Univerzita Karlova v Praze

Přírodovědecká fakulta

Studijní program: Teoretická a evoluční biologie

Mgr. Jakub Rídl

Metagenomické profilování mikrobiálních společenstev

Metagenomic profiling of microbial consortia

Typ závěrečné práce

Disertační

Školitel: Mgr. Jan Pačes, Ph.D.

Ústav molekulární genetiky AV ČR, v.v.i.

Praha, 2016

Prohlášení:

Prohlašuji, že jsem závěrečnou práci zpracoval samostatně a že jsem uvedl všechny použité informační zdroje a literaturu. Tato práce ani její podstatná část nebyla předložena k získání jiného nebo stejného akademického titulu.

V Praze, 7. 6. 2016

Podpis

Abstrakt

Metody molekulární biologie umožňují studium mikrobiální diverzity a analýzu genů kódujících procesy a biochemické dráhy jednotlivých mikroorganismů a celých mikrobiálních společenstev. K tomu zásadně přispěl vývoj technologií masivně paralelního sekvenování DNA. Tyto metody rozšířily možnosti zkoumání diverzity od studia jednotlivých genomů modelových a v laboratoři kultivovatelných mikroorganismů přes jednoduché komunity v extrémním prostředí až k výzkumům komplexních mikrobiálních konsorcií. Tento experimentální přístup je založen na analýze celého metagenomu.

Pozornost je věnována ekosystémům negativně ovlivněným lidskou aktivitou, kde mikroorganismy dokáží nejen přežít, ale také adaptovat metabolismus k využívání a odbourávání látek toxických pro vyšší organismy. Příkladem je bakterie *Achromobacter xylosoxidans* A8 izolovaná z půdy kontaminované toxickými chlorbenzoáty. Sekvenace a analýza genomu *Achromobacter xylosoxidans* A8 umožnila studium genů kódujících enzymy zapojené do degradace chlorbenzoátů v kontextu kompletní genetické informace.

V extrémně kyselém prostředí bývalého dolu ve Zlatých Horách (Česká republika) vznikají zajímavé útvary bakteriálního biofilmu, želatinové stalaktity. Ty jsou tvořené jednou z taxonomicky nejjednodušších komunit s majoritním zastoupením dvou bakterií rodu *Ferrovum* a *Acidithiobacillus*. Sekvenace DNA izolované přímo ze vzorku biofilmu a bionformatická analýza získaných dat nám umožnila sestavení téměř kompletních genomů bez nutnosti pěstování těchto těžko kultivovatelných bakterií. Analýzou RNA jsme identifikovali aktivně transkribované geny konsorcia. Rekonstruované genomy představují unikátní kombinace genů vyvinuté adaptací na konkrétní ekosystém důlní vody ve Zlatých Horách.

Mnohem komplexnější systém představují společenstva půdních mikroorganismů. Sekvenováním celkové „půdní DNA“, amplifikací genů pro 16S rRNA a metagenomickým profilováním jsme odhalili vliv několika druhů rostlin na taxonomické složení a funkční potenciál mikrobiálních společenstev schopných biodegradace polutantů.

Klíčová slova bakteriální genomika, metagenomika, metatranskriptomika, sekvenování DNA, exprese genů, bioinformatika, metabolické dráhy, biodegradace

Abstract

Methods of molecular biology enable studies on microbial diversity based on analysis of genes encoding processes and biochemical pathways of individual microorganisms and also complete microbial consortia. For this a crucial step was elaboration of new technologies of high-throughput DNA sequencing. These methods made it possible to advance studies of diversity from analysis of genomes of model microorganisms easily cultivated in laboratories to simple communities living in extreme environments and further to complex microbial consortia. This experimental approach is based on metagenomic analyses.

Important are studies on ecosystems negatively affected by human activity where microorganisms not only survive but they can convert their metabolism to degrade compounds toxic for higher organisms. An example is bacterium *Achromobacter xylosoxidans* A8 isolated form soils contaminated by toxic chlorobenzoates. Sequencing and analysis of *Achromobacter xylosoxidans* A8 genome made it possible to study genes coding for enzymes that are involved in chlorobenzoates degradation in the context of the complete genetic background.

An interesting microbial biofilm – gelatinous stalactites – was discovered in an extremely acidic environment of the abandoned mine in Zlaté Hory (the Czech Republic). It is formed by a simple consortium with predominantly present bacterial genus *Ferrovum* and genus *Acidithiobacillus*. DNA sequencing of the biofilm sample and bioinformatic analysis of the obtained data enabled us to reconstruct two nearly complete genomes without growing individual bacterial strains that are difficult to cultivate. By RNA analysis expressed genes of the consortium were identified. Thus, metabolic processes of the present bacterial strains can be described. Comparative analysis shows unique properties of individual members of the analyzed consortium that are important in the particular natural conditions.

Much more complex systems are found in soil. We sequenced „soil DNA“, amplified 16S RNA genes and constructed metagenomic profiles of the samples. Using this approach we discovered effect of different plant species on taxonomic composition and functional potential of microbial communities that can degrade pollutants.

Key words bacterial genomics, metagenomics, metatranscriptomics, DNA sequencing, gene expression, metabolic pathways, biodegradation

Poděkování

Rád bych poděkoval svému školiteli Janu Pačesovi a také Čestmíru Vlčkovi, vedoucímu Oddělení genomiky a bioinformatiky, ÚMG AV ČR, kteří umožnili mou práci na prezentovaných projektech. Dále děkuji Ondřejovi Uhlíkovi a Hynkovi Strnadovi, bez kterých bych tuto práci nikdy nenapsal. Můj velký vděk patří také Milušce Hroudové a Marušce Balcové, bez kterých bych u práce nevytrval. Děkuji také kolegům Michalovi Kolářovi, Václavu Pačesovi a Šárce Pinkasové za ochotnou pomoc v mnoha směrech a Šárce Kocourkové a Marcele Vedralové za utváření příjemného pracovního kolektivu. Rád bych poděkoval své rodině a přátelům, kteří nikdy neztráceli víru v mou práci.

Grantová podpora

Projekty byly financovány Grantovou agenturou České republiky (GAČR): 13-28283S; Ministerstvem školství, mládeže a tělovýchovy České republiky (MŠMT ČR): 1M6837805002, 2B080031 a Výzkumným záměrem AV ČR: AV0Z50520514.

Seznam zkratek

RFLP	*restriction fragment length polymorphism*, polymorfismus délky restrikčních fragmentů
T-RFLP	*terminal restriction fragment length polymorphism*, polymorfismus délky terminálních restrikčních fragmentů
WGS	*whole genome shotgun*, celogenomové shotgunové sekvenování
NGS	*next generation sequencing*, sekvenování nové generace
PCR	*polymerarase chain reaction*, polymerázová řetězová reakce
qPCR	*quantitative RCR*, kvantitativní PCR
emPCR	emulzní PCR
dNTP	deoxynukleotid trifosfát
dATP	deoxyadenosin trifosfát
dATPαS	deoxyadenosin alfa-thio trifosfát
ATP	adenosin trifosfát
APS	adenosin fosfosulfát
bp	bazové páry
Mb	*mega base*, milion bazí
Gb	*giga base*, miliarda bazí
SMRT	*single molecule real time*, sekvenační metoda implementovaná společností Pacific Bioscience
CLR	*continuous long read*, souvislé dlouhé čtení produkované sekvenátory RSII (Pacific Bioscience)
CCS	*circular consensus sequence*, korigované konsensuální sekvence produkované sekvenátory RSII (Pacific Bioscience)
CDS	*coding sequence*, kódující sekvence
MID	*multiplex identifier*, identifikátor pro sekvenování různých vzorků během jedné reakce
RDP	*ribosomal database project*, databáze ribosomální RNA
OTU	*operational taxonomic unit*, operační taxonomická jednotka
HMM	Hidden Markovův model
MDS	*multidimensional scaling*, mnohorozměrné škálování
NMDS	*non-metric multidimensional scaling*, nemetrické mnohorozměrné škálování
PCA	*principal component analysis*, analýza hlavních komponent
CB	chlorbenzoát
DCB	dichlorbenzoát
TOC	*total organic carbon*, celkový obsah organického uhlíku
COD	*chemical oxygen deman*, biochemická spotřeba kyslíku
RTA	relativní transkripční aktivita
PCB	polychlorované bifenyly
MDA	*multiple displacement amplification*, metoda amplifikace genomé či metagenomvé DNA

Obsah

1 Úvod

Mikroorganismy obývají veškeré části povrchu planety Země a jejich činnost zásadním způsobem ovlivňuje ekosystémy, ve kterých žijí. Podílejí se na biochemických procesech, jsou zodpovědné za koloběh organických sloučenin a globální koloběh živin, díky čemuž veškeré organismy v biosféře závisejí na mikrobiálních aktivitách. Mikroflóra, která se vyskytuje ve vztahu k vyšším organismům, může přímo ovlivňovat jejich fyziologické procesy. Rozvoj metod molekulární biologie a především vývoj technologií masivního sekvenování v posledním desetiletí umožnil nejen studium nepřeberné míry mikrobiální diverzity v různých prostředích, ale také analýzu genů zodpovědných za kódování funkčních procesů mikroorganismů a celých mikrobiálních společenstev. Zatímco od zveřejnění první kompletní sekvence bakteriálního genomu uplynulo právě dvacet let (Fleischmann et al., 1995), postupné zavedení sekvenačních platforem od roku 2005 výrazně zjednodušilo proces sekvenační analýzy, snížilo náklady a zpřístupnilo genomové projekty jednotlivých mikroorganismů řadě výzkumných skupin. Společně s tím se otevřely nové možnosti ve studiu více či méně komplexních společenstev obývajících nejrůznější ekosystémy. Tato oblast označovaná metagenomika vychází ze dvou jevů: (1) jen méně než procento mikroorganismů dokážeme kultivovat a následně analyzovat v laboratorních podmínkách a (2) na jednotlivých funkčních procesech v ekosystému se podílí celá řada druhů tvořících mikrobiální konsorcia (Handelsman et al., 1998).

Paralelně s překotným nárůstem dat generovaných sekvenačními přístroji vyvstala nutnost výpočetního zpracování velkého objemu informací a značná část nově vzniklé disciplíny nazývané jako bioinformatika se zabývá právě analýzou souborů sekvencí.

Cílem této práce je poskytnout přehled sekvenačních technologií a bioinformatických nástrojů a současných možností jejich využití v bakteriálních genomových projektech a metagenomických analýzách na konkrétních příkladech našich vybraných studií.

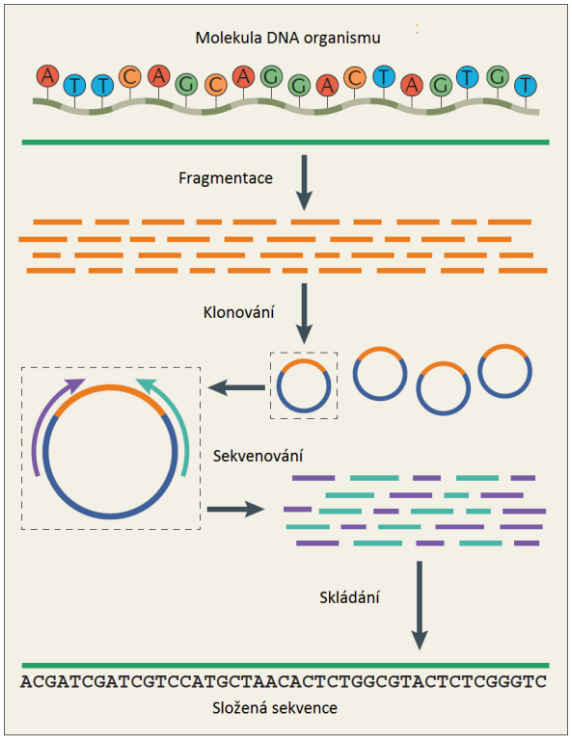
1.1 Studium mikroorganismů v historické perspektivě

Ačkoli moderní metody umožnily skokový nárůst studií mikroorganismů, jejich konsorcií a vztahů s dalšími biotickými i abiotickými složkami ekosystému, stejný zájem provází vědce již stovky let. V roce 1676 uveřejnil Antoni van Leeuwenhoek pozorování bakterií například ve slinách za pomoci jednoduchého mikroskopu, čímž se zapsal do historie jak mikrobiologie, tak mikroskopie. (Schierbeek, 1969). Mikroskop se postupně stal základním

nástrojem mikrobiologie. Další snahou Leeuwenhoekových následovníků byla především izolace a kultivace mikroskopických organismů například na plátcích brambor či želatiny. O dvě stě let později rozpracoval Robert Koch metody pěstování bakterií na pevných médiích pro potřeby jejich vizualizace, mikrofotografie, stanovení počtu buněk a analýzy fyziologických vlastností (Blevins & Bonze, 2010). Rozlišení mikroskopických analýz zásadně zlepšilo zavedení technik barvení, například dle Grama, Ziehl-Nielsena či Schaeffera & Fultona (Beveridge, 2001; Blevins & Bronze, 2010). Evidentní rozdíl mezi počtem mikroorganismů sledovaných v mikroskopu a pěstovaných na miskách (Staley & Konopka, 1985) poukázal na fakt, že ačkoli se v přírodě vyskytují všudypřítomně, potřebují mikroorganismy k růstu speciální podmínky. Sergei Winogradsky ve snaze vytvořit kultivační média, která by co nejlépe simulovala přirozené podmínky života mikroorganismů, položil základy mikrobiální ekologie a studia role mikroorganismů v prostředí (Ackert, 2012). I přes veškerou optimalizaci kultivačních podmínek a rozšíření repertoáru kultivačních metod jsme dnes stále schopni v laboratoři pěstovat jen méně než jedno procento mikroorganismů (Riesenfeld et al., 2004; Lozupone & Knight, 2008; Zhang & Xu, 2008).

Dlouhou dobu zůstávalo studium mikroorganismů omezeno na fenotypové znaky: morfologii, rychlost růstu, formování a vzhled kolonií či selekci některých biochemických vlastností ve specifických médiích. Rok 1977 zaznamenal hned dva milníky. Woerse a Fox publikovali použití prvních molekulárních markerů pro klasifikaci organismů, genů kódujících ribosomální RNA (rRNA) (Woerse & Fox, 1977), které následně umožnily rozlišení třech hlavních větví stromu života (Woerse et al., 1990) a dodnes jsou nástrojem fylogenetických analýz. Ve stejném roce Frederik Sanger uveřejnil spolu s kolegy metodu automatické sekvenace, tedy stanovení pořadí jednotlivých bazí molekul DNA (Sanger et al., 1977). Od 80. let 20. století byly postupně zaváděny další metody molekulární biologie, které vedly nejen ke studiu jednotlivých, v laboratoři pěstovaných mikroorganismů, ale také umožnily posun studia až na úroveň celých populací nekultivovatelných organismů. Významný vliv měly především techniky umožňující získání množství kopií úseků DNA – klonovací techniky a metody amplifikace prostřednictvím polymerázové řetězové reakce (PCR). Pro účel analýzy mikrobiálních populací byly aplikovány metody klonování rRNA genů doprovázené Sangerovým sekvenováním (Leigh, 2006; Zhang a Xu, 2008), fluorescenční *in situ* hybridizace (FISH) (Wagner et al., 2003), metody denaturační a tepelné gradientové gelové elektroforézy (DGGE, TGGE) (Kirk et al., 2004; Zhang a Xu, 2008) a restrikční metody a analýzy délkových polymorfismů vzniklých fragmentů (RFLP, T-RFLP) (Liu et al., 1997; Marsh, 2005).

V molekulárně-genetické sféře dějin mikrobiologie můžeme rozlišit tři hlavní revoluční události (Loman & Pallen, 2016). První je zavedení postupu shotgunového sekvenování (v angličtině *shotgun sequencing* nebo též *whole-genome shotgun sequencing* se zkratkou WGS), kdy je celková DNA izolovaná z buněk štěpena na krátké fragmenty (metaforicky „rozstřelena“, z čehož pramení užití anglického termínu pro brokovnici – *shotgun*), každý fragment je odděleně namnožen klonováním v buňkách hostitelského organismu a sekvenován (obrázek 1.1). Rekonstrukce původních genomů či alespoň jejich delších, souvislých úseků probíhá následně sestavením přečtených fragmentů (jednotlivých čtení, z anglického *read*) podle jejich překryvů. Využití počítačových programů pro sestavení čtení rozpracoval již koncem 70. let Roger A. Staden (Staden, 1979) a software nesoucí jeho jméno je rozvíjen a používán dodnes (Staden, 1996). Metoda shotgunového sekvenování byla následně aplikována také na čtení fragmentů celkové mikrobiální DNA z environmentálních vzorků a pro takové výzkumy se vžil název metagenomika – oblast zaměřená na analýzu metagenomu jakožto souhrnu genomů přítomných organismů lišící se od genomiky, která se zabývá jednotlivými genomy zvlášť (Handelsman et al., 1998).



Obrázek 1.1: Znázornění postupu shotgunového sekvenování (převzato a upraveno z Loman & Pallen, 2016).

Druhou revoluční událostí bylo zavedení metod masivně paralelního sekvenování v angličtině nazývaných buď pojmem *high-throughput sequencing* referujícím k masivní

produkci dat či častěji jako *next-generation sequencing* (NGS), tedy sekvenování nové generace. Metody využívají rovněž shotgunový přístup fragmentace molekul DNA v prvním kroku tvorby tzv. sekvenačních knihoven. Ty jsou však namnoženy metodou PCR a následně analyzovány přístroji schopnými paralelně číst miliony až miliardy sekvencí. Z dnešního pohledu bychom mohli mluvit spíše o druhé generaci sekvenátorů, neboť v současnosti dochází k rozvoji třetí generace platforem schopných dlouhého čtení a/nebo sekvenační analýzy pouze jedné molekuly jako výchozího materiálu. Jejich využití lze považovat za třetí revoluční událost (Loman & Pallen, 2016).

1.2 Vývoj sekvenačních metod

Pojem sekvenování všeobecně popisuje stanovení pořadí základních stavebních složek řetězců biopolymerů, tedy sekvence jejich primární struktury. V případě nukleových kyselin (DNA, RNA) jde o pořadí nukleotidů, u proteinů jsou to jednotlivé aminokyseliny. Zajímavým faktem je, že historicky zavedení sekvenačních technik pro čtení řetězce proteinů předcházelo objevům metod pro sekvenování molekul DNA (Edman, 1949; Sanger & Tuppy, 1951). Vzhledem k tomu, že pořadí aminokyselin v proteinech je dáno pořadím bazí určitých úseků nukleových kyselin (kódujících genů), představovalo sekvenování proteinů možnost, jak se na molekulární úrovni nepřímo dobrat genetického kódu určujícího dědičné znaky organismů. Rovněž metody sekvenování RNA byly vědcům přístupné dříve, než byly představeny hlavní techniky sekvenování DNA v druhé polovině 70. let 20. století. Proto také první kompletní geny, jejichž primární struktura byla odhalena, pocházely z organismů s RNA genomy, tedy z virů. Kulminaci využití tehdejších metod sekvenování RNA představovala analýza genomu bakteriofágu MS2 publikovaná v roce 1976 (Fiers et al., 1976).

1.2.1 Metody první generace

Počátky metod DNA sekvenace jsou spjaty především s trojicí jmen Sanger, Maxam a Gilbert. V roce 1973 popsal britský vědec Frederik Sanger společně s kolegy metodu založenou na *in vitro* syntéze vlákna komplementárního k jednovláknové předloze DNA enzymem polymerázou – obdobně, jako při replikaci DNA v živých buňkách – a využití radioaktivního značení pro detekci bazí nově vznikající molekuly DNA. Postup byl dále modifikován a stal se základem pro tzv. „plus-mínus“ metodu (Sanger & Coulson, 1975), vyvinutou speciálně pro přečtení jednovláknového DNA genomu bakteriofágu phiX174, jehož sekvence byla

publikována v roce 1977 (Sanger et al., 1977b). Ještě v témže roce však Sangerův tým uveřejnil novou metodu terminátorové sekvenace, která opět využívá syntézu DNA polymerázou a značení bazí, ale základem pro detekci jejich pořadí je zastavení polymerace v místě aktuálně čtené baze (Sanger et al., 1977a). V roce 1977 také představili Allan Maxam a Walter Gilbert alternativní metodu sekvenace chemickou modifikací bazí a štěpením řetězce DNA v jejich pozicích (Maxam & Gilbert, 1977). Zpočátku tato metoda převládala, ale postupem času se dostala do popředí Sangerova metoda, jednak díky inovovanému postupu neradioaktivního značení a také postupnou automatizací procesu samotné sekvenace. Princip modifikace bazí a štěpení řetězců DNA vynalezený Maxamem a Gilbertem však našel uplatnění ve výzkumu vazebných míst pro proteiny interagující s DNA a identifikaci metylovaných bazí v rámci epigenetického studia (Ohmori et al., 1978).

Oba přístupy jsou založeny na produkci značného množství různě dlouhých fragmentů analyzované DNA se značenými bázemi na koncích a jejich následné separaci, vizualizaci a analýze. Právě separace DNA molekul pomocí gelové elektroforézy představuje jeden z hlavních předpokladů pro zavedení první generace metod sekvenování, zvláště využití polyakrylamidových gelů umožňujících rozdělit řetězce DNA dle jejich délky s rozlišením na úrovni jednoho nukleotidu. Pozdější automatické sekvenátory uvedené na trh v druhé polovině 90. let 20. století a používané dodnes využívají navíc možnosti kapilární elektroforézy a automatického vyhodnocení dat a jsou schopné sekvenovat desítky různých molekul DNA během jednoho běhu přístroje.

Pro vizualizaci značených nukleotidů je nezbytný dostatečně silný signál. Dalším důležitým předpokladem je proto namnožení výchozího úseku DNA, který je následně podroben sekvenační reakci. Každá značená a analyzovaná baze se tedy vyskytuje na mnoha molekulách DNA, nikoli pouze na jednom vlákně. Takový signál by byl velice slabý a spadal by pod detekční limit tehdejších vyhodnocujících metod. Od počátku proto sekvenační metody závisí na rozvoji klonovacích technik, které umožňují namnožit studovaný úsek DNA v hostujících buňkách bakterie *Escherichia coli* kultivované v laboratoři. Ovšem právě nutnost připravovat každý fragment individuálně je jeden z hlavních limitujících faktorů původního shotgunového sekvenování využívajícího Sangerovu metodu.

1.2.2 Metody druhé generace

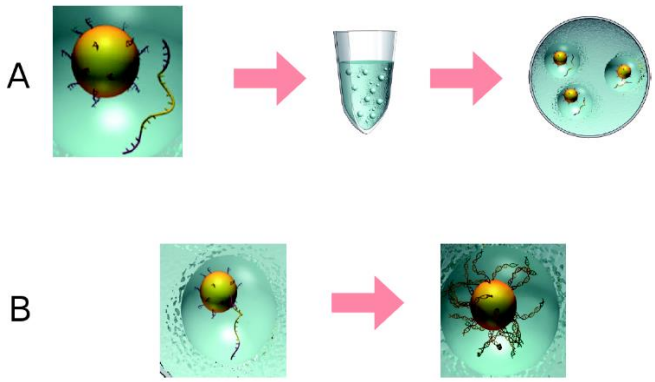
Velká sekvenační centra disponující sály naplněnými řadou nejvýkonnějších kapilárních sekvenátorů umožňují získání velkého objemu dat a analýzu celých genomů i

mnohobuněčných organismů. Příkladem mohou být velké projekty jako sekvenace lidského genomu, který začínal již v roce 1990 (http://web.ornl.gov/sci/techresources/Human_Genome/index.shtml). Pro individuální laboratoře je však využití tradičních metod sekvenace pro analýzu byť i mnohonásobně menších a jednodušších genomů nadmíru pracným a drahým úkolem. Výraznou změnu znamenalo zavedení genomových sekvenátorů nové generace. Od roku 2005 byly postupně uvedeny na trh dvě hlavní: (1) GS (*Genome Sequencer*) přístroj vyvinutý firmou 454 Life Sciences, dnes spadající pod společnost Roche a (2) GA (*Genome Analyzer*) společnosti Solexa, následně koupená firmou Illumina, která postup implementovala v přístrojích HiSeq, MiSeq či NextSeq. Pro úplnost zde uveďme také instrument SOLiD od společnosti Aplied Biosystems (ABI), využívající ojedinělý způsob stanovení bazí (sekvenaci ligací) a platformu PGM od Ion Torent, která umožňuje na mikroskopických potenciometrech detekci změny napětí po uvolnění protonu při polymeraci vlákna DNA. Oba později zmíněné přístroje však stojí spíše v pozadí a v následujícím přehledu jim nebude věnována zvláštní pozornost.

1.2.2.1 Základní principy přípravy sekvenačních knihoven

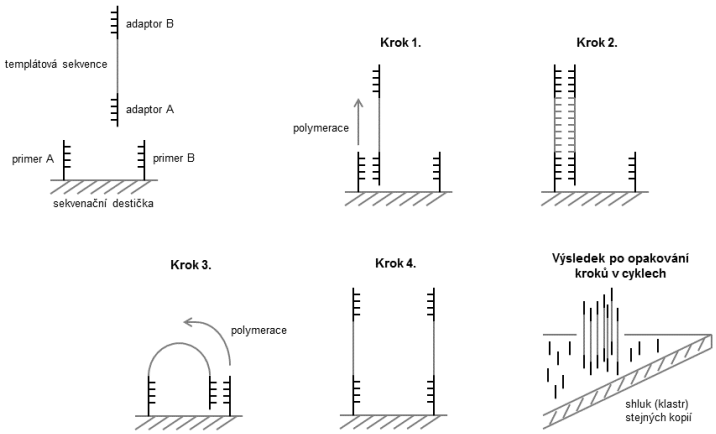
Standardní příprava sekvenačních knihoven pro všechny metody masivního sekvenování sleduje jednotný scénář: fragmentace DNA, ošetření konců fragmentů a navázání sekvenačních adaptorů, amplifikace následovaná samotnou sekvenací. Nejběžnějším způsobem fragmentace DNA je fyzické nalámání delších molekul pod tlakem s využitím stlačeného dusíku (nebulizace) či ultrazvukem (sonikace). Místa, ve kterých ke štěpení dochází, jsou náhodná a u obou řetězců dvouvláknové DNA se mohou o několik bazí lišit. Konce tedy nejsou zarovnané, ale vznikají na nich přesahy. Ty je potřeba odstranit, respektive doplnit kratší vlákna polymerázou. Na zarovnané konce jsou dále připojeny enzymem ligázou sekvenační adaptory, uměle syntetizované oligonukleotidy (krátké molekuly DNA). V následujících krocích slouží pro hybridizaci a stabilizaci na nosné medium, dále jako místa pro nasednutí primerů při PCR amplifikaci a jsou také komplementární pro primery iniciující samotnou sekvenační reakci. Fragmenty DNA jsou nakonec namnoženy PCR amplifikací pro dosažení dostatečné míry signálu při samotné sekvenační reakci. Úkolem však je každou výchozí molekulu společně s nově vznikajícími kopiemi fyzicky separovat od všech ostatních fragmentů. Za tímto účelem byly vyvinuty dva postupy amplifikace: (1) emulzní PCR a (2) můstková amplifikace (*bridge amplification*).

Emulzní PCR je postup využívaný k přípravě vzorků pro instrumenty GS (Roche), SOLiD (ABI) a PGM (Ion Torent). K fyzické separaci fragmentů jsou použity mikroskopické kuličky, které na svém povrchu nesou kovalentně přichycený primer komplementární k adaptorové sekvenci na 5' konci jednovláknové molekuly amplifikovaného fragmentu DNA. Fragmenty sekvenační knihovny jsou hybridizací 5' konce přichyceny k primerům na kuličkách. Elongací primerů v průběhu polymerace vzniká řetězec komplemetární k hybridizované sekvenci, který zůstává kovalentně navázán na kuličce. Aby amplifikace probíhala exponenciální řadou, jsou do reakce přidány také primery komplementární k adaptorové sekvenci na 3' konci. Postupnou elongací primerů navázaných na kuličkách vzniká klonální populace tisíce imobilizovaných kopií pocházející původně z jedné molekuly fragmentu sekvenační knihovny. To, že na jednu kuličku byla hybridizována pouze jedna výchozí molekula DNA a volné produkty amplifikace nemohly uniknout a smíchat se s milióny ostatních reakcí probíhajících paralelně, je zajištěno právě emulzní cestou. Na počátku se amplifikační mix obsahující sekvenační kuličky, molekuly sekvenační knihovny, polymerázu, volné nukleotidy, volné primery a amplifikační pufr roztřepe v oleji. Tím vzniknou malé kapičky reakční směsi, obsahující v ideálním případě jednu kuličku a jednu molekulu DNA. Celá amplifikace pak probíhá v těchto mikroreaktorech obklopených prostředím oleje. Nově vznikající vlákna nejsou schopná uniknout, ale jsou v opakujících se cyklech opět hybridizována na stejnou sekvenační kuličku. Vznik odpovídajících kapiček je zajištěn správným objemem použitého oleje a reakční směsi a vhodným nastavením intenzity třepání. Zachycení pouze jedné kuličky a jednoho fragmentu DNA je ovlivněno poměrem, ve kterém se kuličky a molekuly do směsi přidávají (obrázek 1.2). Kuličky s přichycenými vlákny jsou dále promyty a použity pro sekvenční reakci. Častým jevem je, že do kapky s jednou kuličkou je přece jen zachycena více než jedna molekula DNA a výsledné namnožené sekvence jsou směs z různých templátů. V takovém případě ani signál produkovaný v cyklech sekvenace není jednotný a takové sekvence jsou automaticky rozpoznány a vyřazeny jako chybné.



Obrázek 1.2: Znázornění průběhu emulzní PCR: A) připravené fragmenty DNA jsou smíchány se sekvenačními kuličkami a PCR amplifikačním mixem. Roztřepáním v oleji vziknout kapičky amplifikačního mixu ideálně vždy s jednou kuličkou a jednou sekvencí DNA. B) během samotné PCR dochází k elongaci primeru kovalentně navázaného na kuličku. Po 50 cyklech reakce je povrch kuličky zaplněn kopiemi původní, templátové sekvence (převzato a upraveno z www.454.com)

Můstková (*bridge*) amplifikace byla vyvinuta pro přístroje firmy Illumina. Pro imobilizaci připravených fragmentů DNA slouží destička, na jejímž povrchu jsou přichyceny oba typy primerů komplementární k oběma koncům fragmentu. Ten je proto hybridizací zachycen za oba konce a vlákno vytváří jakýsi můstek mezi dvěma typy primerů v těsné blízkosti u sebe. Nově syntetizované vlákno, které vzniká v jednom cyklu PCR amplifikace elongací jednoho z primerů polymerázou, utváří v dalším cyklu po denaturaci dvouvláknového komplexu opět můstek díky hybridizaci volného konce na nový, sousední primer. Namnožené molekuly tak vytváří na destičce shluky (klastry) tisíce sekvencí v těsné blízkosti. Destička s klastry je poté vložena do sekvenátoru a podle signálu při sekvenaci jsou shluky shodných sekvencí rozpoznány jako jednotlivé body rozmístěné po ploše destičky (obrázek 1.3).

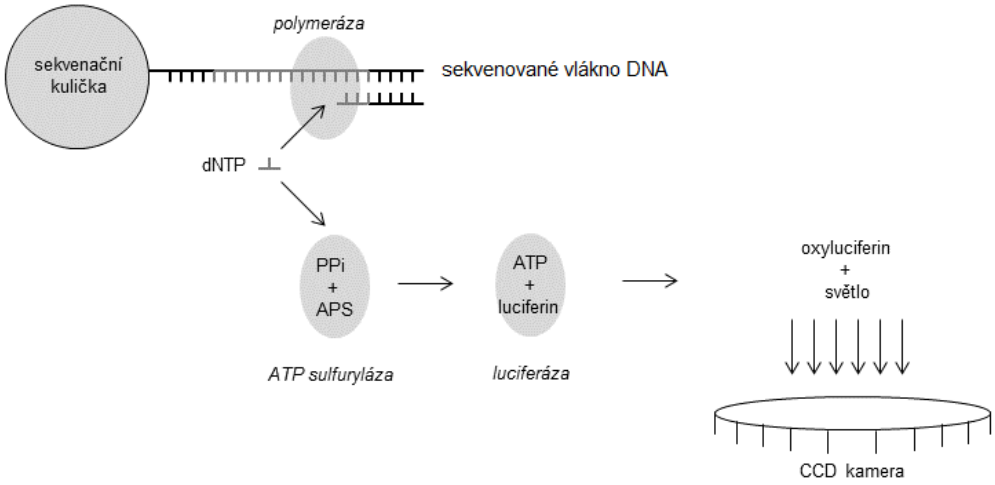


Obrázek 1.3: Grafická reprezentace procesu můstkové amplifikace na příkladu vzniku jednoho shluku stejných kopií templátové DNA.

1.2.2.2 Pyrosekvenace (454 / Roche)

Společnost 454 Life Sciences (později koupená mezinárodním farmaceutickým gigantem Roche) přišla v roce 2005 jako první na trh se sekvenátorem druhé generace umožňujícím masivní paralelní analýzu stovek tisíců krátkých úseků DNA během jedné procedury (Margulies et al., 2005). Genomový sekvenátor s označením zkratkou GS (a doprovázeným identifikátorem jednotlivé verze přístroje) umožňuje provádět v miliónech oddělených mikroreaktorech metodu pyrosekvenace vyvinutou v 90. letech 20. století Palem Nyrénem a Mostafou Ronaghim (Ronaghi et al., 1996). Pyrosekvenace opět využívá syntézu vlákna komplementárního k jednovláknovému templátu obdobně jako Sangerova metoda, ale přidání každého nukleotidu je monitorováno za pochodu skrze sledování uvolnění molekuly pyrofosfátu bez nutnosti separace vláken gelovou elektroforézou. Pyrofosfát (difosfát), přirozený produkt inkorporace volného nukleotidu polymerázou na 3' konci nově syntetizovaného řetězce, vstupuje do enzymatické kaskády, jejímž konečným výsledkem je emitace světelného signálu, který je detekován CCD kamerou. Kaskáda využívající enzymy přidané na počátku do reakce zahrnuje přeměnu pyrofosfátu na molekulu ATP enzymem sulfurylázou; ATP je dále použit enzymem luciferázou k produkci světelného záblesku. Pyrosekvenační reakce běží ve čtyřech opakujících se cyklech (*flows*), kdy je přidán vždy pouze jeden typ nukleotidu (dNTP), zaznamenán signál v mikroreaktorech, kde došlo v dané pozici (podle pořadí cyklu) k inkorporaci nukleotidu a následuje odmytí volných dNTP před přidáním dalšího typu nukleotidu v následujícím cyklu. Aby v cyklu, kdy je přidáván adenosintrifosfát, nedošlo k jeho přímému využití luciferázou pro vysvícení světelného signálu, je místo běžného dATP používán modifikovaný deoxyadenosinthiotrifosfát (dATPαS), který luciferáza není schopna použít jako substrát pro reakci (obrázek 1.4).

Paralelního čtení přibližně dvou milionů sekvencí je dosaženo na speciální pikotitrační destičce. Jde o skleněnou plochu s mikroskopickými jamkami, do každé z nich je schopna zapadnout právě jedna sekvenační kulička s kopiemi DNA fragmentu namnoženými emulzní PCR. Celá destička je snímána kamerou skrze její dno a pozice jamek se zasednutými sekvenačními kuličkami je automaticky rozpoznána podle pozitivního signálu během prvních cyklů díky rozpoznávací klíčové sekvenci, kterou jsou opatřeny sekvenační adaptory pro tvorbu knihoven.



Obrázek 1.4: Znázornění průběhu pyrosekvenační reakce: inkorporace volného nukleotidu (dNTP) polymerázou vede k uvolnění molekuly pyrofosfátu (PPi). Ta je společně s předem přidaným APS (adenosin fosfosulfát) využita enzymem ATP sulfurylázou k produkci ATP. ATP je využit k přeměně přidaného luciferinu na světelný signál.

Pokud se v analyzované sekvenci vyskytují dvě a více shodných bazí (tzv. homopolymery), jsou vždy čteny během jediného cyklu. Jejich přítomnost se projeví větším množstvím uvolněného pyrofosfátu a tím také silnější mírou signálu v daných pozicích. Podle intenzity emitovaného světla je potom vypočtena délka daného homopolymeru (počet opakujících se nukleotidů). Teoreticky si lze představit, že dvě shodné baze za sebou budou znamenat dvojnásobný signál, tři trojnásobný a tak dále. Prakticky se však intenzita signálu nezvyšuje vždy rovnoměrně, převážně vlivem pozadí snímaných obrázků a kolísající efektivity enzymatické kaskády při pyrosekvenaci. To je zdrojem poměrně časté chyby při čtení homopolymerů, která představuje jednu z hlavních nevýhod pyrosekvenace oproti všem ostatním metodám. Chybu tohoto typu však můžeme v datech předpokládat a do určité míry korigovat. Odhlédneme-li od inzercí a delecí vlivem homopolymerů, pak se chybovost přístroje GS pohybuje v rozmezí jedné substituce na 100 – 1.000 nukleotidů. Je tedy vyšší, než u Sangerovy metody (která vykazuje 1 chybu na 10.000 – 100.000 nukleotidů), ale na druhou stranu 454 pyrosekvenování dosahuje jedné z nejnižších chybovostí mezi metodami masivního sekvenování nové generace.

Napříč sekvenačními technologiemi lze obecně sledovat nárůst chybně přečtených bazí spolu s rostoucím počtem cyklů sekvenace, tedy s délkou sekvenovaného úseku. Dochází k tomu vlivem snižující se účinnosti enzymů a/nebo jejich postupnou ztrátou. Nárůst chybovosti s rostoucím pořadím bazí je však v porovnání s ostatními genomovými sekvenátory u pyrosekvenace pozvolnější a dovoluje obdržet správné sekvence o poměrně dlouhé délce.

Nejvýkonnější varianta 454 sekvenátoru s názvem GS FLX Titanium+ (často jen GS FLX+) poskytuje prakticky použitelné sekvence o délce 800 bp – 1.000 bp. Paralelně vyhodnotí kolem 1 milionu sekvencí s celkovým výstupem asi 800 milionů bazí (Mb). Cena se pohybuje kolem 10 amerických dolarů za 1 Mb dat. Firma Roche přišla také se „stolní“ verzí pyrosekvenátoru GS Junior, jehož pořizovací náklady jsou nižší a nabízí možnost sekvenování v menším objemu při nižší spotřebě chemikálií. Výstup jedné procedury GS Junior zahrnuje asi 100 tisíc sekvencí o délce přibližně 400 bp.

1.2.2.3 Reversibilní terminátorová sekvenace (Solexa / Illumina)

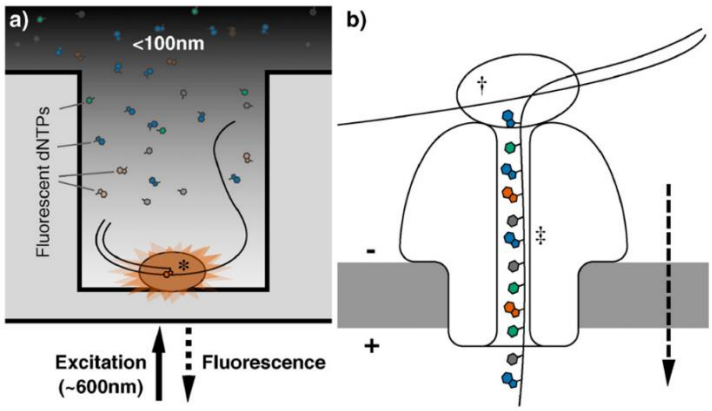
Podobně jako u Sangerovy metody jde opět o princip sekvenace syntézou. Dokonce využívá i obdobné metody detekce individuálních bazí na základě zastavení polymerázové reakce a přečtení fluorescenčně odlišené baze, ovšem opět bez nutnosti gelové elektroforézy. Detekce je prováděna přímo při syntéze vlákna a její zastavení je pouze dočasné a po detekci signálu je opět pokračováno v inkorporaci následující komplementární baze k templátové předloze. Díky této možnosti dočasného přerušení a opětovného pokračování syntézy dostala také název reversibilní terminátorová reakce. Reagencie jsou v cyklech pumpovány rovnou na destičku s klonálně namnoženými klastry, která je výsledkem můstkové amplifikace. Jako templát pro sekvenaci však slouží pouze jedno z vláken. Aby druhé vlákno neinterferovalo s probíhající sekvenací, je odštěpeno na základě modifikace bazí jednoho z primerů navázaných na povrchu destičky a odmyto.

Sekvenační přístroje od firmy Illumina aplikující reversibilní terminátorovou sekvenaci zaznamenávají v průběhu reakce rychlejší nárůst chyb, než je tomu v případě platformy firmy Roche. Vedle úbytku a vyčerpání samotné polymerázy, má technologie také několik specifických zdrojů chyb. Jedním je například právě dočasná terminace syntézy. Při nedostatečném zastavení polymerace v některých pozicích může dojít k inkorporaci dalšího nukleotidu ještě v témže cyklu. Na druhou stranu opačná chyba – trvalá terminace v některém z cyklů – vede k postupnému úbytku míry signálu. K takovému jevu dochází rovněž postupným vysvěcováním fluorescenčních značek. Navíc, přestože Illumina používá čtyři rozdílné fluorescenční značky, jsou excitovány dvěma lasery. Značky pro A/C a G/T jsou excitovány vždy stejným laserem a vykazují signál o podobném spektru. Proto může dojít k chybné substituci mezi A/C a G/T. U produkovaných dat tak můžeme zaznamenat chybovost zhruba o frekvenci 1:10 – 1:100. Rovněž délka jednotlivých čtení je znatelně kratší. Nejdelší možná čtení 300 bp poskytuje instrument MiSeq schopný produkce asi 25 mil. sekvencí v jediném běhu.

Výkonnější přístroj HiSeq přečte 1.000 Gb bazí s cenou pod 50 centů za 1 Mb, ale délka čtení je maximálně 150 bp. Krom toho však umožňují Illumina technologie čtení sekvencí také z druhého směru po několika cyklech dodatečné můstkové amplifikace a denaturace a odštěpení jednoho z vláken. V tomto kontextu používá společnost Illumina označení *paired-end*.

1.2.3 Metody třetí generace

V současnosti je pozornost ve výzkumu sekvenačních technik zaměřena právě na eliminaci fragmentace a amplifikace – tedy na přístupy schopné analyzovat jen jedinou molekulu DNA (tzv. s*ingle-molecule* přístup) a umožňující čtení dlouhých souvislých nukleotidových vláken. Příkladem třetí generace sekvenátorů je instrument PacBio RSII pracující na principu pojmenovaném jako SMRT (*single-molecule real time* sekvenování) od společnostmi Pacific Bioscience. Princip je opět založen na syntéze vlákna komplementárního k sekvenovanému jednovláknovému templátu, ovšem polymeráza je imobilizovaná na dně mikroskopických jamek. Do nich zapadne jediná z předem připravených molekul DNA. Pro syntézu vlákna jsou přidány značené nukleotidy. Při polymeraci jednoho nukleotidu je zaznamenán signál *real-time* přes dno jamky (obrázek 1.5a). RSII dokáže produkovat dlouhá čtení kolem 15 kb (CLR, *continuous long read*), kdy projde templátová molekula polymerázou jen jednou. Taková čtení také vykazují vyšší chybovost v rozmezí 11 až 15 % (Eid et al., 2009; Koren et al., 2012). Metoda však umožňuje ještě druhý typ sekvenace, při kterém je templátová DNA připravena ve formě krátké cirkulární molekuly, což umožňuje její několikanásobné přečtení dokola a výsledkem je korigovaná konsensuální sekvence (CCS, *circular consensus sequence* čtení). Původní studie udávala chybovost pod 1 % (resp. přesnost 99.3 % při 15 násobném průchodu) (Eid et al., 2009). Jiao et al. (2013) popisují chybovost u 1,5 kb dlouhých cirkulárních molekul 1,3 % po aplikaci navrženého algoritmu pro kontrolu kvality a korekci chyb. Přístroj RSII je však nejen velmi velký, ale také drahý (přibližně 700 tis. dolarů) a jsou jím vybavena jen velká sekvenační centra. Právě na letošní rok však společnost Pacific Bioscience ohlásila uvedení menšího a dostupnějšího instrumentu Sequel.



