Uniwersytet Przyrodniczy we Wrocławiu
Wydział Biologii i Hodowli Zwierząt
Kierunek: Bioinformatyka
Studia stacjonarne pierwszego stopnia

Dawid Sikorski
110944

Analiza zmienności genetycznej człowieka na podstawie danych z sekwencjonowania nowej generacji

Analysis of human genetic variability based on data from next-generation sequencing

Praca wykonana pod kierunkiem
dr. Tomasz Suchocki
Katedra Genetyki

Wrocław, 2019

Oświadczenie opiekuna pracy

Oświadczam, że niniejsza praca została przygotowana pod moim kierownictwem i stwierdzam, że spełnia ona warunki do przedstawienia jej w postępowaniu o nadanie tytułu zawodowego.

Data Podpis opiekuna pracy

Oświadczenie autora pracy

Oświadczam, że niniejsza praca dyplomowa została napisana przeze mnie samodzielnie i nie zawiera treści uzyskanych w sposób niezgodny z obowiązującymi przepisami ani też nie była wcześniej przedmiotem procedur związanych z uzyskaniem tytułu zawodowego w wyższej uczelni.

Data Podpis opiekuna pracy

Spis treści

1 Wstęp

Zmienność genetyczna jest podstawowym mechanizmem ewolucyjnym wszystkich organizmów żywych. Efektem takiej zmienności są różnice w budowie białek, przez co ich funkcji a także czasu i miejscu ich powstania, co może prowadzić do powstania różnic fenotypowych. Wiele z takich cech nie ma jednak znaczenia dla przeżycia organizmu. Rzadko takie zmiany mogą faworyzować dany organizm w środowisku, co dalej prowadzi to rozpowszechnienia cechy a w efekcie do powstania nowego gatunku, który lepiej przystosował się do panujących warunków środowiska.
Mimo to większość takich zmian prowadzi do powstania różnych chorób. Badanie zmienności genetycznej człowieka, może dać szansę na poznanie przyczyny ich występowania. Przez badanie różnic w kodzie genetycznym, zauważa się pewne wzory występujących mutacji a pojawiania się pewnych objawów chorobowych. Dzięki takiej wiedzy, przy zastosowaniu technik molekularnego badania genomu, można wykryć mutację która zaszła w DNA i podjąć odpowiednie kroki w celu leczenia objawów danej przypadłości.

1.1 Rodzaje mutacji i ich konsekwencje

Poruszane w tej pracy mutacje to:

- SNP(ang. Single Nucleotide Polymorphism) - jest to mutacja polegająca na substytucji jednego nukleotyda. SNP stanowią dużą część całkowitej zmienności genetycznej.
- Insercja - polega na wstawieniu krótkiego fragmentu nukleotydów lub przynajmniej jednego nukleotydu.
- Delecja - antagonistyczna zmiana do insercji, polegająca na usunięciu conajmniej jednego nukleotydu,

Większość zaobserwowanych mutacji nie wywiera żadnego wpływu, ponieważ te negatywne najczęściej są letalne dla organizmu w związku z czym nie występują często. W pracy zwrócono uwagę na 28 różnych konsekwencji mutacji, które zostaną omówione poniżej.

1.1.1 Mutacje związane z ramką odczytu

Inframe deletion i Inframe insertion nie powoduje frameshifta ale zmienia bialko

Zmiana ramki odczytu (ang. Frameshift variant) Są to mutacje typu delecja lub insercja powodujące zaburzenia w procesie translacji. W zależności od miejsca zajścia zamiany konsekwencją może być całkowita zmiana łańcucha polipeptydowego lub jego części. Wywierają one bardzo duży wpływ na białko, ponieważ taki produkt najczęściej nie jest w stanie pełnić przeznaczonej funkcji przez zmianę jego budowy lub długości, w konsekwencji organizm nie jest w stanie produkować białka lub jego ilość jest niewystarczająca co ma wpływ na różne szlaki metaboliczne i życie samego organizmu. Jednakże aby do tego doszło zmiana nie może dotyczyć trzech lub wielokrotności tej liczby, gdyż taka zmiana uwzględniając cechę kodu genetycznego jaką jest trójkowość, czyli tworzenie przez trzy nukleotydy kodonu, nie spowoduje zmiany ramki odczytu a jedynie w białku pojawi się nieprawidłowa liczba aminokwasów, przy czym takie białko może zachować swoje właściwości i funkcje.

