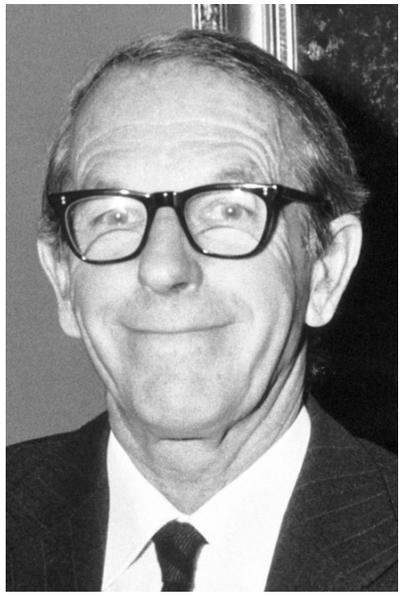
NextGen Sequencing and High-Performance Computing

QB Bootcamp, Day 2
Thursday, 10 September 2020
10:00am - 10:30am

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

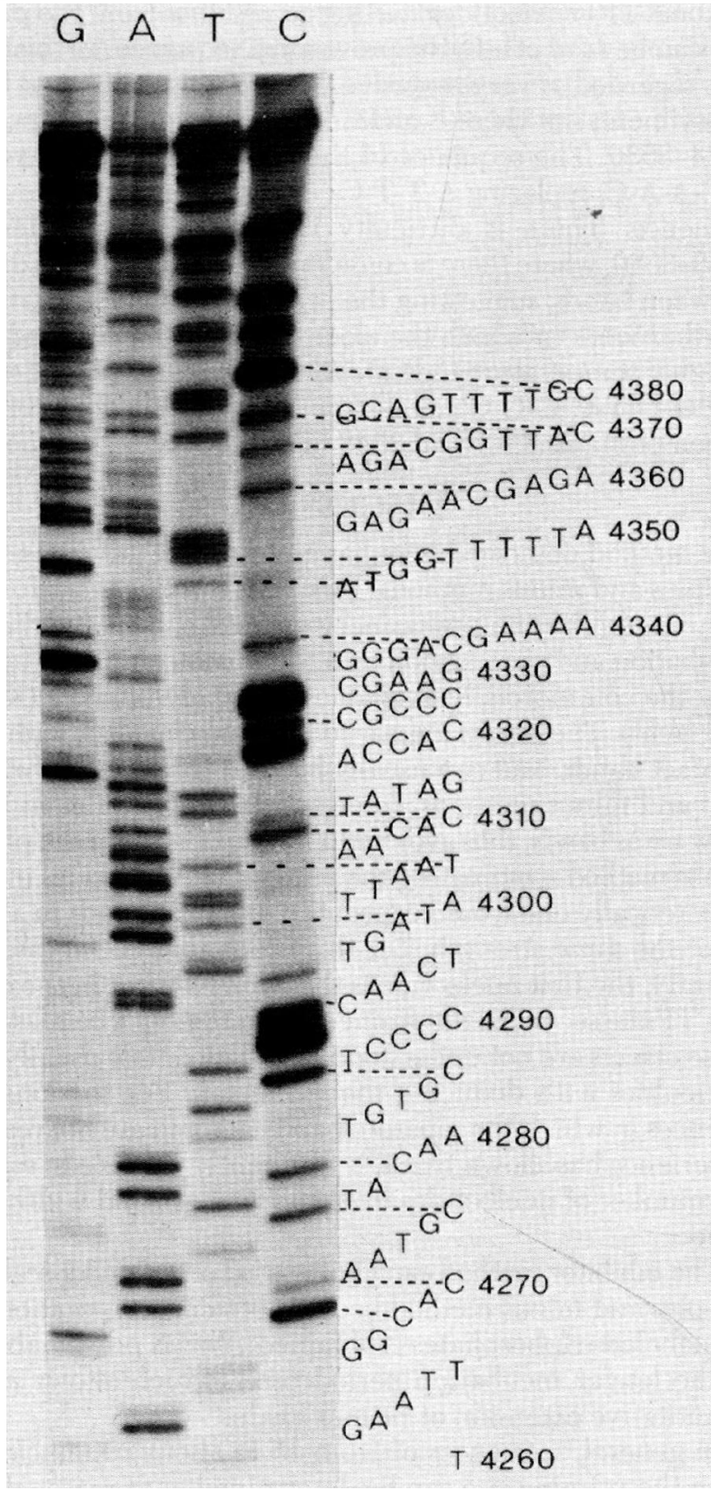


Fred Sanger

Wally Gilbert

1980 Nobel Prize in Chemistry

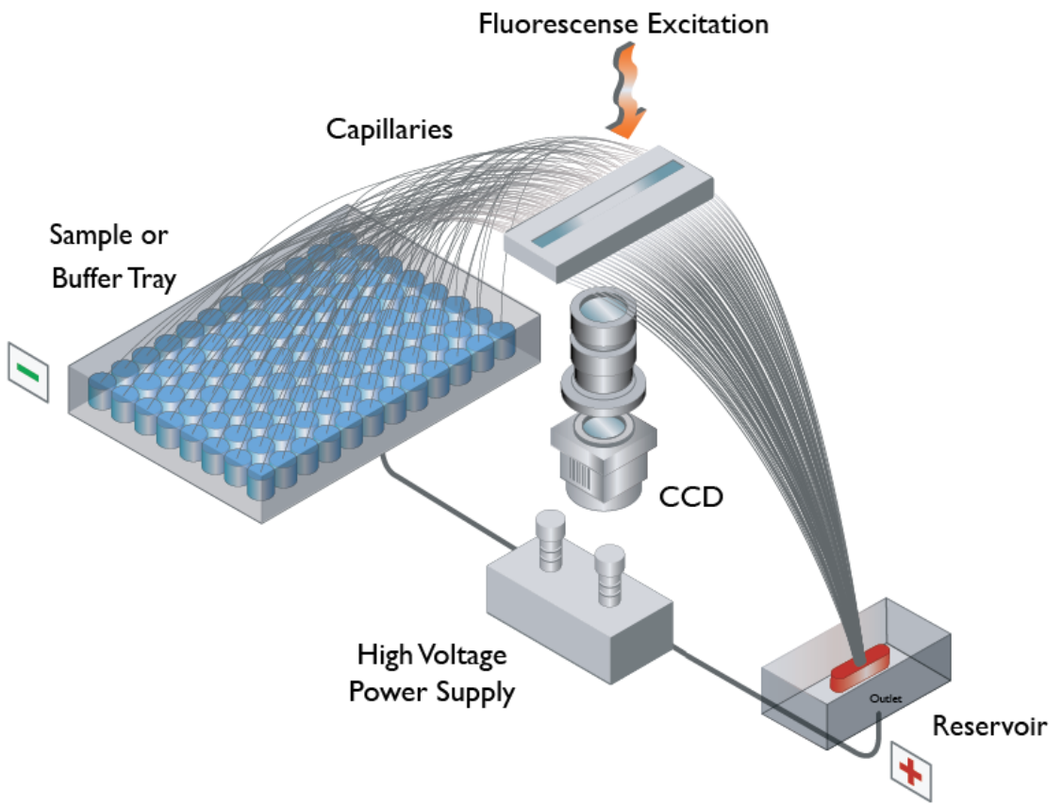
"for their contributions concerning the determination of base sequences in nucleic acids."



