高資料量篩檢技術簡介
An Introdcution of High- throughput Methods and their Applicatons

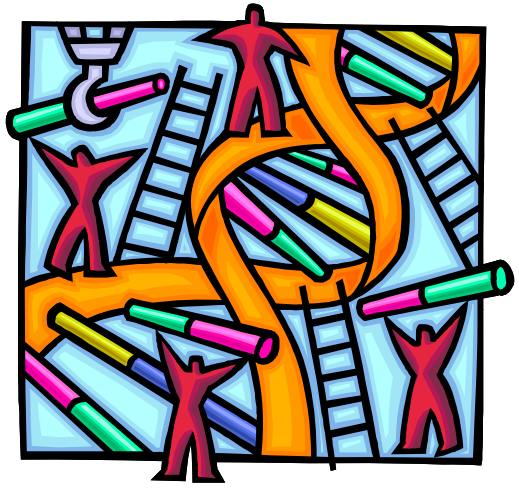
Chen, Shu-Hwa
Institute of Information Science
Academia Sinica
2009.07

Bioinformatics

- Bioinformatics is the application of information technology to the management and analysis of biological data.
- Bioinformatics is an interdisciplinary research area that is the interface between the biological and computational sciences.
- **Bioinformatics** is the field of science in which biology, computer science, and information technology merge to form a single discipline.

Bioinformatics

Biology + Informatics + Statistics



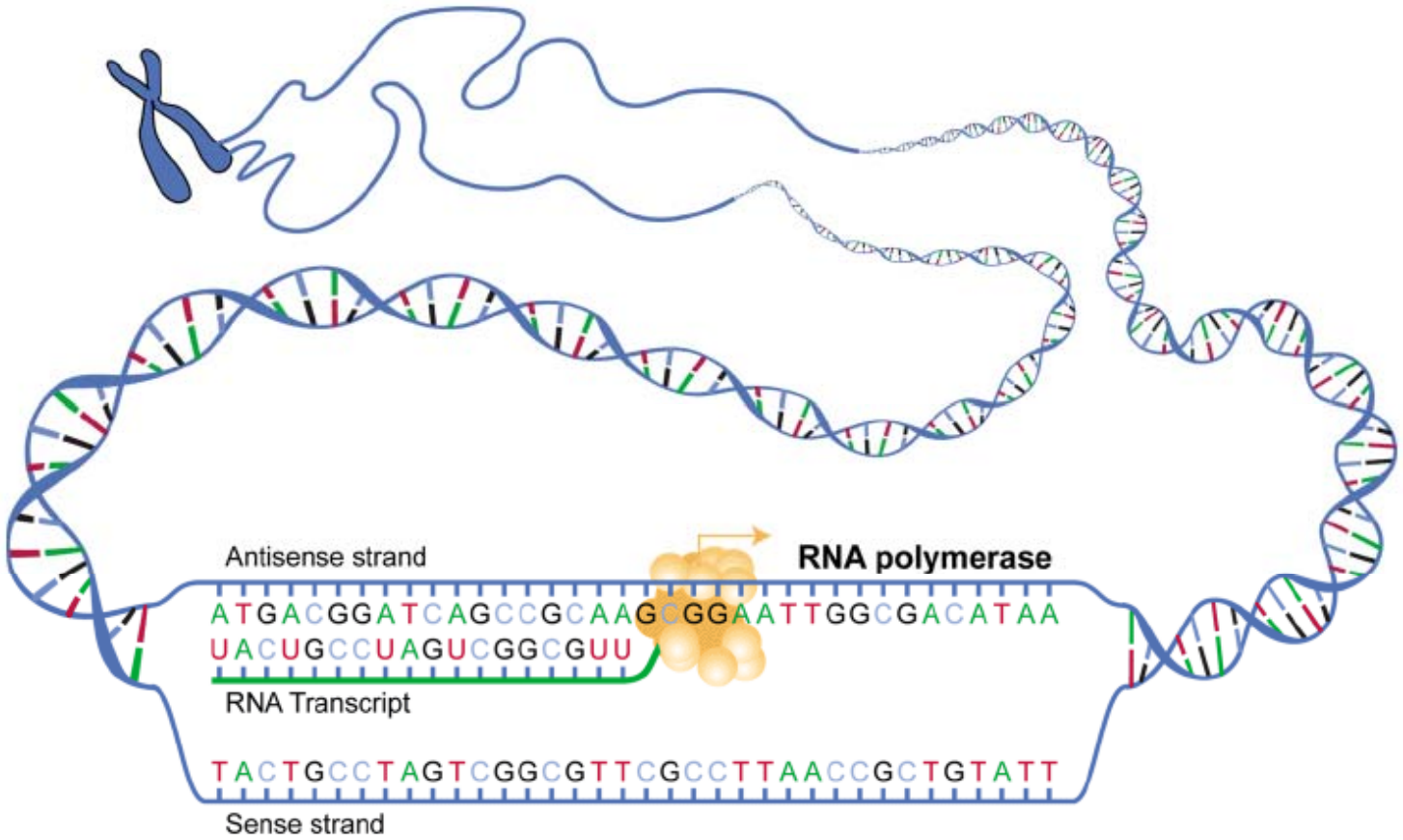
Biological information is coded

Binary Code

```
01000101010001110101 0
10001010100011101010 1
00010
```

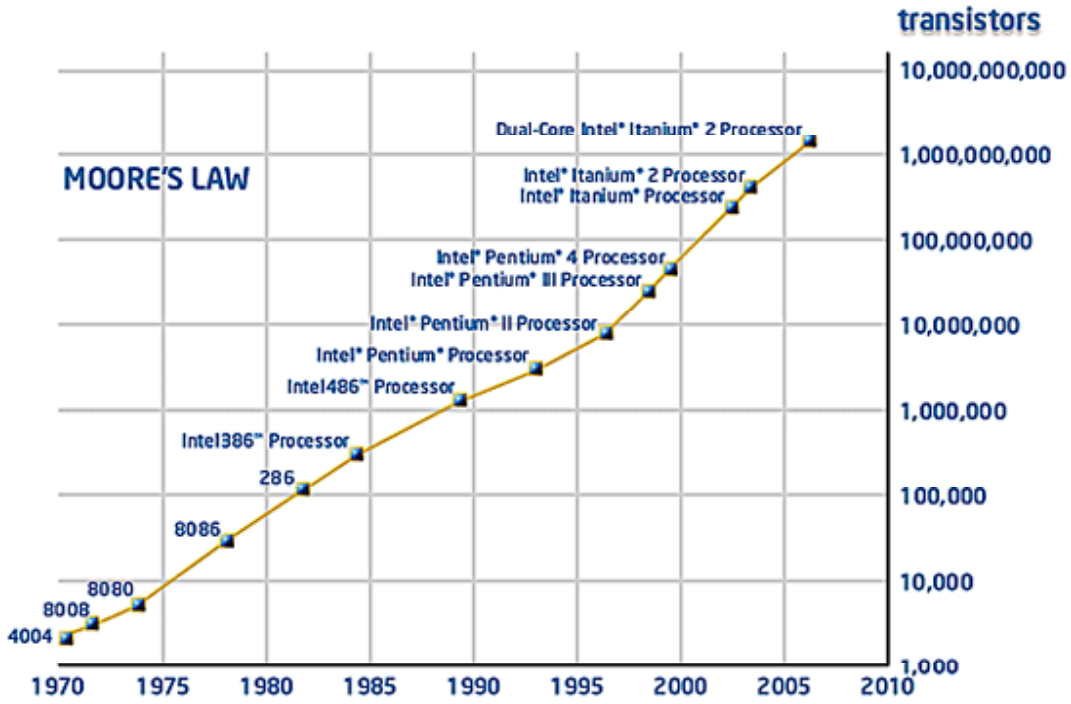
Genetic Code

```
ATTCCATCGGAGTAATTCCATC
GGAGTAATTCCATCGGAGTAAT
```

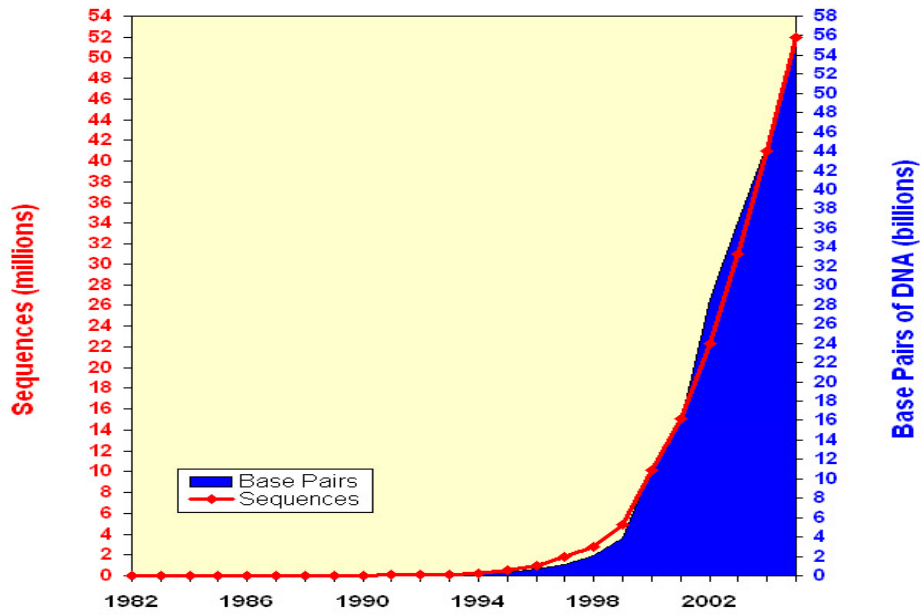


Rising of Bioinformatics

- Internet and WWW
- Computing Power
- Genome Projects
- High-Throughput Technology



Growth of GenBank(1982-2005)

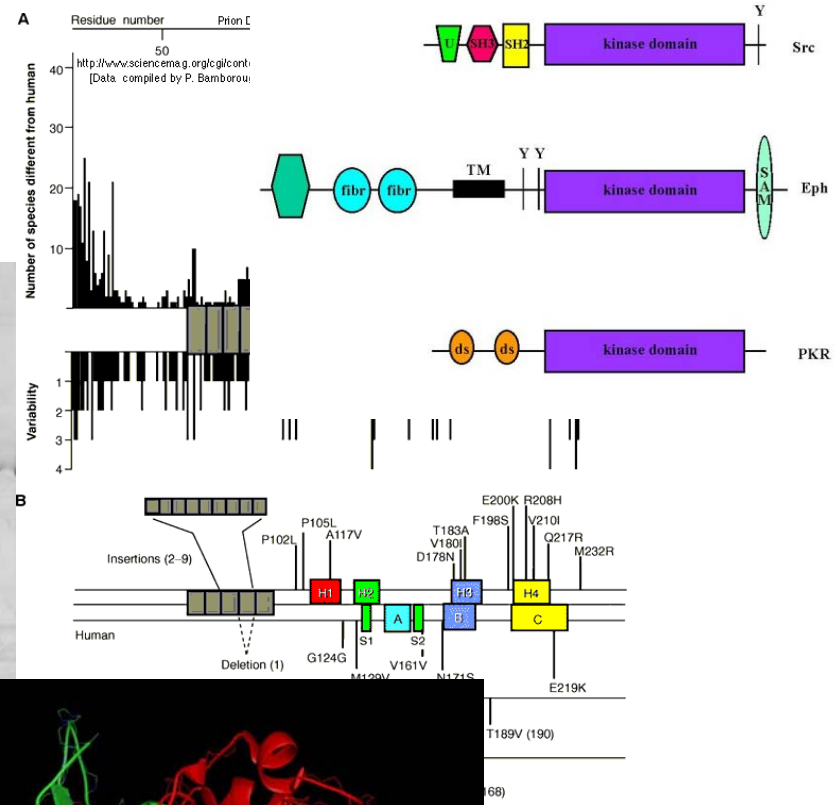


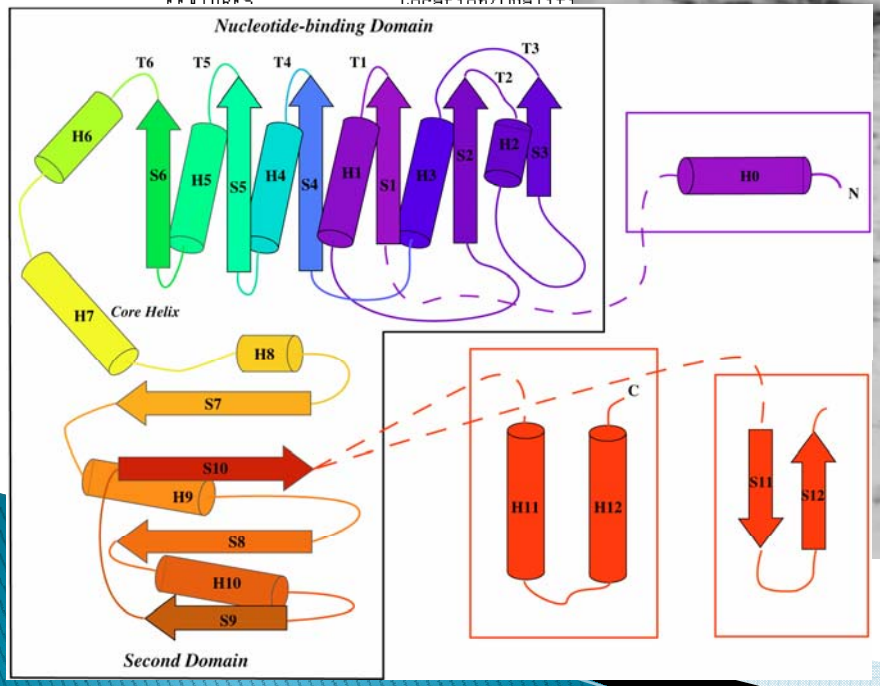
http://www.intel.com/technology/mooreslaw/index.htm

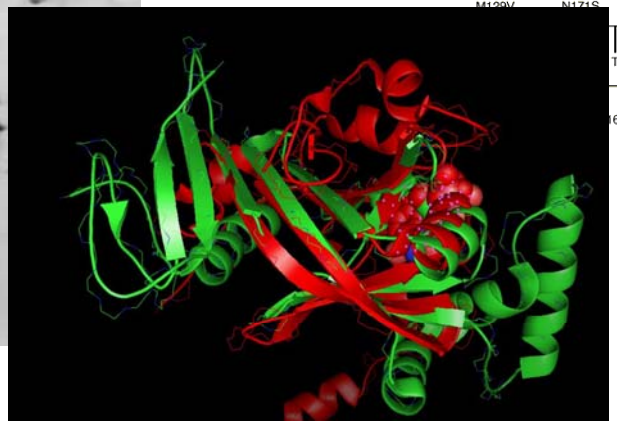
http://www.ncbi.nlm.nih.gov/Genbank/genbankstats.html

