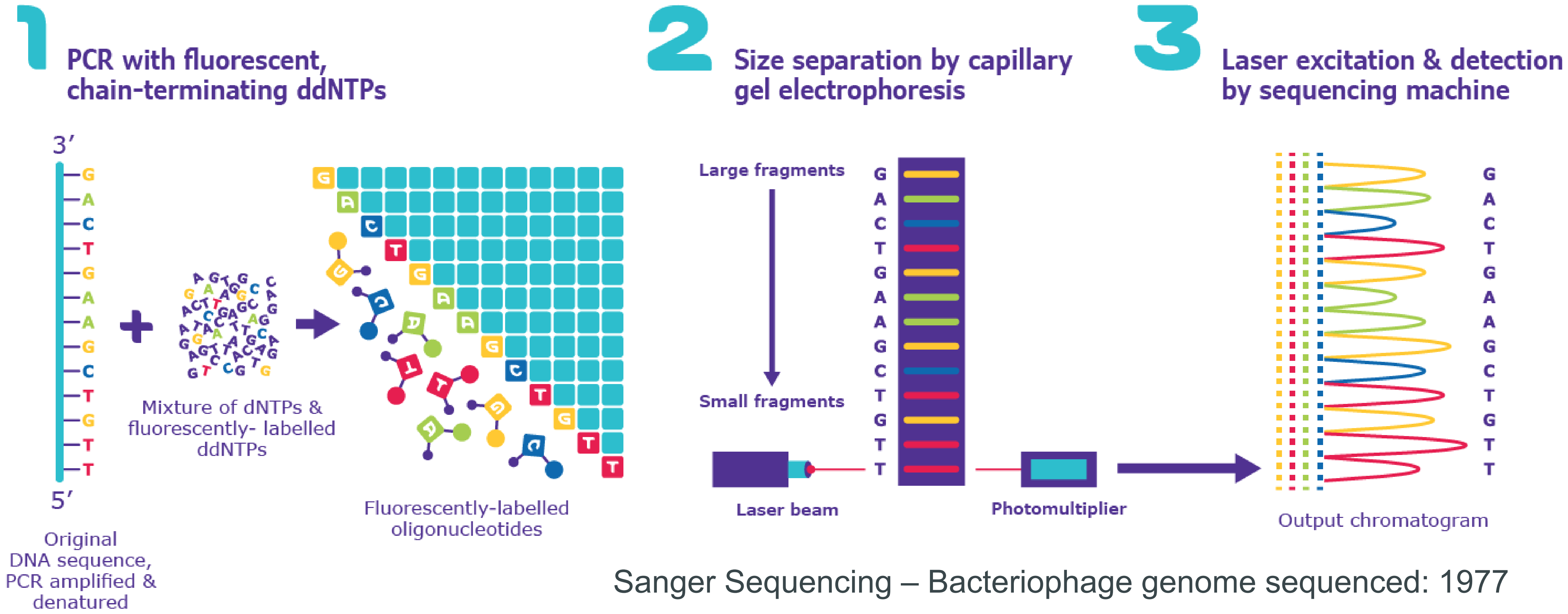
CEITEC

Central European Institute of Technology
BRNO | CZECH REPUBLIC

Introduction to Bioinformatics (LF:DSIB01)

Week 4 : Next Generation Sequencing: techniques and data

Nucleic Acid Sequencing History



1 PCR with fluorescent, chain-terminating ddNTPs

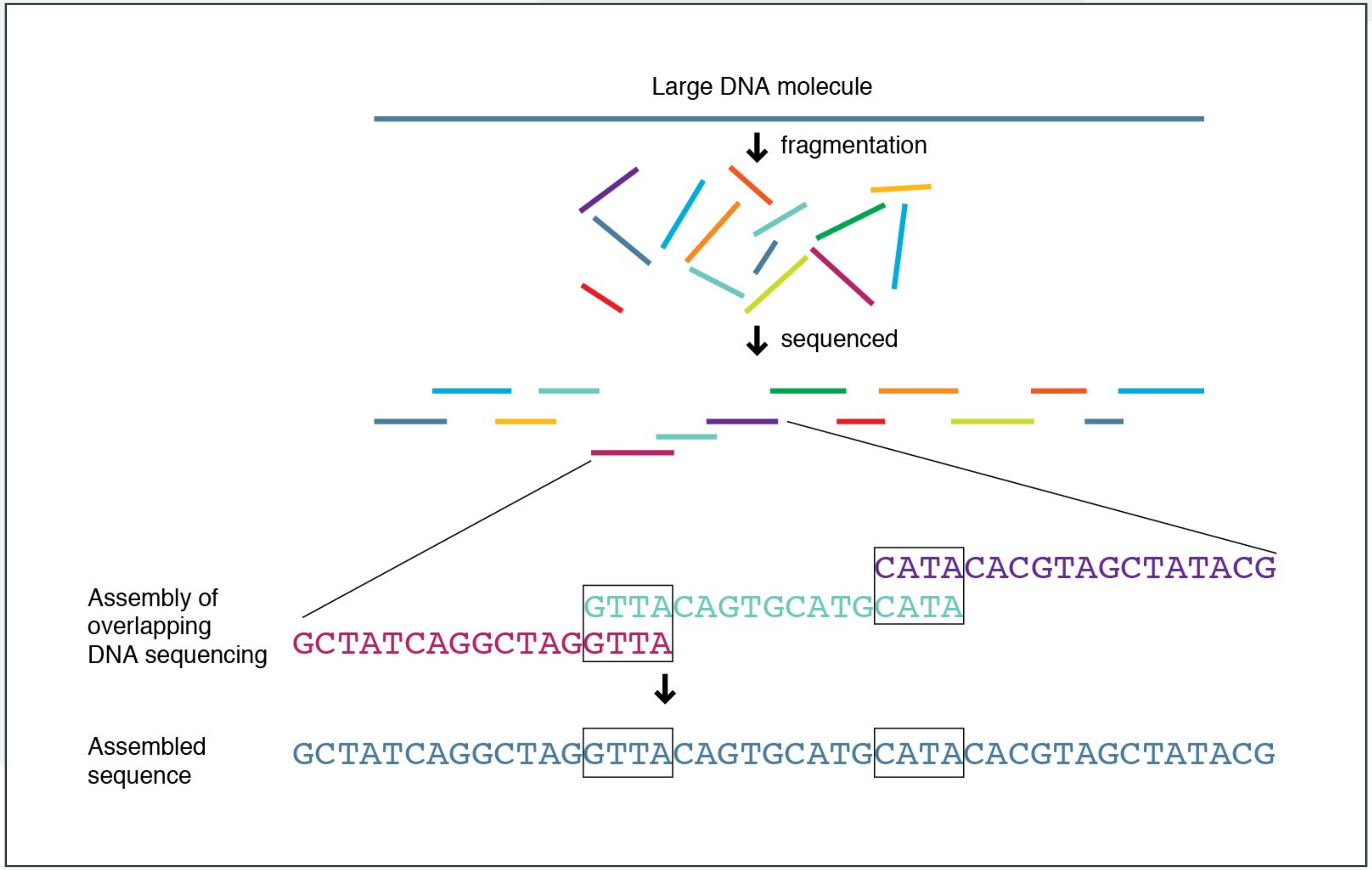
2 Size separation by capillary gel electrophoresis

3 Laser excitation & detection by sequencing machine

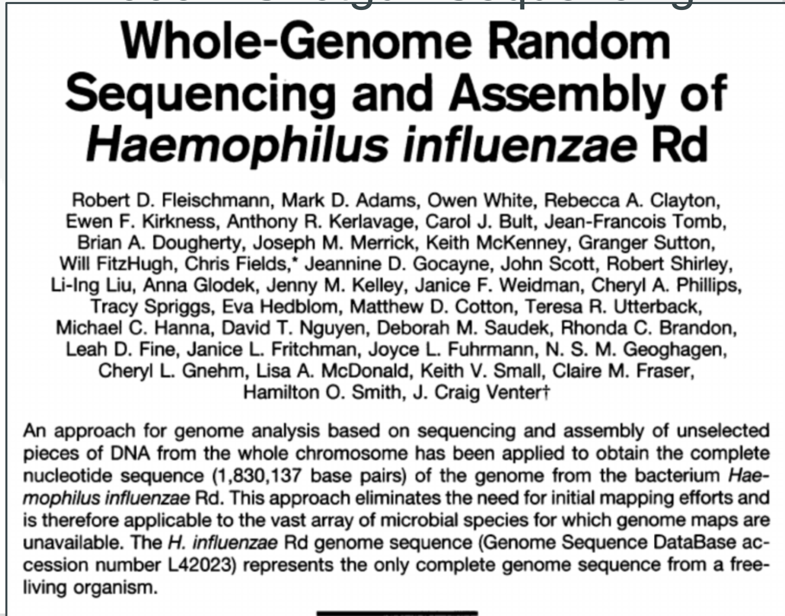
3′
G
A
C
T
G
A
A
G
C
T
G
T
T
5′
Original DNA sequence, PCR amplified & denatured
Mixture of dNTPs & fluorescently- labelled ddNTPs
Fluorescently-labelled oligonucleotides
Large fragments
Small fragments
Laser beam
Photomultiplier
Output chromatogram

Sanger Sequencing – Bacteriophage genome sequenced: 1977

Nucleic Acid Sequencing History



1995 – Shotgun Sequencing



Whole-Genome Random Sequencing and Assembly of *Haemophilus influenzae* Rd

Robert D. Fleischmann, Mark D. Adams, Owen White, Rebecca A. Clayton, Ewen F. Kirkness, Anthony R. Kerlavage, Carol J. Bult, Jean-Francois Tomb, Brian A. Dougherty, Joseph M. Merrick, Keith McKenney, Granger Sutton, Will FitzHugh, Chris Fields,* Jeannine D. Gocayne, John Scott, Robert Shirley, Li-Ing Liu, Anna Glodek, Jenny M. Kelley, Janice F. Weidman, Cheryl A. Phillips, Tracy Spriggs, Eva Hedblom, Matthew D. Cotton, Teresa R. Utterback, Michael C. Hanna, David T. Nguyen, Deborah M. Saudek, Rhonda C. Brandon, Leah D. Fine, Janice L. Fritchman, Joyce L. Fuhrmann, N. S. M. Geoghagen, Cheryl L. Gnehm, Lisa A. McDonald, Keith V. Small, Claire M. Fraser, Hamilton O. Smith, J. Craig Venter†

An approach for genome analysis based on sequencing and assembly of unselected pieces of DNA from the whole chromosome has been applied to obtain the complete nucleotide sequence (1,830,137 base pairs) of the genome from the bacterium *Haemophilus influenzae* Rd. This approach eliminates the need for initial mapping efforts and is therefore applicable to the vast array of microbial species for which genome maps are unavailable. The *H. influenzae* Rd genome sequence (Genome Sequence DataBase accession number L42023) represents the only complete genome sequence from a free-living organism.

Nucleic Acid Sequencing History

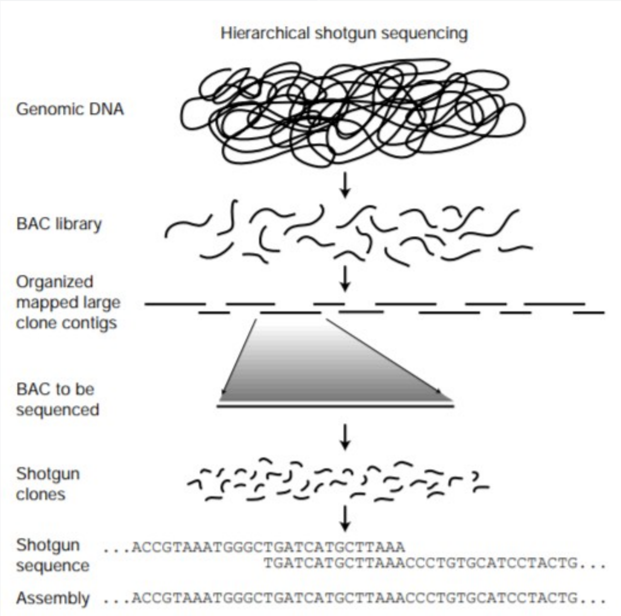
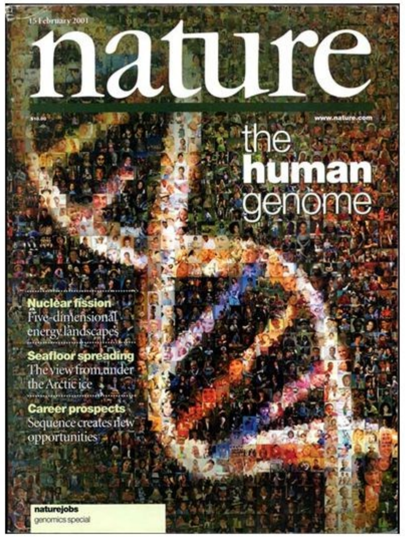
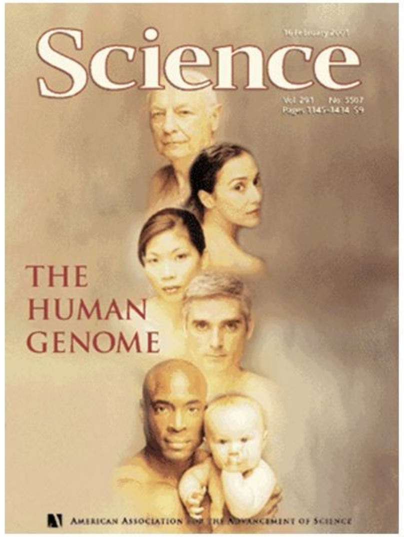

Hierarchical shotgun sequencing
Genomic DNA
BAC library
Organized mapped large clone contigs
BAC to be sequenced
Shotgun clones
Shotgun sequence
...ACCGTAAATGGGCTGATCATGCTTAAA
TGATCATGCTTAAACCCTGTGCATCCTACTG...
Assembly
...ACCGTAAATGGGCTGATCATGCTTAAACCCTGTGCATCCTACTG...

