

BI694
Bioinformatics & Phylogenetics
Winter Semester 2017

WEEK 5

Sequence Databases
Connecting to Remote Machines



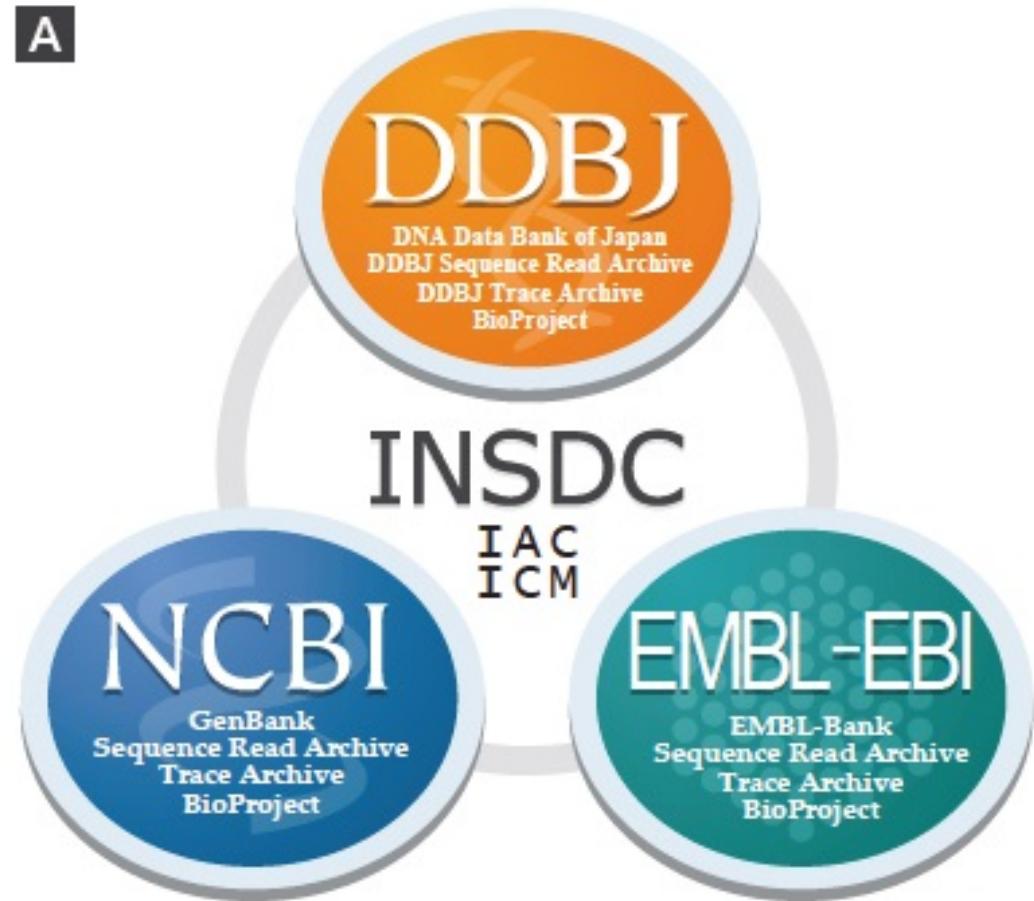
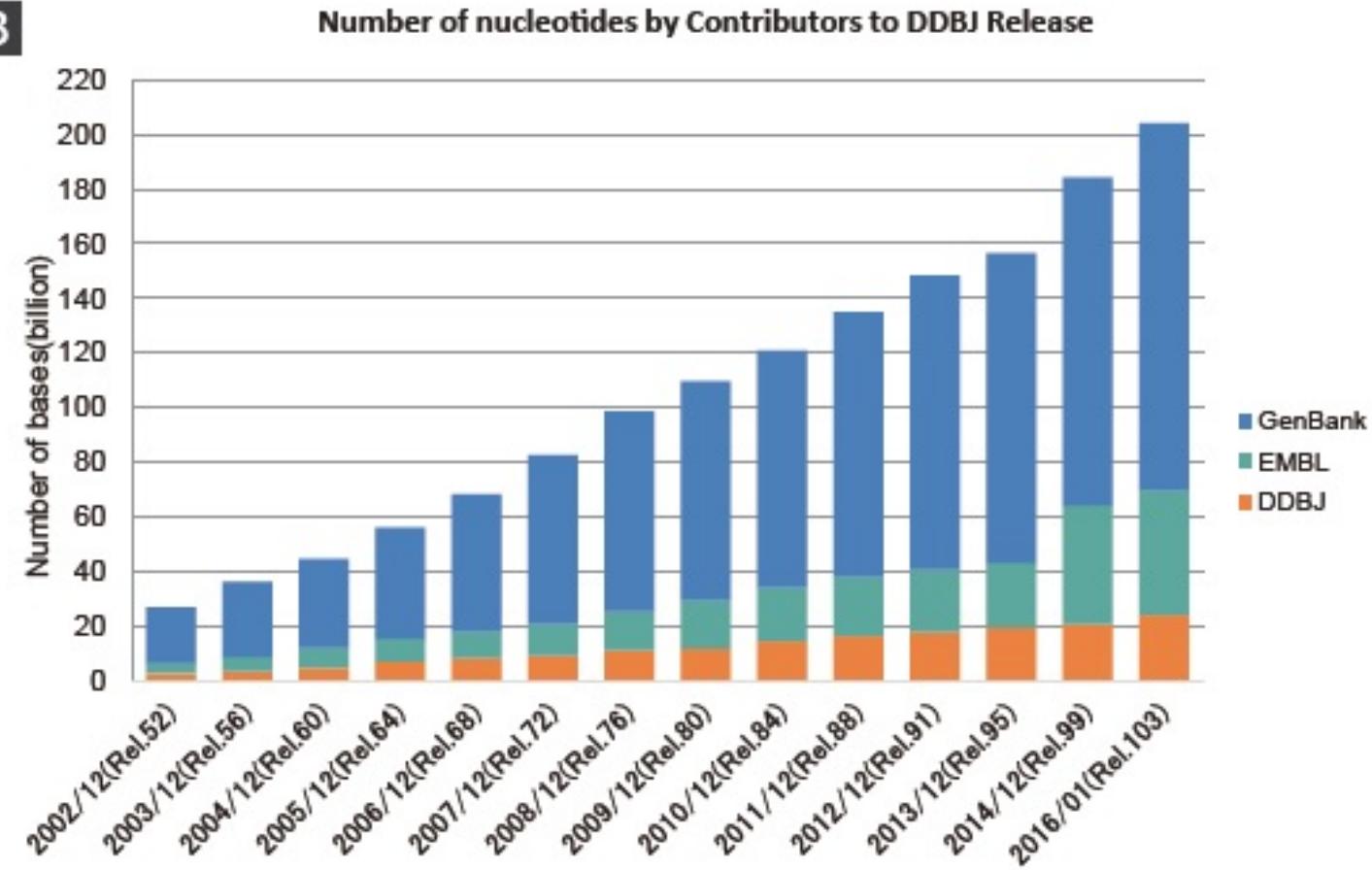
National Center for Biotechnology Information



