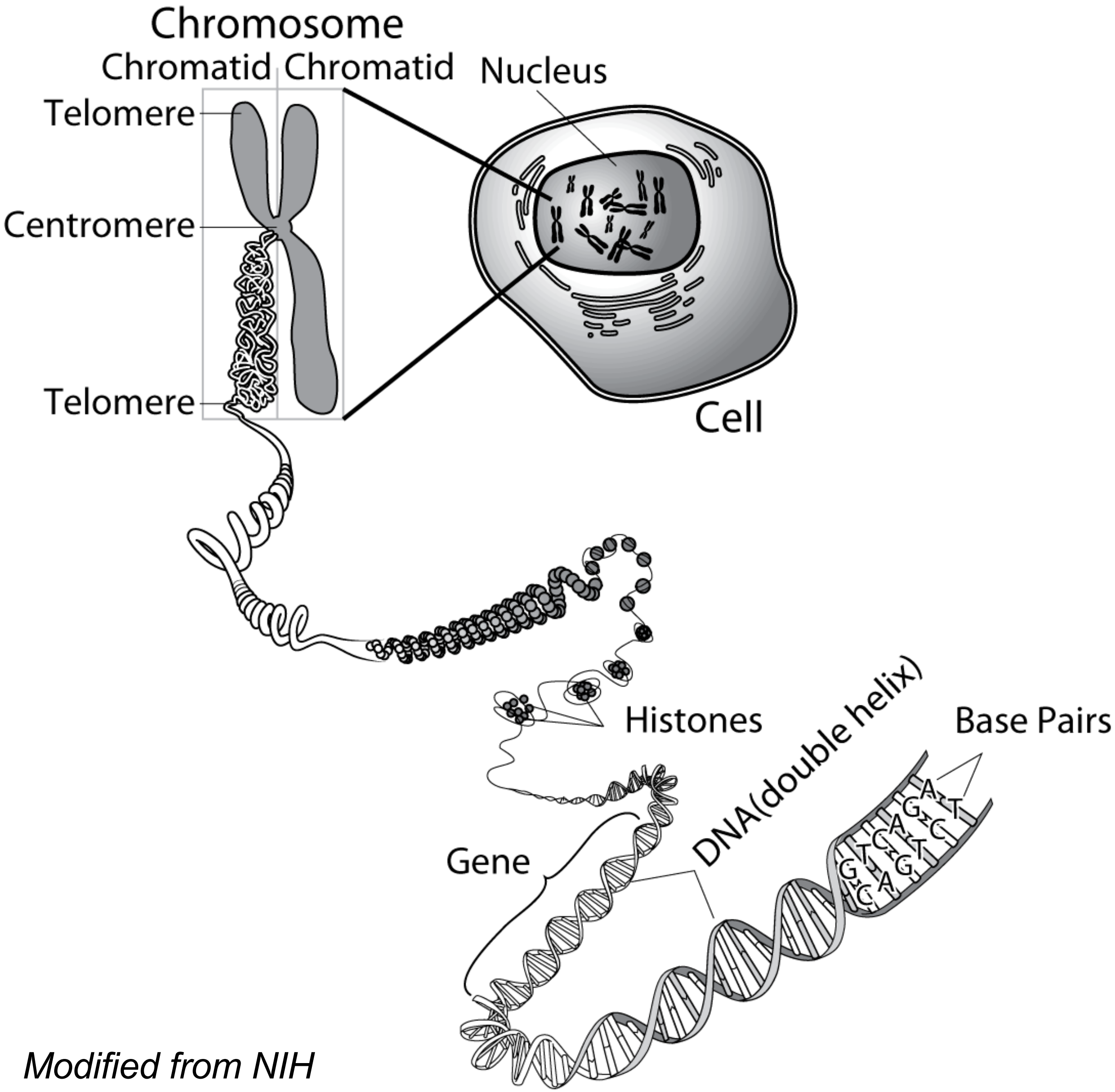
NTNU

Innovation and Creativity

Project – Preprocessor for high throughput sequencing reads

Pål Sætrom

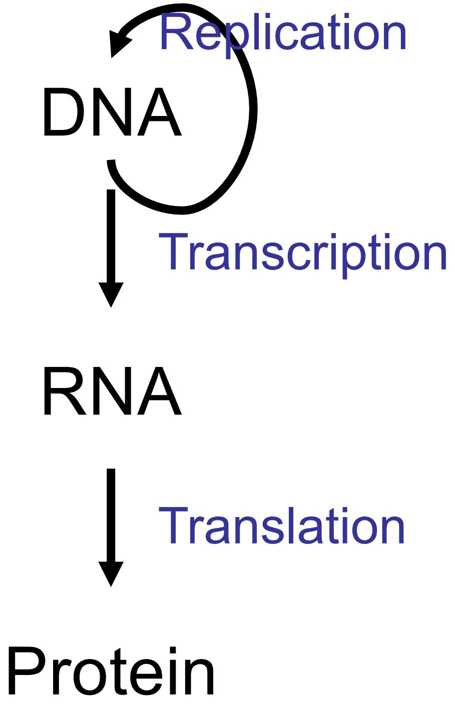
Sequences – basic data structures in cells



Modified from NIH

http://www.accessexcellence.org/RC/VL/GG/chromosome.php

Sequence data



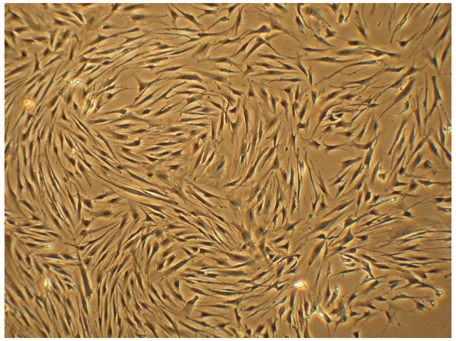
