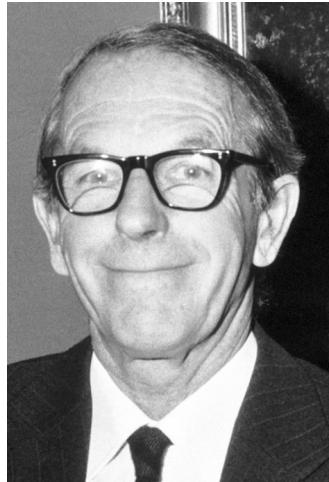


NextGen Sequencing and High-Performance Computing



QB Bootcamp, Day 2
Thursday, 2 September 2021
10:00am - 10:30am

It wasn't until the mid 1970s that efficient methods for sequencing DNA were developed.



Fred Sanger

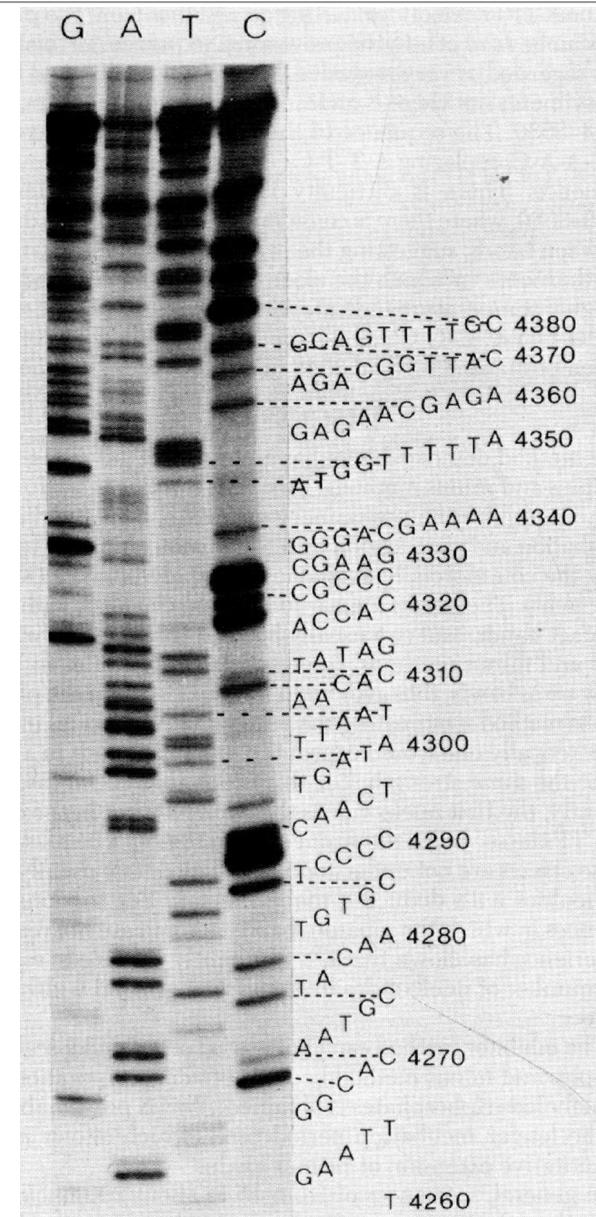


Wally Gilbert



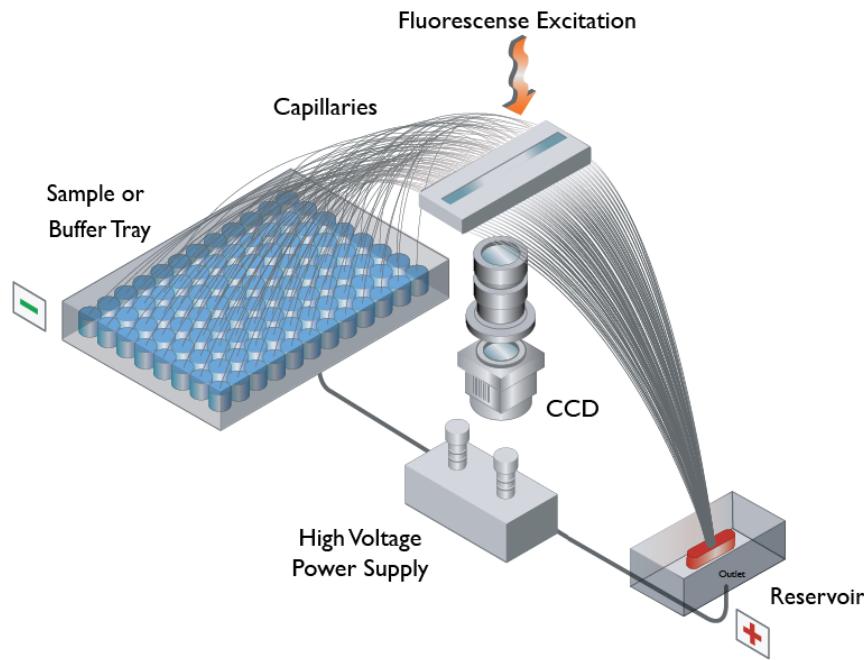
1980 Nobel Prize in Chemistry

"for their contributions concerning the determination of base sequences in nucleic acids."

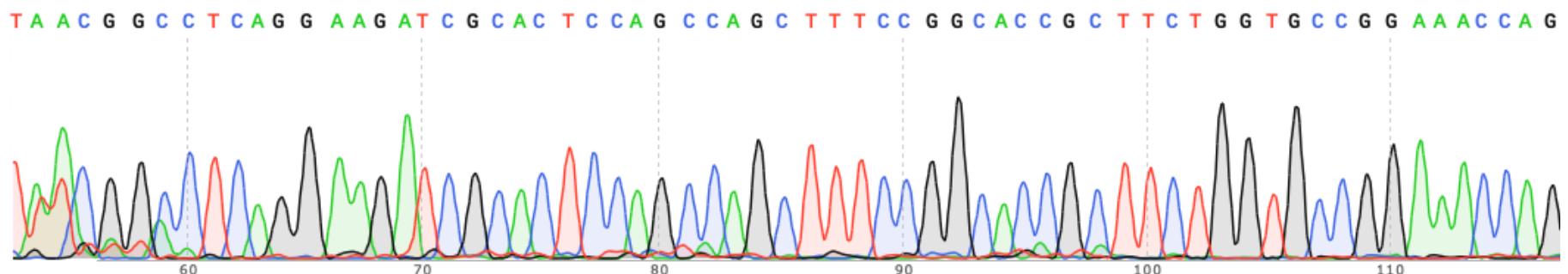


Sanger et al., PNAS, 1977

The efficiency of DNA sequencing increased gradually with the development of fluorescent capillary electrophoresis and with automation



ABI 3730xl Genetic Analyzer
(2304 samples/day)



The human genome was sequenced using Sanger sequencing.

Human genome project (finished in 2003):

3.2 billion nucleotides

\$5 billion (2019 dollars)

That's ~4 million individual Sanger sequencing reactions (not counting overlaps)!



Human genome sequencing facility
at the Whitehead Institute (1994)



The human genome in book form,
Wellcome Collection

Illumina sequencing was announced in 2006. It has become the standard high-throughput DNA sequencing method

NextSeq 500 sequencing run:

reads: 300,000,000

read length: 300 nt

time: 1 day

cost: \$2,000

That's ~**30 human genomes** of DNA!

