

RNA sequencing : library preparation and experimental design

Céline Keime
keime@igbmc.fr

RNA sequencing

- Introduction
- Preparation of RNA-seq libraries
- Design of RNA-seq experiments
- RNA-seq bias already identified

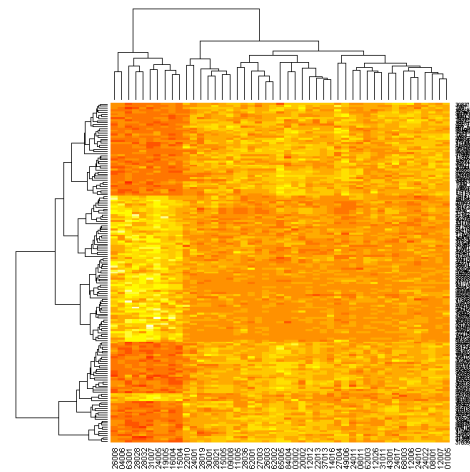
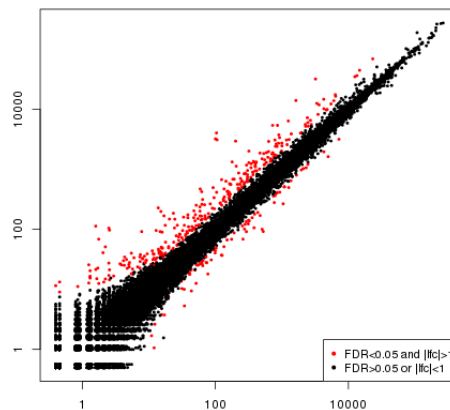
RNA sequencing

- Introduction
- Preparation of RNA-seq libraries
- Design of RNA-seq experiments
- RNA-seq bias already identified

Transcriptome analysis : key aims

■ Quantitative

- Quantify the changes of expression level between different conditions / time points



■ Qualitative

- Catalogue all different transcripts (mRNA, ncRNA)
- Determine the structure of these transcripts
 - TSS, 3' end, splicing patterns, post-transcriptional modifications



Transcriptome analysis : different technologies

- Hybridization-based approach

- Microarrays



- Drawbacks

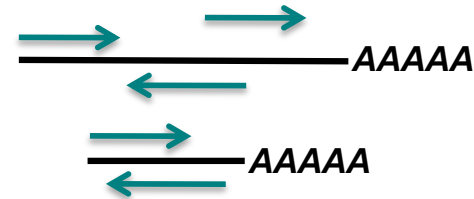
- Reliance upon existing knowledge on transcriptome
 - Poor quantification of lowly (background) and highly (saturation) expressed genes
 - Cross-hybridization

Transcriptome analysis : different technologies

■ Sequence-based approaches

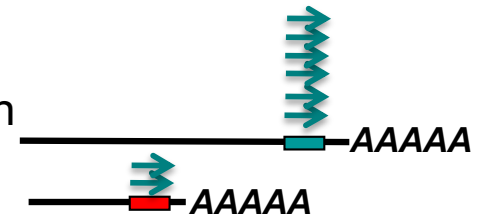
■ EST (Expressed Sequence Tag)

- Sequence of a cDNA fragment
- Drawbacks
 - Sanger sequencing → low throughput
 - Generally not quantitative (normalized libraries)



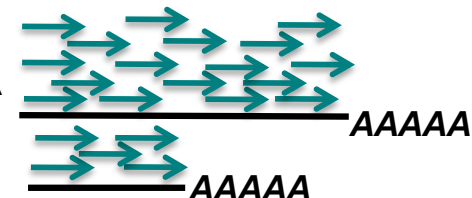
■ SAGE (Serial Analysis of Gene Expression)

- Sequence a tag : short fragment from a specific location of each transcript
- Drawback : only a portion of the transcript is analysed (isoforms are generally indistinguishable from each other)



