Svoluji k zapůjčení své diplomové práce ke studijním účelům a žádám, aby byla vedena přesná evidence vypůjčovatelů. Převzaté údaje je vypůjčovatel povinen řádně odcitovat.

Univerzita Karlova

Přírodovědecká fakulta

Studijní program: Biologie

Studijní obor: Genetika, molekulární biologie a virologie

Bc. Tadeáš Staněk

Evoluce systému RAYT/REP u bakterií rodu *Stenotrophomonas*

Evolution of RAYT/REP system in *Stenotrophomonas*

Diplomová práce

Vedoucí práce: RNDr. Jaroslav Nunvář, Ph.D.

Praha, 2023

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, 8.8.2023

Podpis

Poděkování

Chtěl bych poděkovat především svému školiteli RNDr. Jaroslavu Nunváři, Ph.D. za podporu a vedení během dvou let výzkumu a psaní této diplomové práce. Za jeho ochotu výsledky kdykoliv konzultovat a směřovat další výzkum. Dále chci poděkovat své rodině a přítelkyni, především za psychickou podporu ale i kontroly mého mnohdy nesrozumitelného textu a komplikovaných grafů. Na závěr bych poděkoval své kočce Kikuši, jejíž vrnění mi vždy zlepšilo den.

Abstrakt

REP (repetitivní extragenové palindromy) jsou krátké DNA elementy nacházející se v mezigenových oblastech řady bakterií, běžně ve stovkách kopií na genom. Neschopny vlastní mobilizace jsou REP šířeny pomocí RAYT (REP asociovaná tyrosinová transpozáza). *Stenotrophomonas* jsou gamaproteobakterie s širokou ekologickou nikou. RAYT/REP systémy jsou u těchto bakterií rozšířené, u 102 kmenů bylo celkem nalezeno přes 132 tisíc elementů. V genomech koexistuje několik sekvenčních tříd REP, pro vysokou početnost daného REP je zásadní přítomnost asociovaného RAYT. REP jsou lokalizovány v „REP lokusech“, řadě intergenových oblastí s vysokou evoluční dynamikou – v daném lokusu se mění počty, sekvence i orientace REP. Kromě genomových ostrovů chybí REP v oblasti terminace replikace, v regionu vykazujícím nestandardní evoluční dynamiku (proměnlivý počet a pořadí genů), pravděpodobně související se segregací chromozomu. *Stenotrophomonas* jsou bohaté na inserční sekvence (IS). Bylo identifikováno 35 unikátních IS a ověřen cíl jejich inzerce. Zástupci deseti IS specificky insertují do REP, další čtyři IS cílí do YPAL (*Yersinia* palindromické elementy) – intergenových DNA motivů doposud známých jen u rodu *Yersinia*. Celkem 14 z 35 IS u *Stenotrophomonas* cílí na intergenové motivy, počtem kopií přitom tvoří 75% všech IS. Počet kopií je klíčový parametr stability populace IS, inserční specificita je u *Stenotrophomonas* prediktorem dlouhodobé persistence v genomech.

Klíčová slova:

IS, komparativní genomika, RAYT, REP, *Stenotrophomonas*, YPAL

Abstract

REP (repetitive extragenic palindromes) are short DNA motifs found in the intergenic regions of many bacteria, typically in the hundreds of copies per genome. Unable to mobilize on their own, REP are propagated by RAYT (REP-associated tyrosine transposase). *Stenotrophomonas* are G-bacteria with a broad ecological niche, belonging to the Gammaproteobacteria. RAYT/REP systems are widespread in them, with over 132,000 REP found in 102 strains. Several classes of REP coexist in the genomes, the presence of associated RAYT is essential for high REP abundance. REP are localised in "REP loci", a series of intergenic regions with high evolutionary dynamics - the number, sequence, and orientation of REP vary at a given locus. In addition to genomic islands, REP are absent from the replication termination site, a region showing non-standard evolutionary dynamics (variable number and order of genes), probably related to chromosome segregation. *Stenotrophomonas* are insertion sequences (IS) rich. 35 unique IS were identified and their insertion target was verified. Ten IS insert into REP, and four other target YPAL (Yersinia palindromic elements), DNA motifs so far known only in the genus *Yersinia*. A total of 14 out of 35 IS in *Stenotrophomonas* target intergenic motifs, while their copy number accounts for 75% of all IS. Copy number is a key parameter of IS population stability, and insertion specificity is a predictor of long-term persistence in *Stenotrophomonas*.

Key words:

IS, comparative genomics, RAYT, REP, *Stenotrophomonas*, YPAL

Obsah

Zkratky

AK	Aminokyselina
ANI	Průměrná podobnost DNA dvou porovnávaných genomů
Anti-*ori*	Pozice na DNA nejvzdálenější počátku replikace
BIME	Roztroušené bakteriální mozaikové elementy, anglicky: „Bacterial interspersed mosaic elements“
bp	Pár bází, anglicky „base pair“
core genom	Množina genů sdílená všemi kmeny jednoho druhu
dif site	Oblast nezbytná k rozdělení dceřiných chromozomů, anglicky: „deletion induced filamentation“
dsDNA	Dvouvláknová DNA
FtsK	„Filament temperature sensitive mutant K“ – filamentární teplotně sensitivní mutace K
GI	Genomový ostrov, anglicky: „Genomic island“
HUH motiv	Katalytická triáda: Histidin – Hydrofobní aminokyselina - Histidin
ICE	Integrativní a konjugativní element
Indel mutace	Inzerce/delece nukleotidu/ů v DNA
Intergen	Nekódující oblast mezi geny
Inz1	Inzert1
Inz2	Inzert2
iREP	Inverzní REP element
IS	Inserční sekvence
kDa	kilodalton
KOPS	FtsK orientující polární sekvence
MDR	Multidrug resistence
MGE	Mobilní genetický element
nt	Nukleotid
ORF	Otevřený čtecí rámec, anglicky: „Open Reading Frame“
oriC	Počátek replikace, anglicky: „Origin of Replication“
Pseudogen	Oblast DNA připomínající gen, typicky vzniklá mutací funkčního genu
RAYT	REP asociovaná transpozáza s tyrosinem, anglicky: „REP associated tyrosin transposase“
REP	Repetitivní extragenový palindrom
REPIN	Dvojce REP v inverzní orientaci REP-spacer-iREP, anglicky: „REP doublet forming hairpin“
Replichóra	Jedna polovina kruhového chromozomu, rozděleného dle oriC a oblasti terminace replikace
SNP	Pozice v homologní oblasti dvou DNA s rozdílným nukleotidem, anglicky: „Single Nucleotide Polymorphism“
ssDNA	Jednovláknová DNA
Syntenie	Oblast stálého pořadí genů mezi genomy příbuzných organismů
TE	Transponující element
Terminus	Oblast zlomu GC skew daného genomu
TIR	Terminální inverzní repetice
YPAL	*Yersinia* palindromický element

1. Úvod

Bakterie jsou organismy žijící prakticky ve všech představitelných habitatech. Zde jsou nuceny se přizpůsobit limitům prostředí a zapojit se do kompetice o zdroje. Působí tak na ně řada selekčních tlaků, aby každá část jejich životního cyklu byla efektivní a rychlá. To je jednou z příčin pro bakterie typické vysoké kódující density. Nadbytečná DNA prodlužuje čas replikace a její syntéza stojí bakterii energii. Typicky proto až 90% genomu má kódující kapacitu, zbytek tvoří promotory, rRNA, tRNA, represory, terminátory transkripce ale také i sobecké genetické elementy.

Jedním z nich jsou transpozibilní elementy (TE) - úseky DNA schopné mobilizovat svou DNA a poté ji vložit do jiné pozice v genomu (transpozice). Jedním z nejjednodušších jsou inzerční sekvence (IS). Transpozici provádí pomocí proteinů (transpozáz), které si samy kódují. Technicky tak nejde o nekódující DNA, ale transpozázy nemají ve fyziologii bakterie typicky žádnou roli. Při transpozici se IS často duplikují, nově vzniklé homologní oblasti mohou sloužit pro rekombinační události vedoucí k amplifikaci, přesunu, inverzi či deleci větších úseků DNA. Situaci dále komplikují větší sobecké elementy, ty jsou schopné nést přídavné geny, typicky pro bakterii prospěšné. Mohou dosahovat délky desítek až stovek kbp. Řada z nich je schopna horizontálního přenosu mezi bakteriemi a může zásadně ovlivnit jejich fyziologii.

Na druhé straně spektra velikostí stojí elementy dlouhé desítky či nižší stovky bp. Jsou příliš krátké, aby kódovaly protein, často tak není známo, jak vznikají. Jedním z těchto elementů jsou REP (repetitivní extragenové palindromy), krátké motivy rozšířené do mezigenových oblastí mnoha bakterií. U některých druhů lze nalézt stovky až tisíce těchto elementů per genom. Jsou velice dynamické a jejich počty i pozice mohou být i u příbuzných druhů velmi rozdílné. Dnes je znám protein zodpovědný za diseminaci REP. Tento protein se není schopen sám přesouvat a je stabilní součástí genomu. Na systém proto dlouhodobě působí selekční tlak, který by jej měl formovat tak, aby z něj pro bakterii plynulo zvýšení (či alespoň ne pokles) fitness.

V rámci této práce je studován systém REP elementů a s nimi asociované nukleázy RAYT (REP asociovaná tyrosinová transpozáza) v rámci rodu *Stenotrophomonas*. U těchto bakterií již byl tento systém dříve identifikován a jeví se velmi aktivní (přes tisíc REP per genom).
V posledních letech rychle roste množství plně sekvenovaných genomů *Stenotrophomonas*. Také bylo vypracováno několika komparativně genomických prací, které na základě těchto sekvencí predikují fylogenetické linie rodu. Proto je tento rod vhodný pro komplexní analýzu RAYT/REP systému a jeho evoluci v hostitelských genomech. Dále se zaměřujeme na popsání dynamiky inserčních sekvencí v těchto genomech a jejich vztahu s repetitivními intergenovými motivy.

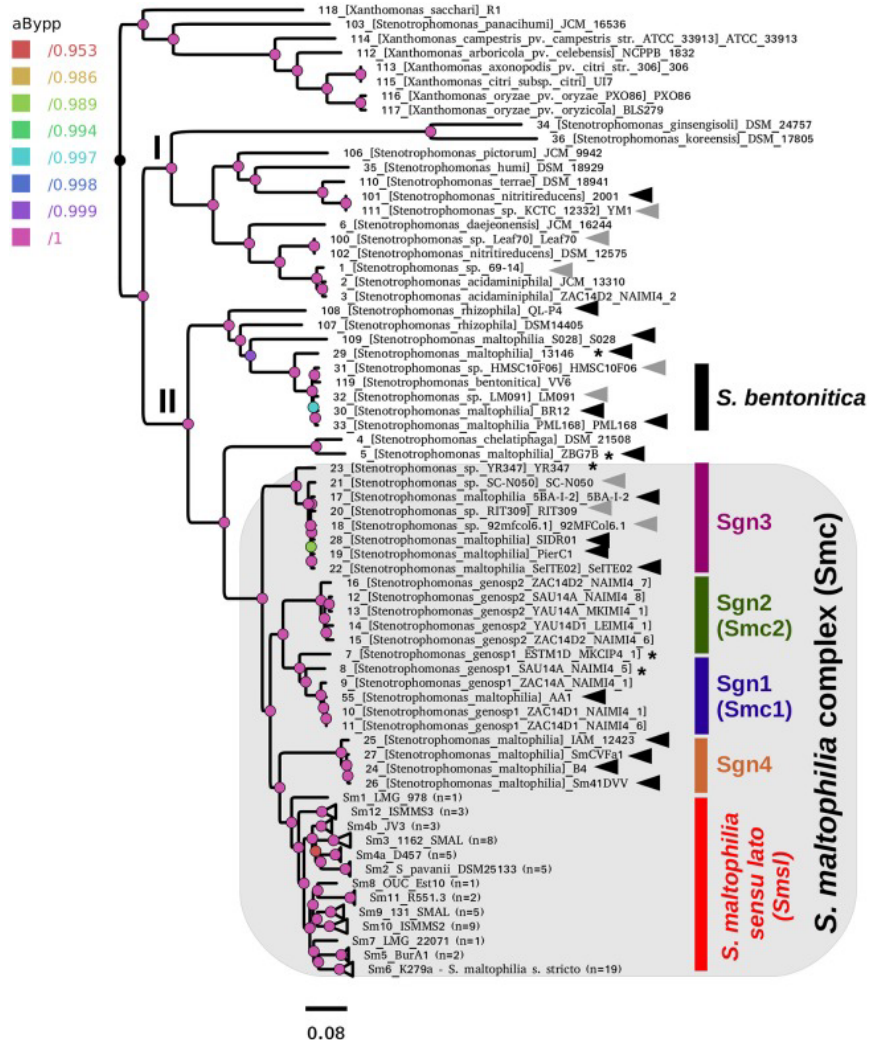
2. Cíle práce

1. Fylogenetická analýza vztahů bakterií rodu *Stenotrophomonas* na základě sekvencí všech kompletně sekvenovaných genomů a srovnání se současnou literaturou.
2. Identifikace všech RAYT přítomných v analyzovaných genomech *Stenotrophomonas*. Jejich rozdělení do skupin na základě umístění v genomu a fylogenetické analýzy.
3. Predikce struktury RAYT proteinů jednotlivých skupin, porovnání s experimentálně získanou strukturou RAYT z *E. coli.*
4. Identifikace všech sekvencí REP v okolí RAYT a hostitelských genomech, rozdělení do skupin na základě asociace s RAYT.
5. Analýza četnosti REP elementů v genomech *Stenotrophomonas* a vztahu k přítomnosti asociovaných RAYT.
6. Analýza distribuce REP elementů v genomech *Stenotrophomonas.* a jejich evoluční stability/variability.
7. *De novo* identifikace inserčních sekvencí v genomech *Stenotrophomonas*. Klasifikace dle databáze ISfinder a homologie transponáz.
8. Identifikace a ověření inzerčních cílů všech IS se zaměřením na inzerce do REP elementů.

3. Přehled literatury

3.1. Rod *Stenotrophomonas*

Stenotrophomonas jsou gram negativní gamaproteobakterie z řádu *Xanthomonadales*. Jde o morfologicky uniformní rod s širokou ekologickou nikou, běžně se vyskytující v půdě i vodě, asociovaný s rostlinami i živočichy. Buňky jsou pohyblivé, využívají „twitching motility“ a vytvářejí biofilmy (Brooke, 2012). Původně byly kmeny řazeny do rodu *Pseudomonas* a dodnes jsou tak některé vedeny (například druh *S. maltophilia* byl řazen jako *Pseudomonas maltophilia*) (HUGH & RYSCHENKOW, 1961). Geneticky jsou *Stenotrophomonas* ve skutečnosti bližší rodům *Xylella* a *Xanthomonas*, jak je vidět i na ilustraci níže.



Ilustrace 1. Fylogenetická analýza rodu Stenotrophomonas připravena pomocí alignmentu kompletních genomů. Na jejím základě je rod rozdělen do skupin se zvláštním zaměřením na definici druhu Stenotrophomonas maltophilia. Převzato z (Vinuesa et al., 2018).

Krom *S. maltophilia* (podkapitola níže) obsahuje rod řadu dalších druhů. Například druh *Stenotrophomonas acidaminiphila* byl poprvé zachycen v roztoku kyseliny tereftalové v petrochemickém závodu (Assih et al., 2002).

Stenotrophomonas nitritireducens byla objevena na biofiltrech, kde redukovala dusitany (Finkmann et al., 2000). *Stenotrophomonas pavanii* je endofytická bakterie cukrové třtiny, které fixuje dusík (Ramos et al., 2011). *Stenotrophomonas rhizophila* se vyskytuje v rhizosférách brambor či řepky, které chrání před plísňovými infekcemi (Wolf et al., 2002). Krom zmíněných existuje řada dalších druhů, většina však dnes nemá sekvenované celé genomy a nejsou více prozkoumány (Heylen et al., 2007; Kim et al., 2010; M. Lee et al., 2011). V desátých letech tohoto století byla pro genotypizaci rodu využívána komparativní genomika na základě sekvencí 16S rDNA a několika dalších genů (například *gyrB*) (Coenye et al., 2004). Později se začaly využívat delší konkatenace desítek až stovek konzervovaných genů (Patil et al., 2018). V posledních letech bylo provedeno několik analýz, které se zaměřují na komparativní analýzu rodu s využitím celogenomových sekvencí. Tyto práce pro zachování kontinuity využívají původních názvů bakterií, ale zařazují je do nových skupin (ilustrace 1).

3.1.1. Stenotrophomonas maltophilia

Stenotrophomonas maltophilia je nejznámějším zástupcem rodu. *S. maltophilia* sensu lato jsou všechny kmeny dnes označené jako Sm1-18, ty mají hodnotu ANI 90-95% (průměrná podobnost nukleotidů). Za *S. maltophilia* sensu stricto se pak považují pouze kmeny skupiny Sm6 (ANI nad 95%) (Gröschel et al., 2020; Xu et al., 2023). Sm6 je korunní skupina celého rodu, která obsahuje také typový kmen *S. maltophilia* 13637 (Davenport et al., 2014) a reprezentativní kmen K279a.

S. maltophilia je z celého rodu nejvíce diverzní a prozkoumaný druh. Většina kmenů vytváří odolné biofilmy, díky kterým kolonizují širokou škálu vlhkých nik. Příkladem budiž rhizosféra rostlin (Berg et al., 2005), vodní tok Niagarských vodopádů (NAKATSU et al., 1995) či prostředí těla široké škály zvířat (E. H. Johnson et al., 2003; Petridou et al., 2010). V medicinálním kontextu pak prostory nemocnic a v nich konkrétně odpadní trubky, ventilaci, čističky vody, dezinfekci rukou či katetry vedoucí fyziologický roztok a mnoho dalších míst (AM et al., 2001; Klausner et al., 1999; Lai et al., 2006; Sui et al., 2012; Wishart & Riley, 1976). Při infekci člověka nejčastěji kolonizuje epitel dýchacích cest, typicky tak infekce vede k pneumonii (Fujita et al., 1996).

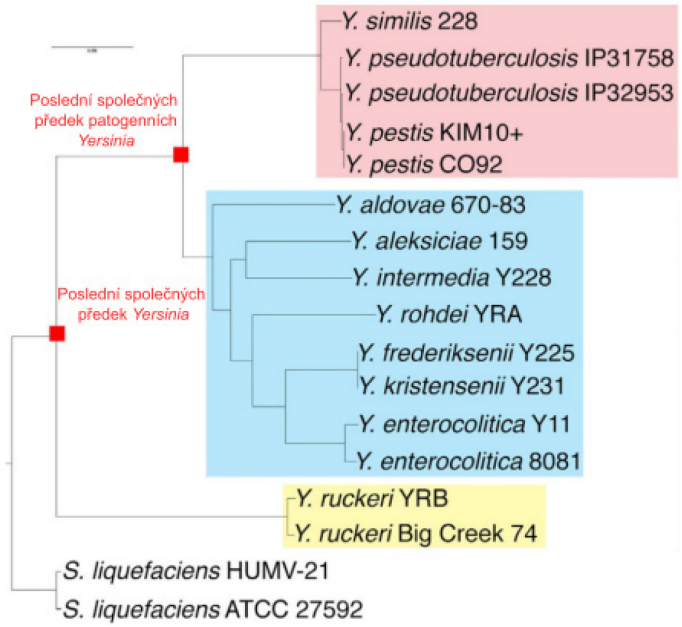
Díky rostoucím schopnostem moderní medicíny udržet na živu a léčit pacienty, kteří mají těžce oslabený imunitní systéme či jsou dlouhodobě na mechanické ventilaci, se zvyšuje nebezpečí nozokomiálních patogenů. Typicky jde o druhy pro zdravého člověka málo nebezpečné, nicméně pro imunodeficientního pacienta smrtící. Jednou z těchto bakterií je *S. maltophilia*. Ta je v takových případech notoricky špatně léčitelná. *S. maltophilia* byla také zachycena u 16% pacientů intenzivní péče (n=123 pacientů) s COVID-19, u kterých propuknul zápal plic. Tito pacienti byly na jednotce intenzivní péči v průměru o 12 dní déle (Raad et al., 2023). Již v minulých letech byla zachycena řada vzorků této bakterie multirezisteních vůči široké škále antibiotik, opravdový rozsah její

antibiotické rezistence byl odhalen při první sekvenace celého jejího genomu (Crossman et al., 2008). Dnes panuje obava z možného vzniku a rozšíření panrezistetních kmenů.

Je třeba poznamenat, že byť existují případy smrti vlivem *S. maltophilia* a mortalita bakteriémie (přítomnosti bakterií v krvi) se pohybuje kolem 40% (Jian et al., 2022; Lai et al., 2004; Muder et al., 1996), jde v současnosti stále o medicinálně málo významnou bakterii. Pacienti, kteří podlehnou této bakteriální infekci jsou ve vážném stavu již před propuknutím infekce (Insuwanno et al., 2020). *S. maltophilia* v současnosti není sama o sobě budoucí globální hrozbou, je ale příkladem rostoucího problému dnešní medicíny, a to sice postupného vzniku rezistence bakterií k většině používaných antibiotik. *S. maltophilia* využívá celé škály mechanismů antibiotické rezistence. Těmi jsou snížená permeabilita buněčné membrány, produkce MDR (Multidrug resistence) pump a enzymů štěpících antibiotikum jako jsou beta-laktamázy, karbapenemázy či aminoglykosidy modifikující enzymy (Zhang et al., 2000).

3.2. Rod *Yersinia*

Yersinia jsou gram-negativní fakultativně anaerobní gamaproteobakterie z čeledi *Enterobacteriaceae*. Nechvalně známé jsou díky patogenním zástupcům, především *Yersinia pestis*. Z genetického hlediska je *Y. pestis* téměř identická s *Y. pseudotuberculosis.* Jde o recentně vzniklý druh, jeho klasifikace je z pohledu genetiky sporná. Průměrná podobnost sekvenovaných genomů (ANI) *Y. pestis* a *pseudotuberculosis* přesahuje 99% (vlastní výsledky), což výrazně převyšuje dnes zavedenou hodnotu 95%, která definuje kmeny jednoho druhu. To je zřejmé z fylogramu na ilustraci 2. *Y. pestis* lze proto považovat za epidemický klon v rámci *Y. pseudotuberculosis*.



Ilustrace 2. Fylogenetický strom kmenů Yersinia, připraven na základě celogenomových sekvencí. Vyznačeny jsou potenciální fylogenetické skupiny. Převzato z (Tan et al., 2016)

Z epidemiologického hlediska je *Y. pestis* unikátní. Nákaza člověka způsobuje mor, onemocnění, které je (bez léčby) letální a jeho pandemie významně proměnily evropskou civilizaci. Z medicínského hlediska se proto *Y. pestis* považuje za vlastní druh. Mor se přenáší kousnutím infikovanou vší, ale také může dojít k přenosu při kontaktu s tělními tekutinami infikovaného zvířete. Vektor se šíří v srsti hlodavců – typicky krys, které ve středověku byly v Evropě hojně rozšířeným druhem. Od té doby populace krys poklesly, zároveň je dnes mor léčitelný antibiotiky. Díky tomu je mor dnes pouze endemickým onemocněním, byť stále ne zcela eradikovaným, neboť se drží v divokých populacích některých druhů zvířat (Kugeler et al., 2015; Salkeld et al., 2010).

Dalšími patogenními druhy jsou *Y. pseudotuberculosis* a *enterocolitica*. Ty způsobují méně závažné gastrointestinální infekce, tzv. yersiniózy. *Y. pseudotuberculosis* se šíří nejčastěji kontaktem s nakaženými zvířaty či jeho fekáliemi. *Y. enterocolitica* se šíří vodou či kontaminovanou potravou, běžným hostitelem je například prase domácí. Nákaza se dokáže rychle rozšířit jak v chovu, tak při porcování masa později na jatkách (Terentjeva & Bērziņš, 2010). Epidemii nejsou běžné, ale při propuknutí může dojít k onemocnění tisíců lidí během několika dní.

Yersinia mají řadu adaptací (tzv. faktorů virulence), které jim dovolují efektivně infikovat hostitele. Jsou to vnitrobuněční paraziti, díky čemuž se vyhnou částí imunitního systému. Všechny tři zmíněné druhy mají plasmid nesoucí sekreční systém typu III (T3SS), který sekretuje šest typů Yop proteinů do nitra parazitované buňky (Shao, 2008). Yop proteiny inhibují fagocytózu a prozánětlivé signály. U *Y. enterocolitica* a *pseudotuberculosis* je tento systém pod kontrolou RNA teploměrů, po transkripci vytvářejí na vláknu RNA vlásenky. Součástí těchto vlásenek je i sekvence Shine-Dalgarno, ta tak není přístupná ribozomům a exprese je inhibována. Vlásenky se rozpadají pouze při zvýšení teploty v buňce na 37 °C (Pienkoß et al., 2021). Krom patogenních kmenů existuje také množství nepatogenních. Zdá se, že řada kmenů *Yersinia* žije v tenkém střevě části lidské populace, a to bez zřejmého vlivu na jejich zdraví (Baut et al., 2018).

3.3. Bakteriální genom

Typicky je genom bakterie tvořen jedním kruhovým chromozomem, jehož průměrná délka se pohybuje okolo 4 Mbp. Nejmenší genomy jsou redukovány pod 0,5 Mbp, ty největší poté dosahují více jak 15 Mbp (Ochman & Caro-Quintero, 2016). Na rozdíl od eukaryot, pro bakterie je typická vysoká kódující denzita s minimem redundantních oblastí. Kódující oblasti jsou odděleny krátkými nekódujícími úseky, které jsou často tvořeny například regulačními oblastmi či repetitivními motivy. Bakteriální genom je velice dynamický a kontinuálně je ovlivňován mutacemi, genomovými přestavbami či horizontálním přenosem.

3.3.1. Genomové ostrovy

Bakterie se typicky množí dělením mateřské buňky. Na rozdíl od eukaryot nedochází k promíchávání alel od rodičů. Dceřiné buňky jsou klony buňky mateřské, během generací přibývá mutací v klonální linii a může na nich probíhat selekce výhodné mutace. Ta se může z klonální linie rozšířit pomocí horizontálního genového přenosu. Horizontální přenos je nejintenzivnější v rámci kmene či druhu (u vzdáleně příbuzných je vzácnější, ale významnější). Vzniká tak kohezní skupina organismů sdílejících mnoho genů. Geny rozšířené ve všech kmenech jsou součástí core genomu, typicky bývají částí konzervovaného genového sousedství (evolučně stabilní oblasti). Opakem core genomu jsou geny sdílené jen částí kmenů jako například genomové ostrovy (GI) (Hacker et al., 1990). Tyto oblasti vznikají selekčním tlakem, který koncentruje užitečné geny jedné funkce k sobě. Oblast poté může být horizontálně přenesena. GI jsou pro bakterie i archea zdrojem metabolické i fenotypové diverzity, která může značně ovlivnit mikroorganismus (Hsiao et al., 2005). Typickým příkladem jsou ostrovy patogenicity umožňující perzistentní kolonizaci hostitelského organismu (Censini et al., 1996; Hacker & Kaper, 2000; Karch et al., 1999).

Core genom je tvořen esenciálními geny a geny zodpovědnými za typický fenotyp druhu. Genomové ostrovy většinou obsahují geny, které v určité nice mohou výrazně zvýšit fitness hostitelské bakterie. GI se udržují v populaci nejčastěji pozitivní selekci. Krom nich existuje i řada dalších sobeckých genetických elementů kolonizujících genom. Charakteristickým zástupcem této skupiny jsou bakteriofágy insertovány v genomu (Bertani & Six, 1958). Jde o lysogenní fázi fága, která v případě aktivace převezme kontrolu nad biosyntetickými drahami buňky pro zajištění vlastní replikace. To vede většinou k usmrcení buňky. Dalšími jsou například toxin-antitoxin systémy, které po inzerci donutí buňku produkovat toxin a protijed (Fozo et al., 2010; Van Melderen & De Bast, 2009). Buňka nemůže systém odstranit, neboť to povede ke ztrátě protijedu (vždy nestabilnější než toxin) a smrti buňky. Toxin-antitoxin systémy jsou často využity v regulačních systémech bakterií, například spuštění fenotypu persistora (Wang & Wood, 2011).

Posledním typem jsou ICE (integrativní a konjugativní elementy). Jedná se o mobilní DNA, která je ale většinu času pouze pasivně propagována během replikace chromozomu se zbytkem genomu. Krom aparátu nutného pro své přežití často ICE nesou geny pro bakterii prospěšné. Příkladem jsou biodegradační dráhy alternativních zdrojů uhlíku či rezistence proti antibiotikům a těžkým kovům. Při mobilizaci ICE dochází k proteosyntéze konjugačního aparátu a potenciálnímu přenosu DNA mezi bakteriemi. V novém hostiteli se ICE ihned vkládá do genomu a opět zahajuje pasivní fázi. Tyto elementy se zdají početné a pravděpodobně mají velký dopad na evoluci bakterií, nicméně dnes nejsou stále dostatečně prozkoumány (Franke & Clewell, 1981; C. M. Johnson & Grossman, 2015).

3.4. Uspořádání genomu související s replikací

V podkapitolách bude rozepsáno několik jevů a mechanismů souvisejících s replikací a které budou více zkoumány v práci. Tyto jevy se nutně nevyskytují u všech bakterií – výjimkou jsou například bakterie s lineárním genomem. Také v případě Xer/*dif* systému existují další systémy, které ho mohou nahradit.

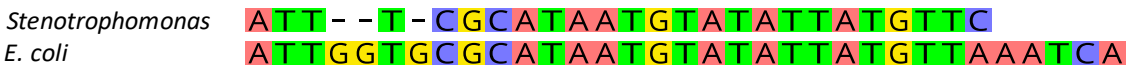
3.4.1. Replikace

Velká většina bakteriálních chromozomů je cirkulární. Klíčovým proteinem pro zahájení replikace těchto bakterií je DnaA. Ten je schopen se vázat na replikační počátek (*oriC*) (Ogasawara et al., 1985; Stuitje et al., 1986). *OriC* je tvořen několika DnaA boxy, které mají sekvenci 5'-TTATNCACA-3' (Schaper & Messer, 1995). DnaA s navázaným ATP na této pozici oligomerizuje a způsobuje tání AT-bohaté oblasti. Zároveň do této oblasti rekrutuje DnaB, což je DNA helikáza. V rámci pre-iniciačního komplexu je na DnaB vázán DnaC, který DnaB inhibuje. Pro zahájení pohybu komplexu je třeba DnaC vyvázat. Toho je dosaženo interakcí DnaG a primerů vedoucího vlákna. Vytvoří se replikační vidlička a zde DnaB interaguje s DNA polymerázou III holoenzymem a rozjíždí se replikace. Replikace se rozjíždí do obou stran chromozomu a dokončuje se přibližně na opačné straně. Existuje více mechanismů, kterými je ukončena replikace, jedním z nich je Xer/*dif* systém.

3.4.2. Xer/*dif* systém

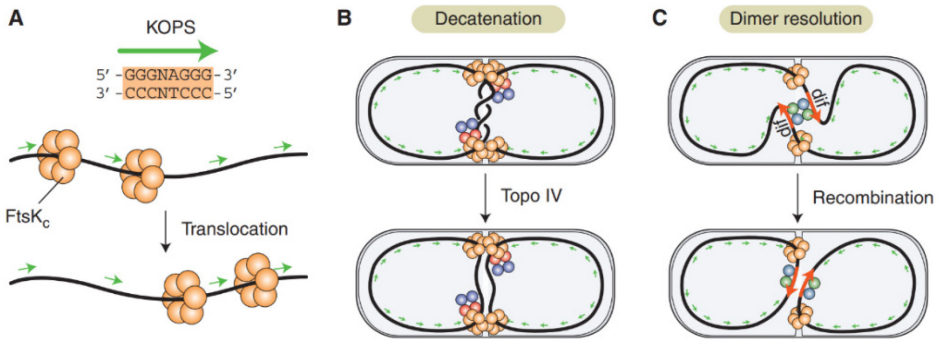
Bakteriální chromozom je replikován pomocí dvou replikačních vidliček jdoucích opačně od *oriC*. Z tohoto pohledu lze genom rozdělit na dvě poloviny, replichóry (Blattner et al., 1997). Při ukončení replikace a spojení replikačních vidliček může docházet k tvorbě takových forem dceřiných chromozomů, které jsou funkčně spojeny a nemohou správně segregovat – v těchto případech je třeba chromozomy od sebe správně oddělit. U *E. coli* k tomu dochází pomocí homologní rekombinace s pomocí dvou tyrozinových rekombináz – systém XerC a XerD (Blakely et al., 1993; Cornet et al., 1996). Tyto rekombinázy operují na konzervované oblasti *dif* site (ilustrace 3) (P. L. Kuempel et al., 1991). S nimi spolupracuje DNA translokáza FtsK (ilustrace 4), ta je nezbytná pro aktivaci rekombinace pomocí XerCD (Hiraga, 1993).

Dif je zkratkou pro anglický termín *deletion induced filamentation*, název odkazuje na filamentární fenotyp *E. coli* při deleci *dif* oblasti (při zároveň inaktivovaném SOS systému). Tyto buňky musí při dělení jako alternativu k XerCD/*dif* využít rekombinační funkce SOS opravných systémů, které nakonec chromozomy oddělí.



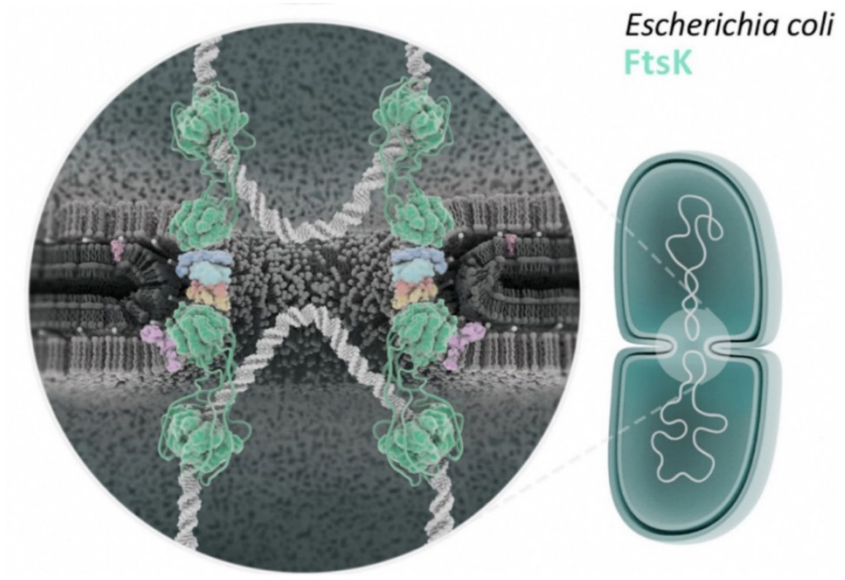
Ilustrace 3. Alignment sekvencí dif site u Stenotrophomonas (definována v rámci práce) a E. coli (Tecklenburg et al., 1995). Připraveno v Geneious prime.

Systém XerCD/*dif* rozdělení chromozomů není jediným systémem, který bakterie využívají. Nicméně, je široce rozšířen – například u rodů *Escherichia*, *Helicobacter*, *Vibrio*, *Stenotrophomonas*, *Pseudomonas*, *Xanthomonas* a dalších (Carnoy & Roten, 2009). Krom bakterií byl prokázán jeho výskyt i u archea (Cortez et al., 2010).



Ilustrace 4. Translokace DNA pomocí FtsK, která se na vláknu orientuje pomocí KOPS elementů. Komplex dekatenuje vlákna dceřiných chromozomů a následně je rekombinací oddělí. Převzato z (Thanbichler, 2010).

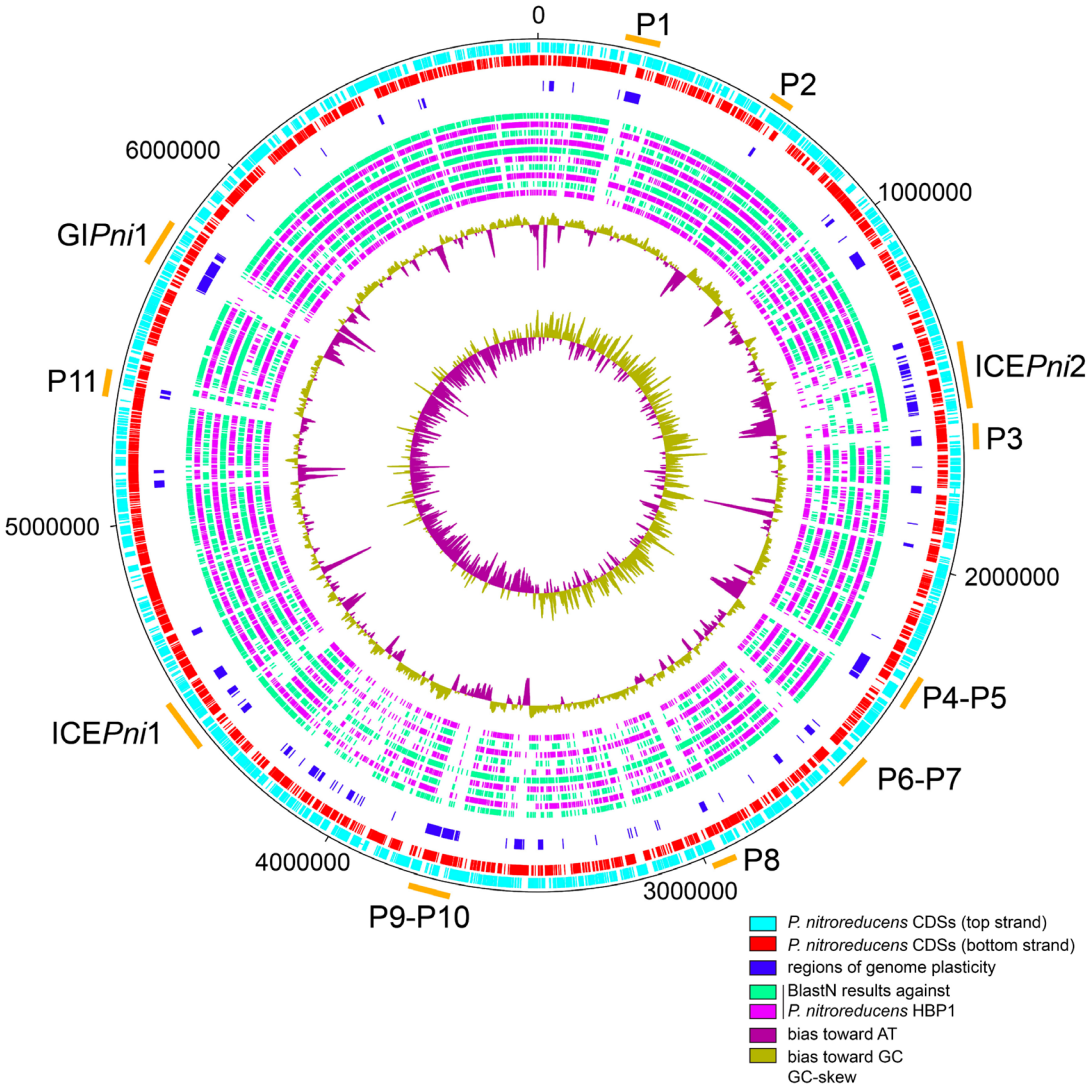
Translokáza FtsK se musí orientovat na DNA vláknu, aby byla schopna nalézt *dif* site. V této oblasti pak pomáhá systému XerCD rozdělit dceřiné chromozomy. K navigaci využívá KOPS elementy (FtsK orientující polární sekvence) (Bigot et al., 2005). KOPS směřují od počátku replikace k *dif* site. Nejčastější sekvence KOPS je 5‘-GGGNAGGG-3‘, některé odchylky jsou tolerovány pro udržení funkce a řada druhů má například přerušené GGG trinukleotidy. V genomu *E. coli* bylo nalezeno necelých 400 KOPS držících se konsensus sekvence (Bigot et al., 2005). FtsK komplex se váže na DNA s KOPS, orientace prvního nalezeného motivu určí směr translokace (J. Y. Lee et al., 2012). Komplex je schopen měnit směr i rychlost translokace, ale molekuly spouštějící tyto procesy nejsou známy (Massey et al., 2006). Krom pohybu po vláknu lze komplex fixovat na membránu, poté slouží jako „DNA pumpa“ (ilustrace 5).



Ilustrace 5. Model segregace chromozomů během dělení E. coli. Již proběhla dekatenace a rekombinace dceřiných chromozomů. Systém FtsK (zeleně) nasedají na membránu, kde jako motory translokují vlákna DNA (šedá). Jde o zjednodušený model, komplex ve skutečnosti tvoří asi 25 FtsK hexamerů. Převzato z (Chan et al., 2022).

3.4.3. GC skew

Zlom v orientaci KOPS elementů se vyskytuje v okolí *dif* site. V těchto místech se také láme hodnota GC skew. GC skew se vyskytuje u všech cirkulárních chromozomů. Jedná se o nerovnoměrné zastoupení GC nukleotidů na vedoucím a opožďujícím se řetězci DNA (Lobry, 1996) (ilustrace 6). Přechody mezi obohacením/ochuzením řetězců jsou lokalizované v oblasti *ori* a *ter*. Z toho důvodu je jev asociován s průběhem replikace. Jeho příčinou je deaminace cytosinu na opožďujícím se vláknu, když je ve formě jednovláknové DNA (Rocha, 2004). V tomto stavu je na deaminace vlákno náchylné (Bhagwat et al., 2016). GC skew lze využít k určení přibližné oblasti terminace replikace.



Ilustrace 6. Cirkulární mapa replikonu Pseudomonas nitroreducens HBP-1. Vnější dva kruhy (tyrkysový a červený) ukazují pozice predikovaných ORF (vedoucí/opožďující vlákno). Kruh níže (modrý) zobrazuje oblasti genomové plasticity, tedy oblasti, které tento kmen odlišují od příbuzných. Nalezeny byly pomocí IslandViewer a jde částečně o inverzi následujících čtyř kruhů homologních oblastí z příbuzných kmenů nalezených BlastnN. Úplně vnitřní kruh měří kumulativní množství GC párů tedy GC skew, kruh nad ním udává celkové % GC párů v DNA. Nad mapou jsou označeny (oranžová) a pojmenovány oblasti genomových ostrovů (GI), profágů (P) a ICE elementů. Převzato z (Carraro et al., 2020)

3.5. REP elementy

Ne všechny mobilní elementy kolonizující genom jsou tak velké jako genomové ostrovy či profágy (běžně desítky až stovky Kbp), některé jsou tak malé, že nemají žádnou vlastní kódující kapacitu. REP jsou krátké intergenové motivy hojně rozšířené v některých bakteriálních druzích. Popsány byly roku 1982 v genomech *Escherichia coli* a *Salmonela typhimurium* (C. F. Higgins et al., 1982). Zkratka REP stojí pro **r**epetitive **e**xtragenic **p**alindromic sequence (Stern et al., 1984). Název je doslovný, REP se vyskytují výhradně mimo kódující oblasti genomu. Mají palindromickou strukturu a vyskytují se ve stovkách kopií na genom. Přestože REP jsou extragenové, většina z nich je transkribovaná (Qian et al., 2015). Mohou tak vytvářet vlásenky na ssRNA, které mají vliv na stabilitu mRNA.

Původně nalezeny u *E. coli*, do konce 90. let byly REP nalezeny u mnoha enterobakterií. V novém tisíciletí také i mimo tuto skupinu, například u *Pseudomonas aeruginosa* a *P. putida* (Weinel et al., 2002). Jak se rozšiřuje množství sekvenovaných druhů napříč bakteriální diverzitou, roste počet druhů s REP. Proteiny RAYT (REP asociovaná tyrosinová transpozáza), které mohou potenciálně šířit REP, jsou rozšířeny mezi proteobakteriemi, zvlášť gammaprotebakteriemi. Dále se vyskytují u *Cyanobacteria*. Naopak zcela chybí u Firmicutes, Deinococcota, *Actinobacteria* či Spirochaetota (Bertels, Gallie, et al., 2017).

3.5.1. Struktura REP

Palindrom je hlavní částí REP, ještě před ním je ale umístěn konzervovaný tetranukleotid 5‘-GTRG-3‘. Celá struktura REP lze rozdělit: GTRG–hlava-variabilní smyčka-ocas. Hlava a ocas jsou rameny palindromu, díky komplementaritě budou vytvářet na ssDNA či RNA vlásenkovou strukturu. Od sebe jsou odděleny variabilní smyčkou. Ta přerušuje palindrom a typicky má délku od 2 do 6 bp. Celková délka REP se pohybuje od 20 do 65 bází.

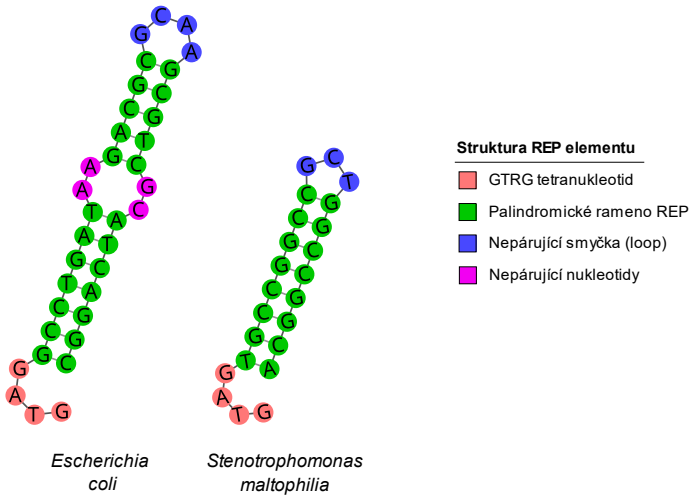
Ne vždy jsou ramena palindromu zcela komplementární, u některých druhů jsou typická krátká přerušení dlouhá 1-2 bp. Tato přerušení jsou konzervovaná u mnoha kopií těchto REP, lze hypotetizovat, že cíleně snižují pevnost REP vlásenky. Nicméně, obecně působí silná selekce na udržení komplementarity ramen (Aranda-Olmedo et al., 2002). Dojde-li k mutaci jednoho ramene, je pozorována selekce na obnovení komplementarity (reverzní mutací či mutací komplementární na druhém rameni). Také je častá mezera mezi GTRG tetranukleotidem a počátek palindromu, obvykle v délce 1-2 bází.

3.5.2. Rozdíly REP mezi druhy

Sekvence a délka REP se liší mezi různými druhy bakterií. Zachována je struktura vlásenky a GTRG tetranukleotid na jejím počátku. Počet kopií i jejich konzervovanost je variabilní. V některých případech i blízce příbuzné bakterie mají zcela odlišné počty i pozice REP elementů (Nunvar et al., 2013). Počet REP jednoho kmene může přesáhnout i tisíc REP

elementů a v příbuzném kmenu nemusí být ani jediný. Není jasné, z jakého důvodu bakterie rozšiřování REP svými genomy tolerují. Situaci dále komplikuje možnost, že některé druhy využívají REP ke specifickému účelu. Nelze tak poznatky z jedné studované bakterie aplikovat na ostatní. Interakce s RAYT a lokalizace v intergenových oblastech jsou pravděpodobně univerzální, může však existovat řada interakčních partnerů (od malých molekul po proteiny), které mohou být druhově specifické.

V rámci *E. coli* byly REP původně objeveny. Většina kmenů má populaci 500 až 600 REP (Bachellier et al., 1999). V porovnání s jinými druhy jsou *E. coli* REP delší, rameno palindromu má 14 bází (ilustrace 7). Stabilita takto dlouhých palindromů je snížena nepárujícím přerušením. Více jak polovina REP je součástí vyšší struktur BIME (Kap. 4.6.2). BIME (anglicky: „Bacterial interspersed mosaic elements“) jsou organizované shluky REP, které lze dle struktury dělit na více typů. Jsou typické pro *E. coli,* u ostatních bakterií jsou vzácné.



Ilustrace 7. Sekundární struktury reprezentativních REP z E. coli a S. maltophilia. Predikce připraveny v Geneious prime, upraveno v Inkscape.

Rod *Pseudomonas* je bohatý na REP elementy. Dobře charakterizovány jsou u druhů *P. fluorescence* a *putida* (Di Nocera et al., 2013). Genomy mají mezi jedním až dvěma tisíci REP. Jde o několik typů, které jsou v mnoha kopiích rozšířeny po genomu. REP u *Pseudomonas* mají navíc konzervované tři báze (5‘-GAA-3‘) na konci palindromu. Také je častá přítomnost dvou až tří bází mezi GTRG a palindromem. Typicky jsou tyto REP součástí vyšší struktury, takzvaných REPIN (anglicky: „REP doublet forming hairpin“, viz. Kap. 4.6.1). Vysoký počet a konzervovanost REP vede k hypotéze, že se u *Pseudomonas* rozšířily relativně nedávno (Nunvar et al., 2013). Nejčastěji jsou umístěny mezi konvergentními geny (transkripce obou genů/operonů je orientována sbíhavě) a ve vzdálenosti do 30 bází od STOP kodonu. Přes tuto blízkost ke kódující oblasti nejsou schopny atenuovat transkripci (Espéli et al., 2001).

Stenotrophomonas jsou dalším rodem s početnou REP populací. Nejprozkoumanější jsou kmeny *S. maltophilia* K279a a R551-3 (Di Nocera et al., 2013). Hyperkonzervovaný je GTRG tetranukleotid, umístěný hned na bázi palindromu (ilustrace 7). Podobně jako u *Pseudomonas* je většina REP konzervovaná a často součástí vyšších struktur (REPIN) (Rocco et al., 2010).

3.5.3. Funkce REP elementů

Univerzální funkce, která by obhájila rozšíření REP mezi bakteriemi, není známa. Nicméně, je popsána řada fyziologických funkcí REP u konkrétních druhů hostitelských bakterií. Umístění v 3‘ nekódující oblasti genu dokáže zvýšit hladinu exprese o desítky procent. Vlásenka (kterou REP v mRNA vytváří) chrání transkript před degradací 3‘-5‘ exonukleáz (Becerril et al., 1985; Khemici & Carpousis, 2004). REP vlásenka nedovolí exonukleáze nasednout a mRNA se zvýšeným poločasem „života" je déle translatována. Pro štěpení sekundární struktury je třeba rekrutovat specifické exonukleázy, např. polynukleotid fosforylázu či RNasu R (Newbury et al., 1987). Bez těchto enzymů by se transkripty kumulovaly v buňce (Cheng & Deutscher, 2005). Umístění v 3‘ nekódující oblasti může transkripci i atenuovat. To ovšem dělá jen část REP, pouze u některých genů a exprese klesá maximálně na polovinu (Espéli et al., 2001).

