POLITECHNIKA POZNAŃSKA

WYDZIAŁ INFORMATYKI I TELEKOMUNIKACJI
Instytut Informatyki

Praca dyplomowa licencjacka

IDENTYFIKACJA MUTACJI W NIEKODUJĄCYCH FRAGMENTACH GENOMÓW PACJENTÓW Z OSTRĄ BIAŁACZKĄ SZPIKOWĄ

Eliza Wielocha, 145171

Promotor
dr hab. Luiza Handschuh

POZNAŃ 2022

Pragnę serdecznie podziękować Pani dr hab.
Luizie Handschuh za umożliwienie mi
przeprowadzenia analizy otrzymanych danych biologicznych,
bez których napisanie niniejszej pracy byłoby niemożliwe.
Dziękuję również za poświęcony mi czas
oraz wszystkie cenne wskazówki.

Pragnę jeszcze osobno wyrazić wdzięczność
Damianowi Trzybińskiemu oraz
Sebastianowi Nawrotowi za okazane wsparcie
podczas pisania niniejszej pracy.

Spis treści

Rozdział 1

Wstęp

1.1 Wprowadzenie

Białaczka jest chorobą nowotworową narządów krwiotwórczych. Charakteryzuje się nadmiernym i nieprawidłowym rozrostem układu białokrwinkowego i pojawieniem się we krwi obwodowej dużej ilości niedojrzałych krwinek białych. Nazwa historyczna białaczki wywodzi się od białawego koloru próbki krwi pacjenta chorego na ostrą białaczkę. W niniejszej pracy skupiono się na jednym z rodzajów białaczek - ostrej białaczce szpikowej (AML). Jej częstość występowania rośnie wraz z wiekiem, dlatego jest ona powszechniejsza u dorosłych.

Do objawów białaczki można zaliczyć małopłytkowość, niedokrwistość i niedobór odporności. Rokowanie przy jej rozpoznaniu jest indywidualne dla każdego pacjenta ze względu na wiele czynników takich jak rodzaj białaczki, wiek czy mutacje genetyczne określające grupy ryzyka. Czas w jakim wykrywa się ten nowotwór jest bardzo istotny dla skuteczności jego leczenia. Jednak samą skuteczność wykrywania ogranicza mała specyficzność objawów, wysokie koszty badań oraz niski wskaźnik uczestnictwa w badaniach przesiewowych.

Nie znamy w pełni mechanizmów patofizjologii białaczek. Z tego powodu poznanie molekularnych aspektów choroby i ich wpływu na postęp nowotworu jest ważne dla odkrycia nowych możliwości personalizacji leczenia.

Sekwencjonowanie nowej generacji (NGS) pozwoliło na dużo lepszą eksplorację genomów nowotworowych. Dzięki temu powstały bazy zawierające zbiory danych dotyczących ludzkich nowotworów złośliwych. Badania w celu diagnostyki białaczek skupiają się w większości na detekcji mutacji w regionach kodujących białka, które stanowią tylko niewielką część genomu. Ostatnio jednak uwagę zaczynają przyciągać niekodujące RNA, które biorą udział w wielu ważnych procesach, w tym prowadzących do onkogenezy. Mogą być to procesy takie jak różnicowanie komórek, ich proliferacja, migrowanie czy apoptoza. Niekodujące RNA mają szansę stać się biomarkerami diagnostycznymi. Wraz z czasem i pojawiającymi się nowymi badaniami stało się jasne, że nie tylko mutacje w genach kodujących białka mają wpływ na przebieg nowotworów, ale także zmiany w obrębie genomu niekodującego.[1]

1.2 Cel i zakres pracy

Celem pracy była identyfikacja mutacji w niekodujących regionach genomów pacjentów z ostrą białaczką szpikową. Następnie wykonane zostało porównanie zidentyfikowanych mutacji między próbkami, aby sprawdzić czy u różnych pacjentów występują te same zmiany. Dodatkowo zostały zidentyfikowane mutacje w próbkach pacjentów po leczeniu białaczki i zostały porównane profile mutacji przed i po leczeniu. To podejście pozwoliło na wykrycie tzw. mutacji somatycznych

powstałych de novo w komórkach białaczkowych. Na koniec porównano zidentyfikowane mutacje z mutacjami z bazy *TCGA*.

Struktura pracy jest następująca: w rozdziale 2 umieszczono wykaz skrótów używanych w pracy. Rozdział 3 zawiera opis roli, jaką pełnią niekodujące regiony genomu i czym są mutacje w nich występujace. Przedstawia również zagadnienia związane z sekwencjonowaniem eksomowym oraz bazą danych *TCGA*. Opisuje też białaczki - czym są, ich klasyikację i patogenezę, ze szczególnym uwzględnieniem ostrej białaczki szpikowej, której dotyczy ninejsza praca. Rozdział 4 jest poświęcony przedstawieniu danych biologicznych. Obejmuje on opis narzędzi wykorzystanych do obróbki próbek jak i format plików. Przedstawione są również role poszczególnych parametrów opisujących próbki. Rozdział 5 zawiera opis identyfikacji mutacji, porównanie zmian między próbkami oraz omówienie wyników. Rozdział 6 stanowi podsumowanie pracy.

Rozdział 2

Wykaz skrótów

3'UTR - 3' untranslated region - rejon 3' niepodlegający translacji
5'UTR - 5' untranslated region - rejon 5' niepodlegający translacji
ALL - Acute lymphocytic leukemia - Ostra białaczka limfoblastyczna/limfocytowa
AML - Acute myeloid leukemia - Ostra białaczka szpikowa
AML M0 - Myeloblastic - Mieloblastyczna ostra białaczka szpikowa
AML M1 - Myeloblastic - Mieloblastyczna ostra białaczka szpikowa bez dojrzewania
AML M2 - Myeloblastic - Mieloblastyczna ostra białaczka szpikowa z dojrzewaniem
AML M3 - Promyeloctic - Promyeloytowa ostra białaczka szpikowa
AML M4 - Myelomonocytic - Mielomonocytowa ostra białaczka szpikowa
AML M5 - Monocytic - Monocytowa ostra białaczka szpikowa
AML M6 - Erythroleukemia - Ostra erytroleukemia
AML M7 - Megakaryocytic - Megakariocytowa ostra białaczka szpikowa
CLL - Chronic lymphocytic leukemia - Przewlekła białaczka limfoblastyczna/limfocytowa
CML - Chronic myeloid leukemia - Przewlekła białaczka szpikowa
DNA - Deoxyribonucleic acid - Kwas deoksyrybonukleinowy
eVai - Expert Variant Interpreter
GATK - Genome Analysis Toolkit
HGVS - Human Genome Variation Society
lncRNA - Long non-coding RNA - Długi niekodujący RNA
miRMut - nazwa narzędzia do annotacji mutacji w genach miRNA
miRNA - MikroRNA
NCI - National Cancer Institute - Narodowy Instytut Raka
ncRNA - Non-coding RNA - Niekodujący RNA
NGS - Next-generation sequencing - Sekwencjonowanie nowej generacji
RBase - Program relacyjnej bazy danych
RNA - Ribonucleic acid - Kwas rybonukleinowy
rRNA - Ribosmomal RNA - Rybosomalny RNA
snoRNA - Small nucleoar RNA - Mały jąderkowy RNA
TCGA - The Cancer Genome Atlas
tRNA - Transfer RNA - Transferowy RNA
UTR - untranslated region - region nieulegający translacji

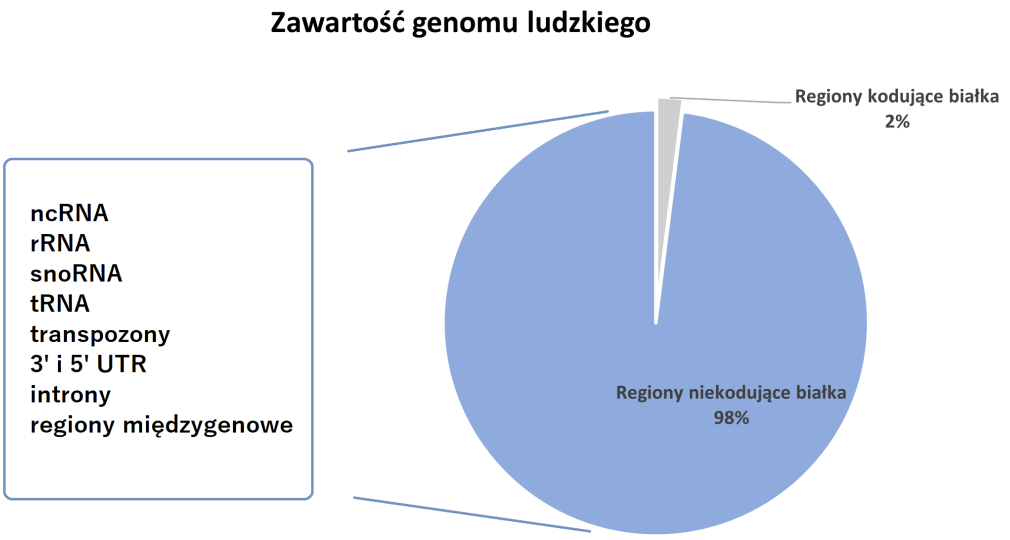
Rozdział 3

Podstawy teoretyczne

W niniejszym rozdziale przedstawione zostały szerzej informacje na temat niekodujących regionów genomu, mutacji oraz białaczek ze szczególnym skupieniem się na ostrej białaczce szpikowej.

3.1 Rola niekodujących regionów genomu w funkcjonowaniu komórek

Geny kodujące białka stanowią zaledwie 1% do 2% ludzkiego genomu. Pozostała część to regiony niekodujące białek (Rys. 3.1) [12].



Rysunek 3.1: Procentowa zawartość genomu człowieka

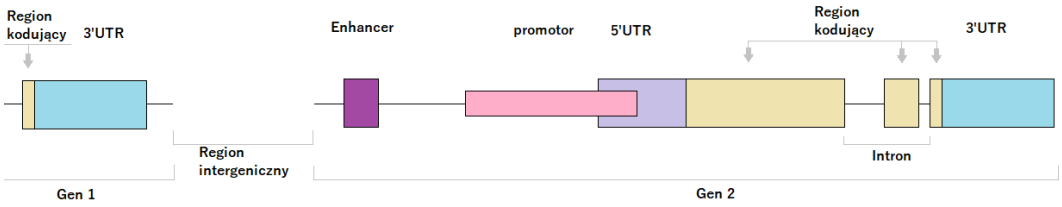
Niekodujący genom jest bogaty w elementy cis-regulatorowe *DNA*, takie jak:

- Promotory (ang. *promotors*), które zapewniają miejsca wiązania dla czynników białkowych. Białka te przeprowadzają transkrypcję. Promotory znajdują się zwykle tuż przed genem na tej samej nici *DNA*.
- Enhancery (ang. *enhancers*), które dostarczają miejsca wiążące dla białek aktywujących transkrypcję. Enhancery mogą znajdować się na nici *DNA* przed lub po kontrolowanym przez nie genie, czasem daleko od niego.

- Wyciszacze (ang. *silencers*), które stanowią miejsca wiązania dla białek hamujących transkrypcję. Mają to samo położenie co enhancery.
- Izolatory (ang. *isolators*), które zapewniają miejsca wiązania dla białek kontrolujących transkrypcję na wiele sposobów. Niektóre z nich blokują wspomaganie transkrypcji enhancerów, a inne zapobiegają zmianom strukturalnym w *DNA*, które hamują aktywność genu.

Regiony niekodujące mogą również dostarczać instrukcji do tworzenia różnych rodzajów *RNA*, takich jak: transferowe *RNA* (*tRNA*), rybosomalne *RNA* (*rRNA*), mikroRNA (*miRNA*) oraz długie niekodujące *RNA* (*lncRNA*). *tRNA* i *rRNA* biorą udział w procesie translacji. *miRNA*, inaczej krótkie *RNA*, uczestniczą w regulacji potranskrypcyjnej blokując proces tworzenia białek. *lncRNA*, czyli długie odcinki nieodującego *RNA*, pełnią funkcje w regulacji aktywności genów.

Niekodujące regiony *DNA* zazwyczaj znajdują się pomiędzy genami i są nazywane regionami między-genowymi. Niektóre z nich, np. introny, mieszczą się w obrębie genów kodujących białka (Rys. 3.2). Są one jednak usuwane podczas obróbki transkryptu, przed wytworzeniem białka. Mogą się w nich znajdować elementy regulatorowe, takie jak enhancery [7].



