Next Generation Sequencing –

An Overview

Olga Vinnere Pettersson, PhD
National Genomics Infrastructure hosted by ScilifeLab,
Uppsala Node (UGC)

Version 5.2.3.b

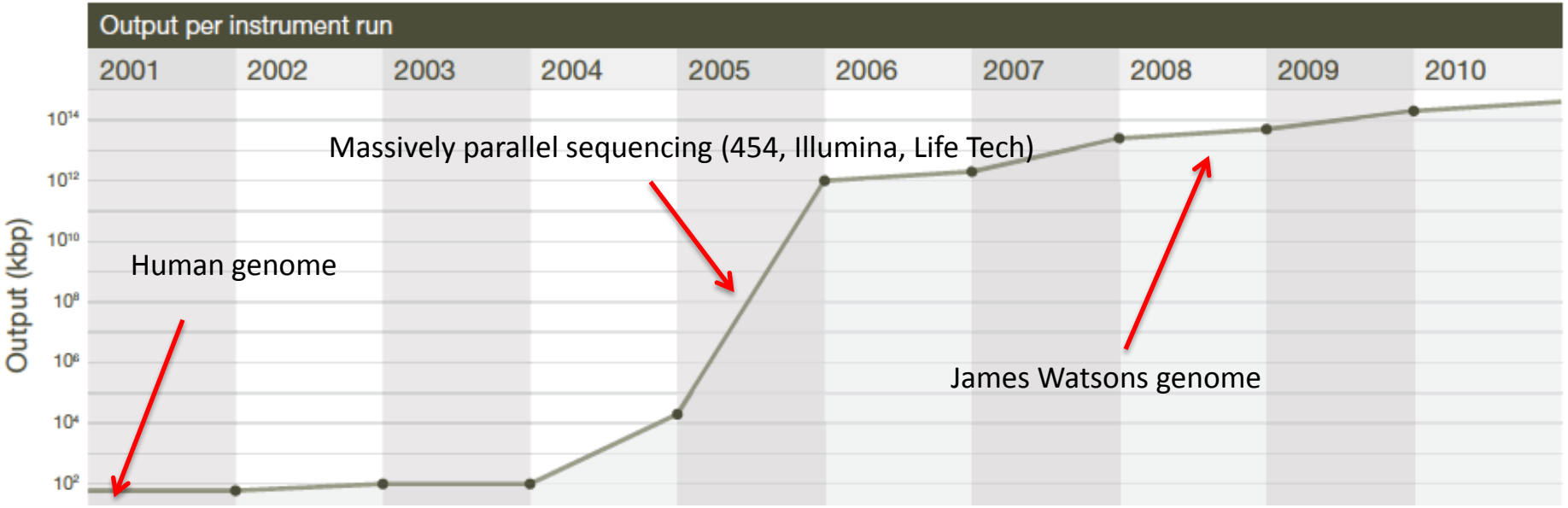
Today we will talk about:

- History and current state of genomic research

- Sequencing technologies:
  - Types
  - Principles
  - Sample prep
  - Their “+” and “-”
  - Couple of pieces of advise

- National Genomics Infrastructure – Sweden

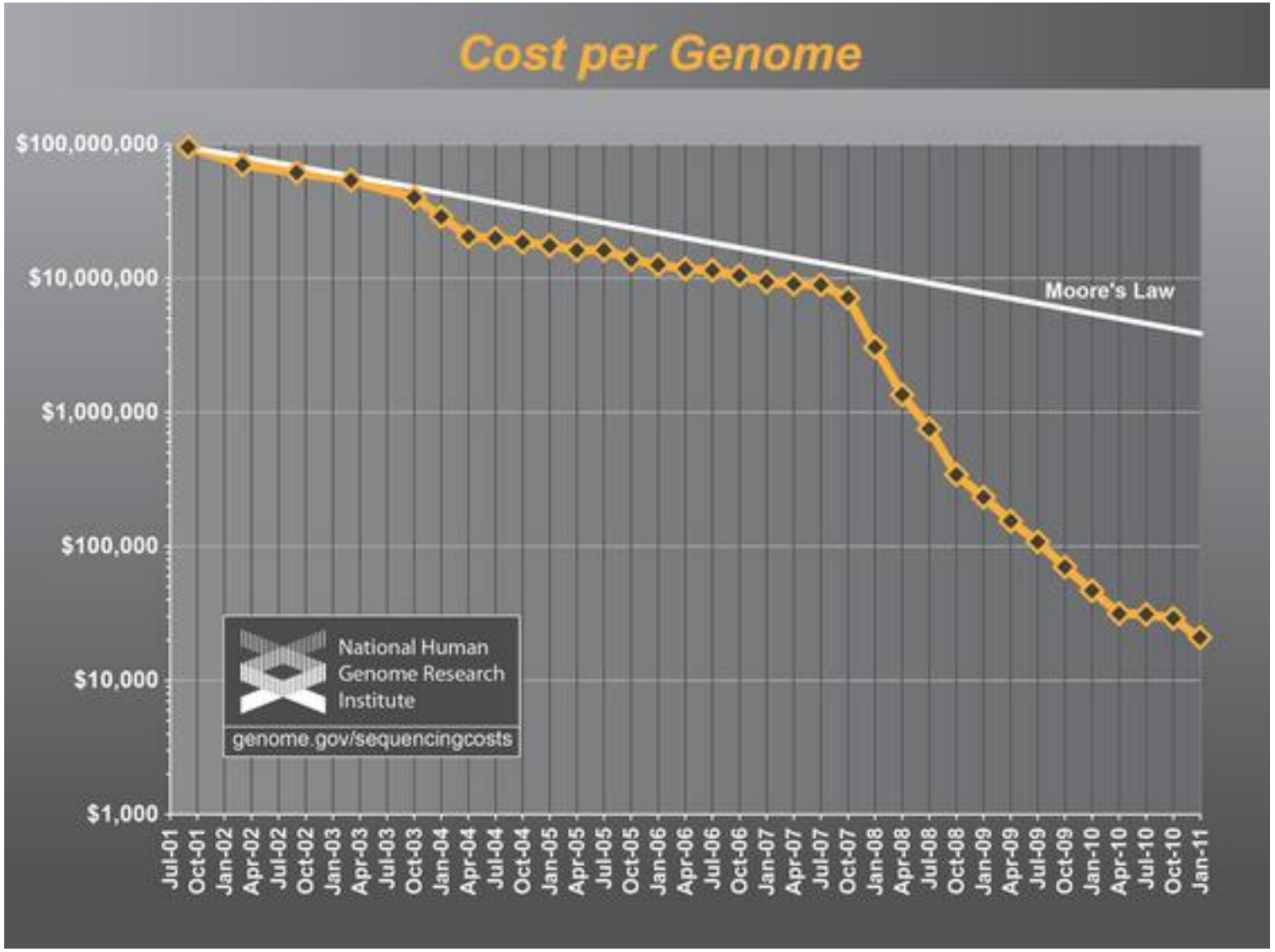
DNA sequencing revolution



Center for Metagenomic Sequence Analysis (KAW)

Swedish National Infrastructure for Large-Scale Sequencing (SNISS)

Science for Life Laboratory (SciLifeLab)


Cost per Genome
$100,000,000
$10,000,000
$1,000,000
$100,000
$10,000
$1,000
Moore's Law
National Human Genome Research Institute
genome.gov/sequencingcosts
Jul-01
Oct-01
Jan-02
Apr-02
Jul-02
Oct-02
Jan-03
Apr-03
Jul-03
Oct-03
Jan-04
Apr-04
Jul-04
Oct-04
Jan-05
Apr-05
Jul-05
Oct-05
Jan-06
Apr-06
Jul-06
Oct-06
Jan-07
Apr-07
Jul-07
Oct-07
Jan-08
Apr-08
Jul-08
Oct-08
Jan-09
Apr-09
Jul-09
Oct-09
Jan-10
Apr-10
Jul-10
Oct-10
Jan-11

What is sequencing?

DEFINITION

- "In genetics and biochemistry, **sequencing** means to determine the primary structure (or primary sequence) of an unbranched biopolymer."

  (http://en.wikipedia.org/wiki/Sequencing)

Once upon a time…

- Fredrik Sanger and Alan Coulson

Chain Termination Sequencing (1977)

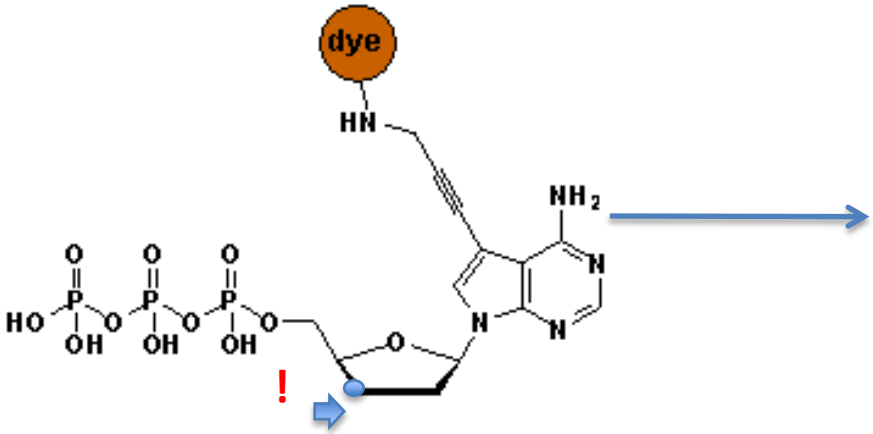
Nobel prize 1980

Principle:

