Next-Generation Sequencing: an overview of technologies and applications

Matthew Tinning
Australian Genome Research Facility

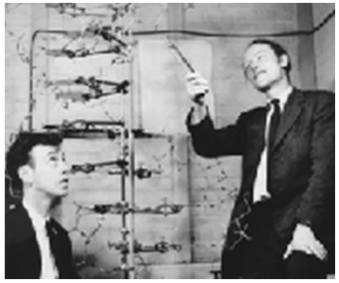
July 2013

www.agrf.org.au

A QUICK HISTORY OF SEQUENCING

A quick history of sequencing

1869 – Discovery of DNA



1909 – Chemical characterisation

1953 – Structure of DNA solved

1977 – Sanger sequencing invented

– First genome sequenced – ΦX174 (5 kb)

1986 – First automated sequencing machine

1990 – Human Genome Project started

1992 – First “sequencing factory” at TIGR

A quick history of sequencing

1995 – First bacterial genome – *H. influenzae* (1.8 Mb)

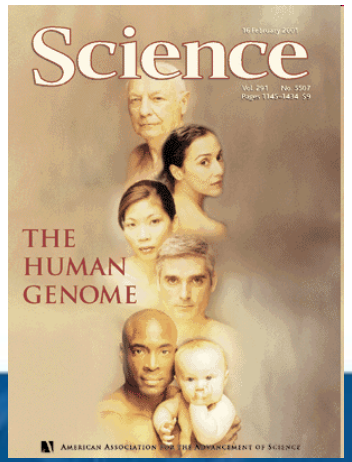
1998 – First animal genome – *C. elegans* (97 Mb)

2003 – Completion of Human Genome Project (3 Gb)

– 13 years, $2.7 bn

2005 – First “next-generation” sequencing instrument

2013– >10,000 genome sequences in NCBI database



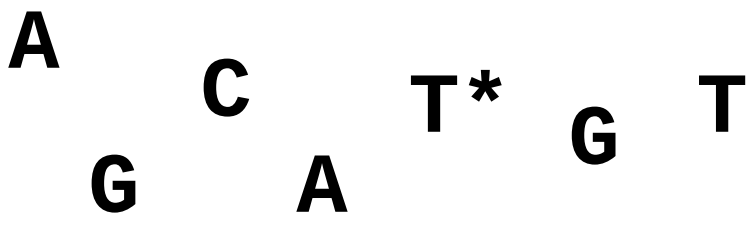
A quick history of sequencing

- 1977
  - First genome (ΦX174)
  - Sequencing by synthesis (Sanger)
  - Sequencing by degradation (Maxam-Gilbert)

Sanger sequencing: chain termination method

- Uses DNA polymerase
- All four nucleotides, plus one dideoxynucleotide (ddNTP)
- Random termination at specific bases
- Separate by gel electrophoresis

Sanger sequencing: chain termination method



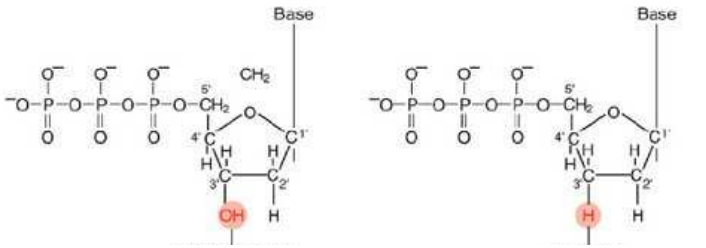
TCTGAT
AGACTACGTACTTGACGAGTAC.....

Incorporation of di-deoxynucleotides terminates DNA elongation

Individual reactions for each base

Sanger sequencing: chain termination method

TCTGATGCAT*
TCTGATGCATGAACT*
TCTGATGCATGAACTGCT*
TCTGATGCATGAACTGCTCAT*
AGACTACGTACTTGACGAGTAC.....

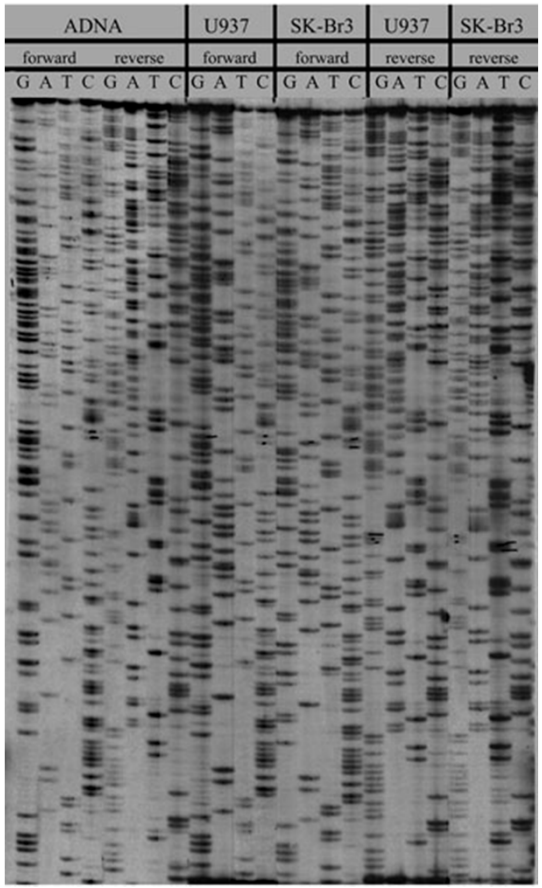


deoxynucleotide

dideoxynucleotide

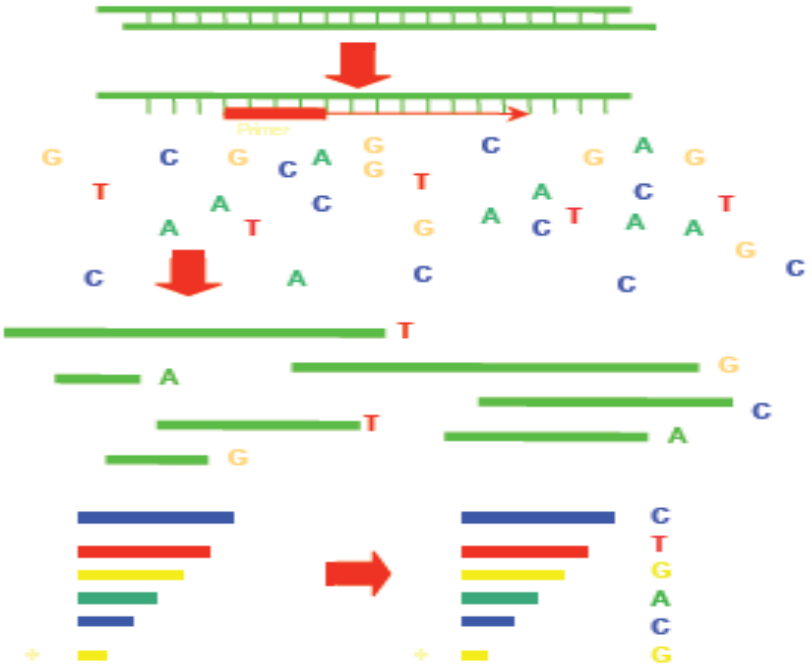
Sanger sequencing: chain termination method

Separation of fragments by gel electrophoresis

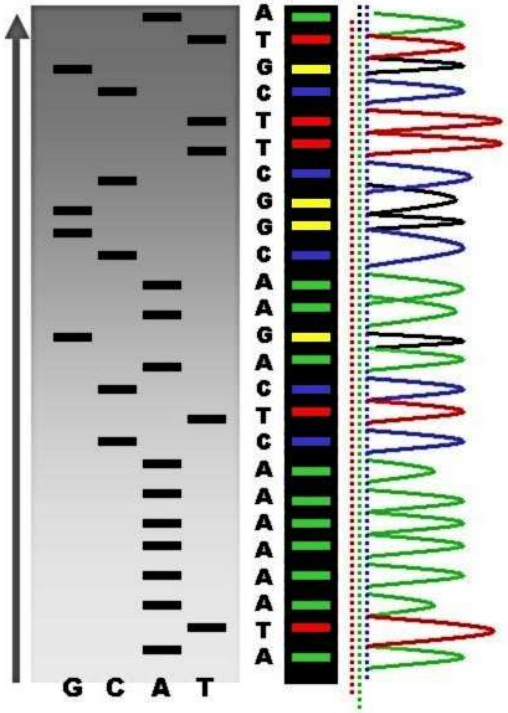


Sanger sequencing: dye-terminator sequencing

1986: 4 Reactions to 1 Lane
fluorescently labelled ddNTPs



Sequencing Reaction Products



Progression of Sequencing Reaction

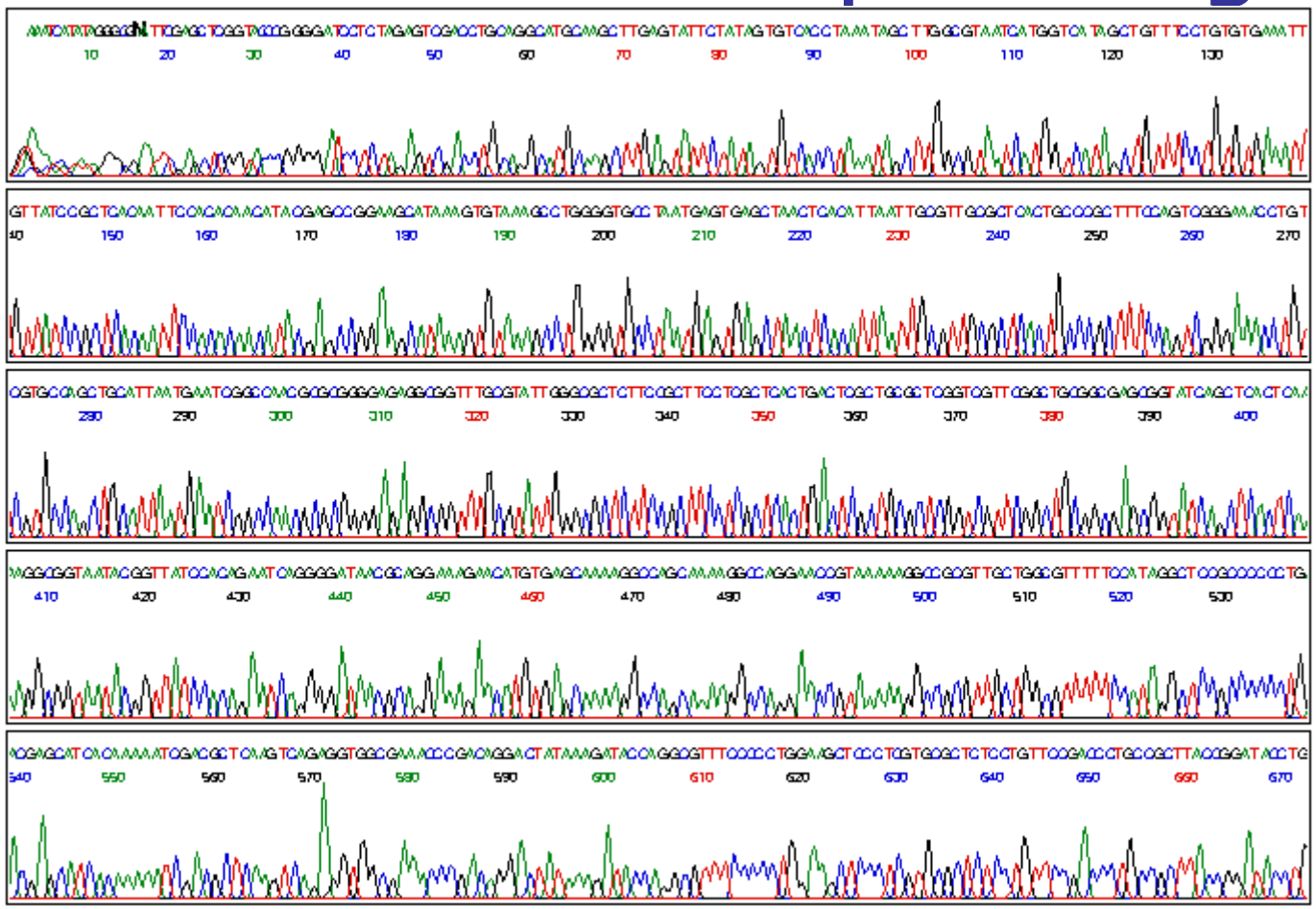
Sanger sequencing: dye-terminator sequencing

