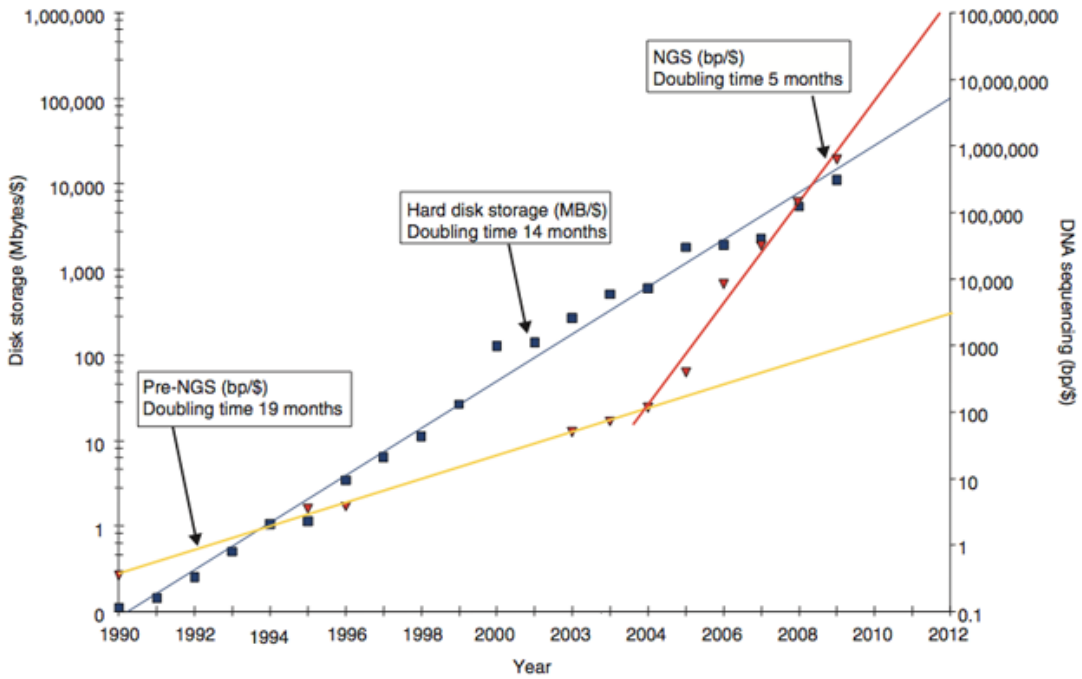
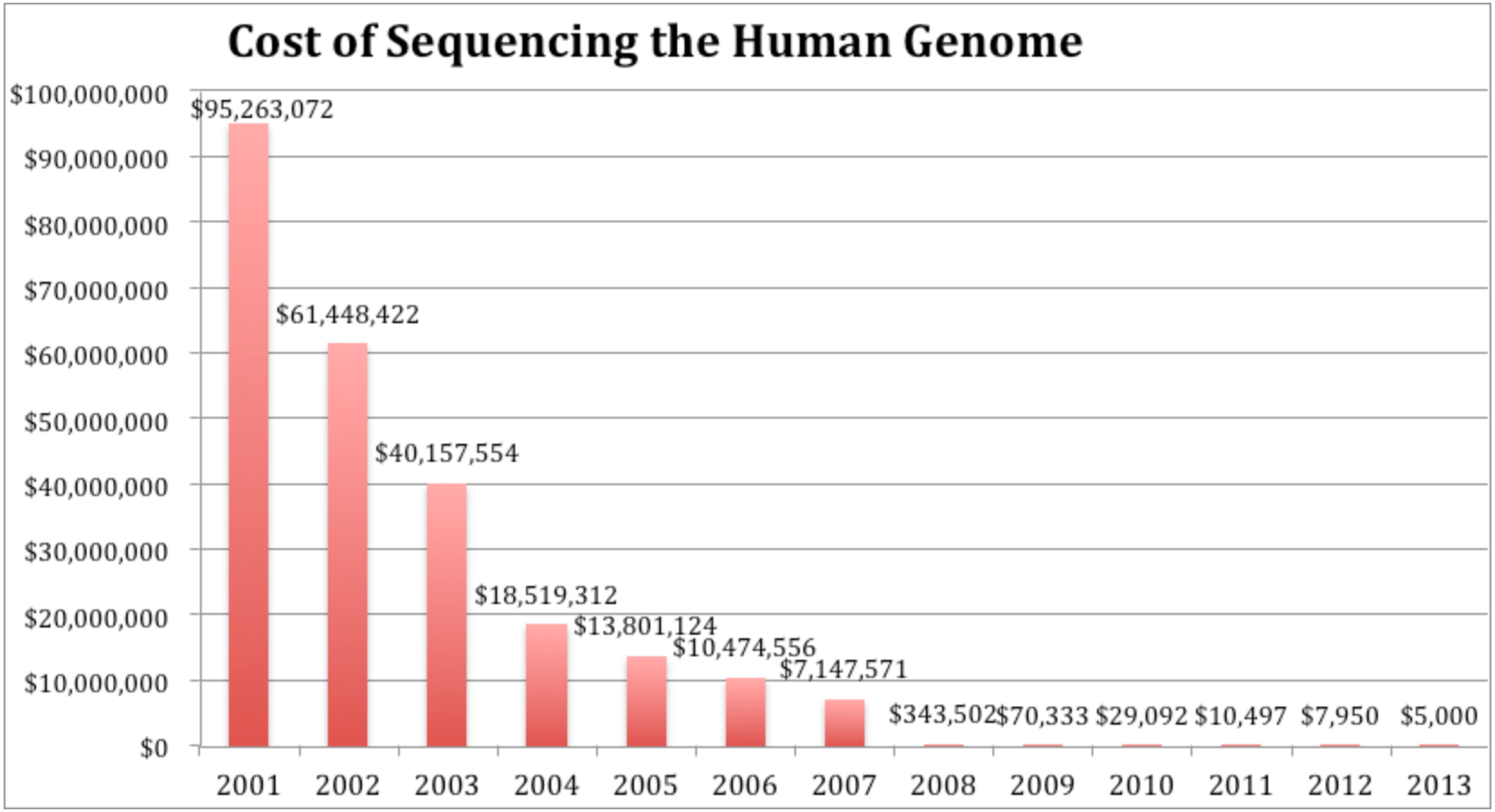
Computational Genomics

Next generation sequencing (NGS)

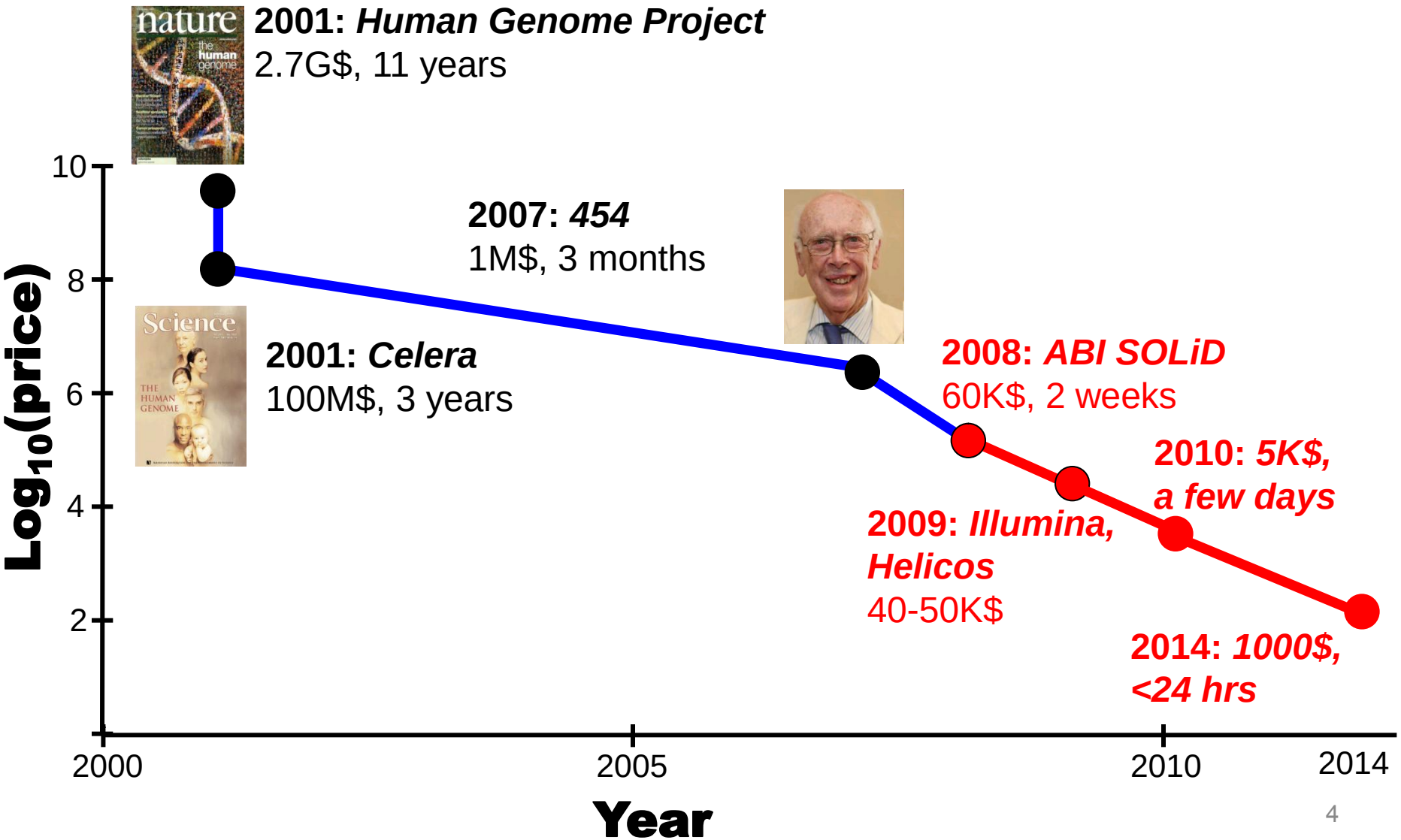
Sequencing technology defies Moore's law



Nature Methods 2011


Cost of Sequencing the Human Genome
$100,000,000
$90,000,000
$80,000,000
$70,000,000
$60,000,000
$50,000,000
$40,000,000
$30,000,000
$20,000,000
$10,000,000
$0
$95,263,072
$61,448,422
$40,157,554
$18,519,312
$13,801,124
$10,474,556
$7,147,571
$343,502
$70,333
$29,092
$10,497
$7,950
$5,000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013

Sequencing the Human Genome



BIOMEDICINE NEWS

2 COMMENTS

Does Illumina Have the First $1,000 Genome?

