NGI, SciLife, NGS, barcoding

and some other stuff

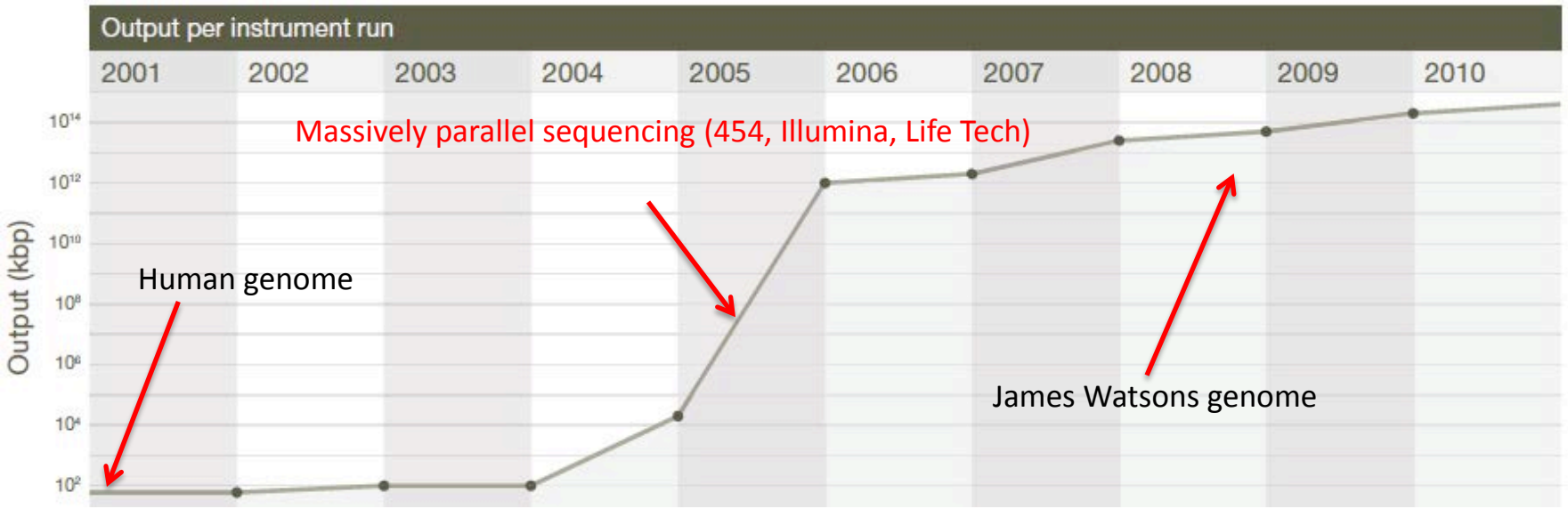
Olga Vinnere Pettersson, PhD
National Genomics Infrastructure hosted by ScilifeLab,
Uppsala Node (UGC)

Ouline

- SciLifeLab & National Genomics Infrastructure (NGI) – structure, mission and services
- Next Generation Sequencing – main technologies
- Sample prep, DNA libraries and barcoding possiblities
- Bioinformatics support at SciLifeLab
- Answers to pre-asked questions

SciLifeLab & National Genomics Infrastructure (NGI) – structure, mission and services

Background – Massively Parallel DNA Sequencing



1998 KTH Genome Center
1996 Uppsala Genome Center
1998 SNP&SEQ Center

SSF Genome Center
Wallenberg Consortium North

Center for Metagenomic
Sequence Analysis (KAW)

VR RFI National Genomics Infrastructure
(former VR RFI SNISS; 1 Jan 2010)

Science for Life
Laboratory (SciLifeLab)

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

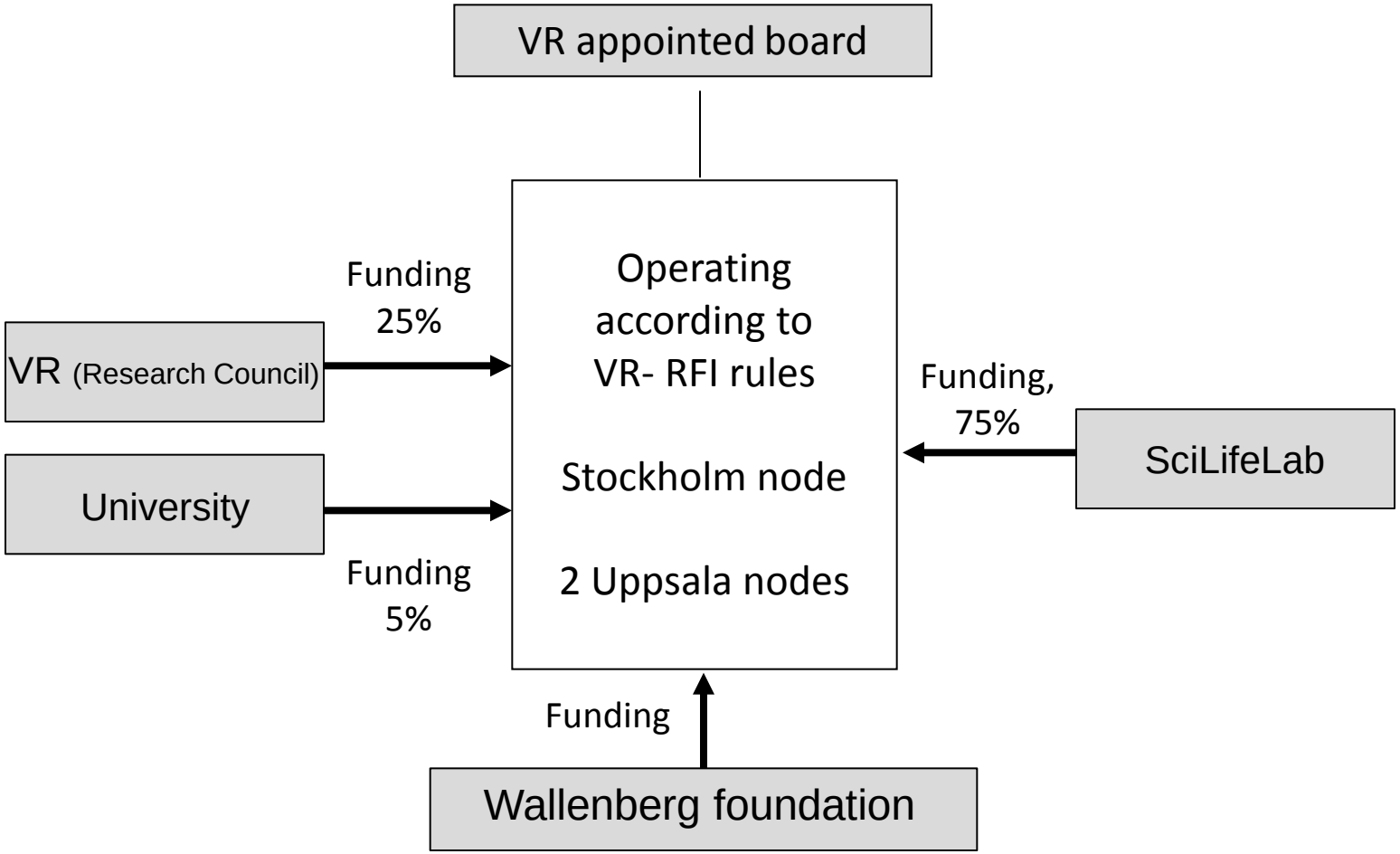
National Genomics Infrastructure - Goals