Automated DNA Sequencers

ABI 377 Plate Electrophoresis

ABI 3730 xl Capillary Electrophoresis

Sanger sequencing: dye-termination sequencing



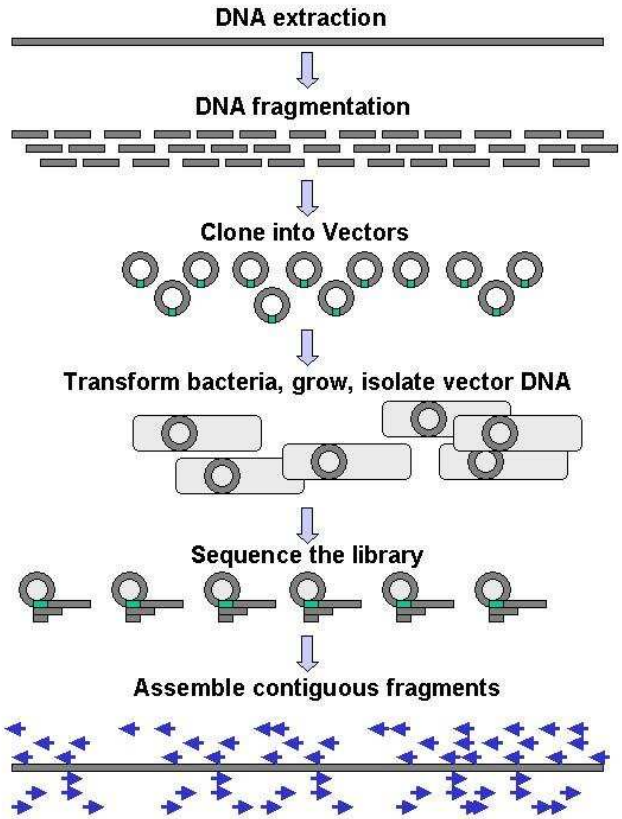
Sanger sequencing: dye-termination sequencing

Performance Specifications	Applied Biosystems 3730 and 3730*x*/ DNA Analyzers				3730	3730*x*/
Selected Applications	Capillary	Runs/Day	ABI Basecaller Phred Q_{20} Bases per Read	LOR**	KB Basecaller Q_{20} Bases per Day	
TargetSeq™ Resequencing	36 cm	72	400	400	1,440,000 bases	2,880,000 bases
Rapid Sequencing	36 cm	40	500	550	1,056,000 bases	2,112,000 bases
Standard Sequencing	36 cm	24	650	700	806,400 bases	1,612,800 bases
Extra Long Read Sequencing	50 cm	8	>800	900	345,600 bases	691,200 bases
DNA Sizing	36 cm	44	Single base resolution up to 500 bases with 0.15 SD		42,254*** genotypes	84,508*** genotypes

- Maximum read length ~900 base
- Maximum yield/day < 2.1 million bases (rapid mode, 500 bp reads)

< 0.1% of the human genome
> 1000 days of sequencing for a 1 fold coverage ...

Sanger sequencing: shotgun library preparation



Human Genome Project

- Launched in 1989 –expected to take 15 years
  - Competing Celera project launched in 1998
- Genome estimated to be 92% complete
  - 1st Draft released in 2000
  - "Complete" genome released in 2003
  - Sequence of last chromosome published in 2006
- Cost: ~$3 billion
  - Celera ~$300 million

Human Genome Project

NEXT GENERATION SEQUENCING

Next-gen sequencing technologies

- Four main technologies
- All massively parallel sequencing
  - Sequencing by synthesis
  - Sequencing by ligation
- Mostly produce short reads- from <400bp
- Read numbers vary from ~ 1 million to ~ 1 billion per run

Next-gen sequencing technologies

- With massively parallel sequencing new methods for sequencing template preparation is required
- Current NGS platforms utilize clonal amplification on solid supports via two main methods:
  - *emulsion PCR (emPCR)*
  - *bridge amplification (DNA cluster generation)*