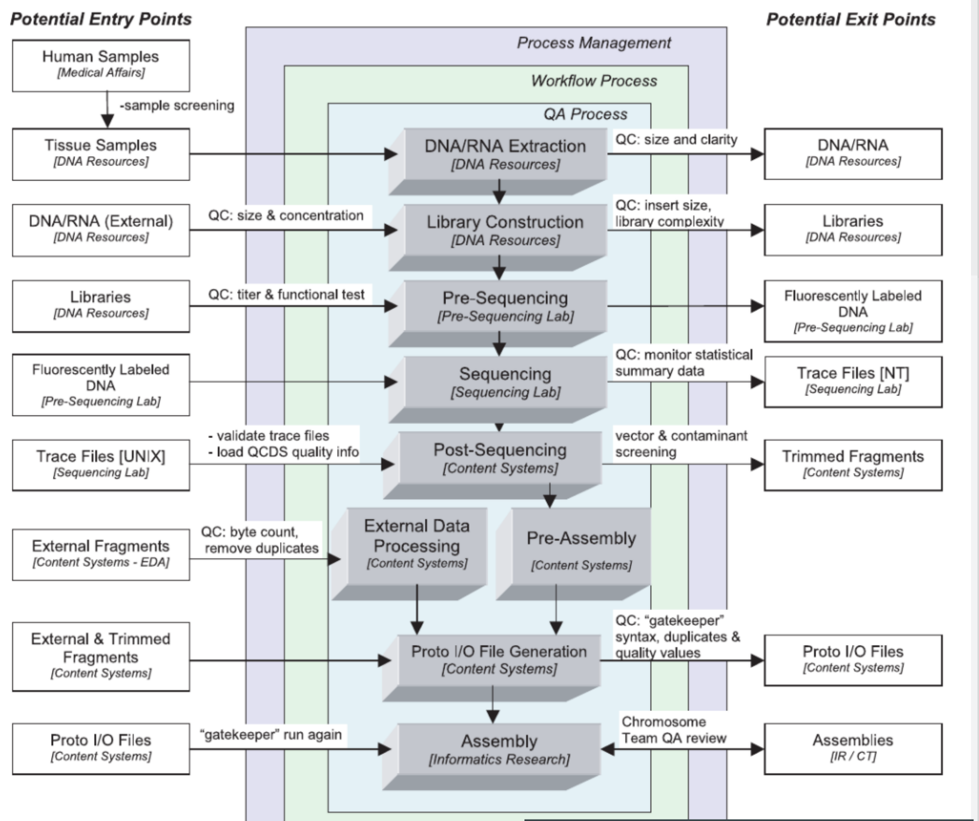
Figure 2 Idealized representation of the hierarchical shotgun sequencing strategy. A library is constructed by fragmenting the target genome and cloning it into a large-fragment cloning vector; here, BAC vectors are shown. The genomic DNA fragments represented in the library are then organized into a physical map and individual BAC clones are selected and sequenced by the random shotgun strategy. Finally, the clone sequences are assembled to reconstruct the sequence of the genome.

Genome sequencing

February 2001 - Publication of the first draft of the human genome







Potential Entry Points

Potential Exit Points

Process Management

Workflow Process

QA Process

Human Samples
[Medical Affairs]

-sample screening

Tissue Samples
[DNA Resources]

DNA/RNA Extraction
[DNA Resources]

QC: size and clarity

DNA/RNA
[DNA Resources]

DNA/RNA (External)
[DNA Resources]

QC: size & concentration

Library Construction
[DNA Resources]

QC: insert size, library complexity

Libraries
[DNA Resources]

Libraries
[DNA Resources]

QC: titer & functional test

Pre-Sequencing
[Pre-Sequencing Lab]

Fluorescently Labeled DNA
[Pre-Sequencing Lab]

Fluorescently Labeled DNA
[Pre-Sequencing Lab]

Sequencing
[Sequencing Lab]

QC: monitor statistical summary data

Trace Files [NT]
[Sequencing Lab]

Trace Files [UNIX]
[Sequencing Lab]

- validate trace files
- load QCDS quality info

Post-Sequencing
[Content Systems]

vector & contaminant screening

Trimmed Fragments
[Content Systems]

External Fragments
[Content Systems - EDA]

QC: byte count, remove duplicates

External Data Processing
[Content Systems]

Pre-Assembly
[Content Systems]

External & Trimmed Fragments
[Content Systems]

Proto I/O File Generation
[Content Systems]

QC: "gatekeeper" syntax, duplicates & quality values

Proto I/O Files
[Content Systems]

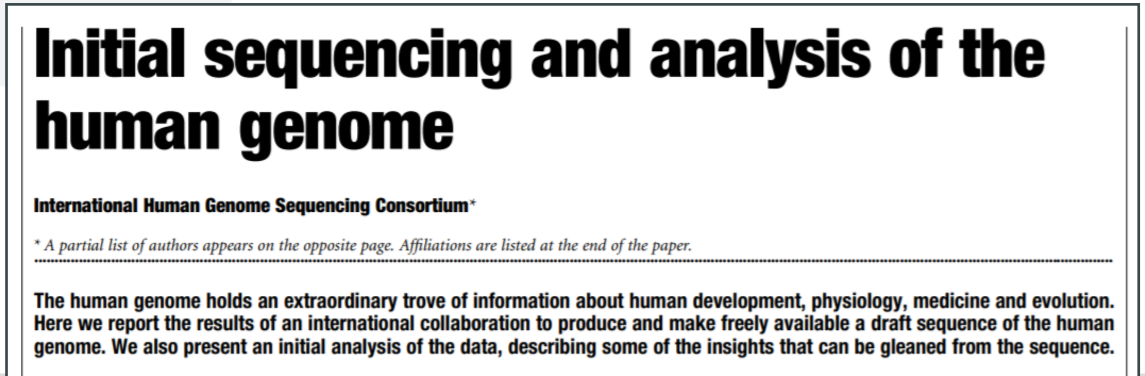
Proto I/O Files
[Content Systems]

"gatekeeper" run again

Assembly
[Informatics Research]

Chromosome Team QA review

Assemblies
[IR / CT]



Initial sequencing and analysis of the human genome

International Human Genome Sequencing Consortium*

* *A partial list of authors appears on the opposite page. Affiliations are listed at the end of the paper.*

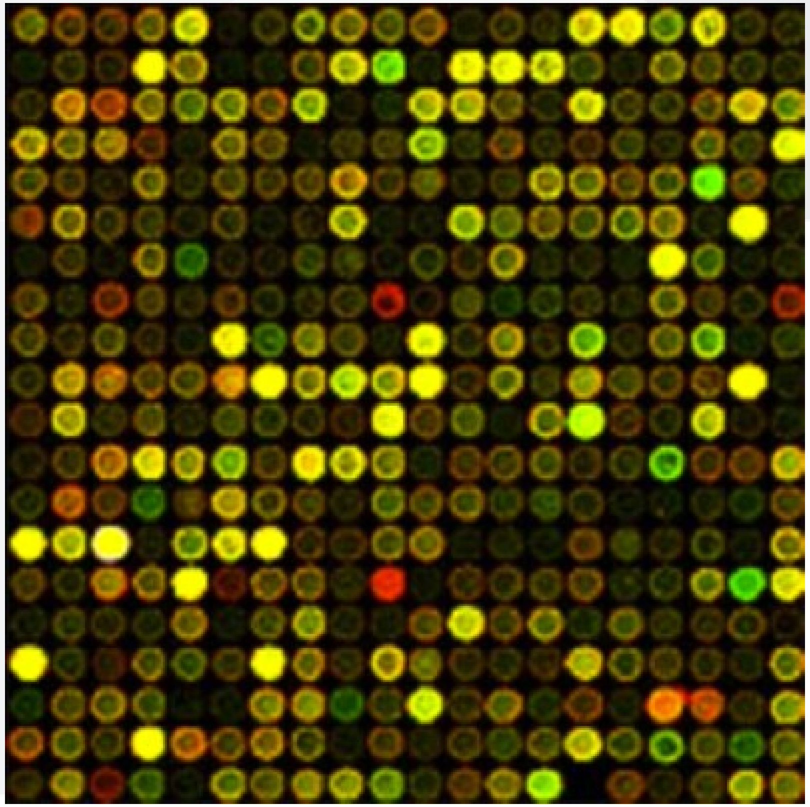
The human genome holds an extraordinary trove of information about human development, physiology, medicine and evolution. Here we report the results of an international collaboration to produce and make freely available a draft sequence of the human genome. We also present an initial analysis of the data, describing some of the insights that can be gleaned from the sequence.

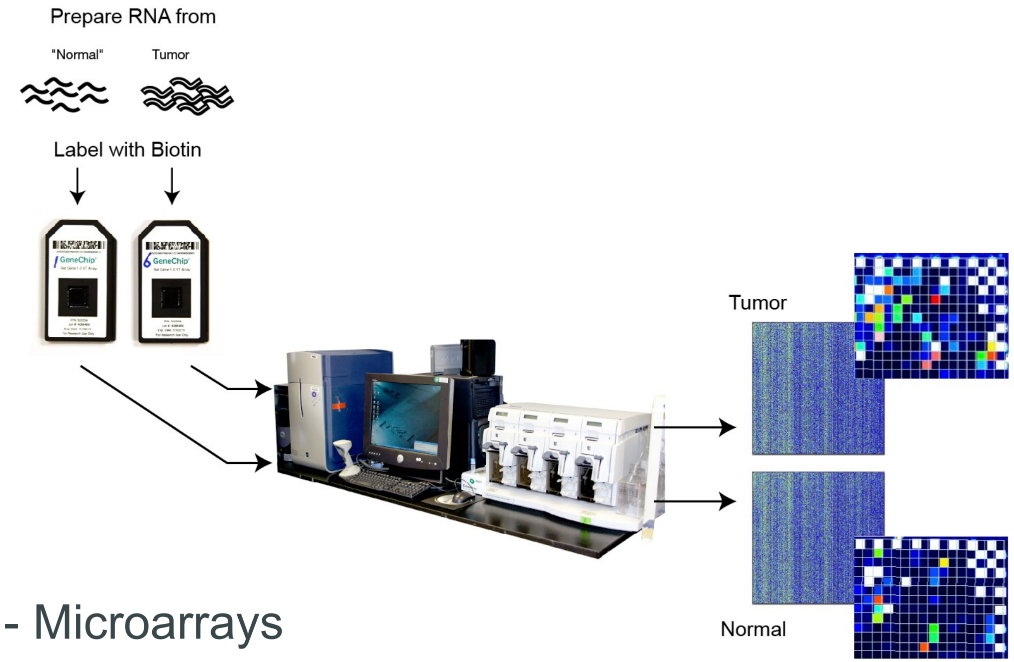
2001
Human Genome Sequenced
Cost: ~300M USD