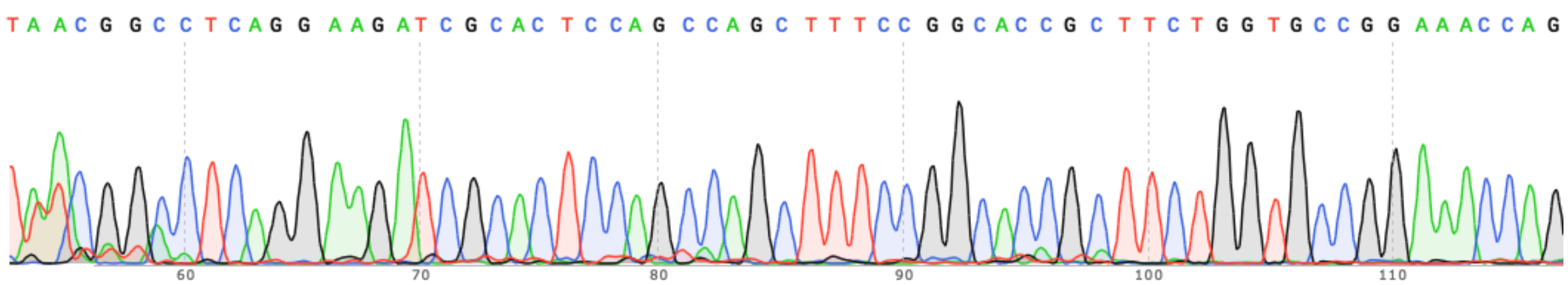
Sanger et al., PNAS, 1977

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)

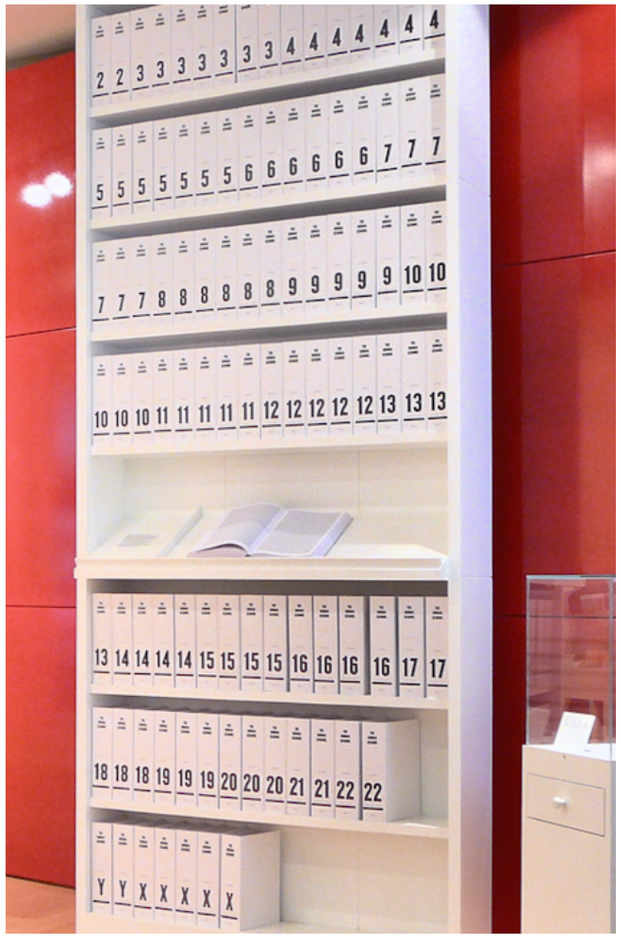


The human genome was sequenced using Sanger sequencing.

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

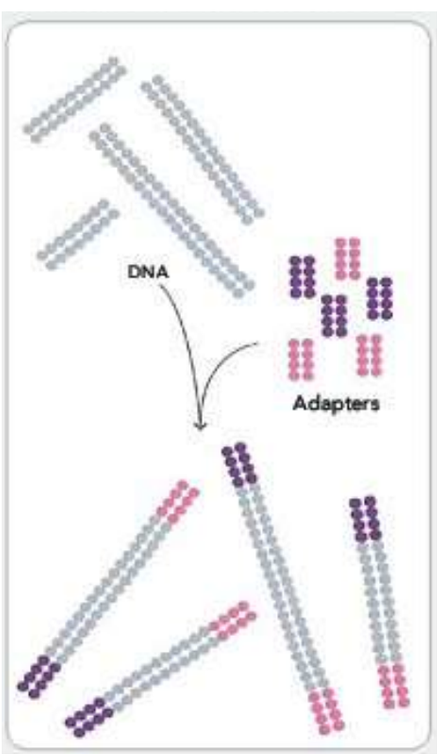
NextSeq 500 sequencing run:
reads: 300,000,000
read length: 300 nt
time: 1 day
cost: $2,000

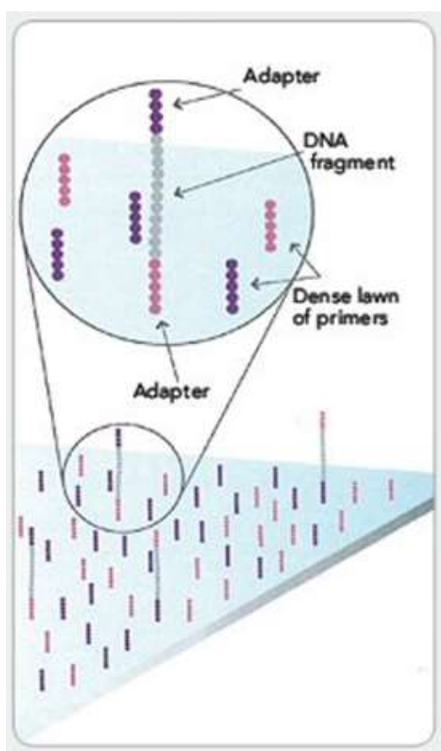
NextSeq 500

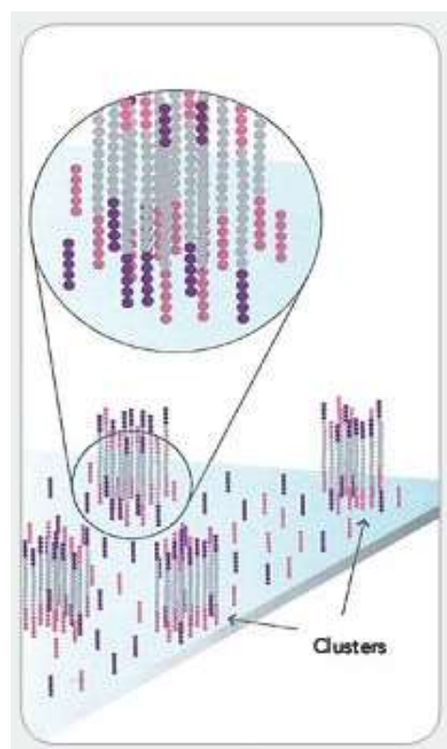
this is where the DNA goes

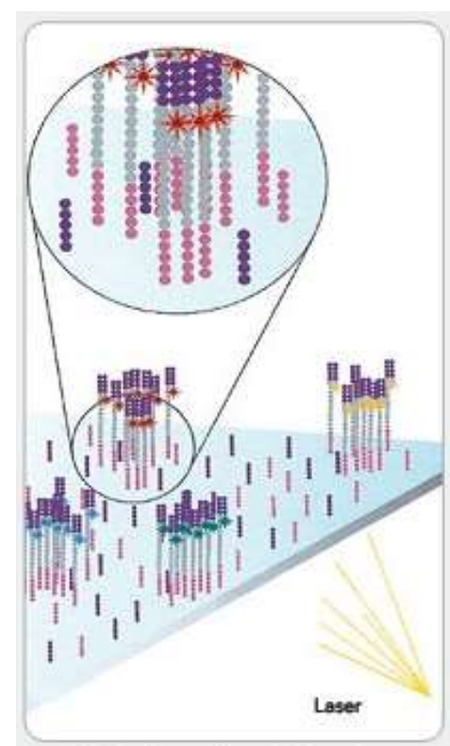
Flow cell

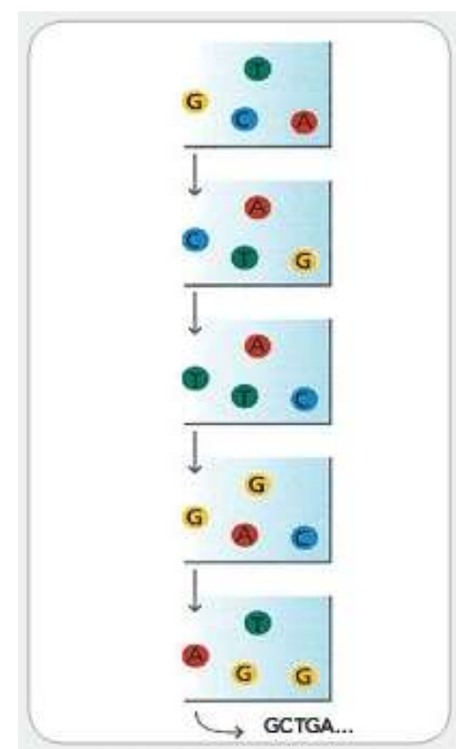
That's ~**30 human genomes** of DNA!



