Next generation sequencers & NGS applications

Speaker: Mei-yeh Lu 呂美曄

PhD, High Throughput Sequencing Core Manager
Academia Sinica
6/10/2010

Outlines

- Evolution of sequencing technologies
  - Sanger vs New Generation Technologies
- NGS platforms and comparisons
- Data output
- Applications
- Public database
- Challenges

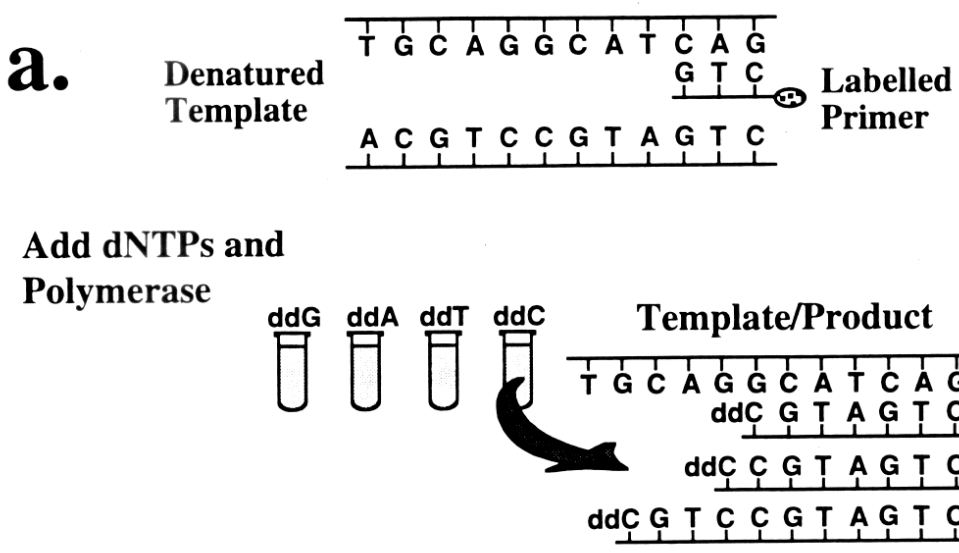
I. Convntional sequencing

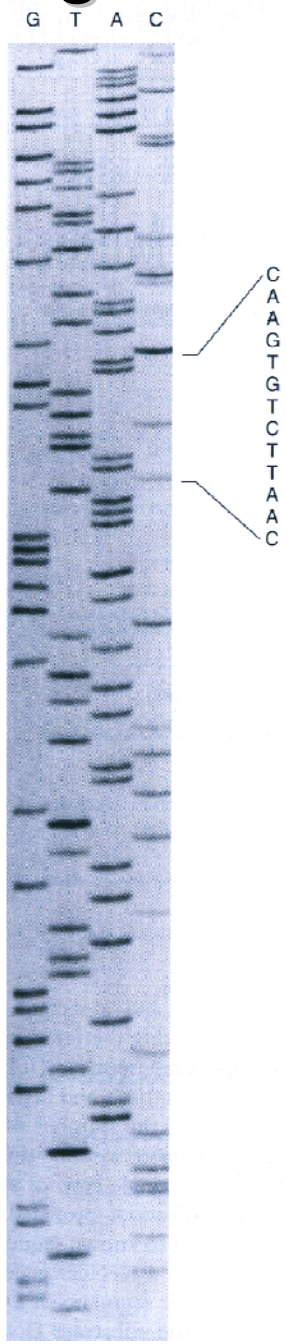
Sanger - Inventor of chain-termination dideoxynucleotide method for sequencing

Frederick Sanger

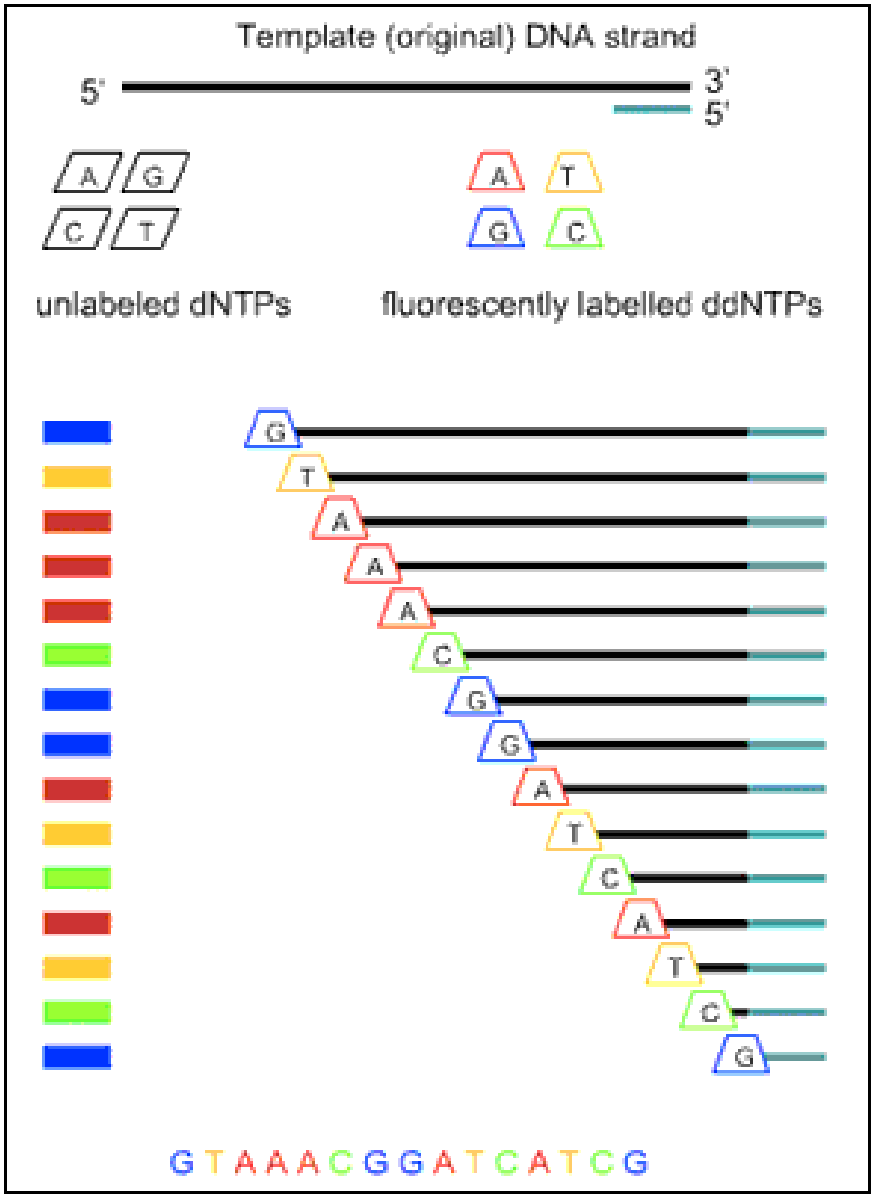
Born	August 13, 1918 Gloucestershire, England
Nationality	United Kingdom
Fields	Biochemist
Institutions	Laboratory of Molecular Biology
Alma mater	St John's College, Cambridge
Notable awards	Nobel Prize in Chemistry (1958) Nobel Prize in Chemistry (1980)

Sanger – radioactive dideoxy sequencing



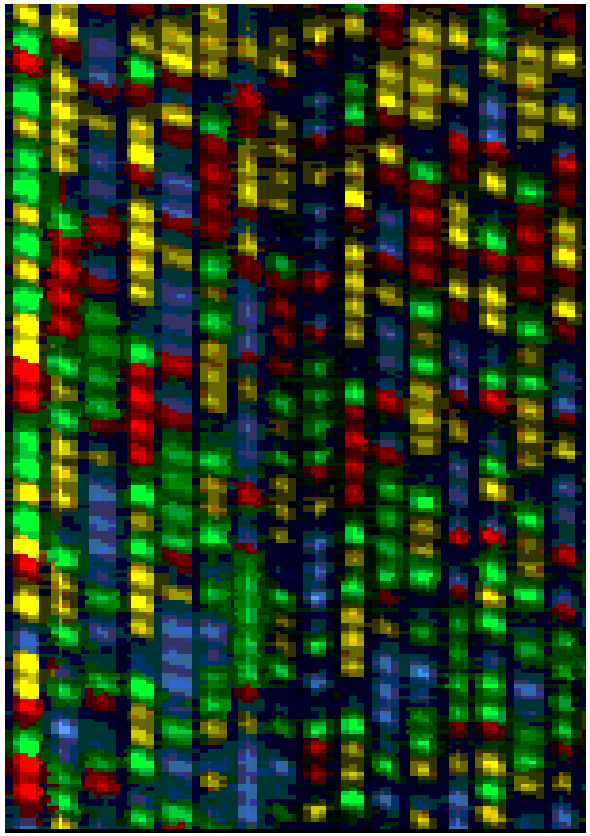


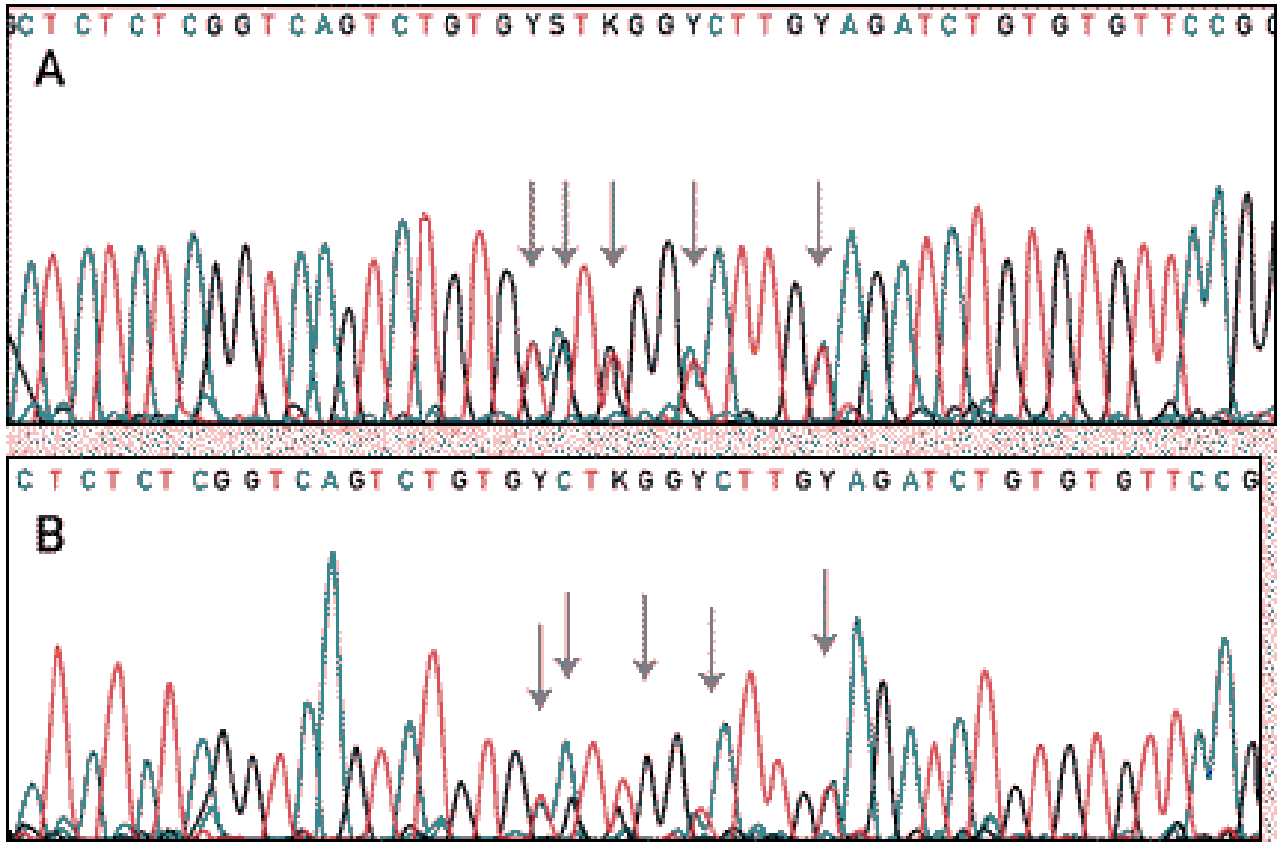
Sanger- Chain-termination by fluorescent dye



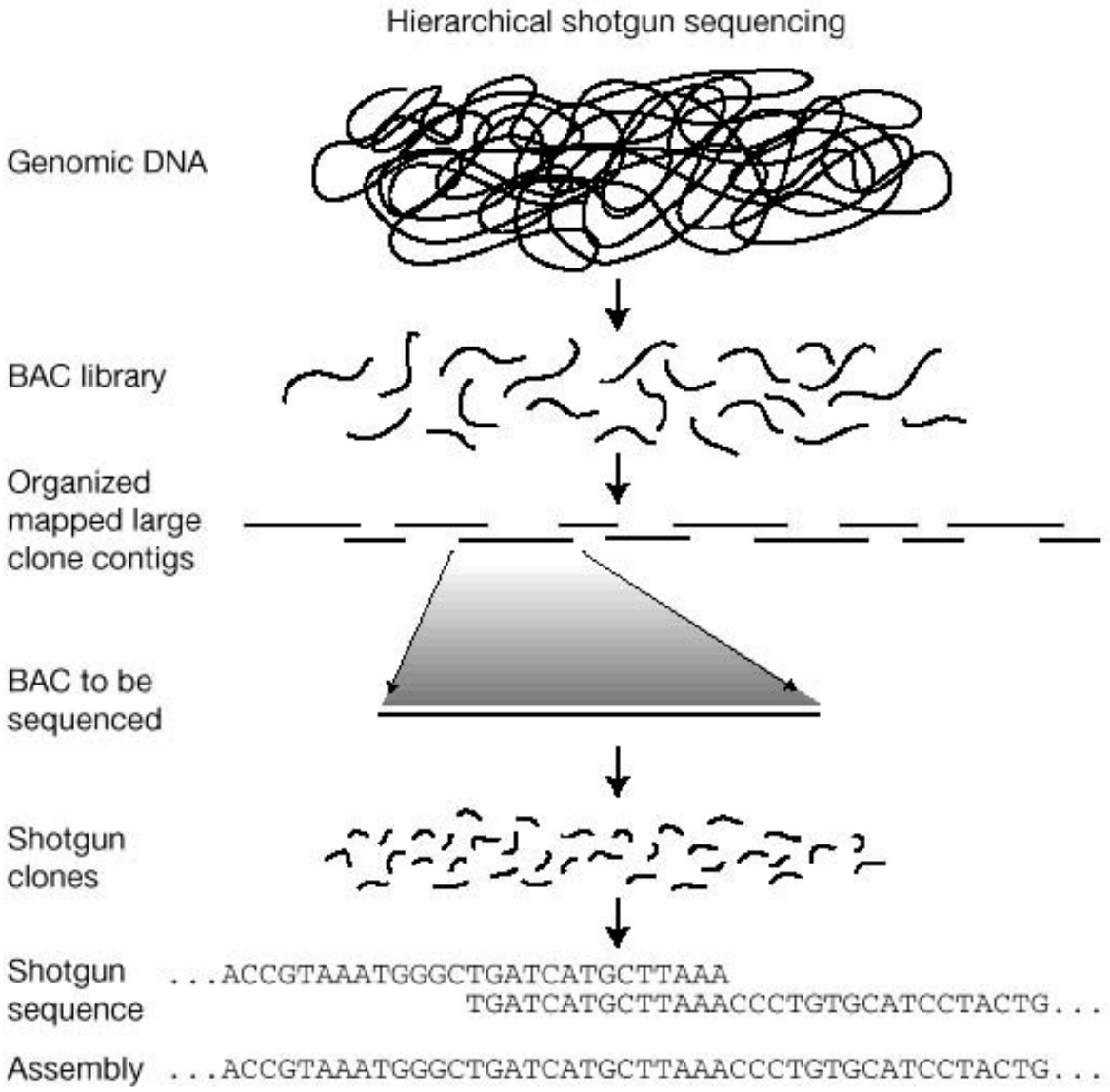
- Fluorescently labeled ddNTP
- **Irreversibly terminated**
- High raw accuracy
- Read lengths: up to 1000
- 384 samples / run

