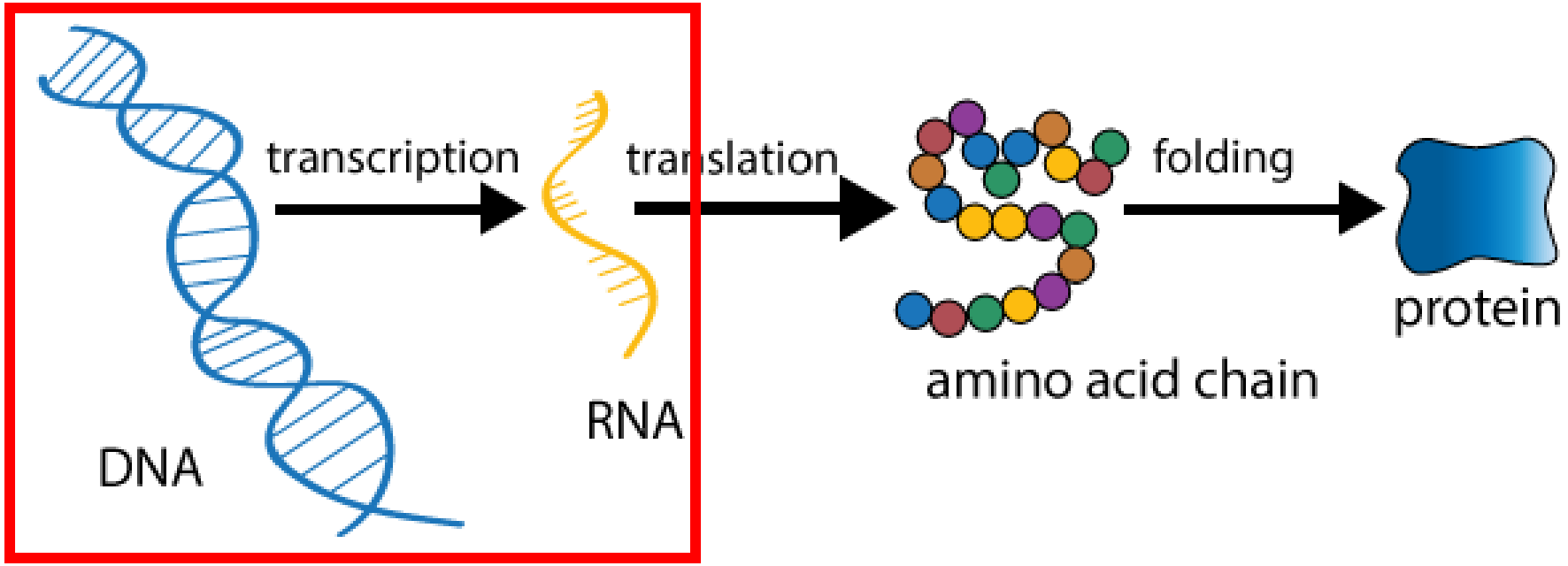
