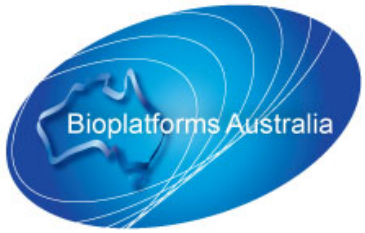


Introduction to Next Generation Sequencing

Philippe Moncuquet| CSIRO, Canberra

Mathias Haimel| EBI, UK
Susan Corley| SBI, Sydney

24th June 2015

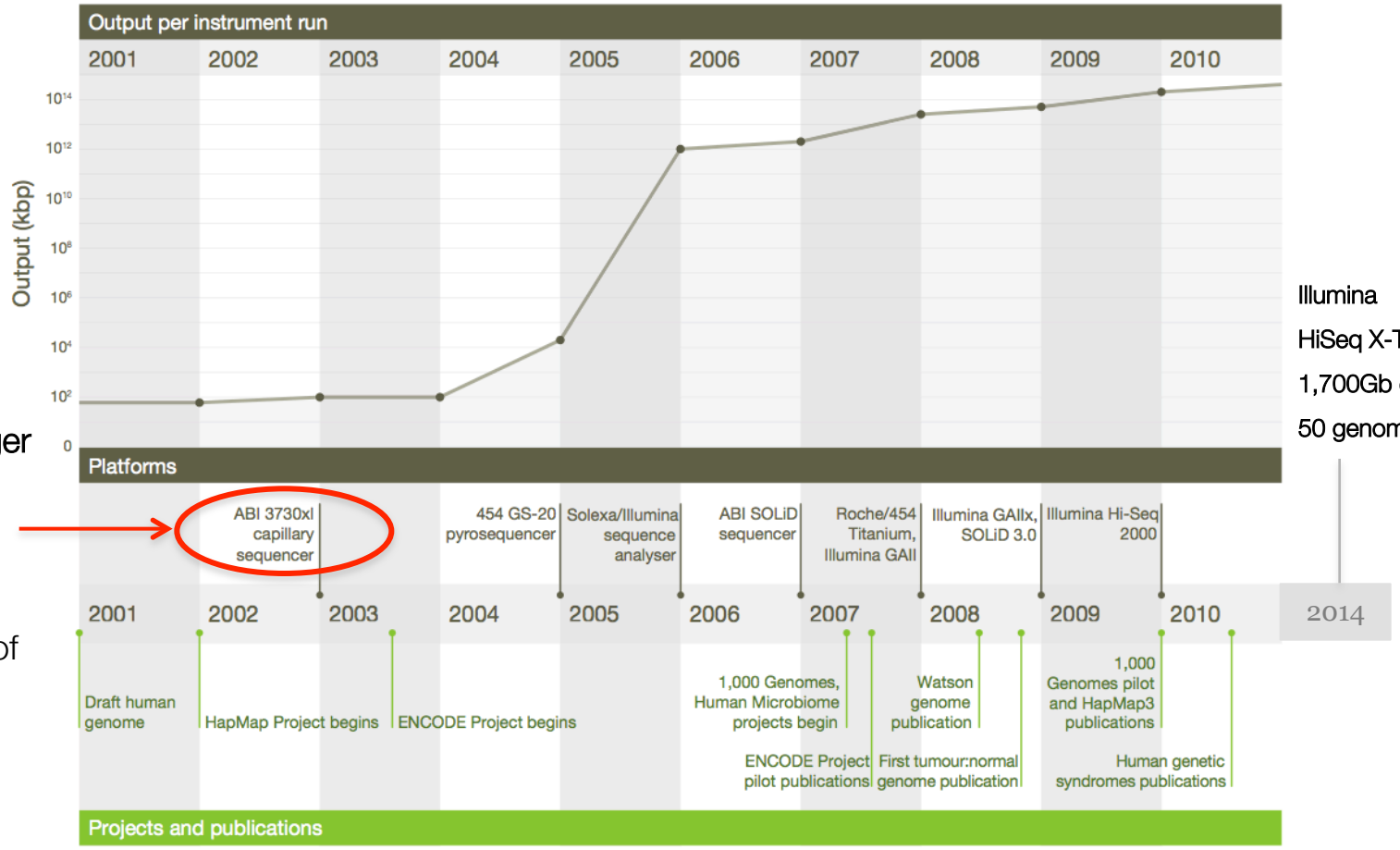
EMBL-EBI

Presentation Overview

1. Next Generation sequence fundamentals
2. NGS Data formats
3. NGS Data Management

Ramaciotti Centre for Genomics

Sequencing – Technology Timeline



Capillary/Sanger

0.000,076 Gb per run

Only 0.026% of genome x 1

Illumina

HiSeq X-Ten

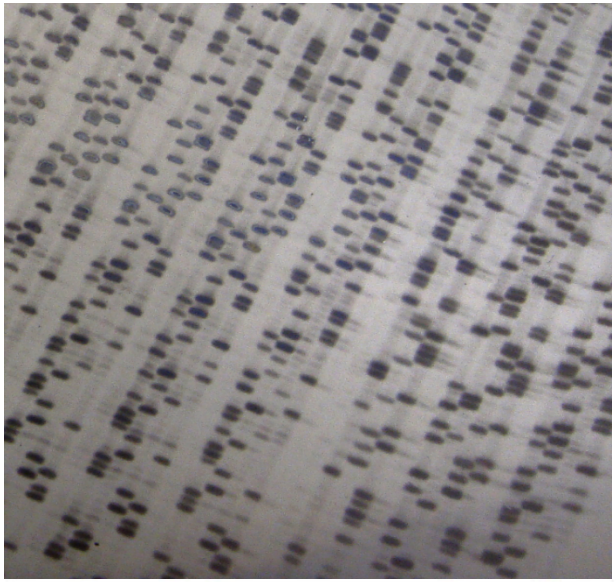
1,700Gb every day

50 genomes a day

