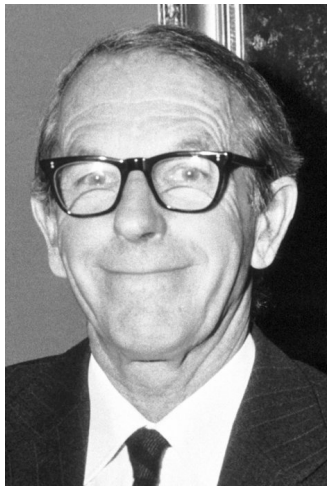
NextGen Sequencing and High-Performance Computing

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.

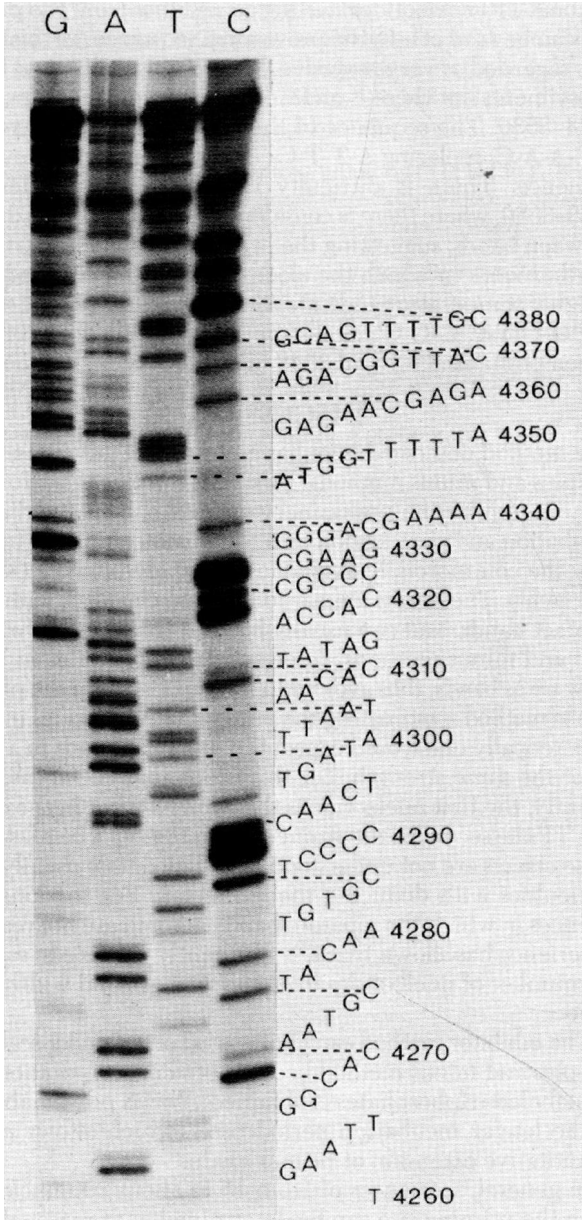


Fred Sanger

Wally Gilbert

1980 Nobel Prize in Chemistry

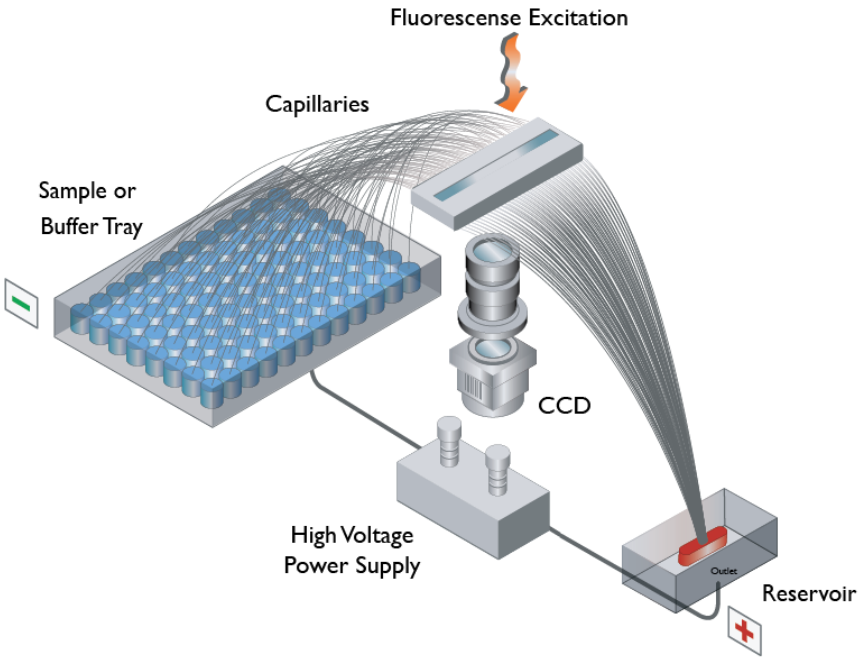
"for their contributions concerning the determination of base sequences in nucleic acids."

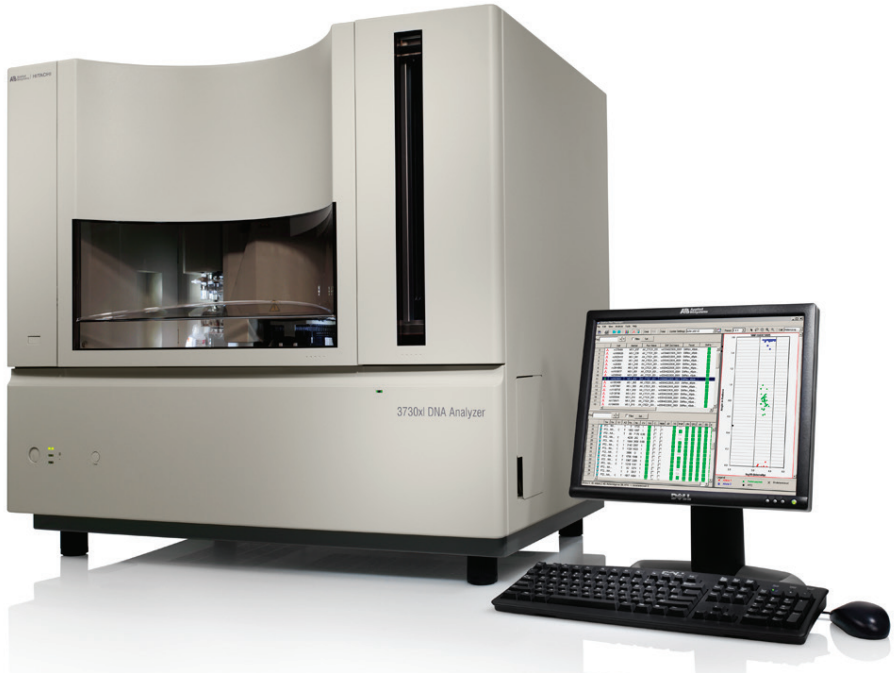


Sanger et al., PNAS, 1977

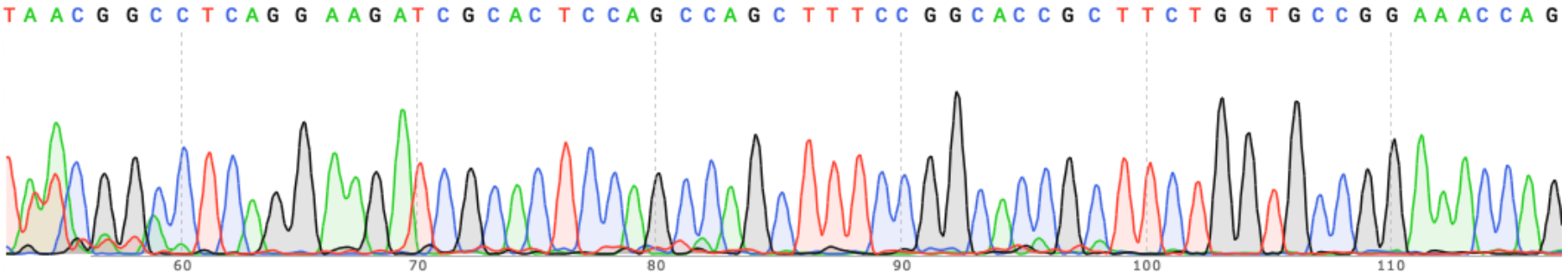
https://www.nobelprize.org/prizes/chemistry/1980/summary/

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.