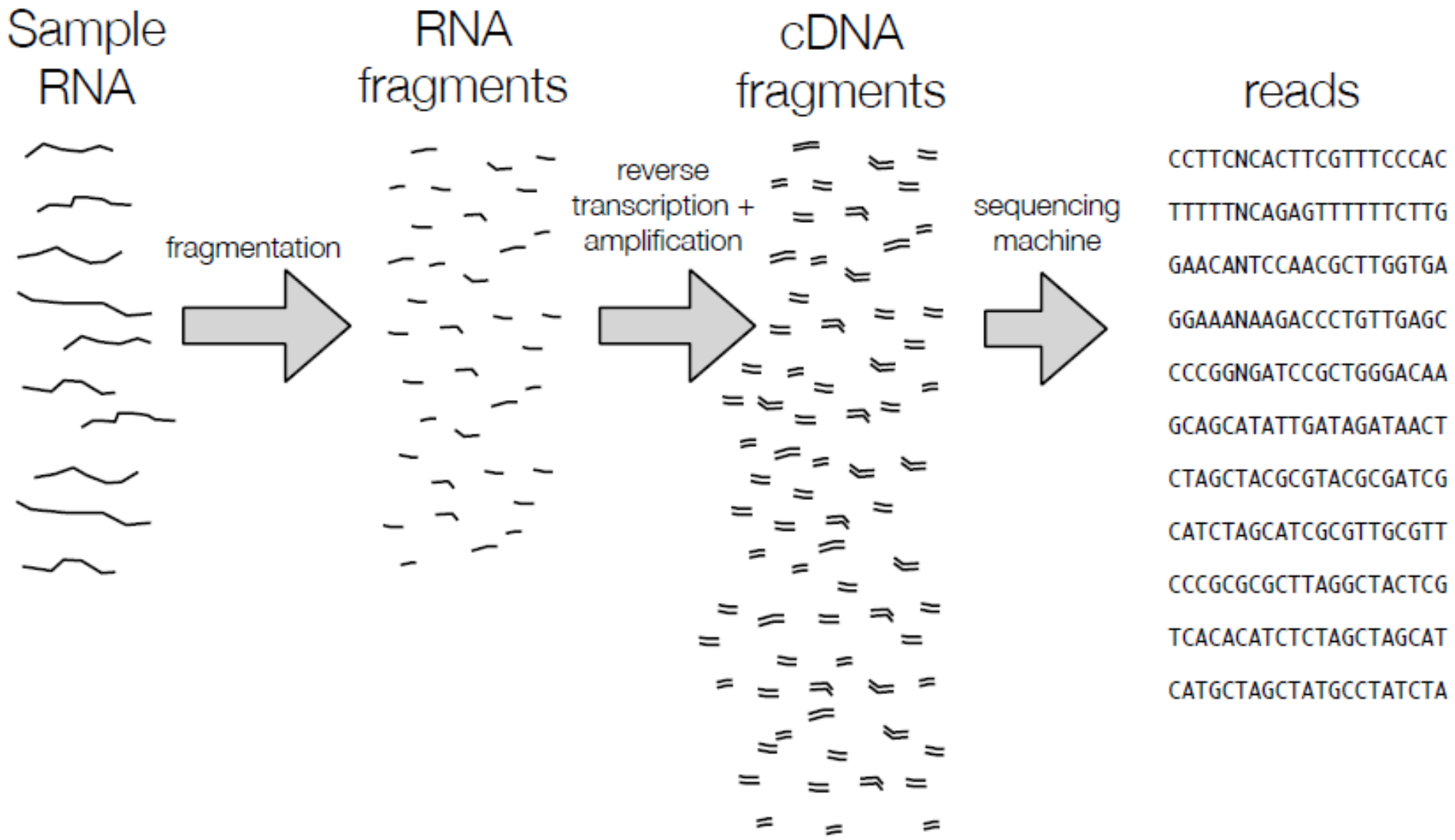
RNA-seq 101

