**Virus genomes and virus-host interactions in human disease** (working title)

Norman Goodacre1, Prajwal Devkota2, Stefan Wuchty2,3,4,5, Peter Uetz6

1

2Dept. of Computer Science, Univ. of Miami, Coral Gables, FL 33146, USA

3Center for Computational Science, Univ. of Miami, Coral Gables, FL 33146, USA

4Dept. of Biology, Univ. of Miami, Coral Gables, FL 33146, USA

5Sylvester Comprehensive Cancer Center, Miller School of Medicine, University of Miami, Miami, FL 33136, USA

6 Center for the Study of Biological Complexity, Virginia Commonwealth University, Richmond, VA 23284, USA

**Introduction**

Viruses are, together with bacteria, the most important pathogens on earth. While most bacteria are still easy to treat with antibiotics, viruses are much harder to control. That is a direct consequence of their nature, being composed of only a few nucleic acids and proteins and sometimes lipids and a few other compounds. Given their simple composition and structure viruses are completely dependend on their hosts, and hence they have to extensively interact with host proteins and other host components.

In this review we provide an overview of virus diversity and how it related to the diversity of host-virus diversity. We surmise that medically important viruses have received more attention and thus more research, hence more interactions should be known and understood. While this is often the case, there is no simple correlation. In fact, being highly pathogenic does not mean a virus is easy to study or it may not be of interest, given a very restricted geographic range or very narrow host range. For some viruses, such as Zika, the threat is relatively recent so that research has only ramped up during the past few years and there is no extensive data available yet.

**Diversity of human viruses**

Most people are infected by one or more viruses. Wylie et al. detected an average of 5.5 viral genera in each of 102 healthy individuals. Given that no more than 5 habitats were screened (nose, skin, mouth, vagina, and stool) we can safely assume that most people have dozens of different viruses in their body. However, only few of them lead to clinical symptoms or disease. We have thus compiled data on the diversity of human viruses by looking both at sequence diversity (Table 1) and epidemiology (Table 2).

Table 1. Top-20 human viruses by genomes sequenced. Sequence numbers as of July, 2016. Clustered sequenced were clustered at ≥98% sequence identity). C/D = un-/ clustered. Genome data from Genbank.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Type | Family name | Sequences (unclust.) | Disease (examples) | C/D | Total complete genomes | Complete genomes (clustered ) |
| RNA | reoviridae | 65870 | Rare diarrhea | 5.78 | 31945 | 5803 |
| RNA | flaviviridae | 225112 | Zika | 3.20 | 7837 | 2019 |
| DNA | hepadnaviridae | 78558 | hepatitis | 8.16 | 7248 | 1946 |
| plant | geminiviridae | 13158 | --- | 3.02 | 6421 | 2316 |
| RNA | picornaviridae | 85636 | Cold etc | 3.60 | 3447 | 1500 |
| RNA | retroviridae | 716088 | AIDS etc | 2.21 | 2890 | 2103 |
| anim | circoviridae | 7838 | --- | 6.92 | 2706 | 542 |
| RNA | phenuiviridae | 4139 | Rift Valley fever | 4.86 | 1678 | 384 |
| RNA | coronaviridae | 19164 | SARS | 5.61 | 1549 | 320 |
| RNA | potyviridae | 16115 |  | 3.63 | 1536 | 843 |
| DNA | papillomaviridae | 17847 | Warts, cancer | 8.14 | 1364 | 359 |
| DNA | polyomaviridae | 8604 | Rare cancers | 10.60 | 1277 | 164 |
| RNA | filoviridae | 2165 | Ebola | 23.53 | 1259 | 37 |
| RNA | togaviridae | 8924 | rubella | 16.25 | 1239 | 137 |
| RNA | pneumoviridae | 22578 | Cold-like | 13.08 | 1231 | 61 |
| plant | nanoviridae | 3110 | --- | 4.92 | 1183 | 282 |
| RNA | caliciviridae | 32405 | gastroenteritis | 5.30 | 1072 | 292 |
| RNA | paramyxoviridae | 29726 | measles | 10.06 | 1008 | 327 |
| RNA | bromoviridae | 4677 | (plants) | 4.08 | 764 | 384 |
| RNA | arenaviridae | 2639 | (animals) | 2.29 | 758 | 469 |

Statistics of viruses known to infect humans, genomes sequenced, genetic diversity?

