

SESSION 8. PRACTICE

Evolution Resolving a criminal case

Phylogenetic analysis of HIVs

DNA samples from Victim, patients, and controls (from Lafayette regions)

PCR DNAs with primers of Env and RT and sequenced them.

Computational analysis of phylogenetic trees with the sequences using ClustalW

Data is publicly available in NCBI Entrez (AY156734-AY156907)

- 132 env sequences
- 42 RT sequences
- clustalw2 rt.fa

How to download Data from NCBI database

The screenshot displays the NCBI website interface. At the top, there is a navigation bar with the NCBI logo, "Resources" (checked), "How To" (checked), and a "Sign in to NCBI" link. Below this is a search bar with a dropdown menu set to "All Databases" and a "Search" button. The main content area is divided into several sections:

- NCBI Home**: A blue arrow-shaped button.
- Resource List (A-Z)**: A vertical list of categories including All Resources, Chemicals & Bioassays, Data & Software, DNA & RNA, Domains & Structures, Genes & Expression, Genetics & Medicine, Genomes & Maps, Homology, Literature, Proteins, Sequence Analysis, Taxonomy, Training & Tutorials, and Variation.
- Welcome to NCBI**: A central section with the text "The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information." and links for "About the NCBI", "Mission", "Organization", "NCBI News", and "Blog".
- Submit**: A section with the text "Deposit data or manuscripts into NCBI databases" and an icon of a document being uploaded.
- Download**: A section with the text "Transfer NCBI data to your computer" and a download icon.
- Learn**: A section with the text "Find help documents, attend a class or watch a tutorial" and an icon of books.
- Develop**: A section with the text "Use NCBI APIs and code libraries to build applications" and an icon of stacked blocks.
- Analyze**: A section with the text "Identify an NCBI tool for your data analysis task" and a network diagram icon.
- Research**: A section with the text "Explore NCBI research and collaborative projects" and a microscope icon.
- Popular Resources**: A list of links including PubMed, Bookshelf, PubMed Central, PubMed Health, BLAST, Nucleotide, Genome, SNP, Gene, Protein, and PubChem.
- NCBI Announcements**: A section with two announcements: "May 4th NCBI Minute: Linking PubMed and ClinicalTrials.gov" (dated 26 Apr 2016) and "Next Wednesday, May 4th, NCBI will present a short tutorial that will teach you" (dated 25 Apr 2016). Below this is a link to a "New NCBI video on YouTube: 'Sequence Viewer: Display dbVar Supporting Calls'" (dated 25 Apr 2016).

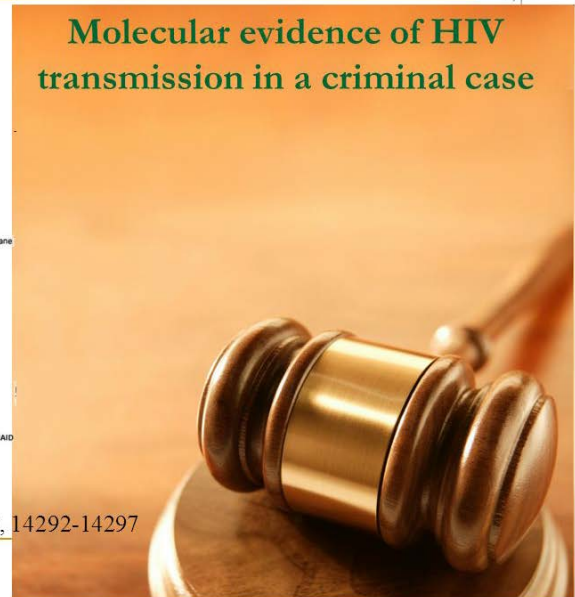
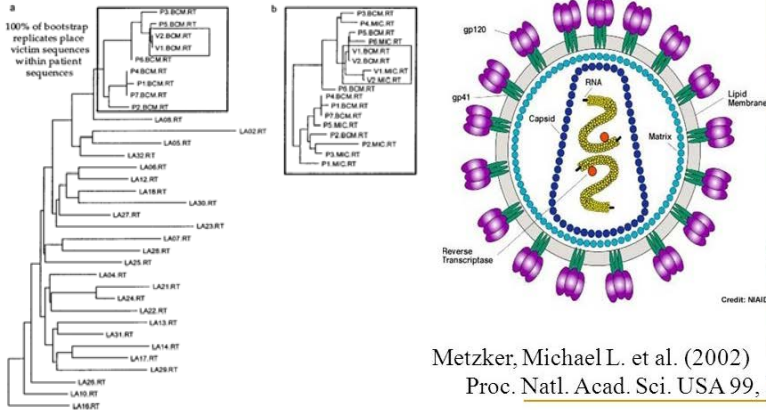
Phylogenetic tree as both scientific and legal evidence in crime scene

