Katedra informatiky
Přírodovědecká fakulta
Univerzita Palackého v Olomouci

BAKALÁŘSKÁ PRÁCE

Vývoj metod pro analýzu dat pořízených technologií „Sekvenování nové generace“

2016

Ondřej Tom

Vedoucí práce: RNDr. Miroslav Kolařík, Ph.D.

Studijní obor: Aplikovaná informatika, kombinovaná forma

Bibliografické údaje

Autor:	Ondřej Tom
Název práce:	Vývoj metod pro analýzu dat pořízených technologií „Sekvenování nové generace“
Typ práce:	bakalářská práce
Pracoviště:	Katedra informatiky, Přírodovědecká fakulta, Univerzita Palackého v Olomouci
Rok obhajoby:	2016
Studijní obor:	Aplikovaná informatika, kombinovaná forma
Vedoucí práce:	RNDr. Miroslav Kolařík, Ph.D.
Počet stran:	32
Přílohy:	1 DVD
Jazyk práce:	český

Bibliograhic info

Author:	Ondřej Tom
Title:	Development of the methods for „Next generation sequencing“ data analysis
Thesis type:	bachelor thesis
Department:	Department of Computer Science, Faculty of Science, Palacký University Olomouc
Year of defense:	2016
Study field:	Applied Computer Science, combined form
Supervisor:	RNDr. Miroslav Kolařík, Ph.D.
Page count:	32
Supplements:	1 DVD
Thesis language:	Czech

Anotace

Byla vytvořena aplikace, která pracuje s nastavením a výstupy tzv. variant callerů (programů pro detekci mutací v analyzované nukleové kyselině, například DNA). V prostředí aplikace je možné spravovat rozsahy parametrů jednotlivých variant callerů a následně generovat skripty pro jejich spuštění ve všech možných nastaveních. Získané výstupy aplikace převádí do standardizovaného formátu a nahrává do své databáze. Na základě srovnání s referenčními daty (ručně ošetřené výsledky) je aplikace schopna určit nejúspěšnější nastavení pro jednotlivé variant callery a také nejúspěšnější kombinace variant callerů samotných.

Synopsis

The aim of this thesis was to create an application that will work with the settings and outputs of variant callers (programs for detecting mutations from DNA next generation sequencing data). The application's environment allows to manage the ranges of selected parameters of each of the variant caller and then generate scripts to run them in all possible settings. The outputs are converted into a standardized format and stored in application's database. Based on a comparison to reference data (manually curated results) the application is able to determine the most successful settings for individual variation scallers and the most successful combination of variant callers themselves.

Děkuji svému vedoucímu bakalářské práce RNDr. Miroslavu Kolaříkovi, Ph.D., za podporu při výběru tohoto tématu. Mé díky patří také Mgr. Nikolovi Tomovi, bez jehož průpravy a odborného dohledu by realizace této práce nebyla možná. V neposlední řadě děkuji také své přítelkyni a rodině za trpělivost a pochopení.

datum odevzdání práce podpis autora

Obsah

Seznam obrázků

Seznam tabulek

1 Úvod

Analýza dat ze sekvenování nové generace (NGS), které se stalo jednou z klíčových metod v biologickém výzkumu i diagnostice, je stále nevyřešený problém. Důvodem je jednak stálý příchod nových technologií na trh, tak i komplexnost a složitost dat ze standardních NGS metod.

Výběr a nastavení série programů (pipeline) pro analýzu NGS dat je obecně složitý proces. Většina bioinformatiků používá přednastavené hodnoty, alternativně testuje několik málo parametrů, nebo se inspiruje literaturou. Jelikož existuje mnoho druhů experimentů vyžadujících specifická nastavení, jsou tyto přístupy velmi často nedostatečné.

Cílem práce bylo vytvořit aplikaci, která pracuje s nastavením a výstupy variant callerů (programů pro detekci mutací), ale také dalších programů, které jsou součástí analytické pipeline.

V prostředí aplikace je možné spravovat rozsahy parametrů jednotlivých variant callerů a následně generovat skripty pro jejich spuštění ve všech možných nastaveních. Získané výstupy aplikace převádí do standardizovaného formátu a nahrává je do své databáze.

Na základě srovnání s referenčními daty (ručně ošetřená data reprezentující výsledky, kterých se snažíme dosáhnout) je aplikace schopna určit nejúspěšnější nastavení pro jednotlivé variant callery a také nejúspěšnější kombinace variant callerů samotných.

1.1 Struktura a význam DNA

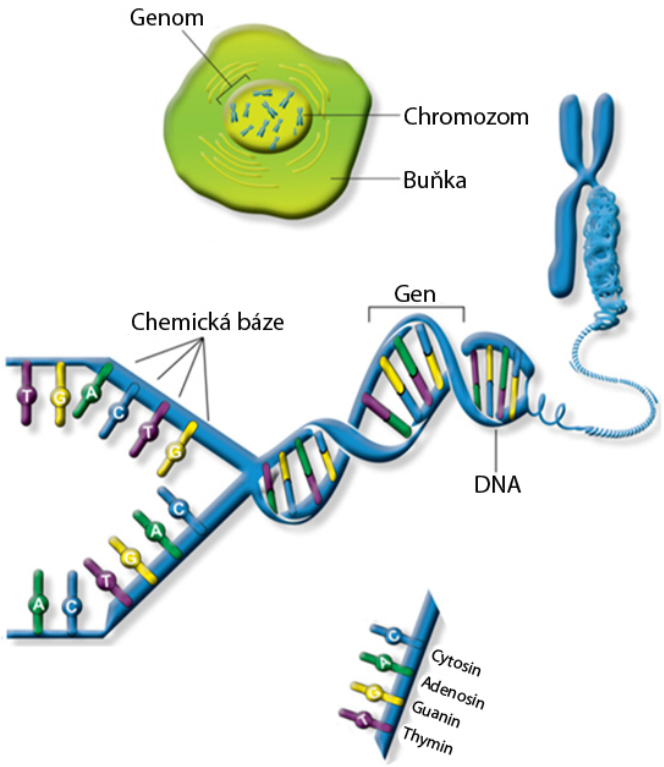
Deoxyribonukleotidová kyselina (DNA) je molekula nesoucí genetickou informaci pro růst, vývoj, fungování a reprodukci všech živých organizmů a mnoha virů. DNA je složena ze čtyř základních stavebních jednotek zvanými též nukleotidy (nt) či bázemi: cytosin (C), guanin (G), adenosin (A) a thymin (T). Jednotlivé nukleotidy jsou spojeny kovalentními vazbami a formují tak jeden ze dvou řetězců DNA. Oba řetězce jsou ve vzájemné anti-paralelní orientaci spojeny vodíkovými vazbami. Jedna molekula dvoušroubovice se označuje jako chromozom. V případě člověka má každá zdravá tělní buňka 23 párů (otec + matka) chromozomů ($3 * 10^9$ nukleotidů) uložených v buněčném jádře [1]. Soubor chromozomů se označuje jako genom.

Chromozomy lze z funkčního hlediska rozdělit na geny a mezigenové oblasti. Gen je tak částí (regionem) molekuly DNA o určité sekvenci nukleotidů (obrázek 1) a lze jej dále rozdělit na exony a introny. Sekvence exonů jednoho genu definují (kódují) složení a pořadí aminokyselin (základních stavebních jednotek) jedné bílkoviny, jež má určitou biologickou funkci jako např. stavební látka, enzym, hormon, biologická molekula zodpovědná za množení/fungování/smrt buňky (potenciální vznik rakoviny), atd. Proces syntézy produktu kódovaného genem se označuje jako genová exprese [1].

Každá buňka lidského těla obsahuje stejné geny (cca 23 000). Soubor všech

genů se označuje jako genotyp. Odlišnosti mezi buňkami/tkáněmi/orgány jsou dány v aktivaci/deaktivaci určitých kombinací genů tudíž, zjednodušeně řečeno, v produkci odlišných bílkovin.

Soubor všech pozorovatelných vlastností a znaků živého organizmu/buněk je označován jako fenotyp, a zároveň je definován jako spolupůsobení genotypu a působením vnějšího prostředí [1].