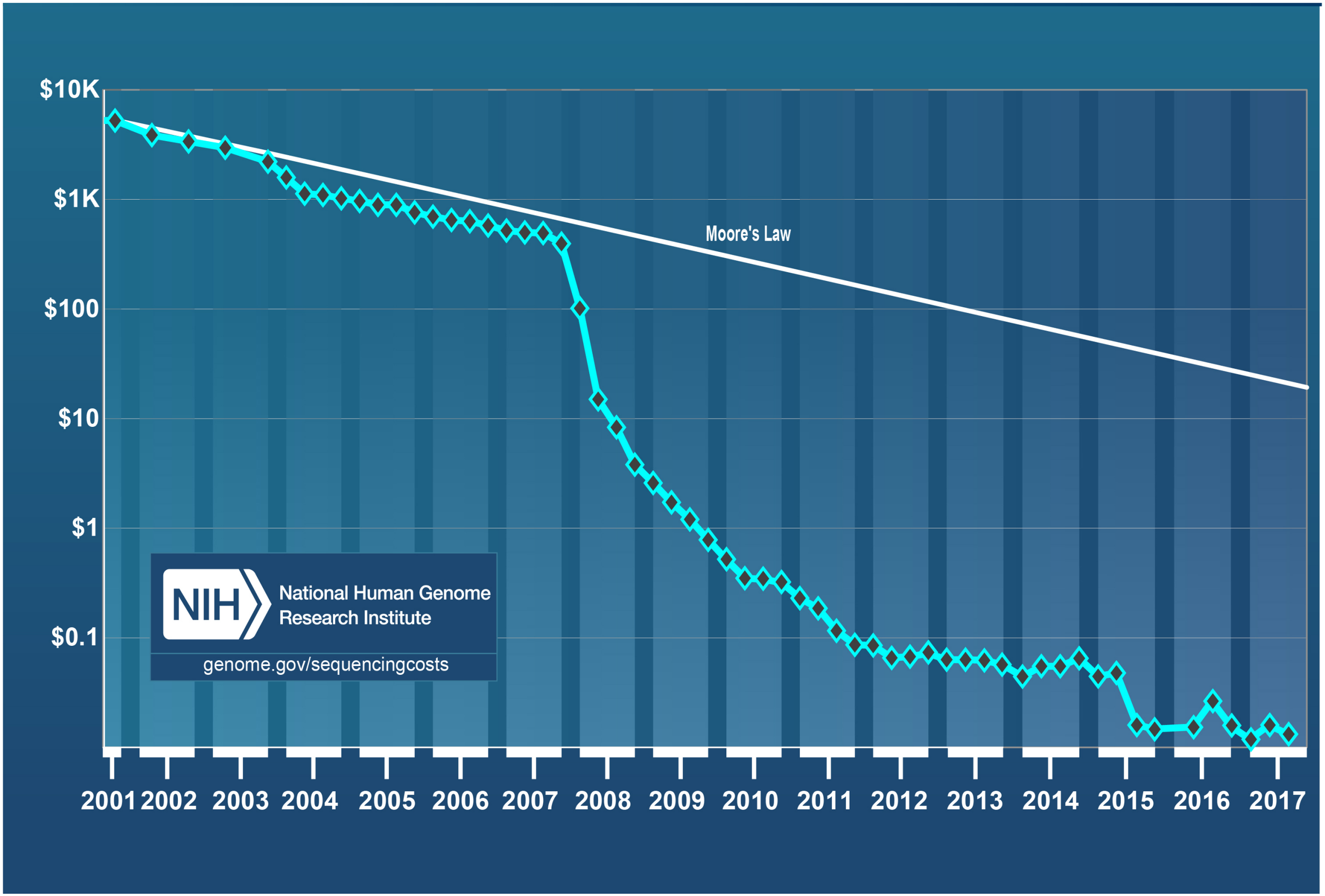








DNA sequencing has undergone an unprecedented technological revolution



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)

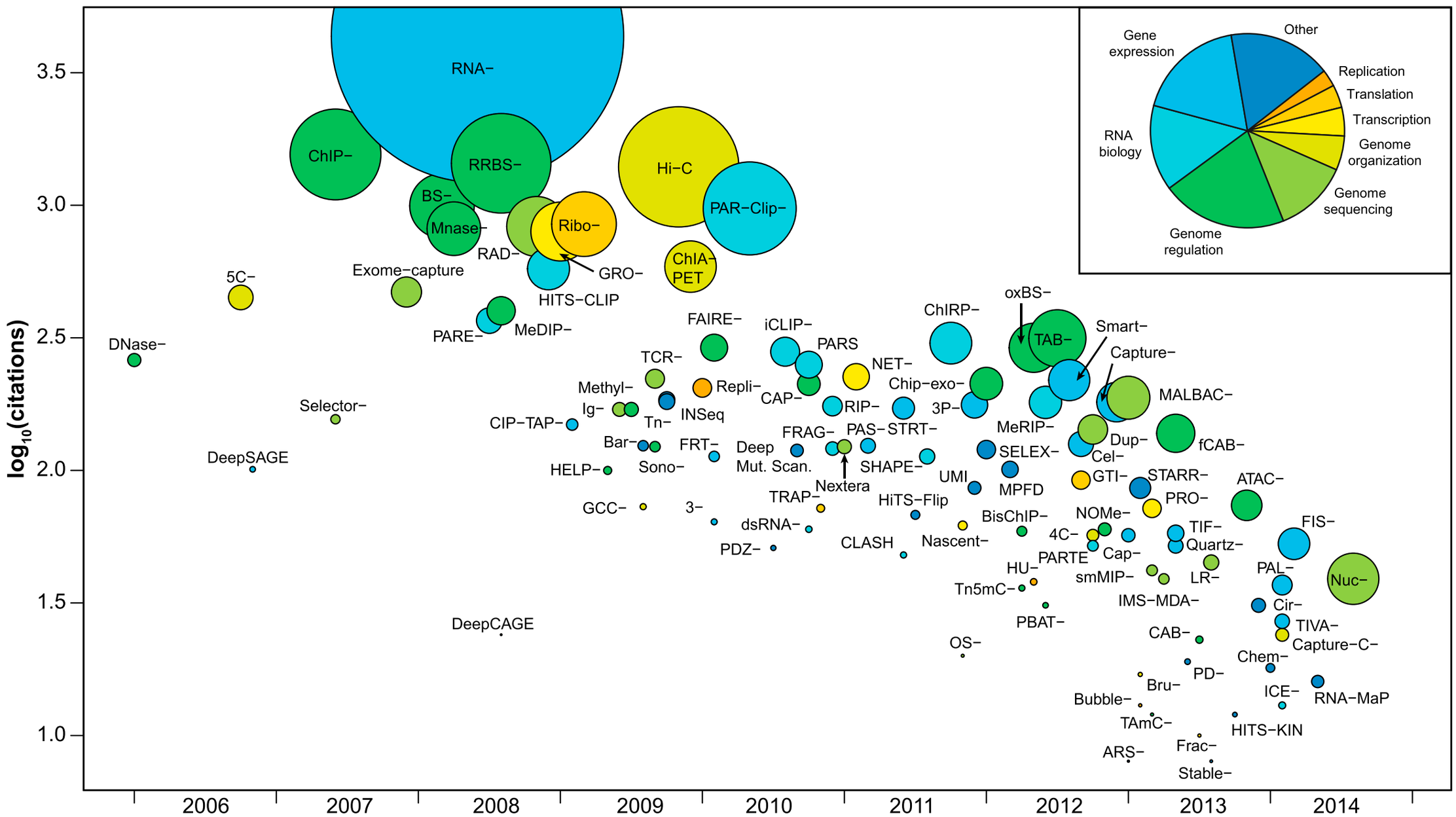
Palm Pilot (1997)

