Univerzita Karlova
Přírodovědecká fakulta

Studijní program: Speciální chemicko-biologické obory
Studijní obor: Molekulární biologie a biochemie organismů

Kateřina Poláčková

Problematika identifikace neznámých virových genomů ve studiích viromu
Problems with identification of unknown viral genomes in the virome studies

Bakalářská práce

Školitel: Mgr. Lenka Kramná, Ph.D.

Praha, 2019

Prohlášení:

Prohlašuji, že jsem závěrečnou práci zpracovala samostatně a že jsem uvedla všechny použité zdroje a literaturu. Tato práce ani její podstatná část nebyla předlo žena k získání jiného nebo stejného akademického titulu.

V Praze 1. 8. 2019 ……

Kateřina Poláčková

Především bych chtěla poděkovat vedoucí mé práce Mgr. Lence Kramné, Ph.D. za užitečné rady, obrovskou trpělivost všechen čas věnovaný konzultacím, aby mohla vzniknout tato bakalářská práce. Dále poděkování patří i mým rodičům a mému nejbližšímu okolí, kteří mě podporují ve studiu a vždy tady pro mě byli nablízku.

Abstrakt

Identifikace virů je problematická, jelikož nemají žádnou sekvenci v genomu, která by byla pro všechny viry společná. Díky NGS a virové metagenomice byla získána spousta sekvencí, ale ne všechny se dosud podařilo identifikovat. Sekvenování s využitím hybridizačních sond se specializuje na vychytávání virových sekvencí právě pomocí specifických sond připravených podle virových genomů a je velmi sensitivní. Při studii viromů ze vzorku stolice, krve a vzorku z dýchacích cest se ukázalo, že identifikace metagenomických dat je nejnáročnější ze vzorku stolice. Povedlo se ale objevit např. nový pikornavirus, nový lidských papillomavirus nebo crAssfág. K identifikaci a analýze těchto dat napomáhají bioinformatické nástroje jako jsou třeba SURPI, Taxonomer, RIEMS a MG-RAST.

Klíčová slova: virom, identifikace, neznámé sekvence, virový genom, „assembly“

Abstract

A virus identification is problematic, because viruses do not have any sequence in genom, which would be common for all of them. Thanks to NGS viral metagenomic is able to gain lots of sequences, but not all of them have been identified yet. Sequencing using hybridization probes specialises in capture of viral genomes and it is very sensitive. Studying viromes from stool and blood samples and samples from respiratory tract has shown that the hardest identification of metagenomic data is from stool samples. Despite this new viruses has successfully been discovered, for example new picornavirus, new human papillomavirus or crAssphage. Bioinformatic pipelines are helpful for identification and analysis of those data, such as pipelines like SRPI, Taxonomer, RIEMS and MG-RAST.

Key words: virome, identification, unknown sequences, viral genome, assembly

Obsah

Abstrakt V

Obsah VI

Seznam použitých zkratek VII

1. Úvod 1

2. Objevování a charakterizace nových virů 2

3. Sekvenování nové generace 3

4. Sekvenování viromu 4

4.1 Sekvenování pomocí náhodných fragmentů 7

4.2 Sekvenování s využitím hybridizačních sond 9

5. Příklady studií viromu v různých typech vzorků s ohledem na podíl sekvencí neznámého původu 10

5.1 Studie viromu ve vzorku stolice 10

5.2 Studie viromu ve vzorku krve 15

5.3 Studie viromu ve vzorcích získaných z dýchacích cest 18

6. Identifikace nových virových genomů z metagenomických dat 21

6.1 Získání celého virového genomu z metagenomu 21

6.2 Příklady nalezení nového viru z metagenomických dat 23

6.2.1 Nový pikornavirus asociovaný s gastroenteritidou 23

6.2.2 Nový lidský papillomavirus asociovaný s febrilním respiračním onemocněním 24

6.2.3 Objev vysoce abundantního bakteriofága 25

7. Bioinformatické nástroje k identifikaci sekvencí 27

8. Závěr 29

Seznam použité literatury 30

Přílohy 35

Seznam použitých zkratek

AIDS	Acquired immune deficiency syndrome	Syndrom získaného selhání imunity
BAL	Bronchoalveolar lavage	Bronchoalveolární laváž
cDNA	Complementary deoxyribonucleic acid	Komplementární deoxyribonukleová kyselina
CRISPR	Clustered regularly interspaced short palindromic repeats	Segmenty nahromaděných pravidelně rozmístěných krátkých palindromatických repetic
DNA	Deoxyribonucleic acid	Deoxyribonukleová kyselina
DNase	Deoxyribonuclease	Deoxyribonukleáza
dsDNA	Double-stranded deoxyribonucleic acid	Dvouvláknová deoxyribonukleová kyselina
dsRNA	Double-stranded ribonucleic acid	Dvouvláknová ribonukleová kyselina
HBoV	Human bocavirus	Lidský bocavirus
HIV	Human immunodeficiency virus	Virus lidské imunitní nedostatečnosti
HPV	Human papillomavirus	Lidský papillomavirus
HRSV	Human respiratory syncytial virus	Lidský respirační syncyciální virus
HRV	Human rhinovirus	Lidský rhinovirus
HTLV	Human T-cell leukemia virus	Lidský T-buněčný lymfotropní virus
IVA	Iterative virus assembler	Programpro skládání virových genomů založený na opakovaném skládání
MCPyV	Merkel cell polyomavirus	Lidský polyomavirus Merkelových buněk
MG-RAST	Metagenomic rapid annotation using subsystems technology	Rychlá metagenomická anotace využívající technologii podsystémů
mRNA	Messenger ribonucleic acid	Mediátorová ribonukleová kyselina
NGS	Next generation sequencing	Sekvenování nové generace
OLC	Overlap-layout-consensus	Konsensus vzniklý dlouhými překryvy
ORFs	Open reading frames	Otevřené čtecí rámce
PBS	Phosphate buffered saline	Solný roztok pufrovaný fosfátem
PCR	Polymerase chain reaction	Polymerázová řetězová reakce
PRICE	Pair-read iterative contig extension	Opětovné prodloužení kontigů párových čtení

rDNA	Ribosomal deoxyribonucleic acid	Ribosomální deoxyribonukleová kyselina
RIEMS	Reliable information extraction from metagenomic sequence dataset	Spolehlivé získání informací z metagenomického sekvenačního datasetu
RNA	Ribonucleic acid	Ribonukleová kyselina
RNase A	Ribonuclease A	Ribonukleáza A
rRNA	Ribosomal ribonucleic acid	Ribosomální ribonukleová kyselina
RT-PCR	Reverse transcription polymerase chain reaction	Polymerázová řetězová reakce spojená s reverzní transkripcí
ssDNA	Single-stranded deoxyribonucleic acid	Jednovláknová deoxyribonukleová kyselina
ssRNA	Single-stranded ribonucleic acid	Jednovláknová ribonukleová kyselina
SURPI	Sequence-based ultrarapid pathogen identification	Velmi rychlá identifikace patogenů na základě sekvencí

1. Úvod

Viry jsou nejhojnějšími entitami na Zemi, v lidském těle se jich vyskytuje až 100násobně více než je lidských buněk (kterých je přibližně 10^{13}). Předpokládá se, že bylo dosud prozkoumáno méně než 1 % všech virů (Mokili, Rohwer, et Dutilh 2012). Historie virologie není tak dlouhá, i když virové onemocnění byly objeveny dávno před samotnými viry (Rascovan, Duraisamy, et Desnues 2016a). Viry hrají velmi důležitou roli v lidském i zvířecím zdraví i v enviromentální ekologii.

Tradiční metody pro detekci virů jako pěstování na buněčné kultuře, elektronová mikroskopie nebo sérologie nejsou vždy dostačující, a tak bylo zapotřebí vyvinout další metody. S rozvojem sekvenování nové generace (NGS) bylo umožněno získat i sekvence zatím neznámých virů, první virová metagenomická studie s lidskými stolicemi byla provedena skupinou Breitbart et al. (2003). Předtím bylo možné sekvenovat viry pomocí Sangerova sekvenování (Kumar, Murthy, et Kapoor 2017).

Cílem této práce bylo shrnout všechny body důležité pro objevení nového viru z metagenomických dat:

- shrnout různé typy sekvenování viromu
- shrnout význam přípravy vzorku na výsledný počet virových čtení
- shrnout bioinformatické strategie a nástroje pro analýzu sekvencí vedoucí k získání nového virového genomu.

2. Objevování a charakterizace nových virů

Prvním objeveným virem byl virus tabákové mozaiky na konci 19. století. Na jeho objevení pracovali nezávisle 3 vědci: Adolf Mayer, Dmitrij Ivanovský a Martinus Beijerinck. V roce 1892 Dmitrij Ivanovský prokázal, že přefiltrovaná šťáva z nemocných rostlin, která by měla být zbavená bakterií, je stále infekční pro zdravé rostliny a v roce 1903 publikoval studii, ve které popsal abnormální intracelulární inkluzní tělíska v hostitelských buňkách nakažených rostlin (Lustig et Levine 1992). O několik let později nizozemský vědec Martinus Beijerinck označil jako původce virus. V roce 1935 dokázal Wendell Meredith Stanley čistý virus vykrystalizovat (Artenstein 2012). K výzkumům virových onemocnění také sloužily zvířecí modely, například při pokusech Louise Pasteura s virem vztekliny na konci 19. století, kdy inokuloval virus přímo do králičího mozku (Fields et Knipe 1991), nicméně až v roce 1940 byly poprvé pozorovány virové částice v elektronovém mikroskopu (H.-W. Ackermann 2011).

Postupem času došlo k rozvoji metod, které umožňují nalezené viry blíže charakterizovat, jakou jsou metody:

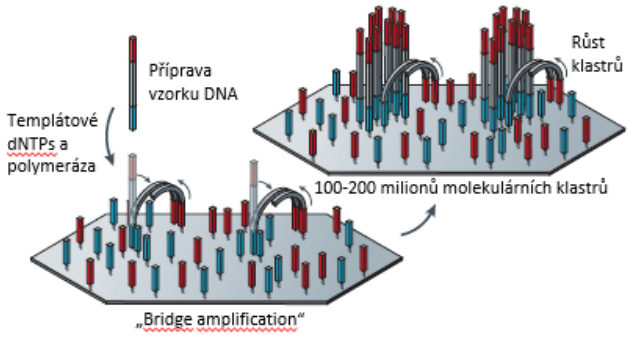
- **sérologické**: umožňují sledovat výskyt protilátek v séru pacienta proti známému, např. virovému antigenu, nebo naopak průkaz výskytu virového antigenu pomocí setu protilátek. Primárním vzorkem je sérum, ale lze použít také vzorky moče, slin, likvoru a jiných (Šmarda et al. 2010).
- **molekulárně-genetické metody:** z těch se ve virologii hojně používá polymerázová řetězová reakce (PCR). Principem je detekovat známý úsek nukleové kyseliny pomocí páru specifických primerů. Tímto způsobem si lze vzorek namnožit a použít k dalším analýzám a reakcím, např. sekvenování nebo detekovat jeho kvantitu v původním vzorku (Waters et Shapter 2014). Použití těchto metod je však vázáno na již známý virus a používají se hlavně v jeho diagnostice a jsou vysoce sensitivní.
- **pěstování viru na buněčných kulturách**: tato metoda je problematická, protože ne všechny viry se v nich množí (Tang et Chiu 2010). Využívá se také k jejich průkazu a množení, ale zejména ke studiu virulence, např. ve studii Anderson et al. (2012).

Až rozvoj NGS a s ním i bakteriální a virové metagenomiky umožnil získat sekvence nových, dosud neznámých virů nebo sekvenčně velmi vzdálených variant známých virů (Rascovan, Duraisamy, et Desnues 2016a). Problémem však zůstává jejich identifikace.

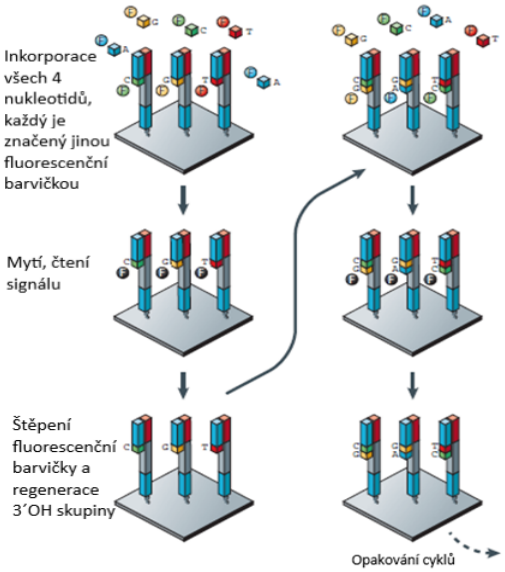
3. Sekvenování nové generace

Sekvenování nové generace umožňuje paralelní sekvenaci velkého množství sekvencí najednou na rozdíl od Sangerovo sekvenování, kde se sekvenuje pouze jeden PCR produkt v reakci. Existuje množství sekvenačních platforem, např. Roche (www.roche.com), Oxford Nanopore (www.nanoporetech.com) a Illumina (www.illumina.com), a každá nabízí množství způsobů přípravy knihovny k sekvenování podle typu výchozího vzorku. Ve všech případech je však nutné výchozí nukleovou kyselinu nafragmentovat a k fragmentům musí být připojeny indexy, které jednotlivé vzorky označí, jako je shrnuto v Metzker (2010).

Dnešní nejčastěji používanou sekvenační technikou je sekvenování na platformě Illumina s fluorescenčně značenými nukleotidy. Tato technika využívá k pomnožení fragmentů destičku („flow-cell"), na které se jednotlivé fragmenty přichytí a pomnoží do tzv. klastrů (tzv. „bridge amplification", **Obrázek 1**), čímž se docílí zesílení signálu. Pro vlastní sekvenování jsou využívány fluorescenčně značené nukleotidy (**Obrázek 2**). Prvním krokem bioinformatické analýzy je tudíž spojení těchto polovin do jedné sekvence. Kvalita sekvencí získaných z platformy Illumina je nižší, což ale vyrovnává množství získaných čtení (až 25 milionů při sekvenování na přístroji MiSeq).



