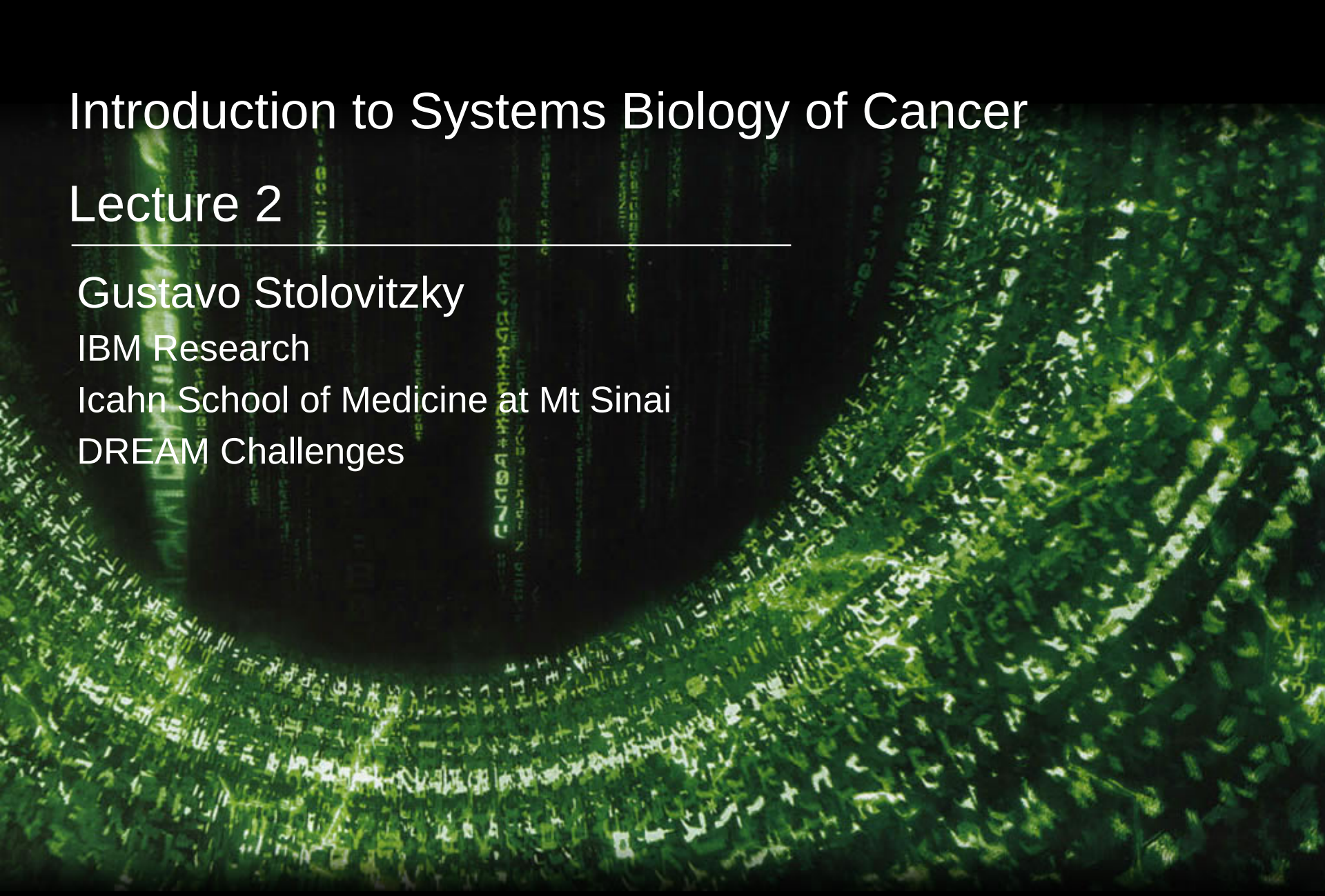
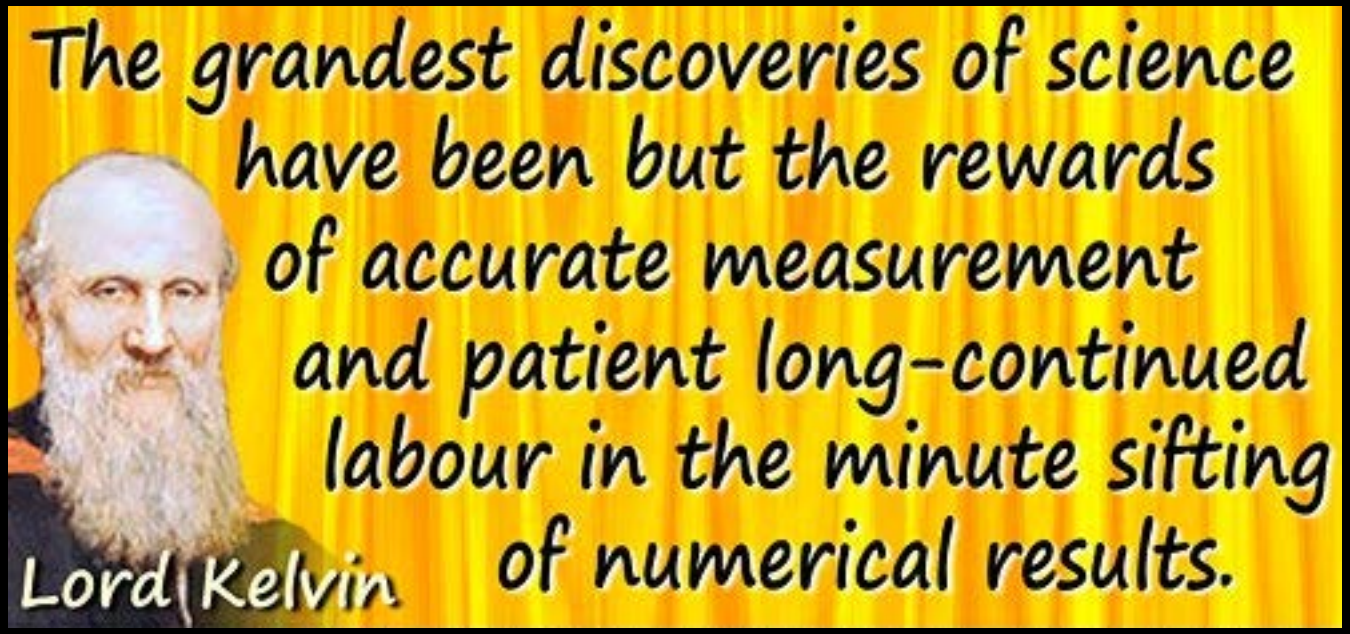
Introduction to Systems Biology of Cancer

Lecture 2

Gustavo Stolovitzky
IBM Research
Icahn School of Medicine at Mt Sinai
DREAM Challenges





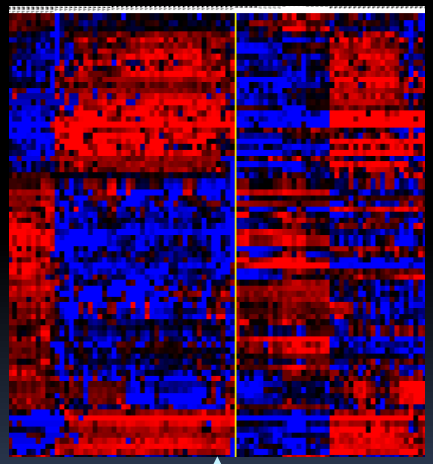
High throughput measurements: The age of omics

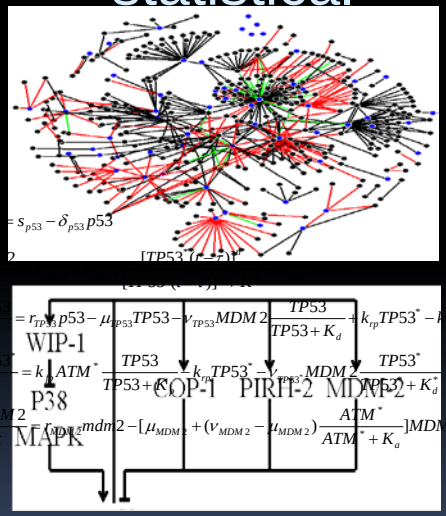
Systems Biology deals with four main tasks

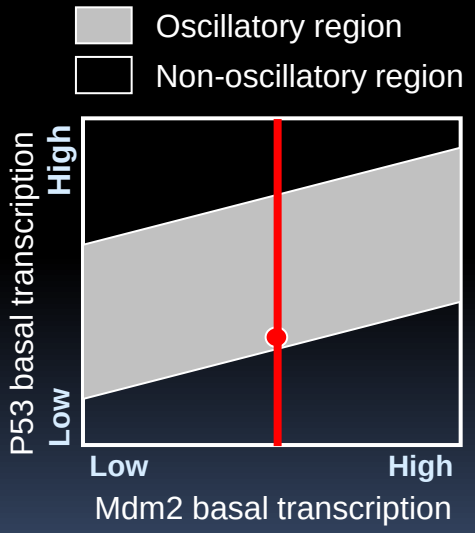
Measurements
New High
Throughput Omics
technologies

Modeling
Data
exploration,
deterministic
statistical

System
Characterization &
Predictions:
Clinical & Biological







Model testing and Validation

What do we need to measure in cancer research

Given what we saw in the Lecture 1, we need to measure the elements of the genome that are disregulated, as well as their functional consequences.

At the DNA level sequence (static)

- Mutations, Copy number alterations, Loss of heterozygosity, Translocations

Epigenetics (static)

- DNA methylation, histone modifications (methylation, acetylation)

At the RNA level, quantify amount (functional)

- Non-coding RNA, microRNA, mRNA, splice variants

What do we need to measure in cancer research

At the protein level

- Protein amounts, phosphorylation and other postranslational modifications.

