xiaomanshawnli@gmail.com

Office hour: MW10:30am-11:30am

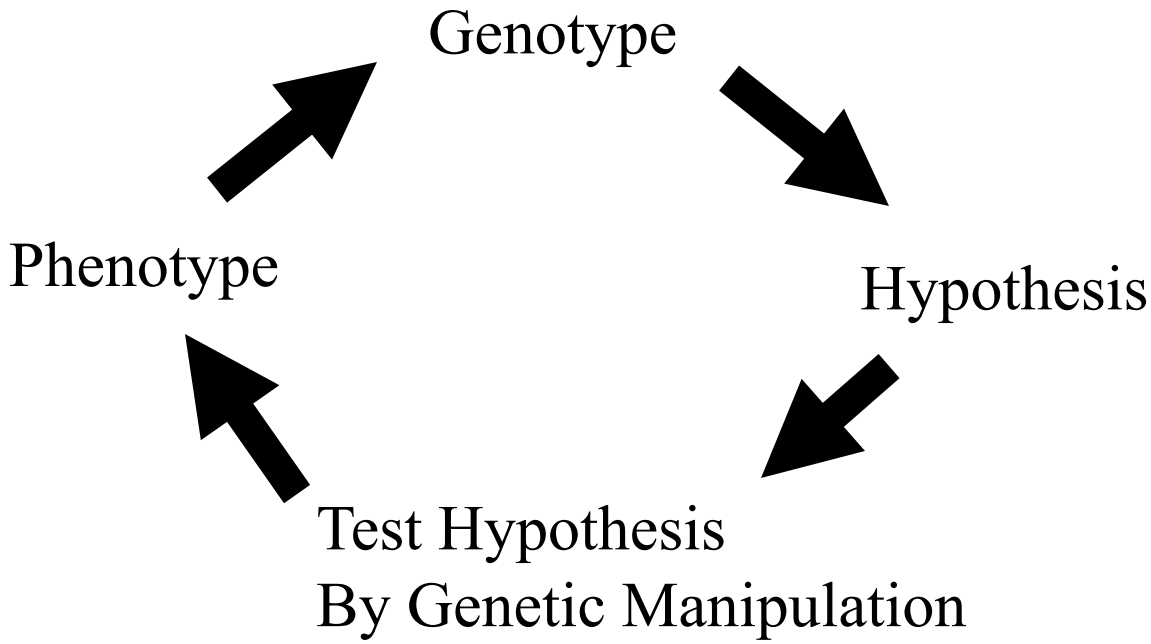
http://www.cs.ucf.edu/~xiaoman/spring/

Next Generation Sequencing Technologies

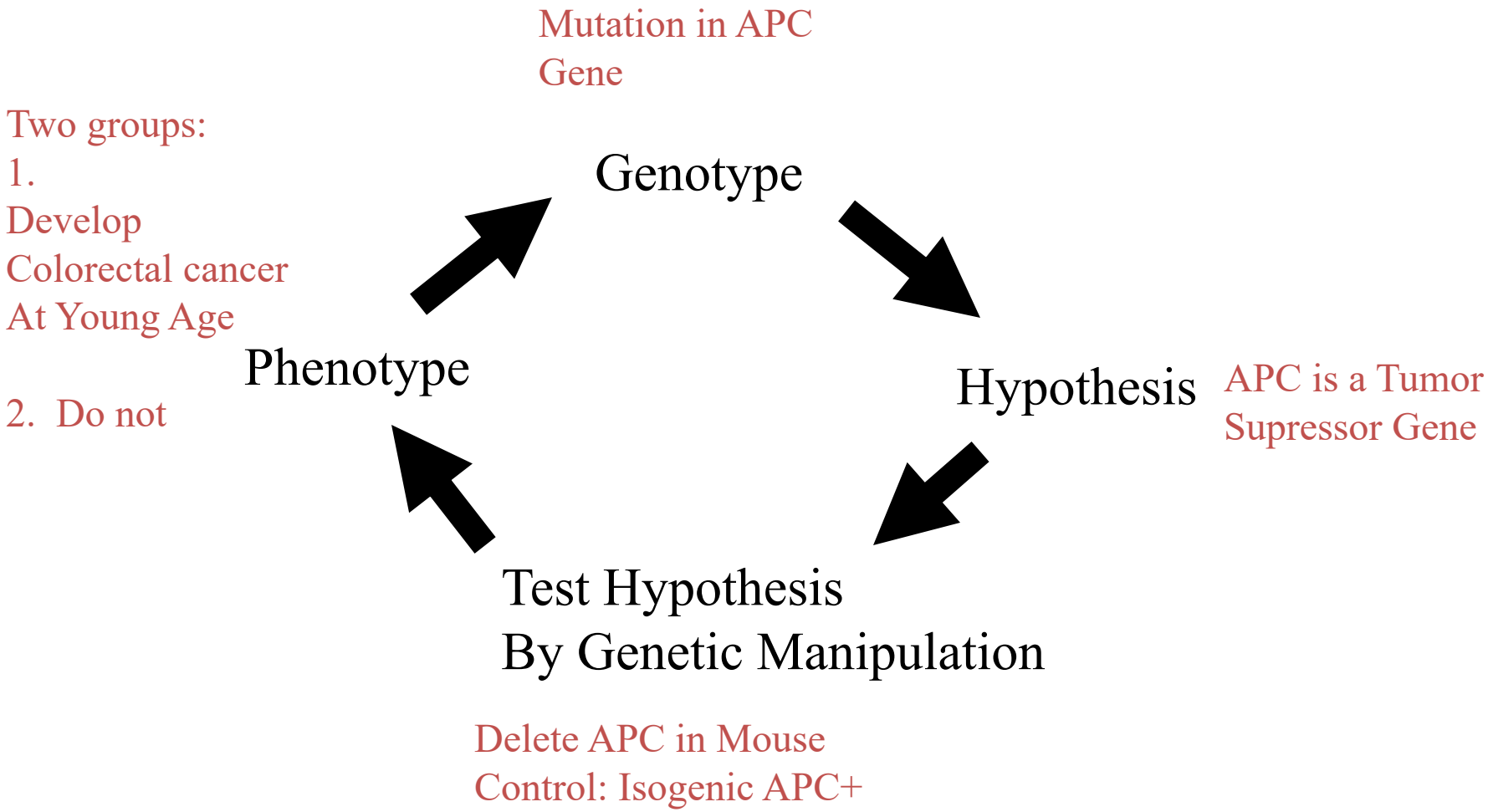
Some slides are modified from Robi Mitra's lecture notes

What will you do to understand a disease?

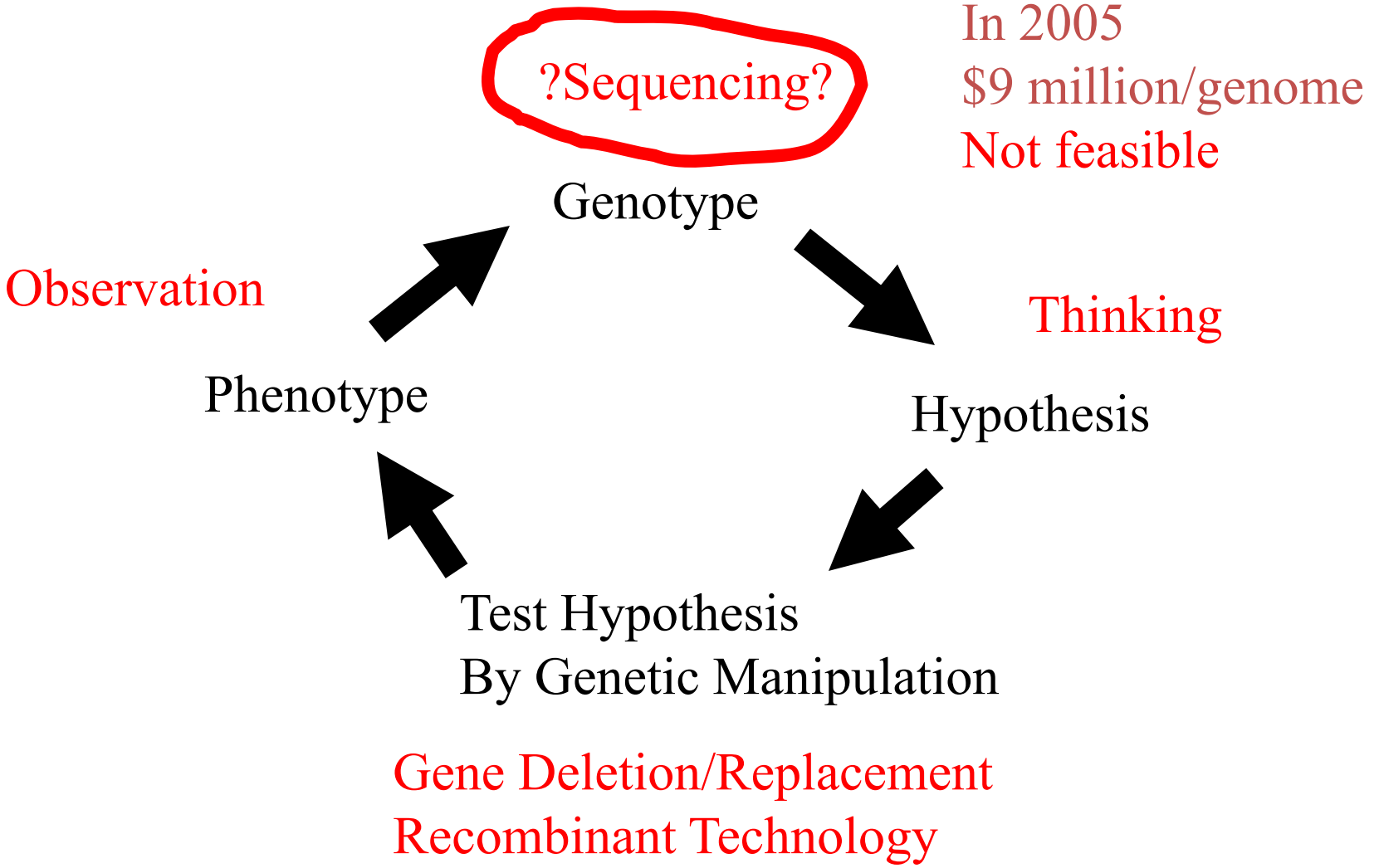
What will you do to understand a disease?



