701-1425-00L - Genetic Diversity: Analysis

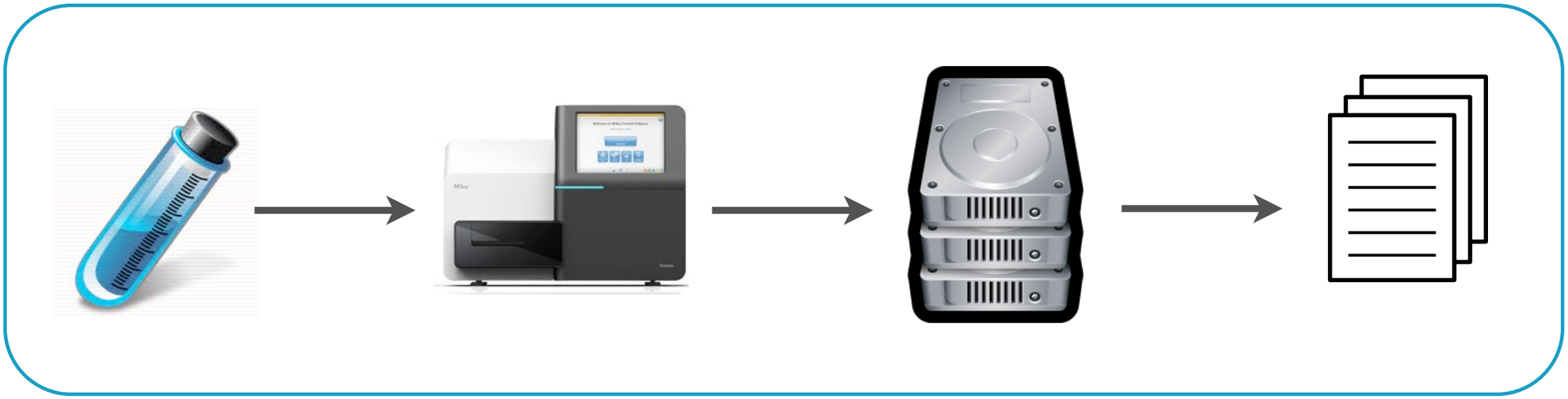
NGS: Introduction

Tuesday, June 18, 2019

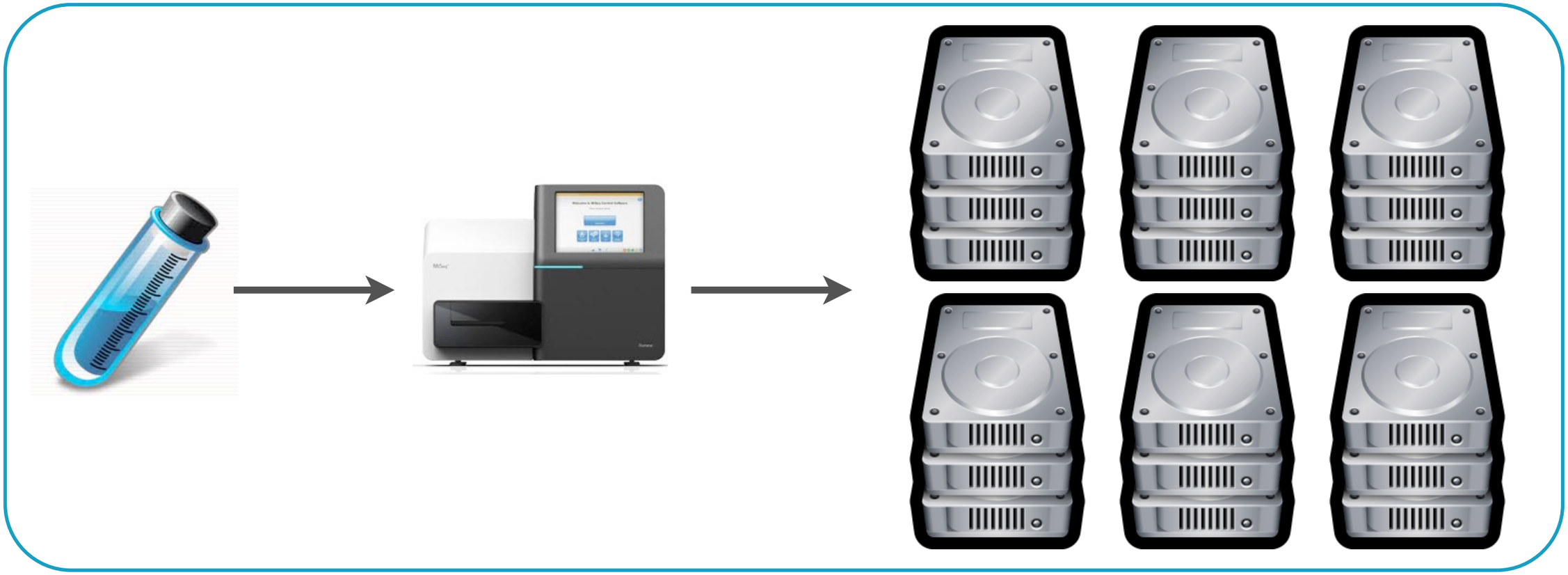
Jean-Claude Walser

jean-claude.walser@env.ethz.ch

Next (Next) Generation Sequencing **Hype**



Next (Next) Generation Sequencing **Reality**

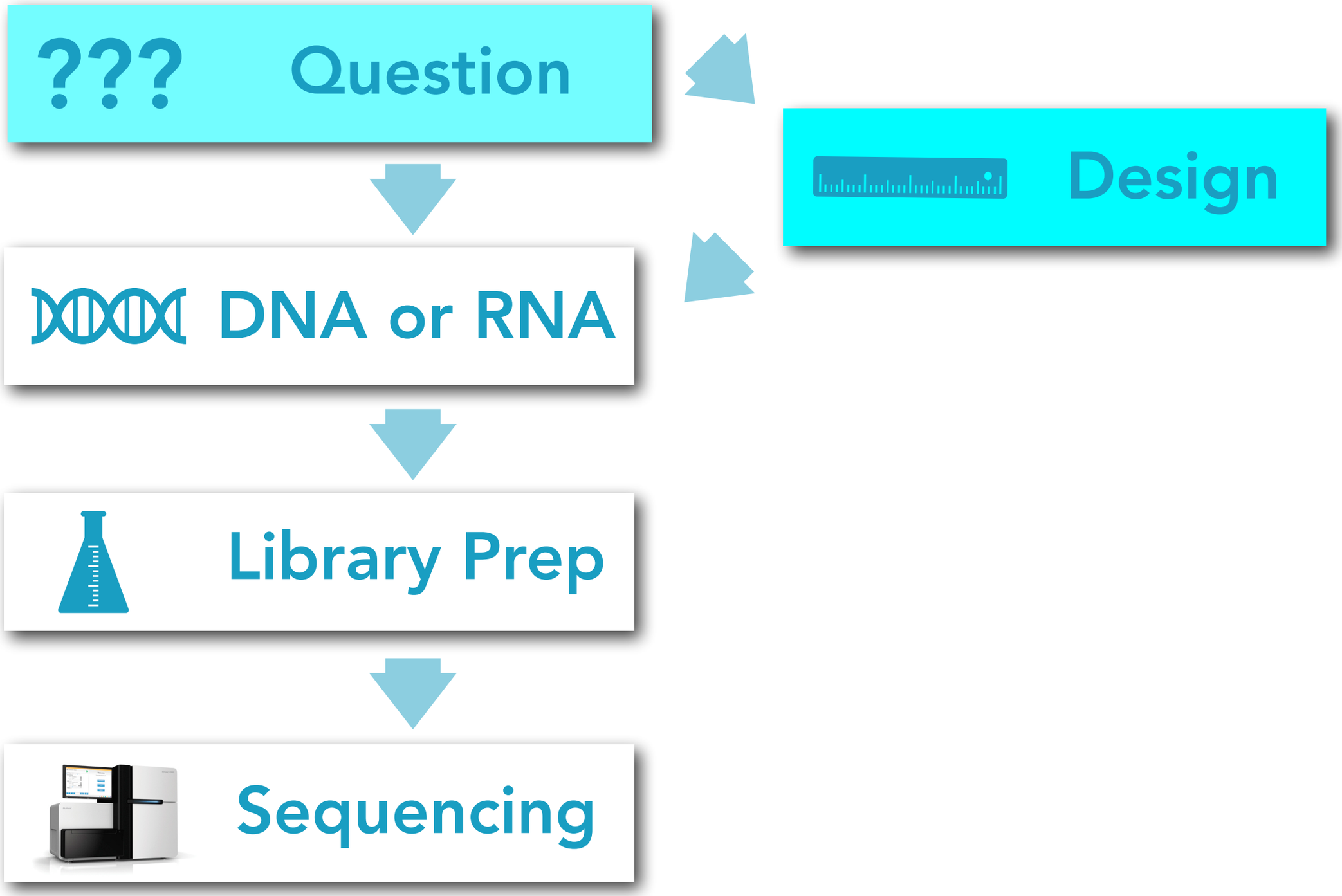


The **First Law of Technology** says we invariably **overestimate** the **short-term impact** of a truly transformational discovery, while **underestimating** its **longer-term effects**.