Molecular evidence of HIV-1 transmission in a criminal case

Michael L. Metzker^{*,†}, David P. Mindell[†], Xiao-Mei Liu^{*,§}, Roger G. Ptak[¶], Richard A. Gibbs^{*}, and David M. Hillis^{**}

^{*}Department of Molecular and Human Genetics, Baylor College of Medicine, Houston, TX 77030; [†]Department of Ecology and Evolutionary Biology and Museum of Zoology, University of Michigan, Ann Arbor, MI 48109-1079; [§]School of Dentistry, Biologic and Materials Sciences, University of Michigan, Ann Arbor, MI 48109; and ^{**}Section of Integrative Biology and Center for Computational Biology and Bioinformatics, University of Texas, Austin, TX 78712

Edited by Walter M. Fitch, University of California, Irvine, CA, and approved September 4, 2002 (received for review May 2, 2002)



Metzker, Michael L. et al. (2002)
Proc. Natl. Acad. Sci. USA 99, 14292-14297

How to download Data from NCBI database (data from the paper in previous slide)

The screenshot shows the NCBI Nucleotide search interface. The search term 'metzker hiv-1 clone USA' is entered in the search bar and is highlighted with a red box. Below the search bar, there is a navigation bar with 'Summary', '20 per page', and 'Sort by Default order'. The search results are displayed as a list of four items, each with a checkbox, a title, and a description. The first item is 'HIV-1 clone V50 from USA envelope glycoprotein (env) gene, partial cds' with a 'FASTA' link highlighted by a red box. The second item is 'HIV-1 clone V49 from USA envelope glycoprotein (env) gene, partial cds'. The third item is 'HIV-1 clone V48 from USA envelope glycoprotein (env) gene, partial cds'. The fourth item is 'HIV-1 clone V47 from USA envelope glycoprotein (env) gene, partial cds'. On the right side, there are sections for 'Results by taxon', 'Find related data', and 'Search details'. The 'Search details' section shows the search query: 'metzker[All Fields] AND ("Human immunodeficiency virus 1"[Organism] OR hiv-1[All Fields]) AND clone[All Fields] AND USA[All Fields]'. The 'FASTA' link in the first search result is highlighted with a red box.

NCBI Resources How To Sign in to NCBI

Nucleotide Nucleotide metzker hiv-1 clone USA Search

NCBI is phasing out sequence GI numbers in September 2016. Please use accession.version! [Read more...](#)

Species Summary 20 per page Sort by Default order Send to Filters: [Manage Filters](#)

Animals (26)
Viruses (806)
Customize ...

Molecule types genomic: DNA/RNA (832)
Customize ...

Source databases INSDC (GenBank) (806)
RefSeq (26)
Customize ...

Sequence length Custom range...

Release date Custom range...

Revision date Custom range...

[Clear all](#)

[Show additional filters](#)

Items: 1 to 20 of 832

<< First < Prev Page 1 of 42 Next > Last >>

[HIV-1 clone V50 from USA envelope glycoprotein \(env\) gene, partial cds](#)
1. 793 bp linear DNA
Accession: AY156905.1 GI: 24210283
[GenBank](#) **FASTA** [Graphics](#) [PopSet](#)

[HIV-1 clone V49 from USA envelope glycoprotein \(env\) gene, partial cds](#)
2. 793 bp linear DNA
Accession: AY156906.1 GI: 24210281
[GenBank](#) [FASTA](#) [Graphics](#) [PopSet](#)

[HIV-1 clone V48 from USA envelope glycoprotein \(env\) gene, partial cds](#)
3. 793 bp linear DNA
Accession: AY156905.1 GI: 24210279
[GenBank](#) [FASTA](#) [Graphics](#) [PopSet](#)

[HIV-1 clone V47 from USA envelope glycoprotein \(env\) gene, partial cds](#)
4. 775 bp linear DNA
Accession: AY156904.1 GI: 24210277
[GenBank](#) [FASTA](#) [Graphics](#) [PopSet](#)

Results by taxon

Top Organisms [\[Tree\]](#)

Human immunodeficiency virus 1 (806)
Homo sapiens (23)
Rattus norvegicus (3)

Find related data

Database: Select

Find items

Search details

metzker[All Fields] AND ("Human immunodeficiency virus 1"[Organism] OR hiv-1[All Fields]) AND clone[All Fields] AND USA[All Fields]

Search See more...