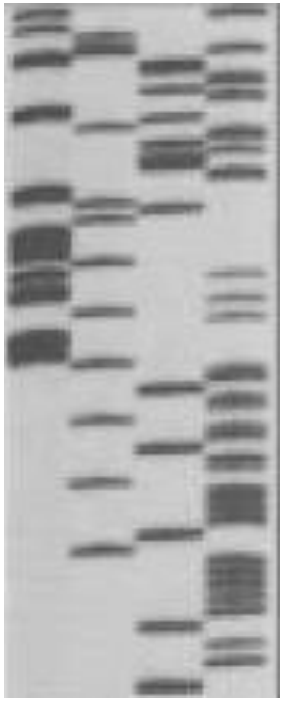
SYNTHESIS of DNA is randomly TERMINATED at different points

Separation of fragments that are 1 nucleotide different in size

Sanger's sequencing

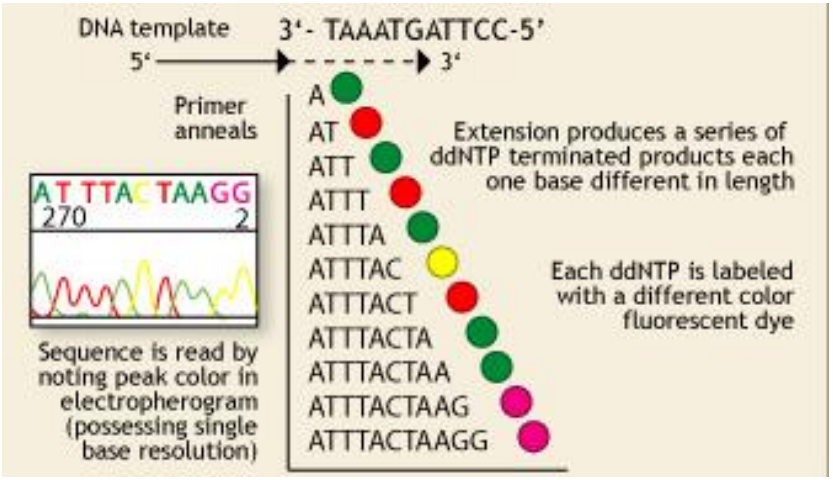


P^{32} labelled ddNTPs



Lack of OH-group at 3' position of deoxyribose

Fluorescent dye terminators



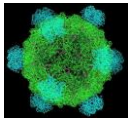
Max fragment length – 750 bp

Sequencing genomes using **Sanger**'s method

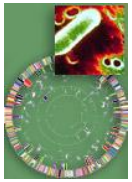
- Extract & purify genomic DNA
- Fragmentation
- Make a clone library
- Sequence clones
- Align sequencies (-> contigs -> scaffolds)
- Close the gaps

- Cost/Mb=1000 $, and it takes TIME

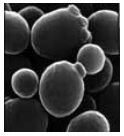
At the very beginning of genome sequencing era…



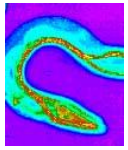
First genome: virus ϕ X 174 - 5 368 bp (1977)



First organism: *Haemophilus influenzae* - 1.5 Mb (1995)



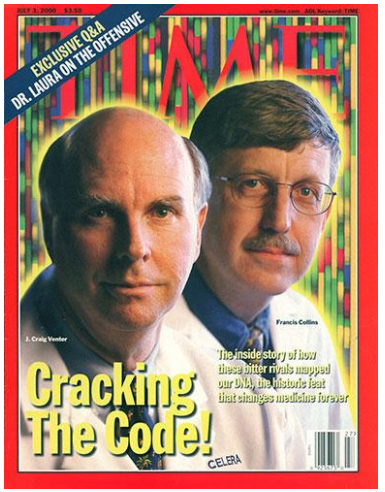
First eukaryote: *Saccharomyces cerevisiae* - 12.4 Mb (1996)



First multicellular organism: *Cenorhabditis elegans* - 100 MB (1998-2002)

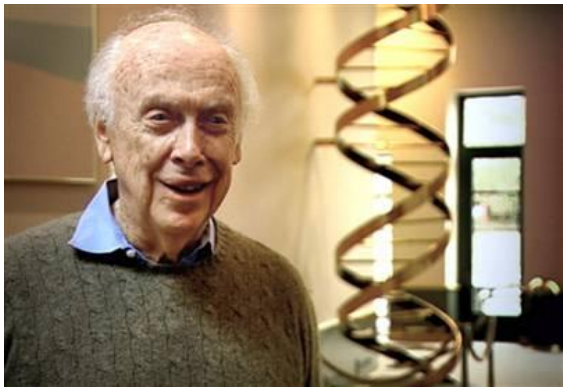
First plant: *Arabidopsis thaliana* - 157 Mb (2000)

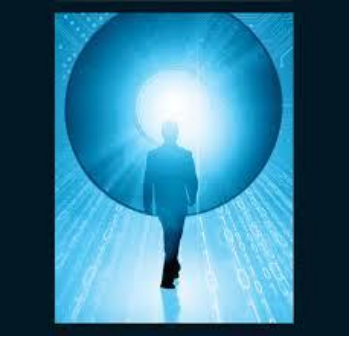
Just an interesting comparison:



- Human genome project, 2007
  - Genome of Craig Wenter costs 70 mln $
    - Sanger's sequencing
  - Genome of James Watson costs 2 mln $
    - 454 pyrosequencing
  - Ultimate goal: 1000 $ / individual

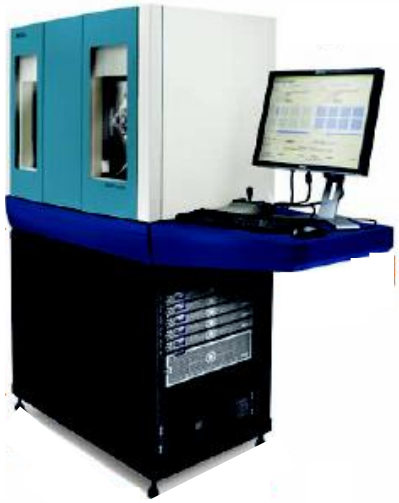
    Almost there!

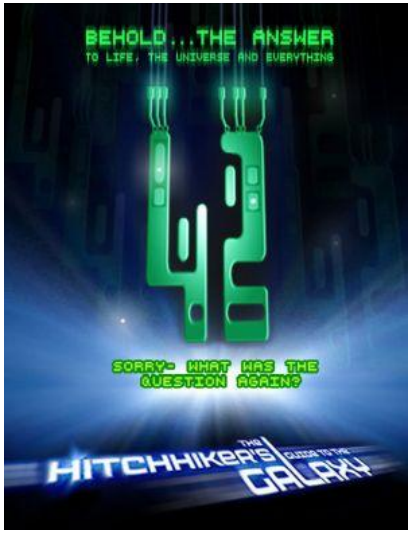




Paradigm change

- From single genes to complete genomes
- From single transcripts to whole transcriptomes
- From single organisms to complex metagenomic pools
- From model organisms to the species you are studying





Science 5 September 1997:
Vol. 277 no. 5331 pp. 1453–1462
DOI: 10.1126/science.277.5331.1453

IF 31.6

‹ Prev | Table of Contents | Next ›

ARTICLES

The Complete Genome Sequence of *Escherichia coli* K-12

Frederick R. Blattner*, Guy Plunkett III*, Craig A. Bloch, Nicole T. Perna, Valerie Burland, Monica Riley, Julio Collado-Vides, Jeremy D. Glasner, Christopher K. Rode, George F. Mayhew, Jason Gregor, Nelson Wayne Davis, Heather A. Kirkpatrick, Michael A. Goeden, Debra J. Rose, Bob Mau and Ying Shao

Journal of Biotechnology
Article in Press, Corrected Proof - Note to users

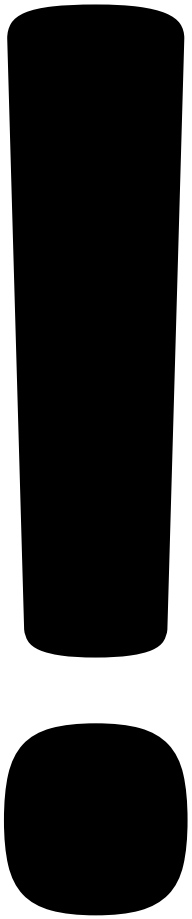
IF 2.9

doi:10.1016/j.jbiotec.2010.12.018 | How to Cite or Link Using DOI
Permissions & Reprints

The complete genome sequence of the dominant *Sinorhizobium meliloti* field isolate SM11 extends the *S. meliloti* pan-genome