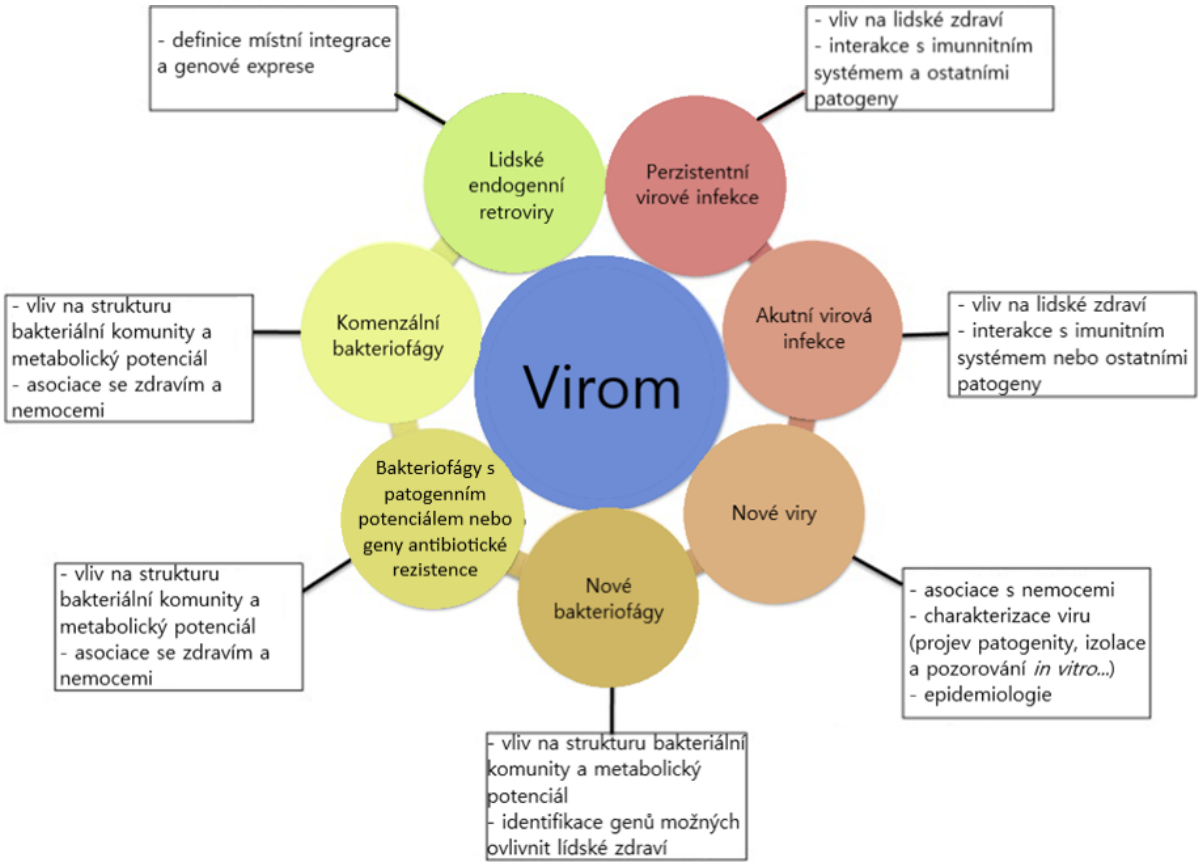
Obrázek 1: Přichycení fragmentů k destičce a jejich následující amplifikace do klastrů, upraveno podle Metzker (2010)



Obrázek 2: Vlastní sekvenování (cyklická reversibilní terminace), upraveno podle Metzker (2010)

4. Sekvenování viromu

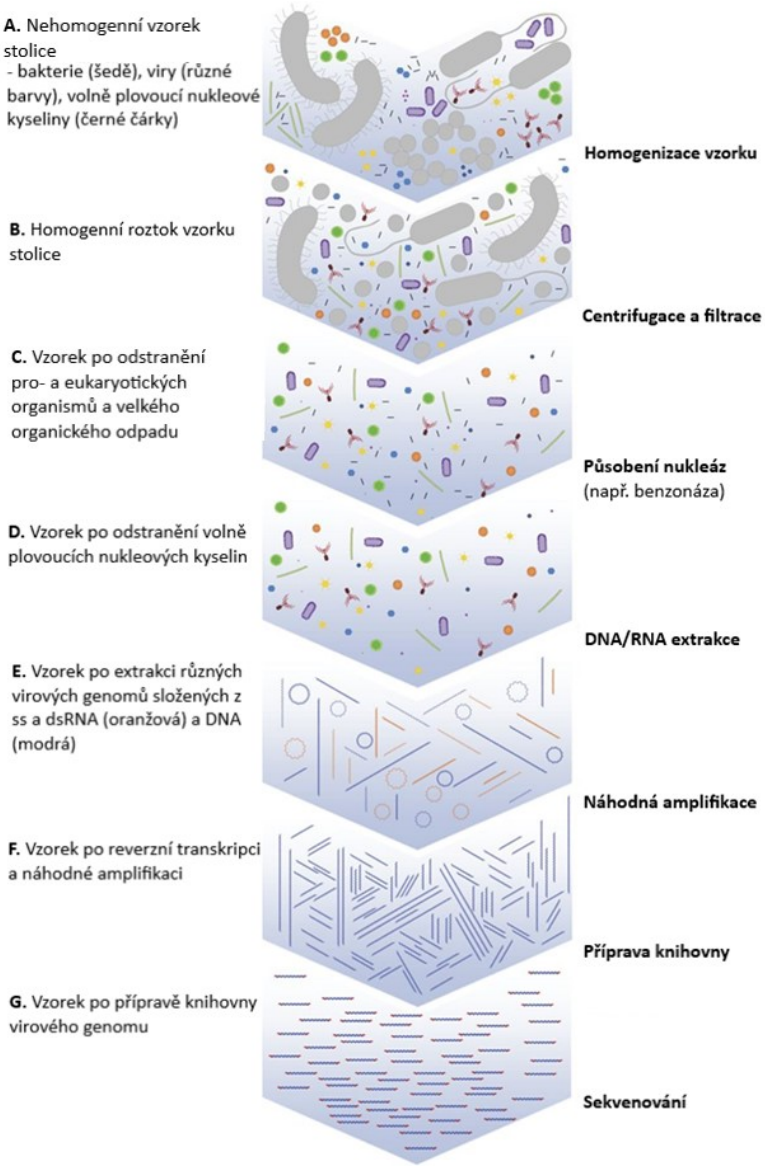
Virom je soubor všech virů, které se nachází v biologickém vzorku a je součástí celého spektra organismů vyskytujících se ve vzorku jako jsou bakterie (bakteriom/mikrobiom) nebo houby (mycobiom). Zahrnuje jak prokaryotické viry (viry bakterií nebo–li bakteriofágy), tak eukaryotické viry jako jsou např. zvířecí a lidské viry (**Obrázek 3**). Soubor genů všech organismů ve vzorku se nazývá metagenom, sekvenování náhodných fragmentů nukleových kyselin pak metagenomické sekvenování.



Obrázek 3: Složky lidského viromu, kruhy představují složky viromu, které mohou být charakterizovány metagenomickým sekvenováním, upraveno podle Wylie, Weinstock, a Storch (2012)

Sekvenová viromu bylo umožněno až s rozvojem sekvenování nové generace. Podle zvolené sekvenační techniky je možné osekvenovat až miliony čtení najednou (Kumar, Murthy, et Kapoor 2017), díky čemuž tak získáme nejen informaci o výskytu virů v daném vzorku, ale i přibližné poměry jejich kvantit, viz. studie Holtz et al. (2014); Moustafa et al. (2017); Yang et al. (2011). Sekvenování viromu je technicky obtížné. Jelikož viry nemají žádnou sekvenci podobnou například genu pro bakteriální 16S ribosomální ribonukleové kyseliny (rRNA), která

by byla přítomna ve všech virech a byla by natolik variabilní, abychom podle ní mohli rozlišovat jednotlivé druhy, musí tak být osekvenován celý virový genom nebo jeho fragmenty, říkáme tedy, že technika je sekvenčně nezávislá (Kumar, Murthy, et Kapoor 2017). Přípravu náhodných fragmentů virového genomu ze vzorku stolice shrnuje **obrázek 4** zabývající se separací virových částic od zbytku vzorku (bakterií, lidských buněk, případně rostlinných zbytků) a následně vznik náhodných fragmentů vyizolovaných nukleových kyselin. V praxi je však toto velmi problematické a výsledné sekvence obsahují množství kontaminujících sekvencí ze zbytku vzorku pocházejících z rozpadlých buněk, které je těžké odseparovat. Separaci virů pomocí gradientu (Zwartouw, Westwood, et Appleyard 1962) také nelze použít, když se snažíme získat viry o různých velikostech, a ještě k tomu se někdy nacházejí ve vzorku v nízkých kvantitách. Proto jsou často ve výsledných sekvencích detekovány velké množství sekvencí nevirového původu (Mokili, Rohwer, et Dutilh 2012; Kumar, Murthy, et Kapoor 2017; Kramná et al. 2015; Lysholm et al. 2012; Popgeorgiev et al. 2013). Na druhou stranu je výhodou, že získáme sekvence neznámých a nepopsaných virů.



Obrázek 4: Schématický stručný popis techniky obohacující virom, upraveno podle (Conceição-Neto et al. 2015); Separace virových částic od zbytků vzorku nejprve začíná homogenizací vzorku, po které jsou centrifugací a filtrací odstraněny např. organické velké odpady, následně působením nukleáz jsou odstraněny volné nukleové kyseliny. Po extrakci různých virových genomů jsou dále náhodně amplifikovány a z těchto náhodných fragmentů je vytvořena knihovna virového genomu, která je nakonec sekvenována.

I přes komplikace však lze touto technikou sledovat komplexní vzorek, detekovat všechny viry, které se ve vzorku nachází, získat jejich sekvence bez nutnosti je amplifikovat na buněčných kulturách. Velký potenciál má metoda při objevování nových virů: např. ve skupině Mokili et al. (2013) se podařilo identifikovat nový lidský papillomavirus při analýze nasálních/oropharyngeálních výtěrů od pacientů s respiračním onemocněním za účelem vyvinout standardní postup pro objevování virů pomocí metagenomiky. Také skupina Popgeorgiev et al. (2013), která se zaměřila na studium virů řádu *Megavirales*, které při analýze

po filtraci vzorku přes 0,22μm filtr mohly být odfiltrovány, popsala nový virus z čeledi *Marseilleviridae* (řád *Megavirales*) nalezený v krvi od zdravých dárců. Metoda byl využita při snaze popsat rozdíly mezi viromem u zdravého a nemocného jedince (např. plazma pacientu s virem lidské imunitní nedostatečnosti (HIV)/syndromem získaného selháním imunity (AIDS) obsahuje velké množství bakteriálních virů, neboli bakteriofágů, endogenní retroviry a také byl nalezen virus běžně se nevyskytující u člověka, zatímco v plazmě u zdravého jedince byly nalezeny pouze některé druhy anellovirů a nebyly nalezeny žádné bakteriofágy (Li et al. 2012)) nebo nově asociovat nalezený virus s určitou nemocí, jako třeba zvýšený výskyt viru hepatitidy C (HCV) u pacientů s diabetem 1. typu (Gebrekristos et al. 2018).

4.1 Sekvenování pomocí náhodných fragmentů

Proces přípravy vzorku a vznik náhodných fragmentů již byl znázorněn na **obrázku 2** v předchozí kapitole. Jednotlivé kroky, které následují po izolaci celkové deoxyribonukleové kyseliny (DNA)/ribonukleové kyseliny (RNA) jako je použili v publikacích Blinkova et al. (2009, 2010) zahrnují:

- **přepis RNA** pomocí náhodných primerů a reverzní transkriptázy do komplementární DNA (cDNA). Pro DNA viry využívá zbytkové aktivity reverzní transkriptázy k DNA. Náhodné primery jsou opatřeny sekvencí nazvanou „tag“, což je vždy stejná sekvence použitá v posledním kroku k amplifikaci těchto fragmentů.

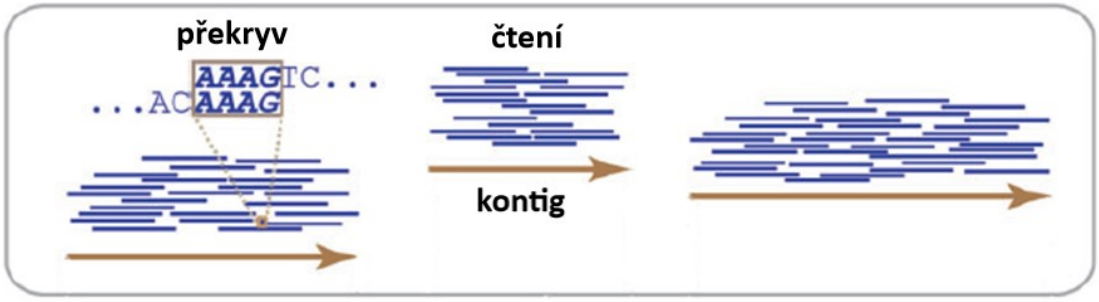
- **syntéza druhého řetězce:** v publikaci Blinkova et al. (2009) k tomu použili Klenowův fragment (fragment DNA polymerázy I *Escherichia coli*, který má 5´-3´ polymerační aktivitu, 3´-5´exonukleázovou aktivitu, ale postrádá 5´-3´exonukleázovou aktivitu (Šmarda et al. 2010)) a náhodné primery (z předchozí reakce).

- **amplifikace náhodných fragmentů:** využívají se připojené sekvence „tagů“, náhodné fragmenty jsou amplifikovány pomocí PCR. Častým problémem sekvenování virů je jejich nízká kvantita nutná pro přípravu knihovny a k NGS. Tento krok překonává tento problém.

- **příprava knihovny podle zvolené sekvenační platformy:** např. NexteraXT Library Preparation kit od firmy Illumina, kdy jsou k nim připojeny indexy, aby byly jednotlivé

vzorky v sekvenačním běhu od sebe rozlišeny a adaptéry. Takto připravená knihovna je poté sekvenována.