Criteria for infrastructures (VR RFI guidelines):

- Being of broad national interest
- Provide essential opportunities for world-leading research
- Can be used by several research groups/users on highly qualified research projects
- Be so comprehensive that individual groups cannot run them independently
- Have a long-term plan for scientific aims, funding and utilisation
- Be openly and easily available to researchers and have a plan for how availability can be improved (applies both to use of the infrastructure, access to data collected and the representation of results)

Joint organisational structure

Old Name: Swedish National Infrastructure for large-Scale Seq, SNISS
New name: National Genomics Infrastructure (1 January, 2013)



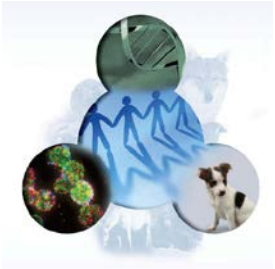
National Genomics Infrastructure

SciLifeLab, Stockholm

SciLifeLab, Uppsala

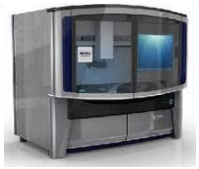
Uppmax, Uppsala

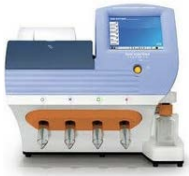
Vetenskapsrådet



NGI Equipment

Illumina HiSeq 2000/2500	17
Illumina Xten	10
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 handling process

Research group

(1) New Users

(2) Large Projects

(3) Frequent User

Research group + platform managers

Project Planning Meeting
(1) Feasability
(2) Study design
(3) Choice of technology
(4) Bioinformatics support

External review panel
- Prof Ivo Gut, Barcelona
- Prof Svante Pääbo, Leipzig

Platform manager

Project Contract

Sequencing and bioinformatics personnel

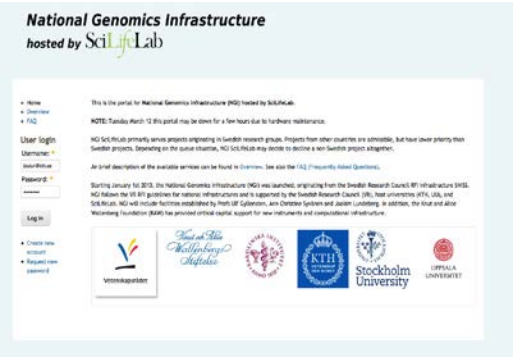
Sample QC, Sequencing, Bioinformatics (primary)

Data delivery to UPPNEX

Research group + platform managers

Follow-up meeting if needed

Point of entrance: https://portal.scilifelab.se/genomics/

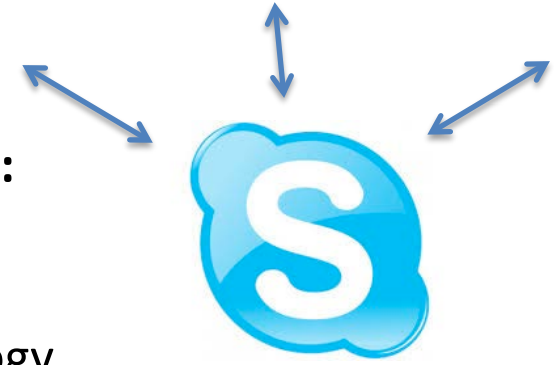


NGI Project coordinators meet every second day via Skype

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

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.

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

Next Generation Sequencing – main technologies

Once upon a time…

- Fredrik Sanger and Alan Coulson

Chain Termination Sequencing (1977)

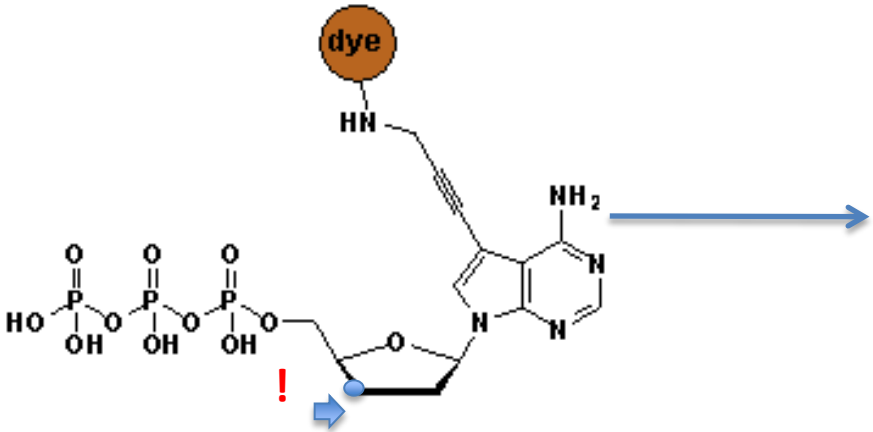
Nobel prize 1980

Principle:

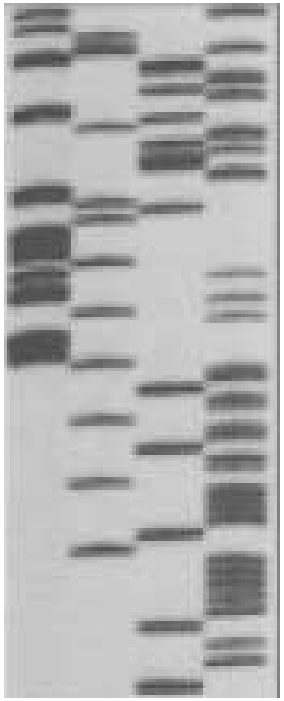
SYNTHESIS of DNA is randomly TERMINATED at different points

