VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

DEPARTMENT OF BIOMEDICAL ENGINEERING

METODY ZPRACOVÁNÍ SEKVENAČNÍCH DAT TECHNOLOGIE OXFORD NANOPORE PRO ÚČELY METAGENOMIKY

METHODS OF PROCESSING OXFORD NANOPORE SEQUENCING DATA FOR METAGENOMICS

BAKALÁŘSKÁ PRÁCE

BACHELOR'S THESIS

AUTOR PRÁCE **Lujza Barilíková**

AUTHOR

VEDOUCÍ PRÁCE **Ing. Kristýna Kupková**

SUPERVISOR

BRNO 2018

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ
FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ

Bakalářská práce

bakalářský studijní obor **Biomedicínská technika a bioinformatika**

Ústav biomedicínského inženýrství

Studentka: Lujza Barilíková ***ID:*** 185945

Ročník: 3 ***Akademický rok:*** 2017/18

NÁZEV TÉMATU:

Metody zpracování sekvenačních dat technologie Oxford Nanopore pro účely metagenomiky

POKYNY PRO VYPRACOVÁNÍ:

1) Vypracujte literární rešerši sekvenačních technologií s primárním zaměřením na technologie třetí generace, zejména na nanopórové sekvenování. 2) Prostudujte současné možnosti vizualizace metagenomických dat. 3) Ve zvoleném programovacím prostředí vytvořte program pro načtení a předzpracování FAST5 souborů získaných nanopórovým sekvenováním. 4) Program doplňte metodami redukce dimezionality umožňujícími vizualizaci metagenomických dat. 5) Diskutujte dosažené výsledky.

DOPORUČENÁ LITERATURA:

[1] LU, Hengyun, Francesca GIORDANO a Zemin NING. Oxford Nanopore MinION Sequencing and GenomeAssembly. Genomics, Proteomics & Bioinformatics [online]. 2016, 14(5), 265-279 [cit. 2017-09-02]. DOI:10.1016/j.gpb.2016.05.004. ISSN 16720229.

[2] LACZNY, Cedric C., Nicolás PINEL, Nikos VLASSIS a Paul WILMES. Alignment-free Visualization of Metagenomic Data by Nonlinear Dimension Reduction. Scientific Reports [online]. 2015, 4(1), - [cit. 2017- 09-02]. DOI: 10.1038/srep04516. ISSN 2045-2322.

Termín zadání: 5.2.2018 ***Termín odevzdání:*** 25.5.2018

Vedoucí práce: Ing. Kristýna Kupková
Konzultant:

prof. Ing. Ivo Provazník, Ph.D.
předseda oborové rady

Abstrakt

Revolučná technológia sekvenovania od spoločnosti Oxford Nanopore Technologies – MinION, predstavuje veľkú nádej v oblasti metagenomiky. Nízka cena, produkovanie dlhých čítaní a prenosnosť, vďaka malým rozmerom, predstavuje len jednu z mnohých výhod tejto technológie. Napriek týmto benefitom je tu však nedostatok dostupných výpočtových nástrojov, ktoré by nám umožnili plne zaobchádzať s produkovanými dátami. V úvode tejto bakalárskej práce sú predstavené terajšie sekvenačné technológie so zameraním sa na technológie tretej generácie, predovšetkým na spomínané nanopórové sekvenovanie. V práci sú uvedené aj súčasné možnosti vizualizácie metagenomických dát. Hlavným cieľom tejto práce je vytvoriť algoritmus, ktorý za pomoci aplikácie metód redukcie dimenzionality priamo na surové dáta, produkované nanopórovým sekvenovaním, bude realizovať tzv. binning metagenomických vzoriek.

Kľúčové slová

Metagenomika, Oxford Nanopore, nanopórové sekvenovanie, redukcia dimenzií

Abstract

The revolutionary sequencing technology introduced by Oxford Nanopore Technologies – MinION holds a great promise in the field of metagenomics. Low cost, produced long reads and portability, due to its small dimensions, represents only one of the many advantages of this technology. Despite the benefits, there is a lack of available computational tools for handling the produced data. The theoretical part of the thesis first introduces current sequencing technologies with main focus on the third-generation sequencing and especially on nanopore sequencing. The recent possibilities of metagenomic data visualization are introduced. The main purpose of the bachelor thesis is to make an algorithm for binning of metagenomic samples based on use of dimensionality reduction techniques straight on raw data produced by nanopore sequencing.

Keywords

Metagenomics, Oxford Nanopore, nanopore sequencing, dimensionality reduction

Bibliografická citácia:

BARILÍKOVÁ, L. Metódy zpracování sekvenačních dat technologie Oxford Nanopore pro účely metagenomiky. Brno: Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, 2018. 48s. Vedoucí práce: Ing. Kristýna Kupková.

Prehlásenie

Prehlasujem, že svoju záverečnú prácu na tému metódy zpracování sekvenačních dat technologie Oxford Nanopore pro účely metagenomiky, som vypracovala samostatne pod vedením vedúceho bakalárskej práce a s použitím odbornej literatúry a ďalších informačných zdrojov, ktoré sú všetky citované v práci a uvedené v zozname literatúry na konci práce.

Ako autorka uvedenej záverečnej práce ďalej prehlasujem, že v súvislosti s vytvorením tejto záverečnej práce som neporušila autorské práva tretích osôb, predovšetkým som nezasiahla nedovoleným spôsobom do cudzích autorských práv osobnostných a som si plne vedomá následkov porušenia ustanovenia § 11 a nasledujúcich autorského zákona č. 121/2000 Sb., vrátane možných trestnoprávnych dôsledkov vyplývajúcich z ustanovenia časti druhej, hlavy VI. diel 4 Trestného zákonníka č. 40/2009 Sb.

V Brne dňa ……

podpis autorky

Poďakovanie

Ďakujem vedúcej mojej bakalárskej práce Ing. Kristýne Kupkovej za skvelé odborné vedenie, poskytnuté konzultácie, ochotu, motiváciu a ďalšie cenné rady pri spracovaní mojej bakalárskej práce. Taktiež by som sa rada poďakovala svojej rodine a priateľovi za podporu v priebehu jej spracovania.

V Brne dňa ……

podpis autorky

Obsah

Zoznam obrázkov

Zoznam tabuliek

ÚVOD

Napriek tomu, že ich nevidíme, sú mikroorganizmy neoddeliteľnou súčasťou nášho života. Každý proces v biosfére je spätý so zjavne nekonečnou kapacitou týchto miniatúrnych stvorení, ktoré transformujú svet okolo nás. Mikroorganizmy obývajú dokonca aj ľudské telo, preto by sme sa mohli zaoberať otázkou ich prepojenia s ochoreniami a rôznymi infekciami postihujúce náš organizmus.

K skúmaniu komunít mikroorganizmov vo veľkej miere prispel postupný rozvoj sekvenačných metód a technológií, ktorých záujmom je štúdium genetickej informácie. Vďaka obrovskému rozvoju dostupných sekvenátorov sa problém množstva produkovaných dát stal minulosťou. V súčasnosti sa preto rozvinulo úsilie tieto dáta spracovať, čo predstavuje vývoj nových algoritmov slúžiacich k ich spoľahlivej analýze a klasifikácii.

Cieľom tejto práce je vizualizácia metagenomických dát, produkovaných zariadením MinION od spoločnosti Oxford Nanopore Technologies, v snahe uskutočniť tzv. binning (roztriedenie) DNA sekvencií podľa organizmov do príslušných skupín na základe ich spoločne zdieľaných znakov získaných z jednotlivých sekvencií. Vizualizácia by mala tiež zjednodušiť interpretáciu, pre človeka nepredstaviteľných, obrovských metagenomických datasetov.

V práci sú podané bližšie informácie o metagenomike, spoločne s vývojom sekvenačných technológií poskytujúcich dáta, ktoré tvoria základ nasledujúcej metagenomickej analýzy. Výklad smeruje k postupnému dopracovaniu sa ku kľúčovému zariadeniu tejto práce, ktorým je nanopórový sekvenátor produkujúci skutočne dlhé sekvencie dát. Dôležitou súčasťou teoretickej časti je aj popis jednotlivých metód vizualizácie dát, ktorých voľba vo veľkej miere ovplyvňuje výslednú kvalitu analýzy.

Praktická časť práce zahŕňa spracovanie metagenomických sekvencií priamo zo signálov obdržaných nanopórovým sekevenovaním a realizáciu ich vizualizácie využitím metód založených na princípe redukcie dimenzií v snahe vytvoriť algoritmus pre roztriedenie týchto dát.

1. METAGENOMIKA

Mikroorganizmy, obklopujúce nás, nepracujú ako individuálne jednotky, ale ako komunity, kde každý z nich vykonáva svoju špecifickú funkciu. Akýkoľvek elementárny cyklus zahŕňa určitý druh mikroorganickej spolupráce, ktorá je prísne regulovaná a spojená interakciami mikrobiálnej komunity. Je dokonca zistené, že bakteriálne organizmy obývajú aj ľudské telo a odhaduje sa, že ich množstvo prevyšuje počet vlastných buniek až 10-krát [1]. Taktiež sa uvažuje o ich spojení s ochoreniami ako sú napríklad rakovina, cukrovka a rôzne črevné infekcie [2]. Vedomosti o mikróboch však pochádzajú z veľkej časti z laboratórneho prostredia a sú získané za abnormálnych a neprirodzených podmienok, kde sú optimálne pestované na umelom médiu v čistej kultúre bez určitého ekologického kontextu [3].

Naproti tomu metagenomika, veda zaoberajúca sa sekvenovaním a analýzou DNA izolovanej priamo z mikrobiálnych vzoriek bez nutnosti ich kultivácie, umožňuje skúmať mikroorganizmy v ich prirodzenom prostredí, v ktorom poväčšine žijú. Metagenomiku, ako vedu, je však len veľmi obťažné vymedziť presnou definíciou, pretože je náročné túto disciplínu v akomkoľvek smere limitovať. Jednoznačne však predstavuje určitý kultivačne nezávislý stupeň klasifikácie jednotlivých komunít, ich členov a umožňuje štúdium týchto komunít na úrovni genómov. Tieto metódy sú základom genomiky a štúdií, ktoré sú zamerané na skúmanie biosféry na tejto úrovni. Z toho dôvodu je možné, že táto veda bude prevratným prínosom v biológii, medicíne, ekológii a v neposlednom rade aj v biotechnológii [3][2].

Štúdium mikrobiálnych komunít bez nutnosti kultivácie je obrovským prínosom, pretože nie všetky mikroorganizmy je možné kultivovať, keďže si to vyžaduje špeciálne, doposiaľ neznáme podmienky. Jeden zo spôsobov analýzy bez potreby kultivovania sa stalo 16S rRNA sekvenovanie [4]. Táto 16S podjednotka ribozómovej RNA, ktorá je prítomná takmer u všetkých prokaryot, je veľmi dobrým identifikátorom, pretože býva pre každý druh baktérií odlišná, vďaka čomu sa stala základom mikrobiálnej taxonómie. Pomocou nej bola taktiež odhalená veľká časť druhovej rozmanitosti mikrobiálneho sveta [5]. Toto cielené sekvenovanie je výborným nástrojom pre posúdenie zastúpenia a relatívneho množstva baktérií v určitej vzorke. Na druhej strane, touto metódou preskúmame len relatívne malú časť každého genómu. Ďalší spôsob analýzy, ktorým naopak získavame prehľad o celom genóme organizmov je náhodné shotgun sekvenovanie. Táto metóda je v porovnaní s cieleným sekvenovaním oveľa menej náchylná na skreslenie, pretože mnohé oblasti genómu disponujú značne vyššou variabilitou ako rRNA, to znamená že produkované dáta poskytujú lepšie rozlíšenie, dokonca aj blízko príbuzných organizmov [6]. Analýza týchto dát však nie je úplne jednoduchá a v súčasnosti je cieľom vynájsť algoritmy pre ich rýchle a efektívne spracovanie, čo je cieľom aj tejto práce.

2. VÝVOJ SEKVENAČNÝCH TECHNOLÓGIÍ

Po objavení terciárnej štruktúry molekuly DNA Watsonom a Crickom, sa vedci rozsiahlo začali zaoberať určovaním poradia báz nukleových kyselín v týchto molekulách, teda sekvenovaním, ktoré má obrovské uplatnenie v medicíne, biológií, biotechnológií a mnohých ďalších vedeckých a výskumných odboroch. Cieľom je predovšetkým skúmanie celých genómov alebo len ich častí, ktoré sú ako súhrn celej genetickej informácie veľmi rozsiahle. Z toho dôvodu, aj napriek najnovším technológiám, nie je proces sekvenovania okamžitou záležitosťou. Doposiaľ bolo vyvinutých niekoľko techník slúžiacich k sekvenovaniu DNA, ktoré sa od seba odlišujú predovšetkým jednotlivými princípmi realizácie, cenou, rýchlosťou a množstvom produkovaných dát, ako je možné vidieť v úvodnom prehľade technológií v Tab.1. [6].

Prevratom skúmania génov, či celých genómov sa stala prvá generácia sekvenovania prinášajúca so sebou dve významné metódy, Sangerova a Maxam-Gilbertova metóda. Obe metódy disponujú určitým obmedzeniam, napriek ktorým ale umožňujú realizovať sekvenáciu genetického materiálu. Pomerne vysoké náklady a časová zdĺhavosť boli však podnetom k vynaliezaniu nových sekvenačných postupov a technológií [7].

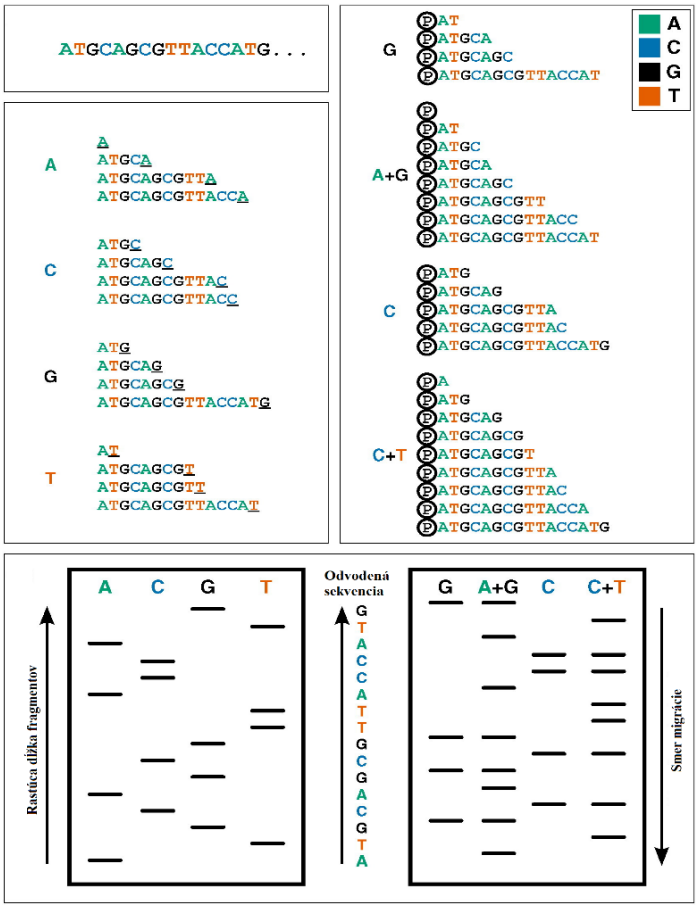
Riešením týchto problémov sa stal nástup druhej generácie sekvenovania, alebo tzv. novej generácie sekvenovania (NGS, z angl. Next-generation sequencing), ktorá vyniká predovšetkým svojou výkonnosťou dosiahnutou využitím paralelizácie sekvenovania.. Po úspešnej druhej generácii sa ďalším prevratom sa stala tretia generácia sekvenovania, ktorá využíva sekvenovanie jednej molekuly v reálnom čase, čím sa radikálne líši od predošlých technológií. Okrem existujúcich technológií sú poniektoré ešte stále v procese zdokonaľovania. Cieľom tohto prístupu je však dosiahnuť čo najrýchlejšie a finančne najnenáročnejšie sekvenovanie, ktoré bude prístupné pre akúkoľvek aplikáciu, či už v teréne alebo v laboratóriu [18].

Tab. 1 Prehľad parametrov sekvenačných technológií [9],[9],[18]

Technológia	***Sanger***	***Roche 454***	***Illumina***	***SOLiD***	***PacBio RS***	***MinION***
Dĺžka čítania	≈ 700 bp	≈700 bp	50-250 bp	≈ 75 bp	≈ 5000 bp (až 40 000 bp)	5-100+kbp
Presnosť	99,9 %	99,9 %	98 %	99,9 %	85 %	85 %
Čas behu	20 min – 3 hod.	24 hod.	1 – 10 dní	1 – 2 týždne	30 min – 4 hod.	< 10 min
Cena /1Mbp	2400 $	10 $	0,05-0,15 $	0,13 $	0,13– 0,60 $	< 0,1 $

2.1 Prvá generácia sekvenovania

V sedemdesiatych rokoch sa začali nezávisle od seba rozvíjať dve metódy sekvenovania DNA, ktorých priebeh je zobrazený na Obr.1 . V prípade Sangerovej metódy, na obrázku vľavo, je DNA vzorka denaturovaná na dve vlákna a templát následne rozdelený do štyroch reakčných zmesí, z ktorých každá obsahuje enzým DNA polymerázu, primer, deoxynukleotidy (A, C, G, T) a jeden zo štyroch dideoxynukleotidov (ddGTP, ddATP, ddTTP, ddCTP), kde každý z nich je značený inou fluroescenčnou značkou a špecifické sú tiež neprítomnosťou hydroxylovej skupiny na 3‘ konci. Proces pokračuje syntézou a predlžovaním komplementárneho reťazca, ktorý je pozastavený náhodným začlenením dideoxynukleotidu, ktorý neumožňuje ďalšie predlžovanie vlákna. Maxam – Gilbertova metóda, na obrázku vpravo, využíva terminálne značené DNA fragmenty pomocou rádiofosforu na 5’ konci, ktoré sú opäť rozdelené do štyroch reakčných zmesí obsahujúcich špecifické chemikálie spôsobujúce štiepenie len v mieste výskytu konkrétnych báz [10]. V oboch metódach sú získané rôzne dlhé fragmenty DNA, ktorú sú nakoniec analyzované pomocou elektroforézy na polyakrylamidovom géle [7]. V súčasnosti je možné analyzovať naraz 384 sekvencií s dĺžkou 600 až 1000 nukleotidov [5].



Obr.1 Sangerova (vľavo) a Maxam – Gilbertova (vpravo) metóda sekvenovania [7]

2.2 Druhá generácia sekvenovania

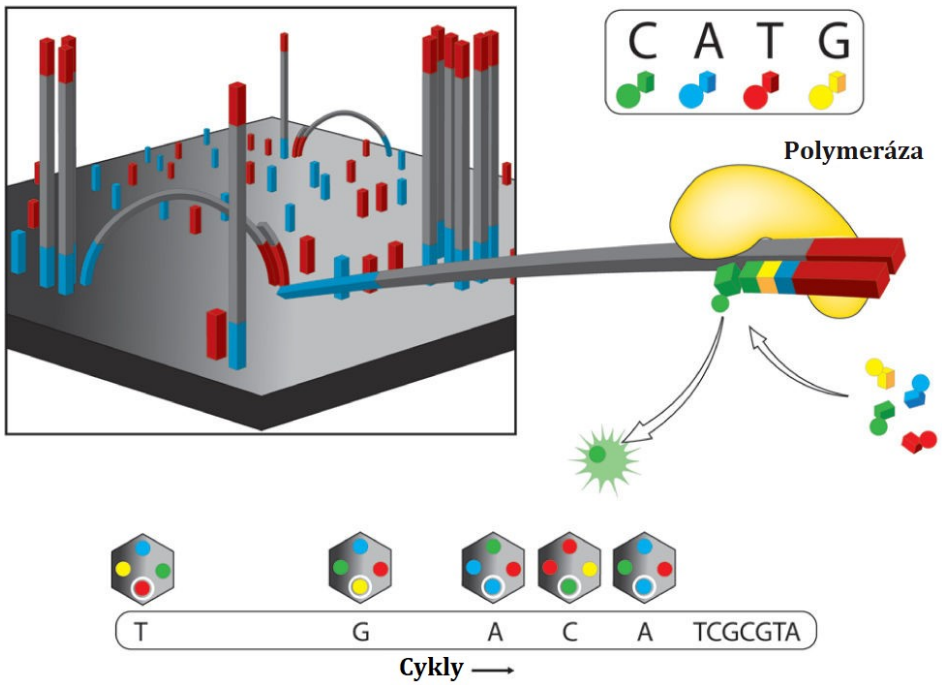
V roku 2005 sa začalo rozvíjať viacero platforiem, ako napríklad Roche/454, SOLiD a Illumina, ktoré využívajú masívne paralelné sekvenovanie. Táto vlna nového prístupu začala byť nazývaná ako druhá alebo nová generácia sekvenovania Ako bolo spomenuté v úvode, revolúcia so sebou prináša zefektívnenie výkonnosti a rýchlosti sekvenovania. Napriek odlišným princípom majú spomínané nástroje sekvenovania spoločné charakteristické postupy, medzi ktoré patrí určitá príprava vzoriek na sekvenovanie, teda vytvorenie knižnice nafragmentovanej DNA a jej amplifikácia. Ďalším spoločným znakom je priebeh a detekcia jednotlivých sekvenačných reakcií, ktoré sú uskutočnené automaticky. Výsledkom týchto reakcií je obrovské množstvo tzv. čítaní (z angl. reads), ktoré sú následne analyzované. Proces analýzy zahŕňa vyvolávanie báz, porovnávanie získaných sekvencií s referenčnými sekvenciami a interpretáciu získaných výsledkov [11].