Po sekvenaci jsou jednotlivé čtení bioinformaticky *de novo* spojovány do kontigů (**Obrázek 5**), které jsou porovnávány s databázemi virových sekvencí, nejčastěji se používá databáze NCBI (www.ncbi.nlm.nih.gov). Sekvence, které nemají žádnou shodu s databází, mohou být nevirového původu, což lze detekovat rozšířením výběru referenčních sekvencí na všechny sekvence v databázi. I přes to se ve vzorcích nachází množství sekvencí neznámého původu (Kramná et al. 2015; Popgeorgiev et al. 2013; Finkbeiner et al. 2008), mezi kterými by mohly být i nové, dosud neobjevené viry.



Obrázek 5: *De novo* „assembly" jednotlivých náhodných čtení do větších celků – kontigů. Sekvence jsou skládány do větších celků pouze na základě identifikace překryvů – shodných míst mezi sebou, jak charakterizuje část „překryv", upraveno podle Sohn a Nam (2016)

4.2 Sekvenování s využitím hybridizačních sond

Princip této metody je založen na vychytávání virových sekvencí ze vzorku pomocí specifických oligonukleotidů (sond). Tento způsob navrhla např. skupina Briese et al. (2015). Vychytávány jsou všechny viry obratlovců, ale už ne bakteriofágy, rostlinné viry nebo viry hub. Sondy byly vytvořeny tak, aby pokrývaly celé genomy virů – stažením všech vybraných virových genomů, klastrováním sekvencí na základě podobnosti a rozštěpením na krátké úseky = sondy. Vysoká senzitivita této metody je založena na 2 milionech sond, které mohou rozeznat genomy 207 známých virových taxonů v různých typech vzorků. Sondy se liší alespoň z 10 % a zachytí sekvence s až 40 % rozdílností, které by mohly být zmutované nebo jinak odlišné (Briese et al. 2015).

Výhodou použití sond pro sekvenování viromu je, že jsou selektivně a přednostně ze vzorků vychytány hlavně virové sekvence. To je výhodou především u vzorků, kde je velká kontaminace jinou DNA/RNA, jako je třeba vzorek tkáně či krve. V takovýchto vzorcích kontaminující nukleové kyseliny natolik převyšují nukleové kyseliny viru, že je problematické zachytit virus ve výsledných sekvencích i při jeho vysoké kvantitě. Briese et al. (2015) uvádí, že s použitím sond můžeme získat 100-10 000x více virových sekvencí než při použití běžného metagenomického sekvenování. Zvyšuje tedy sensitivitu, ale detekuje jen to, s čím sonda hybridizuje, proto to není nejlepší způsob k objevování nových virů, pokud nedetekuje nějaký konzervovaný virový gen.

5. Příklady studií viromu v různých typech vzorků s ohledem na podíl sekvencí neznámého původu

5.1 Studie viromu ve vzorku stolice

Stolice zdravého jedince se skládá převážně ze zhruba 75 % vody a 25 % pevného materiálu (Rose et al. 2015) a obsahuje přibližně 10^{11} buněk/g (Aggarwala, Liang, et Bushman 2017). Ve studii Stephen a Cummings (1980) zjistili, že pevné látky jsou až z 55 % tvořeny bakteriemi. Mezi nejhojnější virové entity nalezené ve stolici patří bakteriofágy, dále se objevují i RNA viry rostlin (získané z potravy) a také eukaryotické (lidské) viry, jejichž diverzita na rozdíl od bakteriofágů v průběhu života jedince klesá (Zárate et al. 2017). Např. ve stolici zdravého muže (Minot et al. 2013) byly po metagenomické analýze nalezeny pouze bakteriofágy a žádné známé viry infikující eukaryotické buňky nebyly detekovány. Ani v některých dalších studiích střevního viromu u zdravých jedinců (Minot et al. 2011; Reyes et al. 2010; Breitbart et al. 2003) se lidské viry nenašly. Nemusí to ale znamenat, že se eukaryotické (lidské) viry u zdravých jedinců nevyskytují, stolice zdravého jedince může obsahovat viry s asymptomatickým průběhem infekce, jako tomu bylo ve studii Kramná et al. (2015), kde byly nalezeny u zdravých dětí např. viry čeledi *Picornaviridae* (enterovirus a parechovirus), které často způsobují infekce u dětí (de Crom et al. 2016a) nebo viry z čeledi *Anelloviridae*, viry se zatím neznámým vztahem k onemocnění (**tabulka 1**).

Nejčastejší bakteriofágové čeledi nalezené ve vzorcích stolice (Minot et al. 2011) jsou:

- *Microviridae* – malé, neobalené fágy s cirkulární jednovláknovým (ssDNA) genomem s T-1 ikosahedrální symetrií bez bičíku. Hostiteli jsou enterobakterie, některé intracelulární bakterie a spiroplasma.
- *Podoviridae, Myoviridae, Siphoviridae* – všechny patří do řádu *Caudovirales* a tvoří nejpočetnější skupiny fágů na světě (H. W. Ackermann, s.d.). Všechny mají bičík a lineární dvouvláknový DNA (dsDNA) genom, hostitelem jsou bakterie a u čeledi *Myoviridae* také archea. Ve střevě jsou to např. *Escherichia coli* nebo salmonely.

Na rozdíl od studií viromu ze stolice zdravých jedinců bylo ve studii Finkbeiner et al. (2008), která metagenomicky analyzovala výskyt virů u dětí s akutním průjmem, nalezeno dokonce 36 % sekvencí virového původu a některé vzorky obsahovaly více než 90 % virových

sekvencí. Tohoto množství bylo dosaženo i přesto, že bylo použito limitované sekvenování (pouze 384 čtení na vzorek) a ve většině případů se podařilo virus zachytit. Ve stolici dětí, která žádnou akutní infekcí neprocházela, bylo z celkového počtu sekvencí pouze 2 % virových a většina z nich patřila bakteriofágům (Kramná et al. 2015). Přítomnost virové složky ve stolici je tedy velmi variabilní (na rozdíl od bakterií, které se jako komenzálové vyskytují ve zdravé stolici). Objev nového lidského viru tedy usnadní výběr vzorku s akutním průběhem infekce, pokud tento virus infekci způsobuje.

Eukaryotní viry, nalezené ve všech zmíněných studiích viromu ze vzorku stolice (Finkbeiner et al. 2008; Kramná et al. 2015; Holtz et al. 2014) byly:

Viry způsobující průjmová onemocnění:

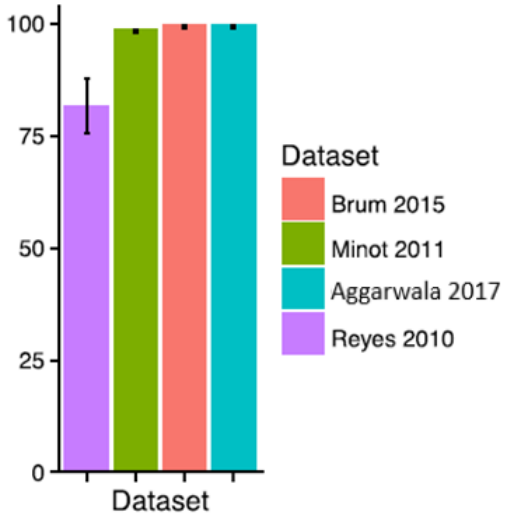
- *Caliciviridae* – neobalené viry s lineární jednovláknovým RNA (ssRNA) (+) genomem. Běžně způsobují gastroenteritidy.
- *Reoviridae* – neobalené viry se segmentovaným genomem s lineární dvouvláknovou RNA (dsRNA), nalezeny také ve studiích (Yinda et al. 2019; Siqueira et al. 2018). Kromě průjmových onemocnění mohou také např. způsobovat onemocnění dýchacích cest.
- *Astroviridae* – neobalené viry s lineárním ssRNA(+) genomem, nalezeny také ve studiích (Yinda et al. 2019; Siqueira et al. 2018). Jsou příčinou gastroenteritidy, hlavně u dětí a starších dospělých, kteří mají oslabený imunitní systém.
- *Adenoviridae* – neobalené viry s lineární dsDNA genomem. Způsobují hlavně infekce horních dýchacích cest, ale mohou být také příčinou gastroenteritidy.
- *Picornaviridae* – neobalené viry s lineárním ssRNA(+) genomem. Způsobují gastroenteritidy, ale také např. i meningitidu, encefalitidu, hepatitidu, infekci horních dýchacích cest atd. Nejčastěji byly nalezeny enteroviry a parechoviry (Finkbeiner et al. 2008; Kramná et al. 2015; Holtz et al. 2014), které jsou hlavními příčinami infekcí u dětí a mají různé projevy od mírných gastrointestinálních infekcí až po vážná onemocnění jako meningitida nebo sepse vedoucí k smrti (de Crom et al. 2016b). Holtz et al. (2014) se podařilo objevit virus, který nebyl v databázi detekován, ale sdílel největší sekvenční podobnost s viry z řádu *Picornavirales.*
- *Parvoviridae* – neobalené viry s lineárním ssDNA genomem. Lidské bocaviry infikují také gastrointestinální trakt, ale gastroenteritidy způsobují jen zřídka (Finkbeiner et al.

2008). V malém množství byl z této čeledi nalezen bocavirus a „adeno-associated“ virus (AAV).

Viry bez známé asociace s gastroenteritidou:

- *Anelloviridae* – neobalené viry s cirkulárním ssDNA(-) genomem. Mají vysokou prevalenci, ale nebyly dosud spojeni s jakoukoliv nemocí. Ve studii Holtz et al. (2014) byla čeleď *Anelloviridae* nejčastěji detekovanou čeledí, a to ve 37 vzorcích z 87, a nechyběla také u ostatních skupin zabývajících se střevním viromem (Yinda et al. 2019; Siqueira et al. 2018).
- *Circoviridae* – neobalené viry s cirkulárním ssDNA genomem. Tato čeleď způsobuje onemocnění hlavně u ptáků a prasat, u lidí je možné, že infikují gastrointestinální trakt (Linlin Li et al. 2010).
- *Orthomyxoviridae* – obalené viry se segmentovaným genomem s lineární ssRNA(-). Není běžné, aby ve vzorcích stolice byly nalezeny viry této čeledi, ale ve dvou studiích se podařilo je detekovat (Holtz et al. 2014; Wootton et al. 2006). Jsou dvě teorie, jak se mohly tyto viry v gastrointestinálním traktu vyskytnout:
    1. virus mohl být produkován uvnitř gastrointestinálním traktu v průběhu časné fáze onemocnění,
    2. méně pravděpodobnou teorií je, že byl virus spolknut s dýchacími sekrety (Wootton et al. 2006).

Studie viromu tedy obsahují různé množství sekvencí nevirového původu. Tyto sekvence mohou být dále identifikovány jako lidská či bakteriální DNA/RNA, nebo nukleové kyseliny hub, jak shrnuje **tabulka č. 1**. Existuje však množství sekvencí neznámého původu, jinak nazývané jako „dark matter“ (Aggarwala, Liang, et Bushman 2017) (**Tabulka 1**). Například ve studii Kramná et al. (2015) nebyli schopni identifikovat až 74 % sekvencí, tzn. sekvence neměly žádnou shodu s databází, se kterou byly porovnávány. Je tedy možné, že vzorky obsahují daleko větší množství virů, jen ještě nebyly objeveny. Procenta sekvencí neznámého původu v různých studiích shrnuje také Aggarwala, Liang, a Bushman (2017) (**Obrázek 6**), kde shrnují výsledky své studie s dalšími studiemi viromu stolice (Reyes et al. 2010; Minot et al. 2011; Brum et al. 2015).



Obrázek 6: Procenta sekvencí neznámého původu v různých studiích viromu stolice (Aggarwala, Liang, et Bushman 2017) – tzv. virová temná hmota („dark matter“), shrnuto ve studii Aggarwala, Liang, a Bushman (2017) (shrnují výsledky ze studií (Minot et al. 2011; Reyes et al. 2010; Brum et al. 2015) a vlastní data)

Tabulka 1: Porovnání tří studií viromu stolice (Finkbeiner et al. 2008; Kramná et al. 2015; Holtz et al. 2014)

typ vzorku	charakteristika kohorty	celkový počet sekvencí	počet virových čtení	celkový počet nalezených virových druhů/rodin	% virových čtení (z toho bakteriofágů)	% nevirových sekvencí			sekvence neznámého původu (nebyl nalezen protějšek v databázi použité k identifikaci)	nejvyskytovanější virové čeledi
						bakterie	lidské	ostatní (houby, rostliny, savčí sekvence, atd..)		
stolice	96 vzorků (od 19 dětí mladších 2 let) [1]	703 242 (průměrný počet čtení/vzorek)	N	N	2 % (většina patřící bakteriofágům)	14%	2%	8%	**74%**	*Picornaviridae*
										Anelloviridae
	87 dětí s průjmem, 5 let a méně [2]	1 052 882	11232	22 rodin	6,9% (4,4% bakteriofágů)	75,8%	7,0%	10,3%	**16,4%**	*Reoviridae*
										Adenoviridae
										Astroviridae
										Circoviridae
	12 dětí, 4-35 měsíců s akutním průjmem [3]	4608 (limitované sekvenování - pouze 384 čtení na vzorek)	519	N	35,6% (1,7% bakteriofágů)	58,8%	1,4%	2,5%	**28%**	*Orthomyxoviridae*
										Caliciviridae
										Parvoviridae

[1] Kramná et al. (2015) – Z celkového počtu sekvencí byly 2 % identifikovány jako virové, přičemž byly nalezeny viry těchto čeledí: *Anelloviridae, Picornaviridae, Parvovirida* a *Caliciviridae.* 74 % sekvencí nemělo žádnou shodu se známou sekvencí v databázi. (N = neuvedeno);
[2] Holtz et al. (2014) – Z klasifikovatelných sekvencí (162 784) identifikovali 6,9 % jako virových (2,5 % patřily eukaryotickým virům a 4,4 % bakteriofágům). Mezi nalezenými virovými sekvencemi byly tyto čeledi: *Reoviridae, Caliciviridae, Adenoviridae, Astroviridae, Anelloviridae, Picornaviridae, Circoviridae* a *Orthomyxoviridae*. 16,4 % sekvencí nebylo identifikováno.
[3] Finkbeiner et al. (2008) – Téměř ve všech 12-ti vzorcích byly nalezeny virové sekvence (vz.č.1: 65 %, vz.č.2: 46 %, vz.č.3: 2,8 %, vz.č.4: 23 %, vz.č.5: 95 %, vz.č.6: 40 %, vz.č.7: 0 %, vz.č.8: 4,8 %, 0,34 %, vz.č.9: 0,68 %, vz.č.11: 100 %, vz.č.12: 3,1 %) a identifikovali tyto čeledi: *Adenoviridae, Reoviridae, Caliciviridae, Parvoviridae* a *Astroviridae*. U 28 % sekvencí nebyla nalezena shoda se známými sekvencemi. (N = neuvedeno)

5.2 Studie viromu ve vzorku krve

Krev je bohatá na lidské krevní buňky, které v ní tvoří 45 % (hlavně erytrocyty, dále leukocyty a trombocyty), a zbylých 55 % doplňuje krevní plazma. Krev zdravého jedince byla často považována za sterilní, ale metagenomické studie prokázali, že se v krvi skrývají značné virové komunity (Rascovan, Duraisamy, et Desnues 2016a; Popgeorgiev et al. 2013; Moustafa et al. 2017) (**Tabulka 2**).