Various Formats for BioDB

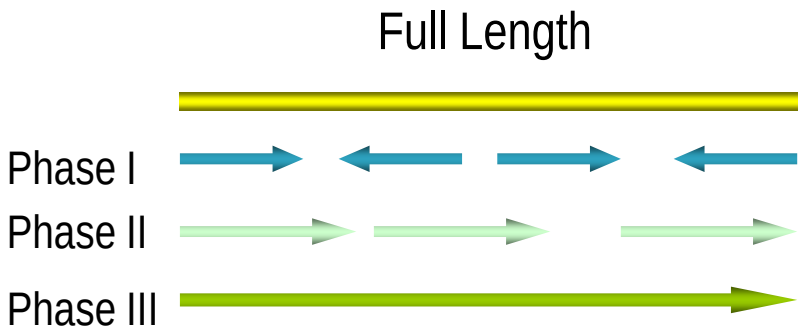
```
LOCUS       eIF4E                   2881 bp    DNA     linear   INV 27-OCT-2005
DEFINITION  Drosophila melanogaster eukaryotic initiation factor 4E (eIF4E)
            gene, alternative splice products, complete cds.
ACCESSION
VERSION
KEYWORDS    .
SOURCE      Drosophila melanogaster (fruit fly)
  ORGANISM  Drosophila melanogaster
            Eukaryota; Metazoa; Arthropoda; Hexapoda; Insecta; Pterygota;
            Neoptera; Endopterygota; Diptera; Brachycera; Muscomorpha;
            Ephydroidea; Drosophilidae; Drosophila.
REFERENCE   1  (bases 1 to 2881)
  AUTHORS   Burnett,F.M., van der Waals J.D. and Szent-Gyorgi A.
  TITLE     Environmental influences
            repeats in several specie
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 2881)
  AUTHORS   Burnett,F.M., van der Wa
  TITLE     Direct Submission
  JOURNAL   Submitted (27-OCT-2005)
            University, 1859 Tennis
FEATURES             Location/Qualifi
```