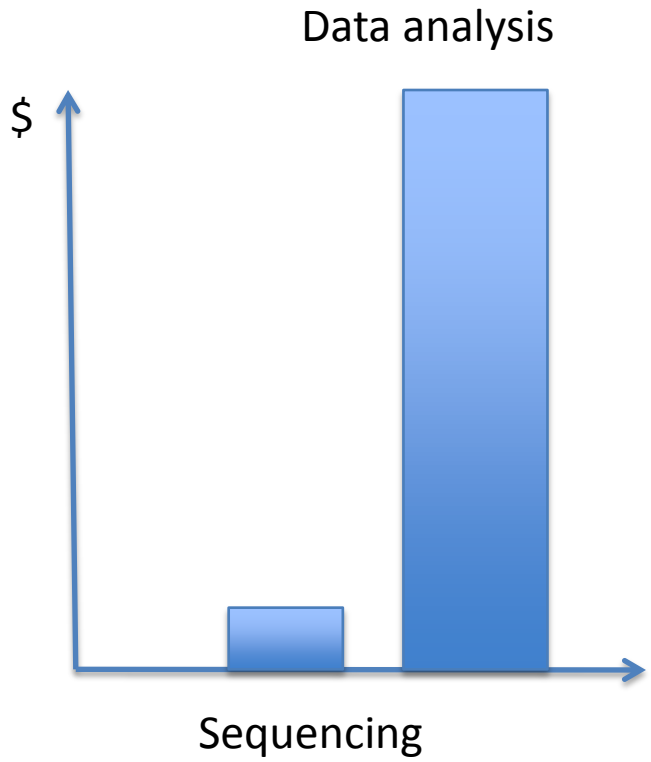
Susanne Schneiker-Bekel[a], Daniel Wibberg[a], Thomas Bekel[b], Jochen Blom[b], Burkhard Linke[b], Heiko Neuweger[b], Michael Stiens[a, c], Frank-Jörg Vorhölter[a], Stefan Weidner[a], Alexander Goesmann[b], Alfred Pühler[a] and Andreas Schlüter[a,]

Main hazard - DATA ANALYSIS



> *"If the data problem is not addressed, ABI's SOLiD, 454's GS FLX, Illumina's GAII or any of the other deep sequencing platforms will be destined to sit in their air-conditioned rooms like a Stradivarius without a bow."*

http://finchtalk.geospiza.com



=> More bioinformaticians to people!

Major NGS technologies

NGS technologies

Company	Platform	Amplification	Sequencing method
Roche	454**	emPCR	Pyrosequencing
Illumina	HiSeq MiSeq	Bridge PCR	Synthesis
LifeTech	SOLiD**	emPCR/ Wildfire	Ligation
LifeTech	Ion Torrent Ion Proton	emPCR	Synthesis (pH)
Pacific Bioscience	RSII	None	Synthesis
Complete genomics	Nanoballs	None	Ligation
Oxford Nanopore*	GridION	None	Flow

RIP technologies: Helicos, Polonator, etc.

In development: Tunneling currents, nanopores, etc.

Differences between platforms

- Technology: chemistry + signal detection
- Run times vary from hours to days
- Production range from Mb to Gb
- Read length from <100 bp to > 20 Kbp
- Accuracy per base from 0.1% to 15%
- Cost per base varies

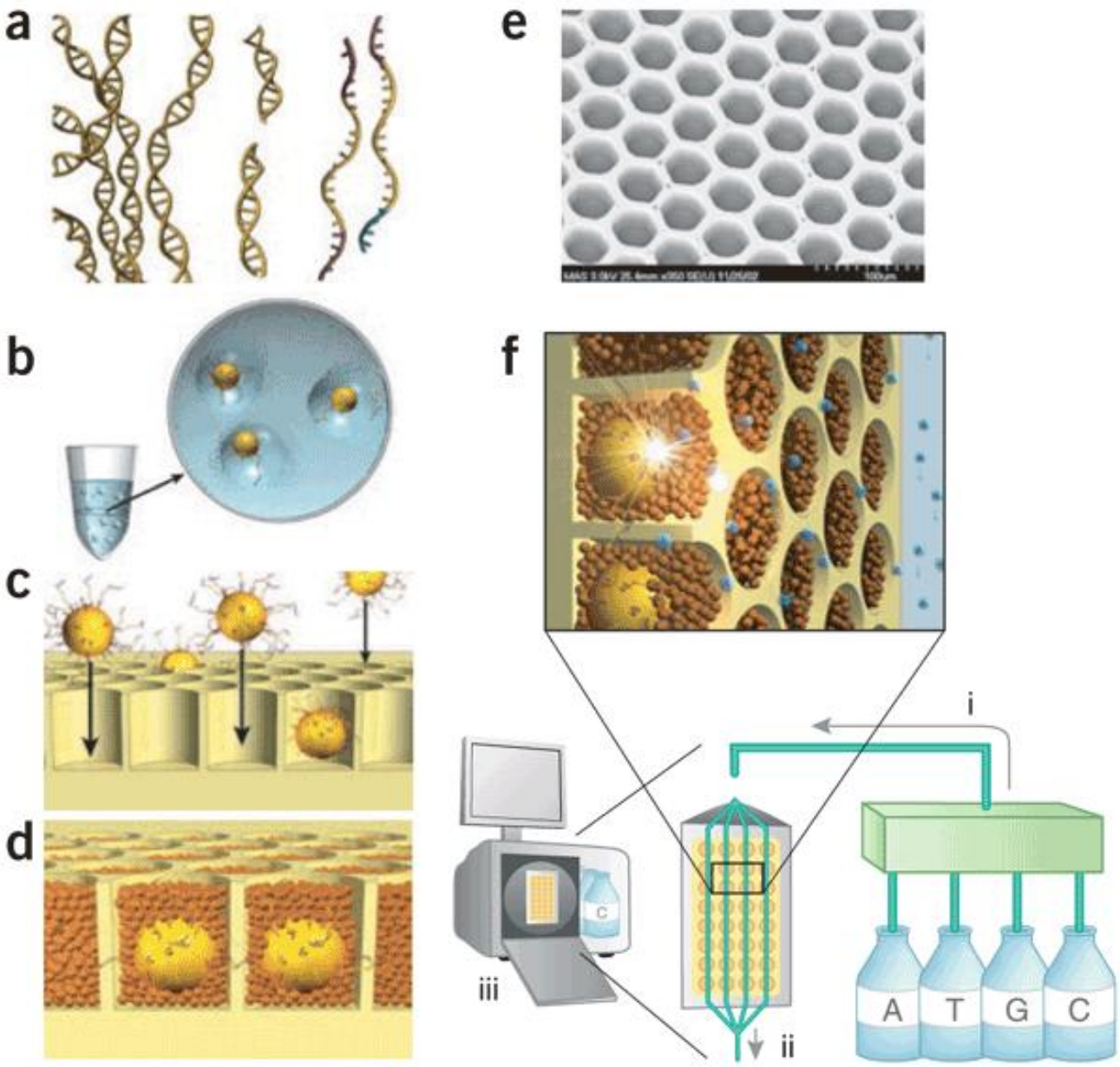
Roche

Instrument	Yield and run time	Read Length	Error rate	Error type
454 FLX+	0.9 GB, 20 hrs	700	1%	Indels
454 FLX Titanium	0.5 GB, 10 hrs	450	1%	Indels
454 FLX Jr	0.050 GB, 10 hrs	400	1%	Indels

Main applications:

- Microbial genomics and metagenomics
- Targeted resequencing

454 Titanium GS FLX



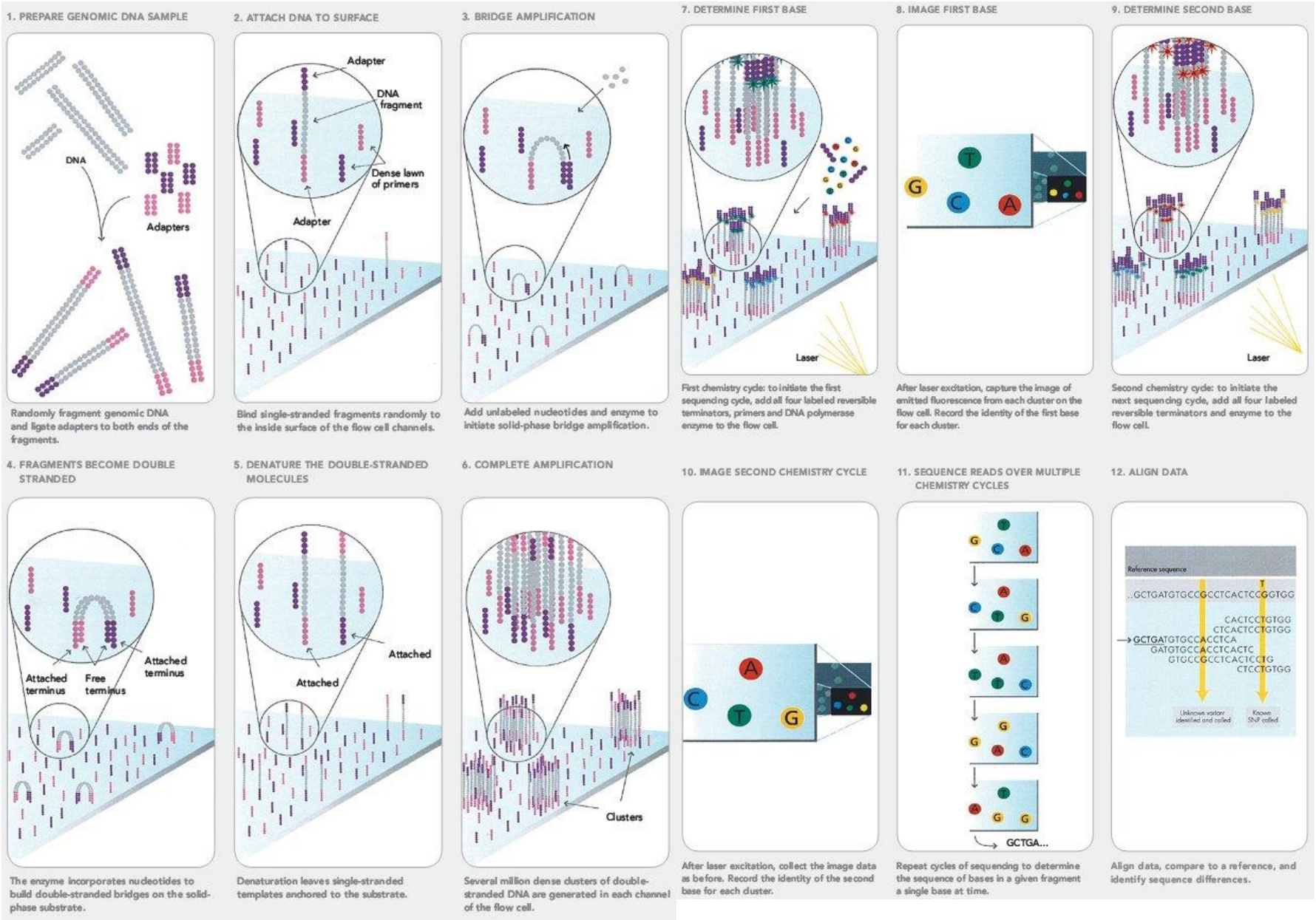
Illumina

Instrument	Yield and run time	Read Length	Error rate	Error type
Upgrade HiSeq2500	120 Gb – 600 Gb 27h or standard run	100x100	0.1%	Subst
MiSeq	540 Mb – 15 Gb (4 – 48 hours)	Up to 350x350	0.1%	Subst
HiSeqXten	800 Gb - 1.8 Tb (3 days)	150x150	"	"

