Univerzita Karlova

Přírodovědecká fakulta

Studijní program: Bioinformatika

Studijní obor: Bioinformatika

Zuzana Halenková

Současné přístupy celogenomového sekvenování a *de novo* sestavení genomu

Current approaches to whole genome sequencing and *de novo* genome assembly

Bakalářská práce

Vedoucí práce: RNDr. Radka Reifová, Ph.D.

Konzultant: Mgr. Jan Pačes, Ph.D.

Praha, 2018

Prohlášení

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

V Praze, 7. 5. 2018

Zuzana Halenková

Poděkování

Ráda bych poděkovala své školitelce, RNDr. Radce Reifové, Ph.D., za odborné vedení práce a čas, který mi věnovala během konzultací. Děkuji také Mgr. Janu Pačesovi, Ph.D. a Mgr. Jakubovi Rídlovi, Ph.D. za cenné připomínky k práci.

Abstrakt

Během uplynulých deseti let klesla díky vývoji sekvenátorů druhé a třetí generace cena sekvenování téměř desettisíckrát. Osekvenování a sestavení celogenomové sekvence organismu je tak čím dál tím dostupnějším nástrojem a může najít využití v mnoha vědních oborech. Na cestě za kompletní sekvencí DNA organismu je nutné učinit několik rozhodnutí, která jsou stěžejní pro úspěšné sestavení kvalitní celogenomové sekvence. Tato rozhodnutí se týkají přípravy vzorků, výběru metody sekvenování a stanovení vhodného přístupu k sestavení sekvence. Bakalářská práce popisuje různé metody, které lze pro jednotlivé části procesu zvolit, a aspekty, které je nutné při rozhodování zohlednit.

Klíčová slova: sekvenování nové generace, sekvenování třetí generace, celogenomové sekvenování, *de novo* sestavení genomu, algoritmy genomového sestavení

Abstract

The cost of sequencing has fallen almost ten thousand times over the past ten years due to the development of second and third generation sequencers. Sequencing and assembling the whole genome sequence of an organism is thus becoming a more affordable tool which can be utilized in many fields of science. On the way to the complete DNA sequence of an organism, multiple important decisions have to be made. These are crucial for the successful assembly of high-quality whole genome sequence and regard sample preparation, choice of sequencing technique and choice of an appropriate approach to whole genome assembly. This bachelor thesis describes various methods which can be utilized in individual steps of the process and aspects to consider while making the decisions.

Keywords: next generation sequencing, third generation sequencing, whole genome sequencing, *de novo* assembly, genome assembly algorithms

Obsah

Seznam zkratek

ATP = adenosintrifosfát

BLAST = *basic local alignment search tool*

CMOS = *complementary metal-oxide semiconductor*

DNAP = DNA polymeráza

dNTP = deoxyribonukleotid

ddNTP = dideoxyribonukleotid

emPCR = emulzní polymerázová řetězová reakce

GEM = *gel bead in emulsion*

ISFET = iontově senzitivní tranzistor s efektem pole (*ion-sensitive field-effect transistor*)

NGS = sekvenování nové generace (*next generation sequencing*)

OLC = *overlap/layout/consensus*

ONT = Oxford Nanopore Technologies

PCR = polymerázová řetězová reakce (*polymerase chain reaction*)

SMRT = Single Molecule Real-Time

ZMW = *zero mode waveguide*

1. Úvod

Rozvoj sekvenačních metod a klesající ceny sekvenování (online zdroj č. 7) přináší zajímavé možnosti pro mnohé genetické obory, například pro populační, ochranářskou nebo lékařskou genetiku. Jednou z možností, k čemu se dnes běžně využívá sekvenování, je i osekvenování a sestavení kompletní genomové sekvence zkoumaného organismu, což je tématem této bakalářské práce.

Před zahájením konkrétních kroků vedoucích k získání celogenomové sekvence je nutné rozmyslet si důkladně jednotlivé fáze tohoto procesu a především techniky, které během nich budou použity. Začíná se odběrem vzorku a izolací DNA. Poté následuje sekvenování. V současnosti používané metody sekvenování jsou označovány jako *shotgun* – místo zpracovávání dlouhé makromolekuly DNA v celku zjišťují sekvence jejích kratších náhodných úseků (Ekblom & Wolf, 2014). Tyto osekvenované úseky DNA se nazývají čtení (anglicky *reads*). Pozice čtení v rámci řetězce jsou ve fázi sekvenování neznámé, ale jejich stanovení je cílem dalšího kroku, tedy genomového sestavení. Výsledkem fáze sestavení by měly být pokud možno kontinuální sekvence všech chromozomů zkoumaného organismu (Miller et al., 2010). K sestavení sekvence je možné přistoupit dvěma způsoby. První přístup využívá již známou sekvenci příbuzného druhu (popsáno například v článku Lischer & Shimizu, 2017). Naproti tomu druhý přístup, označovaný jako *de novo*, staví sekvenci zcela od začátku. Posledním krokem sekvenačního projektu by pak vždy měla být anotace, při níž jsou k jednotlivým úsekům sekvence připojována užitečná data, jako jsou například predikované geny nebo data z RNA sekvenování.

Cílem této bakalářské práce je shrnout v současnosti používané metody sekvenování DNA a principy *de novo* celogenomového sestavení.

2. Vzorek DNA

Kvalita vzorku DNA může výrazně ovlivnit kvalitu výsledné sekvence získané během celogenomového sestavení, je tedy důležité věnovat tomuto kroku dostatečnou pozornost. Výslednou sekvenci ale také pochopitelně ovlivňují i další faktory, především chybovost sekvenačních platforem i sestavovacích algoritmů, a v neposlední řadě také vlastnosti samotného genomu, jako je jeho velikost, duplikovanost nebo množství repetic.

Po odběru vzorku tkáně z organismu dochází činností nukleáz k postupné fragmentaci DNA (Johnson & Ferris, 2002). Je tedy nutné použít pro izolaci DNA čerstvou tkáň, která nám umožní sekvenovat delší úseky (Jain et al., 2018). Pokud je třeba tkáň před sekvenováním delší dobu uchovat, využívá se například zamrazení v tekutém dusíku a následné uchování při alespoň -80 °C nebo konzervace tkáně v 96% etanolu (Genome 10K Community of Scientists, 2009; Wong et al., 2012). Pokud však potřebujeme izolovat dlouhé úseky DNA (v délce několika desítek kilobází), což je důležité pro některé postupy sekvenování genomu, které usnadňují sestavení sekvence, je vždy nejlepší čerstvá tkáň.

Důležitá je i volba tkáně, ze které se bude DNA izolovat. Měla by to být tkáň s vysokým obsahem DNA. Při sekvenování lidského genomu se často používá krev, kvůli jejímu snadnému neinvazivnímu odběru, ačkoliv kvůli bezjadernosti erytrocytů vlastně intuitivně není nejlepší volbou. Oproti tomu při sekvenování například ptačího genomu může být krev vhodnější variantou, jelikož jejich erytrocyty jádro mají. Nejlepší výtěžek DNA je možné získat z varlete a jater (Wong et al., 2012). Při volbě tkáně je nutné zohlednit také její funkci. Například tkáň svalová obsahuje kvůli dobrému zásobení energií velké množství mitochondrií. Dlouhé repetice mitochondriálního genomu poté komplikují sestavení genomové sekvence (Ekblom & Wolf, 2014). Je také vhodné se vyhnout tkáním, které mohou obsahovat cizorodou DNA – například žaludku nebo střevům (Wong et al., 2012).

DNA je z tkáně izolována obvykle pomocí některého z komerčně připravených kitů. Množství DNA potřebného k samotnému sekvenování se mezi jednotlivými platformami liší. Řádově se obvykle pohybuje v jednotkách až desítkách mikrogramů (Ekblom & Wolf, 2014), například u Single Molecule Real-Time sekvenování je to 5 μg a více (Ardui et al., 2018). Rozdíly v množství potřebné DNA mezi jednotlivými platformami jsou způsobené tím, že některé před samotným sekvenováním zařazují amplifikaci DNA pomocí polymerázové řetězové reakce (do této skupiny patří například Illumina sekvenování), zatímco jiné sekvenují rovnou (například Single Molecule Real-Time sekvenování) (Chakraborty et al., 2016).

Posledním krokem před samotným sekvenováním je příprava knihovny, tedy kolekce fragmentů DNA, které bude sekvenátor schopný sekvenovat. Příprava této kolekce se liší mezi jednotlivými platformami a komerční společnosti poskytující sekvenátory nabízí i kity pro tyto účely. Odlišnost mezi jednotlivými platformami je například v požadované délce fragmentů DNA.

3. Sekvenování

V této kapitole budou stručně představeny obvykle používané sekvenační platformy, porovnány jejich mechanismy sekvenování a uvedeny výhody i nevýhody jejich použití.

3.1 Techniky sekvenování

Rozlišujeme tři generace sekvenování. První generaci sekvenování představuje především Sangerovo sekvenování terminací řetězců. Dále se sem řadí i chemické Maxam-Gilbertovo sekvenování, které ale nebylo tak široce používané a dnes se již nevyužívá. Pro metody druhé generace sekvenování (označované též jako metody sekvenování nové generace – NGS) je charakteristická amplifikace DNA pomocí PCR a masivní paralelizace procesu (Srinivasan & Batra, 2014). Do této generace sekvenování patří 454 sekvenování, Illumina sekvenování, SOLiD sekvenování a Ion Torrent sekvenování. Třetí generace sekvenování čte sekvenci DNA na úrovni jednotlivých molekul bez nutnosti předchozí amplifikace (Heather & Chain, 2016). Metodami sekvenování třetí generace jsou Single Molecule Real-Time sekvenování a sekvenování nanopórem. V následujících kapitolách jsou popsány mechanismy sekvenování zde zmíněných technik. Některé z metod jsou známější pod jménem společnosti, která vyrábí sekvenátory je využívající, proto je jméno takové společnosti uvedeno v závorce u názvu sekvenační techniky.

3.1.1 Sangerovo sekvenování (Applied Biosystems)

Sangerovo sekvenování bylo popsáno v článku Sanger et al. (1977) a přineslo průlom v této oblasti výzkumu – s úpravami, které umožnily částečnou automatizaci procesu, bylo možné sekvenovat a sestavovat první kompletní genomové sekvence (Fleischmann et al., 1995).

Mechanismus sekvenování spočívá v syntéze komplementárního vlákna DNA podle templátu DNA polymerázou a její terminaci. Té je možné docílit inkorporací 2‘,3‘-dideoxyribonukleotidu (ddNTP), který oproti přirozeně pro syntézu DNA využívaným 2‘-deoxyribonukleotidům (dNTP) postrádá 3‘-hydroxylovou skupinu, a tak znemožňuje další prodlužování DNA vlákna (Atkinson et al., 1969). Pokud tedy vytvoříme směs DNA templátu, primeru, všech čtyř dNTP a jednoho z ddNTP a budeme ji inkubovat s DNA polymerázou, získáme kolekci fragmentů různých délek se stejným 5‘ koncem a ukončených shodně použitým ddNTP. Tuto kolekci následně separujeme pomocí elektroforézy na denaturujícím akrylamidovém gelu, a tak můžeme pozorovat rozložení inkorporovaného ddNTP napříč

syntetizovanou DNA. Pokud stejný postup uskutečníme pro všechny čtyři ddNTP, je možné následně z gelu přečíst kompletní sekvenci (Sanger et al., 1977).

Značení jednotlivých ddNTP různými fluorescenčními značkami následně umožnilo provádět všechny čtyři reakce zároveň v jedné směsi. Automatická detekce fluorescence (Luckey et al., 1990) pak dala vzniknout první automatizované sekvenační technice schopné zvládnout objemy dat potřebné pro celogenomové sekvenování.

3.1.2 454 sekvenování (454 Life Sciences)

První NGS technikou bylo tzv. pyrosekvenování, známé také jako 454 sekvenování, zavedené v roce 2005 (Margulies et al., 2005). V dnešní době se již tato technika nevyužívá.