Separation of fragments that are 1 nucleotide different in size

Sanger's sequencing

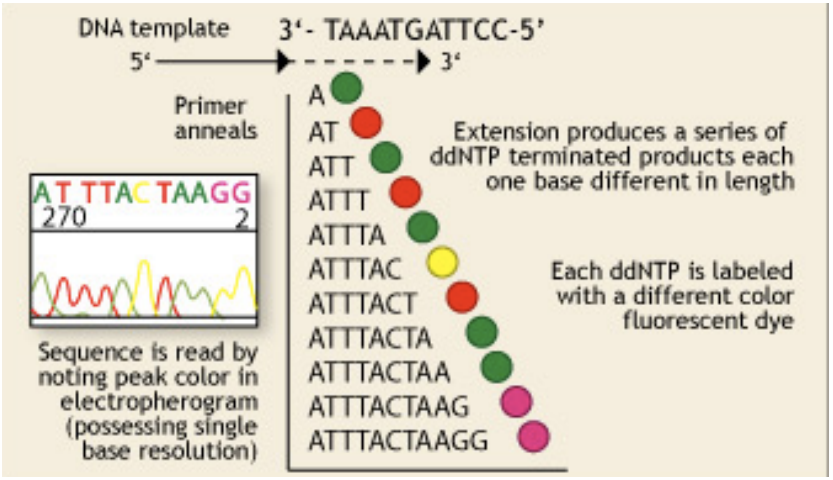


P^{32} labelled ddNTPs



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

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 > 40 Kbp
- Accuracy per base from 0.1% to 15%
- Cost per base varies