Next-gen sequencing technologies

Next-gen sequencing technologies

Roche GS-FLX

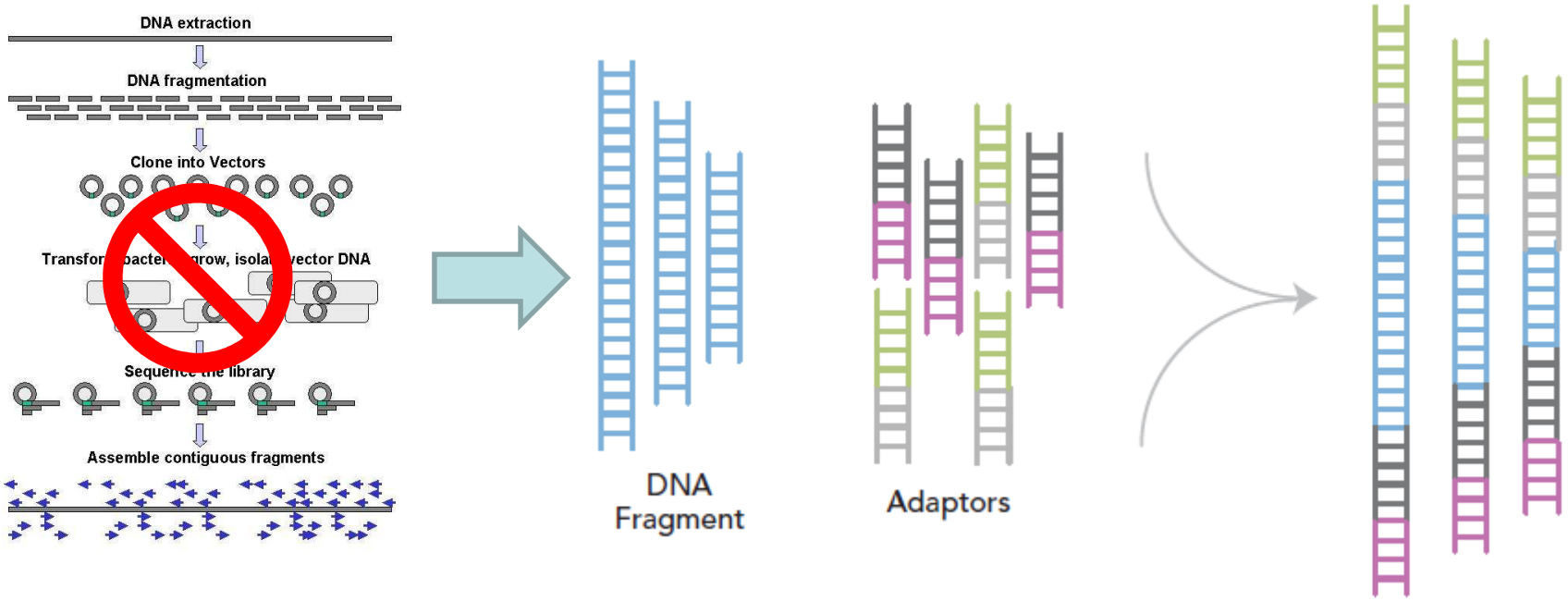
Life Technologies SOLiD

Illumina HiSeq

Life Technologies Ion Torrent/Proton

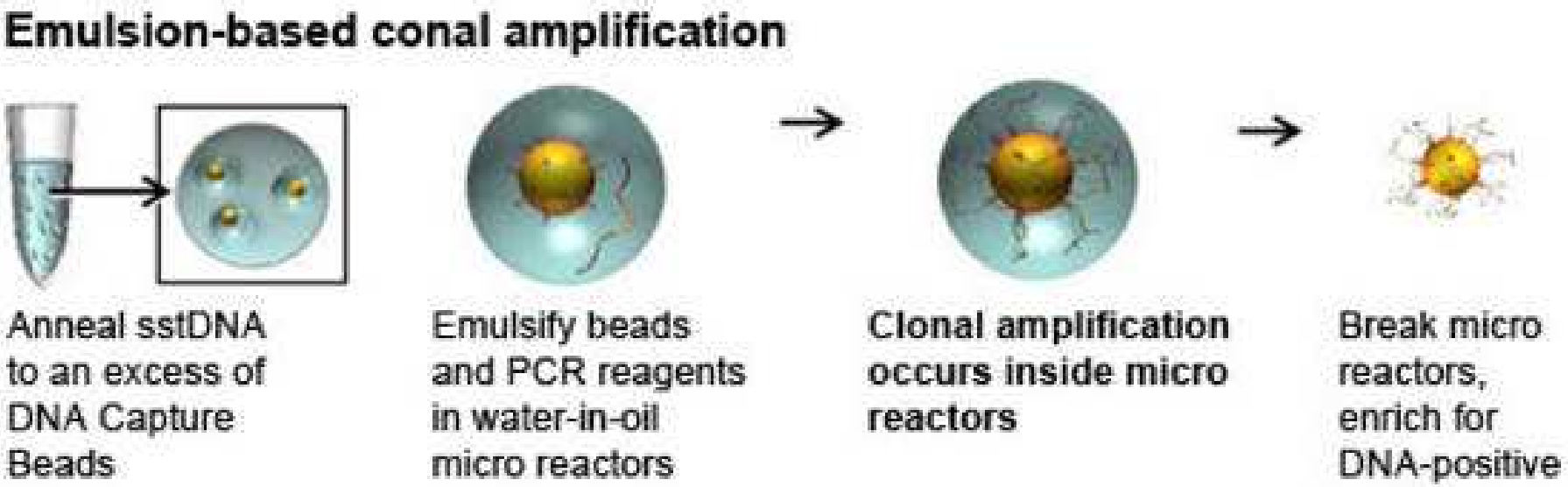
Roche GS-FLX

Next-gen sequencing: shotgun library preparation

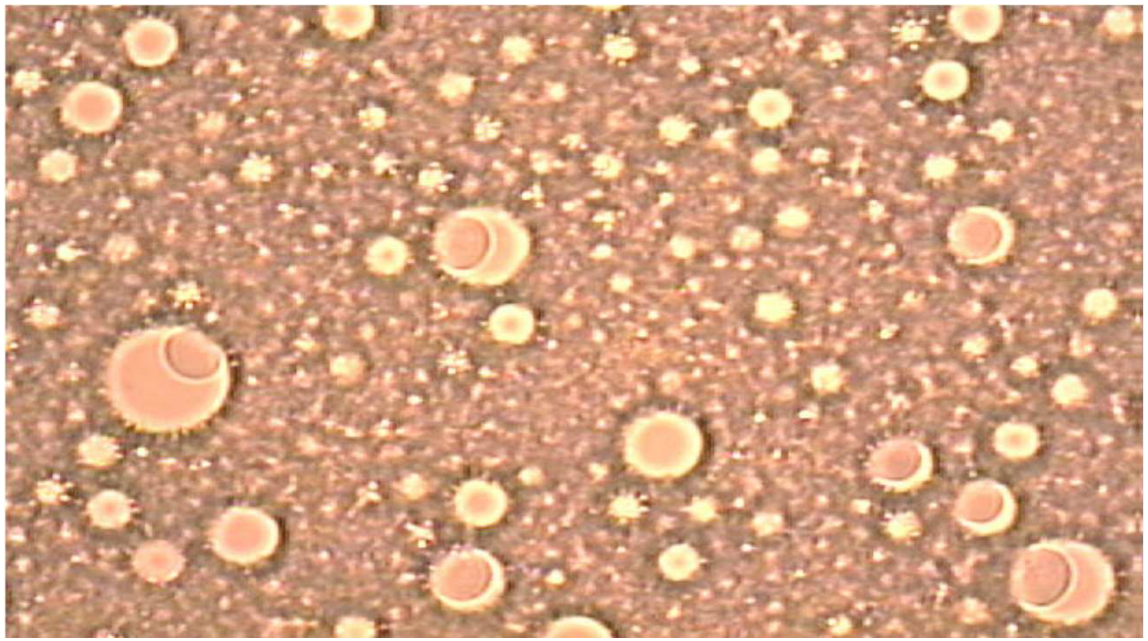


emPCR

Emulsion PCR is a method of clonal amplification which allows for millions of unique PCRs to be performed at once through the generation of micro-reactors.