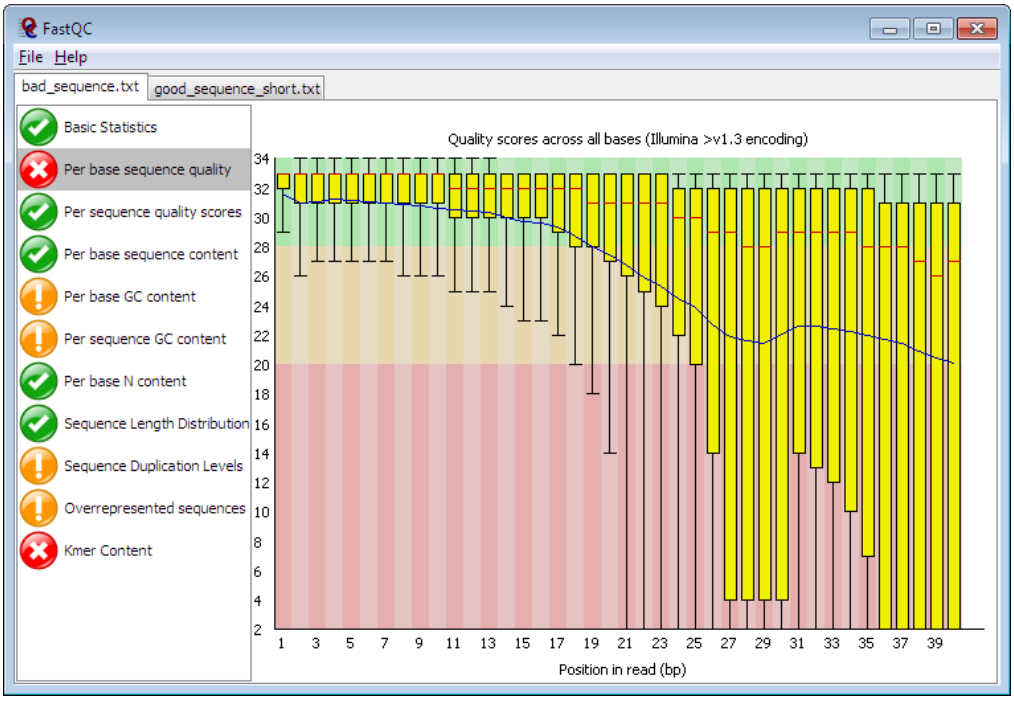
Obrázek 1.5: Znázornění sekvenačních metod třetí generace. A) PacBio SMRT systém s imobilizovanou polymerázou (hnědý ovál) na dně mikrosropické jamky, B) metoda Oxford Nanopore detekující pořadí nukleotidů při průchodu jednovláknové molekuly DNA mikroskopickým pórem (převzato Heather & Chain, 2016).

Zcela odlišnou metodu stanovení pořadí nukleotidů jedné molekuly DNA využívá přístroj MinION od firmy Oxford Nanopores. Proteinové nanopory na syntetické membráně, kterými prochází jednovláknová DNA, umožňují detekovat změnu elektrického proudy podle tvaru procházejících nukleotidů. Jde tedy o první metodu, která nepoužívá polymeraci vlákna DNA (obrázek 1.5b). Detekce ovšem neprobíhá na úrovni jednoho nukleotidu, ale vždy je snímán signál po „slovech“ dlouhých 5 bazí. Společnost vyvíjí rovněž software s názvem Metrichor pro převod signálu do zápisu sekvence jednotlivých nukleotidů. První MinION přístroje byly dostupné omezenému okruhu uživatelů v rámci programu společnosti Oxford Nanopores spuštěnému v roce 2014. Komerčně dostupné jsou od roku 2015. Výhodou přístroje MinION jsou velmi malé rozměry, velikostí jde vlastně o kapesní zařízení, které se připojuje k počítači prostřednictvím USB. Přístroj je navíc schopen číst až desítky tisíc bazí dlouhé sekvence, ovšem chybovost se pohybuje až kolem 30 % (Laver et al., 2015). Je to dáno především metodou detekce po slovech, nikoli jednotlivých nukleotidech. Vylepšováním počítačových programů pro převod do nukleotidové sekvence a modelováním charakteru vznikajících chyb a tvorba algoritmů pro jejich korekci může postupně vézt ke snižování chybovosti produkovaných dat (Jain et al., 2015). Některé pětice nukleotidů lze však odlišit snadněji a některé hůře. Jedinou možností, jak dosáhnout kvalitních dlouhých čtení se zatím zdá být kombinace dat s jinou sekvenační platformou, hlavně od společnosti Illumina (Madoui et al., 2015).

1.3 Kontrola kvality čtení, odstranění chybných sekvencí, trimování, dereplikace

Efektivita enzymů a detekčních metod může během sekvenace kolísat a stanovení konkrétního nukleotidu na konkrétní pozici je ve skutečnosti vyjádřením pravděpodobnosti, kdy je posuzována například intenzita signálu a zároveň šum snímaný na pozadí. Každá výstupní sekvence nukleotidů je tak doprovázena ještě souborem vyčíslených pravděpodobností výskytu jednotlivých bazí.

Samotné další analýzy tak vždy předchází krok kontroly kvality čtení a případně vyřazení chybných sekvencí nebo ořezání chybných úseků (tzv. trimování, *trimming*) převážně na koncích čtení. Asi nejpoužívanějším samostatným programem je FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Ačkoli to většina bioinformatiků neocení, poskytuje také grafické prostředí, které zde uveďme pro lepší představu o funkcích a výstupech programu (obrázek 1.6). Dalšími programy jsou SolexaQA (Cox et al., 2010), Fastx-Toolkit (http://hannonlab.cshl.edu/fastx_ toolkit/index.html), PRINTSEQ (Schmieder & Edwards, 2011), NGS QC Toolkit (Patel & Jain, 2012) a další software určený ke kontrole čtení jen z některých platforem (např. Trimmomatic pro Illumina sekvenátory – Bolger et al., 2014). Důležité je rovněž trimování případných adaptorových sekvencí, vektorových sekvencí či dalších značek uzívaných například pro rozlišení vzorků v jednom běhu sekvenátoru. Pro tento úkol slouží řada programů, vyhledávání zmíněných sekvencí zvládá také již uvedený software FastQC.



Obrázek 1.6: Přehled výstupu programu FastQC pro hodnocení kvality sekvenačních dat (zdroj: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)

Podle charakteru výchozího sekvenovaného vzorku můžeme použít specializované programy na kontrolu kvality čtení. Například pro 454 pyrosekvenování PCR produktů (amplikonů) byl vyvinut algoritmus PyroNoise pro hodnocení chyb na základě porovnávání schémat průběhu intenzity jednotlivých bazí v cyklech sekvenace (flowgramů) (Quince et al., 2011). Tento postup nazývaný jako *denoising* byl dále implementován například v programech Mothur (Schloss et al., 2009) či QIIME (Caporaso et al., 2010). Pro analýzu kvality čtení z RNA sekvenování (tzv. RNA-seq), tedy sekvenování cDNA (komplementární, *complementary* DNA) zpětně přepsané enzymem reverzní transkriptázou z původního vzorku RNA, existuje specializovaný software SEECER (*SEquencing Error CorrEction in Rna-seq data*) (Le et al., 2013).

Při zpracování 454 dat velmi dobře funguje softwarový balík vyvíjený společně s GS instrumentem firmou Roche například pro sestavení sekvencí (program Newbler), mapování (program Mapper) a další aplikace. Ten má v sobě vnitřně zahrnuté postupy hodnocení kontroly sekvencí a umožňuje i filtrování adaptorových či jiných kontaminujících sekvencí na základě uživatelem poskytnuté reference. Řada kroků však nedovoluje nastavení parametrů uživatelem a nemusí poskytovat nejlepší výsledky. Trimování a korekci chyb mají implementované také další programy na tvorbu sestavení sekvencí (např. SOAPdenovo – Luo et al., 2012) a některé dokonce vyžadují hrubá, neprocesovaná čtení (ALLPATHS – Gnerre et al., 2011).

Přestože metody PCR jsou nezbytným základem pro sekvenování nové generace, při průběhu amplifikace může docházet k inkorporaci nesprávných nukleotidů. PCR je také zdrojem jedné specifické chyby v datech, výskytu tzv. duplikátů. Duplikovaná čtení vznikají namnožením jedné shodné templátové molekuly ve více mikroreaktorech či shlucích na destičkách. Ačkoli se vžil pojem duplikáty, nejde vždy jen o zdvojení počtu, ale časté jsou i několikanásobně přečtené shodné sekvence; mohli bychom proto spíše používat označení multiplikáty. Nelze jednoznačně říci, kolik procent unikátních a kolik multiplikovaných sekvencí lze v datech očekávat, a jejich výskyt se liší u každé připravené sekvenační knihovny a u každého sekvenačního běhu. Jako příklad uveďme studii dokládající výskyt multiplikátů v rozmezí 11 % až 35 % u typického metagenomického shotgunového souboru dat (Gomez-Alvarez et al., 2009). Vznik multiplikátů není důsledkem jen emulzní či můstkové amplifikace předcházející samotnou sekvenaci, ale v principu každé PCR amplifikace, která byla použita pro přípravu vzorku či přípravu sekvenační knihovny. Obecně lze říci, že čím je menší vstupní množství DNA a čím je vyšší počet případných amplifikačních kroků během jeho zpracování, tím můžeme očekávat větší procento multiplikátů v datech (Parkinson et al., 2012).

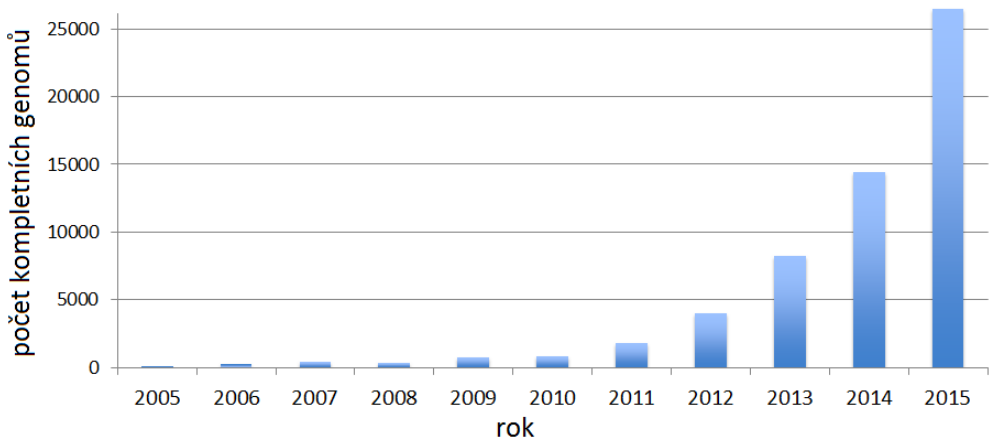
Zmnožená čtení téhož templátu mohou negativně ovlivňovat procesy sestavení sekvencí a hlavně analýzy rozdílně zastoupených genů v metagenomických datech. Naprosto identická čtení by bylo možné snadno identifikovat, multiplikáty se však mohou lišit délkou a také sekvenčně díky vzniku chyb. Ne každá individuální sekvenační reakce dosáhne stejného počtu cyklů se stejnou kvalitou jednotlivých bazí. Detailní analýzu výskytu multiplikovaných čtení nabízí studie Niu et al. (2010), ve které autoři rovněž nabízejí definici pro postup jejich identifikace v datech: multiplikáty jsou buď identická čtení nebo čtení začínající na stejné pozici, jejichž délka se může lišit, avšak kratší sekvence jsou plně zarovnatelné s delšími a vykazují navzájem předem definovanou míru podobnosti. Řada celých softwarových balíků (například již zmíněná sada nástrojů od firmy Roche spjatá se 454 sekvenátory) či online nástrojů pro automatické vyhodnocení sekvenačních dat zahrnuje krok dereplikace, tedy filtrování multiplikátů. Stejný krok mají implementovaný i některé programy na kontrolu kvality NGS dat (například opět FastQC) a lze najít mnoho programů věnovaných cíleně dereplikaci (zde zmiňme například podmnožiny balíku CD-HIT věnované zvlášť 454 a Illumina datům: http://weizhongli-lab.org/cd-hit/). Zajímavé je využití identifikovaných duplikovaných sekvencí zpětně pro odvozování výskytu chybně přečtených bazí. Za tímto účelem byla vyvinuta metoda DRISEE (*detecting errors in metagenomic sequencing data*, Keegan et al.,

2012), která může sloužit k identifikaci problematických souborů dat, u kterých sekvenace neproběhla optimálně.

1.4 Bakteriální genomika

V loňském roce oslavilo celogenomové sekvenování bakteriálních genomů výročí 20 let. Zajímavostí je, že ačkoli první sekvenovanou bakterií byl *Haemophilus influenzae* (Fleischmann et al., 1995), nešlo o patogenní bakterii, jak by se mohlo zdát, ale o linii H. influenzae Rd. dostupnou díky její kultivaci v laboratoři pro izolaci restrikčního enzymu HindIII (Loman & Pallen, 2015). Samotným biologickým vlastnostem kódovaným nově odhalenou genomovou sekvencí byla také věnována poměrně malá pozornost. Zásadní význam však měla nová metoda shotgunového sekvenování s nalámáním genomové DNA na náhodné fragmenty, klonováním a čtením jejich konců prostřednicvím Sangerovy metody (Fleischmann et al., 1995). Během následující dekády umožnila přečtení genomů modelových bakteriálních organismů jako *E. coli* K-12 (Blattner et al., 1997), ale také patogenních druhů jako *Mycobacterium tuberculosis*, což umožnilo odhadovat podstatu patogenních vlastností (Cole et al., 1998). Komparativní genomika bakteriálních linií odhalila roli horizontálního transferu například u *E. coli*, či strukturní vlastnosti některých prokaryotických genomů jako tandemové repetice a inverze (souhrnně Loman & Pallen, 2015). Pozornost byla upřena také na genomy mikroorganismů žijících v lokalitách zatížených lidskou aktivitou a schopných biodegradace polutantů (souhrnně např. Pieper et al., 2004).

Kromě náročné přípravy a sekvenace každého klonovaného fragmentu zvlášť, objevil se později také problém v datech s chybějícími sekvencemi, které nesou geny toxické pro hostitelské buňky E. coli (Kimelman et al., 2012). Metody sekvenování nové generace tyto limitace obchází přesunem od biologického systému namnožení molekul k chemickému přístupu amplifikace sekvencí metodou PCR a masivním automatickým paralelním čtením. Po roce 2005 můžeme díky jejich aplikaci sledovat rapidní vzrůstání počtu kompletních prokaryotických genomů v databázi GenBank (obrázek 1.7). Výhodou Sangerovy sekvenace zůstávala délka spojitého řetězce, který je schopna v jedné reakci přečíst. U Sangera je maximální délka čtení přibližně 1.000 bp, zatímco první genomový sekvenátor GS20 dokázal číst sekvence o délce v průměru 100 bp. I když postupnou snahou dochází k prodlužování čtení, často bohužel za cenu narůstající chybovosti, i tak jsou výchozími daty pro další analýzy pouze krátké fragmenty většinou v rozsahu 150 – 800 bp.



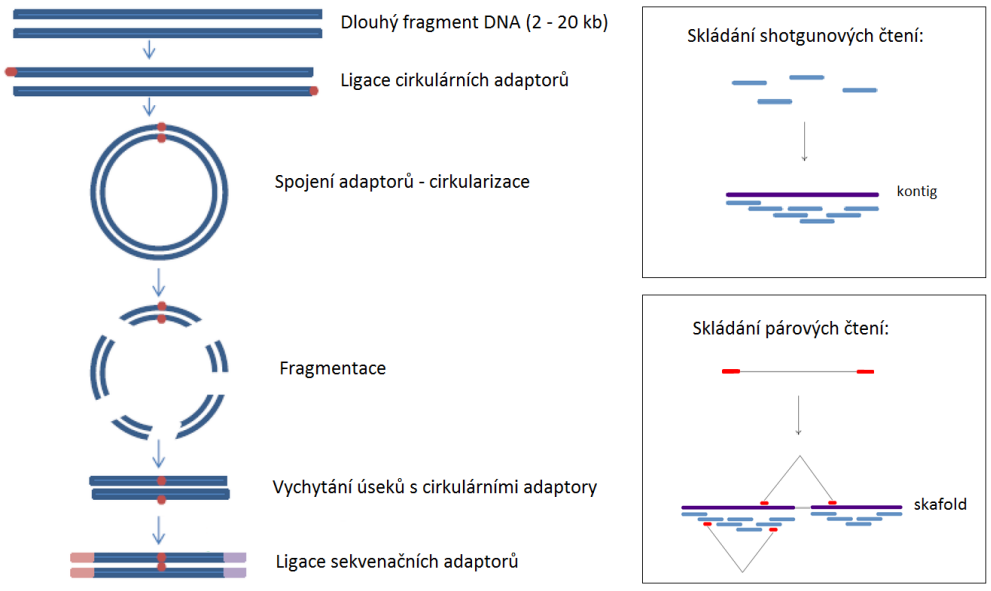
Obrázek 1.7: Počet kompletních sekvenovaných genomů v databázi GenBank do roku 2015 (data z http://ftp.ncbi.nlm.nih.gov/genomes/GENOME_REPORTS/prokaryotes.txt)

1.4.1 Sestavení sekvencí

Velké množství krátkých sekvencí produkovaných genomovými sekvenátory vyústilo v potřebu tvorby sofistikovaných počítačových algoritmů pro jejich sestavení a rekonstrukci původních dlouhých molekul genomů. Proces se v angličtině nazývá *assembly* a popisuje nejen samotný postup, ale i balík výsledných sestavených dat. Počítačový program vykonávající automatické sestavení je pak nazýván jako *assembler*. Vzhledem k tomu, že český pojem „sestavení“ není v tomto ohledu plně dostačující, uchylím se občas v následujícím textu k původním anglickým termínům v zájmu zachování odpovídajícího významu. Nezbytným předpokladem sestavení je, aby každá jednotlivá báze byla pokryta nejlépe desítkami přečtených sekvencí, tedy dosažení dostatečné míry pokrytí (míra pokrytí se vyjadřuje anglickým termínem *coverage*). V případě snahy o přečtení nového neznámého genomu mluvíme o *de-novo* sekvenování. Jednodušší situace nastává, pokud máme k dispozici již stanovenou, referenční sekvenci ať již stejného nebo blízce příbuzného druhu. V takovém případě může být našim úkolem například nalezení polymorfních míst u sledovaných mikrobiálních linií a referenční sekvence slouží jako předloha, na níž můžeme získaná krátká čtení takzvaně mapovat.

Výrazným přispěním k tomu, jak řešit problém s rekonstrukcí původních dlouhým molekul DNA na základě získaných krátkých čtení, je zavedení metod párového sekvenování. Ty následují příkladu původního shotgunového přístupu, kdy klonované fragmenty dlouhé tisíce bazí byly čteny z obou směrů, což poskytovalo jakousi fyzickou mapu a vodítko pro sestavení. V principu jde o postup, kdy se rovněž provádí štěpení výchozích molekul DNA, ale

na fragmenty s předem definovanou délkou výrazně převyšující možnosti jednotlivých forem sekvenace, zpravidla volíme délku 2 kb až 20 kb (kilo bazí). Konce desítky tisíc nukleotidů dlouhých fragmentů jsou nejprve opatřeny speciálními adaptory, které jsou následně spojeny, čímž vznikají cirkulární molekuly. Ty jsou opět naštěpeny, ovšem dále jsou vychytány a sekvenovány jen úseky obsahující cirkulární adaptory (obrázek 1.8). Výsledkem jsou sekvence několika desítek či stovek nukleotidů v závislosti na použitém instrumentu, o kterých víme, že ohraničují úsek, jehož pořadí bazí sice neznáme, ale známe jeho přibližnou délku. V případě platformy od Roche jsou nazývány jako paired-end čtení, společnost Illumina používá v této souvislosti název mate-pairs. Taková data pak obsahují informaci o směrové a prostorové orientaci sekvencí. Zásadní význam hrají v *de-novo* sekvenování pro seřazení již rekonstruovaných spojitých úseků (kontigů, *contigs*) do větších, směrově orientovaných celků (skafoldů, *scaffolds*) (obrázek 1.8). Tak vzniká draftová sekvence, která pokrývá celé výchozí studované molekuly DNA. Obsahuje sice mezery, ale jsou dobře definovány. Párová čtení jsou prozatím jedinou možností, jak řešit rekonstrukci genomů s opakujícími se, tzv. repetitivními úseky. Kromě toho jsou taková data výborným nástrojem pro analýzu strukturních přestaveb během evoluce genomů (viz např. Dempsey et al., 2006).



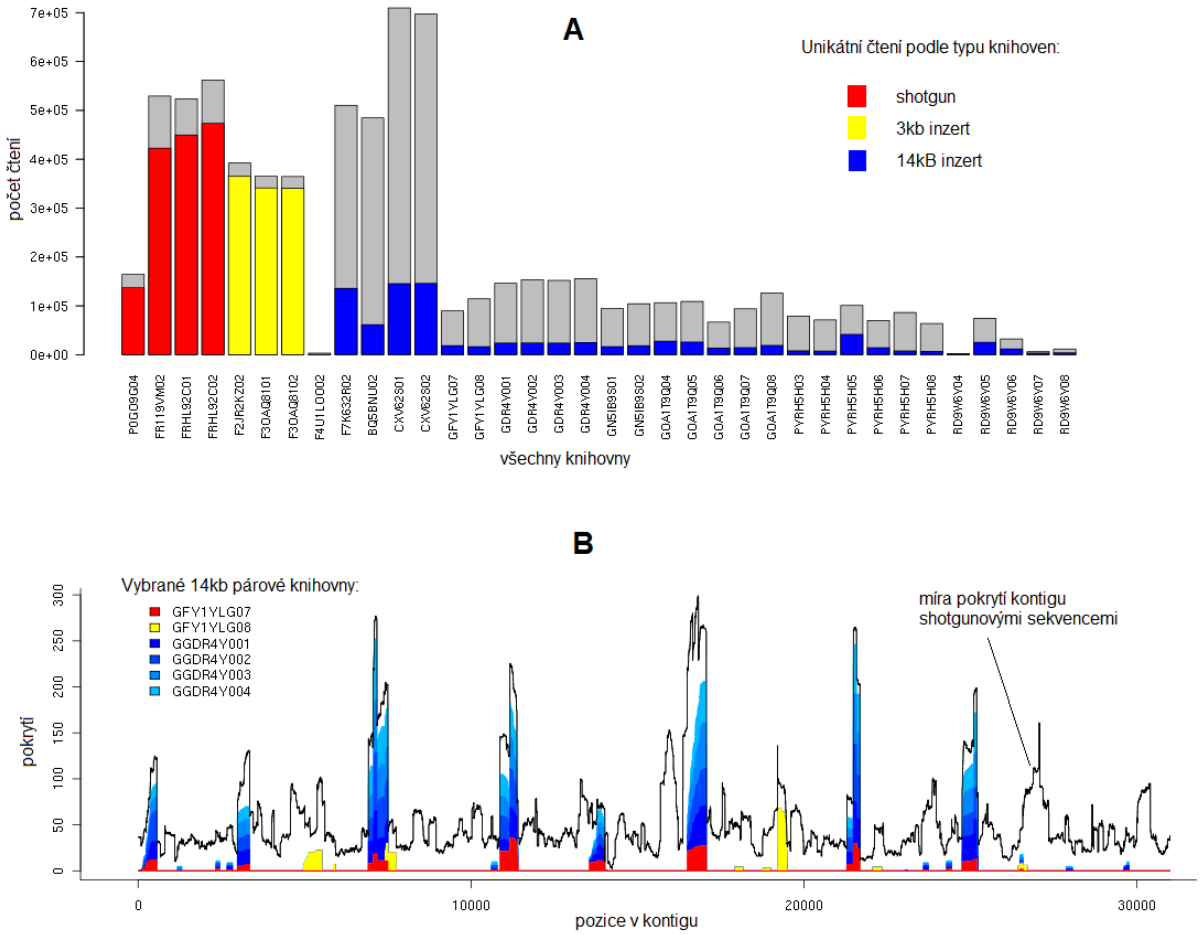
Obrázek 1.8: Postup přípravy vzorku pro párové sekvenování (převzato a upraveno z Gao & Smith, 2015) a znázornění sestavení sekvencí.

Ideálním přístupem pro genomové sekvenování s využitím NGS platforem je kombinovat shotgunové knihovny s párovými knihovnami s různou délkou sekvence inzertu.

Takový přístup jsme úspěšně aplikovali při analýze genomu bakterie *Achromobacter xylosoxidans* A8, která bude popsána v kapitole 3. S neustálým vývojem sekvenátorů a snižování ceny samotných sekvenačních reakcí za jednu přečtenou bazi však dochází k situaci, kdy příprava knihoven převyšuje cenu sekvenace a stává se tak limitujícím faktorem. Dnes publikovaná data tak často vycházejí jen z jedné shotgunové NGS knihovny (Magoc et al., 2013) a snad 90 % bakteriálních genomů v databázi GenBank není kompletních (Land et al., 2015).

1.4.2 Multiplikovaná čtení a význam dereplikace

Počítačové programy využívající algoritmy pro sestavení genomových sekvencí většinou pracují s průměrnou mírou pokrytí nově vznikajících kontigů, což je důvod, proč výsledek assembly může být negativně ovlivněn vysokým podílem multiplikovaných čtení, jak jsme již zmínili v kapitole 1.5. To se obzvláště týká párových knihoven, kdy během jejich přípravy nutně dochází k redukci použitelného materiálu (používáme jen konce dlouhých úseků, navíc dochází ke ztrátám při manipulaci s DNA, jelikož postup zahrnuje více kroků), což vyžaduje mezikrok PCR amplifikace. Z našich zkušeností s pyrosekvenováním párových knihoven vyplývá, že nárůst multiplikátů můžeme sledovat převážně u knihoven s dlouhými inzerty, kde úbytek koncentrace DNA během jejich přípravy je výraznější. Podíl unikátních a multiplikovaných čtení můžeme ukázat na výsledku assembly námi sekvenovaného genomu prvoka *Mastigamoeba balamuthi* (obrázek 1.9, nepublikovaná data). Multiplikovaná čtení nejen zastiňují původní četnosti fragmentů DNA ve vzorku, ale opakovaná sekvenace jedné sekvenační knihovny už díky jejich výskytu nemusí přinášet nové použitelné sekvence (obrázek 1.9a ukazuje na konkrétním příkladu jednoho kontigu z genomu *M. balamuthi* pokrytí jednotlivými typy knihoven, 1.9b dokládá výskyt stejných sekvencí v opakovaně sekvenované knihovně s dlouhým inzertem pokrývající vždy shodné úseky kontigu).



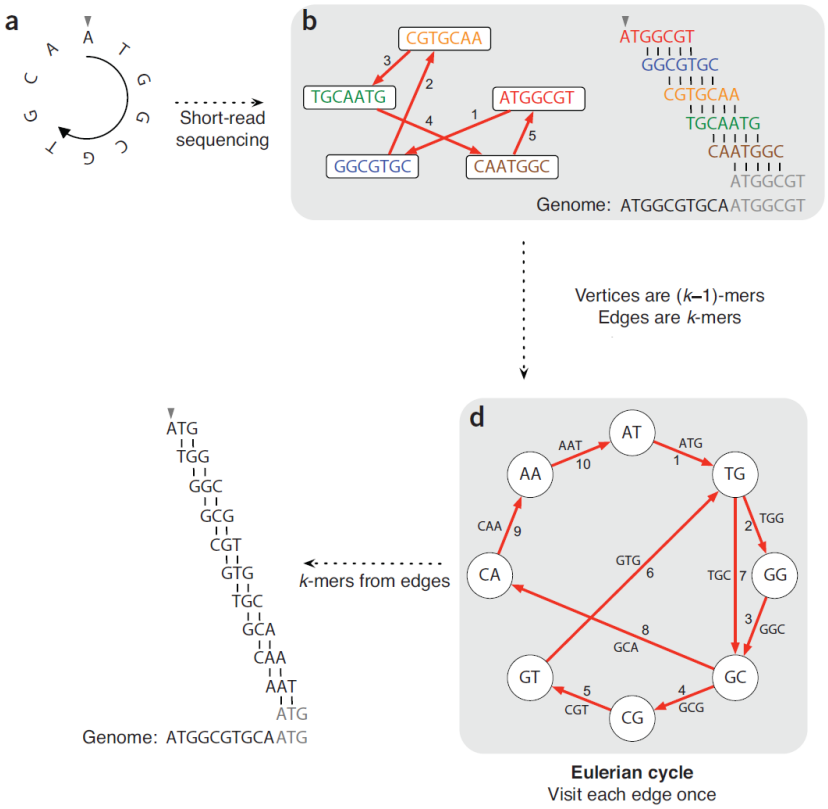
Obrázek 1.9: Výskyt multiplikovaných čtení ve 454 datech: A) podíl unikátních čtení podle různého typu knihoven (barevná výseč sloupců) v celkovém počtu čtení (znázorněném šedivě), B) barevně znázorněné vybrané párové knihovny s inzertem 14 kb pokrývající opakovaně stejné úseky složeného kontigu. Míru pokrytí shotgunovými sekvencemi znázorňuje černá linie.

1.4.3 *De-novo* sestavení

Programů pro sestavení sekvencí je celá řada (viz např. Magoc et al., 2013; Ekblom & Wolf, 2014). Jejich výstupy se liší a také vznikaly za účelem různých aplikací. Některé byly vyvinuté cíleně pro assembly malých, prokaryotických genomů, jako třeba Velvet (Zerbino & Birney, 2008); jiné naopak pro rekonstrukci velkých genomů mnohobuněčných organismů, jako SOAPdenovo (Luo et al., 2012) nebo ALLPATHS (Gnerre et al., 2011), které většinou lze použít rovněž na bakteriální genomy (Magoc et al., 2013). Programy jako Celera assembler (Denisov et al., 2008) či Arachne (Batzoglou et al., 2002) byly původně navrženy pro zpracování delších čtení Sangerovy sekvenace, obdobně jako Newbler (Roche), zaměřený původně na 454 data, používají tradiční přístup tvorby assembly na základě přiřazení (alignmentu) čtení. Zjednodušeně řečeno postup začíná (1) párovým přiřazením mezi všemi

sekvencemi, (2) vyhledání dvojic s nejdelším překryvem, následuje (3) jejich spojení a opětovná tvorba alignmentu od kroku jedna. Můžeme říci, že program pracuje se skutečnými původními sekvencemi. Newbler takto velmi dobře zvládá sestavení 454 čtení, novější verze umožňují také kombinované assembly s daty z jiných platforem, ovšem nastíněný algoritmus je výpočetně náročný a existuje hranice počtu množství čtení (v závislosti na možnostech daného výpočetního klastru), od které již není schopen sestavení dovést do konce (naše vlastní zkušenosti).

Většina programů zaměřených na velké množství krátkých čtení, jaké produkují například sekvenátory od Illuminy, využívá algoritmy z oblasti teorie grafů, tzv. de Bruijnovy grafy (Compeau et al., 2011; Nagarajan & Pop, 2013). Namísto alignmentu celých sekvencí jsou původní čtení rozdělena do slov (*k-mer*) předem definované délky, ty jsou podle shody prefixů (*k-mer* minus poslední baze) a sufixů (*k-mer* bez první baze) pospojovány do grafu, po kterém se pak algoritmus pohybuje posouváním slov o jednu pozici s navštívením každého spoje jedenkrát (obrázek 1.10). Do této kategorie programů spadá například SOAPdenovo (Luo et al., 2012), ALLPATHS (Gnerre et al., 2011), ABySS (Simpson et al. 2009) a Velvet (Zerbino & Birney, 2008). „Hybridní“ způsob sestavení používají programy Atlas (Havlak et al., 2004), Ray (Boisvert et al., 2010) či MaSuRCA (Zimin et al., 2013). Porovnáním assemblerů pro rekonstrukci bakteriálních genomů na základě krátkých čtení se věnuje studie Magoc et al. (2013), přičemž dobrých výsledků dosahovala právě poslední jmenovaná MaSuRCA.

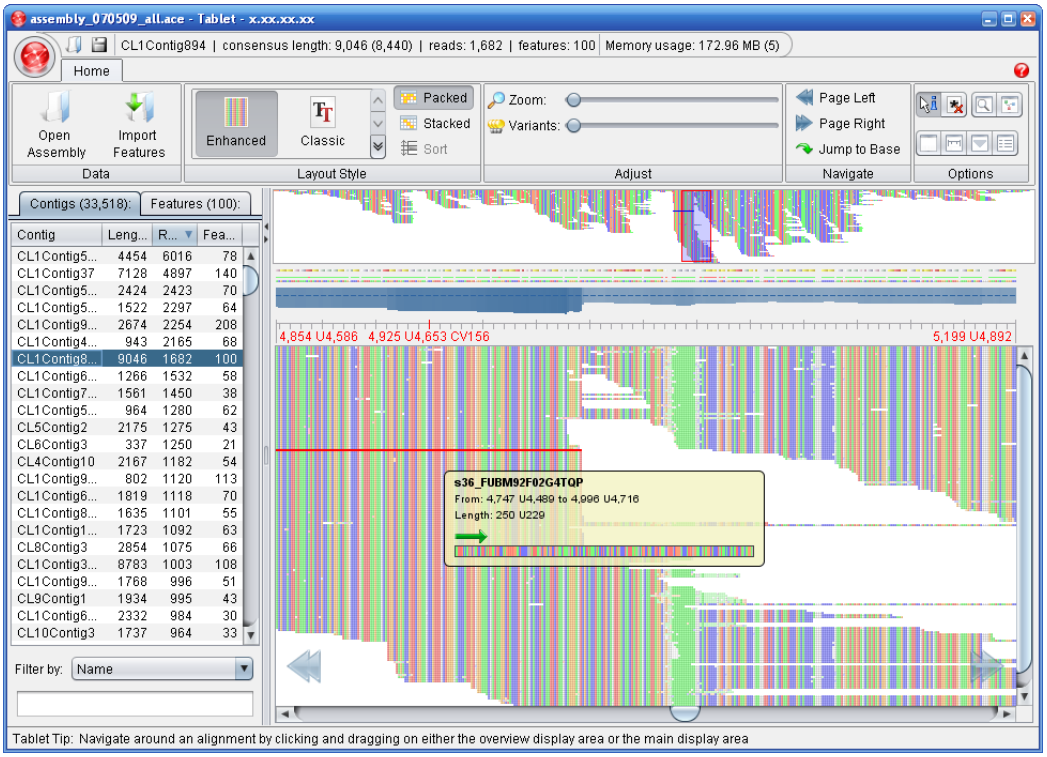


Obrázek 1.10: Principy algoritmů pro sestavení čtení. A) Příklad krátké cirkulární molekuly DNA, která je fragmentována jako při shotgunovém sekvenování. B) Postup sestavení na základě alignmentu překrývajících se úseků, kdy jednotlivá čtení představují uzly grafu a jejich spoje jsou alignmenty dvojice čtení, *genome* reprezentuje sestavenou sekvenci. B) Princip sestavení na základě de Bruijnova grafu, kdy jsou v našem příkladu čtení rozdělena do slov (*k-mer*) o délce 3 nukleotidy. Uzly grafu (*vertices*) jsou reprezentovány všemi prefixy a sufixy, spoj (*edge*) představuje slovo obsahující prefixy a sufixy, které spojuje. Např. slovo ATG spojuje uzly AT a TG. Assembly je konstruováno jako nalezení tzv. euleriánského cyklu procházením spojů grafu, kdy je dosaženo alignmentu sufixu předchozího a prefixu následujícího slova (převzato z Compeau et al., 2011).

Jelikož sekvenační platformy poskytovaly od počátku trochu jiný charakter dat ve smyslu poměru délka versus celková produkce počtu bází, zdálo se nasnadě použití více dostupných sekvenčních dat pro lepší rekonstrukci souvislejších úseků či celých genomů. Logická představa, kterou bylo možné sledovat napříč řadou příspěvků na genomických symposiích, zahrnovala postup, kdy v prvním kroku jsou vygenerovány delší sekvence poskytující sice nižší míru pokrytí, ale představující po složení jakousi kostru, kterou by bylo možné následně doplnit o kratší čtení jiné sekvenační platformy s vyšší mírou pokrytí jednotlivých bází. Prvním assemblerem schopným zpracovat různá data byla MIRA, původně vyvinutá pro sestavení cDNA dat ze 454 sekvenování a umožňující nejprve kombinování sekvencí produkovaných Sangerovou metodou a pyrosekvenováním, později také sekvenátory

firmy Illumina. Dnes většina programů umožňuje zadat data z více než jedné sekvenační platfromy, ze zkušeností však vyplývá, že použití širší škály dat sice může nabízet větší objem vstupních sekvencí a vyšší předpokládanou coverage, nicméně nevede nutně k lepším výsledkům sestavení ve smyslu snížení počtu kontigů či skafoldů, v případě bakteriálního genomu v ideálním případě jedné výsledné cirkulární sekvence. Různá data se zkrátka chovají rozdílným způsobem vlivem rozdílného charakteru sekvenačních chyb.

Alternativní možností je nikoli kombinované sestavení, ale doplnění jednoho assembly jinými sestavenými sekvencemi a nebo dalšími sekvenačními daty. První způsob umožňuje například software minimus2 (součást balíku AMOS, Sommer et al., 2007) nebo Graph Accordance Assembly (GAA) program (Yao, et al., 2012). Druhý postup, například zaplnění mezer v assembly vyprodukovaném programem Newbler s využitím krátkých *paired-end* dat ze sekvenátoru od společnosti Illumina, využívá program GapCloser vyvinutý v rámci SOAPdenovo (Lua et al., 2012). Mapování dlouhých kotigů či skafoldů nebo celých genomů na sebe pro kontrolu assembly nebo porovnání nově získané genomové sekvence s již stanovenou referencí je možné provézt programem MUMmer (respektive podprogramem nucmer – Kurtz et al., 2004). Existují rovněž programy na vizualizaci složených sekvencí, například Consed (Gordon & Green, 2013) nebo Tablet (obrázek 1.11) (Milne et al., 2012).



Obrázek 1.11: Grafické znázornění složených sekvencí v programu Tablet (Milne et al., 2012) poskytující přehled čtení pokrývající různá místa kontigů a míru pokrytí (převzato z https://ics.hutton.ac.uk/tablet)

1.4.4 Predikce kódujících oblastí a anotace

Identifikace kódujících sekvencí (*coding sequences*, CDS) v nově sestavené genomové sekvenci využívá dva přístupy. První je založený na vyhledání struktury v sekvencích odpovídající výskytu genu. Hrubou predikci CDS představuje již vyhledání dostatečně dlouhých otevřených čtecích rámců, tedy úseků nepřerušených stop kodony. Stop kodony by se v náhodné sekvenci vyskytovaly přibližně každých 60 až 75 bp (3 ze 64 možných kodonů jsou stop kodony). Dalším indikátorem je také výskyt promotorových sekvencí, jako je Pribnowův box či vazebná místa pro transkripční faktory, nebo Shine-Dalgarnovy sekvence rozpoznávané na úrovni mRNA ribosomy při translaci. Další metodou je také statistické hodnocení využívaných synonymních kodonů, či ještě citlivější metoda hodnocení dvou po sobě následujících kodonů (dikodonů). Programy, které dokáží automaticky predikovat CDS na základě strukturních informací, jsou například GeneMark (Borodovsky & McIninch, 1993) Glimmer (Delcher et al., 1999) a Prodigal (Hyatt et al., 2010).

Druhý způsob využívá vyhledání podobnosti s databázemi známých genů pro zjištění pozice CDS. Často používaným programem, který kombinuje informace z obou přístupů, je Critica (Badger & Olsen, 1999). Nejlepší přístup pro predikci CDS spočívá ve využití více programů zároveň a sloučení jejich výstupů. Pro kombinování výstupů jsou mocným nástrojem databázové systémy jako například MySql (https://www.mysql.com/), ve kterých lze správně volenými dotazy identifikovat výsledky, ve kterých se jednotlivé programy rozcházejí, a kterým je dále potřeba věnovat pozornost.

Kromě programů pro predikce genů, existuje také software na vyhledávání transferové RNA (tRNA) a transferové-mediátorové RNA (tmRNA) – například tRNAscan (Lowe & Eddy, 1997) a Aragorn (Lalett & Canback, 2004). Pro identifikaci ribosomální RNA (rRNA) lze využít program RNAmmer (Lagesen et al., 2007).

Funční anotace spočívá ve vyhledání podobností mezi predikovanými CDS a geny ve veřejně dostupných databázích. Hlavní databází je GenBank (Benson et al., 2010), provázaná s databázemi EMBL (Kulikova et al., 2004) a DDBJ (Miyazaky et al., 2004), které se průběžně navzájem zálohují. Hlavními zdroji GenBank jsou neredundantní databáze NCBI-NR obsahující aminokyselinové sekvence a NCBI-NT obsahující nukleotidové sekvence. Neredundantní znamená, že záznamy se shodnou sekvencí byly sloučeny do jednoho. GenBank je dále strukturována a poskytuje mnoho dalších databází, jako například databázi celých

genomů, nebo zdroje řazené podle charakteru vstupních dat. Například sekvence získané z environmentálních vzorků (ENV) či celogenomovým sekvenováním (WGS) a další (Benson et al., 2010). GenBank je také propojená s databází SRA (*Sequence Read Archive*), která slouží jako uložiště hrubých sekvenačních dat, jejichž zpřístupnění je často vyžadováno pro přijetí publikace do tisku. GenBank (NCBI-NR a NCBI-NT) slouží jako primární uložiště umožňující deponovat data uživatelům. Díky tomu obsahuje obrovské množství neustále se množících záznamů, ovšem z podstaty je databází nemoderovanou. To znamená, že uživateli uložená data neprocházejí kontrolou, mohou být nepřesně popsaná nebo obsahovat chyby, které se navíc mohou dále množit při využití chybných záznamů pro anotaci dat dalších uživatelů. GenBank se proto rovněž věnuje tvorbě moderované databáze RefSeq (Tatusova et al., 2014).

Obdobou GenBank zaměřenou na proteinové sekvence je databáze UniProt, která vznikla sloučením databází Swiss-Prot, TrEMBL a PIR (The UniProt Consortium, 2015). Mimo jiné nabízí také neredundantní databázi UniRef (Suzek et al., 2007).

Kromě přiřazení funkce odděleně jednotlivým analyzovaným genům existují také snahy o zařarení genů do funkčních systémů, metabolických drah či rodin. Za tímto účelem vznikl například projekt *Gene Ontology*, který si klade za cíl přiřadit genům označení (*GO term*), které by umožňovalo zařasení genu do funkčního procesu. Databáze KEGG (*Kyoto Encyclopedia of Genes and Genomes*) si klade za cíl umístit geny do metabolických a regulačních drah (Kanehisa et al., 2008). Dalším přístupem je identifikace ortologních genů, které se vyvinuli evolucí ze stejného předka. Databázi ortologních genů představuje COG (*Clusters of Orthologous Groups*) (Tatusov et al., 2000; Tatusov et al., 2003) nebo EggNOG (Huerta-Cepas et al., 2016).

Pro vyhlédávání v databázích se nejčastějí používá program BLAST (Camacho et al., 2009). Pro vyhledání konzervovaných domén, pro něž reference nabízejí databáze jako je PFAM (Bateman et al., 2002) či TIGRFAM (Haft), lze použít jeho variantu PSI-BLAST (Altschul et al., 1997). Často se také pro vyhledávání domén využívá tzv. Hiden Markových modelů (HMM), rozšířený je například program HMMER (hmmer.org).

Celý proces anotace řeší také řada dostupných automatických online serverů. Jmenujme například server RAST (Aziz et al., 2008), který využívá vlastní databázi SEED umožňující seskupování anotovaných genů do funkčních subsystémů, nebo server IGS Annotation Engine (http://manatee.sourceforge.net).

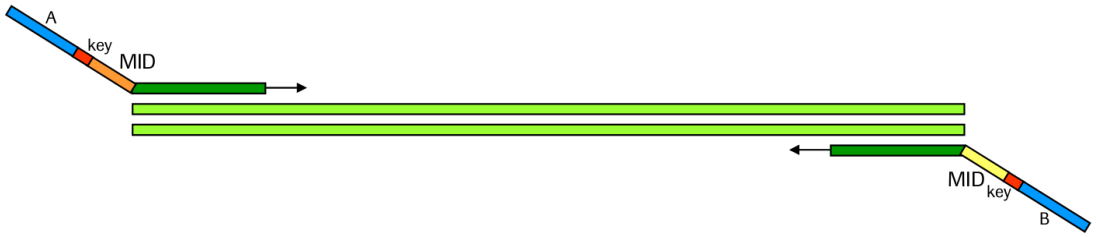
1.5 Studium mikrobiálních společenstev

Celogenomové projekty sice umožňují analýzu kompletní genetické výbavy jednoho organismu, díky čemuž dokážeme popsat jeho funkční procesy a roli v ekosystému, jsou ale z drtivé většiny omezeny pouze na kultivovatelné modelové organismy. Kultivovat v laboratoři však dokážeme jen méně než 1% mikroorganismů (Riesenfeld et al., 2004; Lozupone & Knight, 2008; Zhang & Xu, 2008). Funkčních procesů či metabolických drah se často účastní řada mikroorganismů v ekosystému, z nichž každý ve svém genomu nese jen některé geny kódující proteiny uplatňující se pouze na některé z úrovní systému, nikoli celé funkční dráhy. Izolací jednotlivých mikroorganismů tak postihujeme jen nepatrný zlomek z celkové diverzity a funkčního potenciálu mikroflóry.

V roce 1990 publikoval Giovannoni společně s kolegy první výzkum na úrovni mikrobiální komunity s využitím analýzy genu pro malou podjednotku prokaryotické ribosomální RNA (16S rRNA) jakožto markeru pro taxonomickou klasifikaci členů konsorcia pikoplanktonu Sargasového moře (Giovannoni et al., 1990). Byť tento prvotní report používá metodu hybridizace environmentální DNA s referenční směsí genomové DNA a referenčními sekvencemi 16S rRNA, autoři zároveň navrhují následný postup sekvenování knihovny namnožených úseků 16S rRNA prostřednictvím rekombinantního klonování. Metoda klonování, ať už cíleně některých úseků DNA (16S rRNA genů) nebo s využitím shotgunového přístupu při tvorbě sekvenačních knihoven, sehrála významnou roli v rozvoji výzkumů mikrobiálních konsorcií. Kromě otázky, jaké mikroorganismy jsou v daném prostředí přítomné, vyvstává také dotaz, jak fungují a co dělají. Handelsman s kolegy navrhl v roce 1998 pojem „metagenomika“ pro studium souhrnu teoreticky všech přítomných genomů mikroorganismů, neboli metagenomu mikrobiální komunity, prostřednictvím izolace a analýzy celkové DNA ze vzorků bez nutnosti předchozí kultivace (Handelsman et al., 1998). Zavedení metod masivně paralelního sekvenování umožnilo posun od klonovacích technik k amplifikaci a sekvenaci vybraných genů či přímé tvorbě sekvenačních knihoven rovnou z totální DNA izolované z prostředí. Můžeme dnes tedy rozlišit dva hlavní přístupy ke studiu mikrobiálních komunit: první využívá amplikonové sekvenování, druhý (v pravém slova smyslu metagenomický přístup) sekvenování shotgunové.

