

Background

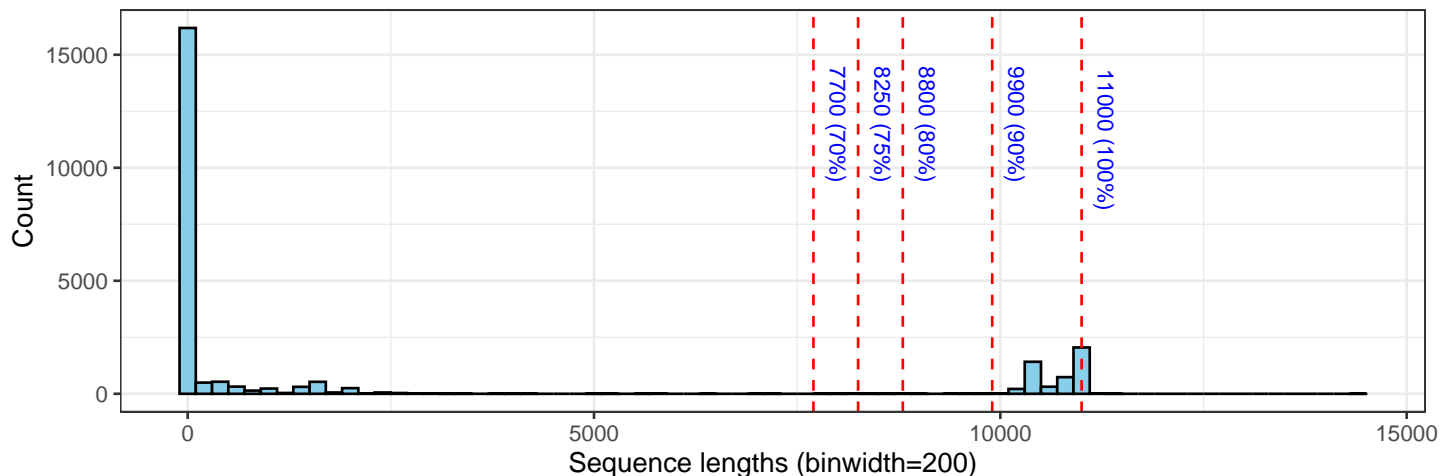
CDC website: <https://www.cdc.gov/west-nile-virus/index.html>

Exploratory Graphics

How long is the WNV genome?

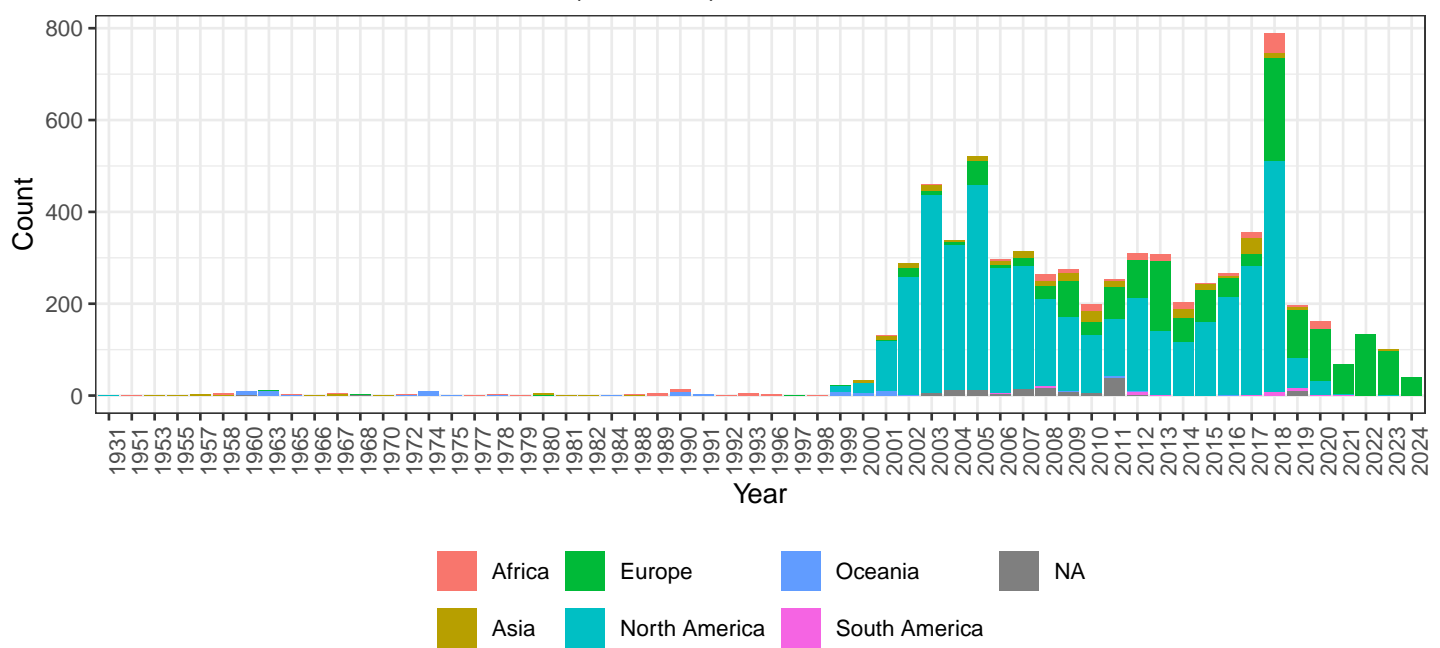
WNV is approximately 11k nt long. The samples pulled from NCBI GenBank have the following distribution of lengths. We can pick a length cut-off for the phylogenetic tree:

