Lars Paulin

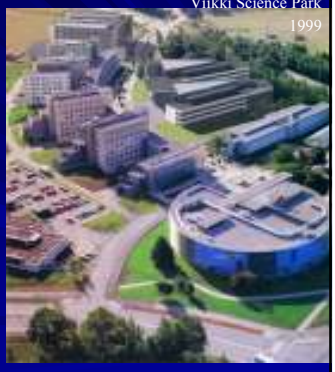
DNA Sequencing Technologies

DNA Sequencing and Genomics Laboratory

Institute of Biotechnology
University of Helsinki

http://www.biocenter.helsinki.fi/bi/dnagen/

Viikki Science Park
1999





## Institute of Biotechnology

- http://www.biocenter.helsinki.fi/bi/
- Independent Research Unit of the University of Helsinki
- About 300 people
- 30 Research groups

- Research Programs :
  - Developmental Biology
  - Cellular Biotechnology
  - Structural Biology and Biophysics
  - Genome Biology Research Program
- Director's Laboratory

- Core Facilities :
  - NMR Laboratory
  - Electron Microscopy
  - Protein Chemistry
  - DNA Sequencing and Genomics Laboratory
  - Transgenic unit
  - Light Microscopy unit

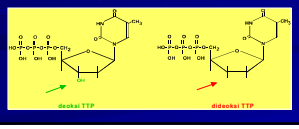


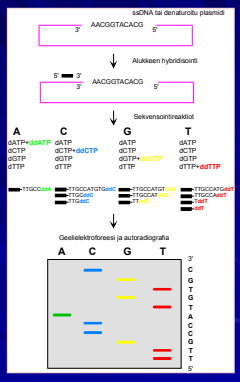
## Short History of DNA Sequencing

- 1977
  - Maxam-Gilbert
  - Sanger
- 1986
  - First Automated DNA Sequencer ABI 370 (373)
- 1988
  - Pharmacia ALF
- 1995
  - ABI 377
    - Up to 96 lanes
- 1996
  - First Capillary DNA Sequencer ABI 310
- 1998
  - First 96 Capillary instruments MegaBace, ABI 3700
- 2000
  - ABI 3100, 16 Capillary
- 2002
  - ABI 3730, 48 or 96 Capillary
- 2005
  - Genome Sequencer GS20, 454 Life Science, Roche
- 2006
  - Genome Analyzer, Solexa/ Illumina
- 2007
  - SOLiD, Applied Biosystems

## Sanger DNA Sequencing

1. Template
   - ssDNA or dsDNA
2. Primer annealing
   - Sequencing primer
3. Elongation
   - DNA polymerase
     - Steps 2 and 3 can be done repeatedly => cycle sequencing
4. Electrophoresis



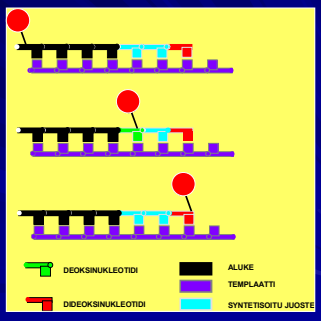

Lars Paulin Institute of Biotechnology University of Helsinki

## Incorporating Labels

Labelled primers
- 1 or 4 labels

Labelled deoxynucleotides
- 1 label

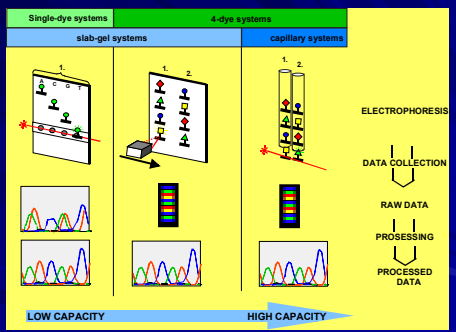
Labelled dideoxynucleotides
- 1 or 4 labels
- BigDye, ET terminators




Sarén, A-M *et.al.* Kemia-Kemi 1996, 23, 724-727

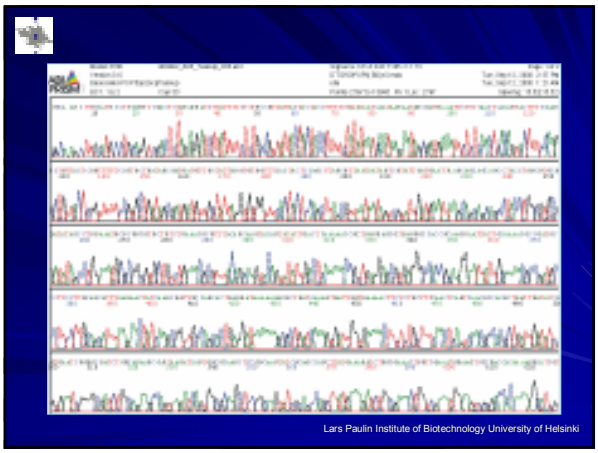
Lars Paulin Institute of Biotechnology University of Helsinki