1.5.1 Amplikonové sekvenování

Pojem amplikonové sekvenování popisuje postup, při kterém je z celkové DNA nejprve metodou PCR amplifikován pouze úsek vybraného genu prostřednictvím specifických primerů a následně sekvenován. Standardní příprava shotgunové sekvenační knihovny pro sekvenátory druhé generace zahrnuje fragmentaci DNA a ligaci adaptorových oligonukleotidů. Při tvorbě amplikonové sekvenační knihovny lze tento krok vynechat a navrhnout primery pro PCR, které kromě samotné sekvence komplementární k cílenému genu nesou rovnou adaptorové sekvence podle zvolené sekvenační platformy. Za adaptorové sekvence lze rovněž zařadit krátké úseky nukleotidů, které pak obsahují všechna příslušná čtení na svém začátku a slouží jako značka (*tag*, nebo též MID z anglického *multiplex identifier*) pro odlišení jednotlivých vzorků sekvenovaných dohromady v jednom běhu sekvenátoru (obrázek 1.12).



Obrázek 1.12: Design primerů pro 454 amplikonové sekvenování. Světle zelená je znázorněna cílená templátová sekvence, tmavě zelené jsou úseky primerů komplementární k jejím ohraničujícím úsekům. Červená představuje klíčovou sekvenci (*key*), kterou rozpoznává instrument GS FLX. MID je identifikátor pro odlišení vzorků sekvenovaných v jedné reakci, modře jsou vyznačeny adaptorové sekvence.

Tyto postupy velmi zefektivňují amplikonové sekvenování v porovnání se shotgunovým přítupem. Výhodou také je, že amplikonová čtení mohou do značné hloubky pokrývat stejný úsek homologních genů a mohou postihnout i geny méně četných členů konsorcia, lze je zarovnat, vytvořit alignment, hodnotit sekvenční rozdíly a provádět fylogenetické analýzy.

Ovšem právě z povahy amplikonových čtení – mají stejný začátek a mohou si být velmi podobná – nelze ve výsledných datech rozlišit chybné PCR duplikáty (či multiplikáty). Jejich výskyt se v jednotlivých sekvenačních reakcích velmi liší (viz kapitola 1.3.1) a možný vliv na pozdější analýzy četností genů ve vzorcích nelze stanovit ani korigovat. Navíc zvolená sada primerů může efektivněji nasedat na genomové sekvence některých organismů a tím je preferovat, čímž vzniká posun v datech a to i s využitím degenerovaných primerů postihujících širokou škálu variant příslušné templátové DNA. Známou chybou při PCR je také vznik tzv. chimerních sekvencí, kdy polymeráza přeskočí během eloganční fáze z jedné templátové

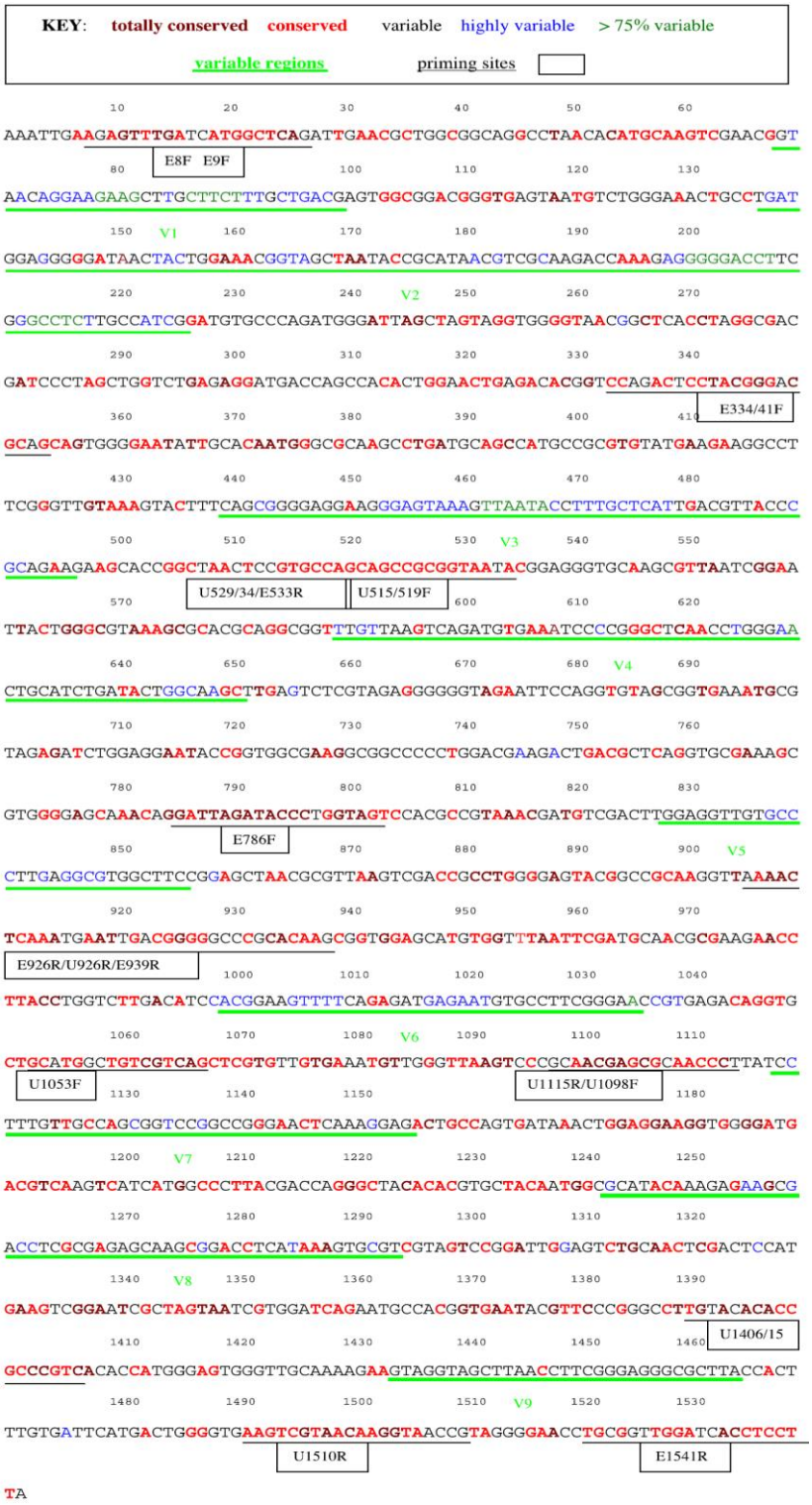
molekuly na jinou a výsledkem je sekvence ze dvou templátů rozdílného původu (Schloss et al., 2011). Za účelem identifikace možných chimérních čtení a jejich odstranění z NGS dat vznikly programy UCHIME (Edgar et al., 2011), Perseus (Quince et al., 2011), ChimeraSlayer (Haas et al., 2011), Dicipher (Wright et al., 2012), Bellerophon (Huber et al., 2004) a B2C2 (Gontcharova et al., 2010).

Amplikonové sekvenování pokrývá jeden či jen několik vybraných genů (markerů), nikoli celé genomy organismů. Na úrovni studia mikrobiálních společenstev bychom tedy mohli mluvit o metagenetice. Jde však často ruku v ruce s metagenomickými výzkumy, jelikož sdílí jednotku výzkumu (celá mikrobiální konsorcia), výchozí materiál (metagenomovou DNA vzorku) a metodiku masivního sekvenování a často tak bývá řazeno přímo pod metagenomiku.

1.5.1.1 16S rRNA gen

Gen pro malou podjednotku ribosomální rRNA (16S rRNA) je nejdůležitější marker pro studium prokaryotické diverzity používaný již přes 40 let. Je velmi konzervovaný, zastoupený napříč prokaryotickými genomy a obsahuje domény vykazující variabilní substituční rychlosti (obrázek 1.13), což umožňuje design primerů do konzervovaných úseků pro studium variabilních regionů. Přestože se při taxonomické klasifikaci neuvažuje vliv horizontálního transferu, přenos mezi bakteriálními liniemi je možný a byl popsán například u streptokoků (Schouls et al., 2003) či pseudomonád (Bodilis et al., 2012).

Díky dlouhodobému a extenzivnímu využívání 16S rRNA markeru existují dnes obsáhlé databáze ribosomální DNA umožňující porovnání a klasifikaci širokého spektra mikroorganismů. Například databáze GreenGenes (DeSantis et al., 2006), Ribosomal Database Project (RDP) (Cole et al., 2014) či SILVA (Quast et al., 2013), která kromě prokaryotické 16S rRNA nabízí také databázi eukaryotických podjednotek ribosomální DNA umožňující klasifikaci členů taxonů houby (fungi) a mnohobuněční (metazoa). Rovněž jsou postupně rozpracovávány a standardizovány postupy (pipeline) pro procesování dat z různých sekvenačních platforem zahrnující počítačové algoritmy pro řešení jednotlivých kroků pipeline a celé softwarové balíky, které umožňují kompletní analýzu od hrubých dat až po výslednou klasifikaci sekvencí a hodnocení diverzity vzorků. Jejich příkladem jsou programy Mothur (Schloss et al., 2009), QIIME (Caporaso et al., 2010) a Seed (Vetsrovský & Baldrian, 2013). Vlastní online nástroje pro analýzu amplikonů nabízí také již zmíněný Ribosomal Database Project a samostatnou pipeline pro 16S rRNA sekvence obsahuje server věnovaný metagenomickým analýzám MG-RAST (Meyer et al., 2008).



Obrázek 1.13: Modelový příklad 16S rRNA *Escherichia coli* znázorňující konzervované a variabilní regiony a místa pro PCR primery – E pro eubakterie, U jako univerzální (převzato z Baker et al., 2003)

Standardní postup zahrnuje tyto hlavní body: (1) kontrolu kvality čtení a trimming adaptorů, MID sekvencí a sekvencí primerů; (2) denoising v případě 454 dat; (3) vyřazení chimerních sekvencí; (4) klastrování sekvencí na základě podobnosti a jejich sloučení do

operačních taxonomických jednotek (OTU); (5) anotace reprezentativních sekvencí jednotlivých OTU vyhledáním podobností v některé ze zmíněných databází. K prohledávání lze využít programy BLAST či FASTA, ovšem záznam s nejvyšší mírou podobnosti nemusí vždy zaručovat správné taxonomické zařazení, zvláště pokud pracujeme s krátkými sekvencemi či sekvencemi z organismu, pro který v databázi neexistují blízce příbuzné druhy. Hierarchický přístup pro klasifikaci sekvencí na základě hodnocení několika nejpodobnějších záznamů nalezených programem BLAST a spolehlivé přiřazení sekvencí do vyšších taxonomických kategorií (na úrovni rodu, třídy či až kmenu) umožňuje software MEGAN (Huson & Weber, 2013). Často se také používá Wangova metoda Naivního Bayesovského algoritmu, která hodnotí podobnost procházením sekvencí po úsecích dlouhých 8 nukleotidů a poskytuje bootstrapový odhad spolehlivosti. Vyvinutá byla původně pro RDP Classifier (Wang et al., 2007), ale je implementovaná například také v softwaru Mothur.

Klasický přístup systematiky bakteriálních druhů zahrnuje fenotypové i genotypové znaky. Molekulární klasifikace využívala standardizované DNA-DNA hybridizační experimenty, kdy jsou bakterie přiřazeny jednomu druhu, pokud vzájemná reasociace jejich genomové DNA nabývá hodnot větší než 70 % při rozdílu denaturačních teplot maximálně 5 °C (Gevers et al., 2005; Achtman & Wagner, 2008). Tento postup lze použít pouze pro kultivovatelné druhy, jejichž čistou genomovou DNA jsme schopni izolovat, hybridizační experimenty se navíc těžko standardizují mezi laboratořemi a nelze klasifikovat jednotlivé sekvence porovnáním s databázemi. Při analýze 16S rRNA genu jsou čtení klastrována do OTU nejčastěji s podobností nejméně 97 % (Gevers et al., 2005), což přibližně odpovídá 70% DNA-DNA podobnosti, tedy hranici pro přiřazení jednomu druhu. Ovšem korelace mezi podobností 16S rRNA sekvencí a DNA-DNA hybridizací se může lišit pro různé bakteriální linie a různé druhy mohou vykazovat až 99% podobnost 16S rRNA genů (například *Bacillus globisporus* a *B. psychrophilus*, Fox et al., 1992). Díky tomu je 97% hranice podobnosti pouze hrubý odhad. Ve většině případů dokážeme klasifikovat bakterie pouze na úrovni rodu, přičemž postrádáme nutné rozlišení pro nižší úrovně (Gevers et al., 2005) a místo označení druh používáme termín OTU pro klastr podobných 16S rRNA sekvencí.

Kromě taxonomického rozlišení představuje problém také rozdílný počet kopií 16S rRNA genů u jednotlivých organismů. Například Větrovský & Baldrian (2013) popisují výskyt 1 až 15 kopií na genom u souboru 1690 bakterií. Kopie v rámci jednoho genomu se navíc mohou vzájemně lišit až o 5%, příkladem je genom *E. Coli* K12 (Eren et al., 2013). Výborný přehled zmíněných nedostatků studia 16S rRNA a analýzu používaných metod představuje

článek Nguyen et al. (2016). Alternativním postupem je analýza několika genů, které se univerzálně vyskytují napříč genomy mikroorganismů pokud možno v jedné kopii na genom, jsou konzervované, ale rychlost jejich evoluce poskytuje větší rozlišení i na úrovni příbuznějších linií. Například přístup původně rozpracovaný v rámci epidemiologie pro odlišení patogenních mikrobů i na nižší než druhové úrovni hodnotí sadu několika housekeepingových genů (Hanage et al., 2006).

Na základě taxonomické klasifikace existuje snaha odvozovat také funkční potenciál mikrobiálních společenstev podle anotace referenčních genomů odpovídajících jednotlivým 16S rRNA sekvencím (De Filippo et al., 2012). Za tímto účelem vznikl software PICRUSt pro odhad metabolických procesů konsorcií (Langille et al., 2013). Postup lze však prakticky použít jen pro komunity s výskytem bakterií, pro něž existují v databázích referenční genomy. Odvozování funkčních informací na základě taxonomie je velmi hrubým odhadem také díky možné ztrátě genů v evoluci jednotlivých mikrobiálních linií nebo naopak jejich získání horizontálním transferem, což zvláště platí například pro geny na plasmidech, které mohou kódovat řadu důležitých funkčních proteinů (jako příklad může sloužit genom bakterie *Achromobacter xylosoxidans* popisovaný v kapitole 3).

1.5.1.2 Funkční geny

Další možností studia mikrobiálních společenstev a jejich genové výbavy je přímá PCR amplifikace a sekvenace funkčních, proteiny kódujících genů (*coding sequences*, CDS). Metody zatím nejsou tak rozpracované jako pro amplikonové sekvenování 16S rRNA. Důležitým tématem jsou chyby v sekvenčních datech vedoucí k posunu čtecího rámce (frameshift), převážně při využití 454 pyrosekvenování. Zatímco frameshift výrazně neovlivní zařazení do OTU a klasifikaci 16S rRNA sekvencí, zásadní význam má posun čtecího rámce na překládání nukleotidové sekvence do sekvence proteinové. Proto existuje snaha korigovat chybné posuny v amplikonových datech zaměřených na funkční geny. Počítačovým nástrojem k tomu určeným je například software FrameBot (Wang et al., 2013), který jsme na základě našich výzkumů byli schopni dále modifikovat (Strejček et al., 2014). Pro zpracování a anotaci funkčních amplikonových dat je výborným nástrojem vyvíjená pipeline a databáze FunGene (Fish et al., 2013), která poskytuje zarovnané sekvence již početné řady kódujících genů, umožňuje alignment s využitím Hidden Markov modelů (HMM) a možnost stažení modelu vytvořeného na základě souboru referenčních sekvencí vybraných genů pro potřeby vyhledávání v uživatelských souborech dat.

1.5.2 Metagenomika

Metagenomika využívá přímé shotgunové sekvenování celkové DNA ze vzorku prostředí. Přestože i zde se používá označení celogenomový shotgun, (*whole genome shotgun*, WGS) pro definici způsobu tvorby sekvenační knihovny například při vkládání dat do databází, mohli bychom spíše hovořit v kontextu metagenomiky o celometagenomovém (*whole metagenome*) sekvenování. Prvotní výzkumy využívající náročné klonování a Sangerovo sekvenování fragmentů DNA byly dostupné velkým sekvenačním centrům. Příkladem je obsáhlý, bezprecedentní metagenomický projekt Graiga Ventera a jeho kolegů na sekvenování mikroorganismů Sargasového moře publikovaný ještě rok před uvedením první NGS platformy (Venter et al., 2004). Prudký vzestup metagenomiky po roce 2005 vyústil ve velké množství studií od výzkumů extremofilních organismů z ekosystémů ovlivněných vulkanickou činností, extrémní teplotou či hodnotou pH (Xie et al., 2011; Pearce et al., 2012; Hua et al., 2015) přes analýzy běžných půdních organismů vykazující největší biodiverzitu na Zemi (Xu et al., 2014) až po studium mikroorganismů asociovaných s povrchem těla či trávicím traktem lidí a dalších vyšších organismů majících značný vliv na jejich fenotypové vlastnosti (Turnbaugh et al., 2009).

Rozmach metagenomiky je umožněn nejen dostupností NGS dat, ale zároveň rozvojem bioinformatických metod zaměřených na metagenomická data. Vznikly dokonce serverové systémy, které umožňují online zpracování od hrubých sekvencí přes anotaci a vizualizaci zastoupení organismů, funkčních genů a metabolických drah až po vzájemné porovnávání vzorků: IMG/M (Malkowitz et al., 2014), METAREP (Goll et al., 2010), CAMERA (Seshadri et al., 2007), RTMg (Edwards et al., 2012) a MG-RAST (Meyer et al., 2008). Ačkoli se může použití online nástrojů jevit jako prosté zjednodušení, kromě využití výpočetní kapacity veřejných serverů, což činí metagenomické studie přístupné i pro skupiny, které nedisponují výpočetními klastry, existují i další důvody jejich využití. Například MG-RAST (metagenomická odnož systému RAST) je všeobecně velmi používaný, což zajišťuje kompatibilitu a možné porovnání dat mezi řadou studií. Pro vyhledávání podobností se známými geny spravuje vlastí neredundantní databázi M5nr, která je souborem většiny hlavních zdrojů: GenBank, SEED, IMG, UniProt, KEGG a eggNOGs (Wilke et al., 2012). Velká obliba systému zároveň vede tým, který ho vyvíjí, k neustálému testování a případné implementaci dostupných metod, takže tato pipeline (viz MG-RAST manual) představuje jakýsi standard.

1.5.2.1 Analýza krátkých čtení versus assembly metagenomických sekvencí

Bioinformatické metody zpracování metagenomických NGS dat jsou úzce spjaté původně se 454 pyrosekvenováním. Hlavním přístupem se stala anotace přímo krátkých čtení bez sestavení, označovaných také jako environmentální genové značky (*environmental gene tags*, EGTs), umožňující analýzu celých mikrobiálních komunit a porovnání různé míry abundance genů mezi vzorky na základě hodnocení počtu klasifikovaných čtení. Hlavními body postupu jsou opět: (1) kontrola kvality a trimování; (2) naprosto nezbytný je zde krok dereplikace a (3) anotace čtení. K anotaci může sloužit vyhledání podobnosti v databázích známých genů pro každou sekvenci zvlášť, například programem BLASTX. Takový postup může být efektivní, pokud máme k dispozici výkonné výpočetní klastry a chceme hodnotit pouze několik souborů dat. Výrazného zrychlení však dosáhneme, pokud vstupní sekvence podrobíme nejprve algoritmům na predikci CDS, předpovězené geny dále seřadíme do klastrů a k prohledávání databází použijeme jen reprezentativní sekvence. Pro predikci genů v krátkých sekvencích DNA (v rozmezí 75-1000bp) jsou vyvíjeny programy MetaGene (Noguchi et al., 2006), Glimmer-MG (Kelley et al., 2012), FragGeneScan (Rho et al., 2010), MetaGeneAnnotator (Noguchi et al., 2008), MetaGeneMark (Zhu et al., 2010), Orphelia (Hoff et al., 2009). Testování většiny z nich se věnuje publikace Trimble et al. (2012). Některé byly implementovány také zmíněnými online anotačními službami. Postup využívající FragGeneScan s následnou tvorbou klastrů přeložených aminokyselinových sekvencí na hranici 90% identity vedl podle vývojářského týmu systému MG-RAST k 750násobnému urychlení pipeline (MG-RAST manual).

Alternativní postup spočívá ve snaze složit z velkého množství dat dlouhé kontigy, či v některých případech celé genomy (Brown et al., 2015), které pak umožňují provádět přesnější predikce genů, přispívají k lepší anotaci a slouží k analýze kontextu sekvencí, například celých operonů. Vzhledem k postupné převaze technologie firmy Illumina dosahující maximální délky čtení jen 300 bp – a to ještě v případě méně výkonného MiSeq přístroje, přičemž velké studie v současnosti vycházejí ze 150bp párových čtení z instrumentu HiSeq (např. Brown et al., 2015) – se sestavení dat v metagenomických projektech stává nezbytností. Ačkoli informace o pokrytí jednotlivých kontigů může sloužit i pro odhad abundance mikrobiálních linií a funkčních genů mezi jednotlivými metagenomy, assembly většinou neslouží ke studiu celé komunity, ale k získání částí genomů nekultivovatelných organismů či kompletní sekvence několika genů pro jejich detailnější analýzu.

Oproti sestavení genomů organismů, jejichž kulturu můžeme pěstovat v laboratoři, čelí sestavení metagenomických čtení několika zásadním problémům. Metagenomové sekvence, byť i jednoho mikrobiálního organismu, nepochází z klonální linie, takže i homologní úseky se mohou lišit, ať již sekvenčně či strukturně. Naopak blízké linie mohou sdílet například repetitivní úseky a také vlivem horizontálního přenosu sdílí někdy stejné geny i vzdálené druhy. V důsledku různého zastoupení organismů se pak velmi liší i pokrytí jejich genomů v metagenomickém vzorku. Standardní genomové assemblery obvykle vycházejí z předpokladu klonální linie a více či méně jednotného sekvenačního pokrytí napříč sestavovanými kontigy. Pokud nás zajímají jen konkrétní genomy, pro které existují v databázích referenční sekvence, můžeme přistoupit k tvorbě asistovaného assembly, kdy programu poskytneme referenční genomovou sekvenci jako vodítko. V tomto případě nejde vlastně o assembly v pravém slova smyslu, spíše o mapování. Použít lze programy Newbler (Roche), MIRA (Chevreux et al., 1999) či MetAMOS (Treangen et al., 2013), jejich výstup lze vizuálně hodnotit v programech jako je Artemis (Carver et al., 2008). Pokud chceme získat složené sekvence dokonce jen jednotlivých genů, můžeme použít program Xander (Wang et al., 2015), který jako vodítko využívá HMM model vytvořený na základě rodiny genů, která nás zajímá. Průběh asistovaných assembly není příliš výpočetně náročný, ale samozřejmě vede k velké redukci metagenomické informace.

Pro *de-novo* assembly nepříliš komplexních komunit fungují i standardní genomové assemblery jako Newbler nebo MIRA. Často však ekosystémy vykazují značnou míru diverzity mikrobiálních populací a pro postihnutí i méně zastoupených druhů je potřeba vygenerovat velké množství krátkých čtení s využitím těch nejvýkonnějších sekvenátorů. Jejich sestavení je mnohem výpočetně náročnější, většinou využívá algoritmy založené opět na de Bruijnových grafech. Nové programy se navíc snaží cíleně řešit charakter metagenomických dat plynoucí z neklonáních kultur a rozdílného pokrytí. Příklady jsou MetaVelvet a MetaVelvet-SL (Namiki et al., 2012; Afiahayati et al., 2015), Meta-IDBA a IDBA-UD (Peng et al., 2011; Peng et al., 2012) či Ray Meta (Boisvert et al., 2012).

Jednotlivá čtení i sestavené sekvence můžeme roztřídit do taxonomické skupiny, v ideálním případě přiřadit genomům jednotlivých organismů. Tato snaha se nazývá *binning* (obrazně roztřídit do košů z anglického *bin*) (souhrnně např. Sharpton, 2014; Escobar-Zepeda et al., 2015; Oulas et al., 2015). U jednoduchých komunit s několika převažujícími liniemi můžeme spočítat obsah GC párů, stanovit pokrytí u složených kontigů, vyhledat homologní geny podle podobnosti v databázích a tato vodítka sloučit dohromady při manuálním postupu

třídění. Pro složitější komunity existují nástroje na automatizaci třídění sekvencí založené na (1) jejich kompozici nebo (2) alignmentu proti referenčnímu záznamu v databázích. První postup hodnotí výskyt sdílených „slov“ na základě porovnání krátkých úseků sekvencí (zvolená délka slova se označuje jako *k-mer*), které představuje jakousi genomovou signaturu. Ten používají například programy TETRA (Teeling et al., 2004), PhyloPhytia (McHardy et al., 2007; Patil et al., 2012), MetaCluster-TA (Wang et al., 2014), S-GSOM (Chan et al., 2008), TACAO (Diaz et al., 2009), PCAHIER (Zheng & Wu, 2010) a ClaMS (Pati et al., 2011). Další nástroje dokáží kombinovat genomovou signaturu s hodnocením více znaků, jako je zmíněný GC obsah, pokrytí kontigů či výskyt genových markerů, například housekeepingových genů (např. MaxBin – Wu et al., 2016 a AMPHORA2 – Wu & Scott, 2012). Některé ze zmíněných programů vyžadují nejprve trénování postupu na sadě dostupných referenčních sekvencí, což sice vede ke zpřesnění hodnocení kompozice sekvencí, ale také ke sklonu lépe seskupovat a také preferovat linie příbuzné k již známým, referenčním organismům.

Druhý přístup využívá vyhledání homologních sekvencí v databázích například s využitím mapovacích programů jako BWA (Li & Durbin, 2009) či Bowtie2 (Langmead & Salzberg, 2012), přístupů založených na vyhledávání s využitím HMM modelů či s použitím programu BLAST. Programy, které dokáží automatizovat tento postup, jsou třeba CARMA (Krause et al., 2008), MetaPhyler (Liu et al., 2011), SOrt-ITEMS (Monzoorul et al., 63). Hybridní postup s využitím obou výše popsaných přístupů následují programy PhymmBL (Brady & Salzber, 2011) a MetaCluster (Wang et al., 2012).

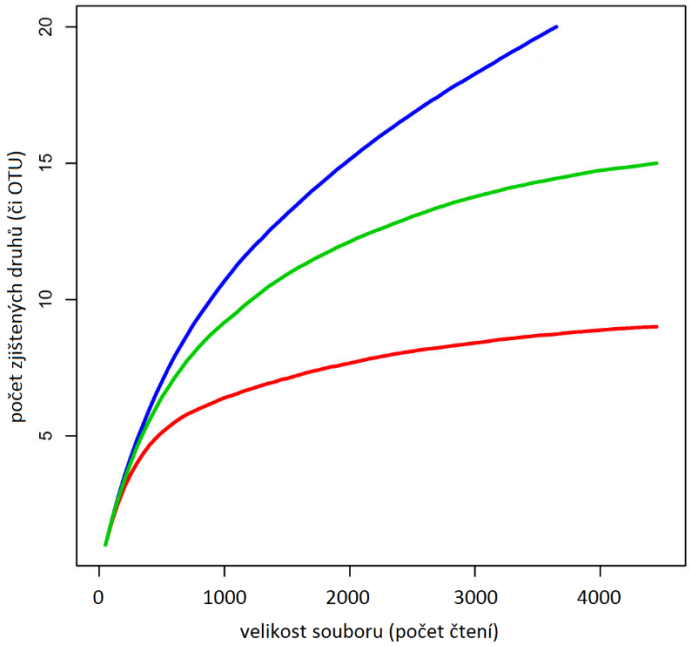
Oba přístupy přirozeně opět čelí problému s horizontálním transferem (Sharpton, 2014). Časté je například kolísání výskytu GC párů v rámci jednotlivých genomů (viz např. genomový projekt *Achromobacter xyloosoxidans* v kapitole 3), značící úseky cizího původu, které mohou být i velmi staré a vlivem evoluce genomu zamaskované a roztříštěné. Jednotlivé organismy preferují jen některé z alternativních kodonů kódujících stejnou aminokyselinu a tato preference může také sloužit jako vodítko pro přiřazení sekvencí k jednotlivému genomu. Ovšem zajímavé je v tomto ohledu zjištění, že používání kodonů může být v rámci bakteriální komunity shodně vyladěné i mezi různými organismy – metagenom se tak chová obdobně jako jeden velký genom (Roller et al., 2013).

1.5.2.2 Taxonomický a funkční profil

Při studiu celých mikrobiálních komunit nás zajímají dvě hlavní otázky: (1) „kdo“ je v daných populacích přítomen a (2) „co dělá“. Odpověď na první nabízí taxonomická

klasifikace čtení a jejím výstupem je taxonomický profil četností identifikovaných organismů. Existují dva přístupy, jak odhadnout zastoupení linií v shotgunových metagenomických datech. První je zaměřen pouze na vyhledávání čtení, která postihují fylogenetické markery, nejčastěji opět 16S rRNA geny. Hodnocení čtení je pak podobné analýze 16S rRNA amplikonů. Ovšem ribosomální geny představují obvykle méně než 0,1% bakteriálních genomů, v metagenomických datech jsou velmi málo zastoupené a na rozdíl od amplikonů pokrývají navíc různé úseky 16S rRNA. Tvorba taxonomického profilu proto většinou vychází z anotace genů kódujících proteiny, pro které v databázích existují nejen záznamy o funkci, ale také informace, z jakého organismu gen pochází. Informace o funkci ze stejné anotace je pak použita pro konstrukci rovněž funkčního profilu abundance CDS, který odpovídá na druhou otázku – jaké procesy je schopné konsorcium vykonávat. Vždy je dobré mít k dispozici různá data pro silnější podporu výsledků a je například užitečné kombinovat shotgunové sekvenování s analýzou amplikonů. Analýzu 16S rRNA PCR produktů lze použít pro získání přehledu o zastoupení organismů před dalšími rozsáhlejšími analýzami nebo pro ověření shotgunového taxonomického profilu. Tyto přístupy jsme zvolili například pro výzkum jednoduché komunity důlního ekosystému (kapitola 4) nebo studium půdních bakterií asociovaných s kořeny rostlin (kapitola 5).

Na základě taxonomických profilů lze počítat odhady diverzity uvnitř metagenomického vzorku, tedy tzv. alfa diverzitu. Je důležité zmínit, že ta je dána dvěma složkami: přirozeně počtem přítomných druhů (či OTU) (*richness*), ale také rozdílnou mírou zastoupení druhů (*evenness*) (Simpson, 1949). Rozdílnou diverzitu tak budou mít komunity se stejným počtem druhů, kdy u jedné z nich budou rovnoměrně zastoupeny, ale u druhé se budou některé vyskytovat ve větším počtu a jiné vzácně. Nejjednodušším výpočetním odhadem, který postihuje obě složky, je Simpsonův index (*D*), který udává pravděpodobnost výběru dvou stejných druhů z daného vzorku (Simpson, 1949). Míru entropie narůstající s počtem druhů ve vzorku pak odhaduje Shannonův index (*H'*) (Shannon, 1948). Představu o diverzitě poskytují také rarefakční křivky (*rarefaction curves*), které zjednodušeně řečeno v tomto kontextu sledují trend přibývajících nově odhalených druhů (případně OTU) v závislosti na vzrůstajícím počtu použitých anotovaných sekvencí (obrázek 1.14). Ty lze rovněž použít pro odhad, do jaké míry jsme „prosekvenovali“ daný vzorek, tedy nakolik postihuje náš soubor sekvencí zastoupení druhů v komunitě, a pro odhad ideální velikosti souboru sekvencí pro postižení většiny (ideálně všech, i vzácných) druhů.

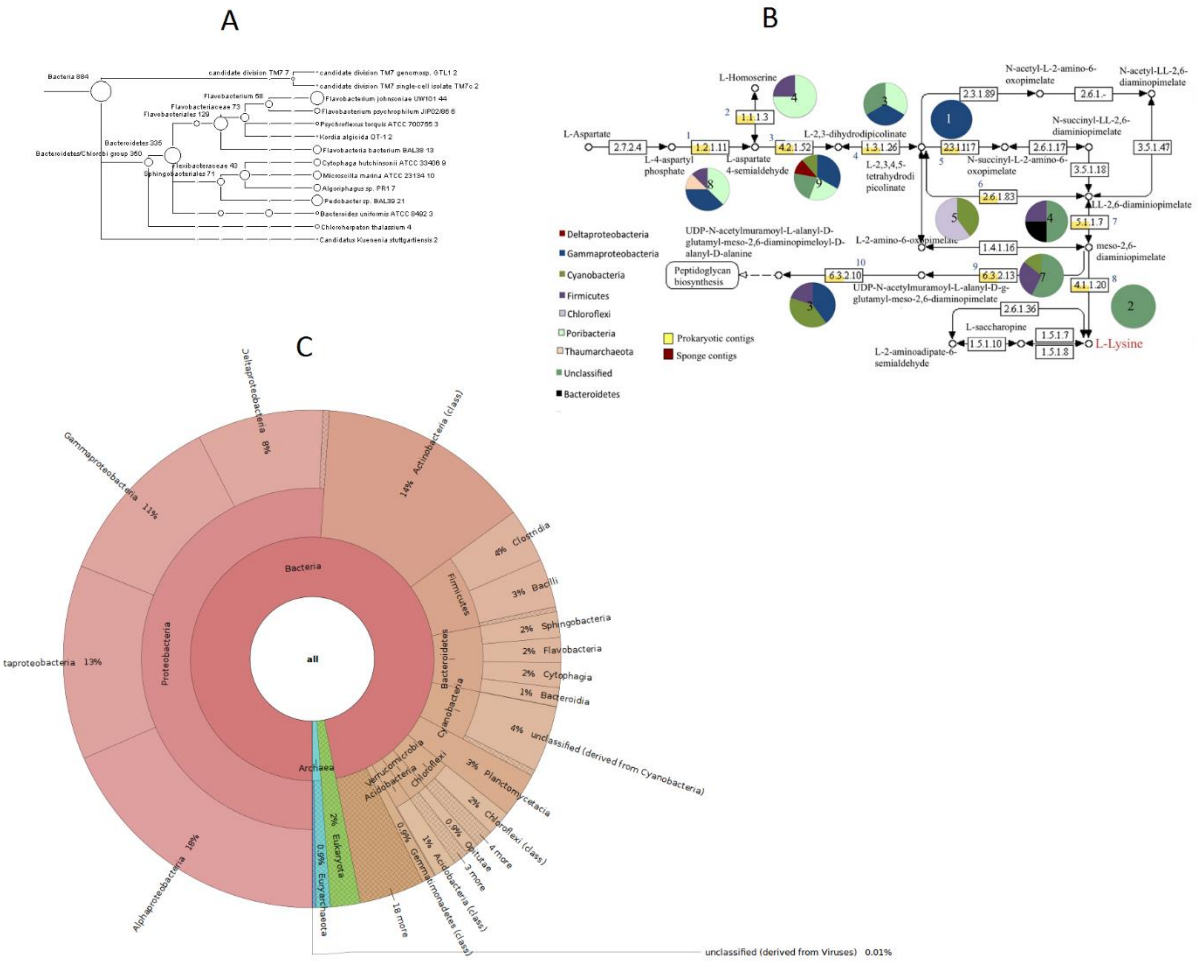


Obrázek 1.14: Příklad rarefakčních křivek znázorňujících vztahy mezi velikostí souboru (počtem použitých čtení) a počtem zjištěných taxonů na základě použitých čtení. Modrá křivka znázorňuje vzorek s velkým počtem různých taxonů. Červená křivka odpovídá vzorku s malým pořtem taxonů – v tomto případě zároveň můžeme říci, že finální velikost souboru čtení (v našem příkladě něco přes 4000 sekvencí) poskytuje dobrou míru „prosekvenování“, kdy získání dalších sekvencí vede k odhalení jen malého množství nových taxonů.

Taxonomické a funkční profily lze vizualizovat ve formě dendrogramů postihující výskyt taxonomických kategorií a funkčních hierarchicky uspořádaných systémů. To umožňuje software MEGAN, původně určený pouze pro taxonomické dendrogramy (obrázek 1.15a), ale novější verze pracují i s funkčními profily u umožňují vizualizovat metabolické dráhy na základě KEGG databáze (obrázek 1.15b). Vizualizaci skladby metagenomického vzorku v podobě interaktivního grafu nabízí nástroj Krona (obrázek 1.15c) (Ondov et al., 2011).

1.5.2.3 Komparativní metagenomika

Taxonomické a funkční profily jsou základem pro komparativní metagenomiku více různých komunit, tedy hodnocení beta diverzity. Opět existuje několik indexů pro vyčíslení beta diverzity (přehledně Tuomisto, 2010), na jejich základě lze sestavit matici vzdáleností jednotlivých souborů, například s využitím indexu Bray-Curtisovy nepodobnosti (Chao et al., 2006) a následně vizualizovat případnou strukturu v datech, kdy některé vzorky mohou být bližší a tvořit klastry prostřednictvím mnohorozměrného škálování (MDS) či analýzy hlavních komponent (PCA).



Obrázek 1.15: Způsoby vizualizace metagenomických dat. A) Dendogram vytvořený v programu Megan, B) vizualizace metabolických drah na základě databáze KEGG a zvýraznění organismů, které jsou schopny jednotlivé stupně dráhy vykonávat (převzato z Fiore et al., 2015) a B) interaktivní koláčový graf pro vizualizaci a procházení zastoupení taxonů vytvořený programem Krona.

Můžeme rovněž hodnotit rozdílnou míru zastoupení jednotlivých taxonomických či funkčních kategorií a identifikovat ty, které se statisticky odlišují mezi vzorky. Za tímto účelem byl například vyvinut software STAMP (Parks et al., 2014), který poskytuje jednak mnohorozměrné metody, tak identifikaci a grafickou vizualizaci statisticky rozdílně četných kategorií, a balíčky do statistického softwaru R (R Development Core Team, 2009) jako phyloseq (McMurdie & Holmes, 2013) a metagenomeSeq (Paulson et al., 2013). Prvním krokem při hodnocení rozdílné míry abundance mezi vzorky je normalizace dat při rozdílném počtu čtení v souborech dat. Tradičním postupem je tzv. *rarefying*, tedy identifikace souboru s nejmenším počtem sekvencí a náhodný výběr stejného množství sekvencí z ostatních vzorků, přičemž jen ty jsou dále používány pro analýzy. Nejjednodušší korekce rozdílné velikosti

souborů je také vztažení počtu identifikovaných kategorií na celkový počet sekvencí, ale v praxi se používají sofistikovanější algoritmy. Obdobné metody jsou velmi dobře rozpracované v oblasti bioinformatického zpracování expresních profilů vycházející z RNA sekvenování (RNA-seq). Nápadná shoda formátu vstupních dat – tedy na jedné straně metagenomických četnostních profilů a na straně druhé expresních profilů udávajících četnosti přepisovaných genů – nás vedla k využití bioinformatického přístupu hodnocení RNA-seq dat pro některé naše metagenomické analýzy, kterým se detailně věnuje kapitola 5 této práce.

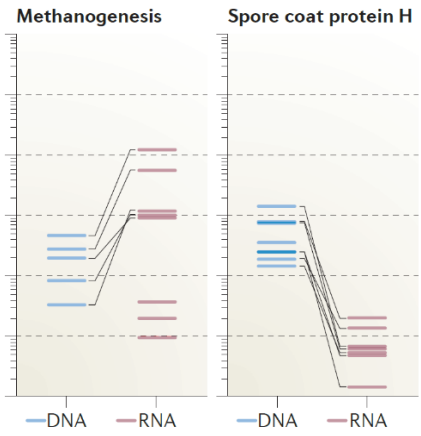
1.5.3 Metatranskriptomika a další „omiky“

Metagenomika postihuje, jaké organismy jsou v komunitě přítomné a jaké procesy mohou být schopny vykonávat. Funkční metagenomický profil představuje pouze funkční potenciál daného konsorcia, ovšem neříká nic o tom, které z genů, funkčních subsystémů a celých drah jsou v daný čas na daném místě skutečně využívány. Až sekvenování RNA zpětně přepsané enzymem reverzní transkriptázou do cDNA s využitím NGS instrumentů umožnilo analýzu skutečně zapnutých, transkribovaných genů a studium právě probíhajících funkcí uvnitř mikrobiálních konsorcií. Obdobně se pro danou oblast výzkumu vžilo pojmenování metatranskriptomika (Bikel et al., 2015; Franzosa et al., 2015).

V době, kdy jsme v roce 2008 začínali s vlastními metagenomickými projekty, byla metatranskriptomika spíše teoretickou, nedosažitelnou možností a to i kvůli tak zdánlivě bazálnímu problému, jako je izolace RNA z environmentálních vzorků v dostatečném množství a kvalitě. Obrovský potenciál sekvenování environmentální RNA však byl velmi rychle rozpoznán a dnes již existují komerčně dostupné sady na izolaci RNA i následné odstranění rRNA, která ve vzorku celkové RNA převládá (Giannoukos et al., 2012) za účelem získání co nejvíce čtení pocházejících z mRNA poskytujících aktuální funkční profil.

Odhady počtu molekul mRNA v jedné buňce u kultivované, exponenciálně se množící *E. coli* dosahují rozmezí asi 1300 až 1800 mRNA molekul v daný čas (Neidhardt, 1996; Taniguchi et al., 2010). To je relativně malé množství oproti jiným makromolekulám (> 2 mil. proteinů na buňku) a oproti samotnému obsahu genů v genomu (> 3000) (Moran et al., 2013). Moran et al. (2013) navíc odhadují pouze v průměru 200 molekul mRNA na jednu buňku u planktonních bakterií z environmetálních vzorků, přičemž autoři předpokládají podobný počet také u bakterií v jiných ekosystémech, jako jsou sladkovodní zdroje či půda. Vedle toho i procesy metabolismu mRNA jsou u bakterií velmi rychlé a doba přetrvání jednotlivých molekul v buňkách krátká (zhruba několik minut – viz Steglich et al., 2010; Taniguchi et al., 2010).

Samotné zpracování vzorků vyžaduje rychlou inaktivaci, aby se zabránilo metabolismu a také degradaci RNA, která je navíc méně stabilní oproti DNA. Během procesu reverzní transkripce doprovázeném amplifikací výsledné cDNA mohou být do sekvencí zanášené chyby. Zpracování sekvenačních dat zahrnuje prvotní trimování podle kvality, například program SEECER je zaměřen na RNA-seq data (Le et al., 2013). Sestavení metatranskriptomových sekvencí vede oproti assembly čtení z genomové či metagenomové DNA k rekonstrukci relativně krátkých původních mRNA molekul (transkriptů), například u bakterií je jejich průměrná délka přibližně 900 nukleotidů (Xu et al., 2006). Pro hodnocení rozdílné abundance transkriptů a tedy odhady diferenční exprese genů mezi mikrobiálními populacemi je potřeba opět pracovat s počty původních, nesestavených čtení. Ty můžeme zpětně mapovat buď právě na cDNA assembly, nebo v případě jejich dostupnosti na referenční genomové sekvence. Nejlépe je kombinovat DNA i RNA sekvenování ze stejných vzorků, přičemž složené DNA sekvence mohou sloužit jako reference pro mapování, ale hlavně počet jednotlivých DNA čtení pokrývající kódující geny poskytuje důležitý odhad množství přítomných genů, vůči kterému je nutné vztáhnout sledované množství transkriptů pro analýzu rozdílné exprese různě četných genů v konsorciu. Takový přístup, kde metagenomická čtení poskytují *baseline* pro odhad upregulovaných a downregulovaných genů, byl využit například pro studium komunit obývajících lidský trávicí trakt (obrázek 1.16) (Franzosa et al., 2014). Obdobně jsme k analýze rozdílně exprimovaných genů přistupovali při výzkumu bakteriální komunity z důlního ekosystému (kapitola 4). Poněkud jiný přístup, a sice paralelní PCR amplifikace genů z environmentálních vzorků DNA i RNA, nám zase umožnil rozlišení přítomných a aktivních mikroorganismů účastnících se procesů dekompozice v půdě (Baldrian et al., 2012).



