RNA sequencing : library preparation and experimental design

Céline Keime
keime@igbmc.fr

RNA sequencing

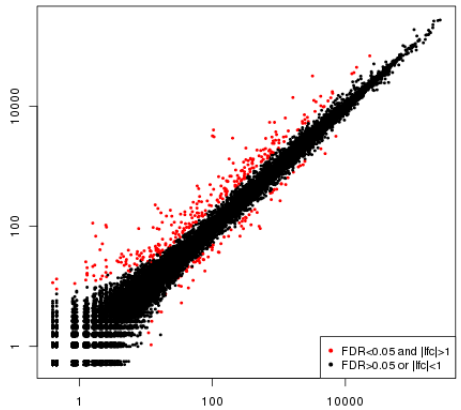
- Introduction : methods for transcriptome analysis
- Preparation of RNA-seq libraries
- Design of RNA-seq experiments
- RNA-seq bias already identified

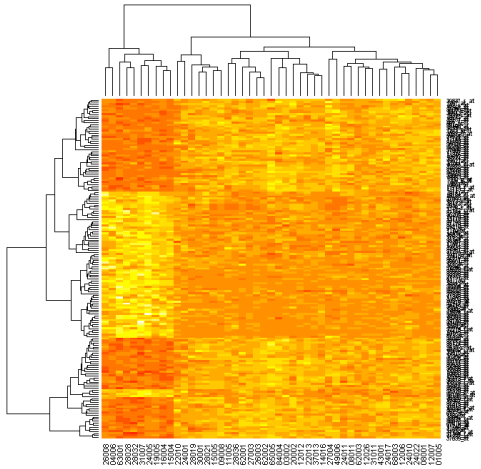
RNA sequencing

- Introduction : methods for transcriptome analysis
- 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



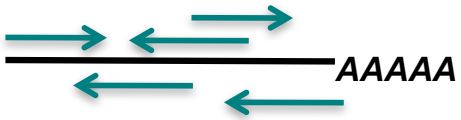


- Qualitative
  - Catalogue all different transcripts (mRNA, ncRNA)
  - Determine the structure of these transcripts
    - TSS, 3'end, splicing patterns, post-transcriptional modifications

Transcriptome analysis : different technologies

- Hybridization-based approaches
  - Microarrays

- Sequence-based approaches
  - EST



    - Generally not quantitative (normalized libraries)
    - Sequenced with the Sanger technique : low throughput
  - SAGE / MPSS / DGE

    - Only a portion of the transcript is analysed : isoforms are generally indistinguishable from each other
  - And now : RNA sequencing

Microarray vs RNA sequencing

	Microarray	RNA sequencing
Quantification of gene expression	Yes	Yes
Reliance upon existing knowledge on transcriptome	Yes	Not necessary ➔ De novo transcriptome assembly, new alternative splicing, RNA start site mapping, RNA editing, transcripts fusion, transcribed SNP identification
Range of detection	Limited Background, saturation	No theoretical limitation
Comparison of different experiments	Not always easy Especially between different labs	Easy
Price (e.g. IGBMC sequencing platform prices per sample for external users)	Affymetrix[1] : 427.7 €	Illumina[2] : 288 €

[1] Affymetrix Gene Array (target labelling + array + hybridization)
[2] Preparation of stranded polyA+ library, Hiseq4000 1x50 sequencing (8 samples per line, i.e.>30M sequences/sample)