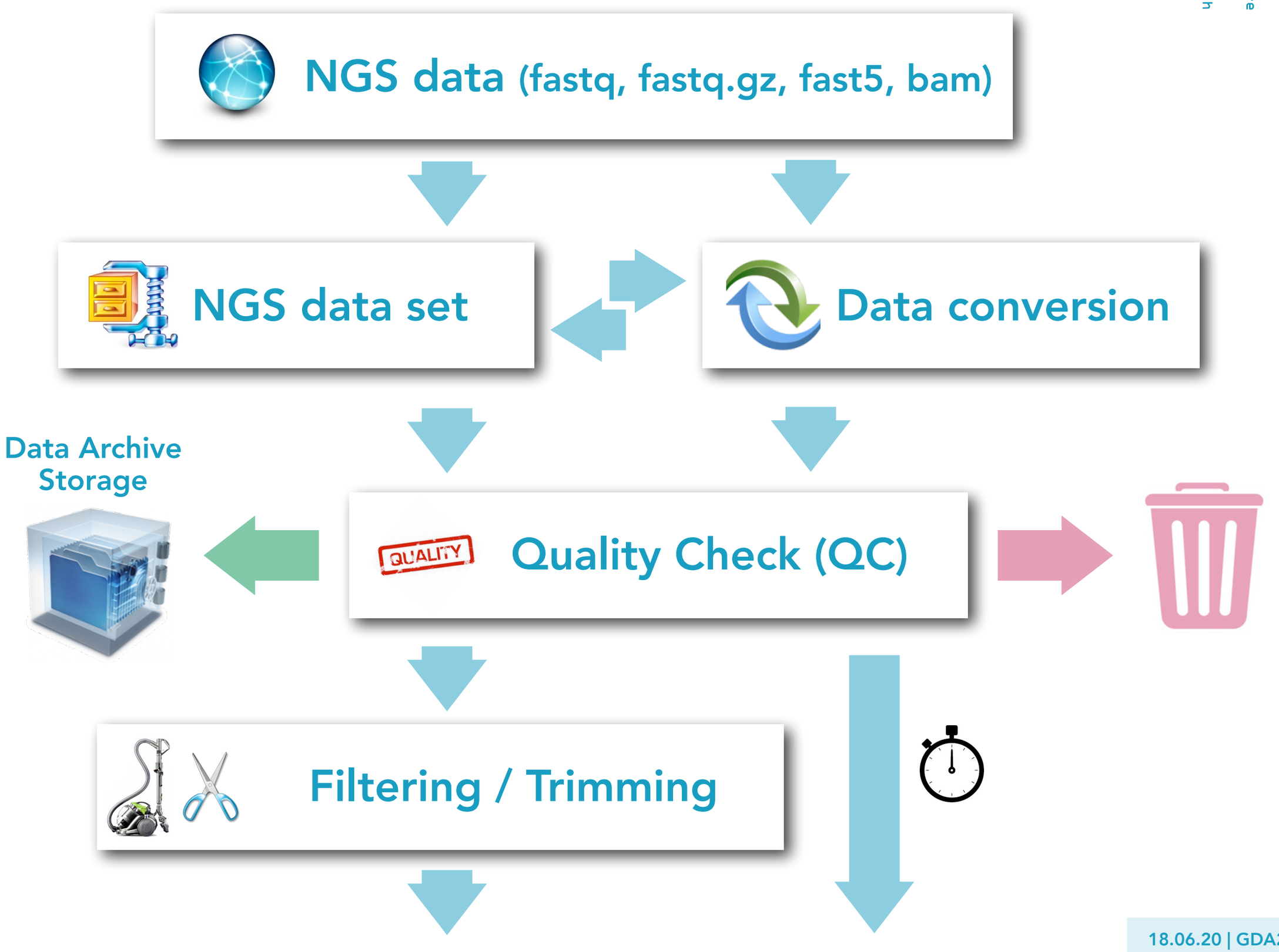
https://www.scientificamerican.com/

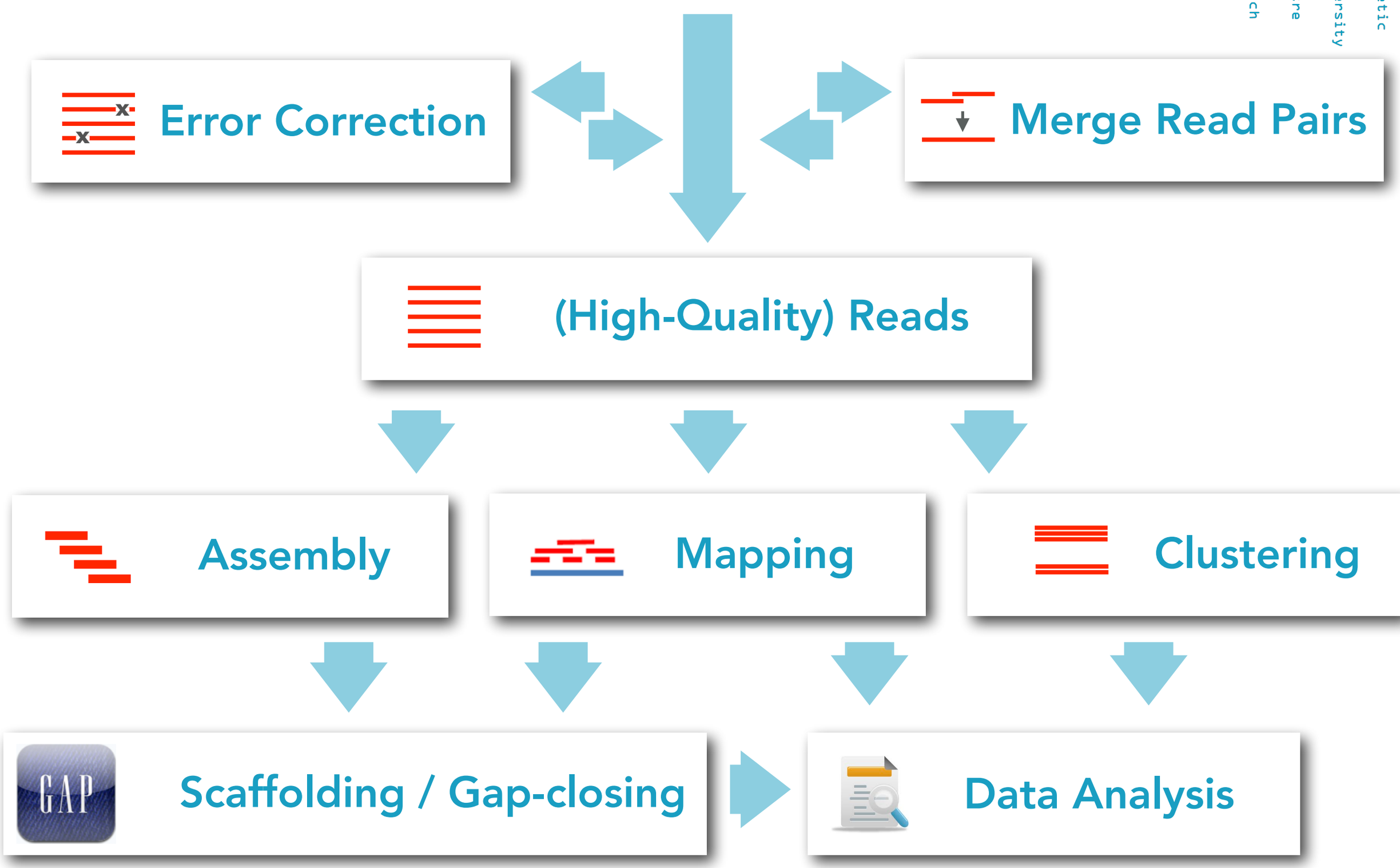
[A-]TYPICAL WORKFLOW

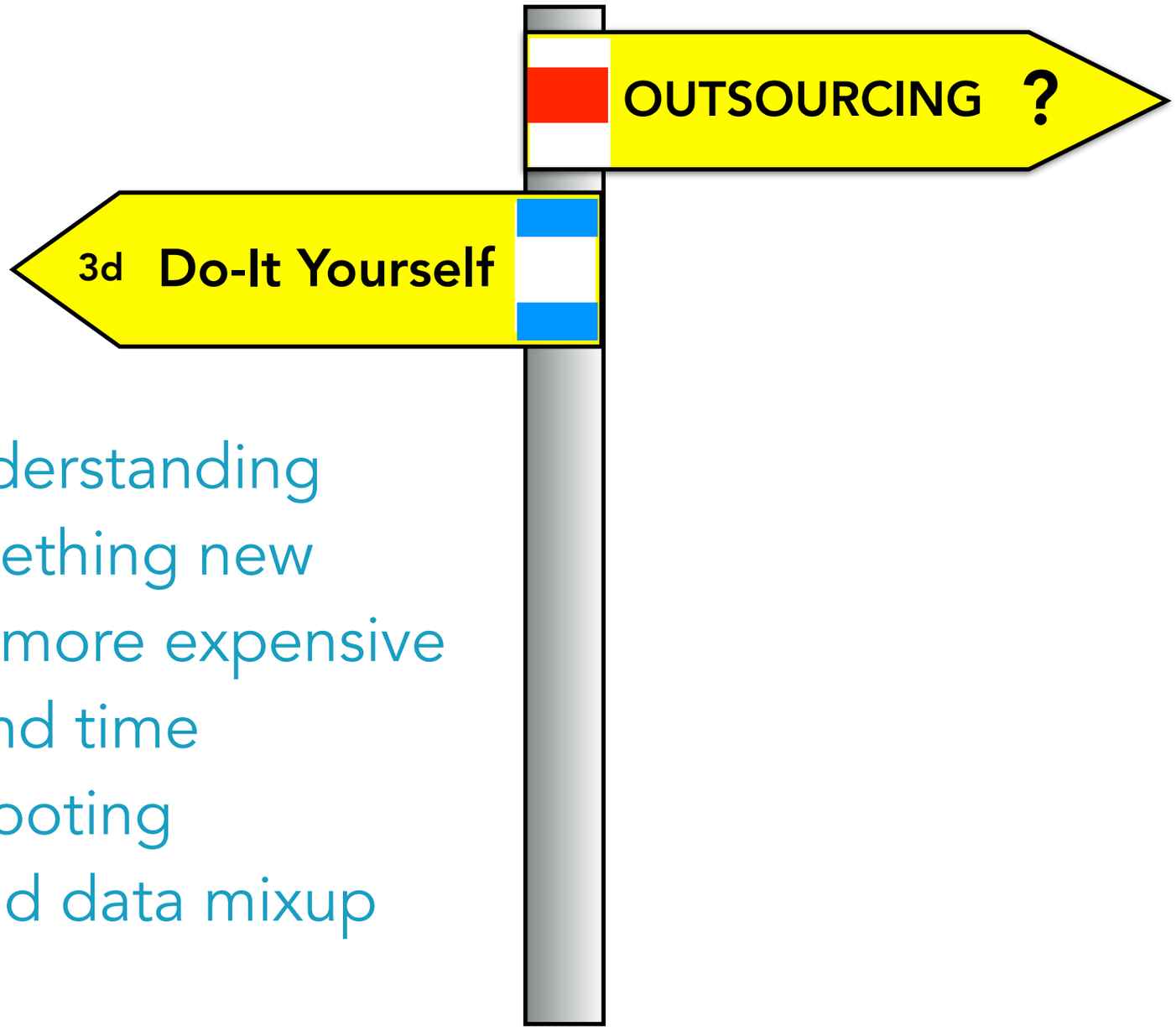
GDC Genetic Diversity Centre Zurich



GDC Genetic Diversity Centre Zurich

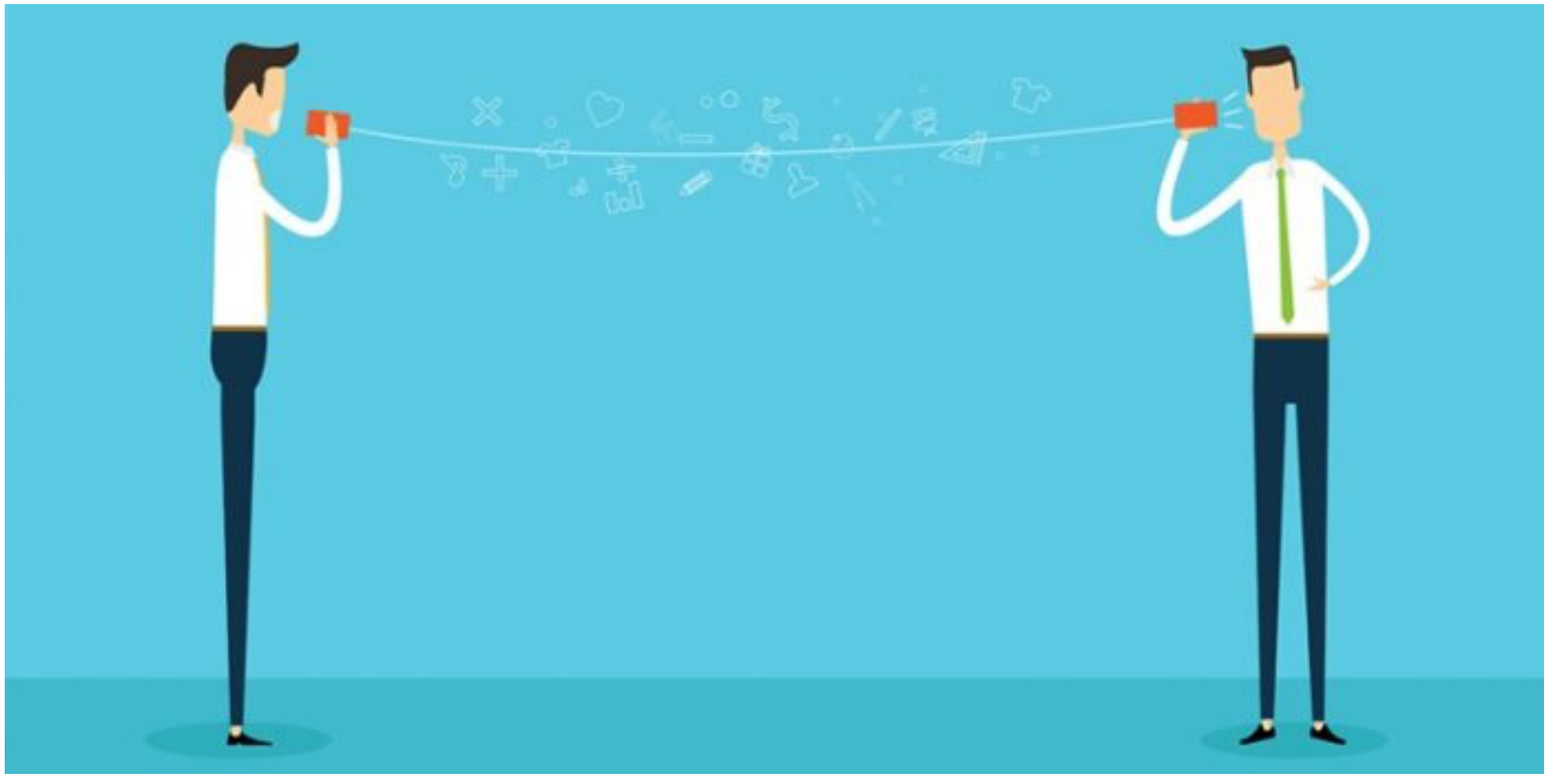


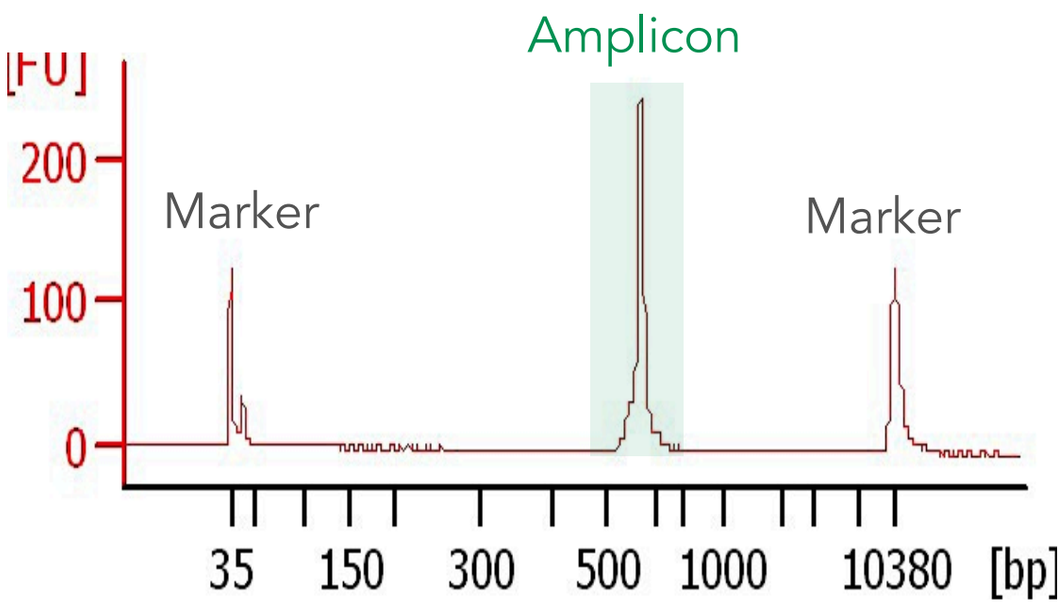

Error Correction
Merge Read Pairs
(High-Quality) Reads
Assembly
Mapping
Clustering
GAP
Scaffolding / Gap-closing
Data Analysis

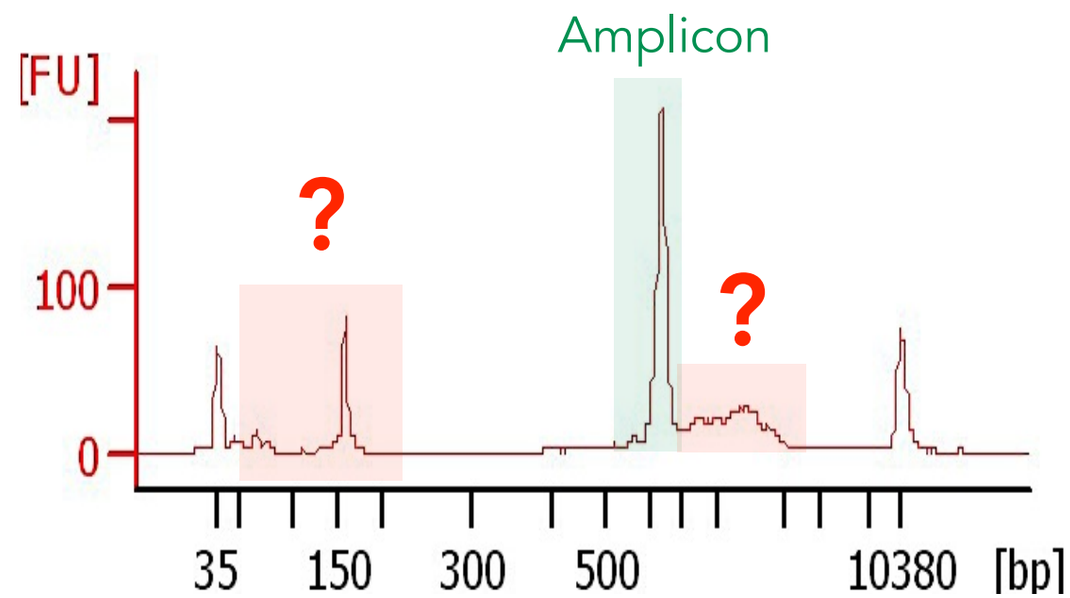


- better understanding
- learn something new
- often not more expensive
- turn-around time
- troubleshooting
- understand data mixup

NGS ▷ Introduction

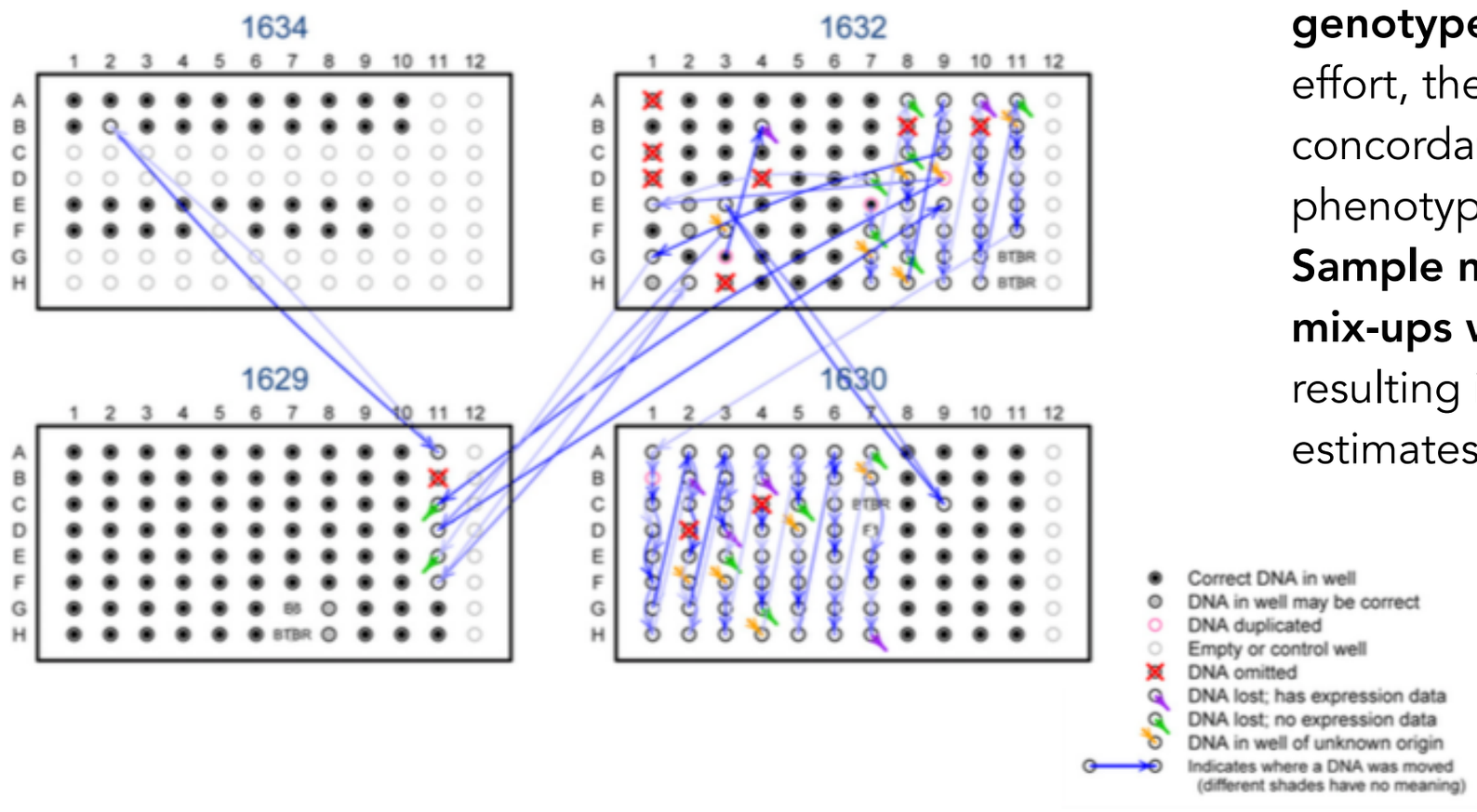



Amplicon
Marker
Marker
[FU]
200
100
0
35
150
300
500
1000
10380
[bp]


Amplicon
?
?
[FU]
100
0
35
150
300
500
10380
[bp]

Identification and Correction of Sample Mix-Ups in Expression Genetic Data: A Case Study

Karl W. Broman,*,2 Mark P. Keller,† Aimee Teo Broman,* Christina Kendziorski,* Brian S. Yandell,‡,§ Śaunak Sen,**,1 and Alan D. Attie†
*Department of Biostatistics and Medical Informatics, †Department of Biochemistry, ‡Department of Statistics, and §Department of Horticulture, University of Wisconsin, Madison, Wisconsin 53706, and **Department of Epidemiology and Biostatistics, University of California, San Francisco, California 94107



"To map the genetic loci influencing a complex phenotype, one seeks to establish an **association between genotype and phenotype**. In such an effort, the maintenance of the concordance between genotyped and phenotyped samples and data is critical. **Sample mislabeling and other sample mix-ups will weaken associations**, resulting in reduced power and biased estimates of locus effects."