How to download Data from NCBI database (data from the paper in previous slide)



Basic Shell Commands

```
$ cd [User_Folder]
```

```
$ mkdir Session8
```

```
$ cd Session8
```

Using clustalw2 for phylogenetic analysis

```
$ ln -s /home/biguser/tutor/Session8/rt.fa .  
$ ll
```

```
[biguser@biglab-master session8]$ ll  
total 36  
-r--r--r-- 1 biguser biguser 34710 Apr 27 10:37 rt.fa
```


Using clustalw2 for phylogenetic analysis

```
$ clustalw2 rt.fa  
$ clustalw2 rt.aln -tree  
$ clustalw2 rt.aln -bootstrap=1000
```

```
[biguser@biglab-master session8]$ ll  
total 108  
-rw-rw-r-- 1 biguser biguser 58796 Apr 27 15:10 rt.aln  
-rw-rw-r-- 1 biguser biguser  1983 Apr 27 15:10 rt.dnd  
-r--r--r-- 1 biguser biguser 34710 Apr 27 10:37 rt.fa  
-rw-rw-r-- 1 biguser biguser  1984 Apr 27 15:10 rt.ph  
-rw-rw-r-- 1 biguser biguser  2172 Apr 27 15:10 rt.phb
```

Using clustalw2 for phylogenetic analysis

```
$ less rt.phb
```

```
(  
(  
gi|24209945|gb|AY156737.1|:0.00000,  
gi|24209939|gb|AY156734.1|:0.00146)  
:0.00000[590],  
gi|24209951|gb|AY156740.1|:0.00000)  
:0.00162[976],  
gi|24210015|gb|AY156802.1|:0.00128)  
:0.00077[431],  
(  
gi|24210009|gb|AY156799.1|:0.00845,  
gi|24210011|gb|AY156800.1|:0.00149)  
:0.00115[318])  
:0.00080[295],  
gi|24209941|gb|AY156735.1|:0.00338)  
:0.00142[389],  
gi|24210007|gb|AY156797.1|:0.00327)  
:0.00151[590],  
(  
(  
(  
gi|24209943|gb|AY156736.1|:0.00267,  
gi|24210013|gb|AY156801.1|:0.00313)  
:0.00070[417],  
(  
(  
(  
gi|24209953|gb|AY156741.1|:0.00000,  
gi|24209955|gb|AY156742.1|:0.00000)  
:0.00018[961],  
gi|24209947|gb|AY156738.1|:0.00127)  
:0.00017[697],  
gi|24210017|gb|AY156803.1|:0.00273)  
:0.00071[400],  
(  
gi|24210019|gb|AY156806.1|:0.00134,  
gi|24210021|gb|AY156807.1|:0.00000)  
:0.00318[959])  
:0.00158[549])  
:0.00245[788],
```

Code 9.1 reformat_giline.ipynb

```
In [4]: import sys
import re

output_open= open("rt_reformat.fa","w")
for line in open("rt.fa", "r"):
    line= line.strip()
    match= re.search(">.*clone (\\S+) ", line)
    if match:
        id= ">" + match.group(1)+ "\\n"
        output_open.write(id)
    else:
        output_open.write(line.rstrip()+ "\\n")
output_open.close()
```

*Regular expression "\\S" - Matches any character which is not a Unicode whitespace character