Ve studii Moustafa et al. (2017), kteří se rozhodli charakterizovat virom krve od 8240 zdravých jednotlivců za účelem zjištění virů u zdravých jedinců s ohledem na darování krve, nalezli nějaký virus u 42 % z nich. Celkem bylo identifikováno 94 různých virů, z toho 19 lidských DNA virů, provirů i RNA viry (herpesviry, anelloviry, papillomaviry, 3 různé polyomaviry, adenovirus, HIV, lidský T-buněčný lymfotropní virus (HTLV), virus hepatitidy B, C, parvovirus B19 a virus chřipky). Zdaleka nejčastějšími viry byly ty z čeledi *Herpesviridae* a *Anelloviridae,* jejich přítomnost v krvi u lidí bez zjevného onemocnění je častá (Moustafa et al. 2017; Segura-Wang et al. 2018). Viry velmi rizikové pro použití krve k transplantaci se vyskytovali spíše ojediněle (nejčastější byl Lidský polyomavirus Merkelových buněk (MCPyV) u 49 jedinců). Sekvence viru chřipky pocházely z DNA vakcíny a hepatitida C se zřejmě vyskytovala v integrované formě. Převážné množství sekvencí v těchto vzorcích krve tvořily lidské buňky a pouze 0,01 % byly viry a 0,2 % bakterie (**Tabulka 2**). Byla také nalezena fágová DNA, která ale byla asociována s fágem phiX174, který se používá ke zvýšení kvality sekvenačního čtení (Moustafa et al. 2017).

Naopak ve studii Popgeorgiev et al. (2013) měly virové sekvence zastoupení 67,7 %, protože získanou krev nejdříve zcentrifugovali a pro svůj výzkum dále použili jen krevní plazmu, která byla už zbavena lidských buněk. Těchto 67,7 % obsahovalo zejména sekvence virů z čeledi *Anelloviridae*, viry s cirkulárním ssDNA genomem, které jsou nalézány v různých typech tkání, zatím s nejasným významem pro lidské zdraví. Další nejčastější sekvence patřily viru z čeledi *Marseilleviridae* který patří do tzv. velkých virů („giant viruses“) a v sekvencích se často nevyskytuje, protože může být zachycen 0,22μm filtry k odstranění bakterií při zpracování vzorku. Autoři poprvé nalezli virus z této čeledi v lidské krvi.

Ve studii Segura-Wang et al. (2018) analyzovali krevní plazmu od jedinců podstupujících transplantaci plic. Imunosuprimovaní pacienti často podléhají virovým infekcím, které se rozvinou z perzistujících virů (zejména herpesviry). U všech jedinců (pacientů i zdravých

kontrol) byly nalezeny viry čeledí *Herpesviridae, Anelloviridae*, ale u imunosuprimovaných pacientů v mnohem větší kvantitě (Segura-Wang et al. 2018). Zajímavé je, že také kvantita anellovirů se také s imunosupresí zvýšila, tudíž by mohla být v budoucnu využívána jako biomarker imunokompetence. Ve vzorcích imunosuprimovaných pacientů byly nalezeny i další viry, nejčastější byla čeleď *Coronaviridae.* Kromě eukaryotických virů byly v této studii nalezeny také bakteriofágy z čeledi *Myoviridae* a *Siphoviridae.*

Další virové čeledi, které se mohou nacházet u asymptomatických jedinců jsou čeledi *Picornaviridae, Poxviridae, Flaviviridae* a *Phycodnaviridae* (shrnuto v Rascovan, Duraisamy, et Desnues (2016b)). A konečně existuje také množství endogenních retrovirů, integrovaných v lidském genomu jako provirus.

Obecná charakteristika nejčastějších nalezených virů z krve u jedinců bez zjevného onemocnění:

- *Herpesviridae* – neobalené viry s lineárním dsDNA genomem. Nejčastější nemocí je opar a jsou také původci onemocnění jako je třeba mononukleóza.
- *Anelloviridae* – viz. kapitola Studie viromu ze vzorku stolice

Hledání nových virů ze vzorků krve komplikuje velké množství lidské DNA z rozpadlých krevních buněk po zamražení a rozmražení. Pokud však máme možnost získat krevní sérum nebo plasmu z nezamražené krve, podíl virových čtení se mnohonásobně zvýší. Množství sekvencí neznámého původu bylo uvedeno pouze ve studii Popgeorgiev et al. (2013), kterým se ze vzorku krevní plasmy nepodařilo identifikovat 14,7 % sekvencí.

Tabulka 2: Porovnání tří studií viromu krve (Moustafa et al. 2017; Popgeorgiev et al. 2013; Segura-Wang et al. 2018)

typ vzorku	charakteristika kohorty	celkový počet sekvencí	počet virových čtení	celkový počet nalezených virových druhů/rodin	% virových čtení (z toho bakteriofágů)	% nevirových sekvencí			sekvence neznámého původu (nebyl nalezen protějšek v databázi použité k identifikaci)	nejvyskytovanější virové čeledi
						bakterie	lidské	ostatní (houby, rostliny, savčí sekvence, atd..)		
krevní plazma	20 zdravých dárců (1)	20 238	13701	N	67,70%	9,50%	N	8%	**14,70%**	*Anelloviridae* *Circoviridae* *Marseilleviridae*
	15 pacientů podstupující transplantaci plic (2)	703 242 (průměrný počet čtení/vzorek)	N	N	3,11%	10,80%	37,20%	0,36% (houby)	**N**	*Polyomaviridae* *Herpesviridae* *Coronaviridae*
krev	8 240 zdravých jedinců, věk od 2 měsíců do 102 let (3)	10^{15}	N	94 různých virů	0,01%	0,20%	95%	N	**N**	*Parvoviridae* *Flaviviridae* *Papillomaviridae* *Adenoviridae*

(1) Popgeorgiev et al. (2013) – Z krevní plazmy od 20 zdravých jedinců bylo získáno po sekvenování 20 238 čtení, ze kterých 14,7 % nebylo indentifikováno. 13 701 (67,7 %) čtení našlo shodu s virovými sekvencemi v databázi z následujících čeledí: *Anelloviridae, Circoviridae* a *Marseileviridae,* a byly také nalezeny sekvence fága λ, jehož hostitelem jsou enterobakterie. (N = neuvedeno)

(2) Segura-Wang et al. (2018) – celkový počet čtení na vzorek činil 703 242, ze kterých po kontrole kvality a filtraci zbylo 39,6 % na další analýzu. Po odstranění lidských, bakteriálních a fungálních čtení byly zbylé čtení porovnávány s virovou databází. Z celkového počtu sekvencí bylo 3,11 % virového původu. Byly nalezeny tyto virové čeledi: *Anelloviridae, Herpesviridae* a *Coronaviridae.* (N = neuvedeno)

(3) Moustafa et al. (2017) – Podařilo se v nich najít 94 různých virů, z nichž nejčastěji zastupovali čeledi jako *Herpesviridae, Polyomaviridae, Parvoviridae, Anelloviridae* a *Papillomaviridae* a v menší míře čeledi *Adenoviridae,*a *Flaviviridae*. (N = neuvedeno)

5.3 Studie viromu ve vzorcích získaných z dýchacích cest

Není jednoduché studovat virom dýchacích cest. Získání vzorku může být někdy problematické, jako např. u dolních dýchacích cest, které nejsou lehce přístupné a někdy je zapotřebí použít invazivní metody pro získání vzorků, jako je např. bronchoalveolární laváž (BAL) nebo-li výplach průdušek a plicních sklípků (Wylie 2017). Naproti tomu stěr z nosní dutiny či z krku nepředstavuje dost reprezentativní vzorek.

Také ve vzorku BALu se vyskytuje množství lidských buněk, proto použili ve studiu Segura-Wang et al. (2018) postup, kde byl vzorek BALu po odebrání zcentrifugován, pelet resuspendován v solném roztoku fosfátem pufrovaným (PBS) a přefiltrován přes 0,45μm filtr. Dále bylo zapotřebí se zbavit volné DNA a RNA pomocí ribonukleázy A (RNasa) a deoxyribonukleázy (DNasy) a po extrakci nukleových kyselin byl vzorek připraven na metagenomické sekvenování. Nasofaryngeální aspirát (výtěr) byl ve studii Lysholm et al. (2012) zpracován podobným způsobem.

Ve studii viromu u jedinců, kteří podstoupili transplantaci plic, byl analyzován vzorek z BALu a nejabundantnějšími čeleděmi byly *Herpesviridae, Anelloviridae* a *Coronaviridae* a bakteriofágy z čeledí *Myoviridae* a *Siphoviridae* (Segura-Wang et al. 2018), stejně jako ve vzorcích krve u stejných pacientů (viz. předchozí kapitola).

Studie, které použili k výzkumu vzorky z nasofaryngeálního aspirátu od pacientů s infekcí dolních dýchacích cest, měly srovnatelné výsledky (Lysholm et al. 2012; Yang et al. 2011), nalezli běžné respirační patogeny:

- *Paramyxoviridae* – obalené viry s lineárním ssRNA(-) genomem. Lidský respirační syncyciální virus (HRSV) patřil mezi nejvíce abundantní viry.
- *Picornaviridae* – lidský rhinovirus (HRV) byl dalším hojně zastoupeným virem.
- *Orthomyxoviridae* – třetí nejabundantnější čeleď, chřipkový virus A, B, C
- Ojediněle byly také nalezeny sekvence lidského bocaviru (HBoV) z čeledi *Parvoviridae* (viz. kapitola Studie viromu ze vzorku stolice) nebo polyomaviru z čeledi *Polyomaviridae* (neobalené viry s dsDNA genomem, nejrozšířenějším lidským polyomavirem je MCPyV).

Virom z dýchacích cest od zdravých pacientů v této kapitole není porovnáván. Dvě ze tří studií shrnuté v **tabulce 3** uvádí pouhých 5 % a 6,8 % sekvencí, které nebyly schopny

identifikovat. V tomto ohledu se zdá, že nejtěžším vzorkem pro identifikaci metagenomických dat je vzorek stolice. I přes snahy o odfiltrování buněk bylo množství sekvencí identifikováno jako lidská DNA.

Tabulka 3: Porovnání tří studií viromu ve vzorku pocházejícího z dýchacích cest (Segura-Wang et al. 2018; Lysholm et al. 2012; Yang et al. 2011)

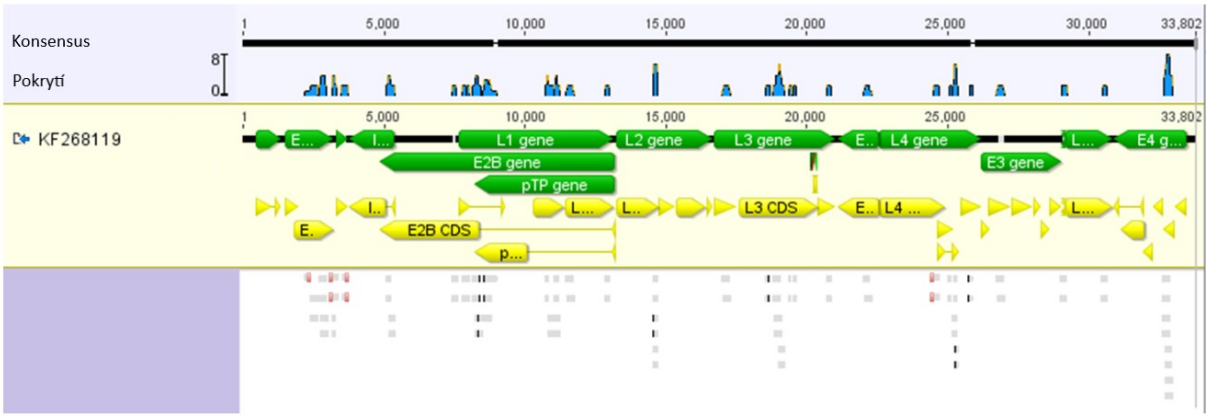
typ vzorku	charakteristika kohorty	celkový počet sekvencí	počet virových čtení	celkový počet nalezených virových druhů/rodin	% virových čtení (z toho bakteriofágů)	% nevirových sekvencí			sekvence neznámého původu (nebyl nalezen protějšek v databázi použité k identifikaci)	nejvyskytovanější virové čeledi
						bakterie	lidské	ostatní (houby, rostliny, savčí sekvence, atd..)		
nasofaryngeální aspirát	210 pacientů s infekcí dolních dýchacích cest vyžadující hospitalizaci [1]	703 790	110 931	39 druhů	15%	9%	60%	11%	**5%**	*Paramyxoviridae* *Orthomyxoviridae* *Picornaviridae*
	16 dětí od 2 měsíců do 10 let s akutní infekcí dolních cest dýchacích [2]	prům. 3,7 milionů sekvencí, 2,9 milionů validních/vzorek	N	> 100 druhů (89 druhů fágových)	0,05%	1%	91,6%	0,5%	**6,8%**	*Parvoviridae* *Anelloviridae*
BAL	15 pacientů podstupující transplantaci plic [3]	703 242 (průměrný počet čtení/vzorek)	N	N	3,11%	11%	37,2%	0,36%	**N**	*Coronaviridae* *Adenoviridae* *Herpesviridae*