Microarray vs RNA sequencing

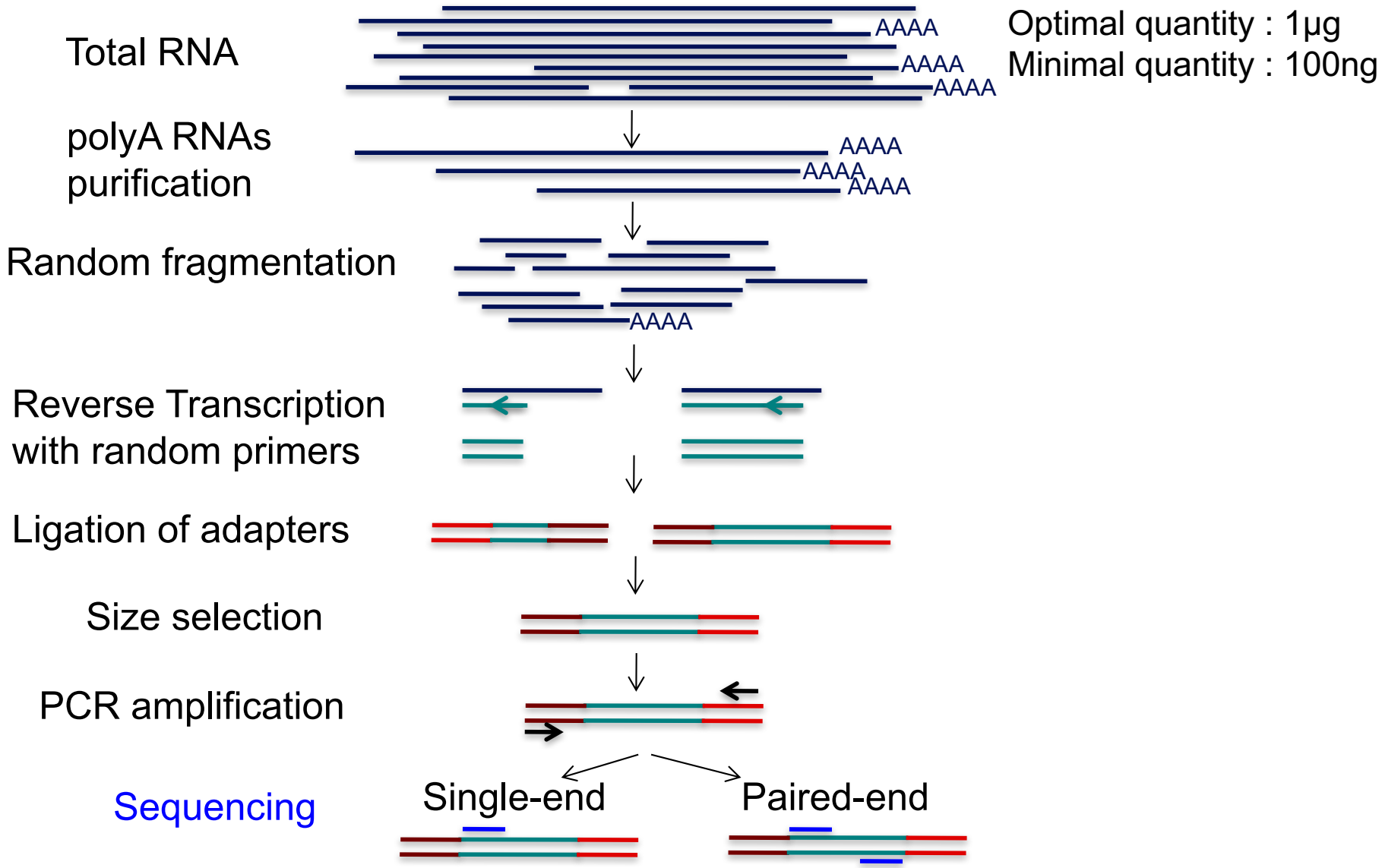
- Marioni et al. (Genome Research 2008;18(9):1509-17)
  - Affymetrix arrays vs Illumina sequencing
  - Main conclusions
    - Good correlation between read counts and normalized array intensities
    - Greater correlation for highly expressed genes
    - When there is a discrepancy between the 2 methods
      - Generally large array intensities and small sequence counts
      - qPCR results agreed more closely with Illumina than with microarray
      - Seems to be mainly due to cross-hybridization on the arrays
    - More information on Illumina dataset
      - e.g. alternative splice variants
- Similar conclusions
  - Mortazavi et al. Nature methods 2008
  - Nagalakshmi et al. Science 2008
  - Toung et al. Genome Research 2011

RNA sequencing

- Introduction : methods for transcriptome analysis
- Preparation of RNA-seq libraries
- Design of RNA-seq experiments
- RNA-seq bias already identified

