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

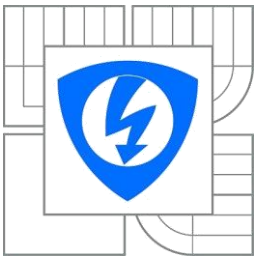
AUTOR PRÁCE — EDUARD TROTT

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.
Předseda oborové rady

UPOZORNĚNÍ:

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.

Abstract

The goal of this thesis is a literary background research on the topic Search LTR retrotransposons in the human genome. It is necessary to characterize the potential problems related to a given topic and to implement appropriate searching algorithm, the result of which is the GFF file that contains all found LTRs.

Keywords: long terminal repeat, LTR, retrotransposon, de novo

Abstrakt

Cílem této semestrální práce je zpracování literárního rešerši o tématu Vyhledávání LTR retrotranspozonů v lidském genomu. Je potřeba popsat možné problematiky navazující na danou tématu a implementovat vhodný algoritm vyhledávání, výsledkem kterého je GFF soubor, který obsahuje všechny nalezeny LTRs.

Klíčová slova: LTR, retrotranspozony

TROTT, E. *VYHLEDÁVÁNÍ LTR RETROTRANSPOZONŮ V LIDSKÉM GENOMU*. BRNO: VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ, FAKULTA ELEKTROTECHNIKY A KOMUNIKAČNÍCH TECHNOLOGIÍ, 2015. 27 S. VEDOUCÍ SEMESTRÁLNÍ PRÁCE ING. KAREL SEDLÁŘ.

Declaration

I declare that my semestral thesis on the topic of Identification of LTR retrotransposons in human genome, I developed independently under the leadership of supervisor of the semestral thesis and using of literature and other information sources, all of which are cited in the work and listed in the end of the work.

As the author mentioned semestral thesis further declare that, in connection with the creation of the semester work I did not break the copyrights of third parties, in particular, I did not intervened illegally in the foreign copyright of personal and / or property and ~ I am fully aware of the consequences of violation of § 11 et seq Act no. 121/2000 Sb., on copyright, rights related to copyright and amending some laws (Copyright Act), as amended, including possible criminal consequences arising from the provisions of Part II, Title VI. Part 4 of the Penal Code no. 40/2009 Sb.

In Brno on ....

(signature of author)

Acknowledgments

I would like to thank and recognize supervisor of this project M.Sc. Karel Sedlář, for his continued support and guidance throughout the completion of this project.

In Brno on ....

(signature of author)

Table of contents

Introduction

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.

In the beginning of this semestral thesis there will be discussed the foundation of genetics, also there will be introduced the fundamental families of transposable elements.

The second part of the thesis will be described one of the families of retrotransposons (mobile genetic elements) - LTR retroelements, their structure and possible involvement in human pathogenesis.

In the next part of the work there will be represented several existing algorithms for de novo searching of these elements.

The result of this work will be the implementation of de novo algorithm to search for LTR elements, which takes aim on the structure of the required elements, their relative positions and their implements to certain families. To determine the quality of the algorithm, the result will be compared with results, which has been taken from another approaches for LTR identifications and reference databases. Just at the end of the work will be provided in the evaluation of the identification of genome browser UCSC.

1 Theoretical background

1.1 Genetics

Since 1953, when Watson and Crick has discovered and deciphered the structure of DNA, started a new era of genetics, a lot of research areas was manifested. For example, molecular genetics, which reveals the chemical nature of heredity, or genetic engineering, dealing with genetic manipulation and introducing them to other organisms, or population genetics, archaeogenetics and many others. Also appeared direction, now representing the field of medicine, which identifies, examines and treats hereditary diseases - a medical genetics.

Big breakthrough of medical genetics is the possibility of sequencing the genome of an individual. This becomes possible by using the development of high-performance sequencing. The cost of genome sequencing is decreasing every year by an exponential scale, which provides opportunities of personal genomics for more people. [23]

Given all of this, research in the field of genetics got a large value. Carried out a number of projects which aimed on studying of the human genome, such as The International HapMap Project or 1000 Genomes Project. Such projects are a key resource for researchers to find genetic mutations which are affecting health, and subsequently to consider options for their treatment. [24]

1.2 Transposable elements

Transposable elements (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.

1.2.1 **CLASS I (retrotransposons)**

Retrotransposons - are mobile genetic elements that use the method of "copy and paste" for propagation in the genome of animals. At least 45% of the human genome is retrotransposons and their derivatives (FIG 1.2). The characteristics of retrotransposons are very similar to retroviruses.

See the next chapter for more details.

1.2.2 **CLASS II (DNA transposons)**

DNA transposons to move inside genome use method "cut and paste" due to the complex enzyme called transposase [1]. Information on the amino acid sequence of the protein encoded transposase to the transposon sequences. Further, this piece of DNA may contain others related to transposon sequence, such genes or their parts. Most DNA transposons have partial sequence. These transposons are not autonomous and move around in the genome due to the transposase, which is encoded by another, complete DNA transposon.

On the ends of the DNA-transposon regions there are located inverted repeats that are specific transposase recognition sites, thereby distinguishing it from the rest of the genome. Transposase is capable of doing double-stranded DNA sections, cut and paste into the target DNA transposon [13]

Different type of transposable elements use a variety mechanisms for their evolutionary survival. LINEs and SINEs rely on vertical transmission within the host genome. DNA transposons are more disorderly, requiring relatively frequent horizontal transfer. LTR retrotransposons use both of previous type of transfer. [2]

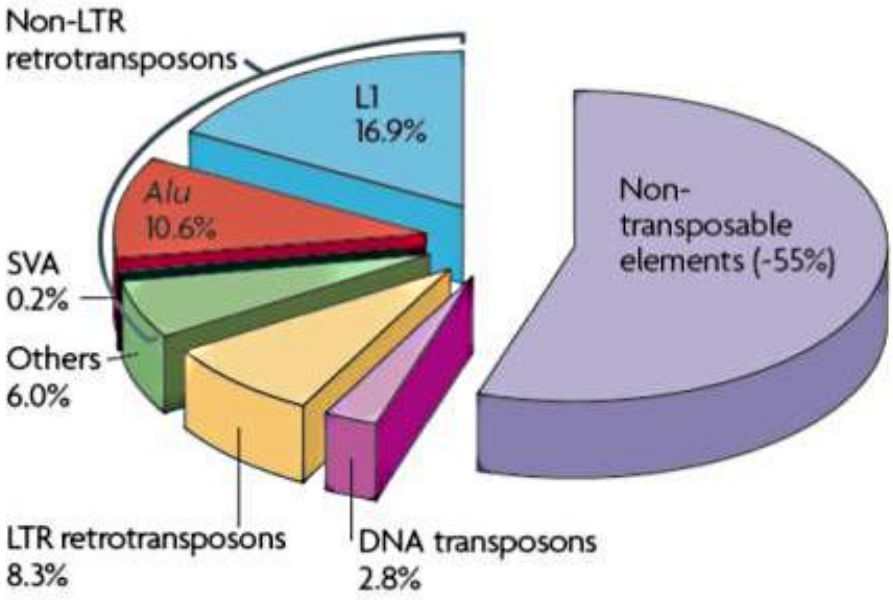


Figure 1.2 The transposable element content of the human genome

1.3 Retrotransposons

Retrotransposons usually consist of three sub-types (FIG 1.3):