The Sequence of the Human Genome

J. Craig Venter,[1]* Mark D. Adams,[1] Eugene W. Myers,[1] Peter W. Li,[1] Richard J. Mural,[1]
Granger G. Sutton,[1] Hamilton O. Smith,[1] Mark Yandell,[1] Cheryl A. Evans,[1] Robert A. Holt,[1]
Jeannine D. Gocayne,[1] Peter Amanatides,[1] Richard M. Ballew,[1] Daniel H. Huson,[1]
Jennifer Russo Wortman,[1] Qing Zhang,[1] Chinnappa D. Kodira,[1] Xiangqun H. Zheng,[1] Lin Chen,[1]
Marian Skupski,[1] Gangadharan Subramanian,[1] Paul D. Thomas,[1] Jinghui Zhang,[1]
George L. Gabor Miklos,[2] Catherine Nelson,[3] Samuel Broder,[1] Andrew G. Clark,[4] Joe Nadeau,[5]
Victor A. McKusick,[6] Norton Zinder,[7] Arnold J. Levine,[7] Richard J. Roberts,[8] Mel Simon,[9]
Carolyn Slayman,[10] Michael Hunkapiller,[11] Randall Bolanos,[1] Arthur Delcher,[1] Ian Dew,[1] Daniel Fasulo,[1]
Michael Flanigan,[1] Liliana Florea,[1] Aaron Halpern,[1] Sridhar Hannenhalli,[1] Saul Kravitz,[1] Samuel Levy,[1]
Clark Mobarry,[1] Knut Reinert,[1] Karin Remington,[1] Jane Abu-Threideh,[1] Ellen Beasley,[1] Kendra Biddick,[1]
Vivien Bonazzi,[1] Rhonda Brandon,[1] Michele Cargill,[1] Ishwar Chandramouliswaran,[1] Rosane Charlab,[1]
Kabir Chaturvedi,[1] Zuoming Deng,[1] Valentina Di Francesco,[1] Patrick Dunn,[1] Karen Eilbeck,[1]
Carlos Evangelista,[1] Andrei E. Gabrielian,[1] Weiniu Gan,[1] Wangmao Ge,[1] Fangcheng Gong,[1] Zhiping Gu,[1]
Ping Guan,[1] Thomas J. Heiman,[1] Maureen E. Higgins,[1] Rui-Ru Ji,[1] Zhaoxi Ke,[1] Karen A. Ketchum,[1]
Zhongwu Lai,[1] Yiding Lei,[1] Zhenya Li,[1] Jiayin Li,[1] Yong Liang,[1] Xiaoying Lin,[1] Fu Lu,[1]
Gennady V. Merkulov,[1] Natalia Milshina,[1] Helen M. Moore,[1] Ashwinikumar K Naik,[1]
Vaibhav A. Narayan,[1] Beena Neelam,[1] Deborah Nusskern,[1] Douglas B. Rusch,[1] Steven Salzberg,[12]
Wei Shao,[1] Bixiong Shue,[1] Jingtao Sun,[1] Zhen Yuan Wang,[1] Aihui Wang,[1] Xin Wang,[1] Jian Wang,[1]
Ming-Hui Wei,[1] Ron Wides,[13] Chunlin Xiao,[1] Chunhua Yan,[1] Alison Yao,[1] Jane Ye,[1] Ming Zhan,[1]
Weiqing Zhang,[1] Hongyu Zhang,[1] Qi Zhao,[1] Liansheng Zheng,[1] Fei Zhong,[1] Wenyan Zhong,[1]
Shiaoping C. Zhu,[1] Shaying Zhao,[12] Dennis Gilbert,[1] Suzanna Baumhueter,[1] Gene Spier,[1]
Christine Carter,[1] Anibal Cravchik,[1] Trevor Woodage,[1] Feroze Ali,[1] Huijin An,[1] Aderonke Awe,[1]
Danita Baldwin,[1] Holly Baden,[1] Mary Barnstead,[1] Ian Barrow,[1] Karen Beeson,[1] Dana Busam,[1]
Amy Carver,[1] Angela Center,[1] Ming Lai Cheng,[1] Liz Curry,[1] Steve Danaher,[1] Lionel Davenport,[1]
Raymond Desilets,[1] Susanne Dietz,[1] Kristina Dodson,[1] Lisa Doup,[1] Steven Ferriera,[1] Neha Garg,[1]
Andres Gluecksmann,[1] Brit Hart,[1] Jason Haynes,[1] Charles Haynes,[1] Cheryl Heiner,[1] Suzanne Hladun,[1]
Damon Hostin,[1] Jarrett Houck,[1] Timothy Howland,[1] Chinyere Ibegwam,[1] Jeffery Johnson,[1]
Francis Kalush,[1] Lesley Kline,[1] Shashi Koduru,[1] Amy Love,[1] Felecia Mann,[1] David May,[1]
Steven McCawley,[1] Tina McIntosh,[1] Ivy McMullen,[1] Mee Moy,[1] Linda Moy,[1] Brian Murphy,[1]
Keith Nelson,[1] Cynthia Pfannkoch,[1] Eric Pratts,[1] Vinita Puri,[1] Hina Qureshi,[1] Matthew Reardon,[1]
Robert Rodriguez,[1] Yu-Hui Rogers,[1] Deanna Romblad,[1] Bob Ruhfel,[1] Richard Scott,[1] Cynthia Sitter,[1]
Michelle Smallwood,[1] Erin Stewart,[1] Renee Strong,[1] Ellen Suh,[1] Reginald Thomas,[1] Ni Ni Tint,[1]
Sukyee Tse,[1] Claire Vech,[1] Gary Wang,[1] Jeremy Wetter,[1] Sherita Williams,[1] Monica Williams,[1]
Sandra Windsor,[1] Emily Winn-Deen,[1] Keriellen Wolfe,[1] Jayshree Zaveri,[1] Karena Zaveri,[1]
Josep F. Abril,[14] Roderic Guigó,[14] Michael J. Campbell,[1] Kimmen V. Sjolander,[1] Brian Karlak,[1]
Anish Kejariwal,[1] Huaiyu Mi,[1] Betty Lazareva,[1] Thomas Hatton,[1] Apurva Narechania,[1] Karen Diemer,[1]
Anushya Muruganujan,[1] Nan Guo,[1] Shinji Sato,[1] Vineet Bafna,[1] Sorin Istrail,[1] Ross Lippert,[1]
Russell Schwartz,[1] Brian Walenz,[1] Shibu Yooseph,[1] David Allen,[1] Anand Basu,[1] James Baxendale,[1]
Louis Blick,[1] Marcelo Caminha,[1] John Carnes-Stine,[1] Parris Caulk,[1] Yen-Hui Chiang,[1] My Coyne,[1]
Carl Dahlke,[1] Anne Deslattes Mays,[1] Maria Dombroski,[1] Michael Donnelly,[1] Dale Ely,[1] Shiva Esparham,[1]
Carl Fosler,[1] Harold Gire,[1] Stephen Glanowski,[1] Kenneth Glasser,[1] Anna Glodek,[1] Mark Gorokhov,[1]
Ken Graham,[1] Barry Gropman,[1] Michael Harris,[1] Jeremy Heil,[1] Scott Henderson,[1] Jeffrey Hoover,[1]
Donald Jennings,[1] Catherine Jordan,[1] James Jordan,[1] John Kasha,[1] Leonid Kagan,[1] Cheryl Kraft,[1]
Alexander Levitsky,[1] Mark Lewis,[1] Xiangjun Liu,[1] John Lopez,[1] Daniel Ma,[1] William Majoros,[1]
Joe McDaniel,[1] Sean Murphy,[1] Matthew Newman,[1] Trung Nguyen,[1] Ngoc Nguyen,[1] Marc Nodell,[1]
Sue Pan,[1] Jim Peck,[1] Marshall Peterson,[1] William Rowe,[1] Robert Sanders,[1] John Scott,[1]
Michael Simpson,[1] Thomas Smith,[1] Arlan Sprague,[1] Timothy Stockwell,[1] Russell Turner,[1] Eli Venter,[1]
Mei Wang,[1] Meiyuan Wen,[1] David Wu,[1] Mitchell Wu,[1] Ashley Xia,[1] Ali Zandieh,[1] Xiaohong Zhu[1]

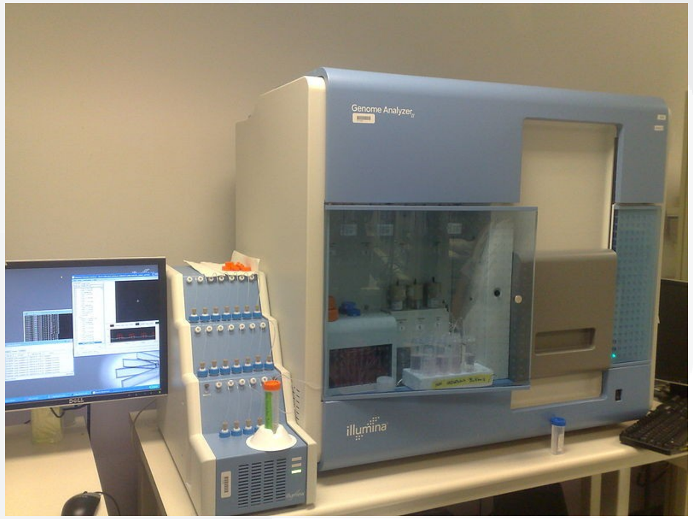
Nucleic Acid (not-)Sequencing History



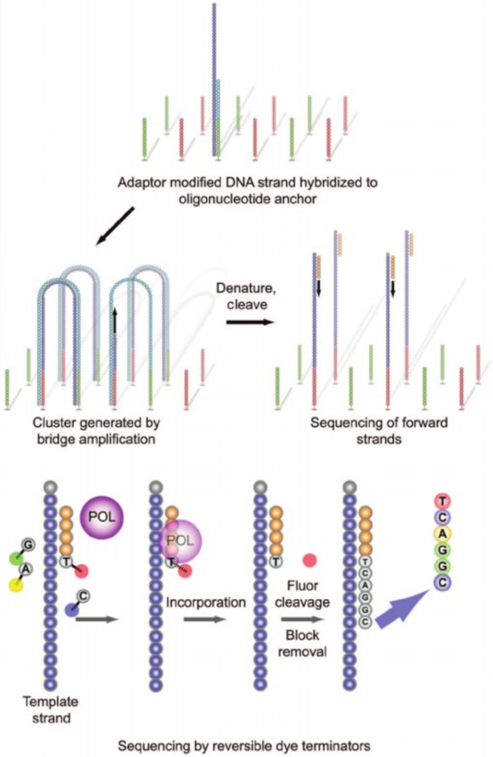


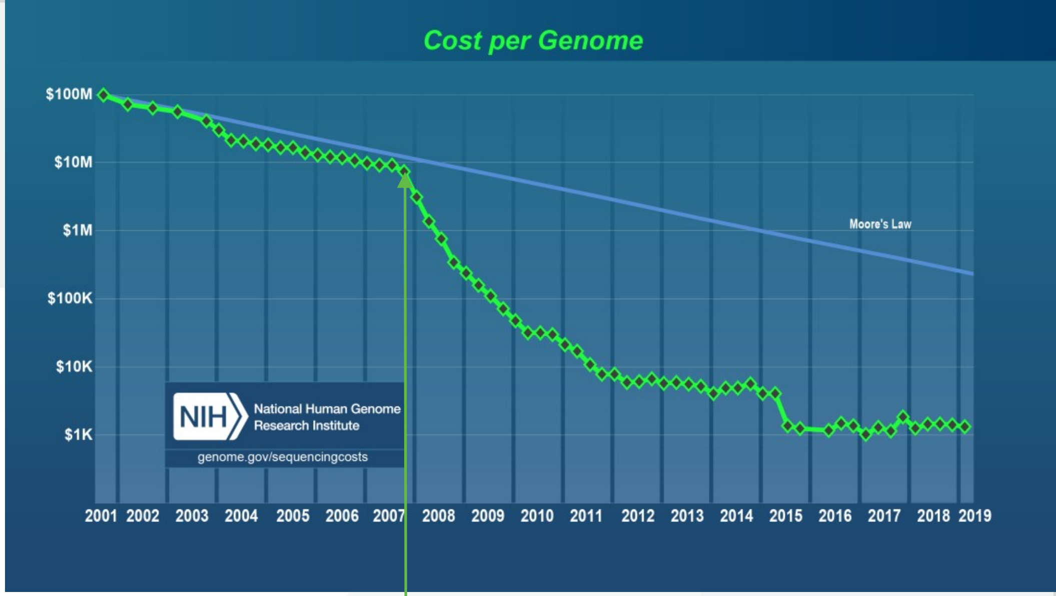
2000s - Microarrays

Nucleic Acid Sequencing History

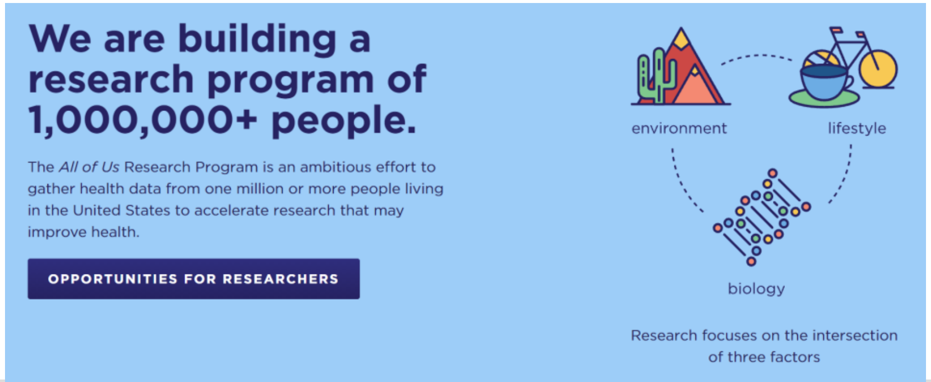


2006 – Solexa Genome Analyser
2007 – Solexa bought by Illumina


Adaptor modified DNA strand hybridized to oligonucleotide anchor
Denature, cleave
Cluster generated by bridge amplification
Sequencing of forward strands
POL
POL
Incorporation
Fluor cleavage
Block removal
Template strand
T
C
A
G
G
C
Sequencing by reversible dye terminators



Next Generation Sequencing
New Generation Sequencing
NGS



We are building a research program of 1,000,000+ people.