Fluorescent Dye-Terminator Cycle Sequencing



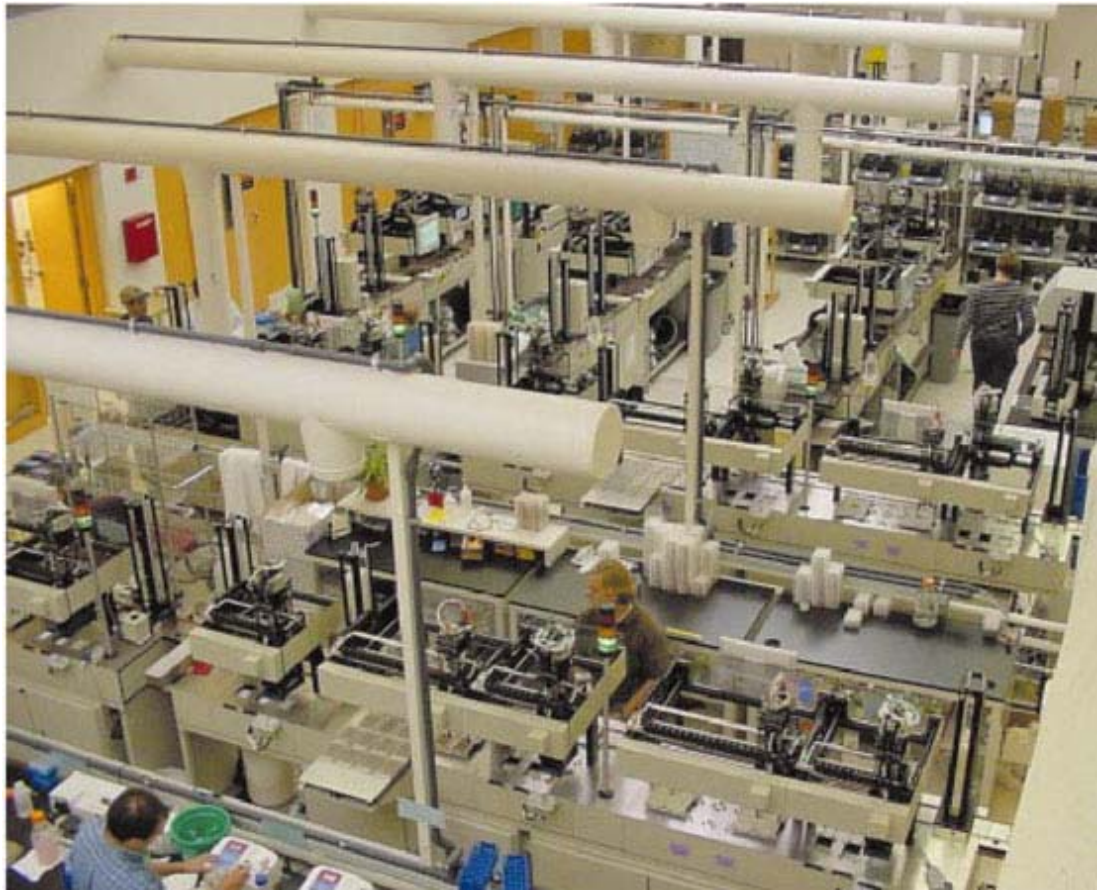


Shotgun Sequencing



Nature 409, 860-921(2001)

Library factory - Whitehead Institute



Sequencing factory - Sanger Institute

Nature 409, 860-921(15 February 2001)

NGS (Next Generation Sequencing) – benefits over Sanger sequencing

- Massive parallel sequencing of shotgun libraries
- Use universal primer on adaptors
  - No need for prior sequence knowledge (good for non-model organisms)
- No bacterial cloning (less representation bias), no walking
- No seq. gel pouring (fast, save labor)
- High throughput (great coverage depth)
- More cost-effective
- Diversified applications
- Various analysis tools available
- Higher sensitivity than array-based detection
- Fast evolving for even greater performance

II. Introduction of NGS

NGS platforms

- **Current technologies**
  - Roche 454
  - Illumina Genome Analyzer
  - Applied Biosystems' SOLiD

- **Emerging sequencing technologies (3rd generation)**
  - eliminate PCR amplification of templates
  - true Single-Molecule-Sequencing (tSMS) technology
  - Helicos, PacificBio, etc
  - stability yet to be verified

Next-generation DNA sequencing

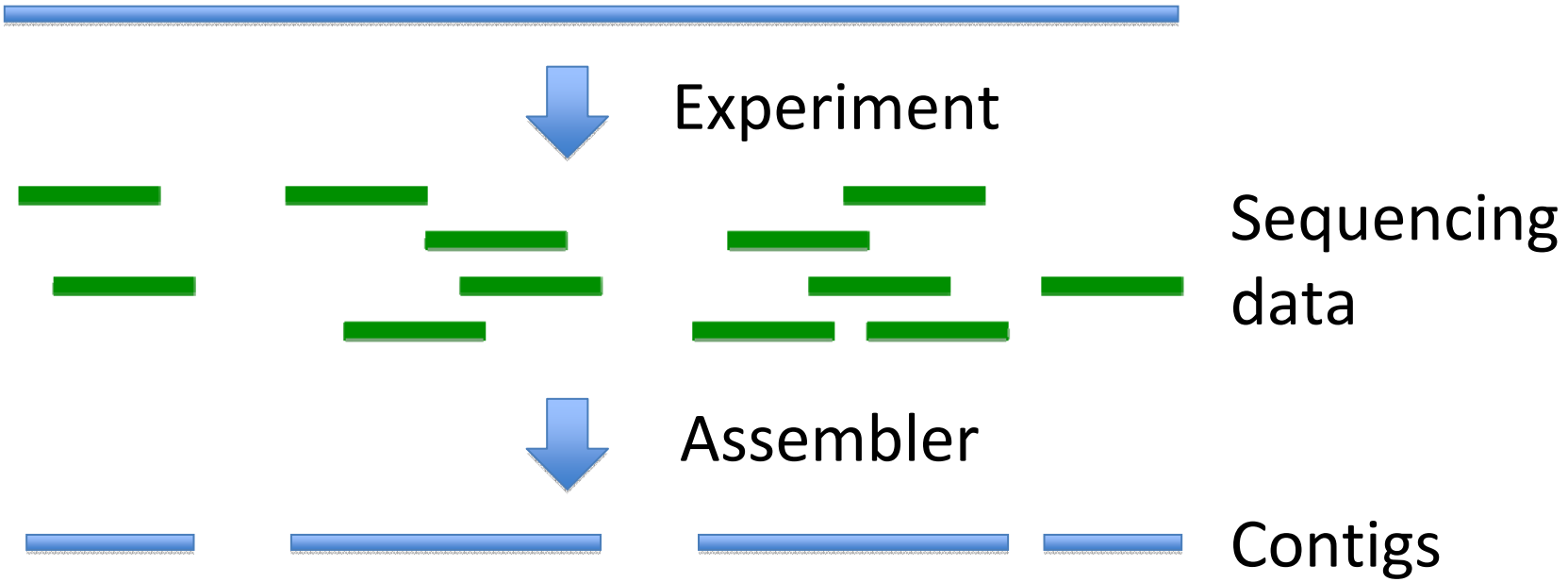
Jay Shendure[1] & Hanlee Ji[2]

***Nature Biotechnology* 26**, 1135 - 1145 (2008)

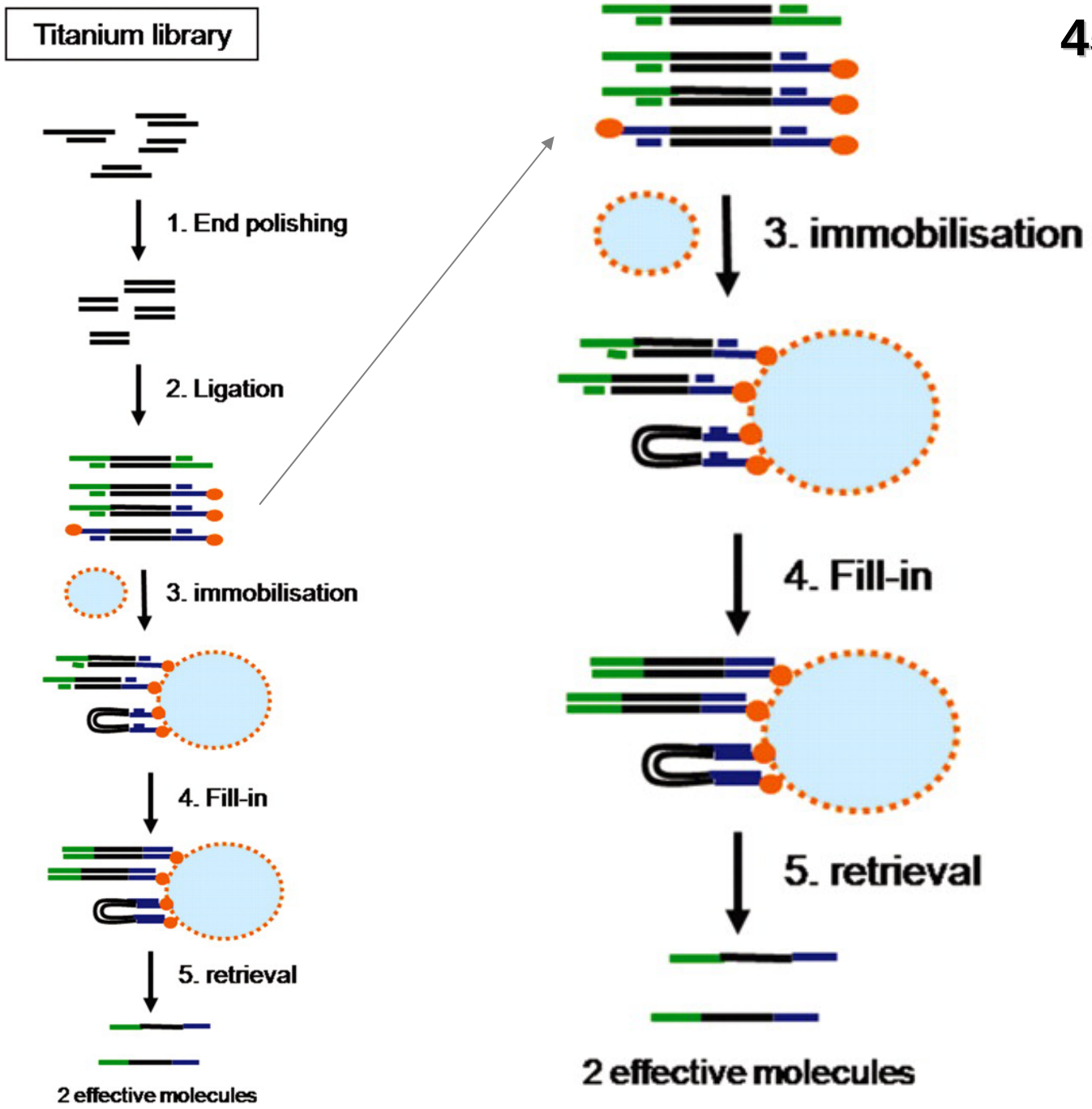
***Nature Review Genetics* 11,** 31-46 (2010)

Shotgun sequencing

- Random samplings
- Read length: Sanger (~1000 bps), NGS: 454 (~400 bps), Solexa (75~110 bps)