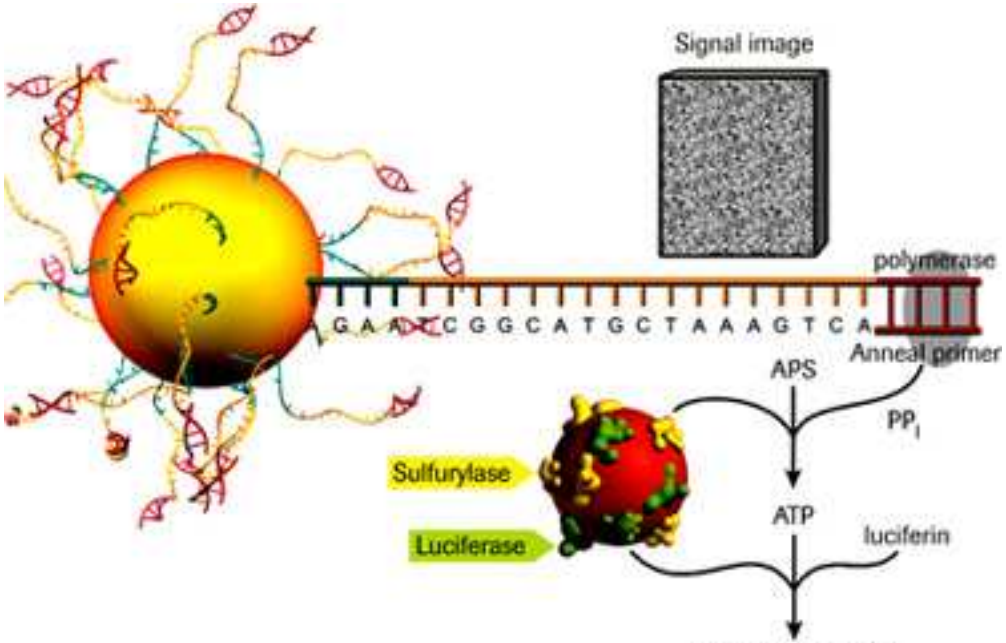


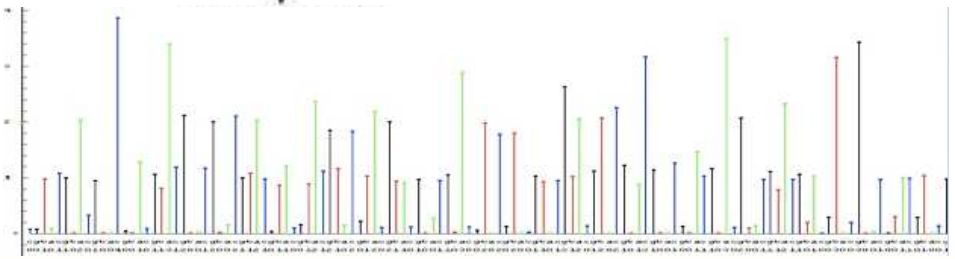
emPCR



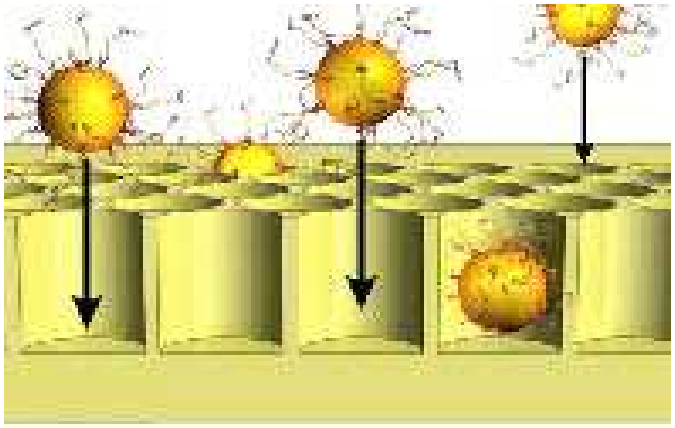
The Water-in-Oil-Emulsion

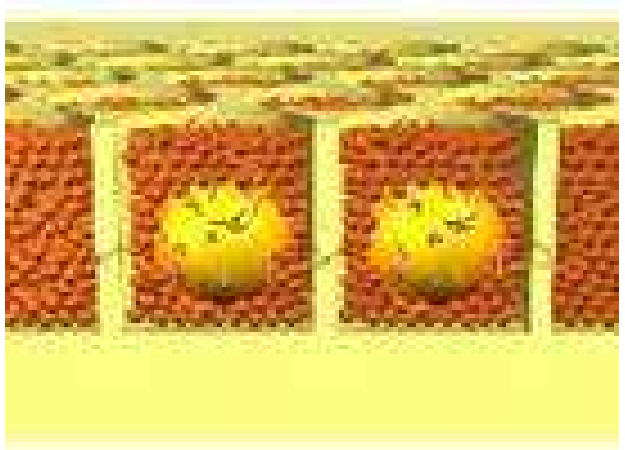
Pyrosequencing



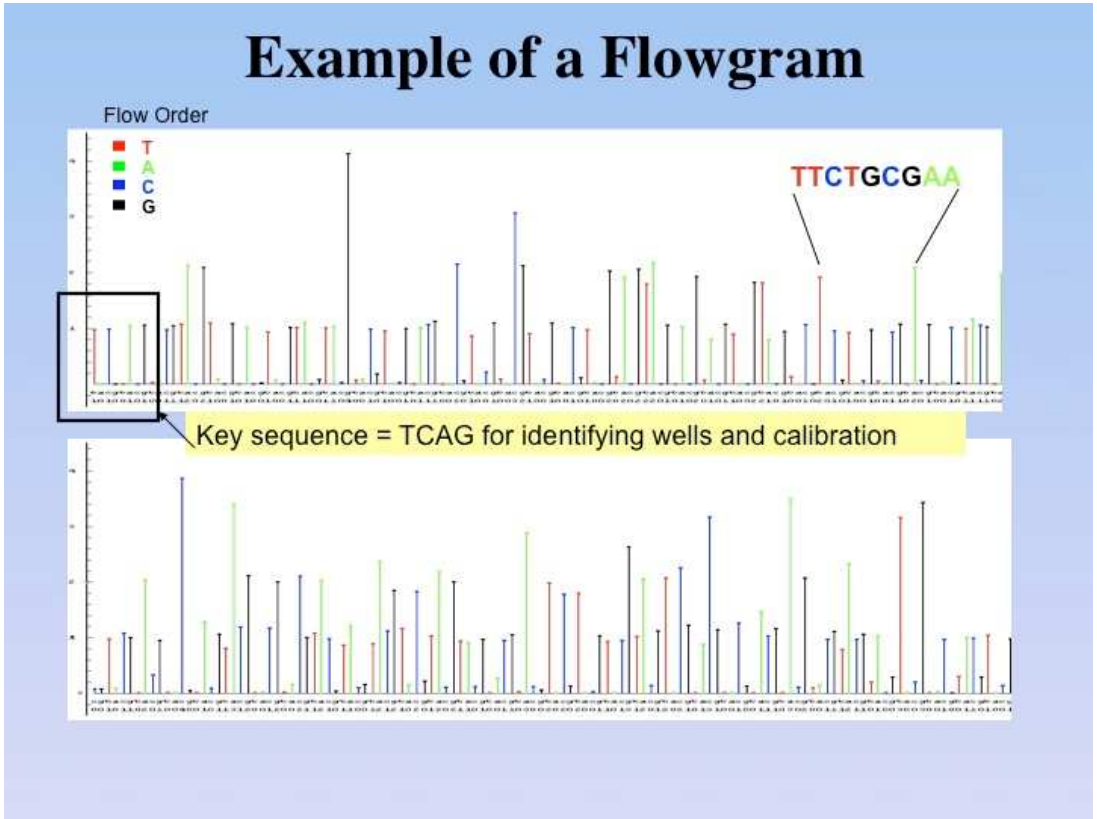


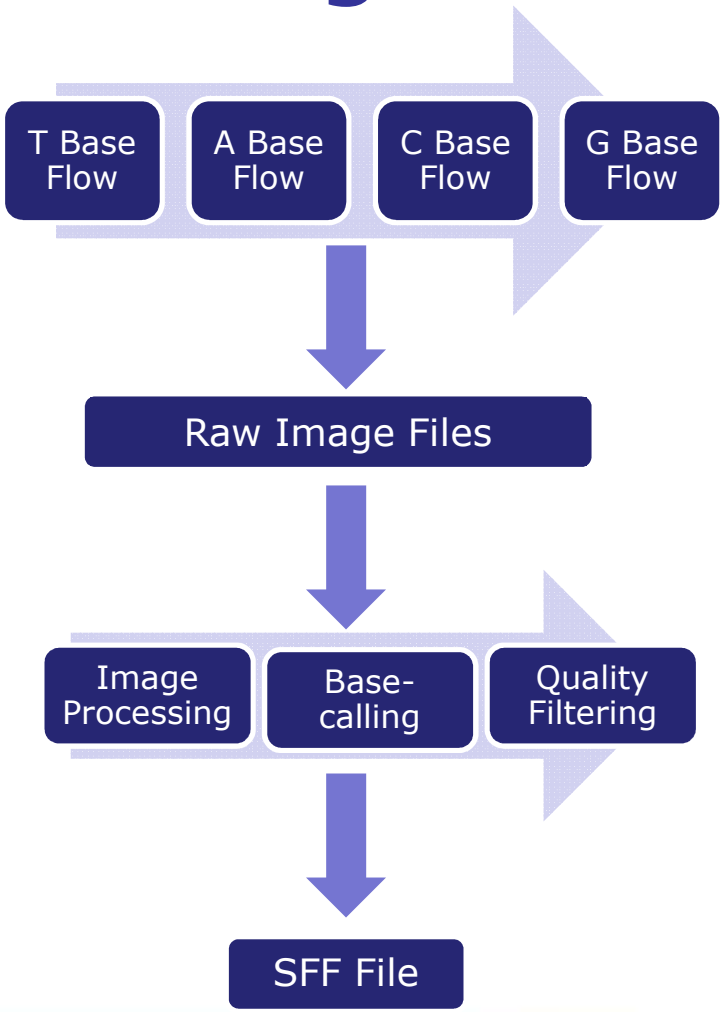
Massively Parallel Sequencing





454: Data Processing



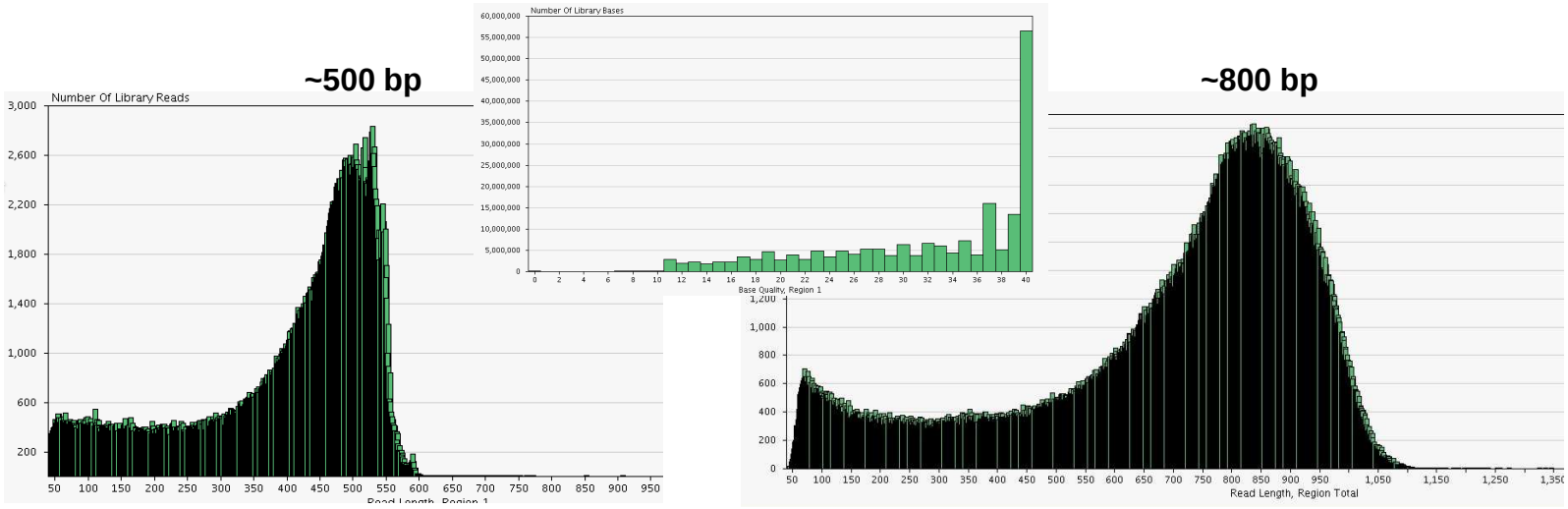


454 Platform Updates

GS20	• 100bp reads, ~20Mbp / run
GS-FLX	• 250bp reads ~100 Mbp / run (7.5 hrs)
GS-FLX Titanium	• 400bp reads ~400 Mbp / run (10 hrs)
GS-FLX Titanium Plus	• 700 bp reads ~700 Mbp/run (18 hrs)
GS Junior	• 400 bp reads ~ 35Mbp/run (10 hrs)

454 Sequencing Output

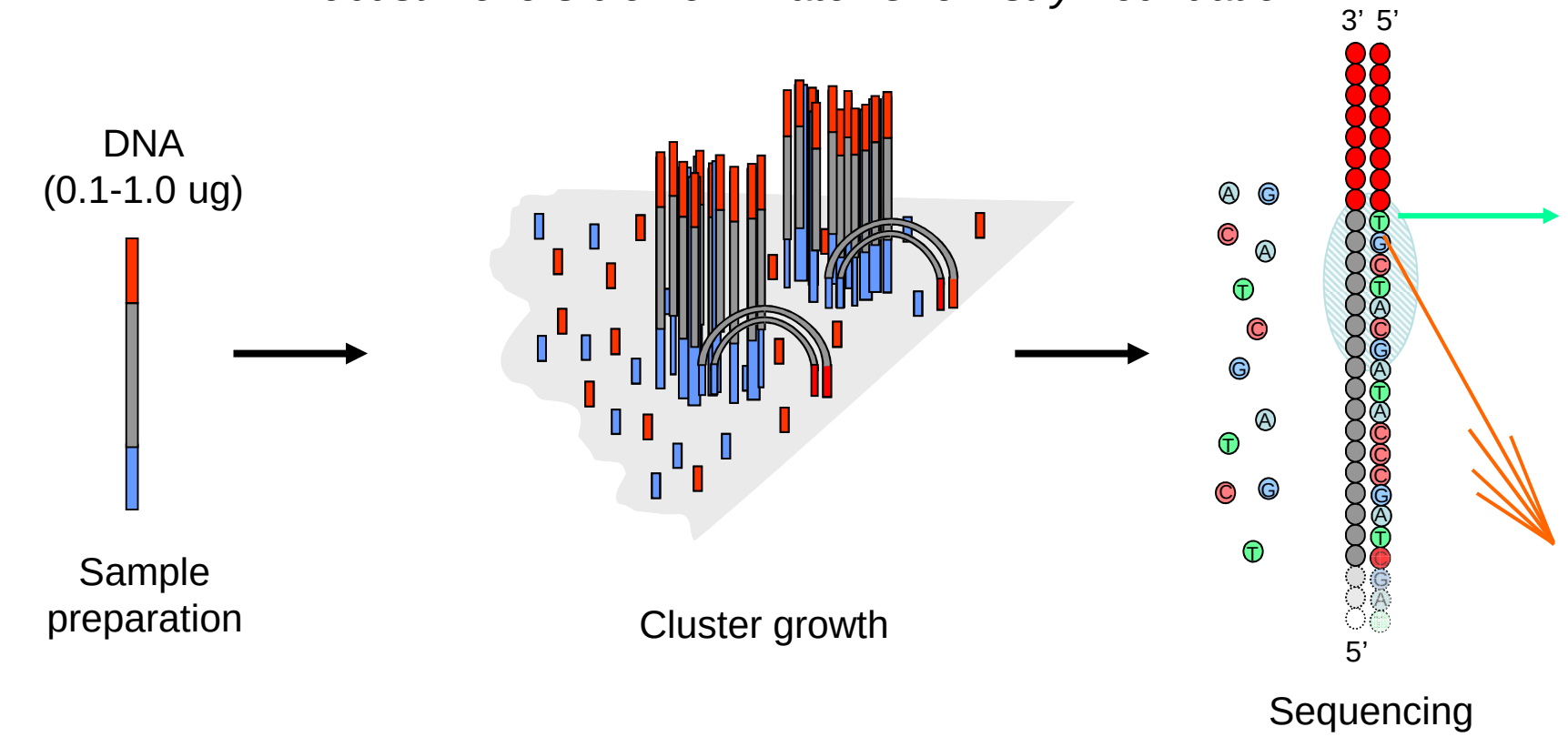
- *.sff *(standard flowgram format)*
- *.fna *(fasta)*
- *.qual *(Phred quality scores)*