The *All of Us* Research Program is an ambitious effort to gather health data from one million or more people living in the United States to accelerate research that may improve health.

OPPORTUNITIES FOR RESEARCHERS

Research focuses on the intersection of three factors

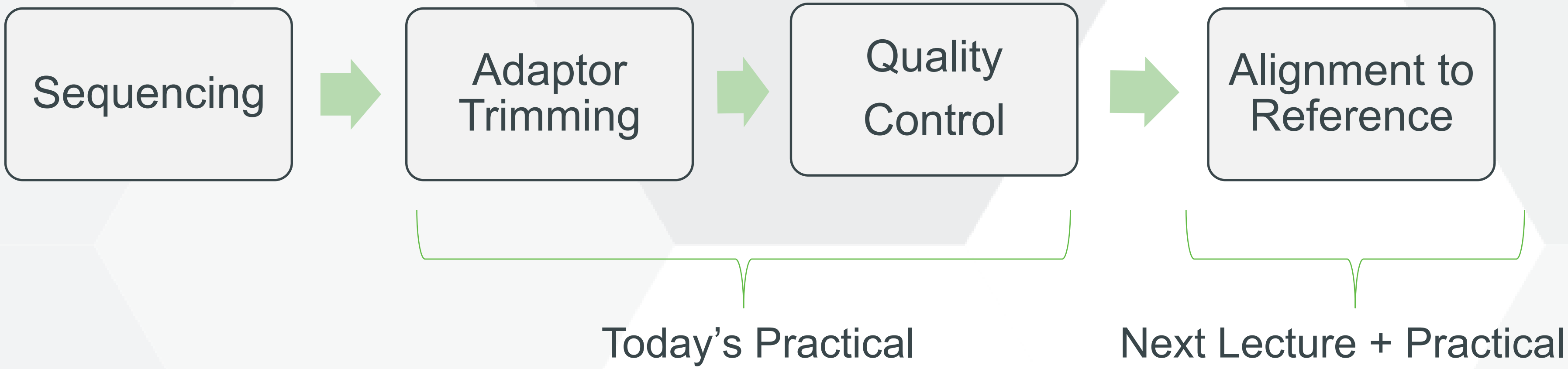
Realistic goal in three-five years

Sequence the entire human genome in a few days for $1000 (Era of Personal Genomics)

HOWEVER, speed of sequencing does not necessarily mean an **understanding** of the genetic information or DNA structure!

2015

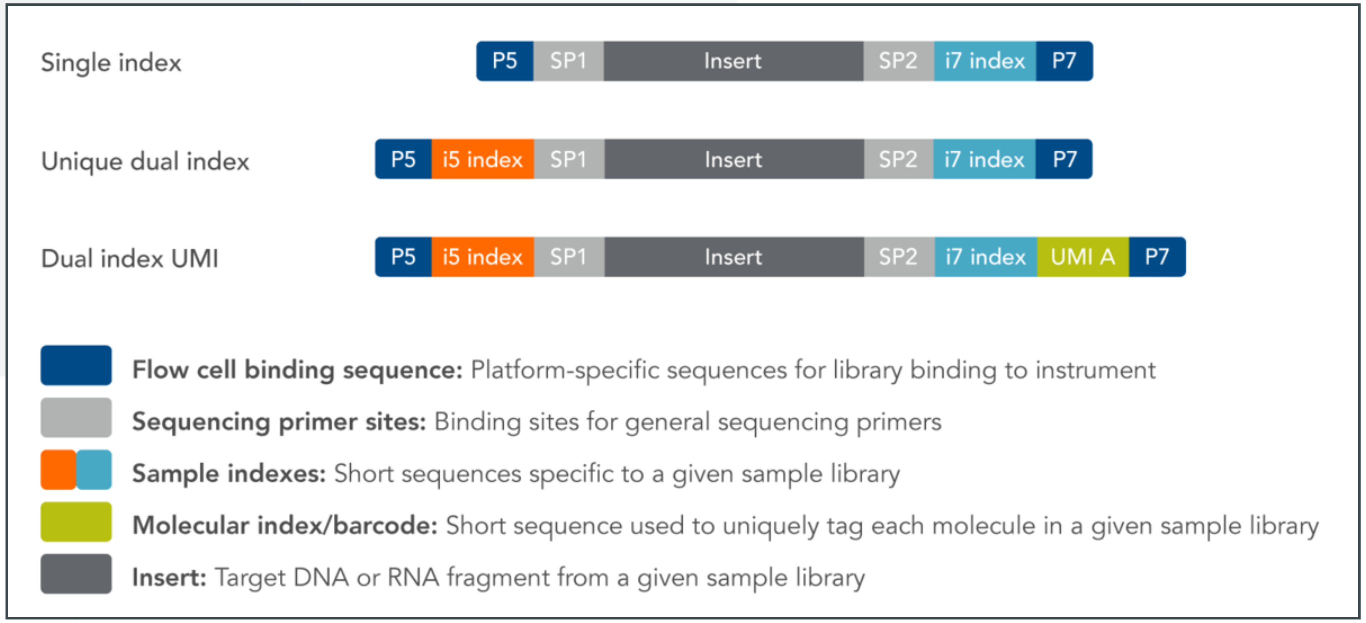
NGS Data analysis workflow



Sequencing → Adaptor Trimming → Quality Control → Alignment to Reference

Today's Practical

Next Lecture + Practical



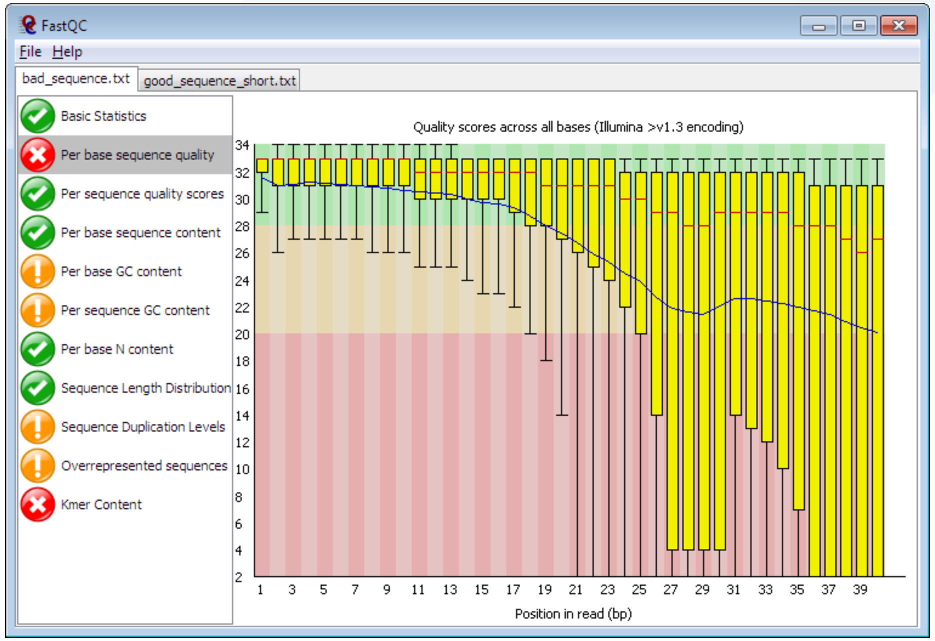
Flow cell binding sequence: Platform-specific sequences for library binding to instrument

Sequencing primer sites: Binding sites for general sequencing primers

Sample indexes: Short sequences specific to a given sample library

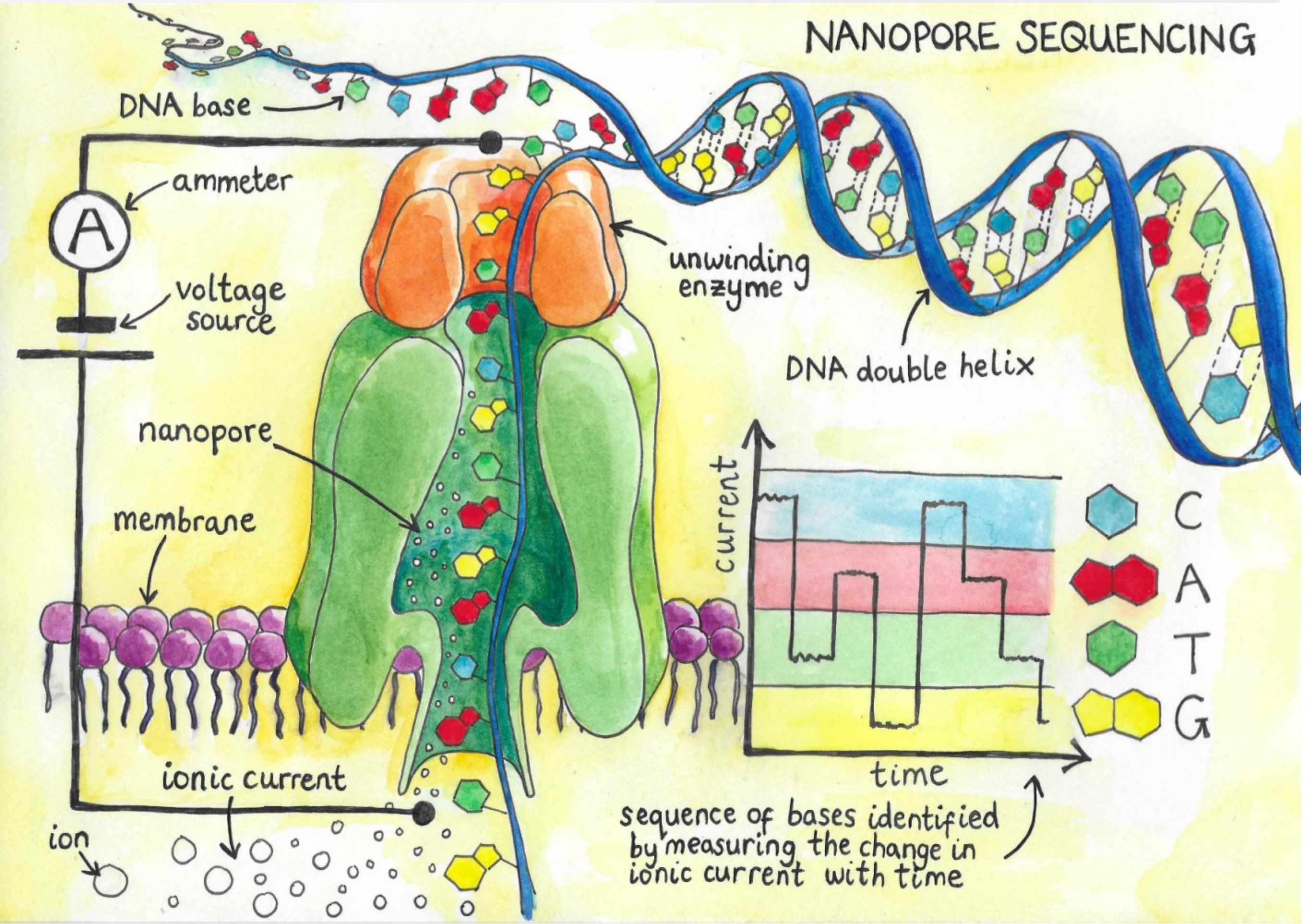
Molecular index/barcode: Short sequence used to uniquely tag each molecule in a given sample library