■ RNA-seq

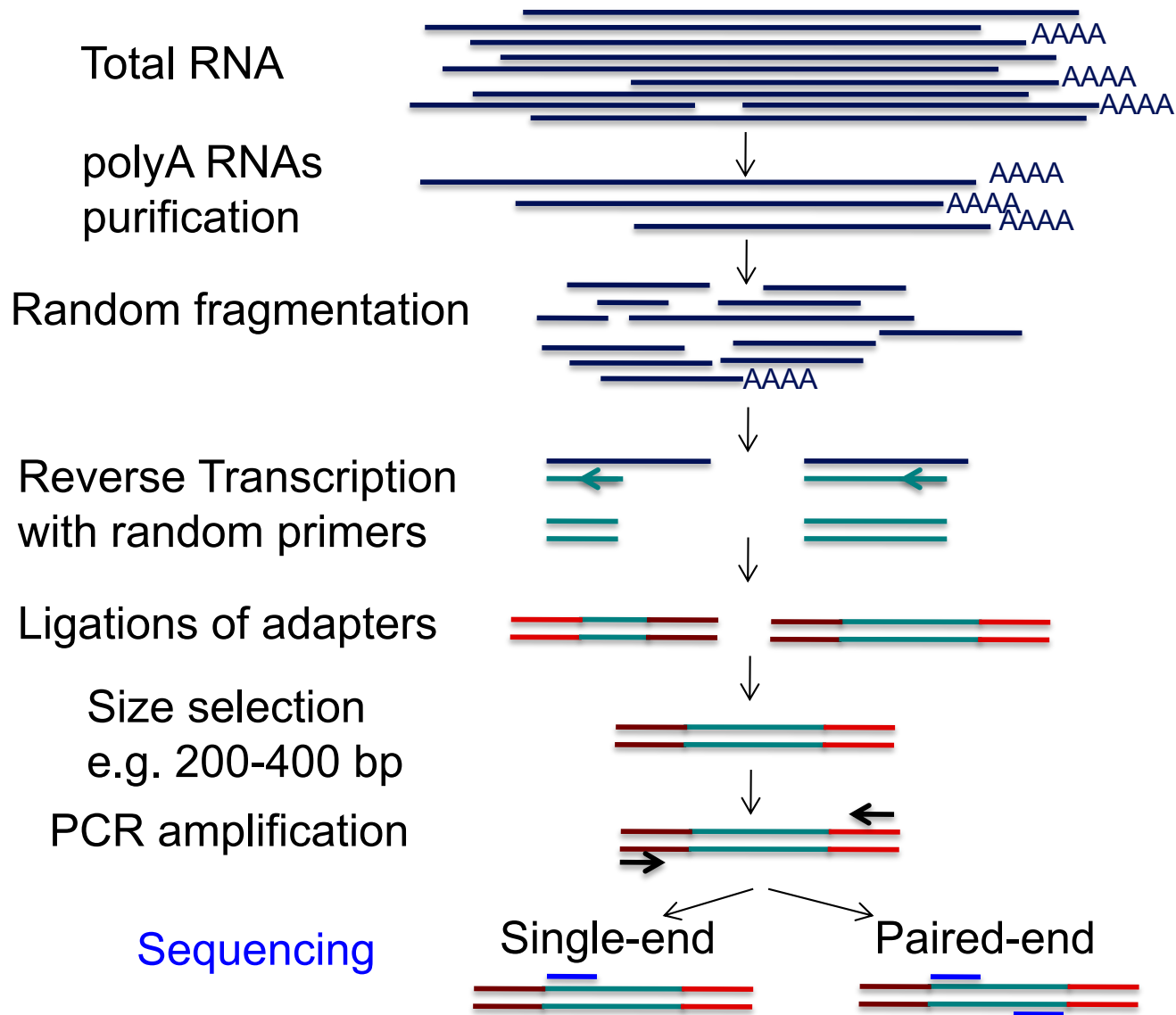
- Sequence cDNA fragments from the whole cDNA
- Qualitative and quantitative information



RNA sequencing

- Introduction
- Preparation of RNA-seq libraries
- Design of RNA-seq experiments
- RNA-seq bias already identified