Roche 454 pyrosekvenovanie

Pyrosekvenovanie predstavuje metódu založenú na princípe sekvenovania syntézou a detekcie uvoľňovaného pyrofosfátu počas syntézy DNA. Celý proces začína nafragmentovaním DNA, ligáciou adaptorov, denaturáciou dvojvláknovej DNA na jednovláknovú, ktorá nasadá na nanoguličku, kde následne prebieha emulzná PCR. Nanoguličky, na ktorých je až 10 miliónov kópií rovnakej DNA, sú dodatočne nanesené do jamiek na pikotitračnej doštičke. Medzi komponenty následnej reakcie patrí primer, ktorý je hybridizovaný k DNA templátu, DNA polymeráza, ATP sulfuryláza, luciferáza, apyráza a substráty adenozín 5‘fosfosulfát (APS) a luciferín. Cyklus každého zo štyroch deoxynukleotidov, ktoré sú pridávané do reakčnej zmesi, prebehne samostatne. Po každom zabudovaní nukleotidu pomocou DNA polymerázy dochádza k uvoľneniu pyrofosfátu, ktorého množstvo je ekvivalentné počtu začlenených nukleotidov. Uvoľnený pyrofosfát je následne konvertovaný do ATP pomocou ATP sulfurylázy za prítomnosti APS. Generovaná ATP aktivuje enzým luciferázu, ktorá sprostredkúva konverziu luciferínu na oxoluciferín za produkcie viditeľného svetla, ktorého intenzita je úmerná počtu molekúl ATP, teda počtu začlenených nukleotidov. Emitované svetlo je dodatočne snímané fotónovým detektorom, napríklad CCD kamerou a konvertované na signál, tzv. pyrogram, ktorého špičky sú úmerné počtu začlenených nukleotidov [12].

Illumina sekvenovanie

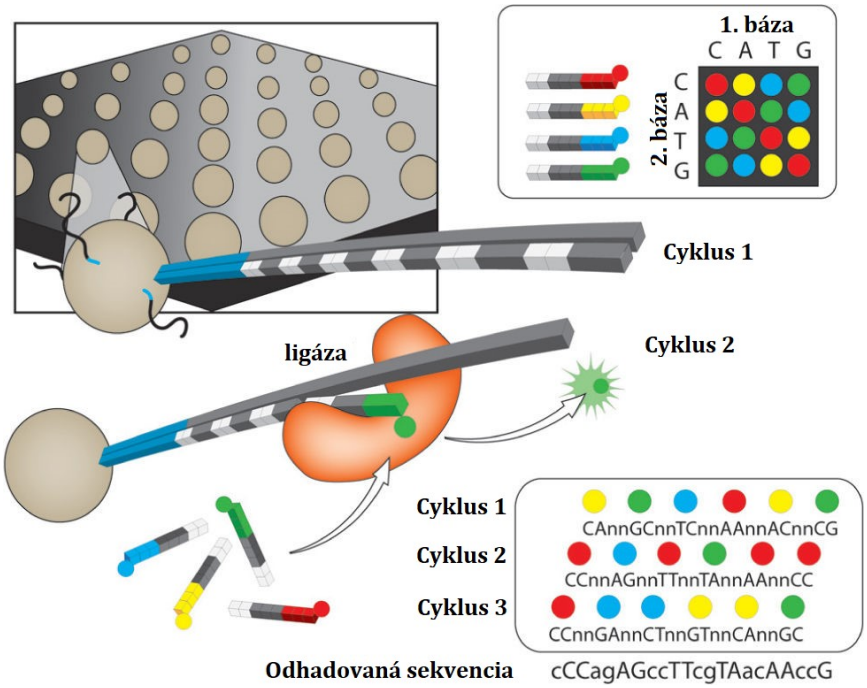
Táto technológia je taktiež založená na princípe sekvenovania syntézou, avšak v kombinácii s mostíkovou PCR amplifikáciou. Opäť prebieha fragmentácia DNA na veľkosť približne 800 báz, ligácia adaptorov k obom koncom fragmentov a následná denaturácia. Výsledná jednovláknová DNA ďalej nasadá na pevný povrch reakčnej komôrky (z angl. flow cell) husto pokrytý sondami, komplementárnymi k adaptorom, kde prebieha spomínaná mostíková PCR a vytvorenie zhlukov jednovláknových DNA, ako je možné vidieť na Obr.2 . Reverzné vlákna sa odštiepia a do reakčnej komôrky sú pridané sekvenačné primery a v každom novom cykle DNA polymeráza, deoxynukleotidy s farebným označením a tzv. reverzibilné terminátory, vďaka ktorým je v každom z cyklov naviazaný iba jeden nukleotid. Po začlenení nukleotidu a excitácii laserom je emitovaná fluorescencia zachytená CCD kamerou. Pred ďalším cyklom sú terminátory s fluorescenčným značením odstránené a cyklus sa opakuje, aby bolo možné určiť celú sekvenciu báz fragmentov [12].



Obr.2 Illumina sekvenovanie s názornou mostíkovou PCR [18]

SOLiD sekvenovanie

V porovnaní s ostatnými platformami, je táto technológia založená na princípe sekvenovania ligáciou a vyznačuje sa vysokou presnosťou. Proces začína prípravou jednej z dvoch typov knižníc, ktorými sú fragmentovaná, kedy získame fragmenty s adaptormi na oboch koncoch a párová (z angl. mate - pair), ktorej výstupom sú fragmenty obsahujúce okrem koncových aj jeden interný adaptor. Nasleduje naviazanie jednovláknovej DNA na magnetickú guličku, na ktorej prebehne emulzná PCR. Gulička s mnohými jednovláknovými DNA nasadá na sklenený povrch reakčnej komôrky. Ďalším krokom je naviazanie primeru na adaptor fragmentu. Súčasťou procesu je aj súbor štyroch fluorescenčne značených sond v podobe oktamérov, ktoré súperia o začlenenie za primer alebo predchádzajúci oligonukleotid. Každá sonda obsahuje dva špecifické nukleotidy, tri ľubovoľné a tri ako reverzibilný terminátor so špecifickou farebnou značkou podľa prvých dvoch špecifických nukleotidov. Na Obr.3 je zobrazený cyklus ligácie, detekcie fluorescenčného signálu a odštiepenia posledných troch nukleotidov, ktorý sa opakuje niekoľkokrát v závislosti od dĺžky čítania. Po dosiahnutí tejto dĺžky sú všetky oligonukleotidy odstránené a celé čítanie sa opakuje s primerom vždy o jeden nukleotid kratší, aby bola postupne prečítaná celá sekvencia. Dekódovanie, ako je opäť možné vidieť na Obr.3 , prebieha po dvoch nukleotidoch, pričom je potrebné poznať prvý nukleotid, ktorý je známy vďaka predom známej sekvencie primeru [12].



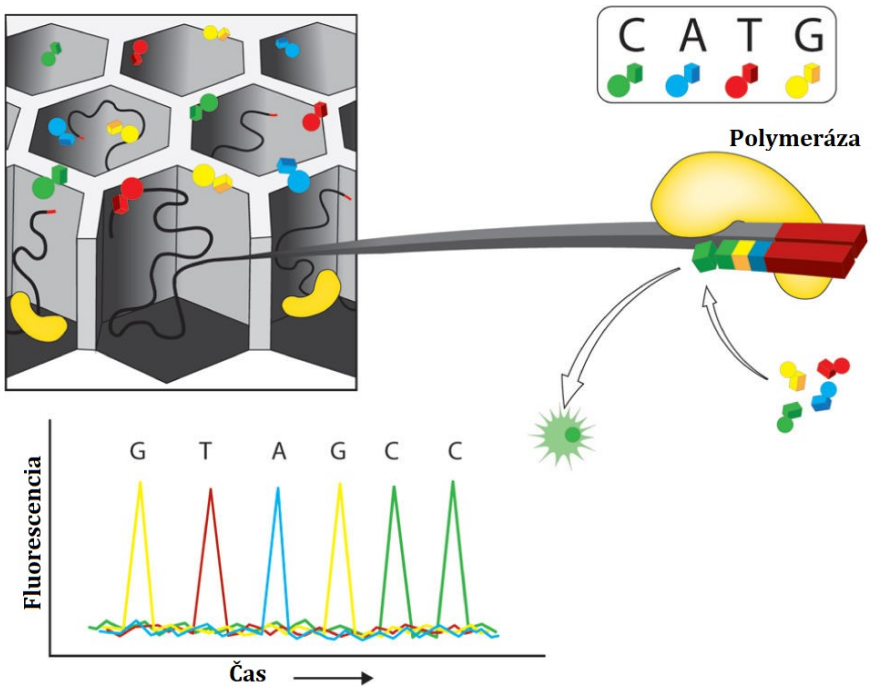
Obr.3 SOLiD sekvenovanie s názorným dekódovaním [18]

2.3 Tretia generácia sekvenovania

Dopyt po technológiách, ktoré by boli schopné pracovať rýchlejšie a produkovali by dlhé čítania vyústil v príchod tretej generácie sekvencovania (TGS, z angl. Third-Generation Sequencing). V porovnaní s NGS sa metódy tretej generácie zameriavajú na jednotlivé molekuly DNA bez nutnosti ich amplifikácie, kde jednotlivé čítania sú dostupné k analýze hneď ako prejdú daným sekvenátorom. Táto generácia poskytuje niekoľko základných pokrokov. Nárast dĺžky čítaní z desiatok báz na desaťtisíc báz na jedno čítanie. Čas sekvenovania sa zredukoval z dní na hodiny či minúty a taktiež dochádza k pomerne významnej eliminácii skresľovania pri sekvenovaní, čo bolo v predošlých metódach, okrem iného, zapríčinené používanou PCR amplifikáciou [8].

PacBio sekvenovanie

Jednou z prvých metód TGS sa stalo SMRT (z angl. Single-Molecule Real-Time) sekvenovanie jednotlivých molekúl v reálnom čase, uvedené spoločnosťou Pacific Bioscience [13]. Templátová DNA (tzv. SMRT bell) je uzavretá, jednovláknová cirkulárna DNA vytvorená ligáciou spínacích adaptorov na oba konce cieľovej dvojvláknovej DNA, čím je umožnené sekvenovanie oboch vláken. Proces, zobrazený na Obr.4 , začína nanesením vzorky molekuly DNA na čip (tzv. SMRT cell) pokrytý nanoštruktúrovým materiálom.

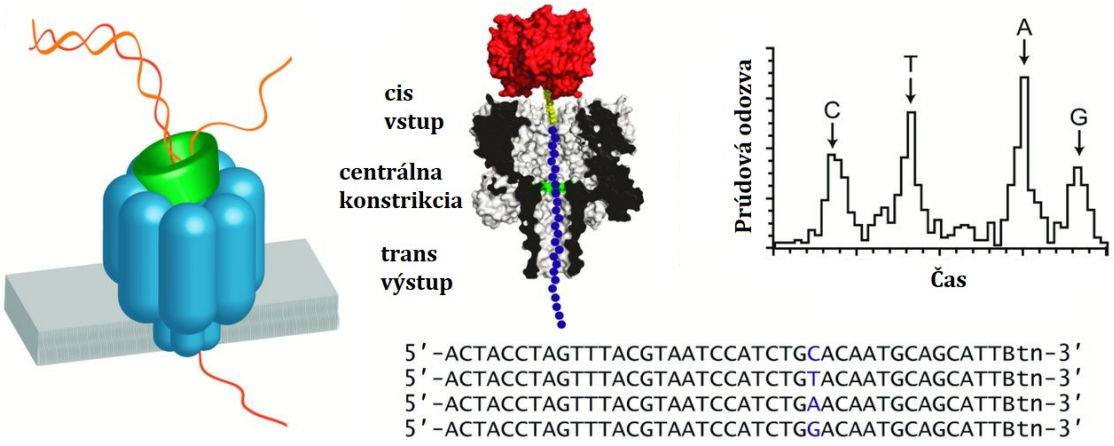


Obr.4 Metóda sekvenovania jednotlivých molekúl v čase [18]

Materiál obsahuje jamky (tzv. ZMW, Zero-Mode Waveguides) s hĺbkou približne 100 nm a priemere 70 nm, do ktorých molekula difunduje. ZMW následne poskytujú detekciu svetla, ktorého vlnová dĺžka je väčšia než samotný objem jamky a taktiež umožňujú sledovať objem o veľkosti vždy maximálne jedného nukleotidu. Na dne každej jamky je ukotvená DNA polymeráza poskytujúca naviazanie spínacieho adaptora a takisto podnecuje začiatok replikácie. K syntéze vlákna sú použité štyri fluorescenčne značené nukleotidy, ktoré pri prechode polymerázou produkujú svetelný impulz indetifikujúci konkrétnu bázu [15].

Nanopórové sekvenovanie

Sekvenovanie pomocou nanopórov, malých otvorov umiestnených v elektricky rezistentnej membráne, ponúka mnoho výhod v porovnaní s už existujúcimi technológiami a je objektom veľkého záujmu súčasnej vedy. Zariadenia pracujúce na princípe nanopóru poskytujú detekciu jednotlivých molekúl a schopnosť analyzovať ich. Inšpiráciou sa stali iónové kanály biomolekúl, ktoré riadia takmer všetky procesy v bunke. Tento nový prístup závisí na elektroforetickej translokácii analytu alebo jeho základných komponentov prostredníctvom tohto miniatúrneho otvoru. Analytom býva vo väčšine prípadov molekula DNA, pričom jej prechodom daným pórom, ako na Obr.5 , dochádza k detegovaniu báz na základe ich efektu na elektrický prúd alebo optický signál. Technológia pracuje s jednotlivými molekulami nemodifikovanej DNA, je pomerne finančne nenáročná a ponúka analýzu v reálnom čase [15].



Obr.5 Sekvenovanie pomocou nanopóru a príslušné výstupné dáta [15]

Spoločnosti Oxford Nanopore sa podarilo komercializovať systém na sekvenovanie DNA, ktorého základom sú tri biologické molekuly zostavené tak, aby pracovali ako jednotný mechanizmus. Biologický nanopór pozostáva predovšetkým z modifikovaného α-hemolyzínu, čo je vo vode rozpustný proteín schopný samovoľne oligomerizovať v lipidových membránach a vytvoriť v nich heptamerný kanál.

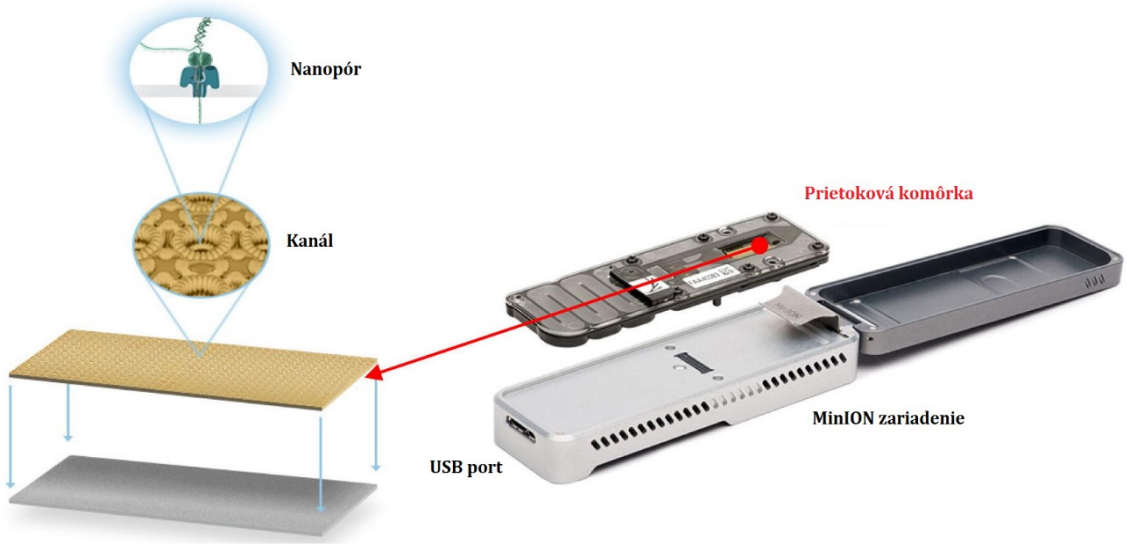
Jeho súčasťou je aj exonukleáza, napojená na extracelulárnej strane póru a syntetický senzor cyklodextrín, kovelentne pripojený k vnútornému povrchu nanopóru. Tento systém je následne súčasťou syntetickej lipidovej membrány, ktorá separuje dva oddiely vyplnené elektrolytom. Po aplikovaní napätia na membránu dochádza k prechodu iónov elektrolytu daným pórom. Konštantný jednosmerný prúd je v tomto prípade považovaný za referenčný. Po prechode analyzovanej molekuly z cis na trans stranu membrány, ako je opäť možné vidieť na Obr.5 , dochádza k čiastočnej zmene jej terciárnej štruktúry a takisto sa mení počet iónov prechádzajúcich pórom, čím dochádza k značnej zmene detegovaného prúdu voči prúdu referenčnému. Zmeny tohto elektrického prúdu, v podobe odporových pulzov, sú zachytávané pomerne senzitívnym zosilňovačom používaným napríklad v patch clamp technikách [17].

3. SEKVENOVANIE TECHNOLÓGIAMI OXFORD NANOPORE

Na základe informácií dostupných z ich internetovej stránky sa spoločnosť Oxford Nanopore Technologies (ONT) snaží o určitý prevrat v oblasti biologickej analýzy už od roku 2005, kedy bola založená. Cieľom tejto spoločnosti je sprístupniť možnosti analýzy DNA prostredníctvom sekvenovania aj tým, ktorí doposiaľ nemali možnosť pracovať so žiadnou zo sekvenačných techník. Snahou je proces zjednodušiť na úroveň, kedy využitím ich technológií bude možné kýmkoľvek analyzovať akékoľvek zvolené vzorky v príslušnom prostredí, či už v laboratóriu alebo v teréne. Ponúkaným zariadeniam od spoločnosti ONT sú venované nasledujúce kapitoly.

3.1 Zariadenie MinION

Myšlienkou sekvenovania DNA pomocou biologického nanopóru sa vedci začali zaoberať už približne pred dvadsiatimi rokmi. Základom tejto metódy je priama elektrická detekcia jednovláknovej DNA, ktorá prichádza do kontaktu s nanopórom. Sekvenované môžu byť extrémne dlhé fragmenty vo veľmi vysokej kvalite, pretože v procese dochádza k detekcii jednotlivých molekúl a vynechaný je krok predchádzajúcej amplifikácie, ktorá býva častým zdrojom skreslenia signálu. V roku 2014 bolo ONT spoločnosťou predstavené prvé komerčne dostupné zariadenie MinION, ktoré je možné vidieť na Obr.6 , umožňujúce nanopórové sekvenovanie.

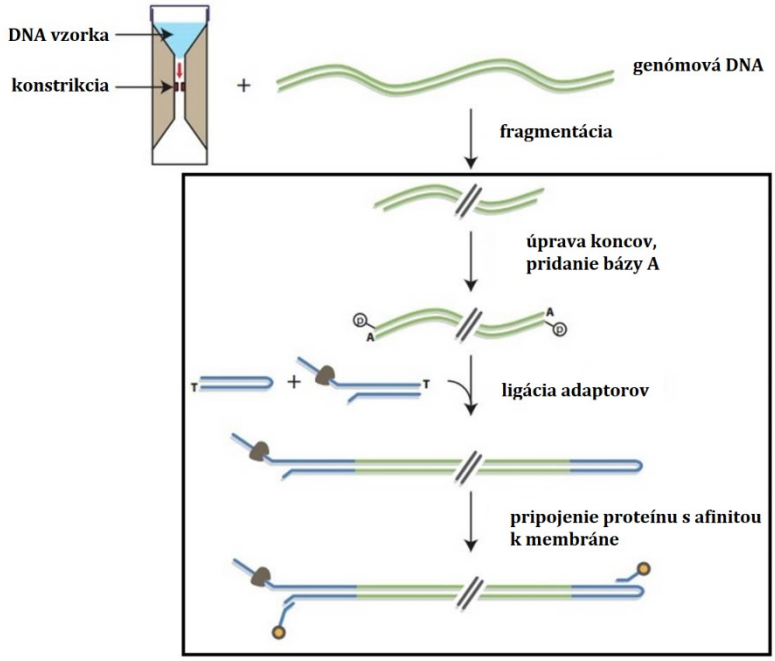


Obr.6 Zariadenie MinION a jeho komponenty [21]

Prenosnosť, malé rozmery, pripojenie cez USB k štandardnému počítaču alebo laptopu s internetovým pripojením predstavuje len jednu z výhod tohto pozoruhodného zariadenia, ktoré priťahuje značnú pozornosť v oblasti genomiky, kde je poskytované sekvenovanie v reálnom čase obrovskou výhodou [19].

3.1.1 Príprava knižnice

Proces prípravy knižnice, zobrazený na Obr.7 , začína nasekaním genómovej DNA použitím zariadenia Covaris g - TUBE na zvolenú dĺžku fragmentov. Následne dochádza k vytvoreniu tupých koncov nasekanej DNA a na 3‘ konce fragmentov je pridaná báza adenozín, aby bolo možné uskutočniť ďalší krok, ktorým je ligácia adaptérov.