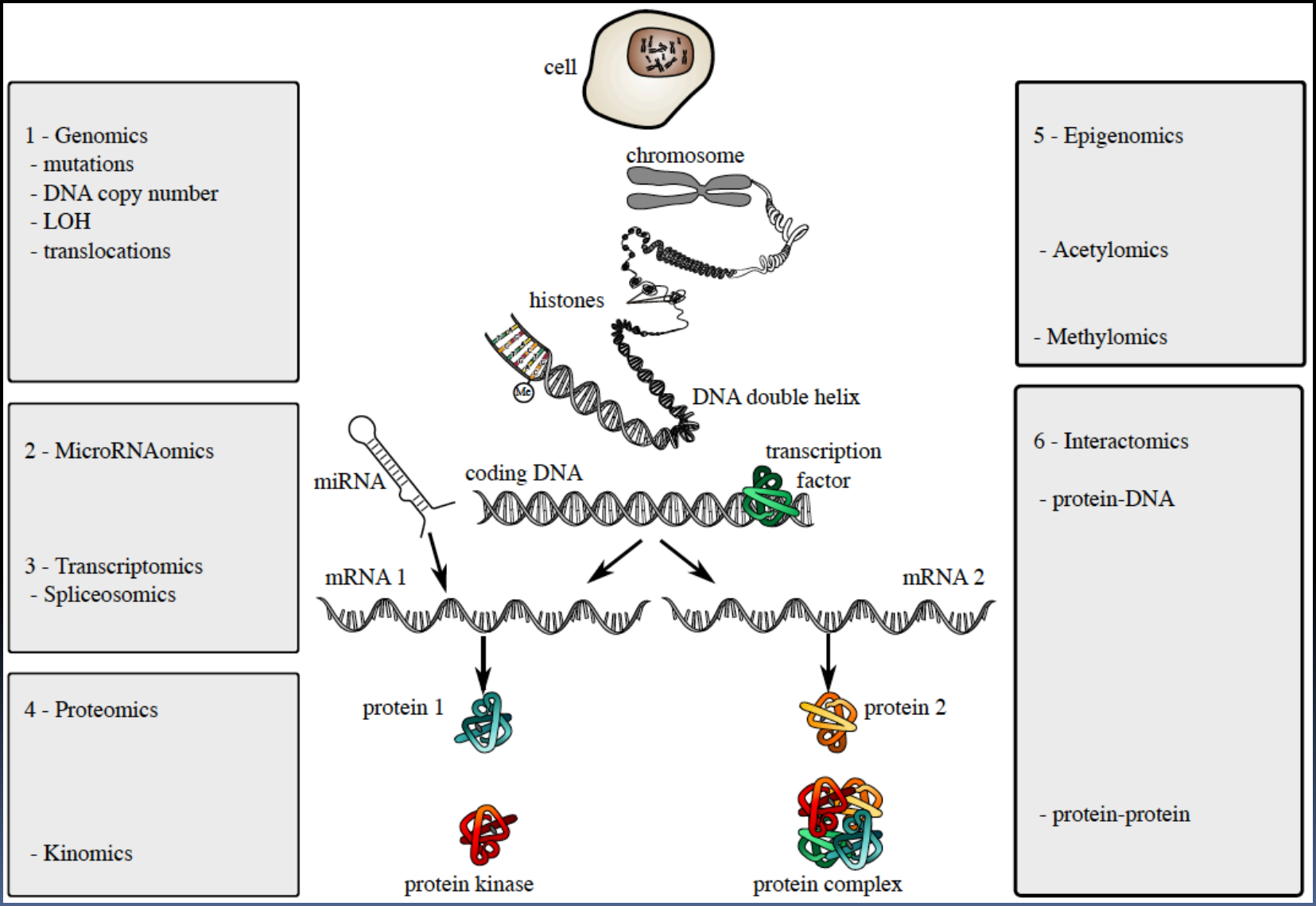
Interactions maps

- Protein (e.g. TF)-DNA interactions, protein-protein interactions

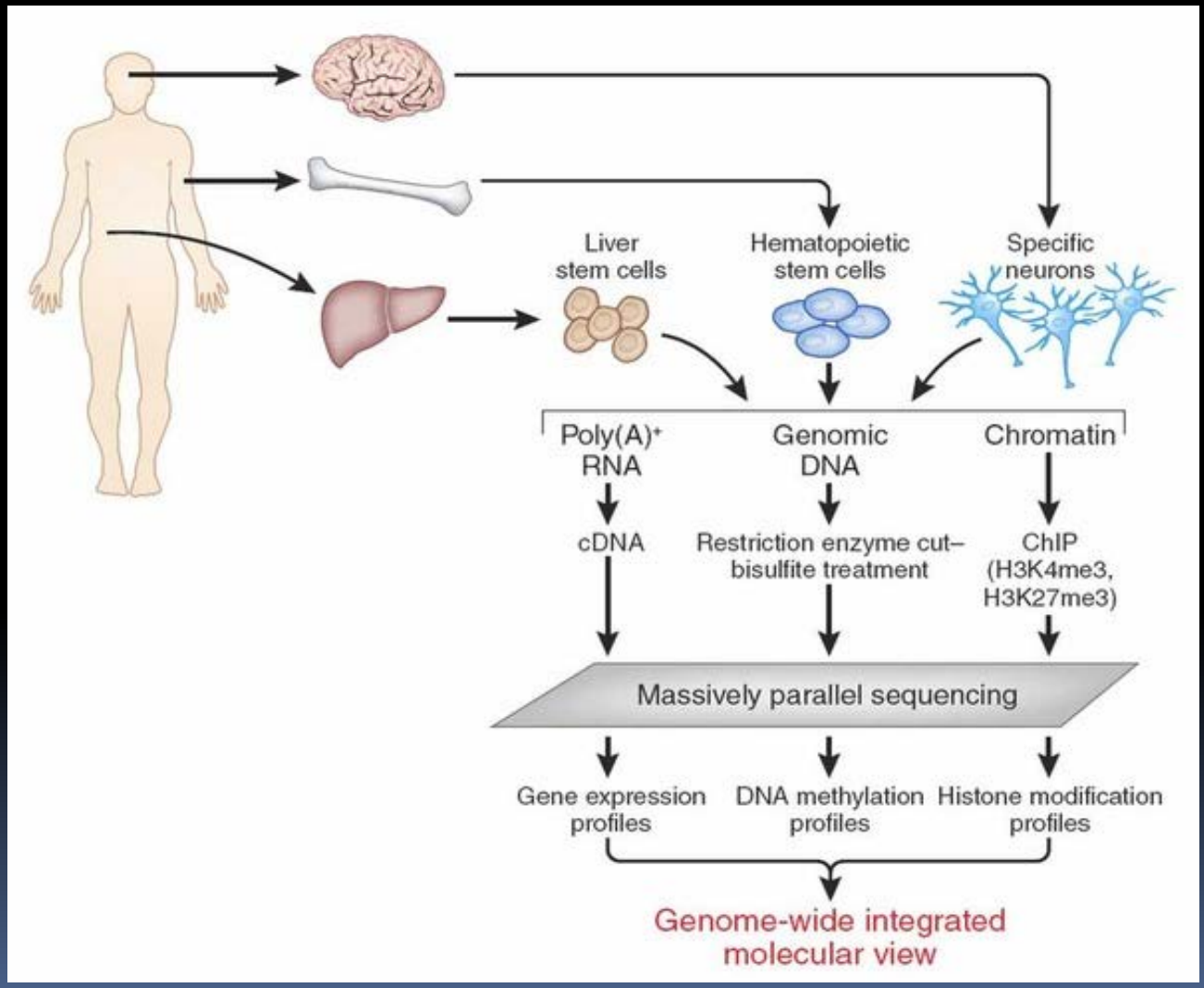
Phenotypes

- Cell viability, patient survival, Patient response to treatment

Omics Technologies



Many biological experiments involve sequencing



DNA Technology Milestones

1952	Electophoresis (Milestone 1)
1967	Discovery of DNA ligase (Milestone 2)
1969	FISH (Milestone 3)
1970	Discovery of restriction enzymes (Milestone 4)
	Discovery of reverse transcriptase (Mile[illegible])
1972	Cloning (Milestone 2)
1975	Southern blot (Milestone 6)
1977	DNA sequencing (Milestone 7)
1980	RFLP concept (Milestone 8)
1982	P-element-mediated manipulation of the fly genome (Milestone 9)
	Whole genome shotgun (Milestone 10)
1983	RFLP realization (Milestone 8)
1985	PCR (Milestone 11)
	DNA fingerprinting (Milestone 12)
1987	YACs (Milestone 13)
	Site-directed mutagenesis of the mouse genome (Milestone 9)

1988	ChIP (Milestone 14)
1990	BLAST — the key to comparative genomics (Milestone 15)
1992	BACs (Milestone 13)
1995	Microarray technology (Milestone 16)
1998	RNAi (Milestone 17)
	Sequencing by synthesis (Milestone 18)
	Full-length cDNA technologies (Milestone 5)
2002	Launch of UCSC Genome Browser (Milestone 19)
2003	DNA assembly programs (Milestone 20)
2004	ENSEMBL — an example of a gene annotation tool (Milestone 21)
2005	HapMap (Milestone 22)
	Sequencing by ligation/polony sequencing (Milestone 18)
2006	Genome-wide maps of DNA methylation (Milestone 23)

From Nature Milestones, DNA Technologies

Sanger Sequencing

Proc. Natl. Acad. Sci. USA
Vol. 74, No. 12, pp. 5463–5467, December 1977
Biochemistry

DNA sequencing with chain-terminating inhibitors

(DNA polymerase/nucleotide sequences/bacteriophage ϕX174)

F. SANGER, S. NICKLEN, AND A. R. COULSON

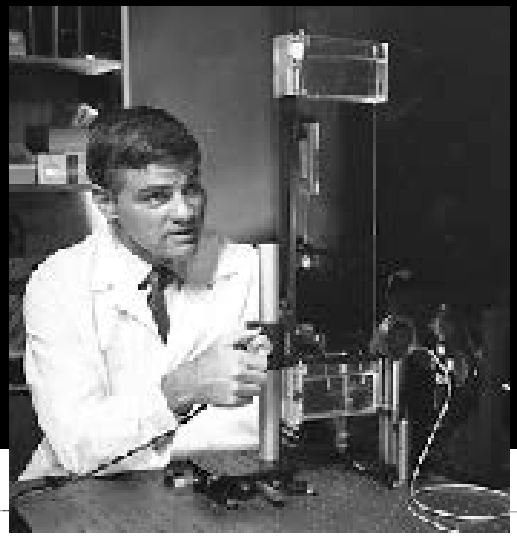
Medical Research Council Laboratory of Molecular Biology, Cambridge CB2 2QH, England

Contributed by F. Sanger, October 3, 1977

ABSTRACT **A new method for determining nucleotide sequences in DNA is described. It is similar to the "plus and minus" method [Sanger, F. & Coulson, A. R. (1975) *J. Mol. Biol.* 94, 441–448] but makes use of the 2′,3′-dideoxy and arabinonucleoside analogues of the normal deoxynucleoside triphosphates, which act as specific chain-terminating inhibitors of DNA polymerase. The technique has been applied to the DNA of bacteriophage ϕX174 and is more rapid and more accurate than either the plus or the minus method.**

a stereoisomer of ribose in which the 3′-hydroxyl group is oriented in *trans* position with respect to the 2′-hydroxyl group. The arabinosyl (ara) nucleotides act as chain terminating inhibitors of *Escherichia coli* DNA polymerase I in a manner comparable to ddT (4), although synthesized chains ending in 3′ araC can be further extended by some mammalian DNA polymerases (5). In order to obtain a suitable pattern of bands from which an extensive sequence can be read it is necessary to have a ratio of terminating triphosphate to normal triphos-

Automatized Sanger Sequencing



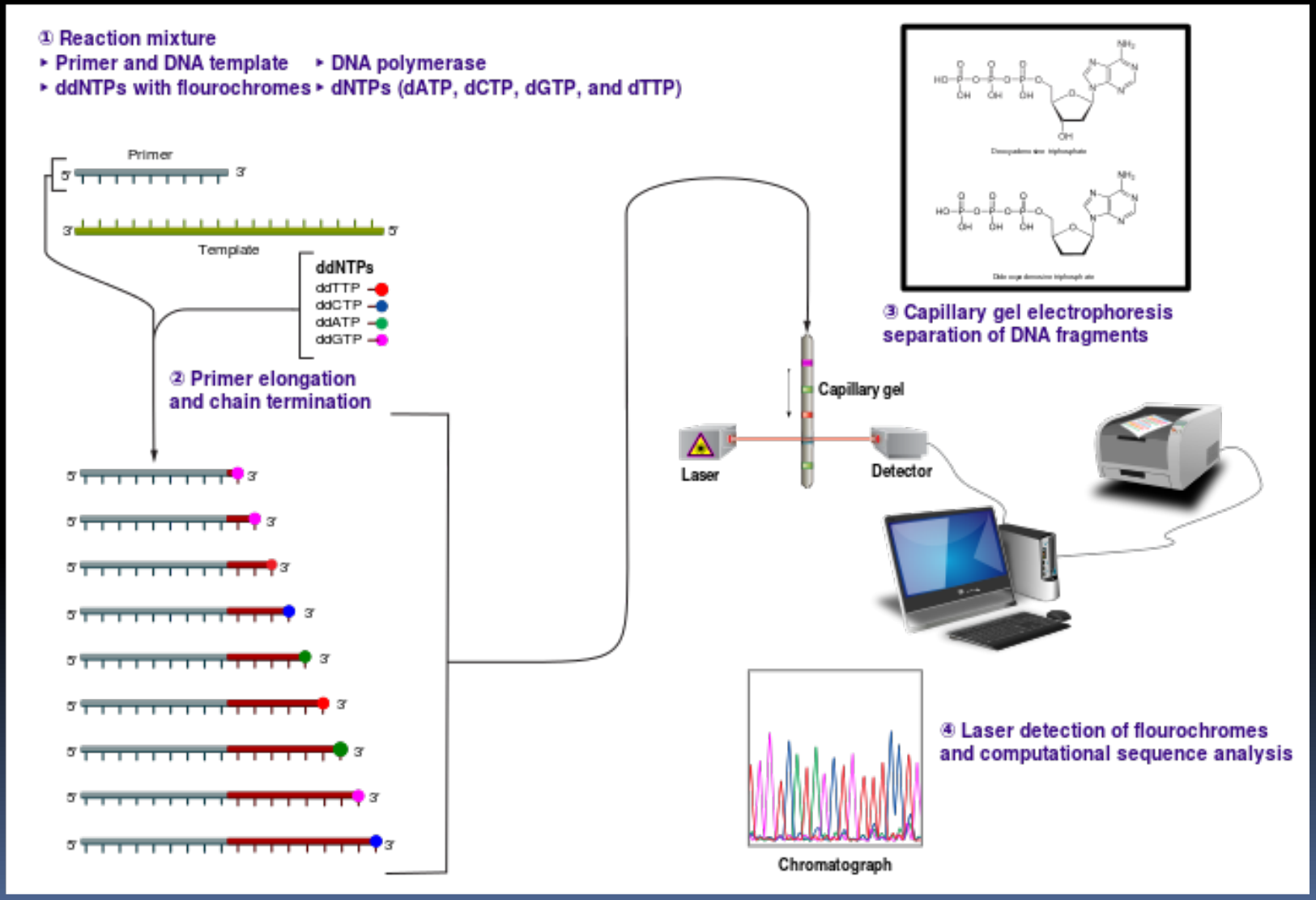
Fluorescence detection in automated DNA sequence analysis

Lloyd M. Smith, Jane Z. Sanders, Robert J. Kaiser, Peter Hughes, Chris Dodd, Charles R. Connell*, Cheryl Heiner*, Stephen B. H. Kent & Leroy E. Hood

Division of Biology, California Institute of Technology, Pasadena, California 91125, USA
* Applied Biosystems, Inc., Foster City, California 94404, USA

We have developed a method for the partial automation of DNA sequence analysis. Fluorescence detection of the DNA fragments is accomplished by means of a fluorophore covalently attached to the oligonucleotide primer used in enzymatic DNA sequence analysis. A different coloured fluorophore is used for each of the reactions specific for the bases A, C, G and T. The reaction mixtures are combined and co-electrophoresed down a single polyacrylamide gel tube, the separated fluorescent bands of DNA are detected near the bottom of the tube, and the sequence information is acquired directly by computer.

Sanger Sequencing



Progress in sequencing

2003 – First genome

- was a mixture of several volunteers
- Took 13 years (1990-2003), 3,000 scientists, $2.7 Billion
- Technology: Sanger Sequencing

2007 – Second Genome

- J.C.Venter's genome
- Took 4 years (2003-2007), 30 scientists, $100 Million
- Technology: Improved Sanger Sequencing

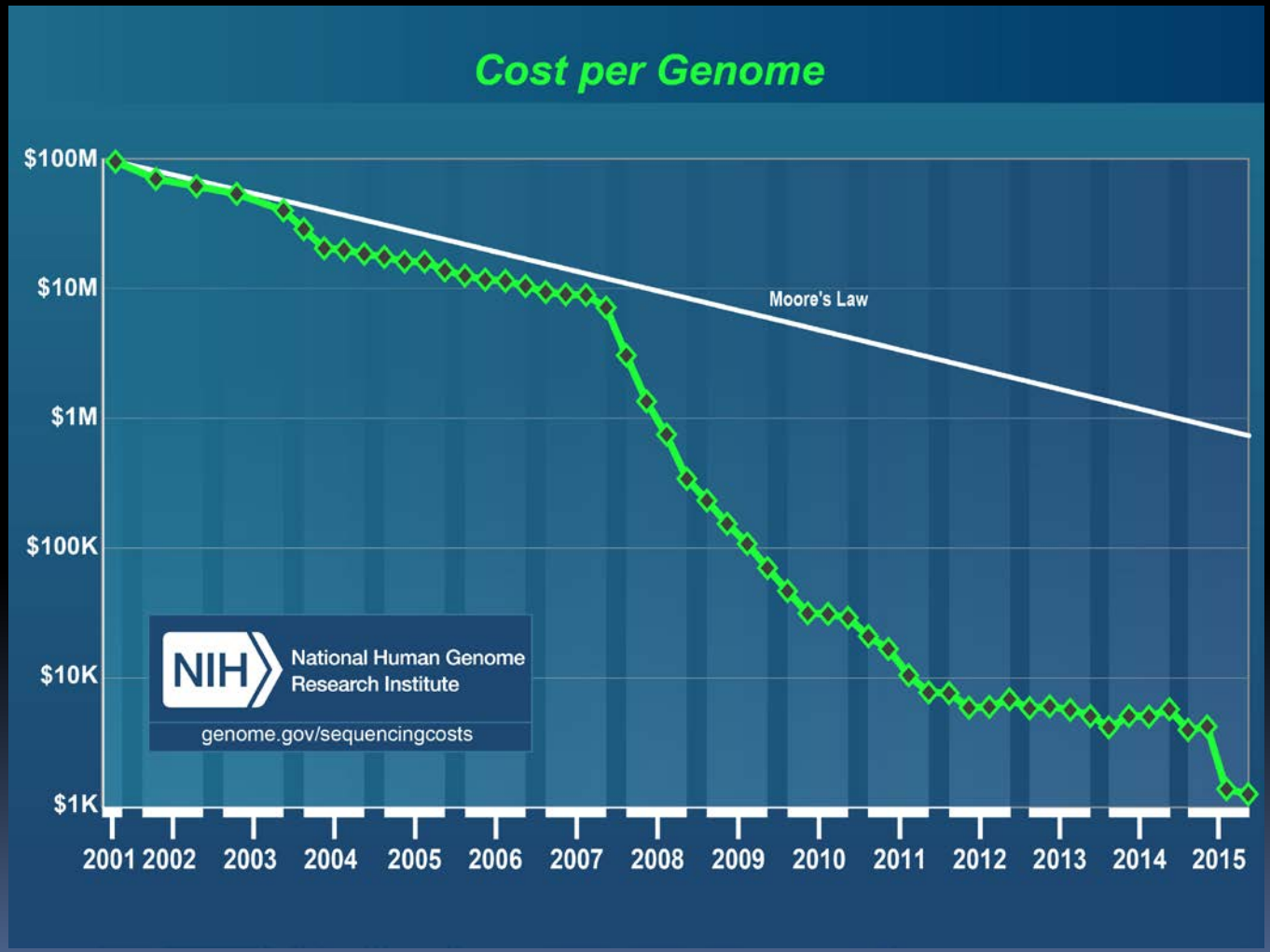
2008 – Third Genome

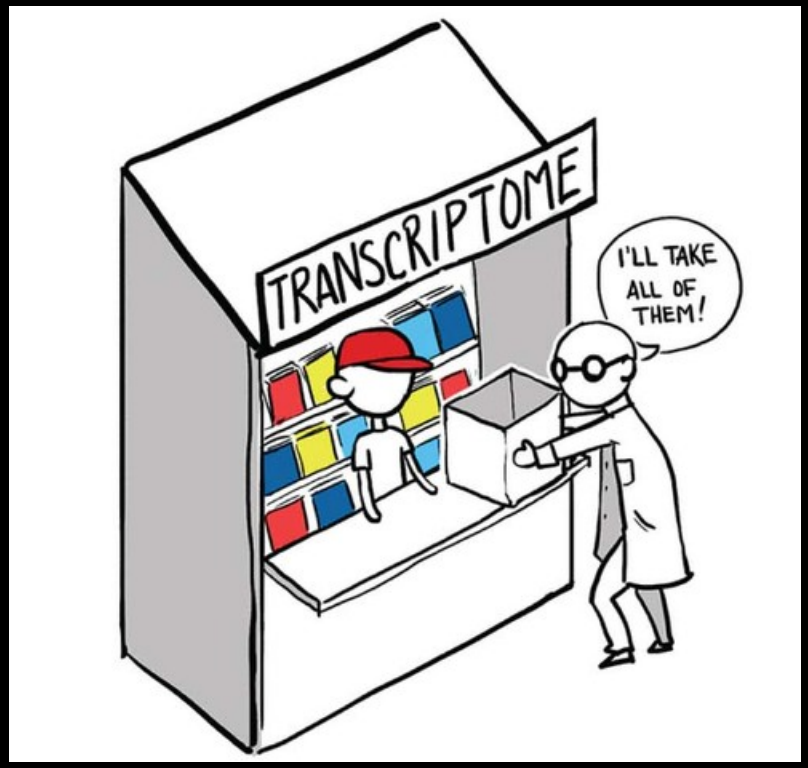
- James Watson
- Took 4.5 months (2008), ~30 scientists, $1.5 Million
- Technology: 454 (second generation, pyrosequencing)

end 2014 – ~ 250,000 Genomes

- Today sequencing costs < $1K
- Second GenerationTechnologies: 454 (defunct), Solid, Illumina (market leader),
- Third Generation Technologies: PacBio, Oxford nanopores