1.1.2 Mutacje missensowne done

- Mutacja zmiany sensu(ang. missense variant)
  Zmiana mogąca prowadzić do zmiany aminokwasu w syntetyzowanym białku. W zależności od substytucji mutacja może nie powodować zmiany funkcji i właściwości posiadanych przez białko (tzw. mutacja konserwatywna), ponieważ nowy aminokwas posiada zbliżone właściwości lub prowadzić do dysfunkcji białka(tzw. mutacja niekonserwatywna)w konsekwencji powstania chorób.

- Niekompletny kodon terminacyjny (ang. incomplete terminal codon varaint)
  Wariant sekwencji. w którym zmienia się co najmniej jedną zasadę końcowego kodonu niekompletnie opisanego.

1.1.3 regulatory region variant done?

Większość wariantów związanych z fenotypem człowieka znajduje się w intronach lub

regionach między genowych, a więc muszą mieć wpływ na regulacje ekspresji genów. W skład takich sekwencji wchodzą wzmacniacze, wyciszacze.... co jeszcze

1.1.4 Mutacje związane ze splicingiem

Splicing, czyli składanie genu jest ważnym procesem zachodzącym podczas dojrzewania mRNA, poprzez odpowiednie wycinanie intronów, a także eksonów podczas alternatywnego splicingu. Proces ten jest katalizowany przez kompleks białkowo-RNA, który nazywany jest spliceosomem. By intron podlegał wycięciu, musi posiadać dwie silnie konserwatywne sekwencję: GU na końcu 5' oraz AG na końcu 3'. Mutacje tych miejsc mogą prowadzić do poważnych konsekwencji dla organizmu. Kompleks może nie rozpoznać miejsca zachodzenia splicingu, w związku z czym ekson nie zostanie poprawnie złożony, może utracić część łańcucha, przez przedwczesne odczytanie kodonu STOP lub nienaturalnie się wydłużyć. Miejsce splicingowe może zostać przesunięte, powodując zmianę ramki odczytu, delecję lub insercję wielu aminokwasów. W przypadku zmiany w sekwencje akceptora(ang. splice acceptor variant) lub donora(ang. splice donor varaint) konsekwencje mutacji są najpoważniejsze, ponieważ dotyczą konserwatywnych sekwencji otaczających ekson. Natomiast wariant regionu splicingowego?(ang. splice region variant) dotyczy samych intronów oraz eksonów. Zajście mutacji w takich sekwencjach może skutkować pojawieniem się nowego miejsca splicingowego, przez co część intronu może zostać potraktowana jako ekson, a sam ekson może zostać uznany za intron. Na skutek insercji lub delecji zmieniona może zostać ramka odczytu eksonu, przez co produkt białkowy będzie posiadał inne właściwości.

1.1.5 Mutacje związane z miejscami inicjacji i terminacji translacji done

W kodzie genetycznym występują kodony odpowiadające za terminację translacji: UGA, UAG i UAA. Napotkanie takiej trójki nukleotydowej przez rybosom, jest sygnałem do zakończenia procesu. W przypadku pojawienia się zmian w kodzie jednego z tych kodonów w postaci substytucji, delecji lub insercji, ich funkcje terminacyjne mogą zostać utracone, przeniesione lub za wcześnie zasygnalizują terminacje. Natomiast kodon AUG pełni podwójną funkcje. Jest zarówno sygnałem inicjacji translacji a jednocześnie odczytywany jest przez

rybosom jako aminokwas metionina.
Wyróżniane są tutaj 4 rodzaje mutacji:

- Mutacja kodonu STOP, nie zmieniająca jego funkcji (ang. stop retained variant) nie ma wpływu na translacje, ponieważ nie jest ważne który z trójki nukleotydowej zostanie użyty do terminacji procesu. Przykładem takiej zmiany może być mutacja w kodonie UAG przez substytucje 3 nukleotydu czyli guaniny(G) na adeninę(A), w wyniku czego uzyskany kodon UAA dalej pełni funkcję zakończenia translacji.