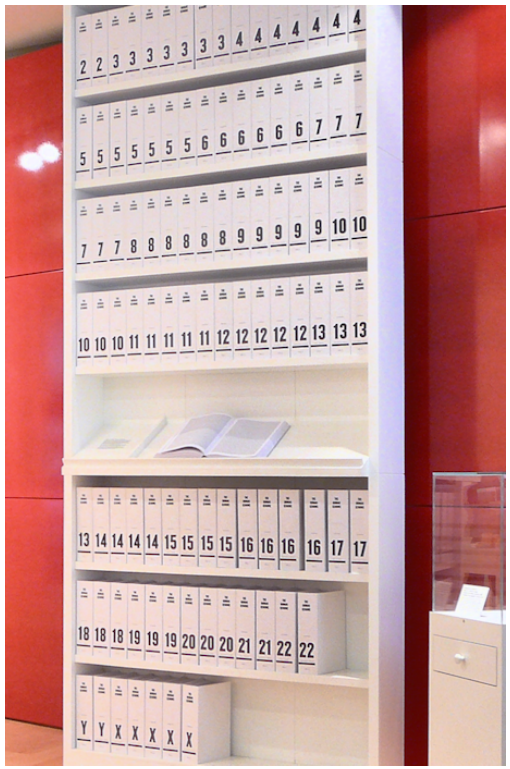
Human genome project (finished in 2003):

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

https://en.wikipedia.org/wiki/Wellcome_Collection

https://www.nature.com/news/human-genome-project-twenty-five-years-of-big-biology-1.18436

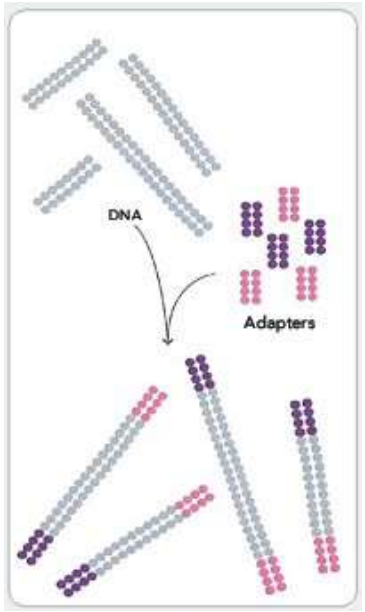
Illumina sequencing was announced in 2006. It has become the standard high-throughput DNA sequencing method

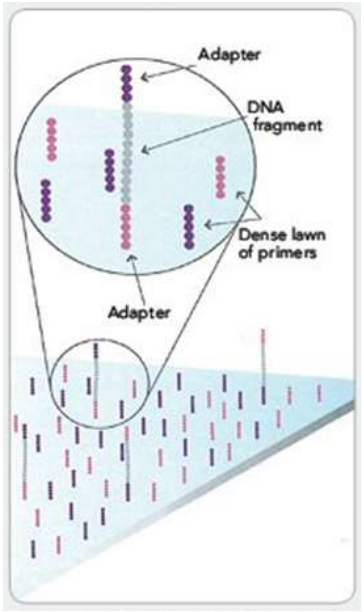
NextSeq 500 sequencing run:
reads: 300,000,000
read length: 300 nt
time: 1 day
cost: $2,000

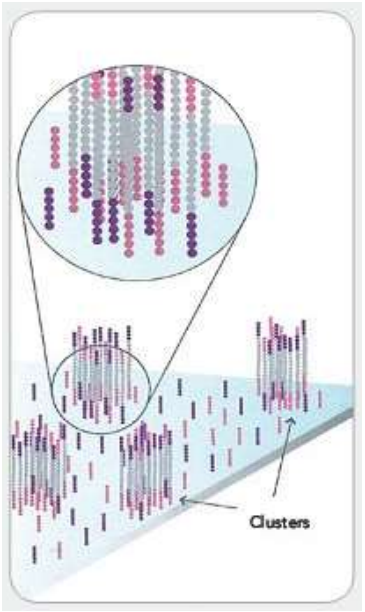
NextSeq 500

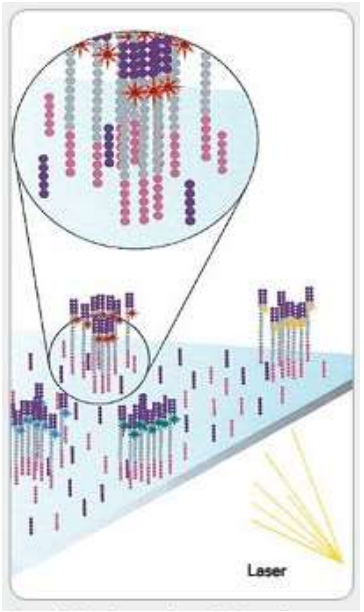
Flow cell

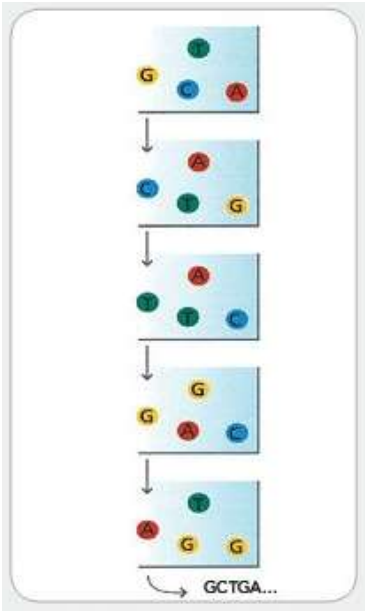
That's ~**30 human genomes** of DNA!