RNA-seq library preparation

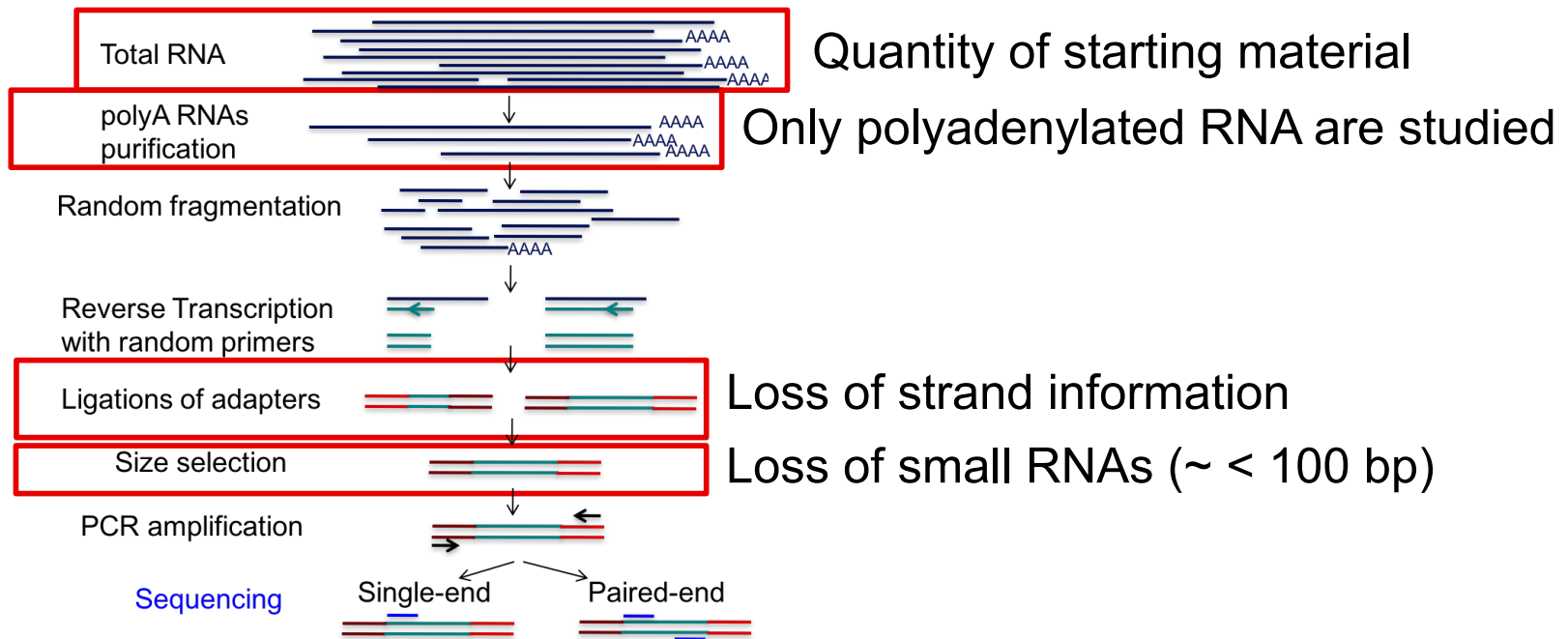


RNA-seq library preparation

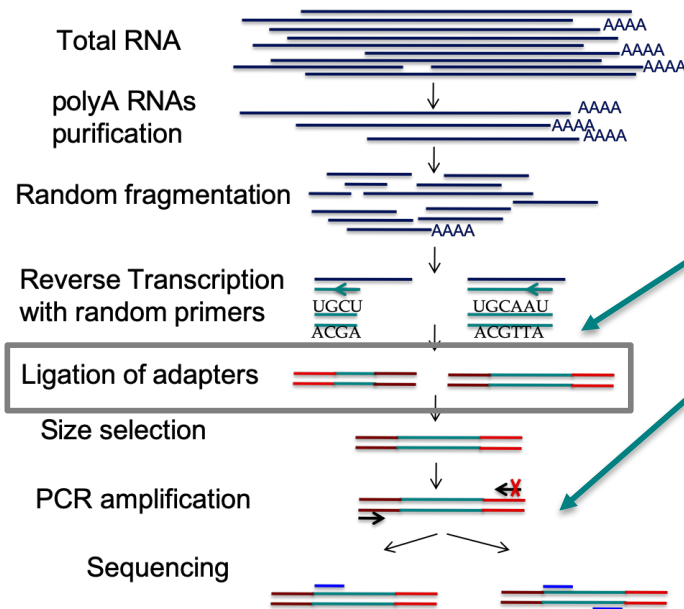
■ Advantages

- Highly reproducible
- High sensitivity
- Allows to study both coding and non-coding polyA+ RNAs expression
- Allows transcript discovery

■ Limitations



RNA-seq library preparation : stranded protocols



Loss of strand information



Stranded (directional) protocols

*Incorporation of dUTP instead of dTTP
in the second strand cDNA synthesis*

*Amplification of reverse strand only
(high fidelity Taq polymerase)*

PolyA+ RNA

→ Illumina TruSeq Stranded mRNA Prep

→ Illumina Stranded mRNAseq Prep,
Ligation

Not limited to polyA+ RNA

→ Illumina TruSeq Stranded Total Prep

→ Illumina Stranded Total RNA Prep Ligation
with Ribo-Zero Plus

Total RNA quantity

Minimal	Optimal
200 ng	1 µg
25 ng	1 µg
100 ng	1 µg
1 ng	1 µg

Advantage : preserves the strand information

→ Allows to determine transcript orientation

→ Important for novel transcript discovery and annotation,
especially for overlapping transcripts

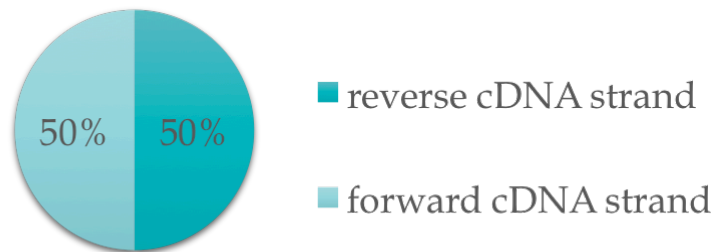
RNA-seq library preparation : stranded protocols

■ Good quality of strand-specificity

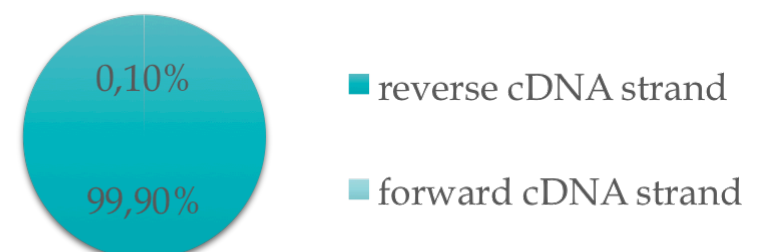
e.g. Results obtained on spike-in RNAs added in 4 libraries prepared with both standard and directional polyA+ RNA-seq protocols (*GenomEast Platform*)

