RNA-Sequencing

Nicolas Balcazar Corinna Blasse An Duc Dang
Hannes Hauswedell Sebastian Thieme

Advanced Algorithms for Bioinformatics (P4)
K. Reinert and S. Andreotti

SoSe 2010
FU Berlin

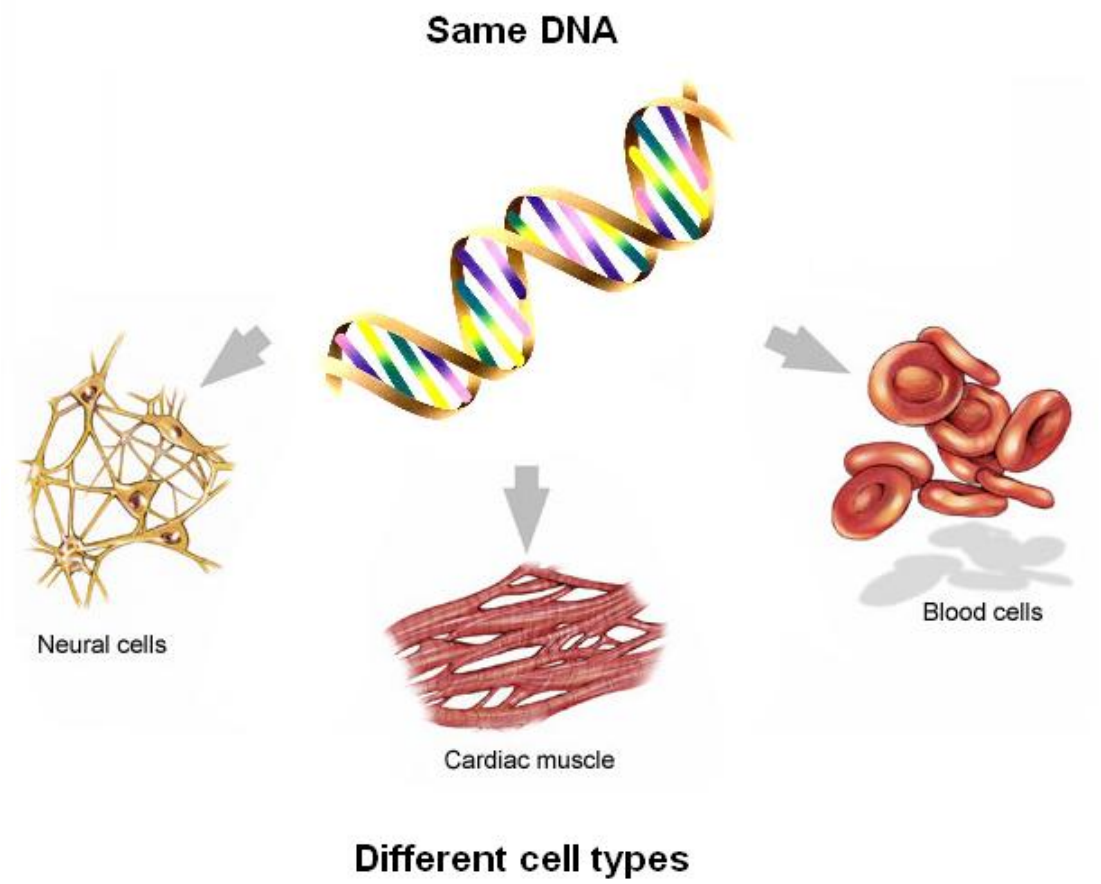
14. April 2010

Outline

Outline

Motivation

- mRNAs specify cells' identity

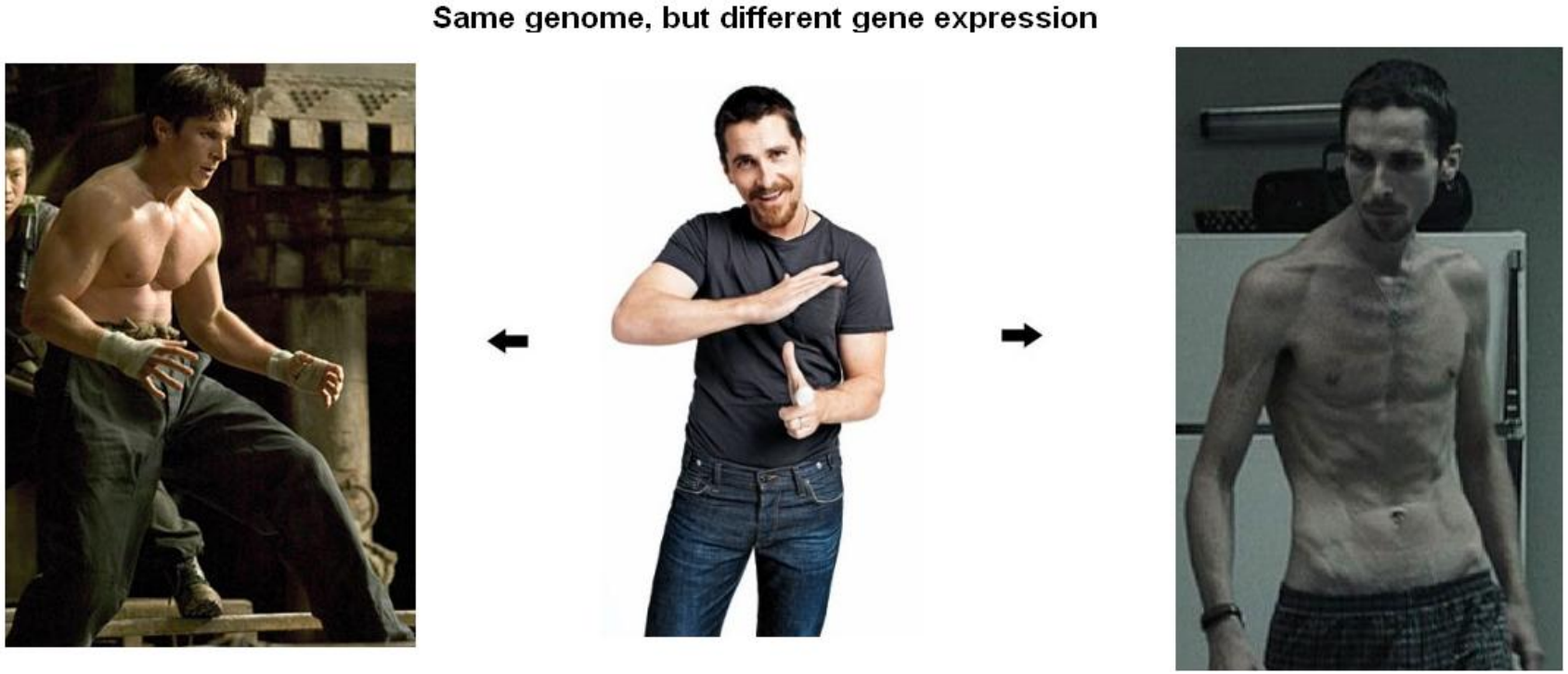


Motivation

- Crucial to understand the processes inside a cell