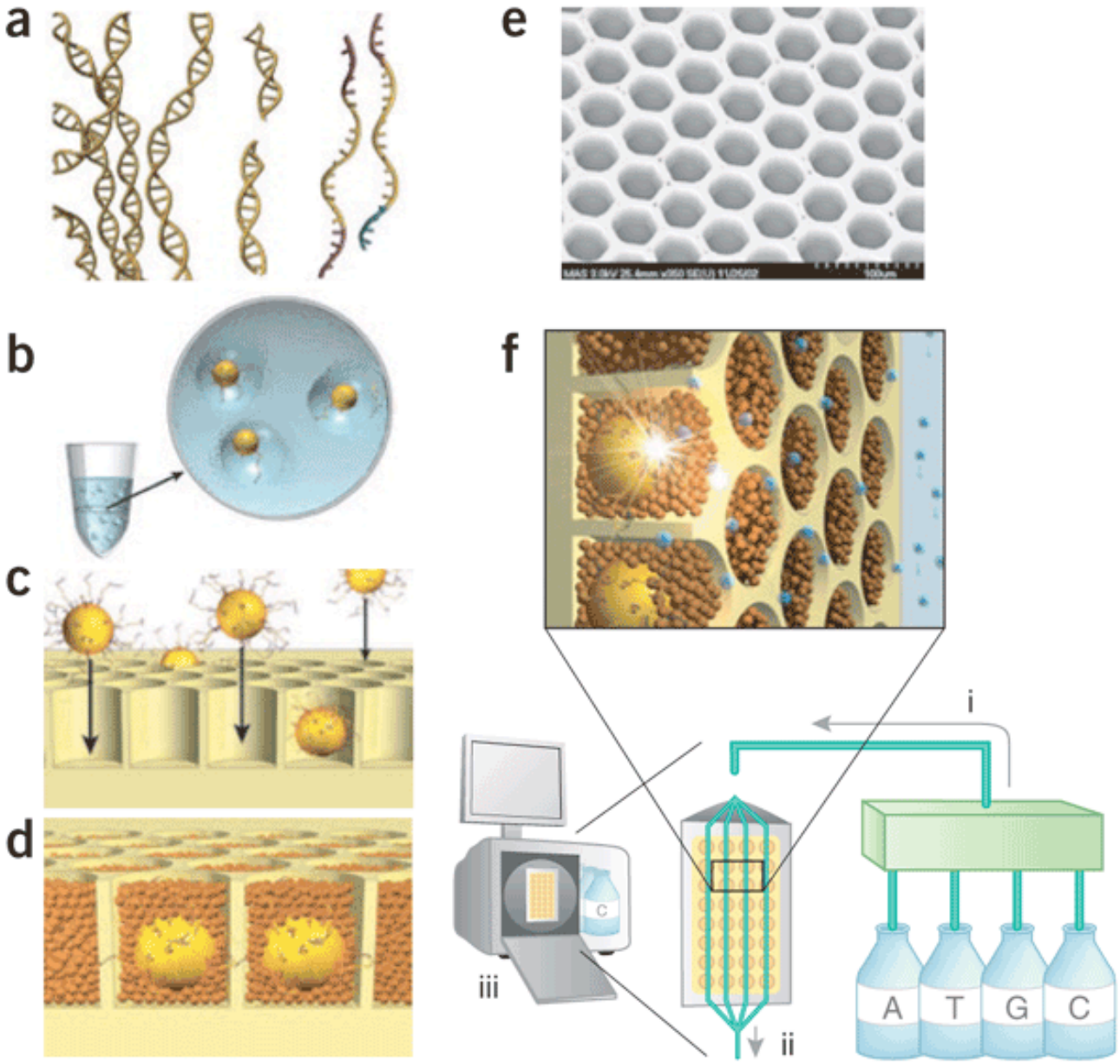
Roche

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

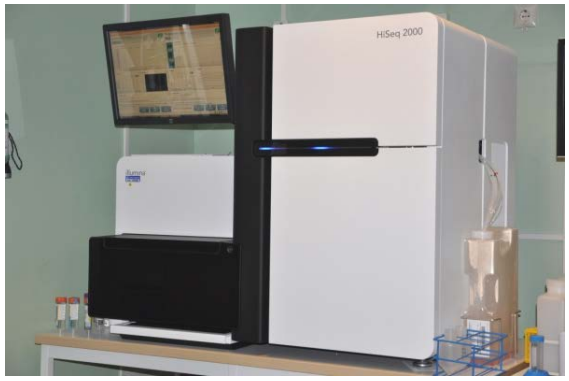


Illumina

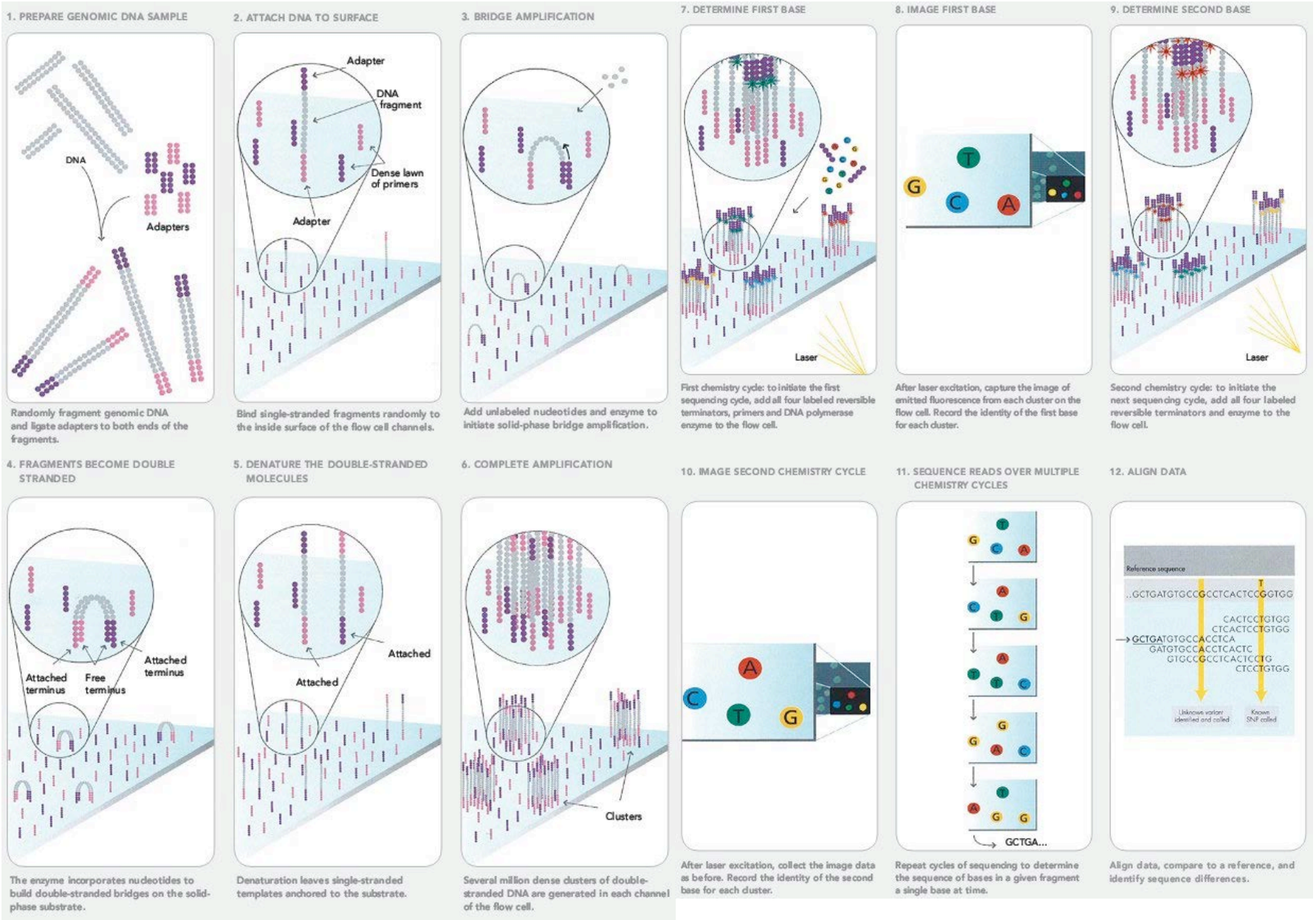
Instrument	Yield and run time	Read Length (bp)	Comment
HiSeq2500, high output	Up to 1.0 Tbases (6 days) or *Up to 120 Gbases (40-70h)*	2x50 2x125 *2x100 2x250*	A lot of data
HiSeq X Ten	Up to 1.6 Tbases (3 days)	2x150	Human genomes
MiSeq	540 Mb – 15 Gbases (4 – 48 hours)	2x75 to 2x300	Low yield but fast

Main applications

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

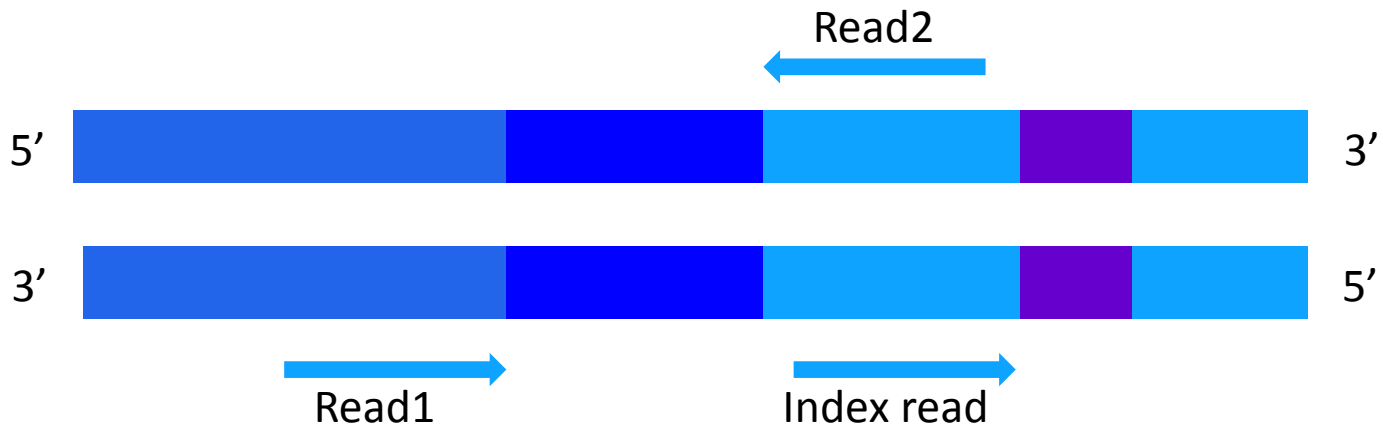


Illumina


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

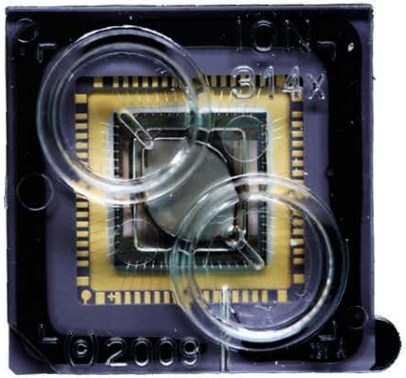


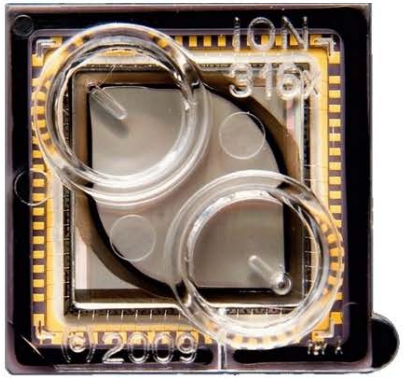
Life Technologies - Ion Torrent & Ion Proton

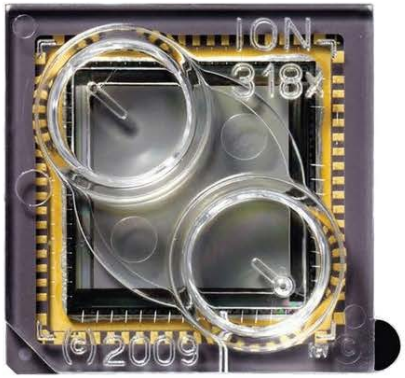
Chip	Yield - run time	Read Length
PGM 314	0.1 GB, 3 hrs	200 – 500
PGM 316	0.5GB, 3 hrs	200 - 500
PGM 318	1 GB, 3 hrs	200 - 500
P-I	10 - 18 GB	200

