

Transcriptomics

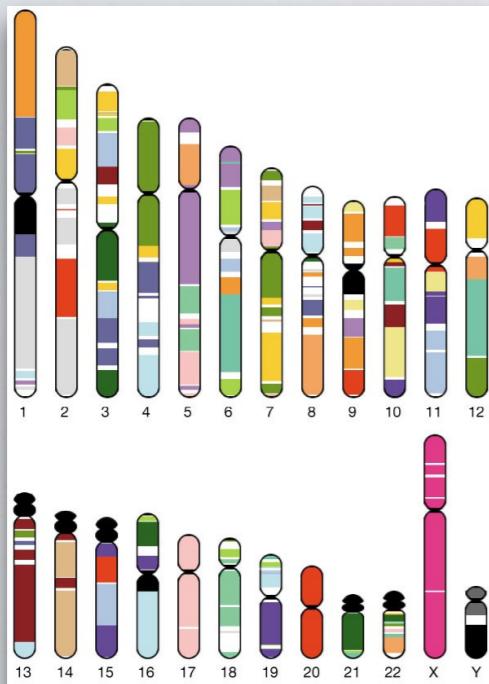


Ryan Lister

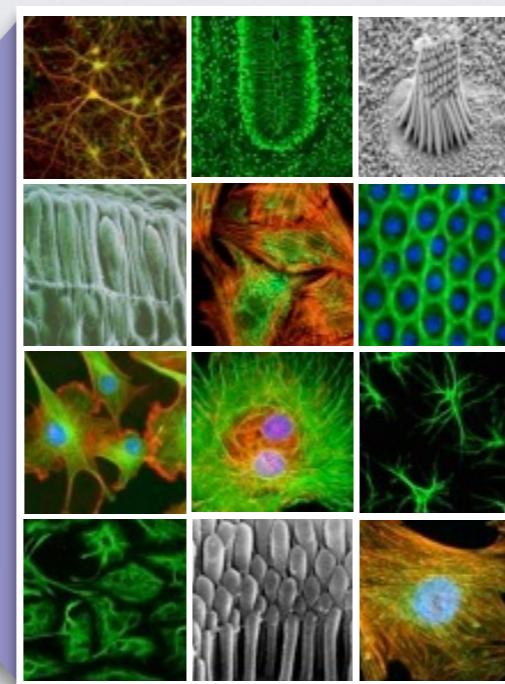
listerlab.org

Transcriptional output of the genome

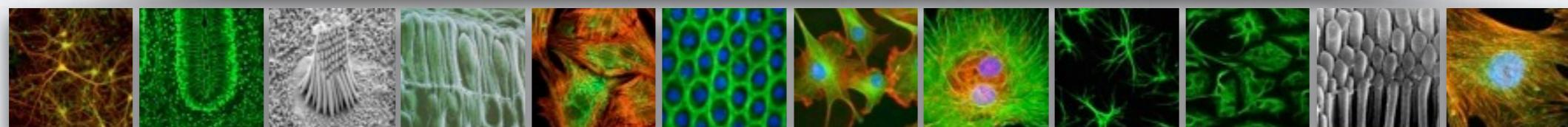
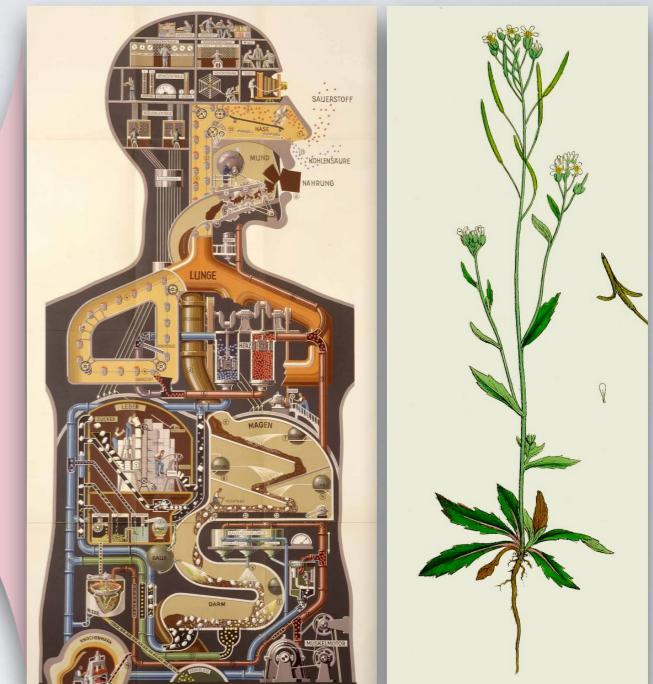
One genome



Hundreds of cell types



Complex development & physiology



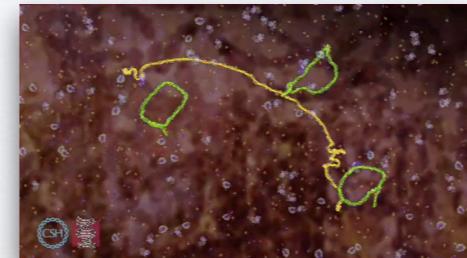
Different transcriptional output of the genome

gene expression



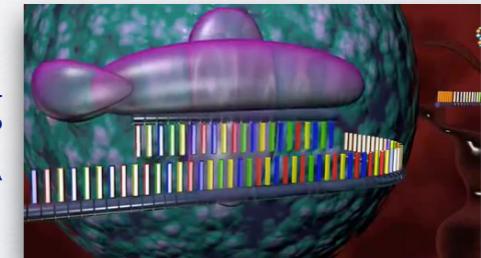
<http://www.dnalc.org/>

splicing



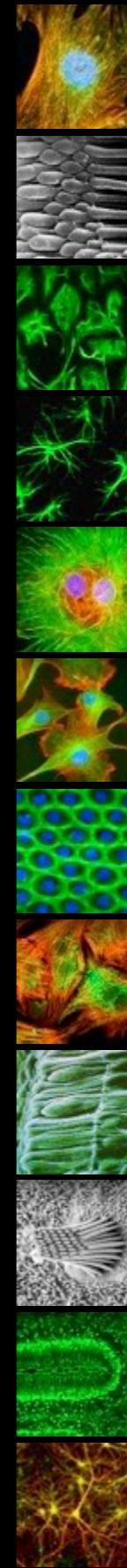
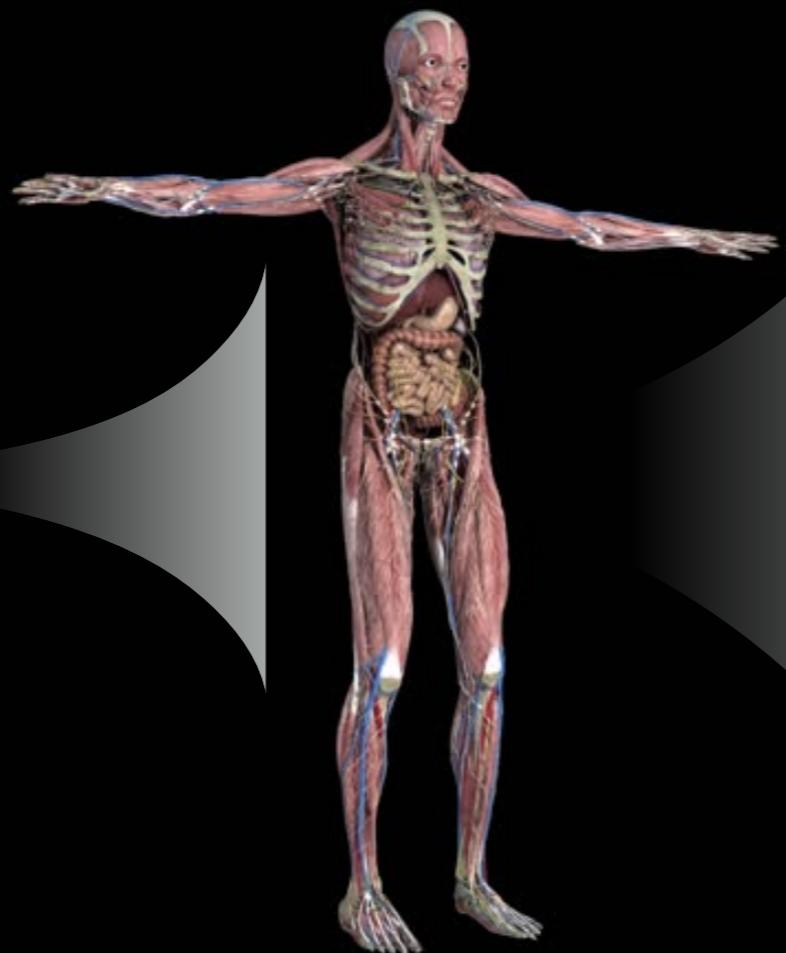
<http://www.dnalc.org/>

non-coding RNA



<https://www.youtube.com/watch?v=gZZyxVP02UU>

One cell One genome



Hundreds
of cell types

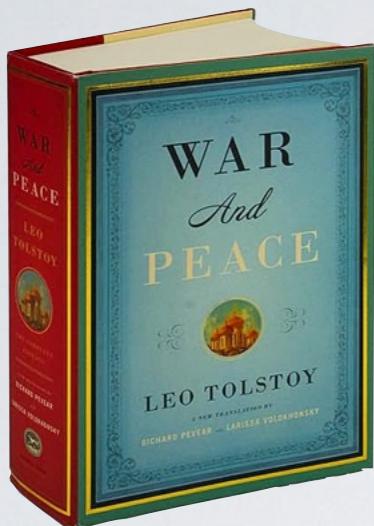
The form, function and behaviour of cells
is underpinned by the genes which
are (or are not) expressed



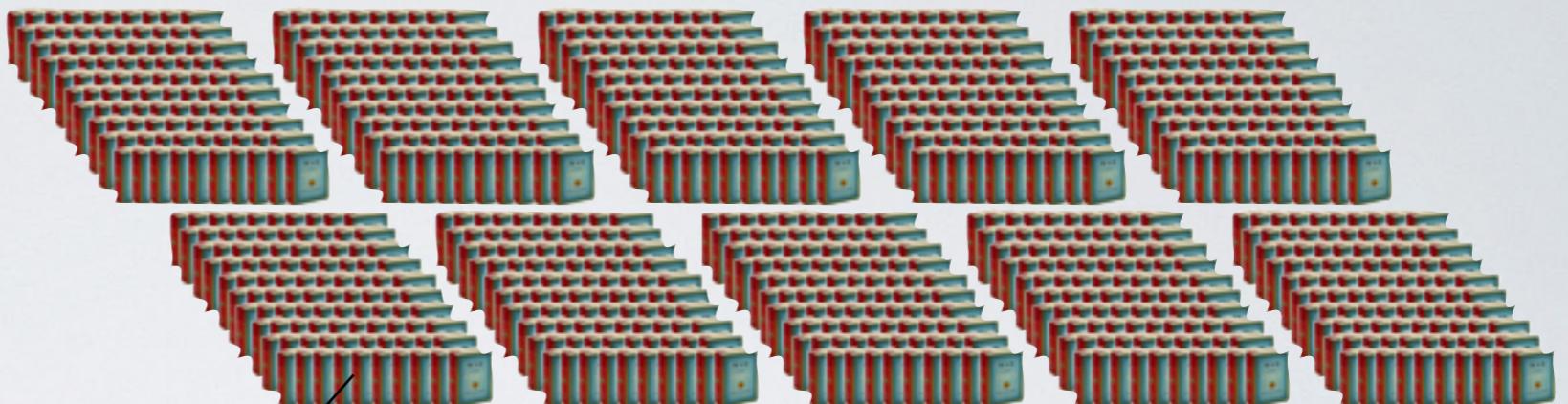
The human genome



Identifying information within a genome sequence



3 million letters

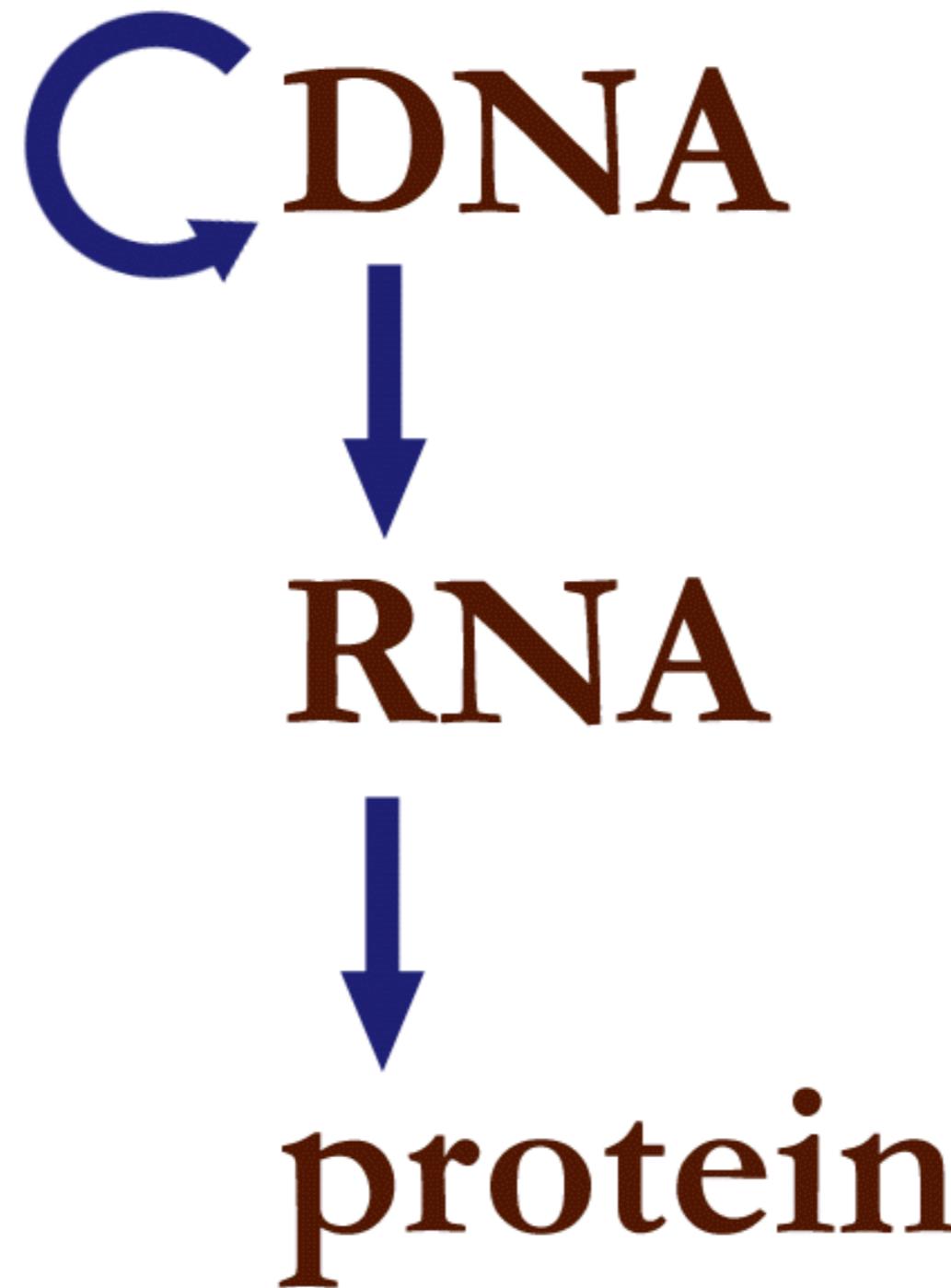


Human genome = ~3 billion letters
= 1000 copies of War and Peace

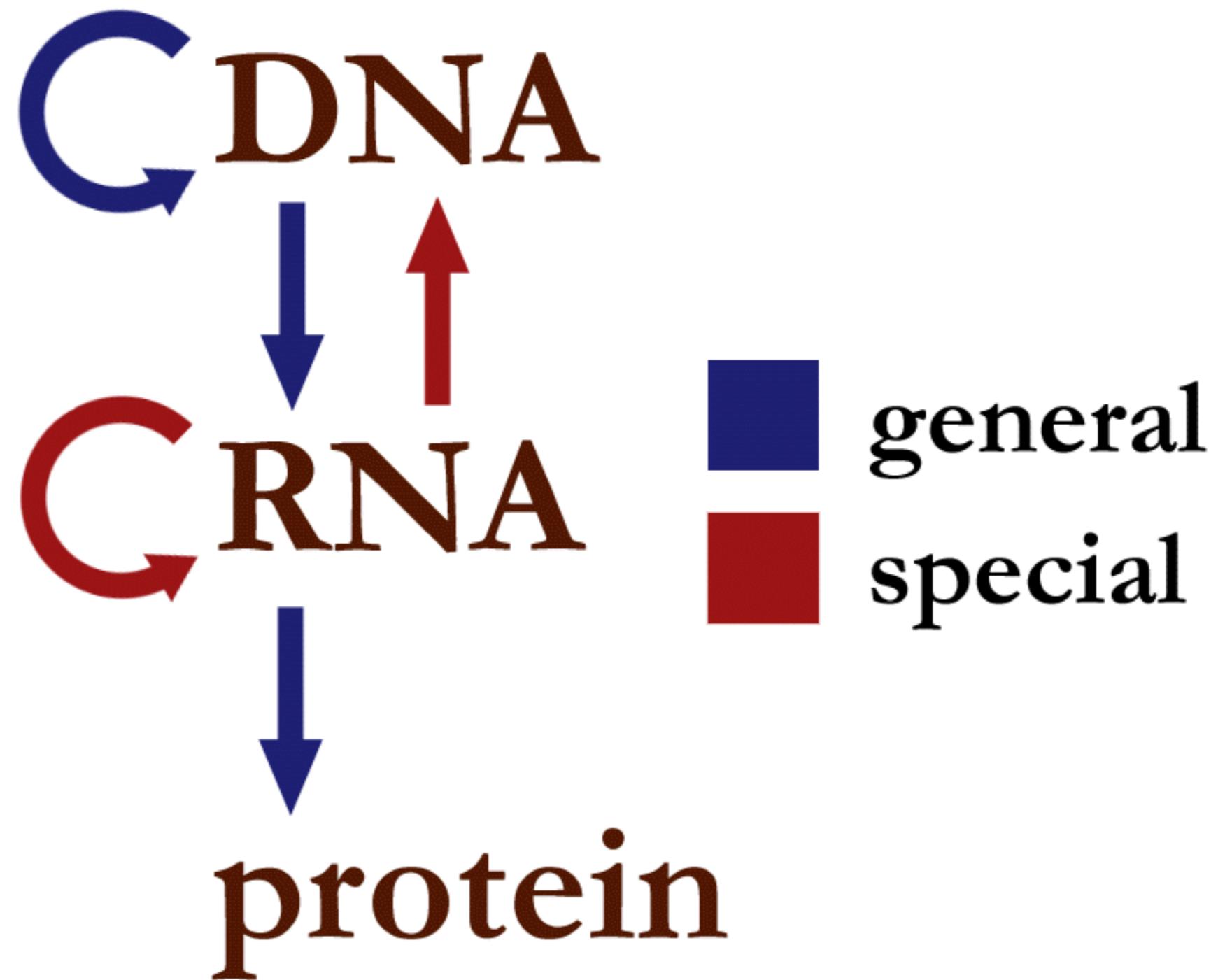
```
ACGCTCAGCGCGAGCTACTATGGACACGGCAGGCAGACTATCAGCGACGAGCTACTAGCAGCAG  
CGCGGCAGGACGTACACTATCGACGCCGATTATAGGCAGACGTATATGGCCGATTATAGCACGC  
GCGGCACACAGCTACTGTACGTAGCTGATCGATCGCGACGAGGCAGGGTCTATTGCGAGGCTAT  
TCTTATGGCAGAGAGGATATATATGCGCCGACAGCGATTACTATCGACGGACTATCTGAGCAGC  
AGCGGACTACTATCTTATTACGGACGAGCGGGGGCAGACGTATATGGCCGATTAGATTATCATATT  
AGATATTAGCGATTATGGCGATATAGGAGAGATTGAGCGACAGCGATTACTATCGAGCAGCAGC  
CTGAGCGACGAGCGGACTACTATTCGAGGCTATTCTTATTATACCGAAGGCAGATTATGGCGATA  
TAGGAGAGATTGAGCGACAGCGATTACTATCGAGCGACTATCTCGACAGCGATTACTATCGAGC  
ACTATCTGAGCGACGAGCGGACTACTATCTTATTACGGACGAGCGGGGGCAGACGTATATGGCCG  
CATTAGAGCGACGAGCGGACTACTATCTTATTACGGACGAGCGGGGGCAGACGTATATGGCCGCA  
TTATAGCACGCGCGTGGCGAGAGAGGATATATGCGCGATTATATTAGATATAGAGGATATAT  
ATGCGCGATTATATTAGATATTAGCGATTATGGCGATATAGGAGAGATTGAGCGACAGCGATT  
ACTATCGAGCGACTATCTGAGCGACGAGCGGACTACTATCTTATTACTAGGCAGATTATGGCGAT  
AGGAGAGATTGAGCGACAGCGATTACTATCGAGCGACTATCTGAGCGACGAGCGGACTACTATCT  
TATTACGGACGAGCGGGGGCAGACGTATATGGCCGATTATAGCACGGCGCTATCTTATTACGG  
ACGAGCGGGGAGGCTATTCTGATATAAAAATGCGAGCAGCTATCTAGCGACGTATATTAAATA  
TTACGATCGTATTACTTATTATCTACGGACGACGTACTGTACGTAGCTGATC  
GATCGCGACGAGCGGGAGCTATTGCGAGGCTATTCTTATGGCGAGAGAGGATATATGCGCG  
ATTCTATAGATATTAGCGATTATGGCGATATAGGAGAGATTGAGCGACAGCGATTACTATCG  
AGCGACTATCTGAGCGACGAGCGGACTACTATCTTATTACGGACGAGCGGGGGCAGACGTATATG  
GCCGATTATAGCACGGCGCACAGCGTACTGTACGTAGCTGATCGATCGGGACGAGCGGGGA  
GTCTATTGCGAGGCATTCTTATGGCGAGAGAGGATATATGCGCGATTATATTAGATATTAGG  
CGATTATGGCGATATAGGAGAGATTGAGCGACAGCGATTACTATCGAGCGACTATCTGAGCGACG  
AGCGGACTACTATCTTATTACGGAAAGGCAGATTATGGCGATATAGGAGAGATTGAGCGACAGCGAT  
TACTATCGAGCGACTATCTGAGCGACGTACTATCGAGCGACTATCTGAGCGACGAGCGGACTAC  
TATCTTATTACGGACGAGATGGGGCAGACGTATATGGCCGATTAGAGCGACGAGCGGACTACTA  
TCTTATTACGGACGAGCGGGGGCAGACGTATATGGCCGATTATAGCACCGCGGCAACAGCTA  
CTGTACGTAGCTGATCGATCGCGACGAGCGGGAGCTATTGCGATATAAAAATGCGAGGAGGAT  
ATATATGCGCGATTATATTAGCGTATATGGCCGATTATAGCACGGCGCTATCTTATTACGG  
CGAGCGGGGGAGGCTAGATATTAGCGATTATGGCGATATAGGAGAGATTGAGCGACAGCGATTAC  
TATCGAGCGACTATCTGAGCGACGAGCGGACTACTATCTTATTACGGACGTACTATCTGAGCG  
ATTAAATATTACGATCGATCGTATTACTTATTATCTATCTACGGACGACGTATACTGTACGT  
GTTATTACGGAAAGGCAGATTATGGCGATATAGGAGAGATTGAGCGACAGCGATTACTATCGAGCG  
CTATCTGACAGCGATTAGTCAGAGCGACTATCTGAGCGACGAGCGGACTACTATCTTATTACGG  
ACGAGCGGGGGCAGACGTATATGGCCGATTAGAGCGACGAGCGGACTACTATCTTATTACGGAC  
GAGCGGGGGGGCAGACGTATATGGCCGATTATAGCACCGCGGCTGATCGATCGCGACGAGCG  
GAGTCTATTGCGAGGCTATTCTTATGGCGAGAGAGGATATATGCGCGATTATATTAGATATT  
GGCGATTATGGCGATATAGGAGAGCGACGAGCGATTACTATCGAGCGACTATCTGAGCGACGCG  
CTACTATCTTATTACGGACGAGCGGGGGGGCAGACGTATATGGCCGATTAGATTGAGCGACAGCG  
ATTACTATCGAGCGACTATCTGAGCGACGAGCGGACTACTATCTGAGCGACGAGCGGACTACTATC
```

sequence that is transcribed

Traditionally, RNAs were mainly thought to form
a bridge between DNA and protein



Traditionally, RNAs were mainly thought to form
a bridge between DNA and protein



Classes of RNA molecules

mRNA	Messenger RNA carries information from DNA to the ribosome, the sites of protein synthesis (translation) in the cell.
rRNA	Ribosomal RNA
tRNA	Transfer RNA
miRNA	Micro RNA
lncRNA	Long non-coding RNA
eRNA	Enhancer RNA: short RNAs transcribed from enhancer regions.
snRNA	Small nuclear RNA: processing pre-mRNA in the nucleus
snoRNA	Small nucleolar RNA: primarily guide chemical modifications of other RNAs

Also siRNA, piRNA, exRNA, scaRNA... and the list goes on

What is a transcriptome?

