Virus geome MSA

SEQUENCE-BASED ANALYSIS OF GENETIC VARIABILITY IN HANTAVIRUS AND INFLUENZA

Table of Contents

Introduction

Hantavirus & Influenza H3N2

한타바이러스(신증후출혈열 원인균)와 인플루엔자 H3N2 바이러스는 높은 유전적 변동성을 보이며 인류 건강을 위협하고 있다.

바이러스의 유전적 변이가 무작위로 발생하는지, 아니면 특정 영역에 집중되는지 확인하고, 유전적 거리와 서열 유사성 간의 관계를 규명하여 바이러스의 진화적 패턴과 군집 구조를 파악하고자 한다.

이 분석에서는 크게 두 가지 가설에 대해 분석하고자 한다.

1. 바이러스의 변이는 특정 위치(Hotspots)에 유의하게 집중될 것이다
2. 계통수(Phylogenetic Tree) 상의 거리는 실제 서열의 차이를 정확히 반영할 것이다

Project information

Data resource: NCBI

Libraries

- Biopython
- Pandas
- Seaborn
- Scipy.stats

Statistics

- Mann–Whitney U test
- Spearman's Rank Correlation Coefficient

Flowchart

1 → Data acquisition

2 → Preprocessing

3 → MSA analysis

4 → Shannon Entropy

5 → Phylogenetic tree
Clustal analysis

Data acquisition & MSA

1

Data acquisition

Biopython을 이용해 NCBI Entrez 접속

300개의 바이러스 게놈 데이터 입수

2

Preprocessing

Hantavirus Segment S DNA filtering

시퀀스 길이가 1600~2000인 데이터만 따로 필터링

3

MSA analysis

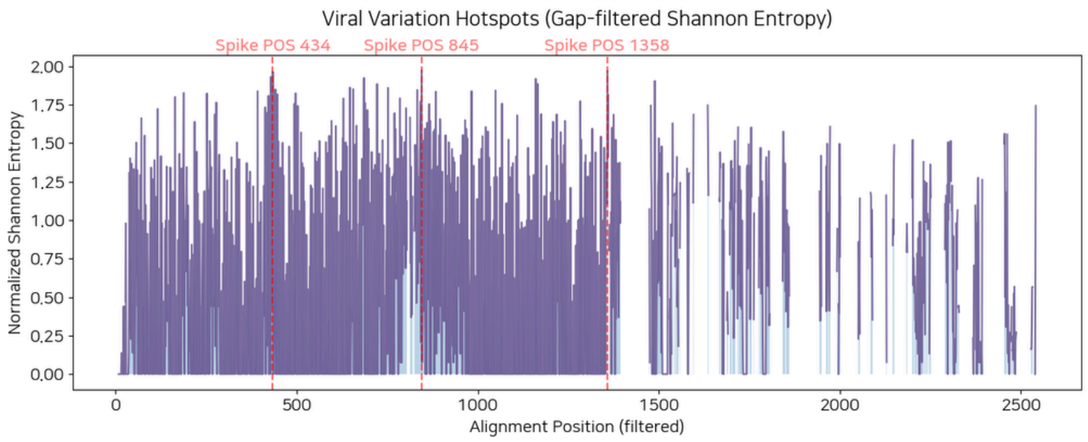
Multiple sequence analysis

Subprocess 및 MUSCLE을 이용한 MSA 진행, 유전적 상동성 확인

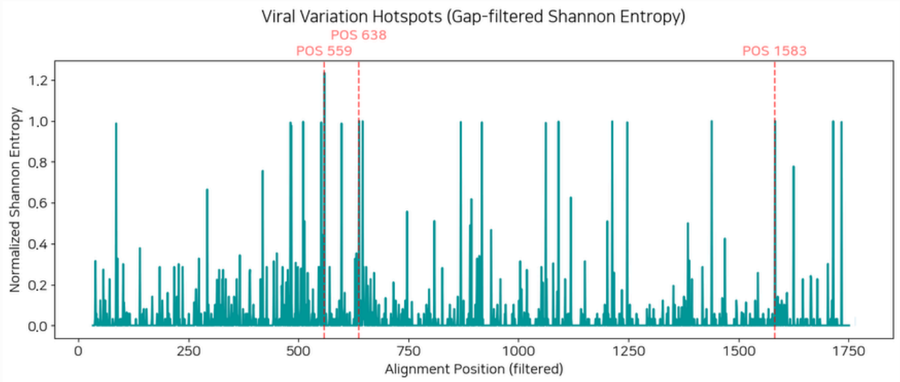
Shannon entropy

변이 민감도를 측정하기 위해 Shannon Entryphy 도출

Window size: 25bp



Hantavirus Shanon entropy (POS: 434, 845, 1358)