Statistics of human disease caused by these viruses, case numbers, mortality, economic damage etc.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Virus (class)** | **Infections** | **Morbidity** | **Mortality** | **Cost** | **Refs** |
| HSV-1/2 | 3.7 B / ~700M | 3M/yr (US) | low | $1B (US) |  |
| HIV-1/2 | 36M (world) |  | 25M total1 |  |  |
| Influenza | >30M (US)3 | 30M (US)3 | 50M 19182 | $10-90B |  |
| Measles | >20 M (ww) ? | <20 M (ww) ? | 140-500k4 | $3-7B (US) |  |
| Hepatitis C | 60-120M (ww) | 4M |  | $10B 5 | R4 |
| Hepatitis B | 248 M ww |  | 600 k ww |  | R1, R2 |
| Zika | 740k S Amer | >2,6k microc. |  |  | R3 |

1 globally, since 1981. 2 Spanish flu of 1918. 3 30 million outpatient visits. 4 The death rate is decreasing, from 535,000 deaths in 2000 to 139,300 deaths in 2010. 5 $10·7 billion in direct medical expenditures in the USA for HCV-related disease from 2010 to 2019

WW = worldwide.

R1 = Maynard JE. Hepatitis B: global importance and need for control. Vaccine 1990; 8 Suppl:S18.

R2 = Ott JJ, Stevens GA, Groeger J, Wiersma ST. Global epidemiology of hepatitis B virus infection: new estimates of age-specific HBsAg seroprevalence and endemicity. Vaccine 2012; 30:2212.

R3 = <https://www.hhs.gov/sites/default/files/Fischer_Zika%20Virus%20Epidemiology%20Update%20Remediated.pdf>

R4 = Wong JB, McQuillan GM, McHutchison JG, Poynard T. Estimating future hepatitis C morbidity, mortality, and costs in the United States. Am J Public Health 2000; 90: 1562–69.

**Virus-host interactions in humans**

Summarizing PPIs infecting human-host interactions

Host-virus interactions: quality over quantity?

Databases:

<http://virhostnet.prabi.fr>

<http://www.ebi.ac.uk/intact/>

<http://virusmentha.uniroma2.it>

**Discussion / Outlook / conclusions**

Which viruses need more effort / discuss mismatch between genome data and PPI data

Phage in human health ?

**Figure 1: (A)** Utilizing a set of human host proteins that are targeted by a variety of different viruses, we determined the enrichment of such target proteins in bins of proteins that have a certain number of interactions in a human protein-protein interaction network. Randomly sampling sets of viral targets, we observed that proteins of increasing degree are preferably targeted by viuses. **(B)** We defined a set of bottleneck proteins in a human protein-protein interaction network as the top 20% of proteins with highest betweeness centrality. In such a set we observed 820 proteis that were targeted by viruses. Randomized sampling of target sets confimed the statistical significance of the observed value, suggesting that bottleneck proteins are prime viral targets. In **(C)** we defined a protein set of top protein rank as the top 20% of proteins with highest protein rank. Randomly sampling viral targets indicated that proteins with protein rank are preferable viral targets as well. **(D)** Determining critical, intermittent and redundant proteins, we observed that critical proteins significantly accrued viral targets while we found the opposite for redundant proteins.



**Figure 2: (A)** Utilizing a set of human protein complexes, we determined the enrichment of viral targets in bins of proteins that appear in a certain number of complexes by randomly sampling viral target sets. Notably, we observed that proteins with increasing occurrence in different complexes appear to be enriched with viral targets. **(B)** Analogously, we determined the enrichmet of viral targets in sets of proteins that appear in different pathways, allowing us to find similar results.