

Working with biological sequence data

Installing Biopython:

We will need a command prompt/terminal

- On OS X:
 - Open terminal window
or
 - Open terminal in Jupyter
- On Windows:
 - Go to "Start"
type "cmd"

Home

localhost:8888/tree

ABP

jupyter

FilesRunningClusters

Select items to perform actions on them.

Home

anaconda

Applications

Box Sync

Desktop

Documents

Downloads

Dropbox

envs

github

Movies

Music

Upload

New

Text File

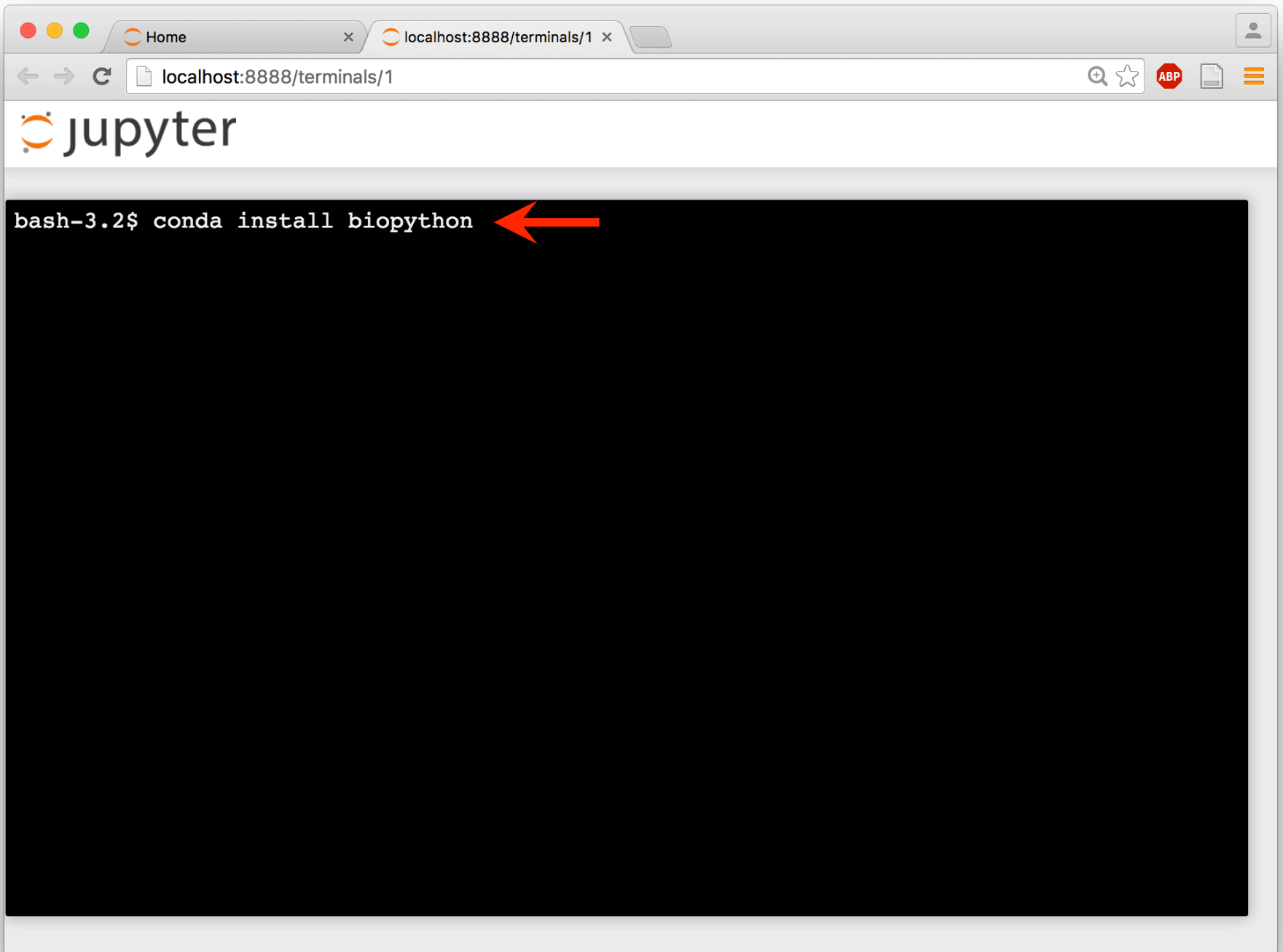
Folder

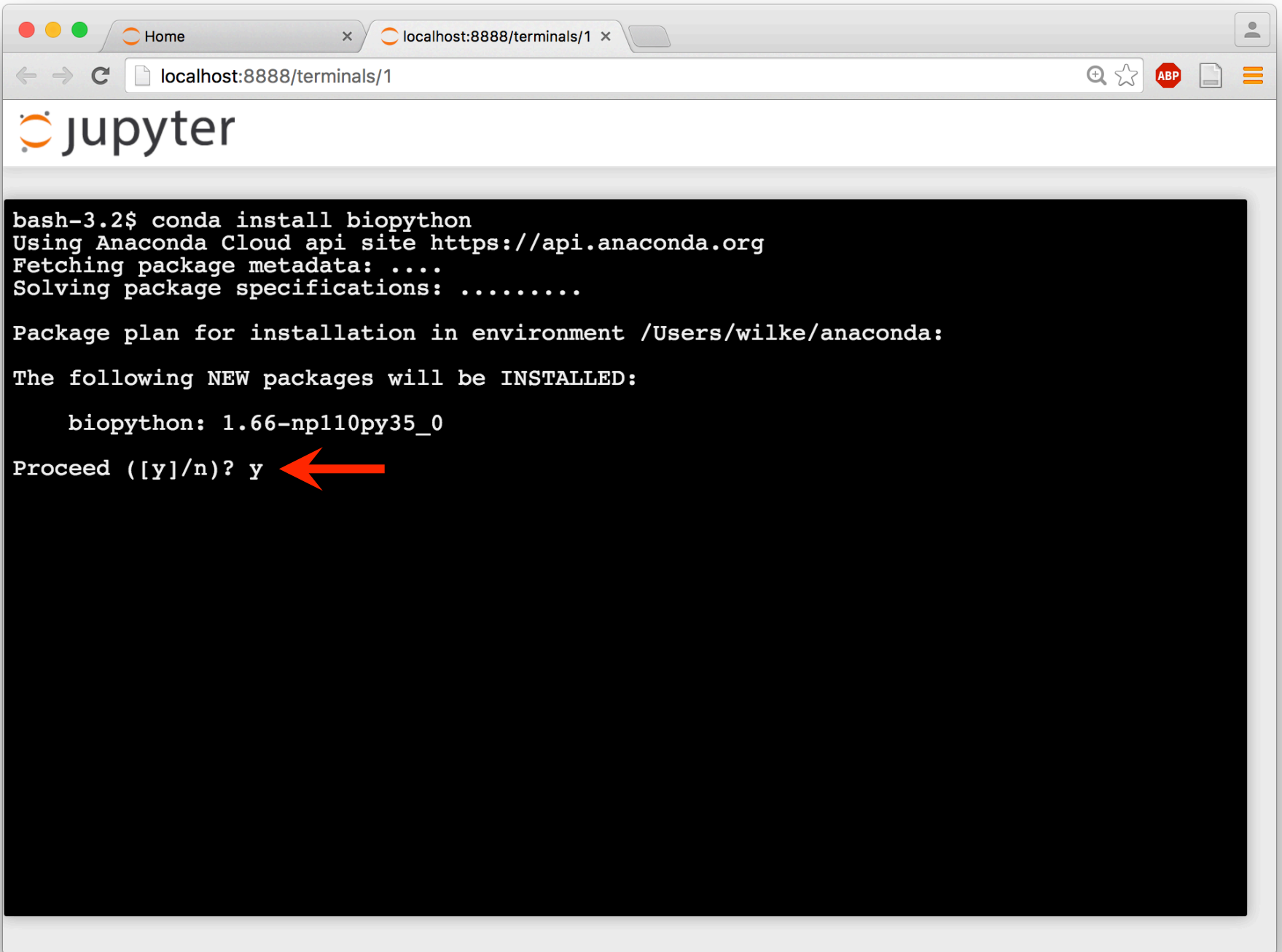
Terminal

Notebooks

Python 3

localhost:8888/tree#





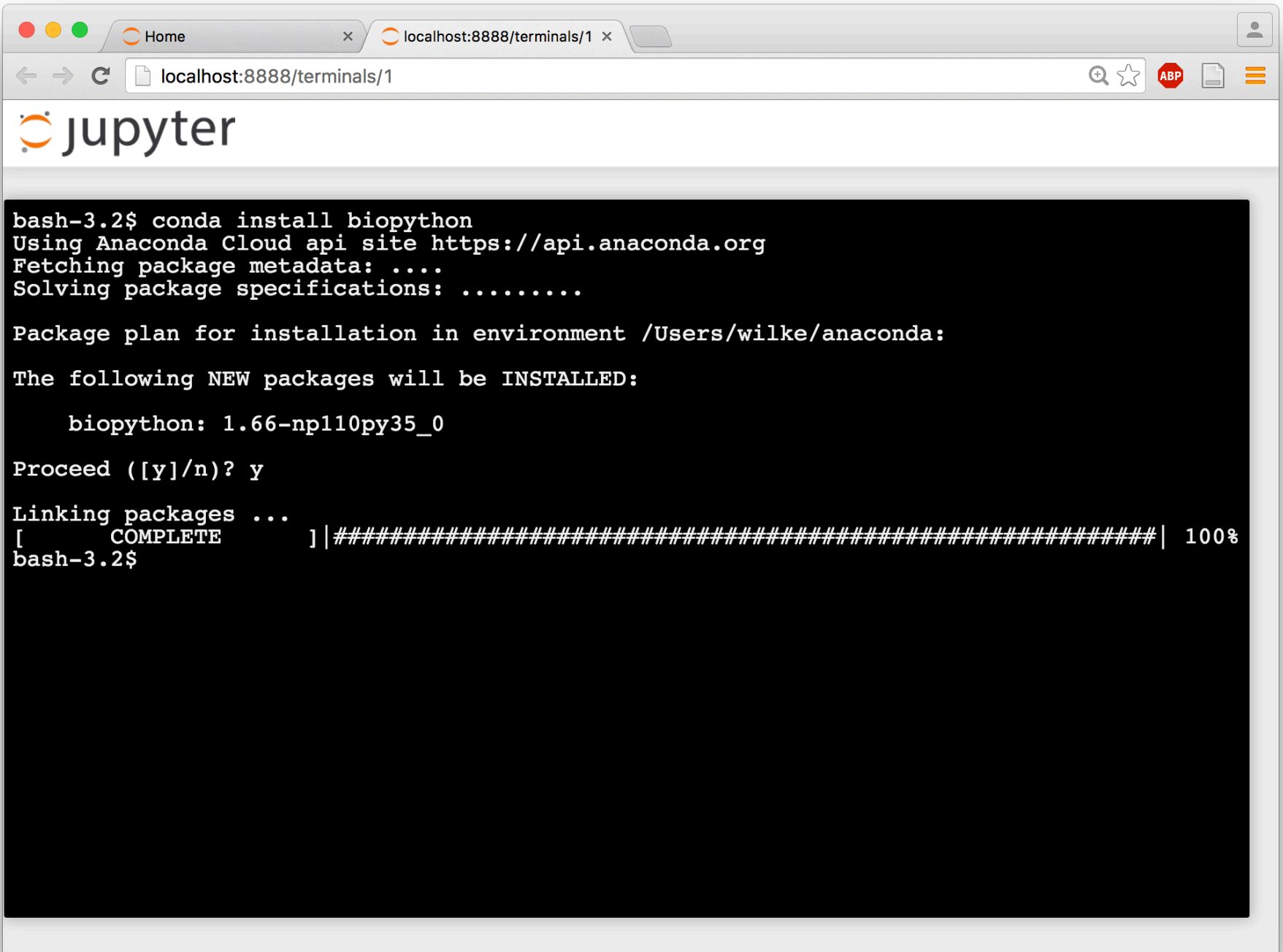
```
bash-3.2$ conda install biopython
Using Anaconda Cloud api site https://api.anaconda.org
