BIOS27815: Infectious Diseases ***– Reading Recap and Discussion –***

Introduction to metagenomic Next Generation Sequencing (mNGS) in Global Health

UChicago Center in Paris

Paris, France

January 2025

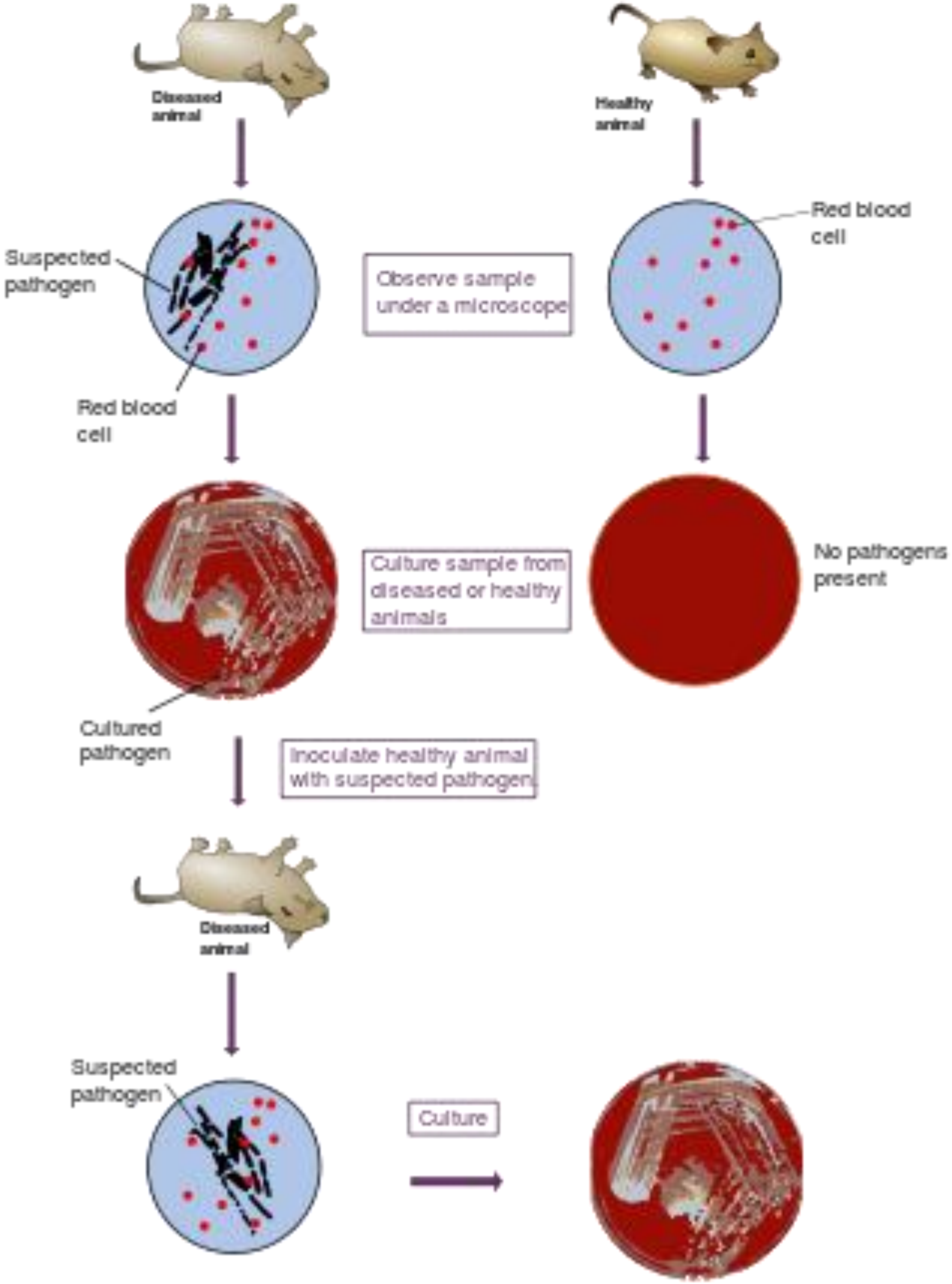
Goals for this lecture

- To introduce mNGS in Global Health
- To introduce and interpret Bohl et al. 2022

Goals for this lecture

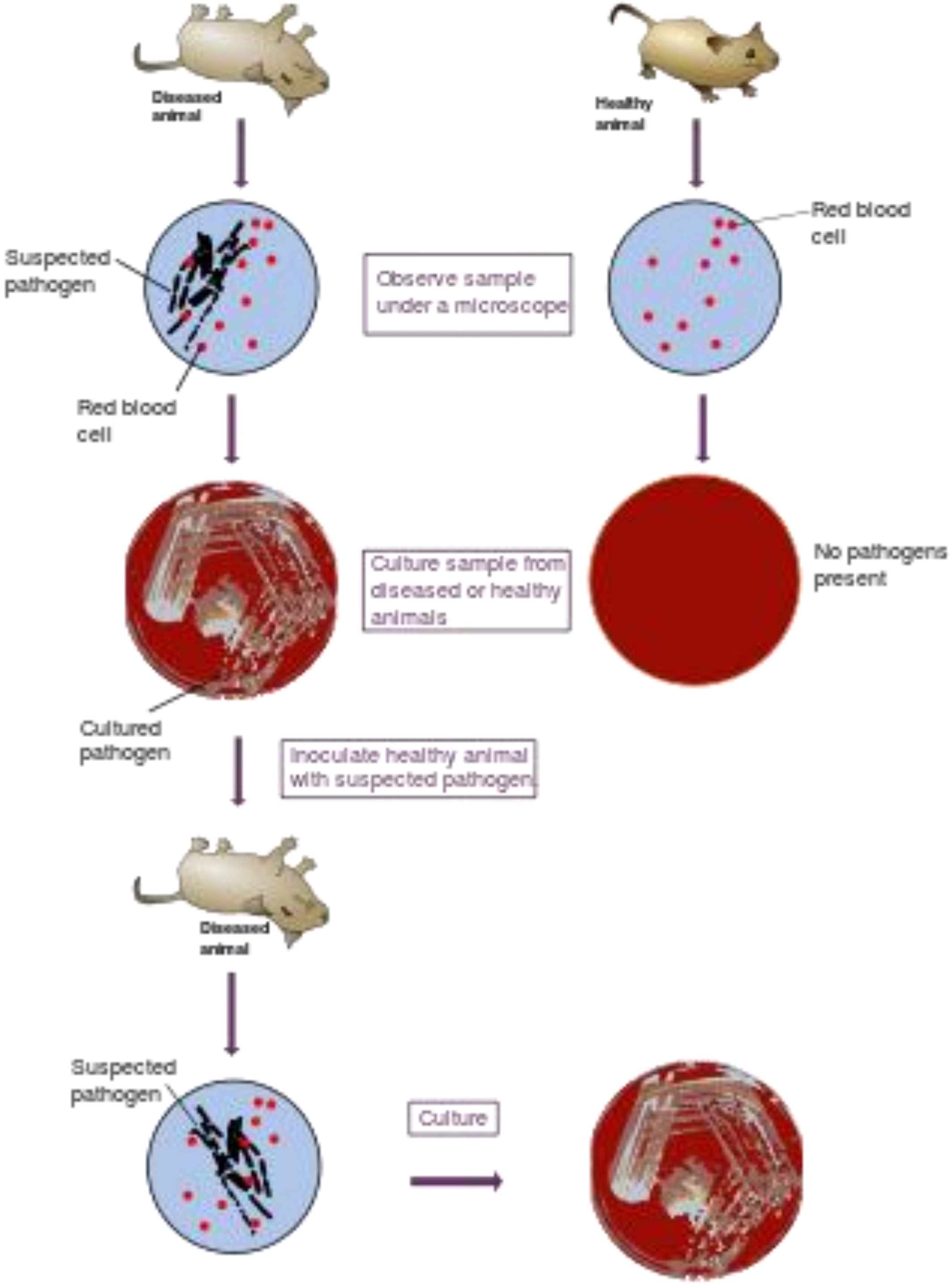
- To introduce mNGS in Global Health
- To introduce and interpret Bohl et al. 2022

Remember Koch’s postulates….



Remember Koch’s postulates….

- Pathogens and parasites cause disease.



Observe sample under a microscope

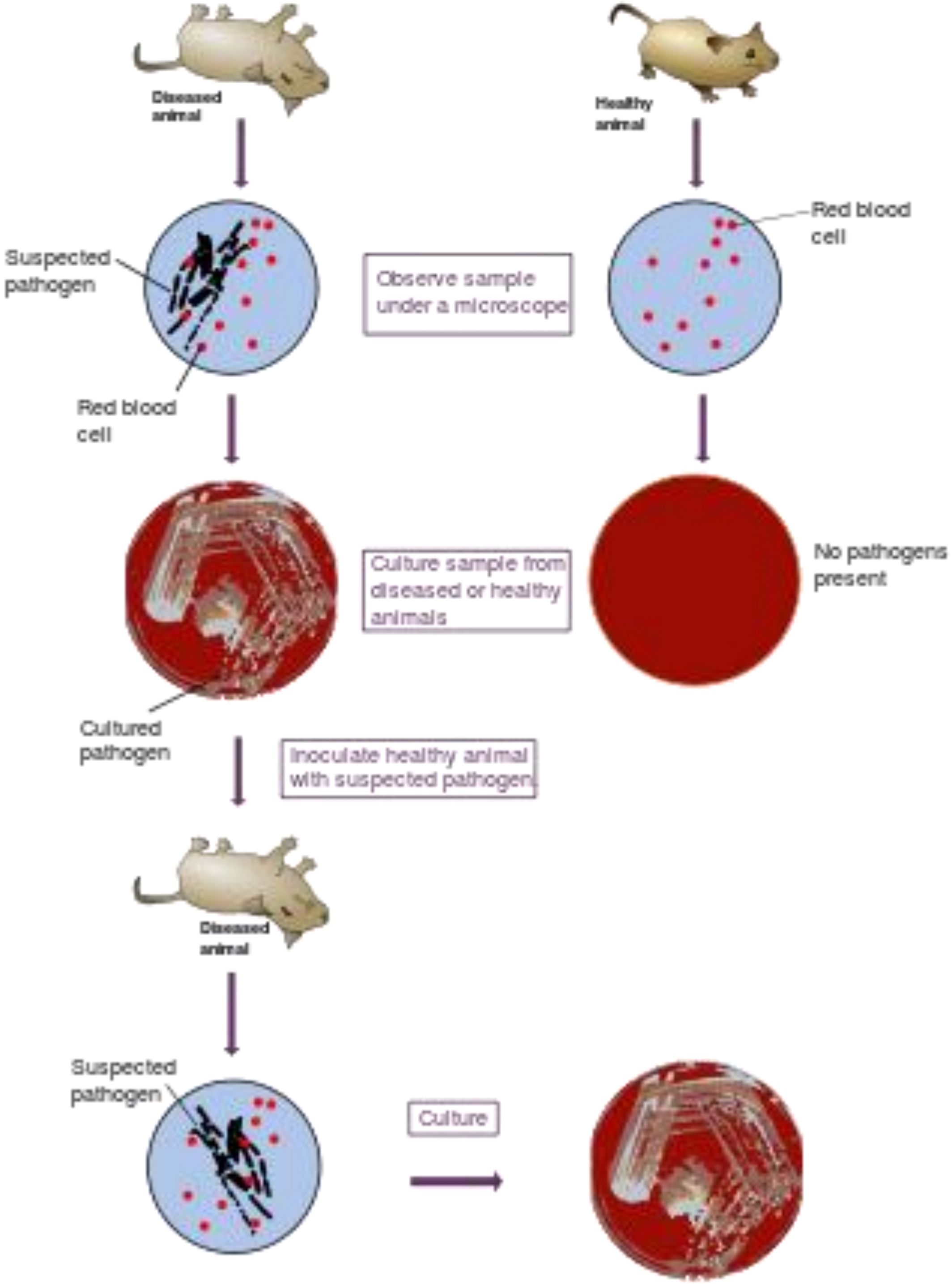
Culture sample from diseased or healthy animals

Inoculate healthy animal with suspected pathogen

Culture

Remember Koch’s postulates….

