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

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?

1.1 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.
liczba rekordów 119794797

1.2 Zawartość pracy?

Co znajduje się w poszczególnych sekcjach pracy

1.3 Użyte oprogramowanie?

2 Przetwarzanie danych

Variant Call Format(VCF) jest to specyficzny format plików tekstowych powszechnie 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.

```
##fileformat=VCFv4.2
##hailversion=devel-4ec53fe2dcdc
##FILTER=<ID=AC0,Description="Allele count is zero after filtering out low-confidence genotypes (GQ < 20; DP
 < 10; and AB < 0.2 for het calls)">
##FILTER=<ID=InbreedingCoeff,Description="InbreedingCoeff < -0.3">
##FILTER=<ID=PASS,Description="Passed all variant filters">
##FILTER=<ID=RF,Description="Failed random forest filtering thresholds of 0.055272738028512555, 0.2064102557
9497013 (probabilities of being a true positive variant) for SNPs, indels">
##INFO=<ID=AC,Number=A,Type=Integer,Description="Alternate allele count for samples">
##INFO=<ID=AN,Number=A,Type=Integer,Description="Total number of alleles in samples">
##INFO=<ID=AF,Number=A,Type=Float,Description="Alternate allele frequency in samples">
##INFO=<ID=rf_tp_probability,Number=1,Type=Float,Description="Random forest prediction probability for a sit
e being a true variant">
```

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.

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
21	9825785	rs1344912965	T	C	1264.93	RF	AC=3;AN=2654;AF=1.13037e-03;

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

```
with open('/home/dawid/Pulpit/Variant_analysis_data/gnomad.exomes.r2.1.sites.chr21.vcf',
#with open('/media/dawid/Wirusy/Variant_analysis_data/gnomad.exomes.r2.1.sites.vcf','r')
    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