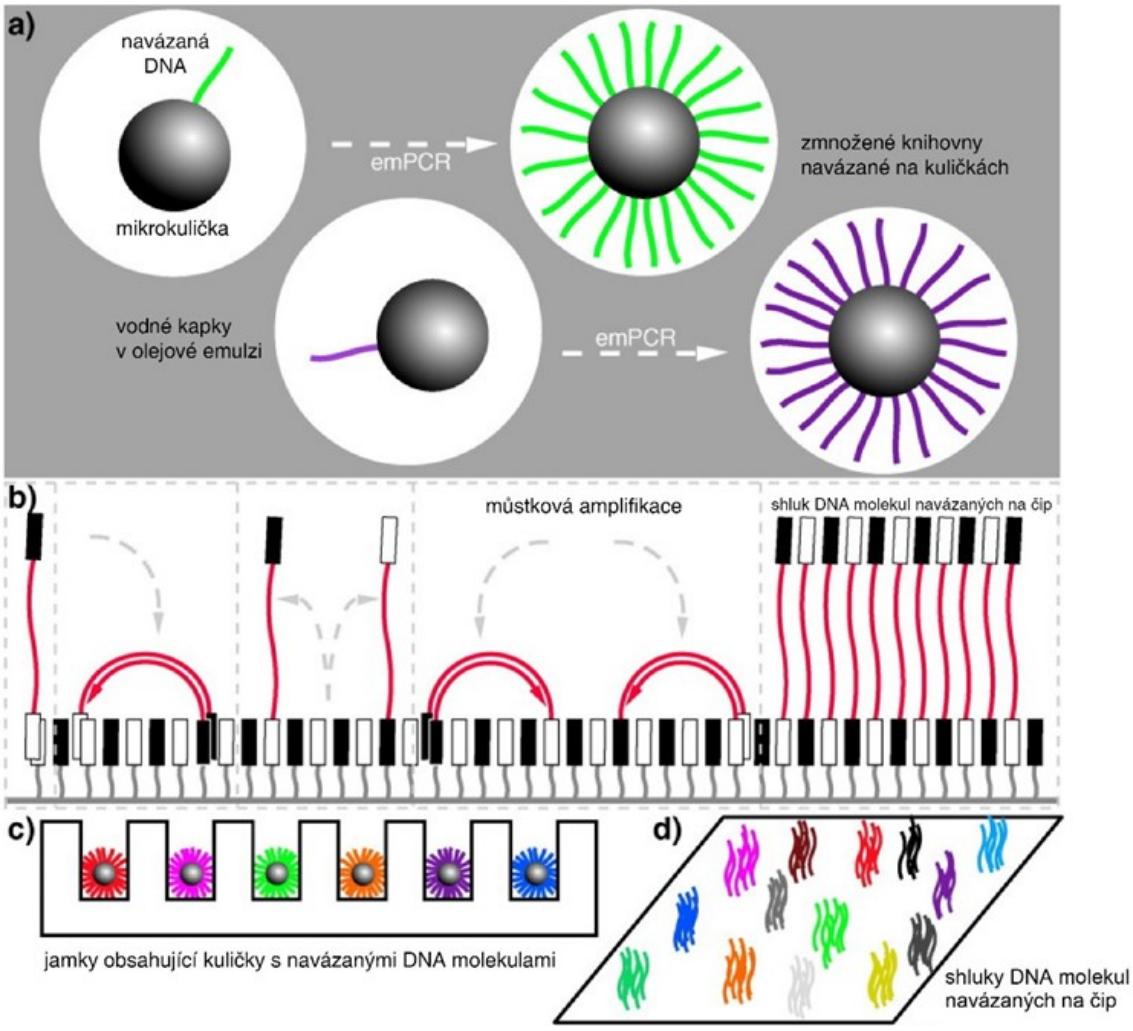
Klíčovým krokem všech NGS technik včetně pyrosekvenování je zavedení nových technik amplifikace fragmentů DNA pomocí polymerázové řetězové reakce (PCR), které umožňují zmnožení mnoha fragmentů DNA paralelně v jedné reakci. To následně umožňuje paralelizaci procesu sekvenování, který přichází na řadu až po této amplifikaci.

V případě pyrosekvenování je jako amplifikační metoda využívána tzv. emulzní PCR (Obr. 1a). Smícháním menšího množství vodného roztoku obsahujícího kuličky pokryté oligonukleotidy, fragmenty DNA a reaktanty klasické PCR s olejem o větším objemu (Dressman et al., 2003) vznikne emulze vody v oleji – kapičky obsahující v ideálním případě právě jednu kuličku a právě jeden DNA fragment. V těchto kapičkách pak proběhne PCR, při které oligonukleotidy navázané na kuličku vystupují jako primery (Dressman et al., 2003). Dojde tak ke vzniku kuličky pokryté kopiemi původního DNA fragmentu. Následně jsou tyto kuličky vyjmuty z oleje pomocí fázové separace.

Takto připravené kuličky s navázanými zmnoženými fragmenty DNA jsou jednotlivě rozmístěny do jamek pikotitrační destičky (Obr. 1c) a samotné sekvenování potom probíhá v každé jamce zvlášť. To umožňuje paralelní sekvenování statisíců fragmentů DNA současně (Margulies et al., 2005). A právě tato možnost paralelizace řadí 454 sekvenování mezi NGS metody.

Mechanismus pyrosekvenování spočívá stejně jako u Sangerova sekvenování v syntéze komplementárního DNA vlákna DNA polymerázou podle templátu, který je ale v tomto případě přichycen na pevné fázi (Ronaghi et al., 1996). Pyrofosfát uvolněný při inkorporaci dNTP je detekován pomocí dvou enzymatických reakcí: za přítomnosti enzymu ATP-sulfurylázy interaguje s adenosin 5‘-fosfosulfátem za vzniku ATP. Enzym luciferáza následně

umožní reakci ATP s luciferinem, při níž dochází k emisi světla (Nyrén, 1987). Pokud tedy zajistíme přítomnost právě jednoho druhu deoxyribonukleotidů ve směsi, můžeme detekovat, zda byl dNTP zařazen DNA polymerázou do nově syntetizovaného vlákna, a dokonce i množství dNTP stejného druhu zařazených po sobě (podle intenzity detekovaného signálu) (Ronaghi et al., 1996). Zajistíme-li střídavé přidávání a následné odmytí jednotlivých druhů dNTP, můžeme podle aktuálního druhu dNTP přítomného ve směsi odečítat sekvenci. Co se



Obr. 1: Metody amplifikace DNA.
a) Emulzní PCR. Molekuly jednořetězcové DNA s označeným 5‘ a 3‘ koncem jsou jednotlivě imobilizovány na mikrokuličkách a tyto jsou vloženy do emulze. V ideálním případě obsahuje každá kapka emulze pouze jednu kuličku. Následně uvnitř každé kapky probíhá emulzní PCR, při které vznikající molekuly DNA zůstávají navázané na kuličce.
b) Můstková PCR. Jednořetězcová molekula DNA má konce označené dvěma typy oligonukleotidů komplementárními k oligonukleotidům připevněným k povrchu čipu. Prostřednictvím párování těchto oligonukleotidů se molekula DNA také připojí k povrchu a následně během PCR vytvoří shluk identických molekul tak, že se ohýbá k sousedním primerům.
c) Kuličky z emulzní PCR jsou rozmístěny do jamek pikotitrační destičky navržených tak, aby se do nich vešla právě jedna kulička.
d) Shluky namnožených fragmentů DNA na čipu.
(Heather & Chain, 2016; převzato a upraveno)

odmytí nepoužitých dNTP týče, výrazný posun přinesl článek Ronaghi et al. (1998), ve kterém bylo navrženo přidání čtvrtého enzymu do směsi. Ten měl schopnost degradovat nukleotidy a ATP, ovšem pomaleji než DNA polymeráza nukleotid inkorporuje. Krok odmytí tak mohl být vynechán. Enzymem splňujícím tyto požadavky je například apyráza (Ronaghi et al., 1998).

3.1.3 Illumina sekvenování (Illumina)

Mezi techniky sekvenování druhé generace se rovněž řadí v současnosti nejpoužívanější sekvenační platforma od společnosti Illumina, původně vyvíjená firmou Solexa. Tato metoda pro amplifikaci sekvenovaných DNA fragmentů využívá tzv. můstkovou amplifikaci (Obr. 1b). Na oba konce fragmentů DNA je připojena adaptorová sekvence (*Y* na 5‘ konec, *Z* na 3‘ konec). Sekvenační čip (*flow cell*) je pokryt dvěma typy oligonukleotidů (*primer 1*, *primer 2*) imobilizovanými na povrchu jejich 5‘ konci. *Primer 1* je komplementární k sekvenci *Z*. 5‘ koncem s *Y* sekvencí je na povrch navázán alespoň jeden fragment DNA. Jeho adaptorová sekvence *Z* potom páruje s *primerem 1* přichyceným k čipu. DNA polymeráza nyní od tohoto primeru nasyntetizuje komplementární vlákno. Denaturací získáme dvě jednořetězcová vlákna DNA připojená k povrchu čipu jejich 5‘ konci a označenými sekvencemi komplementárními k jednomu z primerů na 3‘ konci. Dále je postup (párování s primerem, syntéza, denaturace) opakován (postup patentován – Mayer, 2004).

Samotná metoda sekvenování využívá koncept modifikovaných substrátů pro DNA polymerázu popsaný článkem Canard & Sarfati (1994). Namísto 2‘-deoxyribonukleotid 5‘-trifosfátů s OH skupinou na 3‘ uhlíku je komplementární vlákno podle templátu syntetizováno z 2‘-deoxyribonukleotid 5‘-trifosfátů s fluorescenční značkou na 3‘ uhlíku. Tato značka má tři zásadní vlastnosti: je specifická pro každou bázi, je snadno identifikovatelná (proto fluorescenční) a je možné ji snadno odštěpit při zachování stability DNA a vytvoření 3‘ OH konce, který umožní další syntézu (Canard & Sarfati, 1994). Takto modifikované dNTP jsou nazývány reverzibilními terminátory.

Princip sekvenování je tedy potom následující: jednořetězcové vlákno DNA je přichyceno na pevné fázi, na něm je nasednutý primer poskytující 3‘ OH skupinu pro další syntézu. Po přidání DNA polymerázy a směsi všech čtyř druhů reverzibilních terminátorů dojde k inkorporaci právě jednoho dNTP komplementárního s bází templátu. Zbylé dNTP jsou odmyty. Následně je značka odštěpena a nahrazena hydroxylovou skupinou. Po excitaci laserem je emitována fluorescence, díky které lze identifikovat zařazenou bázi (Heather & Chain, 2016). Tento postup je opakován pro každou bázi.

Illumina sekvenování využívá simultánního sekvenování shluků stejných vláken (Obr. 1d) vygenerovaných pomocí můstkové amplifikace a detekcí emitované vlnové délky a intenzity signálu určuje aktuální bázi za celý shluk. Na čipu se nachází stovky milionů takových shluků, což umožňuje masivní paralelizaci celého procesu sekvenování (Srinivasan & Batra, 2014).

Prvním Illumina sekvenátorem na trhu byl Genome Analyzer představený roku 2006. Dnes Illumina nabízí celou řadu sekvenátorů speciálně uzpůsobených pro různé oblasti využití. Pro celogenomové *de novo* sekvenování malých genomů doporučuje například platformu MiSeq, která umožňuje oproti ostatním sekvenátorům této společnosti číst delší úseky DNA, a pro sekvenování velkých genomů platformu NextSeq, HiSeq nebo NovaSeq (online zdroj č. 3). Srovnání různých platforem je věnována kapitola níže.

3.1.4 Sekvenování ligací SOLiD (Applied Biosystems)

Jak název předesílá, technika představená článkem Shendure et al. (2005) představuje odlišný přístup k sekvenování než všechny výše zmíněné metody. I tato technika však nabízí možnost masivní paralelizace a řadí se mezi metody sekvenování nové generace.