Obrázek 1: Struktura DNA a její uspořádání v buňce [2].

Molekula DNA je uložena v buněčném jádře a skládá se ze 4 chemických bází: adenosin (A), cytosin (C), thymin (T) a guanin (G). Jedna kompletní molekula DNA se označuje jako chromozom. Část DNA kódující bílkovinu se nazývá gen. Soubor chromozomů se označuje jako genom.

1.2 Sekvenování nové generace a bioinformatika

Sekvenování DNA je proces určování posloupnosti nukleotidů v molekule DNA. Znalost sekvence DNA se stala základem výzkumu mnoha biologických a medicínských oborů.

NGS je moderní vysokokapacitní alternativa sekvenování, která se na trhu poprvé objevila v roce 2005 [3] a od té doby prošla značným vývojem: přínos nových technologií, zvýšení kapacity, délky analyzovaných úseků DNA a přesnosti.

Pro výše zmíněné důvody a nižší cenu oproti klasickým metodám se NGS stalo velmi populární metodou jak ve výzkumu, tak v klinické praxi a diagnostice, kde

umožňuje zejména porozumět tomu, jak varianty v genomu ovlivňují fenotyp organizmu, vznik a vývoj nemoci a odpovědi na léčbu. Dnešní cena sekvenace jednoho lidského genomu klesla pod 1 000 $ a je tak 50 000x nižší, než tomu bylo při zahájení „Projektu sekvenování lidského genomu“ [4].

Bioinformatická analýza je stěžejní částí NGS experimentu. Ve většině případů, zejména ve výzkumu, je zapotřebí bioinformatika, který dokáže na základě biologické a technické podstaty experimentu správně vybrat a nastavit sérii programů.

1.3 Základní typy DNA sekvenačních experimentů

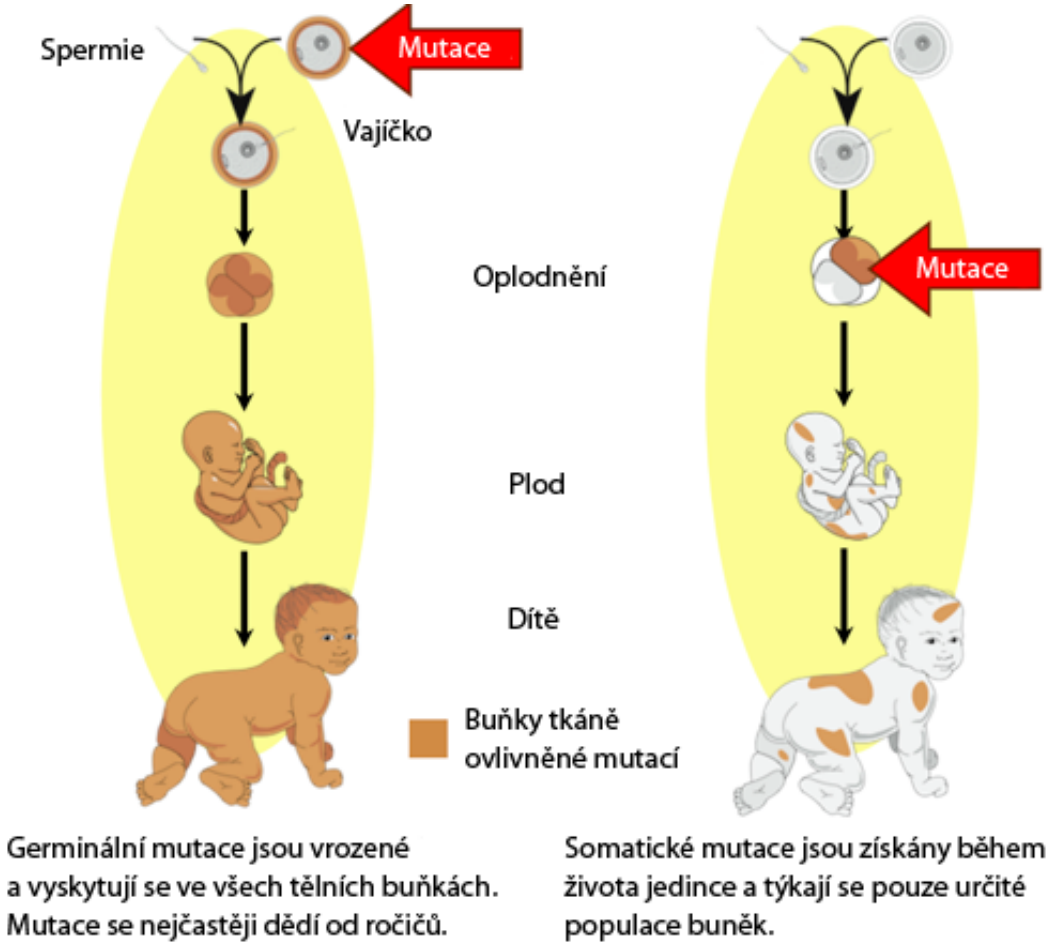
Faktorem pro výběr z těchto designů je jak biologická otázka, tak cena. Obecně lze experimenty rozdělit podle velikosti sekvenovaného regionu na:

1. Cílené — zaměřené na vybrané úseky, či geny.
2. Exomové — analýza všech exonů v genomu.
3. Celogenomové — analýza celého genomu.

Podle nároků na citlivost detekce mutací rozdělujeme sekvenování s nízkým, středním či hlubokým pokrytím. Pro bližší informace o „pokrytí“, viz kapitola 2.1.2.

Mutací se rozumí modifikace jednoho nebo více nukleotidů. V základu je mužeme roztřídit podle:

1. dědičnosti
   - Germinální — zděděná od rodičů, zpravidla se vyskytuje v 50 % nebo 100 % (jeden nebo oba chromozomy) všech buněk.
   - Somatické — získané v průběhu života organizmu, nepřenáší se do potomstva, vyskytují se pouze v určitých populacích buněk, častokrát asociovány s rakovinou.
2. typu záměny
   - Substituce — záměna jednoho či více nukleotidů (záměna jednoho nukleotidu označována jako SNV).
   - Inzerce — přidání jednoho či více nukleotidů.
   - Delece — ztráta jednoho či více nukleotidů (inzerce a delece se hromadně označují jako INDEL).



Obrázek 2: Typy dědičnosti mutací [5].

2 Obecný popis DNA NGS experimentu se zaměřením na pipeline a formáty dat

Analýza NGS dat je komplexní mnohakrokový proces závisející na množství programů, databází a zahrnující práci s velkými heterogenními daty. Pro velký úspěch NGS technologií a na nich postavených projektů existuje a je vyvíjena spousta nástrojů sloužících k analýze specifických částí pipeline [6]. Z tohoto důvodu se stává volba a nastavení optimálního programu složitou otázkou.

2.1 Struktura NGS experimentu

2.1.1 Laboratorní část

1. Izolace DNA, připravení DNA k sekvenaci.
2. Sekvenování DNA a produkce dat ve formátu FASTQ.

FASTQ je textový formát nesoucí informaci o sekvenované DNA ve formě pořadí nukleotidů + informaci o kvalitě každé báze. Jedna biologická sekvence osekvenovaná sekvenátorem a uložená ve formátu FASTQ se označuje jako „čtení“. Kvalita bází odráží pravděpodobnost vzniku chyby (substituce, inzerce, delece) při sekvenaci a je definována jako [7]:

$$Q = -10 \log_{10} P$$

2.1.2 Bioinformatická část – pipeline

Typická pipeline zahrnuje tyto kroky:

1. Kontrola kvality (QC) jak primárních FASTQ, tak i souborů produkovaných v dalších krocích (BAM, VCF).

2. Předzpracování (preprocessing) zahrnuje zejména odstraňování nekvalitních konců čtení (trimming), filtrování čtení (filtering), odstranění kontaminace.