SEQUENCING TECHNOLOGIES

… the all-in-one NGS platform does not exist (yet)!

NGS ▷ Introduction

Sanger (chain termination)

Roche 454 Pyrosequencing (pyrophosphate)

Ion Torrent (semiconductor technology)

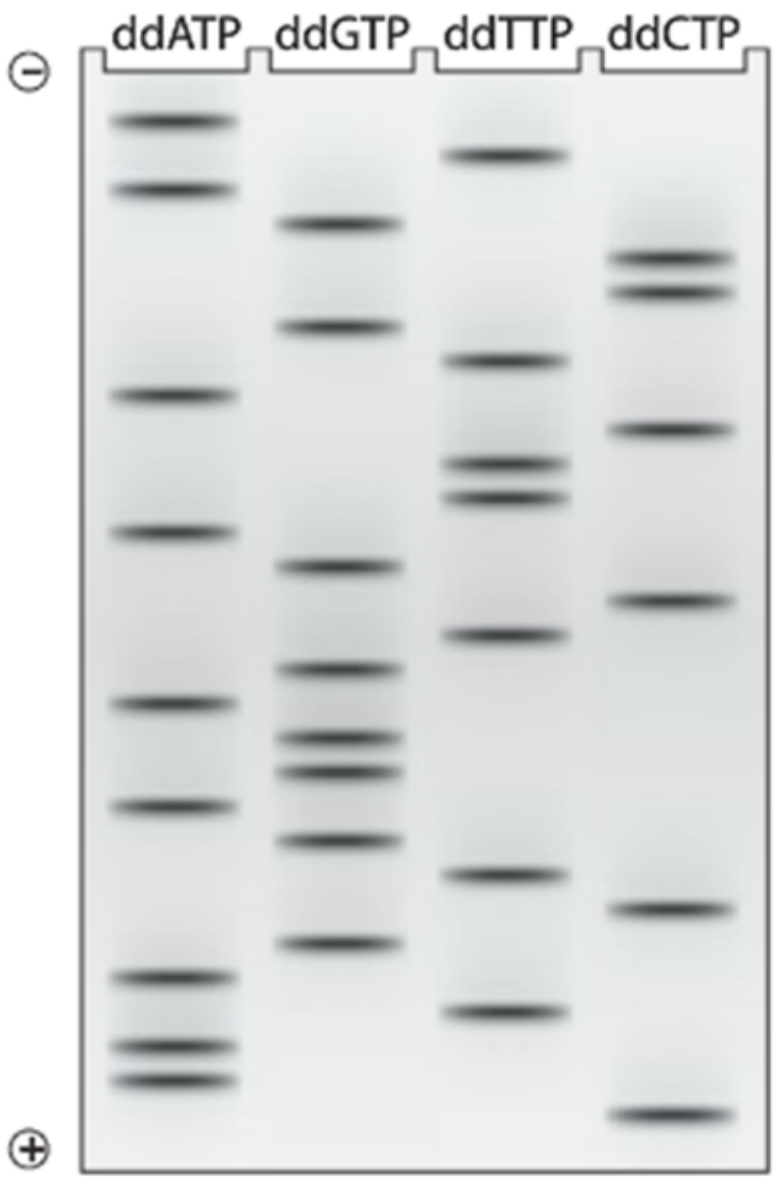
Illumina Sequencing by Synthesis (fluorescent)

PacBio (fluorophore)

Nanopore (ionic current)

Helicos - SeqLL (fluorescent)

Bionano - Saphyr (third-generation optical mapping)



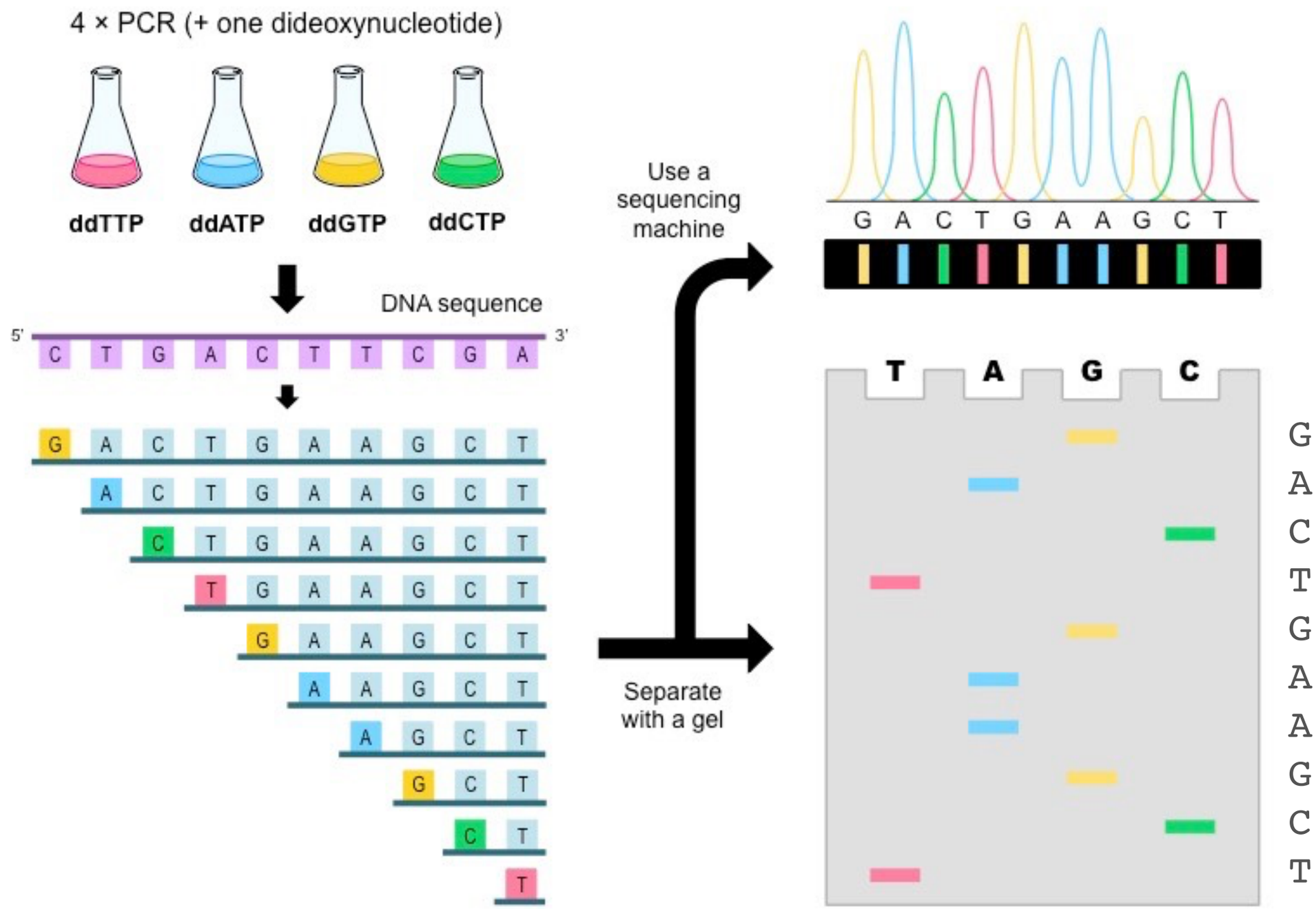
The Nobel Prize in Chemistry 1980
Paul Berg, Walter Gilbert, Frederick Sanger


Paul Berg


Walter Gilbert


Frederick Sanger

The Nobel Prize in Chemistry 1980 was divided, one half awarded to Paul Berg *"for his fundamental studies of the biochemistry of nucleic acids, with particular regard to recombinant-DNA"*, the other half jointly to Walter Gilbert and Frederick Sanger *"for their contributions concerning the determination of base sequences in nucleic acids"*.


4 × PCR (+ one dideoxynucleotide)
ddTTP
ddATP
ddGTP
ddCTP
DNA sequence
5'
3'
C T G A C T T C G A
G A C T G A A G C T
A C T G A A G C T
C T G A A G C T
T G A A G C T
G A A G C T
A A G C T
A G C T
G C T
C T
T
Use a sequencing machine
Separate with a gel
G A C T G A A G C T
T A G C
G
A
C
T
G
A
A
G
C
T

MiniSeq System

1.8-7.5 Gb
8-25 million
2 x 150 bp
50

MiSeq Series

0.3-15 Gb
1-25 million
2 x 300 bp
384

NextSeq Series

20-120 Gb
130-400 million
2 x 150 bp
96

HiSeq Series

125-1500 Gb
2.5-5 billion
2 x 150 bp
12

HiSeq X Series

900-1800 Gb
3-6 billion
2 x 150 bp
16

NovaSeq Series

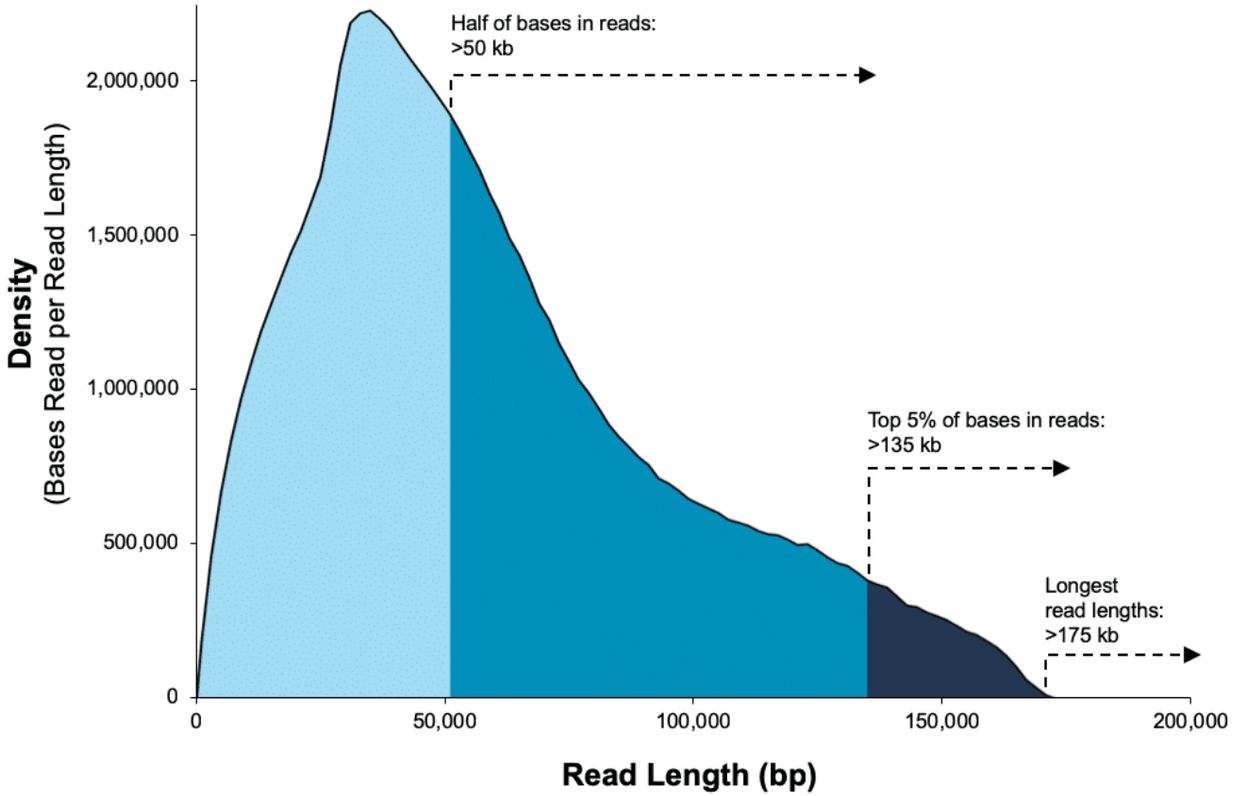
134-6000 Gb
Up to 20 billion
2 x 150 bp
48

http://www.illumina.com

PACBIO® (Pacific Biosciences)

https://www.pacb.com

Sequel



Data from a 35 kb size-selected *E. coli* library using the SMRTbell Express Template Prep Kit 2.0 on a Sequel II System (1.0 Chemistry, Sequel II System Software v7.0, 15-hour movie)*.

Oxford NANOPORE Technologies

https://www.nanoporetech.com

NGS ▷ Introduction

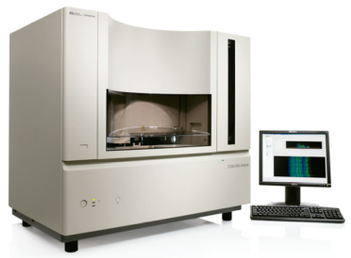

The Economist
Open-source medical devices
Dealing with nuclear waste
Burt Rutan, a maverick in flight
Technology Quarterly
December 1st 2012
The dream of the medical tricorder
the hand held devices inspiring medical add-ons for smartphones


Running
Stop Sequencing

MinION Mk1C

NGS ▷ Introduction

- Research question
- Budged (including storage and anaylsis)
- Read / sequence length
- Number of reads / coverage
- Possible contaminants
- Quality and quantity of template
- Number of samples
- Availability



Few but good sequences

High coverage or many samples but shorter reads

Longer sequences but not so many samples

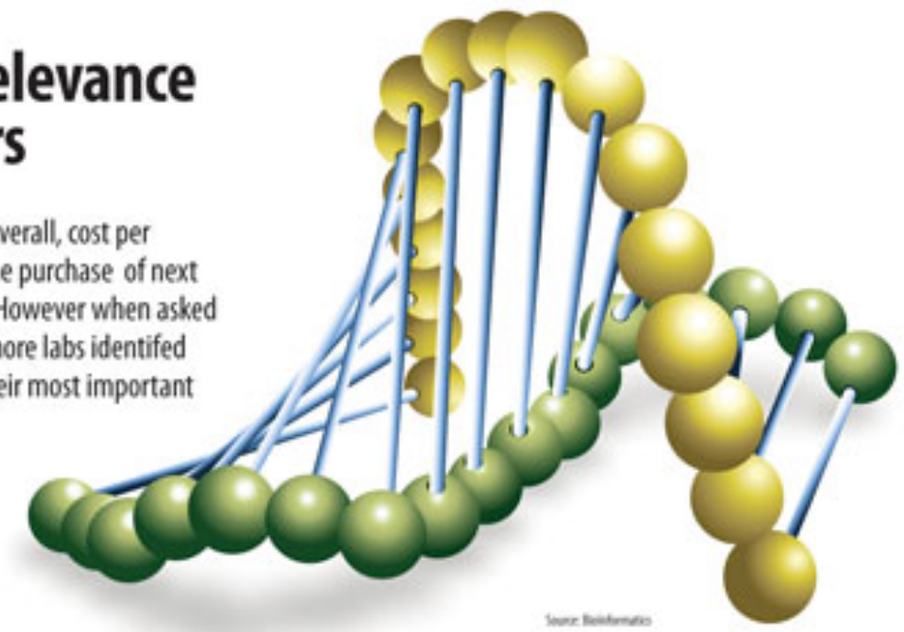
Fast results, long reads but higher error rate

Structural variants and smaller sample size

GEN Genetic Engineering & Biotechnology News

NGS: Cost and Relevance Are Key to Buyers

Recent market research shows that, overall, cost per base was the most cited concern in the purchase of next generation sequencing instruments. However when asked to identify their top three concerns, more labs identifed "Appropriate to My Application" as their most important criteria.