Main applications

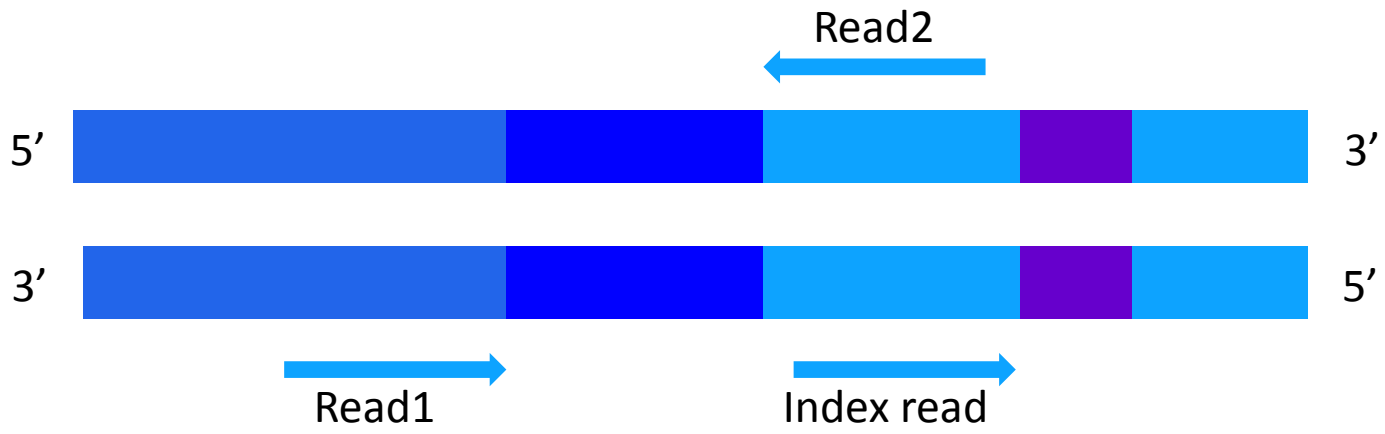
- Whole genome, exome and targeted reseq
- Transcriptome analyses
- Methylome and ChiPSeq
- Rapid targeted resequencing (MiSeq)
- Human genome seq (Xten)

Illumina


1. PREPARE GENOMIC DNA SAMPLE
DNA
Adapters
Randomly fragment genomic DNA and ligate adapters to both ends of the fragments.
2. ATTACH DNA TO SURFACE
Adapter
DNA fragment
Dense lawn of primers
Adapter
Bind single-stranded fragments randomly to the inside surface of the flow cell channels.
3. BRIDGE AMPLIFICATION
Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.
7. DETERMINE FIRST BASE
Laser
First chemistry cycle: to initiate the first sequencing cycle, add all four labeled reversible terminators, primers and DNA polymerase enzyme to the flow cell.
8. IMAGE FIRST BASE
T
G
C
A
After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster.
9. DETERMINE SECOND BASE
Laser
Second chemistry cycle: to initiate the next sequencing cycle, add all four labeled reversible terminators and enzyme to the flow cell.
4. FRAGMENTS BECOME DOUBLE STRANDED
Attached terminus
Free terminus
Attached terminus
The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.
5. DENATURE THE DOUBLE-STRANDED MOLECULES
Attached
Attached
Denaturation leaves single-stranded templates anchored to the substrate.
6. COMPLETE AMPLIFICATION
Clusters
Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.
10. IMAGE SECOND CHEMISTRY CYCLE
A
C
T
G
After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.
11. SEQUENCE READS OVER MULTIPLE CHEMISTRY CYCLES
GCTGA...
Repeat cycles of sequencing to determine the sequence of bases in a given fragment a single base at time.
12. ALIGN DATA
Reference sequence
..GCTGATGTGCCGCCTCACTCCGGTGG
CACTCCTGTGG
CTCACTCCTGTGG
GCTGATGTGCCACCTCA
GATGTGCCACCTCACTC
GTGCCGCCTCACTCCTG
CTCCTGTGG
Unknown variant identified and called
Known SNP called
Align data, compare to a reference, and identify sequence differences.

Illumina reads

Paired-end sequencing



Life Technologies SOLiD

Instrument	Yield and run time	Read Length	Error rate	Error type
SOLiD 5500 wildfire	*600 GB, 8 days*	*75x35 PE* *60x60 MP*	*0.01%*	A-T Bias

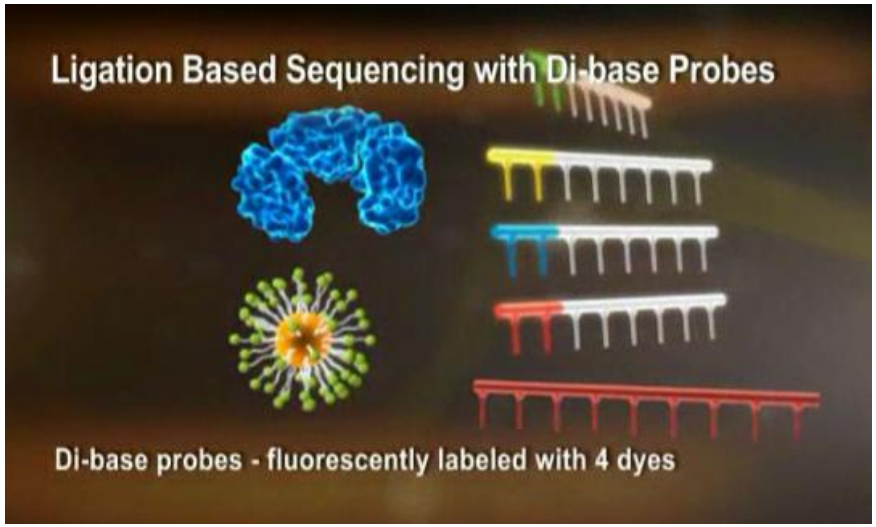
Features

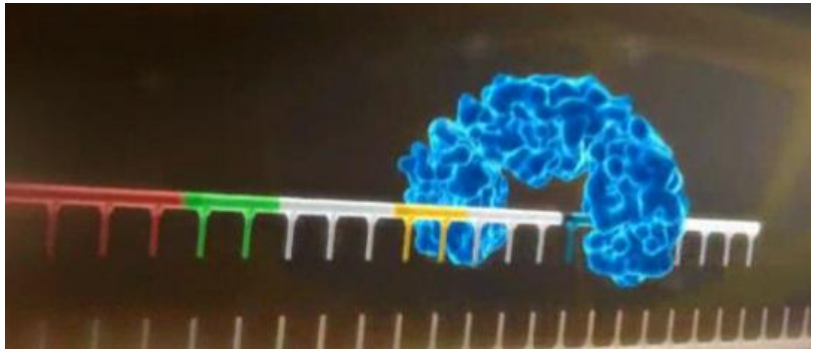
- High accuracy due to two-base encoding
- True paired-end chemistry - ligation from either end
- Mate-pair libraries

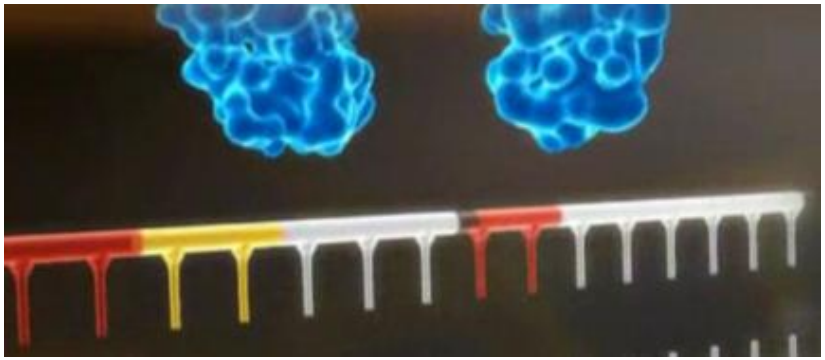
Main applications (currently)