3. Mapování čtení je (spolu s detekcí variant) stěžejní částí analytické pipeline. Účelem mapování je co nejpřesněji zarovnat čtení pořízených sekvenátorem (reprezentující určitou část genomu zkoumaných buněk, většinou dlouhá několik stovek bází) k referenčnímu genomu (miliardy bází). Referenčním genomem se rozumí známá DNA sekvence reprezentující genom určitého druhu. Jelikož se počet čtení produkovaných jedním spuštěním sekvenátoru pohybuje v řádech milionů a sekvence čtení nejsou zcela shodné s referenčním genomem z důvodu mutací a sekvenačních chyb (substituce, inzerce, delece), stává se mapování čtení na referenční genom složitým výpočetním úkonem. Počet čtení pokrývající určitou pozici v genomu se označuje jako hloubka pokrytí. Mapování čtení na referenci je klasicky uchováno ve formátu SAM (sequence alignment/map format), respektive v jeho binární formě BAM (binary SAM) [8].

4. Detekce mutací/variant (variant calling) je rozhodující částí DNA NGS analýzy dat. Funkcí programu pro detekci variant (variant calleru) není pouze detekovat rozdíly od reference, ale zároveň odlišit skutečné mutace od sekvenačních chyb a chyb mapování čtení. Výstupem je typicky soubor ve formátu VCF (variant call format) se seznamem mutací a informací k nim včetně p-hodnoty o spolehlivosti, zda je daná mutace reálná. Nejčastěji využívanými statistikami jsou Bayesian model [9] a Fisher exact test [10]. Pro dosažení kvalitnějších výsledků se také využívají doplňkové anotace a filtry [11]. Nejvyšší senzitivity a specificity při detekci mutací je dosahováno kombinováním více variant callerů dohromady [12].

5. Z důvodu nekonzistence formátů výstupů jednotlivých variant callerů je zapotřebí výsledky v podobě VCF normalizovat [13]. Následně, zejména pro

biologické účely, anotujeme detekované mutace informacemi z nejrůznějších biologických databází a filtrujeme za účelem selekce biologicky významných mutací.

Výběr programů závisí na designu studie a na typech mutací, které chceme detekovat.

Detekce germinálních mutací se využívá zejména ve výzkumu vzácných onemocnění, často se k analýze využívají příbuzní jedinci (matka, otec, dítě), není třeba vysoké hloubky pokrytí.

Pro detekci somatických mutací je žádoucí vyšší pokrytí. Zároveň se v analýze častokrát využívá srovnání vzorků z nádorové tkáně vůči zdravé tkáni jednoho pacienta za účelem odlišení sekvenačních chyb, které se vyskytují u obou vzorků, oproti reálným mutacím, které bývají specifické pro nádor.

Ve většině případů platí, že jeden program vyniká v detekci pouze určitého typu mutace za určité hloubky pokrytí.

3 Metody

3.1 Referenční data

Pro nalezení nejúspěšnějšího nastavení a kombinace variant callerů aplikace porovnává výsledky variant callingu s referenčními daty. Referenční data představují ověřený výstup, kterému se mají výsledky nejúspěšnějších kombinací variant callerů co nejvíce přibližovat.

Referenční data jsou reprezentována 254 mutacemi (236 SNV a 18 INDEL) v exonech genu TP53, které byly detekovány ve 40 vzorcích pacientů trpících chronickou lymfocytární leukemií. DNA byla sekvenována s ultra vysokým pokrytím (průměr 35 000 čtení na pozici) pomocí sekvenátoru Illumina Miseq. Mapování na referenci a variant calling bylo provedeno pomocí program CLC Genomic Workbench. Sekundárně byl pro variant calling použit algoritmus Shearwater R-balíčku DeepSNV [14]. Všechny výsledky prošly manuální inspekcí. Empiricky byla citlivost detekce stanovena na 0.2 % pro SNV a 0.5 % pro INDEL.

3.2 Generování inputu pro variant calling

Data produkována sekvenátorem ve formě FASTQ byla mapována na referenci pomocí nástroje BWA [15], následně byly SAM soubory konvertovány na BAM, sortovány a indexovány pomocí programu Samtools [15]. V této podobě slouží data jako vstup pro variant calling.

3.3 Použité technologie a prostředí vývoje

Většina výkonné části aplikace je naprogramována v jazyce PHP. Pro jeho jednoduchost je vývoj aplikace rychlý a flexibilní. Pro některé klíčové části aplikace

jako je např. paralelizace spouštěných procesů však vhodný není. Pro vývoj těchto částí byl vybrán jazyk Java.

Aplikace je propojena s několika MySQL databázemi, které představují úložiště dat, se kterými aplikace pracuje.

Dalším důležitým vývojovým nástrojem je PHP framework Nette. Zapouzdřuje veškerou funkcionalitu aplikace a je pomocí něj vytvořeno uživatelské prostředí.

Aplikace je spustitelná ve webovém prohlížeči. Grafický výstup je tedy definován pomocí HTML a CSS. Dynamické grafické prvky jsou naprogramovány pomocí javascriptové knihovny jQuery.

3.4 Proces

Procesem v rámci aplikace rozumíme určitou část pipeline, zejména variant callingu. Pro každý proces je třeba zadat:

1. Cestu k vstupním datům.
2. Formát vstupních dat.
3. Formát výstupních dat (výstup procesu je automaticky vstupem pro jeho subproces).
4. Cestu ke spouštěnému programu.
5. Šablonu pro generování skriptu spouštějícího program se specifickou kombinací parametrů.

   Šablonu lze libovolně upravovat. Touto cestou je možné k iteraci volání programu získat vlastní funkcionalitu.
6. Typ a hodnoty testovaných parametrů.

   Parametry dělíme na statické a množinové. Statické se skládají pouze z jedné nebo žádné hodnoty (pouze identifikátor parametru). Množinové se skládají z více hodnot – buď jejich vyjmenováním nebo zadáním rozsahu. Rozsah vymezíme pomocí minimální hodnoty, maximální hodnoty a přírůstku.

3.5 Generování spouštěcích skriptů

Po nakonfigurování procesů následuje automatizované sestavení skriptů se všemi kombinacemi nadefinovaných parametrů, které spouštějí program procesu. Skripty jsou umístěny do adresářové struktury, jejíž příklad vidíme na obrázku 5.

Ve srovnání se strukturou na obrázku 4 vidíme, že ve stromu přibily nové položky. V úrovni pod adresářem se jménem procesu se nacházejí číslované adresáře, kde každý z nich značí větev s unikátní kombinací parametrů procesu. Uvnitř této větve se nachází skript pro spuštění programu s určitou kombinací

```
#!/bin/bash
for file in [input_files_dir]*.[input_files_ext] do
    [cli_call] [params] $file > [output_files_dir]
done
```

Obrázek 3: Příklad šablony.

Text v hranatých závorkách je procesem nahrazován za adekvátní hodnoty.

Obrázek 4: Příklad pipeline.

Položka s ikonou složky reprezentuje kategorii procesů. Kategorie slouží pouze pro logické uspořádání procesů a nemá žádný hlubší význam. Položka s ikonou ozubeného kola reprezentuje proces.

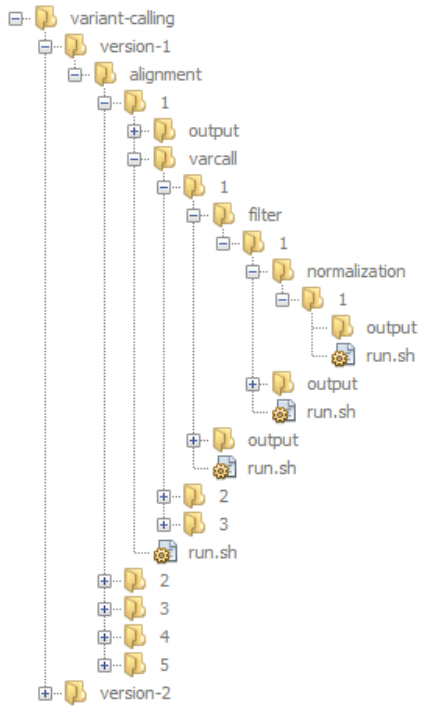
parametrů, adresář pro ukládání výstupních dat a adresáře, které reprezentují subprocesy, které berou vždy jako svá vstupní data výstupní data rodičovského procesu. Adresáře subprocesů mají obdobnou strukturu.