Illumina HiSeq

Illumina Sequencing Technology

Robust Reversible Terminator Chemistry Foundation



Illumina: Data Processing

Nucleotide Flows

↓

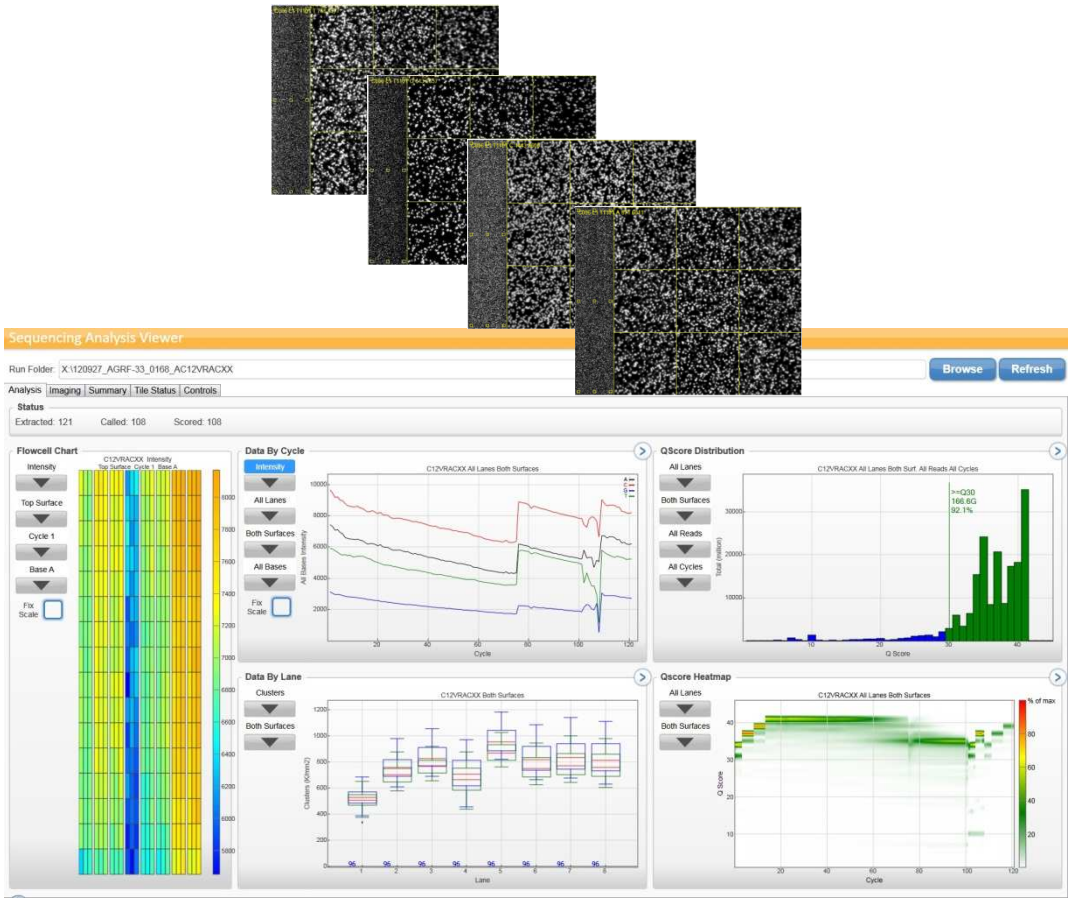
Raw Images

↓

Image Processing → Base-calling → Quality Filtering

↓

.bcl



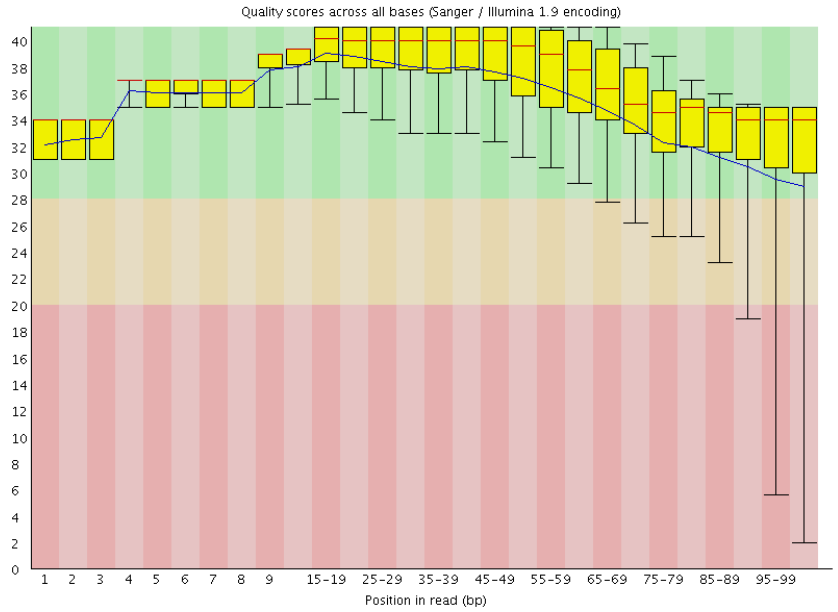
Platform Updates

Solexa 1G	• 18bp reads, ~1Gbp / run
Illumina GA	• 36bp reads ~3Gbp / run
Illumina GAII	• 75bp paired ends ~10Gbp / run (8 days)
Illumina GAIIx	• 75bp paired end reads ~40Gbp / run (8 days)
Illumina HiSeq 2000	• 100 bp paired end reads ~200 Gbp/ run (10 days)
Illumina HiSeq, v3 SBS	• 100bp paired end reads ~600Gbp / run (12 days)
Illumina HiSeq 2500 (Rapid)	• 150 bp paired end reads ~ 180 Gbp/ run (2 days)
MiSeq	• 250 bp paired end reads ~8 Gb/run (2 days)

Maximum yield / day 50,Gbp
~16x the human genome

Illumina Sequencing Output

- *.fastq *(sequence and corresponding quality score encoded with an ASCII character, phred-like quality score + 33)*



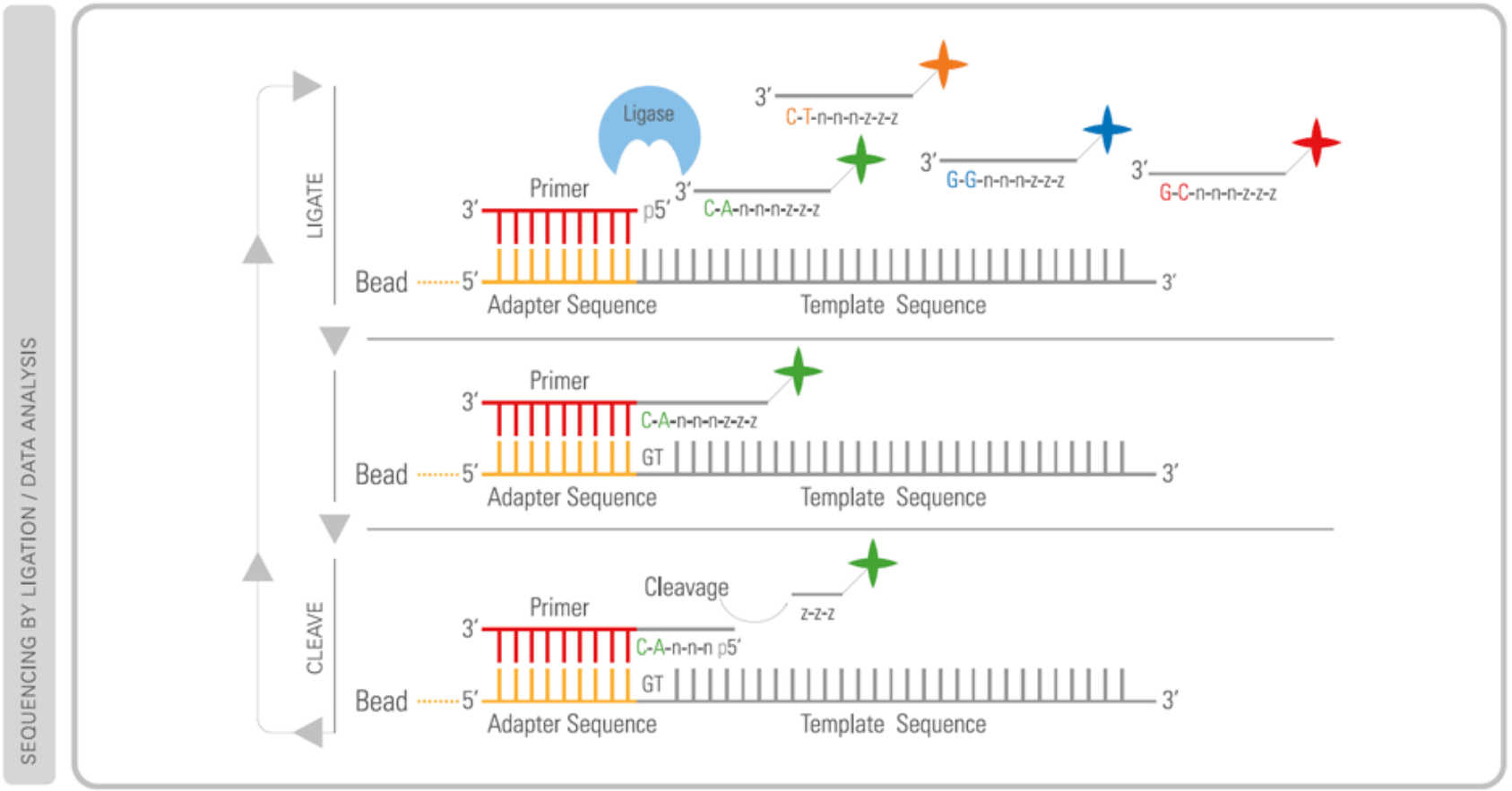
Illumina fastq

```
      1           2       3    4    5   6 7     8
@HWI-ST226:253:D14WFACXX:2:1101:2743:29814 1:N:0:ATCACG
TGCGGAAGGATCATTGTGGAATTCTCGGGTGCCAAGGAACTCCAGTCACATCACGATCTCGTATGCCGTCTTCTGCTT
GAAAAAAAAAAAAAAAAATTA
+
B@CFFFFFHHFFHJIIGHIHIJJIJIIJJGDCHIIIJJJJJJJGJGIHHEH@)=F@EIGHHEHFFFFDCBBD:@CC@C
:<CDDDD50559<B########
```

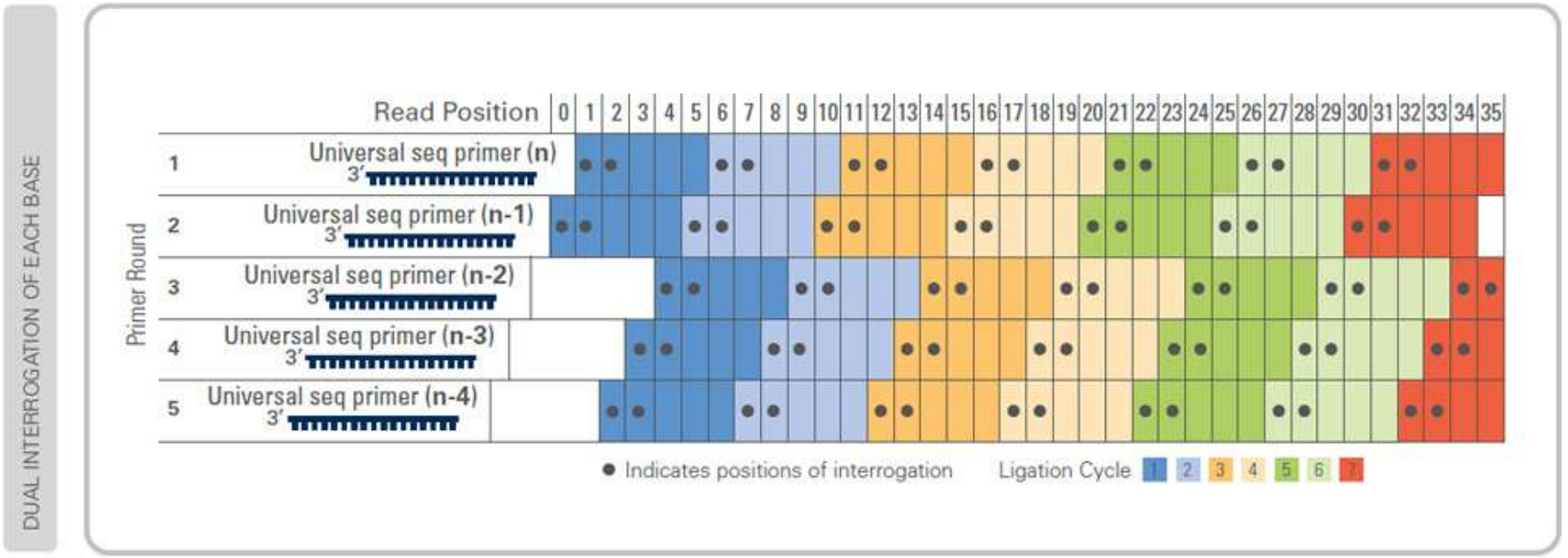
1. unique instrument ID and run ID
2. Flow cell ID and lane
3. tile number within the flow cell lane
4. 'x'-coordinate of the cluster within the tile
5. 'y'-coordinate of the cluster within the tile
6. the member of a pair, /1 or /2 *(paired-end or mate-pair reads only)*
7. N if the read passes filter, Y if read fails filter otherwise
8. Index sequence