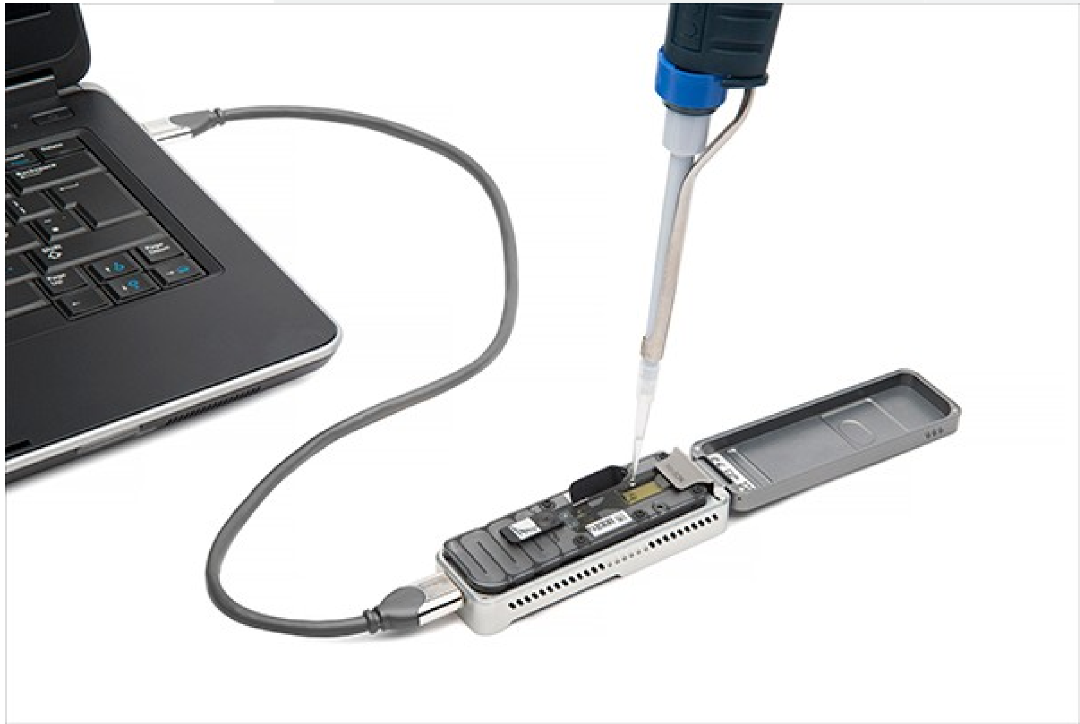
Insert: Target DNA or RNA fragment from a given sample library



- ✓ Basic Statistics
- ✗ Per base sequence quality
- ✓ Per sequence quality scores
- ✓ Per base sequence content
- ! Per base GC content
- ! Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ! Sequence Duplication Levels
- ! Overrepresented sequences
- ✗ Kmer Content

