OBSAH

Poděkování
Na tomto místě bych ráda poděkovala všem, kteří se jakkoli podíleli na realizaci tohoto výukového materiálu. Zvláště bych pak chtěla poděkovat Marianu Novotnému a Michalu Kolářovi za veškerou odbornou kritiku a cenné rady.

1 ÚVOD DO (GENOMICKÉ) BIOINFORMATIKY

Bioinformatika je relativně nový, interdisciplinární vědní obor zabývající se analýzou, interpretací a využitím velkých biologických dat za použití výpočetních technologií a metod. Taktéž se věnuje vývoji nových algoritmů, programů, analytických modelů a databází, ukládání, transformování a přenosu velkých objemů dat. Tento vědní obor zahrnuje studium genomů, signálních a metabolických drah, 3D dat, proteinových domén, i farmakologických složek. Kombinuje znalosti z molekulární biologie a výpočetní informatiky, ale také z matematiky, statistiky, biochemie, chemie a vytěžování dat.

Kromě klasické bioinformatiky, která spíše využívá výpočetní technologie k získání informací z biologických dat, případně vytváří nové algoritmy na zpracování biologických dat, se můžeme setkat s computational biology, která využívá biologická data na vytváření algoritmů a prediktivních modelů. V computational biology se také častěji využívá machine learning, neurální sítě a umělá inteligence. Typickým využitím computational biology je návrh nových léčiv nebo výběr vhodného kmene chřipky na očkování na nadcházející období.

Data science jako obor, který se věnuje zpracování a získání informací ze strukturovaných i nestrukturovaných dat pomocí algoritmů a výpočetních procesů, se pak dá považovat za nadmnožinu bioinformatiky.

1.1 Základní znalosti a předpoklady

Vzhledem k šíři a problematice oboru je kromě pouhého použití již vytvořených softwarových nástrojů i poměrně důležité chápat podstatu využívaných algoritmů, podstatu biologického problému, kterým se zabývá výzkum i jeho technologické zpracování. Z tohoto důvodu je dobré mít základní znalosti v následujících oblastech:

Biologie – závisí na problematice, kterou se výzkum hodlá zabývat. Obecně jsou požadovány znalosti molekulární biologie, genového inženýrství, genetiky, genomiky, evoluce a signálních drah.

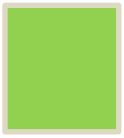
Technologie – vzhledem k původu dat, znalost převážně moderních metod sekvenování, v menší míře i čipových dat.

Bioinformatika – znalost stávajících bioinformatických nástrojů a chápání základních algoritmů (typicky např. lokální a globální zarovnání, skórovací matice), orientace v databázích, práce s různými datovými typy.

Statistika a programování – většinou je vyžadování znalost Pythonu, R, případně Matlabu a Perlu (ten už ale spíše výjimečně). Dále je nezbytná orientace v základní statistice, testování hypotéz a v korelačních analýzách. Rozhodně je dobré mít znalosti i databázových systémů v jazyku SQL nebo SPARQL.

Vytěžování znalostí z dat – hierarchické klastrování, tvorba stromů, orientovaných i neorientovaných grafů, znalosti efektivního vyhledávání v textech (základ pro hledání podobností a zarovnání).

Soft skills – schopnost komplexního myšlení, schopnost pracovat samostatně a učit se samostatně novým věcem, a hlavně znalost anglického jazyka – protože většina nástrojů, článků, příruček a návodů je v anglickém jazyce.



Tento seznam je samozřejmě spíše **ilustrační**, studium bioinformatiky je dlouhodobý proces, ve kterém se spousta poznatků a zkušeností získá praxí (ostatně tak jako ostatně ve všech oborech)

Jelikož drtivá většina bioinformatických nástrojů a softwarových řešení je implementovaná na linuxový systém, je nutná alespoň základní orientace v příkazové řádce.

1.2 Využití bioinformatiky

Bioinformatika má široké využití ve státní i soukromé sféře. Bioinformatici nejčastěji nachází uplatnění ve výzkumu, primárním i jako servisní skupina pro podporu ostatních výzkumů, dále např. ve farmaceutickém průmyslu při hledání nových farmakoaktivních látek, ve šlechtitelství, v potravinářském průmyslu u kontroly falšování potravin, v biotechnologických firmách a ve zdravotnických zařízeních pro detekci potenciálně patogenních variant, personalizovanou medicínu a genové terapie.

Bioinformatika se věnuje všem typům biologických dat, mezi časté oblasti využití spadá např.:

Anotace genomu – lokalizace a charakterizace genů a kódujících oblastí, včetně určení jejich proteinového produktu a funkce, označení hranice intron-exon, určení regulačních oblastí a repetic.

Komparativní bioinformatika – porovnávání genomů (genů, sekvencí, regulačních oblastí) různých organismů s cílem vyhledat sekvence, které sdílejí společného předka, a získat informace o tom, jak jsou dané sekvence zakonzervovány a určit evoluční vztahy mezi nimi.

Analýza variant – často se využívá v medicíně pro odhalení sekvenčních variant (mutací: jednonukleotidových (SNP) i vícenukleotidových polymorfismů (MNP), delece, inzerce), které by mohly způsobovat případně ovlivňovat onemocnění pacienta.

(Diferenční) Analýza genové exprese – typicky zkoumání rozdílu exprese genů mezi zdravou a nemocnou tkání, před, během a po podání léčebné látky, dále charakterizace metabolických a signálních drah.

Predikce struktury proteinů – předpověď trojrozměrné struktury proteinů umožňuje zjistit funkci proteinu, ale ji i vhodnou proteinovou úpravou změnit, určit vazebná místa a molekuly a další proteiny, které se na dané místo váží. Důležitá metoda pro umožnění léčby geneticky podmíněných chorob pomocí proteinové terapie.

Výběr/návrh nových léčiv – prohledávání chemického prostoru a tvorba nových léčiv na základě znalosti biologického cíle a struktury vazebného místa s cílem zlevnit a zrychlit celý proces vývoje nových léčiv.

Bioinformatika obrazu (Bioimage informatics) –využití výpočetních technik k analýze bioobrazů získaných např. z fluorescenční mikroskopie nebo histologie.

Do dalších oblastí spadá analýza proteomických dat, vytváření ontologií, homologní modelování, vazebné interakce, predikce vztahu struktura-funkce a jiné.

1.3 Historický přehled

Pokud vezmeme v potaz základní definici bioinformatiky, tedy že se jedná o analýzu biologických systémů za využití výpočetní techniky, pak se bioinformatika překvapivě datuje už na začátek 60. let 20. století. V této době se myšlenka, že makromolekuly jsou nositelkami dědičné informace, stala středobodem koncepčního rámce molekulární biologie. Důležitou osobou toho období je Margaret Dayhoff, která napsala řadu programů ve FORTRANu na stanovení úplně proteinové sekvence z fragmentů. Tyto programy byly schopny stanovit správnou sekvenci malého proteinu (ribonukleázy) během pár minut.

Rozvoj moderní bioinformatiky ale nastal až později v 80. letech, kdy Marvin H. Carruthers a Leroy Hood vytvořili metodu pro automatizované sekvenování DNA, a vznikly první velké databáze DNA sekvencí jako GenBank, EMBL a DNA Database of Japan. Roku 1988 byla založena Human Genome organization (HUGO). O rok později byla publikována první genomová mapa (pro bakterii Haemophilus influenza) a o další rok později odstartovat projekt sekvenování lidského genomu (Human Genome Project).

Více konkrétních událostí předcházejících i provázejících vývoj interdisciplinárního oboru bioinformatika najdete v Příloze v rámci Časové osy milníků molekulární biologie a výpočetní technologie.

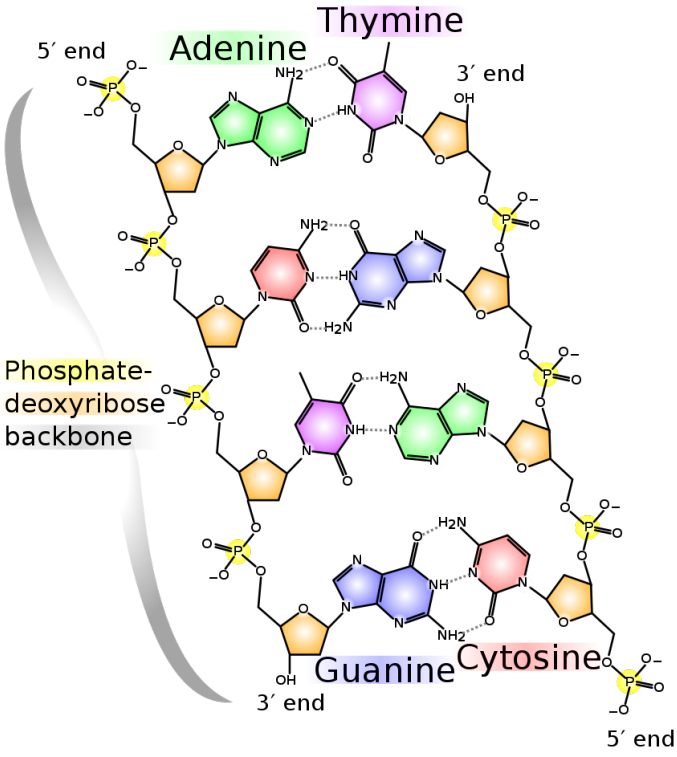
1.4 Zdroje

Fox, J. What is bioinformatics?, 2005. The Science Creative Quarterly. https://www.scq.ubc.ca/what-is-bioinformatics/ (accessed Dec 09, 2020).
Luscombe NM, Greenbaum D, Gerstein M. What is bioinformatics? A proposed definition and overview of the field. *Methods Inf Med*. 2001;40(4):346-358.
Tampi, S. M. Introduction to Bioinformatics, 2009. arxiv.org e-Print archive. https://arxiv.org/ftp/arxiv/papers/0911/0911.4230.pdf (accessed Dec 27, 2020).
Vincent AT, Charette SJ. Who qualifies to be a bioinformatician?. Front Genet. 2015;6:164. Published 2015 Apr 24. doi:10.3389/fgene.2015.00164
Welch L, Lewitter F, Schwartz R, et al. Bioinformatics curriculum guidelines: toward a definition of core competencies. PLoS Comput Biol. 2014;10(3):e1003496. Published 2014 Mar 6. doi:10.1371/journal.pcbi.1003496

2 BIOLOGICKÝ ZÁKLAD

2.1 DNA

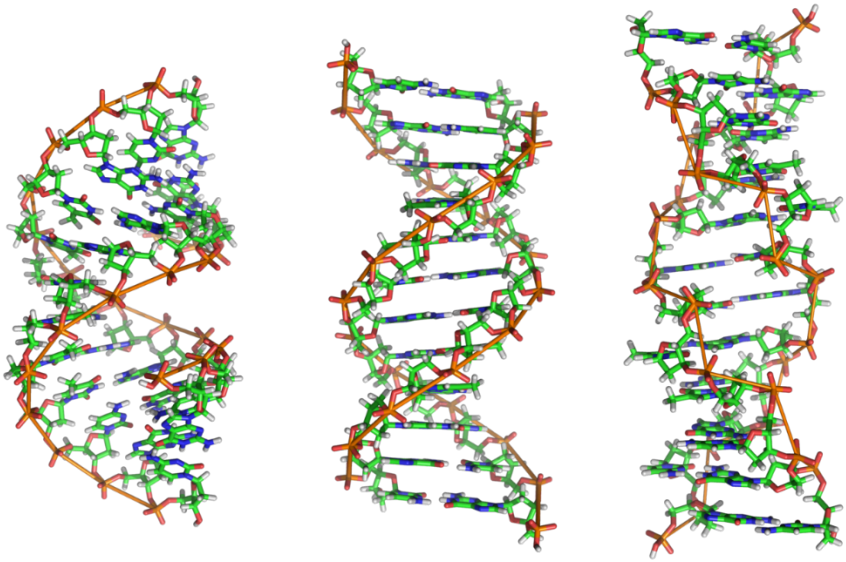
DNA – deoxyribonukleová kyselina, je polymerní biomakromolekula složená z nukleotidů, jejíž hlavní úlohou je uchování genetické informace. Každý nukleotid se skládá z deoxyribózy (sacharidová složka), z fosfátové skupiny a jedné z bazí (hlavní báze jsou 4: 2 puriny adenin A a guanin G, a 2 pyrimidiny cytosin C a thymin T, Obr. 1). Dále se mohou vzácně vyskytovat neklasické báze jako hypoxanthine a inosin a další). Sacharidový zbytek je vázán s fosfátovým zbytkem fosfodiesterovou vazbou, báze je spojena s cukerným zbytkem N-Glykosidovou vazbou.
Charakteristickou strukturou DNA je dvoušroubovice tvořena dvěma navzájem spletenými antiparalelními vlákny. Směr DNA se určuje podle pozice deoxyribózy, rozeznáváme směr 3'→5' a opačně směr 5'→3'. Standardně se báze zapisují ve směru 5'→3'. Nejčastěji se DNA vyskytuje v tzv. B-formě, dalšími formami jsou A-forma a Z-forma (Obr. 2, Tab. 1).



Obr. 1: Báze v DNA, Madprime, CC0, via Wikimedia Commons

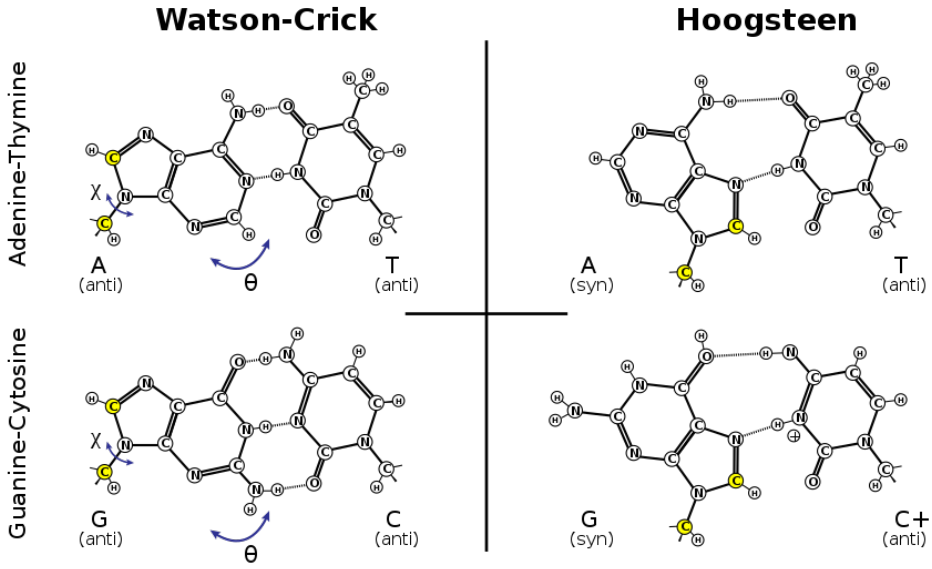
Tab. 1: Konformace DNA

Vlastnost	A-DNA	B-DNA	Z-DNA
Točivost šroubovice (chiralita)	pravotočivá	pravotočivá	levotočivá
Opakování	po každém páru	po každém páru	po každých dvou párech
Průměrný počet párů na jedno otočení šroubovice	11	10,5	12
Sklon páru k ose	20°	−1,2°	-9°
Průměr	23 Å (2,3 nm)	20 Å (2,0 nm)	18 Å (1,8 nm)
Konformace nukleosidu	anti	anti	C: anti, G: syn



Obr. 2: Konformace DNA, zleva A, B a Z DNA, Zephyris, CC BY-SA 3.0, via Wikimedia Commons

Mezi protilehlými bázemi obou vláken se vytvářejí vodíkové můstky (relativně silné nekovalentní interakce). Klasické párování (Watson-Crickovské párování bazí, Obr. 1) nastává mezi guaninem a cytosinem a mezi adeninem a thyminem, tedy dvojice mezi purinem a pyrimidinem. Neklasické párování známe dvojí: Hoogsteenovo párování (Obr. 3) a Wobble párování bazí. Hoogsteenovo párování bazí je stejně jako u kanonického párování mezi adeninem and thyminem, a guaninem a cytosinem, ale v jiné orientaci. Wobble párování je veškeré ostatní párování, které není Watson-Crickovské, hlavními příklady jsou guanine-uracil, hypoxanthine-uracil, hypoxanthine-adenine a hypoxanthine-cytosine páry. Neklasické párování se více než u DNA vyskytuje (i ve funkční podobě) u RNA.

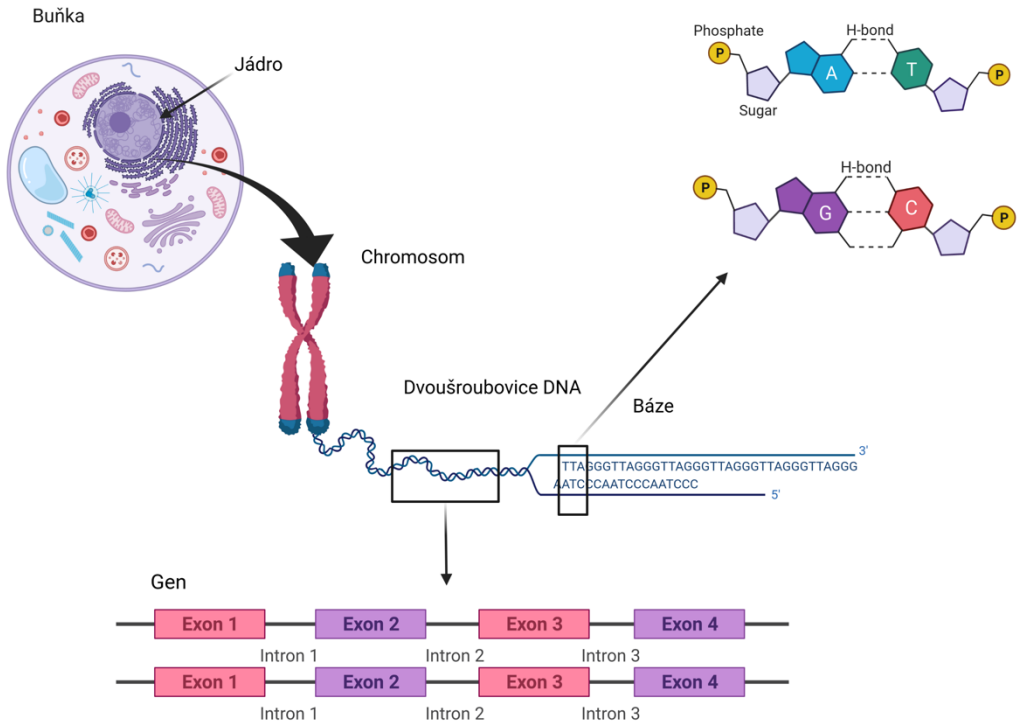


Obr. 3: Ukázka Watson-Crics a Hoogsteen párování, Ian Furst, CC BY-SA 4.0, via Wikimedia Commons

Typy DNA

DNA je v buňce převážně využívána k uložení genetické informace – genu, ve formě jednoho či několika replikonů (chromosomy, plasmidy, Obr. 4).

Intron (eukaryota): úseky genu u eukaryot o kterých se předpokládá, že jsou v zásadě nekódující. V posledních letech se však ukazuje, že i intronová část DNA je velice důležitá.
Exon: úsek genu, který kóduje určitou část funkční RNA. U eukaryot jsou exony přerušovány introny
Enhancer: oblast na DNA u eukaryot, na kterou se váží regulační proteiny a transkripční faktory. Enhancer se může vyskytovat před genem, za genem nebo dokonce uvnitř genu (v intronové oblasti), vzdálen může být desítky tisíc bazí.
Promotor: oblast na DNA, obvykle na začátku genu, na kterou se váže RNA polymeráza nebo transkripční faktor. Navázáním specifického proteinu na promotor obvykle dochází k zahájení transkripce. U prokaryot je na promotor navázána přímo RNA polymeráza, u eukaryot transkripční faktory (TATA-vazebný protein). Obecně důležitou součástí promotoru jsou specifické sekvence nukleotidů – například TATA box (u 40 % genů), CAAT box (u 64 %). Pro jeden gen může existovat více promotorů.
Operon: funkční jednotka DNA, která sdružuje geny pod kontrolou jednoho promotoru.
ORF (open reading frame): otevřené čtecí místo, sekvence, která může být translatovaná, začíná iniciačním kodonem (AUG, u prokaryot možno i GUG) a končí stop kodonem (UAA, UGA nebo UAG)



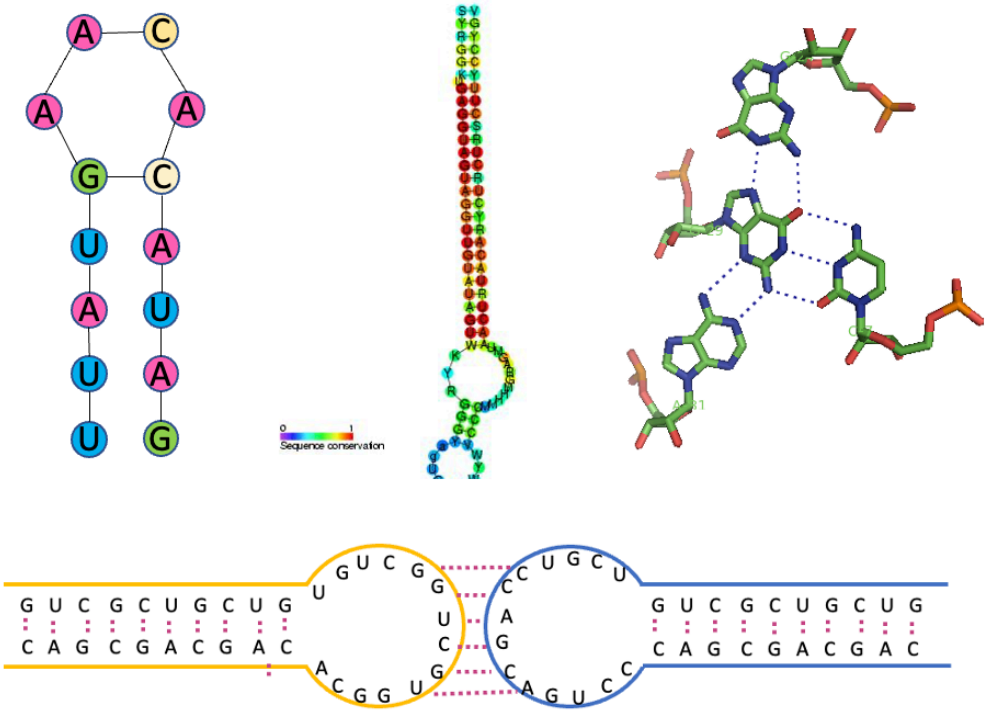
Obr. 4: Pohled na genetickou informaci a její složení. Pfeiferová L., Created with BioRender.com

2.2 RNA

RNA – ribonukleová kyselina, je polymerní biomolekula tvořena nukleotidy podobně jako DNA. Nukleotidy u RNA se skládají z ribózy (cukerná složka, proti deoxyribóze v DNA má na 2' pozici jednu hydroxylovou skupinu), jednoho zbytku kyseliny fosforečné a jedné báze – adeninu, guaninu, cytosinu nebo uracilu (na rozdíl od thyminu u DNA). Zatímco DNA tvoří většinou dvoušroubovici, RNA je flexibilnější a vytváří množství sekundárních a terciárních struktur. Na druhou stranu je RNA v sekundárních strukturách proti DNA méně stabilní. V organismu zastává RNA řadu funkcí, hlavní funkcí je zajištění překladu genetické informace a regulace genové exprese. Dalšími funkcemi jsou například katalytické funkce, řízení modifikace RNA, vazba malých molekul, tvorba jaderných tělísek a kontrola splicingu.

Struktury RNA

Dvoušroubovice: pouze tzv. A-formy (ve srovnání s B-formou širší, má více plochý tvar, hlubší velký žlábek a méně hluboký malý žlábek)
Vlásenka (Obr. 5): vyboulená smyčka vznikající ve dvoušroubovici při lokálním nedokonalém párování bází
Pseudouzel: vytvářející několik interakcí v rámci jedné molekuly
Kissing loops (líbající se smyčky, Obr. 5): interakce vznikající mezi různými molekulami
Trojšroubovice: struktura ve které se do klasické dvoušroubovice připojuje další vlákno, nejčastěji do malého žlábku
G-kvadruplex (Obr. 5): struktura, která se může vyskytnou i v DNA (v promotorových oblastech nebo u telomer), ale mnohem častěji se vyskytuje v RNA. Jedná se o rovinný systém čtyř guaninových bazí, které jsou vzájemně pootočených o 90° a pospojovány vodíkovými můstky. Pro vznik G-kvadruplexů je nutný výskyt Hoogsteenovského párování bazí



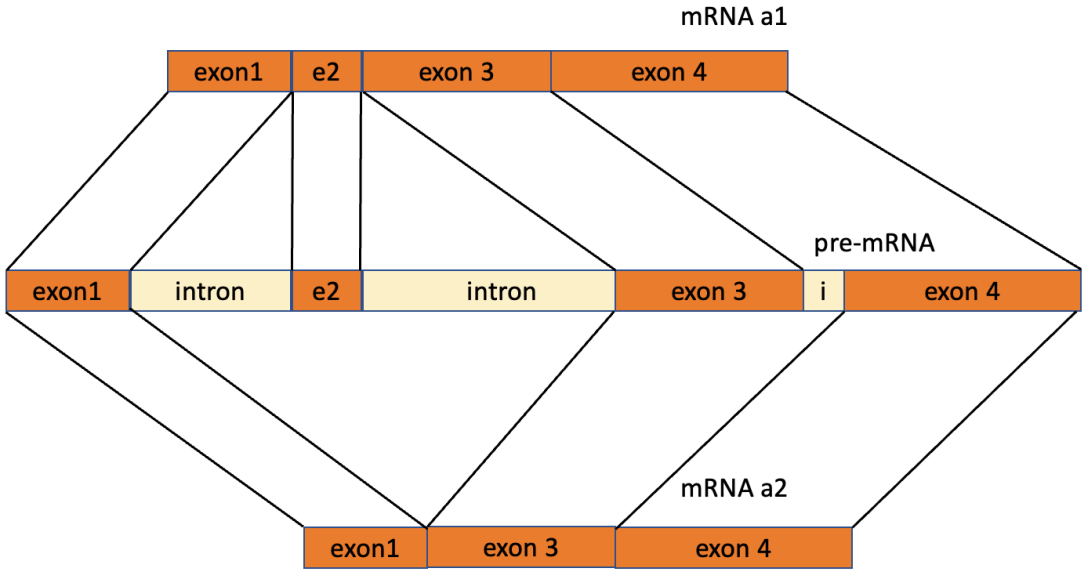
Obr. 5: Vlevo ukázka vlásenky, Pfeiferová L, uprostřed hairpin prekursor pro miR-let-7a, This image is taken from the Rfam database, Public domain, via Wikimedia Commons, vpravo G-kvadruplex, B Benham-Pyle, Public domain, via Wikimedia Commons, dole Kissing loops, Pfeiferová L

Typy RNA

mRNA (messenger RNA, „poslíčková“ RNA)**:** vzniká v buněčném jádře při transkripci. Jedno z vláken DNA je použito jako templát pro syntézu vlákna RNA. Nukleotidy v mRNA jsou komplementární vůči originální DNA.
UTR: nepřekládaná oblast, na každé straně kódující sekvence na řetězci mRNA (na 3' konci mRNA je 3' UTR oblast, na 5' konci mRNA je 5' UTR oblast), může mít řadu regulačních funkcí, například určovat, kdy bude docházet k translaci, případně určovat stabilitu celé mRNA
tRNA (transferová RNA**):** účastní se procesu translace, u kterého dochází k překladu sekvence nukleotidů v mRNA do sekvence aminokyselin v proteinech
rRNA (ribosomální RNA): spolu se specifickými proteiny se podílí na tvorbě ribosomu. rRNA tvoří tzv. ribozym – katalyticky aktivní molekulu RNA s enzymovou aktivitou. U prokaryotických organismů existují 3 různě velké rRNA (23S, 16S, 5S), u eukaryot i 4 druhy (28S, 18S, 5,8S a 5S). Je převážně jednovláknová, ale určité části ribosomu mají strukturu dvojité šroubovice
ncRNA (non-coding RNA, nekódující RNA): (**ncRNA**): obecně jsou molekuly RNA, které nejsou dále překládány na proteiny. Patří k nim piRNA, miRNA, long ncRNA a další. Předpokládá se, že nekódující RNA ovlivňují rozvoj a průběh některých onemocnění, včetně rakoviny, aneurysmat a Alzheimerovy choroby
miRNA: malé nekódující molekuly RNA (22 párů bazí), které mají vliv např. na komunikaci mezi nádorem a nádorovým mikroprostředím
piRNA (Piwi-interacting RNA): piRNA tvoří komplexy RNA-protein prostřednictvím interakcí s proteiny Argonaute podrodiny piwi. Tyto komplexy piRNA se většinou podílejí na epigenetickém a posttranskripčním umlčování transponovatelných prvků a dalších falešných nebo opakovaných transkriptů, ale mohou se také podílet na regulaci dalších genetických prvků v buňkách zárodečné linie
snoRNA: jsou třídou malých molekul RNA, které primárně řídí chemické modifikace jiných RNA, zejména ribosomálních RNA, přenosových RNA a malých jaderných RNA.
snRNA (snurps): malé nukleární ribonukleoproteiny, jsou komplexy RNA-protein, které se kombinují s nemodifikovanou pre-mRNA a různými jinými proteiny za vzniku spliceosomu, velkého molekulárního komplexu RNA-protein, na kterém dochází k sestřihu pre-mRNA
long ncRNA (Long non-coding RNAs, dlouhá nekódující RNA): transkripty s délkou přesahující 200 bazí, které nejsou dále překládány na proteiny

Modifikace

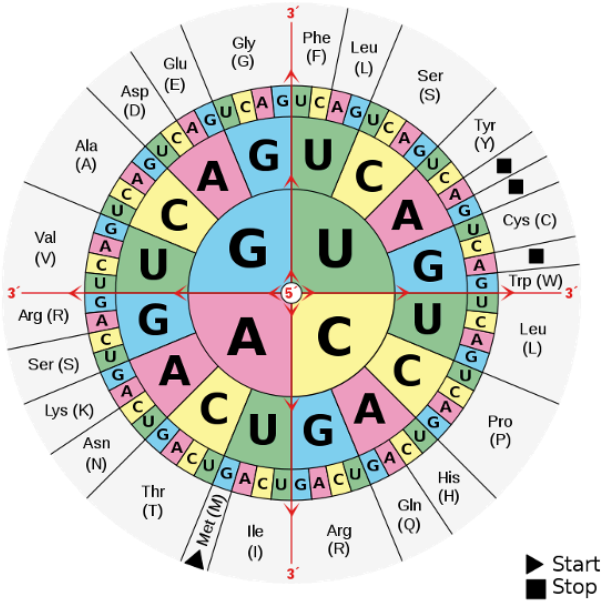
RNA splicing: jedna z posttranskripčních modifikací mRNA
Alternativní splicing: díky různým variantním sestřihům z jednoho genu vzniká více bílkovinných produktů (Obr. 6)
spliceosome: velká a složitá molekulární entita RNA, která se nachází primárně v jádru eukaryotických buněk. Spliceosom je sestaven z malých jaderných RNA (snRNA) a přibližně 80 proteinů. Spliceosom odstraňuje introny z přepsané pre-mRNA, což je typ primárního transkriptu.



Obr. 6: Alternativní splicing, Pfeiferová L

2.3 Proteiny

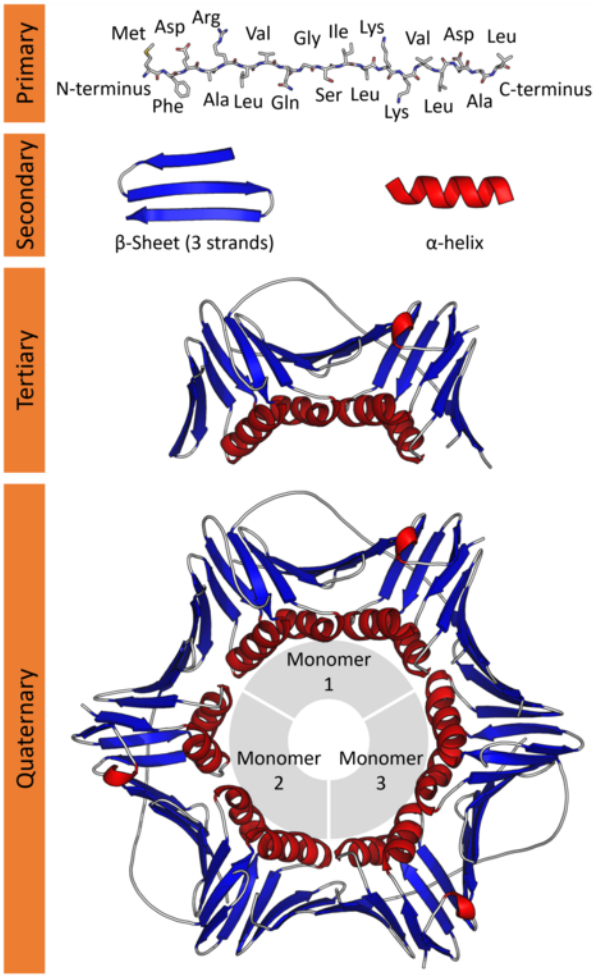
Proteiny jsou polymerní makromolekuly složené z řetězců aminokyselinových residuí. Proteiny zastávají v organismu řadu funkcí, poskytování struktury buňkách a organismu (kolagen, elastin, tubulin, kinesin), transportní (transportin, hemoglobin), pohybové (myosin, aktin), odpovědi na stimuly, katalytické a metabolické (enzymy, hormony, receptory), DNA replikace a obranné (imunoglobulin, fibrin, fibrinogen). Biosyntéza proteinů probíhá na ribozomech procesem translace, kdy jsou jednotlivé aminokyseliny vázaný do peptidových řetězců v pořadí, jaké odpovídá sledu kodónů (tripletů nukleotidů, Obr. 7) na mRNA.



Obr. 7: Nukleotidové triplety – kodony, a jejich proteinová translace, Mouagip, Public domain, via Wikimedia Commons

Struktury

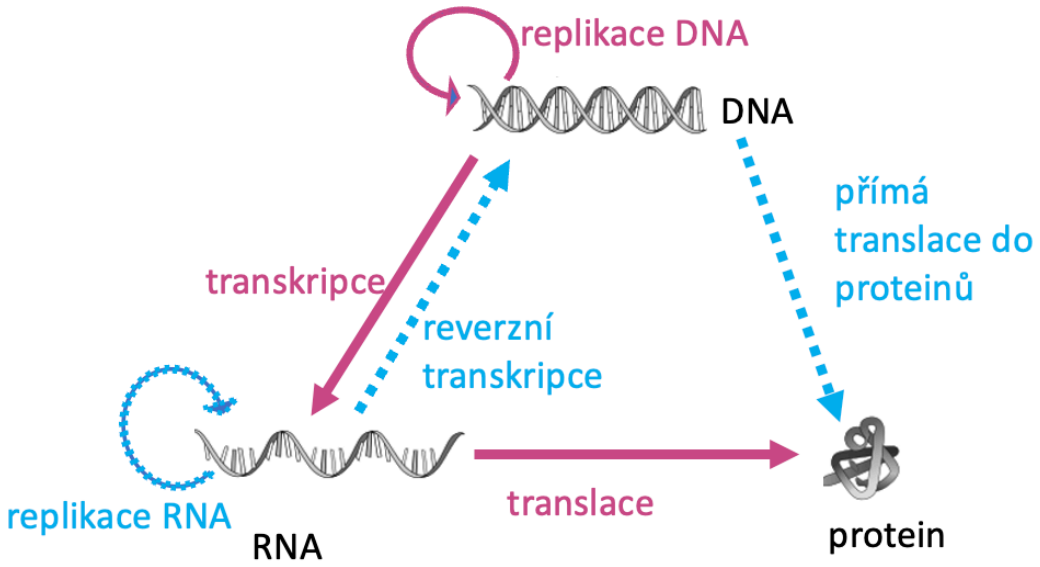
Primární: je daná pořadím aminokyselin v peptidovém řetězci, zapisuje se od N-konce k C-konci
Sekundární: geometrické uspořádání peptidového řetězce, jsou dva (tři) základní motivy, alfa šroubovice (alpha helix) a beta skládaný list (beta-sheet), dále je neuspořádaná struktura (random coil)
Terciální: trojrozměrné uspořádání peptidového řetězce, rozlišujeme globulární strukturu rozpustnou ve vodě a fibrilární, která je ve vodě nerozpustná. Terciální struktura bývá stabilizována kovalentními vazbami
Kvartérní: složitější komplexy proteinů, uspořádání podjednotek jednotlivých polypeptidových struktur, které tvoří jeden funkční protein
Kvintální: vlastnosti povrchu proteinu, determinové pouze v rámci buněčného kontextu (dochází k soustavným přechodných setkáním mezi makromolekulami)



Obr. 8: Členění struktur proteinů, Thomas Shafee, CC BY 4.0 via Wikimedia Commons

2.4 Centrální dogma molekulární biologie

Jako centrální dogma molekulární biologie označujeme cestu přenosu informace mezi biopolymery (Obr. 9). Cesta navržená Francisem Crickem spočívá v replikaci DNA, její transkripci do RNA a z ní pak translace do proteinů. Dlouhou dobu se uvažovaly pouze tyto tři cesty, nicméně byly objeveny tři další (vzácné): reverzní transkripce, kdy jsou retroviry schopny přepsat svou RNA do DNA a začlenit ji do hostitele (využití u reverzní PRC), další přepis RNA-RNA u některých RNA virů, a přepis DNA-protein, který byl pozorován za velice specifických podmínek a pouze v laboratorním in-vitro systému.



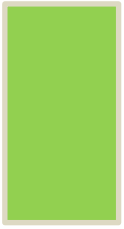
Obr. 9: Centrální dogma molekulární biologie, Pfeiferová L. Hlavní cesty vyznačené fialově, vedlejší tyrkysově

2.5 Zdroje

Klouda, P. *Základy biochemie*; PAVEL KLOUDA, 2013.
Kodíček, M. *Biochemické pojmy*, 2nd ed. [online]; Vydavatelství VŠCHT Praha, http://147.33.74.135/knihy/uid_es-002/index.html
Vodrážka, Z. *Biochemie*, 1st ed.; Academie: Praha, 2002.

3 PCR

Polymerázová řetězová reakce – PCR, je relativně snadná technika, která se používá k amplifikování malého fragmentu DNA, detekci určité sekvence DNA ve vzorku (pomocí komplementárních primerů) a kvantifikaci DNA/transkripce. Přibližný princip metody replikace nukleových kyselin byl popsán již roku 1971 Kjellem Kleppem a Gobindem Khoranou. Za opravdového objevitele metody DNA amplifikace je však považován až Kary Mullis, který v roce 1983 demonstroval její průběh. Metoda klonování fragmentů pomocí PCR byla poprvé publikována v 1985 v Science.

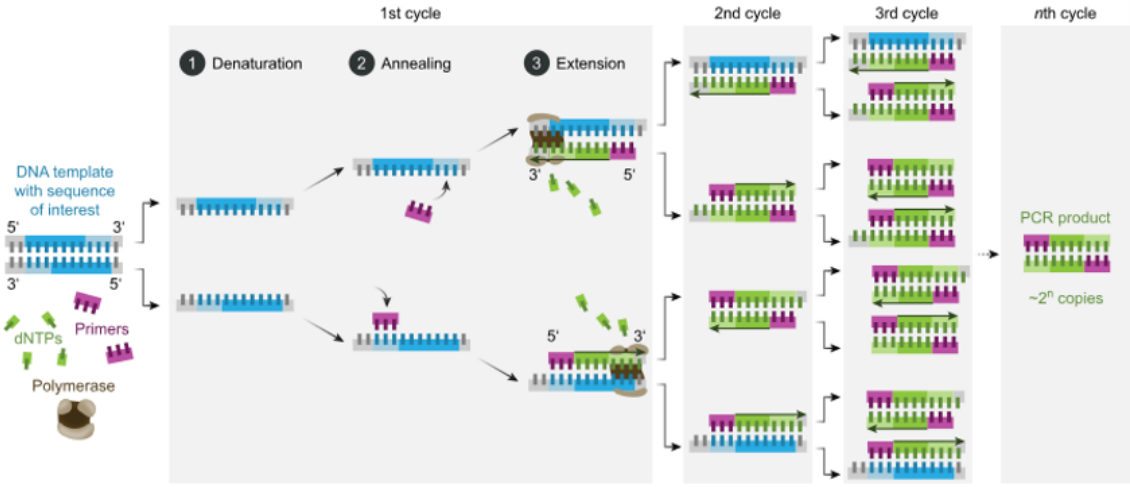


Kary Mullis spolu s Michaelem Smithem získali v roce 1993 Nobelovu cenu za chemii. Michael Smith za jeho zásadní přispění k vytvoření metody specificky řízené mutageneze založené na oligonukleotidech a jejím rozvoji pro studium proteinů a Kary Mullis za invence v metodě polymerázové řetězové reakce.

3.1 Postup

Klasická (**end-point**) PCR se skládá ze tří cyklicky se opakujících kroků; denaturace, nasednutí primerů (primer annealing) a vlastní syntézy DNA (Obr. 10). Pro dostatečné amplifikování fragmentu DNA obvykle stačí 30 cyklů. Teoretický výtěžek jedné molekuly DNA po odběhnutí 32 cyklů PCR je až 1 miliarda nově amplifikovaných molekul. Existuje více technik PCR (Tab. 2).

1. Denaturace – pro rozrušení vodíkových můstků dvoušroubovice DNA, a tím její denaturaci, probíhá rychlé zahřívání molekuly (20–30 sekund) na teplotu 94–98 °C. Díky tomuto kroku vzniká jednovláknová DNA, na kterou se v dalším kroku navazují krátké jednořetězcové oligonukleotidové molekuly se známou sekvencí – sekvenační primery.
2. Annealing – aby bylo umožněno navázání primerů na specifické místo na DNA, musí být teplota v tomto kroku snížena na 50-65 °C. Na vzniklé dvouvláknové úseky DNA-primer je navázána DNA polymeráza.
3. posledním krokem je dosyntetizování fragmentů DNA s navázanými primery ve směru od 5' konce ke 3' konci. Teplota použitá v této fázi se může lišit podle použité DNA polymerázy. Nejčastěji používaná Taq polymeráza má optimum aktivity 72 °C. Výsledkem tohoto kroku je nové vlákno komplementární k původnímu fragmentu DNA a vzniklá dvoušroubovice DNA.



Obr. 10: Postup klasické PCR, Enzoklop, CC BY-SA 4.0, via Wikimedia Commons

Do příchodu termostabilní DNA polymerázy z *Thermophilus aquaticus* (Taq) v roce 1986 se musela v každém cyklu přidávat vždy nová polymeráza, protože byla během denaturační fáze zničena. Díky této modifikaci se mohla PCR zautomatizovat, a tím se stala daleko dostupnější technologií.

Tab. 2: Příklady možností PCR

Polymerázová řetězová reakce	PCR
Reverzní transkripce s následnou polymerázovou řetězcovou reakcí	RT-PCR
Polymerázová řetězová reakce v reálném čase (Real Time)	qPCR
RT-PCR / qPCR kombinovaná technika	qRT-PCR

3.2 Problémy PCR reakcí

Známou chybou při PCR je také vznik tzv. chimerních sekvencí, kdy polymeráza přeskočí během elongační fáze z jedné templátové molekuly na jinou a výsledkem je sekvence ze dvou templátu rozdílného původu. Jako zdroj kontaminace DNA je nejčastěji uváděn přenos kontaminující DNA z dříve aplikovaných produktů PCR, vzájemná kontaminace zdrojových materiálů, kontaminace plasmidem z rekombinantního klonu který obsahuje cílovou sekvenci specifického produktu - nespecifická vazba primeru, vznik sekundárního amplifikačního produktu.

3.3 Otázky k tématu

1. Popište princip PCR reakce
2. Jaké jsou časté problémy v PCR reakcích?

3.4 Zdroje

PCR

Brown, T.; Brown, D.; Brown, T (Jnr).; Brown, A. ATDBio - Sequencing, forensic analysis and genetic analysis. https://atdbio.com/nucleic-acids-book/Sequencing-forensic-analysis-and-genetic-analysis (accessed Feb 14, 2022).
Polymerase Chain Reaction (PCR). Thermo Fisher Scientific. https://www.thermofisher.com/cz/en/home/life-science/pcr.html (accessed Dec 09, 2020).
Saiki RK, Scharf S, Faloona F, Mullis KB, Horn GT, Erlich HA, Arnheim N. Enzymatic amplification of beta-globin genomic sequences and restriction site analysis for diagnosis of sickle cell anemia. Science. 1985 Dec 20;230(4732):1350-4. doi: 10.1126/science.2999980.

4 SEKVENOVÁNÍ

Sekvenování je souhrnný termín pro biochemické metody, jimiž se zjišťuje pořadí nukleových bází (A, C, G, T) v sekvencích biomolekul, a tedy informace v sekvenci uložené. DNA a RNA molekuly (chemické entity) nesou biologickou informaci zakódovanou v pořadí nukleových bazí. Tato informace může kódovat protein (mRNA), gen a jeho regulační oblasti (genomová DNA) nebo například funkční RNA (rRNA). Bylo vyvinuto poměrně velké množství technik sloužících k sekvenování DNA, které se liší některými základními principy a dále především cenou a rychlostí. Sekvenování RNA většinou neprobíhá přímo, ale pomocí komplementární DNA k RNA, tzv. cDNA, kvůli větší chybovosti a menší stabilitě RNA Polymerázy. Nicméně metody 3. generace (dále v textu) již nejsou závislé na PCR, tudíž u nich je možné (a hojně využívané) sekvenování přímo RNA.

Tato část je zaměřena na DNA, cDNA, u metod 3. generace i RNA, ale ne na sekvenování proteinů. Pro náhled na transkriptom se používá RNA-Seq, ačkoli by jistě bylo lepší určovat přímo proteiny, nicméně sekvenování proteinů je technicky značně náročné, jejich sekvence se určuje Edmanovým odbouráváním nebo v současné době pomocí hmotnostní spektrometrie.

4.1 ÚVOD DO SEKVENOVÁNÍ

Většina dat pro bioinformatické zpracování pochází právě ze sekvenačních projektů. Sekvenování se dá chápat jako metoda určení pořadí nukleových bází v určitém úseku řetězce nukleové kyseliny. Pro účely detekce sekvenačních variant a pro určitou část anotace genomu se využívá sekvenování DNA, pro transkriptomiku a analýzu genové exprese se využívá RNA s mezikrokem transkripce na cDNA (komplementární DNA k RNA). RNA se přímo klasickými metodami ani metodami nové generace nesekvenuje kvůli nestabilitě RNA polymerázy, a kvůli její větší chybovosti (RNA polymeráza nemá proofreading aktivitu, tj. exonukleázovou aktivitu ve směru 3'→ 5' pro opravu RNA). Metody třetí generace se díky tomu, že nevyužívají PCR, mohou použít i na přímou sekvenaci RNA.

Podle způsobu sekvenování rozdělujeme sekvenační metody na **Klasické** (První generace), Nové generace **NGS** (Next generation sequencing, Massive parallel sequencing, masivní paralelní sekvenování, MPS) a na **Třetí generace** sekvenování (3^{rd} generation sequencing).

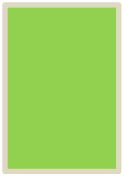
Do klasických sekvenačních metod řadíme **Maxam-Gilbertovu metodu** a **Sangerovu metodu**. Zatímco Maxam-Gilbertova využívala chemickou modifikaci DNA a štěpení na určitých místech a je v současné době překonaná, Sangerova metoda kopíruje přirozený proces syntézy DNA s terminací pomocí značených dideoxynukleotidů, a v lehce modifikované podobě se využívá dodnes, například ve zdravotnictví pro potvrzení patogenních mutací nalezených pomocí metod NGS.

NGS metody **(Illumina-Solexa, 454, Ion-Torrent**) oproti klasickým metodám umožňují rychlou a cenově příznivou produkci velkého množství osekvenovaných vzorků najednou (zisk milionů sekvencí, cena sekvenování za 1 b je až o dva řády nižší oproti Sangerovu s využitím kapilární elektroforézy). Pracují na principu paralelizace procesu sekvenování, kdy dochází k sekvenování tisíců až milionů sekvencí současně. Tímto však vzniká obrovská produkce výstupních dat, s následnou potřebou data utřídit, zanalyzovat a uschovat.

Všechny NGS metody mají podobný technologický průběh. Prvním krokem je fragmentování templátové DNA na úseky dlouhé několika set bází. Konce fragmentů jsou enzymatickou reakcí zatupeny. Dále dochází k napojením krátkých oligonukleotidových sekvencí (adaptérů). Jednotlivé fragmenty jsou amplifikovány pomocí PCR reakce a paralelně sekvenovány se ziskem miliónů sekvencí najednou. Narozdíl od Sangerovy metody, při které se získávají dlouhé sekvence (> 500bp), se délka sekvencí získaných pomocí NGS pohybuje zhruba mezi 20 až 700 párů bazí. Sekvenační výtěžek jednoho běhu sekvenátoru je v řádu tisíce Gb.

Specifickou aplikací NGS metod je sekvenování nukleových kyselin z jedné buňky (**Single-cell** sequencing).

Metody třetí generace (**SMRT, NanoPore, PacBio**) nevyužívají PCR reakci, zaměřují se na sekvenování jedné molekuly. Navíc produkují mnohem delší sekvence (desetitisíce bazí) oproti NGS metodám (maximálně stovky). Porovnání jednotlivých metod je v Tab. 3 a 4.



Donedávna metody třetí generace vykazovaly větší míru chybovosti (single error rate i accuracy) oproti ostatním metodám, proto se stále pracuje převážně s metodami NGS, nicméně v současné době (rok 2020) stávající metody dohání, ne-li předhání.

Funkčně rozlišujeme celou řadu sekvenačních analýz, jako RNA-seq (transkriptom), ATAC-Seq (otevřenost chromatinu), ChIP-Seq (Protein-DNA interakce), Methyl-Seq (místa methylované DNA), Ribo-Seq (stanovení translace), Dup-Seq (metoda pro detekci velmi vzácných mutací, s minimalizováním sekvenačních chyb), Sono-Seq (přístupné chromatinové oblasti) a zhruba desítka dalších.

Dále je možné sekvenovat celý genom (whole genome sequencing, WGS), celý exom (whole exom sequencing, WES) a cíleně sekvenovat podle panelů (targeted sequencing, gene specific sequencing), kdy se sekvenují geny většinou podle funkce (např. *cancer panel*, zaměřené na geny způsobující či ovlivňující rakovinné bujení, nebo *cardio panel* pro onemocnění srdce).

Tab. 3: Srovnání vybraných metod a modelů (Illumina)

Platforma/model	Run time	Výtěžek	Délka readu	Počet sekvencí
MiSeq (Illumina)	4-55 h	0,5-15 Gb	2 × 300 bp	1-25 mil
Nextseq 550 (Illumina)	12–30 h	120 Gb	2 × 150 bp	400 mil
Nextseq 2000 (Illumina)	11-48 h	330 Gb	2 × 150 bp	1.1 miliard
Novaseq 6000 (Illumina)	13-44 h	80–6000 Gb	2 × 250 bp	1.6–40 miliard
Ion 550 Chip (Ion Torrent)	11.5 h	20–25 Gb	200 bp	100–130 mil
Ion 530 Chip (Ion Torrent)	21.5 h	6-8 Gb	400 bp	15–20 mil
Ion 510 Chip (Ion Torrent)	5	0.6-1 GB	400 bp	2–3 mil
MinION Mk1B (Oxford Nanopore, Illumina)	1 min - 72 h	10–50 Gb	>4 Mb	100 tisíc na barcodovaný vzorek
PromethION 48 (Oxford Nanopore, Illumina)	1 min - 72 h	100-300 Gb	>4 Mb	až k 25 mil
PacBio **SequelIIe Systém (PacBio)**	30 h	4 Gb/SMRT Cell	>10 kb	4 mil
Sequel Systém (PacBio)	20 h	4 Gb/SMRT Cell	>10 kb	0.5 mil
SeqStudio Genetic Analyzer (Applied Biosystem, Sangerova metoda)	30 min		350-800 bp	67000
Refreshed 3730 Series Genetic Analyzer (Applied Biosystem, Sangerova metoda)	20 min		400-900 bp	1.38-2.76 mil
5500 SOLiD	24 h pro 35 bp (1 lane), 7 dnů pro 75 bp x 35 bp nebo 60 bp x 60 bp (6 lanes)	90-140 Gb	75 bp (fragment), 75 bp x 35 bp (paired end), 60 bp x 60 bp (mate paired	Exomy do 18, malé RNA do 144

Tab. 4: Přesnost vybraných modelů; consensus accuracy označuje kombinovanou informaci z více readů v datasetu, tedy celkovou přesnost (eliminace náhodné chyby v jednotlivých readech) single errors znamenají chybu v jednotlivých readech. U metod třetí generace můžeme stále vidět vyšší single error rate, nicméně jejich celková přesnost již převyšuje minimálně Illumina instrumenty.

Platforma/model	Consensus accuracy %	Single error rate %
MiSeq (Illumina)	≥ 89.7	0.1
Nextseq 550 (Illumina)	≥ 80	0.34
Nextseq 2000 (Illumina)	≥ 75	0.16
Novaseq 6000 (Illumina)	≥ 80	0.1
Ion 550 Chip (Ion Torrent)	99.97	1.78
Ion 530 Chip (Ion Torrent)	99.97	1.78
Ion 510 Chip (Ion Torrent)	99.97	1.78
MinION Mk1B (Oxford Nanopore, Illumina)	> 97.5	13–15
PromethION 48 (Oxford Nanopore, Illumina)	> 97.5	13–15
PacBio **SequelIIe Systém (PacBio)**	> 99	13
Sequel Systém (PacBio)	> 99,999	13
SeqStudio Genetic Analyzer (Applied Biosystem, Sangerova metoda)	> 99	1
Refreshed 3730 Series Genetic Analyzer (Applied Biosystem, Sangerova metoda)	99	0.125
5500 SOLiD	99.99	0.01

4.1.1 Biologický materiál pro sekvenování

Způsoby a možnosti bioinformatické analýzy jsou silně ovlivněny experimentální částí, od výběru biologického materiálu, přes jeho odběr až po zvolenou metodu sekvenování.

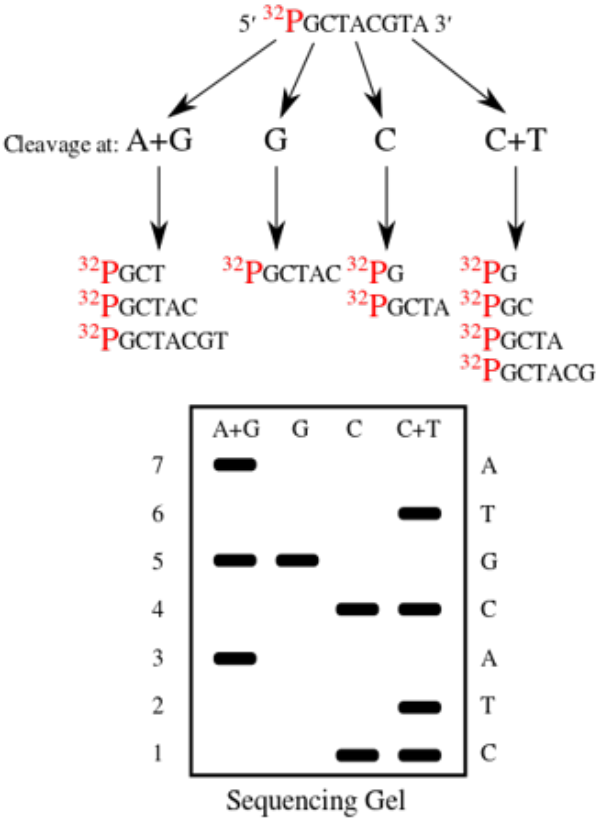
Sekvenovat se dá prakticky veškerý biologický materiál, od slin, přes tkáně, plnou krev, plazmu, sérum, leukocyty a kosti. V onkologii se v současné době využívá tzv. cfDNA (cirkulující volná DNA), což je typ extracelulární DNA vyskytující se v krvi (plazmě).

Obecně nejméně vhodným materiálem pro sekvenování bývá **plná krev.** Je to proto, že obsahuje směsici různých typů buněk, včetně mrtvých buněk a cfDNA. Při použití plné krve může u některých typů studií dojít ke zkreslení výsledků výzkumu.

4.2 KLASICKÉ METODY SEKVENOVÁNÍ

4.2.1 Maxam-Gilbert

Maxam-Gilbertova metoda (Obr. 11) využívá chemické štěpení sekvenované DNA ve specifických místech. Čtený DNA fragment je na 5' konci označen radioaktivní značkou. Pomocí chemické reakce je čtený úsek štěpen na jedné nebo dvou ze čtyř nukleotidových bází v každé ze čtyř reakcí (A+G, G, C, C+T). Fragmenty jsou odděleny podle délek gelovou elektroforézou, zobrazeny jsou pomocí autoradiografie.

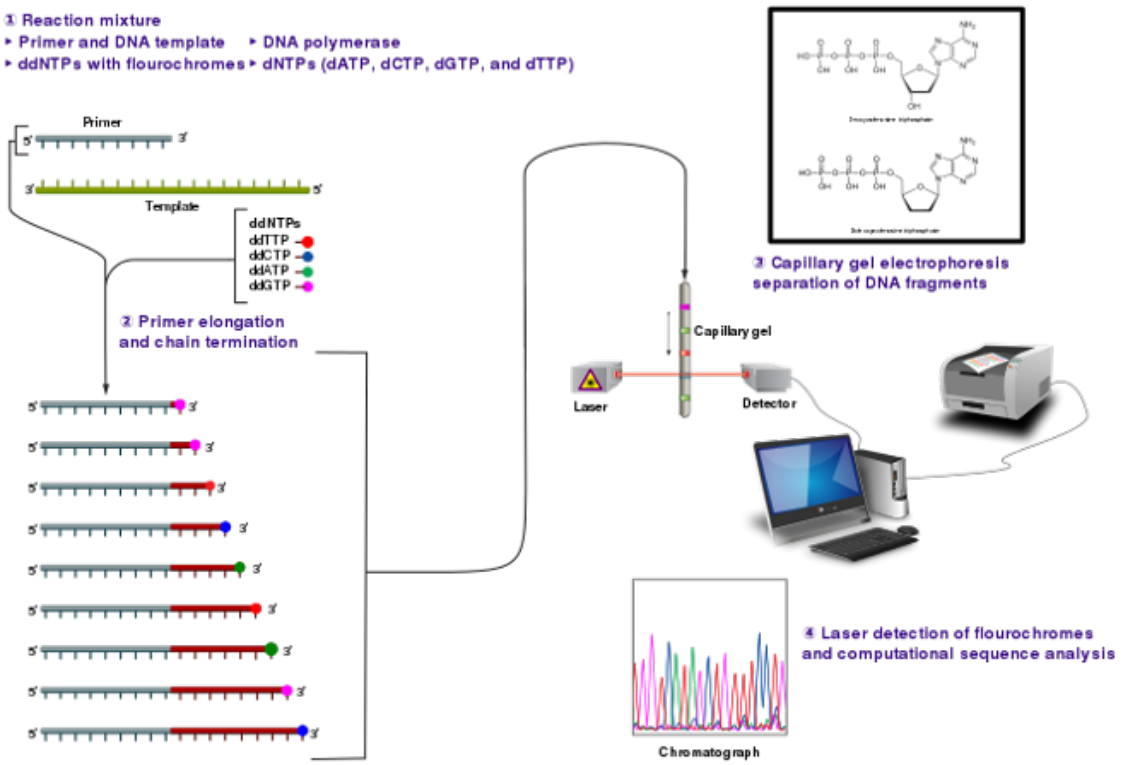


Obr. 11: Maxam-Gilbert, Shakiestone, CC BY-SA 3.0, via Wikimedia Commons

4.2.2 Sangerova metoda

Sangerova metoda (Obr. 12) kopíruje biologický proces replikace DNA. Původně reakce probíhala paralelně ve čtyřech zkumavkách. Vybraná sekvence se vloží do reakční směsi s radioaktivně značeným primerem, DNA polymerázou (Taq polymeráza, T7 polymeráza), s množstvím čtyř základních deoxyribonukleotidů (dATP, dGTP, dCTP, dTTP) a s jedním ze čtyř dideoxynukleotidů (ddATP, ddGTP, ddCTP, ddTTP) v každé reakci. Dideoxynukleotid je schopen se začlenit do replikující se DNA, ale následně zastaví elongaci řetězce, protože nemá OH skupinu, na níž by se připevnil další nukleotid. Všechny replikované sekvence v dané zkumavce jsou ukončeny vloženým dideoxynukleotidem. Výsledkem pak je směs různě dlouhých sekvencí DNA, které začínají radioaktivním primerem a končí daným dideoxynukleotidem. Sekvence jsou děleny pomocí gelové elektroforézy podle délek.

V současné době se místo radioaktivního značení využívá barvení dideoxynukleotidů fluorescenční látkou, reakce taktéž neběží pro jednu sekvenci ve čtyřech zkumavkách. Celý proces je zrychlený automatizací se separací pomocí kapilární elektroforézy (96 vzorků). Sangerova metoda se využívá např. na klinických pracovištích pro verifikaci nalezené mutace z NGS metod, nebo na detekci velkých delecí.