Sep 23, 2019

Study transcription (because we can)



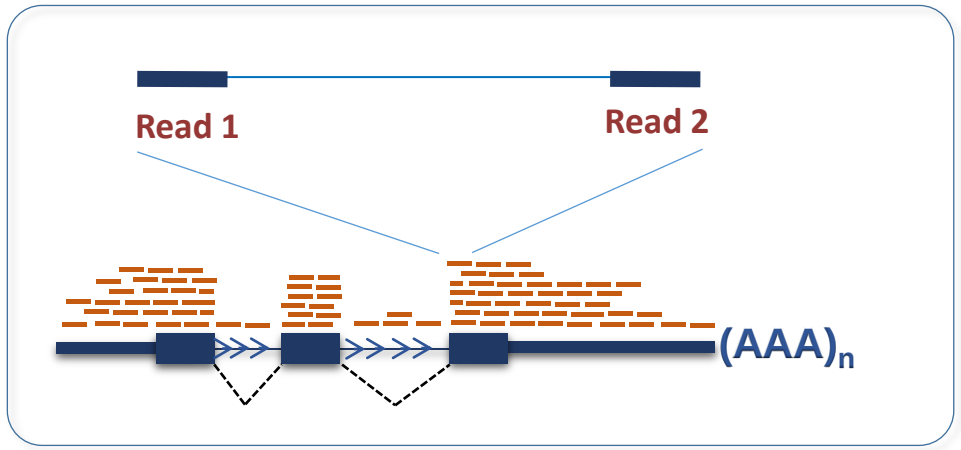
What is RNA-seq



http://www.biostat.wisc.edu/bmi776/lectures/rnaseq.pdf

RNA-seq alignment and quantification

- Reads are mapped (or aligned) to genome
- Expression is quantified based on genome annotation: a map of where the genes are located in the genome



Gene expression table

	title	Ctrl_0hrs_rep1	Ctrl_0hrs_rep2	Ctrl_1hrs_rep1	Ctrl_1hrs_rep2	Ctrl_2hrs_rep1	Ctrl_2hrs_rep2	Ctrl_4hrs_rep1	Ctrl_4hrs_rep2	Ctrl_6hrs_rep1	Ctrl_6hrs_rep2	DI_0hrs_rep1	DI_0hrs_rep2	DI_1hrs_rep1	D
2	treatment	Ctrl	Ctrl	Ctrl	Ctrl	Ctrl	Ctrl	Ctrl	Ctrl	Ctrl	Ctrl	DI	DI	DI	D
3	ENSMUSG00000000001	923	637	305	473	326	578	522	612	347	475	463	714	355	
4	ENSMUSG00000000003	0	0	0	0	0	0	0	0	0	0	0	0	0	
5	ENSMUSG00000000028	29	26	13	30	6	10	2	3	0	0	6	27	8	
6	ENSMUSG00000000037	5	1	2	0	0	0	0	1	0	0	0	0	0	
7	ENSMUSG00000000049	0	0	0	0	0	0	0	0	0	0	0	0	0	
8	ENSMUSG00000000056	435	316	119	212	54	105	12	23	39	60	188	168	163	
9	ENSMUSG00000000058	148	142	82	186	60	95	31	49	40	36	122	214	110	
10	ENSMUSG00000000078	12109	8717	27631	52634	24103	51526	22181	24186	18669	20819	3523	8595	21196	
11	ENSMUSG00000000085	83	68	16	30	13	15	16	16	14	19	62	49	28	
12	ENSMUSG00000000088	1277	1069	586	824	384	938	523	722	379	495	730	526	413	
13	ENSMUSG00000000093	1	2	4	0	0	0	0	0	0	0	0	4	0	
14	ENSMUSG00000000094	0	0	0	0	0	0	0	0	0	0	0	0	0	
15	ENSMUSG00000000103	0	0	0	0	0	0	0	0	0	0	0	0	0	
16	ENSMUSG00000000120	0	0	0	0	0	0	0	0	0	0	0	0	0	
17	ENSMUSG00000000125	0	0	0	0	0	0	0	0	0	0	0	0	0	
18	ENSMUSG00000000126	0	0	3	0	0	0	0	0	2	1	2	0	0	
19	ENSMUSG00000000127	415	286	100	242	136	262	122	95	78	64	114	471	128	
20	ENSMUSG00000000131	530	503	242	314	203	379	232	307	158	248	272	398	195	
21	ENSMUSG00000000134	1088	888	360	423	177	333	261	295	268	767	561	954	316	
22	ENSMUSG00000000142	0	0	0	0	0	1	0	0	0	0	0	0	0	
23	ENSMUSG00000000148	111	80	21	50	26	18	8	21	12	16	77	108	16	
24	ENSMUSG00000000149	1089	1395	426	895	235	495	85	120	38	49	523	734	476	
25	ENSMUSG00000000154	9	12	0	5	4	12	2	6	1	4	0	0	0	
26	ENSMUSG00000000157	5	3	3	9	3	8	2	2	0	4	6	0	1	
27	ENSMUSG00000000159	0	0	0	0	0	0	0	0	0	0	0	0	0	
28	ENSMUSG00000000167	4	5	3	0	2	1	3	6	0	0	0	0	0	
29	ENSMUSG00000000168	205	125	73	83	48	82	33	62	18	37	118	142	63	
30	ENSMUSG00000000171	110	83	40	75	48	70	37	38	29	36	105	88	55	
31	ENSMUSG00000000182	0	0	0	0	0	0	0	0	0	0	0	0	0	
32	ENSMUSG00000000183	0	0	0	0	0	0	0	1	0	0	0	0	0	
33	ENSMUSG00000000184	5147	4766	2571	5587	5149	10619	15481	20315	16075	17627	329	678	548	
34	ENSMUSG00000000194	891	675	237	429	194	454	202	271	220	204	329	692	322	
35	ENSMUSG00000000197	0	0	0	0	0	0	0	0	0	0	0	0	0	

