De novo identification of repeat families in large genomes

1. 요약

반복서열은 진핵 생물의 유전체 서열 중 상당 부분을 차지하고 있다. 예쁜꼬마선충의 20%, 인간 유전체의 50% 이상이 반복 서열로 구성되어 있다. 또한 Transposon elements를 포함하여 simple repeat region, low complexity 영역이 전체 유전체의 약 70-80% 가량 해당된다. 따라서 반복서열을 선별하는 일은 유전체 분석에서 매우 중요한 부분이다. 본 보고서에서는 반복 서열을 정의하기 위한 alignment 기법인 Repeat Scout (De novo identification of repeat families in large genomes, 2005) 알고리즘을 활용, RepeatMasker의 라이브러리를 생성하고 반복 서열의 마스킹 작업을 진행 과정을 알아보기로 한다.

2. 개요

전반부에서는 RepeatMasker 설치하고 해당 프로그램을 실행하기 위한 Repbase을 승인/설치하는 과정을 다루고 후반부에서는 RepeatScout을 설치/실행하여 RepeatMasker에서 활용한 라이브러리 파일을 생성하는 과정을 다룬다. 이후에 RepeatScout 알고리즘을 Motif finding에도 적용 가능한지 시뮬레이션을 실행하여 가능성을 알아본다.

De novo repeat family identification

Introduction : De novo repeat family identification

RepeatScout algorithm을 활용한 De novo repeat family 정의 방법에 대한 논문

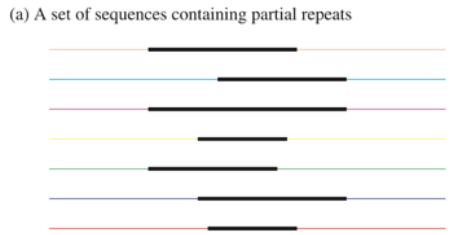
배경 : Repetitive DNA 구성은 다양한 생물에 많은 퍼센테이지를 차지하지만 반복은 알려져있지 않고 따라서 masking할 수 없기 때문에 비슷한 두 종의 유전체 비교/분석에 어려움

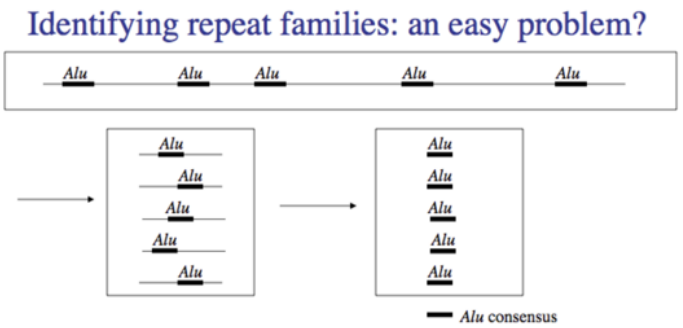
목적 : RepeatScout algorithm을 활용하여 De novo repeat family을 정의하고 이를 통해 repeat family를 마스킹, most single-species 혹은 multi-species 분석을 진행

Ref 01) RepeatMasker 마스킹 툴(http://www.repeatmasker.org)

What is repeat family?

A repeat family is "a collection of similar sequences" which appear many times in a genome.
For example, the Alu repeat family has over 1 million approximate occurrences in the human genome.





- Alignment 이후의 Alu consensus(오른쪽 그림)

3. RepeatMasker 설치 및 실행

RepeatMasker는 유사성 기반의 검색을 통해 반복서열 데이터베이스에 존재하는 서열과 비교하여 유전체 내에 존재하는 transposon element와 retrotransposon element, rolling circles를 추출하고, TRF(tandem repeat finder)라는 서브 프로그램에 의해 단순반복 서열을 규명한다. 이때 종별로 특이적인 패턴을 가지는 반복서열이 존재하므로 주기적으로 최신의 반복서열 데이터베이스를 업데이트하여 분석하는 것이 좋다. 반복서열 데이터 베이스 RepBase는 http://www.girinst.org/ 에서 제공되며, 2018년 6월 현재 Human을 포함하여 모두 43종에 대한 반복서열 데이터베이스를 제공하고 있으며, 연구자가 원하는 형태의 데이터베이스를 따로 구성하여 사용할 수도 있다. RepeatMasker에서 분석가능한 형태의 데이터베이스는 모두 fasta foramt이며 ; 혹은 > 이후 해당 시퀀스가 시작되는 것이 특징이다.