20 years

iPhone X (2017)

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

X-seq experiments



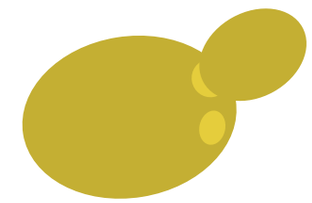
Reuter et al., 2015, Mol. Cell

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

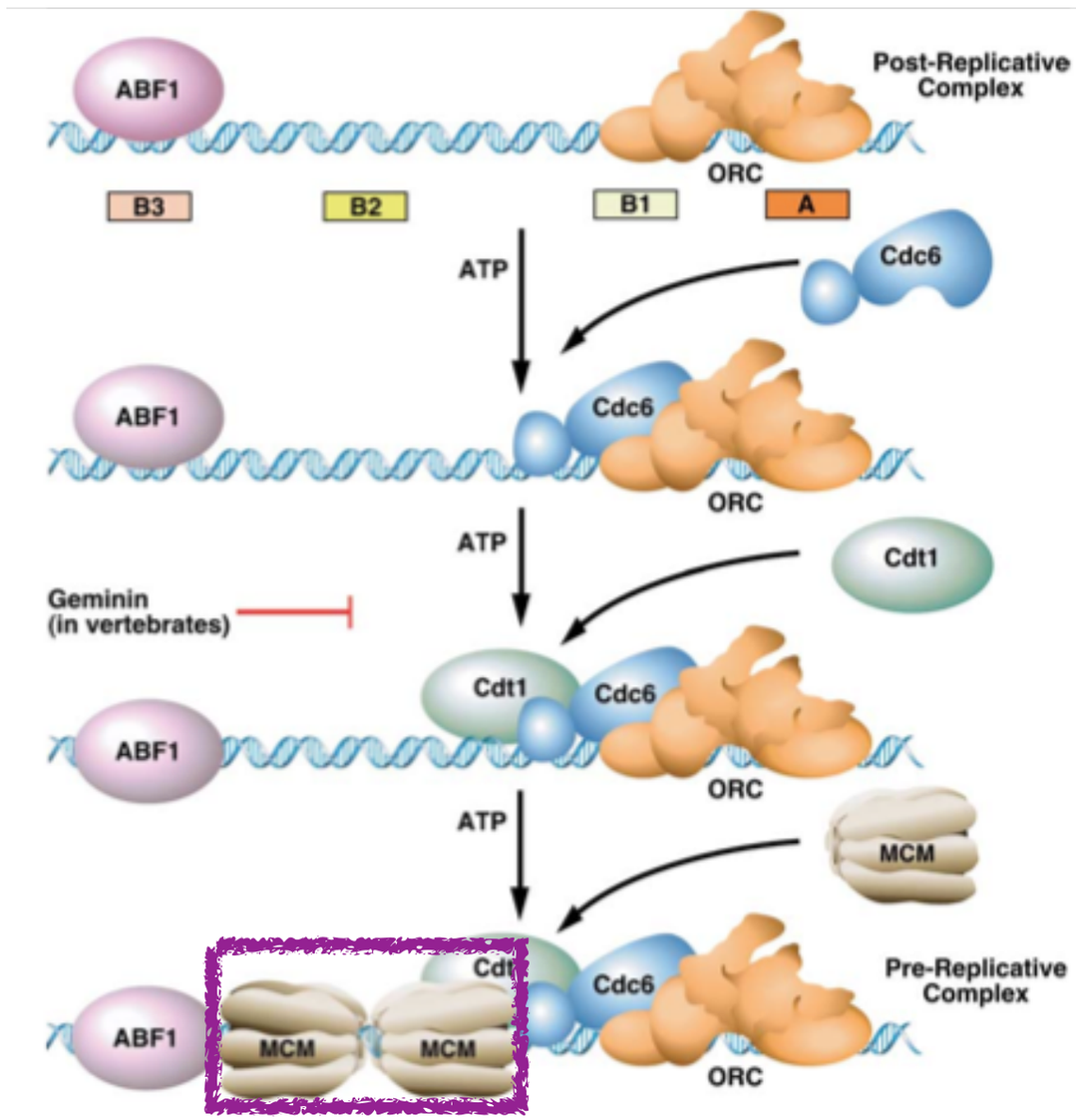
Domain within the helicase subunit Mcm4 integrates multiple kinase signals to control DNA replication initiation and fork progression

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

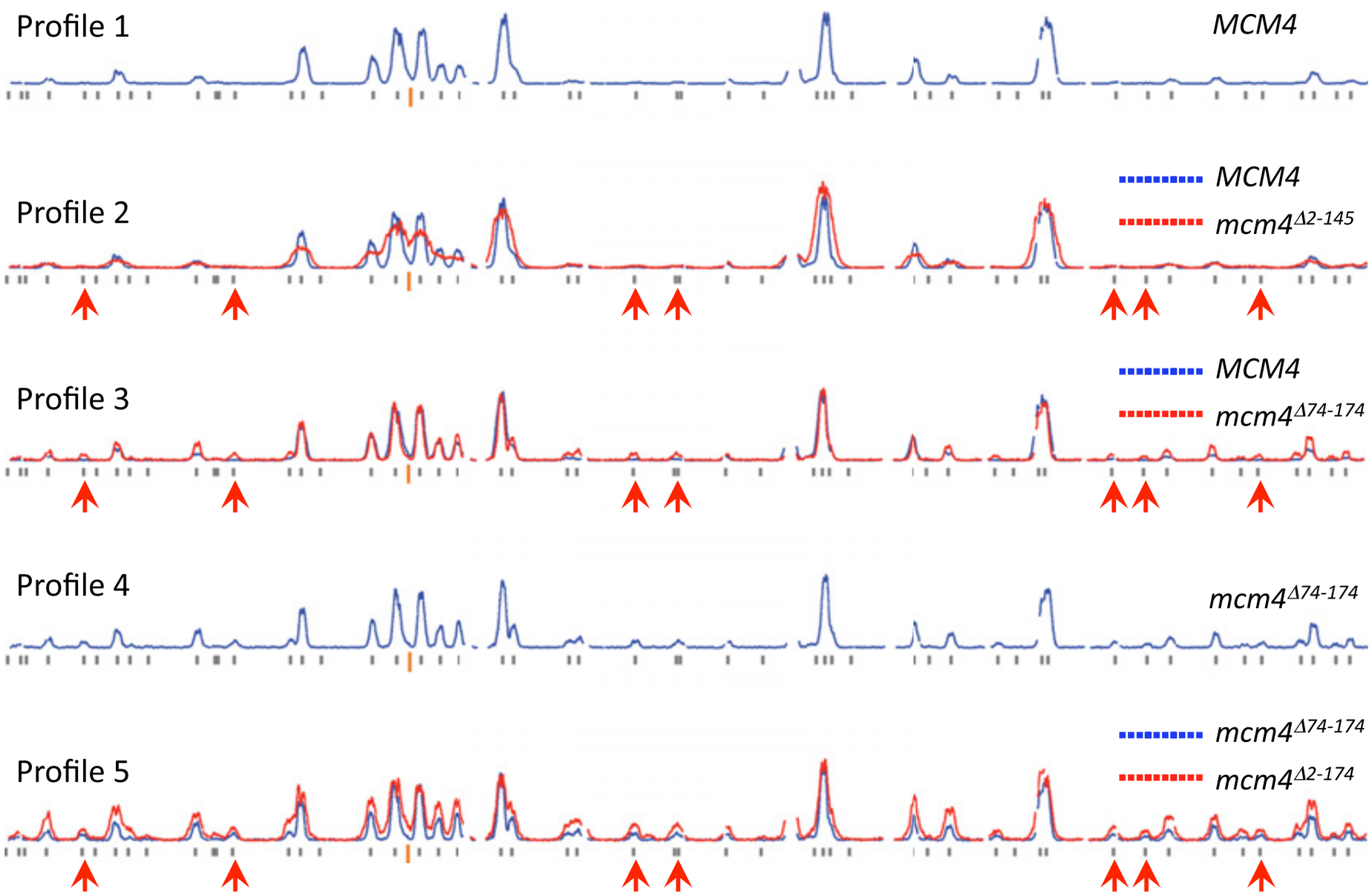
PNAS | Published online April 16, 2014 | E1899–E1908