Transcriptome: the complete set of transcribed sequences

After genomes were sequenced, major projects aimed to determine
the complete set of transcribed sequences

ENCYCLOPEDIA OF DNA ELEMENTS PROJECT

The ENCODE Project Consortium

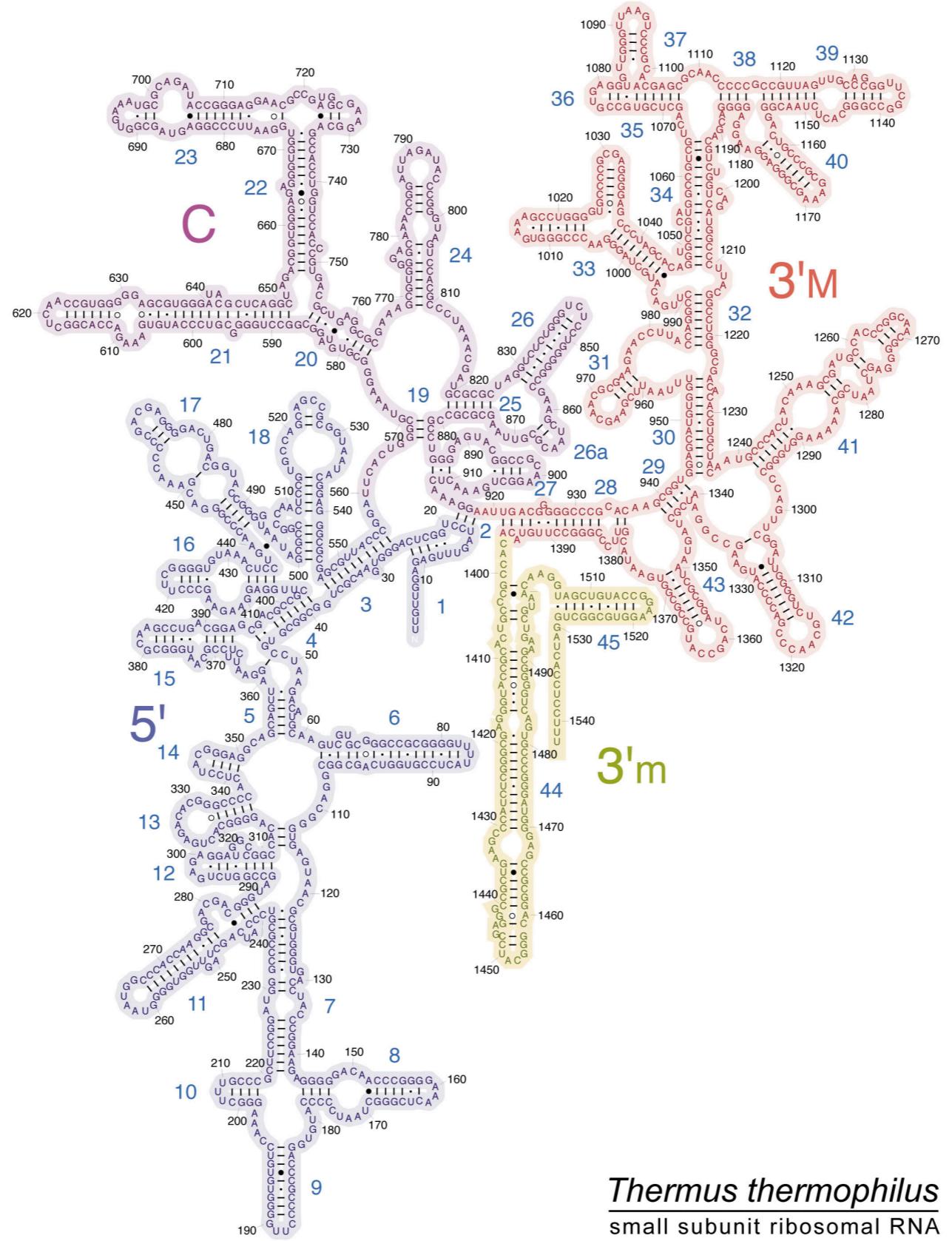
- Exons of protein-coding genes cover 2.9% of the genome
- Protein-coding genes span 39.5% of the genome
- 62,403 transcription start sites (TSSs)
- On average per locus: 6.3 alternatively spliced transcripts
3.9 different protein-coding transcripts
- 8,801 identified small RNAs (piRNAs, miRNAs, siRNAs)
- 9,640 long non-coding RNA (lncRNA) loci
 - transfer RNA, microRNA, small nuclear RNA, small nucleolar RNA
 - antisense RNA, enhancer associated RNA, many types of small RNA

Pervasive transcription:

- > 85% of the genome is transcribed
- > 50,000 long intergenic non-coding RNAs (lncRNAs)

What is the most abundant
form of RNA in a cell?

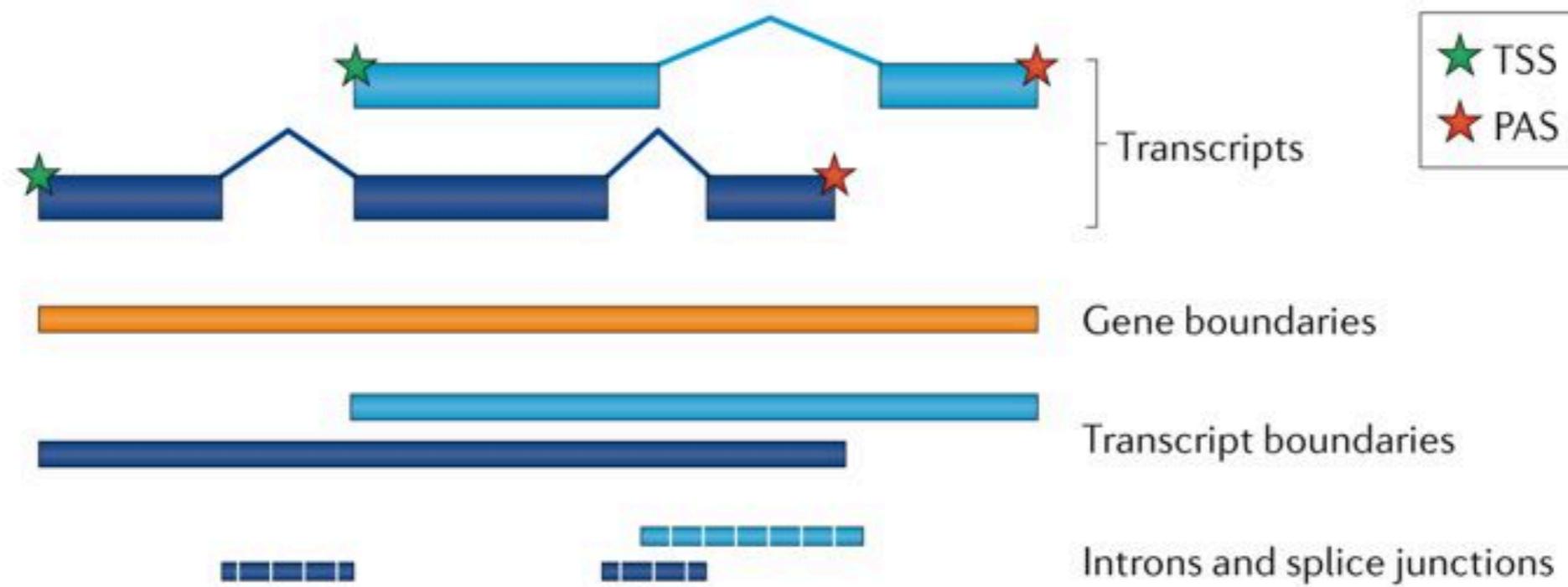
Ribosomal RNA



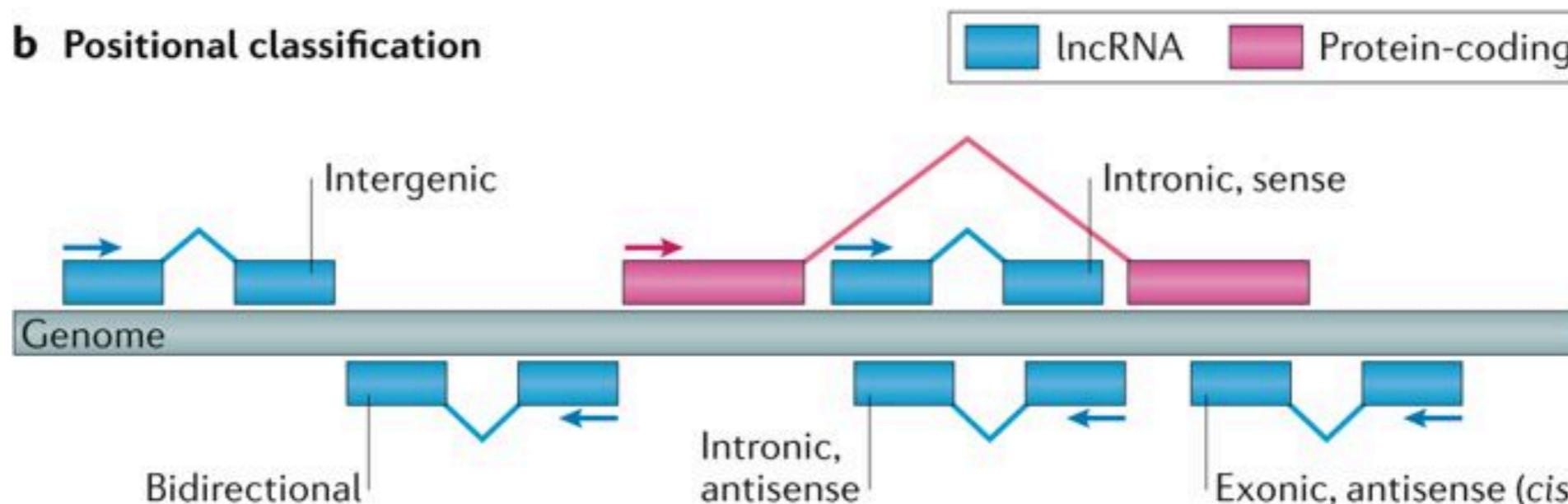
Thermus thermophilus
small subunit ribosomal RNA

Basic concepts of lncRNA annotations

a Definition of annotation structures

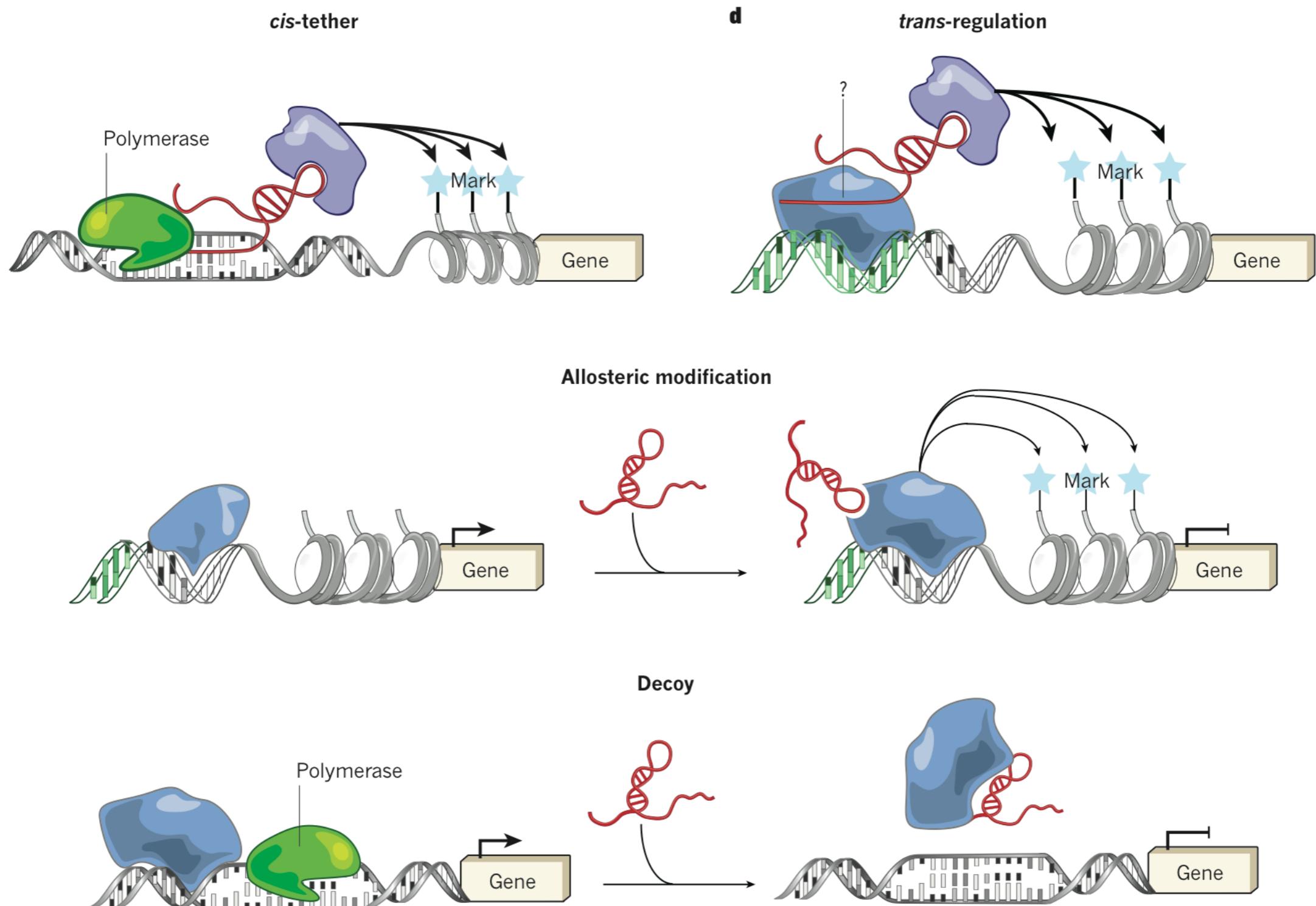


b Positional classification

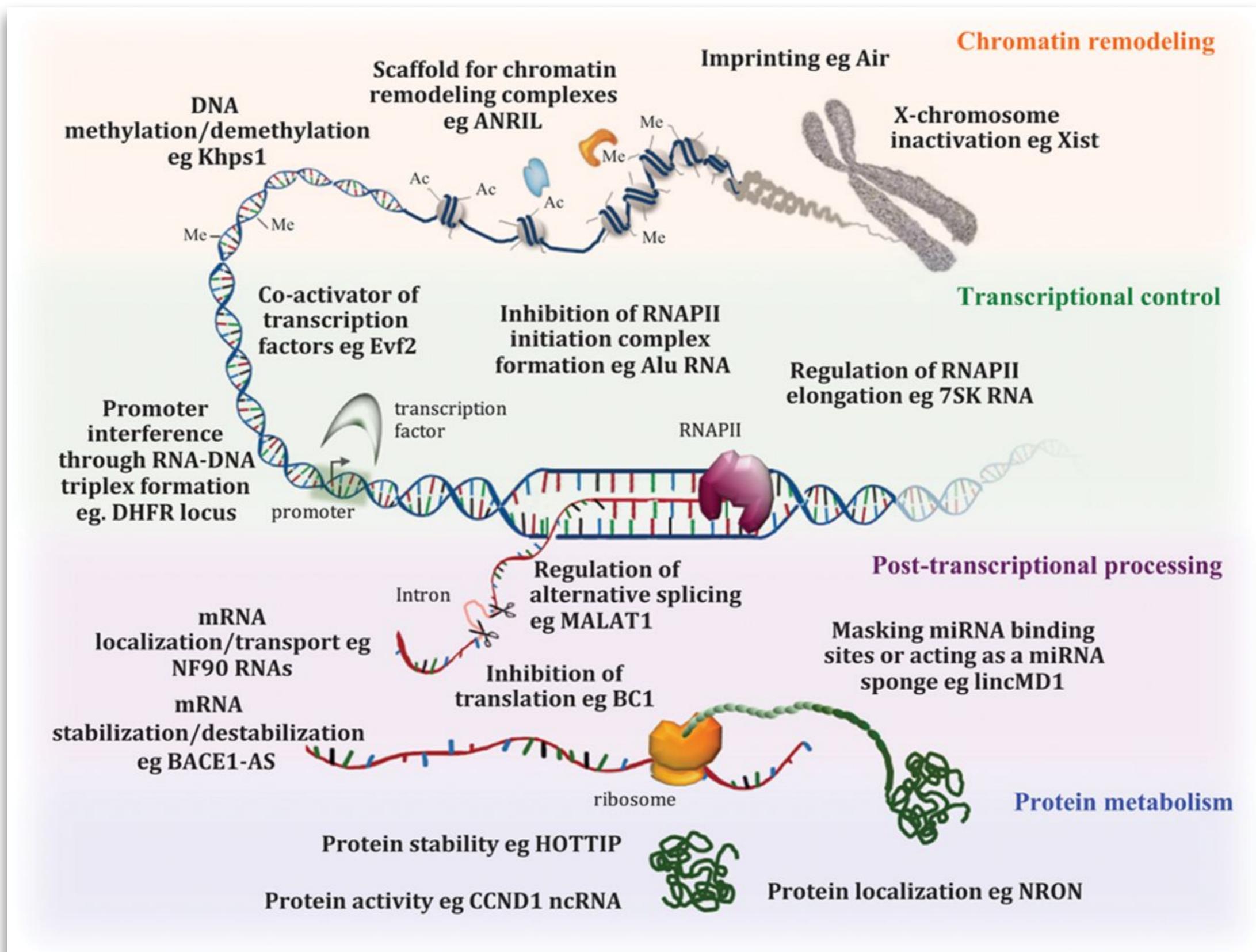


A complex RNA world

Models of regulatory roles for long non-coding RNAs



A complex RNA world



Transcriptome: cellular information in RNA-form

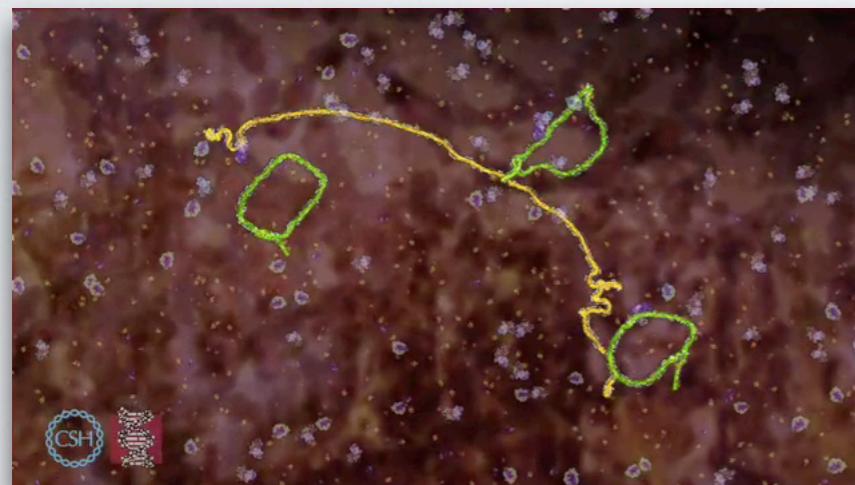
<http://www.dnalc.org/>



gene expression:

- which genes (or genomic sequences) are expressed?
- when are they expressed?
- what is the transcriptional start site?
- how many mRNA copies are transcribed?
- what is the rate of transcription?

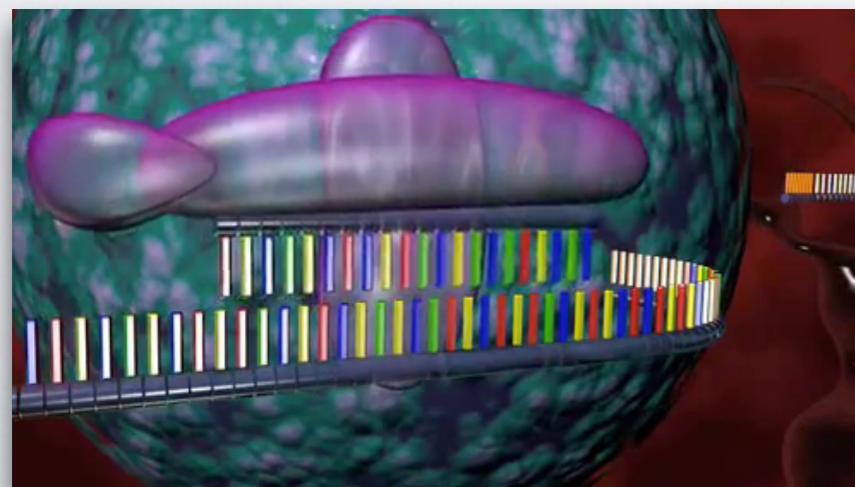
<http://www.dnalc.org/>



splicing:

- what is the pattern of splicing for each mRNA?
- what is the relative abundance of each splice isoform?

<https://www.youtube.com/watch?v=gZZyxWP02UU>



non-coding RNAs:

- which micro RNAs (miRNAs) are produced?
- what are the miRNA targets?
- which mRNAs are being actively degraded by miRNAs?
- what long non-coding RNAs are transcribed?
- what are the interactions between proteins and non-coding RNAs?

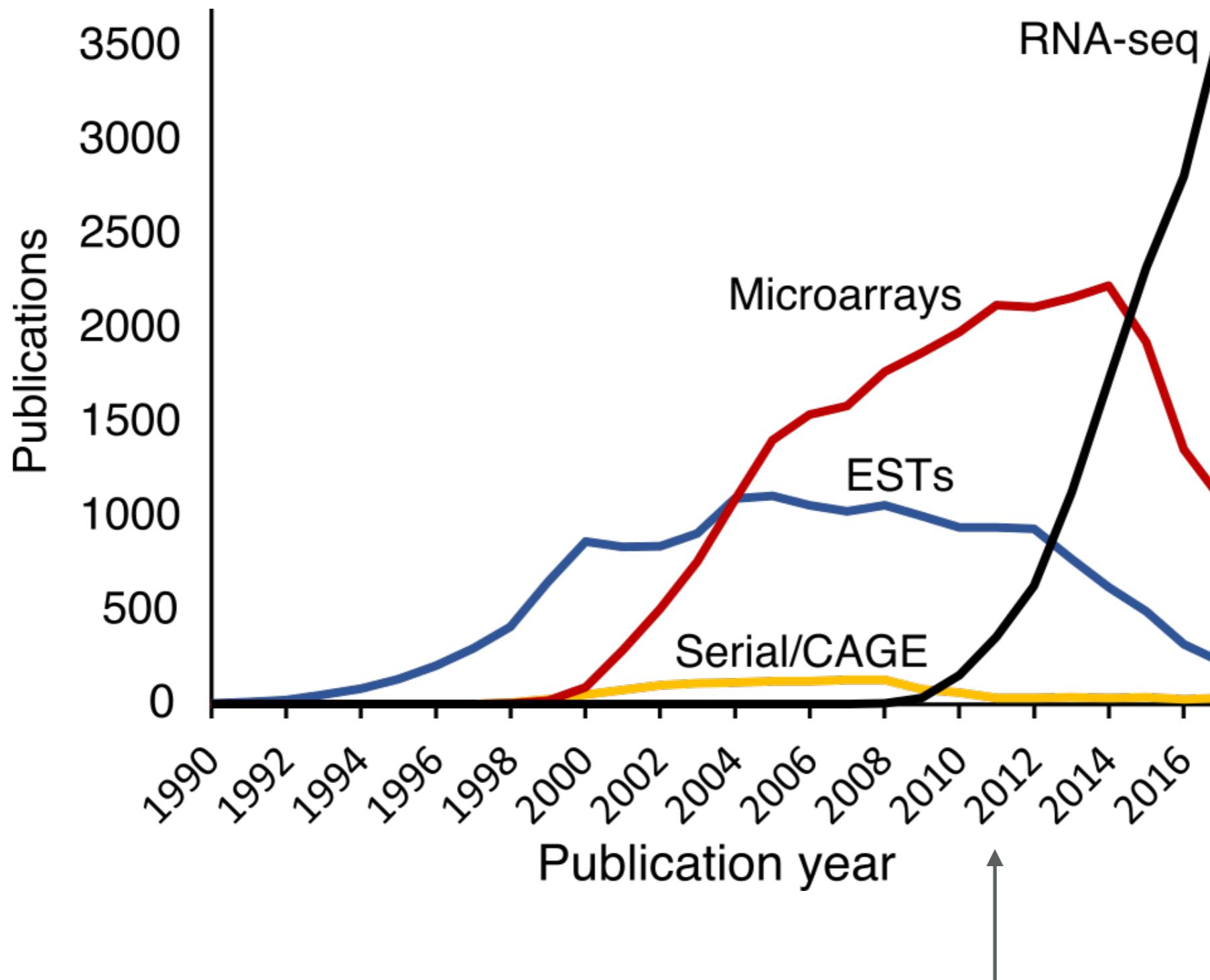
Goal of transcriptome analysis
Identify, characterize and catalogue all transcripts expressed within a specific cell or tissue, at a particular stage, under various conditions

The word *transcriptome* emerged in the 1990's

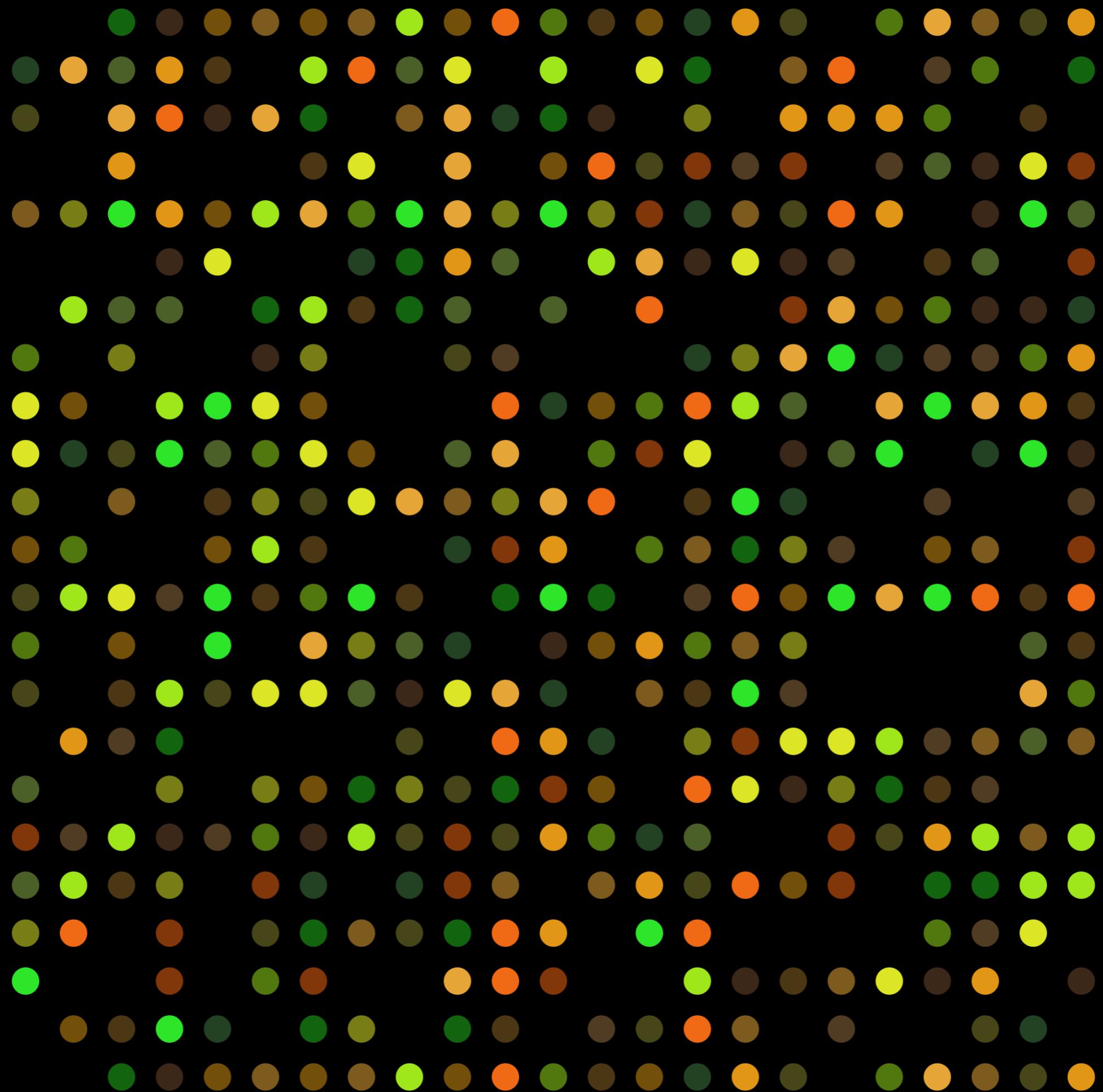
Technique	Method	Emergence	Technology
ESTs	Random	1980-90's	Sanger
SAGE	Random concatenation	Early 90's	Sanger
Microarray	Known	1995	Hybridisation
RNA-seq	Random	2008	Massively parallel sequencing

Transcriptomics is characterised by the development of new techniques that redefine what is possible every decade and render previous technologies obsolete.