3-1. RepeatMasker 설치

Prerequisite : C compiler, make, Sequence Search Engine(BLAT, Cross Match, RMBlast or WU-Blast), Repeat Database(Repbase: https://www.girinst.org)

RepeatMasker-open3.0 와 Prerequisites

3-1-1) RMBlast(http://www.repeatmasker.org/RMBlast.html)

기본 국소정렬 검색 도구인 BLAST(the Basic Local Alignment Search Tool, BLAST)는 서열의 유사성을 밝히는 데 가장 많이 사용되는 방법이며, 블라스트 프로그램은 사용자들이 제공한 검색 대상 서열에 대하여 NCBI 의 전체 데이터베이스를 대상으로 하여 검색을 수행한다.

Method a) source for RMBlast(소스코드 직접 다운로드 받아 컴파일하는 방식)
The BLAST+ source contains both packages in one distribution(ncbi-blast-2.6.0+-src.tar.gz, isb-2.6.0+-changes-vers2.patch.gz)

Method b) pre-compiled binaries for RMBlast(pre-compiled 된 소스 코드를 받아서 설치하는 방식)

1) main BLAST+ toolkit(BLAST+ Binaries:ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/2.2.28/)
2) new RMBlast application(RMBlast Binaries:ftp://ftp.ncbi.nlm.nih.gov/blast/executables/rmblast/2.2.28)

b-1)
BLAST+ toolkit :
curl -O
ftp://ftp.ncbi.nlm.nih.gov//blast/executables/blast+/LATEST/ncbi-blast-2.7.1+-x64-macosx.tar.gz

b-2)
RMBlast application :
curl -O
ftp://ftp.ncbi.nlm.nih.gov//blast/executables/rmblast/LATEST/ncbi-rmblastn-2.2.28-universal-macosx.tar.gz

설치
tar -xvzf ncbi-blast-2.2.28+-universal-macosx.tar.gz
tar -xvzf ncbi-rmblastn-2.2.28-universal-macosx.tar.gz

cp -R ncbi-rmblastn-2.2.28/* ncbi-blast-2.2.28+/
rm -rf ncbi-rmblastn-2.2.28
mv ncbi-blast-2.2.28+ ncbi-rmblastn-2.2.28

3-1-2) Repbase

다음과 같이 Repbase, RepeatMasker, 그리고 RMBlast의 버전을 일치시킨다. 버전간 호환이 불가하기 때문에 Repbase, RepeatMasker, RMBlast의 버전을 아래와 같이 일치시켜야 에러가 발생하지 않는다. 따로 사용자가 설정하지 않으면 Repeatmaskerlibraries-20140131.tar.gz가 RepeatMasker가 참조하는 default library가 된다. 사용자 정의 RepeatMasker library 설정은 프로그램 실행시 -lib 옵션을 통해 가능하다.

Repbase : repeatmaskerlibraries-20140131.tar.gz (52.85 MB)
RepeatMasker : RepeatMasker-open-4-0-5.tar.gz
RMBlast : ncbi-blast-2.2.28+-src.tar.gz
rmblast-2.2.28-src.tar.gz

RepeatMasker를 실행하기위해서는repetitive elements consensus sequences를포함하는Repeat Library가 필요하다. 현재 사용되고 있는 Repeat Librarys는Repbase Update 라이브러리이다. 해당되는 라이브러리는 가장 상용적으로 사용되고 있으며, human, rodent(설치류), zebrafish, Drosophila(초파리), Arabidopsis thaliana의 종을 포함.