S. cerevisiae



Here are some examples of the published replication profiles



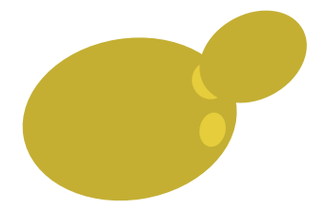
Sheu et al., 2014, PNAS

~300 separate loci direct DNA replication initiation in Saccharomyces cerevisiae

ARS: autonomously replicating sequence

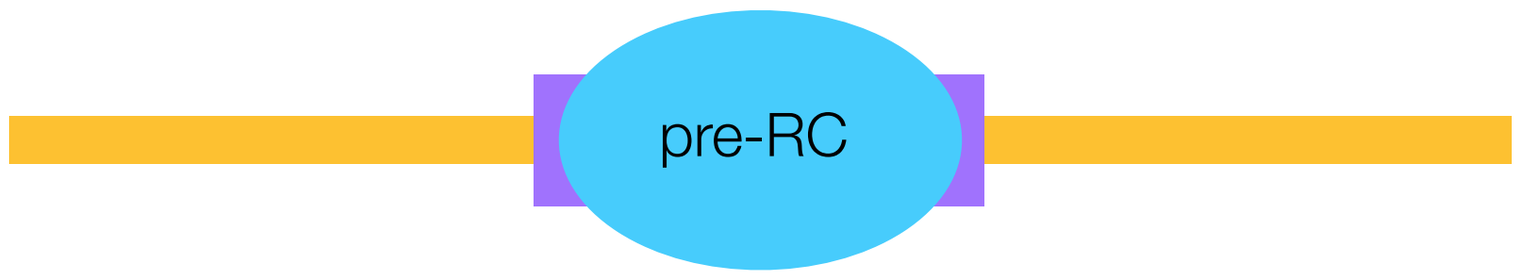
old ssDNA

new ssDNA

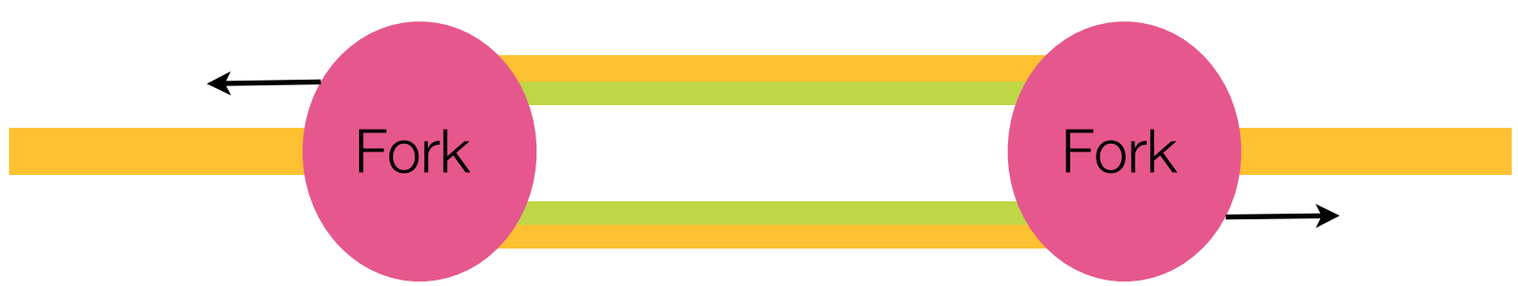


S. cerevisiae

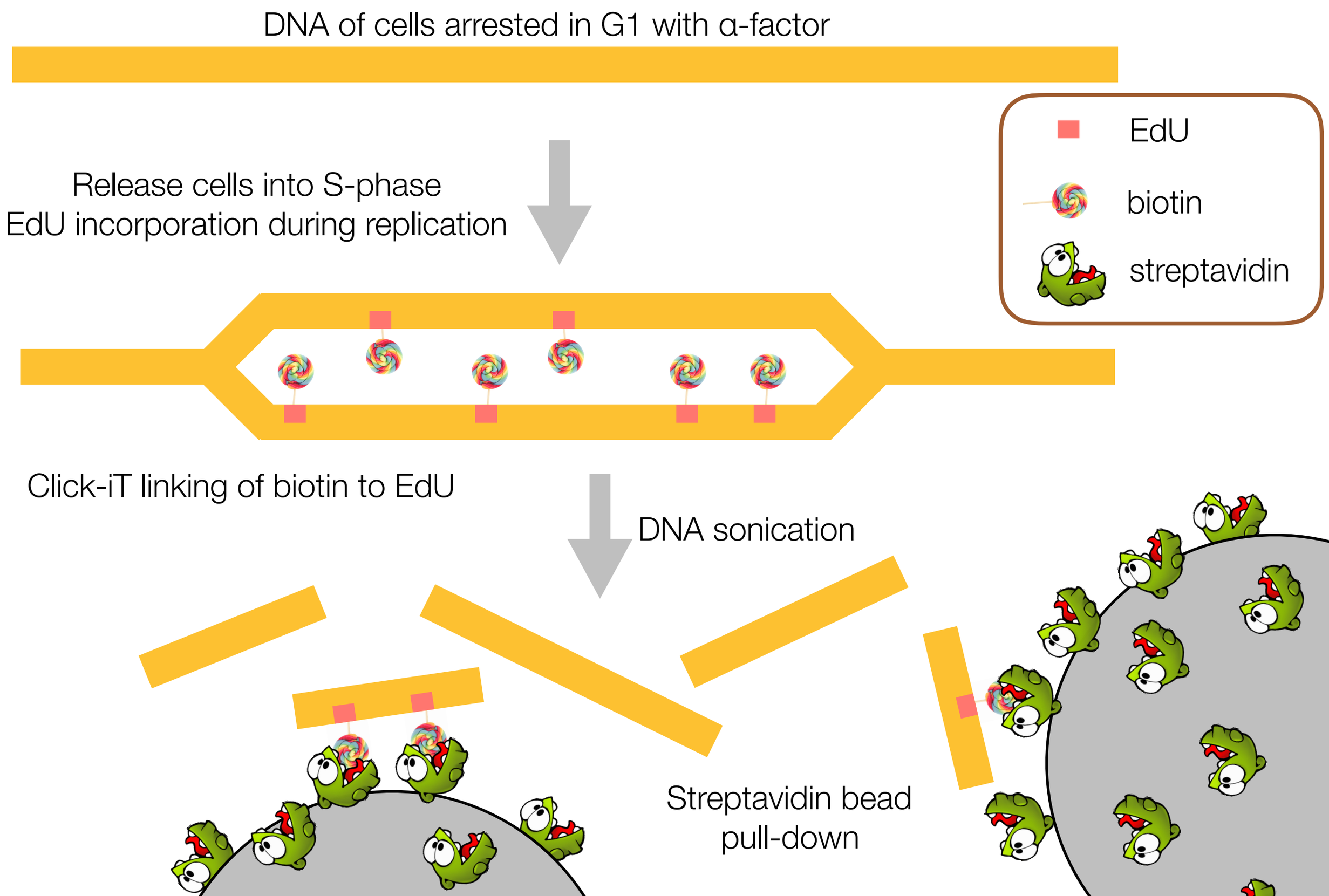
G1 phase



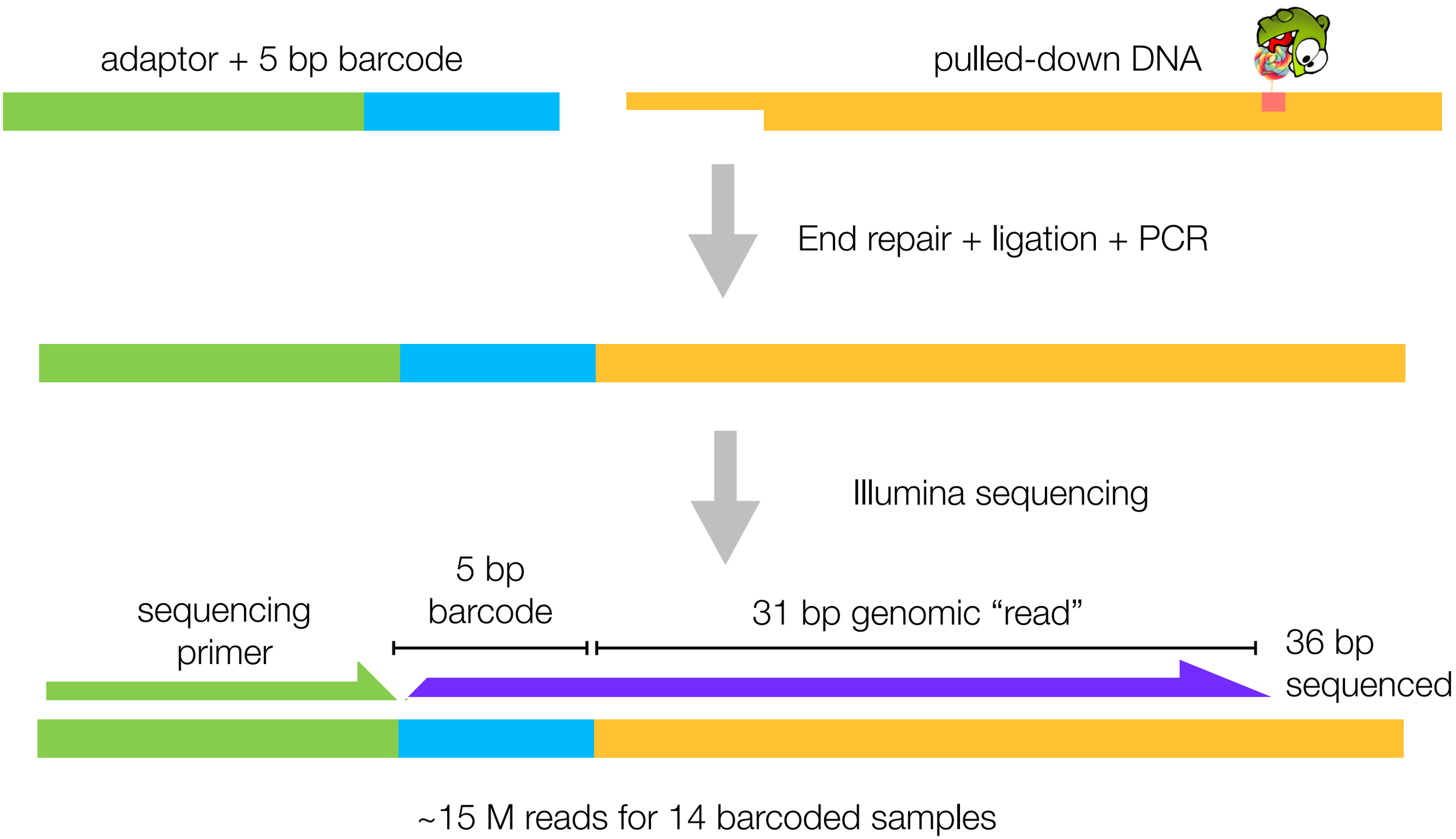
S phase



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



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@bnbdev2:~/bnb_exercise/reads$ ls -lh
total 302M
-rw-r----- 1 jkinstructor WSBSStudents 70M Aug 30  2016 A1.fastq
-rw-r----- 1 jkinstructor WSBSStudents 97M Aug 30  2016 B1.fastq
-rw-r----- 1 jkinstructor WSBSStudents 68M Aug 30  2016 C1.fastq
-rw-r----- 1 jkinstructor WSBSStudents 69M Aug 30  2016 D1.fastq
jkinstructor@bnbdev2:~/bnb_exercise/reads$
```

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

The information for each read is split over 4 lines

```
jkinstructor@bnbdev2:~/bnb_exercise/reads$ head -n 20 A1.fastq
@HANNIBAL_0056:7:1:9620:1049#0/1