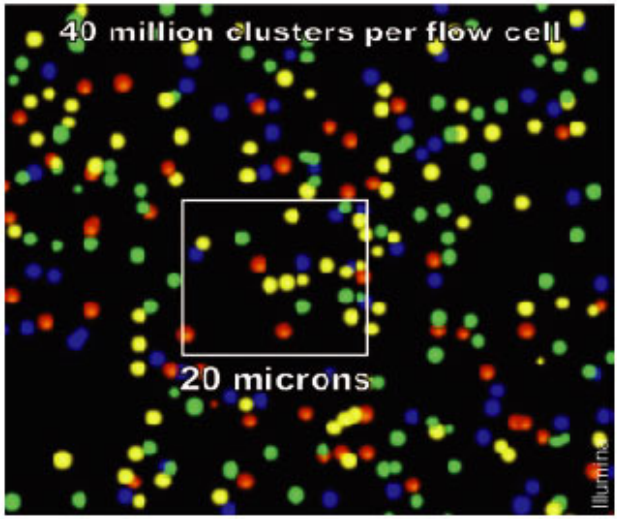
3.2 Gb = 3.2 x 10^9 bp = human genome

Image: E Mardis, Nature, 2011

Sanger versus Next Gen



Capillary electrophoresis (CE)-based Sanger sequencing



In principle, the concept behind NGS technology is similar to CE

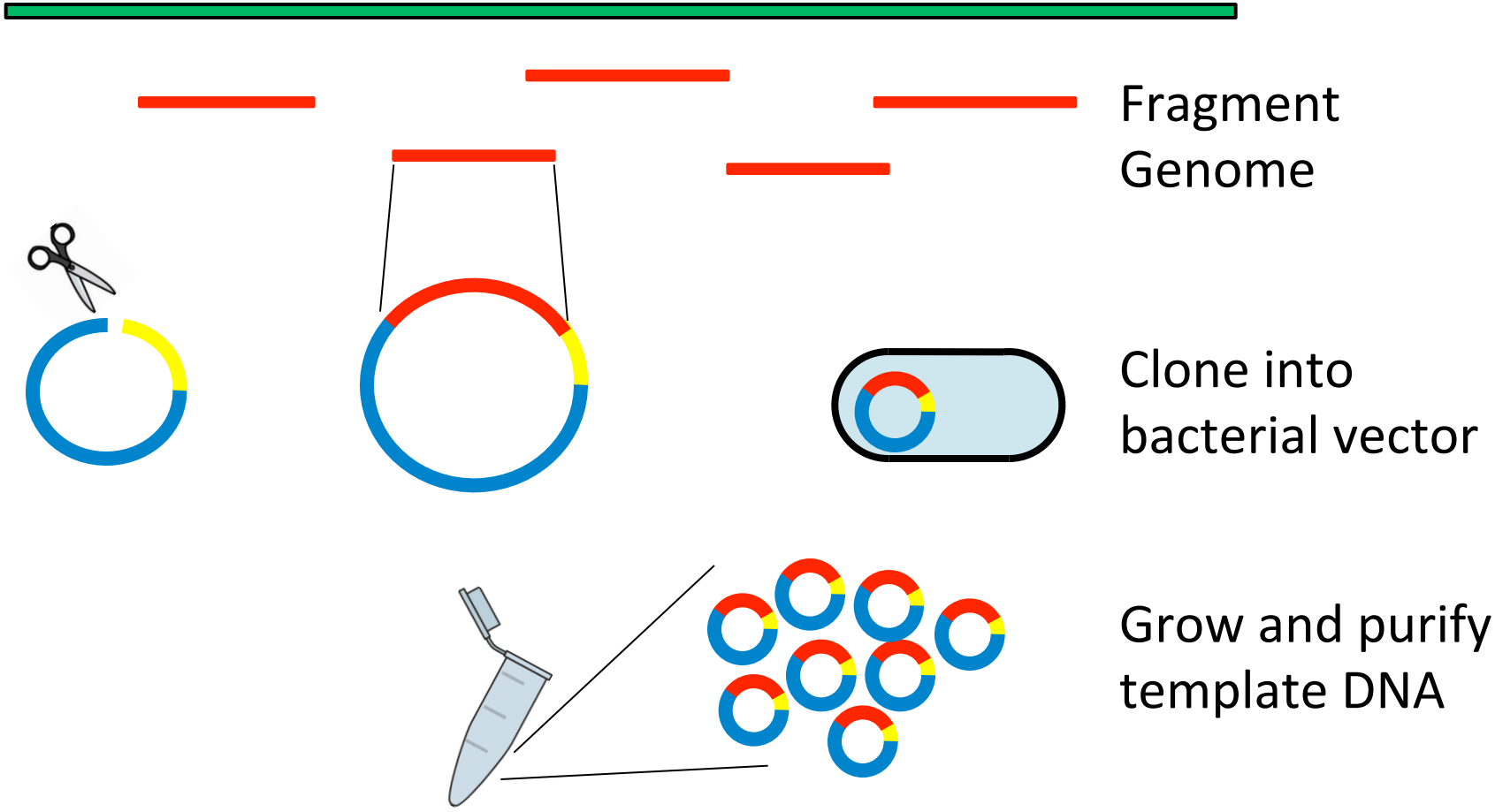
- *Sequentially identify the bases of a small fragment of DNA from signals emitted as each fragment is re-synthesized from a DNA template strand.*

- *NGS extends this process across millions of reactions in a massively parallel fashion, rather than being limited to a single or a few DNA fragments*