## Automated DNA Sequencing




Sarén, A-M *et.al.* Kemia-Kemi 1996, 23, 724-727

Lars Paulin Institute of Biotechnology University of Helsinki



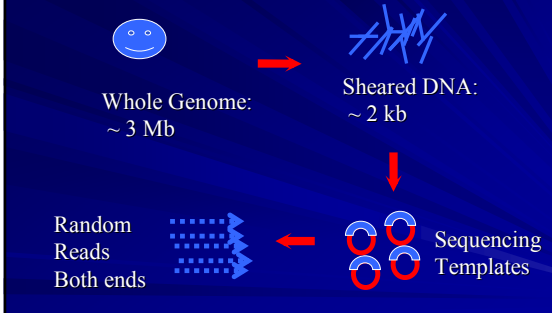
Lars Paulin Institute of Biotechnology University of Helsinki

## Strategies for Genome Sequencing

- Shotgun approach
  - random sequencing of different sized libraries
  - assembly using different software
  - closing of gaps using different methods
- Libraries
  - usually made by random shearing of genomic DNA
  - 2 kb, 4-6 kb, 10 kb plasmid libraries
  - fosmid or cosmid libraries with 30 - 50 kb inserts

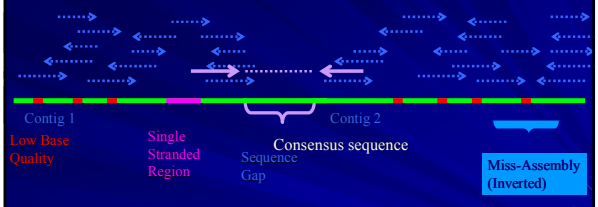
Lars Paulin Institute of Biotechnology University of Helsinki

## Whole Genome Shotgun Sequencing




Lars Paulin Institute of Biotechnology University of Helsinki

## Shotgun Sequencing :ASSEMBLY




- 0.5 -1.0 X (2 reads/kb) - 'Skimming'
- 3.5 - 4.0 X (~9 reads/kb) -'half-shotgun'
- 6.5 - 8.0 X (~18 reads/kb) - 'pre-finished'
- 10 X (22-24 reads/kb) - 'deep shotgun'

Lars Paulin Institute of Biotechnology University of Helsinki

## Phred, Phrap and Staden Package Program

### Phred and Phrap

- University of Washington
- Phil Green, http://www.phrap.org/

Phred quality score:

$QV = - 10 * \log_{10}( P_e )$

where $P_e$ is the probability that the base call is an error.

| Phred score | $P_e$ | Accuracy of the base call |
|---|---|---|
| 10 | 1 in 10 | 90% |
| 20 | 1 in 100 | 99% |
| 30 | 1 in 1,000 | 99.9% |
| 40 | 1 in 10,000 | 99.99% |
| 50 | 1 in 100,000 | 99.999% |

### Staden Program

- Cambridge, Sanger Center
- Roger Staden, http://staden.sourceforge.net/
- Trace editing
- Phrap assembly and Gap4 editing
  - display of traces from sequencers
  - translations, orfs, RE etc.
  - good capacity

Lars Paulin Institute of Biotechnology University of Helsinki

## New DNA Sequencing Technology

(Reviews: Schendure, J & Ji, H Nat Biotechnol 2008,26,1135-1145; Pettersson, E *et.al.* Genomics 2009, 93,105-111; Mardis, E Annu Rev Genomics Hum Genet 2008, 9,387-402; Metzker ML. Nat Rev Genet. 2010, 11,31-46)

Parallel Sequencing Technology

- Massive throughput
- Fast sequencing
- No cloning step
- PCR
- Currently four systems ready
  - Genome Sequencer (www.454.com,www.roche.com)
    - 454 Life Sciences, Roche
    - Launched in October 2005
  - Solexa, Genome Analyzer (www.illumina.com)
    - Illumina
    - Launched 2006
  - SOLiD (www.appliedbiosystems.com)
    - Applied Biosystems
    - Launched in October 2007
  - Helicos (www.helicosbio.com)
    - Helicos
    - Launched in 2009