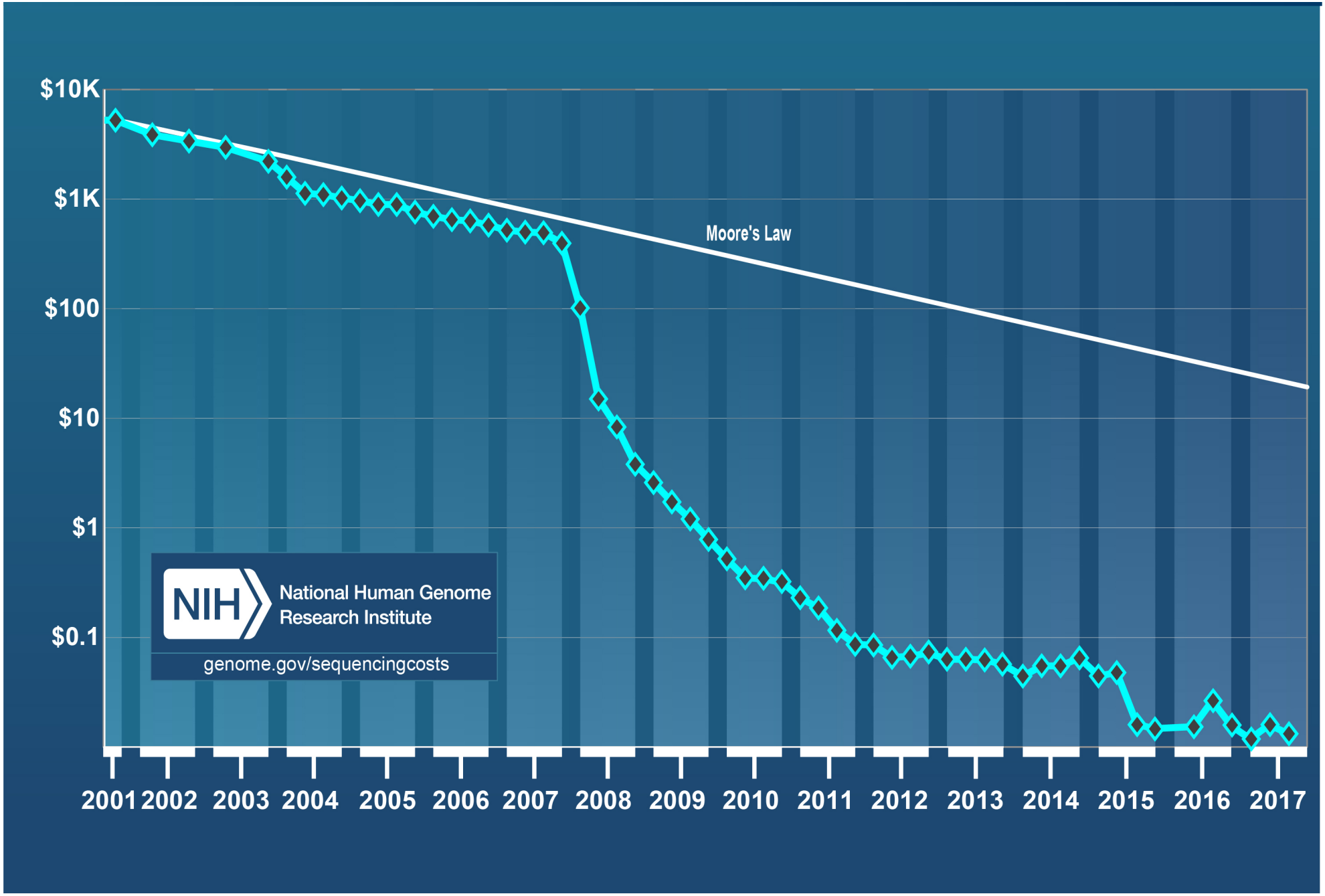






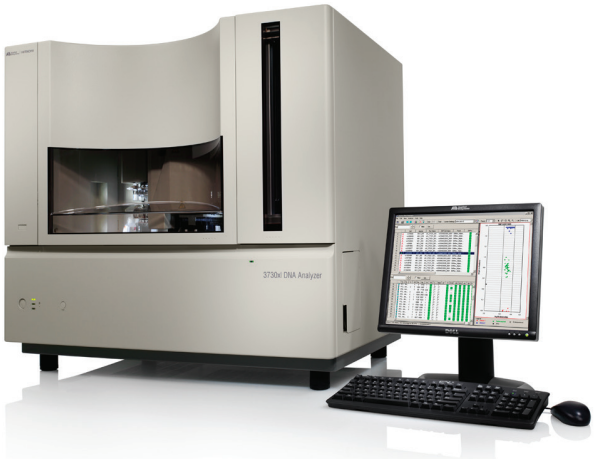


DNA sequencing has undergone an unprecedented technological revolution



https://www.genome.gov/27565109/the-cost-of-sequencing-a-human-genome/

DNA sequencing vs. computing



ABI 3730xl Genetic Analyzer (2007)

5 years

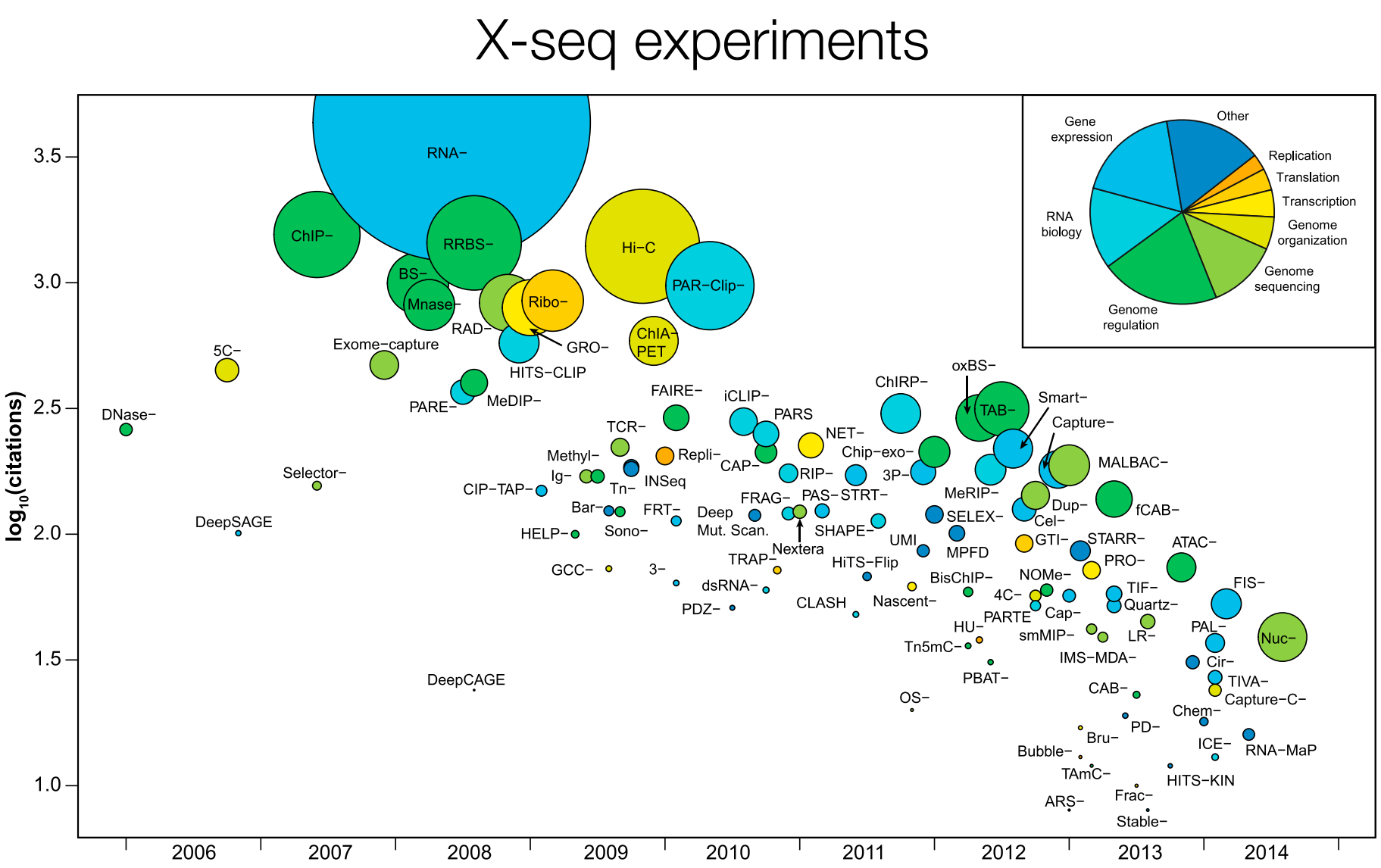
Illumina HiSeq 2500 (2012)

Palm Pilot (1997)

20 years

iPhone X (2017)

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



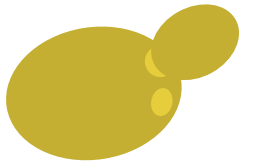
Reuter et al., 2015, Mol. Cell

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

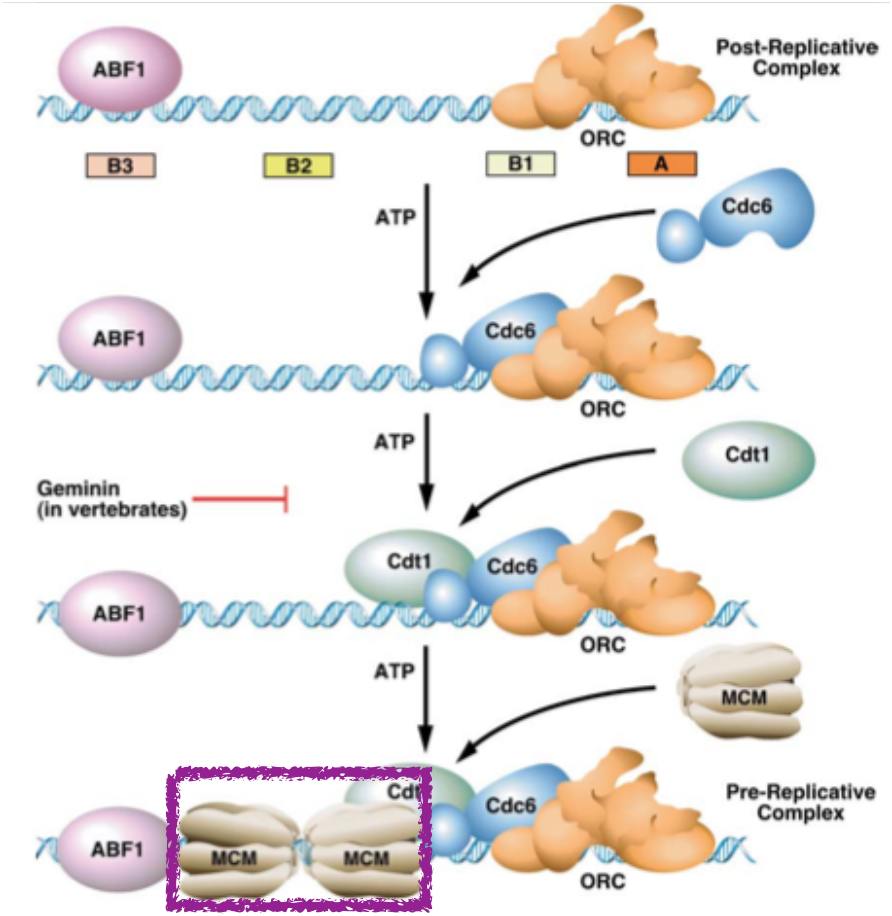
Domain within the helicase subunit Mcm4 integrates multiple kinase signals to control DNA replication initiation and fork progression