Sequence lengths (all data = 24037)



How well sampled is WNV across time and space?

WNV entries with collection date (n = 6692)



The earliest samples seem to be from 1931. However, these are EF631122 and EF631123 and are supposedly collected in Illinois, USA, much earlier than the NY99 event. We might need to dig into publications that contain these GenBanks to

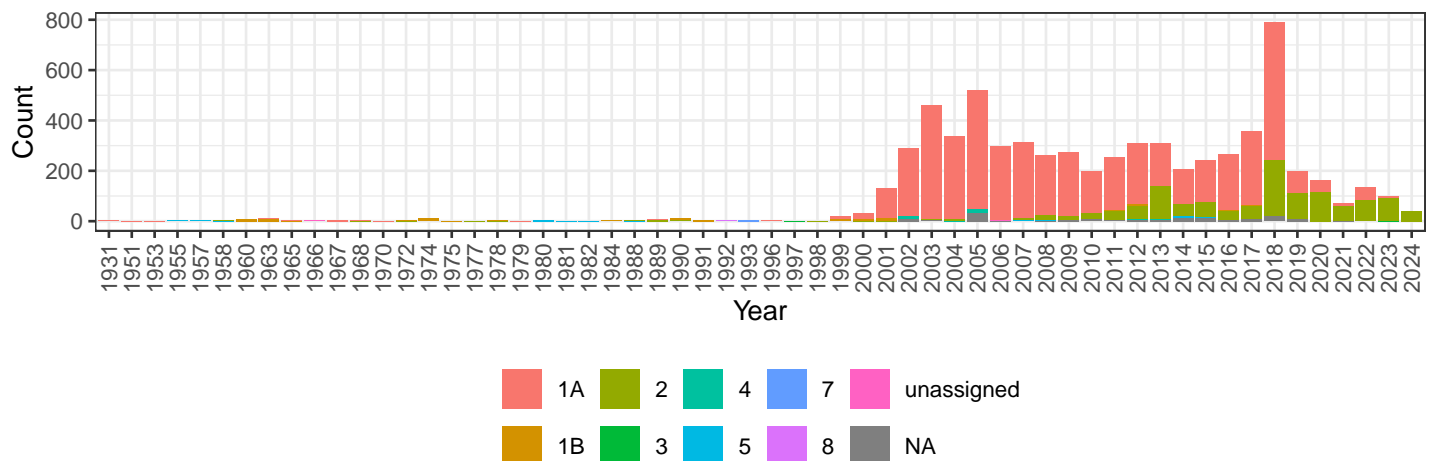
check if this collection date is correct. Such as publication:

- Amore, G., Bertolotti, L., Hamer, G.L., Kitron, U.D., Walker, E.D., Ruiz, M.O., Brawn, J.D. and Goldberg, T.L., 2010. Multi-year evolutionary dynamics of West Nile virus in suburban Chicago, USA, 2005–2007. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 365(1548), pp.1871-1878.