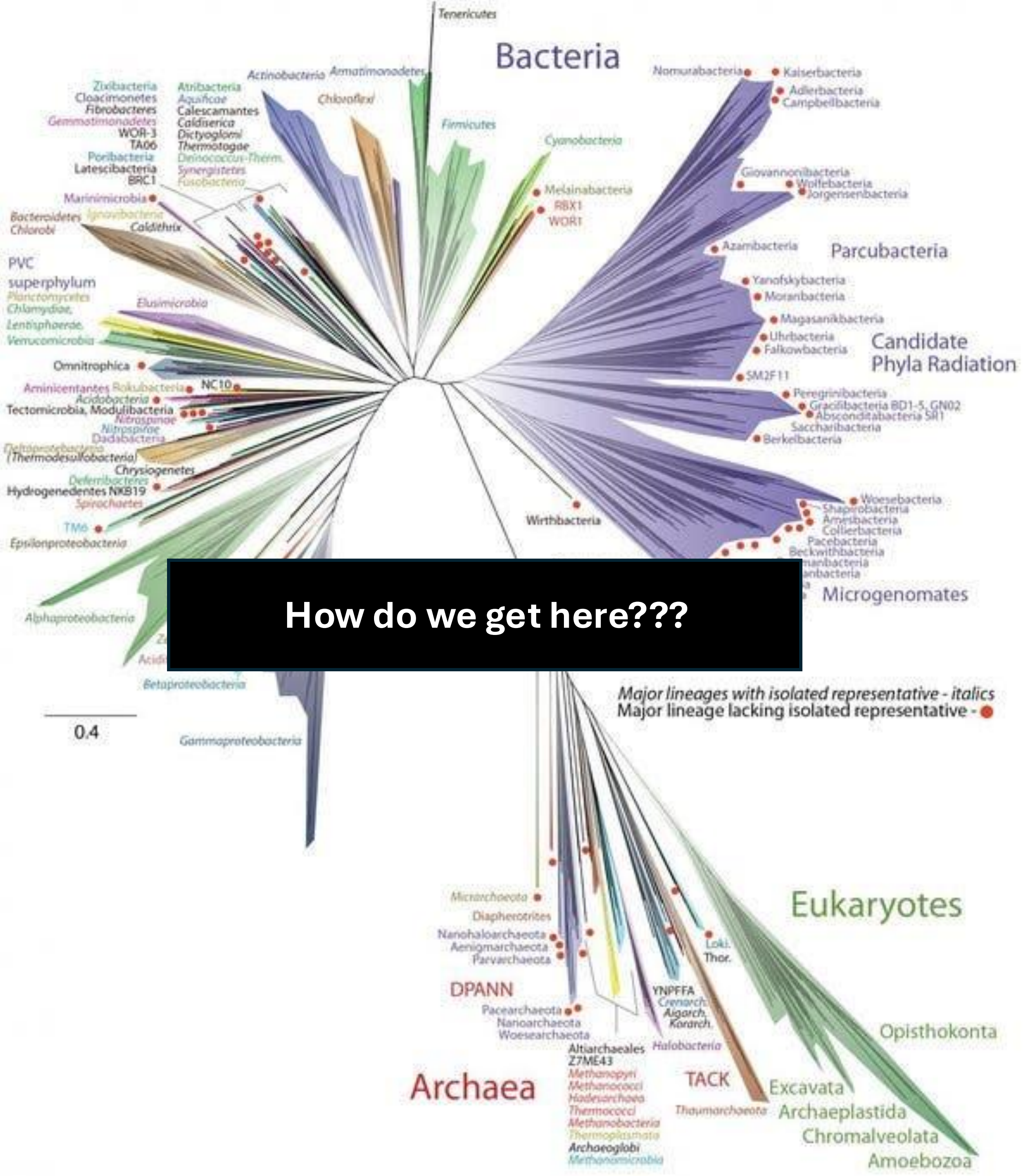
- Pathogens and parasites cause disease.
- But how do we know *what* pathogen causes *which* disease?



Remember Koch's postulates....

- Pathogens and parasites cause disease.

- But how do we know *what* pathogen causes *which* disease?



How do we get here???

How do we know *what* pathogen causes *which* disease?

How do we know *what* pathogen causes *which* disease?

- microscopy
- culture with microscopy
- antibody tests (e.g. ELISAs)
- nucleic acid tests (e.g. PCR)
- sequencing

How do we know *what* pathogen causes *which* disease?

- microscopy
- culture with microscopy
- antibody tests (e.g. ELISAs)
- nucleic acid tests (e.g. PCR)
- sequencing

Only sequencing gives us a unique “name” for each pathogen.

How do we know *what* pathogen causes *which* disease?

- microscopy
- culture with microscopy
- antibody tests (e.g. ELISAs)
- nucleic acid tests (e.g. PCR)
- sequencing

Only sequencing gives us a unique “name” for each pathogen.

*‘**Library prep**’ is the preparation of a sample for sequencing.*

What is sequencing?

What is sequencing?

- Sequencing determines the order of the four nucleotide bases (A, T, G and C) that make up DNA

What is sequencing?

- Sequencing determines the order of the four nucleotide bases (A, T, G and C) that make up DNA
- The first DNA sequences were obtained in the early 1970s. Since then, technology has advanced significantly

What is sequencing?

- Sequencing determines the order of the four nucleotide bases (A, T, G and C) that make up DNA
- The first DNA sequences were obtained in the early 1970s. Since then, technology has advanced significantly
- Human Genome Project: October 1990 – April 2003
  - $2.7 billion
  - Mostly Sanger sequencing
- Today, human genomes are sequenced rapidly and cheaply ($100s) and much less for smaller organisms (e.g. viruses)
  - Often using ‘Next Generation Sequencing’ (NGS) techniques

What is sequencing?

Sanger

Illumina
(short-read)

Nanopore
(long-read)

Sanger Sequencing

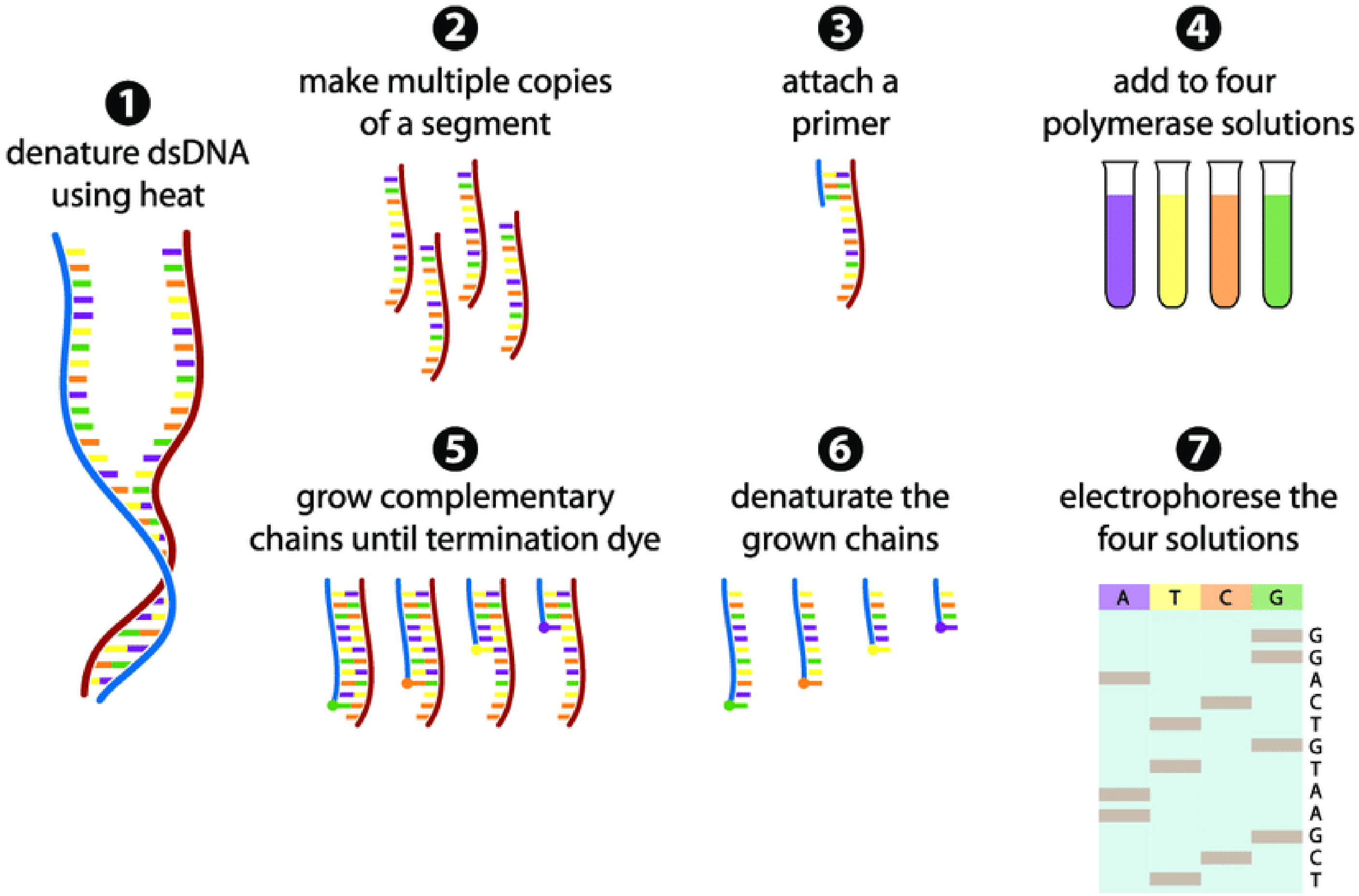
- The DNA sample is divided into four separate reactions, containing all four of the standard deoxynucleotides (dATP, dGTP, dCTP and dTTP) and a DNA polymerase (which attached the dNTPs)

- To each reaction is added only one dideoxynucleotides (ddATP, ddGTP, ddCTP, or ddTTP)

- Four separate reactions are needed in this process to test all four ddNTPs

- The ddNTP stops the DNA polymerase when it comes to a base of that type (e.g. A, T, G, C)

- The fragments are then run on a gel. The smallest move through the gel furthest and the ‘ladder’ shows the sequence of the DNA