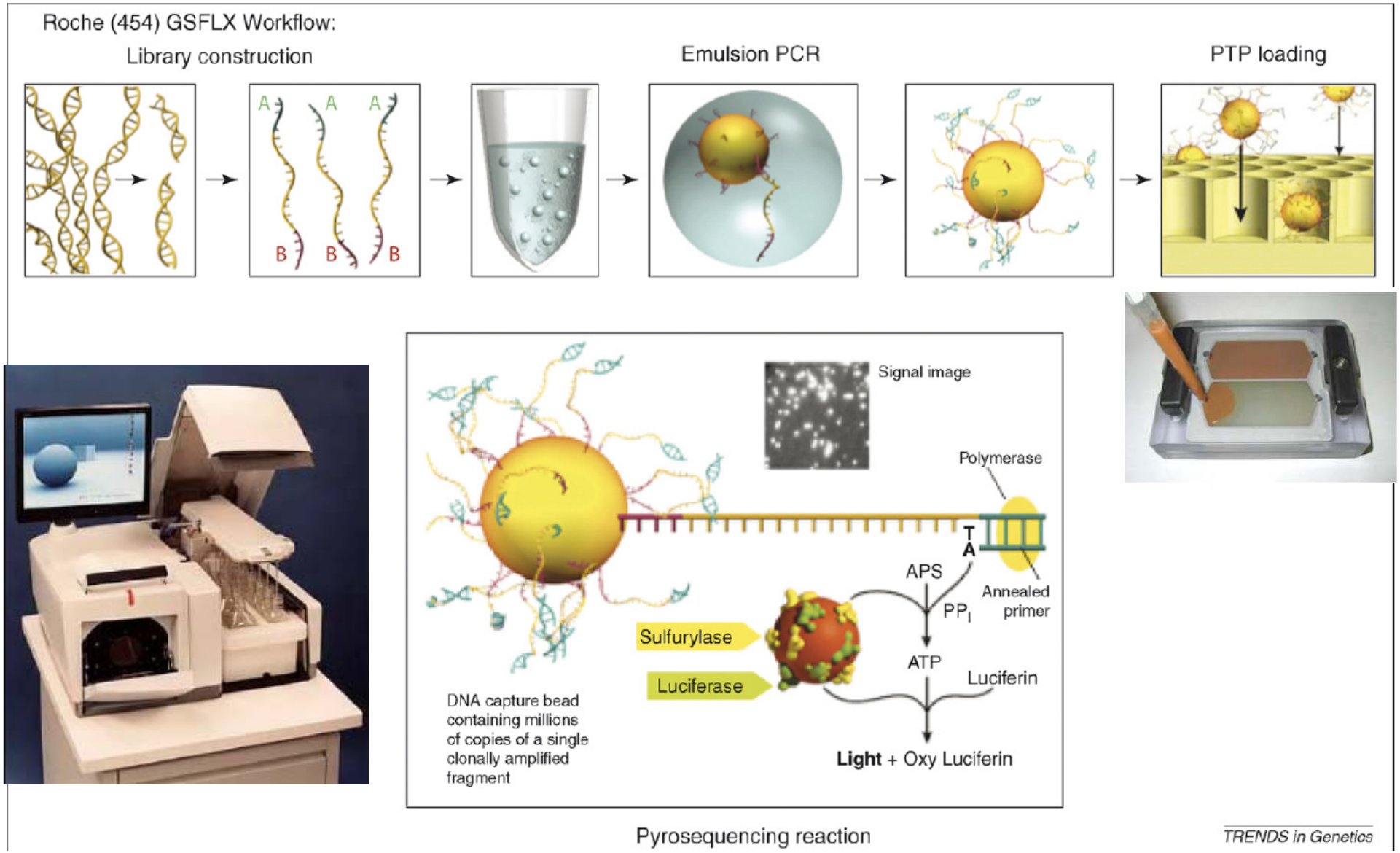


454 Library prep

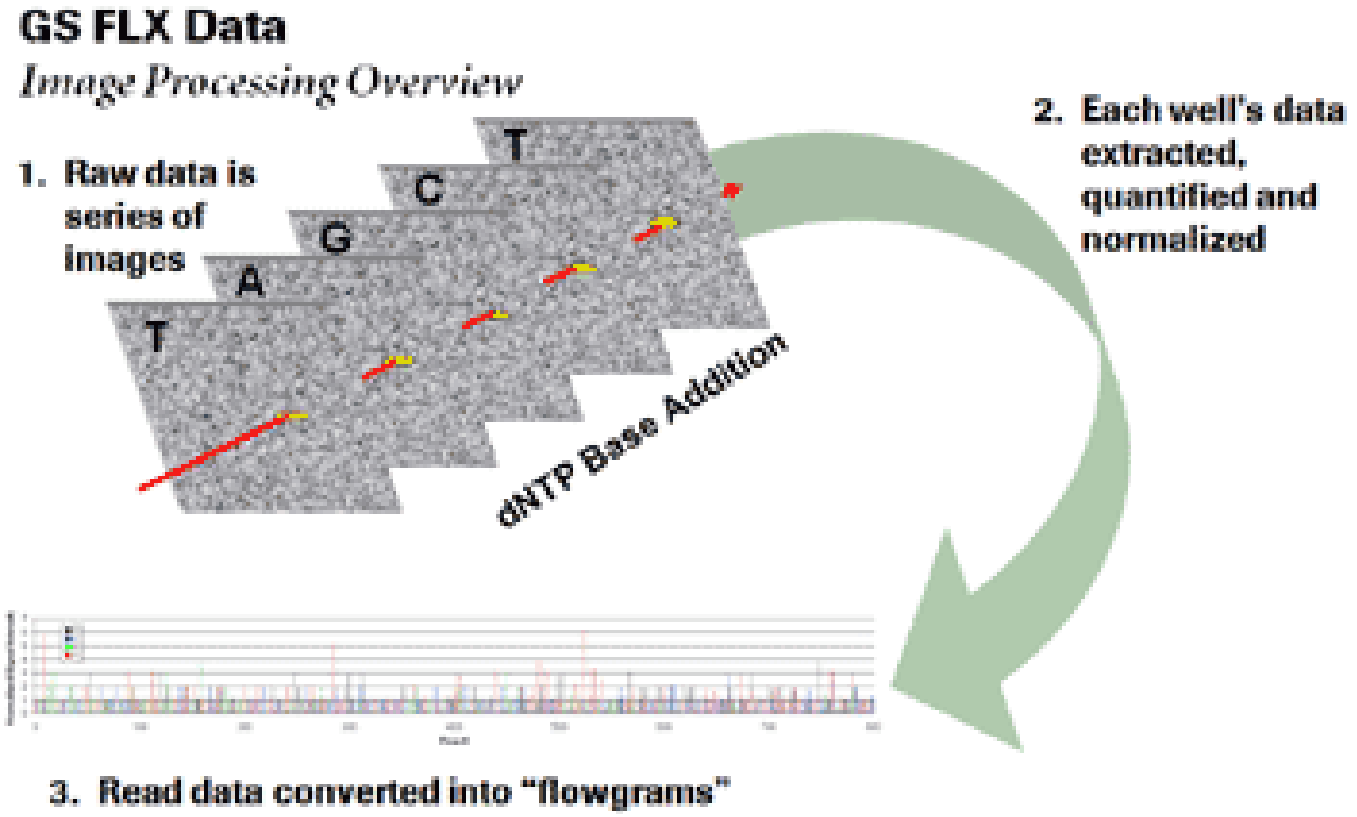


Nucleic Acids Research, 2010, 1–9

454: emPCR/pyrosequencing



Sequential nucleotide flow & full imaging

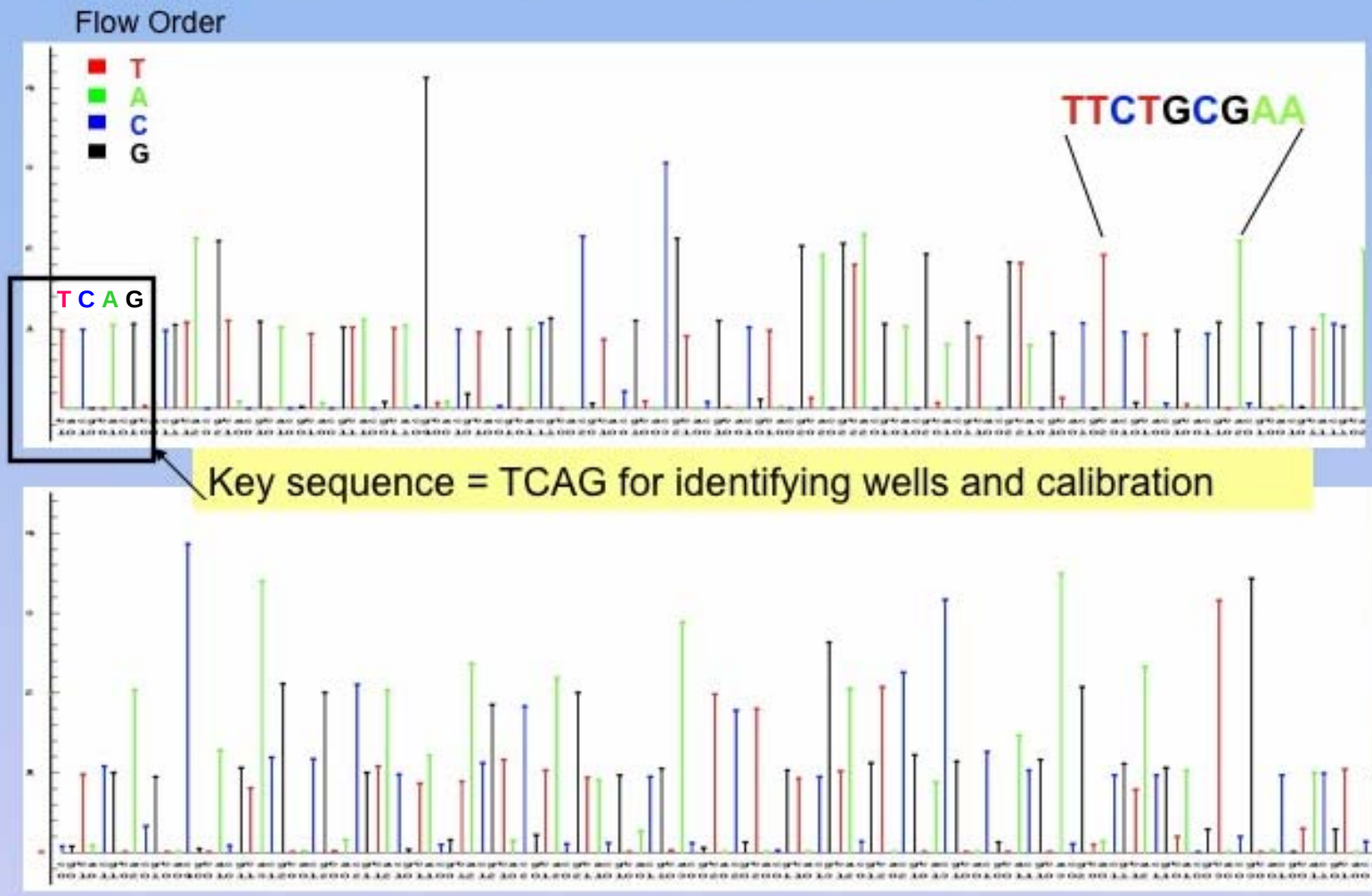


http://www.youtube.com/watch?v=bFNjxKHP8Jc&feature=related

Genome Sequencer FLX Multimedia Presentation: Watch this short flash presentation describing the GS FLX Instrument and an overview of the technology involved in a sequencing run.
http://www.roche-applied-science.com/publications/multimedia/genome_sequencer/flx_presentation/wbt.htm

Genome Sequencer FLX Standard Series Workflow Presentation: Follow this interactive tutorial through the entire GS FLX system workflow, from sample prep to data analysis. (Note: This presentation does not describe GS FLX Titanium series workflow.)
http://www.454.com/products-solutions/multimedia-presentations.asp

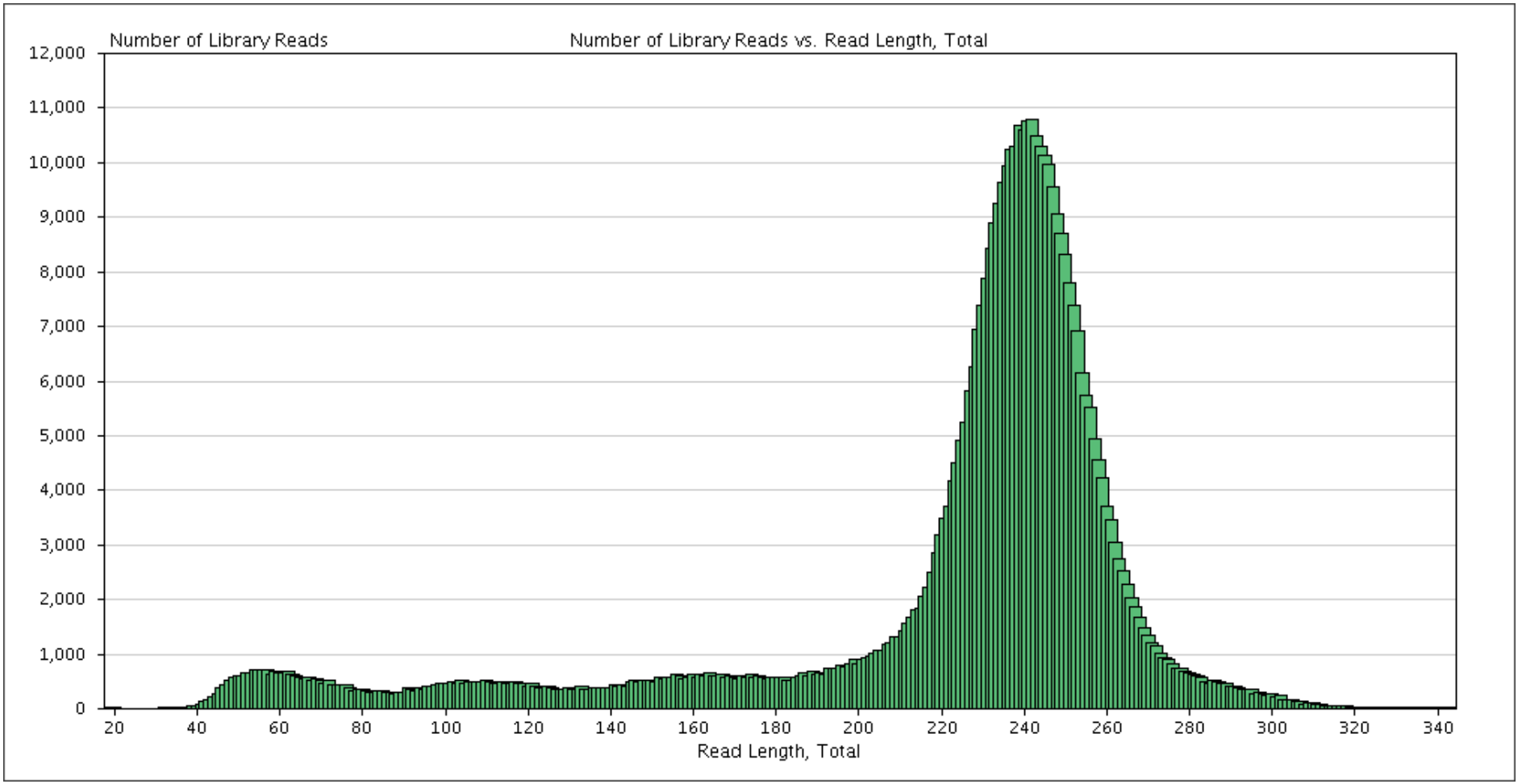
Example of a Flowgram



http://www.youtube.com/watch?v=kYAGFrbGl6E

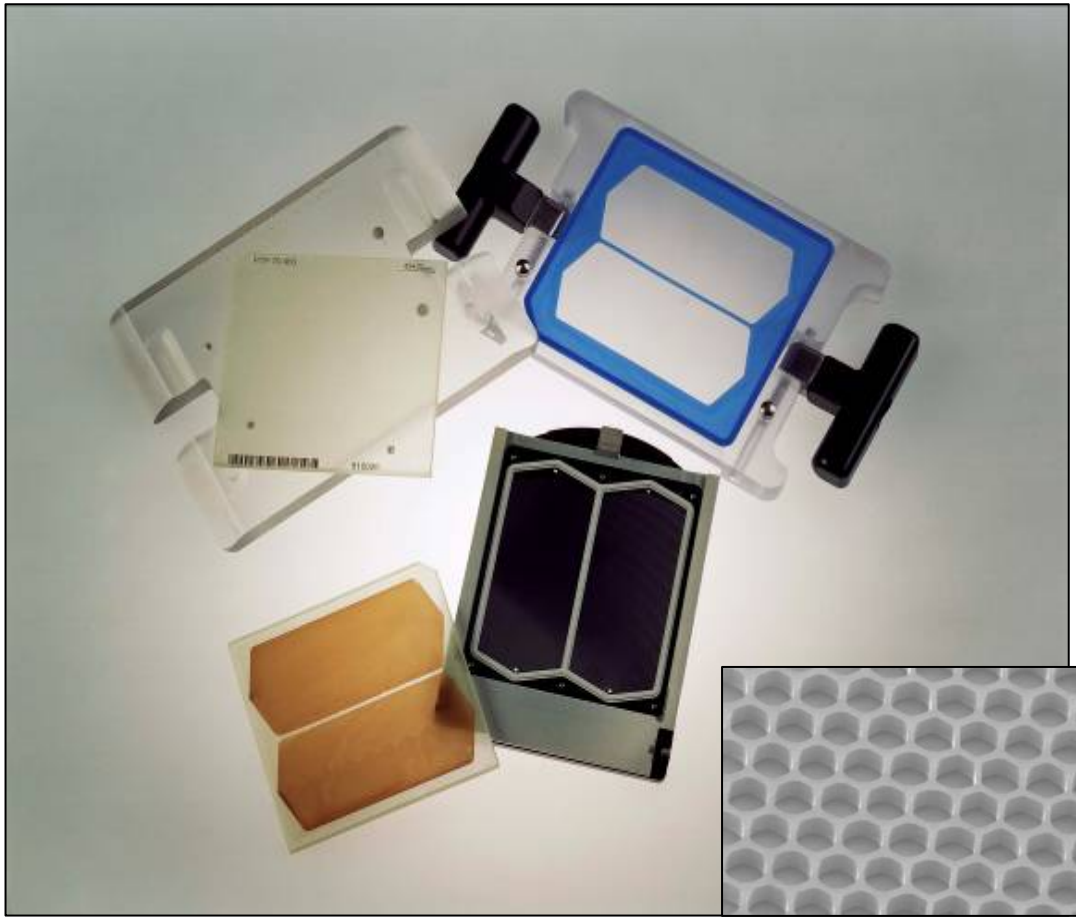
One full run 454 Seq.:

Total = 101 Mb



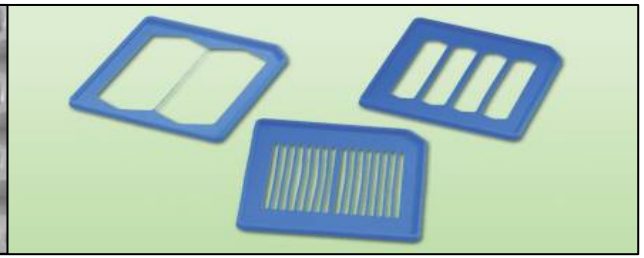
GS FLX Throughput

Multiple Gasket Formats and Plate Sizes Provide Flexibility in Sample Loading and Throughput