Lars Paulin Institute of Biotechnology University of Helsinki

# Genome Sequencer

(www.454.com, www.roche.com)
(Margulies, *etal*. Nature 2005,437, 376)

- Genome Sequencer GS20;FLX; Titanium
  - 454 Life Science, Roche
- Parallel Sequencing
  - Shotgun sequencing
    - No plasmid libraries
    - Linkers ligated to fragments
    - Emulsion PCR
    - Picotiter plate, 1 600 000 wells
    - Titanium 3 400 000 wells
  - Pyrosequencing
    (Nyren, P. et al Anal Biochem. 1993, 208,171-5)
    - Detection with sensitive CCD camera
    - Run time ca. 4,5 h; 7,5 h; 10h
    - Read lenght 100 -120 bp; 250 – 300 bp; 400+ bp
    - Raw sequence ca. 25 – 35 Mb/run; 80 – 100 Mb/run; 400-600 Mb/run



# Genome Sequencer GS 20/FLX

DNA Library Preparation → emPCR → Sequencing → emPCR → Sequencing

**DNA Library Preparation**
1. DNA fragmentation
2. Fragment end polishing
3. Adaptor ligation
4. Library immobilisation
5. Fill-in reaction
6. Single-stranded template DNA (sstDNA) library isolation
7. sstDNA library quality assessment and quantitation

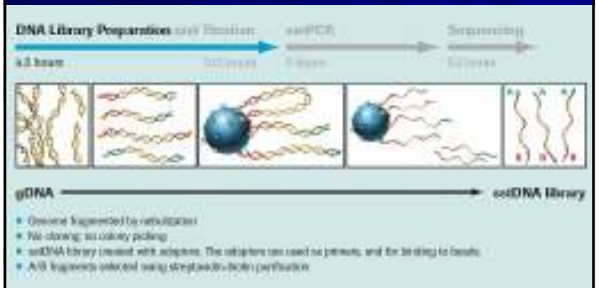
**Emulsion PCR Amplification**
1. Preparation of the live amplification mix
2. sstDNA library capture
3. Emulsification
4. Amplification
5. Bead recovery
6. sstDNA library bead enrichment
7. Sequencing primer annealing

**Sequencing/ Genome Sequencer 20 Operation**
1. The pre-wash Run
2. PicoTiterPlate™ preparation
3. The PREP Run
4. The Sequencing Run

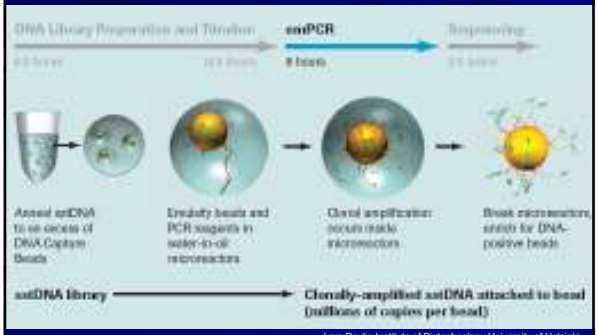
**Output**
1. FASTA file
2. Assembly
3. Mapping