Obrázek 1.16: Příklad analýzy exprese genů na základě počtu RNA-seq čtení, která pokrývají jeden gen (či funkční systém), kdy čtení z genomové DNA poskytují základ pro normalizaci při porovnání více genů (převzato a upraveno z Franzosa et al., 2015),

Nutno připustit, že cesta od transkriptu k funkčnímu proteinu není přímočará a nemůžeme postihnout například možné posttranslační procesy nezbytné pro aktivitu proteinů. Rovněž vztah mezi množstvím mRNA molekul a aktivitou příslušných kódovaných produktů není tak jasný, jak by se mohlo jevit. Roli může hrát třeba čas přetrvání aktivního enzymu v buňce ještě určitou dobu po degradaci samotné mRNA. Z tohoto hlediska tedy stojí metatranskriptomika stále na půl cesty. Ještě přímější způsob k analýze funkční aktivity mikrobiálních komunit tak představuje zkoumání proteinů – metaproteomika. Při zaměření na enzymatické procesy pak můžeme využívat metody metabolomiky na úrovni celých společenstev. V duchu ostatních názvů bychom měli na mysli vlastně meta-metabolomiku, což zní již docela komicky. Podle shodné koncovky -omika (*-omics* v angličtině) se pro využití těchto oblastí při studiu mikrobiálních konsorcií nabízí označení meta-omiky (*meta-omics*), dobrý přehled kombinování různých přístupů poskytují Franzosa et al. (2015).

2 Cíle

- Sekvenační analýzou genomové DNA s využitím GS FLX přístroje stanovit sekvenci genomu bakterie *Achromobacter xylosoxidans* A8, která je schopná využívat chlorbenzoáty jako zdroj uhlíku a energie. Na jejím základě predikovat sekvence kódující geny, vytvořit jejich funkční anotaci a poskytnout základ pro analýzu degradační dráhy chlorbenzoátů v kontextu kompletní genetické informace.
- Aplikovat přístupy ověřené na základě analýzy genomu *A. xylosoxidans* A8 pro výzkum taxonomicky jednoduché komunity tvořící krápníkové biofilmy v prostředí extrémně kyselé důlní vody bývalého dolu ve Zlatých Horách (Česká Republika). Rovněž amplifikací a sekvenací 16S rRNA genů určit složení daného mikrobiálního konsorcia.
- Sekvenovat celkovou metagenomovou DNA ze vzorku biofilmu ze Zlatých Hor, sestavit souvislé sekvence pokrývající genomy majoritně zastoupených mikroorganismů a anotovat predikované geny. Na základě anotace odhadnout fyziologické a metabolické vlastnosti mikroorganismů vyvinuté v konkrétním ekosystému s extrémně nízkou hodnotou pH. Prostřednictvím sekvenování cDNA zpětně přepsané z mRNA izolované z biofilmu identifikovat exprimované geny aktivně využívané zastoupenými organismy.
- Aplikací amplikonového a shotgunového sekvenování a adaptací bioinformatických metod analyzovat komplexní společenstva mikroorganismů v půdě kontaminované polychlorovanými bifenyly. Sekvenovat amplikony 16S rRNA genů a celkovou metagenomovou DNA, anotovat čtení a sestavit taxonomické a funkční profily zkoumaných vzorků.
- Stanovit vliv kořenů rostlin a minerálního hnojiva na strukturu a funkční potenciál konsorcií půdních organismů. Provést komparativní analýzu taxonomických a funkčních profilů konsorcií ovlivněných přítomností křenu (*Armoracia rusticana*), lilku (*Solanum nigrum*) a tabáku (*Nicotiana tobacum*) a vlivem minerálního hnojiva.

3 Stanovení a bioinformatická analýza kompletní genomové sekvence *Achromobacter xylosoxidans* A8

Achromobacter xylosoxidans A8, aerobní, gram-negativní tyčinkovitá bakterie ze třídy Betaproteobacteria, je aktérem v zajímavém příběhu prolínajícím se historií molekulárně-biologických metod na poli studia členů mikrobiálních komunit v prostředí. Původně byla tato konkrétní linie izolována z půdy kontaminované polychlorovanými bifenyly (PCB) a byla popsána v roce 1990 jako *Pseudomonas sp.* A8 společně s dalšími degradéry chlorbenzoátů, produktů dráhy degradace PCB (Pavlů et al., 1999). Bakterie schopná využívat 2-chlorbenzoát (2-CB) a 2,5-dichlorbenzoát (2,5-DCB) jako jediný zdroj uhlíku a energie byla následně na základě sekvence 16S rRNA genu zařazena do druhu *Achromobacter xylosoxidans* s označením linie A8. Geny zodpovědné za degradaci chlorbenzoátů byly identifikovány na jeho plasmidu pA81, jehož úsek dlouhý 12,4 kb byl sekvenován Sangerovou metodou a publikován v roce 2004 (Jenčová et al., 2004). Stejným postupem byla také stanovena celá sekvence plasmidu pA81 dlouhá 98,2 kb (Jenčová et al., 2008). Následná sekvenace genomu *A. xyloosoxidans* A8 byl jeden z našich prvních projektů celogenomového sekvenování s využitím *next-generation* platformy a jeho publikace v roce 2011 představuje první kompletní bakteriální genom stanovený 454 pyrosekvenací v české laboratoři (Strnad & Rídl et al., 2011).

Chlorbenzoáty vznikají při mikrobiální biodegradaci PCB v prostředí a jsou schopné průběh těchto procesů zpětně ovlivňovat. PCB jsou velmi stabilní, byly v minulosti široce užívány pro různé aplikace a představují velmi rozšířené polutanty. Pro průběh bioremediace – procesu, při kterém jsou metabolismem organismů odbourávány polutanty – je důležitá přítomnost mikrobiálního konsorcia, jehož jednotliví členové produkují enzymy schopné vykonávat postupně všechny kroky dráhy, tedy degradovat nejen výchozí PCB, ale rovněž toxické meziprodukty jako jsou chlorbenzoáty (Pavlů et al., 1999). Analýzou celých komplexních mikrobiálních konsorcií, která vznikají v půdě kontaminované PCB, se zabývá kapitola 5 této práce. V případě, kdy se některé z aktivních mikroorganismů podaří izolovat a pěstovat v laboratoři, otevírá se možnost detailního studia fenotypových a genotypových vlastností a tyto informace pak mohou zpětně sloužit k pochopení fungování přinejmenším části spektra daného konsorcia. Příkladem je právě bakterie *A. xylosoxidans* A8 a tato kapitola je věnována sekvenaci a anotaci jeho genomu.

3.1 Materiál a metody

3.1.1 Kultivace, izolace DNA a sekvenování

A. xylosoxidans A8 izolovaný z půdy kontaminové PCB (Pavlů et al., 1990) byl kultivován při teplotě 28°C v ABC minerálním mediu s přidaným 2,5-DCB pro zajištění uchování degradačních genů (podle Jenčová et al., 2004). Celkovou DNA jsme izolovali za použití Gentra Puregene Yeast/Bact. A kitu (Qiagen) dle protokolu výrobce. Pro 454 pyrosekvenování jsme připravili celkem tři typy knihoven: (1) shotgunovou knihovnu, (2) párovou knihovnu s délkou inzertu přibližně 3 kb a (3) párovou knihovnu s dlouhými inzerty (přibližně 8 kb) podle protokolů General Library Preparation Manual, Paired-End Library Preparation Manual – 3kb Span a 8-20kb Span (Roche). Sekvenování jsme provedli na přístroji GS FLX verze Titanium s využitím Titanium reagencií (Roche).

3.1.2 Sestavení čtení

Prvotní sestavení bylo provedeno programem Newbler verze 2.3 (Roche) s použitím přednastavených parametrů pro assembly čtení z genomové DNA. K vizuální kontrole assembly jsme použili program Tablet (Milne et al., 2010). Krátké mezery (*gaps,* gapy) mezi sousedními kontigy ve složených skafoldech nejsou zastoupeny především z důvodu sekvenačních chyb. V takovém případě umožňuje program Tablet vizualizovat čtení, která sice přesahují mezery mezi kontigy, ale vlivem rozdílných chybných sekvencí je program Newbler nedokáže jednoznačně spojit. Tato čtení byla extrahována a složena v programu Staden (Staden, 1996), což umožnilo postupně mezery zaplnit. Druhým případem jsou větší mezery, které plynou často z výskytu repetitivních elementů a duplikovaných genů. Typickým příkladem jsou geny pro 16S rRNA, které bakteriální genomy obsahují i v několika kopiích (Větrovský & Baldrian, 2013). Důsledkem jsou místa ve skafoldech, která počtem a délkou odpovídají duplikovaným oblastem. Sekvence genů program složí do samostatných kontigů, které nejsou jednoznačně lokalizované a jejichž relativní coverage oproti celkovému pokrytí závisí na počtu kopií genů. Ty je možné zařadit na několik odpovídajících míst v sekvenci genomu podle překryvů na koncích kontigů duplikovaných genů a sousedních kontigů ve skafoldech a podle orientace konců párových čtení.

3.1.3 Anotace

Predikci genů kódujících proteiny (*coding sequences*, CDS) jsme provedli paralelně třemi programy: Critica (Badger & Olsen, 1999), Glimmer (Delcher et al., 1999) a Prodigal (Hyatt et al., 2010). Několik případů, kdy se predikce ze všech programů plně neshodovala, jsme výsledky ověřili manuálně. Nejčastěji některý z programů odhadoval jiný začátek či konec téhož otevřeného čtecího rámce. Software tRNAscan (Lowe & Eddy, 1997) a Aragorn (Lalett & Canback, 2004) jsme použili pro nalezení genů pro transferové RNA (tRNA) a transferové-mediátorové RNA (tmRNA); ribosomální geny (rRNA) jsme identifikovali programem RNAmmer (Lagesen et al., 2007). Funkční anotaci jsme přiřadili predikovaným genům kódujícím proteiny na základě vyhledání podobnosti k sekvencím v databázích UniRef90 (Tatusov et al., 1999), NCBI-NR (Benson et al., 2010) a KEGG (Kanehisa et al., 2008) programem BLASTP s nastavením hodnoty očekávaného počtu sekvencí se stejnou či lepší podobností při stejném dotazu na danou databázi (*expectancy, E-value*) menší než 10^{-10}. Zároveň jsme využili pro anotaci automatické online nástroje RAST Annotation Server (Aziz et al., 2008) a IGS Annotation Engine (http://manatee.sourceforge.net) a výsledky porovnali s naší anotací obdobně jako v publikaci Strnad et al. (2010). Vyhledání podobností s databázovými sekvencemi v NCBI-NR jsme podrobili rovněž mezigenové oblasti mezi predikovanými CDS s využitím programu BLASTX s méně přísnějším nastavením kritéria *E-value* < 10^{-5}. Pokud predikované geny sdílely stejnou anotaci se sousedním genem nebo přilehlou mezigenovou oblastí, byla oblast identifikována jako možný výskyt posunu čtecího rámce (frameshift). Výsledky anotace byly ověřeny v programu Artemis (Carver et al., 2008) s vizualizací grafu pro využívání dikodonů (Pačes & Pačes, 2002). Predikované frameshifty byly posouzeny manuálně. V případě, že sousední oblasti obsahovaly části stejného anotovaného genu a čtecí rámec byl přerušen posunem v oblasti homopolymeru, bylo místo identifikováno jako sekvenační chyba a opraveno. Začátek replikace chromosomu *A. xylosoxidans* A8 jsme identifikovali podle průběhu rozdílného obsahu G a C bazí mezi vedoucím a zpožďujícím se vláknem replikace (tzv. GC *skew* analýza) (Lobry, 1998).

3.2 Výsledky a diskuze

3.2.1 Organizace genomu

Celkově jsme získali 1 022 247 shotgunových čtení, 494 741 párových čtení s délkou inzertu přibližně 3 kb a 623 041 párových čtení s inzertem 8 kb, což dohromady představuje přes 680 mil. bazí a poskytuje 92-násobné průměrné pokrytí genomu *A. xylosoxidans*. Výsledkem automatického assembly programem Newbler byly tři cirkulární skafoldy odpovídající molekule genomové DNA (chromosomu) o délce 7 013 095 bp a dvěma plasmidům: pA81 (98 156 bp) a pA82 (247 895 bp) (obrázek 3.1). Skafoldy obsahovaly celkem 65 mezer, které bylo možné všechny manuálně zaplnit lokálním assembly příslušných čtení a podle pozic párových čtení jak bylo popsáno výše. Genom přesahující 7 mil. bazí je relativně velký a blíží se velikosti genomu *Pseudomonas aeruginosa*, který byl popsán jako jakýsi „švýcarský nůž“ mezi prokaryotickými genomy (Aujoulat et al., 2012). Chromosom *A. xylosoxidans* A8 vykazuje vysoký obsah GC párů (66 %), plasmidy v průměru obsahují 62 % GC párů. Začátek replikace chromosomu, který u bakterií bývá konvenčně umísťován před gen *dnaA*, jsme identifikovali analýzou GC skew (obrázek 3.1) a ověřili jeho místo v blízkosti *dnaA* genu.

Nukleotidové sekvence jsou veřejně přístupné v databázi GenBank pod přístupovými kódy CP002287 (chromosom), CP002288 (plasmid pA81) a CP002289 (plasmid pA82).

3.2.2 Identifikované geny

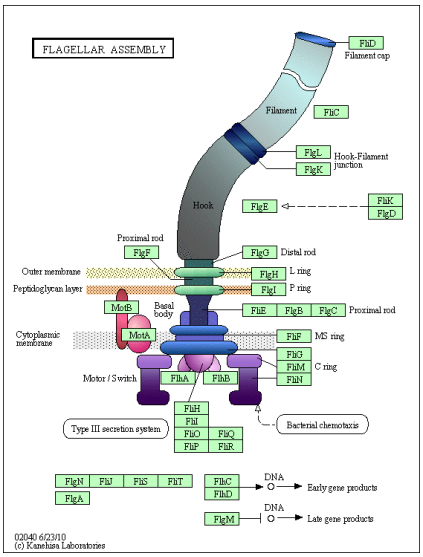
Identifikovali jsme 6 459 CDS na chromosomu *A. Xylosoxidans*, 104 CDS v rámci plasmidu pA81 a 252 CDS v plasmidu pA82. Chromosom dále obsahuje 3 rRNA operony, 60 tRNA pro přepis všech 20 aminokyselin a selenocysteinu a jeden gen pro tmRNA. Kódující oblasti pokrývají 91,3% genomu *A. xylosoxidans*. Na základě anotace byla přiřazena funkce 5 620 nalezeným CDS (82,5%), funkci u 783 (11,5%) CDS se prohledáním databází přiřadit nepodařilo. Ostatních 412 predikovaných CDS nevykazovalo podobnost s žádným záznamem v databázích.

Obrázek 3.1: Grafické znázornění chromosomu *A. xylosoxidans* A8 a dvou plasmidů pA81 a pA82. Cirkulární obrázky znázorňují (od periferie ke středu): (1) geny v (+) orientaci (barevné kódy COG hierarchie viz legenda v pravo), (2) geny v (-) orientaci (barevné kódy COG hierarchie viz legenda v pravo), (3) GC obsah (počítaný v okně 1000 bp), (4) GC skew (počítáno v okně 200 bp).

Genom obsahuje 59 inaktivních a/nebo mutovaných transpozáz a fágových genů, identifikovali jsme rovněž 8 dalších zkrácených pseudogenů. Přibližně tisícovka genů se uplatňuje v procesech buněčného transportu a 675 genů má regulační funkci. Kompletní sada genů pro strukturu bakteriálního bičíku (obrázek 3.2) umožňuje pohyb bakterie.

Nalezli jsme operon *ohbRAB* kódující a regulující enzym dioxygenázu iniciující degradaci chlorbenzoátů na chlorkatechol, kompletní klastr genů *mocpRABCD* pro katabolismus chlorkatecholů a *hyb* operon umožňující degradaci salicylátu, čímž jsme potvrdili jejich umístění na pA81 plasmidu (Jenčová et al., 2008). Dále jsme identifikovali přibližně 70

genů kódujících další dioxygenázy, které mohou být rovněž asociovány s bioremediačními schopnostmi bakterie.



Obrázek 3.2: Grafické znázornění genů pro strukturu bakteriálního bičíku *A. xylosoxidans* A8.

Kontaminace prostředí aromatickými polutanty je často doprovázena také vysokým obsahem těžkých kovů, které ovlivňují složení bakteriálních konsorcií (Gremion et al., 2004; Roane et al., 2001). Několik predikovaných genů *A. xylosoxidans* bylo asociováno s rezistencí vůči Pb^{2+}, Cd^{2+} a Zn^{2+} iontům prostřednictvím aktivního transportu efluxními pumpami, ATPázami typu P. Patří sem skupina *pbt* genů kódovaná rovněž na plasmidu pA81, jejíž funkce byla potvrzena skupinou dr. Kotrby expresí v bakterii *E. coli* (Hložková et al., 2013). Ztráta plasmidu pA81 však vede pouze k poklesu, nikoli ztrátě rezistence u *A. xylosoxidans*, což ukazuje na roli dalších ATPázových pump kódovaných geny v jeho chromosomu (Hložková et al., 2013).

Genom rovněž obsahuje 28 genů spojovaných s patogenními vlastnostmi a 33 genů asociovaných s produkcí toxinů a rezistencí vůči antibiotikům. Bakteriální druh *Achromobacter xylosoxidans* byl původně izolován z pacientů a popsán jako oportunistický patogen u jedinců s oslabenou imunitou (Holmes et al., 1977; Igra-Siegman et al., 1980). Li et al. (2013) provedli komparativní analýzu založenou na námi publikované sekvenci kompletního genomu společně s pěti dalšími dostupnými draftovými assembly genomů rodu *Achromobacter*, která naznačuje fylogenetickou příbuznost s lidskými patogeny rodu *Bordetella*. Studie odhalila unikátní vlastnosti patogenních druhů izolovaných z pacientů s cystickou fibrózou. Oproti environmentálním liniím nesou patogenní druhy rodu *Achromobacter* geny

umožňující adaptaci na hostitelské prostředí spojené například s adhezními a invazními vlastnostmi či potencionální produkcí toxinů (Li et al., 2013).

Vzhledem k tomu, že *Achromobacter xylosoxidans* A8 izolovaný z půdy byl nějakou dobu jediný zástupce druhu *A. xylosoxidans* se stanovenou kompletní sekvencí genomu, sloužil genom jako reference i pro další výzkumy patogenních linií spojených s cystickou fibrózou (Dupont et al., 2015). Dnes (květen 2016) je k dispozici dalších šest kompletně složených genomů druhů izolovaných z pacientů (GenBank BioProject kódy: PRJNA231221, PRJNA231221, PRJNA231221, PRJEB6403, PRJNA288995, PRJNA260837) a řada draftových sekvencí umožňující další studium patogenních vlastností (viz např. Badalamenti et al., 2015; Ridderberg et al., 2015; Ormerod et al., 2015).

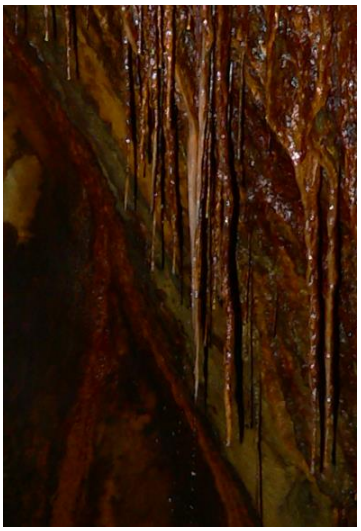
3.3 Závěry

Vhodně zvolená kombinace různých typů sekvenačních knihoven umožňuje sekvenování a assembly kompletních bakteriálních genomů výhradně s využitím 454 pyrosekvenace. Případná problematická místa v assembly a anotaci (mezery a frameshifty) genomu *A. xylosoxidans* A8 bylo možné vyřešit manuálně bez potřeby dalších dat, například bez zdlouhavého sekvenování PCR produktů.

Geny pro utilizaci chlorbenzoátů činí *A. xylosoxidans* A8 vhodným nástrojem pro bioremediační studie s potenciálem tvorby biologického systému k odbourávání polutantů. Kromě toho nese také geny poskytující rezistenci vůči těžkým kovům, jejichž výskyt může omezovat procesy biodegradace aromatických uhlovodíků v prostředí (Gremion et al., 2004; Roane et al., 2001). Kompletní stanovená genomová sekvence *A. xylosoxidans* A8 dovoluje další analýzy v kontextu celého genetického pozadí, ať již genů spojených s bioremediacemi či jinými bakteriálními funkcemi v ekosystémech. Server KEGG umožňuje procházet jednotlivé funkční dráhy *A. xylosoxidans* A8 (http://www.genome.jp/kegg-bin/show_pathway?axy01100), například prostřednictvím interaktivní vizualizace kompletní metabolické mapy (příloha 1). Výborným online nástrojem poskytujícím interaktivní prostředí pro vizualizaci a vyhledávání v genomu (příloha 2) je také server BioCyc (http://biocyc.org/organism-summary?object=AXYL762376).

4 Metagenomická analýza a rekonstrukce genomů členů bakteriálního společenstva z extrémně kyselého ekosystému důlní vody ve Zlatých Horách

Mikroorganismy schopné optimální rychlosti růstu v kyselém prostředí jsou nazývány acidofilní. Extrémně acidofilní zástupci žijící v biotopech s pH < 4 zahrnují fenotypově a fylogeneticky různorodou skupinu prokaryot a některé mikroskopické řasy a protozoa (Johnson, 2009). Ačkoli jednotlivé mikrobiální buňky nejsou pozorovatelné pouhým okem, dokáží některá acidofilní prokaryota formovat velké buněčné celky spojené extracelulárními polymery a jejich přítomnost je na pohled evidentní (Johnson, 2012). Kromě silných nánosů biofilmu jsou to řetězovité útvary ve vodních pramenech (v angličtině nazývané *acid streamers*) a také želatinové stalaktity na stropech jeskyní někdy obklopené tvrdou, mineralizovanou vrstvou tvořící brčka (*snottites*) (obrázek 4.1).



Obrázek 4.1: Krápníkový útvar bakteriálního biofilmu v šachtě bývalého dolu ve Zlatých Horách, Česká republika (autorem snímku je Lukáš Falteisek).

Tyto útvary byly sledovány na mnoha místech po světě v přirozených prostorách pod povrchem Země, v kyselých vyvěrajících pramenech nebo také v dolech, kde vlivem těžby dochází k expozici sulfidů v rudě, které podléhají mikrobiologické a abiotické oxidaci za vzniku solí kyseliny sírové a poklesu pH, čímž vzniká velmi kyselá důlní voda. (Banks et al., 1997; Schippers, 2004; Johnson & Hallberg, 2005). Nejrozšířenějším sulfidem k zemské kůře

je pyrit (disulfid železnatý, FeS_2), který poskytuje zároveň železo i síru mikroorganismům schopným jejich redukované formy využívat jako donor elektronu.

Prvním izolovaným, kultivovaným a popsaným acidofilním mikroorganismem byl *Acidithiobacillus* (původně *Thiobacillus*) *ferrooxidans* (Temple & Colmer, 1951), chemolitoautotrofní gamaproteobakterie schopná oxidace železa a síry jako jediného zdroje energie. Jeho přítomnost v kyselých vodách byla sledována celosvětově, obdobně byl také popsán široký výskyt například acidofilních autotrofních bakterií rodu *Leptospirillum* či acidofilních heterotrofů rodu *Acidiphilium* (Johnson, 1994; Baker & Banfield, 2003; Gonzalez-Toril et al., 2003; Méndez-García et al., 2015). Postupně se ukázalo, že bakterie popisované jako *At. ferrooxidans* často obsahují více než jeden druh a následně byla rozlišena další linie – *At. ferrivorans* (Hallberg et al., 2010). Zatímco *At. ferrooxidans* je schopen růstu v kyselejším prostředí (rozsah pH 1,3 – 4,5 oproti 1,9 – 3,4 u *At. ferrivorans*), je citlivější na nižší teploty a neroste při teplotě pod 10 °C (optimum 30 – 35 °C). *At. ferrivorans* je schopný růst při teplotě v rozmezí 4 – 37 °C, přičemž optimální rychlosti růstu bylo dosaženo za teplot v rozmezí 28 – 33 °C (Hallberg et al., 2009; Hallberg et al., 2010).

Ve velmi kyselých ekosystémech byla po dlouhou sledována také obtížně kultivovatelná bakterie definovaná pouze na základě amplifikace 16S rRNA genu jako člen třídy Betaproteobacteria a pojmenovaná *Ferrovum myxofaciens* (Johnson & Hallberg, 2005; Johnson et al., 2014). Modifikací kultivačních postupů se podařilo vypěstovat typovou linii *Ferrovum myxofaciens* P3G (Johnson et al., 2014). Sekvenací s využitím MiSeq přístroje byla stanovena nekompletní, draftová sekvence jejího genomu (Moya-Beltrán et al., 2014). Sekvence pokrývající téměř kompletní genom zástupce rodu *Ferrovum* (s označením FB7) se podařilo složit rovněž z metagenomických sekvenačních dat z kyselé vody v dole Fankou (Guangdong, jižní Čína) (Hua et al., 2015) a draftový genom linie JA12 rodu *Ferrovum* poskytlo sekvenování DNA jednoduché komunity schopné oxidovat železo s dominantně zastoupenými dvěma bakteriemi ve vodě z lokality Tzschelln (Lusatia, Německo) (Ullrich et al., 2016).

Výskyt či relativní zastoupení identifikovaných druhů acidofilních bakterií na různých lokalitách je dán drobnými rozdíly fyzikálně-chemických vlastností ekosystému. Zástupci rodu *Ferrovum* preferují vyšší pH a vyšší obsah redukovaného železa než bakterie *At. ferrooxidans*, která dokáže tolerovat kyselejší prostředí a velmi efektivně využívat i nízké koncentrace železa (Santofimia et al., 2013; Jones et al., 2014; Wang et al., 2015). Například v uhelném dole v Apalačských horách (Pensylvánie, USA) byl identifikován *At. ferrooxidans* ve vodě s pH < 2,7

a koncentrací železnatých iontů menší než 5 mM, zatímco v pramenech s pH hodnotou kolem 3 a vyšším obsahem železa dominovaly bakterie rodu *Ferrovum* (Jones et al., 2014). V roce 2012 publikovali Lukáš Falteisek a Ivan Čepička studii mikroorganismů v prostředí bývalého dolu na těžbu mědi ve Zlatých Horách (Česká Republika). Celkem 15 prezentovaných vzorků pokrývajících různé části dolu umožnilo analýzy rozdílných biotopů, identifikaci jejich hlavních mikroorganismů a odvození vztahů v rámci jedné lokality. Na základě rozdílného výskytu v několika vzorcích byly vůbec poprvé popsány ekologické rozdíly *At. ferrooxidans* a *At. ferrivorans* v přirozeném prostředí. V rámci studie byl také získán vzorek želatinového krápníkového biofilmu s označením ZH7, který na základě analýzy několika sekvencí 16S rRNA představuje jednoduchou bakteriální komunitu s dominantním zastoupením bakterie *Ferrovum myxofaciens* doprovázené minoritním výskytem *At. ferrrivorans* (Falteisek & Čepička, 2012). Právě tuto bakteriální populaci jsme se rozhodli dále studovat metagenomickým přístupem s využitím sekvenování celkové DNA s potenciálem sestavení kompletních genomových sekvencí přítomných acidofilních druhů bez nutnosti jejich kultivace. Pro odhalení aktivních členů společenstva a analýzu exprimovaných genů jsme sekvenovali rovněž mRNA.

4.1 Materiál a metody

4.1.1 Důlní ekosystém ve Zlatých Horách a vzorek biofilmu

Oblast Zlatých Hor sloužila k těžbě rud již od 14. století. Pyrit, chalkopyrit a pyrhotin s obsahem měděných žil byl extenzivně těžen v letech 1965 až 1990, celkově bylo důlních aktivit zanecháno v roce 1993 (Kotris, 2004). Výsledkem rozpouštění nahromaděného odpadního pyritu vzniká typická kyselá důlní voda. Bývalé šachty jsou dnes přístupné až do hloubky 350 metrů. Celý systém je přirozeně ventilovaný se stálou teplotou vzduchu 8 – 10 °C a teplotou vody 8 – 9 °C (Falteisek & Čepička, 2012).

Několik centimetrů želatinového bakteriálního stalaktitu bylo odebráno do sterilní 15ml falkony a označeno jako vzorek ZH7 v rámci předchozí studie Falteisek & Čepička (2012). Paralelně se vzorkem biofilmu byla pro chemické analýzy odebrána voda, která jím protékala ze stropu šachty v hloubce přibližně 190 m pod zemí. Fyzikálně-chemické vlastnosti byly stanoveny a popsány Falteiskem a Čepičkou (2012), souhrnně jde o kyselou vodu s pH

hodnotou 2,9, teplotou 8,7 °C, s velmi nízkou koncentrací organického uhlíku a celkovou mineralizací přibližně 5,8 g/l, tvořenou převážně sírany, Fe, Al, Mg a Cu (tabulka 4.1).

Tabulka 4.1: Fyzikálně-chemické vlastnosti vody ze vzorku ZH7. Všechny koncentrace jsou uvedeny v mg/l, TOC (*total organic carbon*, celkový obsah organického uhlíku), CODmn (*chemical oxygen demand*, biochemická spotřeba kyslíku – sumární obsahu organických látek manganometrickým stanovením).

pH	Amoniak	Dusitany	Dusičnany	Chloridy	Sírany	Fluoridy	Na	K	Ca	Mg	Fe	Mn	Si	CODmn	Fosfáty	Huminové látky	TOC	As	Ba	Be	Al	Cr	Cd	Co	Cu	Ni	Pb	Ag	Zn	Li
2,7	0	0	2,7	8,3	3509	5,1	2,9	2,5	49,7	97,4	774	26,7	18	71,2	0	0	7,7	0,225	0,82	0,0039	131	0,11	0,048	1,5	90	0,74	0	0	5,9	0,09

4.1.2 Izolace DNA, 454 sekvenace a MiSeq sekvenace

Želatinový stalaktit tvořený bakteriálními buňkami byl v laboratoři rozdělen sterilním skalpelem a pinzetou. Přibližně 250 mg bylo použito pro izolaci v několika paralelních reakcích s využitím ZR Soil Microbe DNA Kitu (Zymo Research) (Falteisek & Čepička, 2012). Izolovaná DNA byla smíchána dohromady pro dosažení dostatečného množství k tvorbě následujících sekvenačních knihoven. Pro 454 sekvenování jsme vytvořili opět tři typy knihoven dle protokolů od firmy Roche: (1) shotgunovou (General Library Preparation Manual), (2) párovou s inzertem 3 kb (Paired-End Library Preparation Manual – 3kb Span) a (3) párovou knihovnu s inzertem kolem 8 kb (Paired-End Library Preparation Manual – 8kb Span). Na základě našich zkušeností s výskytem duplikovaných čtení v párových knihovnách jsme se rozhodli připravit poslední zmiňovanou párovou knihovnu ve 4 oddělených reakcích se skutečnou výslednou délkou inzertů (na základě pozic párových čtení u výsledných kontigů) v rozmezí v průměru 5,8 – 8,8 kb. To umožnilo vícenásobnou sekvenaci a získání většího počtu unikátních čtení (viz kapitola 1.4.2). Sekvenace proběhla na přístroji GS FLX verze Titanium. K odstranění duplikátů a uchování pouze unikátních čtení pro následující analýzy (dereplikaci) jsme použili program CD-HIT (http://weizhongli-lab.org/cd-hit/).

Pro sekvenování na platformě od firmy Illumina jsme izolovanou DNA fragmentovali v 18 cyklech na sonikátoru Bioruptor Next Gen (Diagenode), což vedlo k získání úseků s průměrnou délkou přibližně 400 bp. Ty jsme zpracovali sadou TruSeq DNA Sample Prep Kit (Illumina) dle protokolu výrobce (TruSeq DNA Sample Preparation Guide, Illumina) a výslednou knihovnu jsme podrobili sekvenaci na přístroji MiSeq (Illumina) v režimu délky čtení 250 bp z obou konců (2x250 paired-end). Výsledná čtení jsme procesovali programem

FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) pro kontrolu kvality, trimování nekvalitních úseků a odstranění pozůstatků sekvenačních adaptorů.

4.1.3 16S rRNA amplifikace a sekvenace amplikonů

Úsek 16S rRNA genu byl amplifikován prostřednictvím PCR s využitím sady primerů: *forward* f8-27: 5'- AGAGTTTGATCMTGGCTCA-3' a *reverse* r372-357: 5'- ACGGGCGGTGTGTRC-3' (Shyu et al., 2007). Každý z primerů obsahoval navíc na 5'-konci jeden ze sekvenačních adaptorů (podle 454 Sequencing Application Brief No. 001-2009, Roche), což zajišťuje možnost přímého sekvenování výsledných amplikonů na přístroji GS FLX (Roche). Reakční směs v celkovém objemu 20 µl obsahovala 0,2mM dNTPs (Finnzymes, Finland), 0,25µM primery (Generi Biotech, Česká Republika), 0,1mg/ml BSA (*bovine serum* albumin, New England BioLabs, Great Britain), 0,4 jednotky (U) Phusion Hot Start II DNA Polymerázy (Finnzymes, Finland) s odpovídajícím pufrem a templátovou DNA (10-50 ng). Cykly amplifikace byly následující: 98 °C po dobu 30 s, 35 cyklů 98 °C po dobu 10 s, 60 °C po dobu 30 s, 72 °C po dobu 60 s a závěrečná extenze 72 °C po dobu 10 min. Výsledné produkty byly přečištěny s využitím AMPure XP kuliček (Agencourt, Beckman Coulter, USA) dle manuálu pro odstranění krátkých fragmentů (s délkou pod 200bp), které mají tendenci převládat v následné přípravě sekvenační reakce. Produkty byly následně sekvenovány z jednoho směru (od *forward* primeru) na přístroji GS FLX verze Titanium s využitím Titanium reagencií (Roche).

4.1.4 Izolace RNA a sekvenace

Řetězcovité želatinové bakteriální struktury byly rozrušeny mechanicky roztřepáním v RLT izolačním pufru (Qiagen) společně s přidanou 5mm kovovou kuličkou pomocí přístroje TissueLyser dle doporučení výrobce (Qiagen). Totální RNA byla následně izolována z homogenizovaného vzorku prostřednictvím Qiagen RNeasy Mini Kit dle manuálu výrobce (RNeasy Mini Handbook, Qiagen). K odstranění ribosomální RNA jsme použili sadu Ribo-Zero rRNA Kit dle protokolu výrobce (Epicentre, nyní Illumina) určenou pro vychytání rRNA molekul gramnegativních i grampozitivních bakterií pomocí prób vázaných na magnetické kuličky.

Vzorek RNA s odstraněnou ribosomální RNA byl následně použit pro přípravu sekvenační knihovny sadou NEXTflex RNA-seq Kit (BIOO Scientific) podle protokolu

výrobce, který zahrnuje fragmentaci RNA, reverzní transkripci do cDNA, ligaci sekvenačních adaptorů a PCR amplifikaci. Následná sekvenace proběhla na přístroji MiSeq (Illumina) s nastavením maximální délky čtení 150bp z obou stran fragmentů (2x150 paired-end). Kontrolu kvality, korekci a trimování jsme provedli v programu SEECER (Le et al., 2013). Hodnocení výskytu rRNA, která unikla procesu odstranění na kuličkách, jsme provedli programem SortMeRNA (Kopylova et al., 2012).

4.1.5 Analýza 16S rRNA amplikonů

16S rRNA čtení byla zpracována v softwaru Mothur verze 1.28 (Schloss et al., 2009) následujícím postupem: (1) použili jsme 450 cyklů (*flows*) sekvenace; (2) pro odstranění nekvalitních sekvencí (tzv. odstranění šumu, *denoising*) byly hodnoceny intenzity signálu v sekvenačních cyklech (*flowgramy*) algoritmem PyroNoise (Quince et al., 2009); (3) sekvence primerů a identifikátorů na začátku čtení byly ořezány; (4) čtení byla dále zarovnána podle referenčního alignmentu SILVA (verze 119) a pokud nedosahovala délky minimálně 240 bp, byla z dalších analýz vyřazena; (5) čtení byla klastrována (*single-linkage pre-clustering*) s nastavením minimálního rozdílu jedné záměny na 100 bp; (6) chybná čtení, která vznikají při PCR přeskočením polymerázy z jedné templátové molekuly DNA na jinou (chimerní sekvence) byla identifikována implementovaných programem Perseus (Quince et al., 2011) a odstraněna; (7) odstraněny byly také sekvence, které se vyskytují v souborech jen jednou a nebyly přiřazeny žádnému klastru (*singletony*) a rovněž sekvence identifikované programem Mothur jako kontaminace pocházející z mitochondrií, chloroplastů a z eukaryotických genomů; (8) získané kvalitní sekvence byly klastrovány s podobností minimálně 97 % do operačních taxonomických jednotek. Výsledné OTU byly klasifikovány na základě jejich podobnosti s geny databáze RDP s využitím programu RDP Classifier (Wang et al., 2007). Pro reprezentativní sekvenci každé OTU jsme rovněž vyhledali podobné sekvence v NCBI-NT programem BLASTN.

4.1.6 Sestavení sekvencí

454 čtení z genomové DNA (454 data) jsme sestavili v programu Newbler verze 2.8 (Roche) s požitím přednastavených parametrů pro sestavení sekvencí genomové DNA. Pro zaplnění mezer mezi sousedními kontigy ve skafoldech jsme použili čtení z genomové knihovny (gDNA-MiSeq data) a RNA-seq čtení (RNA-MiSeq data) z přístroje MiSeq, která jsme mapovali na assembly programem GapCloser vyvinutým za tímto účelem v rámci

projektu SOAPdenovo2 (Lua et al., 2012). Toto výsledné sestavení je považováno za finální a sloužilo pro další analýzy, není-li uvedeno jinak. Přestože Newbler umožňuje procesování rovněž dat z MiSeq sekvenátoru, množství získaných čtení přesahuje možnosti algoritmu Newbleru pro dosažení assembly s využitím nám dostupných výpočetních klastrových počítačů při kombinaci všech získaných sekvencí. Kombinované assembly 454 a gDNA-MiSeq čtení bylo proto provedeno v programu MaSuRCA verze 2.1.0 (Zimin et al., 2013). Rozlišení čtení pocházejících z genomu bakterie rodu *Acidithiobacillus* pro oddělené assembly, jsme provedli mapováním na dostupné referenční genomové sekvence *Acidithiobacillus ferrooxidans* ATCC 23270 a ATCC 53993 (RefSeq přístupové kódy NC_011761, NC_011206) a *Acidithiobacillus ferrivorans* SS3 (NC_015942) programem BWA (Li & Durbin, 2009).

Pro assembly RNA-MiSeq čtení jsme použili program Trinity vyvinutý specificky za účelem sestavení cDNA sekvencí (Grabherr et al., 2011).

4.1.7 Predikce genů a anotace

Predikci genů kódujících proteiny (CDS) jsme provedli stejný postupem jako při analýze genomu *Achromobacter xylosoxidans* A8 (kapitola 3) s využitím programů Critica (Badger & Olsen, 1999), Glimmer (Delcher et al., 1999) a Prodigal (Hyatt et al., 2010). V případech, kdy programy predikovaly odlišné 5' konce CDS, jsme pro další analýzy použili delší z predikovaných úseků. Rovněž jsme dále hodnotili i CDS predikované jen dvěma či jedním z programů. Pro predikované CDS jsme vyhledali podobné sekvence v databázi NCBI-NR programem BLASTP (*E-value* < 10^{-10}). Rovněž jsme využili anotační server RAST (Aziz et al., 2008) pro predikci a funkční anotaci genů a za účelem zařazení genů do funkčních drah jsme použili KAAS – KEGG Automatic Annotation Server (Moriya et al., 2007).

4.1.8 Roztřídění skafoldů (binning)

Pro roztřídění sestavených kontigů a přiřazení výsledných skupin jednotlivým genomům přítomných organismů jsme použili kombinovaný přístup založený na (1) kompozici sekvencí, (2) výsledném sestavení a (3) anotaci.

První cesta zahrnovala hodnocení výskytu sdílených „slov" o délce čtyř nukleotidů (tetramerů) pro složené kontigy v programu ClaMS (Pati et al., 2011), který využívá Markovova řetězce pro modelování každé sekvence jako cesty v de Bruijnově grafu, čímž vytváří signaturu, která je porovnaná se signaturou v poskytnutém referenčním setu sekvencí.

Vzhledem k předpokladu majoritního zastoupení bakterií rodu *Ferrovum* a *Acidithiobacillus*, jako referenční data byly použity dostupné kompletní či draftové sekvence genomů *Acidithiobacillus ferrooxidans* ATCC 23270 a ATCC 53993 (RefSeq přístupové kódy NC_011761, NC_011206), *Acidithiobacillus ferrivorans* SS3 (NC_015942), *Ferrovum myxofaciens* P3G (NZ_JPOQ01000001) a *Ferrovum* spp. JA12 (NZ_LJWX01000001).

Dalšími kroky při hodnocení kompozice kontigových sekvencí byl výpočet průměrného obsahu GC párů a hodnocení preference užívání dikodonů, která může být specifická pro různé organismy (Tats et al., 2008). Pro stanovení preferovaných dikodonů pro CDS v rámci jednotlivých kontigů jsme vytvořili vlastní scripty v jazyce Perl. Matice výskytu sdílených dikodonů mezi kontigy byla použita pro vizualizaci struktury v datech pomocí korespondenční analýzy ve statistickém programu R (R Development Core Team, 2009).

Roztřídění na základě parametrů výsledného assembly spočívalo v hodnocení průměrné míry pokrytí kontigových sekvencí jednotlivými čteními. Míra pokrytí zhruba odpovídá zastoupení jednotlivých organismů v populaci, což dovoluje odlišit genomové sekvence více a méně četných organismů. Sofistikovanější metodou bylo využití informací o pozicích konců párových čtení směřujících do různých kontigů. V tomto ohledu je důležité, že jsme měli k dispozici párová čtení i s poměrně dlouhou délkou inzertu (až 8,8 kb). I v případech, kdy vlivem sekvenačních chyb nebo výskytem duplikovaných čtení není schopen assembler spojit sousední kontigy či je dokonce přiřadit stejnému skafoldu, lze použít informace o pozicích párových čtení a jejich postupným procházením vytvořit síť kontigů, které budou s největší pravděpodobností pocházet z jednoho genomu. Vznikající síť jsme vizualizovali s využitím programu Cytoscape (Shannon et al., 2003).

Konečným vodítkem pro roztřídění sestavených sekvencí byla anotace predikovaných CDS přináležejících jednotlivým kontigům na základě vyhledání homologních genů v databázi NCBI-NR a informace o jejich taxonomické příslušnosti. Vzhledem k tomu, že některé predikované geny vykazují podobnost k většímu počtu genů různých organismů a pro některé CDS jsou v databázích jen vzdáleně podobné geny, hodnotili jsme vždy 20 genů s nejvyšší mírou podobnosti k jedné CDS. Anotované CDS tak bylo možné zařadit do vyšších taxonomických kategorií, spolehlivě na úrovni třídy.

4.1.9 Mapování a prohledávání sestavených sekvencí

Mapování RNA-MiSeq čtení na anotované kontigy s predikovanými kódujícími úseky jsme provedli programem STAR (Dobin et al., 2013). Stejným způsobem jsme mapovali RNA-

MiSeq čtení na reprezentativní sekvence jednotlivých OTU vytvořených programem Mothur na základě amplikonového sekvenování 16S rRNA genů.

Pro vyhledání potencionálních homologních CDS k dále popisovaným vybraným genům v sestavených nukleotidových sekvencích kontigů z programu Newbler a kontigů vytvořených assemblerem Trinity z MiSeq-RNA čtení jsme používali program TBLASTN (*E-value* < 0,01).

4.1.10 Analýza transkribovaných genů

Relativní transkripční aktivita (RTA) genů byla stanovena na základě počtu MiSeq-RNA čtení mapovaných do úseků jednotlivých CDS podle vztahu:

$$RTA_{ab} = \frac{cDNA_{ab}}{DNA_{ab}},$$

kde $cDNA_{ab}$ je relativní četnost MiSeq-RNA čtení mapovaných do úseku genu *a* v genomu (skupině roztříděných kontigů) *b*, DNA_{ab} je průměrná míra pokrytí kontigu obsahujího gen *a* v genomu *b*. Relativní četnost byla počítána jako procento čtení mapovaných do genu *a* z celkového počtu všech čtení mapovaných do genomu *b*.

4.2 Výsledky a diskuze

4.2.1 16S rRNA geny a taxonomické složení vzorku

Po zpracování amplikonových sekvencí programem Mothur jsme obdrželi 5191 sekvencí, které byly přiřazeny do 13 OTU na úrovni 97% podobnosti (tabulka 4.2). Z výsledků je patrné, že ve vzorku jsou dominantně zastoupené tři bakterie rodů *Ferrovum, Acidithiobacillus* a třetí neznámé proteobakterii, jejíž 16S rRNA sekvence se shoduje v databázi NCBI-NT s 16S rRNA genem nekultivované bakterie amplifikovaným z environmentálního vzorku. RDP klasifikátor identifikoval vzdálenou příbuznost k 16S rRNA rodu *Sideroxydans*. Řádově nižší zastoupení vykazuje bakterie rodu *Acidiphilium,* jehož acidofilní heterotrofní zástupci byli dříve popsáni v konsorciích tvořených majoritně rodem *Ferrovum* (Ullrich et al., 2016; Johnson et al., 2014). Obdobně jsou zastoupeny další tři dosud nepopsané bakterie, doprovázené ostatními sporadicky zastoupenými OTU.

Reprezentativní sekvence OTU1 vykazovala 100% sekvenční shodu s druhem *Ferrovum myxofaciens* P3G; pro potřeby následujícího textu budeme pro námi analyzovanou bakterii používat označení „*Ferrovum myxofaciens*“ ZH7. Bakterie rodu *Acidithiobacillus,*

která se vyskytuje v našem vzorku (OTU2), nese 16S rRNA gen s 99% podobností k druhu *Acidithiobacillus ferrooxidans* a nadále o ni budeme referovat jako o „*Acidithiobacillus*“ linie ZH7B.