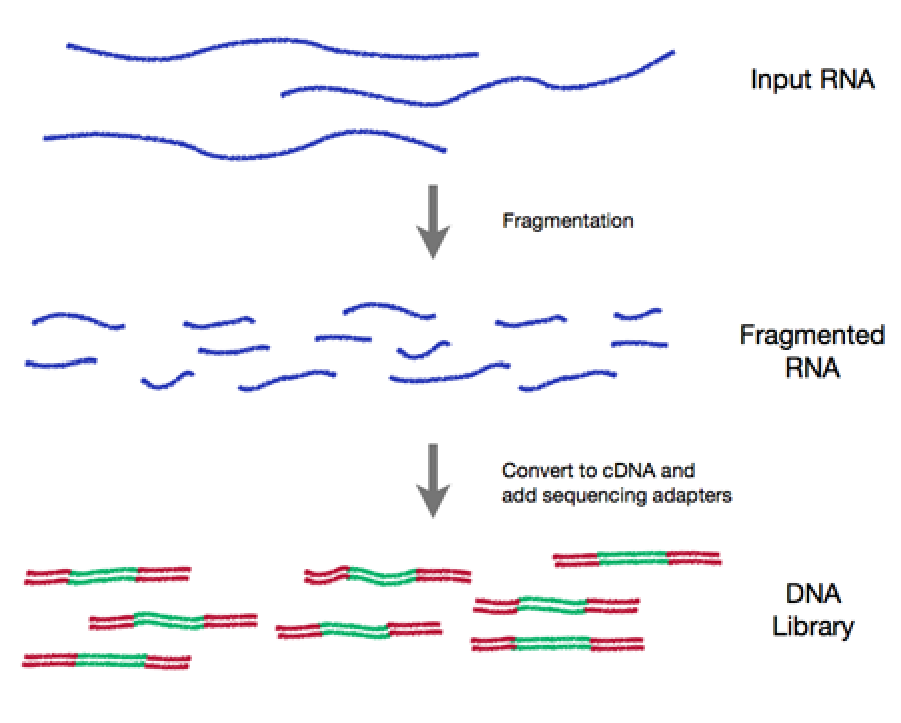
Long Read Sequencing





Long Reads – Low per read accuracy

RNA-Seq



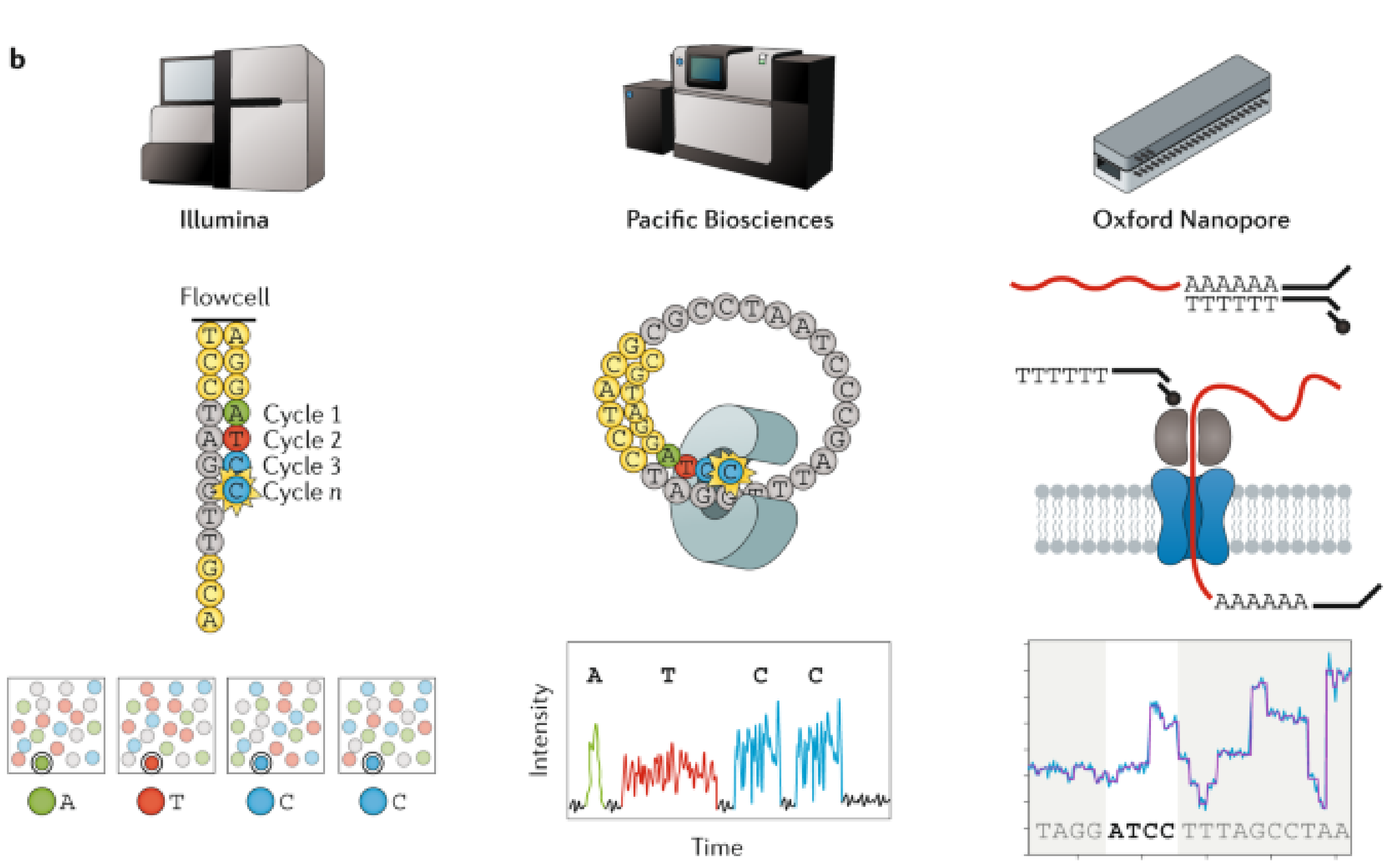
Input RNA

Fragmentation

Fragmented RNA

Convert to cDNA and add sequencing adapters

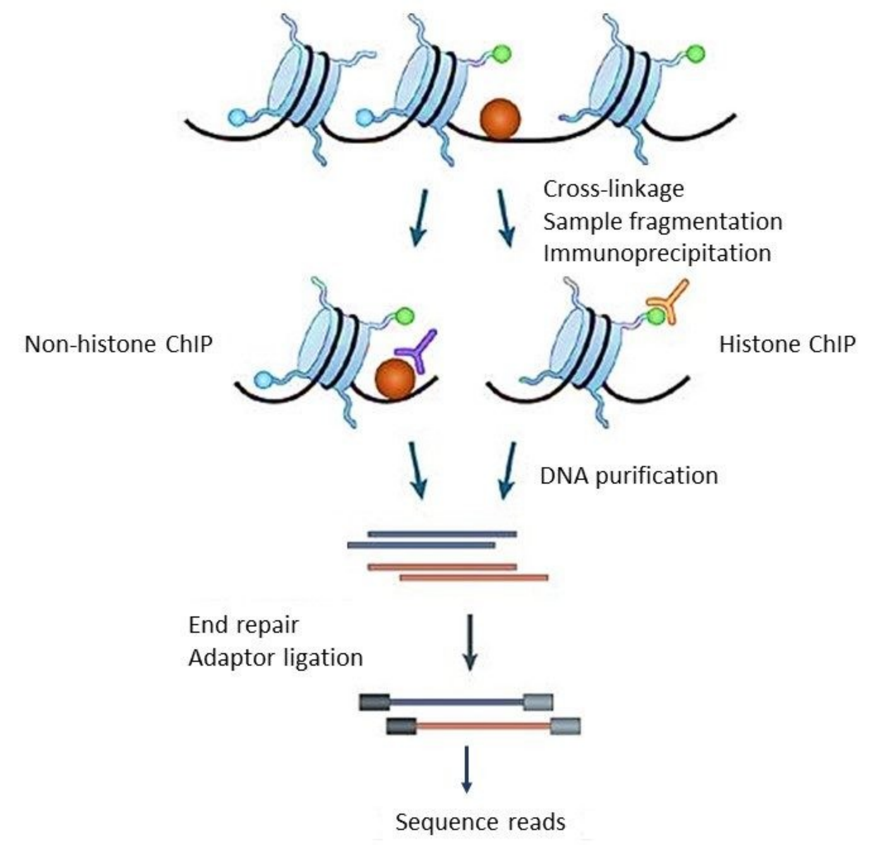
DNA Library



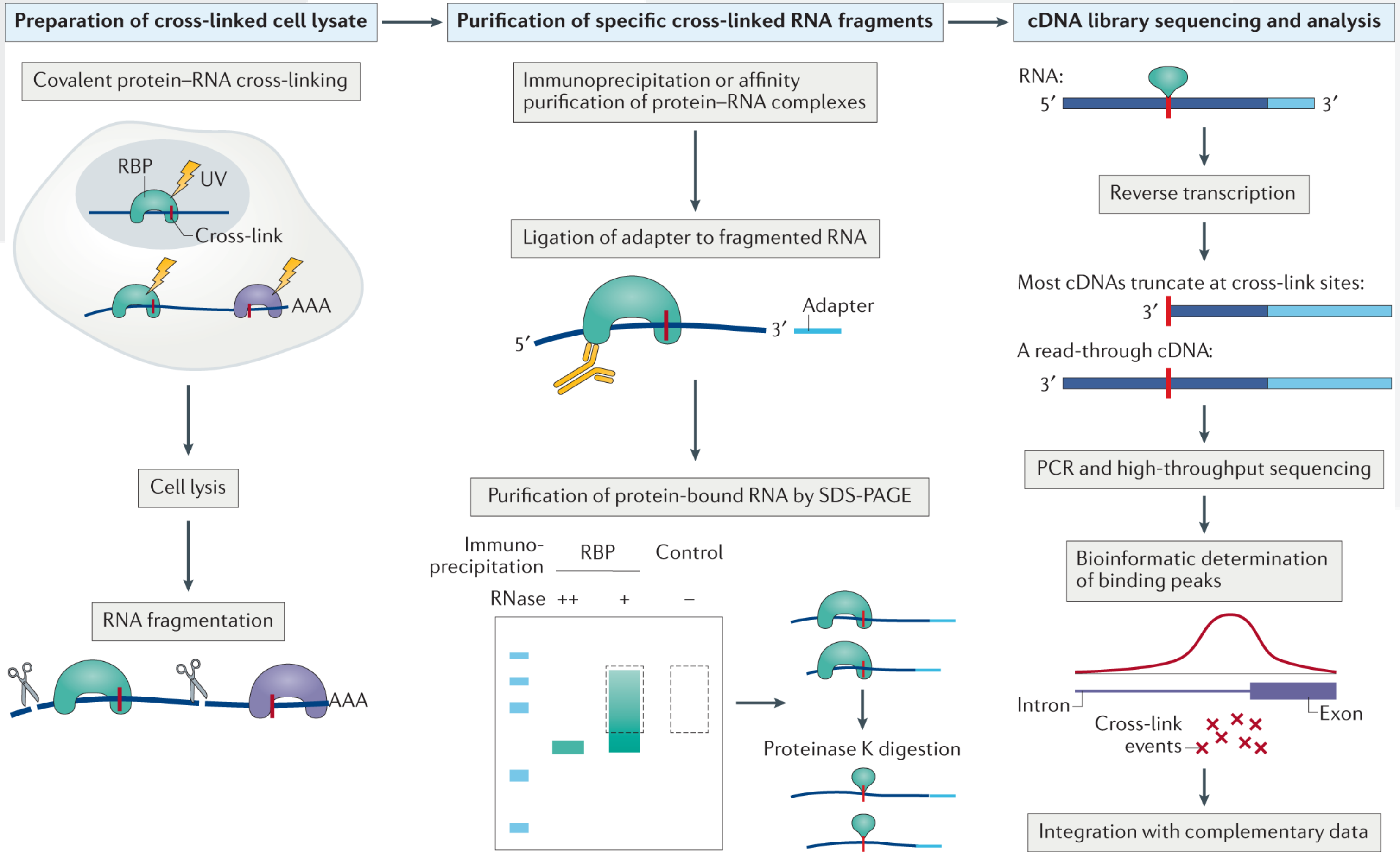
RNA-Seq Analysis

- Alignment to transcriptome or genome (gapped)
- Poly-A selection or Ribosomal RNA depletion
- Can be used to quantify RNA, or to identify structural differences (e.g. splicing)
- Usual downstream analysis: Fold Change between conditions

Immunoprecipitation based techniques



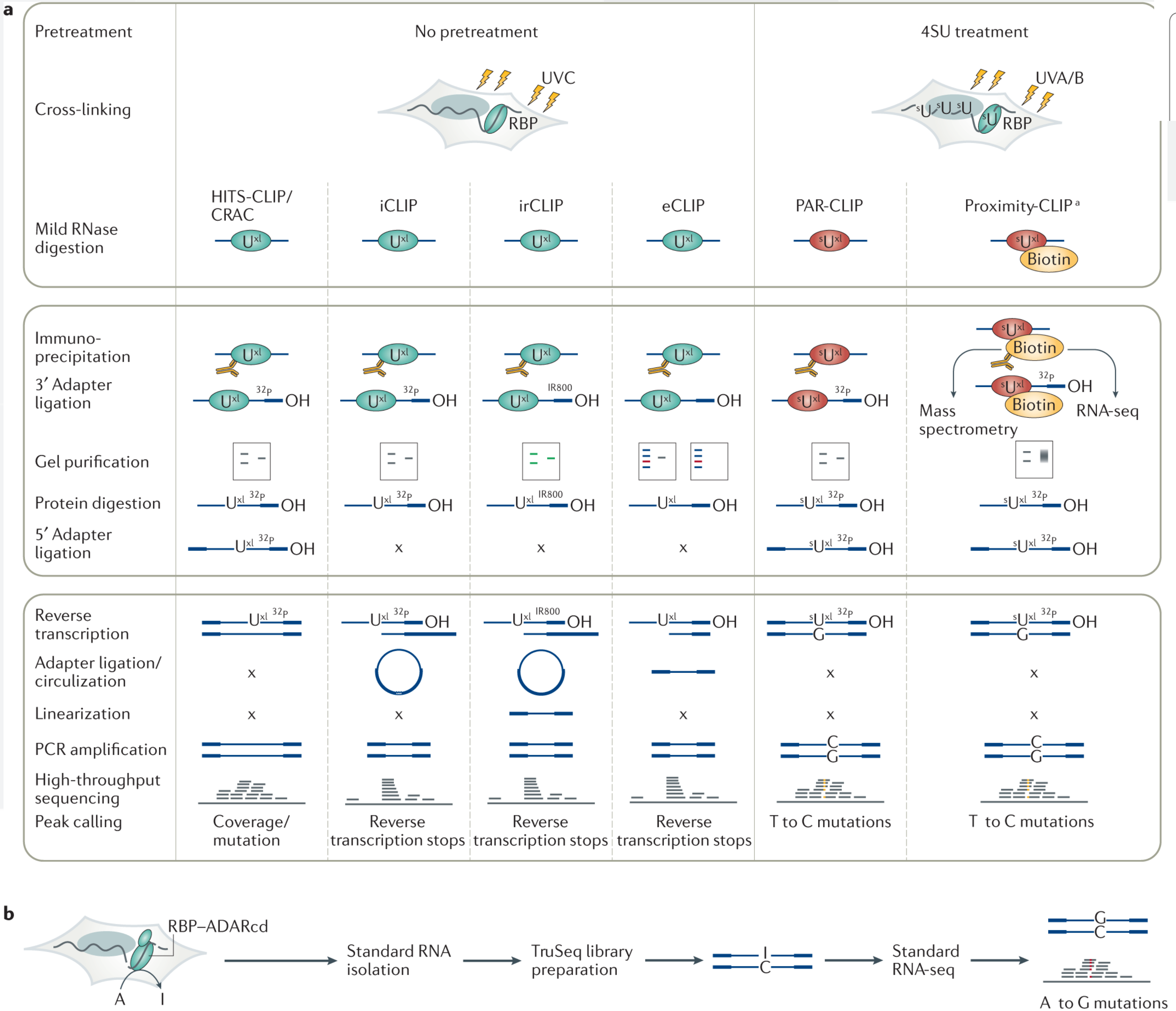
ChIP-Seq : DNA Binding Proteins

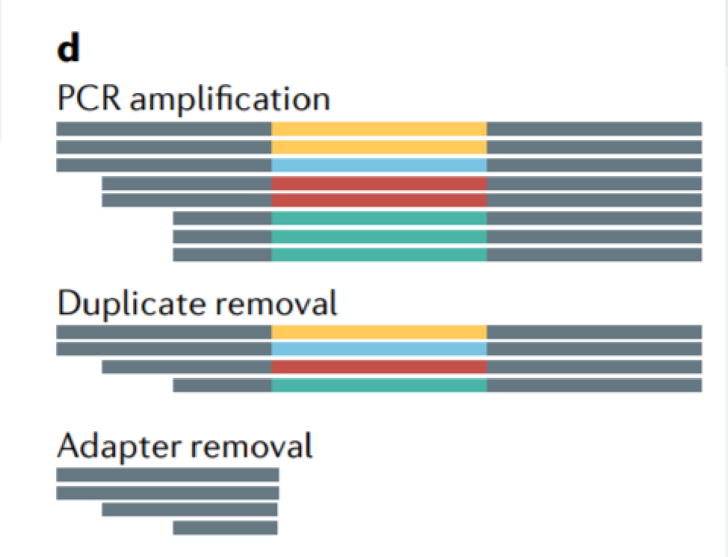

Preparation of cross-linked cell lysate
Purification of specific cross-linked RNA fragments
cDNA library sequencing and analysis
Covalent protein–RNA cross-linking
RBP
UV
Cross-link
AAA
Cell lysis
RNA fragmentation
AAA
Immunoprecipitation or affinity purification of protein–RNA complexes
Ligation of adapter to fragmented RNA
5′
3′
Adapter
Purification of protein-bound RNA by SDS-PAGE
Immuno-precipitation
RBP
Control
RNase ++
+
–
Proteinase K digestion
RNA:
5′
3′
Reverse transcription
Most cDNAs truncate at cross-link sites:
3′
A read-through cDNA:
3′
PCR and high-throughput sequencing
Bioinformatic determination of binding peaks
Intron
Exon
Cross-link events
Integration with complementary data

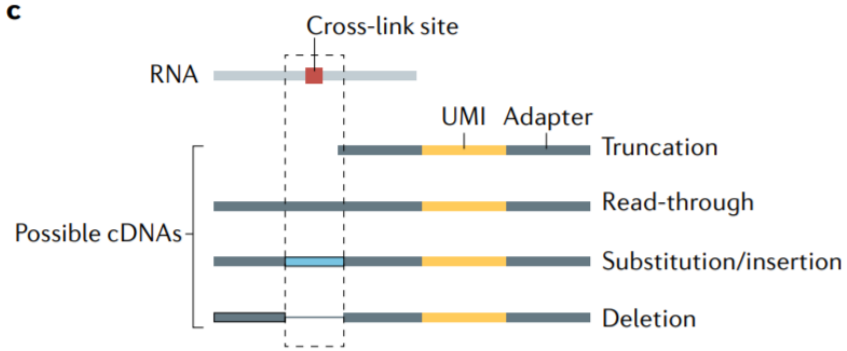
CLIP-Seq : RNA Binding Proteins

CLIP and complementary methods

Markus Hafner[1], *Maria Katsantoni*[2,3], *Tino Köster*[4], *James Marks*[1], *Joyita Mukherjee*[5,6], *Dorothee Staiger*[4], *Jernej Ule*[5,6,7]✉ *and Mihaela Zavolan*[2,3]