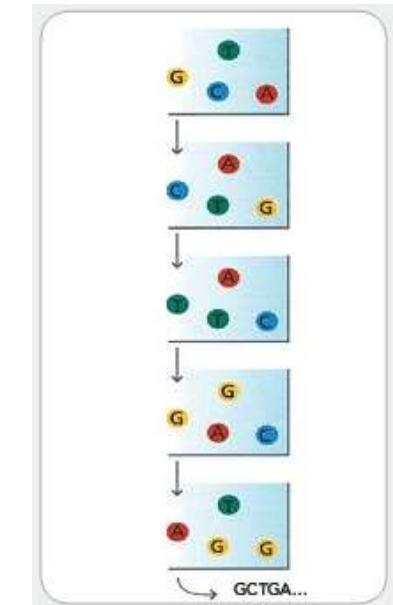
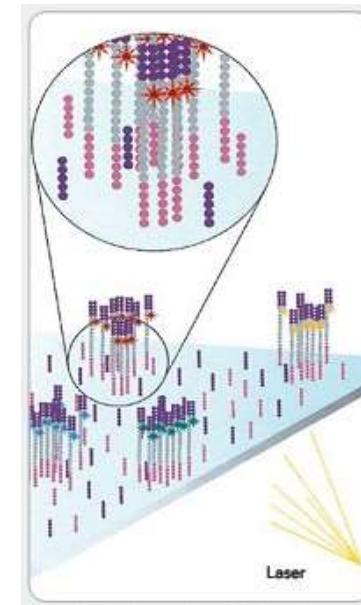
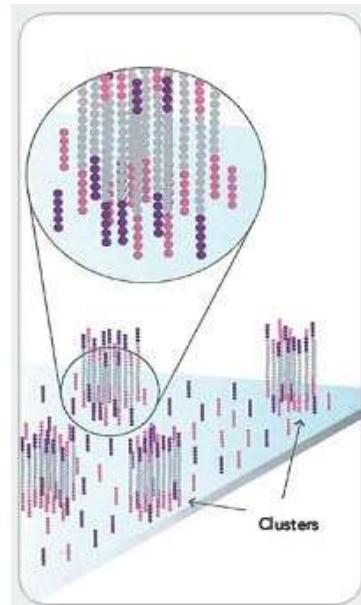
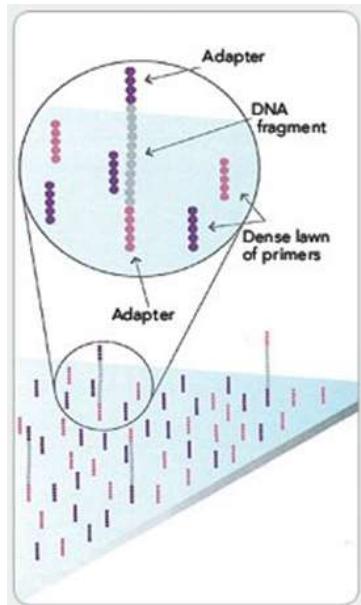
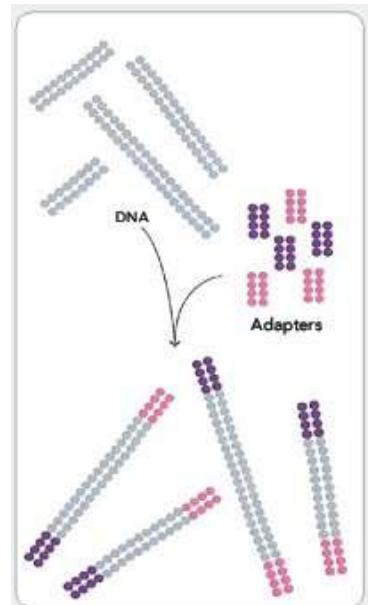


NextSeq 500

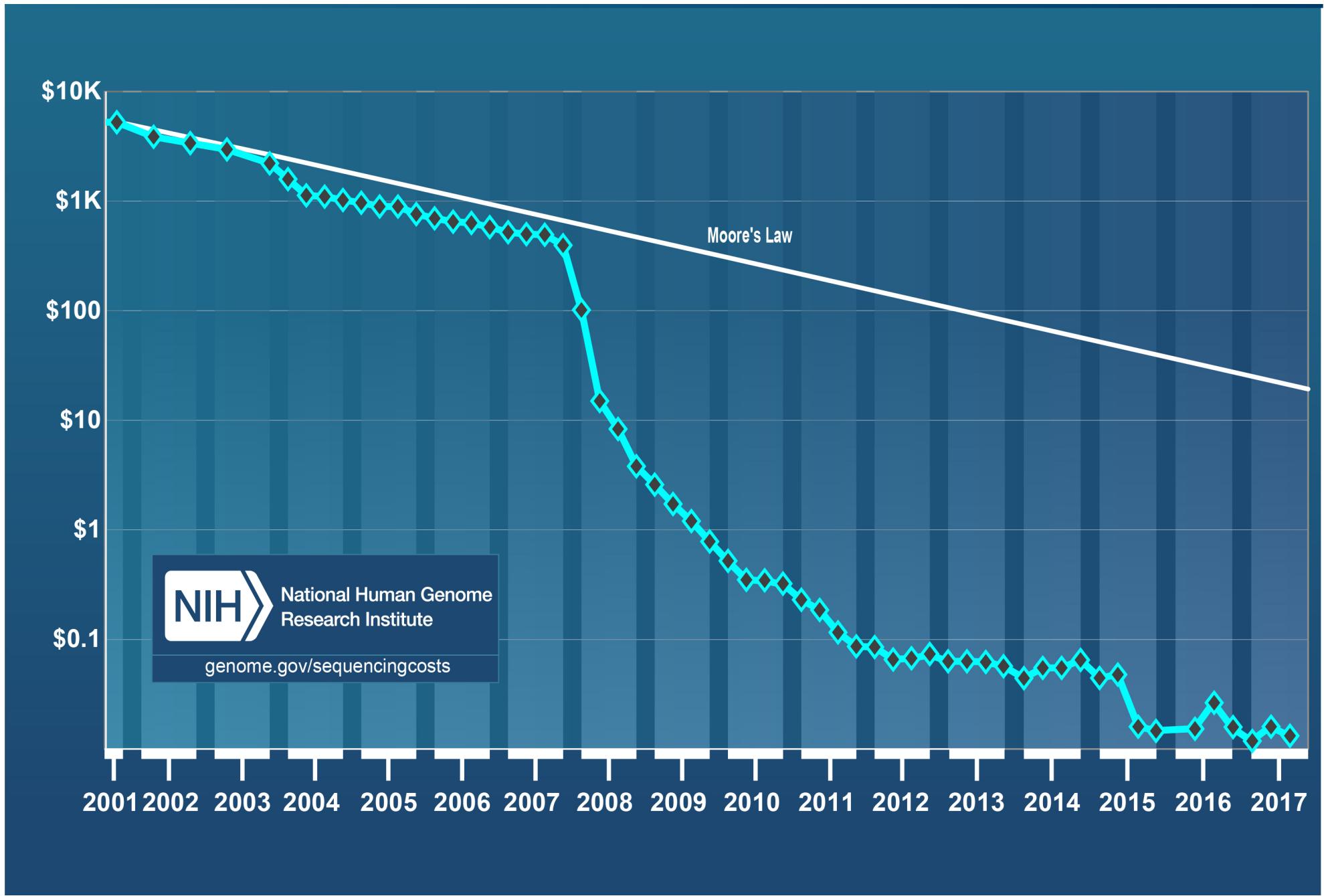
this is where the DNA goes



Flow cell



DNA sequencing has undergone an unprecedented technological revolution



DNA sequencing vs. computing



ABI 3730xl Genetic Analyzer (2007)



5 years



Illumina HiSeq 2500 (2012)



Palm Pilot (1997)