Obr. 12: Průběh Sangerovy metody, Estevezj, CC BY-SA 3.0, via Wikimedia Commons

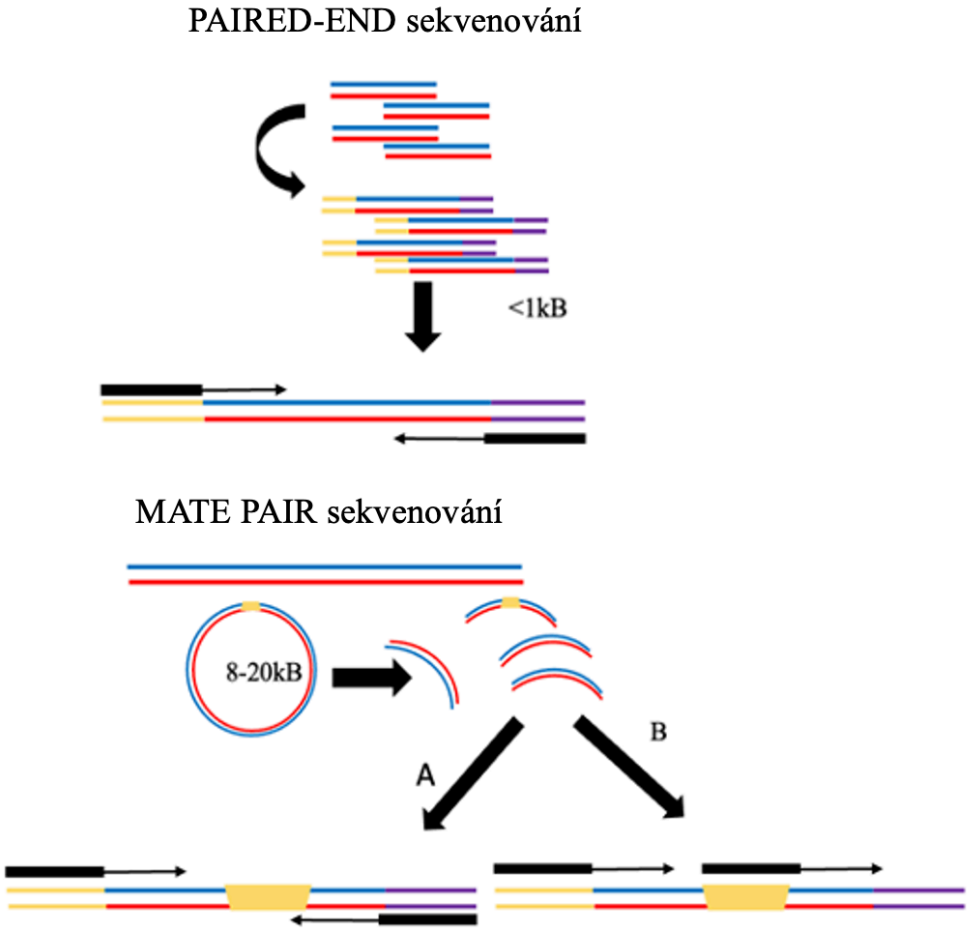
4.3 Sekvenování nové generace

Metody sekvenování nové generace využívají podobný mechanismus, skládající se ze tří hlavních kroků:

- Příprava knihovny – knihovny se připravují náhodnou fragmentací DNA, ať už fyzikální, chemickou, nebo biochemickou metodou, kterou následuje přopojení specifických sekvencí na konce fragmentů
- Amplifikace – knihovna je namnožena (amplifikována) pomocí klonálních amplifikačních metod a PCR, amplifikace probíhá v klastrech nebo na kuličkách
- Samotné sekvenování – DNA je sekvenována za použití různých protokolů, např. identifikace fluorochromových značek, využití biochemické reakce

Sekvenační knihovny je možné připravovat jako single-end nebo paired-end/mate pair čtení:
Single-end: při single-end čtení se fragment čte pouze od jednoho konce k druhému
Paired-end: fragment se čte z obou konců
Mate pair: je specifický typ paire-end sekvenování. Fragment molekuly je cirkularizován, fragmentován, a následná orientace čtení je závislá na použitém kitu pro přípravu. Na obrázku pod textem (Obr. 13) jsou páry připraveny pomocí různých sekvenačních kitů, přičemž se mění orientace čtení ve výsledných datech.

Hezké vysvětlení důvodu použití paire-end čtení poskytují následují videa od Roba Edwarse: https://www.youtube.com/watch?v=WTbnk91e2WU

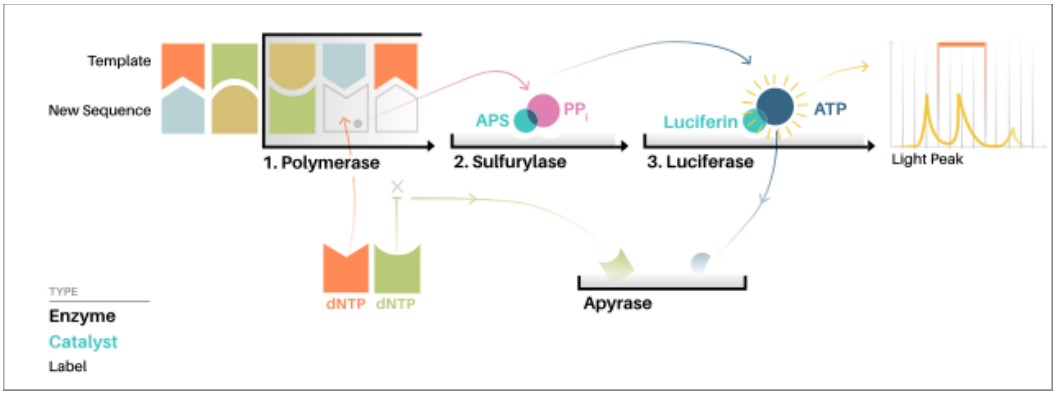


Obr. 13: Při paired-end sekvenování (vlevo) se sekvenují skutečné konce krátkých molekul DNA (méně než 1 kb), zatímco při sekvenování mate pair párů (vpravo) jsou konce dlouhých molekul spojeny, a připraveny podle speciálních sekvenačních knihovnách. Konce dlouhých molekul s vybranými velikostmi jsou spojeny s vnitřní adaptorovou sekvencí (tj. linker, žlutá) v cirkularizační reakci. Kruhová molekula je poté zpracována pomocí restrikčních enzymů nebo fragmentací. Na fragmenty obohacené o linker jsou připojeny adaptéry. Linker pak může být použit jako druhé primární místo pro další sekvenační reakci ve stejné orientaci (A) nebo může být sekvenování provedeno z druhého adaptéru (B), z reverzního vlákna. Pfeiferová L

4.3.1 Technologie 454

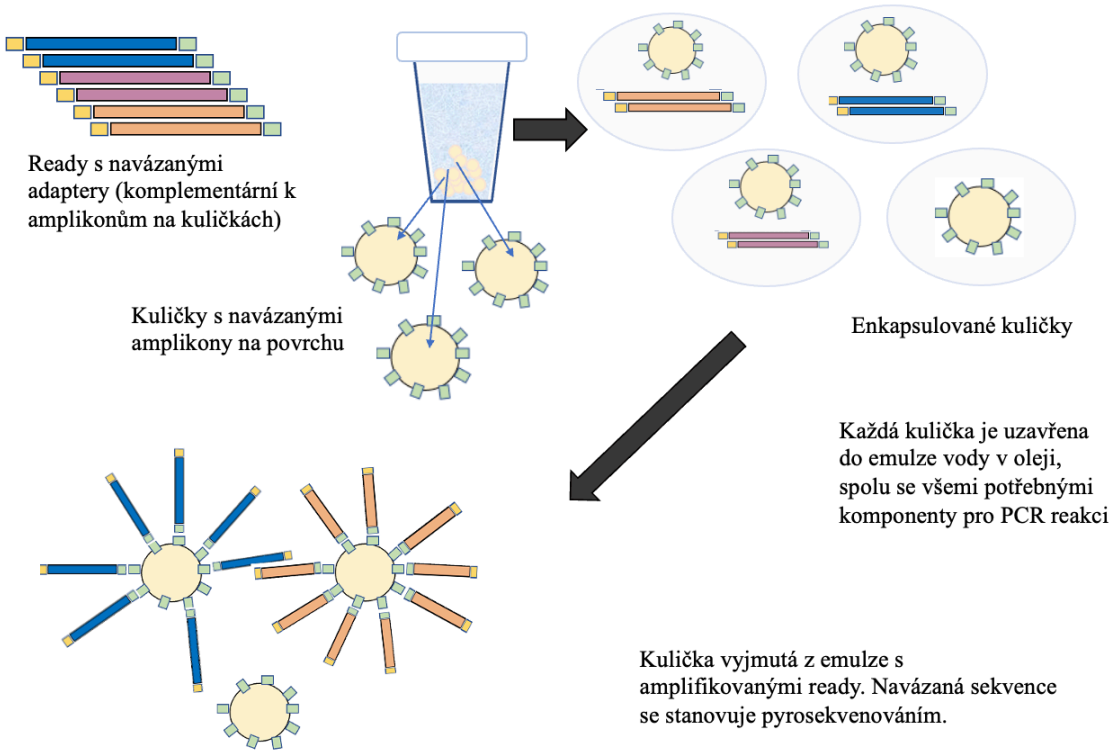
Technologie 454 (2005, Roche 2007) využívala metodu pyrosekvenování fragmentů DNA (série enzymatických reakcí, při kterých dochází k luminiscenci, Obr. 14) připravených pomocí emulzní PCR (Obr. 15). V současné době to je již zastaralá metoda, která se nevyužívá, ale ve své době znamenala významný pokrok v oblasti sekvenování.

Uvedení metody 454 umožnilo první sekvenaci lidského genomu.



Obr. 14: Pyrosekvenování, Technologie 454, This image has been created during "DensityDesign Integrated Course Final Synthesis Studio" at Polytechnic University of Milan, organized by DensityDesign Research Lab in 2015. Image is released under CC-BY-SA licence. Attribution goes to "Jacopo Pompilii, DensityDesign Research Lab"., CC BY-SA 4.0 via Wikimedia Commons

Vysvětlení metody pyrosekvenování v podobě videa lze nalézt na youtube stránce Quick Biochemistry Basics Pyrosequencing, https://www.youtube.com/watch?v=wY8to-_zAEo

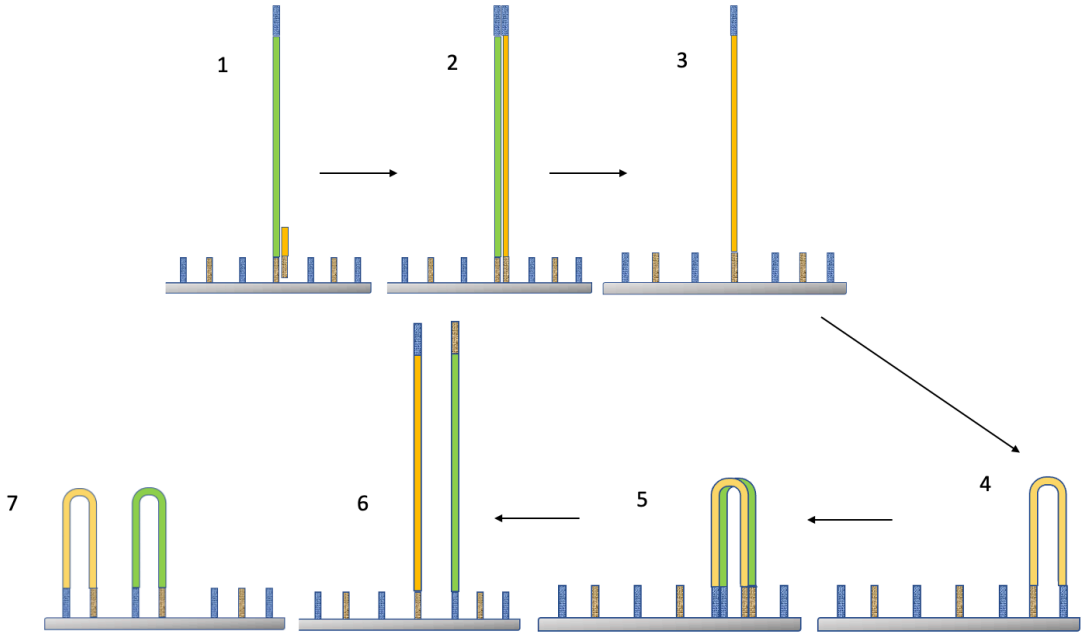


Obr. 15: Postup emulzní PCR technologie 454, Pfeiferová L

4.3.2 Solexa Illumina

Solexa (Solexa 2006, následně Illumina) je metoda založená na klastrové PCR, s použitím na single-end i pair-end knihovny. V prvním kroku dochází k vytvoření sekvenační knihovny náhodným fragmentováním DNA na úseky o velikosti zhruba 200 bp. Fragmentace probíhá většinou ultrazvukově/sonifikací nebo enzymaticky.
Na konce fragmentů jsou ligovány dva různé adaptéry (na jednom konci jeden adaptér, na druhém konci druhý adaptér). Fragmenty jsou následně denaturovány a nově vzniklé jednořetězcové molekuly jsou pomocí adaptérů komplementárně napojeny na oligonukleotidy, umístěné na povrchu reakční komůrky (**flow cell**).
Samotná klastrová (můstková) PCR (Obr. 16) probíhá následovně: na oligonukleotidech imobilizovaných na povrchu reakční komůrky dojde k extenzi oligonukleotidů podle templátové DNA. Původní templátová vlákna jsou odmyta pryč. V komůrce zůstává pouze nově syntetizované vlákno, navázané na povrchu komůrky. Při následném annealingu dochází k navázání volného konce imobilizované molekuly DNA k sousednímu komplementárnímu oligonukleotidu, tím vznikají jednořetězcové můstky. V dalším kole amplifikace, kdy je oligonukleotid prodloužen, se vytváří dvouřetězcový můstek. Dvouřetězcové molekuly jsou opětovně denaturovány na jednořetězcové, jejichž volné konce přisedají k volným oligonukleotidům. Celý proces se cyklicky opakuje. Dvouřetězcové mosty jsou nakonec denaturovány a reverzní vlákna odstraněna.

Sekvenování je založeno na syntéze komplementární DNA DNA-dependentní DNA polymerázou. V každém sekvenačním cyklu se nachází všechny čtyři fluorescenčně značené nukleotidy s terminační značkou (chemicky inaktivovaná OH skupina na 3' konci). Terminační značka znemožňuje navázání více nukleotidů. Po začlenění nukleotidu je na detektoru zachycen signál příslušného fluorochromu, každý nukleotid má svoji značku. Následně dochází k odstranění terminační značky a fluorochromu, a k navázání dalšího nukleotidu. Velkou výhodou je, že v jednom běhu může být analyzováno více knihoven.



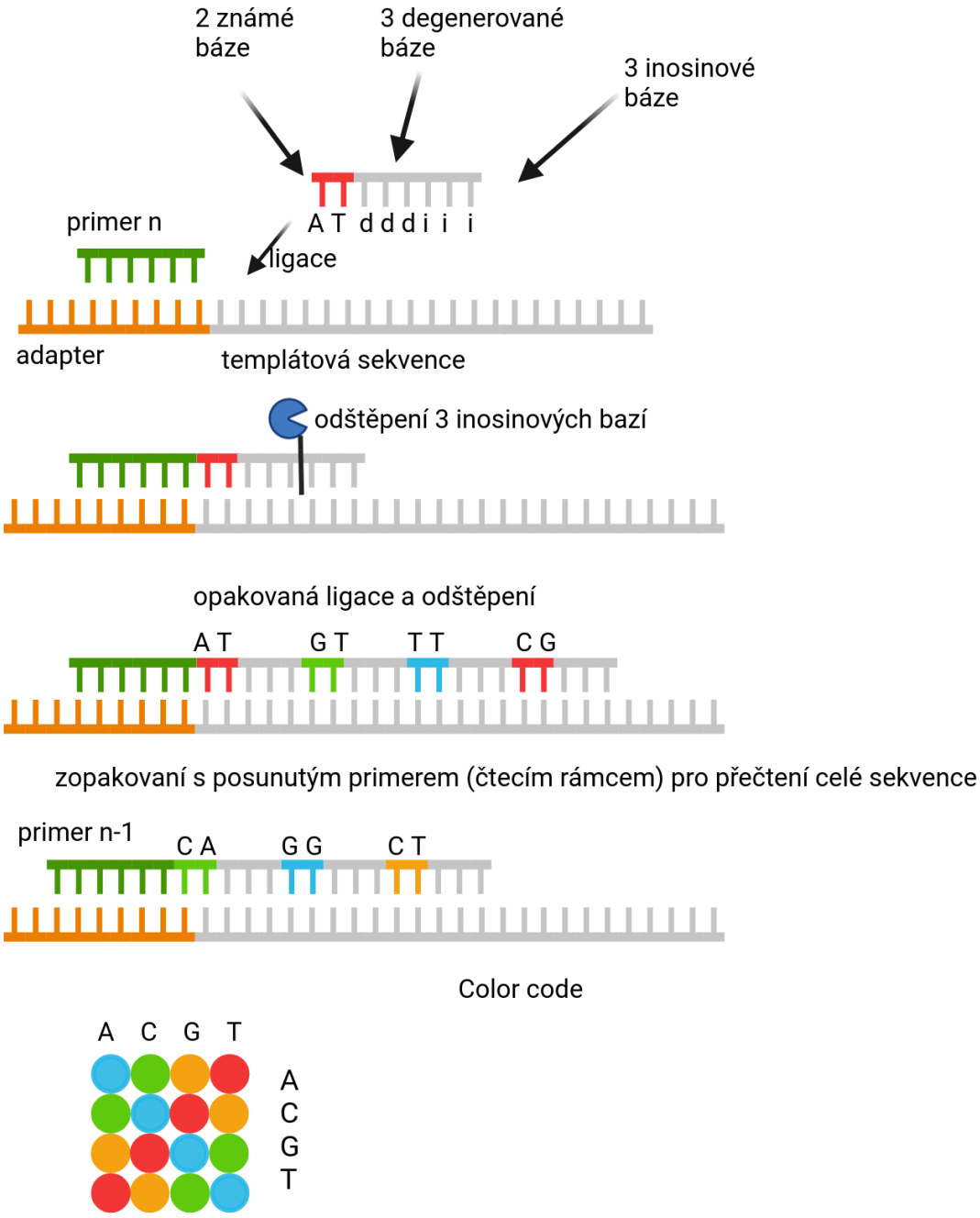
Obr. 16: Můstková PCR, Solexa (Illumina). 1. přisednutí fragmentu ke komplementárnímu oligonukleotidu umístěném na povrchu flow celly. 2. Extenze přisedlého fragmentu. 3. Denaturace vzniklého řetězce na jednořetězcové vlákno, odmytí původního řetězce. 4. Navázání volného konce k sousednímu komplementárnímu oligonukleotidu na povrchu flow celly – vytvoření jednořetězcového můstku. 5. Extenze oligonukleotidu a vytvoření dvouřetězcového můstku. 6. Denaturace na jednořetězcové molekuly. 7. Celý proces se opakuje, od kroku 4 (volné konce řetězců se opět váží na sousední komplementární oligonukleotidy)

Společnost Illumina má vlastní youtube channel, ve kterém lze kupříkladu najít video ukazující můstkovou PCR, https://www.youtube.com/watch?v=fCd6B5HRaZ8&t=112s

4.3.3 SOLiD

Na rozdíl od ostatních metod, SOLiD (Sequencing by Oligo Ligation and Detection, Applied Biosystems 2006, následně Life Technologies, nyní ThermoFisher) při sekvenování využívá DNA ligázu místo DNA polymerázy. Použití je možné na single-end i pair-end knihovny, přičemž pair-endové knihovny jsou tvořené z jednotlivých fragmentů. Metoda SOLiD (Obr. 17) je výrazně přesnější než zbylé NGS metody a je použitelná i na repetitivní oblasti. Na druhou stranu vytváří mnohem kratší sekvence, a je naprosto nevhodná pro sekvenaci palindromních sekvencí. Podobně jako u metody 454, se využívá emulzní PCR na kuličce. Templátová DNA je rozštěpena na velmi krátké úseky o délce několika desítek bazí. Na konce se naváží dva různé adaptéry (na 5' konec jeden adaptér, 3' konec druhý adaptér). V mikroreaktoru se fragmenty naváží na kuličku a dochází k emulzní PCR na povrchu kuličky. Po PCR dojde k denaturaci templátového vlákna (fragmentu) a k modifikaci 3' konce. Tyto kuličky jsou vloženy na skleněnou podložku, která může být členěna do jedné, čtyřech nebo osmi sekcí.

K sekvenování se využívá vazby krátkých fluorescenčně značených sond na templátové vlákno pomocí ligázy. Primer je hybridizován na adaptér, na který se následně naváže jedna ze šestnácti fluorescenčně značených sond. První dvě báze sondy jsou známé, další tři báze jsou degenerované a poslední tři báze jsou inosinové. Pro hybridizaci sondy musí být první dvě báze komplementární ke stanovované sekvenci. Po navázání je fluorescenční značka se třemi posledními bázemi odštěpena, je změřena fluorescence a vyhodnocena pomocí tabulky barev. Tímto způsobem dochází k navázání dalších sond. Aby byla sekvence přečtena celá, je nutné celý postup několikrát opakovat s posunutím čtecího rámce.

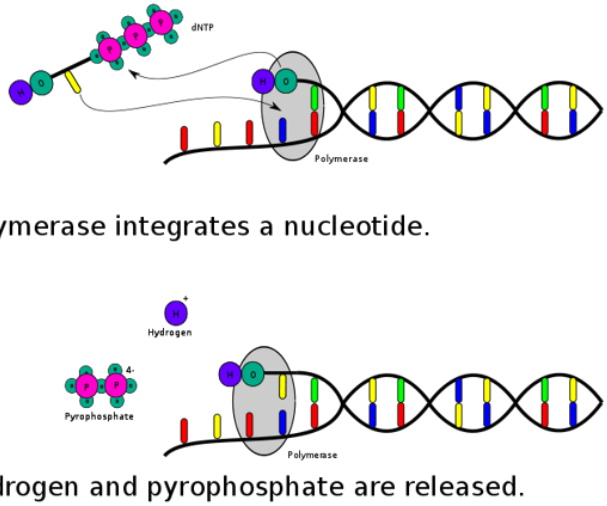


Obr. 17: Barevné schéma a princip sekvenování metody SOLiD, Pfeiferová L

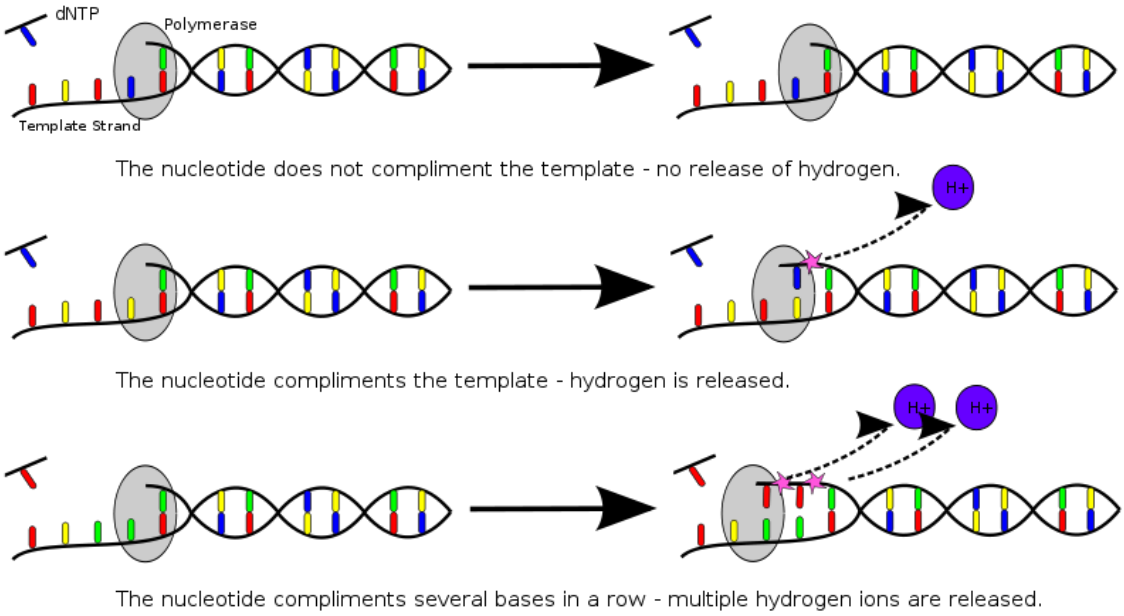
4.3.4 Ion Torrent

Při nasedání báze na DNA polymerázu dochází k uvolnění vodíkového protonu. Toho využívá metoda Ion Torrent (Ion Torrent Systems Inc 2010, později Life Technologies, nyní Thermofisher), která se zaměřuje na detekci elektrochemické změny pH za použití polovodičového čipu při uvolnění vodíkového iontu (Obr. 18, 19, 20).

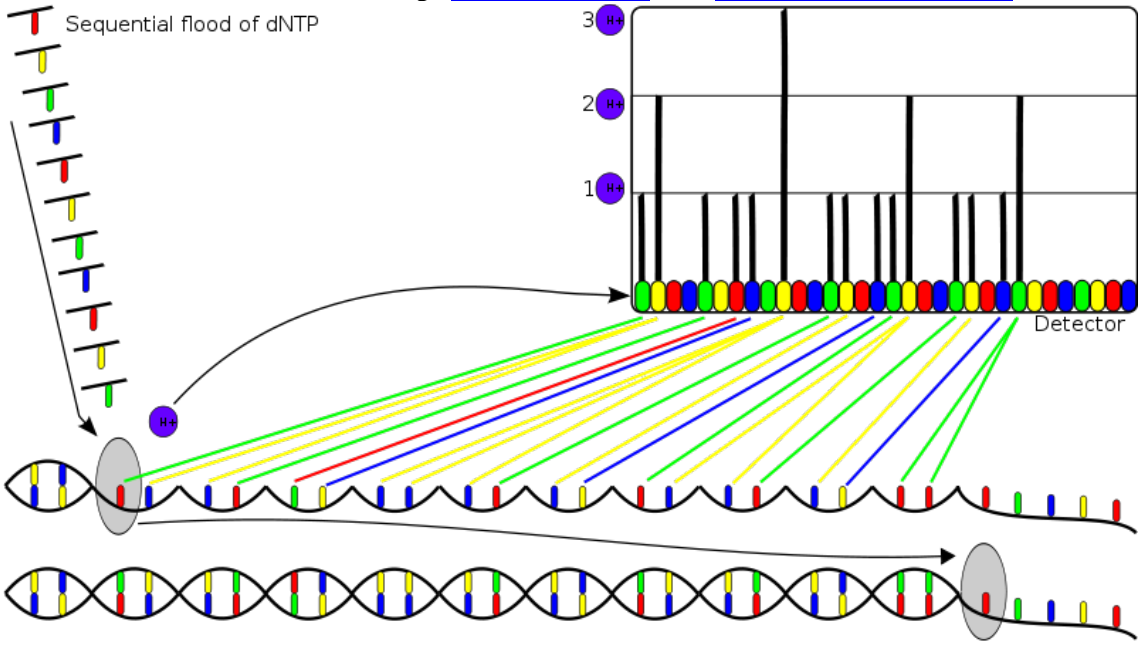
Postup pro vytvoření pair-endové knihovny je následující: DNA je nastříhána na menší fragmenty, a na oba konce fragmentů jsou naligovány adaptéry značené biotinovou značkou. Vložením do velmi zředěného roztoku dochází k samovolné cirkularizaci hybridizací právě pomocí biotinovaných adaptérů. DNA fragment je exonukleázou střižen naproti značkám a na konce fragmentu jsou naligovány templátové adaptéry. Podobně jako u metod 454 a SOLiD probíhá emulzní PCR s cílem amplifikovat značené fragmenty. Detekce probíhá na polovodičovém čipu obsahujícím řadu malých jamek s detektory. Série elektrických impulsů je přenášena z čipu do počítače a převedena do sekvence DNA. V případě, že jsou za sebou dvě stejné báze, dochází k zesílení původního signálu, nikoli k vytvoření dvou individuálních signálů, proto má tato metoda problémz s prosekvenováním dlouhých homopolymerů.



Obr. 18: Ion Torrent, Enzoklop, CC BY-SA 4.0, via Wikimedia Commons



Obr. 19: Ion Torrent, Enzoklop, CC BY-SA 4.0, via Wikimedia Commons



Obr. 20: Ion Torrent, čtení signálu, Enzoklop, CC BY-SA 4.0, via Wikimedia Commons

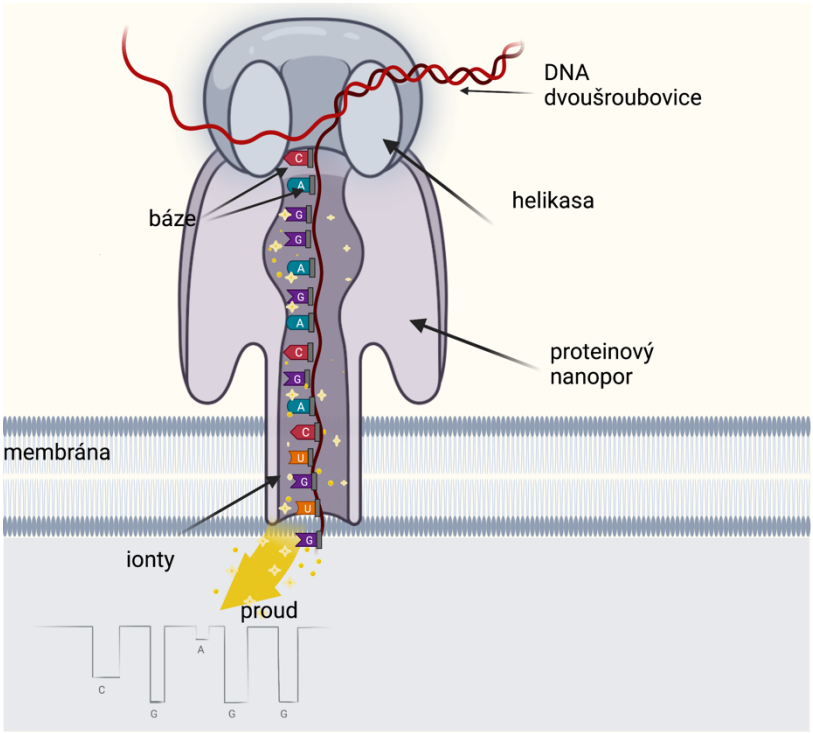
Na youtube channelu společnosti ThermoFisher lze kromě jiného nalézt i video vysvětlují IonTorrent sekvenování, https://www.youtube.com/watch?v=zBPKj0mMcDg

4.4 Třetí generace sekvenování

4.4.1 Nanopore

Nanopore (Oxford Nanopore Technologies Ltd 2014) je metoda sekvenování založena na měření elektrické vodivosti při průchodu báze membránou. Systém tvoří proteinový nanopor (Obr. 21) zabudovaný do artificiální membrány. Membránou prochází elektrický proud. Při průchodu DNA membránou dochází ke změnám proudu. Vzhledem k tomu, že se jednotlivé báze od sebe mírně odlišují, jsou tyto změny patrné i na profilu procházejícího elektrického proudu. Takto je v některých případech umožněno identifikovat i modifikace jednotlivých bazí.

Vzhledem ke kompaktnosti zařízení pro nanoporové sekvenování a nízké pořizovací ceně je možné tuto metodu využít v terénu, jako např. při epidemii Eboly v Libérii, případně pro sekvenování genomu nově objevených rostlin. Další velkou výhodou je možnost přímého sekvenování RNA místo převádění RNA na cDNA, čímž se vyhneme transkripčnímu zkreslení. Mezi hlavní nevýhody pak patří vysoká chybovost, která se pohybuje v řádu desítek procent.



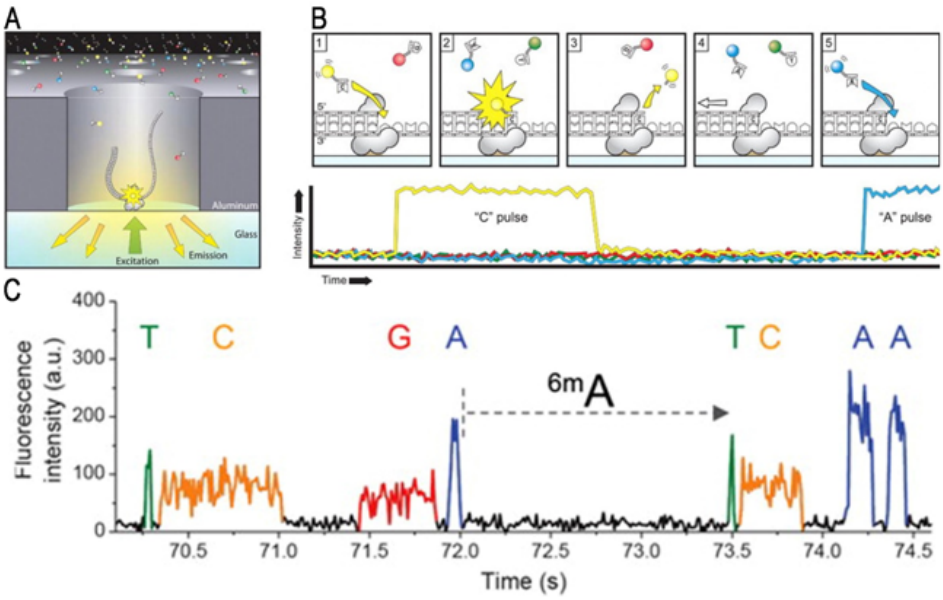
Obr. 21: Nanopore, Pfeiferová L, Created with BioRender.com

Ukázku technologie Nanopore lze nalézt na jejich youtube channelu, Oxford Nanopore Technologies, https://www.youtube.com/watch?v=RcP85JHLmnI

4.4.2 PacBio SMRT

Další metodou, která se zabývá sekvenováním jedné molekuly, je metoda SMRT (Pacific Biosciences 2009) - single molecule, real time (jedna molekula, reálný čas, Obr. 19). Nejedná se o metodu postupného inkorporování jedné značené báze. Ačkoli využívá metodu sekvenování pomocí syntézy, optický systém v zásadě vytváří film tak, jak polymeráza včleňuje fluorescenčně značené nukleotidy. Aby se zamezilo přehlušení signálu začleněné báze volnými, PacBio využívá **zero mode waveguides** (vlnovody s nulovým režimem, ZMWs). ZMWs jsou mikrojamičky o objemu 20 zL (zL = zeptolitr, tj. 10^{-21}L), ve kterých je na dně umístěna jedna molekula polymerázy a jedno cirkularizované vlákno DNA/RNA s adaptorovými sekvencemi. Mikrojamičky jsou navrženy tak, aby zachytávaly signál pouze ze dna jamičky, tedy z toho nukleotidu, který je včleňován do řetězce. Signál ostatních nukleotidů je odfiltrován. Simultánně v jamičkách čipu probíhá stovka tisíc reakcí.

Výhodou této metody je její rychlost, většinou v řádech maximálně několika hodin, tvorba řetězců až o délce desítek kilobází a možnost číst i oblasti s velmi vysokým GC obsahem. Další výhodou je možnost detekovat DNA modifikace přímo během sekvenování vzhledem k použité technologii, která snímá kinetiku inkorporování bazí do řetězce. Nevýhodou je zatížení SMRT metody poměrně velkou chybovostí dosahující až 15 % (single pass error). Na druhou stranu se jedná o náhodné chyby bez systematického efektu. Chybovost je možné dále snížit za použití **Circular Consensus Sequencing** sekvenačního módu, při kterém je vlákno DNA/RNA v mikrojamičce opakovaně kontinuálně sekvenováno.



Obr. 22: Ilustrace sekvenování metodou SMRT, Rhoads a Au, 2015

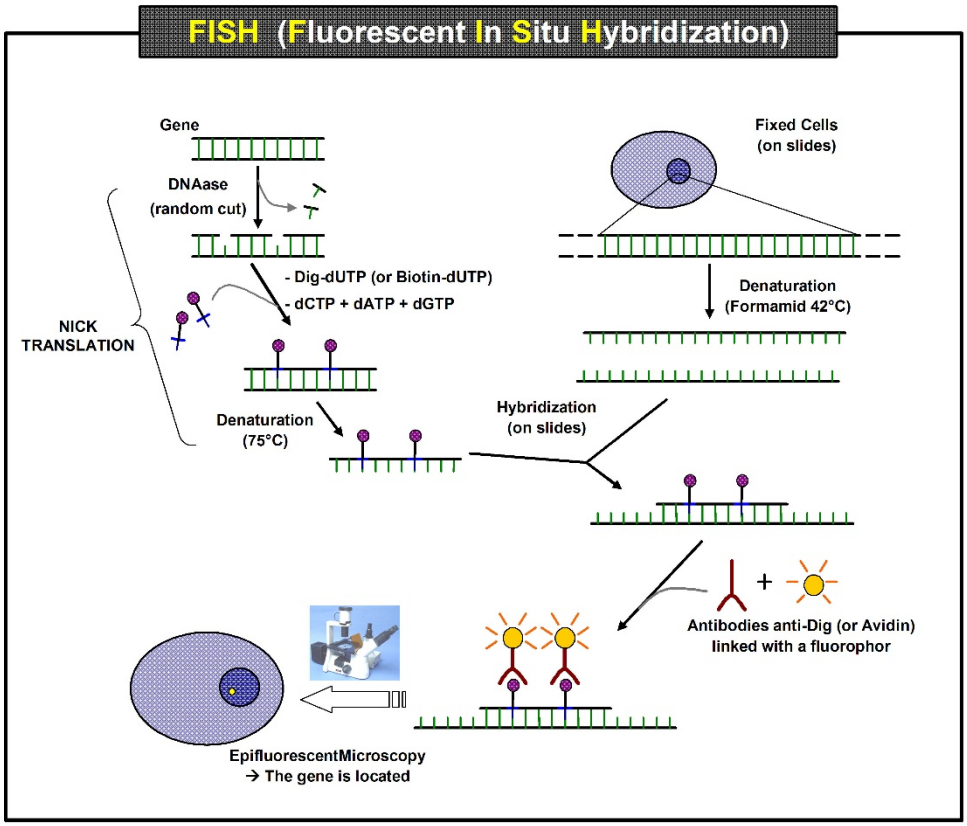
Ukázku SMRT sekvenování lze nalézt na jejich youtube channelu, PacBio PacBio Sequencing, https://www.youtube.com/watch?v=_lD8JyAbwEo

4.5 Velké genomové projekty

4.5.1 HUGO

V roce 1990 odstartoval projekt sekvenování lidského genomu, iniciovaný roku 1988 konsorciem Human Genome Project, který si dal za cíl přečíst do 15 let kompletní genetickou informaci člověka. Mezi další cíle projektu patřilo vytvoření genetické a fyzické mapy, identifikovat geny, rozvinout a inovovat technologie pro DNA sekvenování i zpracování dat a umožnit jednoduchý přístup k datům spolu se zjednodušením přenosu dat a provázáním databází.
Práce probíhala následovně:

- Laboratoře po celém světě si rozdělily úseky v lidském genomu na základě hrubé mapy lidského genomu. Metodou hybridizace (FISH, Obr. 23) byly úseky genomu namapovány na jednotlivé chromosomy, od těchto známých úseků se pak pokračovalo dále.



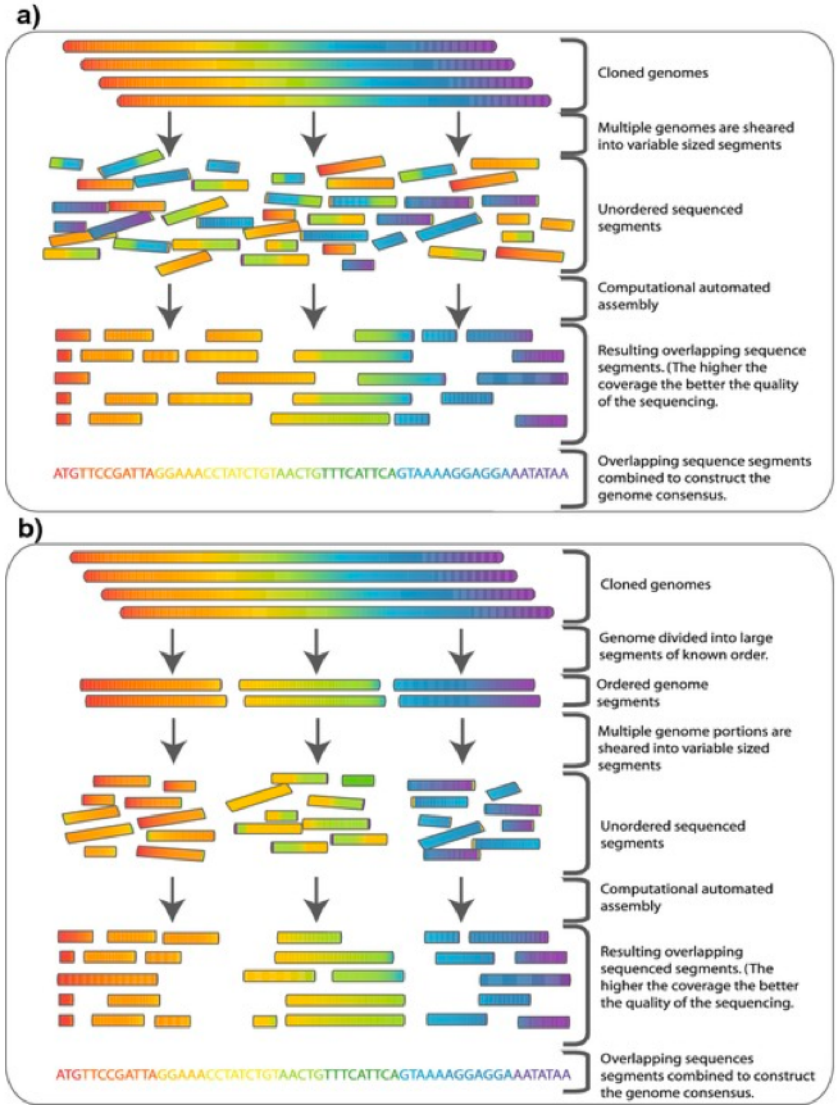
Obr. 23: FISH schéma, MrMatze, CC BY-SA 3.0 , via Wikimedia Commons

- Části byly nastříhány na ještě menší úseky o délce několik desítek až stovek tisíců bazí, těchto úseků byla následně vytvořena genomová knihovna v bakteriích, které nesly tzv. BAC klony (Bacterial Artificial Chromosome)
- Pomocí hybridizace byly identifikovány klony, které v sobě nesly ony známé namapované kousky chromosomů. V tuto chvíli vědci věděli, na jakém místě v určitém chromosomu se nachází malý kus lidského genomu. Tento kus se pak metodou Chromosome Walking mohl postupně prosekvenovat.
- Později se přišlo s rychlejším a efektivnějším řešením, jak tyto kusy lidského genomu v BAC klonech prosekvenovat – metodou Random Shotgun. BAC klon byl při této metodě rozdělen na krátké fragmenty, které se naslepo začaly sekvenovat. Po dostatečné hloubce sekvenování se na základě překryvů zpětně poskládal celý kus BAC klonu.

O této metodě mluvíme jako o tzv. **Hierarchical shotgun** (Obr. 24 b). Ačkoli byly rozděleny velké úseky lidského genomu a tyto velké úseky si následně rozdělily na menší a až ty sekvenovaly, vždy dotyční pracovníci věděli, se kterou oblastí genomu pracují.

4.5.2 Celera

Obchodník a vizionář Craig Venter přišel v roce 1998 s nápadem použít metodu **Random Shotgun** (Obr. 24 a) na celý lidský genom, tj. netvořit BAC klony, které by se mapovaly na jednotlivé chromosomy a postupně k nim přidávat informace, ale naslepo rozbít celou molekulu lidské DNA, hluboce osekvenovat (tisíce vláken DNA současně) a podle překryvů zpětně poskládat celý genom.



Obr. 24: a) Random shotgun, b) hierarchar shotgun, Commins, J., Toft, C., Fares, M. A., CC BY-SA 2.5, via Wikimedia Commons

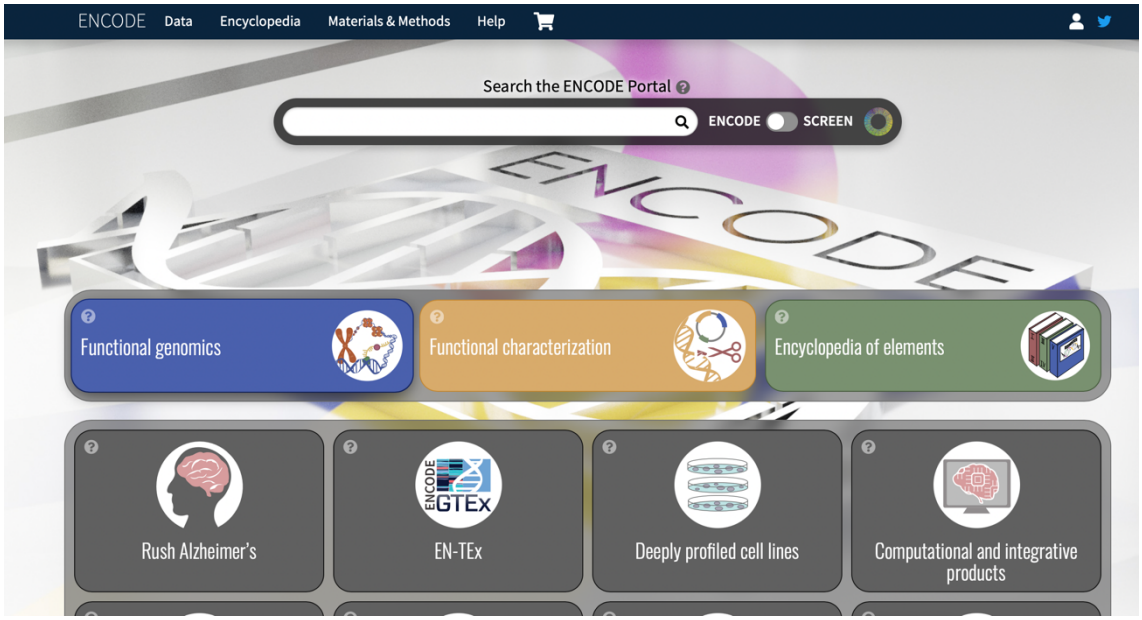
Za tímto účelem nakoupila jeho firma Celera Genomics stovky sekvenátorů, které bez přestání skoro rok chrlily jednu sekvenci za druhou. Nakonec se dostali na zhruba 27 milionů sekvencí o průměrné délce 540 bazí. Tímto způsobem bylo umožněno poskládat 90 % lidského genomu (přesněji řečeno jeho euchromatinové části, tj. aktivní DNA, ve které je naprostá většina genů).
V roce 2001 bylo oficiálně oznámeno úspěšné osekvenování celého lidského genomu, ačkoli byla místa, která osekvenována nebyla (a stále taková místa jsou, např. repetitivně oblasti).

Konsorcium T2T dokončilo sekvenaci lidského genomu v roce 2022.

Celera Genomics opublikovala svoji práci v časopisu Science, zatímco konsorcium HUGO v Nature.

4.5.3 ENCODE – Encyclopedia of DNA Elements

Projekt ENCODE (Obr. 25, 26) je veřejný výzkumný projekt zaměřený na identifikaci funkčních prvků v lidském genomu, včetně promotorů, regulačních sekvencí, hladiny RNA a odhalení okolností, za kterých se aktivuje gen. Dalším cílem projektu bylo vyhodnotit, jak mohou geny ovlivňovat lidské zdraví, a zároveň stimulovat vývoj nových léčebných postupů pro prevenci a léčbu těchto onemocnění Projekt byl zahájen v září 2003, přičemž přímo navazoval na projekt HUGO, který byl zaměřen na sekvenaci lidského genomu.



Obr. 25: Screenshot Encode web page ENCODE: Encyclopedia of DNA Elements. https://www.encodeproject.org/ (accessed Feb 2, 2022)

Na projektu ENCODE pracovalo více než 400 výzkumných pracovníků po dobu 10 let. Během pilotní fáze projektu se testovaly a porovnávaly existující metody pro analýzu na definované části lidského genomu (test probíhal na přibližně 1 % lidského genomu). Tato fáze byla zcela otevřená všem tak, aby se mohlo prozkoumat co nejvíce technik, technologií a strategií. Cílem těchto snah bylo vypracovat sadu postupů, které by umožnily komplexní identifikaci všech funkčních prvků v lidském genomu. Hlavní předpoklad byl, že zhruba 1,5 % DNA tvoří protein kódující geny (na počet přibližně 20 000), a zbytek je junk DNA, DNA bez funkce.
První výsledky, zveřejněné v roce 2007, byly shledány jako poněkud kontroverzní, a staly se terčem kritiky. Mimo jiné bylo zveřejněno

- genom obsahuje nejrůznější formy aktivních elementů, a tedy méně nepoužívaných sekvencí, než se dříve myslelo.
- přepisuje se většina genomu
- většina bází genomu je spojena s alespoň jedním primárním transkriptem a mnoho transkriptů je vzdáleně spojeno s oblastí se zavedenými protein kódujícími oblastmi
- mnoho dříve nerozpoznaných počátečních míst transkripce bylo identifikováno, z nichž mnohé vykazuje strukturu chromatinu a sekvenci specifickou pro vazebné místo proteinu podobnou dobře známým promotorům
- replikace DNA koreluje se strukturou chromatinu

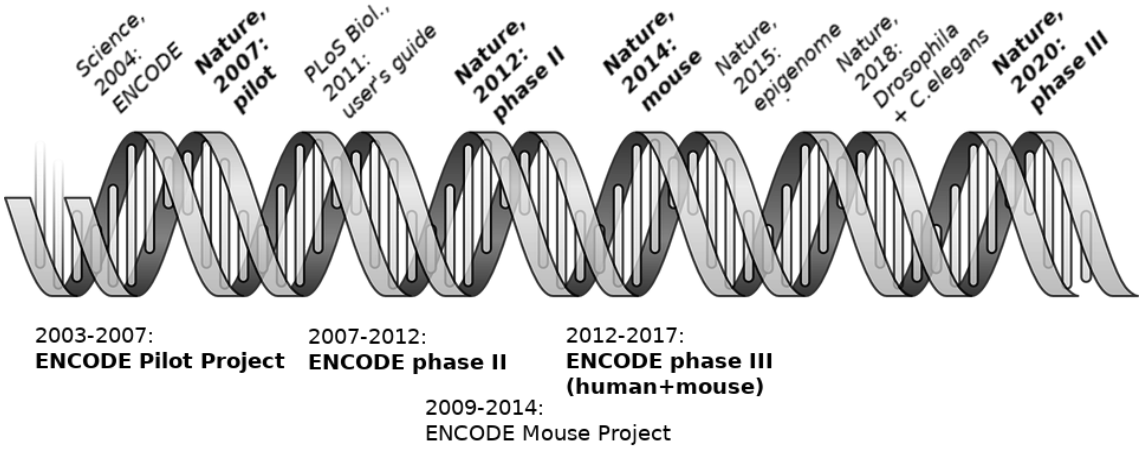
- regulační sekvence, které obklopují počáteční místa transkripce, jsou symetricky distribuovány bez tendence směřovat k upstream regionům
- souhra mezi geny, oblastmi regulujícími genovou aktivitu a dalšími úseky DNA mnohem složitější, než si kdokoliv myslel
- bylo identifikováno mnoho nových protein nekódujících transkriptů, mnohé z nich se překrývaly s protein kódujícími oblastmi a další se nacházely v oblastech genomu, o kterých se dříve myslelo, že jsou transkripčně tiché

V produkční fázi projektu pak byl analyzován celý genom. V průběhu roku 2012 bylo současně publikováno přes 30 prací. Finálně bylo vyhodnoceno, že

- přes 80 % genomu je aktivních (tj. má nějakou funkci), ačkoli se neví přesně mechanismus působení a cíl. V této aktivní části genomu bylo nalezeno přes 70 000 promotorových oblastí, a téměř 40 000 enhancerů.
- S jistotou bylo popsáno, že minimálně 9 % genomu se zapojuje do regulace genové exprese a kontroluje a reguluje syntézu proteinů. Celkově projekt ENCODE identifikoval přes 4 miliony genových spínačů rozptýlených v celém genomu s tím, že genové spínače jsou ve fyzickém kontaktu s jimi kontrolovanými geny, přestože mohou být lineárně vzdáleny stovky kilobází.
- Kromě toho, že víme, které geny ovlivňují určitá onemocnění, byly odhaleny i některé spínače podílející se na regulaci zapínání a vypínání genů. Například se ukazuje, že malá změna v genovém spínači zvaném CARD9 je spojena se zvýšením rizika rozvoje Crohnovy choroby o 20 %.

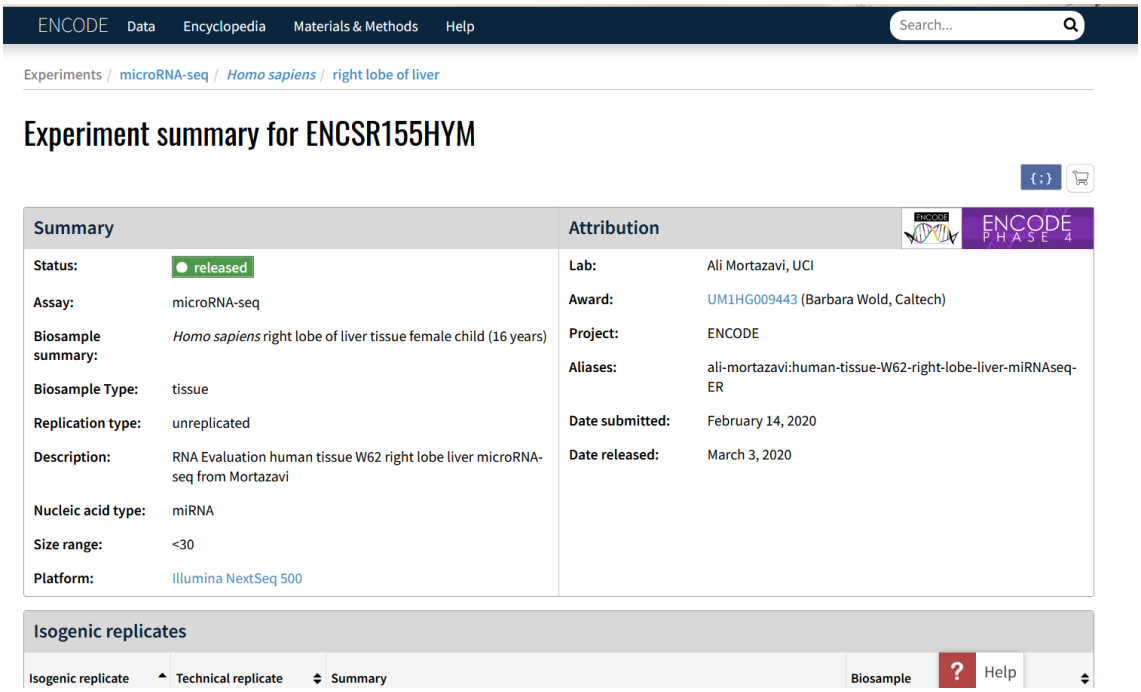
- Poznatek o vlivu genetických spínačů prohlubuje porozumění genové expresi a otevírá nové možnosti pro léčbu nejrůznějších chorob.

V třetí fázi se projekt ENCODE zaměřil na kandidáty cis-regulačních elementů člověka a myši z téměř šesti tisíc nových experimentálních datasetů. Vytvořil online registr SCREEN integrující jednotlivé složky online podoby ENCODE Encyklopedie. Stěžejní publikace jakožto fáze celého projektu ENCODE zpřehledňuje následující časová osa (Obr. 26).



Obr. 26: Přehled ENCODE publikací a jednotlivých fází projektu, Svatoňová P

Kromě publikací byla vytvořena i interaktivní ENCODE Encyklopedie (Obr. 27), organizující nejvýznamnější analytické produkty do anotací a poskytující nástroje pro jejich vyhledávání a vizualizaci. Encyklopedie má dvě úrovně anotací, základní anotace jsou odvozeny přímo z experimentálních dat, typicky vyrobených jednotnou pipeline a anotace integrační, které spojují různé typy experimentálních dat se základními anotacemi.
Základní anotace jsou např. otevřená chromatinová místa (DNAse1-Seq), histonové nabohacení (ChIP-Seq), genová exprese (RNA-seq), DNA methylace a další. Integrační level obsahuje informace z The Registry of Candidate cis-Regulatory Elements (cCRE, integruje všechna vysoce kvalitní data z DNA-seq a H3K4me3, H3K27ac a CTCF ChIP-seq vytvořená konsorcii ENCODE a Roadmap Epigenomics, klasifikuje cCRE do skupin podobných promotorům, enhancerům, DNázám-H3K4me3 a CTCF agnostickým způsobem podle buněčného typu), z FunSequ (další zdroj ENCODE pro anotaci germinálních i somatických variant, zejména v nekódujících oblastech rakovinových genomů), další anotace variant z HaploReg, RegulomeDB a další databáze.



Obr. 27: Screenshot Encode web page, Experiment summary for ENCSR155HYM. ENCODE: Encyclopedia of DNA Elements. https://www.encodeproject.org/experiments/ENCSR155HYM (accessed April 09, 2020)

4.5.4 HapMap

V říjnu 2002 odstartoval mezinárodní vědecký projekt zaměřený na vytvoření mapy haplotypů lidského genomu (haplotyp = skupina alel na určitém místě na chromosomu, která je mezi generacemi přenášena současně) a popsání běžných vzorců variant. HapMap se používá k hledání genetických variant ovlivňujících zdraví, nemoci a reakce na léky a faktory prostředí. První sada výsledků byla zveřejněna v roce 2005, druhá sada výsledků byla zveřejněna v roce 2007. Sada výsledků z fáze III byla zveřejněna v roce 2009.
V pilotní fázi I a II byly charakterizovány haplotypy z populací Yoruby v Nigérii (trio 30 dospělých lidí s oběma rodiči), Japonska (44 lidí bez vzájemných vztahů), Číny (45 lidí bez vzájemných vztahů) a Spojených států (30 trojic, obyvatelé Utahu s předky ze severní a západní Evropy). Ve fázi III byly přidány populace lidí z jihozápadu USA s předky z Afriky, z Denveru v USA s předky z Číny, Gujarati z Houstonu v USA, Luhya a Maasai z Keni, z Los Angeles v USA s předky v Mexiku a lidé z Toskánska v Itálii.

4.5.5 1000 Genome, 1000000 Genomes a 1+ Million Genomes Projects

1000 Genome Project

Mezi lety 2008 až 2015 probíhal mezinárodní výzkumný projekt s cílem vytvořit podrobný katalog lidských genetických variant, který by podporoval budoucí medicínské výzkumné projekty. Cílem projektu 1000 Genome bylo vytvořit katalog téměř všech genetických variant (varianty s frekvencí nejméně 1 % v populaci), včetně SNP a strukturálních variant, a jejich kontext v haplotypu. 1000 Genome Project navázal na projekt HapMan.
Pilotní fáze zahrnovala tři proudy: prvním bylo zjistit, zda je coverage 4x dostatečná pro odhalení hlavních cílů výzkumu, druhý směr se zabýval sekvenací trojic matka-otec-dítě s cílem posoudit pokrytí a platformy a centra a třetí se zabýval osekvenováním 1000 oblastí genů v 900 vzorcích s cílem posoudit metody pro zachycení genové oblasti.
Hlavní fáze projektu se skládala ze 3 částí: fáze 1 a fáze 3 se zabývala tvorbou dat, zatímco fáze 2 technologických vývojem.

Celkově bylo objeveno a charakterizováno více jak 88 milionů variant (z toho 84,7 milionů SNP, 2,6 milionu indelových variant a 60 tisíc strukturních variant).

100000 Genomes Project

Pokračovacím projektem v letech 2015 až 2018 byl britský projekt 100000 Genomes – osekvenování genomů pacientů s rakovinou, vzácným onemocněním nebo infekčním stavem, s cílem propojit sekvenační data do standardizované a rozšiřitelné zprávy ohledně diagnózy, léčby a výsledků.

1+ Million Genomes

Nejnovějších projektem je 1+ Million Genomes, iniciativa se záměrem vytvořit síť genetických a klinických dat v rámci celé Evropy. Cílem projektu je mimo jiné umožnení porovnání genetické a klinické informace lidí, což by mohlo napomoci k rychlejší detekci onemocnění, jeho vývoje a možnostech léčby.

4.5.6 Earth BioGenome Project

Earth BioGenome Project (EBP, 2018-2028) je globální iniciativa, jejímž finálním plánem je osekvenování všech eukaryotických forem života. EBP plánuje připravit kompletní katalog všech druhů rostlin, zvířat, hub a jednobuněčných eukaryot. Mezi cíle iniciativy patří zlepšení života obyvatelstva (vývoj nových léčeb pro infekční a dědičné choroby, vytvoření nových biosyntetických paliv, tvorba nových biomateriálů, odhalení léčiv, která by mohla pomoci zpomalit/zastavit stárnutí a postupy, jak zastavit hladomor), ochrana biologické rozmanitosti a porozumění ekosystémům.

4.6 Otázky k tématu

1. Co mají společného metody sekvenování nové generace?
2. Zjednodušeně popište princip klastrové (můstkové) PCR.
3. Jakou rozdílnou metodu od ostatních NGS metod využívá SOLiD pro sekvenování?
4. Jak se liší metody NGS od Třetí generace sekvenování?
5. Co je nanopor? Kde se využívá nanoporové sekvenování?
6. Popište zjednodušeně metodu SMRT.

4.7 Zdroje

Úvod do sekvenování

Applied Biosystem Genetic Analysis Systems. Thermo Fisher Scientific. https://www.thermofisher.com/cz/en/home/life-science/sequencing/sanger-sequencing/sanger-sequencing-technology-accessories.html (accessed Aug 08, 2021).