Fetching package metadata: ....
Solving package specifications: .....

Package plan for installation in environment /Users/wilke/anaconda:

The following NEW packages will be INSTALLED:

    biopython: 1.66-np110py35_0

Proceed ([y]/n)? y
```



```
bash-3.2$ conda install biopython
Using Anaconda Cloud api site https://api.anaconda.org
Fetching package metadata: ....
Solving package specifications: .....

Package plan for installation in environment /Users/wilke/anaconda:

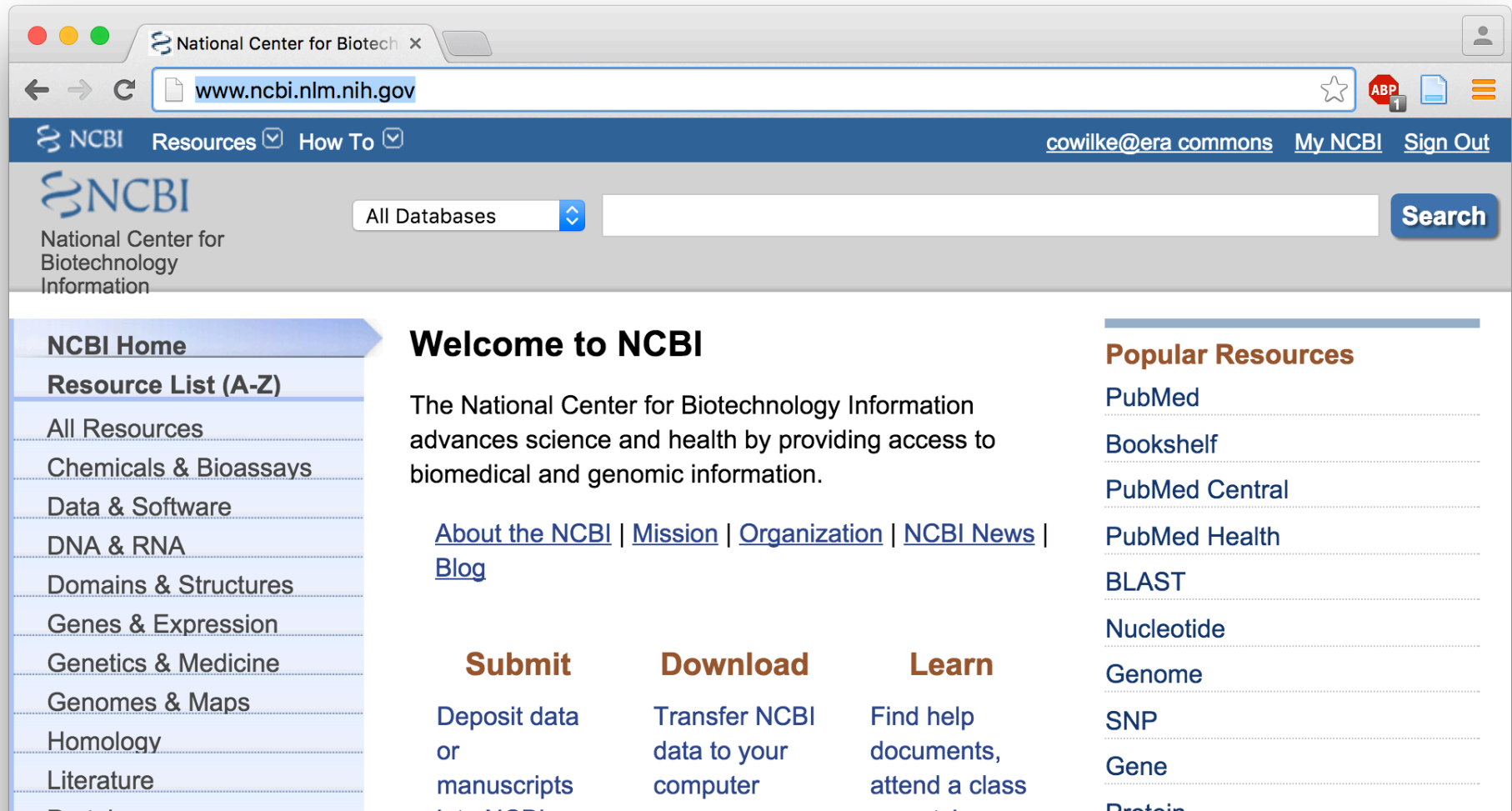
The following NEW packages will be INSTALLED:

    biopython: 1.66-np110py35_0

Proceed ([y]/n)? y
Linking packages ...
[      COMPLETE      ]|#####| 100%
bash-3.2$
```

