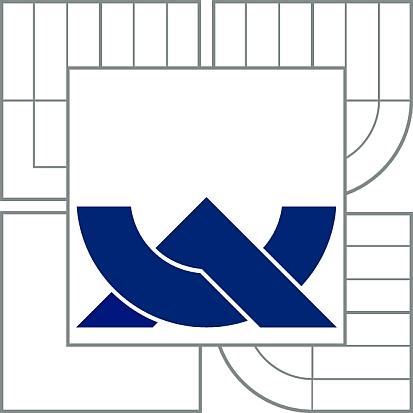
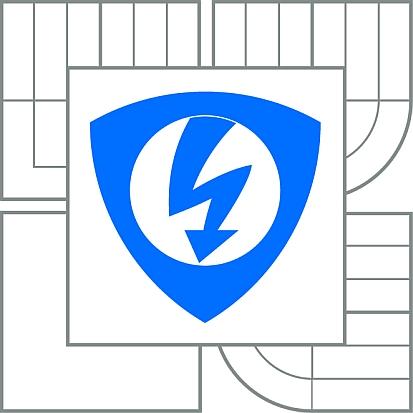
VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ



BRNO UNIVERSITY OF TECHNOLOGY

FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ



ÚSTAV BIOMEDICÍNSKÉHO INŽENÝRSTVÍ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

DEPARTMENT OF BIOMEDICAL ENGINEERING

VYHLEDÁVÁNÍ LTR RETROTRANSPOZONŮ V LIDSKÉM GENOMU

IDENTIFICATION OF LTR RETROTRANSPOSONS IN HUMAN GENOME

SEMESTRÁLNÍ PRÁCE

SEMESTRAL THESIS

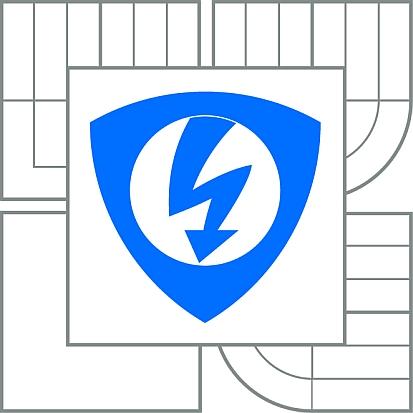
AUTOR PRÁCE EDUARD TROTT

AUTHOR

VEDOUCÍ PRÁCE Ing. KAREL SEDLÁŘ

SUPERVISOR

BRNO 2014

**VYSOKÉ UČENÍ**

**TECHNICKÉ V BRNĚ**

**Fakulta elektrotechniky**

**a komunikačních technologií**

**Ústav biomedicínského inženýrství**

**Semestrální práce**

bakalářský studijní obor

**Biomedicínská technika a bioinformatika**

|  |  |  |  |
| --- | --- | --- | --- |
| ***Student:*** | Eduard Trott | ***ID:*** | 155615 |
| ***Ročník:*** | 3 | ***Akademický rok:*** | 2014/2015 |

**NÁZEV TÉMATU:**

**Vyhledávání LTR retrotranspozonů v lidském genomu**

**POKYNY PRO VYPRACOVÁNÍ:**

1) Zpracujte literární rešerši metod pro vyhledávání LTR retrotranspozonů v DNA, zaměřte se především na metody de novo. 2) Popište jednotlivé části retrotranspozonu a rodiny typické pro lidský genom, včetně jejich možných spojení s onemocněními. 3) Navrhněte a v jazyce R/Bioconductor realizujte nástroj pro vyhledávání LTR retrotranspozonů s vhodným výstupem (gff soubor). Funkčnost ověřte na sekvencích nejnovější dostupné verze lidského genomu. 4) Nástroj doplňte o možnost nalezené elementy rozdělit do rodin a tyto rodiny identifikovat pomocí vhodné referenční databáze. 5) Zhodnoťte úspěšnost vyhledávání pomocí již dostupné anotace, například z genomového prohlížeče UCSC. 6) Výsledky statisticky vyhodnoťte a diskutujte.

Pro splnění semestrálního projektu je nutné vypracování bodů 1) až 3).

