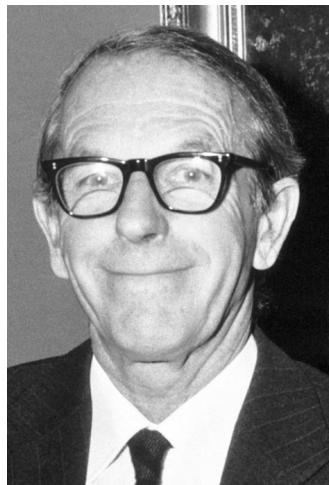


NextGen Sequencing and High-Performance Computing



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

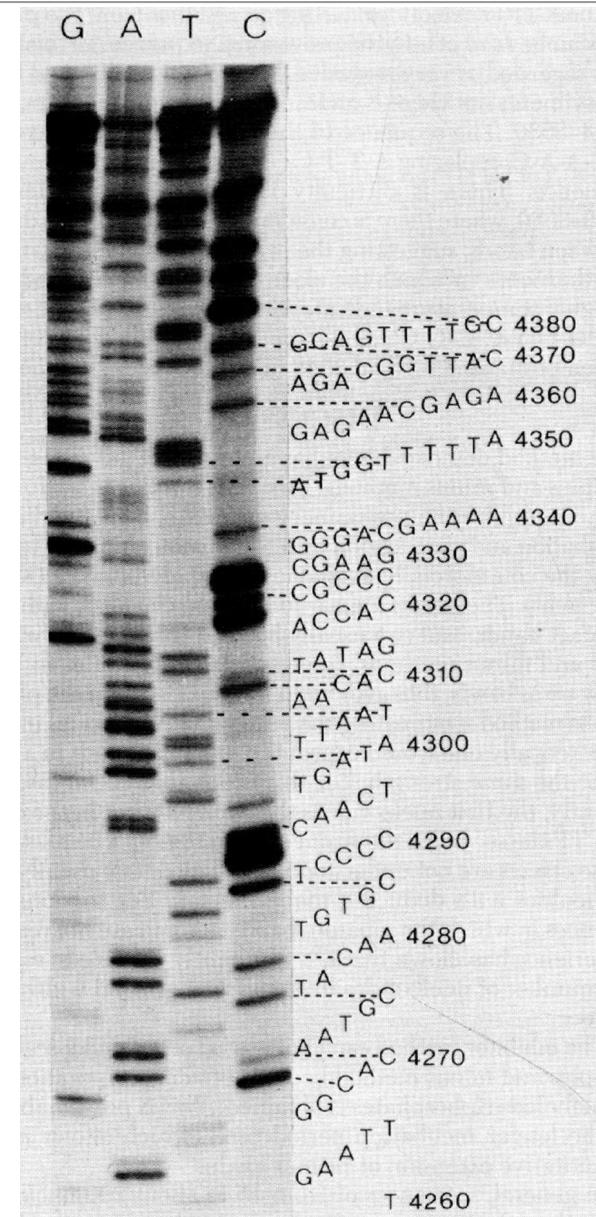


Wally Gilbert



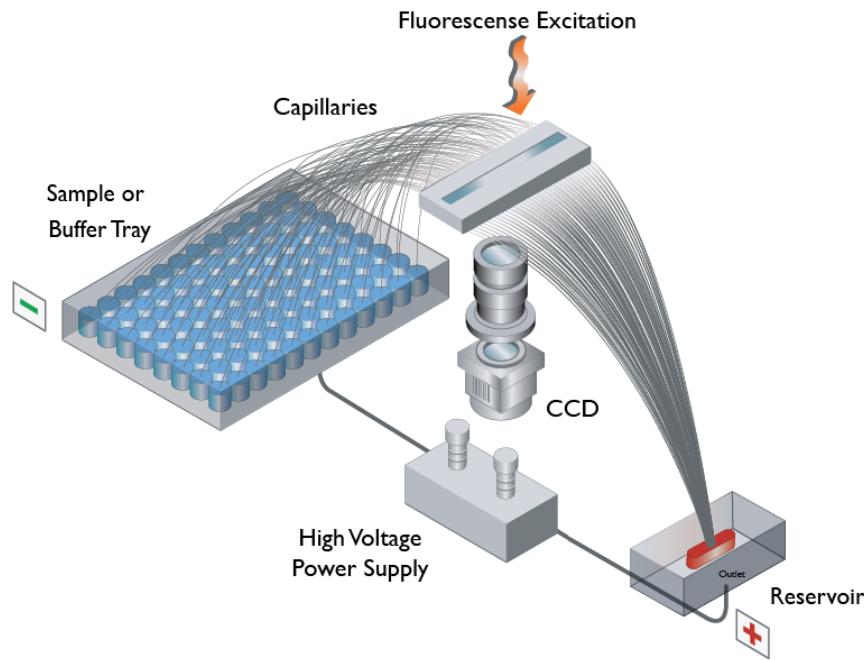
1980 Nobel Prize in Chemistry

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

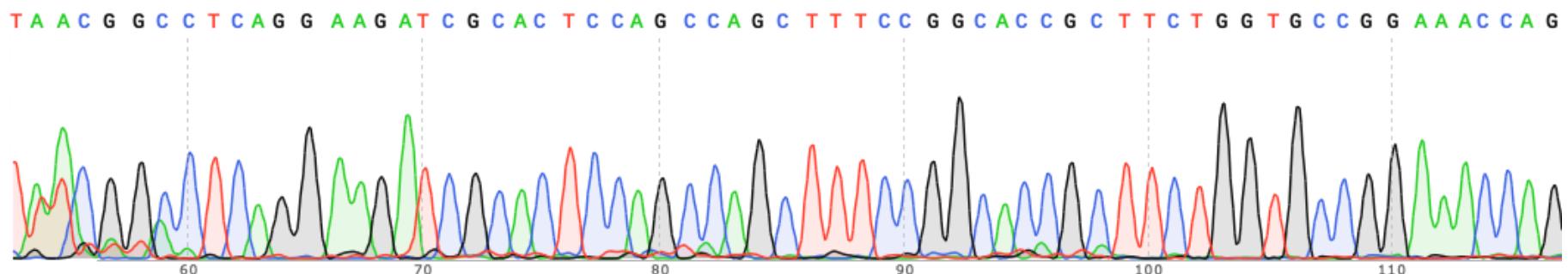


Sanger et al., PNAS, 1977

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!