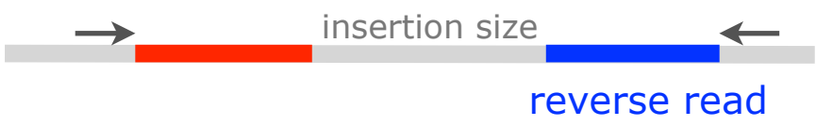
The 10 Most Critical Platform Attributes as Defined by Purchasers

1. Cost per base 43%
2. Sequencing data quality 34%
3. Appropriate for my application 32%
4. Reproducibility/accuracy 31%
5. Amount of DNA/RNA needed per experiment 25%
6. Read length 24%
7. Instrument cost 18%
8. Number of reads 17%
9. Available software analysis tools 16%
10. Instrument reliability 16%

SEQUENCING DATA

Illumina Sequence Read Data

Single Reads (SR)



Paired-End (PE) Reads

Overlapping Paired-End (PE) Reads

Single Reads (SR) with Index

Paired-End (PE) Reads with Index

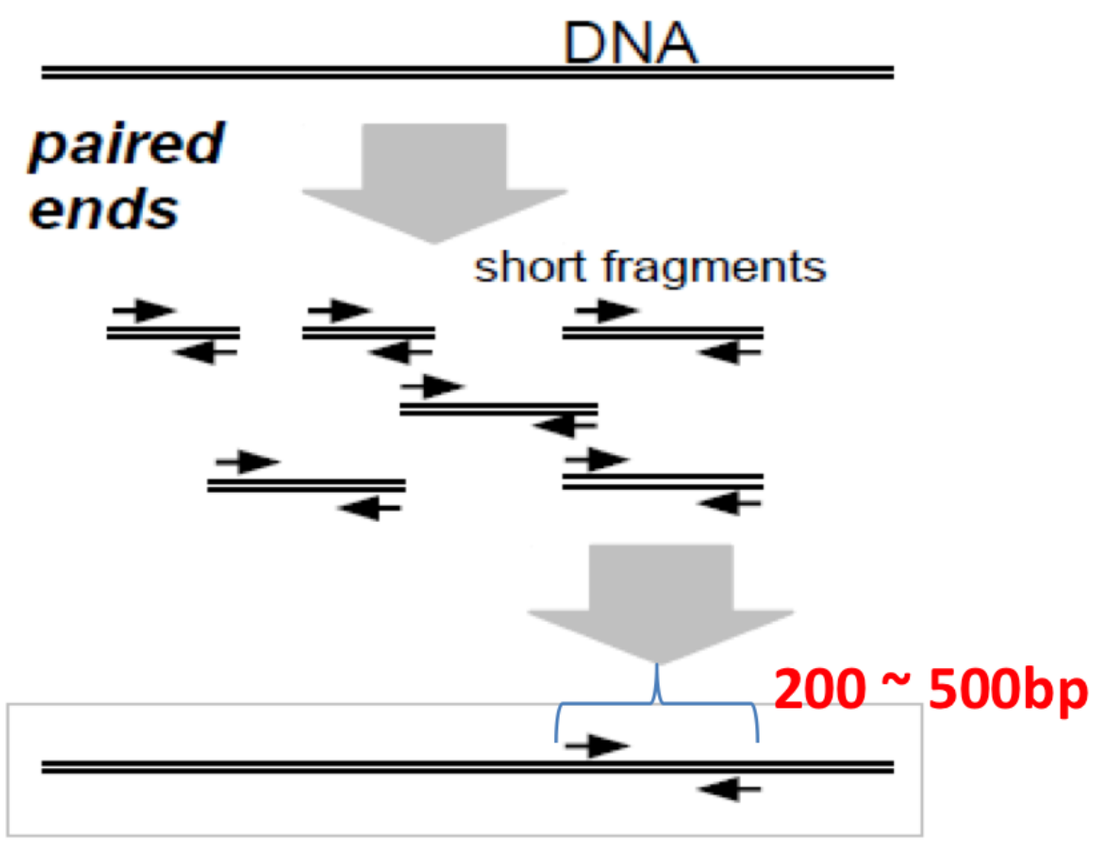
Paired-End (PE) Reads with Dual Indexing

Extended Single Reads (SR) with Index

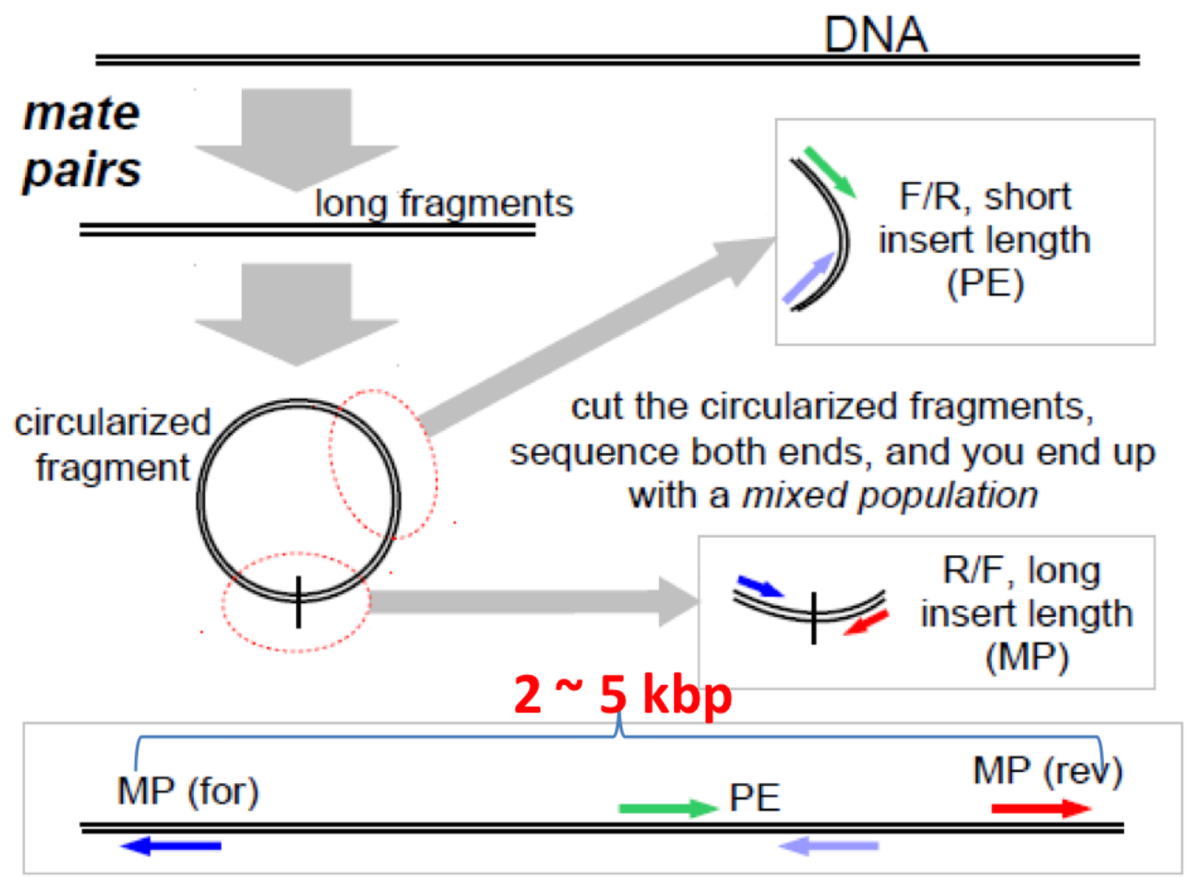
GDC Zurich Centre Diversity Genetic

Illumina Sequence Read Data

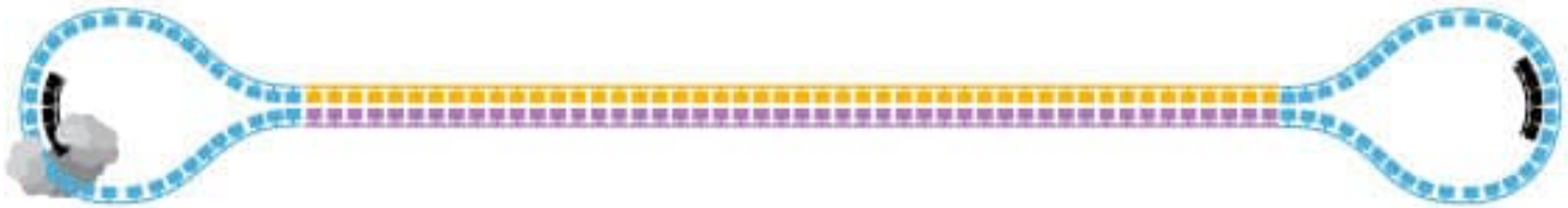
paired-end (PE)



mate-pair (MP)