tar -xvzf repeatmaskerlibraries-20140131.tar.gz
ln -s Libraries/ current

3-1-3) RepeatMasker 설치

curl -O http://repeatmasker.org/RepeatMasker-open-4-0-5.tar.gz
tar -xvzf RepeatMasker-open-4-0-5.tar.gz

curl -O http://repeatmasker.org/RepeatMasker-open-4-0-7.tar.gz
curl -C - -O http://repeatmasker.org/RepeatMasker-open-4-0-7.tar.gz
tar -xvzf RepeatMasker-open-4-0-7.tar.gz

mv RepeatMasker RepeatMasker_4.0.5

ln -s RepeatMasker_4.0.5 current_4.0.5
cd RepeatMasker_4.0.5/Libraries

mv RepeatMaskerLib.embl RepeatMaskerLib.embl.backup
ln -s ../../../Repbase/current/RepeatMaskerLib.embl

기존에 있던 RepeatMaskerLib 파일은 backup으로 이름을 바꾸고 2) Repbase 과정에서 다운 받았던 Repeatmaskerlibraries-20140131.tar.gz의 RepeatMaskerLib.embl로 대체. Repbase의 RepeatMaskerLib의 기본 파일명이 RMRBSeqs.embl인 경우, 파일명을 RepeatMaskerLib.embl로 변경.

3-2. RepeatMasker 실행

RepeatMasker input_genome_sequence.fas -lib output_repeats.fas.filtered_1

time
/Users/inmobi/Desktop/Korflab/Packages/RepeatMasker/RepeatMasker_4.0.5/RepeatMasker
/Users/inmobi/Desktop/Korflab/Packages/RepeatScout-1
/human3M.fa -lib /Users/inmobi/Desktop/Korflab/Packages/RepeatScout-1
/human3M_out.fa human3M_out_RepeatMasker.fas

4. RepeatScout 설치 및 실행

RepeatScout 알고리즘이 적용된 RepeatScout 프로그램은 RepeatMasker가 참고할 라이브러리 파일을 생성한다. 입력 파일은 fasta 형식이다. 서브 프로그램을 순차적으로 실행하여 RepeatMasker의 라이브러리 파일을 출력한다. 이후 RepeatMasker의 -lib 옵션을 통해 해당 라이브러리를 참고 RepeatScout 알고리즘으로 반복서열을 masking 할 수 있다.

4-1. RepeatScout 설치

prerequisite
necessary : make, a C compiler,

perl 5.5, nseg (Wooton and Federhen, 1993), trf (Benson, 1999) to filter your repeat library to remove low-complexity and tandem sequences.
RepeatMasker-open3.0 to filter your repeat library against segmental duplications, exons, or other features.

4-2. RepeatScout 실행

#1. build_lmer_table
Running RepeatScout proceeds in four phases. First, build_lmer_table creates a file that tabulates the frequency(TABLE) of all l-mers in the sequence to be analyzed.

build_lmer_table -sequence input_genome_sequence.fas -freq output_lmer.frequency

ex)
time /Users/inmobi/Desktop/Korflab/Packages/RepeatScout-1/build_lmer_table -sequence human3M.fa -freq human3M.freq

#2. RepeatScout

Second, RepeatScout takes this table(output_lmer.frequency) and the sequence(input_genome_sequence.fas) and produces a fasta file that contains all the repetitive elements that it could find.

RepeatScout -sequence input_genome_sequence.fas -output output_repeats.fas -freq output_lmer.frequency

ex)
time /Users/inmobi/Desktop/Korflab/Packages/RepeatScout-1/RepeatScout -sequence human3M.fa -output human3M_out.fa -freq human3M.freq

#3. filter-stage-1.prl

Third, the "filter-stage-1.prl" script is run on the output of RepeatScout to remove low-complexity and tandem elements;
filter-stage-1.prl output_repeats.fas > output_repeats.fas.filtered_1

ex)
time /usr/local/repeatscout/latest/filter-stage-1.prl human3M_out.fa > human3M_out.fa.filtered_1

