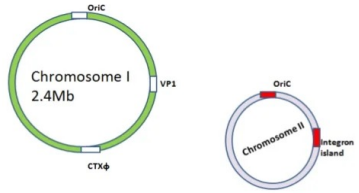
Prior bioinformatics analysis

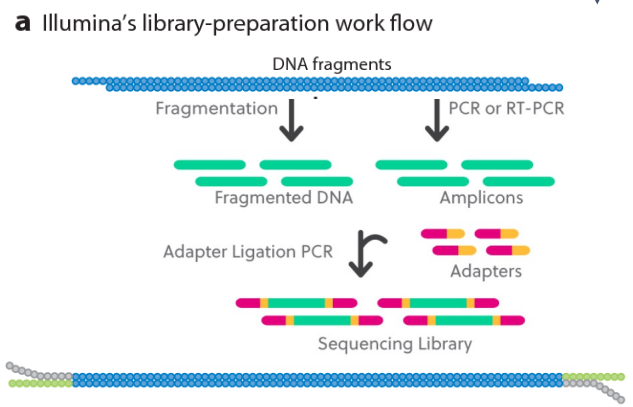
Culturing v. cholerae
In specific media

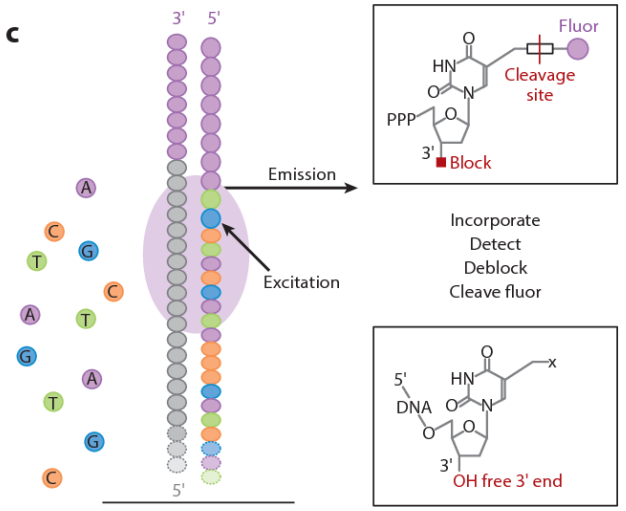
Person infected with *v. cholerae*

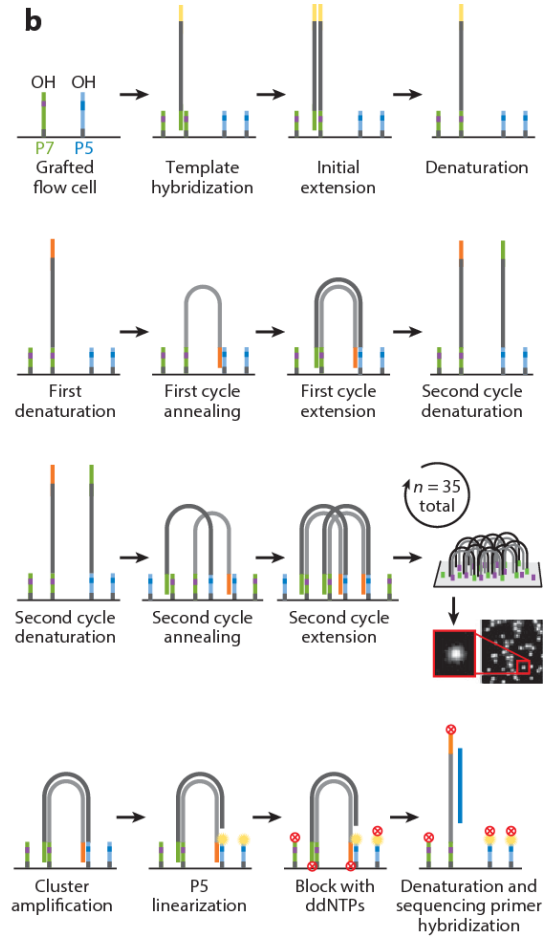
Extraction





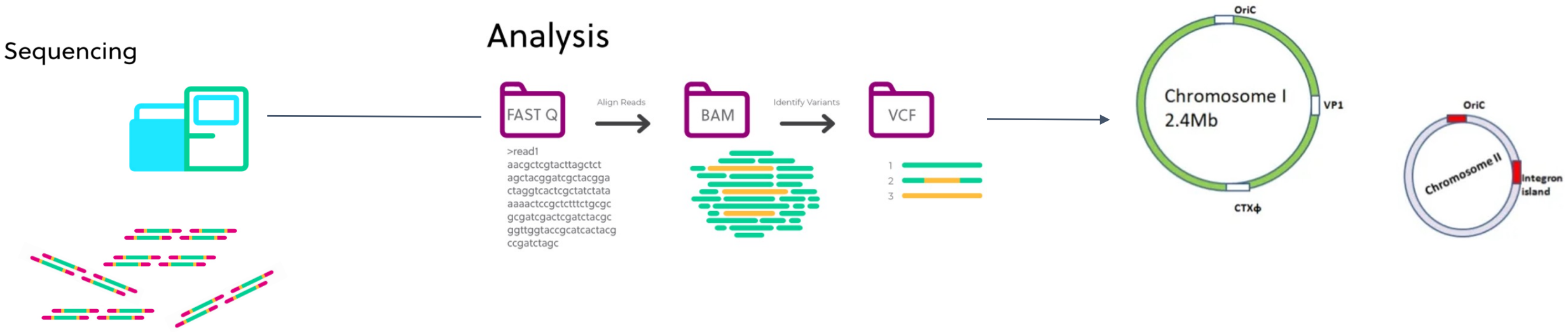