Illumina announces a new high-end sequencer made for "factory-scale" sequencing of human genomes.

By Susan Young on January 14, 2014

The $1,000 genome has been a catchphrase of the sequencing industry for years, but despite bold promises from different companies, this benchmark hasn't been met. Now, thanks to a new sequencing machine from Illumina, it may finally be within reach.

At the J.P. Morgan Healthcare Conference on Tuesday, Illumina CEO Jay Flatley announced a new

WHY IT MATTERS

Genome sequencing is still too costly for many medical applications, or for large-scale sequencing projects that could uncover the genetic basis of disease.

Company Unveils DNA Sequencing Device Meant to Be Portable, Disposable and Cheap

By ANDREW POLLACK
Published: February 17, 2012

DNA sequencing is becoming both faster and cheaper. Now, it is also becoming tinier.

A British company said on Friday that by the end of the year it would begin selling a disposable gene sequencing device that is the size of a USB memory stick and plugs into a laptop computer to deliver its

Next Gen-Omics™

Genome Resequencing

mRNA Tag Profiling

Methylation Analysis

Small RNA Identification

Functional Elements (ChIP-Seq, DNAse-Seq)

Transcriptome Sequencing

Applications of next-generation sequencing

Category	Examples of applications
Complete genome resequencing	Comprehensive polymorphism and mutation discovery in individual human genomes
Reduced representation sequencing	Large-scale polymorphism discovery
Targeted genomic resequencing	Targeted polymorphism and mutation discovery
Paired end sequencing	Discovery of inherited and acquired structural variation
Metagenomic sequencing	Discovery of infectious and commensal flora
Transcriptome sequencing	Quantification of gene expression and alternative splicing; transcript annotation; discovery of transcribed SNPs or somatic mutations
Small RNA sequencing	microRNA profiling
Sequencing of bisulfite-treated DNA	Determining patterns of cytosine methylation in genomic DNA
Chromatin immunoprecipitation–sequencing (ChIP-Seq)	Genome-wide mapping of protein-DNA interactions
Nuclease fragmentation and sequencing	Nucleosome positioning
Molecular barcoding	Multiplex sequencing of samples from multiple individuals

Nature Biotechnology 26 (10): 1135-1145 (2008)

