BASH

- We will go over some basic bash commands
- Download data from ENA

Please open Terminal

Where am I?

$ pwd

‘**p**rint **w**orking **d**irectory’

/Users/childebayeva/agar

What is here?

$ ls

‘list’

How do I know what each command does?

$ man ls

‘manual’

$ man ls -a

Make directories

$ mkdir
‘make directory’

$ mkdir My_directory
$ ls
$ ls -l

$ mv My_directory agar2022
‘move’
$ ls

Play with directories

$ rmdir

‘remove directory’

$ rmdir agar2022

$ ls

$mkdir agar2022_Module1

$ ls

Navigate between directories

$ cd

‘change directory’

$ pwd

$ cd agar2022_Module1

$ cd .. # go back to a parent directory

$ cd ~/agar2022_Module1

$ cd - # go to previous directory

Paths

- **Absolute = from the ‘root’ directory**

/Planet_Earth/Europe/Central_Europe/Germany/Leipzig/Deutscher_Platz_6

Paths

- **Absolute = from the ‘root’ directory**

/Planet_Earth/Europe/Central_Europe/Germany/Leipzig/Deutscher_Platz_6

- **Relative = from current directory**

../Deutscher_Platz_6

Comments

```
$ # This is a Bash comment
$ echo "This is Code" # This is an inline Bash comment

VAR=10
# if [[ $VAR -gt 5 ]]; then
# echo "Variable is greater than 5."
# fi

if [[ $VAR -gt 5 ]]; then
echo "Variable is greater than 5."
fi
```

RESEARCH ARTICLE

The formation of human populations in South and Central Asia

VAGHEESH M. NARASIMHAN, NICK PATTERSON, PRIYA MOORJANI, NADIN ROHLAND, REBECCA BERNARDOS, SWAPAN MALLICK, IOSIF LAZARIDIS, NATHAN NAKATSUKA, IÑIGO OLALDE, [...] DAVID REICH +108 authors Authors Info & Affiliations

SCIENCE • 6 Sep 2019 • Vol 365, Issue 6457 • DOI: 10.1126/science.aat7487

Ancient human movements through Asia

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6822619/

Science

Current Issue | First release papers | Archive | About | Submit manuscript

HOME > SCIENCE > VOL. 365, NO. 6457 > THE FORMATION OF HUMAN POPULATIONS IN SOUTH AND CENTRAL ASIA

RESEARCH ARTICLE

The formation of human populations in South and Central Asia

VAGHEESH M. NARASIMHAN, NICK PATTERSON, PRIYA MOORJANI, NADIN ROHLAND, REBECCA BERNARDOS, SWAPAN MALLICK, IOSIF LAZARIDIS, NATHAN NAKATSUKA, IÑIGO OLALDE, [...] DAVID REICH +108 authors Authors Info & Affiliations

SCIENCE • 6 Sep 2019 • Vol 365, Issue 6457 • DOI: 10.1126/science.aat7487

2,610 134 CHECK ACCESS

Ancient human movements through Asia

ENA European Nucleotide Archive

Enter text search terms | Search

Examples: histone, BN000065

Enter accession | View

Examples: Taxon:9606, BN000065, PRJEB402

Home | Submit | Search | Rulespace | About | Support

We recommend that you subscribe to the ENA-announce mailing list for updates on services.

For SARS-CoV-2 data submissions, users should contact us in advance of submission at virus-dataflow@ebi.ac.uk for specific advice on options and to access the highest levels of support. We have also launched a Drag-and-Drop Data Submission Service (currently in Beta) suitable for certain SARS-Cov-2 submissions. We are inviting submitters to try this out. Please contact us at the email above for details.

European Nucleotide Archive

The European Nucleotide Archive (ENA) provides a comprehensive record of the world's nucleotide sequencing information, covering raw sequencing data, sequence assembly information and functional annotation. More about ENA.

Access to ENA data is provided through the browser, through search tools, through large scale file download and through the API.