Illumina sequencing

Genomics made easy.



Remove adapters

1

Raw reads

Remove bad reads

2

Trimmed reads

Align to reference

3

.bam file

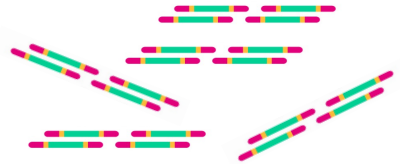
.vcf file

FASTA file

Post-Analyses:

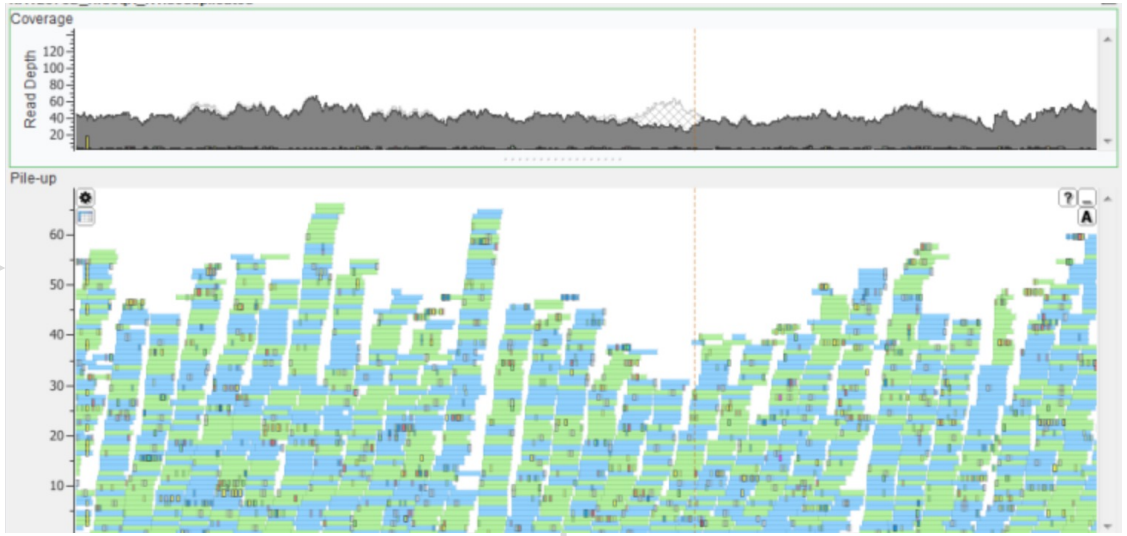
1. ML-tree rec
2. Mut-analys
3. Dated-tree