- 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

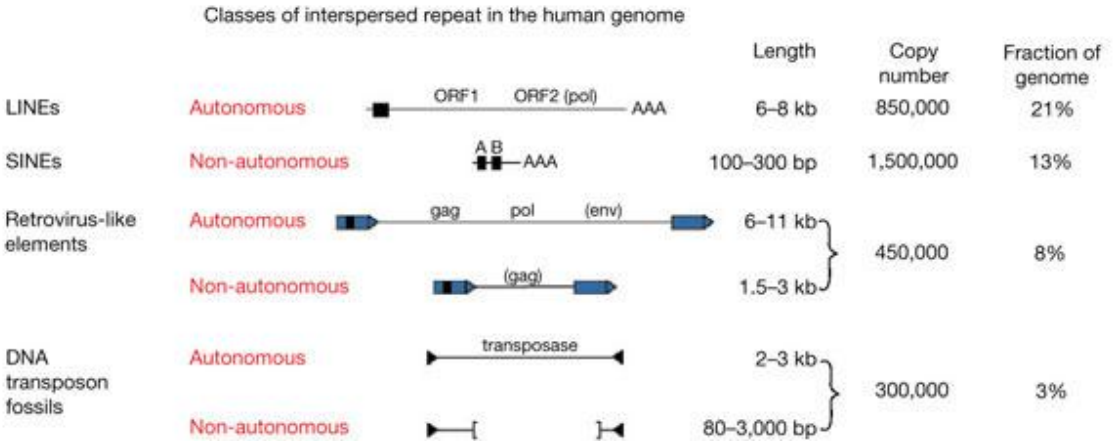


Figure 1.3 Classes of interspersed repeat in the human genome

1.3.1 LINEs (Long INterspersed Elements)

One of the most ancient element in eukaryotic genomes. These transposons are about 6 kb long (FIG 1.3), encode two ORF and contain an internal polymerase II promoter. After translation, a LINE RNA assembles with its own encoded proteins and moves to the nucleus. An endonuclease activity makes a single-stranded nick and the reverse transcriptase uses the nicked DNA to prime reverse transcription from the 3′ end of the LINE RNA [15]. The LINE machinery is probably responsible for most reverse transcription in the genome, including the creation of processed pseudogenes and the retrotransposition of the non-autonomous SINEs. There are three LINE families that have been found in the human genome: LINE1, LINE2 and LINE3, of which LINE1 is still active. [2]

1.3.2 SINEs (Short INterspersed Elements)

Because of SINEs are non-autonomous they require the presence of LINE elements to move. They are short (about 100–300 bp (FIG 1.3)), contain an internal polymerase III promoter and encode no proteins. Indeed, most SINEs ‘live’ by sharing the 3′ end with a resident LINE

element. The promoter regions of all known SINEs are derived from tRNA sequences, with the exception of a single monophyletic family derived from the signal recognition particle component 7SL. This family includes the only active SINE in the human genome: the Alu element. The human genome contains three distinct monophyletic families of SINEs: the active Alu, and the inactive MIR and Ther2/MIR3. [2]

1.3.3 **LTR retrotransposons**

LTR retrotransposons are surrounded by long terminal repeats that contain all of the necessary transcriptional regulatory elements. Exogenous retroviruses seem to have arisen from endogenous retrotransposons by acquisition of a cellular envelope gene (env). Transposition takes place through the retroviral mechanism with reverse transcription. Although a variety of LTR retrotransposons exist, only the vertebrate-specific endogenous retroviruses (ERVs) appear to have been active in the mammalian genome. Mammalian retroviruses fall into three classes (I–III), each comprising many families with independent origins. [2]

2 LTR retrotransposons

2.1 LTR retroelements families

LTR retrotransposons are divided into three subclasses (FIG. 2.1):

- Ty1-copia-like (Pseudoviridae)
- Ty3-gypsy-like (Metaviridae)
- Endogenous retroviruses (ERV)

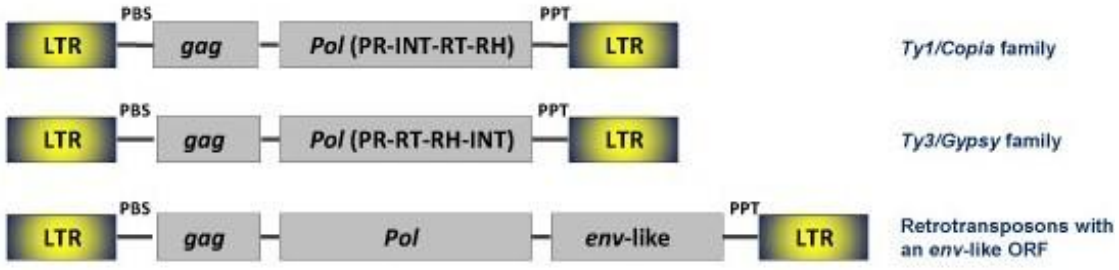


Figure 2.1 Genomic organization of different types of LTR retrotransposons [12]

2.1.1 Ty1/Copia

The Ty1/Copia represents one of the most important families of LTR retroelements in eukaryotes. Representation of Ty1/Copia LTR retroelements in the genomes of animals, fungi, plants, algae and several protists suggests that the ancestors of this family probably co-existed with the ancestors of Ty3/Gypsy LTR retroelements before the split between plants and unikonts. [12].