Ion GeneStudio S5 Specs. Thermo Fisher Scientific. https://www.thermofisher.com/cz/en/home/life-science/sequencing/next-generation-sequencing/ion-torrent-next-generation-sequencing-workflow/ion-torrent-next-generation-sequencing-run-sequence/ion-s5-ngs-targeted-sequencing/ion-s5-specifications.html (accessed Aug 08, 2021).

SOLiD Next-Generation Systems & Accesories. Thermo Fisher Scientific. https://www.thermofisher.com/cz/en/home/life-science/sequencing/next-generation-sequencing/solid-next-generation-sequencing/solid-next-generation-sequencing-systems-reagents-accessories.html.html (accessed Aug 08, 2021).

Product comparison. Oxford Nanopore Technologies. https://nanoporetech.com/products/comparison (accessed Aug 08, 2021).

Data concordance between the NextSeq™ 1000, NextSeq 2000, and NextSeq 550 Sequencing Systems. Illumina Sequencing and array-based solutions for genetic research. https://www.illumina.com/content/dam/illumina-marketing/documents/products/appnotes/nextseq-1000-2000-data-concordance-app-note-970-2020-001.pdf (accessed Aug 08, 2021).

Klasické metody sekvenování

Smith LM, Fung S, Hunkapiller MW, Hunkapiller TJ, Hood LE. The synthesis of oligonucleotides containing an aliphatic amino group at the 5' terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis. *Nucleic Acids Res*. 1985;13(7):2399-2412. doi:10.1093/nar/13.7.2399

Maxam AM, Gilbert W. A new method for sequencing DNA. Proc Natl Acad Sci U S A. 1977;74(2):560-564. doi:10.1073/pnas.74.2.560

Sanger F, Nicklen S, Coulson AR. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A. 1977;74(12):5463-5467. doi:10.1073/pnas.74.12.5463

Yeo, L.Y., Chang, H.-C., Chan, P.P.Y. and Friend, J.R. (2011), Microfluidic Devices for Bioapplications. Small, 7: 12-48. https://doi.org/10.1002/smll.201000946

Sekvenování nové generace

Behjati S, Tarpey PS. What is next generation sequencing?. Arch Dis Child Educ Pract Ed. 2013;98(6):236-238. doi:10.1136/archdischild-2013-304340
Brown, T.; Brown, D.; Brown, T (Jnr).; Brown, A. ATDBio - Sequencing, forensic analysis and genetic analysis. https://atdbio.com/nucleic-acids-book/Sequencing-forensic-analysis-and-genetic-analysis (accessed Feb 14, 2022).
Edwars, R. Why does paired end sequencing help assembly?. https://www.youtube.com/watch?v=WTbnk91e2WU (accessed April 03, 2022).
Head SR, Komori HK, LaMere SA, et al. Library construction for next-generation sequencing: overviews and challenges. Biotechniques. 2014;56(2):61-passim. Published 2014 Feb 1. doi:10.2144/000114133
Illumina Illumina Sequencing by Synthesis. https://www.youtube.com/watch?v=fCd6B5HRaZ8&t=112s (accessed April 03, 2022).
LabGuide.cz - Průvodce laboratoří. https://labguide.cz/ (accessed Dec 27, 2020).
Muzzey D, Evans EA, Lieber C. Understanding the Basics of NGS: From Mechanism to Variant Calling. Curr Genet Med Rep. 2015;3(4):158-165. doi:10.1007/s40142-015-0076-8
Slatko BE, Gardner AF, Ausubel FM. Overview of Next-Generation Sequencing Technologies. Curr Protoc Mol Biol. 2018;122(1):e59. doi:10.1002/cpmb.59
Voelkerding KV, Dames SA, Durtschi JD. Next-generation sequencing: from basic research to diagnostics. Clin Chem. 2009;55(4):641-658. doi:10.1373/clinchem.2008.112789

Třetí generace sekvenování

Ardui S, Ameur A, Vermeesch JR, Hestand MS. Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics. Nucleic Acids Res. 2018;46(5):2159-2168. doi:10.1093/nar/gky066
Delivering highly accurate long reads to drive discovery in life science. PacBio. https://www.pacb.com/smrt-science/smrt-sequencing/ (accessed Dec 09, 2020).
Eisenstein, M. An ace in the hole for DNA sequencing. Nature 550, 285–288 (2017). https://doi.org/10.1038/550285a
How it works. Oxford Nanopore Technologies. https://nanoporetech.com/how-it-works (accessed Dec 09, 2020).
Hoenen T, Groseth A, Rosenke K, et al. Nanopore Sequencing as a Rapidly Deployable Ebola Outbreak Tool. Emerg Infect Dis. 2016;22(2):331-334. doi:10.3201/eid2202.151796

Mondal TK, Rawal HC, Gaikwad K, Sharma TR, Singh NK. First de novo draft genome sequence of Oryza coarctata, the only halophytic species in the genus Oryza. F1000research. 2017 ;6:1750. DOI: 10.12688/f1000research.12414.1.
Oxford Nanopore Technologies How nanopore sequencing works. https://www.youtube.com/watch?v=RcP85JHLmnI (accessed April 03, 2022).
PacBio PacBio Sequencing – How it Works. https://www.youtube.com/watch?v=_lD8JyAbwEo (accessed April 03, 2022).
Rhoads A, Au KF. PacBio Sequencing and Its Applications. Genomics Proteomics Bioinformatics. 2015;13(5):278-289. doi:10.1016/j.gpb.2015.08.002

Velké genomové projekty

About IGSR and the 1000 Genomes Project. IGSR: The International Genome Sample Resource. https://www.internationalgenome.org/about (accessed Dec 09, 2020).
Beyond One Million Genomes (B1MG) Project. https://b1mg-project.eu/ (accessed March 13, 2022).
ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep 6;489(7414):57-74. doi: 10.1038/nature11247.
ENCODE Project Consortium, Birney E, Stamatoyannopoulos JA, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature. 2007;447(7146):799-816. doi:10.1038/nature05874
ENCODE Encyclopedia Version 5: Genomic and Transcriptomic Annotations. ENCODE: Encyclopedia of DNA Elements. https://www.encodeproject.org/data/annotations/ (accessed Dec 09, 2020).
Earth BioGenome Project. https://www.earthbiogenome.org (accessed April 03, 2022).
Genomics England 100,000 Genomes Project | Genomics England. https://www.genomicsengland.co.uk/initiatives/100000-genomes-project (accessed April 03, 2022).
Gonzaga-Jauregui C, Lupski JR, Gibbs RA. Human genome sequencing in health and disease. Annu Rev Med. 2012;63:35-61. doi:10.1146/annurev-med-051010-162644
Green ED, Watson JD, Collins FS. Human Genome Project: Twenty-five years of big biology. Nature. 2015;526(7571):29-31. doi:10.1038/526029a
Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome [published correction appears in Science 2001 Jun 5;292(5523):1838]. Science. 2001;291(5507):1304-1351. doi:10.1126/science.1058040
The Encyclopedia of DNA Elements (ENCODE). National Human Genome Research Institute. https://www.genome.gov/Funded-Programs-Projects/ENCODE-Project-ENCyclopedia-Of-DNA-Elements (accessed Dec 09, 2020).
International HapMap Project. National Human Genome Research Institute Home. https://www.genome.gov/10001688/international-hapmap-project (accessed Dec 09, 2020).

5 DATA A DATOVÉ TYPY

Jelikož nedílnou součástí práce bioinformatika je analýza, zpracování, konverze formátů a ukládání dat, je nezbytné mít přehled o tom, co jsou data.

Data obecně jsou sady charakteristik, které se používají pro popis určitého jevu či efektu. Data se shromažďují a ukládají s cílem vytvořit na základě jejich analýzy nějaký předpoklad či hypotézu. Údaje uložené v datech mohou mít kvantitativní (numerické) i kvalitativní (textové, zvukové, obrazové) vlastnosti. Data, která se v bioinformatice používají, většinou zahrnují sekvence genů, exonových a intronových oblastí, celých genomů. Dále mohou obsahovat informace o proteinových strukturách, včetně terciální struktury, expresi mRNA, protein-protein interakcích, protein-DNA interakcích a další informace z omics oborů (proteomika-proteomics, genomika-genomics, metabolomika-metabolomics a další).

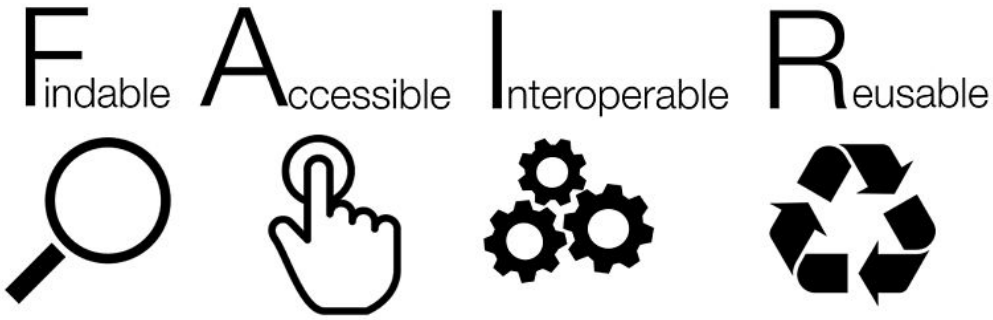
5.1 Raw data

Raw data (primární data) jsou data získaná přímo ze zdroje, bez jakéhokoli zpracování. Existují captured data, která se sbírají přímo za účelem následné analýzy, a exhaust data, která jsou vedlejším produktem hlavních funkční výpočetních systémů. Typická capture data jsou LOG soubory, cookies a dočasné (temporary) soubory.

5.2 FAIR data

Vzhledem k množství dat (a ceně), která jsou generována moderními sekvenačními metodami, vyvstala naléhavá potřeba vytvořit funkční infrastrukturu, která by podporovala opakované použití dat z vědeckých projektů. Byla vytvořena skupina se zástupci akademické obce, průmyslu, grantových agentur i vydavatelů, která se zaměřila na vytvoření jednoduchých zásad pro uložení dat z projektů a umožnění jejich zpracování – FAIR Data Principy. Kromě využitelnosti ostatními vědci je zde kladen důraz i na umožnění automatizovaného strojového zpracování a vyhledávání uložených dat, s cílem podpořit jejich opětovné využití jedinci.

FAIR data (Obr. 28) jsou ta, která splňují podmínku pro vyhledatelnost (**F**indable), přístupnost (**A**ccesible), interoperabilitu (**I**nteroperable) a znovupoužitelnost (**R**eusable). Nejčastěji se jedná o zpřístupnění dat publikovaných v rámci vědeckého projektu, případně i o nepublikované datasety. V případě, že není možné z různých důvodů zveřejnit plná data, je dobré zveřejnit alespoň podrobná metadata.



Obr. 28: FAIR data, Fair data, SangyaPundir, CC BY-SA 4.0, via Wikimedia Commons

FINDABLE: data mají unikátní a trvalý identifikátor, který umožňuje dohledání datasetu. Kromě vyhledání člověkem se jedná i o umožnění automatizovaného strojového čtení. Kromě toho by k datům měla existovat i dohledatelná, a stejným stylem označená metadata.

ACCESIBLE: data jsou přístupná za použití standartních nástrojů, použitý protokol by měl být veřejně dostupný a zdarma. Je možné vyžadovat ověření uživatele a jeho oprávnění. Accesible neznamená, že dataset má být volně a zdarma k dispozici široké společnosti, ale že má přesně určené podmínky, za kterých přístupný je. V ideálním případě je stroj schopen automaticky pochopit požadavky, a pak je buď automaticky provést nebo na ně upozornit uživatele. Typicky se jedná o povinnost použít https nebo ftp, prokázat se přihlašovacími údaji univerzity/vědecké společnost nebo třeba provést ověření po telefonu. Z etického hlediska není možné mít veškerá data naprosto volně přístupná veřejnosti.

INTEROPERABLE: data musí být integrovatelná s dalšími daty, standartními způsoby zpracování a ukládání. Data jsou prezentována standardizovanými a zdokumentovanými postupy, popisy používají standardizované slovníky, terminologie a ontologie. Při navázání dalších dat jsou data popsána tak, aby bylo možné pochopit jejich vzájemný vztah.

REUSABLE: metadata a data by měla být co nejpřesněji popsána. Standardem by měl být komplexní popis podmínek, za jakých došlo k jejich generování, včetně typu přístroje a softwarového vybavení, a informace o původu dat a možnosti použití. Mělo by být umožněno provedení replikace analýzy, případně reanalýza za změněných podmínek.

FAIR data nejsou synonymem pro open acces data!

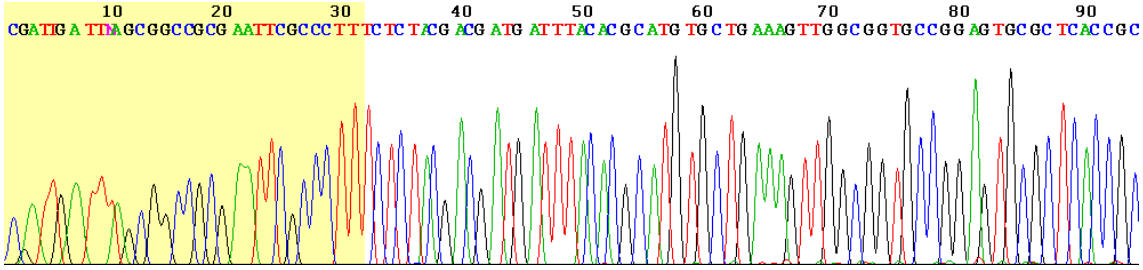
Principy Fair Guiding (Wilkinson et al, 2016),
https://www.nature.com/articles/sdata201618

5.3 Datové typy a formáty

Vzhledem k množství různorodých dat zpracovávaných v bioinformatice a bioinformatických analýzách, jako např. informace o zarovnání sekvencí, anotace, strukturní informace o proteinech, údaje o datové expresi, informace získané z databází a popisů experimentů a platforem, je zřejmé, že bude existovat i velké množství datových typů a datových formátů, které tyto informace ukládají. Běžně se formáty dělí na textové a binární.
Kromě běžných typů formátů (BAM, VCF, GTF), které je možné zpracovat a upravit velkým množstvím nástrojů, se vyskytují i softwarově specifické formáty, které se dají otevřít a zpracovat pouze určitým softwarovým nástrojem (často se to stává u komerčních programů).

5.3.1 SCF (ABI, AB1)

SCF (a formáty ABI, AB, AB1) slouží k ukládání dat z fluorescenčních sekvenátorů (Obr. 29). Vzhledem k tomu, že se dodnes běžně v nemocnicích potvrzují nalezené patologické varianty z NGS metod pomocí upravené Sangerovy metody, je poměrně výhodné mít povědomí i o tomto formátu.
Každý soubor obsahuje data z jediného čtení. Soubor začíná hlavičkou, dále obsahuje sekvenci (sample points), pozice bazí vzhledem k sekvenci a číselné odhady přesnosti každé báze. Volitelně může obsahovat komentáře a privat data (místo k ukládání informací, které nejsou formátem podporovány).



Obr. 29: Zobrazení výstupu chromatografu, Loris, Public domain, via Wikimedia Commons

5.3.2 FASTA

FASTA je jednoduchý textový formát pro reprezentaci nukleotidových nebo peptidový sekvencí. Nukleotidy nebo aminokyseliny jsou zapsány použitím jednopísmenkových kódů. Minimální počet řádku je dva: hlavička/řádek s komentářem, který začíná znakem >, většinou obsahuje identifikátor sekvence, a druhý řádek obsahující sekvenci. Volitelně může obsahovat řádek komentáře začínající znakem **;**, který je při zpracování ignorován. Ve formátu FASTA se např. zapisují referenční sekvence, které můžeme získat z databáze RefSeq pod NCBI, případně mohou být vstupními daty pro algoritmus BLAST.

5.3.3 FASTQ

FASTQ (Obr. 30) je textový formát, který seskupuje dohromady sekvence a jejich odpovídající skóre kvality. Každý záznam tu je rozdělen na 4 řádky:

Řádek 1: začíná znakem @ a jedná se o identifikátor sekvence a volitelný popis

Řádek 2: sekvence uložená ve standardním jednopísmenném kódu

Řádek 3: komentář, začíná znakem +, a za ním případně následuje identifikátor sekvence

Řádek 4: hodnocení kvality sekvence, musí obsahovat stejný počet symbolů jako písmena v sekvenci. Ukázkové Phred score zapsané ve FASTQ souboru může vypadat takto: AAAFFJJJJJJJJJJJJJJJJFJJFJJJJJFJJJJJJJJJJJJJJJ#FJ#JJJJF#F#FJJ#F#JJJFJJJJJ.

Obr. 30: Ukázka FASTQ, vlastní data

Kvalita se určuje podle Phred skóre Q = -10 log(p). Ukazuje, jak si můžeme být jisti, že báze byla sekvenována a identifikována správně. Malé p ve vzorci je pravděpodobnost, že uvedená báze je nesprávná (Tab. 5).

Tab. 5: Phred skore a přesnost zapsané báze

Phred quality skóre	Pravděpodobnost, že je daná báze uvedena nesprávně	Přesnost zapsané báze
10	1 in 10	90 %
20	1 in 100	99 %
30	1 in 1000	99,90 %
40	1 in 10000	99,99 %
50	1 in 100000	~ 100,00 %

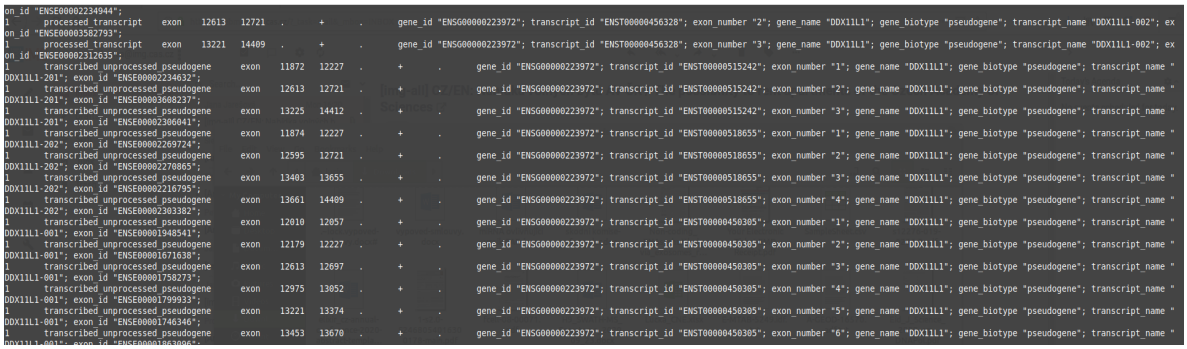
Spíše než číselnými hodnotami je Phred skóre uváděno znaky ASCII od 33 do 126 (33 až 126 jsou kódy pro jednotlivé znaky, takže skóre může být reprezentováno jedním znakem, Tab. 6). FASTQ-Sanger udává škálu Phred skóre od 0-93, FASTQ-Illumina udává Phred skóre mezi 0 a 62.

Tab. 6: Příklad Phred skore

Dec	Char	PHRED	Dec	Char	PHRED	Dec	Char	PHRED
33	!	0	73	I	40	93	q	80
34	“	1	74	J	41	94	r	81

5.3.4 GFF/GTF a GFF3

GFF – Generic Feature Format, je standardní formát souboru pro ukládání genomických funkcí. Standartně je tvořen 9 sloupci oddělenými tabulátory, které nesmí obsahovat prázdnou množinu (místo vynechání místa se používá znak **.**) a případně ještě řádky s komentáři (Obr. 31, Tab. 7 a 8).



Obr. 31: Ukázka GFF, vlastní data

Tab. 7: Povinné sloupce GFF

Pole	Popis
Seqname	název chromosomu nebo scaffoldu; jména chromosomů mohou být uvedena s předponou **chr** nebo bez ní. Seqname by mělo mít jako standartní název nebo Ensembl identifikátor, jako je ID scaffoldu, bez jakéhokoli dalšího přidaného obsahu jako je druh organismu nebo assembly
source	název programu, který tuto feature vygeneroval, nebo zdroj dat (název databáze nebo projektu)
feature type	název feature, typicky gen, exon a další
start	Počáteční pozice * prvku, se sekvenčním číslováním začínajícím na 1
end	Konečná pozice * prvku, se sekvenčním číslováním začínajícím na 1
score	hodnota skóre s desetinnou čárkou
strand	+ (forward) nebo - (reverse)
frame	hodnoty **0**, **1** nebo **2, u protein kódujících sekvencí**. 0 znamená, že první bází prvku je první báze kodonu, 1 znamená, že druhá báze je první bází kodonu atd.
group/feature	group: Všechny řádky se stejnou skupinou jsou spojeny dohromady do jedné položky feature: seznam tag-hodnota oddělený středníkem, poskytující další informace o každé vlastnosti (feature), každá dvojice je oddělena mezerou

GTF – General Transfer Format, je specifický druh GFF verze 2. Osm sloupců je shodných s GFF, sloupec feature podporuje i hodnoty 5UTR, 3UTR, inter, inter_CNS, and intron_CNS. Seznam feature musí začínat jedním ze dvou povolených atributů: gene-id nebo trascript-id.

Tab. 8: Tabulka GFF3, GFFC formát vychází z GTF

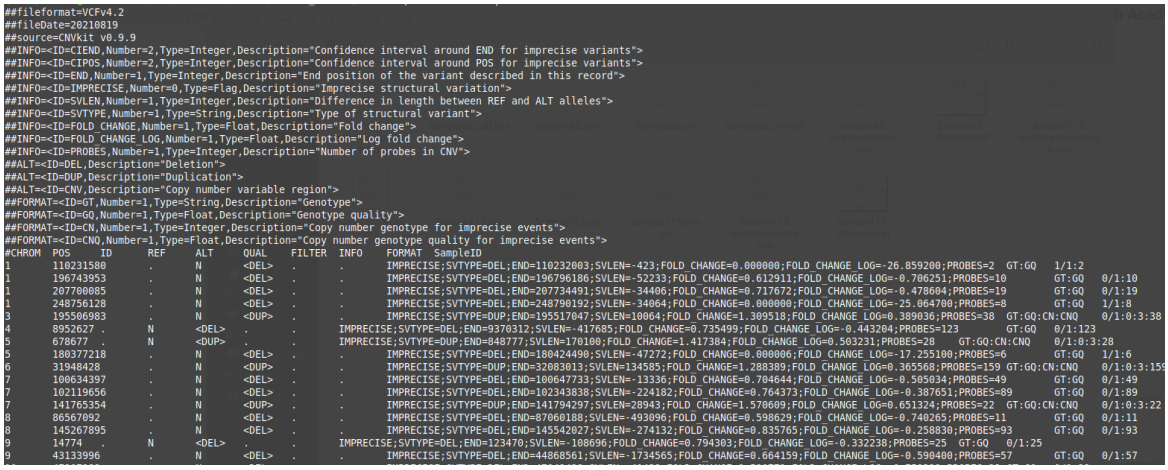
Pole	Popis
seqid	název chromosomu nebo scaffoldu; jména chromosomů mohou být uvedena s předponou **chr** nebo bez ní. Seqname by mělo mít jako standartní název nebo Ensembl identifikátor, jako je ID scaffoldu, bez jakéhokoli dalšího přidaného obsahu jako je druh organismu nebo assembly
source	název programu, který tuto funkci vygeneroval, nebo zdroj dat (název databáze nebo projektu)
type	název vlastnosti, musí to být výraz ze SOFA sekvenční ontologie
start	Počáteční pozice * prvku, se sekvenčním číslováním začínajícím na 1.
end	Konečná pozice * prvku, se sekvenčním číslováním začínajícím na 1.
score	hodnota skóre s desetinnou čárkou
strand	+ (forward) nebo - (reverse)
phase	hodnoty **0**, **1** nebo **2**. 0 znamená, že první bází prvku je první báze kodonu, 1 znamená, že druhá báze je první bází kodonu atd.
attribute	seznam tag-hodnota oddělený středníkem, poskytující další informace o každé vlastnosti (feature), některé tagy jsou předdefinovány: ID, Name, Alias, Parent; každá dvojice je oddělena mezerou

GFF3 souborový formát vychází z GTF. GFF2 může obsahovat pouze dvě vlastnosti hierarchie, zatímco GFF3 může obsahovat libovolné prvky. GFF2 také nevyžaduje, aby sloupec 3, feature type, byl součástí sekvenční ontologie. Může to být jakýkoli řetězec

5.3.5 VCF

VCF – Variant call format, je textový formát používaný v bioinformatice pro ukládání variant genových sekvencí. Variantami se myslí SNP, inzerce, delece, CNV a další strukturní varianty. Na rozdíl od GFF formátu neobsahuje redundantní informace. U VCF stačí uložit variantu spolu s referenčních genomem. Ačkoli byl původně vytvořen pro potřeby velkých projektů jako 1000 Genome Project, ten momentálně využívá svoji vlastní specifikaci pro strukturní varianty, Genomic VCF (gVCF).

Soubor se skládá z hlavičky začínající znaky ## a minimálně 8 tabulátory oddělenými sloupci (Obr. 32, Tab. 9). Informace v hlavičce obsahují meta-informace, jako INFO, FILTER a FORMAT.



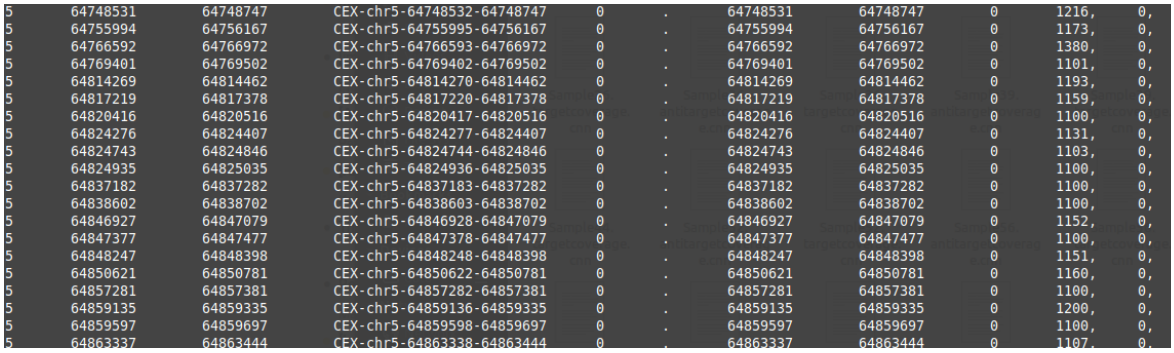
Obr. 32: Ukázka VCF souboru pro zápis CNV variant

Tab. 9: Pole ve formátu VCF

	Pole	Popis
1	CHROM	Jméno sekvence (typicky chromosomu), na kterém se nachází daná varianta. Většinou se jedná o referenční sekvenci
2	POS	Pozice varianty na dané sekvenci
3	ID	Identifikátor varianty, typicky RS číslo (dbSNP), v případě, že je neznámý, použije se zápis pomocí **..** Pokud je zapisováno více identifikátorů, vkládá se mezi ně středník bez mezer
4	REF	Referenční báze/alela
5	ALT	Seznam alternativních alel na dané pozici
6	QUAL	Skóre kvality
7	FILTER	Pole označující, kterým z dané sady filtrů varianta prošla
8	INFO	Seznam párů klíč-hodnota popisující variantu.
9	FORMAT	volitelný sloupec, položky popisující vzorky (např. ADF je hloubka osekvenování pro každou alelu na forwardovém vláknu, GT znamená Genotyp, ...)
+	SAMPLES	volitelný sloupec, pro každý vzorek popsaný v souboru, hodnoty jsou uvedeny podle pole FORMAT

5.3.6 BED

Formát BED – Browser Extensible Data, je formát používaný na definování dat použitých v anotačním tracku v genomových prohlížečích. Každá vlastnost je popsána na samostatném řádku, pro každý řádek je minimální počet sloupců 3 (Obr. 30, Tab. 10), volitelných sloupců je dokonce 9 (Tab. 11), přičemž pořadí volitelných sloupců je neměnné. To znamená, že pokud chci použít volitelný sloupec 8, musím vyplnit hodnoty i pro 1 až 7. Jednotlivé sloupce mohou být odděleny tabulátory nebo mezerami. Soubor může obsahovat tzv. track lines obsahují informace o konfiguraci zobrazení řádků, jméno, popis, nastavení useScore a itemRgb. Track lines by měly být umístěny nad řádky, které chtějí ovlivnit.



5	64748531	64748747	CEX-chr5-64748532-64748747	0	.	64748531	64748747	0	1216,	0,
5	64755994	64756167	CEX-chr5-64755995-64756167	0	.	64755994	64756167	0	1173,	0,
5	64766592	64766972	CEX-chr5-64766593-64766972	0	.	64766592	64766972	0	1380,	0,
5	64769401	64769502	CEX-chr5-64769402-64769502	0	.	64769401	64769502	0	1101,	0,
5	64814269	64814462	CEX-chr5-64814270-64814462	0	.	64814269	64814462	0	1193,	0,
5	64817219	64817378	CEX-chr5-64817220-64817378	0	.	64817219	64817378	0	1159,	0,
5	64820416	64820516	CEX-chr5-64820417-64820516	0	.	64820416	64820516	0	1100,	0,
5	64824276	64824407	CEX-chr5-64824277-64824407	0	.	64824276	64824407	0	1131,	0,
5	64824743	64824846	CEX-chr5-64824744-64824846	0	.	64824743	64824846	0	1103,	0,
5	64824935	64825035	CEX-chr5-64824936-64825035	0	.	64824935	64825035	0	1100,	0,
5	64837182	64837282	CEX-chr5-64837183-64837282	0	.	64837182	64837282	0	1100,	0,
5	64838602	64838702	CEX-chr5-64838603-64838702	0	.	64838602	64838702	0	1100,	0,
5	64846927	64847079	CEX-chr5-64846928-64847079	0	.	64846927	64847079	0	1152,	0,
5	64847377	64847477	CEX-chr5-64847378-64847477	0	.	64847377	64847477	0	1100,	0,
5	64848247	64848398	CEX-chr5-64848248-64848398	0	.	64848247	64848398	0	1151,	0,
5	64850621	64850781	CEX-chr5-64850622-64850781	0	.	64850621	64850781	0	1160,	0,
5	64857281	64857381	CEX-chr5-64857282-64857381	0	.	64857281	64857381	0	1100,	0,
5	64859135	64859335	CEX-chr5-64859136-64859335	0	.	64859135	64859335	0	1200,	0,
5	64859597	64859697	CEX-chr5-64859598-64859697	0	.	64859597	64859697	0	1100,	0,
5	64863337	64863444	CEX-chr5-64863338-64863444	0	.	64863337	64863444	0	1107,	0,

Obr. 33: Ukázka souboru BED, vlastní data

Tab. 10: Povinná pole formátu BED

	Pole	Popis
1	chrom	název chromosomu nebo scaffoldu
2	chromStart	počáteční pozice; první báze v chromosomu je značena **0**
3	chromEnd	koncová pozice

Tab. 11: Volitelná pole formátu BED

	Pole	Popis
4	name	definuje jméno řádku v BED
5	score	skóre mezi 0 a 1000, v případě nastavení v hlavičce tracku na useScore=1 bude zobrazeno jako stupně šedé (čím tmavší, tím vyšší skóre)
6	strand	+ (forward), - (reverse)
7	thickStart	Počáteční pozice, ve které prvek začíná být vykreslen jako plný obdélník. Pokud není určen thickStart thickEnd, nastavují se obvykle do polohy chromStart
8	thickEnd	Koncová pozice, ve které je prvek vykreslen jako plný obdélník
9	itemRgb	hodnota ve tvaru R, G, B (0, 10, 0); pokud v tracku nastavení itemRgb=ON, pak bude tato hodnota určovat barvu zobrazení dat v daném BED řádku (doporučeno používat max. 8 barev na jeden soubor BED)
10	blockCount	počet bloků (exonů) na řádku
11	blockSizes	velikost bloků (exonů)
12	blockStarts	počáteční pozice každého bloku

V případě příliš velkého souboru BED je výhodnější ho transformovat do indexovaného binárního souboru bigBED. Umožňuje zpracování velkých datových souborů mnohem rychleji než u běžného souboru BED.

5.3.7 PDB

PDB – Protein Data Bank formát je druh textového souboru, který masově používal pro zápis sekvence terciální struktury proteinu a nukleových kyselin. Obsahoval informace o atomových souřadnicích, sekundární struktuře a atomové konektivity. Tento formát přestal být podporován jako hlavní formát z důvodu omezeného množství atomů, které v něm mohou být zapsány (nicméně vzhledem k tomu, jak dlouho se využíval, je stále k dispozici, a stále se hojně využívá). V současné době se využívá formát PDBx/mmCIF.

5.3.8 PDBx/MMCIF

PDBx/mmCIF (mmCIF = macro-molecular Crystallographic Information File), je formát zápisu sekvence terciální struktury proteinu a nukleových kyselin, který nahradil dříve používaný formát PDB. Nahrazení proběhlo kvůli větší flexibilnosti formátu: neukládá žádná omezení pro počet atomů, reziduí nebo řetězců, které mohou být zastoupeny v jedné položce; struktury obsahující >62 řetězců a/nebo 99999 ATOM záznamů nemohly být plně zapsán v PDB formátů. Klíčovou složkou tohoto formátu je „slovník“ povolených datových položek – sada názvů dat navržených k popisu makromolekulárního krystalografického experimentu a jeho výsledků. I nadále obsahuje hlavně informace o atomových souřadnicích, sekundární struktuře a spojení atomů.

5.3.9 Binární soubory

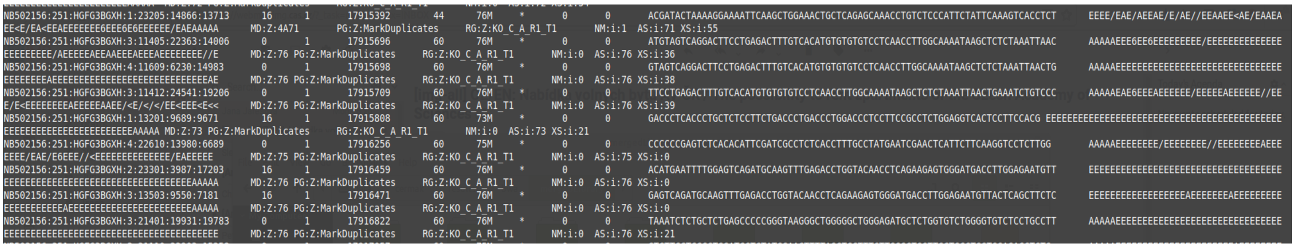
SRA – Sequence Read Archiv je komplexní soubor sekvenačních dat nové generace a zároveň formát k ukládání dat (tedy nejde přímo o nový typ dat). Archiv umožňuje přístup veřejnosti ke genomickým a transkripčním datům z různých projektů. SRA archiv podporuje principy FAIR. SRA formát dat je binární archivační soubor, který ukládá zarovnaná i nezarovnaná data ve formátu BAM, a dále raw data ve formátu FASTQ. Získání dat i jejich konverze do původních formátů je možná díky SRA toolkit.

Formátem, se kterým se setká prakticky každý bioinformatik, je soubor **BAM (**Binary Alignment Map), který je binárním zpracováním textového souboru typu **SAM (**Sequence Alignment Map**).**

5.4 Formát SAM/BAM

5.4.1 SAM/BAM

SAM je textový formát, který ukládá informace o zarovnání různých sekvencí namapovaných proti referenční sekvenci, obsahovat může i nenamalované sekevence. Jedná se o základní soubor, ze kterého je možné vyčíst množství informací, a použít ho na další analýzu dat (pro variant calling, výpočet jednotlivých vlastní, atd). Kromě sekvencí obsahuje informace popisující alignmentu, např. zda se jedná o více fragmentů, zda jsou všechny fragmenty správně namapovány, jestli existuje sekundární alignmentu a další. Skládá se z hlavičky (header) a ze sloupců oddělenými tabulátory (Obr. 34, Tab. 12). Řádky hlavičky začínají znakem @. Hlavička může obsahovat některé z předdefinovaných tagů jako @SQ – slovník pro referenční sekvenci, může obsahovat záznamy o jménu referenční sekvence (SN*), její délce (LN*) nebo třeba použitém organismu (SP). Důležitými jsou pole FLAG (Tab. 13) a pole CIGAR (Tab. 14).



Obr. 34: Zobrazení výstupu souboru BAM, vlastní data

Tab. 12: Sloupce formátu SAM/BAM

	Pole	Charakter	Popis
1	QNAME	String	Zadané jméno záznamu, záznamy se stejným QNAME se považují za pocházející ze stejného templátu.
2	FLAG	Int	bitwise FLAG
3	RNAME	String	Jméno referenční sekvence, pokud jsou k dispozici řádky záhlaví @SQ, musí být v jedné ze značek SQ-SN přítomno RNAME (pokud není RNAME označeno prázdným místem *)
4	POS	Int	1- based leftmost mapping POSition
5	MAPQ	Int	Kvalita mapování, $-10 \log_{10} p$, kde p je pravděpodobnost, že pozice je namapována špatně. Hodnota 255 znamená, že kvalita není k dispozici.
6	CIGAR	String	CIGAR záznam
7	RNEXT	String	Název referenční sekvence primárního zarovnání dalšího načteného readu v templátu.

	Pole	Charakter	Popis
8	PNEXT	Int	Poloha zarovnání dalšího načteného readu v templátu.
9	TLEN	Int	Pozorovaná délka templátu. Pokud jsou všechny segmenty mapovány na stejnou referenci, délka templátu se rovná počtu bází od namapované báze úplně vlevo k namapované bázi úplně vpravo. Levý segment má znaménko plus a pravý kraj má znaménko minus. Znaménko segmentů uprostřed není definováno. Je nastavena na 0 pro šablonu s jedním segmentem nebo když jsou informace nedostupné.
10	SEQ	String	segment sekvence, pokud neobsahuje prázdnou pozici (*), pak se její délka musí rovnat musí rovnat součtu délek operací M / I / S / = / X v CIGAR poli.
11	QUAL	String	ASCII záznam Phred sóre kvality +33

Tab. 13: Bitwise FLAG, součet bitových flagů umožňuje označení více atributů zarovnání čtení

BIT		Popis
1	0x1	templát má více segmentů v sekvenci
2	0x2	každý segment je správně zarovnám vzhledem k aligneru
4	0x4	nenamapovaný segment
8	0x8	další segment v templátu je nenamapovaný
16	0x10	SEQ je reverzně komplementární
32	0x20	SEQ dalšího segmentu v templátu je reverzně komplementární k prvnímu segmentu v templátu
64	0x40	první segment v templátu
128	0x80	poslední segment v templátu
256	0x100	druhý alignment
512	0x200	záznam neprošel filtrem – platform/vendor kontrola kvality
1024	0x400	PCR nebo optický duplikát
2048	0x800	doplňkový alignment

Tab. 14: Cigar, Concise Idiosyncratic Gapped Alignment Report – pole popisující alignment

OP	BAM	Popis
M	0	alignment match (ale možný mismatch v sekvenci)
I	1	inzerce oproti referenci
D	2	delece proti referenci
N	3	přeskočená oblast proti referenci

OP	BAM	Popis
S	4	soft clipping (clipped sekvence se vyskytují v SEQ)
H	5	hard clipping (clipped sekvence není zapsaná v SEQ)
P	6	padding (silent deletion z padded reference)
=	7	match sekvence
X	8	mismatch sekvence

Další informace o formátu SAM/BAM https://samtools.github.io/hts-specs/SAMv1.pdf.
BAM je komprimovaná a binární reprezentace souboru SAM, s možností indexování (soubor BAI). Oba formáty obsahují totožné informace.
Práce s formáty BAM a SAM, a jejich vzájemná konverze, je umožněna pomocí nástrojů SAMtools.

5.4.2 SAMTOOLS

SAMtools obsahují sadu nástrojů pro interakci a zpracování souborů SAM a BAM. Umožňují třídění, srovnávání, extrahování informací a indexování souborů. Pomocí SAMtools je možné zpracovat binární soubor BAM, aniž by byla nutná dekomprese celého souboru.
Základní příkazy v SAMtools:

view: filtruje data ve formátu SAM nebo BAM. Je možné vybrat pouhou část dat, se kterými chceme pracovat, místo všech dat.
sort: třídění souboru BAM podle koordinát případně jiného klíče (volba –n pro jméno). Vzhledem k náročnosti procesu je možné vytvářet více dočasných souborů a později je spojit dohromady (volba –m)
index: příkaz index vytvoří nový indexový soubor, který umožňuje rychlé vyhledávání dat v setříděném SAM/BAM souboru.
flagstats: spočítá počet zarovnání pro každý typ pole FLAG
stats: vytváří komplexní statistiky ze souboru zarovnání, včetně výpočtu vlastností

Další příkazy na tomto odkazu, https://bioinformatics.uconn.edu/resources-and-events/tutorials-2/file-formats-tutorial/#fileformats_sam.

5.5 Otázky k tématu

1. Jaký je rozdíl mezi FAIR a Open acces data
2. Popište soubory GFF a BED.
3. Jaké jsou výhody binárního zápisu souborů?
4. Jak se liší FASTA od FASTQ souboru?
5. Uveďte nějaký příklad a vysvětlení k Phred Score
6. Co vše lze vyčíst ze SAM/BAM souboru?
7. Uveďte příklad příkazu v SAMtools

5.6 Zdroje

FAIR data

Wilkinson MD, Dumontier M, Aalbersberg IJ, et al. The FAIR Guiding Principles for scientific data management and stewardship [published correction appears in Sci Data. 2019 Mar 19;6(1):6]. *Sci Data*. 2016;3:160018. Published 2016 Mar 15. doi:10.1038/sdata.2016.18

Datové typy a formáty

Frequently Asked Questions: Data File Formats. USCS Genome Browser. http://genome.ucsc.edu/FAQ/FAQformat.html (accessed Dec 09, 2020).
Supported file formats. Ensembl Home. http://www.ensembl.org/info/website/upload/index.html#formats (accessed Dec 09, 2020).
File Formats Tutorial This section explains some of the commonly used file formats in bi. Institute for Systems Genomics Computational Biology Core. https://bioinformatics.uconn.edu/resources-and-events/tutorials-2/file-formats-tutorial/ (accessed Dec 09, 2020).
File Format Guide. The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information Web site. https://www.ncbi.nlm.nih.gov/sra/docs/submitformats/ (accessed Dec 09, 2020).
Introduction to Protein Data Bank Format. Resource for Biocomputing, Visualization, and Informatics. https://www.cgl.ucsf.edu/chimera/docs/UsersGuide/tutorials/pdbintro.html (accessed Dec 09, 2020).
Protein Data Bank Contents Guide. wwPDB: Worldwide Protein Data Bank. http://www.wwpdb.org/documentation/file-format-content/format33/v3.3.html (accessed Dec 09, 2020).
Sequence read archiv. The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.https://trace.ncbi.nlm.nih.gov/Traces/sra/ (accessed Dec 09, 2020).

Formát SAM/BAM

Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-2079. doi:10.1093/bioinformatics/btp352
Samtools, 2019. Samtools. http://www.htslib.org/doc/samtools.html (accessed Dec 09, 2020).
File Formats Tutorial. Institute for Systems Genomics Computational Biology Core. https://bioinformatics.uconn.edu/resources-and-events/tutorials-2/file-formats-tutorial/ (accessed Dec 09, 2020).

6 DATABÁZE A ZDROJE DAT

Kromě generování dat z vlastních sekvenačních projektů, je možné získat data i z některého z veřejně dostupných úložišť. Existují stovky databází s uloženými daty jak z mikročipových, tak sekvenačních technologií.

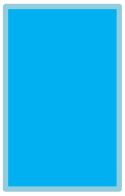
Některé databáze cílí na určitý typ záznamu: proteinové (InterPro, Swiss-Prot), struktur proteinů (PDB), RNA (Rfam), microRNA (MirBase), databáze genových expresí (GEO, ArrayExpress), signálních drah (Kegg pathways, WikiPathways, MSigDb), genových ontologií (Gene Ontology, Human Disease Ontology, Cell Ontology), metabolických drah (MetaCyc) a mnohé další.

Za centrální (primární) databáze by se daly považovat databáze NCBI, ENA (EMBL) a DDBJ, které spadají do dlouhodobé (rok 1987) iniciativy International Nucleotide Sequence Database Collaboration (INSDC), v rámci které si na denní bázi vyměňují nová/aktualizovaná sekvenační data („vzájemně se zrcadlí“), zároveň používají stejnou syntaxi pro ukládání a zpracování záznamů. Primární databáze obsahují experimentálně odvozená data, jako je nukleotidová sekvence, proteinová sekvence nebo makromolekulární struktura. Experimentální výsledky vkládají badatelé přímo do databáze a data mají v zásadě archivní charakter. Jakmile jim je přiděleno přístupové číslo k databázi, data v primárních databázích se nikdy nezmění: jsou součástí vědeckého záznamu.

Sekundární databáze obsahují data odvozená z výsledků analýzy primárních dat, často čerpají informace z mnoha zdrojů, včetně jiných databází (primárních a sekundárních), kontrolovaných slovníků a vědecké literatury. Jsou různě kurátorované, mohou využívat výpočetních algoritmů a manuální analýzy a interpretace k získání nových informací.

Tab. 15: Rozdělení podle https://www.ebi.ac.uk/training/online/courses/bioinformatics-terrified/what-makes-a-good-bioinformatics-database/primary-and-secondary-databases/ (accessed Aug 2 2021)

	Primární databáze	Sekundární databáze
Synonyma	Archivní	Kurátorované, Databáze znalostí
Zdroj dat	Přímé nahrání experimentálně odvozených dat	Výsledky analýzy, literárního výzkumu a interpretace (často dat v primárních databázích)
Příklady	ENA, GenBank a DDBJ (nukleotidové sekvence) ArrayExpress a GEO (funkční data) Protein Data Bank (PDB; koordináty 3D makromolekulárních struktur)	InterPro (proteinové rodiny, motivy) UniProt Knowledgebase (sekvence a fukční informace o proteinech) Ensembl (variace, funkce, regulace, funkční informace celého genomu)



Pokusy o dělení jsou nepřesné (a někdy možná i matoucí) i z toho důvodu, že v současné době už databáze jako uložiště čistě experimentálních dat prakticky neexistují – všechny bývají rozšiřovány o nástroje pro analýzu.

6.1 Často používané databáze

6.1.1 NCBI

NCBI – Národní centrum pro biotechnologické informace, vytváří databázi ze sekvencí předkládaných jednotlivými laboratořemi a ze sekvencí získaných výměnou dat s mezinárodními databázemi nukleotidových sekvencí, Evropskou laboratoří molekulární biologie (EMBL) a DNA databází Japonska (DDBJ). Mezi hlavní databáze NCBI patří GeneBank.

NCBI web portal, https://www.ncbi.nlm.nih.gov/search/, (dříve pod ENTREZ) je rozhraní pro prohledávání databází molekulární biologie, který poskytuje integrovaný přístup k údajům o sekvenci nukleotidů a proteinů, informacím o genovém a genomovém mapování, údajům o 3D struktuře, PubMed MEDLINE a dalším. Vyhledávání pokrývá více než 20 databází včetně kompletních dat o proteinových sekvencích z PIR-International, PRF, Swiss-Prot a PDB a dat o nukleotidových sekvencích z GenBank, které obsahují informace z EMBL a DDBJ.

Vyhledávací systém používá intuitivní uživatelské rozhraní pro rychlé vyhledávání sekvence a bibliografických dat. Jedinečnou vlastností systému je jeho použití předpočítaných hledání podobnosti pro každý záznam k vytváření odkazů na sousedy nebo související záznamy v jiných databázích. Tato propojení usnadňují integrovaný přístup v různých databázích. Výsledky lze prohlížet v různých formátech, včetně FlatFile, FASTA, XML a dalších. Grafické rozhraní umožňuje snadnou vizualizaci úplných genomů nebo chromosomů, jakož i biologickou anotaci jednotlivých sekvencí. Systém také umožňuje hromadné stahování velkých výsledků vyhledávání.

GeneBank

GeneBank, https://www.ncbi.nlm.nih.gov/genbank/ je veřejně přístupná anotovaná sekvenční nukleotidová databáze s nukleotidovými sekvencemi z více než 300 000 druhů organismů. Data do GeneBank jsou získaná primárně prostřednictvím podání od jednotlivých laboratoří a hromadných podání od velkých sekvenčních projektů, včetně Whole Genome Shotgun (WGS) a projekty vzorkování životního prostředí. Přístup a vyhledávaní na GenBank je zajištěno přes NCBI Entrez, který zároveň integruje vyhledávání v hlavních DNA a proteinových strukturních a sekvenčních databázích a v hlavní databázi odborných biomedicínsky zaměřených časopisů PubMed. BLAST poskytuje vyhledávání sekvenční podobnosti GenBank a dalších sekvenčních databází. Kromě přístupu přes webové rozhraní lze také prostřednictvím FTP celou databázi k danému datu bezplatně nainstalovat na konkrétní počítač. V tomto případě je však nutno ji pravidelně aktualizovat – NCBI uvolňuje novou verzi každé 2 měsíce.

Základní typy datových záznamů

- standardní originální nukleotidové sekvence: sekvence získané sekvenováním fragmentů genomové DNA
- sekvence EST (expressed sequence tags): neúplné sekvence konců jinak necharakterizovaných cDNA, data obvykle nižší kvality než standardní sekvence
- sekvence HTGS (high throughput genome sequencing): dosud neposkládané a neanotované sekvence pocházející ze sekvenování genomů
- sekvence WGS (whole-genome shotgun): referenční sekvence již většinou poskládaných a anotovaných kompletních genomů
- sekvence TPA (third party annotation): sekvence anotované jinými než původními autory
- sekvence TSA (transcriptome shotgun assembly sequence): sekvence transkriptomů získané reverzním přepisem z mRNA do cDNA, jedna z nejrychleji narůstajících oblastí dat
- sekvence ENV (Environmental sample sequence): environmentální DNA získaná sekvenováním celých společenstev často nepopsaných organismů, např. metagenomická data získaná z biofilmů, sedimentů, horkých pramenů, povrchu tkání apod.

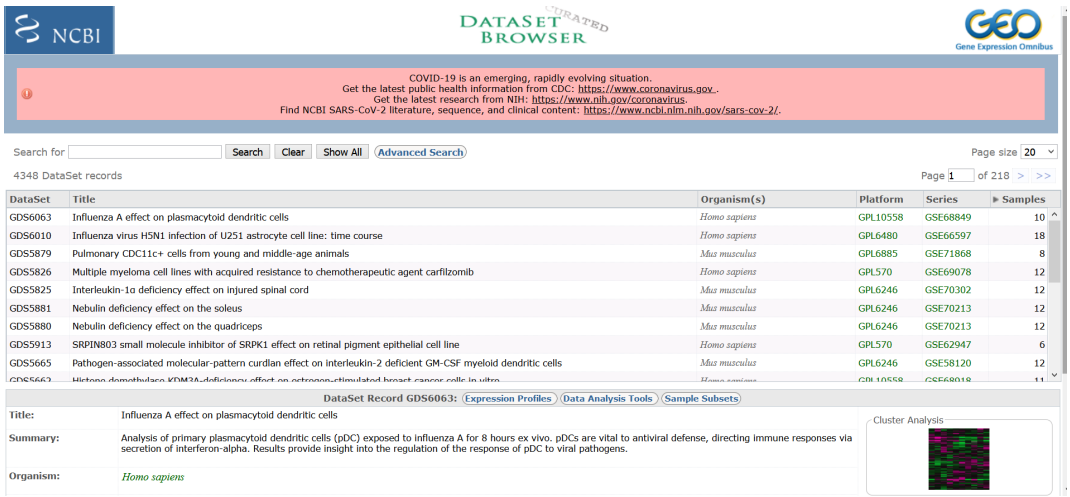
dbSNP

dbSNP, https://www.ncbi.nlm.nih.gov/snp/, je největší světová databáze pro varianty nukleotidů. K tomuto datu se dbSNP skládá z velkého shluku druhově specifických databází, které obsahují více než 12 milionů neredundantních variací sekvencí (polymorfismy s jedním nukleotidem, inzerce či delece a krátké tandemové repetice) a více než 1 miliarda jednotlivých genotypů z HapMap a další rozsáhlé genotypizační aktivity – více než 200 GB dat a každý den roste.

Kromě těchto zpracovává NCBI celou řadu databází pro medicínské a vědecké účely. Toto zahrnuje databázi Online Mendelian Inheritance in Man (**OMIM**), Databázi molekulárních modelování (**MMDB**) 3D proteinových struktur, Genovou mapu lidského genomu, Prohlížeč taxonomie a Cancer Genome Anatomy Project (**CGAP**), na kterém pracuje ve spolupráci s Národním onkologickým institutem.

GEO Datasets

Gene Expression Omnibus, https://www.ncbi.nlm.nih.gov/sites/GDSbrowser/, (GEO, Obr. 35) je veřejně dostupné úložiště, které archivuje a volně distribuuje komplexní sady funkčních genomických dat podaných vědeckou komunitou. Kromě uložení dat je k dispozici také sada webových rozhraní a aplikací, které uživatelům pomohou vyhledávat a stahovat studie a vzorce genové exprese uložené v GEO, a následně je i analyzovat pomocí nástroje GEO2R. Typy dat uložené v databázi mohou pocházet z čipových experimentů zpracovaných technologiemi Affymetrix, Agilent, Nimblegen a Illumina, dále z RT-PCR, High-throughput sekvenování a SAGE (data ze Sangerova sekvenování). Každý GEO záznam obsahuje informaci o organismu, typu experimentu, technickém popisu designu a souhrnu experimentu, jména těch, kteří se podíleli na zpracování, citaci a odkaz na článek, odkaz na platformu, vzorky, analyzační program GEO2R a odkaz na uložená data v BioProject a na stáhnutí raw dat, případně předzpracovaných tabulek a metadat.



Obr. 35: GEO DataSet web page, GEO DataSet Browser. National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/sites/GDSbrowser/ (accessed Dec 29, 2020)

BioProject

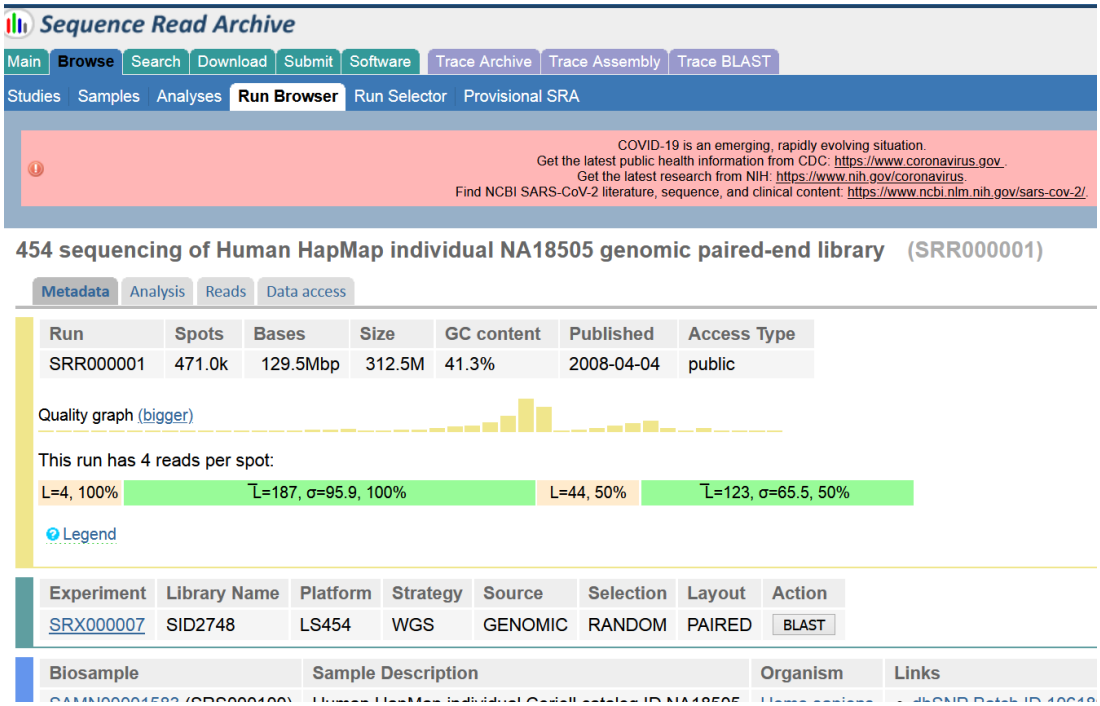
BioProject, https://www.ncbi.nlm.nih.gov/bioproject/browse, shromažďuje kolekce biologických dat z jednotlivých organizací i konsorcií. Typicky záznam v BioProjectu obsahuje informace o typu dat, rozsahu, použitém organismu, odkazu na publikace vztažené k projektu, odkazy na příbuzné databázové zdroje, a hlavně odkazy na data (pokud jsou k dispozici) ve formě SRA experimentu, odkazu na sekvenci zdrojového organismu a na další vztažené datasety (BioSample a Assembly). Registrace v BioProjectu je nutnou součástí vkládání dat do primárních archivů NCBI, tj. do SRA, TSA a WGS (v současné době to není nutnou podmínkou pro vkládání dat do GEO).

Assembly

Assembly, https://www.ncbi.nlm.nih.gov/assembly/organism/2759/all/, je databáze poskytující informace o struktuře assemblovaných genomů, názvech assembly a dalších meta-datech, statistické reporty a odkazy na data genomových sekvencí.

SRA, TSA A WGW (NCBI)

SRA – Sequence Read Archive, https://www.ncbi.nlm.nih.gov/sra (Obr. 36), archiv sekvenčních raw dat z technologií sekvenování nové generace, včetně Illumina, 454, IonTorrent, Complete Genomics, PacBio a OxfordNanopores.



Obr. 36: Ukázka Sequence Read Archive web page Sequence Read Archive Run Browser. National Center for Biotechnology Information. https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=SRR000001 (accessed Dec 29, 2020)

TSA – Transcriptome Shotgun Assembly, https://www.ncbi.nlm.nih.gov/genbank/tsa/, archiv výpočetně sestavených sekvencí transkriptů z primárních dat, jako jsou EST a technologie sekvenování nové generace, místo použití tradičních metod klonování a sekvenování cDNA.

WGS – Whole Genome Shotgun, https://www.ncbi.nlm.nih.gov/genbank/wgs/, archiv genomových assembly neúplných genomů nebo neúplných chromosomů prokaryot nebo eukaryot, které jsou generovány sekvenováním shotgun whole genome strategií. Projekty WGS mohou obsahovat anotace, ale nemusí. Pro prokaryotní organismy má NCBI vytvořenou pipelinu na složení genomu, kdy při nahrání genomu je možné tuto pipelinu spustit, a před finálním odesláním upravit a zkontrolovat.

NCBI datasets
NCBI datasets, https://www.ncbi.nlm.nih.gov/datasets/, je experimentální platforma pro vyhledávání a vytváření datasetů. Zároveň tento projekt usiluje o dodržování principů FAIR. Momentálně umožňuje uživatelům získat sekvence genomů a jejich anotace podle taxonomického jména, ID taxonomie nebo assembly přístupového čísla.

6.1.2 EMBL-EBI

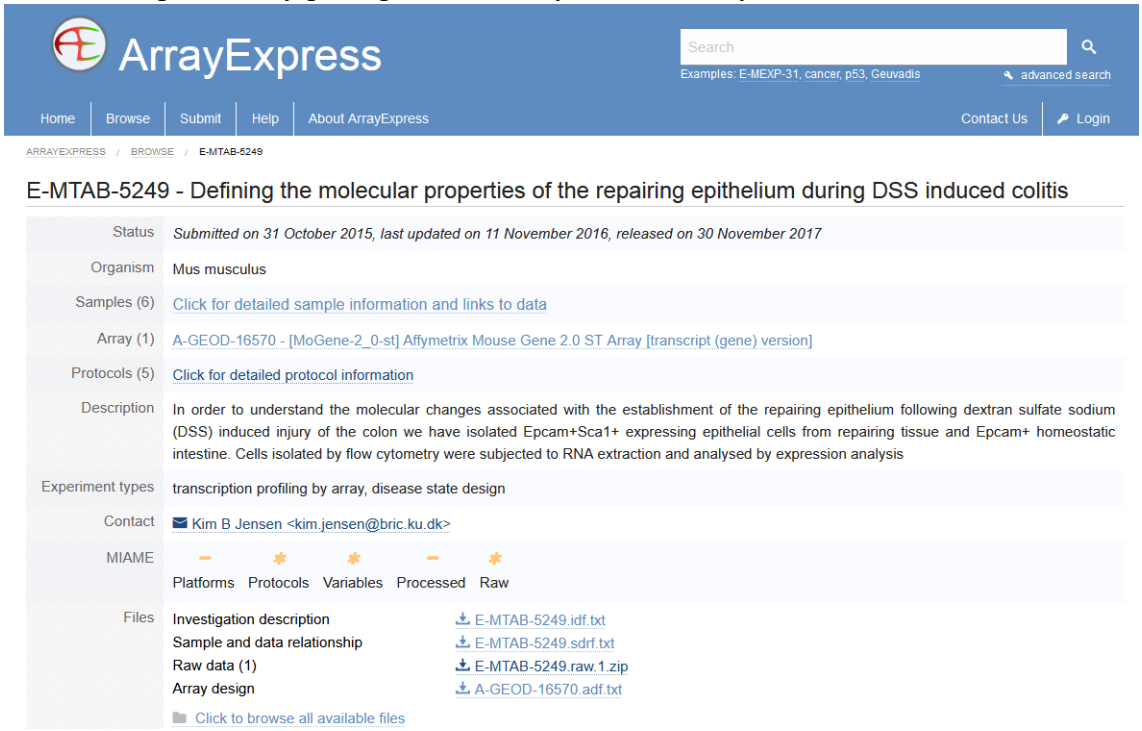
Databáze Evropské laboratoře molekulární biologie (EMBL) je komplexní soubor primárních dat uchovávaných v Evropském bioinformatickém institutu (EBI), https://www.ebi.ac.uk. Data jsou přijímána z center pro sekvenování genomu, od jednotlivých vědců a patentových úřadů. Databáze EMBL jsou uloženy a udržovány v systémech správy dat Oracle, SQL server a Postgres, a lze je prohledávat na internetu pomocí vyhledávacího systému Sequence Retrieval System (SRS), což je vyhledávač EBI pro databáze molekulární biologie. EMBL-EBI je jedním z největších úložišť dat na světě, mimo jiné mají pod svou správou databáze:

ENA
Databáze záznamů a platforma pro správu dat: ENA, https://www.ebi.ac.uk/ena/browser/search, zahrnuje jak globálně komplexní datový zdroj, který zachovává světový veřejný výstup sekvenčních dat, tak bohaté portfolio nástrojů a služeb na podporu správy sekvenčních dat.