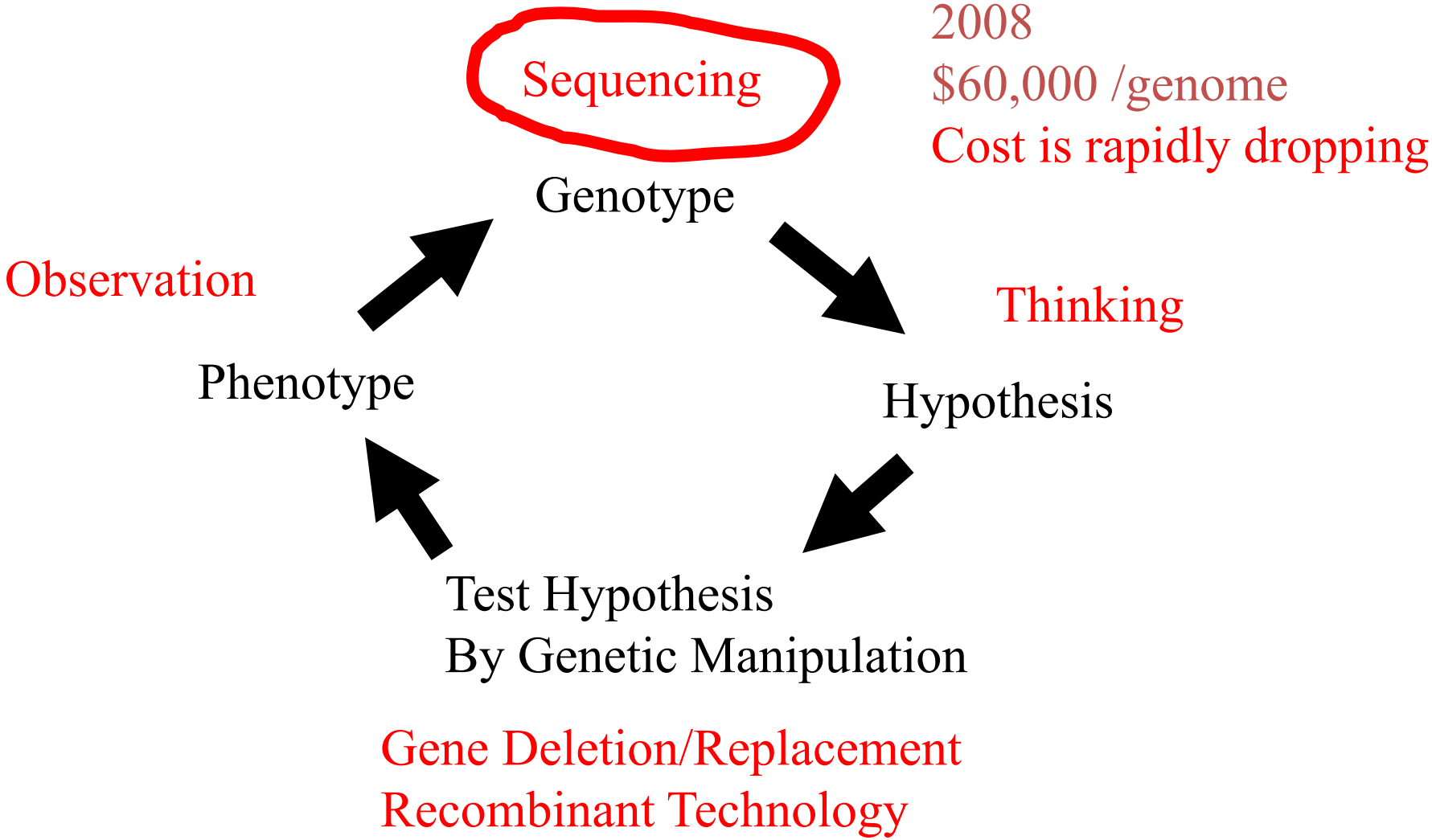
Forward Genetics



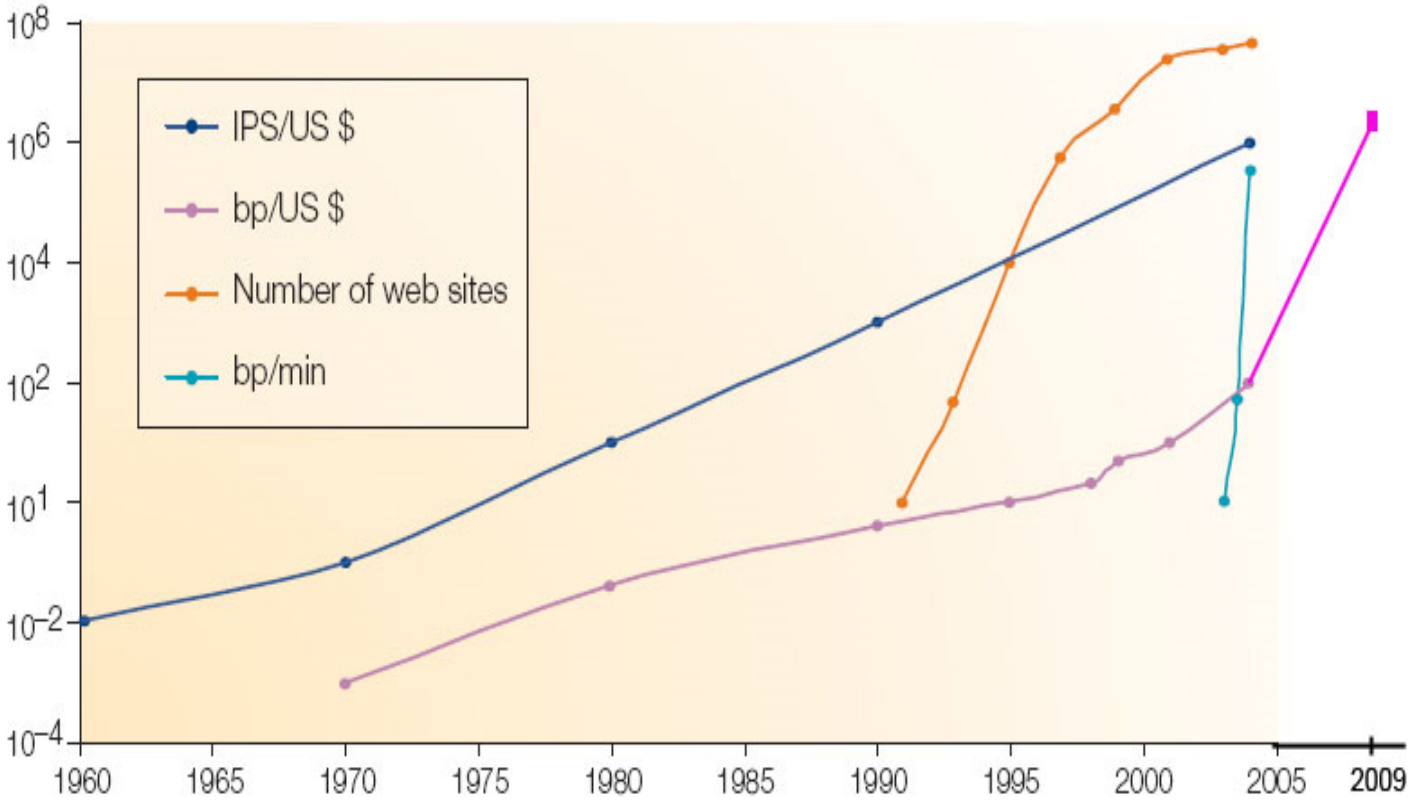
The Cycle of Forward Genetics



The Problem with Forward Genetics



Bp/US dollar: increases exponentially with time



Adapted from Jay Shendure et al 2004

Two questions:

- How was this dramatic acceleration achieved?

- What does it mean?

How was this achieved?