Rysunek 3.2: Rozmieszczenie regionów regulatorowych i kodujących wzdłuż genu[11]

3.2 Mutacje w niekodujących fragmentach genomu

W rozwoju nowotworów cechą charakterystyczną jest nabywanie przez komórki nowotworowe mutacji somatycznych. W prawidłowej tkance nie są one obecne. Niektóre z tych mutacji są tak zwanymi motorami napędowymi (ang. *drivers*) i powodują rozrost komórek nowotworowych. Jednak ich zdecydowana większość to tzw. mutacje pasażerskie (ang. *passengers*), które nie mają wpływu na rozwój nowotworu [2]. Do tej pory wykryto mnóstwo mutacji w kodujących regionach *DNA*, jednak jak już wiadomo jest to zaledwie ułamek genomu. Zmiany w regionach niekodujących mogą również pełnić rolę czynników wywołujących nowotwory lub stymulujących ich rozwój. Mutacje w niekodujących regionach *DNA* mogą modyfikować ekspresję genów *in cis* lub *trans* [10].

W promotorach zmiany wpływają na miejsca wiązania czynników regulujących transkrypcję. W przypadku regionów nieulegających translacji (ang. *untranslated region, UTR*) mutacje mogą zaburzać działanie *miRNA*. Jeśli chodzi o introny to znajdujące się w nich mutacje mogą wpływać na splicing. W regionach intergenicznych zmiany mogą mieć wpływ na geny znajdujące się powyżej lub poniżej miejsca mutacji. [11].

3.3 Sekwencjonowanie eksomowe

NGS umożliwia sekwencjonowanie dużych ilości *DNA*, na przykład pełnych genów lub eksomów, czyli wszystkich tych fragmentów *DNA* danej osoby, które dostarczają instrukcji do tworzenia białek (eksonów). Eksomy stanowią około 1% genomu człowieka. Sekwencjonowanie eksomowe

umożliwia identyfikację zmian w regionie kodującym białka dowolnego genu zamiast tylko w wybranych genach, jak ma to miejsce w sekwencjonowaniu tzw. paneli genów. Większość mutacji chorobotwórczych występuje w eksonach, dlatego też sekwencjonowanie eksomowe jest uważane za skuteczną metodę ich identyfikacji. [6]

3.4 Baza danych *TCGA*

W roku 2006 dwie instytucje - *National Cancer Institute* (*NCI*) i *National Genome Research Institute*, rozpoczęły pracę nad programem badania genomów nowotworowych o nazwie *The Cancer Genome Atlas* (*TCGA*). Program zakładał charakterystykę molekularną wybranych nowotworów pierwotnych na tle odpowiadających im tkanek prawidłowych. Obejmował on 33 typy nowotworów. W ciągu następnych kilkunastu lat *TCGA* wygenerował ponad 2,5 petabajta danych, które przyczyniły się do poprawy możliwości diagnozowania, leczenia oraz zapobiegania nowotworom. Dane te są publicznie dostępne i wykorzystywane przez wielu badaczy [14].

3.5 *MiRMut* - narzędzie do adnotacji mutacji

Narzędzie *miRMut* zostało stworzone przez pracowników Zakładu Genetyki Molekularnej Instytutu Chemii Bioorganicznej Polskiej Akademii Nauk w Poznaniu. Służy ono do adnotacji mutacji w genach *miRNA* na podstawie wyników sekwencjonowania całego eksomu (*WES*) lub sekwencjonowania całego genomu (*WGS*). Istnieje wiele narzędzi do adnotacji wariantów w genach kodujących, lecz *miRMut* jest jednym z pierwszych takich narzędzi do adnotacji mutacji w genach *miRNA*.

Dostęp do niego można uzyskać poprzez publicznie dostępne repozytorium na stronie *github.com* (*https://github.com/martynaut/mirnome-mutations*). Znajdują się tam skrypty odpowiadające za działanie narzędzia, a także referencyjne pliki potrzebne do niektórych etapów adnotowania. Dostępne są również przykładowe dane wejściowe i wyjściowe, aby ułatwić zapoznanie się z poszczególnymi etapami działania *miRMuta*.

Działanie narzędzia składa się z 6 kroków, które obejmują:

- charakterystykę mutacji, w tym identyfikację genów *miRNA* (na podstawie *miRBase* i *MirGeneDB*)
- lokalizację mutacji w strukturze prekursora *miRNA*
- wykrycie potencjalnego zaburzenia motywu wiążącego *RNA*
- przypisanie mutacji zgodnie z nomenklaturą *Human Genome Variation Society* (*HGVS*)
- określenie cech genów *miRNA*
- wizualizację

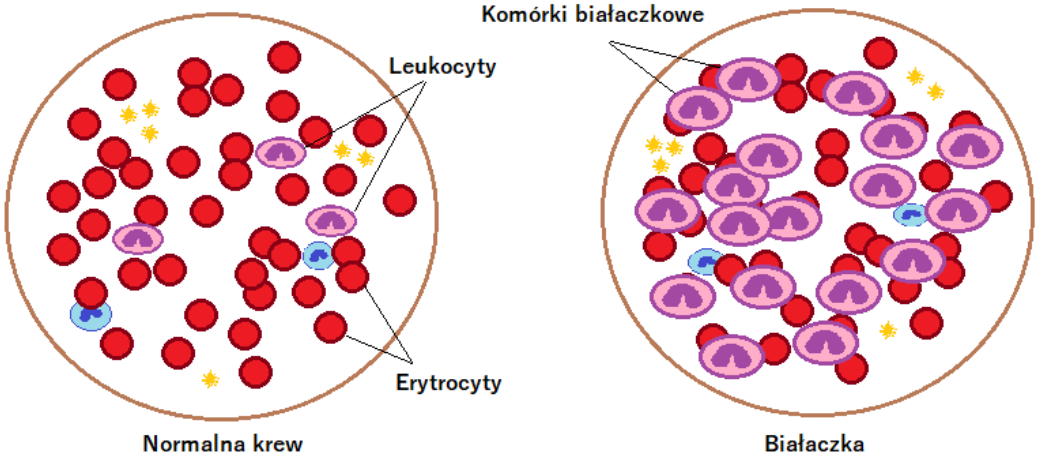
MiRMut został stworzony do analizy wyników sekwencjonowania *WGS* lub *WES* zmapowanych do genomu ludzkiego *GRCh37* (*hg19*) lub *GRCh38* (*hg38*). Wykorzystuje pliki w formatach *.vcf* lub *.csv*. Skrypty zakładają, że wszystkie potrzebne analizy jakościowe zostały wykonane przed ich użyciem [13]. MiRMut i dane wejściowe zawierają wyłącznie te mutacje, które przeszły filtrację.

3.6 Białaczki - czym są, klasyfikacja i patogeneza

Obecnie istnieje 130 typów i podtypów nowotworów hematologicznych. Powstają one z komórek, które pełnią funkcję czerwonych krwi, granulocytów i płytek krwi. Najczęściej występującymi

nowotworami układu krwiotwórczego są białaczki [4]. Charakteryzują się one klonalną ekspansją komórek krwiotwórczych, które naciekają szpik kostny (Rys. 3.3). Mogą również zajmować krew i inne tkanki. Namnażanie się komórek białaczkowych powoduje wypieranie prawidłowych komórek krwiotwórczych i utratę ich funkcji, co prowadzi do ciężkiego przebiegu choroby. Chorzy na białaczkę skarżą się na objawy takie jak:

- niedokrwistość
- małopłytkowość
- granulocytopenia
- chudnięcie, osłabienie
- gorączka, poty
- znacznie powiększone węzły chłonne



Rysunek 3.3: Schematyczny wygląd krwi zdrowej i typowej dla białaczki [5]

Nowotwory układu krwiotwórczego zwykle nie mają uwarunkowań rodzinnych i nie są też chorobami zakaźnymi. Czynniki środowiskowe przyczyniające się do powstawania białaczek to m.in. ekspozycja na promieniowanie i inne czynniki rakotwórcze (np. benzen) w miejscu pracy i domu czy niska częstotliwość pola elektromagnetycznego [1, 3, 9]. Białaczki klasyfikuje się ze względu na przebieg (ostra lub przewlekła) oraz linię rozwojową komórek, z których się wywodzą (mieloidalna/szpikowa lub limfoidalna). Wyróżnia się więc cztery główne podtypy:

- Ostra białaczka szpikowa (ang. *Acute myeloid leukemia, AML*)
- Przewlekła białaczka szpikowa (ang. *Chronic myeloid leukemia, CML*)
- Ostra białaczka limfoblastyczna lub limfocytowa (ang. *Acute lymphocytic leukemia, ALL*)
- Przewlekła białaczka limfoblastyczna lub limfocytowa (ang. *Chronic lymphocytic leukemia, CLL*)

ALL występuje najczęściej u dzieci, natomiast pozostałe wymienione podtypy objawiają się częściej u dorosłych [3]. W niniejszej pracy zajmowano się jedynie wynikami sekwencjonowania eksomów pacjentów z ostrą białaczką szpikową (*AML*). Prowadzonych jest i było mnóstwo badań

mających na celu lepszą klasyfikację białaczek, a co za tym idzie, skuteczniejsze leczenie. Obecnie wskaźnik przeżycia w przypadku ostrej białaczki limfoblastycznej (*ALL*) wynosi ponad 90% dla osób w wieku 14 lat lub poniżej. Niestety przy innych podtypach białaczek, takich jak *AML* lub *ALL* u dorosłych, wskaźnik przeżywalności 5 lat wynosi mniej niż 50%. Spowodowane jest to trudnościami takimi jak:

- Oporność na leczenie
- Heterogenność klonalna choroby podstawowej
- Słaba tolerancja na intensywne leczenie
- Choroby współistniejące osłabiające dodatkowo organizm [12]

3.7 Charakterystyka ostrej białaczki szpikowej

Ostra białaczka szpikowa nazywana jest również ostrą białaczką granulocytową lub ostrą białaczką nielimfocytową i jest najczęściej występującym typem ostrej białaczki.

Pojawia się ona, gdy szpik kostny zaczyna produkować w nadmiarze komórki, które nie są całkowicie dojrzałe. Normalnie przekształcają się one w białe krwinki, jednak w przypadku *AML* ich rozwój jest zablokowany, przez co nie mogą zwalczać infekcji. Szpik kostny wytwarza również nieprawidłowe czerwone krwinki i płytki krwi. W miarę upływu czasu prawidłowe krwinki białe i czerwone oraz płytki krwi są wypierane przez nieprawidłowe komórki.

Cechą odróżniającą *AML* od innych typów białaczek jest to, że istnieje osiem jej podtypów określanych na podstawie rodzaju komórki, z której białaczka się rozwinęła:

- Mieloblastyczna (ang. *Myeloblastic, M0*)
- Mieloblastyczna (ang. *Myeloblastic, M1*) - bez dojrzewania
- Mieloblastyczna (ang. *Myeloblastic, M2*) - z dojrzewaniem
- Promyelocytowa (ang. *Promyeloctic, M3*)
- Mielomonocytowa (ang. *Myelomonocytic, M4*)
- Monocytowa (ang. *Monocytic, M5*)
- Erytroleukemia (ang. *Erythroleukemia, M6*)
- Megakariocytowa (ang. *Megakaryocytic, M7*) [8]

Rozdział 4

Opis danych biologiczych

Na potrzeby pracy otrzymano wyniki z sekwencjonowania eksomów komórek krwi lub szpiku kostnego od pacjentów dotkniętych ostrą białaczką szpikową leczonych w Klinice Hematologii i Transplantacji Szpiku Szpitala Klinicznego Przemienienia Pańskiego Uniwersytetu Medycznego im. Karola Marcinkowskiego w Poznaniu. Materiał pozyskano we współpracy z lekarzami hematologami w ramach projektu pt.: "Zastosowanie współczesnej genomiki funkcjonalnej i bioinformatyki do charakteryzacji i tworzenia modeli procesów biologicznych o istotnym znaczeniu w medycynie i rolnictwie" (nrPBZ-MNiI-2/1/2005), którego kierownikiem był prof. dr hab. Marek Figlerowicz, a wykonawcą dr hab. Luiza Handschuh, promotor niniejszej pracy licencjackiej. Każdy z pacjentów wyraził pisemną, świadomą zgodę na udział w badaniu. Zgodę na prowadzenie badań wyraziła również Komisja Bioetyczna przy Akademii Medycznej im. K. Marcinkowskiego w Poznaniu (nr uchwały 34/06). W ramach kolejnego projektu, pt. „Analiza eksomów pacjentów z ostrą białaczką szpikową" (NCN, nr DEC-2017/01/X/NZ2/01906), którego kierownikiem była dr hab. Luiza Handschuh, próbki DNA 38 pacjentów poddano sekwencjonowaniu eksomów.