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6822619/

https://www.ebi.ac.uk/ena/browser/view/PRJEB32466?show=reads

ENA European Nucleotide Archive

Enter text search terms | Search

Examples: histone, BN000065

PRJEB32466 | View

Examples: Taxon:9606, BN000065, PRJEB402

Home | Submit | Search | Rulespace | About | Support

Project: PRJEB32466

By sequencing 523 ancient humans, we show that the primary source of ancestry in South Asians is an ancient population we detect at sites in cultural contact with the Indus Valley Civilization (IVC) that we show formed a genetic gradient between early hunter-gatherers of Iran as well as hunter-gatherers of South Asia (with a negligible contribution from Central Asia). Following the IVC's decline, people from this population mixed with groups primarily descended from southern Asian hunter-gatherers to form one of the two main sources of South Asian variation, the "Ancestral South Indians" (ASI) whose direct descendants live today in southern India. Around 4000-3500 years ago, people from this same population mixed with descendants of Steppe pastoralists who spread via Central Asia to form the "Ancestral North Indians" (ANI). The Steppe ancestry in the ANI is distinctively similar to that in Bronze Age Eastern Europe, suggesting that it is tracking a movement of people that affected both regions and that likely spread the unique features shared between Indo-Iranian and Balto-Slavic languages. Our results suggest that a language ancestral to Indo-Iranian was spoken on the Steppe ~4000 years before present.

Show More

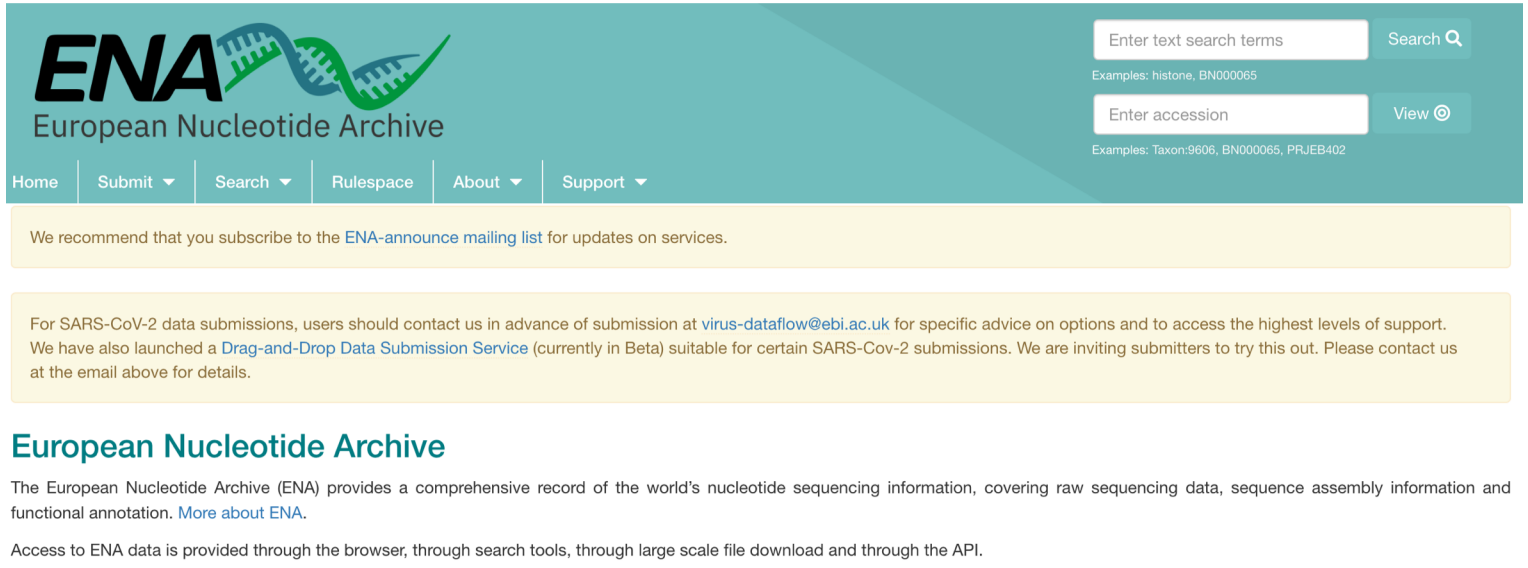
View:	XML
	XML (STUDY)
Download:	XML
	XML (STUDY)
Navigation:	Show
Read Files:	Hide
Publications:	Show
Related ENA Records:	Show