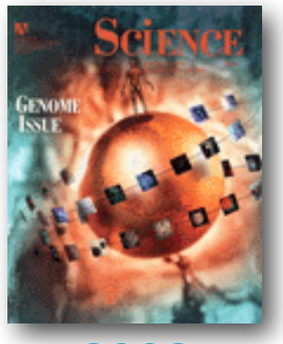
PacBio SMRTbell Library

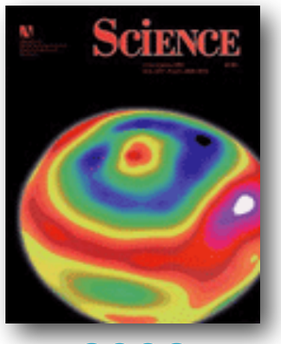


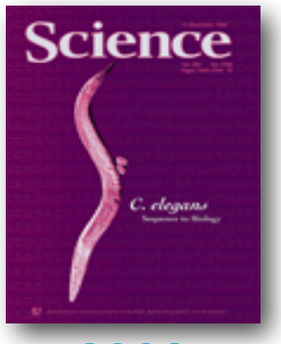
NGS ▷ Introduction

GDC
Zurich Centre Diversity Genetic

1995

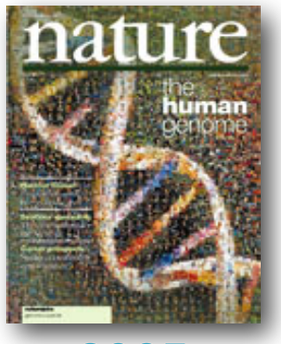

1996


1997


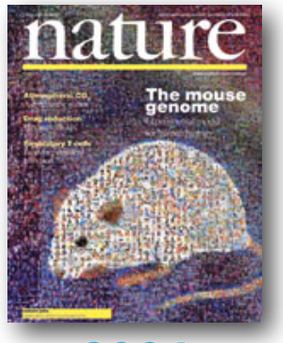
1998


2000


2000


2001


2002


2002


2002

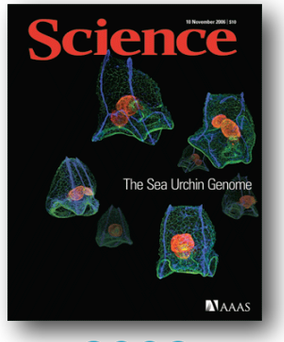

2002


2004


2004


2005


2005


2006


2006


2007


2007


2008


2008


2009


2009


2010


2011

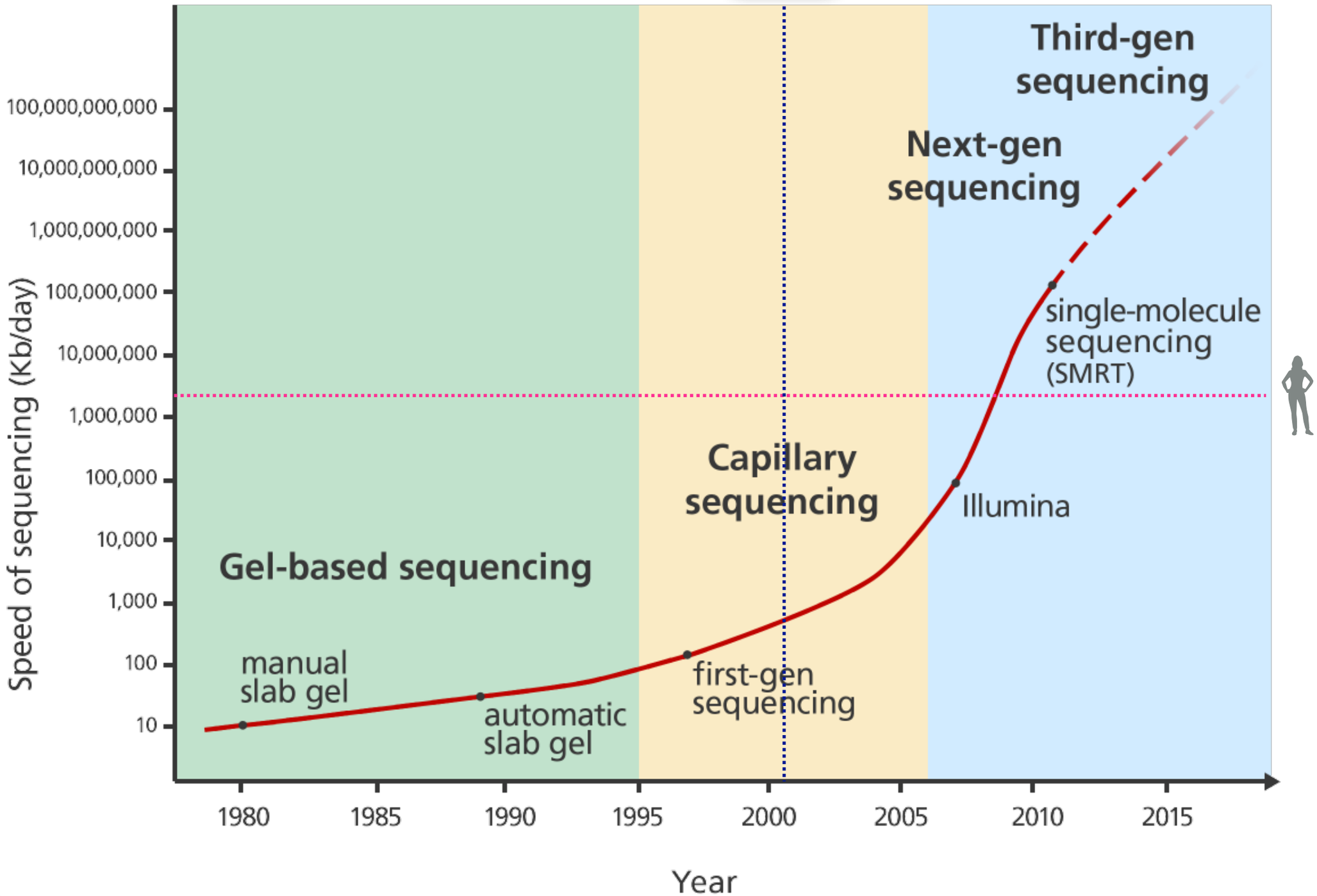

2011

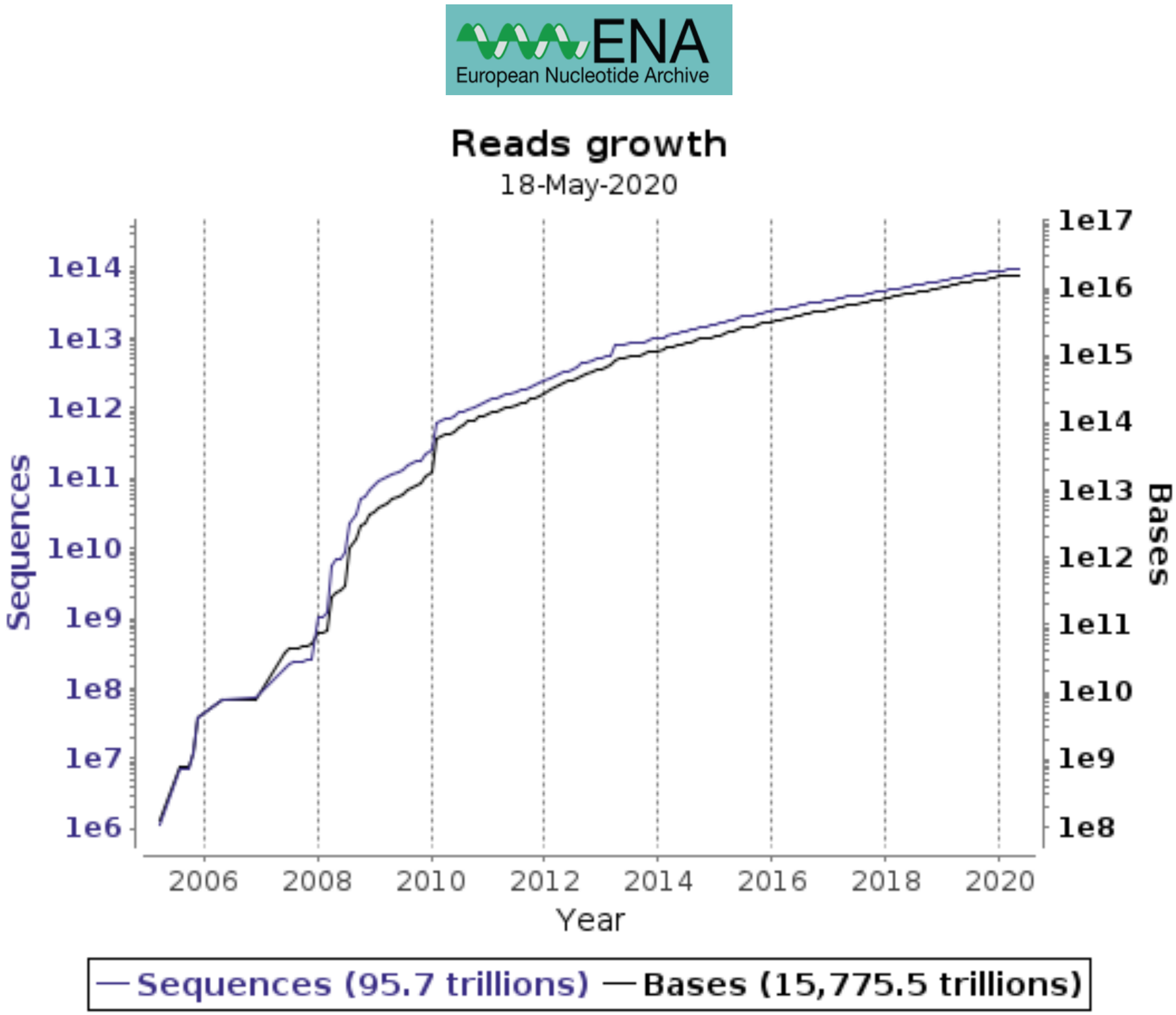

2011

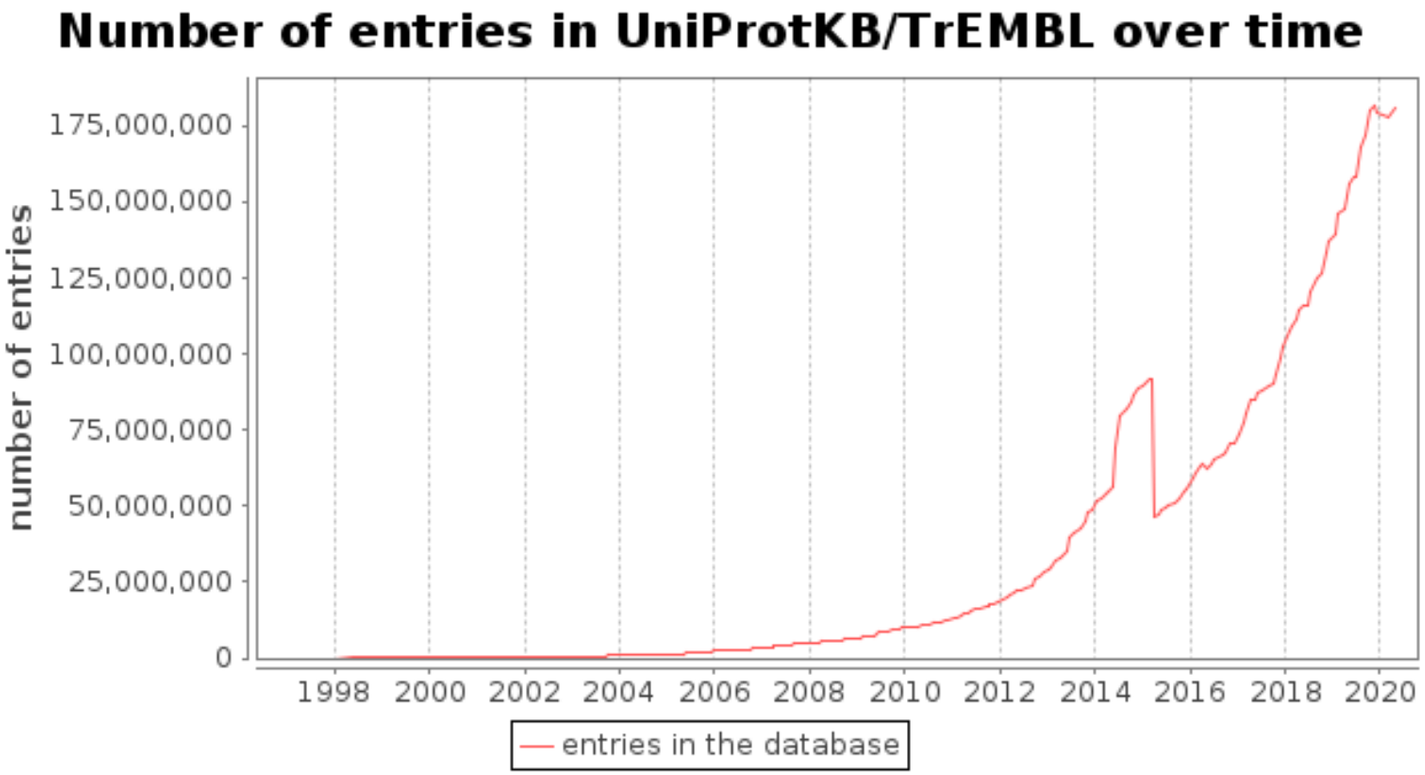

2012

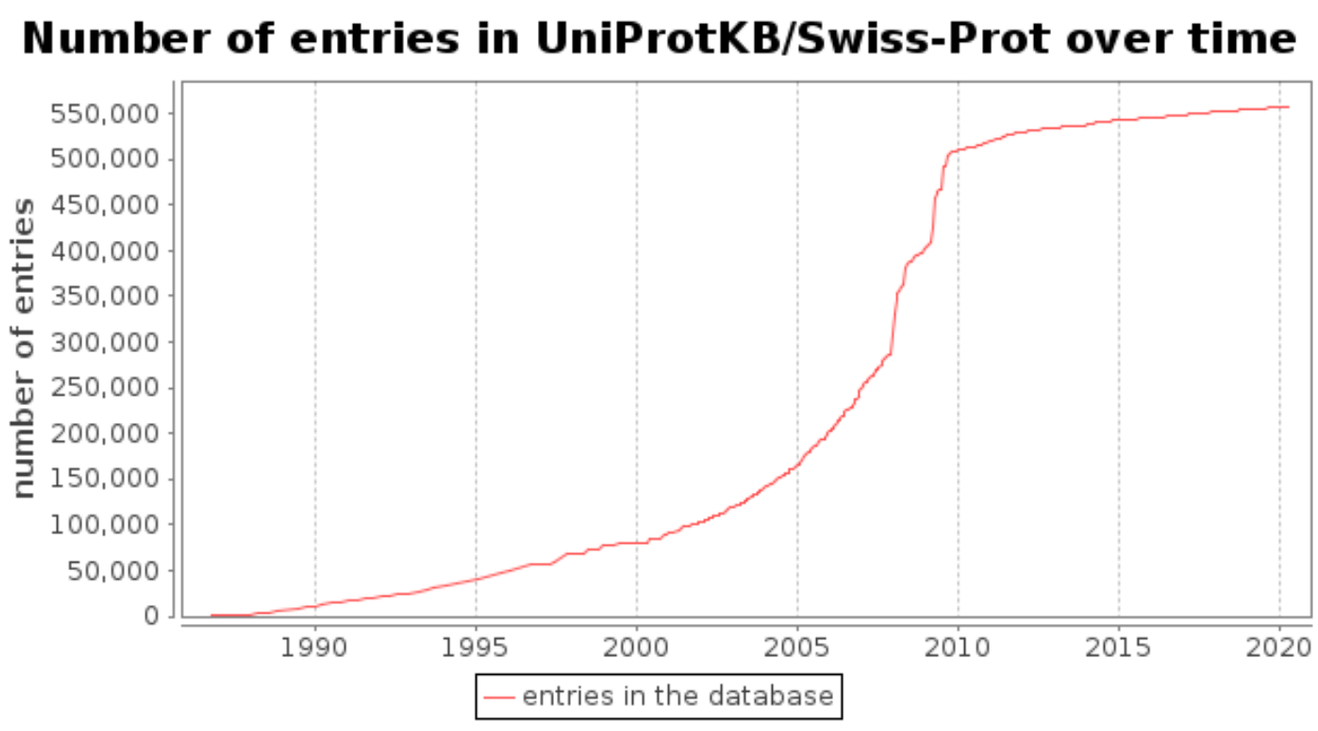

...


nature
the human genome


Speed of sequencing (Kb/day)
100,000,000,000
10,000,000,000
1,000,000,000
100,000,000
10,000,000
1,000,000
100,000
10,000
1,000
100
10
1980
1985
1990
1995
2000
2005
2010
2015
Year
Gel-based sequencing
Capillary sequencing
Next-gen sequencing
Third-gen sequencing
manual slab gel
automatic slab gel
first-gen sequencing
Illumina
single-molecule sequencing (SMRT)


ENA
European Nucleotide Archive
Reads growth
18-May-2020
Sequences
1e14
1e13
1e12
1e11
1e10
1e9
1e8
1e7
1e6
Bases
1e17
1e16
1e15
1e14
1e13
1e12
1e11
1e10
1e9
1e8
2006
2008
2010
2012
2014
2016
2018
2020
Year
Sequences (95.7 trillions)
Bases (15,775.5 trillions)



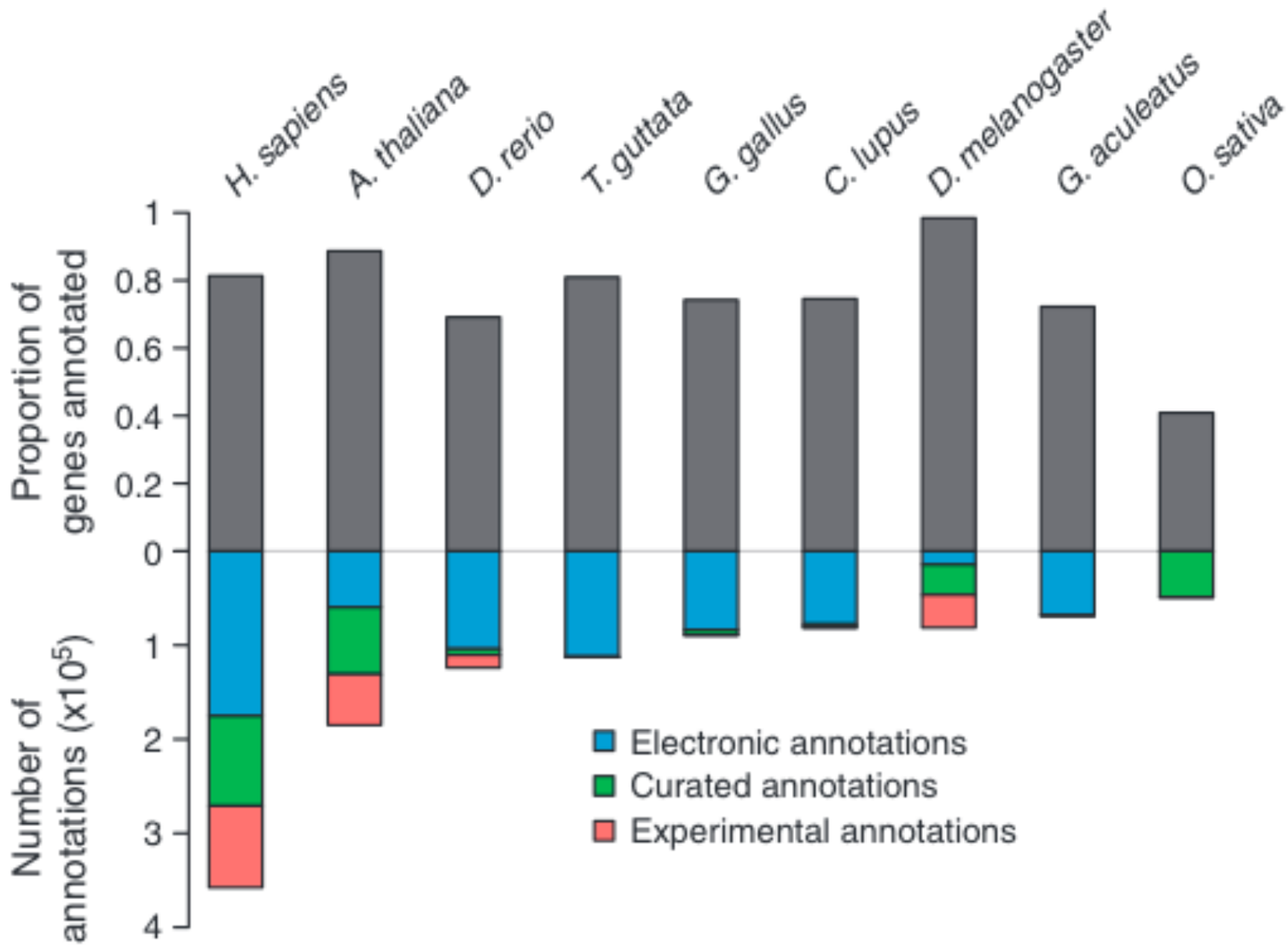


The **UniProt Knowledgebase** (**UniProtKB**) is the central hub for the collection of functional information on proteins, with accurate, consistent and rich annotation.

UniProtKB consists of two sections:

Reviewed (Swiss-Prot) - Manually annotated
Records with information extracted from literature and curator-evaluated computational analysis.

Unreviewed (TrEMBL) - Computationally analyzed
Records that await full manual annotation.



The proportion of annotated genes and their types of annotations for nine sequenced genomes (as of February 2013). Humans (*Homo sapiens*) and *Arabidopsis thaliana* have the highest number of annotations for animals and plants, respectively. They also have the most experimentally derived annotations. Most other species, except *Drosophila melanogaster*, are annotated mostly electronically.

Primmer et al. (2013) Mol Ecol

DATA SUBMISSION

ENA European Nucleotide Archive

The European Nucleotide Archive (ENA) captures and presents information relating to experimental workflows that are based around nucleotide sequencing. A typical workflow includes the isolation and preparation of material for sequencing, a run of a sequencing machine in which sequencing data are produced and a subsequent bioinformatic analysis pipeline. ENA records this information in a data model that covers input information (sample, experimental setup, machine configuration), output machine data (sequence traces, reads and quality scores) and interpreted information (assembly, mapping, functional annotation).