#5. filter-stage-2.prl
cat output_repeats.fas.filtered_1 | filter-stage-2.prl --cat=input_genome_sequence.fas.out > output_repeats.fas.filtered_2

ex)
time /usr/local/repeatscout/latest/filter-stage-2.prl

결과

```
==================================================
file name: angrep.fa
sequences:           382
total length:     960504 bp  (959923 bp excl N/X-runs)
GC level:         43.34 %
bases masked:      39822 bp ( 4.15 %)
==================================================
               number of      length   percentage
               elements*    occupied  of sequence
--------------------------------------------------
SINEs:                0            0 bp    0.00 %
      ALUs            0            0 bp    0.00 %
      MIRs            0            0 bp    0.00 %

LINEs:                0            0 bp    0.00 %
      LINE1           0            0 bp    0.00 %
      LINE2           0            0 bp    0.00 %
      L3/CR1          0            0 bp    0.00 %

LTR elements:         0            0 bp    0.00 %
      ERVL            0            0 bp    0.00 %
      ERVL-MaLRs      0            0 bp    0.00 %
      ERV_classI      0            0 bp    0.00 %
      ERV_classII     0            0 bp    0.00 %

DNA elements:         0            0 bp    0.00 %
     hAT-Charlie      0            0 bp    0.00 %
     TcMar-Tigger     0            0 bp    0.00 %

Unclassified:        73        28426 bp    2.96 %

Total interspersed repeats:    28426 bp    2.96 %


Small RNA:            0            0 bp    0.00 %

Satellites:           0            0 bp    0.00 %
Simple repeats:     207        11583 bp    1.21 %
Low complexity:       0            0 bp    0.00 %
==================================================
```

Figure. RepeatScout

```
==================================================
file name: angrep.fa
sequences:           382
total length:     960504 bp  (959923 bp excl N/X-runs)
GC level:         43.34 %
bases masked:      18050 bp ( 1.88 %)
==================================================
               number of      length   percentage
               elements*    occupied  of sequence
--------------------------------------------------
SINEs:                0            0 bp    0.00 %
      ALUs            0            0 bp    0.00 %
      MIRs            0            0 bp    0.00 %

LINEs:                9         2804 bp    0.29 %
      LINE1           1           89 bp    0.01 %
      LINE2           0            0 bp    0.00 %
      L3/CR1          7          474 bp    0.05 %

LTR elements:        14         3242 bp    0.34 %
      ERVL            0            0 bp    0.00 %
      ERVL-MaLRs      0            0 bp    0.00 %
      ERV_classI      1           57 bp    0.01 %
      ERV_classII     0            0 bp    0.00 %

DNA elements:         4          451 bp    0.05 %
     hAT-Charlie      2          112 bp    0.01 %
     TcMar-Tigger     0            0 bp    0.00 %

Unclassified:         1          128 bp    0.01 %

Total interspersed repeats:     6625 bp    0.69 %


Small RNA:            0            0 bp    0.00 %

Satellites:           0            0 bp    0.00 %
Simple repeats:     173         9888 bp    1.03 %
Low complexity:      28         1537 bp    0.16 %
==================================================
```

Figure. RepBase library

Anopheles gambiae는 말라리아의 전염성 바이러스 유전자이다. Anopheles gambiae의 시퀀스를 분석한 결과 RepeatScout의 경우 세부 항목별 elements까지 세분화 시키지는 못한다. GC level에서는 43.4% 로 Repbase Lib를 사용했을 때와 같지만 based masked를 참고하면 Repbase Lib를 사용했을 때보다 2.27% 가량 더 많은 반복서열을 masking 할 수 있는 것을 확인할 수 있다.

```
==================================================
file name: angrep.fa
sequences:            382
total length:     960504 bp  (959923 bp excl N/X-runs)
GC level:         43.34 %
bases masked:      18050 bp ( 1.88 %)
==================================================
               number of      length   percentage
               elements*    occupied  of sequence
--------------------------------------------------
SINEs:                0           0 bp    0.00 %
      ALUs            0           0 bp    0.00 %
      MIRs            0           0 bp    0.00 %

LINEs:                9        2804 bp    0.29 %
      LINE1           1          89 bp    0.01 %
      LINE2           0           0 bp    0.00 %
      L3/CR1          7         474 bp    0.05 %

LTR elements:        14        3242 bp    0.34 %
      ERVL            0           0 bp    0.00 %
      ERVL-MaLRs      0           0 bp    0.00 %
      ERV_classI      1          57 bp    0.01 %
      ERV_classII     0           0 bp    0.00 %

DNA elements:         4         451 bp    0.05 %
     hAT-Charlie      2         112 bp    0.01 %
     TcMar-Tigger     0           0 bp    0.00 %

Unclassified:         1         128 bp    0.01 %

Total interspersed repeats:     6625 bp    0.69 %


Small RNA:            0           0 bp    0.00 %

Satellites:           0           0 bp    0.00 %
Simple repeats:     173        9888 bp    1.03 %
Low complexity:      28        1537 bp    0.16 %
==================================================
```