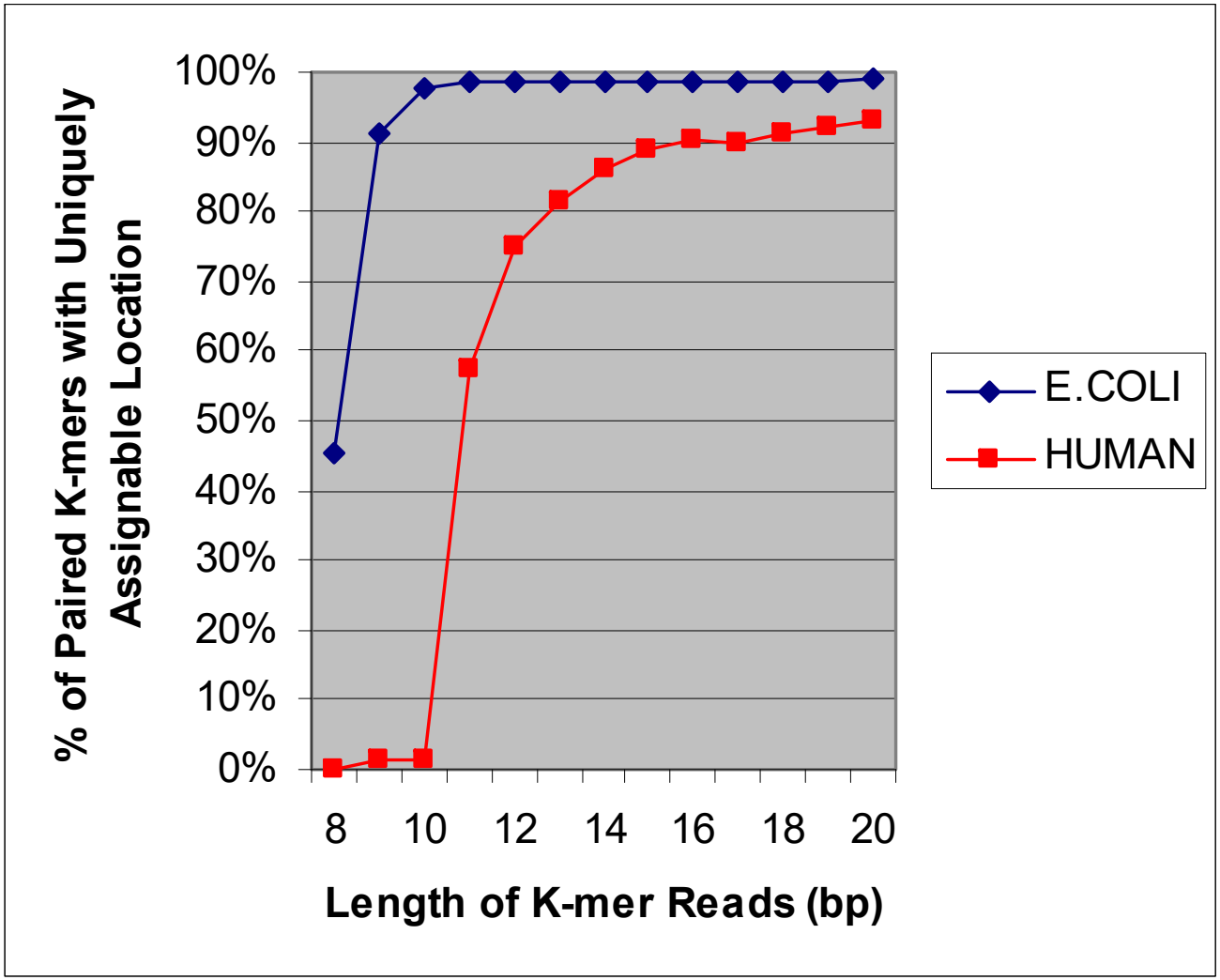
- Integration (Think about sequencing pipeline)
- Parallelization
- Miniaturization

Same concepts that revolutionized integrated circuits

Plus one additional insight

Read length does not matter

Read Length is Not As Important For Resequencing



Jay Shendure

Second generation sequencers

- 454 Life Sciences (Roche Diagnostics)
  - 25-50 MB of sequences in a single run
  - Up to 500 bases in length
- Solexa (Illumina)
  - 1 GB of sequences in a single run
  - 35 bases in length
- SOLiD (Applied Biosystems)
  - 6 GB of sequences in a single run
  - 35 bases in length

Comparing Sequencers

	Roche (454)	Illumina	SOLiD
Chemistry	Pyrosequencing	Polymerase-based	Ligation-based
Amplification	Emulsion PCR	Bridge Amp	Emulsion PCR
Paired ends/sep	Yes/3kb	Yes/200 bp	Yes/3 kb
Mb/run	100 Mb	1300 Mb	3000 Mb
Time/run	7 h	4 days	5 days
Read length	250 bp	32-40 bp	35 bp
Cost per run (total)	$8439	$8950	$17447
Cost per Mb	$84.39	$5.97	$5.81

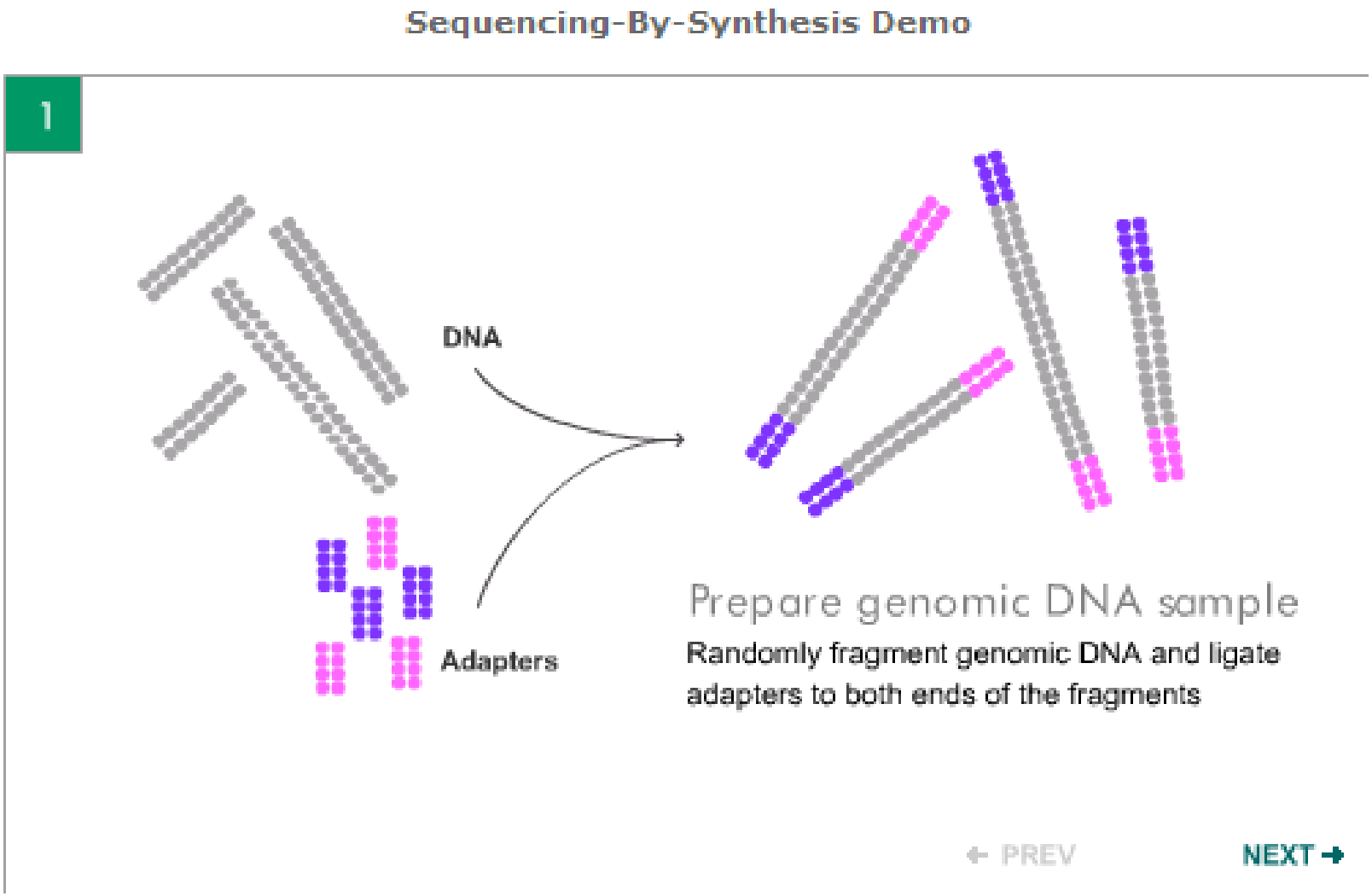
From Stefan Bekiranov, Univ of Virginia, 2008

NGS Technology Comparison

	ABI SOLiD	Illumina GA	Roche FLX
Cost	SOLiD 4: $495k SOLiD PI: $240k	IIe: $470k IIx: $250k HiSeq: $690k	Titanium: $500k
Quantity of Data per run	SOLiD 4: 100Gb SOLiD PI: 50Gb	IIe: 20 - 38 Gb IIx: 50 – 95 Gb HiSeq: 200Gb +	450 Mb
Run Time	7 Days	4 Days	9 Hours
Pros	Low error rate due to dibase probes	Most widely used NGS platform. Requires least DNA	Short run time. Long reads better for de novo sequencing
Cons	Long run times. Has been demonstrated certain reads don't match reference	Least multiplexing capability of the 3. Poor coverage of AT rich regions	Expensive reagent cost. Difficulty reading homopolymer regions