BioSamples
BioSamples, https://www.ebi.ac.uk/biosamples/samples, popisující biologické vzorky a poskytující odkazy na související experimentální data.

ArrayExpress

ArrayExpress, https://www.ebi.ac.uk/arrayexpress/browse.html, (Obr. 37) je repositář funkčně genomických dat z mikročipových a dalších sekvenačních platforem. Část experimentálních dat je importována z NCBI Gene Expression Omnibus databáze. Data uložená v ArrayExpress obsahují informace o použitém organismu, vzorcích, použité čipové technologii a protokolu, popis experimentu a kontakty na účastníky experimentu. Dále mohou obsahovat odkazy na raw data, zprocesovaná data, design čipu a odkaz na detailní popis vztahů mezi daty (pokud jsou data k dispozici). Pro ArrayExpress existuje balíček v R pro snadný přístup k datům a vytváření datových struktur.



Obr. 37: Ukázka repozitáře ArrayExpress pro projekt E-MTAB-5249, Array Expres. The European Bioinformatics Institute. https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-5249/ (accessed Dec 29, 2020)

6.1.3 DDBJ

Japonská DDBJ Center, https://www.ddbj.nig.ac.jp/index-e.html, mezinárodně přispívá jako člen INSDC ke sběru a poskytování nukleotidových sekvenčních dat s ENA / EBI v Evropě a NCBI v USA.
DDBJ Center je oficiálně certifikováno pro shromažďování nukleotidových sekvencí od výzkumných pracovníků a pro vydávání mezinárodně uznávaného přístupového čísla zadavatelům údajů. Přístupové číslo vydané pro každá sekvenční data je v databázi jedinečné a mezinárodně uznávané, aby zaručilo zadavateli vlastnictví předložených a publikovaných údajů. Vzhledem k tomu, že DDBJ Center si denně vyměňuje uvolněná data s ENA / EBI a NCBI, sdílejí tato tři datová centra v daném okamžiku prakticky stejná data. Prakticky sjednocená databáze se nazývá INSD; Mezinárodní databáze nukleotidových sekvencí.
DDBJ shromažďuje sekvenční data hlavně od japonských vědců, ale samozřejmě přijímá data a vydává přístupová čísla výzkumníkům v jiných zemích. 99 % údajů INSD od japonských vědců se předkládá prostřednictvím DDBJ.

INSD obsahuje data nukleotidové sekvence související s patentovými přihláškami shromážděnými patentovými úřady v Japonsku, Koreji, Evropě a USA. DDBJ Center také poskytuje data aminokyselinových sekvencí souvisejících s patentovými přihláškami shromážděnými patentovými úřady v Japonsku a Koreji.

6.2 Další databázové zdroje dat

Odkazy na další zdroje a databáze pod EMBL-EBI
Tools & Data Resources. The European Bioinformatics Institute. https://www.ebi.ac.uk/services/all (accessed Dec 09, 2020).
Odkazy na další zdroje a databáze pod NCBI
All Resources - Site guide. The National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/guide/all/ (accessed Dec 09, 2020).
Odkazy na zdroje a databáze pod správou organizace Elixir
ELIXIR Deposition Databases for Biomolecular Data. ELIXIR. https://elixir-europe.org/platforms/data/elixir-deposition-databases (accessed Dec 09, 2020).

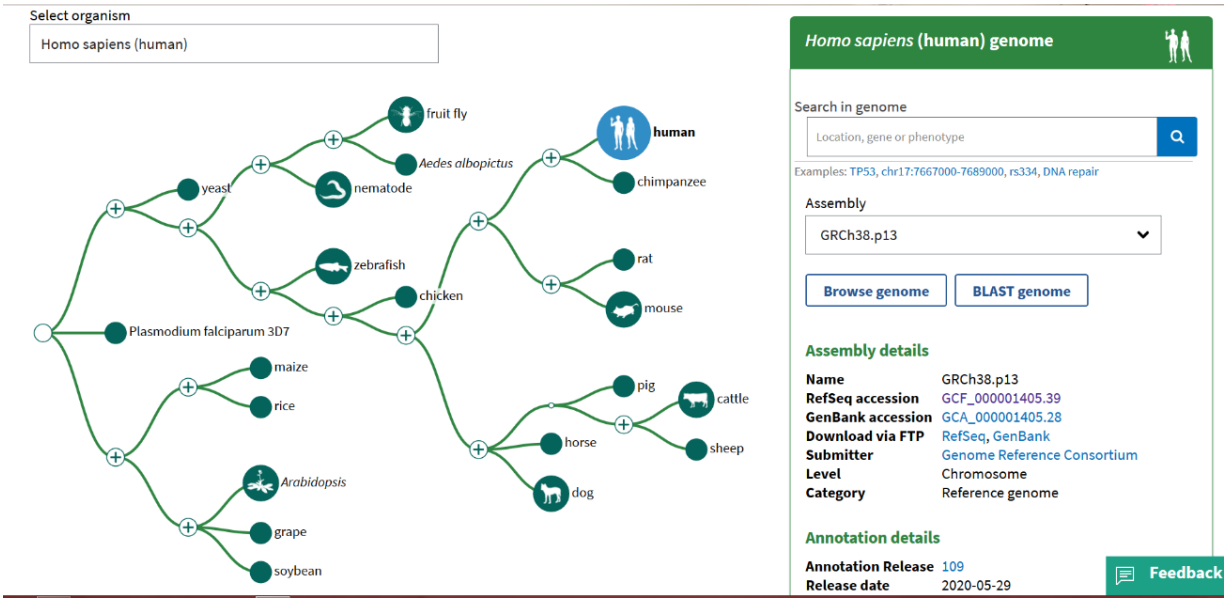
Další zajímavou a vzhledem k širokému obsahu informací nesmírně užitečnou databází je UniProt (se manuálně spravovanou databází SwissProt a automaticky spravovanouTrEmbl), UniProt. https://www.uniprot.org (accessed March 04, 2022).

6.3 Genomové prohlížeče

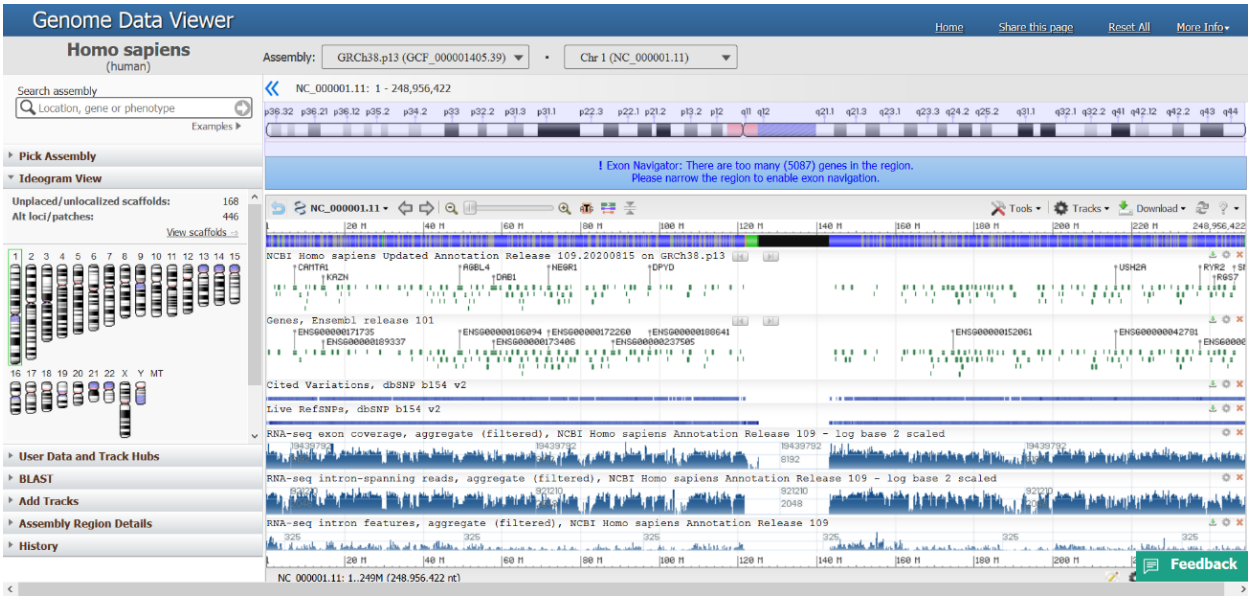
Genomové prohlížeče jsou grafická rozhraní, která slouží k prohlížení informací z genomových databází. Umožňují procházet celé chromosomy i genomy s anotovanými údaji včetně predikce struktury genů, proteinů, exprese, zobrazení regulačních míst, variant, fylogenetického kontextu, srovnávacích analýz a dalších. Mezi nejznámější veřejné genomické prohlížeče patří USCS Genome Browser, NCBI's Genome Browser a Ensembl Genome Browser.

6.3.1 NCBI

NCBI Genome Data Viewer, https://www.ncbi.nlm.nih.gov/genome/gdv/, (GDV, Obr. 38 a Obr. 39) je prohlížeč genomu podporující průzkum a analýzu eukaryotických RefSeq genomů. Uživatelé mohou pomocí GDV vizualizovat různé typy dat souvisejících se sekvencí v kontextu genomu. Genome Data Browser se také používá v různých prostředcích NCBI, jako jsou GEO a dbGaP, k zobrazení datových sad souvisejících se specifickými experimenty nebo vzorky v kontextu prohlížeče genomu.



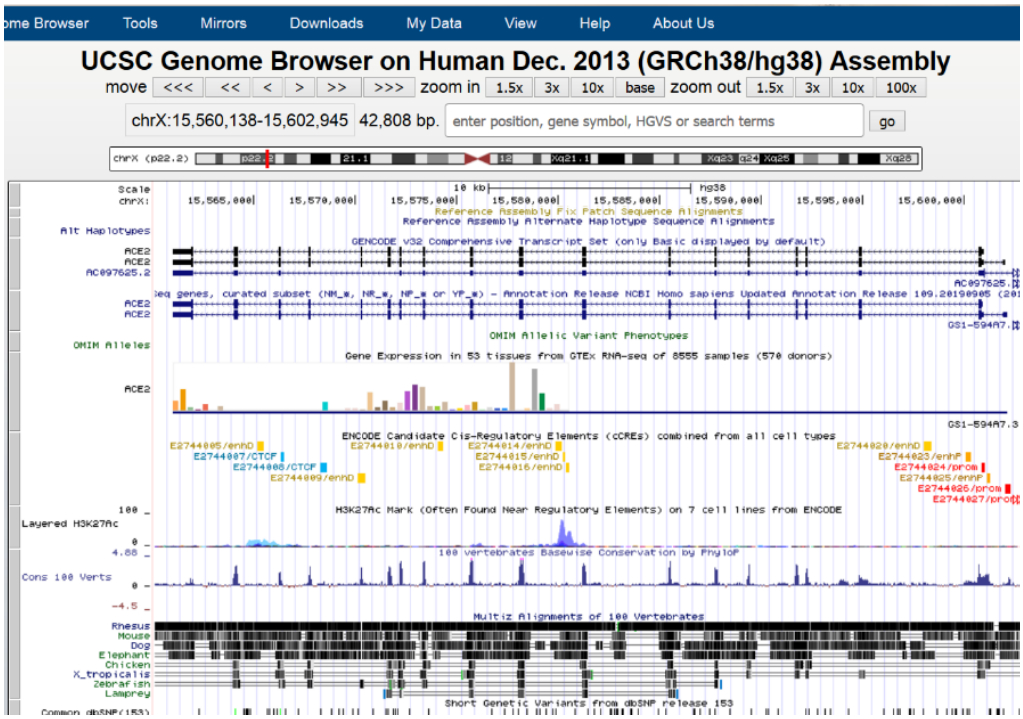
Obr. 38: Screenshot GDV web page, NCBI´s genome browser for human. National Institutes of Health. https://www.ncbi.nlm.nih.gov/genome/gdv (accessed Dec 29, 2020)



Obr. 39: Screenshot web page, Chr1: 1-249.0M. National Institutes of Health. https://www.ncbi.nlm.nih.gov/genome/gdv/browser/genome/?id=GCF_000001405.39 (accessed Dec 29, 2020)

6.3.2 UCSC

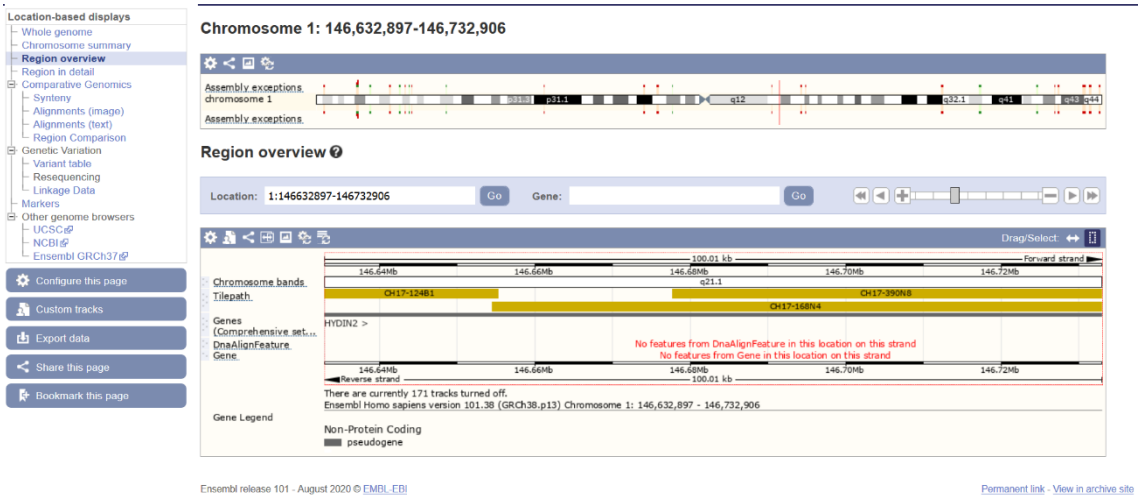
UCSC, https://genome.ucsc.edu/cgi-bin/hgGateway, (Obr. 40) obsahuje širokou sbírku assembly a anotací obratlovců a modelových organismů, spolu s velkou sadou nástrojů pro prohlížení, analýzu a stahování dat. Zobrazuje informace z mapování a sekvencí včetně alternativních haplotypů, geny a genové predikce, fenotypy, mRNA a EST, exprese, regulace, komparativní genomiku, variace, repetice, a dokonce assembly denisovanů a neandrtálců. V UCSC byly ukázány výsledky z HUGO projektu.



Obr. 40: Screenshot UCSC Genome Browser web page, UCSC Genome Browser on Human Dec. 2013 (GRCh38/hg38) Assembly. USCS Genome Browser. https://genome.ucsc.edu/cgi-bin/hgTracks?db=hg38&lastVirtModeType=default&lastVirtModeExtraState=&virtModeType=default&virtMode=0&nonVirtPosition=&position=chrX%3A15560138%2D15602945&hgsid=991914879_5J03eST6a4laZYPeAe0R0V550vNY (accessed Dec 29, 2020)

6.3.3 Ensembl

Ensembl, https://genome.ucsc.edu/cgi-bin/hgGateway, je genomový prohlížeč (Obr. 41) genomů obratlovců, který podporuje výzkum komparativní genomiky, evoluce, sekvenčních variací a transkripční regulace. Sestavuje anotované geny, počítá vícenásobné zarovnání, předpovídá regulační funkce a shromažďuje údaje o nemoci. Ensembl tools zahrnují BLAST, BLAT, BioMart a Variant Effect Predictor (VEP) pro všechny podporované druhy



Obr. 41: Screenshot Ensembl web page, Chromosome 1: 146,632,897-146,732,906. Ensembl genome browser 102. https://www.ensembl.org/Homo_sapiens/Location/Overview?db=core;r=1:146632897-146732906 (accessed Dec 29, 2020)

6.4 Otázky k tématu

1. Jaké znáte hlavní databáze?
2. Co je primární databáze?
3. Jak můžeme dělit databáze?
4. Která databáze ukládá mikročipová data?
5. Co obsahuje záznam GEO?

6.5 Zdroje

Hlavní a často používané databáze

ArrayExpress – functional genomics data. The European Bioinformatics Institute. https://www.ebi.ac.uk/arrayexpress/ (accessed Dec 09, 2020).
Assembly. The National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/assembly/ (accessed Dec 09, 2020).
Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res. 2010;38(Database issue):D46-D51. doi:10.1093/nar/gkp1024
Bioinformation and DDBJ Center. https://www.ddbj.nig.ac.jp/index-e.html (accessed Dec 09, 2020).
BioProject. The National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/bioproject/ (accessed Dec 09, 2020).
BioSamples. The European Bioinformatics Institute. https://www.ebi.ac.uk/biosamples/ (accessed Dec 09, 2020).
Datasets. The National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/datasets/ (accessed Dec 09, 2020).
GEO DataSets. The National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/gds/ (accessed Dec 09, 2020).
Run Browser: Sequence Read Archive. The National Center for Biotechnology Information. https://trace.ncbi.nlm.nih.gov/Traces/sra/sra.cgi?view=run_browser (accessed Dec 09, 2020).
Sequence Set Browser. The National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/genbank/tsa/ (accessed Dec 09, 2021).
Sequence Set Browser. The National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/genbank/wgs/ (accessed Dec 09, 2021).
The National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/ (accessed Dec 09, 2020).
Tools & Data Resources. The European Bioinformatics Institute. https://www.ebi.ac.uk/services (accessed Dec 09, 2020).

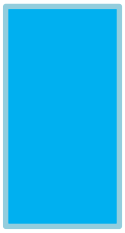
Genomové prohlížeče

Genome data viewer. National Institutes of Health. https://www.ncbi.nlm.nih.gov/genome/gdv/ (accessed Dec 09, 2020).
Genome Browser Gateway. USCS Genome Browser. https://genome.ucsc.edu/cgi-bin/hgGateway (accessed Dec 09, 2020).
Ensembl genome browser. Ensembl genome browser. https://www.ensembl.org/index.html (accessed Dec 09, 2020).

7 KVALITA DAT A VIZUALIZACE

7.1 Odstranění nežádoucích sekvencí

Během přípravy sekvenačních knihoven se na konce všech readů připojují adaptorové sekvence. Ty jsou vyžadovány pro nasedání primerů, pro správné uchycení na flowcellu a pro připojení indexových sekvencí, které umožňují identifikaci vzorku. Jedním ze způsobů zlepšení výsledné kvality raw dat, a posléze i správnosti zarovnání, je trimování (ořezání) těchto adapterových sekvencí, koncových bazí readů, případně dalších sekvencí a bazí např. podle kvality. Adaptorové sekvence by měly být odebrány z readů i vzhledem k možnému ovlivnění zarovnání sekvencí, zvláště pak u de novo assembly.



UMI (unique molecule identifiers, někdy označované jako molecular barcodes) jsou „jedinečné“ artificiální krátké DNA sekvence (indexy), které se využívají pro korekci sekvenačních chyb, zvýšení přesnosti sekvenování a odstranění PCR biasu. Standartně se jich využívá u sekvenování jednotlivých buněk, ale třeba také u sekvenování malých molekul.

UMI samotné ještě nemusejí být dostatečně silné pro identifikaci pre-PCR molekul. Sekvenování může obsahat kolem 30-50 milionů readů, a UMI mají 12 bazí, tj. 4^12 možností, nehledě na to, že i UMI mohou podléhat sekvenačním chybám. Důležité je kolik různých molekul je ve vzorku – single-cell RNA-seq může obsahovat milion biomolekul. Deduplikace readů od technických replikátů proto typicky probíhá až po alignmentu, kdy se kromě sekvence UMI rozhoduje i podle koordinát mapování.

7.1.1 Nástroje určené k trimování sekvencí

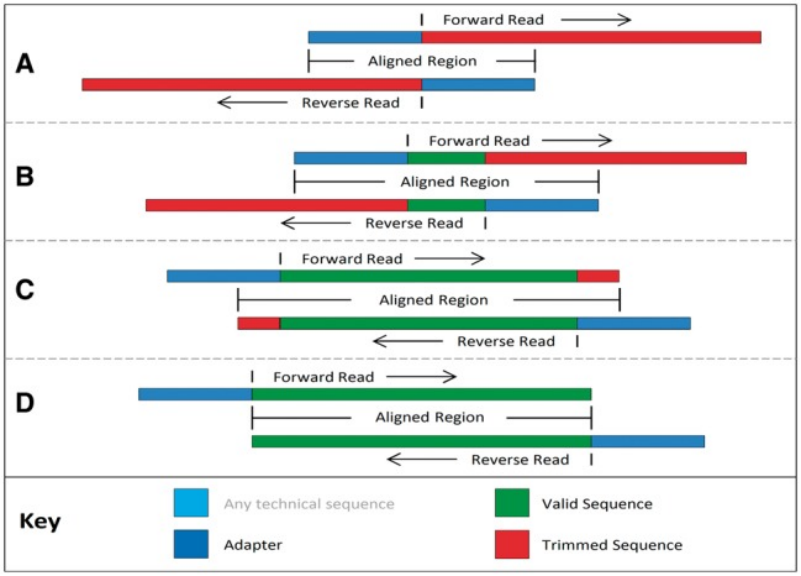
Cutadapt a TrimGalore!

Cutadapt je samostatný nástroj napsaný v Pythonu, ale pro zvýšení rychlosti je algoritmus alignmentu implementován v C jako rozšiřující modul Pythonu. Vyhledává a filtruje adaptorové sekvence, sekvence primerů, poly-A konce a další nežádoucí sekvence z 454 a Illumina dat, ale také z color-space SOLiD readů, a to jak u single-end dat, tak i pair-end dat. Zvláště užitečný je tento nástroj pro trimmování small-RNA seq single readů. Další vlastností je možnost demultiplexování single-end readů (pro pair-end není tato možnost k dispozici), kdy ready jsou zapisovány do různých souborů podle adaptorů, které v nich byly nalezeny. Pro zrychlení práce lze použít Cutadapt simultánně s BWA alignerem. Pomocí pipy je umožněno obejít ukládání souborů, Cutadapt output je rovnou použit jako input pro BWA.

TrimGalore! je tzv. wrapper script kolem nástrojů Cutadapt a FastQC, který automatizuje odstranění adaptorových sekvencí, stejně jako sekvencí a bazí s nízkou kvalitou, a kontrolu kvality výsledného trimování. Zároveň obsahuje přidané funkce umožňující odstranění biasovaných (jak nadhodnocených, tak podhodnocených) methylovaných pozic u bisulfidového sekvenování se sníženým zastoupením.

Trimmomatic

Trimmomatic je nástroj pro zpracování single-end i pair-end Illumina dat, založený na programovacím jazyku Java. Trimmomatic umožňuje řadu různých trimování, jako např. kontrolu readů od 5' konce, dále v okamžiku, kdy kvalita čtení klesne pod stanovenou úroveň, daný read je v daném místě oříznut, oříznutí readu na stanovenou délku odstraněním bazí z konce sekvence nebo odstranění celého readu, pokud průměrná hodnota validity je pod stanovenou mezí, a další. Trimmomatic také umožňuje **palindromní** trimovací strategii pro pair-endová data (Obr. 42) v případě tzv. read-through kdy sekvenovaný insert je kratší než délka čtení. V tomto případě obě čtení musí obsahovat stejný počet platných bazí a kontaminující sekvenci z protilehlého adapteru. Platná sekvence v každém párovém čtení bude reverzním komplementem. Výsledné sekvence jsou zarovnány za použití globálního alignmentu. Zarovnání se skóre větším než stanovený threshold určuje, že první části každého čtení jsou vzájemně reverzními komplementy, zbývající části čtení odpovídají jejich adaptérům, které jsou odebrány. Takto jsou detekovány technické kontaminace ve čteních.

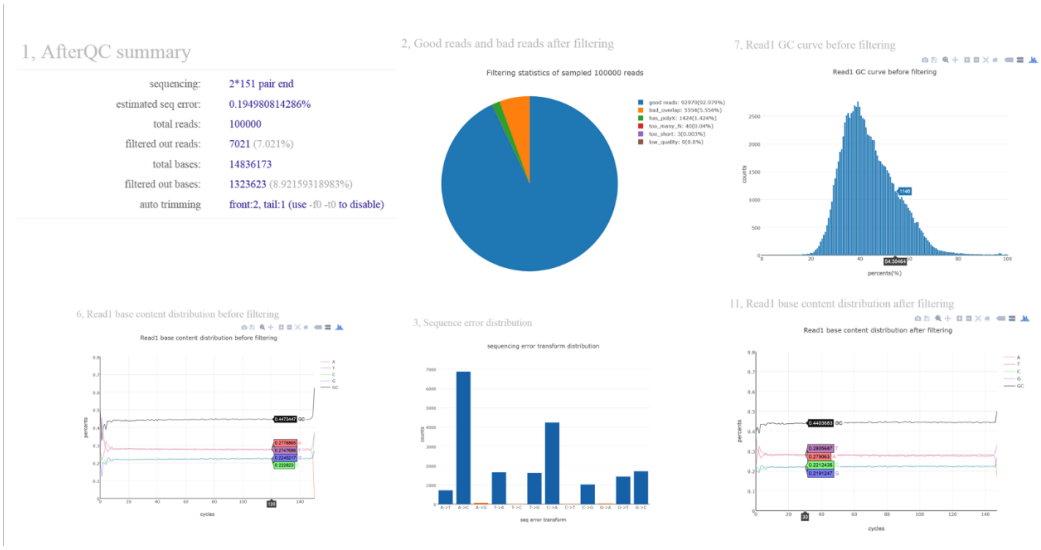


Obr. 42: Předpokládané zarovnání sekvencí palindromním režimu. Proces zarovnání může vést k následujícím možnostem: obsahují pouze adaptorové sekvence (A), obsahují sekvenci insertu a adaptoru (B), obsahují sekvenci insertu a částečně adaptoru (C) anebo neobsahují žádný adaptor (D) (Bolger et al, 2014)

Fastp a AfterQC

Fastp je pythonovský nástroj implementovaný v prostředí C++ určený k předzpracování single i pair-endových dat ve formátu FASTQ spolu s vizualizací jejich kvality. Umožňuje celou řadu operací s daty, jako filtrování readů s nízkou kvalitou nebo velkým obsahem N bazí, odstranění adaptorových sekvencí, identifikace a opravy chybných bazí za předpokladu, že jedna báze (pravděpodobně správná) má vysokou a ostatní báze naopak velmi nízkou kvalitu. Zároveň dokáže zpracovat data s UMI identifikátory, kdy sekvence UMI je vložena do hlavičky readu. Výsledná zpráva o kvalitě sekvencí je vizualizována pomocí html stránky (ukázka reportů Obr. 43), podobně jako u nástroje FastQC, nebo pomocí JSON souboru. Ukázkový report je dostupný na opengene.org/AfterQC/report.html.

AfterQC je starší verzí fastp, která není implementována do prostředí C++.



Obr. 43: Ukázka výstupu AfterQC, http://opengene.org/AfterQC/report.html (accessed Dec 27, 2020)

7.1.2 Nástroje pro odstranění dalších nežádoucích sekvencích

Kromě adaptorových a jiných technických reliktů se v sekvenačních knihovnách mohou objevit další nežádoucí sekvence, vzniklé jak např. špatným odběrem materiálu, kontaminací v laboratoři, tak i biologickými vlastnosti vzorku. V případě předpokládatelného kontaminantu (pracujeme např. se zvířecí DNA, a máme obavy z cross-kontaminace od pracovníka v laboratoři) je možné namapovat ready na známý kontaminant, a nealignované ready zpětně namapovat na požadovaný organismus. Také je takto možné stanovit příbuzné sekvence, kdy ta data, která se namapovala na kontaminant, zpětně namapujeme na požadovaný organismus, a podle míry úspěšného mapování určíme nakolik byl vzorek kontaminován).

SortMeRNA

Kromě tohoto způsobu můžeme využít i nástroj SortMeRNA, pythonovský nástroj kompilovaný v C++, určený pro Linux a Windows. SortMeRNA je nástroj primárně určený na filtrování ribosomální RNA z dat ve formátu FASTA, FASTQ a jejich komprimovaných variant, za použití referenčních databází SILVA a RFAM. Kromě tohoto jej lze použít i na odstranění jiných nežádoucích sekvencích uložených ve formátu FASTA.

7.1.3 Kontrola duplicit

U NGS metod jsou fragmentované kusy (c)DNA obvykle amplifikovány pomocí PCR reakce. Amplifikační proces ale není unikátní a výsledné knihovny pak mohou být zatíženy tzv. amplifikačním (PCR) biasem. To znamená, že jsou pomocí PCR vytvořeny technické replikáty insertu, a nikoli biologické replikáty (různé mRNA/DNA molekuly) ze vzorku. Jedním z možných řešení pro odstranění PCR biasu je ignorování readů začínajících stejnou bazí a majících stejnou délku po dosažení určitého počtu, což ovšem není úplně vhodný přístup v případě cDNA molekul, které se v buňce vyskytují v mnoha kopiích. Další z možností je využití tzv. UMI – unikátních molekulových identifikátorů. Dalším z problému pak může být amplifikování fragmentů bez biologické hodnoty, typicky PCR primerů a adaptorových sekvencí bez navázaných fragmentů, což se řeší typicky trimováním sekvencí a kontrolou kvality readů. Rozdíly v prosekvenovanosti knihoven, délek genů a GC obsahu se typicky řeší normalizací.

Deduplikace readů pomocí UMIs

Prvním krokem v deduplikaci je odstranění sekvence z readů a její vložení do headru. To lze udělat pomocí nástrojů fastp, UMI-tools i bcl2fastq. Dalším krokem je odstranění duplikovaných readů ze souboru. Best practise v tomto případě je použití nástrojů z balíčku Alevin v případě single cell dat a UMI-tools v případě ostatních dat. V případě použití UMI-tools na příkladu dat ze small RNA sekvenování za použití QIAseq miRNA Library Kit vypadá extrakce následovně:

```
umi_tools extract --stdin=in.FASTQ.gz --stdout=out.FASTQ.gz --extract-
method=regex --bc-pattern='.+AACTGTAGGCACCATCAAT{s<=2}(?
P<umi_1>.{12})(?P<discard_2>.*)'
```

- +AACTGTAGGCACCATCAAT ponech vše před touto sekvencí (AACTGTAGGCACCATCAAT, adaptorová sekvence) pro použití dál
- {s<=2} povolujeme 2 chyby v patternu před závorkou (v naší adaptorové sekvenci)
- (? P<umi_1>.{12}) sekvence UMI (bude vložena do headru readu)

- (?P<discard_2>.*)' odstraň veškeré další báze po UMI (umi tools jsou původně určeny pro single cell data, jejichž složení readu je adaptér, pak UMI a pak to, co chceme opravdu osekvenovat, takže umi-tools nechávají 5 konec readu automaticky, pokud není řečeno jinak)

Extrakce je prováděna na hrubých readech, ještě před dalším zpracováním, včetně ořezání. Po mapování jsou z mapování odstraněny duplikované ready pomocí příkazy dedup:

```
umi_tools dedup --method=unique -I in.BAM -S out.BAM
```

7.2 Kontrola kvality dat

Kromě již zmíněného Fastp, existují další volně dostupné nástroje pro kontrolu kvality dat, jedním z nejvíc využívaných je FastQC.

FastQC je nástroj napsaný v prostředí Java, který se používá pro kontrolu kvality NGS dat, určený pro prostředí Linux, Mac i Windows. Jako vstupy mohou sloužit soubory FASTQ, BAM a SAM, včetně color-space variant. Výsledné analýzy jsou přehledně vizualizovány v html reportu. Report obsahuje informace o kvalitě jednotlivých bazí, poměru jednotlivých bazí, počtu N bazí, duplikovaných, overreprezentovaných sekvencích (včetně zápisu nabohacených sekvencí a jejich možného zdroje) a o obsahu jednotlivých adaptorových sekvencí. Adaptorové sekvence jsou defaultně nastaveny na Illumina sekvenování, ale je možné přidat do programu vlastní seznam adaptorových a jiných sekvencí, které chce uživatel detekovat

7.2.1 Qualimap

Qualimap je nástroj nezávislý na platformě (Windows, Mac i Linux), napsaný v jazycích Java a R, který se používá pro kontrolu kvality namapovaných čtení spolu s analýzou jejich vlastností. Umožňuje práci v příkazové řádce i v grafickém prostředí. Výsledný soubor obsahuje informace o celkovém a procentuálním počtu readů (namapovaných i nenamapovaných), délkách readů včetně znázornění jejich distribuce, počtu duplikací, zastoupení GC obsahu a mnoha dalšími.

7.2.2 RSeQC

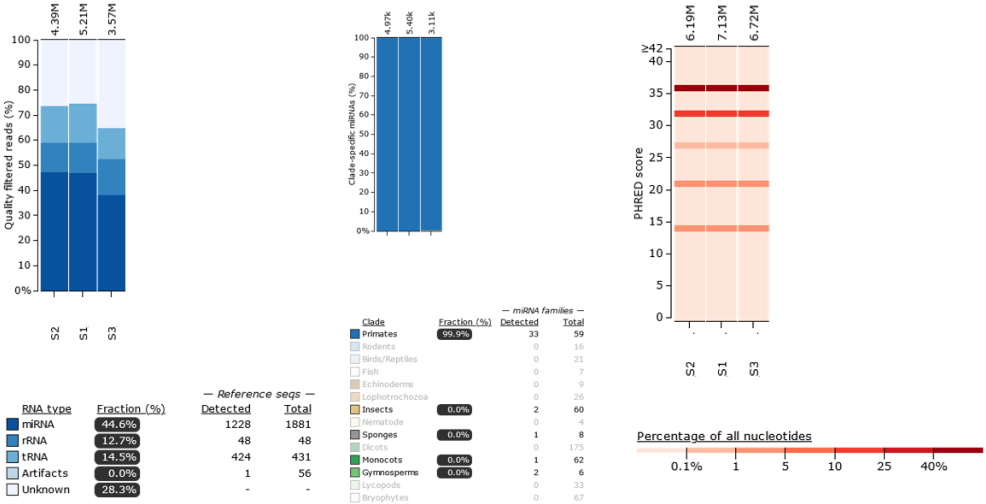
RSeQC je balíček pythonských skriptů pro vyhodnocení dat z masivního paralelního sekvenování. Základní moduly slouží k rychlému vyhodnocení kvality sekvence, PCR a GC biasu a k vizualizaci poměrů nukleotidů. Modul specifický pro RNA-seq slouží k určení směru readů (forward, reverse strandedness), distribuci namapovaných readů, coverage, integrity RNA na úrovni transkriptů a mnoha dalších informací.

7.2.3 Preseq

Preseq je balíček nástrojů v C++ pro Linux a Mac. Vstupními soubory jsou soubory BAM, BED a count histogram – soubor, ve kterém pro každou vlastnost (čtení, gen, cokoli dalšího) je určen počet jeho výskytů. Slouží k vyhodnocení komplexity knihovny, vyhodnocení redundantních readů a k odhadnutí užitečnosti dalšího sekvenování. Je dostupný také jako balíček v R softwaru.

7.2.4 Mirtrace

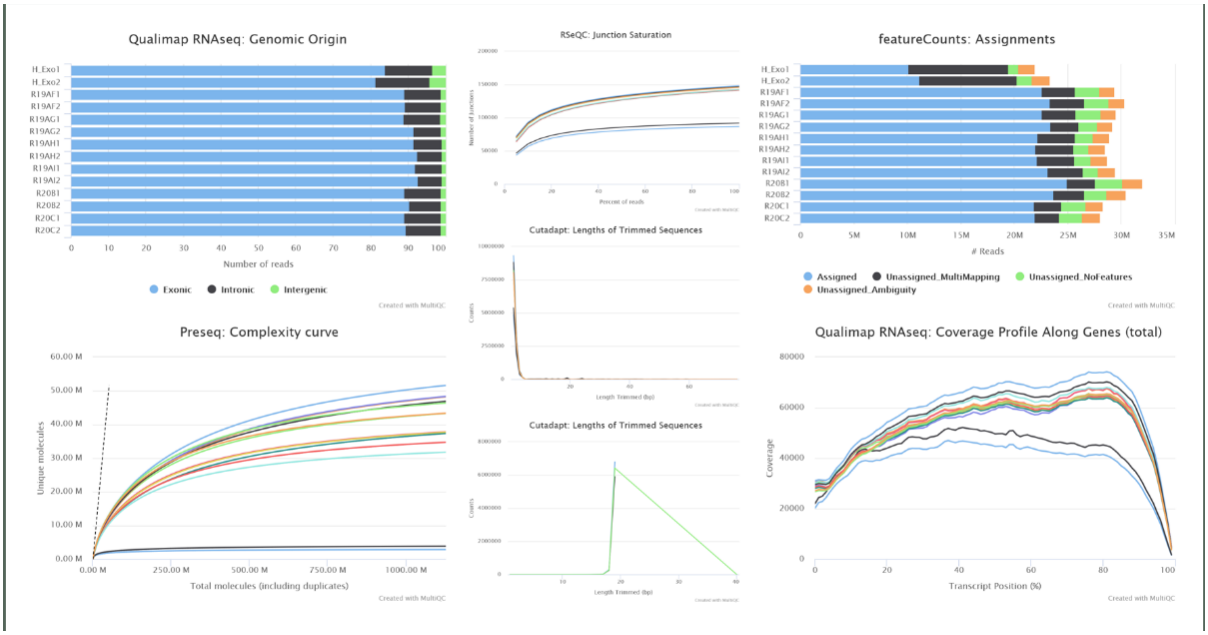
Specializovaným softwarem je Mirtrace. Jedná se o nástroj pro kontrolu kvality microRNAs (QC mode), spolu s určením původu vzorků a případným umožněním odhalení cross-kontaminací (Trace mode, Obr. 44). Výsledky analýzy jsou uloženy do interaktivního html souboru. QC část reportu obsahuje informace o distribuci PHRED skóre, délkách readů, procentuální zastoupení readů, které ne/prošly kontrolou spolu s odůvodněním, znázornění frakcí RNA typu a RNA komplexity. Trace část obsahuje informace o zastoupení druhů v jednotlivých vzorcích, spolu s procentuálním a absolutním vyjádřením frakcí.



Obr. 44: Ukázka výstupů v Mirtrace, vlastní data

7.2.5 MultiQC

MultiQC je nástroj, který shromažďuje výsledky bioinformatických analýz, statistických analýz a kontrol kvality, ze kterých následně tvoří výsledný interaktivní report v html formátu. Momentálně podporuje více jak 110 samostatných bioinformatických nástrojů včetně AfterQC, FastQC, fastp, Cutadapt, SortMeRNA, featurecounts, Trimmomatic, Samtools, Bowtie, Salmon, Star a desítky dalších (Obr. 45).



Obr. 45: Ukázka výstupů v MultiQC, vlastní data, nástroje Qualimap (Coverage a Genomic Origin), RSeQC, featureCounts, Preseq a Cutadapt (délky trimovaných sekvencí)

7.3 Vizualizace sekvencí a zarovnání

Pro vizualizaci získaného zarovnání, sekvencí, a vlastností (variant, inzercí, delecí, apod.) můžeme využít některého z následujících prohlížečů.

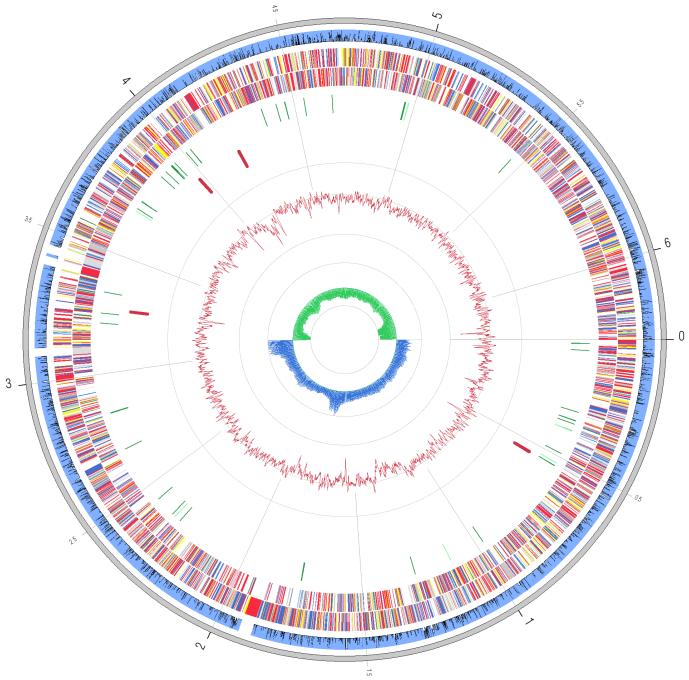
7.3.1 Artemis

Artemis (Wellcome Sanger Institut 2000) je volně dostupný genomový prohlížeč a anotační nástroj, který umožňuje vizualizaci dat a datových analýz ze sekvenování nové generace v rámci kontextu celé sekvence, a také vizualizaci translace ve všech šesti čtecích rámcích. Obzvláště užitečná je pro analýzu malých genomů bakterií, archeí a nižších eukaryot. Artemis umožňuje práci s indexovanými soubory BAM, CRAM, VCF, BCF, GFF3 a FASTA (indexováno pomocí SAMtools), dále s daty zapsanými ve formátu EMBL, GenBank a GFF3. Pro ukládání je možné využít formáty EMBL, GenBank a GFF a Sequincing Table Format. Z vlastností lze zmínit např. vykreslení GC obsahu,

anotování vložené sekvence pomocí EMBL a GenBank, zobrazení proteinového překladu na vymezeném úseku bazí i s filtrací synonymních/nesynonymních variant a zobrazení sekvenčních variant (inzerce, delece, SNP).

7.3.2 Circos

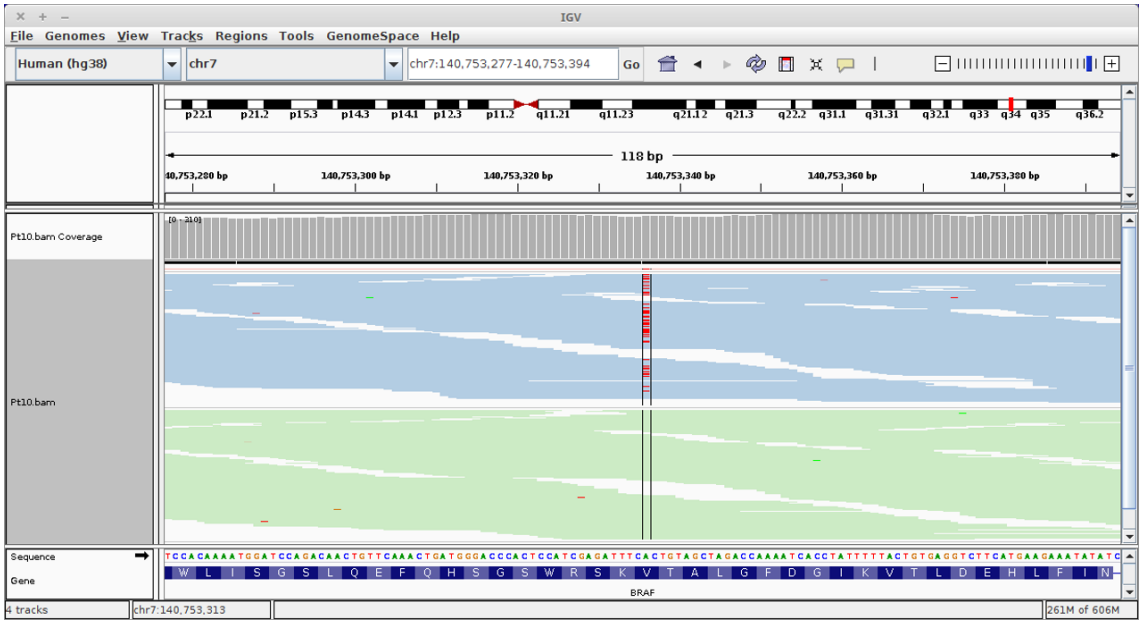
Circos (Circos 2009) je softwarový balíček určený na vizualizaci dat na kruhovém půdorysu. To umožňuje zkoumat vlastnosti mezi objekty a pozicemi. Původně byl určen čistě na vizualizaci genomických dat, jako je zarovnání a strukturní změny, ale umožňuje vykreslovat vztahy v jakémkoli dalším odvětví, např. v matematice, databázových systémech a datové analýze. Je vhodný zejména pro vrstvení různých datových sad a pro vytvoření vysoce informativní infografiky s texturou a vizuální atraktivností. Circos je schopen zobrazovat data také jako 2D bodové a hranové grafy, histogramy, heatmapy, dlaždice, konektory a textové informace. Lze v něm vytvářet bitmapové nebo vektorové obrázky z datových vstupů ve stylu GFF a hierarchických konfiguračních souborů (Obr. 46). Nástroj lze snadno zautomatizovat pomocí jednoduchých textových konfiguračních souborů, které lze začlenit do analyzačních pipeline. Circos se používá k identifikaci a analýze podobností a rozdílů vyplývajících z porovnání genomů, jako je zobrazení genomových přestaveb, SNP variant, vztahů mezi geny a proteiny, identifikace sekvenačních kontigů obsahujících zlomy. Jako vstup slouží data v jednoduchém textovém formátu.



Obr. 46: Rhodococcus erythropolis, Circos, Strnad et al, 2014

7.3.3 Integrative Genomics Viewer IGV

IGV (2012) je volně dostupný vizualizační nástroj pro interaktivní zkoumání velkých, integrovaných genomických datasetů (Obr. 47). Zaměřuje se na vizualizaci a interaktivní průzkum genomických dat v kontextu referenčního genomu. Umožňuje flexibilně vizualizovat mnoho různých typů dat společně, včetně zobrazení informací o vlastnostech vzorku, jako jsou klinické a fenotypové informace a různé anotace. Pro podporu interaktivního průzkumu přímá manipulace v obdobném stylu jako jsou Google Maps. Například klepnutím a tažením se posune pohled přes genom, dvojitým kliknutím na zvolenou oblast se přiblíží pohled pro podrobnější zobrazení. Podporuje interakci v reálném čase na všech stupních rozlišení genomu, od celého genomu až po páry bází, a to i pro velmi velké soubory dat. Datový server Broad IGV je hostitelem mnoha souborů anotací genomu a datových sad z různých veřejných zdrojů.
IGV podporuje širokou škálu datových typů jako namapované sekvenační ready z dat druhé generace, čipová data, mutace, CNV, genové exprese, metylace, a anotace genomu. IGV umožňuje práci s indexovanými i neindexovanými soubory, včetně standartních souborů typy BAM, VCF, BED, také soubory LOG, SEG a WIG. Datové soubory lze načíst z lokálních i vzdálených zdrojů, včetně zdrojů cloudových, což umožňuje porovnávat vlastní genomové datasety s veřejně dostupnými daty. IGV existuje jako desktopová aplikace, webová aplikace, ale i javaskriptová aplikace pro vlastní webovou stránku. Hlavní důraz je však kladen na podporu biomedicínských výzkumných pracovníků, kteří chtějí načíst, vizualizovat a zkoumat vlastní soubory dat, ale mohou také zpřístupnit své datové soubory ostatním pro prohlížení v IGV, a sdílet je s kolegy nebo vědeckou společností obecně.



Obr. 47: V600E mutace v melanomech, IGV, Kolář M

7.3.4 Tablet

Tablet je vysoce výkonný grafický prohlížeč určený pro sekvence a zarovnání NGS dat. Podporuje práci se soubory typu ACE, AFG, MAQ, SOAP2, SAM, BAM, FASTA, FASTQ a GFF3. Umožňuje vyhledat, zobrazit a zvýraznit klíčové vlastnosti souborů GFF3, VCF, GTF a BED. Podporuje zobrazování informací ze CIGAR pole v SAM/BAM souborech, včetně zobrazení pair-endových dat. Další jeho předností je jednoduchá instalace a podpora pro všechny operační systémy, včetně 64bit verzí.

7.3.5 ASCIIGenome

ASCIIGenome je genomový prohlížeč založený na rozhraní příkazového řádku, který zobrazuje data přímo v oknu terminálu. Podporuje formáty GFF, GTF, bigBED, BED, bigwig, BEDGraph, TDF, VCF a BAM. Textové formáty mohou být nahrávány i v GZIP kompresované formě. Většinu zdrojů dat lze číst ze vzdálených adres URL, jako jsou adresy dostupné v prohlížeči genomu UCSC nebo Ensembl.

7.4 Otázky k tématu

1. Co je trimování sekvencí?
2. Z jakého důvodu se odstraňují adaptorové sekvence?
3. Jaké znáte nástroje na odstraňování nežádoucích sekvencí?
4. Co jsou UMI a k čemu se využívají?
5. Jaké jsou hlavní vlastnosti FastQC?
6. Jak se dá využít MultiQC?
7. Jaké znáte nástroje na vizualizaci výsledků?
8. Jaké znáte genomové prohlížeče?

7.5 Zdroje

Odstranění nežádoucích sekvencí

Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*. 2014;30(15):2114-2120. doi:10.1093/bioinformatics/btu170

Cutadapt. Cutadapt documentation. https://cutadapt.readthedocs.io/en/stable/index.html (accessed Dec 09, 2020).

Chen S, Huang T, Zhou Y, Han Y, Xu M, Gu J. AfterQC: automatic filtering, trimming, error removing and quality control for fastq data. *BMC Bioinformatics*. 2017;18(Suppl 3):80. Published 2017 Mar 14. doi:10.1186/s12859-017-1469-3

Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics*. 2018;34(17):i884-i890. doi:10.1093/bioinformatics/bty560

Kopylova E, Noé L, Touzet H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics*. 2012;28(24):3211-3217. doi:10.1093/bioinformatics/bts611

Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet Journal. https://journal.embnet.org/index.php/embnetjournal/article/view/200/479 (accessed Dec 09, 2020).

Smith, T.; Sudbery, I. Taking appropriate QC measures for RRBS-type or other -Seq applications with Trim Galore!. https://github.com/FelixKrueger/TrimGalore/blob/master/Docs/Trim_Galore_User_Guide.md (accessed Dec 09, 2020).

Smith, T.; Sudbery, I. Trim Galore. Babraham Bioinformatics. http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/ (accessed Dec 09, 2020).

Smith, T.; Sudbery, I. UMI-tools Tools for dealing with Unique Molecular Identifiers. UMI tools. https://umi-tools.readthedocs.io/en/latest/index.html (accessed Dec 09, 2020).

Smith T, Heger A, Sudbery I. UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. Genome Res. 2017;27(3):491-499. doi:10.1101/gr.209601.116

Trimmomatic Manual: V0.32. THE USADEL LAB. http://www.usadellab.org/cms/uploads/supplementary/Trimmomatic/TrimmomaticManual_V0.32.pdf (accessed Dec 09, 2020).

Kontrola kvality dat

Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*. 2016;32(19):3047-3048. doi:10.1093/bioinformatics/btw354
FastQC. Babraham Bionformatics. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (accessed Dec 09, 2020).
García-Alcalde F, Okonechnikov K, Carbonell J, et al. Qualimap: evaluating next-generation sequencing alignment data. *Bioinformatics*. 2012;28(20):2678-2679. doi:10.1093/bioinformatics/bts503
Kang, W., Eldfjell, Y., Fromm, B. et al. miRTrace reveals the organismal origins of microRNA sequencing data. Genome Biol 19, 213 (2018). https://doi.org/10.1186/s13059-018-1588-9
Qualimap 2.2.1 documentation. Qualimap Evaluating next generation sequencing alignment data. http://qualimap.conesalab.org/doc_html/index.html (accessed Dec 09, 2020).
Smith, A. D.; Daley, T. Preseq. https://github.com/smithlabcode/preseq (accessed Dec 09, 2020).
Wang, L., Wang, S., & Li, W. (2012). RSeQC: quality control of RNA-seq experiments. Bioinformatics (Oxford, England), 28(16), 2184–2185. http://doi.org/10.1093/bioinformatics/bts356

Vizualizace dat

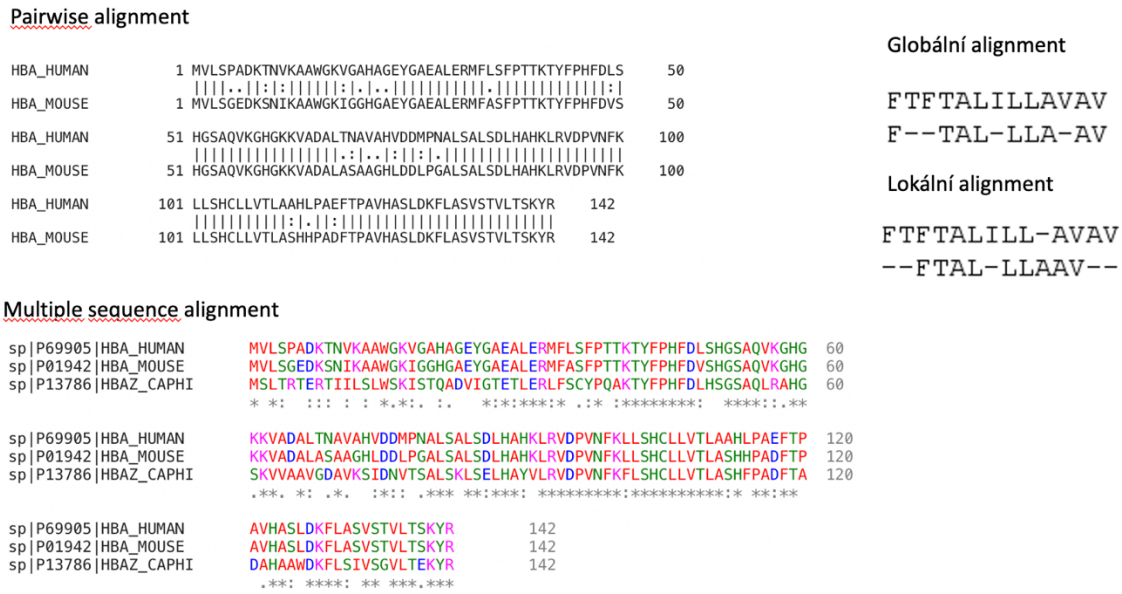
Application of Circos to Genomics. Circos, Circular Genome Data Visualization. http://circos.ca/intro/genomic_data/ (accessed Dec 09, 2020).
Artemis manual. Wellcome Sanger Institute Pathogen Informatics. https://sanger-pathogens.github.io/Artemis/Artemis/artemis-manual.pdf (accessed Dec 09, 2020).
Carver T, Harris SR, Berriman M, Parkhill J, McQuillan JA. Artemis: an integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics*. 2012;28(4):464-469. doi:10.1093/bioinformatics/btr703
Dario Beraldi, ASCIIGenome: a command line genome browser for console terminals, Bioinformatics, Volume 33, Issue 10, 15 May 2017, Pages 1568–1569, https://doi.org/10.1093/bioinformatics/btx007
Krzywinski M, Schein J, Birol I, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009;19(9):1639-1645. doi:10.1101/gr.092759.109
Milne I, Stephen G, Bayer M, et al. Using Tablet for visual exploration of second-generation sequencing data. *Brief Bioinform*. 2013;14(2):193-202. doi:10.1093/bib/bbs012
Milne I, Bayer M, Cardle L, et al. Tablet--next generation sequence assembly visualization. *Bioinformatics*. 2010;26(3):401-402. doi:10.1093/bioinformatics/btp666
Strnad H, Patek M, Fousek J, et al. Genome Sequence of Rhodococcus erythropolis Strain CCM2595, a Phenol Derivative-Degrading Bacterium. Genome Announc. 2014;2(2):e00208-14. Published 2014 Mar 20. doi:10.1128/genomeA.00208-14

Robinson JT, Thorvaldsdóttir H, Wenger AM, Zehir A, Mesirov JP. Variant Review with the Integrative Genomics Viewer. *Cancer Res*. 2017;77(21):e31-e34. doi:10.1158/0008-5472.CAN-17-0337
Rutherford K, Parkhill J, Crook J, et al. Artemis: sequence visualization and annotation. *Bioinformatics*. 2000;16(10):944-945. doi:10.1093/bioinformatics/16.10.944
Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform*. 2013;14(2):178-192. doi:10.1093/bib/bbs017

8 ALIGNMENT

Jako (sekvenční) alignment označujeme zarovnání dvou (pairwise alignment), či více (multiple sequence alignment) sekvencí (aminokyselinových, nukleotidových) na základě jejich vzájemné podobnosti (Obr. 48). Sekvence mohou být zarovnány s úmyslem nalézt co nejlepší shodu napříč celou délkou sekvencí – globální alignment, nebo pouze v rámci určitého/určitých úseků – lokální alignment. Z tohoto důvodu se globální alignment používá spíše pro srovnávání stejně či alespoň podobně dlouhých sekvencí s vyšší mírou podobnosti, zatímco lokální alignment se využije pro naleznutí úseků s vysokou mírou shody v sekvencích, které si nejsou příliš podobné.

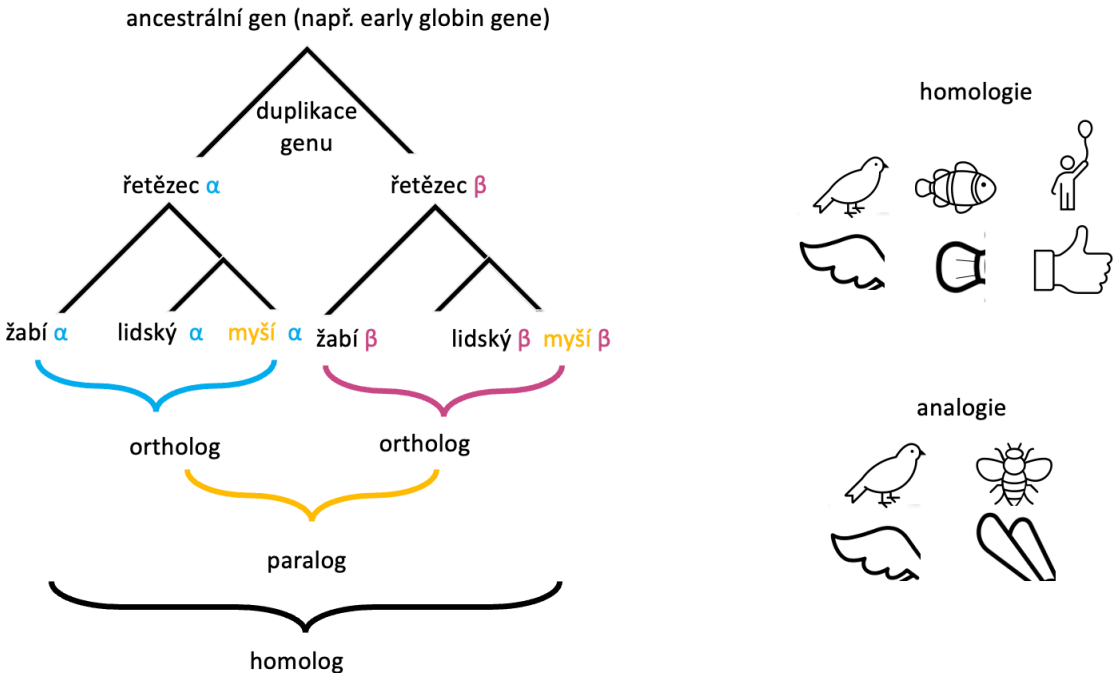
U alignmentu nám jde o vyhodnocení, zda dané sekvence mohou mít společného předka. Chceme proto, aby naše substituční matice přiřadila skóre alignmentu vyhodnocením poměru mezi [pravděpodobnost že sekvence mají společného předka] proti [pravděpodobnost, že sekvence byly zarovnány náhodou] (očekáváme zlepšení oproti náhodnému).



Obr. 48: Členění alignmentů, pairwise, multiple a rozdíl mezi globálním a lokálním, vlastní data

Význam alignmentu nalézáme v následujících oblastech: identifikace sekvence v databázích, vyhledání podobných sekvencí v databázi, detekce mutací (variant calling), hledání konzervovaných částí sekvence (často regiony s určitou funkční), odhalování příbuzných vztahů mezi sekvenci (evoluce, produkce fylogenetických stromů), odhad vztahu struktury a funkce makromolekul a předpověď vyšších struktur.

V rámci evoluční biologie se můžeme setkat s výrazem homologie. Homologie označuje podobnost struktury, fyziologie nebo vývoje různých druhů organismů/znaků na základě jejich původu ze společného evolučního předka, Homologické geny jsou tedy ty, které se v evoluci mění spolu. Jako paralogy pak označujeme historicky homologické geny, které vznikly (opakovanou) duplikací v rámci jednoho genomu. Homologické geny, které se vyskytují u různých organizmů, jsou ortology, analogie („falešná homologie“, homoplasie) označuje podobnost díky nezávislé evoluci nevedoucí ke společnému předkovi (vztah křídla ptáků a hmyzu, Obr. 49). Homologie je základem pro hierarchické srovnávání v evoluční biologii.