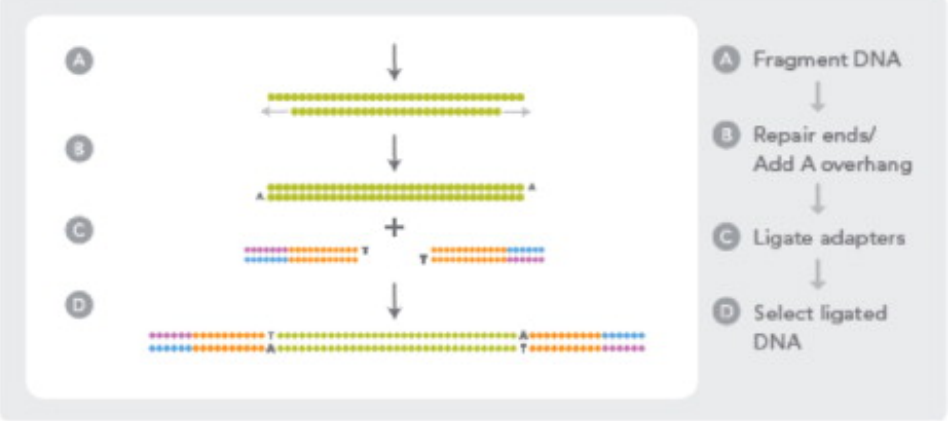
70 x 75 PicoTiterPlate

- 2- lane gasket
- 4- lane gasket
- 8- lane gasket
- 16- lane gasket

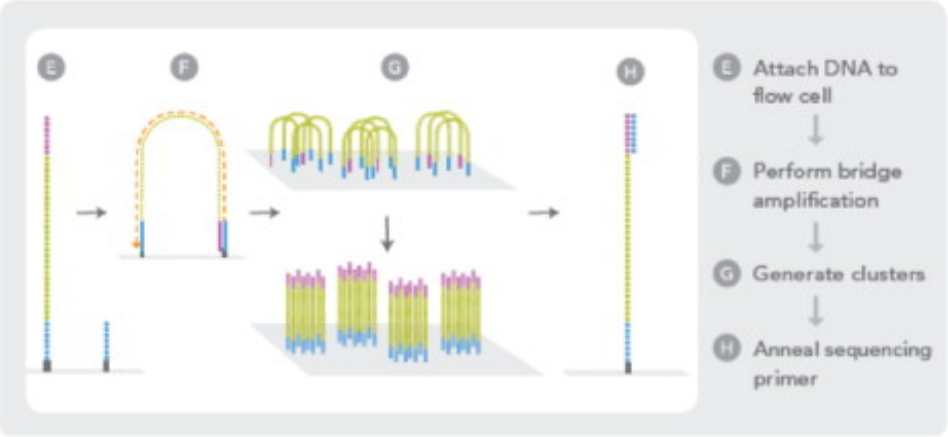


Solexa: Reversible terminator / SBS

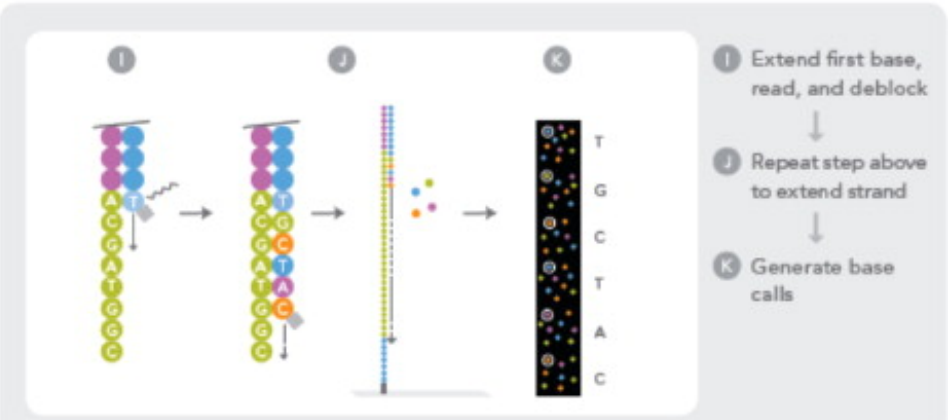
I.

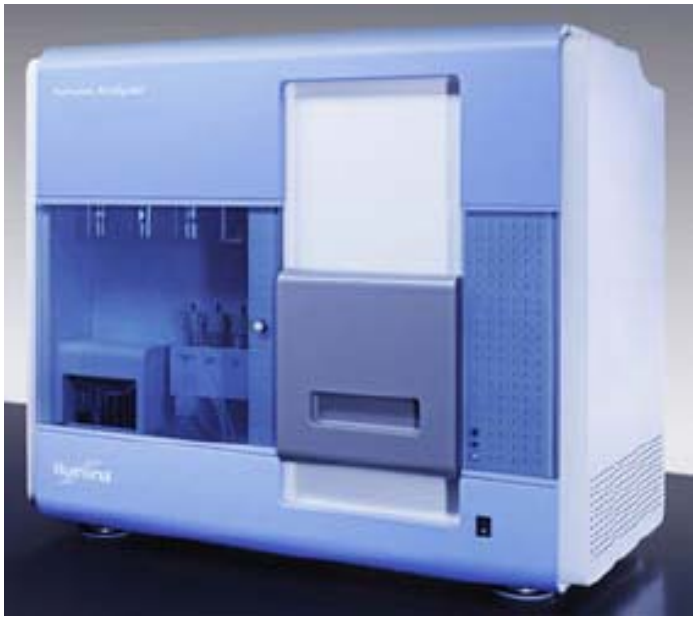


II.

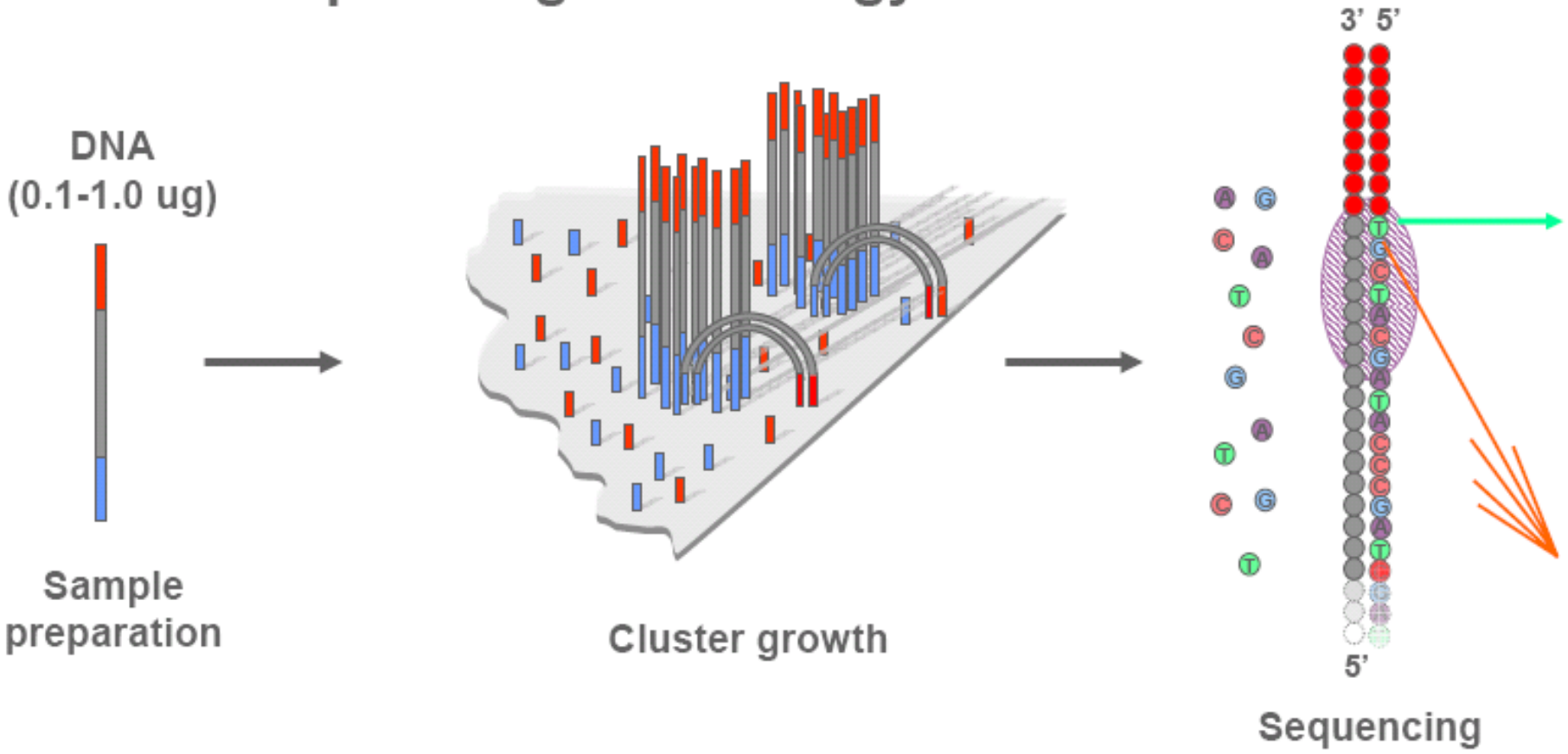


III.



