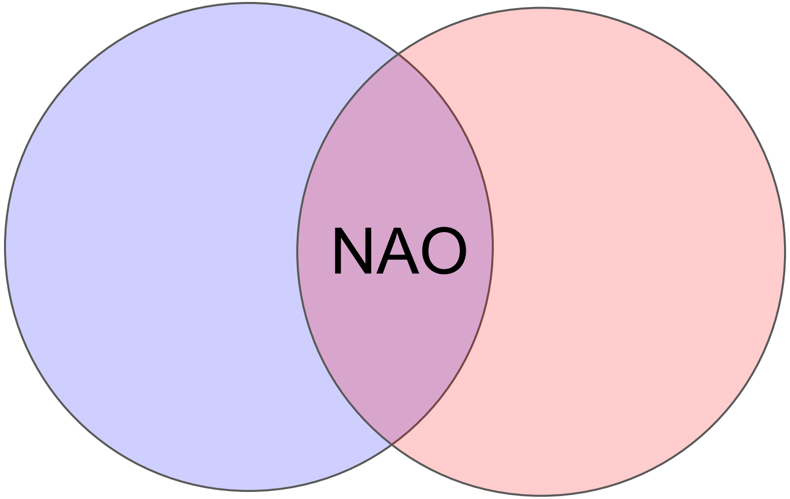
Predicting Relative Abundance

Jeff Kaufman

2023-10-18

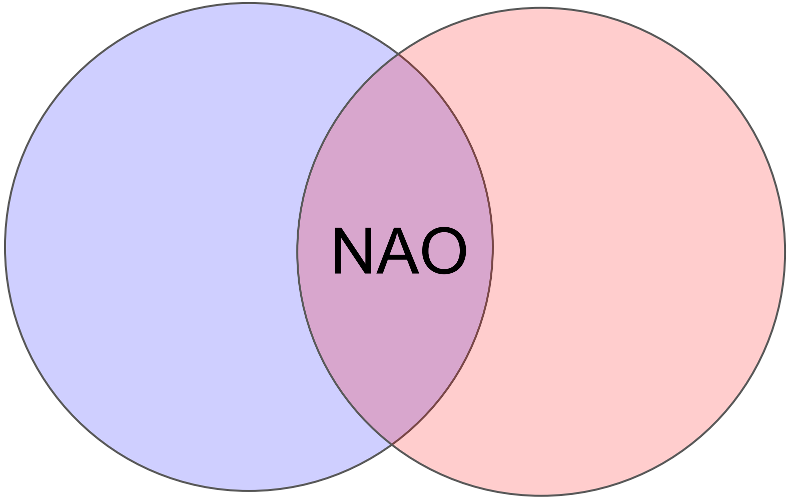
Wastewater Biosurveillance Workshop

Background



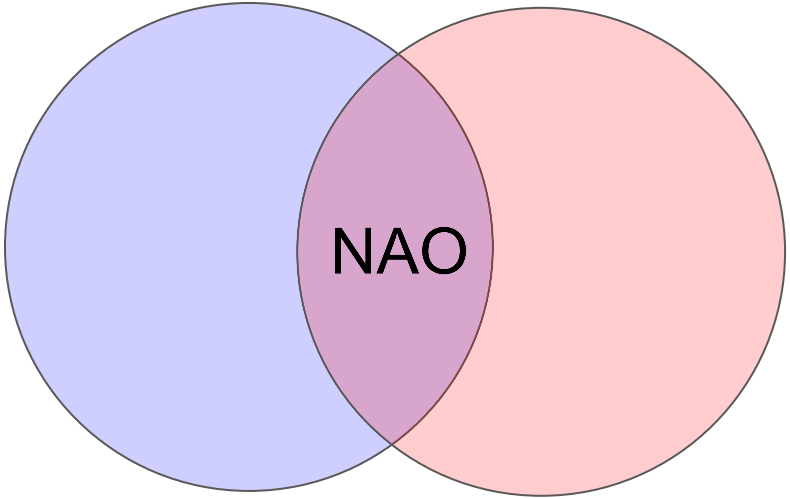
Background

- I'm Jeff Kaufman



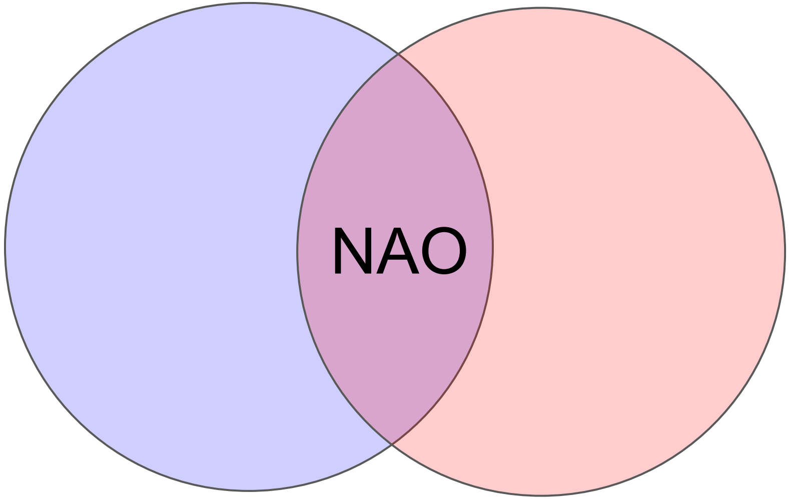
Background

- I'm Jeff Kaufman
- From the Nucleic Acid Observatory



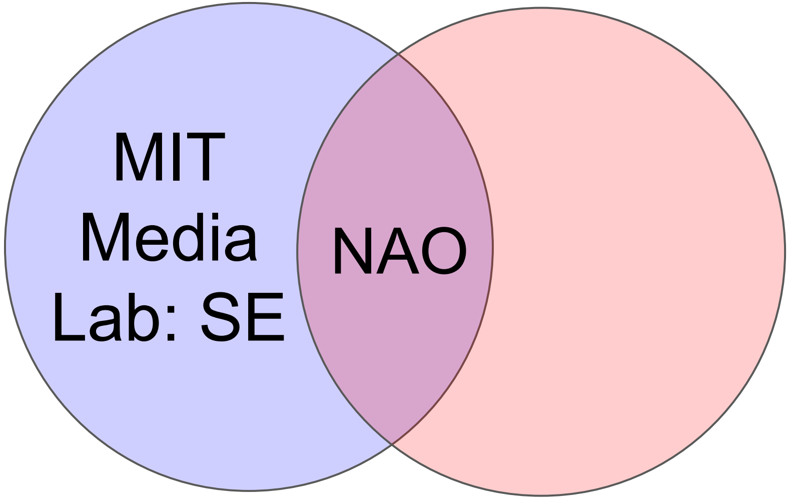
Background

- I'm Jeff Kaufman
- From the Nucleic Acid Observatory
- Collaborative Project



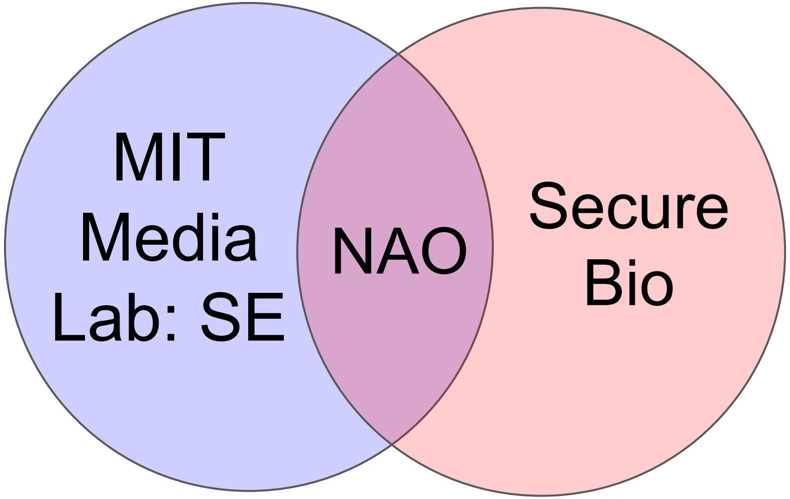
Background

- I'm Jeff Kaufman
- From the Nucleic Acid Observatory
- Collaborative Project
  - MIT Media Lab: Sculpting Evolution (Esvelt Lab)



Background

- I'm Jeff Kaufman
- From the Nucleic Acid Observatory
- Collaborative Project
  - MIT Media Lab: Sculpting Evolution (Esvelt Lab)
  - SecureBio



Background

Background

- Presenting work from a team

Background

- Presenting work from a team

Mike McLaren

Simon Grimm

Dan Rice

Jeff Kaufman

Background

Background

- Summary of an NAO report

Background

- Summary of an NAO report
    - 2023-08-10, "Predicting Virus Relative Abundance in Wastewater"

Background

- Summary of an NAO report
  - 2023-08-10, "Predicting Virus Relative Abundance in Wastewater"
  - `data.securebio.org/p2ra`

Background

Background

- NAO Goal: detect pandemics

Background

- NAO Goal: detect pandemics
    - Even if they don't look like previous ones

Background

- NAO Goal: detect pandemics
  - Even if they don't look like previous ones
  - Even one that's intentionally "stealth"

Background

- NAO Goal: detect pandemics
  - Even if they don't look like previous ones
  - Even one that's intentionally "stealth"
- Wastewater

Background

- NAO Goal: detect pandemics
  - Even if they don't look like previous ones
  - Even one that's intentionally "stealth"
- Wastewater
  - Millions of people → one sample

Background

- NAO Goal: detect pandemics
  - Even if they don't look like previous ones
  - Even one that's intentionally "stealth"
- Wastewater
  - Millions of people → one sample
- Metagenomic sequencing

Background

- NAO Goal: detect pandemics
  - Even if they don't look like previous ones
  - Even one that's intentionally "stealth"
- Wastewater
  - Millions of people → one sample
- Metagenomic sequencing
  - Doesn't require pre-selecting pathogens

Background

- NAO Goal: detect pandemics
  - Even if they don't look like previous ones
  - Even one that's intentionally "stealth"
- Wastewater
  - Millions of people → one sample
- Metagenomic sequencing
  - Doesn't require pre-selecting pathogens
  - But most reads won't match a pathogen

Key question

Key question

- At a given stage of a future viral pandemic, what fraction of wastewater metagenomic sequencing reads match the virus?

Key question

- At a given stage of a future viral pandemic, what fraction of wastewater metagenomic sequencing reads match the virus?
- $\mathrm{RA}(1\%)$: relative abundance of virus, when:

Key question

- At a given stage of a future viral pandemic, what fraction of wastewater metagenomic sequencing reads match the virus?
- $\mathrm{RA}(1\%)$: relative abundance of virus, when:
    - 1% currently infected (prevalence)

Key question

- At a given stage of a future viral pandemic, what fraction of wastewater metagenomic sequencing reads match the virus?
- $\mathrm{RA}(1\%)$: relative abundance of virus, when:
  - 1% currently infected (prevalence), or
  - 1% became infected this week (incidence)

Key question

Key question

- Knowing $\mathrm{RA}(1\%)$ for many viruses would help us estimate:

Key question

- Knowing $\mathrm{RA}(1\%)$ for many viruses would help us estimate:
  - How deep to sequence?

Key question

- Knowing $\mathrm{RA}(1\%)$ for many viruses would help us estimate:
  - How deep to sequence?
  - What would it cost?