Yi-Jun Sheu[a], Justin B. Kinney[a], Armelle Lengronne[b], Philippe Pasero[b], and Bruce Stillman[a,1]

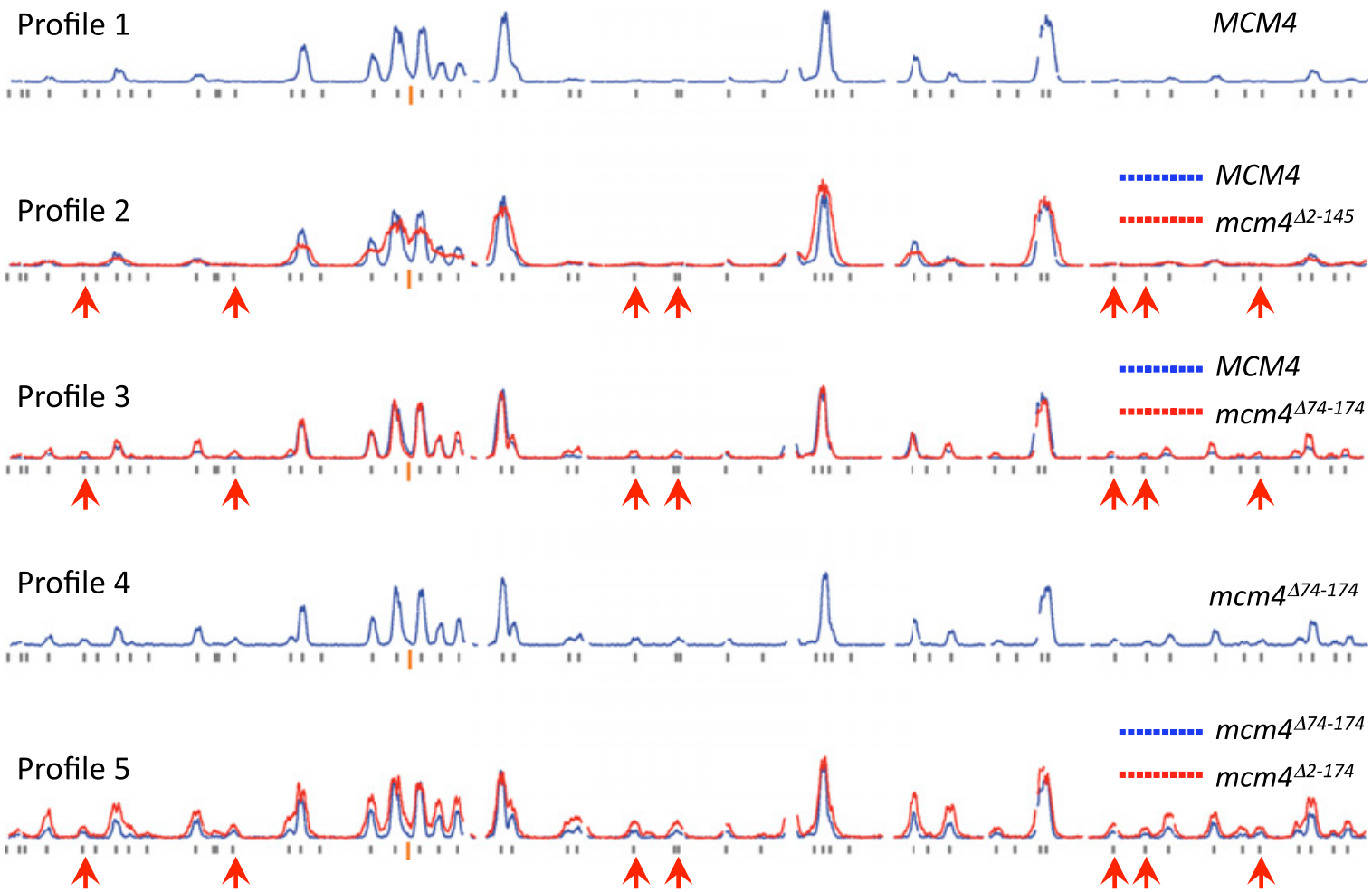
PNAS | Published online April 16, 2014 | E1899–E1908