Code 9.1 reformat_giline.ipynb

jupyter rt_reformat.fa 2분 전

File Edit View Language

```
1 >P1_BCM_RT
2 CCCATAAGTCCATTGAAACTGTACCAGTAAAAATAAGGCCAGGAATGGATGGCCCCAAAAGTTAAACAAAT
3 GGCCACTGACAGAAGAAAAATAAAGCATTAGTAGAAAATTTGTACAGAAAATGGAAAAGGAAGAAAAAAT
4 TTCAAAAAATGGGCCTGAAAATCCATACAATACTCCAGTATTTGCCATAAAGAAAAAAGACAGTACTAAA
5 TGGAGAAAAATAGTAGATTTACAGAGAACTTAATAAGAGAAGAACTCAAGACTCTGGGAAGTTCAATTAGGAA
6 TACCACATCTGTCAGGGTTAAAAAGAAAAAATCAGTAACAGTCTGGATGGGTGATGCATATTTTTTC
7 AGTTCOCTTAGATAAAGAGTTCAGGAAGTACTGCTTTACCATAOCTAGTATAAACAATGAGACACCA
8 GGGATTAGATATCAGTACAATGTGCTTCCACAGGGATGGAAAAGGATCACCAGCAATATTCCAAAGTAGCA
9 TGACAAAAATCTTAGAGCCTTTTAGAAAAACAAAATCCAGACATAGTATCTATCAATACATGGATGATCT
10 GTATGTAGGATCTGACTTAGAAAAATAGGGCAGCATAGAAATAAAAAATAGAGGAACCTAAGACACATCTGTTG
11 AAGTGGGGACTTACCACACAGACAAAAAACATAGAAGGAAACCCCATTCCTTTGGAT
12
13 >P2_BCM_RT
14 CCCATAAGTCCATTGAAACTGTACCAGTAAAAATAAGGCCAGGAATGGATGGCCCCAAAAGTTAAGCAAT
15 GGCCACTGACAGAAGAAAAATAAAGCATTAGTAGAAAATTTGTACAGAAAATGGAAAAGGAAGAAAAAAT
16 TTCAAAAAATGGGCCTGAAAATCCATACAATACTCCAGTATTTGCCATAAAGAAAAAAGACAGTACTAAA
17 TGGAGAAAAATAGTAGATTTACAGAGAACTTAATAAGAGAAGAACTCAAGACTCTGGGAAGTTCAATTAGGAA
18 TACCACATCTGTCAGGGTTAAAAAGAAAAAATCAGTAACAGTCTGGATGGGTGATGCATATTTTTTC
19 AGTTCOCTTAGATAAAGAGTTCAGGAAGTACTGCTTTACCATAOCTAGTATAAACAATGAGACACCA
20 GGGATTAGATATCAGTACAATGTGCTTCCACAGGGATGGAAAAGGATCACCAGCAATATTCCAAAGTAGCA
21 TGACAAAAATCTTAGAGCCTTTTAGAAAAACAAAATCCAGACATAGTATCTATCAATACATGGATGATCT
22 GTATGTAGGATCTGACTTAGAAAAATAGGGCAGCATAGAAATAAAAAATAGAAGAACTAAGACACATCTGTTG
23 AAGTGGGGACTTACCACACAGACAAAAAACATAGAAGGAAACCCCATTCCTTTGGAT
24
25 >P3_BCM_RT
26 CCCATAAGTCCATTGAAACTGTACCAGTAAAAATAAGGCCAGGAATGGATGGCCCCAAAAGTTAAACAAAT
27 GGCCACTGACAGAAGAAAAATAAAGCATTAGTAGAAAATTTGTACAGAAAATGGAAAAGGAAGAAAAAAT
28 TTCAAAAAATGGGCCTGAAAATCCATACAATACTCCAGTATTTGCCATAAAGAAAAAAGACAGTACTAGA
29 TGGAGAAAAATAGTAGATTTACAGAGAACTTAATAAGAGAAGAACTCAAGACTCTGGGAAGTTCAATTAGGAA
30 TACCACATCTGTCAGGGTTAAAAAGAAAAAATCAGTAACAGTCTGGATGGGTGATGCATATTTTTTC
31 AGTTCOCTTAGATAAAGAGTTCAGGAAGTACTGCTTTACCATAOCTAGTATAAACAATGAGACACCA
32 GGGATTAGATATCAATACAATGTGCTTCCACAGGGATGGAAAAGGATCACCAGCAATATTCCAAAGTAGCA
33 TGACAAAAATCTTAGAGCCTTTTAGAAAAACAAAATCCAGACATAGTATCTATCAATACATGGATGATCT
34 GTATGTAGGATCTGACTTAGAAAAATAGGGCAGCATAGAAATAAAAAATAGAAGAACTAAGACACATCTGTTG
35 AAGTGGGGATTTATCACCACAGACAAAAAACACCAAGGAAACCCCATTCCTTTGGAT
36
```