Approach

Approach

- Link public health data to sequencing data

Approach

- Link public health data to sequencing data
  - Collect metagenomic wastewater data

Approach

- Link public health data to sequencing data
  - Collect metagenomic wastewater data
    - Data from published studies (via SRA)

Approach

- Link public health data to sequencing data
    - Collect metagenomic wastewater data
        - Data from published studies (via SRA)
    - Process into per-virus relative abundances

Approach

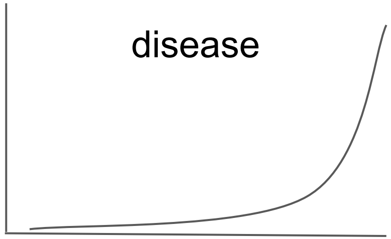
- Link public health data to sequencing data
  - Collect metagenomic wastewater data
    - Data from published studies (via SRA)
  - Process into per-virus relative abundances
  - Select target viruses

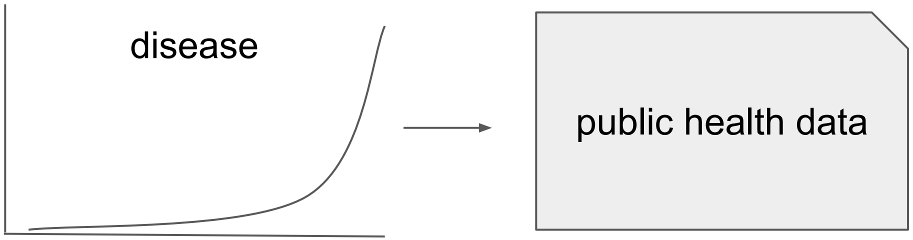
Approach

- Link public health data to sequencing data
    - Collect metagenomic wastewater data
        - Data from published studies (via SRA)
    - Process into per-virus relative abundances
    - Select target viruses
    - Collect public health estimates

Approach

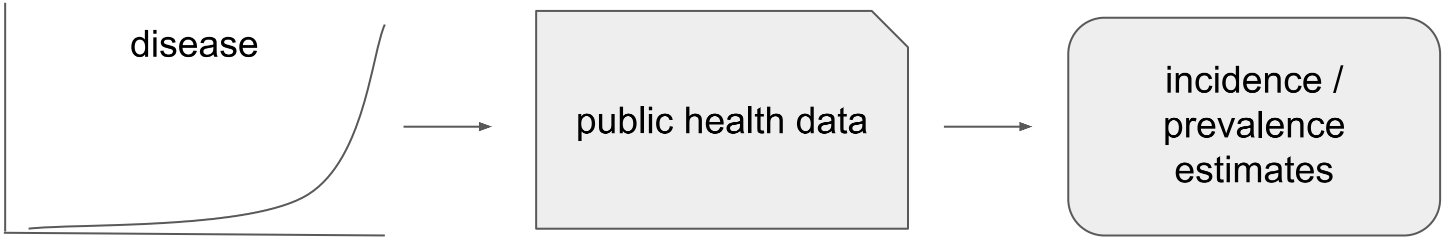
- Link public health data to sequencing data
  - Collect metagenomic wastewater data
    - Data from published studies (via SRA)
  - Process into per-virus relative abundances
  - Select target viruses
  - Collect public health estimates
  - Estimate $\mathrm{RA}(1\%)$