# Library preparation

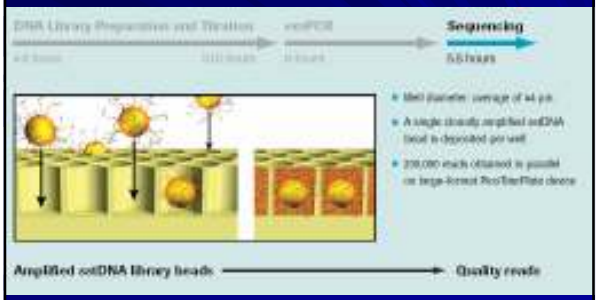


## Emulsion PCR



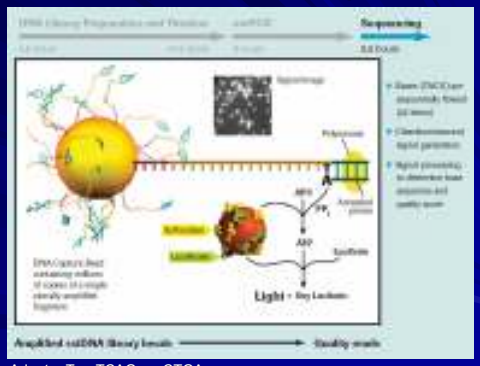

Lars Paulin Institute of Biotechnology University of Helsinki

## PicoTiterPlate (PTP)




Lars Paulin Institute of Biotechnology University of Helsinki

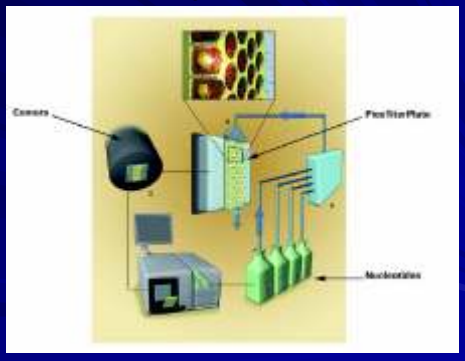
## Pyrosequencing



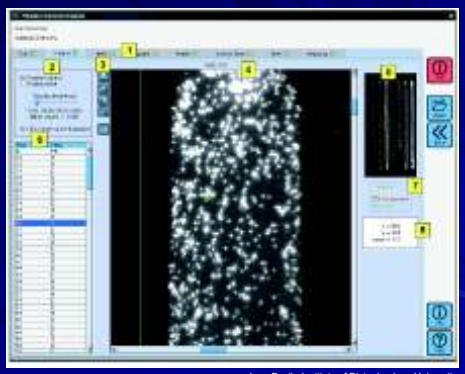

Adaptor Taq TCAG -- CTGA

Lars Paulin Institute of Biotechnology University of Helsinki

## Genome Sequencer GS20/FLX

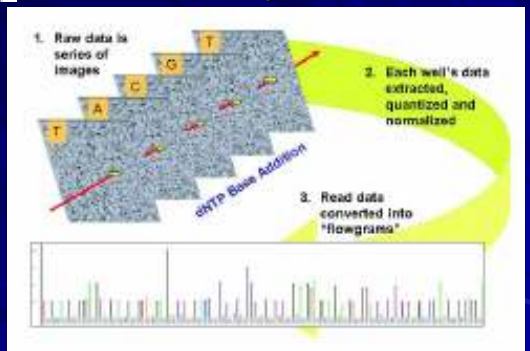


Lars Paulin Institute of Biotechnology University of Helsinki



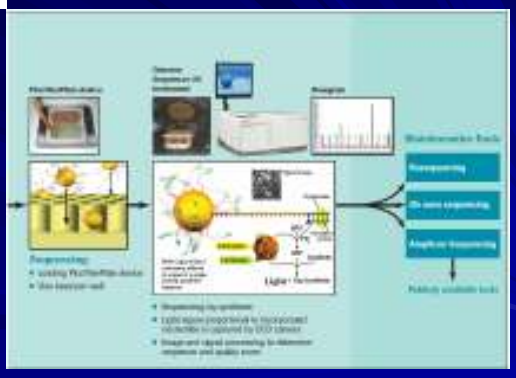
Lars Paulin Institute of Biotechnology University of Helsinki

## Flowgram




Adaptor Taq TCAG -- CTGA

Lars Paulin Institute of Biotechnology University of Helsinki



Lars Paulin Institute of Biotechnology University of Helsinki

## 454 Life Science/Roche

GS Junior

- 100 000 reads
- 400 bp read length
- 35 Mb
- 10 h run time, 2 h data processing

Lars Paulin Institute of Biotechnology University of Helsinki

## Illumina/Solexa Genome Analyzer

(www.illumina.com ; Bentley,DR Curr Opinion Genet Dev 2006, 16,545-552 )

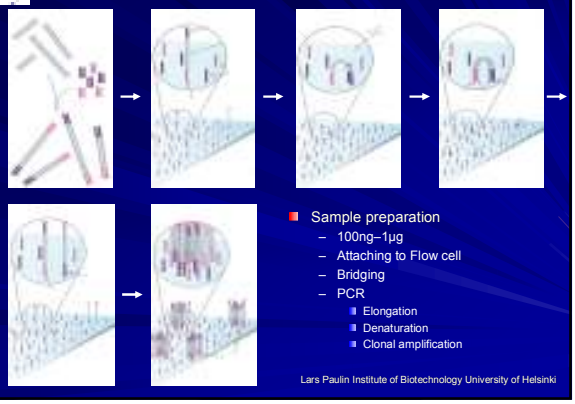
- Clonal Single Molecule Array technology
  - Sequencing-by-synthesis technology
  - Reversible terminator-based sequencing
    - removable fluorescence
  - Flow cell with > 10 million clusters
    - Each cluster ~1,000 copies of template /cm$^2$
  - 1–8 samples / run
  - 3 laser system (660, 635, and 532 nm)
  - Read length 35 - 100 bp, up to 30 Gb / run
    - Run time 3 – 10 days,