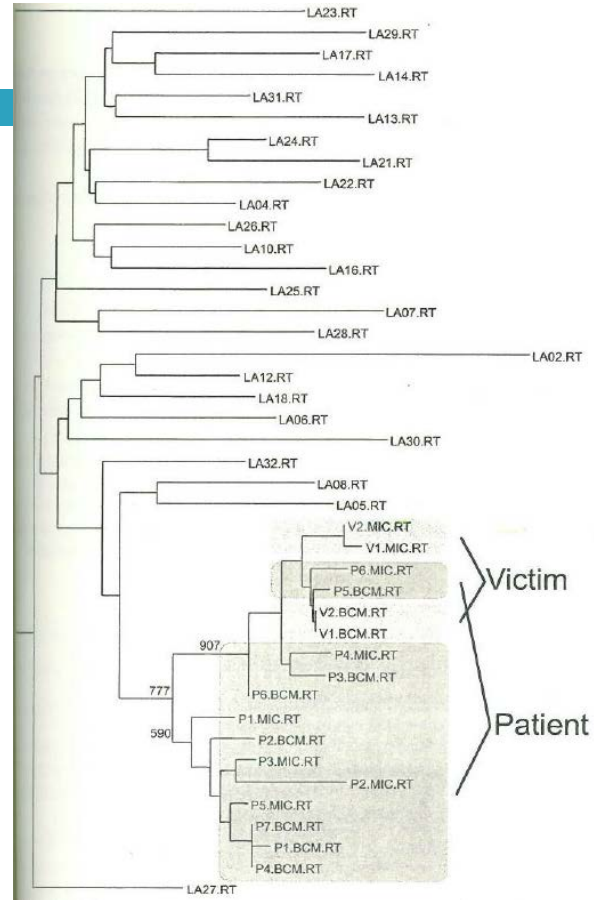
Using clustalw2 for phylogenetic analysis

```
$ clustalw2 rt_reformat.fa  
$ clustalw2 rt_reformat.aln -tree  
$ clustalw2 rt_reformat.aln -bootstrap=1000
```

```
[biguser@biglab-master session8]$ ll  
total 236  
-r--r--r-- 1 biguser biguser 117608 Apr 27 14:00 env.fa  
-r--r--r-- 1 biguser biguser  34710 Apr 27 10:37 rt.fa  
-rw-rw-r-- 1 biguser biguser  44904 Apr 27 15:04 rt_reformat.aln  
-rw-rw-r-- 1 biguser biguser   1219 Apr 27 15:04 rt_reformat.dnd  
-rw-rw-r-- 1 biguser biguser  30762 Apr 27 15:04 rt_reformat.fa  
-rw-rw-r-- 1 biguser biguser   1408 Apr 27 15:05 rt_reformat.phb
```