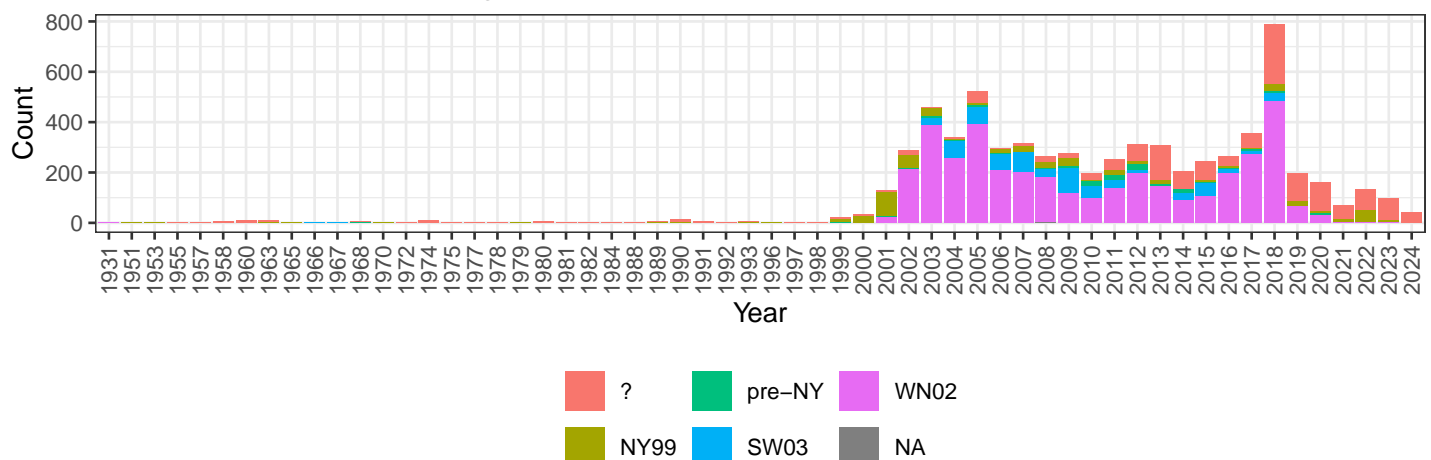
What lineages are available for WNV?

- **Global lineage system:** Koch, R. Tobias, Diana Erazo, Arran J. Folly, Nicholas Johnson, Simon Dellicour, Nathan D. Grubaugh, and Chantal BF Vogels. "Genomic epidemiology of West Nile virus in Europe." *One Health* (2023): 100664.
- **USA-based system:** Hadfield, J., Brito, A.F., Swetnam, D.M., Vogels, C.B., Tokarz, R.E., Andersen, K.G., Smith, R.C., Bedford, T. and Grubaugh, N.D., 2019. Twenty years of West Nile virus spread and evolution in the Americas visualized by Nextstrain. *PLoS pathogens*, 15(10), p.e1008042.