Main applications

- Microbial and metagenomic sequencing
- Targeted resequencing
- Clinical sequencing



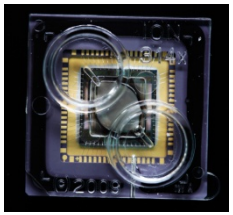


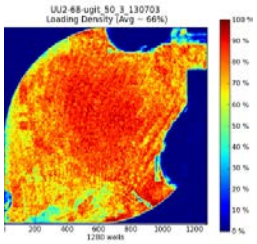


314 chip	316 chip	318 chip	PI chip
10 Mb	**100 Mb**	**1 Gb**	**10 Gb**
	200 – 500 bp		**200 bp**
virus, bacteria, small eukaryote			**eukaryote**

IonTorrent Throughput - **400bp**

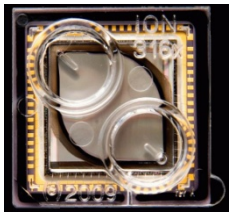
314 chip (10 Mbp)

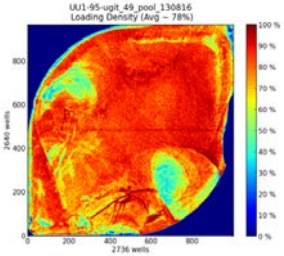




Total Number of Bases [Mbp]	224.64
▸ Number of Q20 Bases [Mbp]	39.50
Total Number of Reads	531,758
Mean Length [bp]	422
Longest Read [bp]	2,676

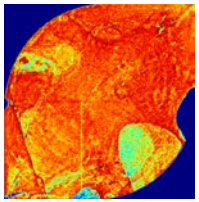
316 chip (100 Mbp)





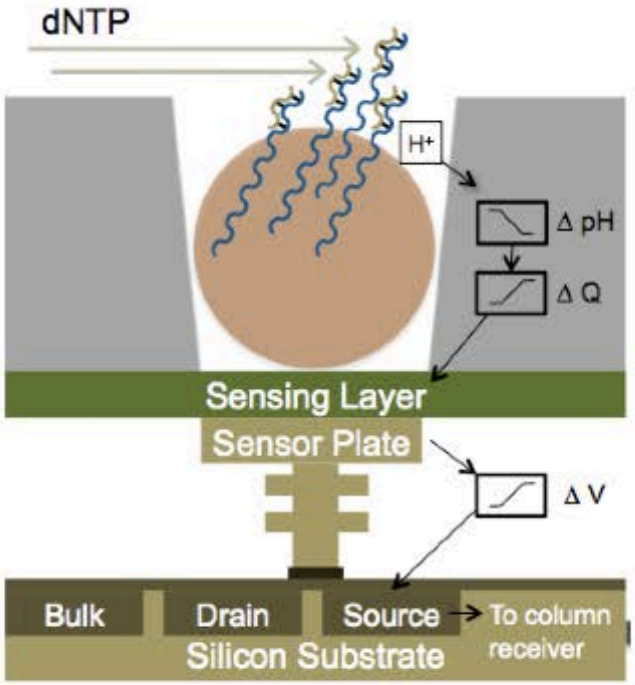
Total Number of Bases [Mbp]	707.33
▸ Number of Q20 Bases [Mbp]	548.84
Total Number of Reads	2,933,870
Mean Length [bp]	241
Longest Read [bp]	619

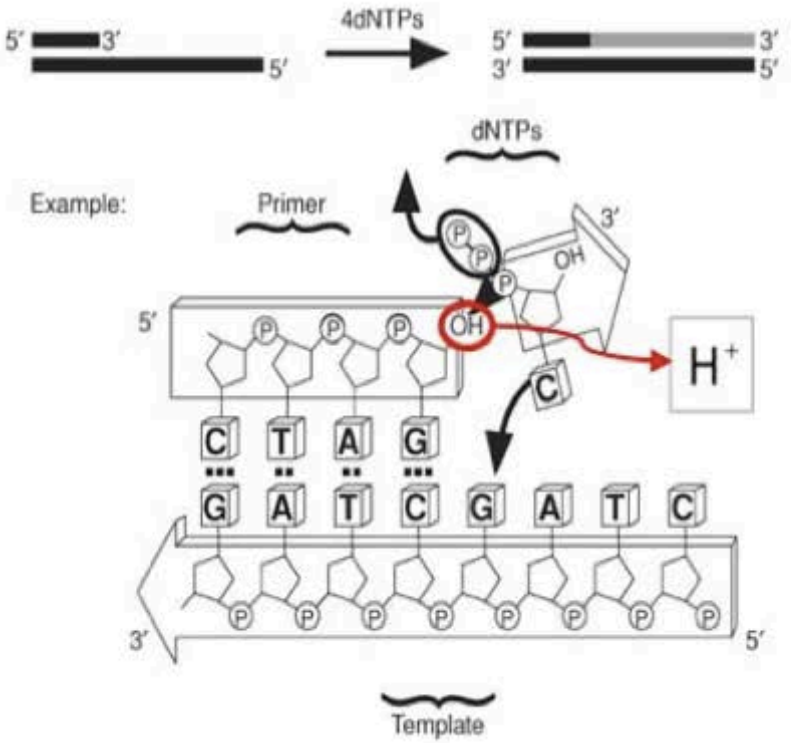
318 chip (1 Gbp)



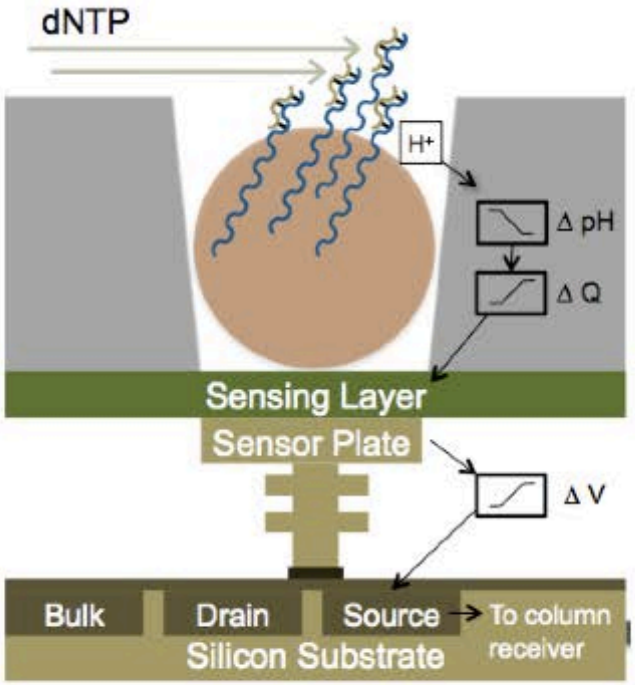
Total Number of Bases [Mbp]	863.08
▸ Number of Q20 Bases [Mbp]	667.99
Total Number of Reads	4,417,950
Mean Length [bp]	195
Longest Read [bp]	682