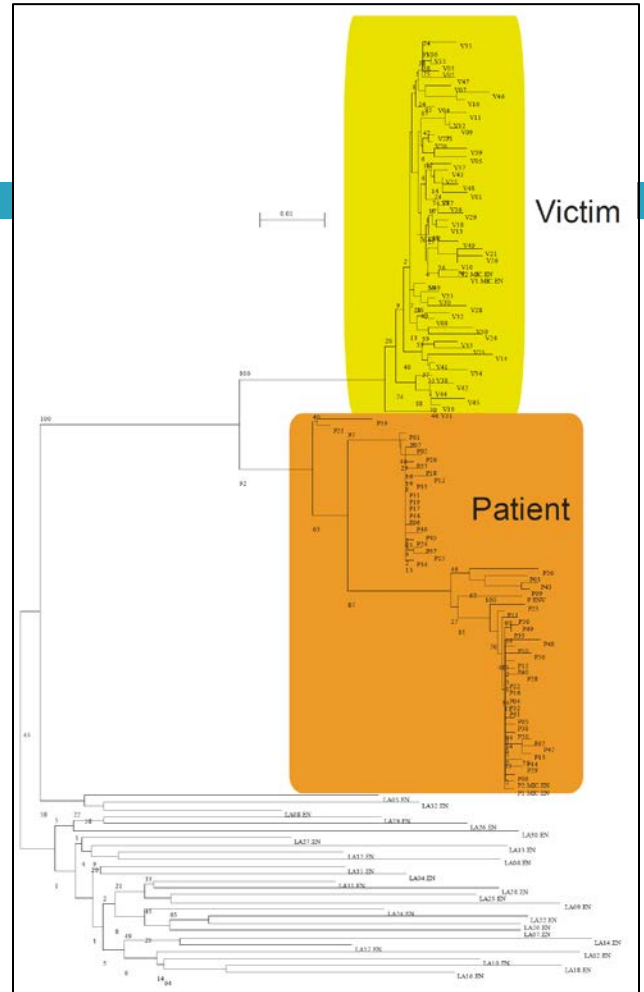
Using clustalw2 for phylogenetic analysis (“rt.fa”)

```
(
(
P4.BCM.RT:0.00000,
P1.BCM.RT:0.00146)
:0.00000[590],
P7.BCM.RT:0.00000)
:0.00162[976],
P5.MIC.RT:0.00128)
:0.00077[431],
(
P2.MIC.RT:0.00845,
P3.MIC.RT:0.00149)
:0.00115[318])
:0.00080[295],
P2.BCM.RT:0.00338)
:0.00142[389],
P1.MIC.RT:0.00327)
:0.00151[590],
(
(
(
P3.BCM.RT:0.00267,
P4.MIC.RT:0.00313)
:0.00070[417],
(
(
(
V1.BCM.RT:0.00000,
V2.BCM.RT:0.00000)
:0.00018[961],
P5.BCM.RT:0.00127)
:0.00017[697],
P6.MIC.RT:0.00273)
:0.00071[400],
```



Using clustalw2 for phylogenetic analysis (“env.fa”)

```
(
(
(
LA16.EN:0.03515,
LA18.EN:0.05200)
:0.00998[68],
LA02.EN:0.06466)
:0.00325[9],
(
(
LA10.EN:0.05145,
LA26.EN:0.04984)
:0.00366[33])
:0.00593[2],
(
(
LA17.EN:0.02509,
LA14.EN:0.06339)
:0.01176[55])
:0.00212[1],
(
(
(
LA07.EN:0.05054,
LA24.EN:0.03889)
:0.00531[25],
(
LA20.EN:0.04773,
LA22.EN:0.04938)
:0.00634[49])
:0.00281[4])
:0.00036,
(
(
LA06.EN:0.06420,
LA05.EN:0.04289)
:0.00278[15],
(
(
LA09.EN:0.06125,
LA25.EN:0.04618)
:0.00343[25],
(
LA28.EN:0.05127,
```



Exercise

- Make a Python script to derive one of the distance matrices shown in Fig. 9.1. The starting point for the script is the multiple alignment in the same figure.

```
>A
GGACCACTACGAGCGCCTACGACGTA
>B
GGACCCCTACGAGCCCCTACGACGTA
>C
GGACCGCTGCGAGCTTCTACGACGTA
>D
GGACCTCTCCGGGCAGCTAGGACGTA
```



	A	B	C	D
A	0	2	4	6
B	2	0	4	6
C	4	4	0	6
D	6	6	6	0

Answer for Exercise

i : row
j : column

for j in xrange

	A	B	C	D
A	0	2	4	6
B	2	0	4	6
C	4	4	0	6
D	6	6	6	0

for i in xrange

```
seqs= ["GGACCACTACGAGCGCCTACGACGTA",  
       "GGACCCCTACGAGCCCCTACGACGTA",  
       "GGACCGCTGCGAGCTTCTACGACGTA",  
       "GGACCTCTCCGGGCAGCTAGGACGTA"]  
for i in xrange(len(seqs)):  
    rowseq= seqs[i]  
    for j in xrange(len(seqs)):  
        colseq= seqs[j]  
        diff=0  
        for k in range(len(colseq)):  
            if rowseq[k] != colseq[k]:  
                diff +=1  
        print diff,  
    print "\n"
```

0 2 4 6

2 0 4 6

4 4 0 6

6 6 6 0

이번주 과제는 어렵습니다. 중간고사 시험 준비 잘 하시기 바랍니다.