Tabulka 4.2: Přehled počtu amplikonavých 16S rRNA čtení (# amplikon) náležejících jednotlivým OTU a počtu MiSeq-RNA čtení (# RNA-seq) mapovaných na reprezentativní sekvence OTU. BLASTN vs. NCBI-NT a podobnost [%] udává klasifikaci na základě vyhledání podobností proti NCBI-NT databázi a míru podobnosti se záznamem v NCBI-NT. RDP klasifikace uvádí taxonomickou klasifikaci reprezentativních sekvencí OTU na základě podobnosti s RDP databází.

OTU	# amplikon	# RNA-seq	BLASTN vs. NCBI-NT	podobnost [%]	RDP klasifikace
OTU1	1736	27034	Ferrovum	100	Proteobacteria;100%; Betaproteobacteria;100%; Ferrovales;100%; Ferrovaceae;100%; Ferrovum;100%
OTU2	1938	3540	Acidithiobacillus	99	Proteobacteria;100%; Gammaproteobacteria;100%; Acidithiobacillales;100%; Acidithiobacillaceae;100%; Acidithiobacillus;100%
OTU3	95	2	nekultivovaná bakterie	100	Acidobacteria;100%; Acidobacteria_Gp1;100%; Granulicella;83%
OTU4	103	12	Acidiphilium	99	Proteobacteria;100%; Alphaproteobacteria;100%; Rhodospirillales;100%; Acetobacteraceae;100%; Acidiphilium;75%
OTU5	12	0	Acidobacteria	99	Acidobacteria;100%; Acidobacteria_Gp1;100%; Gp1;81%
OTU6	44	6	Acidocella	100	Proteobacteria;100%; Alphaproteobacteria;100%; Rhodospirillales;100%; Acetobacteraceae;100%; Acidocella;92%
OTU7	36	4	Alicyclobacillus	98	Firmicutes;83%; Bacilli;70%; Bacillales;62%; Bacillaceae 1;9%; Domibacillus;8%
OTU8	4	46	Leptospirillum	100	Nitrospirae;100%; Nitrospira;100%; Nitrospirales;100%; Nitrospiraceae;100%; Leptospirillum;100%
OTU9	3	0	Sphingomonas	99	Proteobacteria;100%; Alphaproteobacteria;100%; Sphingomonadales;100%; Sphingomonadaceae;100%; Sphingomonas;97%
OTU10	28	8	Thiomonas	100	Proteobacteria;100%; Betaproteobacteria;100%; Burkholderiales;100%; Burkholderiales_incertae_sedis;100%; Thiomonas;100%
OTU11	1020	3420	nekultivovaná bakterie	100	Proteobacteria;100%; Betaproteobacteria;99%; Gallionellales;72%; Gallionellaceae;72%; Sideroxydans;72%
OTU12	107	46	NA	100	Proteobacteria;100%; Gammaproteobacteria;85%; Xanthomonadales;85%; Xanthomonadaceae;85%; Dokdonella;52%
OTU13	65	0	NA	99	Acidobacteria;100%; Acidobacteria_Gp1;100%; Terriglobus;76%

Podle počtu mapovaných MiSeq-RNA čtení na reprezentativní sekvence jednotlivých OTU jsme byli schopni odhalit aktivní členy společenstva. Je patrné, že nejvíc čtení bylo namapováno na referenční sekvenci OTU1 asociovanou s bakterií „*Ferrovum myxofaciens*“ ZH7. Tato čtení počtem řádově převyšují RNA sekvence „*Acidithiobacillus*“ linie ZH7B a třetí bakterie, jejichž míra zastoupení je obdobná (tabulka 4.2). Ačkoli v amplikonových datech reprezentovaný jen 4 sekvencemi (OTU8), aktivním členem se zdá být i zástupce rodu *Leptospirillum*. Výskyt rodu *Leptospirillum* byl dříve rovněž zaznamenán v kyselých důlních vodách (Schrenk et al., 1998; Ferrer et al., 2016). Je důležité upozornit, že ribosomální RNA byla z izolované celkové RNA laboratorně odstraňována a tato metoda mohla specificky preferovat některé bakterie.

4.2.2 Sestavení sekvencí

Přehled získaných čtení z jednotlivých typů knihoven nabízí tabulka (tabulka 4.3). Prvotním sestavením 454 dat v programu Newbler jsme získali 272 skafoldů obsahujících 1330 mezer. S použitím párových MiSeq-DNA a MiSeq-RNA čtení jsme programem GapCloser byli schopni zaplnit 514 mezer. Výsledná verze skafoldů je z hlediska této práce považována za finální (tabulka 4.4) a sloužila pro všechny navazující analýzy.

Tabulka 4.3: Přehled získaných sekvencí podle typu připravených knihoven.

Knihovna	Délka čtení (bp)	Počet čtení	Počet bází (bp)
454 amplikony	359	5 191	1 863 569
454 shotgun	379	882 567	334 720 625
454 3 kb párová	349	1 098 719	383 050 189
454 8 kb párová	340	437 409	148 564 464
MiSeq DNA	2x250	9 632 944	4 835 737 888
MiSeq RNA	2x150	19 595 715	5 917 905 930

Tabulka 4.4: Porovnání výsledků sestavení genomových sekvencí. Celkový počet bazí a délka skafoldů jsou uváděny bez délky mezer.

	Newbler (454 data)	Newbler + GapCloser	MaSuRCA (454 + MiSeq-DNA data)
Počet skafoldů	272	272	17 492
Počet mezer	1 330	514	2 708
Celkový počet bazí (bp)	8 822 016	9 887 094	34 204 767
Celková délka mezer (bp)	1 519 219	370 996	2 463 603
Nejdelší skafold (bp)	763 989	760 723	67 437
Nejkratší skafold (bp)	2 009	2 009	300
Průměrná délka skafoldů (bp)	38 019	37 714	2 096
Medián délky skafoldů (bp)	4 335	4 335	902
N50 (bp)	137 633	140 416	4 797

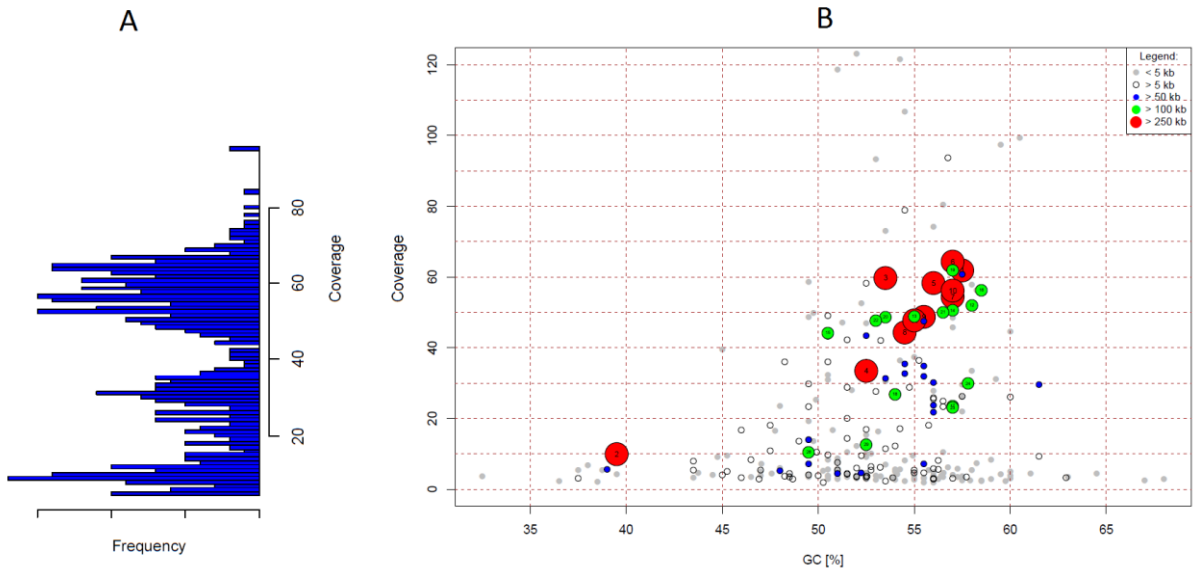
Pro kombinované sestavení genomových sekvencí z obou použitých sekvenátorů jsme zvolili program MaSuRCA na základě studie porovnávající výsledky assembly několika bakteriálních genomů pomocí osmi různých programů (Magoc et al., 2013). Z assembly našich dat vyplývá, že MaSuRCA sestavila řádově větší celkový počet bazí do skafoldů (tabulka 4.4), což odpovídá výrazně většímu množství použitých vstupních dat a díky tomu pravděpodobně více postihuje také genomy méně četných zástupců společenstva. Ovšem s ohledem na počet a délku skafoldů považujeme za lepší výsledné assembly z programu Newbler na základě pouze 454 dat, což dokládá, že software je velmi dobře uzpůsobený právě pro sestavení „nativních" 454 čtení. Především velmi dobře pracuje s párovými daty při tvorbě dlouhých skafoldů. Lze očekávat, že průběh assembly všech genomových sekvencí programem MaSuRCA bude ovlivněn právě jiným charakterem dat, převážně odlišnou chybovostí. Zajímavé ovšem je, že lepších výsledků, co se týče celkového počtu a délky skafoldů, nebylo dosaženo ani v případě, kdy jsme programem MaSuRCA sestavili 454 čtení a MiSeq čtení zvlášť. Rovněž v souhrnu výsledků nemělo pozitivní vliv na assembly ani prvotní filtrování dat z genomu bakterie *Acidithiobacillus* na základě mapování čtení na tři dostupné referenční genomy a následná tvorba dvou assembly: (1) pouze ze čtení mapovaných na genomy *Acidithiobacillus* a (2) filtrovaná čtení bez sekvencí *Acidithiobacillus* (neprezentovaná data).

4.2.3 Roztřídění sestavených sekvencí

Roztřídění (*binning*) kontigů jsme provedli na základě analýzy kompozice sekvencí, výsledných parametrů assembly a anotace predikovaných genů. Všechny tyto metody poskytují indicie, které jsme kombinovaně hodnotili při finálním manuálním roztřídění kontigů.

Průměrnou míru pokrytí v rámci složených sekvencí ukazuje obrázek 4.2a, obrázek 4.2b vizualizuje vztah míry pokrytí a obsahu GC párů v kontigových sekvencích. Výsledek korespondenční analýzy preferovaných dikodonů je prezentován na obrázku 4.3a. Analýza párových čtení propojujících kontigy vedla k tvorbě sítí, z nichž dvě hlavní spojují dohromady asi tři čtvrtiny sestavených kontigů (obrázek 4.3b).

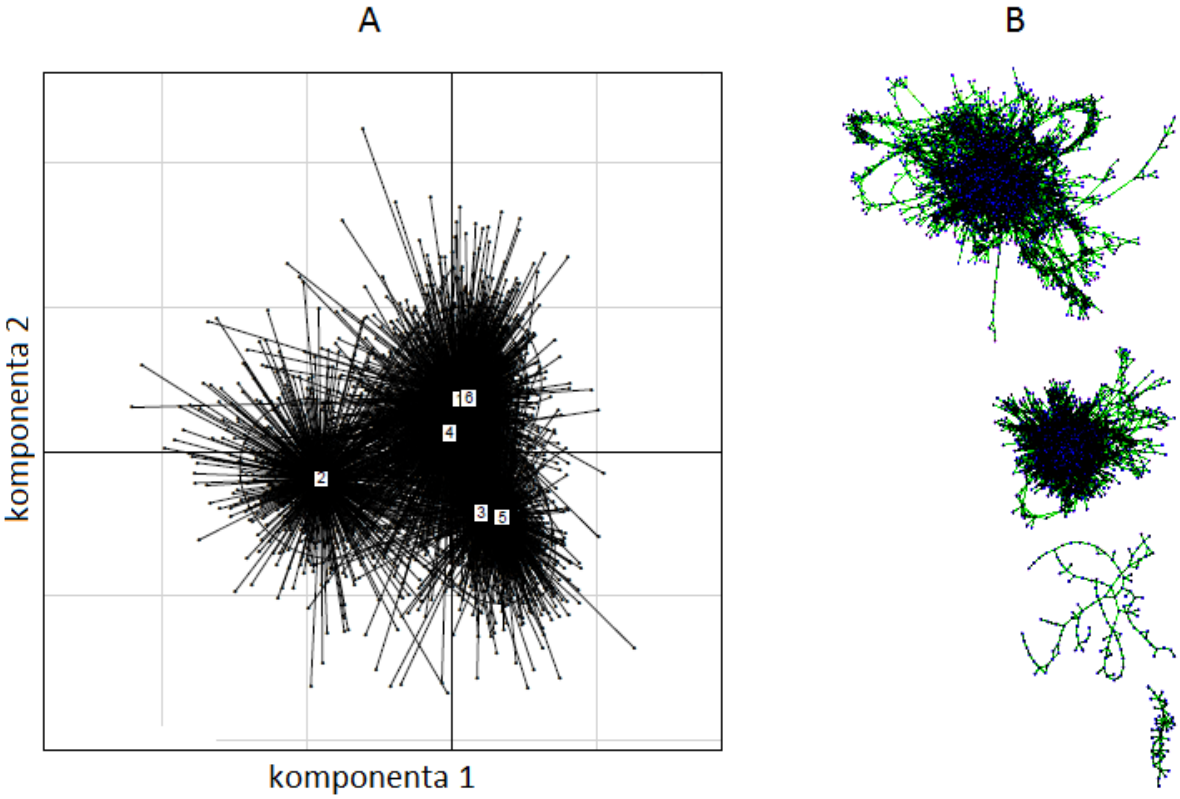
Vzhledem ke zvoleným referenčním sekvencím umožnilo hodnocení sekvenční signatury programem ClaMS predikovat rozdělení kontigů do třech skupin: *Acidithiobacillus, Ferrovum* a nezařazené sekvence. Na základě výsledků 16S rRNA amplikonového sekvenování pak očekáváme přítomnost bakterie třídy Gammaproteobacteria (*Acidithiobacillus*) a dvou linií ze třídy Betaproteobacteria, z nichž jedna se zdá být nejvíce zastoupena (*Ferrovum*).



Obrázek 4.2: Analýza míry pokrytí a obsahu GC párů v sestavených kontigových sekvencích. A) Historgam míry pokrytí (*coverage*), B) vztah míry pokrytí (*coverage*) a obsahu GC párů, zvýrazněny velikost kruhů odpovídá délce kontigů (viz

Manuálním hodnocením výstupů všech popsaných strategií a jejich kombinací s hodnocením taxonomické příslušnosti anotovaných CDS jsme docílili třech množin dat. První množina nejvíce zastoupená v datech a vykazující příslušnost ke třídě Betaproteobacteria byla přiřazena genomu „*Ferrovum myxofaciens*“ ZH7. Druhá početná skupina kontigů asociovaných

se třídou Gammaproteobacteria představuje genom „*Acidithiobacillus*“ linie ZH7B. Do třetí skupiny s relativně nízkým pokrytím spadají všechny ostatní složené metagenomické sekvence, z části obsahují zřejmě genom neznámé proteobakterie popsané na základě 16S rRNA amplikonového sekvenování, ale rovněž úseky genetické výbavy všech dalších málo zastoupených linií. Přehled roztříděných skafoldů a jejich charakteristiky nabízí tabulka 4.5.



Obrázek 4.3: Anlýza preferovaných dikodonů a vizualizace sítí vytvořených na základě párových čtení. A) výsledek korespondenční analýzy na základě matice sdílených preferovaných dikodonů mezi sestavenými kontigy. B) Výsledné sítě vzniklé spojením kontigů podle konců párových čtení vizualizované v programu Cytoscape, zahrnuty jsou jen sítě tvořené alespoň 10 kontigy.

Tabulka 4.5: Výsledné roztřídění kontigů do skupin a jejich charakteristiky

Sk.	Organismus	bp	počet skafoldů	počet kontigů	GC (%)	Průměrné pokrytí	počet CDS
1	*Ferrovum myxofaciens* ZH7	3 261 464	59	186	53	84	3 198
2	*Acidithiobacillus* spp. ZH7B	3 556 813	39	121	56	81	3 692
3	mix	3 068 817	174	479	48	12	3 975

4.2.4 „*Ferrovum myxofaciens*“ ZH7 a „*Acidithiobacillus*“ linie ZH7B

Podařilo se nám rekonstruovat téměř kompletní, draftové sekvence bakterií „*Ferrovum myxofaciens*“ ZH7 a „*Acidithiobacillus*“ linie ZH7B metagenomickým sekvenováním smíšeného společenstva. V průběhu zpracovávání výsledků této studie byly postupně

uveřejněny dalšími výzkumnými skupinami kompletní genomové sekvence dvou bakterií *At. ferrooxidans* () a jedné *At. ferrivorans* (). Vzhledem k obtížné kultivaci bakterií rodu *Ferrovum* (Johnson et al., 2014) byly prozatím publikovány pouze nekompletní, draftové sekvence typové linie P3G (Moya-Beltrán et al., 2014), bakterie *Ferrovum* FB7 z metagenomických sekvenačních dat (Hua et al., 2015) a linie JA12 také z jednoduché komunity (Ullrich et al., 2016). To umožňuje zaměřit se na komparativní analýzu s námi stanovenými draftovými sekvencemi, které jsou zároveň výrazným přispěním do balíku současných omezených genomových dat. V tomto kontextu můžeme na základě genomových sekvencí a analýzy transkribovaných genů popsat funkční procesy bakterií „*Ferrovum myxofaciens*“ ZH7 a „*Acidithiobacillus*“ linie ZH7B uplatňující se v krápníkovém biofilmu.

4.2.4.1 Zdroj uhlíku

Obě námi popisované bakterie používají jako zdroj uhlíku fixaci CO_2 prostřednictvím Calvinova-Bensonova-Basshamova cyklu, což odpovídá popsaným bakteriím z prostředí kyselých důlních vod (Moya-Beltrán et al., 2014; Ullrich et al., 2016). Oba genomy obsahují kompletní klastry genů kódující karboxylázu/oxygenázu RuBisCO, přičemž „*Ferrovum myxofaciens*“ ZH7 obsahuje dva operony oba kódující RuBisCO izoformu I, zatímco „*Acidithiobacillus*“ linie ZH7B má navíc izoformu II, což odpovídá výskytu těchto enzymů v dostupných genomech rodů *Ferrovum* a *Acidithiobacillus* (Ullrich et al., 2016). Obě bakterie navíc mají i klastr genů pro formování struktury karboxysomu, který se podílí na zvýšení efektivity fixace CO2 enzymem RuBisCO (Cannon et al., 2001).

4.2.4.2 Fixace dusíku

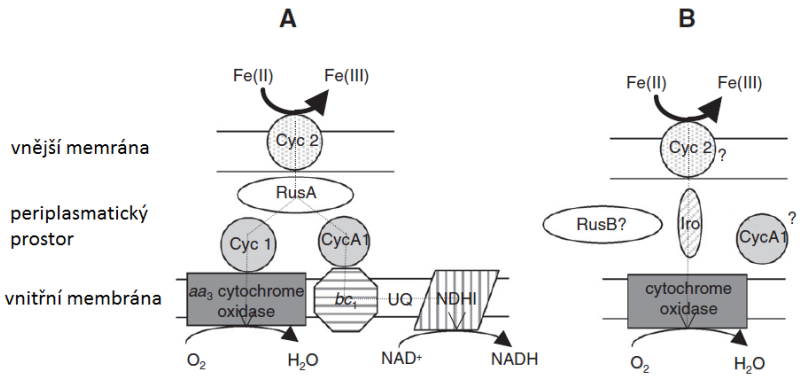
„*Ferrovum myxofaciens*“ ZH7 a „*Acidithiobacillus*“ linie ZH7B mají repertoár genů pro fixaci molekulárního dusíku a rovněž pro využití dusičnanů, dusitanů i amoniaku jako zdrojů dusíku.

Zajímavý je ovšem výskyt kompletního klastru genů *ureABCDEFG* v námi rekonstruovaném genomu „*Acidithiobacillus*“ linie ZH7B, který kóduje dráhu pro syntézu CO_2 a NH_3 z močoviny. Klastr pro metabolismus močoviny byl dříve popsán u dvou bakterií rodu *Ferrovum* na základě draftových sekvencí genomů FKB7 (Hua et al., 2015) a JA12 (Ullrich et al., 2016). Zatím však nikdy nebyl zaznamenán u bakterií rodu *Acidithiobacillus* ani u dalších acidofilních druhů oxidujících železo ze tříd Betaproteobacteria a Gammaproteobacteria. Právě zdroje dusíku mohou být v daném ekosystému limitujícím faktorem (tabulka 4.1) a využití

alternativního zdroje u jedné z dominantně zastoupené bakterie umožňuje soužití v rámci společenstva.

4.2.4.3 Zdroje energie

Na základě vzniku vysráženého minerálu schwermannitu (Fe^{3+} hydroxysíranu), který prostupuje bakteriální buňky uvnitř krápníkových struktur biofilmu (Falteisek & Čepička, 2012), lze usuzovat na probíhající aerobní oxidaci dvojmocného železa Fe^{2+} (Mori et al., 2016), které slouží jako donor elektronu a zdroj energie pro bakteriální členy společenstva (Croal et al., 2004). Modelová dráha aerobní oxidace Fe^{2+} na Fe^{3+} byla poměrně intenzivně studována a detailně popsána u druhu *At. ferrooxidans* (Appia-Ayme et al., 1999; Quatrini et al., 2009). Využívá proteinu cytochrom c (Cyc2) ve vnější membráně buněk k oxidaci Fe^{2+}; k přenosu elektronu dále slouží malý periplasmatický protein rusticyanin. Akceptorem elektronu je buď O_2 redukovaný cytochrom c oxidázou (aa_3 komplex), kdy se pro přenos elektronu uplatňuje další periplasmatický protein kódovaný genem Cyc1, nebo je elektron přenesen prostřednictvím cytochromu c4 kódovaného genem CycA1 přes bc1 komplex k NADH:chinon oxidoreduktáze, kde dochází k redukci NAD+ (obrázek 4.4a) (Ilbert & Bonnefoy, 2013).



Obrázek 4.4: Znázornění dráhy pro oxidaci redukovaného železa. A) *At. ferrooxidans*. B) „*Acidithiobacillus*“ linie ZH7B (převzato a upraveno z Ilbert & Bonnefoy, 2013).

U *At. ferrivorans* nebyl popsán funkční rusticyanin (RusA), ale jeho modifikovaný homolog rusticyanin B (RusB), jehož funkce není známá a není jisté, zda se uplatňuje v oxidaci Fe^{2+} (Hallberg et al., 2010). Zároveň však bakterie *At. ferrivorans* mají gen Iro pro oxidázu železa, která by mohla hrát roli v alternativní dráze oxidace Fe^{2+} (Amouric et al., 2011). V genomu „*Acidithiobacillus*“ linie ZH7B jsme identifikovali gen kódující rusticyanin B, který vykazuje naprostou identitu na úrovni aminokyselinové sekvence ke genu rusB u bakterií *At. ferrivorans* (Hallberg et al., 2010). Rovněž genom obsahuje gen Iro s aminokyselinovou

sekvencí identickou ke genům popsaným u *At. ferrivorans* (Amouric et al., 2011). Výskyt obou genů jsme potvrdili také v našich RNA-MiSeq datech sestavených assemblerem Trinity do kontigů odpovídajících transkriptům. To dokládá jejich expresi a je velmi pravděpodobné, že některý z nich či dokonce oba zároveň se uplatňují v alternativní dráze oxidace železa. Genom „*Acidithiobacillus*“ linie ZH7B obsahuje dále gen podobný Cyc2 (82% podobnost na úrovni aminokyselinové sekvence) a geny vykazující 72% a 84% podobnost ke genu Cyc1 resp. CycA1 v genomu *At. ferrooxidans* ATCC 23270. Expresi všech tří genů jsme ověřili opět prohledáním sestavených transkriptů. Ač přesný mechanismus není znám, domníváme se, že i tyto geny se uplatňují v oxidaci železa u námi popisované bakterie „*Acidithiobacillus*“ linie ZH7B a navrhovaná dráha je obdobou systému bakterie *At. ferrooxidans* (obrázek 4.4b).

Mechanismus oxidace železa u bakterií rodu *Ferrovum* je prozatím prostudován velmi málo. Předchozí studie navrhovaly podobnou dráhu jako u *At. ferrivorans* na základě výskytu predikovaných genů s podobností aminokyselinových sekvencí kolem 30 % ke genům Cyc2 a Iro (Hua et al., 2015; Ullrich et al., 2016). V genomu bakterie „*Ferrovum myxofaciens*“ ZH7 jsme vyhledali CDS podobné ke genům, které se uplatňují v oxidační dráze u *At. ferrooxidans*/*ferrivorans*: Cyc2, Cyc1, CycA1, rusticyanin a Iro. Rovněž jsme v genomu vyhledali případné podobnosti ke genům, které byly popsány jako součást dráhy oxidace železa u dalších organismů: mtr geny bakterie *Shewanella putrefaciens* (Beliaev & Saffarini, 1998) a Pio operon bakterie *Rhodopseudomonas palustris* TIE-1 (Jiao & Newman, 2007). Pro gen rusticyanin ani pro mtr a Pio geny jsme nenašli podobné úseky v genomu „*Ferrovum myxofaciens*“ ZH7 (*E-value* < 10^{-2}). Pro geny Cyc2, Cyc1, CycA1 (*At. ferrooxidans* ATCC 23270) a Iro (*At. ferrivorans* SS3) jsme nalezli pravděpodobné homologní geny s podobností aminokyselinových sekvencí 28 %, 35 %, 44 % resp. 32 %. To odpovídá rovněž rekonstruovaným genomům bakterií *Ferrovum* linie FKB7 (Hua et al., 2015) a JA12 (Ullrich et al., 2016). Tyto geny se opět vyskytují i mezi sestavenými transkripty „*Ferrovum myxofaciens*“ ZH7. Pro oxidaci Fe^{2+} na vnější buněčné membráně a přenos elektronu na úrovni periplasmatického prostoru využívají bakterie rodu *Ferrovum* pravděpodobně obdobnou dráhu jako *At. ferrooxidans*/*ferrivorans*.

Kromě oxidace redukovaného železa byla u bakterie *At. ferrooxidans* popsána také dráha pro oxidaci redukované anorganické síry jako zdroje elektronu a energie (Quatrini et al., 2009). Zatímco námi sestavený genom bakterie „*Acidithiobacillus*“ linie ZH7B obsahuje geny, které se uplatní při oxidaci síry, druhá bakterie „*Ferrovum myxofaciens*“ ZH7 tyto geny postrádá. Nebyly nalezeny ani v publikované sekvenci genomu *Ferrovum* linie JA12 (Ullrich

et al., 2016) a kultivovaná bakterie *Ferrovum myxofaciens* P3G nebyla schopna využívat síru jako donor elektronu (Johnson et al., 2014). To představuje další rozdělení nik dvou dominantně zastoupených bakterií v námi studovaném konsorciu.

4.2.4.4 Bakteriální bičík a schopnost pohybu

Genom „*Ferrovum myxofaciens*“ ZH7 obsahuje celou sadu genů pro tvorbu struktury bakteriálního bičíku, kterou můžeme nalézt také v genomu typové linie P3G, nikoli však bakterie *Ferrovum* linie JA12, jejíž genom byl ostatně popsán jako redukovaný ve více ohledech (Ullrich et al., 2016). Bakterie „*Acidithiobacillus*“ linie ZH7B geny pro strukturu bičíku postrádá. Zajímavostí je, že jako jediný nepohyblivý zástupce rodu *Acidithiobacillus* byl prozatím popsán *At. ferrooxidans,* zatímco například *At. ferrivorans* se od něho odlišuje rychlým pohybem (Hallberg et al., 2009).

4.2.4.5 Bakteriální imunitní systém

Oba rekonstruované genomy obsahují geny pro CRISPR/Cas klastr, který představuje systém bakteriální imunitní odpovědi proti exogenním genetickým elementům, například bakteriofágům (souhrnně např. Horvath & Barrangou, 2010). Tento systém není kódovaný genomy bakterií rodu *Ferrovum* – P3G ani JA12; mezi zástupci rodu *Acidithiobacillus* jsme ho objevili pouze v dostupné genomové sekvenci *At. ferrooxidans* ATCC 23270. Bakterie „*Ferrovum myxofaciens*“ ZH7 a „*Acidithiobacillus*“ linie ZH7B zřejmě získaly systém imunitní odpovědi jako adaptaci na prostředí, což by naznačovalo, že v daném ekosystému jsou vystaveny tlaku bakteriofágových infekcí.

4.2.4.6 Bakteriální kapsule, exopolysacharidy a formování biofilmu

Bakterie „*Ferrovum myxofaciens*“ ZH7 je schopna tvorby vnější ochranné vrstvy díky klastrům genů kps a wbc pro formování kapsule. Kapsule poskytuje další obranný mechanismus, zároveň se však uplatňuje v adhezi a formování biofilmu (Schembri et al., 2004).

Oba námi rekonstruované genomy dále obsahují celou řadu genů zapojených do produkce exopolysacharidů, v důsledku čehož vzniká charakteristický řetězcový makroskopický biofilm.

4.2.5 Vysoce exprimované geny

Na základě stanovení relativní transkripční aktivity (RTA) jsme identifikovali upregulované geny u členů zkoumaného společenstva. V rámci této práce se následně zaměřme na deset nejvíce exprimovaných genů z jednotlivých rekonstruovaných genomů.

4.2.5.1 „*Ferrovum myxofaciens*“ ZH7

Mezi deseti geny s nejvyšší mírou exprese je na prvním místě metalloproteáza FtsH, která je konzervovaná u bakterií a je zodpovědná za kontrolu a případné odbourání nepotřebných či poškozených proteinů (Langklotz et al., 2012). Dále chaperon (GroEL), který v buňkách udržuje správné sbalování proteinů. Oba zmíněné geny představují zástupce řady adaptací na extrémní podmínky důlního prostředí, kdy společným vlivem nízké hodnoty pH a oxidačním stresem hrozí nestabilita buněčných struktur a poškození DNA a proteinů (Hua et al., 2015; Ullrich et al., 2016).

Vysoce exprimované jsou také dva geny pro dvě podjednotky cytochrom c oxidázy cbb_3 komplexu. Ten slouží v dráze oxidace Fe^{2+} pro redukci O_2 na vnitřní buněčné membráně například u neutrofilních Fe^{2+} oxidujících bakterií *Mariprofundus ferrooxidans* PV-1 a *Sideroxydans lithotroficus* ES-1 (Ullrich et al., 2016). Využití cbb3 komplexu jako alternativy k aa3 oxidáze *At. ferrooxidans* bylo již dříve uvažováno u bakterií *Ferrovum myxofaciens* P3G a *Ferrovum* linie JA12 na základě draftových genomových sekvencí (Moya-Beltrán et al., 2014; Ullrich et al., 2016). Mezi nejvíce exprimovanými geny jsou také potencionální homology k CycA1 a Iro, což podporuje hypotézu jejich zapojení do dráhy oxidace železa na úrovni vnější membrány či periplasmatického prostoru, jak bylo diskutováno výše.

Další z deseti nejvíce exprimovaných CDS je gen pro flagelin. Přestože u bakteriálních buněk účastnících se formování nárůstu biofilmu lze předpokládat mechanismus regulace vzniku struktur pro pohyb (Guttenplan & Kearns, 2013), vysoká míra transkripce flagelin genu naznačuje aktivní pohyb bakterií „*Ferrovum myxofaciens*“ ZH7 ze zkoumaného vzorku biofilmu.

Posledním třem z desítky nejvíce exprimovaných CDS bakterie „*Ferrovum myxofaciens*“ ZH7 byl na základě anotace přiřazen pouze status hypoteticky predikovaných genů.

4.2.5.2 „*Acidithiobacillus*“ linie ZH7B

Nejvíce transkribované CDS bakterie „*Acidithiobacillus*“ linie ZH7B spadají rovněž do skupiny genů podílejících se na rezistenci vůči nízkému pH a oxidačnímu stresu. Spadá sem konkrétně chaperon DnaK zodpovědný za údržbu proteinů a DNA. Funkci při opravách DNA lze usuzovat také u upregulovaného genu pro helikázu nadrodiny II (Byrd & Raney, 2012) a identifikovaná ATP-dependentní Lon proteáza je zodpovědná za degradaci poškozených proteinů jako odpovědi na stress (Lee & Suzuki, 2008). Zřejmě sem můžeme zařadit i gen ParB podílející se na rovnoměrné segregaci například plasmidových sekvencí při dělení bakteriálních buněk (Hayes, 2000). Upregulovaný je také membránový porin popsaný u *At. ferrooxidans* zabraňující volnému pohybu protonů přes vnější membránu v kyselém prostředí (Guiliani & Jerez, 2000).

Dalším z desítky genů s vysokou expresí je NADH:chinon oxidoreduktáza, důležitý zástupce dráhy oxidace redukovaného železa. Zprostředkovává redukci NAD+, který slouží jako finální akceptor elektronu (viz obrázek 4.4). Velmi exprimovaný je také další důležitý protein téže dráhy – homolog Cyc2 diskutovaný výše.

Vysoce transkribovaný je také gen pro tvorbu lipoproteinů, uplatňující se v tvorbě biofilmu a dále gen kódující intron skupiny II, mobilní retroelement schopný vlastního vystřižení (souhrnně viz např. Martinéz-Abarca & Toro, 2000). Pro jeden z deseti nejvíce aktivních genů bakterie „*Acidithiobacillus*“ linie ZH7B nebyla nalezena podobnost s geny v databázi NCBI-NR.

4.2.5.3 Nejvíce transkribované geny ostatních členů společenstva

Míru relativní transkripční aktivity jsme stanovili také pro predikované geny, které se nacházejí v sekvencích kontigů nepřiřazených žádné ze dvou dominantních bakterií (skupina kontigů 3). Zajímavé je, že mezi deseti nejvíce transkribovanými geny bylo 8, které vykazovaly shodu pouze s hypoteticky predikovanými geny v databázi NCBI-NT, nebo je nelze anotovat vůbec (neexistují podobné sekvence v NCBI-NT). To vypovídá o výskytu zcela neznámých bakterií a funkčních procesů, které jsou v daném vzorku aktivní. Jeden z genů, které se podařilo anotovat, se uplatňuje při tvorbě bakteriální vnější ochranné kapsule, druhý kóduje mobilní element – intron skupiny II.

4.3 Závěry

Biofilm utvářející makroskopické stalaktity na stropě šachty bývalého dolu ve Zlatých Horách je tvořen taxonomicky velmi jednoduchou bakteriální komunitou. Na základě sekvenování celkové DNA, 16S rRNA amplikonů a RNA jsme odhalili složení konsorcia, funkční potenciál přítomných genomů a aktivně užívané geny s vysokou mírou transkripce. Z dat vyplývá, že konsorcium je tvořeno dvěma dominantně zastoupenými druhy, relativně početná je také třetí dosud nepopsaná bakterie a minoritně se vyskytuje ještě několik dalších bakterií.

Z metagenomických sekvencí získaných bez nutnosti předchozí kultivace jednotlivých bakterií jsme byli schopni sestavit téměř kompletní draftové genomové sekvence dvou majoritně zastoupených bakterií: „*Ferrovum myxofaciens*“ ZH7 a „*Acidithiobacillus*“ linie ZH7B. Na jejich základě jsme odhalili kódované predispozice bakterií.

Bakterie jsou plně vybaveny pro život v daných podmínkách, obě jsou autotrofní a jejich metabolické dráhy jsou zcela oddělené a nepředstavují žádnou formu syntrofie. Obě získávají uhlík fixací CO_2 a energii oxidací železa. „*Acidithiobacillus*“ linie ZH7B je navíc schopen oxidovat redukovanou síru a jako zdroj dusíku dokáže metabolizovat močovinu, což umožňuje dvěma popsaným bakteriím využívat z části oddělené niky.

V porovnání s dalšími zástupci rodu *Ferrovum* je námi popsaná bakterie „*Ferrovum myxofaciens*“ ZH7 schopna fixovat molekulární dusík, je vybavena bakteriálním imunitním systémem CRISPR/Cas a její buňky jsou „obrněny“ kapsulí, což odráží vlastnosti daného ekosystému.

Rekonstruovaný genom „*Acidithiobacillus*“ linie ZH7B se vyznačuje směsicí znaků vyskytujících se u druhů *At. ferrooxidans* a *At. ferrivorans*. 16S rRNA gen vykazuje největší míru podobnosti s 16S rRNA sekvencí genu *At. ferrooxidans*. „*Acidithiobacillus*“ linie ZH7B však využívá geny pro rusB a Iro (pravděpodobně zapojené do dráhy oxidace železa), které byly nalezeny u linií druhu *At. ferrivorans* (Hallberg et al., 2010; Amouric et al., 2011). Zajímavé také je, že *At. ferrivorans* byl popsán jako vysoce pohyblivá bakterie (Hallberg et al., 2010), zatímco rekonstruovaný genom „*Acidithiobacillus*“ linie ZH7B geny pro strukturu bakteriálního bičíku postrádá. Obsahuje však geny pro CRISPR/Cas systém, které v rámci rodu *Acidithiobacillus* lze nalézt pouze v genomu *At. ferrooxidans* ATCC 23270. Tato data dokládají variabilitu linií uvnitř rodu *Acidithiobacillus*. Námi popsaná bakterie představuje unikátní linii,

která se vyvinula prostřednictvím specifických adaptací v konkrétním prostředí kyselé důlní vody ve Zlatých Horách.

5 Metagenomické profilování komplexní komunity půdních bakterií z kontaminované zeminy a analýza vlivu rostlin na jejich složení a funkční potenciál

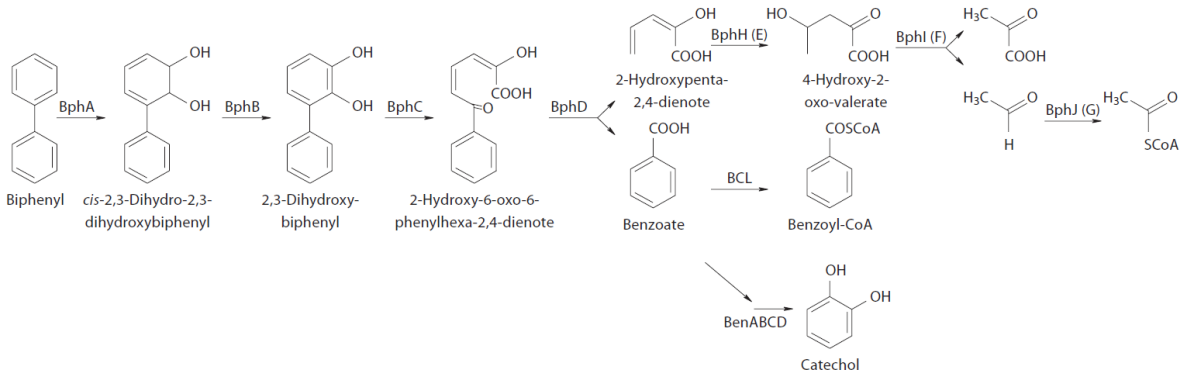
Půda představuje ekosystém s enormní diverzitou mikroorganismů, často uspořádaný do mikroskopických struktur. Jeden gram obvykle obsahuje stovky druhů čítajících miliardy jednotlivých buněk. Tenká vrstva v těsném okolí kořenů rostlin nazývaná rhizosféra je obzvláště aktivní prostředí ovlivněné aktivitou rostlin a přítomností specifických organických sloučenin (Chaudhry et al., 2005; Hartman et al., 2009; Philippot et al., 2013). Molekuly vylučované z kořenů rostlin zahrnují cukry, aminokyseliny, organické kyseliny, mastné kyseliny, steroly, růstové hormony, enzymy, flavonoidy, nukleotidy, rostlinné hormony, alkoholy, alkylsulfáty, anorganické ionty a plynné molekuly (přehledně viz Dennis et al., 2010). Fyzikální a chemické vlastnosti rhizosféry navíc ovlivňuje také samotná růstová aktivita kořenů. To vše v důsledku ovlivňuje přítomná mikrobiální společenstva a často nacházíme větší denzitu mikrobiálních buněk ve vrstvě rhizosféry (Dennis et al., 2010).

V půdě kontaminované důsledkem lidských aktivit hraje rhizosféra obzvláště důležitou roli (Macek et al., 2009; Macková et al., 2006; Macková et al., 2010; Kurzawová et al., 2012). Uhlíkaté sloučeniny produkované kořeny rostlin podporují kometabolismus řady organických polutantů, například flavonoidů (Pham et al., 2012; Toussaint et al., 2012; Pham et al., 2015), a rovněž slouží jako zdroj uhlíku a energie pro mikroorganismy (Leigh et al., 2006). V rhizosféře je obvykle vyšší obsah kyslíku nezbytný pro fungování oxygenáz, enzymů uplatňujících se v biodegradaci řady polutantů (Leigh et al., 2002). Rostliny také produkují biosurfaktanty, které zvyšují dostupnost těžko rozpustných kontaminantů pro mikrobiální metabolismus (Read et al., 2003). Kořeny navíc hrají roli v pohybu molekul a hromadění polutantů v oblasti rhizosféry, která se tím více stává aktivním místem biodegradace (Liste & Alexander, 2000; Yi & Crowley, 2007).

Polychlorované bifenyly (PCB) představují značné riziko jak pro fungování ekosystémů, tak i pro lidské zdraví. Přesto, že jejich výroba byla zastavena již před více než třiceti lety (v bývalém Československu v roce 1984), patří stále mezi velmi rozšířené polutanty ve všech složkách životního prostředí díky své vysoké stabilitě. PCB jsou deriváty bifenylu, jehož dvě jednoduchou vazbou spojená benzenová jádra nesou jeden až deset atomů chloru.

Jejich počet a poloha umožňuje vznik až 209 izomerů (kongenerů) číslovaných podle nomenklatury Ballschmitera a Zella (Ballschmiter & Zell, 1980). Jsou velmi stálé i za vysokých teplot, nehořlavé, odolné vůči kyselinám, zásadám a dalším chemickým látkám, nerozpustné ve vodě a naopak rozpustné v tucích a organických rozpouštědlech. Původně byly PCB směsi vyráběné jako náplně kondenzátorů, dále do hydraulických a setrvačníkových zařízení, ale díky svým vlastnostem dospěly k široké škále využití také jako přísady do barev, lepidel či dokonce rtěnek s podobně.

Ačkoli jsou PCB toxické pro řadu vyšších organismů, některé mikroorganismy jsou schopné je metabolicky transformovat. V aerobních podmínkách jsou například bakterie schopné využívat k metabolismu bifenyl a jeho deriváty s jedním atomem chloru (Sylvestre, 1980), vzácněji také se dvěma atomy (Kim & Pricardal, 2001) jako zdroj uhlíku a energie. Další kongenery s nižším počtem atomů chloru mohou být biodegradovány kometabolicky (neslouží přímo jako zdroj energie) společně s bifenylem jako hlavním substrátem (Furukawa, 2000). Degradační enzymy jsou u bakteriálních genomů kódovány *bph* operonem, degradační dráha PCB je znázorněna na obrázku 5.1.



Obrázek 5.1: Degradační dráha polychlorovaných bifenylů (převzato z Pieper & Seeger, 2008).

Jako zajímavý způsob sanace lokalit znečištěných PCB se ukázalo využití přirozeného bioremediačního potenciálu některých rostlin. Předchozí studie dokládaly vliv výsadby křenu (*Armoracia rusticana*), lilku (*Solanum nigrum*) a tabáku (*Nicotiana tobacum*) na snížení obsahu polychlorovaných bifenylů v kontaminované zemině (Ionescu et al., 2009; Kurzawová et al., 2012). Co však zůstává otázkou, jsou samotné mechanismy vlivu rostlin. Tedy zda fungují pouze skrze vlastní absorpci a transformaci bifenylů (viz například Rezek et al., 2008), nebo se zásadním způsobem podílejí na utváření a podporování mikrobiálních společenstev schopných biodegradace PCB v půdě. S využitím masivního pyrosekvenování jsme se proto rozhodli

zaměřit se na otázky (1) do jaké míry ovlivňuje přítomnost rostlin složení bakteriálních konsorcií v zemině, (2) jak rozdílné složení mají společenstva ovlivněná různými rostlinami, (3) jakou roli hraje NPK minerální hnojivo běžně používané pro podporu růstu daných rostlin a (4) jaké důsledky mají výše zmiňované vlivy na funkční potenciál mikrobiálních společenstev v kontaminované zemině.

5.1 Materiál a metody

5.1.1 Mikroprostředí vzorků vytvořené v laboratoři

Vzorky zeminy byly odebrány ze skládky kontaminované půdy u obce Lhenice v Jižních Čechách, která byla převezena z areálu asfaltového závodu z Milevska a uložena v roce 1996. Kontaminace probíhala od roku 1960 do roku 1990 postupným unikáním teplonosného media Delotherm obsahujícího převážně směs PCB Delor 103, které vyráběl chemický kombinát Chemko Strážné na Slovensku. Agrochemická analýza a stanovení obsahu polutantů byly již dříve publikovány (Pavlíková et al., 2007; Uhlík et al., 2012; Stella et al., 2015). Krátce shrnuto, zemina je písčitohlinitá s poměrem přibližně 60 % písek, 30 % silt a 10 % jíl. Obsah indikátorových kongenerů PCB (28, 52, 101, 118, 138, 153, 180 podle Ballschmiter & Zell nomenklatury) byl dříve stanoven přibližně na 100 μg/gram zeminy (Uhlík et al., 2012), zatímco detailnější analýza kongenerů ukazuje na celkovou kontaminaci přesahující 700 μg/g (Stella et al., 2015). Sledovaný úbytek obsahu PCB v čase naznačuje probíhající procesy přirozené atenuace v dané lokalitě (Uhlík et al., 2012). Nižší celkový obsah PCB byl také sledován v zemině odebrané z okolí kořenů rostlin vyskytující se vegetace, což napovídá o roli fyto/rhizoremediace (Stella et al., 2015). Kromě PCB obsahuje půda rovněž bifenyl, polyaromatické uhlovodíky (PAH), pesticidy (DDT a stopy hexachlorbenzenu a lindanu) a těžké kovy (převážně vysoké koncentrace chromu a zinku).