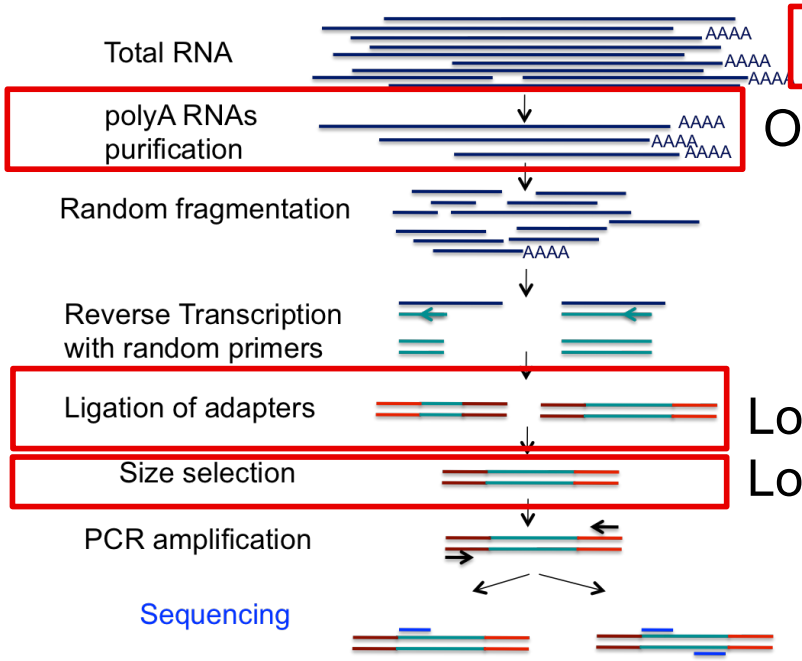
Standard RNA-seq protocol

Illumina TruSeq RNA SamplePrep



Standard RNA-seq protocol

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



Optimal quantity : 1µg
Minimal quantity : 100ng

Quantity of starting material

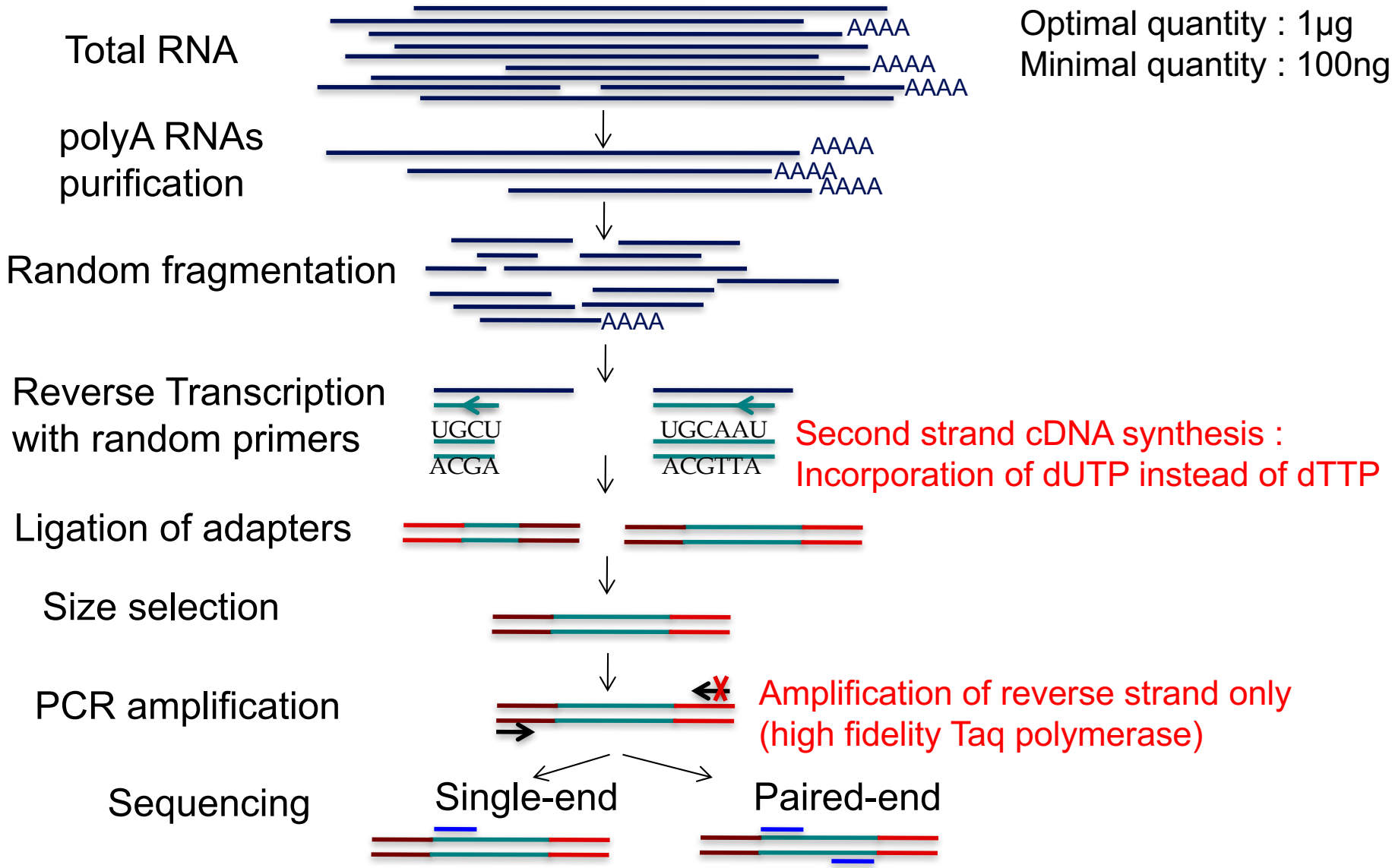
Only polyadenylated RNA are studied

Loss of strand information

Loss of small RNAs (~ < 100 bp)

Directional RNA-seq protocol

Illumina Directional mRNA-Seq SamplePrep

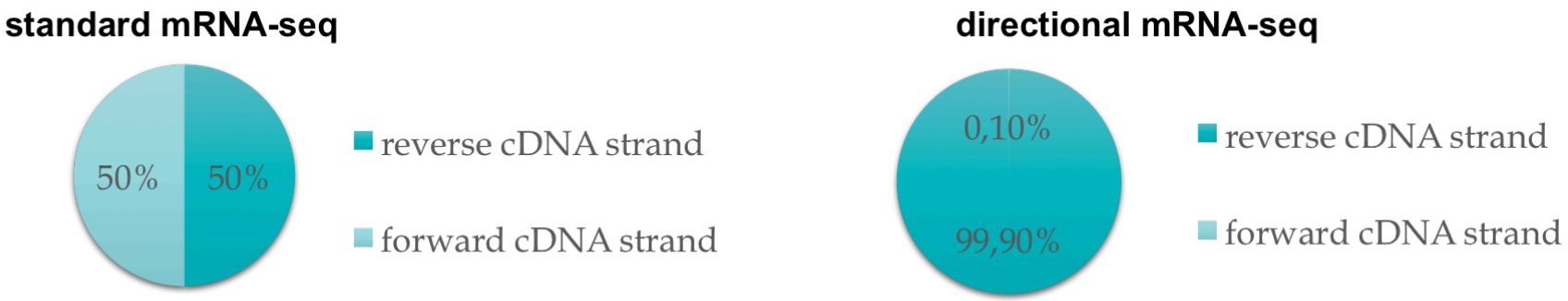


Directional RNA-seq