Cluster Station

Flow cell

Lars Paulin Institute of Biotechnology University of Helsinki

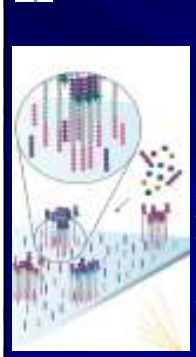
## Illumina/Solexa



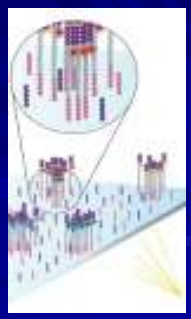
- Sample preparation
  - 100ng–1µg
  - Attaching to Flow cell
  - Bridging
  - PCR
    - Elongation
    - Denaturation
    - Clonal amplification

Lars Paulin Institute of Biotechnology University of Helsinki

## Illumina/Solexa sequencing



Sequencing
- First bases
- Fluorescent reversible terminators
- Detection with laser and CCD camera



Sequencing
- Second bases detected after removal of label and blocking

Lars Paulin Institute of Biotechnology University of Helsinki

## Illumina Instruments

Genome Analyzer IIx
• 2 x 100 bp
• Upto 500 million reads/ flowcell

Genome Analyzer IIe
•2 x 100 bp
• Medium throughput

HiSeq 2000
•2 x 100 bp
•.Upto 300 Gb/run
• 2 billion paired-end reads

Paired-End Module

cBot Cluster Generation System

Lars Paulin Institute of Biotechnology University of Helsinki

## SOLiD, Applied Biosystems

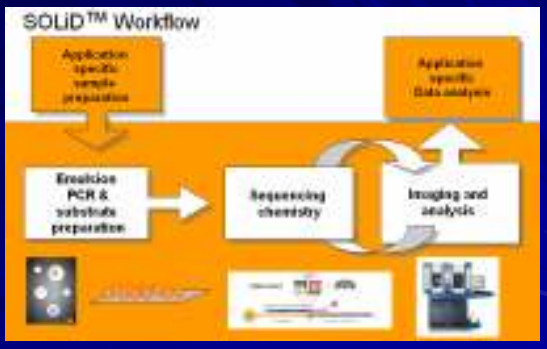
(Life Technologies; www.appliedbiosystems.com)

- Sequencing by Ligation
  - emPCR
    - Small beads, 1μm
  - Attaching to glass slides
  - Labelled probes
    - Fuor colours
    - 2 base encoding system
  - Repeated ligation steps
  - Detection with 4 Mpixel camera
  - Read lenght 50 bp + 35 bp
  - 1-2 slides / run
  - 30 Gb / run / slide
  - Run time 5 -10 days

Shendure, J. *et.al.* Science 2005, 309, 1728-1732

SOLiD v4

Lars Paulin Institute of Biotechnology University of Helsinki




Lars Paulin Institute of Biotechnology University of Helsinki

## SOLiD Libraries


Lars Paulin Institute of Biotechnology University of Helsinki

## Properties of the Probes

Cleavage site is between 5th and 6th base

3' ligation site, cleavage site and dye are spatially separated

Fluorescent dye interrogates base on 1st + 2nd position

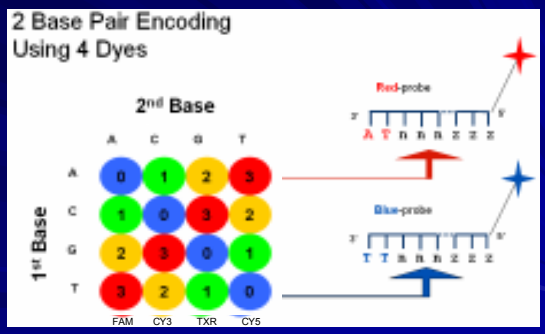
Probes are octamers

N=degenerate bases, Z=universal bases

1024 probes, 256 probes per color

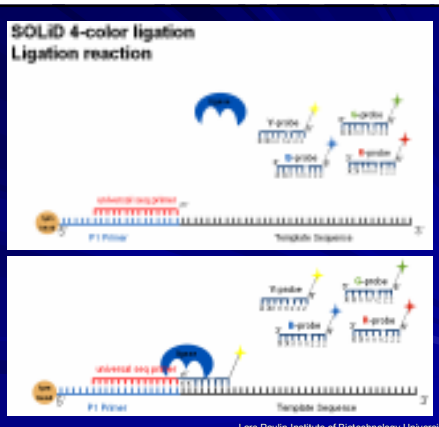
Lars Paulin Institute of Biotechnology University of Helsinki