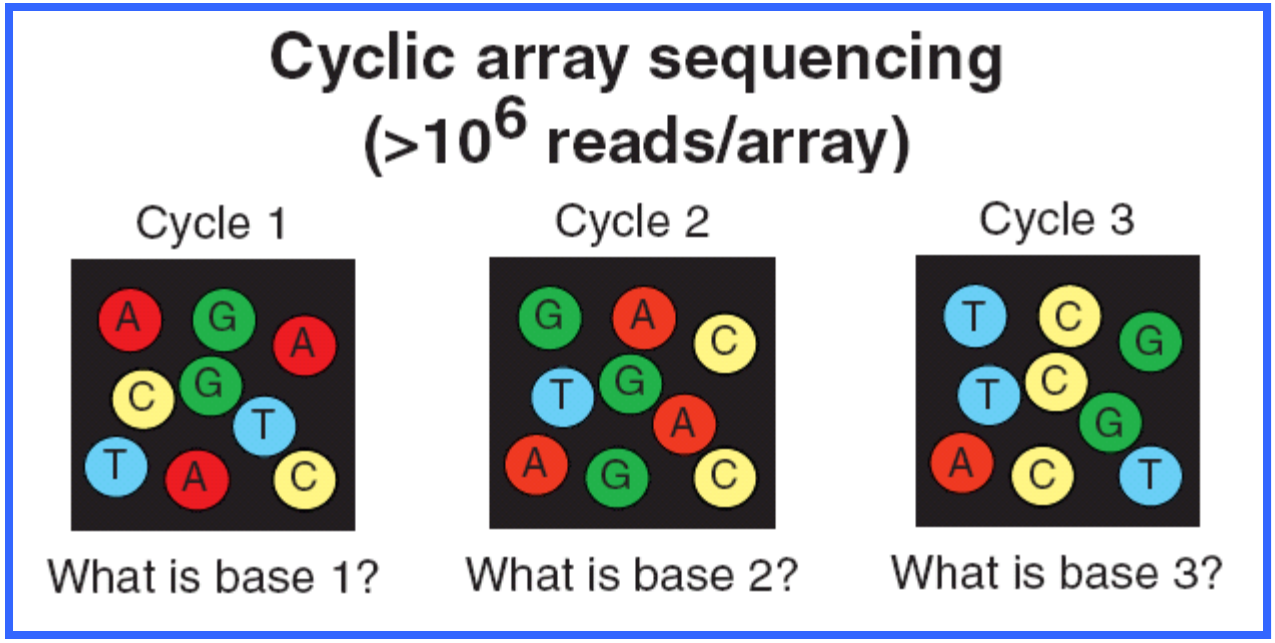
Roche & illumina analyzers

Genome Sequencer FLX

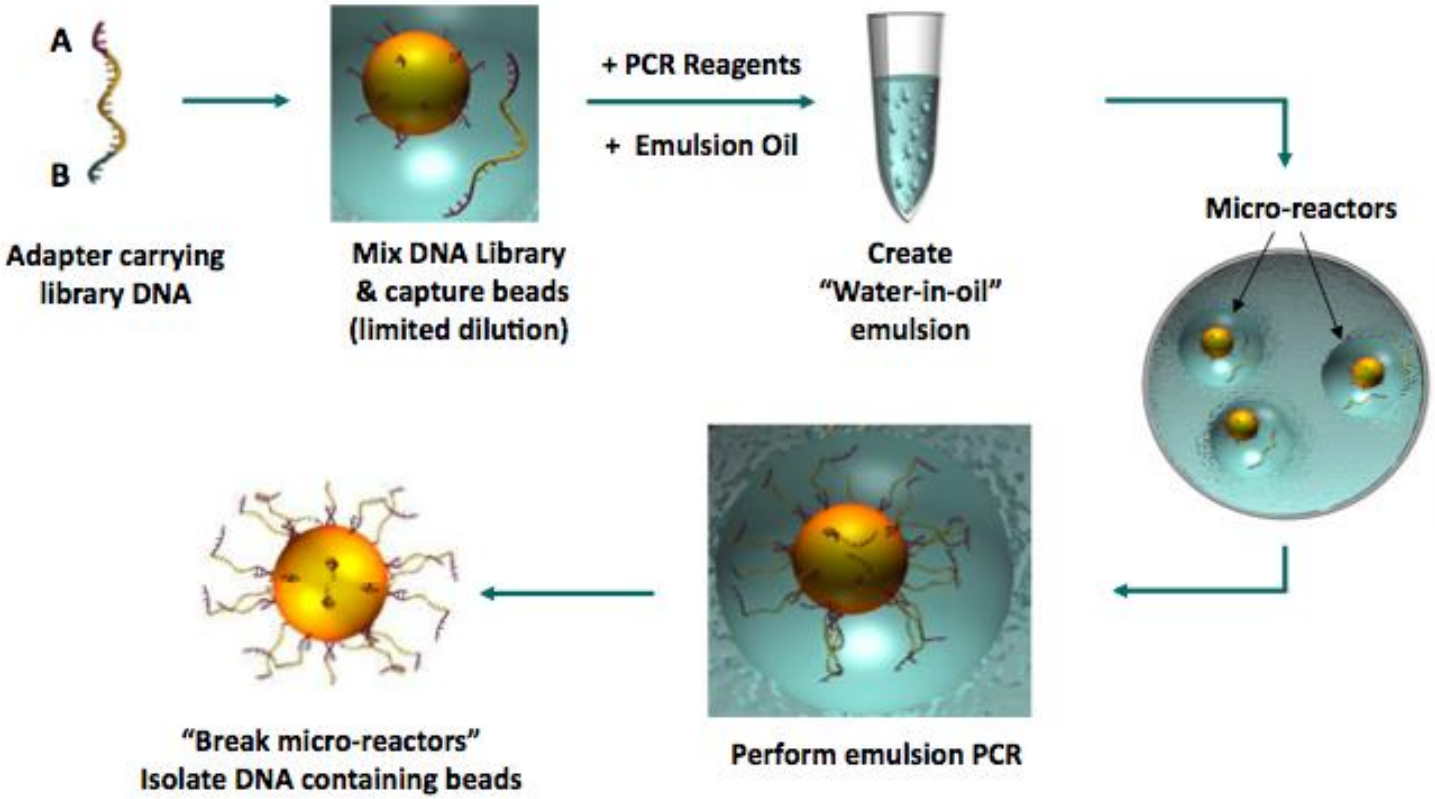
Illumina Genome Analyzer

High Parallelism is Achieved in Polony Sequencing

Sanger **Polony**

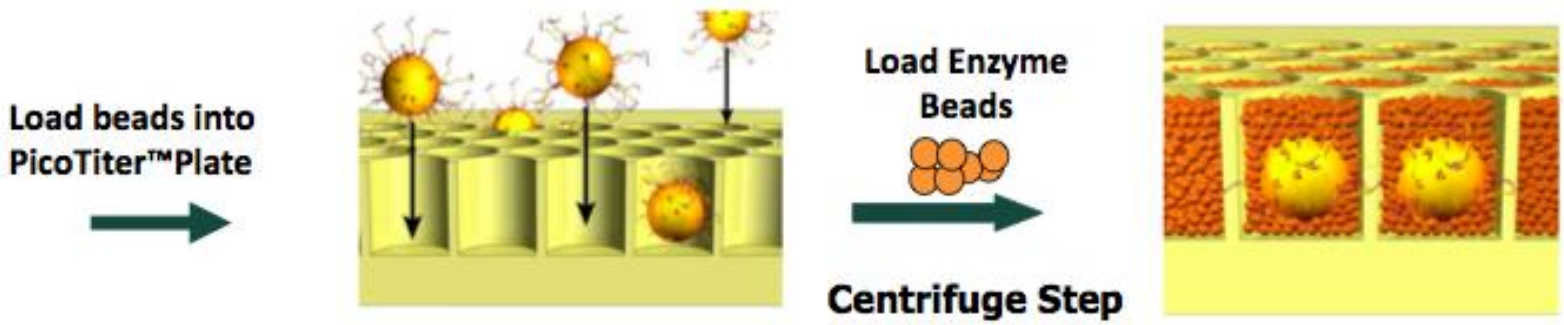


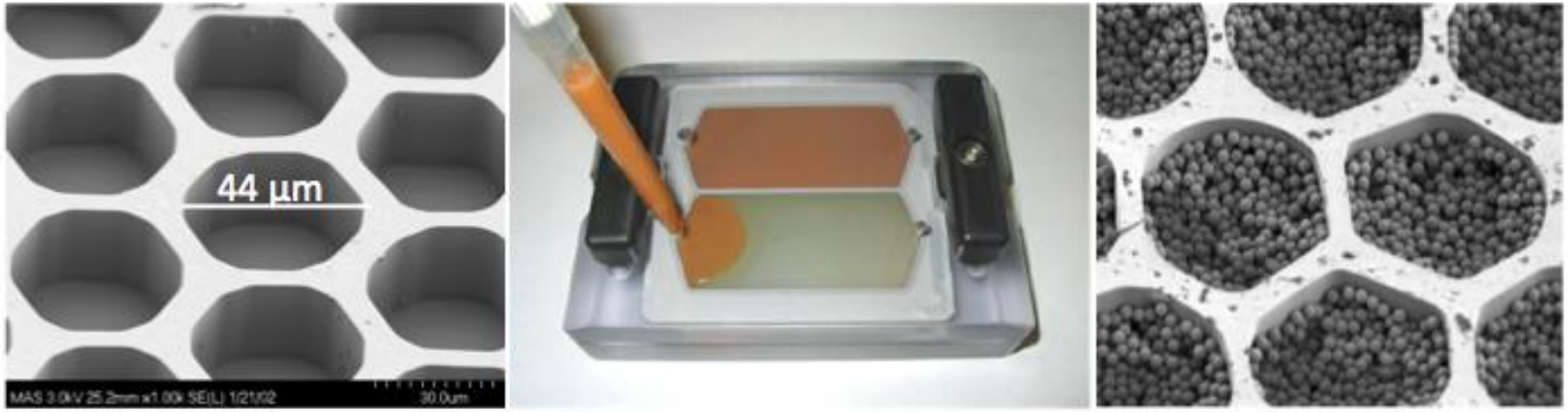
Generation of Polony array: DNA Beads (454, SOLiD)



DNA Beads are generated using Emulsion PCR

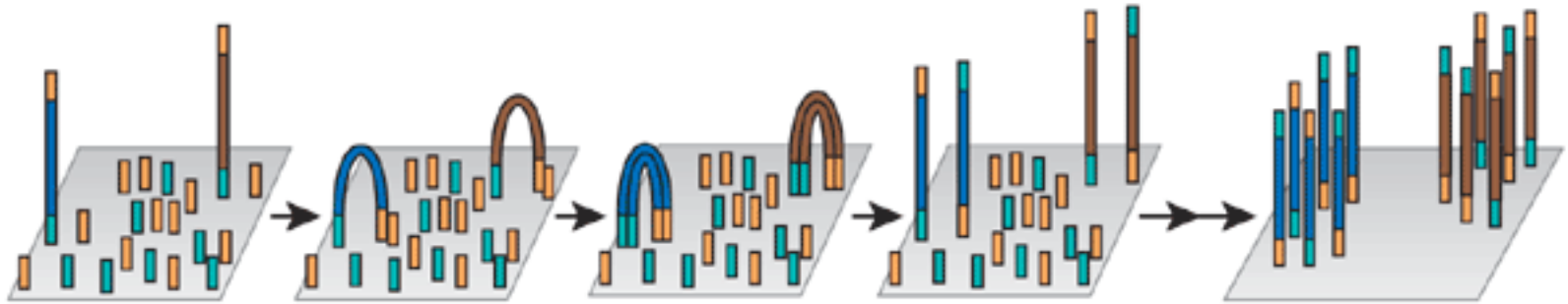
Generation of Polony array: DNA Beads (454, SOLiD)





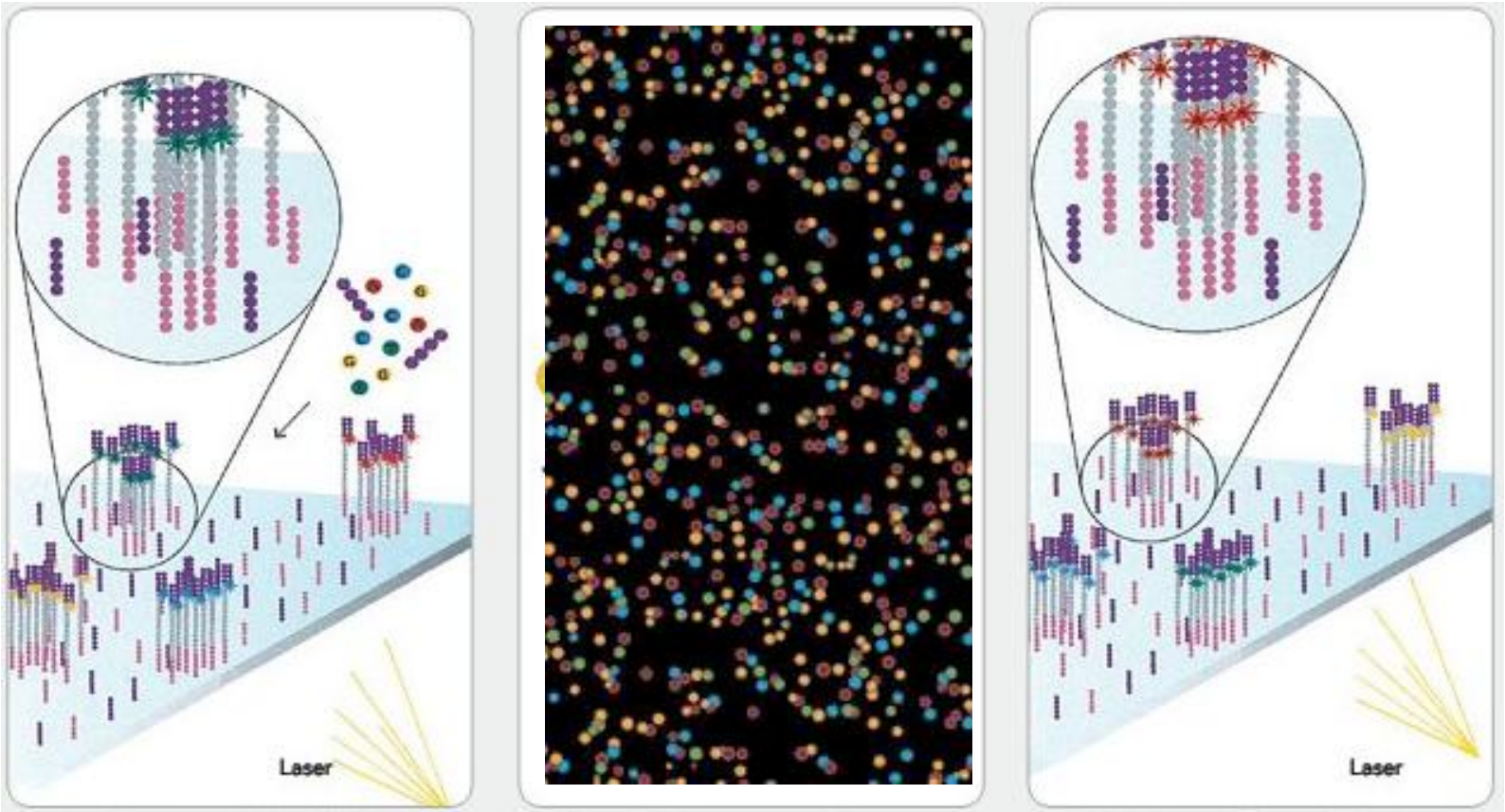
DNA Beads are placed in wells

Generation of Polony array: Bridge-PCR (Solexa)



