Dr. Barbara Uszczyńska-Ratajczak

Abstract

Mammalian genomes produce tens of thousands of long noncoding RNAs (lncRNAs) – long transcripts with limited protein coding potential. Although an increasing number of lncRNAs is linked to fundamental physiological processes in the cell, the vast majority of them (~97%) still remain functionally uncharacterized. Understanding biological roles of lncRNAs requires accurate genome annotations describing their precise location, gene boundaries and transcript structures. Current lncRNA catalogs show evident signs of incompleteness with many gene models being fragmented or uncatalogued. To overcome this issue, the present work aims to advance towards a complete and accurate annotation of lncRNAs in human and mouse genomes. By developing and applying targeted long-read RNA sequencing methodology, this study provides accurate lncRNA annotations at high-throupt rates. Presented methodology eliminates the need for the noisy transcriptome assembly and requires minimal manual curation. Produced transcript models uncover thousands and hundreds of novel, full-length lncRNAs for human and mouse genomes, respectively, also substantially increasing the annotated transcript complexity within targeted loci. Resulting lncRNA catalogs are of quality comparable to present-day manually curated annotations. Moreover, the full-length models enabled to confidently redefine the genomic properties of lncRNAs and show that so far our perception of lncRNA features has been largely driven by the incompleteness of their annotations.

dr Barbara Uszczyńska-Ratajczak

Streszczenie w języku polskim

Genomy ssaków wytwarzają dziesiątki tysięcy długich, niekodujących RNA (lncRNA) - długich transkryptów o ograniczonym potencjale kodowania białek. Chociaż coraz więcej doniesień naukowych potwierdza rolę lncRNA w regulacji podstawowych procesów fizjologicznych w komórce, dla zdecydowanej większości z nich (~97%), ich biologiczna funkcja nadal pozostaje nieznana. Zrozumienie biologicznego znaczenia lncRNA i ich ról w komórce wymaga dostępności wysokiej jakości katalogów genów, które opisują ich dokładną lokalizację w genomie, wskazują granice ich genów oraz określają struktury ich transkryptów. Obecnie istniejące katalogi lncRNA wykazują wyraźne oznaki niekompletności, które obejmują głównie fragmentaryczne modele dla wielu genów oraz brak licznych, niezidentyfikowanych dotąd struktur. Aby przezwyciężyć ten problem, celem tej pracy jest ułatwienie procesu katalogowania lncRNA dla genomów człowieka i myszy poprzez opracowanie ukierunkowanych, wysokoprzepustowych metod opartych na technologii sekwencjonowania RNA za pomocą długich odczytów. Przedstawione metody oferują połączenie przepustowości i dokładności dzięki eliminacji etapu rekonstrukcji transkryptów, który charakteryzuje się wysokim odsetkiem transkryptów o błędnych strukturach. Co więcej, proponowane tutaj metody ograniczają do minimum udział człowieka na etapie budowy i weryfikacji kompletności struktur modeli transkryptów. W efekcie uzyskane dane pozwoliły na identyfikację setek oraz tysięcy nowych lncRNA o pełnej długości dla odpowiednio genomów człowieka i myszy, jednocześnie znacząco zwiększając liczbę transkryptów w obrębie testowanych loci. Ponadto otrzymane katalogi lncRNA charakteryzują się wysoką jakością, porównywalną z tą dla katalogów weryfikowanych manualnie. Dodatkowo modele lncRNA o pełnej długości pozwoliły skutecznie przedefiniować właściwości genomiczne lncRNA i pokazały, że nasze dotychczasowe postrzeganie cech lncRNA w dużej mierze było konsekwencją artefaktów wynikających z niekompletności katalogów lncRNA.