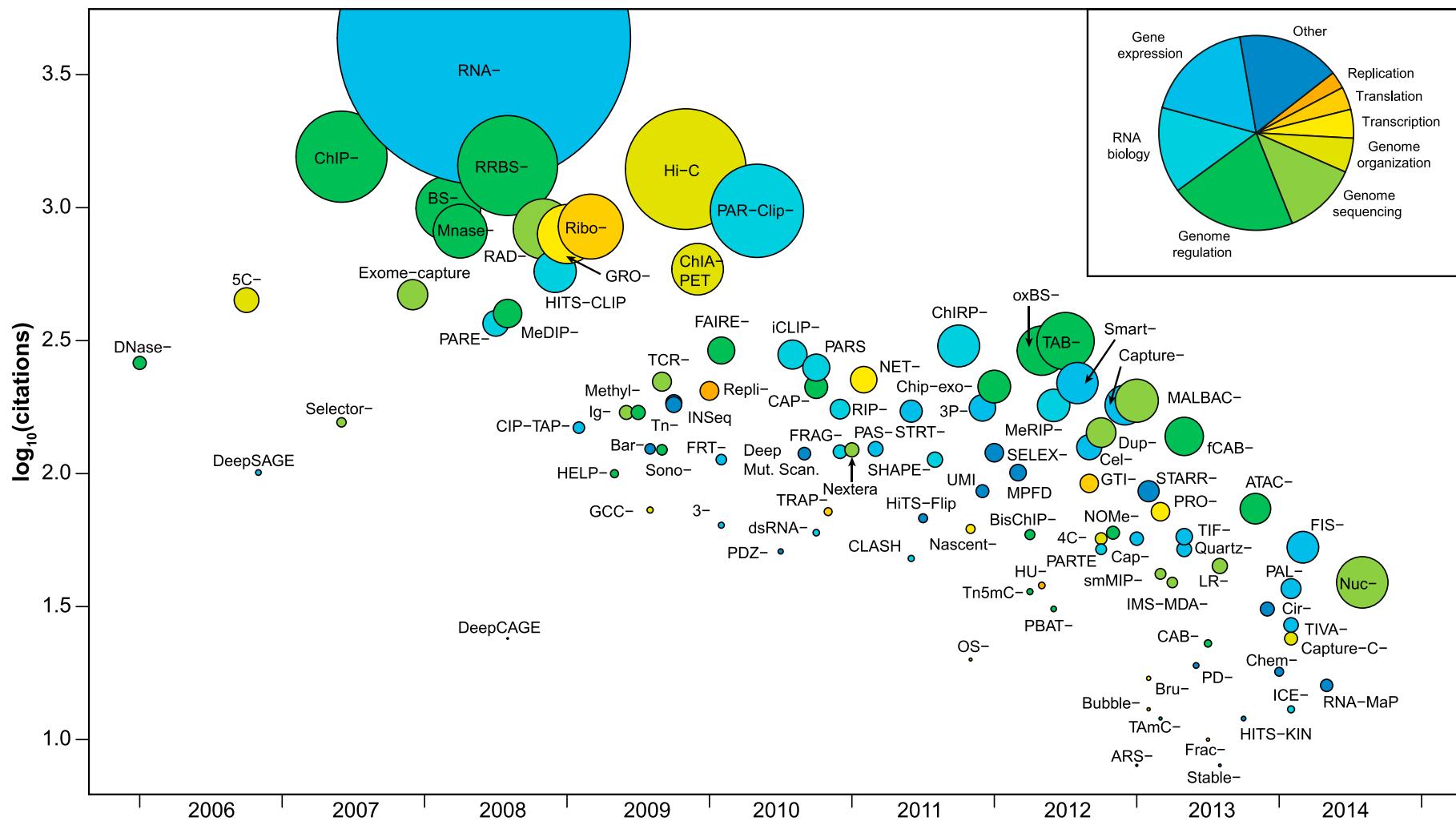
20 years



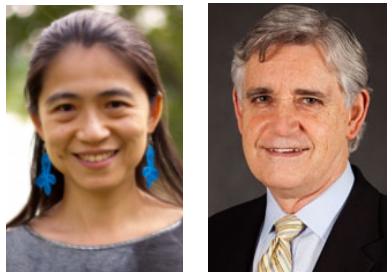
iPhone X (2017)

There are many different ways of using high-throughput sequencing to study biology

X-seq experiments



The Stillman uses high-throughput DNA sequencing to study the dynamics of DNA replication initiation and progression



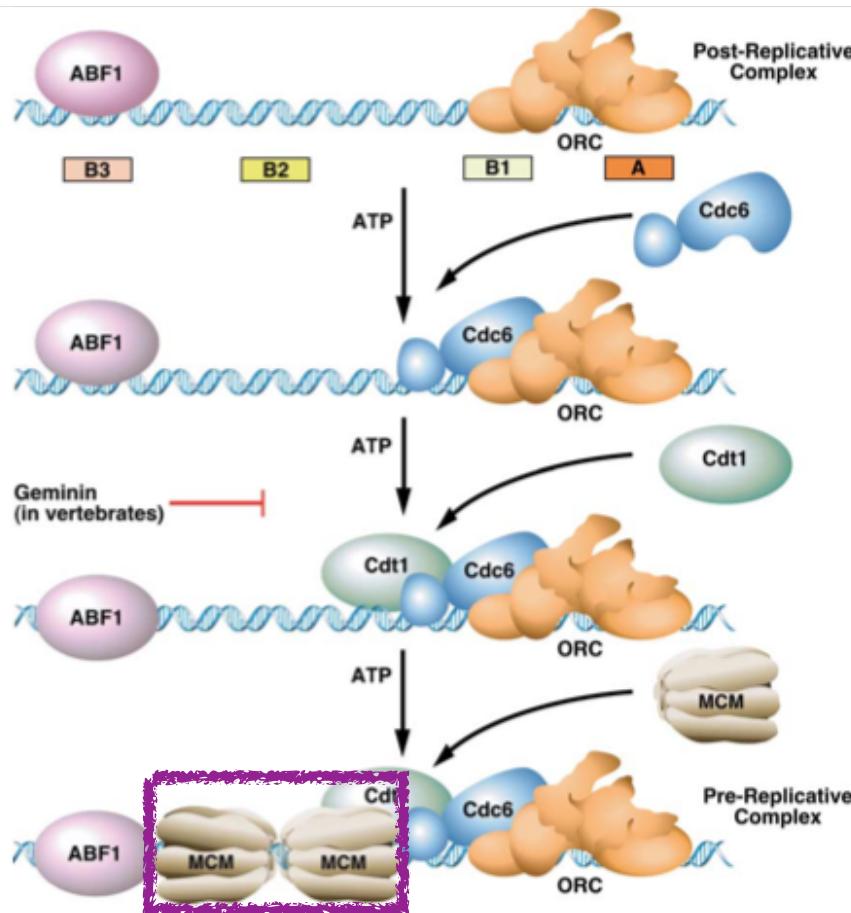
Domain within the helicase subunit Mcm4 integrates multiple kinase signals to control DNA replication initiation and fork progression

Yi-Jun Sheu^a, Justin B. Kinney^a, Armelle Lengronne^b, Philippe Pasero^b, and Bruce Stillman^{a,1}

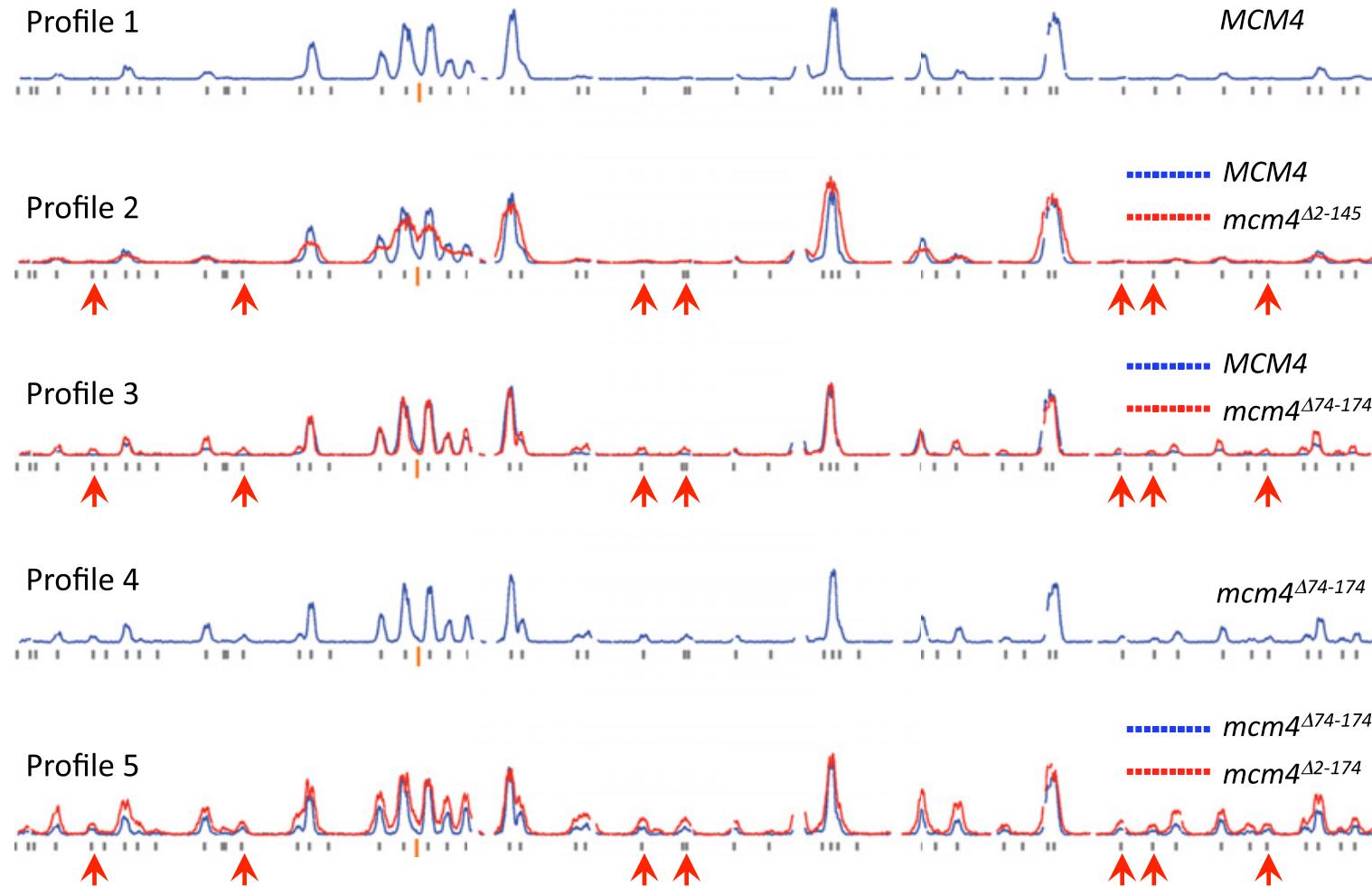
PNAS | Published online April 16, 2014 | E1899–E1908



