Script for master thesis presentation

Slide number

1.

Welcome everyone , thank your for coming and listening to the defense of my master thesis .
The topic of the master thesis is the characterization of the virome in two unprocessed sequencing data sets.
One locally provided and one set downloaded from the Internet.
I utilized the in house pipeline VirusHuntergatherer for Virus discovery and assembly ad conributed my own application called Virusparies.
Lets start.

2.

Here you see the outline. I start with the introduction. I want to talk about the need of virus discovery, to provide motivation for what was done here. Then I will present the goals. I present the VirusHunterGathrer workflow and the output I primarly used. Then I talk about Virusparies, the R package I developed. I present the data and results of the two datasets I analyzed and end on a conclusion and outlook.

3.

Lets start by answering the question. Why do we need virus discovery?

4.

The first reason is of course epidemics and pandemics. Pandemics used to be a phenomenon that occurred once up to thrice per century, however because of increased air travel, and more contact to both people and animals, we saw an increase in outbreaks. The graph here shows the major viral pandemics affecting humans since the industiralization. In the first 20 years of this century, we already saw 6 major outbreaks and in this decade we saw for example the mpox outbreak with a new variant only recently being identified in germany. And according to the virome project , there might be as much as 1.67 million viruses yet unidentified, with some having the capacity to infect humans. **That means, we need to monitor samples and enviroments in a larger scale then ever before to identify disease causing viruses before the outbreak occurs.**

5.

Not all viruses are harmful. We harbor approximately 10^{13} viruses, primarily infecting our microbiota and showing high heterogeneity across body sites. The interactions between viruses, their hosts, and the environment are complex. For example, endogenous retroviruses (HERVs) integrated into our genome can be linked to certain cancers but are also essential for placenta morphogenesis. Meanwhile, herpesvirus 6 is found in nearly every adult and is harmless unless the individual is immunocompromised. On the other hand, Anelloviridae viruses, although present even in immunodeficient individuals, appear to be commensal. This complexity is not completly understood, since only a fraction of the virome is understood and we need virus discovery to understand it.

6.

Lastly we can use virus discovery for research and to improve our understanding of viral history. Take for example Hepatitis C, a virus of the Flaviviridae family. 50 million are infected suffering from liver fibrosis and sometimes liver cancer. The current treatment consists of specific regimens that target specific steps of the HCV lifecycle directly. And those regimens are effective, however also expensive. People of lower income cannot afford repeated treatments and their a certain risk groups such as injection drug users who are likely to have multiple infections. For those people the only long term solution is the development of a vaccine, however chimpansees, which were used previosly for research are now endangeored and can no longer be used. One solution to guarantee further research is the identification of yet unknown hepaciviruses, which are closely related and cause similar symptoms as HCV, but in hosts which are easier to manage. For example the gerge baker virus is the current surrogate for HCV, however it too has limitations. This would also bring a better understanding of the virus history. For example, preior to 2011 there was the assumption that hepaciviruses can be found only in us and chimpansees. Now thanks to the indentification of new hepaciviruses, we know that they infect other mammalians and even non-mammalian hosts.

7.

Here I want to show the goal of my thesis.

First, I got liver sequencing data from patients post transplantation. Some rejected the liver. I want to see if I can identify virus sequences and if those sequences can be linked to the rejection.
Then I got a mammalian samples, where viruses were previosly discovered but not assembled. I performed the assembly and functional annotation, to see if the sequences are genuine virus sequences.
Lastly I developed an R package Virusparies, which I will also present here.

8.

Next, lets talk about the VirusHunterGatherer pipeline and Virusparies.

9.

Here, you see the workflow of VirusHunterGatherer. I did not develop it. I only applied VhVg. It takes locally available files as input or it can download sequencing data from the sequencing read archive. The dat is processed and here we see the first VirusHunter step. VirusHunter performs a sensetive homology search for viral reads in the unprocessed sequencing data. For that it utilizes profile hidden markov models. Profile hidden markov models are build based on conserved proteins characteristic of a specific virus family. That means we can build can compare a sequence against the RNA polymerase of an entire family, instead of 1 sequence against another. Based on that we get reads, which we then filter to retain only viral contigs and we create micro contigs and align them against a viral data base to make sure that we only have viral sequences left and to perform taxonomic annotation.
We then utilize those micro contigs in VirusGatherer as seeds to assemble longer contigs up to complete genomes and perform againt a taxonomic annotation.
The output are the contigs, but also tables, which return the results for the screening of profiles and the taxonomic annotation.
Here we see that if we use data from the SRA we also obtain host and study information, while for local we only get the fastq id. We get the result for the profile search with the best E-value and number of reads for that profile. For taxonomy annotation we get also the best E-value, the

protein sequence identity to the viral database, what viral subject it aligned to and in gatherer we also get the length of the contig.