Main types of RNAs

- rRNA – ribosomal RNA: 80% of the cell RNA
- tRNA – transfer RNA: 15% of the cell RNA
- **mRNA** – messenger RNA for protein coding genes
- Other RNAs: miRNA, lncRNA, …

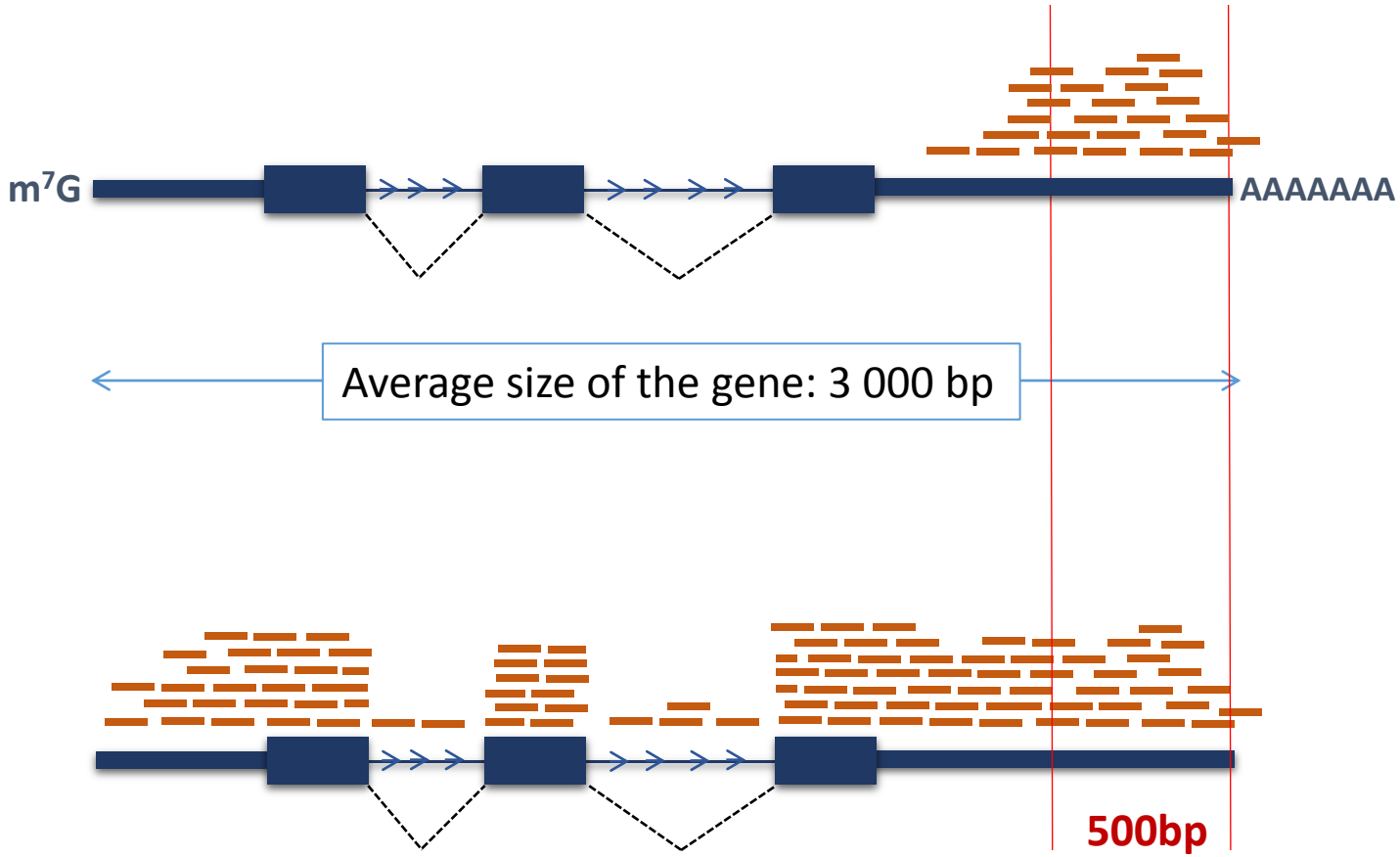
- Some of the RNAs are short: tRNA, miRNA, … and are not getting into normal RNA-seq