disease



disease

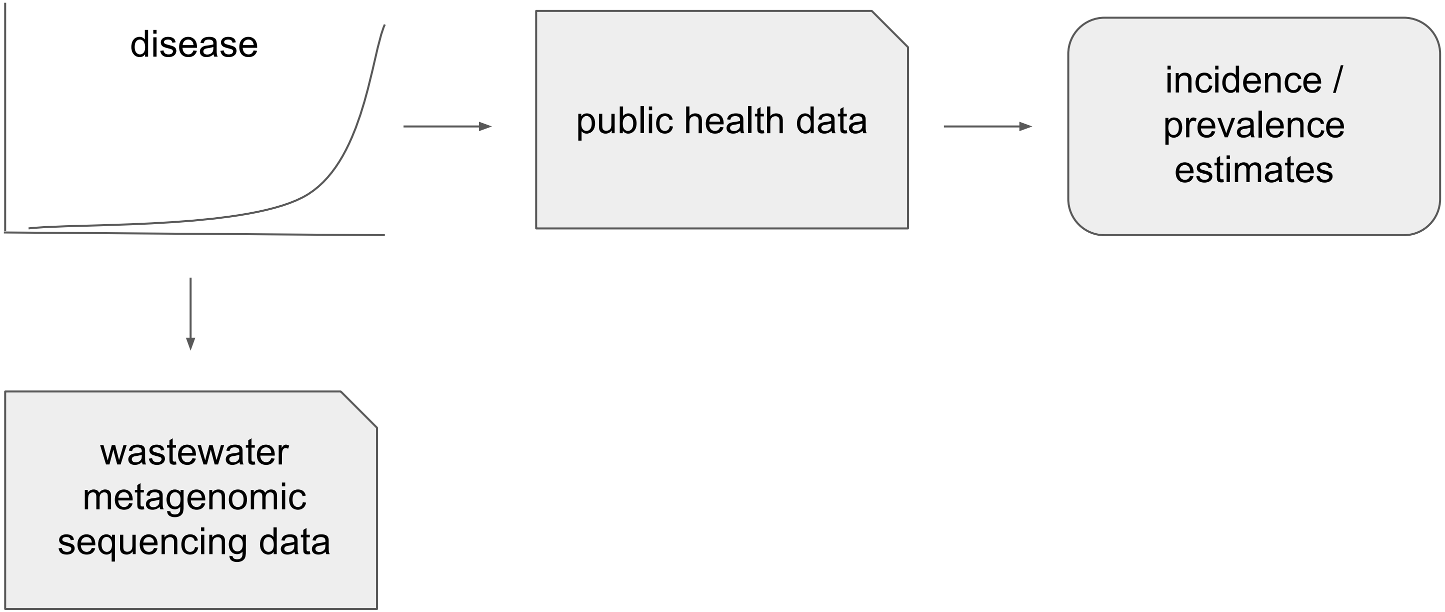
public health data



disease

public health data

incidence / prevalence estimates

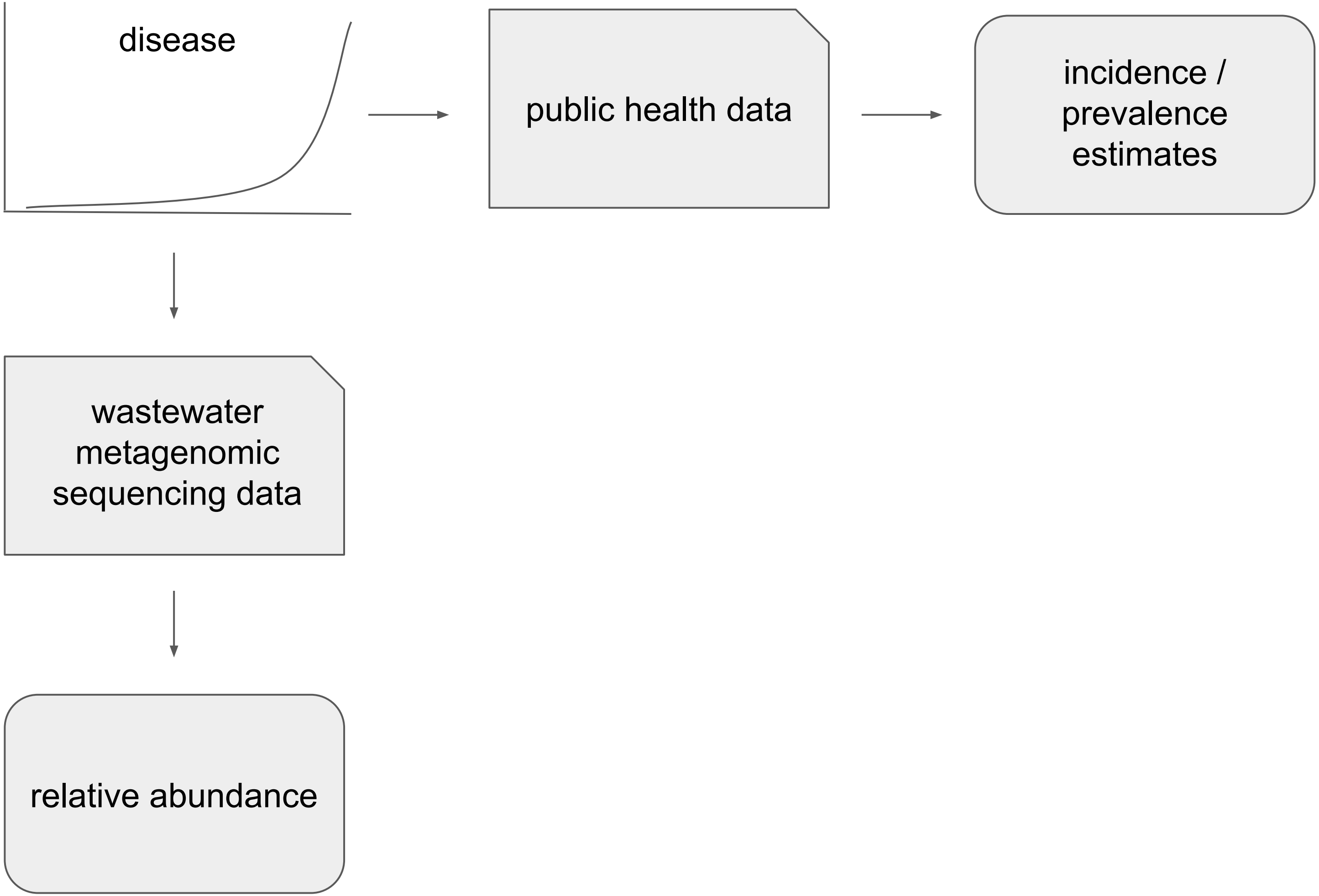


disease

public health data

incidence / prevalence estimates

wastewater metagenomic sequencing data



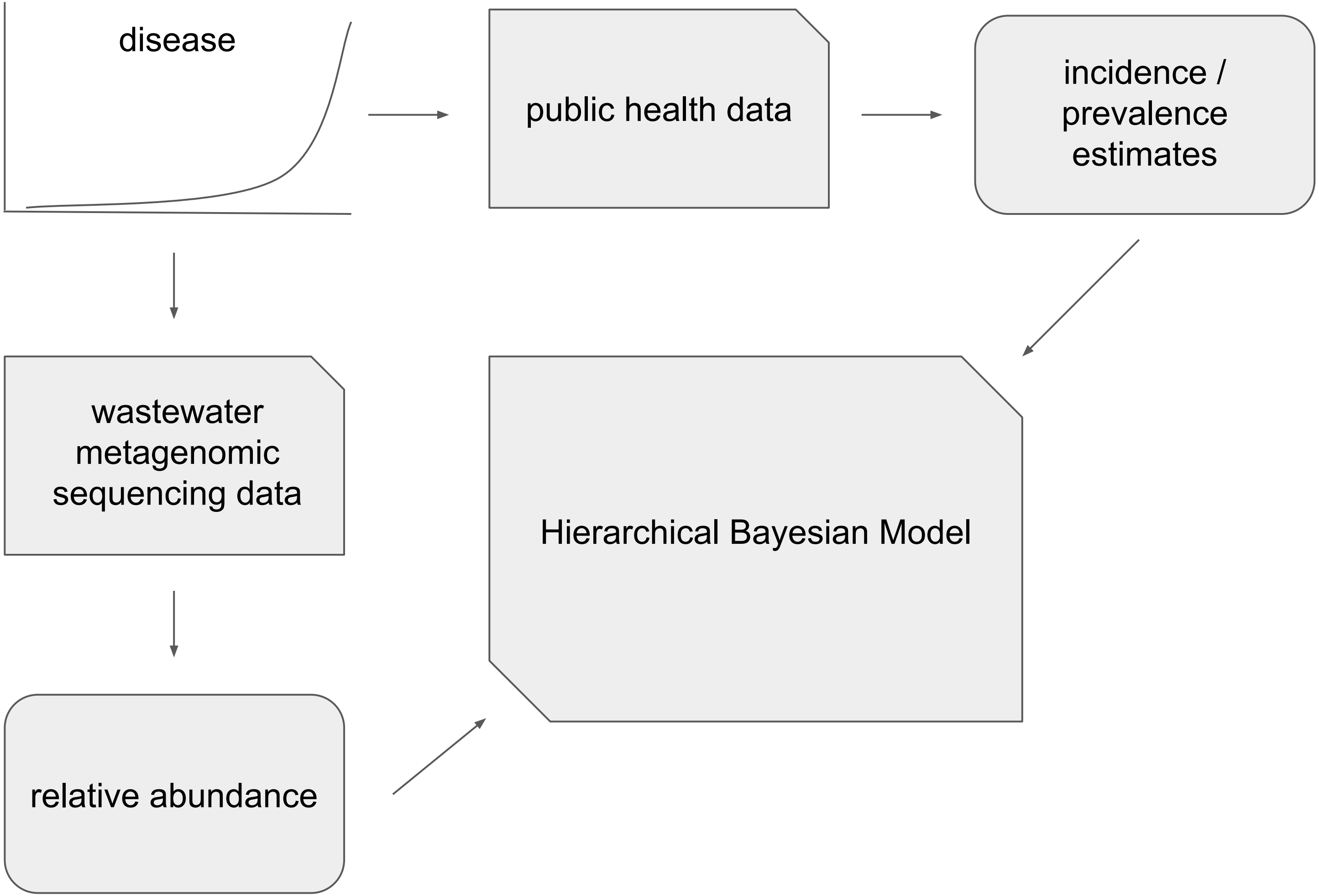
disease

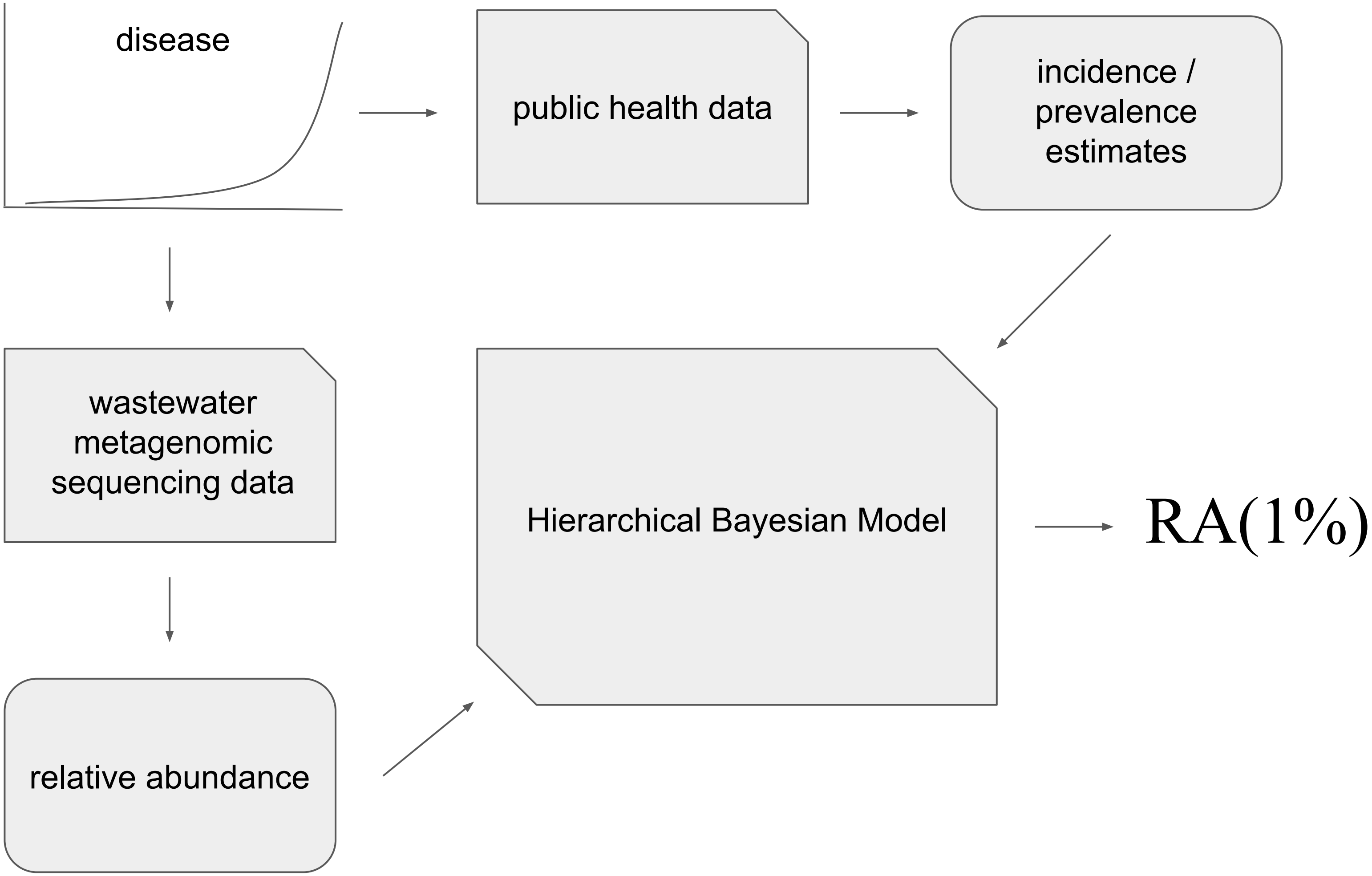
public health data

incidence / prevalence estimates

wastewater metagenomic sequencing data

relative abundance


disease
public health data
incidence / prevalence estimates
wastewater metagenomic sequencing data
Hierarchical Bayesian Model
relative abundance



disease

public health data

incidence / prevalence estimates

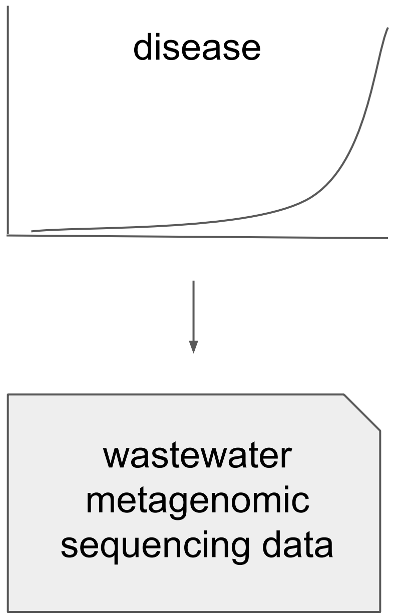
wastewater metagenomic sequencing data

relative abundance

Hierarchical Bayesian Model

RA(1%)

Sequencing Data

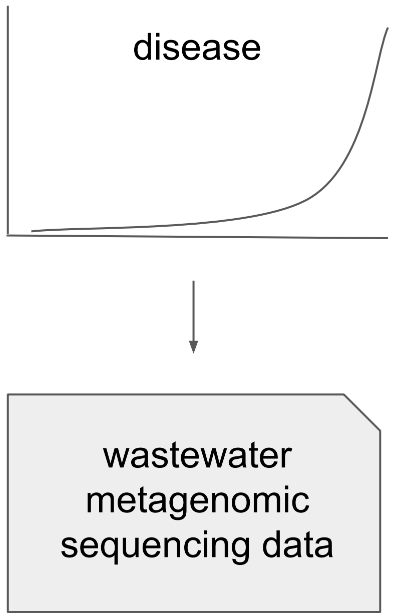


disease

wastewater metagenomic sequencing data

Sequencing Data

- RNA (2.8B read pairs)

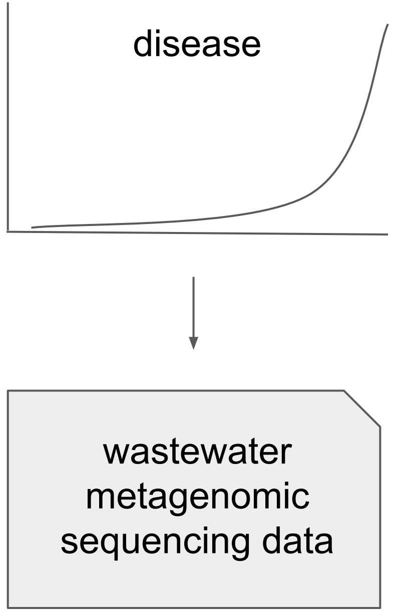


disease

wastewater metagenomic sequencing data

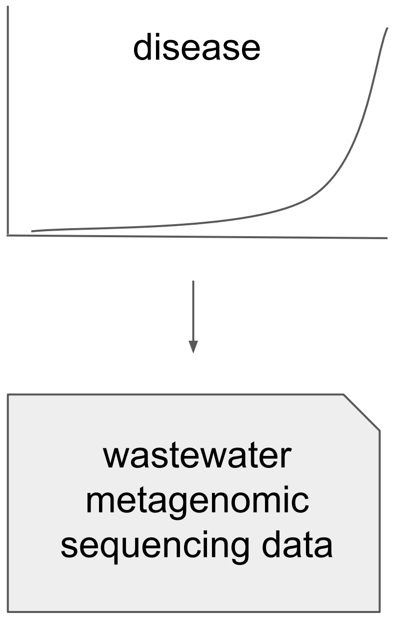
Sequencing Data

- RNA (2.8B read pairs)
  - SF: Crits-Christoph et al. (2021)



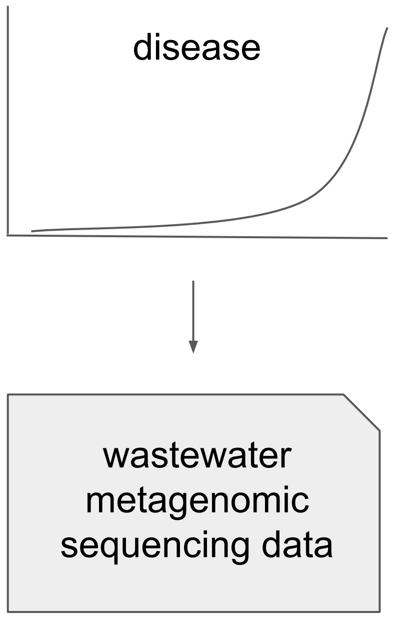
Sequencing Data

- RNA (2.8B read pairs)
  - SF: Crits-Christoph et al. (2021)
  - LA: Rothman et al. (2021)



Sequencing Data

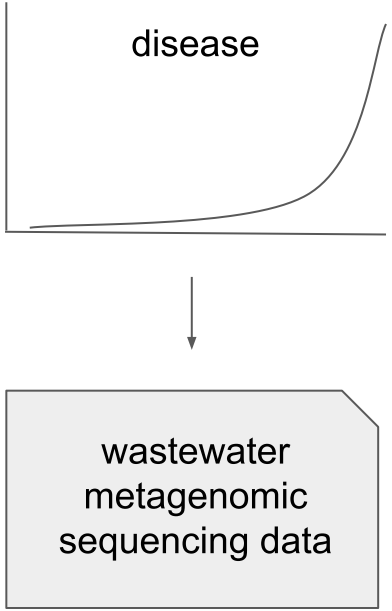
- RNA (2.8B read pairs)
  - SF: Crits-Christoph et al. (2021)
  - LA: Rothman et al. (2021)
  - Ohio: Spurbeck et al. (2023)



Sequencing Data

- RNA (2.8B read pairs)
    - SF: Crits-Christoph et al. (2021)
    - LA: Rothman et al. (2021)
    - Ohio: Spurbeck et al. (2023)

- DNA (4.4B read pairs)



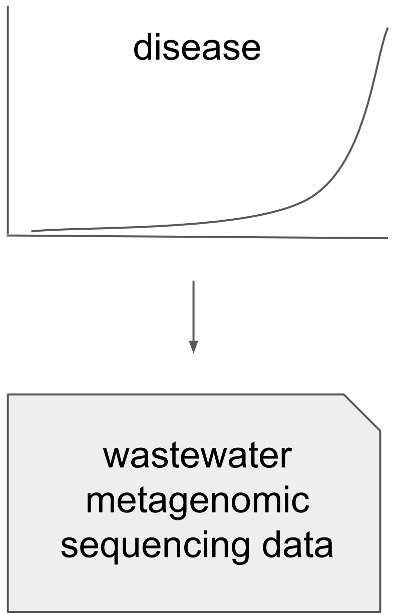
disease

wastewater metagenomic sequencing data

Sequencing Data

- RNA (2.8B read pairs)
  - SF: Crits-Christoph et al. (2021)
  - LA: Rothman et al. (2021)
  - Ohio: Spurbeck et al. (2023)

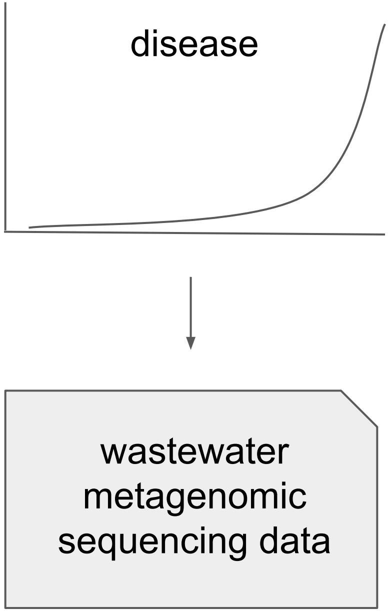
- DNA (4.4B read pairs)
  - Copenhagen: Brinch et al. (2020)



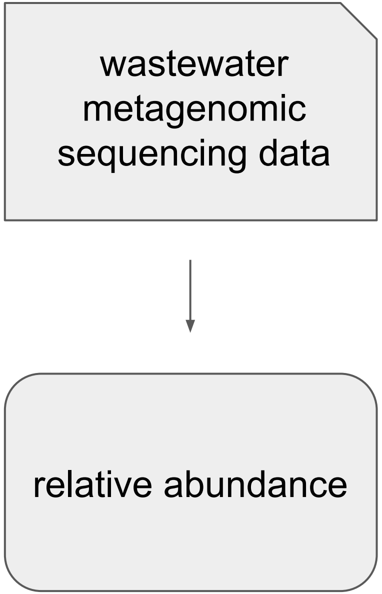
wastewater metagenomic sequencing data

Sequencing Data

- RNA (2.8B read pairs)
  - SF: Crits-Christoph et al. (2021)
  - LA: Rothman et al. (2021)
  - Ohio: Spurbeck et al. (2023)
  - *All during Covid-19*
- DNA (4.4B read pairs)
  - Copenhagen: Brinch et al. (2020)
  - *Pre-Covid-19*