Sequencing is now at ~$1K

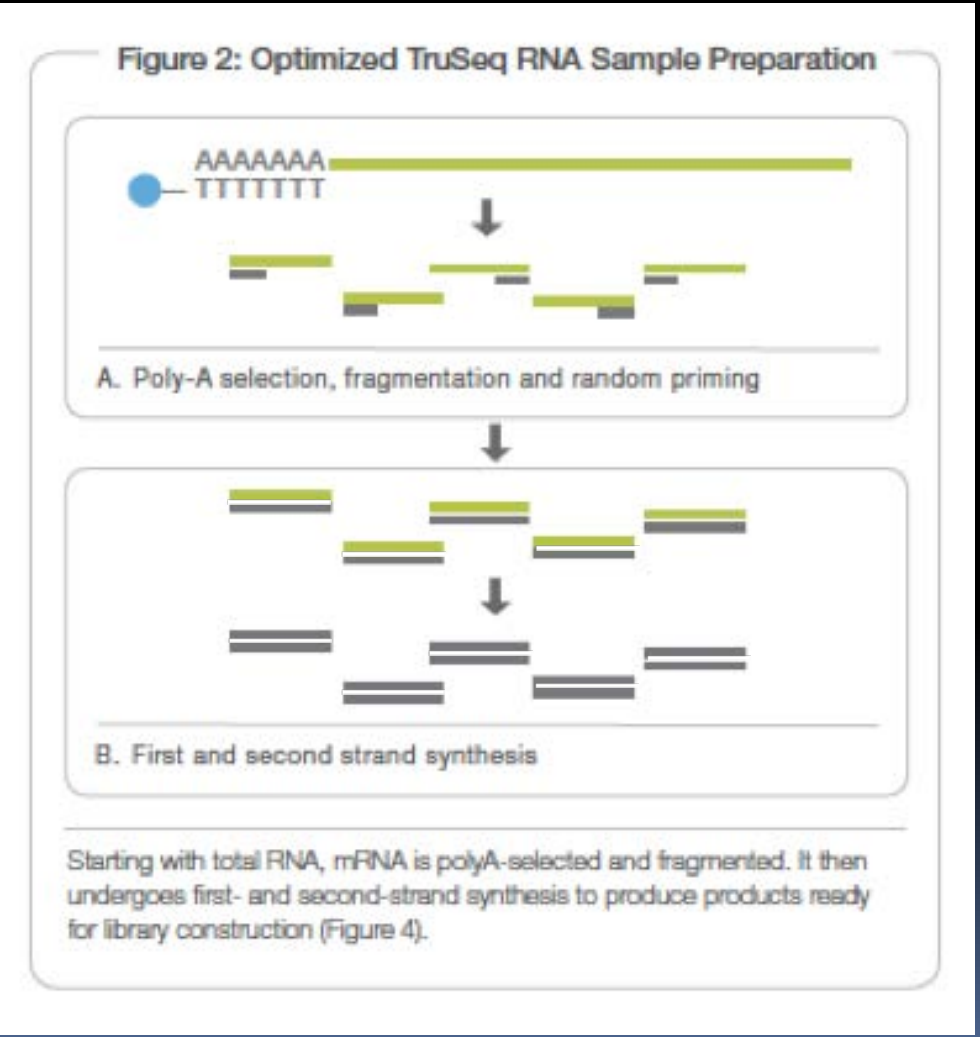




RNA-seq

Illumina sequencing

Library Construction



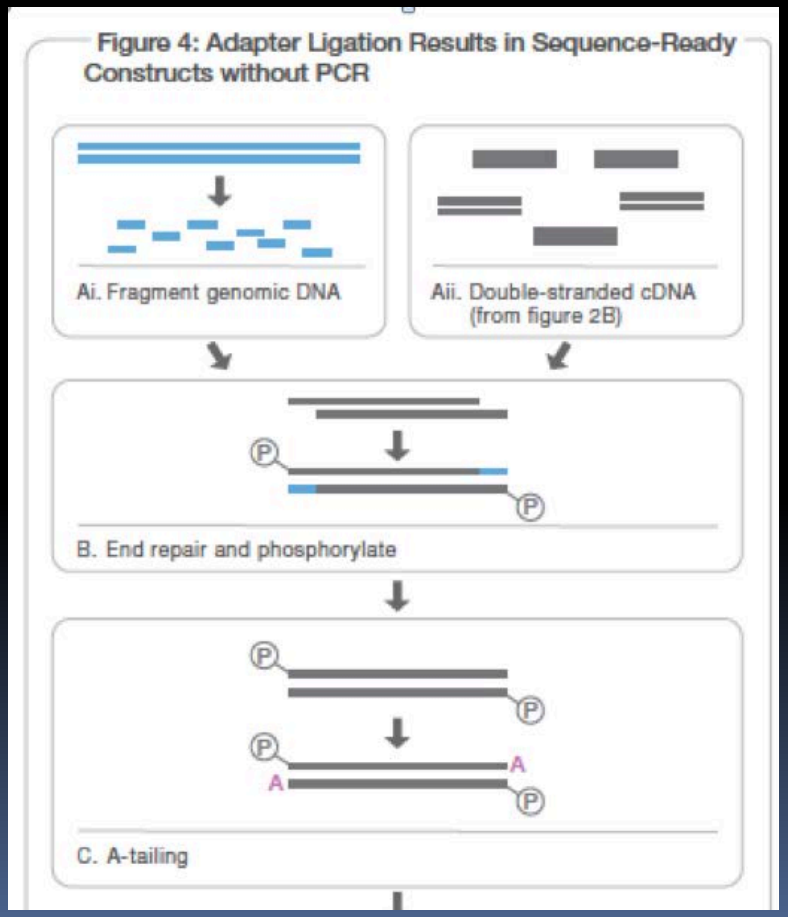
Before Library Construc;on

1. Poly-A Selection (Total RNA→ mRNA)
2. mRNA fragmentiaton
3. First strand synthesis
4. Second strand synthesis

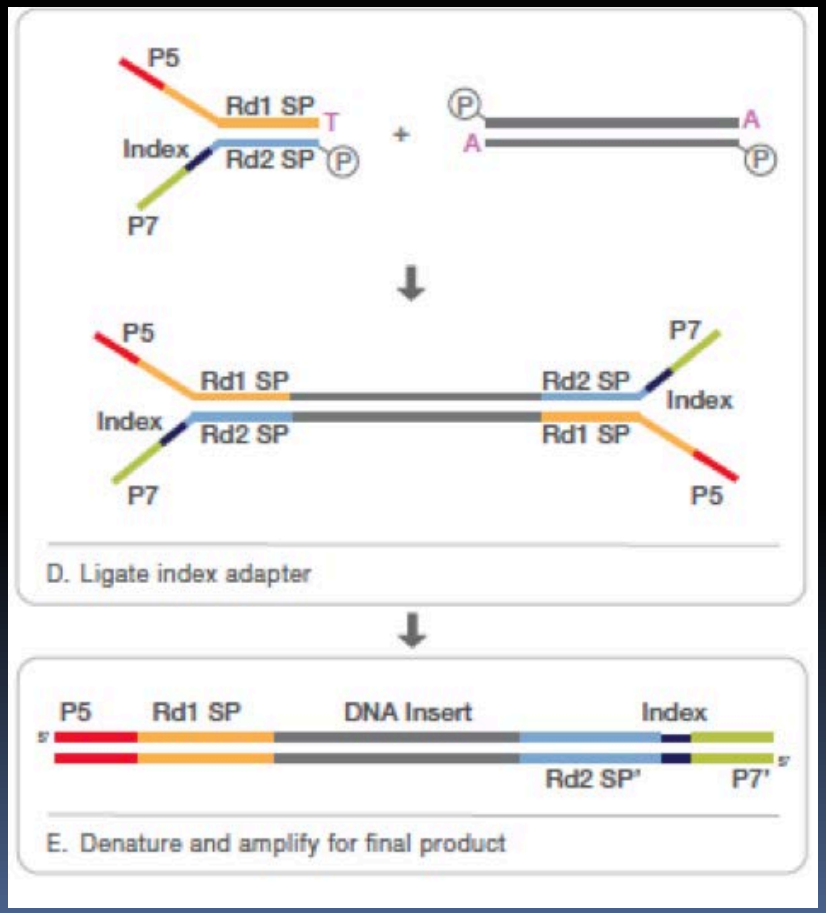
Illumina sequencing

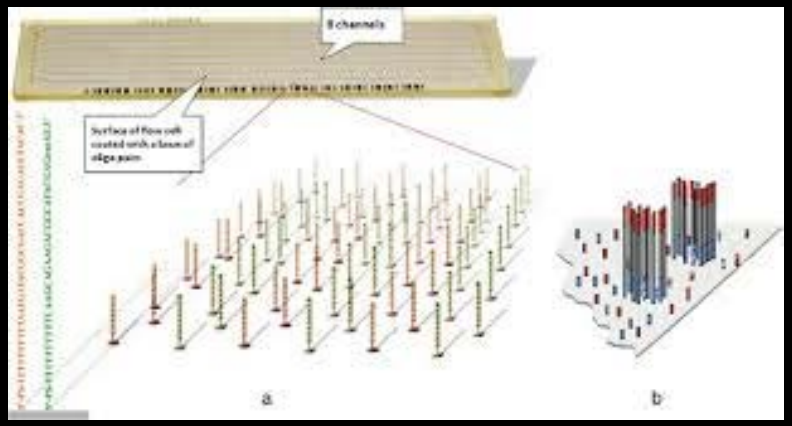
Library Construction

Prepare for adapter ligation



Adapter ligation

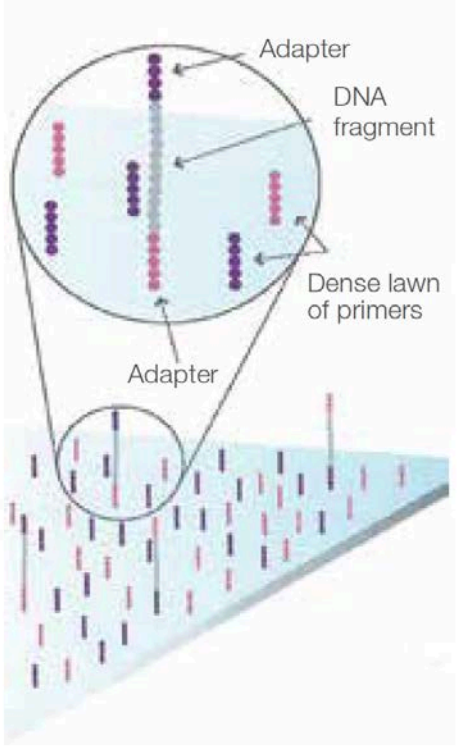




Flow cell with oligos

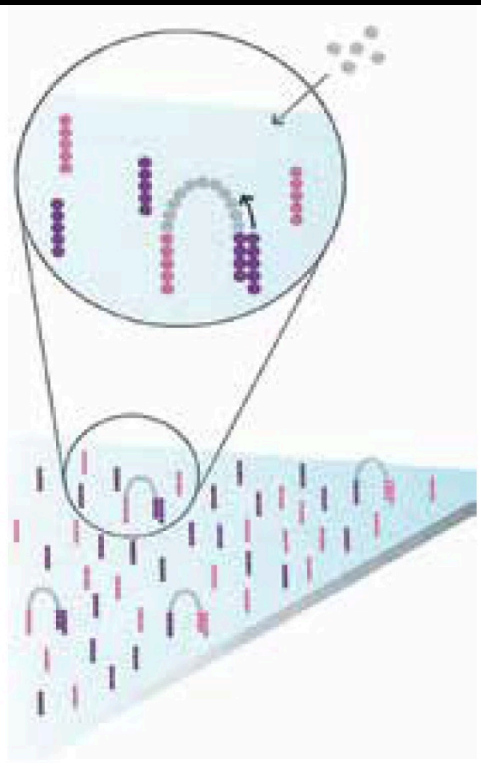
Illumina sequencing

Attach DNA to Surface



Bind single-stranded fragments randomly to the inside surface of the flow cell channels.

Bridge Amplification

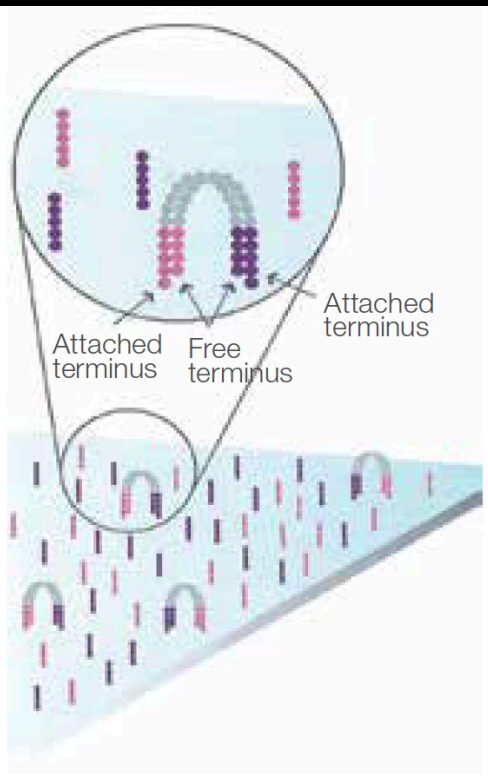


Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.

