Integracyjna analiza danych genomicznych z pojedynczych komórek

Streszczenie popularnonaukowe

Genom zawiera pełną informację genetyczną organizmu. Ta informacja jest interpretowana w procesie ekspresji genów, w którym sekwencje DNA poszczególnych genów są dekodowane do funkcjonalnych produktów, takich jak białka. Prawidłowa ekspresja genów to precyzyjnie sterowany proces kontrolowany przez elementy regulatorowe DNA znajdujące się w niekodujących częściach genomu. Proces ten był szeroko badany za pomocą technik eksperymentalnych wykorzystujących sekwencjonowanie DNA i RNA. Jednakże tradycyjne metody sekwencjonowania pozwalają uzyskać tylko uśredniony sygnał z wielu komórek, gubiąc w ten sposób informację o różnicach między komórkami. Zmieniło się to wraz z powstaniem technologii, które pozwalają na sekwencjonowanie genomu lub transkryptomu pojedynczych komórek (single cell sequencing). Zastosowanie technologii technologii sekwencjonowania pojedynczych komórek ujawnia niejednorodności w populacji komórek i pozwala na wyróżnienie rzadkich typów komórek.

Celem projektu jest integracja danych genomicznych z metod na poziomie pojedynczej komórki i na poziomie populacji, w celu pełniejszego zrozumienia procesu ekspresji genów. Mechanizmy regulujące ekspresję genów funkcjonują w złożonej trójwymiarowej architekturze, pozwalającej zbliżyć w przestrzeni funkcjonalne fragmenty DNA do siebie. Trójwymiarowa architektura genomu dopuszcza, aby elementy regulatorowe znajdowały się daleko na nici DNA od ich regulowanych przez nie genów, z wieloma innymi genami znajdującymi się pomiędzy nimi. Mimo tego, że trójwymiarowa organizacja genomu ma zasadnicze znaczenie dla właściwej interpretacji informacji genetycznej, to metody eksperymentalne pozwalające określić tę organizację są zwykle wykonywane na poziomie populacji, aby zapewnić wystarczającą rozdzielczość. W tym celu uzupełnimy dane o organizacji genomu danymi z eksperymentów sekwencjonowania pojedynczych komórek, mierzących transkrypcję genów (single cell RNA-seq) i dostępność genomu (single cell ATAC-seq).

Trójwymiarowa organizacja genomu zachodzi w wielu skalach. Jedną z nich stanowią dobrze rozgraniczone domeny fizycznych interakcji, znane jako domeny skojarzone topologicznie (Topologically Associating Domains) lub domeny chromatynowe. Struktura domen chromatynowych koreluje z ekspresją genów, co pozwala na zdefiniowanie aktywnych i nieaktywnych domen chromatynowych na poziomie populacji. Jednak trend obserwowany na poziomie populacji może być przeciwny do tego, jaki występuje w niektórych typach komórek. Użyjemy danych z sekwencjonowania RNA w pojedynczych komórkach, aby sprawdzić, jak częste jest to zjawisko. Przyjrzymy się również zmienności ekspresji genów, obserwowanej w poszczególnych typach komórek i domenach chromatynowych, i zbadamy, w jakim stopniu geny o dużej zmienności w ich ekspresji są zlokalizowane w tej samej domenie.

Drobniejszy poziom organizacji genomu zapewnia, by elementy regulatorowe znajdowały się w przestrzennej bliskości promotorów regulowanych przez nie genów. Te kontakty enhancer-promotor można zaobserwować w danych Hi-C na poziomie populacji, ale same te dane nie pozwalają stwierdzić, czy kontakt występuje we wszystkich typach komórek. Aby ujawnić zróżnicowanie w tych kontaktach w ramach populacji oraz przewidzieć, w jakich typach komórek mają one miejsce, wykorzystamy dane ATAC-seq dla pojedynczych komórek i sprawdzimy dostępność chromatyny w oddziałujących regionach. Na koniec zintegrujemy dane dotyczące dostępności chromatyny z danymi RNA-seq dla pojedynczych komórek z tej samej populacji. Korzystając z dodatkowych informacji o miejscach wiązania czynników transkrypcyjnych i metod analizy sekwencji DNA, spodziewamy się zidentyfikować czynniki odpowiedzialne za aktywację lub represję transkrypcji w poszczególnych typach komórek.