Obr. 49: Vztah homolog, paralog, ortholog, Pfeiferová L

Kromě sekvenčního rozeznáváme i strukturní alignment, kterým porovnáváme proteiny (případně i molekuly RNA) ne čistě na základě sekvenční podobnosti, ale převážně podle jejich tvaru a prostorové konformace.

Pro alignment protein kódujících sekvencí je vhodnější použít aminokyselinové sekvence namísto nukleotidů. Existuje několik důvodů, proč je jejich použití vhodnější. Předně existuje 21 (podle jiných teorií/zdrojů můžeme narazit na 20, 21 i 22 základních) aminokyselin, ale pouze 4 různé (základní) nukleotidy. Tím pádem můžeme vypočítat statistiku i pro mnohem kratší alignmenty aminokyselin, oproti nukleotidovým. Dále se při porovnávání aminokyselinových sekvencí bere v potaz i pravděpodobnost záměny jedné aminokyseliny za jinou. Dalším důvodem je degenerovaný genetický kód, kde zhruba 30 % všech nukleotidových substitucí nezpůsobí záměnu aminokyselin, tím pádem nejsou pod selekčním tlakem a vytvářejí šum v datech.

8.1 Výpočet skóre a substituční matice

Každé dvojici sekvencí je po zarovnání přirazeno číslo – skóre, které určuje míru jejich podobnosti. Platí zde jednoduchá úměra: čím vyšší je skóre, tím vyšší je podobnost. Podle použité substituční matice (skórovací funkce) může výsledné skóre dosahovat i záporných hodnot. Příklad jednoduchého výpočtu (Obr. 50).

```
A A E E C C D D E E F
A A D D K K K E F G G
4+4+2+2-3-3-1+2-3-2-3          = -1
```

Obr. 50: Ukázka jednoduchého výpočtu skóre s využitím matice BLOSUM62

8.1.1 Vznik a penalizace mezer

V rámci zarovnání musíme počítat i s možným vznikem mezer, které nám umožňují zarovnat sekvence, ve kterých došlo k inzerci/deleci. Inzerce/delece v alignmentu může vzniknout z různých důvodů, jedním z nejčastějších důvodů je vznik bodové mutace v jedné ze sekvencí. Dalšími důvody jsou: nepřesný crossover při meióze (inzerce nebo delece řetězce bází), DNA slippeage při replikaci (vznik repetice v řetězci) anebo translokace DNA mezi chromozomy. Mezery můžeme v rámci alignmentu nacházet libovolně v řetězci, na začátku, uprostřed i na konci (Obr. 51).

```
MVLSPA-DKTNVKAAWGKVGAHAG-EYGAEALERMFLSFPTTKTYFPHFD
|||| . ||:|:|||||||:|.| | ||||||||||.||||||||||||
MVLS-GEDKSNIKAAWGKIGGH-GAEYGAEALERMFASFPTTKTYFPHFD
```

Obr. 51: Zobrazení možných mezer v alignmentu

Neexistují teorie pro odvození ceny mezer (gap penalty), nicméně obecně platí, že cena za otevření mezery je vyšší než za její prodloužení, i kvůli tomu, že tvorbou mezer se zvyšují pravděpodobnost alignmentu náhodných sekvencí. Navíc, pokud je gap penalty nízká, můžeme sice získat „lepší“ alignment, ale z hlediska biologie také může jít o nesmysl.

Je také zapotřebí si uvědomit, že indely se často objevují v blocích, takže jednoduchá lineární funkce (cena za otevření mezery plus za prodloužení) není dostatečně přesná, proto se využívá realističtější schéma, tzv. affine gap model.

8.1.2 Substituční matice

Jak již bylo řečeno, výběr substituční matice, ovlivňuje výsledné skóre (a další přidanou statistiku). Vycházíme z předpokladu, že sekvence mají nebo mohou mít společného předka. Substitučními maticemi rozumíme teoreticky i empiricky odvozené matice, ve kterých jsou zapsány frekvence, při kterých se sledovaný znak (nukleotidová nebo aminokyselinová substituce) v nukleotidové nebo proteinové sekvenci změnil v rámci evoluce na jiný. Informace v nich jsou často zapsané v matici jako zlogaritmované pravděpodobnosti nalezení dvou specifických stavů zarovnaných znaků (log-odds matice), a jsou závislé na předpokládaném počtu evolučních změn nebo sekvenční nepodobnosti mezi porovnávanými sekvencemi. Známe několik různým substitučních matic, PAM, BLOSUM, GONNET, DNA identity matrix (skoro jediná zaměřená na nukleotidy, vyhodnocení transicí a transverzí), Kimurův model, WAG; my se zde budeme věnovat prvním dvěma.

PAM (Point accepted mutation)

Jedna z vůbec prvních substitučních matic (log-odds matice, teoreticky odvozená) vypočítaná Margaret Dayhoff v 70. letech 20. století, založená na mutacích v rámci globálního alignmentu. Pro výpočet matice PAM1 byla vybrána skupina velmi blízce příbuzných sekvencí s frekvencemi mutací odpovídajícími jedné jednotce PAM (1 % aminokyselinových pozic, které byly změněny). Celkem bylo použito použil 71 proteinových rodin, u kterých Dayhoff vytvořila hypotetické fylogenetické stromy a zaznamenala počet pozorovaných substitucí (podél každé větve stromu) v cílové matrici 20x20. Na základě shromážděných mutačních dat z této skupiny sekvencí pak byla odvozena substituční matice. Matice PAM1 ve výsledku ukazuje, jaká rychlost substituce by se očekávala, kdyby se změnilo 1 % aminokyselin. PAM1 matice se dá použít jako základ pro výpočet dalších matric za předpokladu, že opakované mutace by se řídily stejným vzorem jako ty v matrici PAM1, a s tím, že se na stejném místě mohou vyskytovat vícenásobné substituce (M. Dayhoff spočítala matice až do PAM250). Problémem těchto matic je to, že sekvenční změny v rámci dlouhých evolučních časových měřítcích se nedají dobře aproximovat slučováním malých změn, ke kterým dochází během krátkých časových měřítek. PAM matice se tedy hodí na porovnávání blízce příbuzných proteinů.

BLOSUM (BLOck SUBstitution Matrix)

Matice BLOSUM (log-odds matice, empiricky odvozená), poprvé popsané v 90. letech Stevnem Henikoffem a Jorjou Henikoffovou, napravují problém změn v dlouhých evolučních měřítkách, se kterými se potýkají matice PAM. Vycházejí z většího množství více rozmanitých proteinů. Výpočet prvních matic probíhal následovně: v databázi BLOCKS se vyhledaly celé bloky konzervovaných oblastí v rámci proteinových rodin (bez mezer v zarovnání; předpokládá se, že konzervované oblasti nesou funkci, a proto

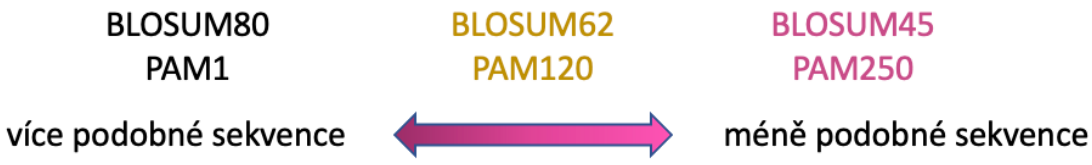
mají nižší míru substitucí). Následně byly v těchto blocích vypočítány relativní zastoupení aminokyselin a pravděpodobnosti jejich substituce. Aby se snížila odchylka od blízce příbuzných sekvencí na rychlosti substituce, segmenty v bloku se sekvenční identitou nad určitým prahem byly seskupeny, čímž se snížila váha každého takového shluku – pro matici BLOSUM62 byl tento práh určen na 62 %, frekvence se tedy počítaly mezi shluky, takže páry byly počítány pouze mezi segmenty s méně než 62 % podobností. Poté se vypočítala logaritmická skóre pro každý ze 190 možných substitučních párů 20 standardních aminokyselin. Matice BLOSUM se využívají v blastp, a jsou vhodné pro identifikaci neznámých nukleotidových sekvencí.

BLOSUM matrices vysokými čísly je dobrá pro porovnaní́ vysoce příbuzných sekvencí, zatímco nízké́ pro relativně́ vzdálené́ podobnosti
Fun fact: U matice BLOSUM62 byla objevena chyba ve výpočtu, avšak po její opravě matice funguje biologicky hůře

Stručné vysvětlení log-odds: log-odds je logaritmická pravděpodobnost s jakou by se aminokyselina **a** mohla změnit (mutovat) na aminokyselinu **b** v rámci evoluce, na základě omezení našeho evolučního modelu

8.1.3 Rozdíly mezi maticemi PAM a BLOSUM

1. Matrice PAM jsou založeny na mutacích pozorovaných během globálního zarovnání, což zahrnuje jak vysoce konzervované, tak vysoce mutabilní oblasti. Matrice BLOSUM jsou založeny pouze na vysoce konzervovaných oblastech v sériích zarovnání (lokální), které neobsahují mezery
2. Matice BLOSUM jsou založeny na pozorovaných zarovnáních; nejsou extrapolovány ze srovnání blízce příbuzných proteinů jako PAM
3. Vyšší čísla ve schématu pojmenování matice PAM znamenají větší evoluční vzdálenost, zatímco vyšší čísla u matice BLOSUM znamenají vyšší sekvenční podobnost, a tedy menší evoluční vzdálenost (Obr. 52).



Obr. 52 Příklad: PAM250 se používá pro vzdálenější sekvence než PAM1; BLOSUM80 se používá pro bližší sekvence než BLOSUM46, Pfeiferová L

Hezké vysvětlení výpočtu a backgroundu za substitučními maticemi PAM a BLOSUM najdete na https://www.youtube.com/watch?v=68lF71zEUF8

8.2 Pairwise aligment

Pairwise alignment se používá k přímému porovnání dvou sekvencí spolu s nalezením nejlépe odpovídajícího lokálního nebo globálního zarovnání, případně overlap a semi-globálního, a k vyhledání podobných sekvencí v databázi. Pro vytváření párového zarovnání můžeme použít následující základní metody: dotplot, a dynamické programování, případně jejich implementace. V rámci dynamického programování využíváme Smith-Watermanův algoritmus pro lokální a Needleman-Wunschův algoritmus pro globální zarovnání. Specifickým problémem je alignment genů, kde musíme u eukaryotických organismů počítat s introny.

8.2.1 Dot plot

Patrně nejjednodušším (ale zároveň poměrně časově náročným) provedením pairwise alignmentu je Dot plot, metoda grafického srovnání dvou biologických sekvencí využívající tvorbu rekurentního grafu. Princip je následující: vytvoří se dvoudimenzionální matice, kdy jedna sekvence se nachází na horizontální ose a druhá na vertikální. Srovnání je pak provedeno jednoduše tak, že postupně pro každý prvek obou sekvencí procházíme řádky/sloupce, ve kterém se nachází a zaznamenáváme shody s druhou sekvencí, takže odpovídající segmenty sekvence se objeví jako běhy diagonálních čar napříč maticí (Obr. 53). Dot plot lze mimo jiné použít k zobrazení repetitivních oblastí. Oblasti, které jsou výrazně podobné, se na grafu projeví jako diagonální čáry mimo hlavní úhlopříčku. Tento efekt může nastat, když se protein skládá z více podobných strukturních domén.
Náhodné shody mohou být pro větší přehlednost odfiltrovány za pomoci nastavení okna (window size), kdy jsou porovnávány celé bloky sekvencí (po 3, 4, 10, více, v zásadě podle délky sekvence), a shoda se zaznamená pouze v případě dosáhnutí určité minimální thresholdové hodnoty. Vysvětlení dotplotu si můžete poslechnout na https://www.youtube.com/watch?v=nnPNwlzX2qg&t=8s

	T	**A**	**T**	**C**	**G**	**A**	**A**	**G**	**T**	**A**
T	X		X						X	
A		X				X	X			X
T	X		X						X	
T	X		X						X	
C				X						
T	X		X						X	
A		X				X	X			X

Obr. 53: Zobrazení dotplotu, vlastní data

8.2.2 Globální alignment

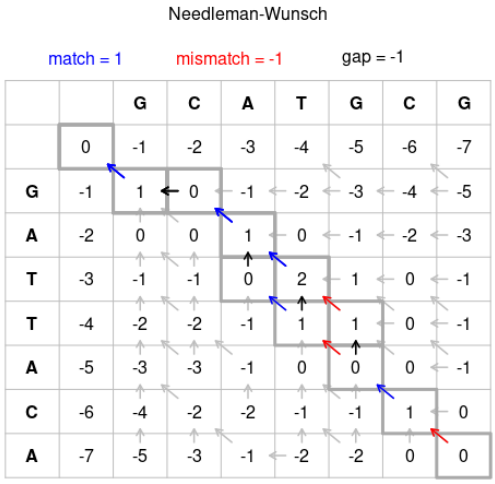
Algoritmus Needleman-Wunch je příkladem globální alignmentu založeném na dynamickém modelování, který zarovnává podobné sekvence v celé jejich délce. Z tohoto důvodu je vhodným algoritmem pro porovnání podobně/stejně dlouhých sekvencí s vysokou podobností. Od algoritmu Smith-Watermanu se liší v tom, že vyžaduje penaltu za mezeru (u Smith-Watermana není povinná, ale lze ji inkorporovat). Ta se započítává, pokud je nejvyšší skóre nad anebo vlevo od hodnocené buňky. Needleman-Wunschův algoritmus zároveň povoluje záporné hodnoty výsledného skóre.

Skóre se počítá podle následující rovnice (2), $A = a_1 a_2 \dots a_n$ a $B = b_1 b_2 \dots b_m$ jsou sekvence o délce n a m, které chceme zarovnat.

$$S_{ij} = max \begin{cases} S_{i-1,j-1} + s(a_i b_j), \\ max_{k \geq 1}\{S_{i-k,j} - W_k\}, \\ max_{l \geq 1}\{S_{i,j-l} - W_l\} \end{cases} \qquad \text{platné pro } 1 \leq i \leq n,\ 1 \leq j \leq m, \quad (2)$$

Algoritmus je následující (Obr.54):

1. Určí se substituční matice a gap penalty, s(a, b) je skóre podobnosti prvků, které představují dvě sekvence, W_k je penalizace mezery o délce k
2. Sestaví se skórovací matice S a inicializuje se její první řádek a sloupec. Velikost skórovací matice je (n+1) x (m+1) s indexováním od 0. $S_{k0} = S_{0l} = 0$ pro $0 \leq k \leq n$ a $0 \leq l \leq m$
3. Vyplní se matice pomocí rovnice (2), kde $S_{i-1,j-1} + s(a_i b_j)$ je skóre alignmentu a_i a b_j, $S_{i-kj} - W_k$ je skóre, pokud je a_i na konci délky mezery o délce k, $S_{i,j-l} - W_l$ skóre, pokud je b_j na konci délky mezery o délce l
4. Traceback, začíná se (odzadu) od nejvyššího skóre S a končí se na buňce matice se skóre 0



		G	C	A	T	G	C	G
	0	-1	-2	-3	-4	-5	-6	-7
G	-1	1	0	-1	-2	-3	-4	-5
A	-2	0	0	1	0	-1	-2	-3
T	-3	-1	-1	0	2	1	0	-1
T	-4	-2	-2	-1	1	1	0	-1
A	-5	-3	-3	-1	0	0	0	-1
C	-6	-4	-2	-2	-1	-1	1	0
A	-7	-5	-3	-1	-2	-2	0	0

Obr. 54: Globální alignment, Slowkow, CC0, via Wikimedia Commons

8.2.3 Lokální alignment

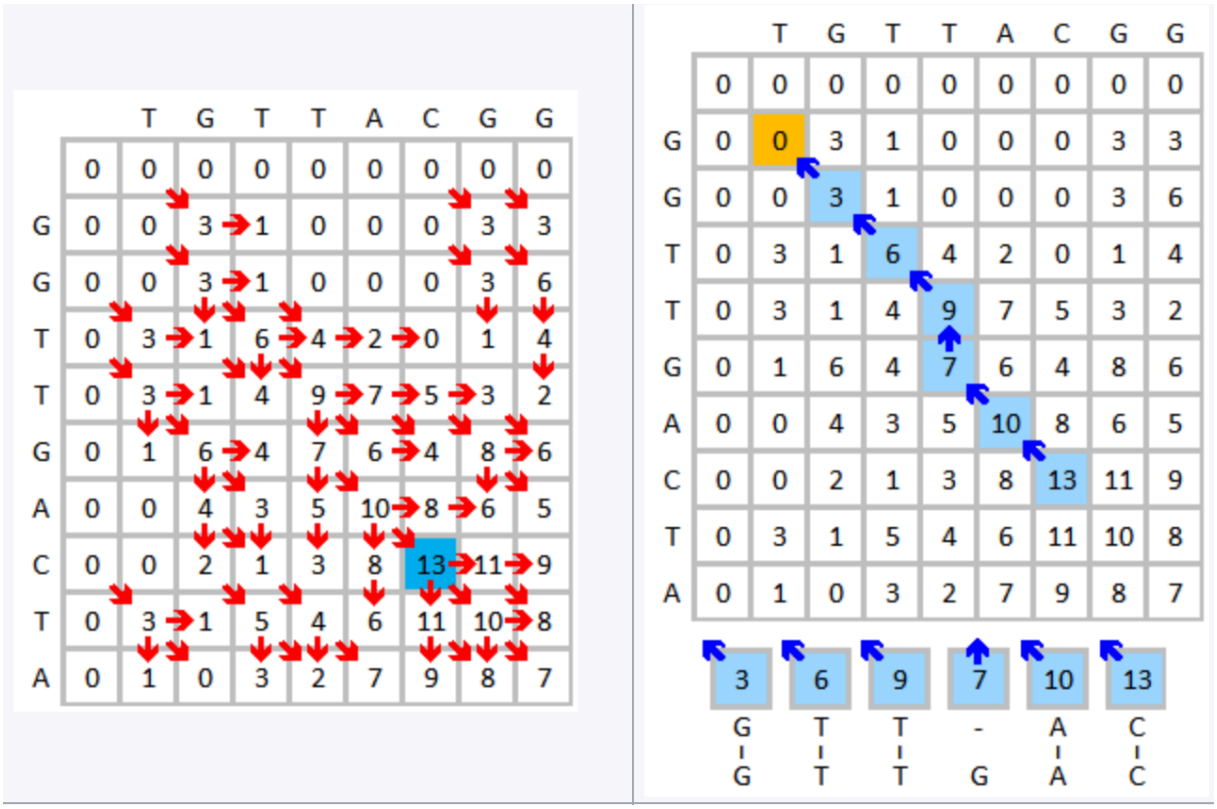
Lokální alignment používá algoritmus Smith-Waterman, který využívá dynamické programování (rozdělení komplexního problému na sérii malých úkolů. To znamená, že se nesnažíme zarovnat celou sekvenci, ale hledáme shodu nebo rozdíl u dvojic nukleotidových pozic). Jedná se o specifický případ algoritmu Needleman-Wunch. U tohoto algoritmu se nemusejí počítat mezery (i když je možné je do algoritmu inkorporovat, a my je počítat budeme), a výsledné skóre nemůže vycházet v záporných číslech.

Skóre se počítá podle následující rovnice (1), $A = a_1 a_2 \dots a_n$ a $B = b_1 b_2 \dots b_m$ jsou sekvence o délce n a m, které chceme zarovnat.

$$S_{ij} = max \begin{cases} S_{i-1,j-1} + s(a_i b_j), \\ max_{k \geq 1}\{S_{i-kj} - W_k\}, \\ max_{l \geq 1}\{S_{i,j-l} - W_l\}, \\ 0 \end{cases} \qquad \text{platné pro } 1 \leq i \leq n,\ 1 \leq j \leq m, \ (1)$$

Algoritmus je následující (Obr. 55):

1. Určí se substituční matice a gap penalty, s(a, b) je skóre podobnosti prvků, které představují dvě sekvence, W_k je penalizace mezery o délce k.
2. Sestaví se skórovací matice S a inicializuje se její první řádek a sloupec. Velikost skórovací matice je (n+1) x (m+1) s indexováním od 0. $S_{k0} = S_{0l} = 0$ pro $0 \leq k \leq n$ a $0 \leq l \leq m$.
3. Vyplní se matice pomocí rovnice (1), kde $S_{i-1,j-1} + s(a_i b_j)$ je skóre alignmentu a_i a b_j, $S_{i-kj} - W_k$ je skóre, pokud je a_i na konci délky mezery o délce k, $S_{i,j-l} - W_l$ skóre, pokud je b_j na konci délky mezery o délce l a 0 pokud není podobnost mezi dvojicí $a_i b_j$ (výsledná hodnota byla záporná).
4. Traceback, začíná se (odzadu) od nejvyššího skóre S a končí se na buňce matice se skóre 0.



Obr. 55: Lokální alignemnt Smith-Waterman algoritmus, skóre za shodu (match) +3 a za rozdíl (mismatch) –3, gap penalty 2 (2k, k značí délku mezery), Yz cs5160, CC BY-SA 4.0

Lokálního alignmentu využívají nástroje BLAST a FASTA (viz níže).

8.2.4 Overlap alignment

Sekvenci si můžeme rozdělit na prefix část („přední“) sufix část („zadní“). Overlap alignment se pak dá popsat jako případ, kdy ignorujeme prefix prvního řetězce a sufix druhého řetězce, pokud zbývající podřetězce (sufix-prefix overlaping) poskytují dobrý alignment. Overlapy jsou důležité pro fragment assembly, kdy se zpětně skládá genomická sekvence z náhodných subsetů podřetězců.

8.2.5 Gene alignment

Ve chvíli, kdy máme proteinovou sekvenci z neznámého genu, a hledáme sekvenci tohoto genu v DNA. U prokaryot použijeme semi-globální alignment, což je specifický podtyp alignmentu, kdy je několik prvních řádků nastaveno na hodnotu 0. Pak můžeme skóre $S_{i-1,j-1} + S(a_i, b_j)$ nahradit za $S_{i-1,j-3} + s(a_i, aa[B[j-2..j]]$, kde A je proteinová sekvence, B je DNA sekvence, aa[xyz] je aminokyselina přepsaná do DNA kodonu. U eukaryot je situace nicméně obtížnější, protože musíme brát v potaz sekvence intronů. Nejběžnějším způsobem, jak je zakomponovat, je použít penalizaci affine gap, a přiřadit velkou cenu otevření mezery a naopak velmi malou cenu za rozšíření mezery na straně DNA, a na straně proteinu pak použít lineární cenu mezery (otevření, plus extenze). Také je dobré si uvědomit, že počet intronů je omezený.

8.2.6 Nástroje využívající lokální alignment

Heuristická aproximace Smith-Watermanova algoritmu se uplatňuje ve vyhledávání podobných sekvencí v sekvenčních databázích. Pro posouzení nejpodobnějších slouží statistické vyhodnocení vzhledem k míře podobnosti dané čistě náhodou (E-Value).

Proč je správnější dívat na E-Value: skóre může být hezké, ale reálně nemusí vypovídat o podobnosti – může se jednat i jen o alignmentem 8 aminokyselin

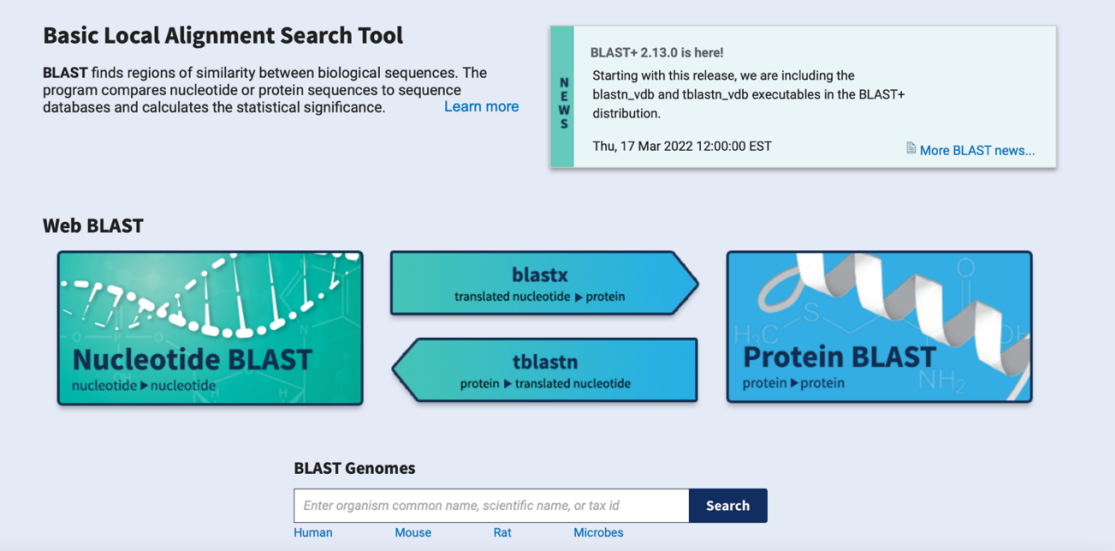
8.2.7 **Důležité pojmy ke statistice**

Skóre S: numerická hodnota, která popisuje celkovou kvalitu zarovnání. Vyšší čísla odpovídají vyšší podobnosti. Stupnice skóre závisí na použitém systému bodování (substituční matice, penalizace mezery).
Pokrytí (Query cover): jaká část zájmové sekvence je porovnávaná s nalezenou sekvencí.
lambda: parametr závislý na substituční matici a na ceně za vytvoření mezery. Používá se na převedení skóre S na normalizované S´ v bit score.
lambda ratio: lambda pro daný bodovací systém v poměru s tím, který používá stejné skóre substituce, ale s nekonečnými náklady na mezeru. Tento poměr ukazuje, jaký podíl informací v zarovnání bez mezer musí být obětován v naději na zlepšení jeho skóre prostřednictvím rozšíření pomocí mezer.
E-hodnota (E-value): kolikrát je možné v prohledávané databázi očekávat sekvenci se stejným (či lepším) skóre alignmentu jako má nalezená sekvence vůči zájmové čistě náhodou. E-hodnota by měla být co nejmenší, ideálně nula. Z definice E-value závisí na velikosti databáze, a proto se v čase a s množstvím sekvencí, které do databází vstupují mění. Pro „normální“ databázi se jako treshold, kdy můžeme výsledek považovat za solidní bere E-value 10^{-5}.
Bit-score: normalizované skóre vyjádřené v bitech, které vám umožní odhadnout velikost vyhledávacího prostoru, kterou byste museli prohlédnout, než byste očekávali, že najdete skóre tak dobré jako nebo lépe než tohle náhodou.

Rozšířené informace ke statistice využívané v Blastu najdete na této stránce, https://www.ncbi.nlm.nih.gov/BLAST/tutorial/Altschul-1.html.

8.2.8 **BLAST**

Jedním z nejčastěji používaných algoritmů pro lokální alingment je BLAST (zkratka z anglického The **B**asic **L**ocal **A**lignment **S**earch **T**ool) od NCBI (Obr. 56, 57). Základní algoritmus je jednoduchý a robustní, má dobrý poměr mezi citlivostí a rychlostí analýzy. Podle použití se porovnává nukleotidová nebo proteinová sekvence s velkými DNA nebo proteinovými sekvenčními databázemi jako je GenBank a EBI (na rozdíl od FASTA má oddělené programy pro práci s DNA resp. proteiny). Vzhledem ke své flexibilitě je možné jej použít jak přímo k identifikaci a anotaci vyhledávaného úseku, tak v případě dosud neanotovaných genomických dat také k odvození funkčních i evolučních vztahů mezi sekvencemi a k zařazení genu do rodiny. Kromě popisu porovnaných sekvencí program dále vypočítává statistickou významnost pomocí bit-skóre, procentuální shody a E-hodnoty (Obr. 58, 59).



Obr. 56: Ukázka hlavní stránky BLASTu, NCBI Basic Local Alignment Search Tool. https://blast.ncbi.nlm.nih.gov/Blast.cgi (accessed April 30, 2022)

BLAST identifikuje shodné sekvence pomocí heuristické metody, která aproximuje Smith-Watermanův algoritmus. BLAST nehledá identitu přímo, ale na základě indexu. Zjednodušeně identifikuje shody mezi dvěma sekvencemi, jejichž skóre již nejde zvýšit rozšířením dané oblasti a označí je jako HSP (High segment scoring pair). Základem jsou 3 algoritmické kroky: K dotazované sekvenci spočítá všechny **slova dané délky se skóre podobnosti lepším než treshold**. Těmito slovy pak prohledává databázi a **hledá hity**. Vezme nejlepší oblast a hledá na obě strany od ní, ideálně dokud nedosáhne **HSP**. Není zaručeno nalezení optimálního zarovnání a veškerých shod, některé hity mohou být vynechány.

Zjednodušený pohled na BLAST algoritmus:

1. Odstranění oblastí s nízkou komplexivitou nebo repetitivní sekvence v dotazované sekvenci;
    - oblasti, které poskytují falešně vysoké skóre.
2. Vytvoření slovníku obsahujícího klíčová slova – k-mery z dotazované sekvence
    - většinou k = 3 pro sekvenci proteinu a,
    - k = 11 pro nukleidovou sekvenci.
3. Vytvoření seznamu pravděpodobných shod;
    - berou se pouze slova s vysokým skórem,
    - pro proteiny standartně použití skórovací matice (BLOSUM62),

- pro nukleitodové sekvence: match +5, misatch -4, případně match +2, mismatch -3,
- úsek hodnocen jako odpovídající (hit), pokud přesáhne předem definovanou prahovou hodnotu (threshold) T.

4. Uspořádání zbývajících slov s vysokým skóre do efektivního stromu vyhledávání.
5. Hledání shody v databázi pro slova s vysokým skóre.
6. Alignment se rozšiřuje na obě strany, dokud skóre přestane dosahovat prahové hodnoty;

    - užití stále stejné skórující matrix,
    - vytvoření HSP, jako High-skóring segment pair,
    - dále se pracuje HSPs které mají větší skóre než empiricky stanovené skóre S.

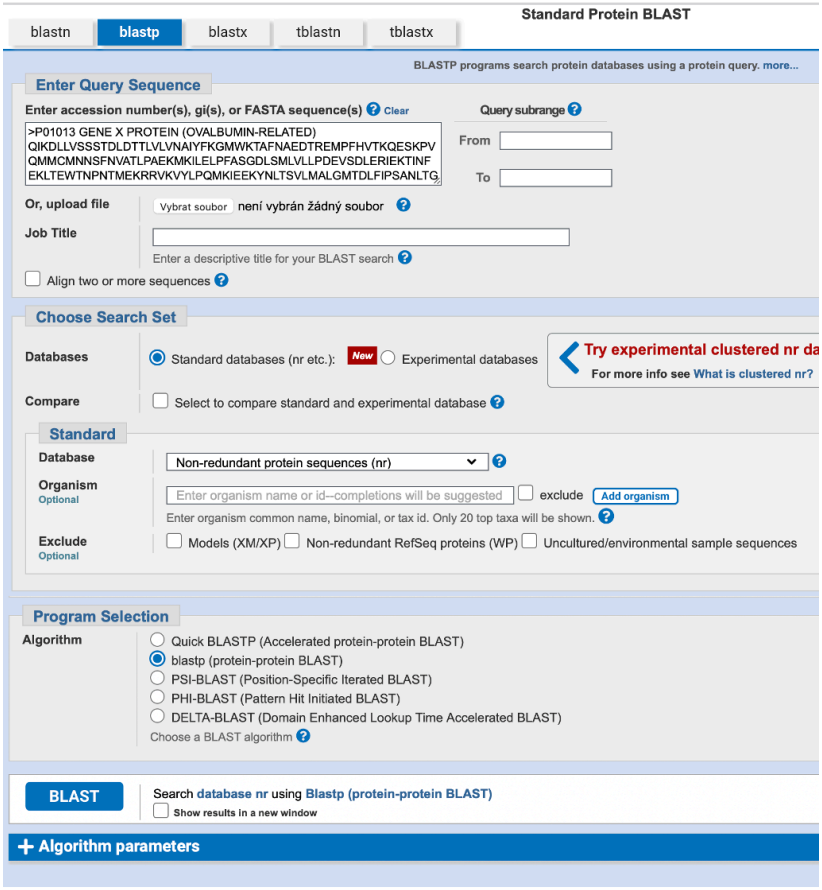
7. Vyhodnocení významnosti HSP skóre;

    - výpočet E-Value,
    - výpočet bit-score.

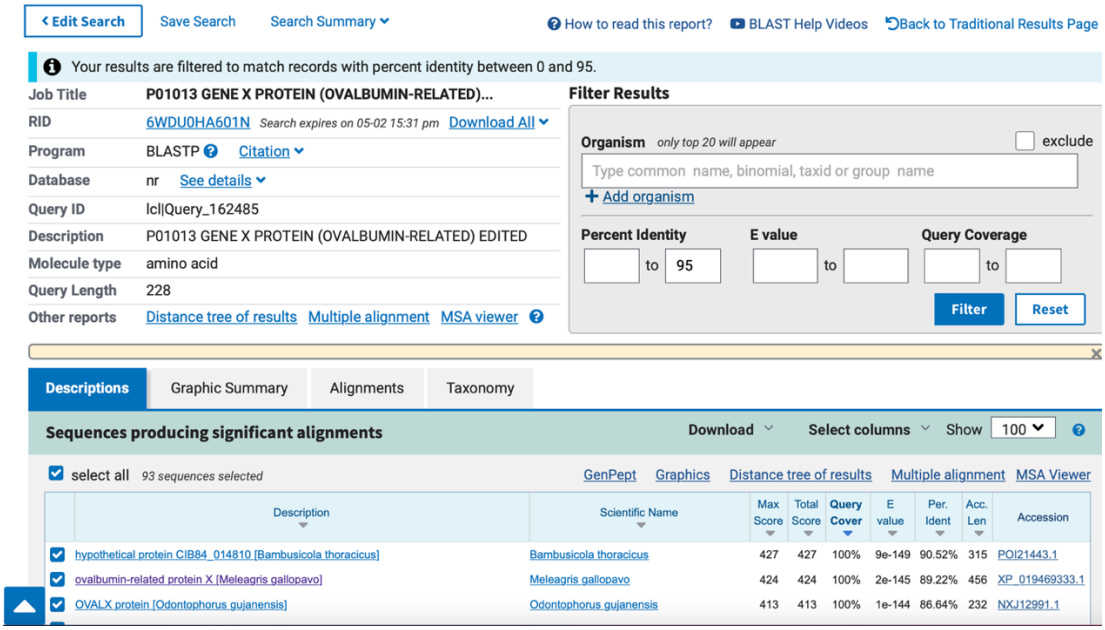
V programu jdou měnit hodnoty pro vyhledávání cílových sekvencí. Běžně jsou potřeba následující parametry:

1. Hodnota hraniční e-val.
2. Délka klíčového slova pro slovník.
3. Skórovací matice (proteiny).
4. Hodnota match/mismatch (nukleotidy).
5. Hodnota vytvoření mezery.
6. Hodnota prodloužení již vytvořené mezery.
7. Minimální skóre pro rozšíření odpovídajícího úseku (hitu).

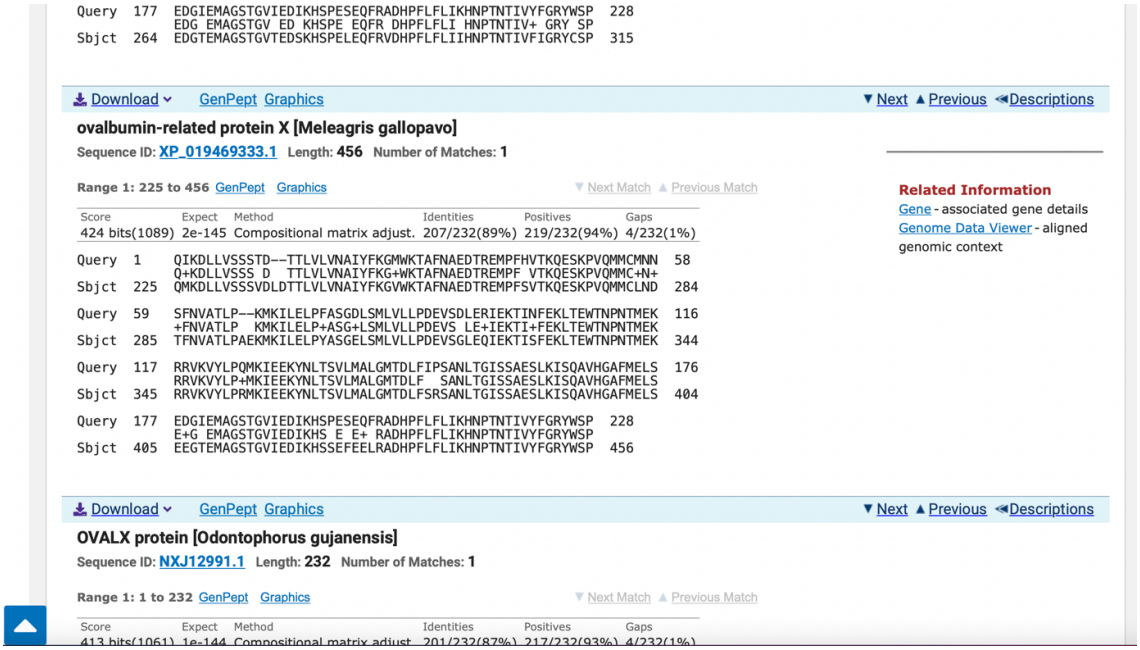
Další vysvětlivky k pojmům používaných nejen v BLASTu (např. i v PAM maticích) najdete zde, https://www.ncbi.nlm.nih.gov/books/NBK62051/, a další užitečné tutoriály, včetně videí pak zde, https://www.ncbi.nlm.nih.gov/home/tutorials/.



Obr. 57: Vyhledávání v BLAStu, aminokyselinová sekvence, NCBI Basic Local Alignment Search Tool. https://blast.ncbi.nlm.nih.gov/Blast.cgi (accessed April 30, 2022)



Obr. 58: Ukázka výsledků vyhledávání aminokyselinového řetězce, výsledek obsahuje maximální skóre, E-value, počet nalezených sekvencí, identifikátory, použitou databázi a jiné, NCBI Basic Local Alignment Search Tool. https://blast.ncbi.nlm.nih.gov/Blast.cgi (accessed April 30, 2022)



Obr. 59: Ukázka výsledků vyhledávání aminokyselinového řetězce, záložka alignment, NCBI Basic Local Alignment Search Tool. https://blast.ncbi.nlm.nih.gov/Blast.cgi (accessed April 30, 2022)

Programy v rámci BLASTu:

Nucleotide-nucleotide BLAST (blastn): podle zadané nukleotidové sekvence vrací nejpodobnější nukleotidové sekvence z nukleotidové databáze (specifikované uživatelem).
Protein-protein BLAST (blastp): podle zadané proteinové sekvence vrací nejpodobnější proteinové sekvence z proteinové databáze (specifikované uživatelem).
Nucleotide 6-frame translation-protein (blastx): prohledává proteinovou databázi pomocí přeloženého nukleotidového dotazu, six-frame translace (oba směry, 3+3).
Nucleotide 6-frame translation-nucleotide 6-frame translation (tblastx): převádí dotazovanou nukleotidovou sekvenci do všech šesti možných čtecích rámců a porovnává ji s překlady šesti rámců databáze nukleotidových sekvencí. Účelem tblastxu je najít velmi vzdálené vztahy mezi nukleotidovými sekvencemi.
Protein-nucleotide 6-frame translation (tblastn): porovnává proteinový dotaz s překlady všech šesti čtecích rámců databáze nukleotidových sekvencí.
Large numbers of query sequences (megablast): zřetězuje mnoho vstupních sekvencí dohromady, aby vytvořily velkou sekvenci před prohledáváním databáze BLAST, a následně analyzují výsledky vyhledávání, aby získaly jednotlivá zarovnání a statistické hodnoty. Je velmi rychlý, nachází efektivně dlouhá zarovnání velmi podobných sekvencí.

8.2.9 PSI-BLAST

Pozičně-Specifický Iterovaný BLAST (PSI-BLAST) (blastpgp) je program k nalezení vzdálených příbuzných proteinu. Hledá senzitivnějším způsobem a je založen na BLAST, ale dále využívá iterace kroků multiple alignmentu a tvorby a použití pozičně specifické substituční matice (PSSM). Nejprve je vytvořen seznam všech úzce příbuzných proteinů. Tyto proteiny jsou sloučeny do obecné profilové sekvence, která shrnuje významné rysy přítomné v těchto sekvencích. Použitím tohoto profilu je poté spuštěn dotaz proti proteinové databázi a je nalezena větší skupina proteinů. Tato větší skupina se používá k vytvoření jiného profilu a proces se opakuje.
Zahrnutím příbuzných proteinů do vyhledávání je PSI-BLAST mnohem citlivější při zachycení vzdálených evolučních vztahů než standardní protein-protein BLAST.

BLAST existuje i ve formě webové aplikace na adrese https://blast.ncbi.nlm.nih.gov/Blast.cgi. Jedinými požadavky na spuštění programu je zadání dotazované sekvence/sekvencí a výběr databáze. BLAST najde v databázi dílčí sekvence, které jsou podobné dílčím sekvencím v dotazu. Podle typu (blastn, blastp, …) se i ve webové verzi BLAST dají měnit některé parametry, jako počet cílových sekvencí k zobrazení, threshold, délka klíčového slova, match/mismach skóre (příp. skórovací matice u proteinů), použití filtrů a masek.

8.2.10 **FASTA**

Alternativní program k programu BLAST od EBI. Na rozdíl od algoritmu BLAST jsou zde tolerovány mezery. FASTA v citlivosti předčí BLAST, ale je výpočetně výrazně náročnější. Výběr možností pro zarovnání v rámci BLAST a FASTA najdeme v Tab. 16.

Vstupní parametry pro FASTA jsou obdobné jako pro BLAST: velikost gap penalty za otevření a prodloužení mezery, zvolená substituční matice a velikost vyhledávaného slova (k-tuple). Nastavení velikosti k-tuple ovlivňuje výslednou rychlost zarovnání, se zvyšující se velikostí slova klesá citlivost (nenalezne všechny záznamy). Úseky o nízké komplexitě, které by mohly znepřesnit výsledky zarovnání, by měly být před vlastním zarovnáním, případně dodatečně, odfiltrovány.

Algoritmus FASTA se skládá z následujících kroků:

1. Umístění porovnávaných sekvencí na vertikální a horizontální ose grafu (obdobně jako u Dot-plotu). Program vyhledá krátké identické sekvence o předem dané délce, které vytvoří diagonály ve skórovací matici.
2. Vybere se předem určený počet nejdelších diagonál a spočítá se nenormalizované skóre alignmentu bez mezer (init1). Dále se vyřadí sekvence, jejichž nejlepší diagonály nedosahující předem dané mezní hodnoty (cutoff), tím dostanemě 1. treshold.
3. Vytvoří se neoptimalizovaný alignment spojením (včetně vzniku mezer) nejlepší diagonály se sousedními, přepočítá se skóre včetně mezer (initn).
4. Alignment je rozšiřován na obě strany pomocí klasického pairwise alignmentu, dokud skóre nepřesáhne zvolenou mezní hodnotu
5. Finální výpočet z-skóre/bit-skóre a E-value.

Programy v rámci FASTA:

FASTA: srovnává DNA (aminokyselinové) sekvence s jinou DNA (proteinovou) sekvencí/databází DNA (proteinových) sekvencí.
SSEARCH: srovnává proteinovou nebo DNA sekvenci s databází, Smith-Waterman bez optimalizací a heuristik (velmi pomalé).
TFASTX/Y: porovnávají proteinové sekvence s DNA sekvencí/databází, translace DNA do proteinové sekvence, včetně posunů čtecího rámce a vyhledávání všemi směry. FASTY je pomalejší, protože povoluje frameshifty v rámci kodonů.
GGSEARCH: global-global search, využívá metodu dynamického programování Needleman-Wunsch pro sestavení globálního alignmentu.
GLSEARCH: global-local search, využívá metodu dynamického programování Needleman-Wunsch pro sestavení globálního alignmentu.

Tab. 16: Přehled programů pro vyhledávání sekvencí na základě podobnosti v sekvenčních databázích. (Cvrčková F, 2016)

	DNA	Protein
DNA	FASTA BLASTN SSEARCH GGSEARCH GLSEARCH	BLASTX
Protein	TBLASTN TFASTX/Y	BLASTP PSI-BLAST FASTA SSEARCH

Webová aplikace FASTA je na stránce https://www.ebi.ac.uk/Tools/sss/fasta/.

8.3 Multiple sequence alignment

Multiple sequence alignment je rozšířením pairwise alignmentu, kdy zarovnáváme několik proteinových či nukleotidových sekvencí. Využití je následující: nalezení „diagnostického vzoru“ (diagnostic patterns) na jehož základě je možné charakterizovat proteinové rodiny, analýza homologie mezi novou sekvencí se sekvencemi v databázích, určení vzájemné příbuznosti sekvencí v rámci skupiny spojené s tvorbou fylogenetických stromů, predikce sekundární a terciární struktury nových proteinů a také navržení primerů či oligonukleotidových sond pro PCR a další technologie.

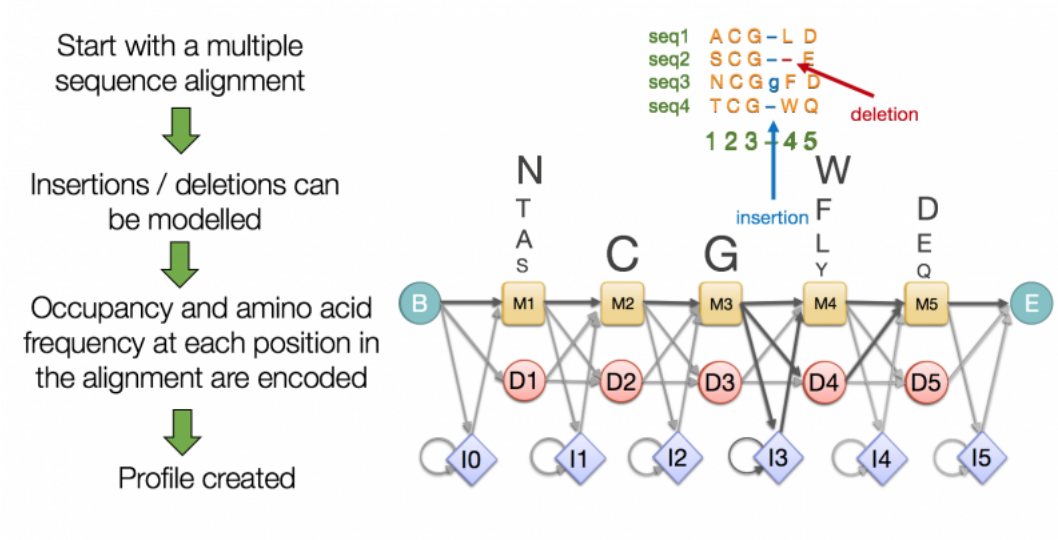
Existuje několik možných metod pro řešení vícenásobného zarovnání:
dynamické programování: přímé rozšíření pairwise alignmentu, simultánní alignment všech sekvencí, silně výpočetně a časově náročné, nevhodné pro více jak 4 sekvence
progresivní alignment: nejčastěji používaný k vytvoření alignmentu; využívá fylogenetické informace – hierarchický, nejdříve identifikuje nejpodobnější sekvence a následně inkorporuje ostatní
iterativní: odstraňuje problémy progresivního alignmentu, který je závislý na prvotním přiložení nejpodobnějších sekvencí pomocí opakovaní alignmentu pro podskupiny sekvencí následující po globálním alignmentu hledání motivů: nalezení částí konzervovaných sekvenčních motivů pomocí globálního přiložení a následně „hodnocení“ těchto úseků nezávisle na celé sekvenci
pravděpodobnostní modely: využití skrytých Markovových modelů

Vzhledem k výpočetní náročnosti (programové řešení je blízko NP kompletních problémů) využívá většina programů pro zarovnání více sekvencí heuristické metody spíše než globální optimalizaci, které ale často neposkytují optimální řešení.

8.3.1 Nástroje

HMMER

HMMER (Obr. 60) se používá k vyhledání homologních sekvencí v sekvenčních databázích. Implementuje metody využívající skryté pravděpodobnostní modely nazývané profilové skryté Markovovy modely (profilové HMM). HMMER umožňuje citlivou detekci vzdálených homologů, u které využívá pravděpodobnostní modely. V nedávné době došlo k upravení výpočetní náročnosti a tím k podstatnému zrychlení (zhruba stejná rychlost jako BLAST). HMMER se často používá společně s databází proteinových profilů, (Pfam), ale obdobně jako BLAST může pracovat se sekvencemi dotazů. Prohledání dotazované proteinové sekvence v databázi může probíhat pomocí phmmer nebo iterativním vyhledáváním pomocí jackhmmer.



Obr. 60: HMM modelování multiple sequence alignmentu, EMBL-EBI Train Online, CC BY-SA 4.0 via Wikimedia Commons

ClustalW

Jednotlivým sekvencím přiřazuje váhy (weight – W) podle četnosti zastoupení (čím více jsou si sekvence podobné, tím nižší mají váhu a naopak) a penalizuje přítomnost mezer v závislosti na jejich pozici (pozičné specifické ceny za mezery), jedna z metod progresivního alignmentu. Zjednodušený algoritmus je následující: pro každou dvojici sekvencí se vytvoří pairwise alignment, následně se sestrojí příbuzenský strom (similarity tree) a v pořadí dle nejbližší příbuznosti se sestaví alignmenty. Jednou vložené mezery v pairwise alignmentu jsou zachovány i v pozdějším sestavení alignmentů.

ClustalOmega

Nejnovější verze programů Clustal. Využívá guide stromu vypočítaného z matice vzdáleností, kterým určuje pořadí jednotlivých párových alignmentů. Vhodný pro středně až dlouhé proteinové alignmenty. Oproti předchozí verzi rychlejší a přesnější.

MUSCLE (MUltiple Sequence Comparison by Log-Expectation)
MUSCLE pro výpočet používá součet výslednách skóre zarovnání párů sekvencí Skóre zarovnání páru sekvencí se vypočítá jako součet skóre substituční matrice pro každý zarovnaný pár zbytků plus penalizace za mezery. Výpočet probíhá ve třech fázích: progresivní návrh, progresivní vylepšení a zpřesňující fáze. Ve fázi progresivního návrhu vytváří algoritmus vícenásobné zarovnání návrhu, přičemž klade důraz na rychlost před přesností. Zjednodušený postup je následující: Sestaví se matice vzdáleností (distance matrix) pro každou dvojici sekvencí, na základě toho je sestaven (první) příbuzenský strom. V pořadí od větví ke kmenu je v každém rozvětvení vytvořen profil, který při dalším porovnávání nahrazuje původní sekvence a vytváří první multiple sequence alignment. Zde dochází k přepočítání vzdáleností a tvorbě druhého příbuzenského stromu a druhého zarovnání. Vzniklý stromu se rozdělí na dvě části a pro každou z nich se vytvoří vícenásobné zarovnání. Pokud je výsledný alignment lepší, je zachován. Toto se iterativně opakuje do konvergence nebo do určeného počtu kroků. MUSCLE dokáže podle zvolených podmínek dosáhnout jak lepší průměrné přesnosti, tak lepší rychlosti než CLUSTALW nebo T-Coffee.

T-Coffee
Pomalejší ale výrazně přesnější než ClustalW. Kombinuje data z více předchozích alignmentů, které mohly být vytvořeny různými postupy (lokální, globální, strukturní podobnost, ...). Místo substituční matice použité v klasickému progresivnímu alignmentu se využívá pozičně specifické skórovací schéma (extended library).

8.4 Otázky k tématu

1. Popište zjednodušeně globální a lokální alignment. Jaké algoritmy se pro ně používají? Co je gene alignment?
2. Co je principem alignmentu, proč se dělá?
3. Zjednodušeně popište princip a použití BLAST a vysvětlete E-value.
4. Vysvětlete rozdíl mezi BLAST a PSI-BLAST.
5. Na [této stránce, https://www.uniprot.org/uniprot/A8DDH8#sequences,](https://www.uniprot.org/uniprot/A8DDH8#sequences) si stáhněte sekvenci ve formátu fasta a vyzkoušejte si zarovnání na webové aplikaci BLAST.

8.5 Zdroje

Výpočet skóre a substituční matice

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol*. 1990;215(3):403-410. doi:10.1016/S0022-2836(05)80360-2

BLAST® Command Line Applications User Manual [Internet]., 2020. National Center for Biotechnology. https://www.ncbi.nlm.nih.gov/books/NBK279684/ (accessed Dec 28, 2020).

BLAST® Command Line Applications User Manual, 2020. National Center for Biotechnology. https://www.ncbi.nlm.nih.gov/books/NBK279690/ (accessed Dec 28, 2020).

Compeau, P., & Pevzner, P. *Bioinformatics algorithms: an active learning approach*, Vol. 1, 2nd ed.; La Jolla, California: Active Learning Publishers, 2015.

Cvrčková, F. *Úvod do praktické bioinformatiky*, 1st ed.; Academia: Praha, 2006.

J. Pačes. Alignment; přednáška pro VŠCHT Praha 2020

Kellis, M. *Computational Biology - Genomes, Networks, and Evolution* [online]; MIT OpenCourseWare, 2020. https://bio.libretexts.org/Bookshelves/Computational_Biology/Book%3A_Computational_Biology_-_Genomes_Networks_and_Evolution_(Kellis_et_al.) (accessed Dec 27, 2020).

Mäkinen, V., Belazzougui, D., Cunial, F., & Tomescu, A. I. *Genome-scale algorithm design*. Cambridge University Press, 2015.

Pearson, W. R., & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*, *85*(8), 2444–2448. https://doi.org/10.1073/pnas.85.8.2444

Pairwise alignment

BLAST® Command Line Applications User Manual [Internet]., 2020. National Center for Biotechnology. https://www.ncbi.nlm.nih.gov/books/NBK279684/ (accessed Dec 28, 2020).

BLAST® Command Line Applications User Manual, 2020. National Center for Biotechnology. https://www.ncbi.nlm.nih.gov/books/NBK279690/ (accessed Dec 28, 2020).

Cvrčková, F. Úvod do praktické bioinformatiky, 1st ed.; Academia: Praha, 2006

Kellis, M. *Computational Biology - Genomes, Networks, and Evolution* [online]; MIT OpenCourseWare, 2020. https://bio.libretexts.org/Bookshelves/Computational_Biology/Book%3A_Computational_Biology_-_Genomes_Networks_and_Evolution_(Kellis_et_al.) (accessed Dec 27, 2020).

Mäkinen, V., Belazzougui, D., Cunial, F., & Tomescu, A. I. *Genome-scale algorithm design*. Cambridge University Press, 2015.

Multiple sequence alignment

Eddy, S. What is a hidden Markov model?. *Nat Biotechnol* 22, 1315–1316 (2004). https://doi.org/10.1038/nbt1004-1315
J. Pačes. Multiple Alignment; přednáška pro VŠCHT Praha 2020
Morgenstern B. DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics. 1999 Mar;15(3):211-8. doi: 10.1093/bioinformatics/15.3.211. PMID: 10222408.
Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol. 2000 Sep 8;302(1):205-17. doi: 10.1006/jmbi.2000.4042. PMID: 10964570.
Sievers F, Higgins DG. Clustal Omega for making accurate alignments of many protein sequences. Protein Sci. 2018 Jan;27(1):135-145. doi: 10.1002/pro.3290. Epub 2017 Oct 30. PMID: 28884485; PMCID: PMC5734385.
Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004 Mar 19;32(5):1792-7. doi: 10.1093/nar/gkh340. PMID: 15034147; PMCID: PMC390337.

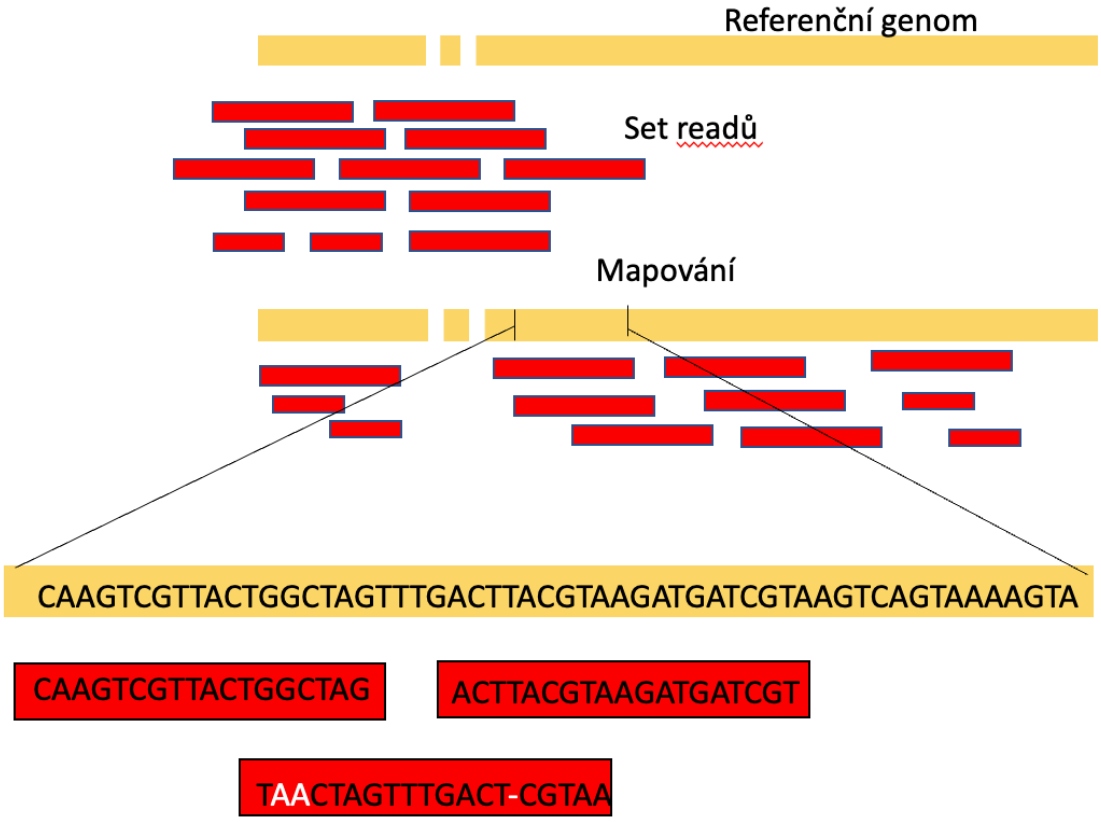
9 MAPOVÁNÍ A ASSEMBLY

NGS metody sekvenování (viz kapitola Sekvenování) generují velké množství krátkých fragmentů bez genomického kontextu. Pokud byl referenční genom pro daný organismus již sestaven, můžeme namapovat jednotlivé ready na referenční sekvenci. V opačném případě je možné využít vysoce příbuzného druhu jako nápovědu a namapovat sekvence na něj, anebo využít de-novo assemblerů.

Pro přiřazení segmentů na referenční genom by bylo možné použít např. BLAST, ale bylo by to nesmírně výpočetně a časově náročné, zarovnání milionů krátkých sekvencí tímto způsobem by trvalo týdny. Další možností by bylo využít multiple sequence alignment, ale vzhledem k tomu, že každá oblast je pokryta několika ready, by tento přístup vygeneroval enormní množství semi-globálních alignmentů. Nehledě na to, že cílem alignmentu je najít přesnou pozici a rozdíl, kdežto zde nám stačí najít přibližnou pozici. Tím pádem se využívá specifických algoritmů a programů, tzv. mapování.

9.1 Mapování

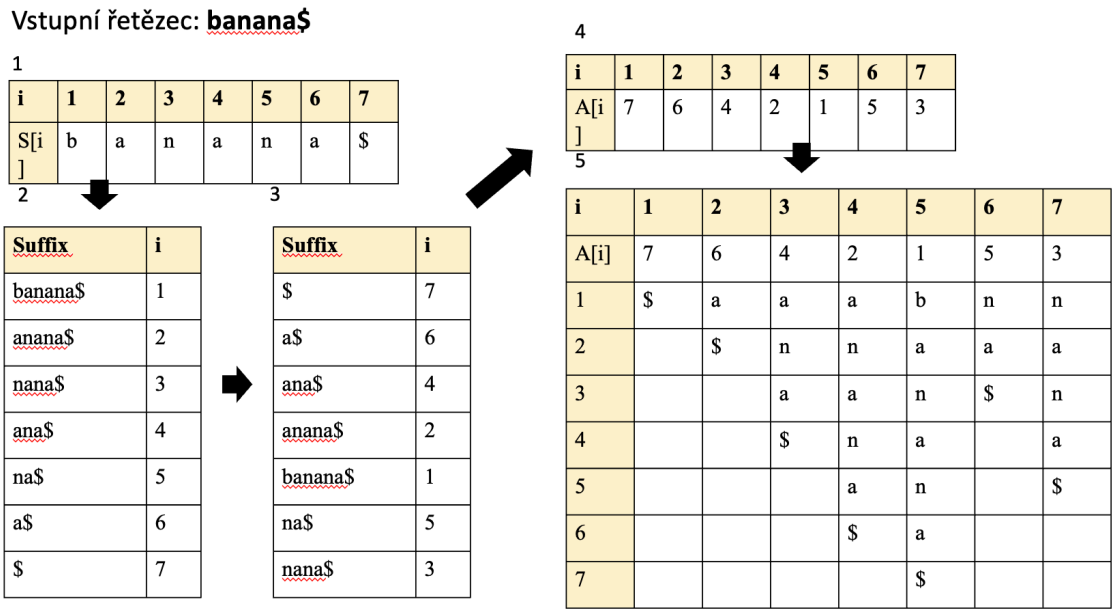
Mapování je proces zarovnání jednotlivých readů na referenční genom (Obr. 61), jedná se o jeden z klíčových kroků v moderní analýze genomických dat. Pomocí mapování jsou ready přiřazeny ke konkrétnímu místu v genomu, což umožňuje získat informace (podle použité techniky) např. o sekvenačních variantách, genové expresi, otevřenosti chromatinu a dalších. Protože jsou naše čtení krátká, může být v referenční sekvenci několik stejně pravděpodobných míst, ze kterých mohly být přečteny, zejména v případě repetitivních oblastí.