1 denature dsDNA using heat

2 make multiple copies of a segment

3 attach a primer

4 add to four polymerase solutions

5 grow complementary chains until termination dye

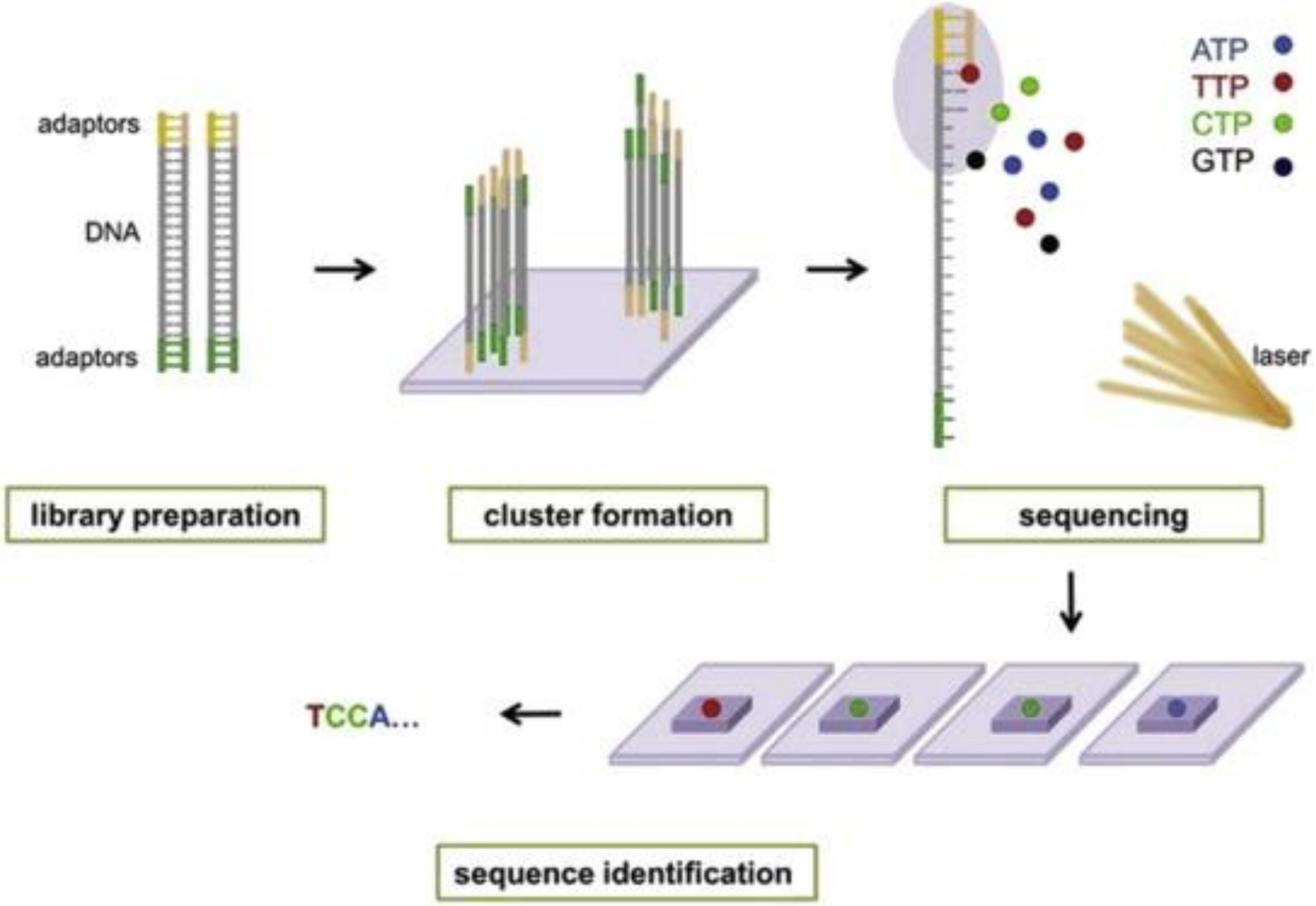
6 denaturate the grown chains

7 electrophorese the four solutions

https://www.youtube.com/watch?v=FvHRio1yyhQ&t=171s

Illumina (short read) Sequencing

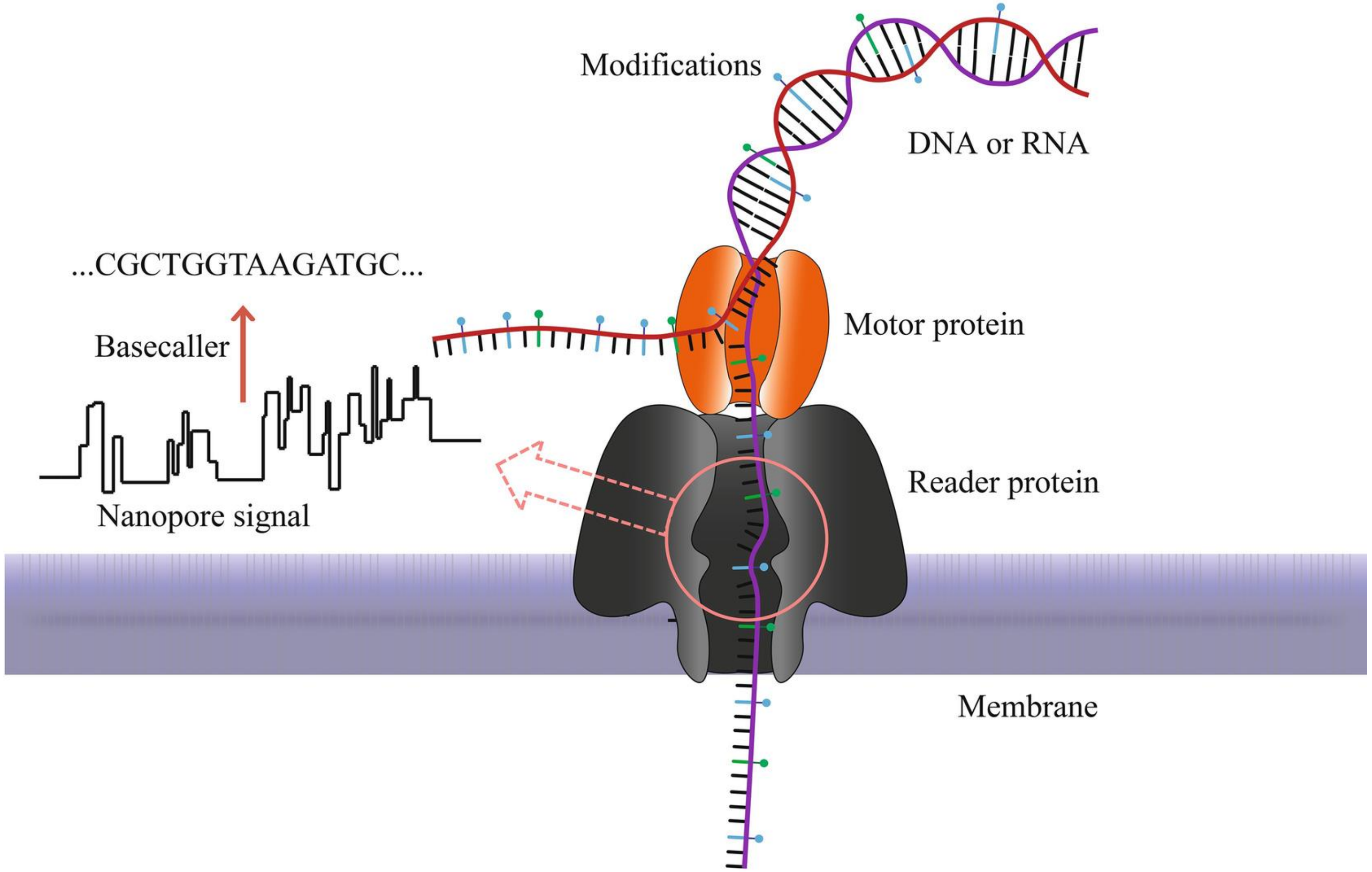
- Sequencing by synthesis
- 'Short read' technology – DNA is cut up into 200-600 bp chunks
- The DNA is amplified, so there are lots of copies of the chunks
- They are denatured, then fluorescent complimentary bases are attached
- These fluoresce different colours, which is recorded, and the sequence is identified



https://www.youtube.com/watch?v=womKfikWlxM

Nanopore (long read) Sequencing

- ‘Long read’ technology
- A DNA library is prepared (proteins are added)
- Nucleic acids are passed through a protein nanopore
- As the different bases move through the nanopore, it creates a different electrical signal
- These resulting changes in the electrical signal is decoded to provide the specific DNA or RNA sequence



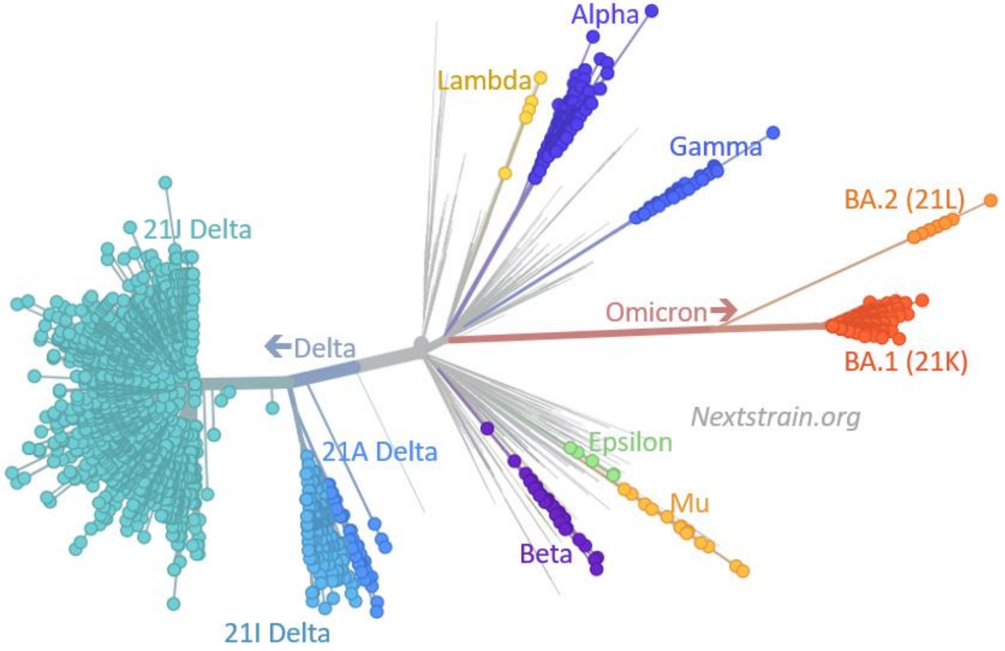
https://www.youtube.com/watch?v=RcP85JHLmnI&t=18s

Where do these sequences go once we have them?

- Nextclade: genome quality and curation
- GISAID: Global Initiative on Sharing All Influenza Data
- NCBI: National Center for Biotechnology Information

Where do these sequences go once we have them?

- Nextclade: genome quality and curation
- GISAID: Global Initiative on Sharing All Influenza Data
- NCBI: National Center for Biotechnology Information



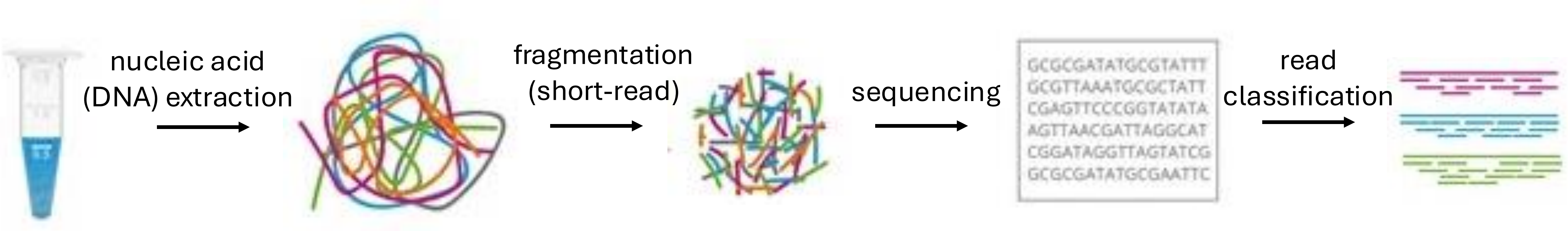
Publicly available 'background' sequences are critical for public health inference.

Lowe 2022

What is **metagenomic NGS** (**mNGS**)?

What is **metagenomic NGS** (**mNGS**)?

- mNGS is the sequencing of ALL genetic material recovered directly from environmental or clinical samples.



nucleic acid (DNA) extraction

fragmentation (short-read)

sequencing

read classification

What is **metagenomic NGS** (**mNGS**)?

- mNGS is the sequencing of ALL genetic material recovered directly from environmental or clinical samples.
- This is in contrast to ‘amplicon’ sequencing which uses primer targets to sequence only material of an organism (e.g. pathogen) of interest.

What is **metagenomic NGS** (**mNGS**)?

- mNGS is the sequencing of ALL genetic material recovered directly from environmental or clinical samples.
- This is in contrast to ‘amplicon’ sequencing which uses primer targets to sequence only material of an organism (e.g. pathogen) of interest.
- mNGS requires both laboratory and bioinformatics expertise.

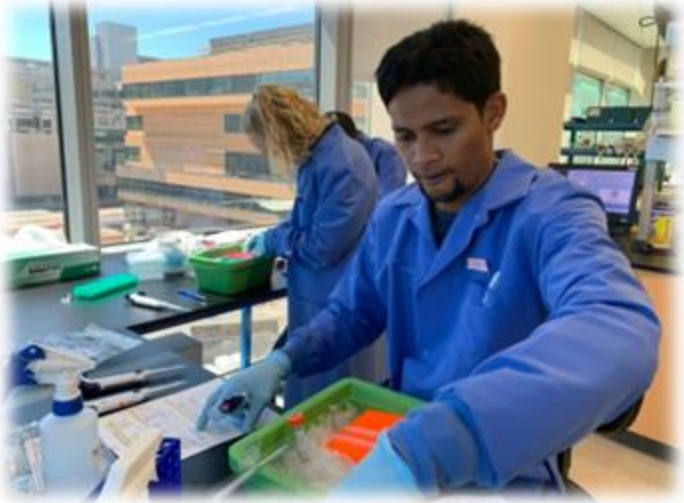
What is **metagenomic NGS** (**mNGS**)?

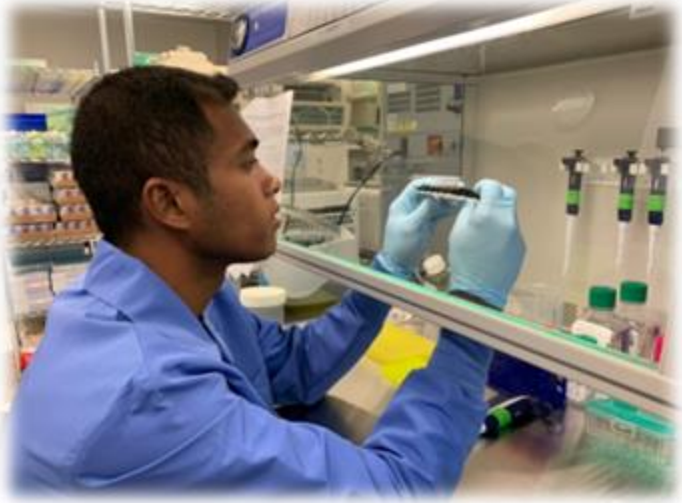
- mNGS is the sequencing of ALL genetic material recovered directly from environmental or clinical samples.
- This is in contrast to 'amplicon' sequencing which uses primer targets to sequence only material of an organism (e.g. pathogen) of interest.
- mNGS requires both laboratory and bioinformatics expertise.
- mNGS is particularly useful for identifying the etiology of viral disease because there is no single gene common across all viral genomes (in contrast to bacteria and protozoa, e.g. 16S)

Why mNGS for LMICs?

BILL & MELINDA GATES foundation

- One protocol with nearly limitless applications!
  - No need for multiple kits and reagents to be ordered and testing
  - Training pipeline is discrete and powerful
- Computational tools for accessibility are becoming more readily available (e.g. CZID).
- Still many challenges:
  - Expense!
  - Supply chain
  - Difficulty of bioinformatics analysis

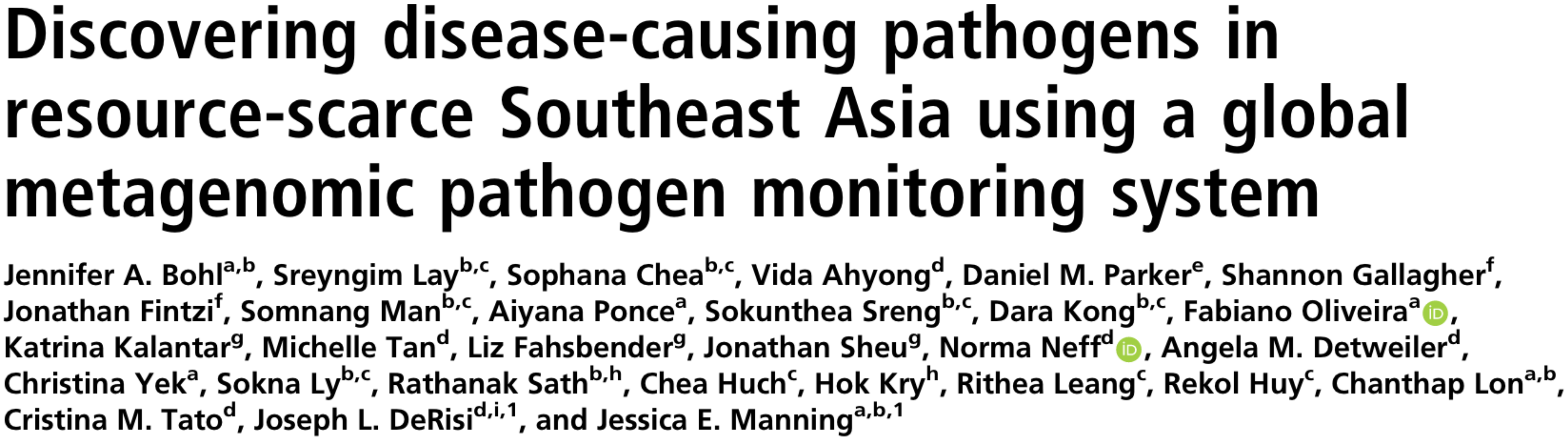




Goals for this lecture

- To introduce mNGS in Global Health
- To introduce and interpret Bohl et al. 2022

mNGS in Global Health:



Discovering disease-causing pathogens in resource-scarce Southeast Asia using a global metagenomic pathogen monitoring system

Jennifer A. Bohl[a,b], Sreyngim Lay[b,c], Sophana Chea[b,c], Vida Ahyong[d], Daniel M. Parker[e], Shannon Gallagher[f], Jonathan Fintzi[f], Somnang Man[b,c], Aiyana Ponce[a], Sokunthea Sreng[b,c], Dara Kong[b,c], Fabiano Oliveira[a], Katrina Kalantar[g], Michelle Tan[d], Liz Fahsbender[g], Jonathan Sheu[g], Norma Neff[d], Angela M. Detweiler[d], Christina Yek[a], Sokna Ly[b,c], Rathanak Sath[b,h], Chea Huch[c], Hok Kry[h], Rithea Leang[c], Rekol Huy[c], Chanthap Lon[a,b], Cristina M. Tato[d], Joseph L. DeRisi[d,i,1], and Jessica E. Manning[a,b,1]

What did Bohl et al. do?