REP může ovlivnit také translaci. Umístěny do 15 bází od STOP kodonu REP vlásenka fyzicky brání ribozomu dokončit translaci (Deng et al., 2019). Nedokončený protein je poté degradován. Dochází tak k poklesu exprese až o dvě třetiny, část ribozomů translaci dokončí díky RNA helikázám, které REP vlásenku rozvolňují. Koncentrace těchto helikáz má přímý vliv na hladinu exprese těchto genů. Množství helikáz stoupá při stresu (například vlivem UV záření). Mohlo by tak jít o mechanismus rychlé regulace exprese při změně podmínek (Liang & Deutscher, 2016). Pozice mnoha REP u *Stenotrophomonas* osciluje kolem 15 bází od STOP kodonu (Rocco et al., 2010). Mutace v této oblasti tak mohou modulovat REP-dependentní regulaci exprese. Obecnou výhodou toho způsobu regulace exprese je, že pouhá přítomnost REP v intergenu má minimální vliv na expresi ostatních genů (je-li dost daleko).

Některé REP mají druhově specifické funkce. Na BIME, strukturu složenou z několika DNA motivů a REP, se u *E. coli* váže IHF protein. IHF je DNA-vazebný strukturní protein nukleoidu (Boccard & Prentki, 1993; Oppenheim et al., 1993). IHF se váže na specifické DNA motivy a poté ohýbá dvoušroubovici až o 180 stupňů (Montaño & Rice, 2011; Rice, 1997; Rice et al., 1996). Ohyb dokáže zajistit interakci vzdálených oblastí genomu (Goosen & van de Putte, 1995), má také vliv na replikaci i genovou expresi (Charlier et al., 1995; Ryan et al., 2002). U *E. coli* byla také pozorována vazba DNA polymerázy I s REP, účel této interakce není jasný (Gilson et al., 1990).

3.5.4. Pozice REP v genomu

Oblasti mezi konvergentními transkripčními jednotkami jsou typickou lokací REP. Pohyb RNA polymerázy při transkripci konvergentních operonů vytváří pozitivní nadobrátky na DNA. DNA gyráza (v *E. coli*) interaguje s REP, může tak snižovat napětí DNA vlákna a podpořit transkripci oblasti (Espéli & Boccard, 1997; Gellert et al., 1976; Reece et al., 1991; Yang & Ferro-Luzzi Ames, 1988). Naopak minimum REP se vyskytuje

v intergenech mezi divergentními geny/operony (Tobes & Ramos, 2005).REP jsou součástí core genomu, téměř se nevyskytují v evolučně mladších oblastech (Segerman, 2012). Důvodem může být jejich relativně pomalá propagace (Loper et al., 2012). Taktéž je jich minimum v oblastech vzniklých recentní integrací mobilních genetických elementů. REP se nešíří prostřednictvím MGE. U *S. maltophilia* K279a je desetina genomu tvořena genomovými ostrovy (41 GI), žádný z nich neobsahuje REP (Rocco et al., 2009, 2010). Dále je známá absence REP v okolí oblasti iniciace replikace a terminace replikace, důvod je neznámý (Bachellier et al., 1999).

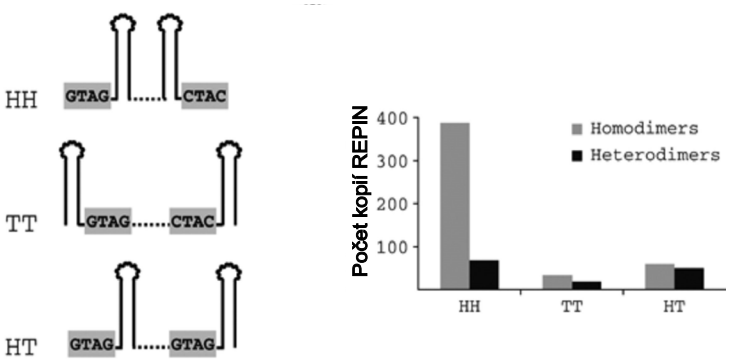
REP jsou většinou transkribovány, mají tak vliv na osud mRNA (jíž jsou součástí) (Espéli et al., 2001). Zatím však nebyla nalezena jasná spojitost mezi REP a funkcí genů, v jejichž downstream oblastech se nacházejí. Regulační role REP je pravděpodobně součástí komplexní sítě obsahující více interakčních partnerů.

3.6. Vyšší struktury REP

REP nebývají v genomech osamoceny. Většina je součástí větších mozaik, BIME a REPIN. Ty bývají druhově specifické, můžou obsahovat dva, ale i desítky REP elementů. Typická je pravidelné vzdálenost mezi REP i jejich typická orientace. Poprvé byly popsány u *E. coli*, kde jsou velmi četné.

3.6.1. REPIN

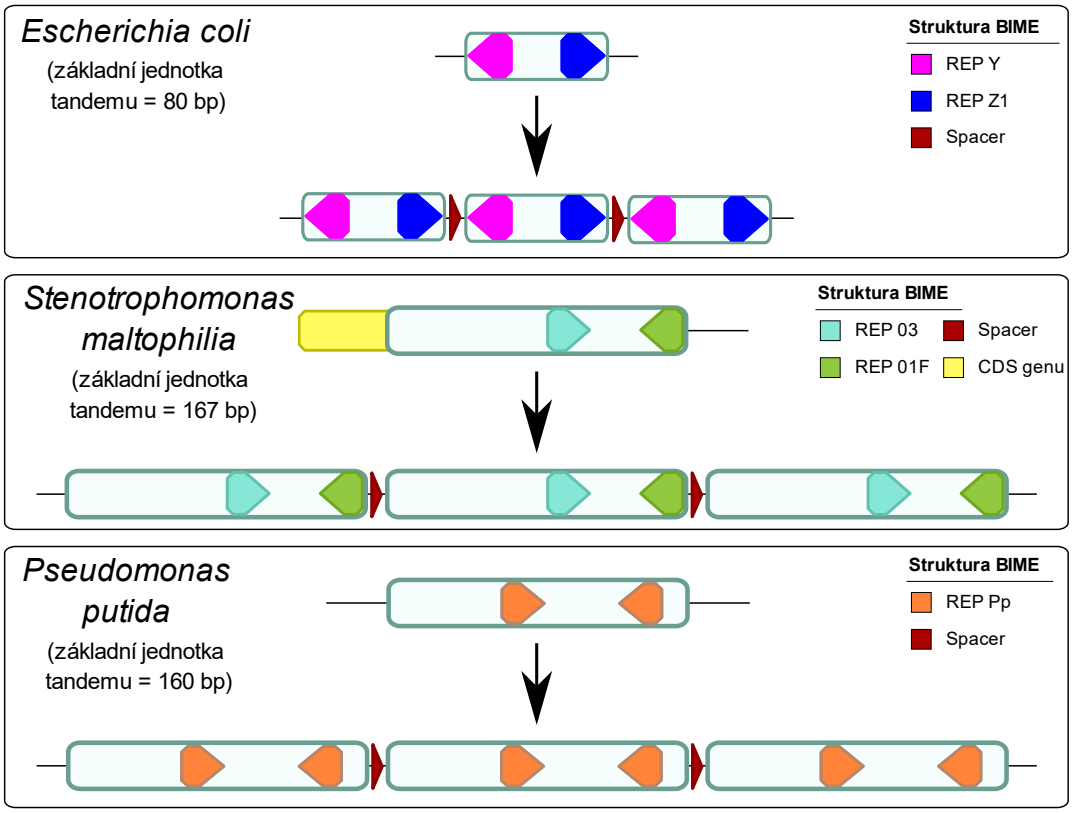
REPIN (anglicky: „REP doublet forming hairpin“) je vyšší strukturou REP. Složena je z dvou REP v inverzní orientaci (ilustrace 8), oddělených mezerou o přibližně konstantní délce (sekvence konzervována nebývá). Právě prostřednictvím REPIN jsou REP rozšiřovány po genomu bakterie (Bertels & Rainey, 2011a, 2011b). Osamocený REP či naopak komplikována mozaika REP elementů (BIME) jsou odvozeny z původního REPIN. U *Stenotrophomonas* je mnoho REP součástí REPIN, jejich struktura je na ilustraci níže. Dále je mnoho REP součástí BIME (viz níže), zbylých přibližně 20% REP u *Stenotrophomonas* jsou monomery (Rocco et al., 2010).



Ilustrace 8. Struktura REPIN v rámci Stenotrophomonas maltophilia. REP v rámci REPIN jsou v různé orientaci, ta je určena dle pozic GTAG tetranukleotidů REP – HH (head-to-head), TT (tail-to-tail) či HT (head-to-tail). Sloupcový graf zachycuje absolutní počty těchto REPIN a navíc rozlišuje, zda jsou REP sekvence identické (homodimer) či rozdílné (heterodimer). Převzato z (Rocco et al., 2010).

3.6.2. BIME

Roku 1982 byly popsány REP u *E. coli*, ale až v roce 1991 bylo zjištěno, že většina je součástí vyšších struktur. BIME (anglicky: „bacterial interspaced mosaic elements“) je v *E. coli* asi 300, mají délku až 500 bází (obvykle jsou ale kratší). Jedná se o mozaiky z kratších elementů, jedním z nich jsou právě REP (Gilson et al., 1991). Je mnoho způsobů, jak je BIME mozaika poskládána, mohou obsahovat dva nebo také přes deset REP. Existuje několik pravidel, kterých se ale všechny drží. REP jsou od sebe vždy odděleny dalším motivem a sousední REP jsou vždy v opačné orientaci. Základní motiv BIME je iREP – mezera – iREP. Tento motiv byl genomem v minulosti rozšířen. Časem došlo k mnoha ztrátám či tandemovým amplifikacím jednotlivých BIME. V dnešní podobě tak lze v *E. coli* najít osamocené REP, tak složité BIME mozaiky obsahující mnoho REP i dalších motivů. Krom *E. coli* jsou BIME přítomny v mnoha dalších druzích, několik příkladů je na ilustraci 9. Jednotlivé REP jsou orientovány dle pozice GTAG tetranukleotidu.



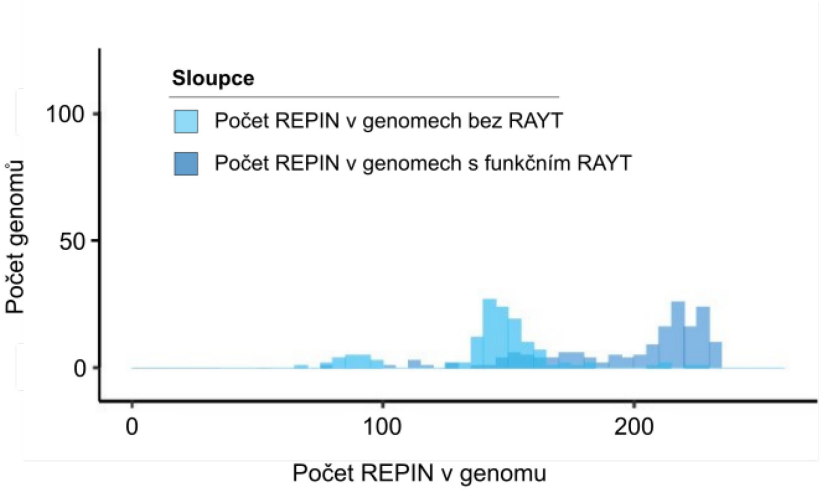
Ilustrace 9. Ukázka BIME mozaik z konkrétních lokusů E. coli (GCF_000005845.2, pozice: 0.7800–0.7805 Mbp), S. maltophilia (GCF_012647025.1, pozice: 3.5425–3.5435 Mbp) a P. putida (GCF_024749025.1, pozice: 1.763 – 1.765 Mbp). Opakované motivy v tandemu mají vždy sekvenční identitu nad 95%. Jsou odděleny spacery, ty mají v BIME tandemu stejnou sekvenci, mezi druhy se liší. REP sekvence jsou převzaty z prací (Aranda-Olmedo et al., 2002; Bachellier et al., 1999; Di Nocera et al., 2013).

3.7. RAYT

Nukleáza zodpovědná za propagaci REP byla objevena teprve v roce 2010 (Nunvar et al., 2010). Byla pojmenována RAYT (anglicky: „REP-associated tyrosine transposase“), sekvencí je příbuzná transpozázám rodiny IS*200*/IS*605*. RAYT nejsou transpozázy (podkapitola 4.9), jelikož nejsou schopny mobilizovat vlastní sekvenci, místo toho mobilizují

REP elementy, nejspíš preferenčně ze svého okolí (nicméně, označení RAYT transpozáza se stále používá). *Rayt* gen je vždy obklopen REP, jejich kopie bývají nejčastější a nejkonzervovanější typ REP v genomu. To je velmi pravděpodobně způsobeno tím, že právě tyto REP jsou RAYT šířeny.

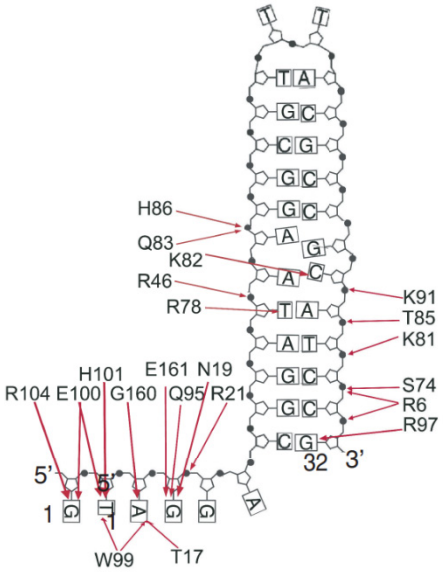
Přesný mechanismus, jakým REP rozeznávají a následně rozšiřují po genomu, není znám. Nicméně mobilizovanou jednotkou není samotný REP, ale vyšší struktura – REPIN. Krom poziční asociace, tedy že REP obklopují RAYT, byla také provedena kvantitativní analýza stovek genomů *E. coli*. Ta prokázala, že přítomnost funkčního RAYT je asociována s vyšším počtem REPIN (ilustrace níže).



Ilustrace 10. Porovnání množství nalezených REPIN v de-replikovaných genomech E. coli s funkčním RAYT (tmavě modrá) a bez RAYT (světle modrá). Převzato a upraveno z (Park et al., 2021).

RAYT jsou schopny vázat a následně štěpit řetězec ssDNA. RAYT rozezná strukturu palindromu REP, pro vazbu je kritická přítomnost GTRG tetranukleotidu u palindromu. U *E. coli* jsou popsány aminokyseliny nezbytné pro tuto vazbu. V oblasti blízko N terminu jde o 17T, 19N, 21R a druhá oblast obsahuje 95Q, 99W, 100E, 101H, 104R (ilustrace 11). Vazba RAYT-palindrom je zprostředkována přes cukr-fosfátovou kostru DNA (Messing et al., 2012). Dále jsou identifikovány ještě dvě AK v C terminální oblasti, ta ale není v rámci RAYT diverzity konzervována. U *E. coli* tato doména pravděpodobně funguje jako auto-inhibitor. Pravděpodobně kompetuje o vazbu se ssDNA REP elementu a je tedy mechanismem snižujícím aktivitu RAYT nukleázy.

Cílem štěpení je 5‘-CT-3‘ dinukleotid, který může být umístěn dále od REP (při jeho posunu bude stále nukleázou rozeznán a štěpen). Ke štěpení ssDNA vlákna RAYT využívají katalytický tyrosin (Y115 u *E. coli*) a HUH motiv (H59, M60, H61 u *E. coli*) (Ilyina & Koonin, 1992), ty jsou pro jejich funkci nezbytné. K reakci potřebují dvojmocný kofaktor, preferenčně Mn^{2+} ale se sníženou efektivitou využijí i Mg^{2+} - tento kationt je koordinován histidiny HUH motivu (Messing et al., 2012). Vyšší afinita RAYT k vzácnějšímu prvku mohla být selektována pro snížení aktivity proteinu.



Ilustrace 11. Sekundární struktura REP, ke které jsou přiřazeny aminokyseliny RAYT. Ty mají naznačeny pozice na nukleotidech, ke kterým se vážou. Převzato z (Messing et al., 2012)

RAYT se vyskytují u 24-26% druhů bakterií, které mají sekvenovaný alespoň jeden genom. Nicméně, nemusí vždy jít o funkční REP/RAYT systémy, ale pouze o sekvenčně podobné transpozázy. RAYT jsou vystaveny diverzifikující selekci (Bertels & Rainey, 2011b) a nikdy se v genomu nevyskytují ve více identických kopiích. Změny AK sekvence mohou ovlivnit, jaké REP budou RAYT rozšiřovat. Geny pro RAYT jsou součástí core genomu, jen minimálně se šíří horizontálním přenosem.

3.8. YPAL

REP a vyšší struktury z nich skládané nejsou jedinými nekódujícími repetitivními sekvencemi vyskytujícími se v bakteriálních genomech. Jedním z dalších jsou YPAL (anglicky: „*Yersinia pal*indromic sequences“). YPAL jsou DNA motivy vyskytující se v mnoha kopiích v genomech *Yersinia*. Poprvé byly popsány v roce 1999 (Bachellier et al., 1999). Jde o 168 bází dlouhé elementy s konzervovanou sekvencí. Jsou blíže zkoumány jen v několika studiích, neboť nejde o příliš rozšířené motivy (zvlášť v porovnání s REP). Známy jsou jen z *Yersinia* (BLAST nalezeny také u *Erwinia)*, kde se vyskytují nejvýše v 150 kopiích na genom (nejčastěji kolem 100). YPAL se pravděpodobně rozšiřují transpozicí. Transpozáza zodpovědná za jejich šíření není známa. Při inzerci dochází k duplikace cíle v délce od 3 po 26 bází. Cíl není sekvenčně konzervován, ale je AT-bohatý a obsahuje přerušený palindrom (De

Gregorio et al., 2006). Pravděpodobně se tak jedná o Rho-nezávislé terminátory transkripce (Farnham & Platt, 1981). YPAL se inzeruje vedle počátku palindromu a celá sekvence palindromu je duplikována (ilustrace 12).

YPAL se nacházejí v intergenových oblastech. Pokud YPAL lokus obsadí, typicky zde perzistuje a je v tomto lokusu přítomen i u příbuzných kmenů. Mobilizace YPAL je (alespoň u části kmenů *Yersinia)* stále aktivní proces (vlastní výsledky). YPAL jsou umístěny nejčastěji jsou downstream od stop kodonu sousedního genu. Díky tomu jsou YPAL transkribovány do mRNA, vytváří na ní vlásenku a zvyšuje tak její stabilitu (až 10x více transkriptu) (De Gregorio et al., 2006).

YPAL obsahují potenciální otevřený čtecí rámec, který kóduje 56 aminokyselin. Tato část je konzervovaná ve všech YPAL elementech *Y. pestis* a *Y. pseudotuberculosis*. Více než 70% identitu si sekvence udržuje u ostatních *Yersinia*. Na C-konci hypotetického proteinu je krátká hydrofobní doména. Ta (při fúzi čtecího rámce YPAL se sousedním proteinem) může sloužit jako transmembránová doména či signální sekvence (Delihas, 2007).

Ilustrace 12. Okolí několika YPAL elementů (schematicky zaznamenány jako šipky). Podtrženy jsou palindromické oblasti Rho nezávislých terminátorů transkripce. Převzato z (De Gregorio et al., 2006)

3.9. Inzerční sekvence

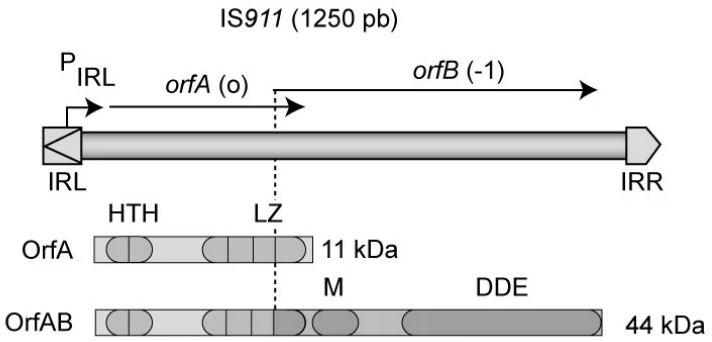
Inserční sekvence jsou sobecké DNA elementy. V porovnání s REP či YPAL jsou podstatně delší a obsahují protein (tzv. transpozázu) kódující oblast, díky němu jsou schopné autonomně měnit svou polohu v genomu. Sekvence kódující transpozázu je obklopena krátkými nekódujícími oblastmi. Ty transpozáza rozezná, celou sekvenci vyštěpí a následně vloží jinde v genomu. U řady IS při tom dochází k duplikaci transponované sekvence. IS lze rozdělit dle cíle inserce. Většina nemá specifický cíl, hostitelská bakterie je tak vystavena riziku inzerce do esenciálního genu (což je pro ni letální). Nezávisle na sobě několik skupin vyvinulo schopnost sekvenčně či strukturně specifické inzerce (W. Y. Hu et al., 2001; Olasz et al., 1998; Williams, 2002). Níže jsou popsány rodiny IS, které se hojně vyskytují u rodu *Stenotrophomonas* a budou analyzovány ve výsledcích (kapitola 6.7).

3.9.1. Rodina IS*3*

IS*3* je rodina inzerčních sekvencí, charakteristická dvěma kódovanými proteiny. Je to jedna z nejstarších definovaných skupin, prvně popsána při inaktivaci *gal* a *lac* operonů (Ahmed & Scraba, 1975; Hu et al., 1975; Ptashne & Cohen, 1975; Rosselin et al., 1968). Později byly IS*3* nalezeny na řadě plazmidů gram negativních bakterií. Dnes je IS*3* jednou z největší a nejrozšířenějších rodin. Vyskytuje se u stovek bakteriálních druhů, nicméně reálná distribuce je pravděpodobně řádově vyšší. Zástupce rodiny, IS*1397,* u *E. coli* specificky inzertuje do všech tří druhů REP, které se v genomu vyskytují (Clément et al., 1999; Wilde et al., 2003).

3.9.1.1. Struktura IS3

Délka se pohybuje od 1200 do 1550 bp, typicky jsou IS ohraničeny konzervovanými inverzními terminálními repeticemi, ty mají od 20 do 40 bp. Téměř vždy začínají 5‘ – TG a častý je v jejich sekvenci tetranukleotid jedné báze (například CCCC) (Zekrí & Toro, 1996). Inverzní repetice bývají přerušovány nepárujícími nukleotidy. Po inzerci vytvářejí 3-4 bp duplikace. IS*3* nesou dva čtecí rámce, OrfA a OrfB (ilustrace 13). OrfA kóduje krátký protein o délce kolem 11,5 kDa. Ten nese HTH motiv (helix-turn-helix), pomocí něhož transpozáza rozeznává inverzní repetice při transpozici. Dále také obsahuje C-terminální leucinový zip, který slouží pro dimerizaci. OrfB nese DDE katalytický motiv nezbytný pro štěpení DNA.



Ilustrace 13. Struktura DNA kódující IS911 z rodiny IS3. Kódující oblast je obklopena terminálními inverzními repeticemi, které jsou vzájemně komplementární (IRL, IRR). Důležité motivy čtecích rámců OrfA a OrfAB jsou níže. Převzato z (Chandler et al., 2015).

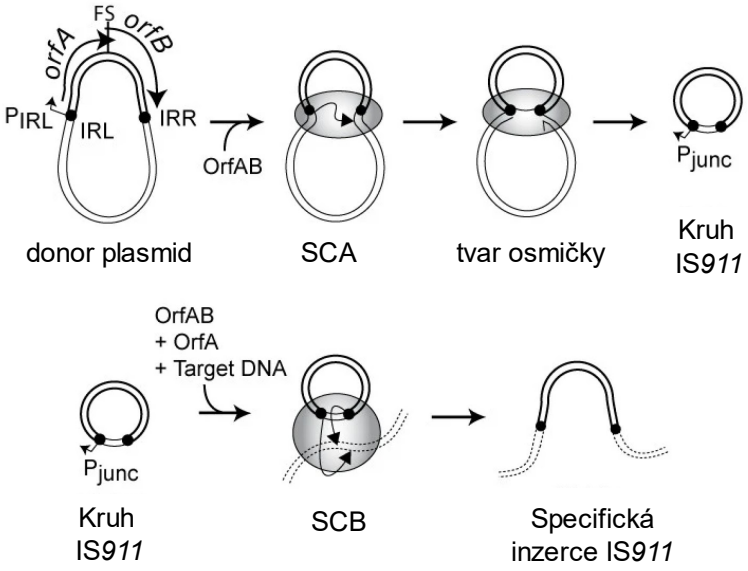
Čtecí rámec OrfB neodpovídá aminokyselinové sekvenci výsledného funkčního proteinu. Při translaci dochází k takzvanému programovanému ribosomálnímu frameshiftu, na sekvenci 5‘-AAAAAAG-3‘ (u IS*911*) (Chandler et al., 2015; Fayet et al., 1990; *ISfinder*, n.d.-a; McAdam et al., 1990; Polard et al., 1991). Motiv A(6)G je typickým pro sklouznutí ribozomu, ale v rámci IS*3* existují i další varianty. Exprese funkční transpozázy je snížená, neboť sklouznutí ribozomu a posunutí čtecího rámce je vzácné. Většinou je vyráběn nefunkční produkt, který je rychle degradován.

OrfAB je výsledkem programovaného translačního frameshiftu, prvně pozorovaného u IS*150* (Farabaugh, 1996). Z OrfA bere důležitý HTH motiv a leucinový zip, z OrfB DDE katalytický motiv. Kombinací tak vzniká protein schopný dimerizace, rozeznání konzervovaných okrajů IS a štěpení DNA. Dimer OrfAB je udržován solnými můstky v kombinaci s vnitřním hydrofobním jádrem a elektrostatickými interakcemi (Haren et al., 1998).

Frekvence frameshiftu se mezi IS*3* značně liší, do velké míry je ovlivněna sekvencí DNA v okolí, která je schopná vytvářet stem-loop struktury, které mají vliv na frameshift. Od nejméně častého frameshiftu během translace u IS*3411* (2%), přes IS*3* (6%), IS*911* (15%) až po IS*150* (50%) (měřeno *in vitro*) (Mazauric et al., 2008). Vliv má také fyziologický stav buňky, zvýšení teploty či vyšší hustota ribozomů na mRNA snižuje frekvenci frameshiftu.

3.9.1.2. Transpozice IS3

„Životní cyklus" IS*3* transpozáz je prozkoumán na modelu IS*911*. OrfAB se váže na nekódující okraje své inserční sekvence (Duval-Valentin & Chandler, 2011). Vazba na DNA imobilizuje OrfAB, stericky je zabráněno transpozici proteinem exprimovaným jinou kopií IS*3* v genomu. Translace je přímo spojena s dimerizací a vzniku takzvaného synaptického komplexu transpozázy s DNA v cirkulární formě. (Haren et al., 2000) Transpozice IS*911* je replikativní excize následována insercí (ilustrace 14). Nejdříve dimer transpozázy drží oba konce IS u sebe v rámci kruhového synaptického komplexu (během štěpení DNA a jejím transportu), poté je jeden konec IS uvolněn. Volný konec je schopen zahájit integraci (Chandler et al., 2015; Turlan et al., 2004).



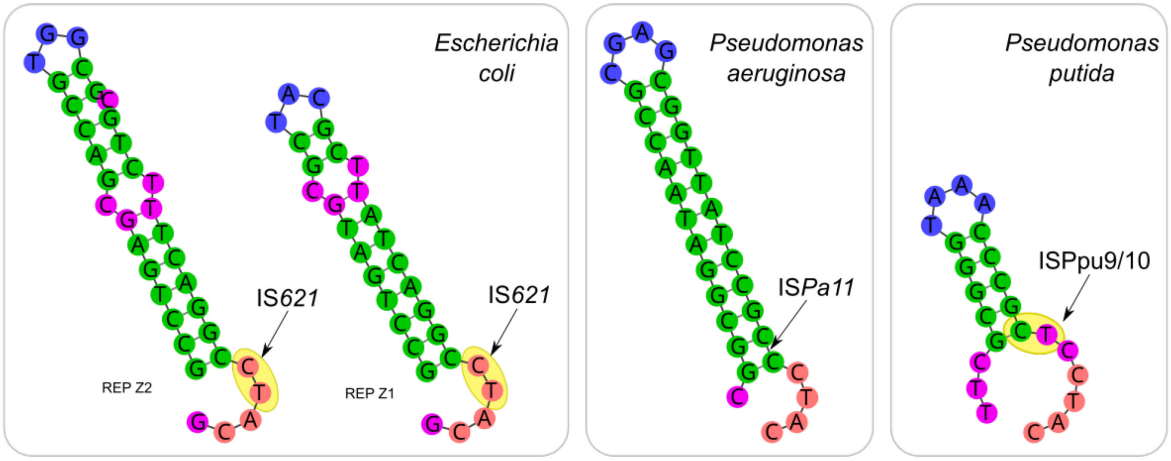
Ilustrace 14. Schéma transpozice IS911. Dimer transpozázy vyštěpí sekvenci IS přes synaptický komplex A (SCA) za vzniku kruhového intermediátu. Při vzniku kruhového intermediátu je sestaven silný promotor Pjunc. Kruhový intermediát je schopen se vložit na nové místo v DNA přes synaptický komplex B (SCB). Malé černé oblasti reprezentují inverzní repetice, poloprůhledná elipsa pak transpozázu. Převzato z (Chandler et al., 2015).

Pravá inverzní repetice nese -35 promotorovou oblast, ta směřuje od ORF transpozázy. Levá inverze obsahuje -10 promotorovou část směřující k ORF. Jejich kombinací vzniká silný promotor, *pJunc*. Ten vzniká pouze, je-li IS*3* ve stavu kruhového intermediátu DNA. (Duval-Valentin et al., 2001; Ton-Hoang et al., 1997). Složený promotor je až 50x silnější než původní promotor transpozázy a hraje roli při kontrole transpozice. Cirkulární forma se silným pJunc promotorem pravděpodobně slouží jako zdroj nových lineárních forem, které poté integrují do genomu (Sekine et al., 1999).

3.9.2. Rodina IS*110*

První IS*110* byly charakterizovány roku 1985 v genomu *Streptomyces coelicolor* (Chater et al., 1985). Dnes jsou popsány u stovek druhů bakterií i archea (Krupovic & Forterre, 2015). Při inzerci část IS*110* vytváří krátké duplikace (nejčastěji dvě báze). Katalytickým AK motivem je DEDD, jde o apomorfii rodiny IS*110* (Buchner et al., 2005). Nekódující okraje IS neobsahují inverzní repetice, které jsou typické pro řadu IS rodin (Siguier et al., 2015). Mechanismus inzerce není detailně prozkoumán, ale probíhá přes kruhový intermediát.

Jedním z více prozkoumaných členů rodiny IS*110* je IS*492*, z *Pseudomonas atlantica*. Tato IS se inzertuje do specifických oblastí, jednou z nich je gen extracelulárních polysacharidů, který tím deaktivuje. Excize poté vede ke zpětnému obnovení funkce (Bartlett & Silverman, 1989). Řada IS*110* jako svůj inzerční cíl využívá REP elementy (ilustrace 15). Konkrétně IS*621* u *E. coli* má všechny kopie (10) vloženy v REP (Choi et al., 2003). Také IS*Ppu10* v genomu *Pseudomonas putida* KT2440 má všech 8 kopií IS vložených do REP. Příbuzná IS*Ppu9* se vkládá do stejného REP a stejné pozice, ale v opačné orientaci. IS*Pa11* se specificky inzertuje do REP *Pseudomonas aeruginosa*.

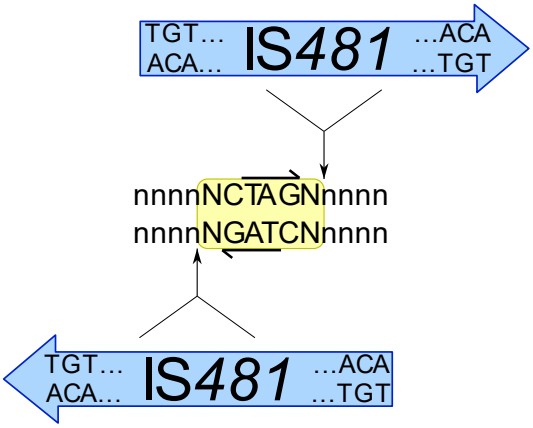


Ilustrace 15. Inzerce IS621, ISPa11, ISPpu9 a ISPpu10 do REP. Struktura REP je zvýrazněna dle schéma na ilustraci 7 a je používána po zbytek práce. Data z (Choi et al., 2003; Tobes & Pareja, 2006). Žlutě jsou zvýrazněny nukleotidy duplikované při inzerci. Sekundární struktury predikovány v Geneious prime, upraveno v Inkscape.

3.9.3. Rodina IS*481*

IS*481* byla poprvé identifikována u *Bordetella pertussis*. Původně byla řazena jako odvozený typ rodiny IS*3* (McLafferty et al., 1988). Sdílí s nimi zkrácený N-terminus a prodloužený C-terminus. Na rozdíl od IS*3*, funkční protein nevzniká frameshiftem a není přítomen leucinový zip. Je tak považována za samostatnou rodinu IS elementů, dnes popsanou ve více než 130 druzích. IS*481* mají terminální repetice (28 bp) a kódují transpozázu dlouhou asi 300 AK s DDE katalytickou doménu. Část IS*481* cílí svou inzerci do 5‘-NCTAGN-3‘ hexanukleotidu (Ilustrace 16).

Po inzerci je 5‘-NCTAGN-3‘umístěn downstream v kódující oblasti a TAG funguje jako STOP kodon transpozázy. Hexanukleotid je při inzerci duplikován (Stibitz, 1998). Blízce příbuzné elementy byly nalezeny v genomech některých eukaryot. Je tak možné, že se tyto transpozázy v minulosti rozšířily i mimo bakterie (Kojima & Bao, 2023).



Ilustrace 16. Inzerce IS481 do NCTAGN hexanukleotidu, který je při transpozici duplikován (podbarveno žlutě). Duplikovaná sekvence slouží také jako STOP kodon transpozázy (podtrženo). IS481 má vždy terminální okraj 5‘-TGT-3‘, který je na druhém okraji komplementován. Připraveno v Inkscape.

Bordetella bronchiseptica je bakterie způsobující chronická respirační onemocnění u řady zvířat. V evoluci prodělala masivní amplifikaci transpozáz IS*481*, které se rozšířily do stovek oblastí v genomu. Tyto homologní oblasti pak posloužily jako klíčové body pro masivní redukci genomu. Během tohoto procesu se z ancestrální *B. bronchiseptica* odštěpily dvě linie, *B. parapertussis* a *B. pertussis. B. pertussis* je striktně lidský patogen způsobující černý kašel. Při porovnání dnešních genomů lze odhadnout, k jak velké redukci došlo. Kmen *B. bronchiseptica* RB50 má genom o délce 5,33 Mbp a kóduje 5007 genů. Genom *B. pertussis* Tohama I je o 1 Mbp kratší (4,09 Mbp) a o asi 600 genů chudší (Parkhill et al., 2003; Weigand et al., 2017). Masivní amplifikací IS tak lze dosáhnout rozsáhlých přestaveb a redukce genomu.

4. Materiál a metody

V práci jsou využity výhradně bioinformatické postupy. Byla použita široká škála nástrojů a bioinformatických „pipelines“ (programy/skripty pracující v sérii po sobě). Největší část analýza probíhala v prostředí Geneious prime, které zahrnuje mnoho nezbytných nástrojů a dovoluje přehledně kombinovat jednotlivé metody a výsledky intuitivně vizualizuje.

4.1. Získání genomů rodu *Stenotrophomonas* a tvorba fylogramu

Z databáze NCBI assembly (*Home - Assembly - NCBI*, n.d.) byly získány všechny kompletní genomové sekvence rodu *Stenotrophomonas*, tedy *S. maltophilia, S. acidaminiphila, S. pavanii, S. indicatrix, S. rhizophila, S. nitritireducens, S. sp.* Byla zpracována aktuální literatura zaměřená na komparativní genomiku rodu a přidány další genomy, které k rodu patří. Z nich plně sekvenovaný byl kmen *Pseudomonas geniculata* E119. Celkem bylo staženo 102 záznamů (září 2021). Spolu s nimi byl vybrán outgroup genom, který je *Stenotrophomonas* příbuzný, takže bude větví, jejímž prostřednictvím lze strom *Stenotrophomonas* zakořenit. Byl vybrán druh *Xanthomonas* translucens, kmen ICMP11055, neboť rod *Xanthomonas* je sesterský k *Stenotrophomonas* (viz podkapitola 4.1).

Fylogram byl připraven s pomocí serveru CSI phylogeny. Bylo použito defaultní nastavení programu, pouze kritérium „minimální vzdálenost SNP pozic“ nebylo aplikováno. Pro konstrukci stromu byla použita metoda Maximum likelihood (ML) s bootstrap 10 000 (počet opakování výpočtů pro kalkulaci jedné větve). Vizualizován a upraven byl v online prostředí ITOL (ilustrace 25). Ke stromu byla také připravena heatmapa průměrné podobnosti nukleotidových sekvencí (ANI), ta byla vytvořena python skriptem využívajícím fastANI.

Jednotlivé klastry kmenů, které fylogram predikoval, byly klasifikovány dle komparativně genomických prací zaměřených na rod Stenotrophomonas (Gröschel et al., 2020b; Mercier-Darty et al., 2020; Vinuesa et al., 2018).

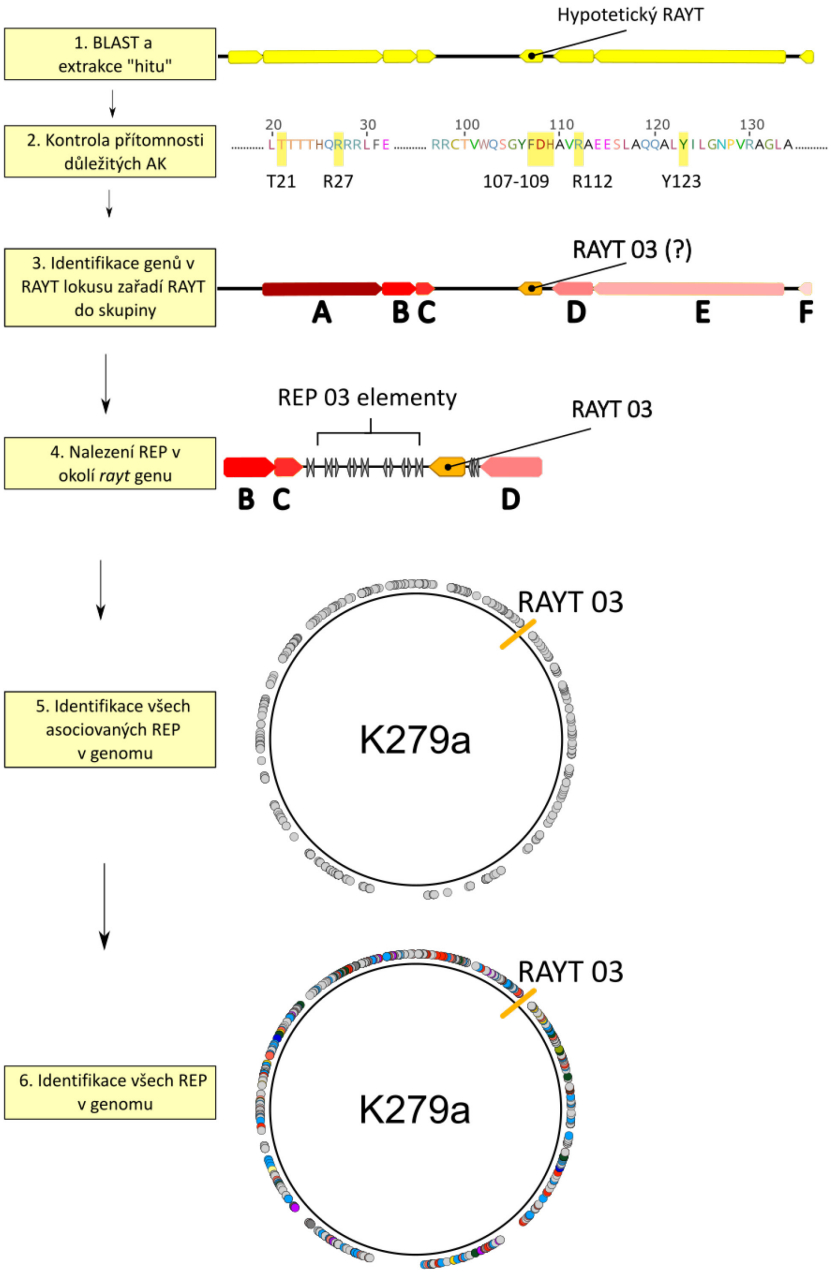
Výsledky CSI Phylogeny

Velikost referenčního genomu	4 761 583 bp
Pozice invariantní ve všech genomech	399 964 bp
Procent referenčního genomu přítomna ve všech genomech	8,3998%
Počet variantních pozic v MEGA*	132 460

*Tabulka 1. Výsledné statistiky celogenomového alignmentu 103 analyzovaných kmenů na serveru CSI Phylogeny. *Tyto pozice sloužily napříč alignmentem jako substrát pro fylogram.*

4.2. Identifikace RAYT proteinů

Pro *de novo* identifikace RAYT byla použita AK sekvence již známého proteinu ze *S. maltophilia* R551-3 (NCBI lokus: ACF53164). BLAST této sekvence odhalil všechny potenciální RAYT u *Stenotrophomonas* (400 potenciálních RAYT). Globální homologie aminokyselinových sekvencí pro identifikaci není dostatečná, vzhledem k existenci blízce příbuzných proteinů, které nejsou asociovány s REP (Bertels, Gallie, et al., 2017) a *bona fide* transpozázám rodiny IS*200*/IS*605*. Potenciální RAYT musejí mít konzervované AK nezbytné pro vazbu GTRG tetranukleotidu. Jedná se o threonin (pozice 21), dále arginin (pozice 27 a 112), glutamin (pozice 103), tyrosin (pozice 123) a motiv na pozici 107-109 (Nunvar et al., 2010) (klíčové AK a postup jsou na ilustraci 17).



Ilustrace 17. Postup de novo identifikace rayt genu a jeho lokusu, včetně asociovaných REP.

Po tomto kroku bylo analyzováno okolí *rayt* genů pro rozdělení do skupin. Genové okolí *rayt* („RAYT lokus“) je v evoluci konzervované – příslušné lokusy jsou syntenní. Bylo identifikováno jedenáct genových oblastí, ve kterých jsou *rayt* přítomny (příloha 10.2). Jednotlivé RAYT byly označeny podle syntenního lokusu, ve kterém jsou kódovány, tj. 01-11. Podle naší hypotézy RAYT rozšiřuje REP vyskytující se v jeho bezprostředním okolí. Tyto REP byly nalezeny vizuálním prohledáváním intergenových oblastí přiléhajících z obou stran k *rayt* (přítomnost GTRG následovaného palindromem) a, paralelně, pomocí serveru Palindrome analyser (Brázda et al., 2016). Většinou platí, že RAYT kódovaný v daném lokusu je asociovaný se stejným typem REP, a jsou proto označeny stejně jako asociovaný RAYT. Výjimkou jsou RAYT/REP 01, které jsou extrémně variabilní, a v menší míře RAYT/REP 02 a 04 (viz podkapitola 6.2.2). Tyto REP jsou proto děleny do tříd 1A-01F, 02A-02B a 04A-04G. Referenční sekvence RAYT jednotlivých lokusů jsou součástí přílohy 10.3.

Po rozdělení a ověření RAYT skupin, byly nalezeny všechny REP ve všech genomech *Stenotrophomonas*. REP byly vyhledávány pomocí funkce Motif finder v Geneious Prime (nastavení: nula chyb, hledání na vedoucím i opožděném vláknu DNA).

4.3. Cirkulární mapa REP

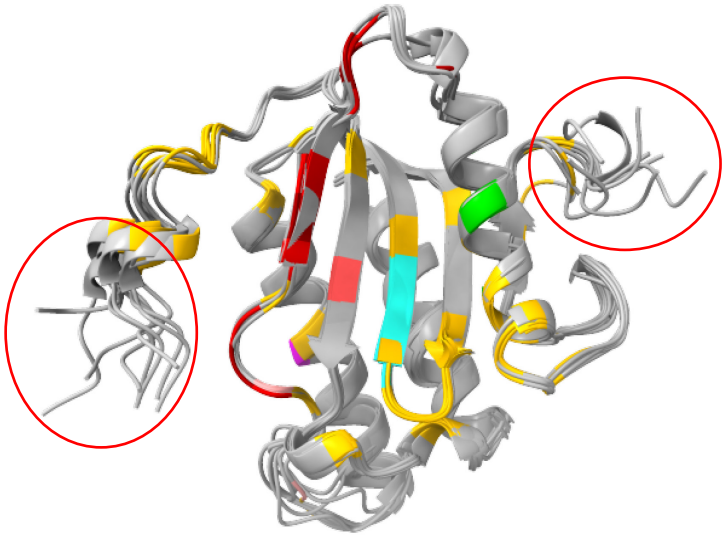
Cílem bylo vytvořit mapu REP, která porovnává globální rozšíření REP napříč všemi genomy. Bude na ní možné odhalit REP bohaté/chudé oblasti a jejich konzervovanost napříč fylogenezí rodu. Prvním krokem bylo zarovnání genomů dle počátku replikace (DnaA). V takto upravených genomech byly anotovány pozice všech REP (Motif finder). Data o REP (počátek, konec a typ) a genomu, ve kterém se nachází (jméno, délka), byla exportována. Vzhledem k tomu, že genomy mají variabilní délku (rozdíl až 1,4 Mbp), bylo nutné data normalizovat. Normalizovaná pozice REP (μ) je definována dle vzorce:

$$\mu = \frac{(počátek\ REP + konec\ REP) * 0{,}5}{délka\ genomu}$$

Kmeny byly uspořádány dle jejich pozice na fylogramu rodu, 1. je bazální *S. nitritireducens* 2001 a 102. je nejodvozenější *S. m.* K279a. Poté byl v R připraven cirkulární graf s využitím knihovny circlize (Gu et al., 2014). Y osa je složena ze 102 koncentrických kruhů, které reprezentují jednotlivé genomy (úplně centrální kruh je *S. n.* 2001). Na X ose jsou REP umístěny dle své normalizované pozice (μ). REP jsou zobrazeny jako body zbarvené podle druhu REP. Výsledkem je ilustrace 36 v podkapitole 5.3. Obdobně byla připravena mapa KOPS elementů (ilustrace 40).

4.4. Predikce struktury RAYT proteinů

Struktury proteinů byly predikovány s pomocí upravené verze AlphaFold 2.3.2 (Jumper et al., 2021). Konkrétně byl využit Google Colab notebook – AlphaFold Colab (*AlphaFold.Ipynb - Colaboratory*, n.d.). Jedná se o zjednodušenou verzi AlphaFold, pracující na stejném principu. Software nevyužívá templátů (homologních struktur) a pracuje se



Ilustrace 18. Porovnání modelů zástupců všech RAYT ze Stenotrophomonas. Nabarveny jsou konzervované oblasti (žlutá), katalytický tyrosin (zelená), HUH motiv (tyrkysová) a REP vazebné AK (červená). Struktury jsou v superpozici dle homologie (algoritmus Matchmaker). Jsou zakroužkovány oblatí konců, které mají dle AlphaFold nižší kvalitu predikce. Data z AplhaFold2, připraveno v ChimeraX.

subsetem BFD databáze (Steinegger & Söding, 2018). Tím výrazně snižuje procesní nároky, při malém snížení přesnosti.

Byl připraven model struktury zástupce každého RAYT z 11 skupin (ilustrace 18). Jako zástupce byl vždy vybrán RAYT z genomu s nejvyšším počtem asociovaných REP. Alignment sekvencí všech 271 RAYT odhalil konzervované pozice, které byly na struktuře zvýrazněny. REP vazebné aminokyseliny byly převzaty z literatury (Messing et al., 2012).

Struktury s výrazněnými klíčovými oblastmi byly vizuálně porovnány v prostředí ChimeraX. Struktury všech zástupců jsou vzájemně velice podobné, pozice konzervovaných aminokyselin pak totožné. Zvlášť potenciální REP vazebné oblasti mají vysokou koncentraci konzervovaných pozice a strukturně jsou identické. Pro další studium byl použit model jednoho zástupce – RAYT 03 z genomu *S. m.* K279a, viz tabulka níže.

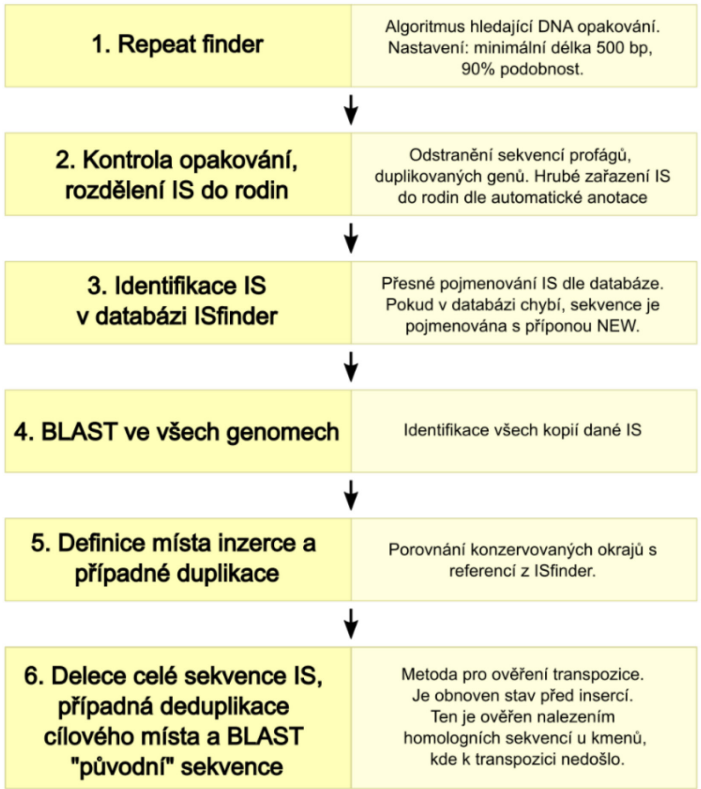
NCBI reference	**Sekvence proteinu**
WP_024958589.1	MSSHRLRLGRHSIIGQSYVLTTTTHQRRRLFESEAAAACVIDQFHYIEQRGLVQSHAWVV MPDHVHWMFELRAAHLPDIARRMKSSSALALNRLVGRRCTVWQSGYFDHAVRAEESL AQQALYILGNPVRAGLAGQIGEYPYAWSVWL

Tabulka 2. AK sekvence RAYT 03 z K279a, použitá k predikci modelu RAYT. Model byl poté využit při dalších strukturních analýzách.