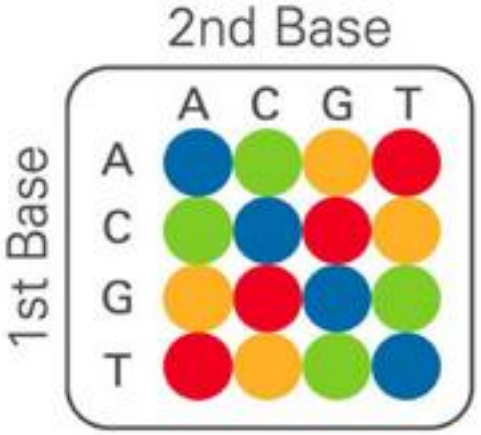
- ChiPSeq

SOLiD - ligation







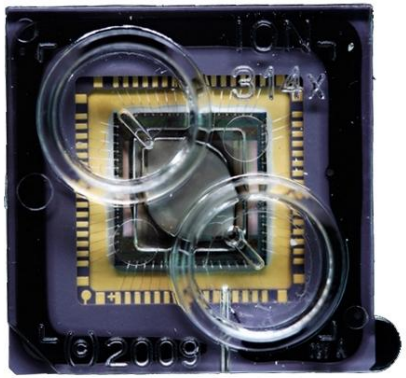


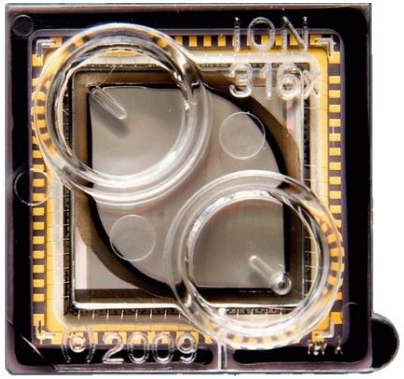
Life Technologies - Ion Torrent & Ion Proton

Chip	Yield - run time	Read Length
PGM 314	0.1 GB, 3 hrs	200 – 400
PGM 316	0.5GB, 3 hrs	200 - 400
PGM 318	1 GB, 3 hrs	200 - 400
P-I	10 GB	200

Main applications

- Microbial and metagenomic sequencing
- Targeted resequencing
- Clinical sequencing

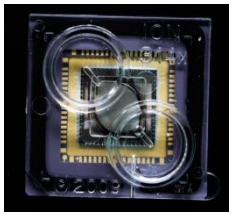


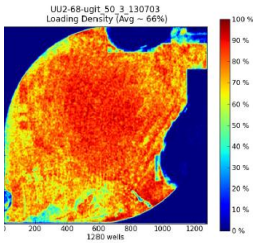


314 chip	316 chip	318 chip	PI chip
10 Mb	**100 Mb**	**1 Gb**	**10 Gb**
	200 – 400 bp		**200 bp**
virus, bacteria, small eukaryote			**eukaryote**

IonTorrent Throughput - **400bp**

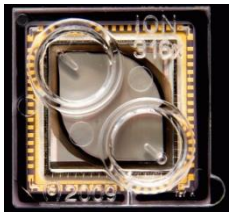
314 chip (10 Mbp)

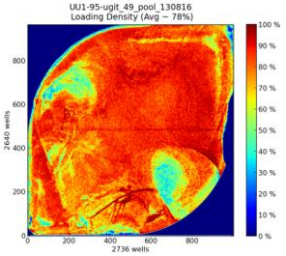




Total Number of Bases [Mbp]	224.64
▸ Number of Q20 Bases [Mbp]	39.50
Total Number of Reads	531,758
Mean Length [bp]	422
Longest Read [bp]	2,676

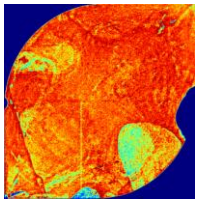
316 chip (100 Mbp)





Total Number of Bases [Mbp]	707.33
▸ Number of Q20 Bases [Mbp]	548.84
Total Number of Reads	2,933,870
Mean Length [bp]	241
Longest Read [bp]	619

318 chip (1 Gbp)

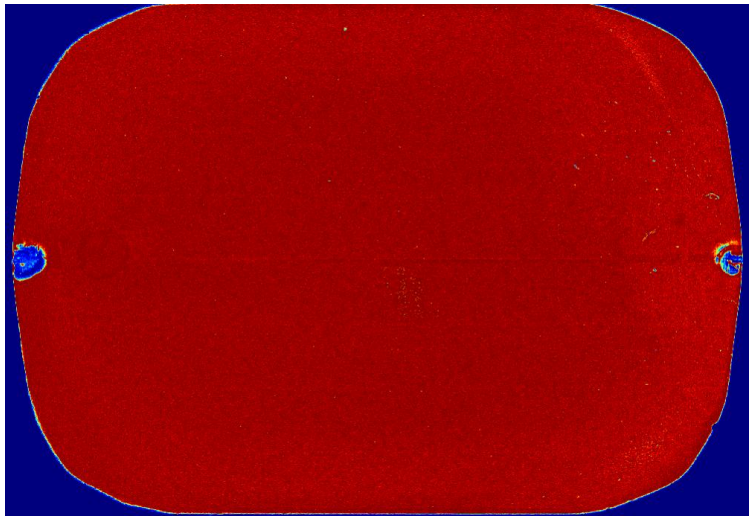


Total Number of Bases [Mbp]	863.08
▸ Number of Q20 Bases [Mbp]	667.99
Total Number of Reads	4,417,950
Mean Length [bp]	195
Longest Read [bp]	682