Zemina byla odebrána z hloubky 1 m, kde nebyly vizuálně zjištěny žádné kořeny přirozené vegetace. Vždy přibližně 350 g bylo umístěno do litrových inkubačních nádob vyložených hliníkovou fólií proti případné sorpci PCB na stěny nádoby. Čtyři měsíce staré sazenice křenu (*Armoracia rusticana*), lilku (*Solanum nigrum*) a tabáku (*Nicotiana tabacum*), jejichž kořeny byly nejprve omyty vodou pro odstranění původní zeminy, byly zasazeny do nádob s kontaminovanou zeminou. Rostliny byly kultivovány při stabilních pokojových podmínkách (25 °C, 12 hodin denního světla) po dobu 6 měsíců. Kromě projevu samotných

rostlin na půdní mikroorganismy byl sledován také vliv komerčně dostupného NPK minerálního hnojiva. Polovina výsadby od každé rostliny byla hnojena za použití Univerzal KH (Nohel Garden, Česká Republika) se složením: 9% N, 4% P, 8% K, 3% Mg, se stopovým obsahem Fe, Zn, Mn, Cu, Mo, B, dále organickou složkou a růstovými hormony. Aplikace probíhala dle návodu výrobce: jednou za dva týdny společně se zaléváním, počáteční dávka činila 15ml na 2l vody s následným dávkováním 5ml na 2l vody. Shodným postupem byla inkubována také kontrolní zemina bez vysazených rostlin, hnojená i nehnojená. Všechny vzorky byly připraveny v triplikátech. Po dosažení šesti měsíců byla zemina z oblasti kořenů a z kontrolních nádob homogenizována a uložena při teplotě −20 °C. Pro izolaci DNA jsme použili PowerMax Soil DNA Isolation Kit (MoBio Laboratories Inc., USA) dle standardního protokolu výrobce s výjimkou závěrečného koncentrování DNA srážením v etanolu s přidaným glykogenem (postup dle Uhlík et al., 2009). DNA ze třech korespondujících triplikátů byla na závěr smíchána dohromady.

5.1.2 Shotgunová sekvenační analýza

Sekvenační knihovny ze vzorků izolované DNA byly připraveny podle návodu Rapid Library Preparation Manual (Roche) pro pyrosekvenování genomové DNA. Každá knihovna odpovídající jednomu vzorku byla sekvenována na jednom regionu velkého formátu (*large region*) pikotitrační destičky s využitím přístroje GS FLX (Roche), verze reagencií Titanium.

5.1.3 Anotace krátkých DNA čtení

Výsledná čtení byla nahrána na online server MG-RAST (Meyer et al., 2008), který umožňuje kontrolu kvality, filtrování chybných a uměle duplikovaných čtení a především automatickou anotaci sekvencí. Ta zahrnuje jednak přiřazení taxonomického původu sekvencí a také funkční anotaci. Pro získání profilu četností taxonomických přiřazení jsme použili prohledávání proti kompozitní neredundatní databázi (*multi-source, non-redundant*) M5NR vytvořené správci serveru MG-RAST. Do profilu byly zařazeny pouze hity s *E-value* < 10^{-10}, identitou ≥ 60% a minimální délkou alignmentu 15 aminokyselin; pro každé anotované čtení byl použit jeden reprezentativní záznam. Stejné parametry byly použity i pro tvorbu funkčních profilů pro každý soubor dat s využitím zařazení funkčních genů do COG hierarchické nomenklatury (Tatusov et al., 2000; Tatusov et al., 2003).

5.1.4 Komparativní analýza metagenomů na základě shotgunových čtení

Abychom odhalili rozdíly v získaných profilech mezi jednotlivými vzorky, rozhodli jsme se adaptovat bioinformatické postupy používané pro analýzy rozdílně exprimovaných genů, které jsou zahrnuty v balíčku DESeq2 (Love et al., 2014) pro statistický software R (R Developmental Core Team, 2009). Postup spočívá v aplikaci lineárního modelu na četnostní profily. Dále je nutné provést normalizaci dat, aby četnosti výskytu taxonomických či funkčních kategorií byly navzájem porovnávatelé, jelikož pro každý vzorek jsme vycházeli z rozdílného počtu čtení. K tomu jsme použili funkci *rlog* rovněž z balíčku DESeq2. Následuje samotné párové porovnání profilů postupně mezi všemi vzorky, které jsou seřazeny podle předem definovaných schémat. V našem případě byly vzorky seřazeny do skupin podle použití hnojiva (hnojené – nehnojené) a podle vegetace (křen – tabák – lilek – kontrola). Rozdíly jsme hodnotili jako statisticky signifikantní, pokud splňovaly kritéria (1) *fold change* práh 1,2 a (2) *false discovery rate* hranice 0,01, což znamená, že jsme dále posuzovaly pouze znaky, které vykazovaly nejméně 20% rozdílnost v alespoň jednom a více vzájemných porovnáních s pravděpodobností falešně pozitivního výsledku 1%. Tyto rozdíly byly vizualizovány formou *heatmap* grafů v softwaru R.

5.1.5 16S rRNA amplifikace a sekvenace

Úsek genů kódujících 16S rRNA byl amplifikován prostřednictvím PCR s využitím sady primerů: *forward* f563-577: 5'-AYTGGGYDTAAAGNG-3' (Cole et al., 2009) a *reverse* r1406-1392: 5'- ACGGGCGGTGTGTRC-3' (Lane, 1991). Každý z primerů nesl na 5'-konci jeden ze sekvenačních adaptorů (podle 454 Sequencing Application Brief No. 001-2009, Roche); *forward* primer navíc obsahoval ještě krátkou sekvenci specifického identifikátoru pro následné rozlišení vzorků po sekvenaci na sdílených regionech pikotitrační destičky (podle 454 Sequencing Technical Bulletin No. 005-2009, Roche). Reakční směs v celkovém objemu 20 µl obsahovala 0,2mM dNTPs (Finnzymes, Finland), 0,25µM primery (Generi Biotech, Česká Republika), 0,1mg.mL^{-1} BSA (*bovine serum* albumin, New England BioLabs, Great Britain), 0,4 jednotky (U) Phusion Hot Start II DNA Polymerázy (Finnzymes, Finland) s odpovídajícím pufrem a templátovou DNA (10-50 ng). Cykly amplifikace byly následující: 98 °C po dobu 30 s, 35 cyklů 98 °C po dobu 10 s, 60 °C po dobu 30 s, 72 °C po dobu 60 s a závěrečná extenze 72 °C po dobu 10 min. Výsledné produkty byly smíchány v ekvimolárním množství a přečištěny s využitím AMPure XP kuliček (Agencourt, Beckman Coulter, USA) dle manuálu pro

odstranění krátkých fragmentů (s délkou pod 200 bp), které mají tendenci převládat v následné přípravě sekvenační reakce. Produkty byly následně sekvenovány z jednoho směru (od *forward* primeru) na přístroji GS FLX s využitím Titanium+ reagencií.

5.1.6 Zpracování 16S rRNA sekvencí a comparativní analýza

Sekvence 16S rRNA amplikonů byly zpracovány v softwaru Mothur verze 1.27 (Schloss et al., 2009). Při zpracování jsme postupovali podle schématu: (1) Použili jsme rozmezí mezi 650 a 800 cykly (*flows*) sekvenace. (2) Pro odstranění nekvalitních sekvencí (tzv. odstranění šumu, *denoising*) byly hodnoceny intenzity signálu v sekvenačních cyklech (*flowgramy*) algoritmem PyroNoise (Quince et al., 2009). (3) Sekvence primerů a identifikátorů na začátku čtení byly ořezány. (4) Čtení byla dále zarovnána podle referenčního alignmentu SILVA (verze 119) a ta, která nedosahovala délky minimálně 400bp, byla vyřazena. (5) Čtení byla klastrována (*single-linkage pre-clustering*) s nastavením minimálního rozdílu jedné záměny na 100 bp. (6) Chybná čtení, která vznikají při PCR přeskočením polymerázy z jedné templátové molekuly DNA na jinou (chimerní sekvence), byla identifikována implementovaných programem Perseus (Quince et al., 2011) a odstraněna. (7) Odstraněny byly také sekvence, které se vyskytují v datasetech jen jednou a nebyly přiřazeny žádnému klastru (singletony), a rovněž kontaminantní sekvence (například původu mitochondriálního, chloroplastového či pocházející z Eukaryotických genomů). (8) Výsledná, validní čtení byla klasifikována podle referenční databáze SILVA (verze 119).

Výsledný taxonomický profil (četnosti klasifikovaných OTU pro každý vzorek) představuje stejný datový formát jako četnostní profily získané anotací shotgunových čtení. Následná komparativní analýza byla provedena stejným způsobem jako komparativní analýza shotgunových profilů s využitím metod pro expresní profilování v balíku DESeq2, jak bylo popsáno výše.

Nemetrické mnohorozměrné škálování za účelem vizualizace případné struktury v datech jsme provedli v balíku vegan (Oksanen et al., 2013) pro statistický software R (R Development Core Team, 2009) s využitím funkce *metaMDS* (s parametry *nonshare* = FALSE a *autotransform* = FALSE) a funkce *envfit*.

5.1.7 Umělá komunita

Paralelně s přípravou 16S rRNA sekvencí ze vzorků zeminy byly prostřednictvím PCR získány 16S rRNA amplikony také z umělé (*mock*) komunity připravené v laboratoři jako mix genomové DNA izolované z osmi kultivovaných bakterií se známou, publikovanou sekvencí genomů (tabulka 5.1). Buňky byly pěstovány přes noc v Luria Bertani mediu (Oxoid, UK), genomovou DNA jsme následně izolovali sadou PureLink Genomic DNA Mini Kit (Invitrogen, USA). Počet kopií16S rRNA genů na jeden nanogram každé genomové DNA byl odhadnut na základě real-time kvantitativní PCR (qPCR) s použitím referenční křivky konstruované z genomové DNA *Pseudomonas stutzeri* JM300, která obsahuje 213tis. 16S rRNA kopií na 1 ng genomové DNA (Ginard et al., 1997). Kultivace a izolace DNA *Pseudomonas stutzeri* JM300 probíhala stejným způsobem jako u členů umělé komunity popsaným výše.

Tabulka 5.1: Bakterie použité pro tvorbu umělé komunity.

Bakterie	BioProject Accesion
Achromobacter xylosoxidans A8	PRJNA59899
Burkholderia xenovorans LB400	PRJNA254
Pseudomonas putida JB	PRJNA84349
Rhizobium radiobacter C58	PRJNA57865
Arthrobacter chlorophenolicus A6	PRJNA58969
Bacillus pumilus SAFR-032	PRJNA59017
Micrococcus luteus NCTC 2665	PRJNA59033
Rhodococcus jostii RHA1	PRJNA58325

Podmínky qPCR byly následující: 12 μl každé reakční směsi obsahovalo 1x DyNAmo Flash SYBR Green qPCR Mastermix (Finnzymes, Finland), 4 pmol každého z páru primerů f786: 5‘-GATTAGATACCCTGG-tag-3’ a r939: 5‘-CTTGTGCGGGCCCCCGTCAATTC-3‘ (Baker et al., 2003) a 2 μl templátové DNA. Cykly programu jsme nastavili na: 95 °C po dobu 5 min, 35 cyklů 95 °C po dobu 20 s, 55 °C po dobu 30 s, 72 °C po dobu 30 s a závěrečná extenze 72 °C s dobou trvání 10 min. Genomovou DNA osmi vybraných bakterií jsme smíchali podle výsledků qPCR tak, aby počet kopií 16S rRNA genů spadal do jednoho řádu. PCR amplifikace úseku 16S rRNA byla provedena stejným způsobem jako u ostatních vzorků použitých ve studii a produkty byly sekvenovány společně se vzorky na obou použitých regionech pikotitrační destičky (viz výše). Získané amplikonové sekvence umělé komunity byly rovněž zpracovány v softwaru Mothur verze 1.27 (Schloss et al., 2009) nejprve samostatně pro ověření zvolených metod a nastavení parametrů a následně paralelně s ostatními vzorky jako vnitřní kontrola.

5.2 Výsledky

Přehled sekvenačních výsledků pro shotgunová data a získaných 16S rRNA čtení poskytuje tabulka 5.2. Všechny sekvence jsou veřejně dostupné v databázi *Sequence Read Archive* (SRA) pod přístupovými kódy ERA596740 (shotgunová data) a ERA168068 (16S rRNA sekvence). Projekt založený na serveru MG-RAST, který nabízí přehled filtrování čtení a kompletní anotovaná data pro shotgunové sekvence, je přístupný na odkazu: http://metagenomics.anl.gov/linkin.cgi?project=271

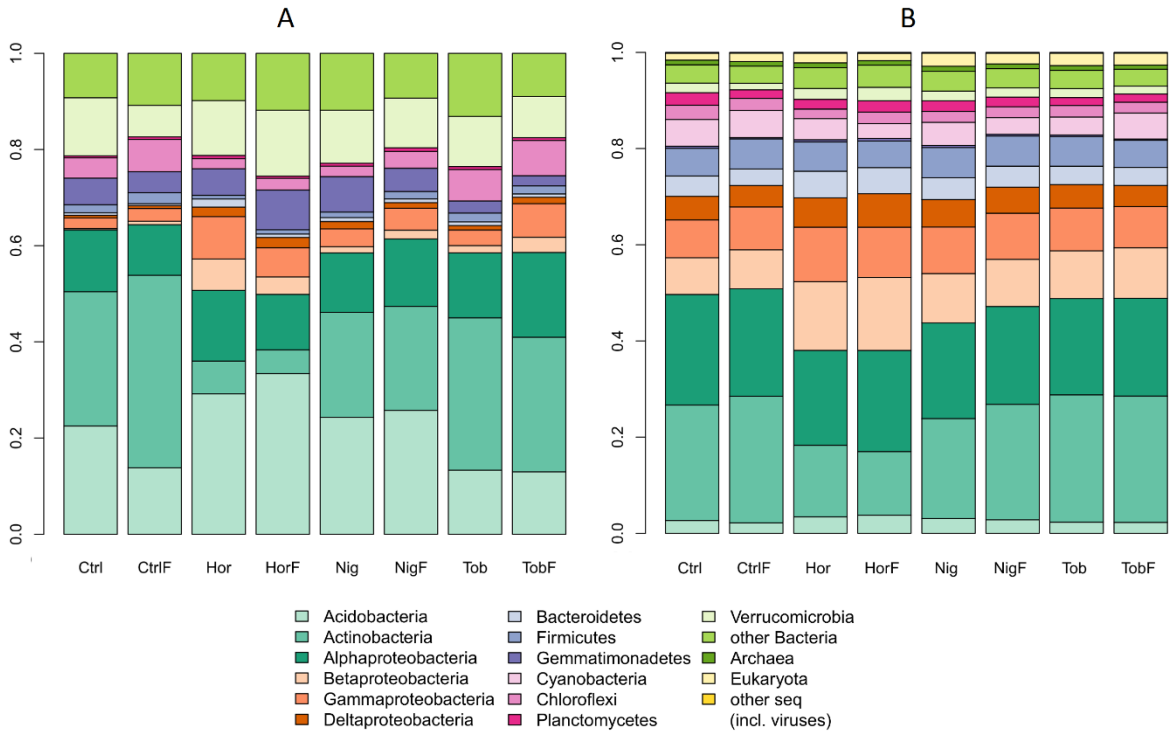
Tabulka 5.2: Kompletní přehled získaných sekvencí. Celkový počet bazí a průměrná délka se vztahují k shotgunovým datům.

Vzorek	Rostlina	Hnojení	# amplikon	# shotgun	bp	Prům. délka (bp)
Hor	kren	-	9 853	712 104	286 480 433	402
HorF	kren	+	14 373	756 433	308 435 859	408
Ctrl	control	-	12 607	727 529	297 444 144	409
CtrlF	control	+	14 259	665 714	276 871 390	416
Tob	tabak	-	16 029	794 668	325 348 558	409
TobF	tabak	+	8 163	735 196	301 519 047	410
Nig	lilek	-	13 302	584 208	245 662 269	421
NigF	lilek	+	9 452	655 217	275 891 786	421

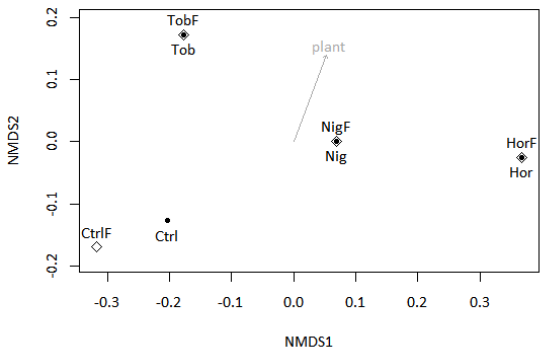
5.2.1 Složení mikrobiálních společenstev v zemině

Na základě taxonomické klasifikace shotgunových dat je zřejmé, že v sekvenovaných metagenomech dominuje bakteriální DNA – čtení klasifikovaná jako bakteriální představují celkově 96.79% ± 0.46% (průměr ± SD, obrázek 5.2a). Z toho důvodu byla následná amplifikace 16S rRNA provedena primery specifickými přednostně pro bakteriální geny, abychom podpořili shotgunová data a získali větší rozlišení pro taxonomickou analýzu v rámci bakteriální domény. Analýza 16S rRNA amplikonů (obrázek 5.2b) ukazuje, že v kontrolních vzorcích zeminy a ve vzorcích osazených tabákem převládají Acinetobacteria. Naproti tomu ve vzorcích zeminy z prostředí kořenů křenu a lilku jsou více zastoupeny Acidobacteria. Tento rozdíl je zvláště patrný u vzorků ovlivněných křenem. V rámci kmene Proteobacteria jsou celkově nejpočetnější Alphaproteobacteria, přičemž Beta-, Gamma- a Deltaproteobacteria jsou nabohacené ve všech vzorcích osazených některou z rostlin. Na základě mnohorozměrného škálování můžeme sledovat, že bakteriální společenstva jsou ovlivněna vegetací spíše než přidáním hnojiva (nemetrické mnohorozměrné škálování založené na Bray-Curtis míře nepodobnosti, obrázek 5.2). Proložením vektoru lze demonstrovat signifikantní vliv

přítomnosti rostlin (*P*-hodnota < 0,05), zatímco použití hnojiva nemá statisticky významný vliv. Taxonomická diverzita bakterií ze vzorků ovlivněných přítomností rostlin je o něco vyšší než u kontrolní zeminy; tento trend, byť slabý, se jeví jako konzistentní napříč vzorky. Rozdíly v diverzitním indexu Simpson jsou pouze nevýrazně signifikantní (Mann-Whitney test, *P*-hodnota = 0.07, příloha - tabulka 5.3).



Obrázek 5.2: Taxonomická struktura půdních mikrobiálních společenstev stanovená na základě A) amplikonových 16S rRNA dat a B) shotgunových dat. Zkratky podle tabulky 5.2.



Obrázek 5.3: Nemetrické mnohorozměrné škálování (NMDS, stress < 0.001) na základě 16S rRNA amplikonových dat s proložením vektoru odpovídajícího vlivu rostlin (*plant*) (*P-value* < 0.05, 40 320 permutací).

Tabulka 5.3: Hodnoty Simpsonova indexu diverzity pro jednotlivé vzorky na základě amplikonových 16A rRNA dat.

Vzorek	Simpson index diverzity
Hor	0.9938542
HorF	0.9926230
Ctrl	0.9923152
CtrlF	0.9914705
Nig	0.9935159
NigF	0.9938309
Tob	0.9924577
TobF	0.9935668

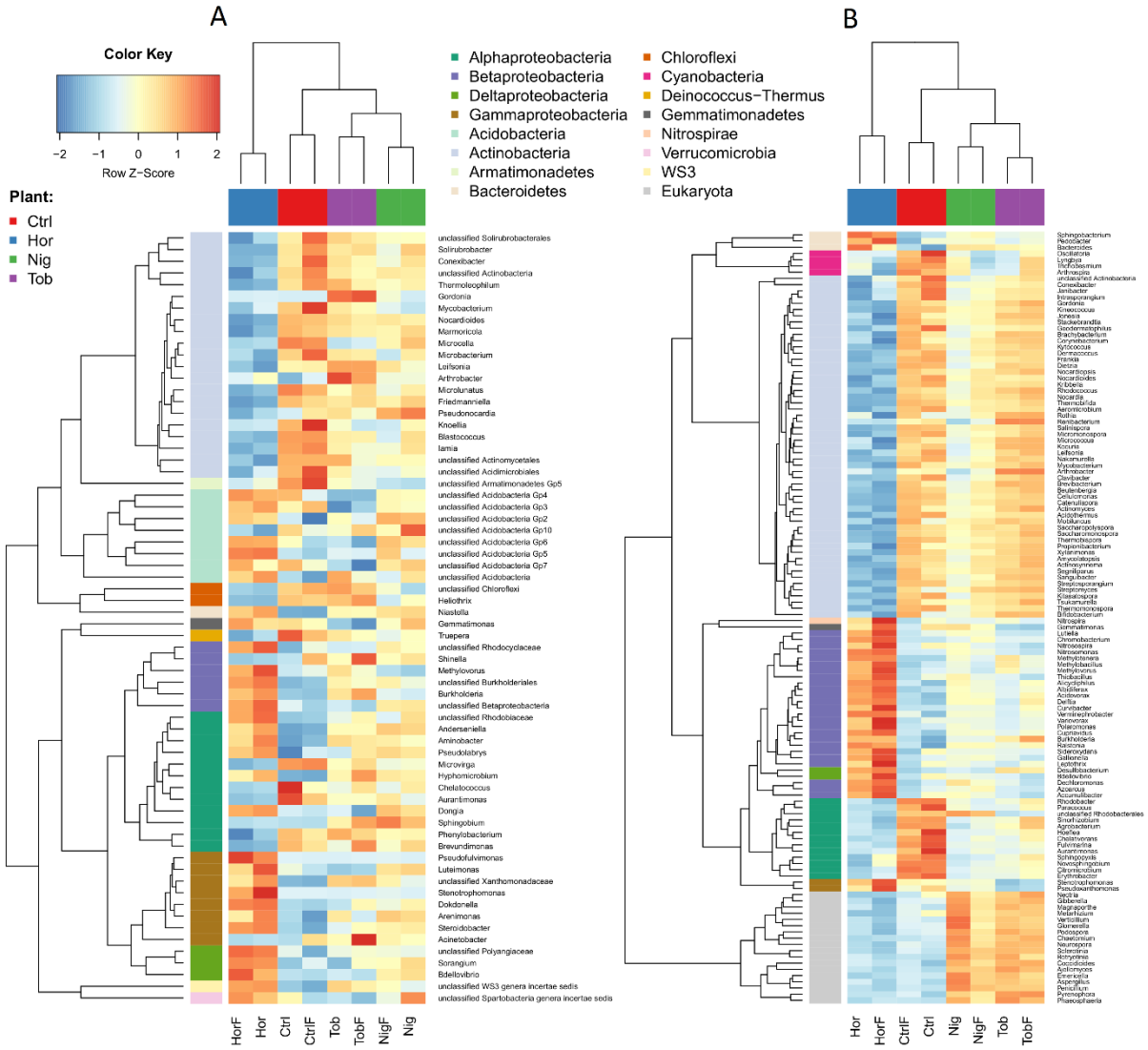
5.2.2 Rozdíly ve složení mikrobiálních společenstev

Párovým porovnáním datasetů sloučených do skupin podle přítomnosti rostliny a aplikace hnojiva jsme byli schopni odhalit členy společenstev, kteří se vyskytovali ve vzorcích v různé míře. Rozdíly v zastoupení identifikované na základě 16S rRNA dat jsou vizualizované prostřednictvím heatmap grafu (Obrázek 5.4a) a odrážejí obdobné trendy jako shotgunová data (Obrázek 5.4b), která zahrnují rovněž populace z říše Fungi. Mikrobiální komunity formované v okolí kořenů křenu se nejvíce odlišují od všech ostatních sledovaných vzorků zeminy. V menším počtu jsou zde zastoupena Actinobacteria, zvláště rod *Streptomyces*. Více četná jsou Gammaproteobacteria (s výjimkou rodu *Acinetobacter*), Deltaproteobacteria a také Betaproteobacteria zahrnující rod *Burkholderia*. Aktinomycety rodu *Arthrobacter* a *Gordonia* obývají naproti tomu ve větším počtu oblast kořenů tabáku. Kontrolní zemina bez vegetace obsahuje nabohacené zástupce taxonu Actinobacteria, zatímco skupina Betaproteobacteria je relativně méně zastoupena (obrázek 5.4). Jak umožňují rozlišit data ze shotgunového sekvenování, zástupci říše Fungi prospívali méně v oblasti kořenů křenu a kontrolní zemině, naproti tomu více početné jsou u vzorků s kultivací lilku a tabáku (obrázek 5.4b).

5.2.3 Funkční potenciál společenstev a jejich vzájemné porovnání

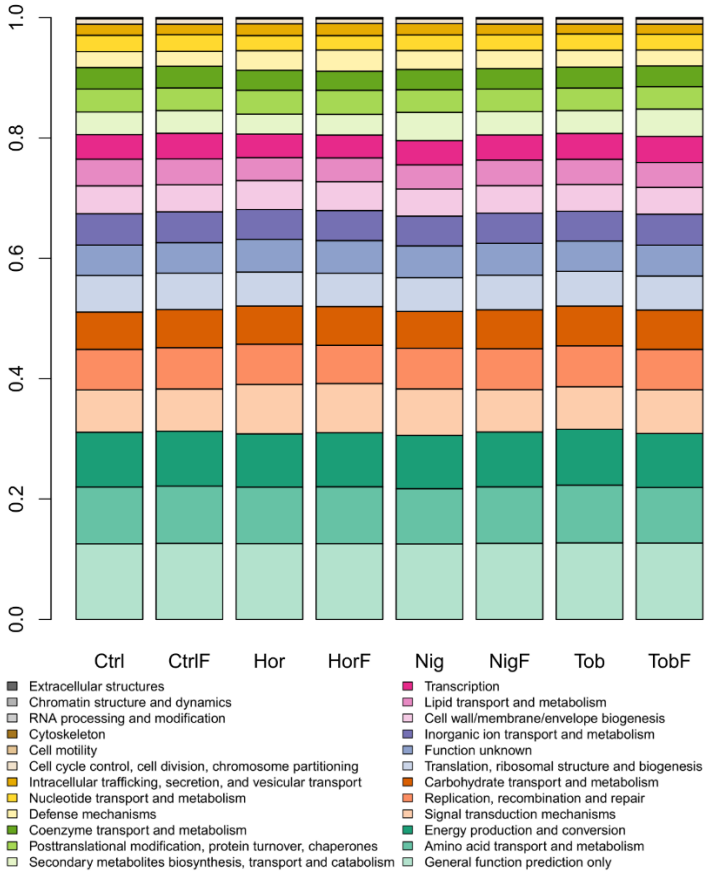
Funkční potenciál jednotlivých metagenomů lze odvodit na základě klasifikace shotgunových čtení do COG kategorií. Na základě analýzy míry četnosti funkčně anotovaných genů je patrné, že funkční potenciál všech vzorků je velmi podobný (obrázek 5.5). Nejčastěji zastoupeným funkčním systém na 1. úrovni COG hierarchie je kategorie obsahující pouze obecný odhad funkce méně charakterizovaných znaků COG systému (*General function prediction only*), za ní dle četností následují systémy *Amino acid transport and metabolism*; *Energy production and conversion*; *Signal transduction mechanisms*; *Replication,*

recombination and repair; *Carbohydrate transport and metabolism*; *Translation, ribosomal structure and biogenesis* a další málo časté systémy.



Obrázek 5.4: Taxony vykazující signifikantní rozdíly v četnosti při párovém porovnání. A) Amplikonová 16S rRNA data. B) shotgunové data.

Při párovém porovnání funkčních profilů na nižších hierarchických úrovních jsme zaznamenali statisticky signifikantní rozdíly, které vykazují podobný trend, jako taxonomická data: méně rozdílů mezi hnojenými a nehnojenými vzorky na rozdíl od většího počtu rozdílů mezi vzorky ovlivněnými přítomností rostlin. Druhá úroveň funkčních subsystémů poskytuje dvě kategorie, které se signifikantně liší v zastoupení: obranné mechanismy (*Defense mechanism*) a geny spojené s funkcí cytoskeletu (*Cytoskeleton*). Obě se více vyskytují ve vzorcích s pěstovanými rostlinami než v kontrolní zemině.

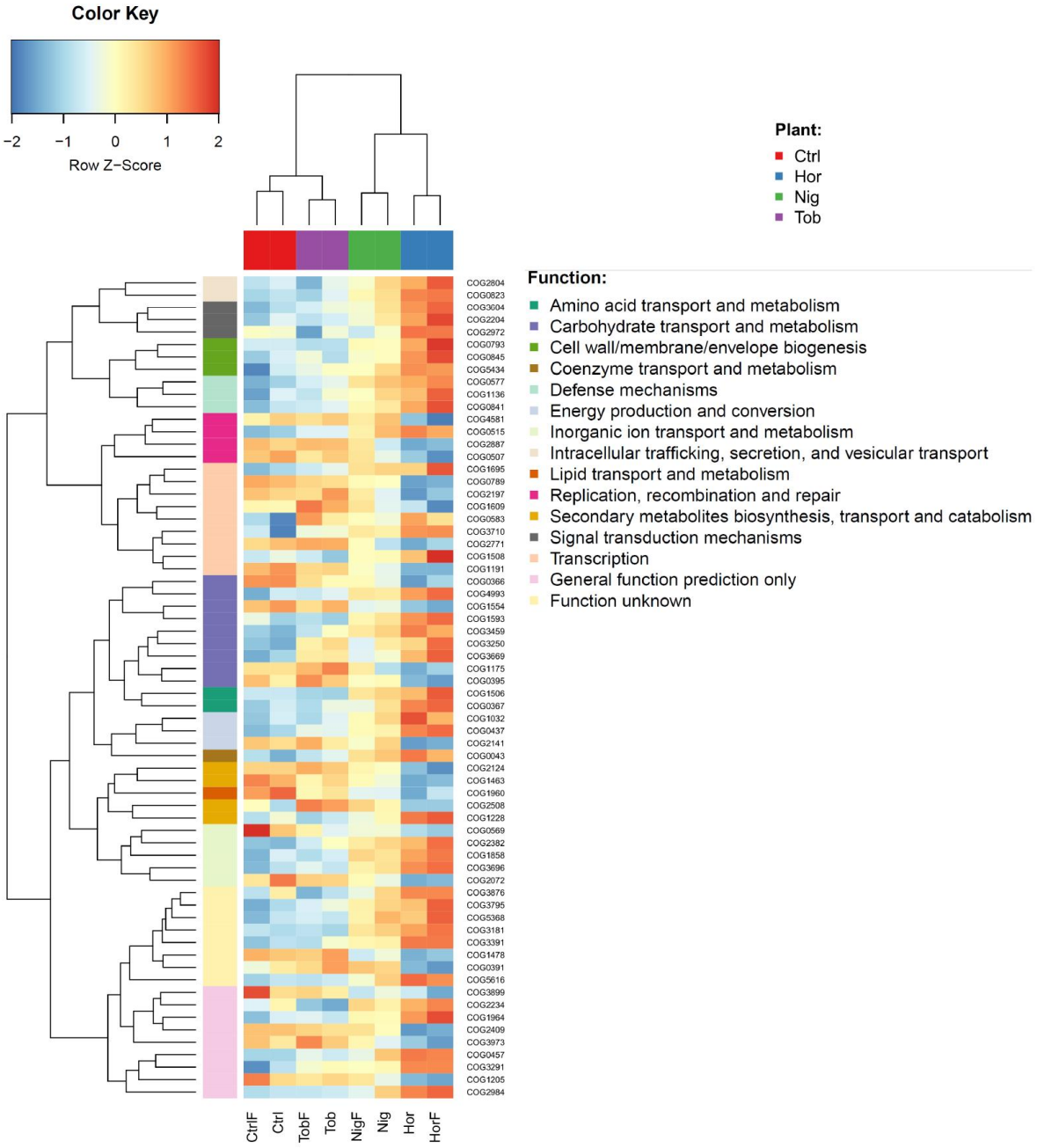


Obrázek 5.5: Zastoupení COG funkčních subsystémů ve vzorcích na základě shotgunových dat.

Více rozdílů přitom vykazují data seskupená na třetí úrovni, kde jsme identifikovali 14 rozdílně četných COG funkcí, přičemž celkově nejvíce se od ostatních vzorků odlišuje zemina s křenem, zatímco prostředí kořenů tabáku spadá nejblíže kontrolním vzorkům. Statisticky významné rozdíly na 4. úrovni jednotlivých COG identifikátorů vykazují stejný trend a jsou prezentovány na obrázku 5.6 společně s vyznačením funkce na třetí úrovni COG hierarchie.

5.3 Diskuze

V prezentované studii jsme se zaměřili na získání nových poznatků o roli pěstované vegetace a s tím spojené aplikace hnojiva na utváření struktury mikrobiálních komunit a jejich funkčního potenciálu ve vzorcích dlouhodobě kontaminované zeminy. Pro postihnutí funkčního potenciálu jsme přistoupili k shotgunovému sekvenování celkové metagenomické DNA. Anotovaná čtení poskytují rovněž taxonomický profil mikrobiálních společenstev, pro hlubší analýzu bakteriálních kmenů jsme se navíc zaměřili na sekvenování 16S rRNA genů amplifikovaných z celkové DNA primery specifickými přednostně pro bakteriální geny.



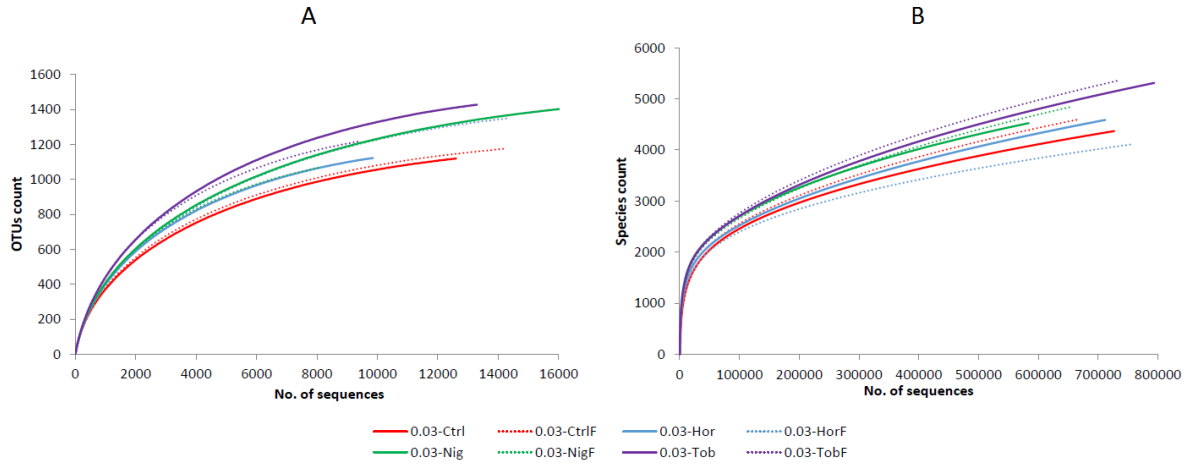
Obrázek 5.6: COG funkce vykazující signifikantní rozdíly mezi vzorky při párovém porovnávání na základě shotgunových dat. Heatmapa ukazuje zastoupení jednotlivých COG na úrovni 4, úroveň 3 je barevně zvýrazněna sloupcem na levé straně grafu, barevné kódování udává legenda (*function*).

Pro bioinformatickou analýzu dat jsme se rozhodli využít přístupy z oblasti RNA sekvenování (RNA-seq) dobře zpracované v balíčku DESeq2 (Love et al., 2014) pro software R. Metody původně navržené na statistické hodnocení rozdílné míry genové exprese na základě cDNA sekvenačních dat jsou podle nás dobře použitelné i na hodnocení různého stupně četnosti operačních taxonomických jednotek, jednotlivých mikrobiálních linií a funkčních genů v amplikonových 16S rRNA a shotgunových metagenomických datech. Využití RNA-seq

postupu nám umožnilo zpracování všech získaných datasetů – taxonomických profilů ze shotgunového a 16S amplikonového sekvenování a funkčních metagenomických profilů – naprosto shodným způsobem. Ukazuje se navíc, že nejsme jediní, kdo rozpoznal potenciál RNA-seq statistických metod pro metagenomiku. McMurdie & Holmes (2014) dokládají na základě analýzy simulovaných i experimentálních metagenomických vzorků, že RNA-seq metody normalizace a aplikace lineárního modelu (zahrnuté např. právě v námi zvoleném balíku DESeq2) vedou k lepším výsledkům při analýze metagenomických vzorků, než rozšířenější metody závisející například na normalizaci dat na jednotnou velikost souboru (tzv. *rarefying*). Především jde o redukci falešně pozitivních výsledků při hodnocení míry četnosti rozdílně zastoupených znaků mezi soubory (McMurdie & Holmes, 2014).

Jelikož v shotgunových čteních dominovaly sekvence bakteriálních genomů (obrázek 5.2b), amlikonové sekvenování jsme zacílily na bakteriální 16S rRNA geny. Hlavní výhoda shotgunového sekvenování oproti amplikonovému je, že při přípravě knihovny vyhází z celkové DNA a nepodléhá sklonu k nabohacení některých sekvencí, které mohou být preferovány při přípravě amplikonů během PCR. Na druhou stranu mezi potenciální nedostatky shotgun přístupu patří menší hloubka pokrytí a případné nezachycení málo četných taxonů, převážně v případě velmi diverzifikovaných biomů. Rovněž preference na straně referenčních databází, které ve větší míře obsahují snáze kultivovatelné bakterie, zatímco řadu taxonů vůbec nepostihují, vede k posunu při anotaci či nedostatečné klasifikaci některých čtení (Shah et al., 2012). Kombinování obou přístupů tak může přispět k zodpovězení důležitých ekologických otázek a to i v případě, kdy jde o studium vzorků obsahujících vysokou mikrobiální diverzitu, jak indikují *rarefaction* křivky na základě sekvencí DNA získaných ze zkoumané zeminy (obrázek 5.7). Mezi shotgunovými daty dominují sekvence klasifikované do kmenů Actinobacteria a Alphaproteobacteria, následované dalšími zástupci Proteobacteria a kmenem Firmicutes (obrázek 1a). Amplikonová data poskytují poněkud jiný obrázek: poměrně nižší zastoupení skupin Alphaproteobacteria, Cyanobacteria a Firmicutes zatímco více zastoupený je kmen Acidobacteria. Chloroplastová DNA sdílí fylogenetický původ se skupinou Alphaproteobacteria, zatímco mitochondriální genomy jsou fylogeneticky příbuzné s třídou Cyanobacteria (Ni et al., 2013). Právě DNA z chloroplastů a mitochondrií je rovněž zastoupená v celkové metagenomické DNA a část shotgunových čtení, klasifikovaných jako zástupci Alphaproteobacteria a Cyanobacteria, ve skutečnosti odráží přítomnost mitochondriální či chloroplastové DNA. Naproti tomu z amplikonových dat jsou tyto sekvence filtrovány, což vysvětluje nižší zastoupení čtení klasifikovaných do skupin Alphaproteobacteria a

Cyanobacteria. Nižší četnost kmene Acidobacteria v shotgunových datech lze naopak vysvětlit obecným nedostatkem známých, sekvenovaných referenčních zástupců této skupiny v sekvenčních databázích (Lee et al., 2015). Z těchto důvodů se nám jeví aplikace rovněž amplikonového sekvenování jako nezbytná pro lepší a přesnější popis bakteriální diverzity ve zkoumaných vzorcích.



Obrázek 5.7: Rarefakční křivky na základě A) amplikonových 16S rRNA a B) shotgunových dat.

5.3.1 Umělá komunita jako vnitřní kontrola

Umělá komunita byla konstruována z DNA vybraných bakterií kmene Proteobacteria, Actinobacteria a Firmicutes. Zástupci těchto kmenů jsou nejčastěji asociováni s degradací aromatických sloučenin. Sekvenování a analýza 16S rRNA genů z osmi bakteriálních linií umělé komunity slouží (1) jako kontrola přesnosti a účinnosti amplifikace a sekvenace samotné a také (2) pro identifikaci správných parametrů pro zpracování amplikonových dat. V ideálním případě bychom měli dostat osm taxonomických jednotech (OTU), které odpovídají bakteriím použitým pro tvorbu definované komunity. Při PCR, konstrukci sekvenačních knihoven, jejich amplifikaci a při samotné sekvenační reakci však dochází k chybám, které mohou zastřít původní složení metagenomického vzorku (Schloss et al., 2011; Uhlík et al., 2012). Z našich zkušeností s pyrosekvenováním vyplývá, že kvalita čtení a rovněž to, jakým způsobem se následně chovají v dalších analýzách a projevují se ve výsledcích, se může odlišovat nejen mezi různými vzorky a různými sekvenačními knihovnami, ale také mezi jednotlivými běhy sekvenátoru a dokonce mezi regiony sekvenační destičky v rámci jedné sekvenační procedury. Naše předchozí analýzy ukázaly, že výsledky amplikonového sekvenování jsou výrazně ovlivněny vznikem chimerních sekvencí při PCR a volbou programu na jejich detekci a

odstranění (Uhlík et al., 2012). Eliminace problematických čtení lze dosáhnout také optimálním nastavením minimálního a maximálního počtu cyklů (*flows*), při kterých je při sekvenaci aplikována reakční směs vždy s jedním konkrétním nukleotidem. Pokud by se v sekvencích nevyskytovaly homopolymery (tedy úseky se dvěma a více stejnými nukleotidy za sebou), odpovídal by počet *flows* délce sekvencí v nukleotidech. Reálné sekvence však homopolymery obsahují a vztah mezi počtem *flows* a délkou sekvencí je rozdílný pro různé sekvence. Pro amplikonová data z umělé komunity jsme testovali různá nastavení počtu cyklů v programu Mothur (parametry *minflows* a *maxflows* příkazu *trim.seq*) a postupně se dopracovali k nastavení minimálního počtu 650 a maximálního 800, kdy počet OTU dosahoval čísla 9, ale do OTU9 přináležela jen dvě čtení, zatímco ostatních 8 taxonomických jednotek odpovídalo použitým bakteriím (nepublikovaná data). Při zpracování všech sekvencí dohromady byl výsledek umělé komunity 8 OTU, které přesně odpovídaly kultivovaným bakteriím. V tomto případě měl software Mothur k dispozici celkově větší počet čtení, průběh algoritmů je tedy přesnější a vnitřní kontrola velmi dobře fungovala pro ověření použitého postupu. Programy na identifikaci chimerních sekvencí Perseus (Quince et al., 2011) a UCHIME (Edgar et al., 2011) dosahovaly v prvním kroku testování velmi podobných výsledků (nepublikovaná data). Náš postup zpracování 16S rRNA sekvencí původně vychází ze standardizované procedury navržené pro *Human Microbiome Project* (Schloss et al., 2011), která zahrnuje UCHIME. V této studii jsme se přiklonili k programu Perseus na základě přechozí analýzy deseti tisíc čtení umělé komunity (Uhlík et al., 2012).

5.3.2 Okolí kořenů rostlin versus kontrolní zemina

Vliv vegetace na mikrobiální společenstva je již dlouhou dobu předmětem mikrobiologického výzkumu (Smalla et al., 2001; Kowalchuk et al., 2002; Berg & Smalla, 2009). Zatímco mikrobiální diverzita bývá sledována vyšší v zemině neovlivněné rostlinami, denzita mikrobiálních buněk je naopak zvýšená v oblasti rhizosféry, tedy úzké vrstvy bezprostředně kolem kořenů rostlin (Dennis et al., 2010; Philippot et al., 2013). Naše data ukazují mírně zvýšenou alfa diverzitu na základě hodnot Simpsonova indexu ve vzorcích osazených rostlinami. Nicméně je nutné zdůraznit, že námi zkoumaná zemina se v zemědělském hledisku odlišuje od běžně zkoumaných půdních typů – je písčitohlinitá, velmi chudá na živiny a organickou složku a navíc znečištěná polychlorovanými bifenyly a vysokým obsahem chromu a zinku (Stella et al., 2015). Právě rhizodepozity, látky produkované kořeny rostlin, zde mohou hrát velkou roli v utváření mikrobiálních komunit a mít za následek

naměřenou vyšší míru diverzity organismů. Obecně je dokládán dramatický účinek rhizodepozice na mikrobiální populace, přičemž kořenům rostlin se připisuje schopnost poskytovat vše potřebné, co okolní zemina jinak postrádá (Hartmann et al., 2009). Sledovaný nárůst diverzity naznačuje potenciál v oblasti kořenů rostlin jakožto místa s vyšší efektivitou bioremediačních procesů, což je v souladu s předchozími pokusy (Macková et al., 2006).