- Zmiana powodująca utratę kodonu stop(ang. stop lost) powoduje nienaturalne wydłużenie białka aż do napotkania sygnału terminacji w kodzie kolejnego białka. Konsekwencją takiej mutacji będzie dysfunkcja dwóch białek.

- Przeciwieństwem do tego jest wariant typy pozyskania kodonu stop(ang. stop gained) w wyniku czego syntetyzowane białko nie osiągnie oczekiwanej długości i właściwości lub powstaną dwa krótkie odcinki polipeptydowe, w przypadku gdy następnym kodenem jest AUG(kodon inicjujący translację), które nie będą posiadały biologicznej funkcji lub będzie ona zaburzona.

- Na skutek mutacji utraty kodonu start(ang. start lost) pominięty zostanie cały fragment mRNA aż kompleks natrafi na kolejny sygnał startu, który może być składnikiem białka, ponieważ AUG jest rozpoznawane również jako sygnał do syntezy metioniny, w wyniku czego powstanie jedynie fragment docelowego białka o właściwościach bardziej lub mniej podobnych. W przeciwnym przypadku inicjacja rozpocznie się dopiero gdy rybosom odczyta sygnał inicjacji kolejnego produktu transkrypcji.

1.1.6 Mutacje związane z czynnikami transkrypcyjnymi done?

Warianty sekwencji wiążących czynniki transkrypcyjne(ang. TF binding site varaint)
W genomie każdego organizmu znajdują się specyficzne sekwencje, do których przyłączają się czynniki transkrypcyjne(ang. TF - transcription factor). Są to tak zwane sekwencje promotorowe, które warunkują miejsce startu transkrypcji. Mutacje w obrębie tych sekwencji mogą prowadzić to zmniejszenia lub całkowitego zaprzestania ekspresji danego genu, ponieważ

czynnik transkrypcyjny nie rozpozna charakterystycznej sekwencji DNA, i w konsekwencji tego nie dojdzie do procesu transkrypcji i dalej translacji.

TFBS ablation

Jest to rodzaj delecji, która zawiera regiony związane z przyłączaniem się czynników transkrypcyjnych. Podobnie jak w przypadku zmian sekwencji wiążących czynniki transkrypcyjne, usunięcie ich powoduje zmianę ekspresji danego genu.

1.1.7 coding sequence varaint

1.1.8 protein altering varaint

1.1.9 Mutacje ciche

- intergenetic variant
- Downstream gene varaint i Upstream gene varaint
- Intron variant - Jest to zmiana pojawiająca się w sekwencji niekodującej DNA. Takie mutacje nie wywierają lub wywierają wpływ na organizm. Jak powszechnie wiadomo, introny są wykorzystywane w genomie jako sekwencja regulujące transkrypcje, regiony związane ze splicingiem
- X5 UTR prime varaint i X3 prime UTR variant done?
  Są to zmiany, które nie podlegają translacji ale są związane z jego kontrolą. Są one położone terminalnie, po obu stronach regionu podlegającemu translacji, od strony 5' oraz 3'. Zmiany w tych sekwencjach powodują zmiany w strukturach mRNA, jego konformacji i budowy przestrzennej, które biorą udział w kontroli translacji. **coś jeszcze**
- non coding transcript varaint
- non coding transcript exon varaint

- synonymous varaint(ang. mutacja cicha)

  Jest to zmiana jednego nukleotydu, który nie powoduje zmiany aminokwasu kodowanego przez triplet. Przykładem takiej mutacji jest zmiana CCC na CCU, w tym przypadku oba kodony odczytywane są jako prolina.