Use of transcriptome profiling technologies

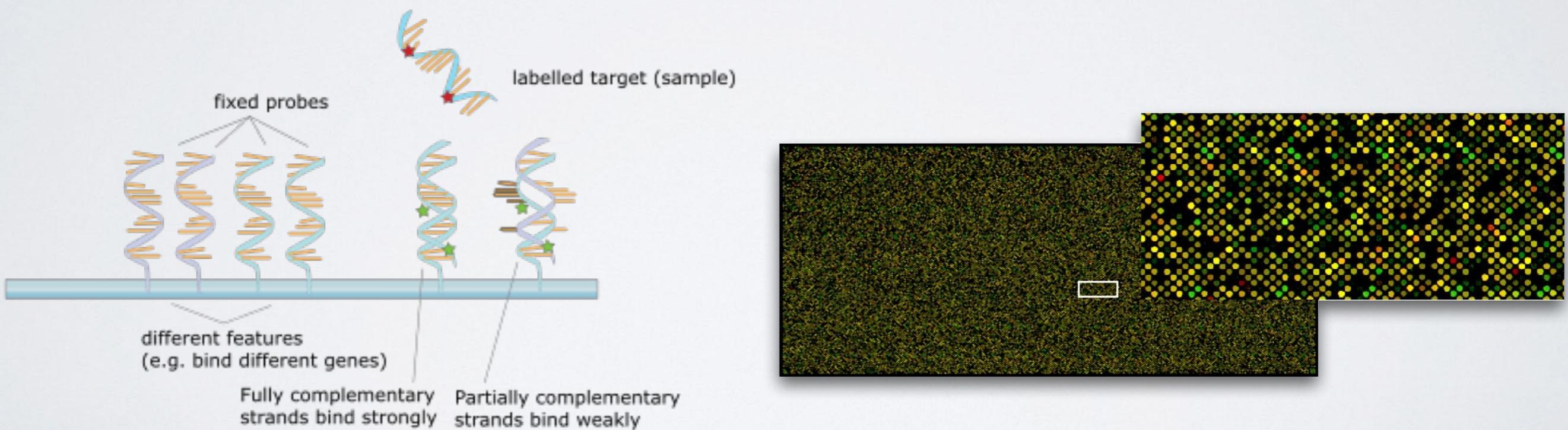


What is a microarray?



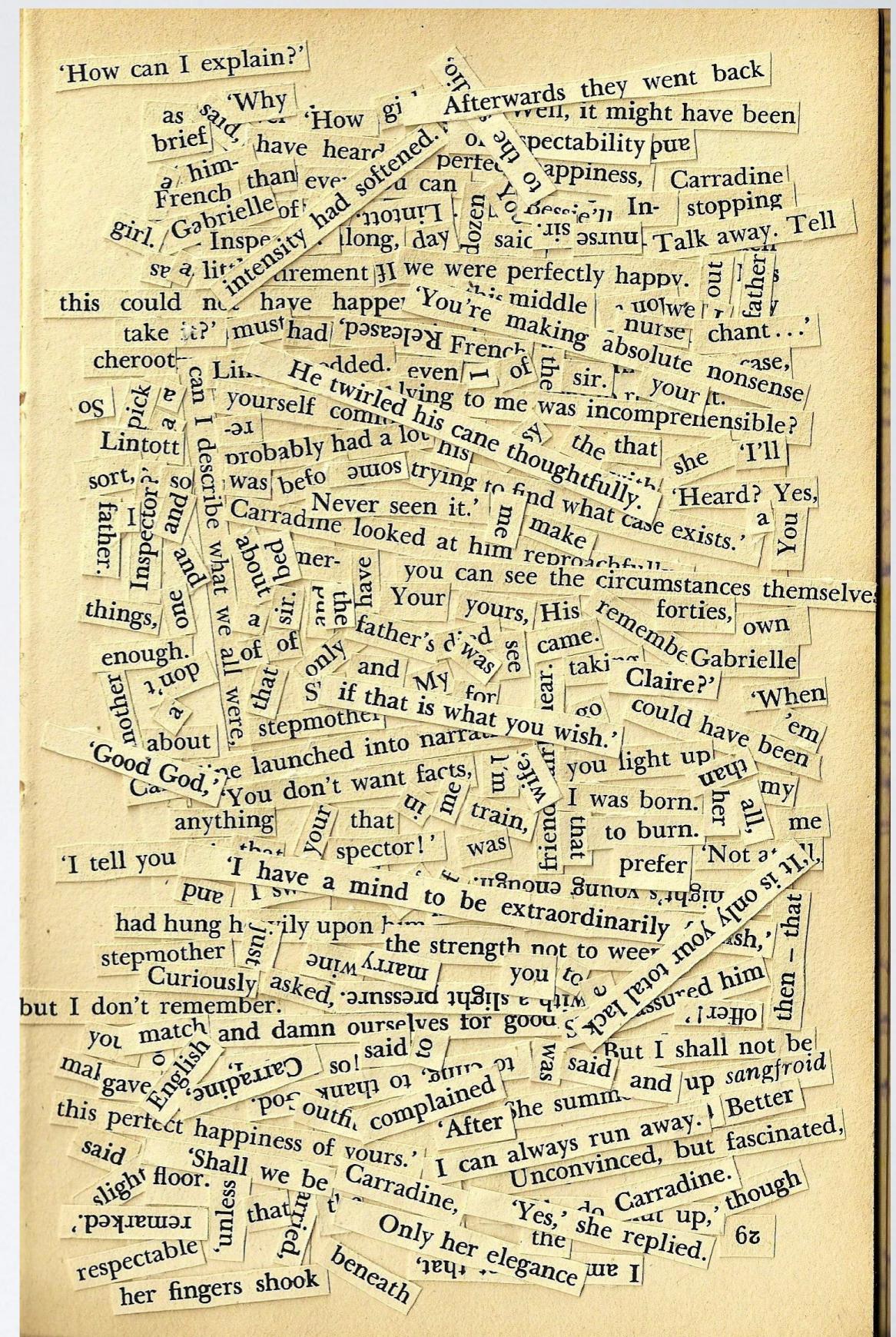
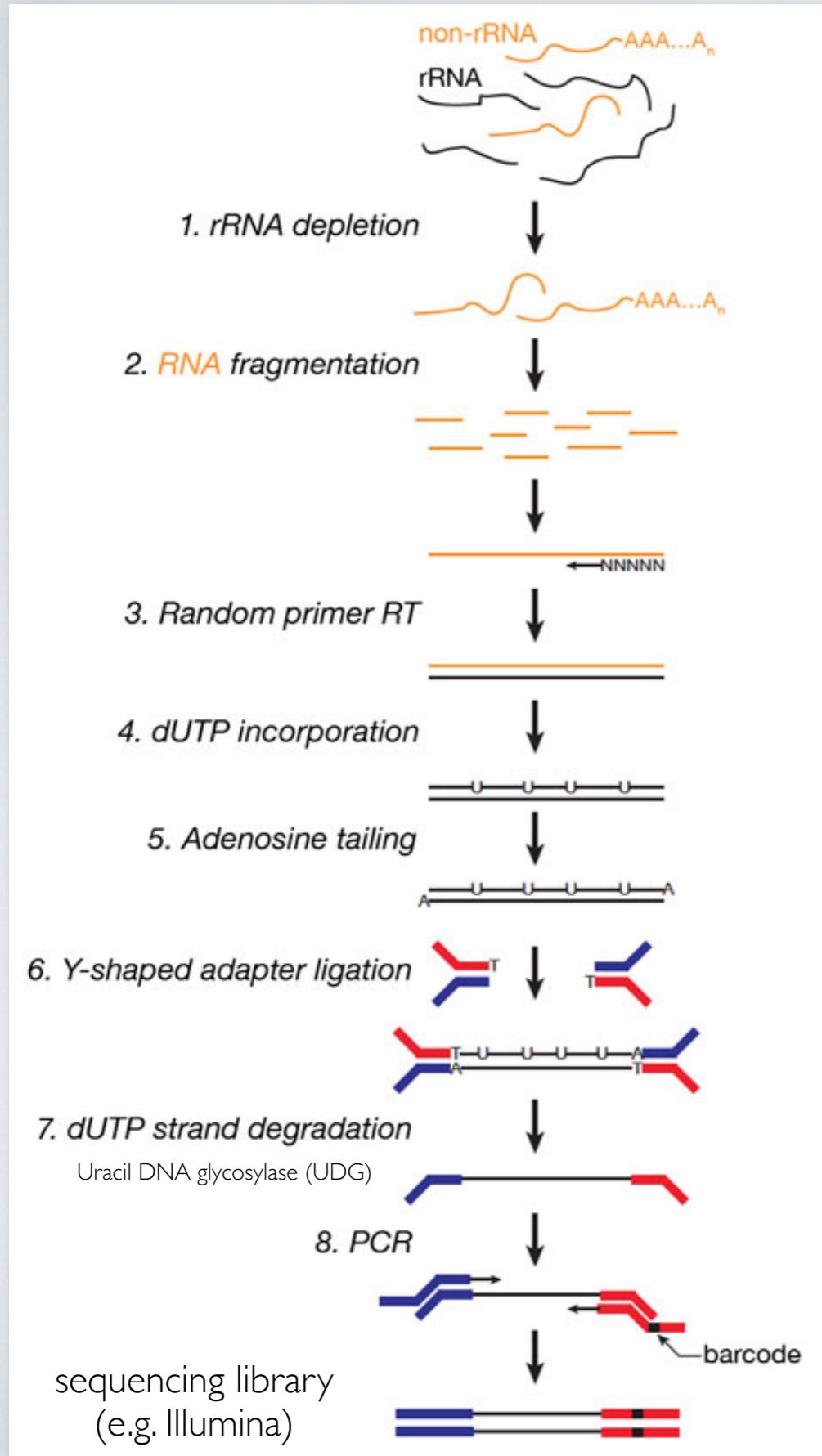
Development of transcriptome analysis

DNA microarrays: microscopic spots of DNA arranged in an ordered grid, each spot containing DNA oligonucleotides of a particular sequence. Hybridize fluorescently labelled cDNA derived from mRNA from a sample to the microarray and measure fluorescent signal from each spot. Intensity of signal is a measure of gene expression.



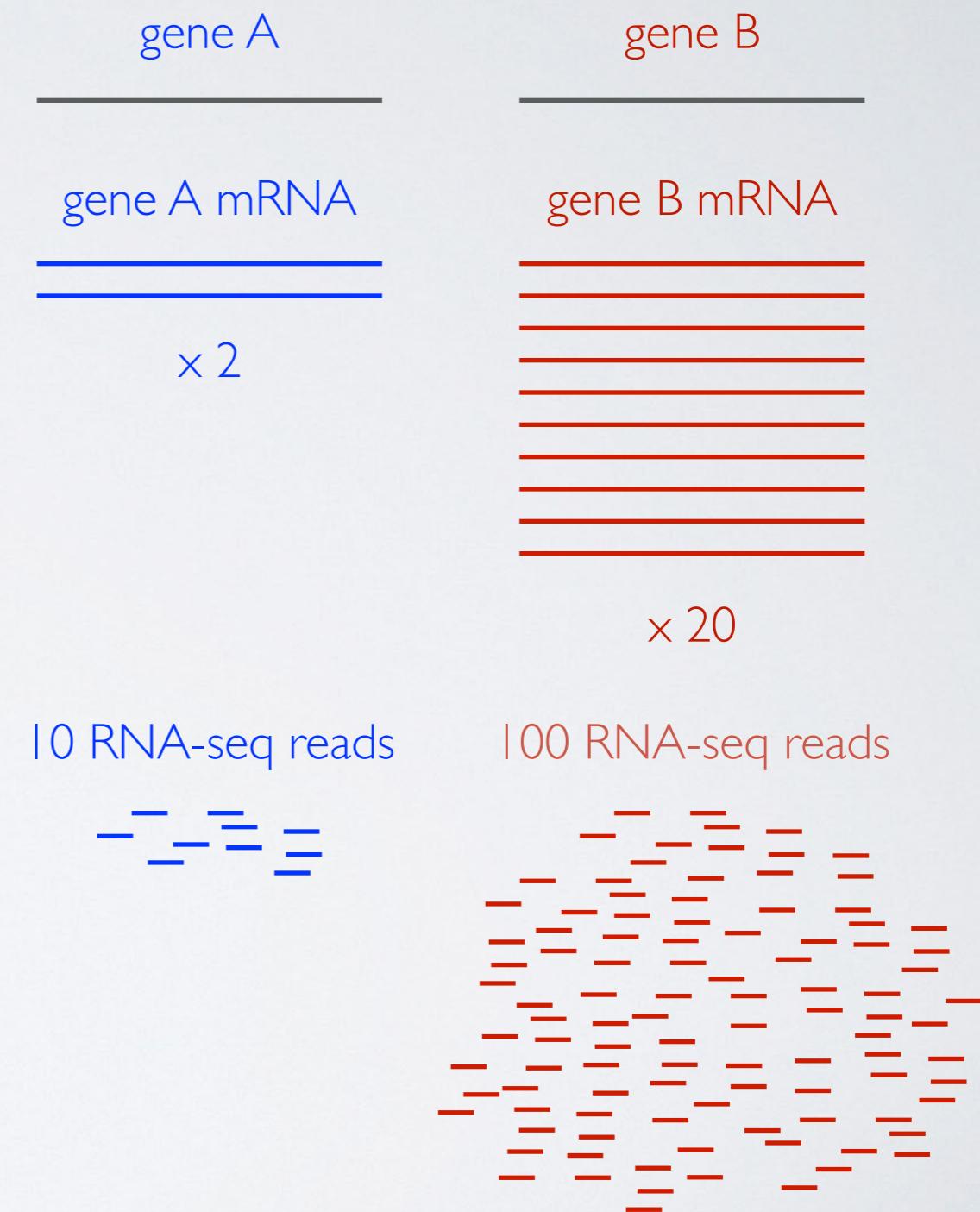
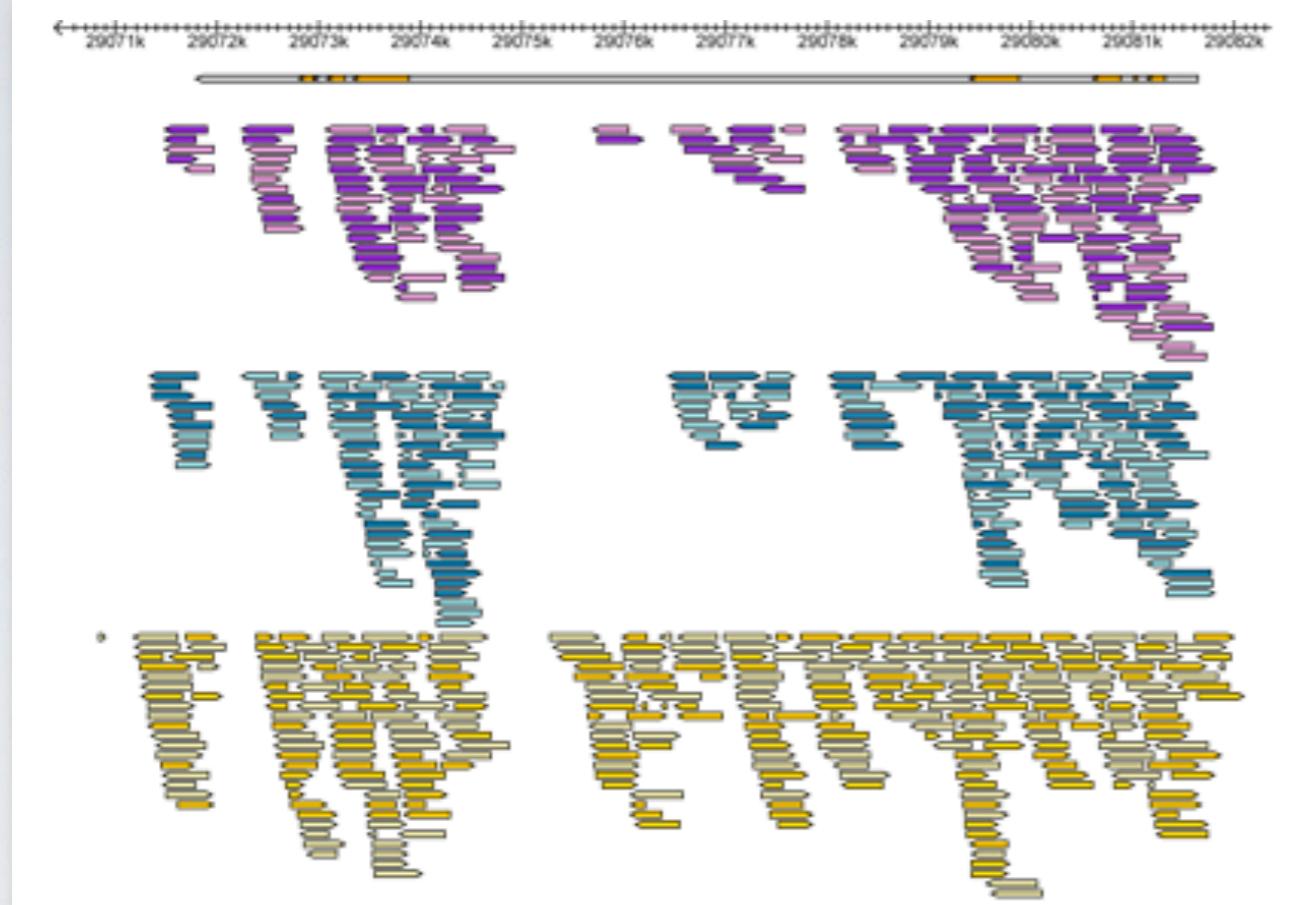
Microarrays are still widely used for high throughput genotyping

RNA-seq: massively parallel sequencing of RNA fragments



RNA-seq: massively parallel sequencing of RNA fragments

Quantitation of abundance of all transcripts



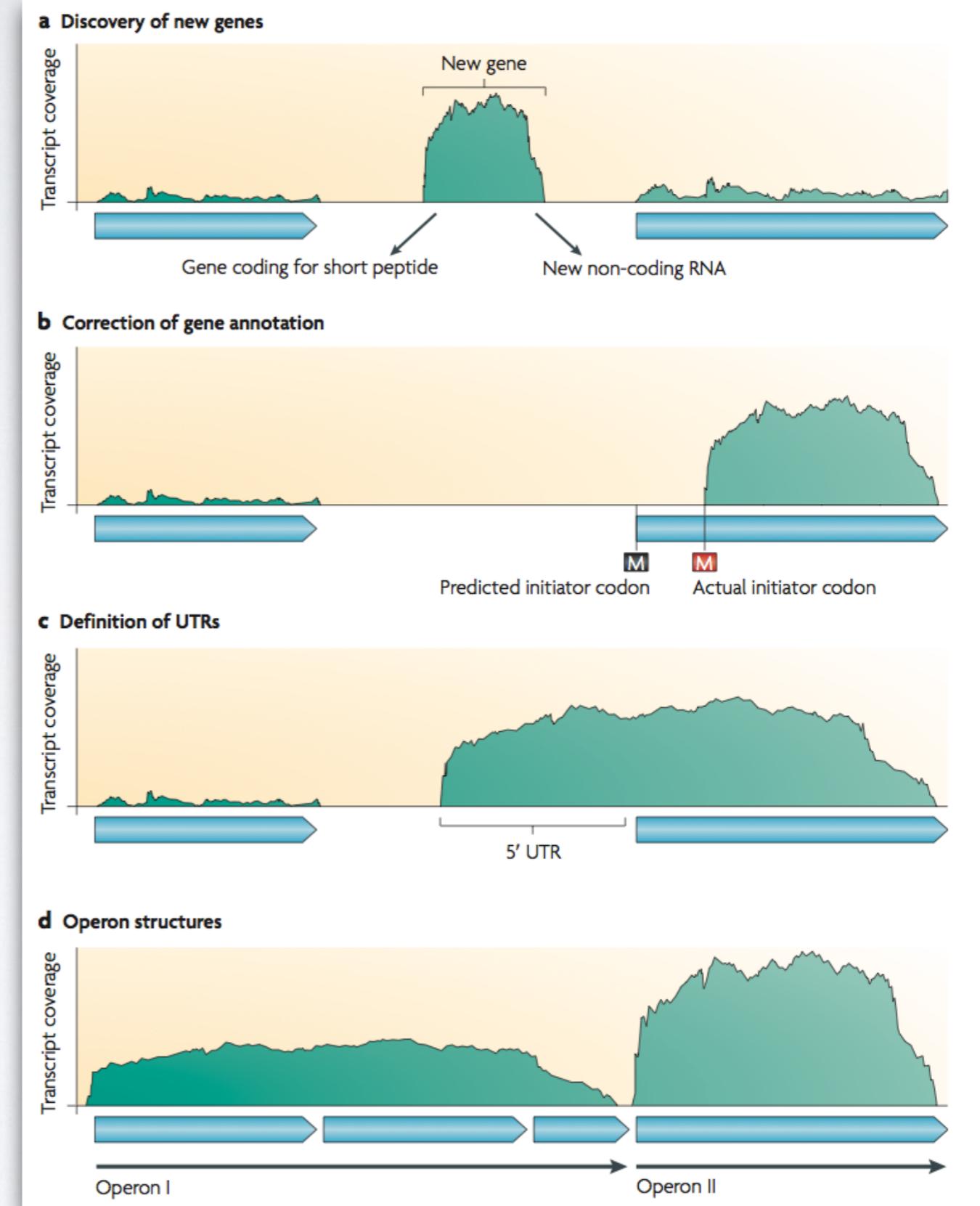
How does RNA-seq compare?