Dane z sekwencjonowania eksomowego zawierały wprawdzie tylko niewielki odsetek regionów niekodujących, ale ze względu na ich położenie w pobliżu genów kodujących białka, istniała większa szansa, że mutacje zidentyfikowane w tych regionach mogą mieć znaczenie dla regulacji ekspresji genów.

W przypadku danych z próbek przed leczeniem, otrzymano zbiór plików bez adnotacji (36 plików) oraz zbiór plików z adnotacją (38 plików). Przy danych z próbek po leczeniu dostępnych było jedynie 7 plików bez adnotacji.

W niniejszej pracy wykorzystano również dane dla *AML* pozyskane z bazy danych *TCGA* przez promotora pracy.

4.1 Opis narzędzi wykorzystanych do obróbki danych oraz ich formatu

Dostępne dane dla próbek przed leczeniem były w dwóch wariantach w osobnych folderach. Pierwszym były pliki potraktowane zestawem narzędzi *Genome Analysis Toolkit* (https://gatk.broadinstitute.org/hc/en-us, *GATK*). Jest to zbiór narzędzi służący do analizy danych z sekwencjonowania wysokoprzepustowego, służący przede wszystkim do wykrywania wariantów. Pliki wyjściowe były w formacie *.vcf*.

Drugi folder zawierał pliki po *GATK*, które dodatkowo przeszły analizę za pomocą programu *Expert Variant Interpreter* (https://www.engenome.com/product/, *eVai*). Są to tak zwane pliki z adnotacją. *eVai* został opracowany w celu wspierania genetyków w interpretacji wariantów germinalnych. Należy on do firmy *enGenome* i stosuje sztuczną inteligencję do klasyfikacji i prio-

rytetyzowania wariantów. Dodatkowo skraca on czas realizacji procesu interpretacji. *eVai* zwraca pliki w formacie *.tsv* jako pliki z adnotacją.

Dane uzyskane z próbek po leczeniu zostały przeanalizowane jedynie za pomocą *GATK*. Nie posiadały one plików z adnotacją.

4.2 Opis poszczególnych parametrów próbek w plikach z adnotacją

Pojedynczy plik *.tsv* (z adnotacją) posiadał 81 parametrów. Parametry te zostały wymienione w tabeli 4.1. Na potrzeby pracy skupiono się na 10 z nich: *CHR, START, END, REF, ALT, EFFECT, GENE, SAMPLE.AO, SAMPLE.RO* oraz *SAMPLE.COV*. Ich opis zamieszczono w tabeli 4.2.

CHR	DBSNP_1TGP_REF_freq	gnomAD_Hom_OTH
START	DBSNP_1TGP_ALT_freq	DANN_score
END	COMMON_1TGP_1_perc	dbscSNV_AB_score
REF	ESP[$version]_EA_freq	dbscSNV_RF_score
ALT	ESP[$version]_AA_freq	PaPI_pred
EFFECT	ESP[$version]_All_freq	PaPI_score
GENE	gnomAD_AF_ALL	PolyPhen-2_pred
ENS_GENE	gnomAD_Hom_ALL	PolyPhen-2_score
ENS_TRANSCRIPT	gnomAD_AF_Male	SIFT_pred
ENS_CANONICAL_TRANSCRIPT	gnomAD_Hom_Male	SIFT_score
RefSeq_map	gnomAD_AF_Female	PseeAC-RF_pred
SEQUENCE_FEATURES	gnomAD_Hom_Female	PseeAC-RF_score
TFBS_Id	gnomAD_AF_NFE	ClinVar_hotSpot
TFBS_name	gnomAD_Hom_NFE	ClinVar_RCV
EXON_INTRON_NUM	gnomAD_AF_AFR	ClinVar_clinical_significance
HGVS_C	gnomAD_Hom_AFR	ClinVar_rev_status
HGVS_P	gnomAD_AF_AMR	ClinVar_traits
CDS_DISTANCE	gnomAD_Hom_AMR	ClinVar_PMIDS
CDS_LEN	gnomAD_AF_EAS	Disease
AA_LEN	gnomAD_Hom_EAS	[$SAMPLE].GENO
OTHER_TRANSCRIPTS	gnomAD_AF_SAS	[$SAMPLE].QUAL
ExAC_AN	gnomAD_Hom_SAS	[$SAMPLE].GENO_QUAL
ExAC_AC	gnomAD_AF_ASJ	[$SAMPLE].FILTER
ExAC_AF	gnomAD_Hom_ASJ	[$SAMPLE].AF
ExAC_isTarget	gnomAD_AF_FIN	[$SAMPLE].AO
DBSNP	gnomAD_Hom_FIN	[$SAMPLE].RO
DBSNP_VERSION	gnomAD_AF_OTH	[$SAMPLE].COV

Tablica 4.1: Lista parametrów znajdujących się w każdym pliku *.tsv*

Numer parametru	Symbol parametru	Znaczenie parametru
1	CHR	Nr chromosomu
2	START	Pozycja początkowa wariantu
3	END	Pozycja końcowa wariantu
4	REF	Allel referencyjny
5	ALT	Allel alternatywny
6	EFFECT	Wpływ wariantu na transkrypt
7	GENE	Symbol genu
79	SAMPLE.AO	Głębokość odczytu dla allelu ALT
80	SAMPLE.RO	Głębokość odczytu dla allelu REF
81	SAMPLE.COV	Całkowita głębokość odczytu dla locus genomowego

Tablica 4.2: Opis parametrów użytych do analizy

Każdy plik posiadał średnio od 200 do 450 tysięcy wierszy (każdy wiersz odpowiadał jednemu wariantowi), które zostały poddane analizie.

4.3 Opis poszczególnych parametrów próbek w plikach bez adnotacji

Na potrzeby analizy otrzymano 36 plików z danymi przed leczeniem i 7 plików z danymi po leczeniu tych samych pacjentów. Oba te zbiory były w formacie *.vcf* (bez adnotacji). Tabela 4.3 przedstawia nazwy tych 7 plików, stanowiące unikatowe identyfikatory próbek.

Próbka przed leczeniem	Próbka po leczeniu
ID012	ID018
ID016	ID020
ID026	ID027
ID041	ID071
ID049	ID065
ID069	ID075
ID087	ID099

Tablica 4.3: Tabela przedstawiająca identyfikatory próbek przed leczeniem i odpowiadających im próbek po leczeniu

W przypadku plików *.vcf* parametrów próbki było jedynie 10: *#CHROM, POS, ID, REF, ALT, QUAL, FILTER, INFO, FORMAT* oraz *nazwa_próbki*. Spośród nich do analizy wykorzystano *#CHROM, POS, REF* i *ALT*. Parametry te były odpowiednikami kolejno *CHR, START, REF* oraz *ALT* z pliku *.tsv* z adnotacją.

4.4 Pliki zawierające dane z bazy *TCGA*

Na potrzeby analizy wykorzystane zostały dane pochodzące z projektu *TCGA LAML*, znajdującego się w bazie *TCGA*. Projekt ten dotyczył ostrej białaczki szpikowej (*AML*) a referencją

był ludzki genom *Hg38*. Folder zawierał 162 pliki z adnotacją, których analiza została wykonana z wykorzystaniem algorytmu *MuTect2*. Algorytm ten bazuje na porównaniu próbki nowotworowej z próbką tkanki prawidłowej pobraną od tego samego pacjenta. Analiza zwraca więc tylko mutacje somatyczne, a nie germinalne. Dodatkowo wyniki analizy poddano filtracji w celu pozbycia się wariantów znajdujących się w regionach kodujących białka.

W plikach znajdowały się więc tylko warianty somatyczne we fragmentach niekodujących genomu u pacjentów chorych na ostrą białaczkę szpikową. Na potrzeby pracy wykorzystano plik łączący wszystkie mutacje ze 162 próbek, przefiltrowane według pokrycia równego co najmniej 20 (warianty o niższym pokryciu były usuwane).

Rozdział 5

Analiza danych

Niniejszy rozdział pracy podzielony został na poszczególne etapy analizy danych. Składa się ona z filtrowania plików, porównania mutacji z poszczególnych próbek między sobą, identyfikacji mutacji somatycznych oraz porównania wyników z danymi z bazy *TCGA*. Zawiera on również dodatek w postaci wykorzystania narzędzia *miRMut* do adnotacji mutacji.

5.1 Wstępna analiza plików

Pliki wykorzystane do tego etapu analizy zawierały wyniki z sekwencjonowania eksomowego pacjentów chorych na ostrą białaczkę szpikową przed leczeniem. Były to dane potraktowane programem *eVai*, a więc z adnotacją, w formacie *.tsv*.

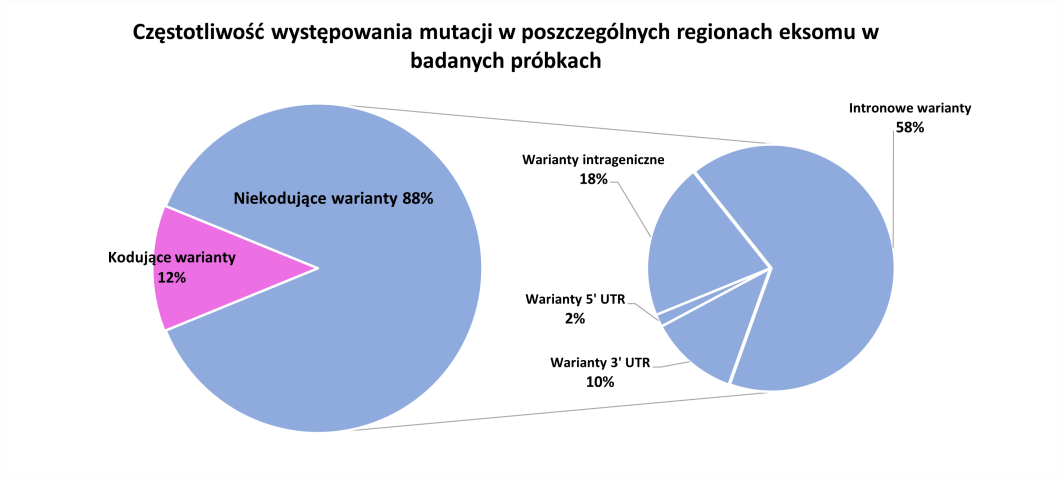
Próbka	Liczba wariantów	Próbka	Liczba wariantów
ID012	180 360	ID069	204 845
ID015	206 594	ID070	305 110
ID016	227 461	ID074	237 803
ID019	330 188	ID075	217 332
ID026	427 091	ID082	129 799
ID030	196 507	ID087	475 747
ID033	41 360	ID090	192 731
ID034	196 704	ID095	184 829
ID035	252 220	ID098	293 077
ID036	410 943	ID100	244 369
ID041	206 512	ID101	201 879
ID045	239 903	ID102	339 870
ID049	236 375	ID105	210 355
ID050	242 199	ID113	198 266
ID051	213 033	ID115	206 344
ID055	188 797	ID117	340 716
ID056	177 640	ID119	270 321
ID058	268 729	ID121	188 347
ID061	213 088	ID135	234 505

Tablica 5.1: Tabela przedstawiająca liczbę wszystkich mutacji w plikach przed filtrowaniem

Powyżej, w tabeli 5.1 przedstawiono liczbę wszystkich wariantów przypadających na próbki zawierające dane jeszcze przed filtrowaniem. Próbka *ID033* zawierała zdecydowanie mniej mutacji

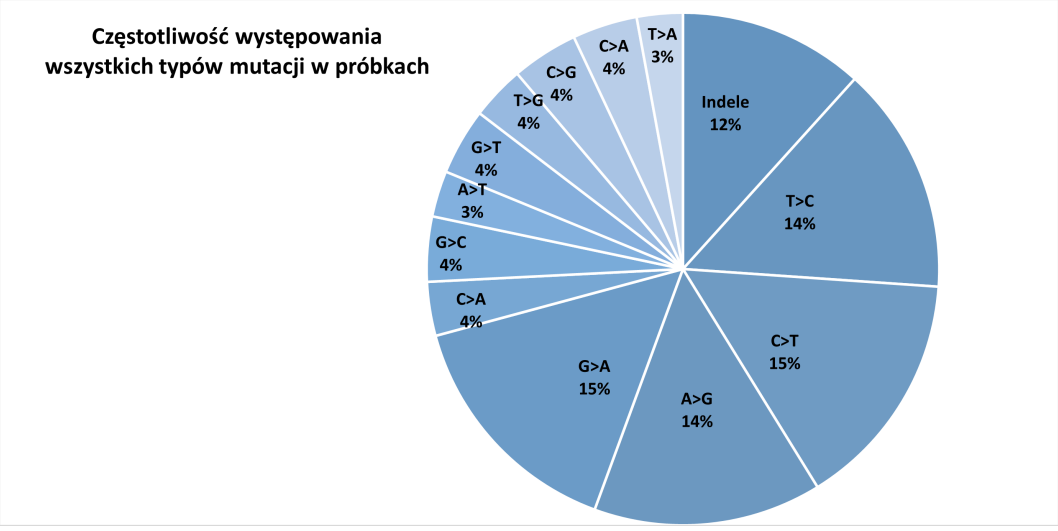
w porównaniu do innych, co wpłynęło na dalsze wyniki analizy.