Secondary Study Accession:	ERP115161
Study Title:	Genome wide ancient DNA from 523 ancient individuals sheds light on genetic exchanges between the St... Show More

Let’s download a file from ENA

```
$ wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR458/009/ERR4589279/ERR4589279.fastq.gz
```

```
$ ls
```



Download multiple files

```
$ touch list.txt # create an empty file
$ cat list.txt

$ nano list.txt # open the list with nano for editing
# 1. copy paste links to files using keyboard shortcuts
# 2. close file ctrl+x > press Y > press ENTER
$ cat list.txt

$ wget -i list.txt
```

Lets look at one of the files

```
$ cat ERR4589279.fastq.gz
```

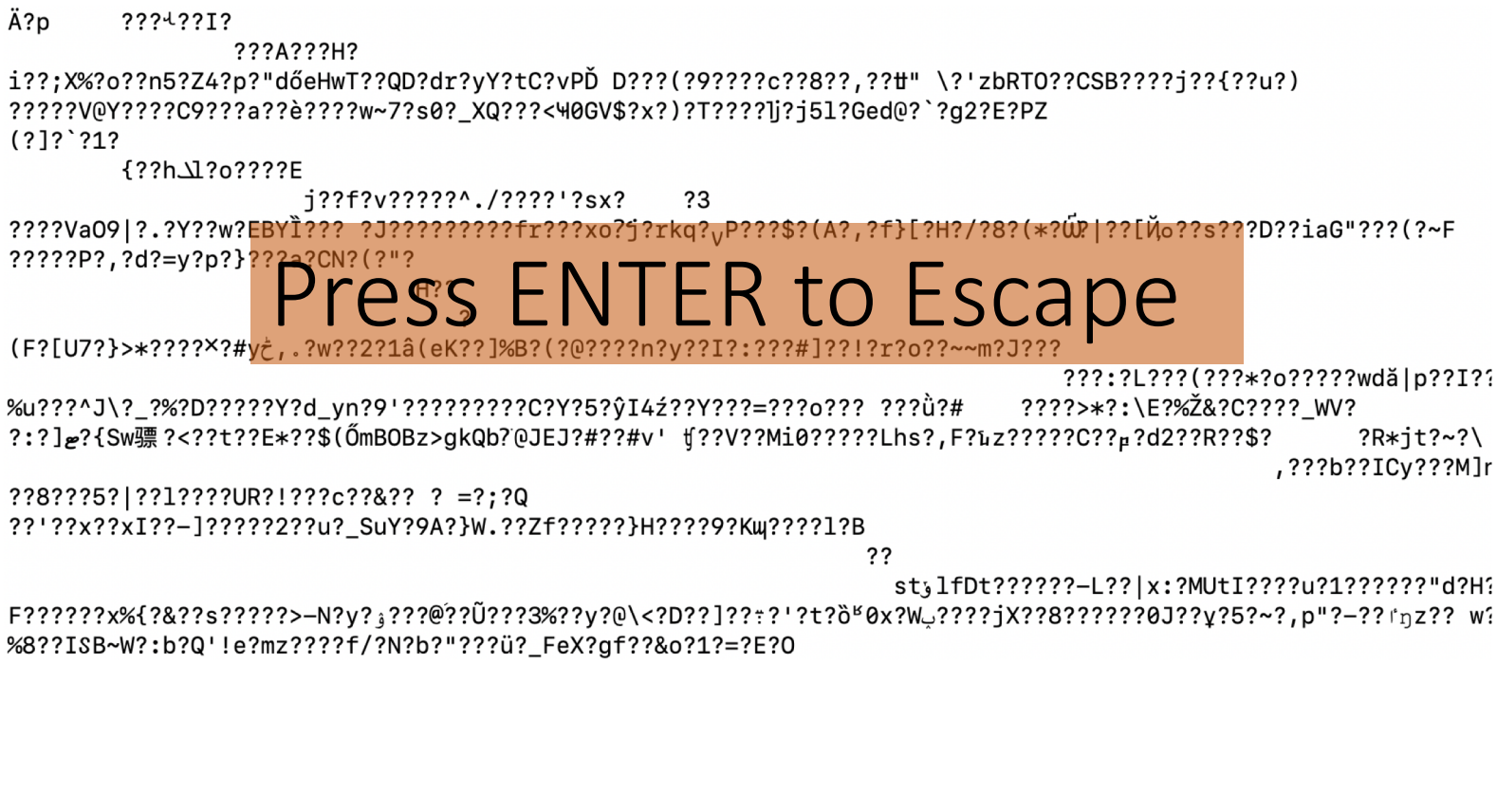
Lets look at one of the files

$ cat ERR4589279.fastq.gz

```
Ä?p     ???ᶜ??I?
                ???A???H?
i??;X%?o??n5?Z4?p?"dőeHwT??QD?dr?yY?tC?vPĎ D???(?9????c??8??,??ŧ" \?'zbRTO??CSB????j??{??u?)
?????V@Y????C9???a??è????w~7?s0?_XQ???<Ч0GV$?x?)?T????ĵ?j5l?Ged@?`?g2?E?PZ
(?]?`?1?
        {??h⊿l?o????E
                     j??f?v?????^./????'?sx?    ?3
????VaO9|?.?Y??w?EBYÌ??? ?J?????????fr???xo?ʝ?rkq?ᵥP???$?(A?,?f}[?H?/?8?(*?Ŵ?|??[Йo??s???D??iaG"???(?~F
?????P?,?d?=y?p?}???a?CN?(?"?
                              H??
                                 ?
(F?[U7?}>*????ˣ?#yč,。?w??2?1â(eK??]%B?(?@????n?y??I?:???#]??!?r?o??~~m?J???
                                                                        ???:?L???(???*?o?????wdă|p??I??
%u???^J\?_?%?D?????Y?d_yn?9'?????????C?Y?5?ŷI4ź??Y???=???o??? ???ù?#    ????>*?:\E?%Ž&?C????_WV?