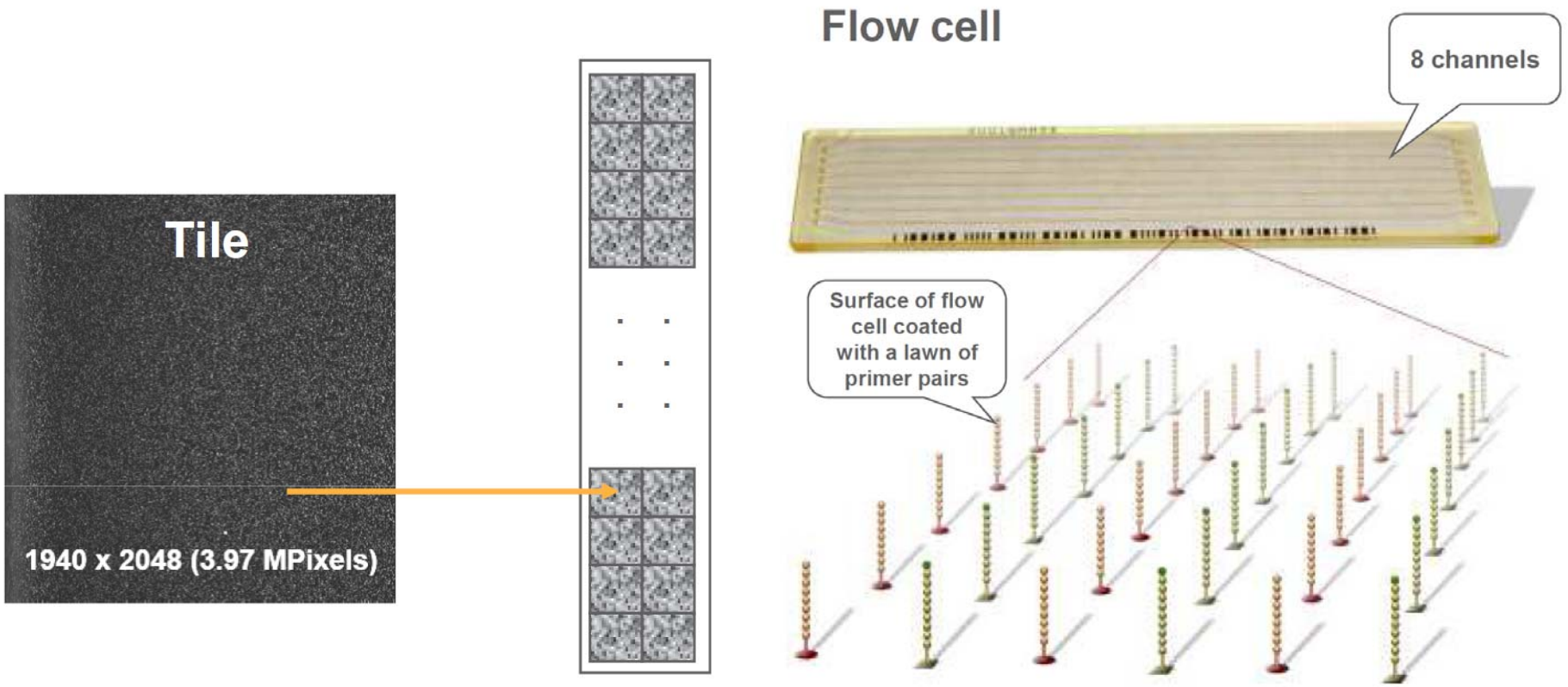


Illumina Sequencing Technology

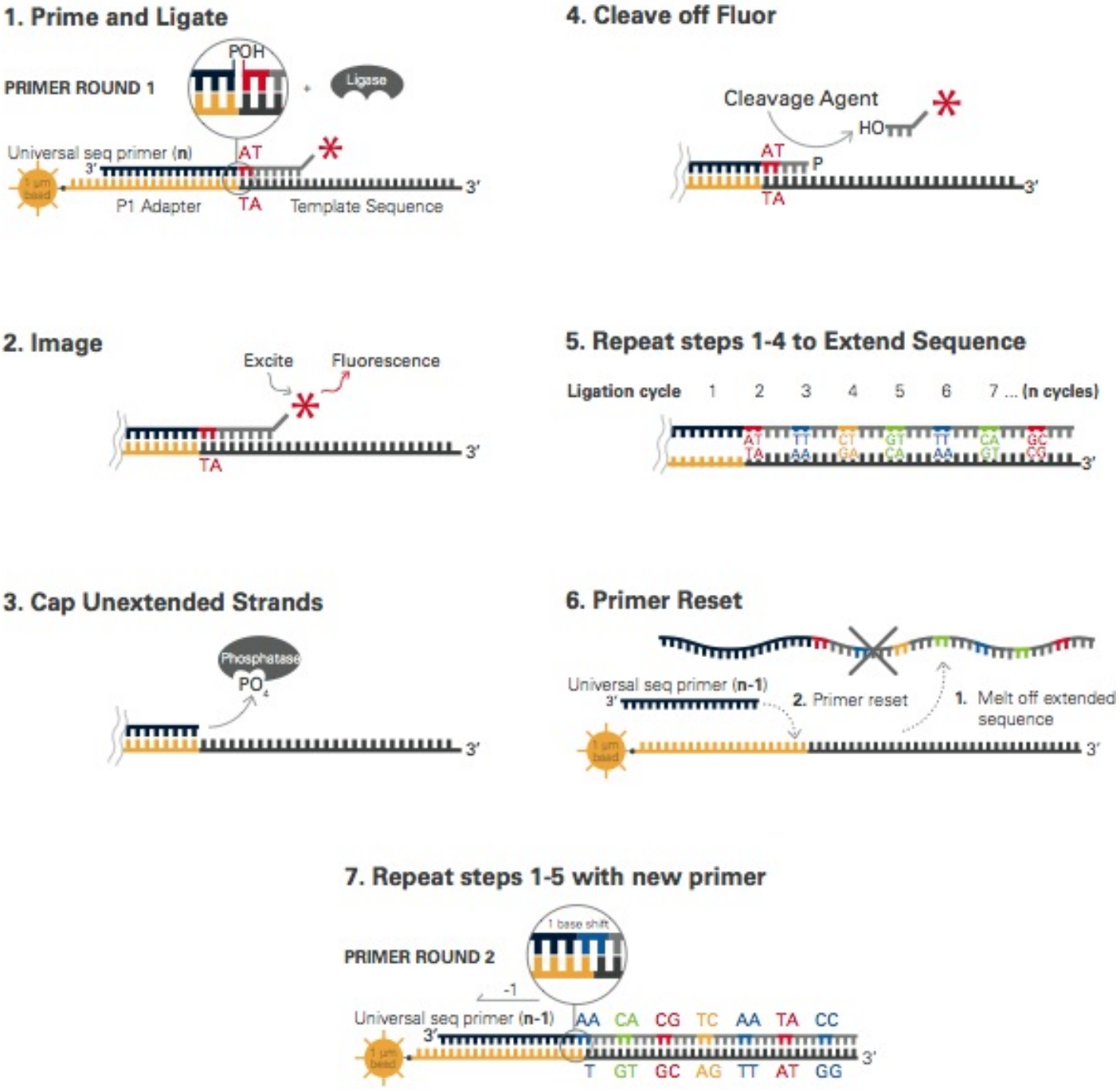


http://www.youtube.com/watch?v=77r5p8IBwJk&feature=related

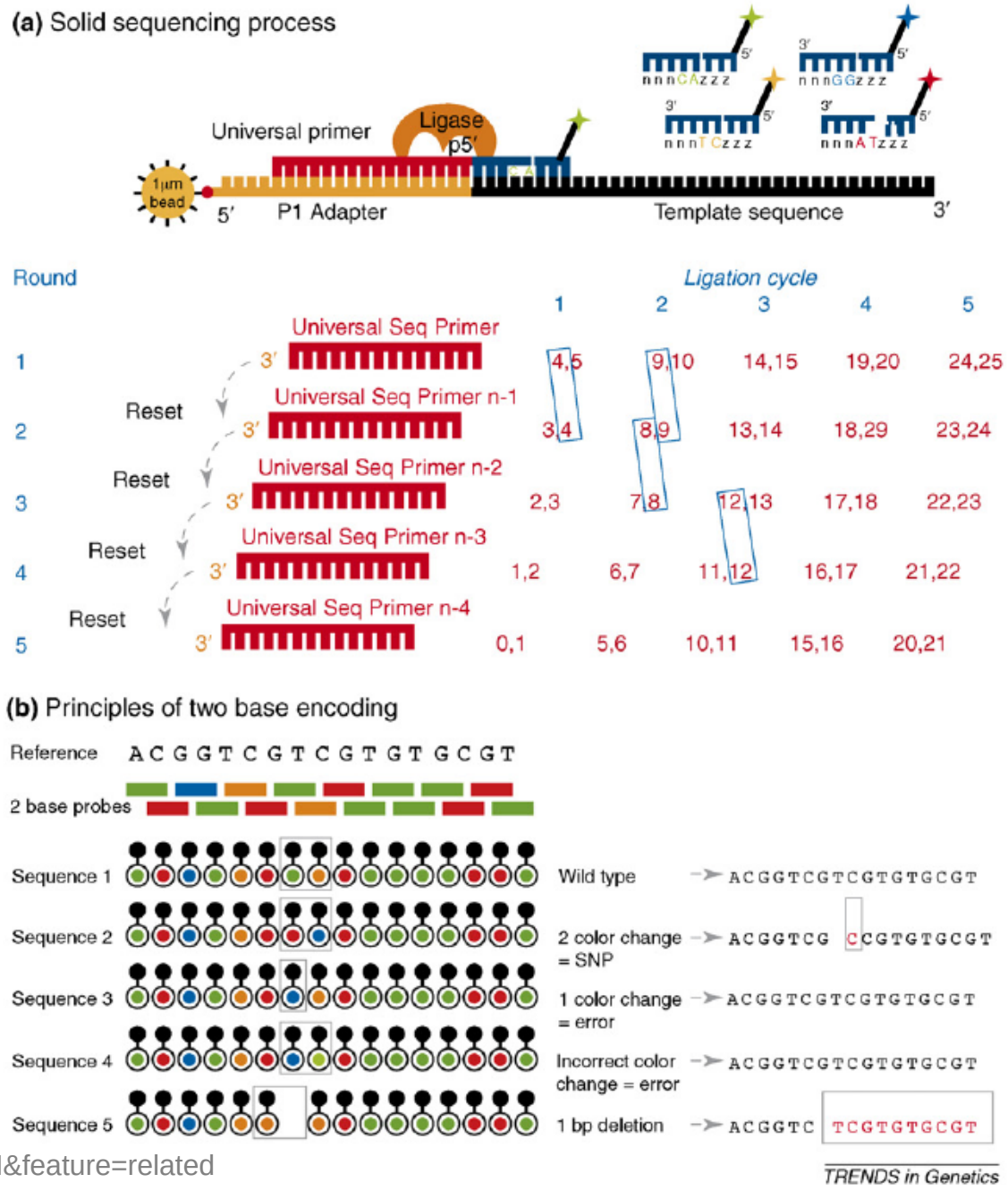
Illumina GA – Flow cell imaging



SOLiD: ligation & 2-base encoding



SOLiD: ligation & 2-base encoding



http://www.youtube.com/watch?v=nlvyF8bFDwM&feature=related

Commercial NGS platforms

	GS FLX Titanium (Roche 454)	Genome Analyzer IIx (Illumina/Solexa)	SOLiD3 (ABI)
Sequencing chemistry	Pyro-sequencing	Polymerase-based sequencing-by-synthesis	sequential ligation with dye-labeled oligonucleotides
Amplification	Emulsion PCR	Bridge amplification	Emulsion PCR
Starting DNA (µg)	3-5 µg	3-5 µg	3-5 µg
Output/run	400-500 Mb	up to 50 Gb/run (PE2*120) (6 Gb/channel)	60 Gb/run
Read length (nt)	avg 400 nt	36, 75, 100 nt	35-50 nt
Time/run	10 h	4, 8, 12 days (PE)	6-12 days
Data processing/run	16 h	2-5 days	2-4 days

Facts about 454 GS FLX

- **Pros:**
  - **Long reads for better assembly; suitable of *de novo* sequencing**
  - **Shorter time required for sequencing and data processing**

- **Cons:**
  - **Error-prone in homopolymeric regions**
  - **Higher cost**
  - **Sample preparation is labor-intensive**

Facts about Illumina GA/Solexa

- **Pros:**
  - **Greater output**
  - **More affordable**
  - **High number of reads provide greater depth and sequence confidence**
  - **Sequence one base a time (no homopolymer concerns)**

- **Cons:**
  - **Short reads; mostly for applications with available genome information**
  - **Takes longer time to run and process data**
  - **Demanding for computing power and data storage**

III. NGS data output and de novo assembly

General terms for NGS

- **Read**: a sequence from one template of the sample library
- **Contig**: a set of overlapping DNA segments derived from a single genetic source; assembled from overlapping reads
- **Coverage**: the average number of reads representing a given nucleotide in the reconstructed sequence

  (= total read length / genome size)

Phred quality scores *Q* (base calling)

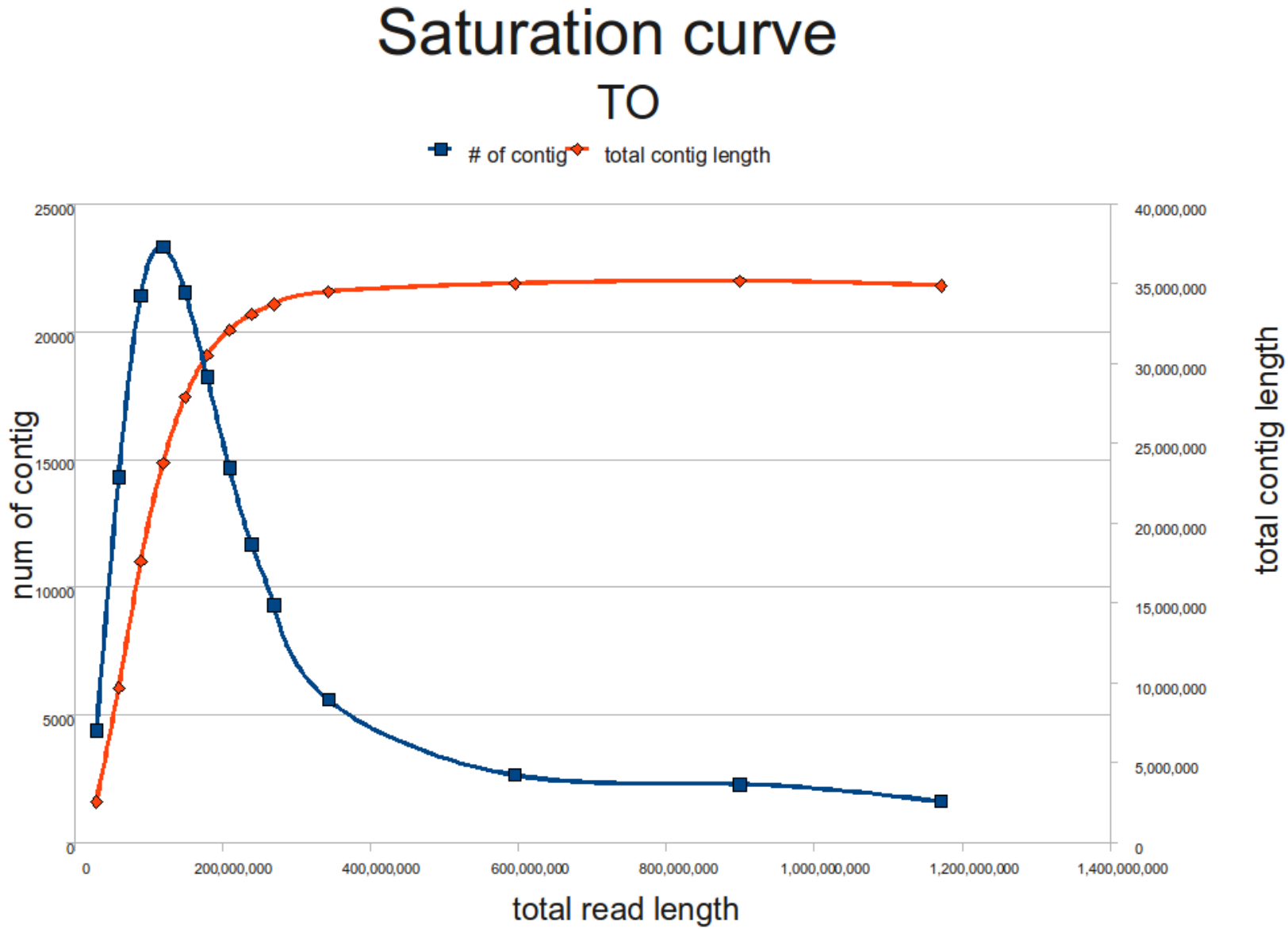
- Phred quality scores *Q* are logarithmically related to error probabilities *P* by $Q = [-10.\log10(P)]$. For example, if Phred assigns a quality score of 30 to a base, the chances that this base is called incorrectly are 1 in 1000. The most commonly used method is to count the bases with a quality score of 20 and above. The high accuracy of Phred quality scores make them an ideal tool to assess the quality of sequences.

Quality of Phred Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90 %
20	1 in 100	99 %
30	1 in 1000	99.9 %
40	1 in 10000	99.99 %
50	1 in 100000	99.999 %

NGS applications

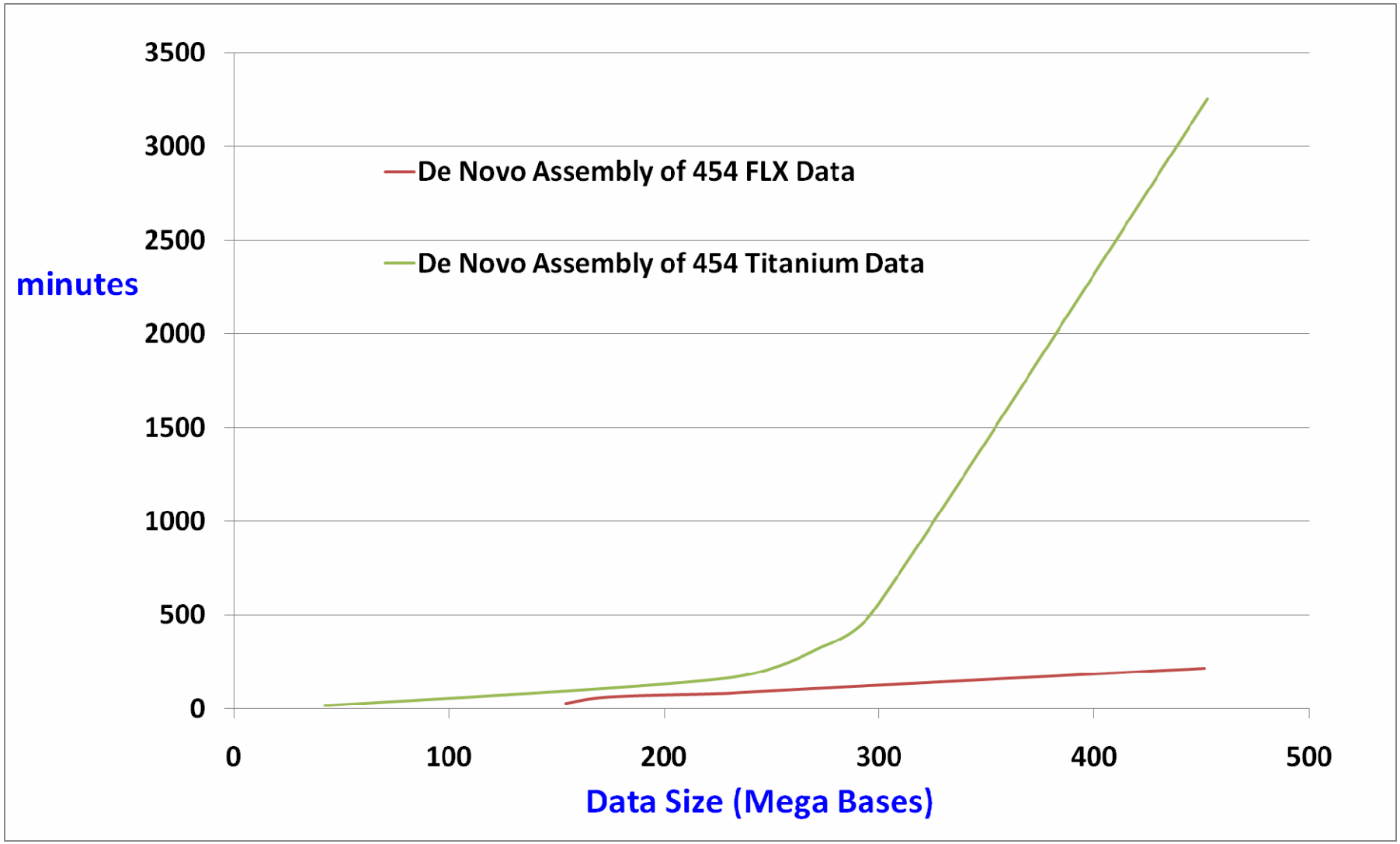
- De novo sequerncing & assembly
  - New genome or transcriptome
  - Alternative splicing
  - Metagenomics
- Re-sequencing with reference genome
  - Variation discovery (SNP, INDEL, CNV)
  - Transcriptome quantification
  - ChIP-sequencing
  - New gene discovery