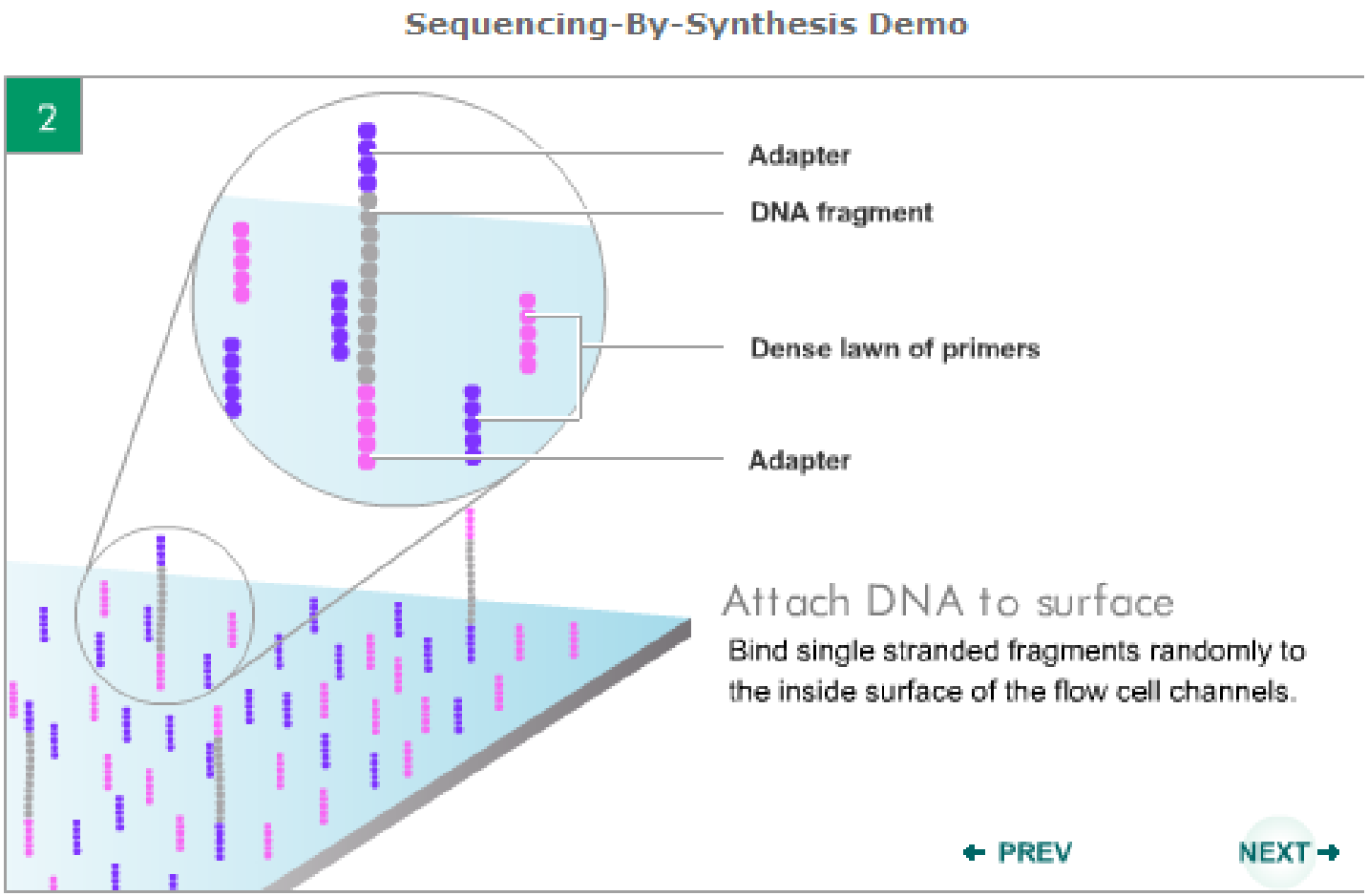
London Local Genomics Center April 1, 2010 NGS meeting