We also care for the protein sequence alignment percentage to the closest virus in the database, because we define the novelty of the viral sequence based on its identity. A 100 % similary would indicate that we identified the same virus as the database. Here I use everything thats below 90 % sequence identity as a novel viral sequence.

I ulized those tables as input in the R package Virusparies, which I developed

10.

The R package Virusparies handles import and export, allows processing and calculations of summary statistics, and the generation of plots and tables which you will see here. That allows us to generate reports which are easier to read than the hittable output.

11.

Now lets start with the first data set. Here, the Taubert working group provided whole transcriptome sequencing data from liver transplant patients.

12.

and some of those patients had either no rejection of the liver or an antibody or t cell based rejection of the liver. As a whole 332 samples were provided spread across 557 FASTQ files. The Taubert working group is interested in antibody mediated rejection and the goal here is to see if we can detect viral sequences in those samples which can be linked to liver rejection, to see if virus presence could be another factor yet missed in their study. I performed a viral search for both DNA and RNA viruses.

13.

Now. I ran the screening for viruses I the samples and the first discovery I made was that in only a small portion of samples were viral sequences identified with a maximum of 165 files files detecting DNA viruses. The number of alignments to viral families was highest among RNA viruses, however in this work I used an E-value cutoff of 1e-5 to define significant hits. And we see that we have only a few significant hits, with the majority being close to the cutoff.

14.

If we look only at the significant hits, we see see that only large DNA and RNA viruses returned significant hits, but only for alignments against 6 families and here againt found in only a maximum of 11 files out of 557. The read coverage was also small with only 36 reads in the search for RNA viruses and 23 for DNA viruses. And except for 1 family, every sequence was closely aligned to the viral database.

15.

Those micro-contigs from Hunter were then used for assembly of longer contigs in gatherer. Here we see that some sequences from Hunter do not appear in Gatherer and some viral sequences such as Pox, Hepadna and Kolmioviridae appear here but not in Hunter. Thats because HunterGatherer uses a small internal cutoff and allowed those sequences through, but we do not consider those significant, even if they appear here. So despite some sequences aligning to hepatitis B and D, those sequences are not considered significant here.

Here again we see that the sequences are closely aligned to the viral database, but we also see that those contigs are really short contigs below 300 nucleotides, giving us little sequence information.

16.

That leads to the conclusion that in 557 files only 6 files returned assembled contigs. So for the remaining 551 no viral sequences were identified. And for the 6 files we found hits which were close to the E-value cutoff or hits which were significant only in Gatherer. Those contigs were short, providing little sequence information. Leading to the conclusion that there is no conclusive evidence linking viral presence to liver rejection.

17.

Next the second dataset. Here I got an existing VirusHunter dataset where RNA virus contigs have been discovered previosly.

18.

The existing hittable contained 34.000 unique SRA ids. Here I filtered for signficant hits based on E-value, picked only those where number of reads was above 4 to guarantee that something got assembled, and I was interested in novel sequences so I picked a cut off below 90% sequence identity.
That returned 1666 datasets, which was still to much. So I setelcted the groups Flavi, Hepe-Virga, Nidoviruses and some negative sense viruses as group of interest and assembled longer contigs based on those groups.

19.

Here you see the distribution of E-values for the assembled contigs from VirusGatherer. I assembled over 8000 contigs from aligning to over 120 families. We can also see that the majority of those contigs – 6000 – returned alignments with significant E-values.

And if wee look at the phyla, we can see that 3 phyla dominate here, with a clear majority – 2553 aligning to Pisuviricota with 86 % being significant.

20.