Pliki zawierały zdecydowanie więcej wariantów niekodujących niż kodujących. Na Rys. 5.1 przedstawiono wykres średniej częstotliwości występowania wariantów kodujących i niekodujących, ze szczególnym uwzględnieniem wariantów w różnych typach sekwencji niekodujących. Wynika z niego, że jedynie 12% stanowiły mutacje w regionach kodujących.



Rysunek 5.1: Średnia częstotliwość występowania mutacji w poszczególnych regionach eksomów w badanych próbkach

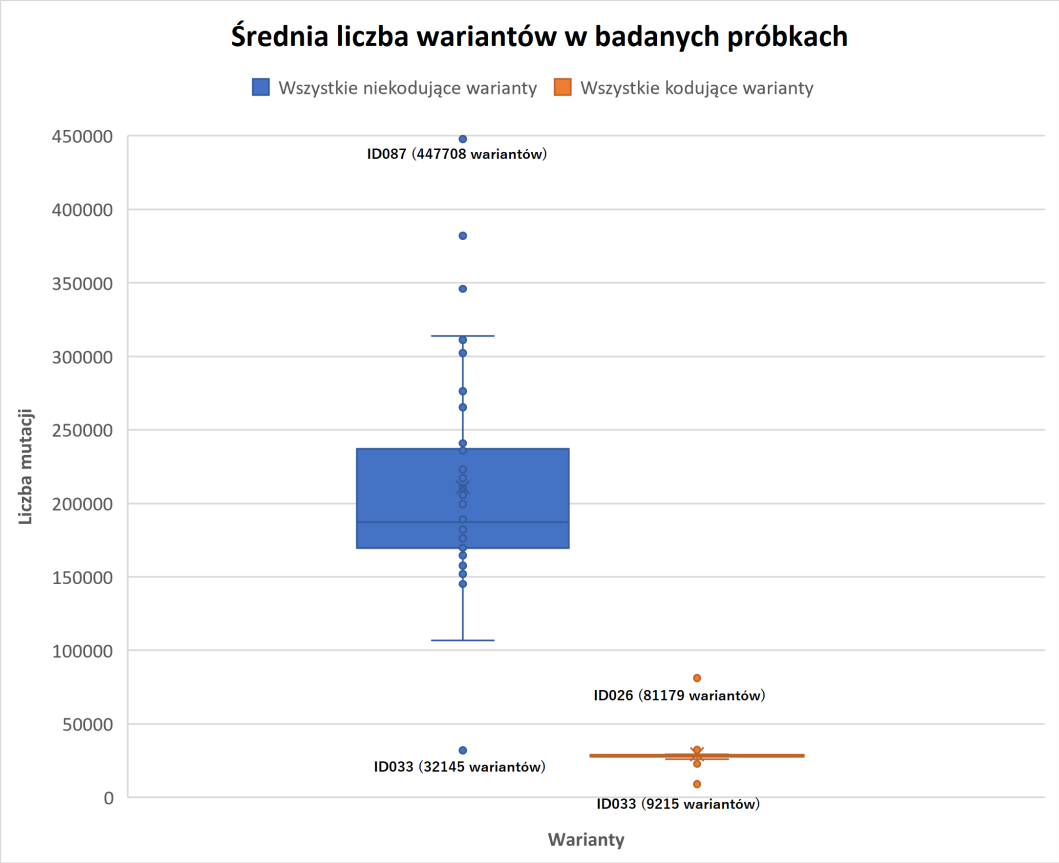
Przed filtrowaniem dodatkowo została zliczona średnia częstotliwość występowania poszczególnych typów mutacji w próbkach. Rezultaty można zobaczyć na Rys. 5.2. Największy odsetek w próbkach (po 14-15%) stanowiły substytucje G>A i C>T oraz ich odwrotności. Reszta zmian pojedynczych nukleotydów zamykała się w częściach stanowiących 3-4%. Natomiast indele stanowiły 12%.



Rysunek 5.2: Średnia częstotliwość występowania wszystkich typów mutacji w badanych próbkach.

Wykres na Rys. 5.3 przedstawia jak wyglądała średnia liczba wariantów w regionach niekodujących i kodujących w danych przed filtrowaniem. Należy pamiętać, że pliki wejściowe zawierały dużą liczbę wariantów o niskim pokryciu, które potem były odfiltrowywane. Dlatego też liczby

na wykresie sięgały wartości 450 tys. a średnia liczba kodujących wariantów jest tak słabo widoczna. Dwoma najmniejszymi wartościami odstającymi były warianty znajdujące się w próbce *ID033* - zarówno dla fragmentów niekodujących jak i kodujących. Jak już wspomniano w podrozdziale dotyczącym wstępnej analizy, próbka *ID033* zawierała bardzo mało wariantów. Dlatego też przy podziale na mutacje we fragmentach kodujących i niekodujących zaniżała ona średnią. W przypadku wartości najwyższych - należały one do próbek *ID087* we fragmentach niekodujących i *ID026* w kodujących, ponieważ te 2 próbki zawierały w sobie najwięcej wszystkich wariantów.



Rysunek 5.3: Średnia liczba wszystkich mutacji w badanych próbkach

5.2 Filtrowanie danych w celu znalezienia mutacji w regionach niekodujących

Na potrzeby filtrowania został stworzony skrypt *filtering_script.py* w języku Python. Skrypt ten ładował plik jako ramkę danych, a następnie na podstawie wybranych wartości parametrów (kolumn) usuwał poszczególne wiersze z ramki. Skrypt brał pod uwagę następujące parametry i ich wartości:

- *CHR*, *START* i *END* - Pozbywanie się takich samych wariantów

  Były to pierwsze 3 kolumny znajdujące się w pliku danych, odpowiadające za nr chromosomu, jego pozycję początkową i końcową. Skrypt sprawdzał czy w pliku istnieje więcej niż jeden wiersz z powtarzającymi się wartościami tych 3 parametrów. Jeśli taka sytuacja miała miejsce, pozostawiany był tylko 1 wiersz a pozostałe usuwano. W ten sposób plik zawierał niepowtarzające się warianty.

- *SAMPLE.AO*, *SAMPLE.RO* oraz *SAMPLE.COV* - Sprawdzenie jakości wariantów

  Te 3 parametry reprezentowały pokrycie (głębokość odczytu) dla locus genomowego. Pokrycie to średnia liczba odczytów reprezentujących dany nukleotyd w zrekonstruowanej sekwencji. Celem filtrowania było pozbycie się ewentualnych błędów sekwencjonowania. Niskie pokrycie poddaje w wątpliwość wiarygodność wyników. Skrypt usuwał wiersze, których wartości parametrów *SAMPLE.AO* i *SAMPLE.RO* były mniejsze niż 10. Jako, że *SAMPLE.AO* i *SAMPLE.RO* składają się na parametr *SAMPLE.COV*, to w ramce zostawały jedynie wiersze o wartościach *SAMPLE.COV* większych lub równych 20.

- *EFFECT* - Rozdzielanie regionów na kodujące i niekodujące

  Parametr *EFFECT* reprezentował region, którego dotyczył dany wariant. Analiza skupiała się na identyfikacji mutacji w regionach niekodujących. Skrypt zostawiał więc jedynie wiersze zawierające wartości parametru *EFFECT*, które zostały wymienione w tabeli 5.2. Indexy innych wierszy były zapisywane do osobnej listy i z nich tworzono plik zawierający fragmenty kodujące.

3_prime_UTR_variant	mature_miRNA_variant
5_prime_UTR_variant	non_coding_transcript_variant
TFBS_ablation	regulatory_region_variant
downstream_gene_variant	splice_acceptor_variant
intragenic_variant	splice_donor_variant
intergenic_variant	splice_region_variant
intron_variant	upstream_gene_variant

Tablica 5.2: Pożądane wartości parametru *EFFECT* w analizie danych

Rezultatem działania skryptu było utworzenie dwóch plików - jeden zawierał przefiltrowane warianty we fragmentach niekodujących, a drugi we fragmentach kodujących. Skrypt znajduje się w załącznikach do pracy. Po jego zastosowaniu w każdym z plików zawierających warianty niekodujące liczba wierszy drastycznie zmalała. Pliki przed filtrowaniem posiadały od 200 tys. do 450 tys. wierszy. Natomiast po filtrowaniu ich liczba wahała się od 167 do prawie 60 tys.

Stworzony został wykres przedstawiający średnią liczbę wariantów w regionach niekodujących i kodujących w danych po filtrowaniu (Rys. 5.4). Widać na nim jak bardzo zmalał zakres liczbowy na osi Y reprezentującej liczbę wariantów. Sięga on tutaj do 60 tys, podczas gdy przed filtrowaniem próbki zawierały do 450 tys. wariantów.

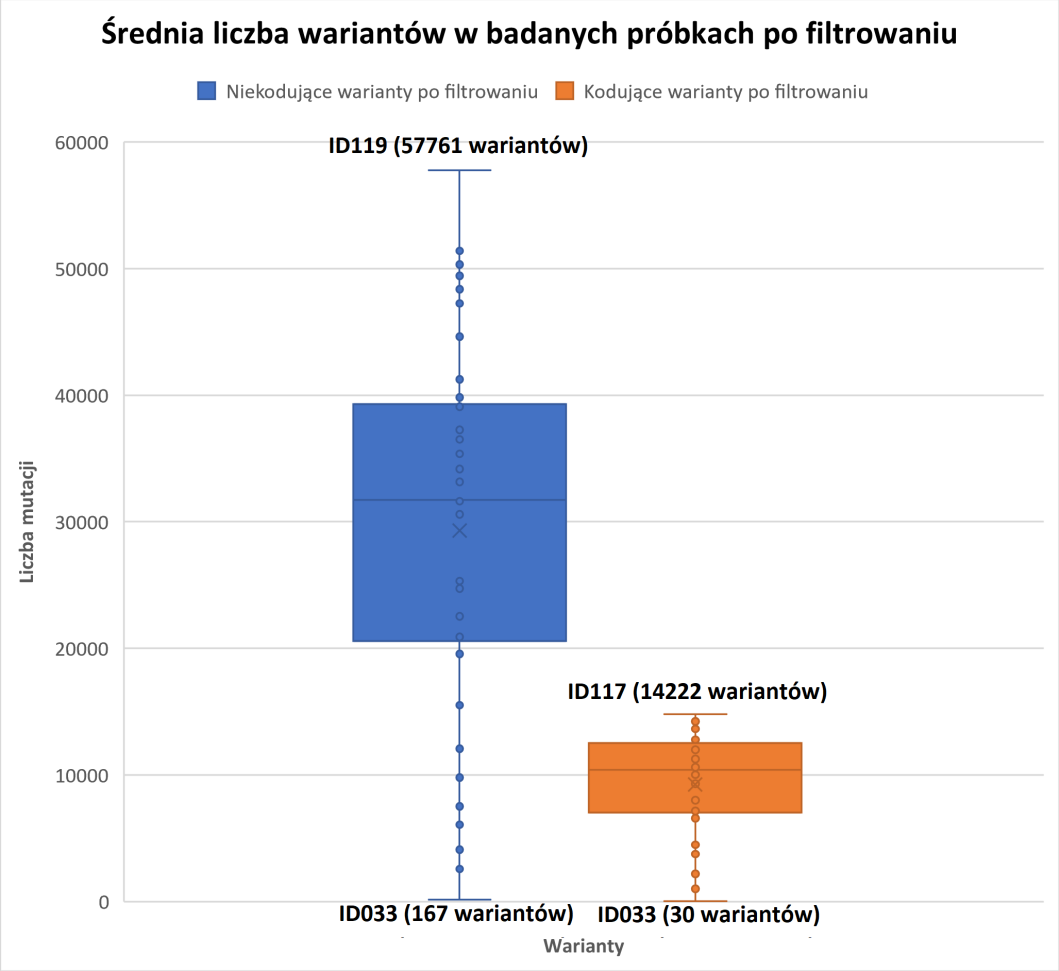
W przypadku regionów kodujących nie obserwowano tak dużej redukcji liczby wariantów po filtrowaniu. Wynika to z faktu, że dane pochodziły z sekwencjonowania eksomów, które skupia się na regionach kodujących, więc ich pokrycie było wyższe.