Getting biological data: The NCBI databases

<http://www.ncbi.nlm.nih.gov/>



The screenshot shows the NCBI website homepage. At the top, there's a browser window with the address bar showing 'www.ncbi.nlm.nih.gov'. Below the browser window, the NCBI logo is on the left, and a navigation bar contains 'Resources' and 'How To' dropdown menus. On the right of the navigation bar, there's a user profile icon, a star, a red 'ABP' icon, a document icon, and a hamburger menu icon. Below the navigation bar, there's a search bar with a dropdown menu set to 'All Databases' and a 'Search' button. On the left side, there's a 'NCBI Home' section with a 'Resource List (A-Z)' and a list of resources: All Resources, Chemicals & Bioassays, Data & Software, DNA & RNA, Domains & Structures, Genes & Expression, Genetics & Medicine, Genomes & Maps, Homology, and Literature. In the center, there's a 'Welcome to NCBI' section with a paragraph about the center's mission and a list of links: About the NCBI, Mission, Organization, NCBI News, and Blog. Below this, there are three columns: 'Submit' (Deposit data or manuscripts), 'Download' (Transfer NCBI data to your computer), and 'Learn' (Find help documents, attend a class). On the right side, there's a 'Popular Resources' section with a list of resources: PubMed, Bookshelf, PubMed Central, PubMed Health, BLAST, Nucleotide, Genome, SNP, Gene, and Protein.

National Center for Biotechnol x

← → ↻ www.ncbi.nlm.nih.gov ☆ ABP 1

NCBI Resources ▾ How To ▾ [cowilke@era commons](#) [My NCBI](#) [Sign Out](#)

NCBI
National Center for
Biotechnology
Information

All Databases ▾ **Search**

NCBI Home
Resource List (A-Z)
All Resources
Chemicals & Bioassays
Data & Software
DNA & RNA
Domains & Structures
Genes & Expression
Genetics & Medicine
Genomes & Maps
Homology
Literature

Welcome to NCBI
The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.
[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News](#) | [Blog](#)

Submit
Deposit data
or
manuscripts
into NCBI

Download
Transfer NCBI
data to your
computer

Learn
Find help
documents,
attend a class

Popular Resources
[PubMed](#)
[Bookshelf](#)
[PubMed Central](#)
[PubMed Health](#)
[BLAST](#)
[Nucleotide](#)
[Genome](#)
[SNP](#)
[Gene](#)
[Protein](#)

Try search for "KT220438"