[1] Lysholm et al. (2012) – Z celkových 703 790 čtení bylo 15 % virového původu. Ve virových sekvencích identifikovali 39 druhů virů, které byly klasifikovány do několika čeledí: *Paramyxoviridae, Picornaviridae* a *Orthomyxoviridae*. 60 % sekvencí mělo původ v lidské DNA, 9 % sekvencí bylo bakteriálních a 5 % se nepodařilo identifikovat.
[2] Yang et al. (2011) – Průměrně bylo získáno 2,9 milionů validních sekvencí na vzorek. Většina sekvencí (91,6 %) měla původ v lidském genomu, pouze 1 % sekvencí bylo z bakteriálního genomu a 0,05 % z virového. 6,8 % z validních sekvencí není známého původu. Identifikované viry patří do čeledí: *Pneumoviridae, Picornaviridae, Orthomyxoviridae, Adenoviridae, Paramyxoviridae, Parvoviridae* a *Herpesviridae*. (N = neuvedeno);
[3] Segura-Wang et al. (2018) – Ze všech čtení se podařilo identifikovat sekvence liského (37,2 %), bakteriálního (11 %) a virového původu (3,11 %) a byly identifikovány tyt virové čeledi: *Anelloviridae, Herpesviridae, Coronaviridae* a *Picornaviridae*. (N = neuvedeno)

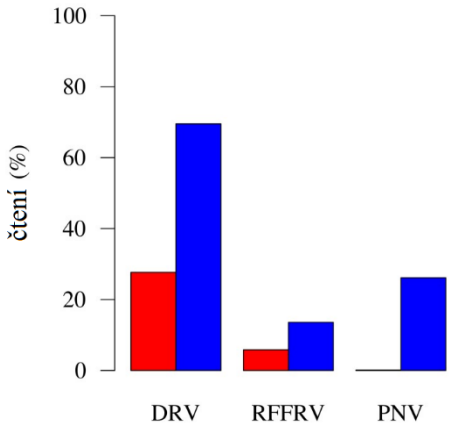
6. Identifikace nových virových genomů z metagenomických dat

6.1 Získání celého virového genomu z metagenomu

Získat celý virový genom z metagenomických dat může být technicky a časově náročné i pro známé virové genomy (vyžaduje většinou dosekvenování mezer Sangerovým sekvenováním, **obrázek 7**). U neznámých virů nebo velmi odlišných variant je nutné ke složení genomu použít další bioinformatické strategie, které se často liší podle typu vzorku i viru. Přitom je získání celého genomu většinou zásadní pro identifikaci jeho čtení ve vzorku (**Obrázek 8**). Smits et al. (2015) dělí tyto strategie do čtyř následujících kroků.



Obrázek 7: Genom adenoviru, v řádku „Pokrytí“ modré místa zobrazují pokrytí referenční sekvence, kde modrá není, tam je mezera. Materiál laboratoře molekulární genetiky (2.LF UK).



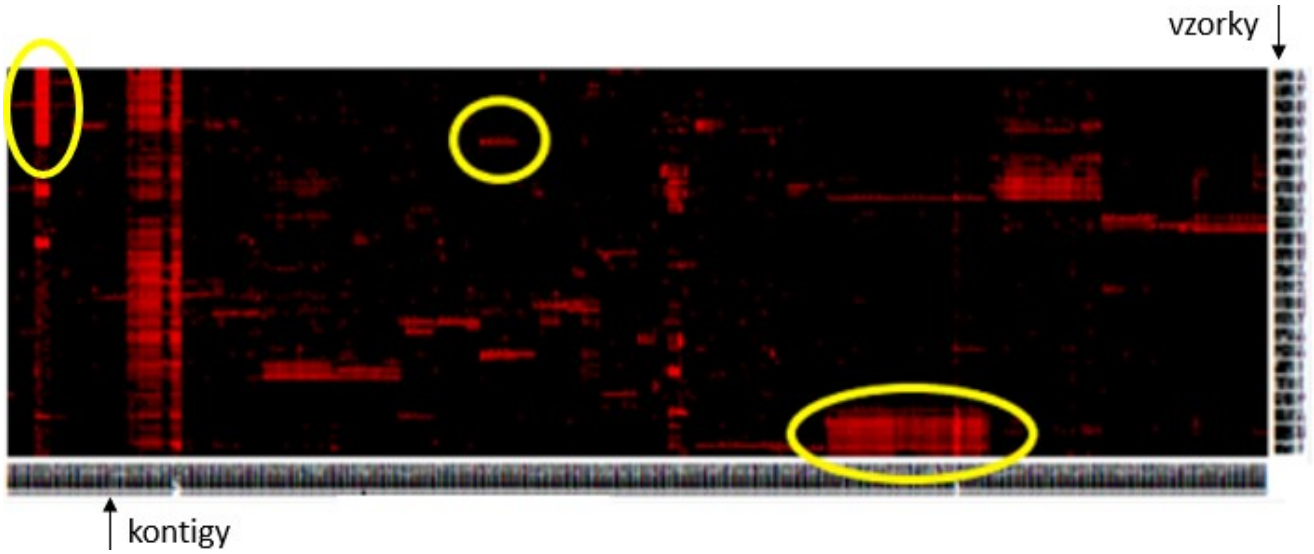
Obrázek 8: Porovnání procent identifikovaných čtení před (červené sloupce) a po sestavení celého genomu (modré sloupce). DRV – delfíní rhabdovirus, RFFRV – fekální rhabdovirus lišky obecné, PNV – nidovirus krajt, upraveno podle (Smits et al. 2015)

- **„Assembly“:** je prvním krokem k nalezení virového genomu. Jedná se o skládání krátkých fragmentů do větších celků = kontigů, po kterém následuje jejich identifikace. Současně používané programy nejsou původně navrženy pro vysoce diverzní virové genomy a neumějí pracovat s nerovnoměrným pokrytím (Smits et al. 2015). Proto, i když je celý virus ve vzorku přítomný ve velké kvantitě, výsledkem *de novo* „assembly“ je spousta krátkých kontigů. Efektivním způsobem, jak zvýšit šanci na složení co nejdelšího kontigu je opakované skládání („iterative assembly“), kdy se opakovaným skládáním prodlužuje kontig na jeho obou koncích, přičemž je v každém kroku možné použít jiné parametry pro skládání (Smits et al. 2015). Většinou se také nehledí na velké pokrytí. Je také dobré, pokud známe sekvence hostitelského genomu a cílový virus není endogenní, že mohou být hostitelské čtení před skládáním odečteny, což může zvýšit délku virových kontigů (Daly et al. 2015). Vhodnými programy opakovaného skládání jsou například PRICE (z anglického „Paired-read iterative contig extension“), který je závislý na kontigu nebo čtení se sekvenční podobností ke známému viru (Ruby, Bellare, et DeRisi 2013), program IVA (z anglického „Iterative virus assembler“) byl vytvořen pro viry s RNA genomy (Hunt et al. 2015) nebo OLC (z anglického „Overlap-layout-consensus), který se používá pro dlouhá čtení generované například sekvenátory firmy Roche.

- **Identifikace „seed“:** cílem tohoto kroku je identifikovat alespoň jeden fragment, který pochází z cílového viru, na základě sekvenční podobnosti se známým virem (Smits et al. 2015). K tomuto je nejvhodnější algoritmus BLAST a databáze známých virů. Prodlužováním nebo zařazením tohoto fragmentu do prodlužujícího se kontigu je možné odtud odvodit identitu viru. Pokud neexistuje identifikovatelný fragment, pak se lze zaměřit na délku kontigu či počet čtení, ze kterého se skládá. Napoví také znaky pro cirkulární genom.

- **Klastrování segmentů či spojování fragmentů jednoho genomu:** existují způsoby jak identifikovat, že nespojené kontigy, které už se dále neprodlužují, pochází z jednoho virového genomu. Tyto způsoby se také hodí pro viry se segmentovaným genomem, kdy je nutné identifikovat, že všechny tyto segmenty pochází z jednoho viru. Jedním z nich je hledání opakujících se sekvenčních motivů, jako jsou například místo začátku mitochondriální RNA (mRNA) a polyadenylový konec (řád *Mononegavirales*), „leader“ sekvence nebo sekvence regulace transkripce (řád *Nidovirales*). Významným

a velmi používaným způsobem je také analýza společného výskytu kontigů („co-occurance analysis“ nebo také „coverage profile binning“, **Obrázek 9**) nebo k-mer profilování, které třídí sekvence podle frekvence k-merů (Smits et al. 2015).



Obrázek 9: Příklad analýzy společného výskytu. Horizontálně jsou seřazeny kontigy a vertikálně jednotlivé vzorky. Čím je pole červenější, tím více čtení kontigu ve vzorku bylo. Předpokládá se, že stejně silné kontigy, vyskytující se společně v jednom nebo více vzorcích (žlutě vyznačené oblasti) by mohly znamenat, že pochází ze stejného organismu (např. viru). Materiál laboratoře molekulární genetiky (2.LF UK).

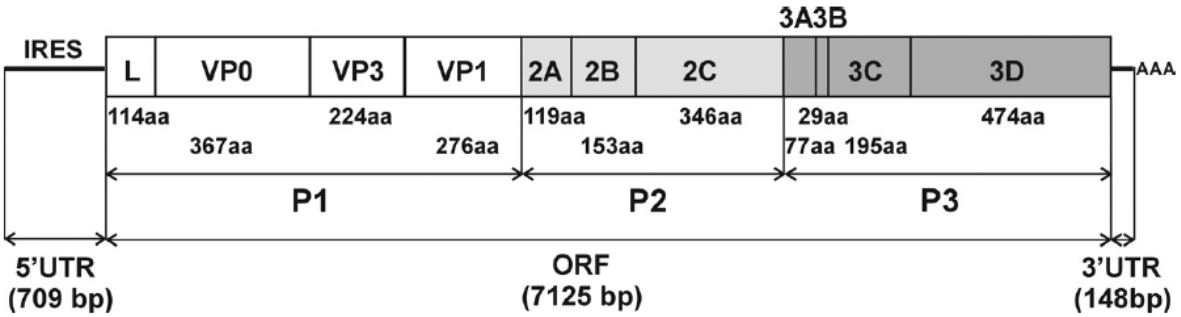
- **Finalizace genomů:** nakonec je potřeba seřadit kontigy tak, aby jejich kódující sekvence odpovídaly pořadí genů v genomu, tzn. vytvořit přesahy mezi kontigy. K tomu lze kombinovat různé typy bioinformatického skládání („assembly“ a „alignment“). Některé kontigy mohou mít chyby na svých koncích, a proto je nelze dále prodlužovat, či přiřadit k jinému kontigu. Toto se dá napravit úpravou konců kontigů

6.2 Příklady nalezení nového viru z metagenomických dat

6.2.1 Nový pikornavirus asociovaný s gastroenteritidou

Skupině L. Li et al. (2009) se podařilo identifikovat nový pikornavirus při analýze virových sekvencí (z 95 vzorků stolice od dětí trpících nonpolio akutní chabou obrnou) pomocí sekvenčně nezávislé PCR amplifikace, pyrosekvenování a vyhledávání sekvenční podobnosti. Následná asociace s gastroenteritidou se ukázala po testování přítomnosti tohoto viru v dalších vzorcích různých populací.

Kontigy dvou pacientů obsahovaly individuální sekvenční čtení, které byly vzdáleně příbuzné k sekvencím aichi viru (rod *Kobuvirus,* čeleď *Picornaviridae*) a byly dočasně pojmenovány jako salivirus (**s**tool **A**ichi-**li**ke virus). Použitím PCR spojenou s reverzní transkripcí (RT-PCR) byly propojeny mezery mezi virovými fragmenty složenými z NGS dat, primery k tomu navržené byly podle konzervovaných oblastí hovězího, prasečího a lidského kobuviru. Genom objeveného viru kódoval prekursor domnělého polyproteinu z 2374 aminokyselin, 5´ nepřekládanou oblast (UTR) a 3´ UTR (**Obrázek 10**). Porovnáním saliviru s ostatními pikornaviry se zjistilo, že má největší sekvenční podobnost s aichi virem: aminokyselinová podobnost P1, P2 a P3 oblastí byla 42,6 %, 35,2 % a 44,6 %. Aby mohl být součástí rodu z pikornavirů, měl by mít podobnost P1, P2 a P3 oblastí 40 %, >40 % a >50 %, je tedy salivirus na hranici mezi novým rodem a vysoce odlišným druhem v rodě *Kobuvirus.*



Obrázek 10: Organizace genomu lidského saliviru (L. Li et al. 2009)

6.2.2 Nový lidský papillomavirus asociovaný s febrilním respiračním onemocněním

Pomocí metagenomického přístupu se podařilo Mokili et al. (2013) identifikovat nový typ lidského papillomaviru (HPV) u pacientů s febrilním respiračním onemocněním. Z nasalfaryngeálních a orofaryngeálních stěrů bylo celkem získáno přes 2 miliony čtení, z nichž bylo 92 % sekvencí neznámého původu, které byly dále analyzovány. Pomocí vlastního počítačového skriptu vybrali ty kontigy, které byly cirkulární (na základě podobnosti mezi přesahujícími fragmenty na 5´ a 3´ koncích). DNA cirkulární strukturu mělo 193 z 2075 kontigů (9,3 %). Protože kromě mnoha virů také i bakteriální plasmidy jsou cirkulární, musely být od sebe rozlišeny. Plasmidy mají charakteristický znak – počátek replikace, pomocí kterého mohou být rozpoznány od virů, u kterých se tyto znaky vyskytují jen zřídka. Žádný takovýto znak však nebyl nalezen, a tak byly dále tyto kontigy analyzovány pomocí algoritmu tBLASTx s velmi relaxovanou E hodnotou (určuje přesnost identifikace) pro určení taxonomie.