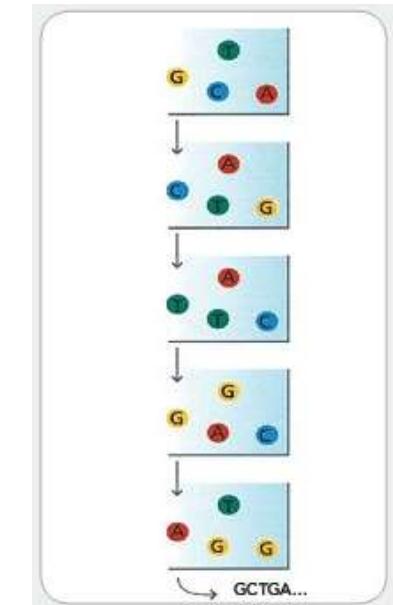
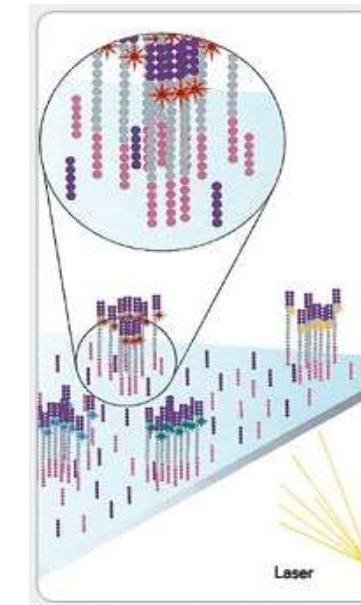
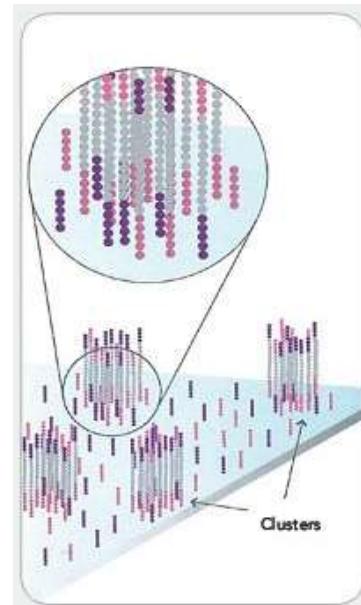
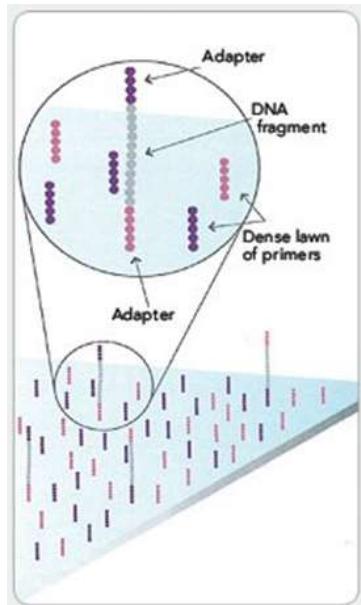
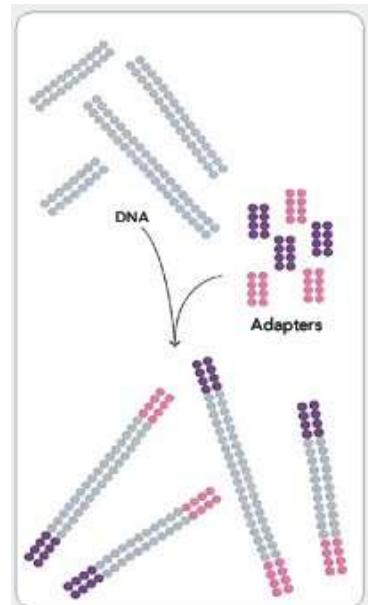