Figure. RepeatScout

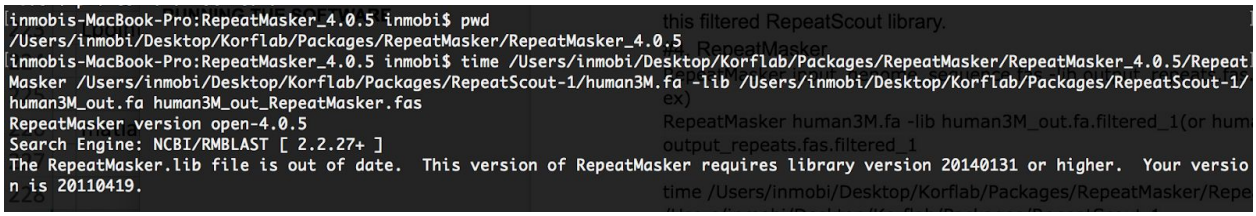
```
==================================================
file name: athrep.fa
sequences:            525
total length:    1625632 bp  (1625632 bp excl N/X-runs)
GC level:         39.29 %
bases masked:      24762 bp ( 1.52 %)
==================================================
               number of      length   percentage
               elements*    occupied  of sequence
--------------------------------------------------
SINEs:                0           0 bp    0.00 %
      ALUs            0           0 bp    0.00 %
      MIRs            0           0 bp    0.00 %

LINEs:                1          69 bp    0.00 %
      LINE1           0           0 bp    0.00 %
      LINE2           0           0 bp    0.00 %
      L3/CR1          1          69 bp    0.00 %

LTR elements:         2         278 bp    0.02 %
      ERVL            0           0 bp    0.00 %
      ERVL-MaLRs      0           0 bp    0.00 %
      ERV_classI      0           0 bp    0.00 %
      ERV_classII     0           0 bp    0.00 %

DNA elements:         2          81 bp    0.00 %
     hAT-Charlie      0           0 bp    0.00 %
     TcMar-Tigger     0           0 bp    0.00 %

Unclassified:         5         831 bp    0.05 %

Total interspersed repeats:     1259 bp    0.08 %


Small RNA:            1          52 bp    0.00 %

Satellites:           0           0 bp    0.00 %
Simple repeats:     380       18913 bp    1.16 %
Low complexity:      86        4566 bp    0.28 %
==================================================
```

Figure. RebBase library

athrep의 경우에는 세부 분류에서 약간의 차이 보인다. 에를 들면 LINEs 에서 RepeatScout의 경우에는 0.35%를 masking하는 RepBase의 경우에는 0.01%에도 못미치는 반복서열을 masking 한다. GC level, bases masked 역시 각각 4.05% 0.36% RepeatScout 알고리즘이 더 많은 반복서열을 masking하는 것을 확인할 수 있다.

errors

01.