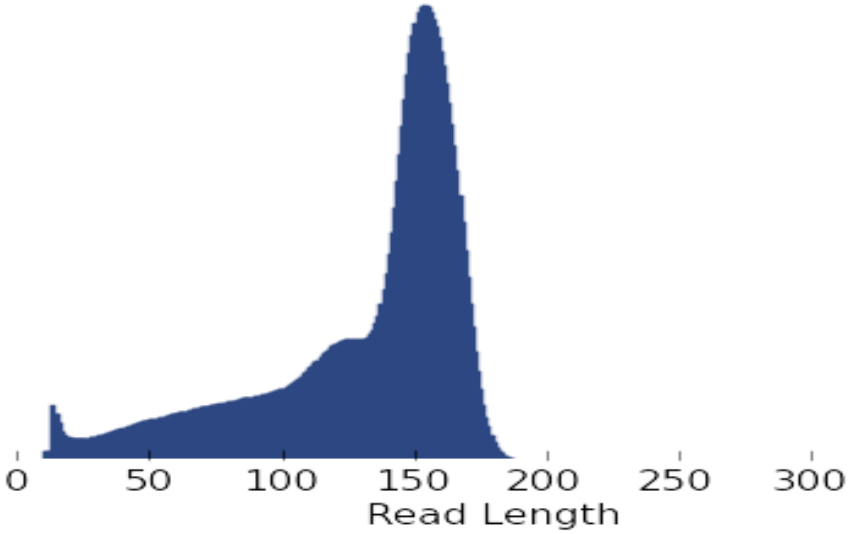
Ion Proton - Throughput

- We now get 10-16GB data from the PI chip

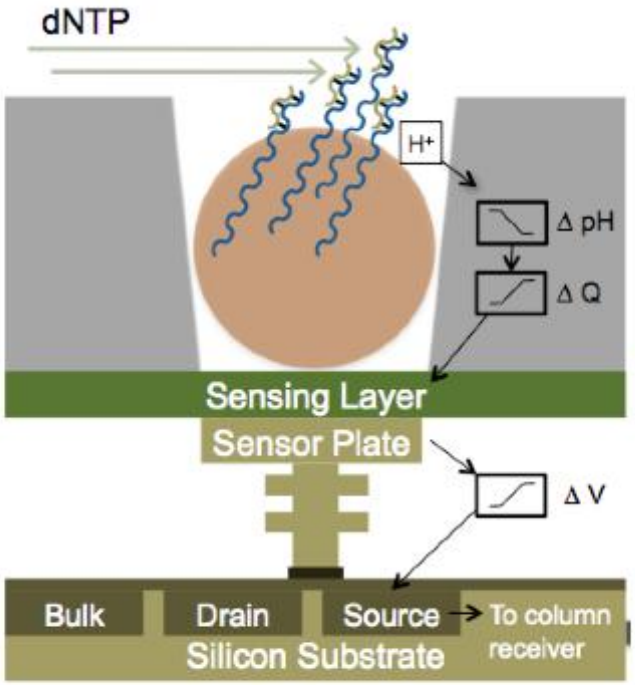
> 90M reads

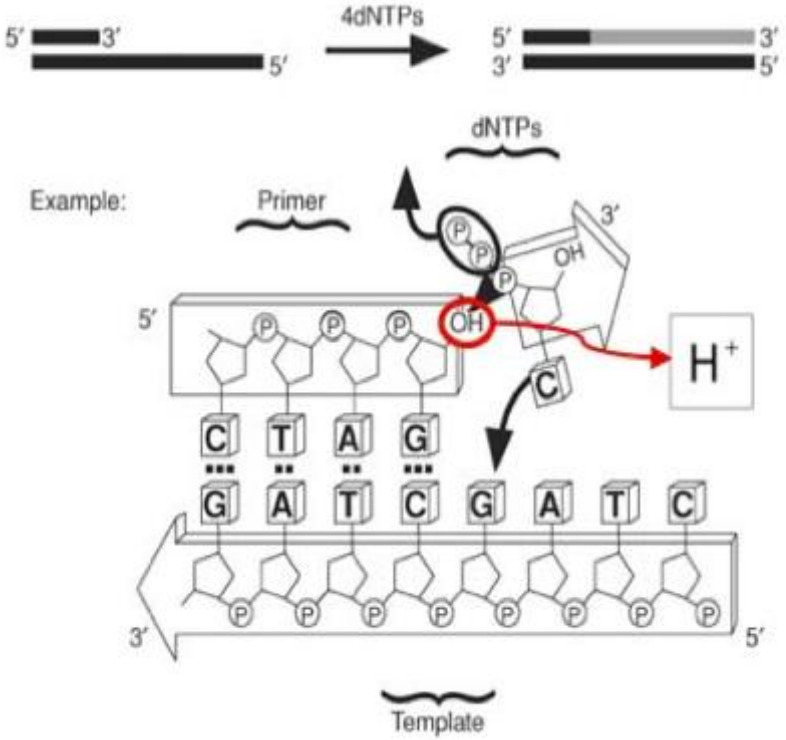


~ 150bp read length



Ion Torrent - H^+ ion-sensitive field effect transistors

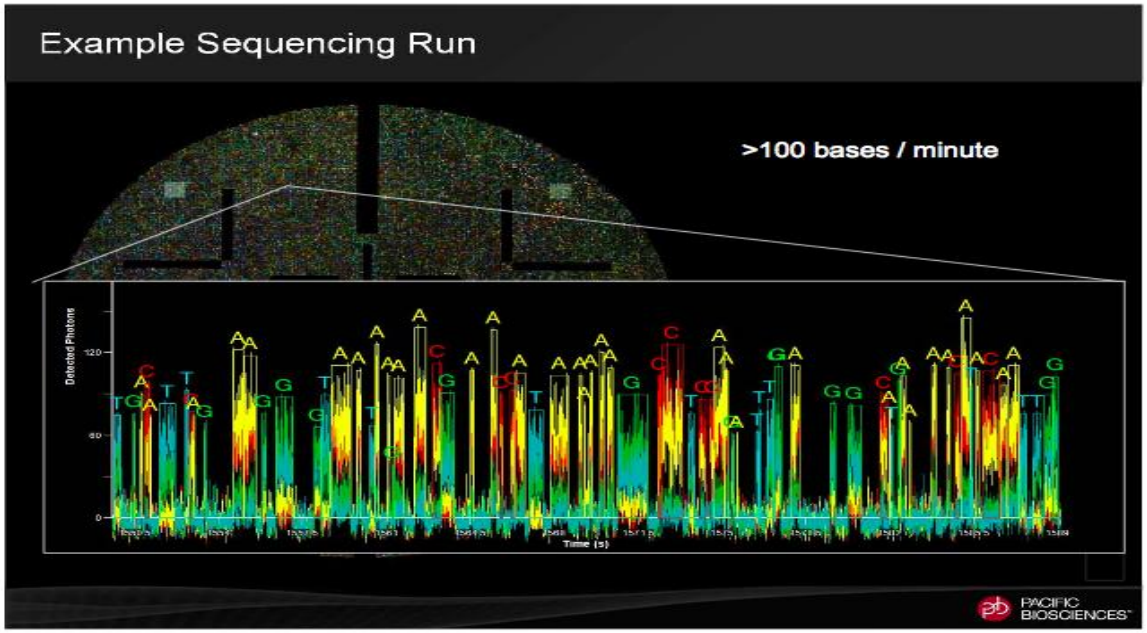




Pacific Bioscience

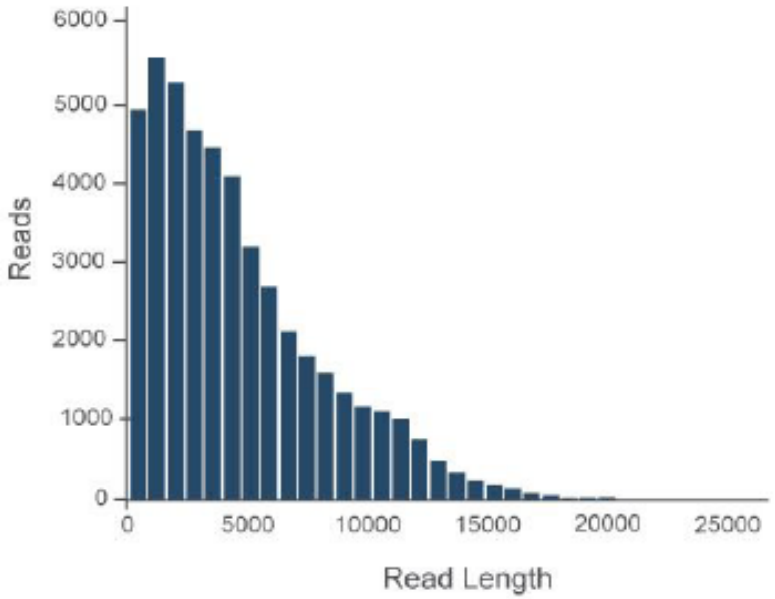
Instrument	Yield and run time	Read Length	Error rate	Error type
RS II	500 Mb – 1.3 Gb /180 - 240 min SMRTCell	250 bp – 20 000 bp (50 000 bp)	15% (on a single passage!)	Insertions, **random**

Single-Molecule, Real-Time DNA sequencing



Typical PacBio® RS II Results

Read Length Distribution

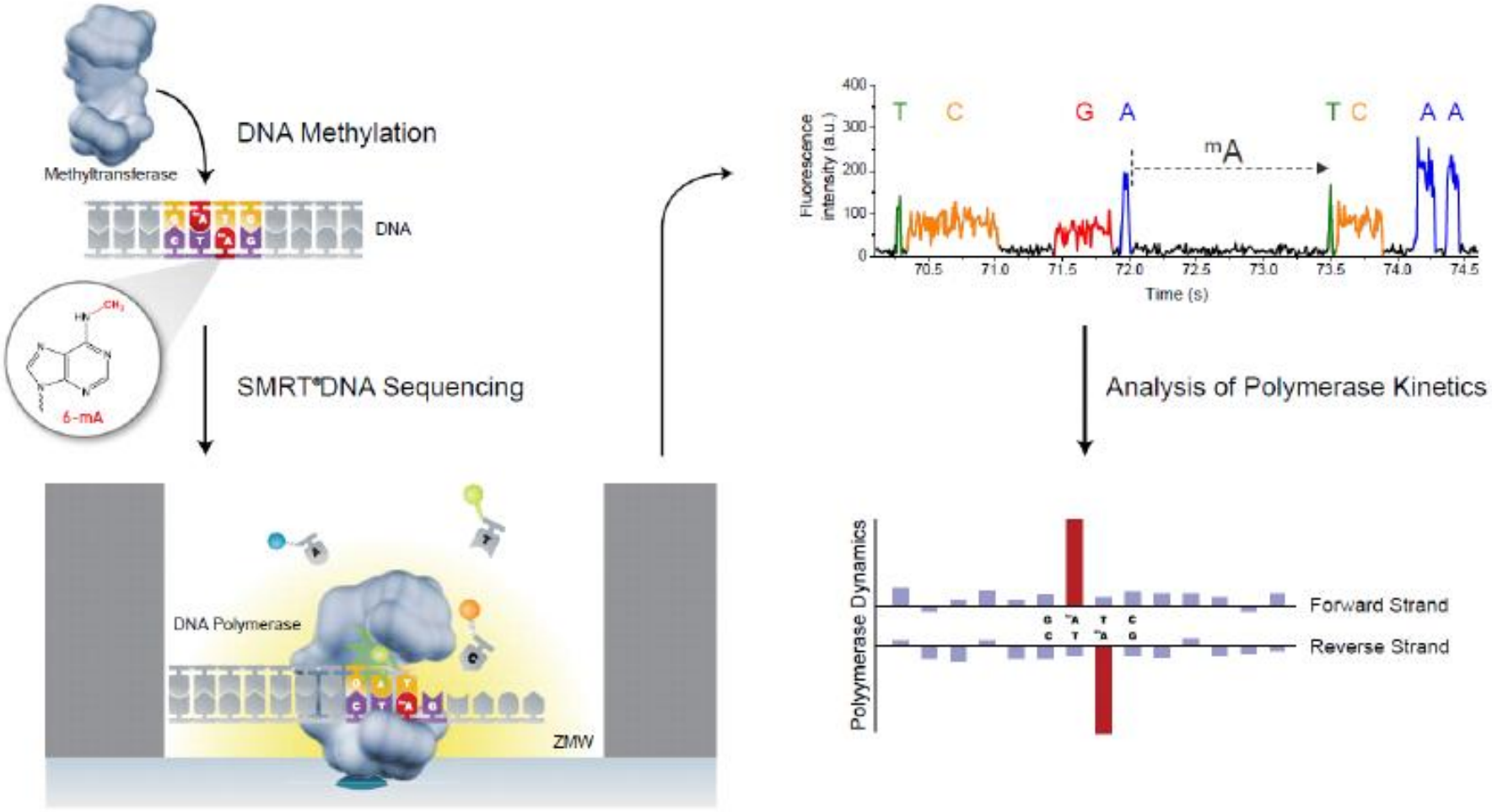


Typical Results

Read Length:	
Average:	4,606 bp
95th Percentile:	11,792 bp
Maximum:	23,297 bp
Throughput per SMRT® Cell:	216 Mb
	47,197 reads

Based on data from 11 kb plasmid library using a 120 minute movie

Base Modification: Discover the Epigenome



Detect base modifications using the kinetics of the polymerization reaction during normal sequencing