Biodegradační potenciál konsorcií ovlivněných přítomností rostlin demonstruje rovněž fakt, že ve větší míře obsahují bakteriální kmeny, které předchozí studie identifikovaly jako PCB degradéry. Zemina z okolí kořenů křenu je nabohacena o rod *Burkholderia*, jehož zástupci byli popsáni jako velmi efektivní degradéři PCB (Mukerjee-Dhar et al., 1998; Tillman et al., 2005; Chain et al., 2006; Uhlík et al., 2013), *Stenotrophomonas*, který byl dříve izolován ze stejné skládkové zeminy z kultury pěstované na PCB (Uhlík et al., 2013) či *Methylovorus*, jehož schopnost metabolismu bifenylů byla přímo detekována metodou SIP ze vzorků rhizosféry křenu (Uhlík et al., 2009). Naše data rovněž dokládají, že zemina s pěstovaným křenem obsahuje mikrobiální konsorcium, které se nejvíce odlišuje od všech ostatních vzorků. *Arthrobacter* a *Gordonia*, bakteriální rody často asociované s biodegradací PCB (Gilbert & Crowley, 1997; Abraham et al., 2005; Leigh et al., 2007; Koubek et al., 2012), jsou naproti tomu nabohacené u vzorku s pěstovaným tabákem.

5.3.3 Kořeny rostlin versus aplikace hnojiva

Aplikace hnojiva má velký význam nejen v zemědělství, ale je mu také věnována pozornost v ekologii. Předmětem studia jsou vlivy chemického hnojení na konsorcia půdních mikroorganismů (Geisseler & Scow, 2014), zvláště pak z hlediska formování funkčních systémů v zemině a jejich vztažení ke specifickým populacím mikroorganismů (Su et al., 2015). Řada studií poukazuje na změny mikrobiálních komunit v důsledku aplikace chemického hnojiva (Leff et al., 2015; Su et al., 2015), zatímco jiné výzkumy dokládají výrazně vyšší účinek samotné vegetace než použitého hnojiva (Benizri & Amiaud, 2005; Liliensiek et al., 2012) a v některých studiích dokonce nebyl prokázán signifikantní efekt hnojení na půdní mikroorganismy (Merschner et al., 2001; Liliensiek et al., 2012). V tomto ohledu přispívají do diskuze výsledky našeho výzkumu: neprokázali jsme signifikantní posun v mikrobiálních konsorciích následkem aplikace chemického, minerálního hnojiva a to jak na úrovni taxonomického složení společenstev tak jejich funkčního potenciálu. Naproti tomu výsadba vybraných rostlin měla za následek změny mikrobiálních populací ve sledovaných vzorcích (obrázek 5.3, 5.4, 5.6).

5.3.4 Kopiotrofní versus oligotrofní mikroorganismy

Mikrobiální konsorcia v zemině bývají selektivně nabohaceny o kopiotrofní či naopak oligotrofní organismy v závislosti na množství přítomného organického uhlíku (Fierer et al., 2007). Prostředí přímo ovlivněné rhizodepozicí skrze kořeny rostlin obecně upřednostňuje výskyt kopiotrofních mikroorganismů (Dennis et al., 2010). Přestože typ růstové strategie nelze zcela jednoznačně vztáhnout k taxonomické příslušnosti organismu na úrovni kategorií kmene či třídy, členové Actinobacteria, Bacteroidetes, Alphaproteobacteria, Betaproteobacteria a Gammaproteobacteria bývají považováni spíše za kopiotrofní, zatímco kmeny Acidobacteria a Planctomycetes za oligotrofní (Fierer et al., 2007; Prober et al., 2015; Leewis et al., 2016). Zvýšená abundance zástupců Alpha-, Beta-, Gammaproteobacteria a kmene Bacteroidetes (obrázek 5.2) ve vzorcích s pěstovanými rostlinami podporuje hypotézu, že oblast kořenů rostlin favorizuje výskyt kopiotrofních organismů. Výskyt oligotrofních tříd bakterií Acidobacteria a Planctomycetes nesleduje tak jednoznačný trend. Zatímco kmen Planctomycetes byl málo četný napříč všemi vzorky, zástupci kmene Acidobacteria vykazují signifikantně odlišnou míru zastoupení ve vzorcích, přičemž největší zastoupení můžeme sledovat v prostředí kořenů křenu. Již dříve byl sledován rozdílný výskyt určitých linií Acidobacteria v rhizosféře specifické vegetace (Nunes de la Rocha et al., 2013) a naše výsledky dokládají pozitivní vliv křenu na vznik podmínek preferovaných pro růst členů třídy Acidobacteria.

Důležitá je v tomto ohledu rovněž skutečnost, že některé funkční kategorie jsou připisovány specificky kopiotrofním či oligotrofním růstovým strategiím. Funkční geny spadající do kategorií jako jsou přenos signálu (*signal transduction mechanisms*), obranné mechanismy (*defense mechanisms*) či transport a metabolismus aminokyselin (*amino acid transport and metabolism*) bývají nabohacené u kopiotrofních organismů (Lauro et al., 2009; Leff et al., 2015). Jejich výskyt se rovněž signifikantně odlišuje mezi našimi vzorky – můžeme sledovat nižší abundanci v kontrolní zemině ve srovnání s ostatními vzorky ovlivněnými přítomností vegetace (obrázek 5.6).

5.4 Závěry

Popsaná studie dokládá, že přítomnost rostlin a nikoli aplikace chemického hnojiva hraje roli ve formování mikrobiálních společenstev v námi studované zemině. Tento efekt se projevuje na úrovni taxonomického složení populací i jejich funkčního potenciálu a míra vlivu

je závislá na konkrétním rostlinném druhu. Obecně můžeme sledovat vznik podmínek upřednostňujících nárůst kopiotrofních organismů v zemině ovlivněné kořeny rostlin. Prezentované výsledky přispívají k poznání ekologických vztahů mezi vegetací a půdní mikroflórou v kontaminovaném prostředí a mohou být dále aplikovány pro porozumění dynamických procesů asociovaných s fyto/rhizoremediacemi na kontaminovaných lokalitách.

6 Směr dalších výzkumů

Metody masivně paralelního sekvenování a s nimi spjatý vývoj bioinformatických nástrojů pro analýzu rozsáhlých souborů sekvenčních dat přispěly ke skokovému rozmachu genomiky, který postupně zasáhl další obory. V rámci mikrobiologie urychlila aplikace těchto metod rozšíření oblasti studia od modelových, v laboratoři pěstovaných druhů až po celá společenstva organismů. Řada přístupů popisovaných v této práci je spjata především se 454 pyrosekvenováním, jehož vývoj byl zvláště zaměřen na postupné prodlužování čtených sekvencí. Delší čtení umožňují nejen sestavení kompletních bakteriálních genomů, ale také jejich analýzu jakožto environmentálních značek v metagenomickém výzkumu. V oblasti provozních nákladů přepočtených na jednu bazi a pracnosti přípravy knihoven a sekvenačního běhu však 454 sekvenátory přestávají konkurovat dalším přístrojům schopným paralelní analýzy řádově většího množství sekvencí ovšem za cenu kratší délky čtení. Většina genomových a metagenomových studií je v současnosti založena na analýze sekvencí dlouhých maximálně 300 bp produkovaných sekvenátory od firmy Illumina. To klade nové nároky na bioinformatické nástroje pro zpracování dat. Řada publikovaných sekvencí však zůstává pouze ve formě nekompletního draftu, a to i v případě studia relativně krátkých a málo komplexních bakteriálních genomů (Land et al., 2015).

Naproti tomu obsáhlé sekvenační projekty ukazují potenciál pro rekonstrukci genomových sekvencí i z komplexnějších metagenomických dat. Příkladem je studie Brown et al. (2015), kdy autoři byli schopni sestavit 8 kompletních a 789 draftových sekvencí do té doby zcela nepopsaných bakteriálních linií na základě 224 miliard krátkých sekvencí. Tento výzkum jasně demonstruje budoucí cestu metagenomických studií.

Některá konsorcia – typicky půdní mikroorganismy – však zahrnují obrovské množství mikrobiálních druhů a shotgunové sekvenování i s využitím nejvýkonnějších platforem nemusí postačovat k odhalení méně zastoupených členů, funkčních genů a jejich rolí v ekosystémech. Možné řešení nabízí metatranskriptomika, která místo veškeré DNA je zaměřena pouze na geny v daném čase a prostředí exprimované (Franzosa et al., 2015). Další možností je zaměřit se jen na vybranou část konsorcií, tedy cílený (*targeted*) přístup.

V oblasti studia bioremediačních procesů je již delší dobu rozpracovávána metoda odhalení metabolicky aktivních mikroorganismů pomocí značeného substrátu, což umožňuje oddělit jejich DNA z celkového izolovaného vzorku (Dumont & Murrell, 2005). Další možností je sortování jednotlivých buněk mikroorganismů pomocí průtokové cytometrie, které vede k

rozdělení buněk do skupin či v konečném důsledku až izolaci jediné buňky (Müller & Nebe-von-Caron, 2010). Takový přístup se nazývá jako *single cell* a umožňuje sekvenování, assembly a analýzu jednotlivých organismů z nejrůznějších ekosystémů (Lasken & McLean, 2014). Při použití NGS platforem představuje takový redukovaný materiál příliš malý zdroj DNA. Jedna z možností je snaha o vyladění postupu pro získání co nejmenšího množství DNA ještě aplikovatelného pro úspěšnou tvorbu sekvenačních knihoven (Džunková et al., 2014). Pokud je výchozím materiálem jen jediná buňka, je nezbytný krok amplifikace DNA. Za účelem namnožení dlouhých molekul genomové DNA byla vyvinuta metoda MDA (*multiple displacement amplification*) (souhrnně Lasken & McLean, 2014). Jakékoli dodatečné kroky v přípravě vzorku pro sekvenování, obzvláště pak ty zahrnující amplifikaci, jsou ovšem častým zdrojem sekvenčních chyb. Našim vlastním zajímavým zjištěním při použití MDA pro amplifikaci DNA z environmetálního vzorku bylo také to, že v datech byly nabohaceny sekvence pocházejí z plasmidů, které krok amplifikace upřednostňoval (nepublikovaná data).

Tento problém efektivně obchází sekvenátory třetí generace umožňující sekvenaci jednotlivých molekul DNA. Ačkoli odpadají kroky přípravy sekvenačních knihoven, vznik chyb se posouvá na úroveň samotné sekvenace. Přestože bylo dosaženo čtení dlouhých i několik tisícovek bazí, vzhledem k vyšší míře chybovosti (Laver et al., 2015) si zatím nelze představit získání dlouhých souvislých sekvencí bakteriálních genomů jen na základě jedné sekvenační analýzy molekuly DNA. Prozatím jedinou možností korekce chyb tak zůstává kombinace s daty produkovanými druhou generací sekvenátorů (Koren et al., 2012; Madoui et al., 2015). Přestože metody sekvenování jedné molekuly DNA jsou rozvíjeny již delší dobu, jejich implementace ve formě komerčně dostupných přístrojů je zatím novinkou. Možnosti využití v genomice a metagenomice jsou autorem této práce s napětím očekávány; především se to týká přístroje MinION, jehož kapesní rozměry mohou umožnit přesun sekvenačních analýz z laboratoře přímo do terénu.

7 Závěr

- Kombinace shotgunových a párových knihoven, jejich sekvenace s využitím GS FLX Titanium (Roche) a assembly doprovázené manuálním zaplněním mezer umožnily přečíst a rekonstruovat kompletní genom bakterie *Achromobacter xylosoxidans* A8. Genom sestává z jednoho chromosomu dlouhého 7 013 095 bp a dvou plasmidů pA81 (98 156 bp) a pA82 (247 895 bp).
- Bioinformatickými metodami jsme identifikovali celkem 6 815 oblastí kódujících geny, 5 620 z nich jsme anotovali na základě podobnosti ke genům ve veřejných databázích. Potvrdili jsme lokalizaci genů pro degradaci chlorbenzoátů na plasmidu pA81. Zároveň jsme identifikovali další geny s potenciálním využití v biodegradaci polutantů.
- Několik identifikovaných genů poskytuje bakterii *A. xylosoxidans* A8 rezistenci vůči těžkým kovům, což z něho činí vhodného kandidáta pro tvorbu bioremediačního nástroje. Vysoký obsah těžkých kovů často doprovází kontaminaci aromatickými polutanty a negativně ovlivňuje bioremediační potenciál mikrobiálních konsorcií.
- Testovaný postup zahrnující sekvenování shotgunových a párových knihoven jsme dále aplikovali při analýze mikrobiálního společenstva tvořícího krápníkové biofilmy v extrémně kyselém prostředí dolu ve Zlatých Horách. Prvotním 454 amplikonovým sekvenováním úseků 16S rRNA genu jsme stanovili složení této velmi jednoduché bakteriální komunity s dominantním zastoupením třech bakterií rodu *Ferrovum, Acidithiobacillus* a třetí dosud nepopsané bakterie.
- 454 sekvenováním shotgunových a párových knihoven a následným sestavením jsme získali 272 skafoldů o celkové délce necelých 9 mil. bp obsahujících 1 330 mezer. Pomocí sekvenování DNA a mRNA ze vzorku biofilmu s využitím přístroje MiSeq (Illumina) se nám podařilo zaplnit 816 mezer a doplnit tak sekvence o více než 1 mil. chybějících nukleotidů. S využitím kombinace různých metod jsme roztřídili sestavené sekvence do skupin, dvě skupiny jsme přiřadili genomům dvou dominantně zastoupených bakterií.
- Podařilo se nám získat téměř kompletní genomové sekvence bakterií označených jako „*Ferrovum myxofaciens*“ ZH7 a „*Acidithiobacillus*“ linie ZH7B ze smíšeného metagenomického vzorku bez nutnosti předcházející kultivace. Anotace

predikovaných genů umožňuje rekonstruovat fyziologické a metabolické vlastnosti jednotlivých organismů. Genomy bakterií „*Ferrovum myxofaciens*“ ZH7 a „*Acidithiobacillus*“ linie ZH7B představují unikátní kombinace genů vyvinuté adaptací na konkrétní ekosystém extrémně kyselé důlní vody ve Zlatých Horách.

- S využitím dat získaných sekvenováním mRNA jsme identifikovali skutečně exprimované geny. Mezi nejvíce exprimovanými jsou geny související s oxidací redukovaného železa jako zdroje energie a geny spojené s adaptací na nízké pH a oxidační stres, kterému jsou bakterie v důlní vodě vystaveny.
- Data velkou měrou připívají k poznatkům o některých prozatím málo zkoumaných acidofilních bakterií, převážně obtížně kultivovatelných linií rodu *Ferrovum*. Představují důležitý základ pro další analýzy dosud neodhalených metabolických drah, například navržené dráhy oxidace železa.
- Amplikonové a shotgunovové sekvenování jsme využili rovněž při výzkumu komplexních mikrobiálních společenstev v půdě kontaminované polychlorovanými bifenyly. Na základě sekvencí 16S rRNA genů a anotovaných shotgunových čtení jsme sestavili taxonomické a funkční profily mikrobiálních komunit v kontaminované zemině ovlivněné výsadbou vybraných druhů rostlin a aplikací minerálního hnojiva.
- Pro bioinformatickou komparativní analýzu rozdílně zastoupených taxonů a funkčních systémů jsme adaptovali metody původně navržené pro analýzu rozdílné míry exprese z RNA-seq dat. To umožnilo odhalit statisticky významný vliv kořenů rostlin na složení a funkční potenciál mikrobiálních společenstev, od kontrolního vzorku se nejvíce odlišovala populace z oblasti kořenů křenu. Aplikace hnojiva používaná pro podporu růstu rostlin neměla významný vliv na mikrobiální komunity v půdě.
- Tyto informace přispívají k pochopení vztahu rostlin a půdních mikroorganismů a mohou být dále využity pro bioremediační výzkumy.

Reference

Abraham, WR, Wenderoth, DF, Glasser, W (2005) Diversity of biphenyl degraders in a chlorobenzene polluted aquifer. Chemosphere 58, 529-533.

Ackert L (2012) Sergei Vinogradskii and the Cycle of Life: From the Thermodynamics of Life to Ecological Microbiology, 1850-1950. Netherlands: Springer Science & Business Media.

Afiahayati, Sato K, Sakakibara Y (2015) MetaVelvet-SL: an extension of the Velvet assembler to a de novo metagenomic assembler utilizing supervised learning. DNA Research 2, 69-77.

Achtman M, Wagner M (2008) Microbial diversity and the genetic nature of microbial species. Nature Review Microbiology 6, 431-440.

Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25, 3389-402.

Amouric A, Brochier-Armanet C, Johnson DB, Bonnefoy V, Hallberg KB (2011) Phylogenetic and genetic variation among Fe(II)-oxidizing acidithiobacilli supports the view that these comprise multiple species with different ferrous iron oxidation pathways. Microbiology 157, 111-22.

Appia-Ayme C, Guiliani N, Ratouchniak J, Bonnefoy V (1999) Characterization of an operon encoding two c-type cytochromes, an aa(3)-type cytochrome oxidase, and rusticyanin in Thiobacillus ferrooxidans ATCC 33020. Applied and Environmental Microbiology 65, 4781–4787.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics 25, 25-9.

Aujoulat F, Roge, F, Bourdier A, Lotthé A, Lamy B, Marchandin H, Jumas-Bilak E (2012) From environment to man: genome evolution and adaptation of human opportunistic bacterial pathogens. Genes (Basel). 3, 191–232.

Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O (2008) The RAST Server: rapid annotations using subsystems technology. BMC Genomics 9, 75.

Badalamenti JP, Hunter RC (2015) Complete Genome Sequence of Achromobacter xylosoxidans MN001, a Cystic Fibrosis Airway Isolate. Genome Announcements 3, e00947-15.

Badger JH, Olsen GJ (1999) CRITICA: coding region identification tool invoking comparative analysis. Molecular Biology and Evolution 16, 512-24.

Baker BJ, Banfield JF (2003) Microbial communities in acid mine drainage. FEMS Microbiology Ecology 44, 139–152.

Baker GC, Smith JJ, Cowan DA (2003) Review and re-analysis of domain-specific 16S primers. Journal of Microbiological Methods 55, 541-55.

Baldrian P, Kolařík M, Stursová M, Kopecký J, Valášková V, Větrovský T, Zifčáková L, Snajdr J, Rídl J, Vlček C, Voříšková J (2012) Active and total microbial communities in forest soil are largely different and highly stratified during decomposition. ISME Journal 6, 248-58.

Ballschmiter K, Zell M (1980) Analysis of polychlorinated bifenyls (PCB) by glass capillary gas chromatography. Fresenius' Zeitschrift für analytische Chemie 302, 20-31

Banks D, Younger PL, Arnesen R-T, Iversen ER, Banks SB (1997) Mine-water chemistry: the good, the bad and the ugly. Environmental Geology 32, 157–174.

Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M, Sonnhammer EL (2002) The Pfam protein families database. Nucleic Acids Research 30, 276-80.

Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S, Mauceli E, Berger B, Mesirov JP, Lander ES (2002) ARACHNE: a whole-genome shotgun assembler. Genome Research 12, 177-89.

Beliaev AS, Saffarini DA (1998) Shewanella putrefaciens mtrB encodes an outer membrane protein required for Fe(III) and Mn(IV) reduction. Journal of Bacteriology 180, 6292-7.

Benizri, E, Amiaud, B (2005) Relationship between plants and soil microbial communities in fertilized grasslands. Soil Biology and Biochemistry 37, 2055-2064.

Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW (2010) GenBank. Nucleic Acids Research 38, D46-51.

Berg, G, Smalla, K (2009) Plant species and soil type cooperatively shape the structure and function of microbial communities in the rhizosphere. FEMS Microbiology Ecology 68, 1-13.

Beveridge TJ (2001) Use of the gram stain in microbiology. Biotechnic & Histochemistry 76, 111-8.

Bikel S, Valdez-Lara A, Cornejo-Granados F, Rico K, Canizales-Quinteros S, Soberón X, Del Pozo-Yauner L, Ochoa-Leyva A (2015) Combining metagenomics, metatranscriptomics and viromics to explore novel microbial interactions: towards a systems-level understanding of human microbiome. Computational and Structural Biotechnology Journal 13, 390-401.

Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y (1997) The complete genome sequence of Escherichia coli K-12. Science 277, 1453-62.

Blevins SM, Bronze MS (2010) Robert Koch and the 'golden age' of bacteriology. International Journal of Infectious Diseases 14, e744-51.

Bodilis J, Nsigue-Meilo S, Besaury L, Quillet L (2012) Variable Copy Number, Intra-Genomic Heterogeneities and Lateral Transfers of the 16S rRNA Gene in Pseudomonas. PLoS ONE 7, e35647

Boisvert S, Laviolette F, Corbeil J (2010) Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. Journal of Computational Biology 17, 1519-33.

Boisvert S, Raymond F, Godzaridis E, Laviolette F, Corbeil J (2012) Ray Meta: scalable de novo metagenome assembly and profiling. Genome Biology 13, R122.

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30, 2114-20.

Borodovsky M, McIninch J (1993) GeneMark: parallel gene recognition for both DNA strands. Computers & Chemistry 17, 123-133.

Brady A, Salzberg S (2011) PhymmBL expanded: confidence scores, custom databases, parallelization and more. Nature Methods 8, 367.

Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, Wilkins MJ, Wrighton KC, Williams KH, Banfield JF (2015) Unusual biology across a group comprising more than 15% of domain Bacteria. Nature 523, 208-11.

Byrd AK, Raney KD (2012) Superfamily 2 helicases. Frontiers in Bioscience (Landmark Ed) 17, 2070-88.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. BMC Bioinformatics 10, 421.

Cannon GC, Bradburne CE, Aldrich HC, Baker SH, Heinhorst S, Shively JM (2001) Microcompartments in prokaryotes: Carboxysomes and related polyhedra. Applied and Environmental Microbiology 67, 5351-5361.

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Peña AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Turnbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R (2010) QIIME allows analysis of high-throughput community sequencing data. Nature Methods 7, 335-6.

Carver T, Berriman M, Tivey A, Patel C, Böhme U, Barrell BG, Parkhill J, Rajandream MA (2008) Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. Bioinformatics 24, 2672-6.

Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, Brown CT, Porras-Alfaro A, Kuske CR, Tiedje JM (2014) Ribosomal Database Project: data and tools for high throughput rRNA analysis. Nucleic Acids Research 42, D633-D642.

Cole ST, Brosch R, Parkhill J, Garnier T, Churcher C, Harris D, Gordon SV, Eiglmeier K, Gas S, Barry CE 3rd, Tekaia F, Badcock K, Basham D, Brown D, Chillingworth T, Connor R, Davies R, Devlin K, Feltwell T, Gentles S, Hamlin N, Holroyd S, Hornsby T, Jagels K, Krogh A, McLean J, Moule S, Murphy L, Oliver K, Osborne J, Quail MA, Rajandream MA, Rogers J, Rutter S, Seeger K, Skelton J, Squares R, Squares S, Sulston JE, Taylor K, Whitehead S, Barrell BG (1998) Deciphering the biology of Mycobacterium tuberculosis from the complete genome sequence. Nature 393, 537-44.

Cole, JR, Wang, Q, Cardenas, E, Fish, J, Chai, B, Farris, RJ, Kulam-Syed-Mohideen, AS, Mcgarrell, DM, Marsh, T, Garrity, GM, Tiedje, JM (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. Nucleic acids research 37, D141-145.

Compeau PE, Pevzner PA, Tesler G (2011) How to apply de Bruijn graphs to genome assembly. Nature Biotechnology 29, 987-91.

Cox MP, Peterson DA, Biggs PJ (2010) SolexaQA: At-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics 11, 485.

Croal LR, Gralnick JA, Malasarn D, Newman DK (2004) The genetics of geochemistry. Annual Review of Genetics 38, 175-202.

De Filippo C, Ramazzotti M, Fontana P, Cavalieri D (2012) Bioinformatic approaches for functional annotation and pathway inference in metagenomics data. Briefings in Bioinformatics 13, 696-710.

Delcher AL, Harmon D, Kasif S, White O, Salzberg SL (1999) Improved microbial gene identification with GLIMMER. Nucleic Acids Research 27, 4636-41.

Dempsey MP, Nietfeldt J, Ravel J, Hinrichs S, Crawford R, Benson AK (2006) Paired-end sequence mapping detects extensive genomic rearrangement and translocation during divergence of Francisella tularensis subsp. tularensis and Francisella tularensis subsp. holarctica populations. Journal of Bacteriology 188, 5904-14.

Denisov G, Walenz B, Halpern AL, Miller J, Axelrod N, Levy S, Sutton G (2008) Consensus generation and variant detection by Celera Assembler. Bioinformatics 24, 1035-40.

Dennis, PG, Miller, AJ, Hirsch, PR (2010) Are root exudates more important than other sources of rhizodeposits in structuring rhizosphere bacterial communities? FEMS Microbiology Ecology 72, 313-327.

DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Applied Environmental Microbiology 72, 5069-72.

Diaz NN, Krause L, Goesmann A, Niehaus K, Nattkemper TW (2009) TACOA: taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. BMC Bioinformatics 10, 56.

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29, 15-21.

Dumont MG, Murrell JC (2005) Stable isotope probing - linking microbial identity to function. Nature Reviews Microbiology 3, 499-504.

Džunková M, Garcia-Garcerà M, Martínez-Priego L, D'Auria G, Calafell F, Moya A (2014) Direct squencing from the minimal number of DNA molecules needed to fill a 454 picotiterplate. PLoS One 9, e97379.

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27, 2194– 2200.

Edman P (1949) A method for the determination of amino acid sequence in peptides. Archives of Biochemistry 22, 475.

Edwards RA, Olson R, Disz T, Pusch GD, Vonstein V, Stevens R, Overbeek R (2012) Real time metagenomics: using k-mers to annotate metagenomes. Bioinformatics 28, 3316-7.

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S (2009) Real-time DNA sequencing from single polymerase molecules. Science 323, 133-8.

Ekblom R, Wolf JBW (2014) A field guide to whole-genome sequencing, assembly and annotation. Evolutionary Applications 7, 1026–1042.

Eren AM, Maignien L, Sul WJ, Murphy LG, Grim SL, Morrison HG, Sogin ML (2013) Oligotyping: Differentiating between closely related microbial taxa using 16S rRNA gene data. Methods in Ecology and Evolution 4. doi: 10.1111/2041-210X.12114.

Escobar-Zepeda A, Vera-Ponce de León A, Sanchez-Flores A (2015) The Road to Metagenomics: From Microbiology to DNA Sequencing Technologies and Bioinformatics. Frontiers in Genetics 6, 348.

Esparza M, Cárdenas JP, Bowien B, Jedlicki E, Holmes DS (2010) Genes and pathways for CO2 fixation in the obligate, chemolithoautotrophic acidophile, Acidithiobacillus ferrooxidans, carbon fixation in A. ferrooxidans. BMC Microbiology 10, 229.

Falteisek L, Cepička I (2012) Microbiology of diverse acidic and non-acidic microhabitats within a sulfidic ore mine. Extremophiles 16, 911-22.

Ferrer A, Bunk B, Spröer C, Biedendieck R, Valdés N, Jahn M, Jahn D, Orellana O, Levicán G (2016) Complete genome sequence of the bioleaching bacterium Leptospirillum sp. group II strain CF-1. Journal of Biotechnology 222, 21-2.

Fierer, N, Bradford, MA, Jackson, RB (2007) Toward an ecological classification of soil bacteria. Ecology 88, 1354-1364.

Fiers W, Contreras R, Duerinck F, Haegeman G, Iserentant D, Merregaert J, Min Jou W, Molemans F, Raeymaekers A, Van den Berghe A, Volckaert G, Ysebaert M (1976) Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. Nature 260, 500-7.

Fiore CL, Labrie M, Jarett JK, Lesser MP (2015) Transcriptional activity of the giant barrel sponge, Xestospongia muta Holobiont: molecular evidence for metabolic interchange. Frontiers in Microbiology 6, 364.

Fish JA, Chai B, Wang Q, Sun Y, Brown CT, Tiedje JM, Cole JR (2013) FunGene: the functional gene pipeline and repository. Frontiers Microbiology 4, 291.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM, Smith H, Venter G (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269, 496-512.

Fox GE, Wisotzkey JD, Jurtshuk P (1992) How close is close: 16S rRNA sequence identity may not be sufficient to guarantee species identity. International Journal of Systematic Bacteriology 42, 166–170.

Franzosa EA, Morgan XC, Segata N, Waldron L, Reyes J, Earl AM, Giannoukos G, Boylan MR, Ciulla D, Gevers D, Izard J, Garrett WS, Chan AT, Huttenhower C (2014) Relating the metatranscriptome and metagenome of the human gut. Proceedings of the National Academy of Sciences USA 111, E2329-38.

Franzosa EA, Hsu T, Sirota-Madi A, Shafquat A, Abu-Ali G, Morgan XC, Huttenhower C (2015) Sequencing and beyond: integrating molecular 'omics' for microbial community profiling. Nature Reviews, Microbiology 13, 360-72.

Furukawa K (2000) Biochemical and genetic bases of microbial degradation of polychlorinated biphenyls (PCBs). J Gen Appl Microbiol 46, 283–296.

Geisseler, D, Scow, KM (2014) Long-term effects of mineral fertilizers on soil microorganisms - A review. Soil Biology & Biochemistry 75, 54-63.

Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, Feil EJ, Stackebrandt E, Van de

Giannoukos G, Ciulla DM, Huang K, Haas BJ, Izard J, Levin JZ, Livny J, Earl AM, Gevers D, Ward DV, Nusbaum C, Birren BW, Gnirke A (2012) Efficient and robust RNA-seq process for cultured bacteria and complex community transcriptomes. Genome Biology 13, R23.

Gilbert, ES, Crowley, DE (1997) Plant compounds that induce polychlorinated biphenyl biodegradation by Arthrobacter sp. strain B1B. Applied and Environmental Microbiology 63, 1933-1938.

Ginard M, Lalucat J, Tummler B, Romling U (1997) Genome organization of Pseudomonas stutzeri and resulting taxonomic and evolutionary considerations. International Journal of Systematic Bacteriology 47, 132–143.

Giovannoni SJ, Britschgi TB, Moyer CL, Field KG (1990) Genetic diversity in Sargasso Sea bacterioplankton. Nature 345, 60-3.

Gnerre S, Maccallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, Berlin AM, Aird D, Costello M, Daza R, Williams L, Nicol R, Gnirke A, Nusbaum C, Lander ES, Jaffe DB (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proceedings of the National Academy of Sciences USA 108, 1513-8.

Goll J, Rusch DB, Tanenbaum DM, Thiagarajan M, Li K, Methé BA, Yooseph S (2010) METAREP: JCVI metagenomics reports--an open source tool for high-performance comparative metagenomics. Bioinformatics 26, 2631-2.

Gomez-Alvarez V, Teal TK, Schmidt TM (2009) Systematic artifacts in metagenomes from complex microbial communities. ISME Journal 3, 1314-7.

Gontcharova V, Youn E, Wolcott RD, Hollister EB, Gentry TJ (2010) Black Box Chimera Check (B2C2): a Windows-based software for batch depletion of chimeras from bacterial 16 S rRNA gene datasets. The Open Microbiology Journal 4, 47–52.

Gonzalez-Toril E, Llobet-Brossa E, Casamayor EO, Amann R, Amils R (2003) Microbial Ecology of an Extreme Acidic Environment, the Tinto River. Applied and Environmental Microbiology 69, 4853–4865.

Gordon D, Green P (2013) Consed: a graphical editor for next-generation sequencing. Bioinformatics 29, 2936-7.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nature Biotechnology 29, 644–652.

Guiliani N, Jerez CA (2000) Molecular cloning, sequencing, and expression of omp-40, the gene coding for the major outer membrane protein from the acidophilic bacterium Thiobacillus ferrooxidans. Applied and Environmental Microbiology 66, 2318-24.

Guttenplan SB, Kearns DB (2013) Regulation of flagellar motility during biofilm formation. FEMS Microbiology Reviews 37, 849-71.

Haas BJ, Gevers D, Earl AM, Feldgarden M, Ward DV, Giannoukos G, Ciulla D, Tabbaa D, Highlander SK, Sodergren E, Methé B, DeSantis TZ; Human Microbiome Consortium, Petrosino JF, Knight R, Birren BW (2011) Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. Genome Research 21, 494-504.

Haft DH, Selengut JD, White O (2003) The TIGRFAMs database of protein families. Nucleic Acids Research 31, 371-3.

Hallberg KB, Amouric A, Brochier-Armanet C, Bonnefoy V, Johnson DB (2009) Physiological and Phylogenetic Heterogeneity among Iron-Oxidizing Acidithiobacillus spp., and Characteristics of the Novel Species Acidithiobacillus Ferrivorans. Advanced Materials Research 71, 167-170.

Hallberg KB, González-Toril E, Johnson DB (2010) Acidithiobacillus ferrivorans, sp. nov.; facultatively anaerobic, psychrotolerant iron-, and sulfur-oxidizing acidophiles isolated from metal mine-impacted environments. Extremophiles 14, 9-19.

Hanage WP, Fraser C, Spratt BG (2006) Sequences, sequence clusters and bacterial species. Philosophical transactions of the Royal Society of London. Series B 361, 1917-1927.

Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. Chemistry & Biology 5, R245-9.

Hartmann, A, Schmid, M, Van Tuinen, D, Berg, G (2009) Plant-driven selection of microbes. Plant and Soil 321, 235-257.

Havlak P, Chen R, Durbin KJ, Egan A, Ren Y, Song XZ, Weinstock GM, Gibbs RA (2004) The Atlas genome assembly system. Genome Research 14, 721-32.

Hayes F (2000) The partition system of multidrug resistance plasmid TP228 includes a novel protein that epitomizes an evolutionarily distinct subgroup of the ParA superfamily. Molecular Microbiology 37, 528-41.

Heather JM, Chain B (2016) The sequence of sequencers: The history of sequencing DNA. Genomics 107, 1-8.

Hložková K, Šuman J, Strnad H, Ruml T, Pačes V, Kotrba P (2013) Characterization of pbt genes conferring increased Pb^{2+} and Cd^{2+} tolerance upon Achromobacter xylosoxidans A8. Research in Microbiology 164, 1009-18.

Hoff K, Lingner T, Meinicke P, Tech MO (2009) Predicting genes in metagenomic sequencing reads. Nucleic Acids Research 37, W101–105.

Holmes B, Snell JJS, Lapage SP (1977) Strains of Achromobacter xylosoxidans from clinical material. Journal of Clinical Pathology 30, 595-601.

Horvath P, Barrangou R (2010) CRISPR/Cas, the immune system of bacteria and archaea. Science 327, 167-70.

Hua ZS, Han YJ, Chen LX, Liu J, Hu M, Li SJ, Kuang JL, Chain PS, Huang LN, Shu WS (2015) Ecological roles of dominant and rare prokaryotes in acid mine drainage revealed by metagenomics and metatranscriptomics. ISME Journal 9, 1280-94.

Huerta-Cepas J, Szklarczyk D, Forslund K, Cook H, Heller D, Walter MC, Rattei T, Mende DR, Sunagawa S, Kuhn M, Jensen LJ, von Mering C, Bork P (2016) eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. Nucleic Acids Research 44, D286-93.

Huber T, Faulkner G, Hugenholtz P (2004) Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. Bioinformatics 20, 2317–2319.

Huson DH, Weber N (2013) Microbial community analysis using MEGAN. Methods in Enzymology 531, 465-85.

Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics 11, 119.

Chain, PS, Denef, VJ, Konstantinidis, KT, Vergez, LM, Agullo, L, Reyes, VL, Hauser, L, Cordova, M, Gomez, L, Gonzalez, M, Land, M, Lao, V, Larimer, F, Lipuma, JJ, Mahenthiralingam, E, Malfatti, SA, Marx, CJ, Parnell, JJ, Ramette, A, Richardson, P, Seeger, M, Smith, D, Spilker, T, Sul, WJ, Tsoi, TV, Ulrich, LE, Zhulin, IB, Tiedje, JM (2006) Burkholderia xenovorans LB400 harbors a multi-replicon, 9.73-Mbp genome shaped for versatility. Proceedings of the National Academy of Sciences of the United States of America 103, 15280-15287.

Chan CK, Hsu AL, Halgamuge SK, Tang SL (2002) Binning sequences using very sparse labels within a metagenome. BMC Bioinformatics 9, 215.

Chao A, Chazdon RL, Colwell RK, Shen TJ (2006) Abundance-based similarity indices and their estimation when there are unseen species in samples. Biometrics 62, 361-71.

Chaudhry, Q, Blom-Zandstra, M, Gupta, S, Joner, EJ (2005) Utilising the synergy between plants and rhizosphere microorganisms to enhance breakdown of organic pollutants in the environment. Environmental Science and Pollution Research 12, 34-48.

Chevreux B, Wetter T, Suhai S (1999) Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB) 99, 45-56.

Ilbert M, Bonnefoy V (2013) Insight into the evolution of the iron oxidation pathways. Biochimica at Biophysica Acta 1827, 161-75.

Ionescu, M, Beranová, K, Dudková, V, Kochánková, L, Demnerová, K, Macek, T, Macková, M (2009) Isolation and characterization of different plant associated bacteria and their potential to degrade polychlorinated biphenyls. International Biodeterioration & Biodegradation 63, 667-672.

Jenčová V, Strnad H, Chodora Z, Ulbrich P, Vlček C, Hickey WJ, Pačes V (2004) Chlorocatechol catabolic enzymes from Achromobacter xylosoxidans A8. International Biodeterioration & Biodegradation 54, 175–181.

Jenčová V, Strnad H, Chodora Z, Ulbrich P, Vlček C, Hickey WJ, Pačes V (2008) Nucleotide sequence, organization and characterization of the (halo)aromatic acid catabolic plasmid pA81 from Achromobacter xylosoxidans A8. Research in Microbiology 159, 118-27.

Jiao X, Zheng X, Ma L, Kutty G, Gogineni E, Sun Q, Sherman BT, Hu X, Jones K, Raley C, Tran B, Munroe DJ, Stephens R, Liang D, Imamichi T, Kovacs JA, Lempicki RA, Huang DW (2013) A Benchmark Study on Error Assessment and Quality Control of CCS Reads Derived from the PacBio RS. Journal of Data Mining in Genomics & Proteomics 4, pii: 16008.

Jain M, Fiddes IT, Miga KH, Olsen HE, Paten B, Akeson M (2015) Improved data analysis for the MinION nanopore sequencer. Nature Methods 12, 351-6.

Jiao Y, Newman DK (2007) The pio operon is essential for phototrophic Fe(II) oxidation in Rhodopseudomonas palustris TIE-1. Journal of Bacteriology 189, 1765-73.

Johnson DB (1995) Acidophilic microbial communities: Candidates for bioremediation of acidic mine effluents. International Biodeterioration & Biodegradation 35, 41–58.

Johnson DB (2012) Geomicrobiology of extremely acidic subsurface environments. FEMS Microbiology Ecology 81, 2-12.

Johnson DB, Hallberg KB (2005) Acid mine drainage remediation options: a review. The Science of the Total Environment 338, 3–14.

Johnson DB, Hallberg KB, Hedrich S (2014) Uncovering a microbial enigma: isolation and characterization of the streamer-generating, iron-oxidizing, acidophilic bacterium "Ferrovum myxofaciens". Applied and Environmental Microbiology 80, 672–680.

Johnson, DB (2009) Extremophiles: acidic environments. In Encyclopaedia of Microbiology, 2nd Edition; Schaechter, M. Ed.; Elsevier: Oxford, UK, pp. 107-126.

Jones DS, Kohl C, Grettenberger C, Larson LN, Burgos WD, Macalady JL (2015) Geochemical niches of iron-oxidizing acidophiles in acidic coal mine drainage. Applied and Environmental Microbiology 81, 1242– 1250.

Jones, D, Nguyen, C, Finlay, R (2009) Carbon flow in the rhizosphere: carbon trading at the soil–root interface. Plant and Soil 321, 5-33.

Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y (2008) KEGG for linking genomes to life and the environment. Nucleic Acids Research 36, D480-4.

Keegan KP, Trimble WL, Wilkening J, Wilke A, Harrison T, D'Souza M, Meyer F (2012) A platform-independent method for detecting errors in metagenomic sequencing data: DRISEE. PLoS Computational Biology 8, e1002541.

Kelley DR, Liu B, Delcher AL, Pop M, Salzberg SL (2012) Gene prediction with Glimmer for metagenomic sequences augmented by classification and clustering. Nucleic Acids Research 40, e9.

Kim S, Picardal F (2001) Microbial growth on dichlorobiphenyls chlorinated on both rings as a sole carbon and energy source. Applied and Environmental Microbiology 67, 1953–1955.

Kimelman A, Levy A, Sberro H, Kidron S, Leavitt A, Amitai G, Yoder-Himes DR, Wurtzel O, Zhu Y, Rubin EM, Sorek R (2012) A vast collection of microbial genes that are toxic to bacteria. Genome Research 22, 802-9.

Kirk JL, Beaudette LA, Hart M, Moutoglis P, Klironomos JN, Lee H, Trevors JT (2004) Methods of studying soil microbial diversity. Journal of Microbiological Methods 58, 169-188.

Kopylova E, Noé L, Touzet H (2012) SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. Bioinformatics 28, 3211-7.

Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED, Adam M Phillippy (2012) Hybrid error correction and de novo assembly of single-molecule sequencing reads. Nature Biotechnology 30, 693-700.

Kotris J (2004) Segmentation of the ore district and overview of the mining performed by Ore Mines Jeseník. In: Pecina V, Večeřa J (eds) Zlaté Hory mine district, proceedings of international conference. Czech geological survey, Jeseník, pp 64–73

Koubek, J, Uhlík, O, Ječná, K, Junková, P, Vrkoslavová, J, Lipov, J, Kurzawová, V, Macek, T, Macková, M (2012) Whole-cell MALDI-TOF: Rapid screening method in environmental microbiology. International Biodeterioration & Biodegradation 69, 82-86.

Kowalchuk, GA, Buma, DS, De Boer, W, Klinkhamer, PGL, Van Veen, JA (2002) Effects of above-ground plant species composition and diversity on the diversity of soil-borne microorganisms. Antonie van Leeuwenhoek 81, 509-520.

Krause A, Ramakumar A, Bartels D, Battistoni F, Bekel T, Boch J, Böhm M, Friedrich F, Hurek T, Krause L, Linke B, McHardy AC, Sarkar A, Schneiker S, Syed AA, Thauer R, Vorhölter FJ, Weidner S, Pühler A, Reinhold-Hurek B, Kaiser O, Goesmann A (2006) Complete genome of the mutualistic, N2-fixing grass endophyte Azoarcus sp. strain BH72. Nature Biotechnology 24, 1385-91.

Krause L, Diaz NN, Goesmann A, Kelley S, Nattkemper TW, Rohwer F, Edwards RA, Stoye J (2008) Phylogenetic classification of short environmental DNA fragments. Nucleic Acids Research 36, 2230-9.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL (2004) Versatile and open software for comparing large genomes. Genome Biology 5, R12.

Kurzawová, V, Štursa, P, Uhlík, O, Norková, K, Strohalm, M, Lipov, J, Kochánková, L, Macková, M (2012) Plant-microorganism interactions in bioremediation of polychlorinated biphenyl-contaminated soil. New Biotechnology 30, 15-22.

Lagesen K, Hallin P, Rødland EA, Staerfeldt HH, Rognes T, Ussery DW (2007) RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Research 35, 3100-8.

Lambers, H, Mougel, C, Jaillard, B, Hinsinger, P (2009) Plant-microbe-soil interactions in the rhizosphere: an evolutionary perspective. Plant and Soil 321, 83-115.

Land M, Hauser L, Jun SR, Nookaew I, Leuze MR, Ahn TH, Karpinets T, Lund O, Kora G, Wassenaar T, Poudel S, Ussery DW (2015) Insights from 20 years of bacterial genome sequencing. Functional & Integrative Genomics 15, 141-61.

Lane, DJ (1991) "16S/23S rRNA sequencing," in Nucleic acid techniques in bacterial systematics, eds. E. Stackebrandt & M. Goodfellow. (New York, NY: John Wiley and Sons) 115-175.

Langille MG, Zaneveld J, Caporaso JG, McDonald D, Knights D, Reyes JA, Clemente JC, Burkepile DE, Vega Thurber RL, Knight R, Beiko RG, Huttenhower C (2013) Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. Nature Biotechnology 31, 814-21.

Langklotz S, Baumann U, Narberhaus F (2012) Structure and function of the bacterial AAA protease FtsH. Biochimica at Biophysica Acta 1823, 40-8.

Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nature Methods 9, 357-9.

Lasken RS, McLean JS (2014) Recent advances in genomic DNA sequencing of microbial species from single cells. Nature Reviews Genetics 15, 577-84.

Laslett D, Canback B (2004) ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Research 32, 11-6.

Lauro, FM, Mcdougald, D, Thomas, T, Williams, TJ, Egan, S, Rice, S, Demaere, MZ, Ting, L, Ertan, H, Johnson, J, Ferriera, S, Lapidus, A, Anderson, I, Kyrpides, N, Munk, AC, Detter, C, Han, CS, Brown, MV, Robb, FT, Kjelleberg, S, Cavicchioli, R (2009) The genomic basis of trophic strategy in marine bacteria. Proceedings of the National Academy of Sciences 106, 15527-15533.

Laver T, Harrison J, O'Neill PA, Moore K, Farbos A, Paszkiewicz K, Studholme DJ (2015) Assessing the performance of the Oxford Nanopore Technologies MinION. Biomolecular Detection and Quantification 3, 1-8.