De novo Assembly / genome coverage

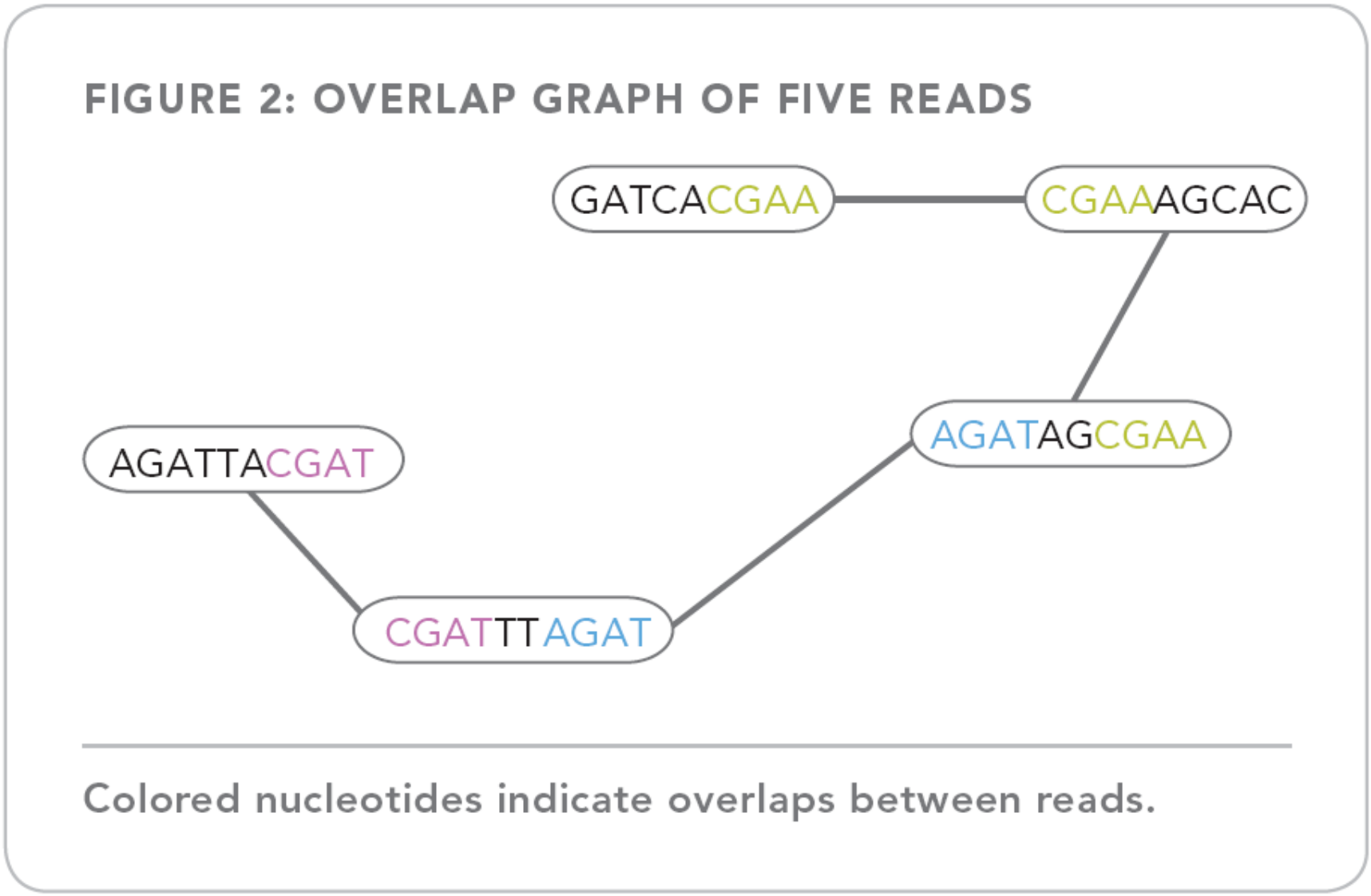


* Newbler v2.3

Time elapsed



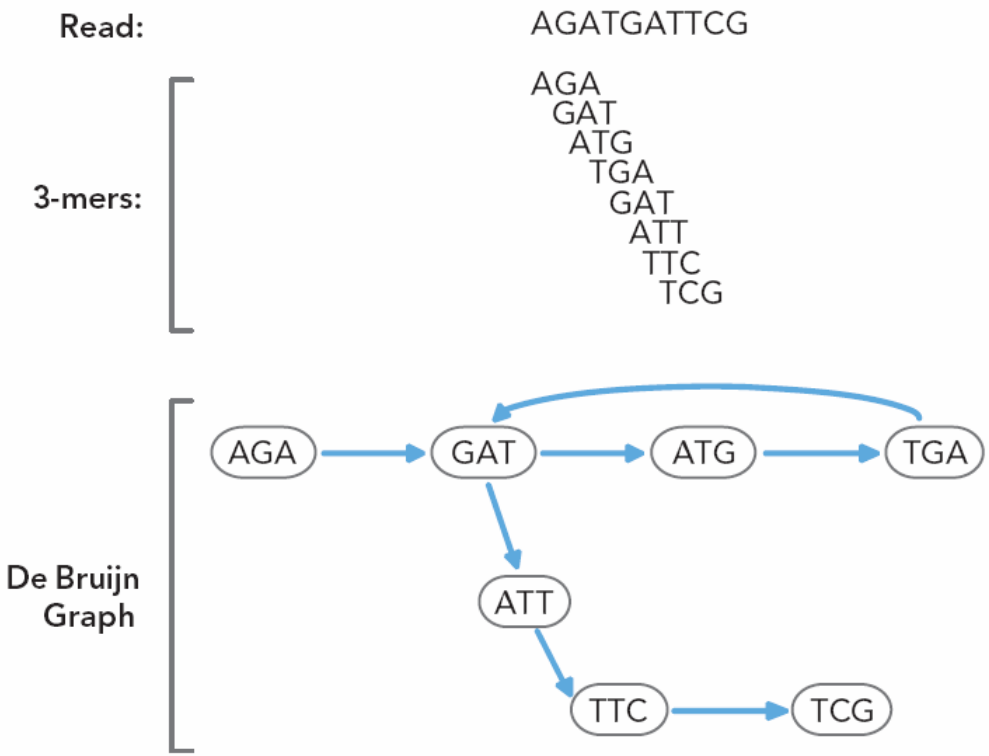
Assembly - overlap graphs



Source: Illumina

Assembly – de Bruijn graph of k-mers

FIGURE 3: DE BRUIJN GRAPH FOR READ WITH K = 3



The length of overlaps is k-1=2. Gray arrows indicate where all the k-mers derived from the one read are placed in the graph. Blue arrows indicate the order of the k-mers and their overlaps.

Source: Illumina

Short-read assemblers

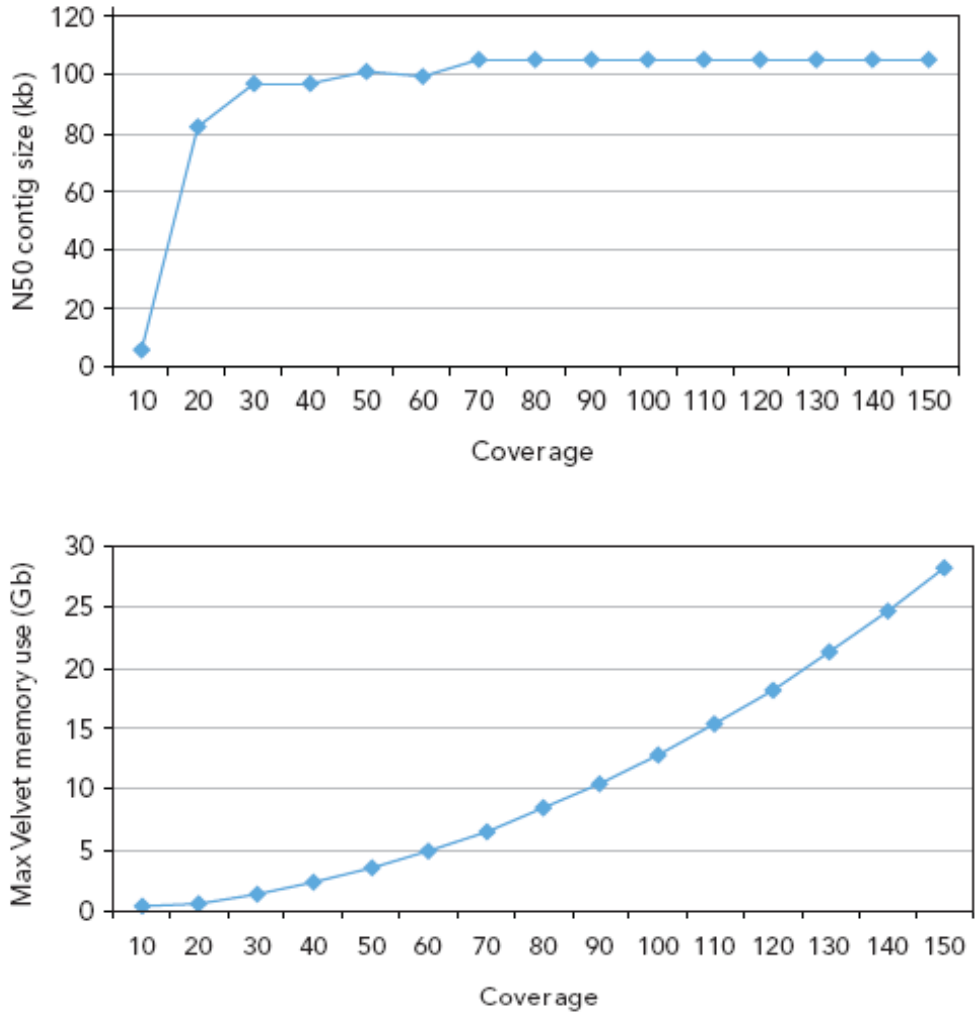
TABLE 1: OVERVIEW OF TESTED ASSEMBLERS

ALGORITHM	DESCRIPTION	STRENGTH	GENOMES ASSEMBLED
Velvet	De Bruijn graph based Error corrections after graph is built	Fast (~30 mins) Easy to use Larger supercontig N50	Bacterial (Ref. 1; this technical note)
SOAPdenovo	De Bruijn graph based Error correction before graph is built	Easy to use Multi-threaded mode	Panda, Bacterial (Ref. 11; this technical note)
ABySS	De Bruijn graph based Can be run in parallel Distributed memory model (efficient)	Easy to use Largest contigs/scaffolds Best suited for large genomes	Human (Ref. 3; this technical note)
Forge	Overlap-layout-consensus method Modifications to accommodate Illumina reads	Largest contigs/supercontigs Good "long read" assembler	Bacterial (this technical note)

Source: Illumina

How much is enough?

FIGURE 4: EFFECT OF COVERAGE



Effect of coverage on N50 contig size and memory requirements in an *E. coli de novo* assembly.

Source: Illumina

How much is enough?

TABLE 2: EFFECT OF COVERAGE ON ASSEMBLY QUALITY

COVERAGE	N50 CONTIG SIZE	LARGEST CONTIG	GENOME COVERAGE
320×	95,313 bp	215,645 bp	99.47%
160×	95,368 bp	209,234 bp	99.72%
50×	97,333 bp	223,793 bp	99.72%
21×	35,828 bp	119,071 bp	99.38%

Source: Illumina

Paired-End vs Single Read

TABLE 3: EFFECT OF READ LENGTH

SAMPLE	N50 CON-TIG SIZE	LARGEST CONTIG	GENOMRE COVERAGE
E. coli, 100 bp pe	132,786 bp	326,886 bp	99.87 %
E. coli, 400 bp sr	22,902 bp	127,976 bp	99.87 %
Chr. 20, 100 bp pe	70,744 bp	484,312 bp	92.69 %
Chr. 20, 400 bp sr	2,319 bp	22,823 bp	92.65 %

Source: Illumina

Bridging helps contig connection

TABLE 6: EFFECT OF INSERT SIZE ON ASSEMBLY

INSERTS	READS (BP)	COV-ERAGE	N50 SUPER-CONTIG	LARGEST SUPER-CONTIG	GENOME COVER-AGE
200 bp	2×75	50×	97 kb	223 kb	99.58%
200 bp + 6 kb	2×75 2×35	50× 28×	1.3 Mb	2.1 Mb	99.07%
200 bp +10 kb	2×75 2×35	50× 28×	4.5 Mb	4.5 Mb	99.69%

Source: Illumina

Clean up the data!

TABLE 7: EFFECT OF FILTERING ON ASSEMBLY QUALITY

FILTERING	READ COVER-AGE	N50 CONTIG SIZE (BP)	LARGEST CONTIG (BP)	GENOME COVER-AGE
No filtering	420×	12,083	62,228	99.37 %
Only PF	328×	95,351	209,222	99.63 %
PF + Ns removed	320×	95,313	215,645	99.47 %
PF + Ns + s35 removed	203x	95,338	268,040	99.58 %

Source: Illumina

IV. Other NGS applications

Re-sequencing apps.

- ChIP-seq (DNA-protein interactions)
  - Regulatory protein binding
  - Chromatin modification and packaging
- Transcriptome sequencing
  - Transriptional networks
  - discover novel genes, splicing variants, and ncRNAs

Short-Read Alignment Tools

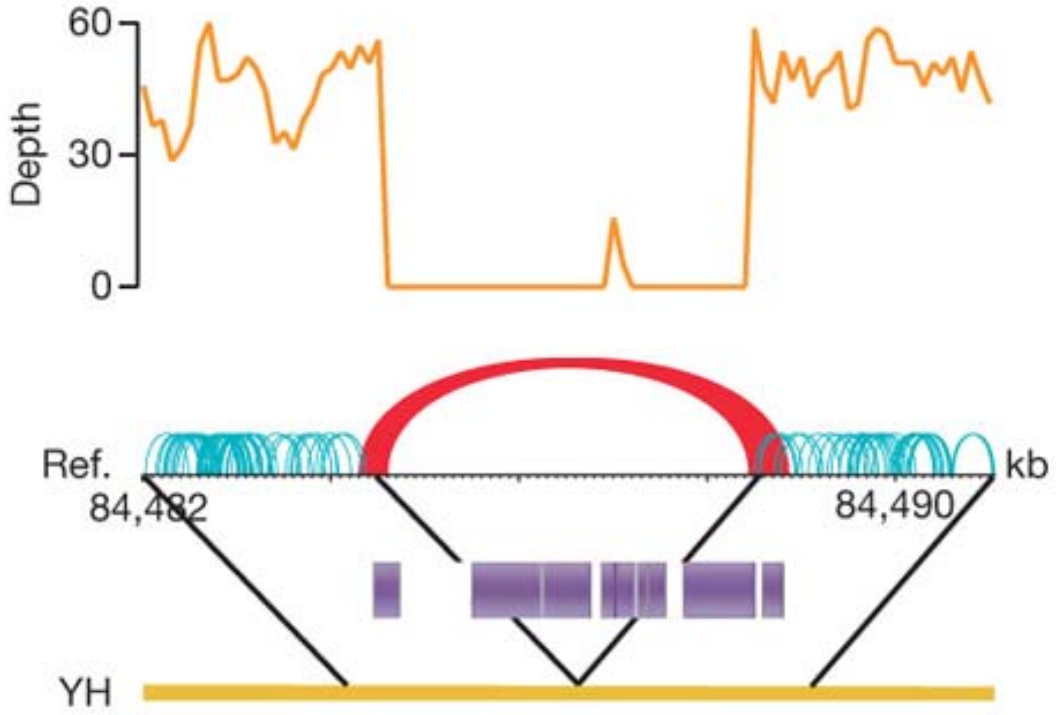
Table 2 A summary of short-read alignment tools

BWA	Illumina	454	SOLiD	S	I	URL
Bowtie *	Y	Y	N	Y	N	http://bowtie-bio.sourceforge.net
ELAND *	Y	N	N	N	N	http://www.illumina.com
Exonerate	Y	Y	N	N	Y	http://www.ebi.ac.uk/~guy/exonerate/
GMAP	Y	N	N	N	N	http://www.gene.com/share/gmap
MOSAIK *	Y	Y	Y	Y	Y	http://bioinformatics.bc.edu/marthlab/Mosaik
MAQ *	Y	N	Y	Y	Y	http://maq.sourceforge.net
MUMer *	Y	Y	N	Y	Y	http://mummer.sourceforge.net/
Novocraft	Y	N	N	Y	Y	http://www.novocraft.com/
RMAP	Y	N	N	N	N	http://rulai.cshl.edu/rmap/
SeqMap	Y	N	N	N	I	http://biogibbs.stanford.edu/~jiangh/SeqMap/
SHRiMP	Y	Y	Y	Y	Y	http://compbio.cs.toronto.edu/shrimp/
SOAP *	Y	N	N	Y	Y	http://soap.genomics.org.cn/
SSAHA2	Y	Y	N	Y	Y	http://www.sanger.ac.uk/Software