?:?]ᵉ?{Sw嘌?<??t??E*??$(ŐmBOBz>gkQb?̀@JEJ?#??#v' ʧ??V??Mi0?????Lhs?,F?ʉz?????C??ᵨ?d2??R??$?      ?R*jt?~?\
                                                                                  ,???b??ICy???M]r
??8???5?|??l????UR?!???c??&?? ? =?;?Q
??'??x??xI??-]?????2??u?_SuY?9A?}W.??Zf?????}H????9?Kщ????l?B
                                                                ??
                                                                  stʒlfDt??????-L??|x:?MUtI????u?1??????"d?H?
F??????x%{?&??s?????>-N?y?ʒ???@??Ũ???3%??y?@\<?D??]??ᵗ?'?t?ȍ0x?W‿????jX??8??????0J??ɣ?5?~?,p"?-??ᵣŋz?? w?
%8??IŚB~W?:b?Q'!e?mz????f/?N?b?"???ü?_FeX?gf??&o?1?=?E?O
```

Lets look at one of the files

$ cat ERR4589279.fastq.gz



Better?

$ gzcat ERR4589279.fastq.gz

```
@ERR4589279.1 NS500217:520:HLYLYBGX5:2:11309:26682:8050
GTGTGGTGGCCCATGCCTGCAATCCCAGCACTT
+
EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
@ERR4589279.2 NS500217:520:HLYLYBGX5:3:12512:4979:3431
GGAGGATCGCTTGAGCCCAGGAGTTCAAGACCAGACTG
+
EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
@ERR4589279.3 NS500217:520:HLYLYBGX5:1:12307:26534:15722
TCACATCACTGCACTCCAGCCTGGATGGCA
+
EEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
@ERR4589279.4 NS500217:520:HLYLYBGX5:4:13410:15867:15793
TGTGGTGGCTCACATCTGTAATCCCAGCACTTTCAGAGGC
+
EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
@ERR4589279.5 NS500217:520:HLYLYBGX5:1:23312:5178:18212
GATCAGGAGTTCGAGACCAGCCTGAT
+
EEEEEEEEEEEEEEEEEEEEEEEEEEE
```

What can we tell from the fastq file?

