Univerzita Karlova v Praze

Přírodovědecká fakulta

Studijní program: Biologie

Studijní obor: Genetika, molekulární biologie a virologie

Bc. Anna Přistoupilová

Využití nových sekvenačních technik v biomedicínském výzkumu

Application of novel DNA sequencing techniques in biomedical research

Diplomová práce

Vedoucí závěrečné práce: Ing. Stanislav Kmoch CSc.

Praha, 2011

Prohlášení:

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

V Praze, 5. 5. 2011

Podpis

Tato diplomová práce byla vypracována v letech 2010 – 2011 na Ústavu dědičných metabolických poruch 1. lékařské fakulty Univerzity Karlovy v Praze a Všeobecné fakultní nemocnice v Praze, v Laboratoři genomiky a bioinformatiky pod odborným dohledem Ing. Stanislava Kmocha CSc.

Práce byla podpořena granty MSM0021620806 a GA UK 250051.

Poděkování

Na tomto místě bych ráda poděkovala svému vedoucímu diplomové práce Ing. Stanislavu Kmochovi CSc. za odborné vedení a za možnost se tímto zajímavým tématem zabývat. Dále bych ráda poděkovala Mgr. Viktorovi Stáneckému za zasvěcení do problematiky analýzy bioinformatických dat a za všechny cenné rady, které mi v průběhu řešení této práce poskytl. V neposlední řadě bych ráda poděkovala Mgr. Lence Noskové a Mgr. Haně Hartmannové, které mi radily a pomáhaly při práci v laboratoři.

Abstrakt

Nové sekvenační techniky v blízké budoucnosti změní přístup, jakým jsou prováděny vědecké experimenty a prováděna diagnostika známých onemocnění. Umožní vznik personalizované medicíny.

Smyslem této diplomové práce je podat přehled nových metodických postupů sekvenování DNA, ukázat možnosti a omezení vhodných bioinformatických nástrojů a demonstrovat přínos těchto technik pro projekty zaměřené na odhalení molekulární podstaty vzácných dědičně podmíněných onemocnění.

V úvodu práce je podán přehled nových, komerčně dostupných metod sekvenace genomu (firem 454 Life Science, Applied Biosystems, Illumina, Helicos), vysvětlen princip na kterém tyto metody pracují a nastíněn směr, kterým se ubírá další vývoj na tomto poli. Úvod se dále věnuje způsobům analýzy dat, která je nedílnou součástí sekvenačních technik.

V praktické části je prezentováno spektrum metod analýzy genomu, které byly úspěšně využity k určení kauzálního genu a mutací způsobujících adultní formu autozomálně dominantní neuronální ceroid lipofuscinózy (ANCL). V rámci tohoto projektu bylo využito v jedné ANCL rodině kombinace celogenomového genotypování (GeneChip® Mapping 10K 2.0 Array), vazebné analýzy (Merlin), analýzy počtu kopií genomové DNA (Genome-Wide Human SNP Array 6.0), analýzy expresních profilů (HumanRef-8 v2 Expression BeadChips) a sekvenování exomu (SOLiD™ 4 System). Získaná data byla podrobena ucelené bioinformatické analýze, v jejímž rámci bylo použito vazebné analýzy a analýzy diferenciálních expresních profilů. Zvláštní důraz byl kladen na porovnání a zhodnocení aktuálně dostupných mapovacích algoritmů pro rekonstrukci, analýzu a anotaci sekvenačních dat a na jejich výběr. Vzájemným propojením výsledků těchto analýz byla definována omezená množina genů, které jsou lokalizovány v kandidátních oblastech definovaných vazebnou analýzou, mají změněný transkripční profil v tkáni pacientů, obsahují unikátní funkčně významnou mutaci, a jejichž biologická funkce naznačuje možnou souvislost se studovaným neurologickým postižením.

Takto definované mutace a jejich segregace se studovaným fenotypem byly následně ověřeny přímým sekvenováním. Konečným výsledkem této studie bylo určení jedné kandidátní mutace, jejíž kauzalita je v současnosti experimentálně prokazována.

Abstract

Application of novel DNA sequencing techniques in biomedical research

Next generation sequencing technologies are changing the way scientific experiments and diseases diagnostics are performed and thus will allow what is called personalized medicine. The sense of presented thesis is to make survey of new approaches to DNA sequencing and demonstrate usage and constraints of bioinformatic analytical tools available to day. Discussed techniques are then applied to the case study of finding molecular basis for rare hereditary disease.

Introductory part deals with overview of commercially available sequencing techniques (454 Life Science, Applied Biosystems, Illumina, Helicos). Fundamentals of each method are described and possible further development is outlined. Post sequencing data analysis is than discussed in details.

In practical section we demonstrate genome analysis techniques successfully used to reveal causal mutation in the gene responsible for adult form of autozomal neuronal ceroid lipofuscinosis (ANCL). Combination of linkage analysis (Merlin), copy number variant analysis (Genome-Wide Human SNP Array 6.0), analysis of expression profiles (HumanRef-8 v2 Expression BeadChips) and exome sequencing (SOLiD™ 4 System) has been applied to members of one ANCL family. We also paid attention to comparison, evaluation and selection of available mapping algorithms used in reconstruction, analysis and anotation of sequencing data. Combined results from different techniques led to definition of small pool of genes localized in candidate areas defined by linkage analysis. Those genes have altered transcript profile in the patient tissue, they posses unique functionally important mutation and their known biological meaning could be linked to neuronal disease studied.

In the next step, candidate mutations have been confirmed by direct sequencing and their fenotype segregation have been checked up. Finally, we have found one mutation probably responsible for ANCL and its relevance is now under further study.

Keywords: next generation sequencing technologies, exome sequencing, mapping algorithms, resequencing, ANCL, neuronal ceorid lipofuscinosis, bioinformatics, SOLiD

Klíčová slova: nové sekvenační techniky, exomové sekvenování, mapovací algoritmy, resekvenování, ANCL, neuronální ceorid lipofuscinóza, bioinformatika, SOLiD

OBSAH

1. ÚVOD

Sekvenování je základní metodou zjištění primární genetické informace, která je určena pořadím bází v molekule DNA. Od roku 1977, kdy Frederic Sanger představil metodu sekvenování, se stalo běžným nástrojem využívaným jak ve výzkumu, tak k diagnostice mnoha onemocnění. Významným mezníkem byl rok 2001, kdy byla publikována první verze lidského genomu (Lander et al., 2001) a rok 2003, kdy byla publikována verze finální (Collins et al., 2003).

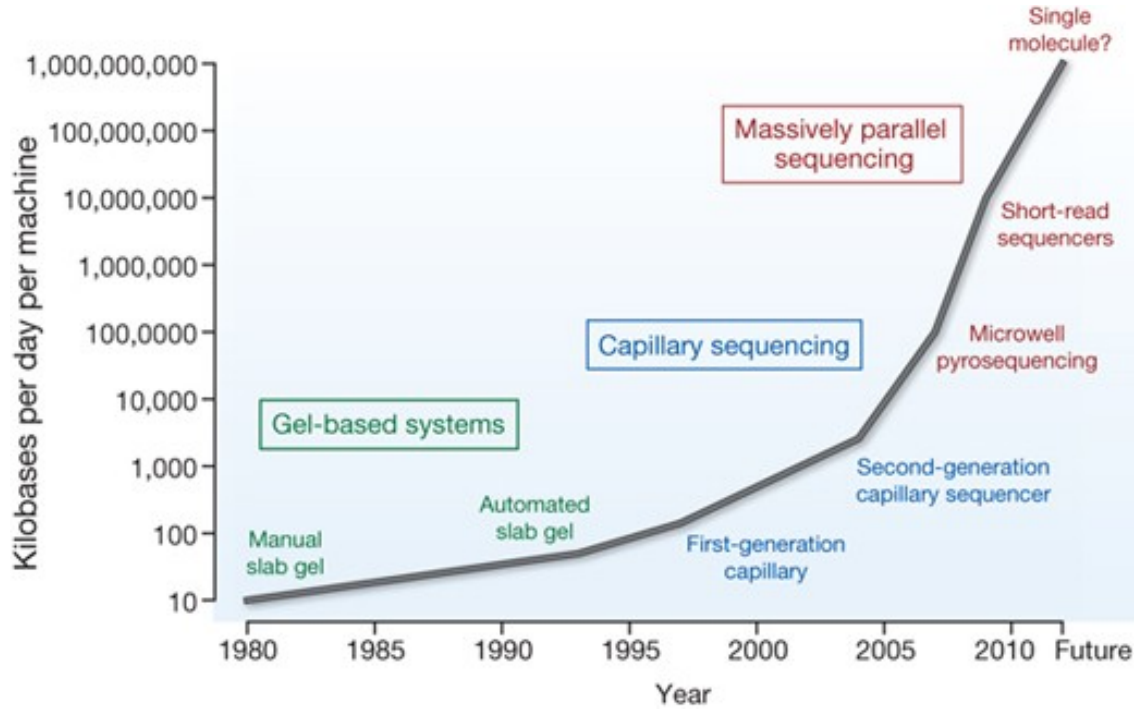
V roce 2005 byl představen první sekvenátor nové generace umožňující sekvenovat obrovské množství dat za mnohem nižší cenu než za použití stávající Sangerovy metody (Margulies et al., 2005). Následovalo představení dalších sekvenátorů nové generace, umožňujících sekvenovat celé genomy. První kompletní lidský genom byl osekvenován v roce 2008. Jednalo se o genom Jamese Watsona. Jeho osekvenování trvalo 2 měsíce a stálo 1 milión dolarů (Wheeler et al., 2008). V současné době trvá sekvenace lidského genomu jeden týden a stojí 20000 dolarů[1]. Ač se cena sekvenování několikanásobně snížila, stále je toto sekvenování nepřijatelně drahé pro rutinní diagnostické použití.

Existuje několik iniciativ, které mají za cíl snížení ceny sekvenace na částku, kterou by byly pojišťovny ochotny zaplatit. Tato částka je stanovena na 1000 dolarů za osekvenovaný genom. Genom za 1000 dolarů[2] je meta, která umožní vznik personalizované medicíny a tím pádem možnost individuální a maximálně účinné léčby, zvolené na základě znalosti genetické informace jedince. Další takovouto iniciativu vyvinula kalifornská nadace X Prize Foundation[3], která vypsala cenu deset miliónů dolarů pro vědecký tým, který jako první osekvenuje genom stovky lidí během deseti dnů, s celkovými náklady nepřevyšující deset tisíc dolarů na jeden genom.

S rozvojem nových sekvenačních technik (NGS) je spojená produkce velkého množství dat, které je potřeba dále uchovávat, zpracovávat a analyzovat (Obr. 1) Stávající metody se pro tyto účely příliš nehodí, protože algoritmy jsou výpočetně náročné a optimalizované pro jiný typ dat. Z tohoto důvodu jsou vyvíjeny nástroje jiné, které lépe odrážejí charakter dat produkovaných NGS technologiemi (Pop and Salzberg,

[1] http://www.genome.gov/27543255
[2] http://www.genome.gov/27541189
[3] http://genomics.xprize.org

2008; Koboldt, 2010).



Obr. 1: Objem dat produkovaný v jednotlivých letech odráží vývoj na poli sekvenačních technologií

Zdroj: http://genome.wellcome.ac.uk/doc_wtx059576.html

V roce 2009 byl poprvé použit postup, který umožňuje vysokokapacitní (high-throughput) sekvenování protein-kódujících úseků DNA, tedy exonů (Ng et al., 2009; Choi et al., 2009). Kódující exony nesou většinu funkčních variant (Ng et al., 2008) a mohou obsahovat jednonukleotidové polymorfismy (SNP, single nucleotide polymorphisms), které jsou zodpovědné za většinu mendelovsky děděných onemocnění (Horner et al., 2010). Soubor všech kódujících exonů se nazývá exom a jeho sekvenování se nazývá exomovým sekvenováním. Exom představuje pouze ~1 % lidského genomu, tedy ~30 Mb rozdělených do ~180000 exonů (Ng et al., 2009). Exomové sekvenování přestavuje nový přístup, který umožňuje osekvenování lidského genomu za 1/20 ceny celogenomového sekvenování. Jedná se o postup, který byl úspěšně použit k objasnění molekulární podstaty několika vzácných, monogenně podmíněných onemocnění (Ng et al., 2010a), stejně tak jako ke zjišťování genetické podstaty komplexních, geneticky heterogenních onemocnění (Sanders, 2011).

Exomové sekvenování, společně s klesající cenou za osekvenovanou bázi umožňuje využívání NGS technik stále většímu okruhu uživatelů. Laboratoř genomiky a bioinformatiky UDMP 1.LF UK a VFN se zaměřuje na studium molekulární podstaty vybraných dědičně podmíněných onemocnění a v současné době se zde začínají využívat nové techniky sekvenace genomu. V rámci studia molekulární podstaty adultní formy neuronální ceroid lipofuscinózy (ANCL) byl osekvenován exom jednoho z pacientů na sekvenátoru SOLiD™ 4 System. Získaná data bylo třeba analyzovat a definovat množinu kandidátních mutací.

1.1. Cíle diplomové práce

1. Porovnat a zhodnotit aktuálně dostupné mapovací algoritmy pro rekonstrukci dat sekvenátoru SOLiD4
2. Provést analýzu dat získaných sekvenací exomu pacienta s adultní formou autozomálně dominantní neuronální ceroid lipofuscinózy
3. Definovat a funkčně anotovat unikátní mutace nalezené v analyzovaném vzorku
4. Propojit výsledky této analýzy s dalšími experimentálními daty z vazebné analýzy, analýzy expresních profilů a funkční anotace
5. Definovat množinu kandidátních mutací a ověřit jejich segregaci s fenotypem ve studované rodině
6. Definovat potenciálně kauzální mutaci(e) pro následné funkční ověření

2. PŘEHLED LITERATURY

2.1. Klasické metody sekvenování

Sekvenováním se rozumí určení přesného sledu bází v molekule DNA. Mezníkem v sekvenaci se stal rok 1977, kdy byly nezávisle na sobě, byly publikovány dvě rozdílné metody sekvenace DNA**,** metoda Sangerova (Sanger et al., 1977a; 1977b) a metoda Maxam-Gilbertova. Sangerova enzymatická metoda postupně vytlačila metodu Maxam-Gilbertovu a až do dneška je nejpoužívanější sekvenační metodou.

2.1.1. Sangerova metoda

Před samotným sekvenováním je potřeba získat dostatečné množství DNA produktu. Toho je možné dosáhnout jeho pomnožením v bakteriálních vektorech nebo od roku 1983 také pomocí polymerázové řetězové reakce (PCR), vynalezené Kary Mullisem.

Získaný úsek DNA je vložen do reakční směsi obsahující následující reakční komponenty:

- Primer komplementární k začátku požadované sekvence DNA
- DNA polymeráza
- Deoxyribonukleotidtrifosfáty (dNTPs)
- Dideoxyribonukleotidtrifosfáty (ddNTPs)- nukleotidy které postrádají 3'- hydroxylovou skupinu potřebnou pro tvorbu fosfodiesterové vazby, a tudíž znemožňují navázání dalšího nukleotidu a tím ukončují syntézu řetězce DNA. ddNTPs jsou radioaktivně či dnes již spíše fluorescenčně označeny.
- Pufr

Reakce probíhá ve čtyřech různých zkumavkách. Každá zkumavka obsahuje primery, dNTPs a jeden z radioaktivně označených ddNTPs (ddATP, ddTTP, ddCTP nebo ddGTP) v poměru 90:10. Samotná reakce je velice podobná PCR reakci. Dochází k cyklickému opakování třech kroků teplotních kroků.

- Denaturace: (95-96 °C), dochází k rozvolnění dvouvláknové DNA na

jednovláknovou

- Annealing: (50-55 °C), primery nasedají na komplementární místa molekuly DNA
- Extenze: (60-70 °C), DNA polymeráza nasedá na dvouvláknový úsek DNA-primer a začíná syntéza nového vlákna na základě komplementarity bází. dNTPs a ddNTPs jsou připojovány na 3´ OH skupinu ribózy.

Tyto tři kroky jsou cyklicky opakovány a dochází ke vzniku stále většího množství fragmentů požadované sekvence. Ve chvíli, kdy je začleněn ddNTP, je syntéza řetězce ukončena. Ve zkumavce vzniknou různě dlouhé fragmenty, vždy zakončené jedním z ddNTP. Ze směsi jsou odstraněny volné nukleotidy, primery i DNA polymeráza, vzniklé fragmenty jsou rozděleny podle svých molekulových hmotnosti pomocí gelové elektroforézy, kdy negativně nabitá DNA putuje ke kladně nabité elektrodě. Rychlost migrace je nepřímo úměrná molekulové hmotnosti DNA fragmentů. Ze získaného elektroforeogramu je sekvence odečítána pomocí radioaktivních značek.

2.1.2. Kapilární sekvenátory

Současné kapilární sekvenátory jsou stále založeny na původním principu, které poprvé použil Frederic Sanger. Došlo pouze k několika drobným změnám:

- Reakce nyní probíhá pouze v jedné zkumavce. To je umožněno použitím různých fluorescenčních značek, kdy každá z nich označuje jeden z ddNTPs.
- K rozdělení fragmentů je využíváno kapilární elektroforézy. Fragmenty procházejí postupně podle své délky miniaturní kapilárou a jejich fluorescenční značky jsou automaticky odečítány a zaznamenávány do počítače .

Kapilární sekvenátory dokáží najednou zpracovat až 384 sekvencí o délce mezi 600 a 1000 nukleotidy. Tyto 384 kapiálrní sekvenátory ale nejsou příliš rozšířené. Mnohem častěji využívané sekvenátory jsou 96 kapilární. Mezi často využívané kapilární sekvenátory se řadí přístroje firmy Applied Biosystems ze série 3xxx a přístroje MegaBACE firmy GE Healthcare.

2.2. Nové techniky sekvenování

Nové techniky sekvenování umožňují sekvenování obrovského množství dat za mnohem nižší cenu, než klasické kapilární sekvenátory. S klesající cenou za osekvenovanou bázi se stávají tyto přístroje dostupnější pro stále širší vědeckou komunitu. Základní princip všech v současnosti rozšířených sekvenátorů nové generace je velmi podobný. Zahrnuje přípravu knihovny, amplifikaci, sekvenaci a analýzu dat.

2.2.1. Příprava knihovny

Před samotným sekvenováním je potřeba připravit genomovou DNA (gDNA) do takové podoby, aby se s ní dalo v pozdějších fázích pracovat. Tomuto procesu se říká příprava knihovny. Příprava knihovny zahrnuje náhodné štěpení gDNA na fragmenty a ligaci adaptérů na oba konce gDNA.

2.2.1.1. Fragmentace DNA

V první fázi je potřeba naštěpit genomovou DNA na kratší úseky. NGS metody k tomuto účelu využívají metody mechanické, jako je sonikace (rozbití DNA na fragmenty pomocí ultrazvuku) a nebulizace (rozprášení DNA na fragmenty působením stlačeného vzduchu v přístroji zvaném nebulizátor).

2.2.1.2. Ligace adaptérů

Existuje několik typů knihoven. Jedním typem knihovny je knihovna fragmentová (fragmented library), kdy jsou na oba konce fragmentu DNA ligovány dva rozdílné adaptéry, přičemž sekvence může být čtena pomocí buď jednoho nebo obou adaptérů (paired-end). Při sekvenování oběma primery vzniká uprostřed neosekvenovaná mezera, tzv. insert size. Tento typ knihovny je vhodný na *de novo* sekvenování. . Jiným typem je mate-paired knihovna. Je čtena z obou stran, nicméně sekvenovaný fragment sestává z obou konců původního DNA fragmentu. Postup přípravy je následující: DNA je naštěpena na fragmenty dlouhé cca 2-5 kb. Oba konce fragmentů jsou označeny biotinylovou značkou, jsou cirkularizovány a poté znovu naštěpeny. Pomocí biotinylové značky jsou vychytány fragmenty, které vznikly spojením obou konců původního fragmentu DNA. Tyto jsou dále upraveny stejným způsobem jako v případě párové knihovny.

2.2.2. Amplifikace

DNA knihovna získaná v předchozích krocích je v tomto kroku pomnožena pomocí PCR reakce. V současných přístrojích se používá můstkové nebo emulzní PCR. Principy těchto dvou metodo jsou podrobně vysvětleny u jednotlivých sekvenátorů.

2.2.3. Sekvenace

Při sekvenování pomocí Sangerovy metody je výsledná sekvence čtena až po skončení celé sekvenační reakce pomocí elektroforetického dělení v závislosti na molekulární hmotnosti fragmentů. V případě nových metod sekvenace tento krok odpadá. Sekvence je čtena okamžitě, pomocí signálu uvolněného vždy po začlenění příslušného nukleotidu. Tomuto principu se také říká sekvenování syntézou (sequencing by synthesis) nebo sekvenace v průběhu extenze (sequencing during extension).

2.2.3.1. Genome Sequencer FLX System (454 Life Sceiences)

V říjnu 2005 byl firmou 454 Life Sciences na trh uveden první sekvenátor nové generace: Genome Sequencer 20 System (Margulies et al., 2005), který dokázal přečíst sekvence dlouhé až 150 bp. Tento přístroj byl v roce 2007 nahrazen přístrojem Genome Sequencer FLX System a délka čtených úseků byla zvětšena až na 300 bp. V roce 2008 firma 454 vyvinula nový reakční kit s názvem Genome Sequencer FLX Titanium Series, který je využitelný na původním přístroji Genome Sequencer FLX System umožňuje číst sekvence dlouhé 400 až 600 bp. Plánované je zvýšení délky čtené sekvence až na 1000 bp.

Princip

Přístroje firmy 454 fungují na principu pyrosekvenace, který v roce 1996 vynalezli Pål Nyrén a Mostafa Ronaghi (Ronaghi et al., 1996).

Příprava knihovny

Vzorek DNA je nejprve naštěpen pomocí nebulizátoru na fragmenty dlouhé 300-800 bp. Na oba konce fragmentu jsou ligovány dva rozdílné dvouvláknové adaptéry A a B, které jsou využívány v následujících krocích k purifikaci, amplifikaci a samotné sekvenaci. Adaptér A má na sobě navázanou biotinovou značku, pomocí které jsou jednotlivé fragmenty přichyceny na streptavidinem obalené magnetické kuličky. Na každou kuličku je v ideálním případě navázán právě jeden jeden fragment DNA.