Currently on HiSeq up to **1.5 x 10^9** clusters per flow cell

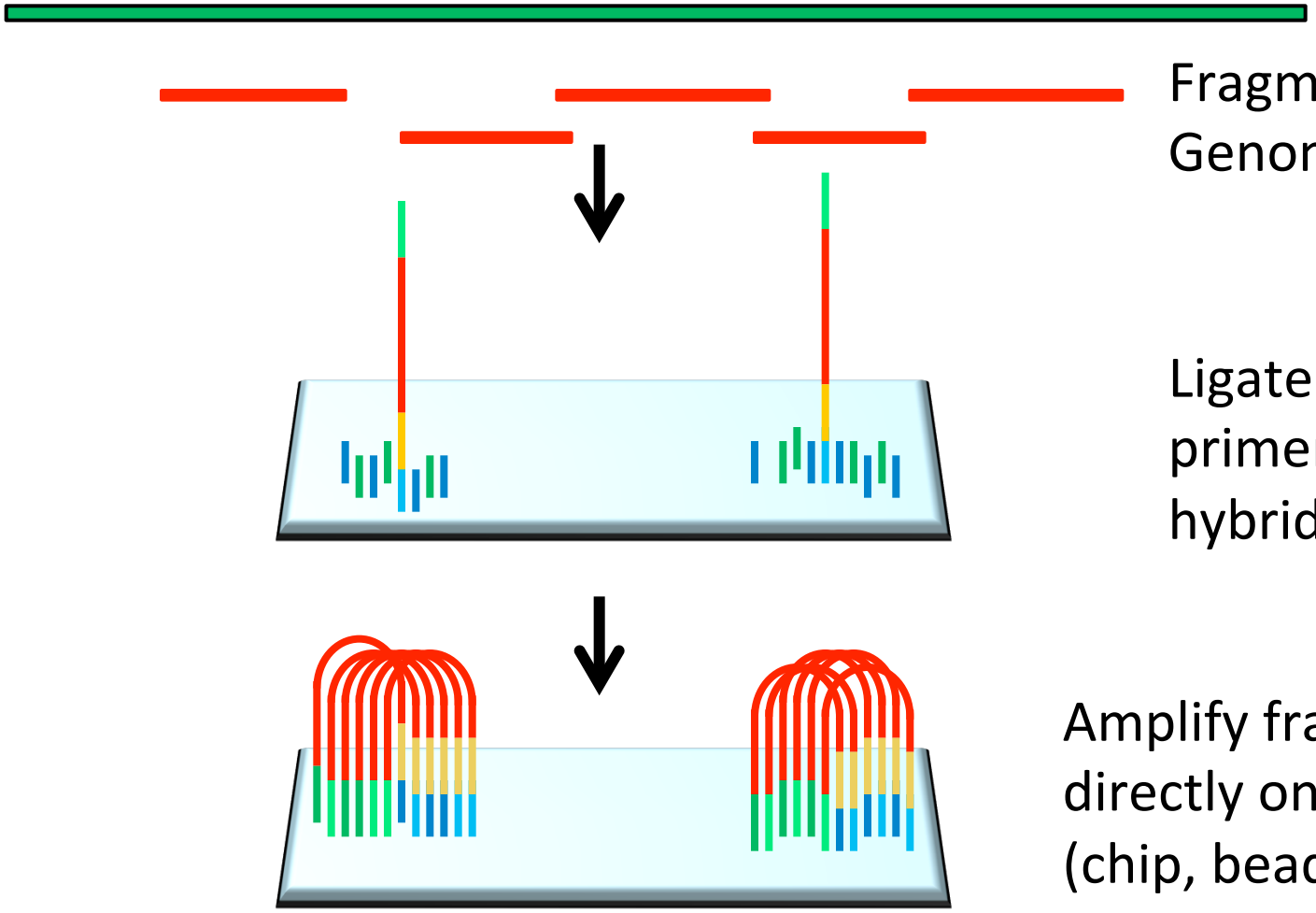
Sample Preparation -Sanger

Library preparation more involved—each sample must contain a single template, requiring purification from single bacterial, yeast colonies, or phage plaques



Sample Preparation – Next Gen

Library preparation more streamlined—sample can consist of a population of DNA molecules that do not require clonal purification

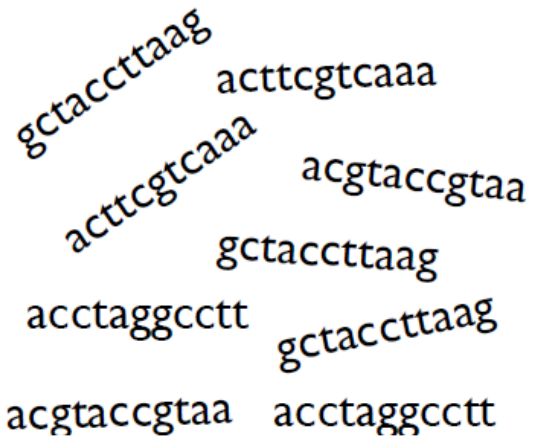


Fragment Genome

Ligate primers and hybridize

Amplify fragments directly on a surface (chip, bead)

NGS Pipeline



Library Preparation → Sequencing → Data Analysis

New Illumina Sequencers: HiSeq X Ten and NextSeq 500

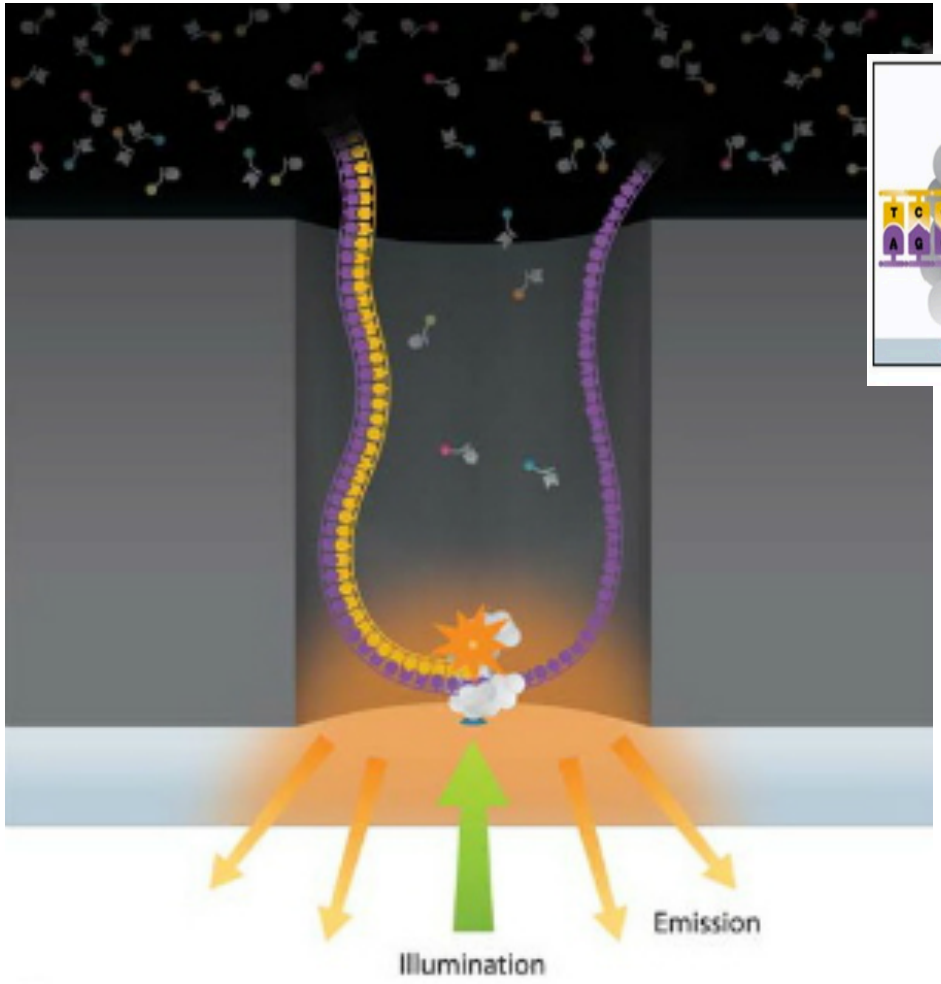
HiSeq X Ten

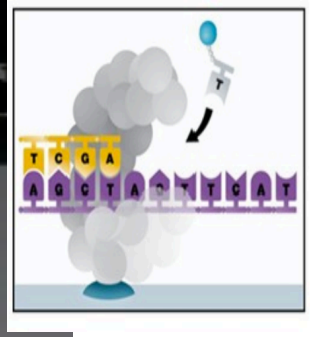
- A factory scale sequencer for population sequencing
- Throughput of 150Gb with 400 Million 2x 150bp reads.

