Genom człowieka składa się z trzech miliardów liter DNA kodujących geny, które odgrywają kluczową rolę w tworzeniu nas takimi, jakimi jesteśmy. Stąd też bardzo często genom określany jest mianem "księgi życia". Tekst "księgi życia" (kolejność liter DNA) od dawna jest znany i choć każdy może go przeczytać, w tej postaci nie daje on żadnych informacji o biologicznych funkcjach poszczególnych regionów. Zatem dla wielu genów ich role w komórce nadal pozostają niezdefiniowane. Geny kodujące białka, których celem jest produkcja białka w komórce są kluczową grupą genów bez których nie możemy istnieć. Dlatego też pierwotnie przewidywano, że to właśnie one budują genom oraz że złożoność organizmu zależy od ich liczby. Zatem zakładano, że im bardziej złożony organizm, tym grubsza jest jego "księga życia". Poznanie sekwencji ludzkiego genomu wiązało się z ogromnym zaskoczeniem, gdyż okazało się, że genom ludzki zawiera niemalże tyle samo genów co genom robaka. Okazało się także, że geny kodujące białka stanowią zaledwie 1-2% ludzkiego genomu. Pozostała część to regiony niekodujące białek, które ze względu na nieodgadnioną rolę jaką pełnią w komórce nazwane zostały „ciemną materią" genomu. Kolejne zdumienie nadeszło wraz z poznaniem tego jak „księga życia" drukowana jest w komórce. Nie tylko stwierdzono, że komórki aktywnie drukują cześć niebiałkową genomu, ale także, że stanowi ona źródło różnego rodzaju jednostek regulatorowych, które często oddziałują z genami kodującymi białka. Największą i najbardziej intrygującą klasą niebiałkowego genomu są długie niekodujące RNA (lncRNAs, ang. long non-coding RNAs). Choć lncRNAs w zasadzie są bardzo podobne do genów kodujących białka, podstawową różnicą jest to, że nie produkują one żadnych białek. Naukowcy przez długie lata zaniedbywali tę część genomu, aż do momentu kiedy ujawniono, że wiele lncRNAs pełni kluczowe funkcje biologiczne w komórce. Niektóre z nich są także zaangażowane w progresję różnych chorób, w tym także nowotworów. Jednak do tej pory jedynie ok. 2% ludzkich lncRNAs (z ok. 19,000) zostało funkcjonalnie scharakteryzowanych. Funkcje dla pozostałej części z nich pozostają nieokreślone.

Jednym z głównych zdanań współczesnej biologii jest zrozumienie, które lncRNA są funkcjonalne i jak te funkcje są przechowywane w genomie. Osiągnięcie tego celu jest złożone. Po pierwsze wymaga znajomości lokalizacji wszystkich lncRNA w genomie. Jest to niezwykle trudne, gdyż lncRNAs nie dostarczają żadnych wskazówek, które pomogłyby ją ustalić. Odnalezienie genów kodujących białka w genomie jest znacznie łatwiejsze dzięki temu, że znamy końcowy produkt ich aktywności - białko zbudowane z aminokwasów. Poszukiwanie białek w sekwencji genomu można porównać do próby znalezienia zdania po hiszpańsku w anglojęzycznej książce. Najpierw należy przetłumaczyć zdanie, a następnie użyć go do przeszukania tekstu książki. Niestety w przypadku lncRNAs ostatecznym produktem jest RNA. Zatem znalezienie ich wymaga przeszukania wszystkich RNA w komórce. Dodatkowo proces ten utrudnia ich tkankowa i komórkowa specyficzność. W niektórych komórkach dany rodzaj lncRNA jest w pełni obecny, natomiast w innych jedynie częściowo lub całkowicie nieobecny. Jest to jak wydruk identycznego tekstu przy użyciu wszystkich drukarek w budynku i uzyskiwanie różnych wyników z poszczególnych biur lub działów. Co więcej, obecność lncRNA w komórce jest w dużej mierze zdominowana przez geny kodujące białka, które są znacznie częściej drukowane, niż niekodujące elementy genomu. Zatem katalogowanie lncRNAs wymaga wyławiania ich z mieszaniny RNA w komórce. Po drugie, charakterystyka lncRNAs wymaga poznania stopnia ich niezmienności ewolucyjnej. Ewolucja zachowuje kluczowe funkcje biologiczne, przekazując je w formie fragmentów DNA od jednego genomu do drugiego na przestrzeni milionów lat. Stąd też niezmienne ewolucyjnie lncRNAs mogą mieć kluczowe znaczenie dla komórki. Do tej pory analiza zachowawczości ewolucyjnej lncRNAs prowadzona była głównie w oparciu o porównania całogenomowe, gdzie sekwencje lncRNAs mapowane były do pełnej sekwencji genomu innego organizmu. Podejście to pozwala jedynie wykryć obecność danego lncRNA w genomie innego organizmu, nie dostarcza natomiast informacji o pokrewieństwie sekwencji lncRNAs u różnych gatunków. To jak wyszukiwanie zdań z jednego rozdziału danej książki używając pełnego tekstu innej książki. Możliwe jest znalezienie części wspólnych pomiędzy książkami, ale niekoniecznie pomiędzy rozdziałami.

Aby lepiej zrozumieć, które lncRNAs są funkcjonalne i jak ta funkcja jest przechowywana w ich sekwencji planujemy: Po pierwsze ulepszyć adnotację lncRNA – szczegółową mapę opisującą lokalizację lncRNA w genomie Danio pręgowanego (*Danio rerio*). Wyzwania związane z identyfikacją lncRNAs przezwyciężymy skupiając się jedynie na fragmentach genomu, które potencjalnie mogłyby je zawierać. Podejście to można porównać do używania zakreślacza w celu zaznaczenia istotnych fragmentów książki, zamiast przeszukiwania całego tekstu. Po drugie, wykorzystamy nasz nowy katalog lncRNA do oceny niezmienności ewolucyjnej pomiędzy sekwencjami ludzkich oraz mysich lncRNAs. Następnie zbadamy wpływ wybranych, niezmiennych ewolucyjnie lncRNAs na rozwój embrionalny Danio pręgowanego. Oczekujemy, iż wyniki tego projektu w znacznym stopniu przyczynią się do zrozumienia biologicznych funkcji lncRNA w komórce.