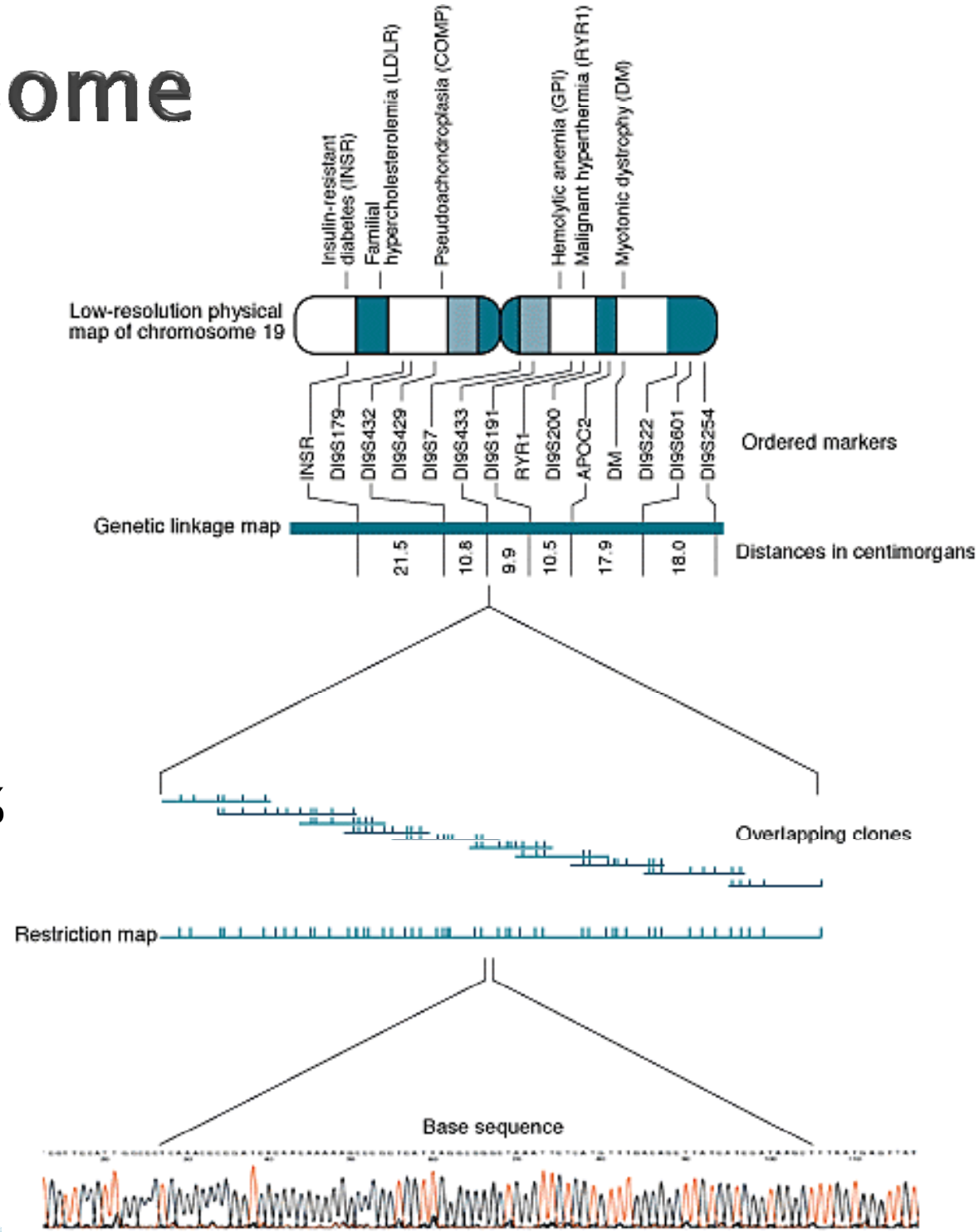




From Chromosome to Sequences

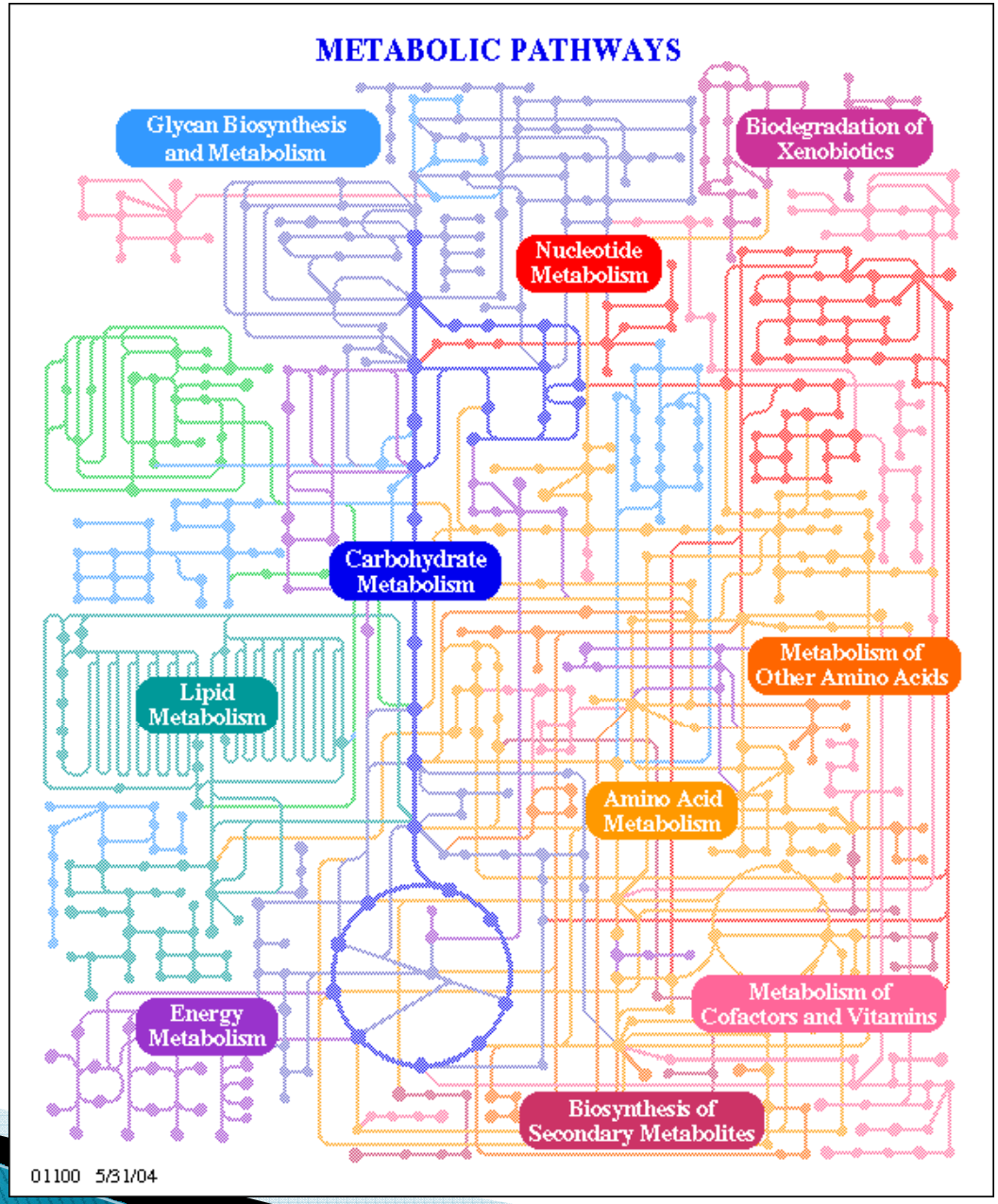


Coverage (8-10X)>> 99%
Error Rate<<0.01%

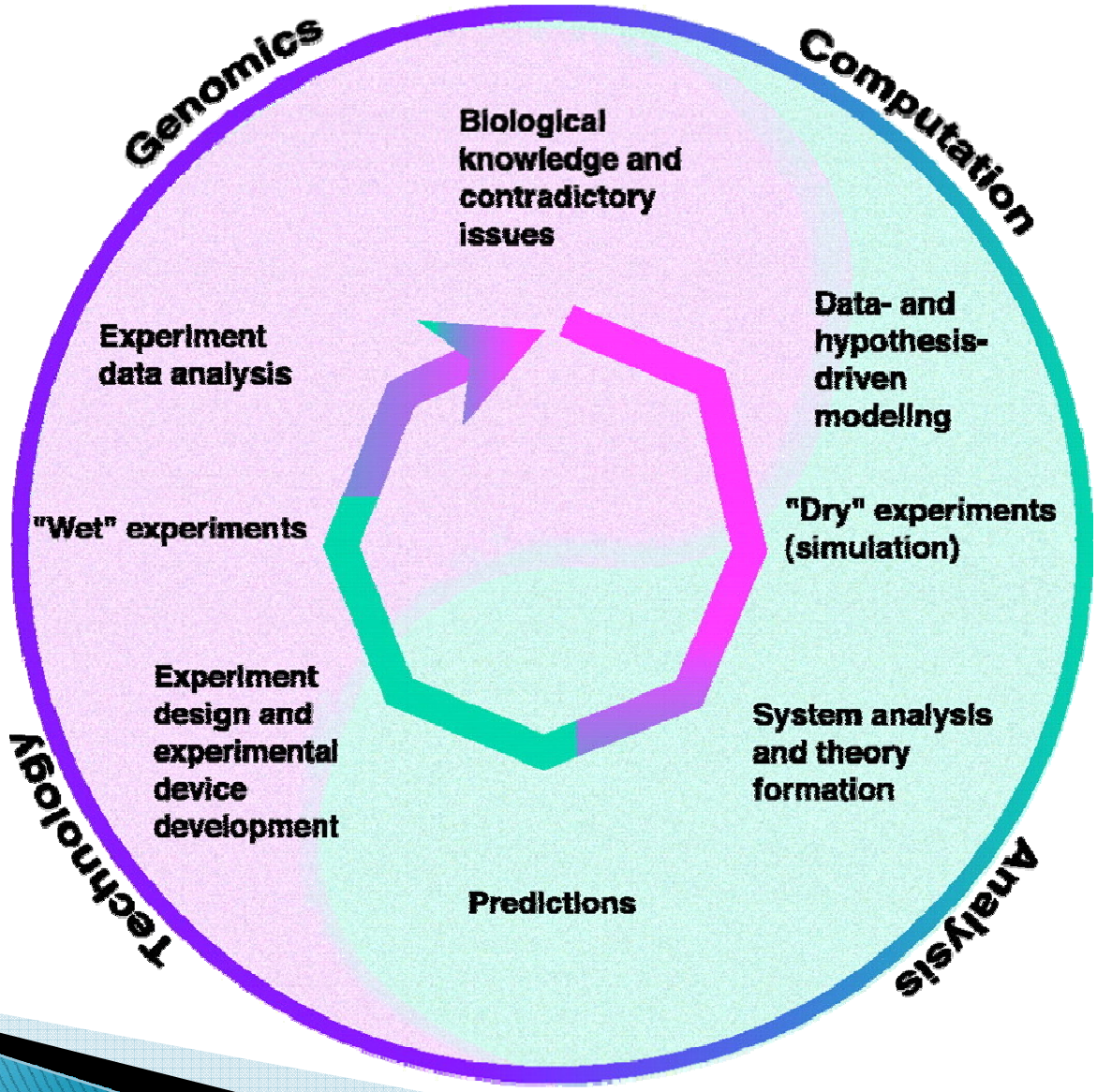


KEGG

Kyoto Encyclopedia of Genes and Genomes



Systems Biology



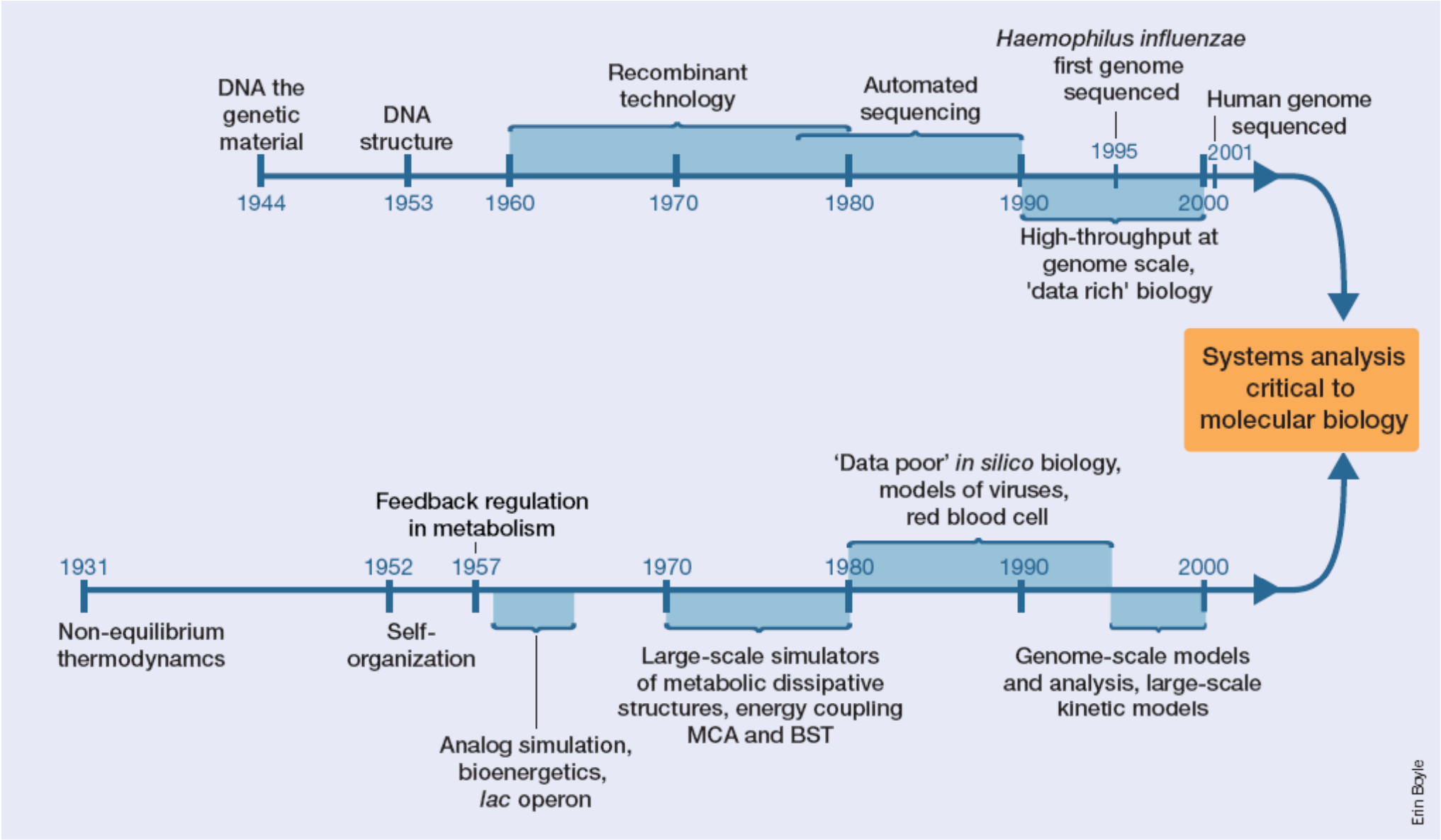


Figure 1 Two lines of inquiry led from the approximate onset of molecular biological thinking to present-day systems biology. The top timeline represents the root of systems biology in mainstream molecular biology, with its emphasis on individual macromolecules. Scaled-up versions of this effort then induced systems biology as a way to look at all those molecules simultaneously, and consider their interactions. The lower timeline represents the lesser-known effort that constantly focused on the formal analysis of new functional states that arise when multiple molecules interact simultaneously.

http://www.nature.com/nbt/journal/v22/n10/full/nbt1020.html

High-Throughput Technology

- Using robotics, data processing and control software, liquid handling devices, and sensitive detectors to quickly conduct millions of biochemical, genetic or pharmacological tests.

AUTOMATION

Brute-Force Approach ➔ Fast and Huge Amount of Data

High-Throughput Technology

- Genomics
  - HTP Sequencing
- Transcriptomics
  - Microarray for gene expression data
- Proteomics
  - Y2H
  - Isotope-Coded Affinity Tags
- Anatomical and Histological Images

High-throughput *omics

Subject	Inclusive set (For an individual)	Statistical Study (many individuals)
Genes	Genome	Genomics
Transcripts	Transcriptome	Transcriptomics
Proteins	Proteome	Proteomics
Metabolites	Metabolome	Metabolomics
Phenotype	Phenome	Phenomics

Basics of DNA Sequencing Method

Sanger Method

Sequencing Method based on DNA polymerase and chain terminator

Single strand template

↓

Primer Annealing

↓

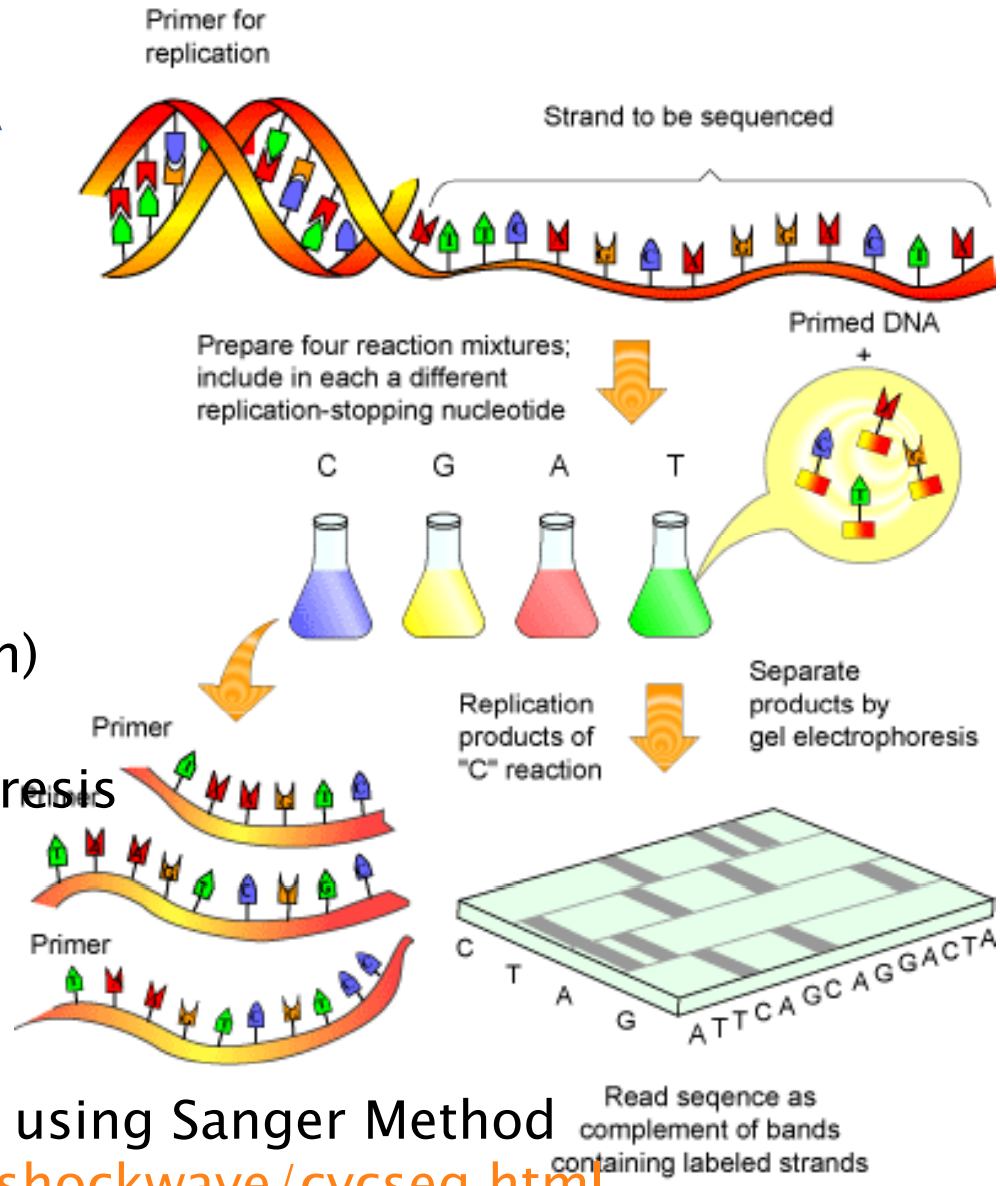
Primer extension by polymerase/ incorporation of ddNTP (stop extension)

↓

Polymerase products sorted by electrophoresis or other equivalent technique

↓

Sequence resolved



Animation of automated DNA sequencing using Sanger Method

http://www.dnalc.org/ddnalc/resources/shockwave/cycseq.html

Next Generation Sequencing

- De Novo Sequencing
  - Genome project, Metagenomics
- Resequencing/Transcriptome sequencing by next-generation technologies
  - Gene expression profiling using novel and revisited sequence census methods
  - Small noncoding RNA profiling and the discovery of novel small RNA genes
  - Protein coding gene annotation using transcriptome sequence data
  - Detection of aberrant transcription events
- Applications of next-generation sequencing for the analysis of epigenetic modifications of histones and DNA
  - DNA methylation profiling by bisulfite DNA sequencing
  - Sequence census applications for mapping histone modifications and the locations of DNA-binding proteins
  - Applications of next-generation sequencers to the study of DNA accessibility and chromatin structure

Reference: Annual Review of Genomics and Human Genetics
Vol. 9: 387-402 (Volume publication date September 2008)
(doi:10.1146/annurev.genom.9.081307.164359)

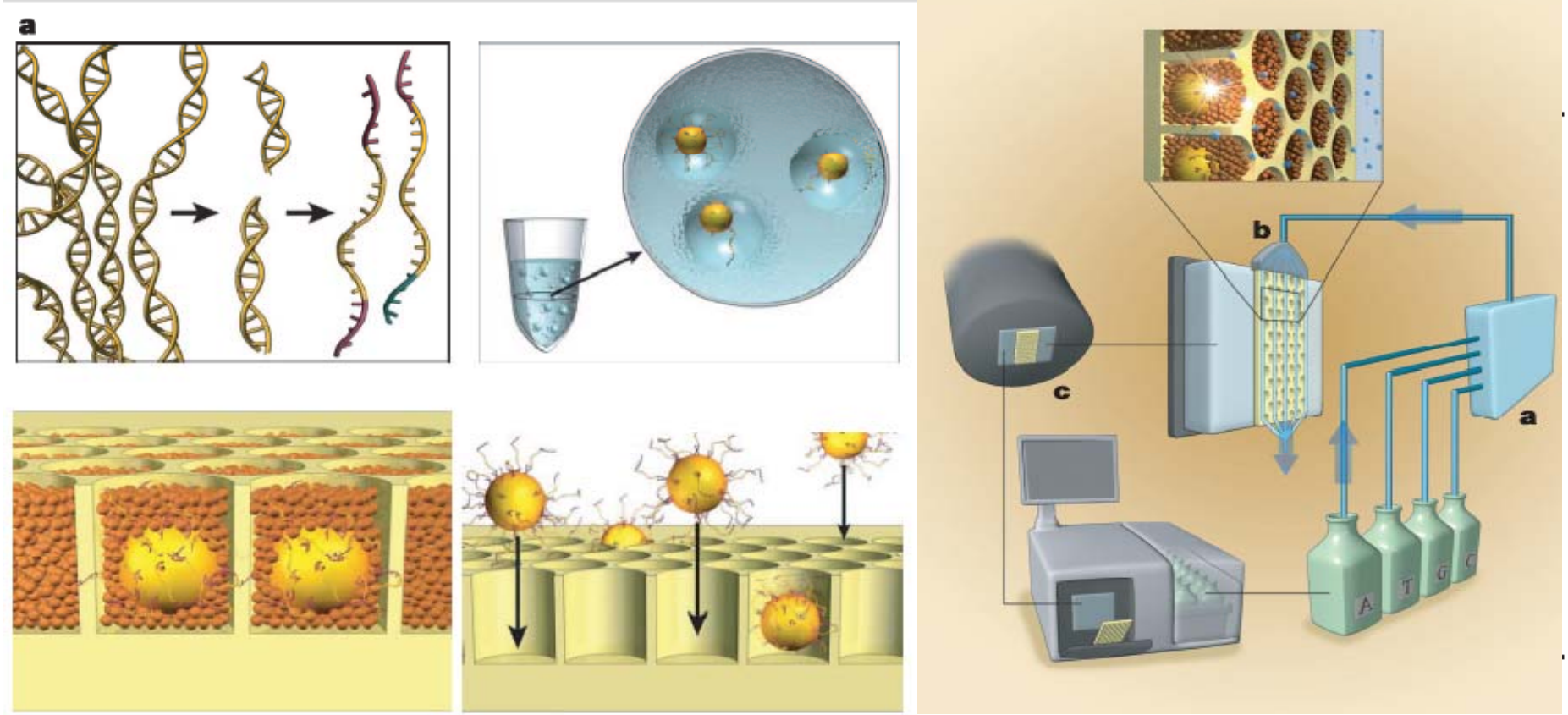
Major Players in NGS

- Three platforms for massively parallel DNA sequencing read production are in reasonably widespread use at present:
- the Roche/454 FLX
- the Illumina/Solexa Genome Analyzer
- the Applied Biosystems SOLiD™ System
- Recently, another two massively parallel systems were announced: the Helicos Heliscope™ (www.helicosbio.com) and Pacific Biosciences SMRT (www.pacificbiosciences.com) instruments.