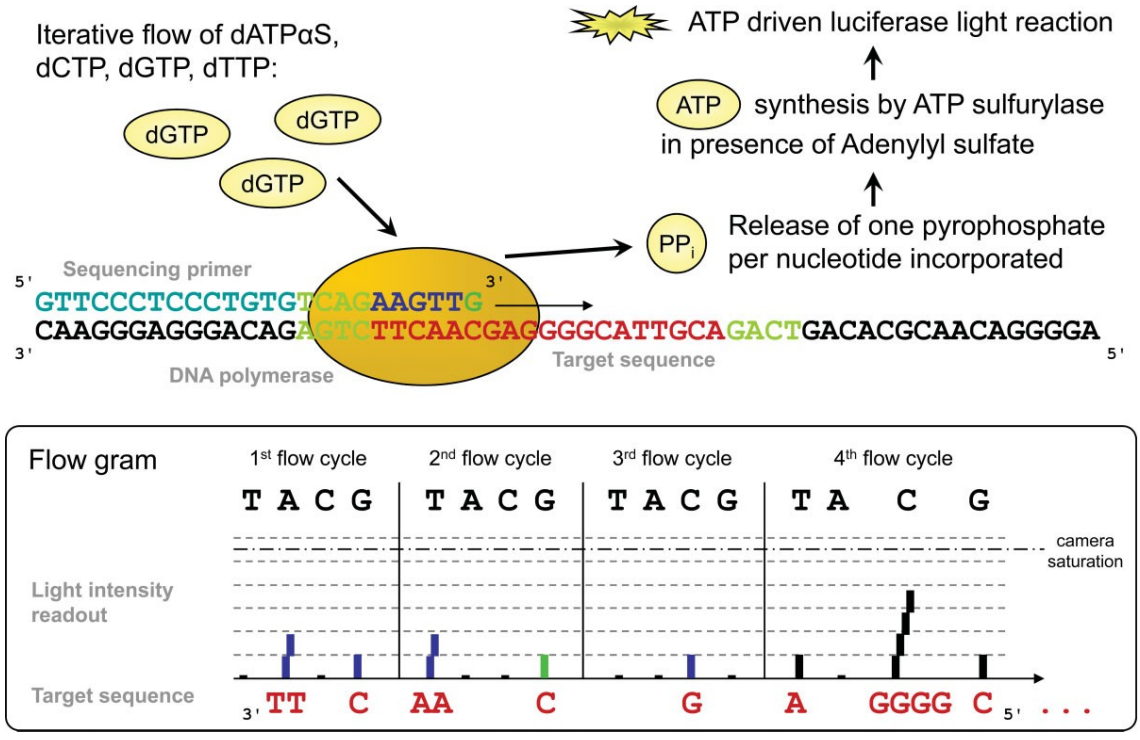
Amplifikace

V následujícím kroku je DNA knihovna namnožena pomocí emulzní PCR. DNA kuličky jsou vloženy spolu s dalšími komponentami reakce (DNA polymeráza, primery, nukleotidy, pufr) do roztoku oleje a vody. Tato směs je promixována a dojde k vytvoření emulze. Kapičky vody vytvořené v olejovém roztoku obsahují vždy jednu kuličku s navázaným fragmentem DNA. Vzniknou mikroreaktory, ve kterých probíhá amplifikační reakce. Následuje rozbití mikroreaktorů a vyjmutí DNA kuliček obsahujících klony vždy jednoho z původních fragmentů. Obohacené kuličky jsou imobilizovány na povrchu titrační destičky (PTP- PicoTiterPlate), která obsahuje stovky tisíc jamek širokých 44μm. Velikost DNA kuličky (cca 26 μm) zajišťuje, že v každé jamce je umístěna právě jedna kulička. Enzymy potřebné pro sekvenační reakci (luciferáza a ATP sulforyláza) jsou navázané na jiných, menších kuličkách, které

zároveň mají za úkol DNA kuličky v jamkách imobilizovat.

Sekvenace

Následuje sekvenační reakce. V každém kroku je titrační destička pokryta roztokem obsahujícím jeden z nukleotidů dCTP, dGTP, dTTP a dATPαS (dATPαS- deoxy-adenosine-5'-(α-thio)-trifosfát). dATPαS se využívá místo dATP z toho důvodu, že na rozdíl od dATP není substrátem luciferázy. Pokud dojde k začlenění daného nukleotidu, je uvolněn fosfát, který je pomocí ATP sulforylázy a adenosin 5'- fosfosulfátu převeden na ATP. ATP dodává energii luciferázové reakci, při které je luciferin oxidován enzymem luciferázou na oxyluciferin, přičemž dojde k uvolnění světelného kvanta, které je detekováno citlivou kamerou umístěnou na spodní straně titrační destičky. Intenzita světla je přímo úměrná množství inkorporovaných nukleotidů. Nevyužité ATP a dNTPs jsou degradovány apyrásou. Výše popsaný princip sekvenátorů firmy 454 je zobrazený na obrázku 2.



Obr. 2: Princip sekvenátorů firmy 454 založený na pyrosekvenační reakci

Zdroj: (Kircher and Kelso, 2010)

2.2.3.2. Genome Analyzer IIx (Illumina/Solexa)

První přístroj firmy Illumina (dříve Solexa) byl představen v roce 2006 a jmenoval se Genome Analyzer. V současné době je na trhu tento přístroj ve verzi Genome Analyzer*IIx*[4] , který dokáže číst sekvence dlouhé až 150 bp.

Princip

Přístroje firmy Illumina jsou založeny na principu sekvenování syntézou.

Příprava knihovny

Odebraný vzorek DNA je pomocí stlačeného vzduchu rozlámán na fragmenty o průměrné délce 800 bp (nebulizace). Na volné konce vzniklých fragmentů jsou navázány dva různé adaptéry a fragmenty jsou elektroforeticky rozděleny podle své molekulové hmotnosti. Pro další zpracování jsou z gelu extrahovány fragmenty o velikosti 150-200 bp.

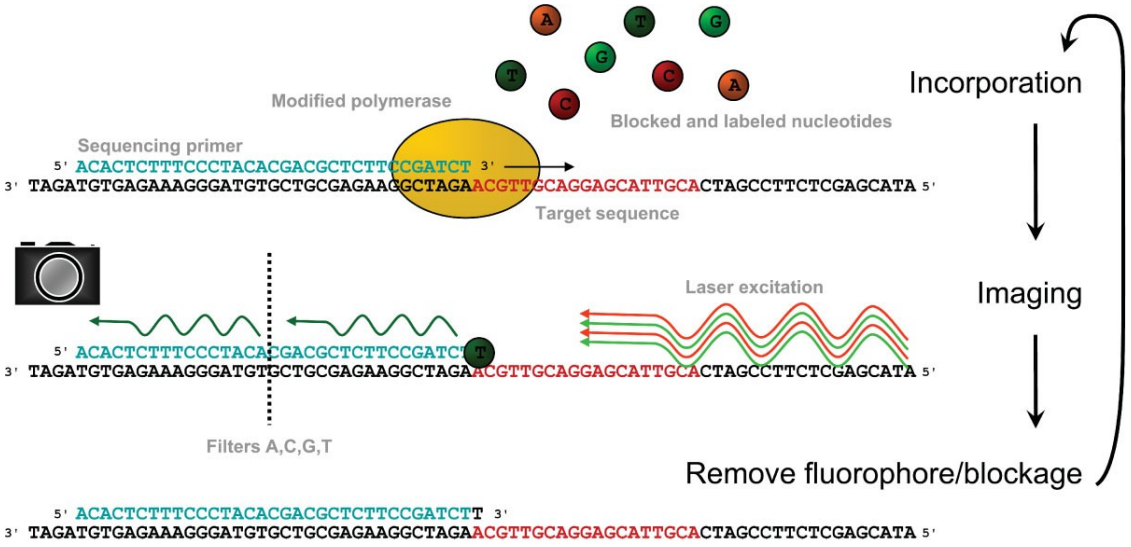
Amplifikace

Fragmenty DNA jsou pomnoženy metodou, která se nazývá můstková amplifikace. Probíhá na sklíčku, jehož povrch je pokryt oligonukleotidy komplementárními k oběma typům adaptérů, navázaných v průběhu přípravy knihovny. Jednovláknové fragmenty jsou přichyceny oběma konci k jednomu z adaptérů, tím jsou imobilizovány a následuje syntéza k nim komplementárního řetězce. Obě vlákna jsou pak oddělena denaturací a celý cyklus se opakuje, dokud nevznikne dostatečné množství kopií původních fragmentů.

Sekvenace

Vlastní sekvenace začíná přidáním čtyř fluorescenčně označených nukleotidů, jejichž 3'-OH konce jsou chemicky inaktivovány a syntéza se tudíž zastaví po připojení jediného nukleotidu. Při tomto připojení vzniká fluorescenční signál, který je detekován za použití laseru. V dalším kroku je chemicky odstraněna fluorescenční značka a tím je odblokovaný 3' konec připraven na začlenění další báze. Princip sekvenátoru firmy Illumina je zobrazen na obrázku 3.

[4] http://www.illumina.com/systems/genome_analyzer_iix.ilmn



Obr. 3: Princip sekvenátoru firmy Illumina

Zdroj: (Kircher and Kelso, 2010)

2.2.3.3. SOLiD (Applied Biosystems)

První NGS sekvenátor firmy Applied Biosystems (SOLiD) byl na trh uveden v říjnu 2007. Tento přístroj četl úseky dlouhé 35 bází. V současnosti je na trhu SOLiD™4 System, který dokáže číst úseky dlouhé až 50 bází.

Princip

Sekvenátor SOLiD pracuje na odlišném principu, než sekvenátory ostatní, a to na principu ligace.

Příprava knihovny

Vzorek DNA je pomocí ultrazvuku rozštěpen na fragmenty dlouhé 150-200bp. Na 5' konec fragmentu je ligován adaptér P1 a na 3´ konec adaptér P2. Takto upravené fragmenty tvoří DNA knihovnu.

Amplifikace

Vytvořená knihovna je amplifikovaná pomocí emulzní PCR, stejně jako u sekvenátoru firmy 454. DNA knihovna je vložena do zkumavky spolu s DNA polymerázou, P1 a P2 primery a kuličkami obsahující oligonukleotidy o sekvenci

komplementární k P1 adaptéru. Výsledkem jsou kuličky obsahující klony vždy jednoho z původních fragmentů. 3´ konce fragmentů jsou modifikovány a kovalentně přichyceny na povrch skla, které může být fyzicky rozděleno až na osm segmentů

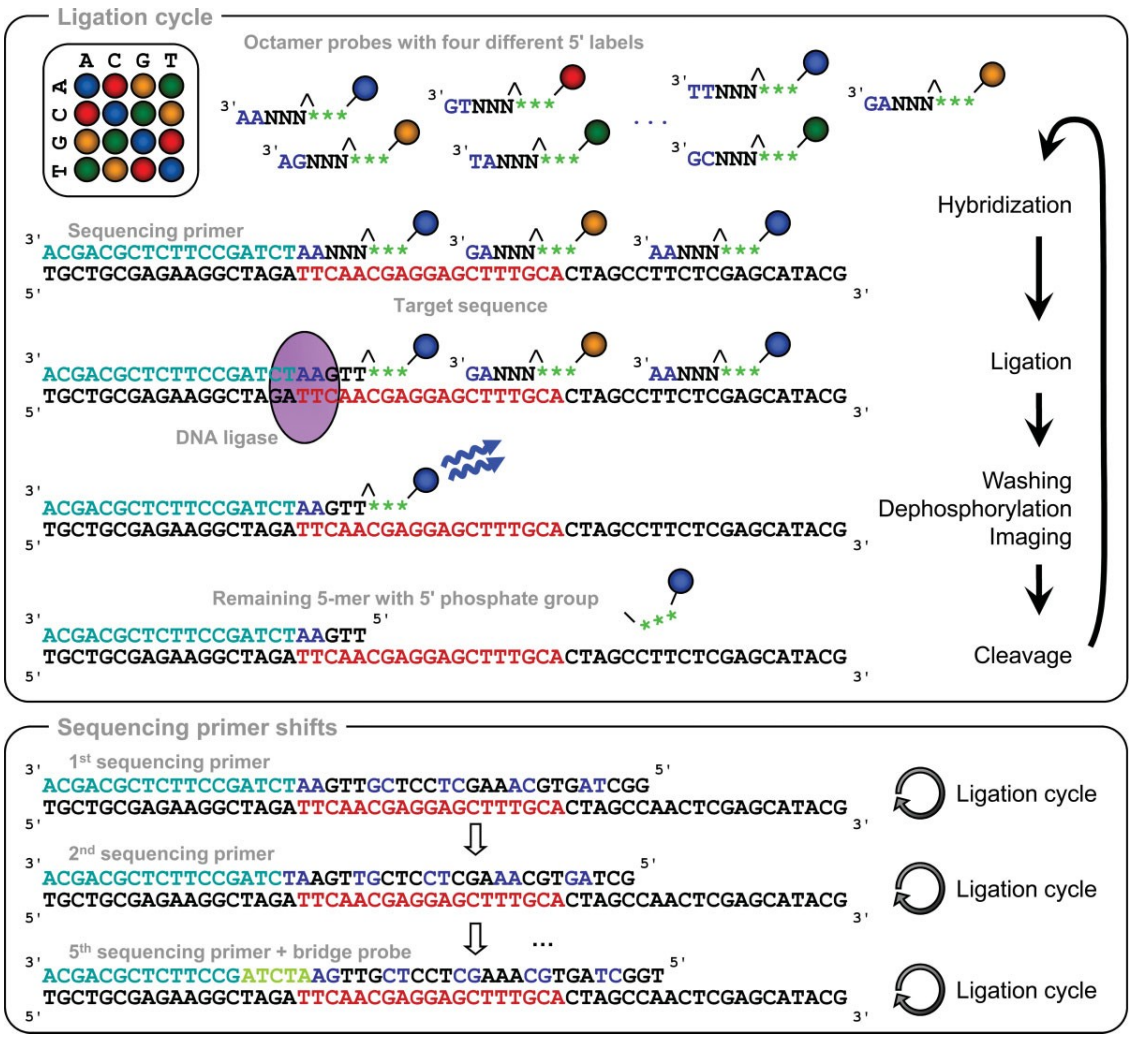
Sekvenace

Na povrchu mikroskopického skla může být navázáno až 800 miliónů kuliček obsahující identické kopie templátů. Na každé kuličce dochází k oddělené sekvenační reakci. Do tohoto kroku jsou si všechny NGS metody podobné. V následujících krocích používá SOLiD System velice odlišnou metodu sekvenování.

Nejprve jsou do reakce přidány primery komplementární k adaptéru navázanému v předchozích krocích. Sekvenace probíhá v pěti až deseti cyklech, během nichž jsou přidávány fluorescenčně označené oktamery, které jsou pomocí ligázy připojeny ke tvořícímu se řetězci. První dvě pozice na 3´ konci jsou báze charakterizované fluorescenční značkou , na pozicích 3, 4 a 5 jsou báze degenerovány tak, aby se mohly navázat na kteroukoliv ze čtyřech bází. Poslední tři pozice jsou univerzální báze nesoucí fluorescenční značku, která charakterizuje první dvě báze. V následujícím kroku dojde k odštěpení posledních třech nukleotidů spolu s fluorescenční značkou a nastává další cyklus.

Po dokončení těchto pěti až deseti cyklů dochází k resetování vlákna, tedy odstranění navázaných primerů i oktamerů. Dochází k navázání primeru o jeden nukleotid kratší a následuje další kolo ligace oktamerů.

Tento systém se nazývá kódováním pomocí dvou bází. Máme 16 možných kombinací dinukleotidů a jen čtyři odlišné fluorescenční značky, takže jednou barvou jsou označeny čtyři různé dvojice bází. Abychom mohli identifikovat bázi, nestačí nám znát pouze jednu barvu. Musíme znát první bázi a pomocí té určíme bázi následující. Výhodou tohoto systému je, že dokážeme odlišit chybu systému (záměna jedné barvy oproti referenční sekvenci) od opravdového polymorfismu (záměna dvou barev oproti referenční sekvenci). Princip sekvenátoru SOLiD je znázorněn na obrázku 4.



Obr. 4: Princip sekvenátrou SOLiD

Zdroj: (Kircher and Kelso, 2010)

2.2.3.4. HeliScope (Helicos)

V roce 2008 byl představen první přístroj, který dokáže číst jednotlivé molekuly DNA.

Princip

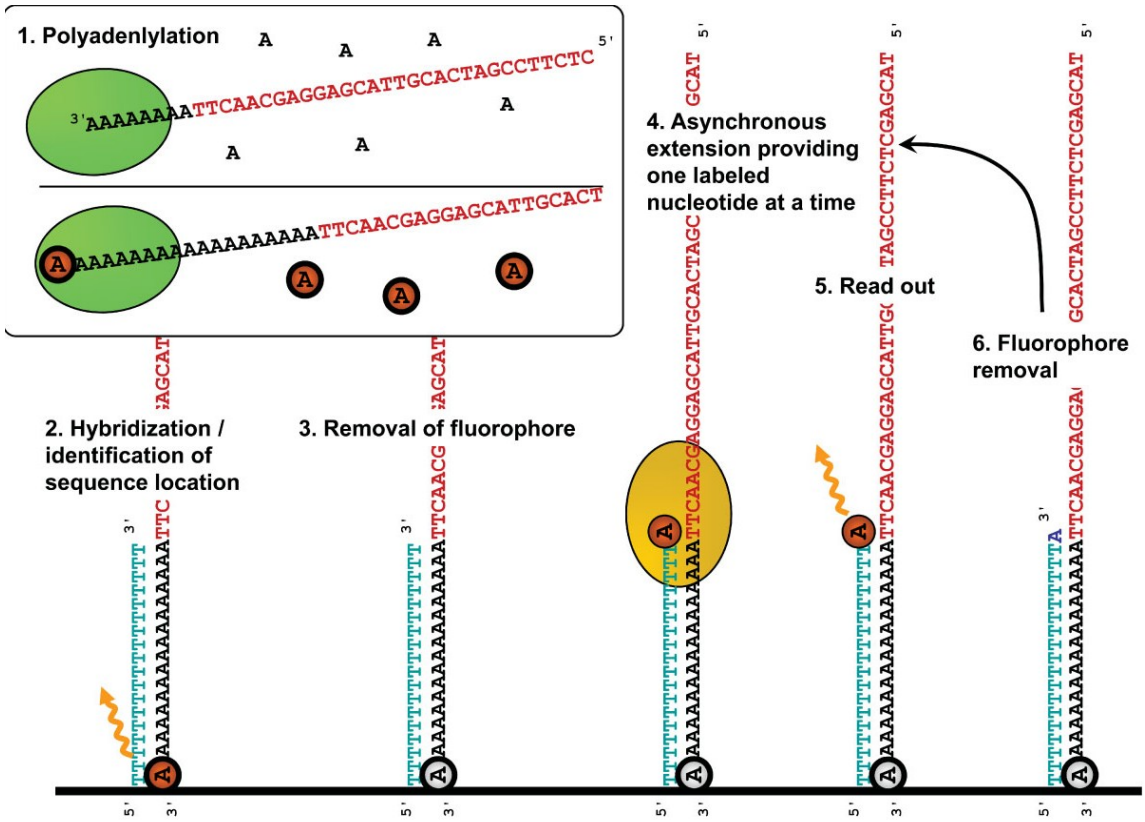
HeliScope funguje na principu asynchronní virtuální terminátorové chemie. Výsledkem jsou sekvence různých délek, tak jako v případě sekvenátoru 454. Pracuje na principu sekvenace syntézou a je to první přístroj, který dokáže číst jednotlivé molekuly DNA nebo RNA. To znamená, že odpadá krok amplifikace vzorků. Tento přístroj čte úseky dlouhé až 70 bází.

Příprava knihovny

Vzorek DNA je rozdělen na fragmenty dlouhé 100-200 bází. Na 3´ konec každého vlákna je navázán adaptér s fluorescenční značkou. Po přidání vzorku DNA na destičku hybridizují vlákna na oligonukleotidy umístěné na jejím povrchu, které jsou komplementární k navázaným adaptérům. Díky tomu, že přístroj čte jednotlivé molekuly, mohou být vlákna DNA rozložena na povrchu s velkou hustotou.

Sekvenace

Laserem jsou detekovány fluorescenční značky, které nám po sejmutí citlivou kamerou určí umístění jednotlivých vláken v dvourozměrném prostoru destičky. Fluorescenční značka je poté odstraněna a následuje první cyklus sekvenování. Je přidána DNA polymeráza s jedním druhem fluorescenčně značeného nukleotidu. Nukleotidy jsou inkorporovány do prodlužujícího se řetězce a poté je odstraněna polymeráza a všechny volné nukleotidy. Pomocí fluorescenčního signálu jsou určena a zaznamenána vlákna, na kterých došlo k začlenění daného nukleotidu. Fluorescenční značka je odstraněna a cyklus může začít znovu, postupně se všemi následujícími nukleotidy. Princip sekvenátoru HeliScope je znázorněn na obrázku 5.



Obr. 5: Princip sekvenátoru firmy Helicos

Zdroj: (Kircher and Kelso, 2010)

2.3. Metody třetí generace

Všechny výše zmíněné metody se v mnoha ohledech od sebe liší - ale jednu věc mají společnou. Tou je mechanismus, kterým je chemická informace uložená v pořadí bází převedena do digitální podoby. Nejprve je tato chemická informace převedena na světelný signál, který je detekován optickou soustavou a následně počítačově zpracován. Tento mezikrok vyžaduje použití drahých reagencií a detekční soustavy sestávající z laseru, který vybudí fluorescenční signál a citlivé kamery, která tento signál dokáže detekovat. V neposlední řadě jsou kladeny vysoké nároky na hardwarové a softwarové vybavení, které je nezbytné pro zpracování velkého množství získaných dat. Tento mezikrok celý sekvenační proces značně zpomaluje a zároveň prodražuje. Další omezení stávajících technologií (kromě sekvenátoru HeliScope) je způsobeno množstvím chyb, které vznikají v průběhu amplifikačního kroku. Z těchto důvodu jsou vyvíjeny nové techniky, které mají za cíl se těmto nedostatkům stávajících sekvenačních technologií vyhnout (Perkel, 2011). Existuje přehled vyvíjených technik a udělených grantů podporujících vývoj těchto nových technik[5]. V nedávné době se staly komerčně dostupné dvě nové technologie, Personal Genome Machine od firmy Ion Torrent a technolgie SMRT (Single Molecule Real Time Sequencing) firmy Pacific Biosciences).

2.3.1. Personal Genome Machine (Ion Torrent / Life Technologies)

Firma Life Technologies uvedla v prosinci 2010 přístroj s názvem Personal Genome Machine (PGM), který je založený na detekci změny pH v důsledku začlenění báze do nově vznikajícího DNA řetězce. Tento přístroj využívá křemíkových polovodičových čipů (CMOS) s mnoha miniaturními jamkami propojenými chodbičkami pro dopravení reagencií do každé z jamek obsahujících molekuly templátové DNA, která byla v průběhu přípravy knihovny amplifikována. V každém kroku je čip zaplaven směsí obsahující DNA polymerázu spolu s jedním z nukleotidů. V případě, že je daný nukleotid komplementární k templátu, je polymerázou začleněn a v důsledku této chemické je z každého začleněného nukleotidu uvolněn jeden vodíkový iont, který způsobí změnu pH. Tato změna je detekovaná pH senzitivní vrstvou, která tvoří dno jamek. Změna pH je pomocí citlivého voltmetru převedena

[5] http://www.genome.gov/27541190

přímo do digitální podoby. Změna pH je přímo úměrná počtu začleněných nukleotidů. Jedná se o první přístroj na trhu, který chemická data převádí přímo na digitální signál[6].

2.3.2. SMRT Sequencing (Pacific Biosciences)

SMRT (Single Molecule Real Time Sequencing), jednomolekulový přístup umožňující sekvenování v reálném čase, je vyvíjený firmou Pacific Bioscience. Tato firma oznámila 27. dubna 2011 prodej prvního přístroje s názvem PacBio *RS*[7]. Jedná se o první komerčně dostupný přístroj třetí generace, který dokáže sekvenovat jednotlivé molekuly DNA v reálném čase.

Technologie SMRT (Korlach et al., 2008) je založena na sledování inkorporace fluorescenčně značených nukleotidů do prodlužujícího se řetězce DNA. DNA polymeráza je imobilizována na dně miniaturní komůrky, jejíž průměr je 70 nm. Každý z nukleotidů je na 5' konci značný jinou fluorescenční značkou a vždy po začlenění jednotlivé báze do prodlužujícího se řetězce dojde k odštěpení flourescenční značky společně s fosfátem. Začleňování báze trvá DNA polymeráze několik milisekund, během kterých je možné flourescenční signál detekovat pomocí laserového paprsku, který prosvětluje skleněné dno komůrky. Rychlost sekvenování je ~10 bází za sekundu. Délka čtených úseků je 250 až 6 kb.