S. cerevisiae



Here are some examples of the published replication profiles

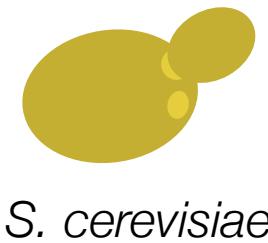


~300 separate loci direct DNA replication initiation in *Saccharomyces cerevisiae*

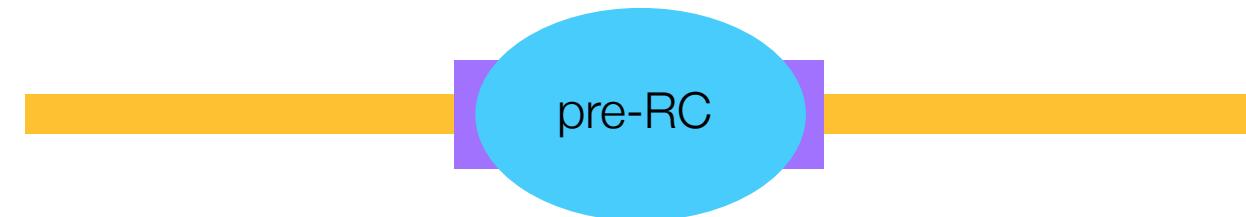
ARS: autonomously replicating sequence

— old ssDNA

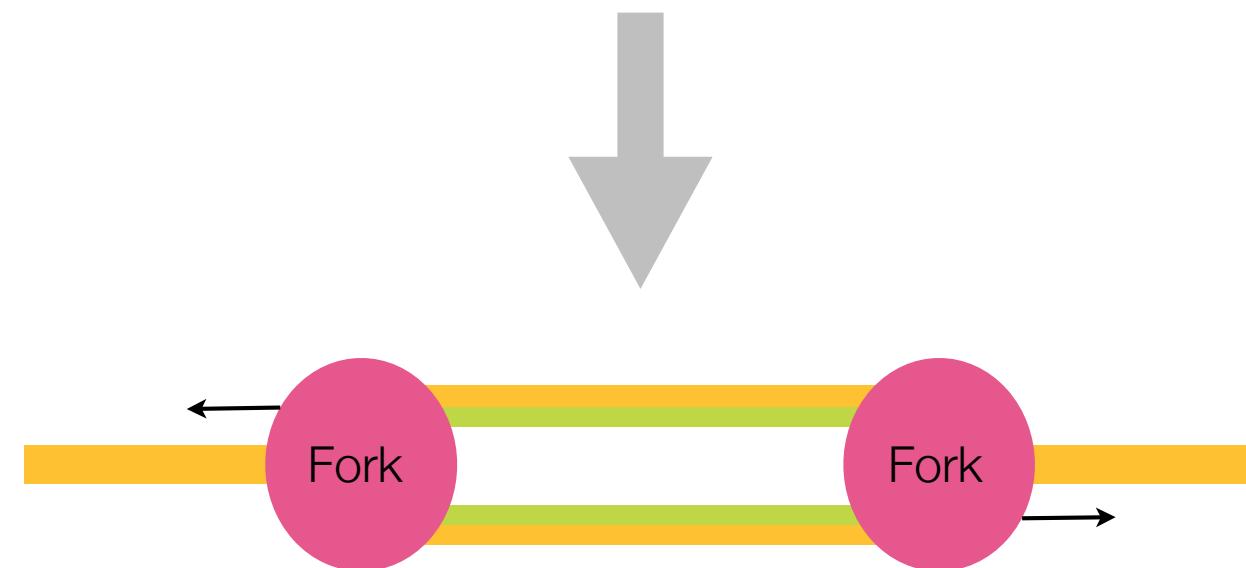
— new ssDNA



G1 phase



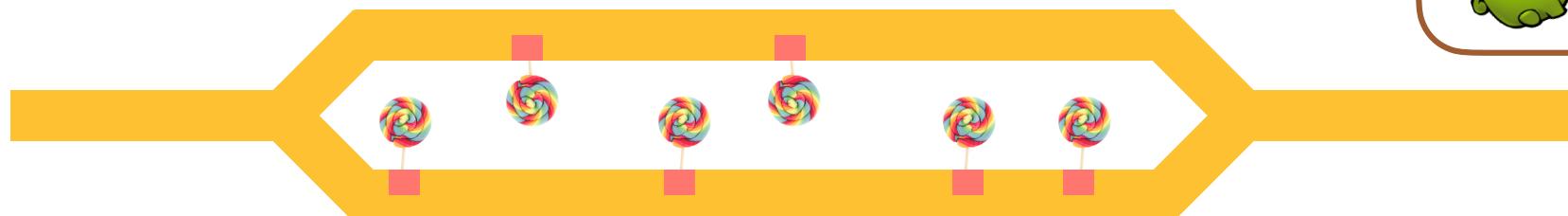
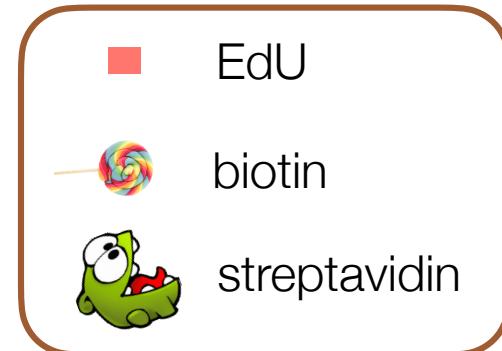
S phase



Newly replicated DNA can be isolated using an EdU pull-down assay

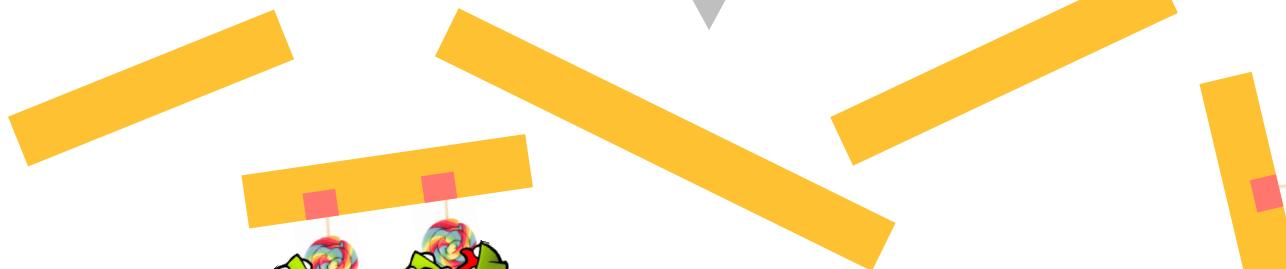
DNA of cells arrested in G1 with α-factor