Technology	Tiling microarray	cDNA or EST sequencing	RNA-Seq
Technology specifications			
Principle	Hybridization	Sanger sequencing	High-throughput sequencing
Resolution	From several to 100 bp	Single base	Single base
Throughput	High	Low	High
Reliance on genomic sequence	Yes	No	In some cases
Background noise	High	Low	Low
Application			
Simultaneously map transcribed regions and gene expression	Yes	Limited for gene expression	Yes
Dynamic range to quantify gene expression level	Up to a few-hundredfold	Not practical	>8,000-fold
Ability to distinguish different isoforms	Limited	Yes	Yes
Ability to distinguish allelic expression	Limited	Yes	Yes
Practical issues			
Required amount of RNA	High	High	Low
Cost for mapping transcriptomes of large genomes	High	High	Relatively low

Wang, Gerstein, Snyder, RNA-Seq: a revolutionary tool for transcriptomics, Nat Rev Genet 10, 57–63 (2009)

Identifying information within a genome sequence

ACGCTCAGCGCAGACTATCGCACCGGCAGGGCAGACTATCGCAGCGAGCTACTAGCAGCAC
CGCGCAGGACGTACACTATCGACGCCGATTATAGCAGACGTATATGCCGCATTATAGCAGC
GCGGCACACAGCTACTGTAGCTAGCTAGCAGCCGACGAGCGGGAGTCATTGCGAGGCTAT
TCTTATGGCGAGAGAGGATATATCGCCGACAGCAGGATTACTATCGAGCAGCTATCGAGCGAC
AGCGGACTACTATCTTATTACGGACGGAGCGGGGGCAGACGTATATGCCGCATTAGATTCA
AGATATTAGCGATTATGGCGATAGGAGAGATTGAGCGACAGCAGGATTACTATCGAGCAGCTAT
CTGAGCGACGAGCGACATCTATTGCGAGGCTATTCTTATTACGAAAGGCGATTATGCCGATA
TAGGAGAGATTGAGCGACAGCAGGATTACTATCGAGCAGCTATCTGACAGCAGGATTACTATCGAGC
ACTATCTGAGCGACGAGCGGACTACTATCTTATTACGGACGAGCGGGGGCAGACGTATATGCCGC
CATTAGAGCGACGAGCGGACTACTATCTTATTACGGACGAGCGGGGGCAGACGTATATGCCGC
TTATAGCAGCGCGTGGCAGAGAGGATATATCGCGATTCTATTAGAGATATAGAGGATATAT
ATGCGCATTCTATTAGATATTAGCGATTATGGCGATAGGAGAGATTGAGCGACAGCAGGATT
ACTATCGAGCGACTACTGAGGACGAGCGGACTACTATCTTATTACTAGCGATTATGCCGATA
AGGAGAGATTGAGCGACAGCAGGATTACTATCTTATTAGCGAGCGGGGAGCTACTATCT
TATTACGGACGAGCGGGGGCAGACGTATATAGCGAGCGGAGGCTACTATCTTATTACG
ACGAGCGGGAGGCTATTCTATCGATA
TTACGATCGATCGTATTACTTATTATCT
GATCGCGACAGCGGGAGTCTATTGCGAAG
ATTCAATTAGATATTAGCGATTATGCCGA
AGCGACTATCTGAGCGACGAGCGGACTACT
GCCGCATTATAGCAGCGCGCACACAGCT
GTCTATTGCGAGGCTATTCTATGGCAGA
CGATTATGGCGATAGGAGAGATTGAGCGACAGCAGGATTACTATCGAGCAGCTATCTGAGCGAC
AGCGGACTACTATCTTATTACGAAAGGC
TACTATCGAGCGACTATCTGAGCGACAGCAG
TACTATCTTATTACGGACGAGATGGGGGAG
TCTTATTACGGACGAGCGGGGGCAGACG
CTGTCAGCTAGCTAGCGATCGCGAGGGGGAGTCTATTGCGATATAAAAAAAATCGAGGAGA
ATATGCGCAGTCTATTAGACGCTATCGCGCATTATAGCAGCGGCACACAGCT
TATCGAGCGACTATCTGAGCGACGAGCGGACTACTATCTTATTACGAGCTATCTGAGCGATTAT
TATCGAGCGACTATCTGAGCGACGAGCGGACTACTATCTTATTACGAGCTATCTGAGCGATTAT
ATTAATATTACGATGGAGCTATTACTCTATTATCTACGGAGCAGCTATCTGAGCGATTACTGAGCG
GTTTATTACGGAAGGCGATTATGGCGATAGGAGAGATTGAGCGACAGCAGGATTACTATCGAGCG
CTATCTGACAGCGATTAGTCAGAGCGACTATCTGAGCGACGAGCGGACTACTATCTTATTACG
ACGAGCGGGGGCAGACGTATATGCCGC
GAGCGGGGGCAGACGTATATGCCGC
GAGTCTATTGCGAGGCTATTCTTATTGCGAGAGAGGATATATGCGCAGTCTATTAGAG
GGCGATTATGGCGATAGGAGAGACGACAGCAGGATTACTATCGAGCGACTATCTGAGCGACGAGCG
CTACTATCTTATTACGGACGAGCGGGGGCAGACGTATATGCCGC
ATTAATCTGAGCGACTATCTGAGCGACGAGCGGACTACTATCTTATTACG
ATTAATCTGAGCGACTATCTGAGCGACGAGCGGACTACTATCTTATTACG

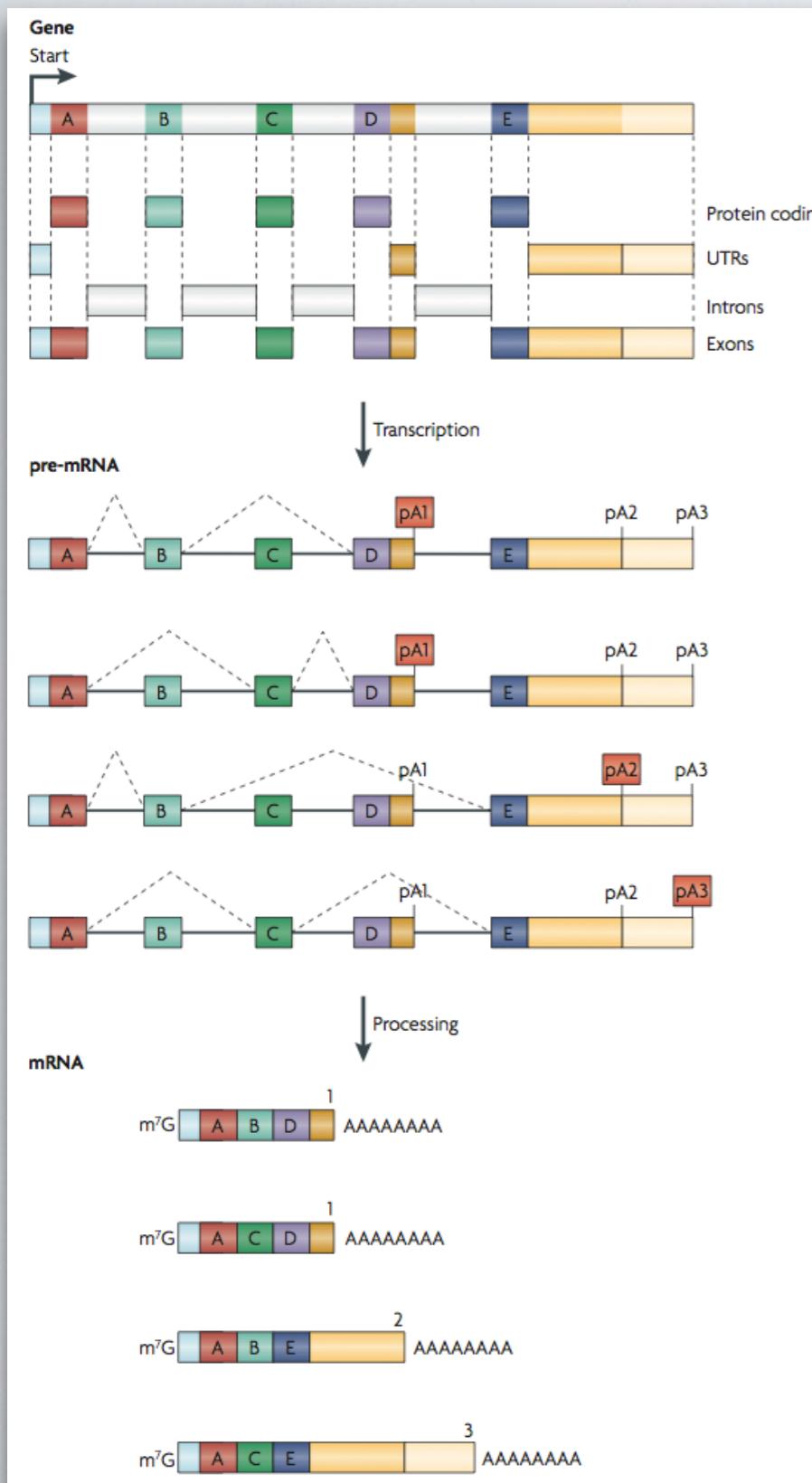


How do we identify information within the sequence of a genome?

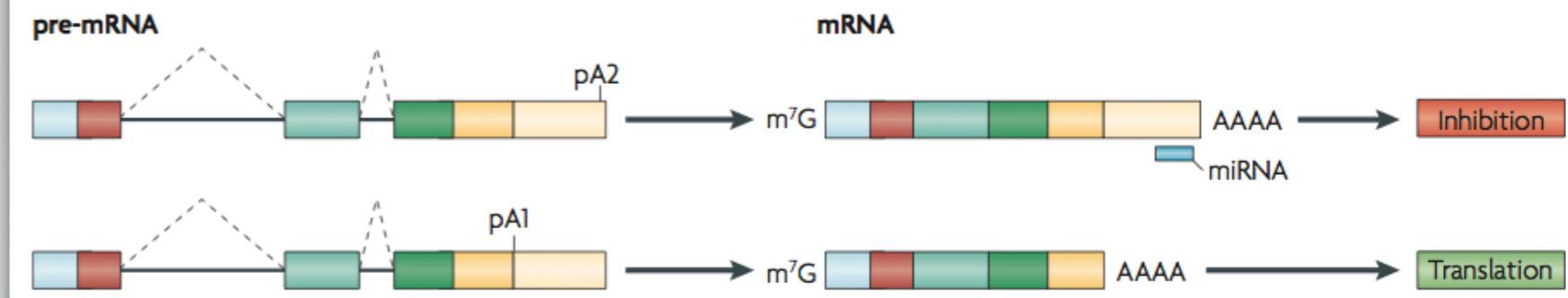
Sequence RNA to identify the genomic sequences that encoded the RNA

- map gene and exon boundaries
 - extensive transcript complexity
 - novel transcription
 - defining transcription level
 - identify splicing isoforms
 - map 5' and 3' boundaries

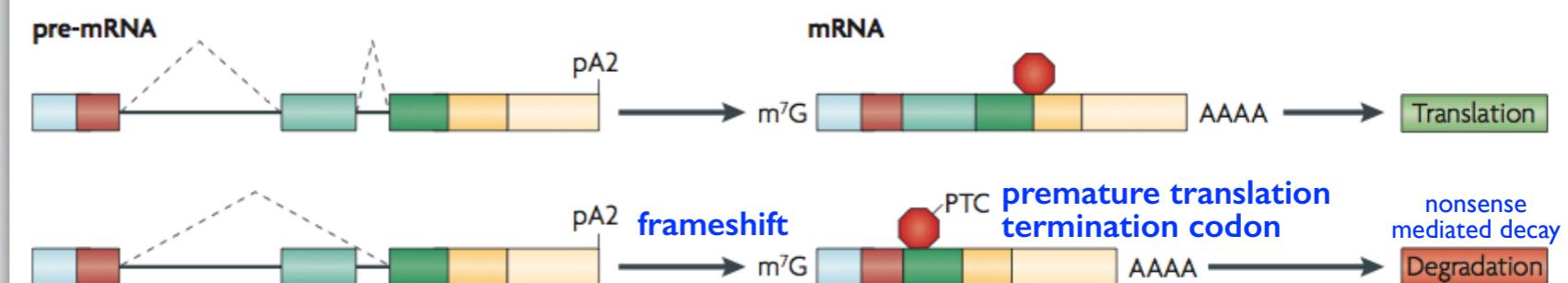
RNA diversity and functional complexity by alternative splicing



a Coupling RNA processing to alternative RNA regulation



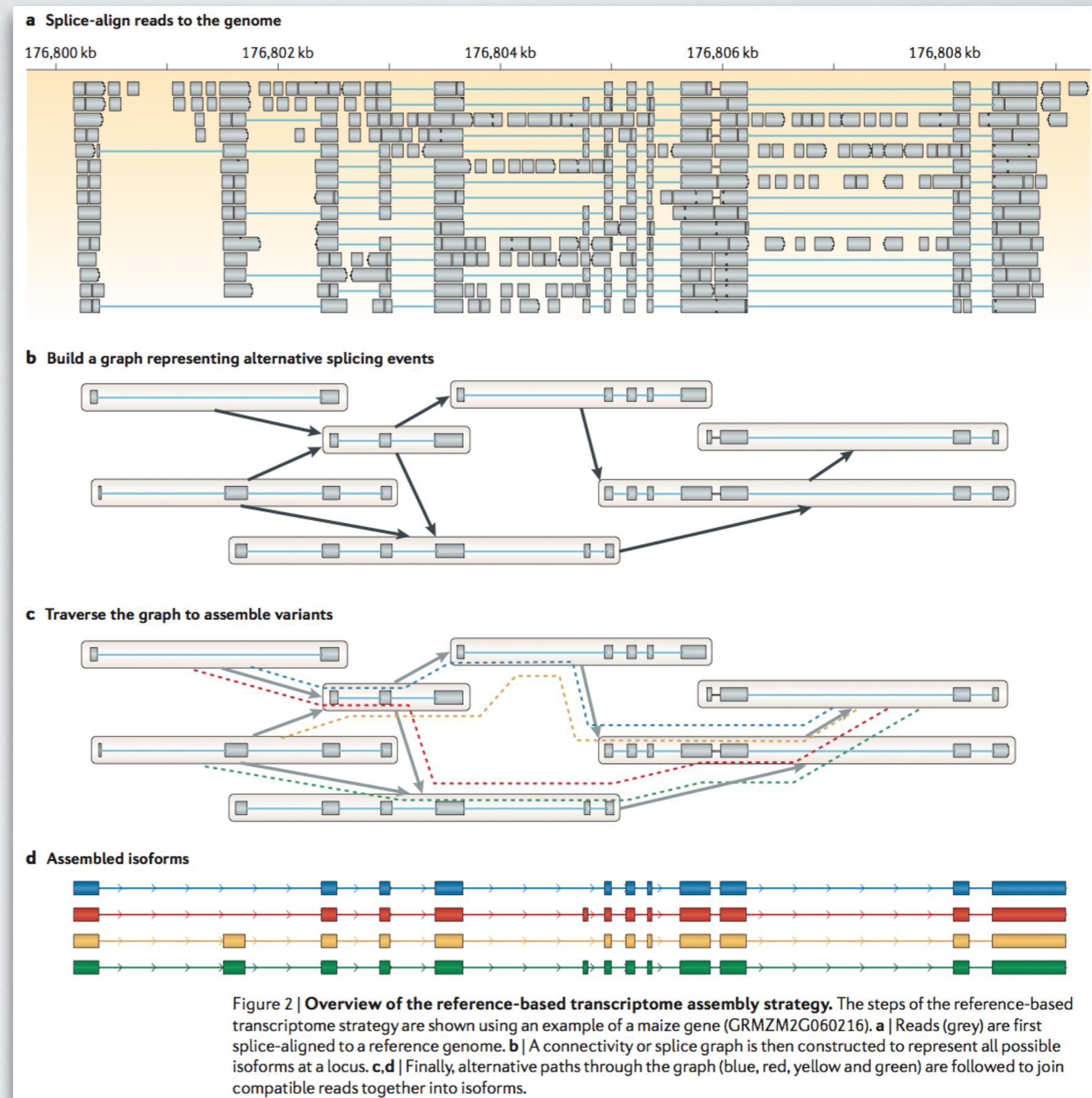
b Coupling alternative splicing with nonsense-mediated decay



Alternative splicing is a major pathway to generate RNA diversity. Comparative analysis of splicing variants yields insights into the biological consequences of this process

Reference based transcriptome assembly with RNA-seq

Uses RNA-seq short read fragments, aligned to the genome sequence, to deduce splice isoforms present in the cell

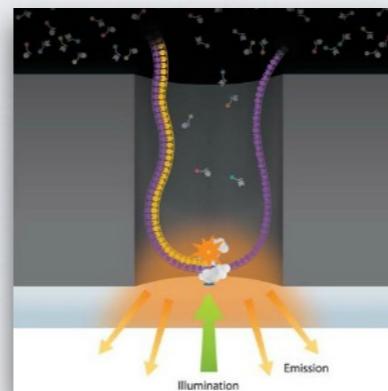


Long read technologies offer distinct advantages

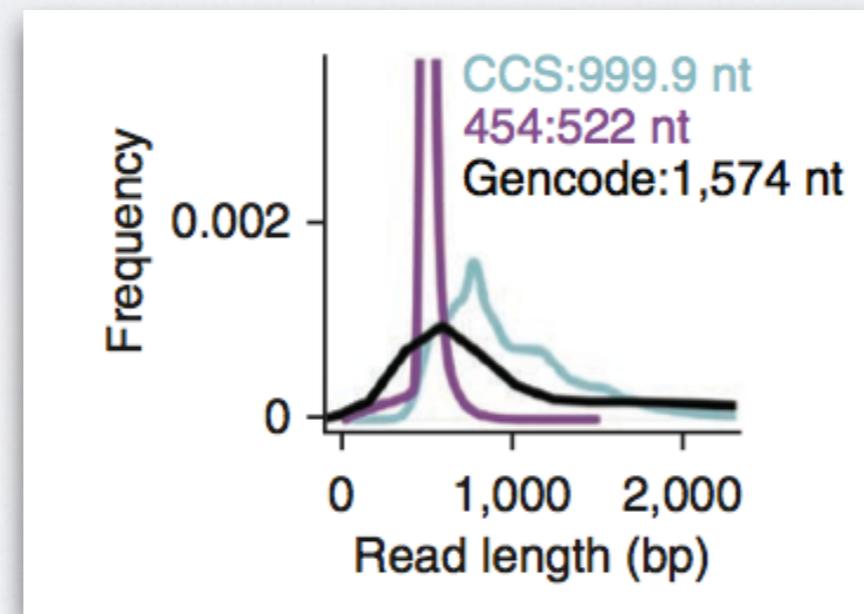
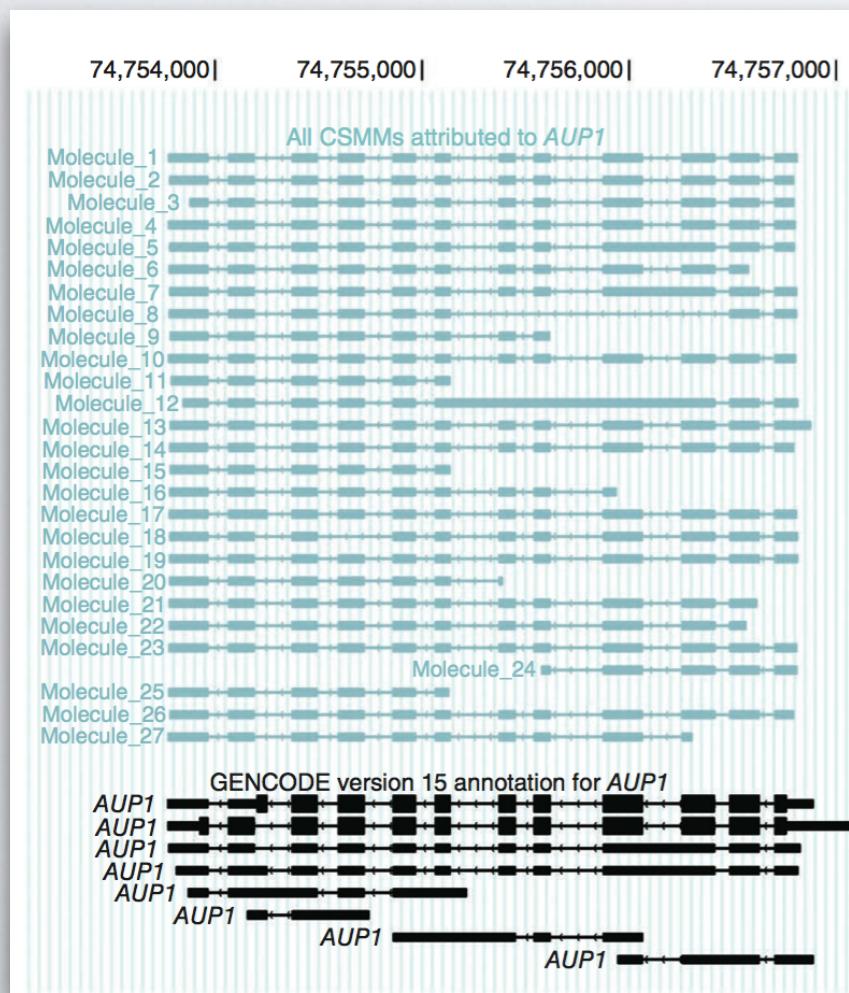
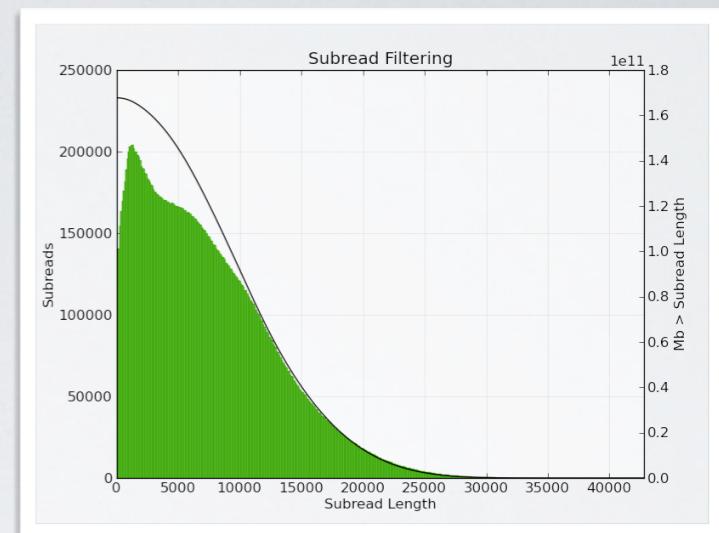
PacBio



Tens of thousands of reads per run
each read typically 5 - 20 kb long



single molecule real
time sequencing

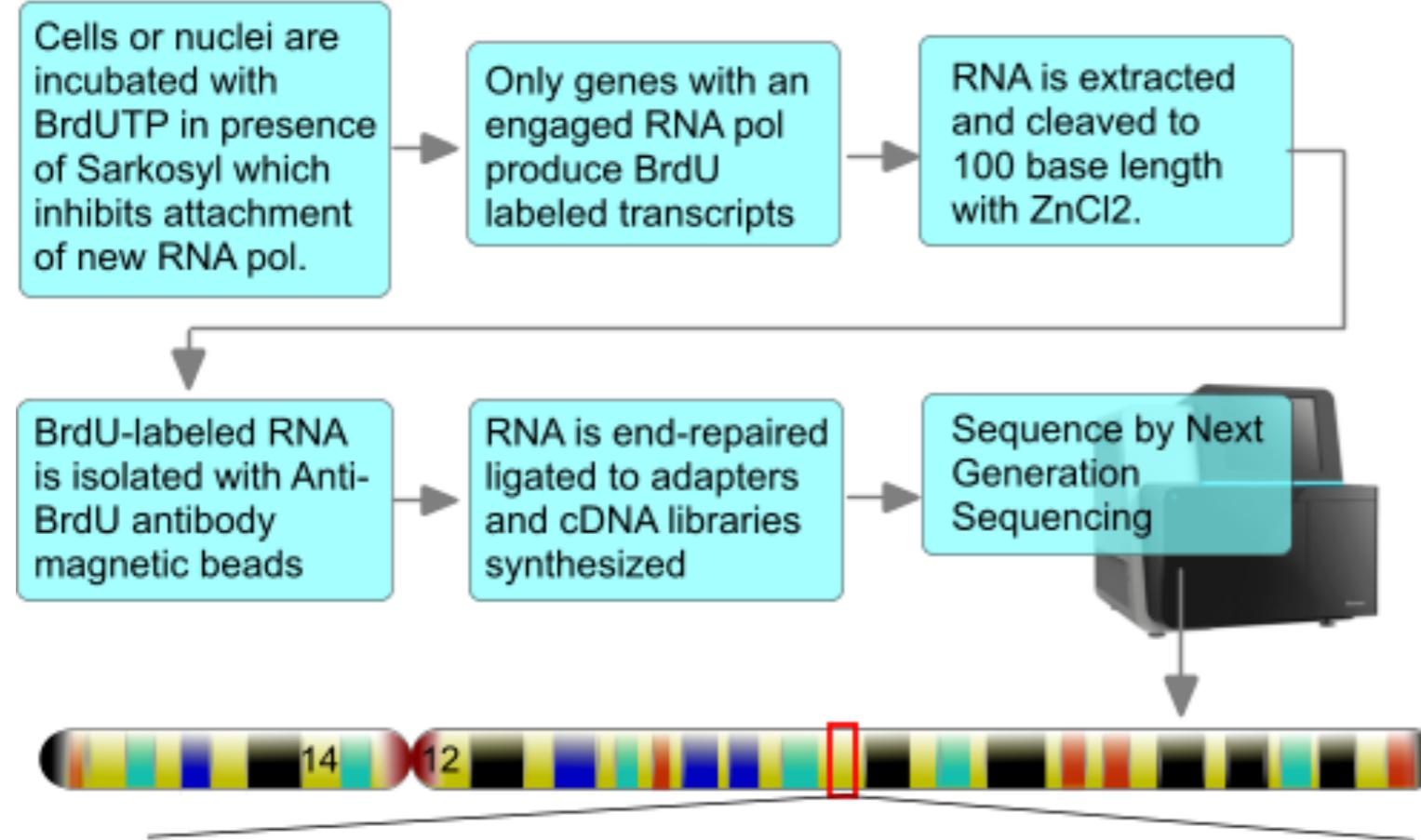


Long read NGS technologies allow sequencing of entire transcripts in a single read, so complex approaches to assemble short fragment reads of RNA are not necessary



Watching transcription: labeling and detecting nascent RNA

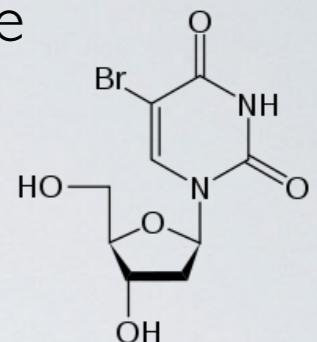
GRO-Seq (Gobal Run On assay)