4.5. Analýza inserční sekvencí

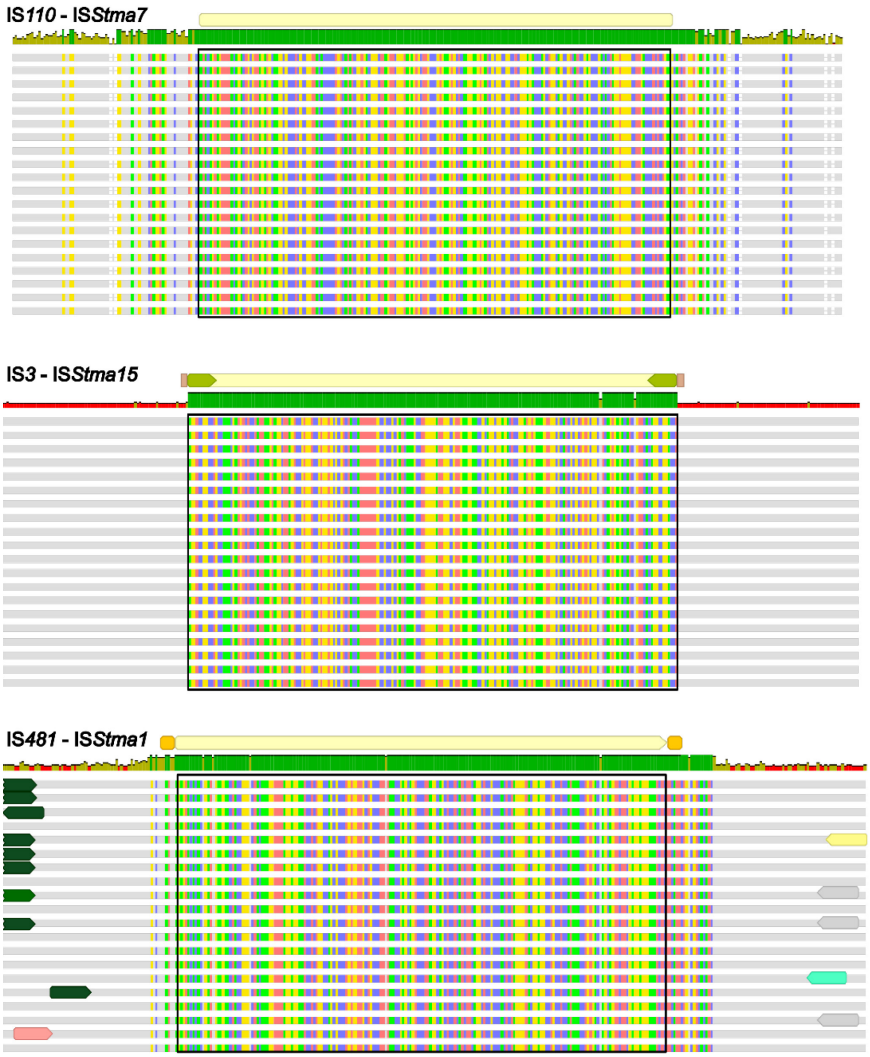
Prvním krokem analýza bylo nalezení a klasifikace všech IS přítomných u *Stenotrophomonas*. Genomy byly z NCBI staženy včetně automatické anotace genů/proteinů, nicméně ta většinu transponáz rozděluje pouze do rodin, jiné transponázy anotuje jen jako „transposase“, případně „hypothetical protein“). Proto byl zvolen *de novo* přístup, jak detekovat IS v genomech (ilustrace 19). Pomocí repeat finder byly nalezeny opakující se sekvence v jednotlivých genomech. Průměrná délka IS je přibližně 1000 bp, nicméně se v nich často vyskytují frameshift mutace. Minimální délka repetice byla proto nastavena na 500 bp. Minimální identita mezi kopiemi repetitivní DNA byla nastavena na 90%, aby byly zachyceny sekvence mírně degenerované ale stále patřící stejné transpozáze.

V každém genomu byly IS/repetitivní sekvence hledány nezávisle. Nalezené repetice byly manuálně zkontrolovány, zda se jedná opravdu o transpozázy. Byly vyřazeny například kopie profágů či běžně nemobilních genů. Jednotlivé IS byly nahrubo zařazeny do rodin dle automatické anotace. Po kontrole všech genomů byla sekvence transpozázy vyhledána v ISfinder databázi. Pokud byl nalezen záznam IS s vysokou sekvenční podobností a původem ze *Stenotrophomonas/Pseudomonas/Xanthomonas*, byla IS pojmenována podle záznamu. Alternativně byla transpozáza pojmenována de novo jako IS*StmaNEW*(*číslo*). Kopie této IS byly pomocí BLAST vyhledávány ve všech genomech. „Cut-off“ bodem BLASTu byla sekvenční podobnost minimálně 90% v 90% délky alignmentu, méně podobné (a kratší) sekvence nejsou považovány za kopie jedné transpozázy. Tyto IS byly klasifikovány a krok byl opakován, dokud nebyly klasifikovány všechny repetice.



Ilustrace 19. Postup de novo identifikace inserčních sekvencí. Provedeno v Geneious prime. Pravá strana je vždy krátký popis kroku.

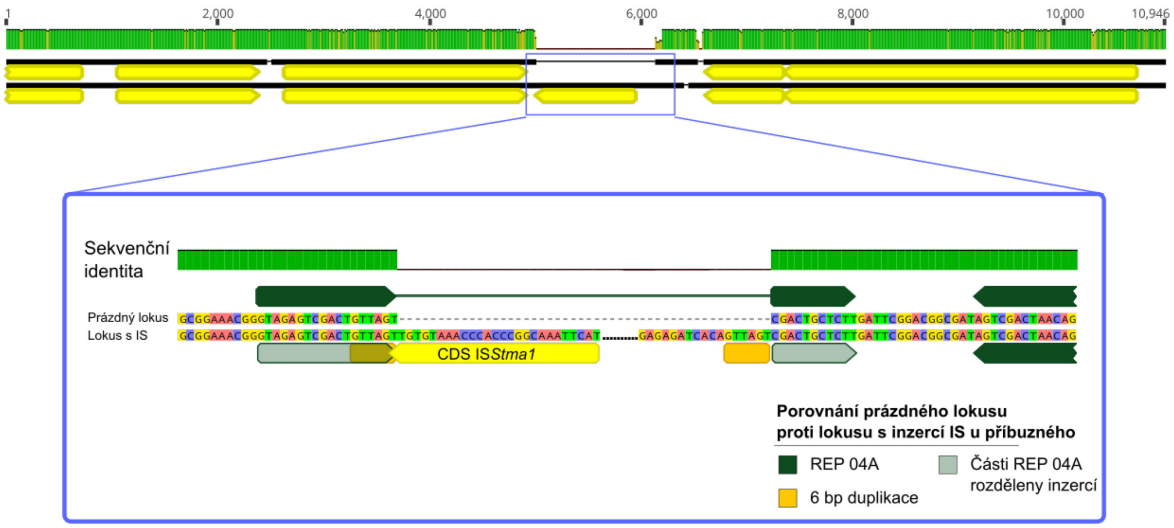
Po identifikaci všech IS byla věnována pozornost jejich transpozici. Ze všech kopií jedné transpozázy byl připraven alignment – metoda Geneious alignment. Bylo použito upravené nastavení s vysokou penalizací za otevření a expanzi mezer v alignmentu. Výsledné alignmenty tak jasně odhalují konce homologií inserčních sekvencí. Několik ukázkových alignmentů znázorňuje ilustrace 20.



Ilustrace 20. Porovnání tří alignmentů inserčních sekvencí. Alignmenty jsou upraveny, délka IS je zkrácena, také počet IS v alignmentu je snížen. Každý alignment porovnává kopie jedné inserční sekvence v Stenotrophomonas, včetně jejích okrajů a okolní sekvence. Zobrazeny nejsou báze, pouze jejich barevný kód, navíc jsou zbarveny jen pozice plně konzervované. Nad alignmentem je vyznačena délka mobilizované DNA inserční sekvence (žlutá), duplikace vzniklé transpozicí (oranžová) a v případě IS3 také inverzní repetice (zelená). REP jsou barveny dle standartního schéma.

Konce oblastí homologie v těchto alignmentech byly dále podrobeny pečlivé vizuální analýze pro determinaci skutečných konců IS, případně konzervovaných cílů inzerce. U IS, které vykazují nízkou či žádnou sekvenční specificitu transpozice (IS*Stma15*, ilustrace 20), odpovídá konec homologní oblasti alignmentu konci samotné IS. Pokud dochází k inzercím do specifických sekvencí DNA, budou také přítomny v oblastech terminální homologie. Proto byly sekvence těchto oblastí porovnávány se sekvencemi REP elementů *Stenotrophomonas*,

což vedlo k odhalení několika IS, které se inzertují do REP (viz 6.7) – terminální homologie v těchto případech přesně odpovídaly dvěma fragmentům původního REP elementu (ilustrace 21). Analogicky, terminální homologie několika dalších IS odpovídaly nově objevenému repetitivnímu elementu, YPAL (viz dále). Na úplném okraji takto definovaných sekvencí IS pak byly presumptivně identifikovány případné duplikace cílového místa (krátké přímé repetice).



Ilustrace 21. Alignment dvou 10 kbp syntenních oblastí – obsazená ISStma1 a bez IS (příbuzný kmen). Dole poté schematicky zkrácená oblast inzerce ISStma1.

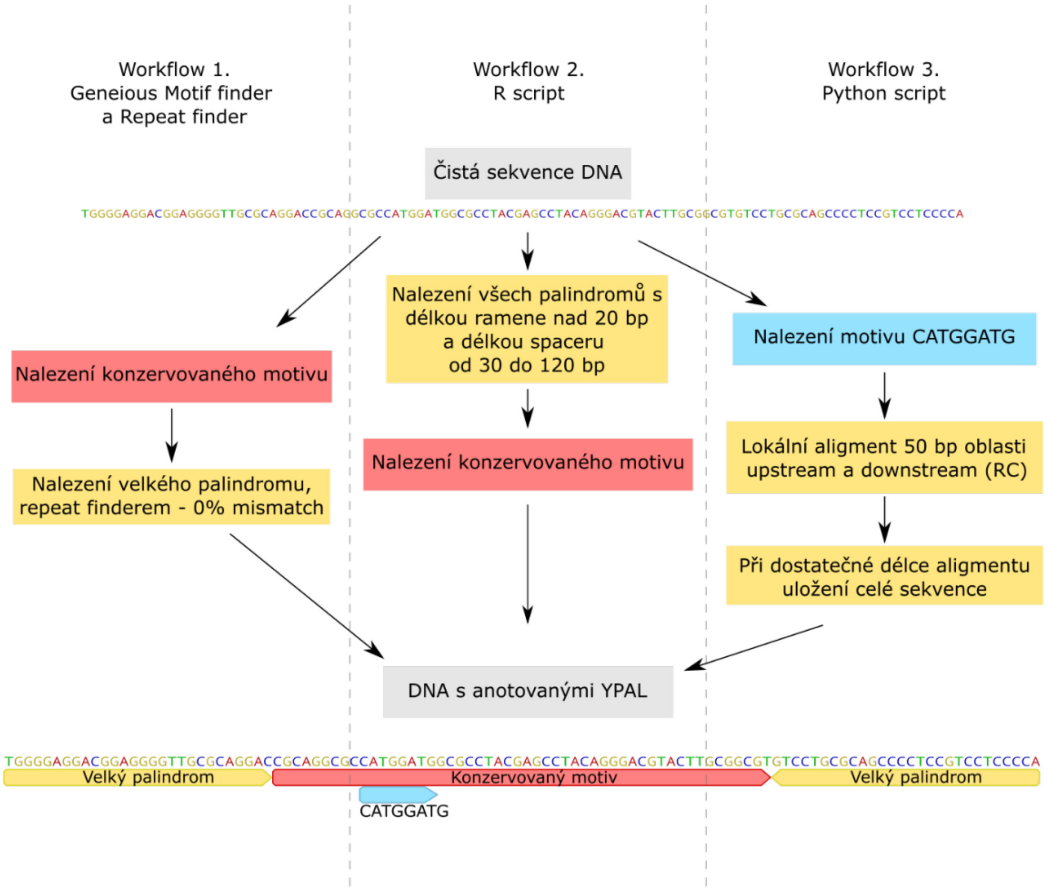
Pro definitivní potvrzení sekvencí IS elementů (včetně přesných konců) a případné sekvenčně-specifické inzerce byla transpozice rekonstruována *in silico*. Z lokusů reálných výskytů IS elementů byla odstraněna předpovězená sekvence IS a případně jedna kopie duplikované sekvence cílového místa inzerce. Tím byla získána hypotetická sekvence daného lokusu před inzercí IS. Tato sekvence byla pomocí BLAST porovnávána s genomy příbuzných kmenů *Stenotrophomonas*. Tímto postupem bylo (u všech analyzovaných IS) potvrzeno, že předpovězený stav před inzercí IS je reálný a tedy, že předpovězené charakteristiky IS platí.

4.6. Analýza YPAL

YPAL elementy nalezené u *Stenotrophomonas* mají typickou charakteristiku a sekundární strukturu jako YPAL u *Yersinia*, nikoliv však sekvenci. Alignment těchto sekvencí odhalil silně konzervovanou sekvence 5‘-CATGGATG-3‘, zbytek sekvence (přes 100 bází) konzervovaný není. Vnitřní část dlouhá 52 bází je částečně konzervovaná a lze ji definovat s pomocí degenerovaných bází (dle pravidel IUPAC).

Tento střední motiv byl dále použit k prohledání genomů a nalezení všech sekvencí YPAL. Pomocí funkce Motif finder (součást in-build funkcí Geneious Prime) byl degenerovaný motiv s povolenými až osmi chybami hledán. Tato hodnota byla stanovena empiricky, jelikož při vyšší hranici jsou již zachyceny i oblasti uvnitř genů, které YPAL nejsou. Nicméně, použití pouze jediné metody se nezdálo jako spolehlivý způsob, jak zachytit

všechny YPAL. Proto bylo navrženo několik různých způsobů, jak je nalézt (ilustrace 22). Výsledné datasety byly poté srovnány a spojeny, díky tomu je zachyceno maximum YPAL a zároveň jsou potlačeny slabiny jednotlivých metod (falešně pozitivní výsledky, anotace v neúplné délce). Z důvodu velkého množství analyzovaných dat, byl využit subset dvaceti kmenů *Stenotrophomonas* (viz metody 5.7).

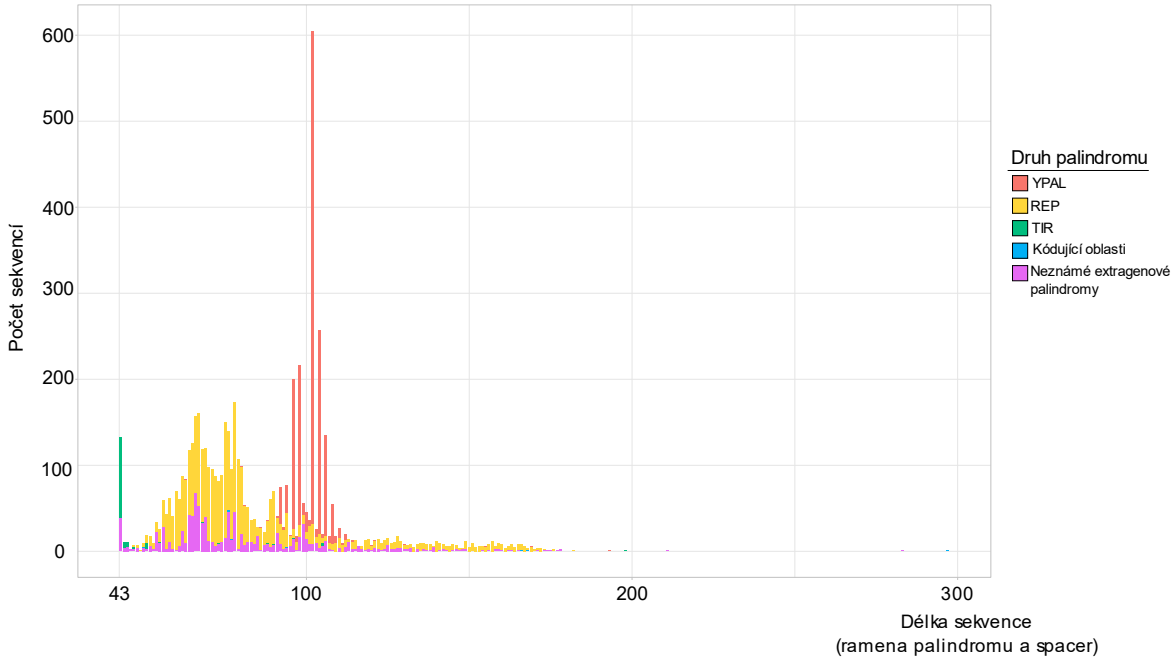


Ilustrace 22. Tři přístupy využité při hledání YPAL. Každý sloupec reprezentuje jednu metodu. Vždy se začíná s čistou sekvencí DNA, v té je různým způsobem identifikována střední část YPAL a poté inverzní ramena. Výsledná data jsou zpětně importována do genomů a všechny YPAL jsou anotovány. Data pak byla využita k dalším analýzám.

První metoda využívá funkcí Geneious prime. Pomocí degenerovaného motivu byly nalezeny vnitřní oblasti YPAL. Ty byly společně s 50 bp upstream a downstream sekvencí extrahovány. Poté byla použita funkce Repeat finder, která nalezne identické oblasti (včetně reverzně komplementárních). Podmínkou bylo, aby délka opakování byla minimálně 20 bp a maximálně 10% rozdílná. Výsledná sekvence je považována za kompletní YPAL. Nevýhodou této metody je, že není schopná odhalit indel mutace. Také neodhalí YPAL s více odvozenou sekvencí.

Z těchto důvodů byl napsán skript v R (**druhá metoda**) využívající knihovny Biostrings a funkce findPalindromes (Pagès H. et al., n.d.). Tato funkce vyhledává přerušené palindromy. Funkce byla nastavena, aby hledala sekvenčně nespecifické palindromy s délkou ramene nad 20 bází a mezerou mezi rameny 30 až 120 bp. Slabinou přístupu je množství redundantních dat. DNA bakterií obsahuje řadu palindromů, vzniklých stochasticky i selekcí. Zachyceny jsou velice často dvojice REP (REPIN či BIME), TIR elementy (terminální inverzní

repetice) (Di Nocera et al., 2013) a další. Také může být zachycena dvojice YPAL jako ramena palindromu. Hrubá data lze použít jako určitou kontrolu rozšíření palindromů mezi kmeny a zda je naše analýza zachytila. Data byla pomocí Geneious anotována a bylo určeno o jaký (nám známý) typ palindromu se jedná (ilustrace 23). Jde o analýzu části kmenů (subset 20 kmenů popsaný v podkapitole 5.7).



Ilustrace 23. Sloupcový graf nalezených palindromů u 20 vybraných kmenů. Kritérii jsou délka palindromu nad 20 bází, mezera mezi rameny mezi 3 až 120 bázemi, nebyly tolerovány žádné chyby v palindromu. Sekvence nalezeny pomocí R skriptu, anotovány pomocí Geneious.

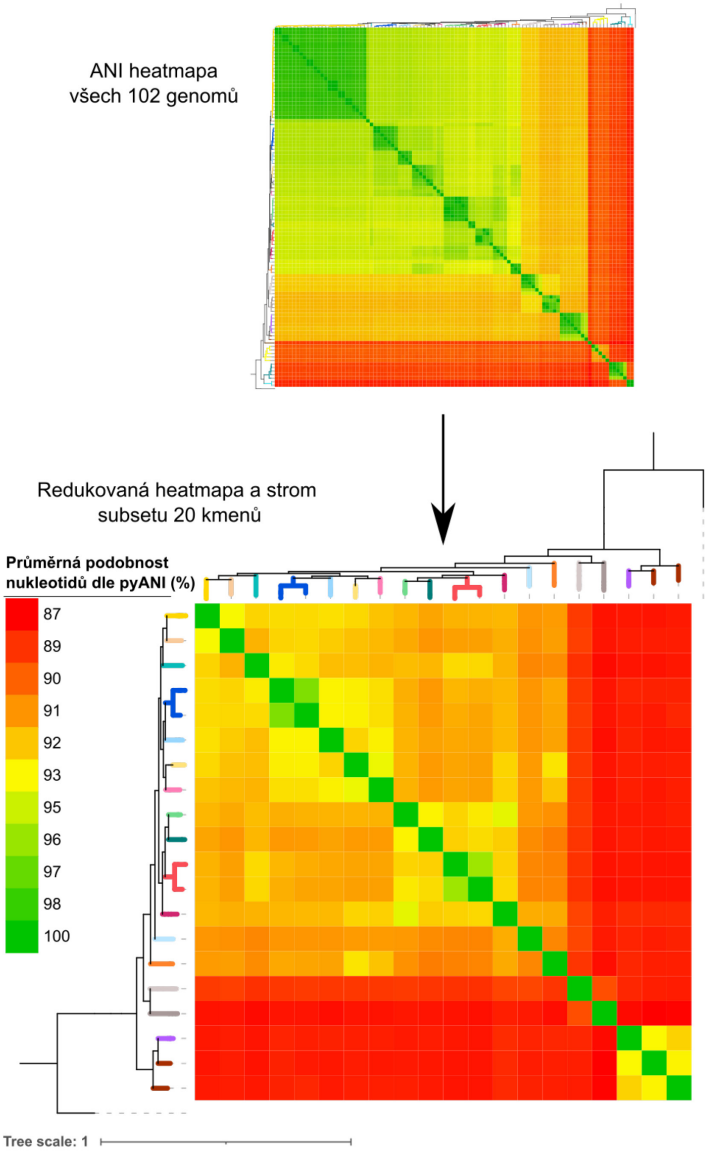
Třetí metoda hledá YPAL pomocí skriptu psaném v Python (skript je součástí přílohy 10.4). Algoritmus nejprve vyhledá všechny 5‘-CATGGATG-3‘ v genomu. Poté zkontroluje, zda se v okolí motivu nachází inverzní repetice – zda oblast -2 až -50 obsahuje kontinuální komplement k oblasti +32 až +82 (v reverse komplementu). Komplementaritu hodnotí Needleman–Wunsch algoritmus (Needleman & Wunsch, 1970), který rozezná jak indel tak změnu bází (sníží přitom hodnocení alignmentu). Musí být přítomen minimálně 20 bází dlouhý komplementární úsek. Výsledný počet YPAL nalezených jednotlivými přístupy je v tabulce níže.

Postup identifikace YPAL	**YPAL nalezeno v subsetu dvaceti kmenů**
Metoda 1. Geneious	2768
Metoda 2. R palindrome finder (0 chyb)	1536
Metoda 2. R palindrome finder (max 3 chyby)	2620
Metoda 3. Python YPAL skript	2460

Tabulka 3. Počet YPAL identifikovaných pomocí jednotlivých postupů.

4.7. Subset dvaceti kmenů *Stenotrophomonas*

Z důvodu snížení výpočetních nároků byl u některých analýz použit subset dvaceti kmenů, jehož strom a ANI heatmapa jsou na ilustraci 24. Strom byl vytvořen redukcí původního fylogramu všech kmenů, jakým způsobem byly kmeny z rodu vybrány, je znázorněno v příloze 10.1. Fylogram subsetu vznikl redukcí fylogenetického stromu všech *Stenotrophomonas*.



Ilustrace 24. Fylogenetický strom dvaceti kmenů subsetu (spodní část). Vytvořen redukcí originálního stromu všech 102 kmenů Stenotrophomonas (horní část). Ke stromu je přidána heatmapa průměrné identity nukleotidů (ANI), připravena pomocí fastANI (Jain et al., 2018). Vizualizováno v ITOL (Letunic & Bork, n.d., 2021), upraveno v Inkscape.

Z kompletních genomů subsetu byl připraven MAUVE alignment. Z něj byly extrahovány jednotlivé kolineární bloky (tedy oblasti, které MAUVE určil jako podobné mezi kmeny) a ty byly konkatenovány do jednoho alignmentu. V tomto alignmentu byly poté manuálně identifikovány REP a YPAL lokusy a dále analyzovány, jak je popsáno v podkapitolách 6.6 a 6.9.4, respektive.

4.8. Seznam využitého softwaru

Název	Popis	Reference
AlphaFold v2.3.2	Google Colab verze AlpaFold pro predikci proteinových struktur	(Jumper et al., 2021)
Biopython 1.79	Knihovna mnoha nástrojů pro bioinformatiku a výpočetní biologii	(Cock et al., 2009)
Biostrings 3.15	Skupina nástrojů specificky zaměřená na analýzu biologických řetězců – DNA, RNA apod.	(Pagès H. et al., n.d.)
circlize 0.4.14	R balíček zaměřený na reprezentaci dat v kruhových diagramech	(Gu et al., 2014)
CSI Phylogeny 1.4	Webový nástroj pro tvorbu fylogenetických stromů analýzu celých genomů	(Kaas et al., 2014)
fastANI 1.34	Rychlý algoritmus odhadující homologii pomocí K-merů unikátních úseků DNA	(*Fastani :: Anaconda.Org*, n.d.; Jain et al., 2018)
Geneious Prime (2021-2023.1)	Prostředí pro *in silico* analýzu DNA. Vhodný ke skládání contigů i široké škále genetických analýz.	(Geneious 2023.0.4 | Bioinformatics Software for Sequence Data Analysis, n.d.)
ggplot2 3.4.0	Sada nástrojů pro vizualizaci široké škály dat	(Gómez-Rubio, 2017)
Google Colab	Cloud-based Jupyter notebook běžící v prostředí Ubuntu	(*Google Colab*, n.d.)
ChimeraX 1.6.1	Program pro vizualizaci molekulárních struktur	(Pettersen et al., 2020)
Inkscape 1.3	Open source grafický editor zaměřený na práci s vektory	(*Draw Freely | Inkscape*, n.d.)
ISfinder	Databáze inserčních sekvencí	(Siguier et al., 2006)
IslandViewer 4	Algoritmus pro *de novo* identifikaci mobilní DNA v genomech	(Bertelli et al., 2017)
ITOL v5	Online prostředí pro práci s fylogenetickými stromy	(Letunic & Bork, 2021)
Jalview 2.11	Platforma pro práci s alignmenty mnoha sekvencí	(Waterhouse et al., 2009)
Mauve 2.4.0	Software zaměřený na alignment mnoha kompletních genomů současně	(Darling et al., 2004)
MEGA-X 10.2	Příprava alignmentů mnoha sekvencí a příprava fylogramů	(Tamura et al., 2021)
Palindrome analyser 2.6.6	Online nástroj pro *de novo* identifikaci palindromů v DNA sekvenci	(Brázda et al., 2016)
pandas 1.5	Populární python knihovna pro práci s velkými datasety	(Mckinney, 2010)
RStudio (2021.09-2023.03)	Vývojové prostředí jazyka R s velkým množstvím nástrojů a knihoven vhodných pro práci s daty.	(RStudio | The Open-Source IDE from Posit, n.d.)
Tidyverse 1.3.2	Kolekce R balíčků pro Data science	(Wickham et al., 2019)
WebLogo3 3.7.11	Webová platforma pro tvorbu WebLogo vizualizace sekvencí	(WebLogo 3 - Create, n.d.)

Tabulka 4. Seznam softwaru, který byl v práci využit

5. Výsledky

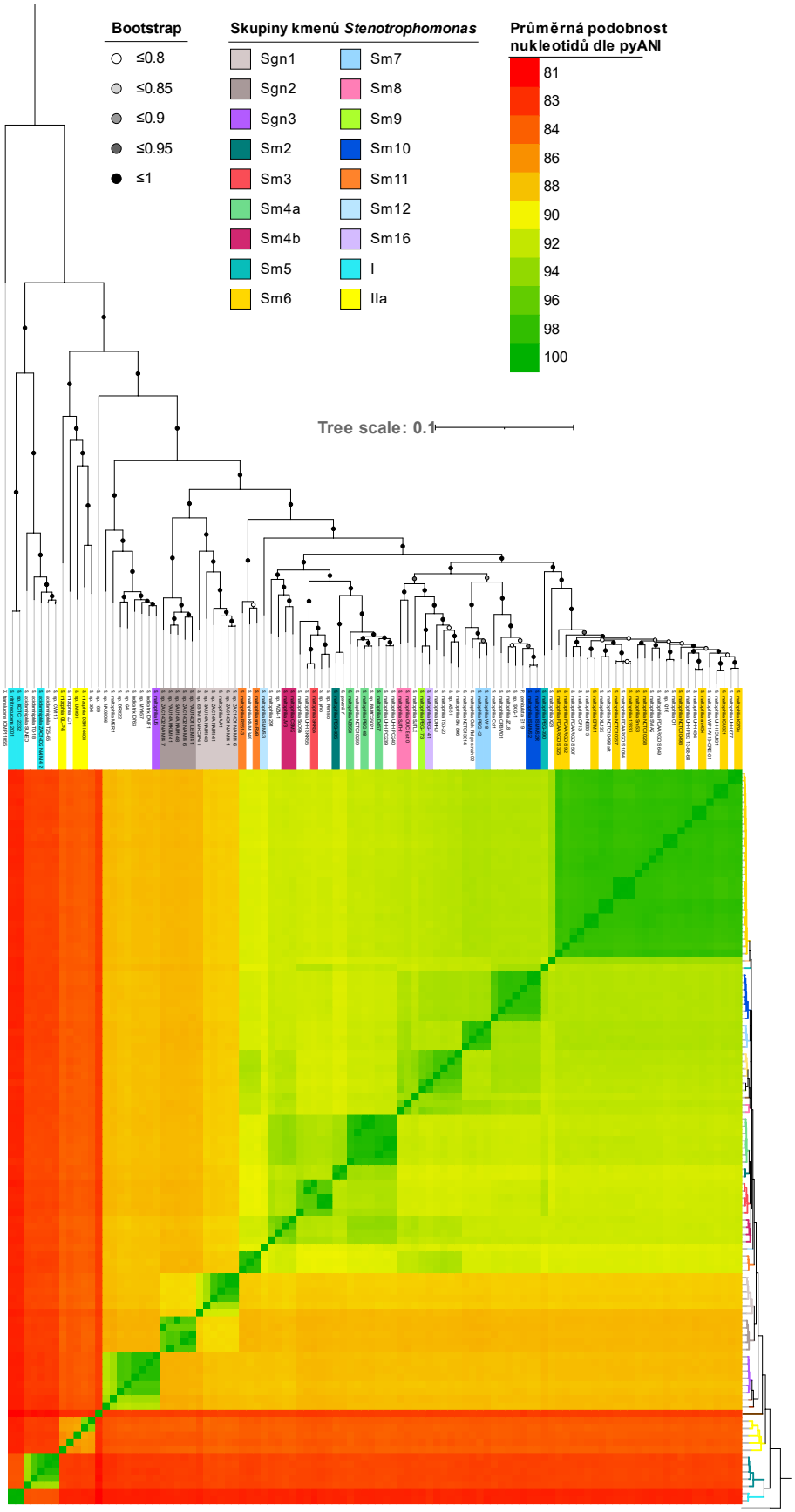
5.1. Fylogenetická analýza bakterií rodu *Stenotrophomonas*

Rod *Stenotrophomonas* byl vybrán pro naši analýzu díky velkému počtu sekvenovaných genomů, významnému posunu znalostí fylogenetiky v posledních letech a přítomnosti velkého počtu REP (Roscetto et al., 2008) a s nimi asociovaných proteinů RAYT (Nunvar et al., 2010). Prvním krokem práce bylo prozkoumat fylogenetické vztahy kmenů s plně sekvenovanými genomy. Proto byl připraven fylogenetický strom na základě celogenomových sekvencí *Stenotrophomonas* (a *Pseudomonas* fylogeneticky náležících do *Stenotrophomonas*). Postup rozepsán viz. Metody 5.1.

Výsledný strom je reprezentován na ilustraci 25. Větve fylogramu mají vprostřed délky tečku, která značí spolehlivosti daného větvení (tzv. hodnotu bootstrap), téměř všechny se blíží 100% spolehlivosti. Barevně jsou zvýrazněny kmeny patřící do skupin charakterizovaných v dřívějších studiích. Proto můžeme zatím nezařazené kmeny přiřadit k těmto skupinám bez potřeby definovat nové skupiny kmenů. Pouze několik velice odvozených kmenů (které nemají žádné příbuzné genomy sekvenovány) bylo zařazeno do nových, vlastních skupin – *S. sp.* 169, *S. m.* X28 a *S. m*. SJTL3.

Na ilustraci níže jsou vyznačeny jen skupiny definované v současné literatuře. Některé skupiny byly ale dále rozděleny, toto dělení je využíváno po zbytek práce (například ilustrace 27). Bazální skupina **I** je tvořena dvěma odvozenými větvemi. Skupina **Ia** je tvořena klonální dvojicí *S. sp.* KCTC 12332 a *S. nitritireducens* 2001. **Ib** pak obsahuje kmeny S. *acidaminiphila* a *S. sp.* CW117. Strom také obsahuje několik klonálních dvojic či dokonce trojic, například: (*S. m.* 454 - *S. m.* UHH454), (S. m. 13637 - *S. m.* FDAARGOS 1044 - *S. m.* NCTC10257) a (*S. m.* ISMMS2 - *S. m.* ISMMS2R).

Korunní skupinou (nejodvozenější monofyletická skupina) rodu je Sm6, ta obsahuje mimo jiné reprezentativní (ale nikoliv typový) kmen K279a a vlastní typový kmen 13637 (Davenport et al., 2014). Pouze kmeny Sm6 jsou dnes považované za druh *Stenotrophomonas maltophilia sensu stricto*. S těmito daty naše analýza souhlasí, kmeny Sm6 jsou fylogeneticky jasně odvozené od ostatních kmenů a zároveň vzájemně příbuzné. Sestavení fylogenetického stromu hostitelského rodu byl nezbytný krok pro analýzu evoluce RAYT proteinů, REP elementů a jejich vztahu s hostitelskými genomy. Všechny další analýzy na těchto poznatcích staví.



Ilustrace 25. Fylogenetický strom rodu Stenotrophomonas. Připraven v CSI Phylogeny a MEGA-X (nastavení viz metody 5.1). Na fylogramu jsou zvýrazněné kmeny, patřící do již definovaných podskupin (Gröschel et al., 2020; Mercier-Darty et al., 2020; Vinuesa et al., 2018). Ke stromu je přidána heatmapa průměrné podobnosti nukleotidů (ANI), ta byla připravena pomocí fastANI (Jain et al., 2018). Vizualizováno v ITOL a Inkscape.

5.2. Fylogenetická analýza RAYT

RAYT nejsou v automatické anotaci genomů identifikovány, typicky jsou značeny jen jako transpozázy/hypotetické proteiny. Pro jejich analýzu bylo nutné *de novo* identifikovat všechny RAYT. Nejprve byly nalezeny potenciální RAYT pomocí BLAST již známého proteinu, u těch byla zkontrolována přítomnost definujících aminokyselin. Nalezené proteiny byly rozděleny do skupin podle svého genového okolí a REP elementů, které je obklopují. Detailní postup je rozepsán v metodách 5.2.

Rayt geny jsou součástí konzervovaných oblastí genomu, celkem 331 nalezených *rayt* bylo rozděleno do 11 lokusů. Dle sekvenční analýzy je 60 RAYT pseudogenizováno, nejčastější je frameshift, který vzniká indel mutací a následným posunutím čtecího rámce. U několika zástupců došlo k nonsense mutaci a vzniku STOP kodonu uvnitř původně kódující oblasti. U tří genů poté došlo k deleci části sekvence, v lokusu tak zbývá pouze část genu obklopená poškozenými REP elementy.

Celkové množství potenciálně funkčních *rayt* u rodu *Stenotrophomonas* je tak 271. REP identifikované v jejich intergenovém okolí jsou považované za asociované s daným *rayt*. Pomocí této asociace je definován RAYT/REP systém. V okolí *rayt* je často více druhů REP (a také degenerované REP). Nicméně vždy byla určena jedna sekvence REP, která je zdaleka nejčastější. Tato asociace pak byla opakovaně potvrzena v dalších genomech. Platí pozitivní korelace, kdy genom s funkčním *rayt* má vždy více REP s ním asociovaných (viz dále, ilustrace 27). *Rayt* je pojmenován podle lokusu, ve kterém se nachází (tedy podle svého genového okolí), REP jsou pojmenovány podle *rayt*, se kterým jsou asociovány. Zároveň, číslování RAYT odpovídá jejich četnosti. RAYT 01 je nejpočetnější skupina (89 zástupců), RAYT 11 je nejvzácnější (jediný zástupce).

Téměř všechny genomy obsahují alespoň jeden *rayt*, přičemž průměr je přes 3 *rayt* per genom. Nikdy nejsou v genomu přítomny dva *rayt* asociované s jednou REP sekvencí. Zároveň nebyly nalezeny duplikace *rayt*. Sekvence každého *rayt* je v daném genomu odlišná.

5.2.1. RAYT lokus a jeho okolí

Rayt geny se vyskytují ve stabilním genovém okolí. Tyto oblasti jsou schematicky znázorněny v příloze 10.2. Každá skupina (RAYT 01 – RAYT 11) je tvořena *rayt*, které jsou ve stejném konkrétním genomovém sousedství (stejné okolní geny – syntenie). Délka intergenových oblastí sousedících s *rayt* v daném lokusu se může značně lišit. Tyto oblasti jsou zaplněny REP elementy. REP jsou zde často ve dvojicích (REPIN) či velkých organizovaných klustrech (BIME). Důvodem, proč dělit RAYT do skupin podle jejich genomového okolí, je vypovídající hodnota pro rekonstrukci evoluce RAYT/REP systému. Podle nejjednoduššího (a tedy nejpravděpodobnějšího) modelu dochází k počátečnímu „obsazení“ daného lokusu *rayt* genem s REP elementy a evoluční perzistenci tohoto stavu, případně sekundárním ztrátám. *Rayt* jsou součástí core genomu (skupiny genů konzervovaných ve většině kmenů či celém druhu).

Nebylo zjištěno, že by některá funkční kategorie genů byla obohacena mezi geny sousedícími s různými RAYT lokusy. Nejvíce RAYT lokusů má obsazen kmen ZAC14A_NAIMI4_1. V tomto kmeni je osm funkčních *rayt*, každý ve vlastním RAYT lokusu a asociovaný s konkrétní sekvencí REP. RAYT/REP systémy zcela chybí u bazální skupiny Ib.

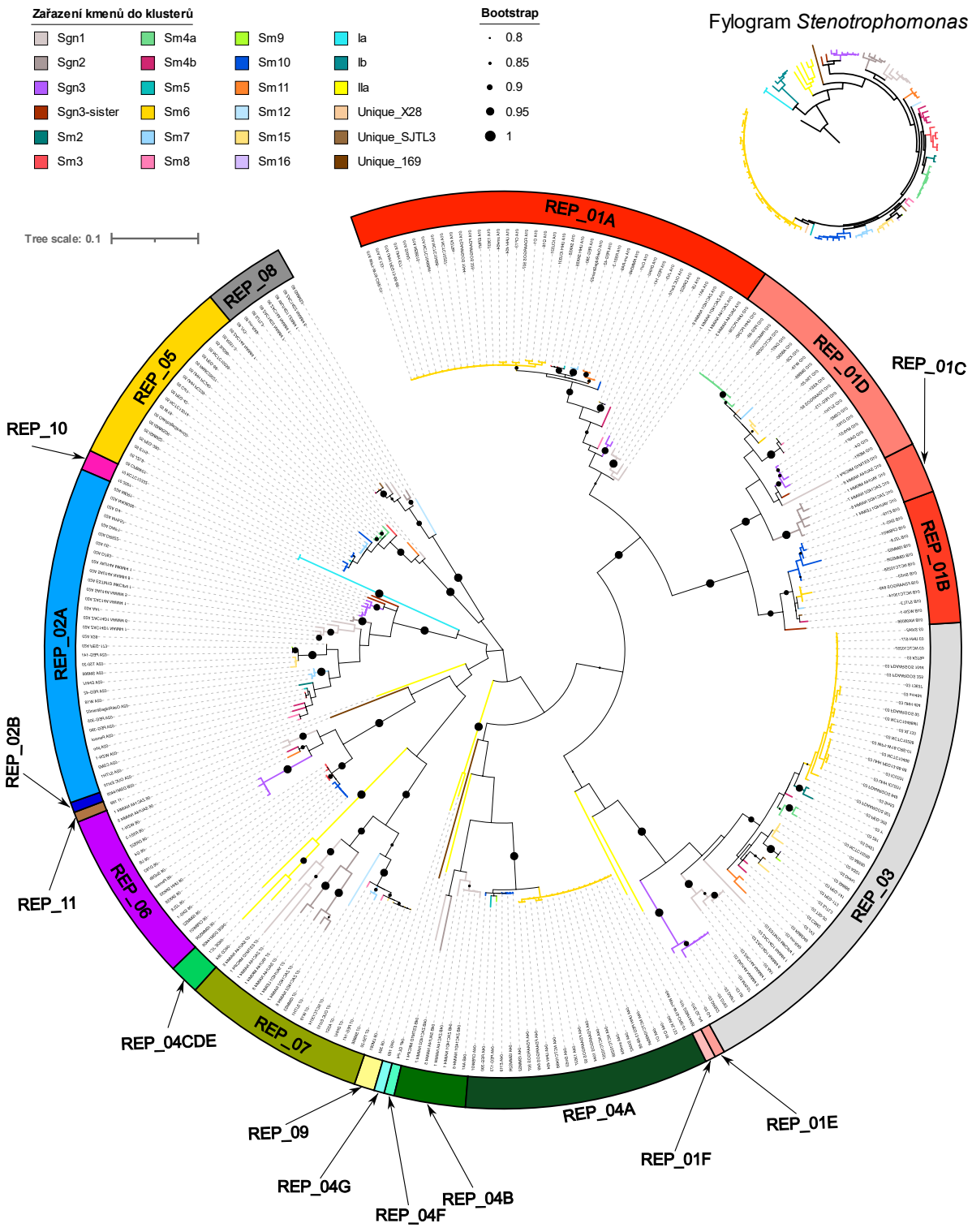
V případech lokusů 01, 02 a 04 byla zjištěna přítomnost různých sekvenčních typů REP elementů v daném lokusu u různých bakterií (kategorizace REP viz dále). Například u RAYT 02 a 04 jde pravděpodobně o běžnou vertikální evoluci spojenou s diverzifikací, původní RAYT v lokusu divergoval natolik, že se ustanovila asociace s jinou sekvencí REP. V těchto případech byly RAYT dále subkategorizovány podle asociovaných REP sekvencí (lokus 01: A-F, lokus 02: A-B, lokus 04: A-G). Situace je jiná u RAYT lokusu 01, zde evidentně dochází k záměnám. Nejde o postupnou evoluci, ale o výměnu celého systému REP/*rayt*. Mechanismus této rekombinace je specifický (dochází k němu pouze u RAYT skupiny 01), nejvíce se vzájemně nahrazují RAYT 01A a 01D. Více v podkapitole 6.2.2.1.

5.2.2. Koevoluce RAYT a REP

Po identifikaci RAYT vyskytujících se u *Stenotrophomonas* a fylogenetické analýze rodu, byla provedena rekonstrukce evoluce samotných RAYT („naslepo“ - pouze na základě samotné sekvence RAYT). Byl připraven alignment všech RAYT (bez pseudogenů) a na jeho základě vytvořen strom (algoritmus maximum likelihood) s bootstrap 10 000 v programu MEGA-X (Tamura et al., 2021).

Na ilustraci 26 je výsledný nezakořeněný strom. Vnější kruh značí, s jakým typem REP elementů je daný RAYT asociován. Je zřejmá vysoká evoluční stálost asociace RAYT s příslušnými REP elementy. Z analýzy nelze vyvodit, zda na změnu sekvence REP postupně reaguje RAYT, či naopak změna sekvence RAYT postupně ovlivní REP.

Strom RAYT lze porovnat se stromem samotných kmenů. Kongruencí, resp. inkongruencí lze odhalit způsob, jakým se v evoluci přenášejí (vertikální vs. horizontální). Příkladem je evoluce RAYT 01 a RAYT 03, které jsou (jak ukazuje fylogram) sekvencemi AK vzdálené ostatním RAYT a každá skupina je fylogeneticky koherentní. Zatímco lokus 01 je obsazený systémem RAYT/REP napříč diverzitou *Stenotrophomonas* (kromě nejbazálnějších skupin), k obsazení lokusu 03 došlo později v evoluci (je prázdný u skupiny IIa – *S. rhizophila*) (viz ilustrace 27). Systém RAYT/REP z lokusu 1 je velmi diverzifikovaný a v rámci této diverzity jak RAYT 01F, tak REP 01F jsou velmi blízce příbuzné RAYT 3, resp. REP 03. RAYT/REP systém z lokusu 03 je tedy paralogního původu z ancestrálního systému lokusu 01. Pro rekonstrukci původu RAYT/REP v ostatních lokusech fylogram bohužel neposkytuje dostatečně jasné indície.

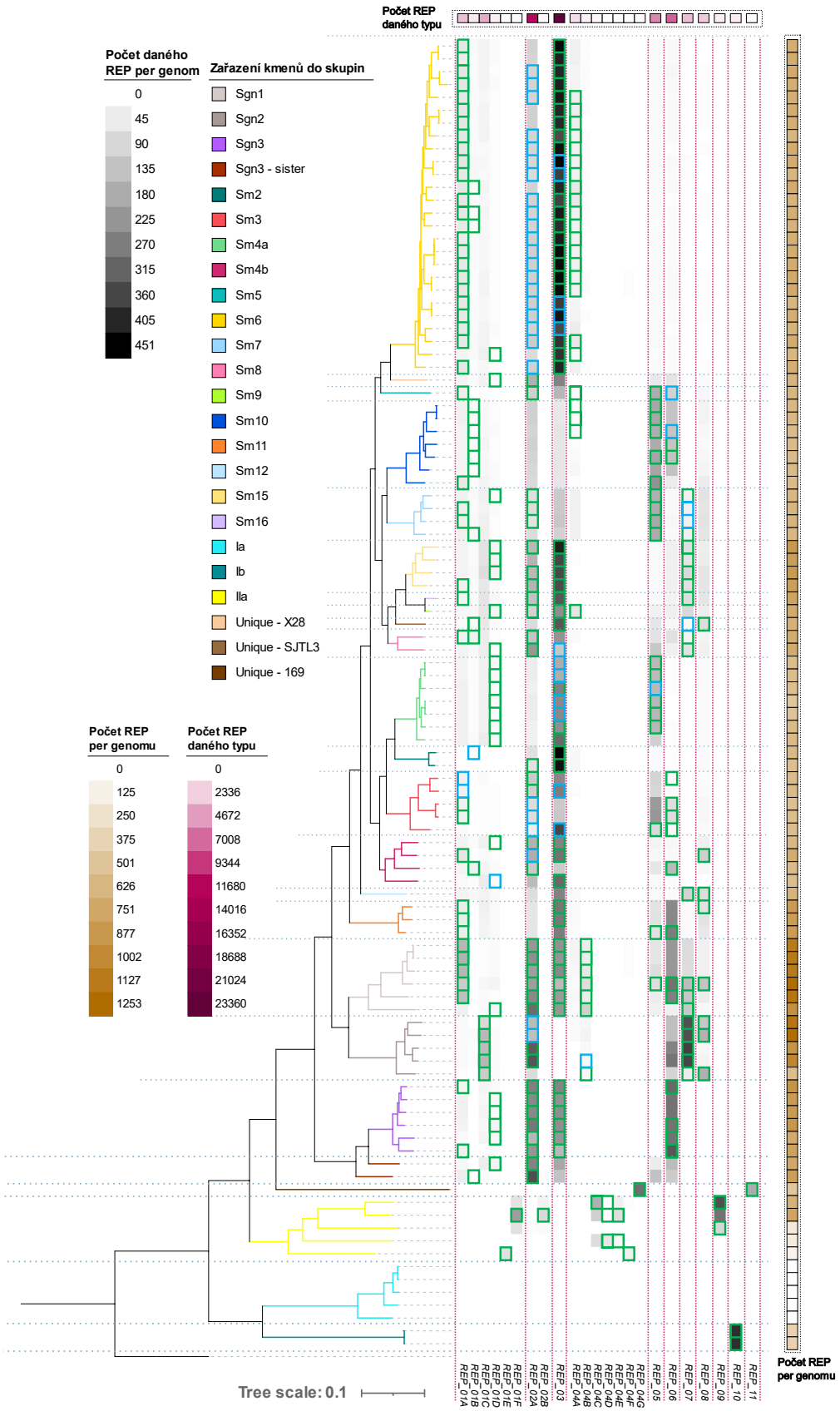


Ilustrace 26. Fylogenetická analýza všech 271 funkčních RAYT nalezených u Stenotrophomonas. Strom vytvořen v MEGA, vizualizován v ITOL. Větve stromu jsou zvýrazněny podle skupin. RAYT jsou označeny podle toho, se kterými REP jsou asociovány.

Dále je patrná pozice RAYT patřících ke kmenům skupiny Sm6 (RAYT 01A, 03 a 04A). Jejich pozice je vždy v rámci parciálního fylogramu dané skupiny RAYT nejodvozenější, což koresponduje s pozicí Sm6 jako korunní skupiny stenotrofomonád. Je to důkazem, že tyto RAYT se v hostitelských bakteriích vyvíjely vertikálním přenosem od společného předka. Jde o důkaz dlouhodobé koevoluce RAYT/REP systémů s hostitelskými kmeny a jejich minimální frekvenci horizontálního přenosu.