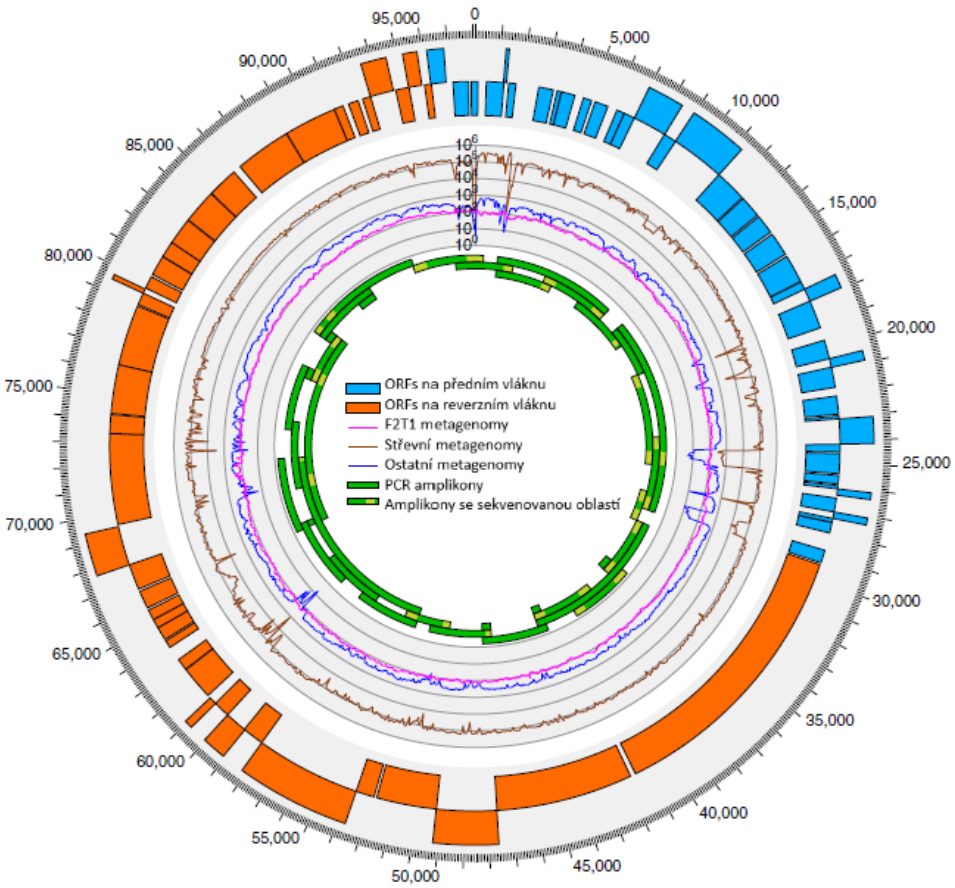
Kontig o délce 7299 bp měl značnou podobnost k HPV sekvencím a byl pojmenován jako HPV_SD2. Jeho genom zahrnuje sedm otevřených čtecích rámců (produkty L2, L1, E6, E7, E1, E2 a E4) stejně jako mají dříve popsané lidské papillomaviry. Fylogenetická analýza celého genomu a L1 genu ukázala, že HPV_SD2 je vzdáleně příbuzný gamma papillomaviru.

6.2.3 Objev vysoce abundantního bakteriofága

Významný objev učinila skupina Dutilh et al. (2014), kteří objevili nového bakteriofága reanalýzou sekvencí získaných dřívějším metagenickým sekvenováním stolice (Reyes et al. 2010). Celkem sekvenovali 12 stolic zdravých subjektů. Novým přístupem bylo kombinovat čtení všech vzorků najednou, použít tzv. „cross-assembly“ (Dutilh et al. 2012). Z virových metagenomů bylo díky *de novo* „cross-assembly“ získáno 7 584 výsledných kontigů, načež jeden z kontigů obsahoval čtení od všech 12-ti jedinců. Mezi tímto kontigem a ostatními kontigy bylo vypočítáno Spearmanovo korelační skóre, jehož výsledkem bylo, že spousta kontigů má velmi podobné struktury a mohly by tedy mít podobný výskyt v originálních vzorcích. Dále kontigy porovnávali s veřejně dostupnými databázemi sekvencí a byly nalezeny shody mezi 1 591 kontigy a sekvencemi v databázi Genbank. Sekvence, která se s těmito sekvencemi kontigů shodovala nejčastěji, pocházela z naprosto nesouvisejícího sekvenování metagenomu stolice. Z toho vyplynulo, že se jedná o kontigy něčeho, co se nachází ve stolici velmi často napříč subjekty.

Dutilh et al. (2014) získali následně genom nového bakteriofága dlouhý cca 97 kbp (**Obrázek 11**), který byl cirkulární a tohoto fága nazvali podle použité metodiky crAssfág (crAssphage). Podle přítomnosti otevřených čtecích rámců (ORFs) určili kódující oblasti a jako kontrolu, že došlo ke správnému složení sekvencí bioinformatickým způsobem následně celý genom přesekvenovali pomocí specifických primerů a Sangerova sekvenování.

Kromě toho se pokusili určit hostitele tohoto fága pomocí analýzy společného výskytu s různými střevními bakteriemi, hledáním homologie a identifikace CRISPR (z anglického „Clustered Regularly Interspaced Short Palindromic Repeats“) „spacers“. Z těchto analýz vyšlo, že hostitelem crAssphaga jsou pravděpodobně bakterie rodu *Bacteroides*. Toto tvrzení se ale zatím nepodařilo nikomu prokázat laboratorně.



Obrázek 11: Schématické zobrazení cirkulárního genomu crAssfága, upraveno podle (Dutilh et al. 2014)

7. Bioinformatické nástroje k identifikaci sekvencí

Ne všechny výzkumné týmy se chtějí zabývat skládáním nových virových genomů. Většině stačí informace, jaký je obsah známých mikroorganismů v získaných sekvencích. Také zde však existují různé přístupy, jak tyto sekvence identifikovat, lze k tomu použít nukleotidové či proteinové databáze, databáze vybraných patogenů či veškerých dostupných sekvencí. Způsobem analýzy můžeme zvyšovat množství identifikovaných sekvencí (sensitivu), ale přitom snižovat rychlost dosažení výsledku, která je však důležitá pro pozdější použití NGS v klinické praxi. Níže jsou uvedeny příklady vybraných dostupných bioinformatických nástrojů, které umožňují automatickou analýzu dat z metagenomického sekvenování:

- **SURPI** (z anglického „Sequence-based ultrarapid pathogen identification"): bioinformatický nástroj navržený skupinou Naccache et al. (2014) k využití v klinické diagnostice. SURPI dokáže detekovat viry a bakterie z 7-500 milion čtení během 11 minut až 16 hodin, podle typu zvolené analýzy. Schéma tohoto nástroje je shrnuto v **příloze č. 1.** Nástroj pracuje s dosud nezpracovanými čteními, ze kterých jsou nejprve odstraněny adaptéry a sekvence s nízkou kvalitou a komplexitou (Naccache et al. 2014). Následně jsou odečteny lidské sekvence a pro zbylá čtení existují dva typy analýz: rychlá a úplná. Při **rychlé analýze** jsou použity pouze nukleotidové databáze virových a bakteriálních sekvencí. Při **úplné analýze** jsou nejprve čtení srovnávány se všemi nukleotidovými sekvencemi v NCBI a jsou identifikovány nejen viry a bakterie, ale i houby a paraziti. Čtení, které nebyly identifikovány jsou následně poskládány *de novo* do kontigů, translatovány do sekvence proteinu a identifikovány pomocí proteinové databáze. Tímto způsobem je možné odhalit i novou variantu viru. Nástroj SURPI využívá k identifikaci databázi NCBI a vlastní programy pro alignment sekvencí (vyhledává shodu). Praktickou nevýhodu zatím představuje nutnost instalace na serveru.

- **Taxonomer:** další nástroj, který umožňuje velmi rychlou analýzu metagenomických sekvencí navrhli Flygare et al. (2016) a jeho název je Taxonomer. Kromě identifikace čtení umožňuje také identifikovat odpověď hostitelského organismu na ně – mRNA profilování. Schéma je shrnuto v **příloze č. 2**. Je rozdělen do 4 modulů, které identifikují čtení pomocí k-merů:

1. „Binner“: třídí sekvence DNA/RNA do širších taxonomických skupin – hostitelské a mikrobiální sekvence (viry, bakterie a houby, ale také neznámé sekvence, špatně zařaditelné sekvence a sekvence patogenů vybraných uživatelem).
2. „Classifier“: klasifikuje jak hostitelskou mRNA pomocí databáze genů, tak geny bakterií a hub. Ke klasifikaci se používají nukleotidové sekvence.
3. „Protonomer“: třídí sekvence virů a neznámé sekvence pomocí proteinové databáze (Uniprot), která obsahuje proteinové produkty genů.
4. „Afterburner“: zvyšuje senzitivitu k identifikaci vzdáleně homologních virových proteinů.

- **RIEMS** (z anglického „Reliable information extraction from metagenomic sequence dataset“): tento nástroj pro analýzu metagenomických dat vyvinuli Scheuch, Höper, et Beer (2015) a je rozdělen na základní analýzu („Basic analysis“) a následnou analýzu („Further analysis“), schéma je shrnuto v **příloze č. 3**. Při **základní analýze** dochází nejprve k identifikaci pozadí vzorku (hostitelské sekvence), ze zbývajících dat potom k identifikaci hojně zastoupených druhů (viry, bakterie, eukaryota) a nakonec analýza pomocí BLASTu se snižujícími se nároky na přesné určení. **Následná analýza** zpracovává všechny neklasifikované sekvence ze základní analýzy. Nejprve dodatečně porovnává neznámá čtení v nukleotidové databázi, a pokud zůstanou nějaké sekvence neklasifikovány, tak jsou převedeny do aminokyselinové sekvence a porovnávány v proteinové databázi.

- **MG-RAST** (z anglického „Metagenomic rapid annotation using subsystems technology“): slouží jako veřejný zdroj pro analýzu a anotaci dat po metagenomickém sekvenování (Meyer et al. 2008) a také veřejný zdroj dat. Jeho výhodou je práce na webovém rozhraní (www.mg-rast.org). Dokáže automaticky zpracovat sekvence přímo ze sekvenátoru. K identifikaci využívá nukleotidové databáze, databáze proteinů, ale také veřejné databáze sekvencí 16S ribosomální DNA (rDNA) k určení taxonomie bakterií, databázi mobilních elementů a jiné.

8. Závěr

Tato práce shrnuje všechny důležité aspekty, které mají vliv na identifikaci virů v sekvencích vzniklých po NGS.

Jsou zde uvedeny způsoby sekvenování viromů, jejich výhody (identifikace viru ve vzorku, kvantifikace viru) a limitace (viry nemají žádný společný gen jako mají např. bakterie gen pro 16S rRNA). Příkladem je sekvenování pomocí náhodných fragmentů nebo sekvenování s využitím hybridizačních sond.

Porovnáním studií viromu z různých vzorků vyplývá, že množství získaných virových sekvencí také záleží na tom, jaký byl výchozí vzorek a jak byl zpracován. V sekvencí nalezených ve zdravé stolici převažují hlavně bakteriální a virové jsou v menšině. V krvi (i ve vzorcích z dýchacích cest) se nachází velké množství lidských buněk, ale pokud se analyzuje krevní plazma, šance na zachycení virových sekvencí roste. Ve všech vzorcích různých výzkumných skupin bylo nalezeno značné množství anellovirů, jejichž význam není dosud jasný. Všechny typy vzorků obsahovaly méně i více sekvencí neznámého původu a je tedy možné, že se v nich vyskytují sekvence neobjevených virů.

Poslední a zřejmě nejtěžší částí při objevování nového viru je vlastní bioinformatické zpracování metagenomických dat. Tyto způsoby se mohou velmi lišit podle povahy virového genomu, jak je ukázáno na příkladech objevení nových virů (crAssphagu, nového lidského papillomaviru nebo nového pikornaviru). Pro rychlé zpracování dat může pomoci velké množství dostupných nástrojů.

Dodnes bylo objeveno pouze malé procento virů (hlavně bakteirofágů) a je možné, že dosud neobjevené viry způsobují různá další onemocnění nejen u lidí, ale i u zvířat a rostlin. Je tedy zapotřebí zdokonalení metod jejich detekce a charakterizace, což může vést k nalezení způsobu léčení takových onemocnění.

Seznam použité literatury

Ackermann, Hans W. s.d. ‘Bacteriophage taxonomy’.

Ackermann, Hans-W. 2011. ‘The first phage electron micrographs’. *Bacteriophage* 1 (4): 225–27. https://doi.org/10.4161/bact.1.4.17280.

Aggarwala, Varun, Guanxiang Liang, et Frederic D. Bushman. 2017. ‘Viral Communities of the Human Gut: Metagenomic Analysis of Composition and Dynamics’. *Mobile DNA* 8 (1): 12. https://doi.org/10.1186/s13100-017-0095-y.

Anderson, Benjamin D., Kelli L. Barr, Gary L. Heil, John A. Friary, et Gregory C. Gray. 2012. ‘A Comparison of Viral Fitness and Virulence between Emergent Adenovirus 14p1 and Prototype Adenovirus 14p Strains’. *Journal of Clinical Virology* 54 (3): 265–68. https://doi.org/10.1016/j.jcv.2012.03.006.

Artenstein, Andrew W. 2012. ‘The Discovery of Viruses: Advancing Science and Medicine by Challenging Dogma’. *International Journal of Infectious Diseases* 16 (7): e470–73. https://doi.org/10.1016/j.ijid.2012.03.005.

Blinkova, O., A. Kapoor, J. Victoria, M. Jones, N. Wolfe, A. Naeem, S. Shaukat, et al. 2009. ‘Cardioviruses Are Genetically Diverse and Cause Common Enteric Infections in South Asian Children’. *Journal of Virology* 83 (9): 4631–41. https://doi.org/10.1128/JVI.02085-08.

Blinkova, O., J. Victoria, Y. Li, B. F. Keele, C. Sanz, J.-B. N. Ndjango, M. Peeters, et al. 2010. ‘Novel Circular DNA Viruses in Stool Samples of Wild-Living Chimpanzees’. *Journal of General Virology* 91 (1): 74–86. https://doi.org/10.1099/vir.0.015446-0.