Poniższy wykres pokazuje jak wiele wariantów zostało odfiltrowanych ze względu na niską wiarygodność wyników sekwencjonowania.

Jak widać najniższe odstające wartości zarówno dla niekodujących jak i kodujących fragmentów należały do próbki *ID033*. Próbka ta zawierała najmniej wszystkich wariantów jeszcze przed filtrowaniem i po filtrowaniu liczba ta dodatkowo spadła.

W przypadku najwyższych wartości odstających, o ile przed filtrowaniem należały one do próbek *ID087* i *ID026*, tak po procesie filtrowania najwięcej wariantów ze fragmentach niekodujących miała próbka *ID119*, a we fragmentach kodujących próbka *ID117*. Oznacza to, że próbki *ID087* i

ID026 musiały zawierać dużą liczbę mutacji o niskim pokryciu, więc zostały one usunięte i liczba ogólna wariantów zmniejszyła się.



Rysunek 5.4: Średnia liczba mutacji w badanych próbkach po filtrowaniu

Różnice między liczbą mutacji we fragmentach niekodujących przed i po filtrowaniu plików przedstawiono w tabeli 5.3. Umieszczono w niej również informacje o liczbie substytucji, delecji i insercji znajdujących się w próbkach po filtrowaniu. Wynika z nich, że największą część mutacji stanowiły substytucje a najmniejszą insercje.

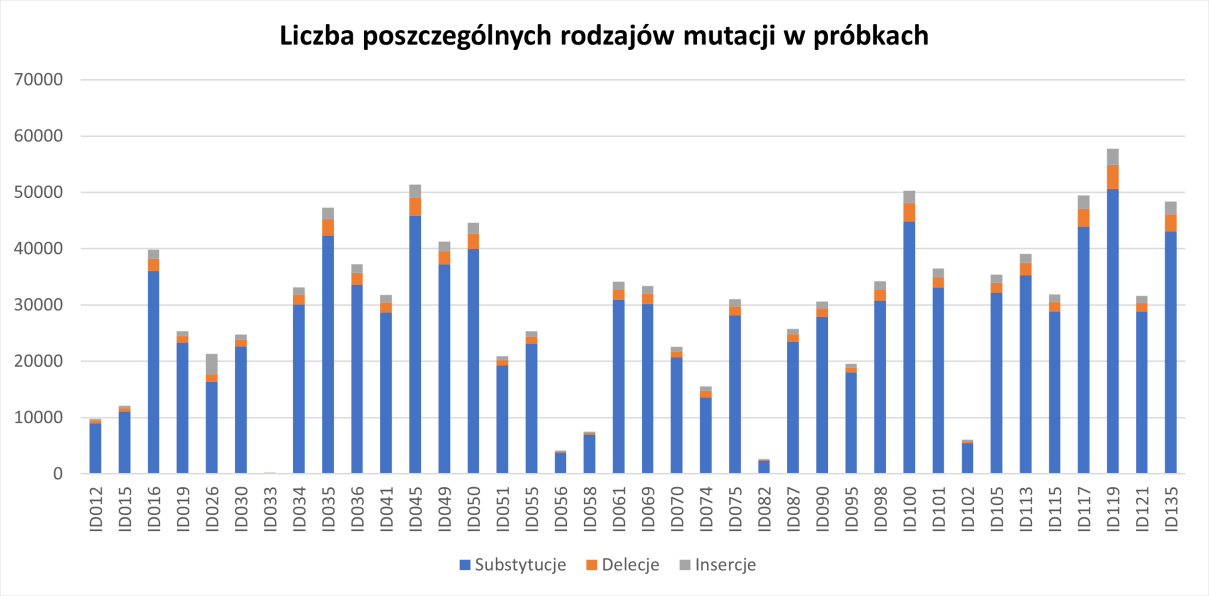
Substytucji było zdecydowanie więcej niż pozostałych typów mutacji w próbkach (87-94%). Insercje i delecje to zaledwie 6-13% próbek, przy czym insercji było nieco mniej niż delecji. Jedynym wyjątkiem była próbka *ID026*, która zawierała dużo więcej insercji (16%) niż pozostałe próbki.

Próbka	Wszystkie mutacje przed filtrowaniem	Mutacje po filtrowaniu	Substytucje	Delecje	Insercje
ID012	180 360	9 789	8 973 (91%)	492 (5%)	324 (3%)
ID015	206 594	12 062	11 116 (92%)	526 (4%)	420 (3%)
ID016	227 461	39 829	36 044 (90%)	2 196 (5%)	1 589 (3%)
ID019	330 188	25 334	23 311 (92%)	1 149 (4%)	874 (3%)
ID026	427 091	21 283	16 382 (76%)	1 294 (6%)	3 607 (16%)
ID030	196 507	24 719	22 664 (91%)	1 167 (4%)	888 (3%)
ID033	41 360	167	156 (93%)	6 (3%)	5 (2%)
ID034	196 704	33 153	30 094 (90%)	1 747 (5%)	1 312 (3%)
ID035	252 220	47 255	42 368 (89%)	2 875 (6%)	2 012 (4%)
ID036	410 943	37 265	33 627 (90%)	2 116 (5%)	1 522 (4%)
ID041	206 512	31 796	28 650 (90%)	1 825 (5%)	1 321 (4%)
ID045	239 903	51 403	45 892 (89%)	3 200 (6%)	2 311 (4%)
ID049	236 375	41 245	37 191 (90%)	2 351 (5%)	1 703 (4%)
ID050	242 199	44 639	39 983 (89%)	2 706 (6%)	1 950 (4%)
ID051	213 033	20 893	19 294 (92%)	911 (4%)	688 (3%)
ID055	188 797	25 317	23 157 (91%)	1 228 (4%)	932 (3%)
ID056	177 640	4 105	3 812 (92%)	142 (3%)	151 (3%)
ID058	268 729	7 524	7 004 (93%)	305 (4%)	215 (2%)
ID061	213 088	34 160	30 954 (90%)	1 850 (5%)	1 356 (3%)
ID069	204 845	33 368	30 201 (90%)	1 831 (5%)	1 336 (4%)
ID070	305 110	22 547	20 709 (91%)	1 061 (4%)	777 (3%)
ID074	237 803	15 511	13 646 (87%)	1 137 (7%)	728 (4%)
ID075	217 332	31 020	28 147 (90%)	1 647 (5%)	1 226 (3%)
ID082	129 799	2 574	2 421 (94%)	94 (3%)	59 (2%)
ID087	475 747	25 774	23 502 (91%)	1 317 (5%)	955 (3%)
ID090	192 731	30 579	27 963 (91%)	1 474 (4%)	1 142 (3%)
ID095	184 829	19 556	18 019 (92%)	882 (4%)	655 (3%)
ID098	293 077	34 179	30 789 (90%)	1 983 (5%)	1 407 (4%)
ID100	244 369	50 321	44 869 (89%)	3 250 (6%)	2 202 (4%)
ID101	201 879	36 507	33 139 (90%)	1 856 (5%)	1 512 (4%)
ID102	339 870	6 067	5 503 (90%)	336 (5%)	228 (3%)
ID105	210 355	35 375	32 188 (90%)	1 758 (4%)	1 429 (4%)
ID113	198 266	39 100	35 275 (90%)	2 216 (5%)	1 609 (4%)
ID115	206 344	31 844	28 866 (90%)	1 694 (5%)	1 284 (4%)
ID117	340 716	49 425	43 923 (88%)	3 230 (6%)	2 272 (4%)
ID119	270 321	57 761	50 637 (87%)	4 232 (7%)	2 892 (5%)
ID121	188 347	31 625	28 853 (91%)	1 507 (4%)	1 265 (4%)
ID135	234 505	48 372	43 094 (89%)	2 976 (6%)	2 302 (4%)

Tablica 5.3: Liczba poszczególnych mutacji w badanych próbkach po filtrowaniu

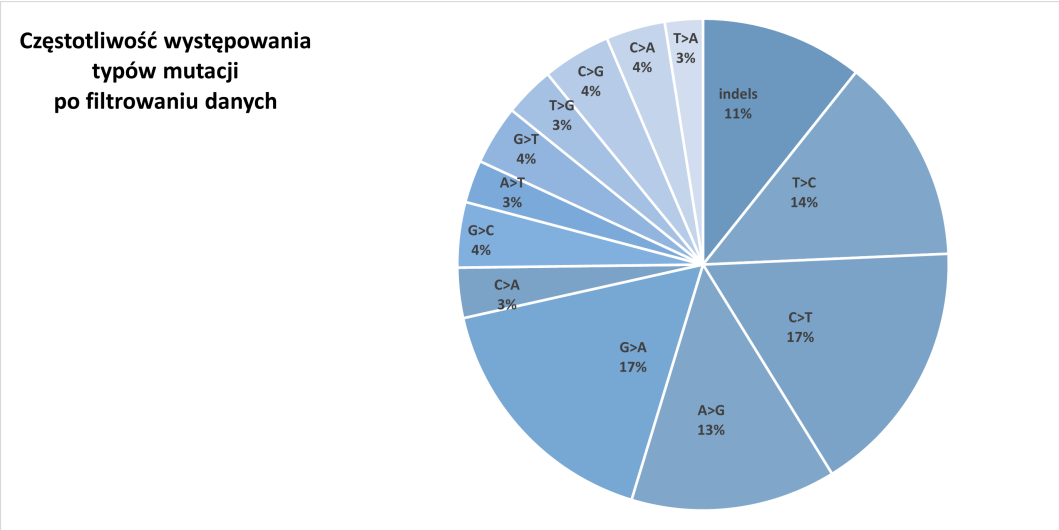
Wykres na Rys. 5.5 przedstawia jak rozkładały się rodzaje mutacji w poszczególnych próbkach po filtrowaniu. Insercje i delecje stanowiły 10% lub mniej składu próbek, przy czym delecji było nieco więcej.

Jak widać próbki *ID033*, *ID056*, *ID058*, *ID082* i *ID102* zawierały mniej niż 10 tys. wariantów po filtrowaniu.



Rysunek 5.5: Liczba rodzajów mutacji w poszczególnych próbkach po filtrowaniu

Stworzono również wykres przedstawiający częstotliwość występowania poszczególnych typów mutacji w próbkach po filtrowaniu (Rys. 5.6). Występowało tutaj więcej substytucji G>A oraz C>T. Oznacza to, że nie odfiltrowano tak dużo wariantów zawierających te substytucje, ponieważ miały one wystarczająco wysokie pokrycie.



Rysunek 5.6: Częstotliwość występowania poszczególnych typów mutacji w badanych próbkach po filtrowaniu danych

5.3 Porównanie mutacji występujących w poszczególnych próbkach

Po filtrowaniu, 38 analizowanych plików zawierało jedynie niekodujące warianty o wysokim pokryciu. Następny etap analizy polegał na porównaniu między sobą mutacji występujących w poszczególnych plikach. W tym celu użyto linii komend Linuxa.

Poniższa komenda scaliła wszystkie pliki *.tsv* zawierające warianty w jeden plik o nazwie *merged.tsv*.

```
cat *.tsv > merged.tsv
```

Następnie plik *merged.tsv* posortowano i wszelkie powtarzające się w nim wiersze zapisano w pliku *duplicates.txt*.

```
sort merged.tsv | uniq -d > duplicates.txt
```

Na koniec każdy wiersz w pliku *duplicates.txt* potraktowano jako osobny łańcuch znaków (*string*). Szukano we wszystkich plikach z danymi w formacie *.tsv* danego łańucha znaków. Jeśli łańcuch zgadzał się z wariantem w którymś innym pliku, zostawał on zapisany w pliku *check.txt* wraz z nazwą pliku, w którym został znaleziony.

```
grep -Fx -f duplicates.txt *.tsv > check.txt
```

Plik wynikowy zawierał 414 powtórzonych wariantów. Za pomocą skryptu w języku Python *repeated_mutations.py* stworzono na ich podstawie tabelę zawierającą informacje o numerze chromosomu i pozycji wariantu, allelu referencyjnym i alternatywanym, genie, liczbie powtórzeń wariantu oraz numerach próbek, w których dany wariant występował. Cała tabela została umieszczona w pliku *table_of_repeated_mutations.csv* znajdującym się w załącznikach do pracy. Zawierała ona 205 mutacji, z których 202 powtórzyły się 2 razy w zbiorze badanych próbek. Pozostałe 3 zostały przedstawione w tabeli 5.4 poniżej. Co ciekawe, najczęściej powtarzającymi się mutacjami były delecje.