Klíčové jsou adresáře s výstupními daty, které se nachází v listových procesech stromu. Tato data reprezentují finální výsledky (seznam detekovaných mutací) pro určitou kombinaci parametrů jednoho variant calleru.

3.6 Paralelní spouštění skriptů

Kvůli vysokému počtu možných kombinací parametrů procesu (v běžných případech jich mohou být více než stovky) je vhodné procesy spouštět paralelně.

Nejdříve jsou paralelně spuštěny všechny procesy první úrovně se všemi kombinacemi svých parametrů. Tyto kořenové procesy mají cesty ke svým vstup-



Obrázek 5: Adresářová struktura vygenerovaných skriptů.

ním datům zadány uživatelem. V okamžiku, kdy je nějaký z procesů dokončen, jsou spuštěny všechny jeho subprocesy (opět se všemi kombinacemi) na jeho výstupních datech. Takto se situace opakuje dokud nejsou všechny subprocesy dokončeny.

Před spuštěním je možné nastavit maximální počet spouštěných vláken (1 běh procesu s jednou kombinací parametrů = 1 vlákno). Lze tím předejít vysokému zatížení procesorů nebo zahlcení operační paměti.

3.7 Import výstupních dat

Po dokončení procesů (doběhnutí pipeline) se výstupní data importují do databáze. Aplikace vyhledá procesy, které se nachází na konci pipeline a pro každý z těchto procesů je následně vytvořena samostatná tabulka, do které jsou vloženy jeho výstupní data. V tabulkách se nachází sloupce udávající informace o mutaci:

- chromosome (číslo chromozomu)
- position (pozice na chromozomu)
- reference (referenční báze)
- alternative (alternativní báze)
- sample (vzorek)

- type (typ záměny)

Struktura této tabulky je totožná se strukturou tabulky, ve které je uložena reference.

V případě cíleného či exomového sekvenování jsou oblastí zájmu pouze určité oblasti genomu. Regiony jsou definovány v tzv. BED (Browser Extensible Data) souboru. Účelem je zejména zrychlení analýzy, šetření místa a odstranění oblastí, které nejsou předmětem zájmu.

3.8 Metriky pro hodnocení konfigurací variant callingu

Na základě porovnání importovaných výsledků procesů s referencí v oblasti zájmu aplikace vypočítá statistické údaje pro:

- TP (true positive) – počet mutací shodných s referencí.
- FP (false positive) – počet mutací, které se v referenci nenachází.
- FN (false negative) – počet mutací, které se vyskytují v referenci, ale ne ve výsledných datech.
- TN (true negative) – počet pozic, na kterých nebyla detekována mutace jak ve výsledných datech, tak v referenci, ale spadají do oblasti zájmu.

Na základě výše uvedených údajů se spočítají následující metriky:

3.8.1 Senzitivita

Senzitivita neboli citlivost nabývá hodnot od 0 do 1 a vyjadřuje úspěšnost, s níž byla zachycena přítomnost mutací.

$$SE = TP/(TP + FN)$$

3.8.2 Specificita

Specificita nabývá hodnot od 0 do 1 a vyjadřuje úspěšnost, s jakou byly vybrány pozice, na kterých se mutace nevyskytují.

$$SP = TN/(TN + FP)$$

3.8.3 PPV

PPV neboli prediktivní hodnota pozitivního testu nabývá hodnot od 0 do 1 a vyjadřuje poměr správně detekovaných mutací vůči všem detekovaným mutacím.

$$PPV = TP/(TP + FP)$$

3.8.4 NPV

NPV neboli prediktivní hodnota negativního testu nabývá hodnot od 0 do 1 a vyjadřuje poměr pozic, na kterých správně nebyla detekována mutace vůči všem pozicím, ve kterých nebyla mutace detekována.

$$NPV = TN/(TN + FN)$$

3.8.5 Ballanced accuracy

Pro vybrání nastavení, která mají zároveň vysokou senzitivitu i specificitu byl použit aritmetický průměr senzitivity a specificity (ballanced accuracy).

$$BA = (SE + SP)/2$$

V některých případech získáváme relevantnější výsledky s použitím PPV a NPV místo senzitivity a specificity.

3.9 Little Profet

Algoritmus pro vyhodnocení nejúspěšnějších kombinací nastavení variant callerů a jejich kombinací nese název „Little Profet“. Průběh algoritmu je rozdělen do několika ucelených částí:

3.9.1 Porovnávání výsledných dat s referencí

V první části Little Profet jsou pro každou tabulku výsledků reprezentující výsledná data jedné určité kombinace parametrů vypočteny základní statistické údaje: TP, FP, FN, TN, SE, SP, PPV, NPV a BA.

3.9.2 Vyhodnocení šampionů

V následující fázi je ze seznamu na základě BA vyčleněno 10 nejúspěšnějších konfigurací, 5 s vyšší senzitivitou a 5 s vyšší specificitou.

3.9.3 Kombinování šampionů

V rámci skupiny 10 šampionů jsou vytvořeny všechny existující kombinace jejich dvojic. U každé z dvojic vytvoříme průnik jimi detekovaných mutací. Průnik v této chvíli reprezentuje výběr (seznam) mutací, který použijeme pro následné srovnání s referencí. Účelem průniku je zbavení se FP.

3.9.4 Zobrazení výsledků

Aplikace po dokončení analýzy zobrazí následující tabulky s výsledky:

- Obecné hodnocení pipeline – obsahuje výsledky srovnání všech pipeline s referencí.

- Hodnocení šampionů – obsahuje výsledky pipeline vyhodnocených jako šampionní.
- Hodnocení kombinací šampionů – obsahuje výsledky všech kombinací šampionů.

V tabulce je zobrazena kromě statistických výsledků také struktura pipeline, ze které je zřejmé, jak je potřeba nastavit konkrétní variant caller (nebo kombinaci variant callerů) pro danou úspěšnost.

4 O aplikaci

4.1 Správa uživatelů

Aplikaci mohou ovládat pouze přihlášení uživatelé. Uživateli je po přihlášení vytvořena „session“, která expiruje společně se zavřením prohlížeče. Správa uživatelů se nachází přímo v aplikaci. Uživateli může být přidělena jedna z těchto rolí:

- Administrátor – bez omezení.
- Uživatel – nemá přístup ke správě ostatních uživatelů.

4.2 Přenositelnost

Aplikace je spustitelná ve všech majoritních operačních systémech (Windows, Linux, OS X). Hlavní omezení zde kladou samotné variant callery, které jsou primárně tvořeny pro unixová prostředí. Aplikace tyto programy neobsahuje jako svou součást, ale jsou v ní definovány pouze jako cesty k jejich spuštění.

4.3 Použité nástroje

4.3.1 NetBeans IDE

Vývojové prostředí NetBeans IDE je robustním nástrojem pro efektivní vývoj, kompilaci a ladění aplikací. Již ve svém základu umožňuje vývoj v mnoha různých jazycích jako je např. C, C++, C#, Java, PHP a mnoho dalších. NetBeans mimo jiné umožňuje instalaci modulů, kterými se dá rozšířit jeho funkcionalita. NetBeans je vyvíjen jako svobodný software pod licencí GNU GPL [16].

4.3.2 Nette Framework

Nette je framework pro tvorbu webových aplikací v jazyce PHP. Zaměřuje se na eliminaci bezpečnostních rizik, podporuje AJAX, DRY, KISS, MVC a znovupoužitelnost kódu. Využívá událostmi řízené programování a je z velké části založen na použití komponent. Nette Framework je rovněž vyvíjen pod licencí GNU GPL [17].

4.3.3 XAMPP

XAMPP je multiplatformní distribuce webového serveru Apache. Kromě něj obsahuje ve svém balíčku také MySQL databázi a nezbytné nástroje pro běh programovacího jazyku PHP. Jedná se o svobodný software [18].

4.3.4 Adminer

Adminer je nástroj napsaný v jazyce PHP umožňující prostřednictvím webového rozhraní jednoduchou správu databáze MySQL, PostgreSQL, SQLite a dalších. Byl napsán jako lehčí alternativa PhpMyAdminu. Je šířený jako jediný zdrojový skript pod licencí Apache [19].