NCBI National Center for Biotechnology Information

Sequence Read Archive (SRA) makes biological sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets. The SRA stores raw sequencing data and alignment information from high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD System®, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®.

Data Deposition Example from the Literature:

Mushegian *et al.* (2018) **Environmental sources of bacteria and genetic variation in behavior influence host-associated microbiota**. AEM doi:10.1128/AEM.01547-18.

Sequence data are deposited in the **European Nucleotide Archive of the EBI under accession number PRJEB30308** (http://www.ebi.ac.uk/ena/data/view/PRJEB30308). Data tables, OTUs sequences and code used for analysis can be found on **Github** at https://github.com/amusheg/Daphnia-microbiota-behavior and will be deposited in **Dryad** upon publication.

EMBL-EBI | Services | Research | Training | About us

ENA European Nucleotide Archive

Search
Examples: BN000065, histone
Advanced
Sequence

Home | Search & Browse | Submit & Update | Software | About ENA | Support

Contact Helpdesk

Study: PRJEB30308

Microbiota associated with Daphnia exhibiting genetic variation in behavior

View: Project XML Study XML

Download: Project XML Study XML

Name
Microbiota of browsing Daphnia

Submitting Centre
Universitaet Basel

Secondary accession(s)
ERP112744

Description
In many organisms, host-associated microbial communities are acquired horizontally after birth. This process is believed to be shaped by a combination of environmental and host genetic factors. We examined whether genetic variation in animal behavior could affect the composition of the animal's microbiota in different environments. The freshwater crustacean Daphnia magna is primarily planktonic, but exhibits variation in the degree to which it browses in benthic sediments. We performed an experiment with clonal lines of D. magna showing different levels of sediment-browsing intensity exposed to either bacteria-rich or bacteria-poor sediment or whose access to sediments was prevented. We find that the bacterial composition of the environment and genotype-specific browsing intensity together influence the composition of the Daphnia-associated bacterial community. Exposure to more diverse bacteria did not lead to a more diverse microbiome, but greater abundances of environment-specific bacteria were found associated with host genotypes that exhibited greater browsing behavior. Our results indicate that, although there is a great deal of variation between individuals, behavior can mediate genotype-by-environment interaction effects on microbiome composition.

Navigation | Read Files | Portal | Attributes

Bulk Download Files (If the downloader app doesn't open, please try using Firefox to launch it.)

Download: 1 - 512 of 512 results in TEXT

Select columns

Showing results 1 - 10 of 512 results

Study accession	Sample accession	Secondary sample accession	Experiment accession	Run accession	Tax ID	Scientific name	Instrument model	Library layout	FASTQ files (FTP)	FASTQ files (Galaxy)	Submitted files (FTP)	Submitted files (Galaxy)	NCBI SRA file (FTP)	NCBI SRA file (Galaxy)	CRAM Index files (FTP)	CRAM Index files (Galaxy)
PRJEB30308	SAMEA5166093	ERS2973813	ERX2993334	ERR2990925	1869227	bacterium	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2	Fastq file 1 Fastq file 2	Fastq file 1 Fastq file 2	File 1	File 1		
PRJEB30308	SAMEA5166094	ERS2973814	ERX2993335	ERR2990926	1869227	bacterium	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2	Fastq file 1 Fastq file 2	Fastq file 1 Fastq file 2	File 1	File 1		
PRJEB30308	SAMEA5166095	ERS2973815	ERX2993336	ERR2990927	1869227	bacterium	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2	Fastq file 1 Fastq file 2	Fastq file 1 Fastq file 2	File 1	File 1		
PRJEB30308	SAMEA5166096	ERS2973816	ERX2993337	ERR2990928	1869227	bacterium	Illumina MiSeq	PAIRED	File 1 File 2	File 1 File 2	Fastq file 1 Fastq file 2	Fastq file 1 Fastq file 2	File 1	File 1		

NCBI Resources How To Sign in to NCBI

BioProject BioProject PRJEB30308 Search

Create alert Advanced Browse by Project attributes Help

Display Settings: Send to:

Microbiota of browsing Daphnia

Accession: PRJEB30308 ID: 516850

Microbiota associated with Daphnia exhibiting genetic variation in behavior

In many organisms, host-associated microbial communities are acquired horizontally after birth. More...

Accession	PRJEB30308
Scope	Monoisolate
Submission	Registration date: 24-Jan-2019 **Universitaet Basel**

Project Data:

Resource Name	Number of Links
SEQUENCE DATA	
SRA Experiments	512
OTHER DATASETS	
BioSample	512

SRA Data Details

Parameter	Value
Data volume, Gbases	22
Data volume, Mbytes	14805

Related information

BioSample

SRA

Recent activity

Turn Off Clear

PRJEB30308 (1)

BioProject

Microbiota of browsing Daphnia

BioProject

The European Nucleotide Archive in 2017

A Benchmark Study on Error Assessment and Quality Control of CCS Reads Derived

Testing the potential of a ribosomal 16S marker for DNA metabarcoding of insects

See more...

1. Choose the NGS technology according to your needs.

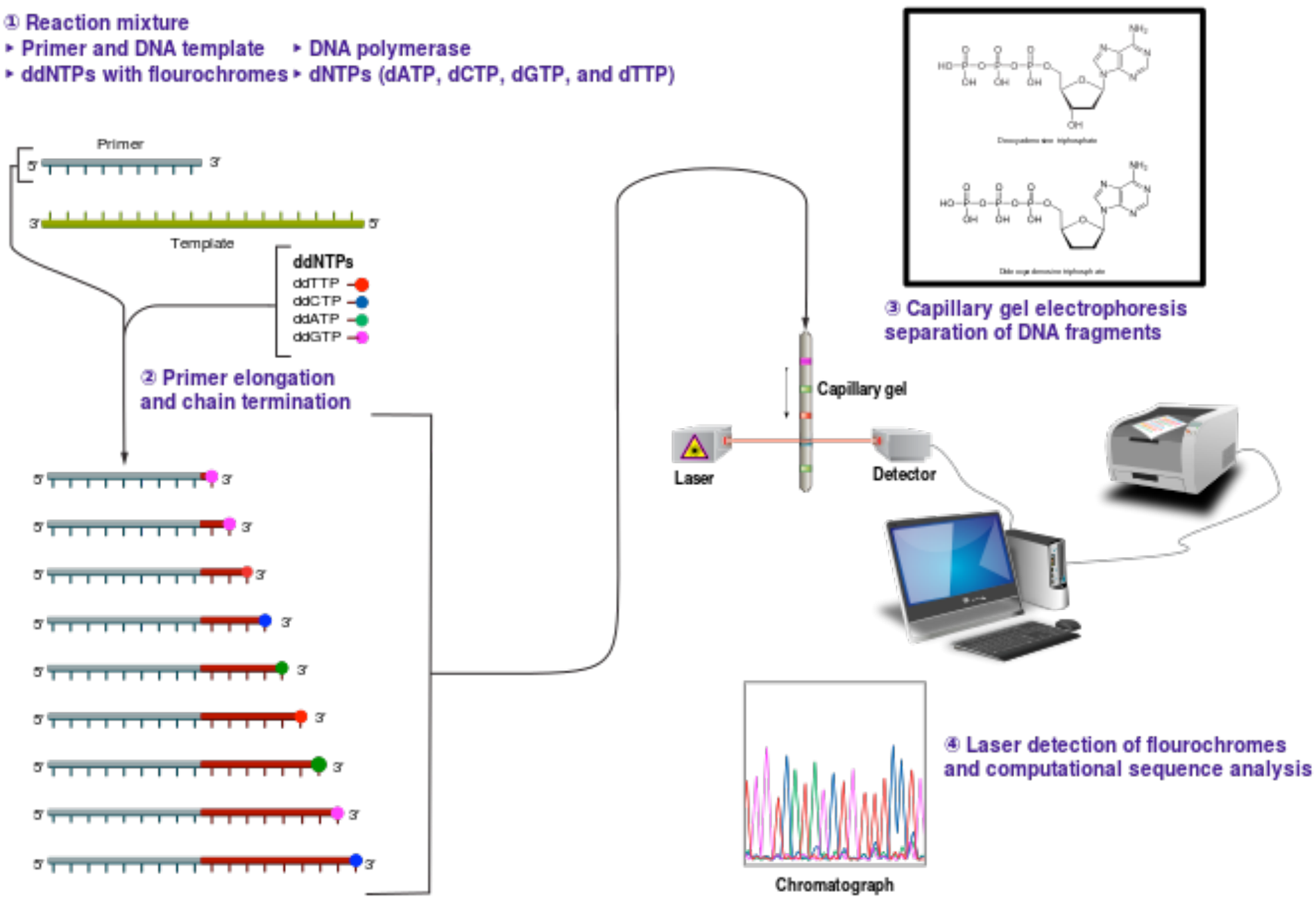
2. Keep your raw data safe and submit it as early as possible.

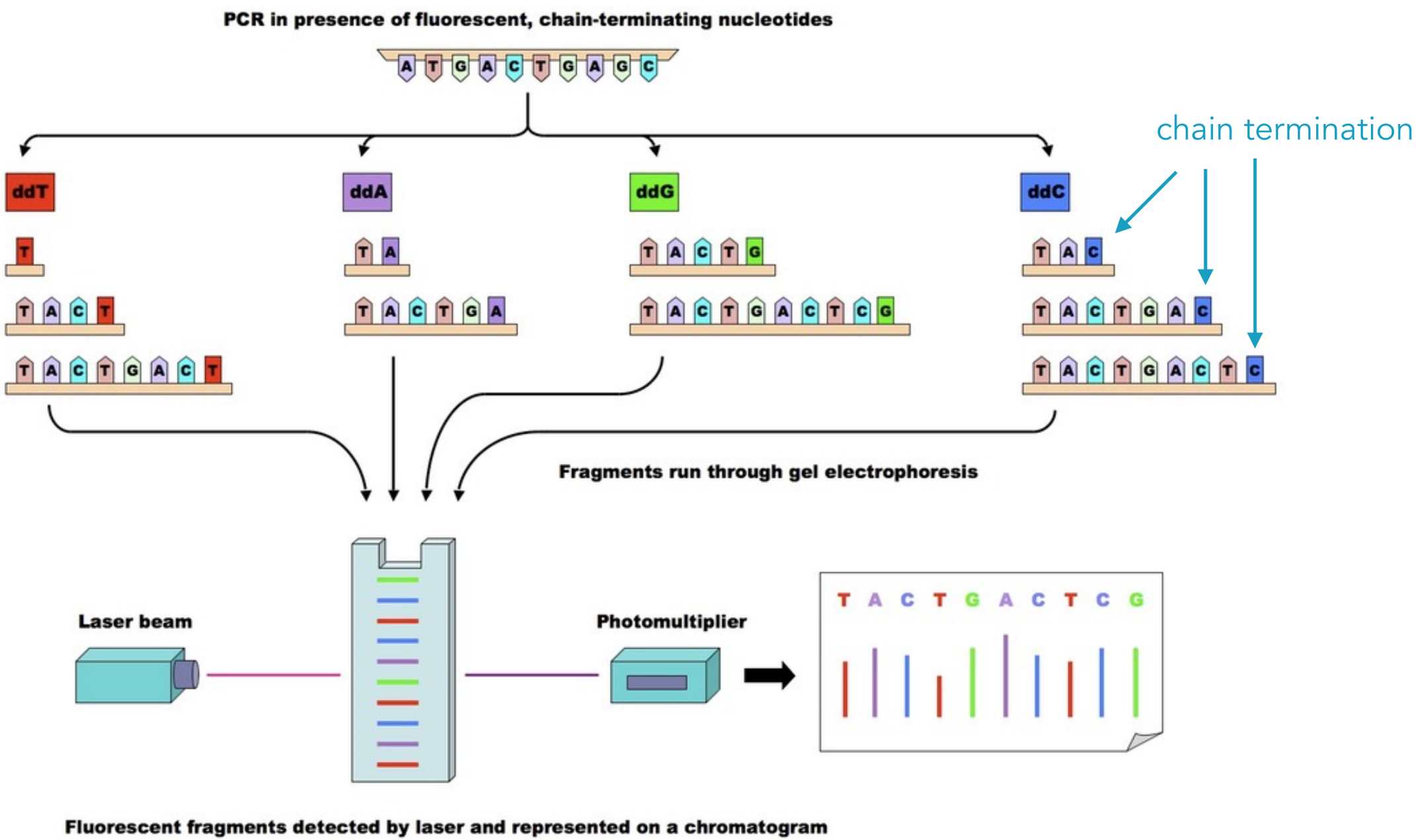
3. Coping one file (archive) is safer than coping multiple files.

EXTRAS

SEQUENCING TECHNOLOGIES EXTENDED

Capillary sequencing




PCR in presence of fluorescent, chain-terminating nucleotides
A T G A C T G A G C
ddT
ddA
ddG
ddC
chain termination
Fragments run through gel electrophoresis
Laser beam
Photomultiplier
T A C T G A C T C G
Fluorescent fragments detected by laser and represented on a chromatogram

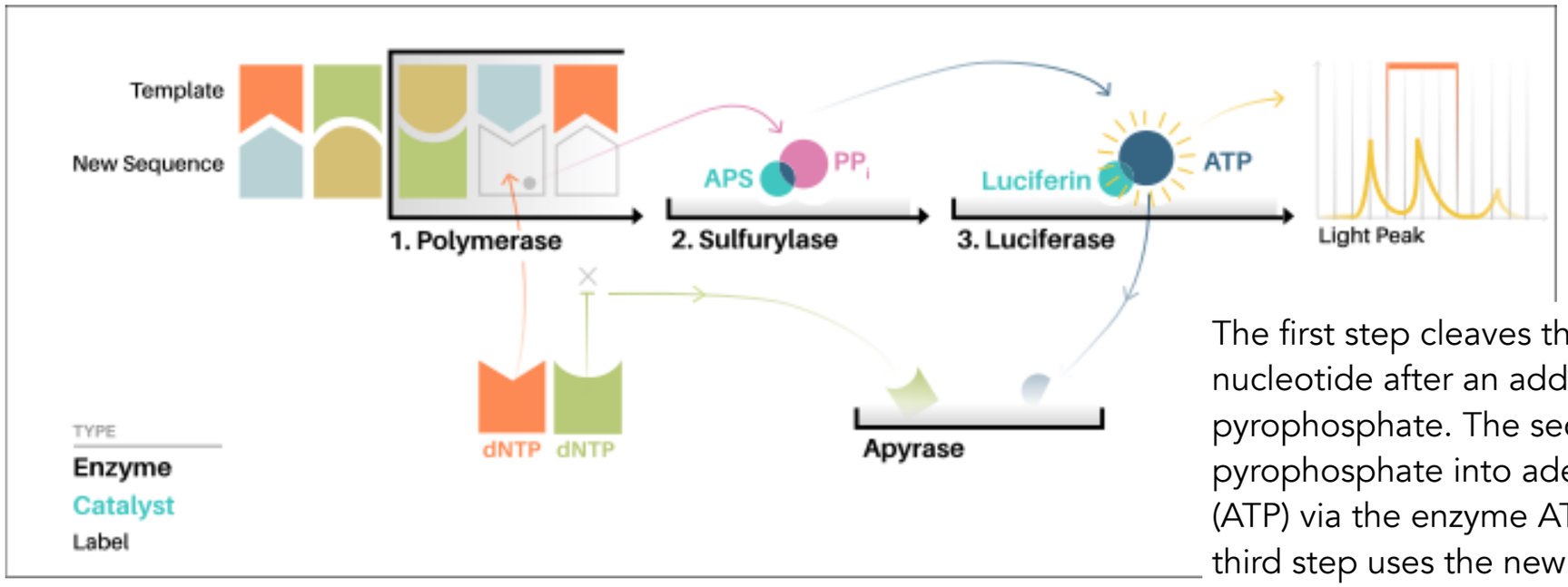
Pyrosequencing

GS Junior

Roche 454

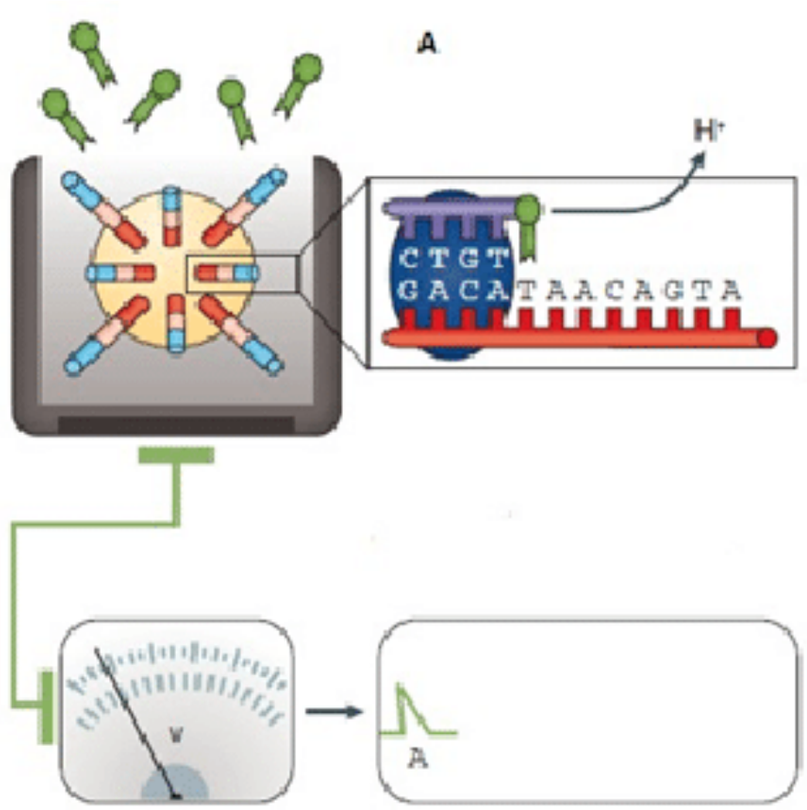
The **PyroMark** uses Pyrosequencing technology for real-time, sequence-based detection and quantification of sequence variants and epigenetic methylation. The PyroMark Q24 is highly suited for the analysis of CpG methylation, SNPs, insertion/deletions, STRs, variable gene copy number, as well as for microbial identification and resistance typing.