DNA Sequencing and Resequencing

454 Sequencing Technology:

emusion-based template/ picoliter reaction well/ pyrophosphate sequencing



Genome sequencing in microfabricated high-density picolitre reactors
Nature. 2005 Sep 15; 437(7057):326-7

454, Next Generation Sequencer (NGS)

Table 1 | Summary of sequencing statistics for test fragments

Size of fibre-optic slide	60 × 60 mm^2
Run time/number of cycles	243 min/42
Test fragment reads	497,893
Average read length (bases)	108
Number of bases in test fragments	53,705,267
Bases with a Phred score of 20 and above	47,181,792
Individual read insertion error rate	0.44%
Individual read deletion error rate	0.15%
Individual read substitution error rate	0.004%
All errors	0.60%

Genome Sequencer FLX System Workflow

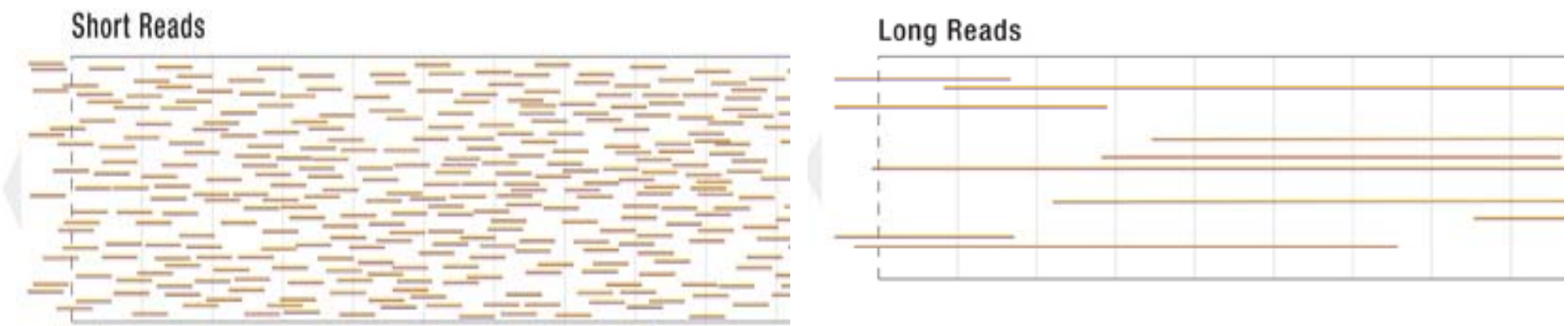
400 bases, Now
~800 bases in near future

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

Sanger Method and NGS

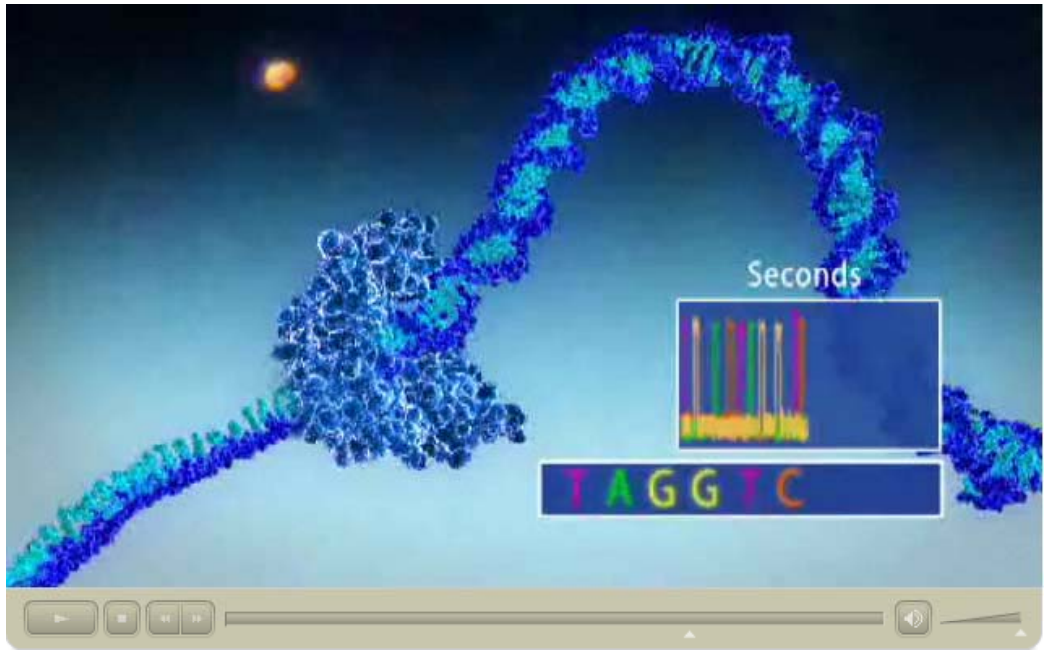
Advances in DNA sequencing technologies

Technology	Approach	Read length	Bp per run
Automated Sanger sequencer ABI3730xl	Synthesis in the presence of dye terminators	Up to 900 bp	96 kb
454/Roche FLX system	Pyrosequencing on solid support	200–300 bp	80–120 Mb
Illumina/Solexa	Sequencing by synthesis with reversible terminators	30–40 bp	1 Gb
ABI/SOLiD	Massively parallel sequencing by ligation	35 bp	1–3 Gb

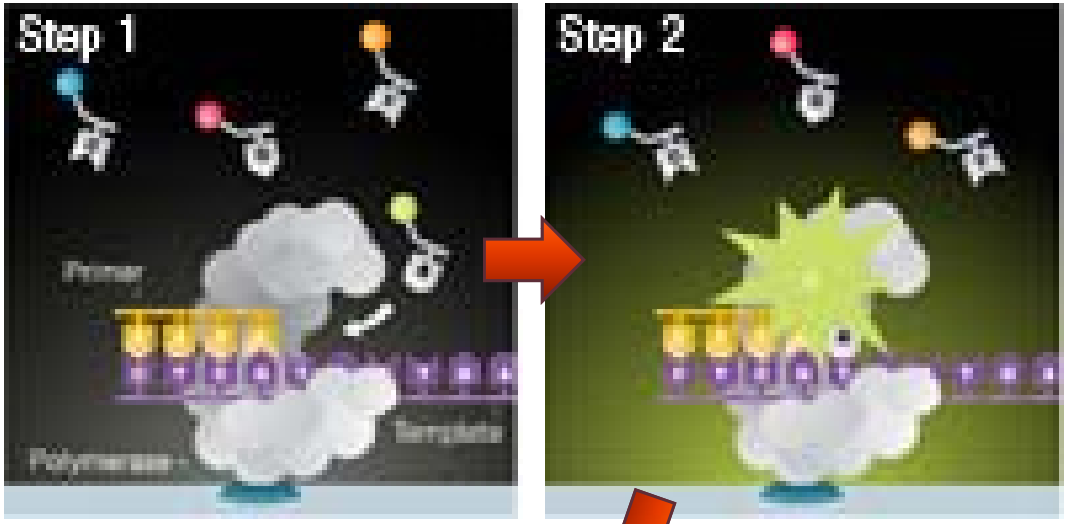


Pacific Sequencer

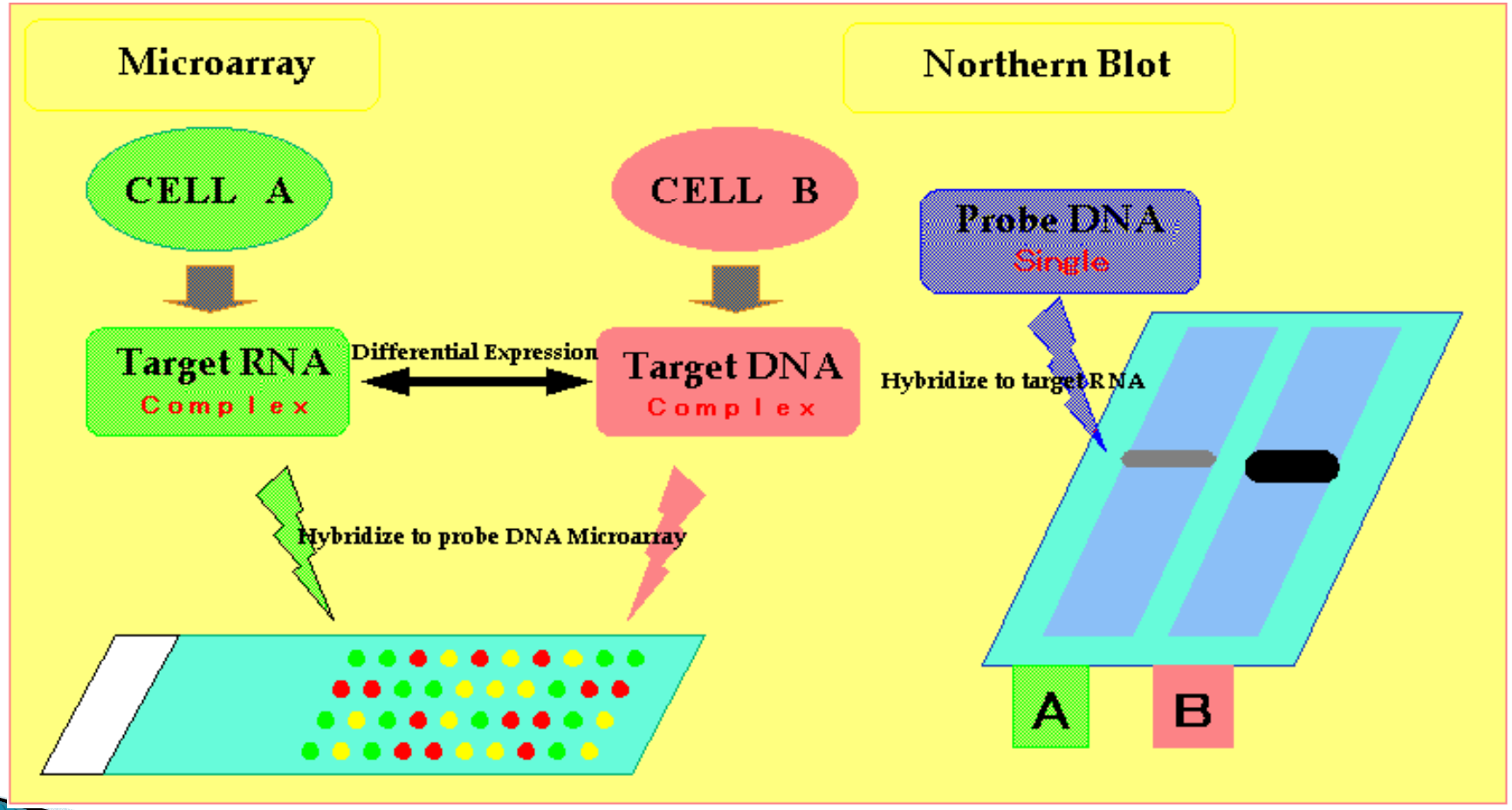
Single Molecule Real Time (SMRT)
DNA sequencing technology,~ 1Kb !!