European Molecular Biology Laboratory



DNA Data Bank of Japan

A**B**

The mission of UniProt is to provide the scientific community with a comprehensive, high-quality and freely accessible resource of protein sequence and functional information.

UniProtKB

UniProt Knowledgebase

Swiss-Prot (555,426)

 Manually annotated and reviewed.

TrEMBL (89,396,316)

 Automatically annotated and not reviewed.

UniRef

Sequence clusters



UniParc

Sequence archive



Proteomes



News



Forthcoming changes

Planned changes for UniProt

UniProt release 2017_08

Curation of human immunoglobulin genes: a fruitful collaboration between UniProtKB/Swiss-Prot and IMGT | Cross-references to ELM

UniProt release 2017_07

A pseudogene turns into an active DNA methyltransferase dedicated to male fertility

News archive

Protein spotlight

A Taste Of Light

August 2017

Light gave life a chance to be. Without it, our planet would not be inhabited by so many living beings of all shapes and sizes. Over time, animals, plants and all sorts of microorganisms have emerged and evolved using this source of photons in different ways. Like hosts of other creatures, we use light for vision so that we can discern

Getting started



UniProt data

Download latest release

Get the UniProt data

Statistics

View Swiss-Prot and TrEMBL statistics

How to cite us

The UniProt Consortium

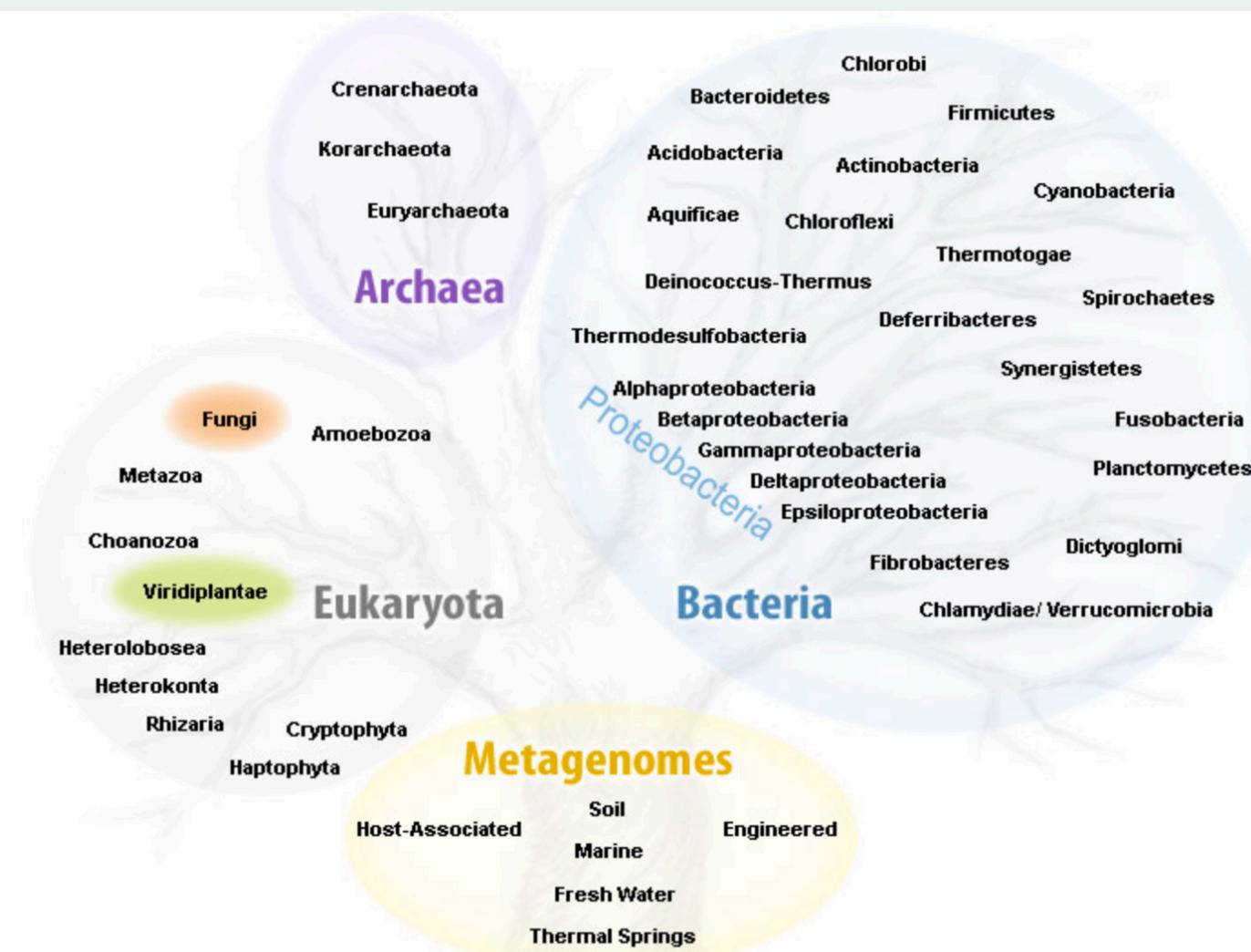
Text search

Our basic text search allows you to search all the resources available

BLAST

Find regions of similarity between your sequences

Sequence alignments

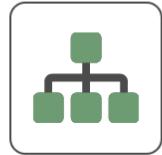
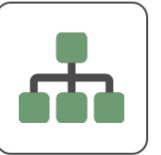
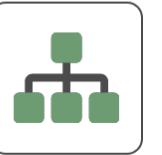


"Tree of Life" drawing by Leila Hornick, copyright 2005

TREE OF LIFE: To use the tree navigation: click a branch name and select the name of the organism of your interest.

[Species ▾](#)[Tools ▾](#)[Info ▾](#)[Download ▾](#)[Help ▾](#)[Cart](#)[Subscribe](#)

Metazome quick search (advanced)

[Flagships](#)[All genomes and families](#)[Early Release Genomes](#)[All released species](#)[Aedes aegypti](#)[Anopheles gambiae](#)[Arthropod](#)[Bilateria](#)[Bombyx mori](#)[Branchiostoma floridae](#)[Caenorhabditis briggsae](#)[Search in](#)[for](#)[GO](#)