- Good quality of strand-specificity

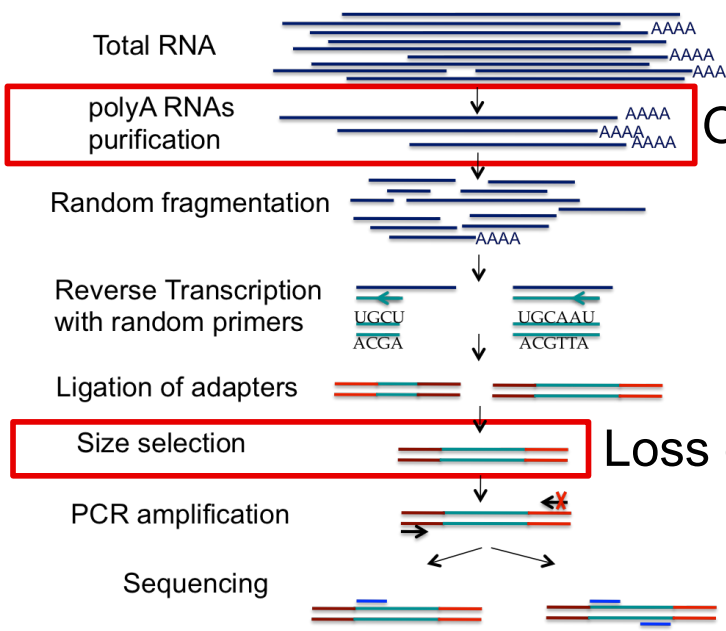
  e.g. Results obtained on spike-in RNAs added in 4 libraries prepared with both standard and directional RNA-seq protocols *(GenomEast Platform)*

  Proportion of reads from each cDNA strand :



Directional RNA-seq

- Advantages
  - Same advantages as standard RNA-seq
  - Preserves the strand information
    - ➔ Allows to determine transcript orientation
    - ➔ Important for novel transcript discovery and annotation, especially for overlapping transcripts
- Drawbacks