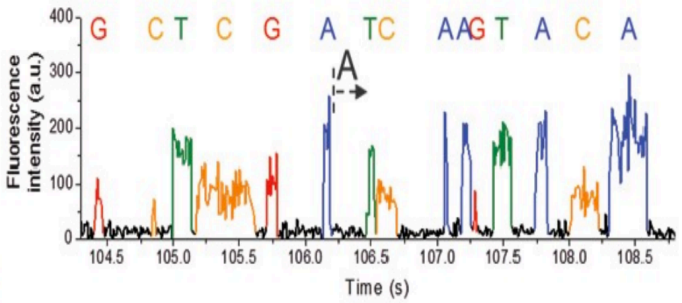
NextSeq 500

- A new desktop high throughput sequencer.
- Power of HiSeq, but with the size of MiSeq.
- Capability to sequence 30X human genome in 30 hours

Pac Bio







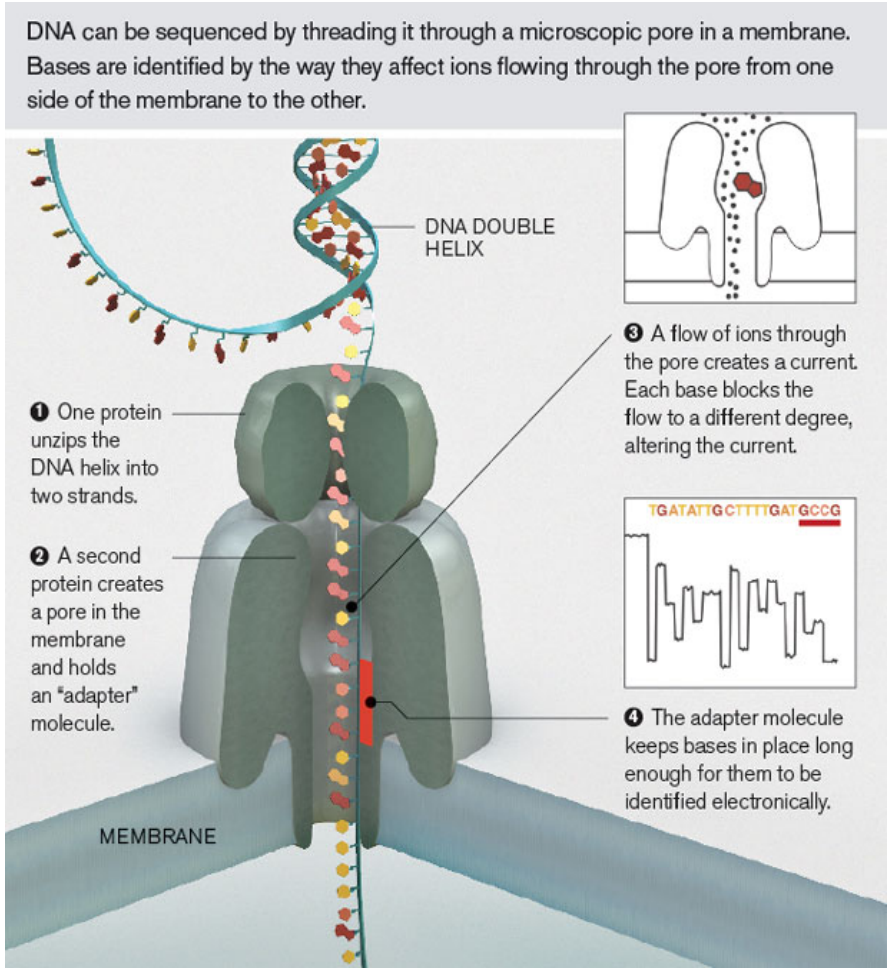
Zero-mode waveguide (ZMW)

- DNA polymerase is affixed to the bottom of a tiny hole (~70nm).
- Only the bottom portion of the hole is illuminated allowing for detection of incorporation of dye-labeled nucleotide.

Pac Bio

- **Advantages:**
  - Really long reads (up to 50kb)
  - Near random distribution of errors
    - which allow correction in high coverage data
  - No PCR bias
  - Direct detection of modified nucleotides
    - A really high coverage is needed for some modification detection.
  - Circular Consensus Reads (CCS)
    - CCS reads have a low error rate and a length sufficient to solve many long repeats in genomes
- **Limitations:**
  - The amount of input materials
  - The error rate
  - The cost

Oxford Nanopore



- Announced Feb. 2012 at ABGT conference.
- Measure changes in ion flow through nanopore.
- Potential for long read lengths and short sequencing times.

http://www2.technologyreview.com

Oxford Nanopore

- **Advantages:**
  - Really long reads (up to 80-100kb)
  - Low-cost, portable instrument
  - Easy sample prep
  - Can repetitively sequence a given molecule to generate higher quality data
- **Limitations:**
  - Just commercially available
  - The error rate
  - Whole-genome sequencing remains a challenge
  - Performance still being tested and optimized
  - Data processing

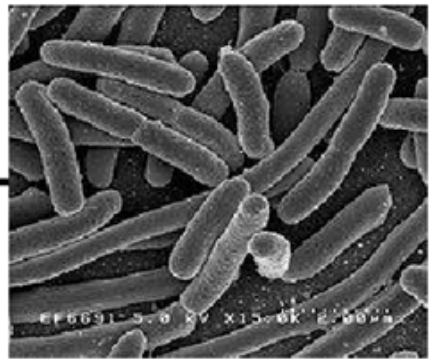
instrument	Nanopore	Pacbio	Ion Torrent	454	Illumina	SOLiD
Method	Single-molecule in real-time	Single-molecule in real-time	Ion semiconductor	Pyro	synthesis	Ligation
Read length	Up to 100kb	Up to 50kb	400 bp	700 bp	50 to 600 bp	50+35 or 50+50 bp
Error type	indel	indel	indel	indel	substitution	A-T bias
single-Pass Error rate %	15	13	~1	~0.1	~0.1	~0.1
Reads per run	~100k	~500k	up to 5M	1M	up to 10G	1.2 to 1.4G
Time per run	Vary	30 minutes to 6 hours	2 hours	24 hours	1 to 10 days,	1 to 2 weeks
Cost per 1 million bases (in US$)	$3	$2	$1	$10	$0.05 to $0.15	$0.13
Advantages	Longest read, ready to use	Longest read length. Fast.	Less expensive equipment. Fast.	Long read size. Fast.	high sequence yield, cost, accuracy	Low cost per base.
Disadvantages	Low yield, cost, errors and stability	Low yield, cost and errors	Errors	Price and errors.	Equipment is expensive. Some restriction for X	Slow, read length, longevity of the plateform