About Metazome

3.2

The Metazome project organizes the proteomes of twenty-four metazoans into gene families defined at nine ancestral nodes on the metazoan evolutionary tree. Families of orthologous and paralogous genes that represent the modern descendants of ancestral gene sets are constructed at key phylogenetic nodes. These families allow easy access to clade-specific orthology/paralogy relationships as well as clade-specific genes and gene expansions. Where possible, each gene has been annotated with PFAM, KOG, KEGG, and PANTHER assignments, and publicly available annotations from RefSeq, UniProt, EnsEMBL, and JGI are hyperlinked and searchable.

News (details...)

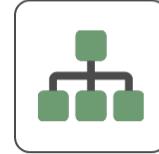


(2016-04-11)
Metazome codebase migrated...

System Status (2017-09-12 04:50)

- ✓ Search
- ✓ BLAST
- ✓ Database
- ✓ BLAT
- ✓ MzMine

Phytozome quick search (advanced)

[Flagships](#)[All genomes and families](#)[Early Release Genomes](#)[All released species](#)[Amaranthus
hypocochondriacus
v1.0](#)[Amborella
trichopoda v1.0](#)[Ananas comosus
v3](#)[Angiosperm](#)[Aquilegia coerulea
v3.1](#)[Arabidopsis halleri
v1.1](#)[Arabidop
v2](#)[Search in](#)[for](#)[GO](#)

About Phytozome

12.1.2

Phytozome, the Plant Comparative Genomics portal of the Department of Energy's Joint Genome Institute, provides JGI users and the broader plant science community a hub for accessing, visualizing and analyzing JGI-sequenced plant genomes, as well as selected genomes and datasets that have been sequenced elsewhere. As of release v12.1, Phytozome hosts 77 assembled and annotation genomes, from 74 viridiplantae species. Forty-three of these genomes have been sequenced, assembled and annotated with JGI Plant Science program resources. By integrating this large collection of plant genomes into a single resource and performing comprehensive and uniform annotation and analyses, Phytozome facilitates accurate and insightful comparative genomics studies.

All gene sets in Phytozome have been annotated with KOG, KEGG, ENZYME, Pathway and the InterPro family of protein analysis tools. Inparanoid pairwise orthology and paralogy groups have

News (details...)