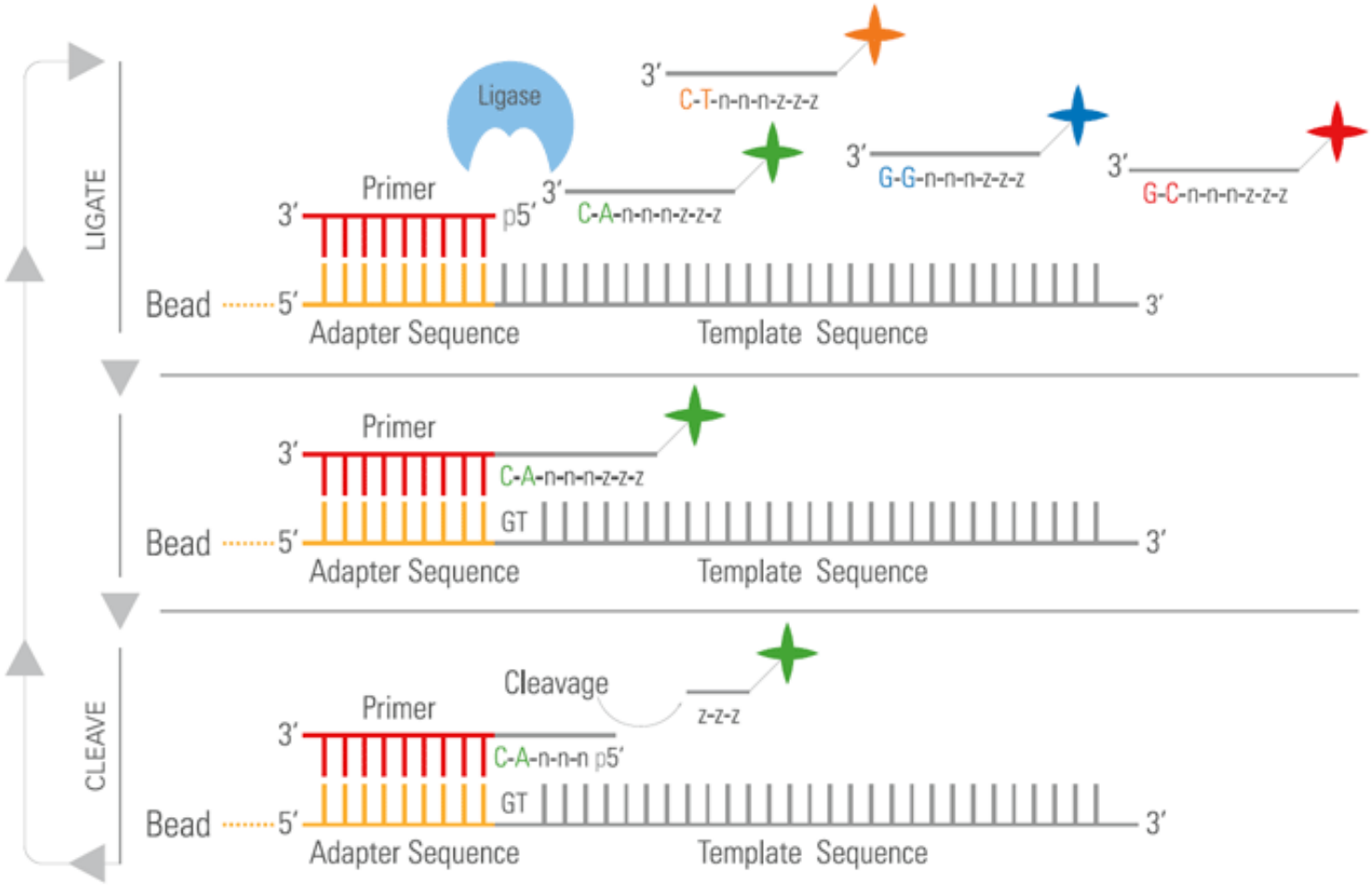
DNA fragments are attached to array and used as PCR templates

Sequencing: Fluorescently labeled Nucleotides (Solexa)



Complementary strand elongation: DNA Polymerase

Sequencing: Fluorescently Labeled Nucleotides (ABI SOLiD)



Complementary strand elongation: DNA Ligase

Single Molecule Sequencing: HeliScope

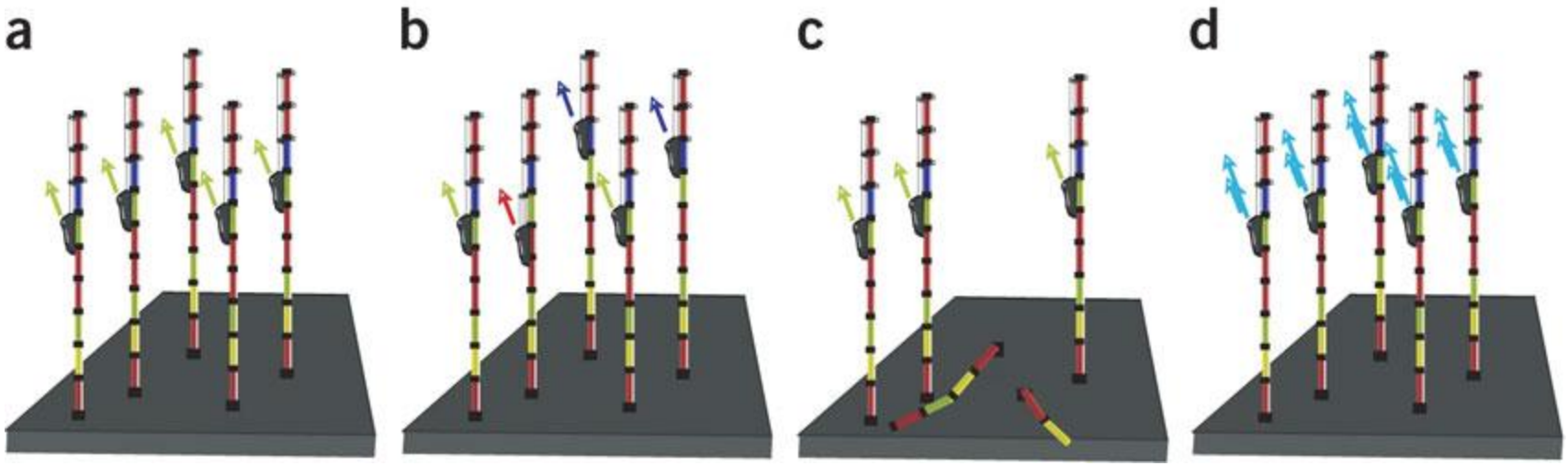
- Direct sequencing of DNA molecules: no amplification stage
- DNA fragments are attached to array
- Potential benefits: higher throughput, less errors

Technology Summary

	Read length	Sequencing Technology	Throughput (per run)	Cost (1mbp)*
Sanger	~800bp	Sanger	400kbp	500$
454	~400bp	Polony	500Mbp	60$
Solexa	75bp	Polony	20Gbp	2$
SOLiD	75bp	Polony	60Gbp	2$
Helicos	30-35bp	Single molecule	25Gbp	1$

*Source: Shendure & Ji, *Nat Biotech*, 2008

Errors



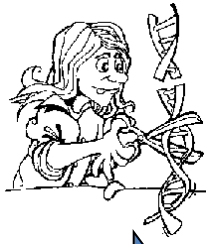
Erlich et al. Nature Methods 5: 679-682 (2008)

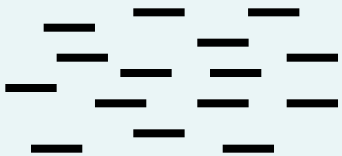
Genome Sequencing

- **Goal**
  - figuring the order of nucleotides across a genome
- **Problem**
  - Current DNA sequencing methods can handle only short stretches of DNA at once (usually between 100-200 bp's)
- **Solution**
  - Sequence and then use computers to assemble the small pieces

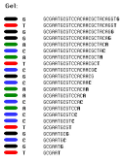
Genome Sequencing

Genome





Short fragments of DNA



AC..GC TG..GT TC..CC
CG..CA
TT..TC
TG..AC
AC..GC GA..GC
CT..TG
GT..GC AC..GC AC..GC
AA..GC AT..AT
TT..CC

Short DNA sequences

ACGTGGTAA CGTATACAC TAGGCCATA
GTAATGGCG CACCCTTAG
TGGCGTATA CATA...

ACGTGGTAATGGCGTATACACCCTTAGGCCATA

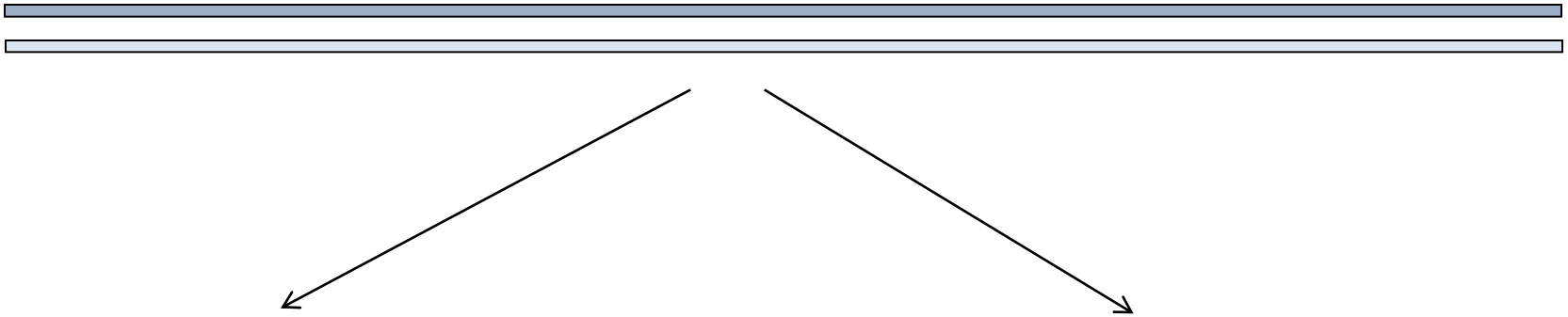
ACGTGACCGGTACTGGTAACGTACA
CCTACGTGACCGGTACTGGTAACGT
ACGCCTACGTGACCGGTACTGGTAA
CGTATACACGTGACCGGTACTGGTA
ACGTACACCTACGTGACCGGTACTG
GTAACGTACGCCTACGTGACCGGTA
CTGGTAACGTATACCTCT...