Proportion of reads from each cDNA strand :

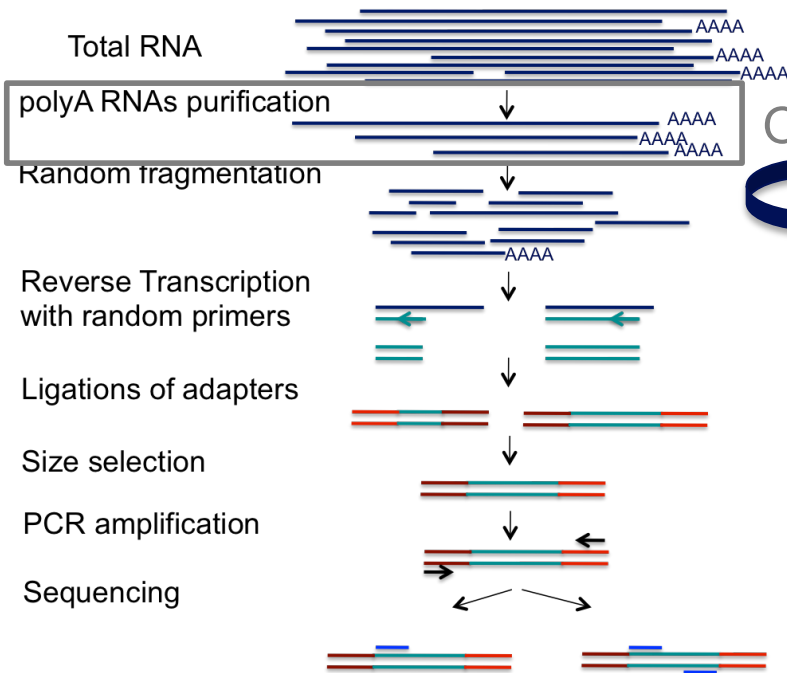
standard mRNA-seq



directional mRNA-seq



RNA-seq library preparation : protocols not limited to polyA+ RNA



Only polyadenylated RNA are studied



Protocols not limited to polyA+ RNA :

- Illumina TruSeq Stranded Total Prep
- Illumina Stranded Total RNA Prep
- Ligation with Ribo-Zero Plus

Total RNA quantity

	Minimal	Optimal
→ Illumina TruSeq Stranded Total Prep	100 ng	1 µg
→ Illumina Stranded Total RNA Prep	1 ng	1 µg

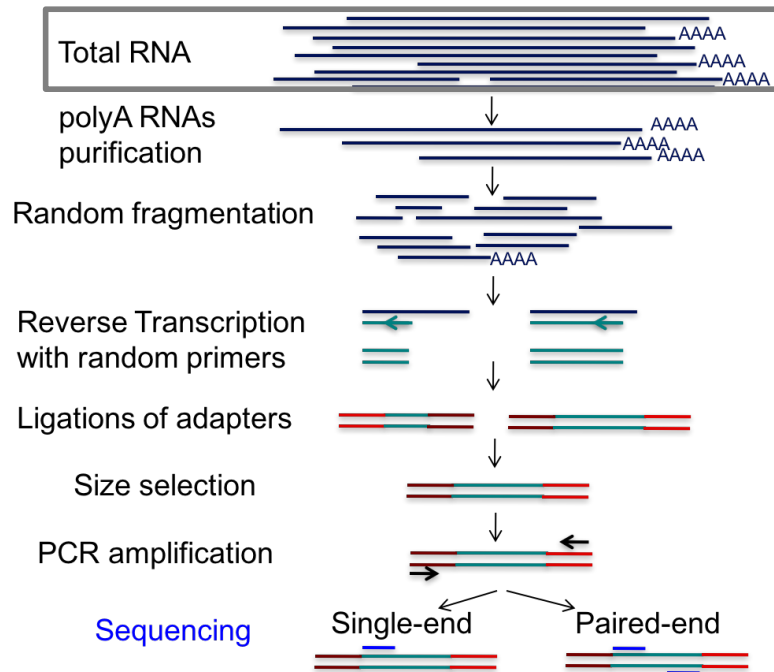
■ Advantage

- Allows to study non-polyadenylated transcripts

■ Drawbacks

- Efficiency of rRNA removal ≠ between samples
- Higher number of RNA molecules sequenced compared to standard RNA-seq
- ➔ More reads needed to achieve the same coverage on polyadenylated RNAs

RNA-seq library preparation : protocol with amplification



Quantity of starting material



Protocol with amplification :

Clontech SMART-Seq v4 UltraLow input RNA
cDNA preparation : cDNA synthesis + amplification
Illumina Nextera XT DNA sample preparation :
library preparation