What did Bohl et al. do?

- Used mNGS to identify the etiology of febrile disease among patients in periurban Kampong Speu province, Cambodia
- Add map from Berkeley talk

How did Bohl et al. do it?

Field

- Collected serum samples from febrile patients in Kampong Speu, Cambodia
  - Patients 6 months – 65 years
  - Fever 38+^{0}C
- Samples derived from two groups:
  - ‘community’ study (childhood cohort study + ‘sick visits’)
  - ‘hospital’ study
- 23 afebrile child control samples from community study

How did Bohl et al. do it?

Field

- Collected serum samples from febrile patients in Kampong Speu, Cambodia
  - Patients 6 months – 65 years
  - Fever 38+^{0}C
- Samples derived from two groups:
  - ‘community’ study (childhood cohort study + ‘sick visits’)
  - ‘hospital’ study
- 23 afebrile child control samples from community study

Lab

1. Sample 5ml whole blood
2. Centrifuge to serum
3. RNA extraction
4. Library preparation
5. mNGS
   - (with host RNA deletion)
6. Clinical validation where possible

How did Bohl et al. do it?

Bioinformatics

- CZID
- Pathogen ID via Z-score criteria
- Collection of geospatial Google Earth data

How did Bohl et al. do it?

Bioinformatics

- CZID
- Pathogen ID via Z-score criteria
- Collection of geospatial Google Earth data

Statistics

- Response: infection with a vector-borne pathogen
- Predictors: Demographic attributes of the patient + geospatial features of the patient’s locality

Why did Bohl et al. perform their study?

Why did Bohl et al. perform their study?

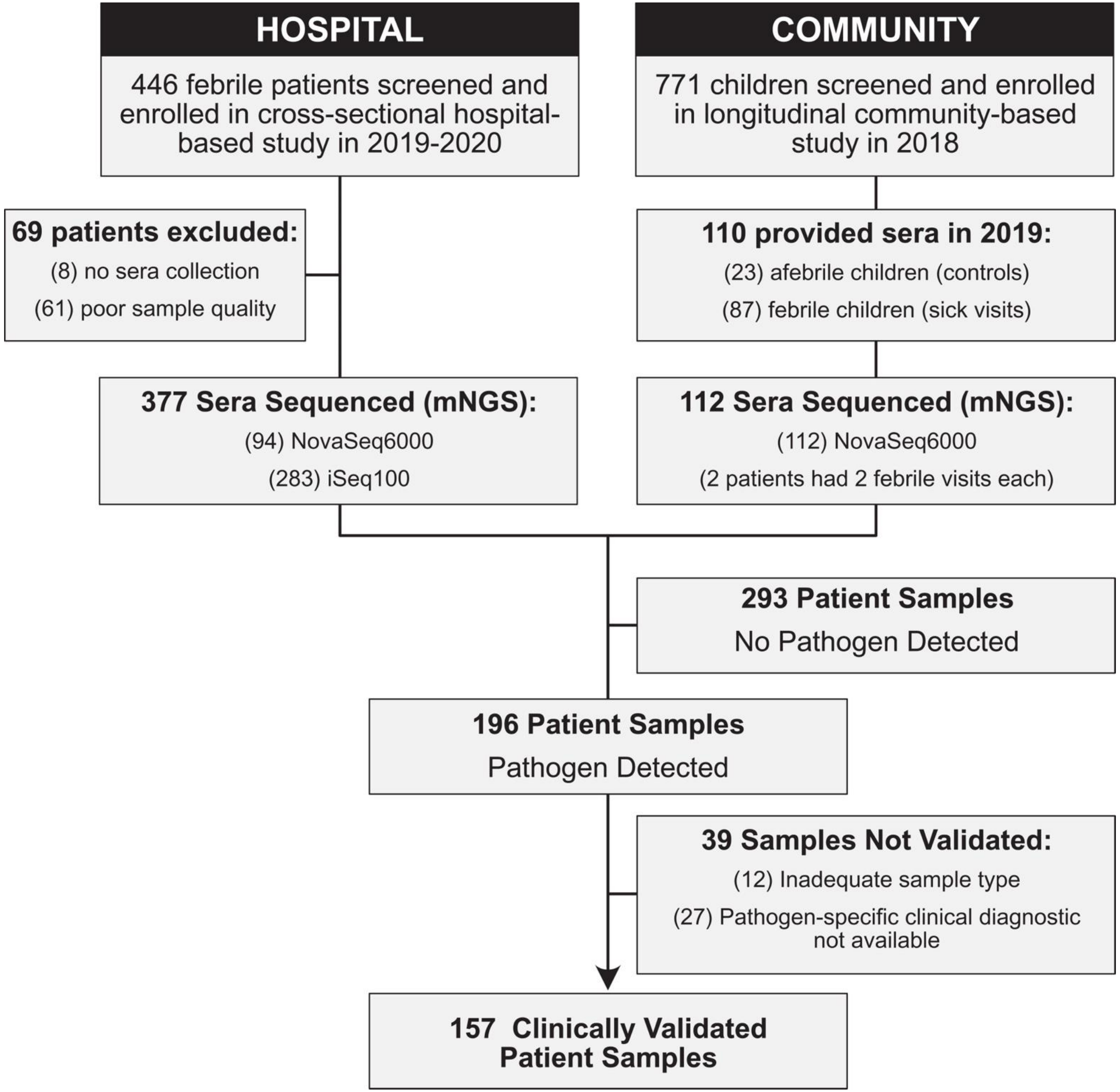
- To identify unrecognized causes of fever in an under-resourced setting

Why did Bohl et al. perform their study?

- To identify unrecognized causes of fever in an under-resourced setting

What did they find?

Results.



HOSPITAL

446 febrile patients screened and enrolled in cross-sectional hospital-based study in 2019-2020

69 patients excluded:

(8) no sera collection

(61) poor sample quality

377 Sera Sequenced (mNGS):

(94) NovaSeq6000

(283) iSeq100

COMMUNITY

771 children screened and enrolled in longitudinal community-based study in 2018

110 provided sera in 2019:

(23) afebrile children (controls)

(87) febrile children (sick visits)

112 Sera Sequenced (mNGS):

(112) NovaSeq6000

(2 patients had 2 febrile visits each)

293 Patient Samples

No Pathogen Detected

196 Patient Samples

Pathogen Detected

39 Samples Not Validated:

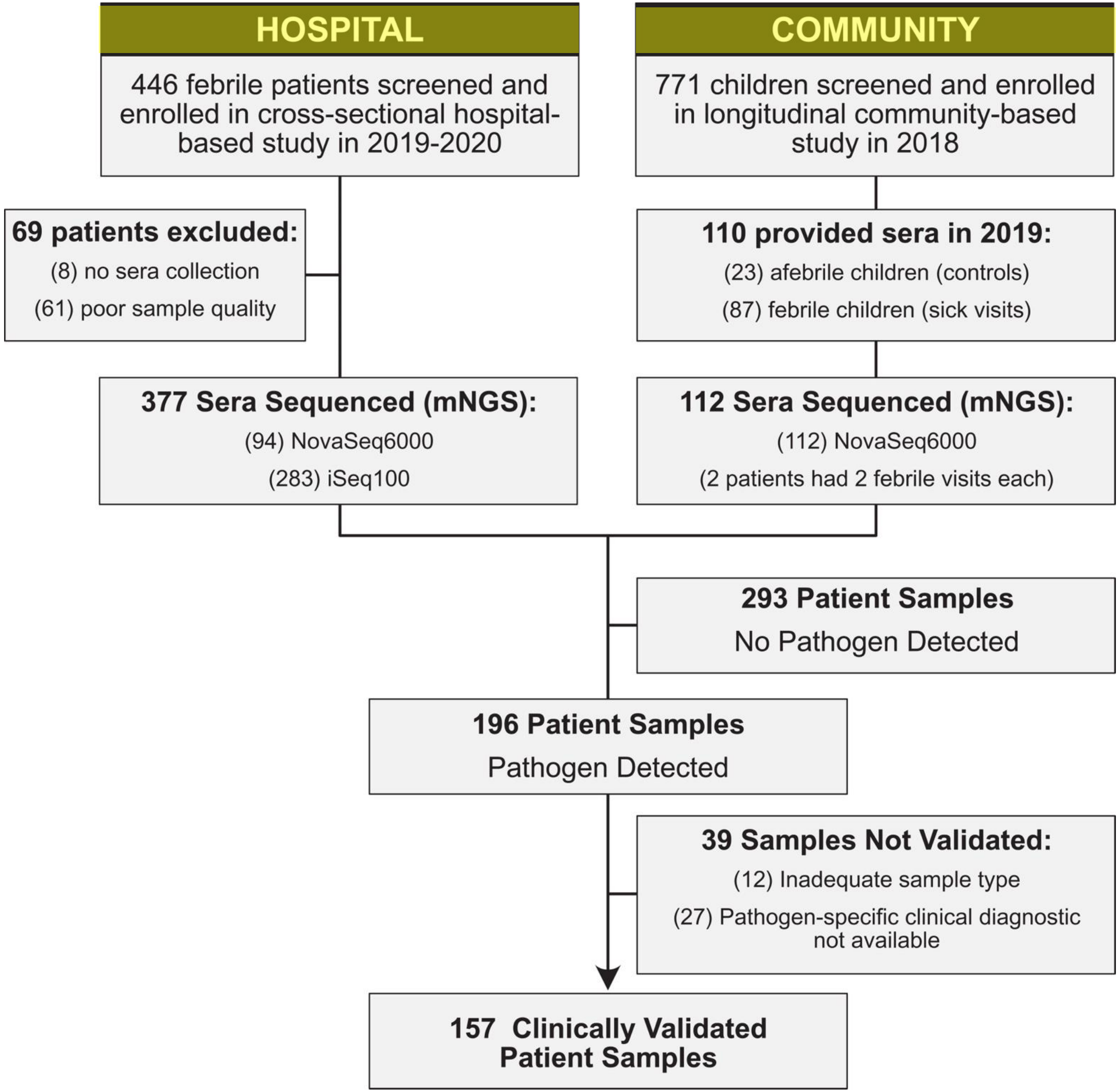
(12) Inadequate sample type

(27) Pathogen-specific clinical diagnostic not available

157 Clinically Validated Patient Samples

Fig. 1: study design

Results.



HOSPITAL

446 febrile patients screened and enrolled in cross-sectional hospital-based study in 2019-2020

69 patients excluded:

(8) no sera collection

(61) poor sample quality

377 Sera Sequenced (mNGS):

(94) NovaSeq6000

(283) iSeq100

COMMUNITY

771 children screened and enrolled in longitudinal community-based study in 2018

110 provided sera in 2019:

(23) afebrile children (controls)

(87) febrile children (sick visits)

112 Sera Sequenced (mNGS):

(112) NovaSeq6000

(2 patients had 2 febrile visits each)

293 Patient Samples

No Pathogen Detected

196 Patient Samples

Pathogen Detected

39 Samples Not Validated:

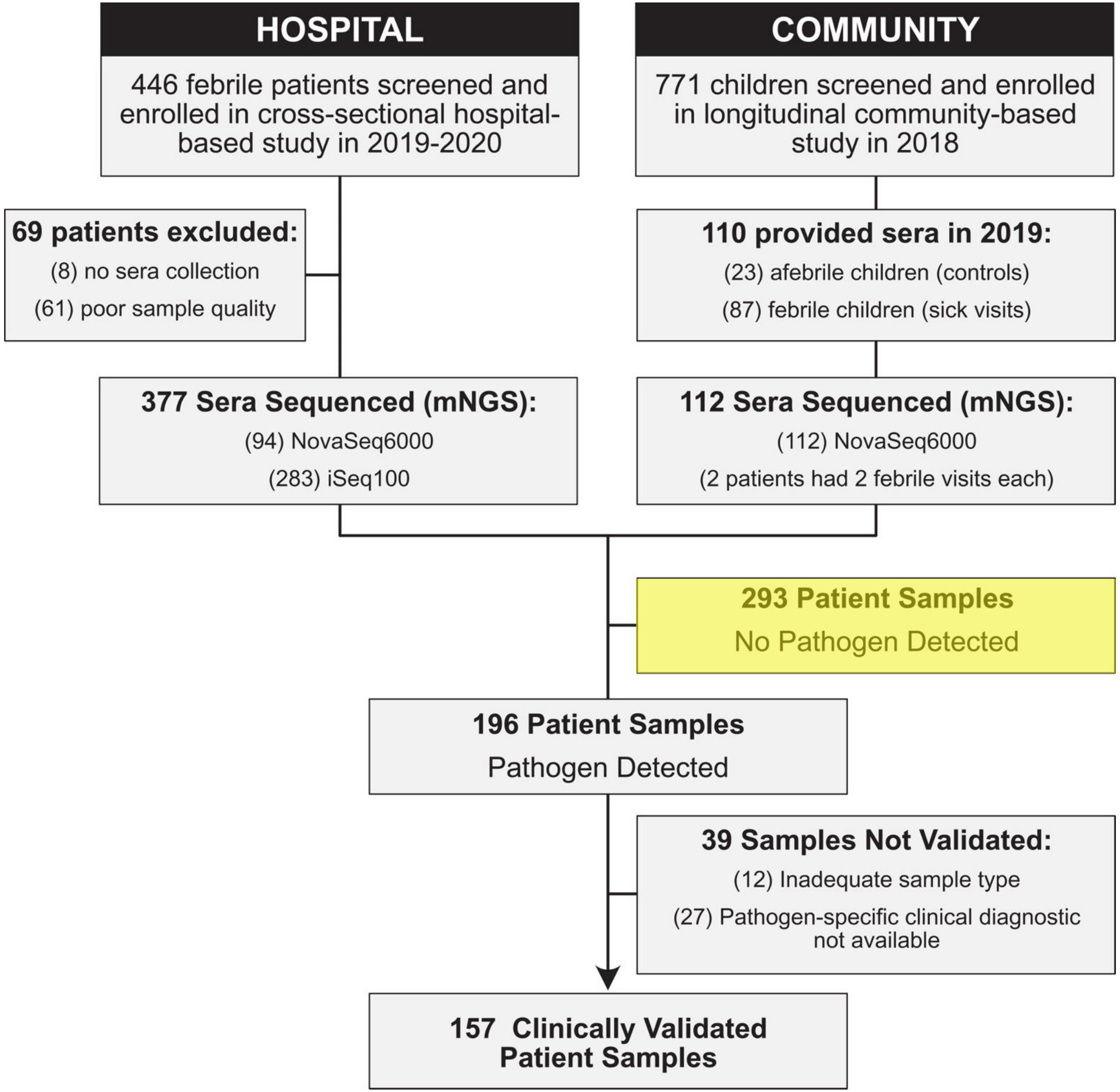
(12) Inadequate sample type

(27) Pathogen-specific clinical diagnostic not available

157 Clinically Validated Patient Samples

Fig. 1: study design

Results.