Sequenced genome

Assembly

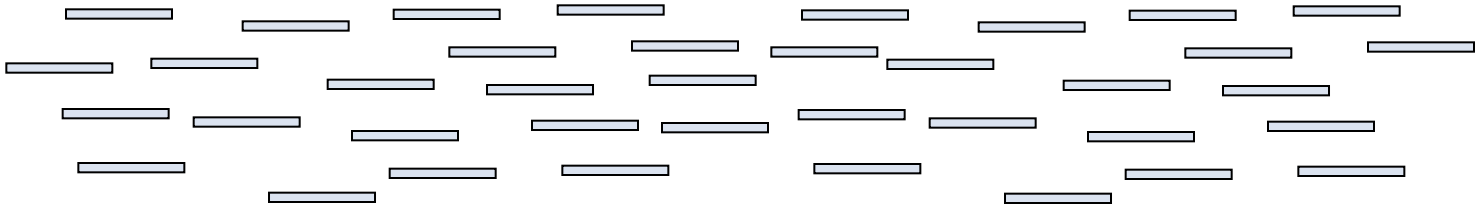
- How do we use the short reads to recover the genome being sequenced?

1: Using a reference genome: Alignment of reads to a reference



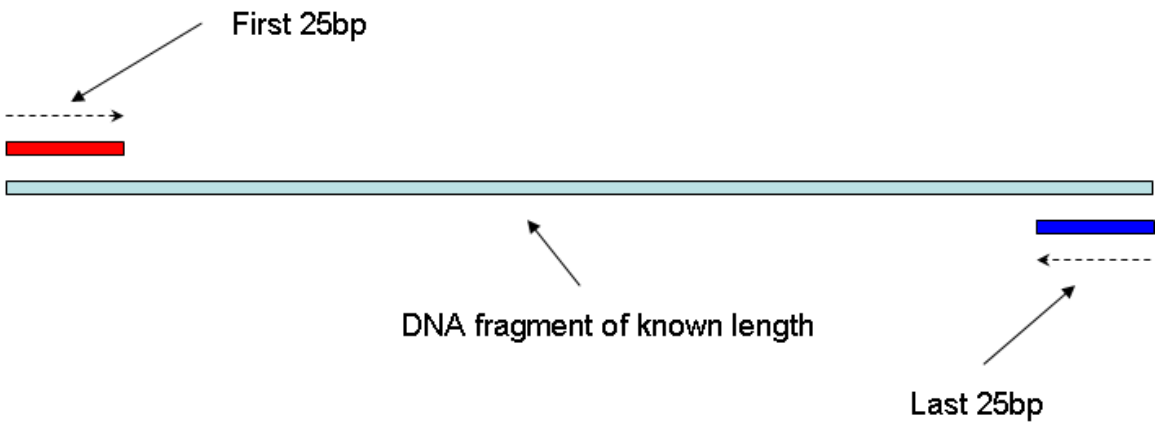
..ACTGGGTCATCGTACGATCGATCGATCGATCGATCGGCTAGCTAGCTA.. Reference

..ACTGGGTCATCGTACGATCGATAGATCGATCGATCGCTAGCTAGCTA.. Sample



2. Assembly without a reference (de novo assembly)

- Paired-End sequencing (Mate pairs)
  - Sequence two ends of a fragment of known size.



  - Currently fragment length (insert size) can range from 200 bps – 10,000 bps

Velvet

- Euler / De Bruijn approach.
- Introduced as a alternative to overlap-layout-consensus approach in capillary sequencing.
- More suited for short read assembly.
- Based on De Bruijn graph.
- Implemented in *Velvet*[1], the mostly used short read assembly method at present.

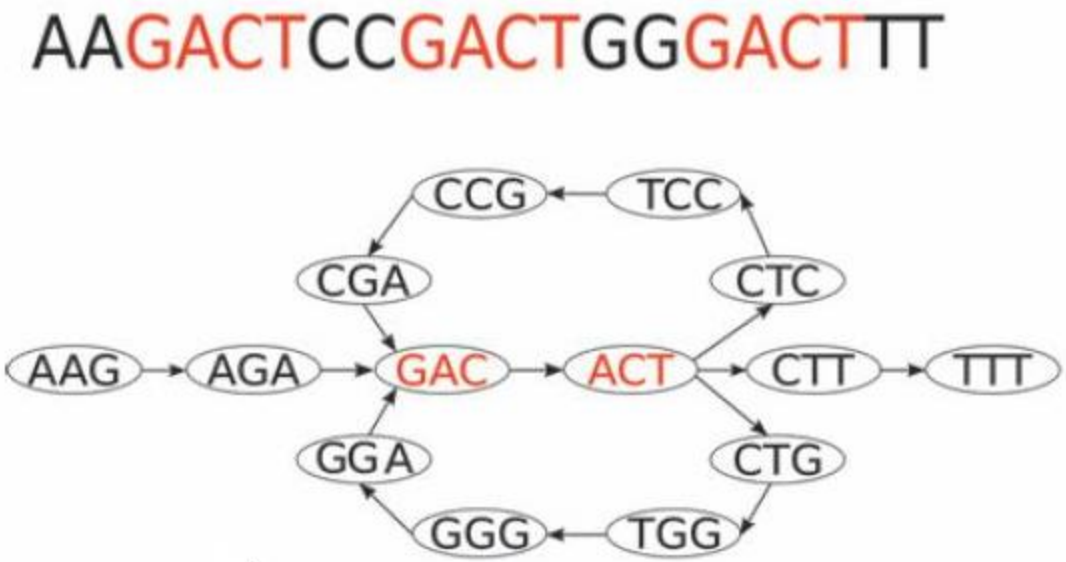
[1]Daniel Zerbino and Ewan Birney. Velvet: Algorithms for De Novo Short Read Assembly Using De Bruijn Graphs. Genome Res. 18: 821-829. 2008

De Bruijn graph method

- Break each read sequence in to overlapping fragments of size *k.* (*k*-mers)
- Form De Bruijn graph such that each (*k-1)*-mer represents a node in the graph.
- Edge exists between node *a* to *b* iff there exists a *k*-mer such that is prefix is *a* and suffix is *b.*
- Traverse the graph in unambiguous path to form contigs.

De Bruijn graph

- K = 4



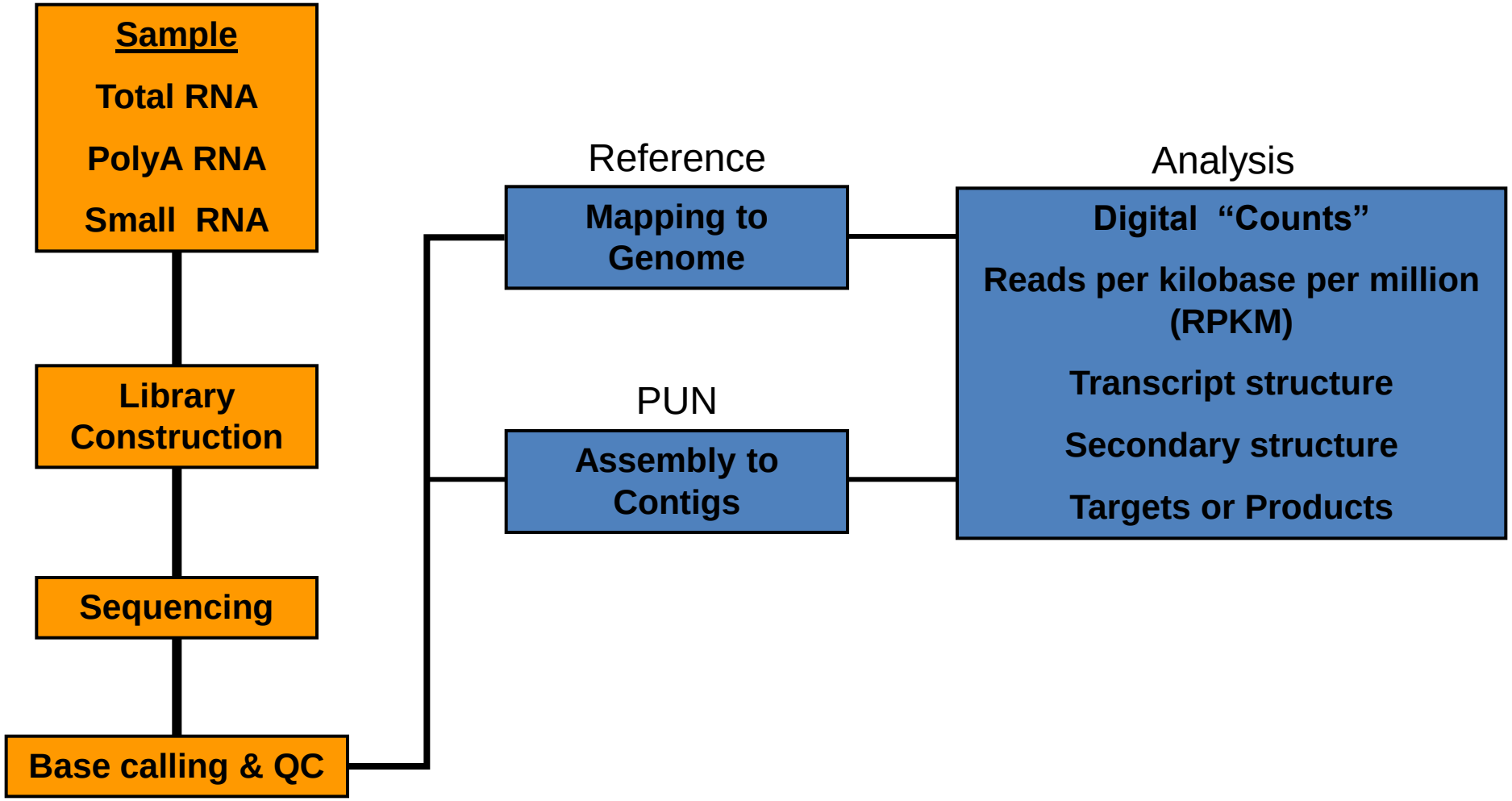
de Bruijn graph of a sequence

De Bruijn graph method / Velvet

- Elegant way of representing the problem.
- Very fast execution.
- Error correction can be handled in the graph.
- De Bruijn graph size can be huge.
  - ~200GB for human genomes.
- Does not use pair information in initial phase, resulting in overly complicated graphs

Applications (beyond DNA)

RNASeq


Sample
Total RNA
PolyA RNA
Small RNA
Library Construction
Sequencing
Base calling & QC
Reference
Mapping to Genome
PUN
Assembly to Contigs
Analysis
Digital "Counts"
Reads per kilobase per million (RPKM)
Transcript structure
Secondary structure
Targets or Products

RNASeq – Compositional properties

Depth of Sequence

- Sequence count ≈ Transcript Abundance
- Majority of the data can be dominated by a small number of highly abundant transcripts
- Ability to observe transcripts of smaller abundance is dependent upon sequence depth

Normalization

- Normalization is the process in which components of experiments are made comparable before statistical analysis.
- It is important in sequencing as well as for microarrays.
- A couple issues in normalization are different sequencing depth (library size) and distributions of reads (long right tails).