NextSeq 500

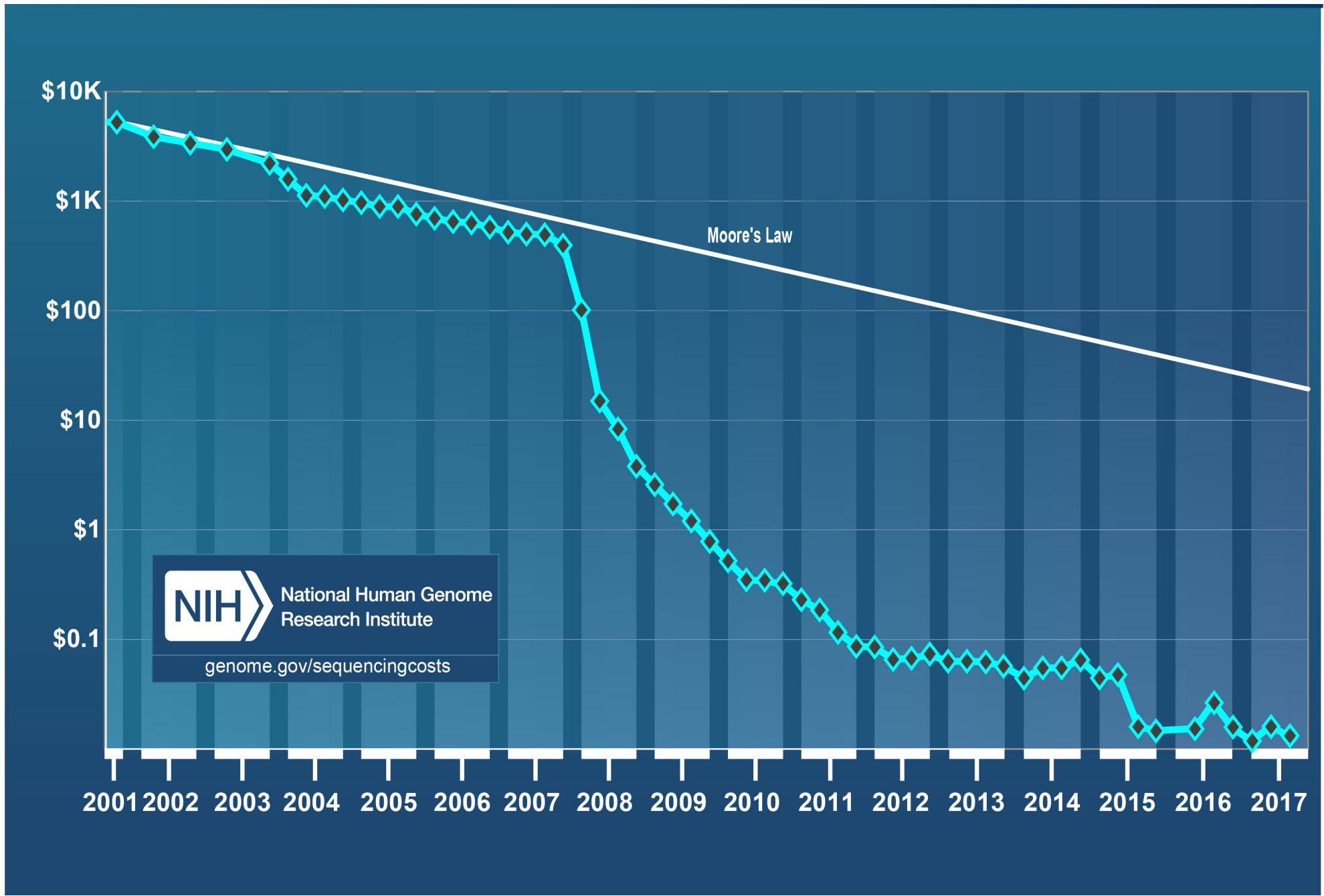
this is where the DNA goes



Flow cell



DNA sequencing has undergone an unprecedented technological revolution



DNA sequencing vs. computing



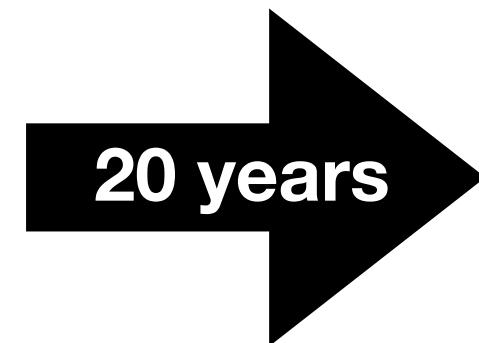
ABI 3730xl Genetic Analyzer (2007)



Illumina HiSeq 2500 (2012)



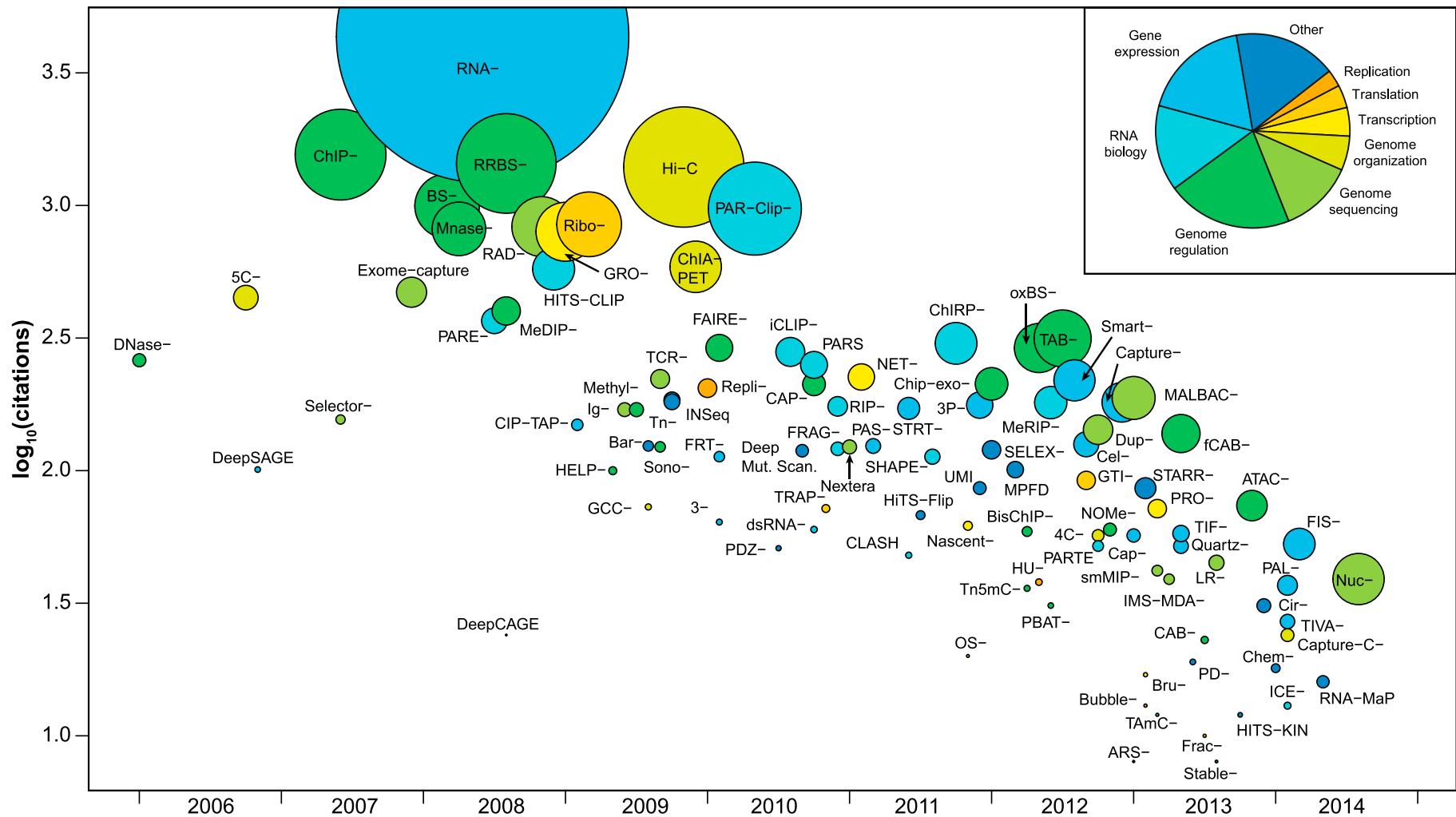
Palm Pilot (1997)



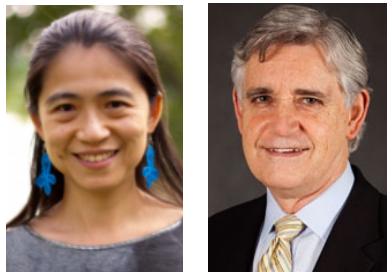
iPhone X (2017)