Breitbart, Mya, Ian Hewson, Ben Felts, Joseph M. Mahaffy, James Nulton, Peter Salamon, et Forest Rohwer. 2003. ‘Metagenomic Analyses of an Uncultured Viral Community from Human Feces’. *Journal of Bacteriology* 185 (20): 6220–23. https://doi.org/10.1128/JB.185.20.6220-6223.2003.

Briese, Thomas, Amit Kapoor, Nischay Mishra, Komal Jain, Arvind Kumar, Omar J. Jabado, et W. Ian Lipkin. 2015. ‘Virome Capture Sequencing Enables Sensitive Viral Diagnosis and Comprehensive Virome Analysis’. A cura Anne Moscona. *MBio* 6 (5): e01491-15. https://doi.org/10.1128/mBio.01491-15.

Brum, J. R., J. C. Ignacio-Espinoza, S. Roux, G. Doulcier, S. G. Acinas, A. Alberti, S. Chaffron, et al. 2015. ‘Patterns and Ecological Drivers of Ocean Viral Communities’. *Science* 348 (6237): 1261498–1261498. https://doi.org/10.1126/science.1261498.

Conceição-Neto, Nádia, Mark Zeller, Hanne Lefrère, Pieter De Bruyn, Leen Beller, Ward Deboutte, Claude Kwe Yinda, et al. 2015. ‘Modular Approach to Customise Sample Preparation Procedures for Viral Metagenomics: A Reproducible Protocol for Virome Analysis’. *Scientific Reports* 5 (nouember): 16532. https://doi.org/10.1038/srep16532.

Crom, S. C. M. de, J. W. A. Rossen, A. M. van Furth, et C. C. Obihara. 2016a. ‘Enterovirus and parechovirus infection in children: a brief overview’. *European Journal of Pediatrics* 175: 1023–29. https://doi.org/10.1007/s00431-016-2725-7.

———. 2016b. ‘Enterovirus and parechovirus infection in children: a brief overview’. *European Journal of Pediatrics* 175: 1023–29. https://doi.org/10.1007/s00431-016-2725-7.

Daly, Gordon M., Richard M. Leggett, William Rowe, Samuel Stubbs, Maxim Wilkinson, Ricardo H. Ramirez-Gonzalez, Mario Caccamo, William Bernal, et Jonathan L. Heeney. 2015. ‘Host Subtraction, Filtering and Assembly Validations for Novel Viral Discovery Using Next Generation Sequencing Data’. A cura I. King Jordan. *PLOS ONE* 10 (6): e0129059. https://doi.org/10.1371/journal.pone.0129059.

Dutilh, Bas E., Noriko Cassman, Katelyn McNair, Savannah E. Sanchez, Genivaldo G. Z. Silva, Lance Boling, Jeremy J. Barr, et al. 2014. ‘A Highly Abundant Bacteriophage

Discovered in the Unknown Sequences of Human Faecal Metagenomes'. *Nature Communications* 5 (1): 4498. https://doi.org/10.1038/ncomms5498.

Dutilh, Bas E., Robert Schmieder, Jim Nulton, Ben Felts, Peter Salamon, Robert A. Edwards, et John L. Mokili. 2012. 'Reference-independent comparative metagenomics using cross-assembly: crAss'. *Bioinformatics* 28 (24): 3225–31. https://doi.org/10.1093/bioinformatics/bts613.

Fields, Bernard N., et David M. Knipe. 1991. *Fundamental virology*. 2nd ed. New York, USA: Raven Press.

Finkbeiner, Stacy R., Adam F. Allred, Phillip I. Tarr, Eileen J. Klein, Carl D. Kirkwood, et David Wang. 2008. 'Metagenomic Analysis of Human Diarrhea: Viral Detection and Discovery'. A cura Edward C. Holmes. *PLoS Pathogens* 4 (2): e1000011. https://doi.org/10.1371/journal.ppat.1000011.

Flygare, Steven, Keith Simmon, Chase Miller, Yi Qiao, Brett Kennedy, Tonya Di Sera, Erin H. Graf, et al. 2016. 'Taxonomer: an interactive metagenomics analysis portal for universal pathogen detection and host mRNA expression profiling'. *Genome Biology* 17 (1): 111. https://doi.org/10.1186/s13059-016-0969-1.

Gebrekristos, Gebretsakan, Mebrahtu Teweldemedhin, Letebrhan Hagos, Tuom Gebrewahid, Berihu Gidey, et Hailay Gebreyesus. 2018. 'Hepatitis C virus infections and associated risk factors in patients with diabetes mellitus; case control study in North West Tigray, Ethiopia'. *BMC Research Notes* 11 (1): 873. https://doi.org/10.1186/s13104-018-3983-4.

Holtz, Lori R., Song Cao, Guoyan Zhao, Irma K. Bauer, Donna M. Denno, Eileen J. Klein, Martin Antonio, et al. 2014. 'Geographic Variation in the Eukaryotic Virome of Human Diarrhea'. *Virology* 468–470 (nouember): 556–64. https://doi.org/10.1016/j.virol.2014.09.012.

Hunt, Martin, Astrid Gall, Swee Hoe Ong, Jacqui Brener, Bridget Ferns, Philip Goulder, Eleni Nastouli, Jacqueline A. Keane, Paul Kellam, et Thomas D. Otto. 2015. 'IVA: Accurate de Novo Assembly of RNA Virus Genomes'. *Bioinformatics* 31 (14): 2374–76. https://doi.org/10.1093/bioinformatics/btv120.

Kramná, Lenka, Kateřina Kolářová, Sami Oikarinen, Juha-Pekka Pursiheimo, Jorma Ilonen, Olli Simell, Mikael Knip, Riitta Veijola, Heikki Hyöty, et Ondrej Cinek. 2015. 'Gut Virome Sequencing in Children With Early Islet Autoimmunity: Figure 1'. *Diabetes Care* 38 (5): 930–33. https://doi.org/10.2337/dc14-2490.

Kumar, Arvind, Satyapramod Murthy, et Amit Kapoor. 2017. 'Evolution of Selective-Sequencing Approaches for Virus Discovery and Virome Analysis'. *Virus Research* 239 (iulius): 172–79. https://doi.org/10.1016/j.virusres.2017.06.005.

Li, L., J. Victoria, A. Kapoor, O. Blinkova, C. Wang, F. Babrzadeh, C. J. Mason, et al. 2009. 'A Novel Picornavirus Associated with Gastroenteritis'. *Journal of Virology* 83 (22): 12002–6. https://doi.org/10.1128/JVI.01241-09.

Li, Linlin, Amit Kapoor, Beth Slikas, Oderinde Soji Bamidele, Chunlin Wang, Shahzad Shaukat, Muhammad Alam Masroor, et al. 2010. 'Multiple Diverse Circoviruses Infect Farm Animals and Are Commonly Found in Human and Chimpanzee Feces'. *Journal of Virology* 84 (4): 1674–82. https://doi.org/10.1128/JVI.02109-09.

Li, S.-K., R. K.-K. Leung, H.-X. Guo, J.-F. Wei, J.-H. Wang, K.-T. Kwong, S.-S. Lee, C. Zhang, et S. K.-W. Tsui. 2012. 'Detection and Identification of Plasma Bacterial and Viral Elements in HIV/AIDS Patients in Comparison to Healthy Adults'. *Clinical Microbiology and Infection* 18 (11): 1126–33. https://doi.org/10.1111/j.1469-0691.2011.03690.x.

Lustig, A, et A J Levine. 1992. 'One hundred years of virology.' *Journal of Virology* 66 (8): 4629–31.

Lysholm, Fredrik, Anna Wetterbom, Cecilia Lindau, Hamid Darban, Annelie Bjerkner, Kristina Fahlander, A. Michael Lindberg, Bengt Persson, Tobias Allander, et Björn Andersson. 2012. 'Characterization of the Viral Microbiome in Patients with Severe Lower Respiratory Tract Infections, Using Metagenomic Sequencing'. A cura Sarah K. Highlander. *PLoS ONE* 7 (2): e30875. https://doi.org/10.1371/journal.pone.0030875.

Metzker, Michael L. 2010. 'Sequencing Technologies — the next Generation'. *Nature Reviews Genetics* 11 (1): 31–46. https://doi.org/10.1038/nrg2626.

Meyer, F., D. Paarmann, M. D'Souza, R. Olson, EM Glass, M. Kubal, T. Paczian, et al. 2008. 'The metagenomics RAST server – a public resource for the automatic phylogenetic and functional analysis of metagenomes'. *BMC Bioinformatics* 9 (1): 386. https://doi.org/10.1186/1471-2105-9-386.

Minot, S., A. Bryson, C. Chehoud, G. D. Wu, J. D. Lewis, et F. D. Bushman. 2013. 'Rapid Evolution of the Human Gut Virome'. *Proceedings of the National Academy of Sciences* 110 (30): 12450–55. https://doi.org/10.1073/pnas.1300833110.

Minot, S., R. Sinha, J. Chen, H. Li, S. A. Keilbaugh, G. D. Wu, J. D. Lewis, et F. D. Bushman. 2011. 'The Human Gut Virome: Inter-Individual Variation and Dynamic Response to Diet'. *Genome Research* 21 (10): 1616–25. https://doi.org/10.1101/gr.122705.111.

Mokili, John L., Bas E. Dutilh, Yan Wei Lim, Bradley S. Schneider, Travis Taylor, Matthew R. Haynes, David Metzgar, et al. 2013. 'Identification of a Novel Human Papillomavirus by Metagenomic Analysis of Samples from Patients with Febrile Respiratory Illness'. *PLOS ONE* 8 (3): e58404. https://doi.org/10.1371/journal.pone.0058404.

Mokili, John L, Forest Rohwer, et Bas E Dutilh. 2012. 'Metagenomics and Future Perspectives in Virus Discovery'. *Current Opinion in Virology* 2 (1): 63–77. https://doi.org/10.1016/j.coviro.2011.12.004.

Moustafa, Ahmed, Chao Xie, Ewen Kirkness, William Biggs, Emily Wong, Yaron Turpaz, Kenneth Bloom, et al. 2017. 'The Blood DNA Virome in 8,000 Humans'. A cura Robert Belshaw. *PLOS Pathogens* 13 (3): e1006292. https://doi.org/10.1371/journal.ppat.1006292.

Naccache, Samia N., Scot Federman, Narayanan Veeraraghavan, Matei Zaharia, Deanna Lee, Erik Samayoa, Jerome Bouquet, et al. 2014. 'A cloud-compatible bioinformatics pipeline for ultrarapid pathogen identification from next-generation sequencing of clinical samples'. *Genome Research* 24 (7): 1180–92. https://doi.org/10.1101/gr.171934.113.

Popgeorgiev, Nikolay, Mickaël Boyer, Laura Fancello, Sonia Monteil, Catherine Robert, Romain Rivet, Claude Nappez, et al. 2013. 'Marseillevirus-Like Virus Recovered From Blood Donated by Asymptomatic Humans'. *The Journal of Infectious Diseases* 208 (7): 1042–50. https://doi.org/10.1093/infdis/jit292.

Rascovan, Nicolás, Raja Duraisamy, et Christelle Desnues. 2016a. 'Metagenomics and the Human Virome in Asymptomatic Individuals'. *Annual Review of Microbiology* 70 (1): 125–41. https://doi.org/10.1146/annurev-micro-102215-095431.

———. 2016b. 'Metagenomics and the Human Virome in Asymptomatic Individuals'. *Annual Review of Microbiology* 70 (1): 125–41. https://doi.org/10.1146/annurev-micro-102215-095431.

Reyes, Alejandro, Matthew Haynes, Nicole Hanson, Florent E. Angly, Andrew C. Heath, Forest Rohwer, et Jeffrey I. Gordon. 2010. 'Viruses in the fecal microbiota of monozygotic twins and their mothers'. *Nature* 466 (7304): 334–38. https://doi.org/10.1038/nature09199.

Rose, C., A. Parker, B. Jefferson, et E. Cartmell. 2015. 'The Characterization of Feces and Urine: A Review of the Literature to Inform Advanced Treatment Technology'. *Critical*

Reviews in Environmental Science and Technology 45 (17): 1827–79. https://doi.org/10.1080/10643389.2014.1000761.

Ruby, J. Graham, Priya Bellare, et Joseph L. DeRisi. 2013. 'PRICE: Software for the Targeted Assembly of Components of (Meta) Genomic Sequence Data'. *G3: Genes|Genomes|Genetics* 3 (5): 865–80. https://doi.org/10.1534/g3.113.005967.

Segura-Wang, Maia, Irene Görzer, Peter Jaksch, et Elisabeth Puchhammer-Stöckl. 2018. 'Temporal Dynamics of the Lung and Plasma Viromes in Lung Transplant Recipients'. A cura Houssam Attoui. *PLOS ONE* 13 (7): e0200428. https://doi.org/10.1371/journal.pone.0200428.

Scheuch, Matthias, Dirk Höper, et Martin Beer. 2015. 'RIEMS: a software pipeline for sensitive and comprehensive taxonomic classification of reads from metagenomics datasets'. *BMC Bioinformatics* 16 (1). https://doi.org/10.1186/s12859-015-0503-6.

Siqueira, Juliana D., Maria Gloria Dominguez-Bello, Monica Contreras, Orlana Lander, Hortensia Caballero-Arias, Deng Xutao, Oscar Noya-Alarcon, et Eric Delwart. 2018. 'Complex Virome in Feces from Amerindian Children in Isolated Amazonian Villages'. *Nature Communications* 9 (1): 4270. https://doi.org/10.1038/s41467-018-06502-9.

Smits, Saskia L., Rogier Bodewes, Aritz Ruiz-González, Wolfgang Baumgärtner, Marion P. Koopmans, Albert D. M. E. Osterhaus, et Anita C. Schürch. 2015. 'Recovering full-length viral genomes from metagenomes'. *Frontiers in Microbiology* 6 (october). https://doi.org/10.3389/fmicb.2015.01069.

Sohn, Jang-il, et Jin-Wu Nam. 2016. 'The Present and Future of *de Novo* Whole-Genome Assembly'. *Briefings in Bioinformatics*, october, bbw096. https://doi.org/10.1093/bib/bbw096.