Reads are mapped to the genome. Location of reads and number of reads indicate genes that have produced Run On transcripts across the whole genome

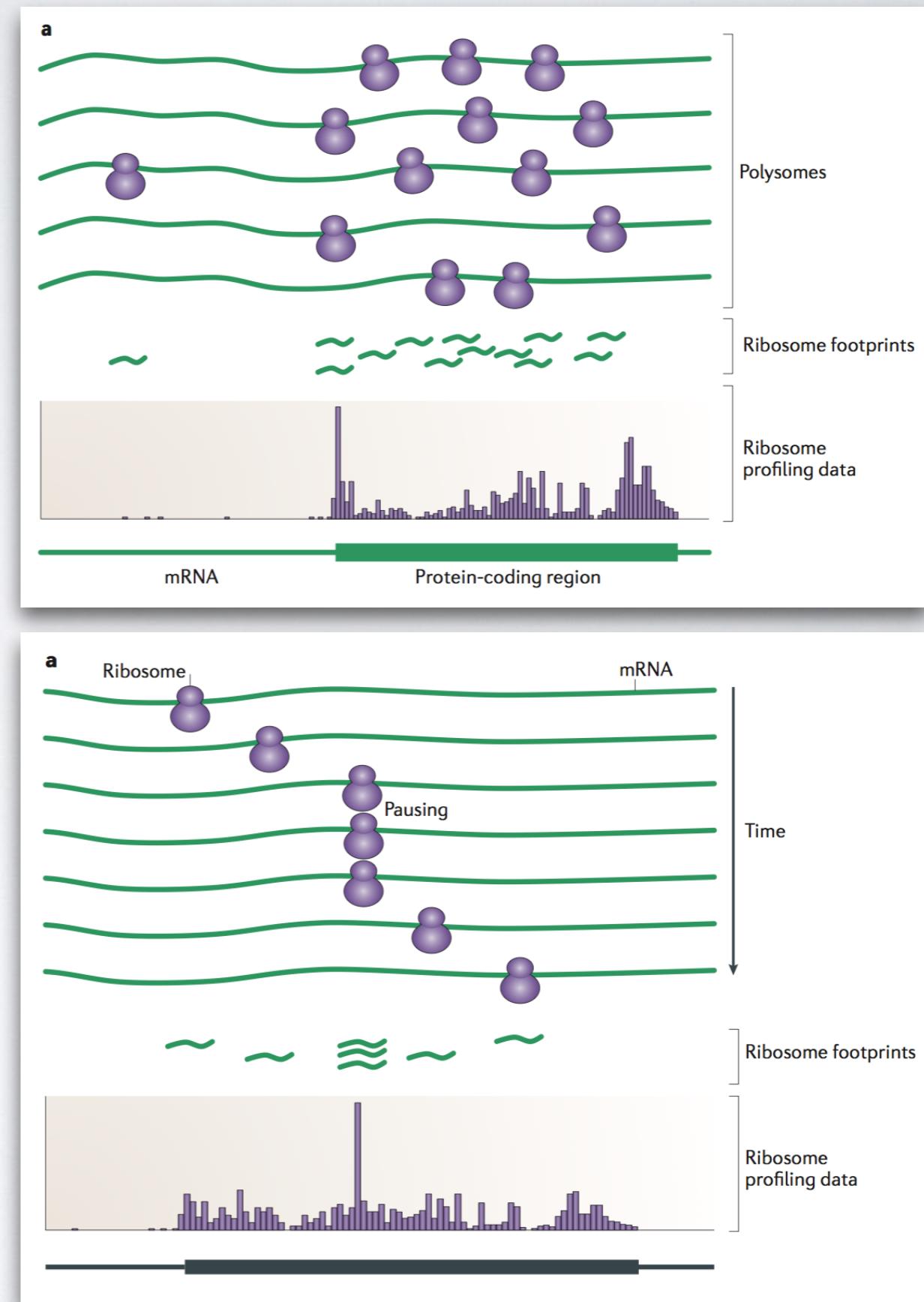
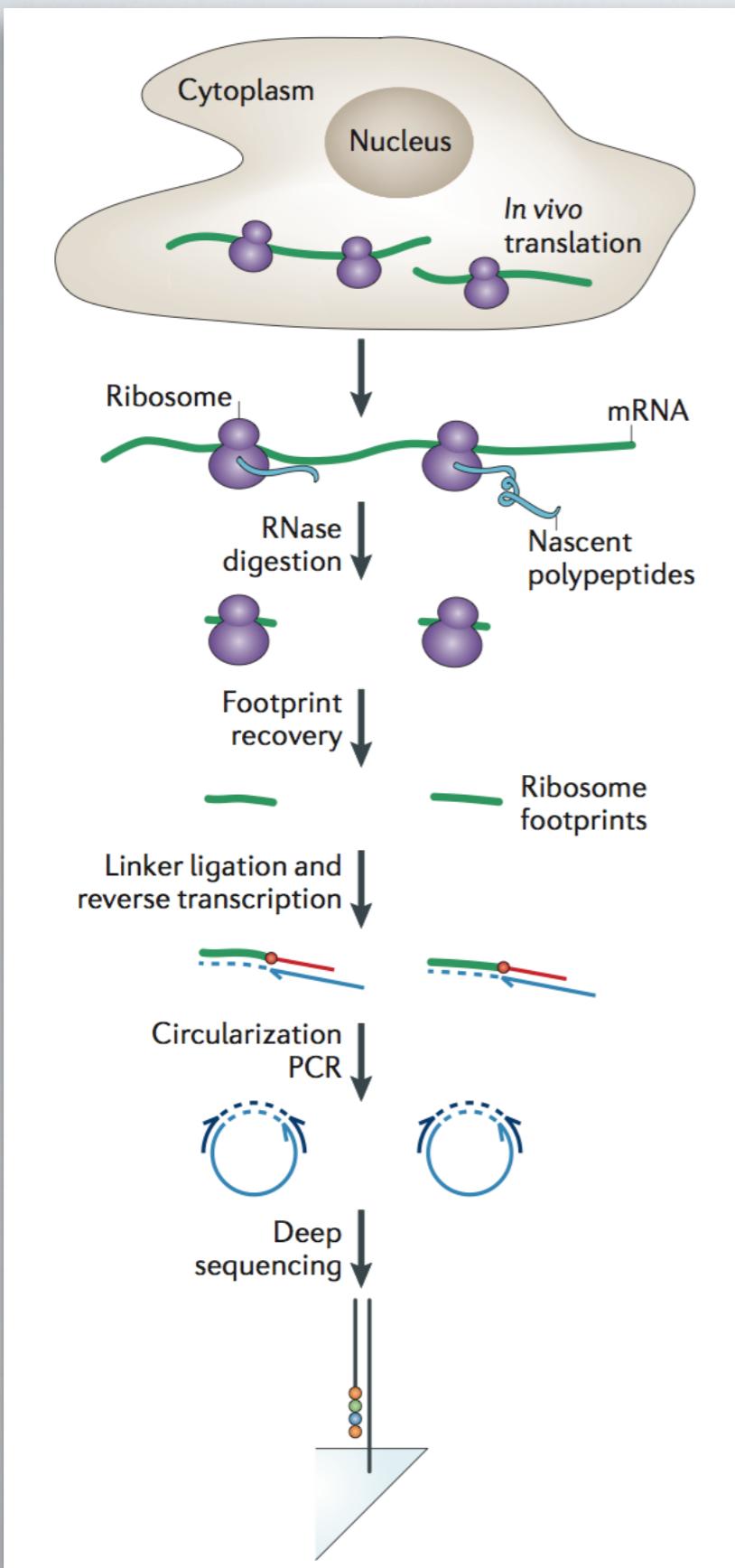
Bromodeoxyuridine (BrdU)

synthetic nucleoside analog of thymidine



incorporated into the newly synthesized RNA strand (in place of thymine)

Watching translation: ribosome footprint profiling



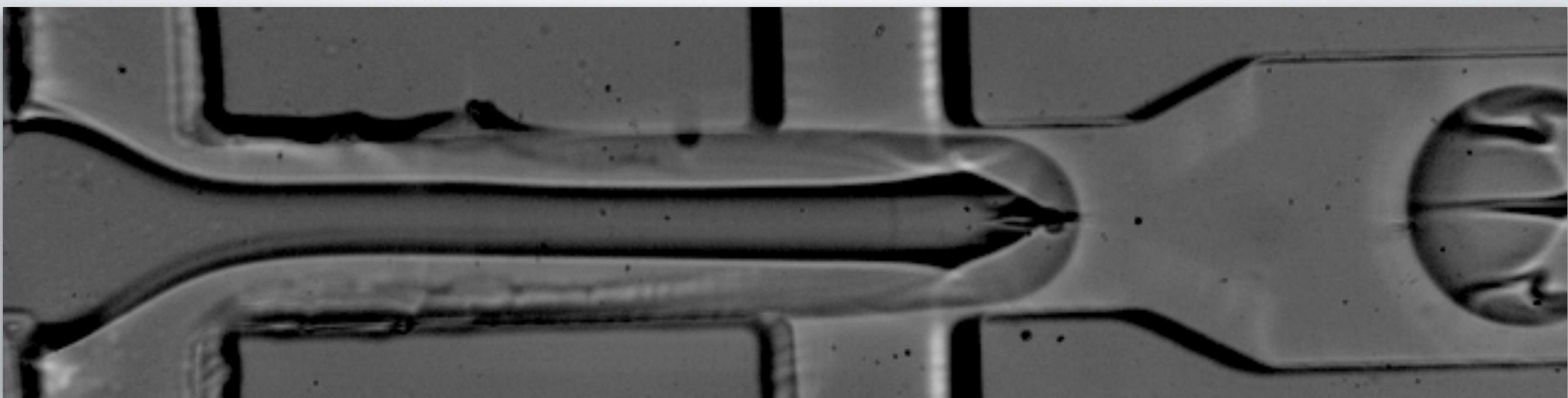
PROBLEM!!!!



In most analyses, cellular heterogeneity is a major confounding issue

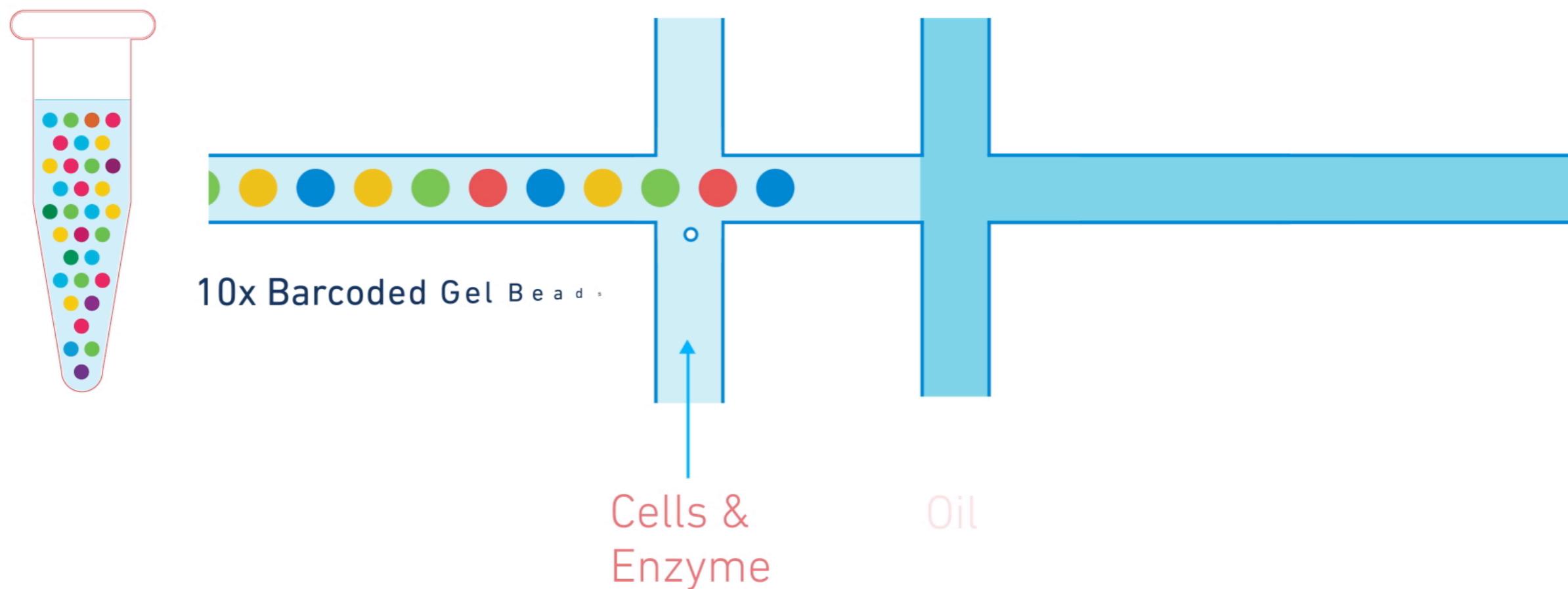
Ideally we would sequence the transcriptome of a single cell

Single cell RNA-seq



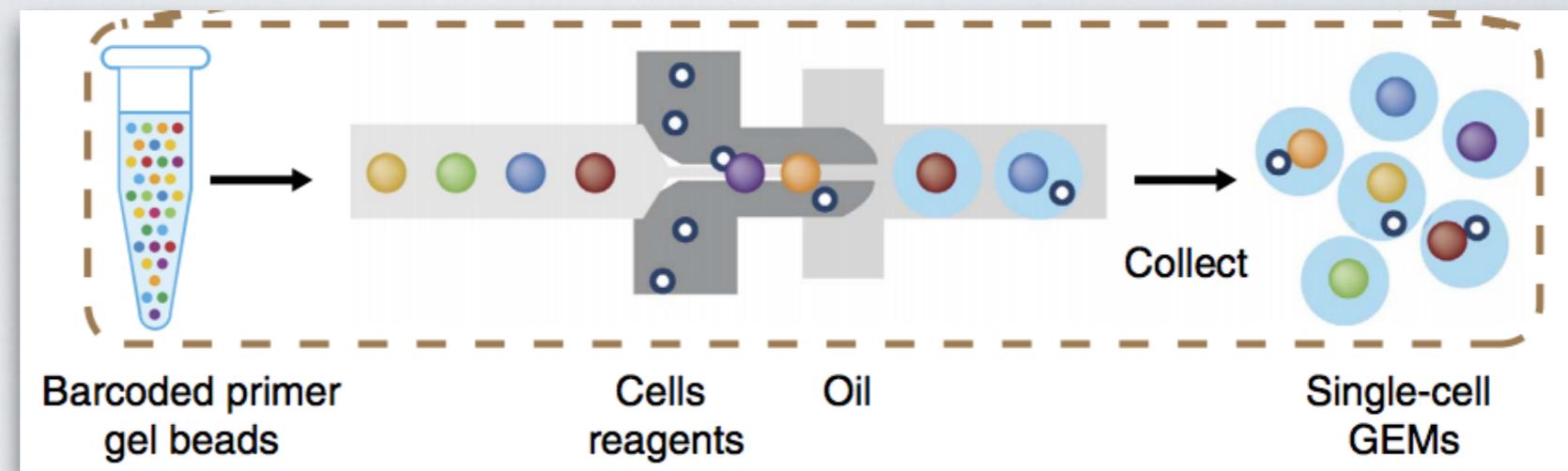
Single cell RNA-seq

10x GemCode™ Technology for Single Cell Partitioning

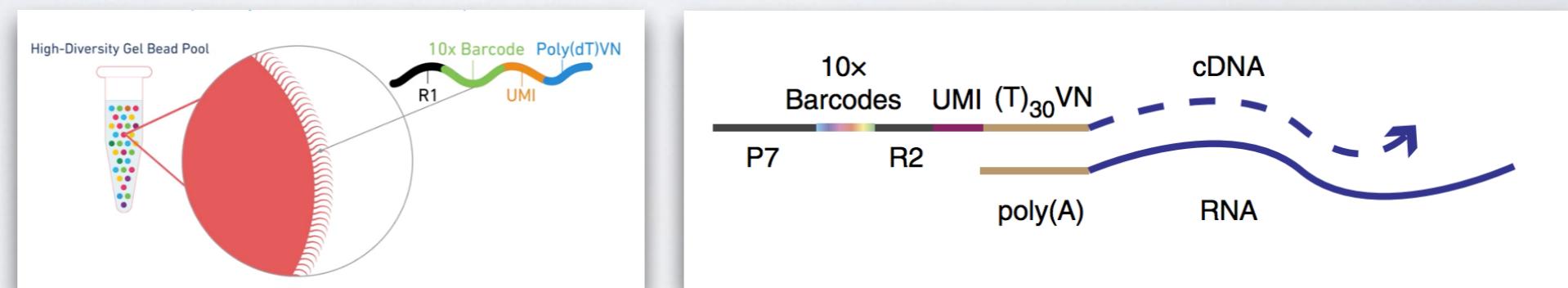


Single cell RNA-seq

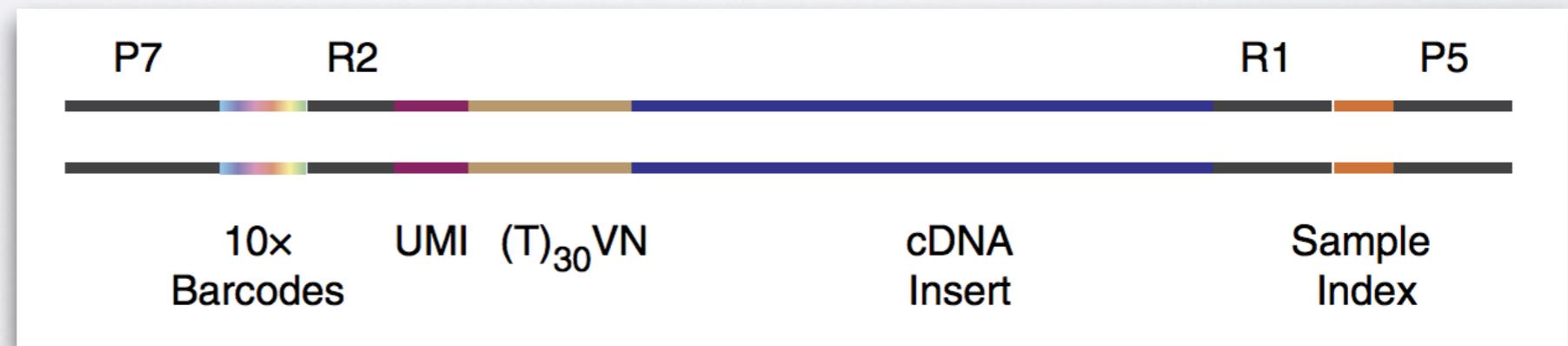
Gel beads loaded with primers and barcoded oligonucleotides are first mixed with cells and reagents, and subsequently mixed with oil-surfactant solution at a microfluidic junction to generate a single cell GEM



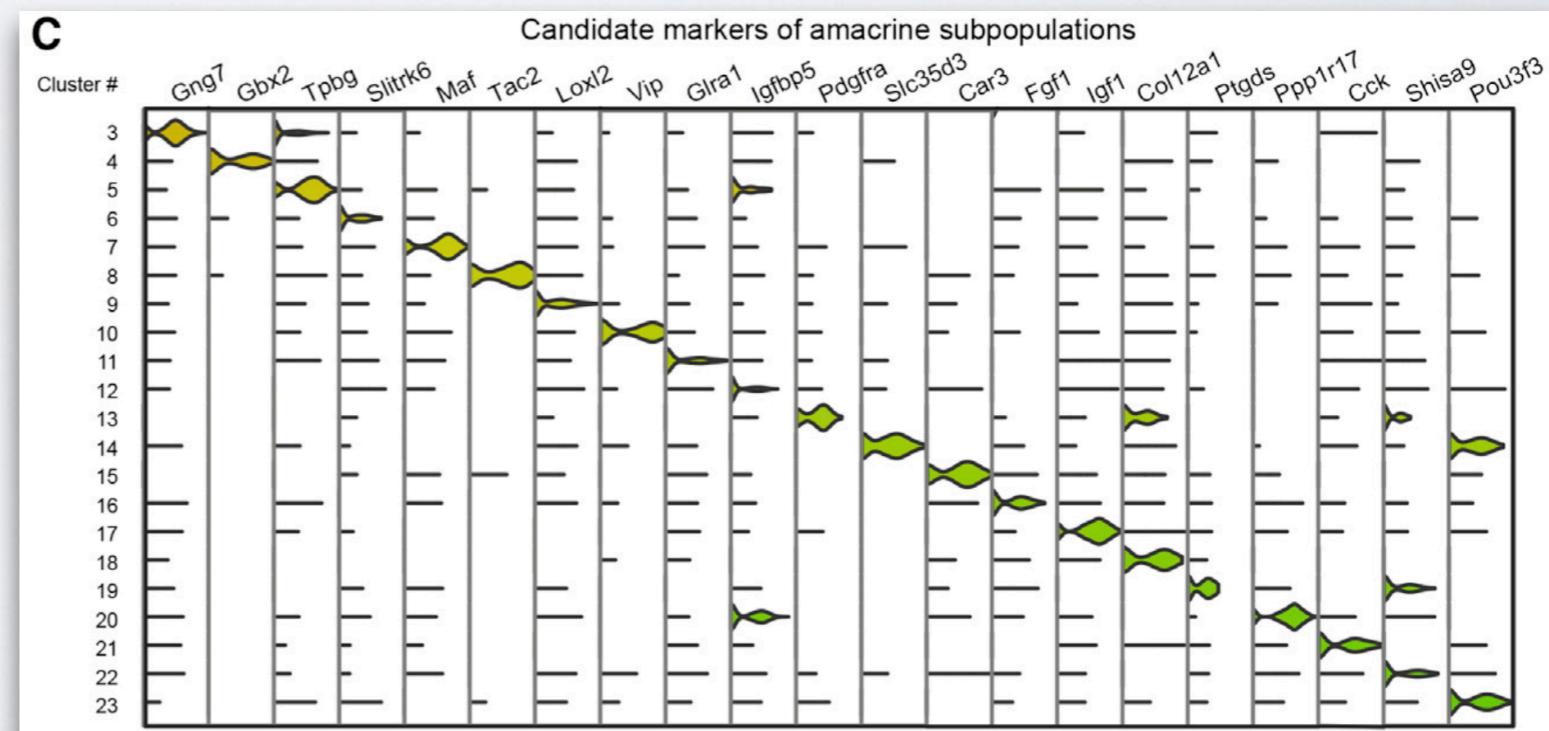
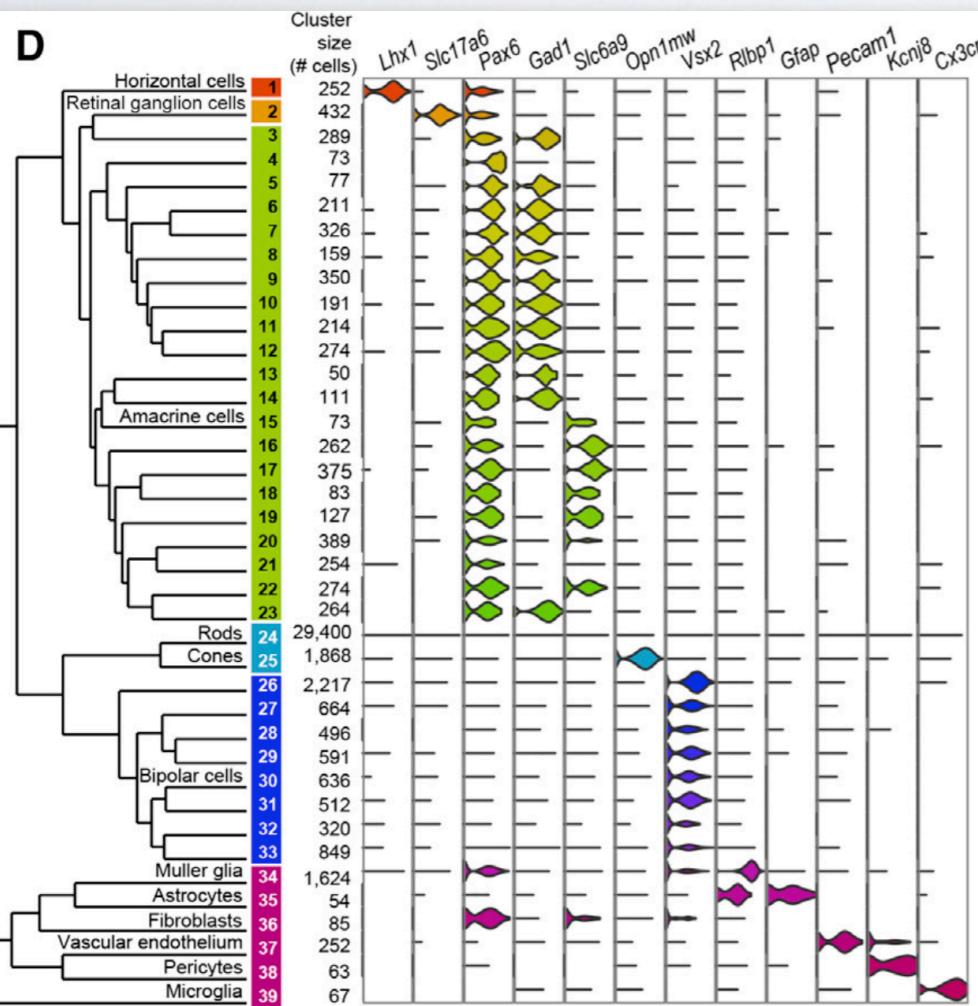
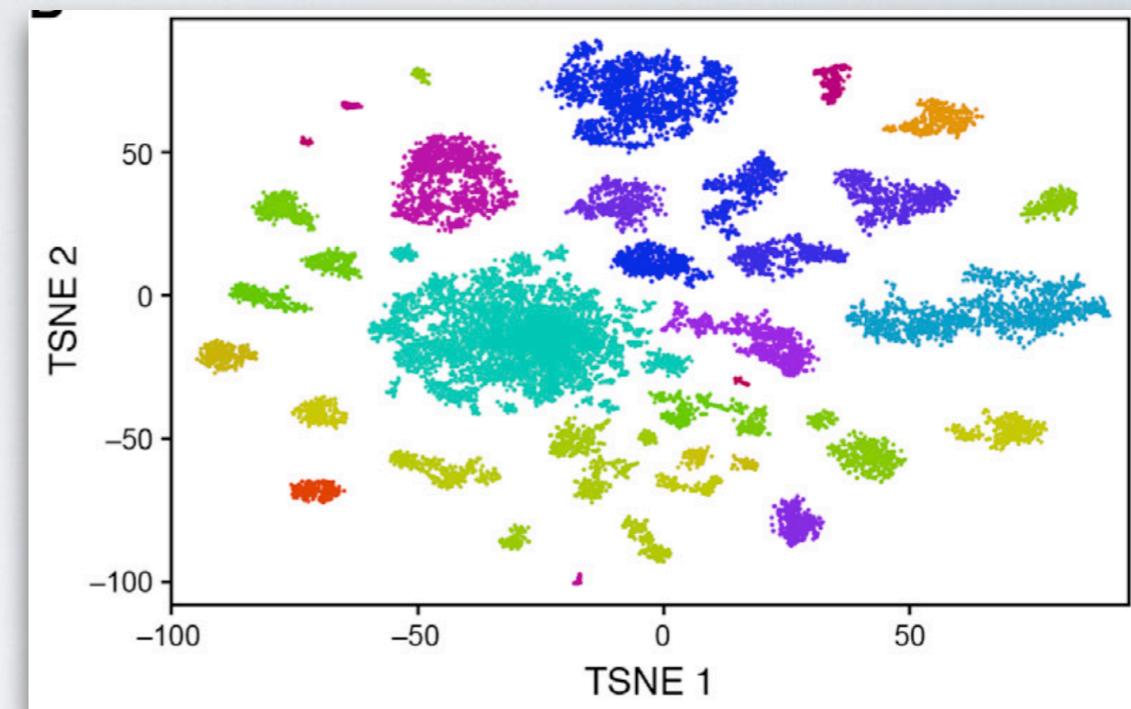
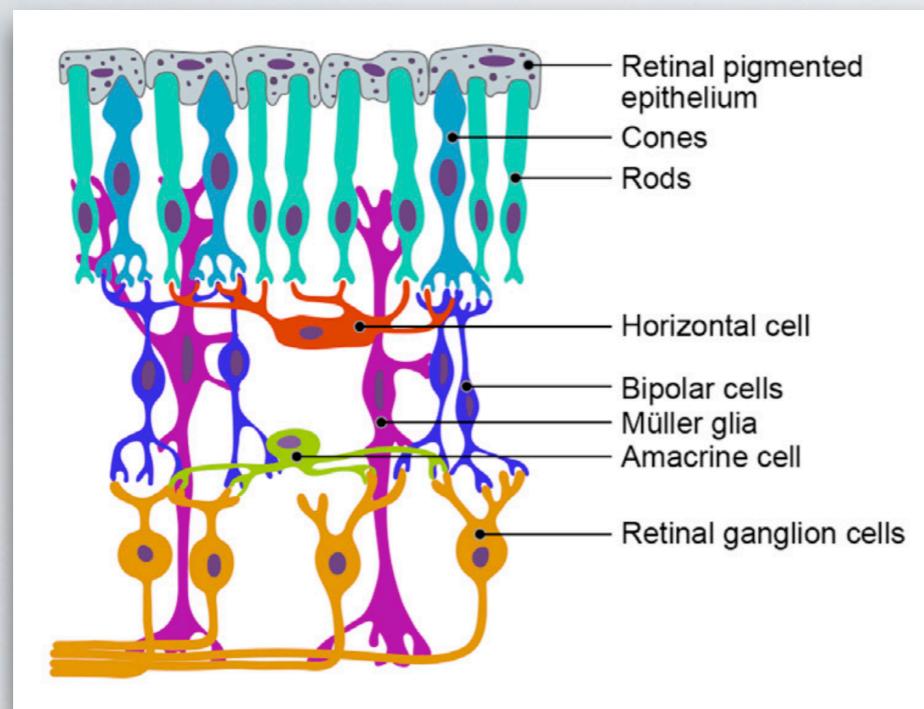
Gel beads contain barcoded oligonucleotides consisting of Illumina adapters, 10x barcodes, UMIs and oligo dTs, which prime RT of polyadenylated RNAs.