Pyrosequencing (pyrophosphate)



The first step cleaves the triphosphate nucleotide after an addition, releasing pyrophosphate. The second step converts pyrophosphate into adenosine triphosphate (ATP) via the enzyme ATP sulfurylase. The third step uses the newly synthesized ATP to catalyze the conversion of luciferin into oxyluciferin via the enzyme luciferase and this reaction generates a quanta of light that is captured from the picotiter plate by a charge- coupled camera.

Ion Torrent (semiconductor technology)



MiniSeq System
1.8-7.5 Gb
8-25 million
2 x 150 bp
50

MiSeq Series
0.3-15 Gb
1-25 million
2 x 300 bp
384

NextSeq Series
20-120 Gb
130-400 million
2 x 150 bp
96

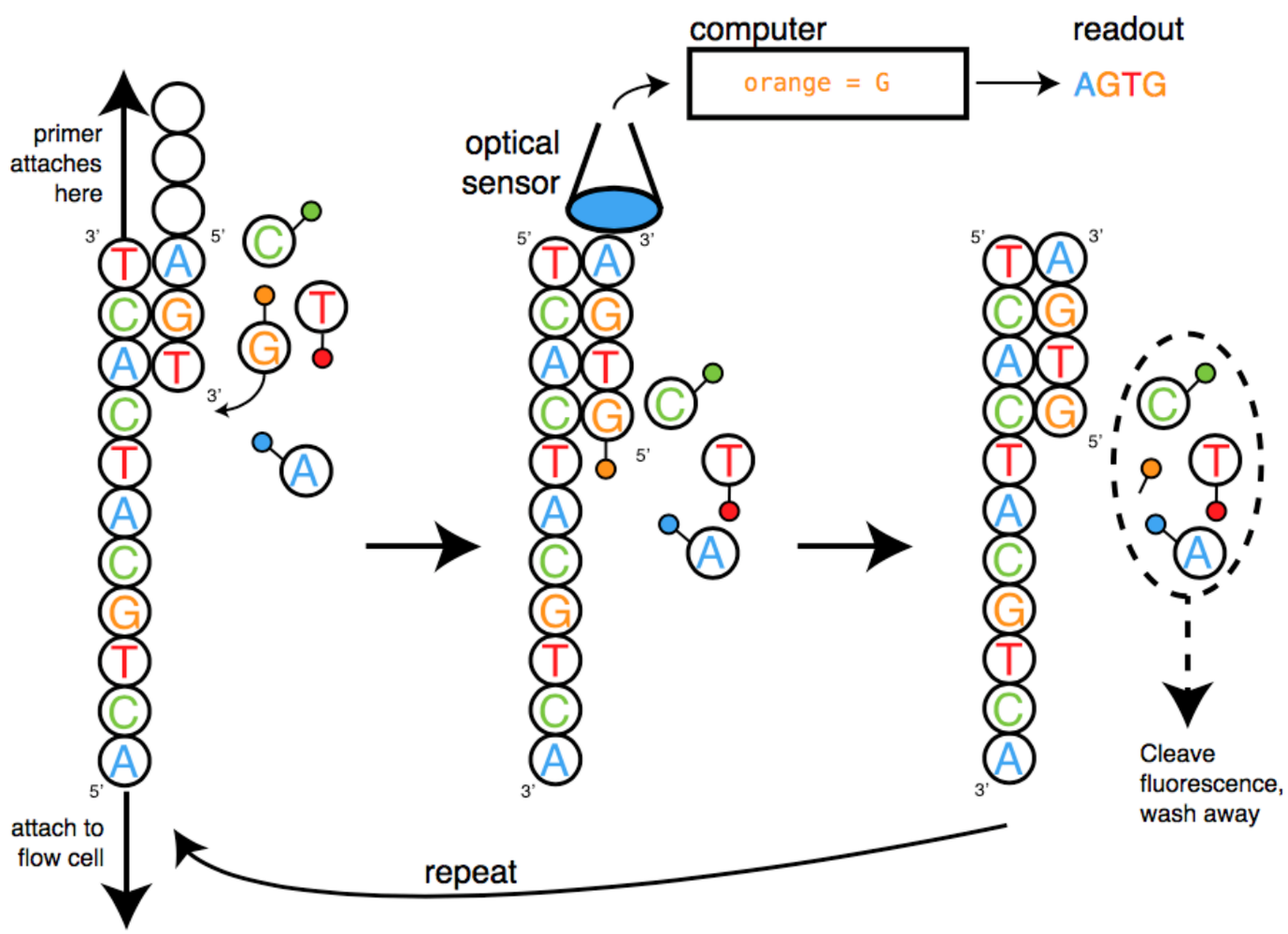
HiSeq Series
125-1500 Gb
2.5-5 billion
2 x 150 bp
12

HiSeq X Series
900-1800 Gb
3-6 billion
2 x 150 bp
16

NovaSeq Series
134-6000 Gb
Up to 20 billion
2 x 150 bp
48

http://www.illumina.com

Sequencing by Synthesis (fluorescent)

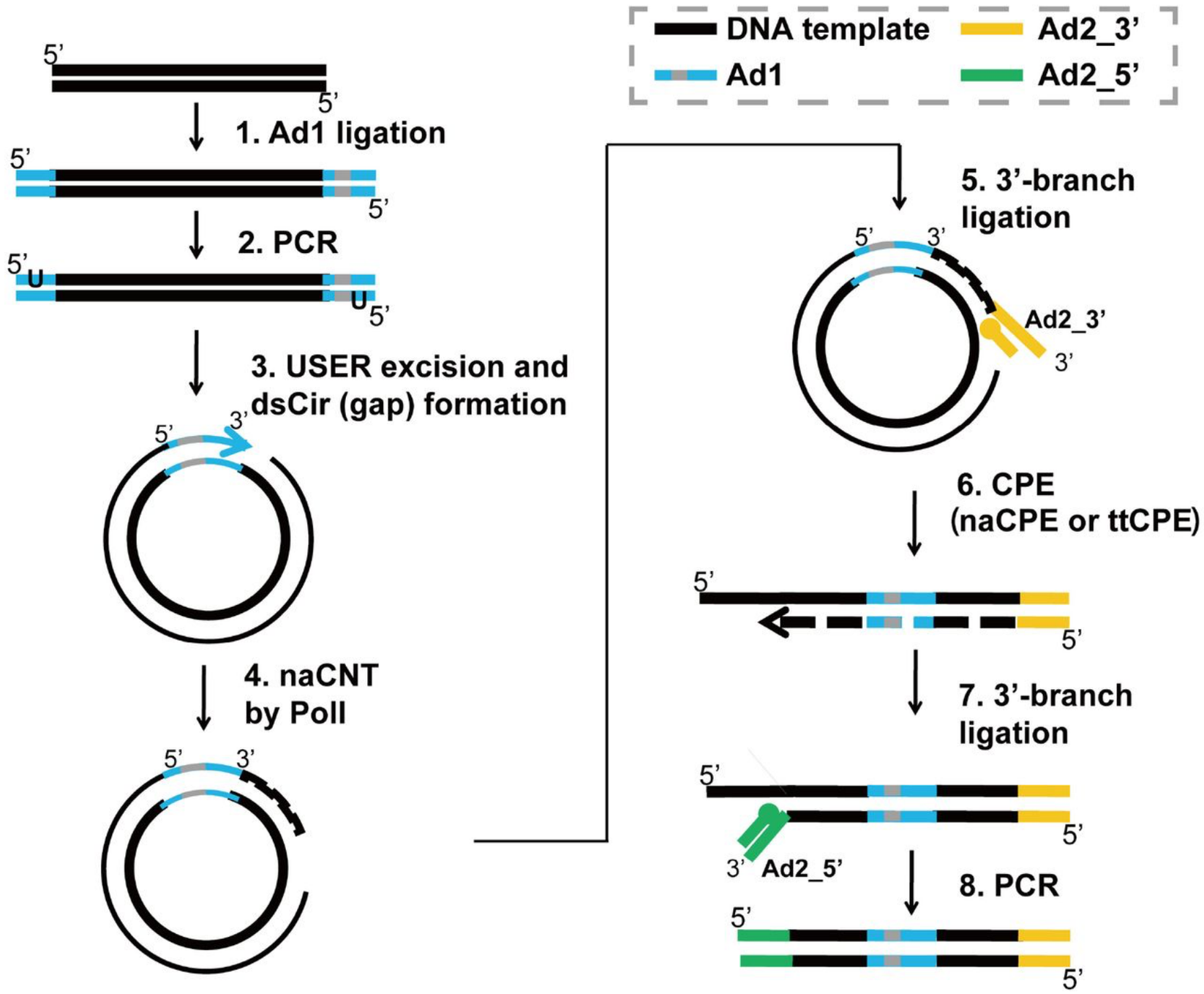


Sequencing by Synthesis. dNTP fluorescence is translated to a base call.

GDC Genetic Diversity Centre Zurich

	 Sequencers +	 Sequencers +	 Sequencers +	 Sequencers +
Product Model	DNBSEQ-T7	DNBSEQ-G400	DNBSEQ-G50	DNBSEQ-G400 FAST
Features	Ultra-high Throughput	Adaptive	Effective	Fast
Applications	Whole Genome Sequencing,Deep Exome Sequencing,Transcriptome Sequencing,and Targeted Panel Projects.	WGS, WES, Transcriptome sequencing and more	Small whole genome sequencing, targeted DNA/RNA panels, low-pass whole genome sequencing	Targeted DNA, RNA, Epigenetics and clinical applications
Flow Cell Type	FC	FCL & FCS	FCL & FCS	FCS
Lane/Flow Cell++	1 lane	4 lane & 2 lane	1 lane	2 lane
Operation Mode	Ultra-high Throughput	High Throughput	Medium Throughput	Medium Throughput
Max. Throughput / RUN	6Tb	1440Gb	150Gb	330G
Effective Reads / Flow Cell	5000M	1500-1800M	500M / 100M	550M
Average run time	PE150 within 24 hours	~38 hours	10-66 hours	12-37 hours
Min. Read Length	PE100	SE50	SE50	SE100
Max. Read Length	PE150	SE400	PE150	PE150

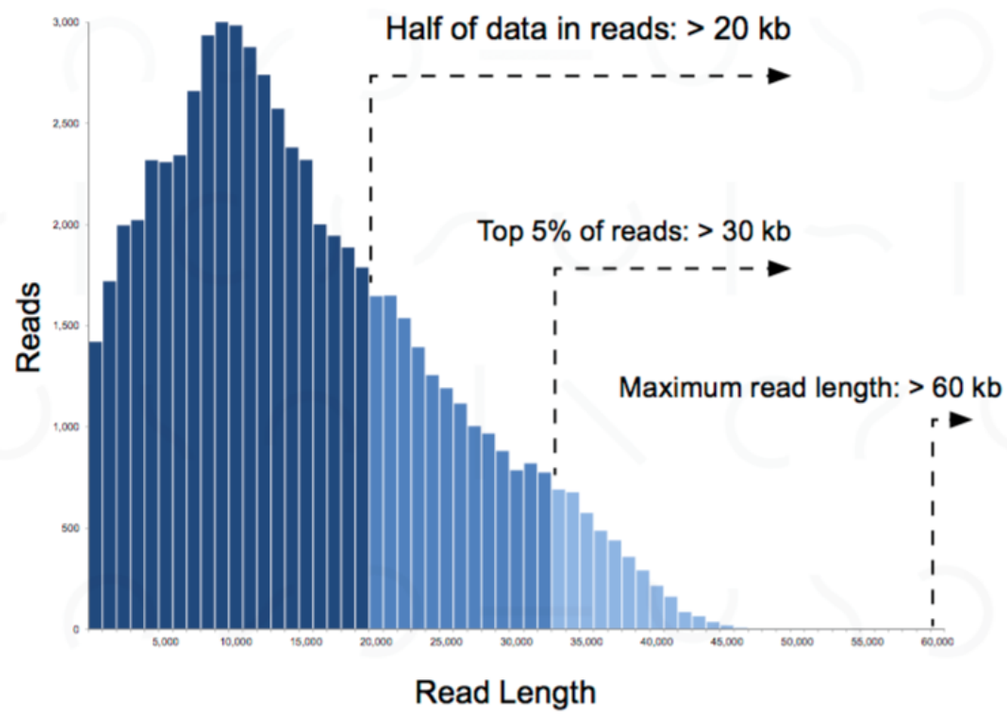
https://en.mgitech.cn


DNA template
Ad1
Ad2_3'
Ad2_5'
1. Ad1 ligation
2. PCR
3. USER excision and dsCir (gap) formation
4. naCNT by PolI
5. 3'-branch ligation
6. CPE (naCPE or ttCPE)
7. 3'-branch ligation
8. PCR
Ad2_3'
Ad2_5'

Long Read Lengths

Read lengths > 20 kb
Data per SMRT Cell: 750 Mb – 1.25 Gb

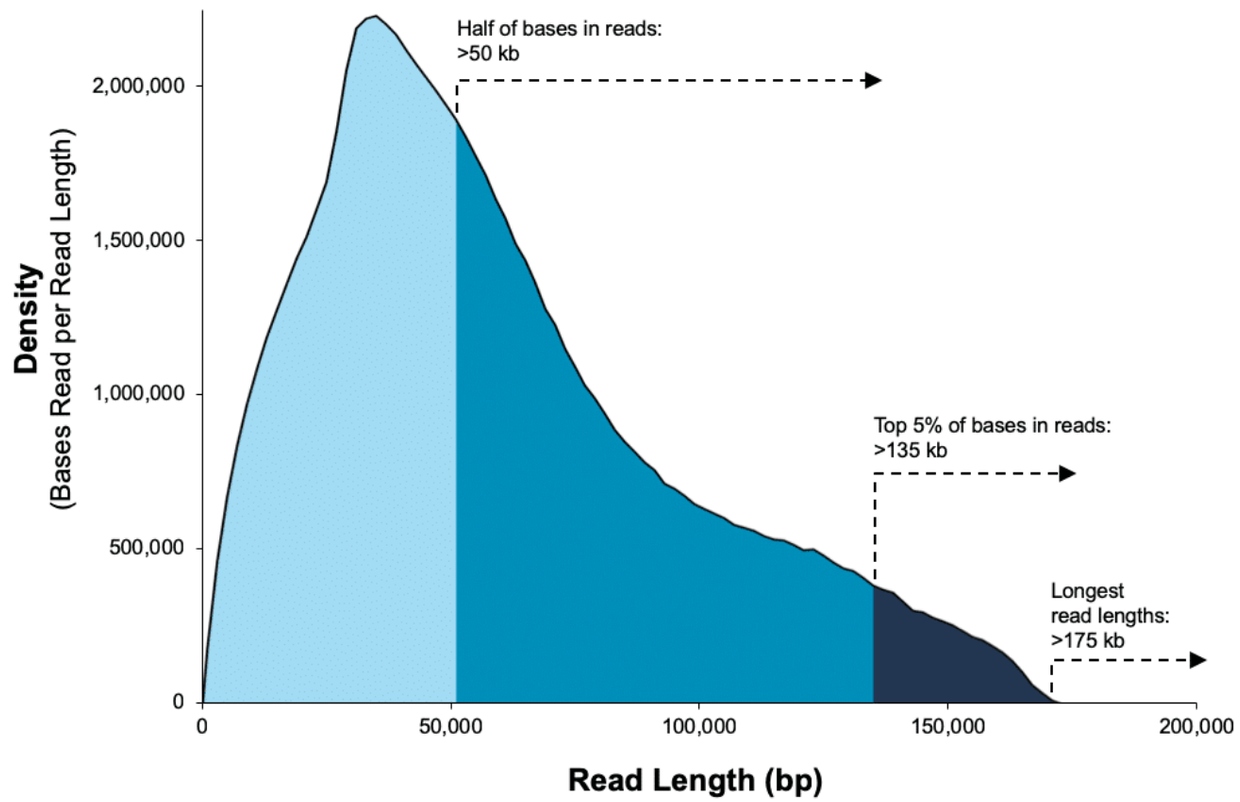
PacBio RS II



Read-length data shown above is from a 20 kb size-selected human library run on a PacBio RS II (6-hour movie, P6-C4 chemistry). The PacBio RS II SMRT Cells generate ~55,000 reads. The Sequel System generates ~370,000 reads per SMRT Cell.