Technology Overview: Solexa/Illumina Sequencing



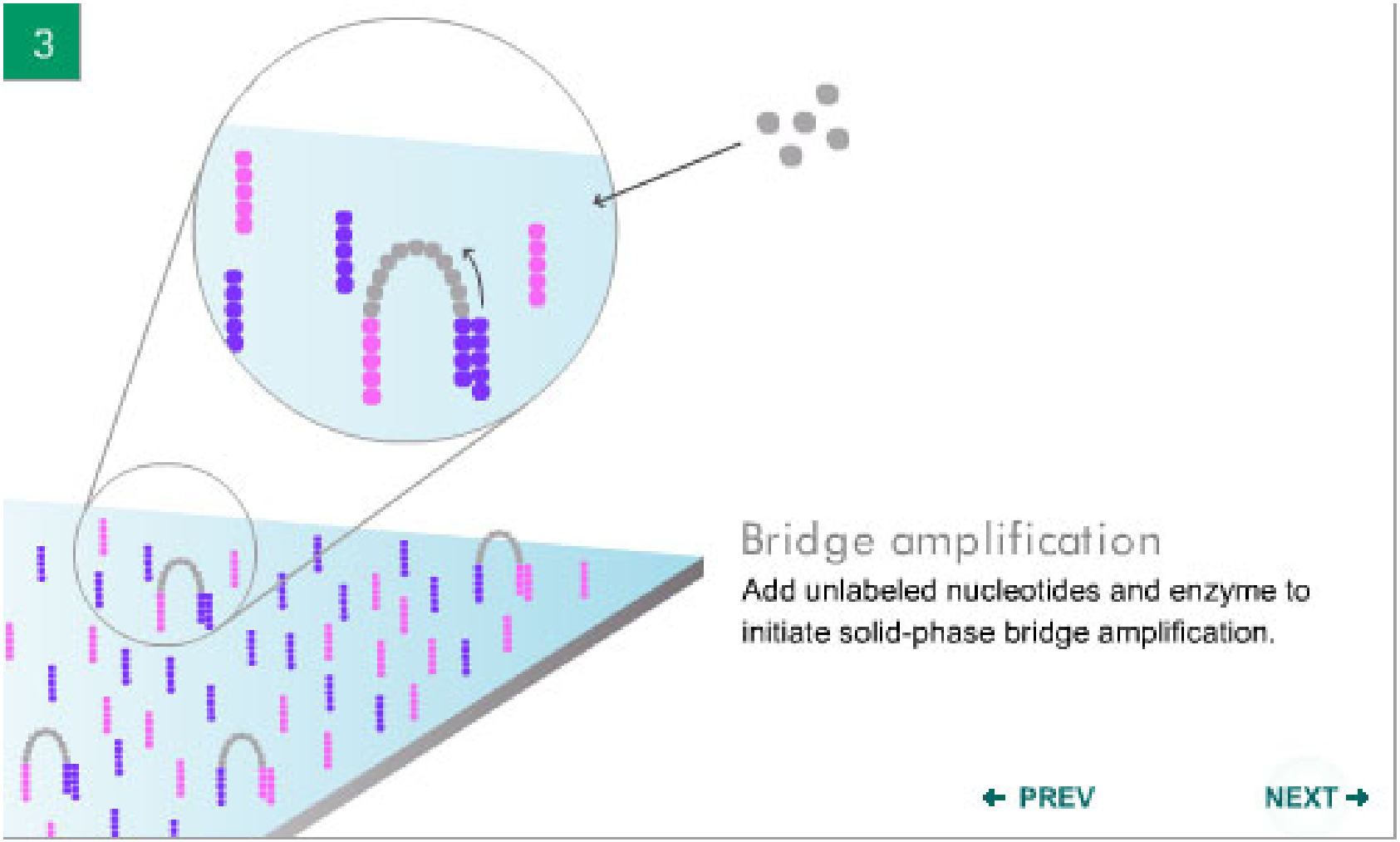
http://www.illumina.com/

Immobilize DNA to Surface

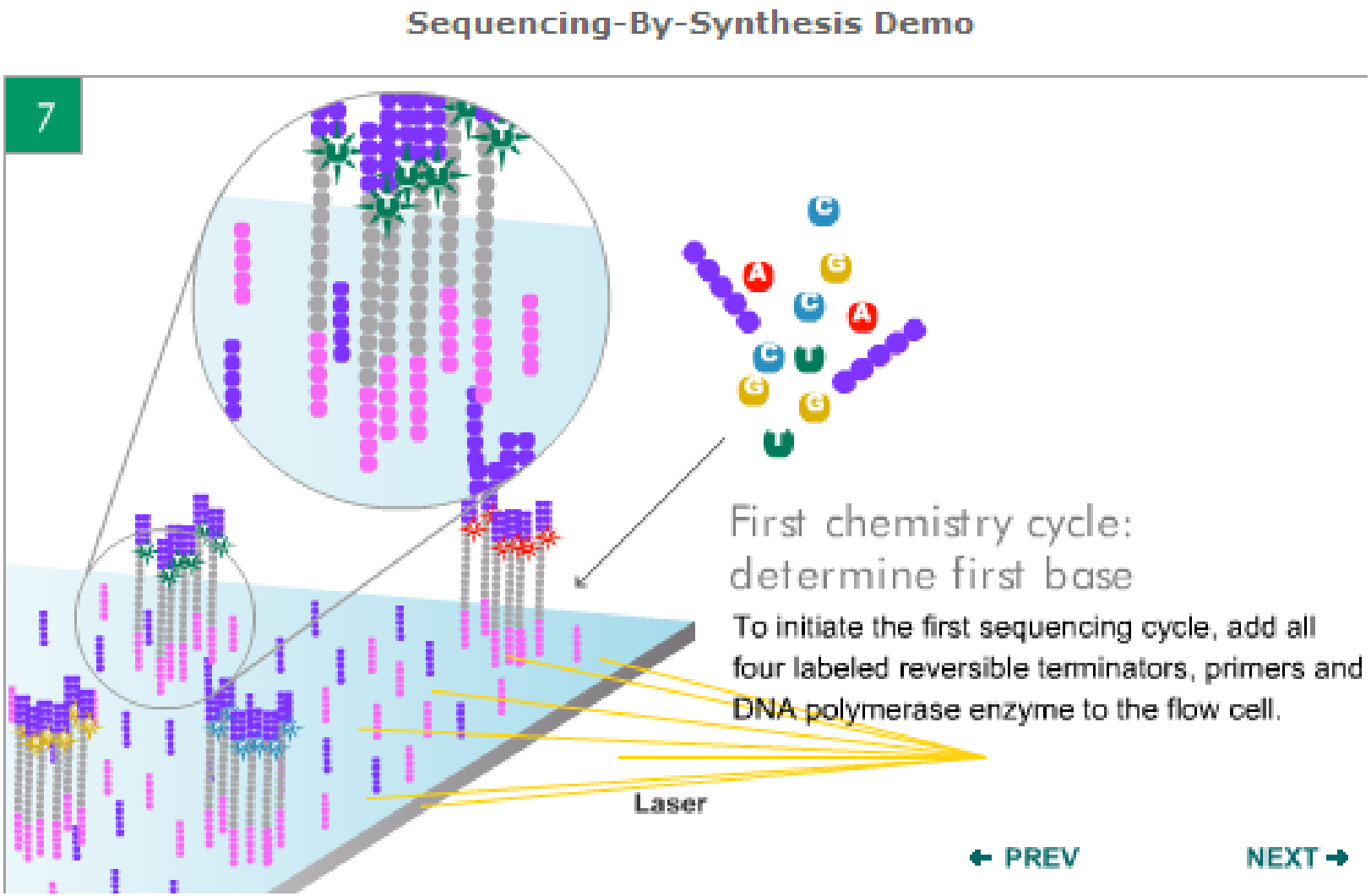


Source: www.illumina.com

Technology Overview: Solexa Sequencing



Sequence Colonies



Sequence Colonies

10

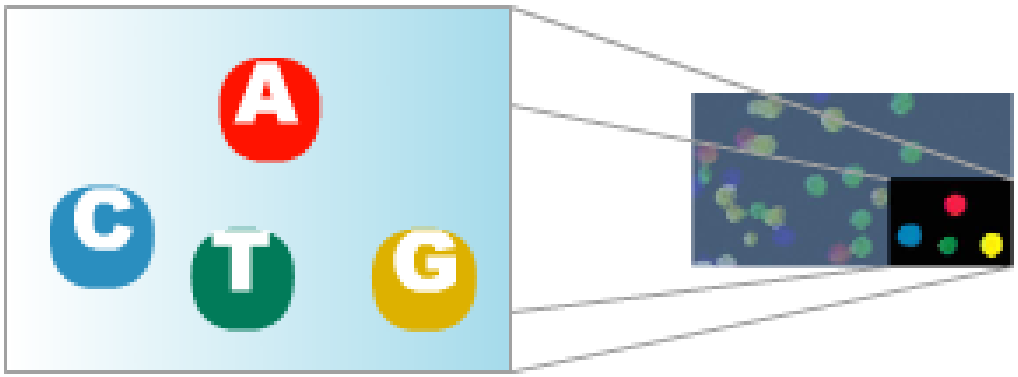
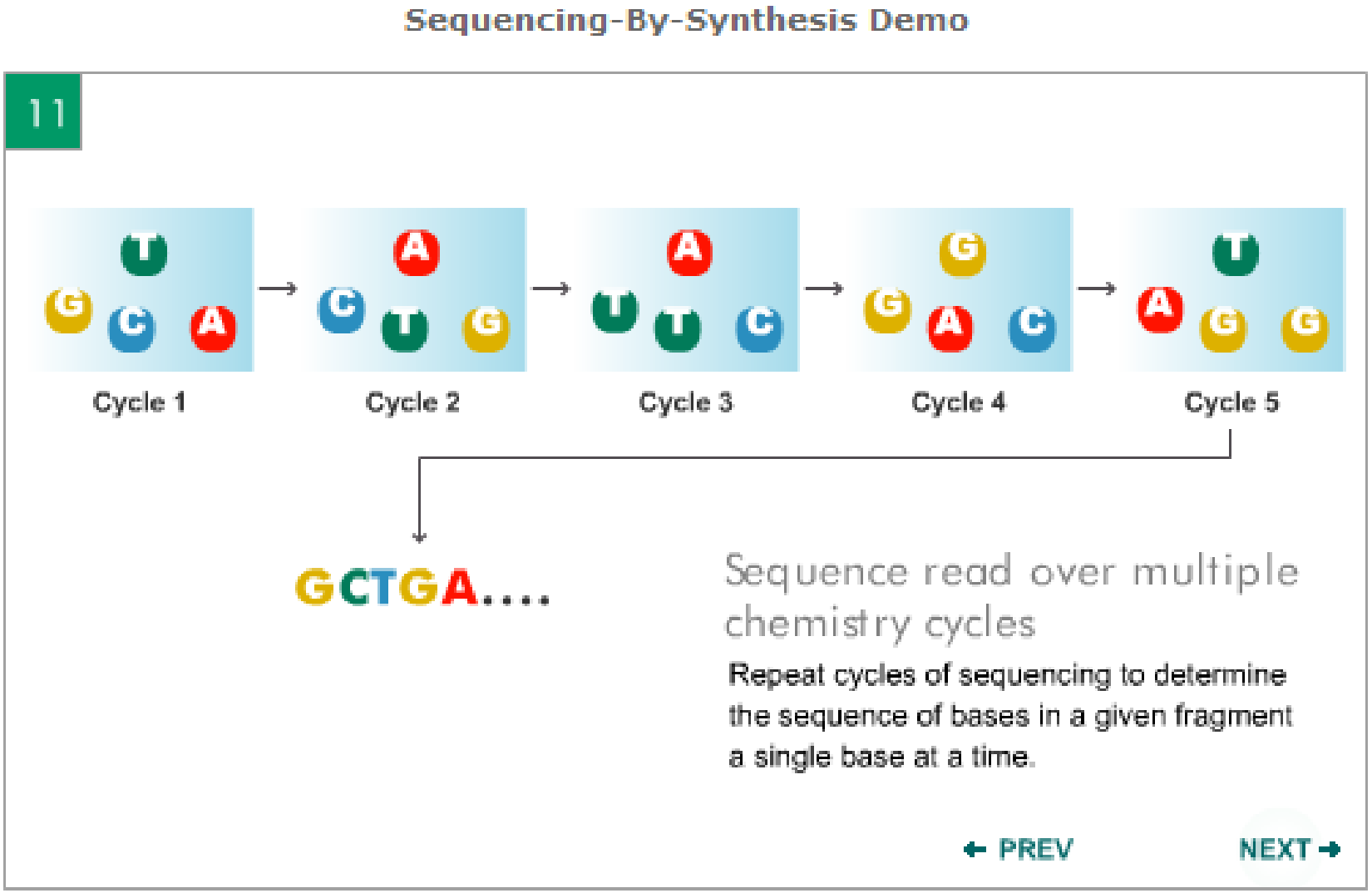


Image of second chemistry cycle is captured by the instrument

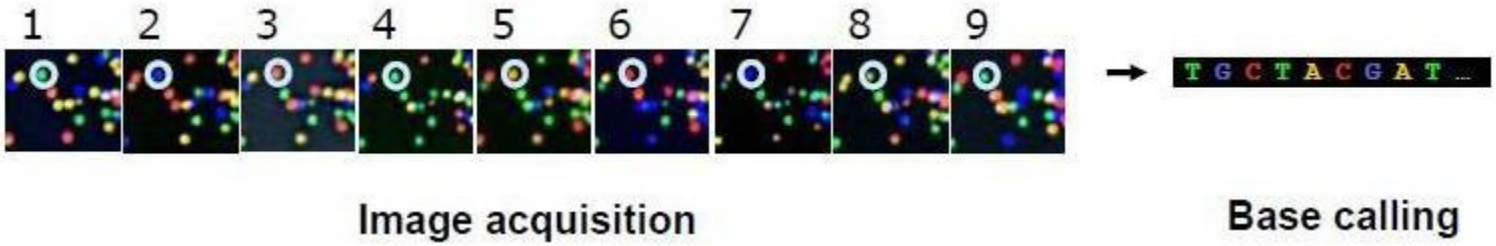
After laser excitation, collect the image data as before. Record the identity of the second base for each cluster.