http://www.pacificbiosciences.com/video_lg.html

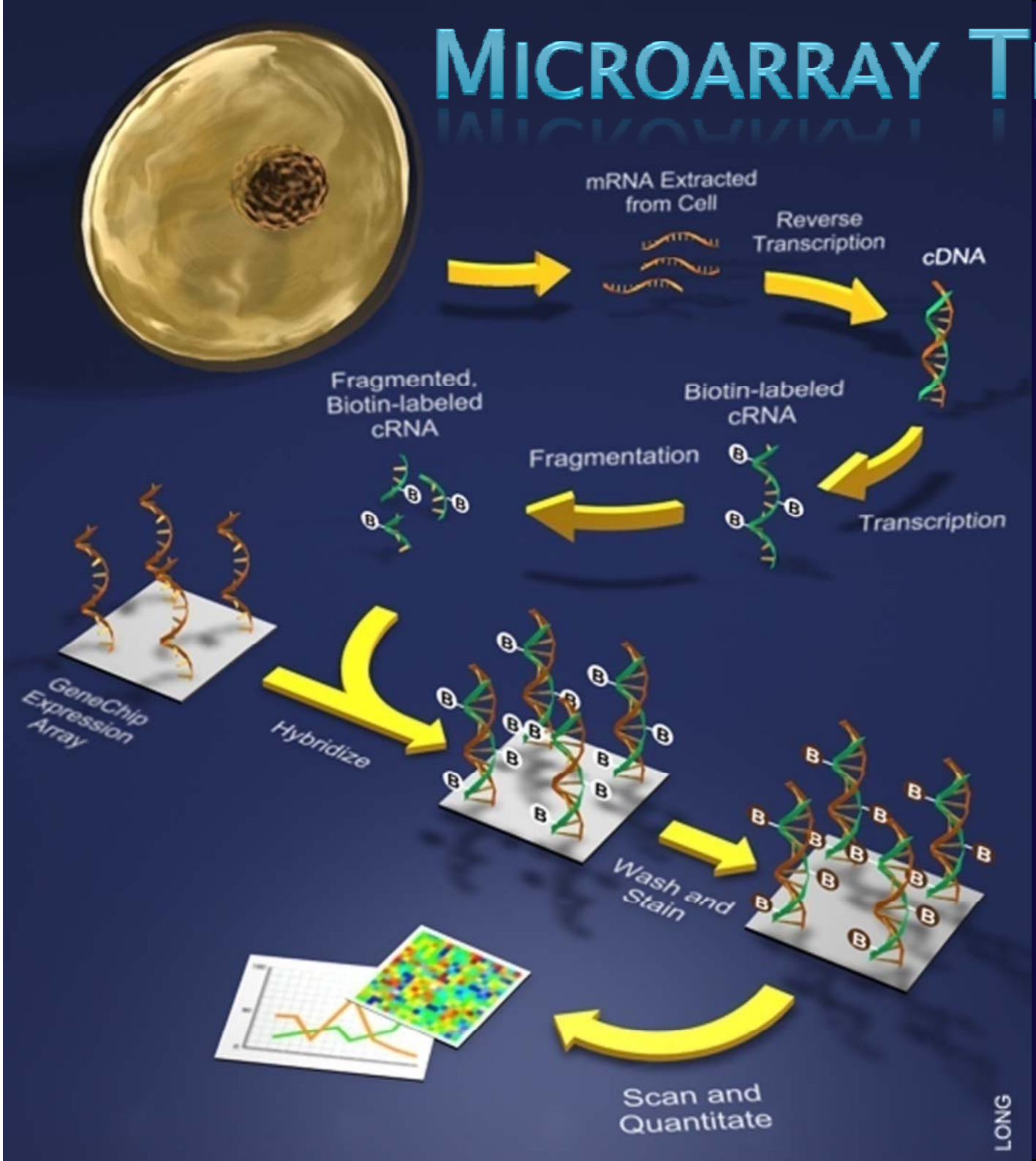


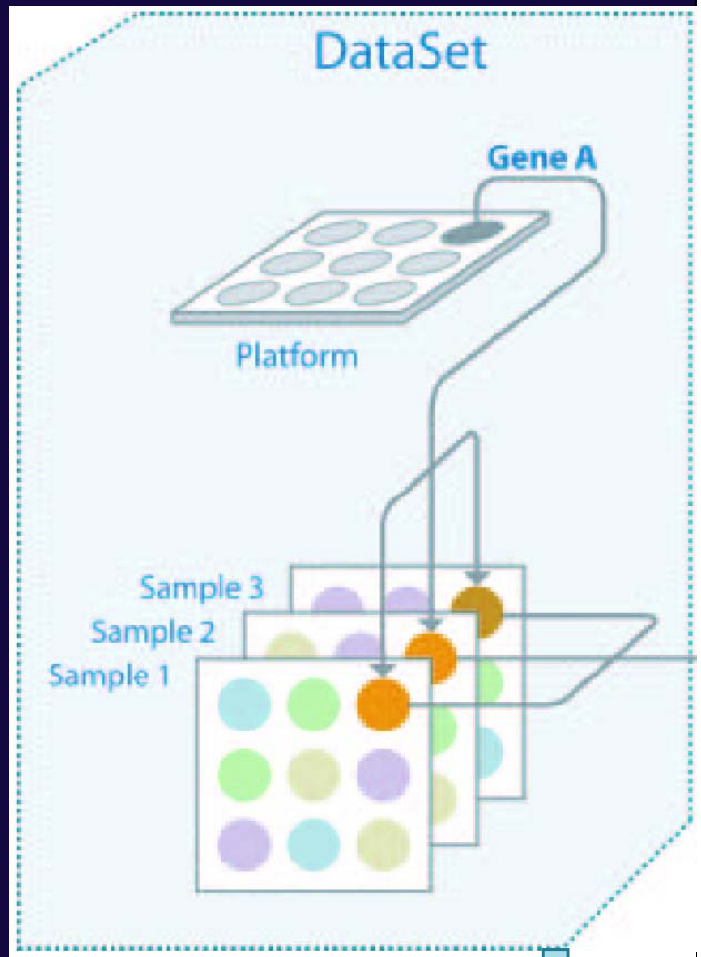
Microarray vs. Northern Blot

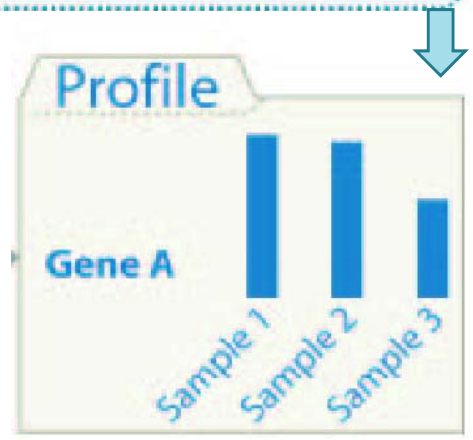


http://cdna01.dna.affrc.go.jp/RMOS/northern_comp_en.html

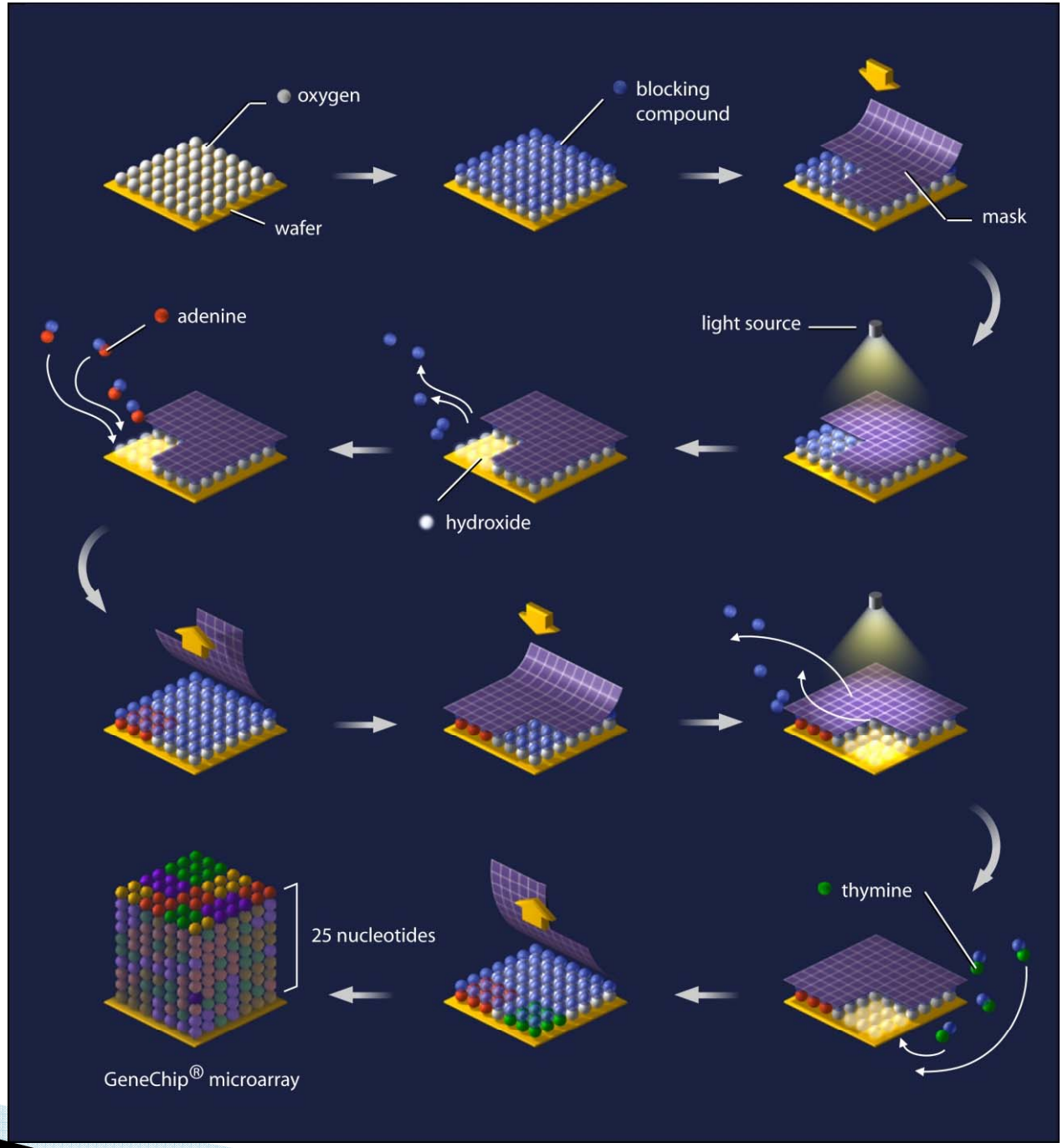
Microarray Technology



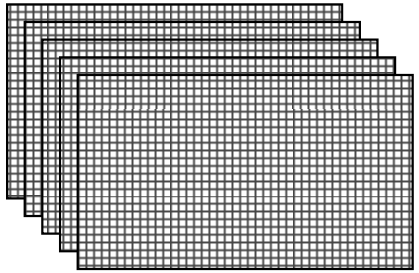




Microarray Chip, made by photo-etching Technology



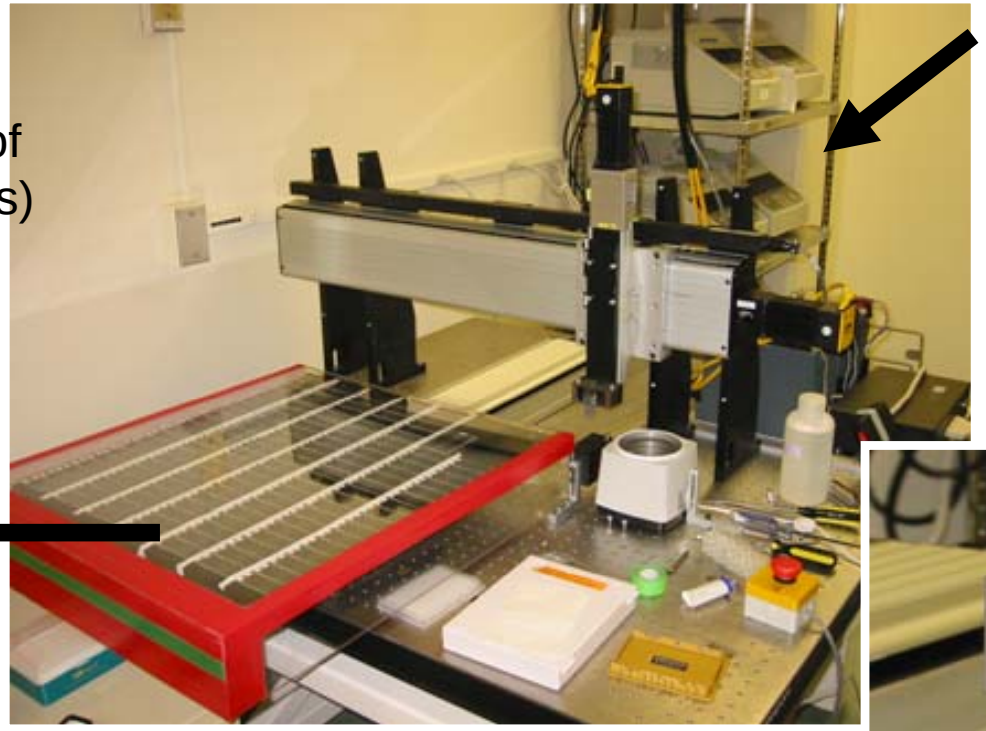
Spotting a Chip

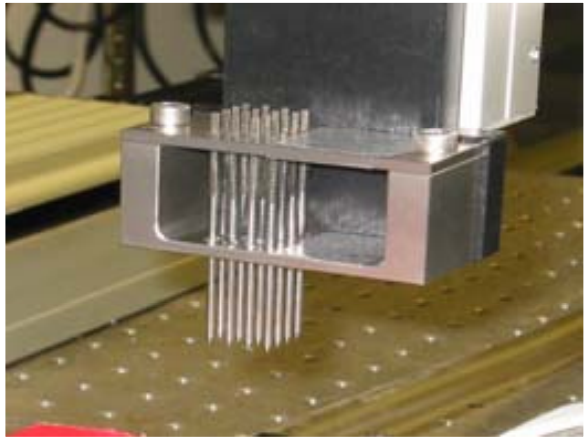


Arrayed Library
(96 or 384-well plates of bacterial glycerol stocks)

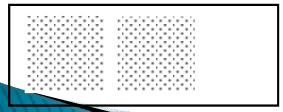
PCR amplification
Directly from colonies with specific primers in 96-well plates