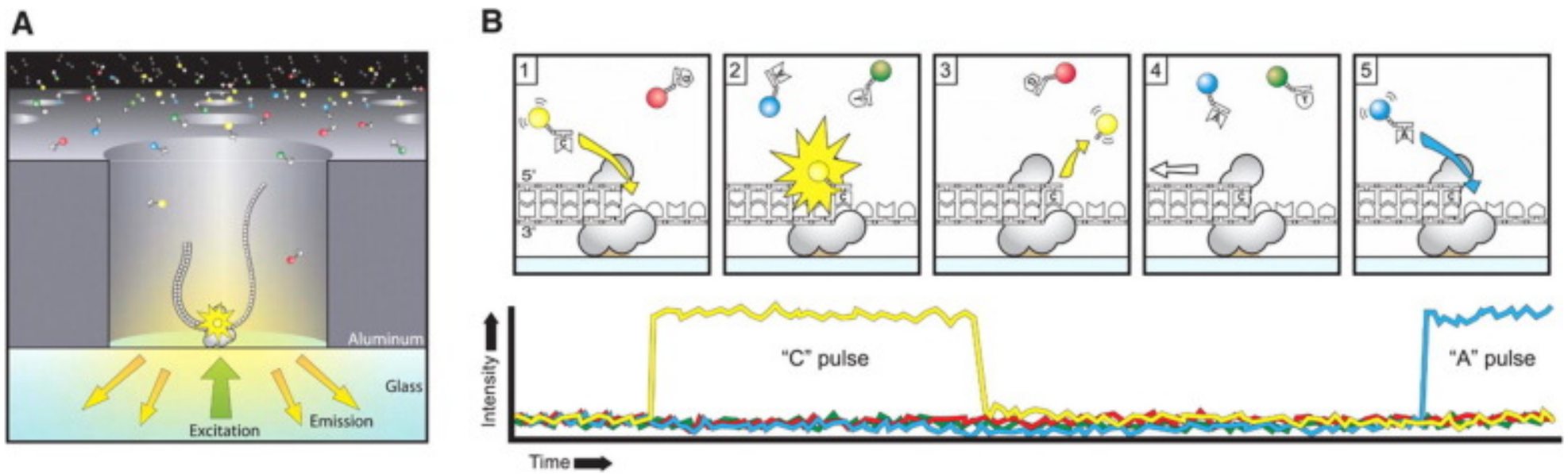
http://www.pacb.com

Sequel



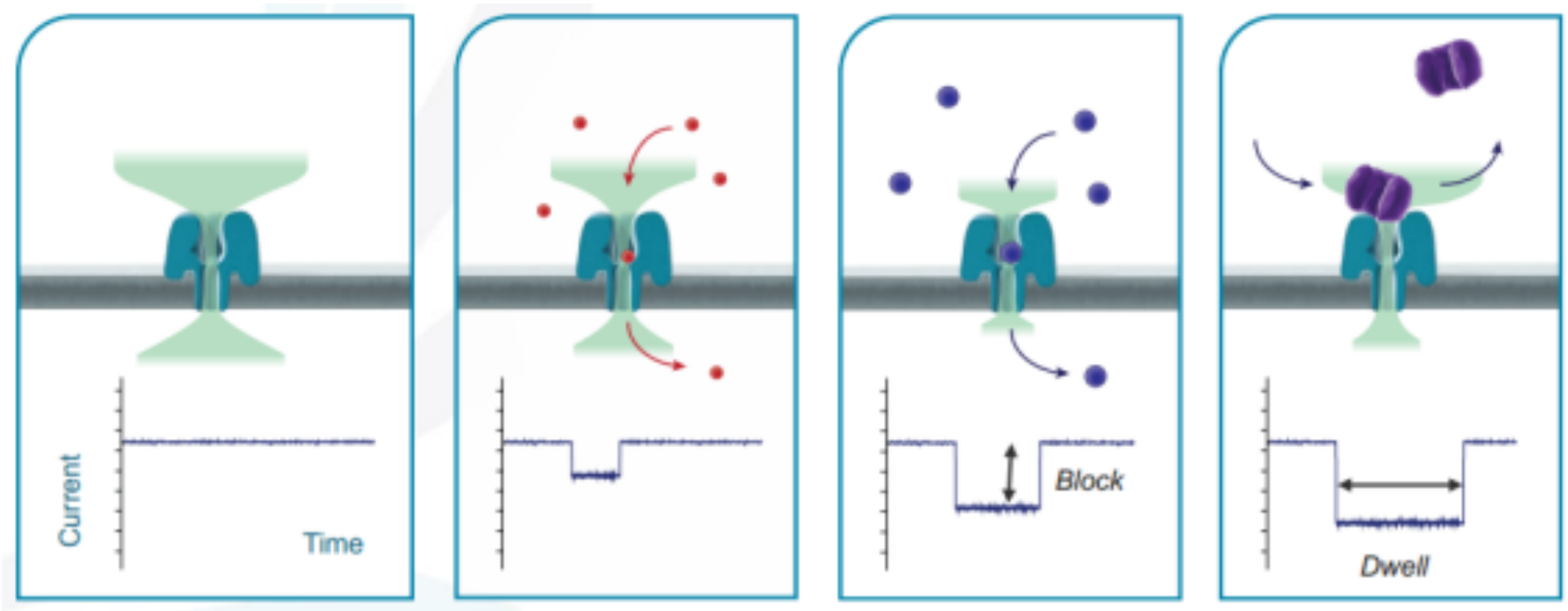
Data from a 35 kb size-selected *E. coli* library using the SMRTbell Express Template Prep Kit 2.0 on a Sequel II System (1.0 Chemistry, Sequel II System Software v7.0, 15-hour movie)*.

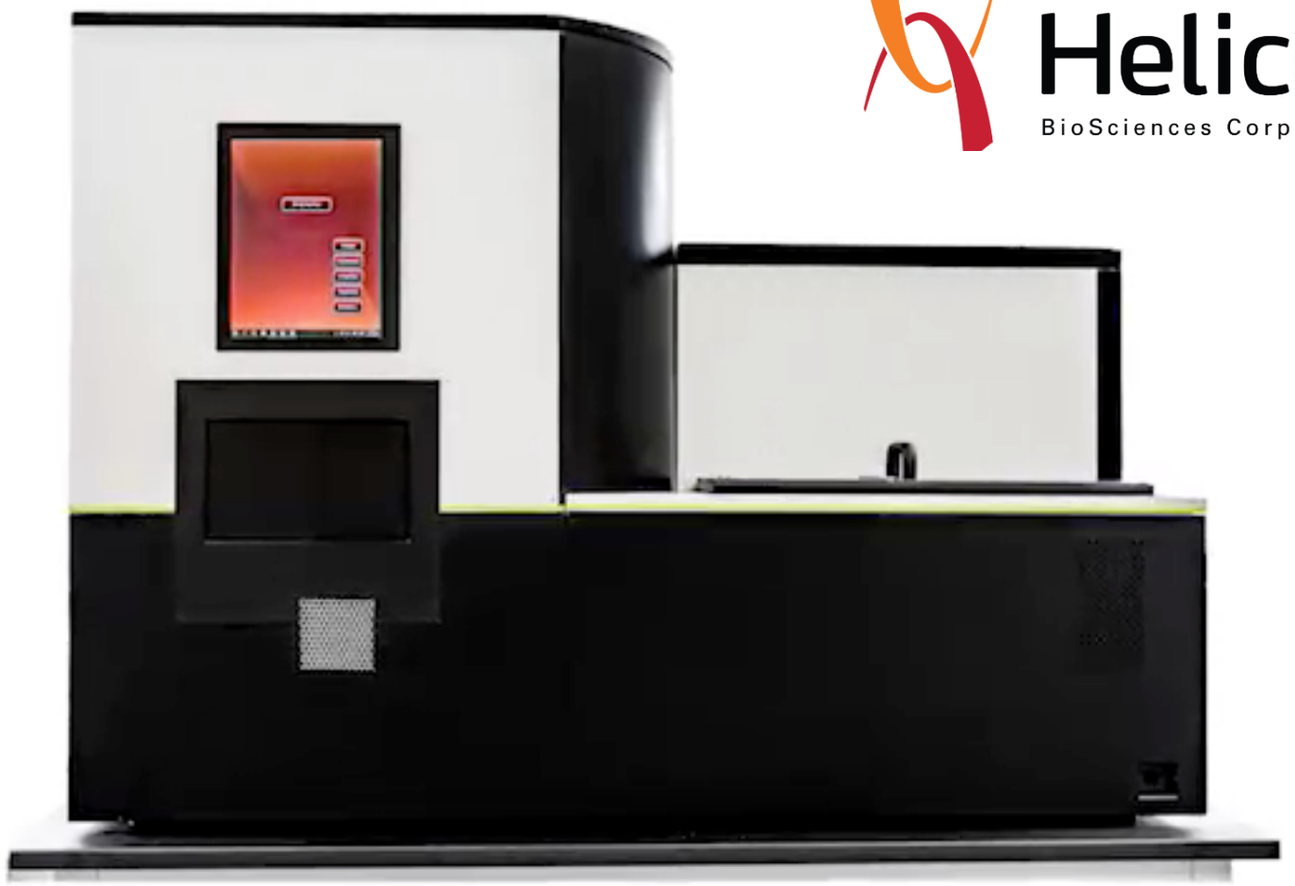
PacBio (fluorophore)



Oxford NANOPORE Technologies

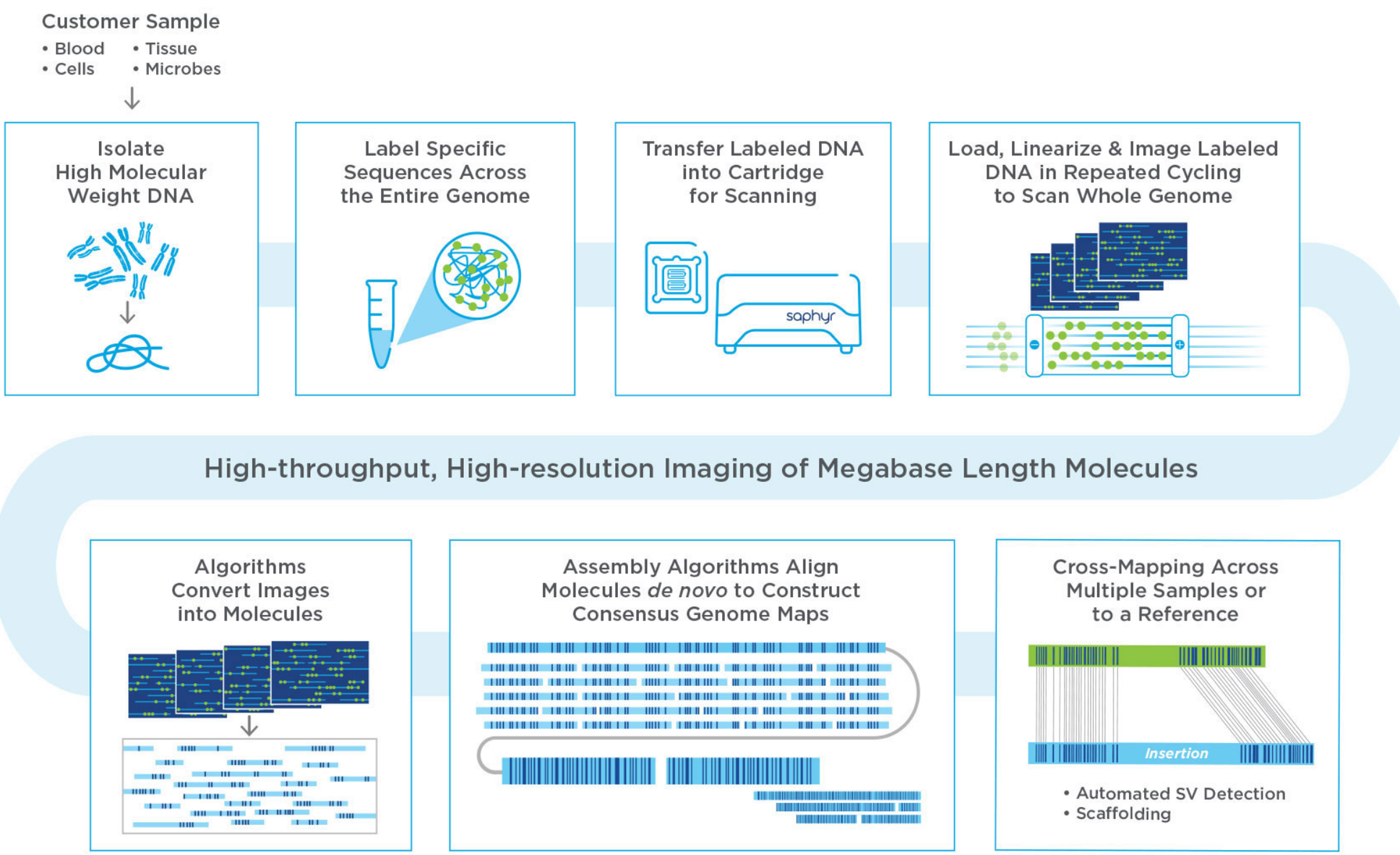
https://www.nanoporetech.com


Current
Time
Block
Dwell



Helicos
BioSciences Corporation


bionano
GENOMICS
saphyr


Customer Sample
• Blood
• Tissue
• Cells
• Microbes
Isolate High Molecular Weight DNA
Label Specific Sequences Across the Entire Genome
Transfer Labeled DNA into Cartridge for Scanning
saphyr
Load, Linearize & Image Labeled DNA in Repeated Cycling to Scan Whole Genome
High-throughput, High-resolution Imaging of Megabase Length Molecules
Algorithms Convert Images into Molecules
Assembly Algorithms Align Molecules *de novo* to Construct Consensus Genome Maps
Cross-Mapping Across Multiple Samples or to a Reference
Insertion
• Automated SV Detection
• Scaffolding