Ilustrace níže znázorňuje heatmapu REP elementů v jednotlivých genomech. Ty jsou seřazeny dle fylogramu (zobrazeno nalevo). Nad heatmapu je přidán samostatný graf celkového počtů REP daného typu (odstíny červené). Další graf je poté vpravo od heatmapy, ten ukazuje celkové počty REP v jednotlivých kmenech (odstíny hnědé). Zelené orámování mají políčka je-li v kmenu přítomen funkční RAYT asociovaný s daným REP. Modré jsou poté rámečky kmenů, které mají daný RAYT pseudogenizovaný.

Heatmapa ukazuje, že počet REP daného typu je typicky vyšší v genomu obsahujícím asociovaný RAYT. To samé platí i pro pseudogenní RAYT, ale v menší míře. Heatmapa také ukazuje rozdělení RAYT/REP v rámci bakterií stejné skupiny a mezi skupinami. Není pravidlem, že by sesterské skupiny nutně sdílely stejné RAYT. Naopak, distribuce RAYT je často diskontinuální. Řada RAYT se vyskytuje pouze v jediné skupině. Tyto tendence značí jednak opakované ztráty RAYT v rámci celých skupin a dále evoluci nových RAYT, které se posléze vertikálně rozšířily v rámci diverzifikace dceřinných linií.

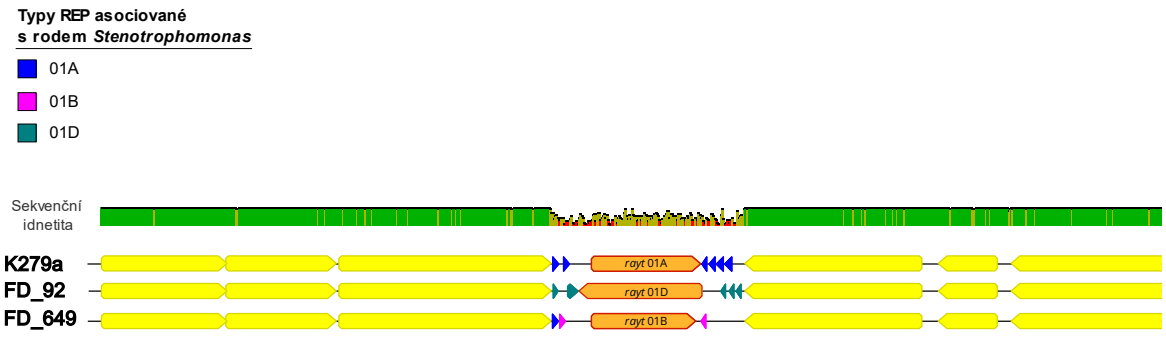


Ilustrace 27. Heatmapa REP elementů rozdělená na sekce po ose Y podle skupin kmenů. Po ose X pak podle druhu REP, každý má unikátní sekvenci, kterou sdílí všechny REP bez odchylky. Odstín bodu značí množství REP elementů v genomu. Zelené ohraničení znamená, že genomu obsahuje funkční RAYT daného typu. Modré ohraničení pak znamená přítomnost pseudogenizovaného RAYT.

5.2.2.1. RAYT/REP 01

Anomální evoluci vykazuje skupina RAYT/REP 01. Krom zcela bazální skupiny Ia-b se vyskytuje ve všech fylogenetických linií v rámci diverzity *Stenotrophomonas*. Z ilustrace 27 je zřejmé, jak v rámci skupiny dochází k „přeskokům“ (potenciální horizontální přenos). Důležité je, že asociace příbuzných RAYT a příslušných REP se nemění (viz fylogram 26), variace tedy zahrnuje celou funkční jednotku RAYT/REP. Nejpočetnější je typ 01A. V rámci skupin u řady kmenů došlo ke změně z RAYT/REP 01A na typy 01B až 01E. Například skupina Sgn2 nemá žádný RAYT 01A, ale pouze 01B. Příkladem aberantní evoluce lokusu je korunní skupina Sm6. Tato skupina je tvořena blízce příbuznými kmeny, většina z nich má RAYT 01A, ale několik kmenů má místo toho 01B a 01D. REP 01A a B jsou jediné REP, kde je GTRG tetranukleotid tvořen GTGG (ilustrace 34).

Tři zástupce Sm06 s rozdílným subtypem RAYT/REP lokusu 1 ukazuje ilustrace 28. Že jde o blízce příbuzné kmeny, je patrné z vysoké sekvenční identity genů v okolí RAYT, které jsou téměř identické. Sekvence *rayt* a jeho blízké okolí, kde jsou REP elementy, je podstatně méně konzervované. Tento předěl je navíc velice ostrý a začíná téměř okamžitě za koncem kódujících oblastí. Anomální evoluci lokusu 01 lze obtížně vysvětlit, jako nejpravděpodobnější mechanismus se jeví opakovaný horizontální přenos, pravděpodobně omezený na daný lokus.



Ilustrace 28. Alignment lokusu RAYT 01 a jeho okolí ze tří kmenů Sm06. Sekvence byly alignovány pomocí Clustal algoritmu s nejvyšším nastavením. Nad geny je přidán graf sekvenční identity a legenda s názvy genů. Zaznamenány jsou REP s maximem tří mismatch. Pro zvýraznění REP není použito barevné schéma využívané ve zbytku práce kvůli snazšímu odlišení. Připraveno v Geneious, vizualizace upravena v Inkscape.

5.2.2.2. RAYT/REP 02

Krom bazálních skupin se RAYT/REP 02 vyskytuje ve většině fylogenetických linií. U linie IIa je přítomen jediný zástupce RAYT 02B. Ten je obklopen REP 02B, lišící se od 02A záměnou báze v rameni palindromu. Po REP 03 jsou REP 02A nejrozšířenějším druhem REP. Vyskytují se běžně jak jednotlivě (31,1%), v REPIN (32,3%) tak součástí amplifikovaných BIME složených z více druhů REP (36,6%). Celkem bylo nalezeno 32 929 kopií (max 3 chyby).

RAYT 02 jsou nejpočetnější skupinou (63 kmenů), z toho 29 zástupců je ale pseudogenizováno a jsou pravděpodobně nefunkční (či musí při translaci dojít k frameshiftu). Pseudogenizace proběhla zřejmě u předka kmenů Sm6, a proto jsou zde všechny RAYT 02 pseudogenizovány (či ztraceny). Dále proběhla nezávisle u skupin Sm03 a Sgn2. I v genomech s pseudogenizovaným či ztraceným RAYT je nicméně stále přítomna velká populace REP 02A (viz dále, kap 6.4).

5.2.2.3. RAYT/REP 03

REP 03 jsou extrémně rozšířené v hostitelských genomech, to je zřejmé z ilustrace 27, kde v grafu celkového počtu elementů REP 03 dominují (graf v odstínech vínové nad heatmapou). Jde o jediné REP vyskytující se vždy ve stovkách kopií na genom. Jde o tak početné elementy, že byla pro vizualizace zvolena šedá barva, aby nezakryly všechny ostatní typy REP. Tento systém se vyvinul z RAYT/REP 01 (kapitola 6.2.2 a ilustrace 26). RAYT 03 je nejrozšířenější skupinou funkčních RAYT. Má 62 zástupců z nichž pouze 12 jsou pseudogeny.

5.2.2.4. RAYT/REP 05 až 08

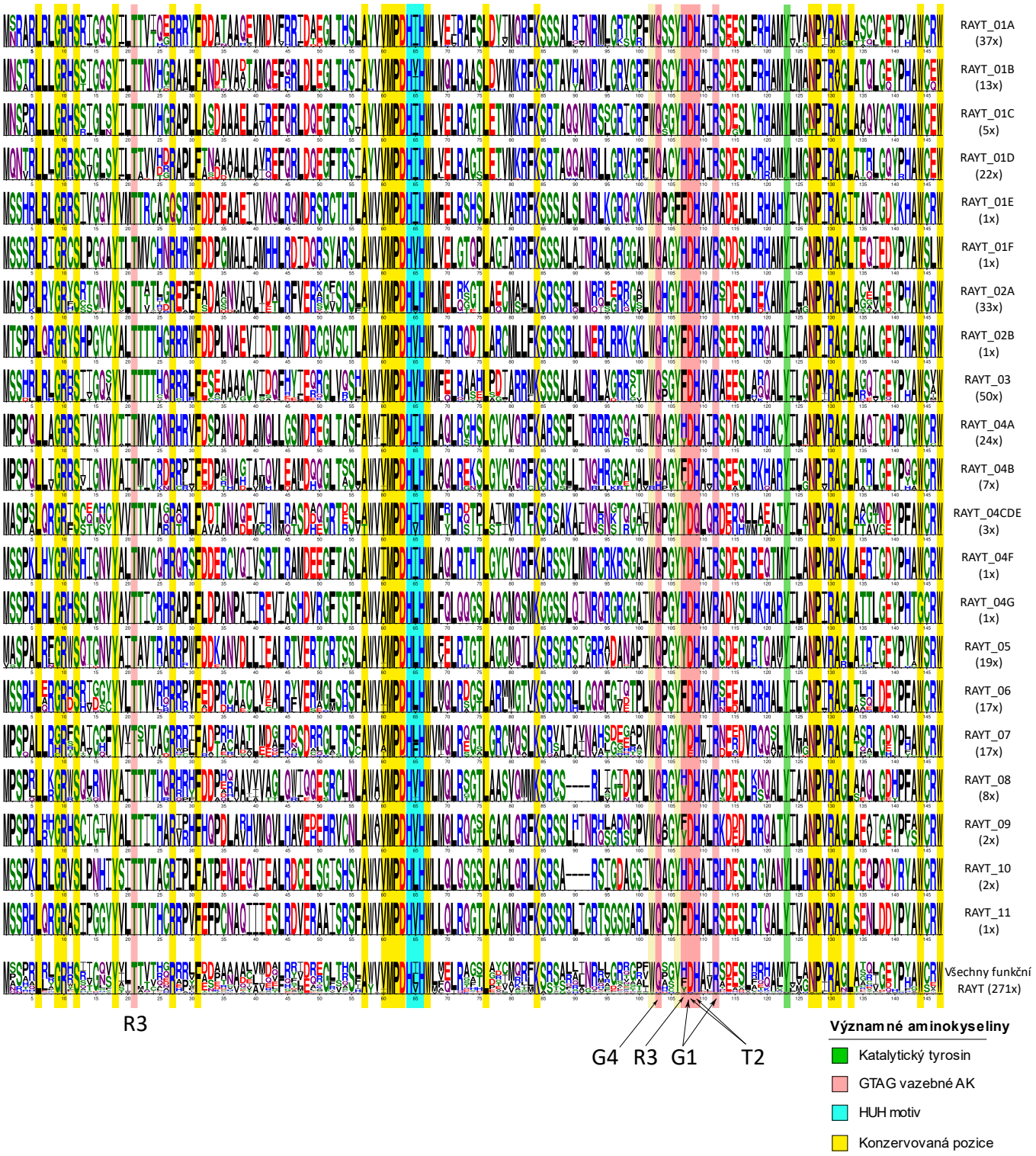
RAYT/REP 05 až 08 jsou méně rozšířené, typicky jen ve dvou či třech skupinách. Jejich evoluce není zcela jasná. V dávné evoluční historii mohlo dojít k jejich horizontální kolonizaci předků různých skupin, což pak vedlo k vertikálnímu rozšíření v celé skupině. Nerovnoměrné obsazení lokusů v rámci skupin pravděpodobně ukazuje na opakované ztráty. Zvlášť REP 05, 06 a 08 jsou často v relativně vysokých počtech přítomny ve skupinách, kde se s nimi asociovaný RAYT vůbec nevyskytuje (viz ilustrace 32).

5.2.2.5. RAYT/REP 09 až 11

Jedná se o systémy unikátní pro jediný kmen (či několik blízce příbuzných) z velmi odvozených linií. Hostitelské kmeny mají vždy značnou populaci asociovaných REP. Tyto REP se nevyskytují v žádných jiných kmenech, k obsazení RAYT lokusu pravděpodobně došlo během divergence od zbytku rodu.

5.3. Struktura RAYT rodu *Stenotrophomonas*

RAYT jsou rozdělené do skupin podle lokusů ve kterých leží, a podle asociovaných REP. Pro každou skupinu byl vytvořen alignment AK sekvencí (ilustrace 29). Pomocí nich byly porovnány sekvence všech RAYT a nalezeny konzervované AK sdíleny všemi skupinami RAYT. Z každé skupiny byl vybrán jeden reprezentativní zástupce, pro který byla vytvořena predikce 3D struktury. Predikce byly modelovány s pomocí AlphaFold2.0 a poté zkoumány v programu ChimeraX. Na ilustraci 30 jsou predikce několik takových proteinů. Výsledné modely byly dále porovnány s dodnes jediným experimentální ověřeným RAYT z *E. coli* K12, jehož struktura byla určena pomocí rentgenové krystalografie (Messing et al., 2012).

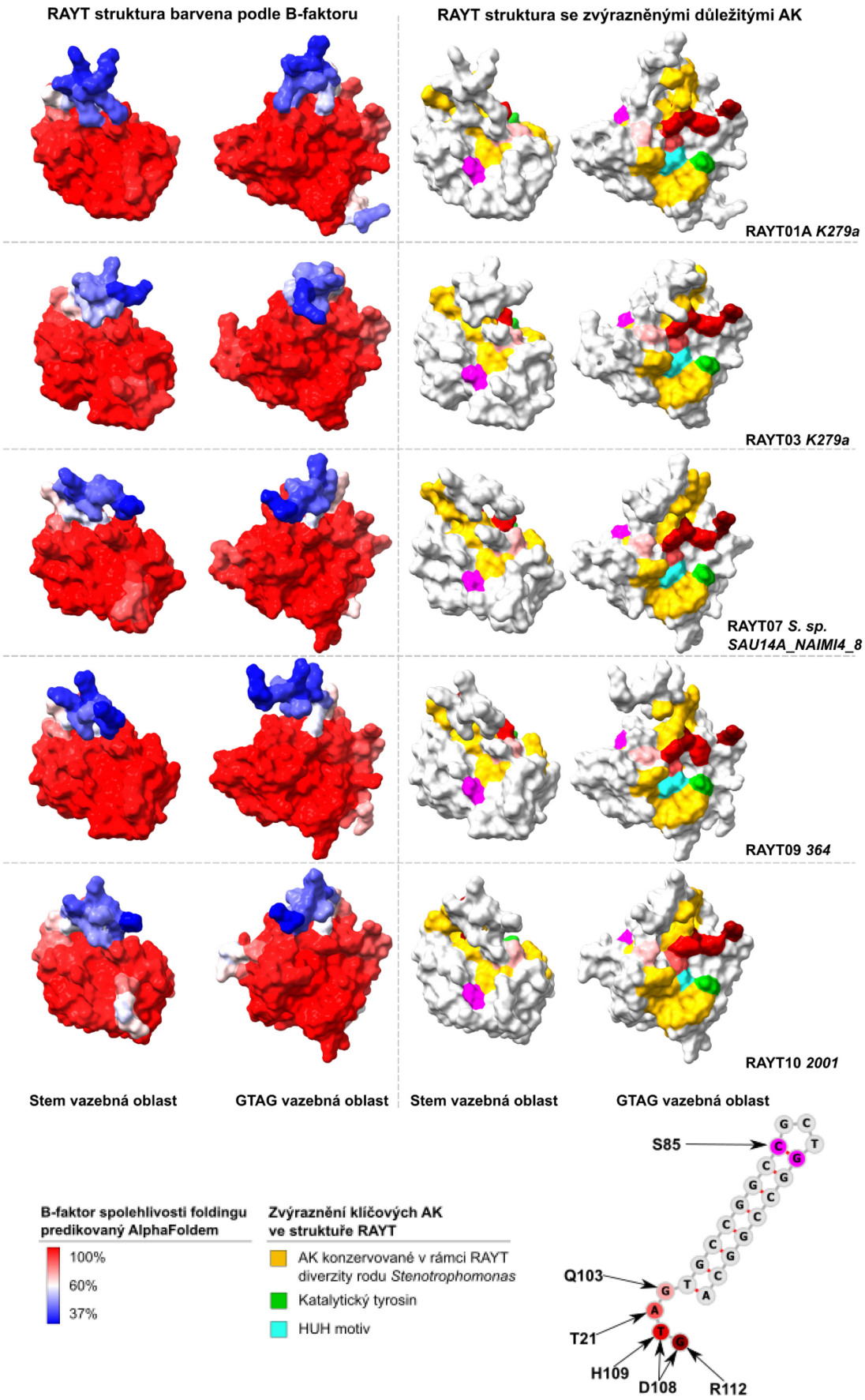


Ilustrace 29. Webloga všech skupin RAYT nacházejících se u Stenotrophomonas. Analýzou alignmentů sekvencí byly identifikovány konzervované aminokyseliny, HUH motivu (červená), katalytický tyrosin (světle zelená), AK vázající GTRG tetranukleotid (modrá) a další (žlutá). Alignment připraven v MEGA (metoda MUSCLE). RAYT 08 a 10 jsou o několik aminokyselin kratší, alignment byl proto ručně upraven. Zároveň je u několika zástupců zkrácena C terminální doména kvůli šířce ilustrace. Připraveno pomocí WebLogo3, upraveno v Inkscape.

U RAYT z *E. coli* jsou známy AK zodpovědné za interakci s REP (Messing et al., 2012). REP *Stenotrophomonas* mají ale odlišnou sekvenci i délku, tudíž nelze všechny výsledky zobecnit. Nicméně oba RAYT interagují s GTRG tetranukleotidy. AK klíčové pro tuto interakci jsou u *Stenotrophomonas* konzervovány a jsou zvýrazněny na ilustraci 30 (pravá polovina, červená). Také byla prozkoumána konzervovanost v sekvenci REP z pohledu interakce se sekvencí RAYT, tedy zda některé báze na konkrétních pozicích v REP nejsou vždy asociovány s konkrétními AK v RAYT. Byl identifikován serin na pozici 85, který je vždy přítomen, pokud je na 8. pozici od GTAG umístěn C (alternativně k C8 komplementární G). Asociovaná oblast **S85** a nukleotid **C8** jsou na ilustraci 30 zvýrazněny fialovou.

Byla tedy připravena predikce struktury RAYT a predikce části vazebných mezi RAYT a REP (ilustrace 30), poslední překážkou pro docking těchto dvou molekul je struktura REP. Což je problém, který se v rámci této práce nepodařilo překonat. Z modelu RAYT *E. coli* lze extrahovat strukturu REP, jenže tento REP má zcela jinou sekvenci i délku než REP *Stenotrophomonas*. Lze snadno připravit strukturu DNA helixu, ale REP nejsou dsDNA vláknem. Modeling s DNA helixem jasně identifikuje předpokládanou vazebnou oblast, ale vazebné interakce nejsou věrohodné. Nebyl nalezen software schopný věrohodně modelovat strukturu ssDNA a ss/dsDNA molekul. Nejbližší alternativou jsou programy na predikci struktur ssRNA a jejich následná validace a převedení na ssDNA. Tyto vícekrokové metody trpí na řadu artefaktů ve výsledném modelu. Prvním krokem je predikce sekundární struktury ssRNA. Dle ní je vytvořena predikce 3D modelu ssRNA a ta je nakonec převedena do DNA. Výsledný model není věrohodný, ssRNA má strukturu vlásenky v A-formě, zatímco ssDNA v B-formě (Chen & Fong, 1969; Eichhorn & Al-Hashimi, 2014). Dále, vazba DNA na RAYT pravděpodobně ovlivňuje strukturu DNA. Většina dokovacího softwaru ale pracuje s rigidní molekulou DNA a pouze částečnou flexibilitou proteinu. Všechny zkoumané programy určily stejnou vazebnou oblast, překrývající se s tou námi predikovanou. Žádný nebyl schopen věrohodně predikovat tvar DNA při navázání do tohoto místa. Největšího pokroku bylo dosaženo s využitím serveru HADDOCK (Van Zundert et al., 2016), který je schopen dockingu proteinu s flexibilní sekvencí DNA, ale ani tyto struktury nejsou dostatečně kvalitní.

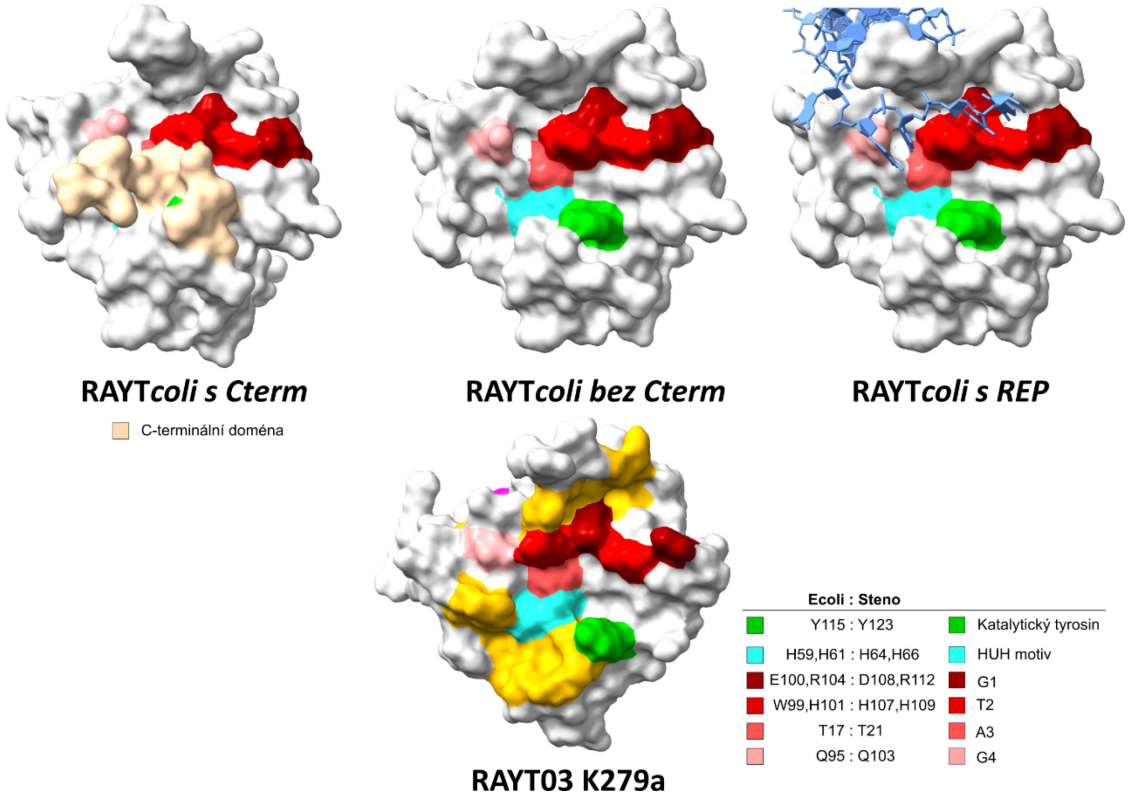
Kvůli výše zmíněným důvodům bohužel nebyl připraven věrohodný model RAYT s navázaným REP. Přesto byla provedena extensivní analýza možných vazebných oblastí a určena vazba konzervovaných nukleotidů REP (díky homologii příslušných GTAG-vazebných aminokyselin RAYT *E. coli*), které jsou na ideálních místech predikovaného proteinu (Ilustrace 30). Okolí katalytického místa proteinu je tvořeno aminokyselinami invariantními v rámci diverzity RAYT u *Stenotrophomonas*, což značí, že okolí aktivního místa je stabilní v rámci diverzity RAYT, a tedy že tyto proteiny jsou katalyticky aktivní.



Ilustrace 30. Predikce RAYT struktur vytvořené v AlphaFold2.0. Vizualizováno v ChimeraX. Pohled je nastaven, aby bylo vidět maximum konzervovaných AK a je vždy stejný. První dva sloupce ukazují pravděpodobnost správné predikce sbalení. Třetí a čtvrtý poté jsou úplně stejné dva pohledy, pouze jsou barveny dle klíčových pozic. Většina proteinu má vysoké hodnocení přesnosti modelu (téměř 100%), přesnost klesá na N a C terminálních doménách, jejichž reálný folding se může podstatně lišit.

AlphaFold2 hodnotil predikce struktury RAYT jako přesné (AlphaFold2 hodnotí většinu proteinu s více jak 99% spolehlivostí). Velký pokles spolehlivosti mají N a C konce. Nízká přesnost predikce N terminu je problematická. Jeho pozice přímo interferuje s vazbou REP elementu, zároveň se zdá, že ji alespoň do nějaké míry může usměrňovat. Tvarem N-terminus připomíná rameno jeřábu, a právě pod tímto ramenem by měl procházet konec ramene REP dále navazující na GTAG tetranukleotid.

Na podobné bázi pracují dimery IS*200* transpozáz, N-terminus zde má funkci podobnou rameni jeřábu (Hyung et al., 2006). Váže ssDNA vlákno a je schopen s ním manipulovat. Hypotézu, že to samé může platit pro RAYT podporuje, že řada AK v N-terminus je v rámci diverzity RAYT konzervovaná. Ilustrace 31 porovnává struktury RAYT *Stenotrophomonas* a *E. coli*. Struktura z *E. coli* byla získána experimentálně (Messing et al., 2012), data stažena ze SWISS-MODEL repositáře (PDB: 4er8) (Bienert et al., 2017). Červenou jsou zvýrazněny AK podílející se na vazbě GTAG, které RAYT *Stenotrophomonas* sdílí s RAYT *E. coli*. Je evidentní jejich prostorová konzervovanost.



Ilustrace 31. Porovnání modelu RAYT E. coli získaného experimentálně a predikce RAYT K279a vytvořené v AlphaFold. První struktura E. coli je kompletní včetně C-terminální domény, ta je u ostatních odstraněna. Struktury byly v ChimeraX pomocí Matchmaker přes sebe překryty (na základě strukturní podobnosti), díky tomu jsou zobrazeny z identického úhlu pohledu. Obarveny jsou konzervované pozice (zlatá) a další důležité AK (viz legenda). Dále jsou v legendě porovnány klíčové pozice u E. coli proti jejich protějškům ze Stenotrophomonas.

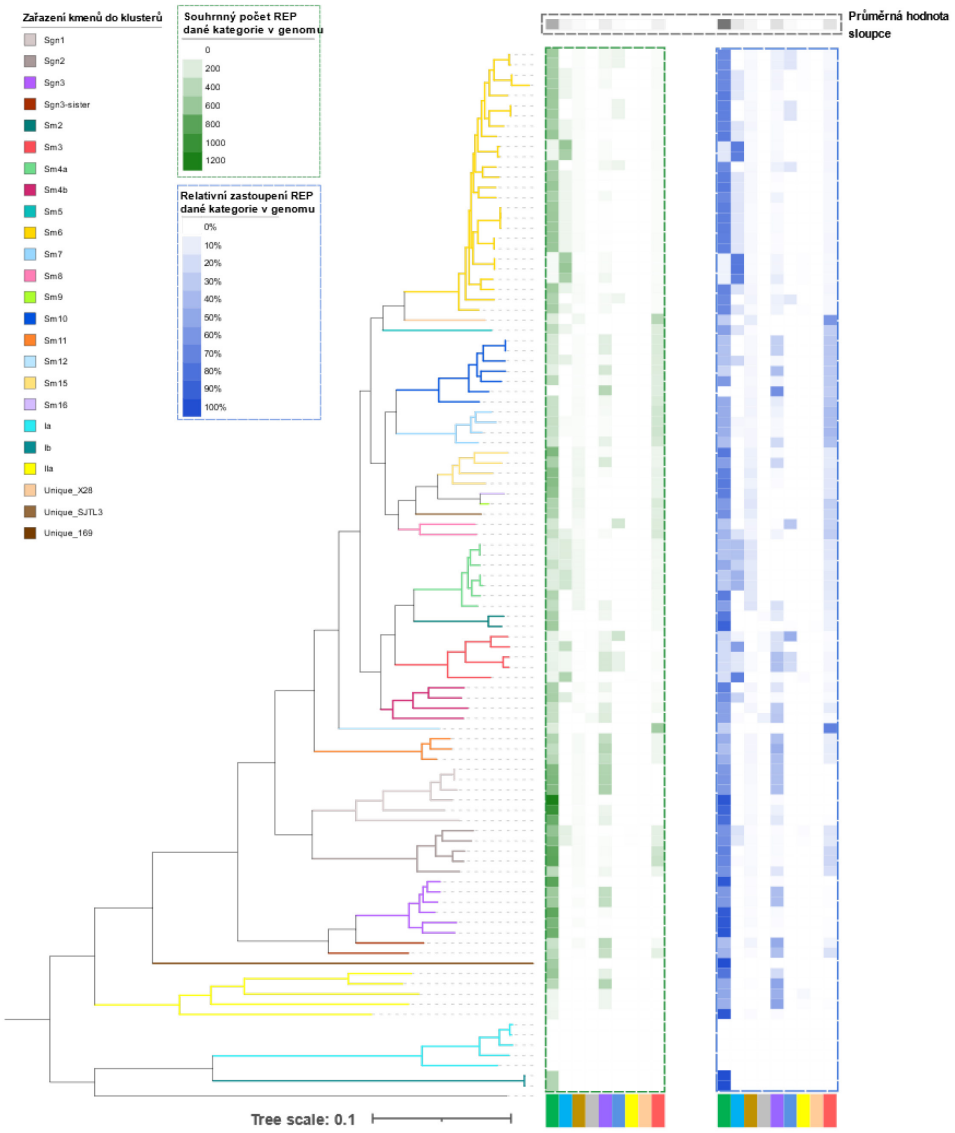
První struktura RAYT *E. coli* je neupravená, ostatní mají odstraněnou C-terminální doménu (C-term je zvýrazněna). Její funkcí je zakrýt katalytickou oblast proteinu a tím regulovat jeho funkci. Po jejím odstranění je při porovnání s RAYT *Stenotrophomonas* zřejmé, že katalytická oblast obou proteinů je prakticky identická. Zároveň je kompletně obklopena zcela konzervovanými aminokyselinami (zlatá barva). Poslední *E. coli* RAYT je zobrazen s navázaným REP elementem (světle modrá), konkrétně jsou vidět nukleotidy GTAG vázané na konzervované AK. Pozice těchto AK je u *Stenotrophomonas* prakticky identická, s velkou pravděpodobností je tak vazba GTRG u obou druhů stejná.

5.4. Asociace RAYT a REP

RAYT u rodu *Stenotrophomonas* jsou úzce asociovány s REP. Sekvenci asociovaného REP lze snadno odhalit, neboť jsou umístěny v intergenovém okolí *rayt* genu – ve více kopiích a u všech lokusů tohoto RAYT. RAYT protein poté tyto REP elementy šíří genomem.

Tento vztah jsme se pokusili kvantifikovat. Ilustrace 32 obsahuje dvě heatmapy počtu REP per genom, zelenou s absolutními hodnotami a modrou s relativními. Na ose x je rozlišené, jaký mají REP vztah k přítomnosti/absenci příslušného RAYT (daného lokusu) v genomu. Možné vztahy RAYT/REP jsou znázorněny na schématu pod ilustrací. REP v genomu, který obsahuje RAYT s těmito REP asociovaný, patří do zelené kategorie. Tyto REP jsou více rozšířené než všechny ostatní kategorie dohromady (jak lze vidět na heatmapách).

Fialová kategorie znamená, že asociovaný RAYT v genomu chybí, ale má ho příbuzný kmen (tzn. ze stejné skupiny). Tato kategorie typicky vypovídá o ztrátách RAYT v rámci skupiny. Vzhledem k tomu, že se jedná o druhou nejčetnější kategorii, ztráta RAYT pravděpodobně nevede ke ztrátě početnosti příslušných REP elementů v hostitelském genomu.



Ilustrace 32. Kombinace dvou heatmap REP elementů. První mapa (zelená) znázorňuje absolutní počty REP jednotlivých kategorií, existuje jen jedna nejvyšší hodnota mezi všemi genomy. Druhá (modrá) mapa pak tato data reprezentuje procentuální zastoupení každé kategorie per genom, součet polí v řádku je tak 100%. REP jsou do kategorií rozděleny dle přítomnosti asociovaného RAYT v genomu (rozepsáno pod heatmapou). Data z Geneious prime, připraveno v ITOL, upraveno v Inkscape.

Alternativní pohled na tato data nabízí ilustrace 33. Jednotlivé body značí počet REP elementů v genomu. Barevné schéma bodů je stejné jako v předchozí ilustraci. Zelené jsou počty REP, kdy genom obsahuje asociovaný RAYT, červené pak počty REP v genomech bez jakéhokoliv RAYT (v genomu i celé skupině). Graf ukazuje, že trendy v globální četnostech REP různých asociačních kategorií, tak jak jsou patrné na ilustraci 32, platí i pro jednotlivé skupiny REP, a vyjadřují tak obecný fenomén. Graf ukazuje řádově nejvyšší počty zelené kategorie REP (příslušný RAYT přítomen), následované fialovou kategorií (RAYT ztracen během diverzifikace bakterií dané skupiny). Na opačném konci spektra četnosti REP elementů je červená kategorie.

Ilustrace 33. Každý bod značí počet REP daného lokusu v jednom genomu. Každý řádek obsahuje 102 bodů (102 studovaných kmenů) a každý bod je zbarvený podle vztahu jaký mají dané REP s RAYT lokusem (v rámci kmene). Osa X je počet REP, na ose Y pak jsou jednotlivé kategorie lokusů. Jde o alternativní pohled na data předchozí ilustrace, barvy se drží stejného schématu. Ilustrace byla vytvořena v R s knihovnou ggplot2.

5.5. REP elementy rodu *Stenotrophomonas*

REP elementy v rámci *Stenotrophomonas* jsou extrémně početné, diverzifikované do mnoha skupin a v rámci těchto skupin sekvenčně konzervované. Pro analýzu byly REP rozděleny do skupin dle jejich asociace s RAYT (viz kap 5.2). V rámci 102 *Stenotrophomonas* bylo nalezeno 63 689 kopií REP shodných se sekvencemi asociovanými s RAYT a 132 039 kopií nesoucích 1-3 mutace (tab. 5).

	Počet	Hustota
100% identické REP	63 689	1 REP na 6 878 bp
REP (až 3 chyby) *	132 039	1 REP na 3 560 bp

*Tabulka 5. Celkové počty REP nalezených ve všech kmenech a jejich průměrná hustota. * tyto REP byly nalezeny blast-like-annotation funkcí Geneious prime, ta krom záměn nukleotidů rozeznává také REP s indel mutací. Jelikož na funkci REP nemá indel větší vliv než záměna báze (nedochází k posunu čtecího rámce) jsou obě změny v principu rovnocenné.*

5.5.1. Struktura REP

REP rodu *Stenotrophomonas* jsou dlouhé 22 až 25 nukleotidů a jsou vysoce konzervované. Pro analýzu konzervovanosti byly nejprve extrahovány všechny sekvence REP z genomu, navíc prodloužené o čtyři nukleotidy z obou stran. K tomu bylo vybráno vždy několik kmenů s nejvyšším počtem kopií daného REP, zároveň šlo o co nejvíce fylogeneticky vzdálené kmeny (Ilustrace 34).

Ilustrace 34. Webloga definující konzervovanost sekvencí všech typů REP, které jsou asociovány s RAYT. Každý typ REP byl prozkoumán u několika (evolučně vzdálených) kmenů. Poté byl vybrán jeden kmen jako reprezentativní pro celý typ REP. Typicky šlo o kmen s nejvíce REP daného typu. Zkratka za weblogem značí_druh REP_název reprezentativního kmene. n značí počet REP v genomu. Zlatě jsou zvýrazněny okolní nukleotidy, které jsou na pozici u více něž poloviny REP.

Tím byl maximálně redukován vliv, že v některém genomu mohou REP vykazovat anomální sekvenci a zkreslit výsledky. Pro každý druh REP bylo připraveno několik nezávislých grafů WebLogo (*WebLogo 3 - Create*, n.d.), zda nebude odhalena konzervovanost REP a jeho okolí specifická pouze pro některé kmeny.

Sekvence konzervovaného tetranukleotidu na začátku REP je téměř vždy GTAG (výjimkou jsou REP 01A, B). Několik typů REP má nepárující první nukleotid po GTAG. Navíc je u mnoha (ale ne všech) REP elementů konzervován před GTAG ještě jeden nukleotid, nejčastěji G (tedy sekvence **GGTAG**), ale i jiné báze. V několika případech jsou konzervovány i další nukleotidy mimo vlastní sekvenci REP. Důvod konzervace těchto přídatných bází není jasný, mohou však mít například funkční úlohu v rozpoznávání elementu pomocí RAYT.

Délka ramene (zelené barva) se pohybuje mezi 7 až 9 nukleotidy a rameno nikdy není přerušeno nepárujícími nukleotidy, na rozdíl od mnoha REP elementů z jiných bakterií (Gilson et al., 1984; Tobes & Ramos, 2005). V sekvenci ramene výrazně převažují GC páry, u některých skupin AT zcela chybí (01E-F, 03, 04E, 06). Díky tomu mají REP u *Stenotrophomonas* pravděpodobně velmi stabilní sekundární struktury. Variabilní smyčka je dva až čtyři nukleotidy dlouhá. Její sekvence není mezi jednotlivými třídami REP konzervovaná.

5.5.2. Genomová lokalizace REP elementů

Po analýzách pracujících pouze s absolutními počty REP v genomu jsme se zaměřili na přesné umístění REP. Pro REP platí, že se vyskytují mimo geny, tedy v intergenových oblastech (tabulka 6). REP v rámci čtecího rámce je nejpravděpodobněji chybou anotačního softwaru, příslušné genové produkty jsou bez výjimky anotovány jako „hypothetical protein“. Část REP také může svým GTAG sloužit jako STOP kodon okolních genů.

Pozice REP	**Počet REP (maximálně 3 chyby)**	**Relativní počet REP**
Intergenová	121 898	92,3%
GTAG jako STOP kodon	8 442	6,4%
Součást čtecího rámce genu	1 699	1,2%

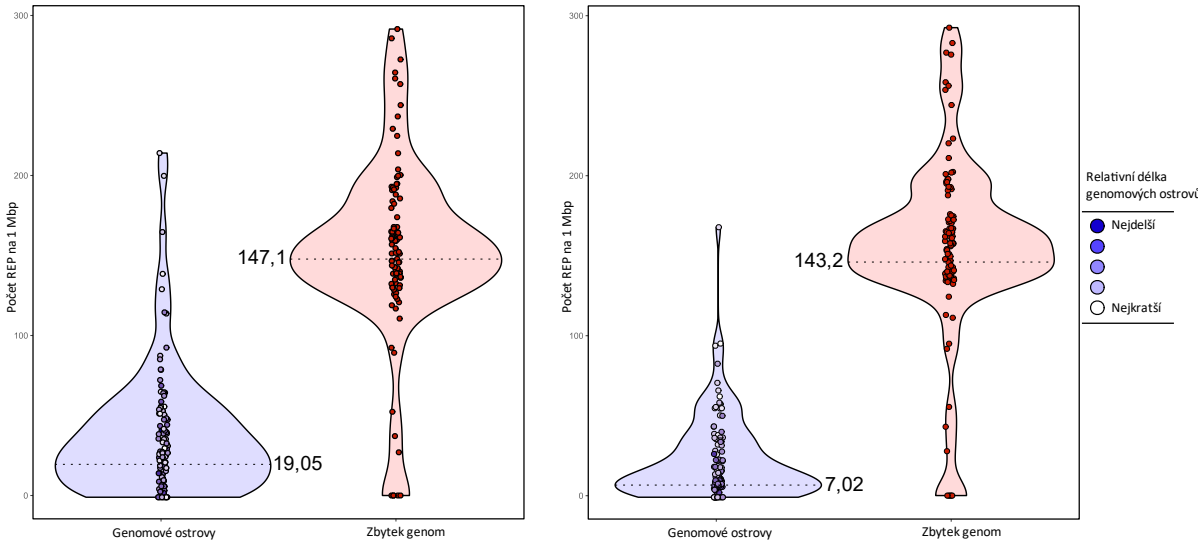
Tabulka 6. Pozice REP vůči okolním genům. Data z Geneious prime.

Pro REP je typické, že chybí na mobilních genetických elementech (dále MGE). To bylo u *Stenotrophomonas* zkoumáno pomocí serveru IslandCompare (Bertelli et al., 2022). Jde o nástroj kombinující několik bioinformatických algoritmů hledajících genomové ostrovy rozdílnými metodami. Jako genomové ostrovy byly určeny ty oblasti, na kterých se shodlo několik nástrojů současně.

Byla spočítána REP hustota v genomových ostrovech a ta pak porovnána s hustotou REP ve zbytku genomu. Data jsou vizualizována jako violin graf (ilustrace 35).

Nečekaným zjištěním je, že IslandCompare považuje některé oblasti RAYT lokusů za genomové ostrovy. RAYT lokusy jsou přitom sdíleny v rámci příbuzných kmenů a jsou považovány za součást stabilního genomu, čemuž pak odpovídají i okolní geny, často jde o house-keeping geny. Byly tak připraveny dvě verze grafu, první s originálními daty (RAYT jsou součást GI) a poté upravená verze, kde RAYT nejsou považovány za součást GI.

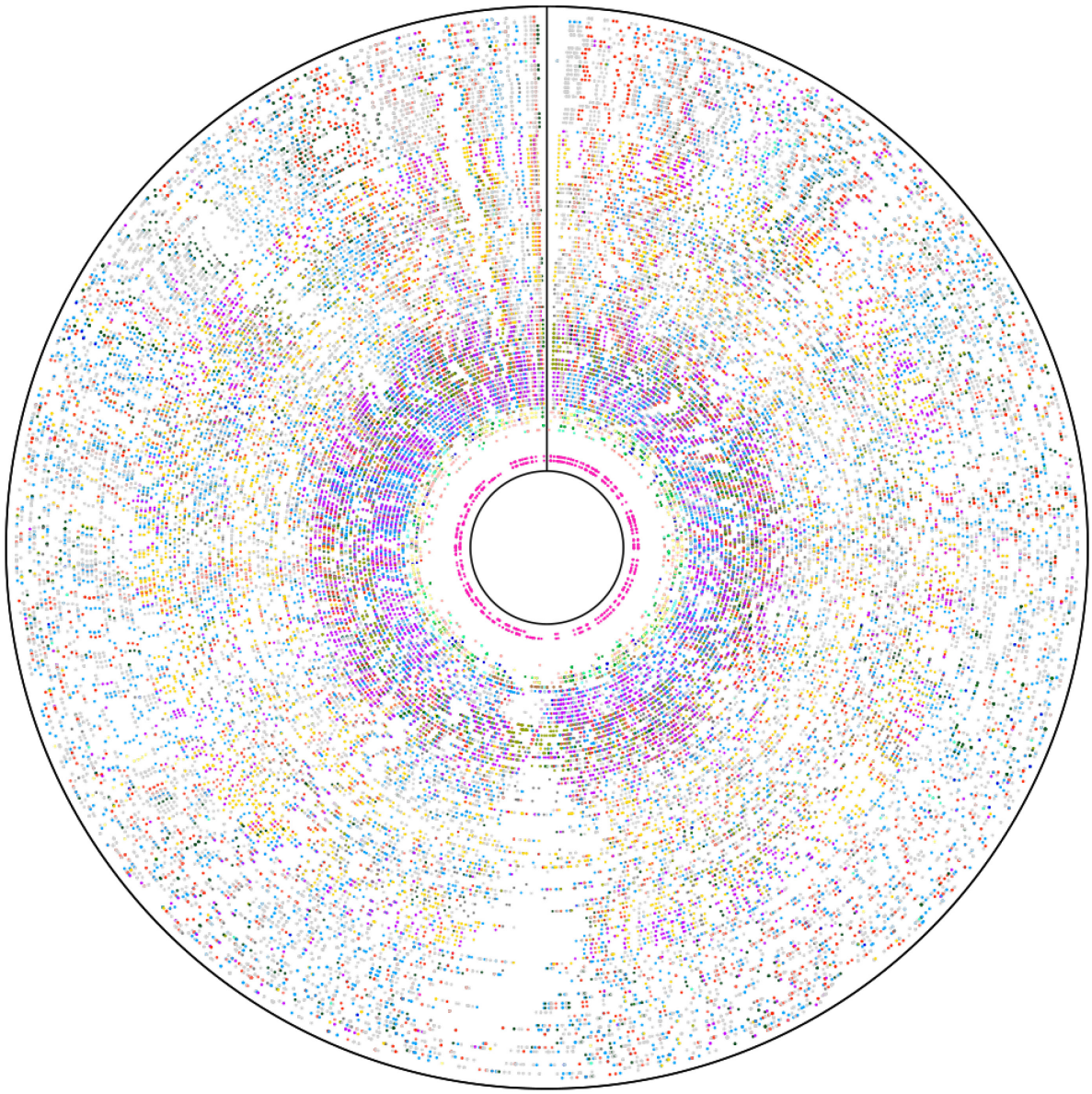
Oba grafy (byť druhý výrazněji) ukazují, že REP elementy jsou v GI podstatně vzácnější. Hlavním důvodem je pravděpodobně to, že expanze REP do nových míst v genomech hostitelských bakterií (intenzita šíření) je nečekaně pomalý proces. REP měly podstatně více času rozšířit se do starších (konzervovanějších) částí genomu než do GI.

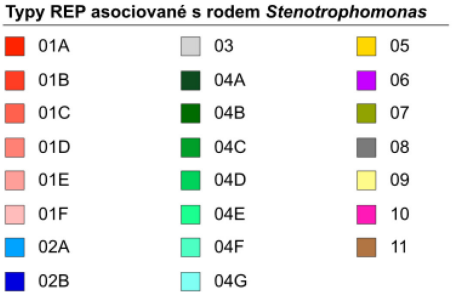


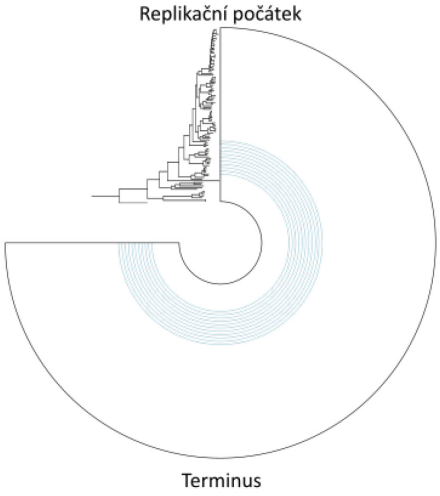
Ilustrace 35. Violin graf hustoty REP elementů v genomových ostrovech a zbytku genomu. Algoritmus zařadil do GI i RAYT lokusy, tato data zobrazuje levá strana ilustrace, na pravé straně jsou RAYT lokusy z GI odstraněny. Body reprezentující hustotu REP u genomových ostrovů jsou zbarveny podle celkové délky GI v genomu. GI nalezeny pomoci IslandCompare. Poté jejich anotace exportovány do Geneious prime. Zde spočítán počet REP každého genomu (v GI a mimo ně). Zároveň zjištěna celková délka GI oblastí. Tato data posloužila pro spočítání hustoty REP elementů na 1 milion bází. Data vizualizována pomocí R a knihoven ggplot2, dplyr.

Dále nás zajímalo, kde se nacházejí REP v rámci celého genomu. Již v 90. letech si vědci zkoumající REP všimli, že výskyt REP se liší v oblastech iniciace a terminace replikace. K tomu došli při zkoumání genomu *E. coli* K12. Měli tak pouze jeden genom a na něm pozorovali asi 500 REP, které definovali pouze jako palindrom přerušený variabilní smyčkou (Bachellier et al., 1999).

V rámci naší analýzy bylo použito přes 100 blízce příbuzných genomů (s prozkoumanými fylogenetickými vazbami) s přesně definovanými REP rozlišenými do řady skupin. Tyto genomy jsou na ilustraci 36 uspořádány jako soustředné kruhy. Kmeny jsou seřazeny dle pozice na fylogramu. Vnitřní kruhy reprezentují kmeny bazální, vnější poté ty odvozené. Jelikož jsou genomy různě dlouhé, kruh vždy reprezentuje procentuální mapu genomu. REP tedy není například 10 000 bází od počátku, ale v 1 % relativní vzdálenost od počátku. Jako počátek je považována pozice DnaA. Tedy v 0 % (na ilustraci vyznačeno jako černá vertikální čára) je DnaA směřující po směru hodinových ručiček. V pozici 50 % leží nukleotid nejvzdálenější od DnaA (dále označováno jako anti-*ori*). V grafu je zachycena pozice více než 68 000 konzervovaných REP elementů.





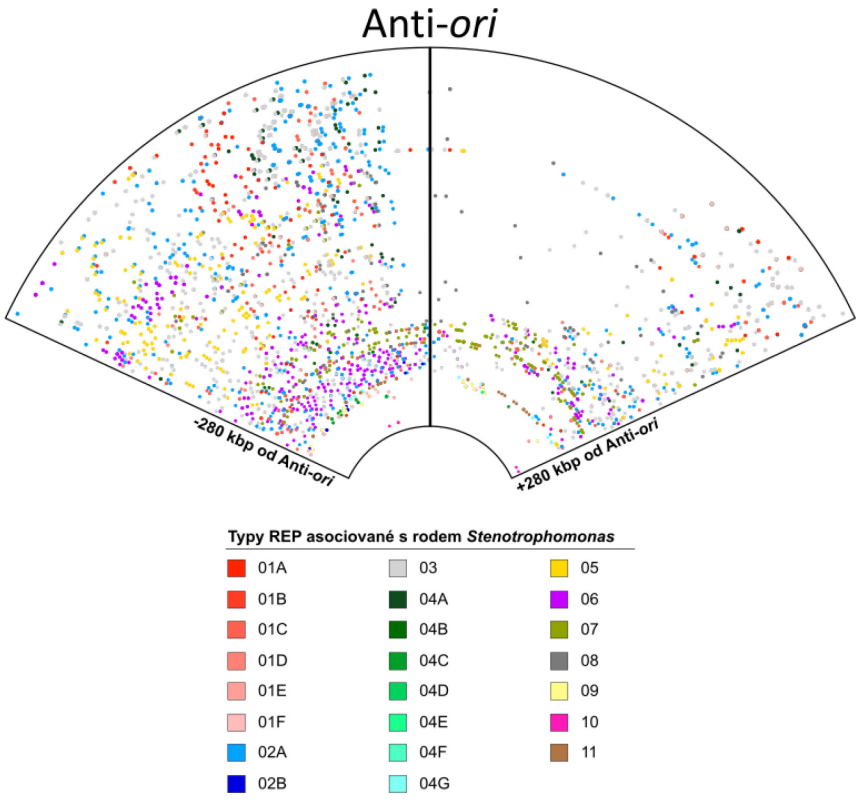


Ilustrace 36. Vizualizace všech REP elementů. Jedná se o koncentrickou reprezentaci všech studovaných genomů. Kruhy jsou seřazeny podle fylogeneze. Vnitřní kruhy reprezentují bazální kmeny, vnější jsou kmeny korunové. REP elementy jsou zbarveny podle barevného klíče, který je součástí legendy (dolní levá část ilustrace). Data o pozici REP, kmeni, typu REP a délce REP byla získána z Geneious prime. V excel dále upravena. V R byl napsán skript pro vytvoření grafu s využitím knihovny circlize.