The screenshot shows a web browser window with the address bar displaying `www.ncbi.nlm.nih.gov/gquery/?term=KT220438`. The page title is "KT220438 - GQuery: Glob: x". The NCBI logo and navigation links are visible. The search results section shows "Results found in 2 databases for 'KT220438'". A highlighted box contains the following information:

[Influenza A virus \(A/NewJersey/NHRC_93219/2015\(H3N2\)\) segment 4 hemagglutinin \(HA\) gene, complete cds](#)
1,701 bp cRNA.
Lab_host: MDCK. Country: USA: New Jersey. Segment: 4. Isolation_source: nasopharyngeal swab. Collection_date: 17-Jan-2015.
Accession: **KT220438.1** GI: 887493048
[GenBank](#) [FASTA](#) [Graphics](#)

Below the highlighted box, there are two sections: "Literature" and "Genes".

Literature		Genes	
Books	0	EST	0
	books and reports		expressed sequence tag sequences
MeSH	0		collected information about
	ontology used for PubMed indexing		

Direct link to search results

<http://www.ncbi.nlm.nih.gov/gquery/?term=KT220438>

A genbank record is just a simple text file

LOCUS KT220438 1701 bp cRNA linear VRL 20-JUL-2015
 DEFINITION Influenza A virus (A/NewJersey/NHRC_93219/2015(H3N2)) segment 4
 hemagglutinin (HA) gene, complete cds.
 ACCESSION KT220438
 VERSION KT220438.1 GI:887493048
 KEYWORDS .
 SOURCE Influenza A virus (A/New Jersey/NHRC_93219/2015(H3N2))
 ORGANISM Influenza A virus (A/New Jersey/NHRC_93219/2015(H3N2))
 Viruses; ssRNA viruses; ssRNA negative-strand viruses;
 Orthomyxoviridae; Influenzavirus A.
 REFERENCE 1 (bases 1 to 1701)
 AUTHORS Sitz,C.R., Thammavong,H.L., Balansay-Ames,M.S., Hawksworth,A.W.,
 Myers,C.A. and Brice,G.T.
 TITLE GEISS Influenza Surveillance Response Program
 JOURNAL Unpublished
 REFERENCE 2 (bases 1 to 1701)
 AUTHORS Sitz,C.R., Thammavong,H.L., Balansay-Ames,M.S., Hawksworth,A.W.,
 Myers,C.A. and Brice,G.T.
 TITLE Direct Submission
 JOURNAL Submitted (29-JUN-2015) Operational Infectious Diseases, Naval
 Health Research Center, 140 Sylvester Rd., San Diego, CA 92106, USA
 COMMENT ##Assembly-Data-START##
 Sequencing Technology :: Sanger dideoxy sequencing
 ##Assembly-Data-END##
 FEATURES Location/Qualifiers
 source 1..1701
 /organism="Influenza A virus (A/New
 Jersey/NHRC_93219/2015(H3N2))"
 /mol_type="viral cRNA"
 /strain="A/NewJersey/NHRC_93219/2015"
 /serotype="H3N2"

FEATURES

source

Location/Qualifiers

1..1701

/organism="Influenza A virus (A/New
Jersey/NHRC_93219/2015(H3N2))"

/mol_type="viral cRNA"

/strain="A/NewJersey/NHRC_93219/2015"

/serotype="H3N2"

/isolation_source="nasopharyngeal swab"

/host="Homo sapiens"

/db_xref="taxon:1682360"

/segment="4"

/lab_host="MDCK"

/country="USA: New Jersey"

/collection_date="17-Jan-2015"

gene

1..1701

/gene="HA"

CDS

1..1701

/gene="HA"

/function="receptor binding and fusion protein"

/codon_start=1

/product="hemagglutinin"

/protein_id="AKQ43545.1"

/db_xref="GI:887493049"