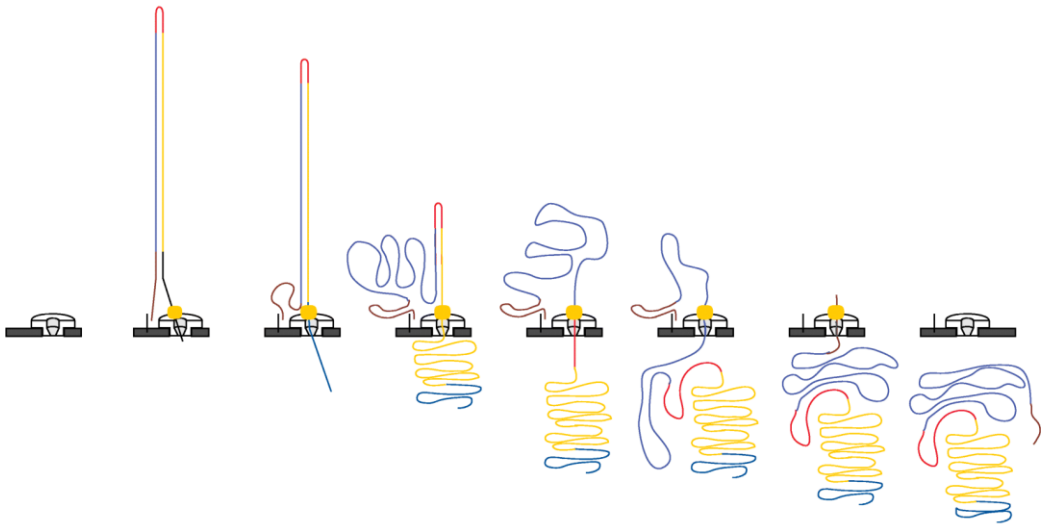


Obr.7 Príprava knižnice DNA [27]

Knižnica vo väčšine prípadov obsahuje dva typy adaptérov, hlavný a spínací. Hlavný adaptér, pozostávajúci z dvoch oligonukleotidov s čiastočnou komplementárnosťou, býva tiež označovaný ako Y adaptér vďaka svojej tvarovej štruktúre. Spínací adaptér, ktorý je tvorený jedným oligonukleotidom s vnútornou komplementárnosťou k vytvoreniu spínacej štruktúry, je označovaný ako HP adaptér. Oba adaptéry sú naviac doplnené motorovými proteínmi, ktoré sprostredkúvajú prechod DNA nanopórom. Funkciou adaptérov je nasmerovanie fragmentov DNA do blízkeho okolia pórov pomocou naviazaných väzbových oligonukleotidov, ktoré majú afinitu k polymérovej membráne. Úlohou spínacieho adaptéra je predovšetkým kovalentné spojenie vláken DNA, čím umožňuje plynulé sekvenovanie dvojvláknovej molekuly [20].

3.1.2 Proces sekvenovania DNA

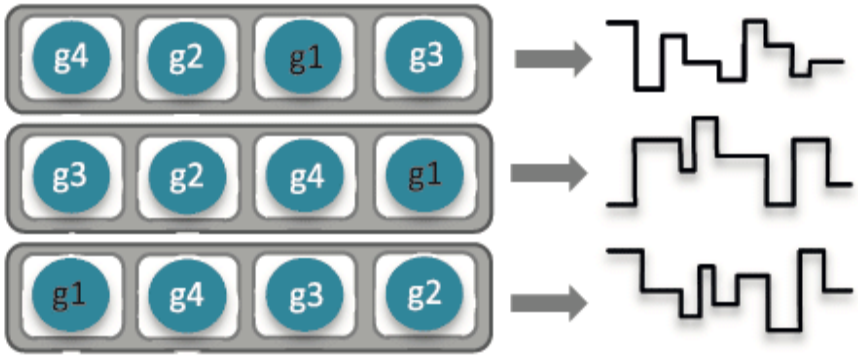
Z pripravenej knižnice sú molekuly DNA koncentrované blízko nanopóru, ktorý je začlenený do nevodivej membrány. Napätie priložené k membráne následne indukuje v nanopóre prúd, ktorý je predmetom záujmu. Celý proces sekvenovania, zobrazený na Obr.8 , začína na jednovláknovom 5‘ konci od hlavného adaptéra, za ktorým nasleduje templátové vlákno, spínací adaptér a následne vlákno komplementárne. Motorový proteín začína rozvoľňovať dvojvláknovú DNA v momente, kedy narazí na komplementárnu oblasť hlavného adaptéra a do nanopóru teda najprv vstupuje templátové vlákno. Nanopórom prechádzajú postupne všetky bázy vlákna rýchlosťou závislou od spomínaného motorového proteínu. Po dosiahnutí spínacieho adaptéra vstupuje pomocou HP motorového proteínu do nanopóru podobným spôsobom aj komplementárne vlákno. Pri prechode molekuly prostredníctvom póru dochádza k zmenám prúdu, ktoré sú charakteristické pre jednotlivé bázy [20].



Obr.8 Prechod molekuly DNA nanopórom [25]

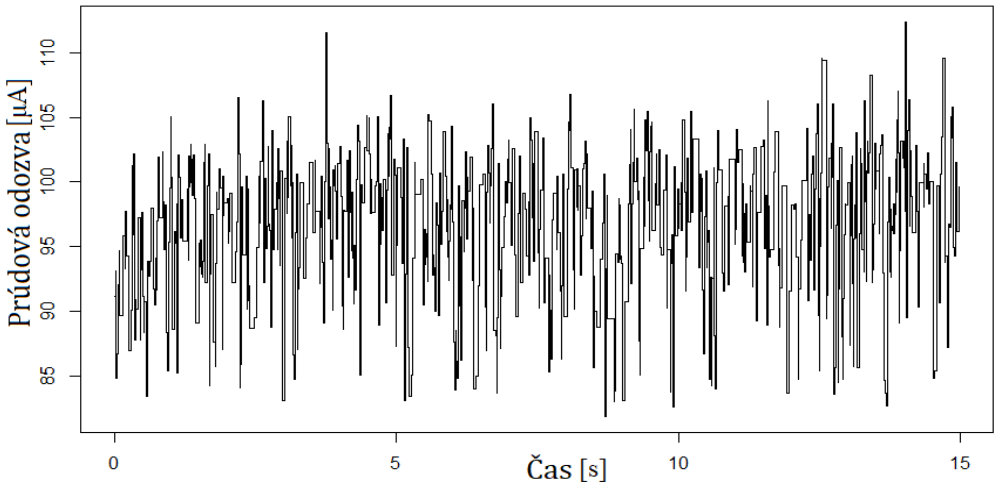
Prúd v nanopóre je meraný pomocou senzora niekoľko tisíc krát za sekundu a produkované dáta smerujú do mikročipu ASIC (z angl. Application Specific Integrated Circuit), ktorý predstavuje integrovaný obvod navrhovaný a vyrábaný pre špecifickú aplikáciu [21]. Nakoniec sú dáta spracované pomocou MinKNOW softvéru, ktorý sa podieľa na zhromaždení a analýze vyprodukovaných dát. Prietoková komôrka (z angl. flowcell) zariadenia MinION, zobrazená na Obr.9 , obsahuje 512 kanálov (šedá), to znamená, že súčasne môže byť sekvenovaných až 512 nezávislých DNA molekúl. Každý kanál naviac obsahuje štyri jamky (biela) a každá z nich obsahuje pór (modrá) a senzor (g1, g2, g3, g4). Kanál je schopný v určitom momente poskytovať dáta z jednej z týchto štyroch jamiek (v tomto prípade je to jamka g1) .

Z hľadiska výkonnosti, teda z hľadiska počtu produkovaných čítaní, sa jednotlivé kanály značne líšia, pretože každý pór môže byť viac či menej aktívny v porovnaní s ostatnými [20].



Obr.9 Prietoková komôrka (z angl. flowcell) s produkovanými dátami [20]

V súčasnosti sú vyvinuté dve významné chémie používané v prietokových komôrkach. Jedná sa o chémiu R7 a R9, ktoré sa predovšetkým líšia v použití konkrétneho póru a od toho sa odvíjajúcich parametrov. Nedávno predstavená R9 chémia sa v porovnaní s nahradenou R7 chémiou vyznačuje väčšou presnosťou, rýchlejším sekvenovaním a generovaním väčšieho množstva dát za jednotku času. Uprednostňuje sa čítanie iba jedného vlákna molekuly DNA, namiesto pomocou spínacieho adaptora spojeného templátu a komplementu. Dosiahnutie vyššej presnosti je spôsobené aj použitím algoritmov na vyvolávanie báz založených na neurónových sieťach, oproti menej presným skrytým Markovovym modelom. Dôvodom získavania oveľa väčšieho množstva dát je potreba sekvenovať iba jedno vlákno molekuly DNA a spomínaná rýchlosť, ktorá sa posunula z hodnoty 70 bps (báz za sekundu) na hodnotu 250 bps [23]. Na Obr.10 sú do grafu (tzv. squiggle plot) vynesené namerané surové dáta v závislosti na čase.



Obr.10 Vyprodukované surové dáta (tzv. squiggle plot) zariadenia MinION

V prípade R7 chémie sú získané surové dáta meraného prúdu dodatočne spracované a konvertované do sekvencie jednotlivých udalostí (tzv. events), ktoré obsahujú informácie o priemernej hodnote prúdu, príslušnej odchýlke a dĺžke trvania [21].

3.1.3 Formát produkovaných dát a vyvolávanie báz

Pre analýzu dát sa v bioinformatike vo väčšine prípadov používa FASTA formát, obsahujúci popis sekvencie a samotnú sekvenciu v znakovej forme, alebo FASTQ formát, naviac obsahujúci parameter popisujúci kvalitu vyvolaných báz. V poslednom období sa pre niektoré aplikácie stávajú však užitočnejšími surové dáta. Výstupom zariadenia MinION je súbor FAST5 z každého čítania. Podobne ako štandardný dátový formát HDF5, aj FAST5 súborový formát je založený na hierarchickom usporiadaní, v ktorom sú skladované metadáta odpovedajúce čítaniu produkovanému jedným z 512 kanálov zariadenia a taktiež udalosti (tzv. events) obsahujúce napríklad dáta o meranom prúde predspracované sekvenátorom [21].

Tento výklad sa zaoberá predovšetkým chémiou R7, keďže tieto dáta sú najviac dostupné a v praktickej časti bude zahrnutá práca s dátami práve v tomto formáte. V tom prípade je vyvolávanie báz (tzv. base-calling) uskutočnené pomocou sofvéru Metrichor, vyžadujúci prístup internetu (tzv. cloud-based software), ktorý je poskytovaný spoločnosťou ONT. Na určenie báz sekvencií z údajov o prúdových zmenách, ktoré poskytuje zariadenie MinION, využíva tento softvér skryté Markovove modely (z angl. Hidden Markov Model, HMM) [24]. Užívateľ používa tento softvér na rovnakom počítači ako zariadenie MinION a jeho úlohou je len načítať súbory FAST5, obsahujúce surový signál, na vzdialený server. Po chvíli sú užívateľovi k dispozícii na stiahnutie súbory FAST5 s už vyvolanými bázami. Využívanie tohto softvéru si však vyžaduje neustály prístup k internetu a taktiež má uzavretý zdrojový kód (tzv. closed source). V zámere poskytnúť otvorený zdrojový kód (tzv. open source) a vyvolávanie báz bez nutnosti pripojenia k internetu bol následne vyvinutý softvér Nanocall. Takisto využíva HMM a poskytuje postačujúce 1D vyvolávanie báz s presnosťou porovnateľnou k sofvéru Metrichor. Ďalšou platformou je Deep - Nano, rekurentná neurónová sieť, ktorá vykazuje pomerne vyššiu presnosť ako softvéry využívajúce HMM [25]. V porovnaní s ostatnými technológiami sa dáta poskytnuté zariadením MinION vyznačujú väčšou nepresnosťou, no napriek tomu sa pomocou tohto zariadenia podarilo detegovať už viacero vírusov, ako napríklad Ebola [28] a Zika [29]. To znamená, že poskytovaná kvalita je vo viacerých prípadoch k identifikácii vírusov, či baktérií postačujúca [30].

3.2 Zariadenie PromethION

Toto stolné zariadenie je príbuzné zariadeniu MinION, ktorému boli venované predošlé podkapitoly. PromethION vlastní 48 samostatných prietokových komôrok, kde každá z nich obsahuje 12-tisíc jamiek, v ktorej je umiestnených 3-tisíc nanopórov. Využívaných môže byť dokopy teda neuveriteľných 144-tisíc nanopórov, takže už nie je potrebné dávkovanie vzoriek, keďže užívateľ môže využiť ľubovoľný počet komôrok a každá z nich môže byť spustená a pozastavená bez narúšania tých zvyšných. Príprava vzoriek je zhodná s prípravou vzoriek pre zariadenie MinION a spracovanie dát, vrátane vyvolávania báz v reálnom čase, je uskutočnené pomocou zabudovaného výpočtového zdroja v systéme. Systém teda poskytuje veľký objem spracovaných dát z veľkého počtu spracovávaných vzoriek v reálnom čase [26].

4. VIZUALIZAČNÉ METÓDY PRE SPRACOVANIE METAGENOMICKÝCH DÁT

Vizualizácie dát sú zásadným nástrojom pri ich analýze, keďže je možné ich skúmať vizuálne, na základe čoho je možné vytvárať takmer okamžité hypotézy, ktoré sa stávajú ďalším predmetom analýzy. Zásadným problémom metagonomiky je správna klasifikácia DNA sekvencií podľa organizmov, z ktorých pochádzajú. V mnohých prípadoch sú sekvencie, pozostávajúce z postupnosti jednotlivých nukleotidov, charakterizované viac ako troma parametrami, to znamená, že nie je vždy jednoduché ich zrozumiteľne a ľudsky prezentovať. Keďže spracovanie dát v znakovej podobe je náročné, je vhodné takéto sekvencie najprv numericky reprezentovať, pričom parametre a pôvodná informácia sekvencie by mali byť zachované. Jedným z riešení spracovania výsledných numerických reprezentácii sa stali metódy vizualizácie dát založené na redukcii dimenzií, ktorým budú venované nasledujúce podkapitoly.

4.1 PCA - analýza hlavných komponentov

Analýza hlavných komponentov, PCA (Principal component analysis), je štatistická metóda využívajúca lineárnu transformáciu pôvodného súboru premenných do podstatne menšieho súboru nekorelovaných premenných, ktoré reprezentujú podstatnú časť informácie obsiahnutej v pôvodnom súbore. Cieľom je teda získať menší počet nekorelovaných premenných, jednoduchšie pochopiteľných, analyzovateľných a vizualizovateľných. Metóda využíva rozptyl ako nositeľa informácie, keďže jednotlivé pôvodné premenné majú v súbore istú variabilitu, ktorá je rozptylom meraná. V procese dochádza k projekcii objektov dát na novú os otočenú okolo počiatku súradnicového systému v smere maximálnej variability. V smere druhej najväčšej variability je vedená druhá os, ktorá je kolmá na prvú. Tieto osi sú definované hlavnými komponentami, ktoré sú lineárnou kombináciou pôvodných premenných, kde prvý z nich predstavuje najväčšiu variabilitu medzi všetkými lineárnymi kombináciami v danom súbore dát. Komponenty sú ortogonálne, teda vzájomne nekorelované, a vytvárané sú postupne s klesajúcim významom dôležitosti. Z toho dôvodu je pár týchto prvých komponentov najvýznamnejších, avšak aj niektoré ďalšie z nich môžu poskytovať relevantné informácie o štruktúre analyzovaných dát [31].

Podľa postupu [32], uvažujeme z matematického hľadiska n-rozmerný vektor získaných dát X, ktorý predstavuje jednu z dimenzií. Následne je od každej hodnoty Xi odčítaný priemer $\overline{X}$ daného vektoru, čím získame množinu dát, ktorých priemer je rovný nule.

$$X_i = X_i - \overline{X} \tag{1}$$

Z centrovaných dát je vypočítaná kovariančná matica. Kovarianca sa vždy počíta medzi dvoma veličinami, to znamená, že v prípade viacrozmerných dát je potrebné počítať viac kovariancií. Jeden zo spôsobov ako zapísať všetky kovariancie, je zapísať ich do matice. Všeobecný zápis kovariančnej matice *C* pre *n* veličín je :

$$C^{n\times n} = (c_{ij}, c_{ij,} = cov(Dim_i, Dim_j)). \tag{2}$$

Nasleduje výpočet vlastných vektorov a vlastných čísel z kovariančnej matice, kde vlastné čísla λ_i štvorcovej matice *A* rádu *n* sú koreňmi rovnice

$$\det(A - \lambda I) = 0, \tag{3}$$

v ktorej *I* predstavuje jednotkovú maticu. Ku každému vlastnému číslu λ_i existuje aspoň jedno nenulové riešenie sústavy rovníc

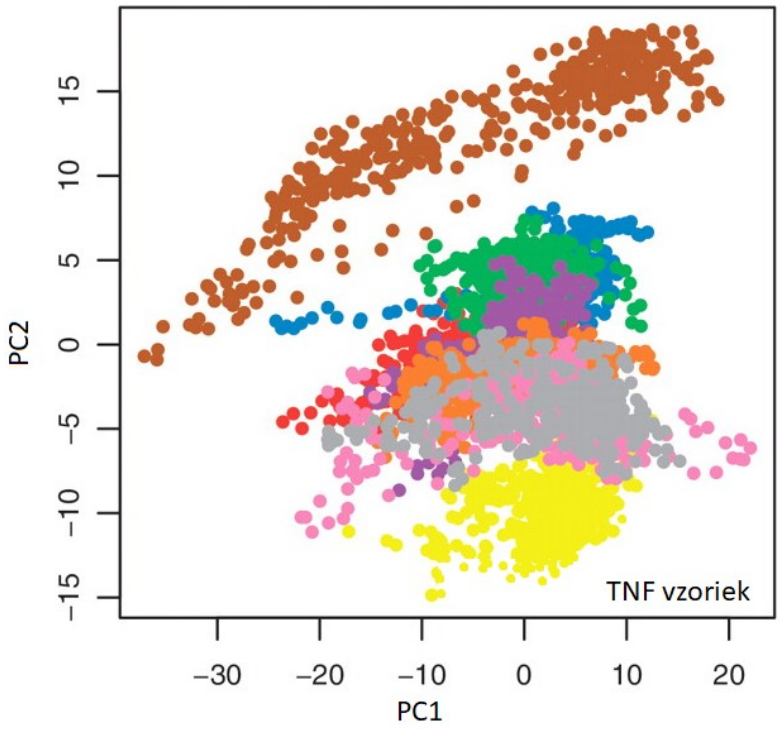
$$Ax = \lambda_i x, \tag{4}$$

v ktorej *x* je pravým vlastným vektorom matice *A*. Posledným krokom je výber hlavných komponentov, čím vznikne opisný vektor *O*, pomocou ktorého získame výsledné dáta *V* na základe rovnice

$$V = O^T \times X^T, \tag{5}$$

kde *X* sú dáta upravené podľa (1). Problémom metódy môže byť zanedbanie relevantných informácii, čím môže dôjsť k skresleniu alebo nedostatočnému charakterizovaniu dát. Výsledná interpretácia je závislá aj na zámere analýzy, ktorej cieľom môže byť odhalenie vzťahov medzi vzorkami alebo odhalenie vzťahov na úrovni druhov [33].

Príkladom využitia PCA pre binning metagenomických dát je metóda 2TBinning, kde sú kombinované viaceré parametre, ako napríklad CG obsah, parameter OFDEG (z angl. Oligonucleotide Frequency Derived Error Gradient) a frekvencia výskytu tetranukleotidov (TNF) náhodných fragmentov z vybranej skupiny genómov, ktorej projekciu pomocou PCA môžeme vidieť na Obr.11 . Z názvu možno odvodiť dvojstupňový prístup metódy, kde v prvom kroku triedenia sú sekvencie zhruba separované do skupín na základe obsahu CG a parametra OFDEG. V druhom kroku je k vylepšeniu roztriedenia použitá frekvencia výskytu tetranukleotidov a následným zhlukovaním sú skupiny rozdelené do menších oddielov [34].

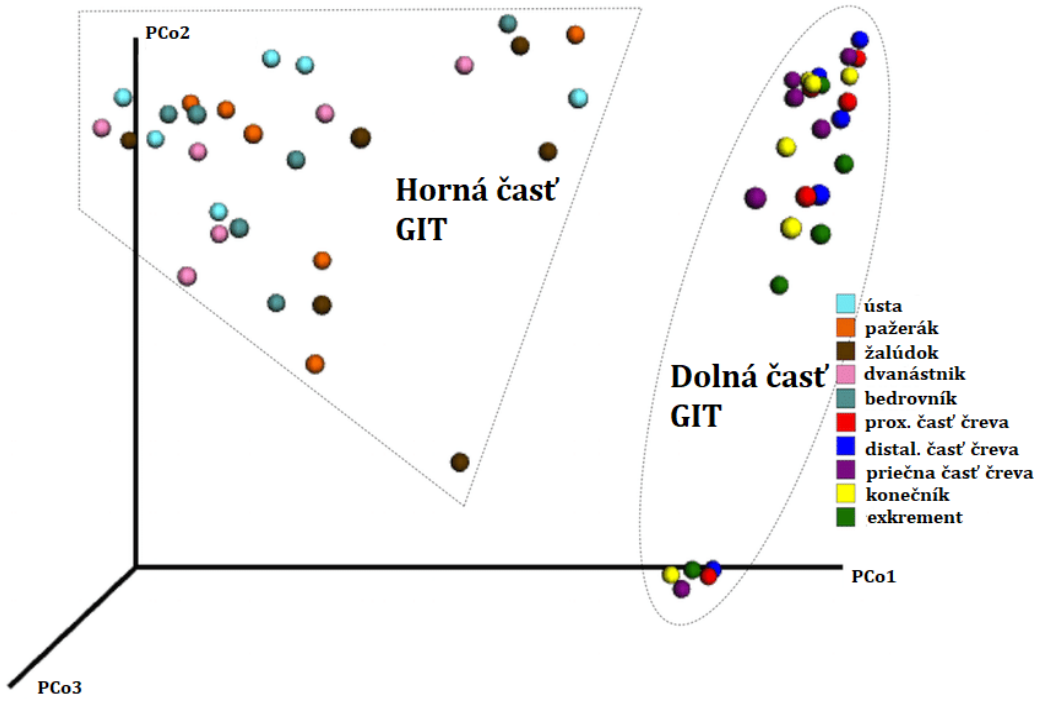


Obr.11 PCA projekcia TNF použitá pre binning metagenomických dát [34]

4.2 PCoA – analýza hlavných koordinátov

V prípade analýzy hlavných koordinát je euklidovská vzdialenosť v priestore stanovená hlavnými koordinátmi, ktoré sú aproximáciou vzdialeností medzi objektami v asociačnej matici odvodenej z pôvodnej matice. Asociačná matica predstavuje maticu vzdialeností vypočítaných pomocou ľubovoľných koeficientov, vyjadrujúcich vzťah, resp. odlišnosť medzi objektami. V metóde ide taktiež o akúsi konfiguráciu OTU (operačná taxonomická jednotka) v euklidovskom priestore za podmienok zachovania vzťahov medzi objektami vyskytujúcich sa v pôvodnej matici vzdialeností. Postup metódy je veľmi podobný metóde PCA, avšak koordináty vystupujúce v tejto metóde sú, v porovnaní s komponentami predstavujúcimi lineárnu kombináciu pôvodných premenných, komplexnými funkciami originálnych premenných sprostredkovaných prostredníctvom zvoleného spôsobu výpočtu odlišnosti objektov [33], [35].