Illumina sequencing

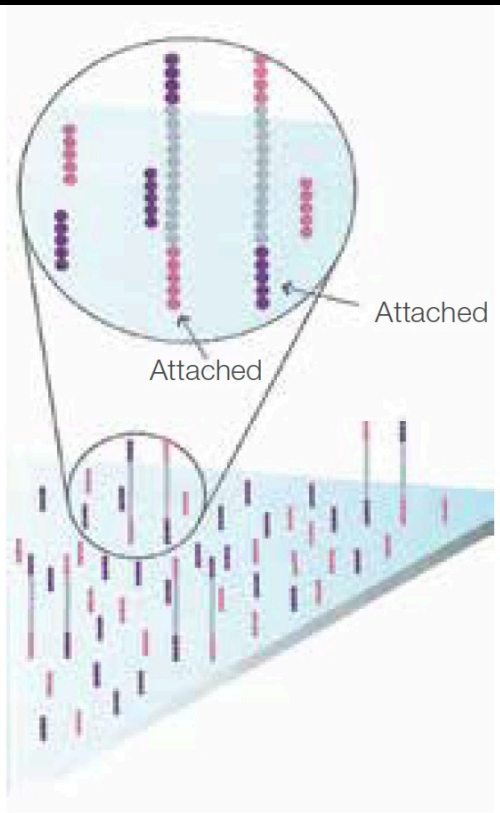
Bridge amplification

Fragments become double stranded



The enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.

Denature the ds molecules

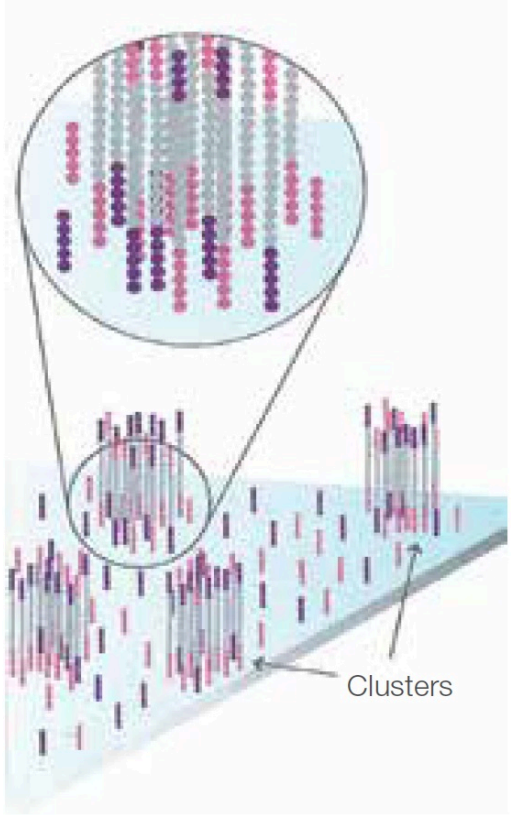


Denaturation leaves single-stranded templates anchored to the substrate.

Illumina sequencing

Bridge amplification

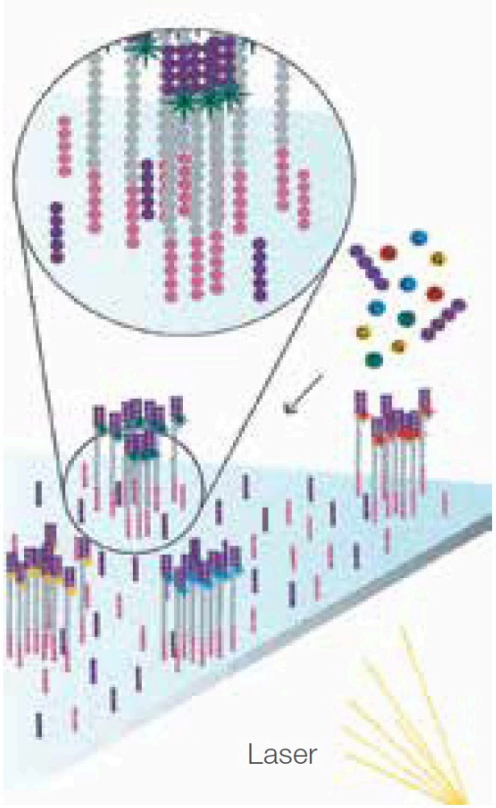
Complete Amplification



Several million dense clusters of double-stranded DNA are generated in each channel of the flow cell.

Sequencing by Synthesis

Determine 1st base

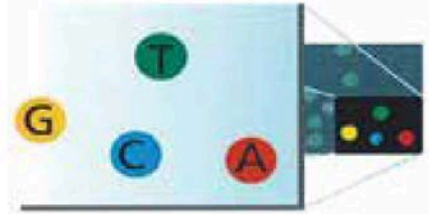


The first sequencing cycle begins by adding four labeled reversible terminators, primers, and DNA polymerase.

Illumina sequencing

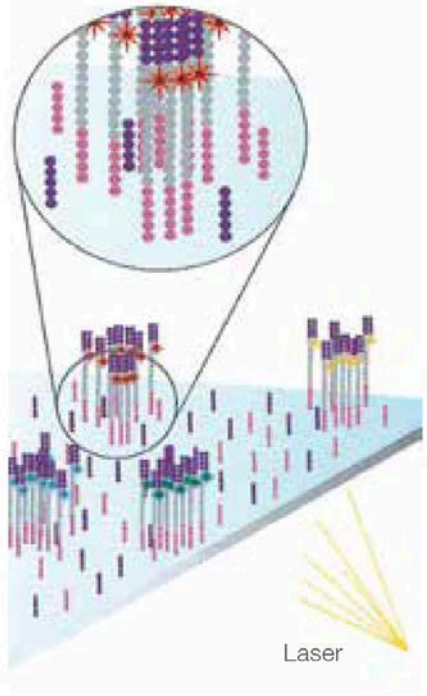
Sequencing by Synthesis

Image 1st base



After laser excitation, the emitted fluorescence from each cluster is captured and the first base is identified.

Determine 2nd base



The next cycle repeats the incorporation of four labeled reversible terminators, primers, and DNA polymerase.

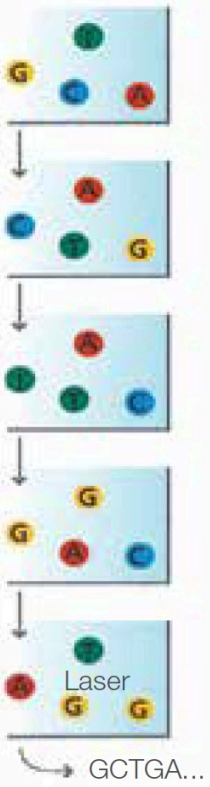
Illumina sequencing

Sequencing by Synthesis

Image 2nd base

After laser excitation, the image is captured as before, and the identity of the second base is recorded.

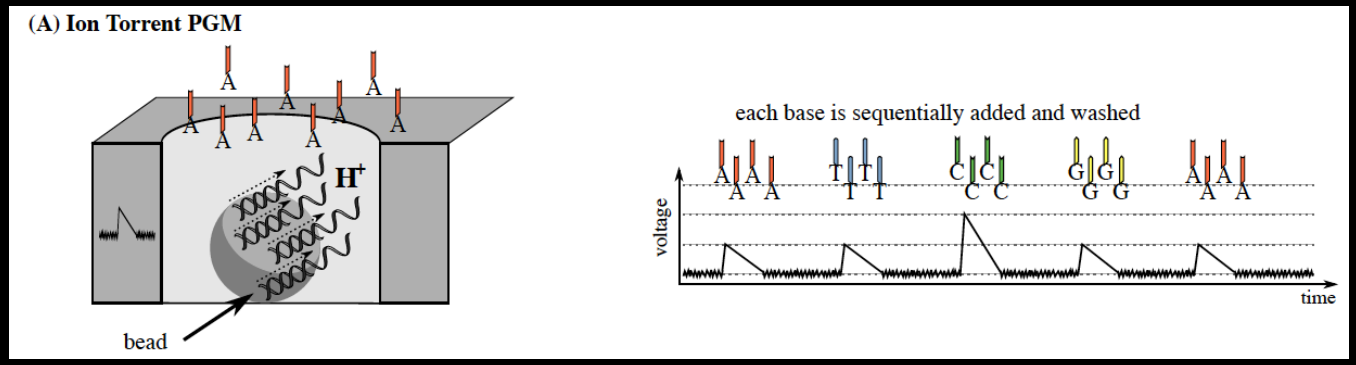
Sequence over multiple Cycles



The sequencing cycles are repeated to determine the sequence of bases in a fragment, one base at a time.

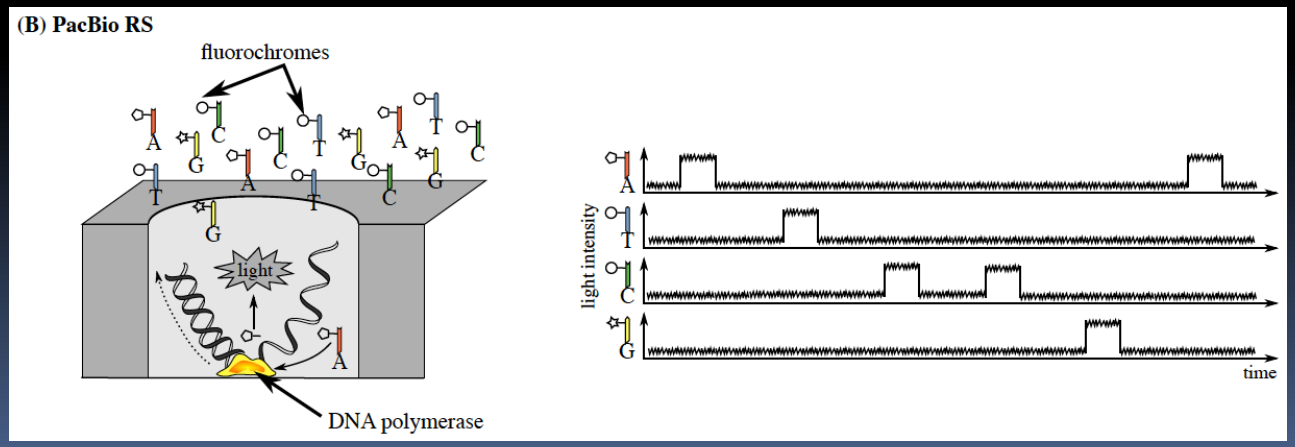
Other Sequencing Technologies

Emulsion PCR, electrical detection of pH change



Ion Torrent

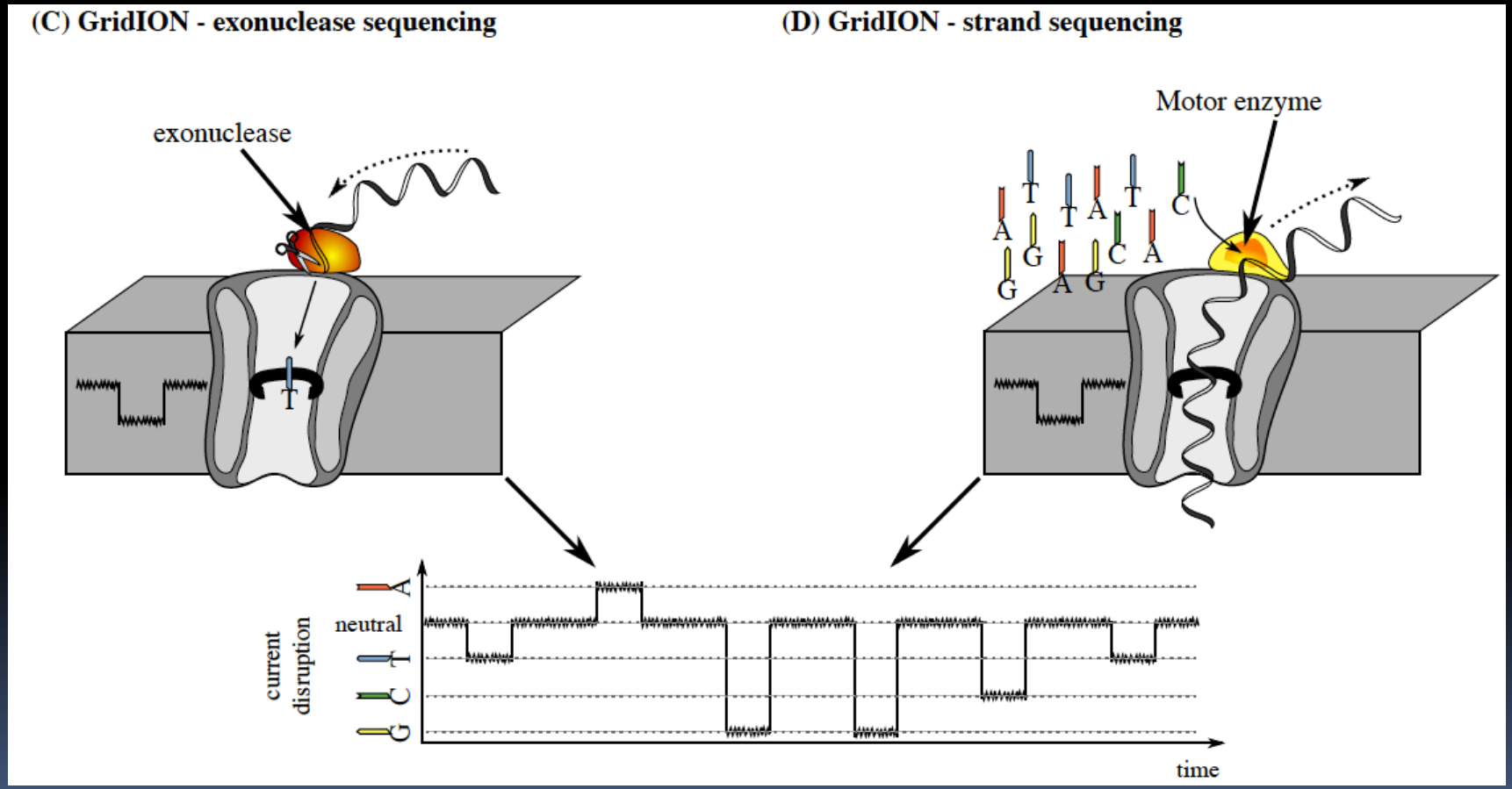
Single cell, optical detection, long reads