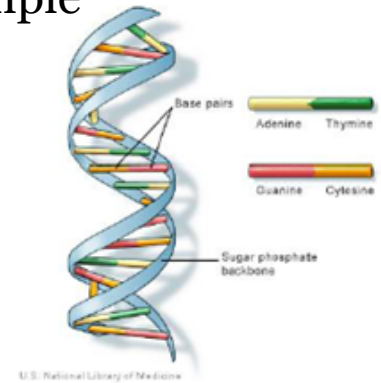
Presentation Overview

1. Next Generation sequence fundamentals
2. **NGS Data formats**
3. NGS Data Management

What data is available

DNA is extracted from a sample



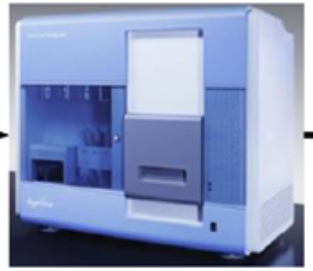


Assembly

DNA is fragmented into small pieces

.....

Fragments



@...../1
ATGC....
+....
UVWZ...
@...../2

The bases of a small fragment of DNA are sequentially identified from signals emitted as each fragment is re-synthesized from a DNA template strand

Data Formats – Read FASTQ format

```
@HWUSI-EAS100R:6:73:941:1973#0/1
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!''*((((***+))%%%++)(%%%%).1***-+*''))**55CCF>>>>>>CCCCCCC65
```

6 - Flowcell lane

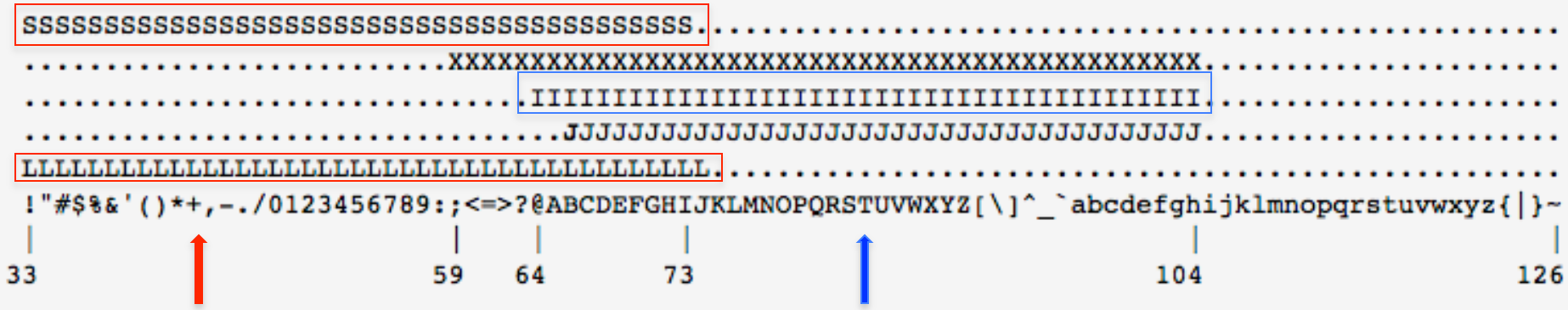
73 - Tile number

941,1973 - 'x','y'-coordinates of the cluster within the tile

#0 - index number for a multiplexed sample (0 for no indexing)

/1 - the member of a pair, /1 or /2 (paired-end or mate-pair reads only)

Data Formats – base quality encoding



S - Sanger Phred+33, raw reads typically (0, 40)
X - Solexa Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64, raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64, raw reads typically (3, 40)
with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
(Note: See discussion above).
L - Illumina 1.8+ Phred+33, raw reads typically (0, 41)

	CHAR	DEC	QUAL	
↑	+	43	10 (43-33)	Phred+33
↑	T	84	20 (84-64)	Phred+64

Data Formats - Sequence Alignment/Map (SAM) format

- Generic alignment format
- Supports short and long reads
- Supports different sequencing platforms
- Header section (optional)
- Alignment section

http://samtools.sourceforge.net/SAM1.pdf

Data Formats – SAM format

Header Section

- Header lines start with @
- @ is followed by TAG
- Header fields are TYPE:VALUE pairs

@RG **ID**:RUN_LANE **CN**:Institute **LB**:LibraryName
PL:Technology **PU**:RunName **SM**:Sample

Example:

@RG **ID**:61DP1AAXX_1 **CN**:AGRF **LB**:Rameses **PL**:ILLUMINA
PU: 61DP1AAXX.1 **SM**: HOLAUSM000A00009637

Data Formats – SAM format

Alignment section- 11 mandatory fields

```
HWI-HI83:6:1101:1210:1974#0/1 99 chr20 287833 30 10M1D25M = 287993 195\
ACCTATATCTTGGCCTTGGCCGATGCGGCCTTGCA\
?8?D?DDDDD8DDDE?E2:<A4CFC?CFB3A?F?C
```

1. QNAME: Query name of the read or the read pair
2. FLAG: Bitwise flag (pairing, strand, mate strand, etc.)
3. RNAME: Reference sequence name
4. POS: 1-Based leftmost position of clipped alignment
5. MAPQ: Mapping quality (Phred-scaled)
6. CIGAR: Extended CIGAR string (operations: MIDNSHP)
7. MRNM: Mate reference name ('=' if same as RNAME)
8. MPOS: 1-based leftmost mate position
9. ISIZE: Inferred insert size
10. SEQQuery: Sequence on the same strand as the reference
11. QUAL: Query quality (ASCII-33 = Phred base quality)

Bitwise flag http://ppotato.files.wordpress.com/2010/08/sam_output.pdf

Data Formats – SAM format

CIGAR operators

- M: match/mismatch
- I: insertion
- D: deletion
- S: softclip
- H: hardclip
- P: padding
- N: skip

```
RefPos:     1  2  3  4  5  6  7     8  9 10 11 12 13 14 15 16 17 18 19
Reference:  C  C  A  T  A  C  T     G  A  A  C  T  G  A  C  T  A  A  C
Read:                  A  C  T  A  G  A  A     T  G  G  C  T
```

```
POS: 5
CIGAR: 3M1I3M1D5M
```