Release cells into S-phase
EdU incorporation during replication

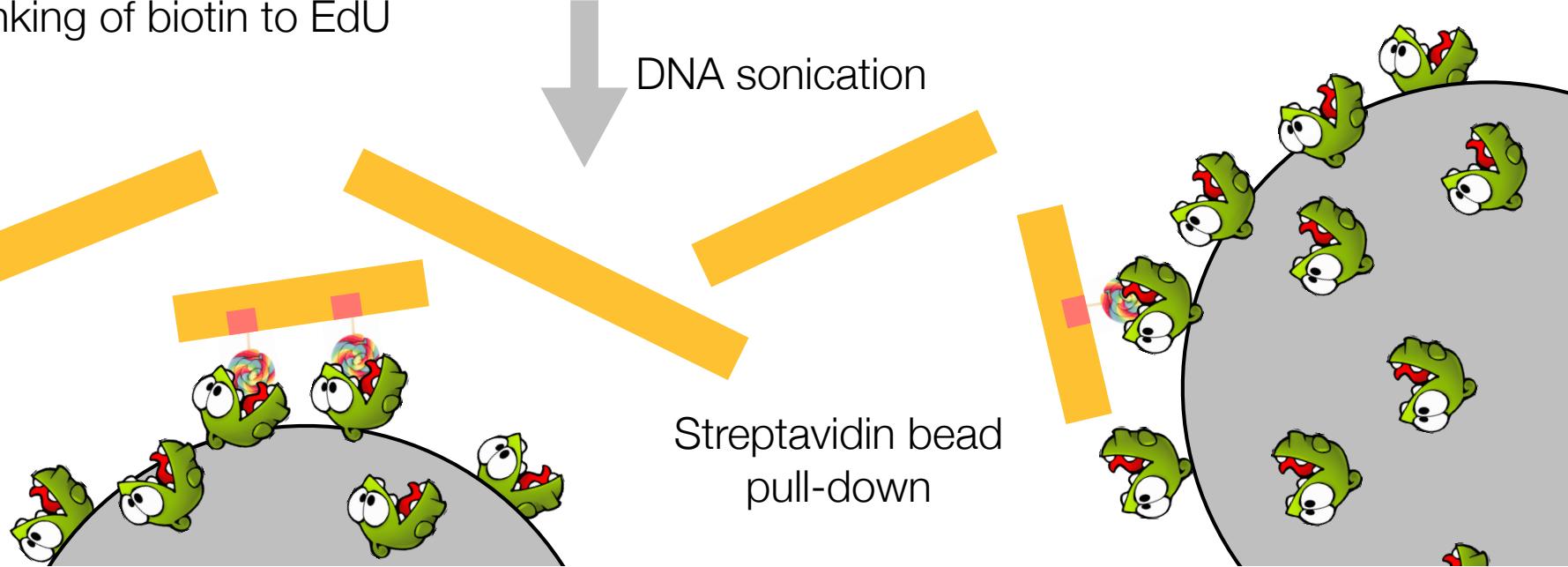


Click-iT linking of biotin to EdU

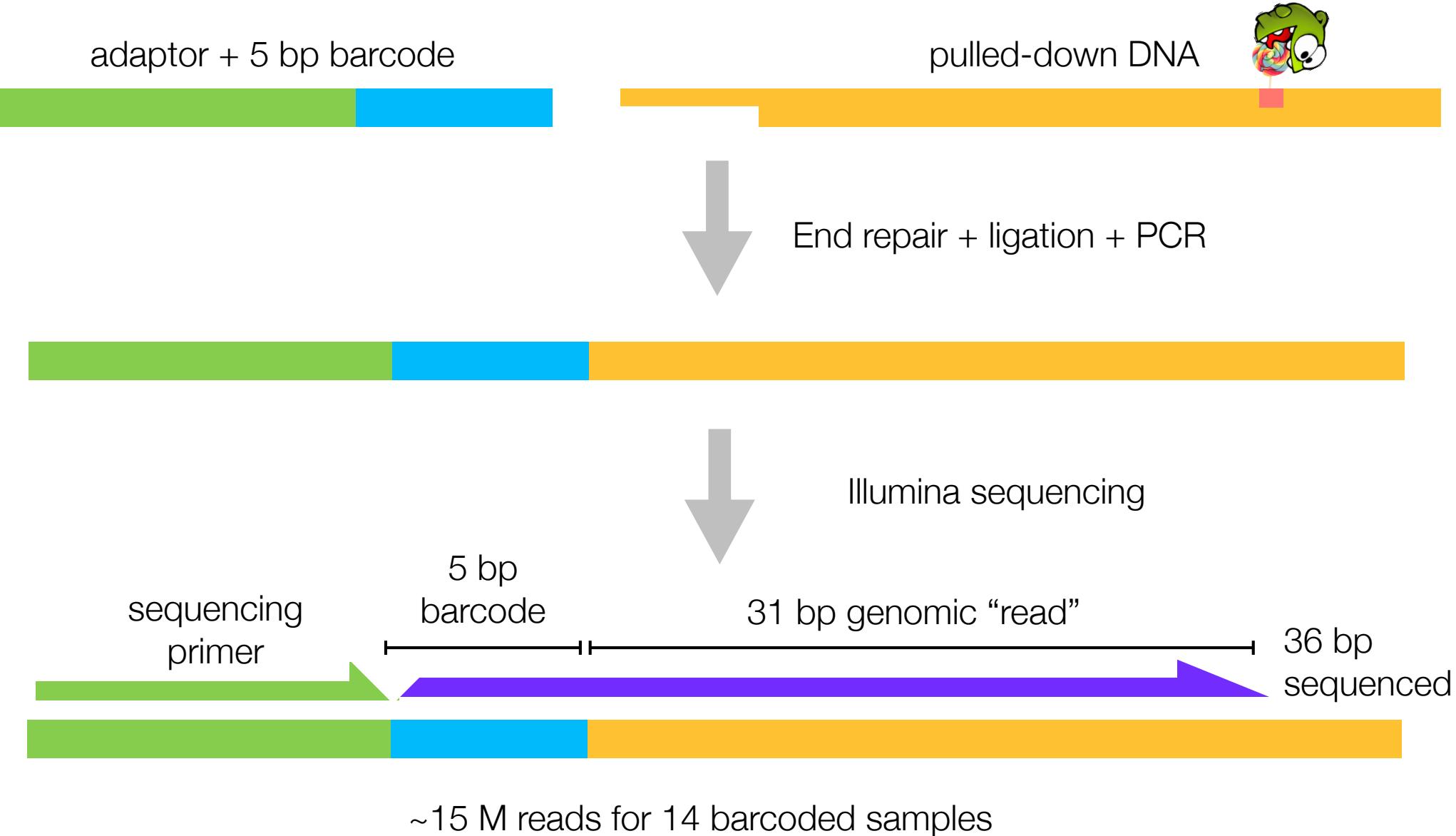
DNA sonication



Streptavidin bead pull-down



Sequencing of pulled-down DNA allows replication to be mapped genome-wide



We will analyze FASTQ read files from 4 different samples

There are four FASTQ files in the reads/ directory

Each FASTQ files is ~60-100 megabytes

```
[[jkinstructor@bamdev1 ~]$ ls
elzar_exercise  elzar_exercise.tar.gz
[[jkinstructor@bamdev1 ~]$ ls -lah elzar_exercise/reads/
total 302M
drwxr-s---  2 jkinstructor wsbs students  4.0K Sep  2 07:05 .
drwxr-sr-x  6 jkinstructor wsbs students  4.0K Sep  2 06:34 ..
-rw-r----- 1 jkinstructor wsbs students 70M Aug 30 2016 A1.fastq
-rw-r----- 1 jkinstructor wsbs students 97M Aug 30 2016 B1.fastq
-rw-r----- 1 jkinstructor wsbs students 68M Aug 30 2016 C1.fastq
-rw-r----- 1 jkinstructor wsbs students 69M Aug 30 2016 D1.fastq
[jkinstructor@bamdev1 ~]$
```