4.3.5 SourceTree

SourceTree je desktopový klient pro verzovací systém GIT a Mercurial. Umožňuje komunikaci s webovou službou pro správu GIT a Mercurial repositářů Bitbucket. Ten na rozdíl od konkurenční služby GitHub nabízí ve své bezplatné verzi možnost vytváření privátních repositářů. Bezplatná verze je limitována pro pět uživatelů. SourceTree je nabízen zdarma.

4.3.6 UnifiedGenotyper

UnifiedGenotyper je součástí balíčku nástrojů GAKT a využívá Bayesian modelu pro detekci nejvíce pravděpodobného genotypu a jeho frekvence [11].

4.3.7 VarDict

VarDict je program pro citlivou detekci mutací (SNV, více nukleotidové substituce, INDEL a strukturální varianty). Od většiny ostatních variant callerů se odlišuje několika funkcemi – provádí lokální zarovnání namapování čtení okolo inzercí a delecí, bere v úvahu tzv. softclipped čtení. Má specifické funkce pro cílené sekvenování a zvládá analyzovat BAM s ultra vysokým pokrytím [10].

4.3.8 VarScan

VarScan je na platformě nezávislý program pro detekci mutací z dat pořízených sekvenováním nové generace. Je vhodný pro detekci somatických, germinálních i strukturálních variant. Nástroj využíva robustního heuristicko-statitistického přístupu k detekci mutací splňující určitá kritéria pro hloubku pokrytí, kvalitu bází, frekvence mutace a statistické významnosti [20].

4.4 Hardwarová konfigurace

Analýza byla provedena na pracovní stanici s těmito parametry:

- Procesor: Intel Xeon 24x 2.40GHz.

- RAM: 64 GB.
- Vnější paměť: 280 GB PCI SSD.
- Operační systém: Ubuntu 14.04 64bit.

5 Výsledky

V rámci bioinformatického experimentu byly testovány 3 variant callery – VarDict, VarScan a UnifiedGenotyper. Analýza byla provedena na 40 reálných vzorcích. Pro účely analýzy byly zkoumány pouze mutace typu substituce.

5.1 Čas analýzy a počet nastavení pro jednotlivé variant callery

Nejvíce kombinací nastavení bylo vygenerováno pro VarDict a to zejména proto, že bylo využito filtrů pro nejrůznější parametry jako subprocesů. Tímto došlo k významnému snížení časových nároků na analýzu (protože se celá analýza nemusela pro každý vzorek spouštět od začátku), viz tabulka 1.

variant caller	počet kombinací	čas analýzy (h)	čas def analýzy (h)
VarDict	2880	40	14
VarScan	320	19	2,5
UnifieldGenotyper	4	10	4,5

Tabulka 1: Celkový počet kombinací a čas analýzy testovaných variant callerů. Vysoký počet kombinací a relativně krátká doba analýzy u VarDict je dána využítím subprocesů, kterým filtrujeme data generovaná v předchozím kroku. Výchozí nastavení (1 kombinace) je označeno jako „def“.

5.2 Detekční schopnost variant callerů

Na základě porovnání detekovaných mutací s referenčními daty (algoritmus Little Profet) byla pro každé nastavení jednotlivých variant callerů odvozena senzitivita a specificita.

V tabulce 2 je uvedena senzitivita a specificita pro nejúspěšnější nastavení každého variant calleru. Pro znázornění funkčnosti a užitečnosti aplikace byly generovány výsledky rovněž pro výchozí nastavení všech variant callerů, viz tabulka 3.

variant caller	senzitivita	specificity
VarDict	0,915	0,977
VarScan	0,944	0,994
UnifieldGenotyper	0,248	1

Tabulka 2: Senzitivita a specificita variant callerů s nejúspěšnějším nalezeným nastavením.
Variant callery se specifickými nastaveními byly seřazeny podle BA. V tabulce jsou pro jednotlivé variant callery zobrazeny senzitivita a specificita jejich nejúspěšnějších nastavení.

variant caller	senzitivita	specificity
VarDict	0,278	1
VarScan	0,218	1
UnifieldGenotyper	0,231	1

Tabulka 3: Senzitivita a specificita variant callerů ve výchozím nastavení.
Variant callery s výchozím nastavením (seřazeno podle BA) vykazují až 4x nižší senzitivitu (VarScan) než nejúspěšnější nastavení.

5.3 Senzitivita a specificita kombinace variant callerů

V případě spuštění klíčového algoritmu Little Profet pro všechny 3 variant callery se do výběru 10 nejlepších (seřazeno podle BA) dostala pouze nastavení variant calleru VarScan. Po sestavení kombinací všech dvojic a následném vytvoření průniku seznamů mutací, dopadla nejlépe kombinace dvou různých nastavení produkujících stejné seznamy mutací (senzitivita 0.944 a specificita 0.994).

6 Diskuse

NGS se stalo v rámci biologického a medicínského výzkumu, vývoje i diagnostiky velmi populární metodou umožňující analyzovat DNA ve vyšších kvalitativních i kvantitativních rozměrech. Primárním výstupem NGS metody jsou sekvence přečtené DNA, které je nutno dále zpracovat – vyčistit od chyb a dát jim hlubší biologický smysl. Tato fáze je úkolem bioinformatiky, která proto hraje v NGS oblasti nepostradatelnou roli.

Jelikož je NGS v dnešní době poměrně velmi rozšířené, existují pro standardní využití, jako např. detekci germinálních mutací u vybraných genů pacientů s určitým onemocněním, ověřené bioinformatické postupy umožňující relativně spolehlivou analýzu. To ovšem neplatí pro detekci somatických mutací. V těchto případech se mohou mutace vyskytovat ve velmi nízkých frekvencích a vyžadují proto nastavení pro velmi citlivou detekci. Úskalí takovéhoto nastavení je

vysoká pravděpodobnost falešné pozitivity, která velmi často odráží přítomnost sekvenačních chyb nebo chyb mapování čtení.

Nastavení pro takovéto typy experimentu je třeba testovat a optimalizovat, což může být velmi časově a intelektuálně náročné. Nejnověji se ukazuje, že nejpřesnější výsledky generují postupy kombinující více variant callerů dohromady. Doposud však neexistuje nástroj, který by dokázal automaticky vygenerovat skripty s velkým množstvím kombinací parametrů a zjistit, které nastavení poskytuje výsledky nejpodobnější referenčním datům.

Čas analýzy závisí na algoritmech samotného variant calleru a na počtu kombinací parametrů, se kterými je spouštěn. Pro generování některých typů parametrů je třeba spouštět celý algoritmus variant calleru znova, jiné typy parametrů mohou být zařazeny v subprocesu a aplikovány jako filtry na data vzniklá v předchozím kroku. Tento způsob nastavení velmi výrazně urychluje analýzu a byl využit pro generování výsledků VarDictu.

Součástí algoritmu Little Profet je porovnání detekovaných mutací s referenčními daty na jehož základě byla stanovena senzitivita a specificita pro všechna nastavení jednotlivých variant callerů, včetně nastavení výchozích. Po srovnání nejúspěšnějších a výchozích nastavení je jasně zřetelné, že nastavení vyhodnocené aplikací jako nejlepší zdaleka překonaly nastavení výchozí (až 4,3x vyšší senzitivita).

V následující fázi bylo ze všech nastavení vybráno 10 takových, které produkovaly nejpřesnější výsledky (seřazeno podle BA). Do výběru se dostaly pouze kombinace nastavení pro variant caller VarScan. Vytvořením průniku 2 nejlepších nastavení bylo dosaženo stejné senzitivity a specificity jako u nastavení nejlepšího, protože obě nastavení generovaly stejné výsledky.

V tomto případě se nepotvrdil autorův předpoklad, že nejlepší výsledky mají produkovat kombinace více druhů variant callerů. Příčin může být několik. Jako nejpravděpodobnější připadá typ dat. Většina experimentů, u nichž se prokázalo, že nejpřesnější výsledky v rámci detekce somatických mutací poskytují pipeline založené na průniku výsledků ze 2 variant callerů bylo postaveno na srovnávání nádorové a zdravé tkáně a výrazně nižší hloubce pokrytí. Zároveň nebylo cílem detekovat mutace o tak nízké frekvenci, jako ve zkoumaném případě. Také referenční data byla vytvořena a ověřena odlišným způsobem. Dalším významným důvodem může být volba odlišných variant callerů a celkově odlišné nastavení pipeline.