Metoda kombinuje emulzní PCR popsanou Dressman et al. (2003) se sekvenováním DNA pomocí ligace oligonukleotidů (Macevicz, 1998). Vzhledem k tomu, že se tato technika v současnosti již nevyužívá, nebudou zde detaily amplifikace fragmentů DNA a jejich sekvenace podrobněji popisovány.

3.1.5 Ion Torrent sekvenování (Life Technologies)

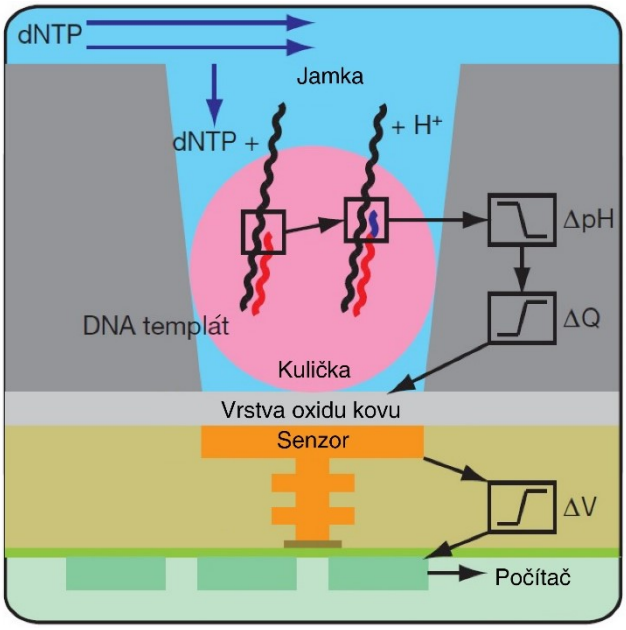
Technika sekvenování druhé generace představená článkem Rothberg et al. (2011) představuje první metodu sekvenování nevyužívající k rozpoznání bází detekci světla. Metoda k amplifikaci fragmentů DNA opět využívá emulzní PCR a svým principem spadá do kategorie sekvenování syntézou. Nejde však již o detekci světelného signálu. Namísto něj se detekuje změna pH vyvolaná uvolněním vodíkového iontu při inkorporaci dNTP DNA polymerázou.

Pro detekci je využito iontově senzitivních tranzistorů s efektem pole (ISFET) integrovaných na CMOS čipu. Sekvenování se opět odehrává v mikrojamkách – jejich dna jsou však tentokrát osazena vrstvou oxidu kovu, jejíž povrchový potenciál se změní spolu se změnou pH roztoku v jamce (Obr. 2). To indukuje změnu napětí na senzoru tranzistoru, který leží pod touto vrstvou (Rothberg et al., 2011). Ta je následně digitalizována a zaznamenána počítačem.

Postup sekvenování je tedy následující. DNA je fragmentována a zmnožena na kuličkách pomocí emulzní PCR. Kuličky jsou následně jednotlivě umístěny do jamek překrývajících čip.

Všechny jamky jsou následně zaplaveny jedním ze čtyř dNTP. Pokud je dNTP přítomný v jamce komplementární k bázi templátu přímo následující primeru, je DNA polymerázou zařazen do syntetizovaného vlákna a je uvolněn vodíkový iont (Rothberg et al., 2011). To změní pH roztoku v jamce a tato změna je detekována, jak už bylo popsáno výše. Zbylé nukleotidy jsou z jamky následně odmyty a nahrazeny dalším druhem dNTP.

Pokud se v sekvenci vyskytuje homopolymer některého z nukleotidů, je DNA polymerázou inkorporováno více dNTP během jednoho zaplavení jamky. Změna pH je poté přímo úměrná počtu zařazených nukleotidů (0,02 za jednu inkorporovanou bázi) (Rothberg et al., 2011), a tak je možné sekvenovat i tyto homopolymery.



Obr. 2: Uspořádání jamky při Ion Torrent sekvenování.
Kulička s DNA templátem je umístěna v jamce, pod kterou se nachází senzor. Uvolňování protonů při prodlužování DNA vlákna mění pH v jamce. Tato změna indukuje změnu v povrchovém potenciálu (Q) vrstvy oxidu kovu, a tak i změnu napětí (V) na příslušném tranzistoru s efektem pole. (Rothberg et al., 2011; převzato a upraveno)

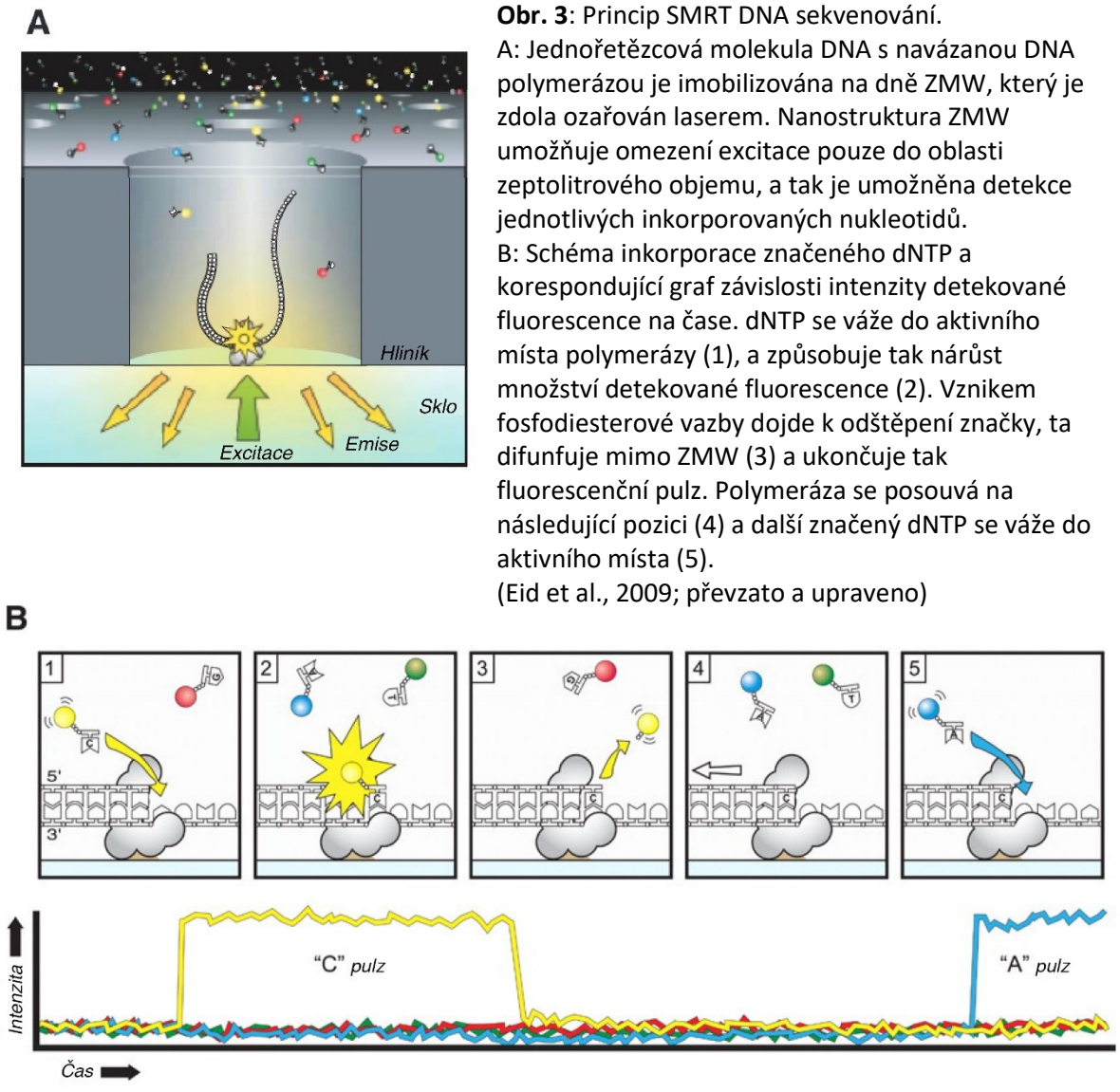
3.1.6 Single Molecule Real-Time sekvenování (Pacific Biosciences)

Single Molecule Real-Time (SMRT) sekvenování se řadí mezi techniky sekvenování třetí generace, pro které je charakteristické čtení sekvence jednotlivých molekul bez nutnosti předchozí amplifikace. SMRT sekvenování umožňuje sledování syntézy DNA v reálném čase, bez nutnosti průběžného odmývání a opětovného přidávání ligandů.

Stěžejním postupem techniky je využití takzvaných *zero-mode waveguides* (ZMW), ve kterých se odehrávají sekvenační reakce. ZMW fungují jako malé jamky a umožňují sledování jednotlivých DNA polymeráz v roztocích o zeptolitrových (1 zl = 10^{-21} l) objemech se zachováním přirozenějších (mikromolárních) koncentrací ligandů (Levene et al., 2003).

ZMW je zdola ozařován světlem o takové vlnové délce, která mu neumožní procházet až na povrch čipu (*SMRT cell*) s desítkami tisíc takových jamek. Signál z každé jamky je zcela izolován od signálu z jamek ostatních a je možné ho ze spodní strany ZMW detekovat a pozorovat tak činnost DNA polymerázy v této jamce. Takto je tedy umožněna i paralelizace procesu (Eid et al., 2009), které bylo dosahováno už metodami sekvenování druhé generace.

Dalšími stěžejními body této techniky jsou povrchová úprava dna ZMW a příprava fluorescenčně značených deoxyribonukleotidů. Povrchová úprava dna ZMW umožňuje ukotvení DNA polymerázy při zachování její aktivity a znemožňuje vazbu deoxyribonukleotidů (Eid et al., 2009). Deoxyribonukleotidtrifosfáty jsou fluorescenčně značeny na terminálním fosfátu (Kumar et al., 2005). Při vzniku fosfodiesterové vazby je tak značka odštěpena a syntetizované vlákno DNA je již nemodifikované.



Obr. 3: Princip SMRT DNA sekvenování.
A: Jednořetězcová molekula DNA s navázanou DNA polymerázou je imobilizována na dně ZMW, který je zdola ozařován laserem. Nanostruktura ZMW umožňuje omezení excitace pouze do oblasti zeptolitrového objemu, a tak je umožněna detekce jednotlivých inkorporovaných nukleotidů.
B: Schéma inkorporace značeného dNTP a korespondující graf závislosti intenzity detekované fluorescence na čase. dNTP se váže do aktivního místa polymerázy (1), a způsobuje tak nárůst množství detekované fluorescence (2). Vznikem fosfodiesterové vazby dojde k odštěpení značky, ta difunfuje mimo ZMW (3) a ukončuje tak fluorescenční pulz. Polymeráza se posouvá na následující pozici (4) a další značený dNTP se váže do aktivního místa (5).
(Eid et al., 2009; převzato a upraveno)

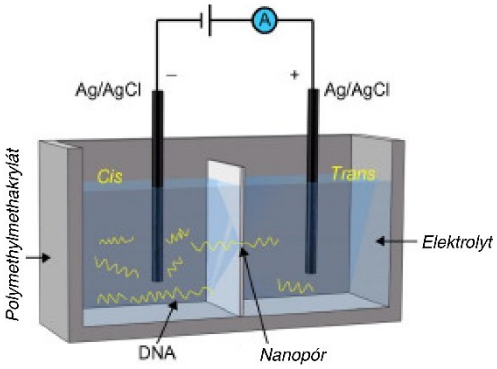
Sekvenování tedy probíhá následovně. DNA polymeráza s jednořetězcovou DNA jako templátem připevněná na dně ZMW syntetizuje komplementární vlákno (Obr. 3A). Fluorescence, pomocí které lze následně určit inkorporovanou bázi, vzniká tak, že polymeráza váže komplementární nukleotid v aktivním místě, a přidržuje ho tak v ozařované oblasti (Obr. 3B; Eid et al., 2009). Když je nukleotid zařazen, značka je spolu s pyrofosfátem odštěpena a difunduje z detekované oblasti. Tím pulz fluorescence pro tuto zařazenou bázi končí.