Improve and Finish Genomes with the PacBio® System

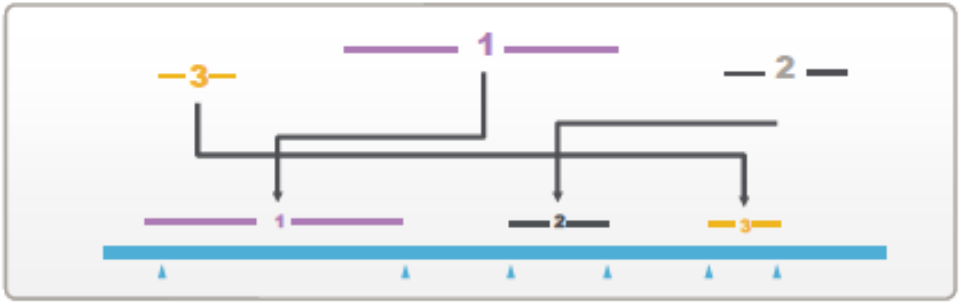
De novo Assembly

Complete genomes with PacBio reads alone

Combine technologies for best of both worlds

Scaffold

Establish framework for genome and resolve ambiguities



Span Gaps

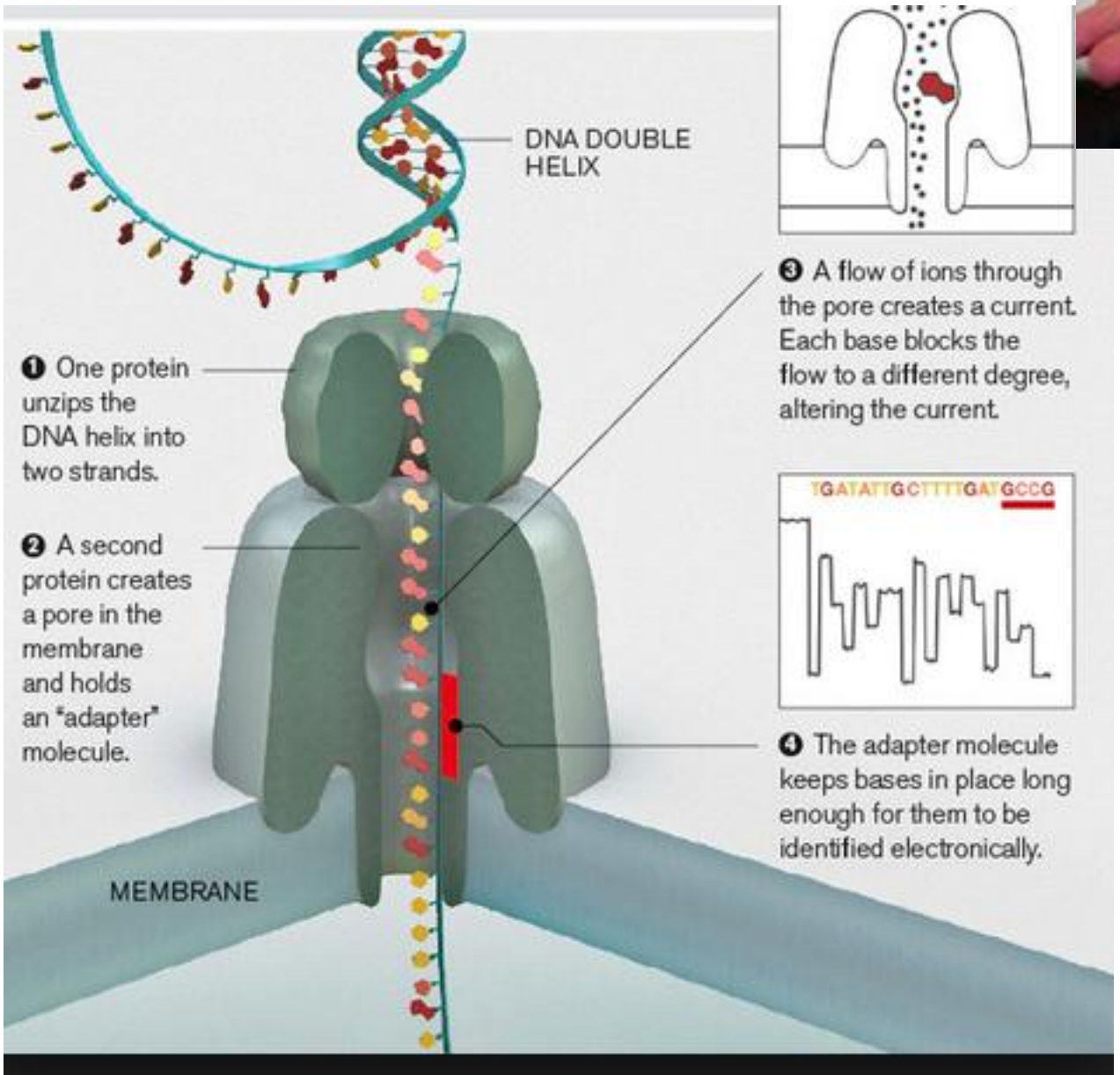
Polish genomic regions with up to 10x improvement

Long-Read Single-Molecule Sequencing at NGI - SciLifeLab

March 17-18
Navet, BMC
Uppsala

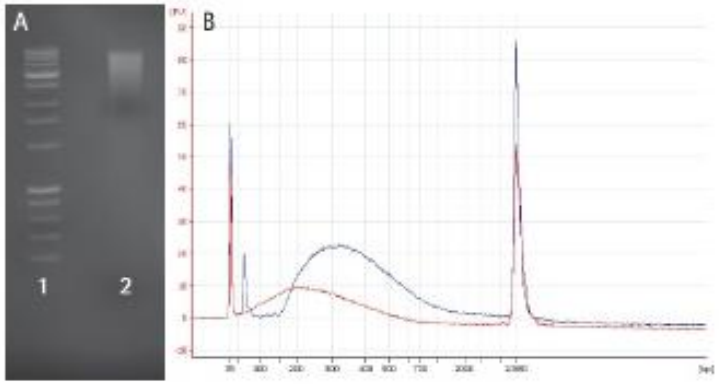
Uppsala Genome Center

Oxford Nanopore MinION

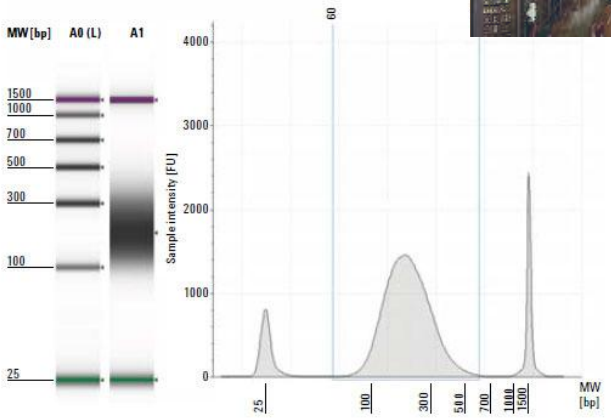


Reads up to 100k
1D and 2D reads
15-40% error rate
Life time 5 days

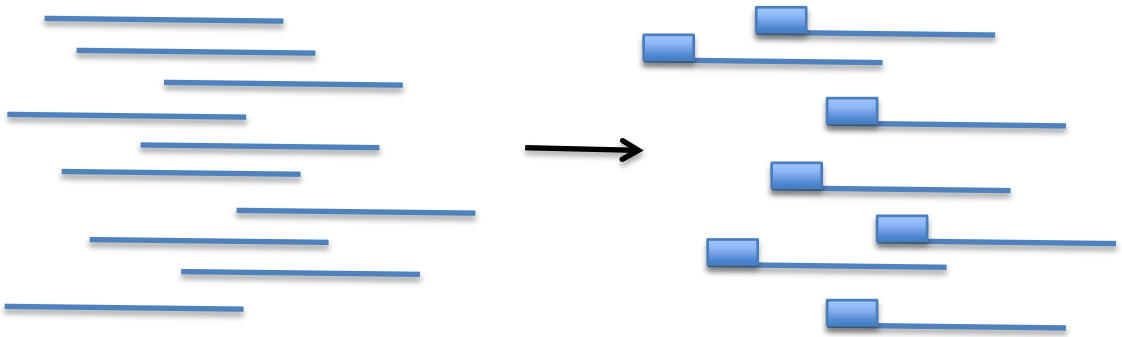
Making a NGS library



DNA QC – **paramount importance**



Sharing & size selection



Amplification

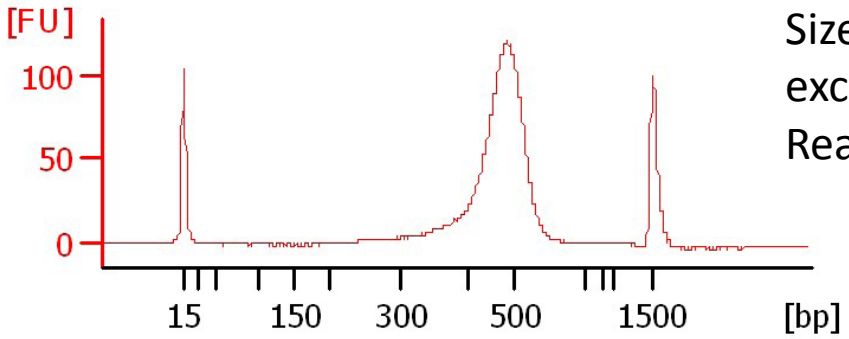
Ligation of sequencing adaptors, technology specific

Input QC control at NGI:

- Qubit for DNA
  - Measures content of dsDNA only
  - Nanodrop & NanoVue overestimate concentrations up to 300%!
- Bioanalyzer for RNA and amplicons
  - RNA: RIN values and concentrations
  - Amplicons: size distribution (extremely important!)