Dosažení 100% shody s referencí nebylo dosaženo z důvodu použití odlišné pipeline včetně variant callerů a ruční kontroly výsledků, kde bylo bráno v úvahu i to, zda se mutace vyskytuje i v jiných časových odběrech u jednoho pacienta (reálná mutace) nebo zda se mutace vyskytuje u více vzorků různých pacientů (sekvenační chyba). Tyto přístupy výrazně ovlivnily přítomnost TP a FP v referenčních datech.

Závěr

V rámci bakalářské práce se mi podařilo vyvinout aplikaci jejíž účelem je automaticky nakonfigurovat pipeline pro detekci mutací z dat pořízených technologií sekvenování nové generace. Splněny byly všechny dílčí cíle a aplikace nabyla plné funkčnosti. Testování proběhlo na reálných datech pacientů s chronickou lymfocytární leukémií.

Výsledkem bylo několikanásobné zvýšení senzitivity detekce mutací oproti standardním nastavením. Principem aplikace je automatické generování skriptů v nadefinovaném rozmezí hodnot vybraných parametrů, jejich paralelní spuštění, nahrání výsledků do databáze, porovnání s manuálně ošetřenými referenčními daty a stanovení nejlepších nastavení a kombinace programů.

Do budoucna je v plánu testovat funkčnost a přesnost aplikace na referenčních genomických datech z „National Institute of Standards and Technology“ a jiných typech dat, jako například syntetických datech určených pro detekci somatických mutací za využití vzorků z nádorové a zdravé tkáně. Zároveň jsou již ve vývoji a testování nové přístupy a algoritmy pro: 1) výběr nastavení, které mají být dále kombinována; 2) metody pro výpis a vykreslení statistik a detailů o mutacích a 3) zrychlení analýzy využitím mezipaměti výsledků (cache). V plánu je také začlenění metod pro detekci mutací využívajících strojového učení.

Na základě inovativnosti (aplikace podobného typu zatím neexistuje), funkčnosti a přehlednosti této aplikace jsem přesvědčen, že bude užitečná pro širokou společnost vědců a lékařských týmů využívajících technologii sekvenování nové generace. Zároveň bude o této aplikaci brzy publikován článek v mezinárodním impaktovaném vědeckém časopise.

Conclusions

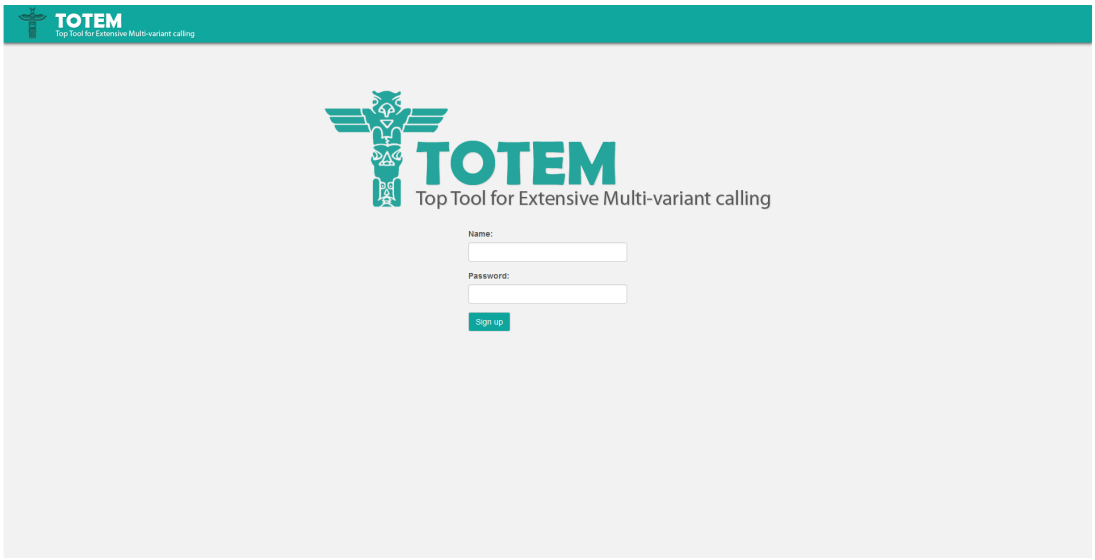
As part of the thesis I have developed an application whose purpose is to automatically configure the pipeline for detecting mutations from the data produced by next generation sequencing technology. All the objectives of the thesis were met and the application acquired full functionality. Testing was done on the real data of patients with chronic lymphocytic leukemia.

As a result, the sensitivity of the variant detection increased multiple times compared to the standard settings. The principle of the application is to automatically generate scripts with a specified range of values of selected parameters, execute them in parallel, upload the results into a database, compare them with a manually curated reference data and determine the best combination of programs settings and programs themselves.

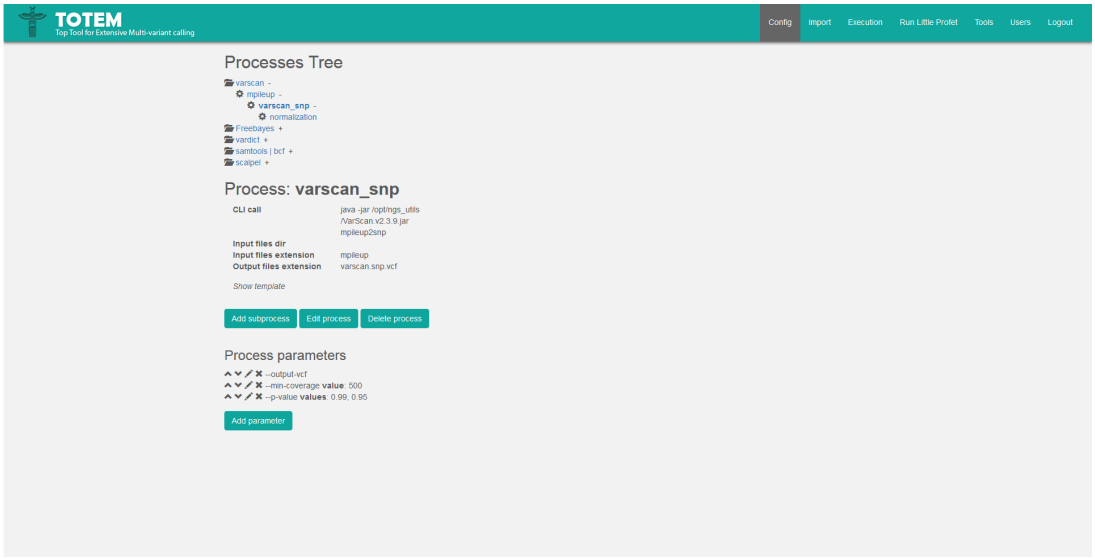
In the future, it is intended to test the functionality and accuracy of the application on a genomic reference data from the "National Institute of Standards and Technology", and also on other types of data, such as the synthetic data dedicated for the detection of somatic mutations using samples from the tumor and healthy tissue. At the same time, new approaches and algorithms for: 1) settings selection; 2) methods for extraction and rendering the statistics and mutations details and 3) acceleration of the analysis using the results cache memory are developed. The plan is also to incorporate methods for mutation detection using machine learning.

For the innovativeness (the application of such a type does not exist yet), functionality and clarity of the application, I believe it will be useful for a wide community of scientists and medical teams using next generation sequencing technology. Furthermore, an article about this application is going to be published soon in an impacted international scientific journal.