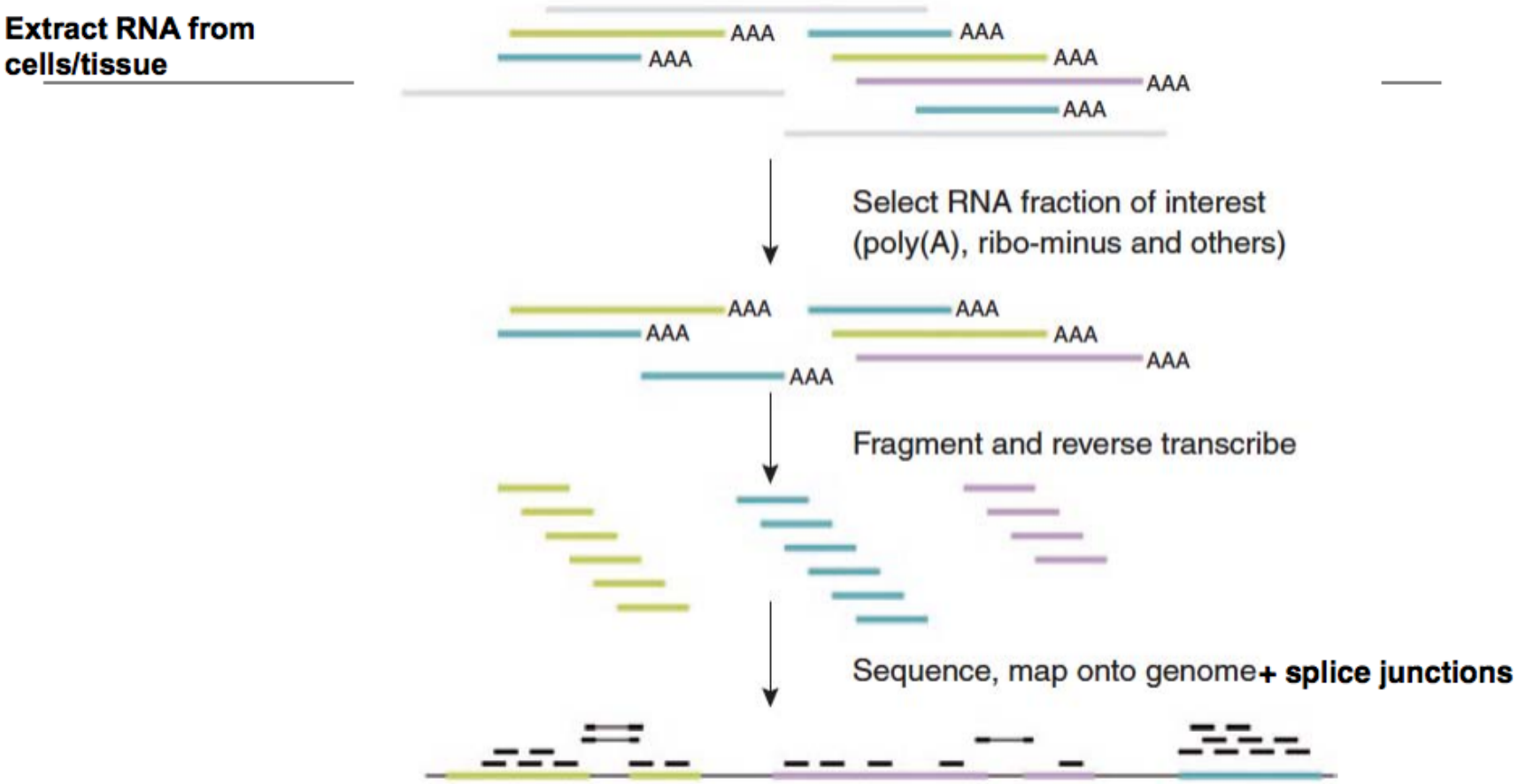
PacBio

Other Sequencing Technologies

Single cell, electrical detection, long reads Oxford Nanopore



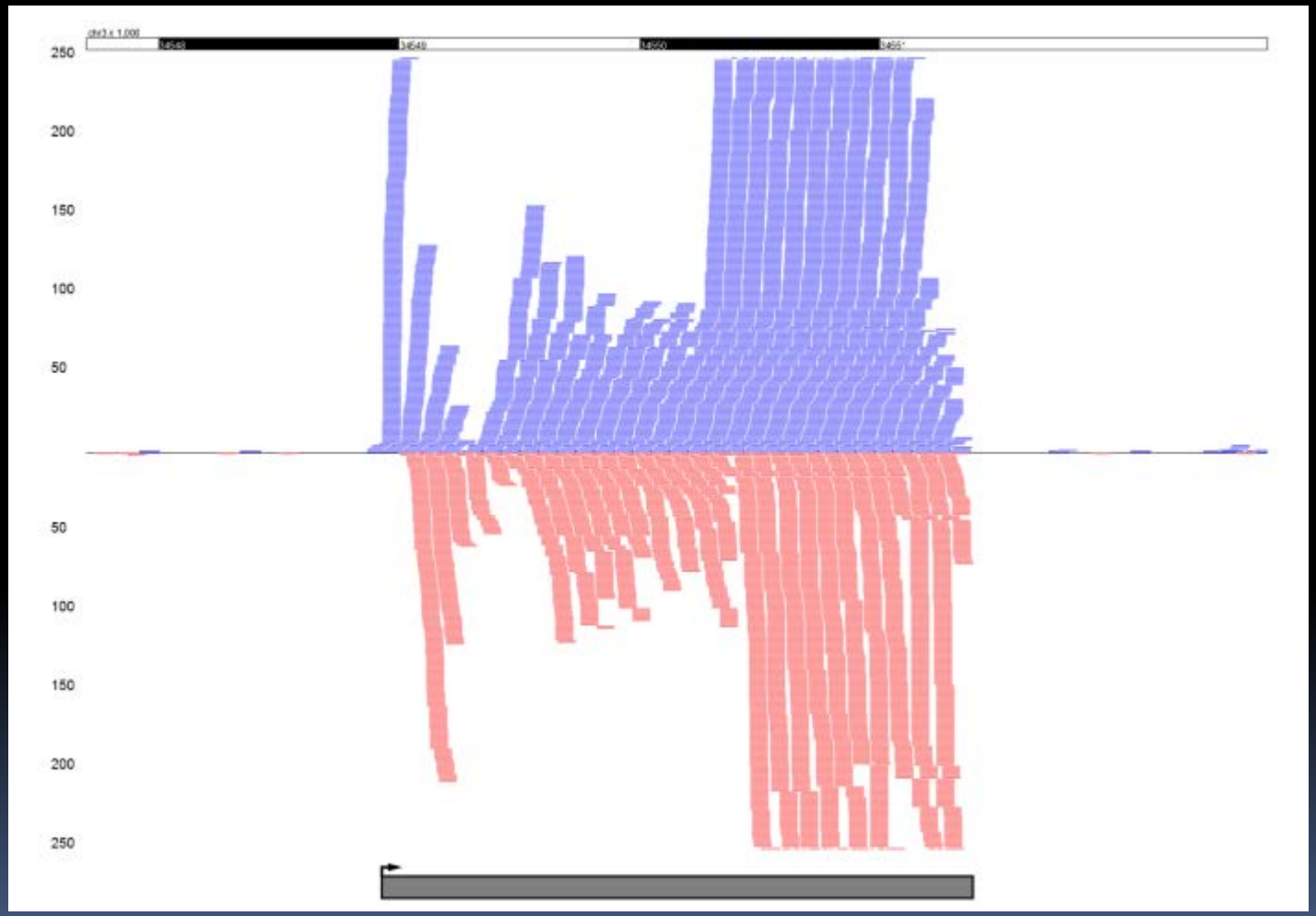
RNA-Seq: millions of short reads from fragmented mRNA



Source: Pepke, Shirley, Barbara Wold, et al. "Computation for ChIP-seq and RNA-seq Studies." *Nature Methods* 6 (2009): S22-32.

Pepke et. al. *Nature Methods* 2009

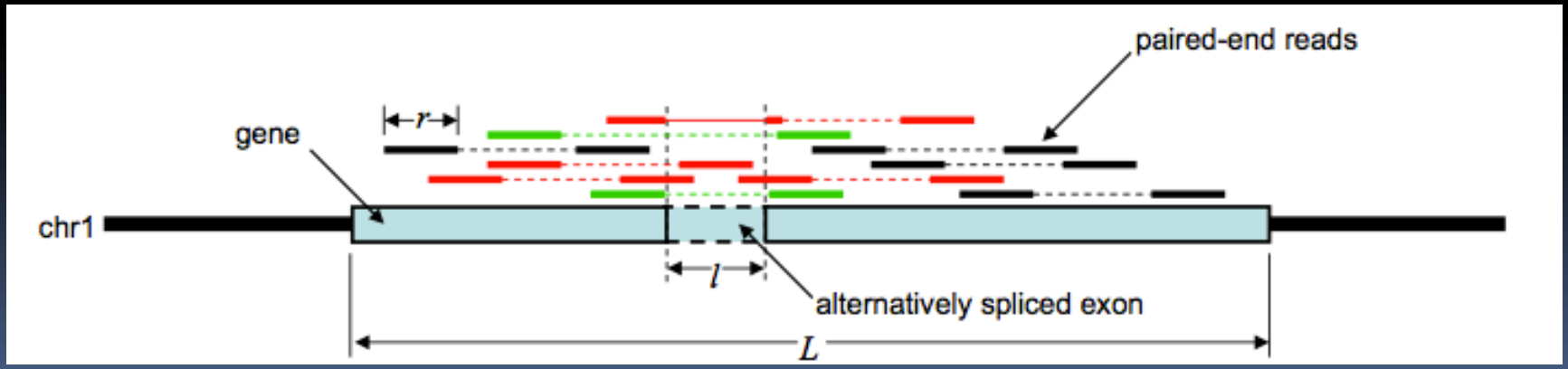
Mapping RNA-seq reads to a reference genome reveals expression



SOX2 Gene

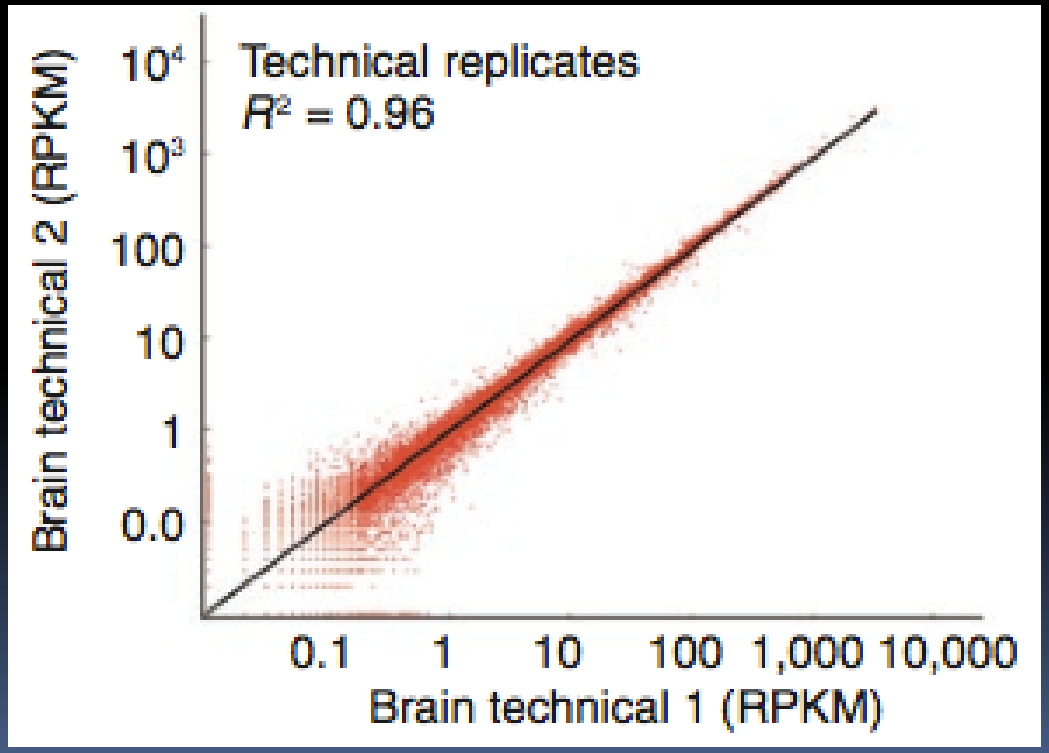
Units of RNA-seq

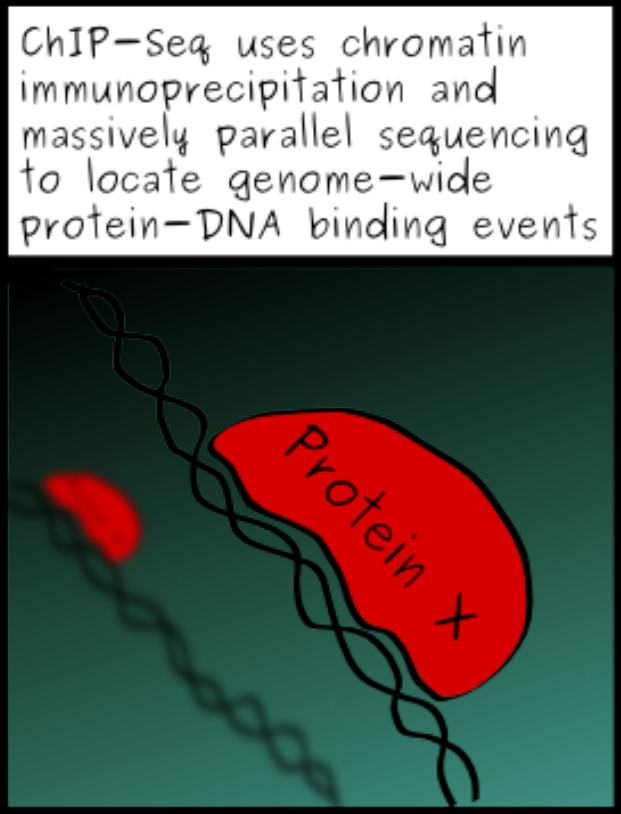
- More reads map to longer genes.
- If comparing different genes, use RPKM: Read Per Kilobase Transcript Per Million Reads.
- If comparing genes to genes across different patients: CPM or Counts Per Million reads (Out of 1M reads, how many mapped to a given gene.)