Summarized Counts

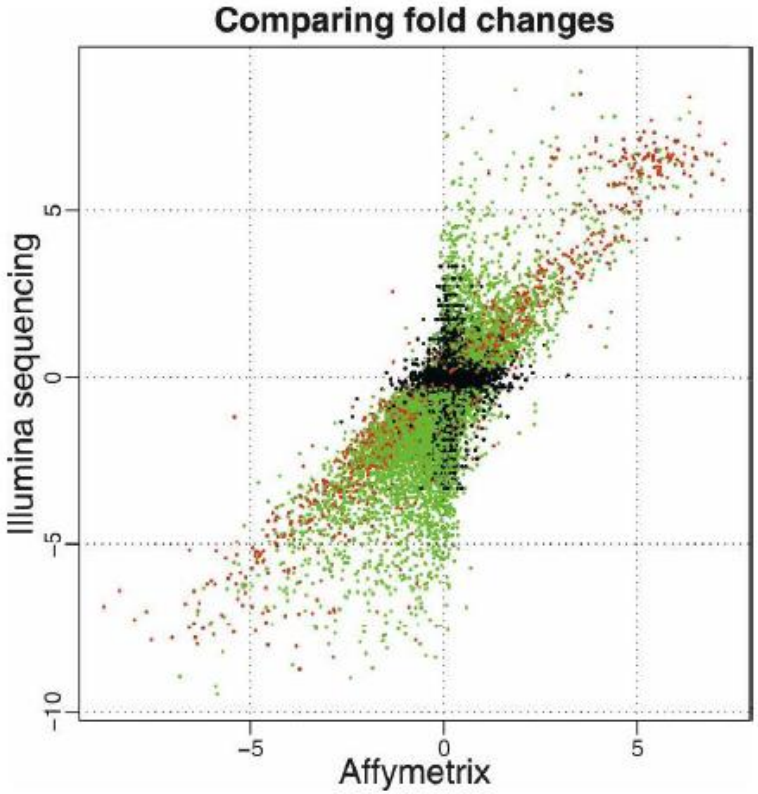
Gene	ORF size	Sample1	Sample2	Sample3	Sample4	Sample5	Sample6
NDUFS2	1861	1286	1700	1404	1485	1409	1161
NDUFS3	883	446	788	530	894	498	642
NDUFS6	417	521	745	722	843	651	609
NFKBIE	1549	59	68	79	52	74	30
NOTCH4	6157	4	2	2	2	5	0
NRTN	971	16	39	22	29	15	21
YBX1	1152	606	626	581	377	578	285
ROR2	3037	0	0	0	0	0	0
OVOL1	1126	0	0	0	0	0	0
PARK2	1541	2	0	0	0	1	0
PCK2	2061	102	171	97	108	89	87
PET112L	1721	184	316	197	280	215	265
PEX14	1161	194	211	259	179	189	115
PFKFB3	1953	155	84	280	109	240	51
PFKFB4	1433	151	79	297	106	213	68
PIGH	670	128	158	159	237	164	92
PIK3C2G	4463	0	0	1	0	0	0
PKNOX1	1517	87	141	109	219	113	156
PKP2	2767	154	150	114	128	122	95
PLCB2	3874	1	3	0	1	1	2
SEPT4	1686	1	6	1	6	0	8
POU4F2	1484	0	0	1	0	0	0
PSPH	1431	323	329	338	438	357	380
RAB4A	871	325	241	332	346	364	258
MAP4K2	2561	218	319	256	298	228	181

Simple RPKM Normalization

- Proportion of reads: number of reads (n) mapping to an exon (gene) divided by the total number of reads (N), n/N.

- RPKM: Reads Per Kilobase of exon (gene) per Million mapped sequence reads, $10^9 n/(NL)$, where L is the length of the transcriptional unit in bp (Mortazavi et al., *Nat. Meth.*, 2008).

RNASeq - Correspondence



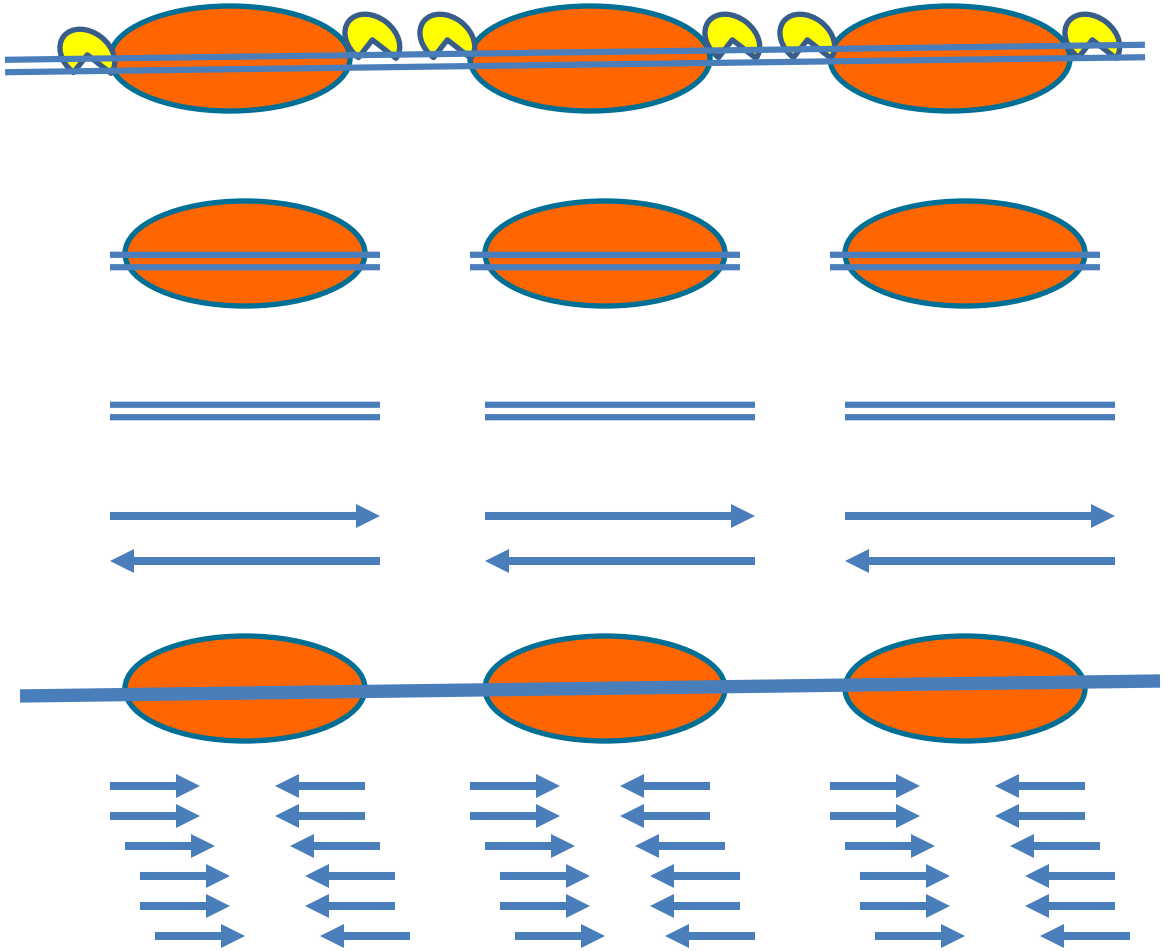
- Good correspondence with :
  - **Expression Arrays**
  - **Tiling Arrays**
  - **qRT-PCR**
- Range of up to 5 orders of magnitude
- Better detection of low abundance transcripts
- Greater power to detect
  - **Transcript sequence polymorphism**
  - **Novel trans-splicing**
  - **Paralogous genes**
  - **Individual cell type expression**