← PREV NEXT →

Call Sequence



Sequencer Output



Quality Scores

Sequence Files

Seque

- ~10 million sequences per lane
- ~500 MB files

1	1	203	322	TGAAAAATTAATGAAATATATGCTATCCGCTCACAC
1	1	229	353	TTAAGATTTTAAATTATTTAGGGTGCATCAGCTTCC
1	1	110	436	TATAAGTTAATATTGTGTATAACCTTTTAGCCACAC
1	1	211	303	TTCTTAACAGGGTGAGTTCCCTGGTTATCCAATACC
1	1	99	329	TTTTATACTTCATGGTTTTTGTGGTGTCAAAAATCT
1	1	221	277	GAATGTATTCCAATATCAAAGAGCAAATTCCACCAC
1	1	225	370	TGATAAGTATAAGTGATTATTGTAATTATGTTTGAG
1	1	243	244	GTTGAACATTCCTTTTCATAGAGCAGTGTTGACACA
1	1	186	360	TATACCACTGTGCATGTTAATAAACGAGGTTGTTTG
1	1	167	333	TAGATAGCTAGGTTGGGAAGTGAAATGATCAGCTTT
1	1	213	333	TTAAAAAAAAAGAAAAAAAAAAGAAAACTAGGCTAC
1	1	244	338	TTCTTTTGTTCCCTAACCTGCCGGACTCCTTCCCAC
1	1	120	382	TAATTTTTATGTTTTTGTAGAGATGGGGTTTCTCCG
1	1	117	201	TCTTTAAAGATCTTTCCGGACACTTTTGGAAAAGAG
1	1	232	321	GATGCATTGCTATGCCTCCCAGTCCGCAACTTCACG
1	1	205	318	TACCCCTTGACTCTTCTTTTGTTACCATTTTTCCCC
1	1	102	449	TTTCAAATTATTATTTATTGCTCATTGTGTTTTTTG
1	1	209	276	TATTCAAAAACAATTTGTTTAAATTTAAAAATGAAC
1	1	197	309	TCATATGAAGCAAATTGTTTTGATCAACTCTCATAT
1	1	209	341	TATGCAAGGAAACAGTTCGCGATGCTCCCGTTTGCG
1	1	201	239	TTTCAATATATGCAGTCTGGTTCCAGAGTTTTTAAT
1	1	247	502	TCTCAATTTGCTATTGTAGTTATTGTTTTACTGTTG
1	1	119	420	GGTGAAGAAACAAAGGCCTGCAAAGTTCCTTCCTAC
1	1	323	445	TTAAGGTACTCAGCACTTTCTACGGCATTACGCGGG
1	1	244	254	GTTAAGTTTGGCCTCTTGCCTGGCATCACTTGCCTT
1	1	233	321	GGACAATTGCAATGCTCACAATTCGGAAACTTCCGC
1	1	95	416	TGGTTGGTACATTTCACATAAATGGAATCACATAAT
1	1	100	587	TAATGTTAAACTGTTAATAATGCTTGCTCCCAGGAA
1	1	119	481	GAACCCAGAAATCACACCTCAGTTTATCCTGGGCCT
1	1	239	312	GGAACCGTCTTCGACTGTGCCGCCTGACGCAAAGGC
1	1	101	312	TGGACAAAGAAGGTGTCTGGGCAATAGAAACAGTGT
1	1	145	341	TCTTCTTGTAAATTTGTTTTAAGTTTTTTATATATG
1	1	530	242	ACCCACACAACTGAACCCACATCACATGACAAGACT
1	1	224	220	TGTTTGTTGAACTCCCGTCATATTGGCTCCCTTGCT
1	1	364	491	TATCTCTTCGTAGCCCCTCTGTGTATGTTCTTCCTC
1	1	214	608	GTTGTGATTGCTCATTAAGACTCTGAACAATACTCA
1	1	196	533	TTCTACGTGTGGCCTTCAGTACTTTTCTTGGGCCTT
1	1	174	351	TCGACGCCGTTTCCCTTCGGGTCCACACGGTGTTTG
1	1	116	344	GAATTGAATCAATTCGGAGACTGTGCGATCGGCCGC
1	1	215	533	TAAGTGTCTATCACGGCCAAGACGCAGGCTGGGTGC
1	1	223	207	TTCTGTTTAAATGCTTGTTCGATGGCTTGTTAGAAG
1	1	121	377	GGCGGGGCGGGGGAGACGCCGGGCCCAGCCCGCCCC

https://www.youtube.com/watch?v=9YxExTSwgPM

Illumina FASTQ format

FASTQ format stores sequences and Phred qualities in a single file. It is concise and compact. FASTQ is first widely used in the Sanger Institute and therefore we usually take the Sanger specification and the standard FASTQ format, or simply FASTQ format. Although Solexa/Illumina read file looks pretty much like FASTQ, they are different in that the qualities are scaled differently. In the quality string, if you can see a character with its ASCII code higher than 90, probably your file is in the Solexa/Illumina format.

Each sequence entry consists of 4 lines:

sequence name after @

sequence

quality score name after + (optional)

quality scores in phred format (http://maq.sourceforge.net/qual.shtml)

FASTQ examples

@EAS54_6_R1_2_1_413_324
CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;;;7;;;;;;;88
@EAS54_6_R1_2_1_540_792
TTGGCAGGCCAAGGCCGATGGATCA
+
;;;;;;;;;;;7;;;;;;;;3;83
@EAS54_6_R1_2_1_443_348
GTTGCTTCTGGCGTGGGTGGGGGGG
+EAS54_6_R1_2_1_443_348
;;;;;;;;;;;9;7;;.7;393333

Quality Score

•Given a character **$q**, the corresponding Phred quality can be calculated with:

$Q = ord($q) - 33; where **ord()** gives the ASCII code of a character.

Solexa/Illumina Read Format

The syntax of Solexa/Illumina read format is almost identical to the FASTQ format, but the qualities are scaled differently. Given a character **$sq**, the following Perl code gives the Phred quality **$Q**:

$Q = 10 * log(1 + 10 ** (ord($sq) - 64) / 10.0) / log(10);

Bioinformatics Challenges

- Rapid mapping of these short sequence reads to the reference genome
- Visualize mapping results
  - Thousand of enriched regions
- Peak analysis
  - Peak detection
  - Finding exact binding sites
- Compare results of different experiments
  - Normalization
  - Statistical tests

Mapping of Short Oligonucleotides to the Reference Genome

- Mapping Methods
  - Need to allow mismatches and gaps
    - SNP locations
    - Sequencing errors
  - Indexing and hashing
    - genome
    - oligonucleotide reads
- Use of quality scores
- Use of SNP knowledge
- Performance
  - Partitioning the genome or sequence reads

Mapping Methods: Indexing the Genome

- Fast sequence similarity search algorithms (like BLAST)
  - Not specifically designed for mapping millions of query sequences
  - Take very long time
    - e.g. 2 days to map half million sequences to 70MB reference genome (using BLAST)
  - Indexing the genome is memory expensive

Mapping Methods: Indexing the Oligonucleotide Reads

- ELAND (Cox, unpublished)
  - "Efficient Large-Scale Alignment of Nucleotide Databases" (Solexa Ltd.)
- SeqMap (Jiang, 2008)
  - "Mapping massive amount of oligonucleotides to the genome"
- RMAP (Smith, 2008)
  - "Using quality scores and longer reads improves accuracy of Solexa read mapping"
- MAQ (Li, 2008)
  - "Mapping short DNA sequencing reads and calling variants using mapping quality scores"

Main features

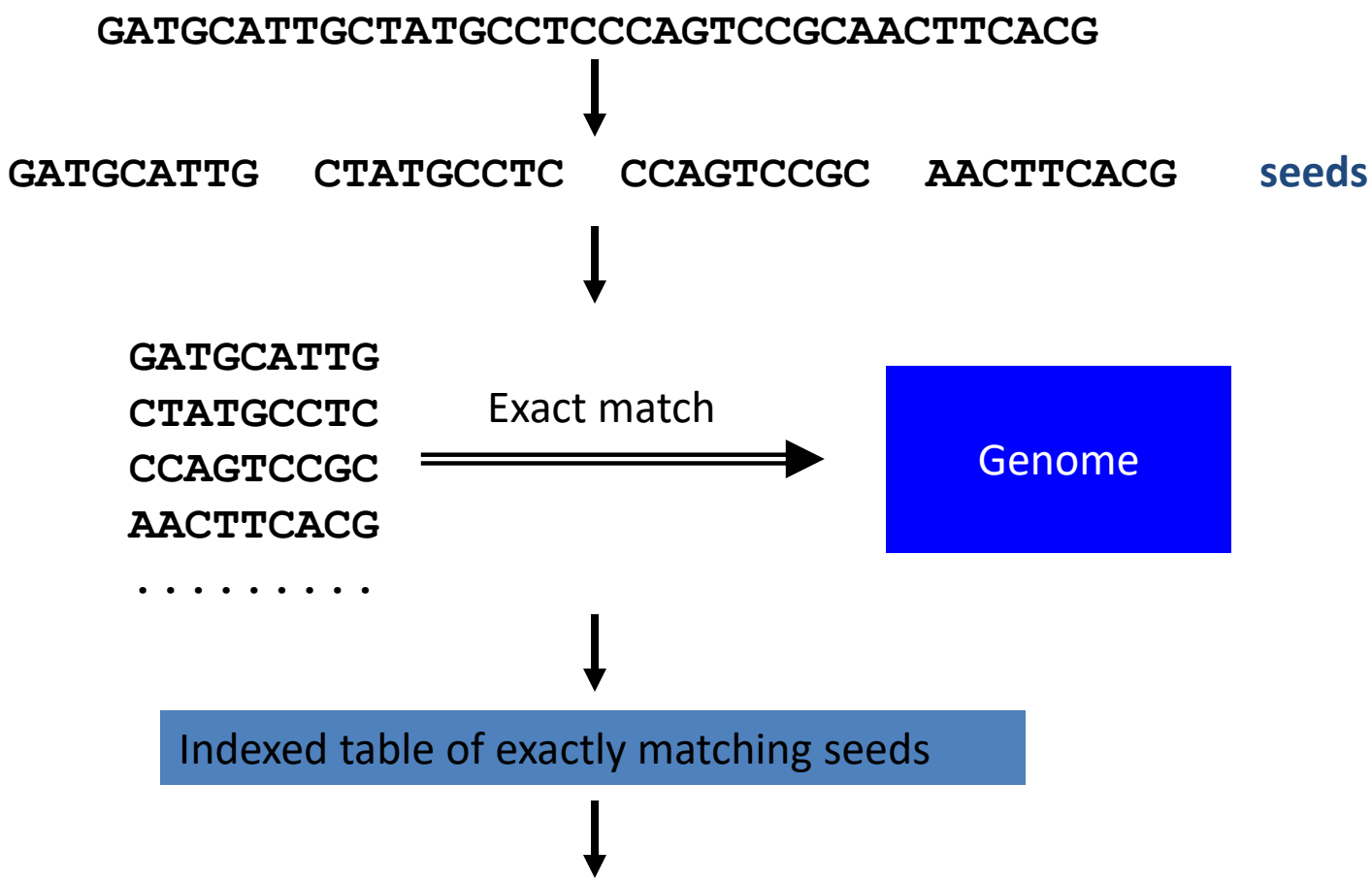
Eland: read length 20-32nt, 2 substitutions