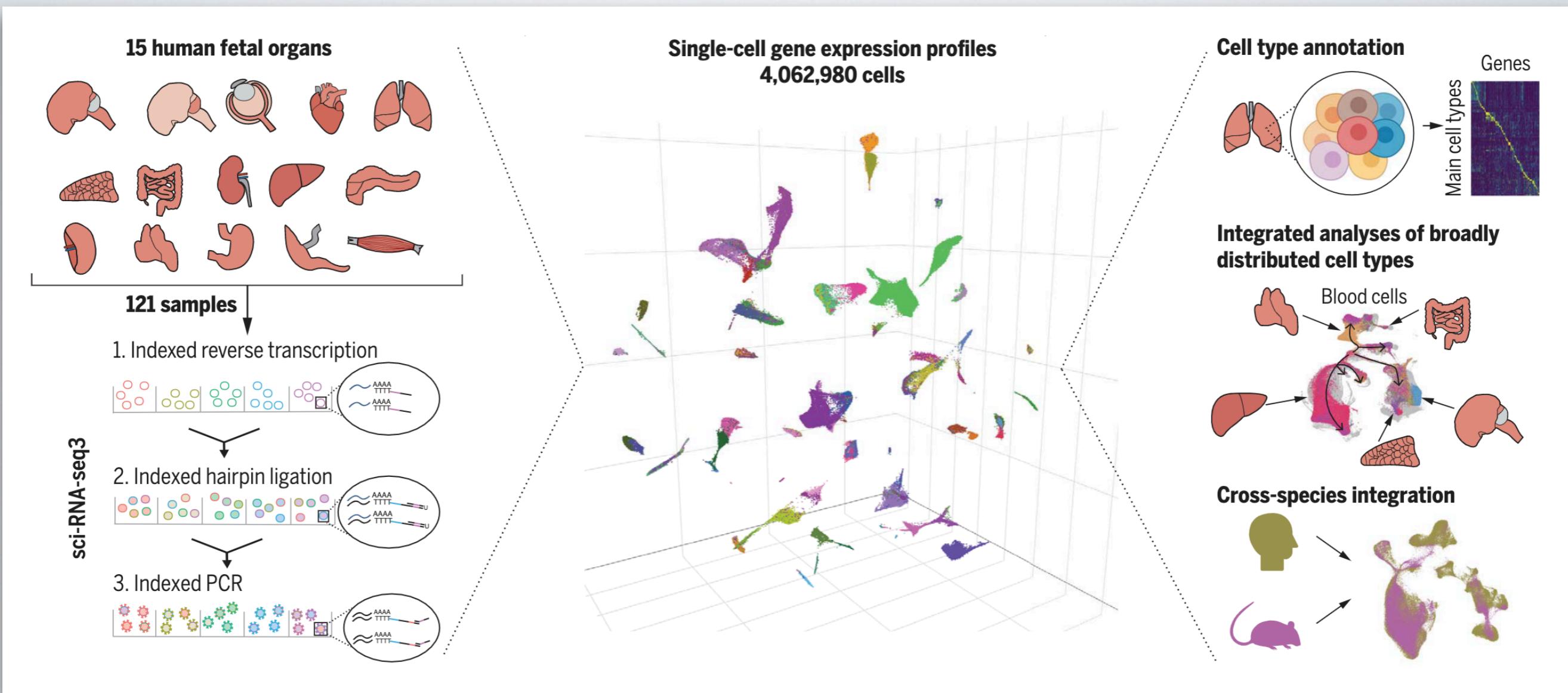
Finished library molecules consist of Illumina adapters and sample indices, allowing pooling and sequencing of multiple libraries on a next-generation short read sequencer.



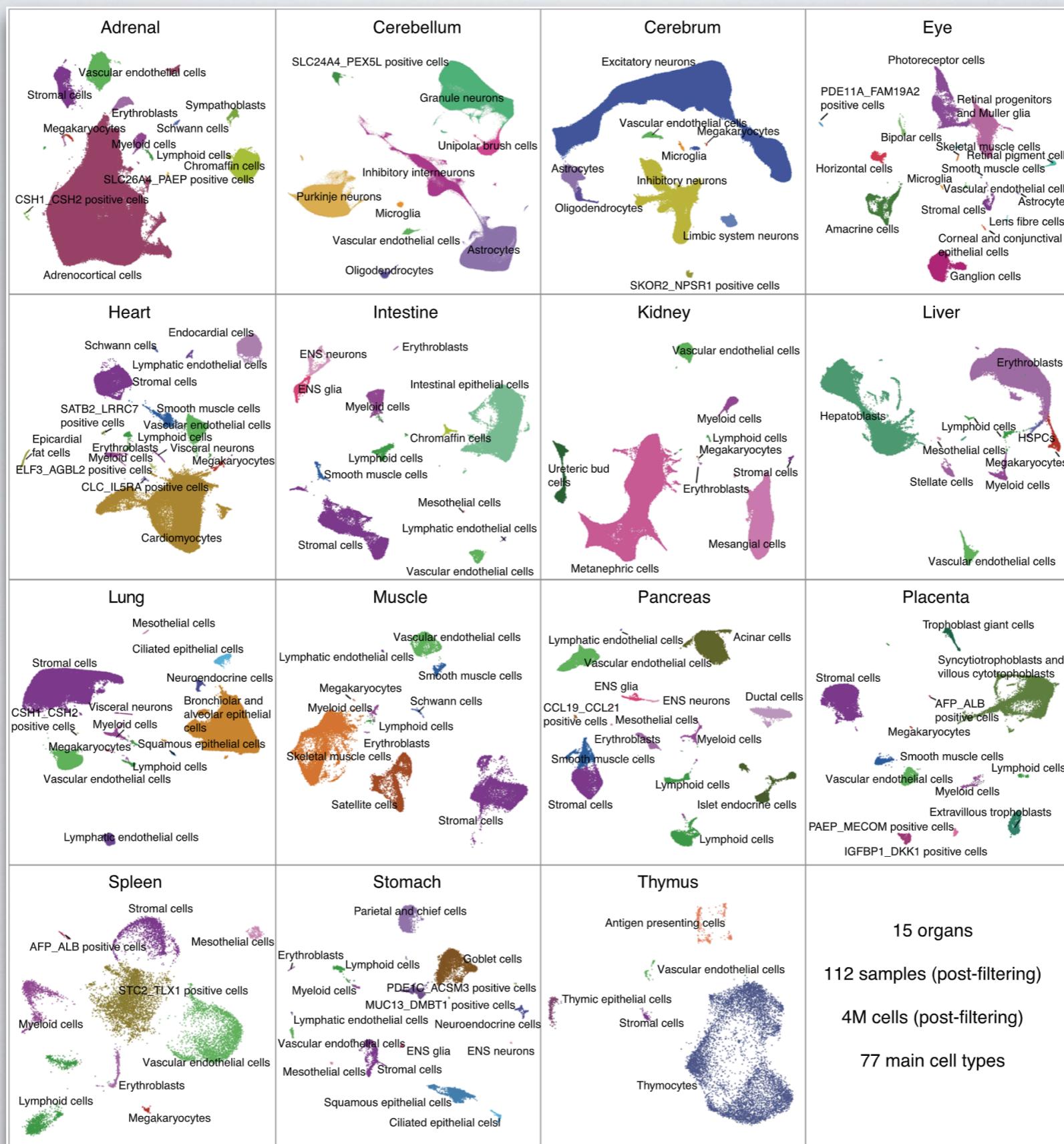
Single cell RNA-seq



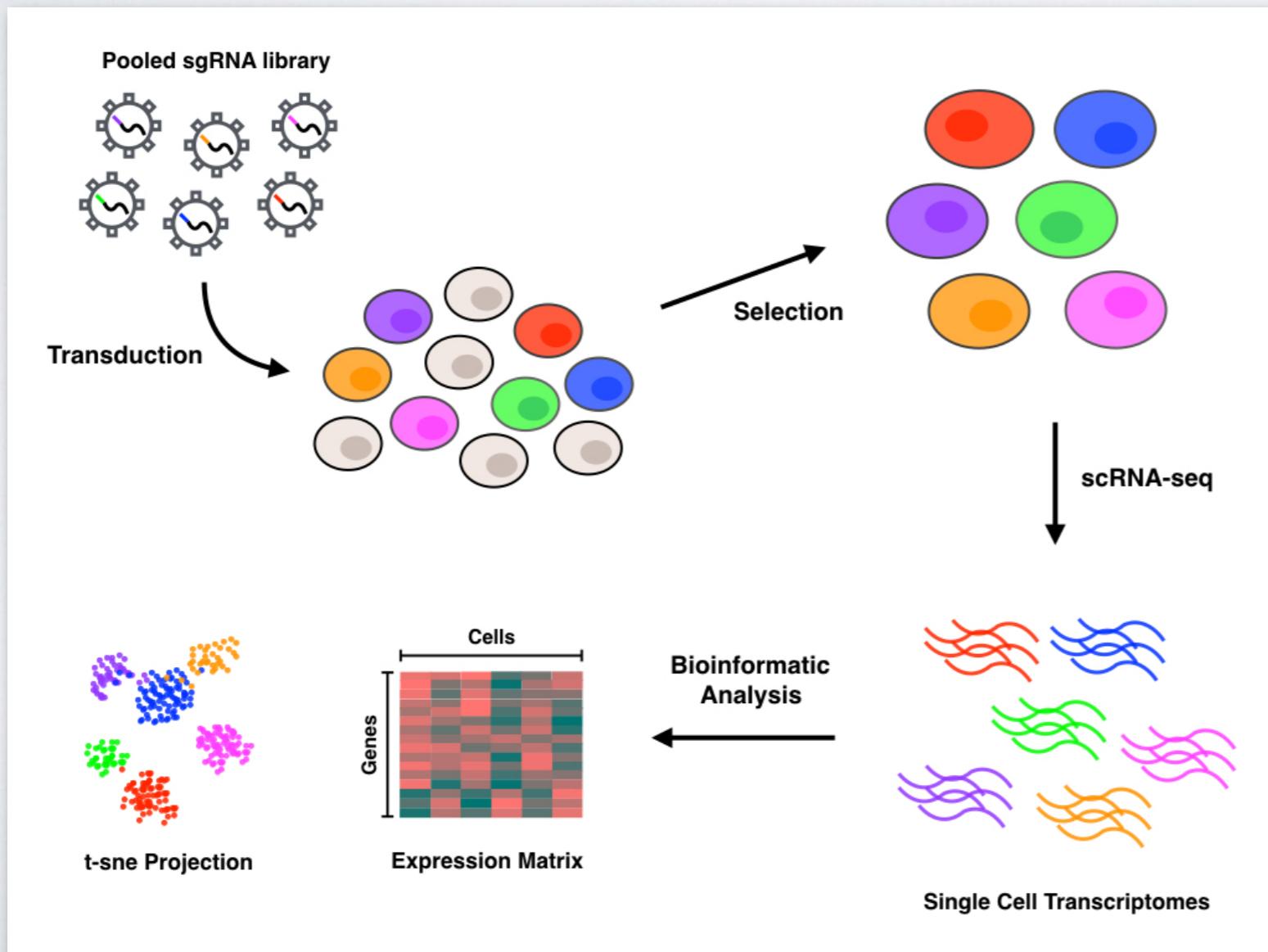
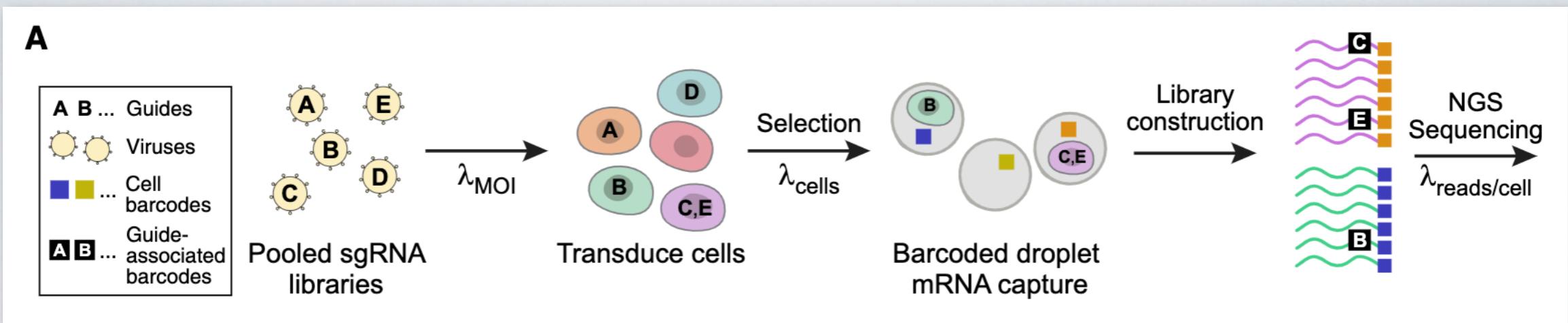
Single cell RNA-seq at increasing scale



Single cell RNA-seq at increasing scale



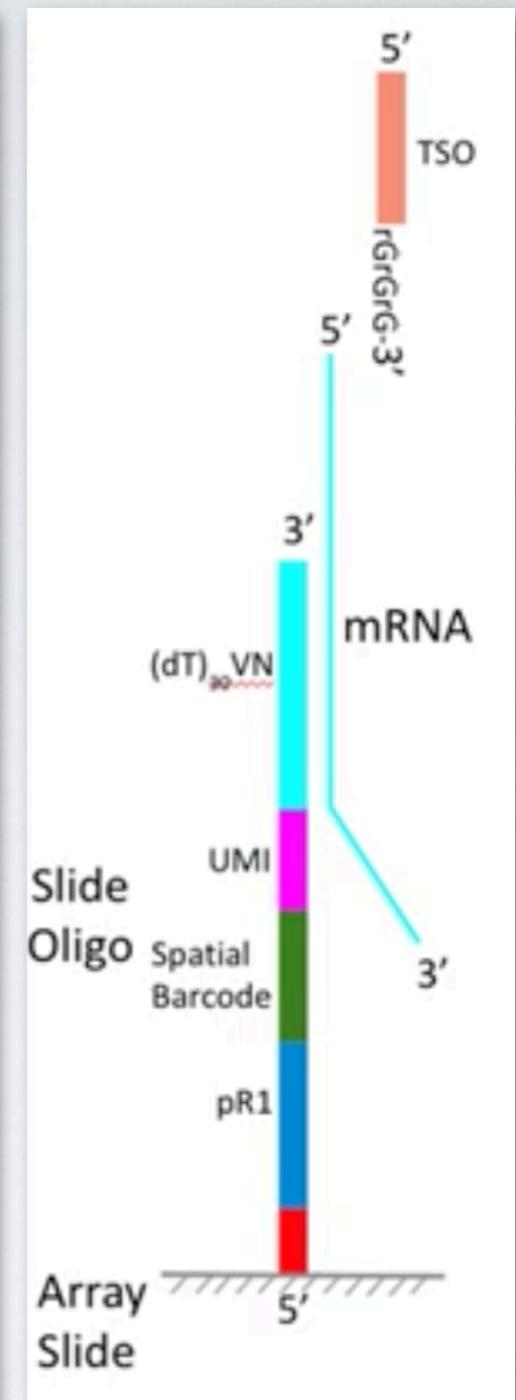
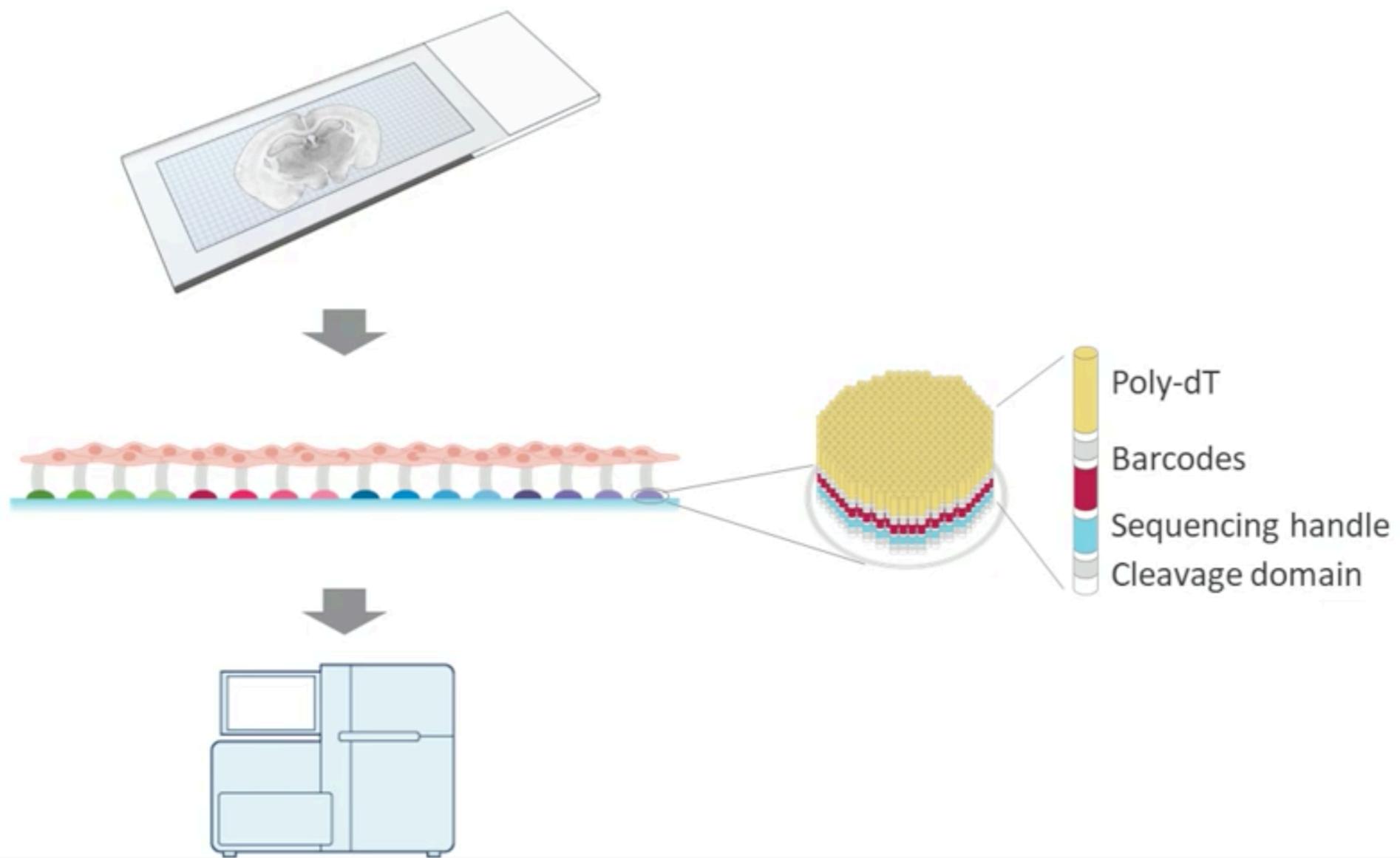
Perturb-seq



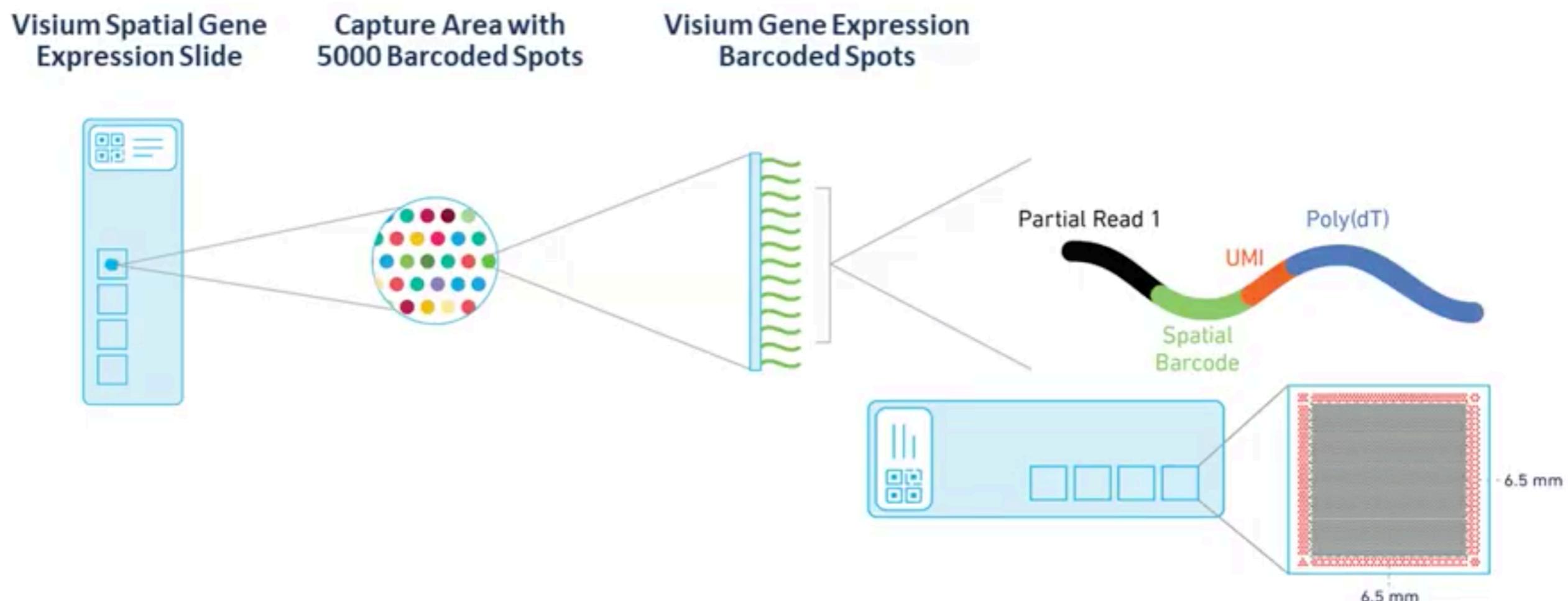
Dixit et al. (2016) Perturb-Seq:
Dissecting Molecular Circuits with
Scalable Single-Cell RNA Profiling of
Pooled Genetic
Screens. *Cell* 167, 0.1016/
[j.cell.2016.11.038](https://doi.org/10.1016/j.cell.2016.11.038)