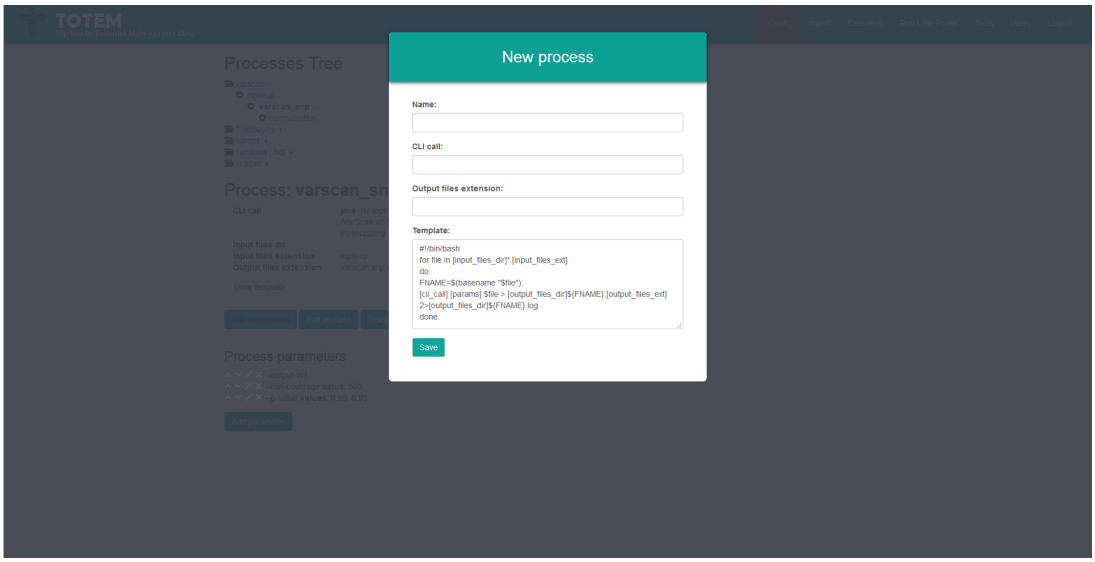
A Uživatelské prostředí aplikace



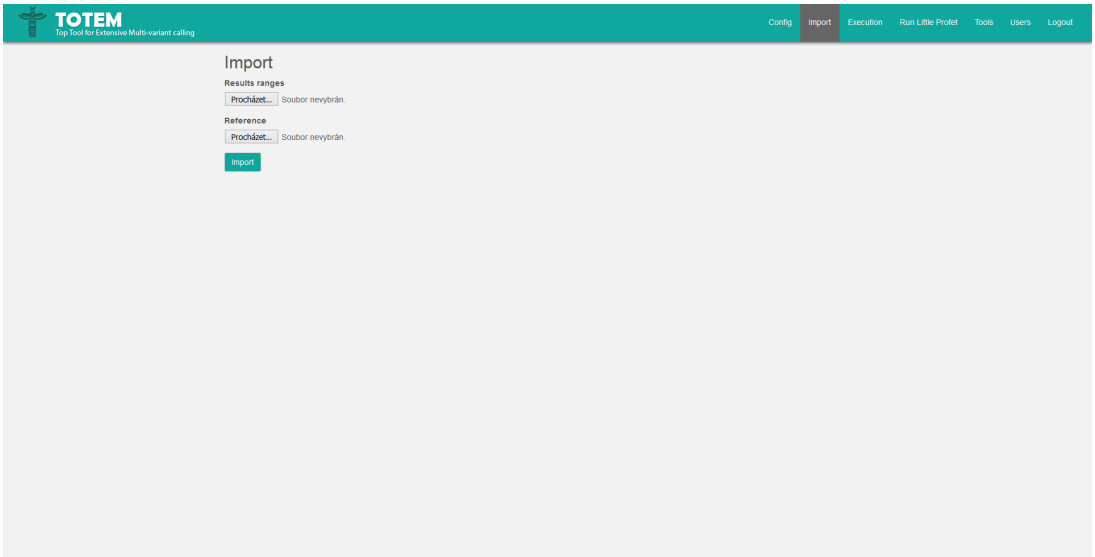
Obrázek 6: Přihlášení uživatele do aplikace.



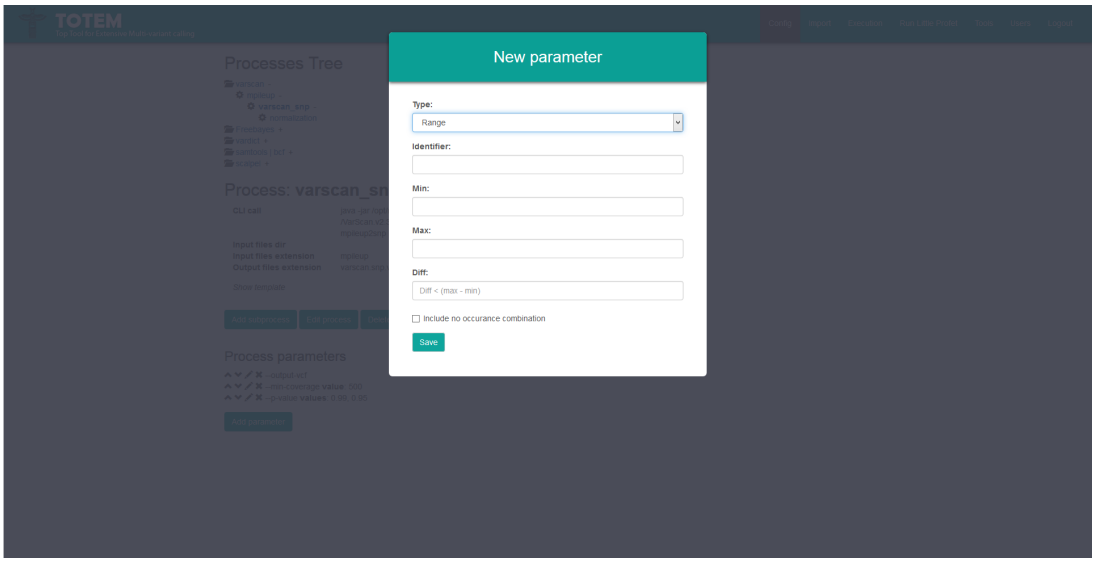
Obrázek 7: Konfigurace pipeline.



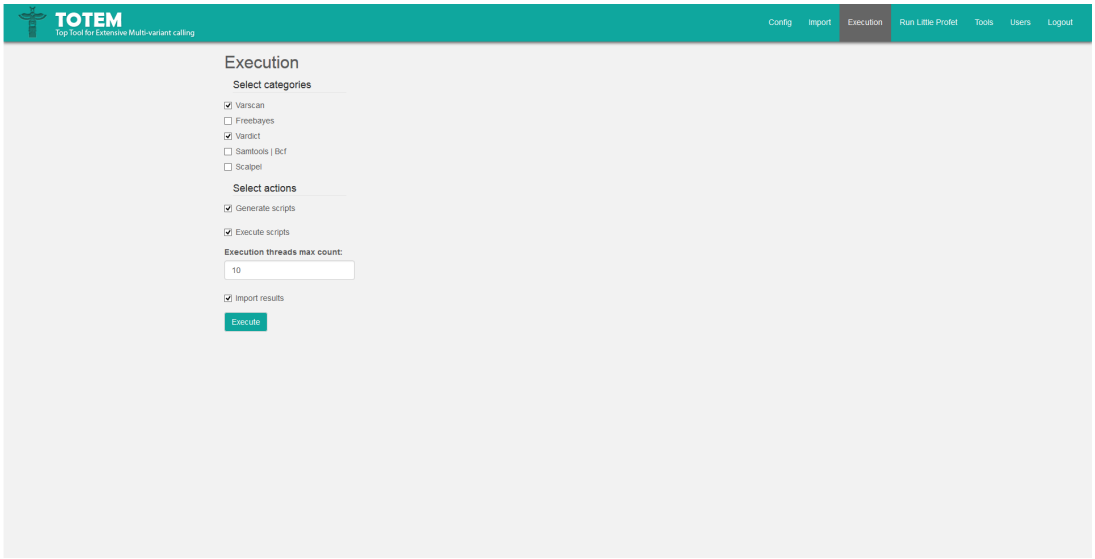
Obrázek 8: Konfigurace procesu.