Genomic structure (FIG 2.1.1):

- A 5' long terminal repeat (LTR) of 100-1300 nt.
- A non-coding region that corresponds to the first portion of the retrotranscribed genome.
- A Primer Binding Site (PBS) of 18 nt.
- Open Reading Frames (ORFs) for gag, pol, (and env in retroviruses) genes.
- A small region of ~10 A/G "Polypurine Tract" (PPT).
- A 3' long terminal repeat (LTR) of 100-1300 nt

Ty1/Copia elements differ from other type of LTR retroelements in the position of integrase (INT) domain within pol polyprotein. While Bel/Pao, Retroviridae LTR retroelements show the INT at the C-terminus of pol (after RNase H), Ty1/Copia elements present INT N-terminal to the reverse transcriptase. [12].

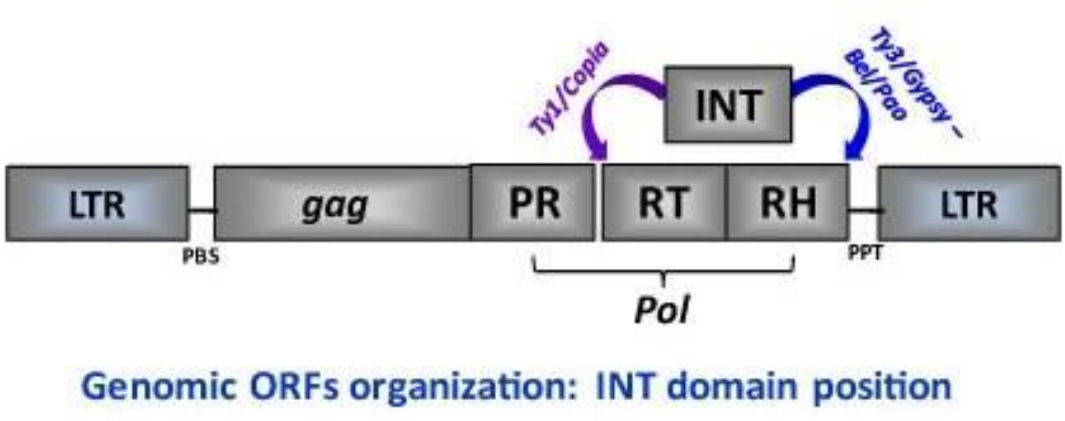


Figure 2.1.1 Genomic organization of Ty1/Copia, Ty3/Gypsy and Bel/Pao [12]

2.1.2 **Ty3/Gypsy**

Ty3/Gypsy LTR retroelements constitute a family of LTR retrotransposons and retroviruses widely distributed in plants, fungi and animals. They are provide a basis for understanding the evolutionary history of the LTR retroelement system. In fact, the main difference between a retrovirus and a LTR retrotransposon is that the retrovirus has an additional Open Reading Frame coding for an envelope (env) polyprotein necessary for transferring retroviruses cell-to-cell. [12].

Genomic structure (FIG 2.1.1):

- A 5' long terminal repeat (LTR) of 100-2000 nt.
- A non-coding region of 75-250 nt.
- A Primer Binding Site (PBS) of 18 nt.
- Open Reading Frames (ORFs) for *gag*, *pol*, (and *env* in retroviruses) genes.
- A small region of ~10 A/G "Polypurine Tract" (PPT).
- A 3' long terminal repeat (LTR) of 100-2000 nt

2.1.3 **Bel/Pao**

Bel/Pao LTR retroelements form a family of LTR retrotransposons and retroviruses described to date only in multicellular genomes.

Similarly to the Retroviridae and the Ty3/Gypsy, Bel/Pao LTR retroelements normally show a gag-pol genome (GAG-PR-RT-RH-INT) plus env (in the case of retroviruses) of variable size (between 4 and 10 Kb), surrounded by LIARs. [12].

Genomic structure (FIG 2.1.1):

- A 5' long terminal repeat (LTR) of 100-900 nt.
- A non-coding region of variable size.
- A Primer Binding Site (PBS) of 18 nt.
- Open Reading Frames (ORFs) for gag, pol, (and env in retroviruses) genes.
- A small region of ~10 A/G "Polypurine Tract" (PPT).
- A 3' long terminal repeat (LTR) of 100-900 nt.

2.1.4 **Retroviridae**

Vertebrate retroviruses (Retroviridae) are restricted to vertebrate animals. They are viral particles that reverse transcribe their RNA genome into a double stranded DNA copy that is inserted into the infected host cell genome. [12]

Retroviridae originally received attention when infectious representatives were characterized in humans. The Retroviridae display a gag-pol structure similar to that presented by Ty3/Gypsy LTR retroelements; the absence or presence of an env gene is the main difference between a Ty3/Gypsy LTR retrotransposon and a potential Ty3/Gypsy or Retroviridae simple retrovirus. [12]

Genomic structure (FIG 2.1.4):

- A 5'direct repeat (R) of 18-250 nt.
- A non-coding region of 75-250 nt (U5).
- A Primer Binding Site (PBS) of 18 nt.
- Open Reading Frames (ORFs) for gag, pol, and env genes and other accessory genes.
- A small region of ~10 A/G "Polypurine Tract" (PPT).
- A non-coding zone of 200-1.200 nt (U3).
- A 3'direct repeat (R) of 18-250 nt.

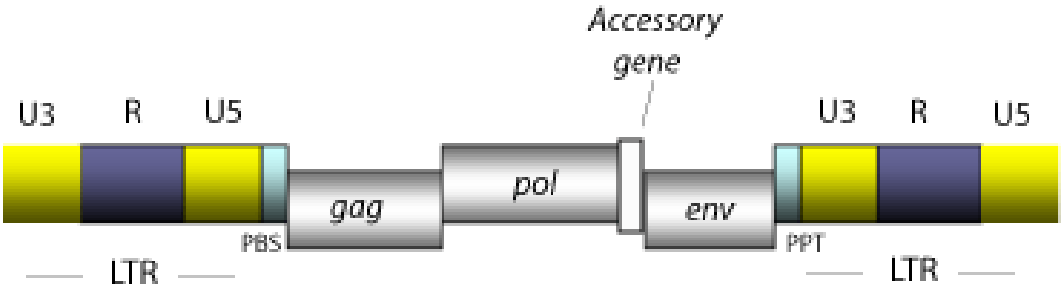


Figure 2.1.4 Genomic organization of Retroviridae [12]

Additional information

The gag gene codifies for a gag polyprotein containing the matrix (MA), capsid (CA) and the nucleocapsid (NC) domains.

