Wprowadzenie do Bioinformatyki
zajęcia laboratoryjne 3

Identyfikacja zmian w sekwencji DNA

Bioinformatyka I rok

1 Wprowadzenie

Zmiany wewnątrz struktury nukleotydowej DNA mogą zachodzić spontanicznie na skutek np. błędów polimerazy DNA lub na skutek czynników mutagennych. Polimorfizm jest to dowolna pozycja w genomie, w której znajdują się co najmniej dwie różne sekwencje, przy czym każda sekwencja występuje u co najmniej 1% populacji. Próg ten jest jednak arbitralny i często do polimorfizmów zaliczane są też rzadsze zmiany, a sam termin jest na ogół używany w odniesieniu do zmian, które bezpośrednio nie prowadzą do chorób o podłożu genetycznym. Mutacje są to znacznie rzadsze zmiany, często mające negatywny wpływ na fenotyp organizmu, przez co są eliminowane z populacji głównie poprzez mechanizm naturalnej selekcji.

Ze względu na możliwość dziedziczenia zmiany w strukturze DNA dzielą się na:

- Somatyczne - pojawiają się podczas mitozy, po zapłodnieniu i w trakcie życia, bardzo rzadko przekazywane są potomstwu.
- Dziedziczne - pojawiają się podczas mejozy w trakcie produkowania jajeczka bądź plemnika, mogą zostać przekazane potomstwu.

Typy zmian, ze względu na budowę:

- Substytucja (podstawienie) - zastąpienie jednego nukleotydu innym (single nucleotide variant - SNV).
- Insercja/delecja - dodanie lub usunięcie co najmniej jednego nukleotydu.
- Inwersja – zmiana kolejności dwóch lub więcej nukleotydów.
- Duplikacja - powielanie jednego lub kilku nukleotydów.
- Translokacja – przemieszczenie jednego lub kilku nukleotydów z jednego miejsca do drugiego w genomie.

Typy zmian ze względu na lokalizację:

- Wewnątrz eksonów genu.
  - W części niekodującej końca 5' lub 3'.
  - W części kodującej:
    - zmiany sensu (ang. missense) - prowadząca do zmiany sekwencji aminokwasów,
    - zmiany synonimiczne (ang. synonymous) – nie prowadzi do zmiany sekwencji aminokwasów (zmieniony kodon odpowiada temu samemu aminokwasowi),
    - prowadzące do przesunięcia ramki odczytu lub wpływające na kodony start/stop.

- Wewnątrz intronów genu.
- W regionach miedzy genowych (ang. intergenic region - IGR).

Zmiany najczęściej oznaczane są z wykorzystaniem nomenklatury HGVS, zawierającej informacje na temat lokalizacji i typu zmiany, jednak bez informacji na temat jej funkcji, np.: "NC_000023.10:g.33038255C>A" oznacza mutację wewnątrz sekwencji NC_000023 z bazy RefSeq (chromosom X) w wersji 10 znajdującej się na pozycji 33038255. Litera g oznacza typ współrzędnej (w tym przypadku sekwencja DNA), C>A opisuje samą zmianę (cytozyna zamieniona na adeninę). Rekord "NP_003997.1:p.Trp24Cys" opisuje zmianę wewnątrz białka NP_003997, znajdującą się w 24 aminokwasie, który został zamieniony z tryptofanu na cysteinę. Zamiast trzy literowych kodów, aminokwasy często są tez zapisane w formie kodów IUPAC, w takiej formie zmiana Trp24Cys jest jednoznaczna z W24C.