Influenza H3N2 Shanon entrophy (POS: 559, 638, 1583)

Statistics

Shannon entropy

Mann-Whitney U test

DNA 염기가 범주형이므로 비모수검정인 Mann-Whitney U test를 진행

귀무가설: 바이러스의 변이는 무작위적으로 발생하며, 특정 위치에 선호적으로 집중되지 않는다.
대립가설: 바이러스의 변이는 무작위가 아니며, 특정 위치(hotspots)에 유의하게 집중된다.

Mann-Whitney U test

	Window size	Threshold*	U-statics	P-value
Hantavirus	25bp	0.750	291600.0	$p < 1e-10$
Influenza	25bp	0.1175	281607.0	$p < 1e-10$

*window-averaged, normalized entropy 상위 10%

바이러스의 변이는 무작위가 아니며, 특정 위치에 유의하게 집중됨을 확인하였다.

Phylogenetic tree: Statistics

Phylogenetic tree

Hantavirus: Mann-Whitney U test
Influenza: Spearman Correlation

귀무가설: 동일 clade 내 서열 유사도와 서로 다른 clade 간 서열 유사도에는 차이가 없다.
대립가설: 동일 clade 내 서열 유사도가 clade 간 서열 유사도보다 유의하게 높다.

Hantavirus

Mann-Whitney U test & effect_size

U-statics	P-value	Effect size	Rank-biserial
1014127.0	p < 1e-10	0.3479	-0.9962

계통수의 동일 clade간 서열 유사도가 clade간 서열 유사도보다 유의하게 높음을 확인

귀무가설: 계통수 구조는 H3N2 해마글루티닌의서열 차이를 반영하지 않는다.
대립가설: 계통수 구조는 H3N2 해마글루티닌의 서열 차이를 반영했다.