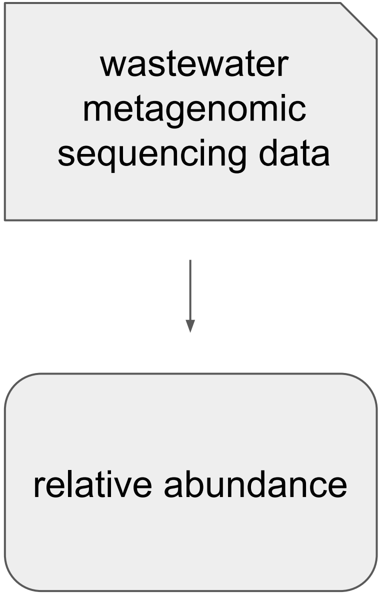


Determine Relative Abundances



Determine Relative Abundances

- Kraken2 to assign reads to species

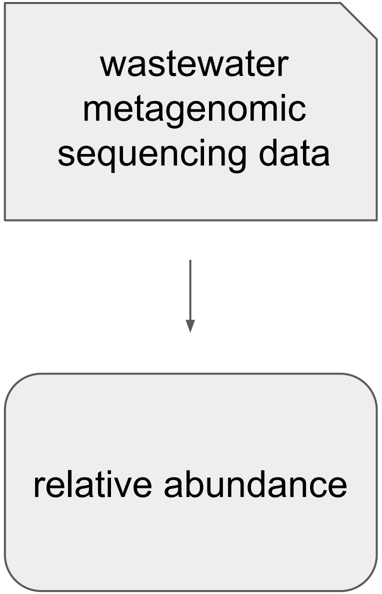


wastewater metagenomic sequencing data

relative abundance

Determine Relative Abundances

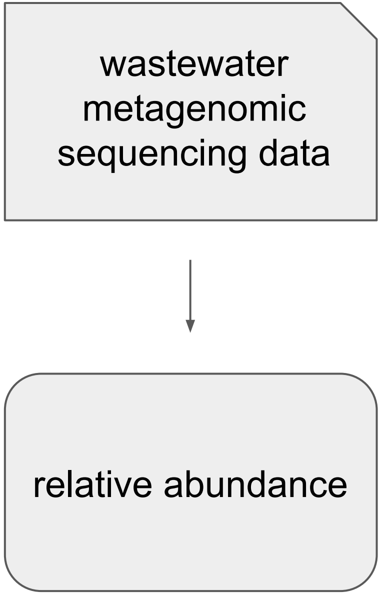
- Kraken2 to assign reads to species
- Alignment to reference genomes to remove false positives



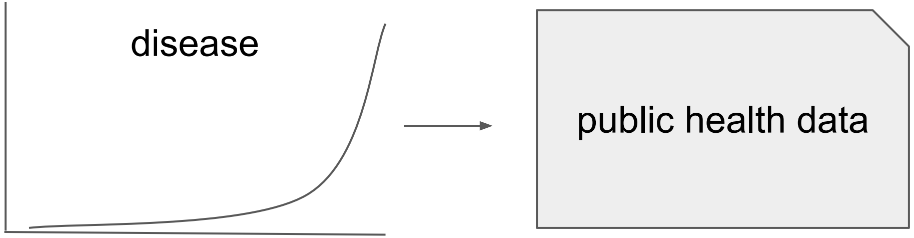
Determine Relative Abundances

- Kraken2 to assign reads to species
- Alignment to reference genomes to remove false positives

- relative abundance = $\frac{\text{reads matching virus}}{\text{reads in sample}}$



Target Viruses

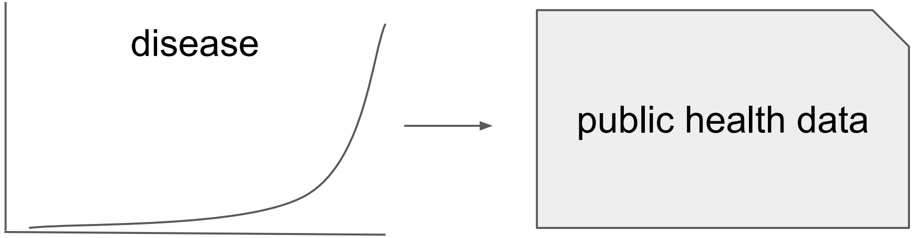


disease

public health data

Target Viruses

- Viruses where we can get public health estimates matching where and when sequencing samples were collected



disease

public health data

Target Viruses

Target Viruses

- Acute

Target Viruses

- Acute
  - Sars-CoV-2

Target Viruses

- Acute
  - Sars-CoV-2
  - Influenza A and B

Target Viruses

- Acute
  - Sars-CoV-2
  - Influenza A and B
  - Norovirus: genogroups I and II

Target Viruses

- Acute
  - Sars-CoV-2
  - Influenza A and B
  - Norovirus: genogroups I and II
- Chronic

Target Viruses

- Acute
  - Sars-CoV-2
  - Influenza A and B
  - Norovirus: genogroups I and II
- Chronic
  - HIV

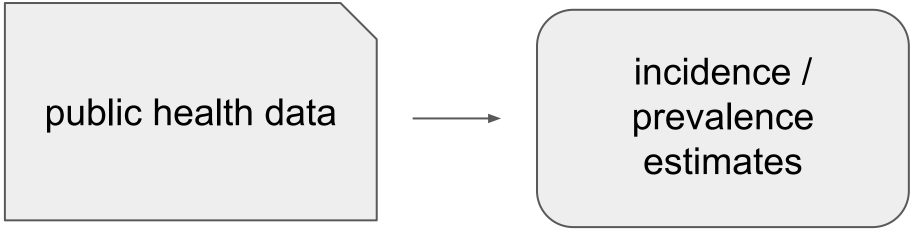
Target Viruses

- Acute
  - Sars-CoV-2
  - Influenza A and B
  - Norovirus: genogroups I and II
- Chronic
  - HIV
  - Herpes viruses: HSV-1, EBV, CMV

Target Viruses

- Acute
  - Sars-CoV-2
  - Influenza A and B
  - Norovirus: genogroups I and II
- Chronic
  - HIV
  - Herpes viruses: HSV-1, EBV, CMV
  - ... eight others