This is what a FASTQ file looks like (circa 2009)

read 1

```
[jkinstructor@bamdev1 ~]$ cd ~/elzar_exercise/reads/
[jkinstructor@bamdev1 reads]$ head -n 20 A1.fastq
@HANNIBAL_0056:7:1:9620:1049#0/1
GTGGTTAGTATATGGTGCAAAAGTGGTATAA
+HANNIBAL_0056:7:1:9620:1049#0/1
ggggggaeadffffccdfaaaaefgfgggg
@HANNIBAL_0056:7:1:1070:1061#0/1
CGAACACAAAGATCTCGTTCTACTTTTTTG
+HANNIBAL_0056:7:1:1070:1061#0/1
f`[facdddfJdcfaa^c fcf dcfffc]
@HANNIBAL_0056:7:1:4279:1052#0/1
TATCCACTACCGCTATACTGGATTCTGACTC
+HANNIBAL_0056:7:1:4279:1052#0/1
hghhhhhhhhhghghghhhhhfhhhfhhhg
@HANNIBAL_0056:7:1:4413:1064#0/1
AAGAAAACGTGCCACCATTGAGTACATCAAC
+HANNIBAL_0056:7:1:4413:1064#0/1
hhhhhhhhcfffffgghhhhgdhffghfb
@HANNIBAL_0056:7:1:5309:1059#0/1
AGTATACTGTGTATATAATAGATATGGAACG
+HANNIBAL_0056:7:1:5309:1059#0/1
bf`ebfcffcfbdbeac^ cfcdffffdf
[jkinstructor@bamdev1 reads]$
```

read 2

read 3

read 4

read 5

The information
for each read is
split over 4 lines

← @name
← sequence
← +name
← quality scores

The yeast genome is in FASTA format

```
[jkinstructor@bamdev1 ~]$ cd elzar_exercise/genome/  
[jkinstructor@bamdev1 genome]$ head genome.fasta  
>1 ref|NC_001133| [org=Saccharomyces cerevisiae] [strain=S288C] [moltype=genomic] [chromosome=I] [note=R64-1-1]  
CCACACCACACCCACACACCCCACACACACACACACACACACACACA  
CATCCTAACACTACCCTAACACAGGCCATAATCTAACCCCTGGCCAACCTGTCTCTCAACTT  
ACCCCTCATTACCCCTGCCCTCACTCGTTACCCCTGTCCATTCAACCATAACCACTCCGAAC  
CACCATCCATCCCTCTACTTACTACCACTCACCCACCGTTACCCCTCCAATTACCCATATC  
CAACCCACTGCCACTTACCCATTACCCATTACCATCCACCATGACCTACTCACCATAAC  
TGTCTTCTACCCACCATATTGAAACGCTAACAAATGATCGTAATAACACACACACGTGCT  
TACCTTACCACTTTATACCACCACTGCCATACTCACCCCTCACTGTATACTGATT  
TACGTACGCACACGGATGCTACAGTATATACCATCTCAAACCTTACCCCTACTCTCAGATT  
CACTTCACTCCATGCCCATCTCACTGAATCAGTACCAAATGCACTCACATCATTATG  
[jkinstructor@bamdev1 genome]$
```

Each header line starts with ‘>’

The corresponding sequence follows,
usually split over lines 80bp long

genome.fasta contains sequences #1 to #16, representing the 16 chromosomes

```
[jkinstructor@bamdev1 genome]$ cat genome.fasta | grep '>'  
>1 ref|NC_001133| [org=Saccharomyces cerevisiae] [strain=S288C] [moltype=genomic] [chromosome=I] [note=R64-1-1]  
>2 ref|NC_001134| [org=Saccharomyces cerevisiae] [strain=S288C] [moltype=genomic] [chromosome=II] [note=R64-1-1]  
>3 ref|NC_001135| [org=Saccharomyces cerevisiae] [strain=S288C] [moltype=genomic] [chromosome=III] [note=R64-1-1]  
>4 ref|NC_001136| [org=Saccharomyces cerevisiae] [strain=S288C] [moltype=genomic] [chromosome=IV] [note=R64-1-1]  
>5 ref|NC_001137| [org=Saccharomyces cerevisiae] [strain=S288C] [moltype=genomic] [chromosome=V] [note=R64-1-1]  
>6 ref|NC_001138| [org=Saccharomyces cerevisiae] [strain=S288C] [moltype=genomic] [chromosome=VI] [note=R64-1-1]  
>7 ref|NC_001139| [org=Saccharomyces cerevisiae] [strain=S288C] [moltype=genomic] [chromosome=VII] [note=R64-1-1]  
>8 ref|NC_001140| [org=Saccharomyces cerevisiae] [strain=S288C] [moltype=genomic] [chromosome=VIII] [note=R64-1-1]  
>9 ref|NC_001141| [org=Saccharomyces cerevisiae] [strain=S288C] [moltype=genomic] [chromosome=IX] [note=R64-1-1]  
>10 ref|NC_001142| [org=Saccharomyces cerevisiae] [strain=S288C] [moltype=genomic] [chromosome=X] [note=R64-1-1]  
>11 ref|NC_001143| [org=Saccharomyces cerevisiae] [strain=S288C] [moltype=genomic] [chromosome=XI] [note=R64-1-1]  
>12 ref|NC_001144| [org=Saccharomyces cerevisiae] [strain=S288C] [moltype=genomic] [chromosome=XII] [note=R64-1-1]  
>13 ref|NC_001145| [org=Saccharomyces cerevisiae] [strain=S288C] [moltype=genomic] [chromosome=XIII] [note=R64-1-1]  
>14 ref|NC_001146| [org=Saccharomyces cerevisiae] [strain=S288C] [moltype=genomic] [chromosome=XIV] [note=R64-1-1]  
>15 ref|NC_001147| [org=Saccharomyces cerevisiae] [strain=S288C] [moltype=genomic] [chromosome=XV] [note=R64-1-1]  
>16 ref|NC_001148| [org=Saccharomyces cerevisiae] [strain=S288C] [moltype=genomic] [chromosome=XVI] [note=R64-1-1]  
[jkinstructor@bamdev1 genome]$
```

We will map reads to the genome on the cluster,
then analyze the resulting .bed files on our local machines

A1.fastq + genome.fasta



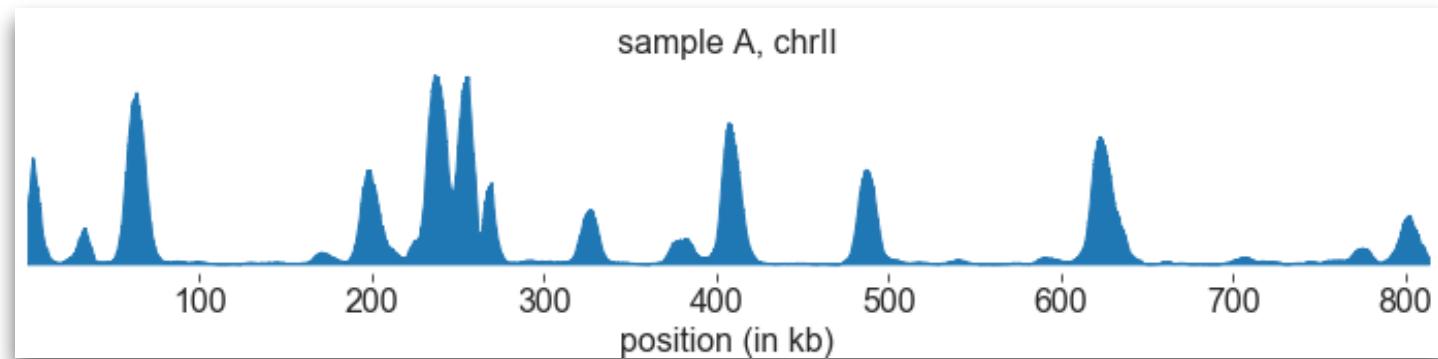
cluster analysis
(bwa + samtools + bedtools)