S. cerevisiae



Here are some examples of the published replication profiles



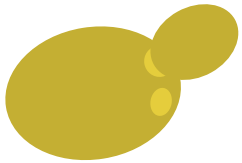
Sheu et al., 2014, PNAS

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

ARS: autonomously replicating sequence

old ssDNA

new ssDNA



S. cerevisiae

G1 phase

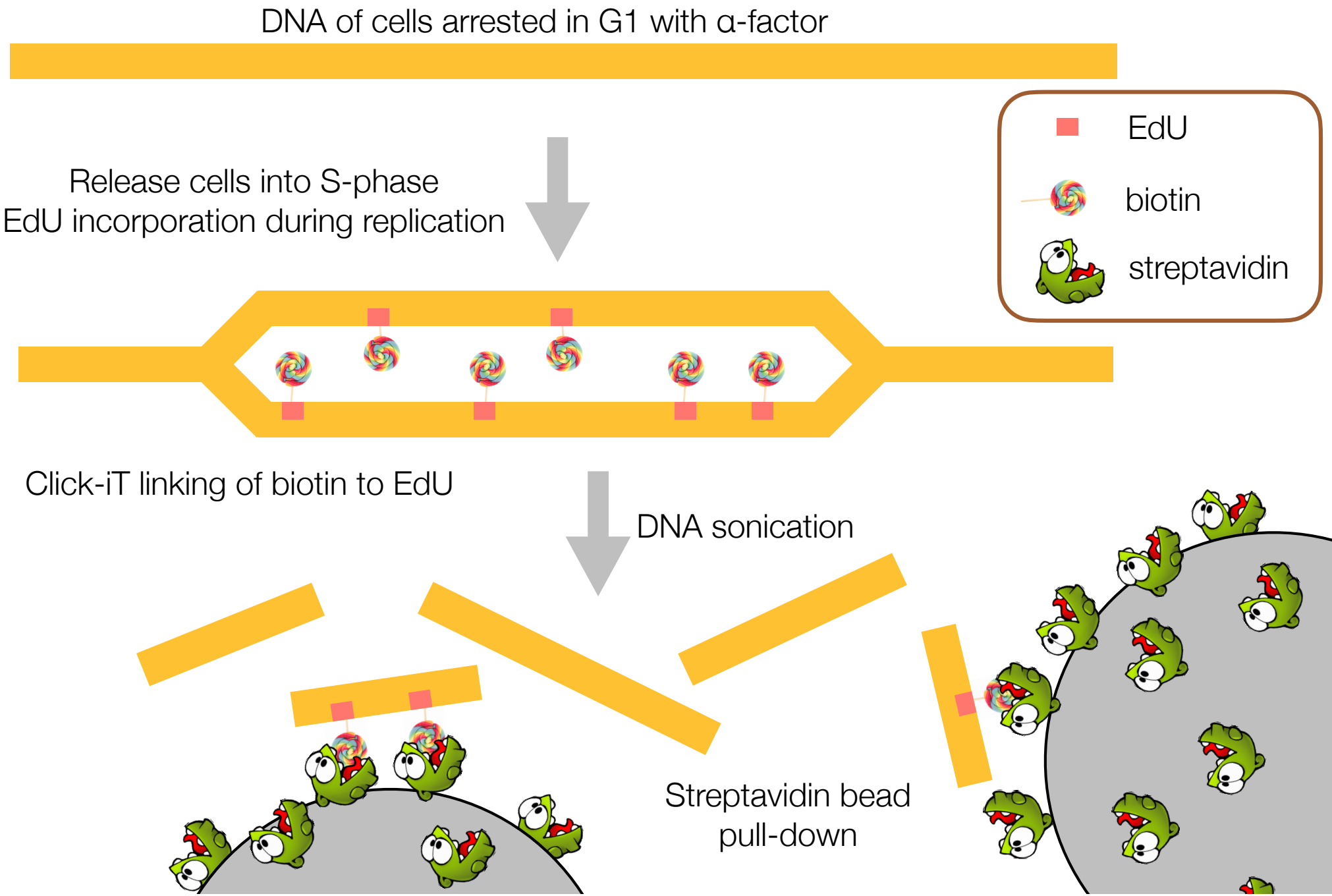
pre-RC

S phase

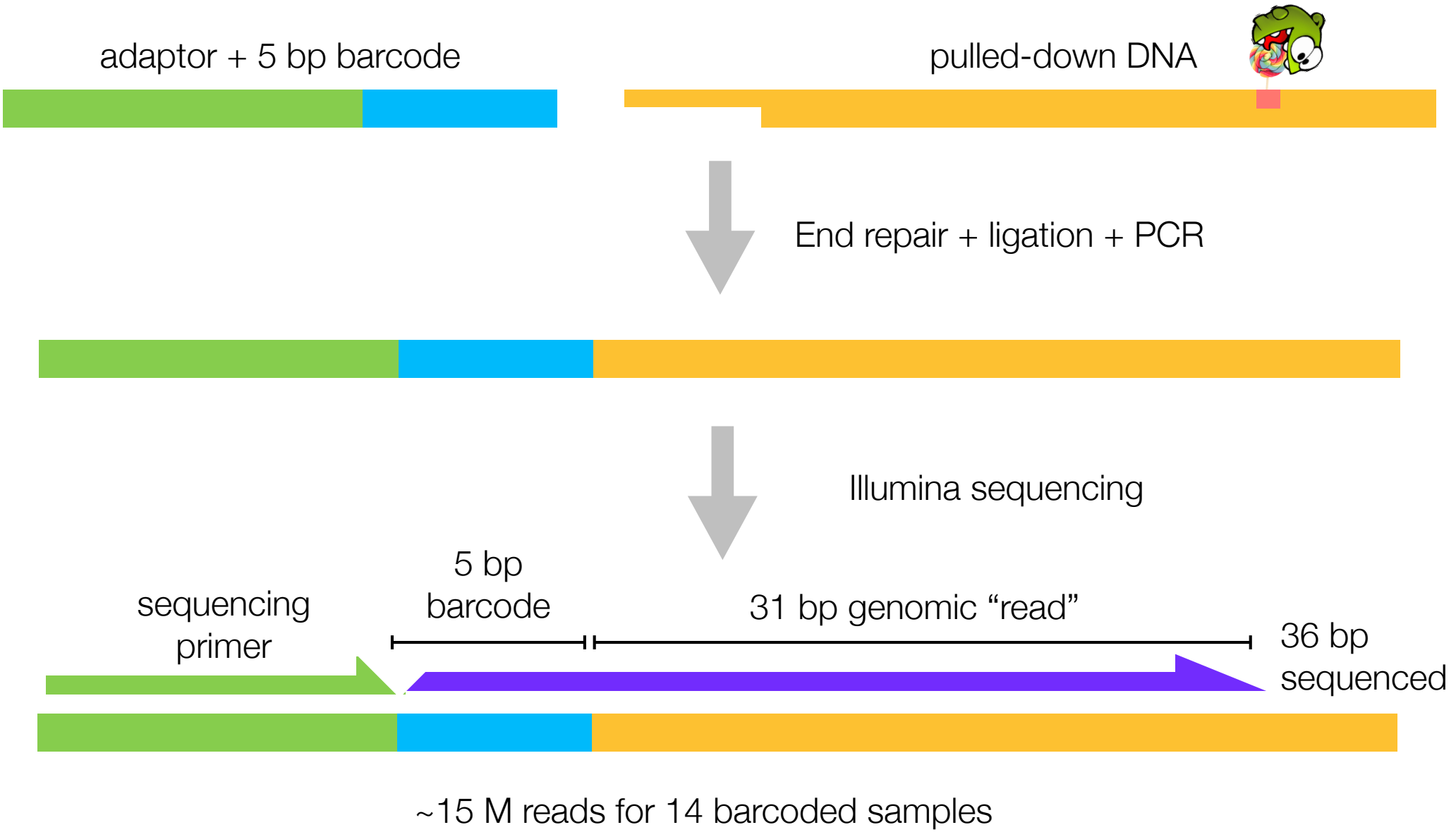
Fork

Fork

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



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

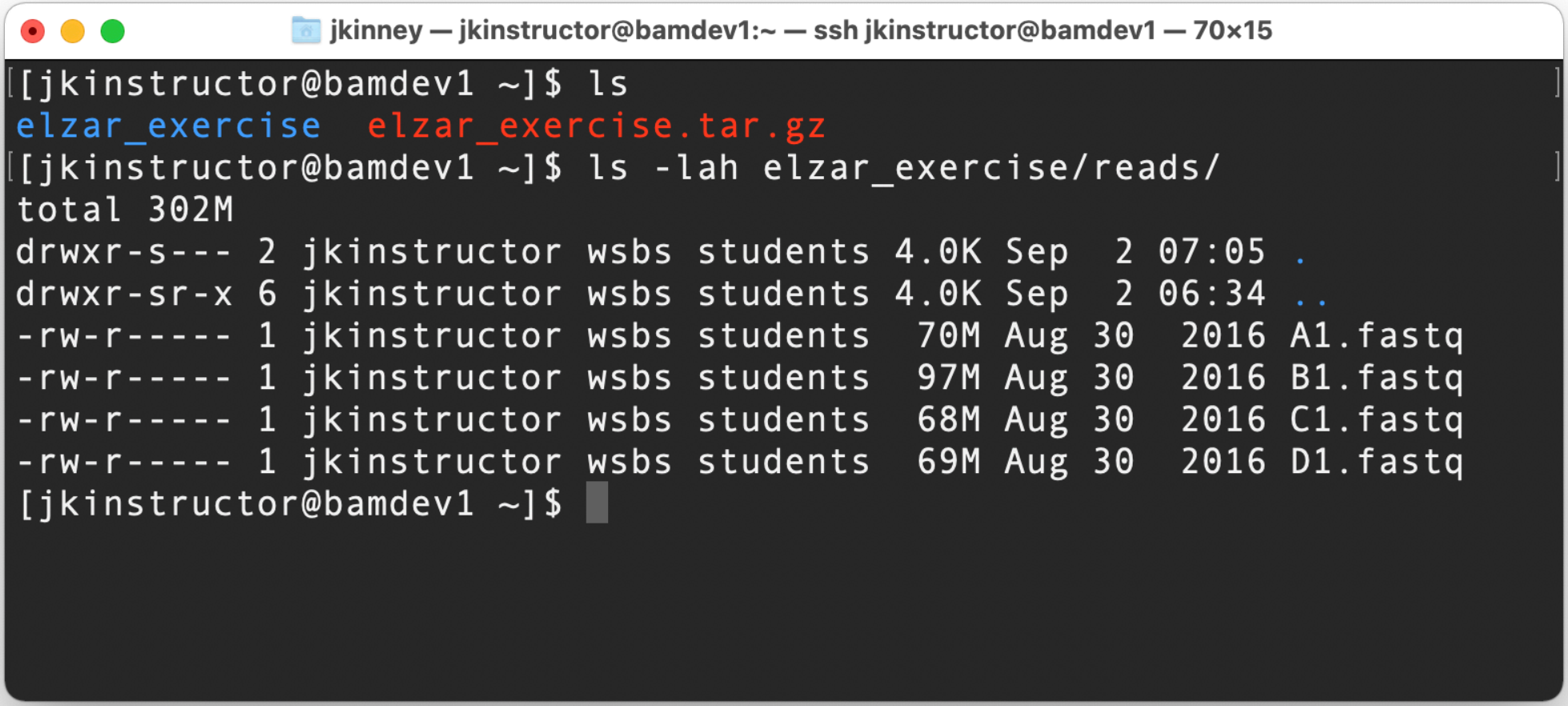


~15 M reads for 14 barcoded samples

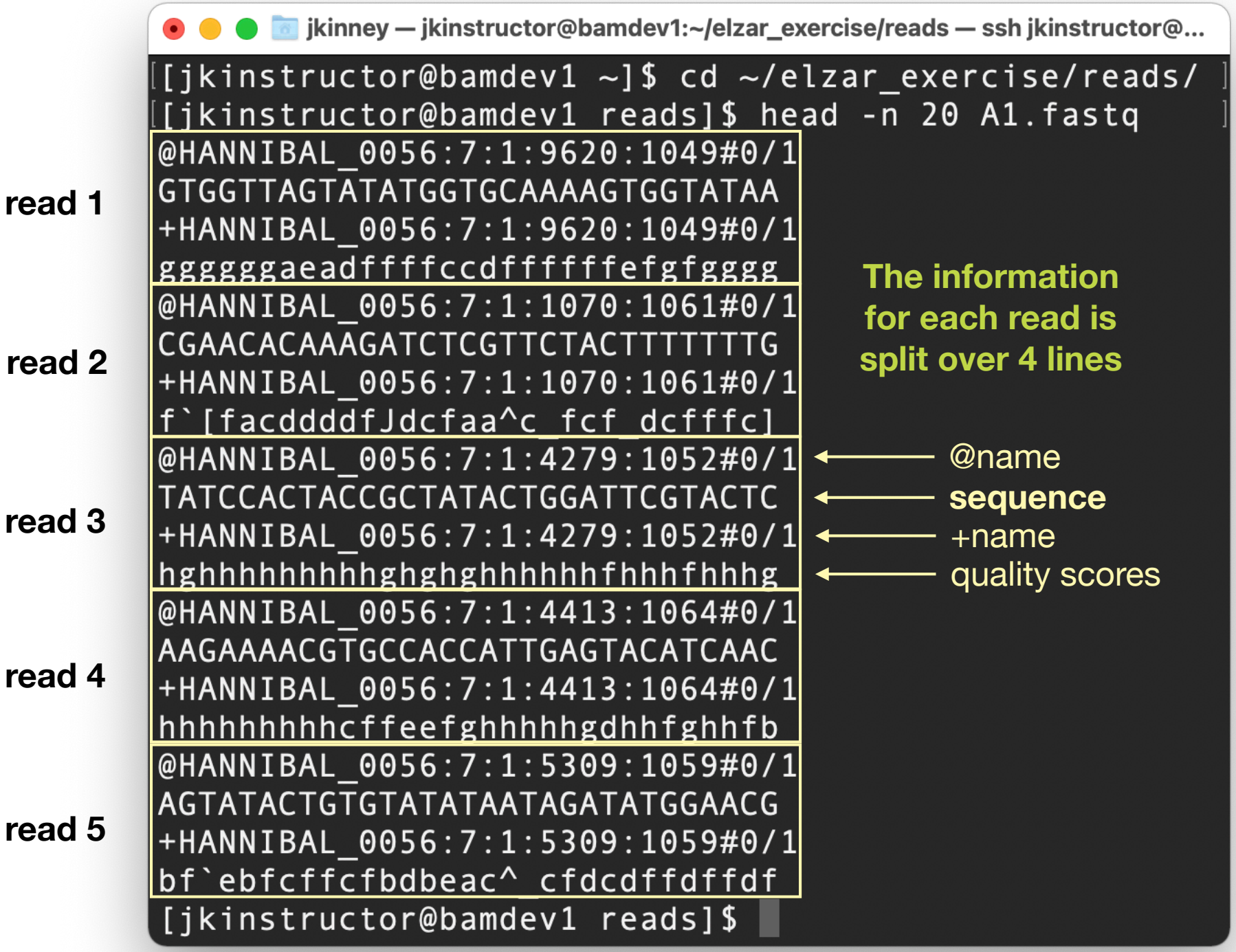
We will analyze FASTQ read files from 4 different samples

There are four FASTQ files in the reads/ directory

Each FASTQ files is ~60-100 megabytes