Spatially resolved transcriptomics

Spatial Transcriptomics *The basics*



Spatially resolved transcriptomics

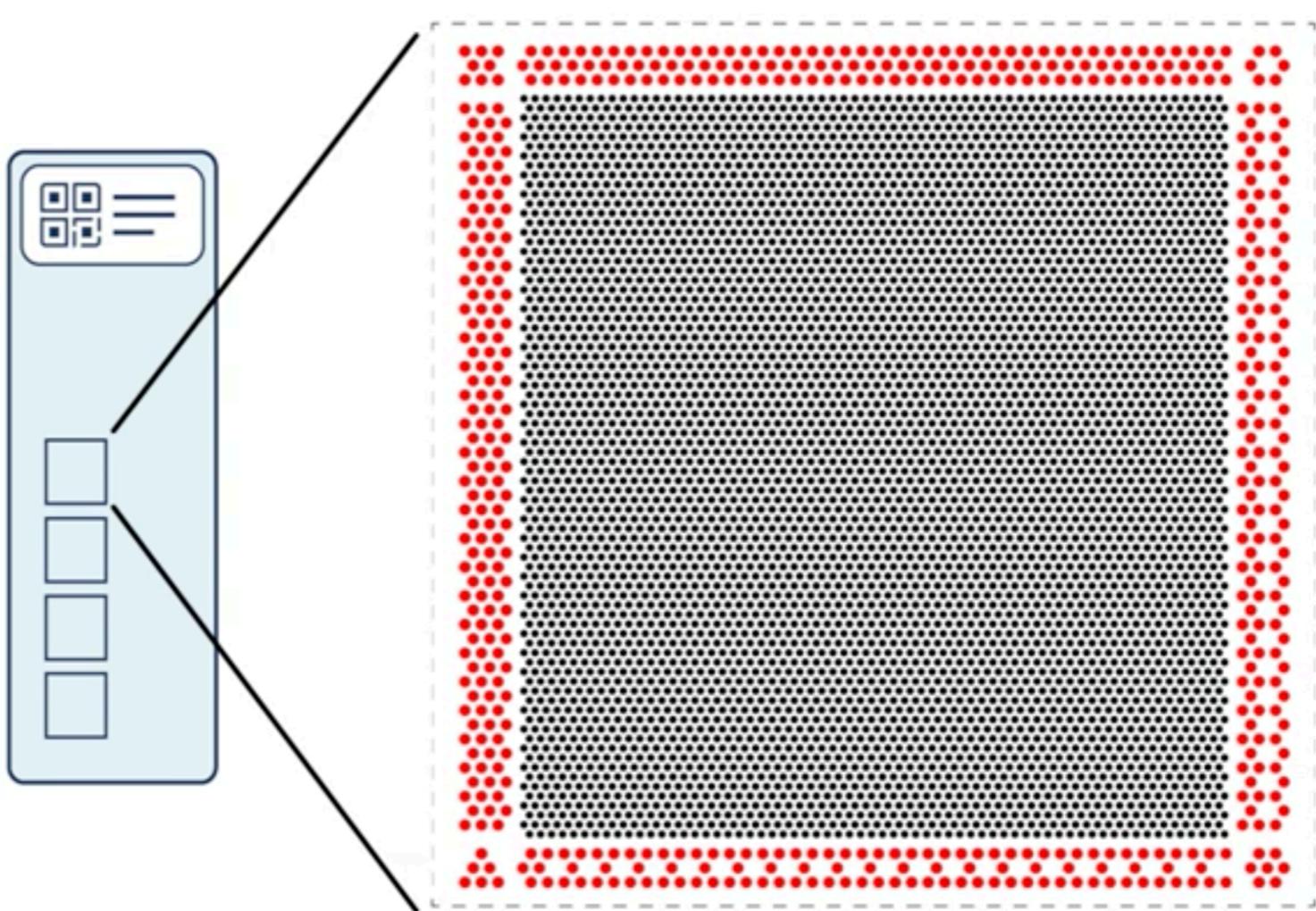


Spatially resolved transcriptomics

The Slides

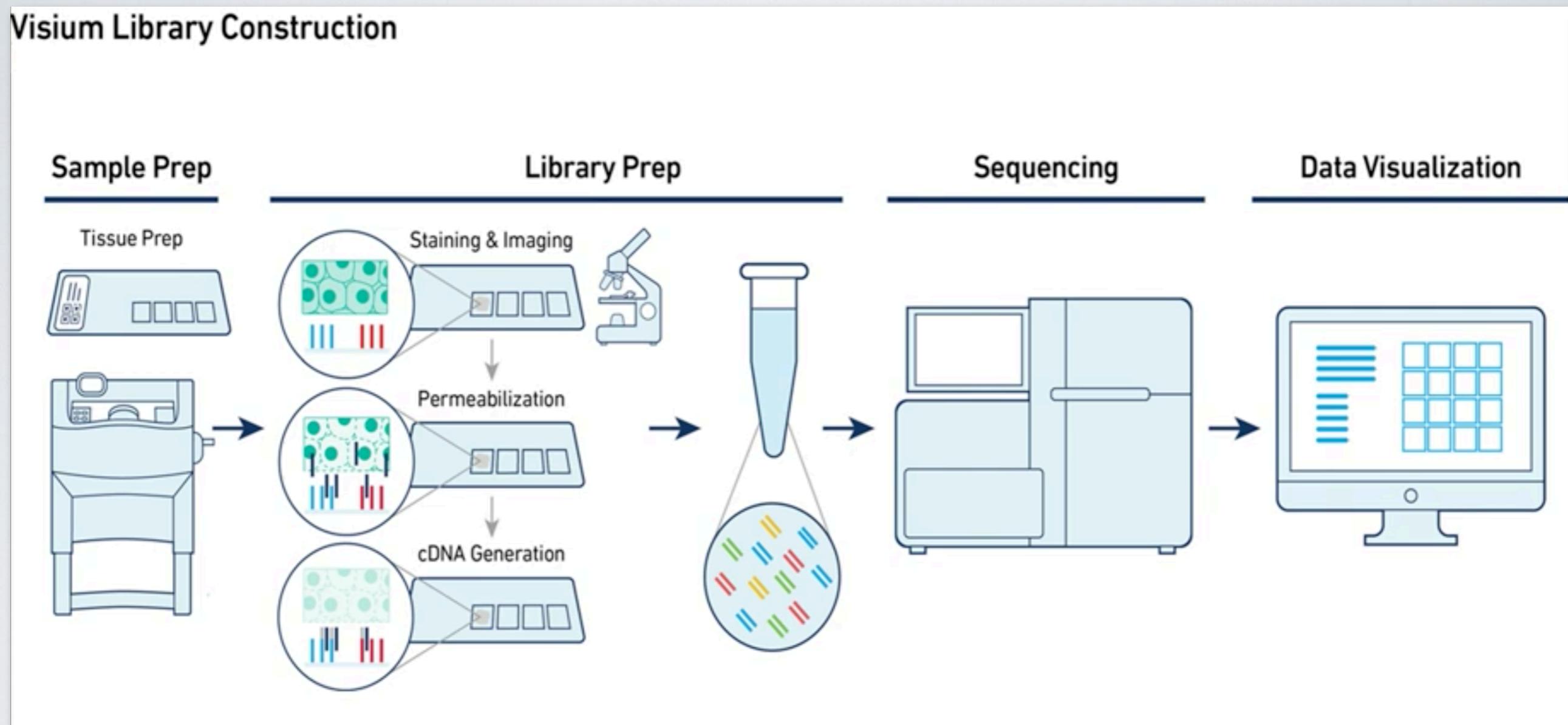
Gene Expression

- Gene Expression Slides
 - Standard 25mm x 75mm glass slide
 - 4 capture areas per slide
 - Standard spacing allows multichannel pipetting
- The Barcodes (6.5mm x 6.5mm)
 - 4992 unique spatially encoded barcodes
 - 55um spot size
 - 100um center to center
 - Dense "orange crate" packing
- The Frame (8mm x 8mm)
 - 100um spots, 200um center to center
 - Visible in brightfield and fluorescent
 - Allows for easy automated or manual alignment of tissues

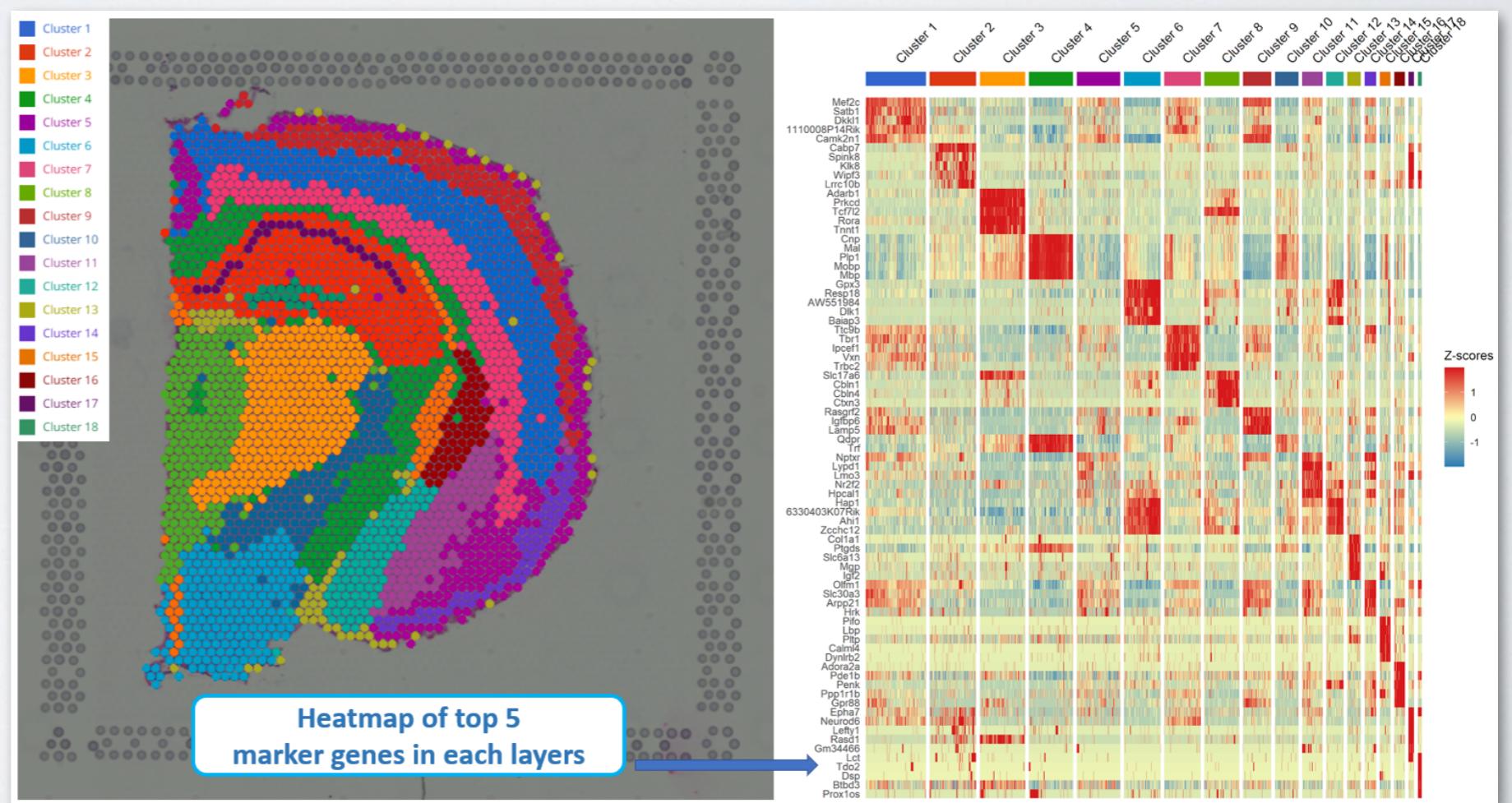
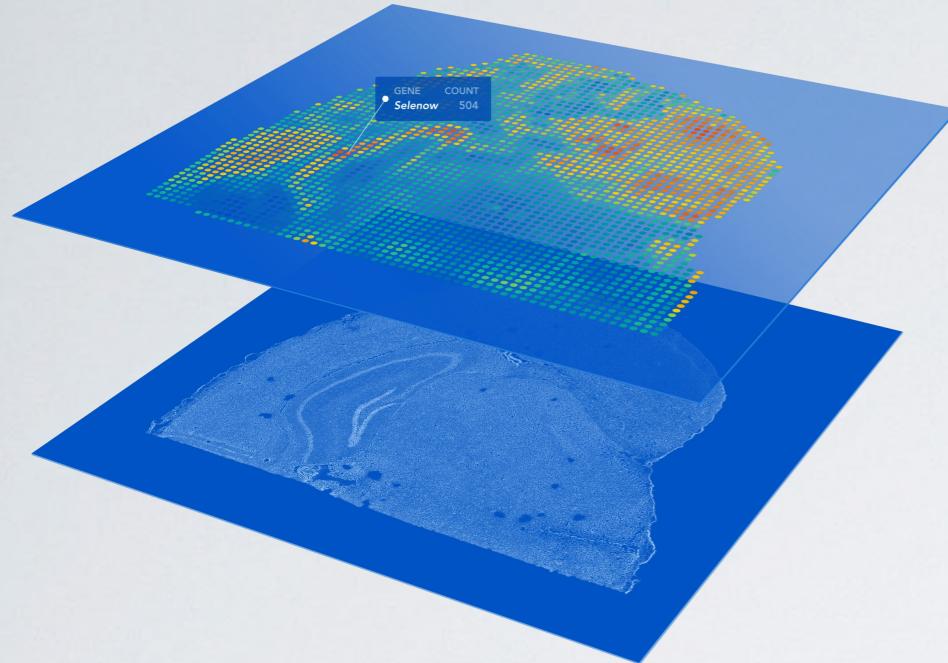


Spatially resolved transcriptomics

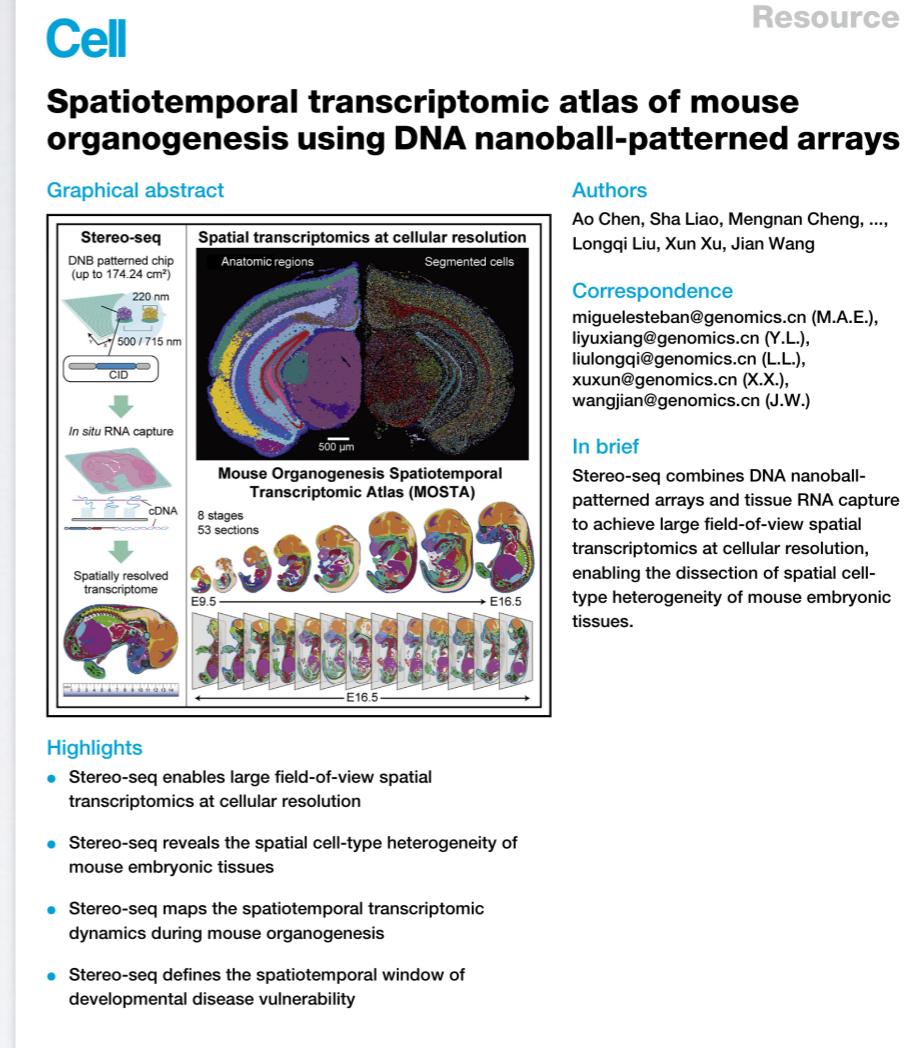
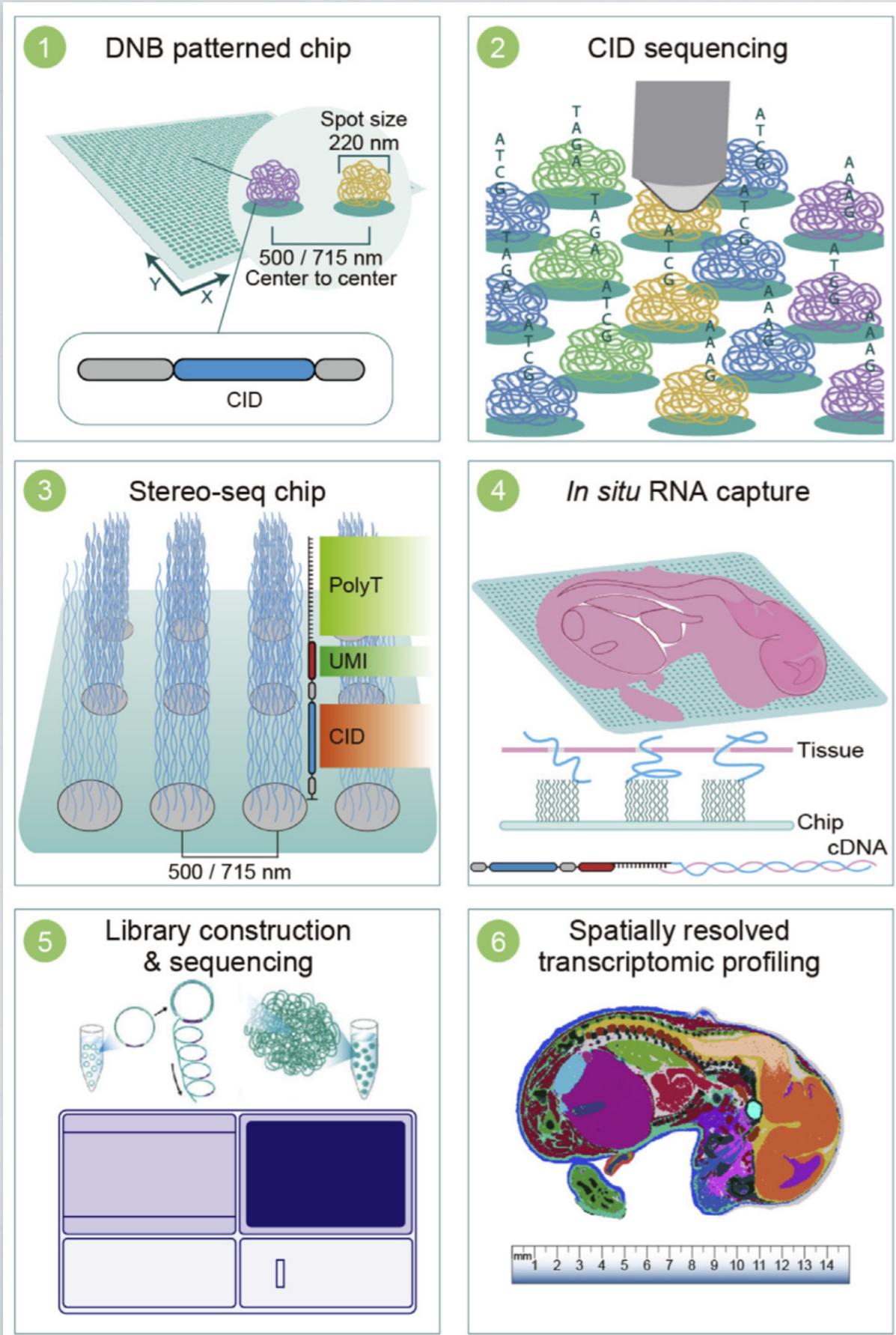
Visium Library Construction



Spatially resolved transcriptomics



Spatially resolved transcriptomics



Chen et al. (2022) Spatiotemporal transcriptomic atlas of mouse organogenesis using DNA nanoball-patterned arrays.
Cell 185, 1777–1792, 10.1016/j.cell.2022.04.003

DNB = DNA NanoBall

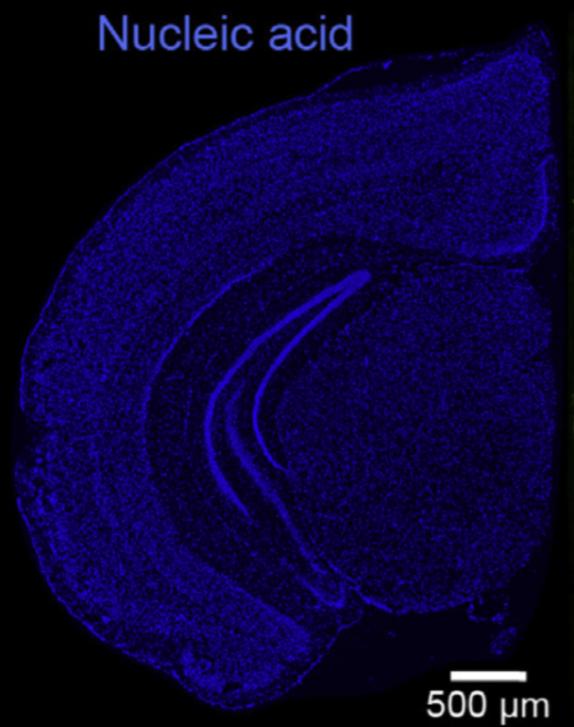
CID = Coordinate IDentity

PolyT and UMI are ligated onto the oligonucleotide containing the CID barcode

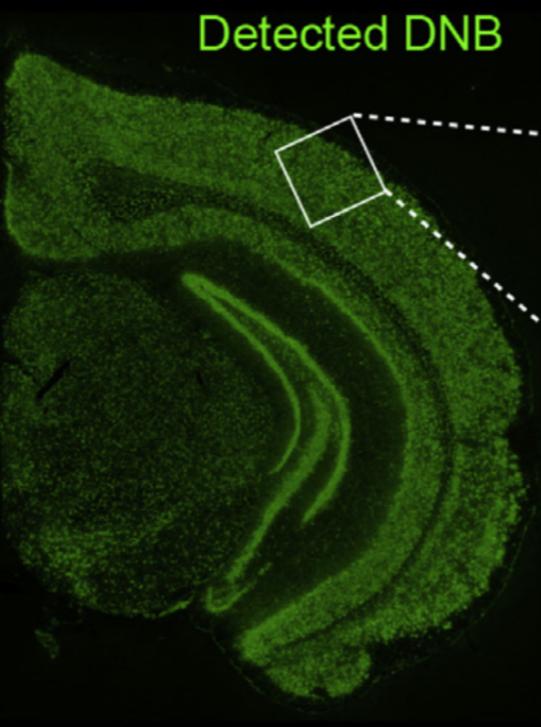
Spatially resolved transcriptomics

A

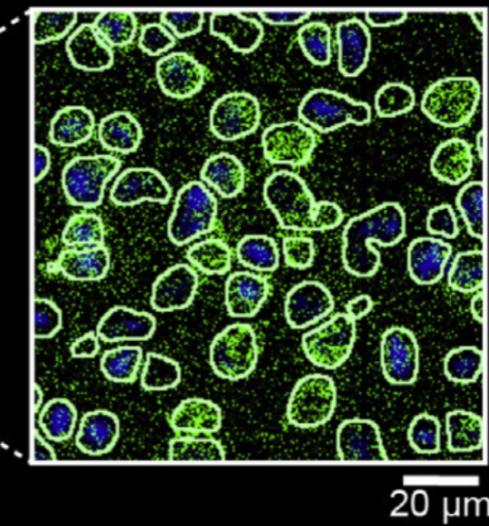
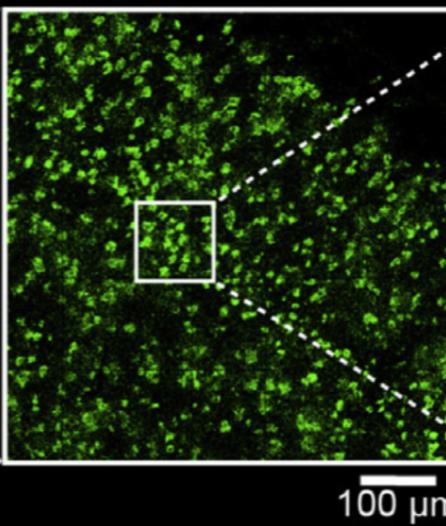
Mouse brain