## 2 Base Pair Encoding Using 4 Dyes

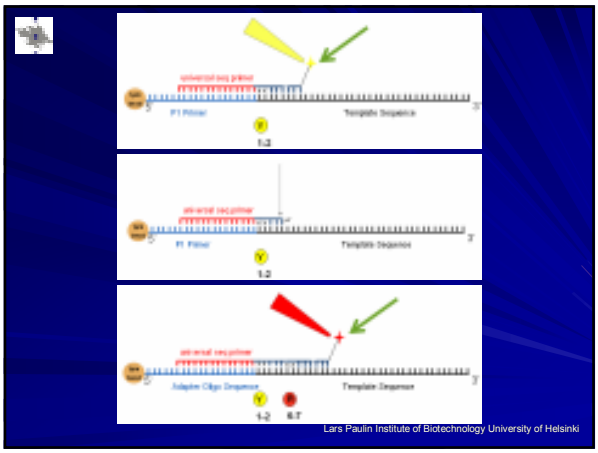


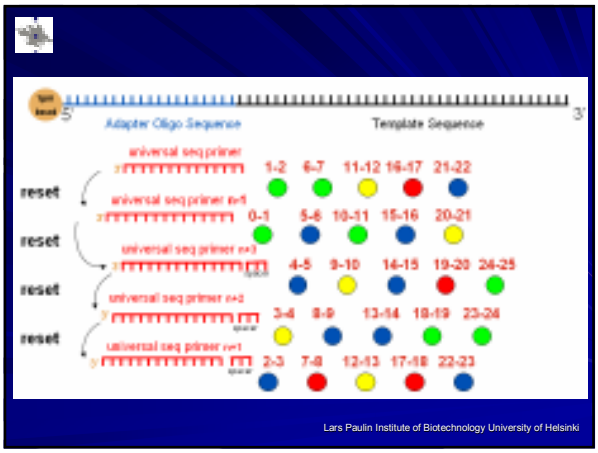

Lars Paulin Institute of Biotechnology University of Helsinki

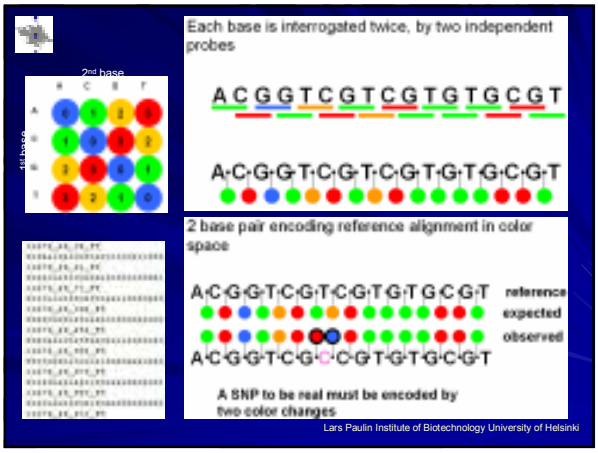
## SOLiD 4-color ligation Ligation reaction



Lars Paulin Institute of Biotechnology University of Helsinki


Lars Paulin Institute of Biotechnology University of Helsinki


Adapter Oligo Sequence
Template Sequence
universal seq primer
reset
1-2 6-7 11-12 16-17 21-22
0-1 5-6 10-11 15-16 20-21
4-5 9-10 14-15 19-20 24-25
3-4 8-9 13-14 18-19 23-24
2-3 7-8 12-13 17-18 22-23
Lars Paulin Institute of Biotechnology University of Helsinki


2nd base
1st base
Each base is interrogated twice, by two independent probes
A C G G T C G T C G T G T G C G T
A·C·G·G·T·C·G·T·C·G·T·G·T·G·C·G·T
2 base pair encoding reference alignment in color space
A·C·G·G·T·C·G·T·C·G·T·G·T·G·C·G·T
reference
expected
observed
A·C·G·G·T·C·G·C·G·T·G·T·G·C·G·T
A SNP to be real must be encoded by two color changes
Lars Paulin Institute of Biotechnology University of Helsinki