**DOPORUČENÁ LITERATURA:**

1. RHO, Mina, Jeong-Hyeon CHOI, Sun KIM, Michael LYNCH a Haixu TANG. De novo identification of LTR retrotransposons in eukaryotic genomes. BMC Genomics. vol. 8, issue 1, s. 90-.
2. KATOH, Iyoko a Shun-ichi KURATA. Association of Endogenous Retroviruses and Long Terminal Repeats with Human Disorders. Frontiers in Oncology. 2013, vol. 3.

|  |  |  |  |
| --- | --- | --- | --- |
| ***Termín zadání:*** | 22.9.2014 | ***Termín odevzdání:*** | 5.1.2015 |
| ***Vedoucí práce:*** | Ing. Karel Sedlář |  |  |
| ***Konzultanti semestrální práce:*** | |  |  |
|  | **prof. Ing. Ivo Provazník, Ph.D.** |  |  |
| **UPOZORNĚNÍ:** | *Předseda oborové rady* |  |  |

Autor semestrální práce nesmí při vytváření semestrální práce porušit autorská práva třetích osob, zejména nesmí zasahovat nedovoleným způsobem do cizích autorských práv osobnostních a musí si být plně vědom následků porušení ustanovení § 11 a následujících autorského zákona č. 121/2000 Sb., včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č.40/2009 Sb.

**Table of contents**

[Abstract 4](#_Toc405763170)

[LTR Retrotransposons 5](#_Toc405763171)

[Участие LTR in human pathogenesis. 6](#_Toc405763172)

[De novo algorithm of searching LTRs 7](#_Toc405763173)

[Results 7](#_Toc405763174)

[References 8](#_Toc405763175)

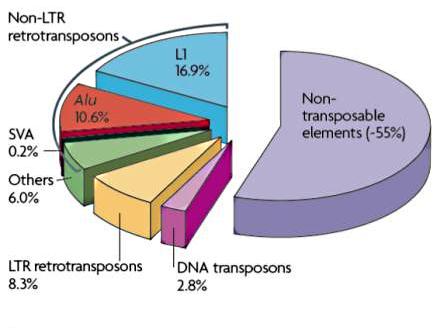
## Abstract

Human LTR elements are endogenous retroviruses which account for ~8% of the human genome. Now most human endogenous retroviruses (HERVs) are traces of viruses, which have been integrated millions of years ago. However HERVs and solitary LTR retrotransposons, not involved in the direct biological processes, may act as additional transcription apparatuses of genes by reactivation in generations or individuals. De novo approaches of searching retrotransposons in the human genome may later lead to finding new retroelements responsible for the cellular biological processes in the causes of which people have not understood at this time.

**Keywords:** long terminal repeat, LTR, retrotransposon

## LTR Retrotransposons

Transposable element (TEs), also known as transposons or "jumping genes", are discrete pieces of DNA sequence that can move in the genome from one location to another. Transposons represent one of types of mobile genetic elements. TEs are allocated to one of two classes, depending on their mechanism of transposition. The first class is retrotransposons, which are copied in two stages: first they are transcribed from DNA to RNA, and then RNA produced is reverse transcribed to DNA. This DNA copy is then inserted at a new position into the genome.



Retrotransposons usually consist of three sub-types:

* LINEs(L1): encode reverse transcriptase, and are transcribed by RNA polymerase II
* SINEs(Alu): transcribed by RNA polymerase III
* LTRs(TEs with long terminal repeats): encode reverse transcriptase, similar to retroviruses

**Figure 1**

**The transposable element content of the human genome**

LTR ретротранспозоны окружены long terminal repeats которые содержат все нужные для регуляции транскрипции элементы. The autonomous elements содержат *gag* and *pol* genes, которые кодируют reverse transcriptase and protease.