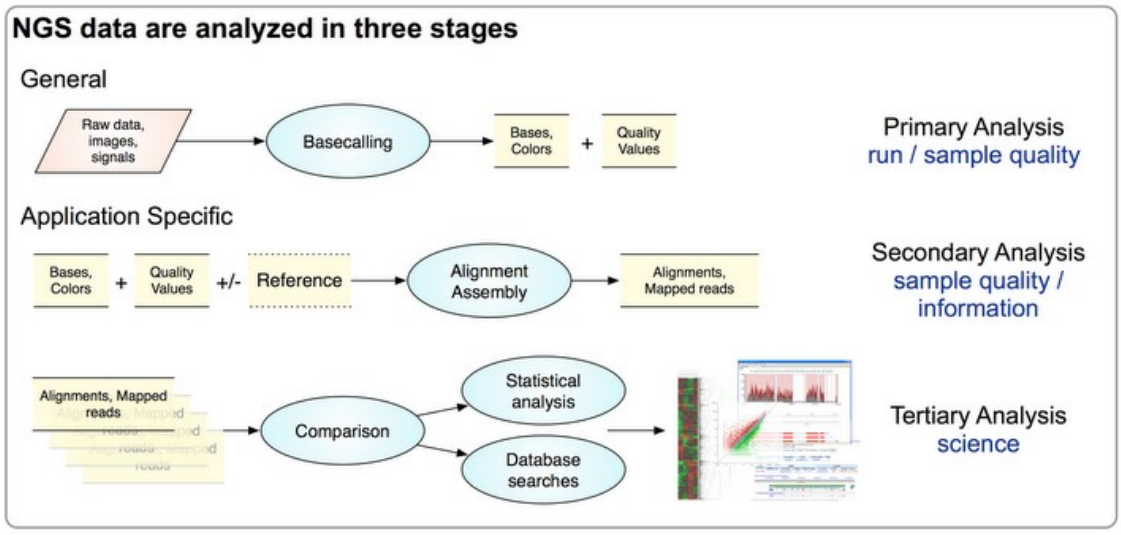
[6] http://www.iontorrent.com/

[7] http://pacificbiosciences.com/sites/default/files/press_release_assets/Commercial%20Shipping%20Announcement_042711_Final_0.pdf

2.4. Analýza dat

Biochemická podstata všech v současnosti rozšířených technologií je různá, ale všechny z nich pracují na základě miniaturizace a paralelizace klasické Sangerovy metody a vizualizace navázání báze pomocí světelného signálu, který je snímán citlivou kamerou. Výstupem ze všech současných NGS sekvenátorů je obrovské množství obrázků, které je nutné podrobit bioinformatické analýze abychom zjistili primární sekvenci čtených úseků. Tomuto kroku se říká primární analýza. Získané čtené úseky jsou poté v sekundární analýze mapovány na referenční sekvenci (resekvenování) nebo je z nich seskládán celý původní genom (*de novo* sekvenování). Poté, co máme čtené úseky namapované na referenční genom, následuje terciální analýza. Jejím cílem je získat informace uložené v pořadí bází a liší se podle toho, zda-li jsme sekvenovali genom, exom, transkriptom, či epigenom. Následují statistické analýzy, porovnávání a anotace nalezených variant a případné propojení s daty uloženými v databázích a propojení s jinými druhy analýz, jako je například vazebná analýza a expresní analýza. Fáze analýzy dat jsou znázorněny na obrázku 6.

Vzhledem k tomu, že praktická část se zabývá zpracováním dat získaných sekvenováním pomocí sekvenátoru SOLiD, je následující popis jednotlivých fází analýzy dat vysvětlován na datech získaných tímto sekvenátorem.



Obr. 6: Znázornění postupu zpracování sekvenačních dat.

Zdroj: http://finchtalk.geospiza.com/2010/04/bloginar-standardizing-bioinformatics.html

2.4.1. Primární analýza dat

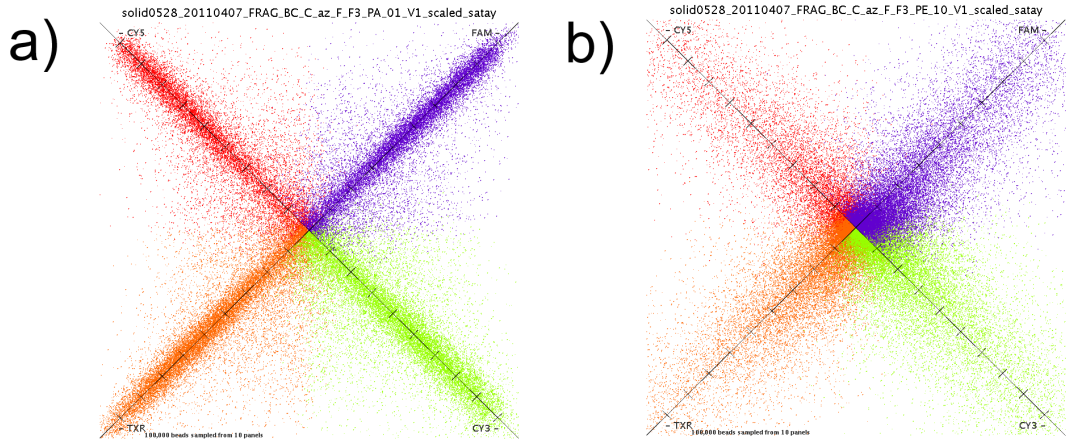
Všechny současně využívané sekvenátory jsou založeny na detekci světelného signálu (chemiluminiscenčního v případě firmy 454 a fluorescenčního v ostatních případech) v jednotlivých sekvenačních cyklech.

Tyto světelné signály jsou snímány citlivou kamerou a zaznamenávány ve formě obrázků. V každém sekvenačním cyklu vznikne obrázek nesoucí informaci o pozicích, ve kterých došlo k inkorporování bazí. Po skončení sekvenace dochází počítačovému zpracování obrázku vzniklých během sekvenace a k rekonstrukci sekvence jednotlivých čtených úseků. Pro tento proces se používá anglický termín „base calling"- tedy v překladu „volání bází".

Nejprve je potřeba identifikovat místa, na kterých se nacházejí kuličky či clustery reprezentující DNA fragmenty. Tato místa jsou zaznamenána do sítě souřadnic a použita pro identifikaci pozic jednotlivých čtených úseků mezi všemi obrázky.

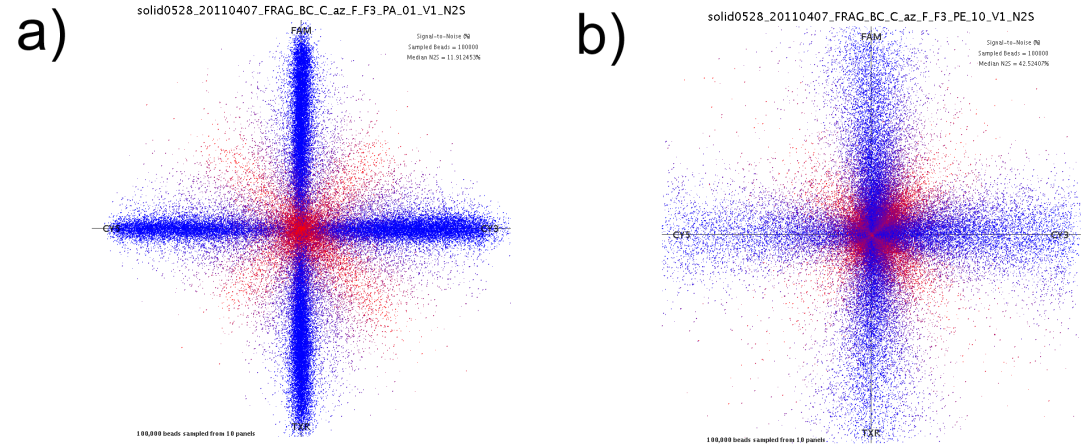
Následuje analýza a normalizace intenzity detekovaného signálu, doby expozice a přiřazování signálů k jednotlivým pozicím v jednotlivých cyklech. Obrázek 7 znázorňuje relativní intenzity signálů detekovaných během jednoho cyklu pomocí jednoho primeru (SOLiD). Osy znázorňují jednotlivé barevné kanály. Body, které jsou mezi osami, mohou představovat polyklonální kuličky, které obsahují více různých fragmentů DNA či kuličky, ve kterých došlo k detekci signálu z vedlejších pozic. Směrem od středu vzrůstá intenzita barevného signálu.

Také je možné vidět, že zastoupení všech čtyřech barev je relativně uniformní. Obrázek 7a) pochází z prvního cyklu ligace prvního primeru, kdežto obrázek 7b) pochází z posledního cyklu ligace posledního primeru.



Obr. 7: Relativní intenzity signálů detekovaných během jednoho cyklu pomocí jednoho primeru sekvenátoru SOLiD. a) intenzity během prvního cyklu ligace prvního primeru b) intenzity během posledního cyklu ligace posledního primeru

Obrázek 8 znázorňují odstup signálu (modře) od šumu (červeně) a ukazují tak podíl kvalitních a nekvalitních kuliček. Obrázek 8a) pochází z prvního cyklu ligace prvního primeru, kdežto obrázek 8b) pochází z posledního cyklu ligace posledního primeru. Z uvedených obrázků 7 a 8 lze vidět, že intenzita fluorescence postupně klesá s každým sekvenačním cyklem, snižuje se odstup signálu od šumu a přesnost určení jednotlivých bází.



Obr. 8: Odstup signálu (modře) od šumu (červeně) v případě a) prvního cyklu ligace prvního primeru b) a v případě posledního cyklu ligace primeru

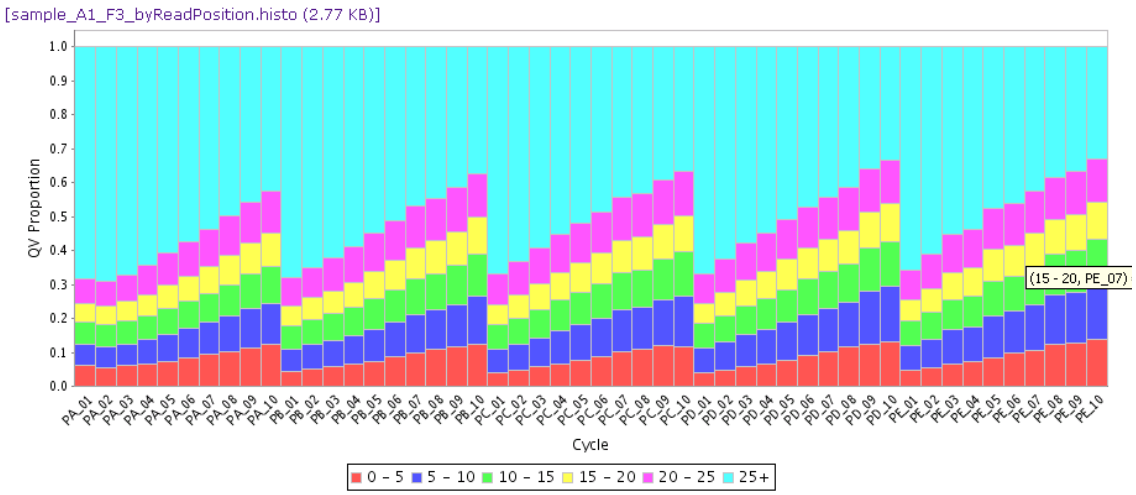
V důsledku snižování odstupu signálu od šumu dochází ke zvýšené pravděpodobnosti špatného určení báze. Pro každou osekvenovanou bázi je vypočteno skóre kvality (QV), často nazýváno jako PHRED skóre (Ewing and Green, 1998). Skóre kvality (QV) je záporný dekadický logaritmus pravděpodobnosti, že je daná báze špatně určená (P), viz. tabulka 1.

$$QV = -10 \cdot \log_{10} P$$

QV (PHRED)	Pravděpodobnost P	Přesnost
10	1 z 10	90,00%
20	1 ze 100	99,00%
30	1 z 1000	99,90%
40	1 z 10000	99,99%

Tab. 1: Skóre kvality a vztah mezi přesností a pravděpodobností špatného určení báze

Počet špatně určených bází má se zvyšujícím se počtem cyklů rostoucí tendenci, nicméně po každém resetování primeru dochází ke snížení počtu sekvenačních chyb (Obr. 9).



Obr. 9: Obrázek znázorňuje počet bází v dané kvalitě (QV) během jednotlivých sekvenačních cyklů

2.4.1.1. Formáty výstupních dat

Mezi nejčastěji využívaný formát výstupních dat primární analýzy patří formáty FASTA a QUAL. První z nich nese informace o sekvenci čtených úseků (FASTA, či pro sekvenci kódovanou v barevném prostoru CSFASTA) a druhý z nich informaci o kvalitě bází v jednotlivých čtených úsecích (QUAL).

FASTA a CSFASTA

Formát FASTA sestává z hlavičky, podle které jsou přečtené úseky identifikované. Hlavička nese informace o sekvenačním panelu, ze kterého pochází (2), souřadních kuličky (souřadnice x=14 a y = 26) a jako poslední údaj zde můžeme najít sekvenační značku (F3). Druhý řádek představuje sekvenci čteného úseku buď v nukleotidovém prostoru (báze A, C, T, G a N pro neurčenou bázi) nebo v barevném prostoru (barvy jsou reprezentovány čísly). T značí poslední bázi primeru, která je v případě sekvenátoru SOLiD nezbytná pro dekódování nukleotidové sekvence z barevného prostoru.

```
>2_14_26_F3
T0112131222002211230321112210212101313322221O1
```

QUAL

QUAL formát je velmi podobný formátu FASTA, ale místo informace o sekvenci nese informaci o kvalitě jednotlivých bází. Hlavička sestává ze stejných částí jako ve formátu FASTA a slouží k jednoznačné identifikaci přečteného úseku a zároveň k následnému spárování se správným čteným úsekem ke kterému hodnoty kvality patří.

```
>2_14_26_F3
24 24 22 27 23 10 13 13 20 19 19 18 24 20 22 12 14 5 20 17 14 20
18 17 19 11 21 19 13 13 12 25 9 19 19 6 5 12 20 13 11 8 12 7 14
```

FASTQ

Informace formátů FASTA a QUAL se často kombinují do jednoho souboru, čímž vznikne formát FASTQ. Obsahuje čtyři řádky. První z nich je název čteného úseku, druhý je jeho sekvence, třetí řádek obsahuje znovu název čteného úseku a čtvrtý řádek nese informace o kvalitě jednotlivých bází. Kvalita je v tomto formátu kódována pomocí ASCII znaků z důvodu menšího datového objemu jednoho kódovaného znaku

```
@1_21_104
T3013210122102202300222021211113102103220320201121O
+
:@&5==?60;/?+;3+;%'5%<&0&&**('540)%7@&'.8+%%'(5)%-
```

2.4.2. Sekundární analýza dat

Primární analýzou jsme získali velké množství krátkých úseků reprezentujících původní genetickou informaci. Následující sekundární analýza se dá rozdělit podle toho, zda-li jsme sekvenovali organismus neznámý (*de novo* sekvenování), či organismus, který byl již v minulosti osekvenován (resekvenování) a tudíž máme k dispozici referenční sekvenci. V případě resekvenování je naším úkolem přiřadit osekvenované úseky ke správným místům referenční sekvence. Jedná se o tzv. alignment či mapování. V případě, že sekvenujeme *de novo*, je náš úkol složitější. Musíme sekvenci zkompletovat z právě osekvenovaných krátkých úseků. Tomuto procesu se říká v angličtině „assembly“ (kompletace). Pro tyto účely je potřeba dosáhnout obrovského pokrytí, aby bylo možné danou sekvenci zpětně zrekonstruovat podle překryvů jednotlivých čtených úseků. Vzhledem k tomu že tato práce je zaměřena na analýzu lidského genomu, jehož referenční sekvence je již známá, tak se zde této problematice nebudu věnovat. Problematiku *de novo* sekvenace přehledně shrnují Miller et al. (2010) a Paszkiewicz a Studholme (2010).

2.4.2.1. Mapovací algoritmy

Nejdůležitějším krokem podmiňujícím úspěšnost sekvenačního experimentu je sekundární analýza dat, ke které slouží algoritmy pro mapování či kompletaci genomů (Flicek and Birney, 2009). Následuje přehled principů, na kterých tyto algoritmy pracují.

Sekvenátory nové generace produkují během jednoho běhu přístroje přibližně až 10[9] čtených úseků dlouhých 25–400 bp. Tyto čtené úseky obsahují jak sekvenační chyby, tak reálné odchylky od referenčního genomu- genetické variace. Následujícím úkolem je namapovat toto obrovské množství čtených úseků k referenčnímu genomu, který má ~3 Gb dat. Tento úkol je nutné zvládnout relativně rychle a zároveň s dobrou přesností.

Lidský diploidní genom sestává z ~6.4 Gb. Pokud sekvenujeme lidský genom, používáme strategii tzv. brokovnice (shotgun). Tato metoda pracuje na principu sekvenování krátkých úseků náhodně po celém genomu. Aby bylo možné z těchto krátkých úseků sestavit původní sekvenci, je nutné osekvenovat každý úsek genomu několikrát tak, aby se všechny krátké úseky navzájem překrývaly. Pro resekvenování je dostačující desetinásobné pokrytí báze (Choi et al., 2009). To pro lidský diploidní genom znamená osekvenovat minimálně 60 Gb dat.

Mapování čtených úseků na referenční sekvenci je řešeno pomocí algoritmů provádějících alignment. Klasické alignovací algoritmy založené na dynamickém programování jako je Smith-Waterman nebo Niedelman-Wunsch, či algoritmy založené na indexování dlouhých úseků v sekvenci templátu (BLAT), či algoritmy založené na kombinaci obou dvou přístupů (např. BLAST) nejsou pro toto obrovské množství dat vhodné (Horner et al., 2010). Důvodem je obrovská výpočetní i paměťová náročnost uvedených algoritmů.

Z tohoto důvodu v současné době vzniká velké množství algoritmů, speciálně upravených pro data produkovaná NGS technologiemi. Základní princip většiny z nich je obdobný.

Alignment je vždy prováděn ve dvou krocích. V první fázi je pro každý čtený úsek identifikováno několik kandidátních pozic (CAL, Candidate Alginment Location), ze kterých by mohl daný úsek pocházet. V této fázi je využíváno heuristických algoritmů, které jsou rychlé, nicméně většinou nezaručují nalezení správného alignmentu. Ve druhé fázi je použito přesnějších, ale výpočetně náročnějších algoritmů (optimální lokální alignment, např. Smith-Waterman či Needelman-Wunsch) pro identifikaci nejlepšího možného alignmentu, který odpovídá původní pozici čteného úseku (Flicek and Birney, 2009).

Přestože existuje velké množství heuristických algoritmů pro mapování krátkých čtených úseků, tak principů na kterých pracují je jen několik. Většinou se mezi sebou liší pouze v programovacích tricích či heuristických metodách, které mají za cíl zvýšit rychlost mapování na úkor co nejmenší ztráty přesnosti (Horner et al., 2010). Nejčastěji využívané metody jsou založeny na hašovacích tabulkách či Burrows-Wheelerově transformaci (BWT). BWT stejně jako hašovací algoritmy mohou být použity pro alignment v prostoru nukleotidovém i prostoru barevném (SOLiD)

Indexování

Obě dvě metody, algoritmy využívající BWT i algoritmy založené na hašovacích tabulkách, využívají v první fázi indexování. Indexování je založené na vytvoření tabulky, která obsahuje „adresy“ krátkých sekvenčních motivů odkazující na místa, odkud tato sekvence pochází. To má výhodu v tom, že neprohledáváme celou sekvenci, ale pouze index. Pokud najdeme záznam v indexu, podíváme se kam nás adresa odkazuje a toto místo označíme jako CAL. Hašovací algoritmy mohou indexovat referenční sekvenci nebo čtené úseky, kdežto BWT algoritmy indexují pouze referenční sekvenci. Indexování referenčního genomu vyžaduje mnoho operační paměti, kdežto indexování čtených úseků vyžaduje operační paměti méně, v závislosti na jejich množství. Velikost indexu závisí vždy na komplexitě sekvence. Pro indexování se využívají následující datové struktury:

a) Sufixové pole

Z řetězce vytvoříme sufixové pole tak, že z něj vytvoříme všechny možné podřetězce (suffixy). Tyto podřetězce abecedně setřídíme a tím dostaneme sufixové pole obsahující indexy všech podřetězců srovnaných v abecedním pořadí (Obr.10). Pro lidský genom je jeho velikost ~12 GB.

Index	Suffix
0	BANANA$
1	ANANA$
2	NANA$
3	ANA$
4	NA$
5	A$
6	$

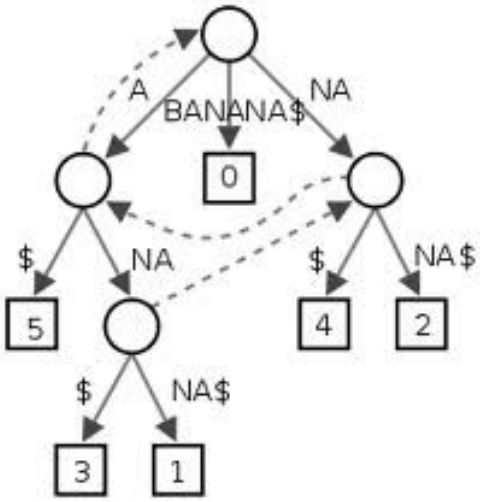
abecedně setříděno →

Index	Suffix
6	$
5	A$
3	ANA$
1	ANANA$
0	BANANA$
4	NA$
2	NANA$

Obr. 10: Sufixové pole na příkladu řetězce BANANA

b) Sufixový strom

Sufixový strom, je stejně jako sufixové pole vytvořen z všech možných podřetězců. Z nich je strom vytvořen tak, že hrany odpovídají podřetězcům daného řetězce. Každý vnitřní vrchol odpovídá nějakému podřetězci (Obr 11). Pro lidský genom je jeho velikost větší než 35 GB.



Obr. 11: Sufixový strom pro textový řetězec BANANA

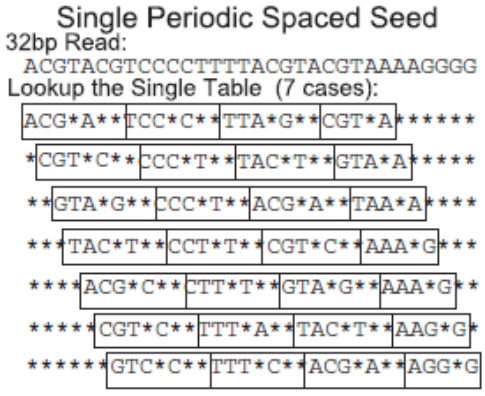
Zdroj: http://upload.wikimedia.org

c) Hašovací tabulka využívající seedy (Seed hash table)

Máme řetězec znaků, ze kterého vytvoříme podřetězce o určité délce. Pro tento podřetězec se využívá anglický pojem seed. Pokud máme řetězec deseti znaků a seed o délce 4, tak ho posouváme postupně po původním řetězci a extrahujeme z něj všechny podřetězce o čtyřech znacích a vždy si k nim připíšeme index vyjadřující při kolikátém posunu seedu byl tento podřetězec vyextrahován. Získané seedy setřídíme podle abecedy, čímž získáme tzv. hašovací tabulku. Velkost této tabulky závisí na délce seedu a na komplexitě vstupního řetězce, nicméně pro lidský genom je jeho velikost ~12 GB.