1

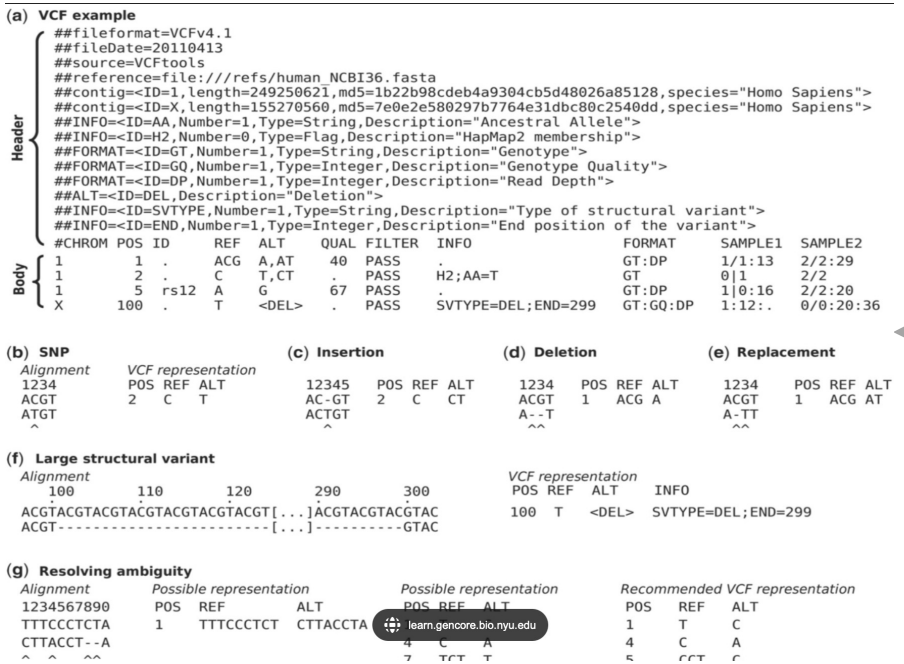


```
Identifier       @SRR566546.970 HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 length=50
Sequence         TTGCCTGCCTATCATTTTAGTGCCTGTGAGGTGGAGATGTGAGGATCAGT
'+' sign         +
Quality scores   hhhhhhhhhhghhghhhhhfhhhhhfffffe'ee['X]b[d[ed'[Y[^Y
Identifier       @SRR566546.971 HWUSI-EAS1673_11067_FC7070M:4:1:2374:1108 length=50
Sequence         GATTTGTATGAAAGTATACAACTAAAACTGCAGGTGGATCAGAGTAAGTC
'+' sign         +
Quality scores   hhhhgfhhcghghggfcffdhfehhhhcehdchhdhahehffffde'bVd
```