Public Health Data

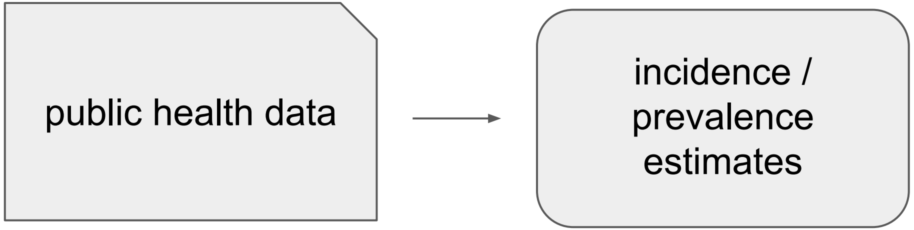


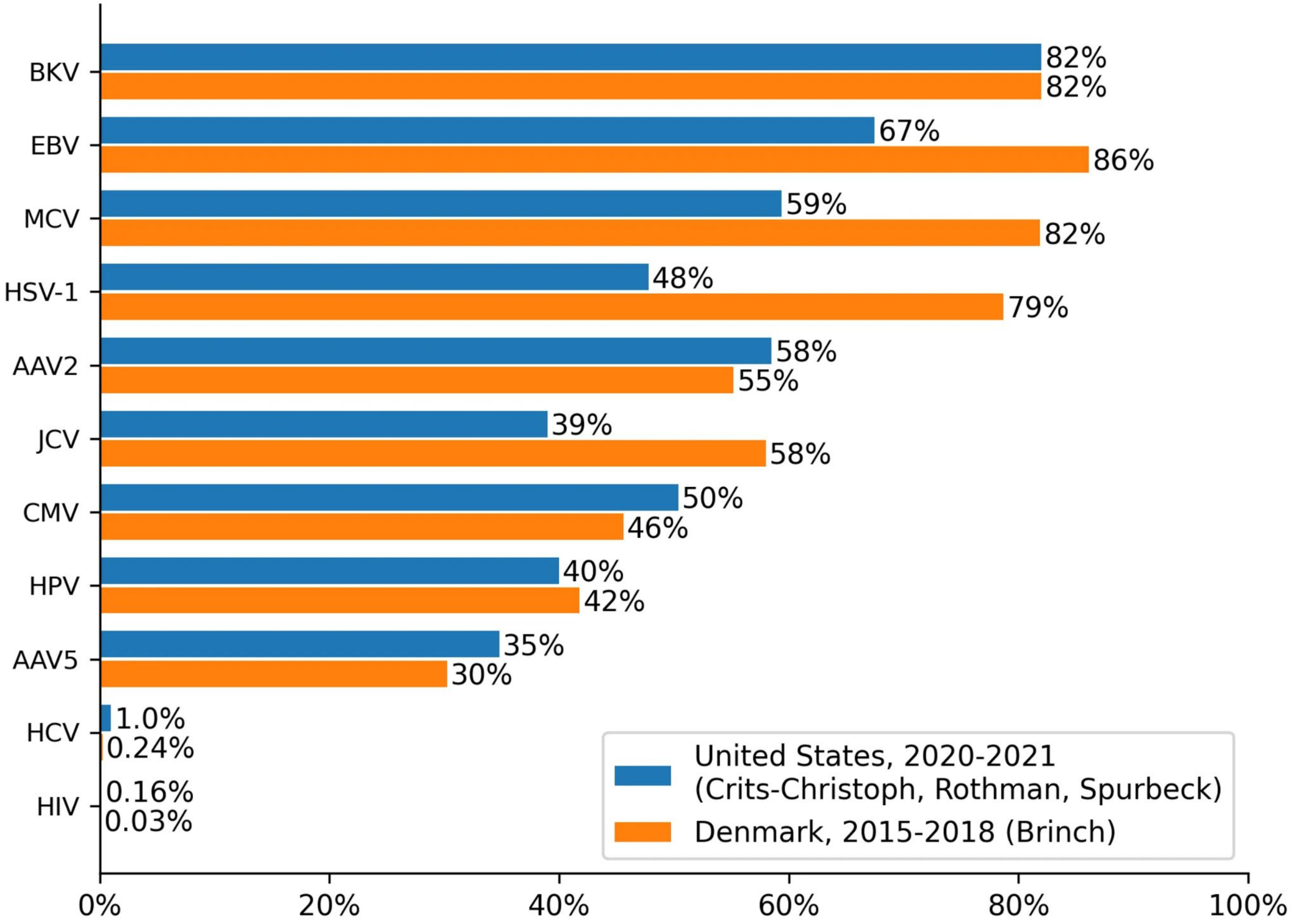
public health data

incidence / prevalence estimates

Public Health Data

- Chronic: estimate prevalence

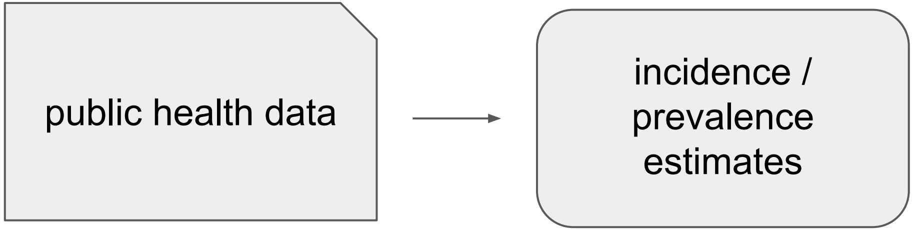




Estimated prevalence of persistent viral infections

Public Health Data

- Chronic: estimate prevalence

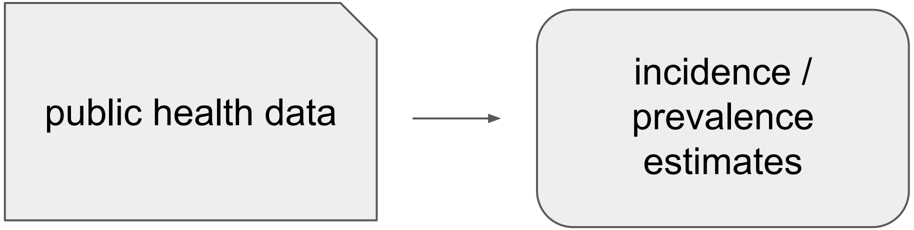


public health data

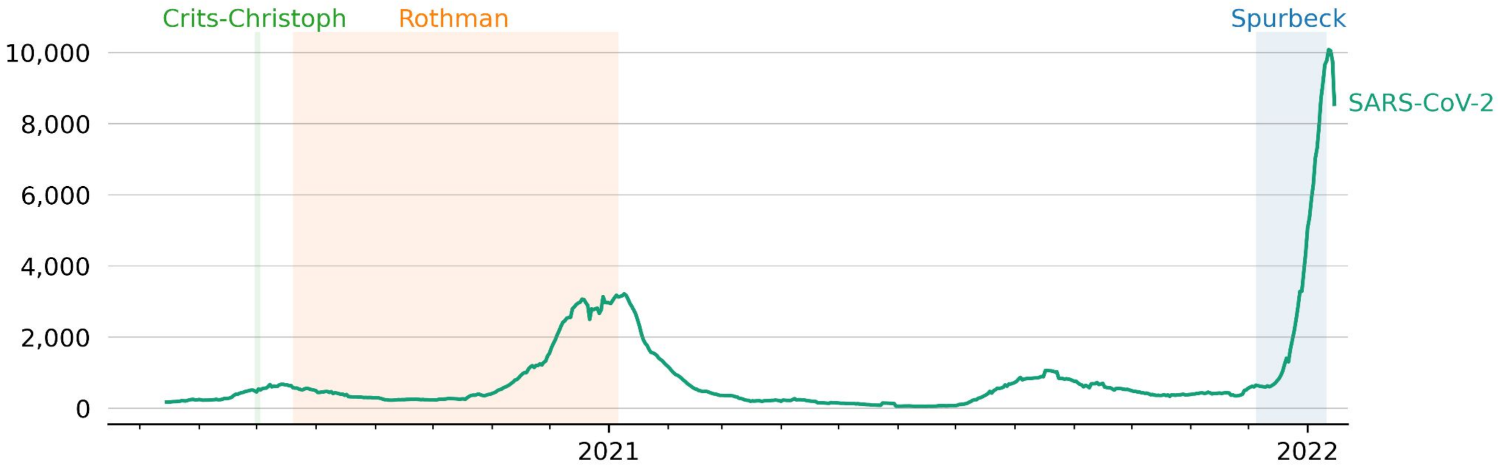
incidence / prevalence estimates

Public Health Data

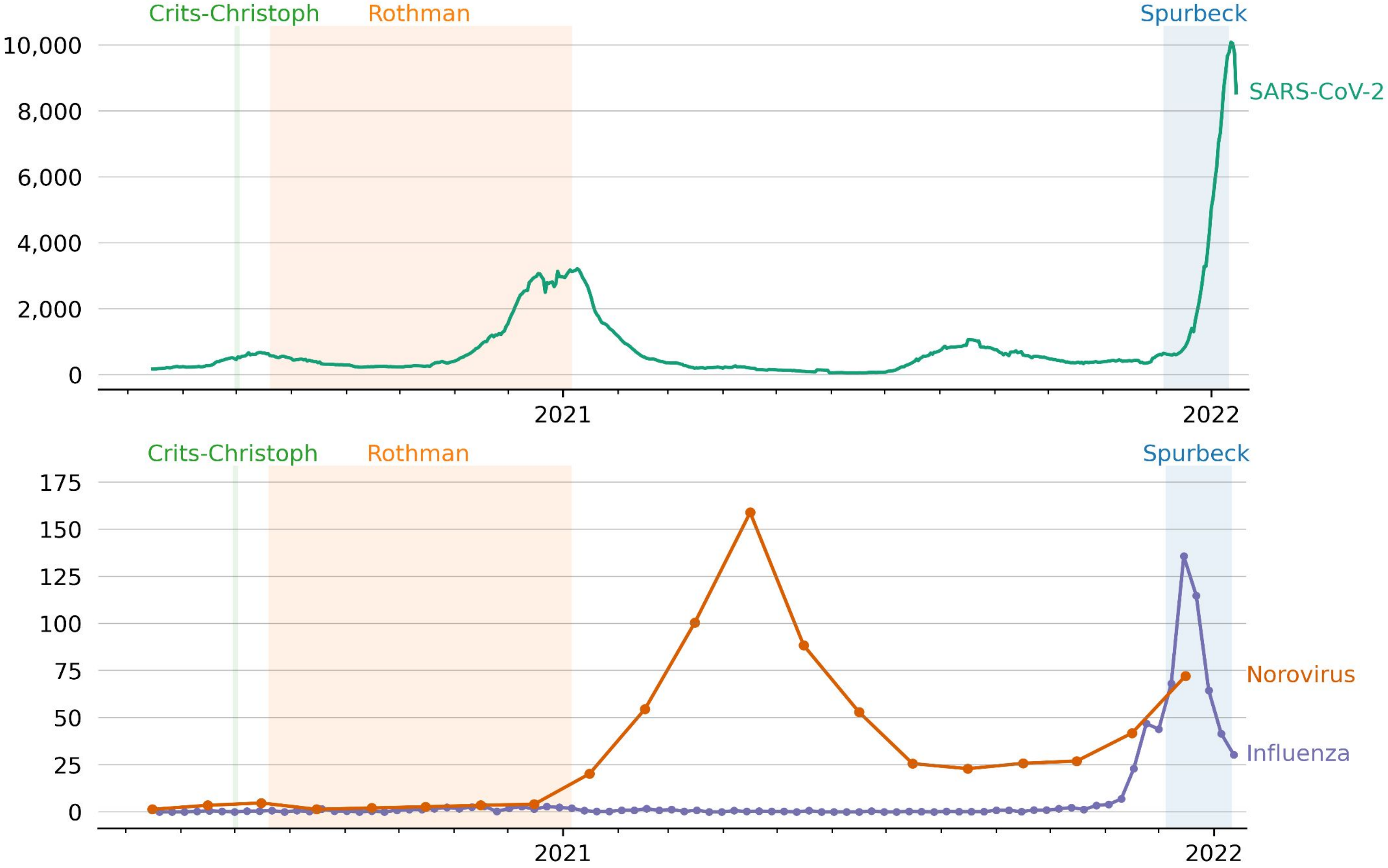
- Chronic: estimate prevalence
- Acute: estimate incidence



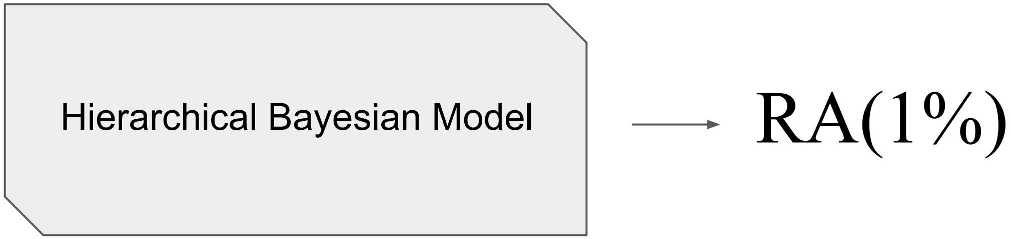
Weekly Infections per 100,000 people



Weekly Infections per 100,000 people



Estimating RA(1%)

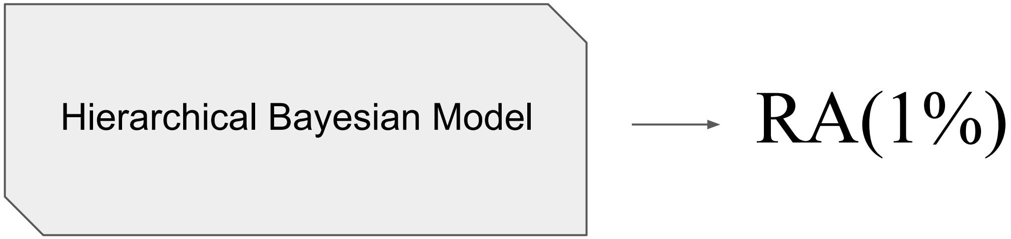


Hierarchical Bayesian Model

RA(1%)

Estimating RA(1%)

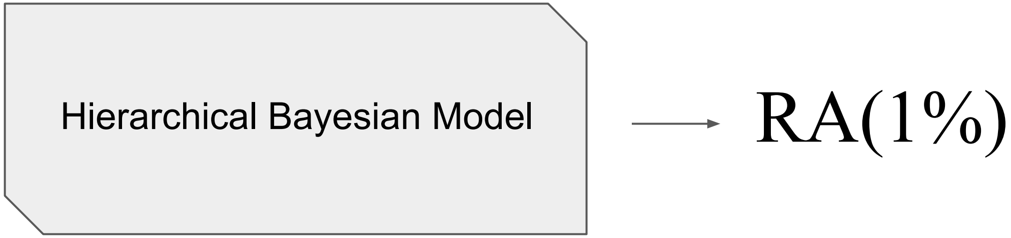
- Chronic: RAp(1%)



Hierarchical Bayesian Model → RA(1%)

Estimating $\mathrm{RA}(1\%)$

- Chronic: $\mathrm{RA}p(1\%)$
  - Relative abundance at 1% prevalence

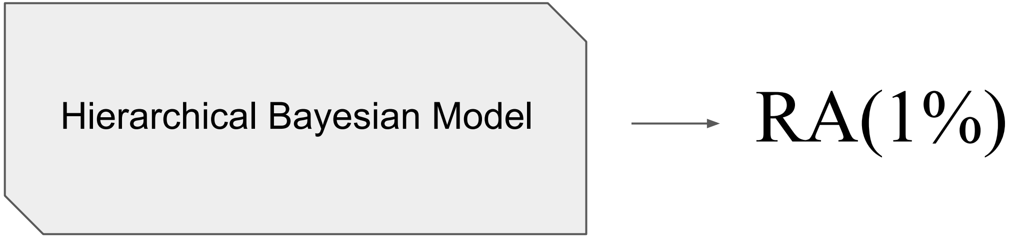


Hierarchical Bayesian Model

$\mathrm{RA}(1\%)$

Estimating $\mathrm{RA}(1\%)$

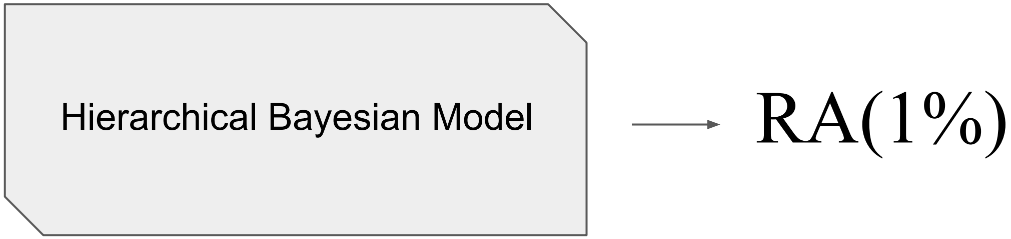
- Chronic: $\mathrm{RA}p(1\%)$
  - Relative abundance at 1% prevalence
- Acute: $\mathrm{RA}i(1\%)$



Hierarchical Bayesian Model → $\mathrm{RA}(1\%)$

Estimating $\mathrm{RA}(1\%)$

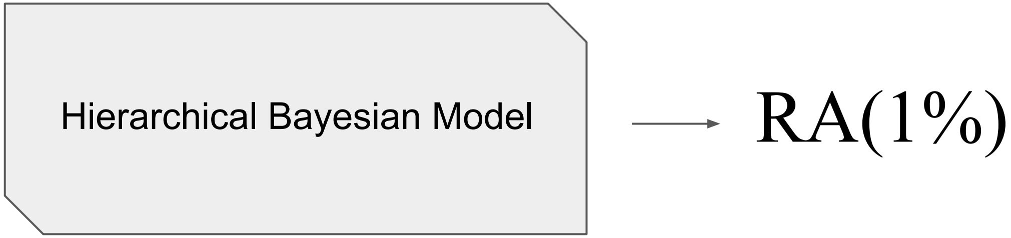
- Chronic: $\mathrm{RA}p(1\%)$
  - Relative abundance at 1% prevalence
- Acute: $\mathrm{RA}i(1\%)$
  - Relative abundance at 1% weekly incidence