Table 3 Bioinformatics tools for short-read sequencing

Program	Categories	Author(s)	Reference	URL
Cross_match	Alignment	Phil Green, Brent Ewing and David Gordon		http://www.phrap.org/phredphrapconsed.html
ELAND	Alignment	Anthony J. Cox		http://www.illumina.com/
Exonerate	Alignment	Guy S. Slater and Ewan Birney	72	http://www.ebi.ac.uk/~guy/exonerate
MAQ	Alignment and variant detection	Heng Li	37	http://maq.sourceforge.net
Mosaik	Alignment	Michael Strömberg and Gabor Marth		http://bioinformatics.bc.edu/marthlab/Mosaik
RMAP	Alignment	Andrew Smith, Zhenyu Xuan and Michael Zhang	73	http://rulai.cshl.edu/rmap
SHRiMP	Alignment	Michael Brudno and Stephen Rumble		http://compbio.cs.toronto.edu/shrimp
SOAP	Alignment	Ruiqiang Li *et al.*	35	http://soap.genomics.org.cn
SSAHA2	Alignment	Zemin Ning *et al.*	36	http://www.sanger.ac.uk/Software/analysis/SSAHA2
SXOligoSearch	Alignment	Synamatix		http://synasite.mgrc.com.my:8080/sxog/NewSXOligoSearch.php
ALLPATHS	Assembly	Jonathan Butler *et al.*	38	
Edena	Assembly	David Hernandez *et al.*	74	http://www.genomic.ch/edena
Euler-SR	Assembly	Mark Chaisson and Pavel Pevzner	75	
SHARCGS	Assembly	Juliane Dohm *et al.*	76	http://sharcgs.molgen.mpg.de
SHRAP	Assembly	Andreas Sundquist *et al.*	39	
SSAKE	Assembly	René Warren *et al.*	40	http://www.bcgsc.ca/platform/bioinfo/software/ssake
VCAKE	Assembly	William Jeck	77	http://sourceforge.net/projects/vcake
Velvet	Assembly	Daniel Zerbino and Ewan Birney	41	http://www.ebi.ac.uk/%7Ezerbino/velvet
PyroBayes	Base caller	Aaron Quinlan *et al.*	34	http://bioinformatics.bc.edu/marthlab/PyroBayes
PbShort	Variant detection	Gabor Marth		http://bioinformatics.bc.edu/marthlab/PbShort
ssahaSNP	Variant detection	Zemin Ning *et al.*		http://www.sanger.ac.uk/Software/analysis/ssahaSNP

Incomplete list compiled from sources, including http://seqanswers.com/forums/showthread.php?t=43 and http://www.sanger.ac.uk/Users/lh3/seq-nt.html.