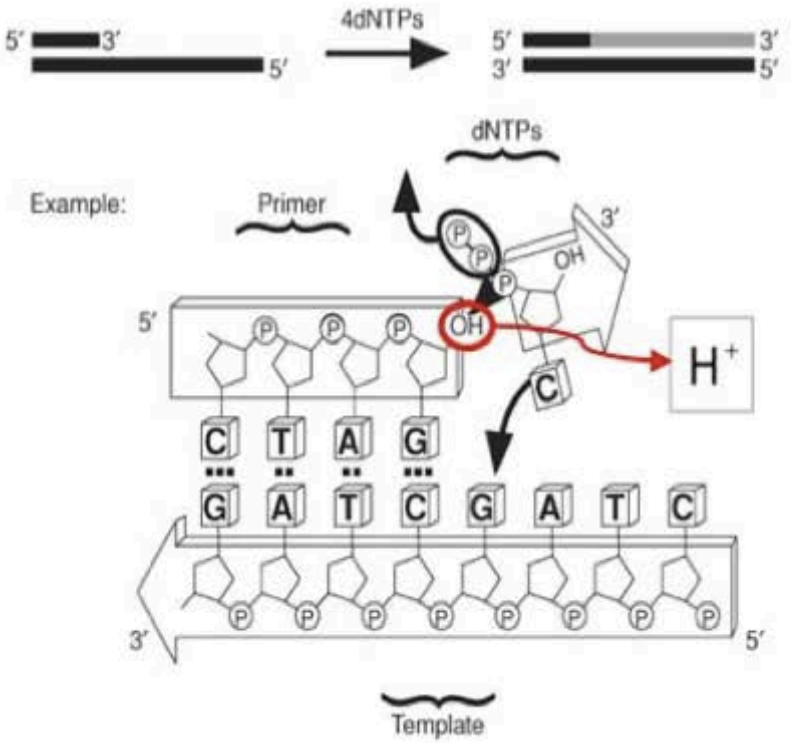
Ion Torrent - H^+ ion-sensitive field effect transistors





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

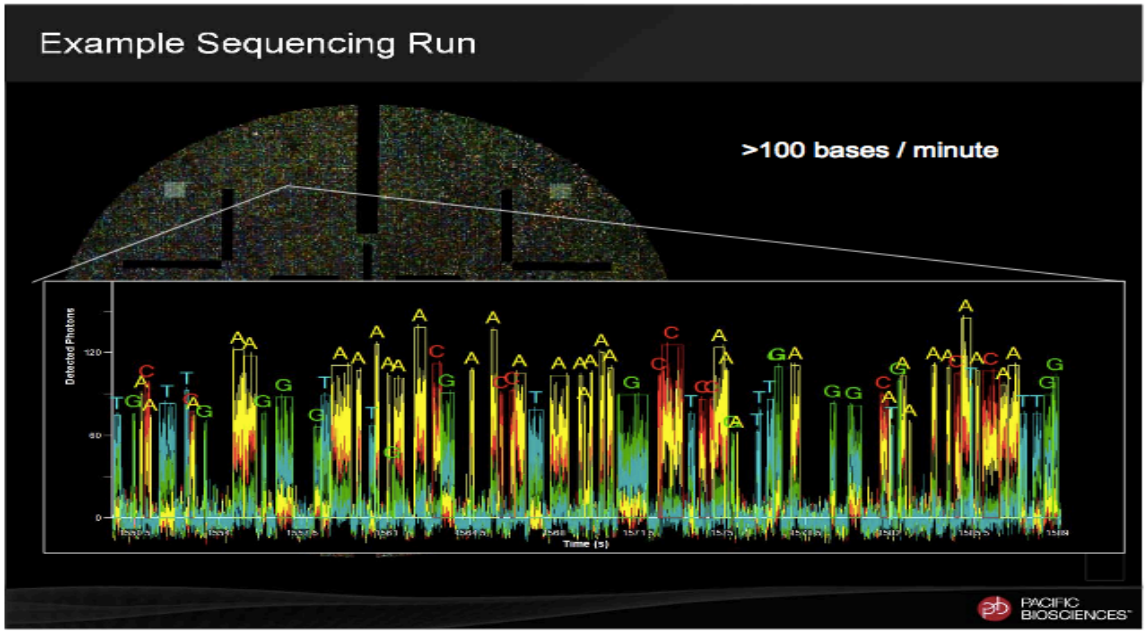




Pacific Bioscience

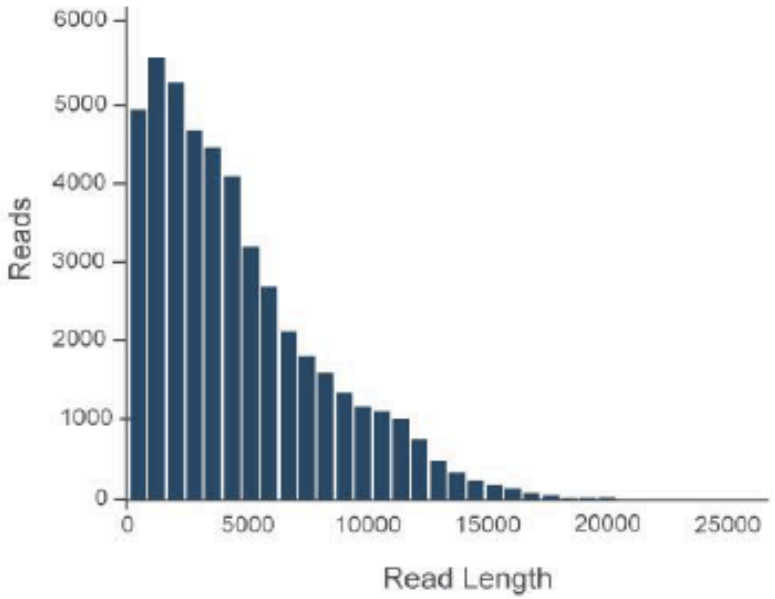
Instrument	Yield and run time	Read Length	Error rate	Error type
RS II	500 Mb – 1.3 Gb /180 - 240 min SMRTCell	250 bp – 20 000 bp (40 000 bp)	15% (on a single passage!)	Insertions, **random**

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

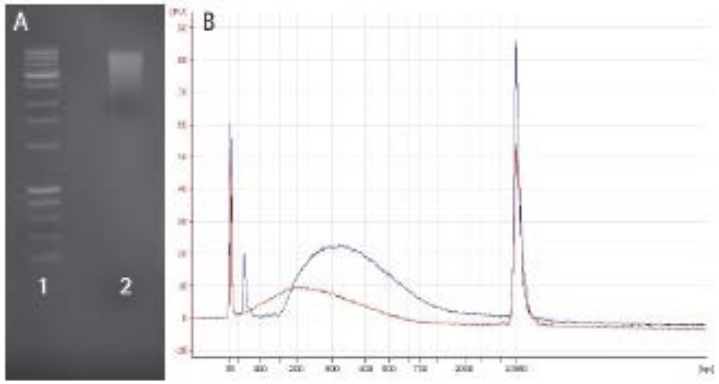
NGS technologies - SUMMARY

Platform	Read length	Accuracy	Projects / applications
454	Medium	Homo- polymer runs	Microbial + targeted reseq
HiSeq MiSeq	Short Medium	High	Whole genome + transcriptome seq, exome
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