We can see here that 332 datasets returned viral sequences, with 102 finding contigs aligning to Flaviviridae, then Astroviridae, Arteriviridae and Picobirnaviridae. We can see also that we get 144 files with Non-RNA viruses. Those are DNA viruses, which are false positives in our data, as their E-values are significantly lower that the rest of the RNA viruses.

21.

Next, I wanted to see how many novel sequences can be found here. Again I define novel here as having a sequence identity below 90 % in comparison to the closest virus from the viral database which we used for taxonomic annotation. And here I identified a lot of novel sequences, with 80% having a sequence identity below 90 %, with most median values falling between 40 and 80 %.

22.

However the majority of those contigs are short. We don't care only for novelty, we want to have as coding complete virus sequences as possible. The length of RNA viruses differ from 2000 – 40000 nt with most being around 10.000. Here we see that only 1225 contigs are longer than 1000

nt, and only 444 being both longer than 1000 nt and novel with ranges from 1000 to 26.000. Here again we see that the 3 previous phyla dominate the long contigs with at least 69% being long and novel and 90 % of long and novel contigs aligning to those 3 phyla.

23.

And here I took the 10 longest contigs from different phyla and performed functional annotation. We can see that we can find the RNA polymerase, which is good, because we used the RNA polymerase for screening. So we expected that. But we can also see that we that the longest contig is a alphacoronavirus found in bats. We can see that all expected open reading frames and porteins such as the non structural proteins, polymerase and Spike protein are found here. This viral contig is special as it is a novel coronavirus sequence found here in a bat. The third largest was also interesting as we can see the expected nucleotide structure for that family, but also on sequence level I identified the RNA editing site typical of the viral subject the sequence aligned too. So we see that we can find viral sequences via VirusHunterGatherer and we can identify close to coding complete sequences.

24.

I mentioned that 3 phyla are dominating the results. So I was curious and looked at the host distribution and and as you can see here the majority of hosts are farm animals such as cattle and pigs, suggesting that SRA is overrepresented by farm hosts. I checked the data, and the 3 dominant phyla were mostly viral families also typically found in farm animals. The rest are rodent or bats. For example we can see picobrinaviruses all of them being found in pigs even if the subject says dog or chicken here. The same goes for Astroviridae and Flaviviridae. One exception however were the green monkeys. Here I found that most green monkey hosts come from 1 study, which studied porciline reproductive and respiratory syndrome virus 1 and 2, not in the natural host but in cell lines. Another discovery was that a chunk of cattle hosts were not cattle but milk products being studied. So we have here a mix of natural hosts, their by products and cell lines.

25.

That leads to the summary. We found over 6000 RNA viruses in 332 samples, with some false positives. Most contigs were novel but short with only 444 being novel and long. We saw that 3 phyla dominated, which can be assossiated with farm animals as hosts. The functional annotation showed that we can find the RdRp which I used for screening but also characteristic proteins of viral families, and we can find novel viruses such as the alphacoronavirus. Of not here is that we can find a mix of natural hosts, cell lines and by products.

27.

And as a concluion. We analyzed the taubert data and got only a small portion of files returing contigs, which were either not significant in Hunter or really short. Leading to the conclusion that there is no conclusive evidence linking viruses to liver rejection.

For mammals we identified over 6000 contigs with 444 being long and novel sequences. We identified charcteristic proteins in those sequences and made a connection to the hosts being mostly farm animals.

Lastly, I developed a small R package, which allows to generate visual and graphical reports for the VirusHunterGatherer output.

28.

The next steps should be the curation of contigs. I mentioned that we have a mix of natural hosts and cells or by products. Depending on what scientific question we want to answer, some of these contigs might be more useful than other. So we have to curate them, and we have to traceback the files to the original study to find out what was done, how it was done and if the virus is new or not.

Here we also performed only taxonomic annotation. No actual taxonomic classification was performed, which should be done next.

And as a whole. The contigs I assembled are a useful resource for future projects. Chong for example studies the spillover risk of RNA viruses and the assembled contigs could be useful data for that project.

29.

That was the presentation.

I would like to thank both my supervisors and Chong for their time and help. Of course I also want to thank the Taubert research group for providing the data, The computational virology research group and the entire institute and twincore community for taking me in and allowing me to do my thesis here. And lastly the scientific community for uploading their data to SRA, which I then used here for the mammal dataset.

30.

Thanks. Any questions?