Applied Biosystems SOLiD

Sequencing by Ligation

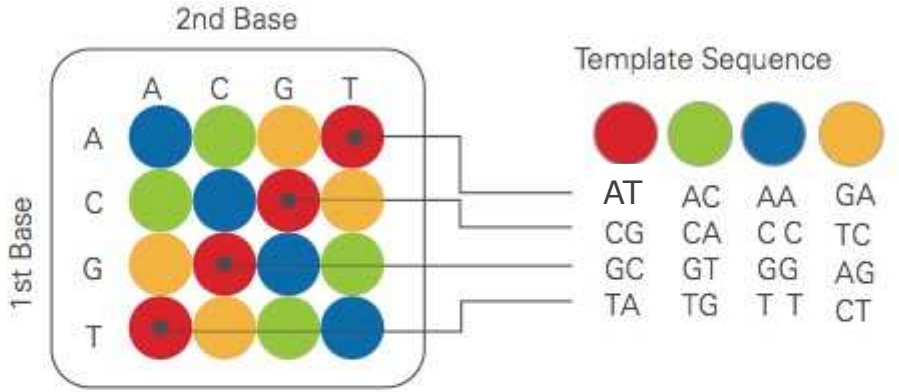


Base Interrogations



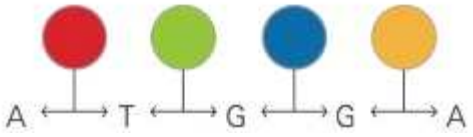
2 Base encoding

Possible Dinucleotides Encoded By Each Color

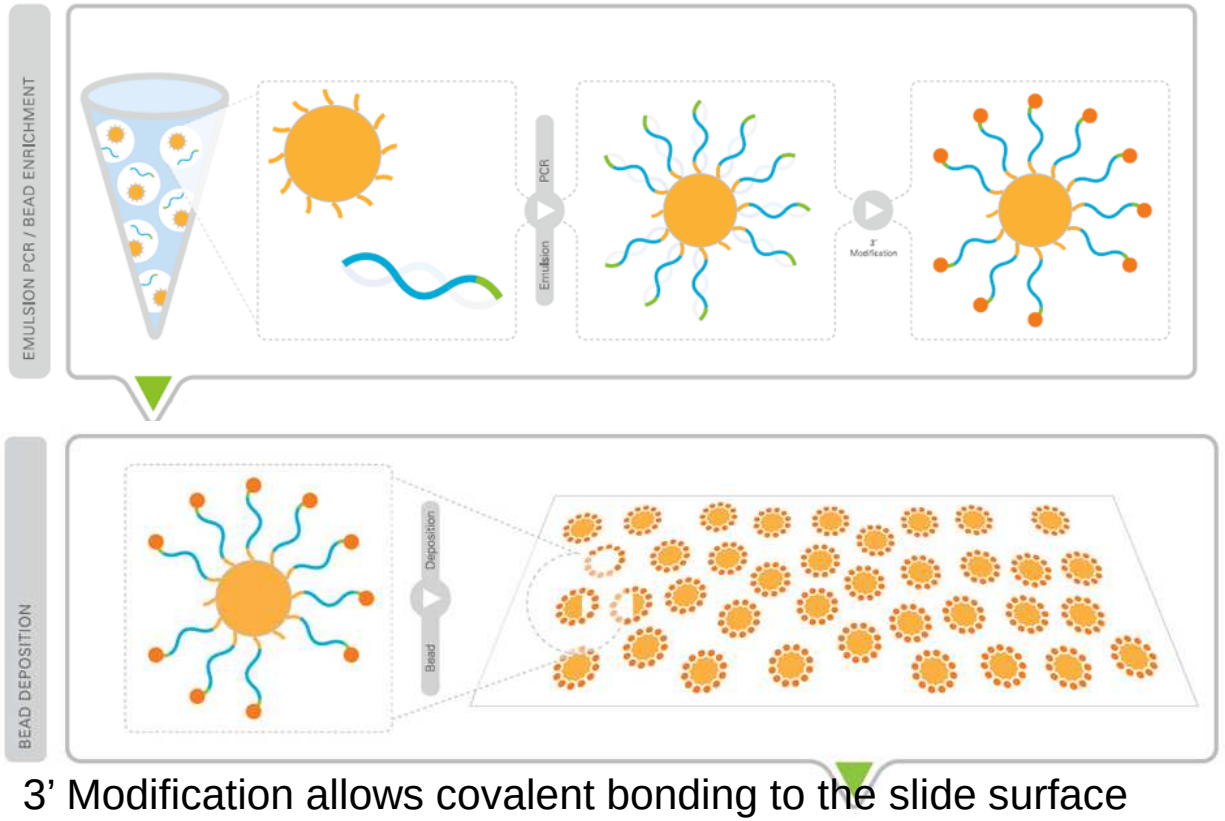


Double Interrogation

With 2 base encoding each base is defined twice



emPCR and Enrichment



3’ Modification allows covalent bonding to the slide surface

Platform Updates

SOLiD 3	• 50bp Paired reads ~50Gbp / run (12 days)
SOLiD 4	• 50bp Paired reads ~100Gbp / run (12 days)
5500xl	• 75bp Paired reads ~300Gbp / run (14 days)

Maximum yield / day 21,000,000,000bp
7x the human genome
3.5 hours of sequencing for a 1 fold coverage.....

SOLiD Colour Space Reads

- *.csfasta *(colour space fasta)*
- *.qual *(Phred quality scores)*

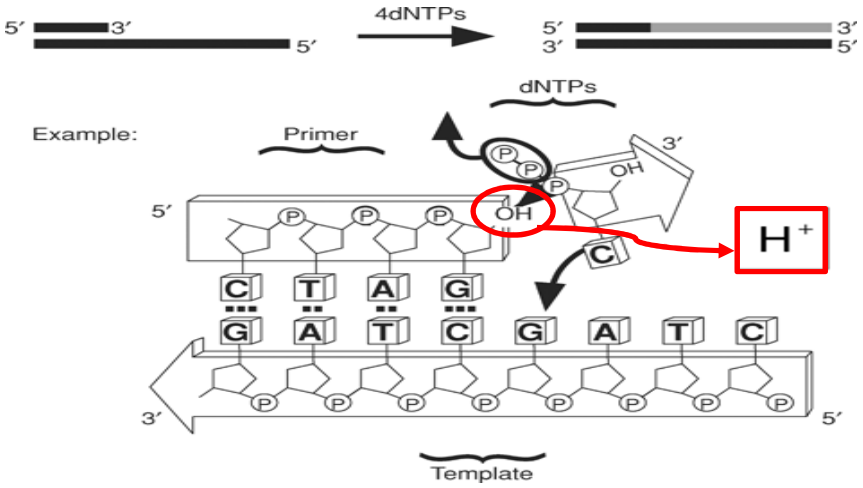
```
>853_17_1660_F3
T32111011201320102312......
```

AA	CC	GG	TT	0	Blue
AC	CA	GT	TG	1	Green
AG	CT	GA	TC	2	Yellow
AT	CG	GC	TA	3	Red

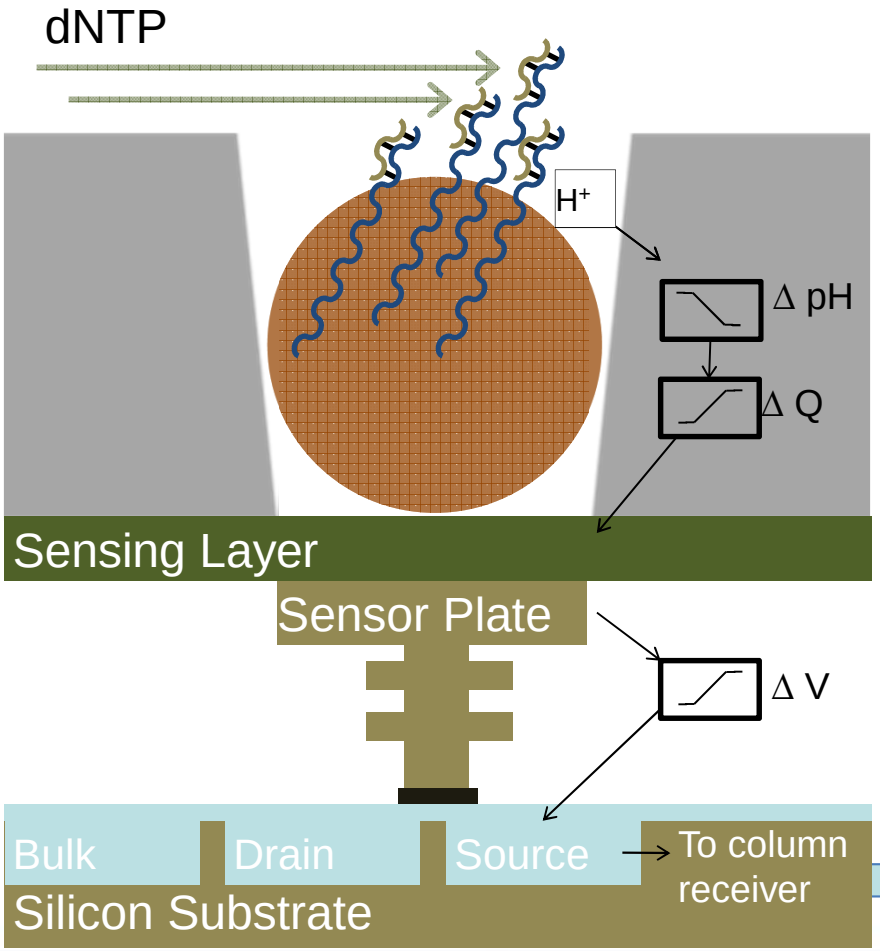
Applied Biosystems: Ion Torrent PGM

Ion Torrent

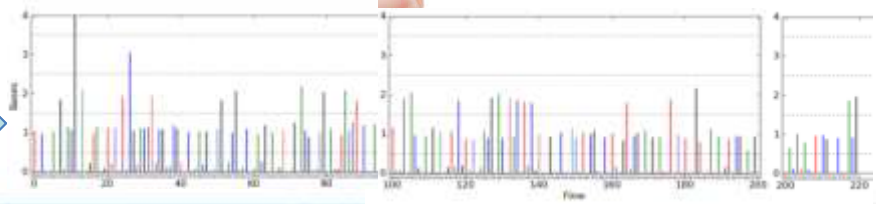
- Ion Semiconductor Sequencing
- Detection of hydrogen ions during the polymerization DNA
- Sequencing occurs in microwells with ion sensors
- No modified nucleotides
- No optics



Ion Torrent



- DNA → Ions → Sequence
  - Nucleotides flow sequentially over Ion semiconductor chip
  - One sensor per well per sequencing reaction
  - Direct detection of natural DNA extension
  - Millions of sequencing reactions per chip
  - Fast cycle time, real time detection

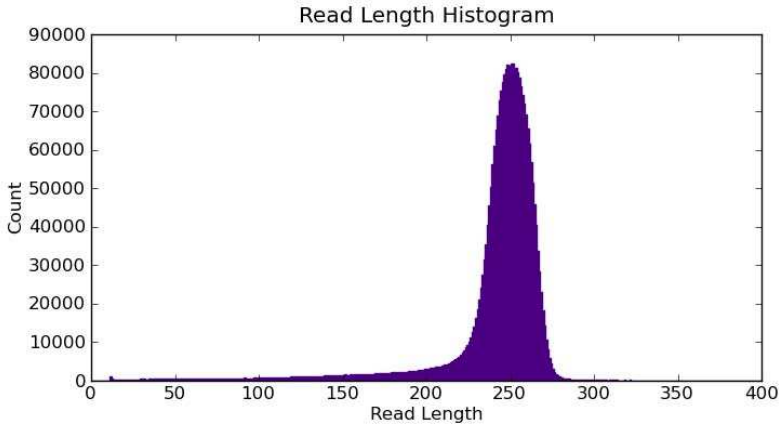


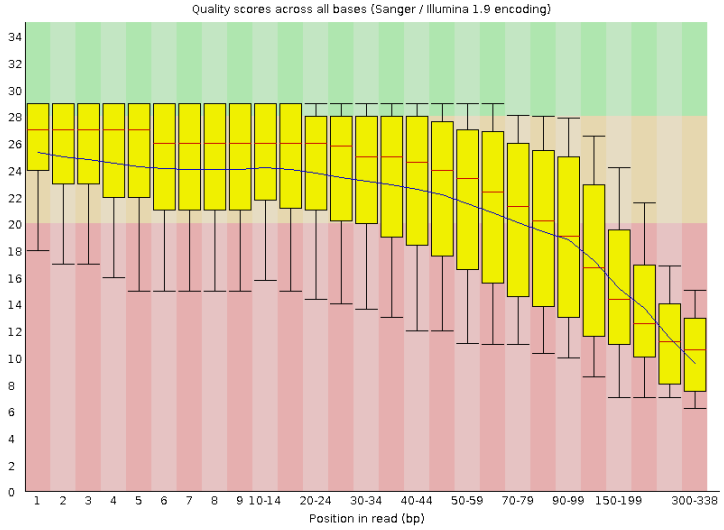
Ion Torrent: System Updates

314 Chip	•100bp reads ~10 Mb/run (1.5 hrs)
316 Chip	•100 bp reads ~100 Mbp / run (2 hrs) •200 bp reads ~200 Mbp/run (3 hrs)
318 Chip	•200 bp reads ~1 Gbp / run (4.5 hrs)
P1 Chip	•100 bp reads ~8 Gbp/run