Optimal quantity : 1µg
Minimal quantity : 100ng

Quantity of starting material

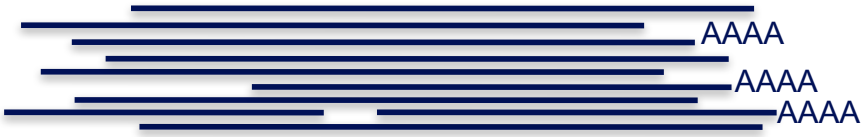
Only polyadenylated RNA are studied

Loss of small RNAs

Total and directional RNA-seq

Illumina Truseq Stranded Total RNA SamplePrep

Total RNA



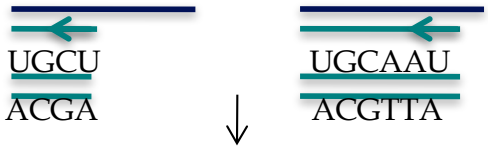
Optimal quantity : 1µg
Minimal quantity : 100 ng

rRNA depletion
(RiboZero)

Several sets of RiboZero probes are available depending on the species
www.epibio.com/rnamatchmaker

Random fragmentation

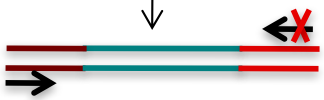
Reverse Transcription
with random primers
Incorporation of dUTP



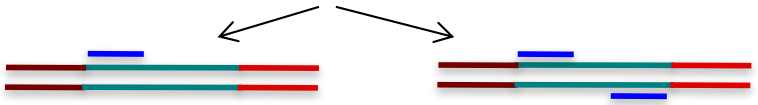
Ligation of adapters

Size selection

PCR amplification
of reverse strand



Sequencing



Total and directional RNA-seq

- Advantages
  - Same advantages as directional RNA-seq
  - Allows to study non-polyadenylated transcripts

- Drawbacks
  - Quantity of starting material
  - Loss of small RNAs
  - 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

Low quantity RNA-seq

NuGEN Ovation RNA-seq system

Total RNA

Optimal quantity : 10 ng
Minimal quantity : 500 pg

Primer annealing
(specific to non-rRNA sequences)