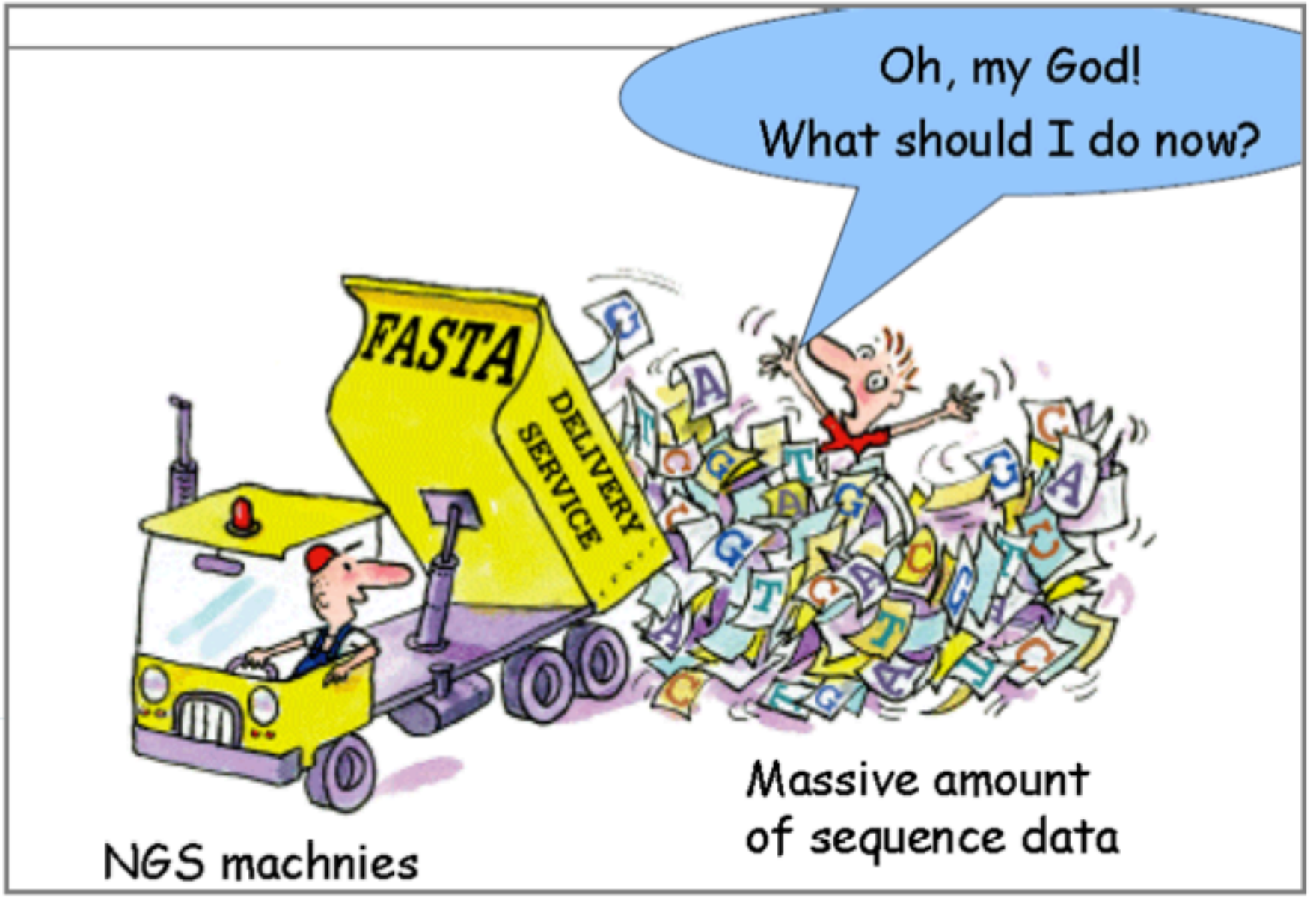
BAM

- Binary compressed version of SAM
- About 1/3 – 1/5 the storage requirements of SAM