/translation="MKTIIALS YILCLVFAQKIPGNDNSTATLCLGHHAVPNGTIVKT
ITNDRIEVTNATELVQNSSIGEICDSPHQILDGENCTLIDALLGDPQCDGFQNKKDWL
FVERSKAYSNCYPYDVPDYASLRSLVASSGTLEFNNE SFNWTGVTQNGTSSACIRRSS
SSFFSRLNWLTHLNYTYPALNVTMPNNEQFDKLYIWGVHHPGTDKDQIFLYAQSSGRI
TVSTKRSQQAVIPNIGSRPRIRDIPSRSISYWTIVKPGDILLINSTGNLIAPRGYFKI
RSGKSSIMRSDAPIGKCKSECITPNGSIPNDKPFQNVNRITYGACPRYVKHSTLKLAT
GMRNVPEKQTRGIFGAIAGFIENGWEGMVDGWYGFRHQNSEGRGQAADLKSTQAAIDQ
INGKLNRLIGKTNEKFHOIEKEEFSEVEGRIODLEKYVEDTKIDIWSYNAELI.VALENO

ORIGIN

```

1  atgaagacta tcattgcttt gagctacatt ctatgtctgg ttttcgctca aaaaattcct
61  ggaaatgaca atagcacggc aacgctgtgc cttgggcacc atgcagtacc aaacggaacg
121 atagtgaaaa caatcacaaa tgaccgaatt gaagttacta atgctactga gctgggttcag
181 aattcctcaa taggtgaaat atgcgacagt cctcatcaga tccttgatgg agaaaactgc
241 acactaatag atgctctatt gggagaccct cagtgtgatg gctttcaaaa taagaaatgg
301 gacctttttg ttgaacgaag caaagcctac agcaactgct acccttatga tgtgccggat
361 tatgcctccc ttaggtcact agttgcctca tccggcacac tggagtttaa caatgaaagc
421 ttcaattgga ctggagtcac tcaaaacgga acaagttctg cttgcataag gagatctagt
481 agtagtttct ttagtagatt aaattggttg acccacttaa actacacata cccagcattg
541 aacgtgacta tgccaaacaa tgaacaattt gacaaattgt acatttgggg ggttcaccac
601 ccgggtacgg acaaggacca aatcttcctg tatgctcaat catcaggaag aatcacagta
661 tctacaaaaa gaagccaaca agctgtaatc ccaaatatcg gatctagacc cagaataagg
721 gatatcccta gcagaataag catctattgg acaatagtaa aaccgggaga catacttttg
781 attaacagca cagggaatct aattgctcct aggggttact tcaaaatacg aagtgggaaa
841 agctcaataa tgagatcaga tgcacccatt ggcaaatgca agtctgaatg catcactcca
901 aatggaagca ttcccaatga caaaccattc caaaatgtaa acaggatcac atacggggcc
961 tgtcccagat atgttaagca tagcactcta aaattggcaa caggaatgcg aaatgtacca
1021 gagaaacaaa ctagaggcat atttggcgca atagcggggt tcatagaaaa tggttgggag
1081 ggaatggtgg atggttggtg cggtttcagg catcaaaatt ctgagggaag aggacaagca
1141 gcagatctca aaagcactca agcagcaatc gatcaaatca atgggaagct gaatcgattg
1201 atcgggaaaa ccaacgagaa attccatcag attgaaaaag aattctcaga agtagaagga
1261 agaattcagg accttgagaa atatgttgag gacactaaaa tagatctctg gtcatacaac
1321 gcggagcttc ttgttgccct ggagaaccaa catacarttg atctaactga ctcagaaatg
1381 aacaaactgt ttgaaaaaac aaagaagcaa ctgagggaaa atgctgagga tatgggaaat
1441 ggttgtttca aaatatacca caaatgtgac aatgcctgca taggatcaat aagaaatgga
1501 acttatgacc acaatgtgta cagggatgaa gcattaaaca accggttcca gatcaaggga
1561 gttgagctga agtcagggtg caaagattgg atcctatgga tttcctytgc catatcatgt
1621 tttttgcttt gtgttgcttt gttggggttc atcatgtggg cctgccaaaa gggcaacatt
1681 aggtgcaaca tttgcatttg a

```