Obr. 61: Vizualizace mapování na referenční sekvenci, Pfeiferová L

9.1.1 Náhled na algoritmy používané v mapování

Vzhledem k velikosti referenční sekvence a všech čtení, není možné pracovat s přímým textovým vyhledáváním jednotlivých čtení v rámci sekvence. Proto se využívá indexování. Indexování genomu si lze představit podobně jako indexování knihy. Pokud chceme vědět, na které stránce se určité slovo vyskytuje nebo kde začíná kapitola, je mnohem efektivnější vyhledat si je v předem vytvořeném rejstříku než procházet každou stránku knihy, dokud ji nenajdeme. Totéž platí pro mapování. Indexy umožňují maperu zúžit potenciální původ dotazované sekvence v genomu, čímž šetří čas i paměť. Příkladem takového indexu je suffix array – seřazená pole (array) všech přípon (suffix) řetězce. Definice suffix array je následující: máme n-stringový řetez S[1]S[2]S[3]…S[n] a substring S od i do j. Suffix array A řetězce S je pole celých čísel poskytujících počáteční pozice přípon S v lexikografickém pořadí. To znamená, že A[i] je počáteční pozice i-tého nejmenšího suffixu v S. Každý suffix S se v A objeví právě jednou. Suffixy jsou jednoduché řetězce, které se třídí, než se jejich počáteční pozice (indexy celých čísel) uloží do A (Obr. 62).



i	1	2	3	4	5	6	7
S[i]	b	a	n	a	n	a	$

Suffix	i
banana$	1
anana$	2
nana$	3
ana$	4
na$	5
a$	6
$	7

Suffix	i
$	7
a$	6
ana$	4
anana$	2
banana$	1
na$	5
nana$	3

i	1	2	3	4	5	6	7
A[i]	7	6	4	2	1	5	3

i	1	2	3	4	5	6	7
A[i]	7	6	4	2	1	5	3
1	$	a	a	a	b	n	n
2		$	n	n	a	a	a
3			a	a	n	$	n
4			$	n	a		a
5				a	n		$
6				$	a		
7					$		

Obr. 62: Tvorba suffix array, 1. indexujeme text „banana“, $ je znak, který je lexikograficky menší než všechny ostatní znaky, označuje konec sekvence, 2. vytvoření suffixů, 3. setřídění suffixů podle vsestupného pořadí, 4. Suffix array A obsahuje počáteční pozice seřazených suffixů, 5. suffix array se suffixy sepsanými přehledně do tabulky, A[3] obsahuje hodnotu 4, která odkazuje na suffix začínající na pozici 4, což je suffix ana$

Vztah suffix array, suffix tree, suffix trie

Trie je nejjednodušší stromová datová struktura používaná k vyhledání konkrétních klíčů ze sady. Tyto klíče jsou nejčastěji řetězce, přičemž vazby mezi uzly nejsou definovány celým klíčem, ale jednotlivými znaky. Aby bylo možné získat přístup ke klíči (obnovit jeho hodnotu, změnit ji nebo odstranit), je trie nejprve procházeno hloubkou, podle vazeb mezi uzly, které představují každý znak v klíči. **Suffix tree** je vylepšením oproti trie (má příponové odkazy, které umožňují lineární vyhledávání chyb, příponový strom ořezává zbytečné větve trie, takže nevyžaduje tolik místa). **Suffix array** je zkrácená datová struktura založená na stromu přípon (žádné odkazy na přípony (pomalé shody chyb), přesto je vyhledávání vzorů velmi rychlé).

Ačkoli suffix array umožňují kompresi dat a efektivnější vyhledávání, v rámci genomů to pořád není dostačující, proto se využívá Burrows-Wheelerova transformace (BWT, Obr. 63). Pro představu, velikost suffix array pro lidský genom, který má zhruba 3.2×10^9 párů bazí, by v 64bit integer systému zabíral přibližně 12 GB, suffix tree dokonce 45 GB. Velikost indexu vytvořeného nástrojem Bowtie2, založeném na upraveném BWT indexaci, zabírá 2.2 GB. BWT lze popsat jako permutaci řetězce T postupným skenováním jeho suffix arrays a zaznamenáním znaku, který předchází každé pozici v suffix array.

Vstupní řetězec: **banana$**

0	B	A	N	A	N	A	$
1	A	N	A	N	A	$	B
2	N	A	N	A	$	B	A
3	A	N	A	$	B	A	N
4	N	A	$	B	A	N	A
5	A	$	B	A	N	A	N
6	$	B	A	N	A	N	A

Rotační matice

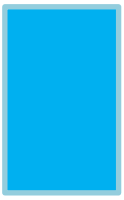
0	A	N	A	N	A	$	B
1	A	N	A	$	B	A	N
2	A	$	B	A	N	A	N
3	B	A	N	A	N	A	$
4	N	A	N	A	$	B	A
5	N	A	$	B	A	N	A
6	$	B	A	N	A	N	A

Seřazená matice

BWT output: **BNN$AAA**

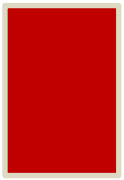
Obr. 63: První sloupec ordered matrix (seřazené matici) se označuje jako F, obsahuje seřazené znaky výchozího řetězce. Poslední sloupec se označuje jako L, obsahuje BWT výchozího řetězce. Ze sloupce L se dá získat původní řetězec

BWT je reversibilní, k původnímu řetězci se můžeme dostat na základě LF mapování, což je funkce, která mapuje z posledního sloupce BWT do prvního sloupce pomocí čísla řádku a počtu výskytu znaku v řetězci.



BWT v podstatě slouží jako bezztrátový slovník čtení, na rozdíl od výsledků heuristického a ztrátového mapování čtení do genomu běžně získaných v prvním kroku sekvenční analýzy. Z BWT se dá získat i suffix array.

Na BWT (a suffix array) je založený FM index (Full-text index in Minute space). Lze jej použít k efektivnímu nalezení počtu výskytů vzorku v komprimovaném textu a také k nalezení pozice každého výskytu. FM index s backtrackingem byl úspěšně použit pro přibližnou shodu řetězců/zarovnání sekvencí. FM index využívá Bowtie.



Suffixy, BWT i FM index jsou exact match. Vzhledem k tomu, že se zajímáme hlavně o rozdíly mezi referenčním a naším genomem (navíc existují diploidní, multiploidní i polyploidní organismy), není možné použít exact pattern matching, ale indexed approximate pattern matching.

Výborným studijním materiálem pro porozumění algoritmům, BWT, BWT backtrackingu, vztahu suffix array a BWT a ostatnímu v mapovacích algoritmech jsou přednášky Bena Langmeada, volně dostupné na https://www.youtube.com/user/BenLangmead

9.1.2 Nástroje pro mapování

Pro mapování (na referenční sekvenci) jsou určeny následující nástroje:

BWA: aligner na principu Burrows-Wheeler transformace s možností volby metody tvorby indexu (FM/BWT). Tři algoritmů dle povahy readů – BWA-backtrack, BWA-SW a BWA-MEM. BWA-MEM je novější a je společností více doporučován. BWA-SW má vyšší citlivost v případě vysoké frekvence mezer v alignmentu.
Bowtie, Bowtie2: rychlé alignery vhodný pro ready o délce stovek až tisíců bp na relativně dlouhé referenční genomy. Pro indexaci genomu využívá FM index. Bowtie je stále považován za nejlepší aligner pro krátké sekvence miRNA.
HISAT2: využívá grafového FM indexu.
STAR: specificky vyvinut pro RNA-seq. Využívá indexace suffix array a konceptu „maximálního mapovatelného prefixu“ (MMP).
GSNAP: určen na detekci variant a sestřihu z krátkých čtení
(A pseudoaligner Salmon, viz níže v kapitole Analýza genové exprese, Nástroje pro získání count table.)

9.1.3 Identifikace variant ze sekvenačních dat

Referenční sekvence reprezentuje (či spíše by měla reprezentovat) nejčastěji se vyskytující varianty v rámci populace (pokud je osekvenovaný dostatečný počet jedinců daného druhu/ údaje o frekvenci v rámci populace jsou k dispozici). Při mapování na referenční genom nám zajímají odchylky od referenční sekvence. Proces identifikace variant z referenční sekvence se nazývá variant calling. Kromě identifikace germinálních (zděděných po rodičích) variant se dá mapování přenést i na problém identifikace somatických (získaných) variant. Ty se mohou vyskytovat v kterékoli buňce těla kromě zárodečných (spermie a vajíčka), a proto se nepředávají dále v potomstvu. Tyto změny mohou (ale ne vždy) způsobit rakovinu nebo jiná onemocnění.

Chyby sekvenování: Sekvenátor může provést chybné volání buď z fyzikálních důvodů, nebo kvůli vlastnostem sekvenované DNA (např. homopolymery). Protože chyby sekvenování jsou často náhodné, lze je odfiltrovat jako jednotlivé čtení během volání variant.
Chyby mapování: Algoritmus mapování může mapovat čtení na nesprávné místo v referenci. To se často děje v oblasti repetic nebo v jiných oblastí s nízkou komplexitou.

Aby byla daná změna vyhodnocena jako varianta, musí být pokryta (coverage) určitým počtem readů. Pod tímto minimálním počtem readů je změna vyhodnocena jako sekvenační chyba. Běžně se také pozoruje kvalita dané záměny, zda se nenachází pouze v jednom směru (u párových čtení) a celková coverage. Vhodné číslo se udává spíše intuitivně, i když může napomoci znalost accuracy a single error rate pro použité metody. Udává se však, že minimálně kvalita varianty by však měla být 20, frekvence 25 a count 5.

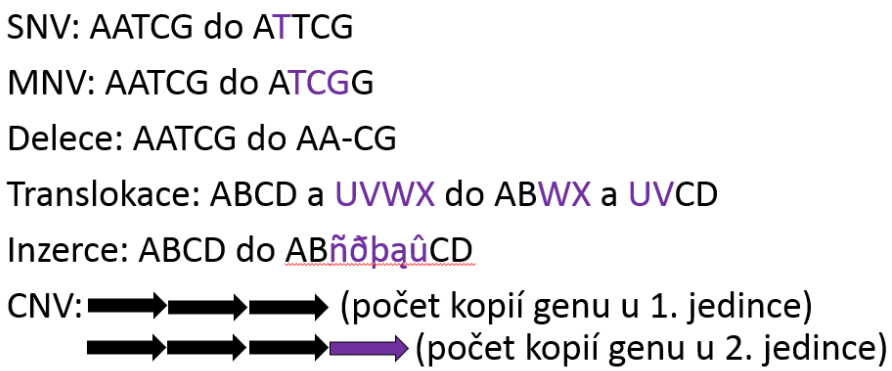
Strukturní varianty (Obr. 64)

SNP (single nucleotide polymorphism, jednonukleotidový polymorfismus): variace v jediném nukleotidu, která se vyskytuje na specifické pozici v genomu; každá taková variace je přítomna v populaci v určité pravděpodobnosti.
MNV (multi-nucleotide polymorphism, vícenukleotidový polymorfismus)
VNTR (Variable Number of Tandem Repeat polymorphism): odlišný počet kopií tandemových repetitivních jednotek v rámci homologních chromosomů.
STR (Short Tandem Repeats, mikrosatelity): 2-16 nukleotidové repetice s vysokou variabilitou opakování, sekvence spolu přímo sousedí, přímé využití ve forenzní analýze a určení otcovství, kdy se sestavuje profil 20 specificky určených STR míst v jednom vzorku a porovnává s druhým vzorkem
CNV (copy number variation): opakování části genomu, počet opakování v genomu se u jednotlivců liší
Delece: vynechání části chromosomu nebo sekvence DNA
Inzerce: přidání jednoho nebo více párů nukleotidových bází do sekvence
Duplikace: duplikace genu/chromosomu
Rekombinance: změny v sekvenci spočívající v jejím rozštípnutí a následném připojení k jinému řetězci
Translokace: přeskupení chromosomů
Chromosomální inverze: segment chromosomu je obrácen od jednoho konce k druhému

Vlivem mutací/strukturních variant v genové sekvenci může dojít k následujícím efektům v proteinovém řetězci.

Nonsense: vznik předčasného stop kodonu v transkribované mRNA, a tedy zkrácený, neúplný a obvykle méně funkční či úplně nefunkční proteinový produkt
Missense: změna v jednom nukleotidu vede ke kodonu, který kóduje jinou aminokyselinu
Conservative: náhrada aminokyseliny v proteinu, který mění danou aminokyselinu na jinou aminokyselinu s podobnými biochemickými vlastnostmi
Silent: bez následné změny v aminokyselině nebo funkci celkového proteinu (obvykle, může vyvolat fenotypickou změnu, např. změna rychlosti exprese, změna rychlosti translace)
Frameshift: typ mutace zahrnující inzerci nebo deleci nukleotidu, ve kterém počet odstraněných/vložených bází není dělitelný třemi, dochází k posunu čtecího rámce

Popsané efekty mohou mít různé stupně závažnosti (impakt) – low, mediocre a high. High impakt efekty často bývají u nonsense a frameshift, low většinou u synonymních mutací. Pro odhad efektu mutací a jejich impaktů se používají prediktory – variant effect predictors, zkráceně VEPs. Zlatým standardem u nástrojů je SnpEff (jednoduchá instalace) a VEP od Ensembl (dostupný také v online verzi). Odhad impaktu pouze ze sekvence a databáze může být někdy nepřesný, i proto se začínají objevovat prediktory na bázi strojového učení.



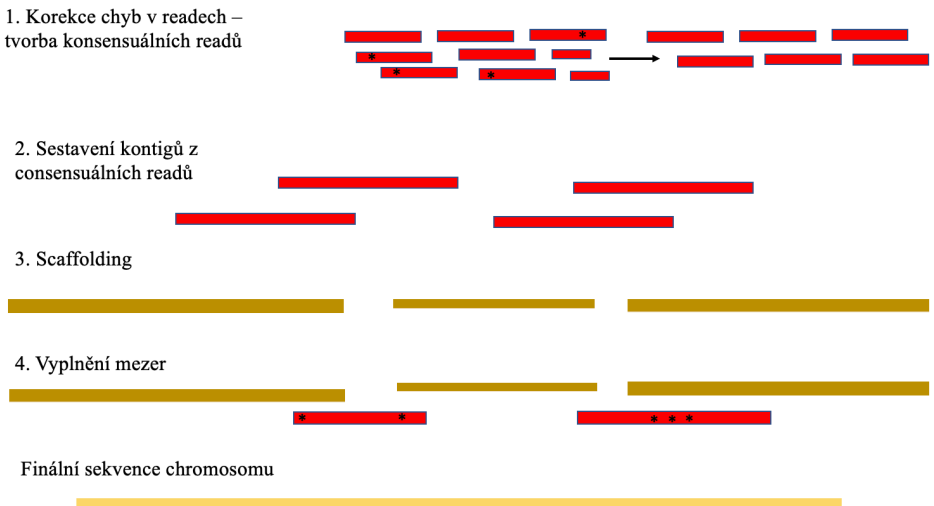
Obr. 64: Možné strukturní varianty, vlastní data

Nástroje pro variant calling

GATK (Genome Analysis Toolkit): soubor nástrojů od Broad Institute s vlastními pipelinami a workflow, obsahuje nástroje na detekci germinálních variant, včetně CNV, v DNA a RNA-Seq datech, zaměřený na člověka
Mutect2: nástroj pod správou GATK zaměřený pouze na detekci somatických variant u člověka
MuSE: nástroj zaměřený na detekci somatických variant u člověka u heterogeních nádorových vzorků. Vyžaduje zpracování dat podle jimi stanovené workflow s využitím (nejenom) GATK.
Freebayes: „a haplotype-based variant detector“, nástroj zaměřený na detekci malých polymorfismů jako SNP, indely, MNP a inserce kratší, než je délka short-read sekvenčního alignmentu, pro každý organismus, ke kterému je referenční sekvence ve formátu FASTA
Strelka2: „Small Variant Caller“, rychlý nástroj optimalizovaný na germinální varianty a na somatické v tumor/normálních párech.
VarScan: nástroj pod Genome Institute at Washington University na detekci variant v NGS datech, zaměřuje se na detekce germinálních varianty (SNP a krátké indely), variant ve více vzorcích, somatických mutací v tumor-normal párech a somatických CNV v tumor-normal vzorcích.

9.2 De-novo assembly

De-novo assembly je metoda pro konstrukci genomových sekvencí z velkého počtu (krátkých nebo dlouhých) fragmentů DNA bez znalosti správné sekvence nebo pořadí těchto fragmentů. Postup de-novo assembly se skládá z následujících kroků: korekce chyb, sestavení kontigů, scaffolding a vyplnění mezer (Obr. 65).



Obr. 65: Vizualizace de-novo assembly, Pfeiferová L

Problematiku de-novo assembly si můžeme představit na následujícím zjednodušeném příkladu. Máme sekvenci T u které známe frekvence k-merů, které se v ní vyskytují. Tím pádem můžeme sestrojit de Bruijnův graf (orientovaný graf reprezentovaný overlapy mezi sekvencemi ve kterém pomocí Eulerovy cesty (každou hranu navštívíme právě jednou, problematiku lze vysvětlit na problému „7 mostů v Königsbergu“) najdeme finální sekvenci (Obr.66).

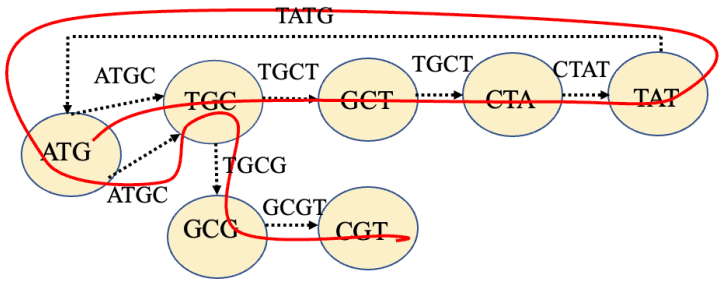
(Neznámá) Finální sekvence:
ATGCTATGCGT

Read 1: ATGCTA

Read 2: CTATGC

Read 3: ATGCGT

Frekvence k-merů (k=4): ATGC (2x), TGCT, GCTA, CTAT, TATG, TGCG, GCGT



Obr. 66: de Bruijnův graf, pro k=4

Při větším počtu k-merů by se nejenom mohlo stát, že nenajdeme pouze 1 řešení Eulerovy cesty, ale že tento problém ani nebude mít řešení v dosažitelném čase. Místo toho se tedy pokoušíme najít co nejmenší set co nejdelších drah – kontigů, v daném assembly grafu. Charakteristickou vlastnosti assembly grafu je, že cestu z vrcholu 1 do vrcholu 2 lze interpretovat jako podřetězec genomu získaný sloučením dílčích řetězců reprezentovaných vrcholy na cestě.

9.2.1 Odstranění chyb v readech

K-mery v de Bruijnově grafu (ready v assembly grafu) mohou obsahovat sekvenační chyby. Ty se mohou projevit jako tip – krátká odbočka v rámci hlavní cesty, nebo jako bubliny (bubbles). Tipů se dá většinou jednoduše zbavit tím, že nastavíme minimální treshold pro délku cesty. U bubbles je situace složitější v tom, že se může jednat o chybové bubbles nebo redundantní, kdy k jejich tvorbě došlo jednoduše tím, že sekvenovaný jedinec měl v daném místě heterozygotní alelu (v případě diploidního organismu). Řešení je nicméně pro oba případy stejné, vybere se jedna cesta (náhodně, nebo podle počtu readů, které ji podporují) a druhý záznam se uloží. Bubbles se také mohou vyskytnout u sekvenčně podobných oblastí.

9.2.2 Sestavení kontigů

Vzhledem k tomu, že počítáme s párovým čtením, a pracujeme s 2 DNA vlákny, máme 8 možností, jak skládat ready do kontigů. Máme read R a S, které se překrývají, respektive ready Ř, Š pro reverzně komplementární vlákno. Pak ready do kontigů můžeme skládat následovně:

(i) Suffix R překrývá prefix S
(ii) Suffix Š překrývá prefix Ř
(iii) Suffix R překrývá prefix Š
(iv) Suffix S překrývá prefix Ř
(v) Suffix Ř překrývá prefix S
(vi) Suffix Š překrývá prefix R
(vii) Suffix Ř překrývá prefix Š
(viii) Suffix S překrývá prefix R

V rámci kontigů je dále zapotřebí zbavit se redundantních readů, které neposkytují žádnou novou informaci (pomocí backtrackingu přes bidirectional BWT index).

9.2.3 Scaffolding

Po vyřešení vzájemné orientace readů nastává obdobný problém ve vzájemné orientaci a vzdálenosti mezi kontigy (NP hard problém).

9.2.4 Gap filling

V rámci de-novo assembly je poměrně jisté, že se nepodaří vytvořit větší sekvenci bez chybějících míst – gaps. Za předpokladu, že máme set readů, které nesedí do žádného z vytvořených kontigů, můžeme vytvořit pomocný overlap graf, ve kterém se budeme snažit nalézt co nejdelší suffix-prefix overlap mezi dvěma nezabudovanými ready (včetně povolení určitého počtu chyb), a tím finálně sestavit kontigy do chromosomů.

V současné době je ale mnohem jednoduší na chybějící části využít metody 3 generace, které produkují mnohem delší ready než NGS metody, a tím tyto části vyplnit.

9.2.5 Overlap-Layout-Consensus

Místo de Bruijnových grafů lze využít Overlap-Layout-Consensus (OLC) přístup. Zjednodušeně pracuje následovně: algortmu se poskytnou vygenerovaná čtení, u kterých jsou identifikovány překrývající se regiony (overlap). Každý read je v grafu vyznačen jako uzel a překryvy jsou reprezentovány jako hrany spojující dva zúčastněné uzly. Algoritmus určí nejlepší cestu podle Hamiltonova principu (každý vrchol/uzel projde právě jednou, pro připomínku Euler prochází každou hranou právě jednou). Hrany a uzle, které nebyly použity (redundantní informace) jsou vyřazeny. Tento proces se iterativně opakuje, výsledné sekvence z každého procesu jsou kombinovány za vytvoření finální konsensuální sekvence (finální genomové sekvence).

9.2.6 Assemblery

Assemblery jsou v drtivé většině založené na de Bruinových grafech – Velvet (balíček algoritmů pro assembly krátkých readů), AbySS (assembler krátkých readů pro sestavení „jakkoli“ velkých genomy, tedy i větších, než je lidský) a SOAPdenovo (short-read assembler pro genomy o přibližné velikosti lidského genomu).

Důvod, proč jsou de Bruijnovy grafy používány místo OLC přístupu je v přístupu ve zpracování redundantních informací z NGS dat (vysoké pokrytí stejných oblastí): vícenásobné překrývání mezi ready zvyšuje frekvenci každého k-meru v překryvu, ale ne počet uzlů v grafu, tudíž nedochází k navýšení výpočetní náročnosti. Nicméně OLC assemblery jsou důležité z historického hlediska.

Algoritmus Celery, která byla použita k sestavení genomu octomilky a částečně při prvních lidských genomech vychází z OLC. Dalším již zastaralých assemblerem je Newbler, také vycházející z OLC principu, který se používal pro ready vzniklé technologií 454. Pomocí Newbleru bylo sestaveno mnoho genů, nyní je však technologicky zastaralý. Výjimkou, která se stále využívá (a používá OLC) je RepeatExplorer, nástroj pro detekci a charaktrizaci repetitivních oblastí v eukaryotických organismech (nicméně obsahuje části silně zaměřené výhradně na rostliny).

9.3 Anotace genomu

Jako anotaci genomu (Obr. 67) označujeme identifikaci interakčních míst DNA s proteiny a RNA, identifikace funkčních oblastí, identifikace hranic intron-exon, protein kódujících částí a genových produktů, regulačních oblastí (viz Encode) a dalších částí.

Anotace genomu se skládá ze 3 základních bodů:

1. Identifikace míst, které *nekódující* proteiny
2. Identifikace možných protein kódujících oblastí
3. Ověření/přidání biologické informace do těchto oblastí

Strukturní anotace obsahuje následující elementy: lokalizace ORFs (open reading frames), genová struktura, kódující oblasti a umístění regulačních motivů. Funkční anotace se skládá z doplňujících biologicky relevantní informací jako je biochemická či biologická funkce, zapojení v regulaci nebo interakcích a exprese.

Jednoduchou metodu anotace představuje BLAST, kdy můžeme najít podobné místo v jiném organismu, a tímto zjistit možnou funkci.

Na rozšíření anotace genomu mají obří podíl velké sekvenovací projekty jako 100,000 Genomes a hlavně ENCODE (popsané výše).

9.3.1 Metody identifikace genů

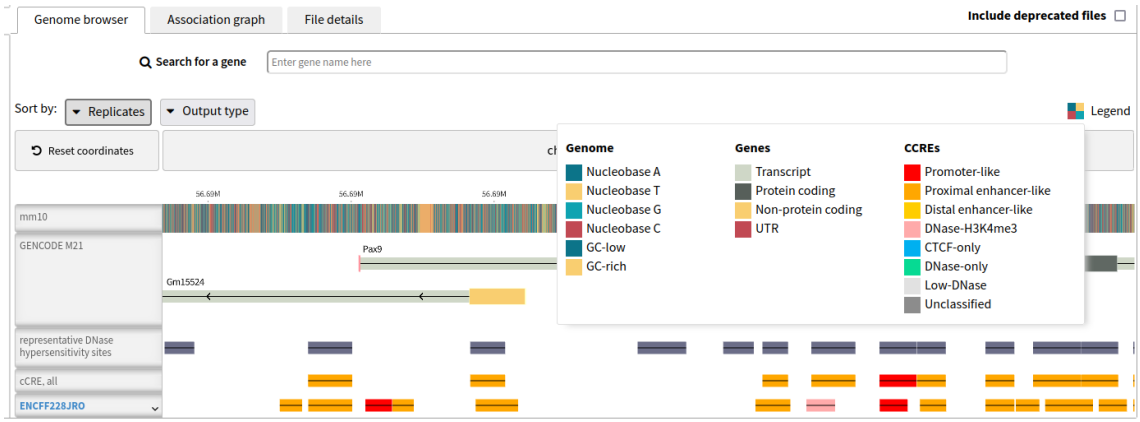
GeneMark: predikce založená na Markovových procesech, předpokládáme, že dlouhé ORF sekvence jsou geny (trénovací dataset). Pak použijeme Markovův model pro ohodnocení pravděpodobnosti, zda daná sekvence odpovídá modelu a vybereme sekvenci s nejvyšší pravděpodobností. Markovův model je sestavován z již anotovaných sekvencí genů.
Prodigal: nástroj založený na dynamickém modelování
BLAST: identifikace homologních oblastí, předpokládáme, že evoluce protein kódujících oblastí je pomalejší, než evoluce nekódujících oblastí.
CEM a **TWINSCAN**: komparativní metody, DNA sekvence je porovnána s podobnou DNA sekvencí z evolučně blízkého organismu. CEM využívá konzervovaných exomových oblastí TWINSCAN Markovovy procesy.

9.3.2 Identifikace nekódujích oblastí

Identifikace nekódujích oblastí je složitější, neboť neobsahují ORF místa a nedá se na ně s úspěchem využít sekvenční podobnosti. Mitochondrie a chloroplasty navíc ukazují nemendelovskou dědičnost. Jedním z nástrojů na identifikaci lncRNA je např. **Lncident**, který využívá vlastnosti vnitřního složení sekvence a informace o otevřeném čtecím rámci založené na modelu podpůrného vektorového stroje.

9.3.3 Identifikace proteinových domén

Identifikace proteinových domén může probíhat např. pomocí strukturního alignmentu, vysoce konzervovaných proteinových oblastí nebo podle Markovových modelů (Pfam kupříkladu využívá jak Markovovy modely, tak multiple sequence alignment).



Obr. 67: Zobrazení anotací v Encode, https://www.encodeproject.org/annotations/ENCSR394RWS/ (accessed April 30, 2022)

9.4 Otázky k tématu

1. Co se dá zjistit pomocí mapování? Jaký je rozdíl mapování a assembly?
2. Co značí v assembly pomocí de Bruijnových grafů bubliny? Bylo by rozumné odstranit vždy všechny bubliny?
3. Jaký je (zjednodušený) postup v de novo assembly?
4. Co je genomová anotace a k jejím se váže velkým sekvenačním projektům?

9.5 Zdroje

Mapování

Compeau, P.; Pevzner, P. *Bioinformatics algorithms: an active learning approach*, 2nd ed.; Active Learning Publishers: California, 2015.

Dokumentace BWA: http://bio-bwa.sourceforge.net/ (accessed April 30, 2022).

Dokumentace Bowtie2: http://bowtie-bio.sourceforge.net/bowtie2/manual.shtml (accessed April 30, 2022).

Dokumentace HISAT2: http://daehwankimlab.github.io/hisat2/manual/ (accessed April 30, 2022).

Dokumentace Salmon: https://salmon.readthedocs.io/en/latest/salmon.html (accessed Dec 28, 2020).

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013 Jan 1;29(1):15-21. doi: 10.1093/bioinformatics/bts635. Epub 2012 Oct 25. PMID: 23104886; PMCID: PMC3530905.

EML-EBI Variant identification and analysis. https://www.ebi.ac.uk/training/online/courses/human-genetic-variation-introduction/variant-identification-and-analysis/ (accessed April 30, 2022).

Kim, D., Paggi, J.M., Park, C. *et al.* Graph-based genome alignment and genotyping with HISAT2 and HISAT-genotype. *Nat Biotechnol* **37,** 907–915 (2019). https://doi.org/10.1038/s41587-019-0201-4

Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics, 25, 1754-1760. [PMID: 19451168]

Li H. and Durbin R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics, 26, 589-595. [PMID: 20080505]

Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012 Mar 4;9(4):357-9. doi: 10.1038/nmeth.1923. PMID: 22388286; PMCID: PMC3322381.

Langmead, B. Ben Langmead - YouTube. https://www.youtube.com/user/BenLangmead (accessed April 30, 2022).

Mäkinen, V., Belazzougui, D., Cunial, F., & Tomescu, A. I. *Genome-scale algorithm design*. Cambridge University Press, 2015.

Patro, Rob, et al. “Salmon provides fast and bias-aware quantification of transcript expression.” Nature Methods (2017). Advanced Online Publication. doi: 10.1038/nmeth.4197.

Wang, L. Cutting Edge Parallel Algorithms Research with CUDA, 2015. nvidia developer. https://developer.nvidia.com/blog/cutting-edge-parallel-algorithms-research-cuda/ (accessed Dec 28, 2020).

Wolff, J.; Batut, B.; Rasche, H. Mapping. https://training.galaxyproject.org/training-material/topics/sequence-analysis/tutorials/mapping/tutorial.html (accessed April 30, 2022).

Wu TD, Watanabe CK. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. Bioinformatics. 2005 May 1;21(9):1859-75. doi: 10.1093/bioinformatics/bti310. Epub 2005 Feb 22. PMID: 15728110.
Wu TD, Nacu S. Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics. 2010 Apr 1;26(7):873-81. doi: 10.1093/bioinformatics/btq057. Epub 2010 Feb 10. PMID: 20147302; PMCID: PMC2844994.

De-novo assembly

Commins, J., Toft, C., & Fares, M. A. (2009). Computational biology methods and their application to the comparative genomics of endocellular symbiotic bacteria of insects. Biological procedures online, 11, 52–78. https://doi.org/10.1007/s12575-009-9004-1
Compeau, P.; Pevzner, P. *Bioinformatics algorithms: an active learning approach*, 2nd ed.; Active Learning Publishers: California, 2015.
Hutwagner, W.; Jain, M.; Yang, Y.; Gupta, S.; Team III Genome Assembly Group. https://compgenomics2019.biosci.gatech.edu/Team_III_Genome_Assembly_Group (accessed Dec 28, 2020).
Mäkinen, V., Belazzougui, D., Cunial, F., & Tomescu, A. I. *Genome-scale algorithm design*. Cambridge University Press, 2015.
S1. A General bioinformatics background of sequence assembly. https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&ved=2ahUKEwir7Oia_8D3AhW4hf0HHe-DBr0QFnoECAcQAQ&url=https%3A%2F%2Fjournals.plos.org%2Fplosone%2Farticle%2Ffile%3Ftype%3Dsupplementary%26id%3D10.1371%2Fjournal.pone.0169662.s001&usg=AOvVaw2KqDR4SagL_t9qEDzeKAqB (accessed April 09, 2022).
The Sequencing Centrum What is de novo assembly?. https://thesequencingcenter.com/knowledge-base/de-novo-assembly/ (accessed April 30, 2022).

Anotace genomu

Introduction to Apollo: i5K E affinis, 2015. https://www.slideshare.net/MonicaMunozTorres/introduction-to-apollo-i5k-e-affinis (accessed Dec 28, 2020).
Han, Siyu & Liang, Yanchun & Li, Ying & Du, Wei. (2016). Lncident: A Tool for Rapid Identification of Long Noncoding RNAs Utilizing Sequence Intrinsic Composition and Open Reading Frame Information. International Journal of Genomics. 2016. 1-11. 10.1155/2016/9185496.
Overview of Structural Variation. National Center for Biotechnology. https://www.ncbi.nlm.nih.gov/dbvar/content/overview/ (accessed Dec 29, 2020).

Stankiewicz P, Lupski JR. Structural variation in the human genome and its role in disease. *Annu Rev Med*. 2010;61:437-455. doi:10.1146/annurev-med-100708-204735

10 ANALÝZA GENOVÁ EXRESE

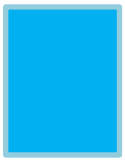
10.1 qPCR a RT-PCR

Metoda qPCR slouží pro kvantifikaci DNA a transkripce. V principu je identická s metodou klasické PCR, ovšem s přidáním fluorescenčního substrátu a s využitím speciálního cykleru, který v průběhu PCR kontinuálně zaznamenává množství DNA. qPCR se obvykle provádí v 96 nebo 384 jamkových destičkách, kdy úroveň fluorescence je zaznamenávaná v každé jednotlivé jamce, krom toho detekce probíhá během každého cyklu. Detekce fluorescence DNA odráží množství přítomné DNA, tj. i množství výchozího templátu. Fluorescence je vyzařovaná substrátem až po jeho navázání na DNA, volný substrát je neaktivní. Detekce DNA je vysoce citlivá i specifická.

U RNA platí, že vzhledem k teplotní nestabilitě RNA polymerázy, a její větší chybovosti při syntéze nového vlákna, se místo přímé amplifikace molekuly RNA využívá reverzní transkripce, při které vzniká komplementární DNA (cDNA) k dané RNA (Obr. 68). Při její přípravě se RNA nejdříve reverzně přepíše do DNA, která je následně duplikovaná na dvojřetězec.

Kvalita RNA bývá určována tzv. RIN číslem (RIN number, RNA integrity number). Tradičně se jedná o vyhodnocení poměru 28S podjednotky rRNA ku podjednotce 18S pomocí kapilární elektroforézy pro celkovou RNA ve vzorku. Standartně elektroforeogram používá desetibodovou stupnici, kde

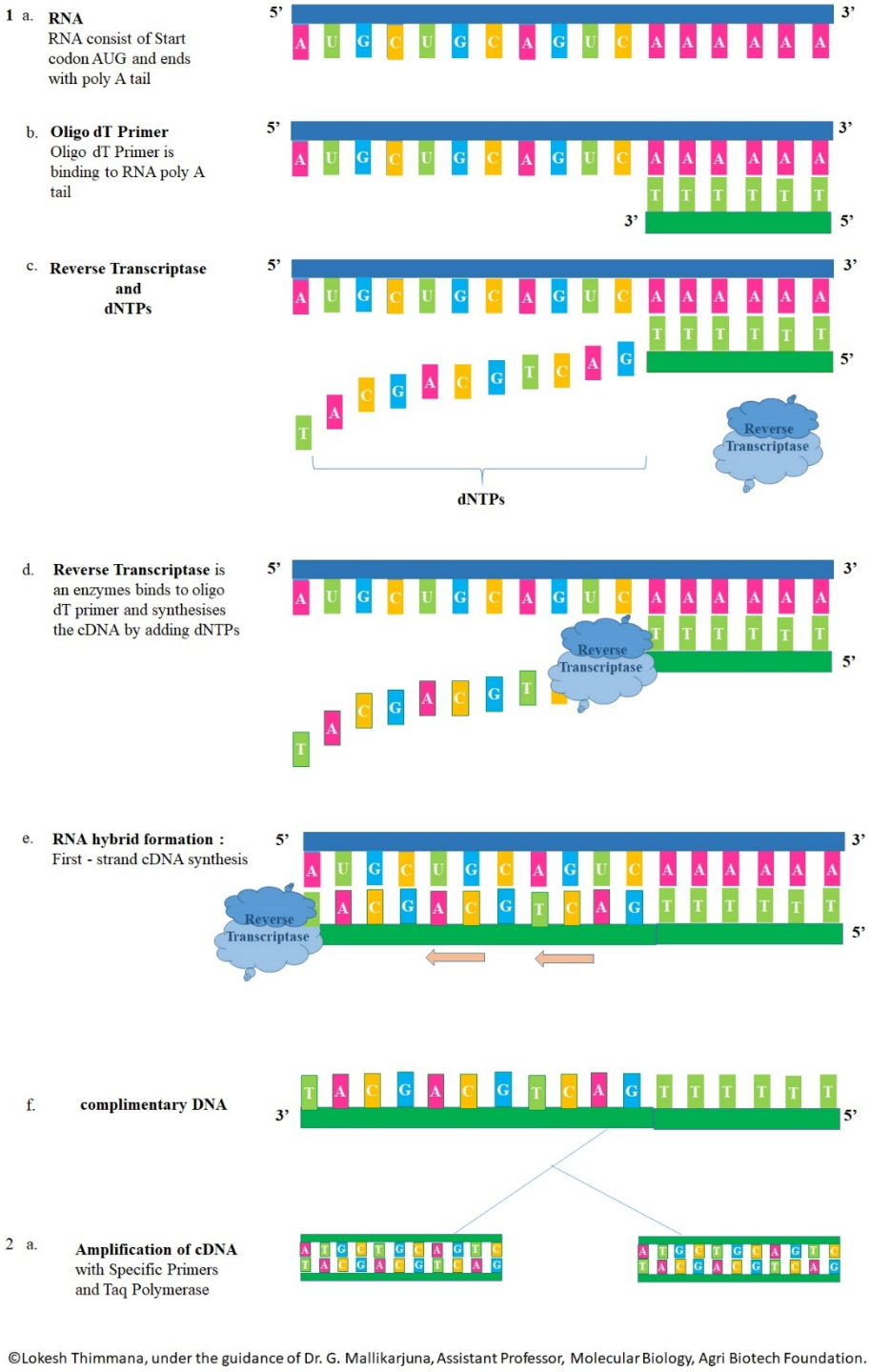
- RIN 1-4 je označení pro silně degradovanou RNA,
- RIN 4-7 pro akceptovatelnou kvalitu RNA,
- RIN 7-10 je označení pro dobrou kvalitu RNA.



I když je RIN považován za zlatý standard ve vyhodnocení kvality RNA, není úplně vhodné ho použít ve všech případech, např. pro vyhodnocení kvality RNA-seq malých molekul – miRNA bývá výrazně stabilnější než ostatní druhy RNA.

4.8 Reverse transcription polymerase chain reaction (RT-PCR)

In RT-PCR, The RNA population is converted to cDNA by reverse transcription (RT), and then the cDNA is amplified by the polymerase chain reaction .The cDNA amplification step provides opportunities to further study the original RNA species, even when they are limited in amount or expressed in low abundance. Common applications of RT-PCR include detection of expressed genes, examination of transcript variants, and generation of cDNA templates for cloning and sequencing.



©Lokesh Thimmana, under the guidance of Dr. G. Mallikarjuna, Assistant Professor, Molecular Biology, Agri Biotech Foundation.

Obr. 68: Průběh reverzní transkripce, Lokeshthimmana, CC BY-SA 4.0, via Wikimedia Commons

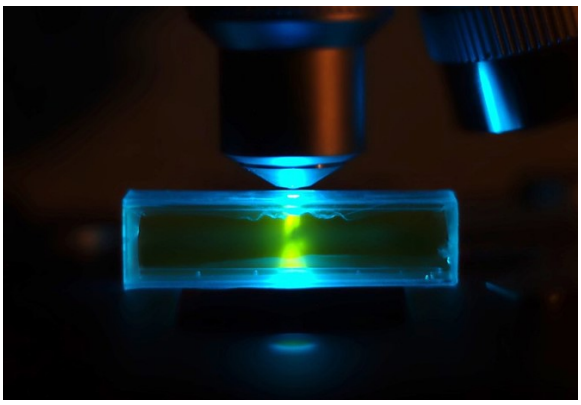
Velice často se využívá PCR metoda kombinující reverzní transkripci s kvantifikací: qRT-PCR

Někdy se lze setkat s označením RT-PCR (PCR v reálném čase) místo qPCR (kvantifikační PCR)!!!

10.1.1 Zdroje fluorescence

SYBR green

Jedním z nejrozšířenějších zdrojů fluorescence pro kvantifikaci qPCR/qRT-PCR je SYBR green (Obr. 69). Řadí se k nespecifickým fluorescenčním substrátům, které vytváří vazbu na dvouvláknovou DNA bez jakékoliv specificity k sekvenci. Zároveň se jedná o tzv. interkalační látku – váže se pouze na dvouřetězcovou DNA, pokud je v reakci přítomna pouze jednořetězcová DNA, molekuly SYBR greenjsou v roztoku ve volné formě a nefluoreskují. V okamžiku, kdy je dvouřetězcová DNA denaturovaná na jednořetězcovou, také dojde k uvolnění SYBR greenu z DNA. V průběhu reakce dochází k nárustům a poklesům fluorescence podle fáze PCR. Jednou z nevýhod použití SYBR greenu je možná nižší přesnost kvantifikace v některých případech než při použití jiných postupů.



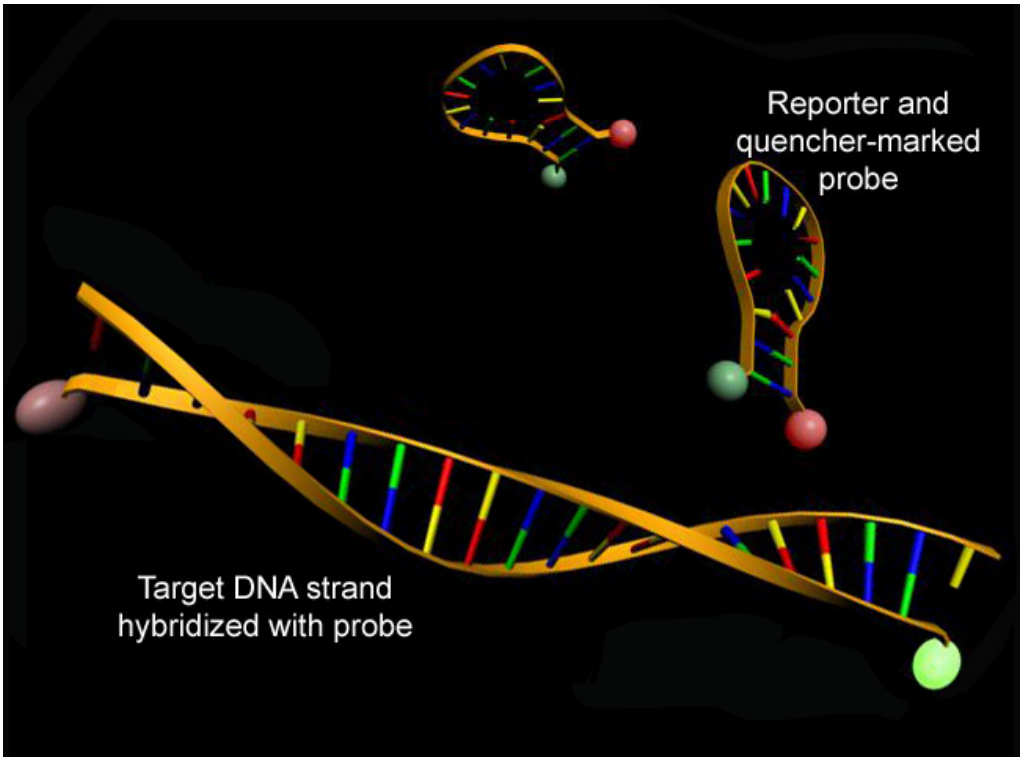
Obr. 69: SYBR green, Zephyris, CC BY-SA 3.0, via Wikimedia Commons

Molecular beacon

Molecular beacons (Obr. 70) jsou jednořetězcové oligonukleotidy se vzájemně komplementární konci, díky kterým se vytvoří tvar vlásenky. Na jednom konci sondy je fluorochrom, na druhém konci je tzv. zhášeč (quencher), který ve vlásence blokuje emisi fluorescence z fluorochromu. Ve chvíli, kdy dojde k denaturaci vlásenky a následně hybridizaci s cílenou sekvencí na DNA, dojde k separaci molekuly fluorochromu od zhášeče, a tím dojde k emisi fluorescence. Pokud se specifická sekvence na DNA nenachází, dojde k opětovnému spojení komplementárních konců sondy do vlásenky, a k emisi fluorescence nedochází.

Molekular beacons jsou tvořeny ze čtyř částí:

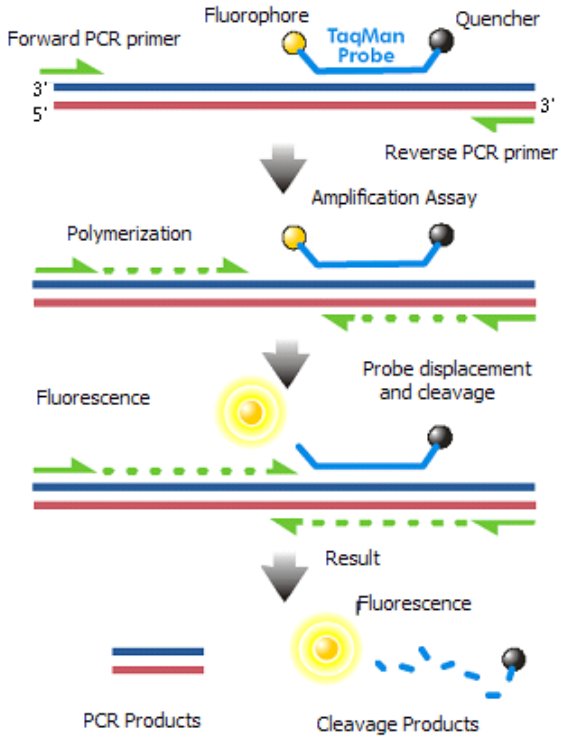
- Očko vlásenky: 18-30 bazí, komplementární ke specifické sekvenci DNA
- Báze vlásenky: 5-7 bazí na koncích oligonukleotidu, vzájemně komplementárních
- 5' fluorofor: molekula fluorochromu, která je umístěna na 5’ konci sondy
- 3' zhášeč: molekula barviva, která je kovalentně vázaná na 3’ konec sondy



Obr. 70: Molecular beacons, S. Jähnichen using 3d studio max and Photoshop Element, GNU Free Documentation License, via Wikimedia Commons

TaqMan

TaqMan sondy (Obr. 71) jsou hydrolyzační sondy o 18–22 párech bazí, které využívají 5' exonukleázové aktivity Taq polymerázy při syntéze DNA. Obsahují sekvence komplementární ke specifické sekvenci na DNA. Na jednom konci sondy se nachází reportérový fluorofor, na druhém je molekula zhášeče. Pokud sonda není činností Taq polymerázy hydrolyzovaná, zhášeč je v blízkosti fluorochromu a k fluorescenci nedochází. Během PCR se sonda váže specificky na komplementární sekvenci DNA mezi forward a reverse primery. Když během syntézy DNA Taq polymeráza narazí na navázanou TaqMan sondu, rozloží ji svojí exonukleázou aktivitou. Tím dojde k separaci fluorochromu od zhášeče, a tím k emisi fluorescence. V průběhu PCR tak dochází k postupnému zvyšování signálu.



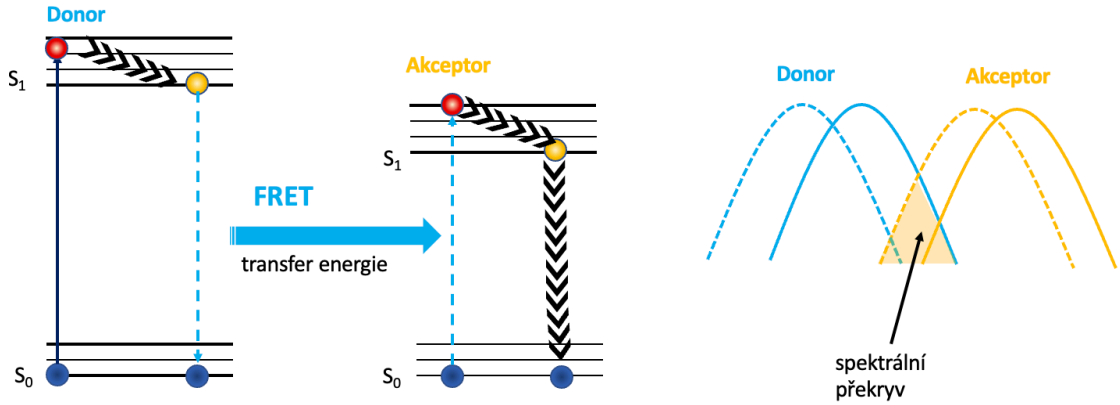
Obr. 71: TaqMan sonda, Braindamaged, Public domain, via Wikimedia Commons

Použití molecular beacon nebo TaqMan sond se zvyšuje specifita RT-qPCR reakcí.

FRET

FRET (Försterův rezonanční přenos energie, Obr. 72) je excitovaný stav vzájemné interakce dvou fluoroforů, kdy emisní energie jednoho fluoroforu (donoru) na 3' konci první sondy je spojena s excitací druhého fluoroforu (akceptoru) na 5' konci druhé sondy. Jedná se o mechanismus nezářivého přenosu energie (bez emise fotonu) mezi dvěma molekulami barviva s elektronově excitovanými stavy. Molekuly akceptoru a donoru musí být umístěn blízko sebe.

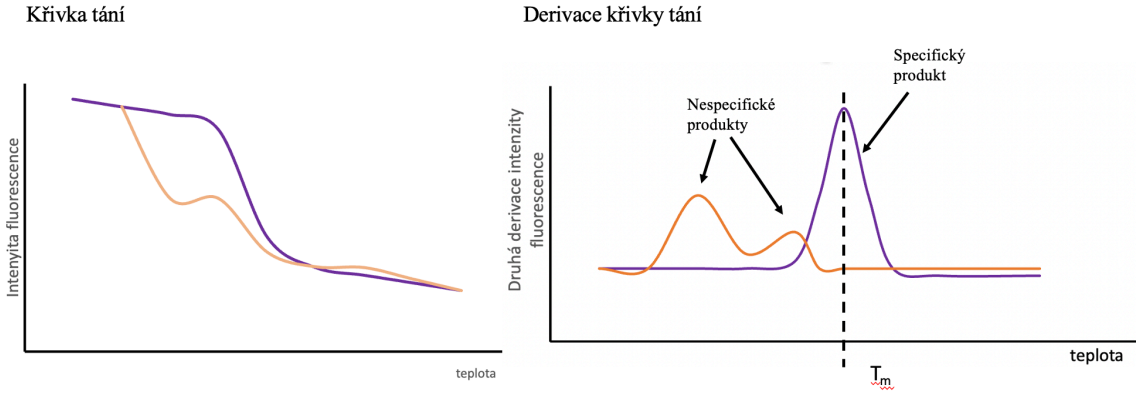
Během PCR sondy hybridizují k cílovým sekvencím na DNA. Donorový fluorochrom, excitovaný externím zdrojem světla, přenese část své excitační energie na akceptorový fluorochrom. Excitovaný akceptorový fluorochrom emituje světlo o jiné vlnové délce, které je následně detekováno detektorem. Fluorochromy jsou cíleně vybírány tak, aby se emisní spektrum jednoho fluorochromu překrývalo s excitačním spektrem druhého fluorochromu.



Obr. 72: Jablonského diagram FRET přenosu mezi donorem a akceptorem

10.1.2 Křivka tání

PCR může být náchylné na tvorbu nespecifických produktů či produktů typu primer-dimer. V případě využití SYBR green můžeme poměrně snadno zjistit jejich možný vznik v reakci. Využívá se k tomu **melting curve** (křivka tání), která ukazuje změnu intenzity fluorescence při různých teplotách (případně její derivaci). Při vzrůstající teplotě dochází ke snížení intenzity fluorescence. V principu jde o to, že různé PCR produkty mají i různé teploty tání (T_m, melting temperature), přičemž nespecifické produkty mají obvykle teplotu tání nižší než specifické. Křivka tání s jedním vrcholem ukazuje na čistě specifickou reakci, křivka s více vrcholy na vznik nespecifických produktů (Obr. 73).



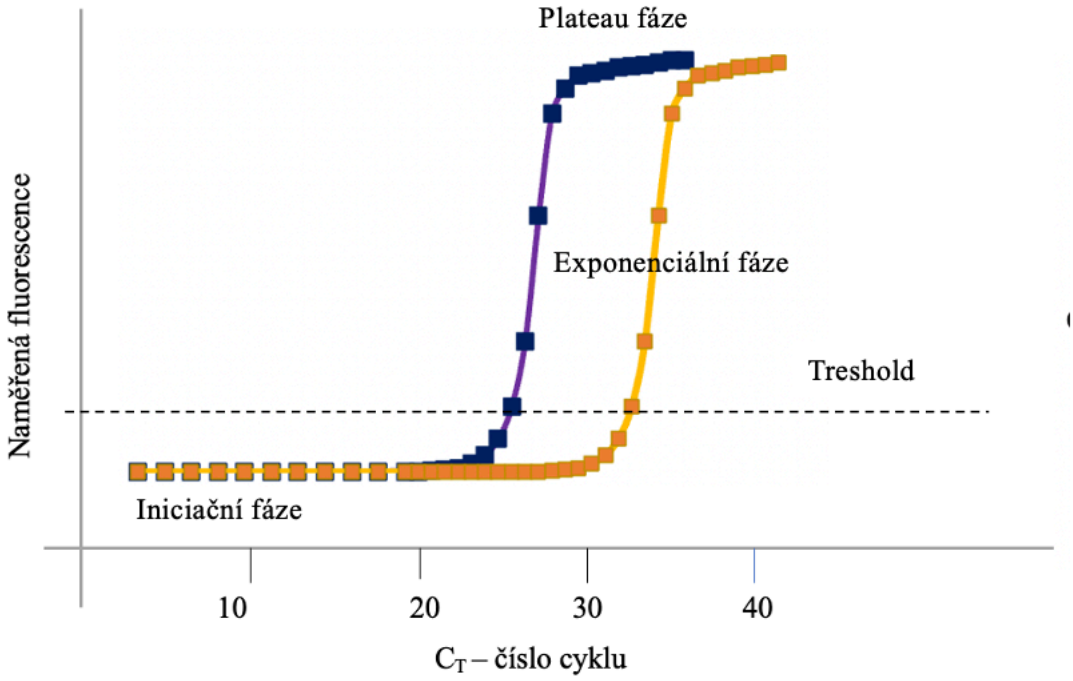
Obr. 73: RT-PCR, křivka tání a její derivace

10.1.3 Reakční křivka

Základní pojmy:

- **Threshold**: hodnota prahu pozadí
- **C_T** (**cycle of treshold**): Jedná se o číslo pořadí cyklu, ve kterém došlo k nárůstu fluorescence nad určenou hodnotu tresholdu (prahu pozadí), kdy, již může být detekována. Hodnota Ct se používá při následné kvantifikaci nebo detekci přítomnosti/nepřítomnosti např. u virových onemocnění (viz. rok 2020 – současnost (únor 2022) v souvislosti s onemocněním Covid-19). Porovnáním hodnot Ct vzorků neznámé koncentrace s řadou standardů lze relativně přesně určit množství templátové DNA v neznámé reakci., platí C_T (cycle of threshold) = Cp (crossing point) = Cq (cycle of quantification).
- **Templát**: sekvence DNA nebo RNA kterou chceme amplifikovat
- **LOD** (Minimální limit detekce): Počet detekovatelných molekul, Signál se statisticky liší od pozadí, ale nelze jej kvantifikovat.
- **LOQ** (Limit kvantifikace): Signál se statisticky liší od pozadí a je možné ho kvantifikovat.

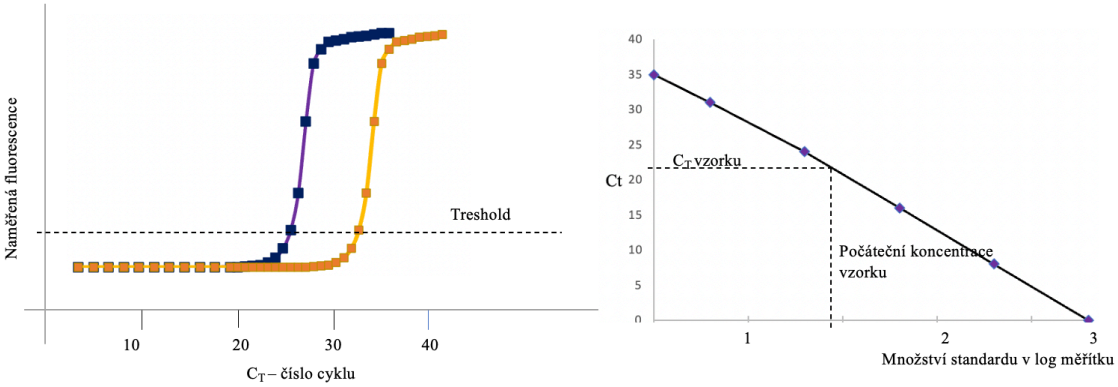
Na začátku je množství templátové DNA nízké, díky čemuž intenzita fluorescenčního substrátu nepřesahuje úroveň pozadí, dochází k nárůstu produktu. Další průběh reakce je **exponenciální**. Intenzita fluorescence přesahuje úroveň pozadí a může být detekována detektorem. V ideálním případě se množství DNA produktu zdvojnásobí podle vzorečku 2^n, kde n je pořadí cyklu. Ke kvantifikaci DNA na základě fluorescence se přistupuje v tomto kroku, protože intenzita fluorescence v tomto kroku poměrně přesně odráží množství DNA v reakci. Finální fází je tzv. **plateau** fáze. Sice je ve vzorku přítomno obrovské množství templátové DNA, ale dochází ke zpomalení nárůstu intenzity fluorescence a ke zploštění reakční křivky (Obr. 74). To může mít příčinu např. v tom, že některé komponenty reakce byly již spotřebovány, čímž dochází k výraznému zpomalení kinetiky reakce. Množství fluorescence produkované v této fázi nemusí vůbec reflektovat množství templátové DNA dodané do reakce, a tak není možné tuto fázi pro kvantifikaci DNA využívat. Přestože mohou mít dvě reakce stejnou koncentraci vstupní DNA, plateau fáze může nastat v jiném čase kvůli jiné kinetice reakce pro daný vzorek.



Obr. 74: RT-PCR, threshold – práh detekce fluorescence, v místě protnutí s křivkou je definováno C_T. C_T (threshold cycle) - číslo cyklu, ve kterém fluorescence překročí limit detekce, Pfeiferová L

10.1.4 Kvantifikace

Absolutní (Obr. 75): využití metody standardní (kalibrační) křivky. Pro každý gen, o který se zajímáme, vytvoříme standartní křivku s vynesením hodnot C_T proti log hodnotám ředění známých koncentračních standartů tohoto genu. Porovnáním neznámé hodnoty se standardní křivkou a extrapolací získané hodnoty dostaneme absolutní kvantifikaci množství genu ve vzorku. V ideálním případě je při každé kvantifikaci vzorku generována nová standardní křivka, ale v praxi se generuje standardní křivku jednou a opakovaně se používá ke kvantifikaci vzorků po jistou dobu. Absolutní kvantifikace se používá u metody Droplet digital PCR technologie.



Obr. 75: Absolutní kvantifikace. a) Amplifikační křivky dvou vzorků o různém počátečním množství. V místě protnutí křivky s tresholdem je hodnota C_T. b) Kalibrační křivka standardů o známé koncentraci

Relativní (častější): změna genové exprese je analyzována ve srovnání s jiným referenčním genem či skupinou genů (například již zmíněnými housekeepingovými geny). Pro základní výpočet se může použít Livakova metoda (komparativní metoda $2^{-\Delta\Delta C_T}$).