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



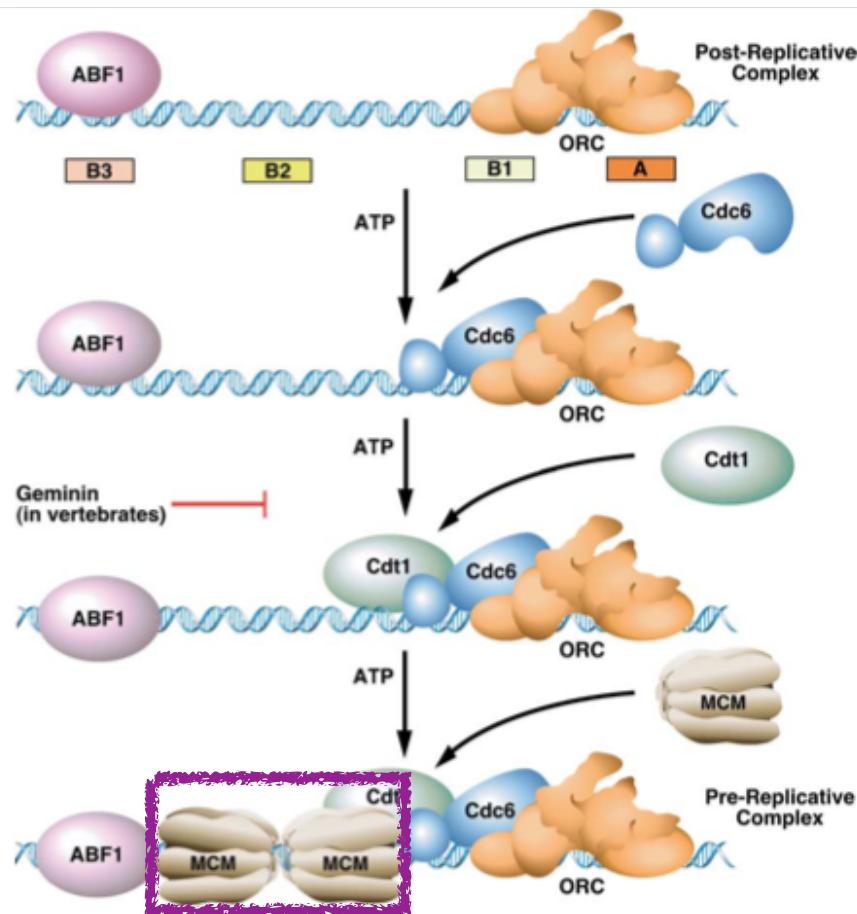
Domain within the helicase subunit Mcm4 integrates multiple kinase signals to control DNA replication initiation and fork progression

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

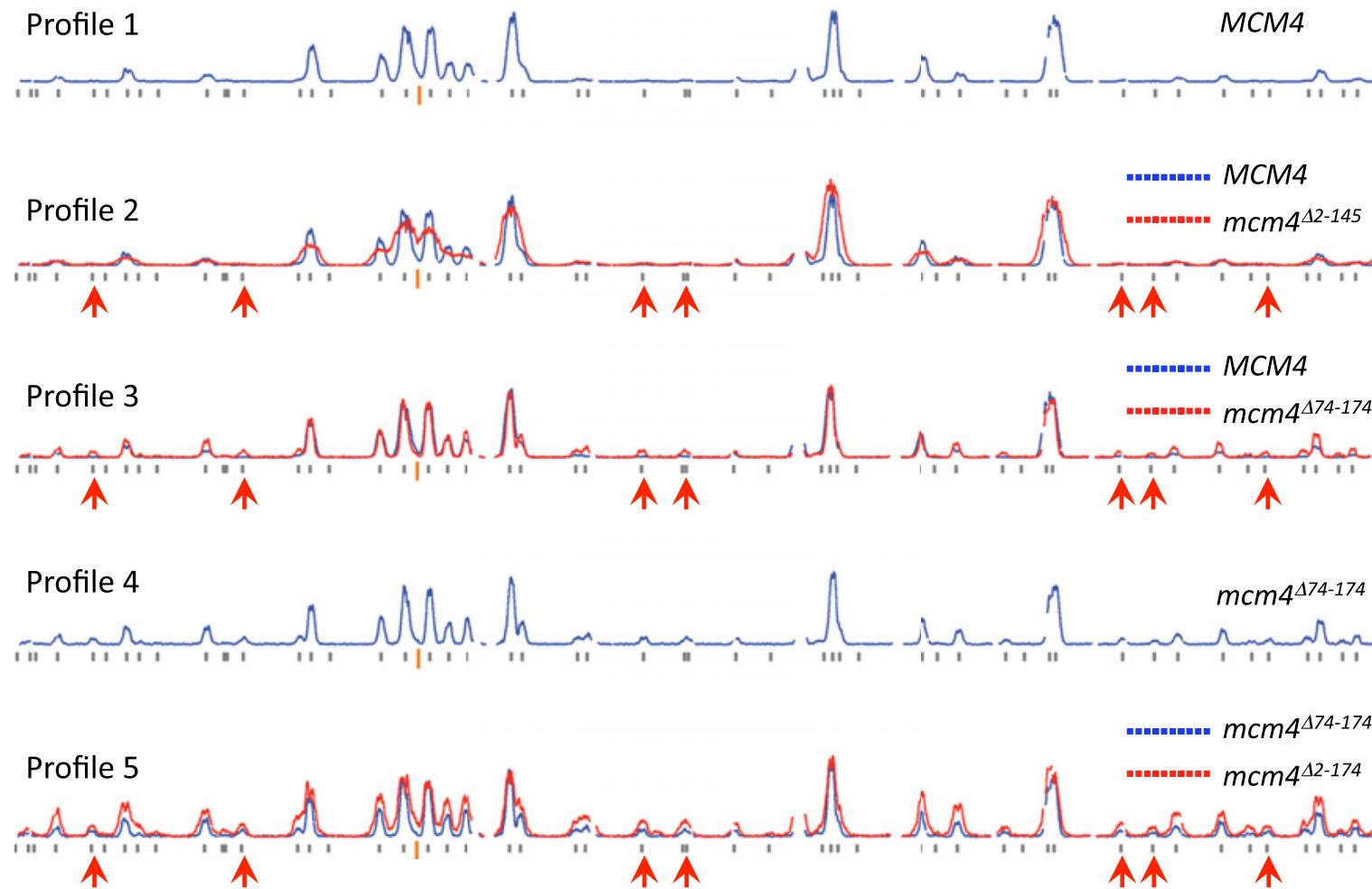
PNAS | Published online April 16, 2014 | E1899–E1908