Tato tabulka může být vytvořena pomocí různých hašovacích funkcí a seedů. Seedy mohou být:

- Kontinuální- posouvají se vždy o jednu pozici a navzájem se tedy překrývají
- Nekontinuální- posouvají se o počet pozic odpovídající délce seedu, takže získané seedy na sebe navazují
- S mezerami- využívající různý počet mezer. Vysvětlení viz. algoritmy využívající hašovací tabulku
- Periodické- jsou tvořeny jedním motivem, který se několikrát za sebou opakuje. Celá tato struktura se postupně posouvá po celé délce čteného úseku (Obr. 12).



Obr. 12: Periodické seedy algoritmu PerM

Zdroj: (Chen et al., 2009)

Algoritmy využívající hašovací tabulku (Hash table algorithms)

Algoritmy založené na hašovacích tabulkách nejčastěji využívají seedů s mezerami. Pro jejich tvorbu se využívají masky. Princip masky uvádí následující příklad:

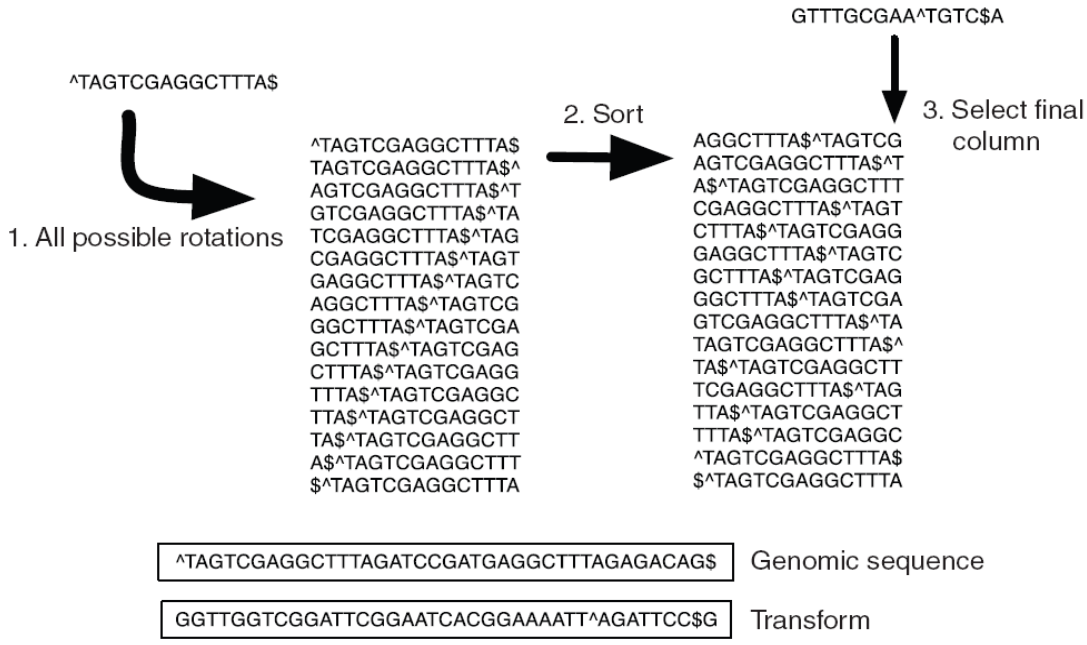
Mějme masku 0011001010 a sekvenci AAGATTACAG. Jedničky určují, ze kterých pozic budou znaky vybrány a nuly určují znaky, které vybrány nebudou. Maska má následující parametry. Velikost (key size, weight), která značí, kolik znaků bude vybráno (key size = 4). Šířka klíče (key width) je počet pozic od prvního vybraného písmene po poslední (key width = 7). Z výše uvedení sekvence tedy budou postupně vybrány znaky na třetí, čtvrté, sedmé a deváté pozici, tedy znaky GAAA (Flicek and Birney, 2009).

Na tomto principu jsou založeny například tyto algoritmy: MAQ, ELAND, SHRiMP ZOOM, Bfast, MOSAIK, SOAP

Burrows-Wheelerova transformace

Druhá hlavní skupina algoritmů je založena na Burrows-Wheelerově transformaci (BWT) a na FM indexování (Ferragina and Manzini, 2000). Koncept sufixového pole je mnohem efektivnější pokud je sufixové pole vytvořeno ze sekvence vytvořené pomocí BWT transformace než ze sekvence původní. FM index umožňuje rychlé vyhledávání podřetězců a pro savčí genomy je stejně velký či menší než je velikost vstupního genomu . To je dáno díky kompresi sufixového pole.

Princip BWT transformace je znázorněn na obrázku 13. V první fázi jsou z referenční sekvence (začínající symbolem ^ a končící symbolem $) vytvořeny všechny její možné rotace (permutace). Tyto rotace jsou abecedně setříděny. Ze setříděného pole rotací původního sekvence se do výstupního zapíše postupně od počátku poslední symbol z každé rotace. Na výstupu je tedy transformovaný vstup rozšířený o ukazatel (symbol $) na konec původního řetězce. Transformovaná sekvence je stejně dlouhá jako ta původní a obsahuje ty stejné znaky, které jsou ale v jiném pořadí. BWT je proces reverzibilní- je tedy možné získat původní sekvenci ze sekvence transformované.



Obr. 13: Princip Burrows-Wheelerovy transformace

Zdroj: (Flicek and Birney, 2009)

Ve druhé fázi je získaná transformovaná sekvence použita pro vytvoření finálního komprimovaného FM indexu. Vytvořený index je prohledáván pomocí jednotlivých čtených úseků a dochází k identifikování několika CAL pro každý z nich. V další fázi jsou použity přesnější alignovací algoritmy pro nalezení nejlepšího výsledného alignmentu odpovídajícímu původní pozici čteného úseku.

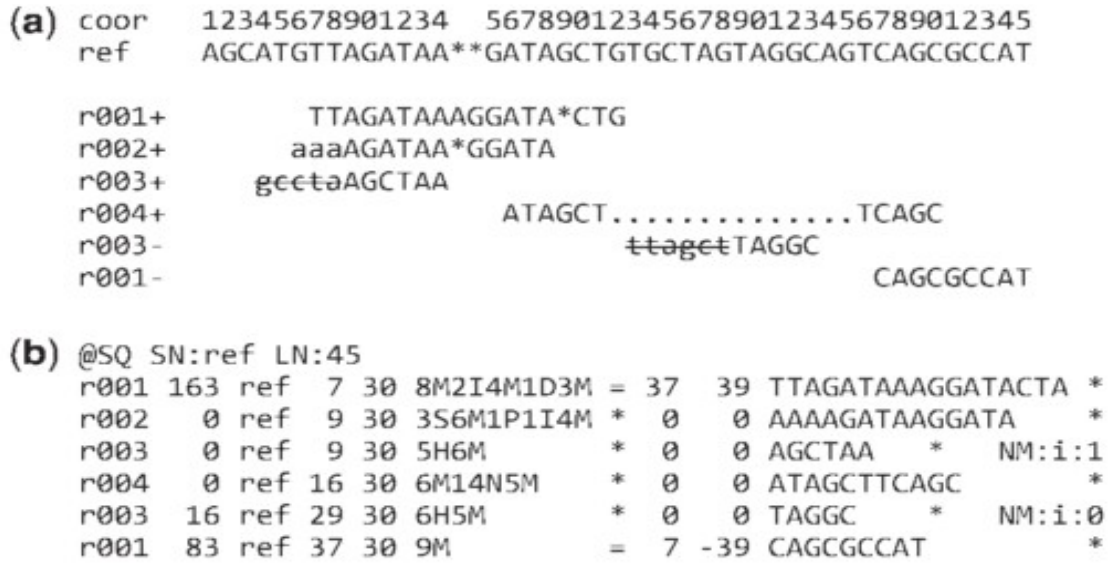
BWT je mnohem rychlejší než hašovací algoritmy při zachování stejné úrovně citlivosti (Flicek and Birney, 2009). Výhodou je také možnost uchovávat kompletní index referenční sekvence na disku a v případě výpočtu ho celý nahrát do paměti počítače (Flicek, 2009). Vzhledem k tomu že všechny mapovací algoritmy se snaží najít nejlepší poměr mezi rychlostí, přesností a nárokem na operační paměť, i u BWT algoritmu najdeme jisté limitace. Nevýhodou například je, že BWA je schopný detekovat alignment pouze v určité editační vzdálenosti (edit distance) která závisí na délce čtených úseků (Li and Durbin, 2009). Editační vzdálenost je počet operací potřebných pro převedení jedné sekvence do podoby sekvence druhé, tedy počet záměna vložených mezer.

Na tomto principu jsou založeny například tyto algoritmy: Bowtie, Bwa, SOAP2.

Formáty výstupních dat

Sekundární analýza slouží k určení původní pozice čteného úseku v referenční sekvenci. Výstupem sekundární analýzy je seznam čtených úseků obohacených o chromozomální souřadnice, skóre kvality alignmentu a další informace. Standartně využívaným formátem pro uchovávání těchto informací je Sequence Alignment/Map formát (SAM) (Li et al., 2009). SAM formát může existovat i v podobě binární, která se nazývá Binary Alignment/Map (BAM).

SAM formát je tabulátory oddělený text, který se skládá ze dvou částí; hlavičky a těla nesoucího informace o samotném alignmentu (Obr14). Jednotlivá pole hlavičky jsou označena symbolem „@“.



Obr. 14: Příklad SAM formátu. a) Alignment čtených úseků a referenční sekvence b) SAM formát. @SQ informace o referenční sekvenci: jméno referenční sekvence (SN) a její délka (LN). Následují povinné položky, jejichž význam je uveden v tabulce X. NM značí počet záměn.

Zdroj: (Li et al., 2009)

Pro každý alignment existuje jedenáct položek povinných (Tab. 2), a neomezený počet položek volitelných.

Název	Popis
QNAME	název čteného úseku
FLAG	informace o čtených úsecích
RNAME	název referenční sekvence
POS	souřadnice na které je čtený úsek namapovaný
MAPQ	kvalita mapování čteného úseku
CIGAR	informace o alignmentu
MRNM	sekvence párové sekvence
MPOS	pozice párové sekvence
ISIZE	velikost inzertu
SEQ	sekvence čteného úseku
QUAL	kvalita jednotlivých bází čteného úseku

Tab. 2: Povinná pole formátu SAM

Položka CIGAR nese informace o alignmentu. Povolené operace jsou shoda (M, match), inzerce (I), delece (D), přeskočená báze (N), báze neshodující se s referenční sekvencí ale přítomná v alignmentu (S, soft clipping), báze neshodující se s referenční sekvencí a nepřítomné v alignmentu (H, hard clipping) a inzerce, která je přítomná i v referenční sekvenci (P, padding).

2.4.3. Terciální analýza dat

Standardním vstupním formátem terciální analýzy je formát SAM nebo BAM vzniklý v průběhu analýzy sekundární. Cílem analýzy terciální je určení míst, kde se námi studovaná sekvence liší od sekvence referenční. Terciální analýza dokáže odhalit SNP a malé inzerce a delece. Algoritmy hledají místa, která se statisticky signifikantně liší od referenční sekvence. Existuje mnoho algoritmů založených na různých principech posuzování kvality čtených sekvencí a kvality jednotlivých bází. Často využívaným nástrojem je programový balík SAMtools, který umožňuje nejen identifikaci odchylek od referenčního genomu, ale má mnoho dalších funkcí, jako je například indexování, odstraňování PCR duplikátů, třídění čtených úseků, převody mezi formáty a filtrování na základě zadaných parametrů. Výstupem terciální analýzy je soubor obsahující nalezené SNP, inzerce a delece. Může se jednat jak o textový soubor oddělený tabulátory, tak často využívaný variant call format (VCF), pro jehož manipulaci se využívá nástrojů VCFtools[8].

[8] http://vcftools.sourceforge.net

3. MATERIÁL A METODY

Abychom prakticky demonstrovali využití NGS technik v biomedicínském výzkumu, aplikovali jsme tyto techniky na reálný případ hledání genu podmiňujícího vzácné dědičné onemocnění, konkrétně autozomálně dominantní formu adultní formy neuronální ceroid lipofuscinózy (ANCL). Toto onemocnění bylo na našem pracovišti zkoumáno již řadu let, pomocí několika nových metod analýzy genomu, jako je vazebná a expresní analýza, genotypování SNP a analýza počtu kopií genomové DNA. Přímým sekvenováním byla ověřena řada kandidátních mutací, z nichž žádná nebyla potvrzena. Zavedení exomového sekvenování do praxe značně snížilo náklady na tuto analýzu a umožnilo využití NGS technik i menším výzkumným laboratořím, jako je ta naše. Z tohoto důvodu jsme se rozhodli vyzkoušet sílu NGS technik při hledání kauzálního genu způsobujícího ANCL.

3.1. Neuronální ceroid lipofuscinóza

Neuronální ceroid lipofuscinózy (NCL) jsou heterogenní skupinou dědičných neurodegenerativních onemocnění. Jedná se o vzácná onemocnění, jejichž incidence je 1- 5:100000 živě narozených a celosvětová prevalence kolem 1- 30:100000 . Podle doby nástupu příznaků se dělí na infantilní, pozdně infantilní, juvenilní a adultní formy. NCL jsou charakterizovány selektivní degenerací neuronů a akumulací autofluorescentního lipopigmentu v lysosomech neuronů i jiných buněk. Střádaný materiál může obsahovat mitochondriální podjednotku c ATP syntázy, případně saponiny A a D (Tyynelä et al., 1993; Elleder et al., 1997). Klinické příznaky zahrnují epileptické záchvaty, křeče, ztrátu zraku, ataxii a progresivní zhoršení pohybových a mentálních schopností.

Dnes je klasifikováno deset různých forem NCL, geny známe u osmi z nich: *CLN1* deficit palmitoyl-protein thioesterázyy *(PPT1), CLN2* deficit tripeptidyl peptidázy (*TPP1), CLN3, CLN5, CLN6, CLN7 (MFSD8), CLN8* and *CLN10* deficit katepsinu D (*CTSD*)[9]. Uvedenými geny kódované bílkoviny fungují jako enzymy (*PPT1, TPP1, CTSD*) nebo jako membránové proteiny, jejichž funkce zatím není detailně známa. Geny způsobující NCL4 a NCL9 (*CLN4* a *CLN9*) nebyly stále identifikovány.

[9] http://www.ucl.ac.uk/ncl/mutation.shtml

Diagnóza je založena na zhodnocení klinického stavu pacienta a průkazu ceroidlipofuscinu v hluboké kožní biopsii. Přínosem může být elektroencefalografické vyšetření (Nijssen et al., 2009) či magnetická rezonance. Pro stanovení jednotlivých typů NCL se provádí enzymatické vyšetření aktivit *PPT1, TPP1* a vyšetření genetické v případě CLN2, CLN3, CLN5, CLN 6, CLN 7, CLN 8.

V současné době neexistuje žádná kauzální léčba, je možná pouze léčba symptomatická. Přesná diagnostika, genetické poradenství a prenatální diagnostika je jediným způsobem jak pomoci postiženým rodinám.

3.1.1. Adultní forma NCL

Adultní forma NCL (ANCL, NCL4) byla popsána v autozomálně dominantní (Kufs, MIM 204300) i recesivní (Parry, MIM 162350) formě dědičnosti. Obě formy jsou jen málo charakterizované a jejich genetická a molekulární podstata je dosud nejasná.

Autozomálně dominantní formu popsal prvně Boehme et al. v roce 1971 u jedenácti pacientů ve čtyřech generacích. Klinicky byla popsána progresivní demence, křeče a myoklonická epilepsie. Patologicky byl popsán úbytek nervové tkáně a akumulace střádavého lipopigmentu ve zbývající nervové tkání. Další výskyt této choroby s familiárním výskytem popsal Ferrer et al. (1980) u šesti jedinců ve dvou generacích. Pacienti trpěli progresivní demencí a mimovolnými pohyby tváří a krku, u jednoho z nich se objevily křeče. Josephson et al. (2001) popsal anglickou rodinu s deseti členy s časným nástupem demence. Nijssen et al. (2002; 2003) popsal holandskou rodinu s autozomálně dominantní NCL se šesti postiženými ve třech generacích a popsal morfologické nálezy v biopsiích a autopsii mozku a v dalších orgánech. Burneo et al. (2003) publikoval familiární výskyt adultní formy u rodiny z Alabamy.

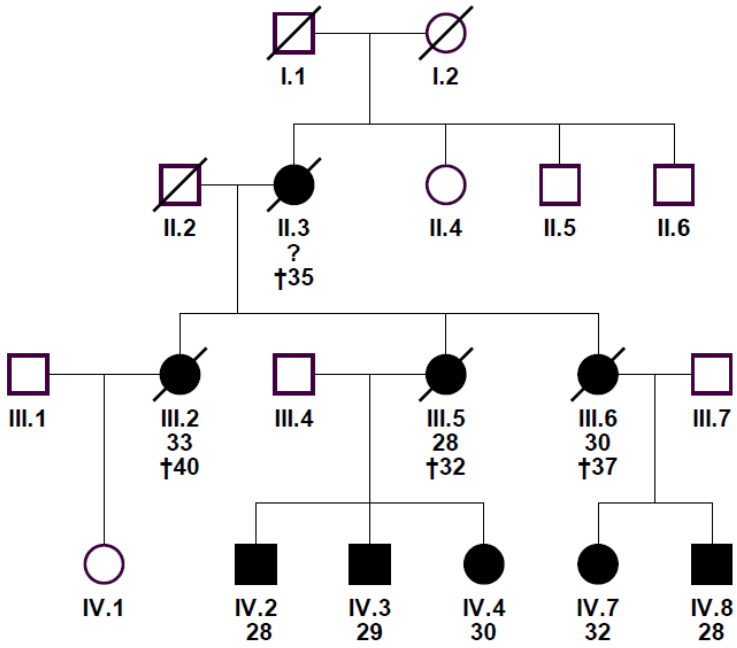
V současné době bylo identifikováno přes 100 pacientů s touto chorobou, převážně se jedná o familiární výskyt. Pokud tedy shrneme klinické příznaky, jedná se o generalizované záchvaty s křečemi, které se objevují mezi 25. - 46. rokem života, myoklonickou epilepsii, ataxii a dysartrii, často se objevují neuropsychiatrické symptomy, především poruchy chování a progresivní demence. Na rozdíl od časných forem se u adultních forem NCL nevyskytují poruchy vidění.

Molekulární podstata této choroby je doposud neznámá.

3.2. Materiál

3.2.1. Biologický materiál

Pro studium molekulární podstaty autozomálně dominantní adultní formy neuronální ceroid lipofuscinózy (ANCL) byly použity biologické vzorky členů rodiny s familiárním výskytem tohoto onemocnění (Obr. 15). Tato česká rodina byla diagnostikována na Ústavu dědičných metabolických poruch Všeobecné fakultní nemocnice v Praze a 1. lékařské fakulty UK v Praze. Diagnóza byla určena na základě klinických příznaků a vyšetření probanda III.6, u kterého se ve 30 letech objevily první příznaky progresivní motorické a kognitivní deteriorace, generalizované záchvaty s křečemi a deprese. Tyto příznaky vedly k úmrtí ve 37 letech. *Post mortem* vyšetření mozkové tkáně prokázalo přítomnost charakteristického střádavého autofluorescenčního lipopigmentu v lysosomech neuronů. V biopsii odebrané z pokožky probanda tento střádavý materiál prokázán nebyl. Diagnóza ostatních členů rodiny byla určena na základě přítomnosti či nepřítomnosti klinických příznaků.



Obr. 15: Rodokmen rodiny postižené ANCL. Černé symboly představují nemocné jedince, otevřené symboly představují jedince zdravé. Věk, kdy se objevily první příznaky onemocnění, je uveden pod symbolem, věk v době úmrtí je označen křížkem (†).

Biologické vzorky, použité pro veškeré níže zmíněné analýzy byly odebrány a zpracovány v laboratoři Ústavu dědičných metabolických poruch Všeobecné fakultní nemocnice v Praze a 1. lékařské fakulty UK v Praze standardními metodami zde používanými.

3.2.2. Chemikálie

Chemikálie na PCR

- 2x Red PCR Master Mix — Rovalab
- MilliQ H_2O

Chemikálie na elektroforézu

- Agaróza — SERVA
- 20x GelRed GelRed^{TM} — *
- GeneRuler™ 100 bp Plus DNA Ladder — Fermentas
- BB pufr — *

* Viz. níže (Příprava roztoků)

Chemikálie na izolaci PCR produktů

- MSB® Spin PCRapace — STRATEC Molecular GmbH
- MilliQ H_2O

Chemikálie na sekvenování DNA

- Sekvenační kit Dye Terminator v3.1 — Applied Biosystems
- MilliQ H_2O

Příprava roztoků

- 10x BB pufr (1 litr) :

100mM tetraboritan sodný dekahydrát (SIGMA)	38,14 g
MilliQ H_2O	1000 ml

- 20x GelRed (500 µl):

100000x GelRed (Biotinum)	2 µl
70% glycerol	71,43 µl
MilliQ H_2O	427,57 µl

3.2.3. Přístroje

Přístroje na PCR

- Laminární box Lamin. Air Typ HV Mini — HOLTEN
- Pipety Research® plus — Eppendorf
- Termocyklér DNA Engine Dyad® — MJ Research
- Vortex-Genie 1 Touch Mixer — Scientific Industries

Přístroje na elektroforézu

- Laboratorní váhy Scout — OHAUSE
- Mikrovlná trouba — Clatronic
- Pipety Research® plus — Eppendorf
- Mikrocentrifuga — Labnet
- Elektroforetický zdroj napětí EV265 — Consort
- Elektroforetická aparatura — Scie-Plas
- UV transiluminátor — Vilber Lourmat
- Fotoaparát PowerShot A520 — Canon

Přístroje na izolaci PCR produktů

- Pipety Research® plus — Eppendorf
- centrifuga 5415 D — Eppendorf
- NanoDrop — NanoDrop Technologies

Přístroje na sekvenování DNA

- Pipety Research® plus — Eppendorf
- Mikrocentrifuga — Labnet
- ABI 3500XL Avant Genetic Analyzer — Applied Biosystems

3.2.4. Software

Genotypování a vazebná analýza

• Sequence Scanner verze 1.0	Applied Biosystems
• Merlin verze 1.1.2.	(Abecasis et al., 2002)
• R-project verze 2.9.2.	http://www.r-project.org
• HaploPainter verze 1.032	(Thiele and Nürnberg, 2005)

Analýza CNV a SNP

• Genotyping Console Software verze 3.02	Affymetrix

Expresní analýza

• Bioconductor verze 2.7.	(Gentleman et al., 2004)
• R-Project verze 2.9.2	http://www.r-project.org
• DAVID verze 6.7	(Huang et al., 2009a; 2009b)

Bioinformatická analýza

• Bfast verze 0.6.4f	(Homer et al., 2009a; 2009b)
• Bioscope v 1.2	Applied Biosystems
• BWA verze 0.5.9-r16	(Li and Durbin, 2009)
• CLCbio Genomics Workbench verze 4	CLCbio
• PerM verze 0.3.3	(Chen et al., 2009)
• Novoalign verze 1.01.08	Novocraft
• SAMtools verze 0.1.8	(Li et al., 2009)
• SeattleSeq Annotation verze 6.0	http://gvs.gs.washington.edu/SeattleSeqAnnotation
• PolyPhen2 verze 2.0.23	(Adzhubei et al., 2010)
• Annovar verze 2011Feb20	(Wang et al., 2010)
• BedTools verze 2.12.0	(Quinlan and Hall, 2010)
• Integrative Genomic Viewer (IGV)	(Robinson et al., 2011)

Ověření kandidátní mutace přímým sekvenováním

• UCSC Genome Browser[10], agloritmus BLAT	(Kent, 2002)
• PrimerPremier	PREMIER Biosoft
• Sequence Scanner verze 1.0	(Applied Biosystems)

[10] http://genome.ucsc.edu

3.3. Metody

3.3.1. Genotypování a vazebná analýza

3.3.1.1. Princip metody

Cílem vazebné analýzy je zjistit, jestli postižení jedinci nesdílejí stejné alely, které by mohly být zodpovědné za vznik onemocnění. Aby mohla být provedena vazebná analýza, je nutná znalost genotypu pro stovky míst v genomu u všech zkoumaných jedinců. Pro genotypizaci se využívají polymorfní genetické markery s vysokou mírou heterozygocie, která je základem pro informativnost analýzy. Používají se například krátké tandemové repetice (STR, Short Tandem Repeat, opakující se jednotky 2 a více nukleotidů) , polymorfismy v délce restrikčních fragmentů (RFLP, Restriction Fragment Length Polymorphism), variabilní počty tandemových repetic (VNTR, Variable Number of Tandem Repeats) či jednonukleotidové záměny (SNP, Single Nucleotide Polymorphism). Vazebnou analýzu lze rozlišit na parametrickou a neparametrickou.