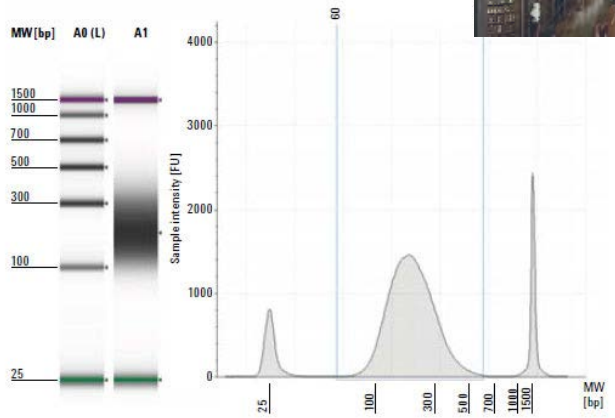
What is The BEST?

Sample prep and DNA libraries

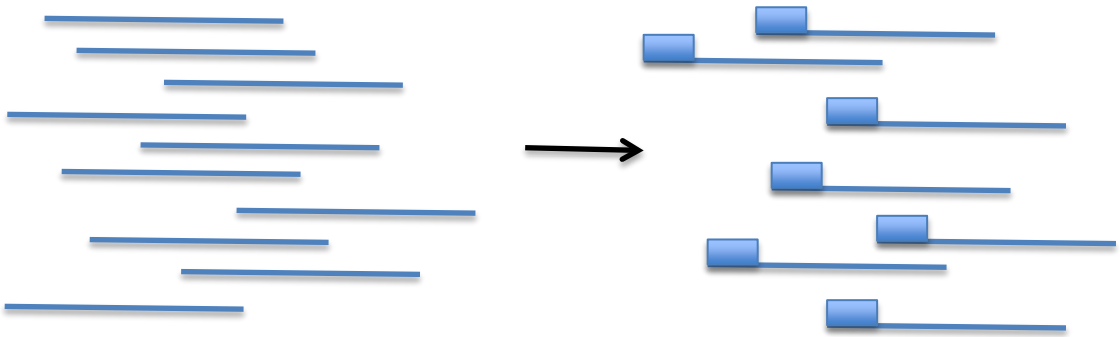
Making a NGS library



DNA QC – **paramount importance**



Sharing & size selection



Amplification

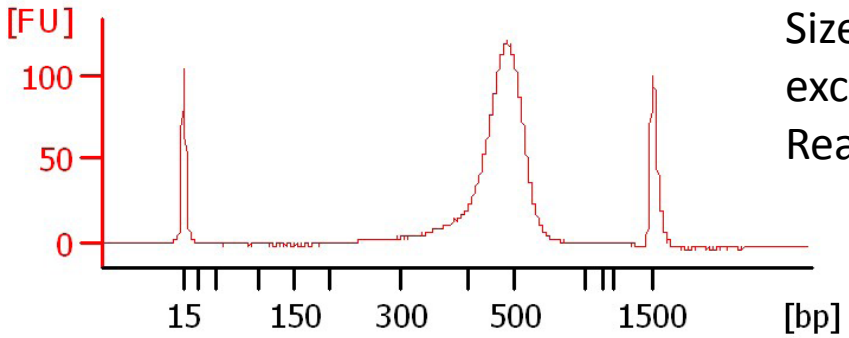
Ligation of sequencing adaptors, technology specific

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!)

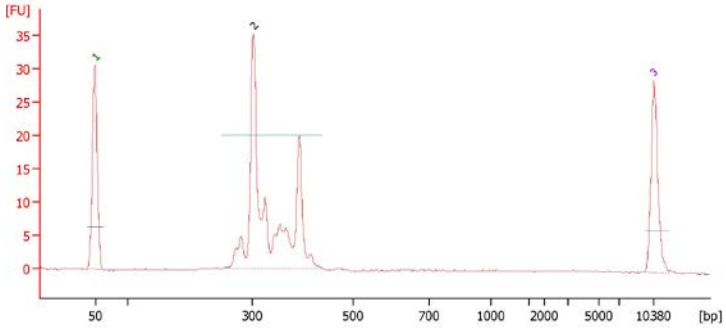
Bioanalyzer: amplicon size check

Example 1: OK size distribution

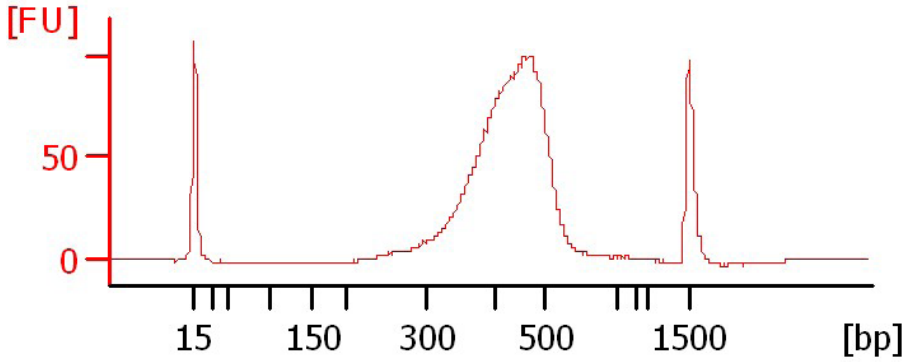


FOR ANY NGS TECHNOLOGY

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

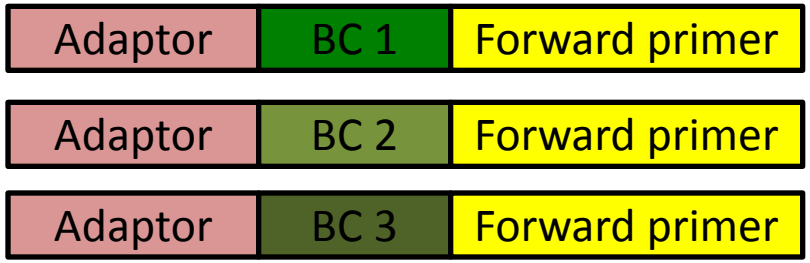
Two types of amplicon libraries

- Conventional library (900 – 2000 kr)

Used when you have a few samples (or internal, sample-specific barcodes)