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

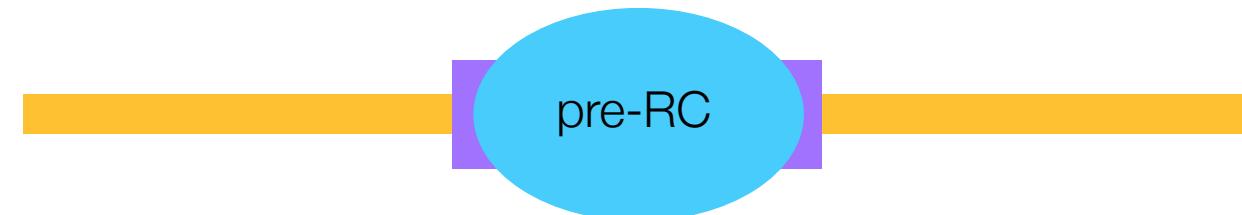
— old ssDNA

— new ssDNA

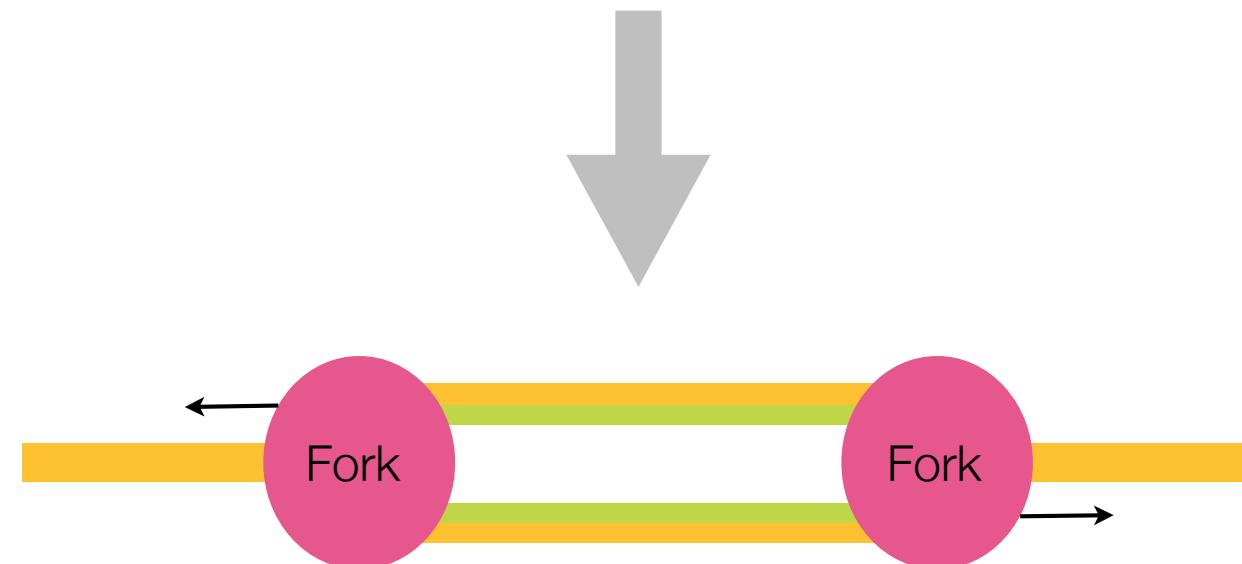


S. cerevisiae

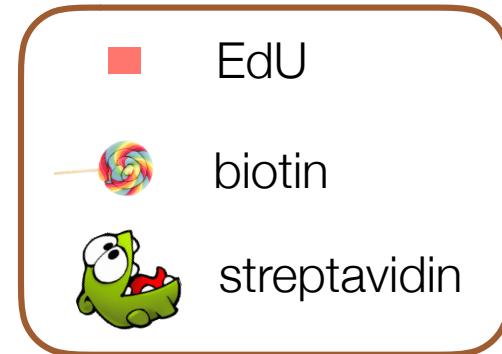
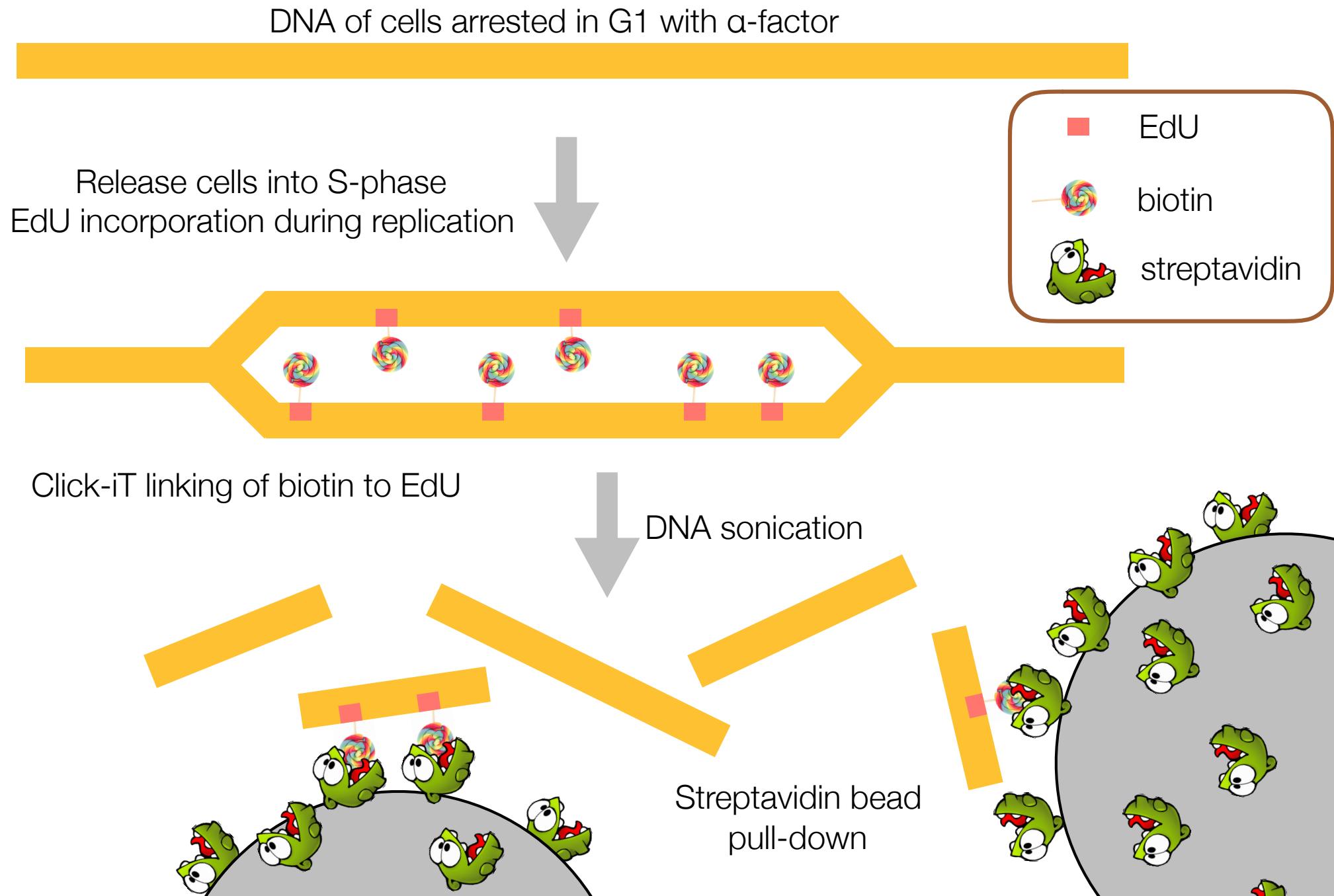
G1 phase