The pol gene codifies for a pol polyprotein containing the protease (PR), reverse transcriptase/ribonuclease H (RT/RNaseH), and integrase (INT) domains.

The env gene codifies for the envelope (env) glycoprotein, which in the maturation process is spliced into the outer surface (SU) membrane protein (the main antigen of the viral envelope), and the transmembrane (TM) protein.

Human LTR elements are endogenous retroviruses which account for ~8% of the human genome. [1]. Retroviruses can transform into LTR retrotransposons by inactivation or by disposal of structures responsible for the extracellular mobility. If a retrovirus infects and then embeds itself into the genome in germ line cells, it can become an Endogenous Retrovirus (ERV). Therefore exogenous retroviruses arose from the acquisition of endogenous retrotransposons cellular envelope gene. [3]

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]

2.2 Participation retrotransposons in human pathogenesis

Over 25 experimentally characterized cellular genes show LTR-mediated evolutionary changes in which are embedded LTRs alternative promoters to provide a new tissue-specificity, play as the major promoters, or promote 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, despite the fact that in general the effect of the LTR insertions moderately manifested in many cases.

Recent studies have shown that HERV-encoded peptide as a tumor-specific antigen is involved 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 [7]. 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. [8]

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. [9] 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 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. [10].

Recent studies identified a human-specific endogenous retroviral insert (hsERV) that acts as an enhancer for human PRODH, hsERV_PRODH. PRODH encodes proline dehydrogenase, which is involved in neuromediator synthesis in the CNS. In this studies they detect high PRODH expression in the hippocampus, which was correlated with the undermethylated state of this enhancer. Because PRODH is associated with several

neurological disorders, they hypothesize that the human-specific regulation of PRODH by hsERV_PRODH may have played a role in human evolution by upregulating the expression of this important CNS-specific gene. [16] As of 2012 documented 96 different human diseases which are caused by de novo introduction of mobile genetic elements. Alu-repeats often cause chromosomal aberrations are the cause of 50 varieties of diseases. [14]

3 Algorithms of searching LTRs

3.1 The algorithm is based on a comparison with the reference database

RepeatMasker is a popular software tool widely used in computational genomics to identify, classify, and mask repetitive elements. RepeatMasker searches for repetitive sequence by aligning the input genome sequence against a library of known repeats, such as Repbase. Sequence comparisons in RepeatMasker are usually performed by the alignment program cross match, which requires significant processing time for larger sequences. [17]

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. [1]

3.2 De novo algorithms

LTR_STRUC is data-mining program that identifies and automatically analyzes LTR retrotransposons in genome databases by searching for structural features characteristic of such elements. LTR_STRUC has significant advantages over conventional search methods in the case of LTR retrotransposon families are having low sequence homology to known queries or families with atypical structure (e.g. non-autonomous elements lacking canonical retroviral ORFs). LTR_STRUC finds LTR retrotransposons by using an algorithm that encompasses a number of tasks that would otherwise have to be initiated individually by the user. For each LTR retrotransposon found, LTR_STRUC automatically generates an analysis of a variety of structural features of biological interest. [18]

RECON - an approach for the de novo identification and classification of repeat sequence families, based on extensions to the usual approach of single linkage clustering of local pairwise alignments between genomic sequences. This extensions use multiple alignment information to define the boundaries of individual copies of the repeats and to distinguish homologous but distinct repeat element families. This approach was able to properly identify and group known transposable elements, when was tested on the human

genome. The program, RECON, should be useful for first-pass automatic classification of repeats in newly sequenced genomes. [19].

RepeatScout algorithm is more sensitive and is orders of magnitude faster than RECON, the dominant tool for de novo repeat family identification in newly sequenced genomes. Using RepeatScout, was estimated that ~2% of the human genome consist of previously unannotated repetitive sequence. [20]

PILER is a new approach to de novo repeat annotation that use characteristic patterns of local alignments induced by certain classes of repeats. Novel repeats found using PILER are reported for Homo sapiens, Arabidopsis thaliana and Drosophila melanogaster. [21]

Many of these methods described above, however, only attempted to identify repeat elements based on their copy numbers in a genome, thus facilitating identification of general repeat elements. Many MGEs indeed appear high copies in the host genome because of their transposition activity. But some MGE families have low copy numbers in some genomes. As a result, successful identification of new MGEs by these bioinformatics approaches requires subsequent manual inspection. [1]

4 Implementation

4.1 Python in bioinformatics

One of the first Open Source languages to gain popularity among biologists was Perl. Perl based the foundation in bioinformatics with a help of strong text processing facilities, which were ideally suited for analyzing early sequence data. Perl has a history of successful use in bioinformatics and is still a very useful tool for biological research. [25]

In comparison to Perl, Python is a relative newcomer to bioinformatics, but is steadily gaining in popularity. A few of the reasons for this popularity are the:

- Readability of Python code
- Ability to development applications quickly
- Powerful standard library of functionality
- Scalability from very small to very large programs

The Python language was designed to be as simple and accessible as possible, without giving up any of the power needed to develop sophisticated applications. Python's clean, consistent syntax leaves it free from the subtleties and nuances that can make other languages difficult to learn and programs written in those languages difficult to comprehend. [25]

Python's dynamic nature adds to its accessibility. Python can be used interactively, allowing you to familiarize yourself with the language of any Python modules in an interactive session where each command produces immediate results.

Python also has excellent support for the object-oriented style of programming. As the data and analytical techniques used in bioinformatics have become more complex, the value of object-oriented language features has risen. [25]

In addition, Python integrates well with systems written in other languages, such as C, C++, Java and Fortran. One of the main benefits of C is speed. When a programmer needs an algorithm to run as fast as possible, they can code it in C or C++ and make it available to Python as an extension module.

So while Perl is more well established in the bioinformatics community, many biologists and bioinformaticians are also turning to Python as it gains in popularity. [25]

For this project used the following Python libraries:

- **Bio**

Biopython is a set of freely available tools for biological computation written in Python.

- **BCBio**

Toolkit for GFF parsing and writing.

- **Time**

Time toolbox used to calculate the time of the pattern search algorithms.

- **Shelve**

Shelve is a persistent, Python dictionary-like object, used to store some intermediate data like DNA sequences or array indices. This allows you to run parts of the program by passing all previous calculations.