(2017-09-07) **Chickpea and *Bsylaticum* added**

(2017-07-31) **4 new genomes released**

(2017-05-09) **Chromochloris genome released**

System Status (2017-09-12 04:55)

- ✓ Search
- ✓ BLAST
- ✓ Database
- ✓ BLAT
- ✓ PhytoMine

Pfam 31.0 (March 2017, 16712 entries)

The Pfam database is a large collection of protein families, each represented by **multiple sequence alignments** and **hidden Markov models (HMMs)**. [More...](#)

QUICK LINKS[**SEQUENCE SEARCH**](#)[**VIEW A PFAM ENTRY**](#)[**VIEW A CLAN**](#)[**VIEW A SEQUENCE**](#)[**VIEW A STRUCTURE**](#)[**KEYWORD SEARCH**](#)[**JUMP TO**](#)**YOU CAN FIND DATA IN PFAM IN VARIOUS WAYS...**

Analyze your protein sequence for Pfam matches

View Pfam annotation and alignments

See groups of related entries

Look at the domain organisation of a protein sequence

Find the domains on a PDB structure

Query Pfam by keywords

Go **Example**

Enter any type of accession or ID to jump to the page for a Pfam entry or clan, UniProt sequence, PDB structure, etc.

Or view the [help](#) pages for more information

Recent Pfam [blog](#) posts[☒ Hide this](#)

[**Pfam 31.0 is released**](#) (posted 8 March 2017)

Pfam 31.0 contains a total of 16712 families and 604 clans. Since the last release, we have built 415 new families, killed 9 families and created 11 new clans. We have also been working on expanding our clan classification; in Pfam 31.0, over 36% of Pfam entries are placed within a clan. The new “stuff” [...]

[**Pfam train online**](#) (posted 8 December 2016)

We now have an online Quick Tour that provides a brief introduction to the Pfam protein families



By sequence

By domain architecture

InterProScan sequence search

This form allows you to scan your sequence for matches against the InterPro protein signature databases, using InterProScan tool.

Enter or paste a protein sequence in FASTA format (complete or not - e.g. PMPIGSKERPTFFEIFKTRCNKADLGPISLN), with a maximum length of 40,000 amino acid long.

Please note that you can only scan one sequence at a time.

Analyse your protein sequence

InterProScan



InterProScan is a sequence analysis application (nucleotide and protein sequences) that combines different protein signature recognition methods into one resource.

[More about InterProScan.](#)

[SEQanswers Home](#)

 User Name User Name Remember Me?
 Password Log in

[Register](#)
[FAQ](#)
[Community ▾](#)
[Calendar](#)
[Today's Posts](#)
[Search](#)

You are currently viewing the SEQanswers forums as a guest, which limits your access. [Click here to register now](#), and join the discussion

» Site Navigation 	
» About SEQanswers	
» Next Gen Summaries	
» Forums	
» Wiki	
» Instrument Map	

» Log in 	
User Name:	<input type="text"/>
Password:	<input type="password"/>
<input checked="" type="checkbox"/> Remember Me?	
<input type="button" value="Log in"/>	
Not a member yet? Register Now!	

» Online Users: 323 	
3 members and 320 guests	

» New Posts 				
Title, Username, & Date	Last Post	Replies	Views	Forum
Searching for microRNAs Bhumika Arora	Today 02:19 AM by Bhumika Arora 	0	37	Bioinformatics
Hisat2 multithreading not working. tirohia	Yesterday 11:10 PM by dpryan 	5	194	RNA Sequencing
Introduction marthastevens47	Yesterday 10:05 PM by marthastevens47 	0	26	Introductions
IndexOutOfBoundsException... ArtVandelay	Yesterday 04:54 PM by ArtVandelay 	0	64	Bioinformatics
Yes .. BBMap can do that! GenoMax	Yesterday 02:35 PM by TomHarrop 	163	33,462	Bioinformatics
what percentage range is normal for undetermined reads in a lane? zeam	Yesterday 01:30 PM by nucacidhunter 	1	67	Illumina/Solexa