S phase

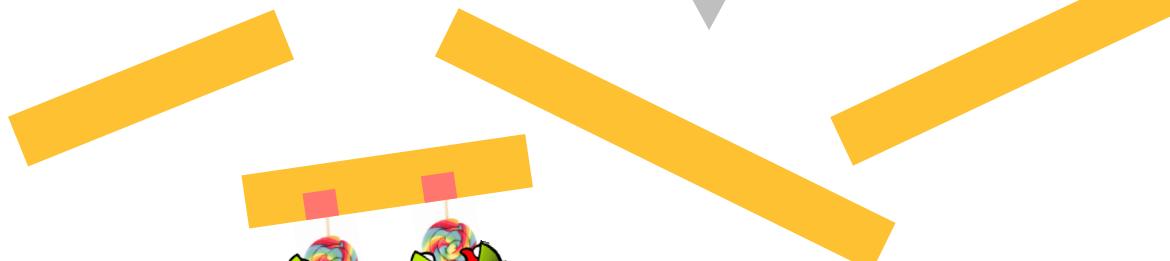


Newly replicated DNA can be isolated using an EdU pull-down assay

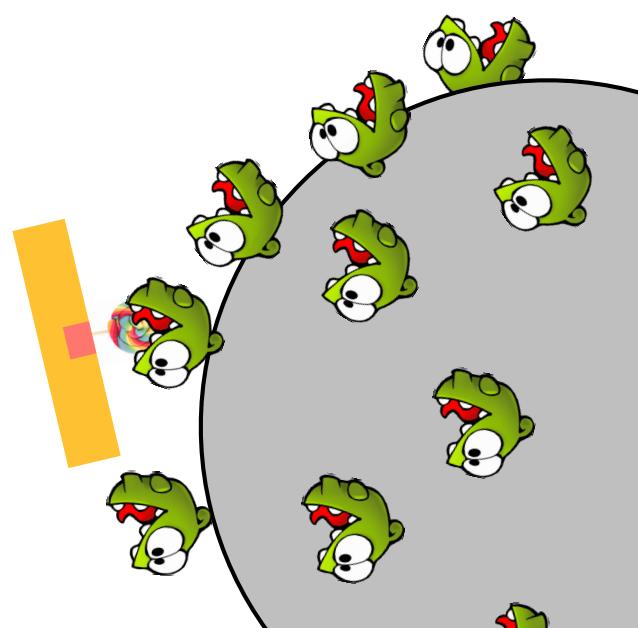
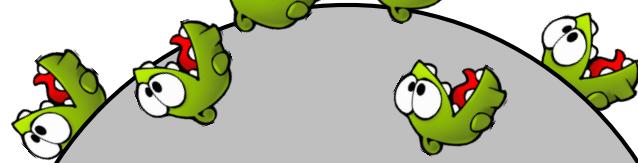


Click-iT linking of biotin to EdU

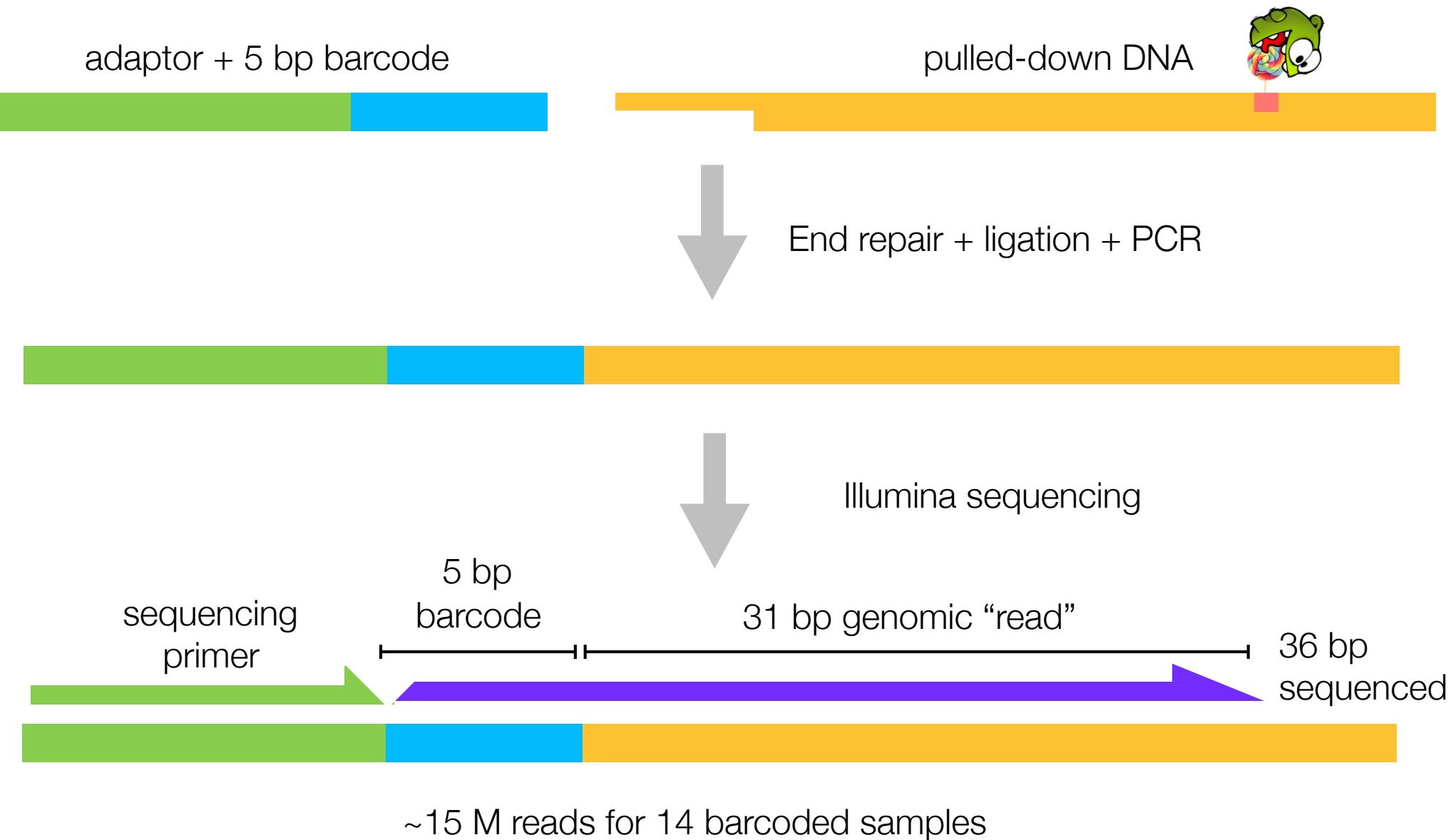
DNA sonication



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

```
jkinney@rpod:~/Desktop/bnb_exercise/reads$ ls  
A1.fastq  B1.fastq  C1.fastq  D1.fastq
```