https://www.ncbi.nlm.nih.gov/books/NBK21729/

Two main approaches for RNA selection

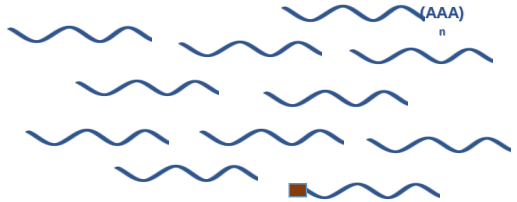
- polyA selection: most standard, relatively cheap and easy protocol, selects mRNAs and some non-coding RNAs
- riboZero: depletes rRNA, works better for degraded RNA, captures all long RNAs

Going after gene expression only

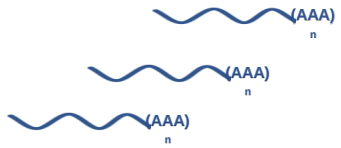


is enough to get accurate level of expression across the whole transcript

3´end RNA-Seq pipeline

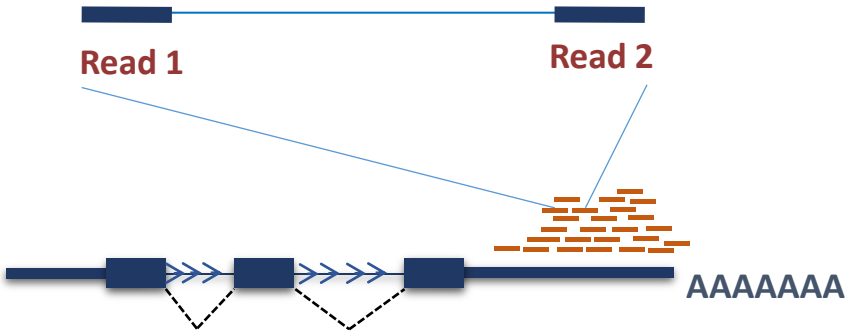


Fragmentation



PolyA selection of mature transcripts

Library prep



Reads, paired

Expression and annotation

RNA specific library prep

Starting from fragmented RNA

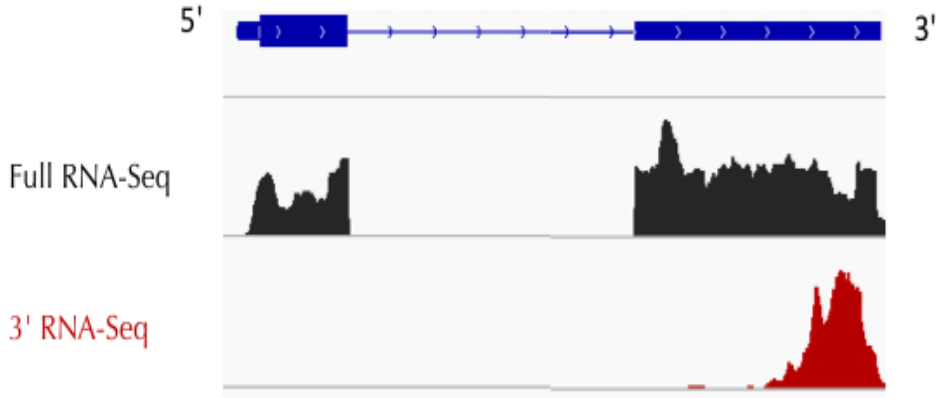
Custom oligo-dT cDNA synthesis

AAAAAAAAAAAAAAA-
LINKERSEQUENCE-TTTTTTTTTTTTTTTTT-

Second linker ligation

Enrichment and extension

3’ vs full RNA-seq data



Early Barcoding Strategy

Starting from fragmented RNA

Custom oligo-dT cDNA synthesis

AAAAAAAAAAAAAAA-

LINKERSEQUENCE-BARCODE-TTTTTTTTTTTTTTTTT-

Second linker ligation

Enrichment and extension

Sensitivity/throughput improved N-fold by early pooling

Sample 1	Sample 2	.....	Sample N
cDNA synthesis with LINKERSEQUENCE-BARCODE1-T15	cDNA synthesis with LINKERSEQUENCE-BARCODE2-T15	.....	cDNA synthesis with LINKERSEQUENCE-BARCODEN-T15