» Our Sponsors 	

» Recent Job Postings 	
Bioinformatician/sta... 08-02-2017 05:30 AM by Bukowski	

» New Ion Torrent Machines...media roundup 	
Sep 01, 2015 - 10:15 AM - by ECO	
Ion just lifted the embargo so all the bloggers are furiously tapping out their preview details and opinions of Thermo's new sequencers... Here are a few:	



Community

Log In

Sign Up

Add New Post

Live search: start typing...

or

Classic search

Limit to: all time ▾

<prev • 51,427 results • page 1 of 1715 • next >

Sort by: update ▾

0 votes	0 answers	2 views	1 (800) 927 2480 QuickBooks POS Support Number QuickBooks Technical Issue Support Number	written just now by ajayrudeelee • 0
0 votes	0 answers	4 views	1 (800) 927 2480 QuickBooks POS Error Support Number QuickBooks Desktop Error Support Number	written 3 minutes ago by ajayrudeelee • 0
0 votes	0 answers	25 views	Generate a reference-ordered data (ROD) file from FASTA	written 38 minutes ago by omer.asr • 0
0 votes	0 answers	29 views	Rn28s1 28S ribosomal RNA [Mus musculus] annotation in GTF file	written 42 minutes ago by efratk • 0
0 votes	0 answers	37 views	Motifs & domains comparisons between proteins	written 1 hour ago by lessismore • 170
0 votes	0 answers	31 views	Mendelian randomization analysis..... OR and Beta value	written 55 minutes ago by Aamiralizai • 0
0 votes	2 answers	140 views	fastQC for Oxford Nanopore reads	written 4 days ago by efratk • 0
0 votes	0 answers	16 views	How to interpret DEXseq results in terms of significance	

Recent Votes

- Trinity : How To Co-Assemble Different Samples (Tumor And Healthy)
- C: Convert sequence file to fasta format using python
- C: BRCA1 and BRCA2 database's for NGS diagnostics purposes
- C: BRCA1 and BRCA2 database's for NGS diagnostics purposes
- A: BRCA1 and BRCA2 database's for NGS diagnostics purposes
- C: Blastn: Blast short reads against genomes and retain only good hits
- A: AWK: if value falls within a range, print sum of values

Recent Locations • All »

- United Kingdom, just now
- Trento, IT, 2 minutes ago
- Belgium, 5 minutes ago

Recent Awards • All »

- Commentator to Titus • 400
- Scholar to Pierre Lindenbaum ♦ 97k
- Teacher to Kevin Blighe • 340
- Scholar to Alex Reynolds ♦ 20k
- Teacher to Alex Reynolds ♦ 20k
- Scholar to Kevin Blighe • 340



National Center for Biotechnology Information

GenBank

Several databases available; the nucleotide (**nt**) and protein (**nr**) sequence collection probably most widely used.

RefSeq provides a curated set of gene/protein sequences from selected genomes.

What else is available? ESTs, SRA, TSA, genomes...



National Center for Biotechnology Information

GenBank

Downloading data:

`ftp://ftp.ncbi.nih.gov/blast/db/`

`update_blastdb.pl --passive --decompress nr`



National Center for Biotechnology Information

GenBank

Downloading data:

`ftp://ftp.ncbi.nih.gov/blast/db/`

`update_blastdb.pl --passive --decompress nr`



National Center for Biotechnology Information

GenBank

Getting custom data – the GenBank query builder



National Center for Biotechnology Information

GenBank

Let's do some sequence searches on our VMs locally
and against the remote database



National Center for Biotechnology Information

GenBank

What about jobs that take a long time?

We can run these on a remote server!