Each FASTQ files is ~60-100 megabytes

```
jkinney@rpod:~/Desktop/bnb_exercise/reads$ ls -lah  
total 616784  
drwxr-x---@ 6 jkinney  staff  192B Aug 28 15:00 ./  
drwxr-x---@ 10 jkinney  staff  320B Sep  1  2016 ../  
-rw-r-----@ 1 jkinney  staff  69M Aug 30  2016 A1.fastq  
-rw-r-----@ 1 jkinney  staff  96M Aug 30  2016 B1.fastq  
-rw-r-----@ 1 jkinney  staff  67M Aug 30  2016 C1.fastq  
-rw-r-----@ 1 jkinney  staff  68M Aug 30  2016 D1.fastq
```

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

The information for each read is split over 4 lines

read 1

```
jkinney@rpod:~/Desktop/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
ggggggaeadffffcccdfffffffefgfgggg
```

read 2

```
@HANNIBAL_0056:7:1:1070:1061#0/1
CGAACACAAAGATCTCGTTCTACTTTTTTG
+HANNIBAL_0056:7:1:1070:1061#0/1
f`[facddddfJdcfaa^c_fcf_dcffffc]
```

← @name
← sequence
← +name
← quality scores

read 3

```
@HANNIBAL_0056:7:1:4279:1052#0/1
TATCCACTACCGCTATACTGGATTGTA
+HANNIBAL_0056:7:1:4279:1052#0/1
hghhhhhhhhhghghghhhhhfhhhfhhhg
```

read 4

```
@HANNIBAL_0056:7:1:4413:1064#0/1
AAGAAAACGTGCCACCATTGAGTACATCAAC
+HANNIBAL_0056:7:1:4413:1064#0/1
hhhhhhhhcffeeffghhhhgdfhhfhhfb
```

read 5

```
@HANNIBAL_0056:7:1:5309:1059#0/1
AGTATACTGTGTATATAATAGATATGGAACG
+HANNIBAL_0056:7:1:5309:1059#0/1
bf`ebfcffcfbdbeac^_cfcdffffdf
```

The yeast genome is in FASTA format

```
jkinney@rpod:~/Desktop/bnb_exercise/genome$ head genome.fasta
>1 ref|NC_001133| [org=Saccharomyces cerevisiae] [strain=S288C] [moltype=genomic] [chromosome=I] [note=R64-1-1]
CCACACCACACCCACACACACACACACACACACACACACACACACACA
CATCCTAACACTACCCCTAACACAGCCCTAACCTAACCCCTGGCCAACCTGTCTCTCAACTT
ACCCTCCATTACCCCTGCCTCCACTCGTTACCCCTGTCCCATTCAACCATAACACTCCGAAC
CACCATCCATCCCTACTTACTACCACTCACCCACCGTTACCCCTCCAATTACCCATATC
CAACCCACTGCCACTTACCCATTACCCATTACCATCCACCATGACCTACTCACCATAAC
TGTTCTTCTACCCACCATATTGAAACGCTAACAAATGATCGTAAATAACACACACACGTGCT
TACCCCTACCACTTTATACCACCAACATGCCATACTCACCCCTCACTTGTATACTGATT
TACGTACGCACACGGATGCTACAGTATATACCATCTCAAACCTACCCCTACTCTCAGATT
CACTTCACTCCATGGCCCATCTCTCACTGAATCAGTACCAAATGCACTCACATCATTATG
```

Each header line starts with '>' time track

19_qboot items
10_qboot items
bnb_exercise 8 items
bnb_exercise 82.1 MB
19_qboot 98.8 MB

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

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

```
jkinney@rpod:~/Desktop/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]
```

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)

sample A, chrII



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

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

The screenshot shows a web browser window for the 'Intranet-New BlackNBlue | HPC' site. The address bar indicates the URL is <http://intranet.cshl.edu/administration/information-technology/hpcc/blacknblue>. The page title is 'BlackNBlue'. The header includes a CSHL logo, navigation links for HOME, GENERAL INFO, ADMINISTRATION, EDUCATION, RESEARCH, and REQUESTS, and a search icon.

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.

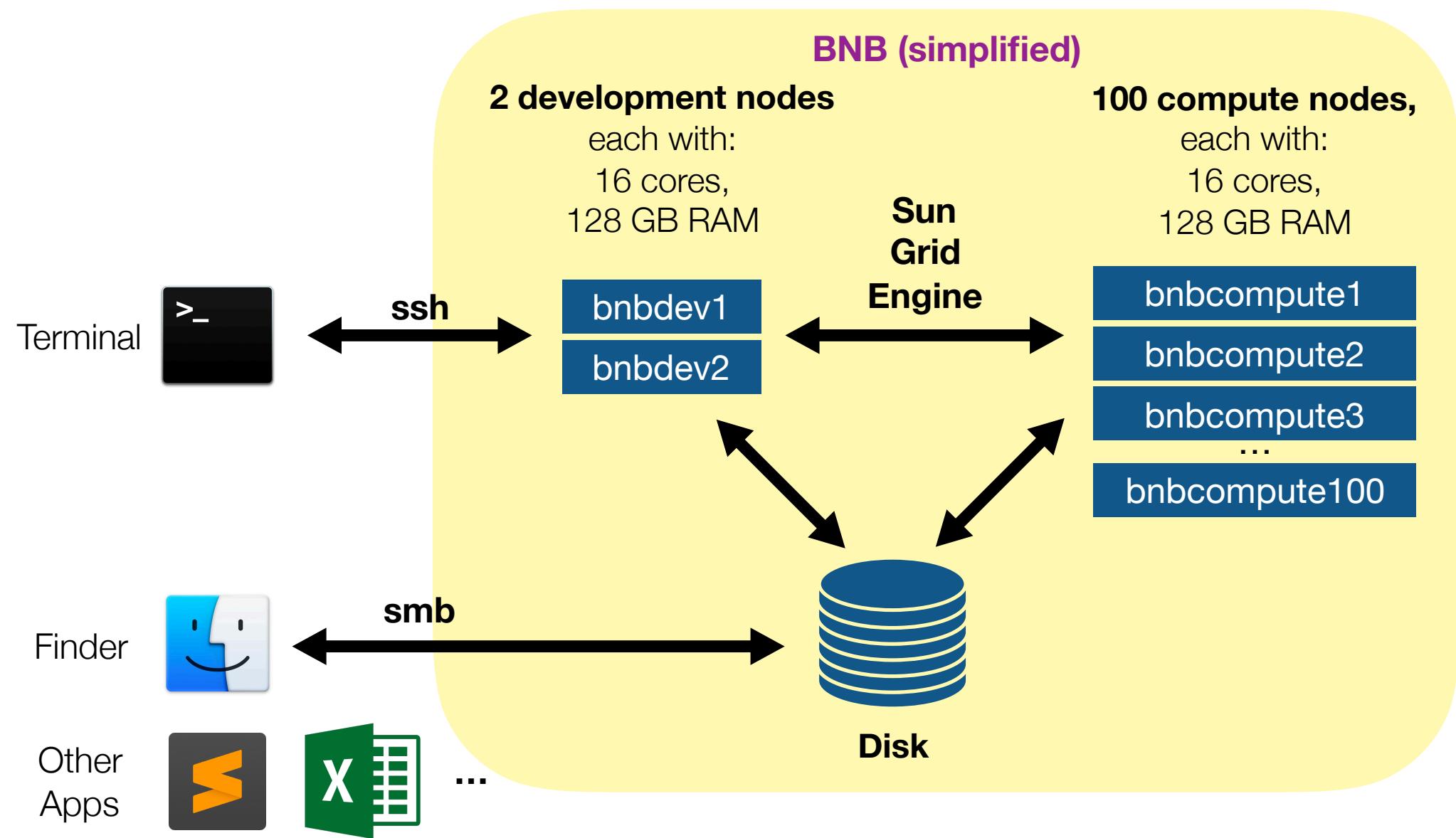
Two development nodes provide the sole point of user access to the cluster and allow for interactive development work as well as submission of jobs to the compute nodes. The cluster is administered from a pair of management nodes running UGE (formerly SGE), a "fair share" resource management system for 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.