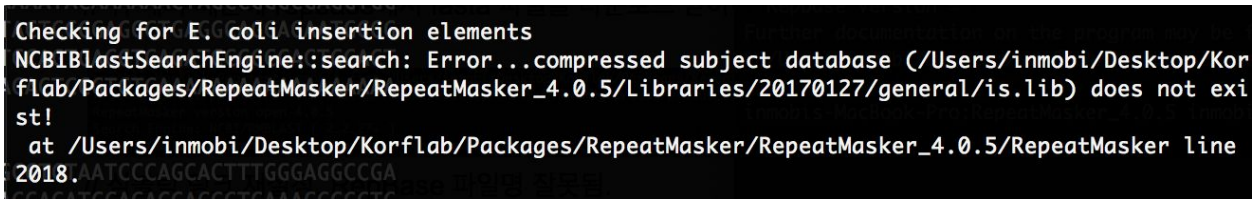


// 위 에러 때문에 repbase에서 fasta 파일을 다운로드 받아야함. 버전 호환 관련 오류

02.

// 심볼릭 링크 재설정. RepBase 파일명 잘못됨.
// ln -s ../../../Repbase/current/RMRBSeqs.embl

03.

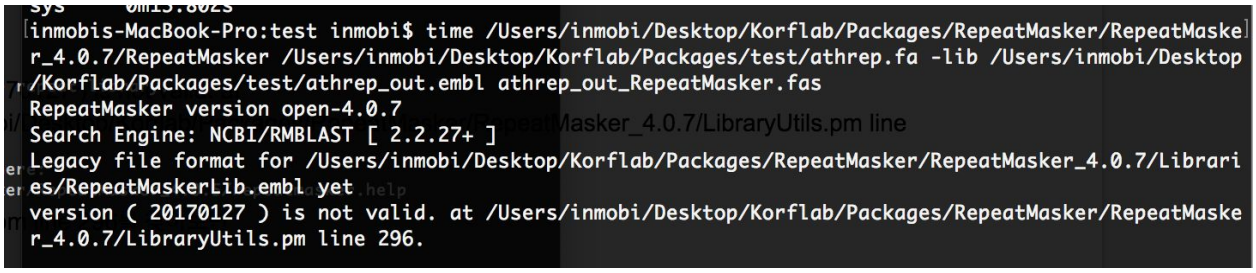


// 특정 파일을 참조하지 못해서 발생하는 에러.

log
Checking for E. coli insertion elements
NCBIBlastSearchEngine::search: Error...compressed subject database (/Users/inmobi/Desktop/Korflab/Packages/RepeatMasker/RepeatMasker_4.0.5/Libraries/20170127/general/is.lib) does not exist!
at /Users/inmobi/Desktop/Korflab/Packages/RepeatMasker/RepeatMasker_4.0.5/RepeatMasker line 2018.
해결방법으로 다음 링크를 참조 해결함.
https://groups.google.com/forum/#!topic/maker-devel/asS5g0f-BE8
I matched the RepeatMasker 4.0.7 with RepBase20170127 and it worked!
현재버전 4.0.5 -> RepeatMasker 4.0.7 재설치 with RepBase20170127

04.

version (20170127) is not valid. at /Users/inmobi/Desktop/Korflab/Packages/RepeatMasker/RepeatMasker_4.0.7/LibraryUtils.pm line 296.



LibraryUtils.pm line 296를 찾아보면

```
    if ( $rmlibVersion && $rmlibVersion <= 20160829 ) {...}
    else {
      die "Legacy file format for $libDir/$mainLibrary yet\n"
}
```

RepeatMaskerLib.embl의 버전이 20160829 이상인 경우에는 스크립트를 종료시킨다.

https://gist.github.com/kbradnam/2632d5ff54b1a9f732e2 참고 다음과 같이 Repbase, RepeatMasker, 그리고 RMBlast의 버전을 일치시킨다.

Repbase : repeatmaskerlibraries-20140131.tar.gz (52.85 MB)
RepeatMasker : RepeatMasker-open-4-0-5.tar.gz
RMBlast : ncbi-blast-2.2.28+-src.tar.gz -> ncbi-blast-2.2.28+-src.tar.gz
rmblast-2.2.28-src.tar.gz -> ncbi-rmblastn-2.2.28-src.zip