Metóda PCoA bola použitá napríklad k testovaniu schopnosti odlišovať zdravých alebo kontaminovaných hostiteľov na základe prítomnosti alebo absencie určitých patogénnych génov [2]. Interpretácia výsledkov prebieha podobne ako v prípade PCA, v tomto prípade však výsledné objekty v súradnicovom systéme reprezentujú celý metagenóm. Na Obr.12 je možné vidieť grafický výstup PCoA pri skúmaní členov mikrobiálnej komunity GIT-u u myši, využitím amplikonového sekvenovania 16S rRNA [36].



Obr.12 Interpretácia výsledkov skúmania mikrobiálnej komunity za použitia PCoA [36]

Matematický postup [35] zahŕňa prácu s asociačnou maticou *D*, predstavujúcou maticu vzdialeností vypočítaných pomocou ľubovoľných koeficientov, vyjadrujúcich vzťah, resp. odlišnosť medzi pôvodnými objektami x_{jh} a x_{ji}. Použitá je napríklad euklidovská vzdialenosť

$$D_{mn} = \sqrt{\sum_j (x_{jm} - x_{jn})^2}. \tag{6}$$

Matica *D* je ďalej transformovaná do matice *A* podľa

$$A_{mn} = -\frac{1}{2} D^2{}_{mn}. \tag{7}$$

Takto získané dáta sú centrované k získaniu matice δ podľa

$$\delta_{mn} = A_{mn} - \overline{A_m} - \overline{A_n} + \overline{A}, \tag{8}$$

kde A_m predstavuje priemer hodnôt v riadkoch, A_n priemer hodnôt v stĺpcoch a $\overline{A}$ priemer hodnôt z celej matice *A*. Následne sú dopočítané vlastné vektory a vlastné čísla, podobne ako v metóde PCA a vybrané hlavné koordináty získané zoradením každého vlastného vektora u_k do dĺžky odpovedajúcej $\sqrt{\lambda_k}$, kde λ_k predstavuje kladné vlastné číslo asociované s príslušným vlastným vektorom.

4.3 NMDS – nemetrické mnohorozmerné škálovanie

NMDS je všeobecne veľmi efektívna metóda analýzy matíc podobností alebo vzdialeností. Na základe tejto analýzy sú vzťahy medzi objektami zobrazené v euklidovskom priestore, v ktorom sú, podobne ako v predošlých metódach, objekty reprezentované bodmi. Vstupom je matica súradníc a parameter o požadovanom výslednom počte dimenzií. Následne prebieha optimálne škálovanie a odhad nových parametrov, resp. výpočet súradníc. V poslednom bode je určená tzv. hodnota *Stress*, kde cieľom metódy je nájsť takú konfiguráciu bodov, ktorá bude túto hodnotu minimalizovať [33]. Hodnota je vypočítaná ako

$$Stress = \sqrt{\sum_{h,i} (d_{hi} - d_{hi}')^2 / \sum_{h,i} d_{hi}^2}, \tag{9}$$

kde d_{hi} je vzdialenosť medzi bodmi h a *i,* parameter d_{hi}' vzdialenosť predikovaná regresiou [37]. Blízkosť objektov v zobrazení odpovedá ich podobnosti, avšak nekorešponduje pôvodným vzdialenostiam medzi objektami.

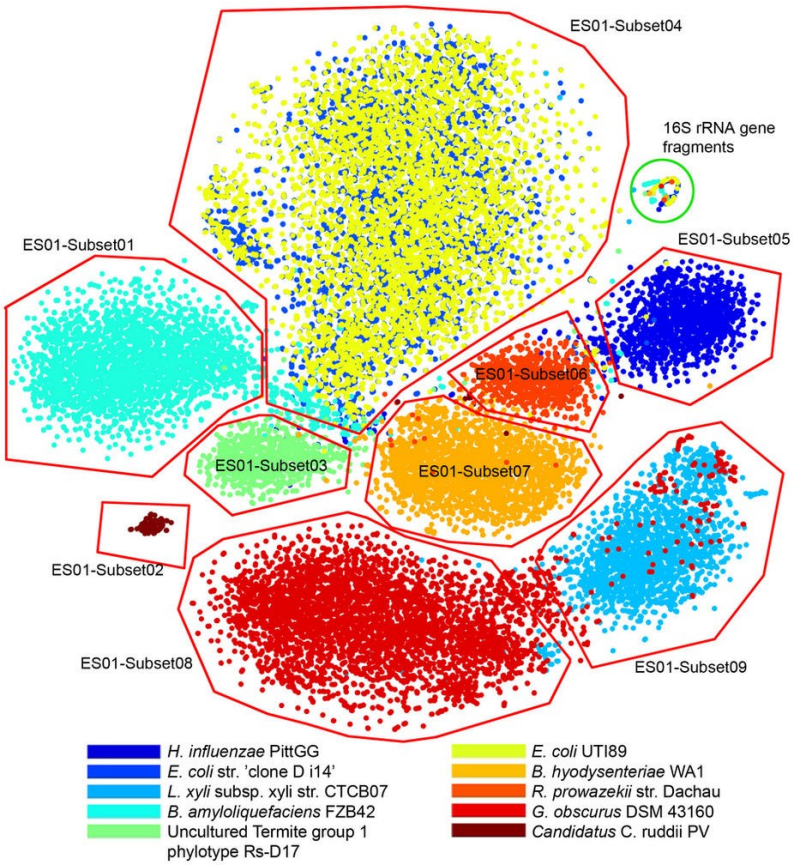
Vďaka zachovaniu poradia vzdialeností medzi objektami je možné podľa potreby osi zobrazenia ľubovoľne prepočítať, rotovať alebo obrátiť k získaniu lepšej vizualizácie alebo interpretácie. Najväčší problém metódy spočíva v správnom odhadnutí počtu dimenzií. Oproti predošlým metódam sa NMDS vyznačuje iteratívnym postupom vďaka čomu rastie aj výpočtová náročnosť [38]. Metóda bola napríklad použitá pri porovnávaní rozmanitosti znakov mikrobiálnych komunít zo vzoriek, ktoré podliehali rôznym postupom hospodárenia s pôdou [39].

4.4 t-SNE metóda

Metóda t-SNE (t-distributed Stochastic Neighbor Embedding) predstavuje metódu nelineárnej redukcie dimenzií a na rozdiel od PCA metódy nie je založená na hľadaní najväčšieho rozptylu v dátach. V tomto prípade je snahou zaistiť, aby objekty, ktoré sú si blízke v originálnom vysoko- dimenzionálnom priestore, boli blízke aj v priestore redukovanom. Oproti SNE metóde, z ktorej je táto metóda odvodená, je t-SNE jednoduchšie optimalizovateľná a poskytuje pomerne lepšiu vizualizáciu dát, pretože redukuje tendenciu nahromadenia bodov uprostred grafu. Metóda umožňuje vizualizáciu vysoko-dimenzionálnych dát, tým že jednotlivým bodom je priradená lokalita v 2D alebo 3D priestore. Vysoko- dimenzionálne dáta euklidovských vzdialeností medzi objektami sú konvertované do matice podobností, ktoré reprezentujú podobnosť dvoch bodov vo viac - dimenzionálnom priestore. Pôvodná SNE metóda využíva na určenie podobností Gaussovo rozdelenie, kde pre blízke body je hodnota relatívne vysoká a pre nízke bude nadobúdať nepatrné hodnoty. Metóda je však obmedzená chybovou funkciou, ktorá je veľmi ťažko optimalizovateľná. Chybová funkcia v prípade t-SNE metódy je symetrickou obdobou chybovej funkcie metódy SNE. Nový prístup využíva na výpočet podobností bodov v nízko-imenzionálnom priestore študentovo rozdelenie, ako tomu nasvedčuje názov, namiesto Gaussovho rozdelenia použitého v predošlej metóde [40].

Rozšírením tejto metódy je BH-SNE metóda využívajúca určitý Barnes-Hut algoritmus, vďaka ktorému sa tento prístup vyznačuje predovšetkým lepšou operačnou náročnosťou, ktorá závisí predovšetkým na množstve vstupných dát. Toto rozšírenie taktiež využíva stromovú dátovú štruktúru VP-stromy (Vantage Point-trees), ktorá k rozloženiu množiny dát využíva vzdialenosť jednotlivých objektov od určitého referenčného bodu na výpočet podobností objektov vstupných dát.

Tento nový algoritmus by mal disponovať viacerými výhodami v porovnaní s predošlou t-SNE metódou, pričom hlavnou z nich je urýchlenie algoritmu, vďaka ktorému je získaná možnosť spracovať obrovské dáta, ako sú napríklad metagenomické dáta získané shotgun sekvenovaním. Príklad takéhoto využitia je na Obr.13 a je vidieť, že metóda funguje, minimálne pre jednoduchšie metagenómy, takmer perfektne [41].



Obr.13 BH-SNE vizualizácia mikrobiálnej komunity simulovaného datasetu genomických fragmentov [41]

Na základe príspevku [42] sú z pôvodných dát vypočítané podmienené pravdepodobnosti $p_{j|i}$, zahŕňajúce výpočet vzdialeností medzi bodmi x_i a x_j za použitia euklidovskej vzdialenosti a rozptylu Gaussovského rozdelenia σ_i.

$$p_{j|i} = \frac{exp(-\left|\left|x_i - x_j\right|\right|^2 / 2\sigma_i^2)}{\sum_{k \neq i} exp(-\left|\left|x_i - x_j\right|\right|^2 / 2\sigma_i^2)} \tag{10}$$

Symetrizáciou dvoch podmienených pravdepodobností je získaná vzájomná podobnosť medzi objektami x_i a x_j v originálnom priestore podľa

$$p_{ji} = \frac{p_{j|i} + p_{j|i}}{2N}, \tag{11}$$

kde N predstavuje celkový počet objektov v tomto priestore. Po konvertovaní objektov do nízko-dimenzionálneho priestoru, je opäť vypočítaná matica podobností $q_{j|i}$ medzi bodmi y_i a y_j podľa

$$q_{j|i} = \frac{\exp(-\left|\left|y_i - \mathrm{y}\right|\right|^2)}{\sum_{k \neq i} \exp(-\left|\left|y_i - y_j\right|\right|^2)}, \tag{12}$$

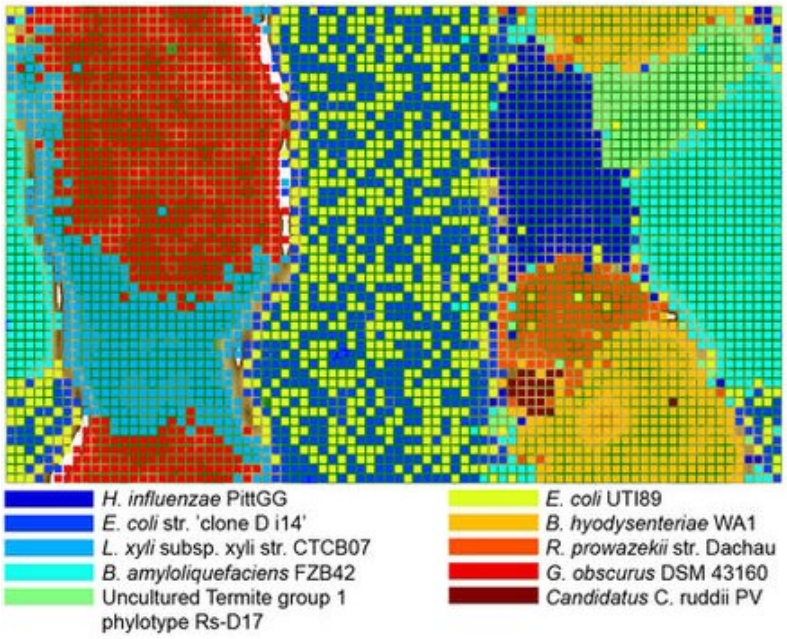
kde v tomto prípade je namiesto Gaussovho rozdelenia použité rozdelenie Študentovo s jedným stupňom voľnosti. Ideálne rozmiestnenie bodov y v redukovanom priestore je získané minimalizovaním hodnoty Kullbuck-Leiblerovej divergencie C medzi vzájomnou distribúciou P a Q podľa

$$C = \sum_i KL(P_i||Q_i) = \sum_i \sum_j p_{i|j} \log \frac{p_{i|j}}{q_{i|j}}. \tag{13}$$

4.5 ESOM – emergentné samo-organizujúce mapy

Emergentné samo-organizujúce mapy, tiež známe ako Kohenove mapy, patria medzi neurónové siete s učením bez učiteľa. Emergencia znamená, že takéto SOM umožňujú vytvoriť vnútorné štrukturálne vlastnosti dát v priestore. ESOM poskytujú premietanie vysoko-dimenzionálnych dát, do 2D mapy pri zachovaní topológie pôvodnej množiny dát. Metóda je výborná pri skúmaní dát a môže byť efektívne využitá pri ich vizualizácii, zhlukovaní a klasifikácii. Pri vizualizácii dát sa do úvahy berie hustota dát v okolí neurónov a taktiež ich vzájomná vzdialenosť. V prípade na Obr.14 je použitá tzv. U-Matrix vizualizácia využívajúca vzdialenosť, najčastejšie euklidovskú, príslušných neurónov.

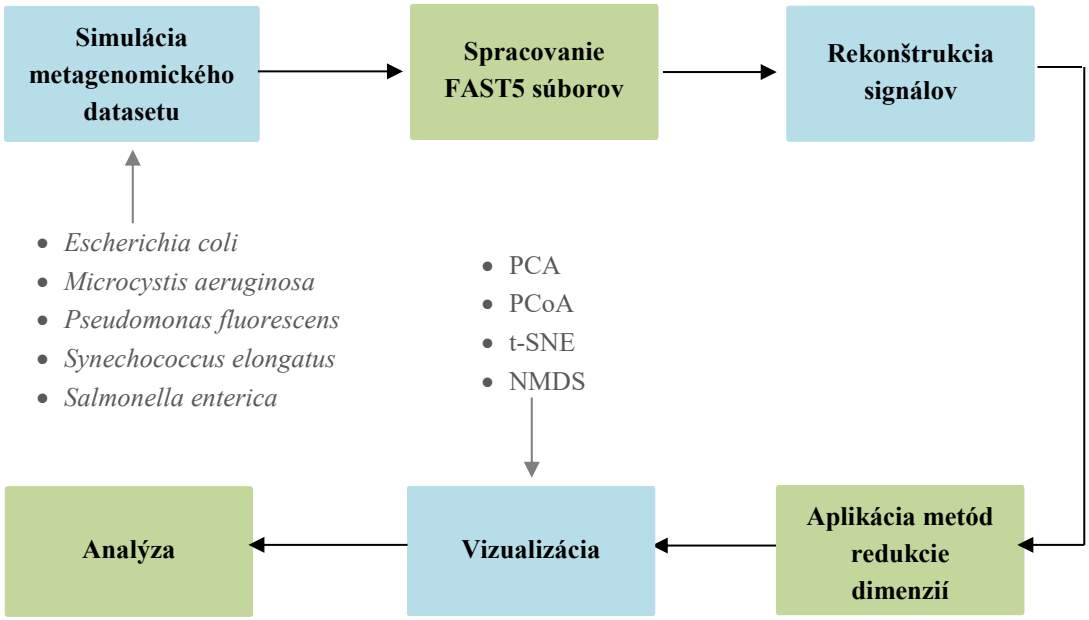
Výsledkom je topografické zobrazenie separovaných oblastí, ktoré v tomto prípade reprezentujú určité mikroorganizmy z tréningovej množiny zvolených genomických dát. Táto vizualizácia môže byť použitá aj k posúdeniu presnosti roztriedenia dát získaných genómov a taktiež k odhaleniu nových regiónov pozostávajúcich zo sekvencií určitých príznakov [41], [43].



Obr.14 ESOM vizualizácia mikrobiálnej komunity simulovaného datasetu genomických fragmentov [41]

5. APLIKÁCIA METÓD REDUKCIE DIMENZIÍ PRE VIZUALIZÁCIU METAGENOMICKÝCH DÁT

Cieľom tejto práce je primárne vytvoriť algoritmus, ktorý bude schopný uskutočniť binning dát na základe vektoru príznakov, odvodených z každej sekvencie, za použitia vhodných metód redukcie dimenzií. Jedná sa o nový prístup spracovania sekvenčných dát využívajúci aplikáciu metód redukcie dimenzií priamo na prúdové signály produkované nanopórovým sekvenovaním. Vo všeobecnosti ide o získanie vektorov signálových sekvencií, ktoré sú následne zobrazené v nízko-dimenzionálnom priestore. Využitím spomínaných metód je získavaný prístup vizualizovať obrovské súbory dát, odpovedajúce zmesi sekvencií z rôznych organizmov tak, že sekvencie pochádzajúce z rovnakého organizmu si sú vo výslednom redukovanom priestore blízke, naopak sekvencie pochádzajúce z rôznych organizmov by mali byť umiestnené ďaleko od seba. Pre názornosť je k dispozícii bloková schéma na Obr.15 , odpovedajúca sľubovanému spracovaniu dát. Aby daná metóda mohla byť otestovaná, v prvom rade je potrebné použiť simulovaný metagenóm, u ktorého je známe, ktorá sekvencia pochádza z ktorého organizmu. Po spracovaní FAST5 súborov získame dáta produkované nanopórovým sekvenovaním, ktoré však neobsahujú priamo meraný signál a preto je potrebné uskutočniť jeho rekonštrukciu. Po získaní týchto signálov je možné na nich aplikovať metódy založené na princípe redukcie dimenzií a následne dáta vizualizovať, čo nám umožní uskutočniť záverečnú analýzu.



Obr.15 Bloková schéma navrhovaného spôsobu spracovania dát

5.1 Voľba metagenomického datasetu

K spracovaniu bol použitý simulovaný metagenomický dataset z verejne dostupných nanopórových sekvencií osekvenovaných bakteriálnych genómov. Dataset pozostáva z dát získaných z archívu ENA (European Nucleotide Archive) a je v ňom zastúpených päť odlišných organizmov. Konkrétne sú to *Escherichia coli*, *Microcystis aeruginosa, Pseudomonas fluorescens, Synechococcus elongatus,* ktoré sú súčasťou metagenómu v archíve pod prístupovým číslom PRJEB8716, a taktiež *Salmonella enterica* s prístupovým číslom metagenómu PRJEB7205. V Tab. 2 sa nachádza prehľad doplňujúcich prístupových čísel ku konkrétnym dátam. V prílohe elektronickej verzie tejto práce sú dáta k dispozícii vo formáte FAST5.