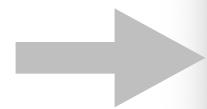
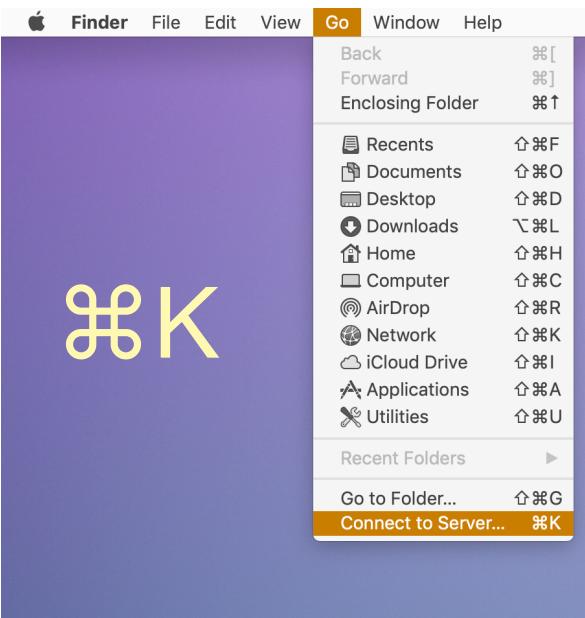
In addition to the standard compute nodes, the cluster has two high memory nodes, each with 1.5TB of memory. The high-memory nodes have Xeon E5-4650 processors running at 2.70 GHz. Each node has 4 sockets with 8 cores per socket, for a total of 32 physical cores. With hyperthreading, users see 64 virtual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

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

Architecture of BlackNBlue (BNB)

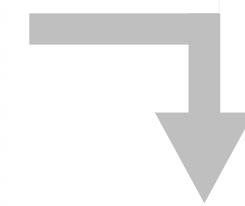
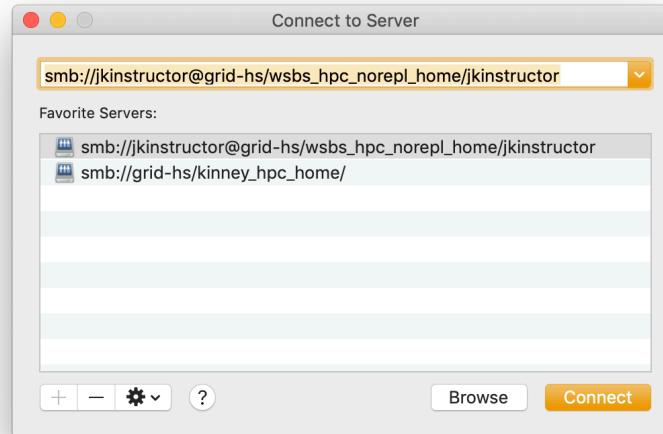


The BNB disk can be mounted using smb



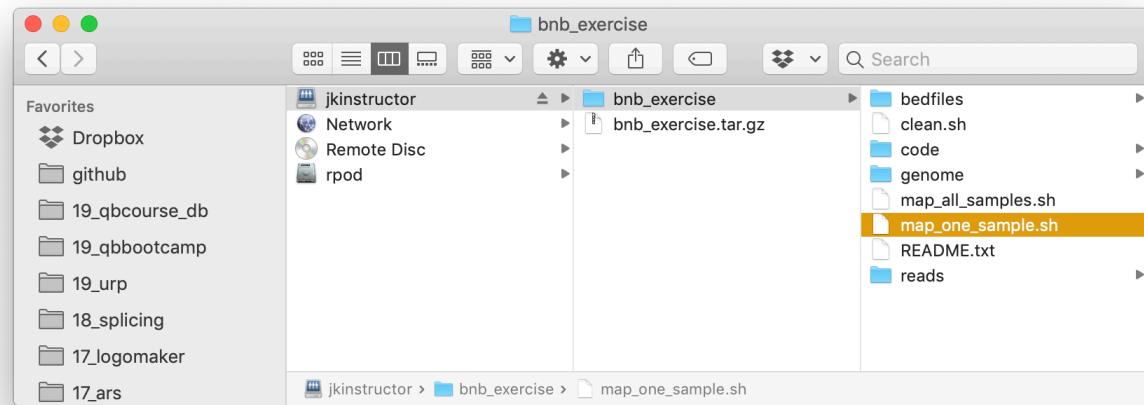
smb://you@grid-hs/wsbs_hpc_norepl_home/you

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



```
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. The script starts with a shebang line, followed by comments and assignments for variables like 'batch' and 'read_length'. It then creates output directories 'mappings' and 'pileups', and finally creates a bwa index for a genome file named 'genome.fasta'.



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.