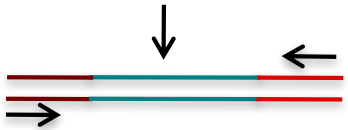
Reverse transcription

Ribo-SPIA amplification

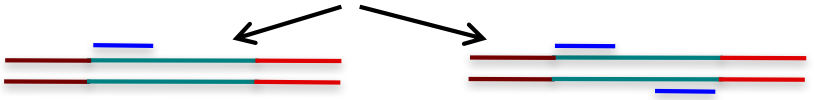
cDNA fragmentation

Ligation of adapters

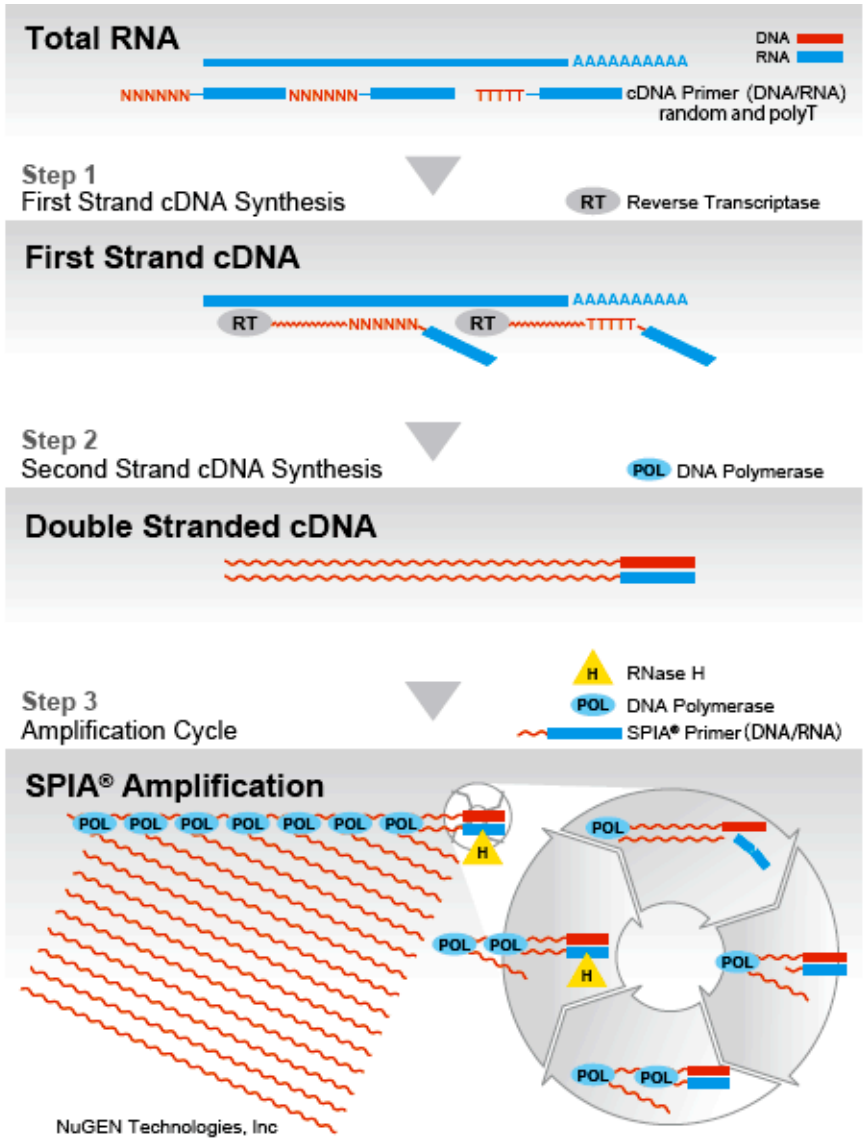
PCR amplification



Sequencing



Ribo-SPIA amplification (NuGEN)



Generation of 1st strand cDNA with a unique RNA sequence at 5' end

Generation of a double-stranded cDNA with a unique DNA/RNA heteroduplex at one end

Degradation of RNA in the DNA/RNA heteroduplex by RNaseH

➔ expose the DNA for binding a second SPIA DNA/RNA chimeric primer

Step repeated several times

➔ linear amplification of the target

Low quantity RNA-seq

- Advantages
  - Low quantity of starting material
  - Study of whole transcriptome
- Drawbacks
  - Bias due to ribo-SPIA primers

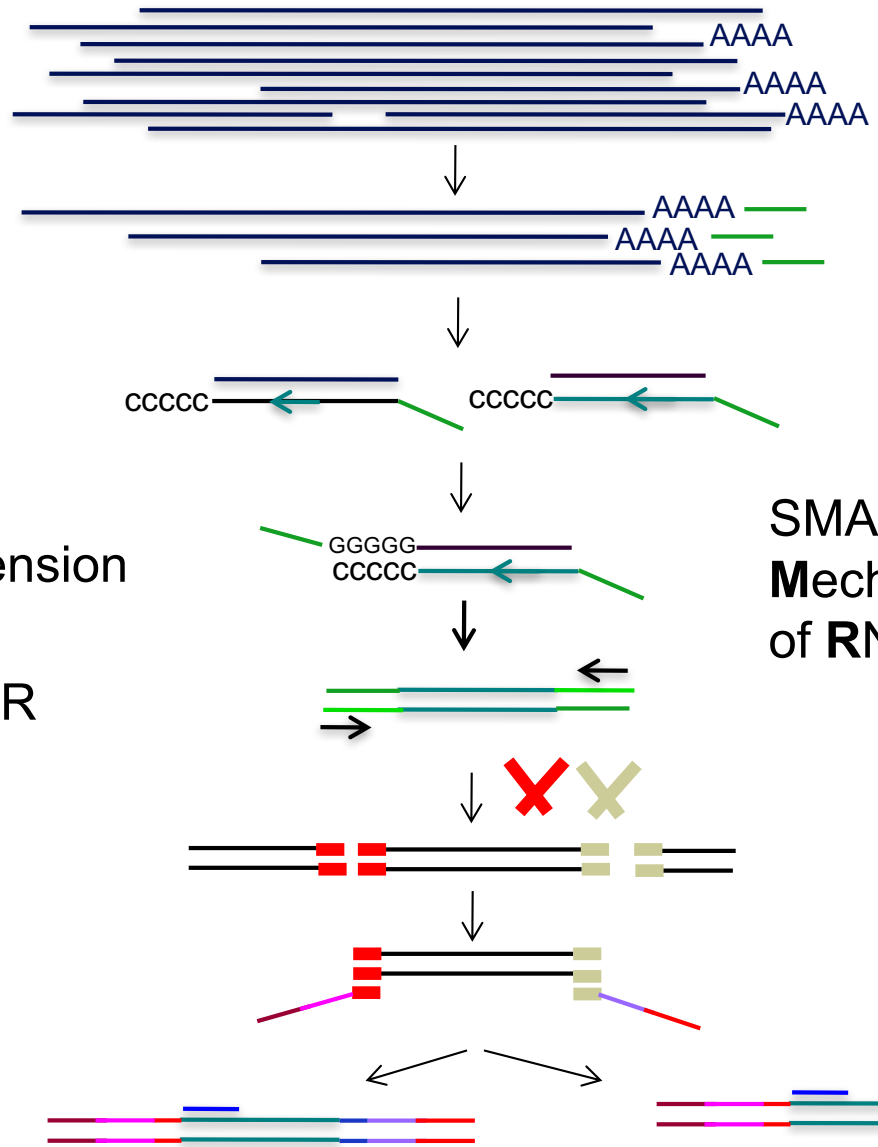
    → some transcripts are not recovered
  - Variable efficiency of rRNA removal
  - Higher number of RNA molecules sequenced compared to standard RNA-seq