GTGGTTAGTATATGGTGCAAAAGTGGTATAA
+HANNIBAL_0056:7:1:9620:1049#0/1
ggggggaeadffffccdfffffefgfgggg
@HANNIBAL_0056:7:1:1070:1061#0/1
CGAACACAAAGATCTCGTTCTACTTTTTTTG
+HANNIBAL_0056:7:1:1070:1061#0/1
f`[facddddfJdcfaa^c_fcf_dcfffc]
@HANNIBAL_0056:7:1:4279:1052#0/1
TATCCACTACCGCTATACTGGATTCGTACTC
+HANNIBAL_0056:7:1:4279:1052#0/1
hghhhhhhhhhghghghhhhhhfhhhfhhhg
@HANNIBAL_0056:7:1:4413:1064#0/1
AAGAAAACGTGCCACCATTGAGTACATCAAC
+HANNIBAL_0056:7:1:4413:1064#0/1
hhhhhhhhcffeefghhhhhgdhhfghhfb
@HANNIBAL_0056:7:1:5309:1059#0/1
AGTATACTGTGTATATAATAGATATGGAACG
+HANNIBAL_0056:7:1:5309:1059#0/1
bf`ebfcffcfbdbeac^_cfdcdffdffdf
jkinstructor@bnbdev2:~/bnb_exercise/reads$
```

read 1

read 2

read 3

read 4

read 5

@name

sequence

+name

quality scores

The yeast genome is in FASTA format

```
jkinstructor@bnbdev2:~/bnb_exercise/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
CACTTCACTCCATGGCCCATCTCTCACTGAATCAGTACCAAATGCACTCACATCATTATC
jkinstructor@bnbdev2:~/bnb_exercise/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@bnbdev2:~/bnb_exercise/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@bnbdev2:~/bnb_exercise/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