Chr	Pozycja	Ref	Alt	Gen	Liczba powtórzeń	Numery próbek
11	33559992	CCACATACTA	-	KIAA1549L	3	016, 051, 069
17	21300363	G	-	MAP2K3	3	019, 055, 115
19	49025629	C	-	AC008687.1	4	026, 051, 070, 119

Tablica 5.4: Mutacje powtarzające się w badanych próbkach częściej niż 2 razy.

Skrypt *repeated_mutations.py* znajduje się w załącznikach do pracy.

5.4 Identyfikacja mutacji somatycznych

Niniejszy etap analizy polegał na porównaniu wariantów z próbek przed leczeniem, które zostały poddane filtrowaniu, z wariantami w próbkach po leczeniu. Miało to na celu zidentyfikowanie mutacji somatycznych, czyli mutacji nabywanych przez komórki nowotworowe, a które nie są obecne w prawidłowej tkance. Jeśli dany wariant znajdował się w próbce przed leczeniem, natomiast nie znaleziono go w próbce po leczeniu, oznaczało to, że jest to mutacja somatyczna.

Analizie poddano 7 par plików, zebranych dla 7 pacjentów z badanej grupy.

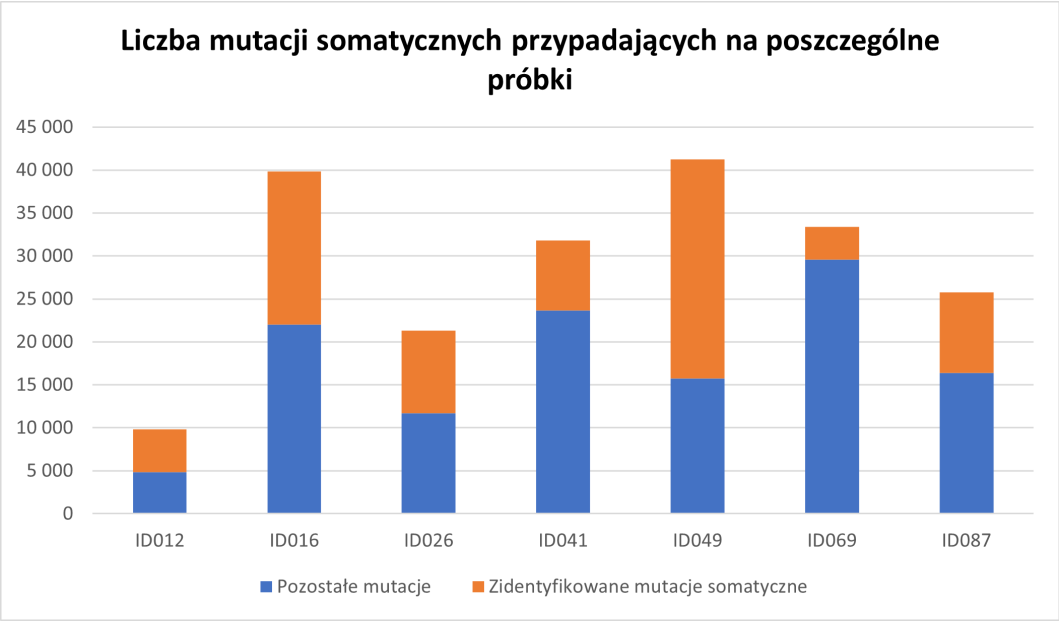
W celu identyfikacji wyżej opisanych wariantów stworzono skrypt *somatic_mutations.py* w języku Python. Pobierał on plik zawierający dane z próbki przed leczeniem po filtrowaniu w formacie *.tsv* i plik z danymi po leczeniu w formacie *.csv* (został zmieniony format z *.vcf* na *.csv*). Oba te pliki ładował w osobne listy, które następnie rozdzielił tak, aby zostały tylko wartości parametrów potrzebne do analizy, to jest *CHROM, START, REF* i *ALT*. Parametry te odpowiadały kolejno numerowi chromosomu, jego pozycji startowej, allelowi referencyjnemu i allelowi alternatywnemu. Następnie przechodził on po każdym jednym wierszu w liście z pliku przed leczeniu i sprawdzał czy wiersz ten odpowiada jakiemukolwiek wierszowi z listy wariantów po leczeniu. Jeśli tak, taki wiersz był ignorowany. Jeśli natomiast znalazł się wiersz, który był w próbce przed leczeniem, ale w próbce po leczeniu już nie, zostawał on zapisany do pliku tekstowego jako mutacja somatyczna. Skrypt znajduje się w załącznikach do pracy.

W tabeli 5.5 przedstawiono jaka część mutacji somatycznych przypadała na poszczególne badane próbki. Średni odsetek tych wariantów wyniósł 39%. W próbce *ID069* zidentyfikowano najmniej mutacji somatycznych (11%). Natomiast w przypadku próbki *ID049*, ten odsetek był najwyższy i stanowił 62%.

Próbka	Wszystkie mutacje niekodujące	Zidentyfikowane mutacje somatyczne
ID012	9789	4981 (51%)
ID016	39829	17835 (45%)
ID026	21283	9609 (45%)
ID041	31796	8164 (26%)
ID049	41245	25488 (62%)
ID069	33368	3799 (11%)
ID087	25774	9380 (36%)

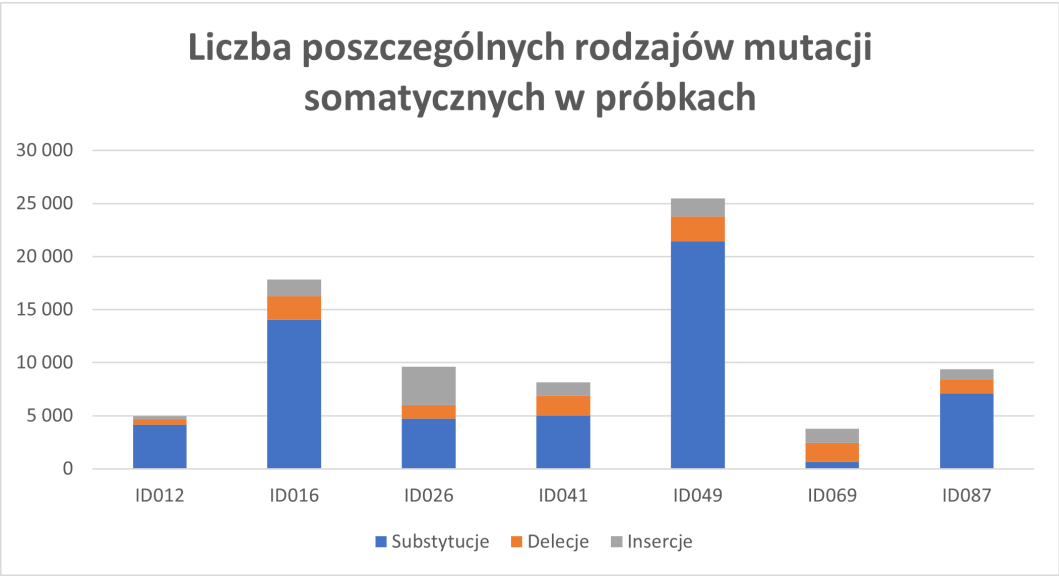
Tablica 5.5: Liczba zidentyfikowanych mutacji somatycznych przypadająca na poszczególne próbki

Poniżej znajdują się 2 wykresy obrazujące liczbę zidentyfikowanych mutacji somatycznych w 7 badanych próbkach. Pierwszy z nich (Rys. 5.7) pokazuje jak dużą część mutacji stanowiły wyżej wspomniane warianty.



Rysunek 5.7: Liczba mutacji somatycznych przypadająca na poszczególne próbki

Wykres na Rys. 5.8 przedstawia z kolei rodzaje mutacji somatycznych i ich liczbę w poszczególnych próbkach. W porównaniu z wykresem przedstawiającym wszystkie substytucje, delecje i insercje (Rys. 5.5), wśród mutacji somatycznych występuje większy udział insercji i delecji, szczególnie w próbkach *ID069*, *ID026* i *ID041*. Z próbki *ID069* wynika, że większość znajdujących się w niej substytucji było germinalnych. W przypadku pozostałych 4 próbek największy odsetek stanowiły substytucje.



Rysunek 5.8: Liczba rodzajów mutacji somatycznych w badanych próbkach

5.5 Porównanie mutacji zidentyfikowanych w badanym zbiorze danych z mutacjami w bazie danych *TCGA*

Przedostatnią częścią tego rozdziału było porównanie danych pochodzących z wcześniejszych etapów analizy z tymi otrzymanymi z bazy *TCGA*. Na potrzeby niniejszego etapu stworzono dwa skrypty: *tcga_data_comparing.py* oraz *counting_mutations.py*. Oba skrypty znajdują się w załącznikach do pracy.

Pierwszy z nich pobierał pliki wyjściowe z wcześniejszych analiz, to jest plik zawierający wszystkie znalezione w regionach niekodujących mutacje somatyczne z 7 próbek i plik zawierający wszystkie mutacje zidentyfikowane w regionach niekodujących z 38 próbek z danymi przed leczeniem. Dodatkowo ładował również plik otrzymany z bazy *TCGA* - wszystkie somatyczne mutacje w regionach niekodujących występujące w 162 próbkach. Z powyższych 3 plików, skrypt stworzył 3 osobne tablice zawierające informacje o numerze chromosomu i pozycji wariantu oraz allelu referencyjnym i alternatywnym. Następnie porównał on tablicę wariantów z bazy danych *TCGA* z pozostałymi dwoma tablicami. Jeśli warianty powtarzały się w tablicach, były zapisywane do osobnych plików. W ten sposób powstały 2 nowe pliki wyjściowe: plik zawierający warianty powtarzające się w bazie *TCGA* i w znalezionych 7 próbkach somatycznych mutacji, oraz plik z wariantami wspólnymi dla tych z bazy *TCGA* oraz 38 badanych próbkami mutacjami w niekodujących regionach.

Finalne pliki zawierały dużo wariantów powtarzających się, dlatego też drugi ze stworzonych skryptów pobierał oba te pliki, wyliczał dla każdego liczbę powtarzających się mutacji, zapisywał je i sam wariant do nowego pliku w kolejności od największej liczby powtórzeń. Oba te pliki pod nazwami *tcga_merged_somatic_counts.txt* i *tcga_merged_all_counts.txt* znajdują się w załącznikach do pracy.

Tabela 5.6 przedstawia ile wspólnych wariantów znaleziono dzięki zastosowaniu powyższych skryptów. Najpierw porównywano tylko mutacje somatyczne, których było zdecydowanie mniej niż mutacji ze wszystkich 38 próbek. Znaleziono ponad 3 tys. mutacji wspólnych dla zbioru mutacji somatycznych w badanych 7 próbach oraz dla bazy danych *TCGA*. Po usunięciu powtózeń zidentyfikowane 2 208 unikatowych wariantów somatycznych. Największa wartość liczby powtórzeń wariantu wyniosła 6. Zbiór mutacji zidentyfikowanych we wszystkich 38 próbkach zawierał ponad milion wariantów w niekodujących fragmentach. Wspólnych wariantów z powtórzeniami dla tego zbioru i *TCGA* znaleziono prawie 49 tys. Było to ponad 16 razy więcej niż przy mutacjach somatycznych. Unikatowych wariantów natomiast pozostało tylko 6 979. Wynika z tego, że wśród tych 49 tys. wariantów zdecydowana większość powtórzyła się wiele razy. Najwyższa wartość liczby powtórzeń danego wariantu wynosiła 38.