Detected DNB

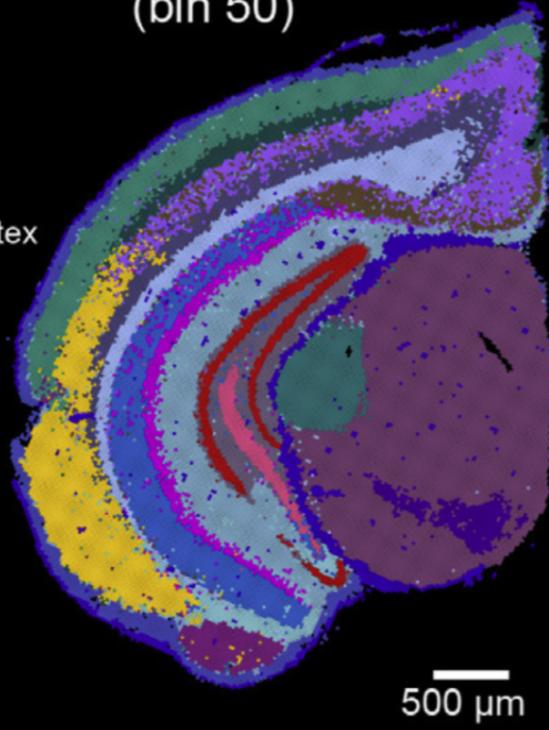
**B**

Detected DNB / Nucleic acid

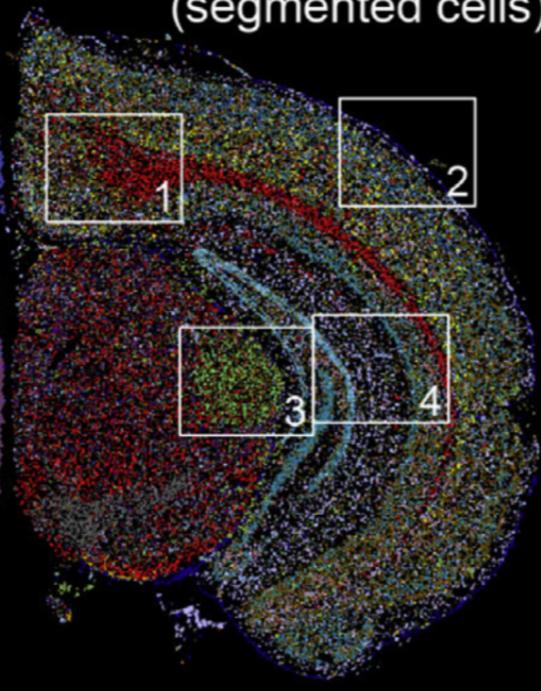
**C**

Anatomic regions
(bin 50)

- Meninges
- Cortex L1
- Cortex L2/3
- Cortex L4
- Cortex L5
- Cortex L6
- Lat-ven cortex
- CAA
- PAN
- FT
- Subiculum
- SO CA1
- CA1
- SL/R CA1
- MLDG
- DG
- CA3
- Thalamus
- Mb
- SN/VTA



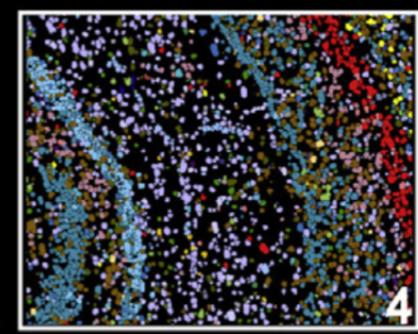
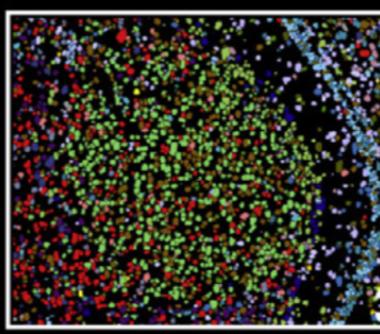
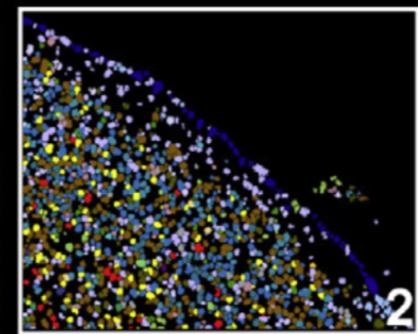
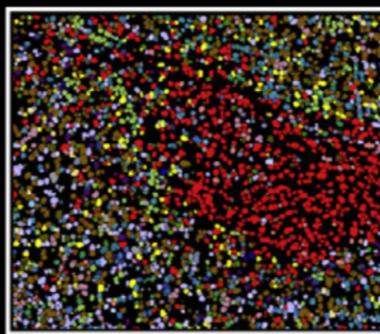
Cell clusters
(segmented cells)



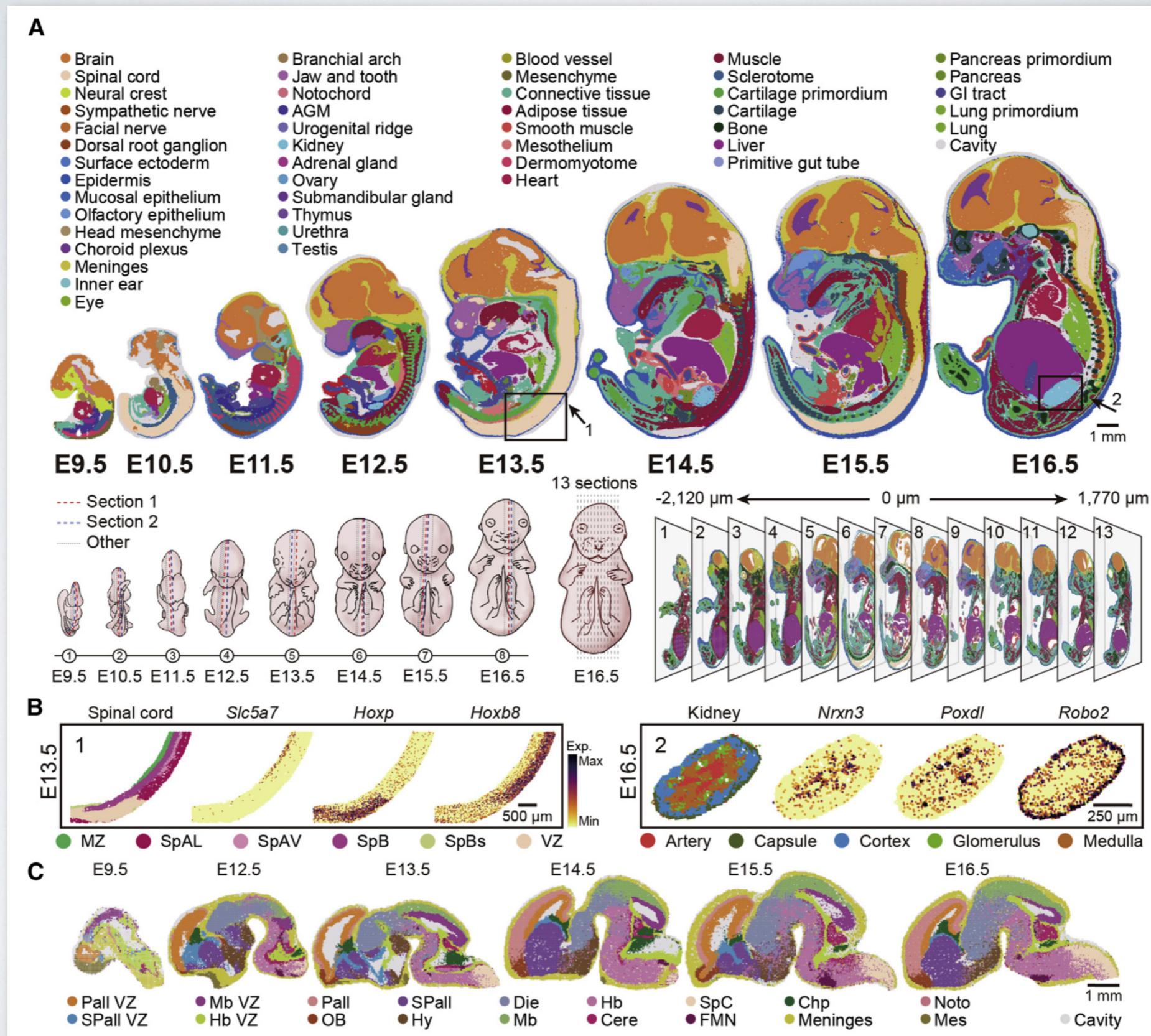
- EX
- EX L2/3
- EX L4
- EX L5/6
- EX L6
- EX Mb
- EX thalamus
- IN thalamus
- DA neuron
- EX CA
- GN DG
- IN Sst⁺
- IN Pvalb⁺Reln⁺
- IN Vip⁺
- IN Pvalb⁺Reln⁻
- Astr1
- Astr2
- Astr3
- Astr4
- Astr5
- Micro
- OPC
- Oligo
- SMC
- Ery
- Meninges
- Endo

D

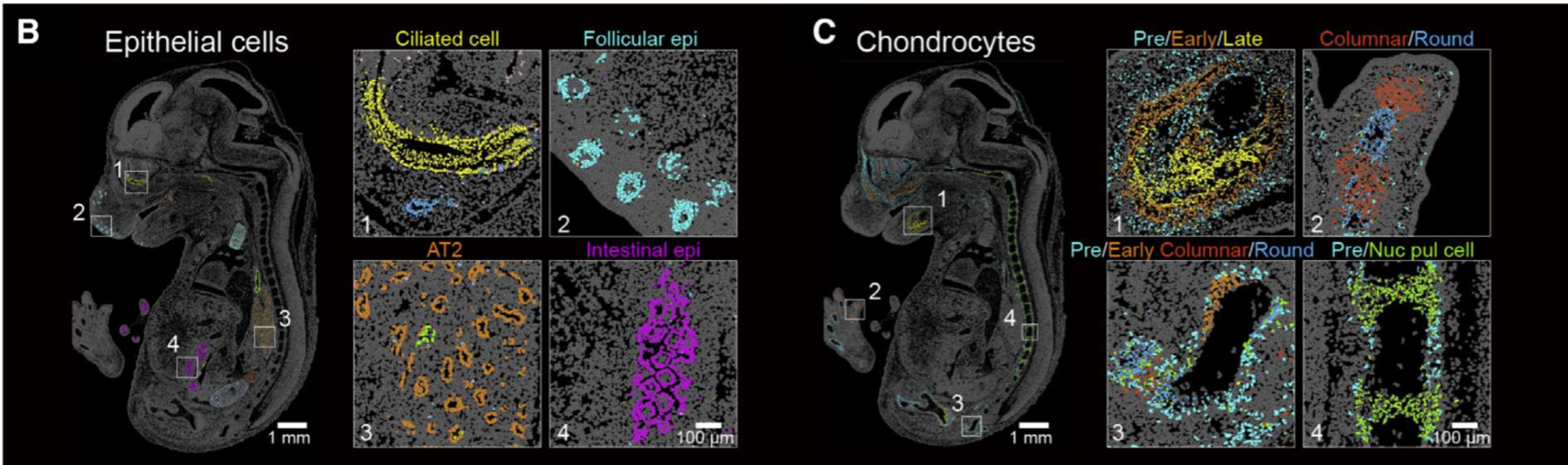
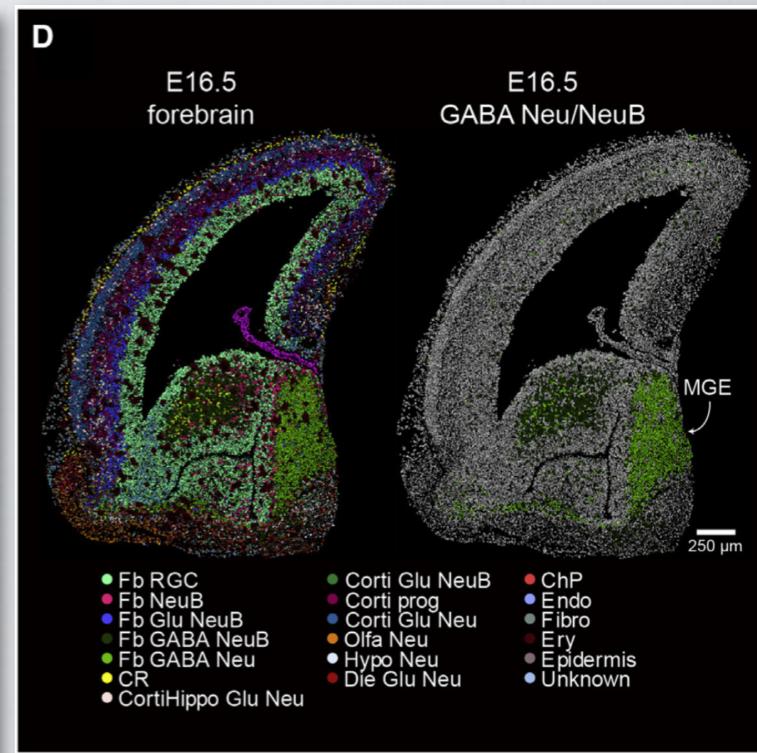
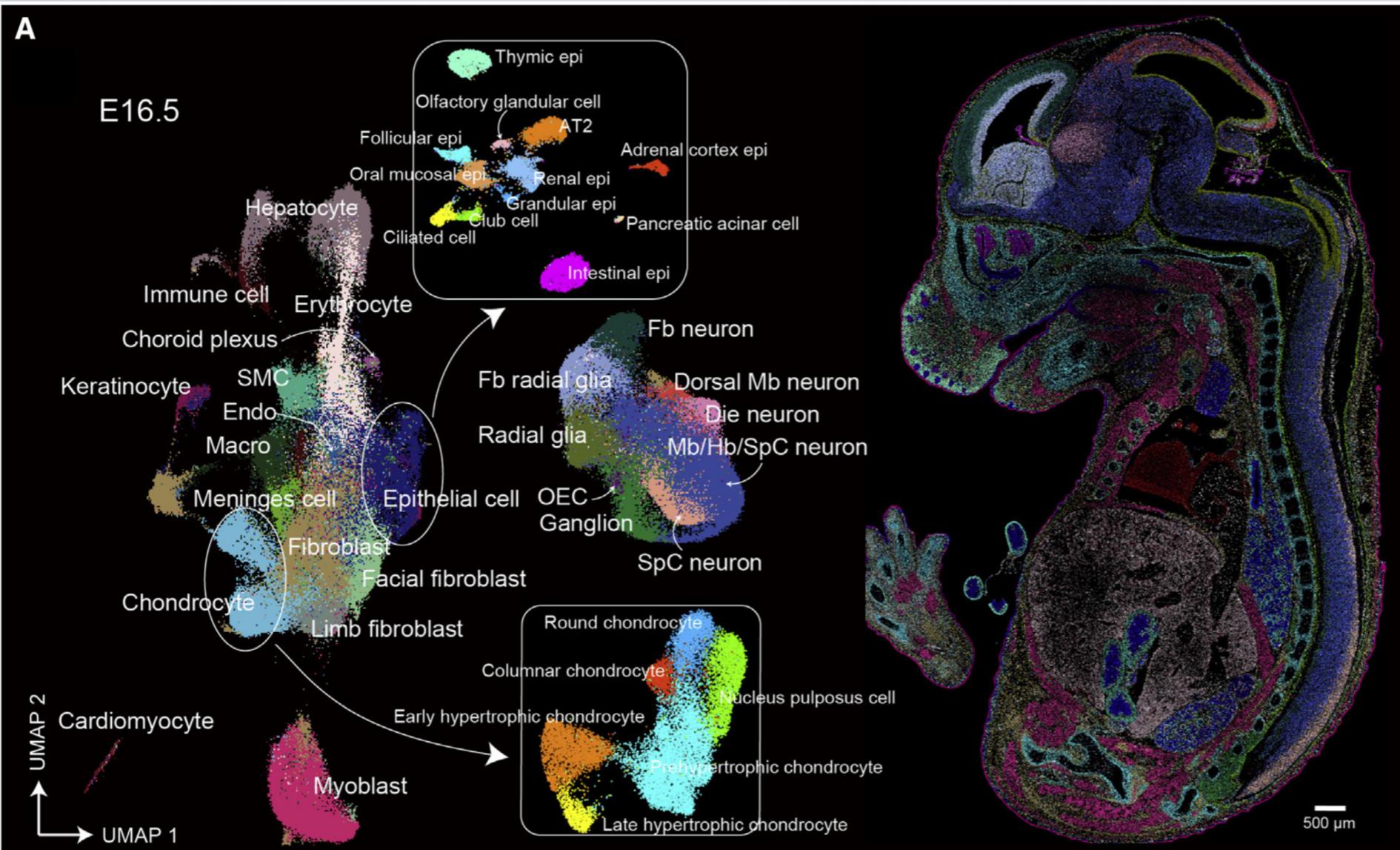
Cell clusters (segmented cells)



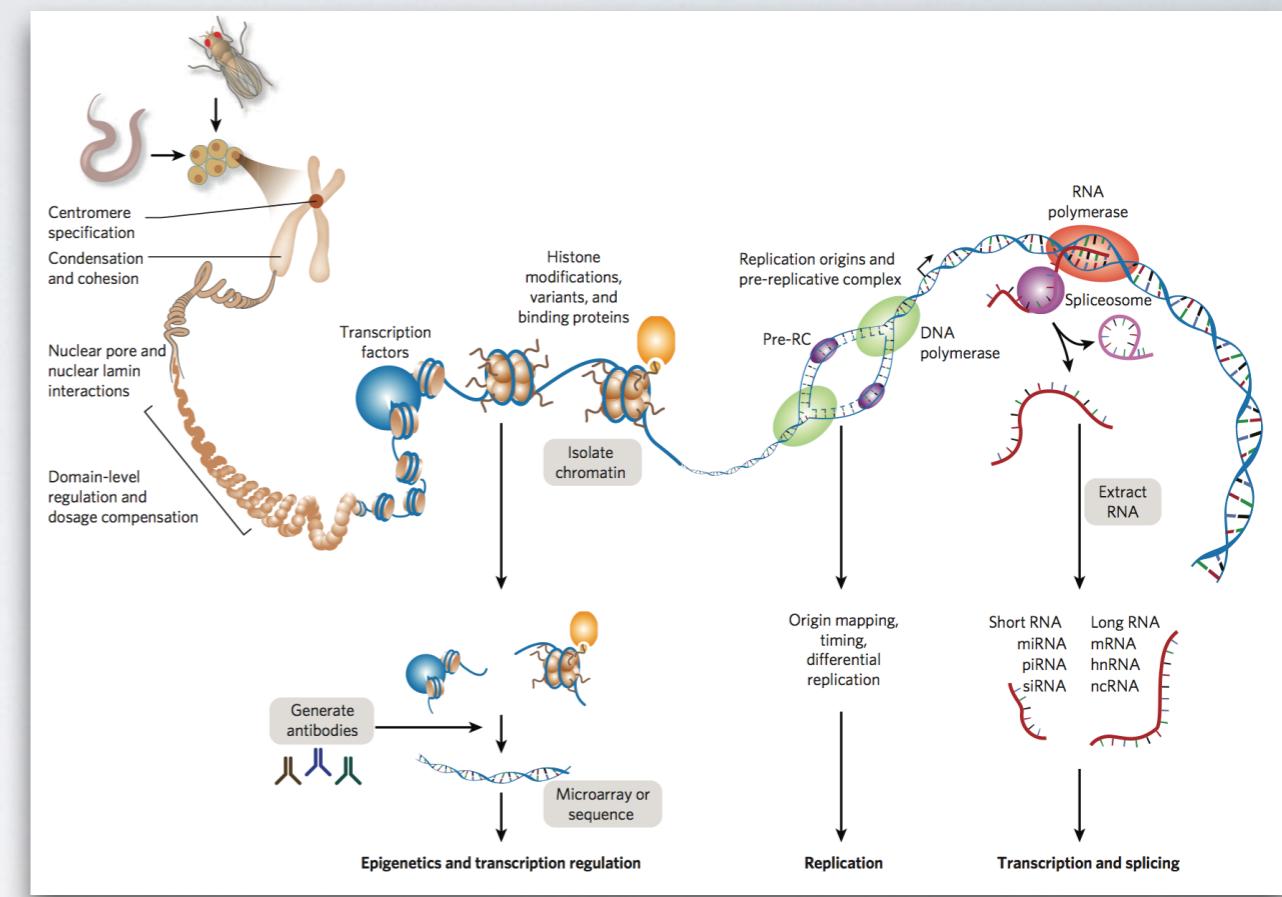
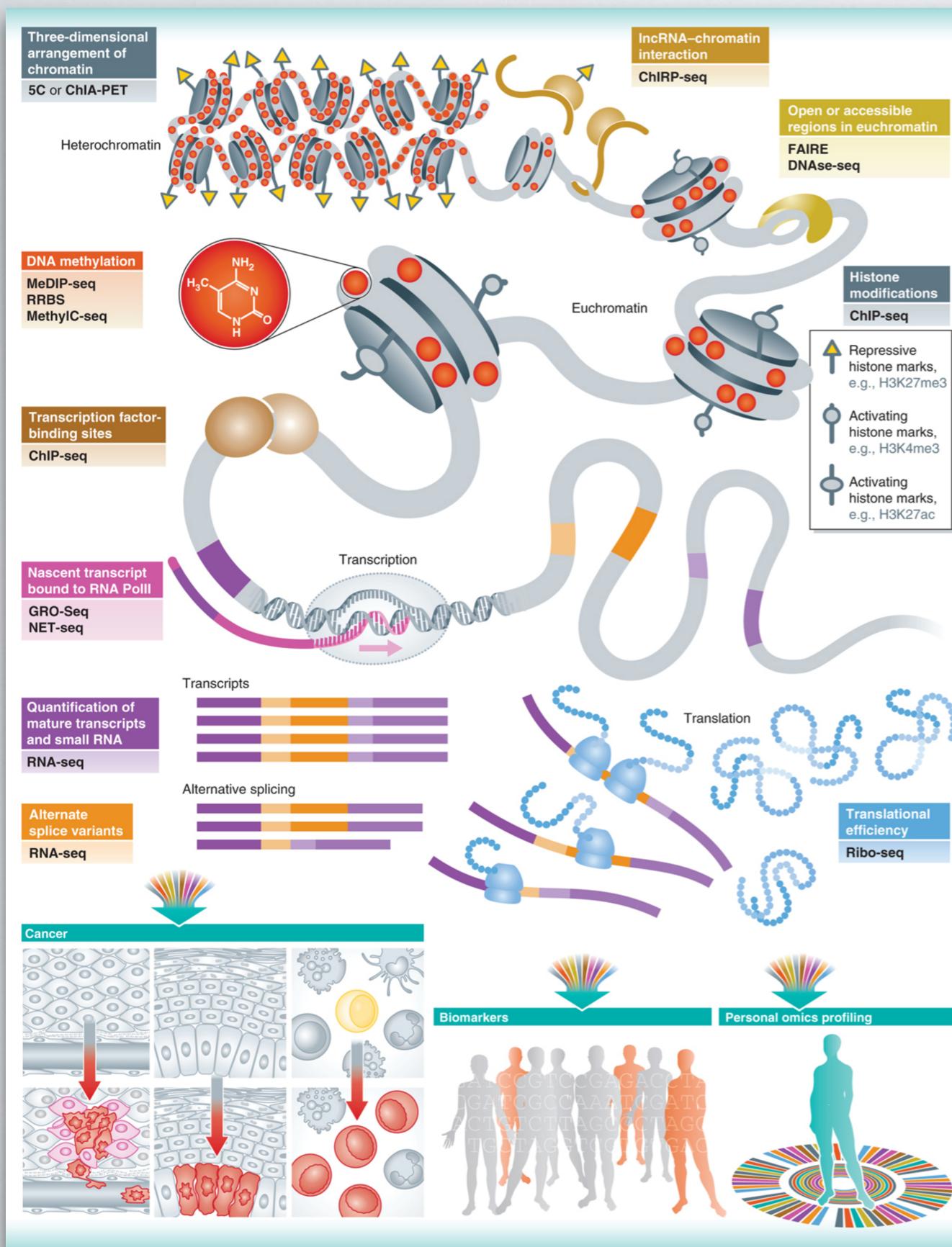
Spatially resolved transcriptomics



Spatially resolved transcriptomics



Exploration of nucleic acid based information through sequencing



Celniker et al., Unlocking the secrets of the genome, Nature 459, 927–930 (2009)

DNA sequencing is an incredibly versatile tool for understanding cellular activity

Anything that exists in nucleic acid form, or that has processes/interactions that can be represented in nucleic acid form, can be investigated by DNA sequencing

It is a powerful tool for reading this biological information

Lister Lab

Harry Perkins Institute of Medical Research
Centre of Excellence in Plant Energy Biology
School of Molecular Sciences
The University of Western Australia

ryan.lister@uwa.edu.au