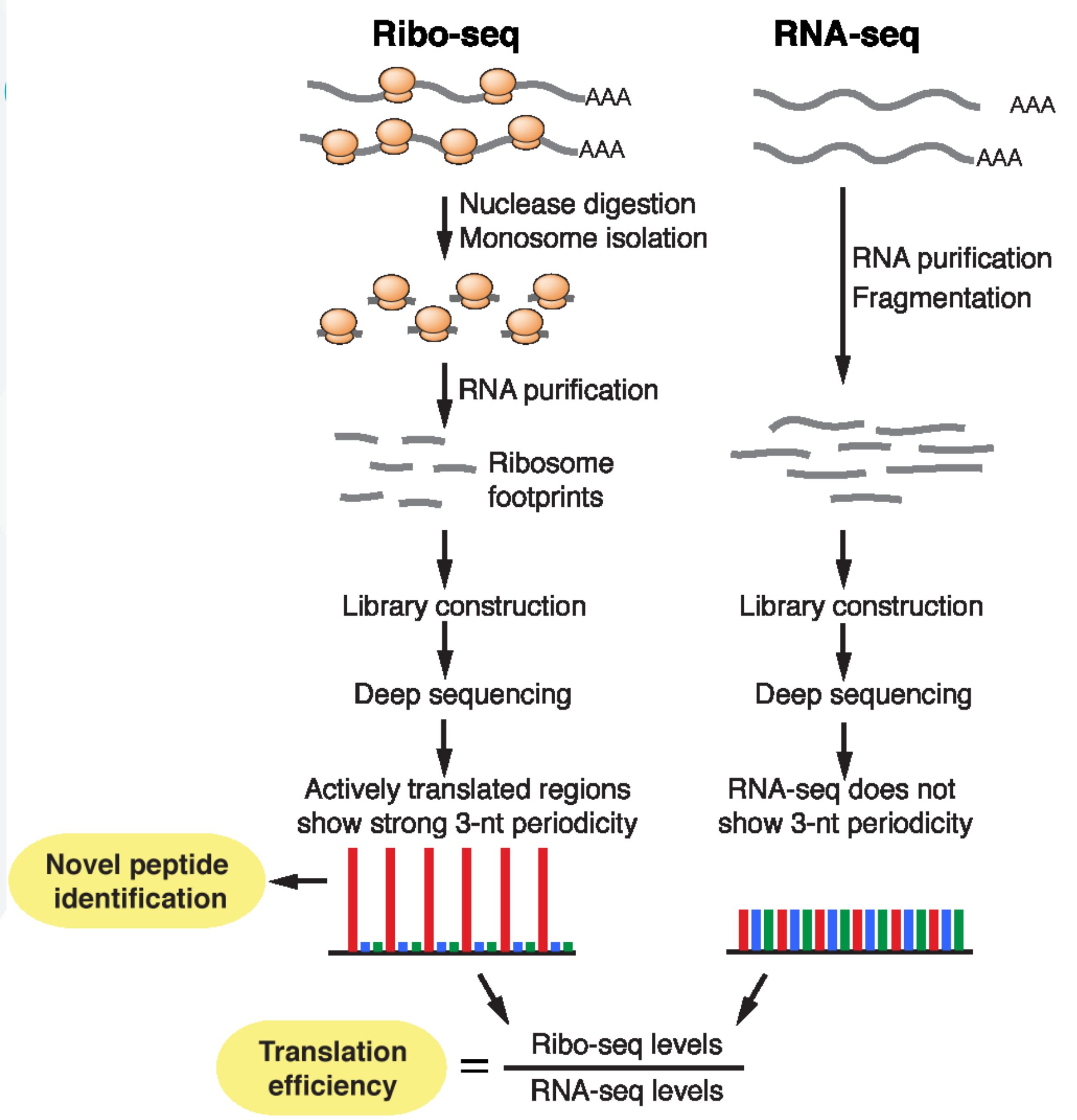






https://www.nature.com/articles/s43586-021-00018-1

Rib



Ribo-seq

RNA-seq

AAA

AAA

AAA

AAA

Nuclease digestion
Monosome isolation

RNA purification
Fragmentation

RNA purification

Ribosome footprints

Library construction

Library construction

Deep sequencing

Deep sequencing

Actively translated regions show strong 3-nt periodicity

RNA-seq does not show 3-nt periodicity

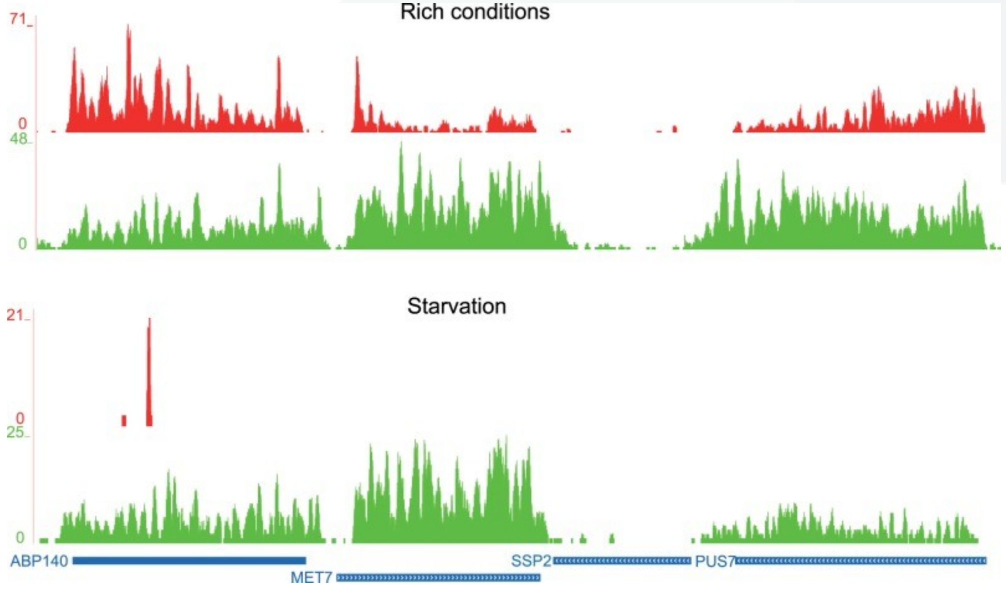
Novel peptide identification

Translation efficiency = Ribo-seq levels / RNA-seq levels

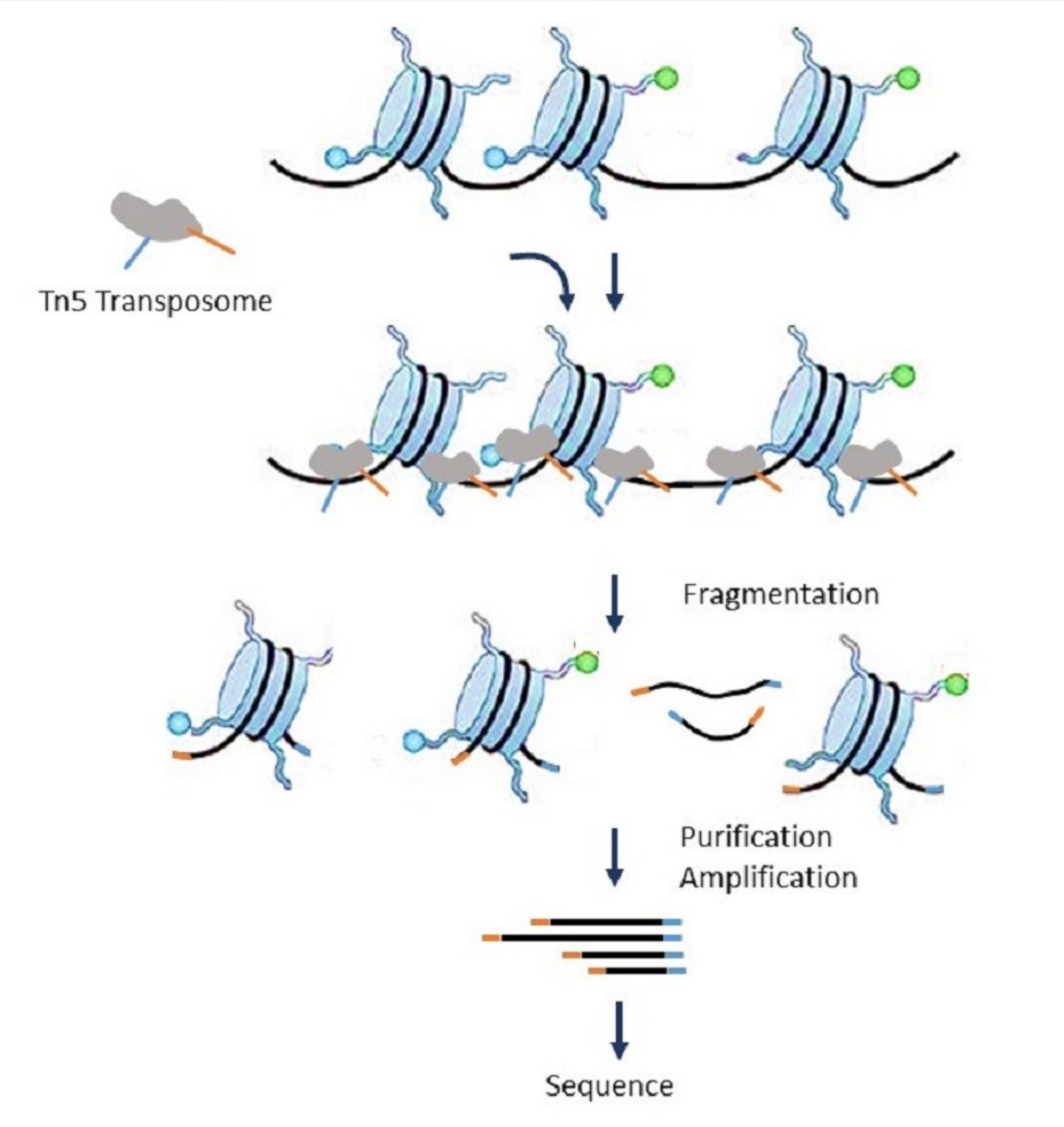
Adapted from Hsu *et al.* 2016

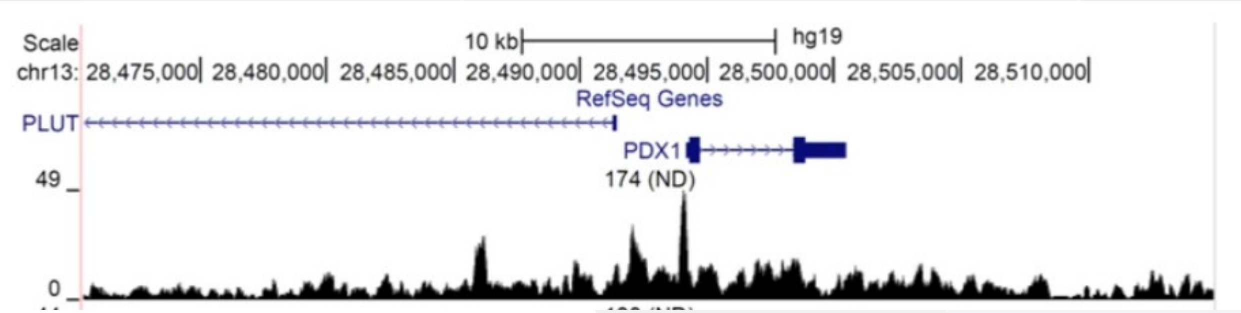
Identification of actively translated RNA