Le HS, Schulz MH, McCauley BM, Hinman VF, Bar-Joseph Z (2013) Probabilistic error correction for RNA sequencing. Nucleic Acids Research 41, e109.

Lee I, Suzuki CK (2008) Functional mechanics of the ATP-dependent Lon protease- lessons from endogenous protein and synthetic peptide substrates. Biochimca et Biophysica Acta 1784, 727-35.

Lee, KCY, Morgan, XC, Power, JF, Dunfield, PF, Huttenhower, C, Stott, MB (2015) Complete genome sequence of the thermophilic Acidobacteria, Pyrinomonas methylaliphatogenes type strain K22T. Standards in Genomic Sciences 10, 1-8.

Leewis, M-C, Uhlík, O, Fraraccio, S, Mcfarlin, K, Kottara, A, Glover, C, Macek, T, Leigh, MB (2016a) Differential impacts of willow and mineral fertilizer on bacterial communities and biodegradation in diesel fuel oil-contaminated soil. Frontiers in Microbiology 7, 837.

Leewis, M-C, Uhlík, O, Leigh, MB (2016b) Synergistic Processing of Biphenyl and Benzoate: Carbon Flow Through the Bacterial Community in Polychlorinated-Biphenyl-Contaminated Soil. Scientific Reports 6, 22145.

Leff, JW, Jones, SE, Prober, SM, Barberán, A, Borer, ET, Firn, JL, Harpole, WS, Hobbie, SE, Hofmockel, KS, Knops, JMH, Mcculley, RL, La Pierre, K, Risch, AC, Seabloom, EW, Schütz, M, Steenbock, C, Stevens, CJ, Fierer, N (2015) Consistent responses of soil microbial communities to elevated nutrient inputs in grasslands across the globe. Proceedings of the National Academy of Sciences 112, 10967-10972.

Leigh MB (2006) Methods for rhizoremediation research. Approaches to experimental design and microbial analysis. In: Macková M, Dowling D, Macek T, eds. Phytoremediation and Rhizoremediation. Theoretical Background. Springer, Dordrecht, Netherlands, pp. 33- 55.

Leigh, MB, Fletcher, JS, Fu, X, Schmitz, FJ (2002) Root turnover: an important source of microbial substrates in rhizosphere remediation of recalcitrant contaminants. Environmental Science & Technology 36, 1579-1583.

Leigh, MB, Pellizari, VH, Uhlík, O, Sutka, R, Rodrigues, J, Ostrom, NE, Zhou, J, Tiedje, JM (2007) Biphenyl-utilizing bacteria and their functional genes in a pine root zone contaminated with polychlorinated biphenyls (PCBs) ISME Journal 1, 134-148.

Leigh, MB, Prouzová, P, Macková, M, Macek, T, Nagle, DP, Fletcher, JS (2006) Polychlorinated biphenyl (PCB)degrading bacteria associated with trees in a PCB-contaminated site. Applied and Environmental Microbiology 72, 2331-2342.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25, 1754-60.

Li K, Bihan M, Yooseph S, Methé BA (2012) Analyses of the Microbial Diversity across the Human Microbiome. PLoS One 7, e32118.

Liliensiek, A-K, Thakuria, D, Clipson, N (2012) Influences of Plant Species Composition, Fertilisation and Lolium perenne Ingression on Soil Microbial Community Structure in Three Irish Grasslands. Microbial Ecology 63, 509-521.

Liste, HH, Alexander, M (2000) Accumulation of phenanthrene and pyrene in rhizosphere soil. Chemosphere 40, 11-14.

Liu B, Gibbons T, Ghodsi M, Treangen T, Pop M (2011) Accurate and fast estimation of taxonomic profiles from metagenomic shotgun sequences. BMC Genomics 12, S4.

Liu WT, Marsh TL, Cheng H, Forney LJ. Characterization of microbial diversity by determining terminal restriction fragment length polymorphisms of genes encoding 16S rRNA (1997) Applied Environmental Microbiology 63, 4516-4522.

Lobry JR (1996) A simple vectorial representation of DNA sequences for the detection of replication origins in bacteria. Biochimie 78, 323-6.

Loman NJ, Pallen MJ (2015) Twenty years of bacterial genome sequencing. Nature Review Microbiology 13, 787-94.

Love, M, Huber, W, Anders, S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biology 15, 550.

Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Research 25, 955-64.

Lozupone CA, Knight R (2008) Species divergence and the measurement of microbial diversity. FEMS Microbiology Review 32, 557-78.

Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu SM, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam TW, Wang J (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1, 18.

Macek, T, Macková, M, Káš, J (2000) Exploitation of plants for the removal of organics in environmental remediation. Biotechnology Advances 18, 23-34. Macková, M., Dowling, D., and Macek, T. (eds.) (2006) Phytoremediation and Rhizoremediation. Theoretical Background. Dordrecht, Netherlands: Springer.

Macková, M, Uhlík, O, Lovecká, P, Viktorová, J, Nováková, M, Demnerová, K, Sylvestre, M, Macek, T (2010) "Bacterial Degradation of Polychlorinated Biphenyls," in Geomicrobiology: Molecular and Environmental Perspective, eds. A. Loy, M. Mandl & L.L. Barton. (Dordrecht, Netherlands: Springer) 347-366.

Madoui MA, Engelen S, Cruaud C, Belser C, Bertrand L, Alberti A, Lemainque A, Wincker P, Aury JM. (2015) Genome assembly using Nanopore-guided long and error-free DNA reads. BMC Genomics 16, 327.

Madsen EL (2008) Environmental microbiology. From genomes to biogeochemistry. Blackwell Publishing, Malden, MA, USA, 479 p.

Magoc T, Pabinger S, Canzar S, Liu X, Su Q, Puiu D, Tallon LJ, Salzberg SL (2013) GAGE-B: an evaluation of genome assemblers for bacterial organisms. Bioinformatics 29, 1718-25.

Magurran AE (2004) Measuring biological diversity. Blackwell Publishing, Malden, MA, USA, 256 p.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM (2005) Genome sequencing in microfabricated high-density picolitre reactors. Nature 437, 376-80.

Markowitz VM, Chen IM, Chu K, Szeto E, Palaniappan K, Pillay M, Ratner A, Huang J, Pagani I, Tringe S, Huntemann M, Billis K, Varghese N, Tennessen K, Mavromatis K, Pati A, Ivanova NN, Kyrpides NC (2014) IMG/M 4 version of the integrated metagenome comparative analysis system. Nucleic Acids Research 42, D568-73.

Marsh TL (2005) Culture-independent microbial community analysis with terminal restriction fragment length polymorphism. Methods in Enzymology 397, 308-329.

Marschner, P, Yang, CH, Lieberei, R, Crowley, DE (2001) Soil and plant specific effects on bacterial community composition in the rhizosphere. Soil Biology and Biochemistry 33, 1437-1445.

Martínez-Abarca F, Toro N (2000) Group II introns in the bacterial world. Molecular Microbiology 38, 917-26.

Maxam AM, Gilbert W (1977) A new method for sequencing DNA. Proceedings of the National Academy of Sciences USA 74, 560–4.

McHardy AC, Martín HG, Tsirigos A, Hugenholtz P, Rigoutsos I (2007) Accurate phylogenetic classification of variable-length DNA fragments. Nature Methods 4, 63-72.

McMurdie PJ, Holmes S (2013) phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data. PLoS One 8, e61217.

Mcmurdie, PJ, Holmes, S (2014) Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. PLoS Computational Biology 10, e1003531.

Méndez-García C, Peláez AI, Mesa V, Sánchez J, Golyshina OV, Ferrer M (2015) Microbial diversity and metabolic networks in acid mine drainage habitats. Frontiers in Microbiology 6, 475.

Meyer, F, Paarmann, D, D'souza, M, Olson, R, Glass, E, Kubal, M, Paczian, T, Rodriguez, A, Stevens, R, Wilke, A, Wilkening, J, Edwards, R (2008) The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics 9, 386.

Milne I, Stephen G, Bayer M, Cock PJ, Pritchard L, Cardle L, Shaw PD, Marshall D (2012) Using Tablet for visual exploration of second-generation sequencing data. Briefings in Bioinformatics 14, 193-202.

Monzoorul Haque M, Ghosh TS, Komanduri D, Mande SS (2009) SOrt-ITEMS: Sequence orthology based approach for improved taxonomic estimation of metagenomic sequences. Bioinformatics 25, 1722-30.

Moran MA, Satinsky B, Gifford SM, Luo H, Rivers A, Chan LK, Meng J, Durham BP, Shen C, Varaljay VA, Smith CB, Yager PL, Hopkinson BM (2013) Sizing up metatranscriptomics. ISME Journal 7, 237-43

Mori JF, Lu S2, Händel M3, Totsche KU3, Neu TR4, Iancu VV5, Tarcea N5, Popp J6, Küsel K7 (2016) Schwertmannite formation at cell junctions by a new filament-forming Fe(II)-oxidizing isolate affiliated with the novel genus Acidithrix. Microbiology 162, 62-71.

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M (2007) KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Research 35, W182-5.

Moya-Beltrán A, Cárdenas JP, Covarrubias PC, Issotta F, Ossandon FJ, Grail BM, Holmes DS, Quatrini R, Johnson DB (2014) Draft Genome Sequence of the Nominated Type Strain of "Ferrovum myxofaciens," an Acidophilic, Iron-Oxidizing Betaproteobacterium. Genome Announcements 2, e00834-14

Mukerjee-Dhar, G, Hatta, T, Shimura, M, Kimbara, K (1998) Analysis of changes in congener selectivity during PCB degradation by Burkholderia sp. strain TSN101 with increasing concentrations of PCB and characterization of the bphBCD genes and gene products. Archives of Microbiology 169, 61-70.

Mukherjee, S, Juottonen, H, Siivonen, P, Lloret Quesada, C, Tuomi, P, Pulkkinen, P, Yrjala, K (2014) Spatial patterns of microbial diversity and activity in an aged creosote-contaminated site. The ISME Journal 8, 2131-2142.

Müller S, Nebe-von-Caron G (2010) Functional single-cell analyses: flow cytometry and cell sorting of microbial populations and communities. FEMS Microbiology Reviews 34, 554-87.

Nagarajan N, Pop M (2013) Sequence assembly demystified. Nature Review Genetics 14, 157-67.

Namiki T, Hachiya T, Tanaka H, Sakakibara Y (2012) MetaVelvet: an extension of Velvet assembler to de novo metagenome assembly from short sequence reads. Nucleic Acids Research 40, e155.

Neidhardt FC (1996) Escherichia Coli and Salmonella: Cellular and Molecular Biology. ASM Press: Washington, DC.

Neumann, G, George, TS, Plassard, C (2009) Strategies and methods for studying the rhizosphere—the plant science toolbox. Plant and Soil 321, 431-456.

Nguyen N-P, Warnow T, Pop M, White B (2016) A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity. Biofilms and Microbiomes 2, 16004.

Ni, J, Yan, Q, Yu, Y (2013) How much metagenomic sequencing is enough to achieve a given goal? Scientific Reports 3, 1968.

Niu B, Fu L, Sun S, Li W (2010) Artificial and natural duplicates in pyrosequencing reads of metagenomic data. BMC Bioinformatics 11, 187.

Noguchi H, Park J, Takagi T (2006) MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. Nucleic Acids Research 34, 5623-5630.

Noguchi H, Taniguchi T, Itoh T (2008) MetaGeneAnnotator: detecting species-specific patterns of ribosomal binding site for precise gene prediction in anonymous prokaryotic and phage genomes. DNA Research 15, 387–396.

Nunes Da Rocha, U, Plugge, CM, George, I, Van Elsas, JD, Van Overbeek, LS (2013) The Rhizosphere Selects for Particular Groups of Acidobacteria and Verrucomicrobia. PLoS ONE 8, e82443.

Ohmori H, Tomizawa JI, Maxam AM (1978) Detection of 5-methylcytosine in DNA sequences. Nucleic Acids Research 5, 1479-85.

Oksanen, J, Blanchet, FG, Kindt, R, Legendre, P, Minchin, PR, O'hara, RB, Simpson, GL, Solymos, P, Stevens, MHH, Wagner, H (2013) vegan: Community Ecology Package [Online]. Available: http://CRAN.R-project.org/package=vegan.

Ondov BD, Bergman NH, Phillippy AM (2011) Interactive Metagenomic Visualization in a Web Browser. BMC Bioinformatics 12, 385.

Ormerod KL, George NM, Fraser JA, Wainwright C, Hugenholtz P (2015) Comparative genomics of non-pseudomonal bacterial species colonising paediatric cystic fibrosis patients. PeerJ 3, e1223.

Oulas A, Pavloudi C, Polymenakou P, Pavlopoulos GA, Papanikolaou N, Kotoulas G, Arvanitidis C, Iliopoulos I (2015) Metagenomics: tools and insights for analyzing next-generation sequencing data derived from biodiversity studies. Bioinformatics and Biology Insights 9, 75-88.

Pačes J, Pačes V (2002) DicodonUse: the programme for dicodon bias visualization in prokaryotes. Folia Biologica (Praha) 48, 246-9.

Parkinson NJ, Maslau S, Ferneyhough B, Zhang G, Gregory L, Buck D, Ragoussis J, Ponting CP, Fischer MD (2012) Preparation of high-quality next-generation sequencing libraries from picogram quantities of target DNA. Genome Research 22, 125-33.

Parks DH, Tyson GW, Hugenholtz P, Beiko RG. (2014) STAMP: Statistical analysis of taxonomic and functional profiles. Bioinformatics 30, 3123-3124.

Patel RK, Jain M (2012) NGS QC Toolkit: a toolkit for quality control of next generation sequencing data. PLoS One 7, :e30619.

Pati A, Heath LS, Kyrpides NC, Ivanova N (2011) ClaMS: a classifier for metagenomic sequences. Standards in Genomic Sciences 5, 248–53.

Patil KR, Roune L, McHardy AC (2012) The PhyloPythiaS web server for taxonomic assignment of metagenome sequences. PLoS One 7, e38581.

Paulson JN, Stine OC, Bravo HC, Pop M (2013) Differential abundance analysis for microbial marker-gene surveys. Nature Methods 10, 1200-2.

Pavlíková, D, Macek, T, Macková, M, Pavlík, M (2007) Monitoring native vegetation on a dumpsite of PCB-contaminated soil. International Journal of Phytoremediation 9, 71-78.

Pavlů L, Vosáhlová J, Klierová H, Prouza M, Demnerová K, Brenner V (1999) Characterization of chlorobenzoate degraders isolated from polychlorinated biphenyl-contaminated soil and sediment in the Czech Republic. Journal of Applied Microbiology 87, 381-6.

Pearce DA, Newsham KK, Thorne MA (2012) Metagenomic analysis of a southern maritime antarctic soil. Frontiers in Microbiology 3, 403–403.

Peer Y, Vandamme P, Thompson FL, Swings J (2005) Re-evaluating prokaryotic species. Nature Review Microbiology 3, 733-739.

Peng Y, Leung HC, Yiu SM, Chin FY (2011) Meta-IDBA: a de Novo assembler for metagenomic data. Bioinformatics 27, i94-101.

Peng Y, Leung HC, Yiu SM, Chin FY (2012) IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28, 1420-8.

Pham, TT, Pino Rodriguez, NJ, Hijri, M, Sylvestre, M (2015) Optimizing Polychlorinated Biphenyl Degradation by Flavonoid-Induced Cells of the Rhizobacterium Rhodococcus erythropolis U23A. PLoS One 10, e0126033.

Pham, TTM, Tu, Y, Sylvestre, M (2012) Remarkable ability of Pandoraea pnomenusa B356 biphenyl dioxygenase to metabolize simple flavonoids. Applied and Environmental Microbiology 78, 3560-3570.

Philippot, L, Raaijmakers, JM, Lemanceau, P, Van Der Putten, WH (2013) Going back to the roots: the microbial ecology of the rhizosphere. Nature Reviews Microbiology 11, 789-799.

Pieper DH, Martins dos Santos VA, Golyshin PN (2004) Genomic and mechanistic insights into the biodegradation of organic pollutants. Current Opinnion in Biotechnology 15, 215-24.

Pieper DH, Seeger M (2008) Bacterial metabolism of polychlorinated biphenyls. Journal of Molecular Microbiology and Biotechnology 15, 121-38.

Prober, SM, Leff, JW, Bates, ST, Borer, ET, Firn, J, Harpole, WS, Lind, EM, Seabloom, EW, Adler, PB, Bakker, JD, Cleland, EE, Decrappeo, NM, Delorenze, E, Hagenah, N, Hautier, Y, Hofmockel, KS, Kirkman, KP, Knops, JMH, La Pierre, KJ, Macdougall, AS, Mcculley, RL, Mitchell, CE, Risch, AC, Schuetz, M, Stevens, CJ, Williams, RJ, Fierer, N (2015) Plant diversity predicts beta but not alpha diversity of soil microbes across grasslands worldwide. Ecology Letters 18, 85-95.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. Nucleic Acids Research 41, D590-6.

Quatrini R, Appia-Ayme C, Denis Y, Jedlicki E, Holmes DS, Bonnefoy V (2009) Extending the models for iron and sulfur oxidation in the extreme acidophile Acidithiobacillus ferrooxidans. BMC Genomics 10, 394.

Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ (2011) Removing noise from pyrosequenced amplicons. BMC Bioinformatics 12, 38.

Quince, C, Lanzén, A, Curtis, TP, Davenport, RJ, Hall, N, Head, IM, Read, LF, Sloan, WT (2009) Accurate determination of microbial diversity from 454 pyrosequencing data. Nature Methods 6, 639-641.

Quince, C, Lanzen, A, Davenport, RJ, Turnbaugh, PJ (2011) Removing noise from pyrosequenced amplicons. BMC Bioinformatics 12, 38.

R Development Core Team (2009) R: A language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing.

Read, DB, Bengough, AG, Gregory, PJ, Crawford, JW, Robinson, D, Scrimgeour, CM, Young, IM, Zhang, K, Zhang, X (2003) Plant roots release phospholipid surfactants that modify the physical and chemical properties of soil. New Phytologist 157, 315-326.

Rezek, J, Macek, T, Macková, M, Tříska, J, Růžičková, K (2008) Hydroxy-PCBs, methoxy-PCBs and hydroxy-methoxy-PCBs: metabolites of polychlorinated biphenyls formed in vitro by tobacco cells. Environmental Science & Technology 42, 5746-5751.

Rho M, Tang H, Ye Y (2010) FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Research 38, e191–e191.

Rhoads A, Au KF (2015) PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics 13, 278-89.

Ridderberg W, Nielsen SM, Nørskov-Lauritsen N (2015) Genetic Adaptation of Achromobacter sp. during Persistence in the Lungs of Cystic Fibrosis Patients. PLoS ONE 10, e0136790.

Riesenfeld CS, Schloss PD, Handelsman J (2004) Metagenomics: genomic analysis of microbial communities. Annual Review of Genetics 38, 525-52.

Roller M, Lucić V, Nagy I, Perica T, Vlahovicek K (2013) Environmental shaping of codon usage and functional adaptation across microbial communities. Nucleic Acids Research 41, 8842-52.

Ronaghi M, Karamohamed S, Pettersson B, Uhlén M, Nyrén P (1996) Real-time DNA sequencing using detection of pyrophosphate release. Analalytical Biochemistry 242, 84-9.

Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M (1977b) Nucleotide sequence of bacteriophage phi X174 DNA. Nature 265, 687-95.

Sanger F, Coulson AR (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. Journal of Molecular Biology 94, 441-446.

Sanger F, Nicklen S, Coulson AR (1977a) DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences USA 74, 5463-7.

Sanger F, Tuppy H (1951) The amino-acid sequence in the phenylalanyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates. Biochemical Journal 49, 463–481.

Santofimia E, González-Toril E, López-Pamo E, Gomariz M, Amils R, Aguilera A (2013) Microbial diversity and its relationship to physicochemical characteristics of the water in two extreme acidic pit lakes from the Iberian Pyrite Belt (SW Spain). PLoS One 8, e66746

Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M (2007) CAMERA: a community resource for metagenomics. PLoS Biology 5, e75.

Shah, N, Tang, H, Doak, TG, Ye, Y (2012) "Comparing bacterial communities inferred from 16S rRNA gene sequencing and shotgun metagenomics," in Biocomputing 2011. (Singapore: WORLD SCIENTIFIC) 165-176.

Shannon CE (1948) Amathematical theory of communication. The Bell System Technical Journal 27, 379–423.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Research 13, 2498-504.

Sharpton TJ (2014) An introduction to the analysis of shotgun metagenomic data. Frontiers in Plant Science 5, 209.

Shyu C, Soule T, Bent SJ, Foster JA, Forney LJ (2007) MiCA: a web-based tool for the analysis of microbial communities based on terminal-restriction fragment length polymorphisms of 16S and 18S rRNA genes. Microbial Ecology 53, 562-570.

Schattner P, Brooks AN, Lowe TM (2005) The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Research 33, W686-9.

Schembri MA, Dalsgaard D, Klemm P (2004) Capsule shields the function of short bacterial adhesins. Journal of Bacteriology 186, 1249-57.

Schierbeek A (1959) Measuring the Invisible World: The Life and Works of Antoni van Leeuwenhoek. London: Abelard-Schuman. (online) https://www.questia.com/library/73684/measuring- the- invisible- world- the- life-and-works

Schippers A (2004) Biogeochemistry of metal sulfide oxidation in mining environments, sediments, and soils. Special Papers-Geological Society Of America 379, 49–62.

Schloss PD, Gevers D, Westcott SL. (2011) Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. PLoS One. 6, e27310.

Schloss, PD, Westcott, SL, Ryabin, T, Hall, JR, Hartmann, M, Hollister, EB, Lesniewski, RA, Oakley, BB, Parks, DH, Robinson, CJ, Sahl, JW, Stres, B, Thallinger, GG, Van Horn, DJ, Weber, CF (2009) Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. Applied and Environmental Microbiology 75, 7537-7541.

Schmieder R, Edwards R (2011) Quality control and preprocessing of metagenomic datasets. Bioinformatics 27, 863-4.

Schouls LM, Schot CS, Jacobs JA (2003) Horizontal transfer of segments of the 16S rRNA genes between species of the Streptococcus anginosus group. Journal of Bacteriolology 185, 7241-6.

Schrenk MO, Edwards KJ, Goodman RM, Hamers RJ, Banfield JF (1998) Distribution of thiobacillus ferrooxidans and leptospirillum ferrooxidans: implications for generation of acid mine drainage. Science 279, 1519-22.

Simpson E (1949) Measurement of diversity. Nature 163, 688.

Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJ, Birol I (2009) ABySS: a parallel assembler for short read sequence data. Genome Research 19, 1117-23.

Singer, AC, Smith, D, Jury, WA, Hathuc, K, Crowley, DE (2003) Impact of the plant rhizosphere and augmentation on remediation of polychlorinated biphenyl contaminated soil. Environmental Toxicology and Chemistry 22, 1998-2004.

Sipila, TP, Keskinen, A-K, Akerman, M-L, Fortelius, C, Haahtela, K, Yrjala, K (2008) High aromatic ring-cleavage diversity in birch rhizosphere: PAH treatment-specific changes of I.E.3 group extradiol dioxygenases and 16S rRNA bacterial communities in soil. The ISME Journal 2, 968-981.

Smalla, K, Wieland, G, Buchner, A, Zock, A, Parzy, J, Kaiser, S, Roskot, N, Heuer, H, Berg, G (2001) Bulk and Rhizosphere Soil Bacterial Communities Studied by Denaturing Gradient Gel Electrophoresis: Plant-Dependent Enrichment and Seasonal Shifts Revealed. Applied and Environmental Microbiology 67, 4742-4751.

Sommer DD, Delcher AL, Salzberg SL, Pop M (2007) Minimus: a fast, lightweight genome assembler. BMC Bioinformatics 8, 64.

Staden, R (1996) The Staden Sequence Analysis Package. MolecularBiotechnology 5, 233-241

Staley JT, Konopka A (1985) Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. Annual Review Microbiology 39, 321-46.

Steglich C, Lindell D, Futschik M, Rector T, Steen R, Chisholm SW (2010) Short RNA half-lives in the slow-growing marine cyanobacterium Prochlorococcus. Genome Biology 11, R54.

Stella, T, Covino, S, Burianová, E, Filipová, A, Křesinová, Z, Voříšková, J, Větrovský, T, Baldrian, P, Cajthaml, T (2015) Chemical and microbiological characterization of an aged PCB-contaminated soil. Science of The Total Environment 533, 177-186.

Strejček M, Wang Q, Rídl J, Uhlík O (2015) Hunting Down Frame Shifts: Ecological Analysis of Diverse Functional Gene Sequences. Frontiers in Microbiology 6, 1267.

Strnad H, Lapidus A, Pačes J, Ulbrich P, Vlček C, Pačes V, Haselkorn R. Complete genome sequence of the photosynthetic purple nonsulfur bacterium Rhodobacter capsulatus SB 1003. Journal of Bacteriology 192, 3545-6.

Strnad H, Rídl J, Pačes J, Kolář M, Vlček C, Pačes V (2011) Complete genome sequence of the haloaromatic acid-degrading bacterium Achromobacter xylosoxidans A8. Journal of Bacteriology 193, 791-2.

Su, J-Q, Ding, L-J, Xue, K, Yao, H-Y, Quensen, J, Bai, S-J, Wei, W-X, Wu, J-S, Zhou, J, Tiedje, JM, Zhu, Y-G (2015) Long-term balanced fertilization increases the soil microbial functional diversity in a phosphorus-limited paddy soil. Molecular Ecology 24, 136-150.

Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics. 23, 1282-8.

Sylvestre M (1980) Isolation method for bacterial isolates capable of growth on p-chlorobiphenyl. Applied and Environmental Microbiol 39, 1223–1224.

Taniguchi Y, Choi PJ, Li G-W, Chen H, Babu M, Hearn J (2010) Quantifying E. coli proteome and transcriptome with single-molecule sensitivity in single cells. Science 329, 533–538.

Tats A, Tenson T, Remm M (2008) Preferred and avoided codon pairs in three domains of life. BMC Genomics 9, 463.

Tatusov, RL, Fedorova, ND, Jackson, JD, Jacobs, AR, Kiryutin, B, Koonin, EV, Krylov, DM, Mazumder, R, Mekhedov, SL, Nikolskaya, AN, Rao, BS, Smirnov, S, Sverdlov, AV, Vasudevan, S, Wolf, YI, Yin, JJ, Natale, DA (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4, 41-41.

Tatusov, RL, Galperin, MY, Natale, DA, Koonin, EV (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Research 28, 33-36.

Tatusova T, Ciufo S, Fedorov B, O'Neill K, Tolstoy I (2014) RefSeq microbial genomes database: new representation and annotation strategy. Nucleic Acids Research 42, D553-9.

Teeling H, Waldmann J, Lombardot T, Bauer M, Glöckner FO (2004) TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. BMC Bioinformatics 5, 163.

Temple KL, Colmer AR (1951) The autotrophic oxidation of iron by a new bacterium, thiobacillus ferrooxidans. Journal of Bacteriology 62, 605-11.

Tillmann, S, Strompl, C, Timmis, KN, Abraham, WR (2005) Stable isotope probing reveals the dominant role of Burkholderia species in aerobic degradation of PCBs. FEMS Microbiology Ecology 52, 207-217.

The UniProt Consortium (2015) UniProt: a hub for protein information. Nucleic Acids Research 43, D204-D212.

Toussaint, J-P, Pham, T, Barriault, D, Sylvestre, M (2012) Plant exudates promote PCB degradation by a rhodococcal rhizobacteria. Applied Microbiology and Biotechnology 95, 1589–1603.

Treangen TJ, Koren S, Sommer DD, Liu B, Astrovskaya I, Ondov B, Darling AE, Phillippy AM, Pop M (2013) MetAMOS: a modular and open source metagenomic assembly and analysis pipeline. Genome Biology 14, R2.

Trimble WL, Keegan KP, D'Souza M, Wilke A, Wilkening J, Gilbert J, Meyer F (2012) Short-read reading-frame predictors are not created equal: sequence error causes loss of signal. BMC Bioinformatics 13, 183.

Tuomisto H (2010) A consistent terminology for quantifying species diversity? Yes, it does exist. Oecologia 164, 853-60.

Turnbaugh PJ, Ridaura VK, Faith JJ, Rey FE, Knight R, Gordon JI (2009) The effect of diet on the human gut microbiome: a metagenomic analysis in humanized gnotobiotic mice. Science Translational Medicine 1, 6ra14.

Uhlík, O, Ječná, K, Macková, M, Vlček, C, Hroudová, M, Demnerová, K, Pačes, V, Macek, T (2009) Biphenyl-metabolizing bacteria in the rhizosphere of horseradish and bulk soil contaminated by polychlorinated biphenyls as revealed by stable isotope probing. Applied and Environmental Microbiology 75, 6471-6477.

Uhlík, O, Musilová, L, Rídl, J, Hroudová, M, Vlček, C, Koubek, J, Holečková, M, Macková, M, Macek, T (2013) Plant secondary metabolite-induced shifts in bacterial community structure and degradative ability in contaminated soil. Applied microbiology and biotechnology 97, 9245-9256.

Uhlík, O, Wald, J, Strejček, M, Musilová, L, Rídl, J, Hroudová, M, Vlček, Č, Cardenas, E, Macková, M, Macek, T (2012) Identification of bacteria utilizing biphenyl, benzoate, and naphthalene in long-term contaminated soil. PLoS ONE 7, e40653.

Ullrich SR, Poehlein A, Tischler JS, González C, Ossandon FJ, Daniel R, Holmes DS, Schlömann M, Mühling M (2016) Genome Analysis of the Biotechnologically Relevant Acidophilic Iron Oxidising Strain JA12 Indicates Phylogenetic and Metabolic Diversity within the Novel Genus "Ferrovum". PLoS One 11, e0146832.

Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO (2004) Environmental genome shotgun sequencing of the Sargasso Sea. Science 304, 66-74.

Větrovský T, Baldrian P (2013) Analysis of soil fungal communities by amplicon pyrosequencing: current approaches to data analysis and the introduction of the pipeline SEED. Biology and Fertility of Soils 49, 1027-1037.

Wagner M, Horn M, Daims H (2003) Fluorescence in situ hybridisation for the identification and characterisation of prokaryotes. Current Opinion in Microbiology 6, 302-309.

Wakai S, Kikumoto M, Kanao T, Kamimura K (2004) Involvement of sulfide:quinone oxidoreductase in sulfur oxidation of an acidophilic iron-oxidizing bacterium, Acidithiobacillus ferrooxidans NASF-1. Bioscience, Biotechnology, and Biochemistry 68, 2519-28.

Wang Q, Fish JA, Gilman M, Sun Y, Brown CT, Tiedje JM, Cole JR (2015) Xander: employing a novel method for efficient gene-targeted metagenomic assembly. Microbiome 3, 32.

Wang Q, Garrity GM, Tiedje JM, Cole JR (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and Environmental Microbiology 73, 5261-7.

Wang Q, Quensen JF 3rd, Fish JA, Lee TK, Sun Y, Tiedje JM, Cole JR (2013) Ecological patterns of nifH genes in four terrestrial climatic zones explored with targeted metagenomics using FrameBot, a new informatics tool. MBio 4, e00592-13.

Wang Y, Leung H, Yiu S, Chin F (2014) MetaCluster-TA: taxonomic annotation for metagenomic data based on assembly-assisted binning. BMC Genomics 15, S12.

Wang Y, Leung HC, Yiu SM, Chin FY (2012) MetaCluster 5.0: a two-round binning approach for metagenomic data for low-abundance species in a noisy sample. Bioinformatics 28, i356-i362.

Wang Y, Yasuda T, Sharmin S, Kanao T, Kamimura K (2014) Analysis of the microbial community in moderately acidic drainage from the Yanahara pyrite mine in Japan. Bioscience, Biotechnology, and Biochemistry 78, 1274–1282.

Wilke A, Harrison T, Wilkening J, Field D, Glass EM, Kyrpides N, Mavrommatis K, Meyer F (2012) The M5nr: a novel non-redundant database containing protein sequences and annotations from multiple sources and associated tools. BMC Bioinformatics 13, 141.

Woese CR, Fox GE (1977) Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proceedings of the National Academy of Sciences USA 74, 5088-90.

Woese CR, Kandler O, Wheelis ML (1990) Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proceedings of the National Academy of Sciences USA 87, 4576-9.

Wright ES, Yilmaz LS, Noguera DR (2012) DECIPHER, a search-based approach to chimera identification for 16S rRNA sequences. Applied and Environmental Microbiology 78, 717-25.

Wu M, Scott AJ (2012) Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. Bioinformatics 28, 1033-4.

Wu YW, Simmons BA, Singer SW (2016) MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics 32, 605-7.

Xie W, Wang F, Guo L (2011) Comparative metagenomics of microbial communities inhabiting deep-sea hydrothermal vent chimneys with contrasting chemistries. ISME Journal 5, 414–26.

Xu L, Chen H, Hu X, Zhang R, Zhang Z, Luo ZW (2006) Average gene length is highly conserved in prokaryotes and eukaryotes and diverges only between the two kingdoms. Molecular Biology and Evolution 23, 1107–1108.

Xu Z, Hansen MA, Hansen LH, Jacquiod S, Sørensen SJ (2014) Bioinformatic approaches reveal metagenomic characterization of soil microbial community. PLoS One 9, e93445.

Yao G, Ye L, Gao H, Minx P, Warren WC, Weinstock GM (2012) Graph accordance of next-generation sequence assemblies. Bioinformatics 28, 13-6.

Yi, H, Crowley, DE (2007) Biostimulation of PAH degradation with plants containing high concentrations of linoleic acid. Environmental Science & Technology 41, 4382-4388.

Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Research 18, 821-9.

Zhang L, Xu Z (2008) Assessing bacterial diversity in soil. Journal of Soils and Sediments 8, 379-388.

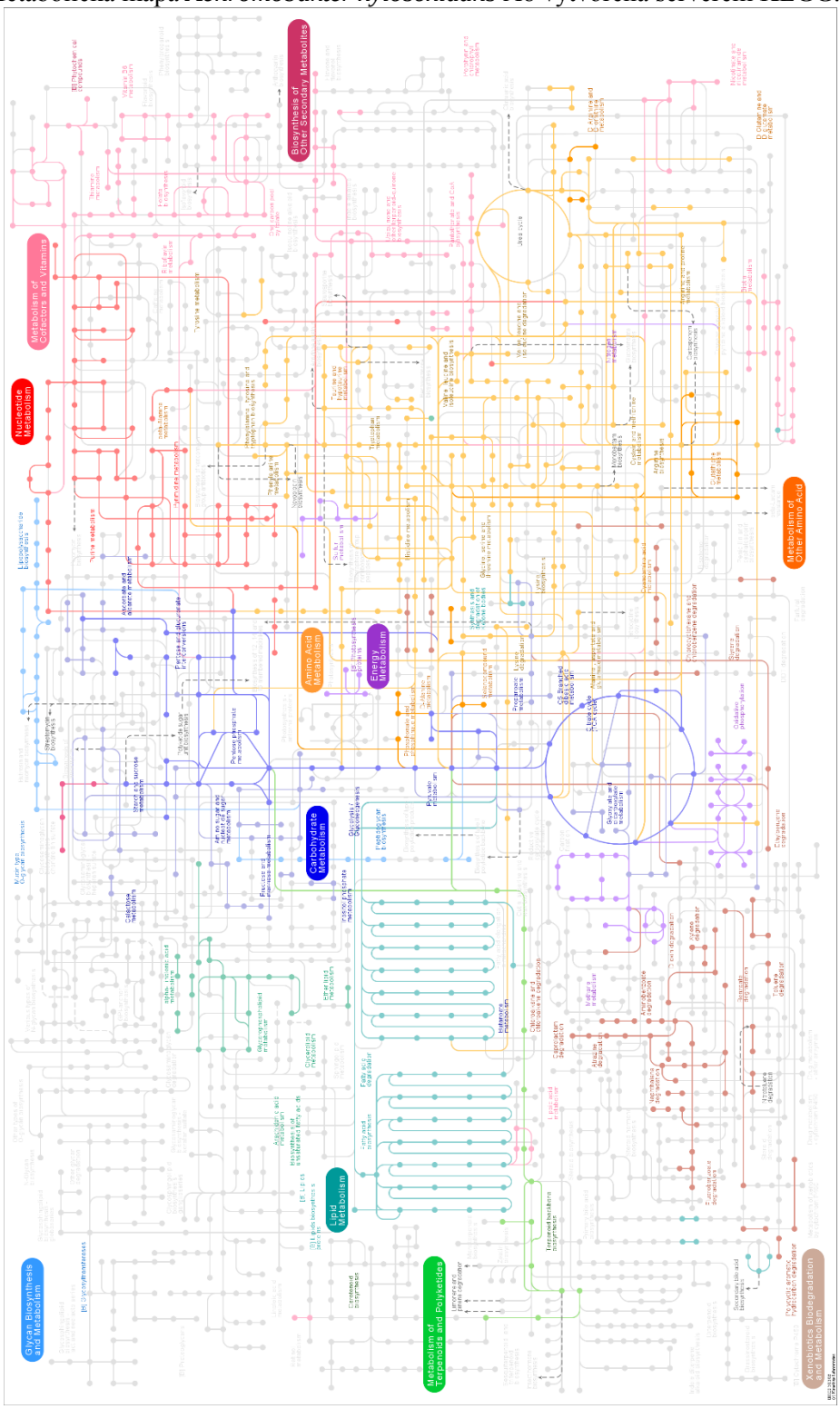
Zheng H, Wu H (2010) Short prokaryotic DNA fragment binning using a hierarchical classifier based on linear discriminant analysis and principal component analysis. Journal of Bioinformatics and Computational Biology 8, 995-1011.

Zhu W, Lomsadze A, Borodovsky M (2010) Ab initio gene identification in metagenomic sequences. Nucleic Acids Research 38, e132–e132.

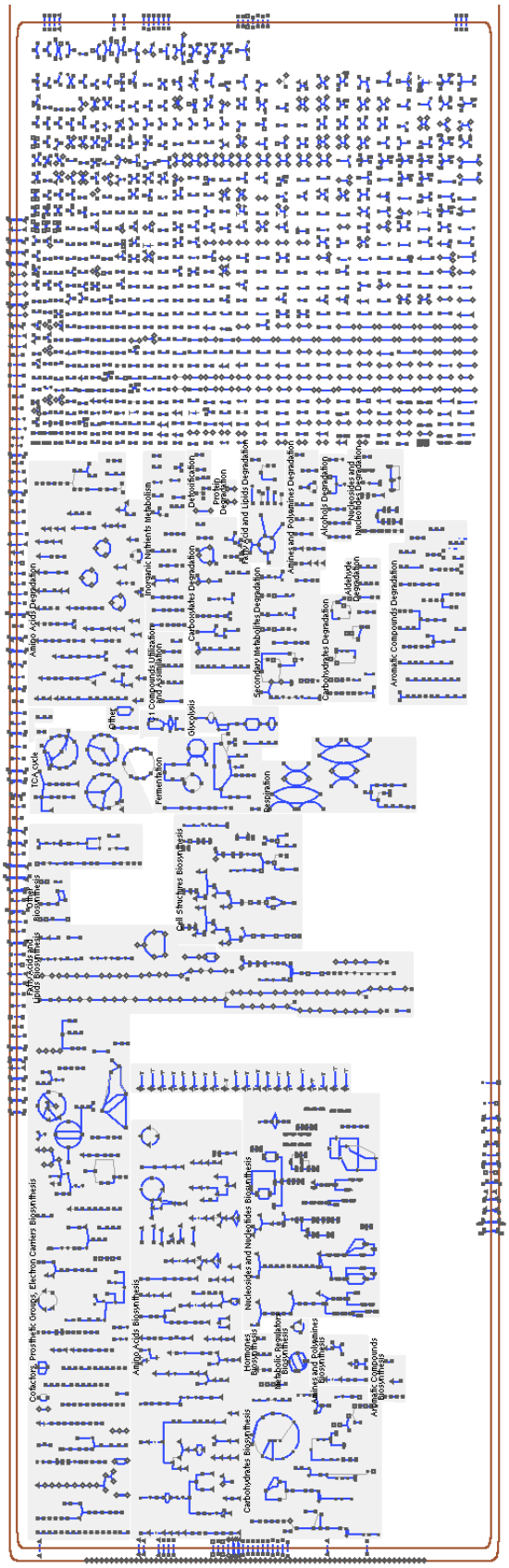
Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA (2013) The MaSuRCA genome assembler. Bioinformatics 29, 2669-77.

Přílohy

Příloha 1: Metabolická mapa *Achromobakter xylosoxidans* A8 vytvořená serverem KEGG.



Příloha 2: Vizualizace metabolických drach *A. xylosoxidans* A8 vytvořená serverem BioCyc (www.biocyc.org).



A. Publikační činnost autora

Rídl J, Kolář M, Strejček M, Strnad H, Štursa P, Pačes J, Macek T, Uhlík O (2016) Plants rather than mineral fertilization shape microbial community. Front Microbiol – ukončeno recenzní řízení, recenzenty doporučeno k publikování.
IF = 3,989

Strnad H*, **Rídl J***, Pačes J, Kolář M, Vlček C, Pačes V (2011) Complete genome sequence of the haloaromatic acid-degrading bacterium Achromobacter xylosoxidans A8. J Bacteriol 193, 791-2.
* autoři přispěli k publikaci rovným dílem
IF = 3,726

Strejček M, Wang Q, **Rídl J**, Uhlík O (2015) Hunting Down Frame Shifts: Ecological Analysis of Diverse Functional Gene Sequences. Front Microbiol 6, 1267.
IF = 3,989

Uhlík O, Strejček M, Vondráček J, Musilová L, **Rídl J**, Lovecká P, Macek T (2014) Bacterial acquisition of hexachlorobenzene-derived carbon in contaminated soil. Chemosphere 113, 141-5.
IF = 3,499

Uhlík O, Wald J, Strejček M, Musilová L, **Rídl J**, Hroudová M, Vlček C, Cardenas E, Macková M, Macek T (2012) Identification of bacteria utilizing biphenyl, benzoate, and naphthalene in long-term contaminated soil. PLoS One 7, e40653.
IF = 4,092

Uhlík O, Musilová L, **Rídl J**, Hroudová M, Vlček C, Koubek J, Holečková M, Macková M, Macek T (2013) Plant secondary metabolite-induced shifts in bacterial community structure and degradative ability in contaminated soil. Appl Microbiol Biotechnol 97, 9245-56.
IF = 3,689

Zubáčová Z, Novák L, Bublíková J, Vacek V, Fousek J, **Rídl J**, Tachezy J, Doležal P, Vlček C, Hampl V (2013) The mitochondrion-like organelle of Trimastix pyriformis contains the complete glycine cleavage system. PLoS One 8, e55417.
IF = 3,730

Zákostelská Z, Kverka M, Klimešová K, Rossmann P, Mrázek J, Kopecný J, Hornová M, Srutková D, Hudcovic T, **Rídl J**, Tlaskalová-Hogenová H (2011) Lysate of probiotic Lactobacillus casei DN-114 001 ameliorates colitis by strengthening the gut barrier function and changing the gut microenvironment. PLoS One 6, e27961.
IF = 4,411

Klimešová K, Kverka M, Zákostelská Z, Hudcovic T, Hrnčíř T, Štepánková R, Rossmann P, **Rídl J**, Kostovčík M, Mrázek J, Kopečný J, Kobayashi KS, Tlaskalová-

Hogenová H (2013) Altered gut microbiota promotes colitis-associated cancer in IL-1 receptor-associated kinase M-deficient mice. Inflamm Bowel Dis 19, 1266-77.
IF = 5,119

Baldrian P, Kolařík M, Stursová M, Kopecký J, Valášková V, Větrovský T, Zifčáková L, Snajdr J, **Rídl J**, Vlček C, Voříšková J (2012) Active and total microbial communities in forest soil are largely different and highly stratified during decomposition. ISME J 6, 248-58.
IF = 3,375

Mořkovský L, Pačes J, **Rídl J**, Reifová R (2015) Scrimer: designing primers from transcriptome data. Mol Ecol Resour 15, 1415-20.
IF = 3,712

Hroudová M, Vojta P, Strnad H, Krejčík Z, **Rídl J**, Pačes J, Vlček C, Pačes V (2012) Diversity, phylogeny and expression patterns of Pou and Six homeodomain transcription factors in hydrozoan jellyfish Craspedacusta sowerbyi. PLoS One 7, e36420.
IF = 4,092

Nývltová E, Stairs CW, Hrdý I, **Rídl J**, Mach J, Pačes J, Roger AJ, Tachezy J (2015) Lateral gene transfer and gene duplication played a key role in the evolution of Mastigamoeba balamuthi hydrogenosomes. Mol Biol Evol. 2015 Apr;32(4):1039-55.
IF = 9,105

Pačes J, Huang YT, Pačes V, **Rídl J**, Chang CM (2013) New insight into transcription of human endogenous retroviral elements. N Biotechnol 30, 314-8.
IF = 1,706

B. Plakátová sdělení

Rídl J, Pačes J, Stavělová M, Kosinová H, Hroudová M, Fousek J, Vlček C. Metagenomic profiling of bacterial consortia inhabiting environments contaminated by chlorinated ethylenes. 5th European Bioremediation Conference, Crete, 2011

Rídl J, Falteisek L, Hroudová M, Pačes J, Strnad H, Vlček C. Analýza taxonomicky jednoduché bakteriální komunity z prostředí hlubinného dolu s využitím metagenomiky a bioinformatiky. ENBIK, 2014