Optimal quantity : 10 ng

Minimal quantity : 150 pg

or cells (100-1000 cells)

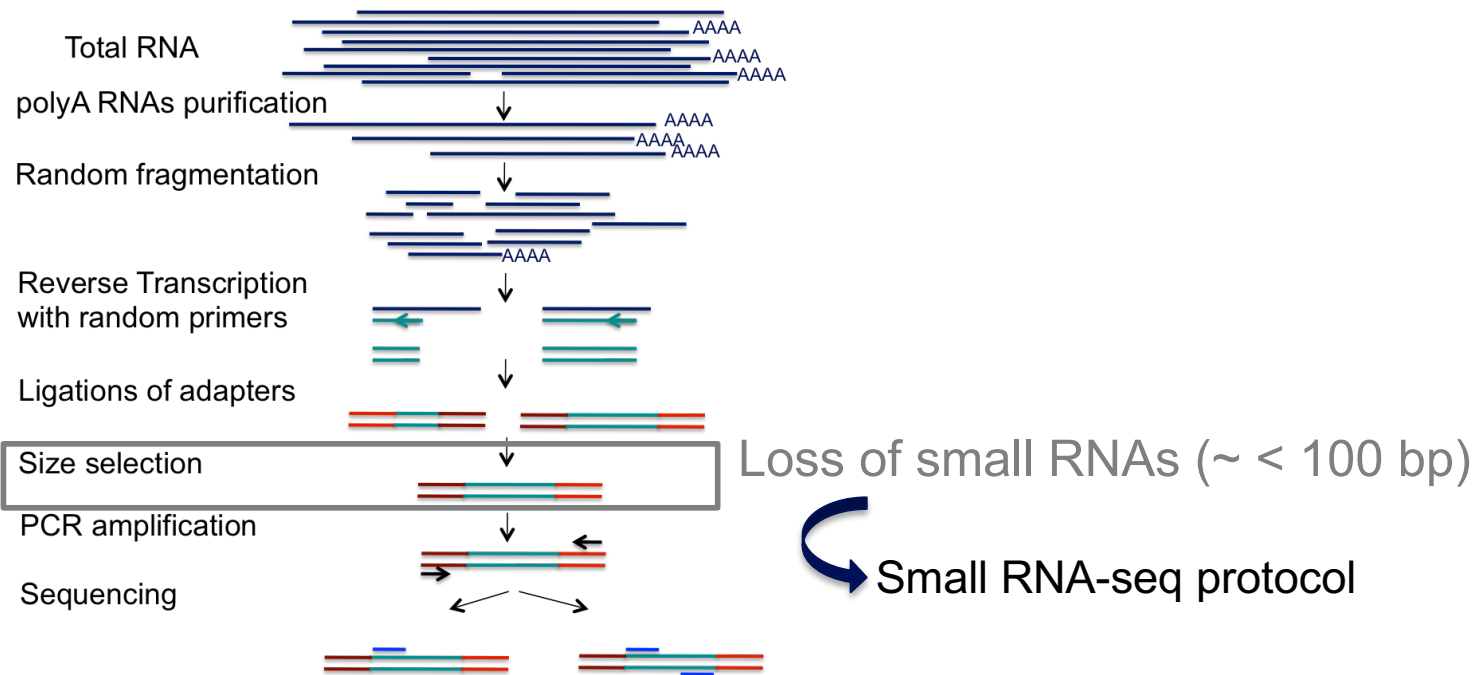
■ Advantage

- Low quantity of starting material

■ Drawback

- Bias due to the amplification

small RNA-seq library preparation



small RNA-seq library preparation

Illumina Truseq smallRNA SamplePrep

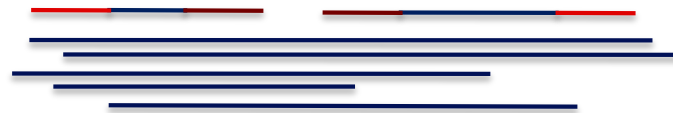
Total RNA



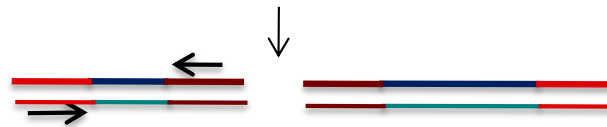
Optimal quantity : 2µg

Minimal quantity : 1µg

**Ligation of adapters
on 5'P and 3'OH**



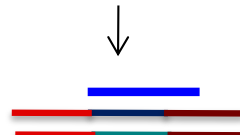
RT 1st strand synthesis
PCR amplification