4.2 De novo algorithm of LTRs search

Represented approach doesn't use the reference database with found LTRs, what allows finding new sites of LTR. The search of young LTR retrotransposons is divided to three stages:

4.2.1 **The first stage**

Searching for identical segments of the length of 40 bp is limited by minimum (1000bp - minimal length of entire of LTR) and maximum (10 kbp maximum length of entire of LTR) distances between them.

For this step was implemented several algorithms for searching the repeated patterns: KnuthMorrisPratt, Binary Search with LCP and Suffix array, searching with regular expressions and searching by built-in functions in Python. It was found that the fastest search function is embedded in Python, because it is C-implementation, on another hand other functions have written in Python.

The following figure shows the difference in execution time of different algorithms. Original algorithm is implemented by built-in Python functions for pattern search. (**FIG 4.2**).

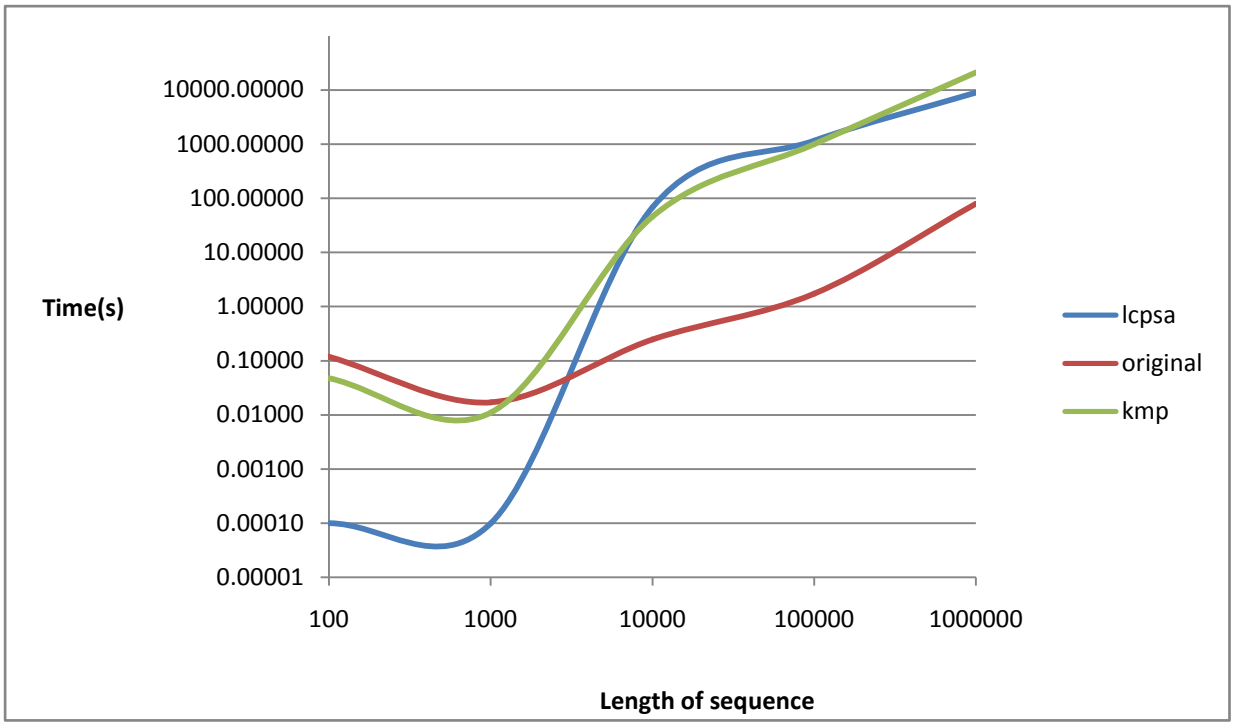


Figure 4.2 Comparison of the time of the algorithms work

Knuth–Morris–Pratt is efficient algorithm for searching for a substring in a string. The running time is linearly dependent on the amount of input data, so is impossible to develop asymptotically more efficient algorithm. The algorithm was developed by D. Knuth and B. Pratt and independently of them, D. Morris. The results of their work they published together in 1977.

The following is a sample pseudocode implementation of the KMP search algorithm:

```
algorithm kmp_search:
    input:
        an array of characters, S (the text to be searched)
        an array of characters, W (the word sought)
    output:
        an integer (the zero-based position in S at which W is found)

    define variables:
        an integer, m ← 0 (the beginning of the current match in S)
        an integer, i ← 0 (the position of the current character in W)
        an array of integers, T (the table, computed elsewhere)
    while m + i < length(S) do
        if W[i] = S[m + i] then
            if i = length(W) - 1 then
                return m
            let i ← i + 1
        else
            if T[i] > -1 then
                let m ← m + i - T[i], i ← T[i]
            else
                let i ← 0, m ← m + 1
    return the length of S
```

Binary search with Long Common Prefix and Suffix Array

The suffix array is a space-efficient data structure used, among others, in full text indices, data compression algorithms and within the field of bioinformatics. The suffix array allows efficient searching of a text for any given pattern, basically it is a sorted array *Pos* of all the suffixes of a text. A suffix array for a text of length *n* can work in O(*n log n*) time, and searching the text for a pattern of length m can be done in O(*m log n*) time by a binary search. When a suffix array is combined with information about the longest common prefixes of elements in the suffix array, string searches can be speeded up to O(*m*+*log n*) time. [22].

The following is a sample implementation of the Binary search algorithm with LCP and Suffix array:

```
def search(P):
     l = 0; r = n
     while l < r:
         mid = (l+r) / 2
         if P > suffixAt(A[mid]):
             l = mid + 1
         else:
             r = mid
     s = l; r = n
     while l < r:
         mid = (l+r) / 2
         if P < suffixAt(A[mid]):
             r = mid
         else:
             l = mid + 1
     return (s, r)
```

Search algorithm by Pythons built-in functions:

The simplest function using the built-in functions of Python language to search for a substring in the text:

```
def original_search(text, pattern)
   if pattern in text:
      return text.index(pattern)
```

4.2.2 The second stage

The main aim of this step is the formation of repetitive sequences in the groups associated with the individual LTR elements. These groups are based on the fact that the structure of formed sections must meet the structure of biological LTR retrotransposons. Further, these groups are written in form of 4 indexes are responsible for: leading the beginning of LTR, the leading end of the LTR, the beginning of the trailing LTR, end of trailing LTR. The algorithm takes into account the gene amplification and minimal/maximal length of LTR parts.