SOAP: 2 substitutions, 1 indel of 1-3nt without any substitutions

RMAP: ungapped mapping, take the tag quality into account

SeqMap: allow <=5 mixture of substitutions and insertions/deletions

Mapping Algorithm (2 mismatches)

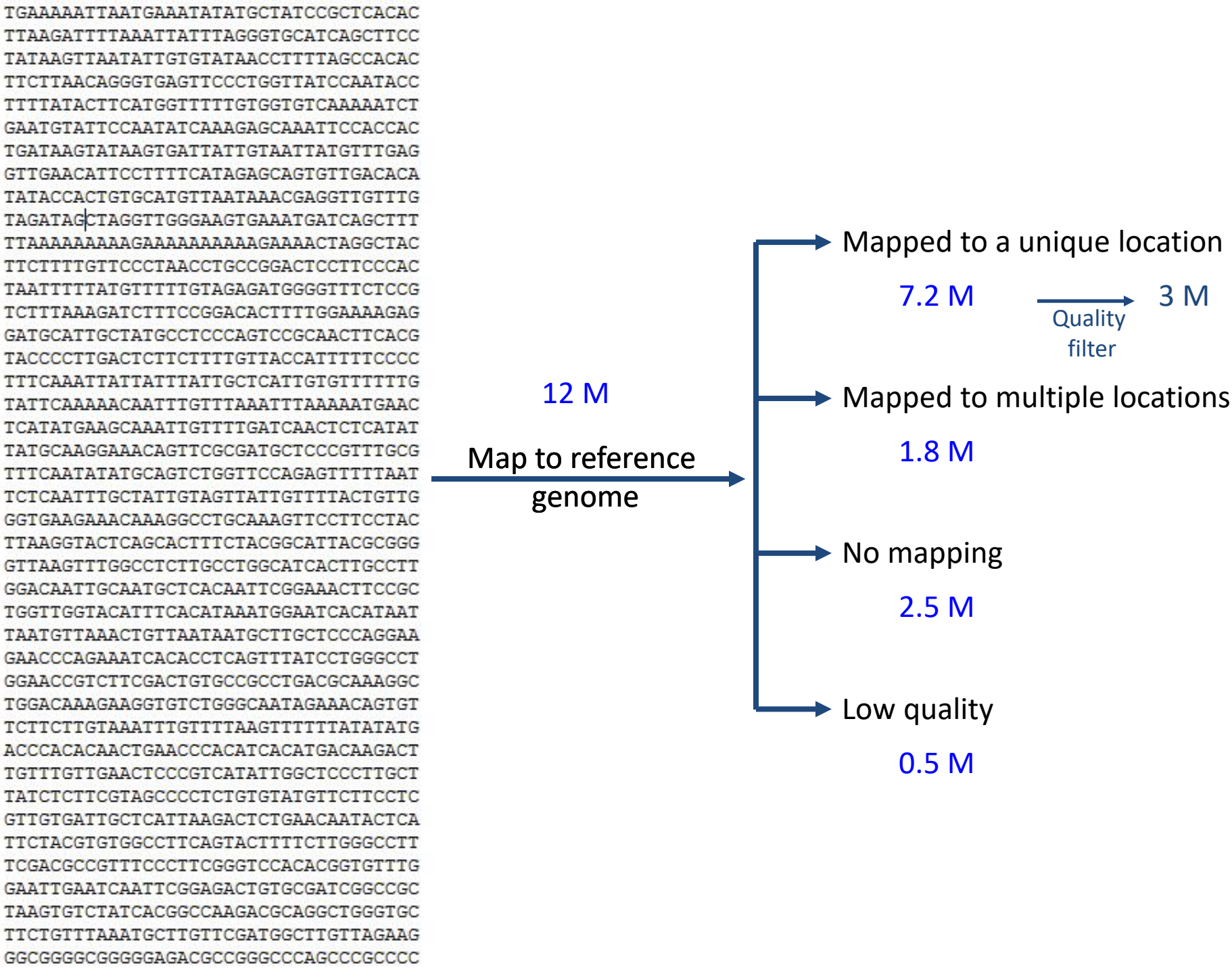


Mapping Algorithm (2 mismatches)

- Partition reads into 4 seeds {A,B,C,D}
  - At least 2 seed must map with no mismatches
- Scan genome to identify locations where the seeds match exactly
  - 6 possible combinations of the seeds to search
    - {AB, CD, AC, BD, AD, BC}
  - 6 scans to find all candidates
- Do approximate matching around the exactly-matching seeds.
  - Determine all targets for the reads
  - Ins/del can be incorporated

- The reads are indexed and hashed before scanning genome

- Bit operations are used to accelerate mapping
  - Each nt encoded into 2-bits


TGAAAAATTAATGAAATATATGCTATCCGCTCACAC
TTAAGATTTTAAATTATTTAGGGTGCATCAGCTTCC
TATAAGTTAATATTGTGTATAACCTTTTAGCCACAC
TTCTTAACAGGGTGAGTTCCCTGGTTATCCAATACC
TTTTATACTTCATGGTTTTTGTGGTGTCAAAAATCT
GAATGTATTCCAATATCAAAGAGCAAATTCCACCAC
TGATAAGTATAAGTGATTATTGTAATTATGTTTGAG
GTTGAACATTCCTTTTCATAGAGCAGTGTTGACACA
TATACCACTGTGCATGTTAATAAACGAGGTTGTTTG
TAGATAGCTAGGTTGGGAAGTGAAATGATCAGCTTT
TTAAAAAAAAAAGAAAAAAAAAAGAAAACTAGGCTAC
TTCTTTTGTTCCCTAACCTGCCGGACTCCTTCCCAC
TAATTTTTATGTTTTTGTAGAGATGGGGTTTCTCCG
TCTTTAAAGATCTTTCCGGACACTTTTGGAAAAGAG
GATGCATTGCTATGCCTCCCAGTCCGCAACTTCACG
TACCCCTTGACTCTTCTTTTGTTACCATTTTTCCCC
TTTCAAATTATTATTTATTGCTCATTGTGTTTTTTG
TATTCAAAAACAATTTGTTTAAATTTAAAAATGAAC
TCATATGAAGCAAATTGTTTTGATCAACTCTCATAT
TATGCAAGGAAACAGTTCGCGATGCTCCCGTTTGCG
TTTCAATATATGCAGTCTGGTTCCAGAGTTTTTAAT
TCTCAATTTGCTATTGTAGTTATTGTTTTACTGTTG
GGTGAAGAAACAAAGGCCTGCAAAGTTCCTTCCTAC
TTAAGGTACTCAGCACTTTCTACGGCATTACGCGGG
GTTAAGTTTGGCCTCTTGCCTGGCATCACTTGCCTT
GGACAATTGCAATGCTCACAATTCGGAAACTTCCGC
TGGTTGGTACATTTCACATAAATGGAATCACATAAT
TAATGTTAAACTGTTAATAATGCTTGCTCCCAGGAA
GAACCCAGAAATCACACCTCAGTTTATCCTGGGCCT
GGAACCGTCTTCGACTGTGCCGCCTGACGCAAAGGC
TGGACAAAGAAGGTGTCTGGGCAATAGAAACAGTGT
TCTTCTTGTAAATTTGTTTTAAGTTTTTTATATATG
ACCCACACAACTGAACCCACATCACATGACAAGACT
TGTTTGTTGAACTCCCGTCATATTGGCTCCCTTGCT
TATCTCTTCGTAGCCCCTCTGTGTATGTTCTTCCTC
GTTGTGATTGCTCATTAAGACTCTGAACAATACTCA
TTCTACGTGTGGCCTTCAGTACTTTTCTTGGGCCTT
TCGACGCCGTTTCCCTTCGGGTCCACACGGTGTTTG
GAATTGAATCAATTCGGAGACTGTGCGATCGGCCGC
TAAGTGTCTATCACGGCCAAGACGCAGGCTGGGTGC
TTCTGTTTAAATGCTTGTTCGATGGCTTGTTAGAAG
GGCGGGGCGGGGGAGACGCCGGGCCCAGCCCGCCCC
12 M
Map to reference genome
Mapped to a unique location
7.2 M
Quality filter
3 M
Mapped to multiple locations
1.8 M
No mapping
2.5 M
Low quality
0.5 M