HOSPITAL

446 febrile patients screened and enrolled in cross-sectional hospital-based study in 2019-2020

69 patients excluded:
(8) no sera collection
(61) poor sample quality

377 Sera Sequenced (mNGS):
(94) NovaSeq6000
(283) iSeq100

COMMUNITY

771 children screened and enrolled in longitudinal community-based study in 2018

110 provided sera in 2019:
(23) afebrile children (controls)
(87) febrile children (sick visits)

112 Sera Sequenced (mNGS):
(112) NovaSeq6000
(2 patients had 2 febrile visits each)

293 Patient Samples
No Pathogen Detected

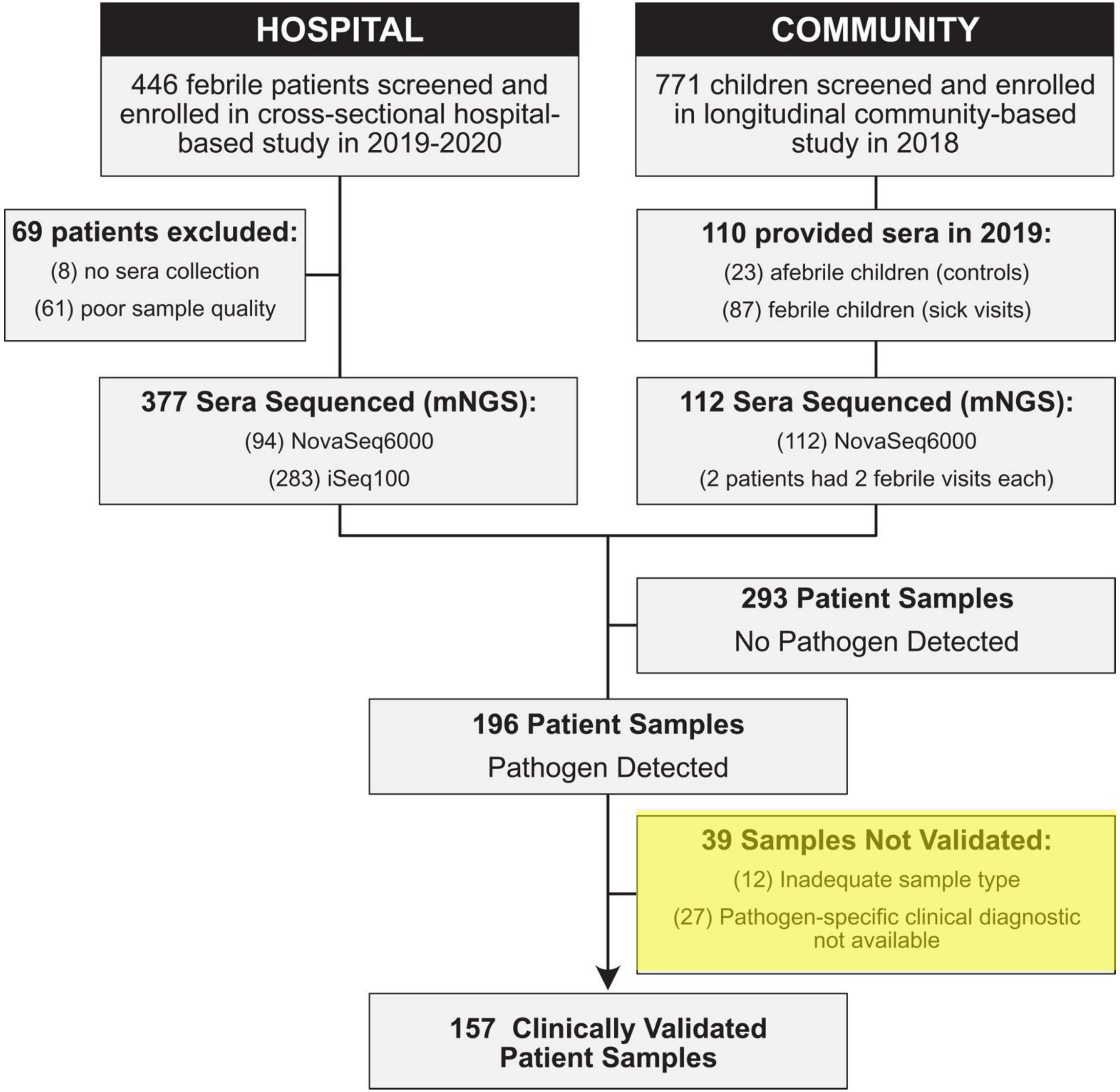
196 Patient Samples
Pathogen Detected

39 Samples Not Validated:
(12) Inadequate sample type
(27) Pathogen-specific clinical diagnostic not available

157 Clinically Validated Patient Samples

Fig. 1: study design

Results.



HOSPITAL

446 febrile patients screened and enrolled in cross-sectional hospital-based study in 2019-2020

69 patients excluded:
(8) no sera collection
(61) poor sample quality

377 Sera Sequenced (mNGS):
(94) NovaSeq6000
(283) iSeq100

COMMUNITY

771 children screened and enrolled in longitudinal community-based study in 2018

110 provided sera in 2019:
(23) afebrile children (controls)
(87) febrile children (sick visits)

112 Sera Sequenced (mNGS):
(112) NovaSeq6000
(2 patients had 2 febrile visits each)

293 Patient Samples
No Pathogen Detected

196 Patient Samples
Pathogen Detected

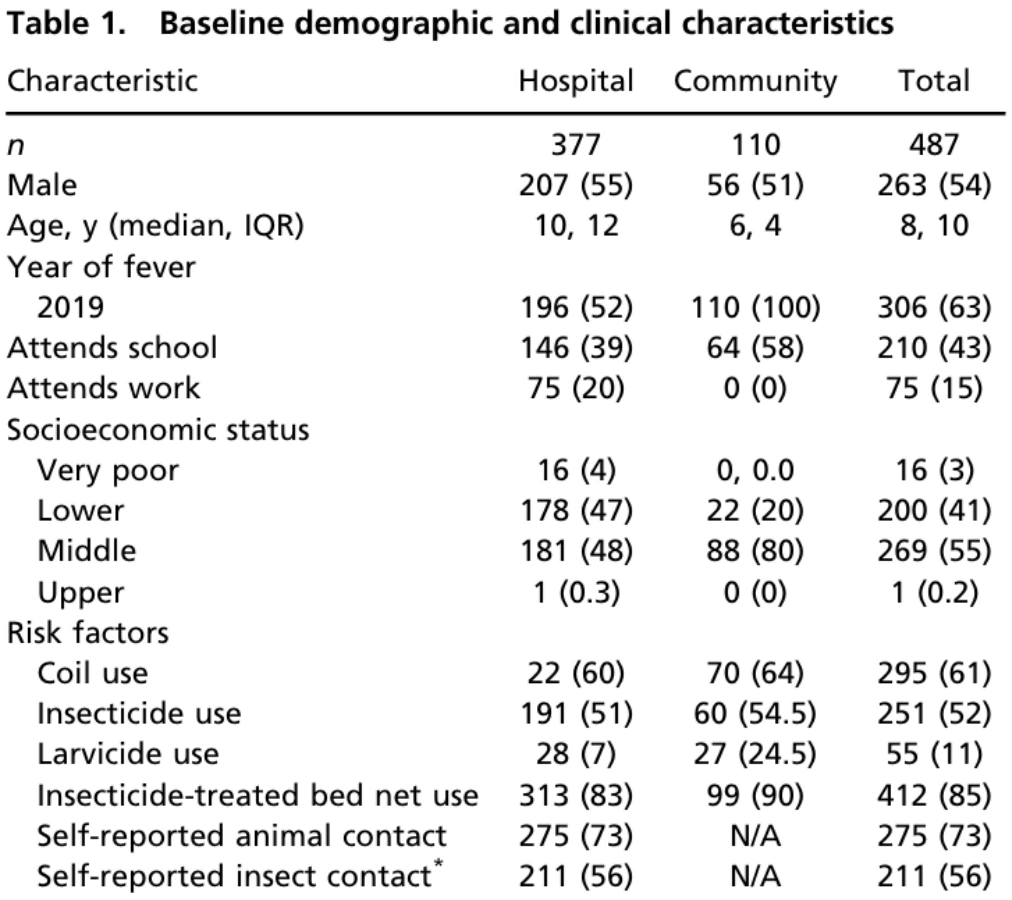
39 Samples Not Validated:
(12) Inadequate sample type
(27) Pathogen-specific clinical diagnostic not available

157 Clinically Validated Patient Samples

Fig. 1: study design

Results.

Table 1: cohort characteristics



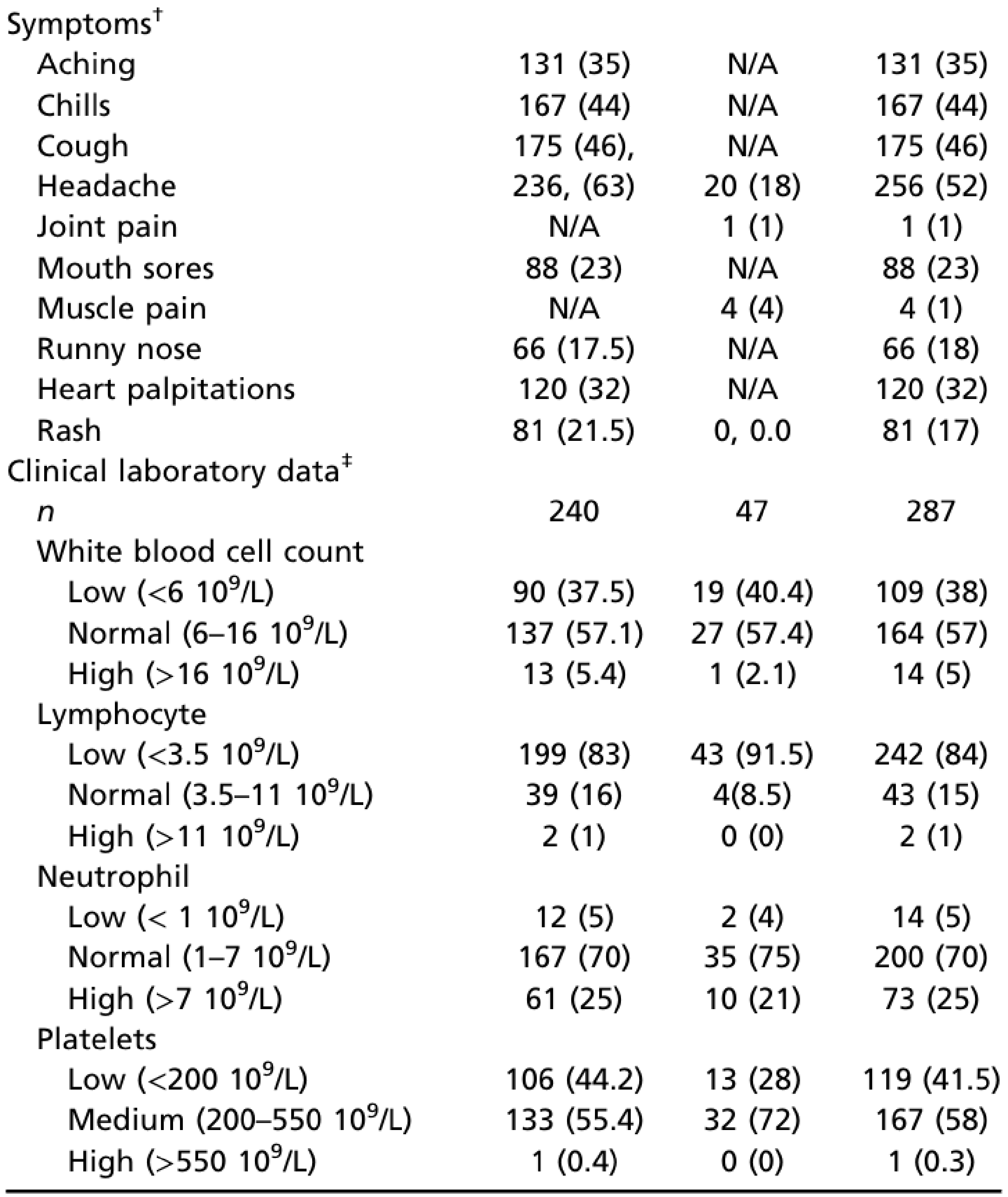
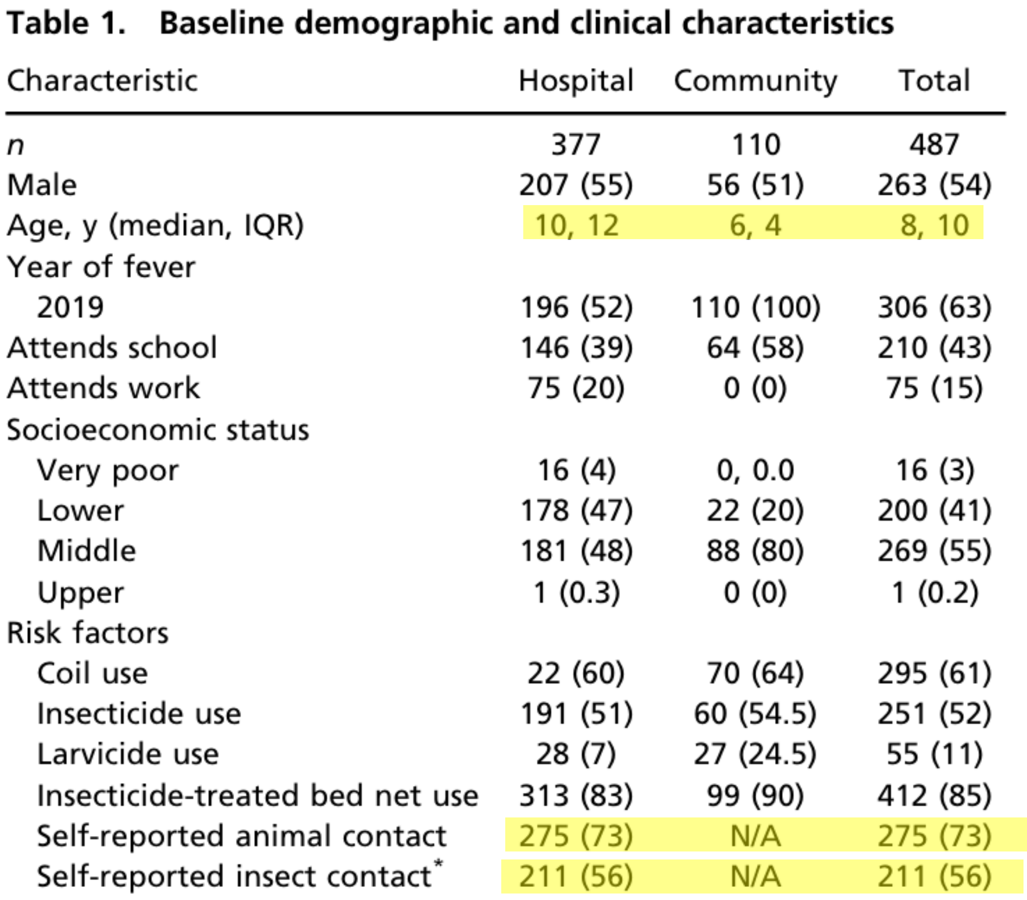