A1.pileup.bed

```
browser position chrIV:1-1531933
track type=bedGraph visibility=2 name="A1" description="A1"
chrI    1      31     2
chrI    32     62     0
chrI    63     93     1
chrI    94     124    0
chrI    125    155    3
chrI    156    186    0
chrI    187    217    0
chrI    218    248    0
```

chromosome # reads
 window

local analysis
(python)



Elzar is the CSHL's high-performance computer cluster (HPCC)

<http://intranet.cshl.edu/administration/information-technology/hpcc/elzar>

The screenshot shows a web browser window with the URL <http://intranet.cshl.edu/administration/information-technology/hpcc/elzar>. The page is titled "The Essence of Elzar". On the left, there is a sidebar for "Information Technology" with links like "Search Menu", "Home", "Divisions" (4), "Systems & Storage" (1), "Elzar" (9), "Contact", "Containers", "GPU nodes", "Jupyter", "Login/Development nodes", "MATLAB", "UGE (Workload Manager)", "User Environment", and "Workflow Tools". The main content area features a cartoon illustration of a scientist holding a brain. Below it, text describes the Elzar cluster as an institutionally shared high performance computing system introduced in 2020. It details the cluster's architecture, consisting of 50 nodes (2 head/management, 2 dev/login, 46 compute) connected via dual 25 Gbps Ethernet networks, and a DDN GridScaler storage system. The cluster is administered by UGE with a "fair share" resource management policy. The compute nodes run Xeon 6252 processors at 2.1GHz with 24 cores per processor and 768GB of memory. The high memory nodes have dual processors. A small "Feedback" icon is in the bottom right corner.



Todd
Heywood

If you use Elzar *at all* for your work, please cite “NIH Grant S10OD028632-01”

Elzar tutorial on YouTube

https://www.youtube.com/embed/D3wfhM_cQPY

The screenshot shows a Mac OS X desktop environment. In the foreground, a terminal window is open with a root shell session on a CSHL cluster node named 'bamhead1'. The user is demonstrating how to use the 'grid' command to access storage. The terminal output includes commands like 'module avail', 'module load <module>', and various 'grid' and 'findmnt' commands. In the background, a web browser window displays the CSHL Intranet homepage. The browser's address bar shows the URL <https://intranet.cshl.edu/administration/information-technology/hpc/elzar-login>. The page content includes sections on simplified paths, direct access from Mac or Windows, and a Slack workspace setup. A video call interface is visible on the right side of the screen, showing a man identified as Todd Heywood.

root@bamhead1:~\$ cutter:~ heywood\$ cutter:~ heywood\$ ssh he
Welcome to Elzar
See <http://intranet.cshl.edu>
Use the following command
'module avail'
'module load <module>'

Last login: Thu Oct 22 1
[heywood@bamdev1 ~]\$ [heywood@bamdev1 ~]\$ [heywood@bamdev1 ~]\$ [heywood@bamdev1 ~]\$ [heywood@bamdev1 ~]\$ pma
/grid/it/home/heywood
[heywood@bamdev1 ~]\$ ls
backup bnb bnlmod.sh
[heywood@bamdev1 ~]\$ [heywood@bamdev1 ~]\$ [heywood@bamdev1 ~]\$ [heywood@bamdev1 ~]\$ fin
|~/grid/it/home
|~/grid/it/home_nlsas
[heywood@bamdev1 ~]\$ ls
total 197
drwxr-xr-x 4 heywood i
drwxr-xr-x 14 heywood i
-rwxr-xr-x 1 heywood w
drwxrwxr-x 3 heywood i
drwxr-xr-x 2 heywood i
drwxr-xr-x 2 heywood i
-rwxr-xr-x 1 heywood i
drwxr-xr-x 3 heywood i
drwxr-xr-x 2 heywood i
-rw-r--r-- 1 root r
-rwxr-xr-x 1 root r
drwxr-xr-x 2 heywood i
-rwxr-xr-x 1 heywood i
drwxr-xr-x 4 heywood i
drwxr-xr-x 2 heywood i
[heywood@bamdev1 ~]\$

CSH Intranet

HOME GENERAL INFO ADMINISTRATION EDUCATION RESEARCH REQUESTS

| grid:/ware/data_nlease_norepl | grid[:/ware/nlease_norepl]/data/data |

The simplified paths on the left map to the "real" paths on the right (where you should substitute "/mnt/grid/" for "grid["). Note that the "hpc" and "nlsas" labels used to distinguish between performance classes of storage, a distinction that no longer applies (the labels are just names now). The ".norepl" label means that the data at the path is not replicated to another physical location to protect against loss in case of a disaster like fire at the main location.

You may access your storage directly from your Mac or Windows machine using volume or share names. On Macs, open Finder and then select "connect to server" under the Go menu. Then enter "smb://grid-hs". You will see a list of volumes corresponding to the right side of the findmnt output described above, except showing all labs, from which you can then select your volume.

On Windows, open Explorer and click in the address bar, or type **ctrl+L**. Then enter "\grid-hs". Then you will see a list of shares for all labs corresponding to the right side of the findmnt output described above, which you can then select.

We plan on adding volume/share names matching the simplified paths on the left side of the findmnt output.

Slack workspace

An Elzar Users workspace has been set up on Slack. This may or may not turn out to be useful for users to share methods, ask questions, and get answers in a group environment. We'll see! If you use a cshl.edu email address, you can sign up at [this link](#). If you want to sign up without a cshl.edu email address, contact [Todd Heywood](#).

HELPFUL LINKS

- Campus Map
- Emergency Information
- Employee Self Service
- Faculty & Staff Directory (FACES)
- Shuttle Schedule

MENUS

- Blackford Hall
- Blackford Bar
- Hillside Cafe
- Genome Center
- Blackford Dinner

EXTERNAL LINKS

- CSHL External Website
- Labdish Blog
- Newsletter Signup
- Harbor Transcript
- Faculty & Staff

CONNECT WITH CSHL

- [Facebook](#)
- [Twitter](#)
- [YouTube](#)
- [LinkedIn](#)
- [Instagram](#)
- [Email](#)
- [RSS](#)