Zbývá vysvětlit, proč není detekována fluorescence z ostatních značených deoxyribonukleotidtrifosfátů přítomných ve směsi. Ta se projevuje pouze jako signál na pozadí, jelikož délka jejího trvání je výrazně kratší než čas, po který je fluoroforová značka přidržována v detekované oblasti DNA polymerázou (Eid et al., 2009).

3.1.7 Sekvenování nanopórem (Oxford Nanopore Technologies)

Dodnes nejnovější metodou sekvenování je sekvenování nanopórem, které se řadí mezi metody sekvenování třetí generace. První data ze sekvenování nanopórem byla veřejnosti představena roku 2012 (Eisenstein, 2012). Technika sekvenování jako taková však byla navržena mnohem dříve, konkrétně článkem Kasianowicz et al. (1996).

Článek Kasianowicz et al. (1996) představil schéma dvou komor naplněných elektrolytem oddělených nepropustnou membránou osazenou alfa-hemolysinovými transmembránovými kanály k měření délky nukleotidového polymeru. Kanál o průměru 2,6 nm umožňuje vedle průchodu iontů také průchod jednořetězcovým molekulám DNA nebo RNA. Pokud tuto soustavu vystavíme elektrickému poli, můžeme zde měřit proud iontů procházejících kanálem. Ve chvíli, kdy je kanál zablokován nukleovou kyselinou, dojde k poklesu měřeného signálu na dobu úměrnou délce polymeru (Kasianowicz et al., 1996). Pokles signálu by navíc mohl reflektovat i velikost a chemické vlastnosti jednotlivých nukleotidů v polymeru, díky čemuž by mohlo být možné získat jeho sekvenci (Kasianowicz et al., 1996).



Obr. 4: Schéma uspořádání sekvenační komory při sekvenování nanopórem.
Komora je membránou přepažena na dvě části, jediný průchod mezi těmito částmi představuje nanopór. Komory jsou naplněny elektrolytem a každá je osazena elektrodou připojenou ke zdroji napětí.
(Feng et al., 2015; převzato a upraveno)

Realizace takového způsobu sekvenování na úsecích známé sekvence DNA se podařila Manrao et al. (2012). Byl použit jediný nanopór MspA zmutovaný tak, aby umožňoval průchod DNA. Ten byl zasazen v lipidové dvojvrstvě oddělující dvě komory (cis a trans) s roztokem KCl a na trans stranu dvojvrstvy bylo aplikováno kladné napětí (Obr. 4). Průchod DNA nanopórem kontrolovala phi29 DNA polymeráza (DNAP) (Manrao et al., 2012), která mimo 5‘-3‘ polymerázové aktivity a 3‘-5‘ exonukleázové aktivity disponuje i aktivitou pro 3‘-5‘ rozplétání DNA dvoušroubovice (Salas et al., 2007). Úseky jednořetězcového DNA templátu byly na 3‘ konci opatřeny vlásenkou a byl k nim připojen komplementární blokující oligomer. Takto upravená vlákna templátové DNA byla přidána do cis komory společně s DNAP a deoxyribonukleotidy (sloužícími jako substráty pro syntézu DNA). DNA templát se navázal jako substrát na DNAP svým 5‘ koncem. Měření proudu iontů skrz nanopór pak vykazovalo symetricky opakující se úrovně signálu odpovídající signálu jednotlivých nukleotidů vlákna DNA procházejícího do trans komory přes nanopór svým 5‘ koncem (pohyb kontrolovaný rozbalováním komplexu DNA templátu a blokujícího oligomeru DNA polymerázou) a následně vracejícího se zpět do cis komory (pohyb zapříčiněný syntézou komplementárního vlákna DNA polymerázou) (Manrao et al., 2012). Článek nastínil, že pro přesné čtení neznámých sekvencí DNA bude pravděpodobně třeba naměřit mapu signálů všech čtveřic nukleotidů (protože měřený signál ovlivňuje kromě nukleotidu nacházejícího se přesně v nejužším místě nanopóru také jeho okolí – Manrao et al., 2011). Taková mapa byla představena Laszlo et al. (2014).

Sekvenátory provádějící sekvenování touto technikou uvedla na trh firma Oxford Nanopore Technologies (ONT). Prvním z těchto sekvenátorů byl roku 2014 MinION. Dnešní generace sekvenátorů společnosti ONT využívají nanopór CsgG (nebo pod jejich firemním názvem R9) (Lu et al., 2016) a pro kontrolu průchodu DNA řetězce nanopórem helikázy (Bowen et al., 2017). Krok syntézy komplementárního vlákna DNA byl tedy z procesu vyřazen a nahrazen sekvenováním 1D nebo $1D^2$. Namísto jednořetězcové DNA je do cis komory přidána klasická dvouřetězcová DNA. U 1D sekvenování je helikázou navázáno a nanopórem provlékáno na trans stranu membrány pouze templátové vlákno, zatímco komplementární zůstává na cis straně. U $1D^2$ sekvenování je komplementární vlákno pomocí adaptoru navázáno na membráně poblíž nanopóru a po dokončení průchodu templátu ho následuje směrem do trans komory. Tím z jedné dvoušroubovice získáme dvě symetrická komplementární čtení, díky čemuž lze opravovat některé chyby vznikající při sekvenování (de Lannoy et al., 2017).

Paralelizace procesu běžná i u dřívějších generací sekvenátorů je samozřejmostí i pro sekvenátory ONT. Například zmíněný MinION, jakožto doposud nejmenší z rodiny sekvenátorů společnosti, umožňuje simultánní sekvenování až 512 DNA molekul najednou (Ip et al., 2015).

Sekvenování nanopórem je stále poměrně nová a vyvíjející se záležitost, ať už mluvíme o nanopórech jako takových nebo i algoritmech dekódujících sekvenci z odečteného signálu. Je to na jednu stranu technika finančně velmi dostupná, na druhou stranu se stále dost vysokou mírou chybovosti (de Lannoy et al., 2017).

3.2 Srovnání sekvenačních technik druhé a třetí generace

Odlišnosti mezi principem sekvenování jednotlivých společností vedou i k rozdílům týkajícím se délky získaných sekvencí, chybovosti, ceny nebo časové náročnosti procesu (Tabulka 1). Trendem dnešního celogenomového sekvenování je kombinace více metod sekvenování za účelem zvýšení kontinuity výsledné sestavené sekvence. Často se jedná o kombinaci Illumina sekvenování generujícího krátká čtení s další technikou, kterou může být buď sekvenování na úrovni celých molekul (tedy SMRT sekvenování nebo sekvenování nanopórem) nebo některá z technik popsaná v následující kapitole (Phillippy, 2017).

Kombinace Illumina sekvenování a SMRT sekvenování přináší výhodu především při překonávání dlouhých repetitivních úseků sekvence. Jak je vidět z Tabulky 1, čtení produkovaná Illumina platformami jsou podstatně kratší než ta z Pacific Biosciences sekvenátorů. Zároveň je ale Pacific Biosciences sekvenování méně přesné a také dražší než Illumina. Pokud k sestavení připravíme čtení z Illumina sekvenování pokrývající genom do velké hloubky a zároveň čtení z Pacific Biosciences sekvenování se střední hloubkou pokrytí, docílíme přesnější a souvislejší sestavené sekvence (Zimin et al., 2017).

Tabulka 1: Srovnání sekvenátorů používaných k celogenomovému sekvenování.

Sekvenátor (technika sekvenování)	Délka čtení	Množství výsledných dat z jednoho běhu	Čas běhu	Počet čtení na běh	Chybovost
MiSeq (Illumina sekvenování)	až 2x300 bp[1,2]	13,2–15 Gb[1,2]	přibližně 56 hodin[1,2,10]	22–25 milionů[1,2]	0,1 %[1,3]
HiSeq 4000 (Illumina sekvenování)	až 2x150 bp[4]	1300–1500 Gb[4,5]	24–84 hodin[4,10]	4,3–5 miliard[4]	0,1 %[3]
Ion GeneStudio S5 Prime System (Ion Torrent sekvenování)	až 600 bp[6,7,8]	0,5–1,5 Gb[6,8] nebo 1,5–4,5 Gb[7,8]	5,5 hodiny[6,8,9] nebo 7 hodin[7,8,9]	3–4 miliony[6,8] nebo 9–12 milionů[7,8]	< 1 %[9]
PacBio Sequel (SMRT sekvenování)	průměrně > 15 000 bp[11]	až 10 Gb[11,12]	0,5–10 hodin[3]	přibližně 400 000[11,12]	13–15 %[3]
MinION Mk 1B (Sekvenování nanopórem)	rovná délce fragmentu, nejdelší 1 Mb[13]	10–20 Gb[13]	1 minuta–48 hodin[13,14]	-	> 1 %[13]

[1] Při použití Miseq Reagent Kit v3.
[2] online zdroj č. 2
[3] Ardui *et al.*, 2018
[4] online zdroj č. 1
[5] Při maximální délce čtení.
[6] Při využití čipu Ion 520.
[7] Při využití čipu Ion 530.
[8] online zdroj č. 6
[9] Čas zahrnuje sekvenování (2,5–4 hodiny) a následnou analýzu.
[10] Čas včetně amplifikace.
[11] online zdroj č. 5
[12] Na každý čip.
[13] online zdroj č. 4
[14] Data jsou získávána hned od začátku běhu.

3.3 Sekvenování usnadňující sestavení genomu

Pro zjednodušení následného celogenomového sestavení byla vyvinuta některá vylepšení *shotgun* sekvenování, která poskytnou kromě informace o sekvenci také informaci o tom, kam mají být v rámci genomu jednotlivá čtení umístěna.

3.3.1 *Paired-end* a *mate-pair* sekvenování

Paired-end sekvenování znamená sekvenování dvouvláknového fragmentu DNA z obou stran, tedy sekvenování *forward* i *reverse* vlákna (Ekblom & Wolf, 2014). Sekvenování druhé generace často využívá DNA polymerázu, která nové vlákno, jehož sekvenci čteme, syntetizuje ve směru od 5‘ k 3‘ konci. Zároveň využívá připevnění templátového vlákna k pevné fázi pomocí adaptéru na jeho konci. Pokud tedy zajistíme, abychom po dokončení sekvenování jednoho vlákna fragmentu byli schopní stejným způsobem na pevnou fázi zachytit konec druhého vlákna, můžeme z každého DNA fragmentu získat sekvenci jeho 5‘ konce v rámci *forward* vlákna i v rámci *reverse* vlákna, a zároveň tak ohraničit inzert známé délky mezi nimi (tedy délky fragmentu).