Pooling samples

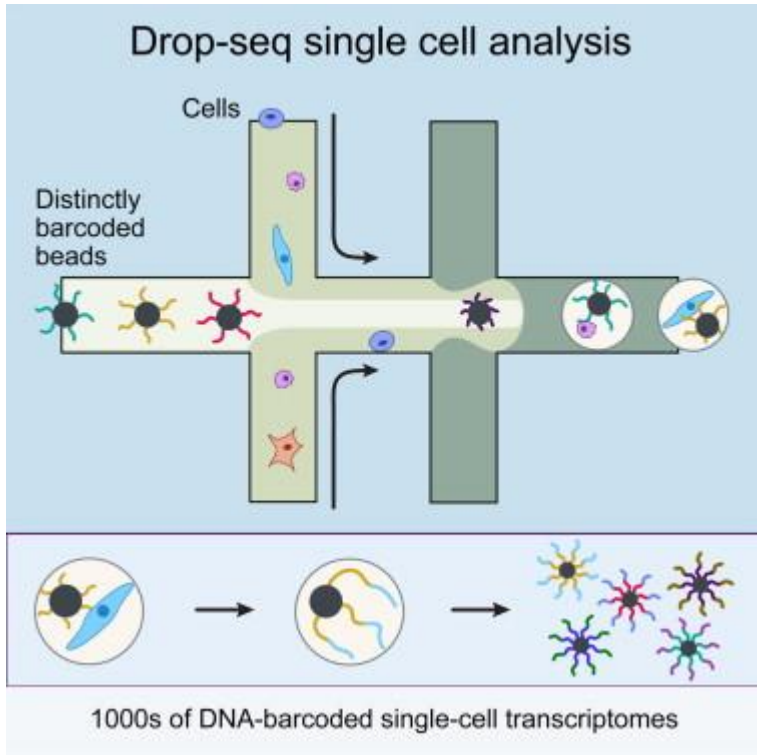
single library prep for N samples

Second linker ligation

Enrichment and extension

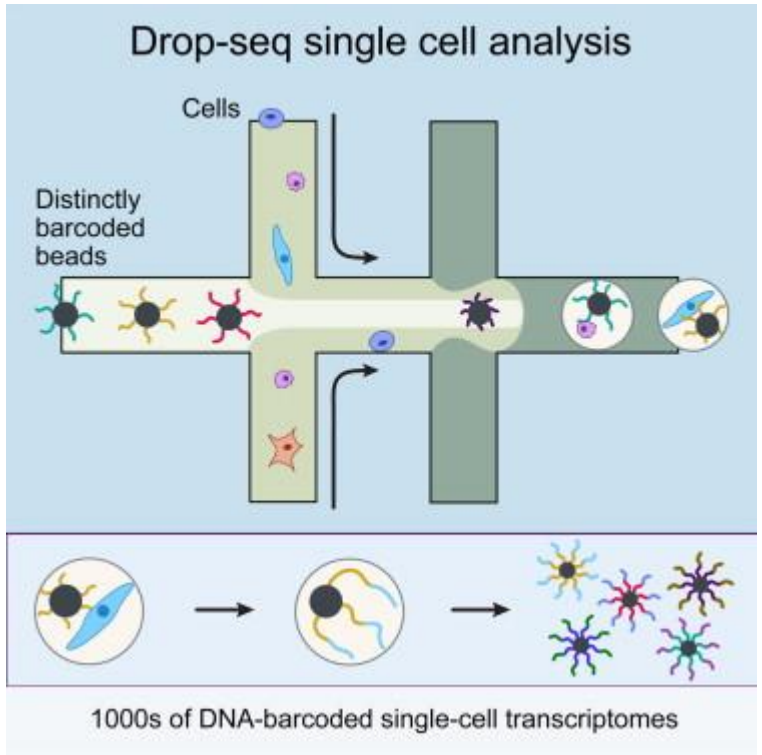
If you use microfluidics, N can be very large and pooling can be automatic!

Sample 1	Sample 2	.....	Sample N
cDNA synthesis with LINKERSEQUENCE-BARCODE1-T15	cDNA synthesis with LINKERSEQUENCE-BARCODE2-T15	.....	cDNA synthesis with LINKERSEQUENCE-BARCODEN-T15



If you use microfluidics, N can be very large and pooling can be automatic!

Sample 1	Sample 2	.....	Sample N