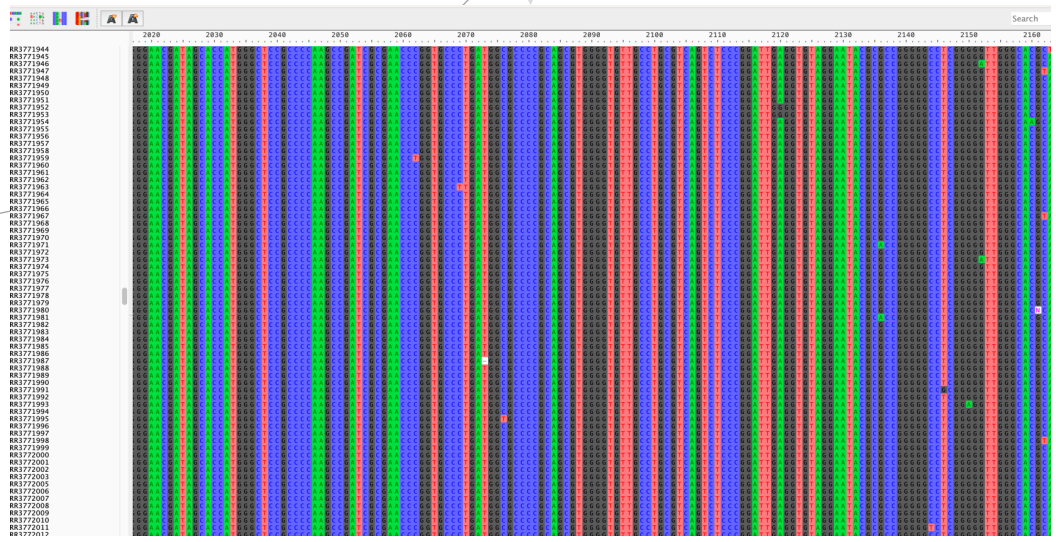
2



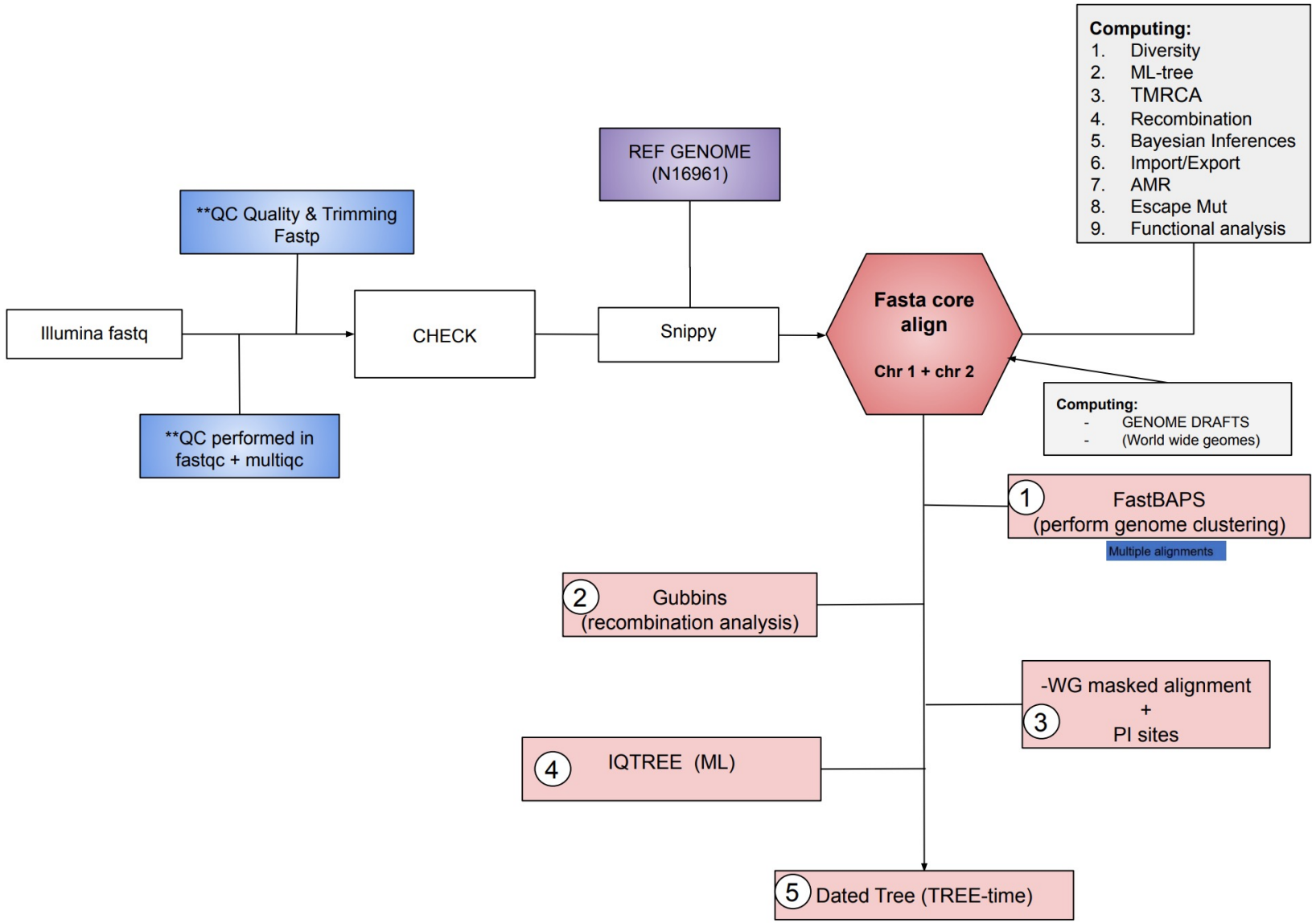
3a

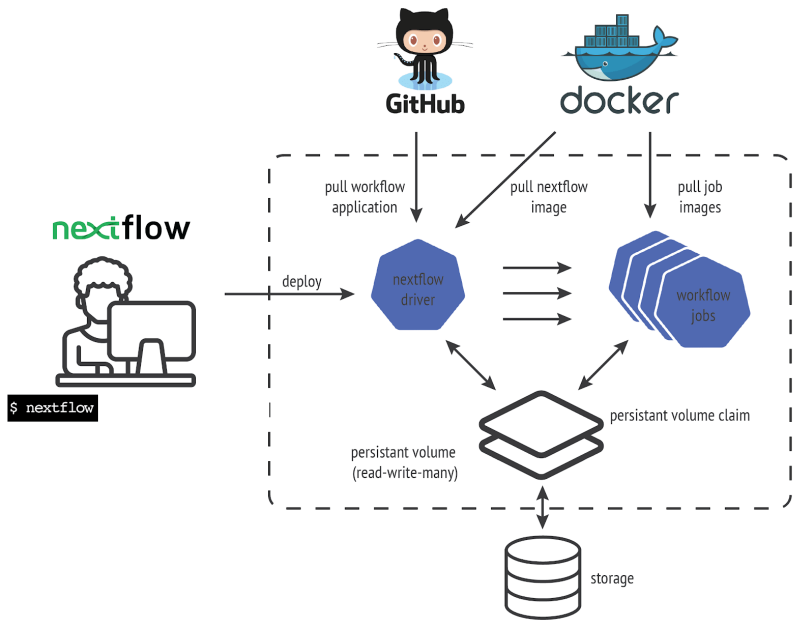


3b



Workflow: Application to cholera





Optimized workflow

- For pathogen genomics we choose workflows: less complicated; reproducible; shortest turn around time

1. fastp: an ultra-fast all-in-one FASTQ preprocessor.

```
usage: fastp -i <in1> -o <out1> [-I <in1> -O <out2>] [options...]
options:
```

2. Snippy: Rapid haploid variant calling and core genome alignment . Finds SNPs between a haploid reference genome and your NGS sequence reads (handles reads >500bp long). It will use as many CPUs as you can give it on a single computer (tested to 64 cores).

```
% snippy --cpus 16 --outdir mysnps --ref Listeria.gbk --R1 FDA_R1.fastq.gz --R2 FDA_R2.fastq.gz
<cut>
Walltime used: 3 min, 42 sec
Results folder: mysnps
Done.
```

% snippy-clean_full_aln core.full.aln > clean.full.aln

% run_gubbins.py -p gubbins clean.full.aln

% snp-sites -c gubbins.filtered_polymorphic_sites.fasta > clean.core.aln

% FastTree -gtr -nt clean.core.aln > clean.core.tree