Table 1. Baseline demographic and clinical characteristics

Characteristic	Hospital	Community	Total
n	377	110	487
Male	207 (55)	56 (51)	263 (54)
Age, y (median, IQR)	10, 12	6, 4	8, 10
Year of fever			
2019	196 (52)	110 (100)	306 (63)
Attends school	146 (39)	64 (58)	210 (43)
Attends work	75 (20)	0 (0)	75 (15)
Socioeconomic status			
Very poor	16 (4)	0, 0.0	16 (3)
Lower	178 (47)	22 (20)	200 (41)
Middle	181 (48)	88 (80)	269 (55)
Upper	1 (0.3)	0 (0)	1 (0.2)
Risk factors			
Coil use	22 (60)	70 (64)	295 (61)
Insecticide use	191 (51)	60 (54.5)	251 (52)
Larvicide use	28 (7)	27 (24.5)	55 (11)
Insecticide-treated bed net use	313 (83)	99 (90)	412 (85)
Self-reported animal contact	275 (73)	N/A	275 (73)
Self-reported insect contact*	211 (56)	N/A	211 (56)
Symptoms†			
Aching	131 (35)	N/A	131 (35)
Chills	167 (44)	N/A	167 (44)
Cough	175 (46),	N/A	175 (46)
Headache	236, (63)	20 (18)	256 (52)
Joint pain	N/A	1 (1)	1 (1)
Mouth sores	88 (23)	N/A	88 (23)
Muscle pain	N/A	4 (4)	4 (1)
Runny nose	66 (17.5)	N/A	66 (18)
Heart palpitations	120 (32)	N/A	120 (32)
Rash	81 (21.5)	0, 0.0	81 (17)
Clinical laboratory data‡			
n	240	47	287
White blood cell count			
Low (<6 10^9/L)	90 (37.5)	19 (40.4)	109 (38)
Normal (6–16 10^9/L)	137 (57.1)	27 (57.4)	164 (57)
High (>16 10^9/L)	13 (5.4)	1 (2.1)	14 (5)
Lymphocyte			
Low (<3.5 10^9/L)	199 (83)	43 (91.5)	242 (84)
Normal (3.5–11 10^9/L)	39 (16)	4(8.5)	43 (15)
High (>11 10^9/L)	2 (1)	0 (0)	2 (1)
Neutrophil			
Low (< 1 10^9/L)	12 (5)	2 (4)	14 (5)
Normal (1–7 10^9/L)	167 (70)	35 (75)	200 (70)
High (>7 10^9/L)	61 (25)	10 (21)	73 (25)
Platelets			
Low (<200 10^9/L)	106 (44.2)	13 (28)	119 (41.5)
Medium (200–550 10^9/L)	133 (55.4)	32 (72)	167 (58)
High (>550 10^9/L)	1 (0.4)	0 (0)	1 (0.3)

Results.

Table 1: cohort characteristics



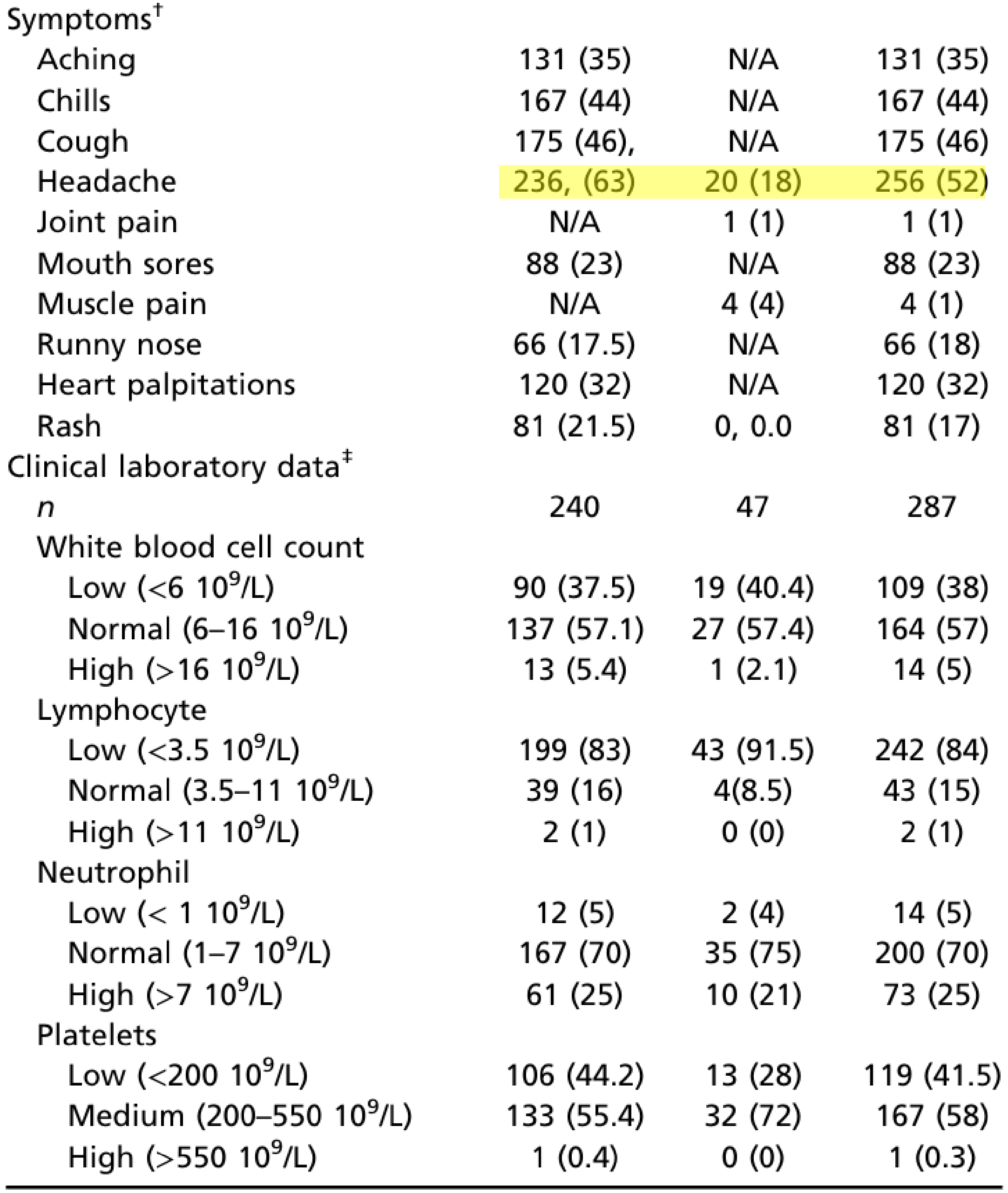
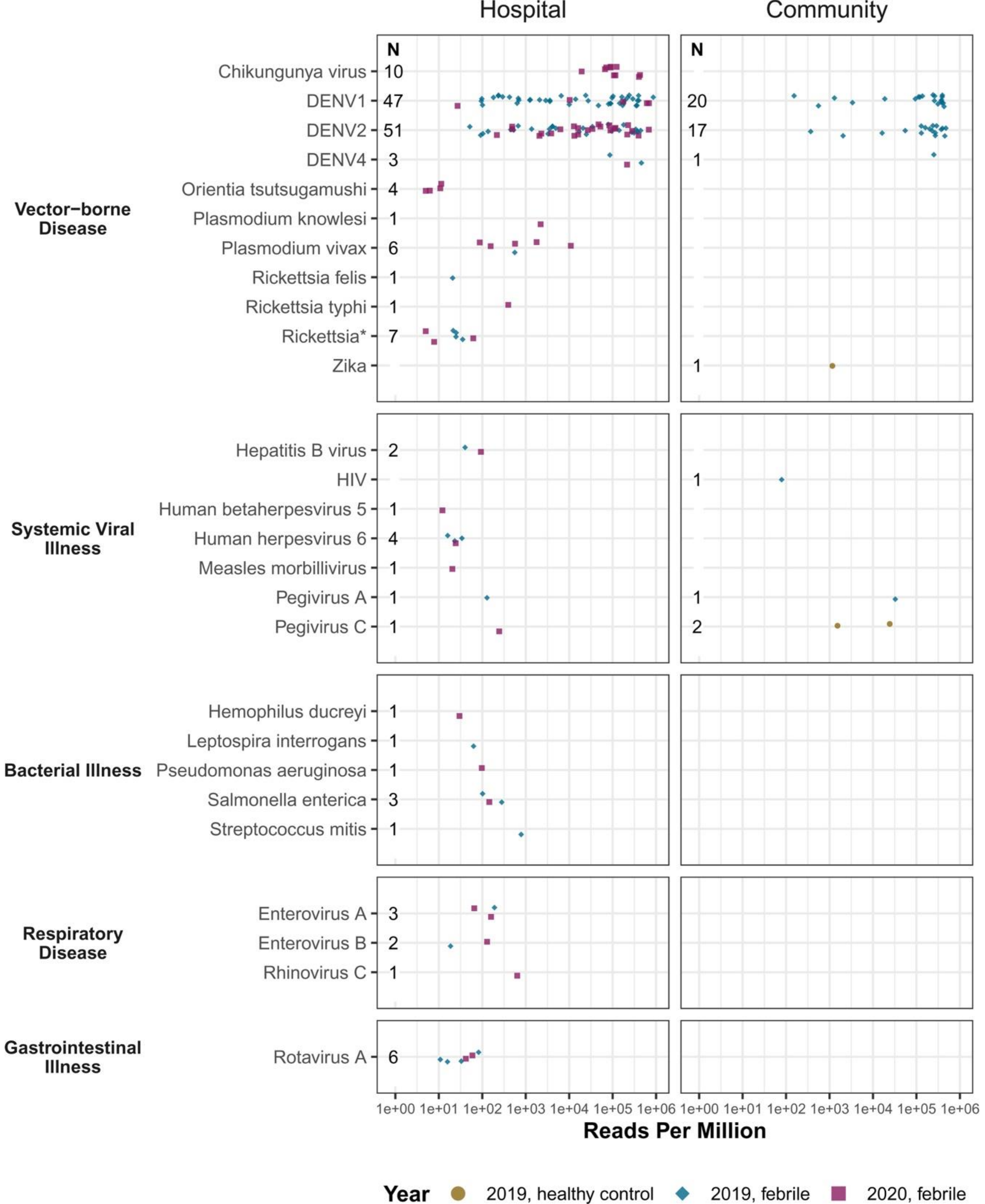