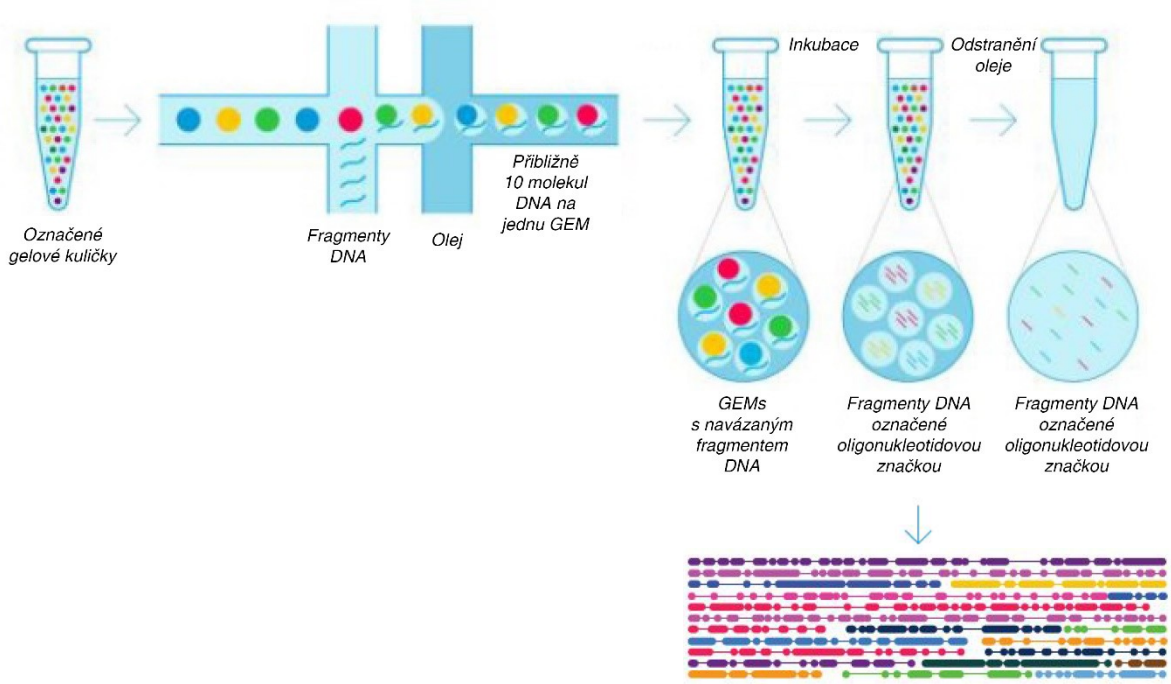
Takový postup je možný s fragmenty o velikosti několik stovek bází (Ekblom & Wolf, 2014). Pokud bychom chtěli ohraničit inzerty delší než tisíc bází, můžeme to udělat pomocí *paired-end* sekvenování dlouhých inzertů neboli takzvaného *mate-pair* sekvenování. Fragment požadované délky cirkularizujeme a pomocí restrikčních endonukleáz, které neštěpí úsek propojení obou konců fragmentu, rozstříháme na kratší úseky (Chen et al., 2009). Ze směsi fragmentů vyizolujeme úsek obsahující propojení obou konců a ten následně sekvenujeme stejně, jako bylo popsáno u klasického *paired-end* sekvenování. Odlišnost přichází v umístění těchto čtení v rámci *forward* a *reverse* vlákna. Kvůli cirkularizaci totiž namísto sekvence 5‘ konců fragmentu dostáváme sekvenci 3‘ konců (Chen et al., 2009).

3.3.2 *Linked-reads* sekvenování (10x Genomics)

Linked-reads sekvenování zjednodušuje sestavení sekvence tak, že ještě před sekvenováním všechny kratší úseky pocházející z jedné delší molekuly DNA označí stejnou značkou. Po osekvenování tak lze čtení z jedné delší molekuly sdružit, což výrazným způsobem usnadní sestavení sekvence (Obr. 5).

Nutností je ovšem izolovat DNA ve formě velmi dlouhých, nepřerušených vláken. Fragmenty DNA o průměrné délce větší než 50 kb (Weisenfeld et al., 2017) jsou spolu s gelovými kuličkami obsahujícími oligonukleotidové značky vloženy do oleje a vytváří takzvané *Gel-beads in emulsion* (GEM). V každé GEM je několik molekul DNA, GEM kuliček

je okolo jednoho milionu a každá z nich obsahuje mnoho kopií unikátní značky o délce 16 bází, která je specifická jen pro tuto jednu kuličku (Weisenfeld et al., 2017). Uvnitř GEM jsou z delších molekul generovány krátké fragmenty označené oligonukleotidovou značkou a primerem (Obr. 5). Část z těchto krátkých fragmentů bude následně použita k sekvenování pomocí některého Illumina sekvenátoru. Po osekvenování můžeme čtení, která původně patřila k jednomu dlouhému fragmentu DNA, rozlišit podle sekvence oligonukleotidové značky.

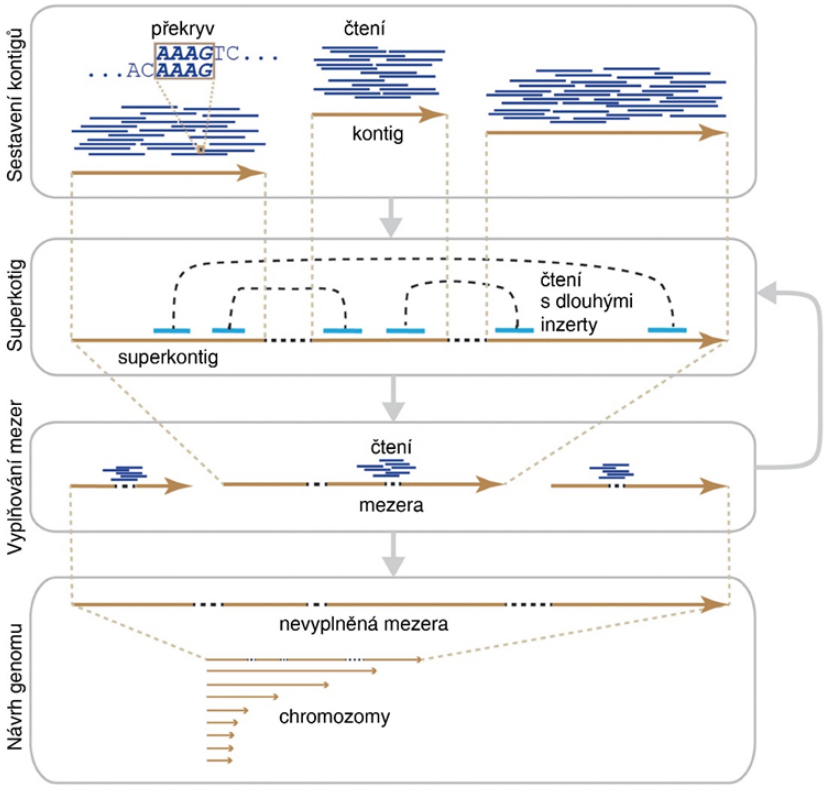


Obr. 5: Generování *linked-reads*.
DNA a označené kuličky tvoří GEMs, inkubací vznikají fragmenty DNA. Všechny fragmenty DNA ve stejné GEM mají shodné značky, což umožňuje znovurozdělení do těchto skupin i po osekvenování.
(Weisenfeld et al., 2017; převzato a upraveno)

4. *De novo* sestavení genomu

Výsledkem sekvenování genomové DNA je soubor krátkých úseků sekvence DNA (čtení) zkoumaného organismu. Ty mohou být různě dlouhé a také různě přesné podle toho, jakou metodou byla DNA sekvenována (Tabulka 1). Každou pozici zatím teoretické sekvence genomu máme v ideálním případě pokrytou více čteními a čtení se vzájemně překrývají (koncová část jednoho čtení je shodná s počáteční částí čtení druhého). Úkolem *de novo* celogenomového sestavení je rekonstruovat z těchto krátkých úseků kompletní sekvenci DNA rozdělenou do jednotlivých chromozomů daného organismu.

De novo celogenomové sestavení lze rozdělit do tří kroků (Obr. 6). V prvním kroku jsou z krátkých čtení budovány o něco delší úseky, takzvané kontigy, což zahrnuje zarovnání sekvencí čtení a vytvoření jejich konsensus sekvence (Miller et al., 2010), ve které se nevyskytují mezery. Výsledkem je obvykle několik stovek až desítek tisíc delších kontigů, jejichž vzájemnou pozici v genomu však neznáme. V druhém kroku jsou kontigy spojovány do delších superkontigů pomocí dlouhých čtení získaných například *mate-pair* sekvenováním. Pomocí v tomto kroku může být i známá genetická mapa studovaného organismu. Tím jsou kontigy uspořádány na chromosomy, ale mohou mezi nimi zůstat mezery. Ty jsou potom vyplněny ve třetím kroku pomocí dalších nezávislých čtení (Sohn & Nam, 2018).



Obr. 6: Obecný postup *de novo* celogenomového sestavení. Z krátkých čtení jsou pomocí překryvů skládány kontigy, které jsou následně pomocí informací z čtení s dlouhými inzerty (čtení z *paired-end* sekvenování) sestavovány do superkontigů. Dále jsou pomocí nezávislých čtení zaplněny zbývající mezery a je vytvořen návrh výsledného genomu. Některé mezery mohou zůstat nevyplněné. (Sohn & Nam, 2016; převzato a upraveno)

Algoritmy *de novo* sestavení genomu můžeme rozdělit do dvou kategorií. První kategorií jsou algoritmy využívající přístup *overlap/layout/consensus*, algoritmy druhé kategorie využívají k sestavení sekvence de Bruijnových grafů. Problém sestavení sekvence řeší algoritmy obou kategorií převedením na problém redukce grafu (Miller et al., 2010). Existuje několik modelů, které lze pro tuto abstrakci využít (budou popsány v následujících kapitolách). Ve všech obvykle využívaných je problém sestavení sekvence NP-těžký (Medvedev et al., 2007), a tedy je nutné při řešení využít nejrůznějších heuristik.

To, že je problém NP-těžký, znamená, že je to problém, na který je převoditelný kterýkoliv problém z třídy NP. Třída NP je tvořena rozhodovacími problémy, jejichž řešení je možné v polynomiálním čase ověřit, ale obecně ne deterministickými algoritmy najít. Problémy, které leží ve třídě NP a zároveň jsou NP-těžké označujeme jako NP-úplné.

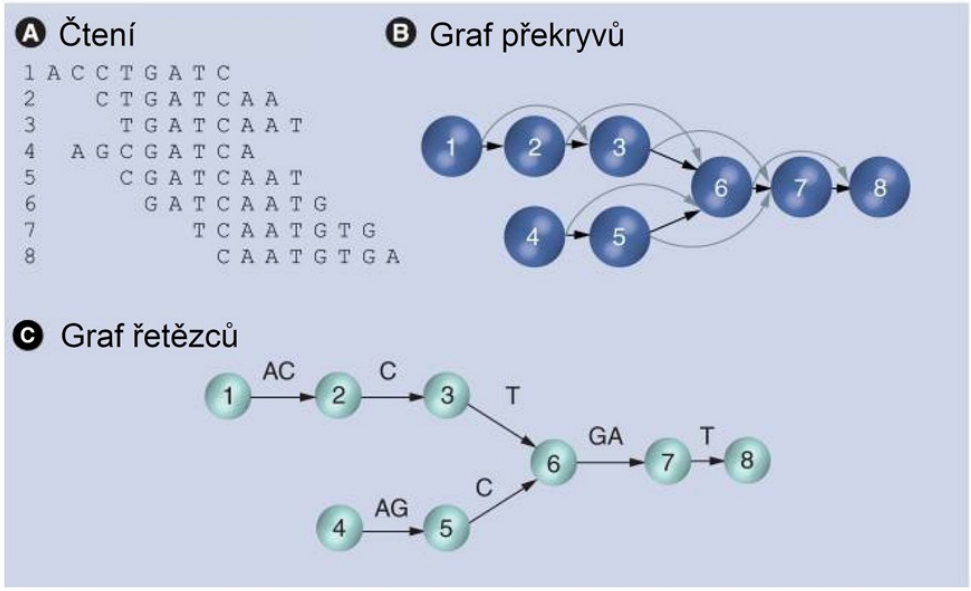
Programy celogenomového sestavení implementují buď jeden z výše uvedených algoritmických přístupů, nebo oba nějakým způsobem kombinují (takový přístup označujeme jako hybridní). Jednotlivé přístupy budou podrobněji představeny v následujících kapitolách.

4.1 *Overlap/Layout/Consensus* přístup k sestavení sekvence

Overlap/Layout/Consensus (OLC) přístup byl využíván programy pro sestavení sekvenačních dat ze Sangerova sekvenování (Miller et al., 2010) a v současnosti je často implementován programy pro sestavení dlouhých čtení (Sohn & Nam, 2018). K sestavení genomové sekvence využívá celých čtení a jejich překryvů, které obvykle modeluje pomocí grafu překryvů nebo grafu řetězců a jeho modifikací. Limitací tohoto přístupu je výpočetní náročnost vyhledávání překryvů mezi čteními a nároky na paměť, což jsou i důvody, proč tyto algoritmy fungují lépe pro dlouhá čtení s relativně nižší hloubkou pokrytí než pro krátká čtení s velkou hloubkou pokrytí, jejichž vysoký počet implikuje větší počet párových sekvenčních zarovnání nutných k sestavení grafu a vyšší komplexitu grafu (Sohn & Nam, 2018).