Count of global-based lineage samples with collection date (n = 6692)



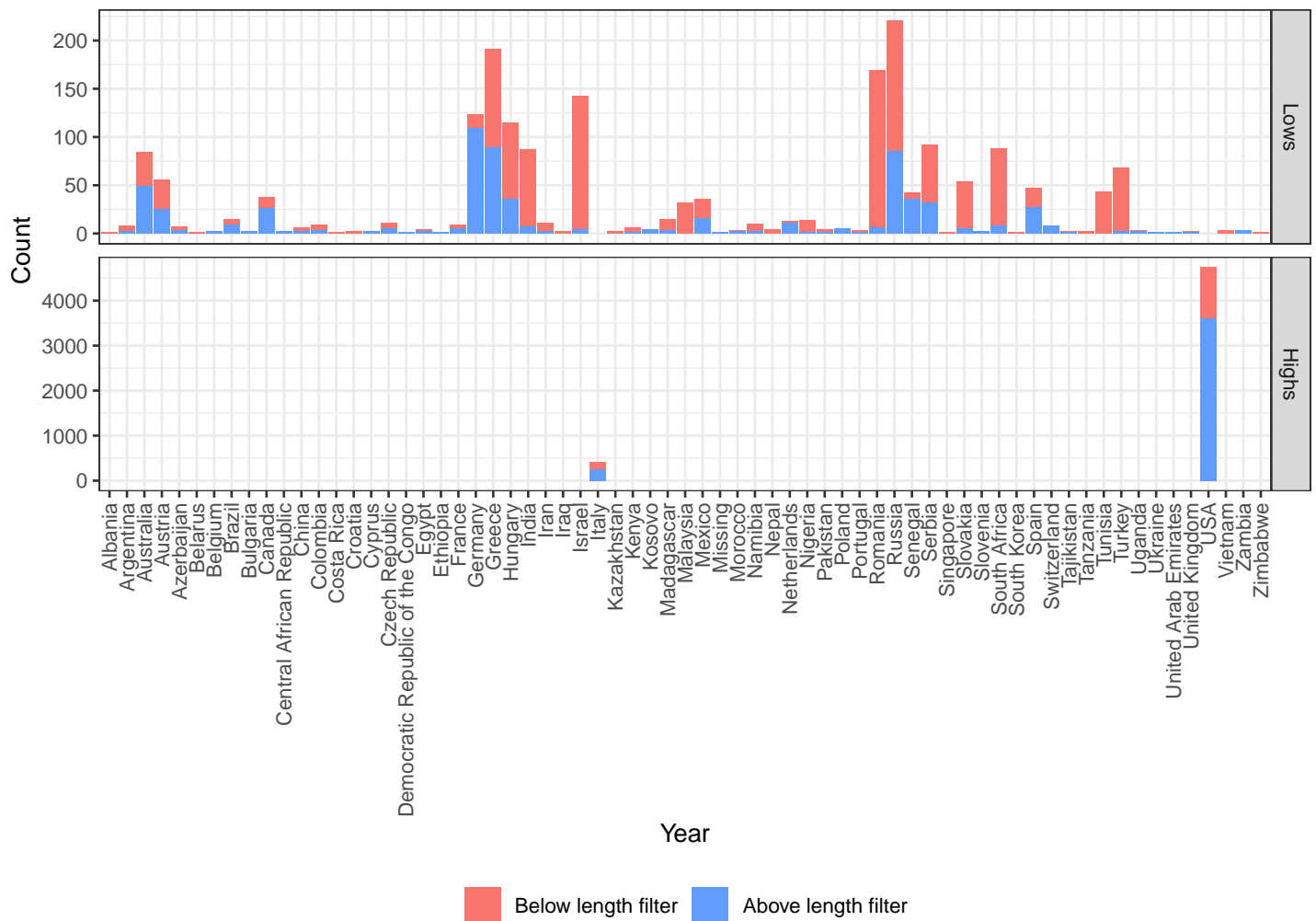
Count of USA-based lineage samples with collection date (n = 6692)



Global subsampling

If min length is set to 80 percent, or 8800nt:

WNV entries that were above or below filter, by country (n = 4780 where min-length=8800)



- Total WNV records = 24037. Note how this includes records without collection date
- Total WNV records that passes the length filter = 4780

If this number is lower than 4k, probably don't need to sub sample further.

Who are the main submitters of sequence data?