Table 1. Baseline demographic and clinical characteristics

Characteristic	Hospital	Community	Total
n	377	110	487
Male	207 (55)	56 (51)	263 (54)
Age, y (median, IQR)	10, 12	6, 4	8, 10
Year of fever			
2019	196 (52)	110 (100)	306 (63)
Attends school	146 (39)	64 (58)	210 (43)
Attends work	75 (20)	0 (0)	75 (15)
Socioeconomic status			
Very poor	16 (4)	0, 0.0	16 (3)
Lower	178 (47)	22 (20)	200 (41)
Middle	181 (48)	88 (80)	269 (55)
Upper	1 (0.3)	0 (0)	1 (0.2)
Risk factors			
Coil use	22 (60)	70 (64)	295 (61)
Insecticide use	191 (51)	60 (54.5)	251 (52)
Larvicide use	28 (7)	27 (24.5)	55 (11)
Insecticide-treated bed net use	313 (83)	99 (90)	412 (85)
Self-reported animal contact	275 (73)	N/A	275 (73)
Self-reported insect contact*	211 (56)	N/A	211 (56)
Symptoms†			
Aching	131 (35)	N/A	131 (35)
Chills	167 (44)	N/A	167 (44)
Cough	175 (46),	N/A	175 (46)
Headache	236, (63)	20 (18)	256 (52)
Joint pain	N/A	1 (1)	1 (1)
Mouth sores	88 (23)	N/A	88 (23)
Muscle pain	N/A	4 (4)	4 (1)
Runny nose	66 (17.5)	N/A	66 (18)
Heart palpitations	120 (32)	N/A	120 (32)
Rash	81 (21.5)	0, 0.0	81 (17)
Clinical laboratory data‡			
n	240	47	287
White blood cell count			
Low (<6 10^9/L)	90 (37.5)	19 (40.4)	109 (38)
Normal (6–16 10^9/L)	137 (57.1)	27 (57.4)	164 (57)
High (>16 10^9/L)	13 (5.4)	1 (2.1)	14 (5)
Lymphocyte			
Low (<3.5 10^9/L)	199 (83)	43 (91.5)	242 (84)
Normal (3.5–11 10^9/L)	39 (16)	4(8.5)	43 (15)
High (>11 10^9/L)	2 (1)	0 (0)	2 (1)
Neutrophil			
Low (< 1 10^9/L)	12 (5)	2 (4)	14 (5)
Normal (1–7 10^9/L)	167 (70)	35 (75)	200 (70)
High (>7 10^9/L)	61 (25)	10 (21)	73 (25)
Platelets			
Low (<200 10^9/L)	106 (44.2)	13 (28)	119 (41.5)
Medium (200–550 10^9/L)	133 (55.4)	32 (72)	167 (58)
High (>550 10^9/L)	1 (0.4)	0 (0)	1 (0.3)

Results.

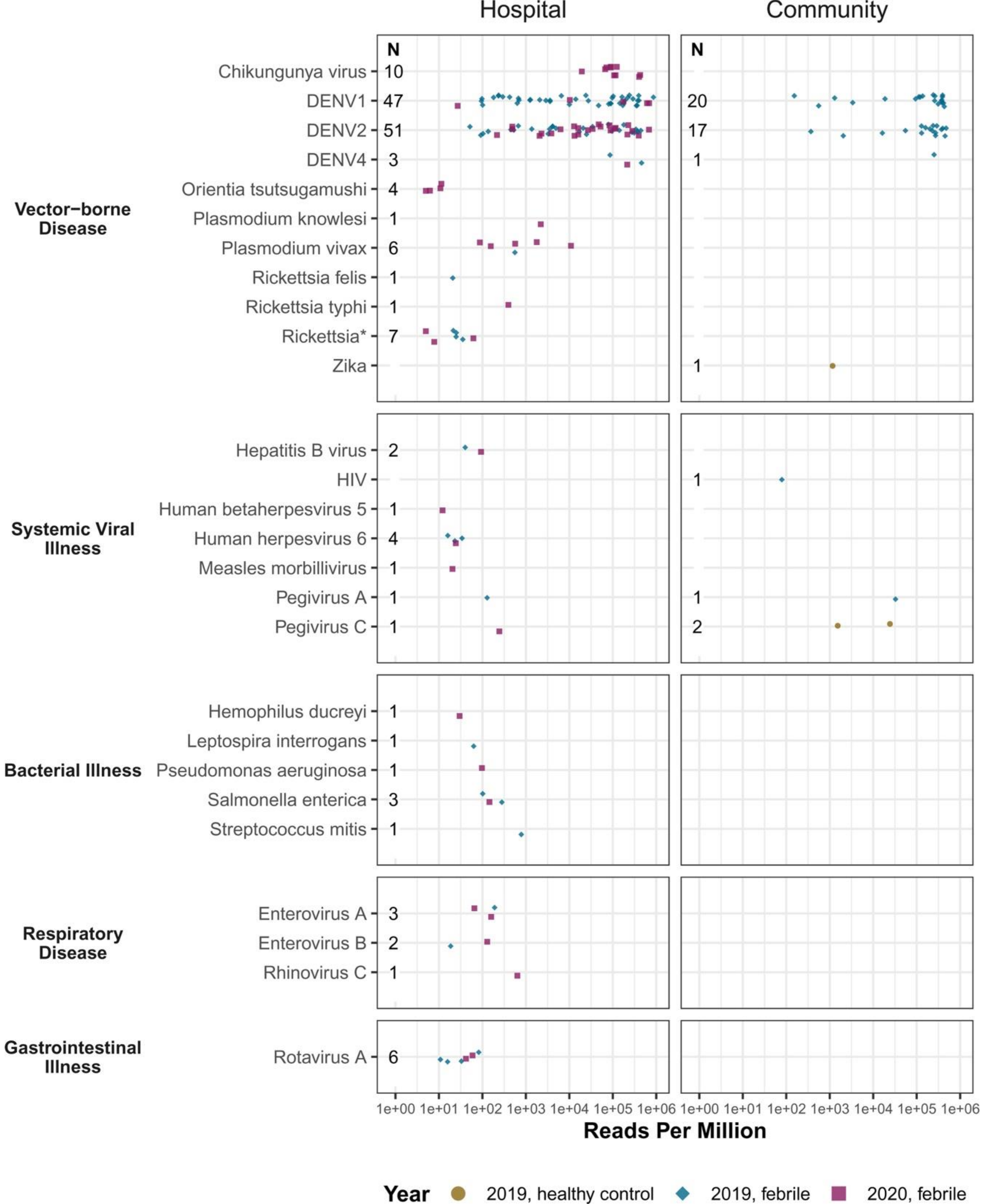
Fig. 2: pathogens identified



Results.

Fig. 2: pathogens identified

- 203 pathogens detected in 489 sera samples (41.5%)
- 7 participants coinfected (1.4%)
- ‘vector-borne disease’, then ‘systemic viral disease’ were the most prevalent clinical categories



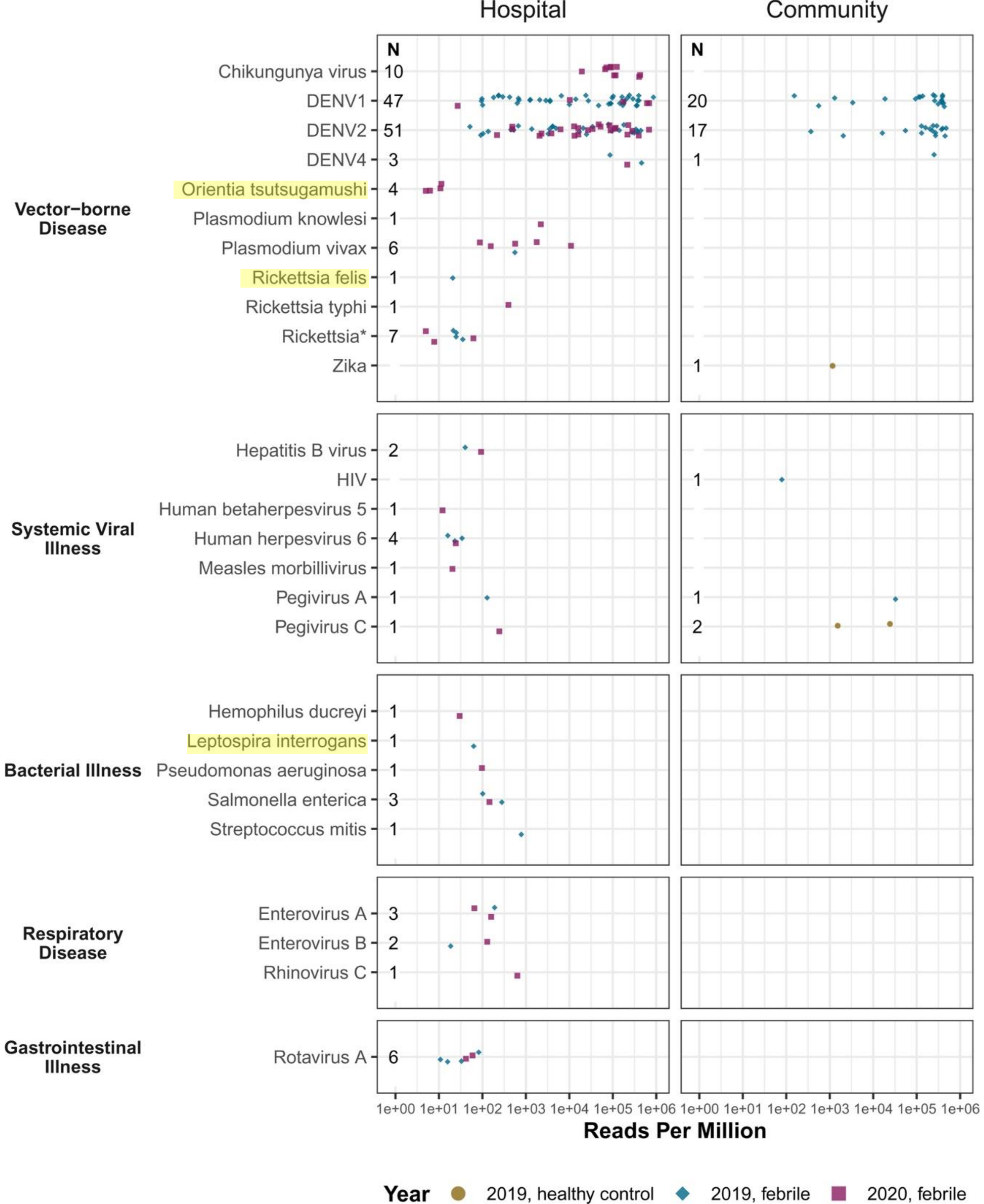
Results.

Fig. 2: pathogens identified

- 203 pathogens detected in 489 sera samples (41.5%)
- 7 participants coinfected (1.4%)
- ‘vector-borne disease’, then ‘systemic viral disease’ were the most prevalent clinical categories

Bacteria

- ***Rickettsia*****:** 4 *Orientia tsutsugamushi*+ cases and 1 *Rickettsia felis*+ case
- ***Leptospira***: 1 *L. interrogans*+ case



Results.

Fig. 2: pathogens identified

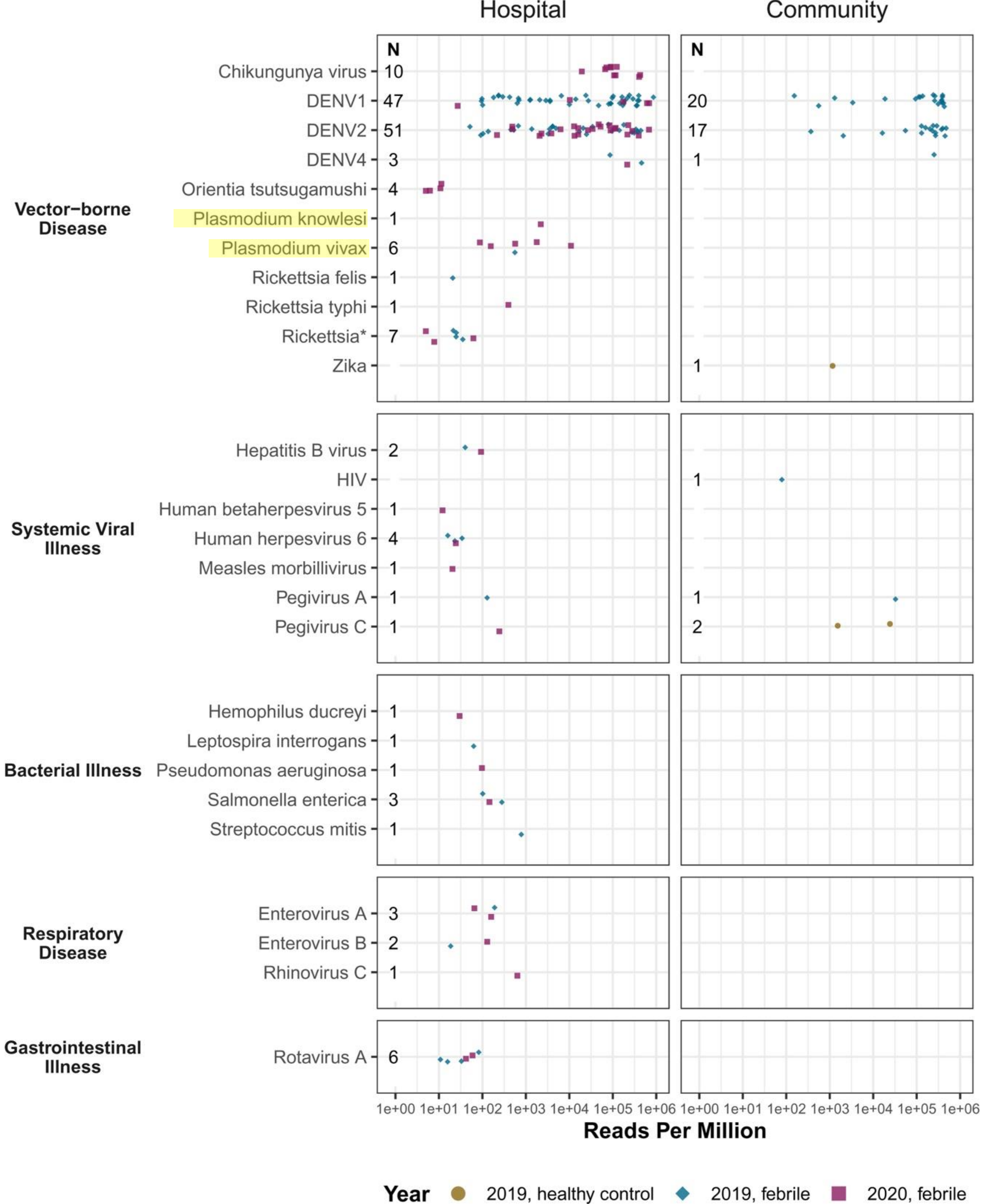
- 203 pathogens detected in 489 sera samples (41.5%)
- 7 participants coinfected (1.4%)
- ‘vector-borne disease’, then ‘systemic viral disease’ were the most prevalent clinical categories

Bacteria

- ***Rickettsia*****:** 4 *Orientia tsutsugamushi*+ cases and 1 *Rickettsia felis*+ case
- ***Leptospira***: 1 *L. interrogans*+ case

Protozoa

- **Malaria:** 6 low parasitemia *Plasmodium vivax*+ cases and 1 *P. knowlesi*+ case



Results.