Tab. 2 Prehľad prístupových čísel organizmov v zvolenom metagenomickom dataste

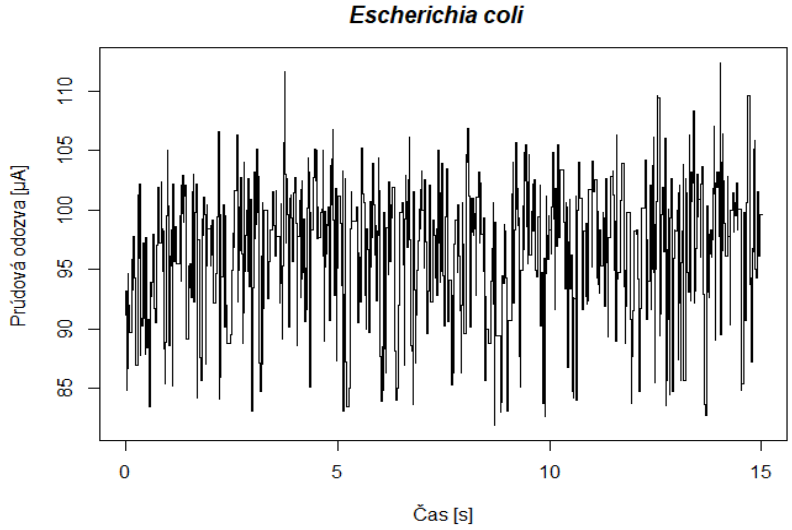
Názov organizmu	*Prístupové číslo vzorky*	*Prístupové číslo behu*
Escherichia coli	SAMEA4528013	ERR1713483
Microcystis aeruginosa	SAMEA4528016	ERR1713486
Pseudomonas fluorescens	SAMEA4528017	ERR1713487
Synechococcus elongatus	SAMEA4528019	ERR1713489
Salmonella enterica	SAMEA2813268	ERR776484

5.2 Predspracovanie dát

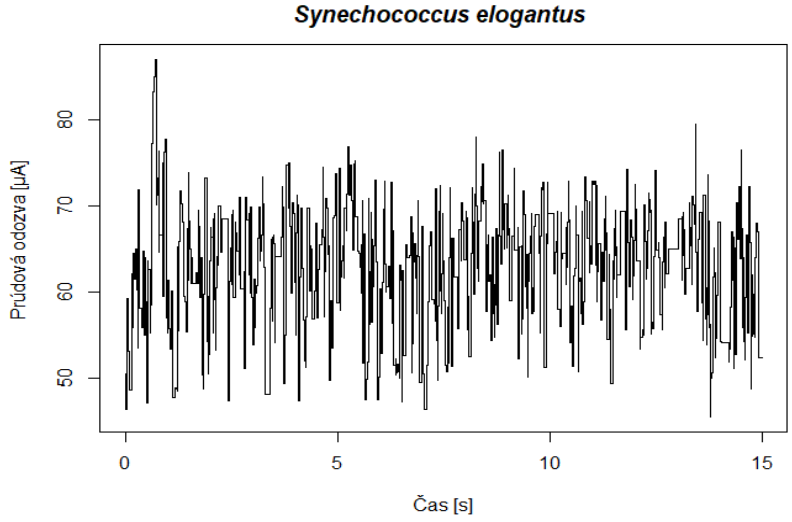
V programovacom prostredí R bol použitý balíček poRe umožňujúci načítanie a analýzu FAST5 súborov, získaných nanopórovým sekvenovaním pomocou zariadenia MinION. Vzhľadom k tomu, že dostupné súbory boli získané sekvenovaním za použitia R7 chémie, výsledné FAST5 súbory neobsahujú priamo nameraný signál, ale informácie o každom „skoku" v signále vo forme tzv. udalostí (events), čo predstavuje určitú nevýhodu, pretože dochádza k strate signálu nášho záujmu. Na základe informácií v týchto udalostiach, ktorými sú stredná hodnota, dĺžka trvania a smerodajná odchýlka, je však možné požadovaný pôvodný signál zrekonštruovať.

Rekonštrukcia signálu vychádza z dopočítania počtu vzoriek pripadajúcich na každú priemernú hodnotu prúdu nameranú nanopórom, vyskytujúcu sa v zložke udalostí, kde každá z týchto hodnôt sa vyznačuje inou dĺžkou trvania. Ukážku takto zrekonštruovaného signálu je možné vidieť na Obr.16 .

a)



b)

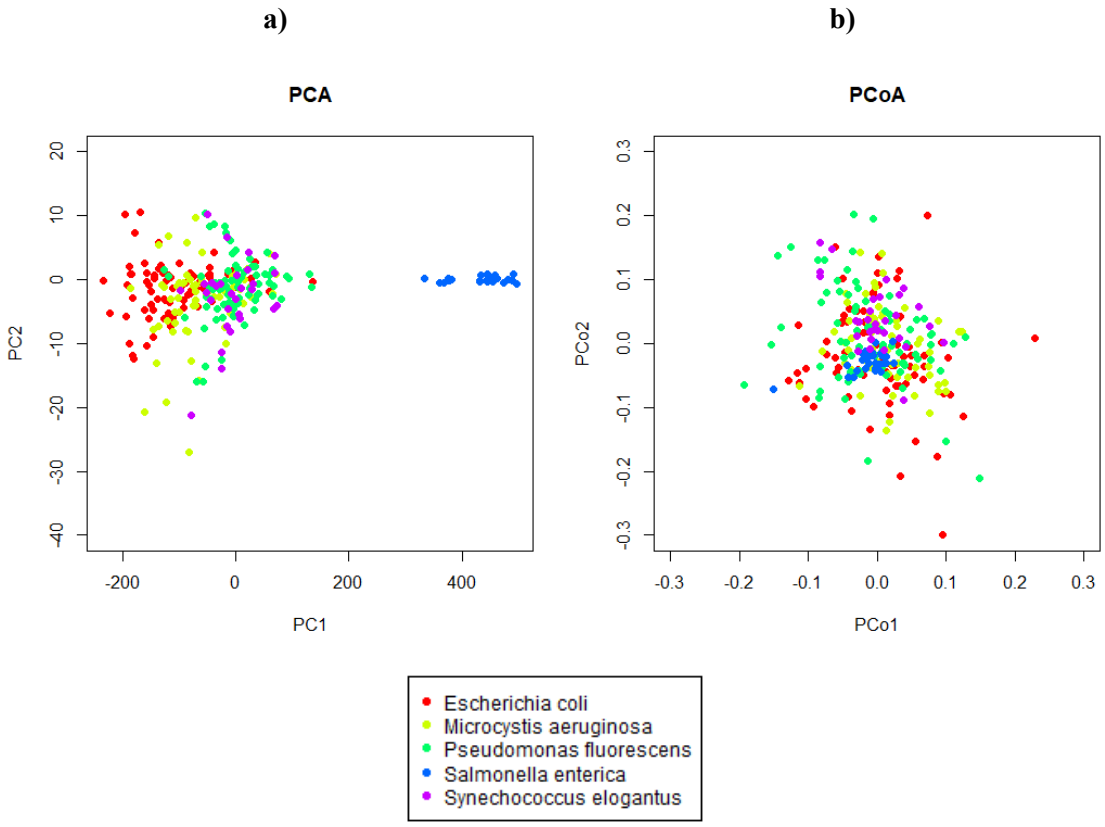


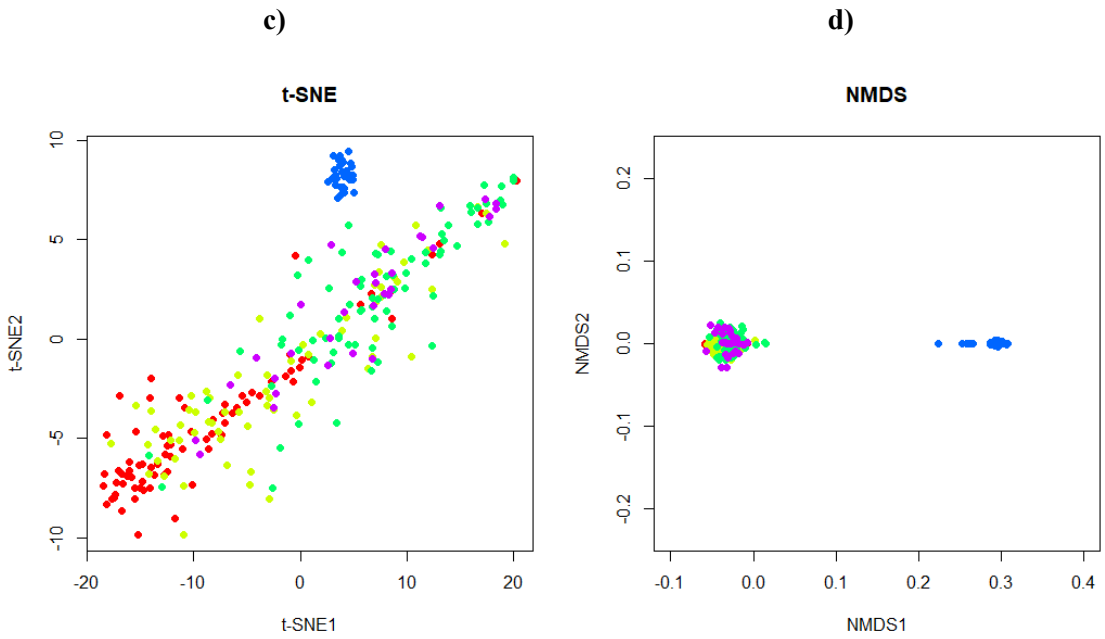
Obr.16 Ukážka zrekonštruovaných signálov pochádzajúcich zo sekvencií a) *Escherichia coli* a *b) Synechococcus elongatus* získaných nanopórovou technológiou

Aj napriek tomu, že signály sú značne odlišné, bolo by veľmi náročné a prácne klasifikovať ich prostým pohľadom. Z toho dôvodu sa obraciame na výpočtové metódy, ktorými sme schopní zamerať sa na najdôležitejšie parametre signálu a ten následne reprezentovať pomocou jediného bodu v nízko-dimenzionálnom priestore.

5.3 Zvolené metódy pre vizualizáciu dát

Obrovskou výhodou vzniknutej signálovej reprezentácie sekvencií, oproti znakovej podobe, je možnosť použiť na tieto dáta už existujúce metódy slúžiace k spracovávaniu signálov. Vektory zrekonštruovaných signálov z jednotlivých čítaní, ktorých rekonštrukcia je popísaná v kapitole 5.2, sú načítané do matice, odpovedajúcej vstupu zvolených metód vizualizácií založených na princípe redukcie dimenzií. Dĺžka produkovaných signálov z nanopóru však nie je rovnaká, to znamená, že pred samotným načítaním signálov do matice je potrebné zvoliť prah ich dĺžky, aby bolo možné na takúto maticu použiť metódy redukcie dimenzií, ktoré predpokladajú vstupy s rovnakou dĺžkou. Medzi zvolené metódy patrí a) PCA, b) PCoA, c) t-SNE a d) NMDS, kde výsledky ich realizácií je možné vidieť na Obr.17 . Po aplikovaní zvolených metód môžeme pozorovať vizualizácie zvoleného metagenomického datasetu, kde vo všetkých grafoch reprezentujú jednotlivé body konkrétnu sekvenciu a naviac farebne odpovedajú organizmu, z ktorého pochádzajú. Na základe toho môžeme jednoducho a jasne hodnotiť výsledné zhluky, ktoré sú v niektorých prípadoch viac, v iných menej odlíšiteľné. Podobné výsledky, pri zahrnutí viacerých organizmov je možné vidieť v prílohách na Príloh. obr.1





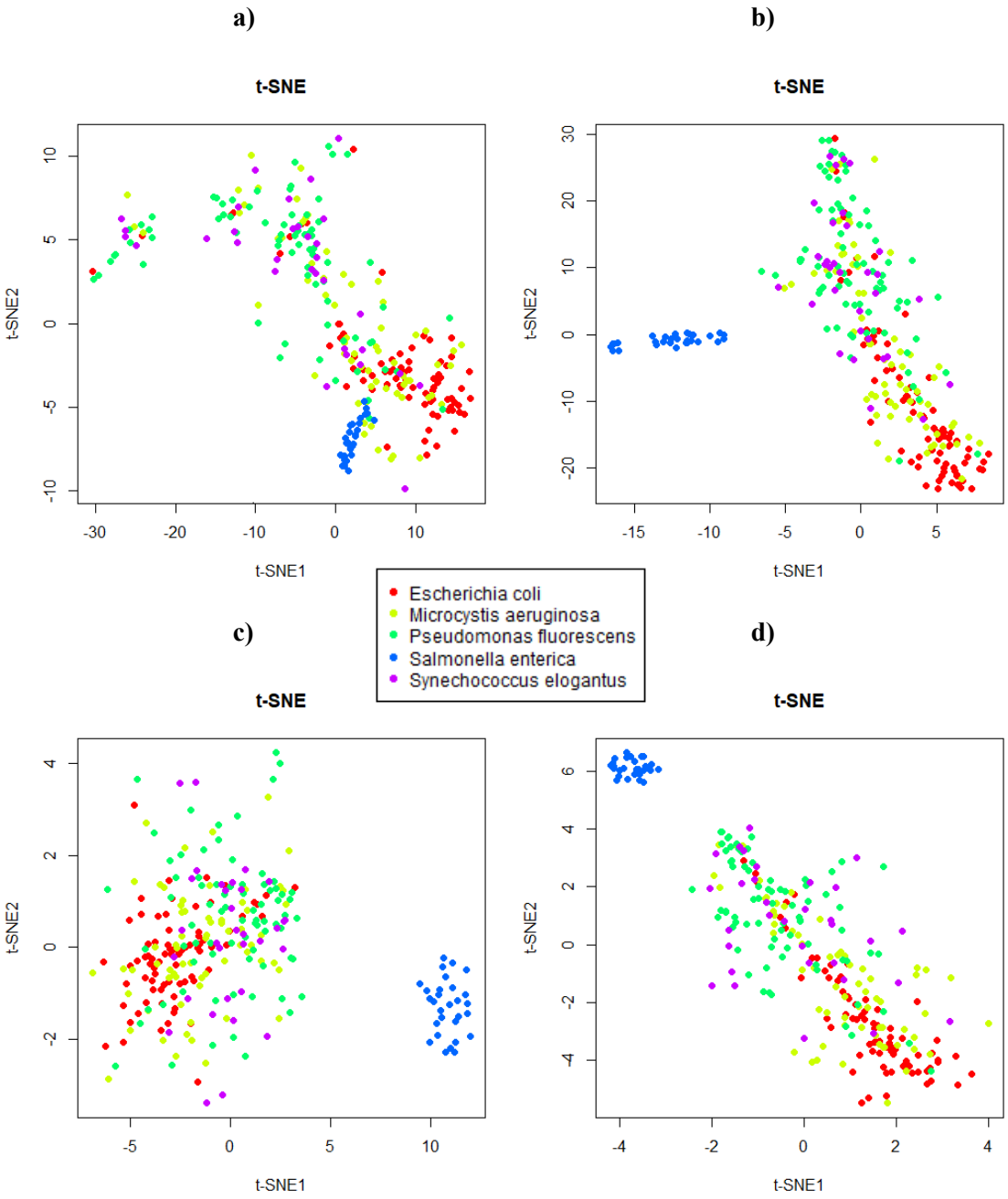
Obr.17 Aplikácia metód a) PCA, b) PCoA, c) t-SNE, a d) NMDS, založených na princípe redukcie dimenzií

Najhorší výsledok poskytuje v tomto prípade metóda PCoA, v ktorej nie sme schopní sa venovať žiadnej skupine jasne separovaných sekvencií určitého organizmu a celkovo vzniká iba jeden neprehľadný zhluk nahromadených bodov. V prípade metódy PCA a NMDS je síce viditeľná separácia sekvencií pochádzajúcich z organizmu *S. enterica*, avšak v prípade zvyšných sekvencií sú jednotlivé body pomerne, v prípade NMDS nadmerne, nahromadené v jednom mieste, a aj napriek tomu, že v prípade metódy PCA sme schopní zachytiť odlíšenie vzniknutých zhlukov, nevykazujú jednoznačnú separáciu. Metóda t-SNE sa svojím výsledkom určite rapídne nevzďaľuje od predošlých dvoch, avšak nahromadenie bodov, mimo tých, ktoré taktiež pochádzajú z organizmu *S. enterica*, nie je natoľko husté a môžeme teda predpokladať, že by táto metóda mohla v analýze ďalších skúmaných parametrov poskytovať relevantné výsledky. V snahe docieliť čo najlepšie výsledky boli skúmané viaceré parametre, značne ovplyvňujúce výsledný binning dát. Medzi skúmané parametre patrí:

- nastavenia v metóde t-SNE,
- počet zahrnutých organizmov,
- minimálna dĺžka trvania sekvencií,
- počet sekvencií pochádzajúcich z každého organizmu,
- taxonomická príbuznosť organizmov.

Nastavenia v metóde t-SNE

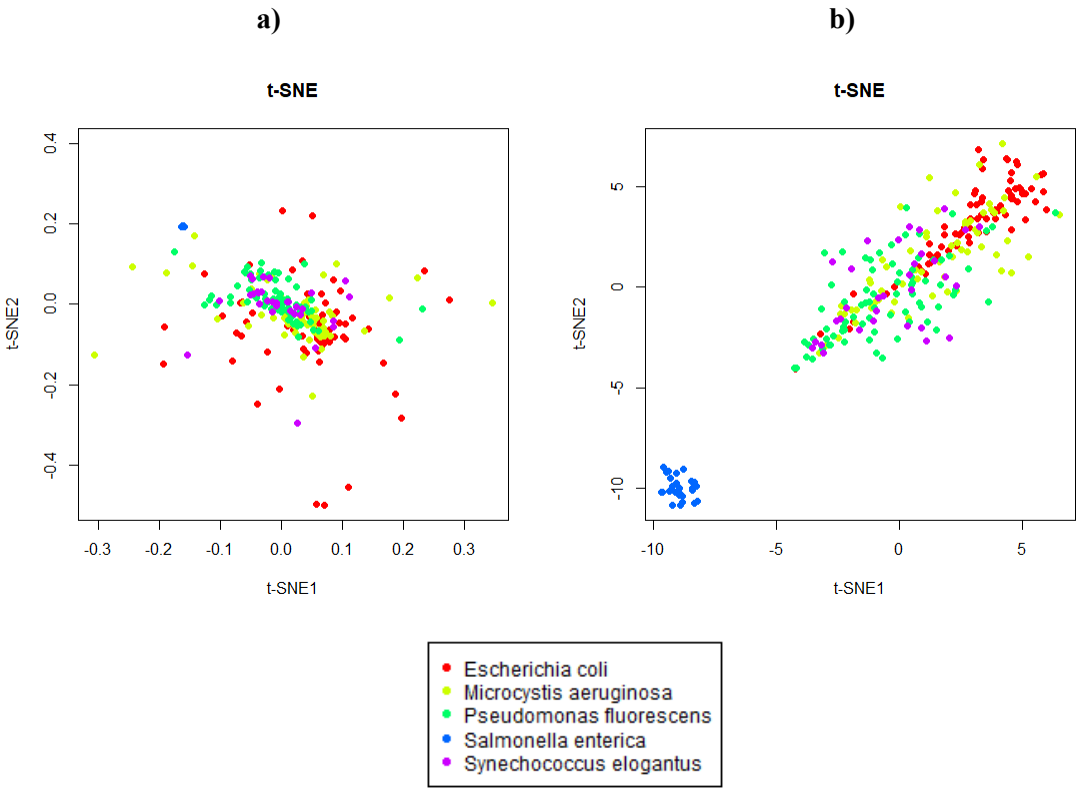
V zvolenej metóde t-SNE hrá, okrem iného, veľkú rolu aj nastavenie parametra perplexity, ktorým odhadujeme počet blízkych susedov každého bodu. Vieme tak určitým spôsobom ovplyvniť mieru dopadu lokálnej a globálnej variability našich dát na výsledné zobrazenia, ktoré môžeme vidieť na Obr.18 .

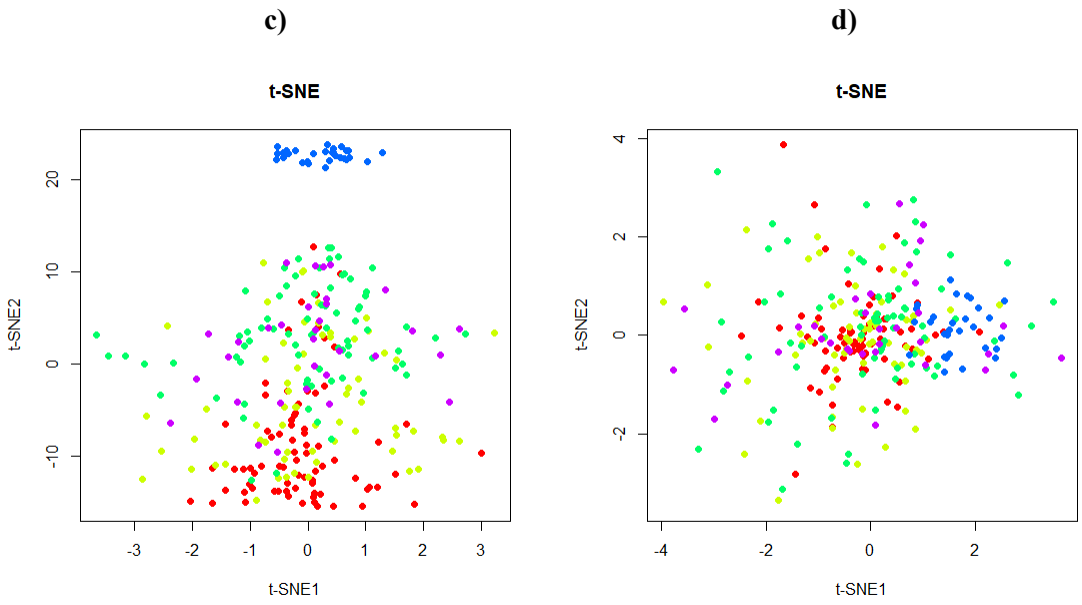


Obr.18 Použitie metódy t-SNE s hodnotami parametra perplexity a) 10, b) 20, c) 30, d) 50

Dosiahnutie čo najideálnejšej hodnoty tohto parametra si vyžaduje analýzu mnohých zobrazení s rôznymi hodnotami tohto parametra. Aby algoritmus pracoval správne, je v prvom rade nutné nastaviť parameter na hodnotu nižšiu ako je celkový počet bodov v zobrazení. Pri voľbe hodnoty je taktiež nutné hľadieť na hustotu dát, ktoré sú k dispozícii. Zjednodušene povedané, čím viac dát je k dispozícii, teda čím sú dáta hustejšie, tým väčšiu hodnotu parametra zadávame. Klasické rozmedzie hodnôt býva od 5 do 50. V zobrazení a) môžeme vidieť príklad, kedy zadaná hodnota nie je postačujúca, alebo vhodná pre dané dáta. S postupným nárastom hodnoty parametra dochádza k lepšiemu rozlíšeniu jednotlivých zhlukov. Metóda t-SNE vykazuje teda značnú flexibilitu, čo sa týka zobrazenia tých istých dát za rôznych podmienok, čo predstavuje značnú výhodu, avšak na druhej strane, jej interpretácia môže byť z tohto dôvodu o niečo zložitejšia.