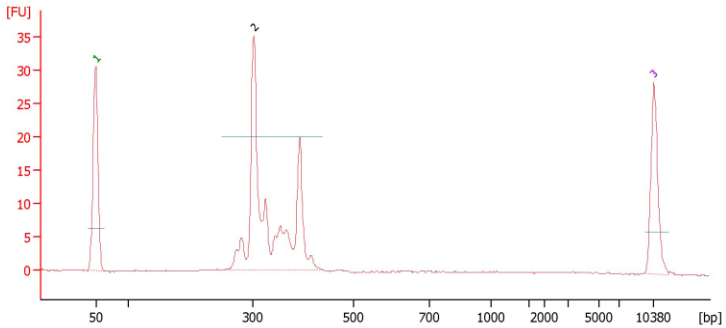
Bioanalyzer: amplicon size check

Example 1: OK size distribution

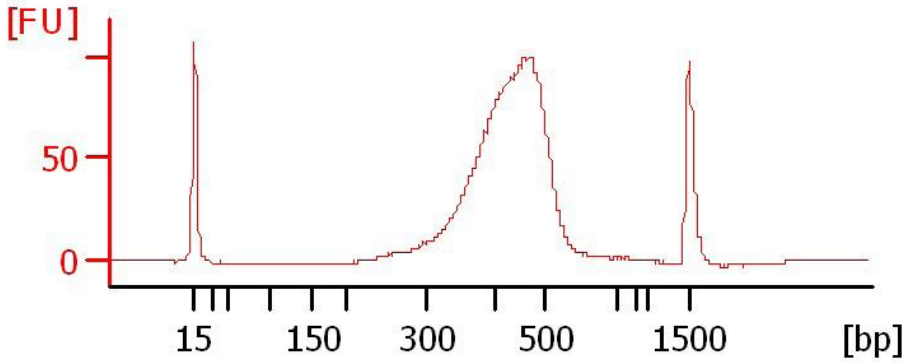


FOR ANY NGS TECHNOLOGY

Size difference among fragments **must not** exceed 80 bp (optimally 50 bp)
Reason – preferential amplification of short fragments



Example 2: several sizes,
fractionation is needed
=> we HAVE to make several libraries



Example 3: broad peak;
size selection is needed

NGS technologies - SUMMARY

Platform	Read length	Accuracy	Projects / applications
454	Medium	Homo- polymer runs	Microbial + targeted reseq
HiSeq MiSeq	Short Medium	High	Whole genome + transcriptome seq, exome
SOLiD	Short	High	Whole genome + transcriptome seq, exome
Ion Torrent	Medium	High	Microbial + targeted reseq
Ion Proton	Short/Mediu m	High	Exome, transcriptome, genome
PacBio	Long	Low – ultra high*	Microbial + targeted reseq Gap closure & scaffolding
MinION	Long	Low	Gap closure, scaffolding structural variants

What is The BEST?

	Illumina HiSeq	Illumina MiSeq	SOLiD Wildfire	Ion Torrent	Ion Proton	PacBio
Read length	100 + 100 bp (150+150 bp)	250 + 250 bp (350+350 bp)	75 bp	200 bp 400 bp (500 bp)	150 bp 200 bp	250 bp – 40 Kbp
WGS: - human - small	 ++++ +++	 ++++	 (+) (+)	 ++++	 + +++	 (+) +++++
De novo	+++	++		+++	++	+++++
RNA-seq miRNA	+++ +++		+++ +++		+++	+++*
ChIP	+++		++++			
Amplicon	++	+++		+++	+++	+++
Metylation	+++					++++*
Target re-seq	++	+++	(+)		+++	+++
Exome	+++		(+)		++++	(+)

Check list:

- Have others done similar work?
- Is your **methodology** sound? Sample size? Repetitions?
- Is there **people** to analyze the data?
- Is there **computer capacity** to analyze the data?
- Will you be able to **publish** NGS data by yourself?
- **PLEASE consult the sequencing facility PRIOR to onset of your project!**

Common pitfalls and a piece of advise:

- If you give us low quality DNA/RNA - expect low quality data
- If you give us too little DNA/RNA – expect biased data

- Do not try to do everything by yourself
- Make sure there is a dedicated bioinformatician available
- Never underestimate time and money needed for data analysis

- Google often!
- Use online forums, e.g. SeqAnswers.com

Summary

- Progress is FAST- keep yourselves updated!

- Chose technology based on:
  - What is most feasible
  - What is most accessible
  - What is most cost-effective

SciLifeLab Genomics & Bioinformatics are here for you!

SciLifeLab TECHNOLOGIES & SERVICES RESEARCH EDUCATION COLLABORATION

Find more information and search for what you need on the page for Technologies & Services

What is the difference between national and regional facilities?

Search for Technologies & Services

National facilities

Affinity Proteomics
Biobank Profiling
Cell Profiling
Fluorescence Tissue Profiling
PLA Proteomics
Protein and Peptide Arrays
Tissue Profiling

Bioimaging
Advanced Light Microscopy
Fluorescence Correlation Spectroscopy

Bioinformatics
Bioinformatics Compute and Storage (UPPNEX)
Bioinformatics Long-term Support (WABI)
Bioinformatics Short-term Support and Infrastructure (BILS)

Chemical Biology Consortium Sweden
Laboratories for Chemical Biology Umeå (LCBU)
The Laboratories for Chemical Biology at Karolinska Institutet (LCBKI)
Uppsala Drug Optimization and Pharmaceutical Profiling (UDOPP)

Clinical Diagnostics
Clinical Biomarkers
Clinical Genomics
Clinical Sequencing

Drug Discovery and Development
ADME (Absorption Distribution, Metabolism Excretion) of Therapeutics (UDOPP)
Biochemical and Cellular Screening
Biophysical Screening and Characterization
Human Antibody Therapeutics
In Vitro and Systems Pharmacology
Medicinal Chemistry – Hit2Lead
Medicinal Chemistry – Lead Identification
Protein Expression and Characterization

Functional Genomics
Karolinska High Throughput Center (KHTC)

National Genomics Infrastructure
NGI Stockholm (Genomics Applications)
NGI Stockholm (Genomics Production)
NGI Uppsala (SNP&SEQ Technology Platform)
NGI Uppsala (Uppsala Genome Center)

Structural Biology
Protein Science Facility

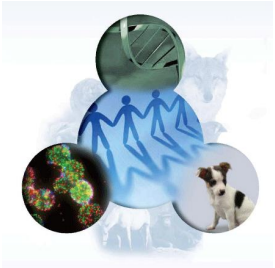
National Genomics Infrastructure

SciLifeLab, Stockholm

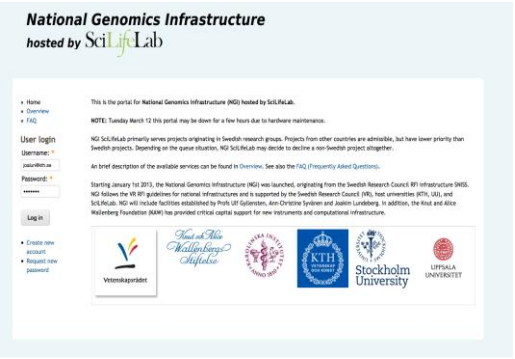
SciLifeLab, Uppsala

Uppmax, Uppsala

Vetenskapsrådet



Portal project flow

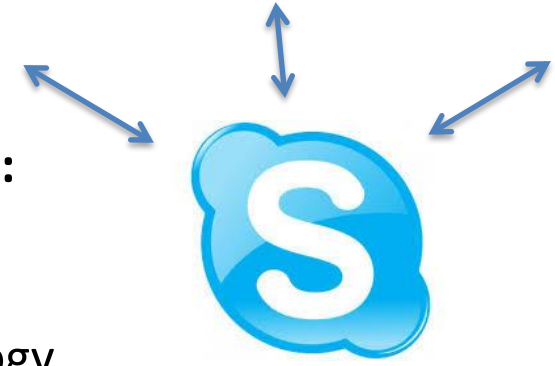


NGI Project coordinators meet every second day via Skype

Ulrika Liljedahl
SNP&SEQ
Uppsala node

Mattias Ormestad
Stockholm Node

Olga Vinnere Pettersson
UGC
Uppsala Node



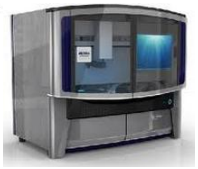
Project distribution is based on:

1. Wish of PI
2. Type of sequencing technology
3. Type of application
4. Queue at technology platforms

Project is then assigned to a certain node and a coordinator contacts the PI

NGI Equipment

Illumina HiSeq 2000/2500	17
Illumina MiSeq	3
Life Technologies SOLiD 5500wildfire	1
Life Technologies Ion Torrent	2
Life Technologies Ion Proton	6
Life Technologies Sanger ABI3730	2
Pacific Biosciences RSII	2
Argus Whole Genome Mapping System	1



One of 5 best-equipped NGS sites in Europe

Project meeting

What we can help you with:

- Design your experiment based on the scientific question.
- Chose the best suited application for your project.
- Find the most optimal sequencing setup.
- Answer all questions about our technologies and applications, as well as bioinformatics.

- Get UPPNEX account if you do not have one.

- In special cases, we can give extra-support with bioinformatics analysis – development of novel methods and applications

Downstream Data Analysis