V grafu jsou patrné REP elementy specifické pro jednotlivé kmeny či skupiny. Například REP 10 (růžová) jsou specifické pouze pro dva bazální kmeny (které jsou navíc klonálně spřízněné). Dále REP 9 (světle/jasně žlutá) jsou přítomné a vysoce abundantní pouze u skupiny **IIa**. Naopak REP 03 jsou tak početné, že bylo nutné zvolit světle šedou barvu, jinak naprosto překryly všechny ostatní typy REP.

Dále jsou patrné čtyři bazální genomy zcela prosty REP elementů (a RAYT), na ilustraci oddělují klonální dvojici **Ia** od zbytku rodu (nejvíce vnitřní kruhy). U té je nápadná identická pozice REP 10, které jsou pro tyto genomy jedinečné. Současně jsou to jediné kmeny, kde se nachází asociovaný RAYT 10.

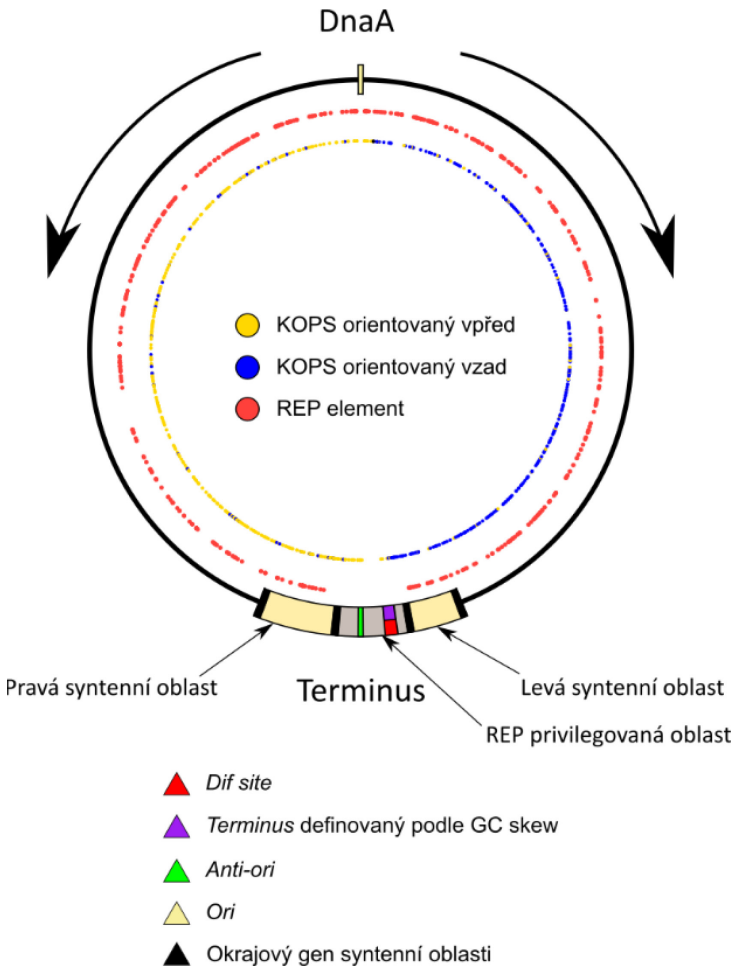
Dále lze z grafu vyčíst REP privilegované oblasti (místa, kde REP chybí). Jednou z menších je například pozice přibližně 350°. Tato oblast je (zvlášť u více odvozených kmenů) zcela bez REP elementů. Po prohledání genomů se ukázalo, že v této oblasti jsou umístěny operony kódující rRNA. Daleko rozsáhlejší je REP privilegovaná oblast v místě terminace replikace (ilustrace 37). Tato oblast přibližně kolokalizuje s místem terminace replikace (*ter*). Důvod k této absenci není znám, navíc oblast terminu ani proces terminace replikace u rodu *Stenotrophomonas* nebyly samostatně zkoumány. Hypotézám, proč je terminus REP privilegovanou oblastí, se bude věnovat následující kapitola.



Ilustrace 37. Výsek z předchozí vizualizace zaměřený na oblast protilehlou k počátku replikace, tzv. Anti-ori. Je zobrazeno 280 kbp na obě strany od Anti-ori, to je zvýrazněno černou přímkou. Připraveno s pomocí Geneious prime, R a knihovny circlize. Upraveno v Inkscape.

5.5.3. Replikační terminus

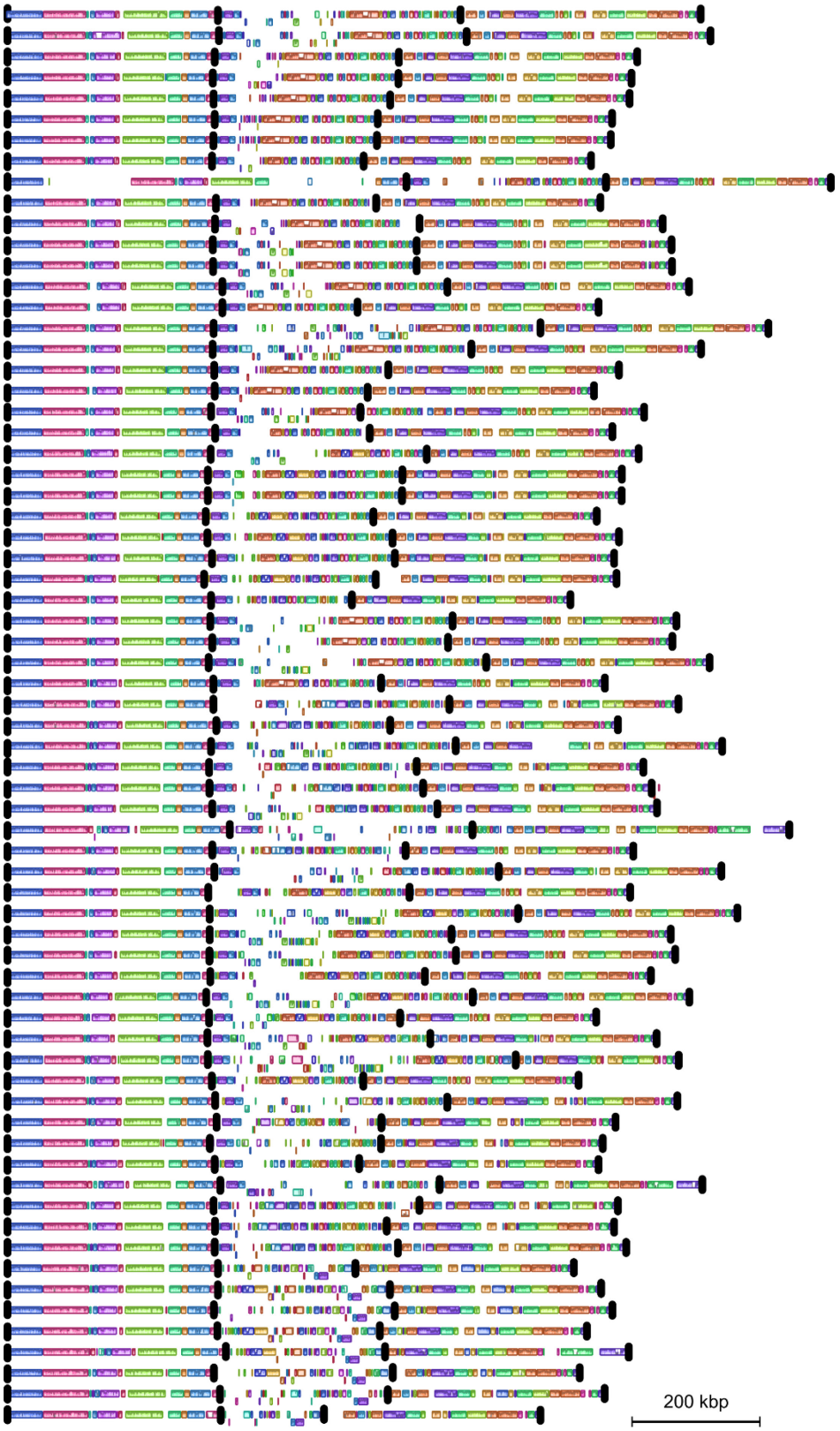
Terminus je oblast, kde dochází k dokončení replikace DNA. Komplex replikačních vidliček se sestavuje v oblasti DnaA. Odtud poté vidličky putují opačným směrem. Za předpokladu, že putují stejnou rychlostí, by se měly setkat v bodě nejvzdálenějším od počátku replikace, ten v rámci této práce nazýváme jako anti-*ori*. Anti-*ori* udává „průměrnou“ pozici v genomu, kde se nejčastěji setkají rovnoměrnou rychlostí cestující replikační vidličky. Skutečný replikační terminus se nemusí v anti-*ori* nacházet, nicméně je rozumné předpokládat jeho pozici v relativní blízkosti. Přibližné schéma struktury genomu se zaměřením na terminus je na ilustraci níže. Jako oblasti syntenie označujeme dlouhé okrajové oblasti terminu (desítky genů), které na rozdíl od střední části mají jednotné pořadí genů (viz níže).



Ilustrace 38. Struktura oblasti terminu. Origin, tedy počátek replikace je umístěn na pozici DnaA. Vnější šipky reprezentují směr pohybu replikačních vidliček po replichórách. Oblast terminu je tvořena dvěma oblastmi syntenie (konzervované pořadí genů), které jsou přerušeny REP privilegovanou oblastí, kde se syntenie rozpadá. Dále je vyznačena pozice dif site a terminus definovaný pomocí zlomu GC skew. Vnitřní kruh reprezentuje KOPS elementy (popsány dále v textu). Prostřední kruh jsou pozice všech REP elementů. KOPS a REP jsou vizualizovány na reprezentativním genomu rodu Stenotrophomonas.

Po zjištění, že terminus je REP privilegovanou oblastí bylo třeba prozkoumat jaké vlastnosti terminus má, kvůli kterým se zde REP nevyskytují. Je otázkou, zda se sem REP nemohou vložit, nejsou schopny se zde dlouhodobě udržet (například díky evoluční nestabilitě této oblasti) nebo jejich inzerce negativně ovlivňuje fitness buňky.

Jako první byla provedena analýza GC skew s pomocí SkewIT (Lu & Salzberg, 2020). Ten nalezl pozice, kde je vrchol kumulativního rozdílu v množství G, resp. C, tato pozice byla nazvána jako terminus. Dále byla provedena analýza využívající maximální množství dat, která by mohla další výzkum nasměrovat. K tomu posloužil program Mauve (Darling et al., 2004). Byla vybrána velká oblast kolem terminu definovaného pomocí GC skew. S pomocí několika dílčích alignmentů byly vytipovány geny, které lze považovat za okraje terminálních syntenních oblastí. Podmínkou bylo, že tyto geny jsou sdíleny všemi analyzovanými kmeny. Z analýzy bylo vyřazeno několik desítek geneticky odvozených bazálních kmenů, které díky divergentním sekvencím mařily syntenní analýzu. Dále byly vyřazeny nesprávně sestavené genomy, které jsme identifikovali nezávisle v několika analýzách. Nakonec bylo použito 69 kmenů, ze kterých byla extrahována oblast terminu (ohraničená dříve definovanými geny) a proveden progresivní Mauve (Darling et al., 2004). Ten odhalil rozsah syntenních oblastí terminu a poskytl informace o jejich konzervovanosti (ilustrace 39).

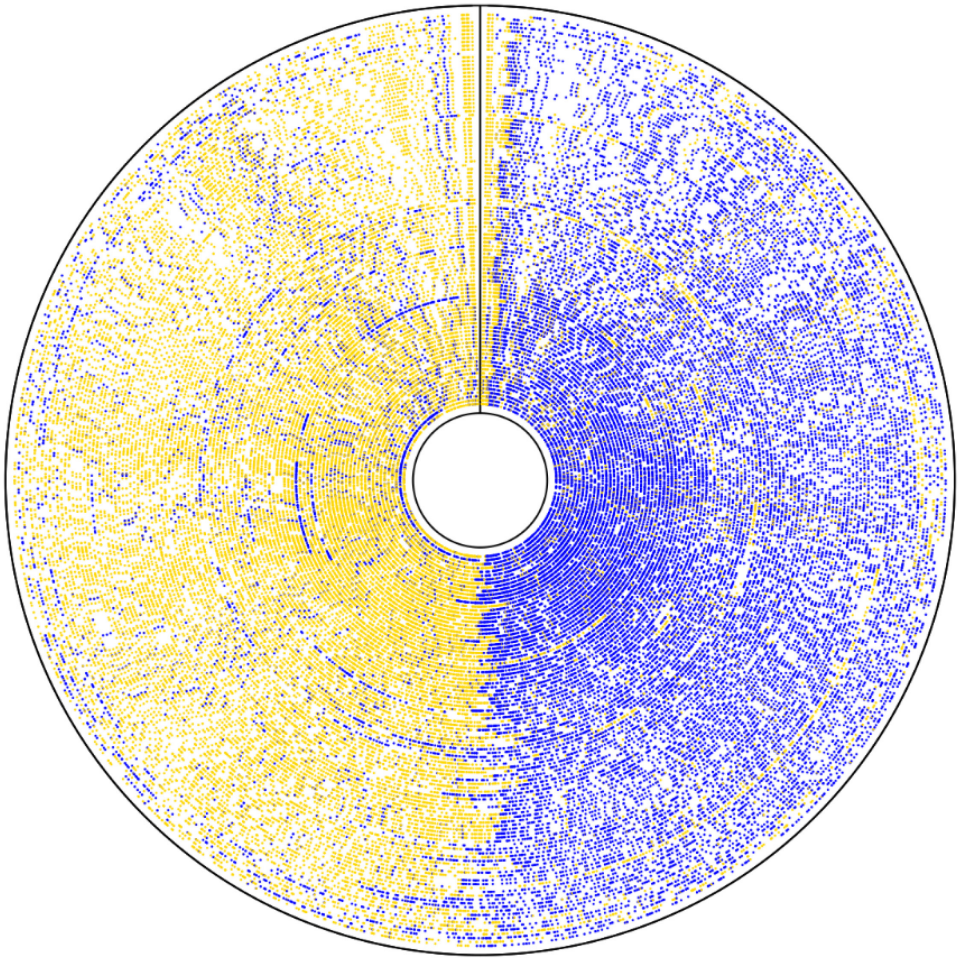


Ilustrace 39. Mauve alignment 69 terminus oblastí. Byly definovány geny ohraničující terminus, které jsou sdíleny mezi kmeny. Poté byla celá oblast extrahována a analyzována progresivním Mauve algoritmem. Geny ohraničující syntenní oblasti jsou zvýrazněny černými čarami. Zároveň jsou sekvence uspořádány podle fylogenetického stromu, první jsou odvozené kmeny (Skupina Sm6), zcela dole jsou pak kmeny bazální (Sm11), ještě bazálnější skupiny byly vynechány. Data z Geneious. Alignment připraven v Mauve, upraven pomocí Inkscape.

Mauve alignment odhalil vysoce konzervovanou oblast levé syntenie, poté o něco méně konzervovanou oblast pravou. Syntenie mezi těmito oblastmi se rozpadá zcela. Právě tato „asyntenní“ oblast je REP privilegovaná. Počátek této oblasti se shoduje s bodem, kde náhle mizí REP (viz ilustrace 41 a 42).

Dále byla prohledána literatura pro sekvence, vyskytující se v oblasti terminu, které jsou asociované s terminací replikace. Jednou z nich je *dif* site (P. Kuempel et al., 1991). Jedná se o 28 bp dlouhý motiv, který se vyskytuje u řady bakterií a je zodpovědný za rozchod chromozomů (po ukončení replikace) do dceřiných buněk. Byly vybrány sekvence *dif* sites z příbuzných rodů *Xanthomonas* a *Pseudomonas* (Mathee et al., 2008; Yen et al., 2002). S jejich pomocí byla nalezena konzervovaná sekvence *dif Stenotrophomonas*. Jde o 25 bp dlouhý motiv, vyskytující se pouze v jedné kopii na genom v blízkosti anti-*ori* všech kmenů.

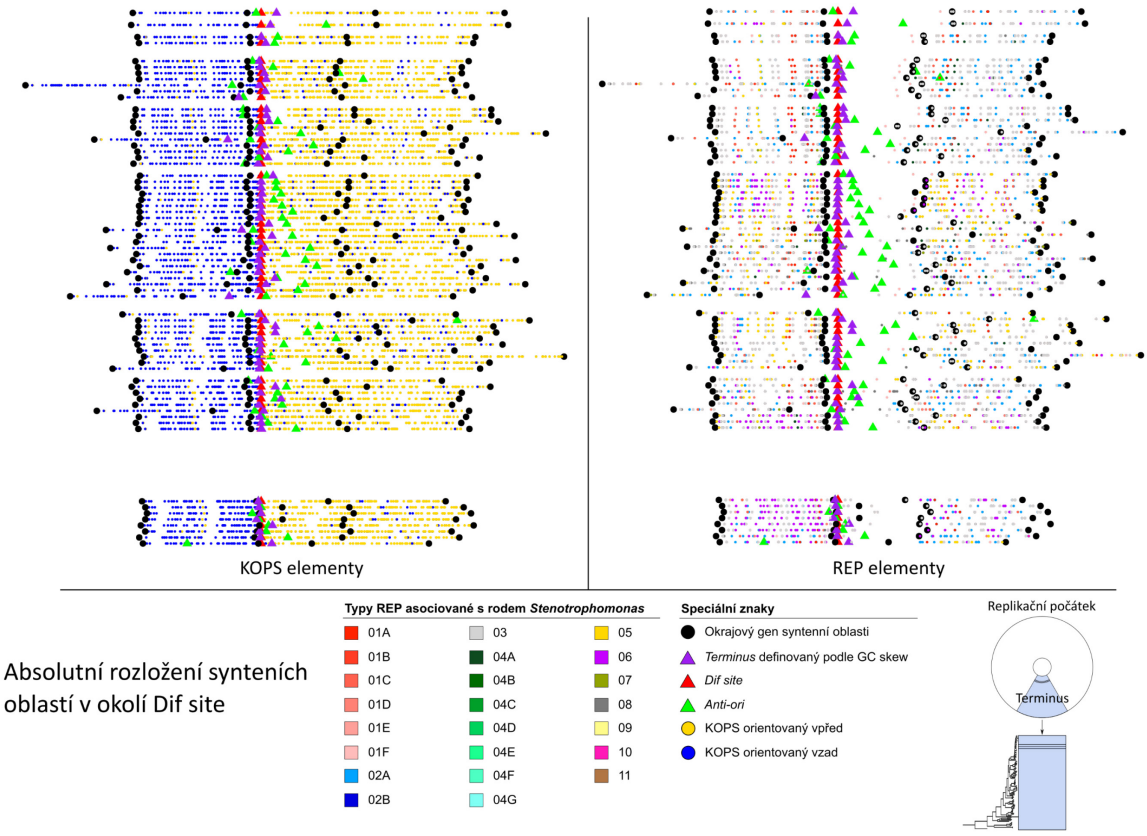
Krom *dif* site byly také identifikovány KOPS elementy (Bigot et al., 2005). Ty pomáhají FtsK vázat se na DNA a zorientovat se na něm (viz kapitola 4.4.2). Sekvence KOPS u rodu *Stenotrophomonas* je GGGCAGGG (identifikováno v práci), tato sekvence splňuje kritéria KOPS konsensu, který je GGGNAGGG. Mapa KOPS na ilustraci níže.



Ilustrace 40. Relativní mapa KOPS ve 102 genomech Stenotrophomonas, které jsou srovnány dle počátku replikace. KOPS jsou zobrazeny jako tečky barveny dle orientace – vpřed (zlatá) a vzad(modrá). Data získána pomocí Geneious prime, vizualizována pomocí R knihovny ggplot2 a circlize.

Konsensus *dif* site platí pro gamaproteobakterie jako *E. coli* či *Pseudomonas*, například u *Bacillus* se však sekvence liší – GAGAAGGG (Nolivos et al., 2012). Elementy jsou rozmístěny relativně rovnoměrně po celém genomu. Jak ukazuje ilustrace 40, jsou dvě oblasti, kde dochází ke zlomu orientace. První zlom překvapivě nenastává přímo na DnaA, ale asi 100 kbp od ní. Od této oblasti směřují elementy k *dif* site. Změna orientace KOPS v okolí *dif* site je ostře vymezená. Mapa KOPS na ilustraci byla připravena obdobně jako mapa REP (ilustrace 36).

Ilustrace 41 a 42 ukazují absenci REP elementů v okolí předpokládané oblasti terminace replikace. REP privilegovaná oblast je na opačné straně od DnaA, nejvzdálenější nukleotid od DnaA je reprezentován zeleným trojúhelníkem jako anti-*ori*. Pozice zlomu GC skew a *dif* site se často překrývají (fialový a červený trojúhelník), zároveň jsou do tohoto bodu orientovány KOPS motivy.

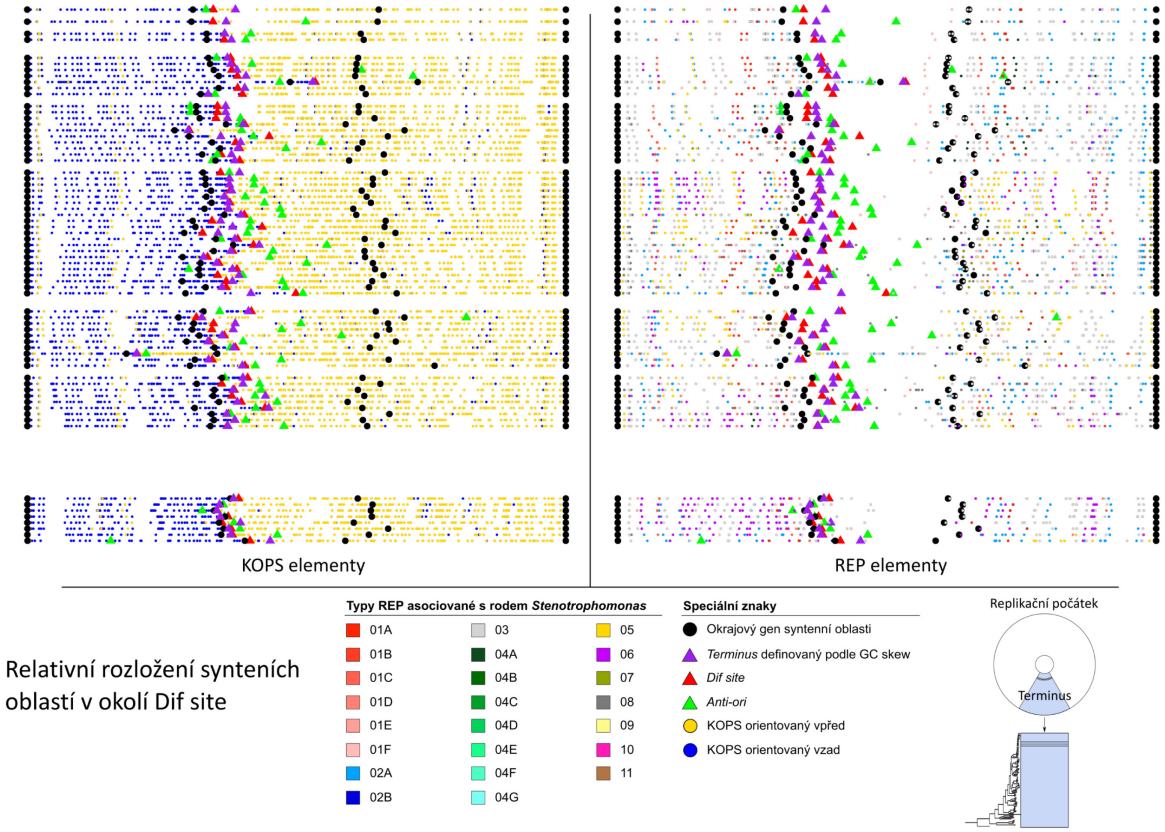


Absolutní rozložení synteních oblastí v okolí Dif site

Ilustrace 41. Vizualizace rozložení KOPS, dif site, REP elementů a GC skew terminu v absolutní velikosti. Jako terminus jsou brány oblasti mezi definovanými okrajovými geny. Ilustrace jsou centrovány podle pozice dif site. KOPS a dif site de novo objeveny s pomocí dat z prací (Bigot et al., 2005; N. P. Higgins, 2007; P. L. Kuempel et al., 1991; Yen et al., 2002). Sekvence terminů jsou seřazeny podle fylogenetického stromu, horní jsou kmeny nejodvozenější, ve spodu jsou kmeny bazální. Mezery jsou prázdná místa po genomech, které byly vyřazeny. Anotováno s pomocí Geneious prime. Exportovaná data byla zpracována pomocí R s knihovnou ggplot2.

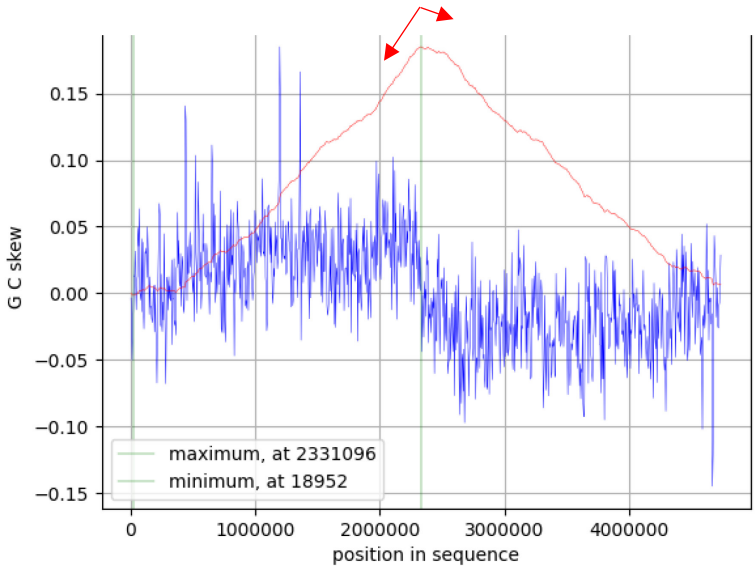
REP privilegovaná je oblast pouze na pravou stranu od *dif* site a překrývá se s oblastí rozpadu syntenie (ohraničena konzervovanými geny – černé tečky na ilustracích). Zdá se tak, že mechanismus zodpovědný za tuto nestabilitu stran přítomných genů a jejich orientace bude, ať přímo či nepřímo, zodpovědný i za absenci REP elementů.

Ilustrace 42 je připravena na základě stejných dat jako ilustrace 41. Rozdílem je, že tato mapa oblasti terminace replikace je srovnána relativně dle okrajů syntenních oblastí. Bioinformatická analýza není schopna odhalit podstatu tohoto jevu. Jednou možností je, že v této oblasti dochází k dekatenaci replikovaných DNA vláken. Při rekombinaci může docházet ke ztrátám DNA, i nízká frekvence stačí, aby tato oblast byla pro každý kmen jedinečná a zbavená všech REP (jejichž frekvence šíření je nízká).



Ilustrace 42. Stejný typ ilustrace jako 41, v této jsou ovšem sekvence terminů relativně srovnány podle okrajových genů syntenie. Sekvence jsou kvůli tomu deformovány, různě dlouhé terminy se zde jeví stejné.

Alternativní hypotézou je, že na anti-*ori* se setkávají replikační vidličky, poté dojde ke zpětnému chodu replikačních vidliček a ty se vrací na *dif* site, kde teprve opouštějí vlákno DNA. REP elementy schopny vytvářet sekundární vlásenkové struktury by mohly tento proces negativně ovlivňovat, což může být důvodem, proč v této oblasti chybí. Pozice *dif* site se prakticky překrývá s terminem definovaným podle GC skew, v této lokaci by se tedy měly setkat replikační vidličky a ukončit replikaci. Vrchol kumulace GC skew je z pravé strany méně ostrý. To znamená, že obohacení/ochuzení o C/G je zde méně výrazné. U některých druhů, spíše než graf s vrcholem, vzniká široké plató (ilustrace 43).



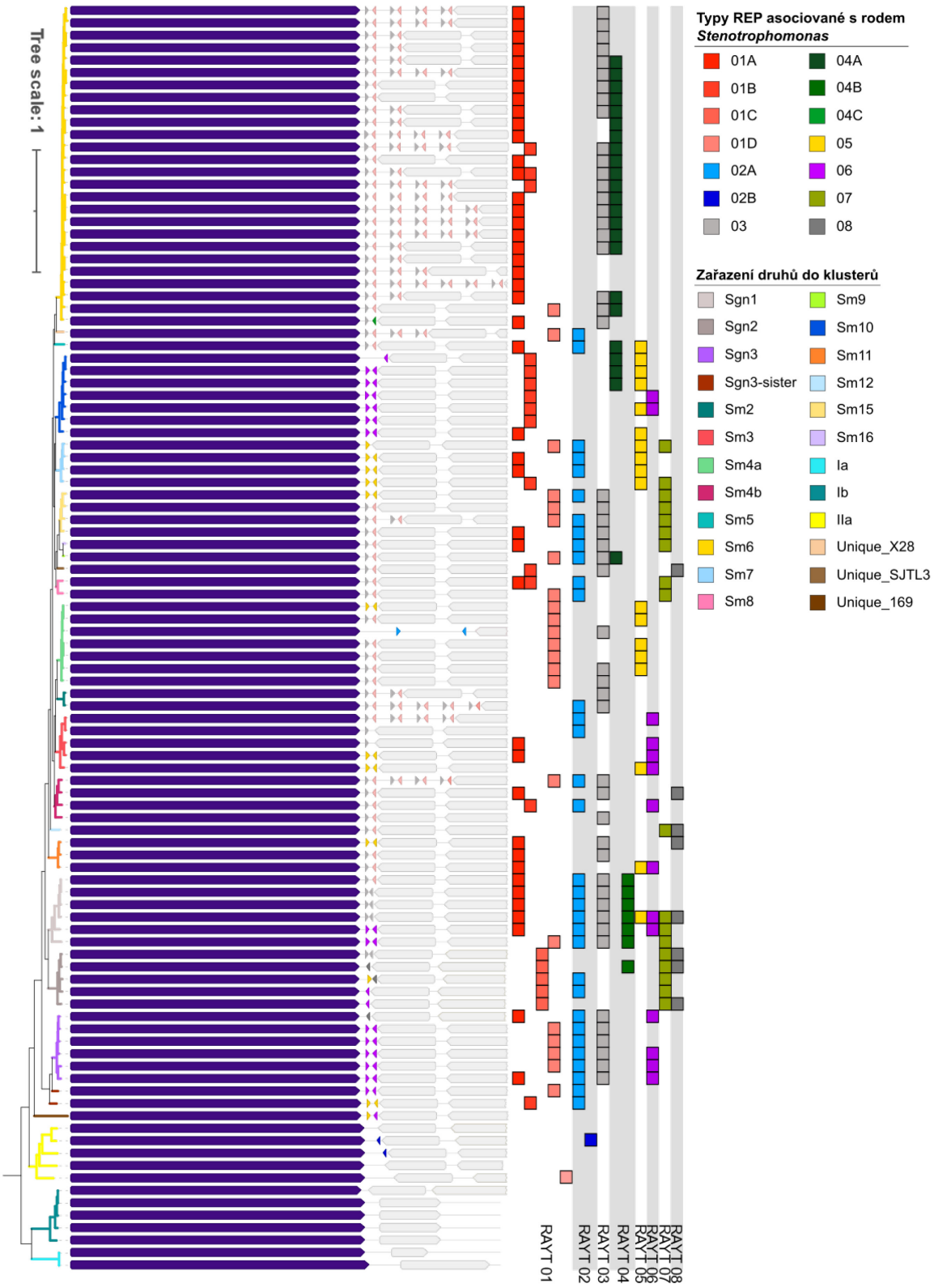
Ilustrace 43. Ukázkový GC skew graf genomu S. m. W18. Místa maxima a minima jsou označena zelenou vertikální čarou. Šipkami nad grafem je ručně naznačen rozdíl ve „směrnici" kolem vrcholu GC skew. Vytvořeno s pomocí SkewIT (Lu & Salzberg, 2020).

Toto plató je u většiny kmenů, zvlášť výrazné je u kmene W18. GC skew je charakteristický pro DNA replikovanou jednosměrně (prostřednictvím definované přítomnosti ssDNA náchylné k deaminaci cytosinu na zpožďujícím se vlákně). Snížená intenzita GC skew v této oblasti proto může souviset právě s pohybem replikační vidličky. Pokud replikační vidličky obrací směr pohybu a vrací se, je příslušný segment DNA replikován oběma směry, což vede k popisovanému pozorování „utlumeného" GC skew.

5.5.4. Analýza variability REP lokusů

V dalším kroku jsme analyzovali variabilitu REP během evoluce hostitelských bakterií v konkrétních intergenových lokusech. Je-li REP přítomen v intergenové oblasti lokusu u ancestrálního kmene, je typické, že jsou REP přítomny u většiny nebo všech kmenů odvozených (viz dále). Často dochází ke změně v sekvence, pozice, orientace i počtu REP.

Bylo identifikováno mnoho REP lokusů (viz kapitola 6.6), jeden z nich prezentuje ilustrace 44. Jde o gen metalopeptidázy rodiny M2 (NCBI lokus tag: G4G30_15005) a downstream intergen. Tento gen se vyskytuje u všech kmenů a je součástí stabilní oblasti DNA. Data pro ilustraci byla získána vyhledáváním BLAST referenčního genu proti všem *Stenotrophomonas*. Bylo extrahováno 102 BLAST „hitů" (sekvencí homologních k referenci) a jejich blízkého okolí (1200 bp). Na ilustraci jsou výsledné oblasti zarovnány podle start kodonu ORF metalopeptidázy a 600 bp downstream (nejde o globální alignment v pravém slova smyslu). Sekvence jsou tak porovnány ve své absolutní délce. Díky tomu lze snadno sledovat dynamiku genového okolí.

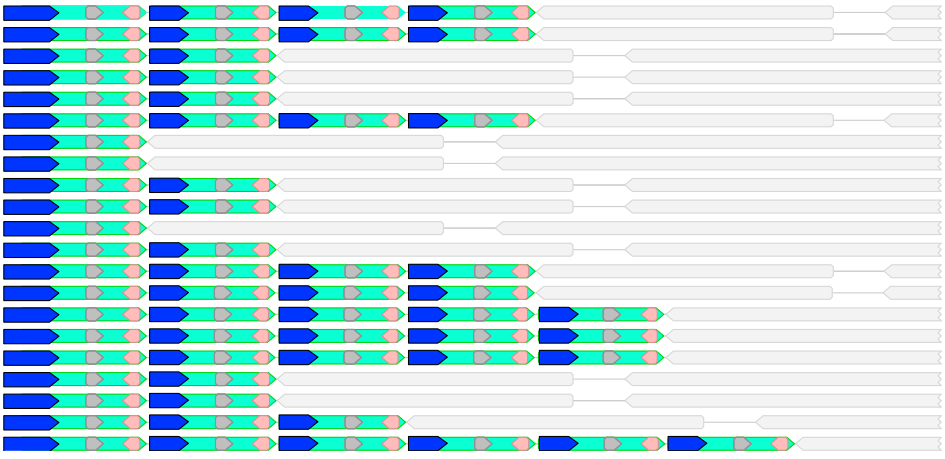


Ilustrace 44. Srovnání downstream oblasti genu metalopeptidázy rodiny M2. V levé části je fylogenetický strom (po ose Y zúžen na 20%) obarven dle standardního schéma. Hlavní část je gen metalopeptidázy (tmavě fialovomodrou) s přidanými 600 bp DNA downstream. Geny kolem metalopeptidázy jsou barveny světle šedou. Jsou vyznačeny REP elementy s maximálně 3 chybami. REP elementy jsou barveny dle legendy, stejné barvené schéma je použito na graf přítomnosti RAYT, pro větší přehlednost. Toto schéma je na pravé straně ilustrace a ukazuje, jaké RAYT jsou v genomu přítomny. Skupiny RAYT, jejichž REP se v lokusu nevyskytují, nejsou uvedeny.

Bazální skupiny (Ia, Ib) nemají RAYT-REP systémy, proto ani v okolí metalopeptidázy nejsou REP přítomny (výjimkou je RAYT 11, jehož REP v lokusu nejsou). Skupina IIa (světle žlutá) má řadu funkčních RAYT-REP. Již v této skupině je první inzerce REP, konkrétně REP 02B. U odvozenějších skupin je patrná výrazná variabilita v počtu, a především typu přítomných REP. Výskyt druhů REP částečně kopíruje fylogenetické vztahy hostitelů a koreluje s přítomností příslušného RAYT v genomech hostitelů dané linie. Nejčastěji se REP elementy vyskytují jako jediný pár typu REPIN, a to homodimer (2xREP 05 či 2xREP 06), nebo heterodimer REP 03/REP 01 (degenerovaný).

U kmenů Sm06 a několika dalších došlo k expanzi REPIN tvořených REP 03 a REP 01F (pravděpodobně jde o poškozený REP 03). Jde o sériové duplikace 167 bází dlouhé oblasti obsahující REPIN a prodloužený 5‘ konec, včetně koncové sekvence ORF metalopeptidázy (světle modrá na ilustraci výše). Tyto oblasti jsou vysoce homologní (nad 90%) a jsou odděleny konzervovanou pět bází dlouhou spacer sekvencí 5‘-AGTCA-3‘. Ilustrace 45 je výsekem několika lokusů z přechozí ilustrace (oblasti downstream od metalopeptidázy) a jsou zde znázorněny tyto duplikované oblasti (tyrkysová).

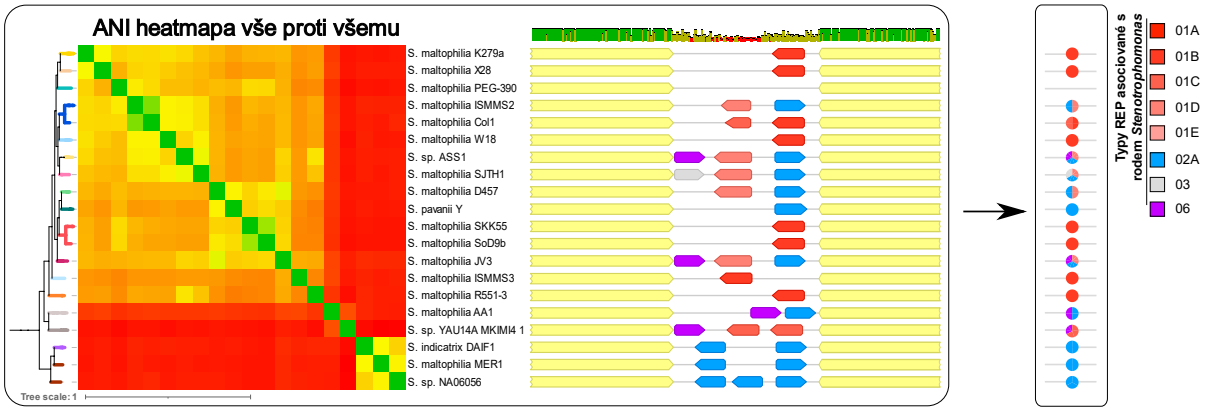
Krom korunní skupiny Sm06 je tato expanze pozorovatelná také u vzdáleně příbuzných Sm02 a Sgn-3 sister. Takto mozaikově uspořádaná struktura je svou architekturou homologická BIME-2 z *E. coli*. Podobné struktury jsou v genomech *Stenotrophomonas* poměrně časté (Nunvar et al. 2010), většina REP elementů se nicméně nachází v samostatných párech typu REPIN. Integrace 3´- koncových sekvencí genů sousedících s REP elementy do BIME je pro *Stenotrophomonas* typická a byla pozorována opakovaně.



Ilustrace 45. Výsek předchozí ilustrace znázorňující pouze pravou stranu od BLAST genu M2 u korunní skupiny Sm6. Přidány jsou pouze 167 bp dlouhé sekvence (tyrkysová) obsahující REPIN. Tyto oblasti jsou odděleny 5 bp mezerami a mají vysokou homologii. 3´-koncová oblast genu, která je součástí repetitivní sekvence, je znázorněna modrou barvou.

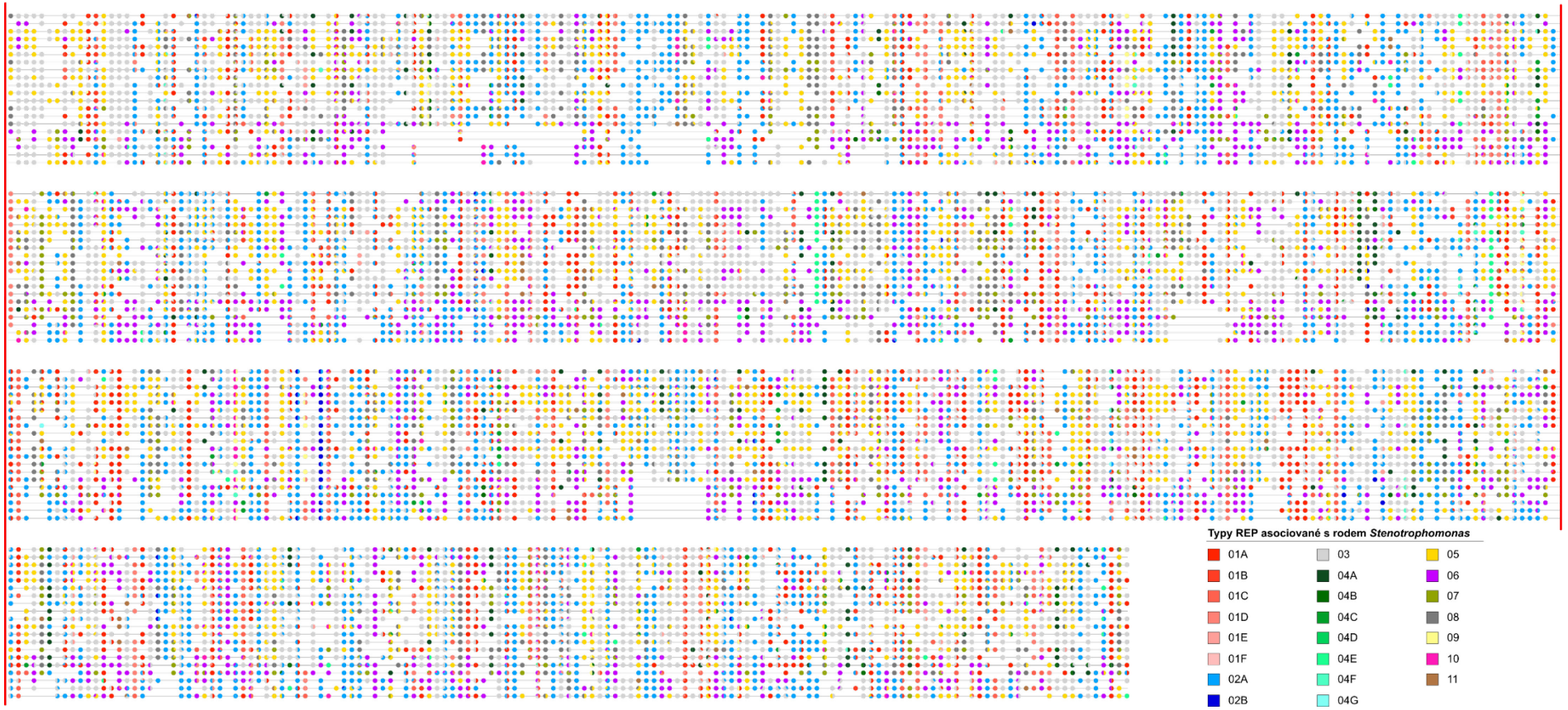
5.6. Globální analýza REP lokusů

Předchozí podkapitola představuje exemplární analýzu dynamiky REP v okolí jednoho konkrétního genu. Abychom zjistili, jaká je typická evoluční dynamika REP v rámci „průměrného" lokusu, bylo potřeba provést analýzu globální. K tomuto účelu jsme zredukovali dataset 102 genomů *Stenotrophomonas* na 20 zástupců, kteří rovnoměrně reprezentují diverzitu rodu (viz příloha 10.1). Homologní oblasti těchto genomů byly nalezeny pomocí Mauve. Vizuálně byly určeny mezigenové oblasti s výskytem REP (přítomny alespoň u poloviny kmenů) – celkem bylo nalezeno 745 takovýchto mezigenových oblastí, pro které bude dále používán termín "REP lokus“. Grafická metodika, jak byla data zpracována ukazuje ilustrace 46.



Ilustrace 46. Postup převodu z alignmentu na globální mapu REP lokusů. V MAUVE alignmentu jsou ručně vybrány REP lokusy (jeden takový je uprostřed ilustrace). Lokus musí obsahovat REP a zároveň vykazovat hodnověrnou homologii okolních genů. Kmeny jsou seřazeny dle fylogenetiky. Lokus každého kmenu je převeden na koláčový graf. Ty tak zobrazují relativní zastoupení REP v tomto lokusu v jednotlivých kmenech. Data byla získána pomocí MAUVE alignmentu v Geneious prime. Data upravena pomocí python (knihovna pandas) a R scriptů (knihovny ggplot2, dplyr, scatterpie). V levé části je přítomen zkrácený fylogram a heatmapa ANI všech 20 kmenů.

Jednotlivé REP lokusy byly extrahovány a REP v každém lokusu pro každý kmen vizualizovány na ilustraci 47. Výsledný graf je kvůli délce rozdělen do čtyř částí pod sebou. Pořadí lokusů je arbitrární, založeno pouze na tom, jak Mauve složil kolineární bloky alignmentu.



Ilustrace 47. Globální mapa REP lokusů v rámci 20 reprezentativních kmenů. Jedná se o MAUVE alignment, ze kterého jsou extrahovány REP bohaté lokusy. Každý lokus je reprezentován sloupcem, kmen je reprezentován řádkem. Jsou použity REP s maximálně třemi chybami. Mapa je vytvořena podle metody popsané předchozí ilustrací. Celkem je zobrazeno 745 REP lokusů. Jeden lokus je zobrazen jako sloupec dvaceti teček, každá tečka je koláčovým grafem relativního zastoupení REP tohoto lokusu v jednom kmenu.

V ilustraci je drtivá většina lokusů obsazená u všech (či téměř u všech) dvaceti reprezentativních kmenů. Menší část lokusů je u bazálních kmenů prázdná, to může být dáno primární absencí REP nebo tím, že tato homologní oblast vznikla později v evoluci. Nepřítomnost REP v daném lokusu v jednom/několika genomech také bývá způsobena absencí celé větší oblasti genomu (včetně REP lokusu + okolních genů). Je přesto zjevné, že REP v lokusech dlouhodobě perzistují. Naopak, minimum lokusů je obsazeno pouze u několika či jediného kmene, což značí, že relativně recentní kolonizace nových lokusů jsou velmi vzácné události. To je v souladu s velmi nízkou četností REP na genomových ostrovech, který jsou typicky recentního původu (viz kap. 6.5.2). Dále je zřejmé, že k záměnám dominantních REP v lokusech dochází poměrně často, a tedy námi pozorovaná dynamika v lokusu metalopeptidázy nepředstavuje v tomto ohledu výjimku, ale běžný stav.
V konkrétních číslech: ve dvaceti kmenech bylo nalezeno 26 999 REP (max 3 chyby), z nichž 91,26% (24 641 REP) je součástí 745 definovaných lokusů.

5.7. Inzerce IS do REP elementů

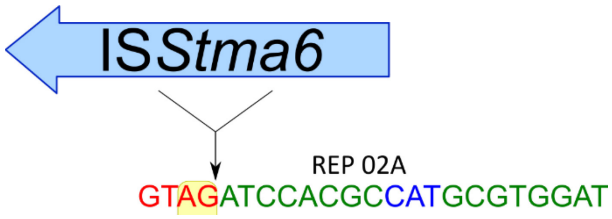
REP elementy *Stenotrophomonas* mají konzervovanou sekvenci. Například REP 03 se vyskytuje 23 360 dokonalých kopií. REP jsou umístěny výhradně v intergenových oblastech, relativně rovnoměrně skrze genom. Ztráta pár REP nemá na fitness buňky zřejmý vliv. Jde tak o ideální cíle pro inzerce IS elementů, přičemž sekvenčně-specifická transpozice do REP byla popsána u IS několika bakterií (kap 4.9.2)

Genomy byly staženy včetně automaticky připravených anotací genů. Automatická anotace identifikovala transponázy řady rodin IS. Konkrétně IS*3*, IS*4*, IS*L3*, IS*21*, IS*91*, IS*110* a IS*481*. Krom nich také pár transpozonů rodiny Tn*3*. Nejpočetnější je rodina IS*3*, následovaná IS*110* a IS*481*. Pro ověření a další zpřesnění znalostí o diverzitě IS *Stenotrophomonas* byly tyto elementy vyhledány nezávisle *de novo*. Byl použit Geneious prime algoritmus pro hledání repetitivních opakování (Repeat Finder). Postup je popsán v metodách, podkapitola 5.5. Výhodou tohoto přístupu je jeho jednoduchost a spolehlivost. Nevýhodou je, že pokud se IS v genomu nachází jen jednou, tak nebude zachycena. Další výhoda přístupu je, že krom samotného genu transponázy zachytí také nekódující terminální sekvence IS, které samotný gen obklopují. Tímto způsobem bylo nalezeno 1086 mobilních elementů. Ty byly (napříč studovanými bakteriálními genomy) uspořádány do skupin podle sekvenční homologie (min. 90% sekvenční identity). Tímto způsobem bylo identifikováno 35 typů IS. Pro klasifikaci IS byl využit server ISfinder (Isfinder, n.d.; Siguier et al., 2006). Některé IS byly již na ISfinder definovány, jiné byly v této práci nalezeny poprvé.