Pro parametrickou analýzu je nezbytné znát model dědičnosti zkoumaného onemocnění. Typ dědičnosti je možné zjistit analýzou rodokmenu či pomocí segregační analýzy, kdy sledujeme segregaci určitého genetického markeru s ohledem na přítomnost či nepřítomnost zkoumaného onemocnění. Dále potřebujeme znát penetranci (neboli pravděpodobnost vzniku nemoci při určitém genotypu), frekvenci alely podmiňující onemocnění v populaci a vliv vnějšího prostředí.

Při odhalování genetické podstaty komplexních znaků jsou nejčastěji užívané neparametrické metody vazebné analýzy, tedy takové, kde nepředpokládáme žádný konkrétní způsob dědičnosti, počet genů nebo míru vlivu negenetických faktorů.

3.3.1.2. Provedení

Analýza genetických markerů byla provedena pomocí genotypovacího čipu GeneChip Human Mapping 10K 2.0 Array (Affymetrix) v Servisní laboratoři funkční genomiky a bioinformatiky na Ústavu molekulární genetiky AV ČR. Pro získání a základní analýzu obrazu byl použit Affymetrix GeneChip Scanner 3000 7G,GeneChip operating Software (GCOS) 1.4. a Affymetrix Genotyping Analysis Software (GTYPE) 4.1.

Pomocí zjištěných genetických markerů byla provedena parametrická vazebná analýza. Při analýze byl předpokládán autozomálně dominantní model dědičnosti s 99% penetrancí, 1% vlivem vnějšího prostředí (phenocopy rate) a předpokládanou frekvencí alely v populaci 0,1 %. Vazebná analýza byla vypočítána za pomoci programu Merlin. Získané výsledky byly zobrazeny pomocí programu HaploPainter a R-project.

3.3.2. Analýza počtu kopií genomové DNA

Analýza počtu kopií genomové DNA (CNV, Copy Number Variation) zkoumá ztráty a duplikace rozsáhlých oblastí genomu (10 kb až 5 Gb). Pokud je frekvence určité varianty v populaci větší než 1 %, jedná se o polymorfismus v počtu kopií (copy number polymorphism).

3.3.2.1. Princip metody

CNV je možné studovat jak pomocí čipových technologií, tak pomocí sekvenování. Nejčastěji používanými metodami pro studium CNV jsou stále čipové technologie (William Blair & Company, 2011). Pro analýzu CNV byl použit SNP genotypovací čip Genome-Wide Human SNP Array 6.0 od firmy Affymetrix. Tento čip obsahuje více než 906600 prób pro detekci SNP a více než 946000 prób pro detekci CNV. Genomová DNA je naštěpena pomocí restrikčních endonukleáz *Nsp I* a *Sty I*. Na konce získaných fragmentů jsou ligovány adaptéry, které jsou komplementární ke kohezním koncům získaných v předchozím kroku. Získané DNA fragmenty jsou amplifikovány pomocí primeru komplementárnímu k sekvenci adaptéru. Podmínky PCR reakce jsou optimalizovány tak, aby byly přednostně amplifikovány fragmenty dlouhé 200 až 1100 bp. Vzorky získané štěpením jednotlivými restikčními endonukleázami jsou smíchány dohromady a vychytány pomocí polystyrénových kuliček s navázanými adaptéry komplementárními k adaptérům navázaných na DNA fragmenty v předchozích krocích. Získaná směs fragmentů DNA je naštěpena, fluorescenčně označena a nalita na čip. Fragmenty DNA hybridizují na próby umístěné na povrchu čipu. Poté je čip zbaven fragmentů, které nehybridizovaly a vyfocen. Následuje analýza získaných obrázků.

3.3.2.2. Provedení

DNA vzorky pacientů II.2, IV.1, IV.2, IV.3, IV.4, IV.7 a IV.8 (Obr. 15) byly genotypovány pomocí čipu Genome-Wide Human SNP Array 6.0 (Affymetrix). Čip byl skenován pomocí GeneChip Scanner 3000 7G (Affymetrix), který je vybavený programem GeneChip Control Console Software 2.01 (Affymetrix). Genotypování bylo provedeno v Servisní laboratoři funkční genomiky a bioinformatiky na Ústavu molekulární genetiky AV ČR.

Získaný CEL soubor byl dále zpracován a analyzován pomocí programu Genotyping Console Software. Signály reprezentující jednotlivé SNP a CNV byly porovnány s referenčním daty která jsou automaticky v uvedeném programu k dispozici. Získaná data byla použita k analýze počtu kopií genomové DNA. Byly uvažovány pouze úseky, které byly pokryty minimálně pěti próbami a byly delší než 10 kb.

3.3.3. Expresní analýza

Genetická informace uložená v jádře buněk nezávisí na typu buněk ani jejich fyziologickém stavu. Úroveň transkripce a translace v jednotlivých buňkách je však závislá na mnoha okolnostech. Abychom pochopili komplexní funkci buněk, tkání a orgánů, je nezbytné zjistit, které geny jsou ve kterých buňkách za různých podmínek exprimovány.

3.3.3.1. Princip metody

Expresi je možné sledovat na dvou úrovních. Jednou možností je analyzovat přítomnost proteinů v buňkách určitého typu za určitých podmínek (proteom). V současné době je studium proteinů relativně složité a nehodí se pro paralelní sledování mnoha proteinů najednou. Druhým způsobem, jak analyzovat genovou expresi je sledování množství mRNA, tedy transkriptomu. Transkriptom je soubor všech molekul mRNA přítomných v buňkách určitého typu za určitých podmínek. Sledování transkriptomu je mnohem jednodušší a vhodnější pro paralelní sledování mnoha molekul mRNA najednou. Sledování genové exprese pomocí analýzy transkriptomu je velmi rozšířené, nicméně je známo, že množství trankriptů v buňce není přímo úměrné množství vzniklých proteinů. Důvodem je kontrola genové exprese na transkripční (splicing, mRNA processing...) a translační úrovni (modifikace proteinů

a jejich degradace). Je tedy důležité si uvědomit, že tímto způsobem neměříme produkt genové exprese, ale transkriptom, který nemusí množství produktu odpovídat.

Základními otázkami, které si v případě analýzy transkriptomu klademe mohou být: je transkript pro gen G v buněčném typu A za podmínek X více nebo méně abundantní než v buněčné typu B za podmínek Y? Buněčnými typy A a B mohou být například normální a rakovinné buňky, podmínky X a Y mohou být identické. V jiném případě lze studovat změnu transkriptomu u stejného buněčného typu za rozdílných podmínek X a Y.

Genovou transkripci je možné sledovat pomocí otevřeného či uzavřeného přístupu. Otevřený přístup je nezávislý na jakékoliv předchozí znalosti, zatímco uzavřený přístup předpokládá znalost sekvence sledovaného souboru genů. Expresní čipy se řadí mezi uzavřené přístupy, což je jeden z důvodů, proč jsou v současnosti nahrazovány sekvenováním, které předchozí znalost sekvence studované molekuly nevyžaduje (William Blair & Company, 2011).

3.3.3.2. Provedení

Pro zjištění, které geny jsou změněně exprimovány ve tkáních pacientů oproti kontrolám byla analyzována RNA odebraná z leukocytů pacientů IV.2, IV.3, IV.7 a IV.8 (Obr. 15).

Expresní analýza byla provedena v Servisní laboratoři funkční genomiky a bioinformatiky na Ústavu molekulární genetiky AV ČR pomocí čipu HumanRef-8 Expression BeadChips (Illumina). Čipy byly skenovány přístrojem Illumina BeadArray Reader a intenzity byly odečteny v programu Illumina BeadStudio Software v3.

Získaná data byla zpracována pomocí programového balíku Bioconductor, který je rozšířením statistického systému R-Project. Data byla normalizována pomocí kvantilové normalizace implementované v balíku Lumi, který je specializován pro zpracování čipových dat firmy Illumina. Pro rozhodování, zda je ten který gen rozdílně exprimován u pacientů oproti kontrolám byl použit modifikovaný t-test implementovaný v balíčku Limma. Korekce na mnohonásobné testování byla provedena metodou False discovery rate (Benjamini and Hochberg, 1995). Za rozdílně exprimované geny jsou považovány ty, jejichž změna byla více než 50%.

Pro funkční anotaci byly uvažovány geny, jejichž exprese byla u pacientů změněna nejméně o 50 % oproti kontrolní skupině a pravděpodobnost, že gen není rozdílně exprimovaný byla menší než 5% (P-value upravená pro mnohonásobná porovnání pomocí Bonferonniho korekce). Pro funkční anotaci změněně exprimovaných genů byla použita databáze DAVID.

3.3.4. Exomové sekvenování

Exomové sekvenování představuje novou metodou, která značně snížila cenu sekvenování pomocí NGS technik, čímž se jejich využití stává dostupnější pro stále větší okruh vědců a v současnosti je častou metodou pro odhalování kauzálních genů vzácných dědičných onemocnění, ať již se jedná o choroby s mendelistickým typem dědičnosti či choroby komplexní. Aby bylo možní studovat exom pacienta, je nejprve potřeba obohatit jeho genomovou DNA o kódující úseky (exony). Existuje několik používaných metod pro obohacení genomu.

3.3.4.1. Princip metody

Pro obohacení genomu se využíván několika různých metod, které jsou schematicky znázorněny na obrázku 16.

(a) Obohacení DNA pomocí čipových technolgií: (Agilent, Roche/Nimblegen a Febit). Próby, které jsou komplementární k cílové sekvenci, jsou imobilizovány na čipu a pomocí hybridizace cíleně vychytávají požadované fragmenty DNA. Poté jsou odmyty fragmenty, které nehybridizovaly a DNA je z čipu uvolněna a dále připravována podle standartního protokolu pro sekvenování.

(b) Biotinylované próby volně plovoucí v roztoku selektivně vychytávají požadovanou DNA. Fragmenty, které hybridizovaly jsou z roztoku vychytány pomocí streptavidinem obalených magnetických kuliček. Próby mohou být jak z DNA (RocheNimbleGen, SeqCap/SeqCap EZ), tak z RNA (Agilent Technologies, SureSelect kit).

(c) Molekulární inverzní proby (MIP, Molecular Infersion Probes). MIP sestávají ze

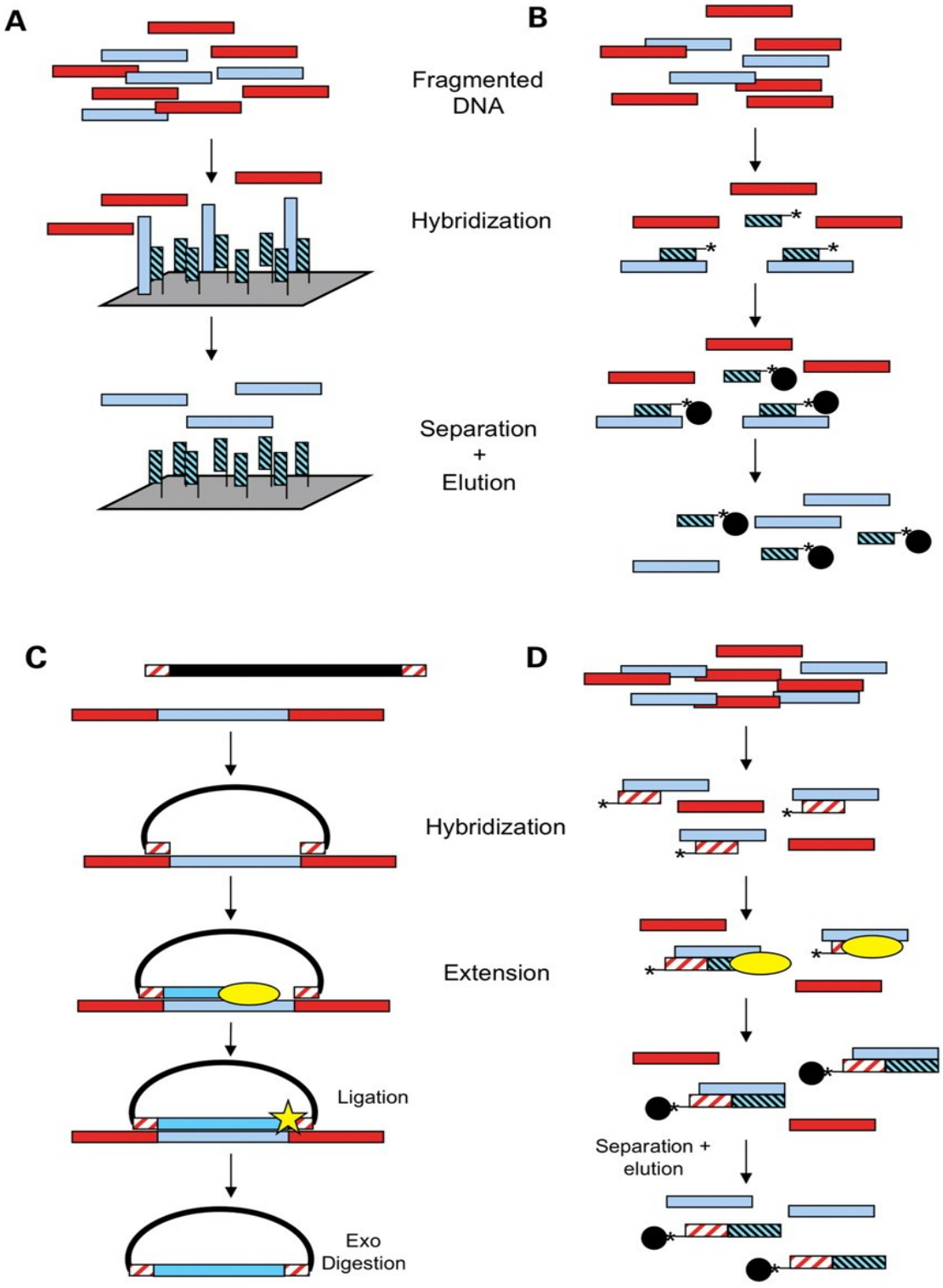
dvou prób komplementárních k požadované sekvenci. Tyto próby jsou spojené univerzálním „spacerem“. Próby nasednou na požadovanou sekvenci a mezera mezi próbami je zaplněna pomocí DNA polymerázy a ligázy. Následuje odbourání zbylé nehybridizované DNA pomocí endonukleáz.

(d) RainStorm PCR - Tato metoda využívá biotinylované primery komplementární k požadované sekvenci. Primery jsou prodlouženy pomocí DNA polymerázy. Výsledná sekvence je zachycena na streptavidinem obalené magnetické kuličky.

Každá z obohacovacích metod je vhodná k jiným účelům v závislosti na velikosti cílové sekvence a v závislosti na počtu vzorků, které potřebujeme obohatit. Pro obohacování lidského exomu (~30 Mb) jsou nejvíce vhodné hybridizační metody založené na čipových technologiích či hybridizaci v roztoku (Mamanova et al., 2010).

3.3.4.2. Provedení

DNA pacienta IV.7 (Obr. 15) byla nejprve obohacena o kódující oblasti pomocí kitu SureSelect Human All Exome Kit (Agilent) a takto získaný vzorek byl osekvenován firmou CeGaT (Germany) na přístroji SOLiD™ 4 System (Applied Biosystems). Získaná data byla následně podrobena bioinformatické analýze.



Obr. 16: Metody obohacování genomu. Světle modré fragmenty reprezentují cílovou genomovou sekvenci, kdežto červené fragmenty reprezentují sekvenci, kterou nechceme obohacovat

Zdroj: (Teer and Mullikin, 2010)

3.3.5. Porovnání mapovacích algoritmů

Abychom zjistili, který mapovací algoritmus je nejvhodnější pro analýzu dat získaných sekvenací exomu pacienta postiženého ANCL, porovnali jsme výsledky získané pomocí různých mapovacích algoritmů. Porovnávané algoritmy jsou uvedeny v tabulce 3.

Algoritmus	Verze	Princip	Zdroj	Dostupnost
Bfast	0.6.4f	hašovací tabulka	http://bfast.sourceforge.net	open source
BioScope	1.2	hašovací tabulka	http://solidbioscope.com	komerční
BWA	0.5.9-r16	Burrows-Wheeler	http://bio-bwa.sourceforge.net	GPL
CLC bio	4.0	nezjištěno	http://www.clcbio.com	komerční
PerM	0.3.3	hašovací tabulka, periodické seedy	http://code.google.com/p/perm	GPL
Novoalign	1.01.08	hašovací tabulka	http://www.novocraft.com	komerční, zdarma pro vědecké účely

Tab. 3: Přehled algoritmů vybraných pro porovnání

Čtené úseky byly namapovány na referenční sekvenci lidského genomu (GRCh37/hg19) pomocí všech výše uvedených nástrojů. Pro alogoritmy PerM, Bfast, BWA a Novoalign byly zvoleny shodné parametry: maximální počet povolených záměn v jednotlivých čtených úsecích oproti referenční sekvenci byl pět a maximální povolený počet míst na které může jeden čtený úsek mapovat byl také pět. Pro výpočet jsme používali počítač vybavený operačním systémem Gentoo se dvěma procesory *Intel® Xeon®* Processor *E5620* a 24 Gb RAM.

Pro identifikaci variant obsažených v genomu byly použity nástroje balíku SAMtools. Jednotlivé čtené úseky byly nejprve setříděny podle chromozomálních pozic (sort) a poté byly odstraněny všechny PCR duplikáty tak, že pouze ty s nejvyšší mapovací kvalitou zůstaly pro další zpracování (rmdup). Pouze čtené úseky s kvalitou mapování větší než 20 (QV > 20) byly použity pro další analýzu. Pro identifikaci inzercí a delecí byla zvolena kvalita vyšší než 50 (Indel QV > 50) a pro identifikaci substitucí kvalita vyšší než 100 (SNP QV > 100). Analýza pokrytí byly provedena pomocí nástrojů BedTools (funkce coverageBed). Pro určení zda-li jsou varianty známé či neznámé posloužil program Annovar kdy anotace probíhala oproti databázi dbSNP verze 131[11].

[11] http://www.ncbi.nlm.nih.gov/projects/SNP/

3.3.6. Bioinformatická analýza

Pro bioinformatickou analýzu dat získaných sekvenací exomu pacienta postiženého ANCL byla vybrána data získaná mapováním pomocí algoritmu Novoalign. Sekvenční varianty byly identifikovány pomocí softwarového balíku SAMtools a varianty, které dosáhly vysokého skóre (QV čteného úseku > 20, QV indel > 50, QV SNP > 100) Vysoce kvalitní varianty byly anotovány pomocí webového serveru SeattleSeq Annotation Server (GRCH37/hg19) který obsahoval varianty obsažené v databázi dbSNP verze 131. Následně byly odfiltrovány varianty, které se nalézaly mimo kódující oblasti genomu. Toho bylo dosaženo porovnáním získaných variant oproti databázi referenčních genů RefSeqGene[12]. Vzhledem k tomu, že podstata onemocnění není známa a incidence onemocnění je velmi malá, lze předpokládat, že ani příčinná mutace nebude uvedena v databázi dbSNP ani databázi 1000 genomů. Predikce vlivu nalezených mutací na strukturu a funkci proteinu byla provedena pomocí nástroje PolyPhen2. Vybrané varianty byly zobrazeny promocí prohlížeče IGV Viewer.

3.3.7. Ověření segregace mutace ve studované rodině přímým sekvenováním

Pro ověření segregace kandidátní mutace v genu *DNAJC5* ve studované rodině přímým sekvenováním byly navrženy primery pokrývající všechny kódující exony kandidátního genu. Pomocí programu Premier Primer byly navrženy primery a jejich specificita byla ověřena oproti referenční sekvenci lidského genomu (GRCh37/hg19) dostupného v prohlížeči UCSC Genome Browser pomocí algoritmu BLAT. Primery byly syntetizovány firmou Generi Biotech (Česká republika). Sekvence použitých primerů jsou uvedeny v tabulce 4.

Název primeru	Sekvence 5' → 3'	Pozice (GRCh37/hg19)	Délka produktu
gDNAJC5_ex2_U	GCCGTATTCTGCCGTCTCAC	20:62559507-62559526	406 bp
gDNAJC5_ex2_L	TCGGCCAGGATAAAGTATGT	20:62559894-62559913	
gDNAJC5_ex3_U	CAGCCCTGGAGAGTCGGACA	20:62560519-62560538	626 bp
gDNAJC5_ex3_L	GGGAACCCTGCAGGCGTGGA	20:62561126-62561145	
gDNAJC5_ex4_U	ATCCCCACCTGGAACGCACCC	20:62562093-62562113	402 bp
gDNAJC5_ex4_L	CCACAAACACTCGCGGCACA	20:62562476-62562495	
gDNAJC5_ex5_U	TTTGTCCAGGTGCCCGAAAG	20:62562723-62562742	655 bp
cDNAJC5_1205L	GAGGCCAAGACGGTAACACA	20:62563359-62563378	

Tab. 4: Sekvence primerů použitých pro ověření kandidátní mutace v genu DNAJC5

[12] http://www.ncbi.nlm.nih.gov/refseq/rsg

Podmínky PCR reakce byly optimalizovány pro celkový objem reakční směsi 25 µl. Složení reakční směsi je uvedeno v tabulce 5 a podmínky reakce pro amplifikaci jednotlivých exonů jsou uvedny v tabulce 6.

Složka	Koncentrace	Objem
dH2O		10,5 µl
1x Red Master Mix	1x	12,5 µl
Upper primer	0,2 pmol/µl	0,5 µl
Lower primer	0,2 pmol/µl	0,5 µl
DNA templát		1 µl
Reakční směs		25 µl

Tab. 5: Složení reakční směsi pro PCR

	Teplota	Čas	Opakování
Počáteční denaturace	94 °C	2min	1x
Denaturace	94 °C	10 s	30x
Hybridizace	T_a	15 s	
Elongace	72 °C	E_t	
Závěrečná elongace	72 °C	10 min	1x
	15 °C	***	

T_a exon 2 = 61 °C
T_a exon 3, 4, 5 = 65 °C
E_t exon 2, 4 = 25 s
E_t exon 3, 5 = 40 s

Tab. 6: Podmínky PCR reakce

Pro kontrolu délky a čistoty získaných PCR produktů bylo provedeno jejich obarvení (5 µl PCR směsi) pomocí 20x GelRed (3 µl) a následně byly elektroforeticky rozděleny v 1% agarózovém gelu (300V, 250mA, 5 minut). Jako marker pro zjištění velikosti PCR produktů byl použit GeneRuler 100 bp Plus.. Získané produkty byly vizualizovány pomocí UV transiluminátoru.