Zbiór danych wejściowych	Zidentyfikowane mutacje w badanym zbiorze danych	Mutacje somatyczne z bazy danych TCGA	Mutacje wspólne dla badanego zbioru oraz TCGA (z powtórzeniami)	Unikatowe warianty wspólne dla badanego zbioru danych oraz TCGA
Mutacje somatyczne z 7 próbek	79 250	112 045	3 243	2 208
Wszystkie mutacje w regionach niekodujących 38 próbek	1 113 480		48 889	6 979

Tablica 5.6: Tabela przedstawiająca liczbę mutacji w analizowanych plikach

Oba zbiory danych porównywane były z somatycznymi mutacjami z bazy *TCGA*. Wynika z tego, że wspólne mutacje, które znaleziono zarówno dla zbioru wariantów somatycznych z 7 próbek, jak i dla wszystkich wariantów z 38 próbek są najprawdopodobniej mutacjami somatycznymi. Należy pamiętać, że dysponowano tylko 7 próbkami, dla których można było zidentyfikować mutacje somatyczne. Pozostała część próbek nie miała odpowiedników dzięki którym można byłoby odrzu-

cić mutacje germinalne. Dlatego też znajdujące się wśród nich mutacje somatyczne, mogły zostać z dużą dozą prawdopodobieństwa wyłonione na podstawie porównania z wariantami somatycznymi z bazy danych *TCGA*.

5.6 Wykorzystanie narzędzia *miRMut* do adnotacji mutacji

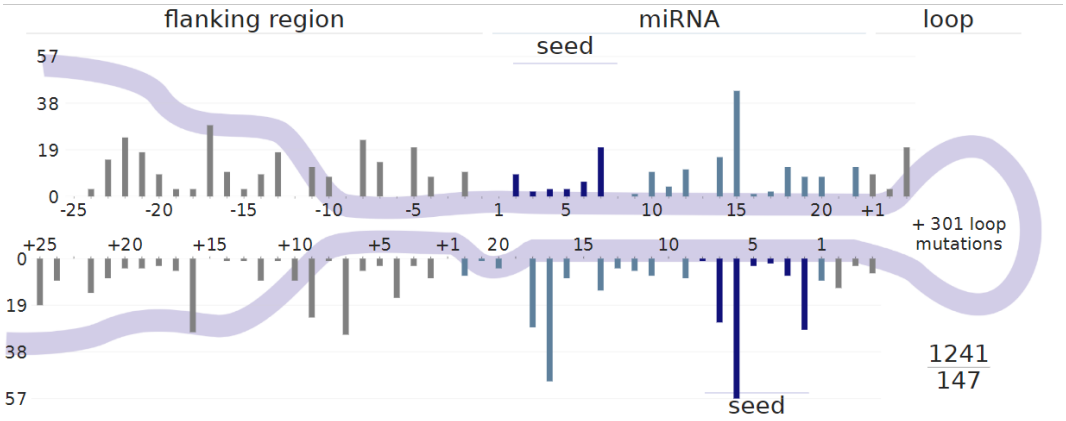
Jako dodatek do części związanej z analizą wykorzystano narzędzie *miRMut* do adnotacji 36 plików w formacie *.vcf* zawierających dane z próbek przed leczeniem.

Wspomniane pliki trzeba było najpierw odpowiednio odfiltrować. Jako, że *miRMut* nie obsługuje wariantów multiallelicznych (tj. wariantów, w których allel alternatywny występuje w więcej niż jednej postaci), przed użyciem go usunięto wszelkie warianty multialleliczne. Każdy z plików zawierał ich niewielką liczbę, dodatkowo były to warianty o zasadniczo niskim pokryciu, dlatego ich brak nie wpływał znacząco na wyniki analizy. Drugim krokiem filtrowania było usunięcie wszystkich wariantów, których pokrycie było mniejsze lub równe 20.

Po przygotowaniu plików uruchomiono *miRMuta* z domyślnymi ustawieniami. Analiza składała się z 6 kroków, z czego 5 pierwszych generowało pliki wynikowe w formacie *.csv*:

- stworzenie plików podsumowujących informacje zawarte w plikach *.vcf*, zawierających między innymi nazwy i identyfikatory pojedynczych próbek, nazwy próbek nowotworowych/normalnych i identyfikatory oraz typy algorytmów użytych do wykrywania mutacji
- sprawdzenie filtrowania plików oraz stworzenie pliku zawierającego listę mutacji opisanych przez chromosom (*chrom*), pozycję nukleotydu (*pos*), allel referencyjny (*ref*), allel alternatywny (*alt*) itd.
- uzupełnienie pliku z poprzedniego podpunktu o informacje o adnotacjach
- generowanie plików z uzyskanymi danymi i dodanie do nich informacji o motywach i wagach, genomowego oznaczenia *HGVS* oraz oznaczenia *RBase HGVS*
- stworzenie pliku, który który sumuje liczbę mutacji znalezionych w każdym genie *miRNA* w analizowanej kohorcie, przy czym dla każdego genu *miRNA* podawana jest liczba unikatowych próbek z co najmniej jedną mutacją, liczba wszystkich mutacji oraz liczba unikatowych pozycji mutacji.

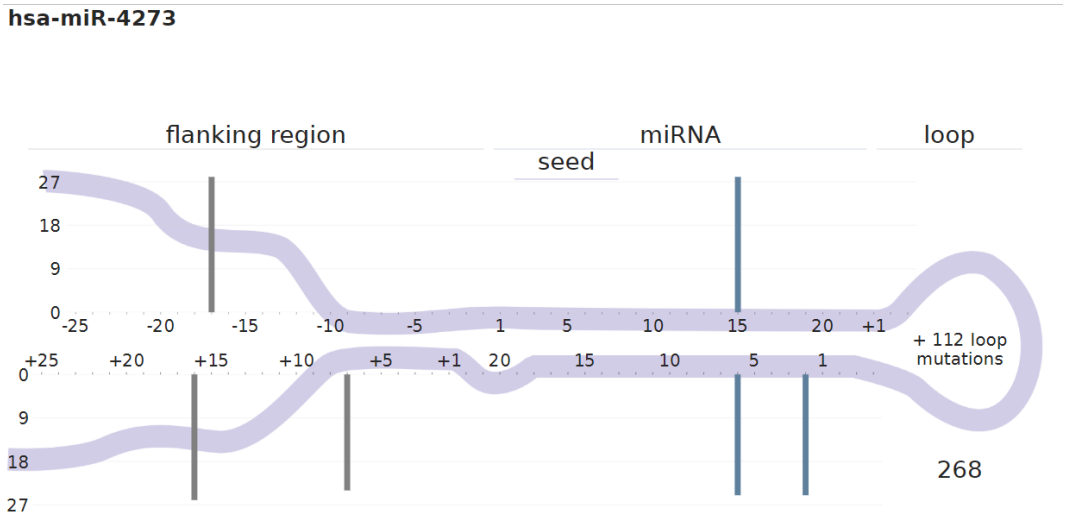
W celu wizualizacji mutacji i umożliwienia interpretacji wariantów w kontekście struktury prekursora *miRNA*, w kroku 6 nałożono zidentyfikowane warianty na konsensusową strukturę prekursora *miRNA* i skategoryzowano je według lokalizacji w podregionach genów *miRNA*. Wszystkie znalezione warianty przedstawia wykres na Rys. 5.9.



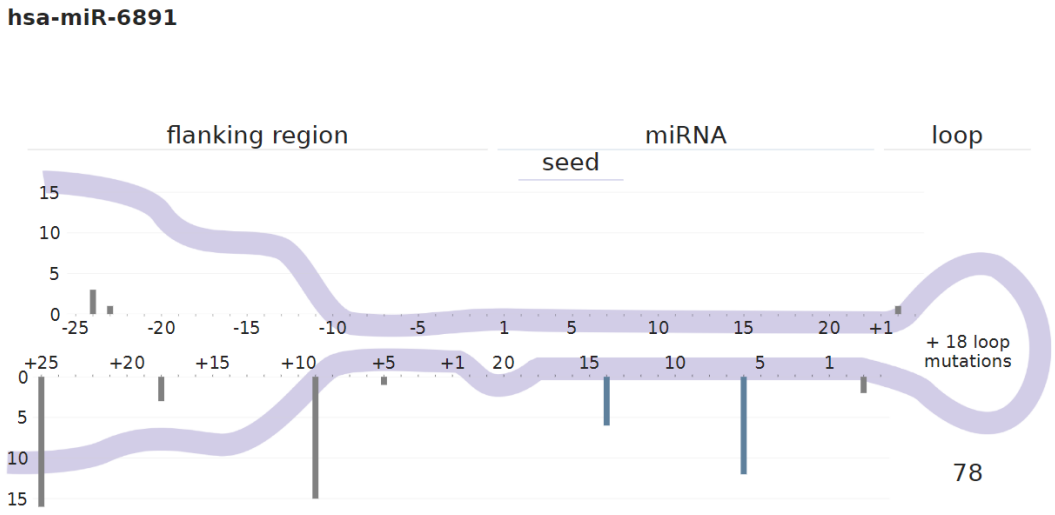
Rysunek 5.9: Lokalizacja wszystkich znalezionych mutacji w strukturze konsensusowej prekursora *miRNA* za pomocą narzędzia *miRMut*

Kolorem niebieskim zaznaczone są pozycje mutacji w regionie dwuniciowym trzonu spinki *miRNA*, natomiast ciemnoniebieskim regiony konserwatywne nici *RNA* (ang. *seed*). Na szaro zaznaczono mutacje otaczające *miRNA* (ang. *flanking regions*) i końcowe pozycje pętli apikalnej. Liczby w prawym dolnym rogu oznaczają liczbę znalezionych mutacji (górna liczba) i liczbę zmutowanych genów *miRNA* (dolna liczba). Wykres przedstawia mutacje tylko w obrębie sześciu pozycji pętli (pierwsze 3 i ostatnie 3 nukleotydy). Liczba pozostałych mutacji pętli (ang. *loop*) jest wypisana w obrębie pętli.

Z wykresu na Rys. 5.9 wynika, że liczba wszystkich znalezionych mutacji w geanch *miRNA* wyniosła 1241. Krok 6 procesu działania *miRMuta* generuje również podobne wykresy dla każdego ze zmutowanych genów *miRNA*. W przypadku niniejszej analizy było to 147 wykresów. Poniżej znajdują się 2 z nich. Przedstawiają one mutacje znalezione w *miR-4273* (Rys. 5.10) oraz w *miR-6891* (Rys. 5.11) w badanym zbiorze danych. Większość wygenerowanych wizualizacji zawierała małą liczbę mutacji w obrębie pojedynczego genu *miRNA* (1-12). Wspomniane wyżej zmutowane geny *miRNA* były jednymi z niewielu, u których wystąpiło więcej mutacji. W przypadku genu *miR-4273* było to 268 mutacji, a w przypadku *miR-6891* - 78.



Rysunek 5.10: Mutacje znalezione w obrębie *miR-4273* w badanym zbiorze danych



Rysunek 5.11: Mutacje znalezione w obrębie *miR-6891* w badanym zbiorze danych

Tabela 5.7 przedstawia identyfikatory zmutowanych genów *miRNA* wraz z liczbą mutacji występujących w obrębie każdego z nich, posortowana od najwyższej wartości. Znajduje się ona również w załącznikach do pracy pod nazwą *miRMut_gene_mutation_table.xls*.

Tablica 5.7: Zmutowane geny *miRNA* wraz z liczbą mutacji w nich występujących

Identyfikator pojedynczego genu miRNA	Liczba mutacji w obrębie genu	Identyfikator pojedynczego genu miRNA	Liczba mutacji w obrębie genu
4273	268	296	3
6891	78	411	3
10527	63	920	3
7705	56	4741	3
6859-4	48	4745	3
1234	22	4800	3
6796	21	6717	3
6751	18	6777	3
27a	17	6797	3
4632	17	1283-2	3
943	16	6859-3	3
412	15	122	2
320e	15	182	2
4751	14	202	2

6892	14	564	2
6859-1	14	658	2
3652	13	662	2
4761	13	944	2
6795	12	1229	2
7844	12	3940	2
550a-3	12	5090	2
6821	11	6776	2
7108	11	6858	2
196a-2	11	6885	2
300	10	7155	2
604	10	122b	2
608	10	194-1	2
6794	10	449b	2
10394	10	509-2	2
423	9	518d	2
618	9	520h	2
938	9	663a	2
3184	9	941-5	2
6499	9	21	1
6730	9	184	1
6811	9	346	1
6886	9	573	1
4315-1	9	577	1
100	8	589	1
126	8	641	1
1307	8	648	1
6084	8	890	1
6823	8	1248	1
6868	8	1270	1

6887	8	1539	1
612	7	3615	1
7107	7	4274	1
6769b	7	4653	1
152	6	4707	1
567	6	4742	1
647	6	6514	1
2110	6	6721	1
3620	6	6744	1
6505	6	6784	1
6801	6	6810	1
6815	6	6826	1
6836	6	6850	1
7162	6	6864	1
196a-1	6	6890	1
492	5	8072	1
595	5	8078	1
596	5	10226	1
624	5	10392	1
3196	5	10399	1
4722	5	146a	1
6726	5	199a-1	1
631	4	302c	1
6763	4	499a	1
6871	4	499b	1
320c-2	4	520f	1
516b-2	4	520g	1
941-1	4	6859-2	1
149	3	941-2	1
187	3		

5.7 Wnioski

Z przeprowadzonej analizy wynika, że badane próbki zawierały mnóstwo słabych jakościowo wariantów (o niskim pokryciu), które należało usunąć na wstępie. Dopiero po odfiltrowaniu wariantów o niskiej jakości, podzielono dane na dwa zbiory - jeden zawierający mutacje niekodujące, oraz drugi, w którym znajdowały się jedynie warianty w regionach kodujących białka. Jako, że próbki poddano sekwencjonowaniu eksomowemu (*WES*), to faworyzowane były warianty kodujące, a mutacje białka niekodującego, należały przede wszystkim do tych znajdujących się blisko regionów kodujących. Z tego też powodu wariaty w regionach kodujących miały wyższe pokrycie.