Ion Torrent Reads

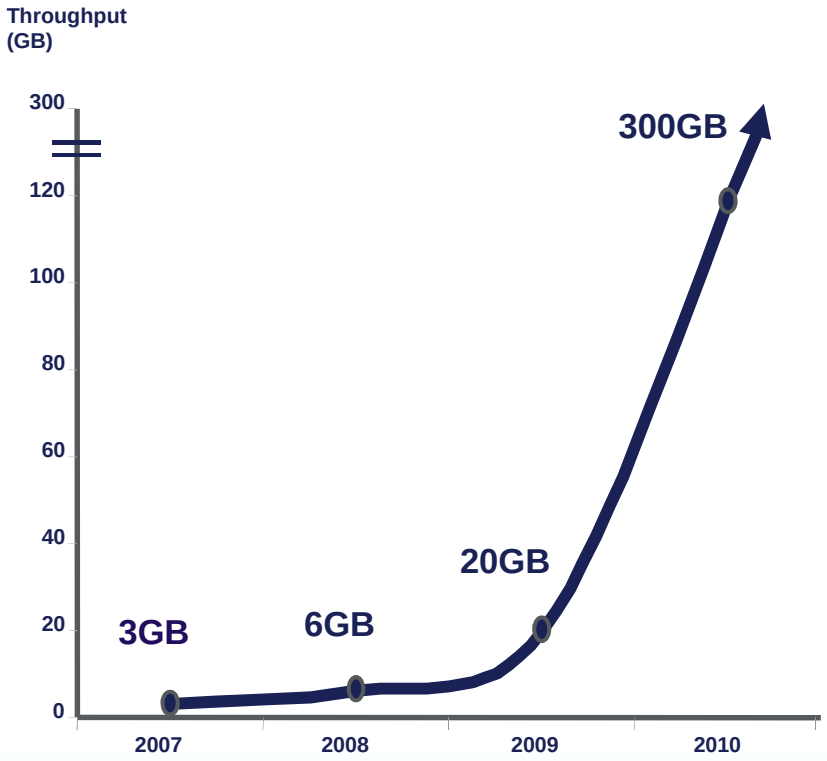
- *.sff *(standard flowgram format)*
- *.fastq (*sequence and corresponding quality score encoded with an ASCII character, phred-like quality score + 33)*



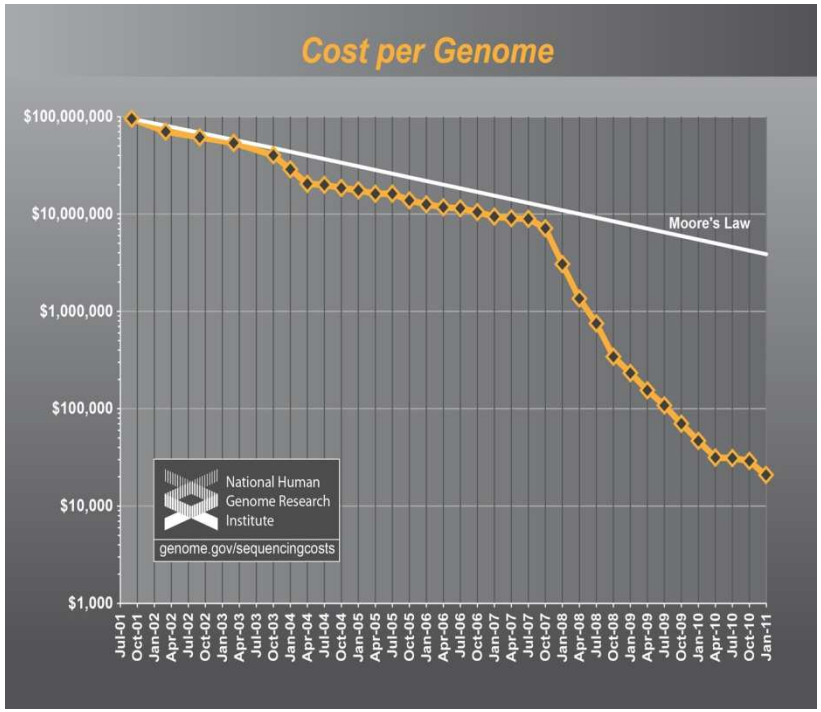


Rapid Innovation Driving Cost Down

Evolution of NGS system output



Cost per Human Genome



Summary of NGS Platforms

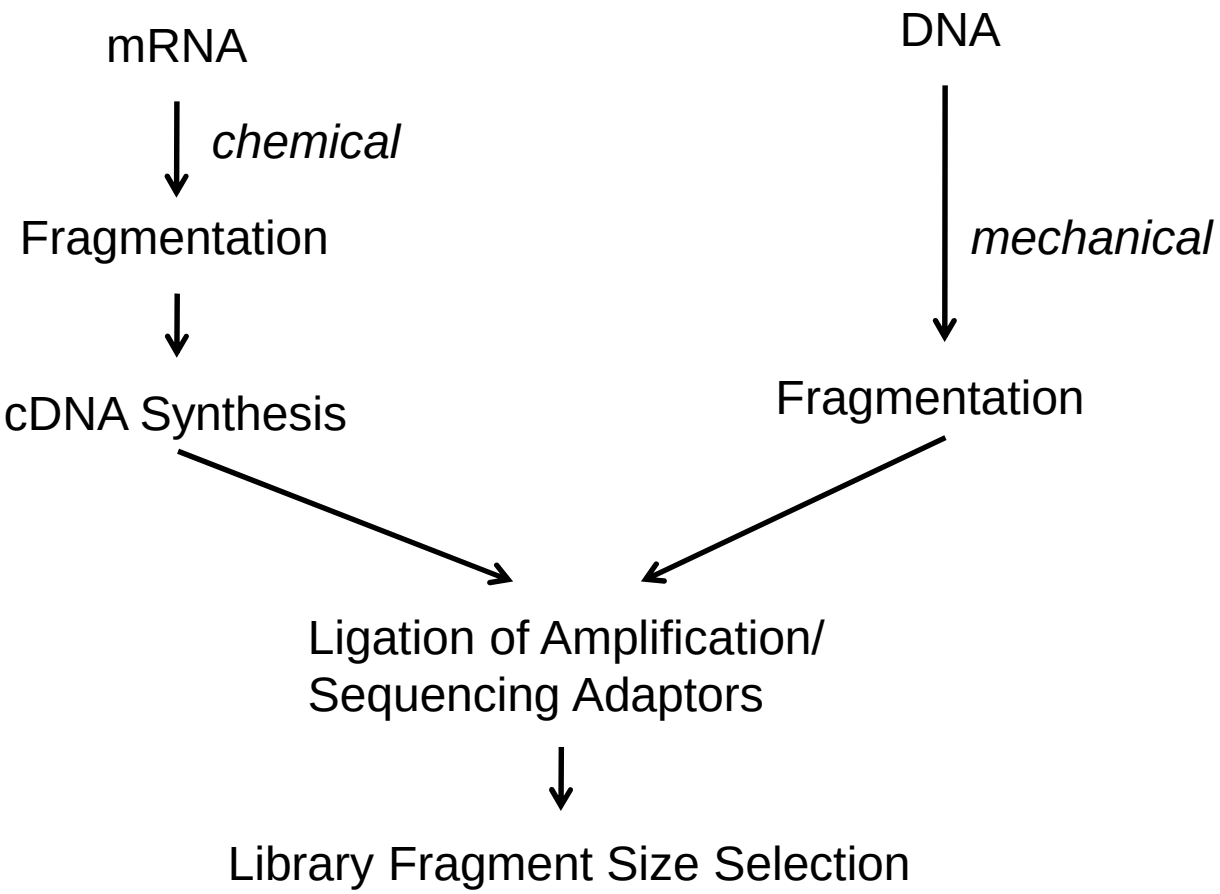
- Clonal amplification of sequencing template
  - emPCR (454, SOLiD and Ion Torrent)
  - Bridge amplification (Illumina)
- Sequencing by Synthesis
  - 454 *Pyrosequencing*
  - Illumina *Reversible Terminator Chemistry*
  - Ion Torrent *Ion Semiconductor Sequencing*
- Sequencing by ligation
  - SOLiD – 2 base encoding
- Dramatic reduction in cost of sequencing
  - GS-FLX provides > 100x decrease in costs compared to Sanger Sequencing
  - HiSeq and SOLiD > 100x decrease in costs over GS-FLX

NEXT GENERATION SEQUENCING APPLICATIONS

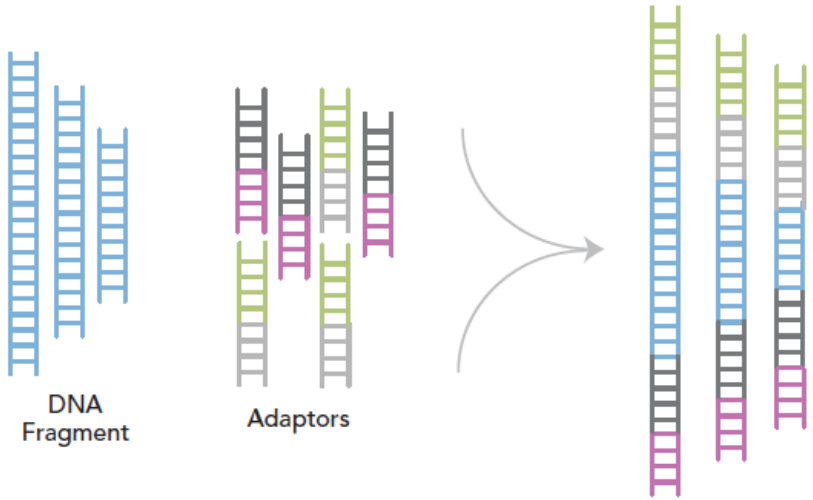
Applications

- DNA
  - whole genome
    - Shotgun & Mate Pair
  - targeted re-sequencing
    - hybrid capture
    - amplicon
  - ChIP-seq
- RNA
  - mRNA
  - whole transcriptome
  - small RNA

Sample preparation



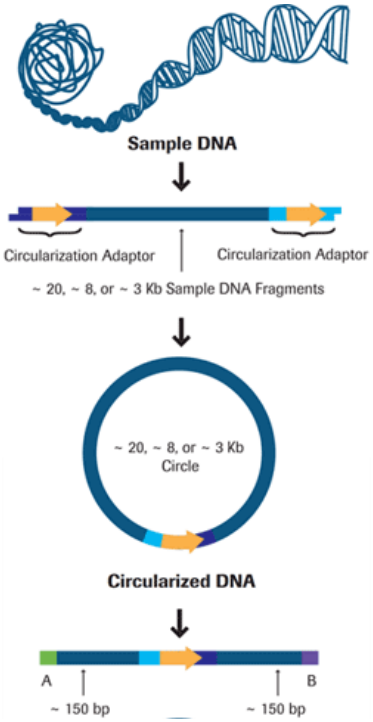
Next-gen sequencing: shotgun library preparation