Adapted size selection

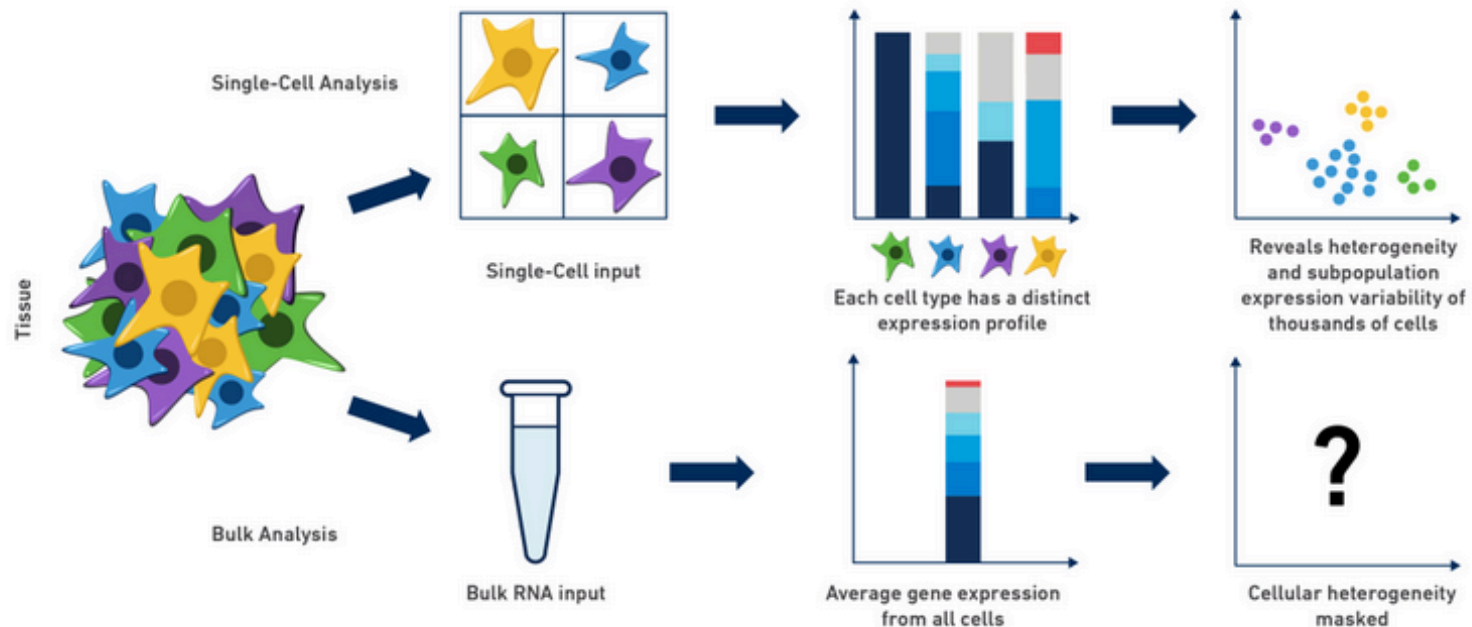


Sequencing



Single-cell RNA-seq

- Majority of RNA-seq experiments : study of a cell population
- Overlooks differences within a cell population that may be important for maintaining normal tissue function or facilitating disease progression
- Single-cell RNA-seq provides the expression profiles of individual cells
 - Allows to characterize the subpopulation structure
 - Allows to study cell heterogeneity

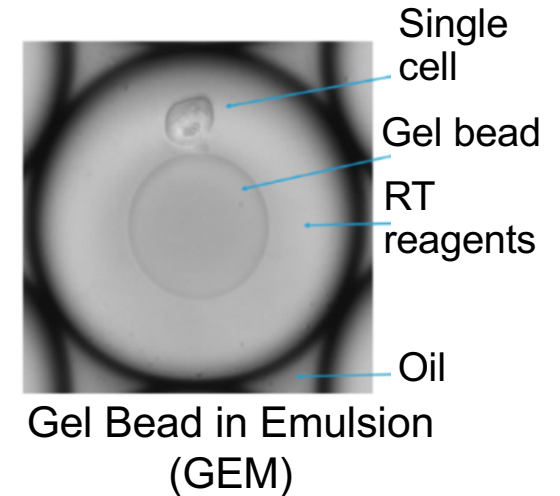
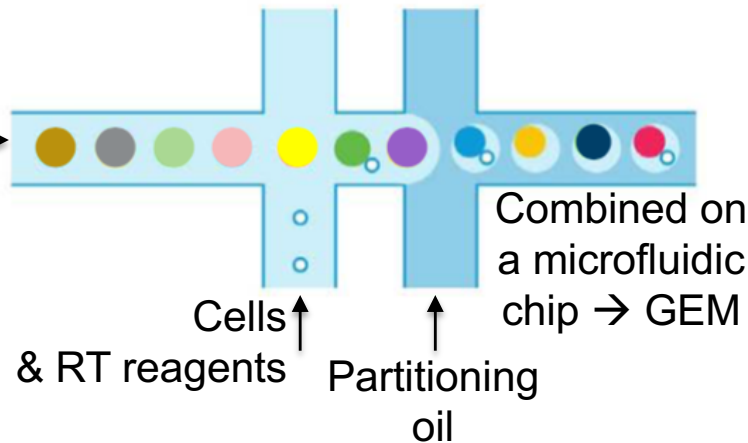
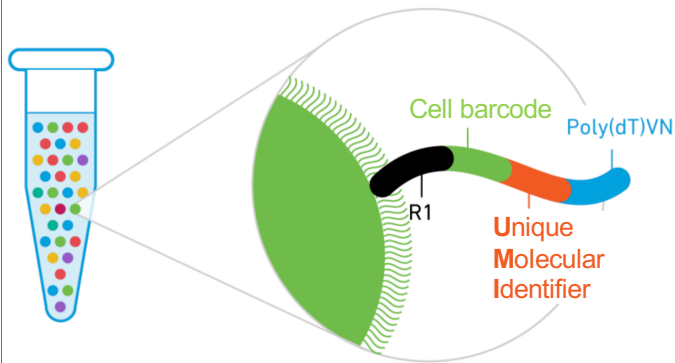