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



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]
CCACACCACACCCACACACCCACACACCACACCACACACCACACCACACCCACACACACA
CATCCTAACACTACCCTAACACAGCCCTAATCTAACCCTGGCCAACCTGTCTCTCAACTT
ACCCTCCATTACCCTGCCTCCACTCGTTACCCTGTCCCATTCAACCATACCACTCCGAAC
CACCATCCATCCCTCTACTTACTACCACTCACCCACCGTTACCCTCCAATTACCCATATC
CAACCCACTGCCACTTACCCTACCATTACCCTACCATCCACCATGACCTACTCACCATAC
TGTTCTTCTACCCACCATATTGAAACGCTAACAAATGATCGTAAATAACACACACGTGCT
TACCCTACCACTTTATACCACCACCACATGCCATACTCACCCTCACTTGTATACTGATTT
TACGTACGCACACGGATGCTACAGTATATACCATCTCAAACTTACCCTACTCTCAGATTC
CACTTCACTCCATGGCCCATCTCTCACTGAATCAGTACCAAATGCACTCACATCATTATG
[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

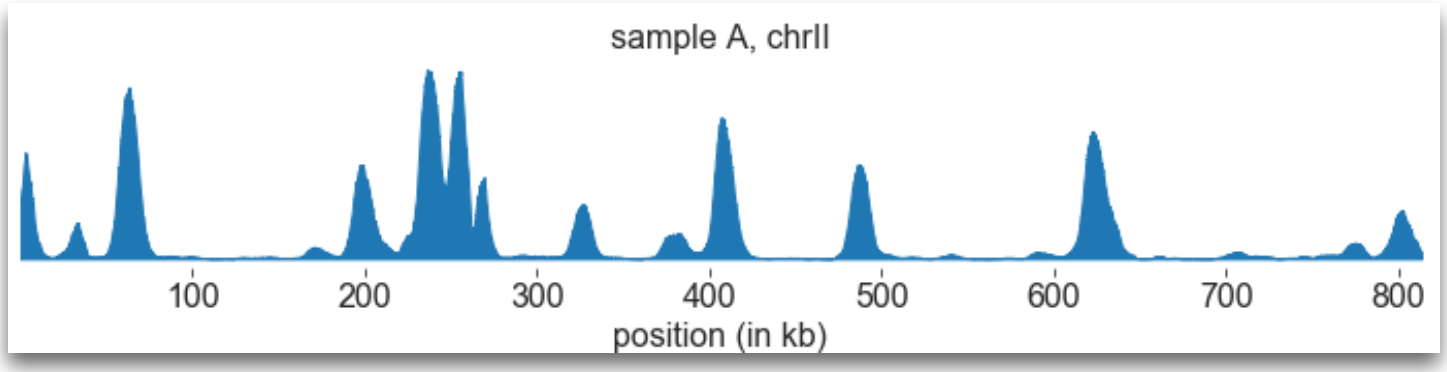
cluster analysis
(bwa + samtools + bedtools)