Obecný vzorec pro výpočet počátečního množství molekul je podle rovnice (3)

$$N(c) = N(0)(1+E)^c \tag{3}$$

kde N_c je počet amplifikovaných molekul po C cyklech, N_0 výchozí počet molekul, E efektivita a C počet cyklů. Vzhledem k tomu, že měříme fluorescenci, a ne počet molekul, používá se vzorec podle rovnice (4)

$$S(c) = K \times A \times F(1+E)^{\mathrm{C}} \tag{4}$$

kde S je naměřený fluorescenční signál, K nastavení přístroje, A je množství vzorku, F je frakce mRNA měřeného genu, AxF je koncentrace měřeného genu. Hodnota C_T je definována dosažením amplifikační křivky k definované „treshold value“ St (rovnice 5 a 6).

$$St = K \times A \times F(1+E)^{C_T} \tag{5}$$

$$\log_2 \frac{S_T}{K} = \log_2 A + \log_2 F + C_T \log_2(1+E) \tag{6}$$

Pokud účinnost (efektivita) reakce není 1, pak provedeme korekci Ct podle rovnice (7)

$$C_T^{corr} \leftarrow C_T \log_2(1+E) \tag{7}$$

Vzhledem k tomu, že se často předpokládá, že E = 1, necháme označení Ct i pro opravenou hodnotu, viz rovnice (8).

$$\log_2 \frac{S_T}{K} = \log_2 A + \log_2 F + C_T{}^{corr} \quad (8)$$

Pokud budeme spolu s měřeným genem t měřit referenční gen r, můžeme vyhodnotit jejich relativní abundanci a odstranit z rovnice konstantu pro nastavení přístroje a vzorek (rovnice 9–13), kde Q je relativní exprese R (v porovnání s R) a $2^{-\Delta C_T}$ je „delta C_T“.

$$C_T^t = \log_2 \frac{S_T}{K} - \log_2 A - \log_2 F^t \quad (9)$$

$$C_T^r = \log_2 \frac{S_T}{K} - \log_2 A - \log_2 F^r \quad (10)$$

$$C_T^r - C_T^t = \log_2 F^t - \log_2 F^r \quad (11)$$

$$\log_2 \frac{F^t}{F^r} = C_T^r - C_T^t \quad (12)$$

$$\frac{F^t}{F^r} = Q = 2^{C_T^r - C_T^t} = 2^{-\Delta C_T} \quad (13)$$

Finálně dostaneme rovnici (14), ve které jsme se zbavili přístrojové konstanty a zkreslení u přípravy vzorků.

$$\Delta C_T = Ct^t - Ct^r \quad (14)$$

Místo jednoho referenčního genu můžeme použít i aritmetický průměr C_T^r více referenčních genů (geometrický průměr jejich frakcí), rovnice (15).

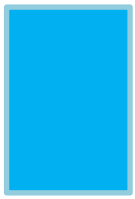
$$\tilde{C}_T^r = \frac{1}{N_R}\sum_{i=1}^{N_R} C_T^i = \log_2 \frac{S_T}{K} - \log_2 A - \frac{1}{N_R}\sum_{i=1}^{N_R} \log_2 F^i \quad (15)$$

Předpokládejme, že referenční gen je exprimován se stejnou intenzitou ve všem vzorcích. Pak můžeme porovnávat hodnoty relativní exprese Q přes všechny vzorky, tj. treatment proti kontrolám, KO vs. WT vzorky, podle porovnání (16).

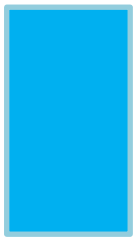
$$RQ = 2^{-\Delta\Delta C_T} \quad (16)$$

Změny v genové expresi jsou často prezentovány na logaritmické škále, rovnice (17)

$$\log_2 FC = \log_2 RQ = -\Delta\Delta C_p \quad (17)$$



Protože je množství DNA v každém cyklu PCR zhruba zdvojnásobeno, je Ct v logaritmickém měřítku. Jako vyjádření poměru exprese se využívá proto log2 fold change hodnota, viz rovnice (17).



$\log_2 FC = 0 \implies$ žádná změna v genové expresi

$\log_2 FC = 1 \implies$ dvojnásobná změna, upregulace

$\log_2 FC = -1 \implies$ dvojnásobná změna, downregulace

10.1.5 Účinnost PCR

Efektivita je teoreticky stoprocentní, ale reálně bývá nižší. To je způsobeno mnoha faktory, jako vazbou primerů, poměrovým složením reakce, možností přítomnosti inhibitorů či enhancerů reakce, u RT-PCR také nízkou kvalitou RNA podle RIN čísla, degradací RNA před transkripcí na cDNA. Podle již zmíněné rovnice (5), naměřený fluorescenční signál S(c) se vypočítá jako $K \times A \times F(1+E)^C$.

Účinnost (efektivita) reakce se dá vypočítat podle rovnic (18) a (19).

$$\log_{10} \frac{S_T}{K} = \log_{10} A + \log_{10} F + C_T \log_{10}(1+E) \tag{18}$$

$$C_T = \frac{\log_{10} {}^{S_T}/_{K}}{\log_{10}(1+E)} - \frac{1}{\log_{10}(1+E)} \times (\log_{10} A + \log_{10} F) \tag{19}$$

Pokud E=1, směrnice $-\frac{1}{\log_{10}(1+E)}$ má přibližnou hodnotu -3,3; číslo menší jak 3,32 (v absolutní hodnotě) značí efektivitu menší jak 100 %, vyšší hodnota ukazuje na nízkou kvalitu vzorku či problémy při přípravě reakce. Jako dobrý výsledek efektivity PCR se udávají hodnoty E mezi 1,75 a 2.

Proměnná F (frakce mRNA měřeného genu) u RNA-Seq zhruba odpovídá hodnotě TPM vynásobené 10^6

10.1.6 Normalizace

Důležitým krokem pro kvantifikaci qPCR a qRT-PCR je normalizace dat podle alespoň jedné z následujících proměnných:

- velikost vzorku, hmotnost nebo objem tkáně,
- celkové množství extrahované RNA,
- celkové množství genomické DNA,
- referenční ribosomální RNAs (18S nebo 28S rRNA),
- referenční mRNA, tzv. vnitřní reference.

Velmi často se pro normalizaci dat využívá referenčních genů, jejichž počet kopií by měl být stejný u všech testovaných vzorků (v rámci odchylky). Jako reference se často využívají tzv. housekeepingové geny, tedy geny kódující proteiny či funkční RNA, které se účastní základních buněčných procesů, které jsou nezbytné pro existenci buňky, bez ohledu na její specifickou roli ve tkáni nebo organismu. Příkladem jsou (u člověka) ACTB (aktin, beta), GAPDH (Glycelaldehyd-3-fosfát dehydrogenáza), PGK1 (Fosfoglycerát kináza 1), B2M (Beta-2-microglobulin), TBP (TATA box vazebný protein), SDHA (Komplex sukcinátdehydrogenázy, podjednotka A, flavoprotein (Fp)), ARBP (Kyselý ribosomální fosfoprotein PO) a další.

Normalizace podle geNorm: Máme n referenčních genů s indexy i, j,… a M vzorků m, n,…, provedeme experiemnt podle rovnice (20)

$$C_T^{mi} = \log_2 \frac{S_T}{K} - \log_2 A^m - \log_2 F^{mi}$$
$$C_T^{mj} = \log_2 \frac{S_T}{K} - \log_2 A^m - \log_2 F^{mj} \quad (20)$$

C_T^{mi} a C_T^{mj} je cyklus, kdy se docílilo tresholdu, $\log_2 \frac{S_T}{K}$ je konstanta, $\log_2 A^m$ je množství vzorku (absorbance), $\log_2 F^{mi}$ a $\log_2 F^{mj}$ jsou relativní výskyty i a j. Nadefinujeme set hodnot A_{ij} pro každý pár genů (i, j), rovnice (21) a (22)

$$A_{ij} = \left\{\log_2 \frac{F_{mi}}{F_{mj}}\right\}, m = 1,2,..M \quad (21)$$

$$A_{ij} = \left\{C_T^{mj} - C_T^{mi}\right\}, m = 1,2,..M \quad (22)$$

Podle rovnice (23) vypočítáme standartní odchyluA_{ij}, kterou označíme jako V_{ij}

$$V_{ij} = stdev(A_{ij}) \quad (23)$$

Finálně definujeme stabilitu genu j jako rovnice (24)

$$M_j = \frac{1}{n\mp} \sum_{i \neq j} V_{ij} \quad (24)$$

Na základě stability genu vybereme ty nejlepší jako referenční geny. Jejich počet se odhaduje, případně existují pseudokritéria.

10.2 RNA-seq

RNA-seq je technika, která se používá na kvantifikaci RNA ve vzorcích pomoci metod sekvenování příští generace. Tímto je umožněno analyzovat a kvantifikovat celý transkriptom, celkový buněčný obsah RNA ve vzorcích, případně je možno se cíleně zaměřit pouze na určitou část celkové RNA (velmi časté je např. cílené zkoumání malých RNA ve vzorcích. Z celkové RNA se většinou odstraňují sekvence ribosomální RNA).

Pochopení transkriptomu je klíčové pro propojení informací o genomu organismu s funkční genovou expresí. RNA-seq nám může říct, které geny jsou v buňce zapnuty, jaká je jejich úroveň exprese, při jaké podmínce jsou aktivovány, vypnuty či jinak změněny, ale také umožňuje náhled na alternativní genové sestřihy, post-transkripční modifikace, jednonukleotidové i vícenukleotidové polymorfismy, somatické mutace, inzerce a delece. Typicky je RNA-seq využíván na vyhodnocení změn genové exprese v experimentech kdy je hodnocena zdravá tkáň proti nemocné, případně kontrolní zdravé/ovlivněné buňky proti buňkách ovlivněným za určitých podmínek, např. léčivem.

Zatímco polymorfismus je spojen s četností výskytu varianty v populaci, somatická mutace je spojena s funkční změnou (vadou, jak u regulačních elementů, tak proteinů u protein kódujících genů).

10.2.1 Workflow RNA-seq

Workflow (Obr. 76) pro zpracování RNA-seq dat kopíruje standardní zpracování sekvenačních dat (pro účely našeho manuálu myšleny standartní FASTQ soubory):

preprocesing: kontrola kvality dat a odstranění nežádoucích sekvencí

mapování: vlastní mapovací proces

post-alingment operace: výpočet vlastností, kontrola mapování, vytváření reportů kvality dat a mapování a jiné

Za zlatý standard pro analýzu NGS dat se mohou např. považovat pipeliny od skupiny nf-core, https://github.com/nf-core.

Preprocesing

Napřed se udělá první kontrola kvality hrubých sekvenačních dat pomocí nástrojů pro kontrolu kvality, jako je FastQC nebo FASTP nebo jiné. V případě, že u přípravy knihoven byly pro zlepšení přesnosti sekvenování a korekci sekvenačních chyb použity sekvence UMI nebo barcody, je třeba je nejpozději v tomto kroku odstranit ze sekvencí např. pomocí UMI-tools, a uložit jejich sekvenci do headrů readů pro budoucí zpracování.

Dalším krokem je odstranění adaptorových sekvencí, bazí z konců readů s nízkou kvalitou a ostatních sekvencí, vzniklých v důsledku sekvenačních chyb, které by mohly zkreslit výsledky mapování pomocí nástrojů jako jsou Cutadapt a Trimmomatic. Po tomto kroku by měla následovat druhá kontrola kvality dat, která by měla ukázat zlepšení. V této části také můžeme odstranit ribosomální sekvence, pokud je nechceme zahrnout do konečného výsledku.

Mapování

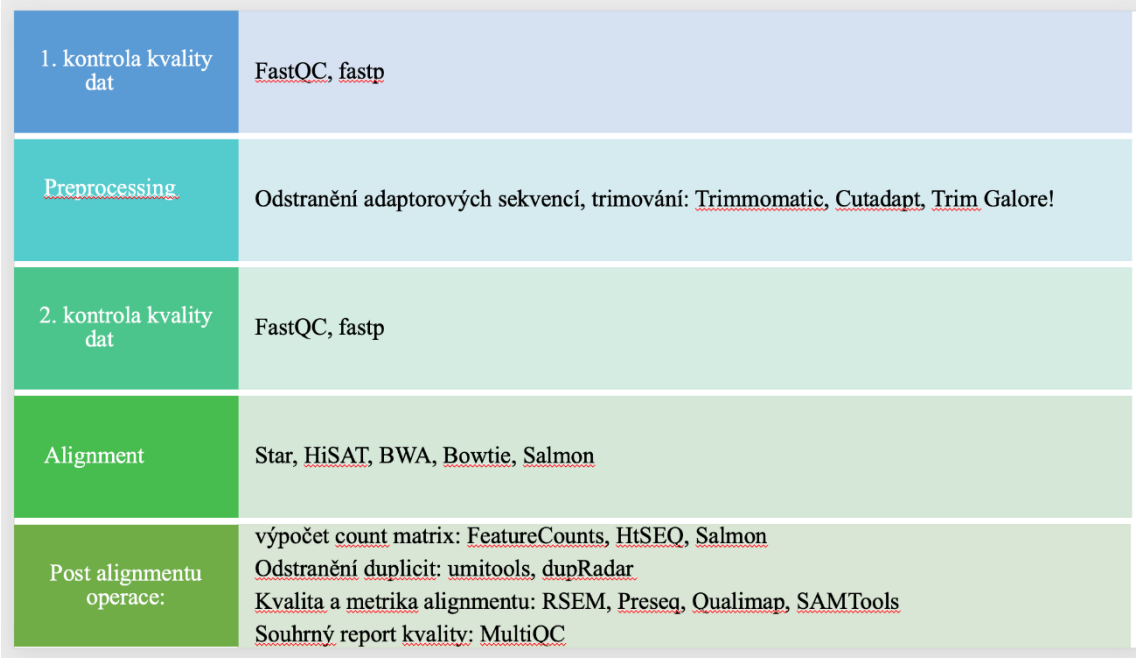
Vlastní mapovací proces Viz Kapitola 9. Mapování.

Post-alignment operace

Do procesů po mapování řadíme všechny procesy vázané na výsledky mapování. Základem je třídění a indexování namapovaných sekvencí pomocí SAMtools. V SAM/BAM souboru namapovaných sekvencí, se alignmenty vyskytují náhodně podle pořadí ve FASTQ souborech. Pro další zpracování je proto minimálně vhodné, v některých případech i nutné, uspořádat alignmenty v genomovém pořadí, tedy podle pořadí chromosomů a podle chromosomových pozic. Dalším důležitým krokem je indexování alignmentů, které umožňuje rychlou manipulaci se zarovnáními, a v některých případech (IGV prohlížeč) umožňují vizualizaci zarovnání proti referenčnímu genomu.

Po seřazení a indexování můžeme provádět další operace s daty. Pokud byly použity UMI/barcody můžeme odstranit duplikované sekvence pomocí UMI-tools podle informací o UMI v hlavičkách readů v alignmentu. Pomocí různých nástrojů můžeme zjistit statistické informace o alignmentu: např. míru duplikací (dupRadar), distribuci namapovaných readů, PCR bias, GC bias, coverage, směrovost vlákna (RSeQC), komplexitu a redundantnost readů pro možné změny ve tvorbě knihovny u příštího sekvenování (Preseq), vyhodnocení readů z hlediska poměrů protein kódujících oblastí, intronů, spojovacích sekvencích a dalších informací důležitých pro možné změny u příští sekvenační knihovny (Qualimap) nebo třeba assemblovat alignmenty do možných transkriptů (StringTie) a zahrnout je do komplexního souboru pomocí MultiQC.

Z hlediska dalšího zpracování pro potřeby analýzy genové exprese je pak nejdůležitější kvantifikace alignmentu: výpočet vlastností v zarovnání.



Obr. 76: Základní workflow pro RNA-seq. Salmon samozřejmě není aligner, ale může sloužit jako tzv. pseudoaligner, je tedy zde zařazen i do alignmentu

10.2.2 Nástroje pro získání count table

FeatureCounts

FeatureCounts je nástroj v balíčku Subread implementovaný v jazyku C, a také funkce v jazyku R, který se používá ke kvantifikaci readů získaných metodami sekvenování příští generace. Přiřazuje čtení ke genomovým vlastnostem jako jsou např. geny, exony, miRNA. Zohledňuje veškeré mezery (inzerce, delece, spojení exon – exon nebo fúze), které se ve čtení nacházejí. Jako hit je označován overlap mezi čtením/fragmentem a danou genomickou vlastností, v případě více překryvů mohou být reportovány všechny vlastnosti a jejich počty za použití -O volby. Může být použit jak na single-end, tak na pair-end data. Jako input může být použít SAM i BAM formát (seřazený i neseřazený), dalším vstupem je soubor s anotacemi (ve formátu GTF nebo SAF). Outputem jsou dva soubory: count table s jmény vlastností, chromosomovými pozicemi vlastností, jejich sumárním počtem a dalšími charakteristikami, a soubor s informaci o počtu a procentuálním poměru přiřazených a nepřiřazených vlastností.

Ukázka kódu:

```
featureCounts -T 16 -p -g gene_name -a Homo_sapiens.GTF -o counts.txt
input.BAM
```

HTseq

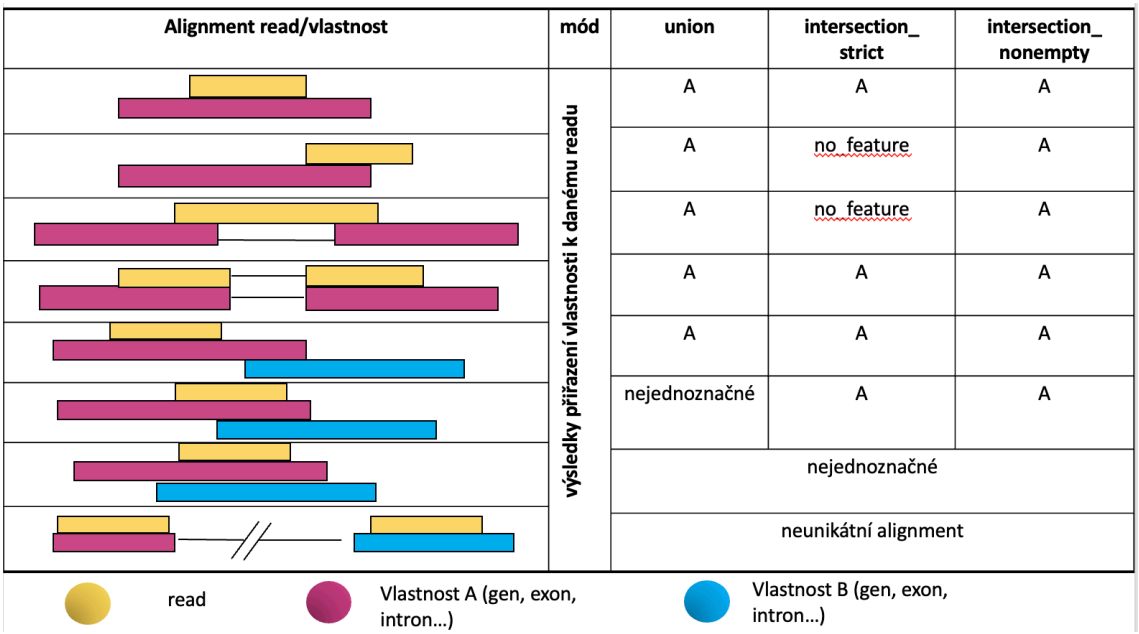
HTseq (obr. 77) je pythonský skript určený pro kvantifikace vlastností RNA-seq a jiných dat z NGS technologií. Jako vstup je použit SAM nebo BAM soubor a anotace ve formátu GTF/GFF. Pro každou anotaci je vyhodnocen počet přiřazených čtení, která překrývají danou vlastnost v anotaci. Vlastností je myšlen interval (rozsah pozic) na chromosomu nebo spojení takových intervalů. V případě RNA-seq jsou to obvykle geny, kde za gen je považováno spojení všech jeho exonů. V jiném případě může být za vlastnost považován jednotlivý exon, např. za účelem kontroly alternativního sestřihu. Pro porovnání u ChIP-Seq se jako vlastnost mohou brát vazebné oblasti z předem určeného seznamu. Pro případ overlapu HTseq umožňuje vybrat si mezi třemi moduly:

1. union: sjednocení všech množin S (i)
2. intersection-strict: průsečík všech množin S (i)
3. intersection-nonempty: průsečík všech neprázdných množin S (i),

kde množina S(i) definována jako množina všech prvků překrývajících polohu i. Pokud S obsahuje přesně jednu vlastnost, bude se pro tuto funkci počítat čtení. Pokud obsahuje více než jednu funkci, čtení se počítá jako nejednoznačné (a nepočítá se pro žádné funkce), a pokud je S prázdné, čtení se počítá jako no_feature.

Ukázka kódu:

```
htseq-count -i gene_name -m intersection-nonempty -f BAM input.BAM
Homo_sapiens.GTF > counts.txt
```



Obr. 77: Zobrazení výsledných efektů v rámci tří zmíněných módů (union, intersection_strict a intersection_nonempty) spolu s –nonunique (all) volbou https://htseq.readthedocs.io/en/master/ (accessed May 29, 2022)

StringTie

StringTie je další nástroj určený na kvantifikaci NGS dat. Jako vstup se vkládá seřazený a indexovaný soubor BAM. Volitelným vstupem je pak anotační soubor ve formátu GTF/GFF. Při využití anotačního souboru, StringTie spočítá pro overlapující ready s anotací coverage, TPM a FPKM (viz. Následující kapitola Normalizace) hodnoty pro daný BAM soubor. Hlavní výhoda tohoto softwaru ale spočívá v možnosti sestavit exprimované vlastnosti přímo z RNA-seq dat, a sestavit nové genové modely. Pro data, která nejsou kryta anotací, může vytvořit dodatečné transkripty (volba -e) a přidat je do kvantifikace.

Výstupní soubory:

- soubor GTF obsahující sestavené transkripty,
- genové abundance v tab-delimited formátů (count table).

Ukázka kódu:

```
stringtie -p 16 -e -G známé_vlastnosti_v_GTF -B -o
nově_nalezené_vlastnosti_v_GTF -A count_table input.BAM
```

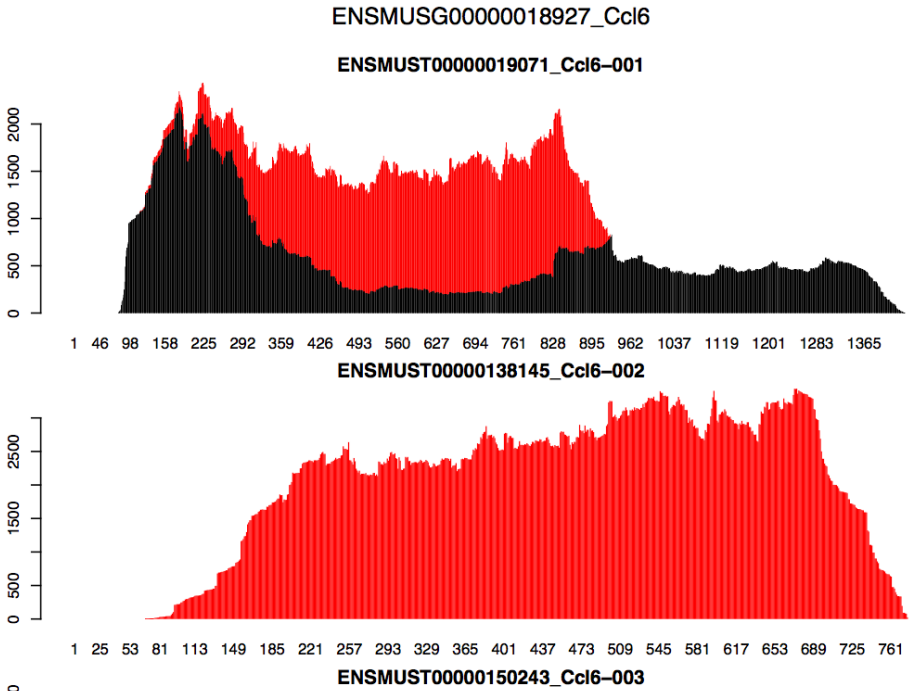
RSEM

RSEM je balíček nástrojů kompilovaný v C, který se používá na odhad hladiny genové a izoformní exprese z dat RNA-seq dat. Oproti předchozím nástrojům, které využívají ready namapované na referenční sekvenci ve formátu BAM, RSEM vyžaduje ready mapované proti referenčnímu transkriptomu. RSEM se dá použít na single-end i pair-end data. Kromě určení skóre kvality, odhadu RSPD (read start position distribution, rozložení počátečních pozic jednotlivých čtení) a proměnlivé délky readů poskytuje informace o úrovni exprese. Pro vizualizaci v UCSC Genome a IGV může vygenerovat soubory BAM a Wiggle v transkripčních i genomových koordinátách. RSEM má také vlastní skripty pro vygenerování grafů hloubky transkripce ve formátu PDF. Jednotlivé grafy lze skládat, přičemž hloubka jedinečných čtení (single mapping) je zobrazena černě a ty, které přispívají k vícenásobným čtením (multiple mapping) jsou zobrazeným červeně (Obr. 78). Kromě toho lze také vizualizovat modely získané z dat.

Ukázka kódu:

```
rsem-calculate-expression -p 8 --paired-end --BAM --estimate-rspd --append-
names \--output-genome-BAM input.BAM Homo_sapines.GTF

rsem-calculate-expression --num-threads 8 --time --star $FASTQ
$(file_directory_output)
```



Obr. 78: Ukázka RSEM hloubky jednotlivých čteních, upraveno, Li, B. A Short Tutorial for RSEM. https://github.com/bli25broad/RSEM_tutorial (accessed Dec 17, 2020)

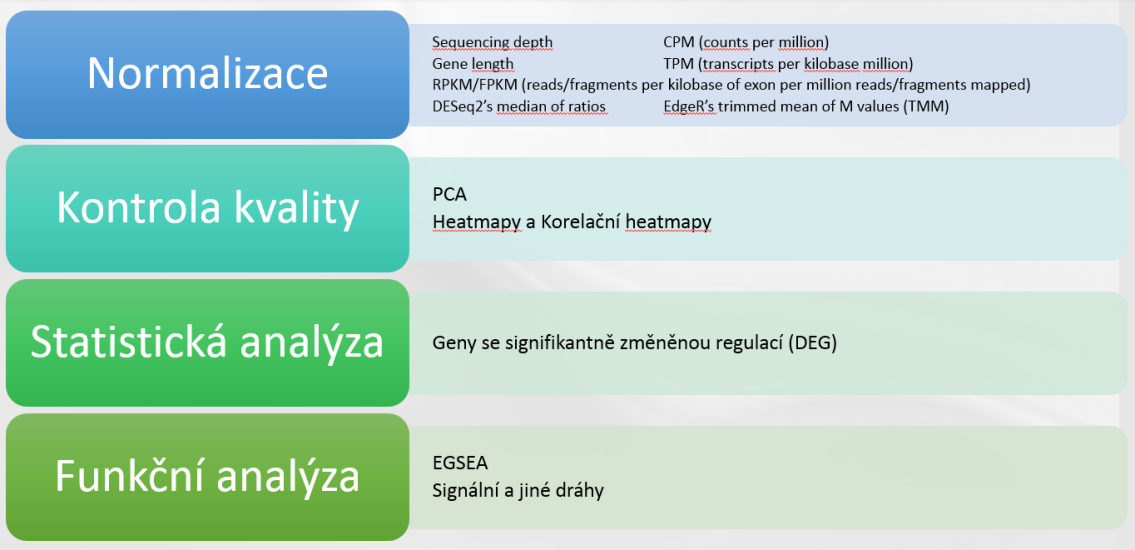
SALMON

Dalším nástrojem, který se využívá ke kvantifikaci exprese transkriptů RNA-seq dat. Využívá konceptu kvazi-mapování s dvoufázovou inferenční (usuzovací) procedurou, což umožňuje velmi rychlý odhad exprese spolu s malým využitím paměti. Inference se provádí pomocí expresivního a realistického modelu dat RNA-seq, který bere v úvahu experimentální atributy a zkreslení běžně pozorované ve skutečných datech RNA-seq. Salmon není mapovací algoritmus v pravém slova smyslu, ale používá se jako pseudoaligner.

Kvantifikace v Salmonu je umožněna ve dvou režimech: první vyžaduje pouze set referenčních transkritpů a soubor typu FASTA/FASTQ s cílovými daty, které přímo zpracuje. V druhém režimu je zapotřebí kromě setu referenčních transkriptů a soubor SAM/BAM s alignmentem. Salmon používá *streaming inference method* na kvantifikaci na úrovni transkripce metody, jejímž základním předpokladem je, že pozorování (tj. čtení nebo zarovnání) jsou prováděna *náhodně*. Z tohoto důvodu alignment NESMÍ být sortovaný – pokud ano, je důležité před spuštěním provést randomizaci.

10.3 Diferenciální genová exprese RNA-seq

Vyhodnocení diferenciální genové exprese probíhá často za pomocí R balíčku DESeq2, limma nebo EdgeR (Obr. 79), kde jako vstup slouží tzv. count-table (viz. Předchozí kapitola)



Obr. 79: Postup analýzy dat RNA-seq

Základní postuláty genové exprese

1. Každé buněčné jádro obsahuje kompletní genom vytvořený v oplodněném vajíčku. Z informačního hlediska je DNA ve všech diferencovaných buněk identická.
2. Nepoužité geny v diferencovaných buňkách si zachovávají potenciál pro expresi.
3. V každé buňce je exprimována pouze malá část genomu a tato část syntetizovaná v buňce je specifická pro daný typ buňky.

Cílem diferenciální exprese je zjištění, který/é z genů jsou exprimované rozdílné v rámci určitých podmínek. Tyto geny pak pomou nastínit biologický pohled na procesy ovlivněné danou podmínkou.

Diferenční (genová) exprese znamená provedení statistické analýzy dat ke zjištění kvantitativních změn exprese mezi dvěma podmínkami. Využíváme testování statistických hypotéz pro rozhodnutí, zda pozorovaný rozdíl v počtu čtení/genů je signifikantně větší nebo menší, než by byla náhodná změna.

10.3.1 Normalizace

Normalizace exprese je nezbytná pro odstranění technických zkreslení v sekvenovacích datech, které mohou být způsobeny například hloubkou prosekvenování a délkou genu. Normalizace může být v rámci vzorku (RPM, RPKM/FPKM, TPM), nebo mezi vzorky (TMM)

Sequencing depth: Vyhodnocení hloubky sekvenování je nezbytné pro porovnání genové exprese mezi vzorky. Při porovnání vzorku A proti vzorku B se může zdát, že se každý gen v A vyskytuje v 2x větší míře jak ve vzorku B, nicméně je to pouze důsledek 2x větší prosekvenovanosti vzorku A proti B.

Gene length: Stejně tak je pro diferenciální genovou expresi důležité vzít v potaz délku genů. Gen X a gen Y mají podobnou míru exprese, ale počet readů namapovaných na gen X může být mnohonásobně větší, než počet namapovaných readů pro gen Y, protože gen X je delší.

CPM/RPM (counts/reads per million, počet/čtení za milion): vydělení počtu čtení vlastnosti (genu) součtem počtu čtení v rámci celé knihovny, a vynásobením výsledků milionem (rovnice 17). Nezohledňuje délky genů.

$$CPM = \frac{počet\ (čtení\ genu) \times 10^6}{celkový\ počet\ (namapovaných\ čtení)} \tag{17}$$

RPK (Reads Per Kilobases, čtení na kilobázi): oproti RPM se získá dělením počtu čtení délkami genů (vyjádřenými v kilobázích).

RPKM/FPKM (reads/fragments per kilobase per million reads/fragments mapped, čtení/fragment na kilobázi na milion, rovnice (34)): RPKM je expresní jednotka normalizovaná na délku genu, která se používá pro identifikaci odlišně exprimovaných genů porovnáním hodnot RPKM mezi různými experimentálními podmínkami. Obecně platí, že čím vyšší je RPKM genu, tím vyšší je exprese tohoto genu. FPKM se používá zejména pro normalizaci počtů pro data RNA-seq s párovým koncem, ve kterých jsou obě (levá a pravá) čtení sekvenována ze stejného fragmentu DNA. Opět platí, že čím vyšší je FPKM genu, tím vyšší je exprese tohoto genu.

$$RPKM = \frac{počet\ (čtení\ genu) \times 10^3 \times 10^6}{celkový\ počet\ (namapovaných\ čtení) \times délka\ genu\ v\ bp} \tag{34}$$

TPM (transcripts per million): TPM je alternativa k RPKM. Vztah k RPKM vysvětluje rovnice (34) a (36), ve které platí, že součet TPM hodnot přes všechny geny je pro každý vzorek konstatní (10^6). Hodnota 10^6 řádově odpovídá počtu mRNA molekul v buňce. Pokud jsou všechna čtení stejně dlouhá, spočítáme TPM následovně (rovnice 37).

$$A = \frac{počet\ (čtení\ genu)}{délka\ genu\ v\ bp} \tag{35}$$

$$TPM = A \times \frac{1}{\sum(A)} \times 10^6 \tag{36}$$

Proporčně k RPKM

$$TPM = \frac{RPKM}{\sum RPKM} \times 10^6 \tag{37}$$

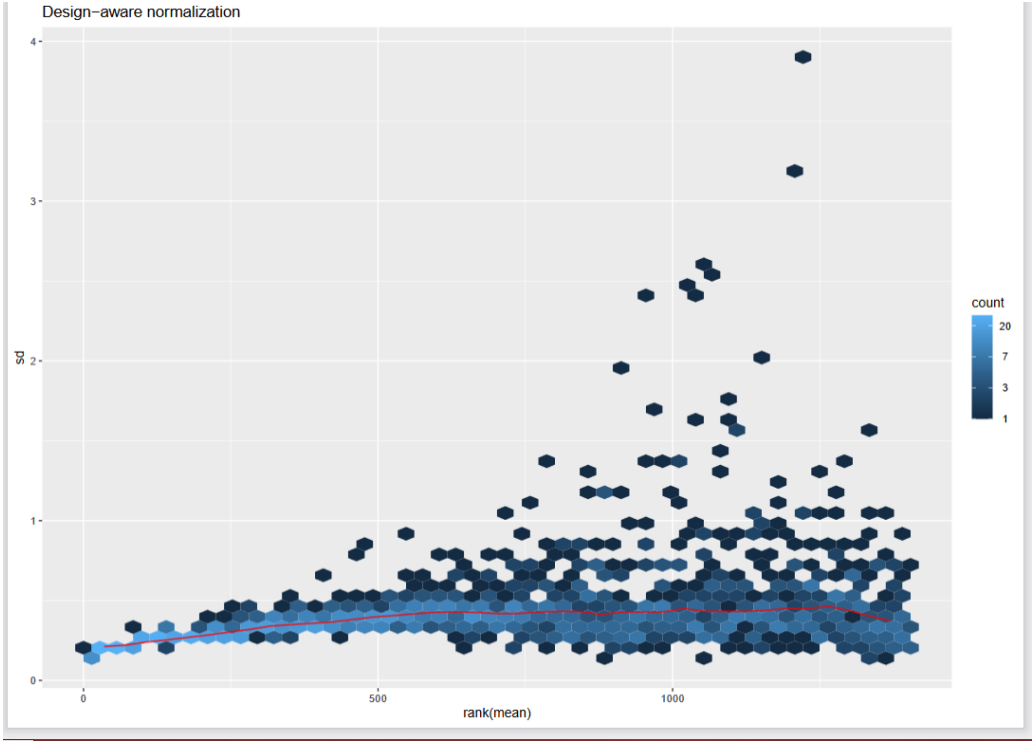
EdgeR's trimmed mean of M values (TMM): používá vážený ořezaný průměr poměrů logaritmických expresí mezi vzorky. TMM předpokládá (stejně jako ostatní metody), že většina genů není diferenčně exprimována. TMM normalizuje celkový výstup RNA mezi vzorky, pro normalizaci nebere v úvahu délku genu nebo velikost knihovny.

DESeq2's median of ratios: je podobná metodě TMM. Normalizace DESeq používá k výpočtu velikostních faktorů medián poměrů pozorovaných počtů. Stručně řečeno, faktor velikosti se vypočítá tak, že se pozorované počty pro každý vzorek vydělí jeho geometrickým průměrem. Velikostní faktor se pak vypočítá jako medián tohoto poměru pro každý vzorek. Tento faktor velikosti se pak použije pro normalizaci nezpracovaných dat počtu pro každý vzorek.

Opět se předpokládá, že většina genů není diferenčně exprimovaná, stejně tak se nebere v úvahu délka genu pro normalizaci, protože předpokládá, že délka genů by měla být mezi vzorky konstantní.

10.3.2 Stabilizace rozptylu

Stabilizace odchylek (transformace stabilizující rozptyl) je krokem v předzpracování dat, který může výrazně zpřesnit statistický model a následnou analýzu. Transformace jsou užitečné při kontrole odlehlých hodnot nebo jako vstup pro techniky strojového učení, jako je shlukování nebo lineární diskriminační analýza. Mnoho statistických metod pracujících s vícerozměrnými daty (typicky PCA, klastrování) pro správné použití vyžaduje homoskedastická data – data, jejichž rozptyl pozorovaných hodnot (jako je exprese genu) nezávisí na průměrné hodnotě. V RNA-seq datech rozptyl závisí na průměru, čím je průměr větší, tím roste i rozptyl. Balíček DESeq2 používá nástroje vst a rlog, které vypočítají transformaci stabilizující odchylku v log2 škále (Obr. 80). Zároveň však neodstraňuje variability, které mohou být spojené s batch efektem nebo s experimentálními proměnnými.



Obr. 80: Stabilizace rozptylu z R balíčku DESeq2, vlastní data

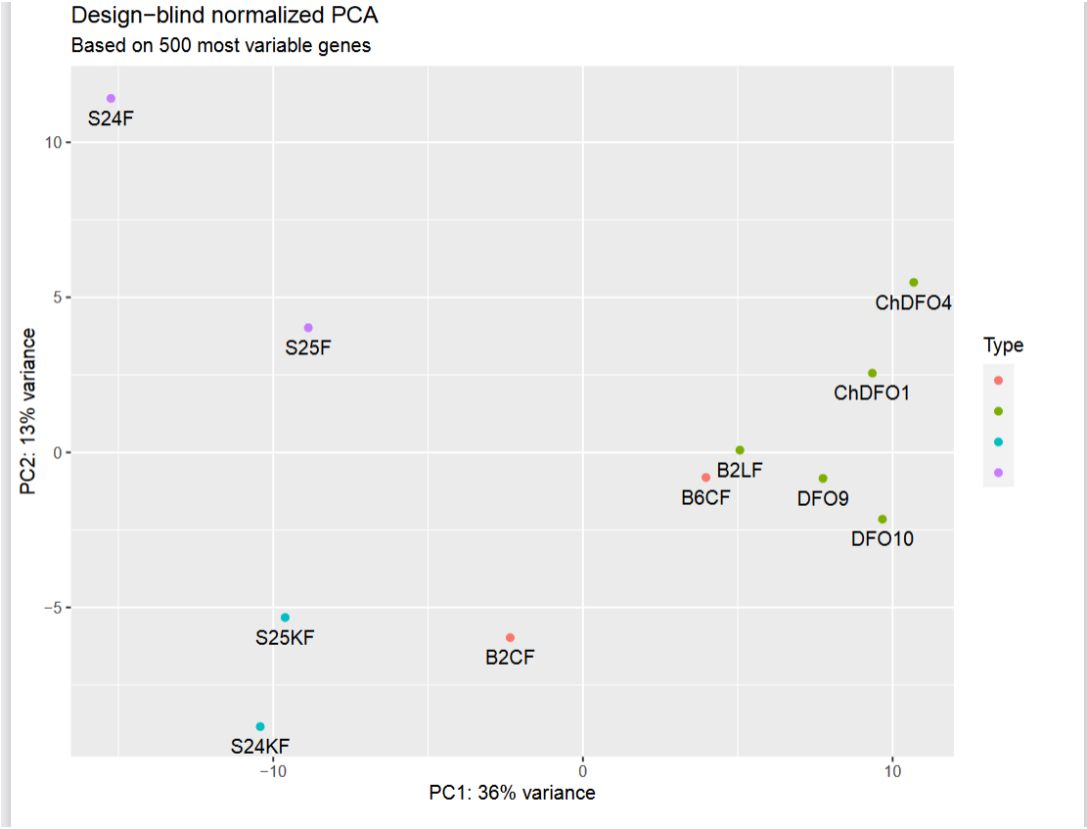
10.3.3 Batch effect

Jako batch effect označujeme faktory jiného než biologického původu, které mohou ovlivnit vyhodnocení experimentálních dat (např. použití jiných metod izolace u dvou skupin, časový posun v treatování kontrolních a měřených buněk…)

ComBat-Seq method: ComBat upravuje vstupní, předem normalizovaná (obecně se doporučuje log normalizace) count data porovnáním/normalizací kvantilů empirických distribucí dat s očekávanou distribucí bez batch efektů v datech. Pro úpravu/odstranění batch efektů z datasetů se používá buď parametrický, nebo neparametrický empirický Bayes. Výstupem je opět count matice s odstraněným batch efektem. ComBat se úspěšně využívá pro spojení dvou (nezávislých) datasetů spolu s umožněním jejich porovnání.

10.3.4 Kontrola kvality

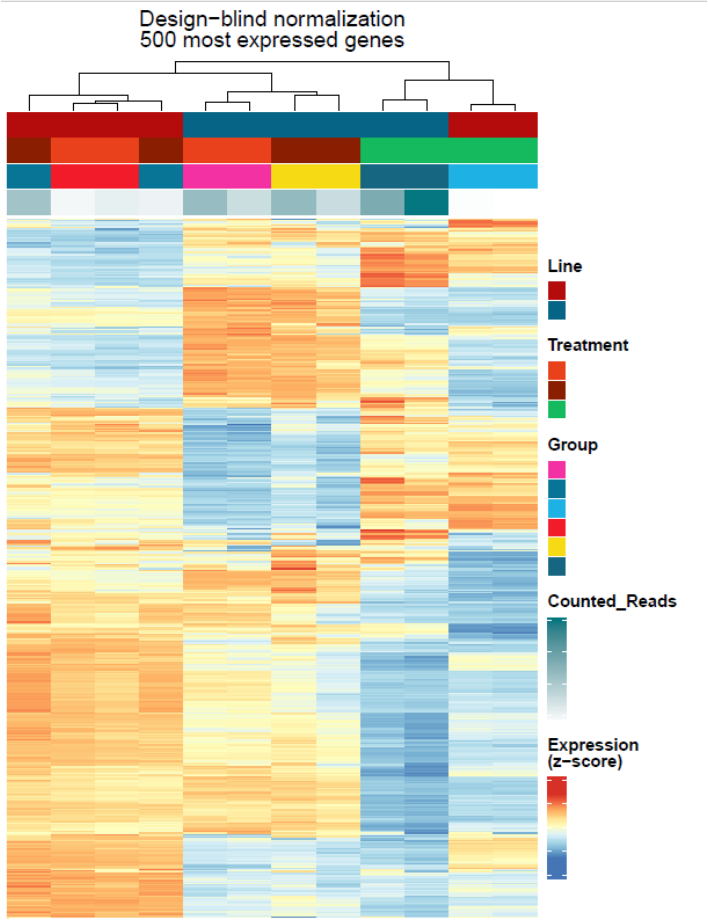
PCA (principal component analysis): Analýza hlavních komponent (PCA, Obr. 81) je technika redukce rozměrů, která najde směry největších změn v datové sadě/v rámci skupin dat a přiřadí je hlavním komponentám. Hlavní složkou (PC) vysvětlující největší množství variability v datové sadě je PC1, PC vysvětlující druhý největší podíl k variabilitě je PC2.



Obr. 81: Ukázka PCA, R balíček DESeq2, vlastní data

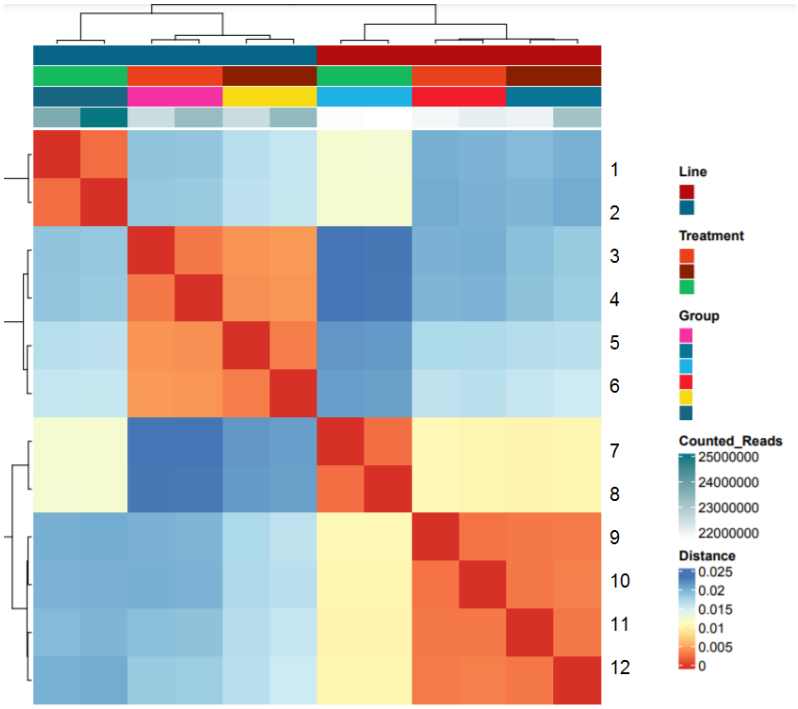
Heatmap: způsob zobrazení dat v mřížce, kde každý řádek představuje gen a každý sloupec představuje vzorek. Barva a intenzita polí se používá k reprezentaci genové exprese nebo jejich změn (Obr. 82).

Pro tvorbu heatmap nedoporučujeme používat R balíček Heatmap.2, který nesprávně používá a vizualizuje klastrování. Místo toho doporučujeme mnohem příjemnější balíček ComplexHeatmap.



Obr. 82: Ukázka heatmapy, R balíček ComplexHeatmap, vlastní data

Korelační Heatmap: zobrazuje korelaci genové exprese pro všechny párové kombinace vzorků v datové sadě (Obr. 83). Protože většina genů není odlišně exprimována, mají vzorky obecně vysoké vzájemné korelace (hodnoty vyšší než 0,80). Vzorky nižší než 0,80 mohou znamenat extrémní hodnotu v datech anebo ukazovat na kontaminaci vzorků. V korelační heatmapě se většinou uvádí hodnota 1-|cor|.



Obr. 83: Ukázka korelační heatmapy, R balíček ComplexHeatmap, vlastní data

10.3.5 Statistická analýza

Cílem tohoto kroku je získat pomocí statistického testování hypotéz na získaných datech na základě (lineárních) modelů informace o signifikantně deregulovaných (upregulovaných či downregulovaných, spolu s hodnotami log2 fold change spojené s další statistikou jako p-hodnota, korigovaná p-hodnota a další) genech/vlastnostech v experimentu. V případě RNA-Sequ využíváme pro statistické modelování přímo hrubé počty čtení, protože obsahují nejvíce informací.

Linerání modely

Modely využívané v kvantitativně genetických analýzách jsou ve většině případů lineární. Předpokládáme, že veličiny, které zkoumáme (nezávisle proměnná X a závisle proměnná Y) jsou spojité, anebo mohou být kódovány jako kvantitativní. Prvním krokem v regresní analýze by mělo být zakreslení jejich numerických hodnot do grafu. Tento krok nám může pomoci ukázat, zda existuje vztah mezi oběma proměnnými:

- zda existují trendy: rostou či klesají obě v jednom směru, nebo jedna klesá, když druhá roste
- zda je lineární závislost vhodným modelem pro vyjádření vztahu mezi těmito dvěma veličinami

Použití lineární modelů je u většiny biologických jevů buď dostačující nebo vynucené nedostatkem dat. Ačkoli nelineární modely mohou být důležité, často je možné je aproximovat zvoleným lineárním modelem. V tradičním smyslu jsou lineární modely složeny ze tří částí:

1. **Vlastní regresní model**: Rovnice modelu definuje efekty, které mohou mít vliv na pozorovanou vlastnost. Maticový zápis obecné modelové rovnice (38) je:

   $$\text{Link}(Y) = \beta X + \varepsilon, \quad (38)$$

   kde y je vektor pozorovaných hodnot vlastnosti, β je vektor pevných efektů, ε je vektor náhodných reziduálních efektů, X je nezávislá proměnná, link je nadefinovaná fuknce, která převádí měřené hodnoty na hodnoty, které můžeme modelovat (linker function)
2. **Daty**
3. **Očekávaným rozdělením reziduí,** $\varepsilon \in N(0, \sigma)$

Lineární regrese

Lineární regrese je matematická metoda, která zahrnuje proložení dat přimkou/rovinou/nadrovinou, a následnou analýzu statistických vlastností tohoto proložení, přičemž předpokládáme, že x-ové souřadnice jsou přesné, zatímco y-ové souřadnice mohou být zatíženy náhodnou chybou. Předpokládáme, že závislost y na x lze graficky vyjádřit přímkou/rovinou/nadrovinou takovou, že při odečítání z grafu bude mezi y-ovou hodnotou měřeného bodu a y-ovou hodnotou ležící na přímce odchylka. Podstatou lineární regrese je nalezení takové přímky, aby součet druhých mocnin těchto odchylek byl co nejmenší pomocí aproximaci daných hodnot přímky metodou nejmenších čtverců.

10.3.6 Testování statistických hypotéz

Statistická hypotéza je tvrzení, které lze či nelze zamítnout na základě experimentálních dat a statistických metod. Testováním hypotéz je možné posoudit vyjádřené předpoklady o (experimentálně) získaných datech. Častým typem experimentu je porovnání mezi skupinami (komparativní experiment), kdy jedna skupina je vystavena vlivu, který je zkoumán (např. vliv léčiva na buňku), druhá skupina je kontrolní bez daného pokusného zásahu. Experimenty je možné provádět mezi více skupinami. Efekt zkoumaného vlivu je číselná hodnota, u které předpokládáme náhodné rozdělení, na jejíž určení použijeme parametrické testy. Důležitým cílem statistických analýz je kvantitativní zjištění rozdílů mezi skupinami vzájemným porovnáním výběrových souborů vzorků v experimentu.

Testováním statistických hypotéz porovnáváme dvě hypotézy: nulovou hypotézu (H_0) a alternativní hypotézu (H_1, H_A, A, tvrzení, jehož platnost chceme testovat). Základním předpokladem je, že nulová hypotéza platí.

U testování hypotéz je nutné stanovit si hladinu spolehlivosti α, pravděpodobnost, že zamítneme nulovou hypotézu, i když platí (chyba I. druhy). Standardně se α stanovuje jako velmi malá hodnota, typická hodnota je 0,1, 0,05, v některých případech i 0,01. Nulová hypotéza je zamítnuta ve chvíli, kdy získaná p-hodnota, tj. pravděpodobnost, že pozorovaná, nebo lepší statistika, nastane, když je platná H_0, je menší než stanovená hladina významnosti α.

Chyby při testování hypotéz

Při testování hypotéz se můžeme dopustit jedné ze 2 chyb (Tab. 18):
chyba I. druhu α (false positive): zamítáme hypotézu H_0, když platí a neměli bychom ji zamítnout
chyba II. druhu β (false negative): nezamítáme hypotézu H_0, když neplatí a měli bychom ji zamítnout

Tab. 17: Chyby u testování hypotéz

		Náš výsledek	
		H_0 platí	**H_0 neplatí**
Skutečnost	**H_0 platí**	správný	chyba I. druhu
	H_0 neplatí	chyba II. druhu	správný

Snažíme se zvolit test tak, aby pravděpodobnost obou chyb byly co nejmenší, avšak takový test není jednoduché sestavit, jelikož chyby spolu souvisí: čím menší je hodnota α, tím větší je chyba β, a naopak. Zpravidla se volí dostatečně nízká hodnota α, chyba β je pak dána velikostí zvolené chyby α. Sílu testu (rozlišovací schopnost testu) definujeme jako pravděpodobnost 1-β. Síla testu určuje pravděpodobnost, že správně zamítneme nulovou hypotézu, když neplatí. S klesající hladinou významnosti α síla testu klesá.

Korekce mnohonásobných testů

Často se setkáme s nutností testovat více statistických hypotéz zároveň (stanovení jak a jestli se liší skupina 1 od 2, 2 od 3, …) S narůstajícím počtem testovaných hypotéz roste zároveň pravděpodobnost označení falešně pozitivního výsledku. To znamená, že se při testování zmýlíme, a vyhodnotíme jako statisticky významný rozdíl tam, kde ve skutečnosti žádný neexistuje. Nastavení hodnoty p pro mnohonásobné testy hypotéz se řídí chybou 1. druhu.

„Můžeme si představit modelovou situaci, kdy provedeme zároveň 60 testů, což v době běžného provádění biochemických a genetických experimentů není zase tolik. Použijeme-li standardní hladinu významnosti α = 0,05, máme pro každý test 5% riziko získání falešně pozitivního výsledku. Vynásobíme-li 60 a 0,05, vyjde nám, že zhruba u 3 testů bychom měli dospět k falešně statisticky významnému závěru. V případě genomických analýz, kde jsou často různé testy pouze formou explorativní a popisné analýzy, nemusí být přítomnost falešně pozitivních výsledků fatální, v klinické praxi to však může vést k zavádějícím výsledkům a mylným interpretacím. Z tohoto důvodu je nutné při násobném statistickém testování uvažovat tzv. ***korekční procedury*** *(correction for multiple testing), které by měly brát v úvahu celkový počet provedených testů.*“ Pavlík, T. *Biostatistika pro matematickou biologii - Problém násobného testování hypotéz* [online]; https://portal.matematickabiologie.cz/index.php?pg=aplikovana-analyza-klinickych-a-biologickych-dat--biostatistika-pro-matematickou-biologii--uvod-do-testovani-hypotez--poznamky-k-testovani-hypotez--problem-nasobneho-testovani-hypotez (accessed Dec 17, 2020).

Family-wise error rate (FWER): kontroluje pravděpodobnost vzniku alespoň jedné chyby prvního druhu. FWER je vhodné, pokud chceme zabránit jakémukoli falešně pozitivnímu výsledku, i za cenu toho, že některé skutečně změny označíme jako falešně negativní. V mnoha případech však určitý počet falešně pozitivních výsledků není velký problém. **Jeden z FWER přístupů je Bonferonniho korekce**: snižujeme hladinu významnosti dílčích testů tak, aby bylo dosaženo rozumné celkové hladiny významnosti. U Bonferonniho korekce zamítáme nulovou hypotézu, pokud je p-hodnota menší nebo rovna hodnotě α/m (α je zvolená hladina významnosti testu, m počet provedených testů).

False discovery rate (FDR): volíme očekávaný podíl falešně negativních nulových hypotéz pro danou množinu testů místo toho, abychom upravili globální hladiny statistické významnosti. Takto definovaná chyba je ekvivalentem ke globální hladině chyby prvního druhu pro daný experiment v případě, že všechny nulové hypotézy platí, ale je nižší v ostatních případech. Kontroluje očekávaný podíl falešně pozitivních (FP) výsledků mezi všemi zamítnutými hypotézami. FDR je méně přísné než FWER. Používáme ho, pokud chceme objevit co nejvíce skutečně pozitivních výsledků i za cenu několika falešně pozitivních, které udržíme v rozumném poměru oproti pravdivým pozitivním výsledkům. FDR je mnohem silnější než FWER vzhledem k chybě II. druhu. **Metoda FDR se také nazývá Benjamini-Hochbergova korekce.** Jednoduchý postup, k výpočtu FDR porovnává setříděné hodnoty p s přímkou se směrnicí α/m. Jednotlivé hypotézy jsou nejprve uspořádány a poté odmítnuty nebo přijaty na základě jejich p-hodnot. Pokud hypotéza s i-tou nejmenší p-hodnotou má p-hodnotu menší než $\frac{i}{m}\alpha$, je H_0 zamítnuta.

Problém násobného testování hypotéz (multiple testing problem) spočívá v tom, že s narůstajícím počtem testovaných hypotéz nám roste také pravděpodobnost získání falešně pozitivního výsledku, tedy pravděpodobnost toho, že se při našem testování zmýlíme a ukážeme na statisticky významný rozdíl tam, kde ve skutečnosti žádný neexistuje. Z tohoto důvodu je nutné při násobném statistickém testování uvažovat tzv. korekční procedury (correction for multiple testing), které by měly brát v úvahu celkový počet provedených testů.

Desing experimentu

Správně zvolený design experimentu je klíčový pro správné vyhodnocení analýzy. Experiment může být zatížen množstvím chyb a technických artefaktů, které negativně ovlivňují následné statistické vyhodnocení. Hlavní myšlenky návrhu experimentů jsou komparace (posouzení efektu intervence), randomizace (znáhodnění) a replikace (opakování).

Komparace: porovnání objektů vystavených zkoumanému jevu s kontrolní skupinou. Nejdříve se proměří všechny uvažované proměnné a poté změníme podmínky experimentu. Usuzujeme, jaký efekt měla změna podmínek na cílové proměnné. Pro spolehlivé porovnání výsledků je důležité k sobě řadit podobné vzorky (např. z hlediska pohlaví zkoumaných pacientů).

Randomizace: náhodné zařazení do skupin, zajistíme objektivitu zařazení do určité skupiny. Randomizace odstraňuje technické artefakty (vzniklé např. sekvenováním kontrol v dubnu, zatímco sekvenování vzorků až v červenci). Korelované technické artefakty (např. pokaždé izolujeme ovlivněné vzorky v pondělí a sekvenujeme v úterý, zatímco izolace kontrol probíhá ve středy a sekvenují se ve čtvrtky) mohou úplně znemožnit interpretaci výsledků experimentu.

Replikace: vícenásobné opakování měření umožní posoudit náhodnou variabilitu měřených veličin, a tím i určit přesnost měření, je nezbytná pro odhad experimentální chyby a směrodatné odchylky efektů pro potřeby testování. Správný experiment by měl být opakovatelný kvůli ověření výsledků. V případě replikace upřednostňujeme biologické replikáty před technickými. V každém případě by měly v každé skupině být minimálně tři biologické replikáty, aby bylo možné odhadnout odlehlé vzorky. Také závisí na použitém testu, v případě nutnosti použít neparametrický test by počet vzorků na skupinu měl být minimálně 10.

Biologické replikáty: z odlišných biologických vzorků

Technické replikáty: odvozeny ze stejného výchozího biologického vzorku, rozdíl mezi dvěma vzorky je spíše v použité technice, než v biologii

10.3.7 Nástroje

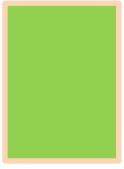
Nástroje pro zpracování RNA-Seq ve statistickém prostředí R.

DESeq2

Tento nástroj bere jako vstup count table spolu s popisem vzorků a informacemi o skupinách. DESeq2 provádí vnitřní normalizaci, kde se pro každý gen ve všech vzorcích vypočítá geometrický průměr. Počty čtení pro každý gen se poté vydělí tímto průměrem. Medián těchto poměrů přese všechny geny ve vzorku je faktor velikosti pro tento vzorek. Tento postup opravuje zkreslení velikosti knihovny a složení RNA, ke kterému může dojít například v případě, že je malý počet genů vysoce exprimován v jedné podmínce experimentu, ale ne v ostatních. Pro odhad disperze a fold changes využívá statistické smrštění (shrinkage estimation). Hodnota disperze se odhaduje pro každý jednotlivý gen a upravuje se fitováním modelu na základě informací i od ostatních genů. Pro správný odhad disperze jsou zapotřebí biologické replikáty ve všech podmínkách experimentu. DESeq2 používá negativní binomický generalizovaný lineární model pro každý gen, pro testování významnosti používá Waldův test a pro odhadnutí a odstranění odlehlých vzorků Cooperovu vzdálenost. Také automaticky odstraní geny, jejichž průměr normalizovaných počtů je pod prahovou hodnotou určenou optimalizačním postupem. Odstranění těchto genů s nízkými počty zlepšuje detekční schopnost tím, že snižuje počet testovaných genů a tak zvyšuje sílu experimentu (viz. Multiple testing issue).

edgeR

Stejně jako DESeq2, edgeR bere jako vstup count table, spolu s popisy vzorků a přidanými informacemi o skupinách v experimentu. edgeR využívá jako normalizaci trimmed mean M-hodnot (TMM), která se podobně jako u DESeq2 používá k výpočtu normalizačních faktorů tak, aby došlo k redukci vlivu složení RNA na experiment ve chvíli, kdy je malý počet genů velmi vysoce exprimován v jedné podmínce experimentu, ale ne v ostatních. Disperze je určena pomocí Cox-Reid profile-adjusted podobnostní metody (CR). Stejně jako DESeq2 používá negativní binomický generalizovaný lineární model pro každý gen, po nafitování dat na model se pomocí F-testu kvazi-věrohodnosti (QL) stanoví diferenciálně exprimované geny. Statistická analýza k identifikaci odlišně exprimovaných vlastností (geny, miRNA, ...) se provádí pomocí vícerozměrného regresního modelu. Pro model lze zadat maximálně tři různé proměnné a jejich interakce. Je doporučeno mít vždy biologické replikáty pro každou podmínku experimentu, ale edgeR, narozdíl od DESeq2, je schopen pracovat i bez biologických replikátů. edgeR také neprovádí filtrování výsledků ve výsledkové tabulce. Tabulku je možné dodatečně filtrovat pomocí jiných nástrojů.