Fig. 2: pathogens identified

- 203 pathogens detected in 489 sera samples (41.5%)
- 7 participants coinfected (1.4%)
- ‘vector-borne disease’, then ‘systemic viral disease’ were the most prevalent clinical categories

Bacteria

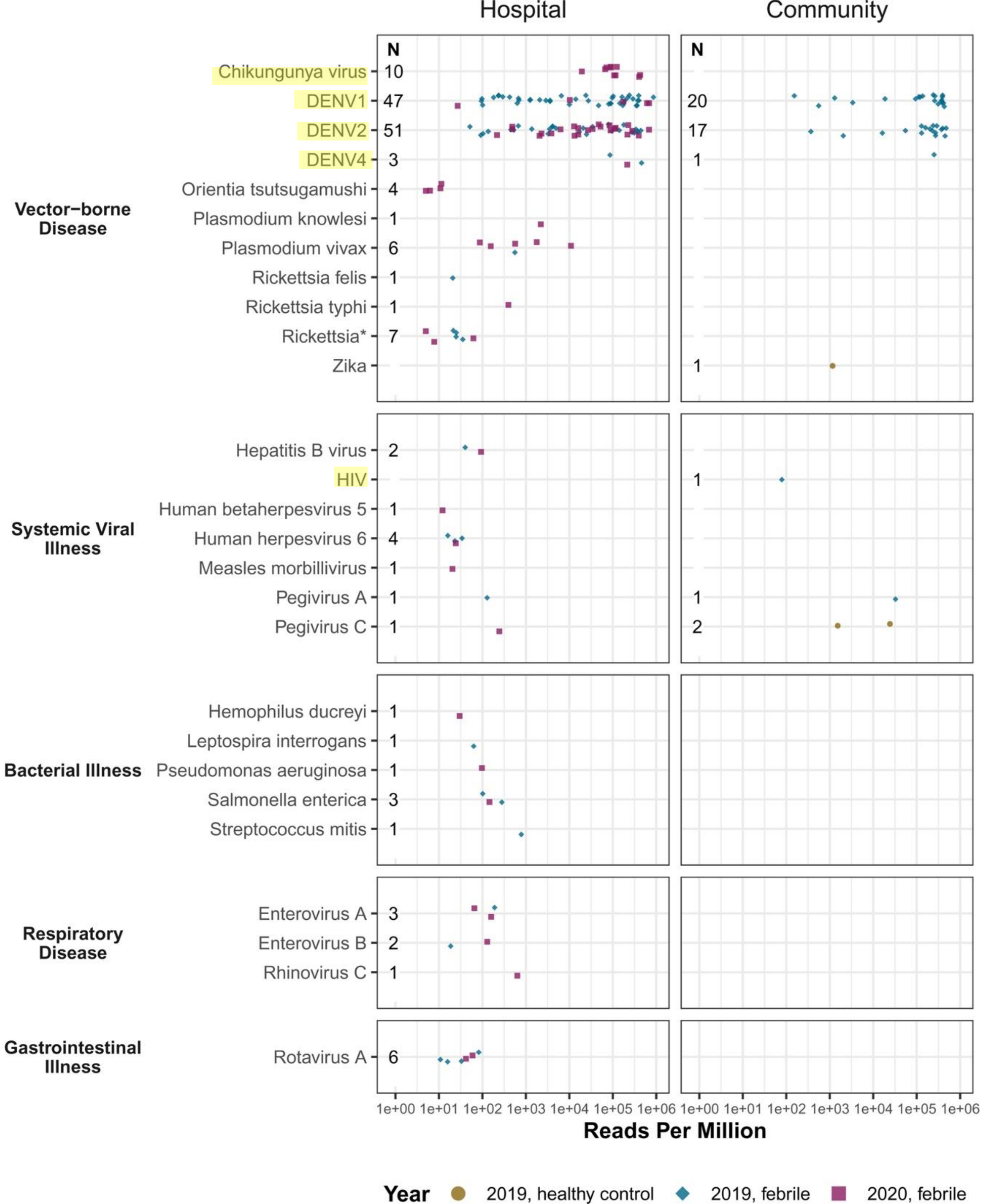
- ***Rickettsia***: 4 *Orientia tsutsugamushi*+ cases and 1 *Rickettsia felis*+ case
- ***Leptospira***: 1 *L. interrogans*+ case

Protozoa

- **Malaria:** 6 low parasitemia *Plasmodium vivax*+ cases and 1 *P. knowlesi*+ case

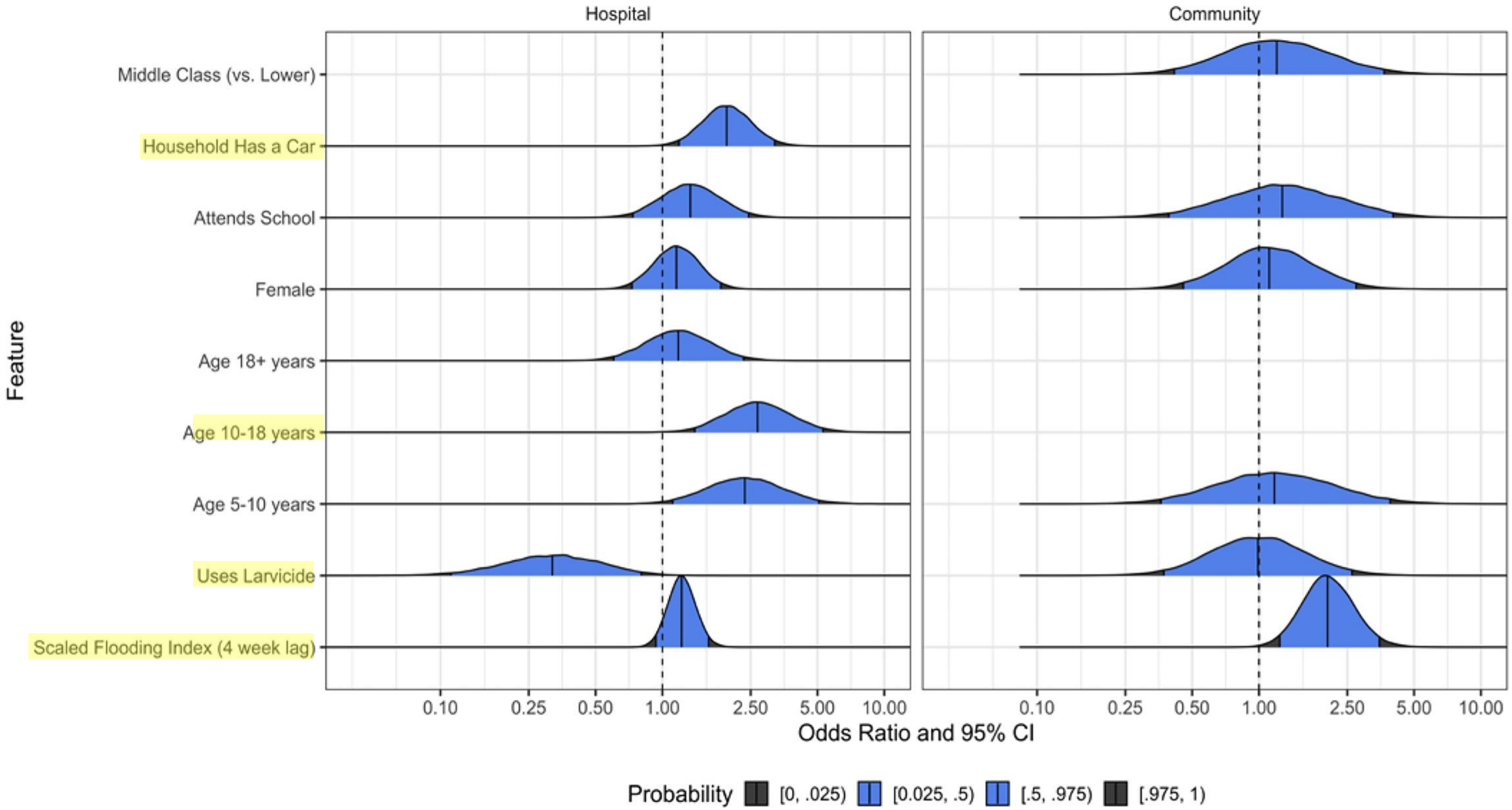
Viruses

- 138 DENV+ cases
- 10 CHIKV+ cases → added to routine PCR testing
- 1 ZIKV+ case
- ***HIV:*** 1 HIV-DENV2 coninfection → linked to ART



Results.

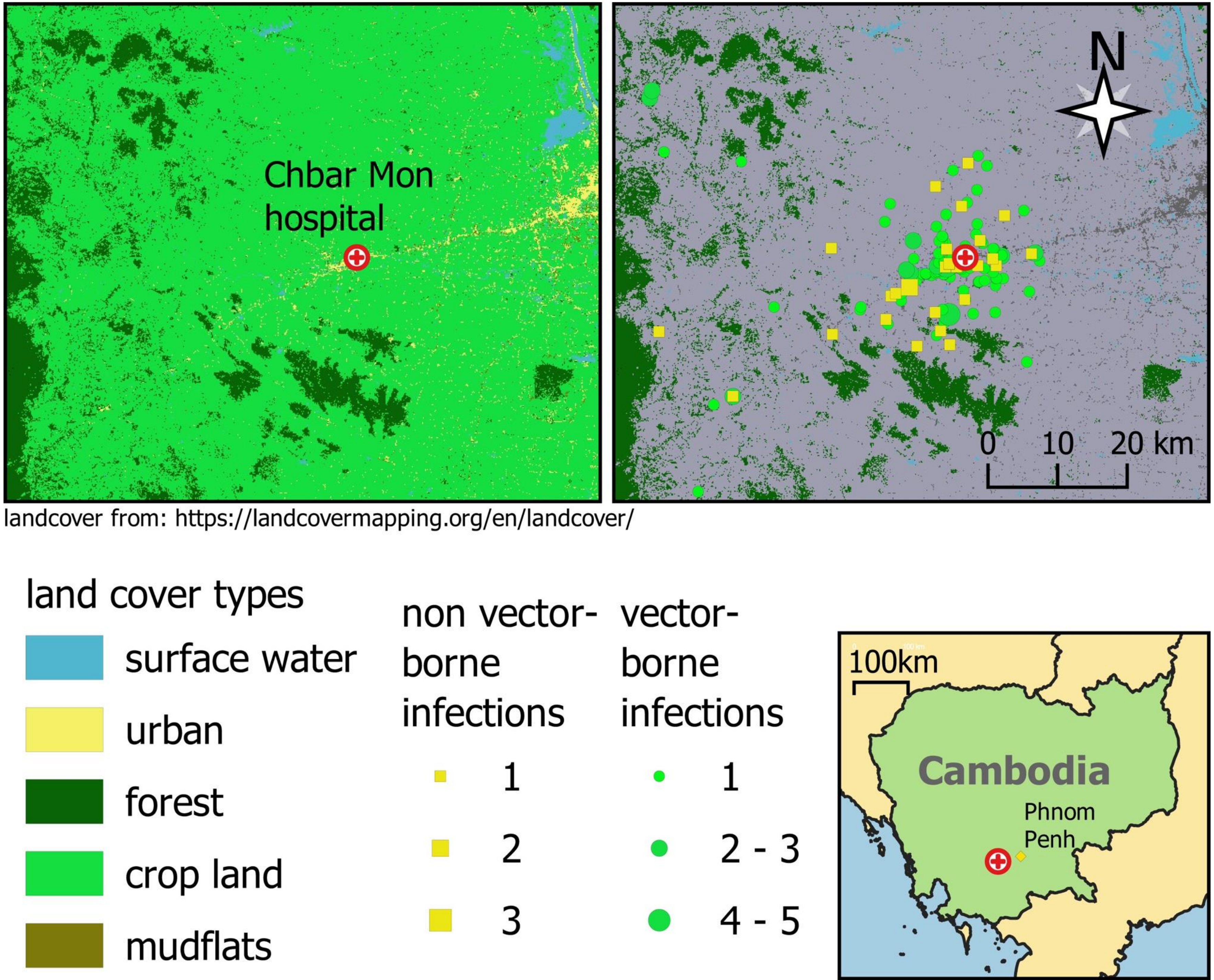
Fig. 3: correlates of VBD



Results.

Fig. 4: geospatial setting

- Crop land = predominant land-cover type (89%)
- Urban = 2nd most (10%)
- Urban participants more likely to have non-VBD (13%) vs. VBD (9%)
- Still urban cases of CHIK, DENV1, DENV2, ZIKV
- 92% DENV cases from crop land



landcover from: https://landcovermapping.org/en/landcover/