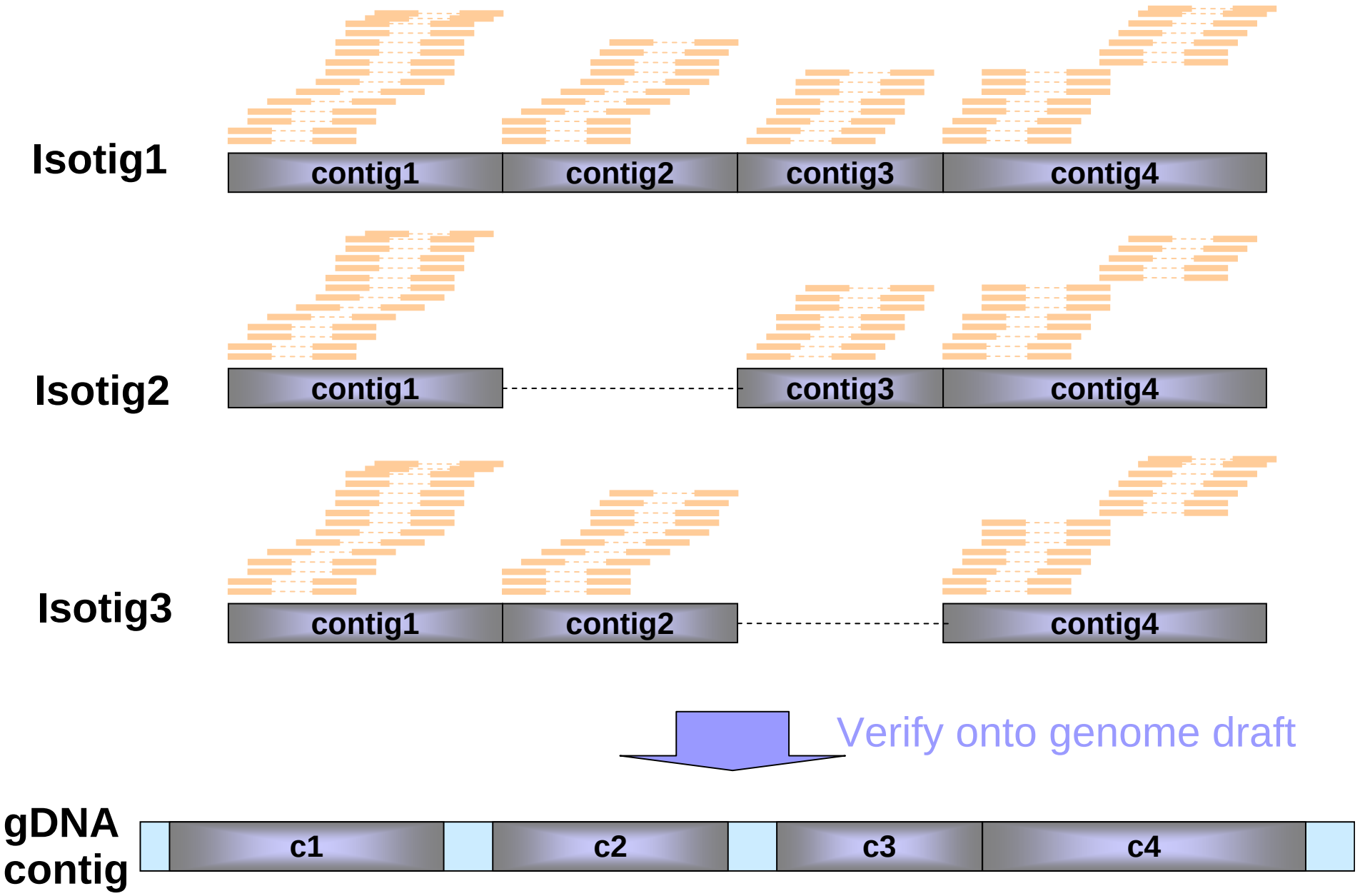
S outputs SNPs, I outputs short insertion deletions (indels)

D. J. Turner et al.: Next-generation sequencing of vertebrate experimental organisms

Structural variations



RNA-seq.


Isotig1
contig1
contig2
contig3
contig4
Isotig2
contig1
contig3
contig4
Isotig3
contig1
contig2
contig4
Verify onto genome draft
gDNA contig
c1
c2
c3
c4

Genome Annotation

- Predict ORF (start and stop codons)
- Predcit exon vs introns
- Six-frame translation
- Assign gene function by known biological function of homologues (BLAST)
- Rely heavily on existing biological information
- Pathway clustering
- Require manual inspection and evaluation

Human Genome Projects (diploid)

- 2007: Sanger capillary sequence
- 2008: James Watson (454; 7.5X genome coverage)
- 2008: a Chinese & an African (Illumina; >30X coverage)

-> all found ~ 3M SNPs (3/4 are previously known sites)

nature Vol 456 | 6 November 2008 | doi:10.1038/nature07484

ARTICLES

The diploid genome sequence of an Asian individual

Jun Wang[1,2,3,4]*, Wei Wang[1,3]*, Ruiqiang Li[1,3,4]*, Yingrui Li[1,5,6]*, Geng Tian[1,7], Laurie Goodman[1], Wei Fan[1], Junqing Zhang[1], Jun Li[1], Juanbin Zhang[1], Yiran Guo[1,7], Binxiao Feng[1], Heng Li[1,8], Yao Lu[1], Xiaodong Fang[1], Huiqing Liang[1], Zhenglin Du[1], Dong Li[1], Yiqing Zhao[1,7], Yujie Hu[1,7], Zhenzhen Yang[1], Hancheng Zheng[1], Ines Hellmann[9], Michael Inouye[8], John Pool[9], Xin Yi[1,7], Jing Zhao[1], Jinjie Duan[1], Yan Zhou[1], Junjie Qin[1,7], Lijia Ma[1,7], Guoqing Li[1], Zhentao Yang[1], Guojie Zhang[1,7], Bin Yang[1], Chang Yu[1], Fang Liang[1,7], Wenjie Li[1], Shaochuan Li[1], Dawei Li[1], Peixiang Ni[1], Jue Ruan[1,7], Qibin Li[1,7], Hongmei Zhu[1], Dongyuan Liu[1], Zhike Lu[1], Ning Li[1,7], Guangwu Guo[1,7], Jianguo Zhang[1], Jia Ye[1], Lin Fang[1], Qin Hao[1,7], Quan Chen[1,5], Yu Liang[1,7], Yeyang Su[1,7], A. san[1,7], Cuo Ping[1,7], Shuang Yang[1], Fang Chen[1,7], Li Li[1], Ke Zhou[1], Hongkun Zheng[1,4], Yuanyuan Ren[1], Ling Yang[1], Yang Gao[1,6], Guohua Yang[1,2], Zhuo Li[1], Xiaoli Feng[1], Karsten Kristiansen[4], Gane Ka-Shu Wong[1,10], Rasmus Nielsen[9], Richard Durbin[8], Lars Bolund[1,11], Xiuqing Zhang[1,6], Songgang Li[1,2,5], Huanming Yang[1,2,3] & Jian Wang[1,2,3]

Here we present the first diploid genome sequence of an Asian individual. The genome was sequenced to 36-fold average coverage using massively parallel sequencing technology. We aligned the short reads onto the NCBI human reference genome to 99.97% coverage, and guided by the reference genome, we used uniquely mapped reads to assemble a high-quality consensus sequence for 92% of the Asian individual's genome. We identified approximately 3 million single-nucleotide polymorphisms (SNPs) inside this region, of which 13.6% were not in the dbSNP database. Genotyping analysis showed that SNP identification had high accuracy and consistency, indicating the high sequence quality of this assembly. We also carried out heterozygote phasing and haplotype prediction against HapMap CHB and JPT haplotypes (Chinese and Japanese, respectively), sequence comparison with the two available individual genomes (J. D. Watson and J. C. Venter), and structural variation identification. These variations were considered for their potential biological impact. Our sequence data and analyses demonstrate the potential usefulness of next-generation sequencing technologies for personal genomics.

NGS - a replacement for microarrays? *Not exactly.*

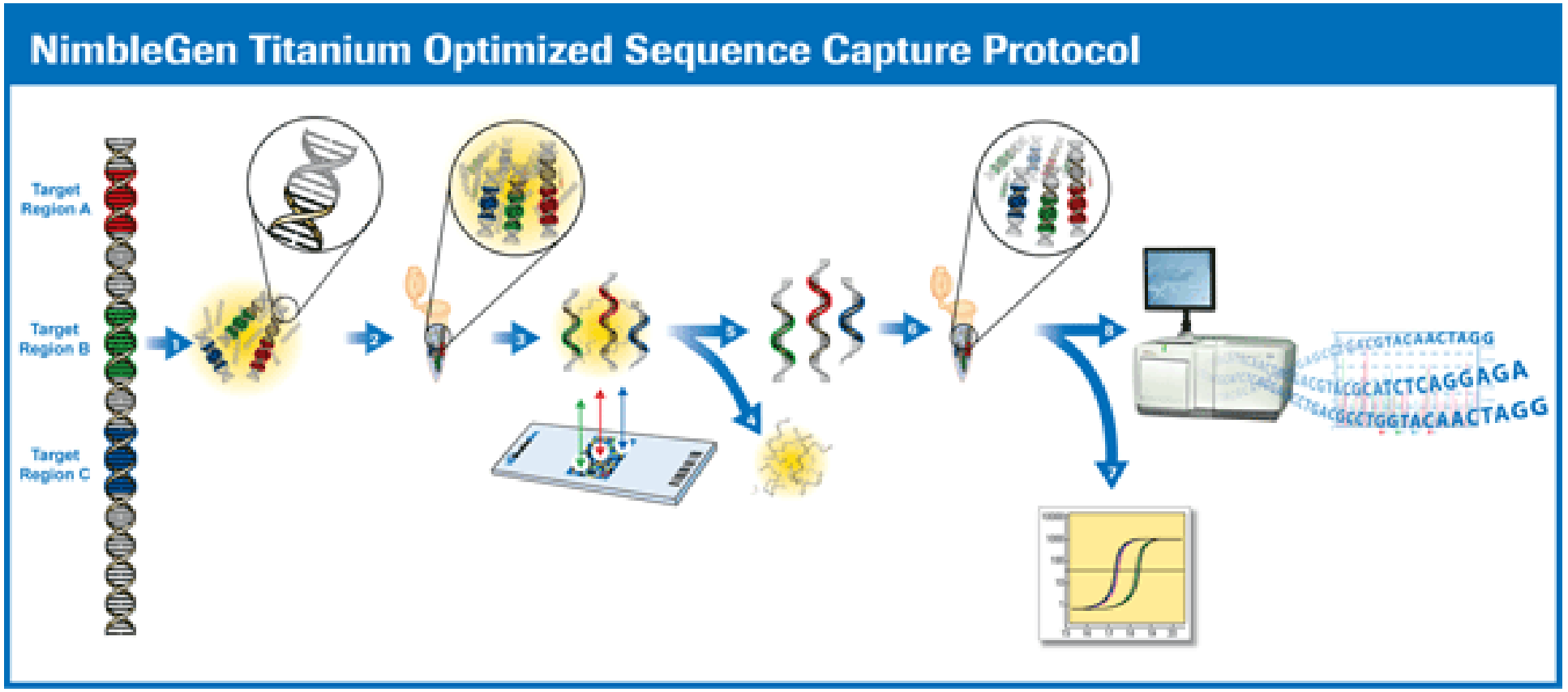
NGS applications:

- Comparative genomics
- ChIP-seq
- Sequencing novel transcripts
- Digital Gene Expression
- RNA-seq:
  - Wider detection dynamic range than arrays
  - Highly reproducible

•screening by array is fast; various chip selections for model organisms

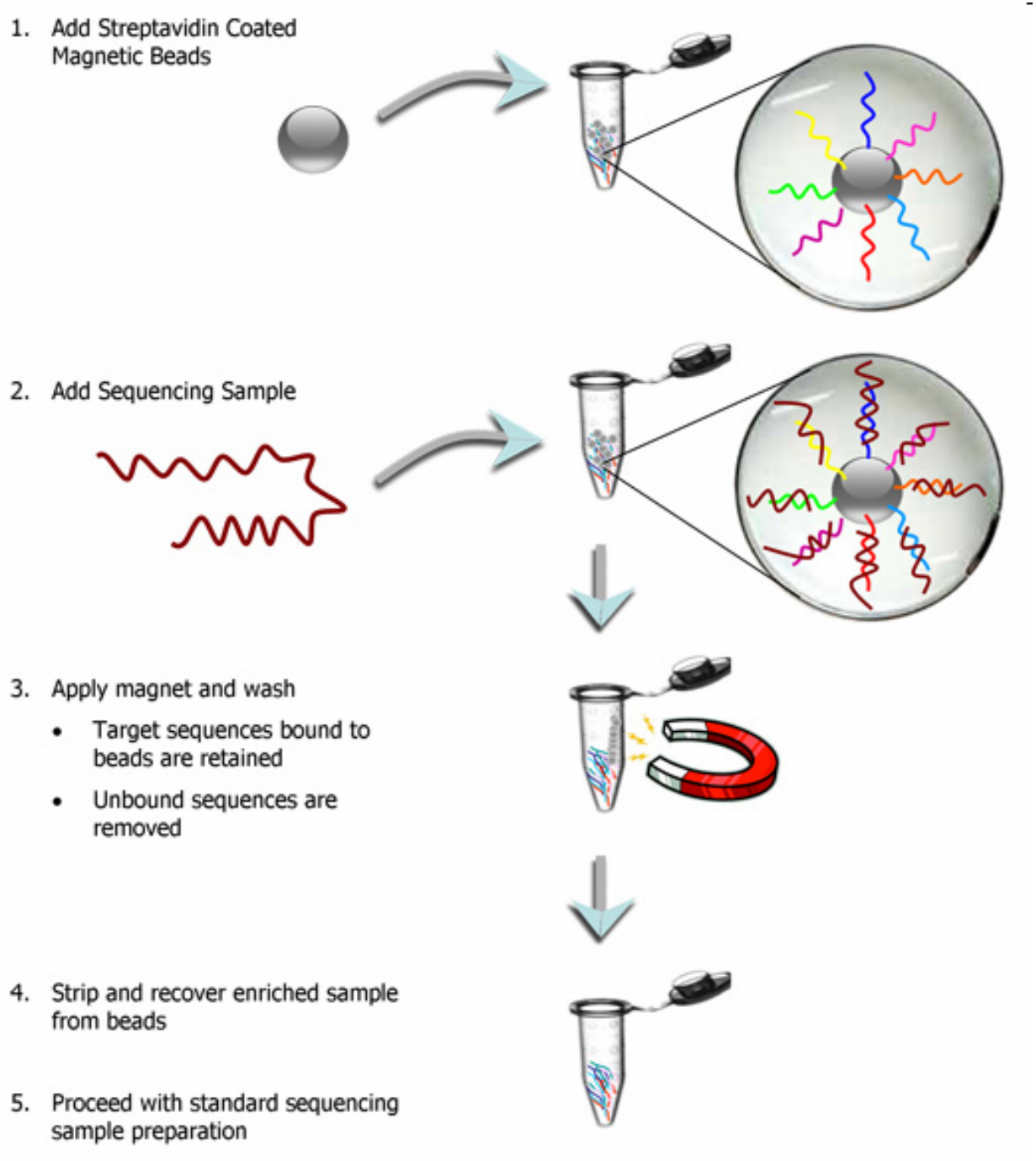
•combine with NGS to enrich fragments of interest for sequencing (Capture-sequencing)