1.1.10 Warianty dojrzałęgo miRNA(ang. mature miRNA variant) done?

miRNA lub microRNA jest to mały fragment niekodującego RNA zawierający do 22 nukleotydów. Jest integralną częścią ekspresji mRNA, wpływając na nią przez obniżanie jej poziomu. Działają poprzez komplementarne łączenie się z określonymi sekwencjami na mRNA a w efekcie rozerwanie nici mRNA lub jego destabilizacje poprzez skrócenie ogonu poli (A). Zmiany w ci sekwencji mogą powodować hamowanie translacji białek, które docelowo nie miały nimi być lub dojdzie do nadekspresji jednego z białek.

1.1.11 transcript ablation

1.1.12 Konsekwencja mutacji a wpływ na produkt

Konsekwencja mutacji	wpływ na produkt
X3 prime UTR varaint	Średni
X5 prime UTR varaint	Średni
downstream gene varaint	Średni
frameshift variant	Wysoki
inframe deletion i inframe insertion	Umiarkowany
intron variant	Średni
missense varaint	Umiarkowany
non coding transcript exon variant	Średni
non coding transcript variant	Średni
regulatory region variant	Średni
splice acceptor varaint	Wysoki
splice region varaint	Niski
splice donor varaint	Wysoki
start lost	Wysoki
stop gained	Wysoki
stop lost	Wysoki
synosymus varaint	Niski
TF binding site varaint	Średni
upstream gene variant	Średni
coding sequence varaint	Średni
incomplete terminal codon variant	Niski
protein altering variant	Umiarkowany
stop retained variant	Niski
mature miRNA varian	Średni
TFBS ablation	Średni
intergenic variant	Średni
transcript ablation	Wysoki

1.2 Cel pracy

Celem pracy jest analiza zmienności genetycznej człowieka. Praca omawia zależności pomiędzy ilością wariantów a ich lokalizacją w kariotypie oraz konsekwencją mutacji, a także rozważania na temat samych konsekwencji mutacji. Przedstawione zostaną rodzaje konsekwencji, jakie wywierają zmiany w DNA oraz dlaczego niektóre mutacje są akceptowane przez organizm a inne powodują jego śmierć.

1.3 Baza danych

Niniejsza praca oparta jest na zestawie danych pochodzących z gnomAD(genome aggregation database). Jest to baza danych opracowana przez międzynarodową koalicję badaczy, w celu agregowania i ujednolicenia danych dotyczących szerokiej gamy projektów sekwencjonowania genomu oraz udostępniania tych danych dla społeczności naukowej.
Baza danych obejmuje 125 748 sekwencji egzomowych oraz 15 708 sekwencji całego genomu

od niepowiązanych osobników zsekwencjonowanych w ramach różnych badań genetycznych specyficznych dla choroby i populacji.

Użyty zestaw danych zawiera informacje o wariancjach z 22 chromosomów autosomalnych, oraz zawiera 119794797 indywidualnych rekordów. Dostarczają one informacji o miejscu wystąpienia danej mutacji, czyli chromosomie i jego pozycji w nim, porównuje zmieniony fragment z próbą kontrolną, a także dostarcza informacji o jego frekwencji i ilości z jaką występował podczas tworzenia bazy.

1.4 Użyte oprogramowanie?

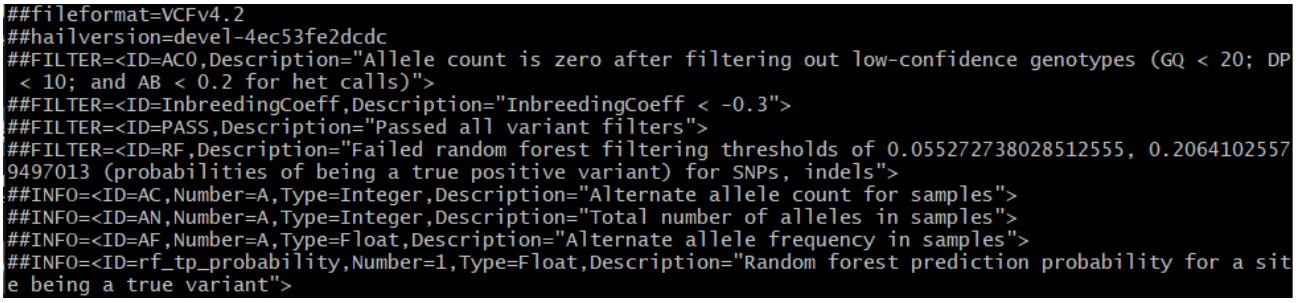
2 Przetwarzanie danych

Variant Call Format(VCF) jest to unikalny format plików tekstowych używany w bioinformatyce do przechowywania danych genomu dotyczących mutacji. Format został opracowany

wraz z rozpoczęciem projektów sekwencjonowania genomów na dużą skalę, takich jak 1000 Genomes Project.