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
```

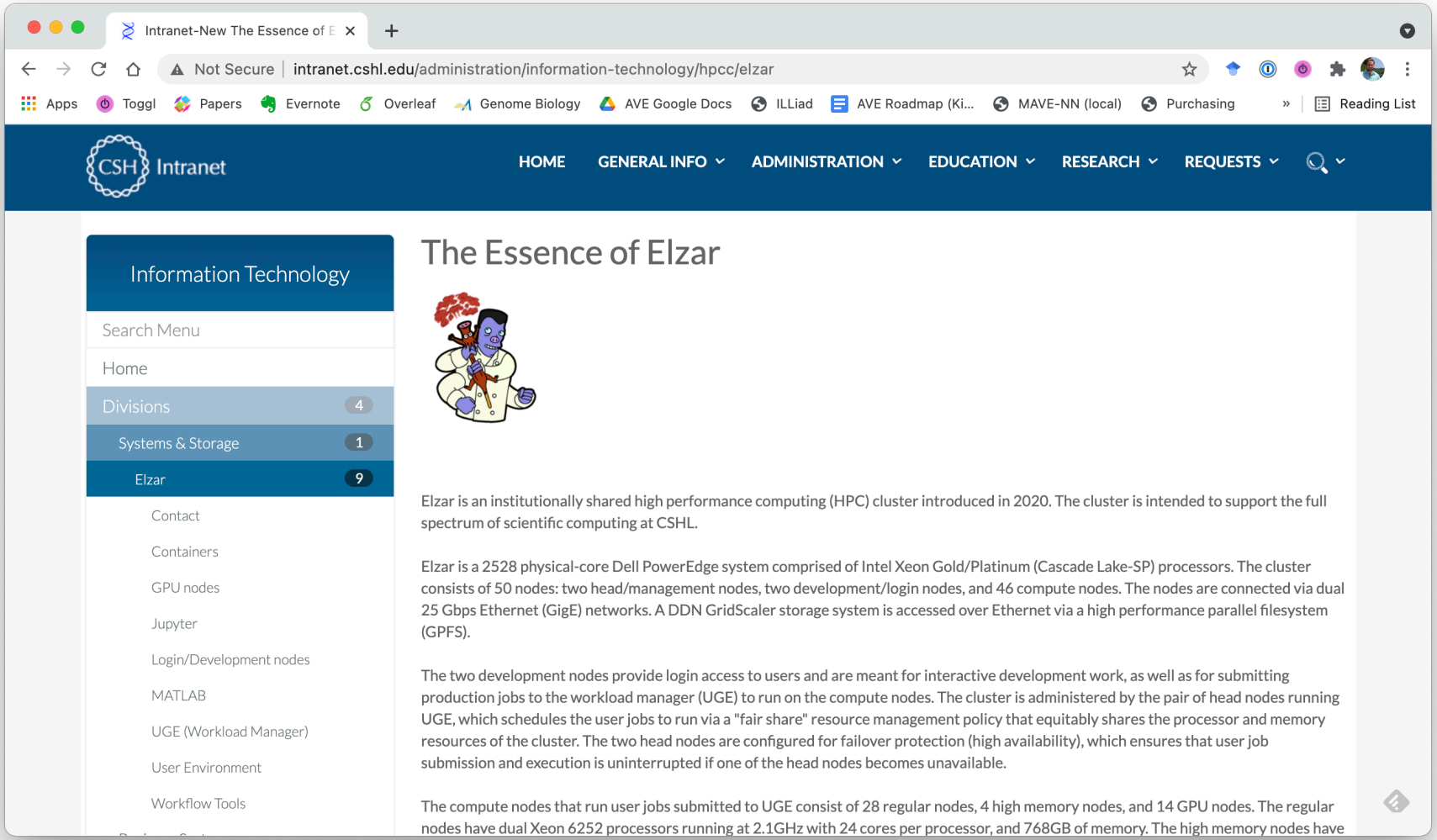
chromosome **window** **# reads**

local analysis
(python)



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

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

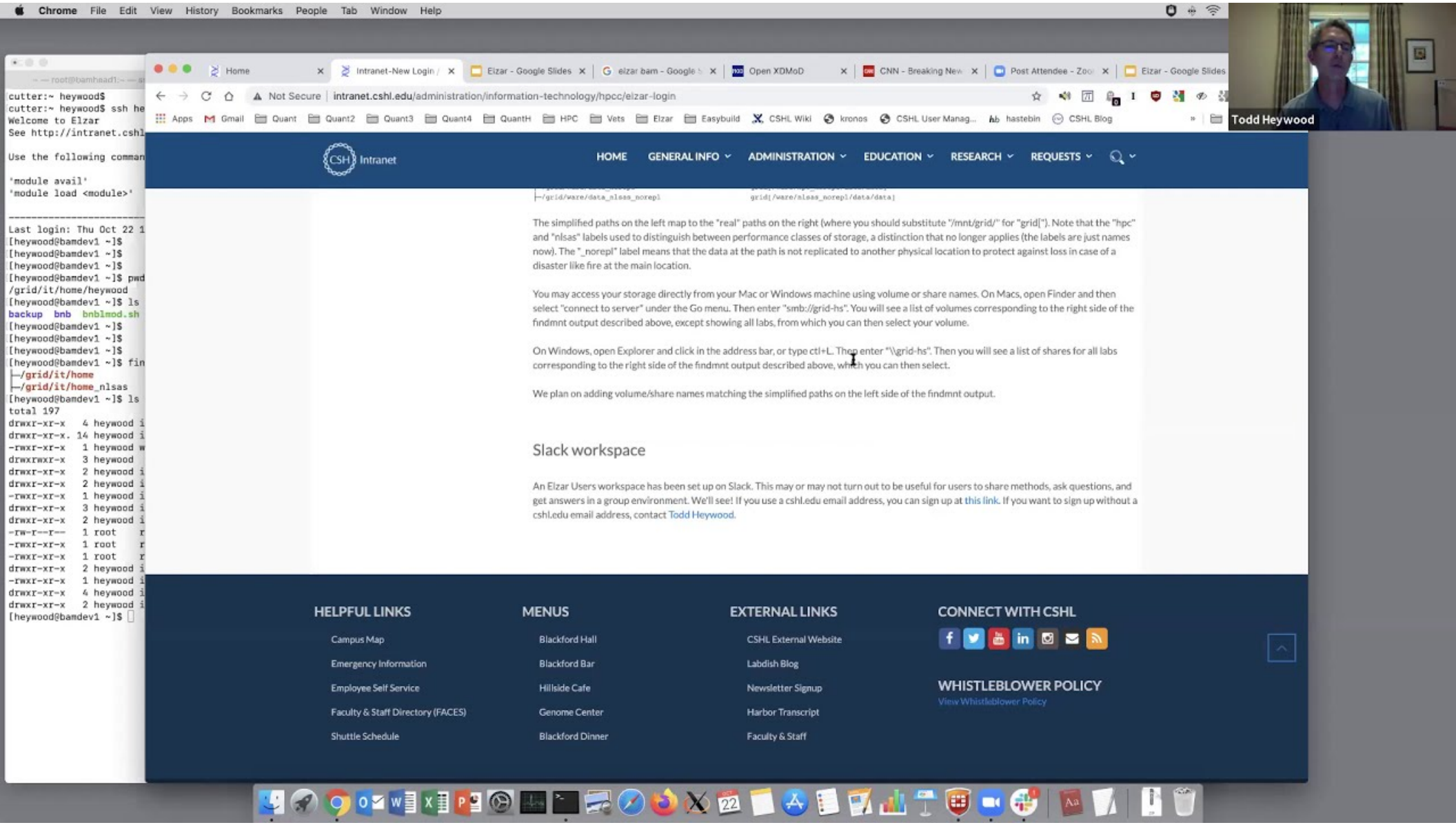


Todd Heywood

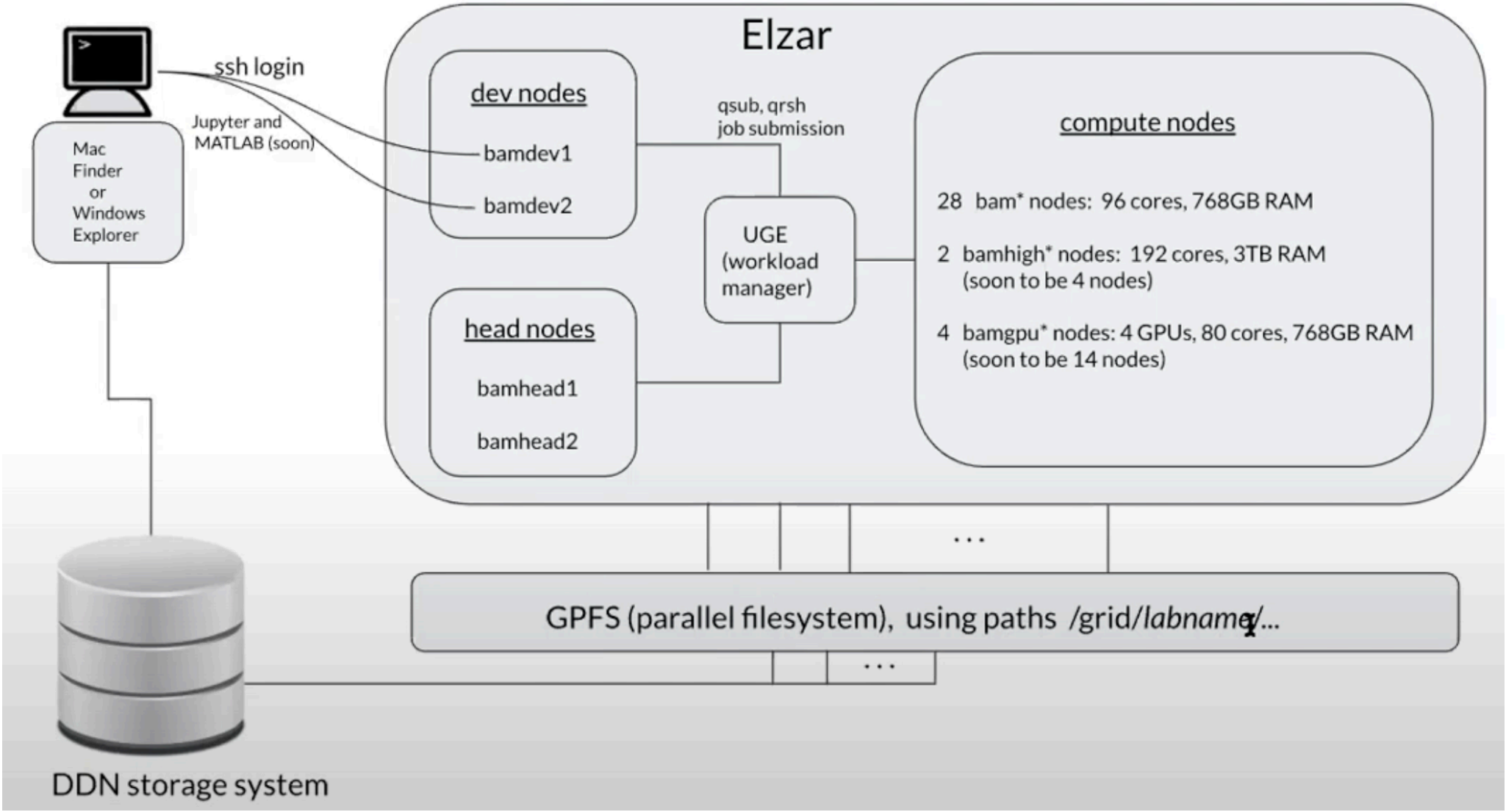
If you use Elzar ***at all*** for your work, please cite "NIH Grant S10OD028632-01"

Elzar tutorial on YouTube