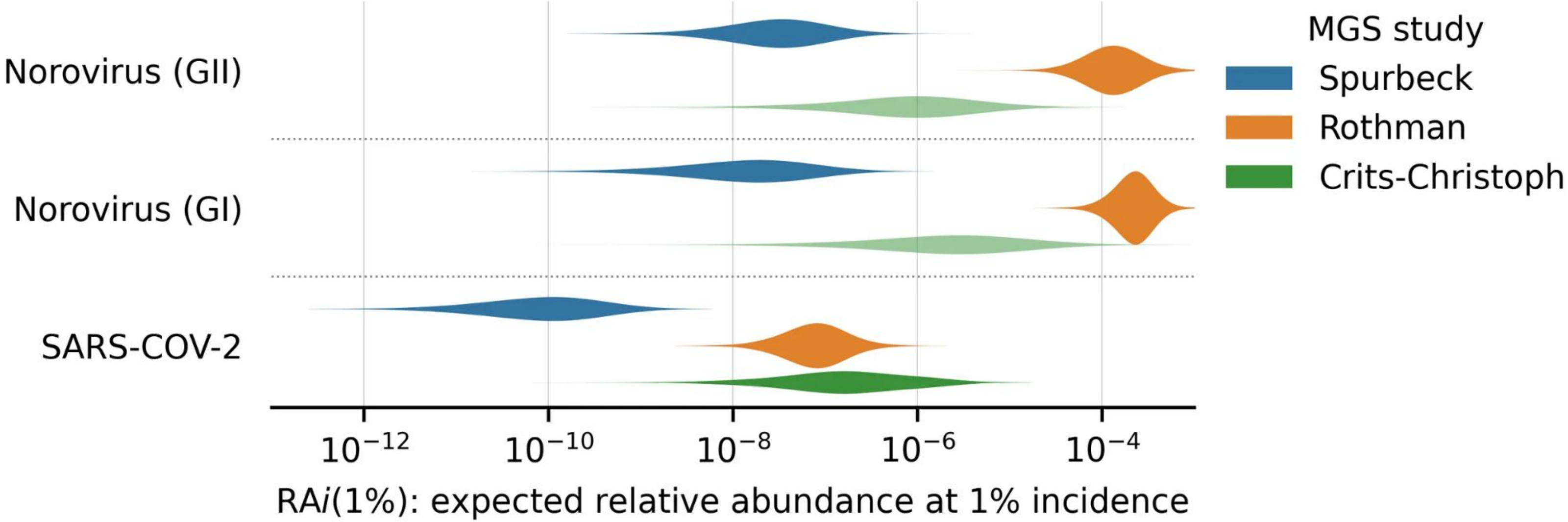


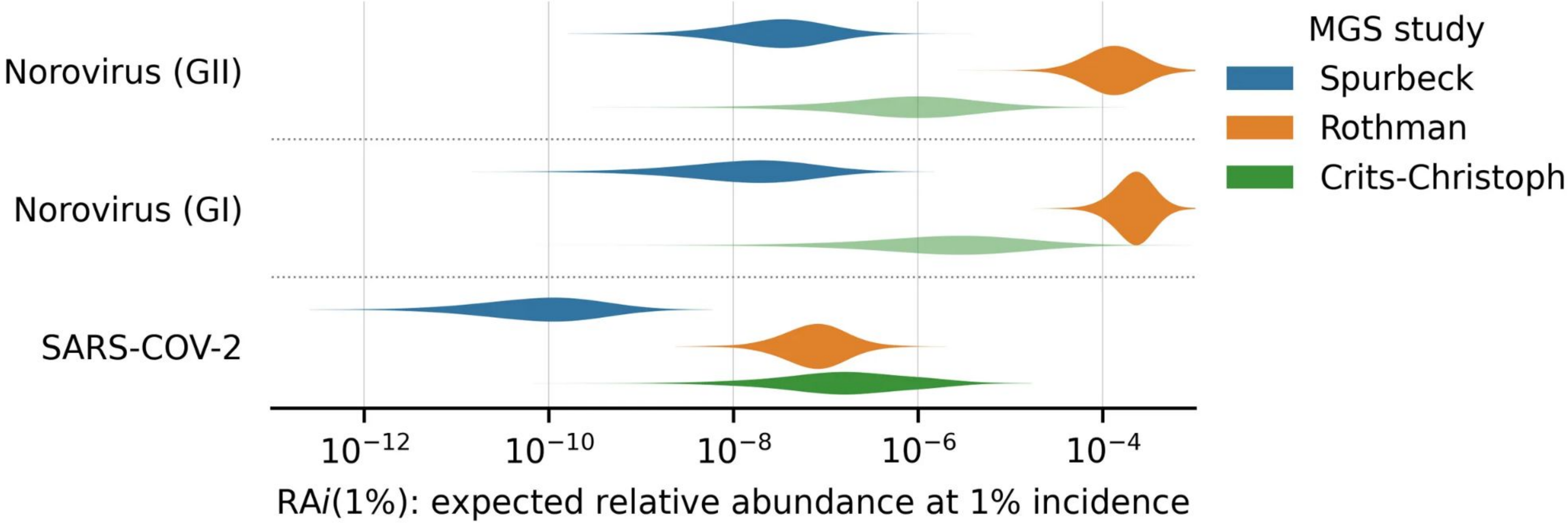
Hierarchical Bayesian Model → $\mathrm{RA}(1\%)$

Estimating $\mathrm{RA}(1\%)$

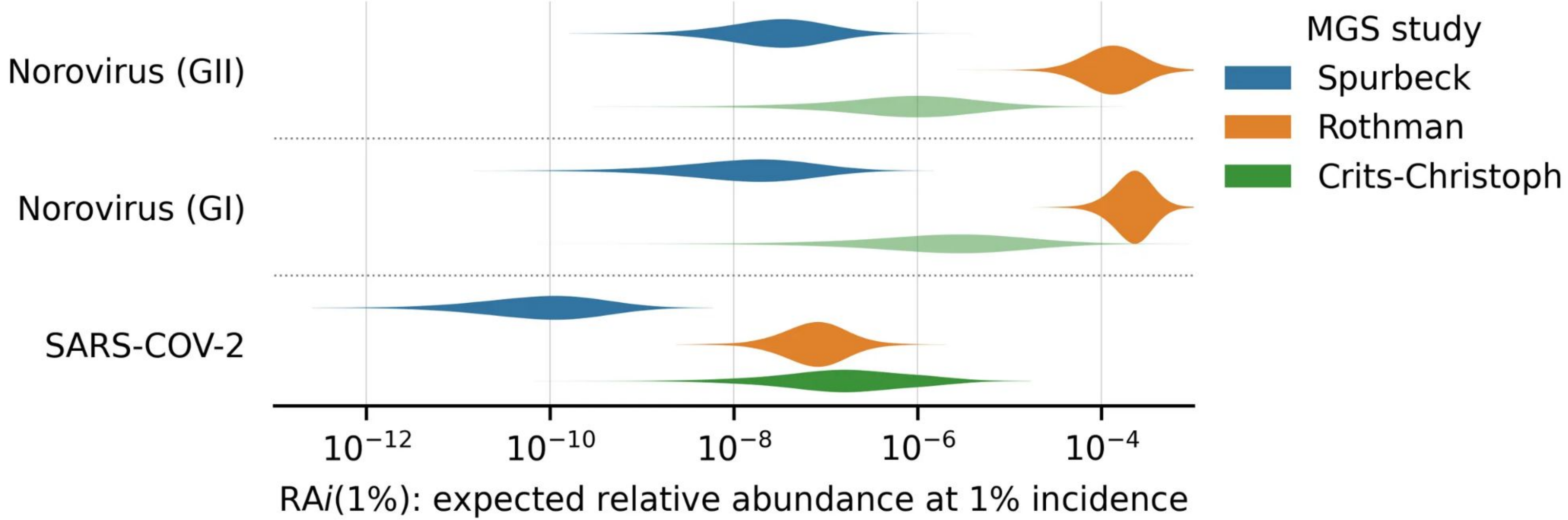
- Chronic: $\mathrm{RA}p(1\%)$
  - Relative abundance at 1% prevalence
- Acute: $\mathrm{RA}i(1\%)$
  - Relative abundance at 1% weekly incidence
- Hierarchical Bayesian logistic regression model



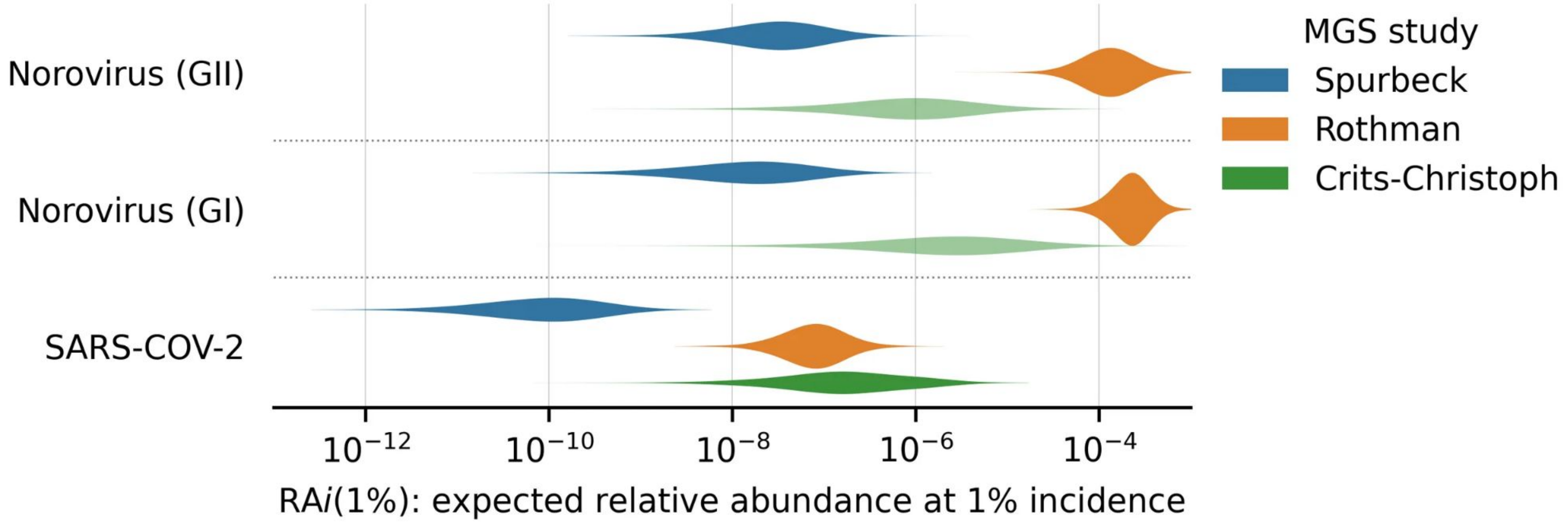

Norovirus (GII)
Norovirus (GI)
SARS-COV-2
MGS study
Spurbeck
Rothman
Crits-Christoph
10^{-12}
10^{-10}
10^{-8}
10^{-6}
10^{-4}
RA*i*(1%): expected relative abundance at 1% incidence