Ďalší parameter, ktorý značne ovplyvňuje výsledné zobrazenia je maximálny počet iterácii, ktorým sa postupne približujeme k požadovanému riešeniu a podobne ako v prípade parametra perplexity, aj v tomto prípade nie je hodnota parametra presne definovaná. Opäť je potrebné preskúmať zobrazenia pri rôznych hodnotách tohto parametra, až do dosiahnutia určitej stability zobrazenia, kedy sme schopní odlíšiť nejaké zhluky. Výsledky použitia rôznych hodnôt tohto parametra môžeme vidieť na Obr.19 .



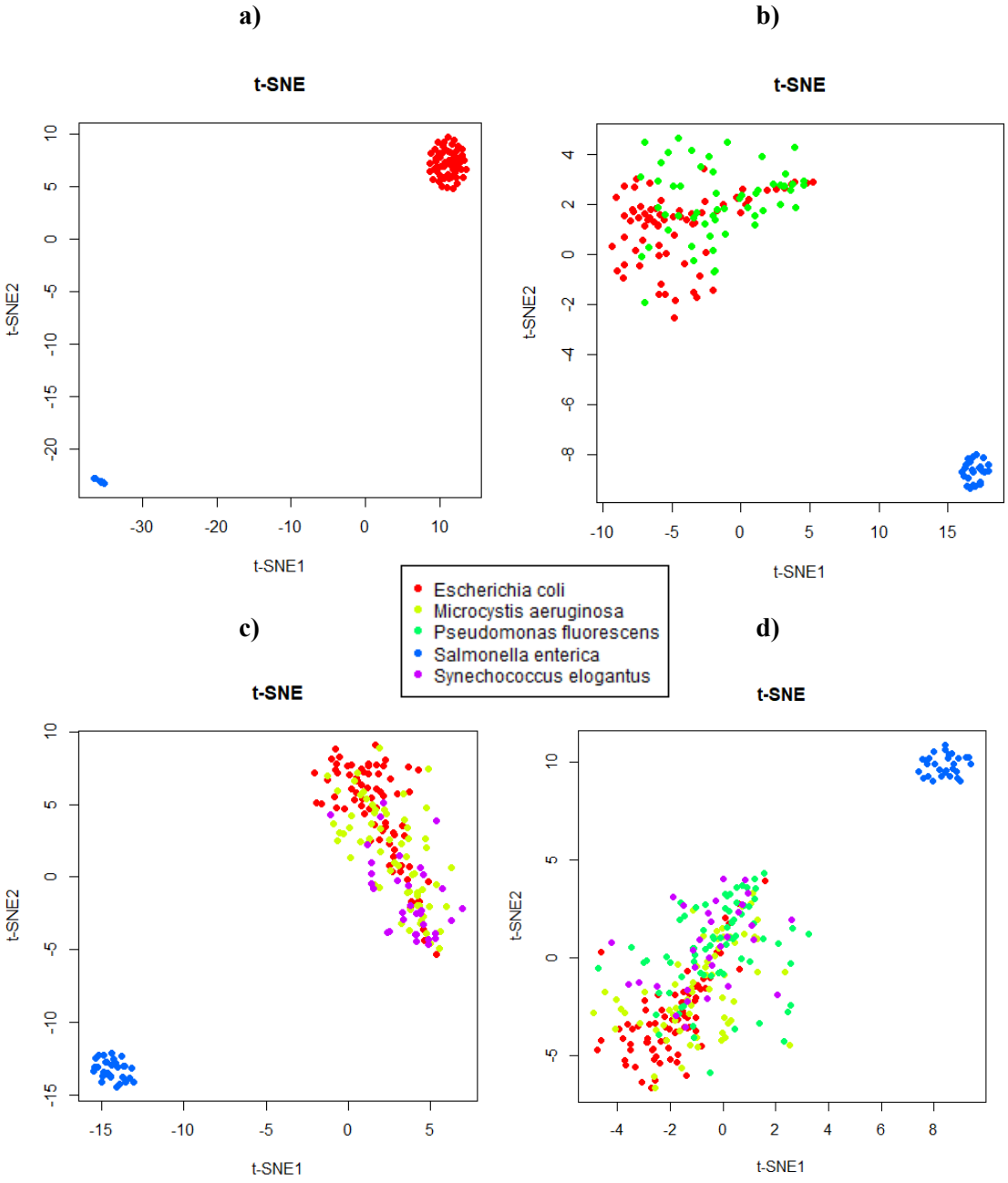


Obr.19 Použitie metódy t-SNE s počtom iterácií a) 250, b) 500, c) 1000 a d) 2000

S parametrom perplexity, ktorý sa na základe zobrazení na Obr.18 javí ako vhodne nastavený na hodnotu v rozmedzí od 30 do 50, je možné skúmať už iba nastavenia počtu iterácií. V prípade použitia a) 250 iterácií, je vidieť, že celý proces bol zastavený veľmi skoro a nedošlo k vytvoreniu takmer žiadnych zhlukov. V prípade použitia b) 500 iterácií, je zobrazenie najideálnejšie, keďže nám vzniká ukážkovo oddelený zhluk, ktorý by bolo možné identifikovať akoukoľvek doplňujúcou zhlukovou analýzou. Postupným zvyšovaním počtu iterácii na hodnoty c) 1000 a d) 2000 je vidieť, že dochádza k nežiadúcemu rozptyľovaniu vzniknutých zhlukov, pretože najskôr dochádza k nadmernému prepočítavaniu jednotlivých vzdialeností. Pri skúmaní ďalších parametrov sa teda budeme držať v okolí hodnoty, ktorá sa javí ako najvhodnejšia, to znamená 500 iterácii.

Počet zahrnutých organizmov

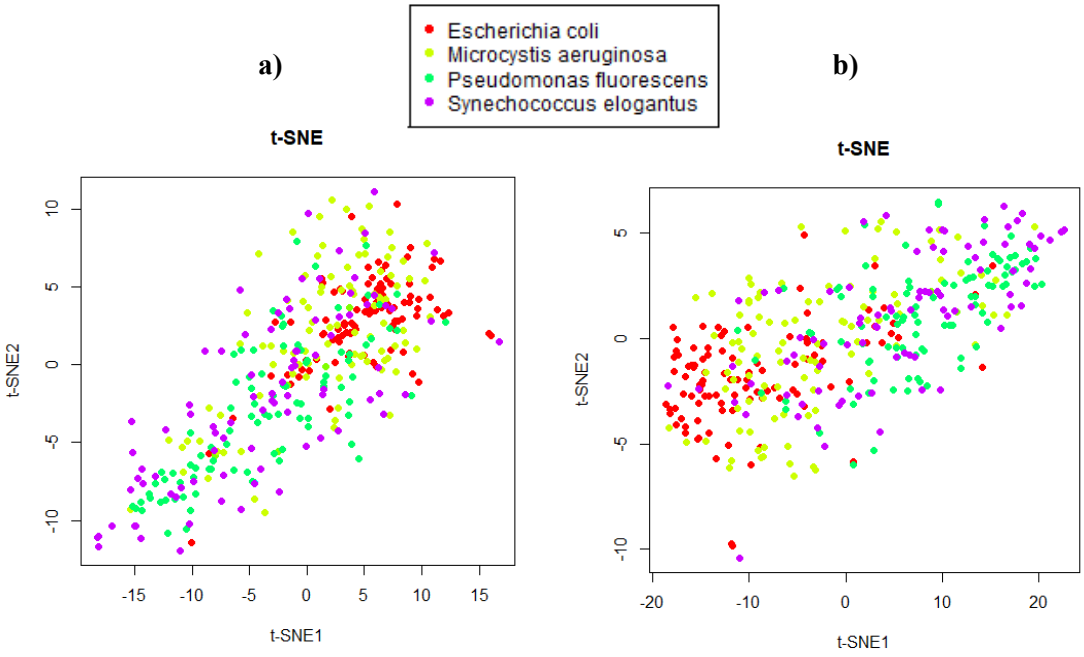
Je značné, že kvalitu výsledného zhlukovania ovplyvňuje počet zahrnutých organizmov a čím je organizmov menej, tým sú zhluky prehľadnejšie a nedochádza k ich nahromadeniu ako je možné vidieť na Obr.20 . V praxi by sme však určite prijali možnosť, kedy by bolo možné zahrnúť čo najviac organizmov, pretože v reálnych vzorkách by tomu taktiež nebolo inak. Na škodu však nemusí byť ani porovnávanie takéhoto malého počtu konkrétnych organizmov pri cielenej analýze ich príbuznosti.



Obr.20 Výsledok aplikovanej t-SNE metódy na a) dva, b) tri, c) štyri a d) päť rôznych organizmov

V každom z týchto zobrazení môžeme pozorovať odľahlý zhluk sekvencií, pochádzajúcich z organizmu *S. enterica*. V takýchto prípadoch by bolo ideálne takto vzniknutý odľahlý zhluk odfiltrovať a po odfiltrovaní realizovať druhotnú vizualizáciu zvyšných sekvencií.

Takýmto iteračným postupom by sme mohli dosiahnuť lepšie rozloženie prvotných nahromadených zhlukov. Výsledné vizualizácie po odfiltrovaní odľahlého zhluku je možné vidieť na Obr.21 .



Obr.21 Výsledok aplikovanej t-SNE metódy po odfiltrovaní odľahlého zhluku pri nastavení parametrov a) perplexity = 30, 1000 iterácií; b) perplexity = 36, 2000 iterácií

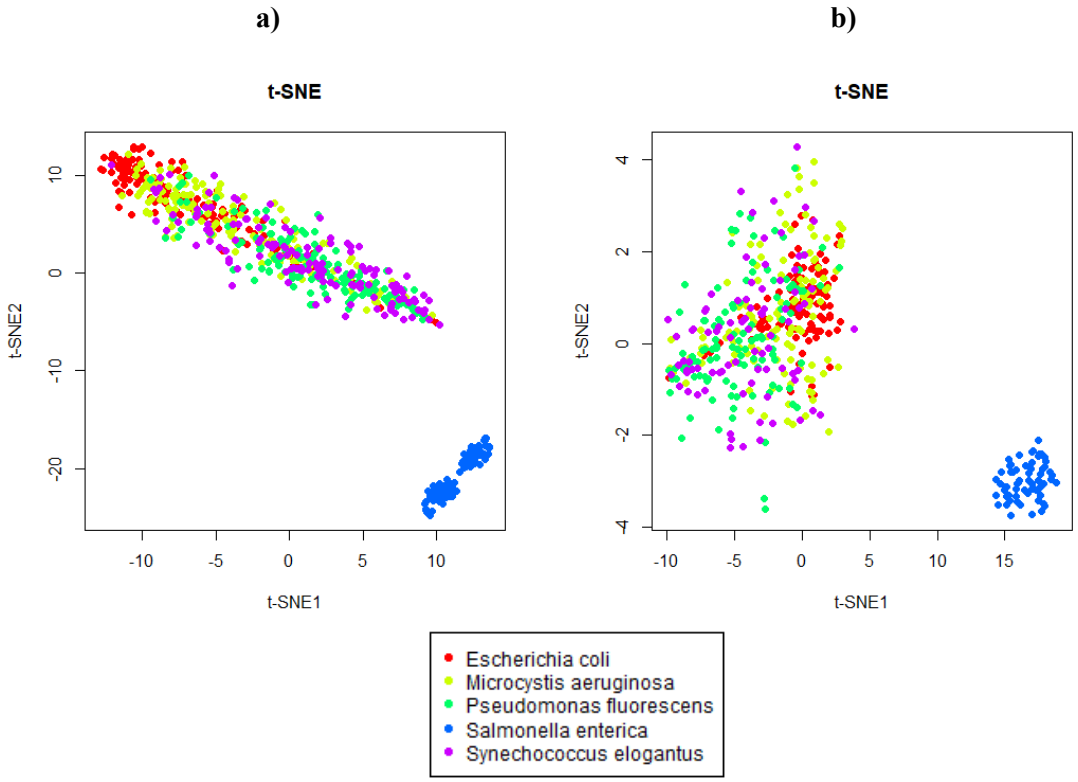
Po analýze vizualizácií pri rôznych nastaveniach parametra perplexity a počtu iterácií už nedošlo k vzniku ďalšieho výrazne odľahlého zhluku. Napriek tomu však získavame lepší prehľad o rozložení zvyšných sekvencií. Pri tejto aplikácií boli brané v úvahu len sekvencie, ktorých dĺžka trvania je väčšia ako priemerná dĺžka trvania všetkých zahrnutých sekvencií v danom datasete. Preto v prípade zahrnutia len dvoch organizmov môžeme vidieť mierny nedostatok bodov v prípade *S. enterica*, čo môže byť spôsobené nedostatočnou variabilitou dĺžok trvania sekvencií. Postupným pridávaním ďalších organizmov sa variabilita dĺžok trvania zväčšuje, tým pádom získavame lepší podiel sekvencií pochádzajúcich z jednotlivých organizmov, čo je prijateľné, pretože v reálnom metagenóme nebude možnosť vyberať si počet zahrnutých organizmov a zahrnuté budú všetky prítomné. Z toho dôvodu by preto bolo vhodnejšie zaoberať sa optimálnou hraničnou dĺžkou trvania sekvencií a jej znížením na určitú prijateľnú hodnotu, tak aby bol súčasne zahrnutý čo najväčší počet sekvencií z jednotlivých organizmov.

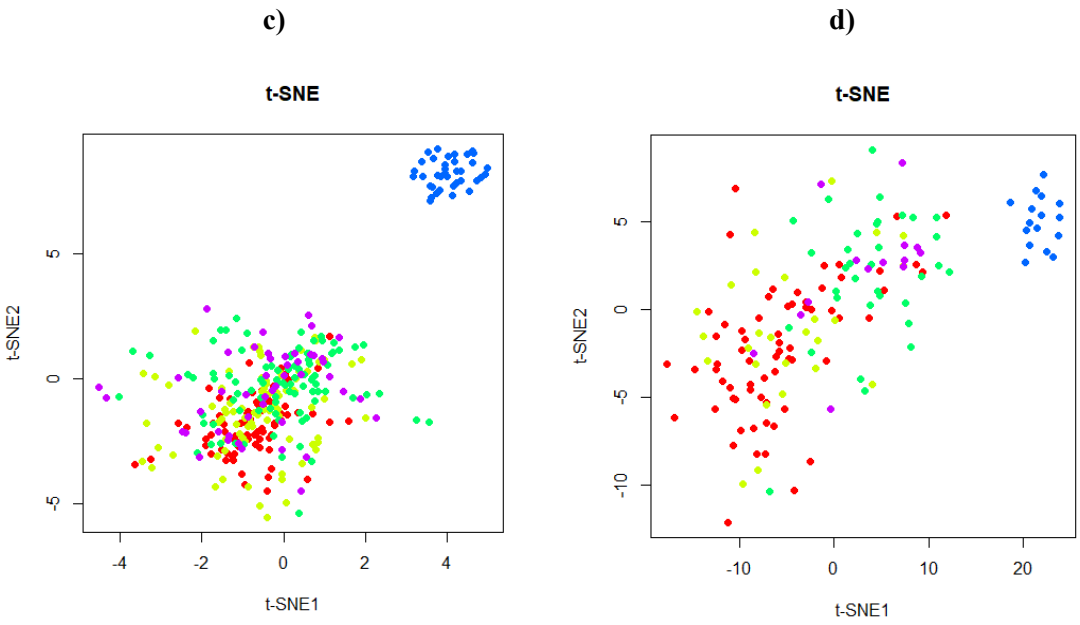
Minimálna dĺžka trvania sekvencií

Pre podrobnejší popis vplyvu minimálnej dĺžky trvania sekvencií, ktoré budú zahrnuté vo výslednom zobrazení boli skúmané štyri rôzne dĺžky a jednotlivé výsledky môžeme vidieť na Obr.22 . Je potrebné ešte spomenúť, že vypočítaná priemerná dĺžka trvania sekvencií v tomto prípade nadobúdala hodnotu 133 sekúnd. Priamo si teda môžeme overiť, či je používanie tejto dĺžky relevantné, alebo len zbytočne zvyšuje výpočtovú a časovú náročnosť daného algoritmu.

Zo zobrazení môžeme usúdiť, že použitie dopočítanej priemernej dĺžky trvania sekvencií, ako tomu bolo doposiaľ, nie je v tomto prípade úplne najlepšou alternatívou. Hoci zobrazenie d) vyzerá obstojne, je značné, že počet zahrnutých sekvencií je pomerne malý, pretože väčšina sekvencií v tomto prípade nepresahuje dĺžku trvania 200 sekúnd.

Zhluk sekvencií, pochádzajúcich zo *S. enterica* je odlíšiteľný pomerne vo všetkých prípadoch, avšak čím je zvolená dĺžka kratšia, tým krajšie a lepšie oddeliteľné zhluky sme schopní pozorovať, dokonca aj v prípade, kedy zhluk z tohto organizmu nadobúda nezvyčajný rozdvojený tvar.



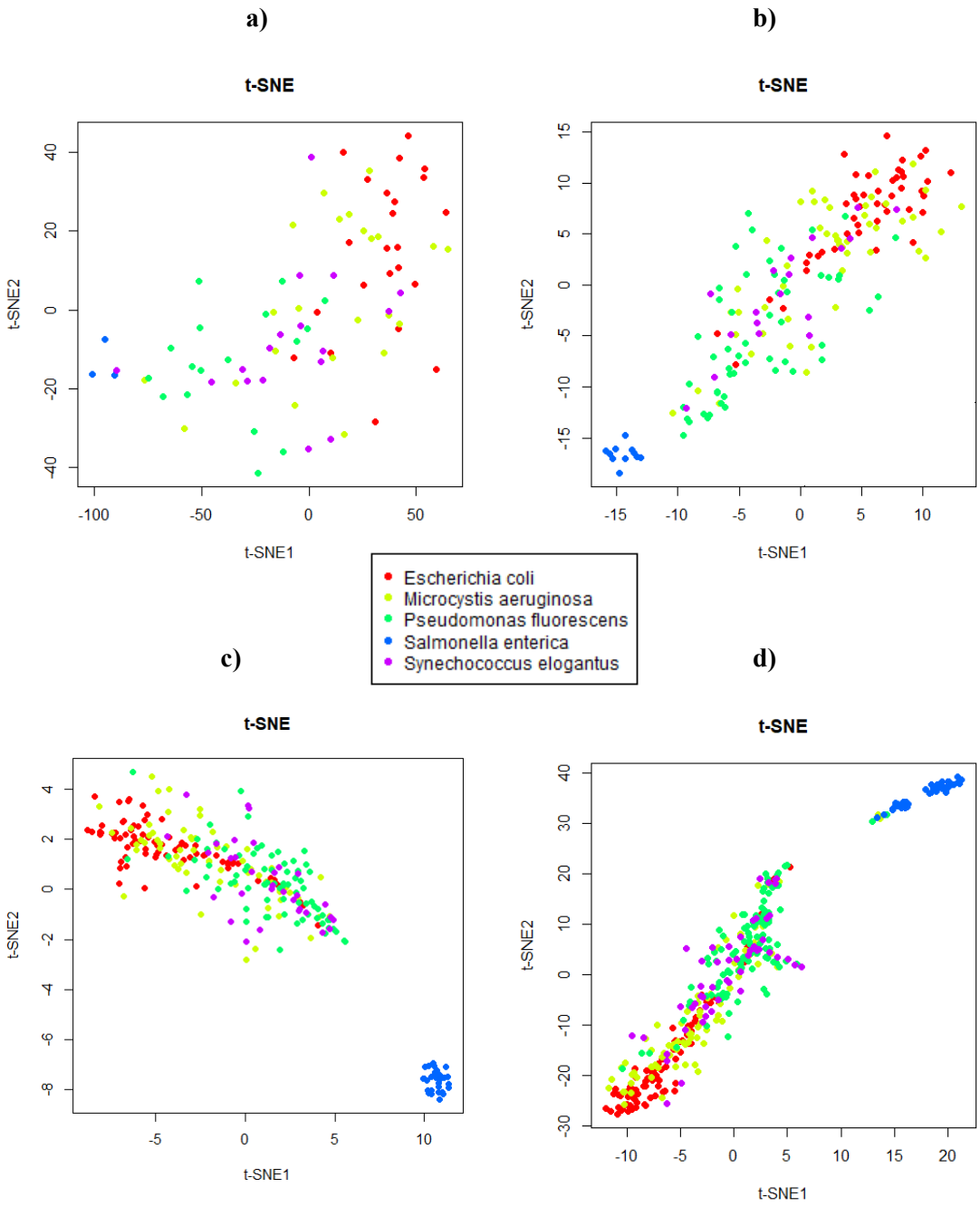


Obr.22 Výsledok aplikovanej t-SNE metódy pri volení dĺžky trvania zahrnutých sekvencií na a) 10, b) 50, c) 100 a d) 200 sekúnd

Je teda možné vyvodiť záver, že pri skúmaní ďalších parametrov teda nebude problém vynechať dopočítavanie priemernej dĺžky trvania zahrnutých sekvencií a príslušnú dĺžku si voliť podľa potrieb, aby bolo zahrnutých čo najviac sekvencií s vynechaním len tých najkratších.