Influenza

Spearman Correlation

Spearman ρ	P-value
-0.9680	p < 1e-10

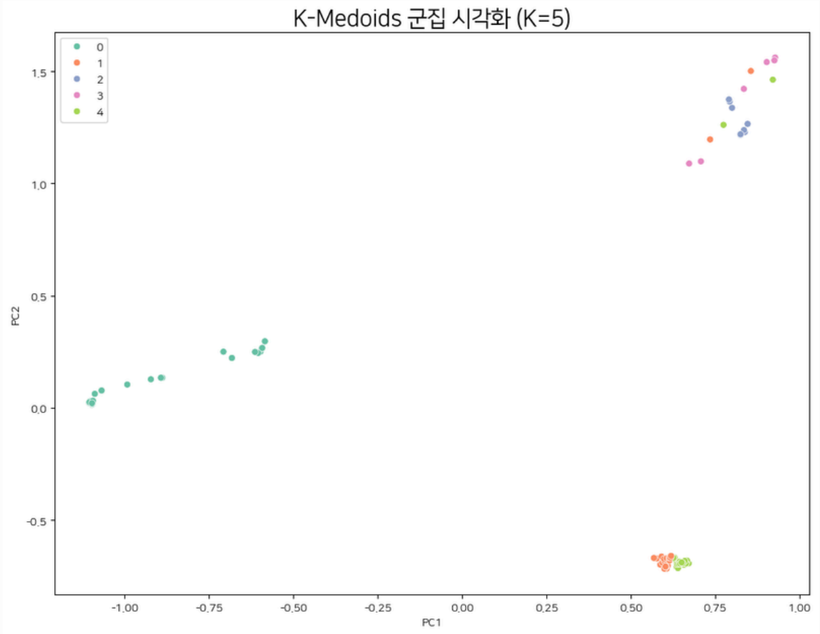
계통수 구조가 H3N2 해마글루티닌의 서열 차이를 반영함을 확인

Phylogenic tree: Clustal analysis

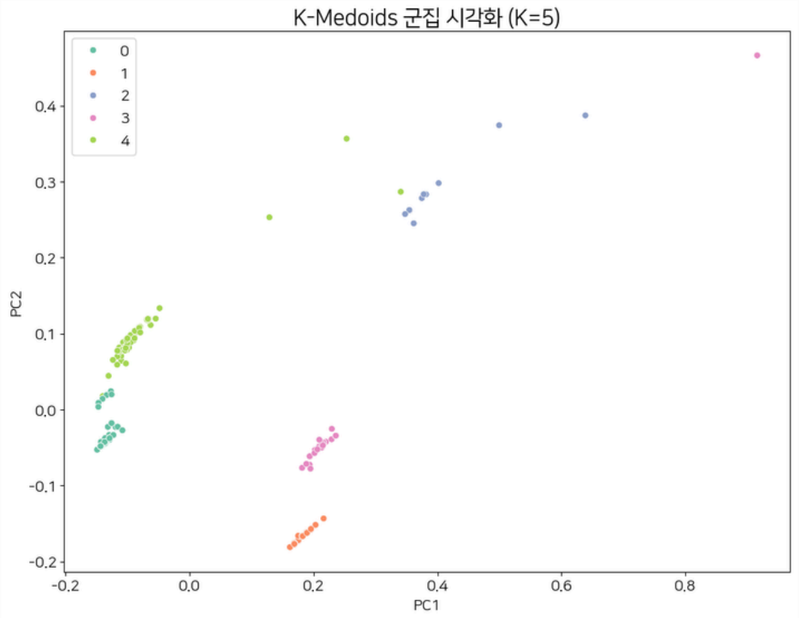
Clustal analysis

k-medoid (k=5), Silhouette Score = 0.63

Phylogenetic tree를 그릴 때 도출된 거리행렬을 기반으로 k-medoid 진행

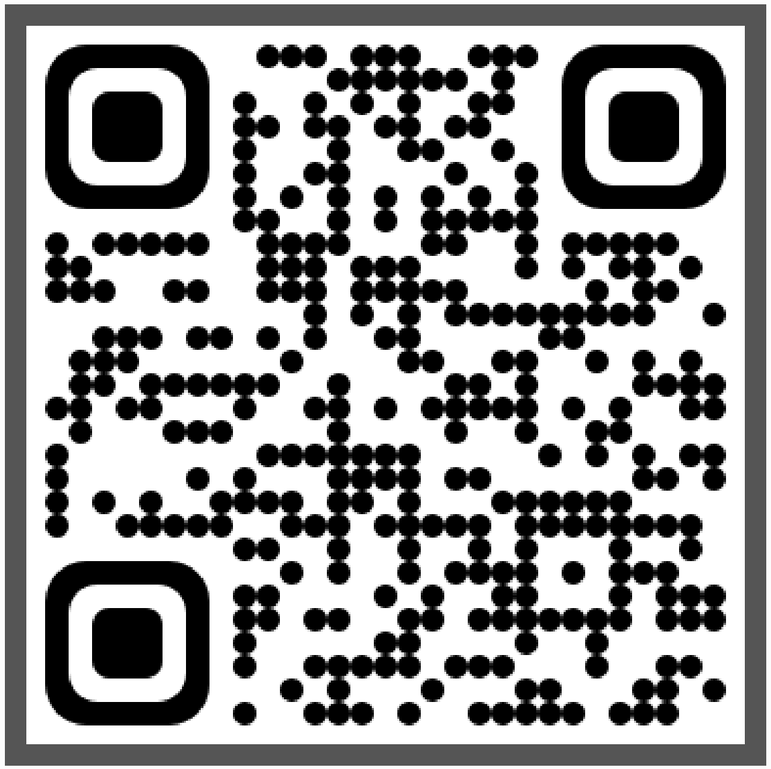


Hantavirus



Influenza H3N2

Contact



Mail

pokemonms@naver.com
blackholekun@gmail.com

Cellular

+82-10-5027-0328

Github

https://github.com/koreanraichu

Blog

https://koreanraichu.tistory.com/