This is a part of *grouping.py* function, which shows the base of algorithm, which has been described above.

```
for lcp_part in db['young_lcp_parts'][1:]:
    if lcp_part[0] + min_pattern_len + min_distance > groups_of_ltrs[-
1][1][0]:
        if lcp_part[0] > groups_of_ltrs[-1][1][1]:
            if duplicates\
                    or (groups_of_ltrs[-1][0][1] - groups_of_ltrs[-
1][0][0]) < min_ltr_len \
                    or (groups_of_ltrs[-1][1][1] - groups_of_ltrs[-
1][1][0]) < min_ltr_len:
                groups_of_ltrs[-1] = [[lcp_part[0], lcp_part[0] +
min_pattern_len],
                                      [lcp_part[1], lcp_part[1] +
min_pattern_len]]
                duplicates = False
            else:
                groups_of_ltrs.append([[lcp_part[0], lcp_part[0] +
min_pattern_len],
                                       [lcp_part[1], lcp_part[1] +
min_pattern_len]])
        else:
            duplicates = True
    elif (lcp_part[0] - groups_of_ltrs[-1][0][0] < max_ltr_len) or \
            (lcp_part[1] - groups_of_ltrs[-1][1][0] < max_ltr_len):
        groups_of_ltrs[-1][0][1] = lcp_part[0] + min_pattern_len
        groups_of_ltrs[-1][1][1] = lcp_part[1] + min_pattern_len
    else:
        duplicates = True
if duplicates or (groups_of_ltrs[-1][0][1] - groups_of_ltrs[-1][0][0]) <
min_ltr_len \
        or min_ltr_len > (groups_of_ltrs[-1][1][1] - groups_of_ltrs[-
1][1][0]):
    del groups_of_ltrs[-1]
```

The main cycle iterates through found repetitive fragments (`db['young_lcp_parts']`) and adds them to current LTR element (`groups_of_ltrs[-1]`) until it's satisfies with the structure of real LTR retroelements. Further is going the forming of the next elements, while all of allocated fragments (`lcp_part`) will not be completed.

4.2.3 **The third stage**

The last step is the calculation of identity between LTRs for each set of indices obtained in the second stage using BLAST algorithm (NcbiblastnCommandline function from Biopython toolbox) and formation GFF file. GFF file contains the beginning and end of each LTR parts of retrotransposon and percentage identity between these LTRs.

Example of first several LTRs from GFF file for chrX (Feb. 2009 GRCh37/hg19):

```
##gff-version 3
##sequence-region chrX 1 155270560
```

```
chrX  ltrfind     SO:0000186  458782      462270      .     +     .
      ID=UnknownLTR_1;Note=identity 96.1538 %
chrX  ltrfind     SO:0000186  2120124     2130102     .     +     .
      ID=UnknownLTR_2;Note=identity 93.2203 %
chrX  ltrfind     SO:0000186  2349094     2359925     .     +     .
      ID=UnknownLTR_3;Note=identity 100 %
chrX  ltrfind     SO:0000186  2859399     2874732     .     +     .
      ID=UnknownLTR_4;Note=identity 97.5 %
chrX  ltrfind     SO:0000186  3271550     3287635     .     +     .
      ID=UnknownLTR_5;Note=identity 100 %
```

4.3 Results

Verification of the algorithm was carried out on human chrX (Feb. 2009 GRCh37/hg19). A total 396 LTR elements were identified. Histogram below represents the number of found elements within certain ranges of length (**FIG 4.3**):

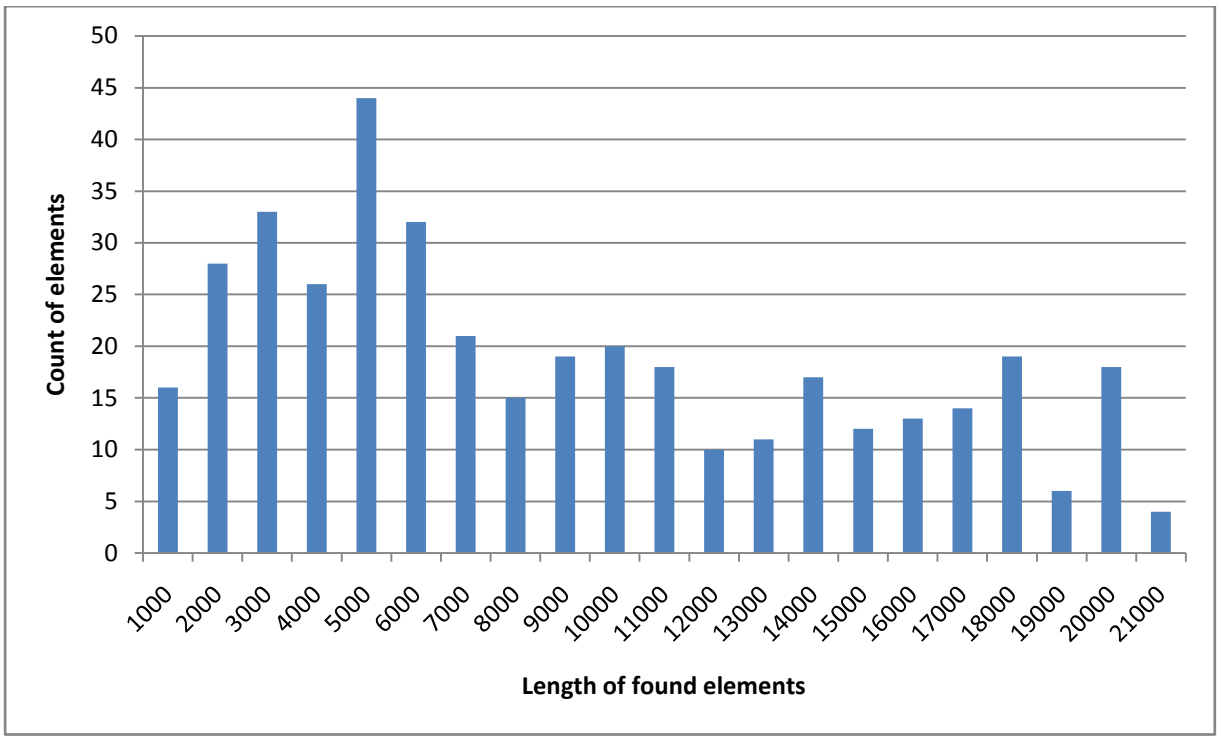


Figure 4.3 Histogram of found elements

The vast majority of found items have a length from 2 to 7 kbp, that approximately corresponds to the structure of LTR elements. Longer sites represent modified elements such as embedded LTRs, whose length can be much longer than in younger items.