Vrcholy grafu překryvů (*overlap graph*, Obr. 7B) reprezentují jednotlivé úseky sekvence (čtení), hrany potom překryvy mezi nimi. Tyto překryvy je nutné předpočítat pomocí párového zarovnání (Myers, 1995). Cesty v tomto grafu reprezentují potenciální kontigy.

Graf řetězců (*string graph*, Obr. 7C) se od předchozího grafu překryvů se liší tím, že neuvažuje vrcholy reprezentující čtení, která vytváří s jiným čtením překryv o maximální délce (Myers, 2005). Jeho vrcholy reprezentují začátky a konce jednotlivých čtení a hrany překryvy mezi těmito čteními. Hrany mezi vrcholy jsou orientované a označené začátkem sekvence čtení reprezentovaného počátečním vrcholem této hrany, který nevytváří překryv



Obr. 7: Graf překryvů a graf řetězců.
Reprezentace množiny osmi čtení (jejichž zarovnání je zobrazeno v A) grafem překryvů (B) a grafem řetězců (C). V grafu překryvů (B) vrcholy odpovídají jednotlivým čtením a hrany překryvům mezi nimi o délce alespoň pěti bází. V grafu řetězců (C) je topologie stejná jako v grafu překryvů (až na tranzitivní hrany), ale vrcholy reprezentují začátky čtení a hrany reprezentují sekvenci, která dvě čtení spojená hranou odlišuje.
(Henson et al., 2012; převzato a upraveno)

se sekvencí čtení reprezentovaného koncovým vrcholem této hrany (Myers, 2005). Graf neobsahuje tranzitivní hrany, tedy pokud čtení *u* vytváří překryv se čtením *v* a zároveň i se čtením *w* a čtení *v* vytváří překryv se čtením *w*, pak graf neobsahuje hranu mezi vrcholy reprezentujícími čtení *u* a *w* (Henson et al., 2012).

Celý proces se při OLC přístupu k sestavení sekvence skládá ze tří kroků. V prvním kroku je pomocí překryvů vybudován graf. Ve druhém kroku jsou v grafu vyhledávány takové cesty, které odpovídají částem genomové sekvence. V ideálním případě by měla být nalezena hamiltonovská cesta grafem, tedy cesta, která prochází všemi vrcholy grafu právě jednou (Pop, 2009). Její nalezení obecně spadá do kategorie NP-úplných problémů, tedy jednotlivé programy využívají pro tento krok různých heuristik. Třetím krokem je určení konsensus sekvence DNA ze čtení uspořádaných podél nalezené cesty (Pop, 2009).

Do této kategorie patří například program Celera, který byl původně vytvořen pro sestavování čtení ze Sangerova sekvenování (Miller et al., 2010). Ze souboru všech čtení nejprve filtruje známé repetitivní úseky a poté buduje graf překryvů pomocí algoritmu podobného BLASTu (Myers et al., 2000). V grafu překryvů pak vytváří takzvané *unitigy*. *Unitig* (*uniquely assemblable contig*) je soubor fragmentů, jejichž uspořádání je očividné – tedy

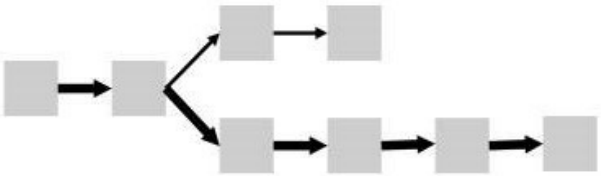
jeho sekvence je tvořena cestou v grafu překryvů, jejíž všechny vnitřní vrcholy jsou stupně dva a koncové vrcholy stupně většího než dva. O každém takovém *unitigu* je rozhodnuto, zda se jedná o unikátní nebo o repetitivní úsek sekvence. Na koncích unikátních *unitigů* se nachází potenciální hranice opakujících se úseků sekvence (Myers et al., 2000). Unikátní *unitigy* jsou dále pomocí dat získaných z *mate-pair* sekvenování spojovány do superkontigů (Myers et al., 2000). Program Celera byl následně upraven pro data z 454 sekvenování, a tak vznikl algoritmus CABOG (Miller et al., 2008).

Dalším z programů implementujících OLC přístup ke genomovému sestavení je Miniasm, program pro sestavování dlouhých čtení ze třetí generace sekvenování. Ten pracuje nad sestavovacím grafem (*assembly graph*), jehož vrcholy reprezentují jednotlivá čtení a hrany překryvy mezi nimi. Sestavovací graf zároveň splňuje následující dvě podmínky: žádný z vrcholů nereprezentuje sekvenci, která by tvořila podřetězec sekvence jiného z vrcholů, a pro sekvenci každého vrcholu nebo hrany existuje v grafu vrchol nebo hrana reprezentující komplementární sekvenci (Li, 2016). Topologie grafu je shodná s grafem řetězců. Ze zkonstruovaného grafu jsou odstraněny tranzitivní hrany a prvky odpovídající svou topologií sekvenačním chybám (podrobněji popsány v kapitole Korekce sekvenačních chyb rozpoznáváním typických topologií v grafu). U větvících vrcholů jsou odstraněny hrany reprezentující ty překryvy, jejichž délka je v poměru k délce překryvu reprezentovaného jinou hranou dostatečně malá (Li, 2016). Dále jsou obdobně jako v programu Celera sestavovány *unitigy*.

Program Canu navržený pro sestavování dlouhých čtení ze třetí generace sekvenování napřed soubor čtení podrobí korekci a filtru sekvenačních chyb (Koren et al., 2017). Z takto opraveného souboru čtení je následně vybudován graf nejlepších překryvů, tedy graf, jehož vrcholy reprezentují jednotlivá čtení a hrany jejich nejlepší překryvy. Nejlepší překryv je definován pro každé čtení následovně: jedná se o nejdelší takový překryv s jiným čtením, který splňuje, že se čtení překrývají svými konci a zároveň délka tohoto překryvu není rovna délce ani jednoho ze čtení (Koren et al., 2017). Pro každé čtení existují dva nejlepší překryvy, jeden na každém konci. Následně jsou budovány kontigy obdobným způsobem, jakým je konstruuje program CABOG (Koren et al., 2017).

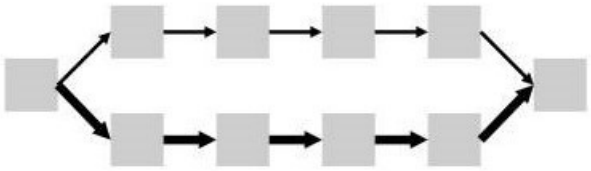
4.1.1 Korekce sekvenačních chyb rozpoznáváním typických topologií v grafu

Některé chyby sekvenování lze v sestrojeném grafu rozpoznat identifikací topologií pro ně typických. Takto lze odhalit například sekvenační chybu v koncové oblasti čtení. Projeví se jako posloupnost vrcholů, pro kterou platí, že z jednoho z jejích koncových vrcholů nevede žádná další orientovaná hrana (Zerbino & Birney, 2008), jedná se tedy o slepou cestu v grafu (Obr. 8). Pokud je délka sekvence tvořené touto posloupností vrcholů menší než nějaká předem daná hodnota a zároveň je hrana připojující tuto slepou odbočku ke zbytku grafu méně početně podpořena daty ze sekvenování než jiná hrana vedoucí z příslušného větvícího vrcholu, je tato posloupnost vrcholů z grafu odstraněna (Li et al., 2010).



Obr. 8: Slepá cesta v grafu.
Sekvenační chyba v koncové části čtení zapříčiní v grafu výskyt zobrazené topologie. V ilustraci jsou hrany zastupující větší množství čtení zakresleny silnější šipkou.
(Miller et al., 2010; převzato a upraveno)

Sekvenační chyby ve vnitřních částech čtení způsobují v grafu vznik takzvaných bublin (Obr. 9), neboli dvojic cest, které začínají a končí stejným vrcholem a mají podobné sekvence (Zerbino & Birney, 2008). Ty lze v grafu najít a odstranit například pomocí algoritmu *Tour Bus* (Zerbino & Birney, 2008), který hledá takové dvojice cest, aby následně provedl párové zarovnání jejich sekvencí. Pokud jsou tyto sekvence dostatečně podobné, jsou dvě cesty sloučeny do jedné, přičemž jako konsensus cesta je použita ta s větším pokrytím sekvenačními daty (Zerbino & Birney, 2008).



Obr. 9: Bublina v grafu.
Sekvenační chyba ve vnitřní části čtení zapříčiní v grafu výskyt zobrazené topologie. V ilustraci jsou hrany zastupující větší množství čtení zakresleny silnější šipkou.
(Miller et al., 2010; převzato a upraveno)

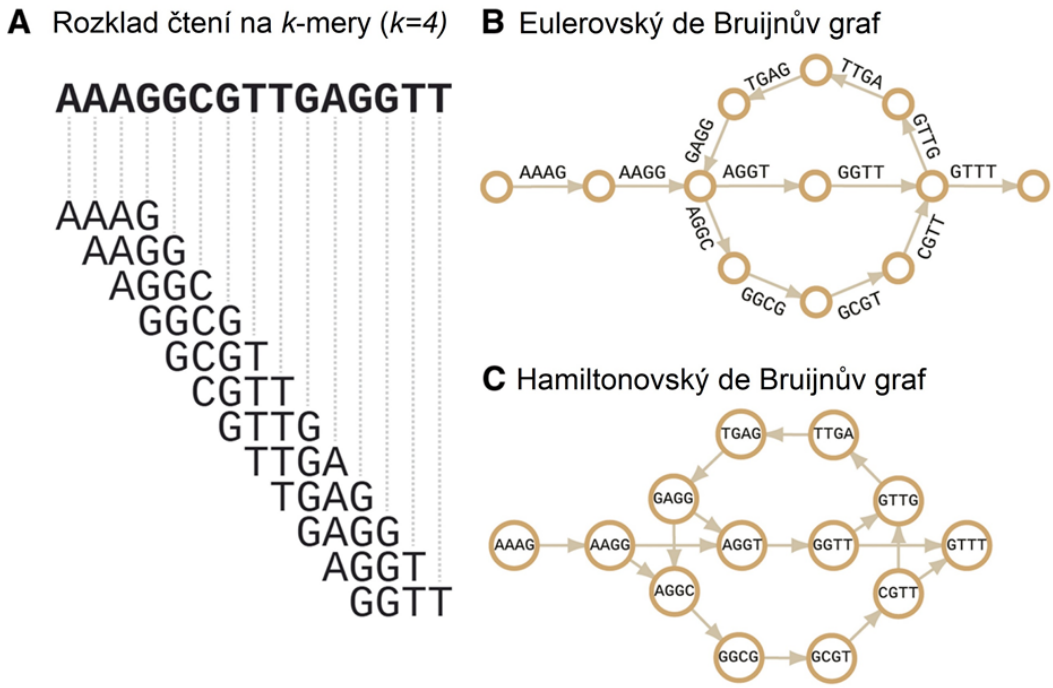
4.2 Sestavení sekvence pomocí de Bruijnových grafů

Při sestavování sekvence z krátkých čtení generovaných sekvenačními technikami druhé generace bývají obvykle využívány de Bruijnovy *k*-mer grafy (Miller et al., 2010). Ty namísto celých čtení využívají množinu všech z nich generovaných podřetězců o délce *k* (*k*-merů)

(Obr. 10A). Limitací tohoto přístupu je krátká délka *k*-merů, která komplikuje sestavení repetitivních úseků (Sohn & Nam, 2018).

V hamiltonovských de Bruijnových grafech (Obr. 10C) je pro každý *k*-mer vyskytující se v některém ze čtení vytvořen vrchol a orientovaná hrana vede z vrcholu *A* do vrcholu *B* v tom případě, že přípona délky *k−1* *k*-meru odpovídajícího vrcholu *A* je shodná s předponou délky *k−1* *k*-meru odpovídajícího vrcholu *B* (Compeau et al., 2011). Výsledná sekvence v takovém grafu by měla, obdobně jako u OLC přístupu, odpovídat hamiltonovské cestě grafem, jejíž nalezení je NP-úplný problém.