Single-cell RNA-seq

- Different technologies for single cell collection
 - Droplets (e.g. Chromium, 10X Genomics)
 - Microfluidics (e.g. C1 single-cell auto-prep system, Fluidigm)
 - Microwells (e.g. Rhapsody single-cell analysis System, BD)
- Different protocols for RNA-seq
 - 3' counting or full-length
 - With or without Unique Molecular Identifiers (UMI)
 - Random sequences used to tag each molecule prior to library amplification
 - 2 reads align to the same location and have the same UMI
→ highly likely PCR duplicates
- Limits
 - Technical noise due to amplification and dropout

10X Genomics single-cell RNA-seq

Gel beads containing millions of barcoded oligos

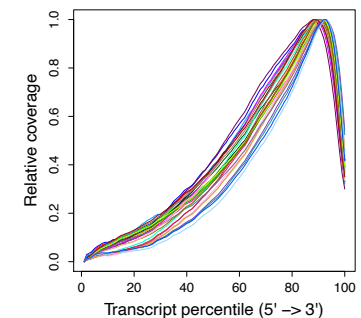
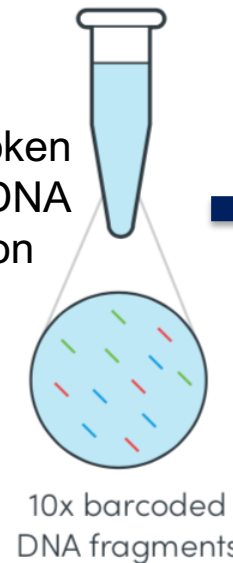
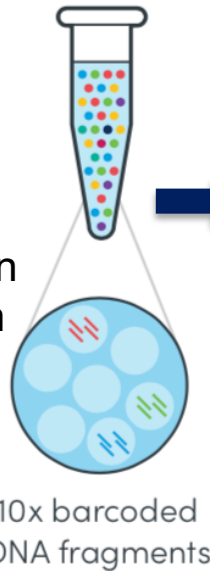


Inside individual GEM:

- Gel bead dissolved
- Cell lysis
- Reverse transcription → Barcoded cDNA from polyadenylated RNA

GEM are broken → Pooled cDNA amplification

Library construction → 3' end of the transcript



From www.10xgenomics.com

RNA sequencing

- Introduction
- Preparation of RNA-seq libraries
- Design of RNA-seq experiments
- RNA-seq bias already identified

Experimental design

1. Define your biological questions of interest
 2. Define the best appropriate experimental design to answer these questions :
 - Library preparation protocol
 - Sequencing strategy
 - Number of reads
 - Number of replicates
- Define a detailed experimental plan in advance of doing the experiment
 - Try to reduce batch effects
 - ENCODE guidelines (mammalian tissues)
<https://www.encodeproject.org/about/experiment-guidelines/>

Which protocol for which application ?

- Choice depend on
 - Quantity of starting material
 - Type of RNA studied (small/long, polyA+/-)
 - Biological questions of interest
 - e.g. new transcript identification → directional protocol
- Keep the same protocol for all samples within a project

Which protocol for which application ?

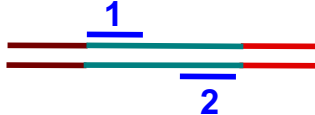
Kit used by the platform	Total RNA quantity		Type of studied RNA	Stranded ^a
	Minimal	Optimal		
Truseq Stranded mRNA Prep Illumina	200 ng	1 µg	Only polyA+ RNA of size > 100 b	Yes
Illumina Stranded mRNA Prep, Ligation	25 ng	1 µg	Only polyA+ RNA of size > 100 b	Yes
SMART-Seq v4 UltraLow Input RNA kit (Clontech) + Nextera XT DNA sample preparation Kit (Illumina)	100 cells	10 ng	Only polyA+ RNA of size > 100 b	No
Chromium Next GEM Single Cell 3' Reagent Kits	1 cell	1 cell	Only polyA+ RNA	No
Truseq Stranded Total RNA Sample Prep Illumina	100 ng	1 µg	All RNA of size > 100 b	Yes
Illumina Stranded Total RNA Prep Ligation with Ribo-Zero Plus	1 ng	1 µg	All RNA of size > 100 b	Yes
Ovation RNA-Seq System V2 + Ovation SP Ultralow Library systems (NuGEN)	500 pg	10 ng	All RNA of size > 100 b	No
Truseq SmallRNA Sample Prep (Illumina)	1 µg	2 µg	All small RNAs with 5'P and 3'OH (desired size can be chosen by the project manager)	Yes

Which sequencing strategy ?