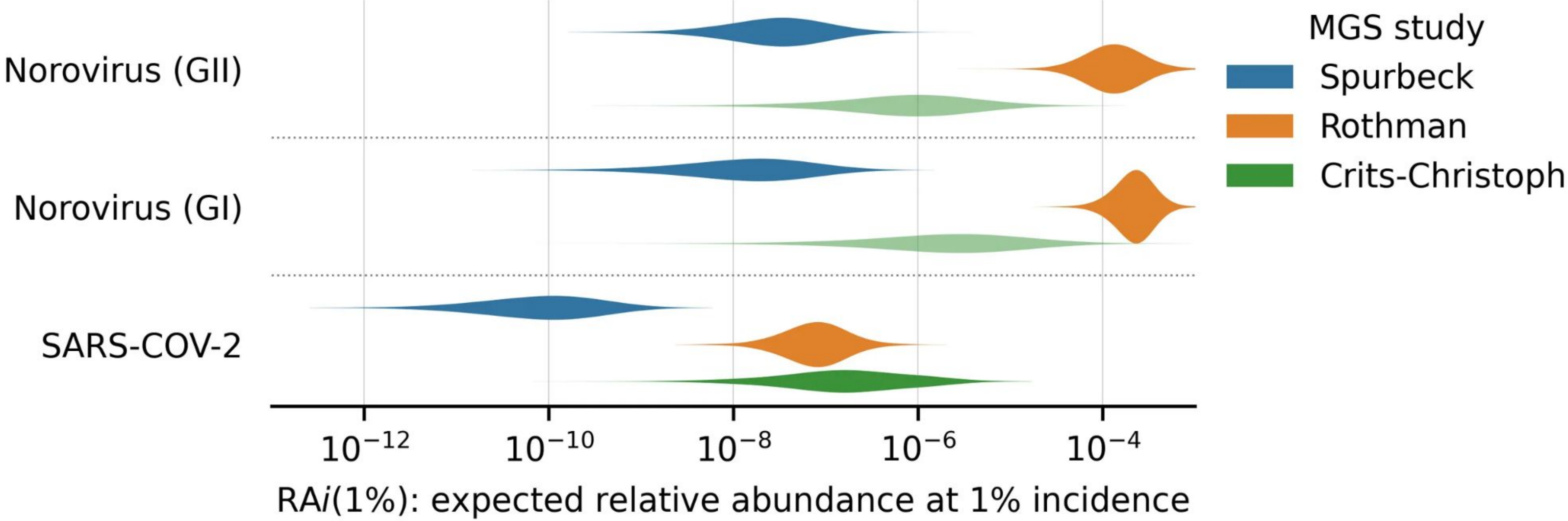
- Lots of variation by virus



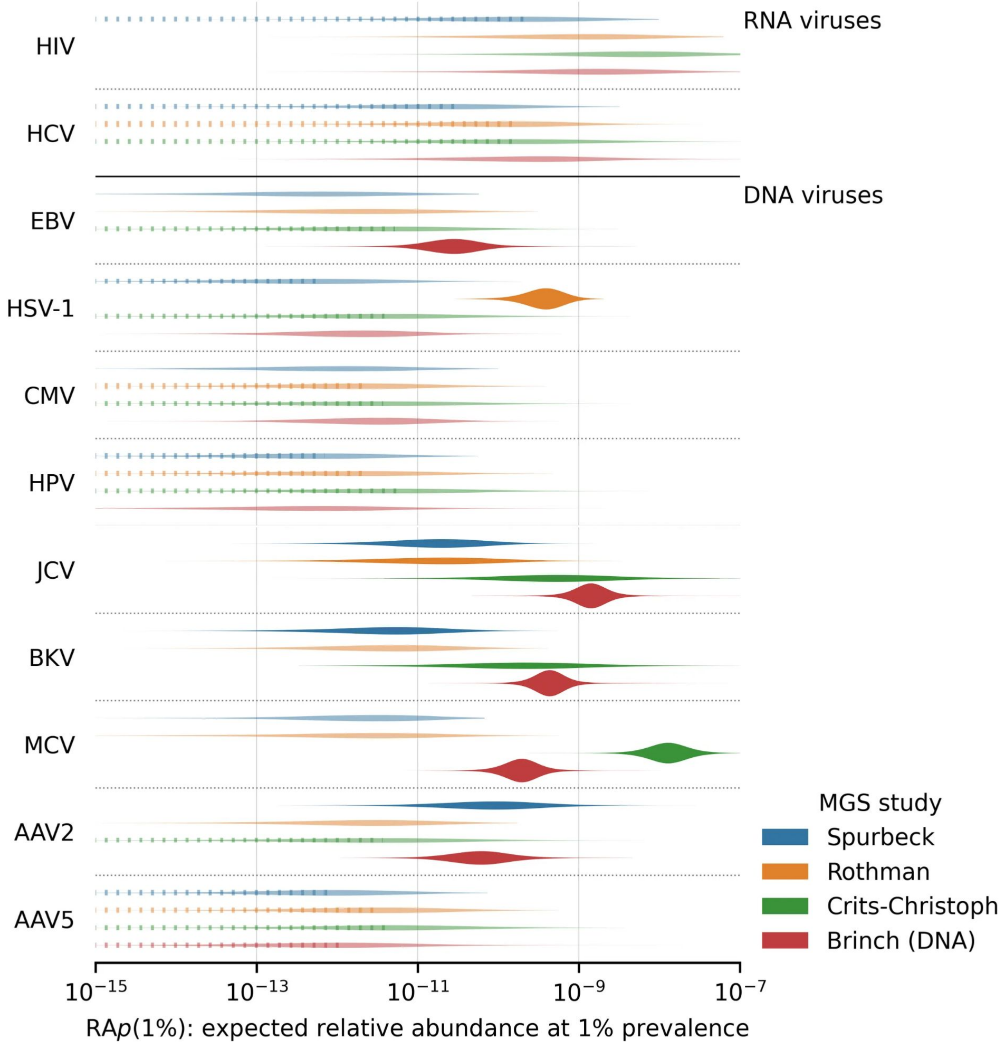
- Lots of variation by virus
  - Norovirus (~1e-4) vs Sars-Cov-2 (~1e-7) in Rothman

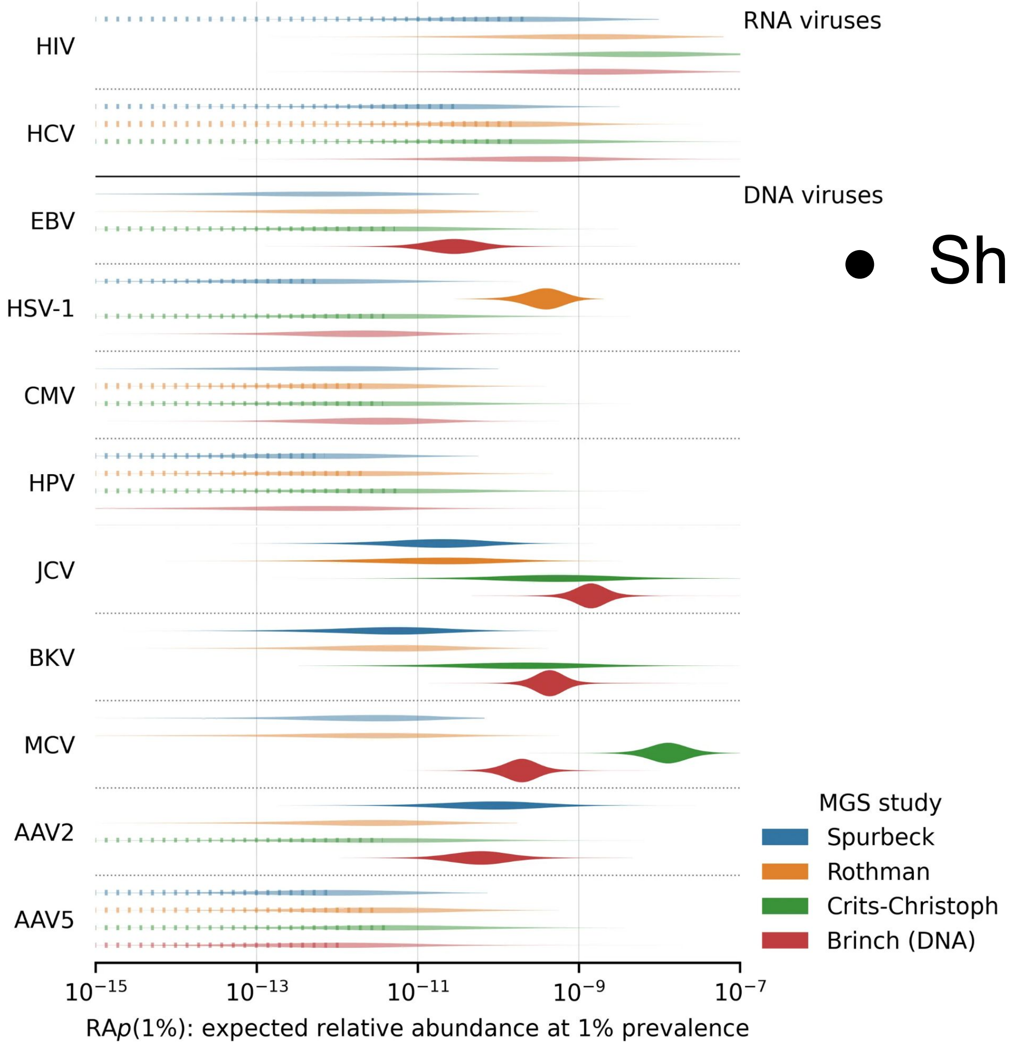


- Lots of variation by virus
  - Norovirus (~1e-4) vs Sars-Cov-2 (~1e-7) in Rothman
- And by study

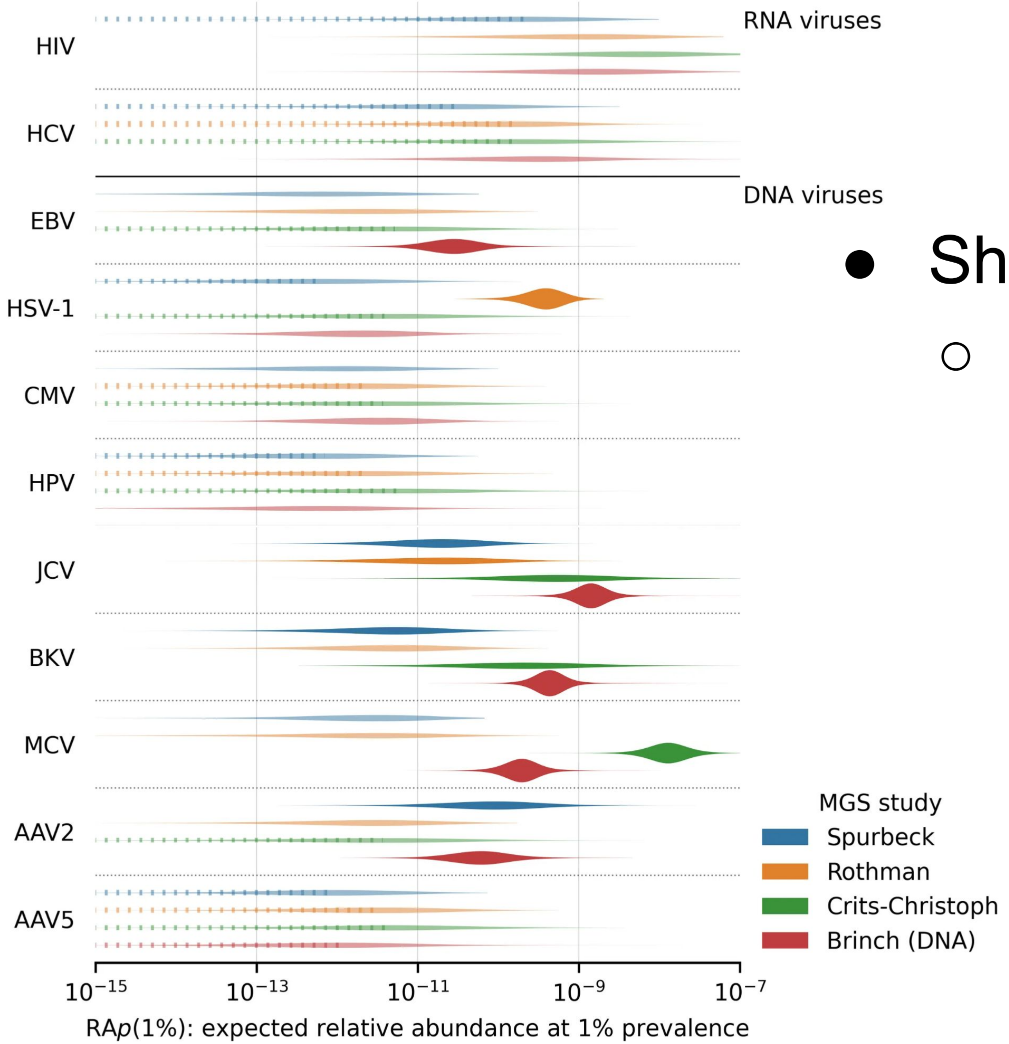


- Lots of variation by virus
  - Norovirus (~1e-4) vs Sars-Cov-2 (~1e-7) in Rothman
- And by study
  - Spurbeck is consistently lower

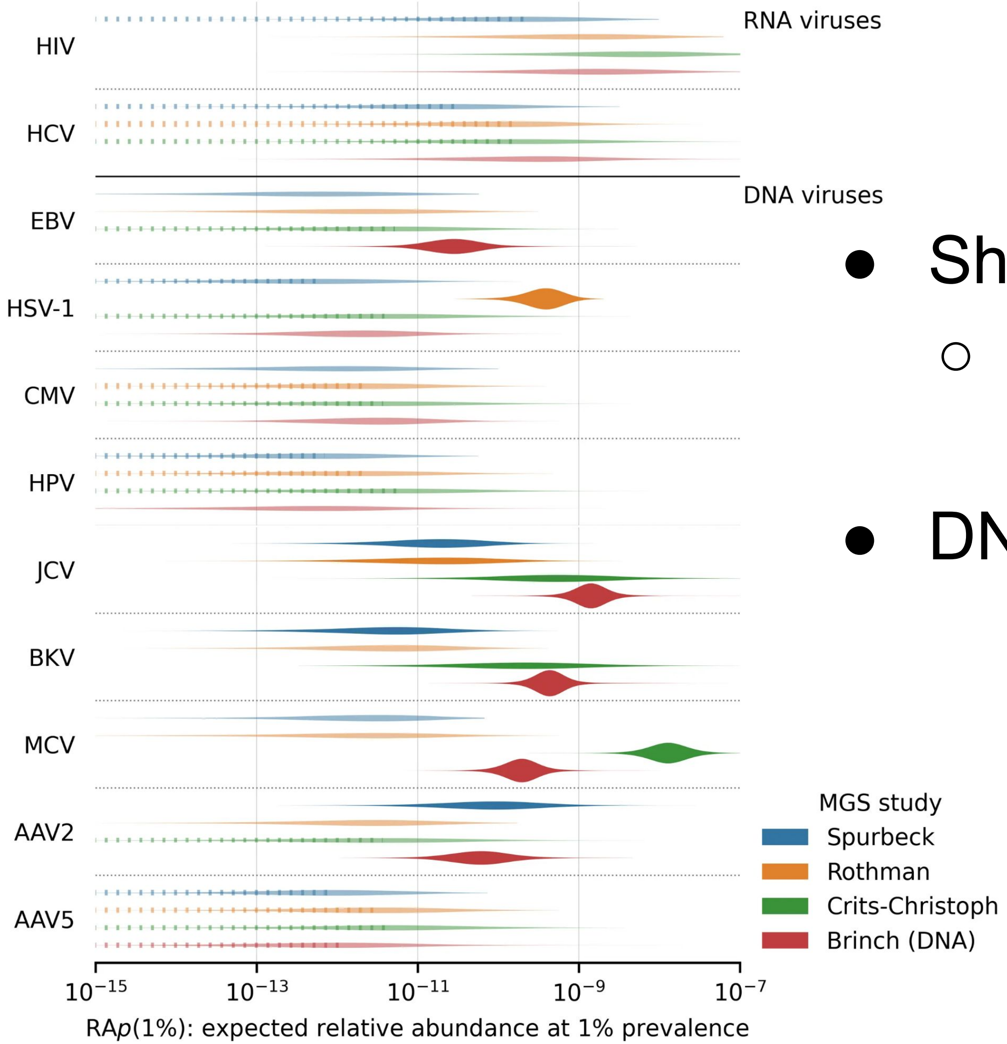

RNA viruses
HIV
HCV
DNA viruses
EBV
HSV-1
CMV
HPV
JCV
BKV
MCV
AAV2
AAV5
10^{-15}
10^{-13}
10^{-11}
10^{-9}
10^{-7}
RAp(1%): expected relative abundance at 1% prevalence
MGS study
Spurbeck
Rothman
Crits-Christoph
Brinch (DNA)