U eulerovských de Bruijnových grafů (Obr. 10B) jsou vrcholy tvořeny naopak všemi možnými různými *k−1* dlouhými předponami nebo příponami *k*-merů tak, že každou *k−1* dlouhou sekvenci může reprezentovat maximálně jeden vrchol grafu. Orientovaná hrana z vrcholu *A* do vrcholu *B* následně vede, pokud existuje *k*-mer, jehož předpona odpovídá *(k−1)*-meru, který je reprezentován vrcholem *A*, a přípona odpovídá *(k−1)*-meru, který je reprezentován vrcholem *B* (Compeau et al., 2011). Problém sestavení sekvence v takovém grafu vybudovaném nad perfektními daty potom odpovídá problému nalezení cesty procházející všemi hranami grafu, tedy nalezení eulerovského tahu (Pevzner et al., 2001), které



Obr. 10: De Bruijnovy *k*-mer grafy.
Krátká čtení jsou rozdělena na *k*-mery (A). Ty potom v hamiltonovském de Bruijnově grafu (C) reprezentují vrcholy, zatímco v eulerovském de Bruijnově grafu (B) hrany.
(Sohn & Nam, 2016; převzato a upraveno)

lze provést v lineárním čase. Data získaná sekvenováním ale obsahují chyby, a tak je i tento typ algoritmů genomového sestavení závislý na heuristikách.

Mezi často využívané programy pro sestavení genomu využívající hamiltonovské grafy patří například ABySS popsaný článkem Simpson et al. (2009). Ten aplikuje rozdělení de Bruijnova grafu mezi několik počítačů, a umožňuje tak paralelizaci celého procesu sestavení genomu. Co se samotného algoritmu týče, po sestavení grafu následuje korekce sekvenačních chyb rozpoznáváním pro ně typických topologií (podrobněji v kapitole Korekce sekvenačních chyb rozpoznáváním typických topologií v grafu). Dále jsou z grafu odstraněny všechny hrany vedoucí z větvících vrcholů. Vrcholy propojené zbylými hranami jsou následně sloučeny, a vytvoří tak kontigy (Simpson et al., 2009). K nim jsou zarovnána čtení z *paired-end* sekvenování. Pro každý kontig je tak sestrojena kolekce kontigů, které jsou s ním propojené skrze určitý počet *paired-end* čtení. V grafu je potom vyhledávána cesta začínající v tomto kontigu, která by obsahovala všechny prvky kolekce (Simpson et al., 2009).

Programem uplatňujícím k sestavení genomu eulerovské de Bruijnovy grafy byl například EULER popsaný Pevzner et al., 2001. Pracuje s množinou *k*-merů podrobenou filtru sekvenačních chyb. Vedle eulerovského grafu *G* vybuduje také kolekci cest *P* tímto grafem, které reprezentují jednotlivá čtení. Následně řeší problém eulerovského supertahu, tedy hledá takový tah eulerovským grafem, který je konzistentní s vybudovanou kolekcí cest (Pevzner et al., 2001). To uskutečňuje postupnými ekvivalentními transformacemi *G* a *P* tak, aby každá cesta v *P* byla nakonec reprezentována pouze jednou hranou v *G*. Eulerovský tah nalezený v takto transformovaném grafu bude řešením problému eulerovského supertahu (Pevzner et al., 2001), a tak i řešením problému *de novo* celogenomového sestavení.

Dalším z řady programů genomového sestavení využívající eulerovské de Bruijnovy grafy je program Velvet. Jeho implementace grafu sdružuje do jednoho vrcholu více překrývajících se *k*-merů tak, že jako sekvenci vrcholu používá sekvenci jejich posledních nukleotidů (Zerbino & Birney, 2008). Dále algoritmus uplatňuje krok zjednodušení grafu, ve kterém sloučí všechny takové dvojice vrcholů, které splňují následující podmínku: z prvního vrcholu vede pouze jedna hrana, a to právě do druhého vrcholu této dvojice, a zároveň do druhého vrcholu vede pouze tato jedna hrana (Zerbino & Birney, 2008). Následně vyhledává v grafu topologické znaky spojené s výskytem sekvenačních chyb – nejprve slepé cesty, poté graf opět zjednoduší, a dále bubliny (tyto topologie byly podrobněji popsány v kapitole Korekce sekvenačních chyb rozpoznáváním typických topologií v grafu). Odstraní z grafu všechny vrcholy, které svým pokrytím nedosahují určité uživatelem nastavené hranice (Zerbino & Birney, 2008). Pomocí

informací z *paired-end* sekvenování pak algoritmem Pebble (Zerbino et al., 2009) spojuje kontigy do superkontigů. Podle míry pokrytí jednotlivých kontigů čteními určí, zda jde o kontig unikátní či nikoliv, a následně spojuje unikátní kontigy propojené *paired-end* informací (Zerbino et al., 2009).

4.3 Hybridní sestavení genomu

Hybridní sestavení genomu kombinuje oba výše uvedené přístupy. Nejprve ze čtení nageneruje krátké *k*-mery a pomocí de Bruijnových grafů je sestaví do kontigů, čímž dojde k redukci celkového objemu dat. Potom z těchto kontigů OLC přístupem sestaví superkontigy.

Takovým způsobem pracuje například program MaSuRCA. Ten nejprve generuje ze souboru čtení všechny možné *k*-mery. Dále prodlužuje čtení o takové *k*-mery, které vytvářejí *k–1* dlouhé perfektní překryvy s *k*-mery na 3‘ nebo 5‘ konci čtení a zároveň lze tyto *k*-mery podle *k–1* dlouhé části jednoznačně určit ve vygenerovaném souboru všech *k*-merů. Tímto prodloužením každého čtení na obou koncích do maximální možné míry vzniknou takzvaná super-čtení (*super-reads*) (Zimin et al., 2013). Super-čtení by bylo možné ekvivalentně budovat i pomocí de Bruijnových grafů. Sestavení super-čtení následně probíhá OLC přístupem (modifikovanou verzí algoritmu CABOG) s pomocí dat z *mate-pair* sekvenování (Zimin et al., 2013).

Hybridní přístup k sestavení sekvence umožňuje i sestavení genomu pomocí čtení pocházejících z různých sekvenačních platforem. Konkrétně pro program MaSuRCA byla článkem Zimin et al. (2017) představena kombinace krátkých čtení z Illumina sekvenování a dlouhých čtení ze SMRT sekvenování. Super-čtení jsou sestavována výše uvedeným způsobem z krátkých čtení a následně umisťována podél dlouhých čtení na základě sekvenční podobnosti (Zimin et al., 2017). Z takto umístěných super-čtení je budován graf, jehož vrcholy reprezentují jednotlivá super-čtení a hrany překryvy mezi nimi. V něm je celogenomová sekvence sestavena opět algoritmem CABOG (Zimin et al., 2017).

4.4 Evaluace kvality sestavení genomu

Jak již bylo řečeno výše, k sestavení sekvence dochází na základě různých heuristik. Je proto žádoucí moci ohodnotit, jak dobře tyto heuristiky fungují. To je možné mnoha různými způsoby.

Intuitivním způsobem je určení procenta genomu pokrytého sestavenou sekvencí (Ekblom & Wolf, 2014). K tomu je nutné určit očekávanou velikost námi zkoumaného genomu, čehož lze dosáhnout například metodou počítání *k*-merů (Sohn & Nam, 2018).

Nejpoužívanější statistikou popisující kvalitu genomového sestavení je N50. Udává se vzhledem k délce kontigů nebo superkontigů. Hodnota N50 vzhledem ke kontigům označuje délku nejkratšího kontigu ze souboru nejdelších sestavených kontigů, jejichž suma délek tvoří dohromady polovinu délky celkově sestavené sekvence (Yandell & Ence, 2012). Hodnota N50 vzhledem ke superkontigům stejným způsobem udává délku superkontigu.

Pro ohodnocení kvality genomových sestavení byly vyvinuty různé programy, jedním z nich je například QUAST. Ten sdružuje metriky, kterými je možné kvalitu sestavení popsat. Řada z nich vyžaduje referenční genom, ale některé z nich lze stanovit i pro sestavené genomy druhů, ke kterým referenční genom dosud neexistuje (Gurevich et al., 2013). Sledovanými metrikami je například počet kontigů, délka nejdelšího kontigu a celková délka sestavené sekvence (počet bází sestaveného genomu). Dále to jsou Nx statistiky (pro x od 0 do 100), mezi které patří i výše zmíněná statistika N50, obsah bází guaninu a cytosinu v genomu nebo počet predikovaných genů (Gurevich et al., 2013).

5. Závěr

Přístupy k celogenomovému sekvenování a *de novo* sestavení genomu podléhají rychlému vývoji. Roku 2001 byla jakožto výsledek několikaletého snažení publikována první sekvence lidského genomu pořízená sestavením čtení získaných Sangerovým sekvenováním pomocí algoritmu Celera (Venter et al., 2001). Dnes lze díky paralelizaci sekvenování a neustále se zvyšující výpočetní kapacitě počítačů realizovat i velké sekvenační projekty. Takovým byl například 1000 Genomes Project, během kterého bylo mezi lety 2008 a 2015 osekvenováno a sestaveno 2504 lidských genomů (The 1000 Genomes Project Consortium, 2015). Mezi stále běžící sekvenační projekty patří například Genome 10K Project, Bird 10K Project nebo 100 000 Genomes Project.

V předchozích kapitolách bylo představeno mnoho různých možností pro sekvenování a *de novo* sestavení genomu. Vybrat z nich nejlepší kombinaci, která zaručeně povede ke spolehlivé celogenomové sekvenci, je ale složité – už jen proto, že sestavená sekvence je vždy dílem heuristických procesů.

Co se kroku sekvenování týče, mezi v současnosti často používané techniky patří sekvenování Illumina, sekvenování nanopórem a SMRT sekvenování. Výhody Illumina sekvenování spočívají především v nízké ceně a vysoké přesnosti, nevýhodu ale představuje krátká délka čtení. Využití sekvenování nanopórem je díky neustálému vylepšování softwaru vyhodnocujícího sekvence z naměřených spekter, a tedy snižující se chybovosti, na vzestupu. Moderní je také využití technologie *linked-reads*, které sice oproti samotnému Illumina sekvenování ztrácí výhodu nízké ceny, nicméně umožňuje přesnější sestavení sekvence.

Volba algoritmu pro celogenomové sestavení do značné míry souvisí s výběrem sekvenační techniky. Sestavení z dlouhých čtení (ze SMRT sekvenování nebo sekvenování nanopórem) pomocí algoritmu s OLC přístupem vyžaduje identifikaci překryvů pomocí párového alignmentu. Ten je nutné hledat pro každou dvojici čtení, a to určuje vysokou výpočetní náročnost těchto algoritmů. Výsledná sekvence je navíc výrazně poznamenána sekvenačními chybami kvůli relativně vysoké chybovosti sekvenačních platforem generujících dlouhá čtení. Sestavení z krátkých čtení (Illumina sekvenování) pomocí algoritmu využívajícího de Bruijnovy grafy může být sice vlivem různých redukcí grafu výpočetně výhodnější, ale není schopné tak dobře sestavit repetitivní úseky. Rozpor mezi těmito přístupy může být řešen programy sestavení genomu kombinujícími krátká i dlouhá čtení. Jednou variantou, jak mohou takové programy fungovat, je sestavení kontigů z krátkých čtení

a následné budování superkontigů podle dlouhých čtení. Druhou variantou je korekce dlouhých čtení pomocí krátkých čtení a následné sestavení opravených dlouhých čtení (Sohn & Nam, 2018).

Poznatky o metodách a principech celogenomového sekvenování a *de novo* sestavení genomu nabyté během psaní této práce bych ráda využila v praxi v rámci navazující diplomové práce týkající se osekvenování a *de novo* sestavení genomu slavíka obecného (*Luscinia megarhynchos*).

6. Seznam literatury

Ardui, S., Ameur, A., Vermeesch, J. R., & Hestand, M. S. (2018). Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. *Nucleic Acids Research*, *46*(5), 2159–2168.