3. IQTREE: takes as input a multiple sequence alignment and will reconstruct an evolutionary tree that is best explained by the input data..

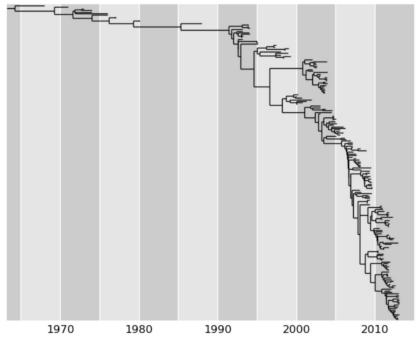
```
iqtree -s example.phy -m TIM2+I+G -alrt 1000 -B 1000
# for version 1.x change -B to -bb
```

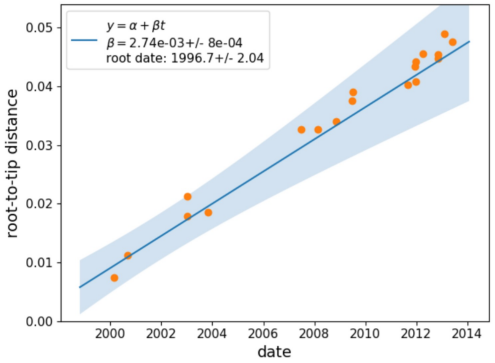
4. TreeTime: maximum likelihood dating and ancestral sequence inference

TreeTime provides routines for ancestral sequence reconstruction and inference of molecular-clock phylogenies, i.e., a tree where all branches are scaled such that the positions of terminal nodes correspond to their sampling times and internal nodes are placed at the most likely time of divergence.

```
treetime clock --tree <input.nwk> --aln <input.fasta> --dates <dates.csv> --reroot least-squa
```

```
treetime mugration --tree <input.nwk> --states <states.csv> --attribute <field>
```





Summary

- V, cholera is large (~> 4MB), hence large computation power is required to analysis the genome.
- NGS Workflow for bacteria such as v. cholerae is a cascade of multiple tools.
- We are working on Nexflow (automated workflow) to easily analyse sequences from fastq to dated tree.

Thank you..