The following are some figures of comparing sites of chromosome and the output of algorithm, which was described above, in UCSC genome browser (**FIG 4.4**):

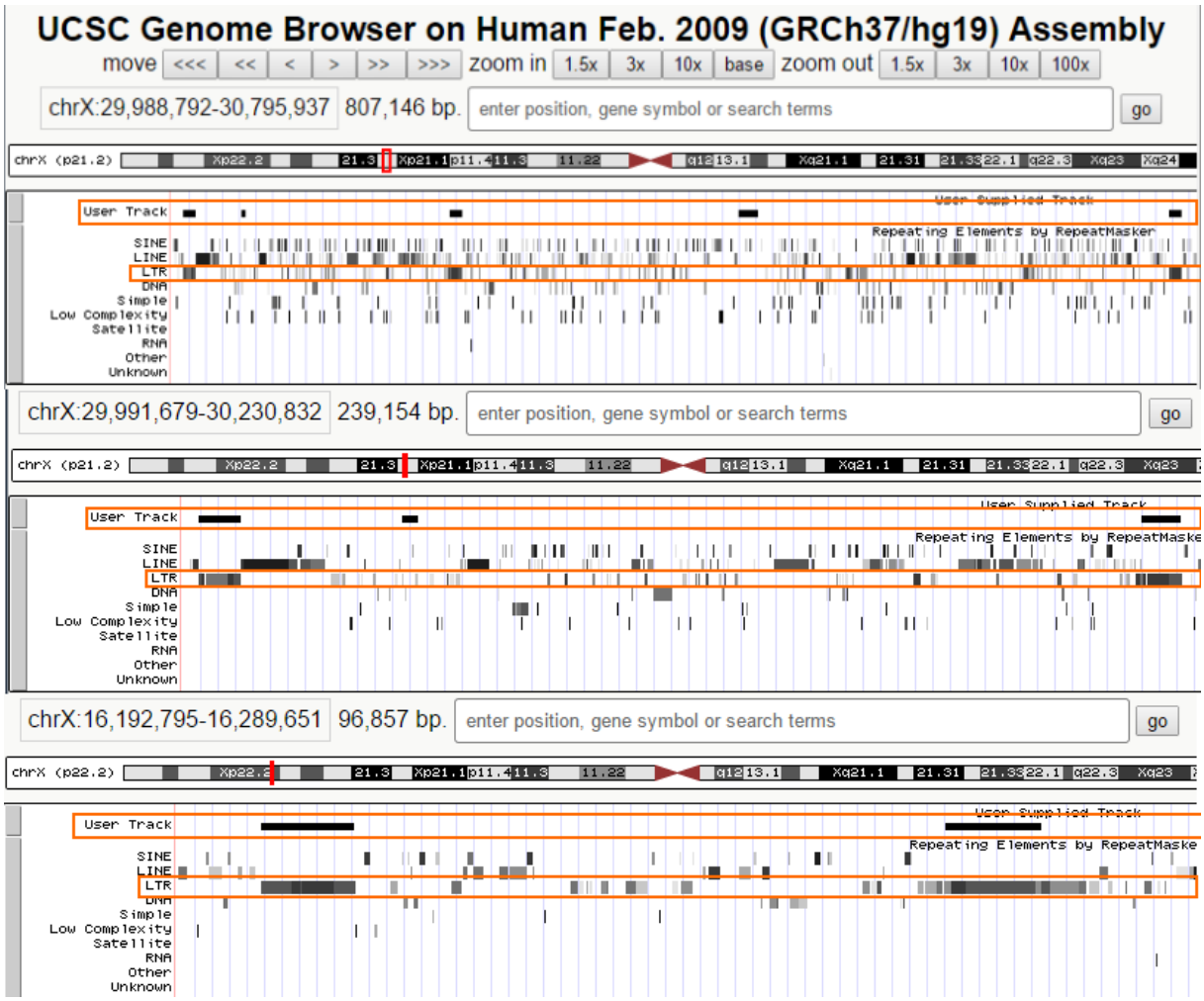


Figure 4.4 UCSC genome browser comparison

As you can see, there is detected the correct elements and elements that were not found or were incorrectly identified like LTRs.

Conclusions

In the first part I'm focusing on describing the genetic material. This is important for understanding the disease, which can cause by human endogenous retroviruses. A good knowledge of their structure and origin are essential for the design of efficient de novo searching algorithm.

In the most important part, I implemented de novo method to search for young LTR retroelements. As the basis of the algorithm searching for repetitive patterns and their subsequent association in segments corresponding to the structure of the required elements was taken. The vast majority of detected LTR retrotransposons have identities more than 80%. This shows the flexibility of the algorithm to search for the young LTRs. This algorithm is aimed only to search young LTRs, that is a small number of elements relative to already found (references database). However, there is still place for improvement, such as increasing speed of the patterns search or increase the number of found retroelements by finding the optimal parameters to search.

In continuation of this project will be compared the structure of found elements with a structure of separate families of LTR retrotransposons for the distribution to certain families (Ty1/Copia, Ty3/Gypsy, Retroviridae). To improve the searching algorithm the basic function will be interpreted in C ++. The statistical evaluation of the results will be achieved by comparing with results from other approaches for LTR identifications and reference databases.

References

[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-. DOI: 10.1186/1471-2164-8-90.

[2] LANDER, Eric S. et al. Initial sequencing and analysis of the human genome. Nature. 2001-2-15, vol. 409, issue 6822, s. 860-921. DOI: 10.1038/35057062.

[3] MALIK, H. S. Poised for Contagion: Evolutionary Origins of the Infectious Abilities of Invertebrate Retroviruses. Genome Research. vol. 10, issue 9, s. 1307-1318. DOI: 10.1101/gr.145000.

[4] COHEN, Carla J., Wynne M. LOCK a Dixie L. MAGER. Endogenous retroviral LTRs as promoters for human genes: A critical assessment. Gene. 2009, vol. 448, issue 2, s. 105-114. DOI: 10.1016/j.gene.2009.06.020.

[5] BIECHE, I. Placenta-Specific INSL4 Expression Is Mediated by a Human Endogenous Retrovirus Element. Biology of Reproduction. 2002, vol. 68, issue 4, s. 1422-1429. DOI: 10.1095/biolreprod.102.010322.