Dalším krokem bylo prozkoumat inzerční specifitu jednotlivých IS – zda transponují do nespecifické DNA či do konzervovaných motivů jako jsou REP.

5.7.1. IS*110*

IS*110* jsou u *Stenotrophomonas* značně rozšířenou rodinou. Již dřívější práce objevily asociaci rodiny IS*110* s REP, kdy členové IS*110* specificky cílí svou inzerci přímo do REP palindromu nebo do jeho blízkosti (Choi et al., 2003). Bylo proto prozkoumáno, zda tento vztah mají IS*110* a REP i zde. Bylo nalezeno sedm zástupců IS*110*, kteří transponují do REP (a jeden zástupce s REP asociovaný). Tyto IS byly rozděleny do několika skupiny, třech již popsaných – ISStma4, IS*Stma6* (ilustrace 48) a IS*Stma7* a pěti námi definovaných – ISStmaNEW1-*5*.

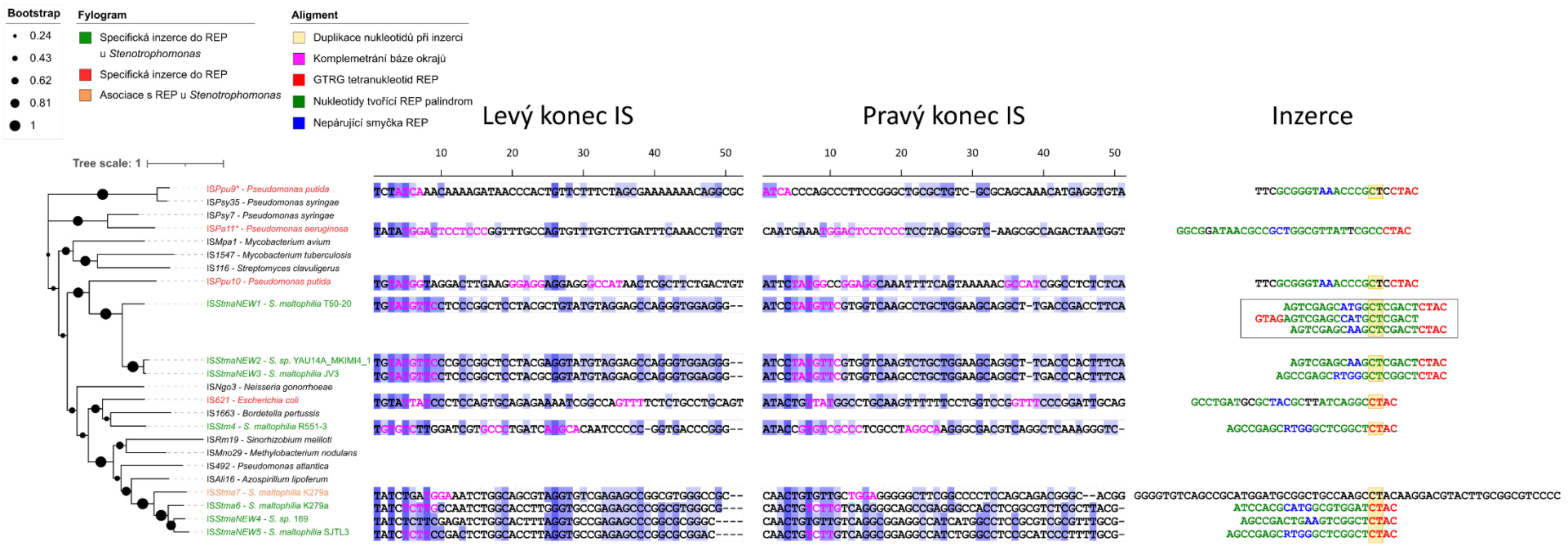


Ilustrace 48. Inzerce ISStma6 (rodina IS110) do REP 02A. Tyto IS cílí svou inzerci specificky mezi konec GTAG tetranukleotidu a počátek palindromu REP. Sekvence REP je zvýrazněna podle stejného klíče jako u ostatních REP ilustrací. Červenou je konzervovaný tetranukleotid, zelenou je palindrom a modrou nepárující loop, zvýrazněn je duplikovaný dinukleotid. Ilustrace připravena v Inkscape.

Byla provedena fylogenetická analýza nalezených členů IS*110* asociovaných s REP u *Stenotrophomonas*, a jiných bakterií (Ramos-González et al., 2006) a všech sekvenčně podobných IS (identifikované BLAST v databázi ISfinder). Z muscle alignmentu AK sekvencí byl připraven strom (metoda maximum likelihood, bootstrap 500) v programu MEGA a vizualizovány v ITOL (ilustrace 49).

IS transponující do REP jsou rozprostřeny skrze celý strom. Lze tak usoudit, že k evoluci inzerční specificity do REP došlo několikrát nezávisle v rámci celé skupiny. IS*Stma4* se vyskytuje pouze u skupin Sm8 a Sm11, ve kterých cílí do REP 05. RAYT/REP 05 přitom není v těchto skupinách dominantně rozšířen. IS*Stma6* je nejpočetnější skupinou IS*110* asociovaných s REP, přes 100 zástupců cílí do REP 02A. IS*Stma6* inzertuje do GTAG tetranukleotidu, přičemž duplikuje CT, analogická je inzerce IS*Stma4*.

IS*Smta07* necílí do REP elementů, cílem transpozice jsou repetitivní sekvence DNA, později identifikované jako YPAL (více v kapitole 6.9). IS*Stma7* je přesto asociovaná s REP, součástí pravého okraje (downstream od ORF) je konzervovaný REP 03. Ten je umístěn 27bp downstream od STOP kodonu a směřuje ke konci kódující oblasti.



*IS*Ppu9* a IS*Pa11* mají přehozený levý a pravý konec

Ilustrace 49. Fylogenetický strom IS110 u Stenotrophomonas, včetně těch nejpodobnějších nalezených v Isfinder. Obarveny jsou názvy IS přítomných u rodu Stenotrophomonas. Ke každému IS je přiřazena sekvence levého a pravého konce. Pravý konec je v reverse komplementu. Alignment těchto sekvencí je zbarven podle míry konzervace. Fialovou je zvýrazněna komplementární sekvence. U inzerčních cílů je žlutou podbarvena oblast, která se při inzerci duplikuje. REP má sekvenci nabarvenou dle legendy. Data z Geneious, připraveno s ITOL, Jalview a Inkscape.

Tři IS (IS*StmaNEW1-3*) pravděpodobně pochází ze společného předka. Jde o příbuzné IS*Ppu10* z *P. putida* (také transponuje do REP). Všechny ostatní IS*110* u *Stenotrophomonas* cílí do CTAC (reverzně komplementární GTAG) a duplikují CT, respektive AG. Nicméně tyto tři IS (stejně jako příbuzný IS*Ppu10*) necílí do CTAC, ale do sekvence samotného palindromu REP 05 a 07, konkrétně do GCTC v rámci palindromu. IS*StmaNEW2* cílí do REP 07 a IS*StmaNEW3* do REP 05. Může jít o divergentní evoluci, při které došlo ke mutaci REP vazebné oblasti (ale možný je i vznik *de novo*). K evoluci pravděpodobně došlo u skupiny Sgn1, tyto kmeny jsou bohaté na REP 05 i REP 07. IS*StmaNEW1* je liberálnější s cílem inzerce. Tento „generalista" cílí jak do REP 07, tak REP 08, do REP 08 přitom v obou orientacích. Důvodem bude identická sekvence palindromických ramen REP 07 a 08. Jediný rozdíl mezi těmito REP je v sekvenci loop.

Z linie vedoucí k IS*Stma6* odvětvují dvě menší větve – IS*StmaNEW4* a IS*StmaNEW5*. IS*StmaNEW4* má pouze 2 zástupce a vyskytuje se jen u *S. sp*. 169. Cílí do REP 04G, které jsou jedinečné pro tento kmen a jsou zde hojně zastoupeny (295 kopií). IS*StmaNEW5* je po IS*Stma4* druhá podskupina, která cílí do REP 05, a to dokonce do stejného místa palindromu. Má 15 kopií, rozšířených ve třech nepříbuzných kmenech.

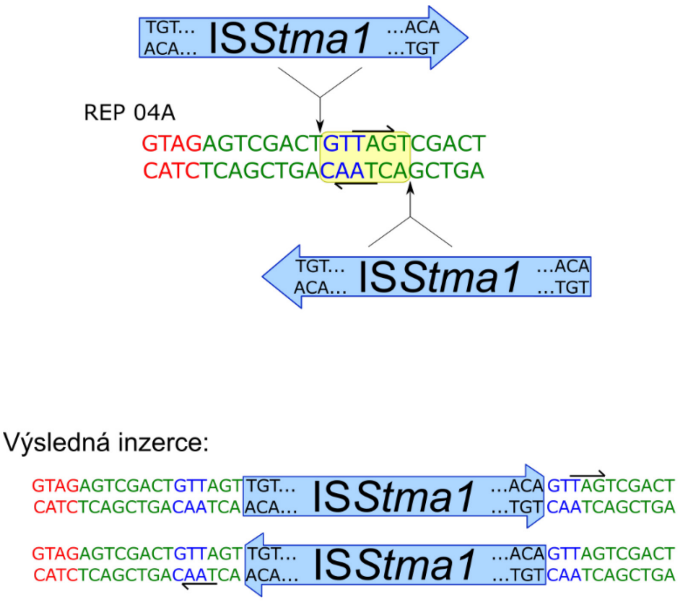
Z analýzy vyplývá, že všechny IS*110* nalezené u *Stenotrophomonas* jsou asociovány s REP, všechny (krom IS*Stma7*) je využívají jako cíl pro svou inzerci. Při inzerci dochází k duplikaci CT. Většina zástupců inzertuje do GTRG tetranukleotidu, nicméně skupina tří příbuzných IS cílí na AG (respektive CT) v palindromu. Díky této flexibilitě nabízí REP (vztaženo na obě orientace) 70 potenciálních inzerčních cílů v rámci 19 (z 23) typů REP.

5.7.2. IS*481*

U *Stenotrophomonas* se vyskytují čtyři zástupci IS*481*. Z AK sekvence byl připraven alignment (muscle) a poté fylogenetický strom (maximum likelihood, bootstrap 500) (ilustrace 51). Dle větvení fylogramu se zdá, že původním inzerčním cílem je NCTAGN hexanukleotid. Ten je při inzerci duplikován a TAG slouží jako STOP kodon transponázy (ilustrace 50). Například IS*481*, podle které je celá rodina pojmenovaná, využívá tento DNA motiv (Stibitz, 1998). IS*StacX*, identifikovaný v této práci má cíl stejný jako původní IS*481*. Je přítomen u bazálních kmenů *S. acidaminiphila* T0-18 a SUNEO (skupina Ib), v těchto kmenech chybí RAYT/REP systémy.

Tři zástupci IS*481* u *Stenotrophomonas* již byli popsáni, jde o IS*Stma1, 3* a *12*. Jsou přítomni u odvozenějších kmenů, ve kterých transponují do REP 04A. Konkrétně do hexanukleotidu složeného z loop a tří nukleotidů ramene. Inzerce na vedoucím vláknu duplikuje 5'-GTTAGT-3', na opožďujícím vláknu je duplikována sekvence 5'-ACTAAC-3' Stejně jako u ostatních IS*481* při inzerci do REP je duplikovaná oblast využita jako STOP kodon (podtržená část).

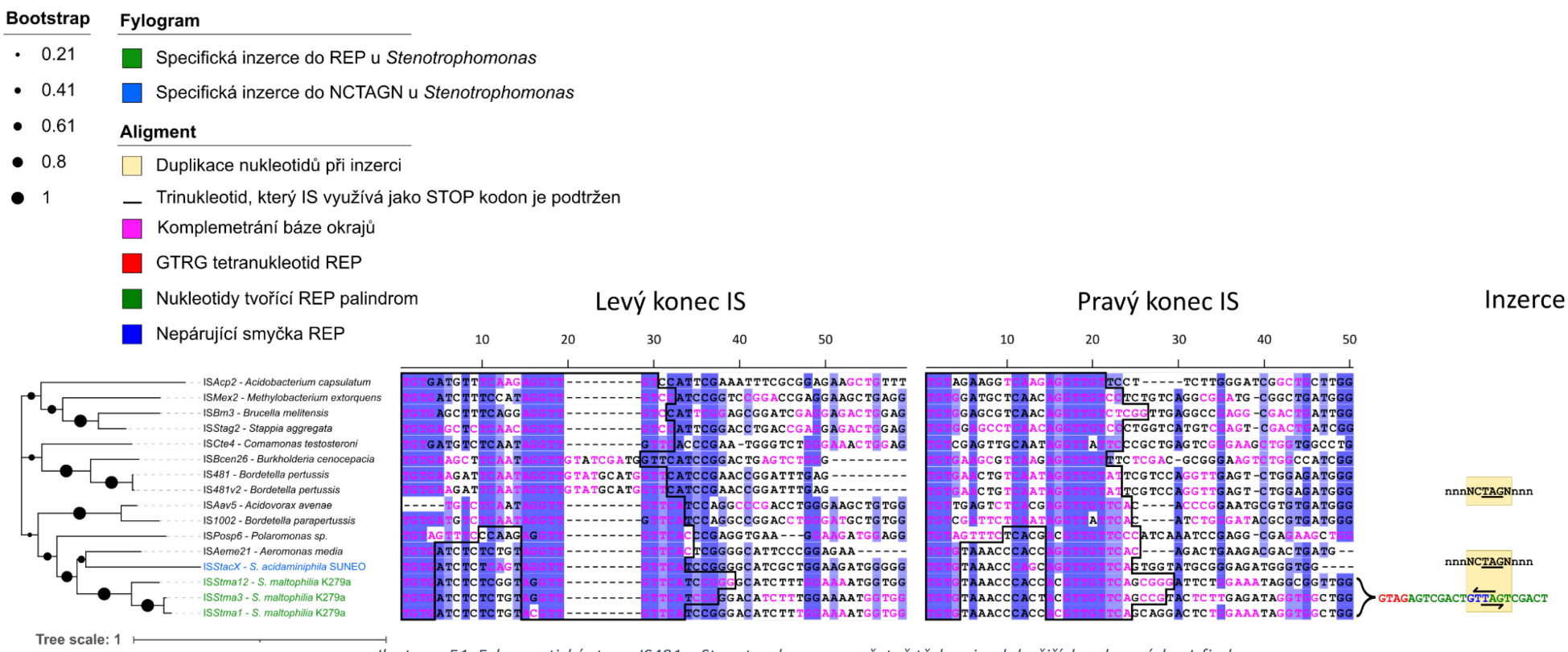
Inzerce IS*481* se zdají starší než například inzerce IS*110*. Části REP elementů, zbylé po transpozici jsou více degradované. Mutace jsou jak v sekvenci palindromu, tak například velmi často z GTAG tetranukleotidu chybí G1. Oblast REP, která je součástí STOP kodonu bývá více konzervovaná. Pravděpodobně právě kvůli své funkci STOP kodonu inserční sekvence.



Ilustrace 50. Schéma inzerce ISStma1 do REP 04A. IS je zobrazena včetně konzervovaného TGT terminálního trinukleotidu. REP je obarven dle standardního schématu, duplikovaná sekvence je zvýrazněna. V druhé části ilustrace je schéma výsledné inzerce, je podtržen STOP kodon kódující sekvence transponázy. Připraveno v Inkscape.

Součástí ilustrace 51 je fylogenetický strom (maximum likelihood, bootstrap 1000) a alignment levého a pravého konce IS. IS*StacX* transponuje do NCTAGN, stejně jako původní IS*481*. Inzerční cíle příbuzných IS definované v ISfinder použity nebyly. IS*Stma1, 3* a *12* jsou sekvenčně homologní, v dalších částech práce jsou považovány za jednu skupinu.

Alignment konců odhaluje, že nejvíce konzervované jsou úplné okraje IS, trinukleotid TGT na samém konci je přítomen u všech zástupců. Krom něj je ještě konzervováno 7 nukleotidů v pozici 15 až 21 od konce. Tyto dvě konzervované oblasti jsou přítomny na obou koncích a jsou vzájemně komplementární, což je stav typický pro většinu IS. IS*110* je v tomto ohledu výjimkou.



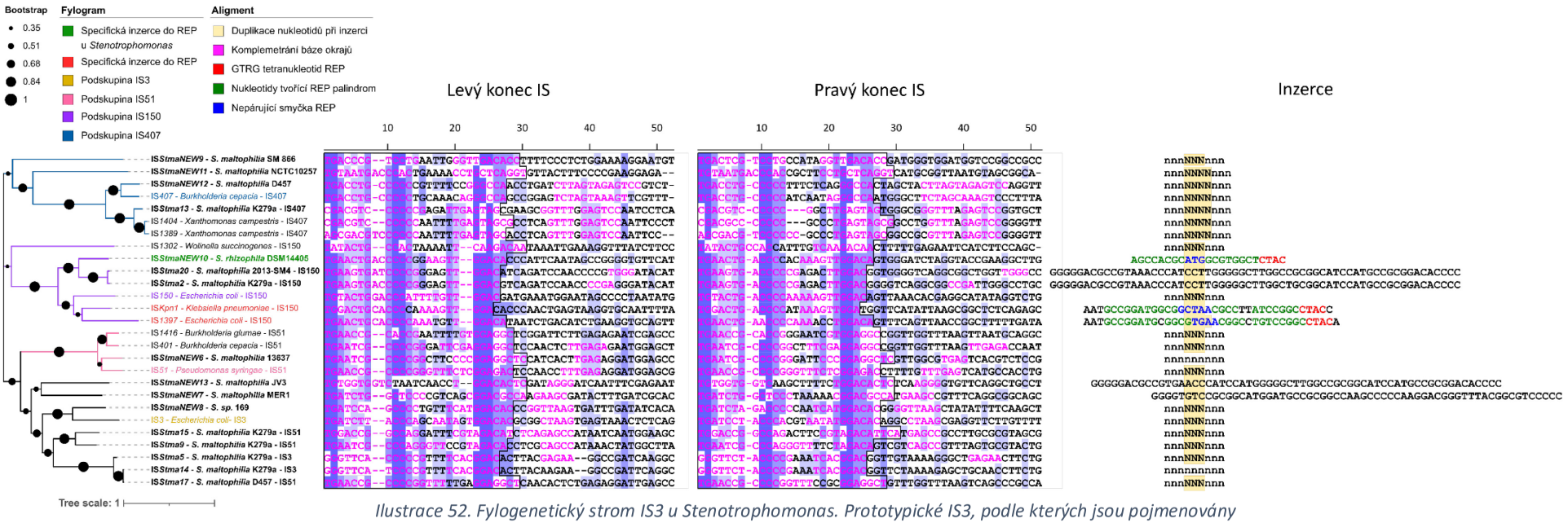
Ilustrace 51. Fylogenetický strom IS481 u Stenotrophomonas, včetně těch nejpodobnějších nalezených v Isfinder. Obarveny jsou názvy IS přítomných u rodu Stenotrophomonas. Ke každému IS je přiřazena sekvence levého a pravého konce. Pravý konec je v reverse komplementu. Alignment těchto sekvencí je zbarven podle míry konzervace. Rámečky zvýrazňují konzervovanou komplementární sekvenci. U inzerčních cílů je žlutou podbarvena oblast, která se při inzerci duplikuje. Podtržen je trinukleotid, který IS po vložení využívá jako svůj STOP kodon. REP má sekvenci nabarvenou dle legendy. Data z Geneious, připraveno s ITOL, Jalview a Inkscape.

5.7.3. IS*3*

Největší známou rodinou inserčních sekvencí je IS*3*. Rodina se vyskytuje u řady bakteriálních druhů a je diverzifikovaná do několika podskupin, jejichž zástupci se hojně vyskytují také u *Stenotrophomonas*. V ISfinder jsou záznamy osmi členů IS*3* u *Stenotrophomonas*. Jde o zástupce čtyř podskupin, konkrétně IS*3*, IS*51*, IS*150* a IS*407*. Krom těchto IS bylo nalezeno osm nových členů IS*3*. Celkem bylo u *Stenotrophomonas* identifikováno 416 kopií různých IS*3*, což je činí nejčetnější přítomnou rodinou.

Automatické anotace řady IS*3* mají polohu frameshiftů určenou nepřesně. K přípravě spolehlivého fylogenetického stromu bylo potřeba manuálně definovat konec OrfA a začátek OrfB. Byl připraven alignment (algoritmus muscle) pro porovnání sekvencí s blízce příbuzným IS*3*, které mají správně určený posun čtecího rámce OrfB (tedy bez předčasných STOP kodonů).

Výhodou práce s IS*3* je přítomnost inverzních opakování, která ohraničují IS. V případě IS*3* ze *Stenotrophomonas* se délka inverzní repetice pohybuje od 12 do 20 bází, někdy dále následuje mezera a inverze pokračuje. U IS*3* typicky dochází k duplikaci tří bází (vzácně čtyř) při inzerci. Lze tak snadno identifikovat konce IS, poté lze odvodit stav DNA před inzercí transpozázy. Fylogram IS*3* byl připravený stejnou metodikou jako v případě IS*110* a IS*481* (alignment muscle, fylogram metodou maximum likelihood a bootstrap 500).



Ilustrace 52. Fylogenetický strom IS3 u Stenotrophomonas. Prototypické IS3, podle kterých jsou pojmenovány podskupiny mají přiřazenou vlastní barvu. Pokud větev obsahuje členy pouze jedné IS3 podskupiny, je celá větev obarvena. Inserční sekvence přítomny u Stenotrophomonas jsou tučně zvýrazněny. Název sekvence je vždy ID – bakteriální druh – zařazení do IS3 podskupiny (je-li definováno). Po fylogramu následuje alignment levého a pravého konce (jeho sekvence je reversně komplementární). V případě inzerce do REP je jeho sekvence obarvena podle legendy. Fylogram vytvořen v MEGA. Vizualizováno v ITOL. Alignmenty připraveny pomocí algoritmu MUSCLE a vizualizovány v Jalview. Upraveno v Inkscape.

Většina IS*3* nemá specifický cíl inzerce. Pro srovnání, IS*110* mají všechny specifický cíl (REP, YPAL), stejně tak IS*481* (REP, NCTAGN). Rodina IS*3* je v tomto ohledu odlišná, pouze jeden člen IS*3* cílí do REP a tři do YPAL. Do REP cílí IS*StmaNEW10* vyskytující se pouze u *S. rhizophilia* DSM14405. Tento kmen má jako jediný funkční RAYT asociovaný s REP 02B, do kterého tato transpozáza cílí. Kmen má také zdaleka nejvíce kopií tohoto REP (87x). Šest kopií IS*StmaNEW10* inzeruje do trinukleotidu 5'-ATG-3', který tvoří loop REP. K inzerci může dojít v obou orientacích a vždy dochází k duplikaci loop. To je stejná situace, jako u již dříve definovaných IS*3* cílících do REP – IS1397 (u *E. coli*) a IS*Kpn1* (u *Klebsiella pneumoniae*) (Wilde et al., 2001, 2003). Tyto IS transponují také do loop oblasti REP, duplikují při tom tři až čtyři nukleotidy.

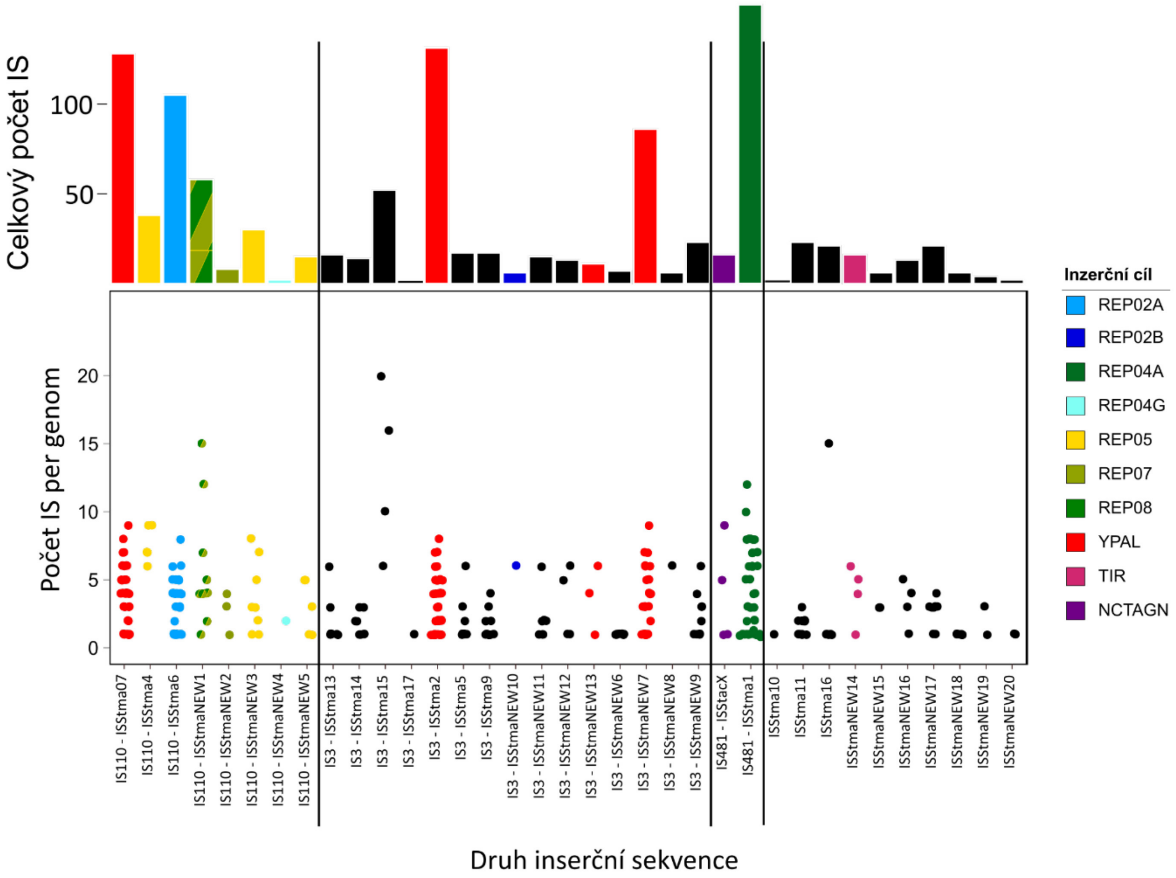
Všechny IS*3* transponující do REP jsou součástí podskupiny IS*150*. V rámci IS*3* analyzovaných v této práci má právě tato podskupina nejčastěji specifický inzerční cíl. Krom inzerce do REP byla nalezena nová konzervovaná sekvence, do které IS*3* cílí (YPAL, kapitola 6.9). Do této repetitivní sekvence rovněž inzertuje IS*Stma7* z rodiny IS*110* (viz výše). Inzertují do ní čtyři inserční sekvence této rodiny, které jsou ze dvou vzdálených linií. Jedná se tedy o nezávislou konvergentní evoluci. Konkrétně jde o IS*Stma2* a *20*, které jsou sesterské (ISfinder udává 95% homologii AK sekvence) a inzerci mají stejnou. Druhou skupinou jsou IS*StmaNEW7* a IS*StmaNEW13*. Ty jsou vzájemně méně příbuzné, pozice inzerce do YPAL se mezi nimi liší.

Ve *Stenotrophomonas* se vyskytuje více druhů IS*3* s nespecifickým inzerčním cílem (11 z 15 IS*3* zástupců) než specifickým. IS se sekvenčně specifickou transpozicí sice tvoří méně jak třetinu z celkové diverzity IS*3*, ale díky vysokému počtu kopií jsou IS*3* s cílenou transpozicí výrazně početnější (celkem 239 z 416 identifikovaných IS*3*).

5.8. Srovnání inzerčních sekvencí

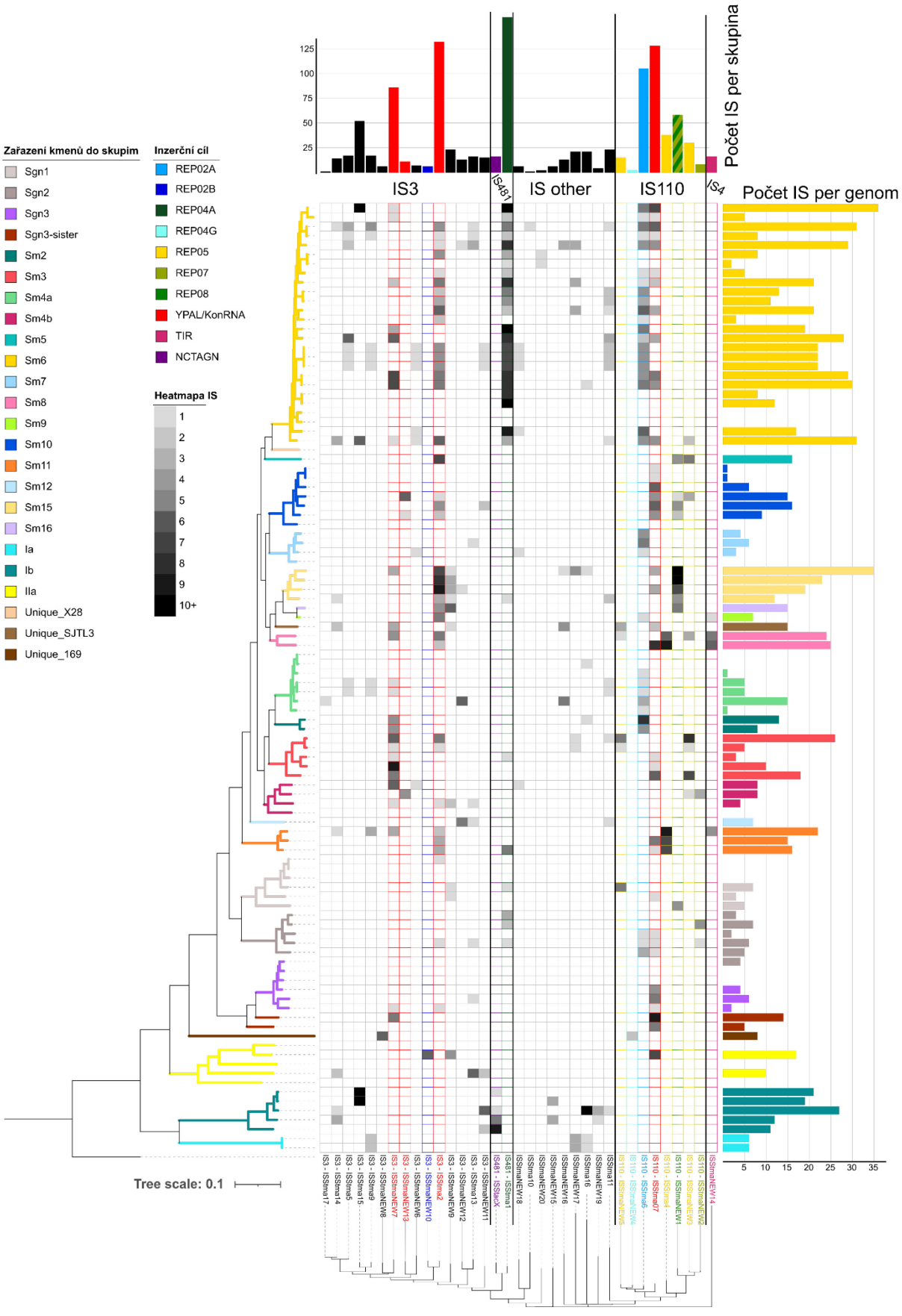
Velké množství inzerčních sekvencí může pro bakterii znamenat značnou zátěž. IS se mohou vkládat do kódujících či regulačních oblastí, což vede nejčastěji ke snížení fitness až smrti buňky. Ilustrace 53 porovnává počty IS v každém kmenu *Stenotrophomonas*. Černé tečky (a sloupce) reprezentují počty IS s nespecifickou inzercí. Ostatní jsou obarveny dle inzerčního cíle (viz legenda). Z dat je zřejmé, že počty IS s nespecifickým cílem (jak celkové, tak v jednotlivých genomech) jsou nižší. Výjimkou jsou IS*Stma15* (rodina IS*3*) a IS*Stma16* (rodina IS*21*). Ty se v několika genomech vyskytují ve velkém počtu. Pravděpodobně jde o expanzi inserčních sekvencí, ke které dochází například při reduktivní evoluci genomu (Ran et al., 2010; Vigil-Stenman et al., 2015). I přes tyto výjimky je zřejmé, že IS s nespecifickou inzercí jsou v rámci jednotlivých genomů méně početné. Pokud jsou v genomu přítomny, vyskytují se typicky pouze v jedné kopii a jsou tak vystaveny většímu riziku extinkce v případě mutace či ztráty.

Oproti tomu IS se specifickou inzercí jsou početnější. Cílem transpozice jsou REP (známo z dřívějších studií i této práce) (Wilde et al., 2003), TIR či YPAL (v této práci i dřívější studie) (De Gregorio et al., 2006). Všechny motivy sdílí vlastnosti, které z nich dělají ideální cíle IS. Vyskytují se v intergenových oblastech ve vysokém počtu vysoce homologních kopií a jejich rozrušení pravděpodobně nevede k vážnému snížení fitness hostitelské bakterie.



Ilustrace 53. Distribuce jednotlivých IS u Stenotrophomonas. Graf s dvěma oblastmi dat. Dolní boxplot znázorňuje počet IS sekvencí v genomech. Pokud je v genomu IS přítomen, je v jeho sloupci tečka. Osa Y pak značí, kolik kopií tohoto IS v daném genomu je. Horní sloupcový graf je součtem všech kopií daného IS. Data jsou obarvena (dle legendy), když má IS specifický inzerční cíl. ISStma2 a 20 jsou spojeny do jedné skupiny, stejně tak ISStma1, 3 a 12.

Alternativní pohled na získaná data nabízí ilustrace 54. Ta k datům přidává kontext fylogeneze kmenů. Ilustrace demonstruje (ne)stálost IS. V rámci jedné fylogenetické linie je IS rozšířena často do všech kmenů, byť v různém počtu. Mezi jednotlivými skupinami je rozložení IS zcela náhodné, bez ohledu na jejich příbuznost. Na heatmapě jsou zvýrazněny (sloupce) IS se specifickým inzerčním cílem. Krom již zmíněných zástupců rodin IS*3*, IS*110* a IS*481* byla identifikována řada inserčních sekvencí z dalších rodin. Ty nejsou početné a ani nevykazují inzerční specifitu, proto nebyly v rámci práce více zkoumány.


Zařazení kmenů do skupim
Sgn1
Sgn2
Sgn3
Sgn3-sister
Sm2
Sm3
Sm4a
Sm4b
Sm5
Sm6
Sm7
Sm8
Sm9
Sm10
Sm11
Sm12
Sm15
Sm16
Ia
Ib
IIa
Unique_X28
Unique_SJTL3
Unique_169
Inzerční cíl
REP02A
REP02B
REP04A
REP04G
REP05
REP07
REP08
YPAL/KonRNA
TIR
NCTAGN
Heatmapa IS
1
2
3
4
5
6
7
8
9
10+
125
100
75
50
25
Počet IS per skupina
IS3
IS481
IS other
IS110
IS4
Počet IS per genom
5
10
15
20
25
30
35
Tree scale: 0.1

Ilustrace 54. Heatmapa inserčních sekvencí rodu Stenotrophomonas. Kmeny jsou srovnány podle fylogenetiky (větve obarveny dle legendy). Inserční sekvence jsou srovnány podle vlastního fylogenetického stromu (zobrazen pod heatmapou). IS se specifickým inzerčním cílem mají obarven název i políčka heatmapy dle legendy. Dále jsou přiloženy sloupcové grafy s absolutním počtem kopií jedné IS (nad heatmapou) a absolutním počtem všech IS v jednom kmeni (vpravo od heatmapy). Data byla získána manuální identifikací IS v Geneious prime. Upravena a vizualizována ITOL. Fylogram všech IS byl připraven v MEGA-X (muscle alignment, maximum likelihood strom s bootstrap 300).

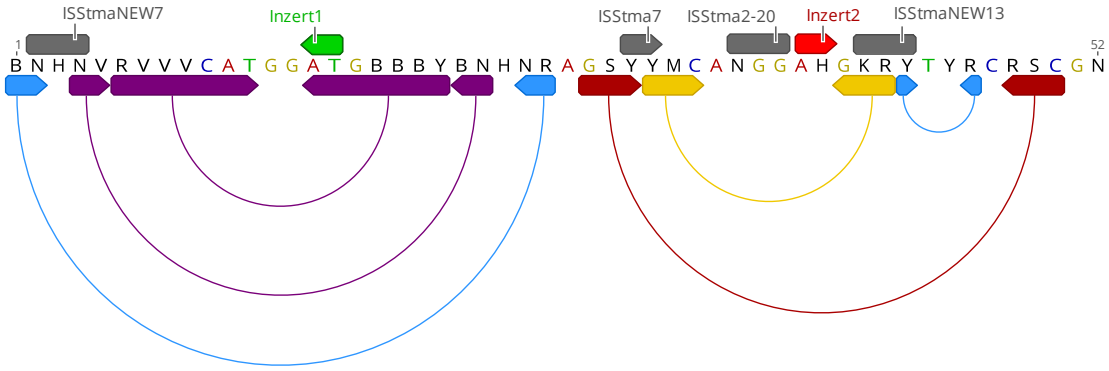
5.9. YPAL

Při analýze IS*3* vyskytujících se u *Stenotrophomonas* pouze jedna transponovala specificky do REP elementů. Tři různé IS*3* pak transponovaly do evidentně konzervované sekvence, kterou se nepovedlo přiřadit k již známým DNA motivům. IS*3*, které cílí do těchto motivů, jsou zdaleka nejpočetnější zástupci rodiny. Krom IS*3* do motivu inzertuje také IS*Stma7*, zástupce IS*110*. Po odvození stavu před inzercí IS byly kopie této sekvence hledány v genomech (pomocí BLAST) pro analýzu konzervovanosti. Z nalezených sekvencí byl připraven alignment a z něj byla určena konzervovaná sekvence tohoto motivu.

Pro podrobnější analýzu těchto repetitivních sekvencí byla použita podskupina dvaceti genomů (seznam těchto kmenů je v metodách 5.7). Celkový počet kopií YPAL v těchto kmenech se pohybuje kolem dvou tisíc (sto per genom). Záleží přitom, jakou metodou byly sekvence hledány a kde přesně byla určena hranice, co se již za YPAL nepovažuje (metody 5.6). Otázkou bylo, zde jde o zatím nepopsaný motiv a zda se vyskytuje pouze u *Stenotrophomonas*. BLAST podobnou sekvenci odhalil (ve vyšším počtu kopií) u několika dalších taxonů bakterií, např. *Marinobacter* a *Yersinia*. Po důkladném prohledání literatury bylo odhaleno, že sekvence homologní těmto repetitivním elementům *Stenotrophomonas* byly u rodu *Yersinia* popsané jako takzvané YPAL elementy (více v podkapitole 4.8) (Bachellier et al., 1999).

5.9.1. Struktura YPAL

Z důvodů relativně nízké sekvenční konzervovanosti je obtížné tyto repetitivní elementy jasně definovat. Tím se odlišují od REP elementů – jeden typ REP je až v tisících identických kopiích rozšířen napříč diverzitou kmenů *Stenotrophomonas*, část kopií pak má jednu až tři mutace (viz tabulka 5). REP se obecně udržují kolem dobře definovatelné „master“ sekvence. YPAL z *Yersinia* jsou v tomto obdobné. YPAL nalezené u *Stenotrophomonas* naopak obsahují minimum absolutně konzervovaných pozic. Celková délka se pohybuje kolem 100 až 120 bp a absolutně sekvenčně konzervovaný je jen zlomek z celkové délky (ilustrace 55).



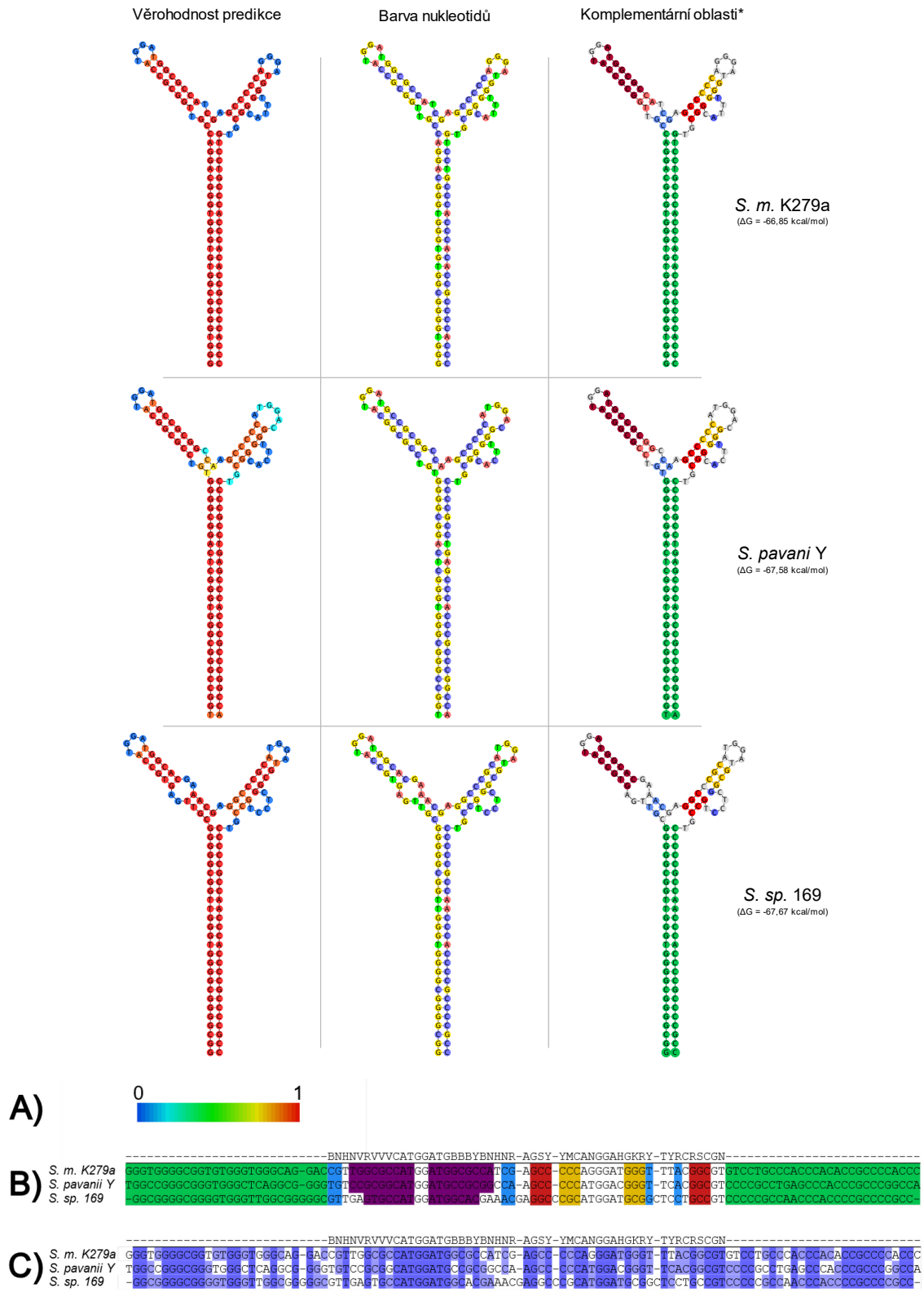
Ilustrace 55. Střední část YPAL sekvence dlouhá 52 bází, z obou stran je obklopena nekonzervovanými rameny, která jsou palindromická (na ilustraci chybí). Pomocí clustal alignmentu YPAL byla získána sekvence nejčastějších nukleotidů, ta byla ručně upravena na degenerované báze, aby obsáhla sekvenci všech YPAL. Na motivu jsou anotovány oblasti, které jsou vždy vzájemně komplementární. Zároveň jsou vyznačeny oblasti, které jsou cílem inserčních sekvencí, Inz1 a Inz2.

Střední část (mezi inverzními rameny) obsahuje řadu komplementárních oblastí, u kterých je mutace jedné vždy kompenzována opačnou mutací v oblasti druhé. Celkově se ze sekvenční analýzy zdá, že motiv nepodléhá náhodné mutagenezi, ale působí na něj silná selekce pro zachování komplementarity zmíněných částí. Střední část motivu obsahuje invariantní motiv 5‘-CATGGATG-3‘, který tvoří střed levé palindromické oblasti.

Střední část YPAL (ilustrace 55) je vždy obklopena velkými komplementárními rameny. Ta nemají konzervovanou sekvenci, nicméně jsou bohatá na GC páry (ilustrace 56). Levé rameno (blíže 5‘-CATGGATG-3‘) obsahuje hlavně guanosin, pravé poté cytosin. Délka ramene se pohybuje od 21 po 30 bp. Víc jak 70% YPAL má ramena zcela komplementární, dalších 20% poté obsahuje maximálně jednu nepárující bázi.

Ilustrace 56. Weblogo sestavené z alignmentu (muscle) 4349 YPAL elementů. Modré šipky naznačují komplementární ramena obklopující středné část. Komplementární oblasti zvýrazněny jako na předchozí ilustraci. Data z Geneious prime, připraveno s pomocí Weblogo3.

Komplementární ramena, která obklopují střední část, nemají konzervovaný motiv. Proto byl napsán skript hledající ramena YPAL nezávisle na jejich sekvenci (viz metody 5.6). Ze sekvencí YPAL v plné délce byly poté predikovány sekundární struktury mnoha těchto motivů. Vybrané reprezentativní YPAL jsou zobrazeny na ilustraci 57.



Ilustrace 57. Predikce sekundárních struktur typických zástupců YPAL ze vzdáleně příbuzných kmenů Stenotrophomonas (na předchozí straně) a legenda. V řádku je vždy stejný element, pouze je jiné schéma barev. První sloupec je dle věrohodnost predikce daného nukleotidu (měřítko A). Druhý sloupec je barven dle bází a třetí dle oblastí, které byly dopředu vyhodnoceny jako komplementární. B) a C) jsou alignmenty zobrazených YPAL. V B) jsou obarveny oblasti, které jsou vždy komplementární. V C) je alignment barven odstínem modré dle konzervovanosti. Sekundární struktury byly připraveny v prostředí Geneious fold prediction algoritmem pro DNA (model 2004) za teploty 25°C. Alignmenty zpracovány v Jalview.

Je zřejmá identická sekundární struktura napříč YPAL z různých kmenů. Již při analýze alignmentu sekvencí jsme odhalili řadu komplementárních oblastí a predikovali selekci nikoliv na konzervaci sekvence, ale sekundární struktury. Naši hypotézu modely sekundárních struktur potvrdily.

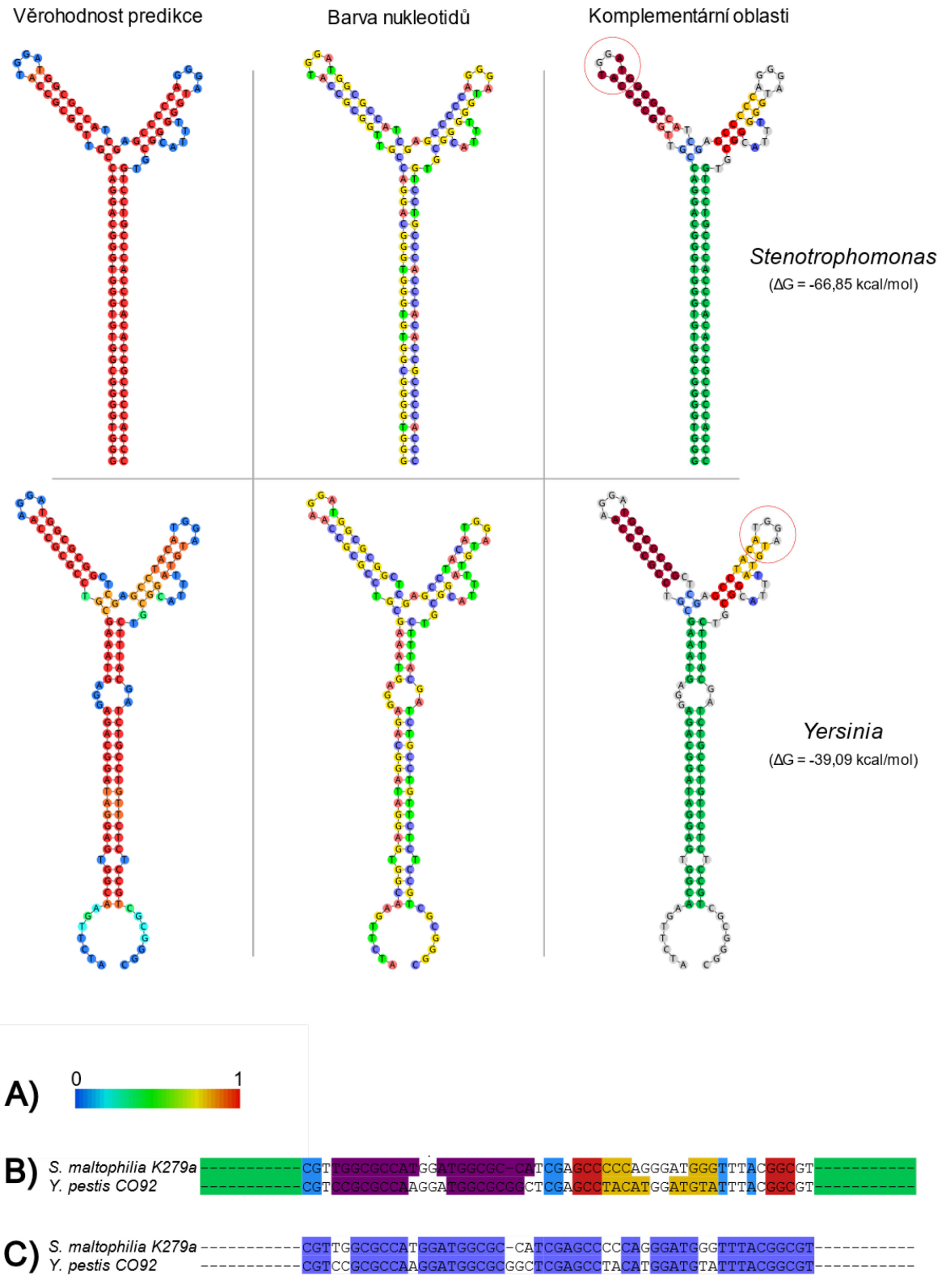
Řada YPAL obsahuje krátké indel mutace, které se ale jen vzácně vyskytují v oblastech komplementarity. Místo toho jsou přítomny v nepárujících oblastech, které na predikované struktuře vytváří malá očka. Vysoce konzervovaný motiv 5‘-CATGGATG-3‘ vytváří v predikci vrchol levé vlásenky, podtržené nukleotidy tvoří nepárující smyčku. Pravá vlásenka a její smyčka jsou méně konzervovány. Výsledkem je motiv podobný vidlici. Obě vlásenky mají ve vrcholu smyčku GG, které jsou přítomny téměř vždy. Možná právě zde dochází k interakci s hypotetickým proteinem zodpovědným za šíření/udržování či jinou interakci s YPAL.