Exact position of Ribosome



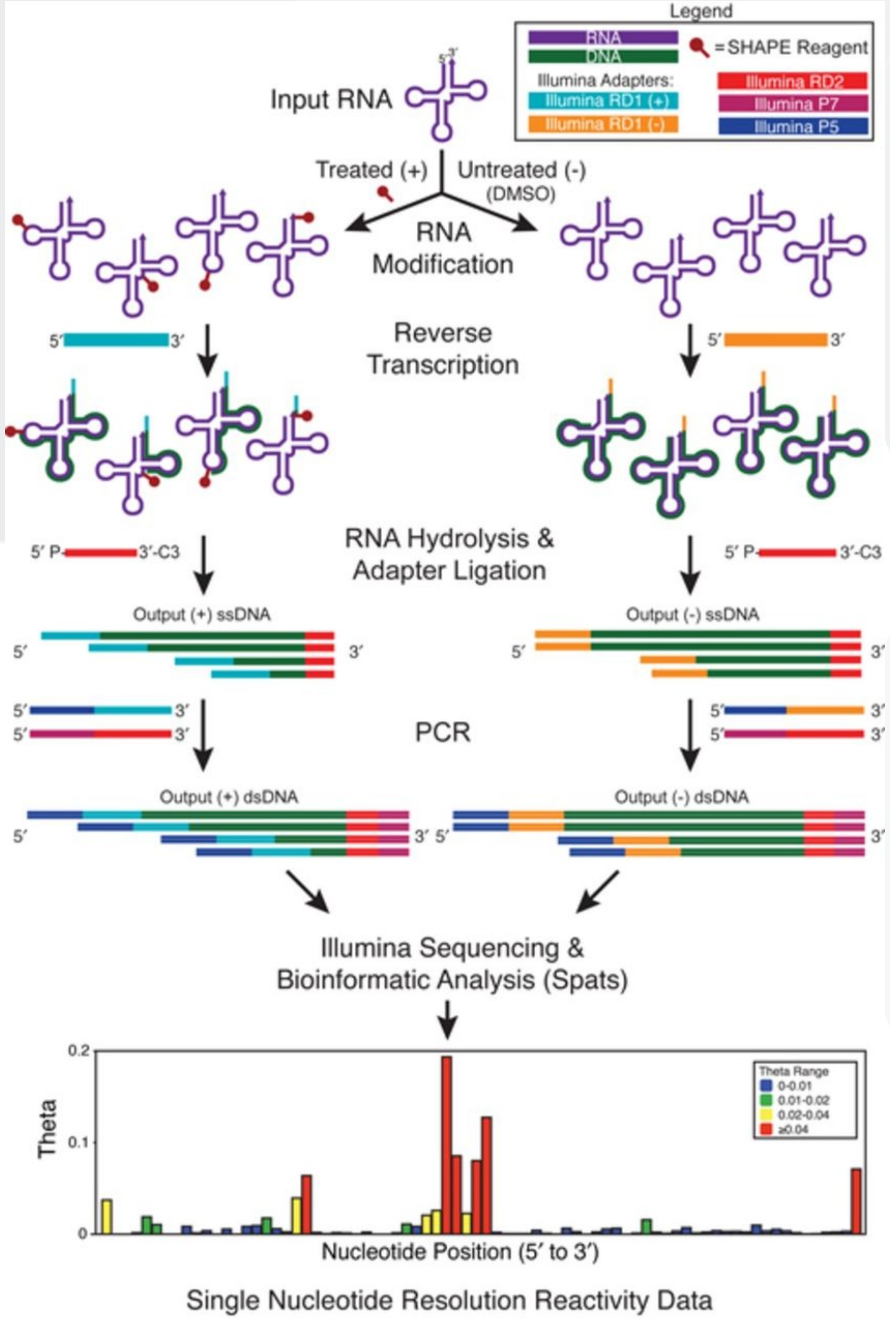
ATAC-Seq



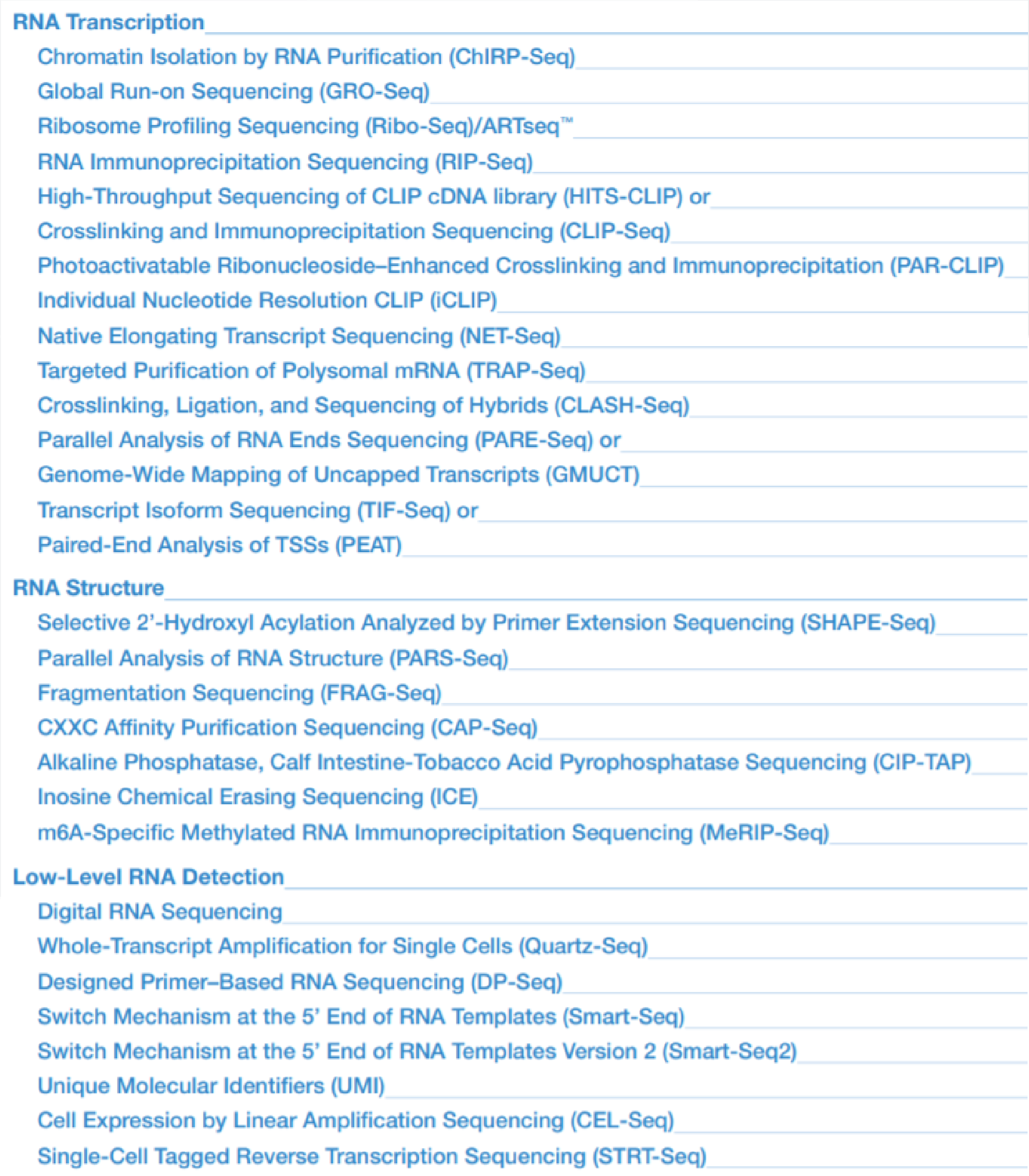


Identification of accessible chromatin areas

Shape-Seq


Legend
RNA
DNA
= SHAPE Reagent
Illumina Adapters:
Illumina RD1 (+)
Illumina RD1 (-)
Illumina RD2
Illumina P7
Illumina P5
Input RNA
Treated (+)
Untreated (-)
(DMSO)
RNA Modification
Reverse Transcription
5′
3′
RNA Hydrolysis & Adapter Ligation
5′ P
3′-C3
Output (+) ssDNA
Output (-) ssDNA
PCR
Output (+) dsDNA
Output (-) dsDNA
Illumina Sequencing & Bioinformatic Analysis (Spats)
Theta
0.2
0.1
0
Theta Range
0-0.01
0.01-0.02
0.02-0.04
≥0.04
Nucleotide Position (5′ to 3′)
Single Nucleotide Resolution Reactivity Data

… and others



RNA Transcription

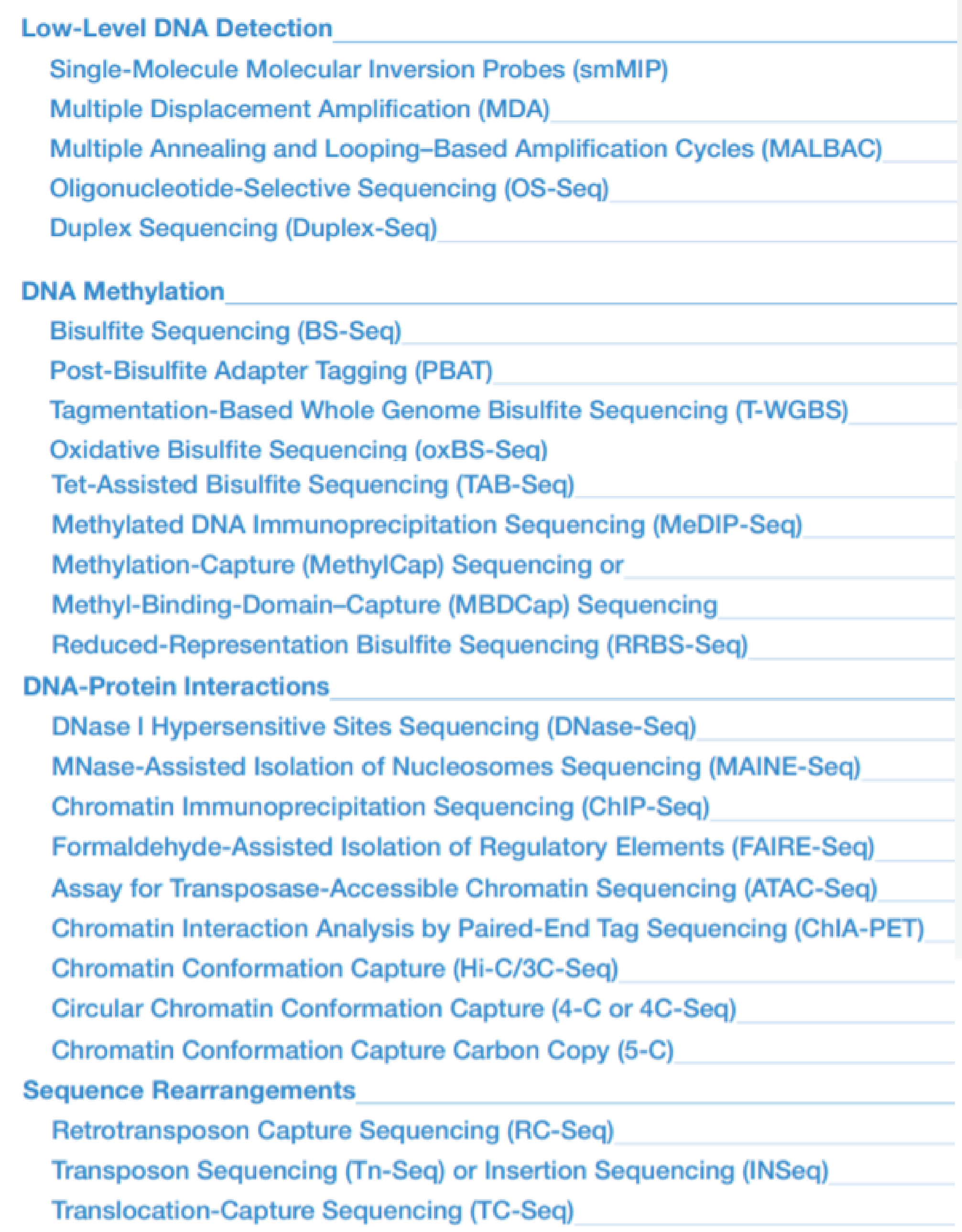
- Chromatin Isolation by RNA Purification (ChIRP-Seq)
- Global Run-on Sequencing (GRO-Seq)
- Ribosome Profiling Sequencing (Ribo-Seq)/ARTseq™
- RNA Immunoprecipitation Sequencing (RIP-Seq)
- High-Throughput Sequencing of CLIP cDNA library (HITS-CLIP) or
- Crosslinking and Immunoprecipitation Sequencing (CLIP-Seq)
- Photoactivatable Ribonucleoside–Enhanced Crosslinking and Immunoprecipitation (PAR-CLIP)
- Individual Nucleotide Resolution CLIP (iCLIP)
- Native Elongating Transcript Sequencing (NET-Seq)
- Targeted Purification of Polysomal mRNA (TRAP-Seq)
- Crosslinking, Ligation, and Sequencing of Hybrids (CLASH-Seq)
- Parallel Analysis of RNA Ends Sequencing (PARE-Seq) or
- Genome-Wide Mapping of Uncapped Transcripts (GMUCT)
- Transcript Isoform Sequencing (TIF-Seq) or
- Paired-End Analysis of TSSs (PEAT)

RNA Structure

- Selective 2'-Hydroxyl Acylation Analyzed by Primer Extension Sequencing (SHAPE-Seq)
- Parallel Analysis of RNA Structure (PARS-Seq)
- Fragmentation Sequencing (FRAG-Seq)
- CXXC Affinity Purification Sequencing (CAP-Seq)
- Alkaline Phosphatase, Calf Intestine-Tobacco Acid Pyrophosphatase Sequencing (CIP-TAP)
- Inosine Chemical Erasing Sequencing (ICE)
- m6A-Specific Methylated RNA Immunoprecipitation Sequencing (MeRIP-Seq)

Low-Level RNA Detection

- Digital RNA Sequencing
- Whole-Transcript Amplification for Single Cells (Quartz-Seq)
- Designed Primer–Based RNA Sequencing (DP-Seq)
- Switch Mechanism at the 5' End of RNA Templates (Smart-Seq)
- Switch Mechanism at the 5' End of RNA Templates Version 2 (Smart-Seq2)
- Unique Molecular Identifiers (UMI)
- Cell Expression by Linear Amplification Sequencing (CEL-Seq)
- Single-Cell Tagged Reverse Transcription Sequencing (STRT-Seq)



Low-Level DNA Detection

- Single-Molecule Molecular Inversion Probes (smMIP)
- Multiple Displacement Amplification (MDA)
- Multiple Annealing and Looping–Based Amplification Cycles (MALBAC)
- Oligonucleotide-Selective Sequencing (OS-Seq)
- Duplex Sequencing (Duplex-Seq)

DNA Methylation

- Bisulfite Sequencing (BS-Seq)
- Post-Bisulfite Adapter Tagging (PBAT)
- Tagmentation-Based Whole Genome Bisulfite Sequencing (T-WGBS)
- Oxidative Bisulfite Sequencing (oxBS-Seq)
- Tet-Assisted Bisulfite Sequencing (TAB-Seq)
- Methylated DNA Immunoprecipitation Sequencing (MeDIP-Seq)
- Methylation-Capture (MethylCap) Sequencing or
- Methyl-Binding-Domain–Capture (MBDCap) Sequencing
- Reduced-Representation Bisulfite Sequencing (RRBS-Seq)

DNA-Protein Interactions

- DNase I Hypersensitive Sites Sequencing (DNase-Seq)
- MNase-Assisted Isolation of Nucleosomes Sequencing (MAINE-Seq)
- Chromatin Immunoprecipitation Sequencing (ChIP-Seq)
- Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE-Seq)
- Assay for Transposase-Accessible Chromatin Sequencing (ATAC-Seq)
- Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET)
- Chromatin Conformation Capture (Hi-C/3C-Seq)
- Circular Chromatin Conformation Capture (4-C or 4C-Seq)
- Chromatin Conformation Capture Carbon Copy (5-C)

Sequence Rearrangements

- Retrotransposon Capture Sequencing (RC-Seq)
- Transposon Sequencing (Tn-Seq) or Insertion Sequencing (INSeq)
- Translocation-Capture Sequencing (TC-Seq)

Link

CEITEC

@CEITEC_Brno

www.ceitec.eu