Align/Assemble to a reference

* Bowtie - Ultrafast, memory-efficient short read aligner. It aligns short DNA sequences (reads) to the human genome at a rate of 25 million reads per hou workstation with 2 gigabytes of memory. Link to discussion thread here. Written by Ben Langmead and Cole Trapnell.
* ELAND - Efficient Large-Scale Alignment of Nucleotide Databases. Whole genome alignments to a reference genome. Written by Illumina author Anthony Solexa 1G machine.
* EULER - Short read assembly. By Mark J. Chaisson and Pavel A. Pevzner from UCSD (published in Genome Research).
* Exonerate - Various forms of alignment (including Smith-Waterman-Gotoh) of DNA/protein against a reference. Authors are Guy St C Slater and Ewan Bii EMBL. C for POSIX.
* GMAP - GMAP (Genomic Mapping and Alignment Program) for mRNA and EST Sequences. Developed by Thomas Wu and Colin Watanabe at Genentec. C.
* MOSAIK - Reference guided aligner/assembler. Written by Michael Strömberg at Boston College.
* MAQ - Mapping and Assembly with Qualities (renamed from MAPASS2). Particularly designed for Illumina-Solexa 1G Genetic Analyzer, and has preliminar handle ABI SOLiD data. Written by Heng Li from the Sanger Centre.
* MUMmer - MUMmer is a modular system for the rapid whole genome alignment of finished or draft sequence. Released as a package providing an efficien library, seed-and-extend alignment, SNP detection, repeat detection, and visualization tools. Version 3.0 was developed by Stefan Kurtz, Adam Phillippy, A Michael Smoot, Martin Shumway, Corina Antonescu and Steven L Salzberg - most of whom are at The Institute for Genomic Research in Maryland, USA. P0 required.
* Novocraft - Tools for reference alignment of paired-end and single-end Illumina reads. Uses a Needleman-Wunsch algorithm. Available free for evaluatio use and for use on open not-for-profit projects. Requires Linux or Mac OS X.
* RMAP - Assembles 20 - 64 bp Solexa reads to a FASTA reference genome. By Andrew D. Smith and Zhenyu Xuan at CSHL. (published in BMC Bioinforma OS required.
* SeqMap - Works like ELand, can do 3 or more bp mismatches and also INDELs. Written by Hui Jiang from the Wong lab at Stanford. Builds available for m
* SHRiMP - Assembles to a reference sequence. Developed with Applied Biosystem's colourspace genomic representation in mind. Authors are Michael Bru Stephen Rumble at the University of Toronto.
* Slider- An application for the Illumina Sequence Analyzer output that uses the probability files instead of the sequence files as an input for alignment to a sequence or a set of reference sequences.. Authors are from BCGSC. Paper is here.
* SOAP - SOAP (Short Oligonucleotide Alignment Program). A program for efficient gapped and ungapped alignment of short oligonucleotides onto referenc Author is Ruiqiang Li at the Beijing Genomics Institute. C++ for Unix.
* SSAHA - SSAHA (Sequence Search and Alignment by Hashing Algorithm) is a tool for rapidly finding near exact matches in DNA or protein databases usi Developed at the Sanger Centre by Zemin Ning, Anthony Cox and James Mullikin. C++ for Linux/Alpha.
* SXOligoSearch - SXOligoSearch is a commercial platform offered by the Malaysian based Synamatix. Will align Illumina reads against a range of Refseq genome builds for a number of organisms. Web Portal. OS independent.

***de novo* Align/Assemble**

* MIRA2 - MIRA (Mimicking Intelligent Read Assembly) is able to perform true hybrid de-novo assemblies using reads gathered through 454 sequencing te or GS FLX). Compatible with 454, Solexa and Sanger data. Linux OS required.
* SHARCGS - De novo assembly of short reads. Authors are Dohm JC, Lottaz C, Borodina T and Himmelbauer H. from the Max-Planck-Institute for Molecul
* SSAKE - Version 2.0 of SSAKE (23 Oct 2007) can now handle error-rich sequences. Authors are René Warren, Granger Sutton, Steven Jones and Robert Canada's Michael Smith Genome Sciences Centre. Perl/Linux.
* VCAKE - De novo assembly of short reads with robust error correction. An improvement on early versions of SSAKE.
* Velvet - Velvet is a de novo genomic assembler specially designed for short read sequencing technologies, such as Solexa or 454. Need about 20-25X co paired reads. Developed by Daniel Zerbino and Ewan Birney at the European Bioinformatics Institute (EMBL-EBI).

SNP/Indel Discovery

Bioinformatics Challenges

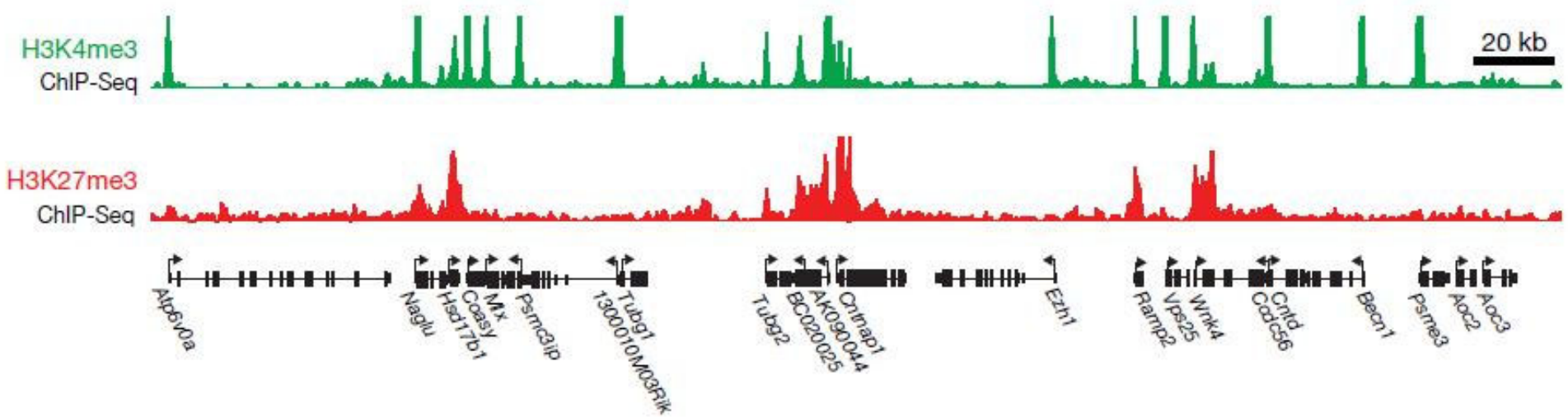
- Rapid mapping of these short sequence reads to the reference genome

- Visualize mapping results
  - Thousand of enriched regions

- Peak analysis
  - Peak detection
  - Finding exact binding sites

- Compare results of different experiments
  - Normalization
  - Statistical tests

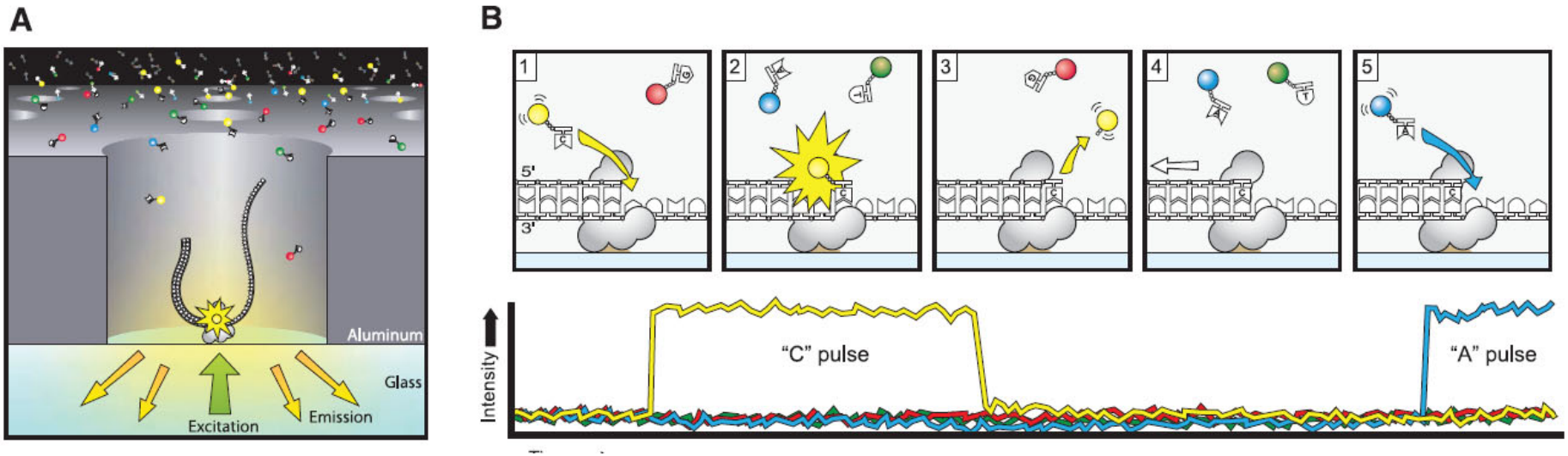
Visualization: Custom



<u>300 kb region</u> from mouse ES cells

Mikkelsen,T.S. *et al*. *Nature* **448**, 553-562 (2007)

Pacific Biosciences: A Third Generation Sequencing Technology



Eid et al 2008

Applications

- Genomes
- Re-sequencing Human Exons (Microarray capture/amplification)
- small (including mi-RNA) and long RNA profiling (including splicing)
- ChIP-Seq:
    - Transcription Factors
    - Histone Modifications
    - Effector Proteins
- DNA Methylation
- Polysomal RNA
- Origins of Replication/Replicating DNA
- Whole Genome Association (rare, high impact SNPs)
- Copy Number/Structural Variation in DNA
- ChIA-PET: Transcription Factor Looping Interactions
- ???

references

- *Epigenetics meets next-generation sequencing*. Park PJ. Epigenetics. 2008 Nov;3(6):318-21.
- *Next-generation DNA sequencing*. Jay Shendure, Hanlee Ji. Nature Biotechnology 26, 1135 - 1145 (2008)