Noise characteristics

- Low technical noise (~Poisson distribution)
- Biological noise can be big

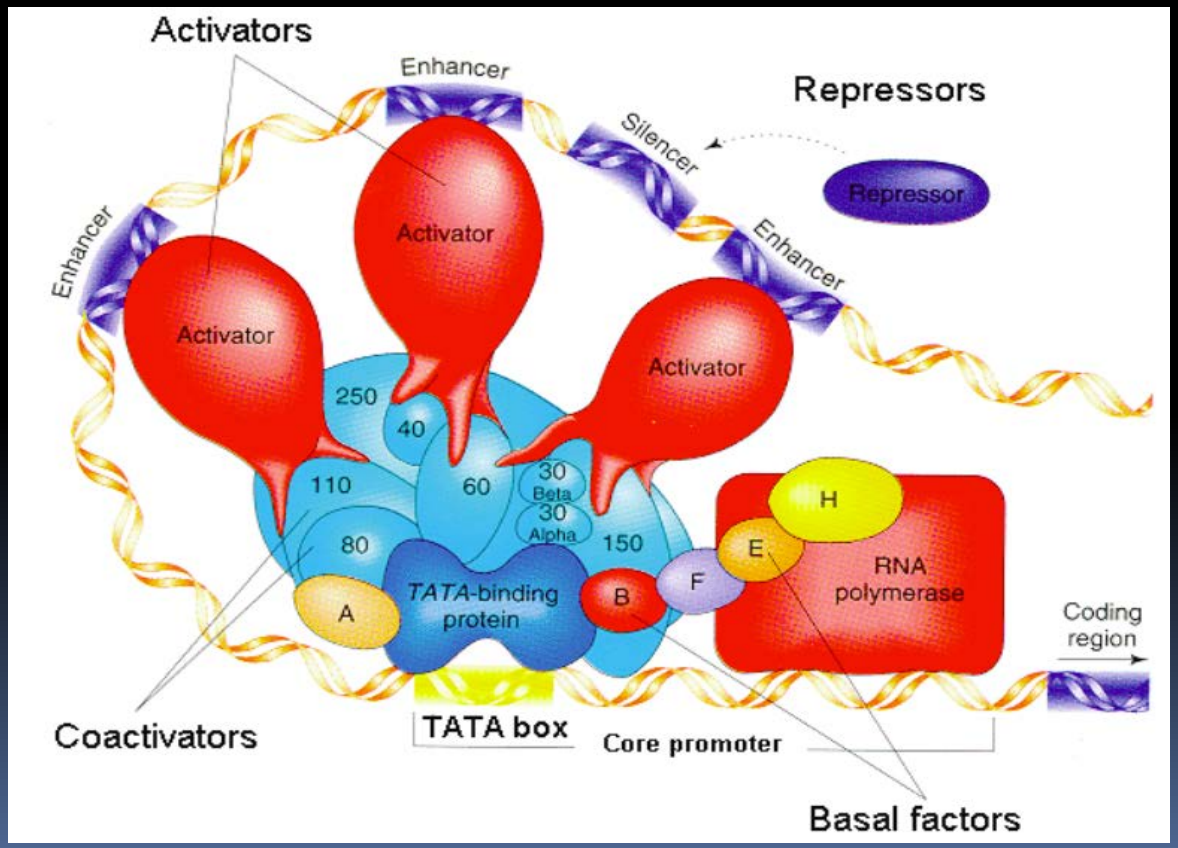



ChIP-Seq uses chromatin immunoprecipitation and massively parallel sequencing to locate genome-wide protein-DNA binding events
Protein X

ChIP-seq

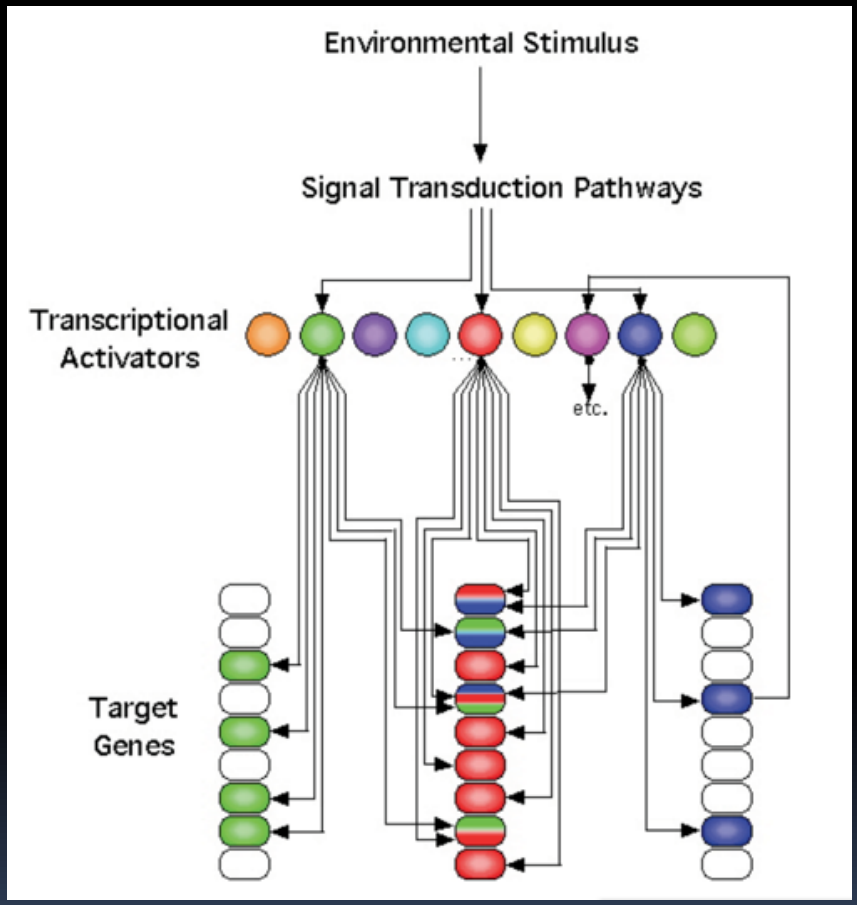
Regulatory Genomics and the Biology of Transcription Factors

- There are 1,500 TF in humans
- Transcription factor (TF) binds to DNA and controls transcription:
- promotes or represses the recruitment of the RNA polymerase



TF determine gene regulatory circuits

- There are 1,500 TF in humans
- They activate or silence target genes
- The connectivity of TFs to targets defines transcriptional regulation networks
- Many *network motifs* present such as:
  - Feed-forward loops (ensure signals)
  - Fan-outs (amplify signals)
  - Feed-back loops (create pulses)
  - see Uri Alon's work
- Networks reveal cell logic

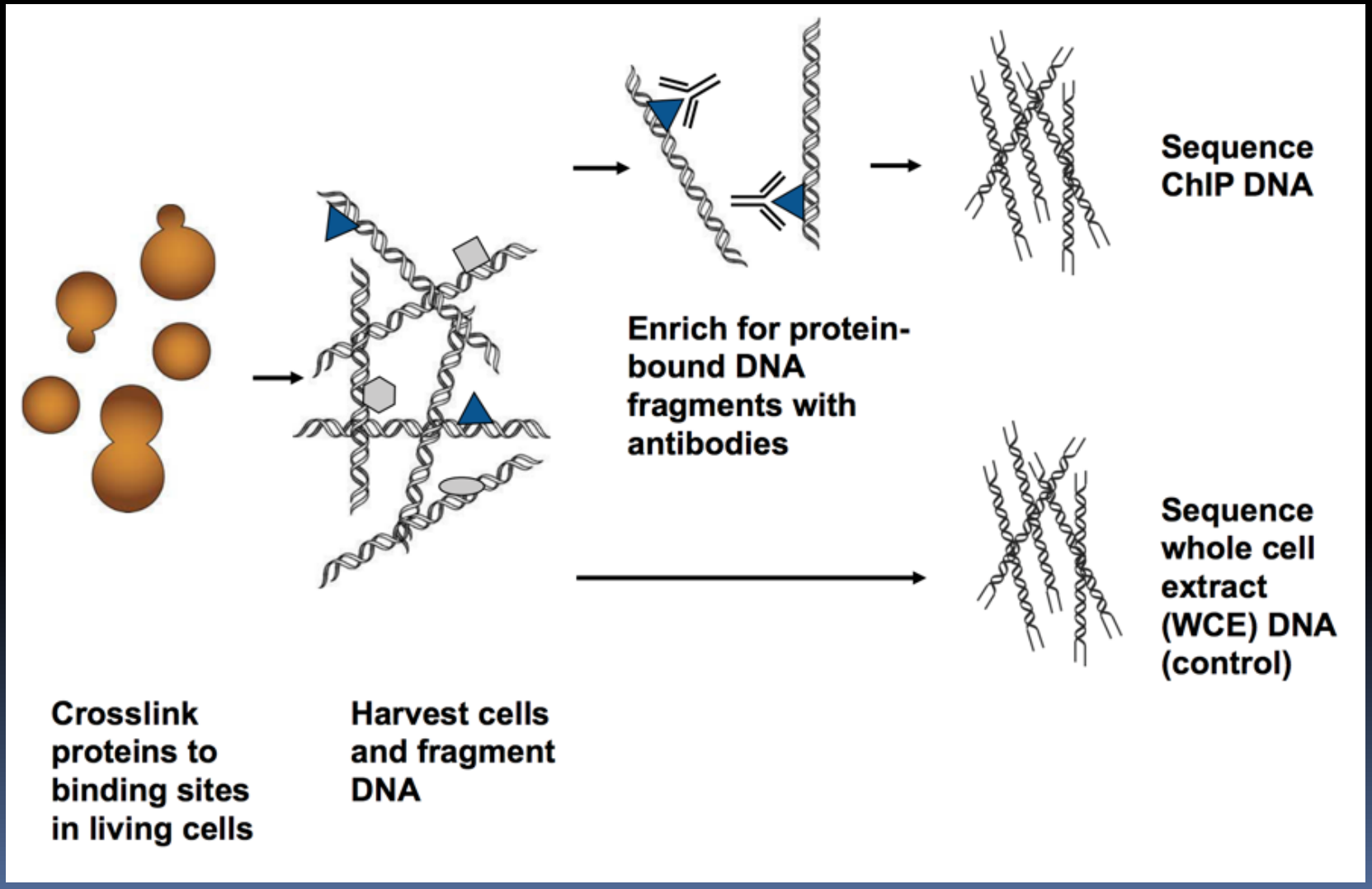


Rick Young, MIT
(Pioneer of ChIP-chip & ChIP-Seq)

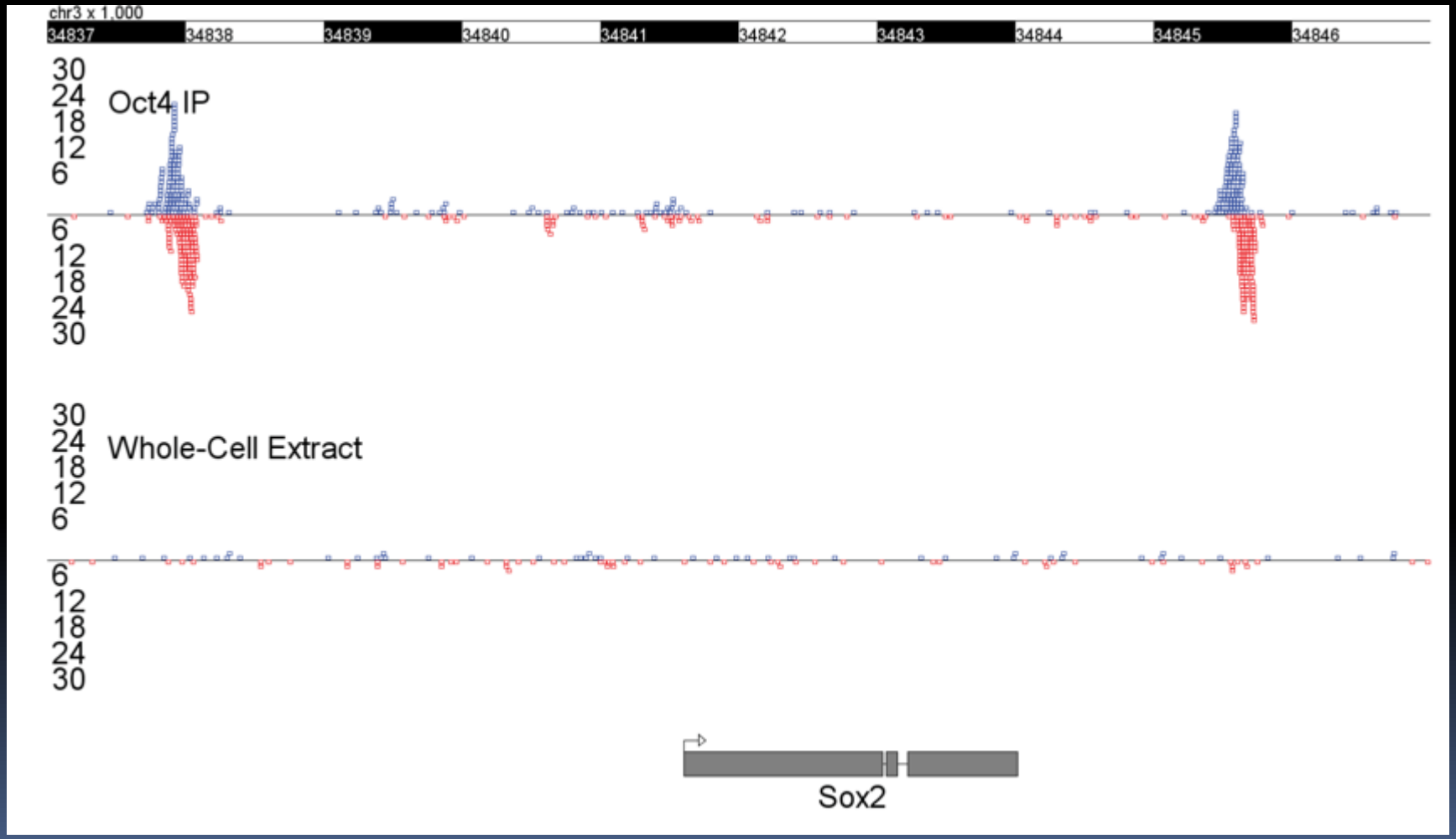
ChIP-Seq: study TF-DNA interactions

- ChIP-Seq: Chromatin Immuno-precipitation followed by sequencing
- Selects proteins out with an antibody specific to that protein
- Sequences any of the DNA that is “sticking” to the selected proteins.
- From the reads, can we identify where the proteins are binding

ChIP-Seq protocol



ChIP-Seq Example: OCT4 binding in *SOX2* Region in mouse ES cells



Slide from David Gifford, MIT OpenCourseWare

The ENCODE Project https://www.encodeproject.org

ARTICLE

doi:10.1038/nature11247

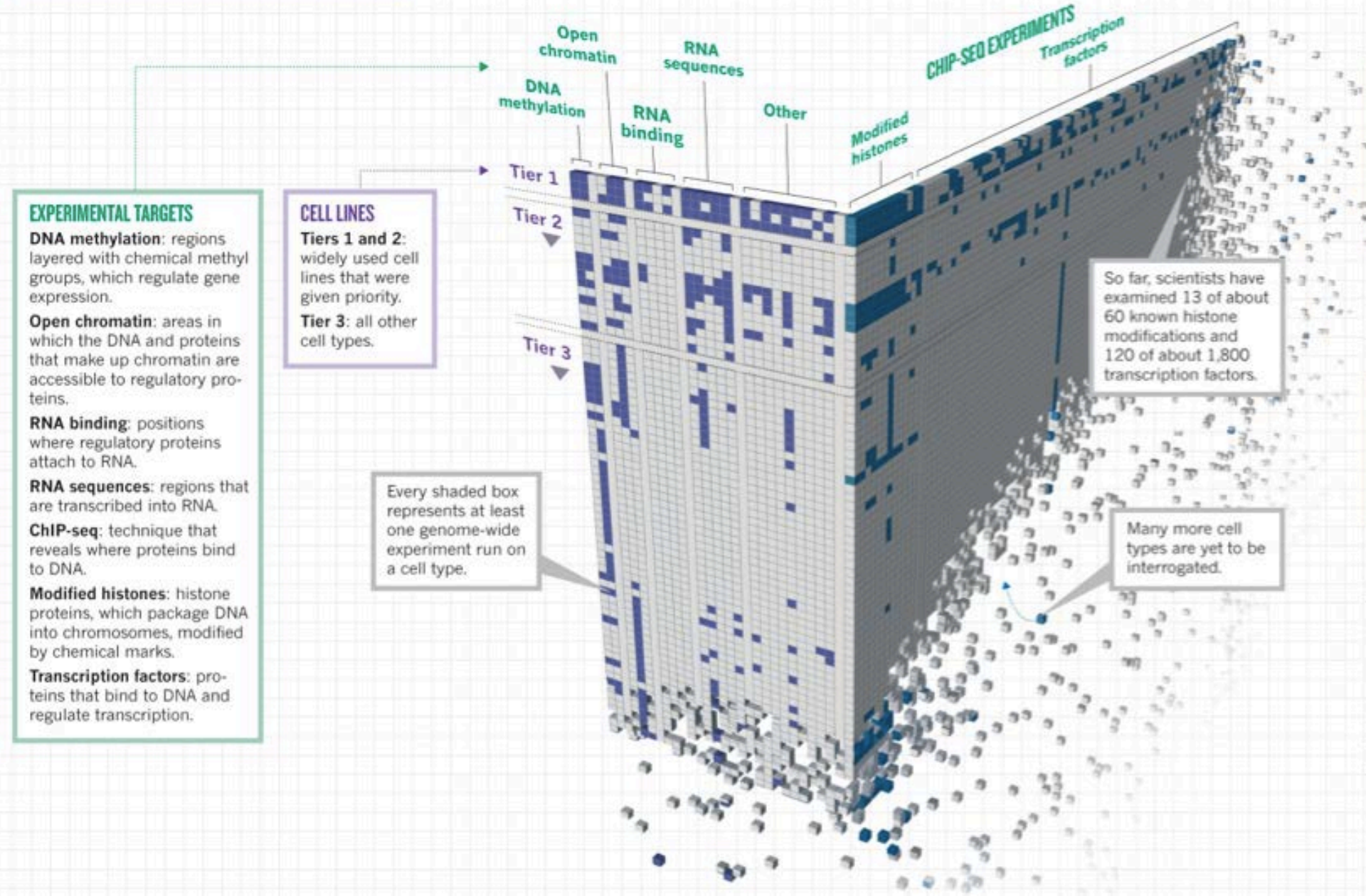
An integrated encyclopedia of DNA elements in the human genome