cDNA synthesis with LINKERSEQUENCE-BARCODE1-T15	cDNA synthesis with LINKERSEQUENCE-BARCODE2-T15	.....	cDNA synthesis with LINKERSEQUENCE-BARCODEN-T15



More on this in single-cell module from Kostya tomorrow

Sequencing: HiSeq 2500 ->HiSeq 4000 -> Hiseq X

- Output is in millions of reads
- Reads are typically of 50bp and 100bp
- Single end or paired end

To run sequencing you have to fill flow cell:

- Rapid Run Mode:
  - Flow Cell consists of two lanes,
  - ~120 M reads from each lane
- Productivity Mode:
  - Flow Cell Consists of 8 lanes
  - ~200-250M reads from each lane

Sequencing depth/replication

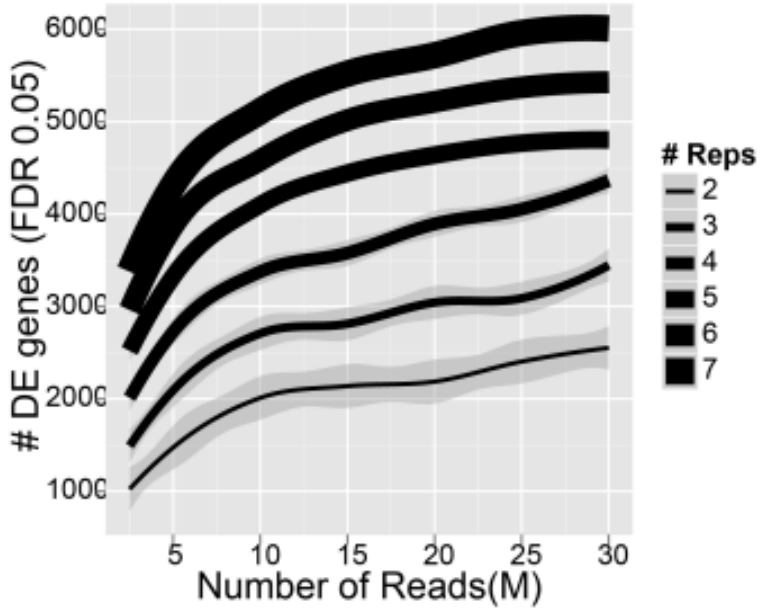
The more the merrier!