Identyfikacja mutacji w regionach niekodujących wykazała, że największą ich część stanowiły substytucje, a insercje i delecje zajmowały jedynie około 10%. Zarówno dla danych przed jak i po filtrowaniu, najwięcej występowało substytucji G>A i C>T oraz ich wersji komplementarnych.

Kilka próbek utraciło dużo wariantów w wyniku filtrowania. Wśród nich wyróżniła się próbka *ID033*, dla której liczba wariantów znacznie spadła w porównaniu z resztą. Dla tej próbki uzyskano mniej danych z sekwencjonowania niż dla pozostałych próbek, stąd też spora część wariantów nie spełniała parametrów jakościowych (zbyt niskie pokrycie). Nie oznacza to, że ten pacjent miał mniej mutacji, tylko że dla tej próbki eksperyment się nie udał, a nie zostało już materiału na jego powtórkę.

Podczas porównania mutacji występujących w poszczególnych próbkach, znaleziono powtarzające się warianty, jednak większość z nich stanowiła jedynie duplikaty (była obecna tylko w dwóch próbkach). Mutacje tworzą się losowo, więc powtórzenie się dwukrotnie jednego wariantu mogło wynikać z przypadku. Należy również zaznaczyć, iż badano stosunkowo niewielki zbiór danych. Prawdopodobnym jest, że w większych kohortach powtórzeń wariantów byłoby o wiele więcej.

W przypadku analizy mutacji somatycznych, należy zaznaczyć, że brano pod uwagę mały zbiór danych - jedynie 7 próbek, dla których posiadano odpowiadające im próbki zawierające dane po leczeniu tych samych pacjentów. Ta część analizy była więc ograniczona do małej ilości danych.

Wyniki pokazały, że średnio 39% wszystkich mutacji stanowiły mutacje somatyczne. Jednak ta wartość wynikała z dużej różnicy między próbką z najmniejszą liczbą mutacji somatycznych (*ID069* - 11%), a próbką z największą jej liczbą (*ID049* - 62%). Próbkę *ID069* wyróżnił dodatkowo fakt posiadania znikomej liczby substytucji wśród zidentyfikowanych mutacji somatycznych. Wynika z tego, że większość znajdujących się w niej substytucji stanowiły mutacje germinalne, czyli dziedziczne mutacje, a nie te specyficzne dla nowotworu. Z analizy rodzajów mutacji somatycznych wynika, że w prawie połowie próbek wiele substytucji zostało uznanych za mutacje germinalne, podczas gdy to indele zostały zidentyfikowane jako mutacje somatyczne.

Podczas analizy skorzystano z większego zasobu danych, jakim był zbiór 162 plików z adnotacją, pochodzący z projektu *TCGA LAML* znajdującego się w bazie TCGA, ponieważ analizowany zbiór danych nie był liczny. W wyniku porównań otrzymano pliki zawierające mutacje powtarzające się w próbkach oraz zbiór unikatowych mutacji wraz z liczbą ich powtórzeń. Porównywano dwa zbiory osobno z wariantami z bazy danych *TCGA*. Jeden z badanych zbiorów zawierał jedynie warianty z 7 próbek, a w drugim znajdowały się warianty ze wszystkich 38 próbek. Analiza wykazała, że ze zidentyfikowanych mutacji somatycznych ponad 2 tys. było wspólnych ze zbiorem z bazy *TCGA* i powtarzało się najwięcej 8 razy. Natomiast w przypadku wszystkich mutacji w niekodujących regionach, znaleziono ponad 6 tys. powtarzających się wariantów a liczba ich powtórzeń wynosiła najwięcej 38 raza. Wynika z tego, że zwiększenie liczby analizowanych próbek powoduje wzrost

liczby identyfikowanych mutacji, ale nie jest wzrost proporcjonalny do wzrostu liczby próbek . Zwiększa się natomiast liczba mutacji powtarzających się w całym zbiorze.

Jako, że mutacje powstają przypadkowo, fakt że znaleziono taki sam wariant zarówno wśród danych z bazy *TCGA* jak i w wynikach z analizowanego w niniejszej pracy zbioru, może wykazywać na związek tego wariantu z nowotworem, ale może też wynikać z losowości. Zwłaszcza, że wspólne mutacje somatyczne dla 38 próbek i dla 162 próbek z projektu *TCGA LAML* stanowiły jedynie 6% wszystkich somatycznych wariantów z bazy *TCGA*.

Dodatkowym elementem pracy było wykorzystanie narzędzia *miRMut* do adnotacji mutacji znajdujących się w genach *miRNA* dla 36 próbek przed leczeniem. Narzędzie to przeprowadziło proces adnotowania w 6 krokach, z których ostatni wygenerował wykresy pokazujące warianty zidentyfikowane w *miRNA*. Wynika z nich, że znaleziono 1 241 takich mutacji oraz 147 zmutowanych genów *miRNA*.

Podsumowując, analiza wykonana w niniejszej pracy wykazała, że mimo że dane wejściowe pochodziły z sekwencjonowania eksomów, zidentyfikowano z nich liczne mutacje znajdujące się w regionach niekodujących, w tym w genach kodujących *miRNA*.

Tylko niewielki odsetek mutacji w regionach niekodujących powtarzał się w dwóch lub więcej próbkach, co oznacza że większość z nich powstaje przypadkowo i prawdopodobnie nie ma bezpośredniego związku z transformacją nowotworową (są to tak zwane mutacje pasażerskie).

W zidentyfikowanym małym zbiorze mutacji powtarzających się, w wielu próbkach można poszukiwać potencjalnych mutacji sprawczych (ang. *drivers*), ale do tego celu potrzebne jest przeprowadzenie dodatkowych badań biologicznych.

Rozdział 6

Zakończenie

Celem pracy była identyfikacja mutacji w niekodujących fragmentach genomu u pacjentów z ostrą białaczką szpikową. Aby to zrobić odfiltrowano najpierw wszystkie warianty, których jakość była niewystarczająca. Następnie w plikach zostawiono jedynie warianty we fragmentach niekodujących, a resztę mutacji zapisano w osobnych plikach. Potem porównano warianty niekodujące między próbkami aby sprawdzić, czy pacjenci posiadali wspólne mutacje. Następnie zidentyfikowano warianty somatyczne poprzez porównanie plików z próbkami przed leczeniem z plikami zawierającymi dane z próbek po leczeniu. Na koniec analizy porównano wyniki z wcześniejszych jej etapów z danymi otrzymanymi z bazy *TCGA*.

Zaplanowana analiza powiodła się i zostały znalezione liczne mutacje w niekodujących fragmentach genomów u pacjentów z ostrą białaczką szpikową. Mutacje somatyczne stanowiły część wszystkich mutacji, jednak nadal była to dość niewielka liczba. Analiza wykazała również, że większość mutacji nie powtarza się między próbkami. Warianty pacjentów były różne i rzadko udawało się znaleźć te powtarzające się 2 lub więcej razy. Potwierdza to, że mutacje pojawiają się losowo. Należy pamiętać, że analizę ograniczała mała liczba próbek, zwłaszcza tych po leczeniu pacjentów. Porównanie z większym zbiorem danych, jakim były wyniki z projektu *TCGA LAML* wykazało, że warianty powtarzały się między próbkami. Był to jednak mały odsetek mutacji dla całej kohorty.

Niniejsza praca skupiała się na identyfikacji wszystkich mutacji w niekodujących regionach, a włączenie do analizy danych z bazy *TCGA* pozwoliło wytypować te mutacje, które były najprawdopodobniej somatyczne. Uzyskane w pracy wyniki mogą stać się punktem do dalszych analiz, zakładających dodatkowe eksperymenty biologiczne i skupiające się na konkretnych wariantach.

Literatura

[1] Ajaz A. Bhat, Salma N. Younes, Syed Shadab Raza, Lubna Zarif, Sabah Nisar, Ikhlak Ahmed, Rashid Mir, Sachin Kumar, Surender K, Sharawat, Sheema Hashem, Imadeldin Elfaki, Michal Kulinski, Shilpa Kuttikrishnan, Kirti S. Prabhu, Abdul Q. Khan, Santosh K. Yadav, Wael El-Rifai, Mohammad A. Zargar, Hatem Zayed, Mohammad Haris, and Shahab Uddin. Role of non-coding rna networks in leukemia progression, metastasis and drug resistance. *biomedcentral.com*, 2020.

[2] Felix Dietlein, Alex B. Wang, Christian Fagre, Anran Tang, Nicolle J. M. Besselink, Edwin Cuppen, Chunliang Li, Shamil R. Sunyaev, James T. Neal, and Eliezer M. Van Allen. Genome-wide analysis of somatic noncoding mutation patterns in cancer. *www.science.org*, 2022.

[3] hematoonkologia.pl. Białaczki, co o nich powinniśmy wiedzieć. [on-line] `https://hematoonkologia.pl/informacje-dla-chorych/news/id/3139-bialaczki-co-o-nich-powinnismy-wiedziec`.

[4] hematoonkologia.pl. Rozpoznanie i leczenie nowotworów układu krwiotwórczego. [on-line] `https://hematoonkologia.pl/informacje-dla-chorych/news/id/3003-rozpoznanie-i-leczenie-nowotworow-ukladu-krwiotworczego`.

[5] https://bdbioscences.com. Leukemia. [on-line] `https://www.bdbiosciences.com/zh-cn/learn/clinical/blood-cancers/leukemia`.

[6] https://medlineplus.gov. What are whole exome sequencing and whole genome sequencing? [on-line] `https://medlineplus.gov/genetics/understanding/testing/sequencing/`.

[7] https://medlineplus.gov. What is noncoding dna? [on-line] `https://medlineplus.gov/genetics/understanding/basics/noncodingdna/`.

[8] https://www.cancercenter.com. Acute myeloid leukemia. [on-line] `https://www.cancercenter.com/cancer-types/leukemia/types/acute-myeloid-leukemia`.

[9] onkologia.pl. Białaczki. [on-line] `http://onkologia.org.pl/bialaczki/`.

[10] Adrián Mosquera Orgueira, Beatriz Rodríguez Antelo, José Ángel Díaz Arias, Nicolás Díaz Varela, Natalia Alonso Vence, Marta Sonia González Pérez, and José Luis Bello López. Novel mutation hotspots within non-coding regulatory regions of the chronic lymphocytic leukemia genome. *nature.com*, 2020.

[11] Minal B. Patel and Jun Wang. The identification and interpretation of cis-regulatory noncoding mutations in cancer. *mdpi.com*, 2018.

[12] Sunniyat Rahman and Marc R Mansour. The role of noncoding mutations in blood cancers. *pubmedcentral.com*, 2019.

[13] Martyna O Urbanek-Trzeciak, Piotr Kozlowski, and Paulina Galka-Marciniak. mirmut: Annotation of mutations in mirna genes from human whole-exome or whole-genome sequencing. *STAR protocols*, 2022.

[14] www.cancer.gov. Tcga. [on-line] `https://www.cancer.gov/about-nci/organization/ccg/research/structural-genomics/tcga`.

Instytut Informatyki, Wydział Informatyki i Telekomunikacji
Politechnika Poznańska

Skład przy użyciu systemu LaTeX na platformie Overleaf.