Znaczna większość zmian zachodzących w sekwencji DNA nie ma wyraźnego przełożenia na fenotyp organizmu, cześć z nich może jednak prowadzić do groźnych chorób genetycznych takich jak choroba Huntingtona, spowodowana wielokrotnym powieleniem kodonu CAG wewnątrz genu HTT, zespół Downa, będący rezultatem trisomii (obecność dodatkowej kopii) chromosomu 21, czy choroby nowotworowe, będące rezultatem kilku współwystępujących zmian. Niektóre zmiany mogą mieć także korzystny wpływ na życie organizmu, chociaż czasem towarzyszą im także skutki uboczne. Przykładami genów wewnątrz których specyficzne zmiany mogą prowadzić do korzystnych efektów, chociaż nieraz wysokim kosztem, są:

- LRP5 – gen regulujący gęstość mineralną kości, większość mutacji wewnątrz tego genu prowadzi do osteoporozy jednak zmiana Gly171Val (G171V) ma przeciwny efekt znacznie zwiększając ich gęstość a tym samym wytrzymałość (Boyden et al. 2002).
- ACTN3 - gen pomaga wytwarzać specjalne białko zwane alfa-aktyniną-3, która kontroluje szybko kurczące się włókna mięśniowe, mając pozytywny wpływ na sprawność fizyczną organizmu, większość osób ma jednak zmianę Arg577Ter (A577T), która niweluje ten efekt. Brak tej mutacji w obu kopiach genu może w znaczący sposób zwiększać sprawność fizyczną (Yang et al. 2003).
- BHLHE41 (DEC2) - zmiana Pro384Arg (P384A) wewnątrz tego genu jest związana z fenotypem FNSS (Familial Natural Short Sleepers), osoby z tym fenotypem mają tendencję do wysypiania się po około 6 godzinach bez żadnych widocznych skutków ubocznych (He et al. 2009).
- HBB - zmiana heterozygotyczna Glu7Val (E7V) w genie łańcucha β hemoglobiny prowadzi do anemii sierpowatej skracając średnią długość życia chorej osoby, ale jednocześnie zwiększając odporność na malarię (Allison 1954).
- CCR5 - gen odpowiedzialny za produkcje receptora pełniącego ważną role w wielu procesach związanych z układem odpornościowym. Utrata 32 nukleotydowego fragmentu w obu kopiach tego genu (delecja) daje niemal całkowitą odporność na wirusa HIV (Huang et al. 1996).
- MCM6 – zmiana wewnątrz elementu regulatorowego znajdującego się w intronie tego genu zwiększa aktywność genu LCT prowadząc do długotrwałej produkcji laktazy i zdolności do trawienia laktozy przez całe życie. U osób bez tej zmiany (znaczna część światowej populacji) produkcja LCT zmniejsza się po okresie niemowlęcym prowadząc do nietolerancji laktozy (Mattar et al. 2012).

2 Wykorzystywane narzędzia i bazy danych

1. Variant Effect Predictor – narzędzie określające wpływ wariantów (SNV, insercje, delecje, warianty strukturalne) na geny, transkrypty i sekwencję białkową oraz regiony regulatorowe. Ponadto pozwala określić czy określone zmiany występują w innych bazach danych, np. bazie znanych polimorfizmów: https://www.ensembl.org/Homo_sapiens/Tools/VEP?db=core
2. Gene Ontology – baza danych informacji o genach zbudowana z terminów ontologicznych o strukturze grafowej, do których przypisane są specyficzne geny: http://geneontology.org
3. The Genome Aggregation Database (gnomAD) – baza danych obejmująca wyniki sekwencjonowania genomu i egzomu z wielu różnych projektów naukowych (w tym 1000 Genomes Project), obejmująca informacje na temat wariantów odnalezionych u kilkuset tysięcy zdrowych lub chorych osób: https://gnomad.broadinstitute.org/

4. Functional Analysis through Hidden Markov Models (fathmm) – narzędzie do przewidywania konsekwencji funkcjonalnych wariantów kodujących i niekodujących na fenotyp organizmu, poprzez ich związek z określonymi chorobami genetycznymi: http://fathmm.biocompute.org.uk/