DESeq2 a edgeR jsou určené na count data, na počty vlastností z HTS experimentů jako je RNA-seq, Limma je užitečná hlavně na zpracování spojitých dat, jako jsou intenzity čipových dat, případně proteomická data. Nicméně při využití limma-voom je možné zpracovávat i count data.

limma

Narozdíl od nástrojů DESeq2 a edgeR, které se využívají na zpracování HTS dat (RNA-seq, ChIP-Seq, obecně count table s hodnotami počtu výskytu dané informace) se Limma používá **spíše** na vyhodnocení **intenzit** čipových dat (v log2 škále). Jako normalizaci používá metody podle zvolených čipů (single color, two color), buď normalizaci mezi skupinami nebo v rámci skupin. Základní statistikou používanou pro analýzu významnosti je moderovaná t-statistika, která se počítá pro každou sondu a pro každý kontrast s použitím jednoduchého Bayesovského modelu. Moderované t-statistiky vedou nejvyšší hodnoty stejným způsobem jako běžné t-statistiky, kromě toho, že se zvyšují stupně volnosti, což odráží větší spolehlivost spojenou s vyhlazenými standardními chybami.

limma voom

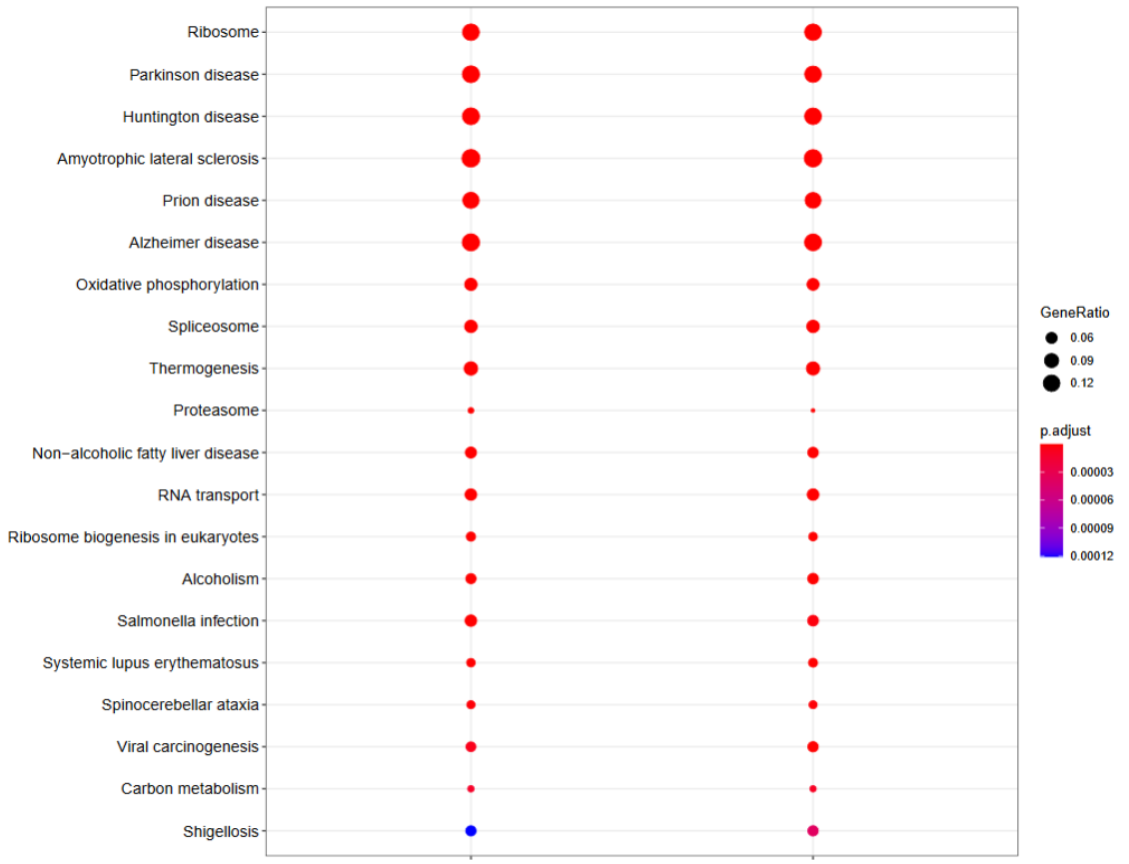
RNA-seq data mohou být zpracována v limma s využitím metody voom (variance modeling at the observational level). Hodnoty v count table jsou transformovány na log2 počty per million reads (CPM), kde „per million reads" je definováno na základě dříve vypočtených normalizačních faktorů. Pro každou log2 CPM hodnotu pro každý gen je nafitován lineár model, a jsou vypočteny residuals. Na odmocninu standartní odchylky reziduálů je nafitována hladká křivka na základě průměrných expresí. Hladká křivka se používá k získání vah pro každý gen a vzorek, které jsou předány do limmy spolu s log2 CPM hodnotami.

Výsledky statistické analýzy dat, převážně informace o log fold change spolu s p hodnotami (adjustive p-val) se finálně využijí ve Funkční analýze dat.

10.4 Funkční analýza dat

Funkční analýza dat zodpovídá otázku, co znamenají změny v expresi genů na biologické úrovni? Protože jsou geny, respektive proteiny, vzájemně propojené do změti drah (signálních, metabolických, ...), jakákoliv odchylka v deregulaci genů může být příčinou ke vzniku onemocnění. Znalost změn na úrovni drah může zrychlit a zpřesnit diagnostiku určitých onemocněních a poskytnout klíč k nalezení správné terapie.
Existuje dvojí řešení pro funkční analýzu dat:

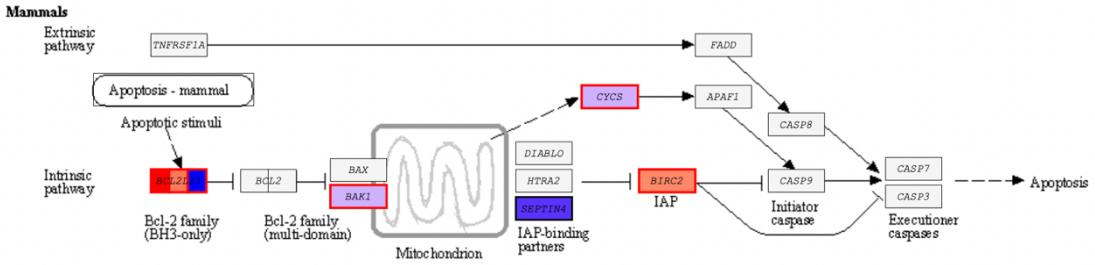
1. Geny, které najdeme deregulované mezi skupinami můžeme vložit do databáze a podívat se kam patří (KEGG, MsigDB, ...), nevýhodou však je to, že nemáme informace o statistické významnosti
2. Můžeme však také porovnávat všechny geny se skupinami genů v jednotlivých drahách s pomocí statistických přístupů, které se obecně nazývají metody analýzy genových sad (**GSEA, Gene Set Enrichment Analysis**). V této metodě pracujeme již s definovanými skupinami genů pro jednotlivé dráhy (Obr. 84).



Obr. 84: Ukázka signifikantních KEGG výrazů ve dvou vzorcích, R balíček ClusterProfiler, vlastní data

10.4.1 Biologické, signální a metabolické dráhy

Biologická dráha je specifický popis biologické funkce, jedná se o řadu interakcí mezi molekulami, které vedou k určitému produktu nebo změně v buňce. Signální dráhy (Obr. 85) jsou sekvence interakcí, které umožňují buňce přijmout signál (informaci vedoucí od jedné buňky ke druhé, změny v prostředí) a biologicky na něj reagovat. Informace se mezi buňkami se šíří pomocí signálních molekul (např. hormony, cytokiny, růstové faktory, malé molekuly jako NO a CO, ...), které tvoří vazbu se specifickými receptorovými molekulami na povrchu buňky (např. GPCR: G-protein–coupled receptor), které umožňují signál přijmout a odpovědět na něj.
Vazbou signální molekuly na receptor dojde na molekulách receptoru ke strukturním změnám. Tato změna umožní navodit sérii reakcí uvnitř buňky (signální transdukce), které v konečném důsledku vyústí v alteraci určité buněčné činnosti.
V biochemii jako metabolické dráhy označujeme metabolické procesy, což jsou sledy enzymy řízených reakcí vyskytujících se v buňce, které vedou k tvorbě konkrétních produktů. Tyto dráhy jsou popisovány enzymy a metabolity. Dráha také může popisovat genetické regulační sítě nebo proteinové kaskády. Ty obsahují informaci o spojení mezi geny, možných důsledcích a směrech efektů.



Obr. 85: Signální dráha apoptózy (zjednodušená), vlastní data, balíček keggsvg J. Novotný

10.4.2 GSEA

Gene Set Enrichment Analysis (GSEA) je metoda, která vyhodnocuje data na úrovni genových sad. Jednotlivé geny jsou seřazeny do genových sad na základě dřívějších biologických znalostí, jako jsou publikované informace o biochemických drahách. Jednolitvé sady se řadí v rámci seznamu podle exprese genů. Cílem metody GSEA je určit, které geny/genové sady se mají tendenci vyskytovat nahoře nebo dole v seřazeném seznamu. V takovém případě je sada genů korelována s rozdílem ve fenotypu mezi danými skupinami (Obr. 86). Pokud genová sada spadá nahoru, jedná se o upregulovaný set, pokud dolů, tak dochází ke snížené/nulové expresi genů v daném setu.
Kroky GSEA:

1. **Výpočet skore nabohacení:** vypočítá se skóre nabohacení (ES, enrichment score), které odráží míru, na kterou je set S nadhodnocený v extrémech (nahoře nebo dole) na celém seznamu. Skóre se vypočítá procházením seznamu a navýšením průběžné statistiky součtu, když narazíme na gen v S a jeho snížením ve chvíli kdy narazíme na geny, které nejsou v S. Velikost přírůstku závisí na korelaci genu s fenotypem. Enrichment score je maximální výchylka od nuly zjištěná náhodným procházením, a zhruba odpovídá Kolmogorově–Smirnově statistice.
2. **Odhadnutí míry signifikantnosti (významnosti) ES:** odhadujeme míru významnosti (*P* value) ES použitím empirického permutačního testu založeném na fenotypu, který zachovává komplexní korelační strukturu dat genové exprese. Konkrétně permutujeme fenotypové značky a přepočítáváme ES sady genů pro permutovaná data, což generuje nulovou distribuci pro ES. Empirická nominální hodnota P pozorovaného ES se poté vypočítá vzhledem k této nulové distribuci. Důležité je, že permutace značení tříd zachovává mezigenové korelace, a tak poskytuje biologicky přijatelnější hodnocení významnosti, než jaké by bylo dosaženo permutací genů.
3. **Úprava pro mnohonásobné testování hypotéz:** po vyhodnocení celé databáze genových sad, upravíme odhadovanou hladinu významnosti tak, aby zohledňovala mnohonásobné testování hypotéz. Nejprve normalizujeme ES pro každou sadu genů na základě velikosti sad, čímž získáme normalizované skóre obohacení (NES). Poté kontrolujeme podíl falešně pozitivních výsledků výpočtem míry falešných objevů (FDR) odpovídající každému NES.

Jaký je rozdíl mezi genovou sadou a signální dráhou?

Signální dráhy jsou sekvence interakcí, které umožňují buňce přijmout signál.
Oproti tomu genová sada je jakákoliv množina genů, například všechny geny patřící do jedné dráhy ale třeba i všechny geny, které mají podobnou funkci.
Genová sada tedy není signální dráha; jedná se o mnohem všeobecnější a méně specifický pojem

Rank	Type	ID	Description	GeneRatio	BgRatio	OddsRatio	pvalue	p.adjust	qvalue	Count
1	1	GO:0003735	structural constituent of ribosome	133/1628	202/17696	7.16	2.76e-88	2.47e-85	1.94e-85	133
2	1	GO:0045296	cadherin binding	122/1628	331/17696	4.01	8.13e-44	3.65e-41	2.87e-41	122
3	1	GO:0050839	cell adhesion molecule binding	132/1628	499/17696	2.88	4.55e-30	1.36e-27	1.07e-27	132
4	1	GO:0045182	translation regulator activity	56/1628	141/17696	4.32	1.53e-22	3.43e-20	2.70e-20	56
5	1	GO:0003954	NADH dehydrogenase activity	29/1628	46/17696	6.85	2.62e-19	3.35e-17	2.64e-17	29
6	1	GO:0008137	NADH dehydrogenase (ubiquinone) activity	29/1628	46/17696	6.85	2.62e-19	3.35e-17	2.64e-17	29
7	1	GO:0050136	NADH dehydrogenase (quinone) activity	29/1628	46/17696	6.85	2.62e-19	3.35e-17	2.64e-17	29
8	1	GO:0016655	oxidoreductase activity, acting on NAD(P)H, quinone or similar compound as acceptor	32/1628	60/17696	5.80	4.18e-18	4.69e-16	3.69e-16	32
9	1	GO:0090079	translation regulator activity, nucleic acid binding	42/1628	109/17696	4.19	1.11e-16	1.11e-14	8.71e-15	42
10	1	GO:0009055	electron transfer activity	42/1628	114/17696	4.00	7.37e-16	6.61e-14	5.20e-14	42

Showing 1 to 10 of 151 entries ← Previous 1 2 3 4 5 Next →

Obr. 86: Ukázka možných výsledků získaných metodou GSEA

10.4.3 Databáze genových sad

Databáze genových sad jsou zdroje, které shromažďují a vizualizují získané znalosti z medicíny a biologických věd, shromážděné podle charakteristik do skupin např. podle povahy dráhy (metabolická, signální) nebo interakcí (protein-protein interakce).
Máme dva způsoby konstrukce dráhy:

- **Na základě dat**: tato dráha je využívaná ke studiu vztahů mezi geny nebo proteiny identifikovanými v konkrétním experimentu, jakým je například mikročipová studie.
- **Na základě znalostí**: tato dráha vyžaduje vytvoření podrobné znalostní databáze o biologických dráhách jednotlivých domén, jakými jsou typ buňky, nemoci nebo systému.

Genové ontologie

Ontologie jsou kontrolované slovníky (obsahují termy či klíčová slova), které definují základní pojmy a vztahy oblasti tématu, ke kterému náležejí. Gene Ontology (GO) začalo jako konsorcium roku 1998, kdy se vědci, kteří studovali genom tří modelových organismů – Drosophila melanogaster (octomilka), Mus musculus (myš) a Saccharomyces cerevisiae (pivovarské/pekařské kvasnice), dohodli na společné práci na klasifikačním schématu pro funkci genů. V současné době se počet různých organismů zastoupených v GO počítá v tisících. GO umožňuje flexibilním a dynamickým způsobem poskytnout srovnatelné popisy homologních genových a proteinových sekvencí napříč fylogenetickým spektrem, a zvláště pokud studované organismy studované geny zdědily od společného předka.
GO nabízí dva hlavní zdroje, **GO samotné a GO Annotations**.

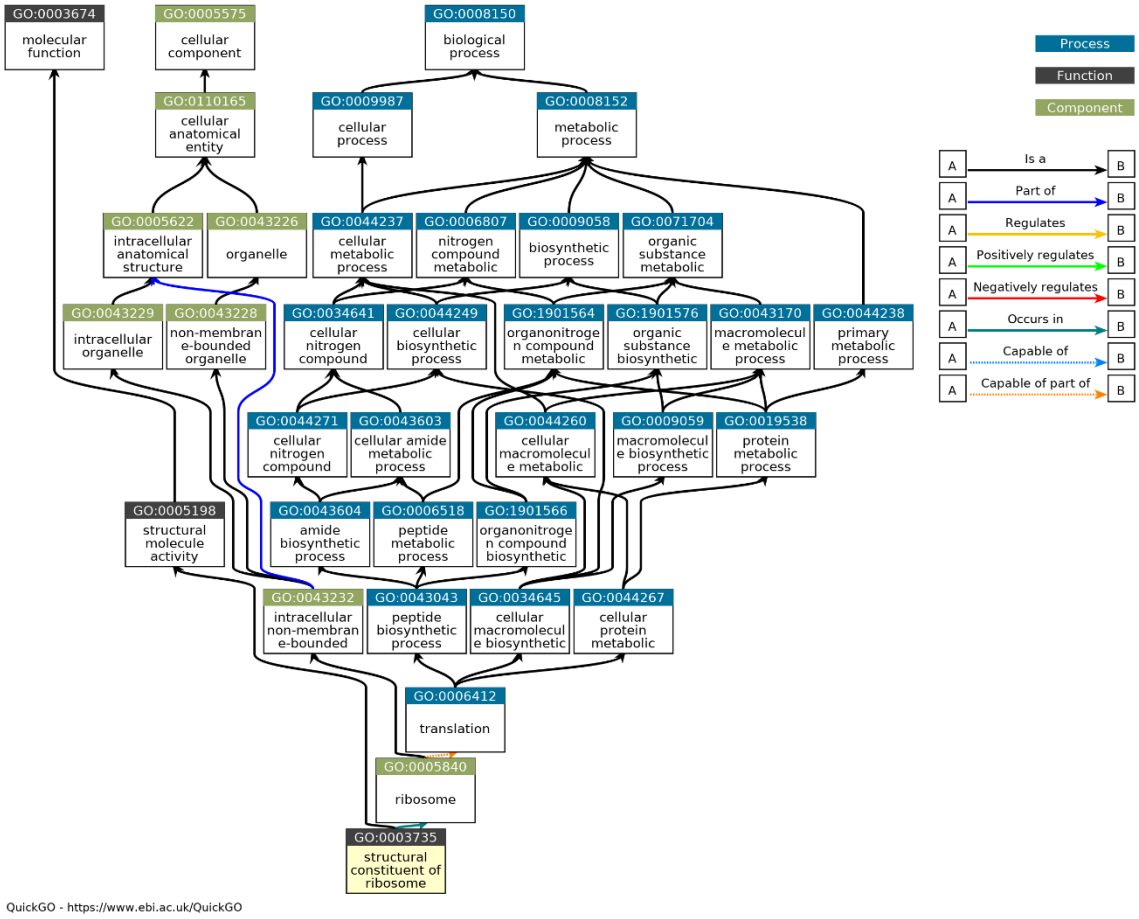
GO

GO ontologie jsou strukturovány jako orientovaný acyklický graf, přičemž GO termy jsou uzly v grafu a vztahy (vlastnosti objektu, **is a**, **part of**, **has part**, **regulates**) jsou hrany mezi termy. Rodičovské uzly jsou obecnější termíny, naopak dceřiné uzly jsou více specifické. Rodičovský uzel obsahuje všechny geny a vlastnosti svého potomka, ale může obsahovat také geny, které u potomka nejsou.

Gene Ontology (GO) popisuje naše znalosti biologické domény s ohledem na tři aspekty (Obr. 87):

- **Biologické procesy (Biological Processes):** Větších procesů neboli biologických programů se dosahuje spřažením více molekulárních aktivit, jako například oprava DNA nebo signální transdukce. Biologický proces tak, jak je popsán v GO však není ekvivalentem k pathways, v současné době se GO nepokouší reprezentovat dynamiku nebo závislosti, které by byly nutné k úplnému popisu cesty.
- **Buněčné komponenty (Cellular Component)**: Umístění relativně k buněčným strukturám, ve kterých genový produkt vykonává funkci, buď buněčné kompartmenty (např. celá mitochondrie), nebo stabilní makromolekulární komplexy, jejichž jsou součástí (např. ribosom). Na rozdíl od ostatních aspektů GO se třídy buněčných komponent nevztahují na procesy, ale spíše na buněčnou anatomii.
- **Molekulární funkce (Molecular function)**: Činnosti na molekulární úrovni prováděné genovými produkty. Pojmy molekulární funkce popisují činnosti, které se vyskytují na molekulární úrovni (katalýza, transport). Termíny GO molekulární funkce představují aktivity spíše než entity (molekuly nebo komplexy), které provádějí akce, a nespecifikují, kde, kdy nebo v jakém kontextu se akce odehrává. Molekulární funkce obecně odpovídají aktivitám, které mohou být prováděny

jednotlivými genovými produkty (tj. proteinem nebo RNA), ale některé aktivity jsou prováděny molekulárními komplexy složenými z více genových produktů. Příklady širokých funkčních termínů jsou katalytická aktivita a aktivita transportéru; příklady užších funkčních výrazů jsou aktivita adenylátcyklázy nebo vazba na Toll receptor. Aby se zabránilo záměně mezi názvy genových produktů a jejich molekulárními funkcemi, jsou molekulární funkce GO často spojovány se slovem aktivita (proteinová kináza by měla aktivitu proteinové kinázy GO molekulární funkce).



Obr. 87: Příklad anotace GO: buněčný komponent ribosom lze zároveň popsat biologickým procesem translace a molekulární funkcí struktrutní složky ribosomu (akce molekul, které přispívají k integritě ribosomu), QuickGo::Term GO:0005840. QuickGO. https://www.ebi.ac.uk/QuickGO/term/GO:0005840 (accessed Dec 27, 2020).

GO Annotations

Určují funkce konkrétního genu. GO Annotations se vytvářejí přidružením genu či produktu genu ke GO termu. Společně tato tvrzení představují „snímek“ současných biologických znalostí. Anotace GO proto zachycují výroky o tom, jak gen funguje na molekulární úrovni, kde v buňce funguje a jaké biologické procesy pomáhá provádět. Genový produkt může být z každé ontologie anotován ničím nebo více výrazy. Každá anotace je podporována GO Evidence Codes z Evidence a Concepts Ontology a referencí. Genové produkty jsou anotovány nejpodrobnějším pojmem v ontologii, který je podporován dostupnými důkazy. Podle principu přechodnosti anotace k výrazu GO implikuje anotaci všem jeho rodičům. Anotace GO mají odrážet nejaktuálnější pohled na roli genového produktu v biologii. Protože se mění biologické znalosti, mohou se anotace pro daný genový produkt měnit, aby odrážely změny ve znalostech a/nebo změny v ontologii. Pokud je genový produkt bez poznámek, jeho role je stále neznámá.

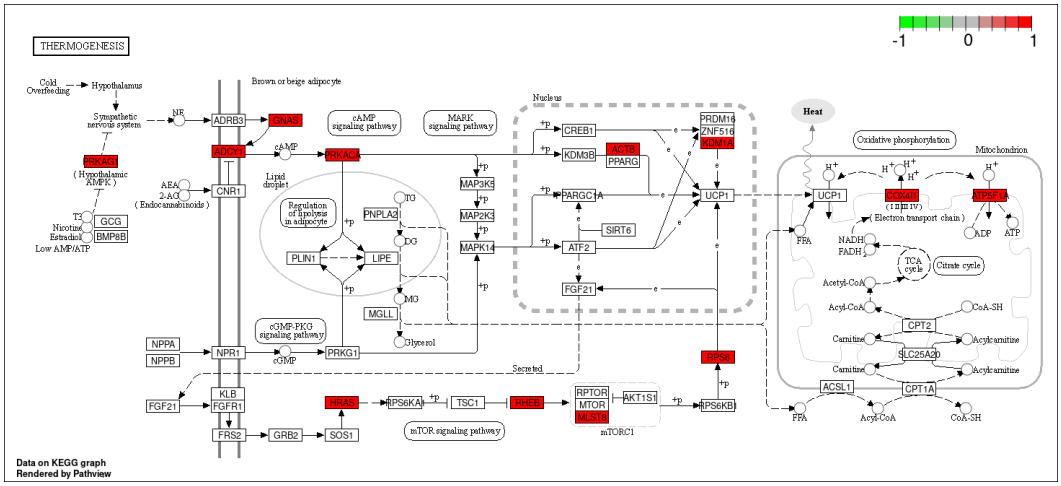
KEGG

KEGG (Kyoto Encyclopedia of Genes and Genomes) je databáze znalostí sloužící k systematické analýze funkcí a vlastností velkých biologických systémů, jako jsou buňky, organismy a celé ekosystémy, na molekulární úrovni. Jedná se o počítačovou reprezentaci biologického systému, který se skládá z molekulárních stavebních bloků genů a proteinů (genomová informace) a chemických látek (chemická informace), které jsou integrovány se znalostmi o molekulárních schématech zapojení interakčních, reakčních a relačních sítí (systémové informace). Informace o vlastnostech byly převážně získány z molekulárních datových souborů generovaných sekvenováním genomu a dalšími vysoce výkonnými experimentální technologiemi.
V současné době se integrovaná databáze KEGG skládá z 18 jednotlivých databází. Unikátním objektem KEGG databáze jsou molekulární sítě (Obr. 88 a 89), které pomocí interakčních, reakčních a relačních sítí představují systémové funkce buňky a organismu. Vizualizace sítí a popisované vztahy jsou uspořádány v následujících databázích:

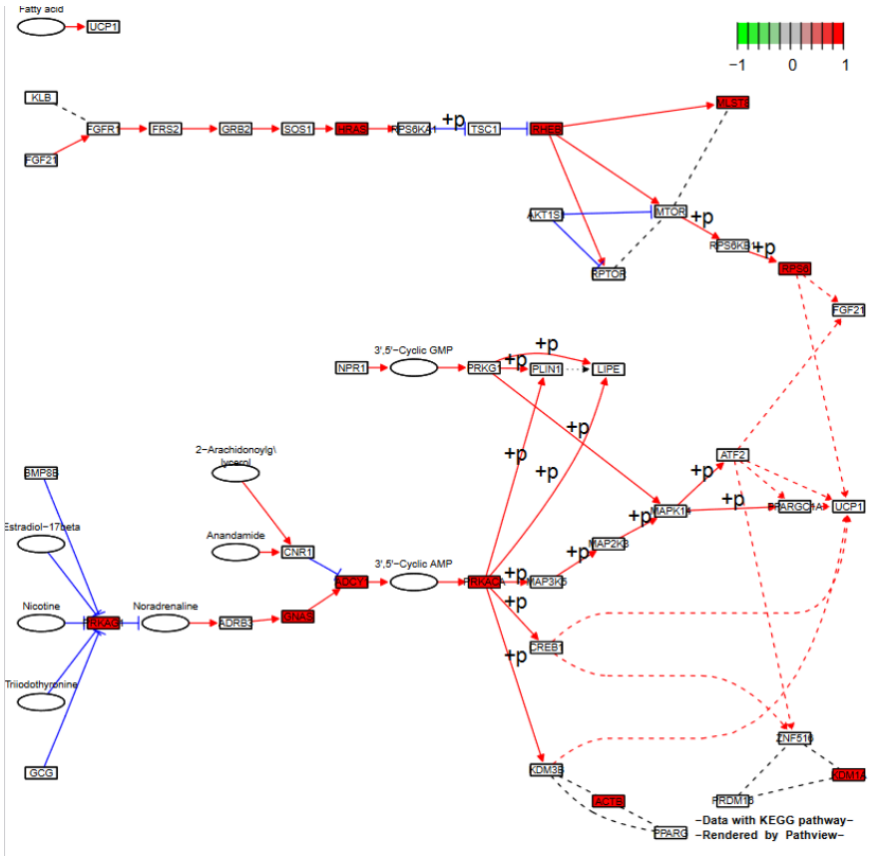
- Mapa sítí: KEGG PATHWAY
- Hierarchie a tabulky: KEGG BRITE
- Členství (logické výrazy): KEGG MODULE
- Členství (jednoduchý seznam): KEGG DISEASE

Zobrazení metabolických a signálních cest je jednou z nejdůležitějších částí KEGG databáze. Mapa drah (KEGG Pathway) se skládá z diagramů, obsahujících výrazy v Kegg ontologii, které zobrazují interakce/reakce spřažené v orientovaných sítích. Díky tomu lze experimentální důkazy ve specifických organismech zobecnit pro využití jinými organismy prostřednictvím genomové informace. Základní referenční mapa je ručně kreslená, ostatní mapy drah (např. ty specifické pro organismus) jsou generovány výpočetně.

KEGG je veřejně dostupná databáze, některé části jsou ale dostupné pouze v placené verzi, jako například stáhnutí celé databáze.



Obr. 88: Změněné geny v KEGG dráze Thermogenesis, R balíček ClusterProfiler, vlastní data



Obr. 89: Změněné geny v KEGG dráze Thermogenesis, Graphviz view, R balíček Pathview, vlastní data

WikiPathways

WikiPathways je databáze založená na dynamickém přístupu podobném Wikipedii, čímž zůstává ve spojení s neustále se rozvíjejícími biologickými znalostmi v literatuře. Jejich cílem je dát dohromady definice drah, které jsou obecně uznávané vědeckou komunitou. Zobrazení drah je jednodušší než v případě KEGG Pathways, což umožňuje snazší úpravu, na druhou stranu dráhy mají neinteraktivní formu statických obrázků, a nelze se tedy přímo z dráhy odkázat na konkrétní gen a získat tak snadno dodatečné informace. Každá dráha uložená ve WikiPathways má ale vyhrazenou svoji wiki stránku zobrazující proudový diagram, popis, reference, vývoj změn a dílčí seznamy genů, proteinů a metabolitů.

MSigDB (Molecular Signature Database)

Databáze MSigDB je spojena s metodou GSEA. Místo grafického znázornění drah (jako u KEGG databáze) používá obecnější pojetí genových sad. Genové sady dělí do pěti hlavních kategorií:

C1 (poziční genové sady): sady reprezentované geny na stejném chromosomu nebo jeho části
C2 (vytvořené genové sady): z online databází drah, publikací v PubMed a znalostí expertů
C3 (tematické genové sady): založené na konzervativních cis-regulačních tématech ze srovnávacích analýz genomů člověka, myši, krysy a psa
C4 (vypočítané genové sady): definované expresí sousedících genů, které jsou spjaty s rakovinou)
C5 (GO genové sady): složené z genů anotovaných stejnými GO-termíny

Uživatel si také může definovat vlastní genové sady týkající se jeho zájmového procesu nebo fenotypu.

10.5 Otázky k tématu

1. Popište amplifikační křivku PCR. Co je cDNA?
2. Co jsou referenční geny? Na co se využívají?
3. Co je to log2 fold change? Co znamená LFC = -2?
4. Co je RNA-seq a k čemu se využívá?
5. Co je count table?
6. Proč je důležité třídění a indexování BAM souborů?
7. Co je lineární regrese?
8. Jaký je postup zpracování RNA-seq dat?
9. Jaké druhy normalizace znáte?
10. Co je genová exprese?
11. Jak se od sebe liší DESeq2 a Limma? Jak a k čemu se využívají?

10.6 Zdroje

Kvantifikace qPCR a qRT-PCR

Bachman J. Reverse-transcription PCR (RT-PCR). Methods Enzymol. 2013;530:67-74. doi:10.1016/B978-0-12-420037-1.00002-6

Boulter, N., Suarez, F.G., Schibeci, S. et al. A simple, accurate and universal method for quantification of PCR. BMC Biotechnol 16, 27 (2016). https://doi.org/10.1186/s12896-016-0256-y

Clark, D. P.; Pazdernik, N. J.; McGehee, M. R. *Molecular biology Chapter 6 - Polymerase Chain Reaction*, 3rd ed. [online]; Academic Cell, 2018; pp 168–198. https://www.sciencedirect.com/science/article/pii/B9780128132883000069?via%3Dihub (accessed Dec 28, 2020).

Dheda K, Huggett JF, Bustin SA, Johnson MA, Rook G, Zumla A. Validation of housekeeping genes for normalizing RNA expression in real-time PCR. Biotechniques. 2004;37(1):112-119. doi:10.2144/04371RR03

Didenko VV. DNA probes using fluorescence resonance energy transfer (FRET): designs and applications. Biotechniques. 2001;31(5):1106-1121. doi:10.2144/01315rv02

Essentials of Real-Time PCR. ThermoFisher Scientific. https://www.thermofisher.com/cz/en/home/life-science/pcr/real-time-pcr/real-time-pcr-learning-center/real-time-pcr-basics/essentials-real-time-pcr.html (accessed Dec 09, 2020).

Elsayed S, Plewes K, Church D, Chow B, Zhang K. Use of molecular beacon probes for real-time PCR detection of Plasmodium falciparum and other plasmodium species in peripheral blood specimens. J Clin Microbiol. 2006;44(2):622-624. doi:10.1128/JCM.44.2.622-624.2006

Freeman WM, Walker SJ, Vrana KE. Quantitative RT-PCR: pitfalls and potential. Biotechniques. 1999;26(1):112-125. doi:10.2144/99261rv01

Khan-Malek R, Wang Y. Statistical Analysis of Quantitative RT-PCR Results. Methods Mol Biol. 2017;1641:281-296. doi:10.1007/978-1-4939-7172-5_15

History of PCR Discoveries. Sigma Aldrich. https://www.sigmaaldrich.com/technical-documents/articles/biology/pcr-introduction-and-historical-timelines.html?gclid=CjwKCAiAtej9BRAvEiwA0UAWXic8a1iR2sR_9yHoqJ6xURgpH7dHdktjPGwwXQRBFCiINkA41K0QkBoCWEUQAvD_BwE (accessed Dec 09, 2020).

LabGuide.cz - Průvodce laboratoří. https://labguide.cz/ (accessed Dec 27, 2020).

Pfaffl, M. W. Relative quantification. Real-Time PCR: Current Technology and Applications, 1.st ed.; Caister Academic Press, 2009; Chapter 3, pp 63–82.

Yuan, J.S., Reed, A., Chen, F. et al. Statistical analysis of real-time PCR data. BMC Bioinformatics 7, 85 (2006). https://doi.org/10.1186/1471-2105-7-85

TaqMan. Science Direct. https://www.sciencedirect.com/topics/biochemistry-genetics-and-molecular-biology/taqman (accessed Dec 09, 2020).

RNA-seq

Anders S, Pyl PT, Huber W. HTSeq--a Python framework to work with high-throughput sequencing data. *Bioinformatics*. 2015;31(2):166-169. doi:10.1093/bioinformatics/btu638
Bendre, R. Gene expression units explained: RPM, RPKM, FPKM, TPM, DESeq, TMM, SCnorm, GeTMM, and ComBat-Seq. Data science blog. https://www.reneshbedre.com/blog/expression_units.html (accessed Feb 04, 2022).
Ewels, P.A., Peltzer, A., Fillinger, S. *et al.* The nf-core framework for community-curated bioinformatics pipelines. *Nat Biotechnol* **38,** 276–278 (2020). https://doi.org/10.1038/s41587-020-0439-x
Fabio, Z.; et al. HTSeq: Analysing high-throughput sequencing data with Python. HTSeq 0.13.1 documentation. https://htseq.readthedocs.io/en/master/overview.html (accessed Dec 28, 2020).
Li, B.; et al. RSEM: accurate quantification of gene and isoform expression from RNA-seq data. https://github.com/deweylab/RSEM#acknowledgements (accessed Dec 28, 2020).
Liao Y, Smyth GK, Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics*. 2014;30(7):923-930. doi:10.1093/bioinformatics/btt656
Love, M.; Anders, S.; Kim, V.; Huber, W. RNA-seq workflow: gene-level exploratory analysis and differential expression. Bioconductor. https://bioconductor.org/packages/release/workflows/vignettes/rnaseqGene/inst/doc/rnaseqGene.html#preparing-quantification-input-to-DESeq2 (accessed Dec 28, 2020).
Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550. doi:10.1186/s13059-014-0550-8
StringTie Transcript assembly and quantification for RNA-seq. StringTie. http://ccb.jhu.edu/software/stringtie/index.shtml?t=manual (accessed Dec 28, 2020).
Wang Z, Gerstein M, Snyder M. RNA-seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*. 2009;10(1):57-63. doi:10.1038/nrg2484

Diferenciální genová exprese

Biological interpretation of gene expression data. Functional genomics II Common technologies and data analysis methods. https://www.ebi.ac.uk/training-beta/online/courses/functional-genomics-ii-common-technologies-and-data-analysis-methods/biological-interpretation-of-gene-expression-data-2/ (accessed Dec 28, 2020).
Differential gene expression (DGE) analysis. https://hbctraining.github.io/Training-modules/planning_successful_rnaseq/lessons/sample_level_QC.html (accessed Dec 28, 2020).
Gentleman RC, Carey VJ, Bates DM, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*. 2004;5(10):R80. doi:10.1186/gb-2004-5-10-r80

Gilbert, S. Differential Gene Expression, 2000. Developmental Biology. 6th edition. https://www.ncbi.nlm.nih.gov/books/NBK10061/ (accessed Dec 28, 2020).
Hansen, K. limma. http://ccb.jhu.edu/software/stringtie/index.shtml?t=manual (accessed Dec 28, 2020).
Love, M.; Anders, S.; Kim, V.; Huber, W. RNA-seq workflow: gene-level exploratory analysis and differential expression. Bioconductor. https://bioconductor.org/packages/release/workflows/vignettes/rnaseqGene/inst/doc/rnaseqGene.html#preparing-quantification-input-to-DESeq2 (accessed Dec 28, 2020).
Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 2014;15(12):550. doi:10.1186/s13059-014-0550-8

Testování statistických hypotéz

Goldman, M. Section0402. Statistics for Bioinformatics. https://www.stat.berkeley.edu/~mgoldman/Section0402.pdf (accessed Dec 28, 2020).
Holčík, J., Komenda, M. (eds.). Matematická biologie: e-learningová učebnice [online].1. vydání. Brno: Masarykova univerzita, 2015. ISBN 978-80-210-8095-9
Testování hypotéz ve statistice. https://cit.vfu.cz/statpotr/POTR/Teorie/Predn3/hypotezy.htm (accessed Dec 28, 2020).
Pavlík, T. *Biostatistika pro matematickou biologii - Problém násobného testování hypotéz* [online]; https://portal.matematickabiologie.cz/index.php?pg=aplikovana-analyza-klinickych-a-biologickych-dat--biostatistika-pro-matematickou-biologii--uvod-do-testovani-hypotez--poznamky-k-testovani-hypotez--problem-nasobneho-testovani-hypotez (accessed Dec 17, 2020).

Funkční analýza dat

Ashburner M, Ball CA, Blake JA, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*. 2000;25(1):25-29. doi:10.1038/75556
Carbon S, Ireland A, Mungall CJ, et al. AmiGO: online access to ontology and annotation data. *Bioinformatics*. 2009;25(2):288-289. doi:10.1093/bioinformatics/btn615
Gentleman RC, Carey VJ, Bates DM, et al. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol*. 2004;5(10):R80. doi:10.1186/gb-2004-5-10-r80
Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M. The KEGG resource for deciphering the genome. *Nucleic Acids Res*. 2004;32(Database issue):D277-D280. doi:10.1093/nar/gkh063
KEGG Overview. https://www.genome.jp/kegg/kegg1a.html (accessed Dec 28, 2020).
Luo W, Brouwer C. Pathview: an R/Bioconductor package for pathway-based data integration and visualization. *Bioinformatics*. 2013;29(14):1830-1831. doi:10.1093/bioinformatics/btt285

Subramanian A, Tamayo P, Mootha VK, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A*. 2005;102(43):15545-15550. doi:10.1073/pnas.0506580102
Yu G, Wang LG, Han Y, He QY. clusterProfiler: an R package for comparing biological themes among gene clusters. *OMICS*. 2012;16(5):284-287. doi:10.1089/omi.2011.0118

11 PŘÍLOHA

11.1 Časová osa

Důležité milníky v molekulární biologii, výpočetní biologii a bioinformatice

1854-1864 Gregor Mendel, zakladatel genetiky, dělá pokusy s křížením hrachu, aby zjistil, jak se dědí jejich jednotlivé znaky

1869 Friederich Miescher poprvé izoluje DNA

1909 Wilhem Johannsen poprvé používá výraz gen

1911 Thomas Morgan prokazuje, že geny jsou umístěny na chromosomech

1941 George Beadle a Edward Tatum potvrzují hypotézu „jeden gen, jeden enzym“ (s objevem alternativního splicingu překonáno)

1944 Oswald Avery, Colin Munro MacLeod a Maclyn McCarty dokazují, že DNA je nositelem genetické informace. Tato informace byla přijata až roku 1952, kdy Alfred Hershey and Martha Chase ukázali, že do organismu (bakterie) musí vstoupit nukleová kyselina, aby došlo k infekci, kdežto proteinová kapsida zůstává na povrchu. Původní tvrzením bylo, že nositelem genetické informace jsou proteiny

1952 Martha Chase a Alfred Hershey definitivně dokazují, že geny jsou složeny z DNA

1952 Grace Hopper dokončuje programový linker (kompilátor), který byl napsán pro systém A-0

1953 Watson a Crick navrhují model dvojité šroubovice DNA na základě dat získaných Rosalind Franklin a Mauricem Wilkinsem

1954 Max Perutz vyvinul metodu těžkých atomů k řešení fázového problému v proteinové krystalografii

1955 Frederick Sanger stanovuje první sekvenci proteinu (hovězí inzulin)

1957 Francis Crick navrhuje Centrální dogma molekulární biologie

1965 Margaret Dayhoff propaguje využití výpočetní technologie k porovnání proteinových sekvencí a znázornění jejich evolučních vztahů ze zarovnání, publikuje knihu Atlas of Protein Sequence and Structure, ve které uvádí všechny v té době známé proteinové sekvence

1966 Marshall Nirenberg, Har Khorana, Severo Ochoa a další rozluštili genetický kód

1969 Je vytvořena ARPANET (prapůvodní počítačová síť) provázáním počítačů na Standfordu a Kalifornské univerzitě v Los Angeles

1970 Jsou zveřejněny podrobnosti algoritmu Needleman-Wunsch pro srovnání sekvencí

1971 Založení databáze PDB (Protein Data Bank)

1972 Paul Berg, David Jackson, Robert Symons vytváří první molekulu rekombinantní DNA

1973 Robert Metcalfe popisuje Ethernet

1974 Vint Cerf a Robert Khan rozvíjejí koncept propojení sítí počítačů do „internetu“ a rozvíjejí Transmission Control Protocol (TCP)

1975 Založení společnosti Microsoft Corporation Billem Gatesem a Paulem Allenem. P. H. O´Farrel popisuje dvojrozměrnou elektroforézu, kde je separace proteinů na polyakrylamidovém gelu SDS kombinována se separací podle izoelektrických bodů

1975-77 Frederick Sanger, Allan Maxam a Walter Gilbert popisují metody DNA sekvenování (Sangerova metoda, Maxam-Gilbertova metoda)

1976 Je založena první geneticko-inženýrská společnost Genentech

1977 Je uvedena nukleotidová sekvence Phi_X_174 (dodnes srovnávací sekvence)

1977 Phillip Allen Sharp a Richard J. Roberts objevili nekódující sekvence, introny, přerušující sekvence s biologickou funkcí (exony)

1978 Margaret Dayhoff zveřejňuje PAM matici (point accepted mutation) - substituční matici ke skórování alignmentu proteinových sekvencí

1980 Walter Gilbert a Frederick Sanger získávají Nobelu cenu za chemii za průkopnickou práci v metodice sekvenování

1982 Je založena databáze GenBank

1983 Kary Mullis vyvíjí metodu PCR reakce

1983 Richard Stallman zakládá projekt GNU zaměřený na svobodný software, inspirovaný operačními systémy unixového typu

1988 Je založen The National Centre for Biotechnology Information (NCBI) v rámci National Cancer Institute

1988 Je zahájena inciativa Human Genome, Pearson a Lipman publikují algoritmus FASTA pro srovnání sekvencí

1988 Je vytvořen nový specifický počítačový program: první počítačový virus, navržený R. Morrisem, infikuje 6 000 vojenských počítačů v USA

1990 Je implementován algoritmus BLAST

1991 CERN oznamuje vytvoření protokolů, které tvoří celosvětovou webovou síť. Je popsána tvorba a použití značek exprimovaných sekvencí (EST). Jsou osekvenování následující důležité geny: BRCA1, BRACA1, CHD1, MMAC1, MMSC1, MMSC2, CtIP, p16, p19 and MTS2.

1993 Affymetrix zahajuje nezávislé operace v Santa Clara v Kalifornii.

1993 Objev microRNA u Caenorhabditis elegans

1994 Burrows a Wheeler přichází s algoritmem na kompresi dat (Burrows-Wheeler Transformation = BWT)

1994 Společnost Netscape Communications Corporation založila a vydala Navigator, komerční verzi Mozilly od NCSA. Je publikována databáze proteinových motivů – PRINTS, Attwoodem a Beckem

1995 Je osekvenován genom Haemophilus influenzea a Mycoplasma genitalium (vedeno Craigem Venterem)

1996 Je osekvenován genom Saccharomyces cerevisiae. Bairoch a kolektiv uvádí databáze PROSITE

1996 Affymetrix ukazuje první komerční DNA čip

1997 Je publikován genom E.coli. LION bioscience AG založily integrovanou genomickou společnost se zaměřením na bioinformatiku. Společnost sestávala z IP z EMBLu, EBI, a DKFZ (The German Cancer Research Center), a University of Heidelberg

1998 Je publikován genom pro Caenorhabitis elegans

1998 Craig Venter tvoří společnost Celera v Rockville v Marylandu

1998 V San Diegu vznikla společnost GeneFormatics, společnost zabývající se analýzou a predikcí struktury a funkce proteinů

2000 Jsou publikovány sekvence genomů Pseudomonas aeruginosa a D.melanogaster

2000 Paolo Ferragina a Giovanni Manzini ukazují použití BWT jako indexu (FM index)

2001 Je publikován lidský genom

2003 Začíná ENCODE Projekt

2005 Je publikována metoda microfluidního pyrosekvenování 454, první z metod Nové generace, která umožnila sekvenování v řádku tisíců až milionů sekvencí

2006 Solexa (Illumina) sekvenování nové generace

2009 První projekt Single cell sekvenování (myší blastomera)

2011 Jennifer Doudna a Emmanuelle Charpentier vyvíjí CRISPR/Cas9 systém specifické editace DNA

2012 ENCODE konsorcium publikuje, že více než 80 % protein nekódující DNA vykazuje určitou biologickou funkci

2018 společnost DeepMind publikuje nástroj na predikci 3D struktur proteinů AlphaFold využívající umělou inteligenci, způsobuje revoluci v proteinových predikcích

2020 Je publikován AlphaFold2 – vylepšená verze AlphaFold

11.2 Další aplikace NGS

11.2.1 Single cell sekvenování a spatial transcriptomics

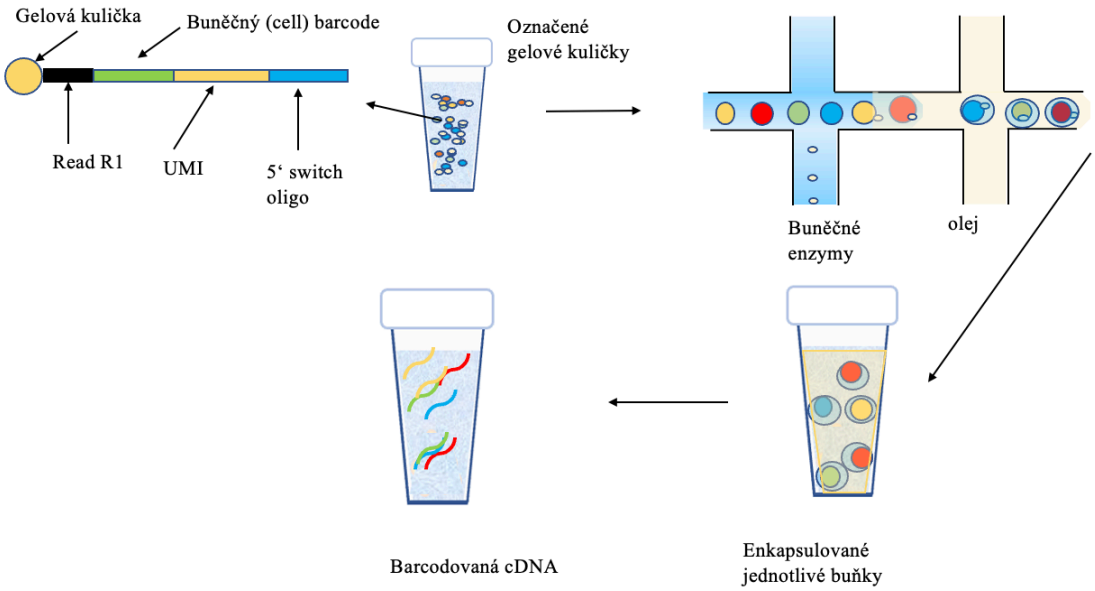
Specifickým způsobem sekvenování je **single cell sequencing** (sekvenování jedné buňky, scRNA-seq), které využívá lehce modifikovaných metod sekvenování příští generace, konkrétně využívá Illuminy (Solexa) a Ion Torrentu. Tato technologie se s úspěchem používá při sekvenování transkriptomů jednotlivých buněk pro zjištění rozdílů v buněčné populaci a charakterizaci buněčných evolučních vztahů. Oproti klasickým způsobům sekvenování, single cell umožňuje detekovat heterogenitu mezi jednotlivými buňkami, rozlišit druhy buněk v různých stádiích vývoje (např. vývoj rakovinné buňky) a v různých fázích buněčného cyklu, a konečně vytvoření celobuněčných map.

Typická workflow pro scRNA-seq zahrnuje následující kroky (s možnými úpravami i vynecháním kroků):

1. izolace a třídění jednotlivých buněk
2. lýze buněk při zachování mRNA
3. zachycení mRNA a přepis na cDNA
4. cDNA amplifikace a příprava knihovny cDNA
5. vlastní sekvenování za využití metod NGS
6. specifikována bioinformatická analýza (R balíčky Seurat, Velocyto, bioinfocz/scdrake)

Rozdělení buněk: Existuje několik možností, jak rozdělit jednotlivé buňky: pomocí aktivované fluorescence, magnetických kuliček, mikroselekce, mikrofluidic či manuální selekce. Momentálně se však nejvíce využívá technologie droplets (Obr. 90): každá buňka je enkapsulována do emulze s gelovou kuličkou na mikrofluidním čipu. Každá buňka je označena specifickým molekulárním barcodem, každá molekula obvykle označena unikátním identifikátorem (UMI). RNA je reverzně transkribována na cDNA (10X Chromium Countner od 10X Genomics).

***Aktivovaná fluorescence**: suspendované značené buňky v kapičkách, kdy každá obsahuje právě jednu buňku, vedeny před laser. Na základě fluorescence a rozptylu světla aplikuje přístroj náboj na kapičku obsahující požadovanou buňku a elektrostatický vychylovací systém usměrní kapičky do různých sběrných zkumavek.*
***Magnetické kuličky**: buňky jsou specificky značeny značkou navázanou na magnetickou kuličku. K oddělení kuliček se používá magnetické pole.*
***Mikroselekce**: laser je vystřelen přes víčko na buňky, které jsou předmětem zájmu. Teplo roztaví membránu a buňky přilnout k roztavené membráně. Po odstranění víčka jsou zachycené buňky odstraněny a ostatní buňky nechány stranou.*
***Manuální selekce**: pomocí skleněné pipety připojené k mikromanipulátoru, jednotlivé buňky tak mohou být vybrány a přeneseny pro analýzu.*
***Microfluidic**: buňky jsou napřed disociovány a poté proudí k čipu. Takto mohou být buňky rozděleny do různých kompartmentů (jamek, kapiček) obsahujících pouze jednu buňku.*



Obr. 90: 10X Chromium příprava buněk pro scRNA-seq

Příprava knihovny a sekvenování probíhá obdobně jako u ostatních NGS metod, velmi obdobně probíhají i některé kroky bioinformatické analýzy (kontrola kvality, mapování). Následně dochází k detekci vysoce variabilních genů, vyhodnocení efektu buněčného cyklu, normalizaci, shlukování a redukci dimenzionality.

Shluková analýza se obvykle počítá na základě transkripčních profilů jednotlivých buněk, a je následována anotací např. pomocí skórovací matice a výběrem markerových genů, které jsou rozdílně exprimované mezi klastry.

Spatial transcriptomics je metoda molekulárního profilování, která umožňuje měřit veškerou aktivitu genů ve vzorku tkáně (FFPE bločky, nebo laserový řez čerstvou zmraženou tkání) a mapovat, kde k aktivitě dochází. Klasická spatial transcriptomics nemá rozlišení na úrovni buněk, ale objevila se technika – single-cell Stereo-seq (rok 2022), která to je schopna umožnit.

11.2.2 Amplikonové sekvenování

Amplikonové sekvenování je jednou z technik využívajících technologie NGS. Využívá se pro paralelní sekvenování několika konkrétních genů/libovolných úseků DNA, kdy můžeme sledovat změny v konkrétních genech spojených typicky s určitým fenotypem. Technika využívá přesně navržených PCR primerů nebo capture probes pro dané úseky na DNA. Typickým použitím v klinickém prostředí je screening onemocnění, kdy se používá kupříkladu pro odhalení somatických mutací ve vzorcích, které obsahují jak normální, tak nádorovou tkáň (typicky vzorky získané biopsií). Nevýhodou je nutnost provést PCR všech úseků ve všech vzorcích. V okamžiku, kdy bychom měli 14 genů spojených s onemocněním, a 96 různých pacientů, bude nutné provést 14 x 96 PCR reakcí, a následně smíchat amplikony pro daný vzorek dohromady. Výsledkem by byla zaplněná 96 jamková destička, kde by bylo v každé jamce 14 různých amplikonů. Tyto amplikony dále sekvenujeme Sangerovským nebo NGS sekvenováním. Amplikonové sekvenování se velmi často používá v genomice, protože výrazně snižuje cenu za sekvenování.

11.2.3 ChIP-Seq, CLIP-Seq, ATAC-Seq

ChIP-Seq se primárně používá k určení, jak transkripční faktory a další proteiny spojené s chromatinem ovlivňují mechanismy ovlivňující fenotyp. Určení toho, jak proteiny interagují s DNA za účelem regulace genové exprese, je nezbytné pro plné pochopení mnoha biologických procesů, stejně jako chorobných stavů. Kombinuje imunoprecipitaci chromatinu (ChIP) s NGS metodami. ChIP vytváří knihovnu cílových míst DNA navázaných na sledovaný protein. NGS metody se používají ve spojení s celogenomovými sekvenčními databázemi k analýze vzoru interakce jakéhokoli proteinu s DNA nebo vzoru jakýchkoli epigenetických modifikací chromatinu. Jako alternativa k závislosti na specifických protilátkách byly vyvinuty různé metody k nalezení nadmnožiny všech aktivních regulačních oblastí v genomu s depletovaným nebo narušeným nukleozomem, jako je DNase-Seq a FAIRE-Seq.

CLIP-Seq (známé také jako HITS-Clip) se používá pro zjištění RNA interagující s RNA-vazebným proteinem (případně jinou RNA). Využívá crosslinking („propojení“) mezi RNA a proteinem, následovanou imunoprecipitací s protilátkami pro daný protein.

ATAC-Seq (Assay for Transposase-Accessible Chromatin with high-throughput sequencing) je metoda k určení otevřenosti/dostupnosti chromatinu napříč genomem pomocí NGS knihoven vytvořených s využitím hyperaktivní transpozázy Tn5. Adaptéry jsou navázány na Tn5 transpozázu, což umožňuje současnou fragmentaci chromatinu a integraci těchto adaptérů do otevřených oblastí chromatinu. ATAC-Seq se využívá k pochopení jak chromatin a další faktory ovlivňují genovou expresi, např. u komplexních onemocnění, embryonálního vývoje, aktivace T-buněk a rakoviny. K určení

chromatinových stavů jednotlivých subpopulací buněk je možné využít ATAC-Seq na úrovni jedné buňky – Single-Cell ATAC-Seq.

11.3 Čipové technologie

Základním principem DNA čipů (mikročipů, microarrays) je hybridizace – schopnost spontánního navázání fluorescenčně značeného vzorku na sondy imobilizované na povrchu čipu (sonda = navržený unikátní úsek DNA sekvence). Hlavní využití čipů je charakterizace apoptózy, buněčné signalizace, diferenciální (genové) exprese, povrchových receptorů, metylace DNA, na genotypizaci (SNP, STR) a pro komparativní genomiku (změny v počtu kopiích genů, porovnání dvou příbuzných genomů, z nichž jeden musí být znám).

Specifickou metodikou je chromatinová imunopercipitace na čipu (ChIP-on-chip), metoda, které slouží ke stanovení epigenetických změn chromatinu či TFBS (Transcription factor binding site, vazebné místo transkripčního faktoru). Epigenetický stav příslušné kódující oblasti a promotoru, je zásadní pro regulaci transkripční aktivity genů. DNA sekvenci náležící k TFBS konkrétního proteinu lze izolovat pomocí imunoprecipitace daného proteinu a získané fragmenty se následně naváží na čip.

V současné době se od použití čipových technologií pomalu upouští, ale v nemocnicích pořád nachází své uplatnění (např. při charakterizaci typu leukémií).

Nejběžnějším nosným podkladem čipu je mikroskopické sklíčko. Povrch čipu musí být upraven pomocí hydrofobních polymerů (poly-L-lysin, modifikovaný silan) poskytujících reakční skupiny pro navázání sond. Fragmenty DNA sond mohou být 5' koncích značeny fluorescenční značkou, nebo nukleotidy jejímž prostřednictvím dochází k vazbě (typicky biotin).

Existují různé technologie pro přípravu DNA sond na povrchu čipu:
Off-line (spotted arrays): mechanické nanášení již presyntetizovaných sond, a to buď cDNA klonů, PCR produktů nebo chemicky připravených oligonukleotidů. Přenos sond probíhá přes čipovací přístroje (spotter, microarrayer, mikropipetor) pomocí hrotu na povrch čipu.
Ink-jet: bezkontaktní výroba čipů, kdy se prostřednictvím elektromagnetického nebo piezoelektrického systému nanášejí stříkačkou kapky těsně nad povrch čipu.
In-situ: přímá chemická syntéza na povrchu čipu pro výrobu libovolných sekvencí oligonukleotidových sond.

Mikročipy jsou pouze semi-kvantitativní metodou, kdy nespecifické hybridizace vedou ke špatným signálům (cross-hybridization). Pro stanovení hodnoty přirozeného fluorescenčního pozadí bez cross-hybridizací jsou čipy vybaveny sondami negativních kontrol, jejichž sekvence se nevyskytují v analyzovaném genomu (např. gen u jiného

druhu organismu, který se nevyskytuje ve studovaném organismu), mismatch proby, které jsou velmi podobné, ale ne dokonale komplementární k cílovému transkriptu a/nebo sondy s housekeepingovými geny. Pomocí intenzity těchto kontrolních sond je možné odhadnout statistickou významnost měřené intenzity signálu pro každý cíl například pomocí p-hodnoty. Také pomáhají odhadnout kvalitativní a kvantitativní limity čipů.

Každý výrobce využívá pro čipy vlastní datový formát. Výstupní soubor z analyzování Affymetrixových čipů jsou soubory CEL, u Illumina čipů IDAT soubory, které obsahují matici dat skládající se z identifikátorů prob, odhad intenzit signálů spolu s intenzitou pozadí nebo odhadem chyby signálů, a z experimentálních metadat. Výstupní soubory mohou být analyzovány pomocí speciálních softwarů daných firem, případně pomocí různých Bioconductor balíčků v softwaru R (Limma, Affy, Oligo).

11.4 Zdroje

Časová osa

Gauthier J., Vincent A.T, Charette S J, Derome N, A brief history of bioinformatics, Briefings in Bioinformatics, Volume 20, Issue 6, November 2019, Pages 1981–1996, https://doi.org/10.1093/bib/bby063

Tampi S. M. Introduction to Bioinformatics, 2009. arxiv.org e-Print archive. https://arxiv.org/ftp/arxiv/papers/0911/0911.4230.pdf (accessed Dec 27, 2020).

Timelines, 2021. Timelines. https://www.genome.gov/about-genomics/educational-resources/timelines (accessed March 03, 2022).

Další aplikace NGS

Ay-Berthomieu, A-S. Beginner's Guide to Understanding Single-Cell ATAC-Seq. https://www.activemotif.com/blog-single-cell-atac-seq (accessed Feb 03, 2022).

Buenrostro J., Giresi P., Zaba L. et al. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat Methods 10, 1213–1218 (2013). https://doi.org/10.1038/nmeth.2688

Haque A., Engel J., Teichmann S.A. *et al.* A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. *Genome Med* **9,** 75 (2017). https://doi.org/10.1186/s13073-017-0467-4

Hu P, Zhang W, Xin H, Deng G. Single Cell Isolation and Analysis. *Front Cell Dev Biol*. 2016;4:116. Published 2016 Oct 25. doi:10.3389/fcell.2016.00116

Introduction to Amplicon Sequencing. Illumina Sequencing and array-based solutions for genetic research. https://www.illumina.com/techniques/sequencing/dna-sequencing/targeted-resequencing/amplicon-sequencing.html (accessed Dec 28, 2020).

Liu F, Zhang Y, Zhang L, et al. Systematic comparative analysis of single-nucleotide variant detection methods from single-cell RNA sequencing data. *Genome Biol*. 2019;20(1):242. Published 2019 Nov 19. doi:10.1186/s13059-019-1863-4

Novotná, M. Princip NGS metody, 2018. Generi Biotech. https://www.generi-biotech.com/cs/princip-ngs-metody/ (accessed Dec 28, 2020).
Ziegenhain C, Vieth B, Parekh S, et al. Comparative Analysis of Single-Cell RNA Sequencing Methods. *Mol Cell*. 2017;65(4):631-643.e4. doi:10.1016/j.molcel.2017.01.023
Xia K, Sun HX, Li J, et al. The single-cell stereo-seq reveals region-specific cell subtypes and transcriptome profiling in Arabidopsis leaves. Dev Cell. 2022;57(10):1299-1310.e4. doi:10.1016/j.devcel.2022.04.011
10X Chromium System. Bauer Core Facility. https://bauercore.fas.harvard.edu/10x-chromium-system (accessed Dec 28, 2020).

Čipové technologie

Bumgarner R. Overview of DNA microarrays: types, applications, and their future. *Curr Protoc Mol Biol*. 2013;Chapter 22:Unit-22.1.. doi:10.1002/0471142727.mb2201s101
DNA Microarray Technology Fact Sheet. National Human Genome Research Institute Home. https://www.genome.gov/about-genomics/fact-sheets/DNA-Microarray-Technology (accessed Dec 28, 2020).
Govindarajan R, Duraiyan J, Kaliyappan K, Palanisamy M. Microarray and its applications. *J Pharm Bioallied Sci*. 2012;4(Suppl 2):S310-S312. doi:10.4103/0975-7406.100283
Li X, Gu W, Mohan S, Baylink DJ. DNA microarrays: their use and misuse. *Microcirculation*. 2002;9(1):13-22. doi:10.1038/sj.mn.7800118
Loewe RP, Nelson PJ. Microarray bioinformatics. *Methods Mol Biol*. 2011;671:295-320. doi:10.1007/978-1-59745-551-0_18