Consolidate into 384-well plates

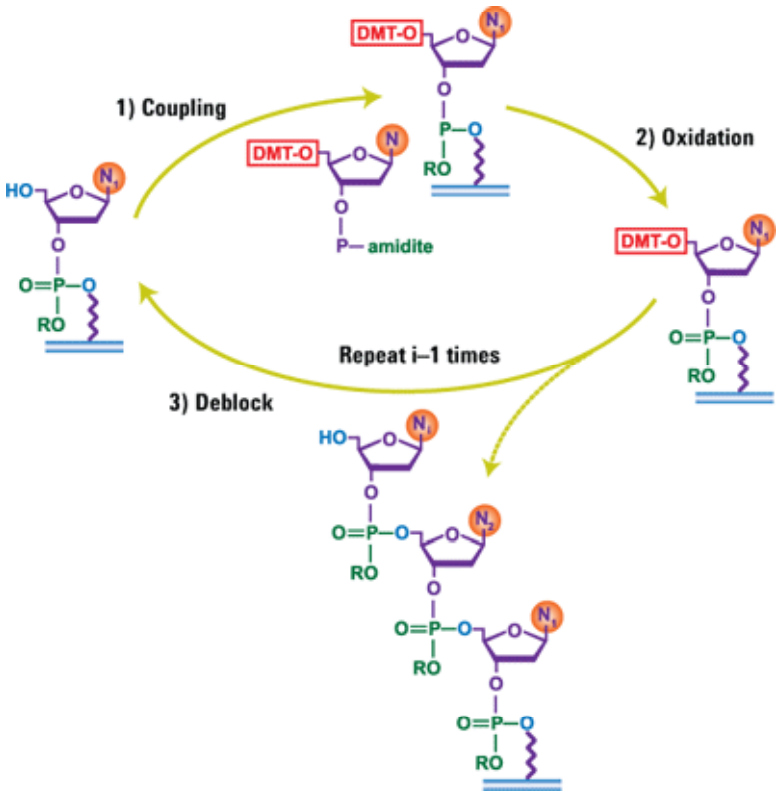


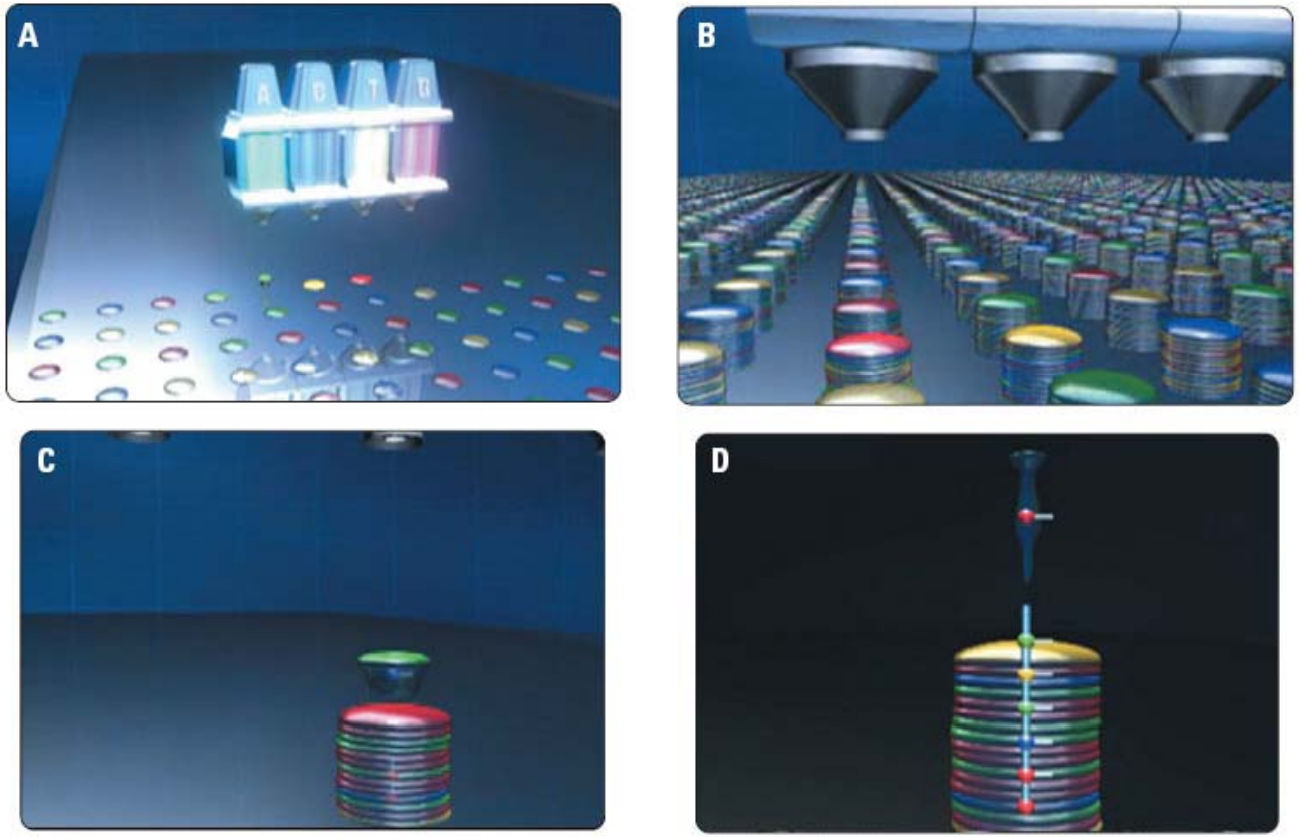


Spot as microarray on glass slides



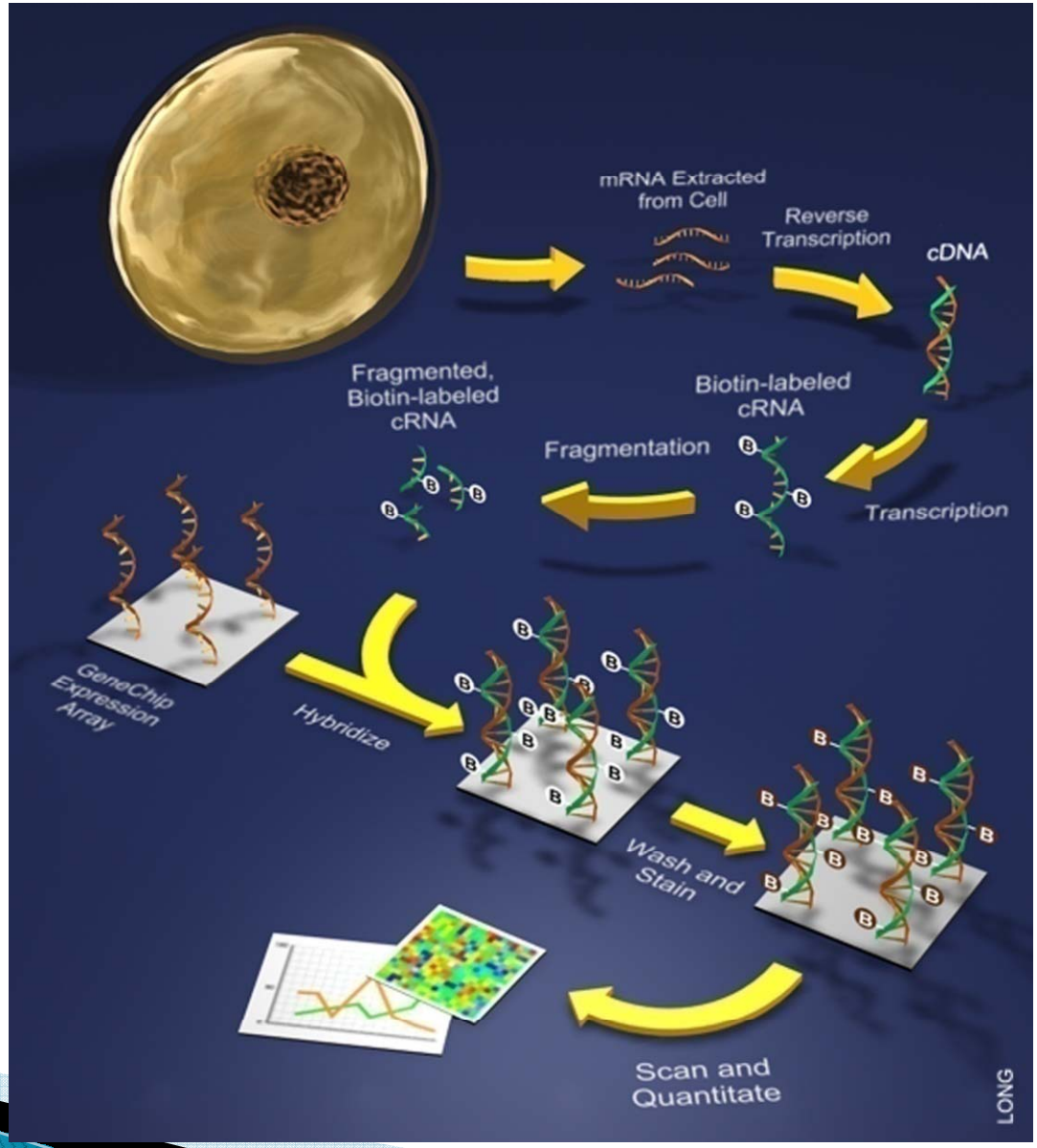
in Situ Synthesis : the Printing Process



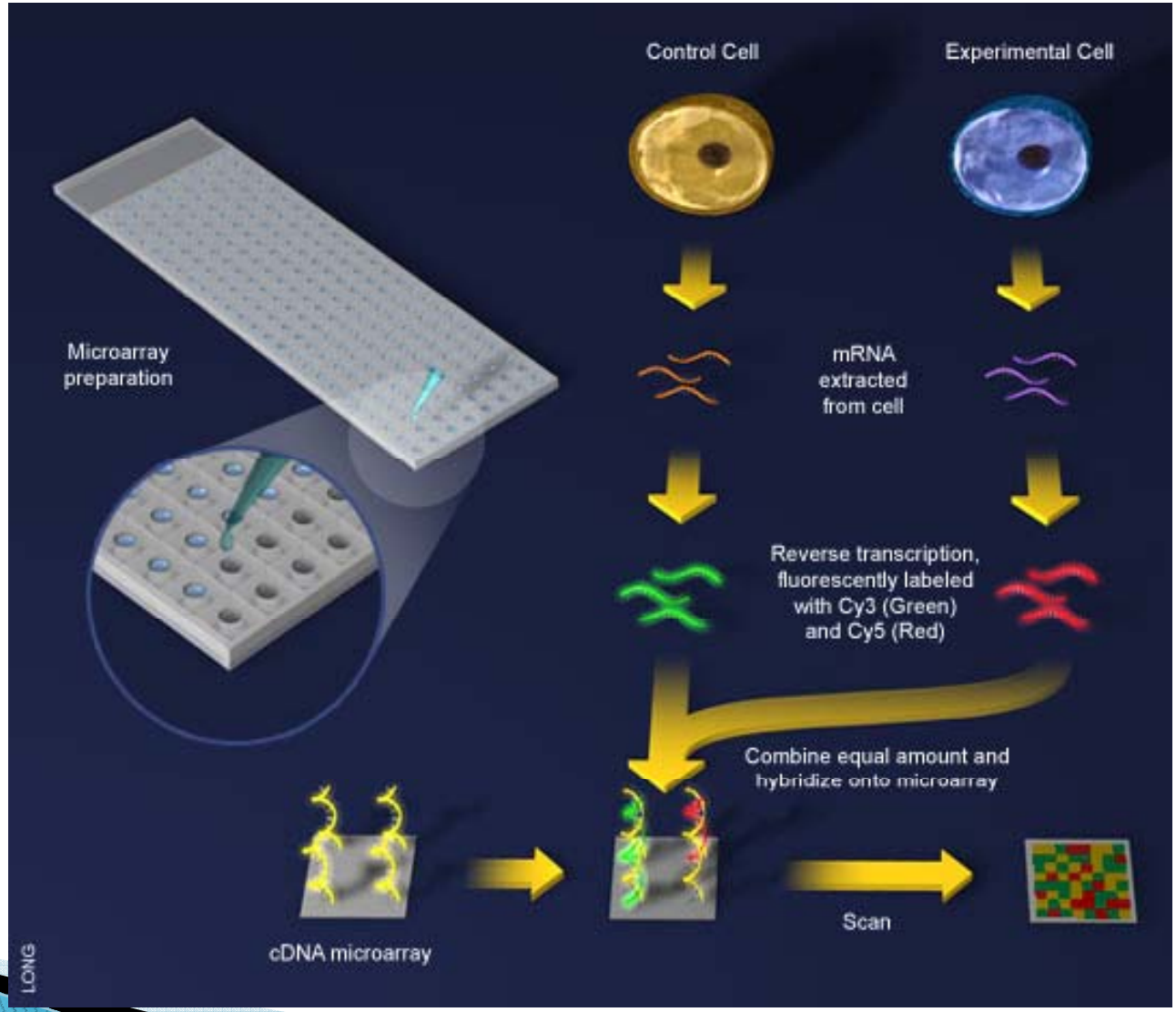


http://www.chem.agilent.com/en-US/Products/Instruments/dnamicroarrays/Pages/gp557.aspx

A Single-Color Array Experiment Workflow



A Two-Color Array Experiment Workflow



cDNA	Array	Ratio	Log_2(Ratio)
		5.0	**2.3**
		2.0	**1.0**
		2.0	**1.0**
		1.0	**0.0**
		0.5	**-1.0**
		0.2	**-2.3**

Animation of Microarray Exp

This animation will demonstrate how DNA microarray experiments are performed.

Throughout the animation, you may use the mouse to identify components of the experiment. Try the yeast cell below for starters.

We will use yeast as a model system to illustrate one use of microarrays, sometimes called DNA chips.

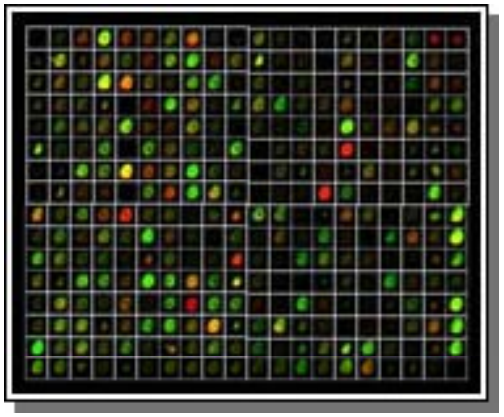
http://www.bio.davidson.edu/Courses/genomics/chip/chip.html

The Workflow

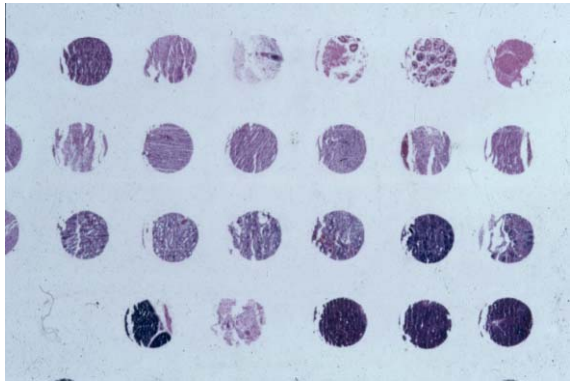
- Image processing
  - Spot identification, intensity
- Normalization
  - within array: background correction, print tip effects...etc
- Normalization between arrays:
  - global normalization, spike-in controls, internal controls
  - mean/median expression level
- Identifying significantly expressed genes:
  - fold changes
  - Statistical analysis through replicated experiments
- Application specific analysis:
  - clustering, regulatory networks.... etc

Types of "BioChip"