Will be automatically de-multiplexed

- Fusion primer – based library



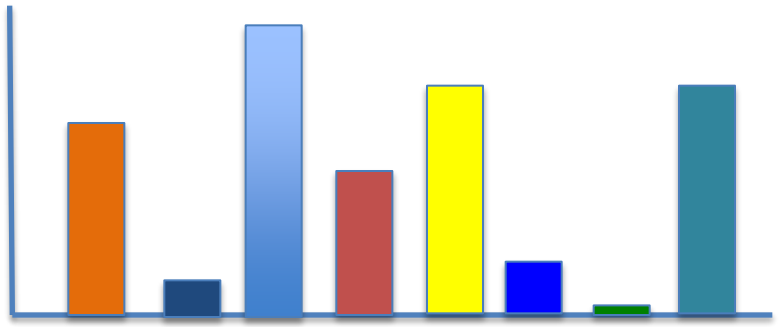
Used when you have many samples

Fusion primers must be of HPLC-quality

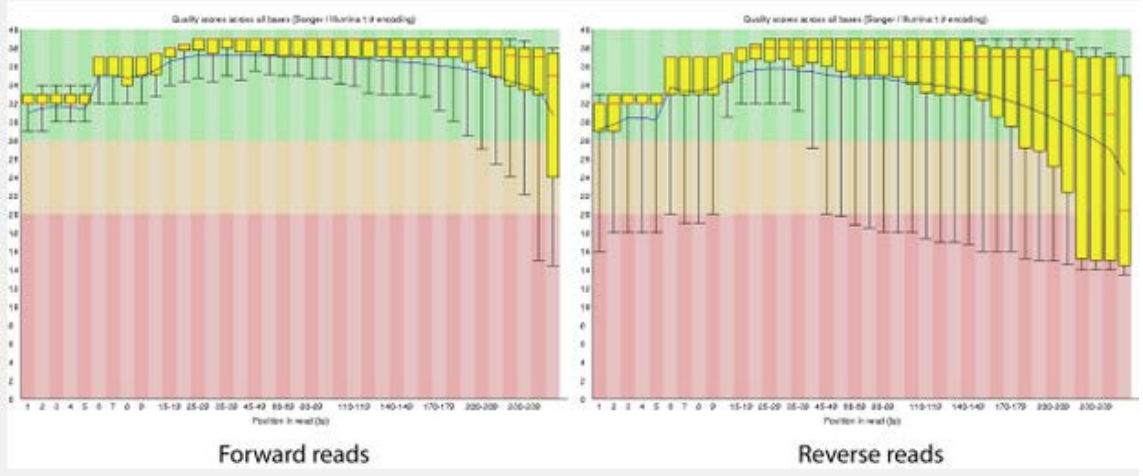
Users must de-multiplex their own barcodes

Barcoding

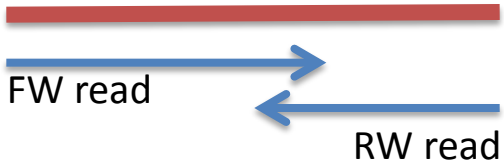
- It is possible to barcode up till 96-plex
- Libraries are pooled together equimolarily
- Concentrations determined by qPCR or Fragment Analyzer
- Multiplexing heavier than 32 external barcodes is not recommended – barcode imbalance

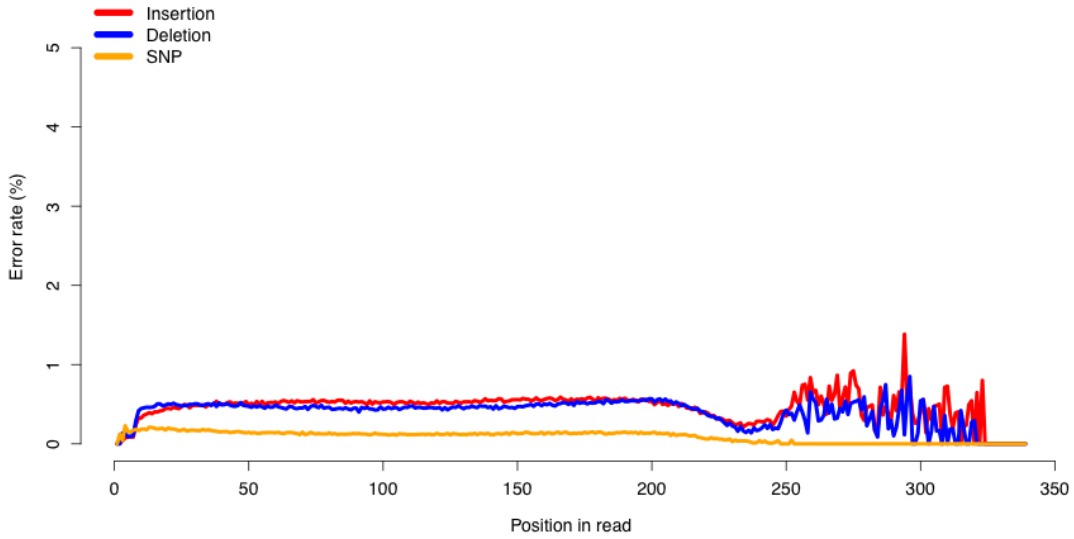


When you sequence an amplicon…



On MiSeq





On Ion

Bioinformatics support at SciLifeLab

NGI offers only a limited support

- NGI is a sequencing service only

- We provide best practice analysis for:
  - De-multiplexing of libraries constructed by NGI
  - UGC: quality and adaptor trimming of Ion reads
  - Some applications – mapping to reference, SNP calling, base-call statistics, etc.
  - De novo assembly of DNA sequences, if needed

SciLife Bioinformatics: short-term support - **BILS**

Bioinformatics Infrastructure for Life Sciences

- Originally VR infrastructure
- More than 40 employees around Sweden
- 80 hours support free of charge
- First come – first serve basis
- Additional support can be purchased for
  - 2000 kr per 8 hours

SciLife Bioinformatics: long-term support - **WABI**

Wallenberg Advanced Bioinformatics Infrastructure

- Originally KAW infrastructure
- 15 employees in Uppsala - Stockholm
- 6 months support free of charge
- 4 calls a year, reviewed by external expert panel
- There MUST be recipient in the group
- Work in collaboration with research groups

Questions

How much does approximately cost to run different types of anlysis, SNP's e-Dna, etc. *It all depends on a scientific question. Every new project must be discussed.*

Can SciLife lab provide Bioinformatic services? i.e help you analyse your sequences, and if so how much do they charge? *Yes: BILS and WABI.*

How do you apply for analyses by Scilife lab? *https://portal.scilifelab.se/genomics/*

Is it possible to buy even DNA extraction according to a certain protocol? And if so, how much is that? *No. We have no resources for that*.

Shotgun-sequencing. Is that more expensive/complicated as earlier techniques, and what are your experiences with this technique? *Hmm…* ☺

Do you have a guarantee, i.e. if the sequencing does not give sufficient bp, or if there is a contamination during the preparation, are you doing a second run for free? *If we fail – you get free re-run. If trouble is due to sample quality – sorry.*