High throughput sequencing – reading the cell’s RNA/DNA

Procedure

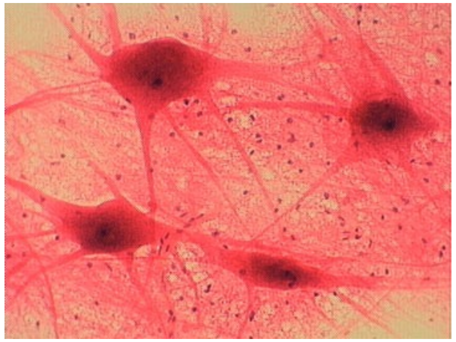
1. Isolate RNA/DNA
2. Prepare sequencing library
3. Sequence
4. Analyze data

1. Isolate RNA/DNA

Connective tissue



Brain



Muscle

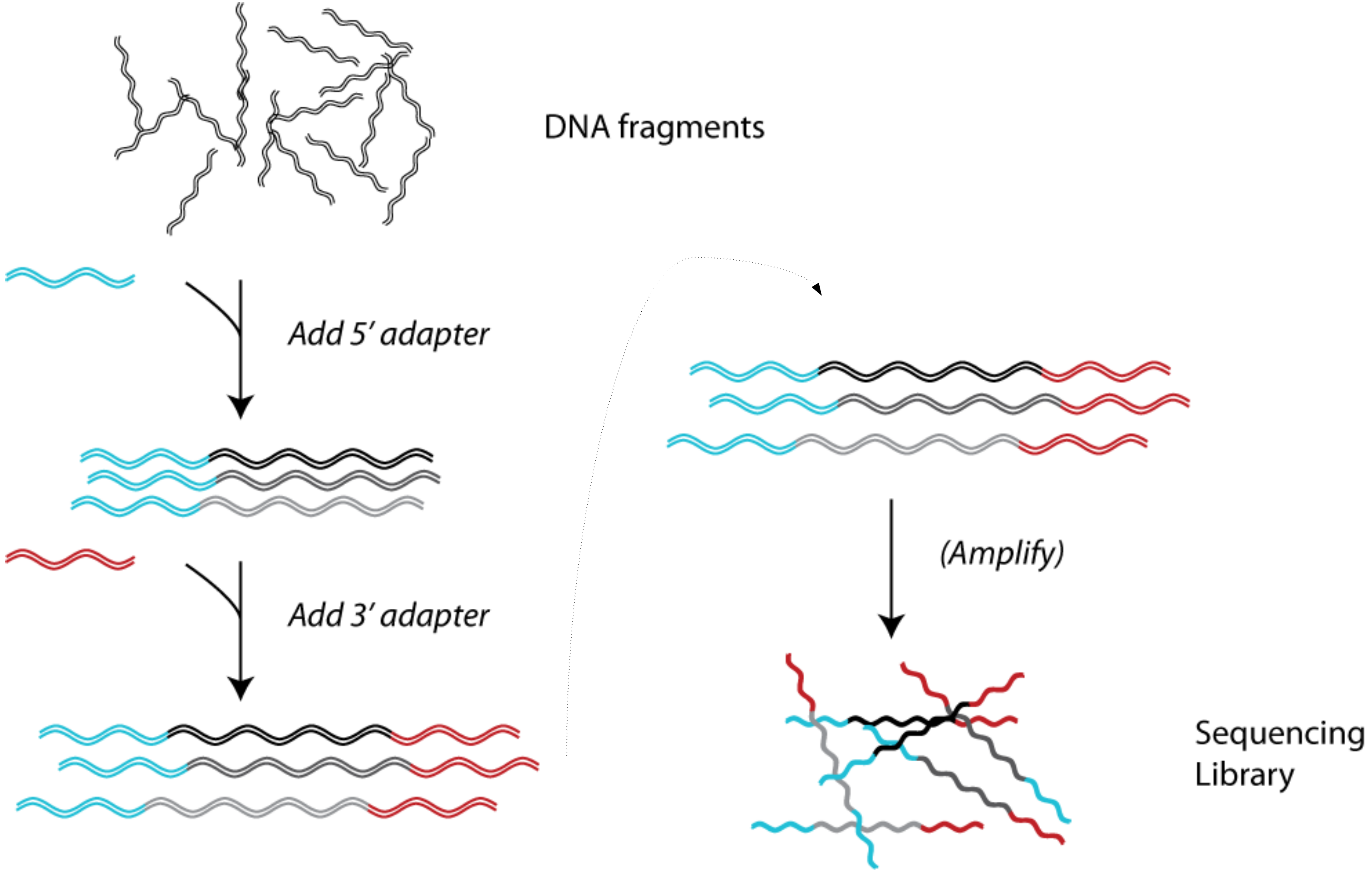
1. Tissue sample
2. “Break” cells (liq. N, blender, chemicals)
3. Chemical reactions to isolate RNA or DNA