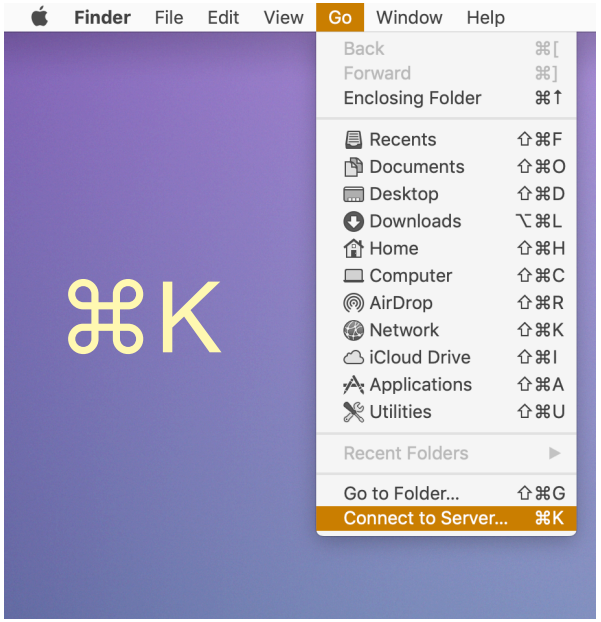
https://www.youtube.com/embed/D3wfhM_cQPY



Architecture of Elzar

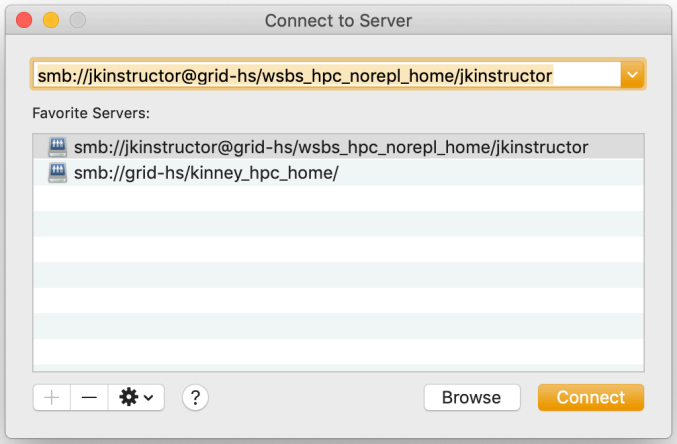


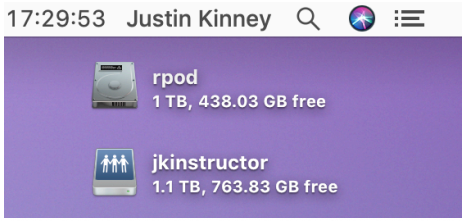
The Elzar disk can be mounted using smb

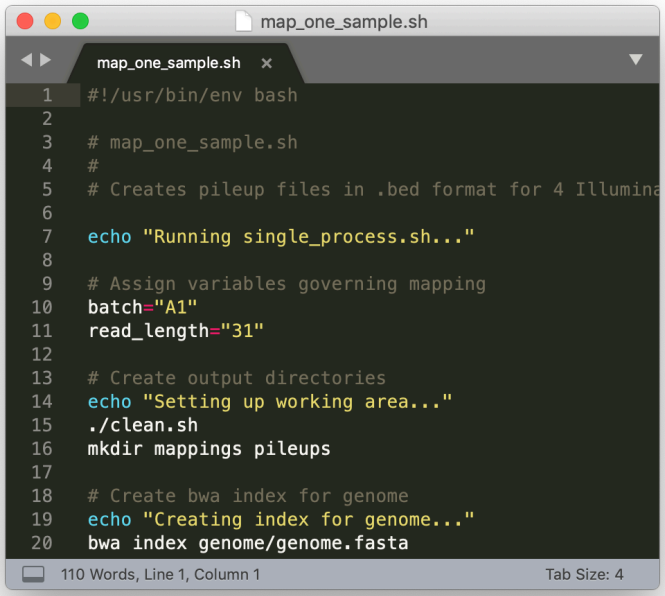


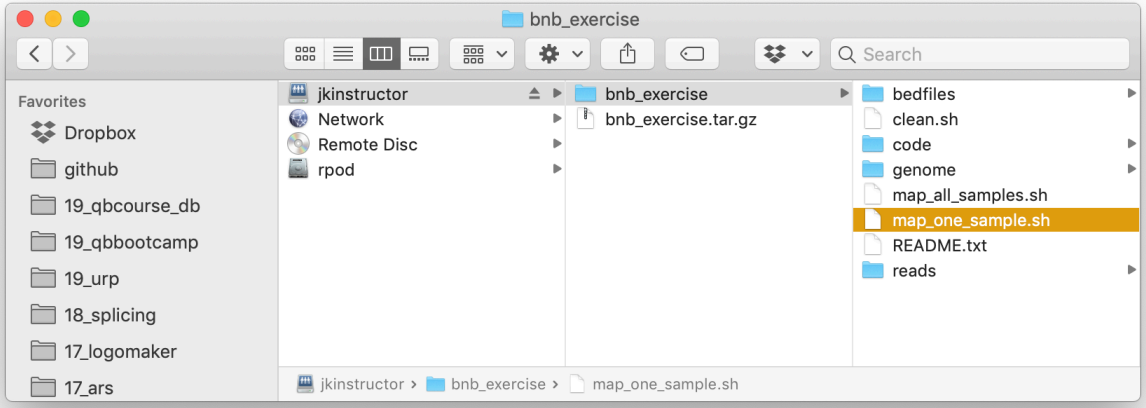
smb://you@grid-hs/wsbs_hpc_norepl_home/you

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









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.