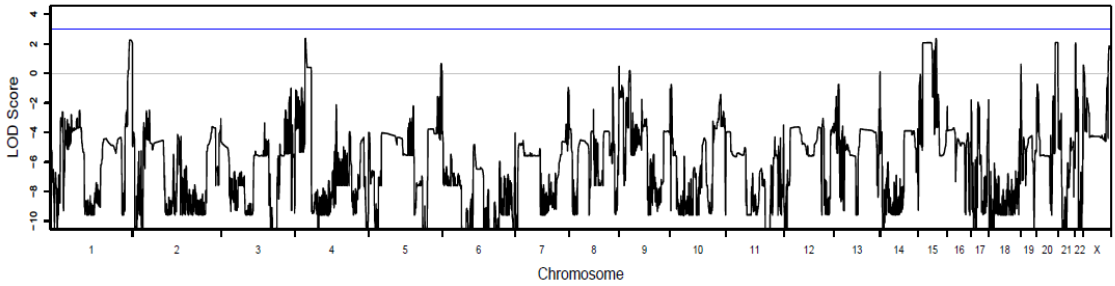
Produkty byly přečištěny pomocí izolačního kitu MSB® Spin PCRapace a koncentrace DNA byla stanovena spektrofotometricky na přístroji NanoDrop.

Získané produkty byly sekvenovány na přístroji ABI 3500XL Avant Genetic Analyzer pomocí sekvenačního kitu Dye Terminator. Pro sekvenační reakci byly použity stejné primery jako pro PCR reakci (Tab. 4). Jejich koncentrace byla 3,2 μM. Získané sekvence byly analyzovány pomocí programu Sequence Scanner.

4. VÝSLEDKY

4.1.1. Vazebná analýza

Výsledky vazebné analýzy jsou graficky zobrazeny na obrázku 17. Abychom s jistotou mohli říci, že jsou dané oblasti ve vazbě, mělo by být LOD skóre > 3 (pravděpodobnost vazby je 1000x pravděpodobnější, než pravděpodobnost nulové hypotézy). Pokud je testovaná rodina malá, je obtížné dosáhnout takto vysokého skóre. Z tohoto důvodu byly za vazebné oblasti pokládány všechny, které dosáhly LOD skóre většího než 2. Bylo nalezeno sedm vazebných oblastí na chromozomech 1, 4, 15, 20 a 22. Tyto oblasti jsou znázorněny na obrázku 18 a jejich chromozomální souřadnice (GRCH37/hg19) jsou uvedeny v tabulce 7.



Obr. 17: Výsledky vazebné analýzy znázorněné pomocí LOD skóre

chromozom	začátek	konec
1	233697529	end
4	23561661	28920119
15	39049915	61382423
15	65139935	67296086
15	71515415	78819152
20	53448624	konec
22	začátek	21449028

Tab. 7: Chromozomální souřadnice oblastí s LOD skóre > 2

Obr. 18: Vazebné oblasti na chromozomech 1, 4, 15, 20 a 22

4.1.2. Analýza počtu kopií genomové DNA

Abychom zjistili, zda-li není ANCL způsobena rozsáhlou inzercí nebo delecí, byla provedena analýza změn počtu kopií genomové DNA. Pro zkoumaný soubor pacientů nebyla nalezena žádná systematická změna v počtu kopií genomové DNA.

4.1.3. Expresní analýza

Bylo nalezeno 904 genů se změněnou expresí u pacientů oproti kontrolám. Abychom zjistili, jakých procesů se tyto geny v buňce účastní, provedli jsme funkční anotaci. V databázi KEGG PATHWAY bylo nalezeno 330 genů z našeho seznamu účastnících se šestnácti různých metabolických drah, z nichž čtyři jsou statisticky signifikantní i po korekci na mnohonásobná porovnání (Bonferonniho korekce, $p < 0,01$). Výsledky jsou uvedeny v tabulce 8.

KEGG PATHWAY	Po-čet genů	%	P-value	Genů v dráze	Fold Enri-chment	Bonferroni
hsa00190:Oxidative phosphorylation	29	3,21	6,77E-10	115	3,81	**1,12E-07**
hsa05012:Parkinson's disease	27	2,99	6,90E-09	111	3,67	**1,14E-06**
hsa05016:Huntington's disease	33	3,65	3,20E-08	169	2,95	**5,28E-06**
hsa05010:Alzheimer's disease	29	3,21	5,87E-07	154	2,84	**9,68E-05**
hsa03040:Spliceosome	21	2,32	1,04E-04	121	2,62	1,70E-02
hsa03010:Ribosome	14	1,55	3,10E-03	84	2,52	4,01E-01
hsa04621:NOD-like receptor signaling pathway	11	1,22	6,75E-03	62	2,68	6,73E-01
hsa04640:Hematopoietic cell lineage	13	1,44	1,02E-02	86	2,28	8,15E-01
hsa04620:Toll-like receptor signaling pathway	13	1,44	3,28E-02	101	1,94	9,96E-01
hsa04260:Cardiac muscle contraction	10	1,11	4,57E-02	72	2,1	1,00E+00
hsa04062:Chemokine signaling pathway	19	2,1	5,84E-02	184	1,56	1,00E+00
hsa05130:Pathogenic Escherichia coli infection	8	0,88	6,26E-02	54	2,24	1,00E+00
hsa05340:Primary immunodeficiency	6	0,66	7,73E-02	35	2,59	1,00E+00
hsa05020:Prion diseases	6	0,66	7,73E-02	35	2,59	1,00E+00
hsa04060:Cytokine-cytokine receptor interaction	24	2,65	8,63E-02	259	1,4	1,00E+00
hsa04120:Ubiquitin mediated proteolysis	14	1,55	9,40E-02	132	1,6	1,00E+00

Tab. 8: Metabolické dráhy kterých se účastní rozdílně exprimované geny. Statisticky významné dráhy jsou označeny tučně.

4.1.4. Porovnání mapovacích algoritmů

4.1.4.1. Porovnání doby mapování

Porovnávali jsme čas, který byl potřebný pro mapovaní pomocí jednotlivých algoritmů. Všechny použité algoritmy umožňují výpočetní úlohu rozdělit do několika podúloh, z nich každá je zpracovávána na jiném procesoru (tzv. multithreading). Dobu potřebnou k výpočtu uvádíme jako dobu, kterou daný proces vytěžoval procesor (CPU čas). Reálný čas získáme vydělením CPU času počtem využívaných výpočetních jader (reálný čas). Výsledky porovnání doby mapování jsou uvedeny v tabulce 9.

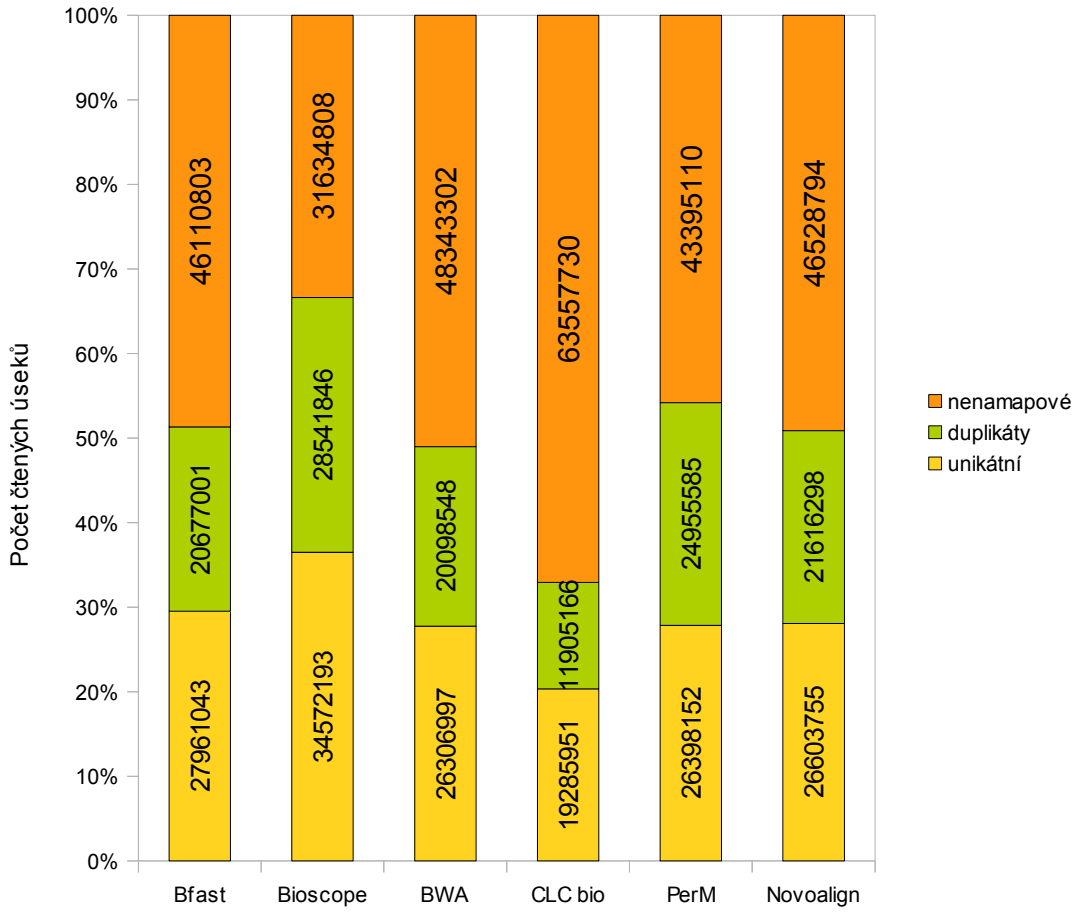
Výpočetní náročnost programu Bioscope neznáme z důvodu zpracování dat firmou CeGaT a náročnost programu CLC bio neuvádíme, protože analýza byla provedena s využitím odlišného serveru na platformě Windows.

	Bfast	Bioscope	BWA	CLC bio	PerM	Novoalign
CPU čas (s)	1074894	N/A	124551	N/A	20876	3302820
Reálný čas (s)	67181	N/A	7784	N/A	1305	206426

Tab. 9: Doba potřebná pro mapování čtených úseků na referenční genom

4.1.4.2. Porovnání počtu čtených úseků

Zajímalo nás, jaké množství čtených úseků dokáží namapovat jednotlivé algoritmy. Celkový počet čtených úseků byl 94748847. Největší počet čtených úseků, tedy 66,61 % z celkového množství, bylo namapován algoritmem Bioscope. Následující algoritmy namapovaly 54,20 % (PerM), 51,33 % (Bfast), 50,89 % (Novoalign) a 32,92% (CLC bio) z celkového množství čtených úseků. Při PCR reakci vzniká mnoho duplikátů; úseků které mají naprosto stejnou sekvenci. Abychom získali pouze unikátní čtené úseky, použili jsme filtr na odstranění duplikátů (SAMtools, rmdup). Po aplikaci tohoto filtru nám zůstalo namapovaných čtených úseků algoritmem Bioscope (36,49 %), Bfast (29,51 %), Novoalign (28,08 %), PerM (27,86 %), BWA (27,76 %), CLC bio (20,35 %). Výsledky jsou znázorněny na obrázku 19.



Obr. 19: Přehled mapování čtených úseků. Graf ukazuje počty čtených úseků nenamapovaných, unikátních a PCR duplikátů z celkového počtu čtených úseků

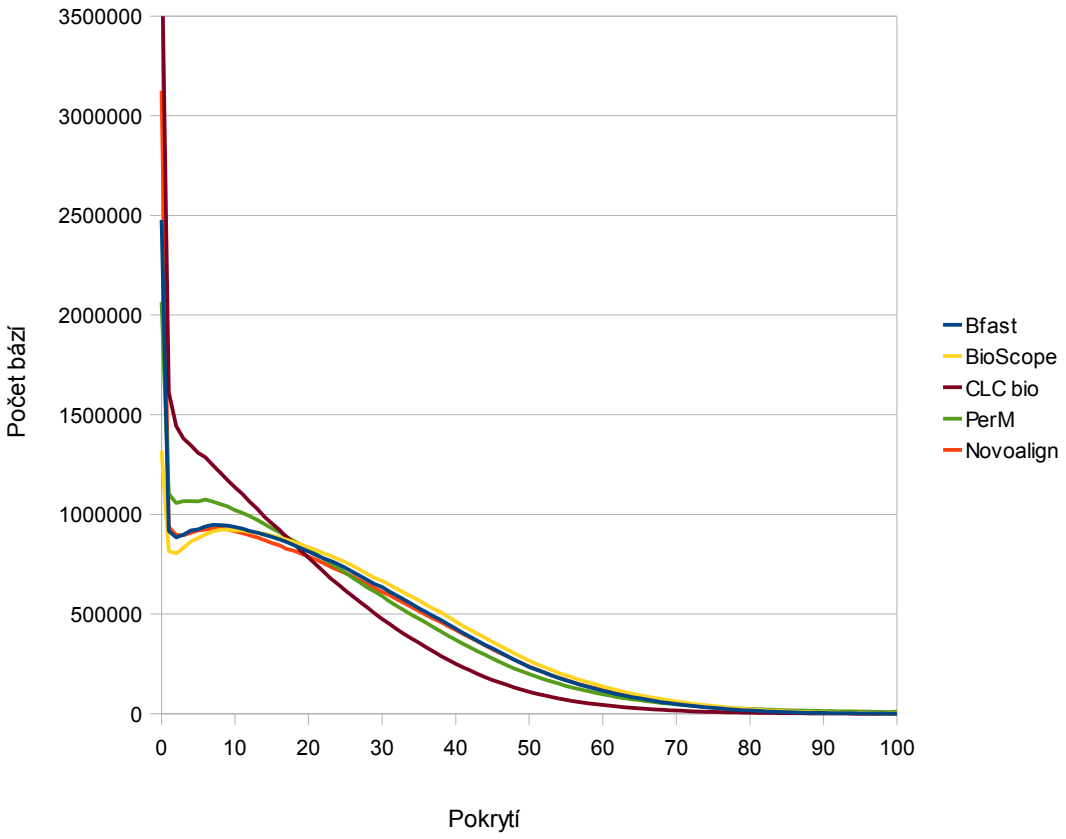
4.1.4.3. Analýza pokrytí

Abychom si udělali představu o tom, jak jsou čtené úseky rozloženy na referenční sekvenci, provedli jsme analýzu pokrytí. Nejprve nás zajímalo, zda-li jsou čtené úseky namapovány na správných souřadnicích; tedy jestli mapují na ty pozice, které byly obohaceny pomocí prób kitu SureSelect Human All Exome Kit (Agilent). Cílová oblast daná návrhem obohacovacího kitu měla dohromady 38815064 bází. Tabulka 10 uvádí přehled pokrytí kterého dosáhly jednotlivé algoritmy. Ve druhém řádku tabulky je

uveden celkový počet bází, které byly namapovány do cílových oblasti vymezeného návrhem obohacovacího kitu. Následují informace o průměrném pokrytí a o mediánu pokrytí a počtu bází, které byly pokryty minimálně jednou, pětkrát a desetkrát. Graf na obrázku 20 znázorňuje počty bází, které dosáhly určitého pokrytí.

	Bfast	Bioscope	BWA	CLC bio	PerM	Novoalign
Báze v cílové oblasti	860940724	932966918	797894910	622176130	826343510	851369753
Průměrné pokrytí	22,18	24,04	20,56	16,03	21,29	21,93
Medián pokrytí	19	21	17	13	18	19
Báze pokryté min 1x	36337575	37495856	36901855	34791309	36749017	35688209
Báze pokryté min 5x	32719666	34179970	32549939	29009014	32455742	32053137
Báze pokryté min 10x	28019653	29636163	27087629	22788231	27161210	27428910

Tab. 10: Analýza pokrytí dosaženého jednotlivými algoritmy



Obr. 20: Počet bází o určitém pokrytí

4.1.4.4. Efektivita obohacení

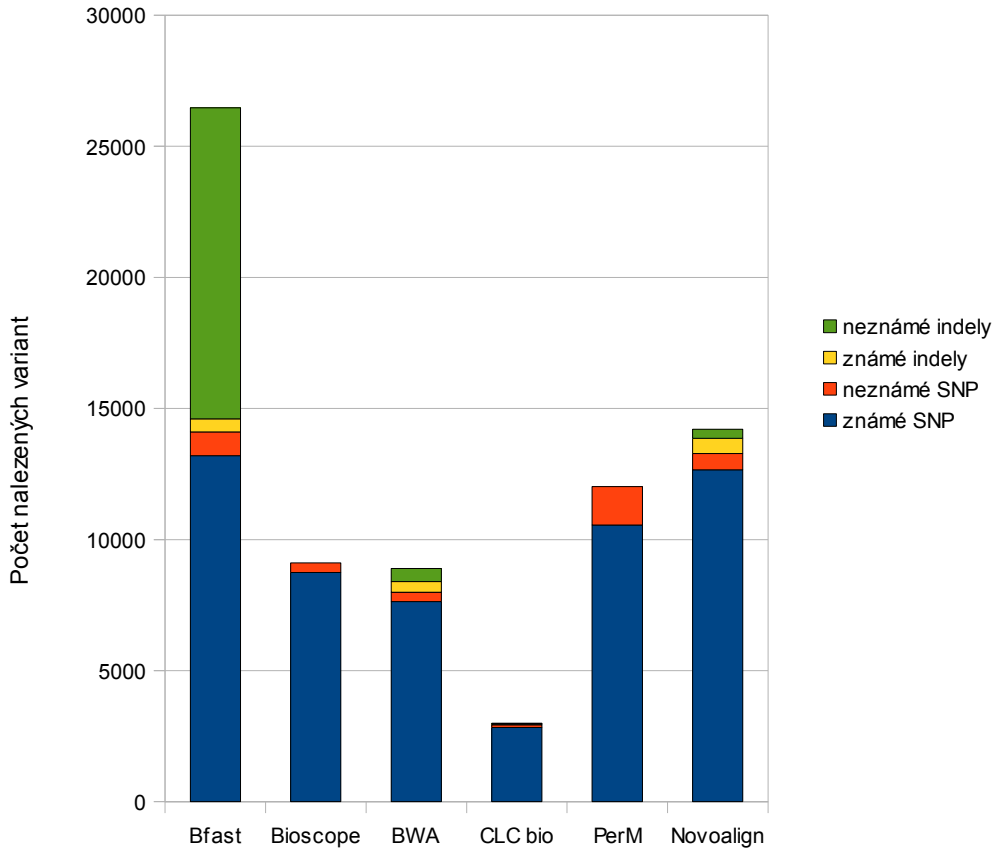
Efektivita obohacení studovaného vzorku o kódující sekvence je počítána jako poměr počtu unikátních čtených úseků lokalizovaných v oblasti dané návrhem obohacovacího čipu ku celkovému počtu namapovaných čtených úseků. Efektivitu obohacení zobrazuje tabulka 11.

	Bfast	Bioscope	BWA	CLC bio	PerM	Novoalign
Efektivita obohacení	61,58%	53,97%	60,66%	64,52%	62,61%	64,00%

Tab. 11: Efektivita obohacení studovaného vzorku

4.1.4.5. Počet variant nalezených jednotlivými algoritmy

Následující analýza měla za cíl odhalit SNP a malé inzerce a delece. Tedy místa, kde se čtené úseky liší od referenční sekvence. Obrázek 21 znázorňuje počty nalezených variant, známých i neznámých SNP, inzrecí i delecí.



Obr. 21: Počet známých a neznámých variant nalezených jednotlivými algoritmy

4.1.4.6. Hodnocení kvality nalezených variant

Předpokládaný počet nalezených variant:

Předpokládaný počet variant (polymofních míst) byl vypočítán podle vzorce na obrázku 22. L značí cílovou oblast, θ frekvenci heterozygotů v dané populaci a N počet analyzovaných vzorků (Depristo, 2010).

$$\text{Number of polymorphic sites} \approx L \cdot \theta \sum_{i=1}^{2N} 1/i$$

Obr. 22: Vzorec pro výpočet předpokládaného počtu nalezených variant

Pro náš případ, kdy máme jeden vzorek, tedy 2N= 2, počet bází v cílové oblasti L = 38815064 a frekvence heterozygotů v evropské popululaci pro exom odpovídá $\theta = 0.42 \times 10^{-3}$ je očekávané množství polymorfních míst 24453. Tabulka 12 uvádí celkový počet polymorfních míst (SNP i insercí a delecí).

	Bfast	Bioscope	BWA	CLC bio	PerM	Novoalign
Polymorfních míst v cílové oblasti	26472	9111	8901	2990	12024	14204

Tab. 12: Počet polymorfních míst nalezených jednotlivými algoritmy

Poměr známých a neznámých variant:

Většina variant je známých, obsažených v databázích genetických variant, jako je například dbSNP. Pro jeden vzorek by mělo být ~90% variant známých, obsažených v dbSNP. Výsledky shrnuje tabulka 13.

SNPs	Bfast	Bioscope	BWA	CLC bio	PerM	Novoalign
známých SNP	93,58%	95,90%	95,60%	96,52%	87,80%	95,35%
známých variant	51,77%	95,90%	90,36%	96,22%	87,80%	93,26%

Tab. 13: Počet známých SNP a známých variant zahrnujících i inserce a delece

4.1.4.7. Korelace zjištěných SNP s genotypy zjištěnými pomocí genotypovacího čipu

Abychom si udělali představu, jak jsou mapovací algoritmy přesné, analyzovali jsme, jaký počet heterozygotních SNP zjištěných genotypováním byl nalezen i pomocí sekvenace (Tab. 14). Celkový počet heterozygotních SNP zjištěných pomocí genotypovacího čipu byl 1743.