LTR ретротранспозоны в свою очередь делятся на три подкласса:

* Ty1-copia-like ([Pseudoviridae](http://en.wikipedia.org/wiki/Pseudoviridae))
* Ty3-gypsy-like ([Metaviridae](http://en.wikipedia.org/wiki/Metaviridae))
* BEL-Pao-like

Ретровирусы могут трансформироваться в LTR ретротранспозоны путем инактивации или удалением структур ответственных за внеклеточную мобильность. Если ретровирус заражает и впоследствии встраивает себя в геном in germ line cells, то может стать an Endogenous Retrovirus (ERV). Therefore Exogenous retroviruses возникли из endogenous retrotransposons приобретением cellular *envelope* gene [3]

Human LTR elements are endogenous retroviruses which account for ~8% of the human genome. [1]

In general, most (85%) of the LTR retrotransposon-derived parts consist only of an isolated LTR, with the internal sequence having been lost by homologous recombination between the flanking LTRs.[2]

## Участие LTR in human pathogenesis.

Более 25 экспериментально охарактеризованных клеточных генов показывают  LTR-mediated эволюционные изменения, в которых встроенные LTRs являются  alternative promoters для предоставления новой tissue-specificity, play as the major promoters, or promotes only minor effects. [4]. For example, A HERV-K(HML-5) LTR plays as the major promoter of INSL4, a insulin-like growth factor gene expressed in placenta. [5]. A HERV-E family LTR plays as an alternative tissue-specific promoter of the endothelin B receptor (EDNRB) gene, by which the gene expression increased ∼15% in placenta. [6]. LTR-derived promoters often increase placenta-specific gene expression, несмотря на то что в общем эффект от LTR insertions проявляется умеренно в многих случаях.

Последние исследования показали, что HERV-encoded peptide as a tumor-specific antigen участвует in the hematopoietic stem cell transplantation  for the therapy of renal cell carcinoma (RCC) [7]

A pioneering study investigate that HERV-E is activated in RCC and that it encodes an overexpressed immunogenic antigen, therefore providing a potential target for cellular immunity [8]. The tumor antigen, CT-RCC-1, recognized by RCC-specific CD8+ T cells is encoded by novel spliced variants of the HERV-E.

A study on tumorigenesis of Hodgkin’s lymphoma provided evidence that aberrant LTR activation contributes to lineage-inappropriate gene expression in transformed human cells and that such gene expression is central for tumor cell survival. They show that B cell–derived Hodgkin's lymphoma cells depend on the activity of the non-B, myeloid-specific proto-oncogene colony-stimulating factor 1 receptor (CSF1R). *CSF1R* transcription in these cells initiates at an aberrantly activated endogenous LTR of the MaLR family (*THE1B*). They conclude that LTR derepression is involved in the pathogenesis of human lymphomas. [9]

Human endogenous retroviruses are remnant forms of infectious retroviruses that integrated into the chromosomal DNA of germ-line cells of human ancestors, increased their copy numbers and have been inherited by present-day humans. Most HERVs are merely traces of original viruses, having first integrated millions of years ago. Within the published human genome sequence, there are over 98,000 human endogenous retroviruses (HERVs), but all are defective, containing nonsense mutations or major deletions. No replication-competent HERVs have been identified to date. [10] However, solitary LTRs derived from HERVs and MaLRs dominate the provirus forms in the copy numbers, and can serve as redundant enhancer-promoter sequences for nearby cellular genes. When the DNA methylation-mediated suppression system becomes compromised, HERVs and LTRs can cause detrimental and/or self-protecting effects. Two prominent examples of the clinically significant HERV/LTR activation have been reported: CSF1R oncogene activation by a MaLR LTR in Hodgkin’s lymphoma and RCC-specific novel HERV-E antigen expression facilitating the immunotherapy. Future researches in oncology and immunogenetics will unveil more details about the endogenous LTR functions in human pathogenesis.

## De novo algorithm of searching LTRs

The conventional approach to annotating MGEs in genomic sequences is based upon homology searching against a well-updated library of known MGEs, e.g. Repbase , using a fast searching program, e.g. RepeatMasker. This approach, however, is limited to annotating those known MGE families, and thus cannot identify new elements. Furthermore, it sometimes even overlooks known elements, because the repetitive nature of MGE elements may confuse the statistical methods (e.g. E-values) that are commonly used in genome annotation

В моем de novo подходе не используются референчные датабазы уже найденных LTR, что позволяет находить новые участки LTR. Нахождения молодых LTR retrotransposons разделена на три стадии:

Первая стадия это нахождение одинаковых участков длинной 40 bp ограниченных минимальным(1000bp – minimal length of entire of LTR) и максимальным (10 kbp maximum length of entire of LTR) расстоянием между ними.