Table 1: Top 25 most frequent sequence submitters with their region and countries

| authors | n | regions | countries |
|-------------------|-------|-------------------|------------------------|
| blatt et al. | 15308 | NA | NA |
| grubaugh et al. | 1180 | North America | USA |
| feriancek et al. | 627 | North America | USA |
| newman et al. | 479 | North America | USA |
| shabman et al. | 458 | North America | USA |
| linnen et al. | 342 | NA | NA |
| henn et al. | 273 | North America, NA | USA, NA |
| bertolotti et al. | 208 | North America | USA |
| herring et al. | 199 | North America | USA, Canada |
| derby et al. | 171 | NA | NA |
| dinu et al. | 161 | Europe | Romania |
| ebel et al. | 141 | NA, North America | NA, USA |
| duggal et al. | 132 | North America, NA | USA, NA |
| amore et al. | 126 | North America | USA |
| anderson et al. | 126 | North America | USA |
| phillips et al. | 112 | North America | USA |
| nagy et al. | 109 | Europe | Hungary, Serbia |
| platonov et al. | 109 | Europe, Asia | Russia, Azerbaijan |
| davis et al. | 108 | North America, NA | USA, NA |
| armstrong et al. | 104 | North America, NA | USA, NA |
| swetnam et al. | 91 | North America | USA |
| bernardin et al. | 88 | North America | USA |
| ? | 87 | Europe, Asia | Germany, India |
| papa et al. | 84 | Europe | Greece |
| monaco et al. | 83 | Europe, Africa | Italy, Turkey, Tunisia |

What hosts are represented?

WNV entries with collection date (n = 6692)

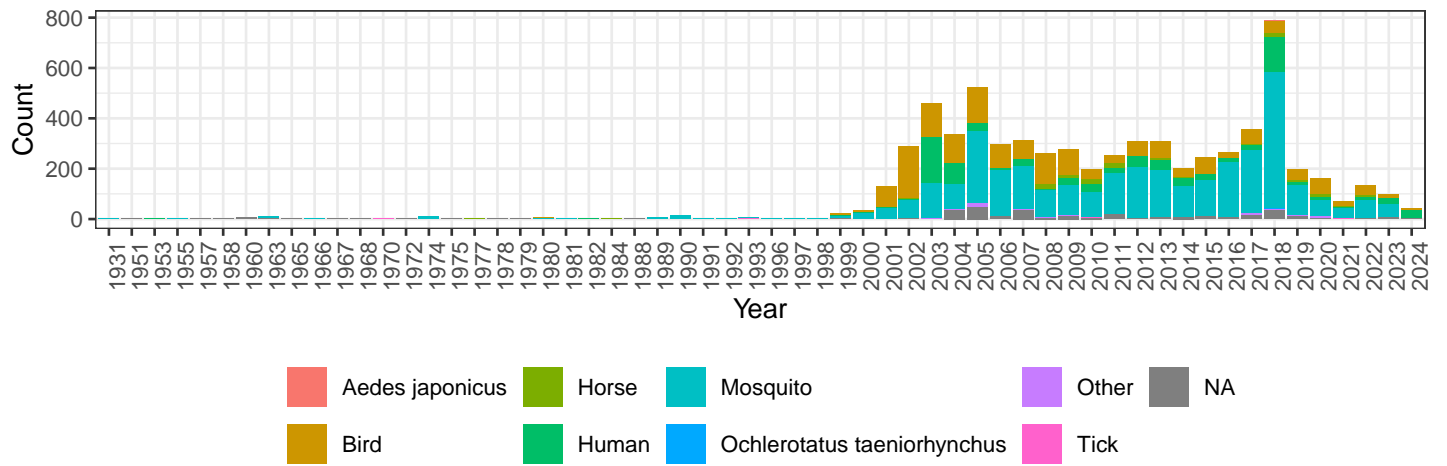


Table 2: Host counts

| host_type | n |
|-----------------------------|-------|
| NA | 17130 |
| Mosquito | 3676 |
| Bird | 2040 |
| Human | 910 |
| Horse | 146 |
| Other | 115 |
| Tick | 17 |
| Aedes japonicus | 2 |
| Ochlerotatus taeniorhynchus | 1 |