Počet sekvencií pochádzajúcich z každého organizmu

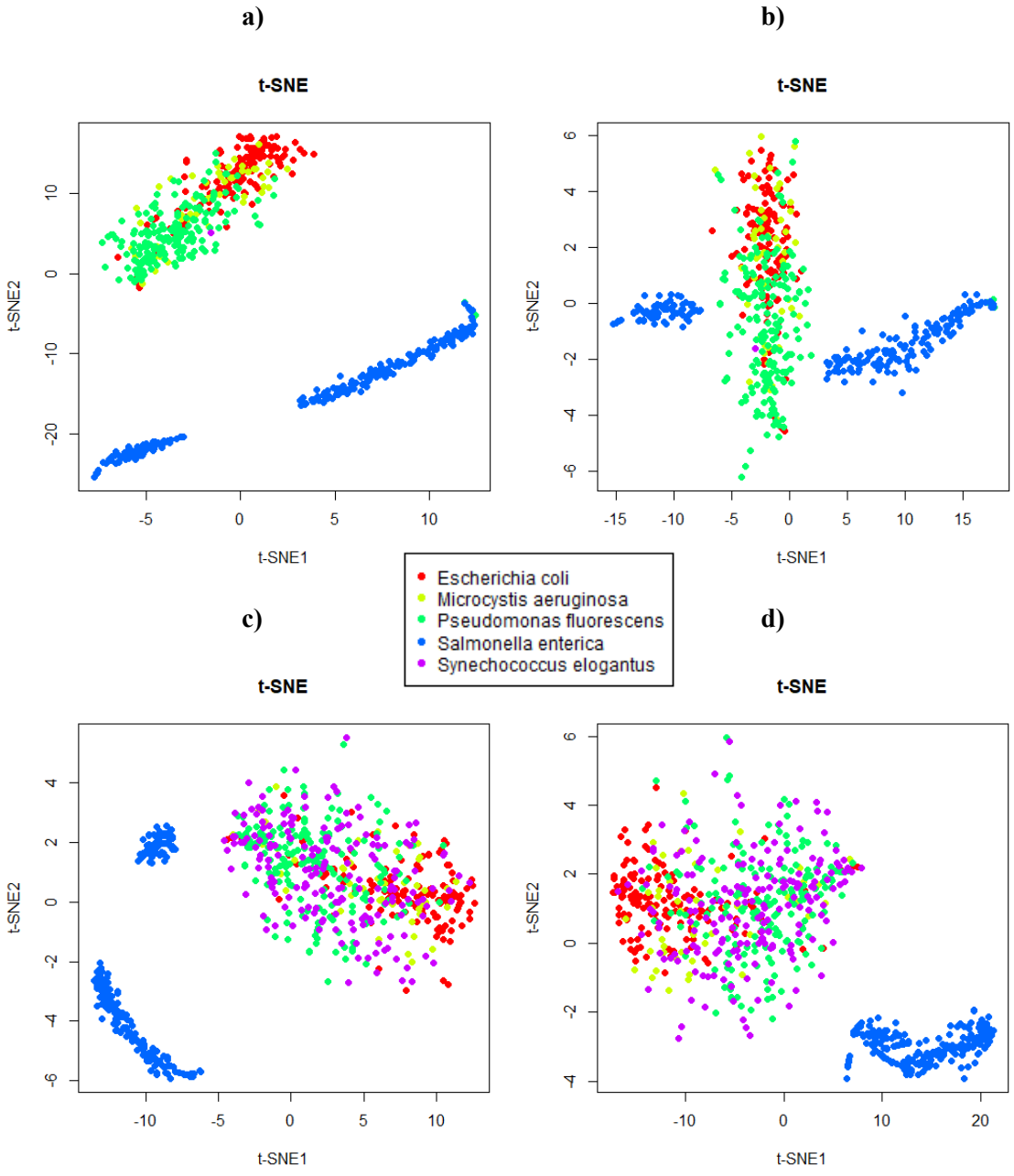
Čo sa týka počtu zahrnutých sekvencií, pochádzajúcich z jednotlivých organizmov, bolo by ideálne, ak aby táto hodnota nadobúdala čo najvyššiu možnú hodnotu. Testovaných bolo niekoľko hodnôt zvoleného počtu sekvencií z každého organizmu vo vzostupnom poradí od 50 po 200 sekvencií. Zdá sa, že veľmi malý počet sekvencií vedie k malému rozlíšeniu vytvorených zhlukov, kdežto vyšší počet vedie k žiadúcemu nahromadeniu bodov, čím vznikajú dostatočne jasne odlíšiteľné zhluky ako je možné vidieť na Obr.23 d).



Obr.23 Výsledok aplikovanej t-SNE metódy pri počte a) 50, b) 100, c) 150 a d) 200 zahrnutých sekvencií z jednotlivých organizmov

Pri skúmaní predošlých parametrov bola intuitívne zvolená hodnota 150 sekvencií z každého organizmu, ktorá sa z týchto zobrazení zdá byť vcelku prijateľná. Podobné výsledky, ktoré je možné nájsť v prílohách na Príloh. obr.2 , boli získané aj za použitia iba troch organizmov. V reálnom metagenóme však bude počet sekvencií z každého organizmu rozdielny. Takýto prípad je taktiež nasimulovaný a zobrazený na Obr.24 .

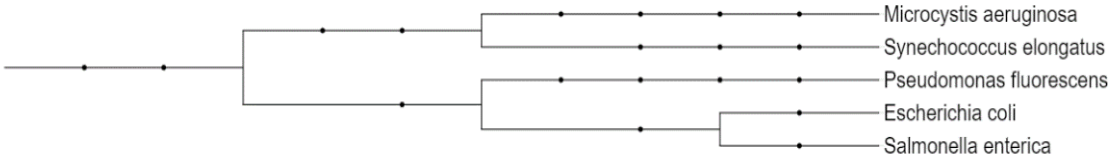
Viditeľné sú jasne odlíšiteľné zhluky a dá sa povedať, že výsledky sú dokonca lepšie, než v prípade neprirodzeného vyberania rovnakého počtu sekvencií, ako tomu bolo v predošlom prípade. Tento výsledok je jednoznačne prijateľný, pretože v reálnom metagenóme je výskyt nerovnakého počtu sekvencií jednoznačne pravdepodobnejší. Ešte pravdepodobnejší je výskyt viacerých organizmov a takýto prípad je možné vidieť v prílohách na Príloh. obr.4 .



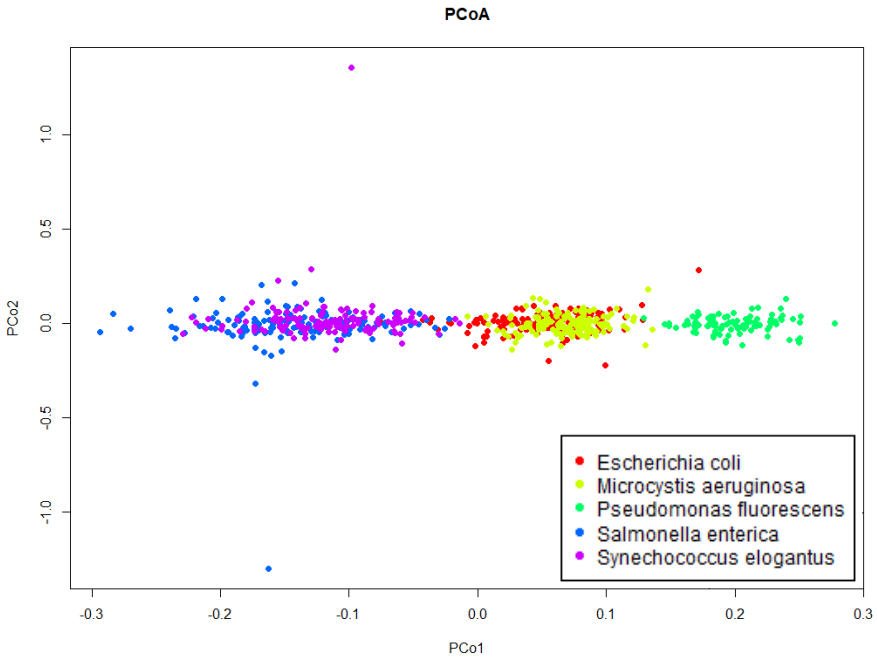
Obr.24 Výsledok aplikovanej t-SNE metódy pri náhodnom počte zahrnutých sekvencií z každého organizmu v poradí podľa legendy a) 150, 50, 200, 300 , 50 sekvencií pri perplexity = 54, 600 iterácií; b) 150, 50, 200, 300 , 50 sekvencií pri perplexity = 30, 500 iterácií; c) 150, 50, 200, 300 , 150 sekvencií pri perplexity = 54, 600 iterácií ; d) 150, 50, 200, 300 , 150 sekvencií pri perplexity = 44, 500 iterácií

Taxonomická príbuznosť organizmov

Pri skúmaní vplyvu taxonomickej príbuznosti organizmov na výsledné zobrazenie sekvencií z nich pochádzajúcich, bolo vhodné si na približnú predstavu vytvoriť fylogenetický strom, zobrazený na Obr.25 , odpovedajúci práve ich spomínanej príbuznosti. V tomto prípade sa však osvedčila ako najviac použiteľná metóda PCoA, ktorej výsledky vo veľkej miere ovplyvňuje dĺžka trvania jednotlivých sekvencií, čo potvrdzujú vizualizácie v prílohách na Príloh. obr.3 . Zo stromu vyplývajúce prvé dva príbuzné organizmy *M. aeruginosa* a *S. elongatus* patria medzi sinice a ich prípadná genómová podobnosť bola zistená pri vytváraní fylogenetického stromu metódou spájania susedov z preložených proteínových domén podľa [44], z čoho môže vyplývať aj ich asociácia v zobrazení na Obr.26 . *E. coli* a *S. enterica* patria medzi enterobaktérie, ktoré sú okrem vody, pôdy a iných oblastí súčasťou aj prirodzenej črevnej mikroflóry u zvierat a ľudí. Pri porovnávaní genómov [45] *E. coli* a *S. enterica* bolo zistené, že viac ako 1100 génov prítomných v genóme *S. enterica* chýbajú v genóme *E. coli,* v ktorom sa naopak nachádza viac ako 800 génov neprítomných v genóme *S. enterica*. Naopak pri porovnávaní genómov *E. coli* a *P. fluorescens,* v príspevku [46], bolo zistené, že sú takmer identické, čomu celkom názorne odpovedá aj táto vizualizácia. Môžeme sa teda domnievať, že práve z tohto dôvodu sa zhluk sekvencií, pochádzajúcich z baktérie *S. enterica*, separuje od zhluku, tvoreného sekvenciami pochádzajúcimi z organizmov *E.coli* a *P. fluorescens*, no napriek tomu si od tohto zhluku zachováva svoju príbuzenskú vzdialenosť vyplývajúcu z fylogenetického stromu.



Obr.25 Fylogenetický strom vyjadrujúci príbuzenské vzťahy medzi organizmami



Obr.26 Aplikovanie PCoA metódy pri skúmaní súvislosti taxonómie s vytváraním zhlukov

ZÁVER

Cieľom tejto bakalárskej práce bolo vytvoriť algoritmus pre binning metagenomických dát, získaných sekvenačnou technológiou Oxford Nanopore. Jedná sa o nový prístup spracovania sekvenčných dát využívajúci aplikáciu metód redukcie dimenzií priamo na prúdové signály produkované nanopórovým sekvenovaním. Výsledný produkt by mal umožniť rýchle spracovanie poskytnutých surových dát a ich efektívnu klasifikáciu vrátane vizualizácie obrovských metagenomických datasetov. V snahe docieliť čo najlepšie výsledky bolo potrebné preskúmať rôzne metódy a ich nastavenia, poprípade predspracovania dát s postupným dopracovaním sa k ideálnej vizualizácii. Tá by mala poskytovať jasné rozlíšenie zhodných sekvencií blízko príbuzných taxónov, mala by byť aplikovateľná na dlhé fragmenty a taktiež by mala poskytovať účinnú analýzu rastúceho množstva dát genómových sekvencií.

Prvým praktickým krokom k dosiahnutiu tohto cieľa bolo vytvorenie simulovaného metagenómu, získanie potrebných dát a na základe informácií v týchto dátach uskutočniť rekonštrukciu k získaniu vektorov signálových sekvencií. Obrovskou výhodou vzniknutej signálovej reprezentácie sekvencií, oproti znakovej podobe, je možnosť použiť na tieto dáta už existujúce metódy slúžiace k spracovaniu signálov.

Využitím metód redukcie dimenzií je získavaný prístup vizualizovať obrovské súbory dát, odpovedajúce zmesi týchto signálových sekvencií z rôznych organizmov v nízko-dimenzionálnom priestore. Miernou nevýhodou týchto metód je nutnosť položiť na vstup maticu signálov, ktorých dĺžka musí byť rovnaká, to znamená, že tieto signály je potrebné orezať, keďže dĺžka signálov produkovaných nanopórovým sekvenovaním sa zakaždým líši. Predstavených bolo niekoľko metód, z ktorých pre analýzu ďalších skúmaných parametrov bola zvolená t-SNE metóda, predovšetkým vďaka vykazujúcej flexibilite pri zobrazeniach tých istých dát za rôznych podmienok.

Na základe doposiaľ získaných výsledkov je možné vyvodiť záver, že použité algoritmy a predspracovanie dát poskytujú obstojné vizualizácie s obrovskou budúcnosťou v oblasti klasifikácie sekvencií získaných predovšetkým technológiou nanopórového sekvenovania. Pomerne prínosné je zistenie, že algoritmus poskytuje veľmi dobré výsledky pri zvolení pomerne krátkych dĺžok trvania sekvencií, vďaka čomu sú zahrnuté takmer všetky sekvencie s vynechaním len tých najkratších. Po doladení by táto metóda mohla mať potenciál pre spracovanie obdobných dát v reálnom čase.

Literatúra

[1] PETERSON, Jane, et al. The NIH human microbiome project. *Genome research*, 2009, 19.12: 2317-2323. [cit. 2017-02-10].

[2] CHAPMAN, T. A., X.-Y. WU, I. BARCHIA, K. A. BETTELHEIM, S. DRIESEN, D. TROTT, M. WILSON a J. J.-C. CHIN. Comparison of Virulence Gene Profiles of Escherichia coli Strains Isolated from Healthy and Diarrheic Swine. *Applied and Environmental Microbiology* [online]. 2006, **72**(7), 4782-4795 [cit. 2017-12-28]. DOI: 10.1128/AEM.02885-05. ISSN 0099-2240. [cit. 2017-02-10].

[3] The New Science of Metagenomics [online]. Washington, D.C: National Academies Press, 2007 [cit. 2017-11-11]. ISBN 978-0-309-10676-4.

[4] WEINSTOCK, George M. Genomic approaches to studying the human microbiota. Nature [online]. 2012, 489(7415), 250-256 [cit. 2017-11-12]. DOI: 10.1038/nature11553. ISSN 0028-0836.

[5] UHLÍK, Ondřej, Strejček MICHAL, Hroudová MILUŠE, Demnerová KATEŘINA a Macek TOMÁŠ. IDENTIFIKACE A CHARAKTERIZACE BAKTERIÍ S BIOREMEDIAČNÍM POTENCIÁLEM – OD KULTIVACE K METAGENOMICE. Chemické listy [online]. 2013, 107(8), 614–622 [cit. 2017-11-12].

[6] LIU, Bo, et al. Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. *Genome biology*, 2011, 12.1: P11.MOROZOVA, Olena a Marco A. MARRA. Applications of next-generation sequencing technologies in functional genomics. *Genomics* [online]. 2008, **92**(5), 255-264 [cit. 2017- 12- 21].

[7] HEATHER, James M. a Benjamin CHAIN. The sequence of sequencers: The history of sequencing DNA. *Genomics* [online]. 2016, **107**(1), 1-8 [cit. 2017-11-12]. DOI: 10.1016/j.ygeno.2015.11.003. ISSN 08887543.

[8] REIS-FILHO, Jorge S. Next-generation sequencing. *Breast Cancer Research* [online]. 2009, **11**(S3), - [cit. 2017-12-21]. DOI: 10.1186/bcr2431. ISSN 1465-542x.

[9] MINAKSHI P., RANJAN K., BRAR B., AMBAWAT S., SHAFIQ M. , ALISHA A., KUMAR P., GANESHARAO JV., JAKHAR S., BALODI S., SINGH A., PRASAD G. (2014). New approaches for diagnosis of viral diseases in animals. Adv. Anim. Vet. Sci. 2 (4S): 55 – 63.

[10]GULLAPALLI, Rama R., et al. Next generation sequencing in clinical medicine: Challenges and lessons for pathology and biomedical informatics. *Journal of pathology informatics*, 2012, 3. [cit. 2017-12-21].

[11] SHENDURE, Jay, Shankar BALASUBRAMANIAN, George M. CHURCH, Walter GILBERT, Jane ROGERS, Jeffery A. SCHLOSS a Robert H. WATERSTON. DNA sequencing at 40: past, present and future. *Nature*[online]. 2017, **550**(7676), 345-353 [cit. 2017-11-12]. DOI: 10.1038/nature24286. ISSN 0028-0836.

[12] MARDIS, Elaine R. The impact of next-generation sequencing technology on genetics. *Trends in Genetics* [online]. 2008, **24**(3), 133-141 [cit. 2017-12-21]. DOI: 10.1016/j.tig.2007.12.007. ISSN 01689525.

[13] KOUBKOVÁ, L., B. VOJTĚŠEK a R VYZULA. Sekvenování nové generace a možnosti jeho využití v onkologické praxi: Next Generation Sequencing – Application in Clinical Practice. Klinická onkologie[online]. 2014, 27(3), 61-68 [cit. 2017-12-28]. ISSN 1802-5307.

[14] SCHADT, E. E., S. TURNER a A. KASARSKIS. A window into third generation sequencing. *Human Molecular Genetics* [online]. 2011, **20**(4), 853-853 [cit. 2017-11-18]. DOI: 10.1093/hmg/ddq481. ISSN 0964-6906.

[15] RHOADS, Anthony a Kin Fai AU. PacBio Sequencing and Its Applications. *Genomics, Proteomics & Bioinformatics* [online]. 2015, **13**(5), 278-289 [cit. 2017-12-22]. DOI: 10.1016/j.gpb.2015.08.002. ISSN 16720229.

[16] BAYLEY, H. Nanopore Sequencing: From Imagination to Reality. *Clinical Chemistry* [online]. 2014, **61**(1), 25-31 [cit. 2017-12-22]. DOI: 10.1373/clinchem.2014.223016. ISSN 0009-9147.

[17] DERRINGTON, I. M., T. Z. BUTLER, M. D. COLLINS, E. MANRAO, M. PAVLENOK, M. NIEDERWEIS a J. H. GUNDLACH. Nanopore DNA sequencing with MspA. *Proceedings of the National Academy of Sciences* [online]. 2010, **107**(37), 16060- 16065 [cit. 2017-11-19]. DOI: 10.1073/pnas.1001831107. ISSN 0027-8424.

[18] ESCALANTE, Ana E., Lev JARDÓN BARBOLLA, Santiago RAMÍREZ-BARAHONA a Luis E. EGUIARTE. The study of biodiversity in the era of massive sequencing. *Revista Mexicana de Biodiversidad* [online]. 2014, **85**(4), 1249-1264 [cit. 2017-12-22]. DOI: 10.7550/rmb.43498. ISSN 18703453.

[19] LOMAN, N. J. a A. R. QUINLAN. Poretools: a toolkit for analyzing nanopore sequence data. Bioinformatics [online]. 2014, 30(23), 3399-3401 [cit. 2017-11-10]. DOI: 10.1093/bioinformatics/btu555. ISSN 1367-4803.

[20] IP, Camilla L.C., Matthew LOOSE, John R. TYSON, et al. MinION Analysis and Reference Consortium: Phase 1 data release and analysis. *F1000Research* [online]. , - [cit. 2017-11-10]. DOI: 10.12688/f1000research.7201.1. ISSN 2046-1402.

[21] SMITH, Michael John Sebastian. *Application-specific integrated circuits*. Reading, MA: Addison-Wesley, 1997.

[22] LU, Hengyun, Francesca GIORDANO a Zemin NING. Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics, Proteomics & Bioinformatics* [online]. 2016, 14(5), 265-279 [cit. 2017-11-10]. DOI: 10.1016/j.gpb.2016.05.004. ISSN 16720229.

[23] Update: New 'R9' nanopore for faster, more accurate sequencing, and new ten minute preparation kit. In: *Oxford Nanopore Technologies* [online]. [cit. 2017-12-23].

[24] JAIN, Miten, Ian T FIDDES, Karen H MIGA, Hugh E OLSEN, Benedict PATEN a Mark AKESON. Improved data analysis for the MinION nanopore sequencer. *Nature Methods* [online]. 2015, **12**(4), 351-356 [cit. 2017-11-19]. DOI: 10.1038/nmeth.3290. ISSN 1548-7091.

[25] JAIN, Miten, Hugh E. OLSEN, Benedict PATEN a Mark AKESON. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology* [online]. 2016, **17**(1), - [cit. 2017-11-19]. DOI: 10.1186/s13059-016-1103-0. ISSN 1474-760x.

[26] About PromethION: System overview & technical specifications. In: Oxford Nanopore Technologies[online]. [cit. 2017-12-23].

[27] Advanced Sequencing Kits: Ligation Sequencing Kit 2D. In: *Oxford Nanopore Technologies: Store*[online]. [cit. 2017-12-23].

[28] HOENEN, Thomas. Sequencing of Ebola Virus Genomes Using Nanopore Technology. *BIO-PROTOCOL* [online]. 2016, **6**(21), - [cit. 2017-12-27]. DOI: 10.21769/BioProtoc.1998. ISSN 2331-8325.

[29] FARIA, N. R., J. QUICK, I.M. CLARO, et al. Establishment and cryptic transmission of Zika virus in Brazil and the Americas. *Nature* [online]. 2017, **546**(7658), 406-410 [cit. 2017- 12-27]. DOI: 10.1038/nature22401. ISSN 0028-0836.