More than 60M reads is
almost never needed

15-20M reads per sample is
good sequencing depth

Even 3-4 million is enough to estimate
differential expression well

Usually it's better to increase the number of
biological replicates instead of library depth

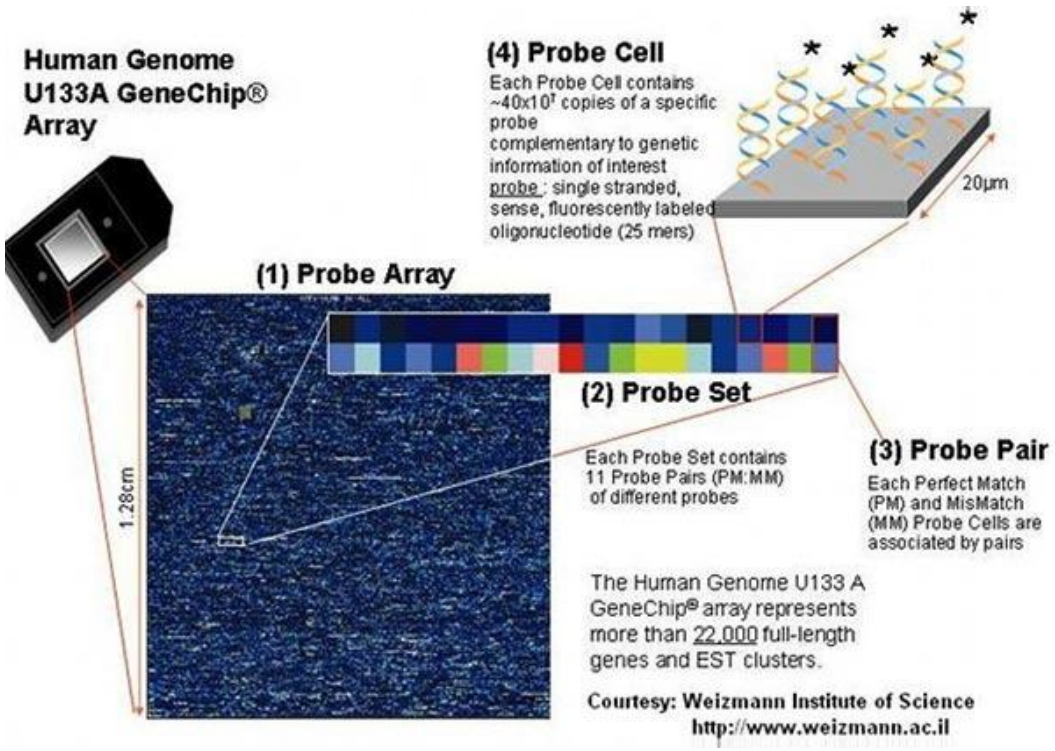


Number of differentially expressed
genes increases as sequencing depth increases

Measuring gene expression using microarrays

- Instead of sequencing, expression is measured via hybridization and fluorescence
- Can be 50k to 1M probes per array
- Probes should be annotated: what genes they are assigned to

- Were developed earlier than RNA-seq but still used today
- Can measure only predefined things



https://www.oceanridgebio.com/affymetrix_genechip