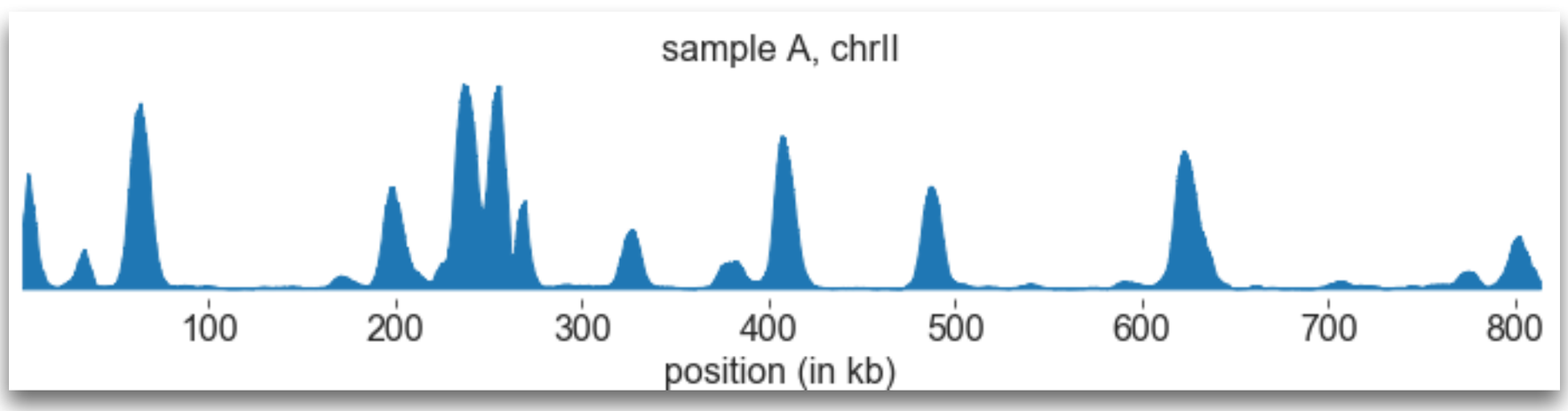
cluster analysis (bwa + samtools + bedtools)

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)



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

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

Intranet-New BlackNBlue | HPC

Not Secure | intranet.cshl.edu/administration/information-technology/hpcc/blacknblue

CSH Intranet

HOME GENERAL INFO ADMINISTRATION EDUCATION RESEARCH REQUESTS

Information Technology

Search Menu

Home

Divisions 5

Systems & Storage 1

BlackNBlue 7

Login & File Transfer

User Environment 2

UGE

Hadoop

Storage & I/O

Contact

Helpful Links

Business Systems

Desktop Support

BlackNBlue

BlackNBlue is an institutionally shared compute cluster introduced in 2012. The cluster is intended to support the full spectrum of CSHL research computing.

BlackNBlue is a 1,696-core IBM System x solution based on the M4 server line with Intel Xeon E5 (Sandy Bridge-EP) processors. The cluster was designed from 106 servers, configured as development, compute, and management nodes, using 10 Gigabit per second Ethernet networking.

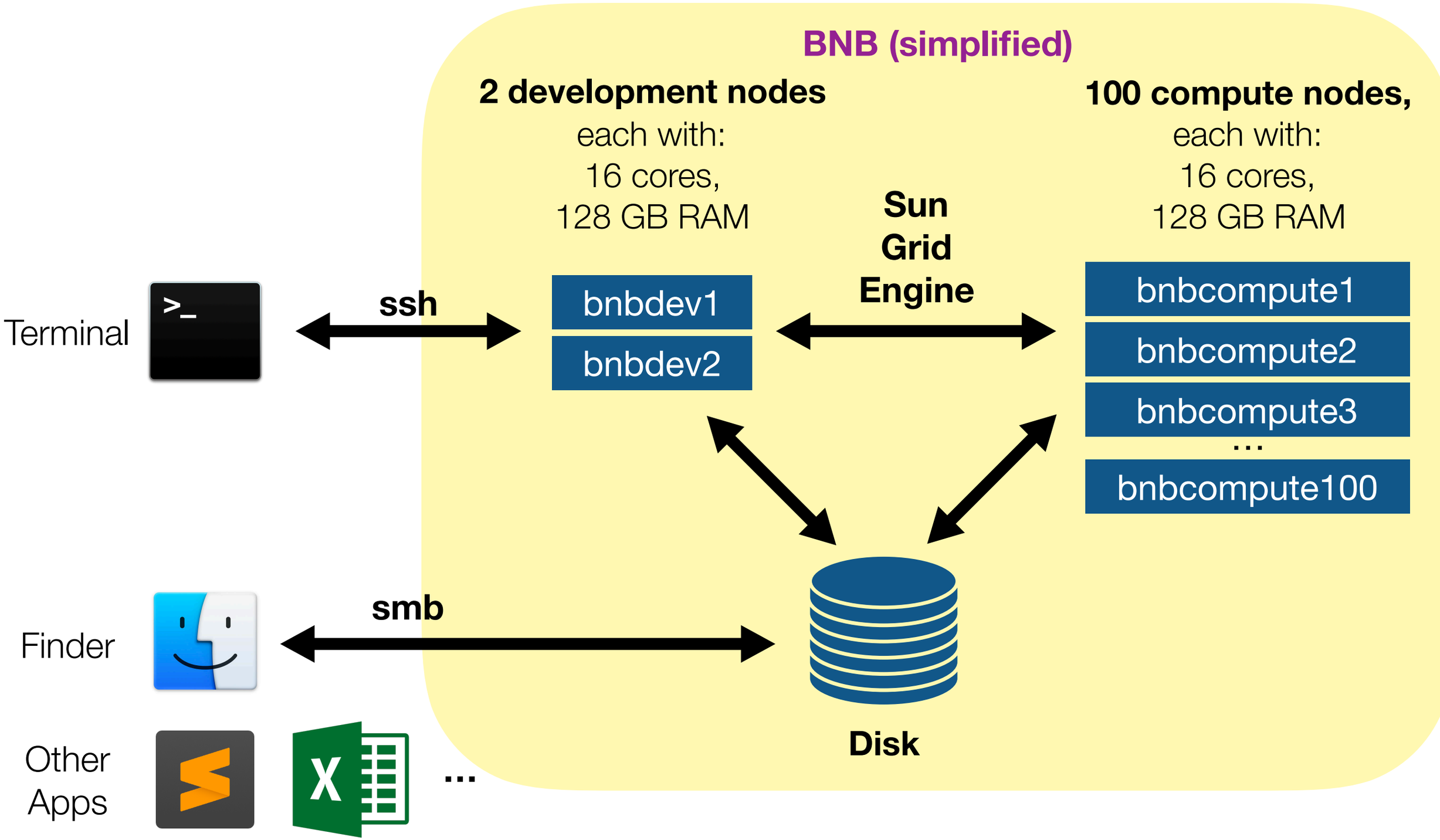
Two development nodes provide the sole point of user access to the cluster and allow for interactive development work as well as submission of jobs to the compute nodes. The cluster is administered from a pair of management nodes running UGE (formerly SGE), a "fair share" resource management system for equitable allocation of compute resources. The management nodes are configured for failover protection that ensures uninterrupted execution of batch jobs in the event that the primary management node becomes unavailable.

The development nodes and 100 compute nodes have Xeon E5-2665 processors running at 2.40 GHz. The development nodes have 64GB of memory, the compute nodes 128GB. Each node has two sockets with 8 cores per socket, for a total of 16 physical cores. Hyperthreading doubles the number of physical cores, resulting in 32 virtual cores per node, which provides a total of 3,200 UGE job slots over the 128GB compute nodes.

In addition to the standard compute nodes, the cluster has two high memory nodes, each with 1.5TB of memory. The high-memory nodes have Xeon E5-4650 processors running at 2.70 GHz. Each node has 4 sockets with 8 cores per socket, for a total of 32 physical cores. With hyperthreading, users see 64 virtual cores, or UGE job slots, for each high-memory node.

BackNBlue is connected to the DDN GridScaler storage system via GPFS, and to IBM SONAS and Isilon storage systems via NFS.