- Expression quantification on annotated transcripts

- Single-end sequencing provides good results 

- Alternative splicing analysis, fusion transcript detection, mapping over repetitive regions, de novo transcriptome assembly

- Paired-end sequencing is needed 

How many reads are needed ?

- Transcriptome coverage as a function of sequencing depth: highly dependant on transcriptome complexity
- Sequencing depth should be determined by the goals of the experiment
- General recommendations for typical mammalian tissues
 - > 30 million reads with polyA+ protocols
 - > 50 million reads with total protocols
 - ... if the goal is to quantify expression of annotated genes
- Higher sequencing depth needed if
 - the sensitivity of detection is important
 - the purpose is to discover novel transcripts
 - the purpose is to precisely quantify transcript isoforms

Examples
on our Hiseq4000

Application	Suggested multiplexing <i>for standard experiments on mammalian genomes</i>
small RNA-seq	20 samples / lane
mRNA-seq with polyA selection → for gene expression quantification → for alternative splicing analysis	8 samples / lane 3 samples / lane
RNA-seq with ribodepletion → for gene expression quantification	5 samples / lane

How many replicates are needed ?

- Low technical variability
and technical variability \ll biological variability
(Marioni et al. Genome Research 2008. Bullard et al. BMC Bioinformatics 2010)
➔ Technical replicates not required
- But “sequencing technology does not eliminate biological variability”
(Hansen et al. Nat Biotechnol. 2011)
 - **Biological replicates are fundamental !**
 - How many ?
 - Highly dependant on the correlation between replicates and on the difference between the compared conditions
 - If possible, prepare more samples for low-input RNA-seq

RNA sequencing

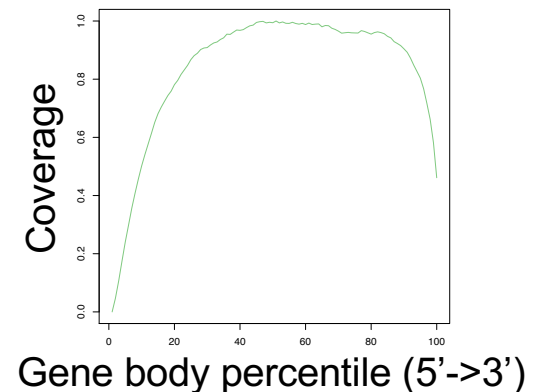
- Introduction
- Preparation of RNA-seq libraries
- Design of RNA-seq experiments
- RNA-seq bias already identified

RNA-seq bias / sources of variability

- As all techniques, RNA-seq present bias affecting expression estimates and subsequent statistical analysis
- Identification of bias in RNA-seq protocol
 - Use of synthetic spike-in standards
(Jiang et al. Genome Research 2011;21(9):1543-51)
 - Provided by ERCC (External RNA Control Consortium)
 - 92 sequences
 - Minimal sequence homology with endogenous transcripts from sequenced eukaryotes
 - Various lengths and GC content, large range of concentrations

RNA-seq bias / sources of variability

- Composition bias of the first 13 nucleotides due to a non-random hexamer priming
(Hansen et al. 2010;38(12):e131. Li et al. Genome Biology 2010;11(5):R50)
- Bias during library amplification (Kozarewa et al. 2009;6(4):291-5)
 - Over-amplification of GC-rich regions
 - Generation of duplicate sequences
- Read coverage bias (Jiang et al. Genome Research 2011;21(9):1543-51)
 - Unevenness in read coverage along transcripts



- Variability in RNA-seq data (Marioni et al. Genome Research 2008;18(9):1509-17. Bullard et al. BMC Bioinformatics 2010;11:94)
 - Biological condition >> library preparation > run > lane

RNA-seq bias / sources of variability

- Transcript abundance
 - Low abundance transcripts more affected by sampling error : more bias in the estimation of their expression level
 - Highly dependant on the sequencing depth :
 - A question of cost, not due to the technique
- Transcript length (Oshlack et al. Biology Direct 2009;4:14)
 - The ability to call differentially expressed genes between samples is associated with the length of the transcript :
 - more statistical power to detect differential expression for long transcripts compared to short ones
- Mappability bias
 - Uniquely mapping reads are typically summarized over genomic regions → regions with lower sequence complexity will tend to end up with lower sequence coverage
 - Reads corresponding to longer transcripts have a higher mappability