[30] BATOVSKA, Jana, Stacey E LYNCH, Brendan C RODONI, Tim I SAWBRIDGE a Noel OI COGAN. Metagenomic arbovirus detection using MinION nanopore sequencing. *Journal of Virological Methods*[online]. 2017, **249**, 79-84 [cit. 2017-12-27].

[31] LEVER, Jake, Martin KRZYWINSKI a Naomi ALTMAN. Points of Significance: Principal component analysis. *Nature Methods* [online]. 2017, **14**(7), 641-642 [cit. 2017-12-27]. DOI: 10.1038/nmeth.4346. ISSN 1548-7091.

[32] SMITH, Lindsay I. A tutorial on principal components analysis. 2002.

[33] RAMETTE, Alban. Multivariate analyses in microbial ecology. *FEMS Microbiology Ecology* [online]. 2007, **62**(2), 142-160 [cit. 2017-12-27]. DOI: 10.1111/j.1574- 6941.2007.00375.x. ISSN 01686496.

[34] SAEED, Isaam, Sen-Lin TANG a Saman K. HALGAMUGE. Unsupervised discovery of microbial population structure within metagenomes using nucleotide base

composition. *Nucleic Acids Research* [online]. 2012, **40**(5), e34-e34 [cit. 2018-01-03]. DOI: 10.1093/nar/gkr1204. ISSN 1362-4962.

[35] LEGENDRE, Pierre; ANDERSON, Marti J. Distance-based redundancy analysis: testing multispecies responses in multifactorial ecological experiments. *Ecological monographs*, 1999, 69.1: 1-24.

[36] SUZUKI, Taichi A., Michael W. NACHMAN a Erwin G ZOETENDAL. Spatial Heterogeneity of Gut Microbial Composition along the Gastrointestinal Tract in Natural Populations of House Mice. *PLOS ONE* [online]. 2016, **11**(9), e0163720- [cit. 2017-12-27].

[37] HOAND, Steven M. Non-metric multidimensional scaling (MDS). 2008.

[38] HARUŠTIAKOVÁ, Danka. *Vícerozměrné statistické metody v biologii*. Brno: Akademické nakladatelství CERM, 2012. ISBN 978-80-7204-791-8.

[39] MILLS, DeEtta K., James A. ENTRY, Joshua D. VOSS, Patrick M. GILLEVET a Kalai MATHEE. An assessment of the hypervariable domains of the 16S rRNA genes for their value in determining microbial community diversity: the paradox of traditional ecological indices. *FEMS Microbiology Ecology* [online]. 2006, **57**(3), 496-503 [cit. 2017-12-27].

[40] MAATEN, Laurens van der; HINTON, Geoffrey. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 2008, 9.Nov: 2579-2605. - [cit. 2017-12-27].

[41] LACZNY, Cedric C., Nicolás PINEL, Nikos VLASSIS a Paul WILMES. Alignment-free Visualization of Metagenomic Data by Nonlinear Dimension Reduction. *Scientific Reports* [online]. 2015, **4**(1), - [cit. 2017-12-27].

[42] VAN DER MAATEN, Laurens. Accelerating t-SNE using tree-based algorithms. *Journal of machine learning research*, 2014, 15.1: 3221-3245. - [cit. 2018-5-12].

[43] DICK, Gregory J, Anders F ANDERSSON, Brett J BAKER, Sheri L SIMMONS, Brian C THOMAS, A Pepper YELTON a Jillian F BANFIELD. Community-wide analysis of microbial genome sequence signatures. *Genome Biology* [online]. 2009, **10**(8), R85- [cit. 2017-12-27].

[44] EHRENREICH, Ian M.; WATERBURY, John B.; WEBB, Eric A. Distribution and diversity of natural product genes in marine and freshwater cyanobacterial cultures and genomes. *Applied and environmental microbiology*, 2005, 71.11: 7401-7413. [cit. 2018- 12- 27].

[45] WINFIELD, Mollie D.; GROISMAN, Eduardo A. Role of nonhost environments in the lifestyles of Salmonella and Escherichia coli. *Applied and environmental microbiology*, 2003, 69.7: 3687-3694.

[46] STOVER, C. K, et al. Complete genome sequence of Pseudomonas aeruginosa PAO1, an opportunistic pathogen. *Nature*, 2000, 406.6799: 959.

Zoznam skratiek

ONT	Oxford Nanopore Technologies
NGS	Next-generation Sequencing
bp	Base Pair
PCR	Polymerase Chain Reaction
TGS	Third-generation Sequencing
SMRT	Single-Molecule Real-Time
ZMW	Zero-Mode Waveguides
ASIC	Application Specific Integrated Circuit
HMM	Hidden Markov Model
PCA	Principal Component Analysis
PCoA	Principal Coordinates Analysis
OTU	Operational Taxonomic Unit
NMDS	Non-Metric Multidimensional Scaling
t-SNE	t-distributed Stochastic Neighbor Embedding
BH-SNE	Barnes-Hut Stochastic Neighbor Embedding
VP-stromy	Vantage Point-trees
ESOM	Emergent Self-Organizing Maps
OFDEG	Oligonucleotide Frequency Derived Error Gradient
TNF	Tetranukleotidová Fekvencia

Zoznam príloh

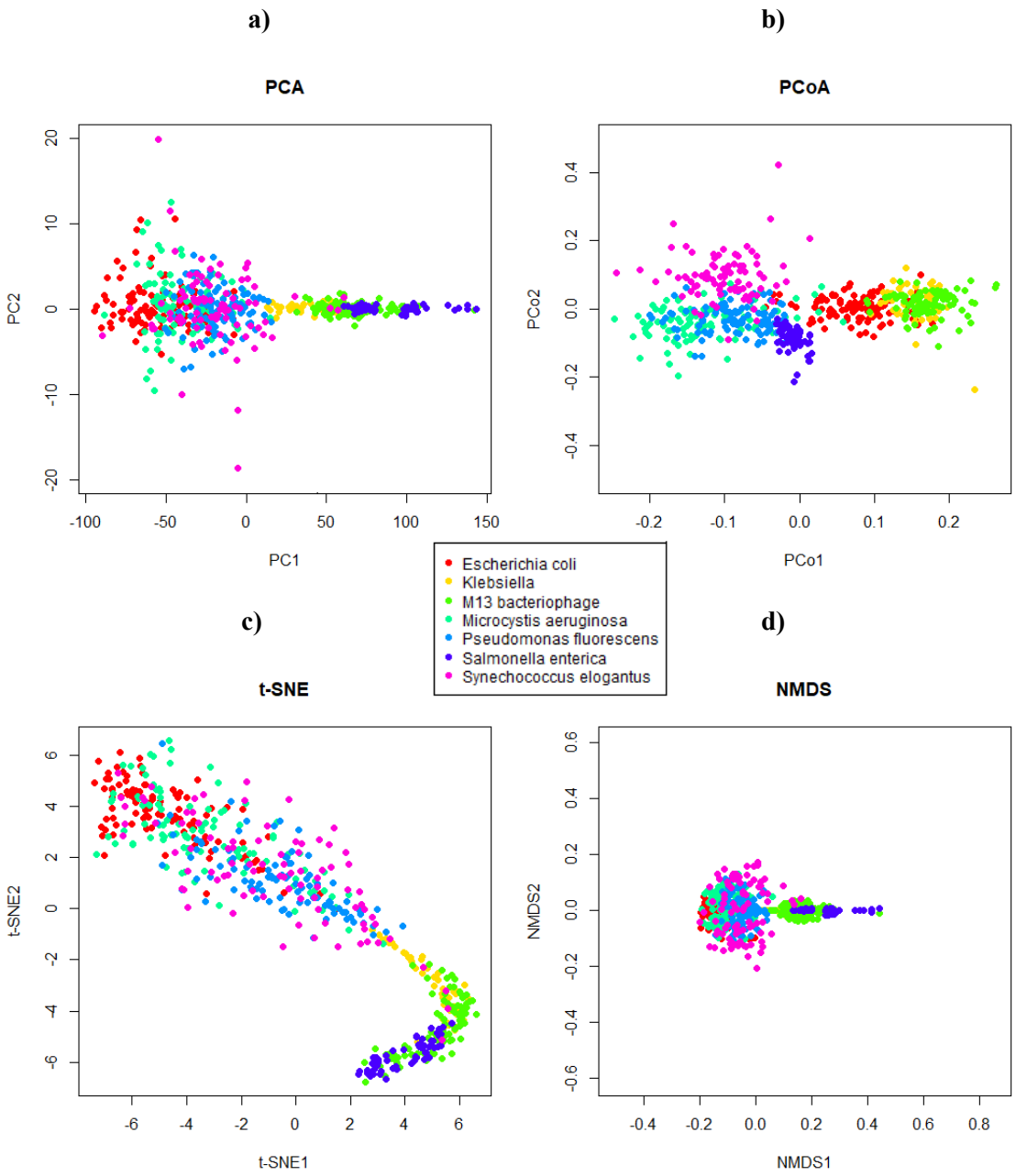
A. Príloha obrázkov vizualizácií
B. Príloha tabuliek
C. CD obsahujúce súbory dát, program a pdf verziu bakalárskej práce

A. Príloha tabuliek

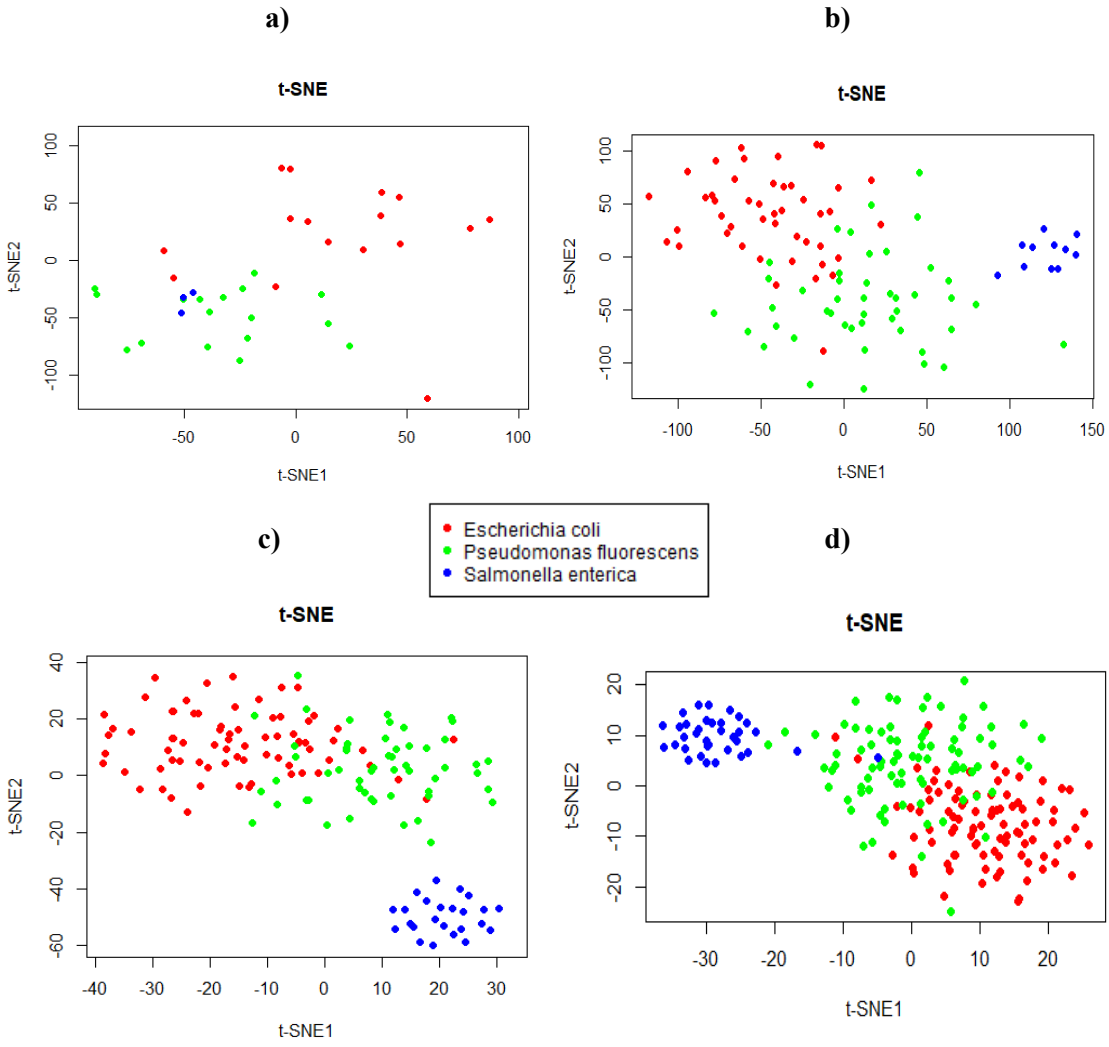
Príloh. tab. 1 Prehľad prístupových čísel doplňujúcich organizmov v simulovanom metagenomickom dataste

Názov organizmu	*Prístupové číslo*	*Prístupové číslo vzorky*	*Prístupové číslo behu*
Phage M13mp18	PRJEB8318	SAMEA3220411	ERR739513
Klebsiella pneumoniae	PRJEB14532	SAMEA4052062	ERR1474979

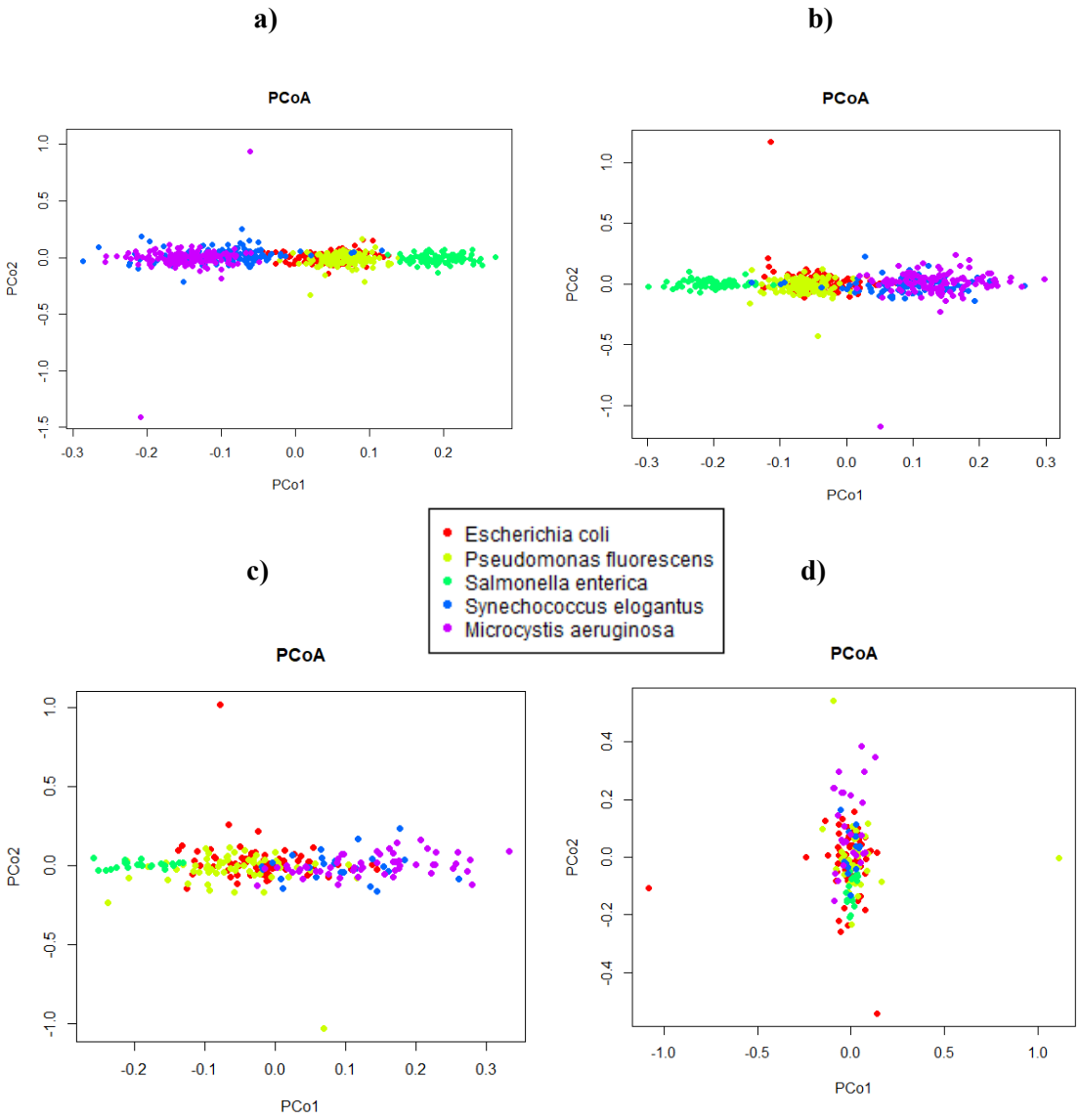
B. Príloha obrázkov vizualizácií



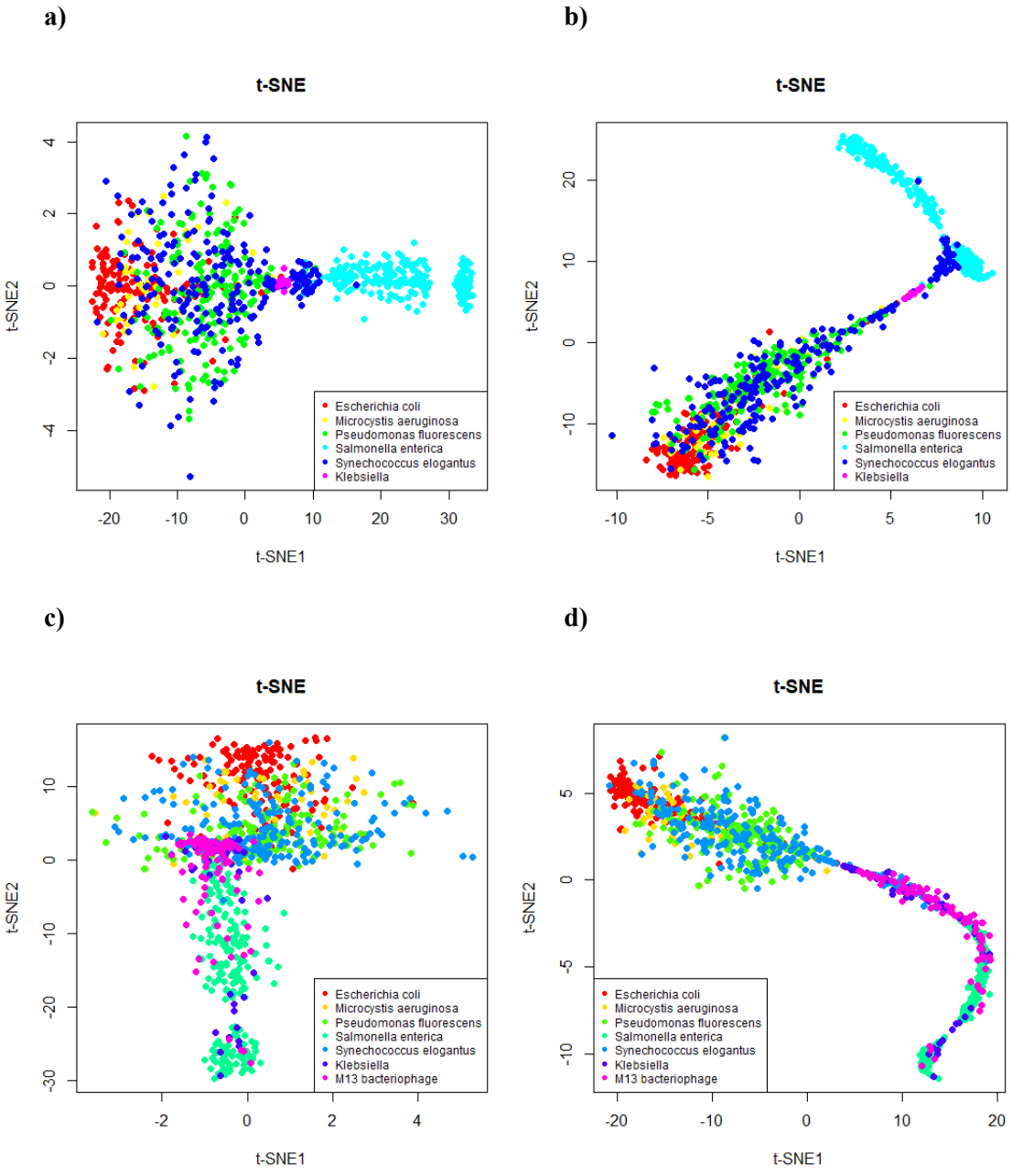
Príloh. obr.1 Aplikácia metód a) PCA, b) PCoA, c) t-SNE, a d) NMDS pri doplnení organizmov



Príloh. obr.2 Výsledok aplikovanej t-SNE metódy pri počte a) 50, b) 100, c) 150 a d) 200 zahrnutých sekvencií z jednotlivých organizmov



Príloh. obr.3 Výsledok aplikovanej PCoA metódy pri volení dĺžky trvania zahrnutých sekvencií na a) 10, b) 50, c) 150 a d) 200 sekúnd



Príloh. obr.4 Výsledok aplikovanej t-SNE metódy pri náhodnom počte zahrnutých sekvencií z každého organizmu v poradí podľa legendy 150, 50, 200, 300 , 50, 100 a 150 sekvencií pri nastaveniach a) perplexity = 64, 600 iterácií; b) perplexity = 54, 700 iterácií; c) perplexity = 64, 700 iterácií ; d) perplexity = 64, 500 iterácií