The ENCODE Project Consortium*

The human genome encodes the blueprint of life, but the function of the vast majority of its nearly three billion bases is unknown. The Encyclopedia of DNA Elements (ENCODE) project has systematically mapped regions of transcription, transcription factor association, chromatin structure and histone modification. These data enabled us to assign biochemical functions for 80% of the genome, in particular outside of the well-studied protein-coding regions. Many discovered candidate regulatory elements are physically associated with one another and with expressed genes, providing new insights into the mechanisms of gene regulation. The newly identified elements also show a statistical correspondence to sequence variants linked to human disease, and can thereby guide interpretation of this variation. Overall, the project provides new insights into the organization and regulation of our genes and genome, and is an expansive resource of functional annotations for biomedical research.

MAKING A GENOME MANUAL

Scientists in the Encyclopedia of DNA Elements Consortium have applied 24 experiment types (across) to more than 150 cell lines (down) to assign functions to as many DNA regions as possible — but the project is still far from complete.



EXPERIMENTAL TARGETS

DNA methylation: regions layered with chemical methyl groups, which regulate gene expression.

Open chromatin: areas in which the DNA and proteins that make up chromatin are accessible to regulatory proteins.

RNA binding: positions where regulatory proteins attach to RNA.

RNA sequences: regions that are transcribed into RNA.

ChIP-seq: technique that reveals where proteins bind to DNA.

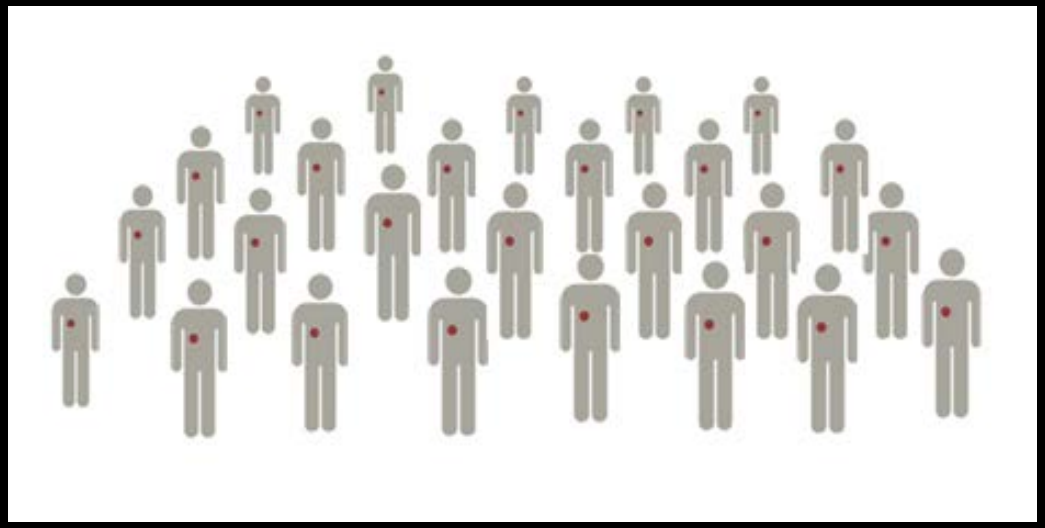
Modified histones: histone proteins, which package DNA into chromosomes, modified by chemical marks.

Transcription factors: proteins that bind to DNA and regulate transcription.

CELL LINES

Tiers 1 and 2: widely used cell lines that were given priority.

Tier 3: all other cell types.



Cancer omics: Learning from patient cohorts

The Cancer Genome Atlas (TCGA)

A resource of matched tumor and normal tissues from 11,000 patients with 12 cancer types

- Cervical cancer
- Cholangiocarcinoma
- Esophageal carcinoma
- Liver hepatocellular carcinoma
- Mesothelioma
- Pancreatic ductal adenocarcinoma
- Paraganglioma & Pheochromocytoma
- Sarcoma
- Testicular germ cell cancer
- Thymoma
- Uterine carcinosarcoma
- Uveal melanoma

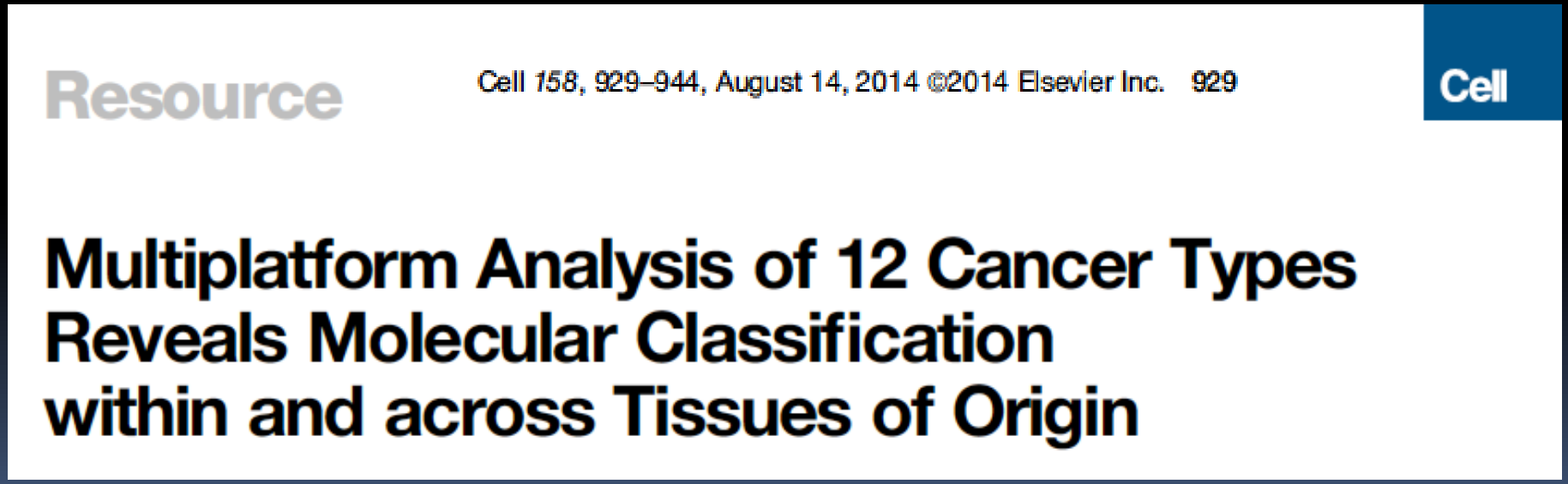
A lot of data available. Go to

https://tcga-data.nci.nih.gov/tcga/tcgaDownload.jsp

To explore data download

The Cancer Genome Atlas (TCGA)

- The Cancer Genome Atlas (TCGA) Research Network has reported integrated genome-wide studies of twelve distinct malignancies in 3,527 cases


Resource

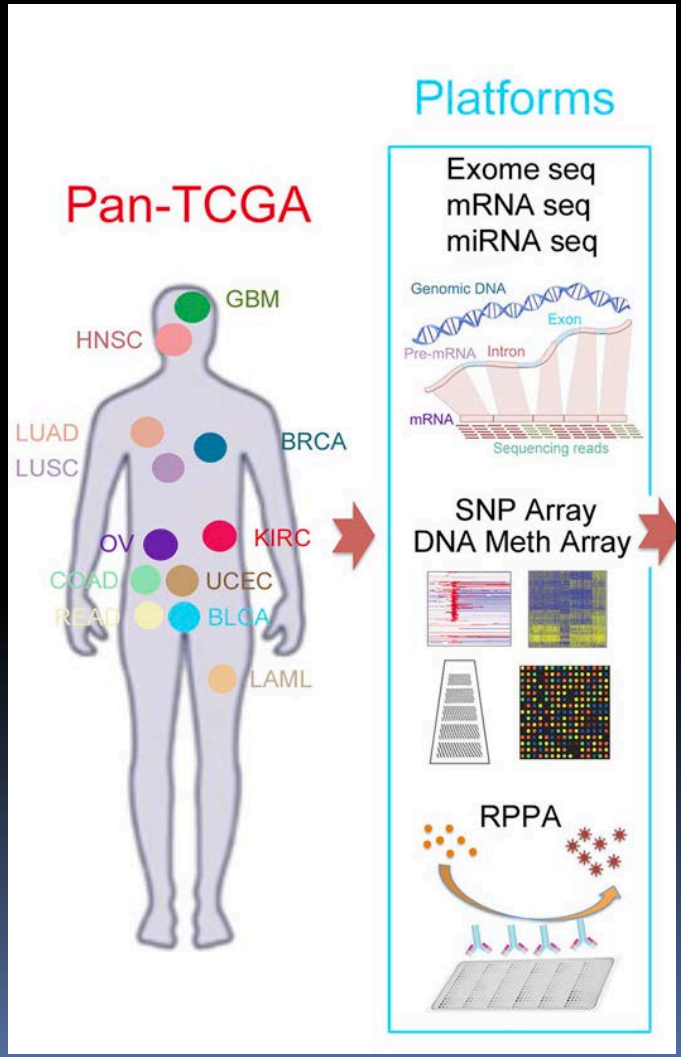
Cell *158*, 929–944, August 14, 2014 ©2014 Elsevier Inc. 929

Cell

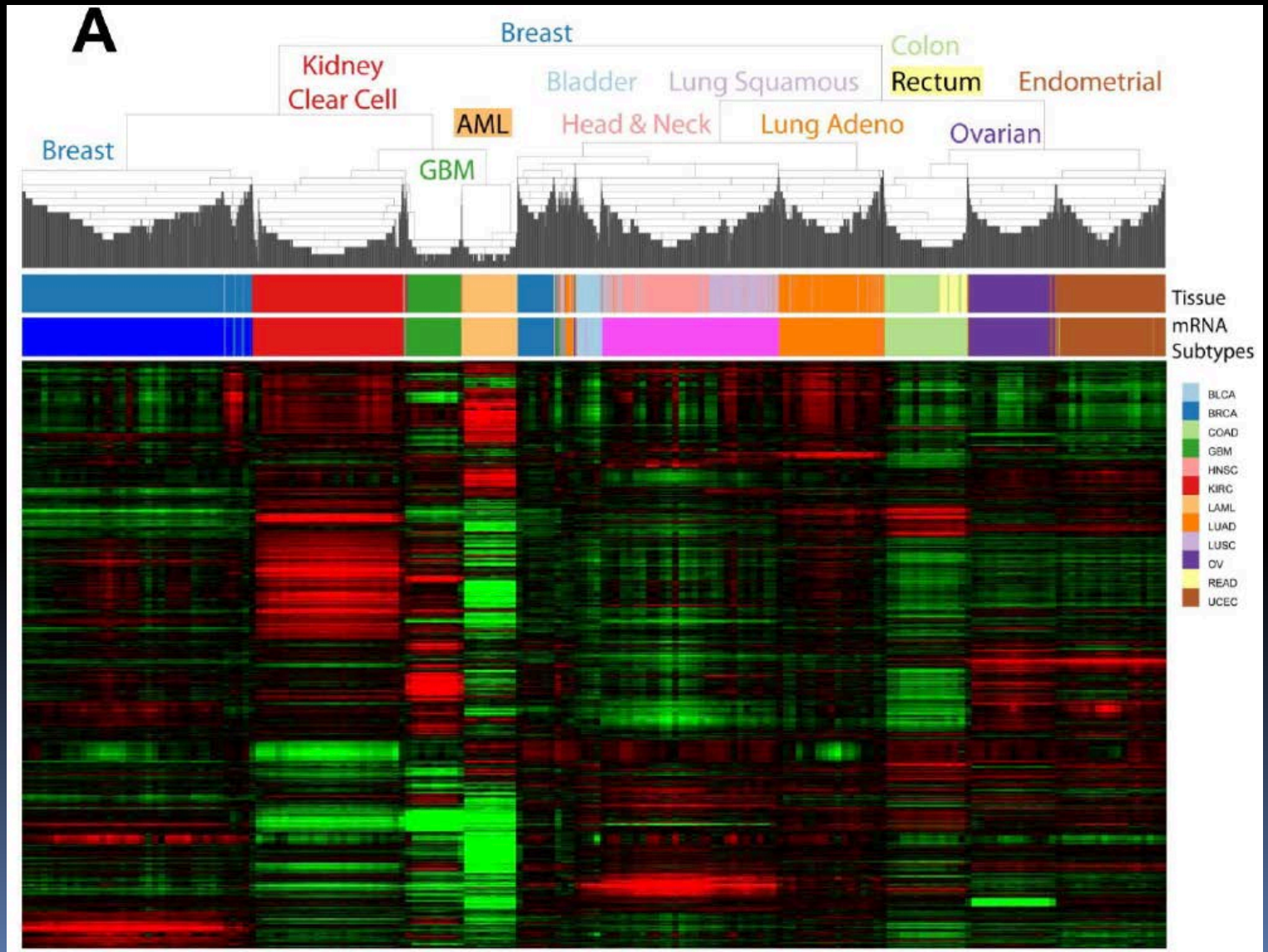
Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin

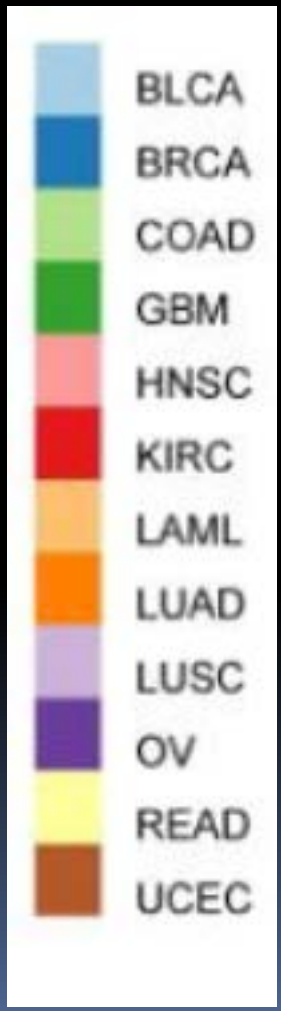
- Classical classification of cancer is based on cell of origin.
- Cancer genomics has found, additionally, that each tissue type can be further divided into 3 to 4 molecular subtypes

- This paper asks the question: Is there an alternative taxonomy beyond the tissue of origin? Based on 6 omics platforms:
- A pan-cancer classification.