    → more reads needed to achieve the same coverage on polyadenylated RNA

Single-cell RNA-seq

Clontech SMART-Seq Ultra Low Input RNA

Quantity : one cell

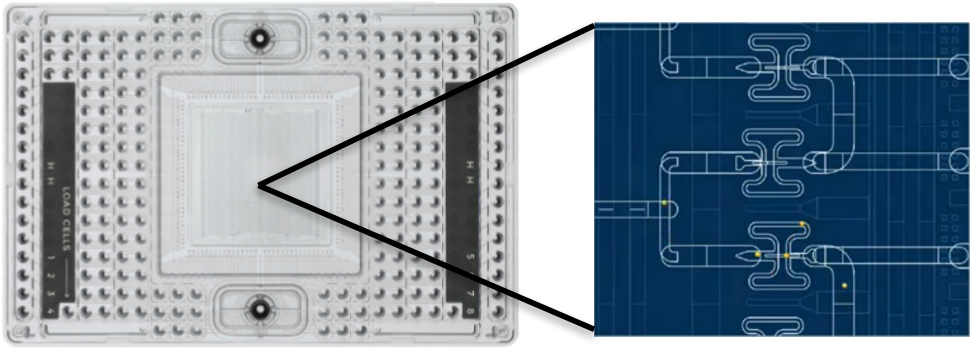


SMART = **S**witching **M**echanism **a**t 5' End of **R**NA **T**emplate

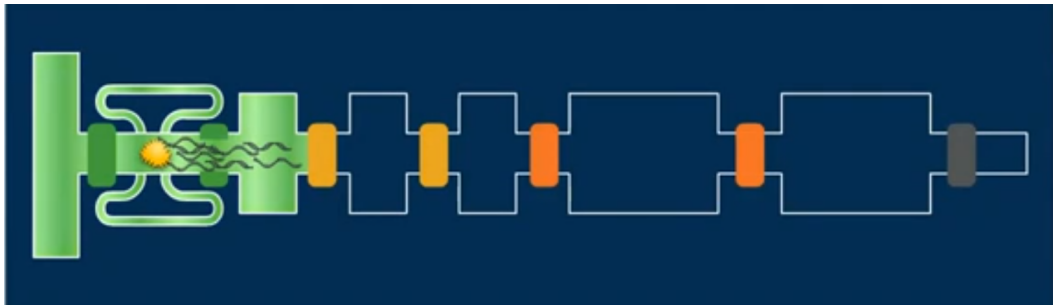
Single-cell RNA-seq
cDNA preparation using Fluidigm C1

- C1 single cell auto-prep system (Fluidigm)

  - Load ~ 1000 cells
  - Capture individual cells



  - Cell lysis
  - cDNA synthesis
  - Pre-amplification
  - Harvest the cDNA of each single cell