Capture sequencing: array-based



www.nimblegen.com

Capture Sequencing: bead-based



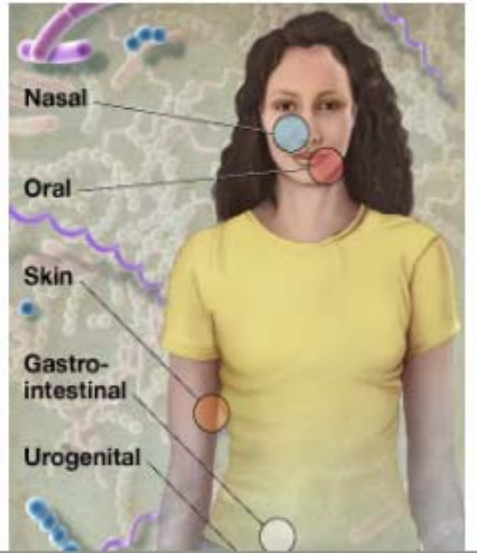
Metagenomics

- “Who’s there?”
- Environmental samples
- Infer relative abundance in the ecosystem
- “human microbiome”
  - NIH “RoadMap” medical research

Human Microbiome Project (HMP)

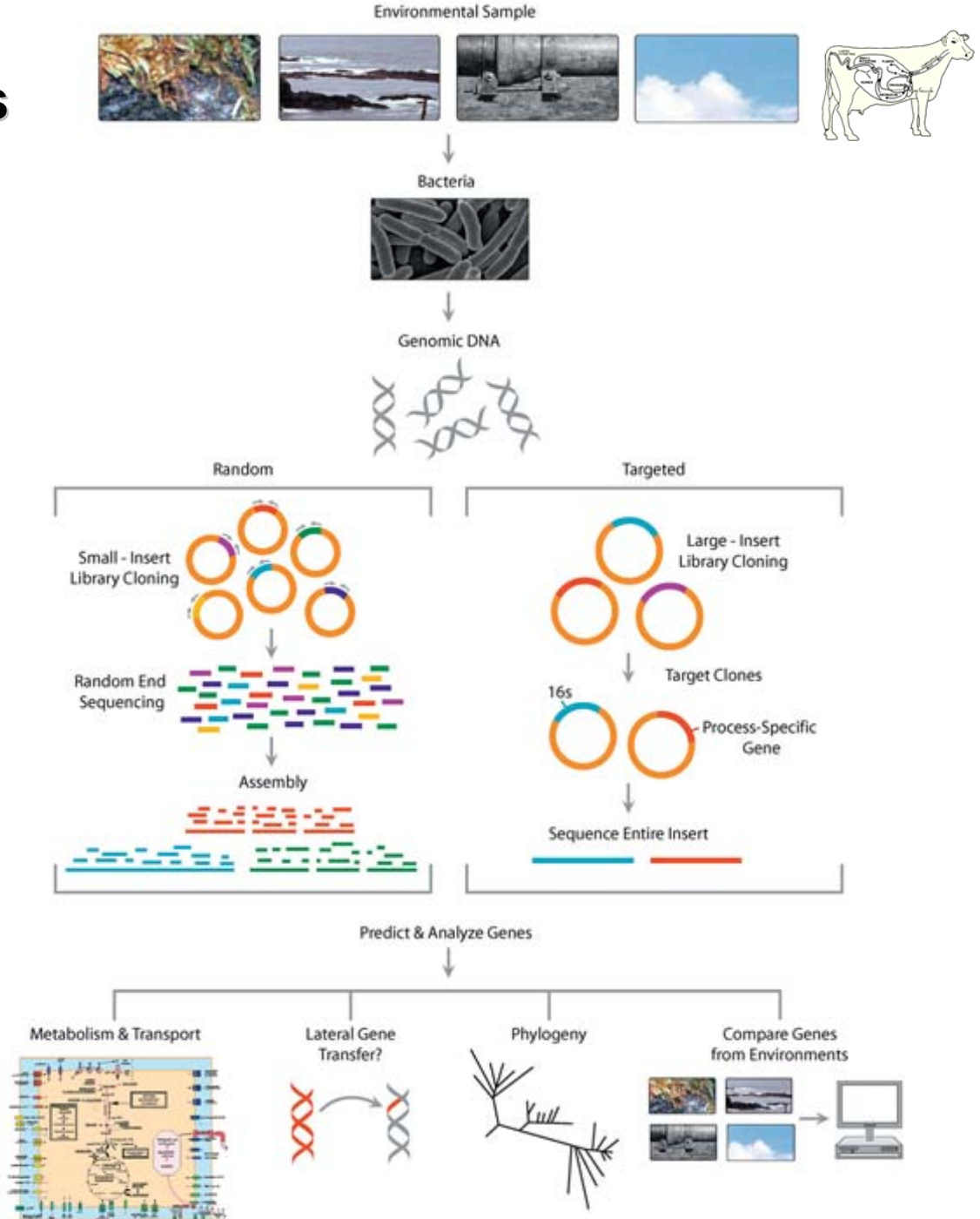
OVERVIEW

Within the body of a healthy adult, microbial cells are estimated to outnumber human cells by a factor of ten to one. These communities, however, remain largely unstudied, leaving almost entirely unknown their influence upon human development, physiology, immunity, and nutrition. To take advantage of recent technological advances and to develop new ones, the NIH Roadmap has initiated the Human Microbiome Project (HMP) with the mission of generating resources enabling comprehensive characterization of the human microbiota and analysis of its role in human health and disease.



Traditional microbiology has focused on the study of individual species as isolated units. However many, if not most, have never been successfully isolated as viable specimens for analysis, presumably because their growth is dependant upon a specific microenvironment that has not been, or cannot be, reproduced experimentally. Among those species that have been isolated, analyses of genetic makeup, gene expression patterns, and metabolic physiologies have rarely extended to inter-species interactions or microbe-host interactions. Advances in DNA sequencing technologies have created a new field of research, called metagenomics, allowing comprehensive examination of microbial communities, even those comprised of uncultivable organisms. Instead of examining the genome of an individual

Metagenomics



camera.calit2.net/education/what-is-metagenomics

Challenges and future directions

- Ever-increasing demand for computation resource
  - Data management (analysis & archiving)
- New sequencing technologies

  (greater data output & longer read length)
- Requires rapid evolution in software modules and pipelines
- Build up open-source community for the tools

International consortium announces the 1000 Genomes Project.

Any two people have 99% identical DNA.

http://www.wired.com/wiredscience/2008/01/the-genome-in-h/

Public release date: 22-Jan-2008

Genome sequences are important

1001 Genomes

A Catalog of *Arabidopsis thaliana* Genetic Variation

Home Collaborators Accessions Tools Downloads Data Center Gallery About
Help desk

Welcome to the 1001 Genomes Project

The 1001 Genomes Vision

The 1001 Genomes Project, launched at the beginning of 2008, has a simple goal: to discover the whole-genome sequence variation in 1001 strains (accessions) of the reference plant Arabidopsis thaliana. The resulting information will pave the way for a new era of genetics that combines large-scale association studies in wild strains with forward genetic analyses in experimental crosses, in order to identify alleles underpinning phenotypic diversity across the entire genome and the entire species. The analyses enabled by this project will have broad implications for areas as diverse as evolutionary sciences, plant breeding and human genetics.

This 1001 Genomes Project is particularly timely because the current technological revolution in sequencing means that it is now feasible to resequence large numbers of genomes. Indeed, a **1000 Genomes** project for humans has just been launched in early 2008 as well. There are, however, several important differences between the two projects. The most important one is that each of the accessions in the Arabidopsis 1001 Genomes project is an inbred line with seeds that will be freely available from the stock centre to all our colleagues. Unlimited numbers of plants with identical genotype can be grown and phenotyped for each accession, in as many environments as desired, and so the sequence information we collect can be used directly in association studies at biochemical, metabolic, physiological, morphological, and whole plant-fitness levels.

As of earl 2010, the complete genome sequences of over 80 accessions have already been released by the Max Planck Institute. There are commitments for the remaining accessions, primarily from the Salk Institute, the Gregor Mendel Institute and Monsanto, and we are hoping for completion of the 1001 Genomes project in the first half of 2011.

Progress as of June 2, 2010:

Commitments:	1001
Sequencing underway:	99
Finished genomes:	157
Released genomes:	91

Links

NCBI SRA Genomes Project

Map resource for 1001 Genomes

News

Mai 3, 2010

A collection of 80 *A. thaliana* accessions sequenced as part of the 1001 Genomes Project is available from ABRC. Each of the accessions is an inbred line that can be ordered as an individual line or as a set (**CS76427**). These stocks can be found using the **ABRC catalog.**

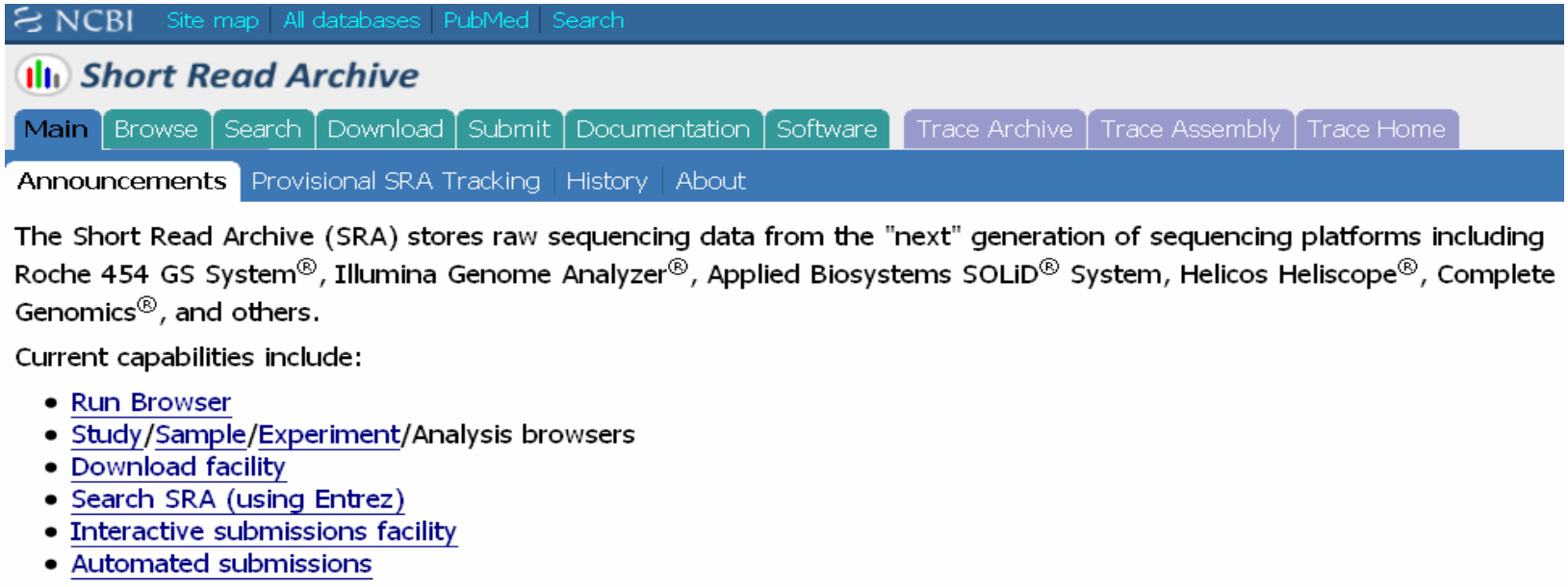
February 2, 2010

The Weigel laboratory has just released 80 *Arabidopsis thaliana* genomes sequenced with paired end Illumina short reads. SNPs and structure variants (SVs) **are now available online.**
For more details, please read the **README** file.

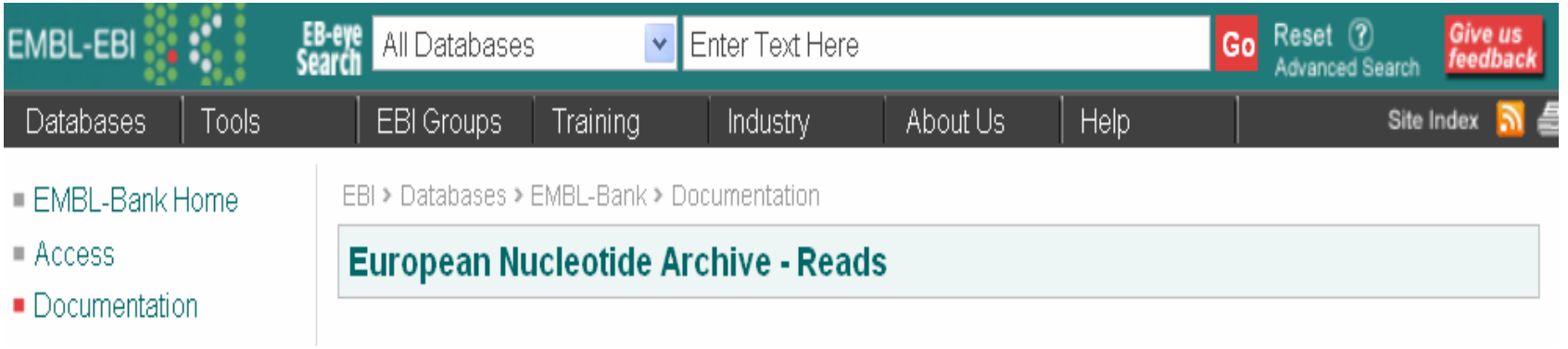
>> News archive...

Public NGS Databases

NIH Short Read Archive

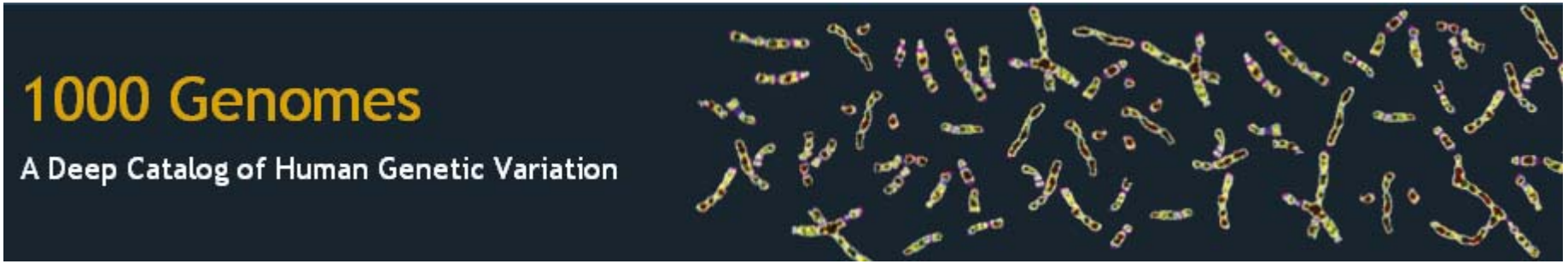


ERA (European Short Read Archive)




wellcome trust
sanger
institute
RSS
Human Genetics
Model Organisms
Pathogens
Bioinformatics
Sequencing
Scientific Divisions
Model Organisms
All | Mouse | Zebrafish | Worm | Yeast | Xenopus


MAMMOTH GENOME PROJECT
PENNSYLVANIA STATE UNIVERSITY
2008
Mammoth Genome Project
Pennsylvania State University


1000 Genomes
A Deep Catalog of Human Genetic Variation

Picture from google image

Solexa vedio

- http://www.illumina.com/media.ilmn?Title=Sequencing-Workflow-Video&Cap=&Img=spacer.gif&PageName=illumina%20sequencing%20technology&PageURL=203&Media=10

AB SOLiD video

http://marketing.appliedbiosystems.com/images/Product/Solid_Knowledge/flash/102207/solid.html