	Bfast	Bioscope	BWA	CLC bio	PerM	Novoalign
shodně nalezených SNP	1444	1072	922	383	1042	1394
shodně nalezených SNP	82,85%	61,50%	52,90%	21,97%	59,78%	79,98%

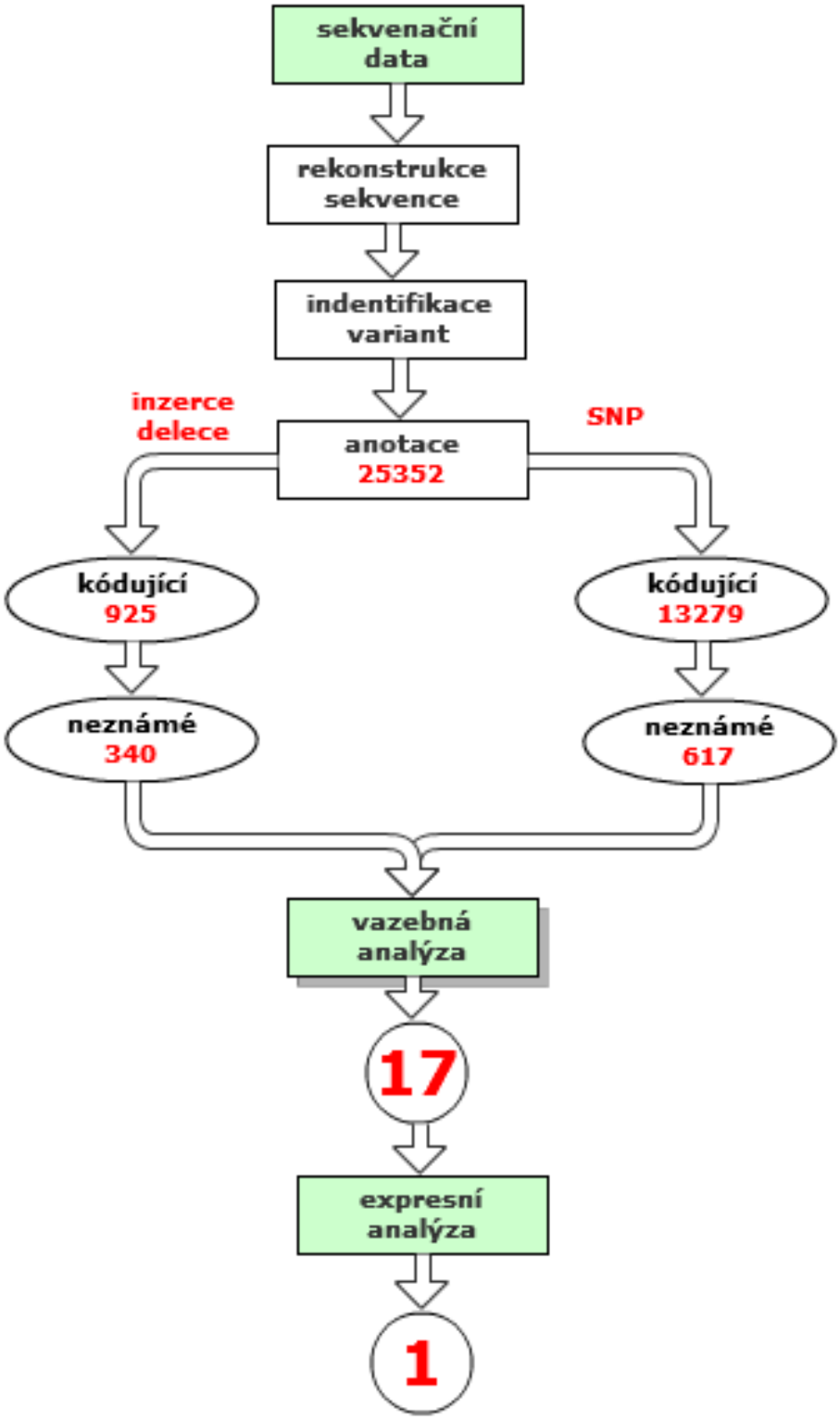
Tab. 14: Počet heterozygotních SNP, které byly shodně nalezeny také pomocí jednotlivých algoritmů

4.1.5. Bioinformatická analýza

Pro hledání mutace způsobující ANCL jsme zvolili algoritmus Novoalign. Identifikovali jsme 25352 vysoce kvalitních sekvenčních variant, které jsme funčně anotovali. Pomocí postupné filtrace, jejíž detaily jsou uvedeny v kapitole 3.3.6, jsme získali nejprve seznam variant kódujících a poté variant neznámých, které nejsou uvedeny v databázi dbSNP ani databázi 1000 Genomů Získali jsme seznam unikátních variant čítající 617 SNP a 340 inzercí a delecí. Dále jsme nalezené varianty propojili s výsledky vazebné analýzy, čím se seznam nalezených mutací snížil na 17. Pouze jeden z genů, obsahující zjištěné mutace, měl ve tkáni pacientů signifikantně změněnou expresi. Uvedený postup je znázorněn na obrázku 23. Nalezené změny shrnuje tabulka 15.

Chromozom	Pozice	Ref. báze	Genotyp	Alely	Poly-Phen2	Gen	Funkce	AMK	cons Score GERP	Změněná exprese
1	235715487	C	Y	C/T	possibly damaging	GNG4	missense	ARG,GLN	2,02	N/A
1	236987511	C	Y	C/T	benign	MTR	synonymous	none	-1,75	0,05
1	247835884	G	S	C/G	benign	OR13G1	synonymous	none	-4,26	N/A
4	25678161	TGC	D3	-TGC	N/A	SLC34A2	coding	none	4,25	N/A
15	41347434	C	Y	C/T	benign	INO80	intron	none	-8,1	N/A
15	43552699	G	K	G/T	benign	TGM5	missense	HIS,ASN	3,91	N/A
15	43900152	C	Y	C/T	benign	STRC	synonymous	none	-3,94	N/A
15	45028846	G	K	G/T	benign	TRIM69	utr-5	none	0,45	N/A
15	59500165	A	R	A/G	benign	MYO1E	missense	ILE,VAL	-0,27	0,94
15	65555517	A	R	A/G	benign	PARP16	synonymous	none	2,01	0,51
15	66857720	C	Y	C/T	benign	LCTL	utr-5	none	2,51	N/A
15	75116808	G	R	A/G	benign	LMAN1L	missense	VAL,MET	-3,55	N/A
20	60884826	G	R	A/G	benign	LAMA5	synonymous	none	-3,36	0,24
20	62562227	CTC	D3	-CTC	N/A	DNAJC5	coding	none	5,34	0,03
22	20097642	C	Y	C/T	benign	DGCR8	utr-3	none	-4,4	0,97
22	20106680	G	R	A/G	benign	RANBP1	intron	none	-6,55	0,07
22	21138486	C	Y	C/T	benign	D1	synonymous	none	4,4	N/A

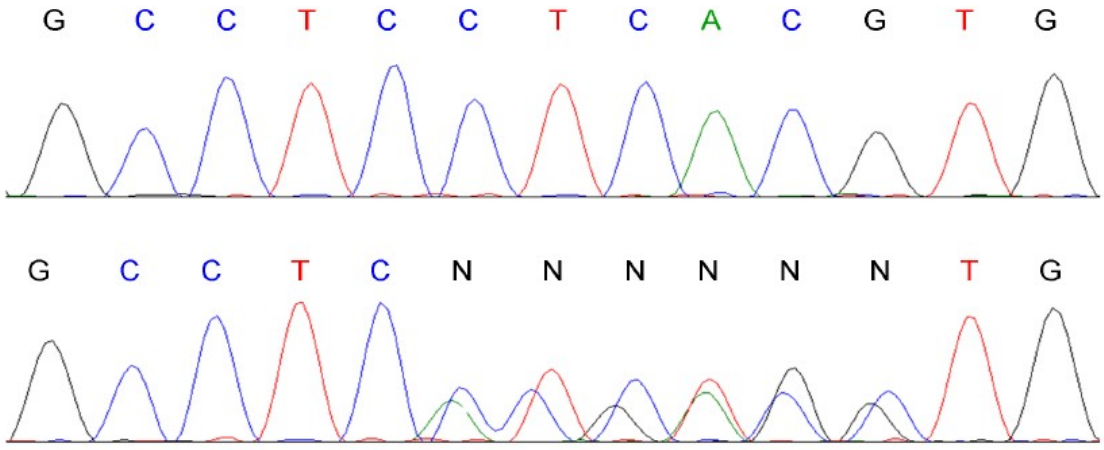
Tab. 15: Seznam neznámých mutací nacházejících se v oblasti vymezené vazebnou analýzou. Hodnota N/A značí, že gen nebyl ve tkáni pacientů exprimován



Obr. 23: Schéma postupu hledání mutace způsobující ANCL. Červeně jsou uvedeny počty nalezených variant v jednotlivých filtračních krocích

4.1.6. Ověření segregace mutace ve studované rodině přímým sekvenováním

Abychom ověřili segregaci kandidátní mutace v genu *DNAJC5* ve studované rodině, osekvenovali jsme kódující exony daného genu u všech členů rodiny od kterých jsme měli k dispozici vzorky DNA. U všech postižených pacientů byla nalezena delece trinukleotiu CTC na pozici 116 která vede k deleci leucinu, kdežto zdraví jedinci tuto mutaci nemají (Obr. 24). Všichni postižení jedinci jsou pro uvedenou mutaci heterozygotní. Tímto byla ověřena segregace mutace se studovaným fenotypem.



Obr. 24: Ověření mutace přímým sekvenováním. V horní části obrázku je uvedena sekvence genu DNAJC5 u zdravých jedinců. V dolní části obrázku je sekvence postižených pacientů s heterozygotní mutací Leu116del v genu DNAJC5

5. DISKUZE

Nejdůležitějším krokem podmiňujícím úspěšnost sekvenačního experimentu je počáteční krok mapování čtených úseků na referenční sekvenci (Flicek and Birney, 2009). Díky rychlému vývoji na poli nových sekvenačních technik jsou neustále vyvíjeny algoritmy nové, které mají za cíl zvýšení rychlosti mapování při zachování co největší přesnosti (Flicek, 2009).

V rámci této diplomové práce byly hodnoceny a porovnávány mapovací algoritmy použité na reálných datech získaných sekvenací exomu pacienta trpícího ANCL. Sekvenování bylo provedeno na přístroji SOLiD™4 System. Porovnávali jsme dobu mapování, počty namapovaných čtených úseků a přesnost mapování. Pro následnou analýzu dat pacienta s ANCL jsme vybrali algoritmus Novoalign, který nejlépe vyhovoval našim požadavkům. Pro zjištění kauzálního genu jsme použili kombinovaného přístupu založeného na filtraci dat a na propojení získaných výsledků s daty vazebné, expresní a funkční analýzy (Ng et al., 2010b). Tímto způsobem jsme objevili kandidátní mutaci v genu *DNAJC5*, jejíž segregaci se studovaným fenotypem ANCL jsme ověřili přímým sekvenováním. V současné době probíhají funkční studie s cílem osvětlit molekulárně biologickou příčinu vzniku onemocnění.

5.1.1. Porovnání mapovacích algoritmů

5.1.1.1. Výběr algoritmů

Pro porovnání byly vybrány algoritmy umožňující alignment čtených úseků produkovaných sekvenátorem firmy SOLiD, tedy algoritmy umožňující mapování v barevném prostoru. Algoritmy jsme vybírali jak podle informací v literatuře, tak podle doporučení uživatelů sekvenačních fór SeqAnswers[13] a BioStar[14]. Algoritmy jsme vybrali tak, aby zde byly zastoupeny oba dva nejčastěji využívané mapovací přístupy (Burrows-Wheelerova transformace a hašovací tabulka).

[13] http://seqanswers.com
[14] http://biostar.stackexchange.com

5.1.1.2. Výpočetní náročnost

První parametr, který jsme porovnávali, byla rychlost mapování čtených úseků na referenční sekvenci. Čas potřebný k výpočtu se lišil mezi nejrychlejším algoritmem (PerM) a nejpomalejším (Novoalign) o ~38 dní v případě, že uvažujeme CPU čas a o ~57 hodin v případě, že uvažujeme čas reálný.

Pokud máme k dispozici výkonný počítač s více procesory umožňující tzv. multithreading, není tento časový rozdíl tak dramatický; pokud ale tento počítač nemáme, je tento rozdíl značný. V případě, že bychom měli k dispozici pouze méně výkonný počítač, bylo by užitečné porovnávat nároky jednotlivých mapovacích algoritmů na operační paměť počítače. V tomto případě by bylo do výběru vhodné zařadit i další mapovací algoritmy, jako je například Bowtie (Langmead et al., 2009), který dle systémové specifikace potřebuje pro uchovávání indexu lidského genomu pouze 2 GB data a pro mapování čtených úseků na referenční sekvenci mu stačí 1,3 GB operační paměti, takže je možné mapování provádět i na stolním počítači. Bowtie je možné využívat i přes webové rozhraní poskytované projektem Galaxy (Goecks et al., 2010; Blankenberg et al., 2010). Jeho nevýhodou je, že neumožňuje vyhledávat inzerce a delece.

Porovnávaný čas zahrnuje všechny kroky potřebné pro mapování čtených úseků, od indexování, přes převody formátů a samotné mapování až po vyhledávání variant. Nutno podotknout, že v případě, že pracujeme stále se stejnou referenční sekvencí, není nutné index referenční sekvence vytvářet stále znovu, ale je vhodné využít indexu vytvořeného. Z tohoto důvodu by bylo vhodné také porovnat pouze dobu potřebnou pro samotné mapování.

Algoritmus Bfast a Novoalign byly nejpomalejší, kdežto mezi ty rychlejší se řadí algoritmus BWA a PerM. Algoritmus Bfast, Novoalgin a PerM jsou založeny na principu hašovacích tabulek, kdežto BWA na Burrows-Wheelerově transformaci. Rozdíl mezi porovnávanými algoritmy využívající hašovací tabulku je ten, že PerM na rozdíl od Novoalignu a Bfastu využívá jiný typ seedů. PerM využívá periodických seedů s mezerami a s velkou váhou. Velká váha je potřebná obzvlášť pro genomy s mnoha repetetivními úseky. Tento přístup umožňuje rychlý výpočet při zachování velké senzitivity (Chen et al., 2009).

Algoritmy založené na BWT transformaci jsou obecně považovány při zachování stejné senzitivity jako rychlejší než algoritmy založené na hašovacích tabulkách (Flicek and Birney, 2009). Naše porovnání algoritmů nicméně prokázalo, že PerM je rychlejší než BWA. Pro zobecnění tohoto pozorování na tvrzení, že algoritmy využívající hašovací tabulky s periodickými seedy jsou rychlejší než algoritmy založené na BWT by bylo potřeba provést porovnání více různých algoritmů.

5.1.1.3. Porovnání počtu namapovaných čtených úseků

Abychom mohli výsledky získané jednotlivými algoritmy navzájem porovnávat, bylo třeba nastavit stejné parametry. Vzhledem k tomu, že každý z algoritmu pracuje na mírně odlišném principu, nebylo možné sjednotit veškeré parametry. U programů Bfast, BWA, PerM a Novoalign jsme nastavili shodně tyto parametry: maximální počet povolených záměn v jednom čteném úseku byl pět a maximální počet pozic, na které může jeden čtený úsek mapovat byl také pět. Hodnoty ostatních parametrů byly výchozí. Algoritmy CLCbio a Bioscope byly spuštěny s rozdílným nastavením a z tohoto důvodu jsou všechna data uváděna jen pro hrubé porovnání.

Algoritmy Novoalign, PerM, BWA a Bfast se v počtu unikátních namapovaných úseků příliš nelišily. Namapovaly 27,86–29,51 % čtených úseků z celkového počtu 94748847. Algoritmus Bioscope namapoval 36,49 % čtených úseků, kdežto algoritmus CLCbio pouze 20,35 %. Tyto rozdíly mohou být dány rozdílným nastavením parametrů mapování. S velkou pravděpodobností se tedy algoritmy téměř neliší v počtu čtených úseků které jsou schopny namapovat, nicméně tuto hypotézu by bylo třeba ověřit pomocí jiného nastavení parametrů.

5.1.1.4. Analýza pokrytí

Analýza pokrytí ukazuje, že počet bází s daným pokrytím je obdobný u všech algoritmů, pouze algoritmus CLCbio mírně zaostává za ostatními; má více bází s menším pokrytím a méně bází s vyšším pokrytím než ostatní algoritmy. Průměrné pokrytí bází je u algoritmů Bfast, BWA, PerM a Novoalign v úzkém rozpětí od 20,59 do 22,18. Průměrné pokrytí bází namapovaných algoritmem Bioscope je 24,04 a programem CLCbio pouze 16,03. Tento rozdíl může být opět dán rozdílným nastavením programů Bioscope a CLCbio.

Následně nás zajímalo, kolik bází je pokryto minimálně 10x, protože osmi až desetinásobné pokrytí už je dostatečné i pro detekci heterozygotních SNP (Ng et al., 2009; Choi et al., 2009). Z výše uvedeného je patrné, že z hlediska pokrytí rozdíly mezi algoritmy nejsou nijak výrazné. Pouze CLC bio mírně zaostává za ostatními algoritmy.

5.1.1.5. Efektivita obohacení

Efektivita obohacení genomu o kódující sekvence byla největší u algoritmu Bfast, následovaly algoritmy CLCbio (64,52 %), Novoalign (64,00 %), PerM (62,61 %), BWA (60,66 %) a nejhoršího obohacení bylo dosaženo pomocí algoritmu Bioscope, a to pouze 53,97 %. Z předchozích výsledků víme, že namapoval největší počet unikátních úseků, nicméně se jich mnoho nacházelo mimo kódující oblasti.

5.1.1.6. Počet variant nalezených jednotlivými algoritmy

Porovnávali jsme počet nalezených variant, inzercí a delcí v cílové oblasti vymezené návrhem obohacovacího kitu. Následně nás zajímalo, kolik z těchto variant je známých, obsažených v dbSNP a kolik je neznámých.

Zde byly rozdíly značné. Algoritmus Bfast nalezl variant nejvíce (26472), nicméně se převážně se jednalo o neznámé inserce a delece. Následoval algoritmus Novoalign (14204), PerM (12024), Bioscope (9111), BWA (8901) a CLCbio (2990).

Pokud porovnáme počty nalezených variant s počtem namapovaných čtených úseků, došlo zde k překvapivé změně pořadí mezi jednotlivými algoritmy. Nejméně variant nalezl opět algoritmus CLCbio, který namapoval i nejméně čtených úseků. Překvapivý byl propad algoritmu Bioscope, který namapoval čtených úseků nejvíce, nicméně variant nalezl méně, než algoritmy s menším počtem namapovaných čtených úseků (PerM, Novoalign). V závislosti na počtu nalezených variant a efektivitě obohacení se domníváme, že algoritmus Bioscope mapuje s menší přesností než algoritmy ostatní, kdežto Novoalign naopak používá pro mapování přísnější kritéria.

5.1.1.7. Hodnocení nalezených variant

Falešně pozitivní a falešně negativní výsledky jsou kritickým bodem všech resekvenačních experimentů (Ng et al., 2009). Neznámé varianty obsahují jak nové, reálné varianty, tak varianty falešně pozitivní, způsobené chybami vzniklými při

sekvenování i v průběhu analýzy dat (Schwartz et al., 2011).

Kvalitu exomových dat jsme ověřili několika způsoby: pomocí predikce počtu očekávaných variant, pomocí poměru známých a neznámých SNP a porovnáním s genotypy získanými pomocí genotypovacího čipu Affymetrix SNP 6.0.

Předpokládaný počet nalezených variant

Předpokládaný počet nalezených variant závisí na frekvenci hetorozygotů v populaci, velikosti cílové oblasti a počtu analyzovaných vzorků. Pro náš vzorek byl předpokládaný počet nalezených variant 24453. Algoritmus Bfast nalezl 26472, což je více než počet očekávaný, kdežto ostatní algoritmy nalezly variant méně.

Poměr známých a neznámých variant

Většina variant je již známých, obsažených v databázi dbSNP. Předpokládaný počet známých variant je ~90 % ze všech nalezených (Depristo, 2010; Kiialainen et al., 2011).

Pokud porovnáváme pouze počet nalezených známých SNP, tak PerM jich nalezl nejméně (87,80 %), následoval Bfast (93,58 %), Novoalign (95,35 %), BWA (95,60 %), Bioscope (95,90 %) a CLCbio (96,53 %).

Pokud ovšem porovnáváme počty všech nalezených variant, tedy jak SNP, tak inzercí a delecí, tak Bfast nalezl nejmenší počet známých variant a to pouze 51,77 %. Algoritmus Novoalign i Bfast nalezli o něco méně známých variant než v případě SNP a to 93,26 % v případě Novoalignu a 90,36 % v případě BWA. U algoritmů PerM a Bioscope zůstal tento počet stejný, neboť rozdíl je dán poměrem známých a neznámých insercí a delecí, které neumí tyto algoritmy vyhledávat.

Překvapivé je nízké procento známých variant nalezených algoritmem Bfast. Poměr známých a neznámých SNP odpovídá výsledkům očekávaným. Z toho plyne, že nízký počet celkových variant je dán vysokým počtem neznámých nisercí a delecí. Z tohoto důvodu se domníváme, že velký počet neznámých inzercí a delecí je falešně pozitivních

Korelace SNP s genotypy zjištěnými pomocí genotypovacího čipu

Zajímalo nás, jaké množství variant bylo shodně nalezených i pomocí genotypovacího čipu.

Celkový počet heterozygotních SNP v kodujících oblastech určených pomocí genotypovacího čipu Affymetrix SNP 6.0 čipu byl 1743. Nejvíce shodně nalezených SNP měl algoritmus Bfast (82,85 %), následoval Novoalign (79,98 %), Bioscope (61,50 %), PerM (59,78 %), BWA (52,90 %) a CLCbio (21,97 %). Zjištěný počet shodně nalezených SNP odpovídá průměrnému sekvenčnímu pokrytí.

Další možnosti ověření kvality exomových dat

Další možností, jak ověřit kvalitu exomových dat by bylo porovnat poměr tranzic a transverzí. Tranzice jsou dvakrát frekventovanější než transverze (Ebersberger et al., 2002). Poměr tranzic a transverzí (Ti/Tv) by měl být pro celogenomové sekvenování ~2,0 a pro sekvenování exomu ~2,8. Falešně pozitivní SNP by měly mít poměr Ti/Tv rovný přibližně 0,5 (Depristo, 2010). Pro analýzu poměru Ti/Tv a rozložení variant na plusovém a mínusovém vlákně lze využít programu GATK, který tyto analýzy umožňuje (McKenna et al., 2010; Depristo et al., 2011).

Bylo by také vhodné ověřit kvalitu nalezených dat tak, že bychom porovnali shodu genotypů nalezených pomocí genotypovacího čipu a pomocí sekvenování. Očekávaná shoda nalezených genotypů by měla být >99,5 % (Depristo, 2010).

Kvalitu mapovacích algoritmů by bylo také možné ověřit na uměle vytvořených sekvenačních datech (Li et al., 2008). Vzhledem k tomu, že bychom věděli, ze kterého úseku referenční sekvence čtené úseky pocházejí, mohli bychom přesně určit, zda-li jsou nalezené varianty reálné a určit počet falešně pozitivních a falešně negativních. Bylo by však potřeba zajistit, aby porovnávaná data co nejlépe odrážela charakter reálných dat.

5.1.1.8. Shrnutí výsledků

Bfast

Algoritmus Bfast je vhodný k vyhledávání SNP, protože jich nalezl největší počet a zároveň nalezl největší počet shodných SNP s genotypovacím čipem. Algoritmus Bfast nicméně na základě našich výsledků není vhodný pro vyhledávání inzercí a delecí z důvodu vysoké falešné pozitivity. Algoritmus Bfast umožňuje nastavení velkého množství parametrů. To poskytuje možnost dobře uzpůsobit podmínky mapování pro různý charakter dat. Pro začínající uživatele může být nicméně obtížná orientace

v tomto množství nastavitelných parametrů.

Novoalign

Algoritmu Novoalign se co do počtu nalezených variant umístil na druhém místě, hned za algoritmem Bfast. Poměr známých a neznámých variant odpovídal očekávanému poměru. Přesnost mapování je na úkor doby výpočtu, která je několikanásobně delší než u algoritmů ostatních. Nevýhoda algoritmu Novoalign je ta, že alignment v barevném prostoru je k dispozici pouze v komerční verzi. Pro využití algoritmu Novoalign pro mapování v barevném prostoru je nutné zvolit komerční verzi tohoto programu. Stejně tak pouze komerční verze podporuje multithreading, což je funkce téměř nezbytná v případě mapování velkého objemu dat, z důvodu časové náročnosti výpočtu.

PerM

PerM je několikanásobně rychlejší, než ostatní algoritmy a počet nalezených variant je jen o něco nižší než u algoritmů Bfast a Novoalign. Poměr známých a neznámých variant odpovídá očekávanému poměru, z čehož můžeme usuzovat, že množství falešně pozitivních variant není vysoké. Algoritmus PerM tedy nabízí nejlepší poměr rychlosti a přesnosti. Jeho nevýhoda je ovšem ta, že neumí nalézt inzerce a delece.