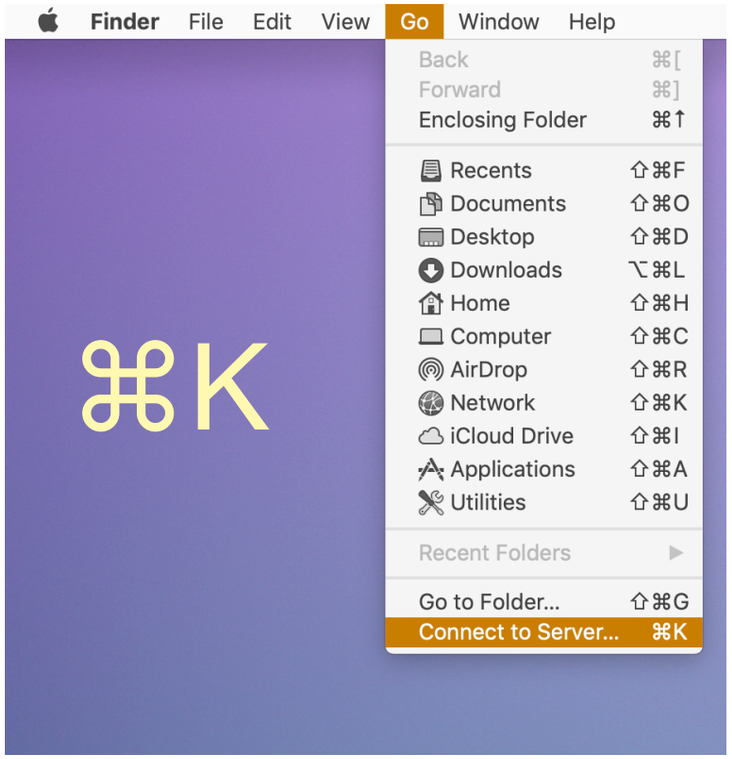
Architecture of BlackNBlue (BNB)

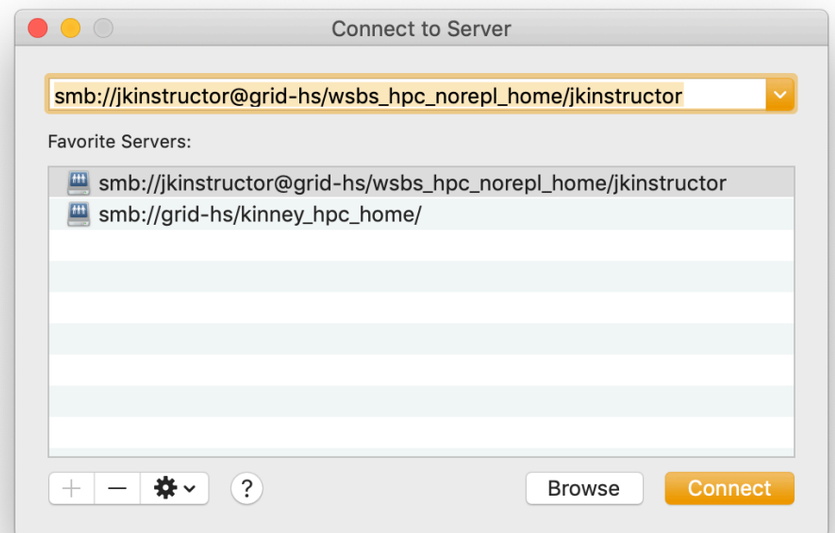


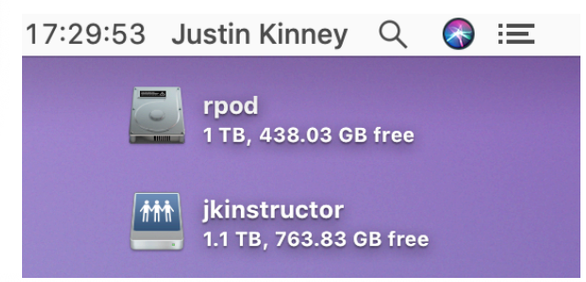
The BNB disk can be mounted using smb

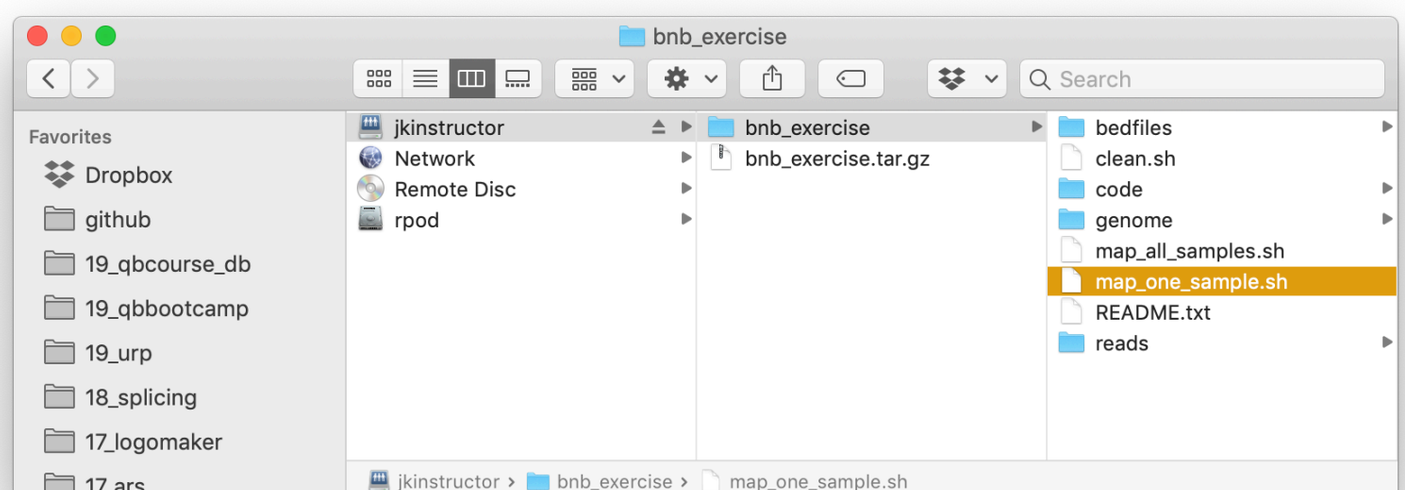
smb://you@grid-hs/wsbs_hpc_norepl_home/you

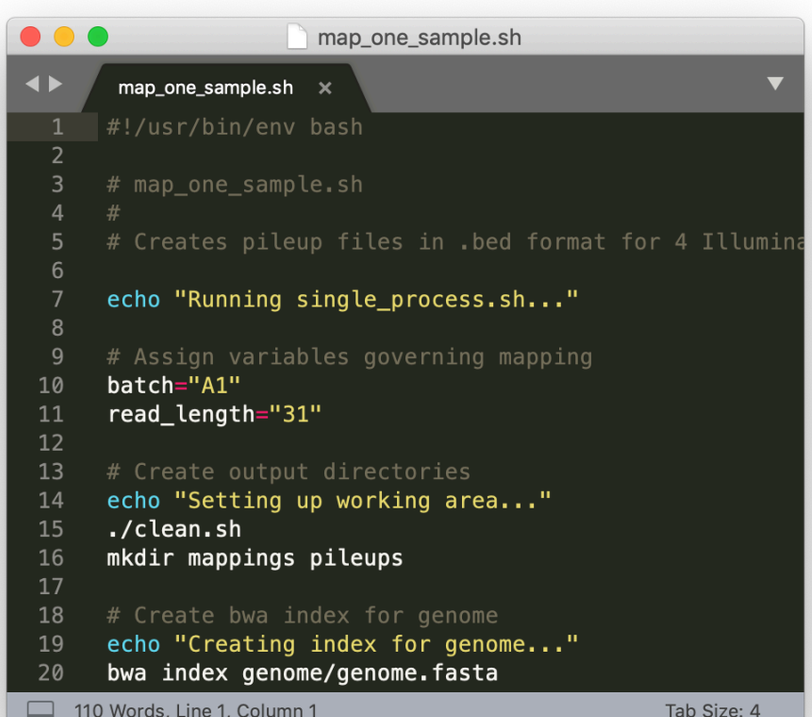
or smb://you@grid-hs/yourlab_hpc_home/you











To do this morning:

1. Copy **bnb_exercises.tar.gz** from **20_qbbootcamp/** to BNB

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.