- @unique read name
- Actual sequence of nucleotides
- Always +
- ASCII encoded base quality scores

```
@ERR4589279.1 NS500217:520:HLYLYBGX5:2:11309:26682:8050
GTGTGGTGGCCCATGCCTGCAATCCCAGCACTT
+
EEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEEE
Quality character   !"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJ
                    |         |         |         |         |
ASCII Value         33        43        53        63        73
Base Quality (Q)    0         10        20        30        40
```

Saving storage space via gzip

- Gzip is used for data compression

```
$ gzip -l ERR4589279.fastq.gz

$ gunzip ERR4589279.fastq.gz
# 19,713,281 bytes

$ gzip ERR4589279.fastq
# 3,743,402 bytes
```

Saving storage space via gzip

- Gzip is used for data compression

```
compressed uncompressed  ratio uncompressed_name
   3743402     19713281  81.0% ERR4589279.fastq
```

$ gzip -l ERR4589279.fastq.gz

$ gunzip ERR4589279.fastq.gz
# 19,713,281 bytes

$ gzip ERR4589279.fastq
# 3,743,402 bytes

Piping

- Lets you pass a message to the next command

```
$ ls -l

$ ls -l | sed -e "s/[aeio]/u/g"
```

Piping

```
$ gzcat ERR4589279.fastq.gz | head
```

Piping

```
$ gzcat ERR4589279.fastq.gz | head

$ gzcat ERR4589279.fastq.gz | head -n 5
```

Piping

```
$ gzcat ERR4589279.fastq.gz | head

$ gzcat ERR4589279.fastq.gz | head -n 5

$ gzcat ERR4589279.fastq.gz | tail -n 5
```

Piping

```
$ gzcat ERR4589279.fastq.gz | head

$ gzcat ERR4589279.fastq.gz | head -n 5

$ gzcat ERR4589279.fastq.gz | tail -n 5

$ gzcat ERR4589279.fastq.gz | head -n 20 | tail -n 5
```

Counting

```
$ man wc

$ gzcat ERR4589279.fastq.gz | wc -l

$ gzcat ERR4589279.fastq.gz | head -n 20 | tail -n 5 | wc -l
```

Grep

- Search for a particular character or string in a text file

```
$ gzcat ERR4589279.fastq.gz | grep @
```

Grep

- Search for a particular character or string in a text file

```
$ gzcat ERR4589279.fastq.gz | grep @

$ gzcat ERR4589279.fastq.gz | grep @ERR | wc -l
```

Exercise

- 1. Go to this article https://www.science.org/doi/10.1126/sciadv.aaz5344
- 2. Find ID's of the two youngest individuals in this publication
- 3. Search for an ENA accession number # hint they start with prj
- 4. Go to ENA and download mitochondrial fastq files for the youngest individuals # mtDNA files will have MT in the file name
  - How many lines does each file have?
  - How many reads does each file contain?
  - Count the number of reads that contain the sequence TGCACTAC

Let’s visualize the fastq file

- Download fastqc
- https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
- Should be quick ~50 MB
- Install

$ wget ftp://ftp.sra.ebi.ac.uk/vol1/run/ERR345/ERR3457596/LBG002.A0101.1_S0_L003_R1_001.fastq.gz

- File > Open > LBG002.A0101.1_S0_L003_R1_001.fastq.gz

Questions

- How many total sequences are there?
- What is the sequence length observed?
- How do quality scores vary along the read?

Questions for me?

- ainash_childebayeva@eva.mpg.de
- ainash.childebayeva@gmail.com