BWA

Algoritmus BWA je druhým nejrychlejším algoritmem, nicméně variant nalezl méně než algoritmy ostatní. Poměr známých variant odpovídá očekávanému poměru. Algoritmus může být vhodný v případě, že potřebujeme rychle vyhledat inzerce a delece.

Bioscope a CLCbio

Vzhledem k tomu, že u algoritmu Bioscope a CLCbio nebyly nastaveny shodné parametry jako u algoritmů ostatních, není možné ze zjištěných výsledků vyvodit relevantní závěry. Výhodou programu CLCbio je grafické prostředí, které umožňuje zpracování dat i příležitostným uživatelům. Pro tento typ lidí je vhodný také algoritmus Bioscope, který je možné ovládat jako pomocí příkazové řádky, tak pomocí grafického rozhraní. Program Bioscope nalezl s danými parametry také větší množství variant než

CLCbio

5.1.1.9. Závěr

Volba vhodného algoritmu závisí na mnoha aspektech, jako jsou použitá data, výkonnost počítače, uživatelské schopnosti a cíl experimentu (Picardi, 2009). Z tohoto důvodu nelze jednoznačně říct, který algoritmus je nejlepší. Možným přístupem je také vyzkoušet kombinaci více algoritmů[15]. V prvním kroku použít algoritmus, který mapuje rychle, ale méně přesně a ve druhém kroku pro čtené úseky, které nebyly namapované použít algoritmus přesnější, ale pomalejší. Pro tento účel se hodí PerM, který má nejlepší poměr počtu nalezených variant a doby potřebné pro alignment a zároveň umožňuje uložení nenamapovaných čtených úseků do zvláštního souboru, který je poté možné využít pro mapování pomocí přesnějšího algoritmu.

5.1.2. Bioinformatická analýza

Dalším cílem bylo provést analýzu dat získaných sekvenací exomu pacienta s ANCL s cílem odhalit kauzální mutaci tohoto onemocnění.

Na základě výsledků porovnání mapovacích algoritmů byla tato data analyzována programem Novoalign.

Zjištěné varianty jsme funkčně anotovali pomocí dat obsažených v databázi dbSNP. Pomocí nástroje PolyPhen2 jsme provedli analýzu vlivu zjištěných mutací na strukturu a funkci proteinu. Pro určení kandidátních mutací jsme získané varianty postupně filtrovali, což je postup který byl použit v řadě experimentů které vedly k odhalení mutací způsobující vzácná onemocnění (Biesecker, 2010).

V první fázi jsme odfiltrovali všechny varianty, které se nacházely mimo oblasti vymezené návrhem čipu, tedy varianty v nekódujících oblastech. Dalším krokem byla filtrace variant, které jsou neznámé, které nejsou evidované v databázi dbSNP ani databázi 1000 Genomů. Získaný seznam byl příliš dlouhý na to, abychom z něj určili kauzální mutaci. Pokud bychom osekvenovali více pacientů, mohli bychom kauzální mutaci určit hledáním mutace, která by byla společná pro všechny postižené jedince.

Získaná data jsme propojili s daty získanými vazebnou analýzou a nalezli jsme celkem 17 unikátních mutací. Dalším předpokladem bylo, že exprese genu je

[15] http://biostar.stackexchange.com

u postižených pacientů signifikantně změněna, ať již zvýšena nebo snížena. Toto kritérium splňovala pouze jedna mutace, a to mutace v genu *DNAJC5*, který kóduje protein CSPα.

CSPα patří mezi rodinu J proteinů. Obdobně jako ostatní J proteiny, CSP působí v komplexu s dalšími proteiny jako molekulární chaperon. Chaperonový komplex sestává z CSPα, Hsc70 a SGT (small glutamine-rich tetratricopeptide repeat domain protein) (Tobaben et al., 2001). CSP aktivuje ATPázu proteinů Hsc70, čímž umožňuje správné sbalování cílových proteinů (Wilbanks, 1996). CSPα se nachází hlavně v nervové tkáni, převážně v oblasti synaptických váčků. Ostatní proteiny nedokáží jeho nepřítomnost nahradit, což naznačuje vysoce specifickou roli při sbalování protein (Johnson et al., 2010).

Zjištěná mutace vede k deleci leucinu na pozici 116 v centrální doméně bohaté na cystein. Tato hydrofóbní doména hraje klíčovou roli v transportu CSPα převážně přes membránu endoplasmatického retikula (ER). Je zodpovědná za rozpoznávání a asociaci CSPα s membránou ER a následnou palmitoylaci, která je potřebná k uvolnění CSPα z ER. V případě, že nedojde k palmitoylaci, zůstává protein zadržován na membráně, což může mít za následek nedostatek tohoto proteinu v cílových tkáních (Greaves and Chamberlain, 2006). V důsledku nedostatku CSPα dochází k narušení formace SNARE-komplexu, což má negativní vliv na uvolňování presynaptických váčků do synaptické štěrbiny a tím pádem negativní dopad na správnou funkci neuronů (Johnson et al., 2010).

Špatné sbalování proteinů je příčinou řady neurodegenerativních onemocnění, jako je například Alzheimerova, Huntingtonova a Parkinsonova choroba (Johnson et al., 2010).

Bylo prokázáno, že CSPα hraje roli v Huntingtonově chorobě. Při huntigtonově chorobě dochází k tvorbě inkluzních tělísek obsahujících mutovaný huntingtin. Bylo prokázáno, že mutovaný huntingtin váže CSPα, kdežto u nemutované formy huntingtinu k této vazbě nedochází. Je možné, že v důsledku vychytávání CSPα nastává nedostatek tohoto proteinu a dochází k neurodegeneraci (Miller et al., 2003).

U myší s delecí genu DNAJC5 byla prokázána progresivní neurodegenerace a předčasné úmrtí (Fernández-Chacón et al., 2004), přičemž zvýšená exprese

α-synucleinu dokáže u těchto myší zastavit neurodegeneraci (Chandra et al., 2005).

Neuropreotektivní role tohoto proteinu dává naději, že ho v budoucnosti bude možné využít pro léčbu neurodegenerativních onemocnění (Johnson et al., 2010).

6. SOUHRN

Tato diplomová práce podává přehled o historii, vývoji a budoucnosti nových sekvenačních technik. Věnuje se možným přístupům analýzy dat a prakticky porovnává jednotlivé mapovací algoritmy na reálných datech.

Vzhledem k cílům práce byly v této diplomové práci porovnány a zhodnoceny aktuálně dostupné mapovací algoritmy Bfast, Bioscope, BWA, CLC bio, PerM a Novoalign. Jako nejpřesnější z nich se ukázal algoritmus Novoalign a jako nejrychlejší algoritmus PerM. Pro analýzu dat získaných sekvenací exomu pacienta s adultní formou autozomálně dominantní neuronální ceroid lipofuscinózy byl použit algoritmus Novoalign, který dokázal odhalit 14204 odchylek od referenčního genomu v kódujících oblastech. Zjištěné mutace byly funkčně anotovány a propojeny s daty získanými vazebnou a expresní analýzou. Byla nalezena jedna mutace, která splňovala podmínky dané charakterem onemocnění: neznámá, funkčně významná mutace v genu, který se nachází v oblasti vymezené vazebnou analýzou, jehož exprese je signifikantně změněna a porucha jeho funkce může způsobit neurodegenerativní onemocnění.

Konečným výsledkem této studie je určení jedné kandidátní mutace, která vede k deleci leucinu na pozici 116 v genu *DNAJC5*. Segregace této mutace s fenotypem ve studované rodině byla ověřena přímým sekvenováním. V současné době probíhají funkční studie vlivu této mutace na vznik onemocnění.

Výše uvedený experiment potvrzuje, že exomové sekvenování je vhodným postupem pro nalezené příčin vzácných dědičných onemocnění a že propojením výsledků sekvenování s dalšími experimenty jako je vazebná a expresní analýza umožňuje určení kauzální mutace i při osekvenování exomu pouze jednoho jedince.

Zavedené postupy jsou v současnosti využívány pro zpracování dalších sekvenačních dat v Laboratoři funkční genomiky a informatiky na Ústavu dědičných metabolických poruch 1.LF UK a VFN s cílem odhalit příčiny vzácných, dědičně podmíněných onemocnění.

Výsledky této práce jsou připravovány jako součást publikace v odborném zahraničním časopise a budou prezentovány na konferenci Dědičné metabolické poruchy – 26. pracovní dny a na bioinformatické konferenci ISMB/ECCB ve Vídni.

7. SEZNAM ZKRATEK

- ATP adenosintrifosfát
- ANCL adultní forma neuronální ceroid lipofuscionózy
- ASCII znaková sada využívaná pro zapisování skóre kvality ve formátu FASTQ
- b bází (bases), analogicky kb, Mb
- BAM binární forma formátu SAM (Binary Alignment/Map)
- bp páry bazí (base pairs)
- BWT Burrows-Wheelerova transformace
- CAL kandidátní pozice (Candidate Alignment Location)
- CEL datový soubor pro ukládání naměřených intenzit
- CIGAR položka formátu SAM nesoucí informace o alignmentu
- CNV variace v počtu kopií genomové DNA (Copy Number Variations)
- CSFASTA datový formát pro ukládání sekvencí v barevném prostoru sekvenátoru SOLiD
- CTSD katepsin D
- dATPαS dATPαS- deoxy-adenosine-5'-(α-thio)-trifosfát
- DNA deoxyribonukleová kyselina
- ddNTP dideoxyribonukleotidtrifosfát
- dNTP deoxyribonukleotidtrifosfáty
- ER endoplazmatické retikulum
- FASTA datový formát pro ukládání nukleotidových sekvencí
- FASTQ datový formát pro ukládání sekvencí a jim odpovídajícímu skóre kvality, je kombinací datových formátů FASTA a QUAL

- FM — typ indexu využívaný společně s Burrows-Wheelerovou transformací
- gDNA — genomová DNA
- LOD — dekadický logaritmus poměru dvou modelů: modelu s vazbou a modelu s nezávislou segregací (logarithm of odds ratio)
- MIP — molekulární inverzní próba (Molecular inversion probe)
- NCL — neuronální ceroid lipofuscinóza
- NCL4 — neuronální ceroid lipofuscinóza andultního typu
- NGS — nové sekvenační techniky (Next Generation Sequencing Technologies)
- PCR — polymerázová řetězová reakce (Polymerase Chain Reaction)
- PGM — Personal Genome Machine, nový sekvenátor pracující na principu detekce změny pH
- PHRED — skóre kvality báze, počítáno jako záporný dekadický logaritmus pravděpodobnosti, že je daná báze špatně určená, také viz. QV
- PPT1 — palmitoyl-protein thioesteráza
- PTP — sekvenační destička využívaná firmou 454 (PicoTiterPlate)
- QUAL — datový formát pro ukládání skóre kvality odpovídajícím
- QV — skóre kvality báze, počítáno jako záporný dekadický logaritmus pravděpodobnosti, že je daná báze špatně určená, také viz. PHRED skóre
- RFLP — polymorfiysmus délky restrikčních fragmentů (Restriction Fragment Length Polymorphism)
- RNA — ribonukleotidová kyselina
- SAM — formát využívaný pro uchovávání alignmentu (Sequence Alignment/Map)

- SMRT Single Molecule Real Time Sequencing, technologie umožňující sekvenovat jednotlivé molekuly v reálném čase
- SNP jednonukleotidové záměny (Single Nucleotide Polymorphisms)
- STR krátké tandemové repetice (Short Tandem Repeat)
- T_i tranzice
- TPP1 tripeptidyl peptidáza
- T_v transverze
- VCF datový formát pro ukládání odchylek od referenčního genomu (Variant Call Format)
- VNTR variabilní počty tandemových repetic (Variable Number of Tandem Repeats)

8. SEZNAM LITERATURY

- Abecasis GR, Cherny SS, Cookson WO, Cardon LR (2002): Merlin--rapid analysis of dense genetic maps using sparse gene flow trees. Nature genetics 30:97-101
- Adzhubei I a, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR (2010): A method and server for predicting damaging missense mutations. Nature methods 7:248-249
- Benjamini Y, Hochberg Y (1995): Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. Journal of the Royal Statistical Society 57:289-300
- Biesecker LG (2010): Exome sequencing makes medical genomics a reality. Nature genetics 42:13-14
- Blankenberg D, Von Kuster G, Coraor N, Ananda G, Lazarus R, Mangan M, Nekrutenko A, Taylor J (2010): Galaxy: a web-based genome analysis tool for experimentalists. Current protocols in molecular biology 89:19.10.1–19.10.21
- Boehme DH, Cottrell JC, Leonberg SC, Zeman W (1971): A dominant form of neuronal ceroid-lipofuscinosis. Brain 94:745-760
- Burneo JG, Arnold T, Palmer C a, Kuzniecky RI, Oh SJ, Faught E (2003): Adult-onset neuronal ceroid lipofuscinosis (Kufs disease): with autosomal dominant inheritance in Alabama. Epilepsia 44:841-846
- Chandra S, Gallardo G, Fernández-Chacón R, Schlüter OM, Südhof TC (2005): Alpha-synuclein cooperates with CSPalpha in preventing neurodegeneration. Cell 123:383-396
- Chen Y, Souaiaia T, Chen T (2009): PerM: efficient mapping of short sequencing reads with periodic full sensitive spaced seeds. Bioinformatics 25:2514-2521
- Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, Nayir A, Bakkaloğlu A, Ozen S, Sanjad S, Nelson-Williams C, Farhi A, Mane S, Lifton RP (2009): Genetic diagnosis by whole exome capture and massively parallel DNA

sequencing. PNAS 106:19096-19101

- Collins FS, Morgan M, Patrinos A (2003): The Human Genome Project: lessons from large-scale biology. Science (New York, N.Y.): 300:286-290
- Depristo M (2010): Data processing and analysis of genetic variation using next-generation sequencing. Dostupný z WWW: http://www.broadinstitute.org/gsa/wiki/index.php/File:Ngs_tutorial_depristo_1210.pdf.
- Depristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, Del Angel G, Rivas MA, Hanna M, McKenna A, Fennell TJ, Kernytsky AM, Sivachenko AY, Cibulskis K, Gabriel SB, Altshuler D, Daly MJ (2011): A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nature genetics 43:491-498
- Ebersberger I, Metzler D, Schwarz C, Pääbo S (2002): Genomewide comparison of DNA sequences between humans and chimpanzees. American journal of human genetics 70:1490-1497
- Elleder M, Sokolová J, Hřebícek M (1997): Follow-up study of subunit c of mitochondrial ATP synthase (SCMAS): in Batten disease and in unrelated lysosomal disorders. Acta neuropathologica 93:379-390
- Ewing B, Green P (1998): Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome research 8:186-194
- Fernández-Chacón R, Wölfel M, Nishimune H, Tabares L, Schmitz F, Castellano-Muñoz M, Rosenmund C, Montesinos ML, Sanes JR, Schneggenburger R, Südhof TC (2004): The synaptic vesicle protein CSP alpha prevents presynaptic degeneration. Neuron 42:237-251
- Ferragina P, Manzini G (2000): Opportunistic data structures with applications In 41st Annual Symposium on Foundations of Computer Science IEEE Comput. Soc, p. 390-398.
- Ferrer I, Arbizu T, Peña J, Serra JP (1980): A golgi and ultrastructural study of a dominant form of Kufs' disease. Journal of neurology 222:183-190

- Flicek P (2009): The need for speed. Genome biology 10:212
- Flicek P, Birney E (2009): Sense from sequence reads: methods for alignment and assembly. Nature methods 6:S6-S12
- Gentleman RC et al. (2004): Bioconductor: open software development for computational biology and bioinformatics. Genome biology 5:R80
- Goecks J, Nekrutenko A, Taylor J (2010): Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome biology 11:R86
- Greaves J, Chamberlain LH (2006): Dual Role of the Cysteine-String Domain in Membrane Binding and Palmitoylation-dependent Sorting of the Molecular Chaperone Cysteine-String Protein. Molecular Biology of the Cell 17:4748 - 4759
- Homer N, Merriman B, Nelson SF (2009a): BFAST: an alignment tool for large scale genome resequencing. PloS one 4:e7767
- Homer N, Merriman B, Nelson SF (2009b): Local alignment of two-base encoded DNA sequence. BMC bioinformatics 10:175
- Horner DS, Pavesi G, Castrignanò T, De Meo PD, Liuni S, Sammeth M, Picardi E, Pesole G (2010): Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing. Briefings in bioinformatics 11:181-197
- Huang DW, Sherman BT, Lempicki RA (2009a): Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols 4:44-57
- Huang DW, Sherman BT, Lempicki RA (2009b): Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic acids research 37:1-13 .
- Johnson JN, Ahrendt E, Braun JEA (2010): CSPa : the neuroprotective J protein 1. Cell 165:157-165

- Josephson S a, Schmidt RE, Millsap P, McManus DQ, Morris JC (2001): Autosomal dominant Kufs' disease: a cause of early onset dementia. Journal of the neurological sciences 188:51-60
- Kent WJ (2002): BLAT---The BLAST-Like Alignment Tool. Genome Research 12:656-664
- Kiialainen A, Karlberg O, Ahlford A, Sigurdsson S, Lindblad-Toh K, Syvänen A-C (2011): Performance of Microarray and Liquid Based Capture Methods for Target Enrichment for Massively Parallel Sequencing and SNP Discovery P. Tan, ed. PLoS ONE 6:e16486
- Kircher M, Kelso J (2010): High-throughput DNA sequencing--concepts and limitations. BioEssays : news and reviews in molecular, cellular and developmental biology 32:524-36
- Koboldt DC (2010): Challenges of sequencing human genomes. Briefings in bioinformatics 11:484-498
- Korlach J, Marks PJ, Cicero RL, Gray JJ, Murphy DL, Roitman DB, Pham TT, Otto G a, Foquet M, Turner SW (2008): Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures. PNAS 105:1176-1181
- Lander ES et al. (2001): Initial sequencing and analysis of the human genome. Nature 409:860-921
- Langmead B, Trapnell C, Pop M, Salzberg SL (2009): Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome biology 10:R25
- Li H, Durbin R (2009): Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754-1760
- Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R (2009): The Sequence Alignment/Map format and SAMtools. Bioinformatics: 25:2078-2079
- Li H, Ruan J, Durbin R (2008): Mapping short DNA sequencing reads and

calling variants using mapping quality scores. Genome research 18:1851-1858

- Mamanova L, Coffey AJ, Scott CE, Kozarewa I, Turner EH, Kumar A, Howard E, Shendure J, Turner DJ (2010): Target-enrichment strategies for next-generation sequencing. Nature methods 7:111-118
- Margulies M et al. (2005): Genome sequencing in microfabricated high-density picolitre reactors. Nature 437:376-380
- McKenna AH, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, Depristo M (2010): The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20:1297-1303
- Miller JR, Koren S, Sutton G (2010): Assembly algorithms for next-generation sequencing data. Genomics 95:315-327
- Miller LC, Swayne LA, Chen L, Feng Z-P, Wacker JL, Muchowski PJ, Zamponi GW, Braun JE a (2003): Cysteine string protein (CSP): inhibition of N-type calcium channels is blocked by mutant huntingtin. The Journal of biological chemistry 278:53072-53081
- Ng PC, Levy S, Huang J, Stockwell TB, Walenz BP, Li K, Axelrod N, Busam DA, Strausberg RL, Venter JC (2008): Genetic variation in an individual human exome. PLoS genetics 4:e1000160
- Ng SB, Buckingham KJ, Lee C, Bigham AW, Tabor HK, Dent KM, Huff CD, Shannon PT, Jabs EW, Nickerson D a, Shendure J, Bamshad MJ (2010a): Exome sequencing identifies the cause of a mendelian disorder. Nature genetics 42:30-35
- Ng SB, Nickerson DA, Bamshad MJ, Shendure J (2010b): Massively parallel sequencing and rare disease. Review Literature And Arts Of The Americas 19:119-124
- Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J (2009): Targeted capture and massively parallel sequencing of 12 human

exomes. Nature 461:272-276

- Nijssen PCG, Brekelmans GJF, Roos RAC (2009): Electroencephalography in autosomal dominant adult neuronal ceroid lipofuscinosis. Clinical neurophysiology : official journal of the International Federation of Clinical Neurophysiology 120:1782-6
- Nijssen PCG, Brusse E, Leyten ACM, Martin JJ, Teepen JLJM, Roos RAC (2002): Autosomal dominant adult neuronal ceroid lipofuscinosis: parkinsonism due to both striatal and nigral dysfunction. Movement disorders 17:482-487
- Nijssen PCG, Ceuterick C, Diggelen OP van, Elleder M, Martin J-J, Teepen JLJM, Tyynelä J, Roos RAC (2003): Autosomal dominant adult neuronal ceroid lipofuscinosis: a novel form of NCL with granular osmiophilic deposits without palmitoyl protein thioesterase 1 deficiency. Brain pathology 13:574-581
- Paszkiewicz K, Studholme DJ (2010): De novo assembly of short sequence reads. Briefings in bioinformatics 11:457-472
- Perkel J (2011): Making Contact with Sequencing's Fourth Generation. BioTechniques 50:93-95
- Picardi E (2009): Bioinformatics tools for fast and accurate reads mapping In Next generation sequencing workshop Bari. Available at: http://mi.caspur.it/workshop_NGS09/docs/Picardi_NGS09.pdf.
- Pop M, Salzberg SL (2008): Bioinformatics challenges of new sequencing technology. Trends in genetics 24:142-149
- Quinlan AR, Hall IM (2010): BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841-842
- Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP (2011): Integrative genomics viewer. Nature Biotechnology 29:24-26
- Ronaghi M, Karamohamed S, Pettersson B, Uhlén M, Nyrén P (1996): Real-time DNA sequencing using detection of pyrophosphate release. Analytical biochemistry 242:84-89

- Sanders SS (2011): Whole-exome sequencing: a powerful technique for identifying novel genes of complex disorders. Clinical genetics 79:132-133
- Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M (1977a): Nucleotide sequence of bacteriophage phi X174 DNA. Nature 265:687-695
- Sanger F, Nicklen S, Coulson AR (1977b): DNA sequencing with chain-terminating inhibitors. PNAS 74:5463-5467
- Schwartz S, Oren R, Ast G (2011): Detection and Removal of Biases in the Analysis of Next-Generation Sequencing Reads P. Lopez-Garcia, ed. PLoS ONE 6:e16685
- Teer JK, Mullikin JC (2010): Exome Sequencing: The Sweet Spot Before Whole Genomes. Human molecular genetics 19:R1-R7
- Thiele H, Nürnberg P (2005): HaploPainter: a tool for drawing pedigrees with complex haplotypes. Bioinformatics 21:1730-1732
- Tobaben S, Thakur P, Fernández-Chacón R, Südhof TC, Rettig J, Stahl B (2001): A trimeric protein complex functions as a synaptic chaperone machine. Neuron 31:987-999
- Tyynelä J, Palmer DN, Baumann M, Haltia M (1993): Storage of saposins A and D in infantile neuronal ceroid-lipofuscinosis. FEBS letters 330:8-12
- Wang K, Li M, Hakonarson H (2010): ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic acids research 38:1-7
- Wheeler D a et al. (2008): The complete genome of an individual by massively parallel DNA sequencing. Nature 452:872-876
- Wilbanks SM (1996): The Cysteine String Secretory Vesicle Protein Activates Hsc70 ATPase. Journal of Biological Chemistry 271:25989-25993
- William Blair & Company (2011): Next-Generation Sequencing Survey. Dostupný z WWW: http://www.genomicslawreport.com/wp-

content/uploads/2011/04/William-Blair-NGS-Report.pdf.