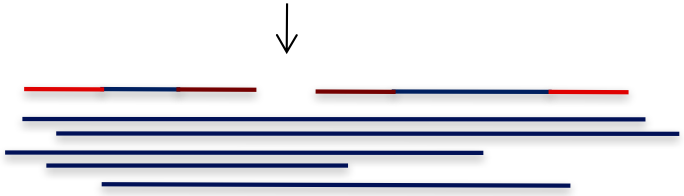
Small RNA-seq

Illumina Truseq smallRNA SamplePrep

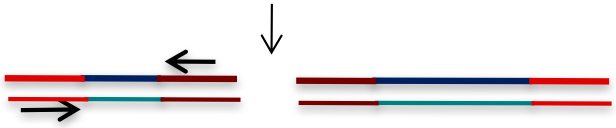
Optimal quantity : 2µg
Minimal quantity : 1µg

Total RNA

Ligation of adapters

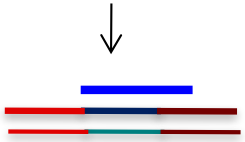


RT 1st strand synthesis
PCR amplification



Adapted size selection

Sequencing



Summary of RNA-seq library preparation methods

Library preparation	Total RNA quantity		Type of studied RNA	Stranded
	Minimal	Optimal		
Standard RNA-seq	100 ng	1 µg	Only polyA+ RNA of size > 100 bp	No
Directional RNA-seq	100 ng	1 µg	Only polyA+ RNA of size > 100 bp	Yes
Total and directional RNA-seq	250 ng	1 µg	All RNA of size > 100 bp	Yes
Low quantity RNA-seq	500 pg	10 ng	All RNA of size > 100 bp	No
Single cell RNA-seq	1 cell	1 cell	Only polyA+ RNA of size > 100 bp	No
Small RNA-seq	1 µg	2 µg	small RNAs (desired size can be chosen)	Yes

RNA sequencing

- Introduction : methods for transcriptome analysis
- 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

Which protocol for which application ?

- Choice depend on
  - Quantity of starting material
  - Type of RNA studied (small/long, polyA+/-)
  - Biological questions of interest
    - Expression quantification on annotated transcripts
      - polyA+ : standard or directional protocol, single-end sequencing
      - polyA+/- : total directional protocol, single-end sequencing
    - Alternative splicing analysis, fusion transcripts detection, mapping over repetitive regions
      - (total) directional protocol, paired-end sequencing
    - New transcripts identification
      - Total and directional protocol, paired-end sequencing
- Compare data from the same protocol

Power analysis

- If similar data are available either from a pilot study or in public repositories
- Scotty
  - http://bioinformatics.bc.edu/marthlab/scotty/scotty.php
  - Busby et al. Bioinformatics 2013;29(5):656-7
  - Allows to design an experiment with an appropriate sample size and read depth to satisfy the user-defined experimental objectives
- Otherwise use the following guidelines…

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
- ENCODE guidelines (mammalian tissues)
  - Gene expression comparison between different polyA+ samples: ~ 30 million paired-end reads of length > 30 nt
  - Discovery of novel transcripts, experiments where the sensitivity of detection is important, precise quantification of transcripts isoforms ➔ higher sequencing depth needed

    > 100-200 million paired-end reads of length > 75 nt
  - https://www.encodeproject.org/about/experiment-guidelines/

How many replicates are needed ?

- Low technical variability
  and technical variability << biological variability
  (Marioni et al. Genome Research 2008;18(9):1509-17. Bullard et al. BMC Bioinformatics 2010;11:94)
  ➔ Technical replicates not required
- But “sequencing technology does not eliminate biological variability”
  (Hansen et al. Nat Biotechnol. 2011;29(7):572-3)
  - **Biological replicates are fundamental !**
  - How many ?
    - Highly dependant on the correlation between replicates

RNA sequencing

- Introduction : methods for transcriptome analysis
- Preparation of RNA-seq libraries
- Design of RNA-seq experiments
- RNA-seq bias already identified

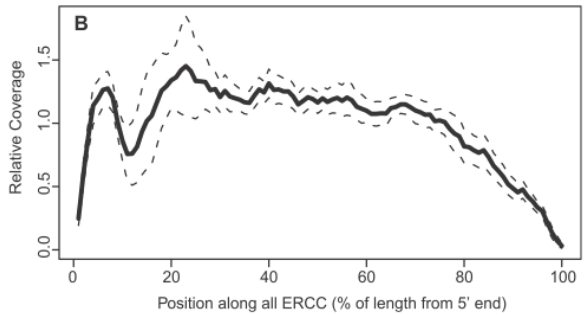
RNA-seq bias / sources of variability

- 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

Information and updates on RNA-seq

- RNA-Seq blog
  - http://www.rna-seqblog.com/