WHISTLEBLOWER POLICY

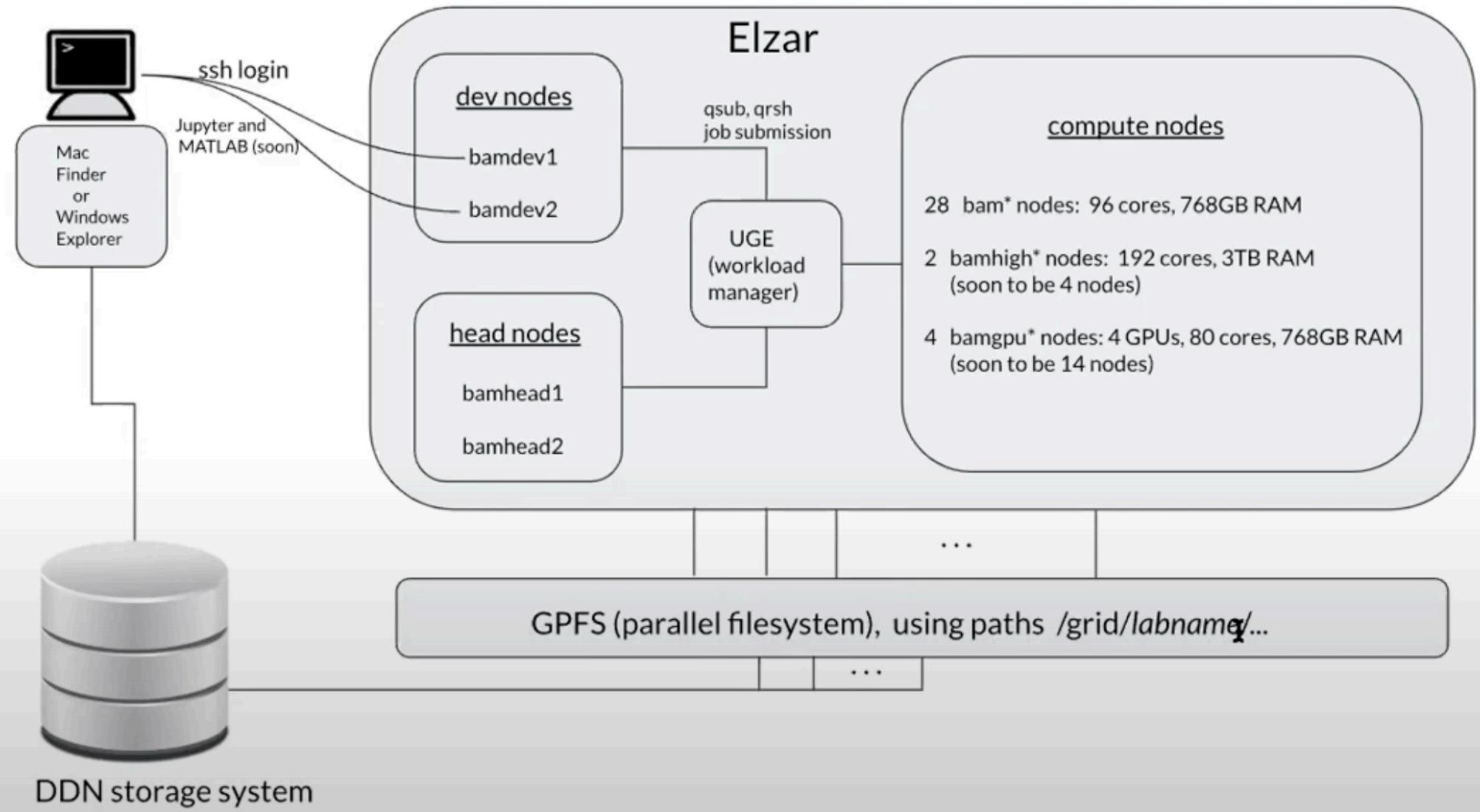
[View Whistleblower Policy](#)

Elzar - Google Slides | elzar barn - Google | Open XDMoD | CNN - Breaking News | Post Attendee - Zoom | Elzar - Google Slides

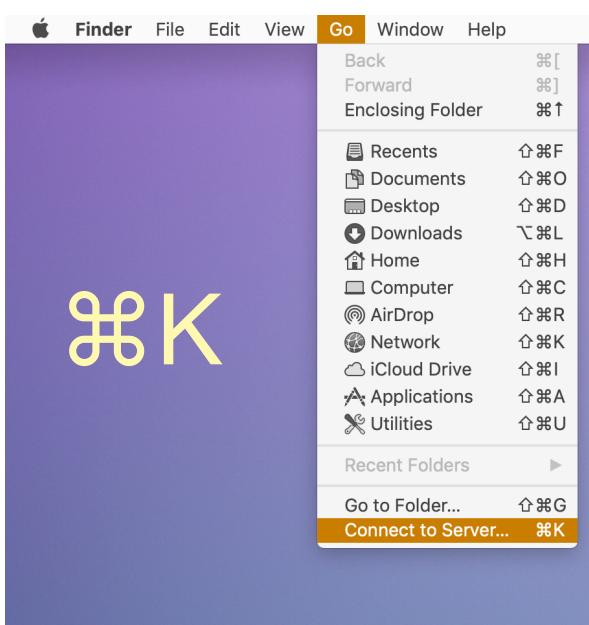
Apps Gmail Quant Quant2 Quant3 Quant4 QuantH HPC Vets Elzar Easybuild CSHL Wiki knomes CSHL User Manager hastebin CSHL Blog

Todd Heywood

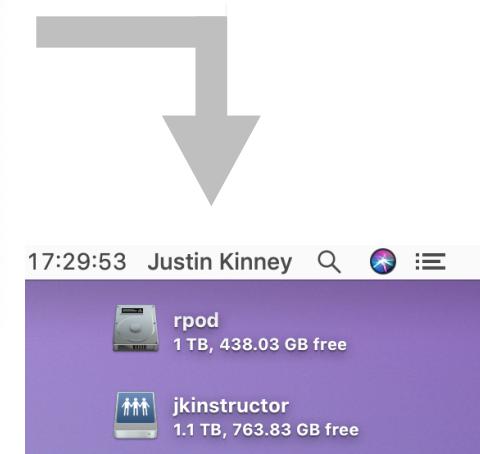
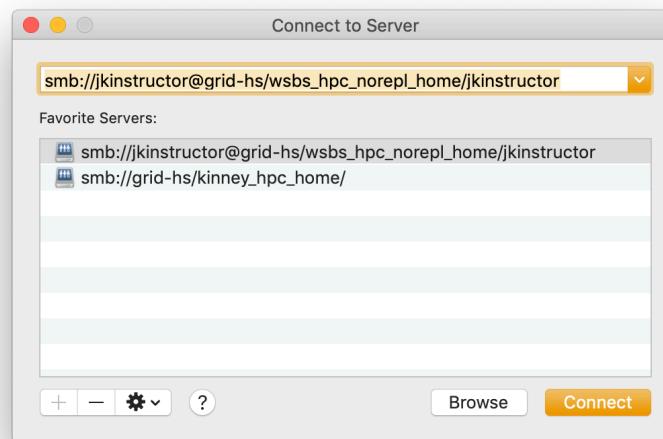
Architecture of Elzar



The Elzar disk can be mounted using smb

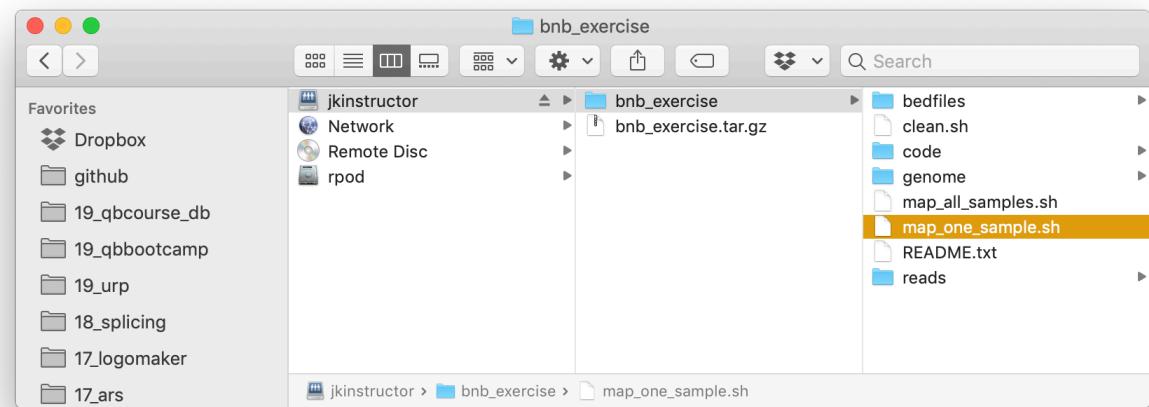


smb://you@grid-hs/wsbs_hpc_norepl_home/you
or
smb://you@grid-hs/yourlab_hpc_home/you



```
map_one_sample.sh
1 #!/usr/bin/env bash
2
3 # map_one_sample.sh
4 #
5 # Creates pileup files in .bed format for 4 Illumina
6 # samples
7 echo "Running single_process.sh..."
8
9 # Assign variables governing mapping
10 batch="A1"
11 read_length="31"
12
13 # Create output directories
14 echo "Setting up working area..."
15 ./clean.sh
16 mkdir mappings pileups
17
18 # Create bwa index for genome
19 echo "Creating index for genome..."
20 bwa index genome/genome.fasta
```

A screenshot of a terminal window titled 'map_one_sample.sh'. The window displays a shell script with numerous lines of code. At the bottom left, it says '110 Words, Line 1, Column 1'. At the bottom right, it says 'Tab Size: 4'.



To do this morning:

1. Copy **elzar_exercises.tar.gz** from **21_qbbootcamp/** to your home directory on Elzar
2. Map one sample of reads to genome using **map_one_sample.sh**
3. Submit four mapping jobs to cluster using **map_all_samples.sh**
4. Copy .bed files to local machine
5. This afternoon: Visualize replication profiles using Python.