Je třeba podotknout, že bylo predikováno více typů alternativních sekundárních struktur. Nicméně prezentovaná „vidlicovitá“ struktura byla predikována zdaleka nejčastěji a jde o jediný tvar, který byl predikován konsistentně napříč sekvenční diverzitou YPAL, vč. sekvenčně odvozených YPAL.

5.9.2. Porovnání YPAL *Yersinia* a *Stenotrophomonas*

Extragenové motivy nalezené u *Stenotrophomonas* byly identifikovány jako YPAL elementy původně nalezené u rodu *Yersinia*. Zde se jedná o 168 bází dlouhé motivy, vyskytující se v intergenových oblastech řady kmenů ve vysokém počtu (přes sto kopií na genom). YPAL u *Stenotrophomonas* mají podobnou střední část, ta je obklopena perfektně komplementárními inverzními rameny (u *Yersinia* tato inverzní ramena obsahují nekomplementární báze). Také jde o sekvence kratší, o délce lehce přes 100 bází (u *Yersinia* pak 168 bází). Pro YPAL u *Stenotrophomonas* je typický výskyt v dubletech – dva elementy jsou vůči sobě v opačné orientaci a jsou odděleny 100 až 300 bp. Tyto agregované YPAL se u *Yersinia* nevyskytují.

YPAL *Yersinia* a *Stenotrophomonas* také mají řadu podobností. Krom již zmíněné podobnosti středních částí se obě nacházejí v intergenových oblastech a v podobném počtu kopií na genom. U bohatších genomů je to přes sto kopií. Klíčovým sdíleným rysem je stejná sekundární struktura, jak ukazuje ilustrace 58.



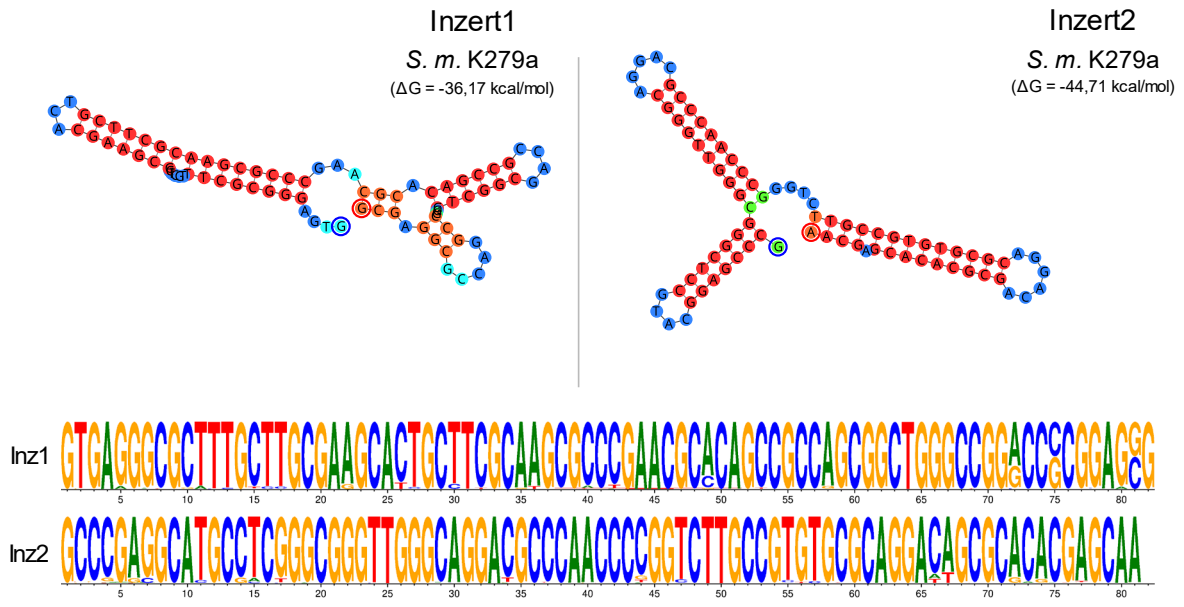
Ilustrace 58. Porovnání sekundárních struktur YPAL mezi Yersinia a Stenotrophomonas. V případě Stenotrophomonas je použita plná sekvence YPAL, u Yersinia je motiv z obou stran zkrácen. Důvodem je, že pouze střední část je vzájemně podobná a vytváří identické sekundární struktury. A) kvalita predikce, B) a C) jsou alignmenty obarveny dle komplementárních oblastí a homologie, respektive. Zakroužkován je CATGGATG motiv. Připraveno v Geneious prime, Jalview a Inkscape.

U *Yersinia* jsou YPAL jednoznačně mobilní a jejich inzerce cílí na sekvenčně nekonzervované palindromy, které jsou uprostřed přerušeny krátkou spacer sekvencí. Za palindromem následuje T-bohatá oblast, jedná se tak pravděpodobně o Rho nezávislé terminátory transkripce. Při inzerci je navíc tato struktura duplikována (De Gregorio et al., 2006). Zda jsou YPAL u *Stenotrophomonas* mobilní a jaká je jejich inzerční specificita, není zcela jasné (předběžné výsledky nejsou jednoznačné) a bude vyžadovat další studium.

5.9.3. Inzert1 a Inzert2

Při analýze YPAL lokusů *Stenotrophomonas* byly nalezeny elementy 82 a 81 bází dlouhé v celkovém počtu 1 011 a 988 kopií, respektive. Všechny kopie každého z inzertů se inzertují vždy do stejné pozice v YPAL (ilustrace 55). Díky jejich charakteristické vlastnosti byly pojmenovány Insert1 (Inz1) a Insert2 (Inz2). Jejich sekvence jsou téměř zcela konzervované. Na rozdíl od YPAL se inzert zdají stále aktivně mobilní – jejich přítomnost v různých YPAL elementech se liší mezi příbuznými kmeny *Stenotrophomonas*. Dokonce jsme pozorovali řadu YPAL elementů, do kterých se vložily oba inzerty. BLAST neodhalil sekvence homologické Inz1 ani Inz2 mimo *Stenotrophomonas* – tyto elementy jsou tedy pro rod absolutně specifické.

Byly připraveny predikce struktury obou inzertů (Ilustrace 59). Díky přítomnosti komplementárních úseků v jejich sekvencích se jak Inz1, tak Inz2 skládají do poměrně stabilních sekundárních struktur. YPAL s vloženým Inz1 či Inz2 (ev. oběma) tak reprezentuje sekundární struktury „vyšší úrovně“.



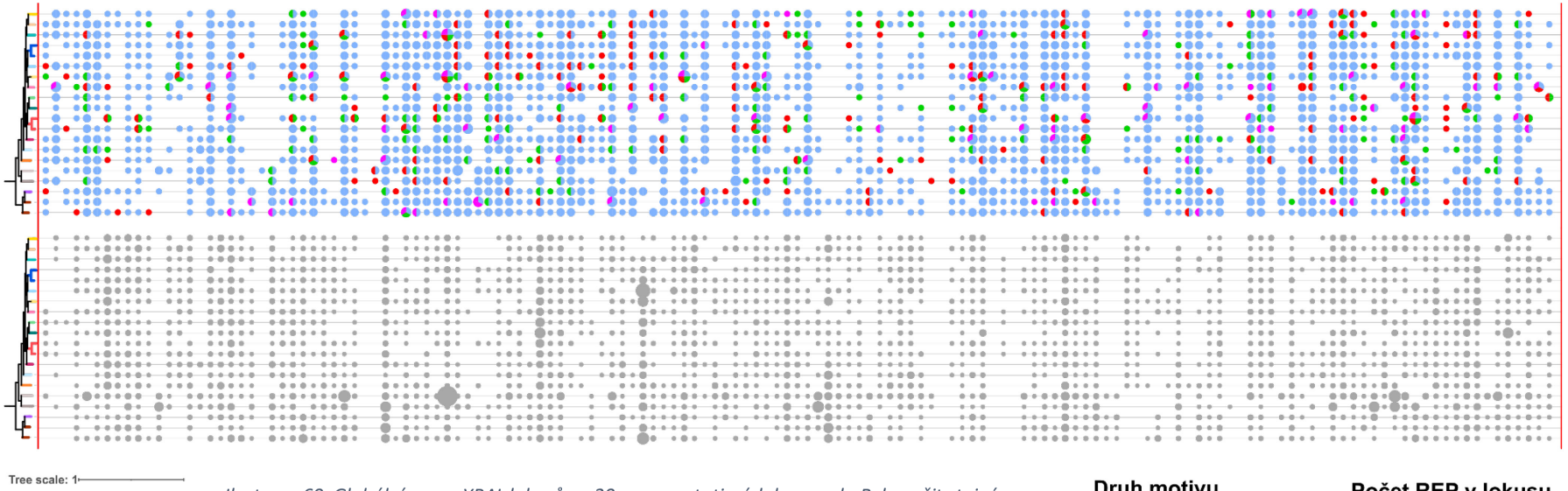
Ilustrace 59. Porovnání sekundárních struktur Inzertů 1 a 2 mezi vybranými kmeny Stenotrophomonas. V dolní části ilustrace poté WebLoga připravená ze všech nalezených inzertů (1011 a 988 kopií, respektive). Webloga připravena v (WebLogo 3 - Create, n.d.). Sekundární struktury připraveny v Geneious prime.

Nelze si nevšimnout paralel mezi Inz1, Inz2 a samotnými YPAL elementy, do nichž se inzertují. Všechny tyto elementy jsou schopné tvořit extenzivní sekundární struktury. Všechny jsou dle *in silico* predikcí (komparativní genomika) mobilní, byť v různých hostitelských bakteriích (YPAL: *Yersinia*, Inz1/Inz2: *Stenotrophomonas*). Mechanismus mobilizace je neznámý a lze těžko odhadovat – elementy mohou být mobilizované v různé formě (DNA, RNA) a s různým katalyzátorem reakce (enzym, ribozym).

5.9.4. Evoluční dynamika YPAL lokusů

Podobným způsobem jako pro REP elementy (viz Kap. 6.6) byla studována evoluční dynamika YPAL v rámci lokusů, které okupují. Ilustrace 60 ukazuje, že většina YPAL se nachází v rámci evolučně konzervovaných lokusů – 89,1% (2524 z 2834). Jde o dvě vyobrazení stejných 147 manuálně nalezených YPAL lokusů (z Mauve alignmentu subsetu dvaceti genomů). V horní části je mapa YPAL lokusů, včetně mobilních elementů s nimi asociovaných (IS, Inz1, Inz2). Dolní část zachycuje tytéž lokusy, ale mapuje pouze počty REP v nich. Většina YPAL lokusů je konzervovaná napříč diverzitou dvaceti vybraných kmenů. Lokusy okupované pouze u několika kmenů jsou typicky součástí větších oblastí, které u zbytku kmenů chybí.

Většina YPAL (přibližně 70%) se vyskytuje v dubletech, tyto YPAL vždy směřují vůči sobě opačně. Vzdálenost (spacer) mezi elementy v dubletu se pohybuje od 100 do 300 bp. Rozdíl je daný dynamikou počtu REP, které mezi YPAL hojně vyskytují. REP zde nemají preferovanou pozici, orientaci ani sekvence. Nikdy však nenarušují integritu YPAL, vždy je zachována určitá vzdálenost mezi YPAL a REP motivy. U zkoumaných kmenů má přes polovinu YPAL do vzdálenosti 400bp přítomen REP. REP jsou v těchto kmenech rozšířeny v méně jak 20% všech intergenových oblastí. Téměř polovina REP sousedících s YPAL jsou REP 02A. Přitom REP 02A tvoří jen 15% všech REP u těchto kmenů. Existuje tak zřejmá korelace společného výskytu REP a YPAL. Vzhledem k relativní stabilitě výskytu YPAL v příslušných lokusech napříč diverzitou hostitelských kmenů se domníváme, že tyto elementy jsou, alespoň v recentní evoluční historii, imobilní. Otázkou je, zda dávné inzerce YPAL směřovaly do lokusů již obsazených REP, či naopak.



Ilustrace 60. Globální mapa YPAL lokusů ve 20 reprezentativních kmenech. Byl použit stejný MAUVE alignment jako u mapy REP lokusů. Na něm byly ručně anotovány YPAL lokusy. Vrchní část ukazuje zastoupení YPAL, inzertů a inserčních sekvencí v YPAL lokusech. Velikost teček se liší, dle počtu prvků v daném lokusu. Dolní část ukazuje ty samé oblasti DNA, ale zobrazeny jsou jen počty REP elementů (dle velikosti teček). Data připravena pomocí Geneious prime, python (knihovna pandas) a R (knihovna ggplot2, scatterpie, dplyr, ComplexHeatmap) (Gu et al., 2016; Hadley Wickham et al., 2023; Wickham et al., 2019; Yu, 2023).

6. Diskuse

Díky rozvoji sekvenačních technologií nové generace a snižujícím se finančním nákladům intenzivně roste počet kompletně sekvenovaných bakteriálních genomů. S pokrokem bioinformatiky lze stále přesněji predikovat kódující oblasti a jejich potenciální produkty. Analýzou těchto dat lze bakterie detailně popsat a predikovat jejich vlastnosti. To je zvláště užitečné v enviromentální mikrobiologii, analýzou DNA z prostředí lze predikovat složení společenstva a identifikovat druhy s unikátními biosyntetickými drahami formujícími celý ekosystém (a také bakterie s biotechnologickým potenciálem). Komparativní analýzou kompletních (tj. uzavřených) genomových sekvencí příbuzných bakterií lze detailně studovat strukturální změny během evoluce genomu – např. horizontální přenos či chromozomové přestavby. S dostatkem genomických dat lze spolehlivě rekonstruovat fylogenezi bakterií jak v rámci rodů, tak s rozlišením jednotlivých kmenů. Analýza celogenomových sekvencí také umožňuje detekovat epidemické klony a monitorovat jejich evoluci během epidemií. To je jen část důvodů, proč si analýza kompletních bakteriálních genomů zaslouží vědeckou pozornost.

Základem našeho výzkumu byla fylogenetická analýza rodu *Stenotrophomonas*. Dnes je k dispozici přes sto kompletních genomů rodu *Stenotrophomonas*. Na jejich základě jsme sestavili fylogenetický strom rodu s vysokou spolehlivostí větvení. Námi predikovaná větvení kmenů do skupin byla podpořena daty z komparativně genomických studií *Stenotrophomonas* provedených v minulých letech (Gröschel et al., 2020; Mercier-Darty et al., 2020; Vinuesa et al., 2018).

Přítomnost REP u *Stenotrophomonas* již byla popsána (Rocco et al., 2010), obdobně jsou u rodu známy funkční RAYT/REP systémy (Nunvar et al., 2013). Řada konkrétních REP zkoumaných v naší analýze tak již popsána byla, nicméně autoři měli k dispozici méně kompletních genomů a studie méně stavěly na evolučním vývoji rodu. V práci byla zkoumána evoluce RAYT/REP systémů v hostitelských genomech. Z 331 nalezených RAYT je část pseudogenizovaná, celkem jsme tedy definovali 271 potenciálně funkčních RAYT. *rayt* geny jsou součástí stabilního genomu bakterie – mají vlastnosti house-keeping genů (Bertels, Gallie, et al., 2017). Mezi nepříbuznými bakteriemi se RAYT šíří primárně horizontálním přenosem (Bertels & Rainey, 2011b). Poté co se RAYT/REP systém ustanovil u předka moderních *Stenotrophomonas*, perzistuje v genomech hostitelských bakterií a je součástí všech (plně sekvenovaných) kmenů (s výjimkou skupin Ia, b). Rekordmanem v tomto ohledu je *S. sp.* ZAC14A_NAIMI4_1, tento kmen je osídlen rovnou osmi pravděpodobně funkčními RAYT/REP systémy, což je dvakrát více než bylo dosavadní popsané maximum. Nebyl nalezen jediný genom s duplikovaným *rayt*, zároveň ani genom, kde by *rayt* byly úplně ztraceny (po prvotním rozšíření). Naše výsledky souhlasí s teorií, že RAYT/REP systémy jsou stabilně udržované vertikálním přenosem a pozitivní selekcí, nikoliv horizontálním přenosem (Bertels, Gokhale, et al., 2017).

Není jasné jak RAYT/REP systémy původně vznikly, dnes nejpravděpodobnější se jeví teorie domestikace inserčních sekvencí rodiny IS*200*/IS*605* (Nunvar et al., 2010). S těmi sdílí RAYT řadu homologních oblastí (Barabas et al., 2008). Původně sobecká transpozáza pravděpodobně ztratila schopnost transpozice vlastní sekvence, ale stále byla schopna šířit krátké ssDNA motivy s inverzním okrajem – REP. Regulovaná diseminace REP je pro buňky zřejmě výhodná a systém se tak stal stabilní součástí genomu. Stabilita neznamená stálost, ve skutečnosti na RAYT působí divergentní selekční tlak. Odlišnost RAYT v rámci příbuzných kmenů byla pozorovány i u dalších druhů (Bertels, Gallie, et al., 2017; Bertels & Rainey, 2011a). U *Stenotrophomonas* jsme identifikovali celkem 11 RAYT lokusů. Sekvence RAYT ze stejného lokusu jsou vzájemně příbuznější než mezi lokusy. Zároveň se mezi lokusy mění sekvence REP, se kterou je RAYT asociovaná a kterou pravděpodobně šíří. RAYT vždy šíří ty REP, které jsou v okolí jeho genu – často jsou to nejčetnější REP v genomu. Je-li v genomu současně více RAYT v různých lokusech, vždy jde o sekvenčně odlišné proteiny šířící různé REP. To bylo potvrzeno u všech námi nalezených RAYT. Zvláštností jsou systémy z lokusů 01, 02 a 04, zde RAYT (v rámci jednoho lokusu) divergovaly natolik, že šíří odlišné REP. Zvlášť RAYT 01 vykazuje abnormální evoluci a zdá se, že neznámým způsobem opakovaně dochází k nahrazení *rayt* genu v lokusu (včetně nahrazení REP), aniž by bylo ovlivněno bezprostřední genové okolí. Tento typ sekvenčně-specifické evoluce DNA nebyl dříve pozorován a není jasné, na jakém principu funguje. Většina RAYT lokusu 02 je pseudogenizována, může jít o mechanismus, jak buňky zabraňují excesivnímu rozšíření REP skrze genom. Obdobný stav byl popsán u *Pseudomonas* (Nunvar et al., 2013).

REP nejsou šířeny jednotlivě, ale v rámci REPIN, tedy REP dimerů. Časem může dojít ke ztrátě jednoho REP či naopak k amplifikaci REPIN do BIME mozaik složených z tandemově amplifikovaných REPIN. V práci jsme nově zjistili, že na první pohled chaotické rozšíření REP v genomu je ve skutečnosti částečně statické. Šíření REP je limitováno do relativně malého množství intergenových oblastí neboli REP lokusů. U subsetu dvaceti kmenů jsme definovali 745 lokusů, ve kterých se nacházelo 91,26% všech přítomných REP (24 641 z 26 999, tj. průměrně 1232 kopií REP na genom). K osidlování nových lokusů dochází, ale s nízkou frekvencí. Teprve při porovnání konkrétního REP lokusu napříč diverzitou hostitelských bakterií se projevuje opravdová diverzita REP – liší se zde počty, pozice, orientace i sekvence REP. Velice časté je nahrazení („přepis") REPIN novým párem REP, který přesně nahradí sekvenci starého REPIN bez vlivu na okolní DNA. Nový REPIN dále může být amplifikován v BIME mozaiky, často složené z různých REP. Je otázkou, proč jsou REP lokusy tak variabilní. Zda přítomnost prvního REP stimuluje aktivitu RAYT v okolí, nebo jsou zde REP udržovány díky pozitivnímu vlivu na fitness hostitele při zároveň uvolněné selekci na jejich počet/sekvenci. Omezení aktivity RAYT výhradně na tyto lokusy může být také mechanismem zabraňujícím nekontrolovanému rozšíření REP skrze genom.

Již dlouho dobu jsou známy rozdíly v globálním rozložení REP, konkrétně jejich absence na mobilních elementech (Loper et al., 2012) a v oblasti terminace replikace (Bachellier et al., 1999). Absenci v rámci mobilní DNA jsme potvrdili, nalezli jsme však výjimku v podobě IS*Stma7*, tato inserční sekvence nese REP 03 přímo v nekódujícím pravém okraji. Do hloubky jsme dále analyzovali REP privilegovanou oblast *ter* domény (region terminace replikace). Zde se REP privilegovaná oblast téměř přesně překrývá s oblastí rozpadu syntenie u všech zkoumaných kmenů. Levý okraj této „asyntenní" REP privilegované oblasti přitom začíná na pozici, kde pravděpodobně dochází k ukončení replikace. Kolokalizuje zde *dif* site, zlom GC skew a také zlom orientace KOPS elementů. Pravý okraj oblasti je pak definovaný pouze ztrátou syntenie. Délka asyntenní oblasti se pohybuje kolem 150 až 250 kbp. Předpokládáme, že REP zde chybí kvůli potenciálně negativnímu vlivu na terminaci replikace. Z určitých indicií (směrnice GC skew) hypotetizujeme, že po setkání replikačních vidliček („pre-terminace") v místě asyntenní oblasti dochází ke zpětnému chodu replikačních vidliček při segregaci dceřiných chromozomů směrem k *dif* site. Tento stav byl již byl popsán například u *Pseudomonas* (Bhowmik et al., 2018). Možná právě pohyb replikačního komplexu po DNA nepřipouští přítomnost motivů tvořících sekundární struktury. Alternativně, v této oblasti může být zvýšená frekvence rekombinace mezi dceřinými chromozomy. Ty jsou zde dekatenovány s pomocí topoisomerázy IV a systému Xer rekombináz (Aussel et al., 2002; Blakely & Sherratt, 1994). Zvýšená míra rekombinace by vysvětlovala i ztrátu syntenie oblasti. Absence REP by mohla být způsobena vyšší frekvencí ztrát, než je rychlost šíření REP.

S pomocí AlphaFold2 byly predikovány strukturní modely zástupců jednotlivých RAYT skupin. Ty byly porovnány se strukturou RAYT z *E. coli* získanou experimentálně v dřívějších pracích (Messing et al., 2012). Analýza odhalila, že proteiny sdílí řadu konzervovaných aminokyselin v okolí REP vazebné oblasti. Samotná struktura katalytického místa, kde dochází k vazbě a štěpení ssDNA REP, je také konzervovaná. Katalytická oblast je však u *E. coli* zakryta C-terminální doménou. Ta u *Stenotrophomonas* chybí, její absence může být zodpovědná za vyšší aktivitu těchto RAYT, která se projevuje mj. rozšiřováním početné a diverzifikované populace REP, popsané výše.

Stenotrophomonas jsou bohaté na inserční sekvence, především rodiny IS*3*, IS*110* a IS*481* jsou v genomech hojně rozšířeny. *De novo* bylo identifikováno 1086 kopií inserčních sekvencí. IS byly následně klasifikovány dle sekvence kódovaných transponáz (databáze ISfinder) a rozděleny na 35 jedinečných zástupců. Byly určeny jejich nekódující okraje, inzerční místa a případné duplikace produkované během transpozice. Z dat byla provedena fylogenetická analýza jednotlivých rodin IS, zaměřená především na analýzu evoluce transpozičních cílů.

Nejvíce IS náleží rodině IS*3*, z 15 nalezených zástupců jsme jich 8 identifikovali v této práci *de novo*. Pouze jedna IS*3* cílí svou transpozicí do REP elementů, konkrétně IS*StmaNEW10* cílí do REP 02B. Systém RAYT/REP 02B je aktivní právě u jediného kmene, kde je přítomno 6 kopií

této IS. Jsou popsáni již dva zástupci IS3 inzerující do REP – IS1397 a IS*Kpn1.* Ty se vyskytují u *E. coli* a *K. pneumoniae*, respektive (Wilde et al., 2003). Všechny tři IS cílí do nukleotidů tvořící smyčku vlásenky, ta je při inzerci duplikována. Další tři IS mají také specifické cíle transpozice. IS*StmaNEW7*, *13* a IS*Stma2* transponují do YPAL, což jsou mimo rod *Yersinia* zatím nepopsané DNA motivy (více níže). Tyto tři IS celkovým počtem kopií značně převyšují ostatní IS*3*.

Rodina IS*110* má nejdynamičtější evoluci inzerčních cílů, všech 8 identifikovaných členů cílí do specifické DNA sekvence. Všechny při inzerci duplikují CT dinukleotid, ovšem zbytek cíle je variabilní. Již popsaná IS*110* cílící do REP je IS*621* z *E. coli* (Choi et al., 2003). U *Stenotrophomonas* tři IS*110* transponují do REP 05, dvě konkrétně do GTAG a jedna do ramene palindromu. Další tři IS*110* pak transponují do REP 02A, 04G a 07. Zvláštností je IS*StmaNEW1*, neboť ta je schopna inzerce do REP 07 a 08 (do toho navíc v obou orientacích).

7 z 8 IS*110* cílí do REP, výjimkou je IS*Stma7*. Ta transponuje do YPAL elementů, IS*Stma7* je nejčetnější IS*110* rozšířenou napříč *Stenotrophomonas*. U *Yersinia* je znám zástupce IS*110* cílící také do YPAL, konkrétně IS*Yen1* (De Gregorio et al., 2006). IS*Stma7* sice necílí do REP, ale je s nimi přesto asociována, neboť v pravém nekódujícím okraji nese konzervovaný REP 03 umístěný 27 bází od konce čtecího rámce, REP je více konzervovaný než zbytek okraje IS*Stma7*. I Přítomnost REP může mít pozitivní vliv na regulaci genů v okolí inzerce a tím vést k šíření/udržení IS. Alternativně se může RAYT jako DNA nukleáza vázat na REP vedle konce kódující oblasti IS a neznámým způsobem, například prostřednictvím interakce s transponázou, ovlivnit transpozici IS*Stma7*.

Z rodiny IS*481* byly identifikováni dva zástupci, námi definovaný IS*StacX* a již známý IS*Stma1*. Oba mají specifické inzerční cíle. IS*StacX* cílí na krátké hexanukleotidy NCATGN a vyskytuje se u bazálních kmenů. IS*Stma1* transponuje od REP 04A a jde o nejrozšířenější IS, jakou jsme v práci identifikovali. Důvodem je hojné rozšíření u *S. maltophilia sensu stricto* (skupina Sm6), ze které je dnes plně sekvenováno nejvíce kmenů. IS*Stma1* využívá nukleotidy REP zároveň jako vlastní STOP kodon. V tomto ohledu se zdá evoluce inzerce této IS limitována. Pouze REP 04A a F mají potenciální STOP kodony na obou komplementárních vláknech DNA současně. Tyto dva REP se tak zdají jako jediné potenciální cíle IS*Stma1*. REP 04F je navíc minoritní, aktivní RAYT má pouze jeden kmen a zde je pouze 33 kopií tohoto REP. Poslední IS se specifickým cílem je námi popsaný zástupce IS*4*, který transponuje do TIR elementů (Di Nocera et al., 2013).

Překvapivým výstupem práce je diverzita inserčních sekvencí se specifickým cílem transpozice a jejich dominance v rámci všech IS u *Stenotrophomonas*. Tento jev nebyl zatím v literatuře popsán. Množství kopií je pro inzerční sekvenci klíčová vlastnost – jelikož jde o sobecké elementy, nevzniká při poškození selekční tlak na opravu a pravděpodobně tak dojde k zániku IS. Platí tak úměra, kdy více kopií znamená vyšší vitalitu IS (Bichsel et al., 2013; Doolittle & Sapienza, 1980). Proliferace IS je u bakterií redukována celou řadou mechanismů, příkladem může být frameshift kódující sekvence, antisenseRNA se silnými promotory, slabá exprese

transponázy, nestabilní mRNA či rychlá degradace proteinu transpozázy. Sobecká DNA spotřebovává zdroje na svou replikaci a zpomaluje replikaci chromozomu. Nespecifická transpozice potenciálně poškozuje geny: typická kódující denzita bakteriálního genomu je 80-90% a tedy velká většina inzercí takových IS inaktivuje hostitelské geny – regulace transpozice je tak v „životním zájmu“ buňky. Specifická transpozice do intergenových DNA motivů možnost inzerční inaktivace genů redukuje až eliminuje. Například REP nejsou esenciální, jsou sekvenčně konzervované a je jich v genomu přítomno velké množství. Často je v intergenové oblasti přítomna řada REP, mají-li zde regulační funkci, inzerce do jednoho ji nemusí vždy ovlivnit. Podobná situace je u IS cílících do YPAL, ty jsou dominantní stejně jako IS cílící do REP. YPAL jsou opět intergenové konzervované motivy a také se u *Stenotrophomonas* typicky vyskytují v dubletech, kdy opět inzerce do jednoho z nich nemusí teoreticky ovlivnit jejich funkci v daném intergenovém lokusu. I pokud dojde vlivem inzerce IS do intergenových elementů k redukci/eliminaci jejich regulační funkce, výsledný vliv na fitness hostitelské buňky bude výrazně menší než inzerční inaktivace genů.

Celkem jsme identifikovali deset IS transponujících do REP, čtyři do YPAL a jednu do TIR. Celkem je těchto 15 IS (z 35 unikátních IS) rozšířeno napříč genomy v 808 kopiích (z 1086 všech IS). Žádná IS necílí do nejčetnější skupiny REP (REP 03), ani do REP 01, které, byť méně početné, jsou všudypřítomné. Pouze IS*Stma06* cílí na velkou skupinu REP 02A, které jsou rozšířeny skrze celou diverzitu *Stenotrophomonas* (stejně jako tato IS). Místo nejpočetnějších a nejrozšířenějších REP jsou preferované méně četné elementy – REP 05, 07 a 08, pro které je typická přítomnost v jedné či několika skupinách *Stenotrophomonas* a minimum kopií ve zbytku rodu. Některé IS dokonce cílí do kmenově unikátních REP – 02B a 04G. Obdobná situace je při inzerci do YPAL, tyto motivy jsou přítomné téměř ve všech genomech ve stálém počtu kolem sta kopií. IS, které do nich transponují, jsou tak rozšířeny napříč rodem.

Při analýze inserčních cílů IS*3* jsme identifikovali neznámý DNA motiv, který byl posléze identifikován jako YPAL – extragenový palindromický motiv známý u rodu *Yersinia* (Bachellier et al., 1999; De Gregorio et al., 2006). Motivy nalezené u *Stenotrophomonas* vytvářejí téměř identické sekundární struktury a sdílejí sekvenčně podobnou střední oblast. Oba motivy jsou extragenové a jejich počty dosahují asi sta kopií na genom. U obou rodů jsou tyto sekvence cílem transpozice IS. Řada charakteristik se ale liší – analýza naznačuje, že sekvence nejsou u *Stenotrophomonas* mobilní. Oproti tomu u *Yersinia pestis* jsme potvrdili recentní mobilitu YPAL (De Gregorio et al., 2006). S tím může souviset sekvenční homologie – u *Yersinia* jsou motivy velmi konzervované. V porovnání se *Stenotrophomonas* je evidentní selekce na zachování sekundární struktury. Okolí YPAL je bohaté na REP, zvláště častý je zde REP 02A (téměř polovina REP). U *Yersinia* bylo dále experimentálně prokázáno, že YPAL elementy jsou transkribovány do mRNA, kterou díky složitým sekundárním strukturám stabilizují. Jde tak o potenciální funkci i u *Stenotrophomonas*. Krom YPAL jsme nalezli dva kratší motivy – Inz1 a Inz2, které se specificky

inzertují do YPAL u *Stenotrophomonas* (u všech ostatních sekvenovaných bakterií chybí). Oba motivy jsou mobilní a jejich počty i pozice se tak mezi kmeny značně liší. Vykazují vysokou sekvenční konzervaci a vytvářejí vlásenkovité sekundární struktury. V literatuře tyto motivy nalezeny nebyly. Nejsou známé proteiny zodpovědné za mobilizaci YPAL, inzertu1 ani inzertu2.

Naše práce je první, která YPAL motivy nalezla a popsala mimo rod *Yersinia*. Krom *Stenotrophomonas* jsme YPAL identifikovali také u *Xanthomonas* a několika druhů čeledi *Alteromonadaceae*. Identifikace YPAL v dalších bakteriálních genomech s využitím specifičtějších nástrojů – například se zaměřením na hledání konzervované struktury (nikoliv sekvence), jsou potenciální cestou dalšího výzkumu. Budoucí výzkum YPAL by se poté mohl věnovat komparativní analýze motivů mezi bakteriálními druhy. Krom konzervovanosti sekvence a struktury, zůstává otázkou celková diverzita inserčních sekvencí cílících do YPAL. Dále také zda existují krátké elementy analogické Inz1/Inz2 a inzertující do YPAL i u jiných druhů bakterií. Klíčovou neznámou zůstává mechanismus mobilizace YPAL – analýza genomických dat může poskytnout cenné indicie k identifikaci genu „mobilizátora“ (konzervovaná genetická asociace s YPAL, korelace s počtem/mobilitou YPAL).

Výstupem naší analýzy jsou čtyři předpoklady, které by měly formovat „evolučně-funkční krajinu“ systémů RAYT/REP u *Stenotrophomonas*: i. buňka je pod silnou selekcí na udržení vysoké kódující hustoty, ii. REP diverzita se odehrává primárně v rámci omezeného množství lokusů, iii. RAYT přepisují stávající REP, iv. hlavní funkcí RAYT je šíření REP asociovaných s jejich geny. Výsledkem by měl být intenzivní „boj“ o REP lokusy doprovázený frekventovaným přepisem REP. Tento stav je zřejmý jak při analýze vzdáleně příbuzných kmenů, tak v rámci jedné skupiny. Více systémů v genomu se zdá být redundantními a vlivem kompetice by mohl zůstat jediný dominantní systém – pravděpodobně RAYT/REP 03, jehož asociované REP jsou napříč diverzitou rodu nejčetnější. K tomu částečně dochází, intenzivní kompetice na úrovni REP lokusů se promítá do redukce počtu RAYT u pokročilých skupin (tato práce a (Park et al., 2021)). Ke ztrátě jednotlivých RAYT dochází, nedochází však ke konkurenčnímu vyloučení a dominanci jediného systému. Extrémním příkladem je již zmíněný kmen ZAC14A_NAIMI4_1 s osmi potenciálně funkčními RAYT, v jehož genomu je rozšířeno mezi 100 až 400 kopií každého s RAYT asociovaného REP elementu (celkem 1914 kopií). Jde o jeden z nejvyšší zaznamenaný počet REP v genomu v rámci studovaných kmenů – výjimečnost takto vysoké koncentrace systémů REP/RAYT naznačuje, že se nejedná o optimální stav z hlediska dlouhodobé evoluční stability.

Zdá se, že primární funkcí RAYT/REP není *de novo* šíření REP do nových oblastí, ale diverzifikace REP v rámci existujících REP lokusů. Šíření REP do nových oblastí probíhá, ale ne frekventovaně. Vlastnosti REP různých typů v lokusech jsou obdobné – všechny jsou běžně součástí REPIN/BIME. Příčiny a funkční důsledky neobyčejné evoluční dynamiky těchto lokusů a důvody jejich kolonizace REP elementy jsou otázky pro budoucí výzkum.

7. Souhrn

1. Na základě kompletních sekvencí 102 genomů *Stenotrophomonas* byly určeny jejich fylogenetické vztahy. Kmeny byly rozděleny do již definovaných taxonomických skupin.
2. Celkem bylo identifikováno 271 funkčních *rayt* (331 včetně pseudogenizovaných). Geny byly rozděleny do 11 skupin na základě lokusu, ve kterém se nachází. Fylogenetická analýza potvrdila monofylii v rámci RAYT kódovaných konkrétním lokusem.
3. Pomocí AlphaFold2 byl připraven model vybraných RAYT proteinů *Stenotrophomonas*. Struktury byly porovnány s experimentálně získaným modelem RAYT z *E. coli*. Proteiny vykazují téměř identickou prostorovou konfiguraci katalytické oblasti, která souvisí s absolutní konzervací příslušných aminokyselin.
4. Analýzou okolí *rayt* byly identifikovány asociované sekvence REP elementů. Asociace je evolučně stabilní, ve třech lokusech však došlo k diferenciaci REP elementů u různých linií *Stenotrophomonas*.
5. Kvantifikovali jsme, že zdaleka nejpočetnější jsou REP v genomech, které mají funkční asociovaný RAYT. Méně četné jsou REP v přítomnosti RAYT pseudogenizovaného, nebo když byl RAYT ztracen. Minimální počet REP je v genomech, kde asociovaný RAYT dlouhodobě chybí.
6. REP elementy asociované s RAYT jsou přítomné ve vysokých počtech kopií u všech kmenů s výjimkou podskupiny Ib, v celkovém počtu 63 689 přesně konzervovaných kopií (132 039 REP s maximálně třemi sekvenčními odchylkami). Zdaleka nejpočetnější je REP 03 vyskytující se ve stovkách kopií téměř ve všech genomech. Méně početná, zato všudypřítomná je skupina REP 01. Zbytek REP se vyskytuje spíše v rámci jedné či několika skupin genomů.
7. V globálním rozložení jsou REP vzácné na mobilních elementech, velmi konzervovaných regionech (rDNA, tRNA…) a ve velké oblasti v okolí terminace replikace. Ta je kromě absence REP charakterizovaná rozpadem syntenie genů. Přes 90% REP je přítomno v rámci necelých 800 REP lokusů. V průběhu evoluce *Stenotrophomonas* dochází v rámci lokusu často k nahrazování jinými typy REP.
8. Bylo identifikováno 35 unikátních inserčních sekvencí, 15 z nich má specifický inserční cíl. V počtu kopií však tyto IS dominují a tvoří téměř 75% všech IS. Jako inzerční cíle slouží několik typů REP (REP 02A a B, 04A a G, 05, 07 a 08) a nově objevená repetitivní sekvence YPAL.

Popsali jsme neznámé sekvenční motivy vyskytující se v intergenových oblastech *Stenotrophomonas*. které jsme posléze určili jako homology elementů YPAL, doposud popsaných jen u rodu *Yersinia*. Do YPAL se kromě řady IS specificky inzertují i dva krátké mobilní DNA motivy tvořící extenzivní sekundární struktury. Mechanismus jejich mobility, jakož i mobility YPAL, je neznámý.

8. Reference

Ahmed, A., & Scraba, D. (1975). The nature of the *gal3* mutation of *Escherichia coli*. *Molecular & General Genetics : MGG*, *136*(3), 233–242. https://doi.org/10.1007/BF00334018

AlphaFold.ipynb - Colaboratory. (n.d.). Retrieved July 27, 2023, from https://colab.research.google.com/github/deepmind/alphafold/blob/main/notebooks/AlphaFold.ipynb#scrollTo=VzJ5iMjTtoZw

AM, R., J, M., A, A., C, F., H, B., S, S., C, B., G, J., & JP, G. (2001). Electronic ventilator temperature sensors as a potential source of respiratory tract colonization with *Stenotrophomonas maltophilia*. *The Journal of Hospital Infection*, *49*(4), 289–292. https://doi.org/10.1053/JHIN.2001.1099

Aranda-Olmedo, I., Tobes, R., Manzanera, M., Ramos, J. L., & Marqués, S. (2002). Species-specific repetitive extragenic palindromic (REP) sequences in *Pseudomonas putida*. *Nucleic Acids Research*, *30*(8). https://doi.org/10.1093/nar/30.8.1826

Assih, E. A., Ouattara, A. S., Thierry, S., Cayol, J. L., Labat, M., & Macarie, H. (2002). *Stenotrophomonas acidaminiphila sp*. nov., a strictly aerobic bacterium isolated from an upflow anaerobic sludge blanket (UASB) reactor. *International Journal of Systematic and Evolutionary Microbiology*, *52*(2), 559–568. https://doi.org/10.1099/00207713-52-2-559/CITE/REFWORKS

Aussel, L., Barre, F. X., Aroyo, M., Stasiak, A., Stasiak, A. Z., & Sherratt, D. (2002). FtsK Is a DNA Motor Protein that Activates Chromosome Dimer Resolution by Switching the Catalytic State of the XerC and XerD Recombinases. *Cell*, *108*(2), 195–205. https://doi.org/10.1016/S0092-8674(02)00624-4

Bachellier, S., Clément, J. M., & Hofnung, M. (1999). Short palindromic repetitive DNA elements in enterobacteria: a survey. *Research in Microbiology*, *150*(9–10), 627–639. https://doi.org/10.1016/S0923-2508(99)00128-X

Barabas, O., Ronning, D. R., Guynet, C., Hickman, A. B., Ton-Hoang, B., Chandler, M., & Dyda, F. (2008). Mechanism of IS*200*/IS*605* Family DNA Transposases: Activation and Transposon-Directed Target Site Selection. *Cell*, *132*(2). https://doi.org/10.1016/j.cell.2007.12.029

Bartlett, D. H., & Silverman, M. (1989). Nucleotide sequence of IS*492*, a novel insertion sequence causing variation in extracellular polysaccharide production in the marine bacterium *Pseudomonas atlantica*. *Journal of Bacteriology*, *171*(3), 1763. https://doi.org/10.1128/JB.171.3.1763-1766.1989

Baut, G. Le, O'brien, C., Pavli, P., Roy, M., Seksik, P., Tréton, X., Nancey, S., Barnich, N., Bezault, M., Auzolle, C., Cazals-Hatem, D., Viala, J., Allez, M., Hugot, J. P., & Dumay, A. (2018). Prevalence of *Yersinia* Species in the Ileum of Crohn's Disease Patients and Controls. *Frontiers in Cellular and Infection Microbiology*, *8*(SEP), 336. https://doi.org/10.3389/FCIMB.2018.00336/BIBTEX

Becerril, B., Valle, F., Merino, E., Riba, L., & Bolivar, F. (1985). Repetitive extragenic palindromic (REP) sequences in the *Escherichia coli gdhA* gene. *Gene*, *37*(1–3), 53–62. https://doi.org/10.1016/0378-1119(85)90257-4

Berg, G., Eberl, L., & Hartmann, A. (2005). The rhizosphere as a reservoir for opportunistic human pathogenic bacteria. *Environmental Microbiology*, *7*(11), 1673–1685. https://doi.org/10.1111/J.1462-2920.2005.00891.X

Bertani, G., & Six, E. (1958). Inheritance of prophage P2 in bacterial crosses. *Virology*, *6*(2), 357–381. https://doi.org/10.1016/0042-6822(58)90089-8

Bertelli, C., Gray, K. L., Woods, N., Lim, A. C., Tilley, K. E., Winsor, G. L., Hoad, G. R., Roudgar, A., Spencer, A., Peltier, J., Warren, D., Raphenya, A. R., McArthur, A. G., & Brinkman, F. S. L. (2022). Enabling genomic island prediction and comparison in multiple genomes to investigate bacterial evolution and outbreaks. *Microbial Genomics*, *8*(5), 000818. https://doi.org/10.1099/MGEN.0.000818

Bertelli, C., Laird, M. R., Williams, K. P., Lau, B. Y., Hoad, G., Winsor, G. L., & Brinkman, F. S. L. (2017). IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. *Nucleic Acids Research*, *45*(W1), W30–W35. https://doi.org/10.1093/NAR/GKX343

Bertels, F., Gallie, J., & Rainey, P. B. P. B. (2017). Identification and characterization of domesticated bacterial transposases. *Genome Biology and Evolution*, *9*(8), 2110–2121. https://doi.org/10.1093/gbe/evx146

Bertels, F., Gokhale, C. S., & Traulsen, A. (2017). Discovering complete quasispecies in bacterial genomes. *Genetics*. https://doi.org/10.1534/genetics.117.201160

Bertels, F., & Rainey, P. B. (2011a). Curiosities of REPINs and RAYTs. *Mobile Genetic Elements*, *1*(4). https://doi.org/10.4161/mge.18610

Bertels, F., & Rainey, P. B. (2011b). Within-Genome evolution of REPINs: A new family of miniature mobile DNA in bacteria. *PLoS Genetics*, *7*(6). https://doi.org/10.1371/journal.pgen.1002132

Bhagwat, A. S., Hao, W., Townes, J. P., Lee, H., Tang, H., & Foster, P. L. (2016). Strand-biased cytosine deamination at the replication fork causes cytosine to thymine mutations in *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(8), 2176–2181. https://doi.org/10.1073/PNAS.1522325113

Bhowmik, B. K., Clevenger, A. L., Zhao, H., & Rybenkov, V. V. (2018). Segregation but Not Replication of the *Pseudomonas aeruginosa* Chromosome Terminates at *Dif*. *MBio*, *9*(5), 1–13. https://doi.org/10.1128/MBIO.01088-18

Bichsel, M., Barbour, A. D., & Wagner, A. (2013). Estimating the fitness effect of an insertion sequence. *Journal of Mathematical Biology*, *66*(1–2), 95–114. https://doi.org/10.1007/S00285-012-0504-2

Bienert, S., Waterhouse, A., De Beer, T. A. P., Tauriello, G., Studer, G., Bordoli, L., & Schwede, T. (2017). The SWISS-MODEL Repository—new features and functionality. *Nucleic Acids Research*, *45*(D1), D313–D319. https://doi.org/10.1093/NAR/GKW1132

Bigot, S., Saleh, O. A., Lesterlin, C., Pages, C., El Karoui, M., Dennis, C., Grigoriev, M., Allemand, J. F., Barre, F. X., & Cornet, F. (2005). KOPS: DNA motifs that control *E. coli* chromosome segregation by orienting the FtsK translocase. *The EMBO Journal*, *24*(21), 3770. https://doi.org/10.1038/SJ.EMBOJ.7600835

Blakely, G., May, G., McCulloch, R., Arciszewska, L. K., Burke, M., Lovett, S. T., & Sherratt, D. J. (1993). Two related recombinases are required for site-specific recombination at *dif* and *cer* in *E. coli* K12. *Cell*, *75*(2), 351–361. https://doi.org/10.1016/0092-8674(93)80076-Q

Blakely, G., & Sherratt, D. (1994). Interactions of the site-specific recombinases XerC and XerD with the recombination site *dif*. *Nucleic Acids Res*, *22*.

Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B., & Shao, Y. (1997). The Complete Genome Sequence of *Escherichia coli* K-12. *Science*, *277*(5331), 1453–1462. https://doi.org/10.1126/SCIENCE.277.5331.1453

Boccard, F., & Prentki, P. (1993). Specific interaction of IHF with RIBs, a class of bacterial repetitive DNA elements located at the 3' end of transcription units. *EMBO Journal*, *12*(13), 5019–5027. https://doi.org/10.1002/j.1460-2075.1993.tb06195.x

Brázda, V., Kolomazník, J., Lýsek, J., Hároníková, L., Coufal, J., & Šťastný, J. (2016). Palindrome analyser – A new web-based server for predicting and evaluating inverted repeats in nucleotide sequences. *Biochemical and Biophysical Research Communications*, *478*(4), 1739–1745. https://doi.org/10.1016/J.BBRC.2016.09.015

Brooke, J. S. (2012). *Stenotrophomonas maltophilia*: an Emerging Global Opportunistic Pathogen. *Clinical Microbiology Reviews*, *25*(1), 2. https://doi.org/10.1128/CMR.00019-11

Buchner, J. M., Robertson, A. E., Poynter, D. J., Denniston, S. S., & Karls, A. C. (2005). *Piv* site-specific invertase requires a DEDD motif analogous to the catalytic center of the RuvC Holliday junction resolvases. *Journal of Bacteriology*, *187*(10), 3431–3437. https://doi.org/10.1128/JB.187.10.3431-3437.2005

Carnoy, C., & Roten, C. A. (2009). The *dif*/Xer recombination systems in proteobacteria. *PLoS ONE*, *4*(9). https://doi.org/10.1371/JOURNAL.PONE.0006531

Carraro, N., Sentchilo, V., Polák, L., Bertelli, C., & van der Meer, J. R. (2020). Insights into Mobile Genetic Elements of the Biocide-Degrading Bacterium *Pseudomonas nitroreducens* HBP-1. *Genes*, *11*(8), 1–19. https://doi.org/10.3390/GENES11080930

Censini, S., Lange, C., Xiang, Z., Crabtree, J. E., Ghiara, P., Borodovsky, M., Rappuoli, R., & Covacci, A. (1996). cag, a pathogenicity island of *Helicobacter pylori*, encodes type I-specific and disease-associated virulence factors. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(25), 14648–14653. https://doi.org/10.1073/PNAS.93.25.14648/ASSET/63AA5537-94DB-4242-9022-5973EA71F047/ASSETS/GRAPHIC/PQ2563032004.JPEG

Chan, H., Mohamed, A. M. T., Grainge, I., & Rodrigues, C. D. A. (2022). FtsK and SpoIIIE, coordinators of chromosome segregation and envelope remodeling in bacteria. *Trends in Microbiology*, *30*(5), 480–494. https://doi.org/10.1016/J.TIM.2021.10.002

Chandler, M., Fayet, O., Rousseau, P., Ton Hoang, B., & Duval-Valentin, G. (2015). Copy-out–Paste-in Transposition of IS*911*: A Major Transposition Pathway. *Microbiology Spectrum*, *3*(4). https://doi.org/10.1128/MICROBIOLSPEC.MDNA3-0031-2014/ASSET/61B768C8-C477-4791-BC06-35B08A52943F/ASSETS/GRAPHIC/MDNA3-0031-2014-FIG5.GIF

Charlier, D., Gigot, D., Huysveld, N., Roovers, M., Piérard, A., & Glansdorff, N. (1995). Pyrimidine regulation of the *Escherichia coli* and *Salmonella typhimurium* carab operons: CarP and integration host factor (IHF) modulate the methylation status of a GATC site present in the control region. In *Journal of Molecular Biology* (Vol. 250, Issue 4). https://doi.org/10.1006/jmbi.1995.0384

Chater, K. F., Bruton, C. J., Foster, S. G., & Tobek, I. (1985). Physical and genetic analysis of IS*110*, a transposable element of *Streptomyces coelicolor* A3(2). *Molecular & General Genetics: MGG*, *200*(2), 235–239. https://doi.org/10.1007/BF00425429

Chen, R. L. W., & Fong, P. (1969). Helical configuration of single-stranded DNA. *Journal of Theoretical Biology*, *22*(1), 180–187. https://doi.org/10.1016/0022-5193(69)90086-1

Cheng, Z. F., & Deutscher, M. P. (2005). An important role for RNase R in mRNA decay. *Molecular Cell*, *17*(2). https://doi.org/10.1016/j.molcel.2004.11.048

Choi, S., Ohta, S., & Ohtsubo, E. (2003). A Novel IS Element, IS*621*, of the IS*110*/IS*492* Family Transposes to a Specific Site in Repetitive Extragenic Palindromic Sequences in *Escherichia coli*. *Journal of Bacteriology*, *185*(16), 4891. https://doi.org/10.1128/JB.185.16.4891-4900.2003

Clément, J. M., Wilde, C., Bachellier, S., Lambert, P., & Hofnung, M. (1999). IS*1397* is active for transposition into the chromosome of *Escherichia coli* K-12 and inserts specifically into palindromic units of bacterial interspersed mosaic elements. *Journal of Bacteriology*, *181*(22), 6929–6936. https://doi.org/10.1128/JB.181.22.6929-6936.1999

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & De Hoon, M. J. L. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11). https://doi.org/10.1093/bioinformatics/btp163

Coenye, T., Vanlaere, E., LiPuma, J. J., & Vandamme, P. (2004). Identification of genomic groups in the genus *Stenotrophomonas* using *gyrB* RFLP analysis. *FEMS Immunology & Medical Microbiology*, *40*(3), 181–185. https://doi.org/10.1016/S0928-8244(03)00307-9

Cornet, F., Louarn, J., Patte, J., & Louarn, J. M. (1996). Restriction of the activity of the recombination site *dif* to a small zone of the *Escherichia coli* chromosome. *Genes & Development*, *10*(9), 1152–1161. https://doi.org/10.1101/GAD.10.9.1152

Cortez, D., Quevillon-Cheruel, S., Gribaldo, S., Desnoues, N., Sezonov, G., Forterre, P., & Serre, M. C. (2010). Evidence for a Xer/*dif* System for Chromosome Resolution in Archaea. *PLOS Genetics*, *6*(10), e1001166. https://doi.org/10.1371/JOURNAL.PGEN.1001166

Crossman, L. C., Gould, V. C., Dow, J. M., Vernikos, G. S., Okazaki, A., Sebaihia, M., Saunders, D., Arrowsmith, C., Carver, T., Peters, N., Adlem, E., Kerhornou, A., Lord, A., Murphy, L., Seeger, K., Squares, R., Rutter, S., Quail, M. A., Rajandream, M. A., … Avison, M. B. (2008). The complete genome, comparative and functional analysis of *Stenotrophomonas maltophilia* reveals an organism

heavily shielded by drug resistance determinants. *Genome Biology*, *9*(4). https://doi.org/10.1186/GB-2008-9-4-R74

Darling, A. C. E., Mau, B., Blattner, F. R., & Perna, N. T. (2004). Mauve: Multiple Alignment of Conserved Genomic Sequence With Rearrangements. *Genome Research*, *14*(7), 1394. https://doi.org/10.1101/GR.2289704

Davenport, K. W., Daligault, H. E., Minogue, T. D., Broomall, S. M., Bruce, D. C., Chain, P. S., Coyne, S. R., Gibbons, H. S., Jaissle, J., Li, P. E., Rosenzweig, C. N., Scholz, M. B., & Johnson, S. L. (2014). Complete Genome Sequence of *Stenotrophomonas maltophilia* Type Strain 810-2 (ATCC 13637). *Genome Announcements*, *2*(5), 974–988. https://doi.org/10.1128/GENOMEA.00974-14

De Gregorio, E., Silvestro, G., Venditti, R., Carlomagno, M. S., & Di Nocera, P. P. (2006). Structural organization and functional properties of miniature DNA insertion sequences in *Yersiniae*. *Journal of Bacteriology*, *188*(22), 7876–7884. https://doi.org/10.1128/JB.00942-06

Delihas, N. (2007). Enterobacterial Small Mobile Sequences Carry Open Reading Frames and are Found Intragenically – Evolutionary Implications for Formation of New Peptides. *Gene Regulation and Systems Biology*, *1*, 117762500700100. https://doi.org/10.1177/117762500700100017/ASSET/IMAGES/LARGE/10.1177_117762500700100017-FIG7.JPEG

Deng, C., Lv, X., Li, J., Liu, Y., Du, G., Amaro, R. L., & Liu, L. (2019). Synthetic repetitive extragenic palindromic (REP) sequence as an efficient mRNA stabilizer for protein production and metabolic engineering in prokaryotic cells. *Biotechnology and Bioengineering*, *116*(1). https://doi.org/10.1002/bit.26841

Di Nocera, P. P., De Gregorio, E., & Rocco, F. (2013). GTAG- and CGTC-tagged palindromic DNA repeats in prokaryotes. *BMC Genomics*, *14*(1). https://doi.org/10.1186/1471-2164-14-522

Doolittle, W. F., & Sapienza, C. (1980). Selfish genes, the phenotype paradigm and genome evolution. *Nature*, *284*(5757), 601–603. https://doi.org/10.1038/284601A0

Download RStudio | The Popular Open-Source IDE from Posit. (n.d.). Retrieved March 20, 2023, from https://posit.co/products/open-source/rstudio/

Draw Freely | Inkscape. (n.d.). Retrieved July 15, 2023, from https://inkscape.org/cs/

Duval-Valentin, G., & Chandler, M. (2011). Cotranslational control of DNA transposition: a window of opportunity. *Molecular Cell*, *44*(6), 989–996. https://doi.org/10.1016/J.MOLCEL.2011.09.027

Duval-Valentin, G., Normand, C., Khemici, V., Marty, B., & Chandler, M. (2001). Transient promoter formation: a new feedback mechanism for regulation of IS*911* transposition. *The EMBO Journal*, *20*(20), 5802–5811. https://doi.org/10.1093/EMBOJ/20.20.5802

Eichhorn, C. D., & Al-Hashimi, H. M. (2014). Structural dynamics of a single-stranded RNA–helix junction using NMR. *RNA*, *20*(6), 782. https://doi.org/10.1261/RNA.043711.113

Espéli, O., & Boccard, F. (1997). *In vivo* cleavage of *Escherichia coli* BIME-2 repeats by DNA gyrase: Genetic characterization of the target and identification of the cut site. *Molecular Microbiology*. https://doi.org/10.1046/j.1365-2958.1997.6121983.x

Espéli, O., Moulin, L., & Boccard, F. (2001). Transcription attenuation associated with bacterial repetitive extragenic BIME elements. *Journal of Molecular Biology*, *314*(3). https://doi.org/10.1006/jmbi.2001.5150

Farabaugh, P. J. (1996). Programmed translational frameshifting. *Microbiological Reviews*, *60*(1), 103. https://doi.org/10.1128/MR.60.1.103-134.1996

Farnham, P. J., & Platt, T. (1981). Rho-independent termination: dyad symmetry in DNA causes RNA polymerase to pause during transcription *in vitro*. *Nucleic Acids Research*, *9*(3), 563–577. https://doi.org/10.1093/NAR/9.3.563

Fastani:: Anaconda.org. (n.d.). Retrieved June 14, 2023, from https://anaconda.org/bioconda/fastani

Fayet, O., Ramond, P., Polard, P., Prère, M. F., & Chandler, M. (1990). Functional similarities between retroviruses and the IS*3* family of bacterial insertion sequences? *Molecular Microbiology*, *4*(10), 1771–1777. https://doi.org/10.1111/J.1365-2958.1990.TB00555.X

Finkmann, W., Altendorf, K., Stackebrandt, E., & Lipski, A. (2000). Characterization of N2O-producing *Xanthomonas*-like isolates from biofilters as *Stenotrophomonas nitritireducens sp*. nov., *Luteimonas mephitis g*en. nov., sp. nov. and *Pseudoxanthomonas broegbernensis* gen. nov., sp. nov. *International Journal of Systematic and Evolutionary Microbiology*, *50*(1), 273–282. https://doi.org/10.1099/00207713-50-1-273/CITE/REFWORKS

Fozo, E. M., Makarova, K. S., Shabalina, S. A., Yutin, N., Koonin, E. V., & Storz, G. (2010). Abundance of type I toxin–antitoxin systems in bacteria: searches for new candidates and discovery of novel families. *Nucleic Acids Research*, *38*(11), 3743. https://doi.org/10.1093/NAR/GKQ054

Franke, A. E., & Clewell, D. B. (1981). Evidence for a chromosome-borne resistance transposon (Tn*916*) in *Streptococcus faecalis* that is capable of "conjugal" transfer in the absence of a conjugative plasmid. *Journal of Bacteriology*, *145*(1), 494. https://doi.org/10.1128/JB.145.1.494-502.1981

Fujita, J., Yamadori~, I., Xu, G., Hojo, S., Negayama, K., Miyawaki, H., Yamaji, Y., & Takahara, J. (1996). Clinical features of *Stenotrophomonas maltophilia* pneumonia in immunocompromised patients. *Respiratory Medicine*, *90*, 35–38.

Gellert, M., Mizuuchi, K., O'Dea, M. H., & Nash, H. A. (1976). DNA gyrase: an enzyme that introduces superhelical turns into DNA. *Proceedings of the National Academy of Sciences of the United States of America*, *73*(11). https://doi.org/10.1073/pnas.73.11.3872

Geneious | Bioinformatics Software for Sequence Data Analysis. (n.d.). Retrieved March 23, 2023, from https://www.geneious.com/

Gilson, E., Clément, J. M., Brutlag, D., & Hofnung, M. (1984). A family of dispersed repetitive extragenic palindromic DNA sequences in *E. coli*. *The EMBO Journal*, *3*(6). https://doi.org/10.1002/j.1460-2075.1984.tb01986.x

Gilson, E., Perrin, D., & Hofnung, M. (1990). DNA polymerase I and a protein complex bind specifically to *E. coli* palindromic unit highly repetitive DNA: Implications for bacterial chromosome organization. *Nucleic Acids Research*, *18*(13). https://doi.org/10.1093/nar/18.13.3941

Gilson, E., Saurin, W., Perrin, D., Bachellier, S., & Hofnung, M. (1991). The BIME family of bacterial highly repetitive sequences. *Research in Microbiology*, *142*(2–3). https://doi.org/10.1016/0923-2508(91)90033-7

Gómez-Rubio, V. (2017). ggplot2 - Elegant Graphics for Data Analysis (2nd Edition). *Journal of Statistical Software*, *77*(Book Review 2). https://doi.org/10.18637/jss.v077.b02

Google Colab. (n.d.). Retrieved March 20, 2023, from https://research.google.com/colaboratory/faq.html

Goosen, N., & van de Putte, P. (1995). The regulation of transcription initiation by integration host factor. In *Molecular Microbiology* (Vol. 16, Issue 1). https://doi.org/10.1111/j.1365-2958.1995.tb02386.x

Gröschel, M. I., Meehan, C. J., Barilar, I., Diricks, M., Gonzaga, A., Steglich, M., Conchillo-Solé, O., Scherer, I. C., Mamat, U., Luz, C. F., De Bruyne, K., Utpatel, C., Yero, D., Gibert, I., Daura, X., Kampmeier, S., Rahman, N. A., Kresken, M., van der Werf, T. S., … Kohl, T. A. (2020). The phylogenetic landscape and nosocomial spread of the multidrug-resistant opportunist *Stenotrophomonas maltophilia*. *Nature Communications 2020 11:1*, *11*(1), 1–12. https://doi.org/10.1038/s41467-020-15123-0

Gu, Z., Eils, R., & Schlesner, M. (2016). Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *Bioinformatics*, *32*(18), 2847–2849. https://doi.org/10.1093/BIOINFORMATICS/BTW313

Gu, Z., Gu, L., Eils, R., Schlesner, M., & Brors, B. (2014). circlize Implements and enhances circular visualization in R. *Bioinformatics (Oxford, England)*, *30*(19), 2811–2812. https://doi.org/10.1093/BIOINFORMATICS/BTU393

Hacker, J., Bender, L., Ott, M., Wingender, J., Lund, B., Marre, R., & Goebel, W. (1990). Deletions of chromosomal regions coding for fimbriae and hemolysins occur *in vitro* and *in vivo* in various

extraintestinal *Escherichia coli* isolates. *Microbial Pathogenesis*, *8*(3), 213–225. https://doi.org/10.1016/0882-4010(90)90048-U

Hacker, J., & Kaper, J. B. (2000). Pathogenicity islands and the evolution of microbes. *Annual Review of Microbiology*, *54*, 641–679. https://doi.org/10.1146/ANNUREV.MICRO.54.1.641

Hadley Wickham, Romain François, Lionel Henry, Kirill Müller, & Davis Vaughan. (2023). *dplyr: A Grammar of Data Manipulation* (https://dplyr.tidyverse.org, https://github.com/tidyverse/dplyr).

Haren, L., Normand, C., Polard, P., Alazard, R., & Chandler, M. (2000). IS*911* transposition is regulated by protein-protein interactions via a leucine zipper motif. *Journal of Molecular Biology*, *296*(3), 757–768. https://doi.org/10.1006/JMBI.1999.3485

Haren, L., Polard, P., Ton-Hoang, B., & Chandler, M. (1998). Multiple oligomerisation domains in the IS*911* transposase: a leucine zipper motif is essential for activity. *Journal of Molecular Biology*, *283*(1), 29–41. https://doi.org/10.1006/JMBI.1998.2053

Heylen, K., Vanparys, B., Peirsegaele, F., Lebbe, L., & De Vos, P. (2007). *Stenotrophomanas terrae sp. nov*. and *Stenotrophomonas humi sp. nov*., two nitrate-reducing bacteria isolated from soil. *International Journal of Systematic and Evolutionary Microbiology*, *57*(9), 2056–2061. https://doi.org/10.1099/IJS.0.65044-0/CITE/REFWORKS

Higgins, C. F., Ames, G. F. L., Barnes, W. M., Clement, J. M., & Hofnung, M. (1982). A novel intercistronic regulatory element of prokaryotic operons. *Nature*. https://doi.org/10.1038/298760a0

Higgins, N. P. (2007). Mutational bias suggests that replication termination occurs near the *dif* site, not at Ter sites: What's the Dif? In *Molecular Microbiology* (Vol. 64, Issue 1). https://doi.org/10.1111/j.1365-2958.2007.05641.x

Hiraga, S. (1993). Chromosome partition in *Escherichia coli*. *Current Opinion in Genetics & Development*, *3*(5), 789–801. https://doi.org/10.1016/S0959-437X(05)80100-5

Home - Assembly - NCBI. (n.d.). Retrieved March 24, 2023, from https://www.ncbi.nlm.nih.gov/assembly

Hsiao, W. W. L., Ung, K., Aeschliman, D., Bryan, J., Finlay, B. B., & Brinkman, F. S. L. (2005). Evidence of a large novel gene pool associated with prokaryotic genomic islands. *PLoS Genetics*, *1*(5), e62. https://doi.org/10.1371/JOURNAL.PGEN.0010062

Hu, S., Otsubo, E., Davidson, N., & Saedler, H. (1975). Electron microscope heteroduplex studies of sequence relations among bacterial plasmids: identification and mapping of the insertion sequences IS*1* and IS*2* in F and R plasmids. *Journal of Bacteriology*, *122*(2), 764–775. https://doi.org/10.1128/JB.122.2.764-775.1975

HUGH, R., & RYSCHENKOW, E. (1961). *Pseudomonas maltophilia*, an alcaligenes-like species. *Journal of General Microbiology*, *26*(1), 123–132. https://doi.org/10.1099/00221287-26-1-123/CITE/REFWORKS

Hyung, H. L., Ji, Y. Y., Hyoun, S. K., Ji, Y. K., Kyoung, H. K., Do, J. K., Jun, Y. H., Mikami, B., Hye, J. Y., & Se, W. S. (2006). Crystal structure of a metal ion-bound IS*200* transposase. *The Journal of Biological Chemistry*, *281*(7), 4261–4266. https://doi.org/10.1074/JBC.M511567200

Ilyina, T. V., & Koonin, E. V. (1992). Conserved sequence motifs in the initiator proteins for rolling circle DNA replication encoded by diverse replicons from eubacteria, eucaryotes and archaebacteria. *Nucleic Acids Research*, *20*(13). https://doi.org/10.1093/nar/20.13.3279

Insuwanno, W., Kiratisin, P., & Jitmuang, A. (2020). *Stenotrophomonas maltophilia* infections: Clinical characteristics and factors associated with mortality of hospitalized patients. *Infection and Drug Resistance*, *13*, 1559–1566. https://doi.org/10.2147/IDR.S253949

ISfinder. (n.d.-a). Retrieved June 4, 2023, from https://isfinder.biotoul.fr/scripts/ficheIS.php?name=IS911

ISfinder. (n.d.-b). Retrieved June 24, 2023, from https://isfinder.biotoul.fr/

Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T., & Aluru, S. (2018). High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications*, *9*(1). https://doi.org/10.1038/S41467-018-07641-9

Jian, J., Xie, Z., & Chen, L. (2022). Risk Factors for Mortality in Hospitalized Patients with *Stenotrophomonas maltophilia* Bacteremia. *Infection and Drug Resistance*, *15*, 3881–3886. https://doi.org/10.2147/IDR.S371129

Johnson, C. M., & Grossman, A. D. (2015). Integrative and Conjugative Elements (ICEs): What They Do and How They Work. *Annual Review of Genetics*, *49*, 577. https://doi.org/10.1146/ANNUREV-GENET-112414-055018

Johnson, E. H., Al-Busaidy, R., & Hameed, M. S. (2003). An outbreak of lymphadenitis associated with *Stenotrophomonas (Xanthomonas) maltophilia* in Omani goats. *Journal of Veterinary Medicine. B, Infectious Diseases and Veterinary Public Health*, *50*(2), 102–104. https://doi.org/10.1046/J.1439-0450.2003.00643.X

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature 2021 596:7873*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Kaas, R. S., Leekitcharoenphon, P., Aarestrup, F. M., & Lund, O. (2014). Solving the problem of comparing whole bacterial genomes across different sequencing platforms. *PLoS ONE*, *9*(8). https://doi.org/10.1371/JOURNAL.PONE.0104984

Karch, H., Schubert, S., Zhang, D., Zhang, W., Schmidt, H., Ölschläger, T., & Hacker, J. (1999). A genomic island, termed high-pathogenicity island, is present in certain non-O157 Shiga toxin-producing *Escherichia coli* clonal lineages. *Infection and Immunity*, *67*(11), 5994–6001. https://doi.org/10.1128/IAI.67.11.5994-6001.1999

Khemici, V., & Carpousis, A. J. (2004). The RNA degradosome and poly(A) polymerase of *Escherichia coli* are required *in vivo* for the degradation of small mRNA decay intermediates containing REP-stabilizers. *Molecular Microbiology*, *51*(3). https://doi.org/10.1046/j.1365-2958.2003.03862.x

Kim, H. Bin, Srinivasan, S., Sathiyaraj, G., Quan, L. H., Kim, S. H., Bui, T. P. N., Liang, Z. Q., Kim, Y. J., & Yang, D. C. (2010). *Stenotrophomonas ginsengisoli sp. nov*., isolated from a ginseng field. *International Journal of Systematic and Evolutionary Microbiology*, *60*(7), 1522–1526. https://doi.org/10.1099/IJS.0.014662-0/CITE/REFWORKS

Klausner, J. D., Zukerman, C., Limaye, A. P., & Corey, L. (1999). Outbreak of *Stenotrophomonas maltophilia* bacteremia among patients undergoing bone marrow transplantation: association with faulty replacement of handwashing soap. *Infection Control and Hospital Epidemiology*, *20*(11), 756–758. https://doi.org/10.1086/501578

Kojima, K. K., & Bao, W. (2023). IS481EU Shows a New Connection between Eukaryotic and Prokaryotic DNA Transposons. *Biology*, *12*(3), 365. https://doi.org/10.3390/BIOLOGY12030365/S1

Krupovic, M., & Forterre, P. (2015). Single-stranded DNA viruses employ a variety of mechanisms for integration into host genomes. *Annals of the New York Academy of Sciences*, *1341*(1), 41–53. https://doi.org/10.1111/NYAS.12675

Kuempel, P., Henson, J., Dircks, L., Tecklenburg, M., & Lim, D. (1991). *dif*, a recA -independent recombination site in the terminus region of the chromosome of *Escherichia coli* . *New Biol*, *3*.

Kugeler, K. J., Staples, J. E., Hinckley, A. F., Gage, K. L., & Mead, P. S. (2015). Epidemiology of Human Plague in the United States, 1900–2012. *Emerging Infectious Diseases*, *21*(1), 16. https://doi.org/10.3201/EID2101.140564

Lai, C. H., Chi, C. Y., Chen, H. P., Chen, T. L., Lai, C. J., Fung, C. P., Yu, K. W., Wong, W. W., & Liu, C. Y. (2004). Clinical characteristics and prognostic factors of patients with *Stenotrophomonas maltophilia* bacteremia. *Journal of Microbiology, Immunology, and Infection = Wei Mian Yu Gan Ran Za Zhi*, *37*(6), 350–358.

Lai, C. H., Wong, W. W., Chin, C., Huang, C. K., Lin, H. H., Chen, W. F., Yu, K. W., & Liu, C. Y. (2006). Central venous catheter-related *Stenotrophomonas maltophilia* bacteraemia and associated relapsing bacteraemia in haematology and oncology patients. *Clinical Microbiology and Infection : The Official*

Publication of the European Society of Clinical Microbiology and Infectious Diseases, *12*(10), 986–991. https://doi.org/10.1111/J.1469-0691.2006.01511.X

Lee, J. Y., Finkelstein, I. J., Crozat, E., Sherratt, D. J., & Greene, E. C. (2012). Single-molecule imaging of DNA curtains reveals mechanisms of KOPS sequence targeting by the DNA translocase FtsK. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(17), 6531–6536. https://doi.org/10.1073/PNAS.1201613109/SUPPL_FILE/SD01.TXT

Lee, M., Woo, S. G., Chae, M., Shin, M. C., Jung, H. M., & Ten, L. N. (2011). *Stenotrophomonas daejeonensis sp. nov.*, isolated from sewage. *International Journal of Systematic and Evolutionary Microbiology*, *61*(3), 598–604. https://doi.org/10.1099/IJS.0.017780-0/CITE/REFWORKS

Letunic, I., & Bork, P. (2021). Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research*, *49*(W1), W293–W296. https://doi.org/10.1093/NAR/GKAB301

Liang, W., & Deutscher, M. P. (2016). REP sequences: Mediators of the environmental stress response? *RNA Biology*, *13*(2). https://doi.org/10.1080/15476286.2015.1112489

Lobry, J. R. (1996). Asymmetric substitution patterns in the two DNA strands of bacteria. *Molecular Biology and Evolution*, *13*(5), 660–665. https://doi.org/10.1093/OXFORDJOURNALS.MOLBEV.A025626

Loper, J. E., Hassan, K. A., Mavrodi, D. V., Davis, E. W., Lim, C. K., Shaffer, B. T., Elbourne, L. D. H., Stockwell, V. O., Hartney, S. L., Breakwell, K., Henkels, M. D., Tetu, S. G., Rangel, L. I., Kidarsa, T. A., Wilson, N. L., van de Mortel, J. E., Song, C., Blumhagen, R., Radune, D., … Paulsen, I. T. (2012). Comparative genomics of plant-associated *Pseudomonas spp*.: Insights into diversity and inheritance of traits involved in multitrophic interactions. *PLoS Genetics*, *8*(7). https://doi.org/10.1371/journal.pgen.1002784

Lu, J., & Salzberg, S. L. (2020). SkewIT: The Skew Index Test for large-scale GC Skew analysis of bacterial genomes. *PLOS Computational Biology*, *16*(12), e1008439. https://doi.org/10.1371/JOURNAL.PCBI.1008439

Massey, T. H., Mercogliano, C. P., Yates, J., Sherratt, D. J., & Löwe, J. (2006). Double-stranded DNA translocation: structure and mechanism of hexameric FtsK. *Molecular Cell*, *23*(4), 457–469. https://doi.org/10.1016/J.MOLCEL.2006.06.019

Mathee, K., Narasimhan, G., Valdes, C., Qiu, X., Matewish, J. M., Koehrsen, M., Rokas, A., Yandava, C. N., Engels, R., Zeng, E., Olavarietta, R., Doud, M., Smith, R. S., Montgomery, P., White, J. R., Godfrey, P. A., Kodira, C., Birren, B., Galagan, J. E., & Lory, S. (2008). Dynamics of *Pseudomonas aeruginosa* genome evolution. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(8), 3100–3105. https://doi.org/10.1073/PNAS.0711982105/SUPPL_FILE/11982TABLE3.XLS

Mazauric, M. H., Licznar, P., Prère, M. F., Canal, I., & Fayet, O. (2008). Apical loop-internal loop RNA pseudoknots: a new type of stimulator of -1 translational frameshifting in bacteria. *The Journal of Biological Chemistry*, *283*(29), 20421–20432. https://doi.org/10.1074/JBC.M802829200

McAdam, R. A., Hermans, P. W. M., Van Soolingen, D., Zainuddin, Z. F., Catty, D., Van Embden, J. D. A., & Dale, J. W. (1990). Characterization of a *Mycobacterium tuberculosis* insertion sequence belonging to the IS*3* family. *Molecular Microbiology*, *4*(9), 1607–1613. https://doi.org/10.1111/J.1365-2958.1990.TB02073.X

Mckinney, W. (2010). *Data Structures for Statistical Computing in Python*.

McLafferty, M. A., Harcus, D. R., & Hewlett, E. L. (1988). Nucleotide sequence and characterization of a repetitive DNA element from the genome of *Bordetella pertussis* with characteristics of an insertion sequence. *Journal of General Microbiology*, *134*(8), 2297–2306. https://doi.org/10.1099/00221287-134-8-2297

Mercier-Darty, M., Royer, G., Lamy, B., Charron, C., Lemenand, O., Gomart, C., Fourreau, F., Madec, J. Y., Jumas-Bilak, E., Decousser, J. W., Aberanne, S., Belmonte, O., Blondiaux, N., Cattoir, V., Dekeyser, S., Delarbre, J. M., Corlouer, C., Haenni, M., Jaouen, A. C., … Vachee, A. (2020). Comparative Whole-Genome Phylogeny of Animal, Environmental, and Human Strains Confirms the Genogroup

Organization and Diversity of the *Stenotrophomonas maltophilia* Complex. *Applied and Environmental Microbiology*, *86*(10). https://doi.org/10.1128/AEM.02919-19

Messing, S. A. J., Ton-Hoang, B., Hickman, A. B., McCubbin, A. J., Peaslee, G. F., Ghirlando, R., Chandler, M., & Dyda, F. (2012). The processing of repetitive extragenic palindromes: the structure of a repetitive extragenic palindrome bound to its associated nuclease. *Nucleic Acids Research*, *40*(19), 9964–9979. https://doi.org/10.1093/NAR/GKS741

Montaño, S. P., & Rice, P. A. (2011). Moving DNA around: DNA transposition and retroviral integration. In *Current Opinion in Structural Biology* (Vol. 21, Issue 3). https://doi.org/10.1016/j.sbi.2011.03.004

Muder, R. R., Harris, A. P., Muller, S., Edmond, M., Chow, J. W., Papadakis, K., Wagener, M. W., Bodey, G. P., & Steckelberg, J. M. (1996). Bacteremia due to *Stenotrophomonas (Xanthomonas) maltophilia*: A prospective, multicenter study of 91 episodes. *Clinical Infectious Diseases*, *22*(3), 508–512. https://doi.org/10.1093/CLINIDS/22.3.508

NAKATSU, C. H., FULTHORPE, R. R., HOLLAND, B. A., PEEL, M. C., & WYNDHAM, R. C. (1995). The phylogenetic distribution of a transposable dioxygenase from the Niagara River watershed. *Molecular Ecology*, *4*(5), 593–604. https://doi.org/10.1111/J.1365-294X.1995.TB00259.X

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, *48*(3), 443–453. https://doi.org/10.1016/0022-2836(70)90057-4

Newbury, S. F. S. F., Smith, N. H. N. H., Robinson, E. C. C., Hiles, I. D. I. D., & Higgins, C. F. C. F. (1987). Stabilization of translationally active mRNA by prokaryotic REP sequences. *Cell*, *48*(2), 297–310. https://doi.org/10.1016/0092-8674(87)90433-8

Nolivos, S., Touzain, F., Pages, C., Coddeville, M., Rousseau, P., El Karoui, M., Le Bourgeois, P., & Cornet, F. (2012). Co-evolution of segregation guide DNA motifs and the FtsK translocase in bacteria: identification of the atypical *Lactococcus lactis* KOPS motif. *Nucleic Acids Research*, *40*(12), 5535–5545. https://doi.org/10.1093/NAR/GKS171

Nunvar, J., Huckova, T., & Licha, I. (2010). Identification and characterization of repetitive extragenic palindromes (REP)-associated tyrosine transposases: Implications for REP evolution and dynamics in bacterial genomes. *BMC Genomics*, *11*(1). https://doi.org/10.1186/1471-2164-11-44

Nunvar, J., Licha, I., & Schneider, B. (2013). Evolution of REP diversity: A comparative study. *BMC Genomics*, *14*(1). https://doi.org/10.1186/1471-2164-14-385

Ochman, H., & Caro-Quintero, A. (2016). Genome Size and Structure, Bacterial. *Encyclopedia of Evolutionary Biology*, 179–185. https://doi.org/10.1016/B978-0-12-800049-6.00235-3

Ogasawara, N., Moriya, S., von Meyenburg, K., Hansen, F. G., & Yoshikawa, H. (1985). Conservation of genes and their organization in the chromosomal replication origin region of *Bacillus subtilis* and *Escherichia coli*. *The EMBO Journal*, *4*(12), 3345–3350. https://doi.org/10.1002/J.1460-2075.1985.TB04087.X

Oppenheim, A. B., Rudd, K. E., Mendelson, I., & Teff, D. (1993). Integration host factor binds to a unique class of complex repetitive extragenic DNA sequences in Escherichia coli. *Molecular Microbiology*, *10*(1). https://doi.org/10.1111/j.1365-2958.1993.tb00908.x

Pagès H., Aboyoun P., Gentleman R., & DebRoy S. (n.d.). *Biostrings: Efficient Manipulation of Biological Strings*. Retrieved August 6, 2023, from https://bioconductor.org/packages/release/bioc/html/Biostrings.html

Park, H. J., Gokhale, C. S., & Bertels, F. (2021). How sequence populations persist inside bacterial genomes. *Genetics*. https://doi.org/10.1093/genetics/iyab027

Parkhill, J., Sebaihia, M., Preston, A., Murphy, L. D., Thomson, N., Harris, D. E., Holden, M. T. G., Churcher, C. M., Bentley, S. D., Mungall, K. L., Cerdeño-Tárraga, A. M., Temple, L., James, K., Harris, B., Quail, M. A., Achtman, M., Atkin, R., Baker, S., Basham, D., … Maskell, D. J. (2003). Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica. Nature Genetics 2003 35:1*, *35*(1), 32–40. https://doi.org/10.1038/ng1227

Patil, P. P., Kumar, S., Midha, S., Gautam, V., & Patil, P. B. (2018). Taxonogenomics reveal multiple novel genomospecies associated with clinical isolates of *Stenotrophomonas maltophilia*. *Microbial Genomics*, *4*(8). https://doi.org/10.1099/MGEN.0.000207

Petridou, E., Filioussis, G., Karavanis, E., & Kritas, S. K. (2010). *Stenotrophomonas maltophilia* as a causal agent of pyogranulomatous hepatitis in a buffalo (*Bubalus bubalis*). *Journal of Veterinary Diagnostic Investigation: Official Publication of the American Association of Veterinary Laboratory Diagnosticians, Inc*, *22*(5), 772–774. https://doi.org/10.1177/104063871002200522

Pettersen, E. F., Goddard, T. D., Huang, C. C., Meng, E. C., Couch, G. S., Croll, T. I., Morris, J. H., Ferrin, T. E., & Thomas Ferrin, C. E. (2020). *UCSF ChimeraX: Structure visualization for researchers, educators, and developers*. https://doi.org/10.1002/pro.3943

Pienkoß, S., Javadi, S., Chaoprasid, P., Nolte, T., Twittenhoff, C., Dersch, P., & Narberhaus, F. (2021). The gatekeeper of *Yersinia* type III secretion is under RNA thermometer control. *PLOS Pathogens*, *17*(11), e1009650. https://doi.org/10.1371/JOURNAL.PPAT.1009650

Polard, P., Prère, M. F., Chandler, M., & Fayet, O. (1991). Programmed translational frameshifting and initiation at an AUU codon in gene expression of bacterial insertion sequence IS*911*. *Journal of Molecular Biology*, *222*(3), 465–477. https://doi.org/10.1016/0022-2836(91)90490-W

Ptashne, K., & Cohen, S. N. (1975). Occurrence of insertion sequence (IS) regions on plasmid deoxyribonucleic acid as direct and inverted nucleotide sequence duplications. *Journal of Bacteriology*, *122*(2), 776–781. https://doi.org/10.1128/JB.122.2.776-781.1975

Qian, Z., Macvanin, M., Dimitriadis, E. K., He, X., Zhurkin, V., & Adhya, S. (2015). A new noncoding RNA arranges bacterial chromosome organization. *MBio*, *6*(4). https://doi.org/10.1128/mBio.00998-15

Raad, M., Abou Haidar, M., Ibrahim, R., Rahal, R., Abou Jaoude, J., Harmouche, C., Habr, B., Ayoub, E., Saliba, G., Sleilaty, G., Mounzer, K., Saliba, R., & Riachy, M. (2023). *Stenotrophomonas maltophilia* pneumonia in critical COVID-19 patients. *Scientific Reports 2023 13:1*, *13*(1), 1–12. https://doi.org/10.1038/s41598-023-28438-x

Ramos, P. L., Van Trappen, S., Thompson, F. L., Rocha, R. C. S., Barbosa, H. R., de Vos, P., & Moreira-Filho, C. A. (2011). Screening for endophytic nitrogen-fixing bacteria in Brazilian sugar cane varieties used in organic farming and description of *Stenotrophomonas pavanii sp. nov*. *International Journal of Systematic and Evolutionary Microbiology*, *61*(4), 926–931. https://doi.org/10.1099/IJS.0.019372-0/CITE/REFWORKS

Ramos-González, M. I. M. I., Campos, M. J. M. J., Ramos, J. L. J. L., & Espinosa-Urgel, M. (2006). Characterization of the *Pseudomonas putida* mobile genetic element IS*Ppu10*: An occupant of repetitive extragenic palindromic sequences. *Journal of Bacteriology*, *188*(1), 37–44. https://doi.org/10.1128/JB.188.1.37-44.2006

Ran, L., Larsson, J., Vigil-Stenman, T., Nylander, J. A. A., Ininbergs, K., Zheng, W. W., Lapidus, A., Lowry, S., Haselkorn, R., & Bergman, B. (2010). Genome erosion in a nitrogen-fixing vertically transmitted endosymbiotic multicellular cyanobacterium. *PloS One*, *5*(7). https://doi.org/10.1371/JOURNAL.PONE.0011486

Reece, R. J., Maxwell, A., & Wang, J. C. (1991). DNA gyrase: Structure and function. *Critical Reviews in Biochemistry and Molecular Biology*, *26*(3–4). https://doi.org/10.3109/10409239109114072

Rice, P. A., Yang, S. W., Mizuuchi, K., & Nash, H. A. (1996). Crystal structure of an IHF-DNA complex: A protein-induced DNA U-turn. *Cell*, *87*(7). https://doi.org/10.1016/S0092-8674(00)81824-3

Rocco, F., De Gregorio, E., Colonna, B., & Di Nocera, P. P. (2009). *Stenotrophomonas maltophilia* genomes: A start-up comparison. *International Journal of Medical Microbiology*, *299*(8). https://doi.org/10.1016/j.ijmm.2009.05.004

Rocco, F., De Gregorio, E., & Di Nocera, P. P. (2010). A giant family of short palindromic sequences in *Stenotrophomonas maltophilia*. *FEMS Microbiology Letters*, *308*(2). https://doi.org/10.1111/j.1574-6968.2010.02010.x

Rocha, E. P. C. (2004). The replication-related organization of bacterial genomes. *Microbiology (Reading, England)*, *150*(Pt 6), 1609–1627. https://doi.org/10.1099/MIC.0.26974-0

Roscetto, E., Carlomagno, M. S., Casalino, M., Colonna, B., Zarrilli, R., & Di Nocera, P. P. (2008). PCR-based rapid genotyping of *Stenotrophomonas maltophilia* isolates. *BMC Microbiology*, *8*, 202. https://doi.org/10.1186/1471-2180-8-202

Rosselin, E. G., Drouet, J., & Drouet, A. (1968). Specific activity of 131 IS*3* measured from the labelling of the insulin molecule. *Revue Francaise d'etudes Cliniques et Biologiques*, *13*(8), 812–814. https://pubmed.ncbi.nlm.nih.gov/5714877/

Ryan, V. T., Grimwade, J. E., Nievera, C. J., & Leonard, A. C. (2002). IHF and HU stimulate assembly of pre-replication complexes at *Escherichia coli* oriC by two different mechanisms. *Molecular Microbiology*, *46*(1). https://doi.org/10.1046/j.1365-2958.2002.03129.x

Salkeld, D. J., Salathé, M., Stapp, P., & Jones, J. H. (2010). Plague outbreaks in prairie dog populations explained by percolation thresholds of alternate host abundance. *Proceedings of the National Academy of Sciences of the United States of America*, *107*(32), 14247–14250. https://doi.org/10.1073/PNAS.1002826107/SUPPL_FILE/PNAS.201002826SI.PDF

Schaper, S., & Messer, W. (1995). Interaction of the initiator protein DnaA of *Escherichia coli* with its DNA target. *The Journal of Biological Chemistry*, *270*(29), 17622–17626. https://doi.org/10.1074/JBC.270.29.17622

Segerman, B. (2012). The genetic integrity of bacterial species: the core genome and the accessory genome, two different stories. *Frontiers in Cellular and Infection Microbiology*, *2*, 116. https://doi.org/10.3389/FCIMB.2012.00116/BIBTEX

Sekine, Y., Aihara, K., & Ohtsubo, E. (1999). Linearization and transposition of circular molecules of insertion sequence IS*3*. *Journal of Molecular Biology*, *294*(1), 21–34. https://doi.org/10.1006/JMBI.1999.3181

Shao, F. (2008). Biochemical functions of *Yersinia* type III effectors. *Current Opinion in Microbiology*, *11*(1), 21–29. https://doi.org/10.1016/J.MIB.2008.01.005

Siguier, P., Gourbeyre, E., Varani, A., Ton-Hoang, B., & Chandler, M. (2015). Everyman's Guide to Bacterial Insertion Sequences. *Microbiology Spectrum*, *3*(2). https://doi.org/10.1128/MICROBIOLSPEC.MDNA3-0030-2014/ASSET/AAEAF137-E6C3-45EF-8B4D-4B497F16E17D/ASSETS/GRAPHIC/MDNA3-0030-2014-FIG2.GIF

Siguier, P., Perochon, J., Lestrade, L., Mahillon, J., & Chandler, M. (2006). ISfinder: the reference centre for bacterial insertion sequences. *Nucleic Acids Research*, *34*(Database issue). https://doi.org/10.1093/NAR/GKJ014

Steinegger, M., & Söding, J. (2018). Clustering huge protein sequence sets in linear time. *Nature Communications 2018 9:1*, *9*(1), 1–8. https://doi.org/10.1038/s41467-018-04964-5

Stern, M. J., Ames, G. F. L., Smith, N. H., Clare Robinson, E., & Higgins, C. F. (1984). Repetitive extragenic palindromic sequences: A major component of the bacterial genome. *Cell*, *37*(3). https://doi.org/10.1016/0092-8674(84)90436-7

Stibitz, S. (1998). IS*481* and IS*1002* of *Bordetella pertussis* Create a 6-Base-Pair Duplication upon Insertion at a Consensus Target Site. *Journal of Bacteriology*, *180*(18), 4963. https://doi.org/10.1128/JB.180.18.4963-4966.1998

Stuitje, A. R., Wind, N. de, Spek, J. c. van Der, Pors, T. H., & Meijer, M. (1986). Dissection of promoter sequences involved in transcriptional activation of the *Escherichia coli* replication origin. *Nucleic Acids Research*, *14*(5), 2333–2344. https://doi.org/10.1093/NAR/14.5.2333

Sui, Y. S., Wan, G. H., Chen, Y. W., Ku, H. L., Li, L. P., Liu, C. H., & Mau, H. S. (2012). Effectiveness of bacterial disinfectants on surfaces of mechanical ventilator systems. *Respiratory Care*, *57*(2), 250–256. https://doi.org/10.4187/respcare.01180

Tamura, K., Stecher, G., & Kumar, S. (2021). MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Molecular Biology and Evolution*, *38*(7), 3022–3027. https://doi.org/10.1093/MOLBEV/MSAB120

Tan, S. Y., Tan, I. K. P., Tan, M. F., Dutta, A., & Choo, S. W. (2016). Evolutionary study of *Yersinia* genomes deciphers emergence of human pathogenic species. *Scientific Reports*, *6*. https://doi.org/10.1038/SREP36116

Terentjeva, M., & Bērziņš, A. (2010). Prevalence and Antimicrobial Resistance of *Yersinia enterocolitica* and *Yersinia pseudotuberculosis* in Slaughter Pigs in Latvia. *Journal of Food Protection*, *73*(7), 1335–1338. https://doi.org/10.4315/0362-028X-73.7.1335

Thanbichler, M. (2010). Synchronization of chromosome dynamics and cell division in bacteria. *Cold Spring Harbor Perspectives in Biology*, *2*(1). https://doi.org/10.1101/CSHPERSPECT.A000331

Tobes, R., & Pareja, E. (2006). Bacterial repetitive extragenic palindromic sequences are DNA targets for Insertion Sequence elements. *BMC Genomics*, *7*(1), 1–12. https://doi.org/10.1186/1471-2164-7-62/FIGURES/2

Tobes, R., & Ramos, J. L. (2005). REP code: Defining bacterial identity in extragenic space. *Environmental Microbiology*. https://doi.org/10.1111/j.1462-2920.2004.00704.x

Ton-Hoang, B., Bétermier, M., Polard, P., & Chandler, M. (1997). Assembly of a strong promoter following IS*911* circularization and the role of circles in transposition. *The EMBO Journal*, *16*(11), 3357–3371. https://doi.org/10.1093/EMBOJ/16.11.3357

Turlan, C., Loot, C., & Chandler, M. (2004). IS*911* partial transposition products and their processing by the *Escherichia coli* RecG helicase. *Molecular Microbiology*, *53*(4), 1021–1033. https://doi.org/10.1111/J.1365-2958.2004.04165.X

Van Melderen, L., & De Bast, M. S. (2009). Bacterial Toxin–Antitoxin Systems: More Than Selfish Entities? *PLoS Genetics*, *5*(3). https://doi.org/10.1371/JOURNAL.PGEN.1000437

Van Zundert, G. C. P., Rodrigues, J. P. G. L. M., Trellet, M., Schmitz, C., Kastritis, P. L., Karaca, E., Melquiond, A. S. J., Van Dijk, M., De Vries, S. J., & Bonvin, A. M. J. J. (2016). The HADDOCK2.2 Web Server: User-Friendly Integrative Modeling of Biomolecular Complexes. *Journal of Molecular Biology*, *428*(4), 720–725. https://doi.org/10.1016/J.JMB.2015.09.014

Vigil-Stenman, T., Larsson, J., Nylander, J. A. A., & Bergman, B. (2015). Local hopping mobile DNA implicated in pseudogene formation and reductive evolution in an obligate cyanobacteria-plant symbiosis. *BMC Genomics*, *16*(1). https://doi.org/10.1186/S12864-015-1386-7

Vinuesa, P., Ochoa-Sánchez, L. E., & Contreras-Moreira, B. (2018). GET_PHYLOMARKERS, a Software Package to Select Optimal Orthologous Clusters for Phylogenomics and Inferring Pan-Genome Phylogenies, Used for a Critical Geno-Taxonomic Revision of the Genus *Stenotrophomonas*. *Frontiers in Microbiology*, *9*(MAY). https://doi.org/10.3389/FMICB.2018.00771

Wang, X., & Wood, T. K. (2011). Toxin-Antitoxin Systems Influence Biofilm and Persister Cell Formation and the General Stress Response. *Applied and Environmental Microbiology*, *77*(16), 5577. https://doi.org/10.1128/AEM.05068-11

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., & Barton, G. J. (2009). Jalview Version 2—a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, *25*(9), 1189–1191. https://doi.org/10.1093/BIOINFORMATICS/BTP033

WebLogo 3 - Create. (n.d.). Retrieved December 3, 2022, from https://weblogo.threeplusone.com/create.cgi

Weigand, M. R., Peng, Y., Loparev, V., Batra, D., Bowden, K. E., Burroughs, M., Cassiday, P. K., Davis, J. K., Johnson, T., Juieng, P., Knipe, K., Mathis, M. H., Pruitt, A. M., Rowe, L., Sheth, M., Tondella, M. L., & Williams, M. M. (2017). The History of *Bordetella pertussis* Genome Evolution Includes Structural Rearrangement. *Journal of Bacteriology*, *199*(8). https://doi.org/10.1128/JB.00806-16

Weinel, C., Ussery, D. W., Ohlsson, H., Sicheritz-Ponten, T., Kiewitz, C., & Tümmler, B. (2002). Comparative Genomics of *Pseudomonas aeruginosa* PAO1 and *Pseudomonas putida* KT2440: Orthologs, Codon Usage, Repetitive Extragenic Palindromic Elements, and Oligonucleotide Motif Signatures. *Genome Letters*, *1*(4). https://doi.org/10.1166/gl.2002.021

Wickham, H., Averick, M., Bryan, J., Chang, W., D', L., Mcgowan, A., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Lin Pedersen, T., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open-Source Software*, *4*(43), 1686. https://doi.org/10.21105/JOSS.01686

Wilde, C., Bachellier, S., Hofnung, M., & Clément, J. M. (2001). Transposition of IS*1397* in the family *Enterobacteriaceae* and first characterization of IS*Kpn1*, a new insertion sequence associated with *Klebsiella pneumoniae* palindromic units. *Journal of Bacteriology*, *183*(15), 4395–4404. https://doi.org/10.1128/JB.183.15.4395-4404.2001/ASSET/130845CA-0CBE-4B70-B3FC-22119564B938/ASSETS/GRAPHIC/JB1510216010.JPEG

Wilde, C., Escartin, F., Kokeguchi, S., Latour-Lambert, P., Lectard, A., & Clément, J. M. (2003). Transposases are responsible for the target specificity of IS*1397* and IS*Kpn1* for two different types of palindromic units (PUs). *Nucleic Acids Research*, *31*(15), 4345–4353. https://doi.org/10.1093/NAR/GKG494

Wishart, M. M., & Riley, T. V. (1976). Infection with *Pseudomonas maltophilia* hospital outbreak due to contaminated disinfectant. *The Medical Journal of Australia*, *2*(19), 710–712. https://doi.org/10.5694/J.1326-5377.1976.TB128238.X

Wolf, A., Fritze, A., Hagemann, M., & Berg, G. (2002). *Stenotrophomonas rhizophila sp. nov.*, a novel plant-associated bacterium with antifungal properties. *International Journal of Systematic and Evolutionary Microbiology*, *52*(6), 1937–1944. https://doi.org/10.1099/00207713-52-6-1937/CITE/REFWORKS

Xu, Y., Cheng, T., Rao, Q., Zhang, S., & Ma, Y. ling. (2023). Comparative genomic analysis of *Stenotrophomonas maltophilia* unravels their genetic variations and versatility trait. *Journal of Applied Genetics*, *64*(2), 351–360. https://doi.org/10.1007/S13353-023-00752-0/FIGURES/6

Yang, Y., & Ferro-Luzzi Ames, G. (1988). DNA gyrase binds to the family of prokaryotic repetitive extragenic palindromic sequences. *Proceedings of the National Academy of Sciences of the United States of America*, *85*(23), 8850–8854. https://doi.org/10.1073/pnas.85.23.8850

Yen, M. R., Lin, N. T., Hung, C. H., Choy, K. T., Weng, S. F., & Tseng, Y. H. (2002). oriC region and replication termination site, *dif*, of the *Xanthomonas campestris pv. campestris* 17 chromosome. *Applied and Environmental Microbiology*, *68*(6), 2924–2933. https://doi.org/10.1128/AEM.68.6.2924-2933.2002

Yu, G. (2023). *Scatter Pie Plot [R package scatterpie version 0.2.1]*. https://CRAN.R-project.org/package=scatterpie

Zekrí, S., & Toro, N. (1996). Identification and nucleotide sequence of *Rhizobium meliloti* insertion sequence IS*Rm6*, a small transposable element that belongs to the IS*3* family. *Gene*, *175*(1–2), 43–48. https://doi.org/10.1016/0378-1119(96)00118-7

Zhang, L. I., Xian-Zhi, L. I., & Poole, K. (2000). Multiple Antibiotic Resistance in *Stenotrophomonas maltophilia*: Involvement of a Multidrug Efflux System. *Antimicrobial Agents and Chemotherapy*, *44*(2), 287. https://doi.org/10.1128/AAC.44.2.287-293.2000