2.1 Specyfikacja formatu VCF

Nagłówek rozpoczyna plik i opisuje jego zawartość. Linie nagłówka są oznaczone jako #. Specjalne słowa kluczowe w nagłówku są oznaczone # #. Zawierają one informacje o pochodzeniu pliku, gatunku oraz objaśniają użyte w nim skróty.



Rysunek 1: Fragment nagłówka z użytego zestawu danych

	Nazwa	Krótki opis.
1	Chrom	Numer chromosomu, na którym znajduje się dany wariant.
2	POS	Pozycja zmiany w sekwencji.
3	ID	Identyfikator zmiany.
4	REF	Allel referencyjny, czyli allel według genomu referencyjnego.
5	ALT	Allel alternatywny, czyli allel jaki pojawił się u osobnika o danym wariancie.
6	QUAL	Wynik jakości allelu?
7	FILTER	Zawiera informacje na temat których filtrów dany rekord nie przeszedł.
8	INFO	Lista zawierająca zastawy klucz-wartość, opisująca wariant.

Pole info zawiera szczegółowe informacje rekordu. Wykorzystane w pracy kategorie to: AC=3, liczba alleli w genotypach AN=2654, całkowita liczba alleli w użytych genotypach

AF=1.13037e-03, częstość występowania danego allelu vep, zawierający szczegółowe informacje na temat wariantu takie jak:
Konsekwencja mutacji, wpływ na produkt, identyfikatory dla biologicznych baz danych(NCBI oraz ENSEMBL) i wiele innych, które nie zostały użyte na potrzeby pracy.

Rysunek 2: Fragment pliku przedstawiający dane

2.2 Przetwarzanie pliku

Z racji dużego rozmiaru pliku oraz zawartości niepotrzebnych informacji, plik ten musiał zostać odpowiednio przygotowany zanim zostanie użyty do przeprowadzenia analizy.
W tym celu został stworzony krótki skrypt w Pythonie, którego zadaniem była ekstrakcja kluczowych dla analizy danych.

2.3 Kod skryptu użytego do wydobycia danych

potrzebne?

```
with open
('/home/dawid/Pulpit/Variant_analysis_data/gnomad.exomes.r2.1.sites.chr21.vcf' ,'r')
as input_file,
open('/home/dawid/Pulpit/Variant_analysis_data/data_1.txt','w')
as output_file:
    for line in input_file:
        if line[0][0] == '#':
            continue
        line = line.split('\t')
        temp = [line[i] for i in [0,1,3,4]]
        info = line[7].split(';')
        info[0] = info[0][len('AC='):]
```

```
info[1] = info[1][len('AN='):]
info[2] = info[2][len('AF='):]
vep = info[-1].split(',')
if line[6] == 'PASS':
    for ele in vep:
        vep_info = ele.strip().split('|')
        vep_info = [vep_info[i] for i in [1,2,3,4,5,6,22]]
        output_file.write('\t'.join(temp+info[0:3]+vep_info)+'\n')
```

3 Analiza danych

3.1 Warianty dotyczące białek

wykres1

wykres2

3.2 Wszystkie warianty genomu

wykres1

wykres2

4 Bibliografia

https://en.wikipedia.org/wiki/Variant_Call_Format

https://www.ensembl.org/info/genome/variation/prediction/predicted_data.html

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5635616/

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4445073/

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4797991/

(How do microRNAs regulate gene expression?

Ian G. Cannell1, Yi Wen Kong1 and Martin Bushell2

School of Pharmacy, Centre for Biomolecular Sciences, University of Nottingham, Universi

https://onlinelibrary.wiley.com/doi/abs/10.1002/wrna.1451