[6] LANDRY, J.-R. a D. L. MAGER. Functional Analysis of the Endogenous Retroviral Promoter of the Human Endothelin B Receptor Gene. Journal of Virology. 2003, vol. 77, issue 13, s. 7459-7466. DOI: 10.1128/jvi.77.13.7459-7466.2003.

[7] TAKAHASHI, Yoshiyuki et al. Regression of human kidney cancer following allogeneic stem cell transplantation is associated with recognition of an HERV-E antigen by T cells. Journal of Clinical Investigation. s. -. DOI: 10.1172/JCI34409.

[8] LAMPRECHT, Björn et al. . Derepression of an endogenous long terminal repeat activates the CSF1R proto-oncogene in human lymphoma. Nature Medicine. 2010-5-2, vol. 16, issue 5, s. 571-579. DOI: 10.1038/nm.2129.

[9] BELSHAW, R., A. L. A. DAWSON, J. WOOLVEN-ALLEN, J. REDDING, A. BURT a M. TRISTEM. Genomewide Screening Reveals High Levels of Insertional Polymorphism in the Human Endogenous Retrovirus Family HERV-K(HML2): Implications for Present-Day Activity. Journal of Virology. 2005-09-13, vol. 79, issue 19, s. 12507-12514. DOI: 10.1128/JVI.79.19.12507-12514.2005.

[10] KATOH, Iyoko a Shun-ichi KURATA. Association of Endogenous Retroviruses and Long Terminal Repeats with Human Disorders. Frontiers in Oncology. 2013, vol. 3, s. -. DOI: 10.3389/fonc.2013.00234.

[11] CORDAUX, Richard a Mark A. BATZER. The impact of retrotransposons on human genome evolution. Nature Reviews Genetics. 2009, vol. 10, issue 10, s. 691-703. DOI: 10.1038/nrg2640.

[12] LLORENS, C., R. FUTAMI et al.. The Gypsy Database (GyDB) of mobile genetic elements: release 2.0. Nucleic Acids Research. 2010-12-22, vol. 39, Database, D70-D74. DOI: 10.1093/nar/gkq1061.

[13] VAN OPIJNEN, Tim a Andrew CAMILLI. Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms. Nature Reviews Microbiology. 2013-5-28, vol. 11, issue 7, s. 435-442. DOI: 10.1038/nrmicro3033.

[14] HANCKS, Dustin C a Haig H KAZAZIAN. Active human retrotransposons: variation and disease. Current Opinion in Genetics. 2012, vol. 22, issue 3, s. 191-203. DOI: 10.1016/j.gde.2012.02.006.

[15] REILLY, M. T., G. J. FAULKNER, J. DUBNAU, I. PONOMAREV a F. H. GAGE. The Role of Transposable Elements in Health and Diseases of the Central Nervous System. Journal of Neuroscience. 2013-11-06, vol. 33, issue 45, s. 17577-17586. DOI: 10.1523/JNEUROSCI.3369-13.2013.

[16] SUNTSOVA, M. et al. Human-specific endogenous retroviral insert serves as an enhancer for the schizophrenia-linked gene PRODH. Proceedings of the National Academy of Sciences. 2013-11-26, vol. 110, issue 48, s. 19472-19477. DOI: 10.1073/pnas.1318172110.

[17] RepeatMasker [online]. [cit. 2015-01-17]. Available from: http://www.repeatmasker.org/

[18] MCCARTHY, E. M. a J. F. MCDONALD. LTR_STRUC: a novel search and identification program for LTR retrotransposons. Bioinformatics. 2003-02-12, vol. 19, issue 3, s. 362-367. DOI: 10.1093/bioinformatics/btf878.

[19] BAO, Z. Automated De Novo Identification of Repeat Sequence Families in Sequenced Genomes. Genome Research. vol. 12, issue 8, s. 1269-1276. DOI: 10.1101/gr.88502.

[20] PRICE, A. L., N. C. JONES a P. A. PEVZNER. De novo identification of repeat families in large genomes. Bioinformatics. 2005-06-16, vol. 21, Suppl 1, i351-i358. DOI: 10.1093/bioinformatics/bti1018.

[21] EDGAR, R. C. a E. W. MYERS. PILER: identification and classification of genomic repeats. Bioinformatics. 2005-06-16, vol. 21, Suppl 1, i152-i158. DOI: 10.1093/bioinformatics/bti1003.

[22] Kasai T, Lee G, Arimura H, Arikawa S, Park K: Linear-time longest common-prefix computation in suffix arrays and its applications.: Jerusalem, Israel. Volume 2089. Springer-Verlag; 2002::181-192. [Lecture Notes in Computer Science]

[23] SERVICE, R. F. GENE SEQUENCING: The Race for the $1000 Genome. *Science*. 2006-03-17, vol. 311, issue 5767, s. 1544-1546. DOI: 10.1126/science.311.5767.1544.

[24] BUCHANAN, C. C., E. S. TORSTENSON, W. S. BUSH a M. D. RITCHIE. A comparison of cataloged variation between International HapMap Consortium and 1000 Genomes Project data: The Race for the $1000 Genome. *Journal of the American Medical Informatics Association*. 2012-03-01, vol. 19, issue 2, s. 289-294. DOI: 10.1136/amiajnl-2011-000652.

[25] O'BRIEN, Patrick. Beginning Python for Bioinformatics. [online]. [cit. 2015-02-01]. Available from: http://www.oreilly.com/pub/au/966/

List of abbreviations

DNA	Deoxyribonucleic acid
ERVs	Endogenous retroviruses
GFF	Gene-finding format, generic feature format
HERVs	Human endogenous retroviruses
IN	Integrase
LARDs	Large retrotransposons derivates
LCP	Long common prefix
LINE	Long interspersed elements
LTR	Long terminal repeats
MaLRs	Mammalian apparent LTR retrotransposons
MGE	Mobile genetic elements
ORF	Open reading frame
PBS	Primer-binding site
PPT	Polypurine tract
PR	Protease
RCC	Renal cell carcinoma
RNA	Ribonucleic acid
RT	Reverse transcriptase
SINE	Short interspersed elements
TEs	Transposable elements
TRIMs	Terminal-repeat retrotransposons in miniature

List of figures

List of attachments

AttachmentA - source code of whole project

- run.py – main file
- arguments.py - processing command line arguments
- algorithms.py - basic used algorithms
- grouping.py - grouping of patterns
- README.md – program instructions

AttachmentB – GFF output of program for human chrX (Feb. 2009 GRCh37/hg19)