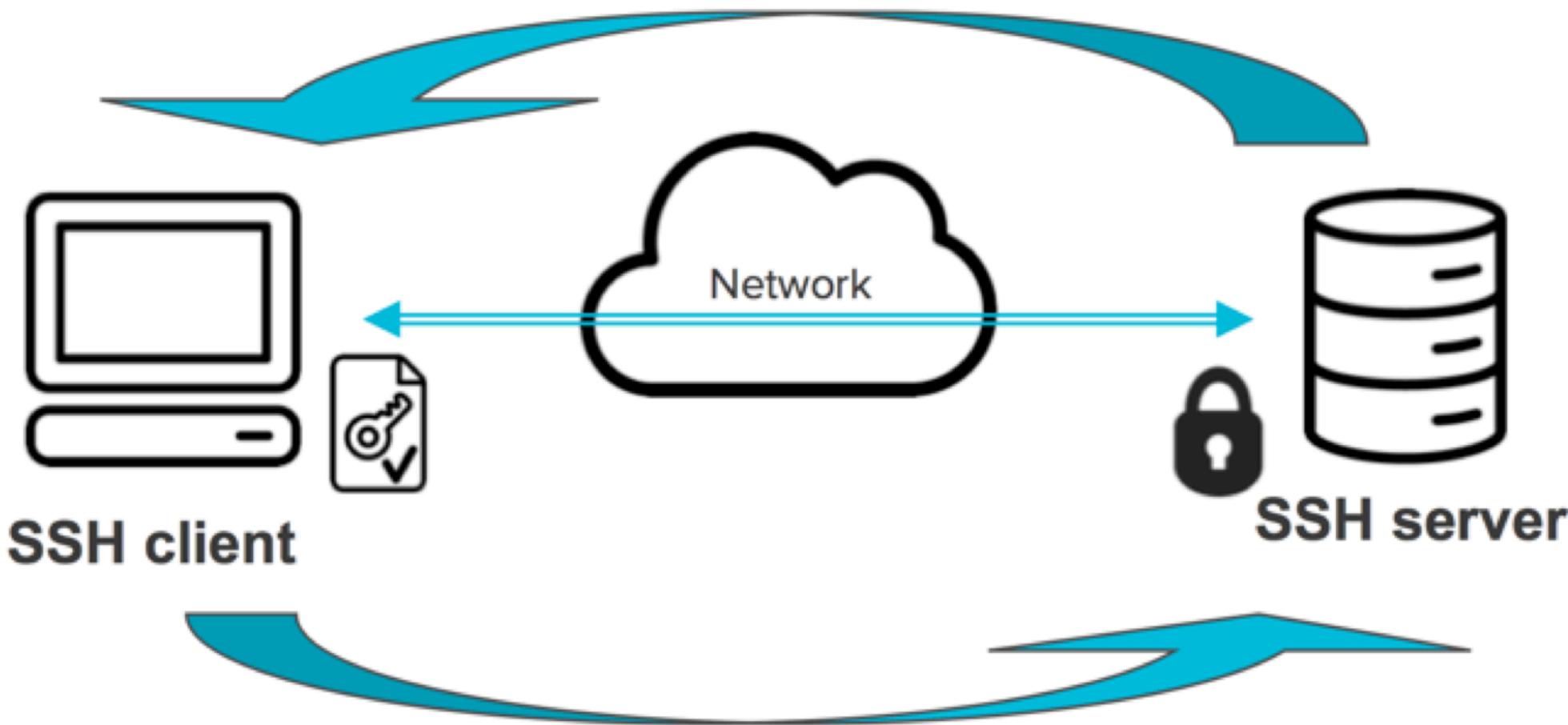
SSH...

OpenSSH

KEEPING YOUR COMMUNIQUÉS SECRET

1) **Server** authentication:

Server proves its identity to the client



2) **User** authentication:

Client proves user's identity to the server

ssh username@168.123.185.xxx

Change your password!

scp – securely copy files over the network

Check the system using htop

Where are we at?