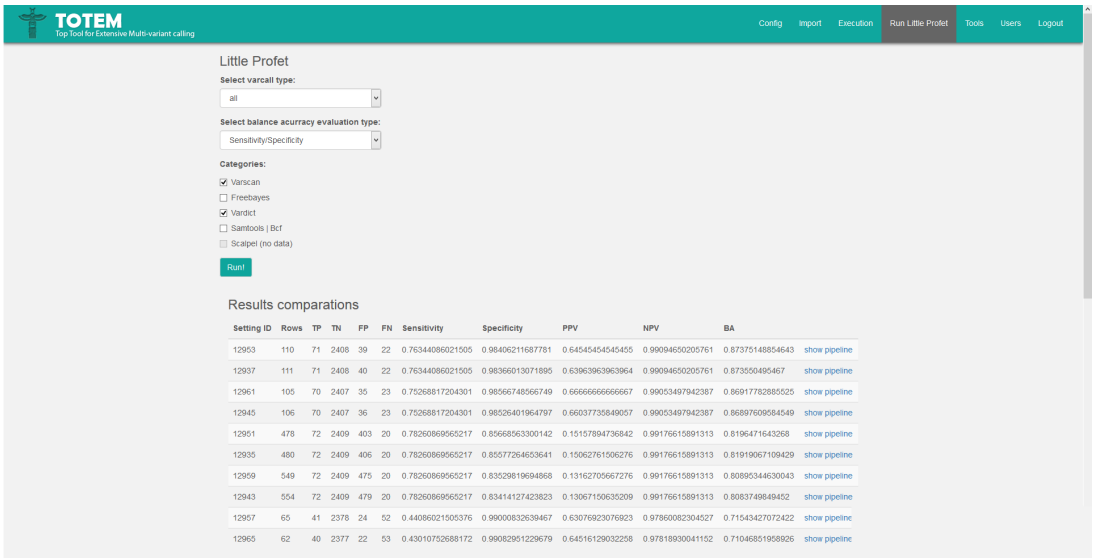
Obrázek 9: Import referenčních dat a oblastí zájmu.



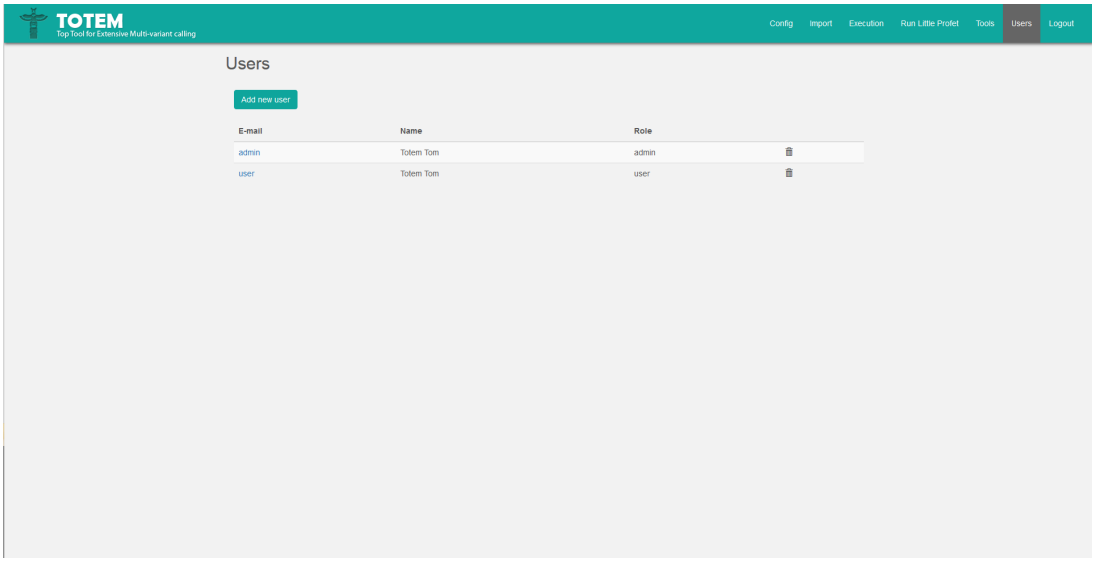
Obrázek 10: Konfigurace parametru procesu.



Obrázek 11: Paralelní spouštění pipeline a import výsledných dat do databáze.



Obrázek 12: Analýza výsledných dat – Little Profet.



Obrázek 13: Správa uživatelů.

B Obsah přiloženého DVD

DVD obsahuje soubory potřebné ke spouštění aplikace. Ve výchozím nastavení jsou aplikací spouštěny tři variant callery. Jejich parametry jsou nastaveny tak, aby bylo možné provést analýzu v krátkém čase.

`src/`
: Kompletní adresářová struktura aplikace.

`sql/`
: SQL soubory pro import potřebných dat do databáze.

`ext_totem/`
: Soubory spouštěných variant callerů.

`readme.txt`
: Instrukce pro instalaci a spuštění aplikace, včetně všech požadavků pro její bezproblémový provoz.

Literatura

[1] ALBERTS; JOHNSON; LEWIS aj. *Molecular Biology of the Cell*. 6. vyd., 2014.

[2] CORIELL INSTITUTE. *DNA, Genes, and SNPs*. Dostupný z: ⟨https://www.coriell.org/personalized-medicine/dna-genes-and-snps⟩.

[3] MARGULIES; EGHOLM; ALTMAN; ATTIYA. Genome sequencing in microfabricated high-density picolitre reactors: Nature. *Nature Publishing Group*. 2005.

[4] GOODWIN; MCPHERSON; MCCOMBIE. Coming of age: ten years of next-generation sequencing technologies: Nature Reviews Genetics. *Nature Publishing Group*. 2016.

[5] HIMME, Ben (ed.). *Genetic vs. somatic mutations*. Dostupný z: ⟨https://www.pathwayz.org/Tree/Plain/GAMETIC+VS.+SOMATIC+MUTATIONS⟩.

[6] PABINGER; DANDER; FISCHER aj. A survey of tools for variant analysis of next-generation genome sequencing data. *Oxford Journals*. 2013. Dostupný také z: ⟨http://bib.oxfordjournals.org/content/15/2/256⟩.

[7] COCK; FIELDS; GOTO; HEUER; RICE. The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. *Oxford Journals*. 2010. Dostupný také z: ⟨http://nar.oxfordjournals.org/content/38/6/1767⟩.

[8] LI; HANDSAKER; WYSOKER aj. The Sequence Alignment/Map format and SAMtools. *Oxford Journals*. 2009.

[9] GARRISON; MARTH. Haplotype–based variant detection from short–read sequencing. *Cornell University Library*. 2012.

[10] LAI; MARKOVETS; AHDESMAKI aj. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.* 2016.

[11] MCKENNA; HANNA; BANKS aj. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Cold Spring Harbor Lab Press*. 2010. Dostupný také z: ⟨http://genome.cshlp.org/content/20/9/1297⟩.

[12] ALIOTO; BUCHHALTER; DERDAK aj. A comprehensive assessment of somatic mutation detection in cancer using whole-genome sequencing. *Nature Communications*. 2015.

[13] CLEARY; BRAITHWAITE; GAASTRA aj. Comparing Variant Call Files for Performance Benchmarking of Next-Generation Sequencing Variant Calling Pipelines. *Cold Spring Harbor Laboratory*. 2015.

[14] GERSTUNG; BEISEL; RECHSTEINER aj. Reliable detection of subclonal single-nucleotide variants in tumour cell populations. *Nature Communications*. 2012.

[15] LI; DURBIN. Fast and accurate short read alignment with Burrows–Wheeler transform. *Oxford Journals*. 2009.

[16] WIKIPEDIA. *NetBeans*. Dostupný z: ⟨https://cs.wikipedia.org/wiki/NetBeans⟩.

[17] WIKIPEDIA. *Nette Framework*. Dostupný z: ⟨https://cs.wikipedia.org/wiki/Nette_Framework⟩.

[18] WIKIPEDIA. *XAMPP*. Dostupný z: ⟨https://en.wikipedia.org/wiki/XAMPP⟩.

[19] WIKIPEDIA. *Adminer*. Dostupný z: ⟨https://cs.wikipedia.org/wiki/Adminer⟩.

[20] KOBOLDT; ZHANG; LARSON aj. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. *Genome Research*. 2012.