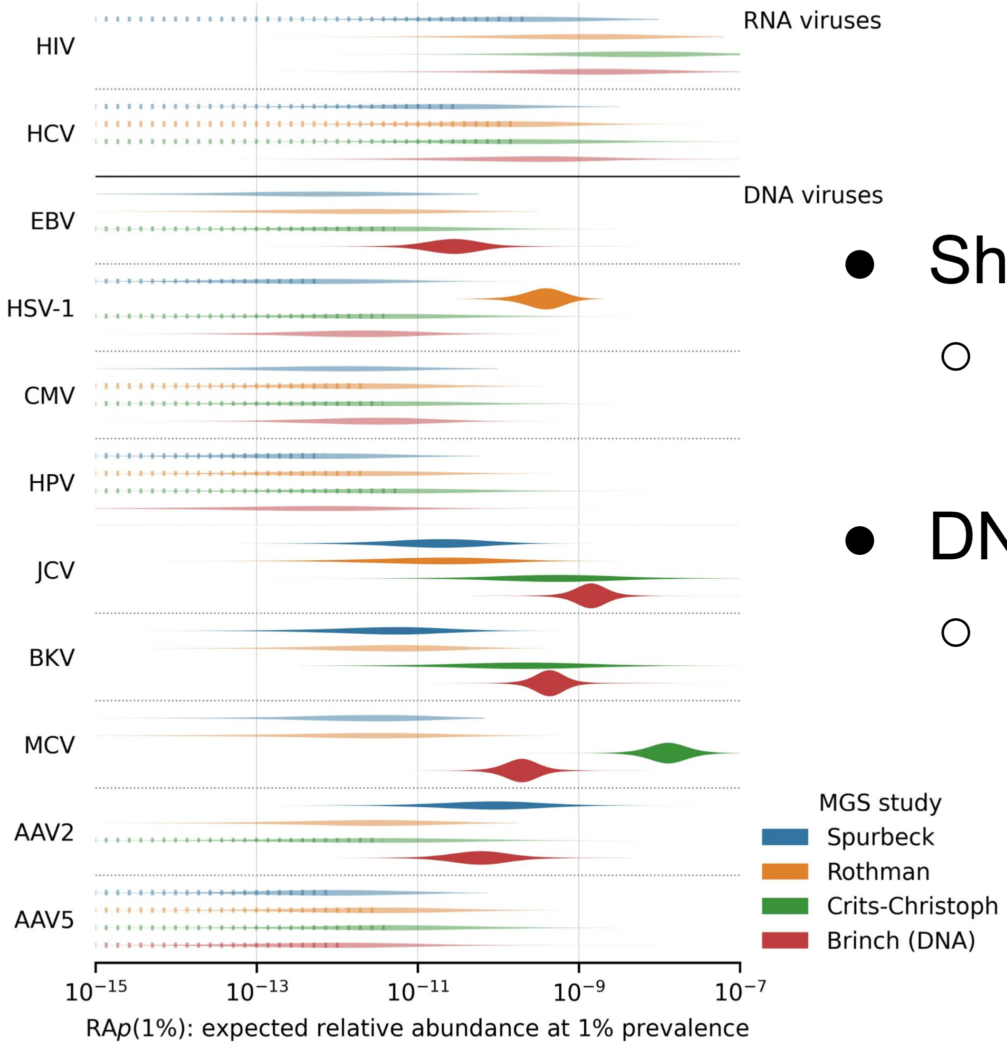
- Sharper estimates in Brinch



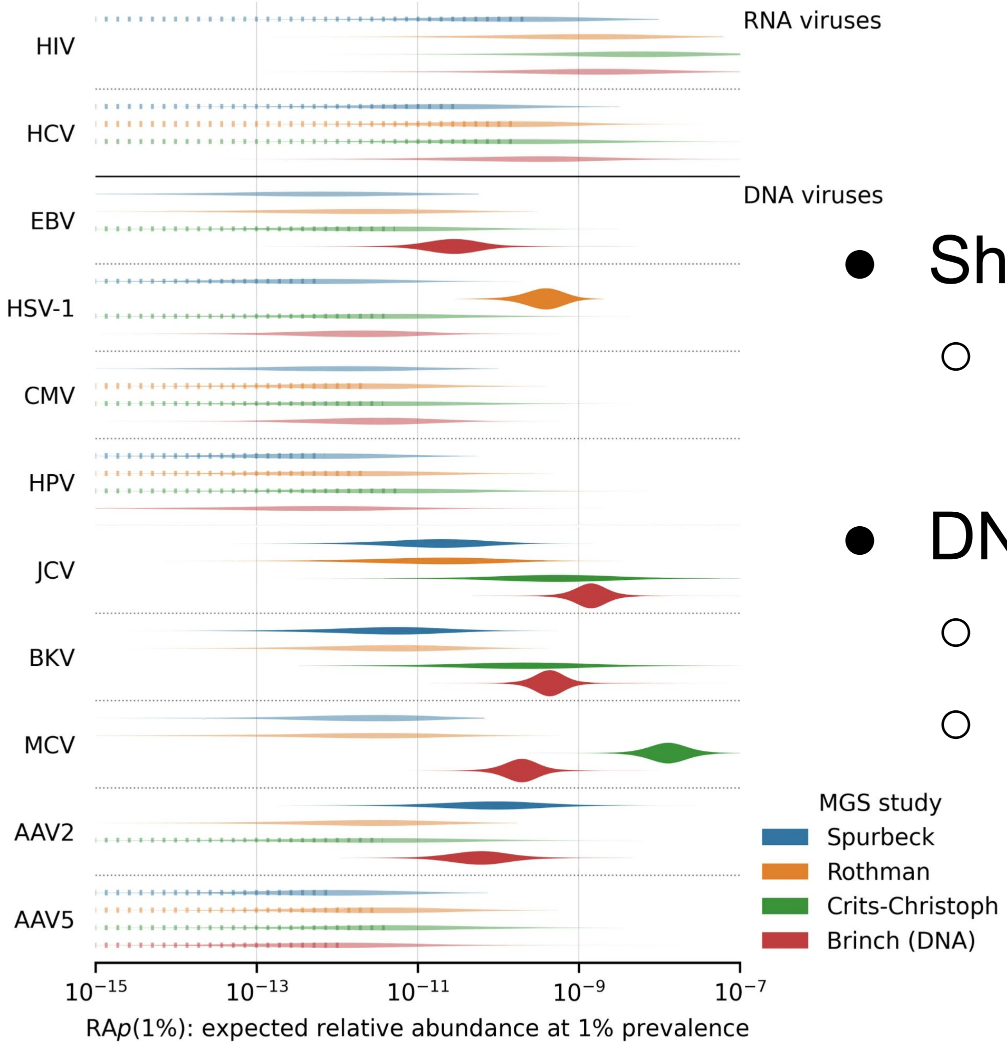
- Sharper estimates in Brinch
  - Many more reads than the three other studies



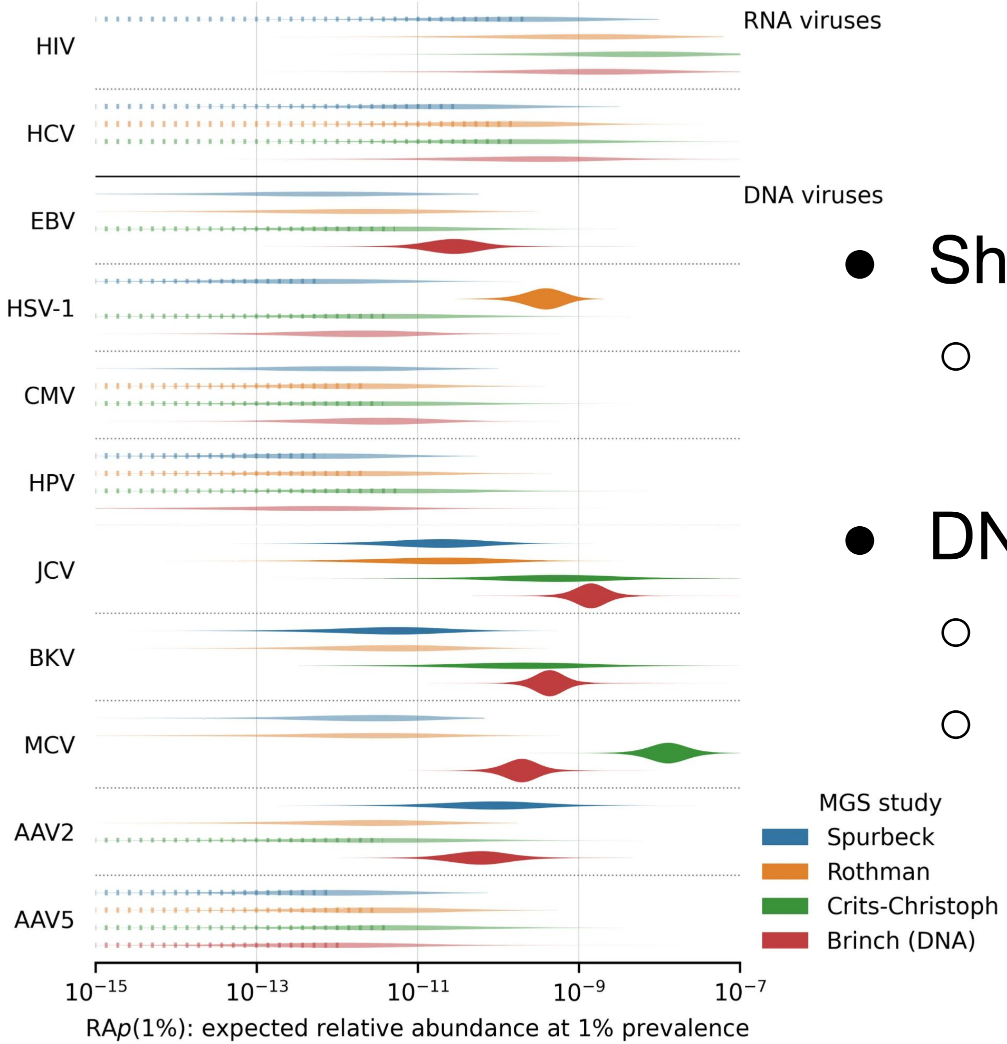
- Sharper estimates in Brinch
  - Many more reads than the three other studies
- DNA viruses in RNA data



- Sharper estimates in Brinch
  - Many more reads than the three other studies
- DNA viruses in RNA data
  - Usually lower $\mathrm{RA}p(1\%)$



- Sharper estimates in Brinch
  - Many more reads than the three other studies
- DNA viruses in RNA data
  - Usually lower RAp(1%)
  - But not always



- Sharper estimates in Brinch
  - Many more reads than the three other studies
- DNA viruses in RNA data
  - Usually lower $\mathrm{RA}p(1\%)$
  - But not always
    - ex: MCV and HSV-1

Limitations

Limitations

- Public health estimates include underreporting factors, which are not very reliable

Limitations

- Public health estimates include underreporting factors, which are not very reliable
- Seasonal viruses were suppressed by Covid-19 response

Limitations

- Public health estimates include underreporting factors, which are not very reliable
- Seasonal viruses were suppressed by Covid-19 response
- Studies were generally underpowered for this purpose

Limitations

- Public health estimates include underreporting factors, which are not very reliable
- Seasonal viruses were suppressed by Covid-19 response
- Studies were generally underpowered for this purpose
    - Deep sequencing (more reads) during a higher-infection time would allow better estimates

Conclusion

Conclusion

- Linked public health data to sequencing data to get estimates of relative abundance as a function of incidence or prevalence: $\mathrm{RA}(1\%)$

Conclusion

- Linked public health data to sequencing data to get estimates of relative abundance as a function of incidence or prevalence: $\mathrm{RA}(1\%)$
- Useful for people modeling detection approaches

Conclusion

- Linked public health data to sequencing data to get estimates of relative abundance as a function of incidence or prevalence: $\mathrm{RA}(1\%)$
- Useful for people modeling detection approaches
- Let's extend this approach and get better estimates!

Conclusion

- Linked public health data to sequencing data to get estimates of relative abundance as a function of incidence or prevalence: $\mathrm{RA}(1\%)$
- Useful for people modeling detection approaches
- Let's extend this approach and get better estimates!

- Full report: `data.securebio.org/p2ra`