Я имплементировал несколько алгоритмов для поиска повторяющихся паттернов, KnuthMorrisPratt , Binary Search with LCP and Suffix array and searching by встроенными функциями языка Python. Быстрее всего оказался поиск встроенными Python функциями из-за того что он C implementation searching algorithm when остальные написаны на Python.

Во второй стадии идет формирование повторяющихся последовательностей в группы, ассоциированные с отдельными LTR элементами. Эти группы так же строятся на основе того что должны удовлетворять структурному строению LTR retrotransposons. Далее эти группы записываются форме 4 индексов отвечающих за начало ведущего LTR, конец ведущего LTR, начало trailing LTR, end of trailing LTR. Алгоритм учитывает **gene amplification**, minimal and maximal length of LTR parts.

В третьей стадии идет формирование вычисление идентичности LTR для каждого набора индексов полученного во второй стадии и формирование GFF file. В GFF file содержится начало и конец каждого из LTR parts of retrotransposon and percentage identity between endings.

## Results

Подавляющее большинство найденных LTR retrotransposons have identities more than 80%. Это показывает достаточную гибкость алгоритма для поиска молодых LTRs. Данный алгоритм нацеленный только на поиск молодых LTRs находит малое количество элементов относительно уже найденных, есть еще возможности для его улучшения, как со стороны увеличения быстродействия, так и со стороны увеличения количества найденных retroelements.

Toolboxes:

Shelve

## References

[1] <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2884099/>

[2] <http://www.nature.com/nature/journal/v409/n6822/full/409860a0.html>

[3] Malik, H. S., Henikoff, S. & Eickbush, T. H. Poised for contagion: evolutionary origins of the infectious abilities of invertebrate retroviruses.*Genome Res.* **10**, 1307–1318

[4] Cohen CJ, Lock WM, Mager DL. Endogenous retroviral LTRs as promoters for human genes: a critical assessment. Gene (2009) 448:105–1410.1016/j.gene.2009.06.020

[5] Bièche I, Laurent A, Laurendeau I, Duret L, Giovangrandi Y, Frendo J-L, et al. Placenta-specific INSL4 expression is mediated by a human endogenous retrovirus element. Biol Reprod (2003) 68:1422–910.1095/biolreprod.102.010322

[6]  Landry J-R, Mager DL. Functional analysis of the endogenous retroviral promoter of the human endothelin B receptor gene. J Virol (2003) 77:7459–6610.1128/JVI.77.13.7459-7466.2003

[7] Takahashi Y, Harashima N, Kajigaya S, Yokoyama H, Cherkasova E, McCoy JP, et al. Regression of human kidney cancer following allogeneic stem cell transplantation is associated with recognition of an HERV-E antigen by T cells. J Clin Invest (2008) 118:1099–10910.1172/JCI34409

[8] Takahashi Y, Harashima N, Kajigaya S, Yokoyama H, Cherkasova E, McCoy JP, et al. Regression of human kidney cancer following allogeneic stem cell transplantation is associated with recognition of an HERV-E antigen by T cells. J Clin Invest (2008) 118:1099–10910.1172/JCI34409

<http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2248804/>

[9]  Lamprecht B, Walter K, Kreher S, Kumar R, Hummel M, Lenze D, et al. Derepression of an endogenous long terminal repeat activates the CSF1R proto-oncogene in human lymphoma. Nat Med (2010) 16:571–910.1038/nm.2129

<http://www.ncbi.nlm.nih.gov/pubmed/20436485/>

[10]  Belshaw R, Dawson AL, Woolven-Allen J, Redding J, Burt A, Tristem M (Oct 2005).["Genomewide Screening Reveals High Levels of Insertional Polymorphism in the Human Endogenous Retrovirus Family HERV-K(HML2): Implications for Present-Day Activity"](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1211540).*J Virol.* **79** (19): 12507–14. [doi](http://en.wikipedia.org/wiki/Digital_object_identifier):[10.1128/JVI.79.19.12507-12514.2005](http://dx.doi.org/10.1128%2FJVI.79.19.12507-12514.2005). [PMC](http://en.wikipedia.org/wiki/PubMed_Central) [1211540](http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1211540).[PMID](http://en.wikipedia.org/wiki/PubMed_Identifier) [16160178](http://www.ncbi.nlm.nih.gov/pubmed/16160178).