## 5500 Series SOLiD Sequencers

**5500**
"lab scale"

- **Lowest start up cost**
- **Seamless upgrade and scale up to 5500xl**

**5500xl**
"Production-scale"

- **Lowest price per sample**
- **Nano-bead ready for maximum throughput**

Lars Paulin Institute of Biotechnology University of Helsinki

## 5500 Series SOLiD™ Sequencer FlowChip

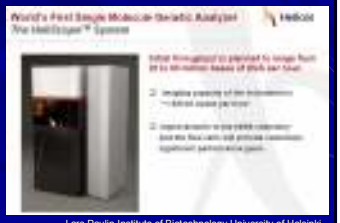
- FlowChip
  - 6 lanes
  - Multiple experiments per lane
  - Barcodes
    - Upto 1152 multiplexing
      - 12 lanes, 96 BC
- Throughput
  - 5500 model 10-15 Gb/day
  - 5500xl model 20-30 Gb/day (30-45 Gb/day nano-beads)
  - Read length
    - Fragment 75 bp
      - 1 day 35 bp, 1 lane
    - Paired-end 75 bp + 35 bp
      - 7 days 6 lanes
    - Mate-pair 60 bp + 60 bp
      - 7 days 6 lanes

Lars Paulin Institute of Biotechnology University of Helsinki

## Helicos

(www.helicosbio.com)
(Harris, T *etal* Science 2008, 320; 106-9

- HeliScope™ Single Molecule Sequencer
  - True Single Molecule Sequencing (tSMS)
  - Sequencing-by-synthesis
  - Template 100 – 200 bp
    - Addition of polyA
  - No PCR amplification
  - 1 000 000 000 reads / experiment
  - 25-90 Mb / h
  - 2 + Gb / day

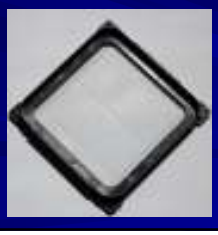


Lars Paulin Institute of Biotechnology University of Helsinki

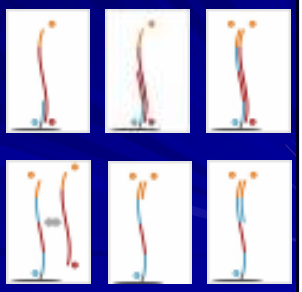
# Helicos

Flow cell

- 25 discrete channels per flow cell
- Single molecule capture by hybridization, allowing densities of 100 million strands of DNA per square centimeter or higher



Paired-end Sequencing
(100 – 200 bp)



Lars Paulin Institute of Biotechnology University of Helsinki

# Applications

- Whole genome sequencing
  - *de novo* sequencing
    - Genome Sequencer FLX
- Comparative sequencing
  - All four systems
- Metagenomics
  - Genome Sequencer FLX
- Amplicon sequencing
  - Mutations / SNP
  - All four systems
- Transcriptome sequencing
  - cDNA
    - All four systems
  - Small RNA
    - All four systems
- ChIP sequencing
  - All four systems
- Methylation sequencing
  - All four systems
- Sequence capture
  - SureSelect or NimbleGen
    - Genome Sequencer FLX
    - SOLiD
    - Illumina

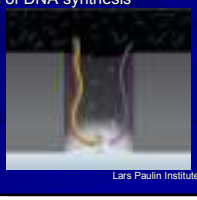
Lars Paulin Institute of Biotechnology University of Helsinki

# Pacific Biosciences

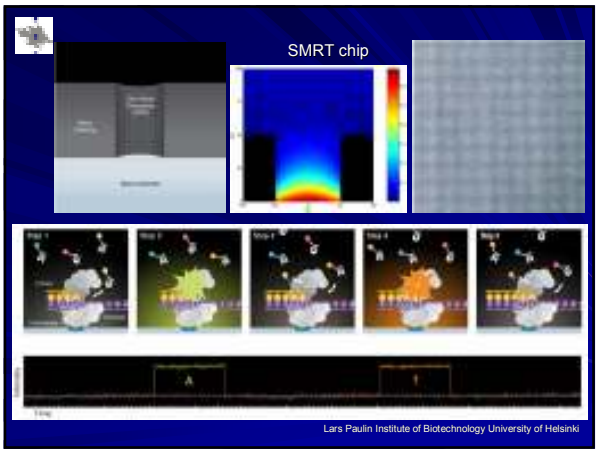
(www.pacificbiosciences.com)