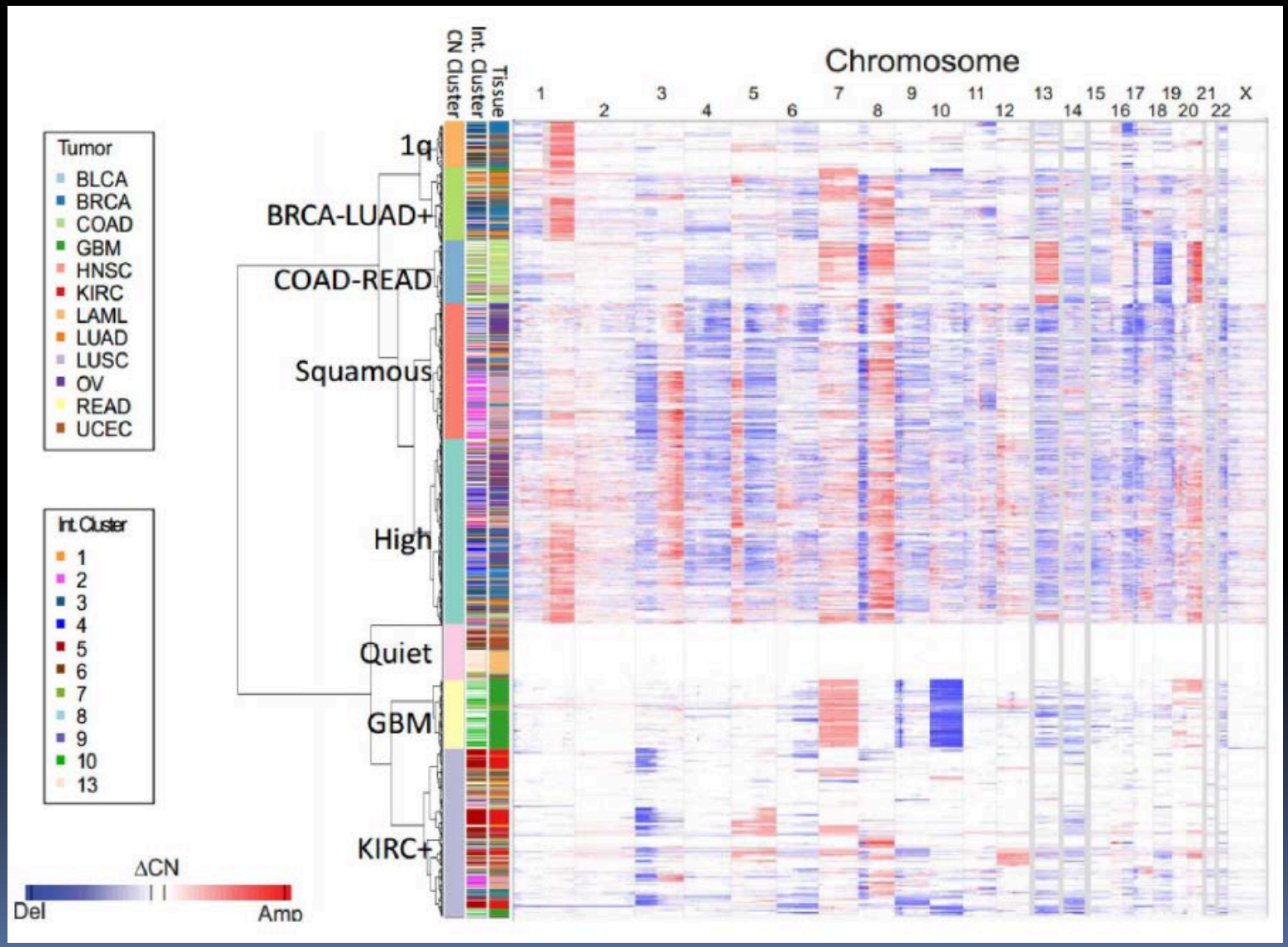


mRNA expression yielded 16 clusters of patients amongst the 12 tumor types





CNV yielded 8 clusters of patients amongst the 12 tumor types



How did they clustered using the 6 genomic platforms?

For each sample (patient) and each genomic platform the authors created a binary vector of size = # of clusters

Patient k cluster assignment in each platform

	CNV 1	CNV 1	CNV 3	CNV 4	CNV 5	CNV 6	CNV 7	CNV 8
CNV	0	0	0	0	1	0	0	0

RNA-seq	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0

……

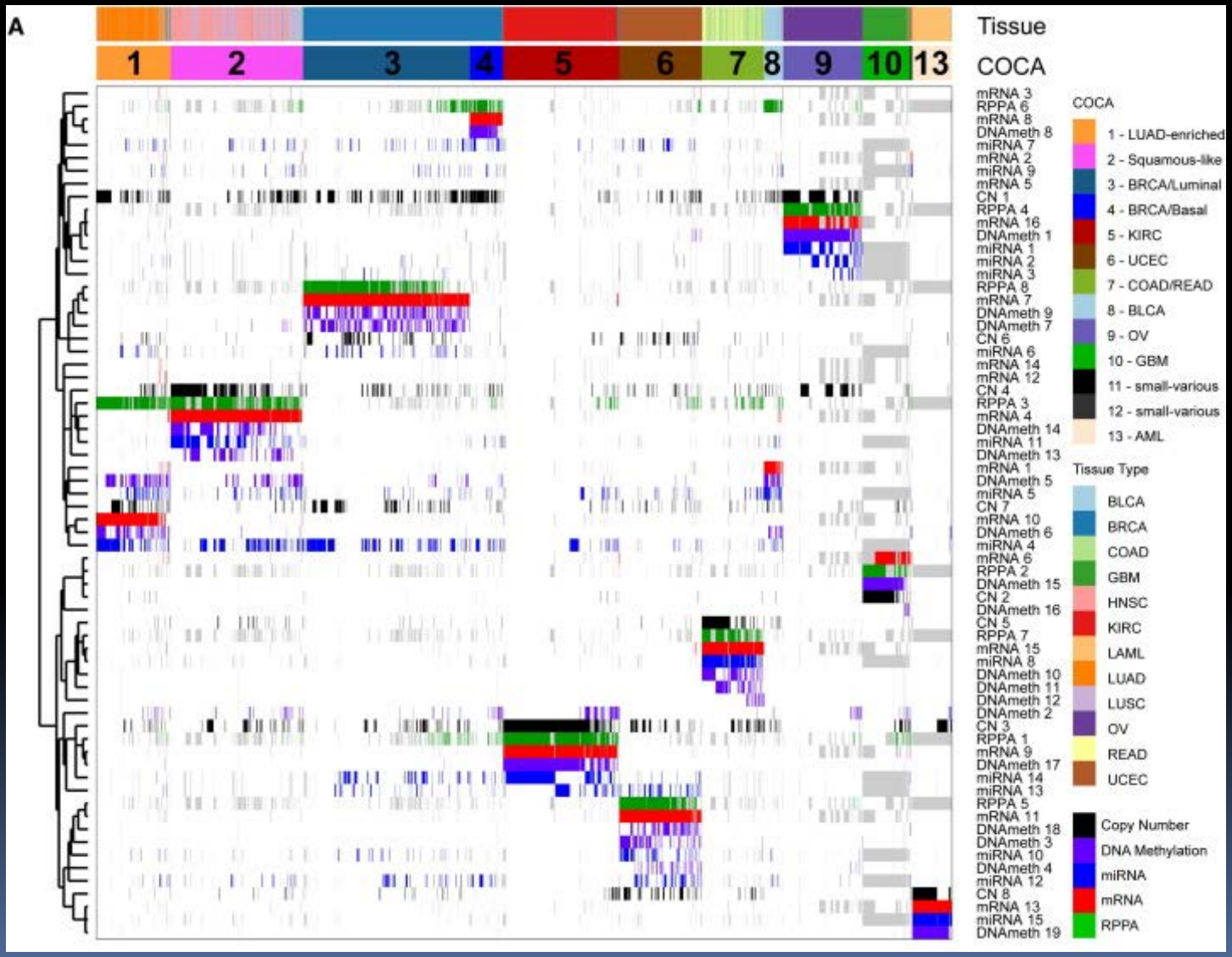
Then concatenate the clusters

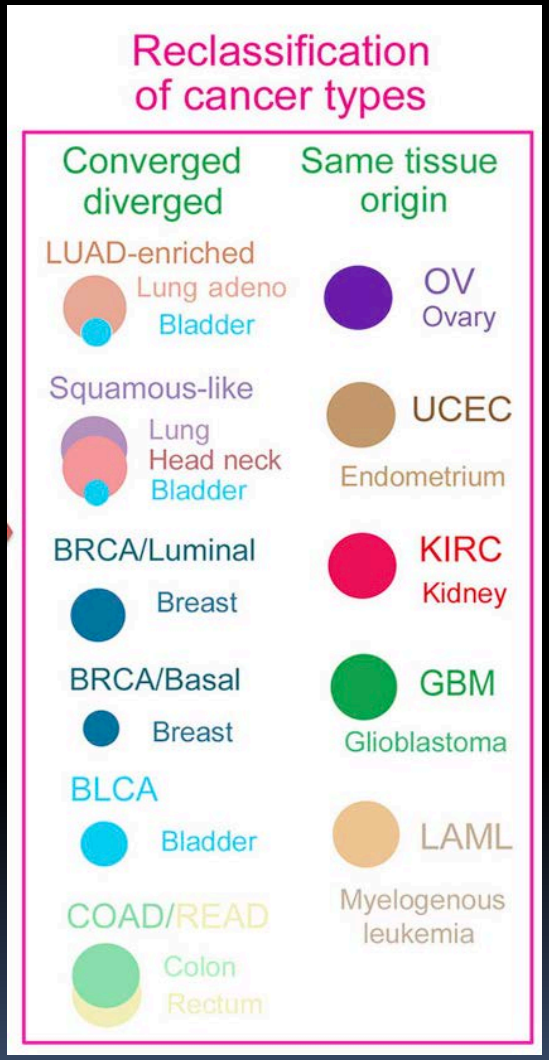
Patient k represented by binary vector across platforms

0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0

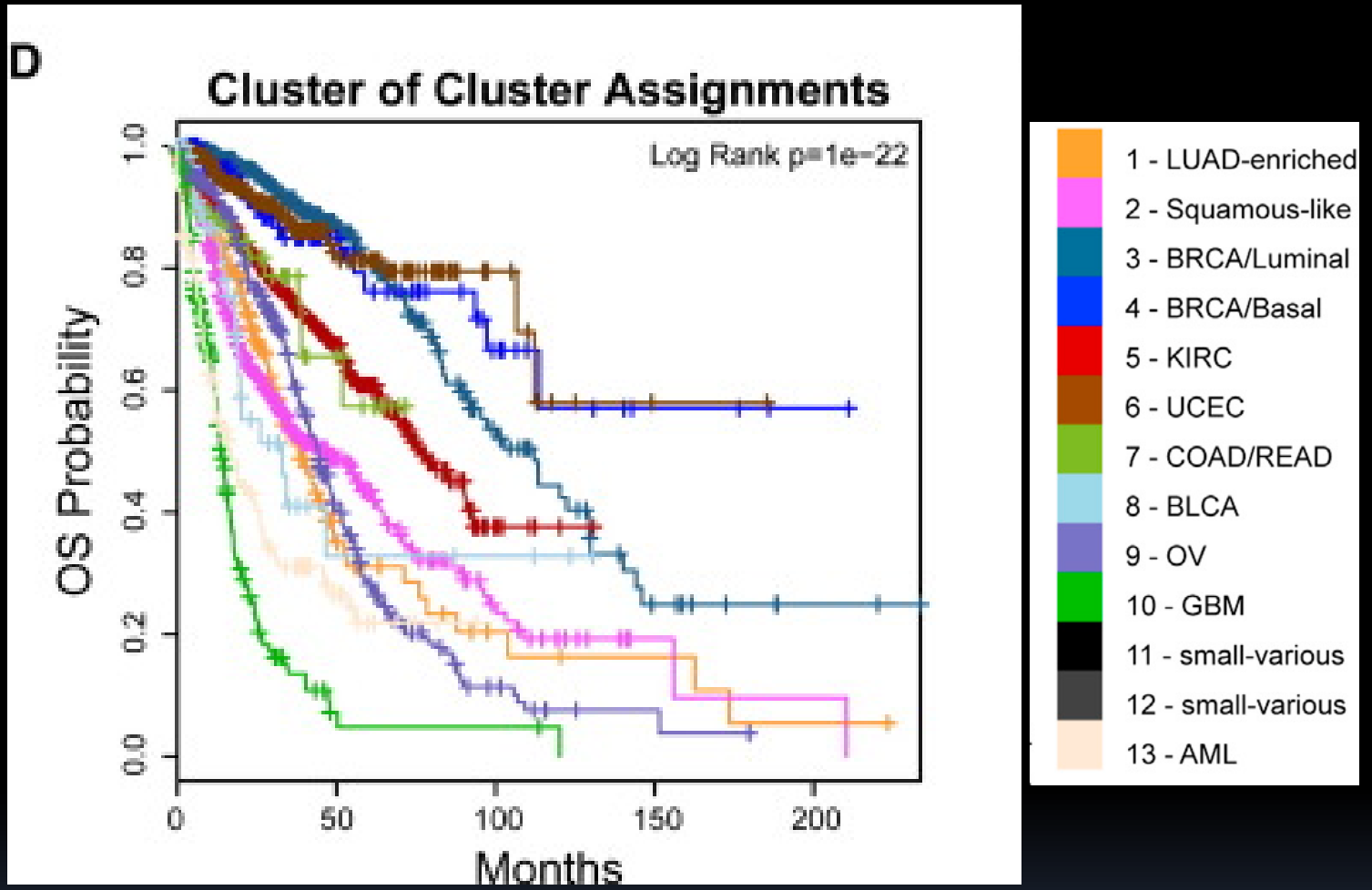
Perform patient clustering on the binary vectors

Consensus Clustering yielded 13 Pan Cancer clusters





- This paper's results suggest that "cell-of-origin" rather than pathway based features dominate the molecular taxonomy of diverse tumor types.
- However, based on this study, one in ten cancer patients would be classified differently by this new molecular taxonomy versus our current tissue-of-origin tumor classification system.



- If used to guide therapeutic decisions, this reclassification would affect a significant number of patients to be considered for nonstandard treatment regimens.

Proposed homework

Read: The Cancer Genome Atlas Research Network, *Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin*, Cell 158, 929–944, August 14, 2014. Bring 1 important take home message

Or

Read: Trapnell et. al, Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks, Nat Protoc. 1;7(3):562-78, March 2012. Try to make sense of the RNA-seq.

Or

Explore the TCGA (The Cancer Genome Atlas) (cancergenome.nih.gov) Data Portal (tcga-data.nci.nih.gov/tcga/tcgaHome2.jsp) dataportal. Try to download some files.