Shotgun libraries

- Whole genome sequencing
  - Input: 100-1,000 ng of DNA
  - shear DNA (<1,000 bp)
  - *End repair*
  - *A-tailing*
  - *Ligation of sequencing adapters*

Next-gen sequencing: shotgun library preparation

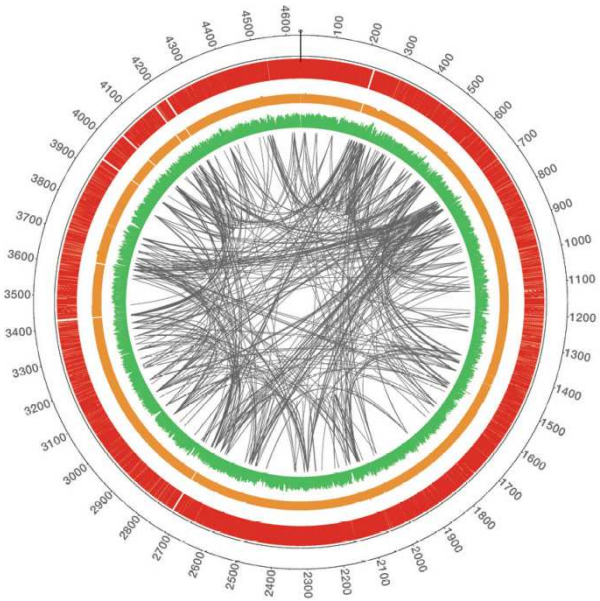


Mate pair libraries

- scafolding and structural variation
  - Input: 5-20 ug of DNA
  - Shear DNA to 3kb, 8kb and 20Kb fragments
  - Ligation of biotinylated circularization adapters
  - Shear circularized DNA
  - Isolate biotinylated mate pair junction
  - Ligate sequencing adapters

Whole Genome Sequencing

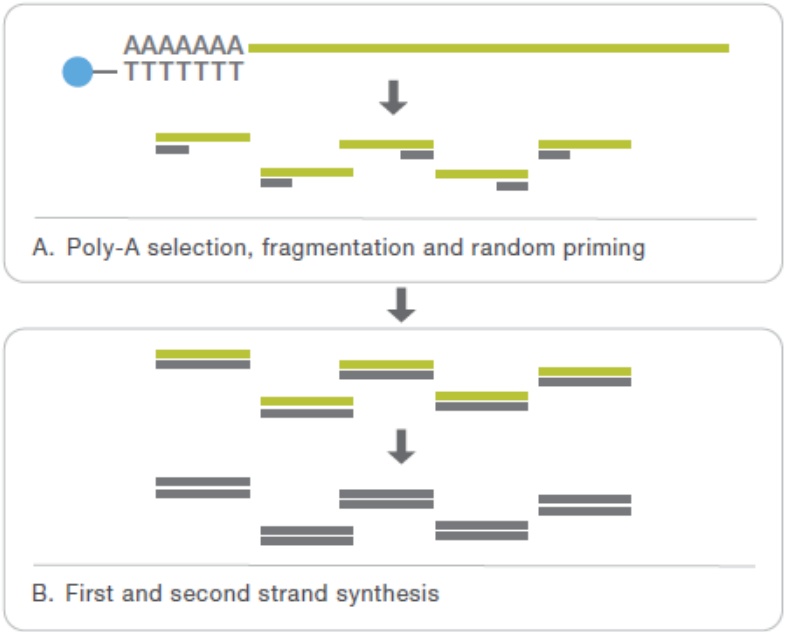
- *de novo* assembly
- Reference Mapping
  - SNVs, rearrangements
- Comparative genomics



***E. coli* assembly from MiSeq Data**
Illumina application notes

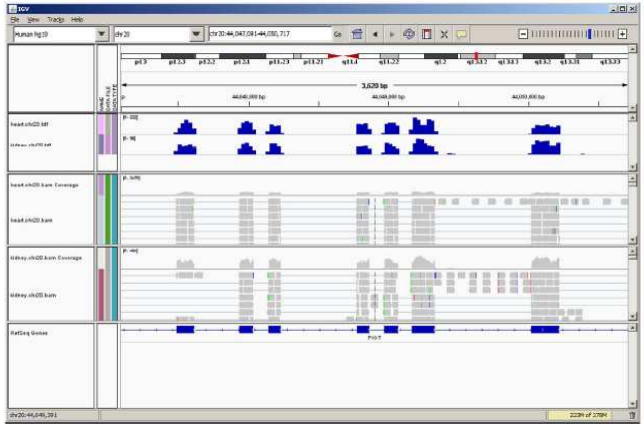
RNA-seq (cDNA libraries)

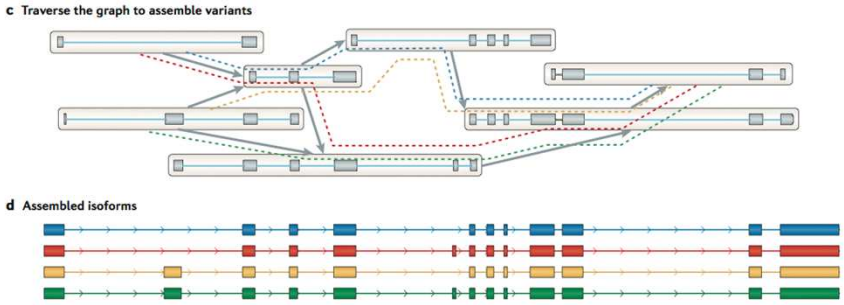
- Shotgun cDNA library of
  - Isolation of Poly(A) RNA or removal of rRNA
  - (100 ng – 4 ug of total RNA)
  - Chemical fragmentation of RNA
  - Random primed cDNA Synthesis & 2nd strand Synthesis
  - Follows standard "DNA" library protocol
- Stranded cDNA libraries
  - 2nd Strand "Marking" incorporation of dUTP in place of dTTP during second strand synthesis.
  - Selective enrichment for non-uracil containing 1st cDNA strand by
    - Use of a polymerase that cannot amplify uracil containing templates
- Small RNA Sample Preparation
  - RNA-adaptor ligation before cDNA synthesis
  - Small RNA size selection via PAGE
    - Library fragment ~145-160bp (insert 20-33 nucleotides)



RNA-seq applications

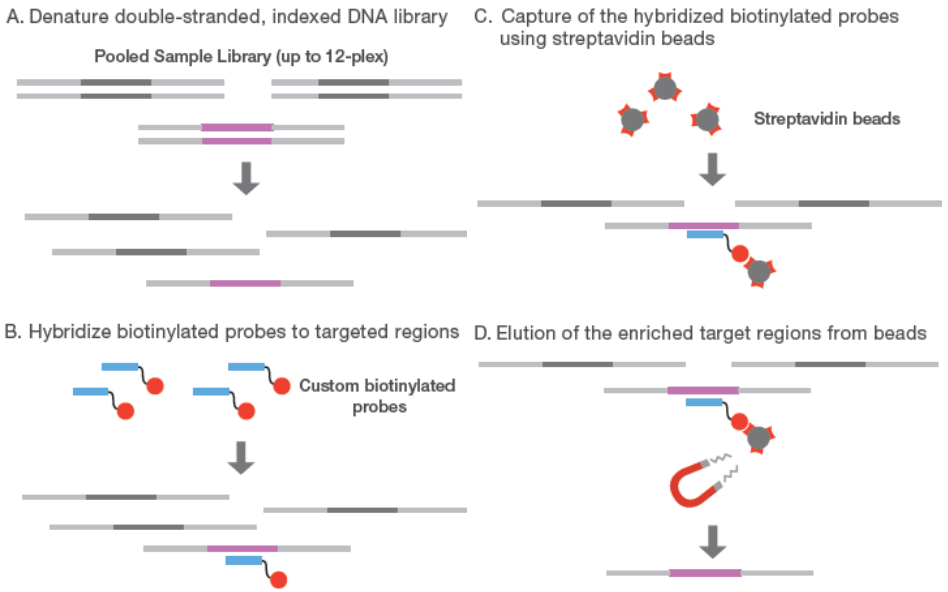
- Gene Expression
- Alternative Splicing & Allele Specific Expression
- Transcriptome Assembly





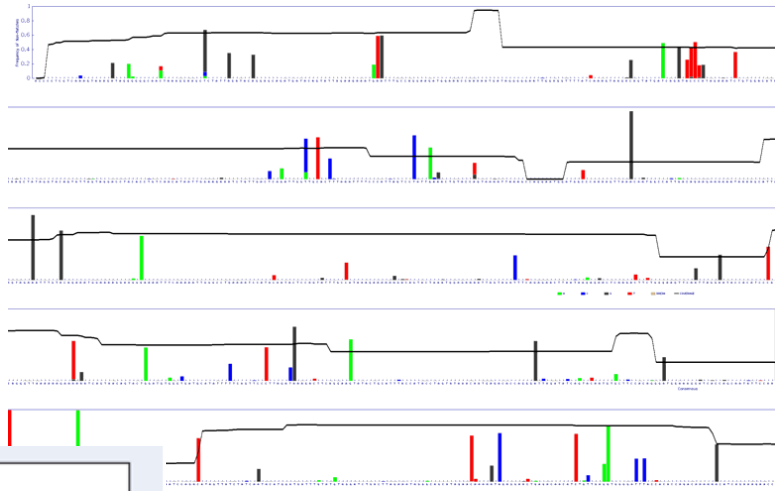
Targeted re-sequencing: hybrid capture

- Enrichment for specific targets via capture with oligonculeotide baits
  - Exome Capture
    - *Capture 40-70 Mb of annotated exons and UTRs*
  - Custom Capture
    - *up to 50 Mb of target sequences*



Targeted re-sequencing: amplicons

- Preparation of amplicons tagged with sequencing adapters
  - Well suited for 454 and bench top sequencers
  - Deep sequencing for detection of somatic mutations
  - 16S Sequencing for microbial diversity



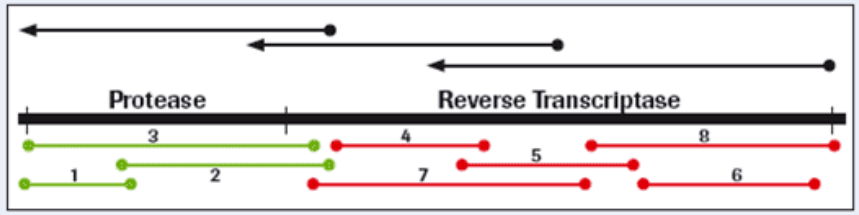


Figure 1: Schematic showing the 8 partially overlapping amplicons positioned across the HIV Protease (PR) and Reverse Transcriptase (RT) coding regions.

SUMMARY

Summary

- Next generation sequencing (NGS) is massively parallel sequencing of clonally amplified templates on a solid surface
- NGS platforms generate millions of reads and billions of base calls each run
- There are four main sequencing methods
  - Pyrosequencing (454)
  - Reversible terminator sequencing (Illumina)
  - Sequencing by ligation (SOLiD)
  - Semiconductor sequencing (Ion Torrent)
- NGS reads are typically short (<400 bp)
- Next generation sequencing is used for a range application including
  - sequencing whole genomes
  - sequencing specific genes or genomic reagions
  - gene expression analysis
  - study of epigenetics