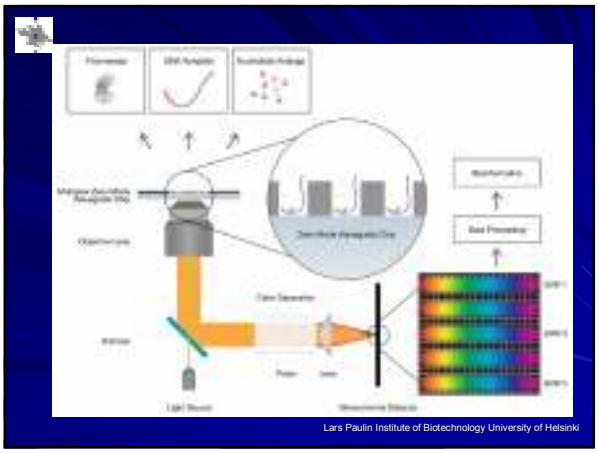
(Korlach, J. *et.al.* PNAS 2008, 105, 1176-81, Levene, MJ. *et.al.* Science 2003, 299, 682-86
Eid, J etal. Science 2009, 323, 133 - 138)

- Technology
  - Single-Molecule Real-Time (SMRT) DNA sequencing technology
  - SMRT chip
    - Thousands of zero-mode waveguides (ZMWs)
    - Holes 100 nm metal film, 20 zeptoliters ($10^{-21}$ liters)
  - Real-time detection of DNA synthesis
    - Fluorescent dNTPs



Lars Paulin Institute of Biotechnology University of Helsinki

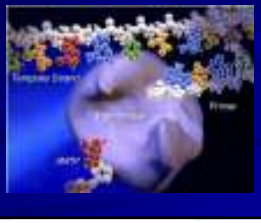



## VisiGen

(www.lifetechnologies.com)

- Technology
  - No cloning or amplification
  - Intact DNA fragments
  - Real-time detection of DNA synthesis, FRET
  - Fluorescent donor on tip of the Polymerase attached on a glass slide
  - Acceptor fluorescent moiety on the nucleotides
    - On the gamma-phosphate
  - 1Mb/sec/machine



Lars Paulin Institute of Biotechnology University of Helsinki

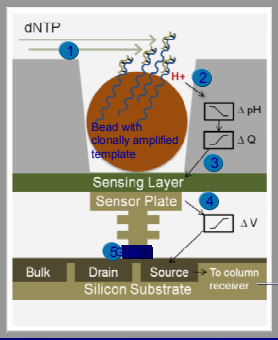
## Ion Torrent

(www.iontorrent.com, www.lifetechnologies.com)

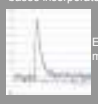
- Semiconductor sequencing
  - Real-time detection of Polymerase activity
  - Use normal dNTPs
  - Nucleotide incorporation releases an $H^+$ ion
  - Detection of pH change with the ion detector
  - Cheap to use
  - Currently 100-200 bp reads

Lars Paulin Institute of Biotechnology University of Helsinki

## Ion Torrent, Detection Chemistry




1. Nucleotide bases (dNTP's) are sequentially flowed into well one at a time
2. Upon incorporation, the nucleotide releases a hydrogen ions which creates a pH change
3. Sensing layer binds to the hydrogen ions
4. Sensing plate transmits ions to the field effective transistor (FET) gate
5. Gate registers voltage change between source and drain proportional to number of bases incorporated



Electric voltage measurement

Lars Paulin Institute of Biotechnology University of Helsinki

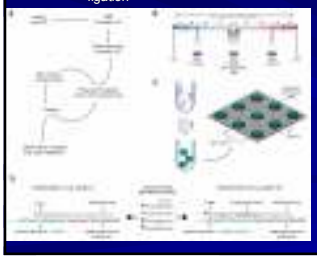
## Other Sequencing Technolgies

Complete genomics

- www.completegenomics.com
- Science 2010,327, 78 - 81
  - DNA Nanoballs on array
  - Combinatorial probe-anchor ligation

Oxford Nanopore Technologies

- www.nanoporetech.com
- Nat Nanotechnol 2009, 4, 265 - 270
  - Exonuclease DNA sequencing
  - Strand sequencing



Lars Paulin Institute of Biotechnology University of Helsinki

(www.genomics.xprize.org)

## $10M to the First Team to Sequence 100 Human Genomes in 10 Days

RegisteredTeams

- 454 Life Sciences (Roche) (www.454.com )
- VisiGen (www.visigenbio.com )
- FfAME (www.ffame.org )
- Reveo (www.reveo.com)
- Base4innovation (www.base4innovation.co.uk )
- Personal Genome X-Team (PGx) (www.personalgenomes.org)
- ZS Genetics, Inc. (www.zsgenetics.com)
- Cracker (www.crackerbio.com)

Lars Paulin Institute of Biotechnology University of Helsinki