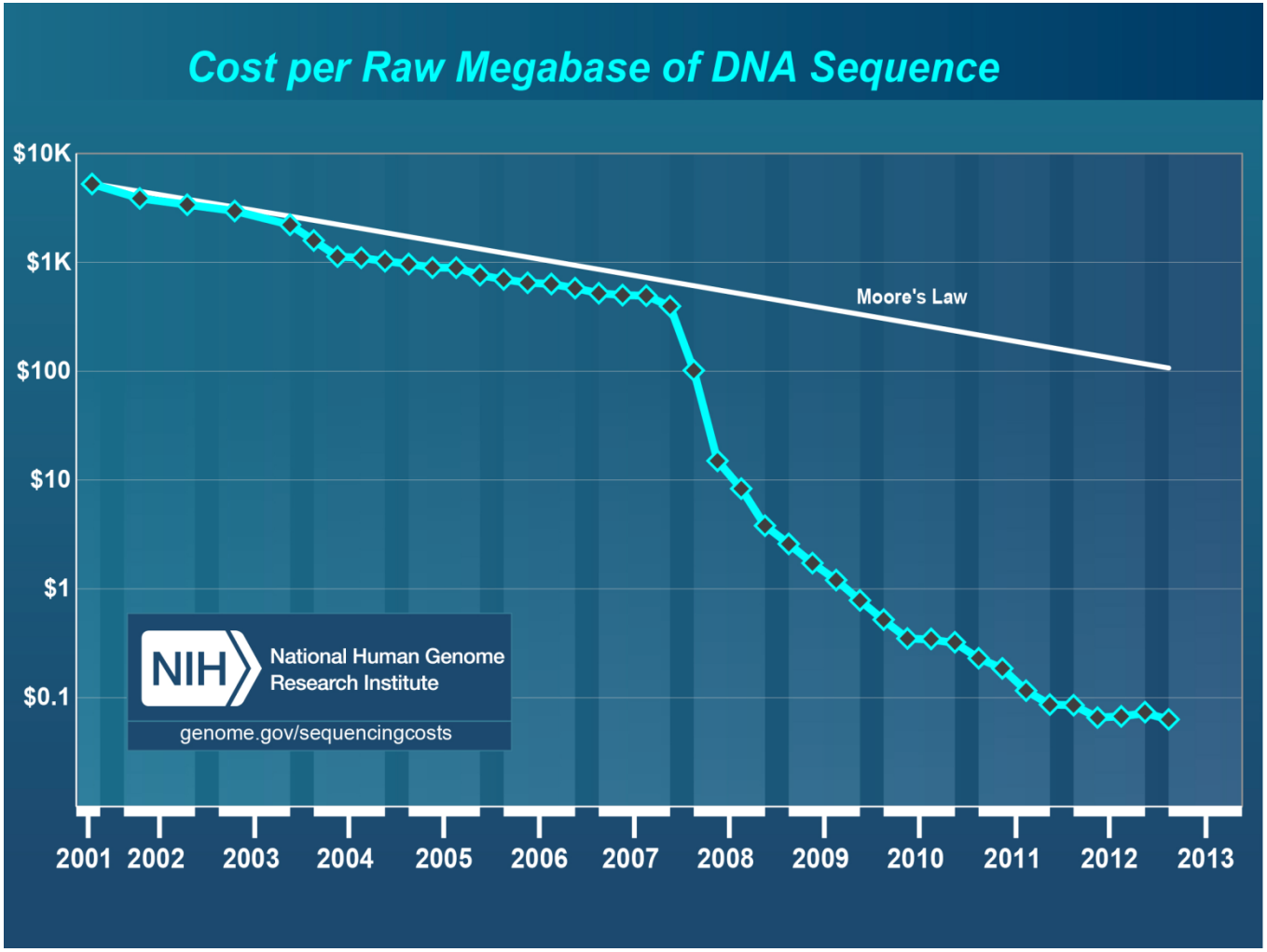
http://samtools.sourceforge.net/

Presentation Overview

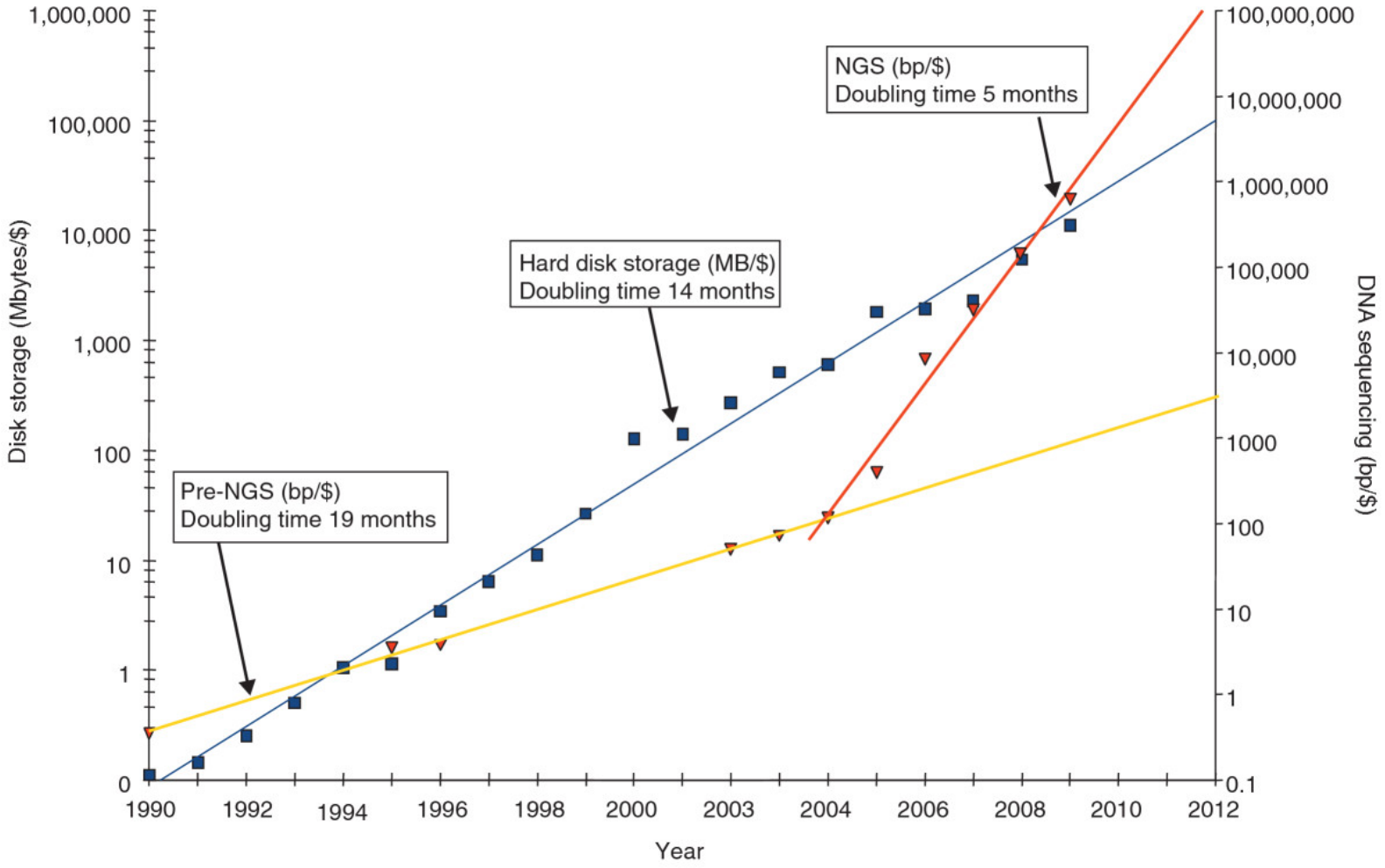
1. Next Generation sequence fundamentals
2. NGS Data formats
3. **NGS Data Management**


Oh, my God!
What should I do now?
FASTA
DELIVERY SERVICE
NGS machnies
Massive amount
of sequence data

BioInfo Moore’s law



What about the storage



Data Volume

- Raw data size could range in terabytes
- FASTQ files 10-100GB
- Mapped data in sam/bam formats

Data Complexity

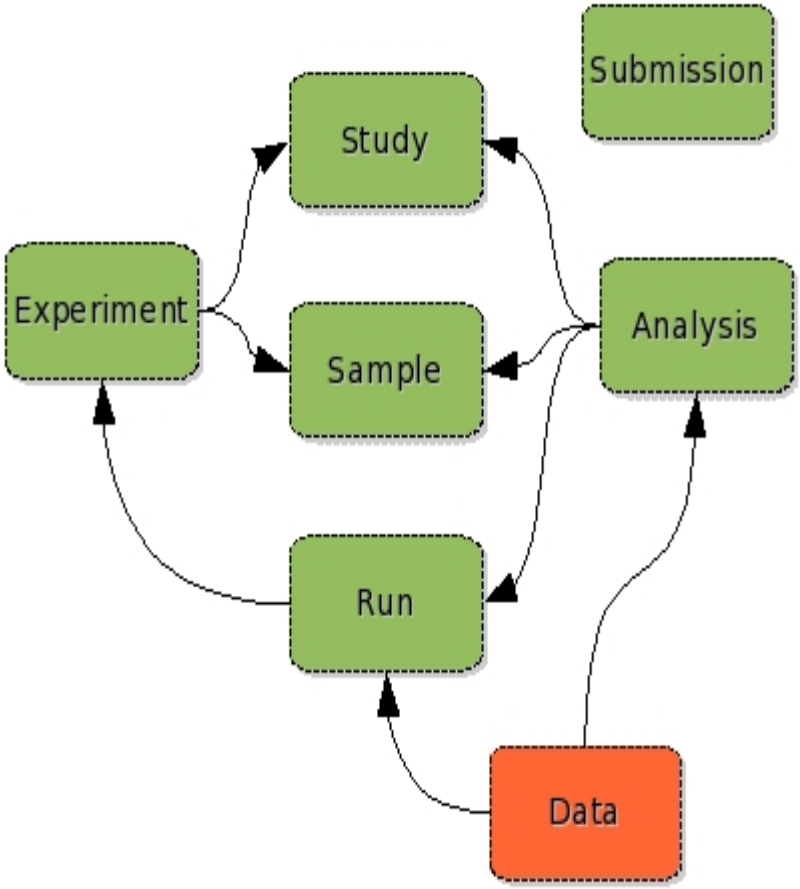
- Multiple samples
- Multiple runs
- Multiple platforms
- Metadata is important