Stephen, A. M., et J. H. Cummings. 1980. 'THE MICROBIAL CONTRIBUTION TO HUMAN FAECAL MASS'. *Journal of Medical Microbiology* 13 (1): 45–56. https://doi.org/10.1099/00222615-13-1-45.

Šmarda, Jan, Jiří Doškař, Roman Pantůček, Vladislava Růžičková, et Jana Koptíková. 2010. *Metody molekulární biologie*. Masarykova univerzita.

Tang, Patrick, et Charles Chiu. 2010. 'Metagenomics for the Discovery of Novel Human Viruses'. *Future Microbiology* 5 (2): 177–89. https://doi.org/10.2217/fmb.09.120.

Waters, Daniel L. E., et Frances M. Shapter. 2014. 'The Polymerase Chain Reaction (PCR): General Methods'. In *Cereal Genomics*, a cura Robert J. Henry et Agnelo Furtado, 1099:65–75. Totowa, NJ: Humana Press. https://doi.org/10.1007/978-1-62703-715-0_7.

Wootton, Susan H., David W. Scheifele, Annie Mak, Martin Petric, et Danuta M. Skowronski. 2006. 'Detection of Human Influenza Virus in the Stool of Children': *The Pediatric Infectious Disease Journal* 25 (12): 1194–95. https://doi.org/10.1097/01.inf.0000245097.95543.11.

Wylie, Kristine M. 2017. 'The Virome of the Human Respiratory Tract'. *Clinics in Chest Medicine* 38 (1): 11–19. https://doi.org/10.1016/j.ccm.2016.11.001.

Wylie, Kristine M., George M. Weinstock, et Gregory A. Storch. 2012. 'Emerging view of the human virome'. *Translational Research*, In-Depth Review: Of Microbes and Men: Challenges of the Human Microbiome, 160 (4): 283–90. https://doi.org/10.1016/j.trsl.2012.03.006.

Yang, J., F. Yang, L. Ren, Z. Xiong, Z. Wu, J. Dong, L. Sun, et al. 2011. 'Unbiased Parallel Detection of Viral Pathogens in Clinical Samples by Use of a Metagenomic Approach'. *Journal of Clinical Microbiology* 49 (10): 3463–69. https://doi.org/10.1128/JCM.00273-11.

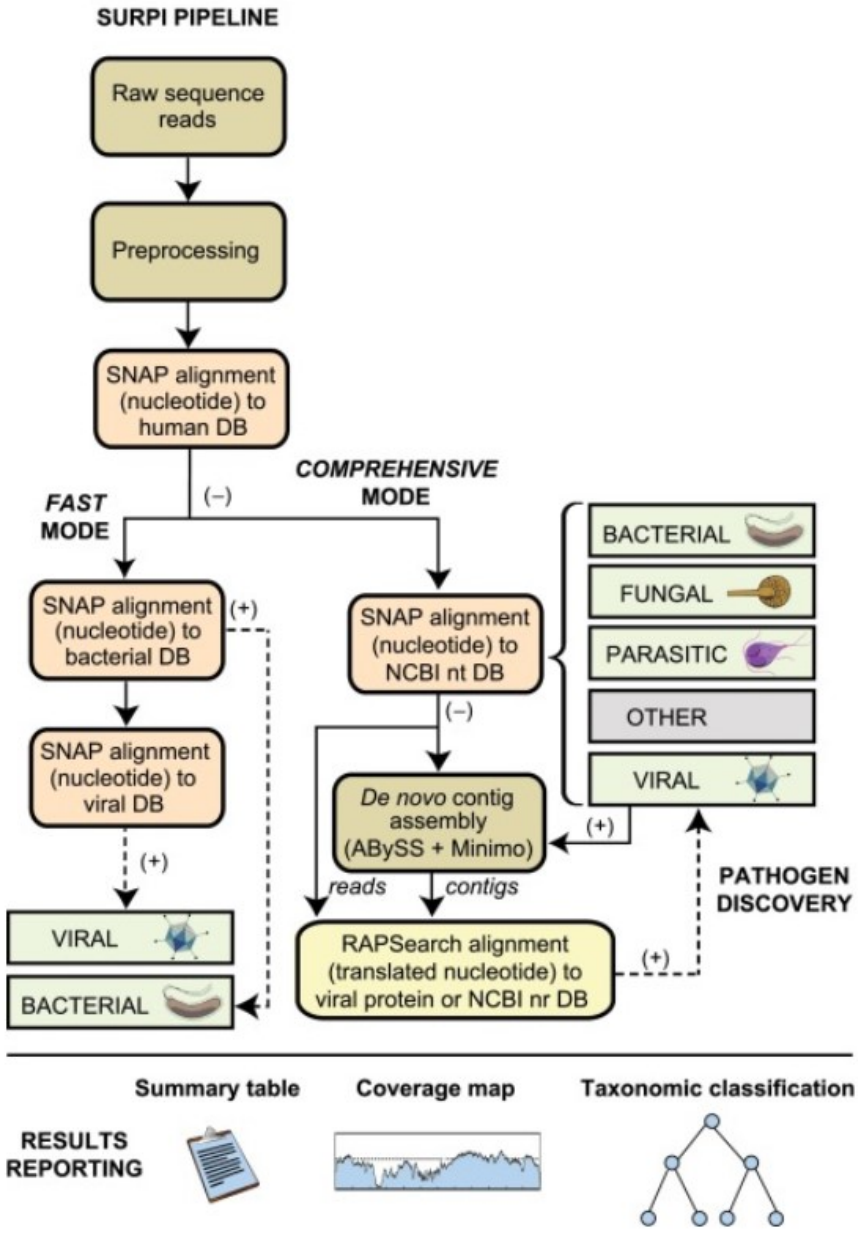
Yinda, Claude Kwe, Emiel Vanhulle, Nádia Conceição-Neto, Leen Beller, Ward Deboutte, Chenyan Shi, Stephen Mbigha Ghogomu, Piet Maes, Marc Van Ranst, et Jelle

Matthijnssens. 2019. ‘Gut Virome Analysis of Cameroonians Reveals High Diversity of Enteric Viruses, Including Potential Interspecies Transmitted Viruses’. A cura Marilyn J. Roossinck. *MSphere* 4 (1): e00585-18, /msphere/4/1/mSphere585-18.atom. https://doi.org/10.1128/mSphere.00585-18.

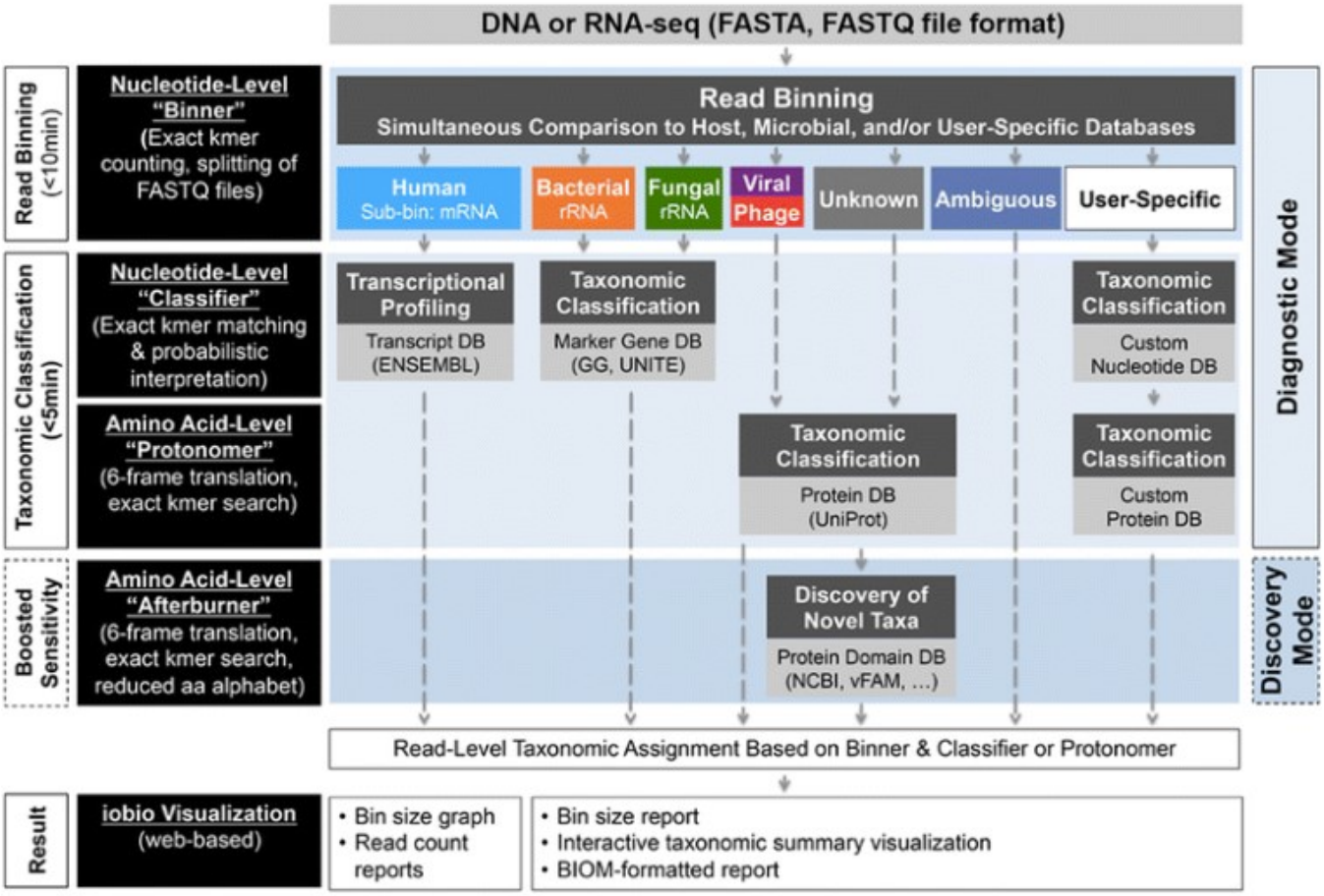
Zárate, Selene, Blanca Taboada, Martha Yocupicio-Monroy, et Carlos F. Arias. 2017. ‘Human Virome’. *Archives of Medical Research*, Microbiota in Health and Disease, 48 (8): 701–16. https://doi.org/10.1016/j.arcmed.2018.01.005.

Zwartouw, H. T., J. C. N. Westwood, et G. Appleyard. 1962. ‘Purification of Pox Viruses by Density Gradient Centrifugation’. *Journal of General Microbiology* 29 (3): 523–29. https://doi.org/10.1099/00221287-29-3-523.

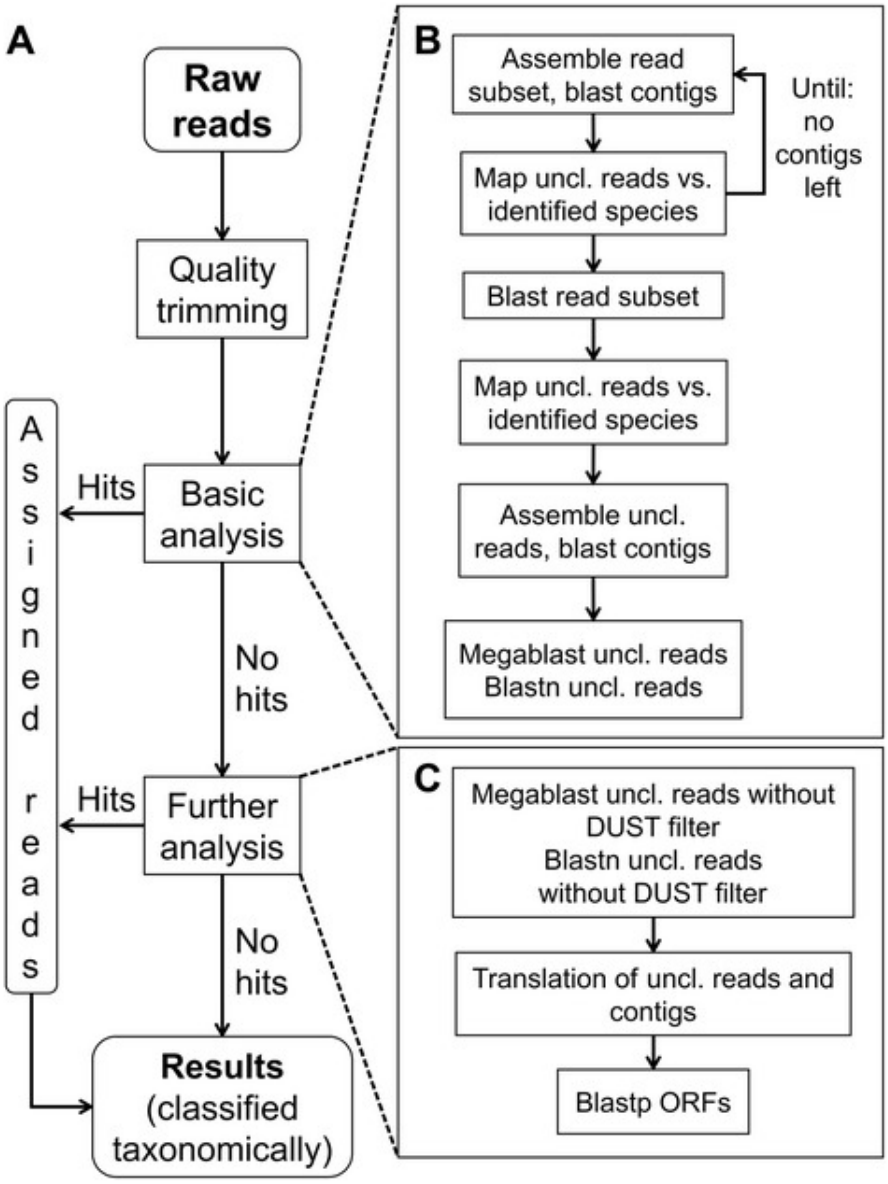
Přílohy



Příloha č. 1: Schématické shrnutí nástroje SURPI (Naccache et al. 2014)



Příloha č. 2: Schématické shrnutí nástroje Taxonomer (Flygare et al. 2016)



Příloha č. 3: Schématické shrnutí nástroje RIEMS (Scheuch, Höper, et Beer 2015)