→ RNA/DNA sample

Simple DNA: Salt, soap, alcohol

Pictures:
Photo Researchers, Inc., Iowa State Univ,.
Stephanie Saade, USA Today

2. Prepare sequencing library



DNA fragments

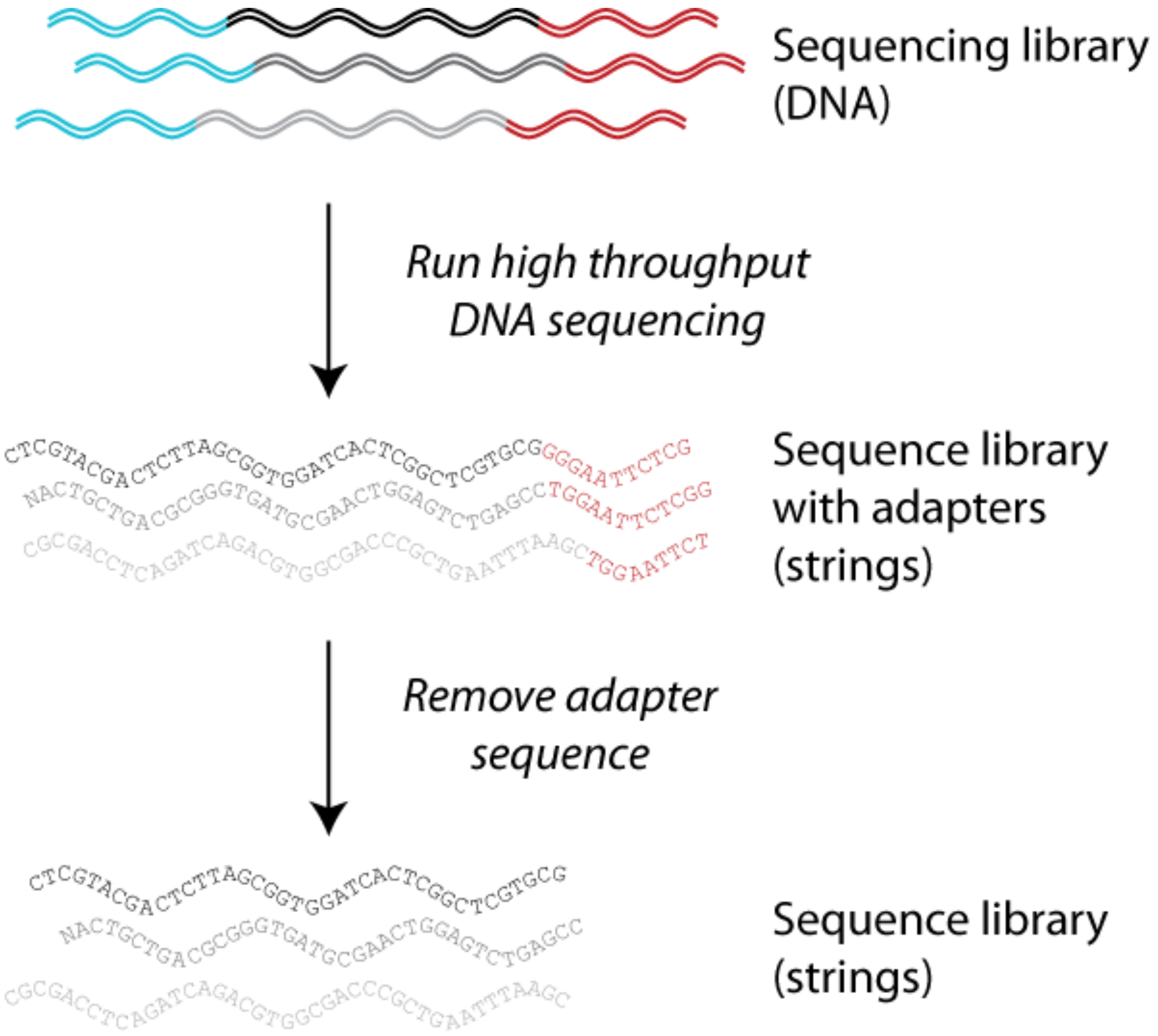
Add 5′ adapter

Add 3′ adapter

(Amplify)

Sequencing Library

3. + 4. Sequence and analyze data


CTCGTACGACTCTTAGCGGTGGATCACTCGGCTCGTGCGGGGAATTCTCG
NACTGCTGACGCGGGTGATGCGAACTGGAGTCTGAGCCTGGAATTCTCGG
CGCGACCTCAGATCAGACGTGGCGACCCGCTGAATTTAAGCTGGAATTCT
CTCGTACGACTCTTAGCGGTGGATCACTCGGCTCGTGCG
NACTGCTGACGCGGGTGATGCGAACTGGAGTCTGAGCC
CGCGACCTCAGATCAGACGTGGCGACCCGCTGAATTTAAGC

Sequencing library (DNA)

Run high throughput DNA sequencing

Sequence library with adapters (strings)

Remove adapter sequence

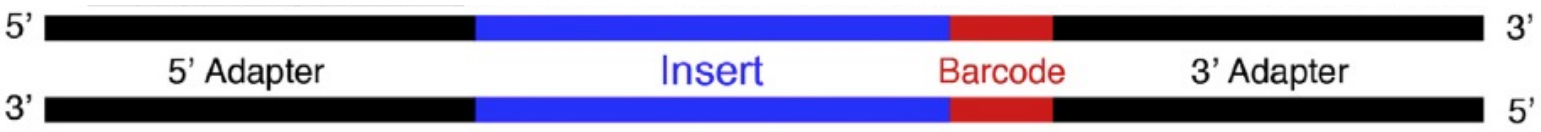
Sequence library (strings)

Barcode sequencing - unique adapter per sample

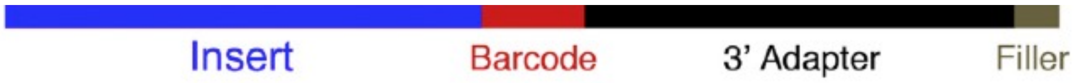
- Sequencing reaction produces lots of data
  - ~ 300 * 10^6 sequences (reads)
  - Cost: ~ NOK 8000-16000
  - Default: single sample per reaction
- Some applications (RNA-seq.) require less data per sample
  - Small RNAs: ~ 10 * 10^6 reads sufficient
- Using unique adapter per sample
  - Allows multiplexing multiple samples
  - “Barcode” read during sequencing

Barcode sequencing – Resulting data

Sequencing library:



Sequencing read:



Read without adapter and barcode:

Project

- Task 1 – Perfectly matching adapter fragments
- Task 2 – Imperfectly matching adapter fragments
- Task 3 – Sequencing errors and error distributions
- Task 4 – Finding the adapter sequence
- Task 5 – De-multiplex barcoded library

- Individually or in pairs

- Deliverable 1: Project report
- Deliverable 2: Oral presentation

Project report

- Parts: Introduction, Methods, Results and Discussion, References
- Figures and Tables to present results
- Pseudo code to describe algorithms
- Follow standard for scientific reports
  - Clear, consistent, unambiguous presentation
  - Consistent (standard) formatting

Deadline: October 31, 23:59.