ChIPSeq

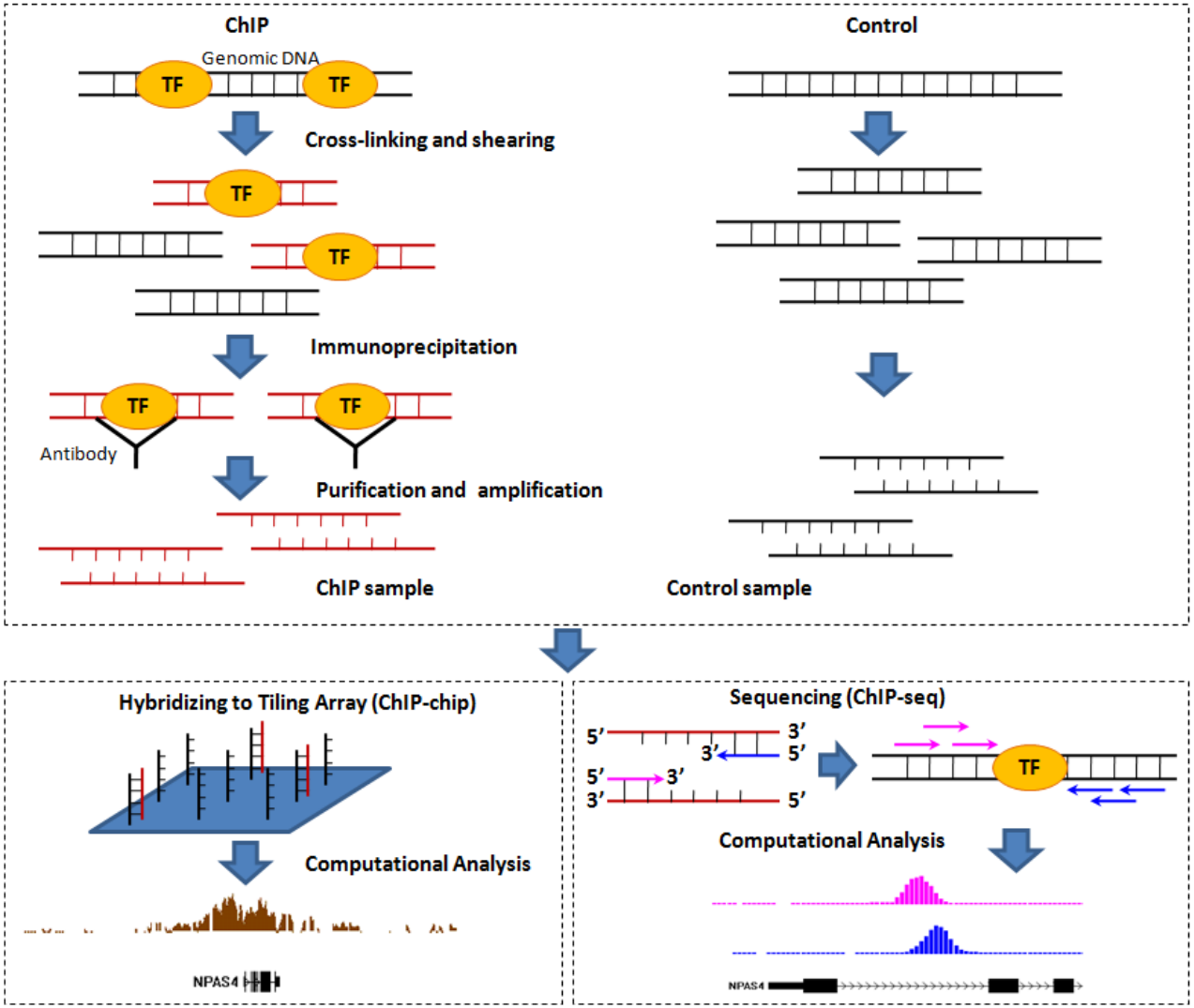


MNase
Linker Digest

Remove
Nucleosomes

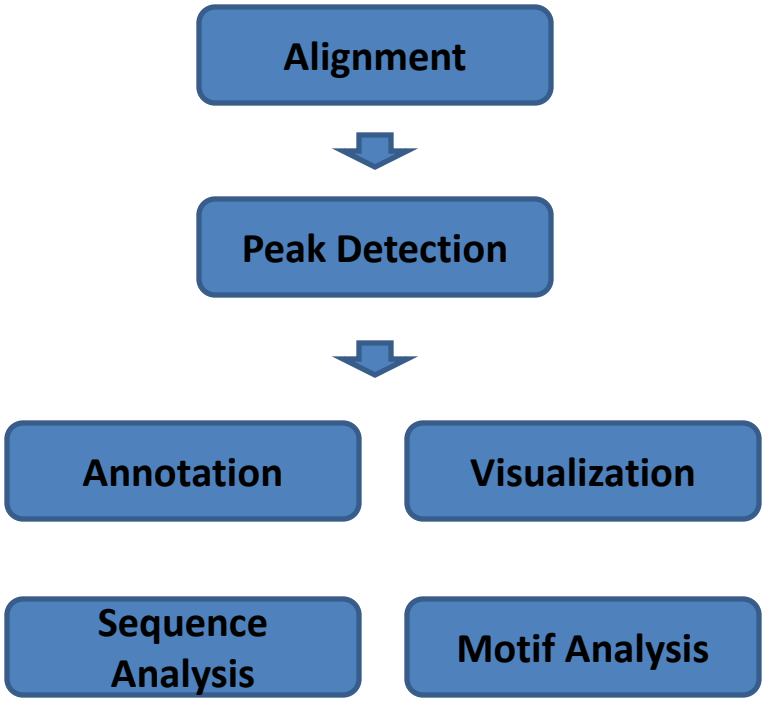
Sequence &
Align

ChIP-Seq



Hongkai Ji et al. Nature Biotechnology 26: 1293-1300. 2008

ChIP-Seq Analysis



Alignment

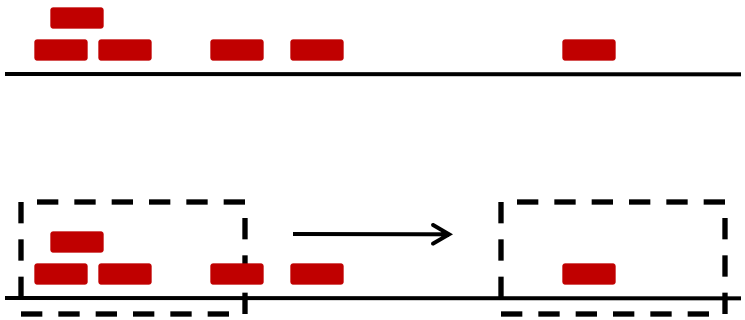
- ELAND
- Bowtie
- SOAP
- SeqMap
- …

Peak detection

- FindPeaks
- CHiPSeq
- BS-Seq
- SISSRs
- QuEST
- MACS
- CisGenome
- …

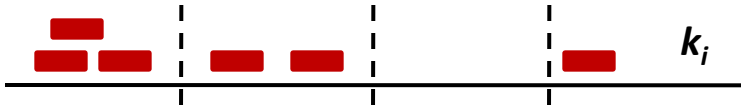
One sample analysis

A simple way is the sliding window method



<u>Poisson background model is commonly used to estimate error rate</u>

$k_i \sim \text{Poisson}(\lambda_0)$

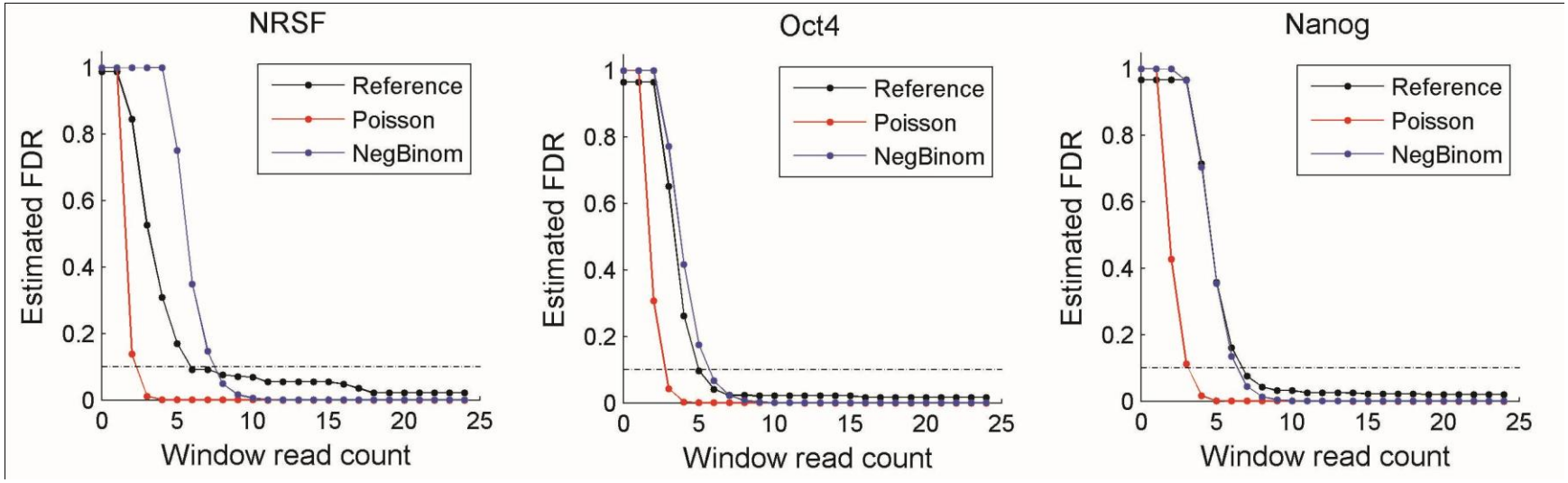


<u>Or people use Monte Carlo simulations</u>

Both are based on the assumption that read sampling rate is a constant across the genome.

Hongkai Ji et al. Nature Biotechnology 26: 1293-1300. 2008

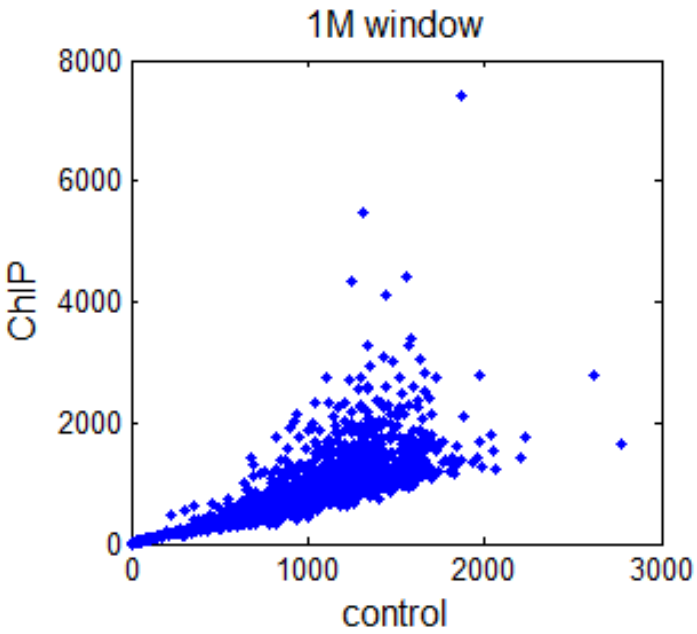
FDR estimation based on Poisson and negative binomial model

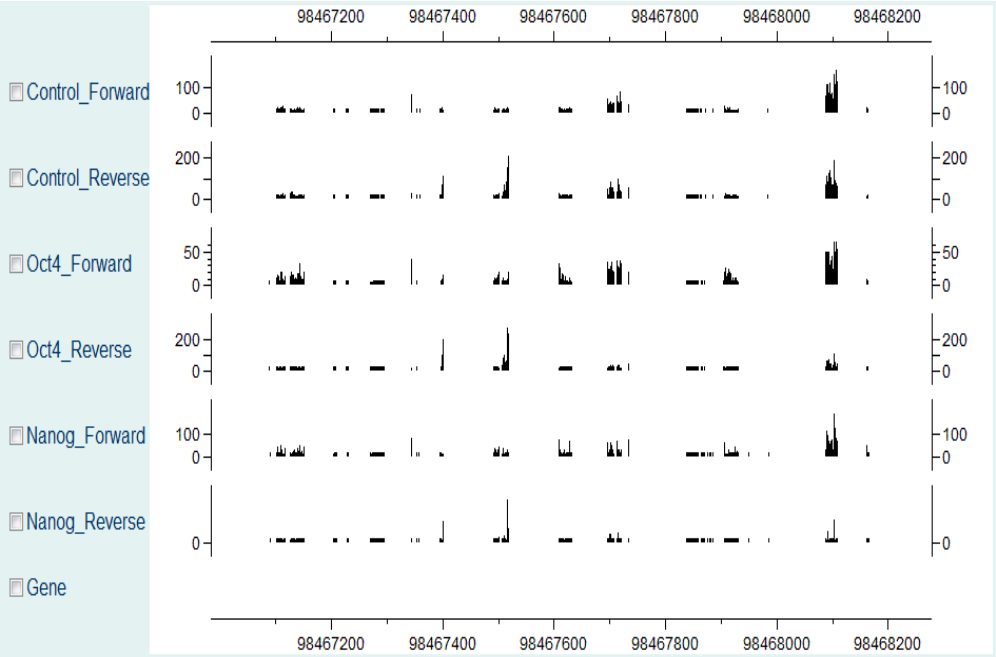


Hongkai Ji et al. Nature Biotechnology 26: 1293-1300. 2008

Two sample analysis

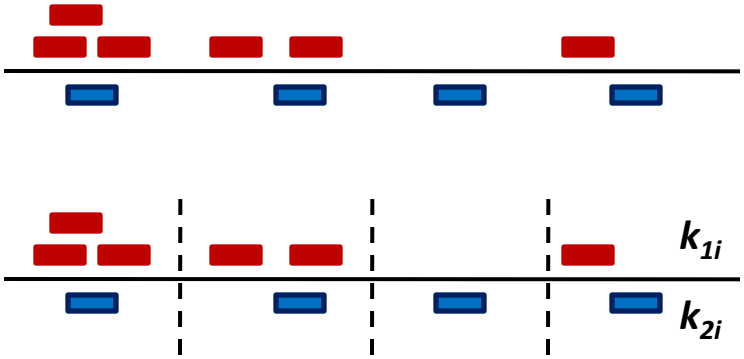
Reason: read sample rates at the same genomic locus are correlated across different samples.





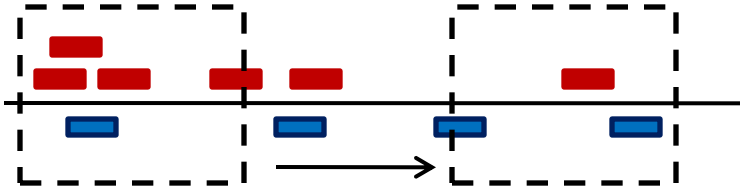
Hongkai Ji et al. Nature Biotechnology 26: 1293-1300. 2008

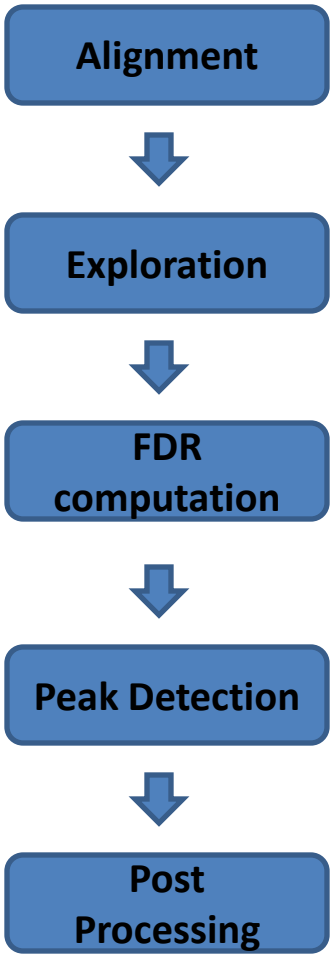
CisGenome two sample analysis



$n_i = k_{1i} + k_{2i}$

$k_{1i} \mid n_i \sim \text{Binom}(n_i, p_0)$





Hongkai Ji et al. Nature Biotechnology 26: 1293-1300. 2008

Epigenome

- Protein-DNA interactions [ChIPSeq]
  - Nucleosome positioning
  - Histone modification
  - Transcription factor interact

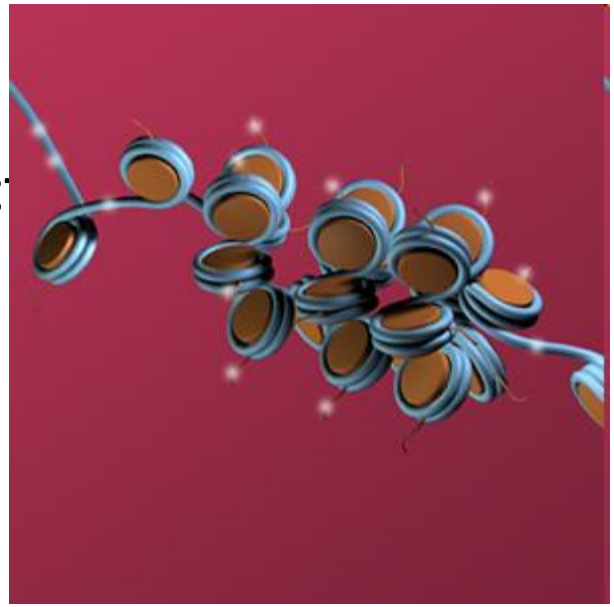


Image : ClearScience

- Methylation [MethylSeq]

- Impact of NextGen
  - Whole genome profiling
  - Resolution

Third generation sequencing

3rd generation

- Single Molecule Sequencing Technology (tSMS)
- No amplification
- 10Gb os sequence data per 8 day run
- Single Molecule Real Time (SMRT) sequencing technology (PacBio RS)

PacBio RS

Nanopore sequencing