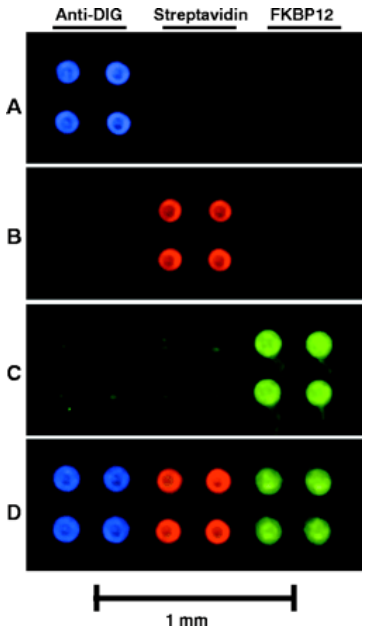
DNA Chip



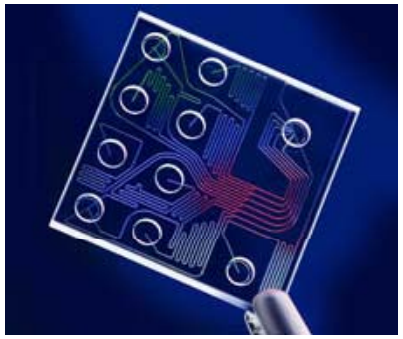
Tissue Array



Protein Chip

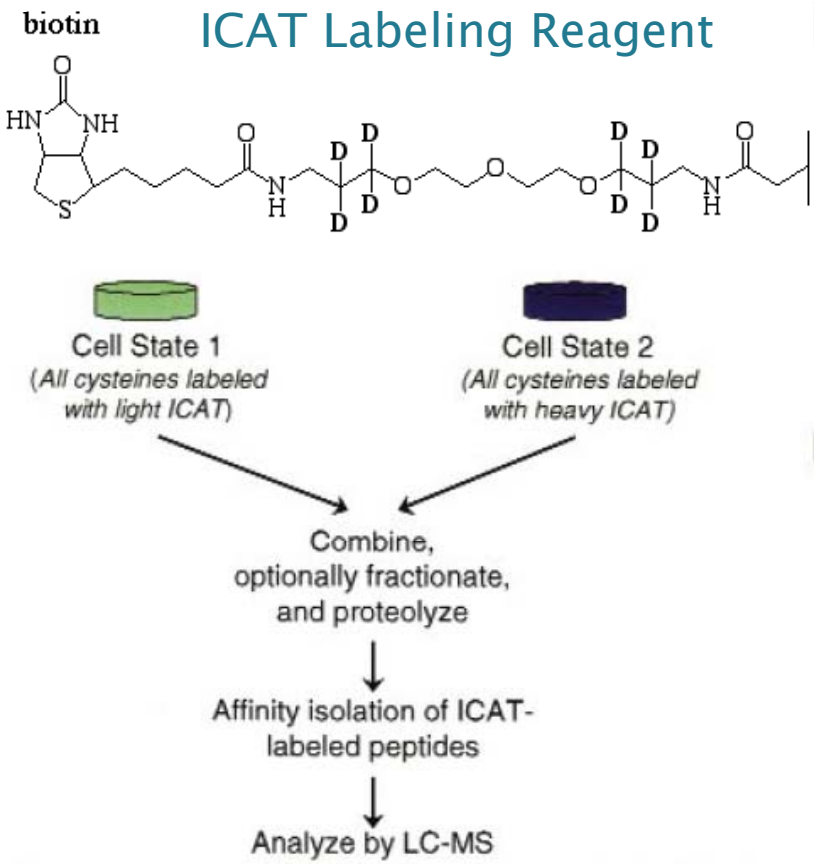


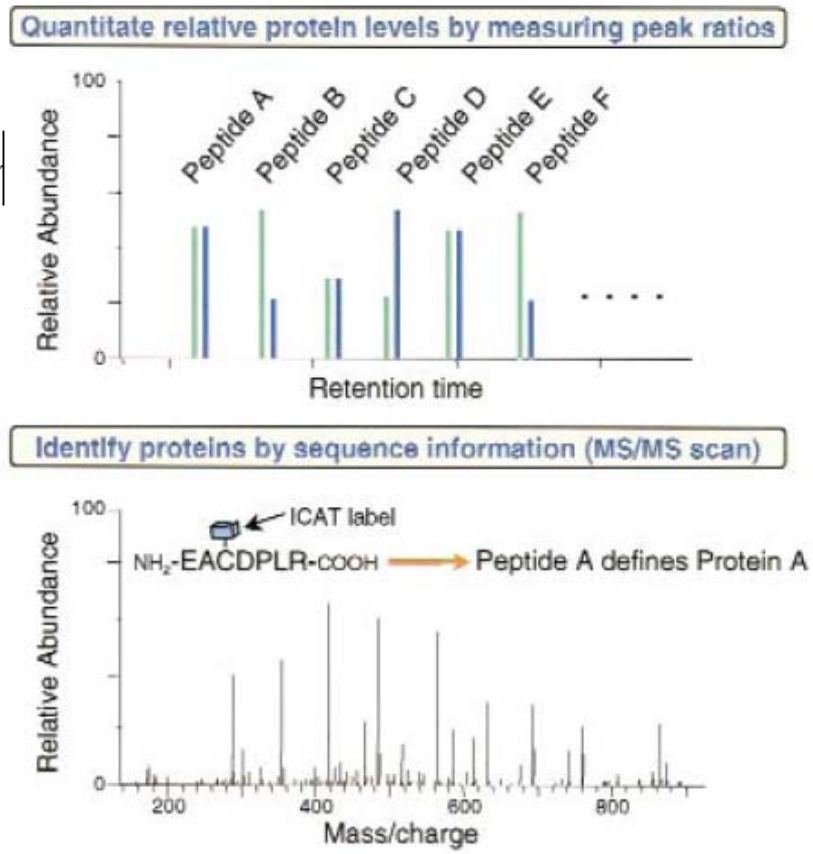
Lab on Chip



Isotope-Coded Affinity Tags

http://www.proteomecenter.org/PDFs/Gygi.NatBiotech.99.pdf

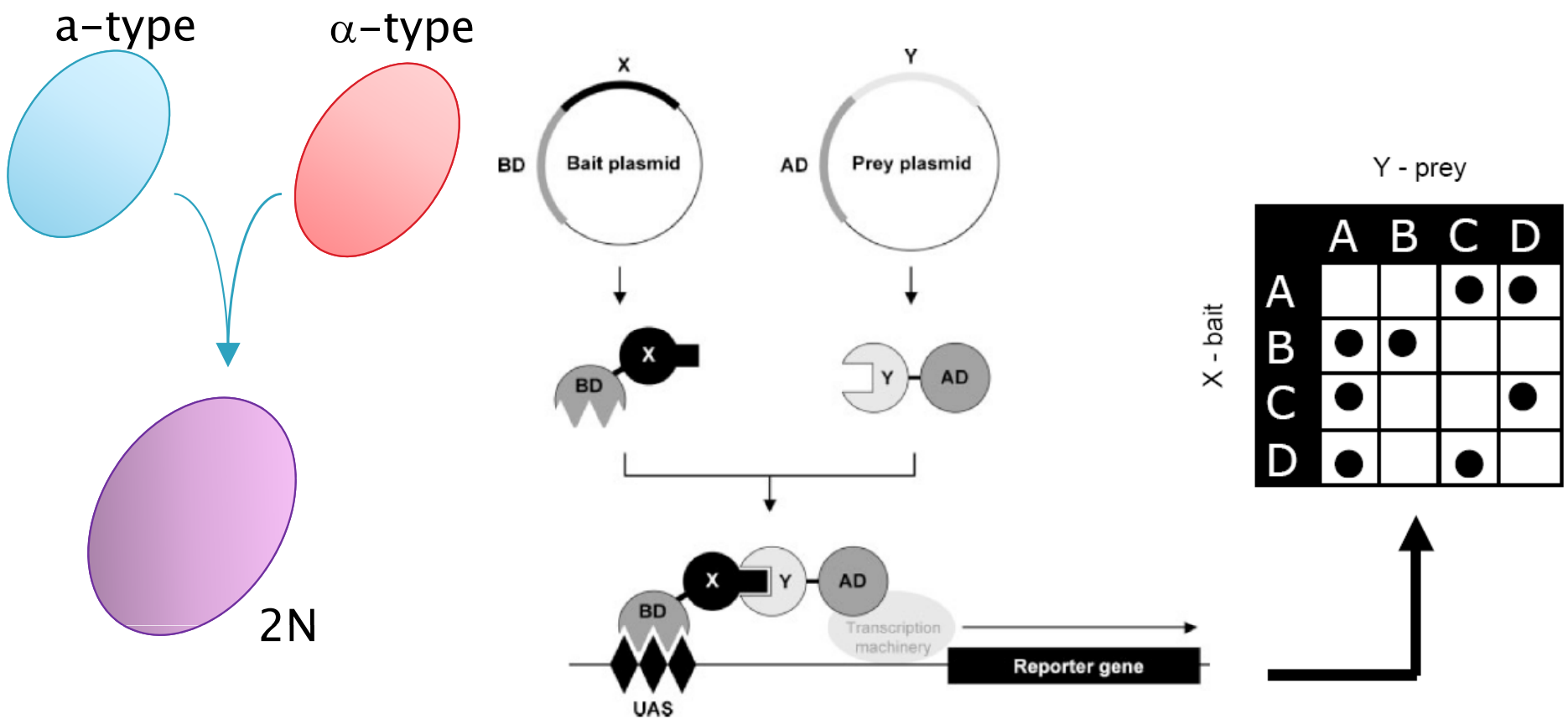




Animation Demo of ICAT Technology

http://www.bio.davidson.edu/Courses/genomics/ICAT/ICAT.html

Yeast Two-Hybrid



An Example of a Systems Biology Approach

articles

Global analysis of protein localization in budding yeast

Won-Ki Huh[1*], James V. Falvo[1*], Luke C. Gerke[1], Adam S. Carroll[1], Russell W. Howson[1], Jonathan S. Weissman[1,2] & Erin K. O'Shea[1]

[1]*Howard Hughes Medical Institute, University of California–San Francisco, Department of Biochemistry and Biophysics, and* [2]*Department of Cellular and Molecular Pharmacology, 600 16th Street, San Francisco, California 94143-2240, USA*

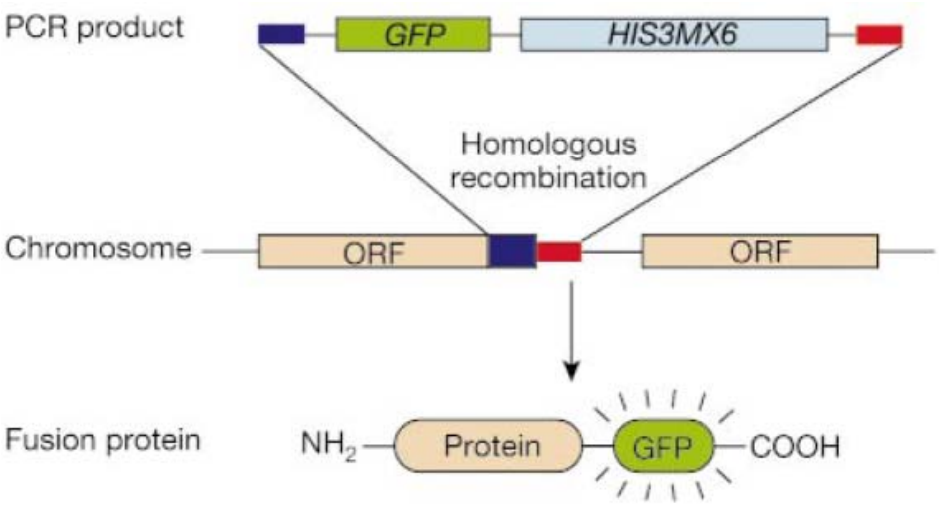
*These authors contributed equally to this work

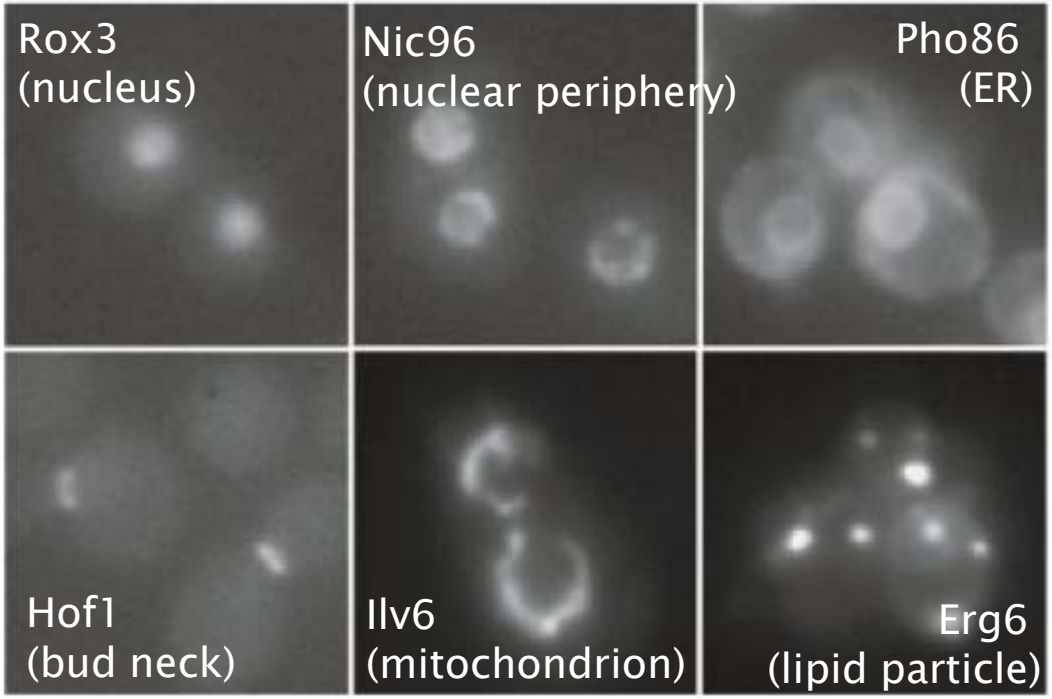
A fundamental goal of cell biology is to define the functions of proteins in the context of compartments that organize them in the cellular environment. Here we describe the construction and analysis of a collection of yeast strains expressing full-length, chromosomally tagged green fluorescent protein fusion proteins. We classify these proteins, representing 75% of the yeast proteome, into 22 distinct subcellular localization categories, and provide localization information for 70% of previously unlocalized proteins. Analysis of this high-resolution, high-coverage localization data set in the context of transcriptional, genetic, and protein–protein interaction data helps reveal the logic of transcriptional co-regulation, and provides a comprehensive view of interactions within and between organelles in eukaryotic cells.