Sequence archive/storage

- Publish data
- Query and download data
- Requirement for publication

- Sequence Read Archive SRA
  - Primary Sequence
  - Small Assemblies

- ArrayExpress
- Gene Expression Omnibus GEO
  - Gene expression



Possible computational infrastructure required for NGS data analysis

A local computing cluster

- Multiple nodes (servers) with multiple cores
- High performance storage
- Fast networks (10 Gb ethernet,infiniband)
- Enough space and conditions for the equipment ("servers room")
- Skilled people (sys admin, developers)

Cloud Computing

Pros

– Flexibility.

– You pay what you use.

– Don‘t need to maintain a data center.

Cons

– Transfer big datasets over internet is slow.

– You pay for consumed bandwidth. That is a problem with big datasets.

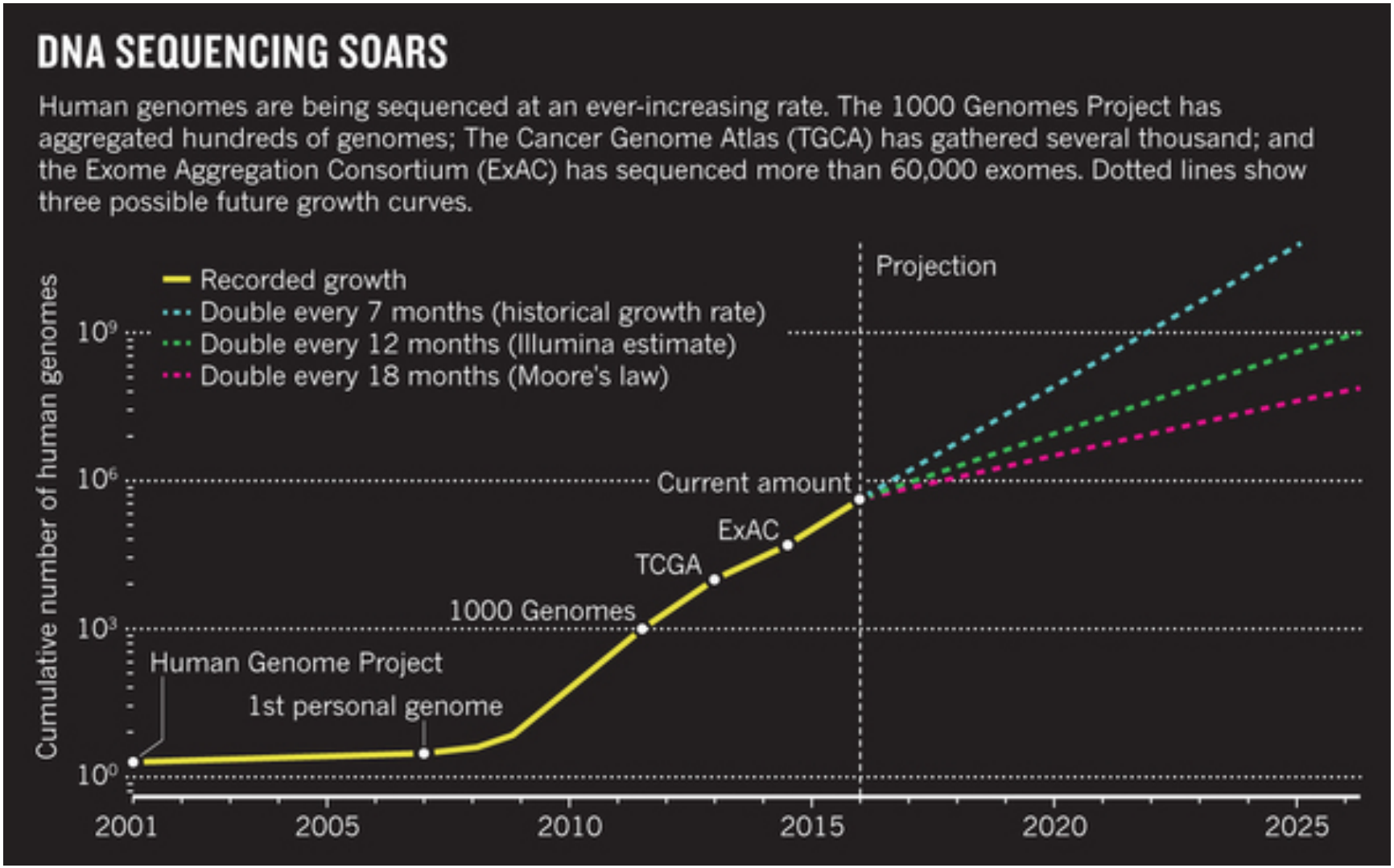
– Privacy/security concerns.

– More expensive for big and long term projects.

Conclusion

- Sequencing costs have dropped but the volume and complexity of data has increased
- Good data and analysis management is critical
- Standard data formats exist for NGS raw data

Future


DNA SEQUENCING SOARS
Human genomes are being sequenced at an ever-increasing rate. The 1000 Genomes Project has aggregated hundreds of genomes; The Cancer Genome Atlas (TGCA) has gathered several thousand; and the Exome Aggregation Consortium (ExAC) has sequenced more than 60,000 exomes. Dotted lines show three possible future growth curves.
Recorded growth
Double every 7 months (historical growth rate)
Double every 12 months (Illumina estimate)
Double every 18 months (Moore's law)
Projection
Cumulative number of human genomes
10^9
10^6
10^3
10^0
Current amount
ExAC
TCGA
1000 Genomes
Human Genome Project
1st personal genome
2001
2005
2010
2015
2020
2025