Atkinson, M. R., Deutscher, M. P., Kornberg, A., Russell, A. F., & Moffatt, J. G. (1969). Enzymatic synthesis of deoxyribonucleic acid. XXXIV. Termination of chain growth by a 2',3'-dideoxyribonucleotide. *Biochemistry*, *8*(12), 4897–4904.

Bowen, R. V., Brown, C. G., Bruce, M., Hedon, A. J., Wallace, J. E., White, J., … Soeroes, S. (2017). U.S. Patent No. 2017/0002406 A1.

Canard, B., & Sarfati, R. S. (1994). DNA polymerase fluorescent substrates with reversible 3'-tags. *Gene*, *148*(1), 1–6.

Compeau, P. E. C., Pevzner, P. A., & Tesler, G. (2011). How to apply de Bruijn graphs to genome assembly. *Nature Biotechnology*, *29*(11), 987–991.

de Lannoy, C., de Ridder, D., & Risse, J. (2017). The long reads ahead: de novo genome assembly using the MinION. *F1000Research*, *6*, 1083.

Dressman, D., Yan, H., Traverso, G., Kinzler, K. W., & Vogelstein, B. (2003). Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(15), 8817–8822.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., … Turner, S. (2009). Real-time DNA sequencing from single polymerase molecules. *Science*, *323*(5910), 133–138.

Eisenstein, M. (2012). Oxford Nanopore announcement sets sequencing sector abuzz. *Nature Biotechnology*, *30*(4), 295–296.

Ekblom, R., & Wolf, J. B. W. (2014). A field guide to whole-genome sequencing, assembly and annotation. *Evolutionary Applications*, *7*(9), 1026–1042.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., … Merrick, J. M. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, *269*(5223), 496–512.

Genome 10K Community of Scientists. (2009). Genome 10K: a proposal to obtain whole-genome sequence for 10 000 vertebrate species. *Journal of Heredity*, *100*(6), 659–674.

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072–1075.

Heather, J. M., & Chain, B. (2016). The sequence of sequencers: the history of sequencing DNA. *Genomics*, *107*(1), 1–8.

Henson, J., Tischler, G., & Ning, Z. (2012). Next-generation sequencing and large genome assemblies. *Pharmacogenomics*, *13*(8), 901–915.

Chakraborty, M., Baldwin-Brown, J. G., Long, A. D., & Emerson, J. J. (2016). Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Research*, *44*(19), e147.

Chen, Z., Godwin, B. C., Ferreri, G. C., & Riches, D. R. (2009). U.S. Patent No. 2009/0233291 A1.

Ip, C. L. C., Loose, M., Tyson, J. R., de Cesare, M., Brown, B. L., Jain, M., … MinION Analysis and Reference Consortium. (2015). MinION Analysis and Reference Consortium: Phase 1 data release and analysis. *F1000Research*, *4*, 1075.

Jain, M., Koren, S., Miga, K. H., Quick, J., Rand, A. C., Sasani, T. A., … Loose, M. (2018). Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nature Biotechnology*, *36*(4), 338–345.

Johnson, L. A., & Ferris, J. A. J. (2002). Analysis of postmortem DNA degradation by single-cell gel electrophoresis. *Forensic Science International*, *126*(1), 43–47.

Kasianowicz, J. J., Brandin, E., Branton, D., & Deamer, D. W. (1996). Characterization of individual polynucleotide molecules using a membrane channel. *Proceedings of the National Academy of Sciences of the United States of America*, *93*(24), 13770–13773.

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, *27*(5), 722–736.

Kumar, S., Sood, A., Wegener, J., Finn, P. J., Nampalli, S., Nelson, J. R., … Fuller, C. W. (2005). Terminal phosphate labeled nucleotides: synthesis, applications, and linker effect on incorporation by DNA polymerases. *Nucleosides, Nucleotides & Nucleic Acids*, *24*(5–7), 401–408.

Laszlo, A. H., Derrington, I. M., Ross, B. C., Brinkerhoff, H., Adey, A., Nova, I. C., … Gundlach, J. H. (2014). Decoding long nanopore sequencing reads of natural DNA. *Nature Biotechnology*, *32*(8), 829–833.

Levene, M. J., Korlach, J., Turner, S. W., Foquet, M., Craighead, H. G., & Webb, W. W. (2003). Zero-mode waveguides for single-molecule analysis at high concentrations. *Science*, *299*(5607), 682–686.

Li, H. (2016). Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences. *Bioinformatics (Oxford, England)*, *32*(14), 2103–2110.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., … Wang, J. (2010). De novo assembly of human genomes with massively parallel short read sequencing. *Genome Research*, *20*(2), 265–272.

Lischer, H. E. L., & Shimizu, K. K. (2017). Reference-guided de novo assembly approach improves genome reconstruction for related species. *BMC Bioinformatics*, *18*(1), 474.

Lu, H., Giordano, F., & Ning, Z. (2016). Oxford Nanopore MinION Sequencing and Genome Assembly. *Genomics, Proteomics & Bioinformatics*, *14*(5), 265–279.

Luckey, J. A., Drossman, H., Kostichka, A. J., Mead, D. A., D'Cunha, J., Norris, T. B., & Smith, L. M. (1990). High speed DNA sequencing by capillary electrophoresis. *Nucleic Acids Research*, *18*(15), 4417–4421.

Macevicz, S. C. (1998). U.S. Patent No. 5750341.

Manrao, E. A., Derrington, I. M., Laszlo, A. H., Langford, K. W., Hopper, M. K., Gillgren, N., … Gundlach, J. H. (2012). Reading DNA at single-nucleotide resolution with a mutant MspA nanopore and phi29 DNA polymerase. *Nature Biotechnology*, *30*(4), 349–353.

Manrao, E. A., Derrington, I. M., Pavlenok, M., Niederweis, M., & Gundlach, J. H. (2011). Nucleotide discrimination with DNA immobilized in the MspA nanopore. *PLoS ONE*, *6*(10), e25723.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., … Rothberg, J. M. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, *437*(7057), 376–380.

Mayer, P. (2004). U.S. Patent No. 2004/0096853 A1.

Medvedev P., Georgiou K., Myers G., Brudno M. (2007). Computability of models for sequence assembly. *Lecture Notes in Computer Science*, *4645*, 289–301.

Miller, J. R., Delcher, A. L., Koren, S., Venter, E., Walenz, B. P., Brownley, A., … Sutton, G. (2008). Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics*, *24*(24), 2818–2824.

Miller, J. R., Koren, S., & Sutton, G. (2010). Assembly algorithms for next-generation sequencing data. *Genomics*, *95*(6), 315–327.

Myers, E. W. (1995). Toward simplifying and accurately formulating fragment assembly. *Journal of Computational Biology*, *2*(2), 275–290.

Myers, E. W. (2005). The fragment assembly string graph. *Bioinformatics*, *21*(Suppl 2), ii79-ii85.

Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., … Venter, J. C. (2000). A whole-genome assembly of Drosophila. *Science*, *287*(5461), 2196–2204.

Nyrén, P. (1987). Enzymatic method for continuous monitoring of DNA polymerase activity. *Analytical Biochemistry*, *167*(2), 235–238.

Pevzner, P. A., Tang, H., & Waterman, M. S. (2001). An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the United States of America*, *98*(17), 9748–9753.

Phillippy, A. M. (2017). New advances in sequence assembly. *Genome Research*, *27*(5), xi–xiii.

Pop, M. (2009). Genome assembly reborn: recent computational challenges. *Briefings in Bioinformatics*, *10*(4), 354–366.

Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M., & Nyrén, P. (1996). Real-time DNA sequencing using detection of pyrophosphate release. *Analytical Biochemistry*, *242*(1), 84–89.

Ronaghi, M., Uhlén, M., & Nyrén, P. (1998). A sequencing method based on real-time pyrophosphate. *Science*, *281*(5375), 363, 365.

Rothberg, J. M., Hinz, W., Rearick, T. M., Schultz, J., Mileski, W., Davey, M., … Bustillo, J. (2011). An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, *475*(7356), 348–352.

Salas, M., Blanco, L., Lázaro, J. M., & de Vega, M. (2008). The bacteriophage ϕ29 DNA polymerase. *IUBMB Life*, *60*(1), 82–85.

Sanger, F., Nicklen, S., & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, *74*(12), 5463–5467.

Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., … Church, G. M. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, *309*(5741), 1728–1732.

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J. M., & Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. *Genome Research*, *19*(6), 1117–1123.

Sohn, J., & Nam, J.-W. (2018). The present and future of de novo whole-genome assembly. *Briefings in Bioinformatics*, *19*(1), 23–40.

Srinivasan, S., & Batra, J. (2014). Four generations of sequencing - is it ready for the clinic yet? *Journal of Next Generation Sequencing & Applications*, *1*, 107.

The 1000 Genomes Project Consortium. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., … Zhu, X. (2001). The sequence of the human genome. *Science*, *291*(5507), 1304–1351.

Weisenfeld, N. I., Kumar, V., Shah, P., Church, D. M., & Jaffe, D. B. (2017). Direct determination of diploid genome sequences. *Genome Research*, *27*(5), 757–767.

Wong, P. B., Wiley, E. O., Johnson, W. E., Ryder, O. A., O'Brien, S. J., Haussler, D., … G10KCOS. (2012). Tissue sampling methods and standards for vertebrate genomics. *GigaScience*, *1*, 8.

Yandell, M., & Ence, D. (2012). A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics*, *13*(5), 329–342.

Zerbino, D. R., & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research*, *18*(5), 821–829.

Zerbino, D. R., McEwen, G. K., Margulies, E. H., & Birney, E. (2009). Pebble and rock band: heuristic resolution of repeats and scaffolding in the velvet short-read de novo assembler. *PloS One*, *4*(12), e8407.

Zimin, A. V, Marçais, G., Puiu, D., Roberts, M., Salzberg, S. L., & Yorke, J. A. (2013). The MaSuRCA genome assembler. *Bioinformatics (Oxford, England)*, *29*(21), 2669–2677.

Zimin, A. V, Puiu, D., Luo, M.-C., Zhu, T., Koren, S., Marçais, G., … Salzberg, S. L. (2017). Hybrid assembly of the large and highly repetitive genome of Aegilops tauschii, a progenitor of bread wheat, with the MaSuRCA mega-reads algorithm. *Genome Research*, *27*(5), 787–792.

7. Online zdroje

1. Illumina, Inc. (2015). HiSeq 3000/HiSeq 4000 sequencing systems [brožura]. [cit. 26. 4. 2018]. Dostupné z: www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/hiseq-3000-4000-specification-sheet-770-2014-057.pdf.

2. Illumina, Inc. (2016). MiSeq system [brožura]. [cit. 26. 4. 2018]. Dostupné z: www.illumina.com/content/dam/illumina-marketing/documents/products/datasheets/datasheet_miseq.pdf.

3. Illumina, Inc. (2018). Sequence platform comparison tool [online]. [cit. 8. 4. 2018]. Dostupné z: www.illumina.com/systems/sequencing-platforms/comparison-tool.html.

4. Oxford Nanopore Technologies. (2017). Nanopore sequencing device comparison [online]. [cit. 27. 4. 2018]. Dostupné z: www.nanoporetech.com/products/comparison.

5. Pacific Biosciences of California, Inc. (2018). SMRT sequencing: read lengths [online]. [cit. 27. 4. 2018]. Dostupné z: www.pacb.com/smrt-science/smrt-sequencing/read-lengths.

6. Thermo Fischer Scientific Inc. (2018). Introducing the Ion GeneStudio S5 series for next-generation sequencing [brožura]. [cit. 26. 4. 2018]. Dostupné z: www.thermofisher.com/content/dam/LifeTech/Documents/PDFs/IPAC/PG1720-PJT2769-COL05762-Ion-GeneStudio-S5-Series-Flyer.pdf.

7. Wetterstrand, K. A. (2018). DNA sequencing costs: data from the NHGRI genome sequencing program (GSP) [online]. [cit. 28. 4. 2018]. Dostupné z: www.genome.gov/sequencingcostsdata.