5. GDC Data Portal – baza danych wyników eksperymentów przeprowadzonych na ludzkich nowotworach, obejmuje dane uzyskane w projektach TARGET i TCGA (The Cancer Genome Atlas): https://portal.gdc.cancer.gov/

3 Zadania do wykonania

W wyniku sekwencjonowania DNA wyizolowanego z ludzkich komórek, zidentyfikowano kilka zmian w strukturze nukleotydowej o nieznanej funkcji. Celem ćwiczenia jest określenie ich potencjalnego wpływu na badane komórki. Dla zestawu pozycji wariantów znajdujących się wewnątrz ludzkiego DNA (genom referencyjny w wersji GRCh38/hg38) należy wykonać poniższe zadania. Odpowiedzi do poniższych zadań zapisz w pliku tekstowym do okazania pod koniec zajęć.

1. Korzystając z narzędzia Variant Effect Predictor określić, które zmiany położone są wewnątrz genów kodujących białka oraz jaka jest ich konsekwencja dla kodowanego białka. Należy ograniczyć liczbę rekordów do jednego na daną pozycję – tylko najbardziej istotna konsekwencja zmiany (opcja Filtering options → Restrict results: Show one selected consequence per variant). Należy zachować identyfikatory rs z bazy dbSNP, jeśli są dostępne (pole Existing variant) oraz symbole genów.

2. Która ze zmian w specyficznych genach opisanych we wprowadzeniu znajduje się na liście uzyskanej w punkcie 1? Należy zwrócić uwagę na pozycje wewnątrz białka i zmianę aminokwasu.

3. Korzystając z bazy GeneOntology (narzędzie do wyszukiwania AmiGO) należy odnaleźć 5 pierwszych terminów opisujących procesy biologiczne, w które jest zaangażowany gen z punktu 2. Po wpisaniu symbolu genu wyniki można zawęzić filtrem "Organism" tylko dla człowieka. Po wejściu do opisu genu należy zawęzić wyniki do procesów biologicznych filtrem "Ontology (aspect) → P". Należy wyświetlić szczegółowe informacje na temat jednego z terminów, klikając na jego nazwę w kolumnie "GO class (direct)" (wybierz pierwszy możliwy termin). Jaka jest jego definicja i identyfikator? Należy określić jego terminy nadrzędne i podrzędne w hierarchii (przez podrzędne rozumie się bezpośrednich potomków) – wykorzystaj link "See term history for at QuickGO", a w ramach odpowiedzi zapisz wykres przodków ("Ancestor Chart").

4. W oparciu o wyniki punktu 1 należy wybrać zmiany położone wewnątrz genów kodujących białka, a następnie określić częstotliwość ich występowania w całej populacji korzystając z bazy gnomAD. Należy w tym celu użyć identyfikatorów rs. Jeśli gen nie posiadał identyfikatora rs to pomiń go. Wykorzystaj wersję bazy danych v2.1.1. Może się zdarzyć, że nie będzie informacji dla identyfikatora rs, jeśli tak, to wyróżnij, który to identyfikator.

5. Korzystając z narzędzia fathmm (wariant Protein Missense Variants) oraz identyfikatorów rs z punktu 1, należy dokonać predykcji wpływu zmian na fenotyp (pole prediction). Czy są zmiany, które nie wywołują uszkodzeń? Jeśli zmiany mogą powodować uszkodzenia, należy podać przykłady chorób w jakie dana zmiana może być zaangażowana.

6. [**Zadanie dodatkowe dla chętnych**]
Korzystając z wyszukiwarki GDC Data Portal należy określić czy geny kodujące białka zidentyfikowane w punkcie 1 mogą mieć istotny wpływ dla procesu karcynogenezy poprzez sprawdzenie czy występują w nich zmiany charakterystyczne dla ponad 5% przypadków specyficznego nowotworu. Przy wyszukiwaniu należy użyć jedynie symboli genów uzyskanych w zad 1.