NATURE (2003), 425:686

Microscopic analysis of yeast strains expressing GFP-tagged proteins

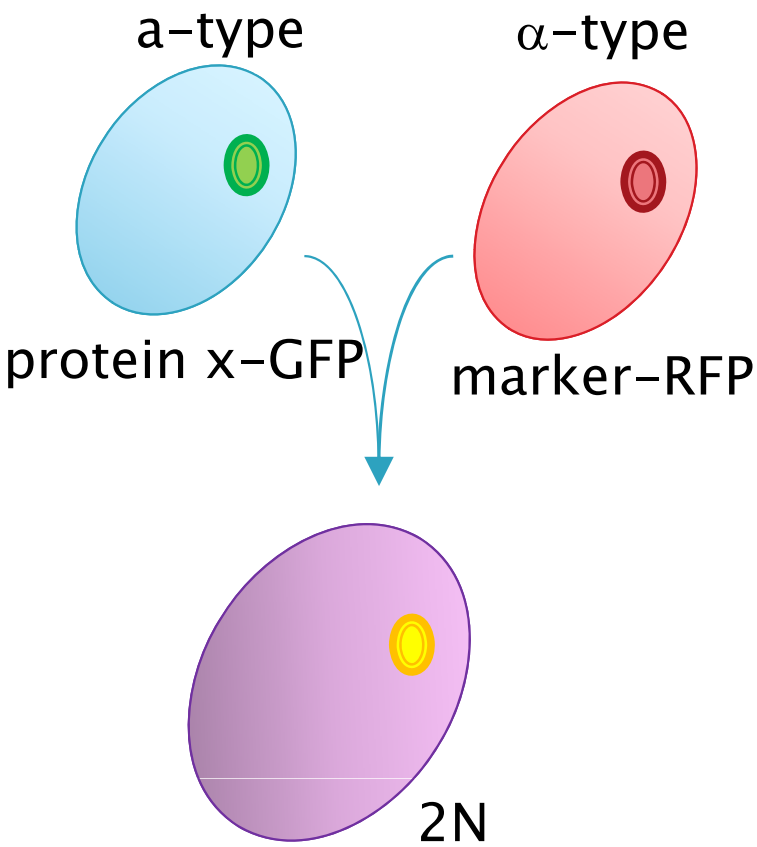
- Constructing recombinant strains

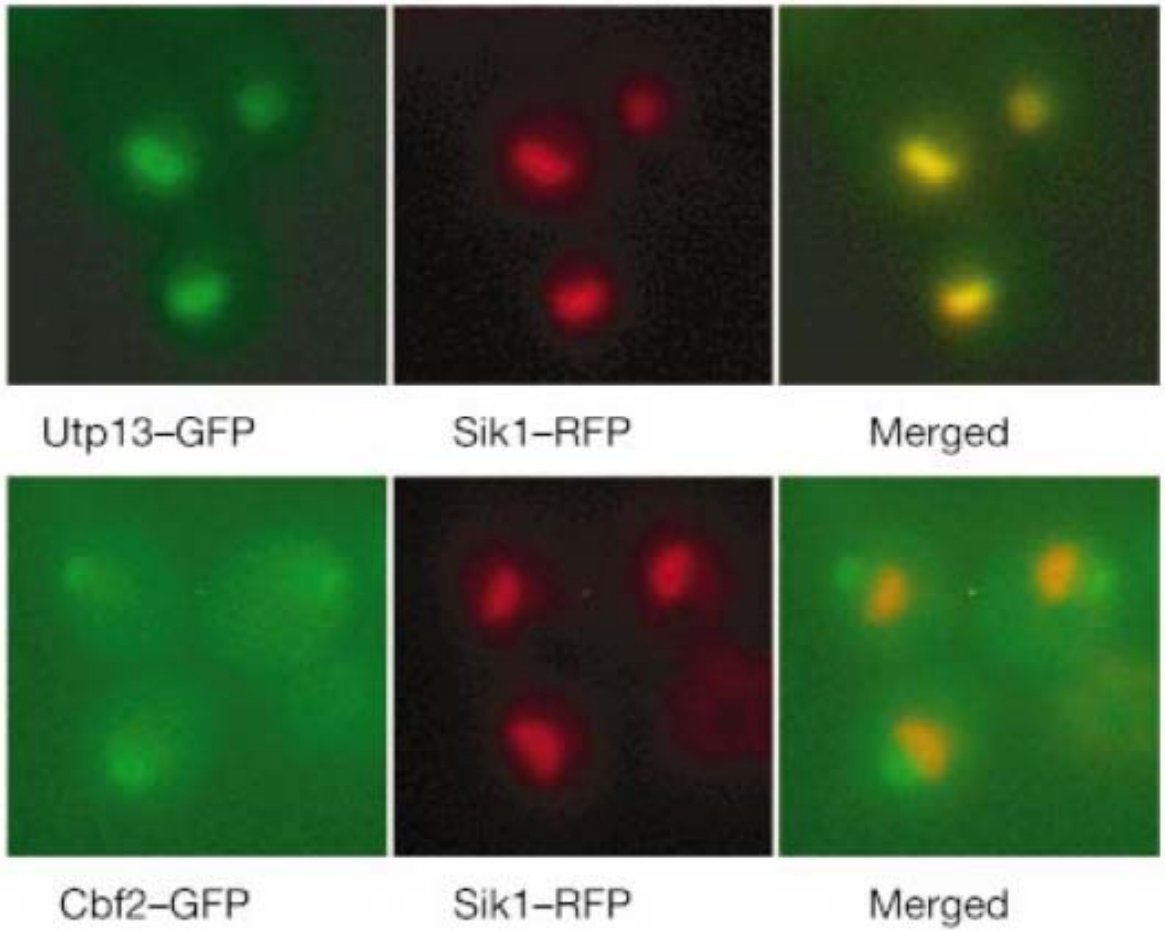




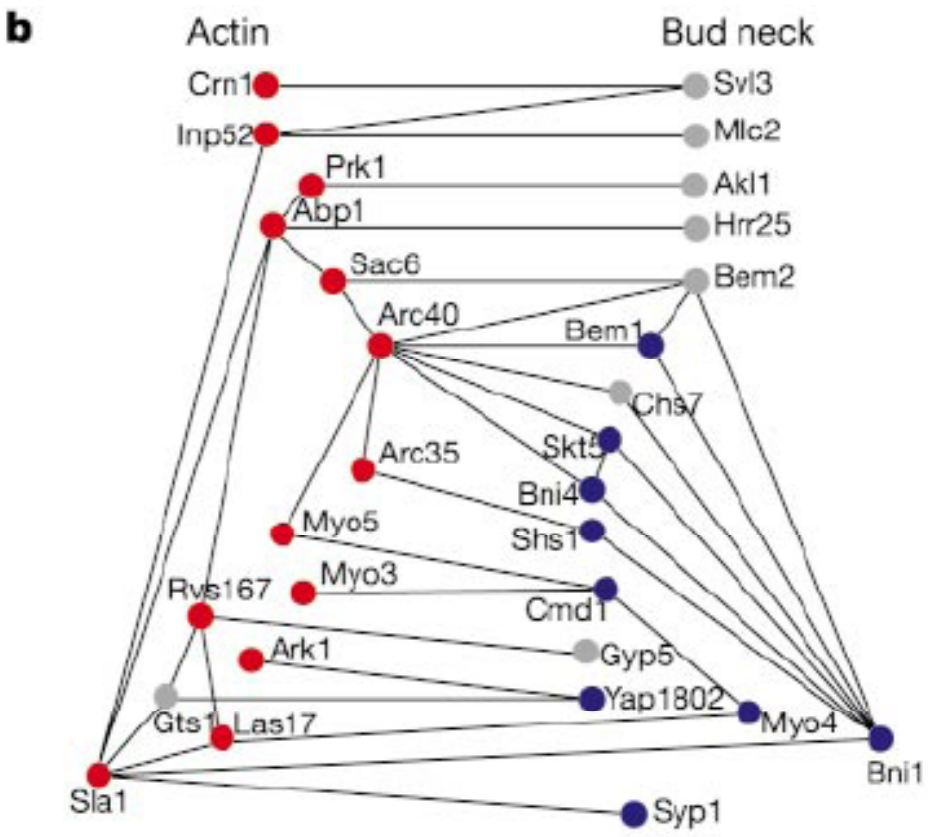
Huh WK et al., 2003. NATURE 425:686

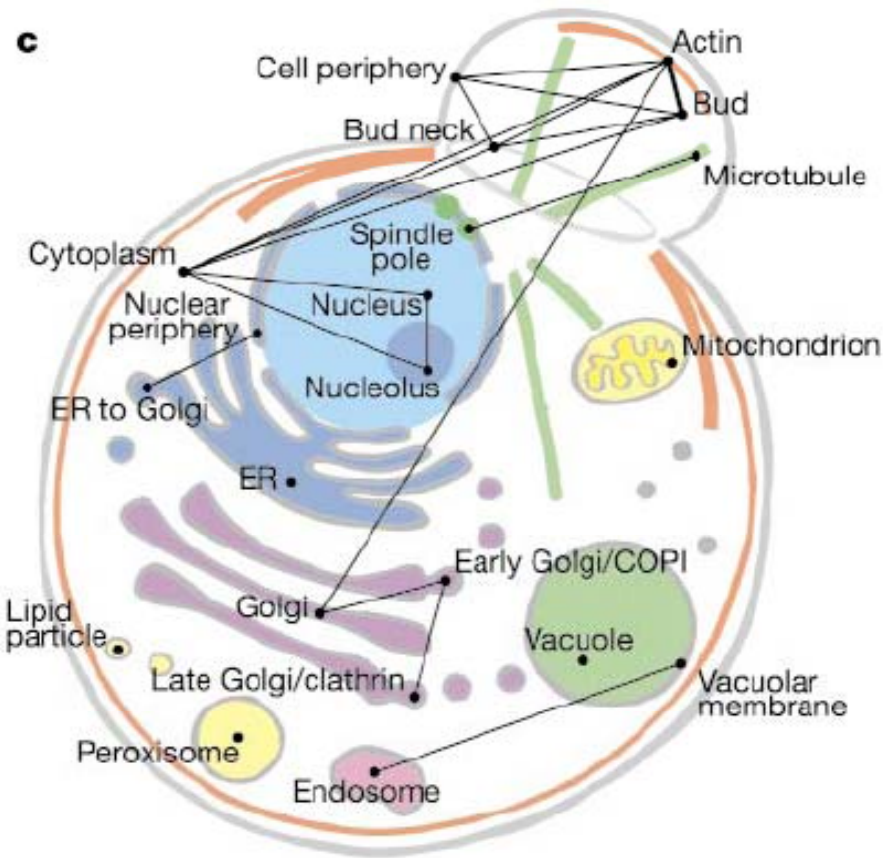
Representative co-localization experiment





Relationship between genetic and physical interactions and subcellular localization

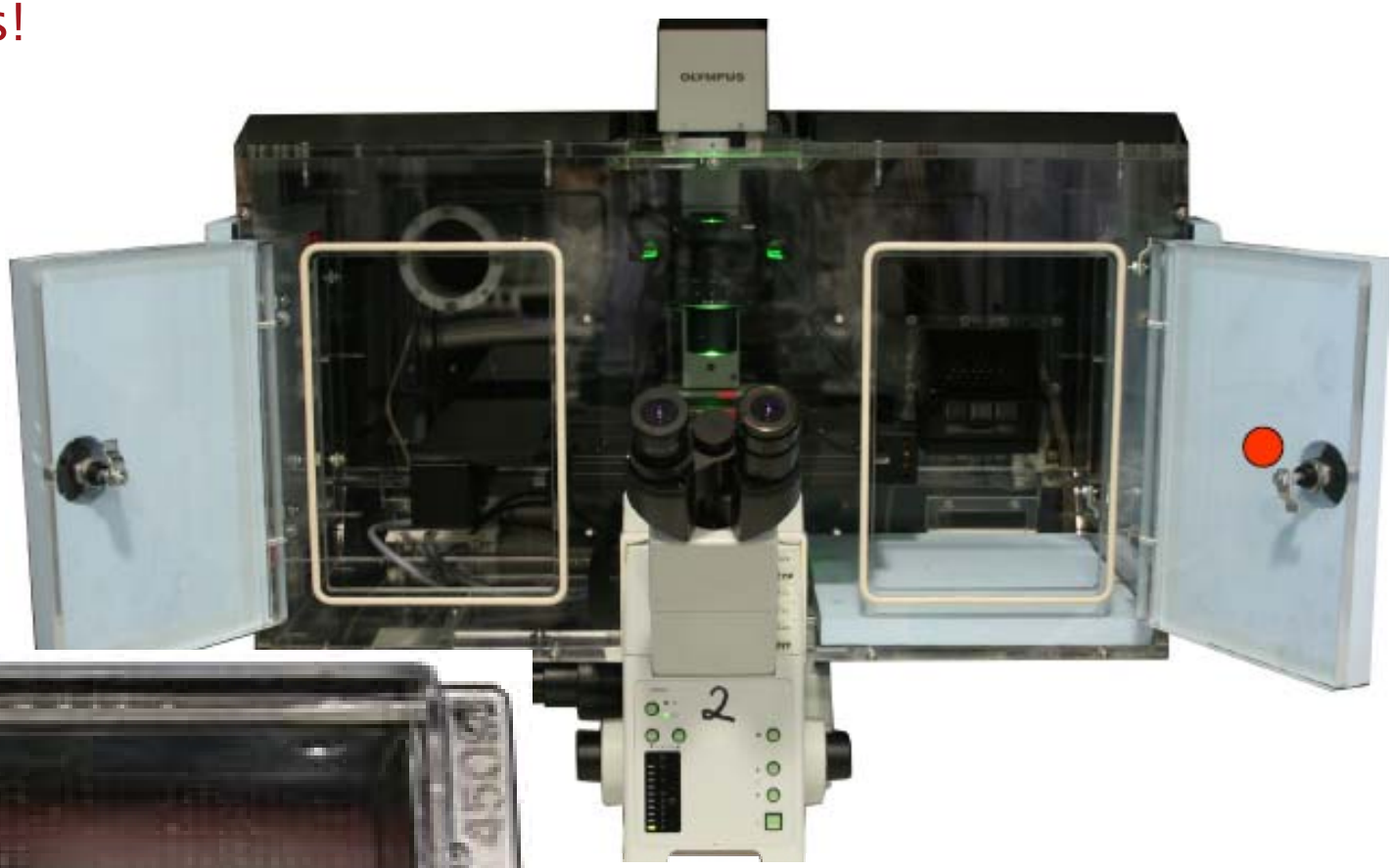




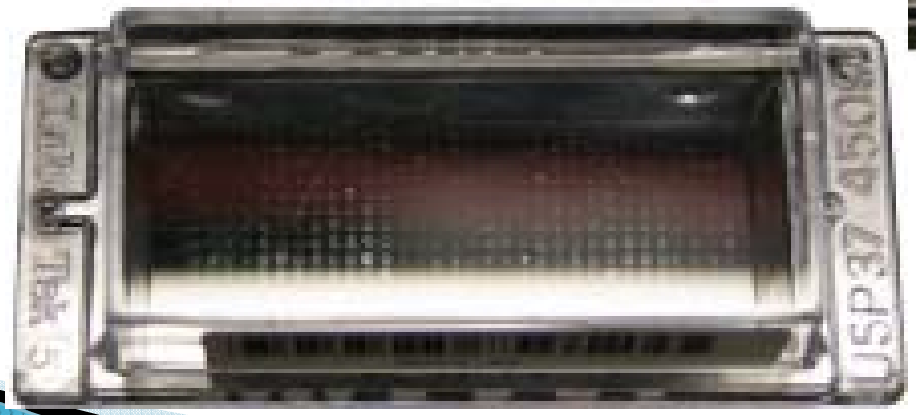
Huh WK et al., 2003. NATURE 425:686

High-Content Microscopy

[384 x 4 =1536 experiments] x # of Focal panels x # of Time-lapse design
➔ innumerous Images!



384-well slide



Adopted from a speach of Urban Liebel, Mitocheck project group,EMBL Heidelberg http://harvester.embl.de/media/2006-01-13-dresden.pdf

High-Throughput Microscopy

4 cell arrays / microscope: currently 1536 spots

time-lapse: 30 min
total assay time: 48 h
data/microscope: 350 GByte (x3 microscopes)

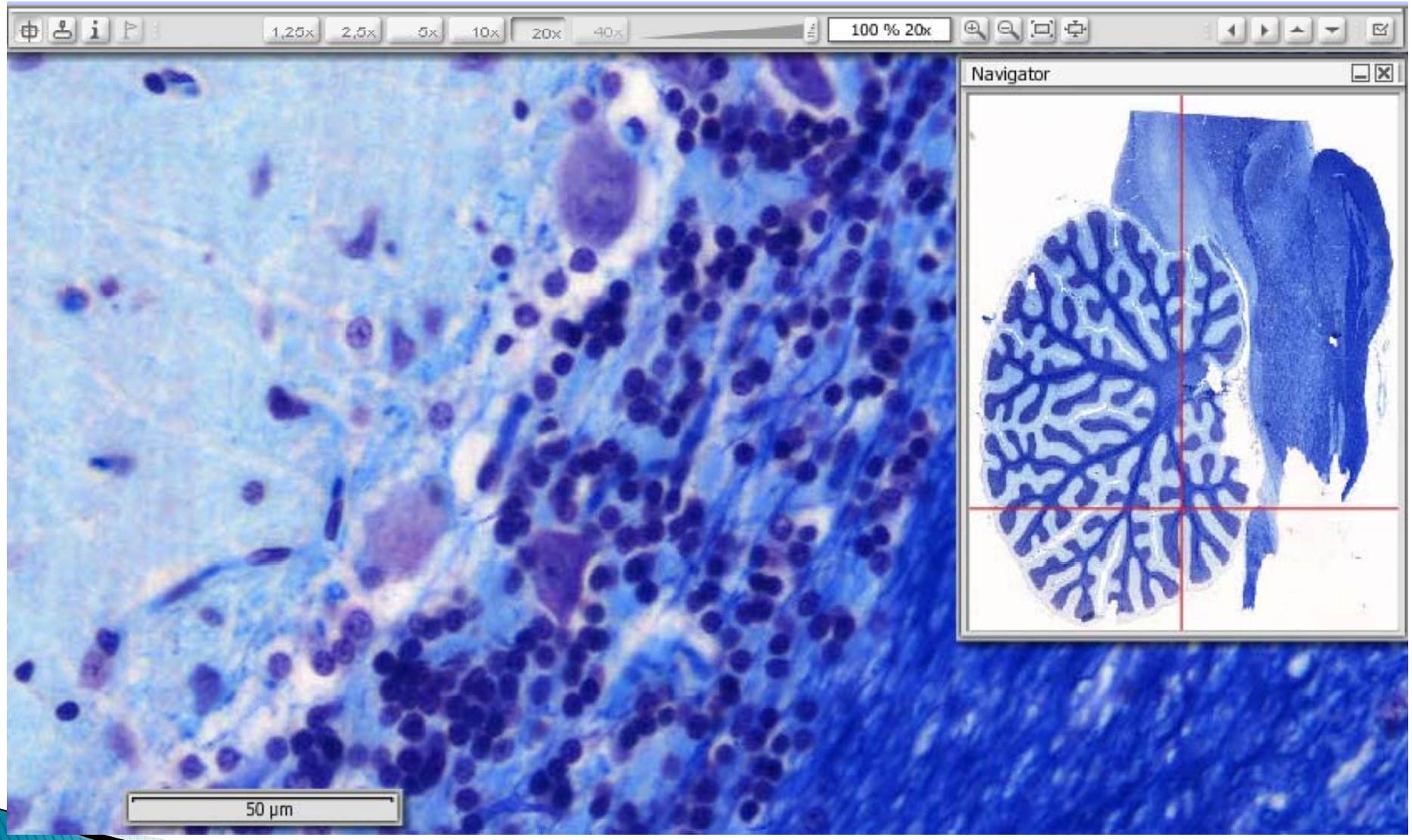
data flow: 1- 1,5 TByte / week (compressed)

data/genome: ~132 000 movies (xVID codec)
~360 arrays w/ replicates ~32 TByte

time/genome: ~100 days (3 microscopes)

Adopted from a speach of Urban Liebel, Mitocheck project group,EMBL Heidelberg http://harvester.embl.de/media/2006-01-13-dresden.pdf

A Whole View on a Slice of the Body



http://dotslide.olympus-sis.com/

FREE YOUR MIND