Comperative Genomics

이승우

Comperative Genomics

Comparative Genomics 란 무엇인가?

다른 종과 다른 종을 분석하는 것

여기에서는 인간의 유전체와 다른 동물(혹은 종) 간에 유전체를 비교하는것을 의미!

?

인간과 동물의 게놈을 비교 분석함으로써 질병 치료는 물론 게놈 염기서열을 통해

인간과 동물의 진화 등을 밝히려는 학문

feat. 네이버 두산지식 백과

그래서...

1. 여기에서는 Fully Sequenced 된 종들을 가져다가 분석한다

   (그 예로 쥐와 인간을 예로 들수 있다)

2. 앙상블 데이터베이스를 사용하고 (많은 종을 Seq. 한 데이터베이스들이 있음)

3. 맥주효모균과 작은 수의 Org. Animals(예쁜꼬마선충, 노랑초파리)

4. 여기에서는 biomaRt Package를 이용할것이다.

Source 1.

```
library("biomaRt") # R에서 바이오마트 Package를 로드
listmarts() # 사용할 수 있는 데이터베이스를 print한다
```

	biomart
1	ensembl
2	snp
3	functional_genomics
4	vega
5	bacterial_mart_9
6	fungal_mart_9
7	fungal_variations_9
8	metazoa_mart_9
9	metazoa_variations_9
10	plant_mart_9
11	plant_variations_9
12	protist_mart_9
13	protist_variations_9
14	msd
15	htgt
16	REACTOME
17	WS220-testing

...

	version
1	ENSEMBL GENES 62 (SANGER UK)
2	ENSEMBL VARIATION 62 (SANGER UK)
3	ENSEMBL FUNCTIONAL GENOMICS 62 (SANGER UK)
4	VEGA 42 (SANGER UK)
5	ENSEMBL BACTERIA 9 (EBI UK)
6	ENSEMBL FUNGI 9 (EBI UK)
7	ENSEMBL FUNGI VARIATION 9 (EBI UK)
8	ENSEMBL METAZOA 9 (EBI UK)
9	ENSEMBL METAZOA VARIATION 9 (EBI UK)
10	ENSEMBL PLANTS 9 (EBI UK)
11	ENSEMBL PLANTS VARIATION 9 (EBI UK)
12	ENSEMBL PROTISTS 9 (EBI UK)
13	ENSEMBL PROTISTS VARIATION 9 (EBI UK)
14	MSD (EBI UK)
15	WTSI MOUSE GENETICS PROJECT (SANGER UK)
16	REACTOME (CSHL US)
17	WORMBASE 220 (CSHL US)

Source 2.

```
ensemblprotists <- useMart("protist_mart_9") # R에서 biomaRt를 통하여 protist_mart_9 을 로드
listDatasets(ensemblprotists) # listDatasets() 함수를 사용하여, ensemblprotists 를 print
```

```
        dataset                                    description
1      pramorum_eg_gene         Phytophthora ramorum genes (Phyra1_1)
2        pvivax_eg_gene                Plasmodium vivax genes (EPr 2)
3   pfalciparum_eg_gene           Plasmodium falciparum genes (2.1.4)
4  ptricornutum_eg_gene      Phaeodactylum tricornutum genes (Phatr2)
5     pchabaudi_eg_gene          Plasmodium chabaudi genes (May_2010)
6   ddiscoideum_eg_gene Dictyostelium discoideum genes (dictybase.01)
7        lmajor_eg_gene    Leishmania major strain Friedlin genes (1)
...
  version
1      Phyra1_1        #Plasmodium vivax & Plasmodium Falciparium
2         EPr 2        #Cause Malaria.
3         2.1.4
4        Phatr2
5      May_2010
6  dictybase.01
7             1
...
```

```
leishmaniaattributes <- listAttributes(ensemblleishmania)
#leishmaniaatributes 변수에 listAttributes func() 를 이용하여 ensemblleishmania를 태움

attributenames <- leishmaniaattributes[[1]]
# attribuenames 에다가 leishmaniaattributes[[1]] 행을 불러옴

attributedescriptions <- leishmaniaattributes[[2]]
# attributedescriptions에다가 leishmaniattributes[[2]] 행을 불러옴

length(attributenames) # Vector Attributenames 의 길이를 print
[1] 292

attributenames[1:10] #처음 1에서 10까지의 entri를 print

[1] "ensembl_gene_id"               "ensembl_transcript_id"
[3] "ensembl_peptide_id"            "canonical_transcript_stable_id"
[5] "description"                   "chromosome_name"
[7] "start_position"                "end_position"
[9] "strand"                        "band"
```

getBM() 함수를 이용하여 Leishmania Major Data 를 불러오기

```
leishmaniagenes <- getBM(attributes = c("ensembl_gene_id"), mart=ensemblleishmania)
leishmaniagenenames <- leishmaniagenes[[1]] # GET 벡터 of 이름 of all L. major genes
length(leishmaniagenenames) #leishmaniagenenames 길이 print
[1] 9379 # 9379개의 다른 Leishmania major Gene들이 L. major 앙상블 데이터셋에서 찾음.
leishmaniagenenames[1:10] #열개의 Gene들을 print
[1] "LmjF.01.0010" "LmjF.01.0020" "LmjF.01.0030" "LmjF.01.0040" "LmjF.01.0050"
[6] "LmjF.01.0060" "LmjF.01.0070" "LmjF.01.0080" "LmjF.01.0090" "LmjF.01.0100"
```

Note that this includes various types of genes including protein-coding genes (both “known” and “novel” genes, where the “novel” genes are gene predictions that don’t have sequence similarity to any sequences in sequence databases), RNA genes, and pseudogenes.

아직 밝혀지지 않은 Gene과 비슷한 종의 Gene를 비교해서, 각 Gene의 역할을 밝혀내는 것.

단백질로 코딩된 Gene들에게만 관심이 있다면?

```
leishmaniagenes2 <- getBM(attributes = c("ensembl_gen_id", "gene_biotype"), mart=ensemblleishmania))
leishmaniagenenames2 <- leishmaniagens2[[1]] # 모든 백터를 L. major genes 들에서 불러냅니다
leishmaniagenebiotypes2 <- leishmaniagenes2[[2]] # 벡터 of BioTypes of 모든 Gene들에게서
```

table() 함수를 이용하여 뽑아보면...

```
table(leishmaniagenebiotypes2) # table() 함수를 이용하여 leishmaniagenebiotype의 테이블을 생성
leishmaniagenebiotypes2
      ncRNA nontranslating_cds protein_coding pseudogene
         84                 2           8310         90
       rRNA             snoRNA          snRNA       tRNA
         63                741              6         83
```

Combine With 2 Genes Using merge() method

```
leishmaniagenes3 <- merge(leishmaniagenes2, leishmaniagenes) # merge함수를 이용하여 Gene들을 합침.
leishmaniagenenames <- leishmaniagenes3[[1]] # leishmaniagenes3의 1번항을 leishmaniagenenames 에 지정
leishmaniagenebiotypes <- leishmaniagenes3[[2]] # leishmaniagenes3의 2번항을 leishmaniagenebiotypes 에 지정
```