- DNA cannot give all that information, but transcriptome

Motivation

- mRNAs govern cells' activity due to environmental conditions



Motivation

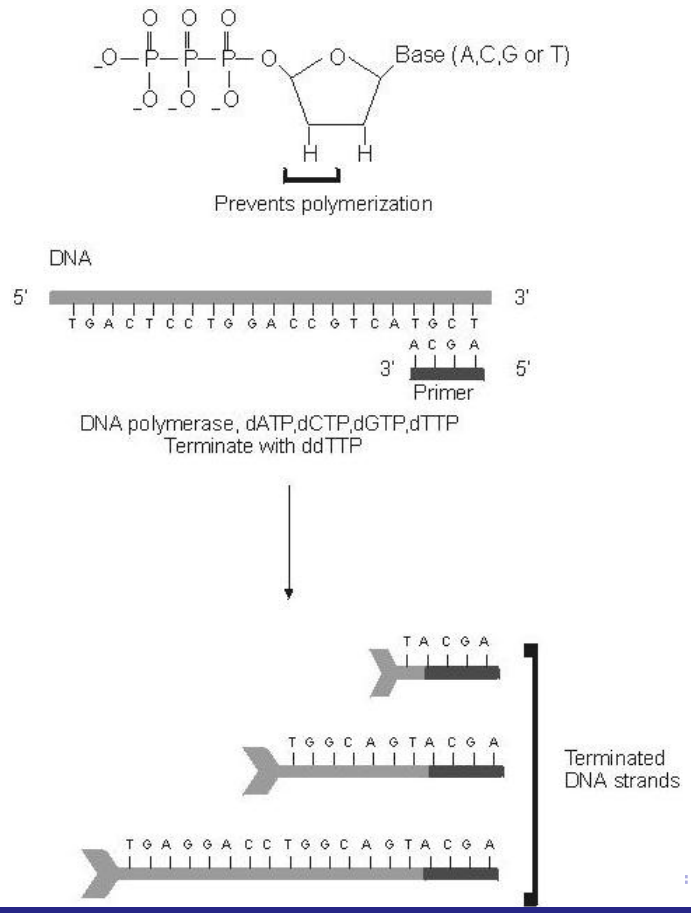
- Crucial to understand the processes inside a cell
- DNA alone doesn't give us all that information, but transcriptome does:
  - mRNAs specify cells' identity
  - mRNAs govern cells' activity due to environmental conditions

For transcriptome analyses RNA sequencing is needed!

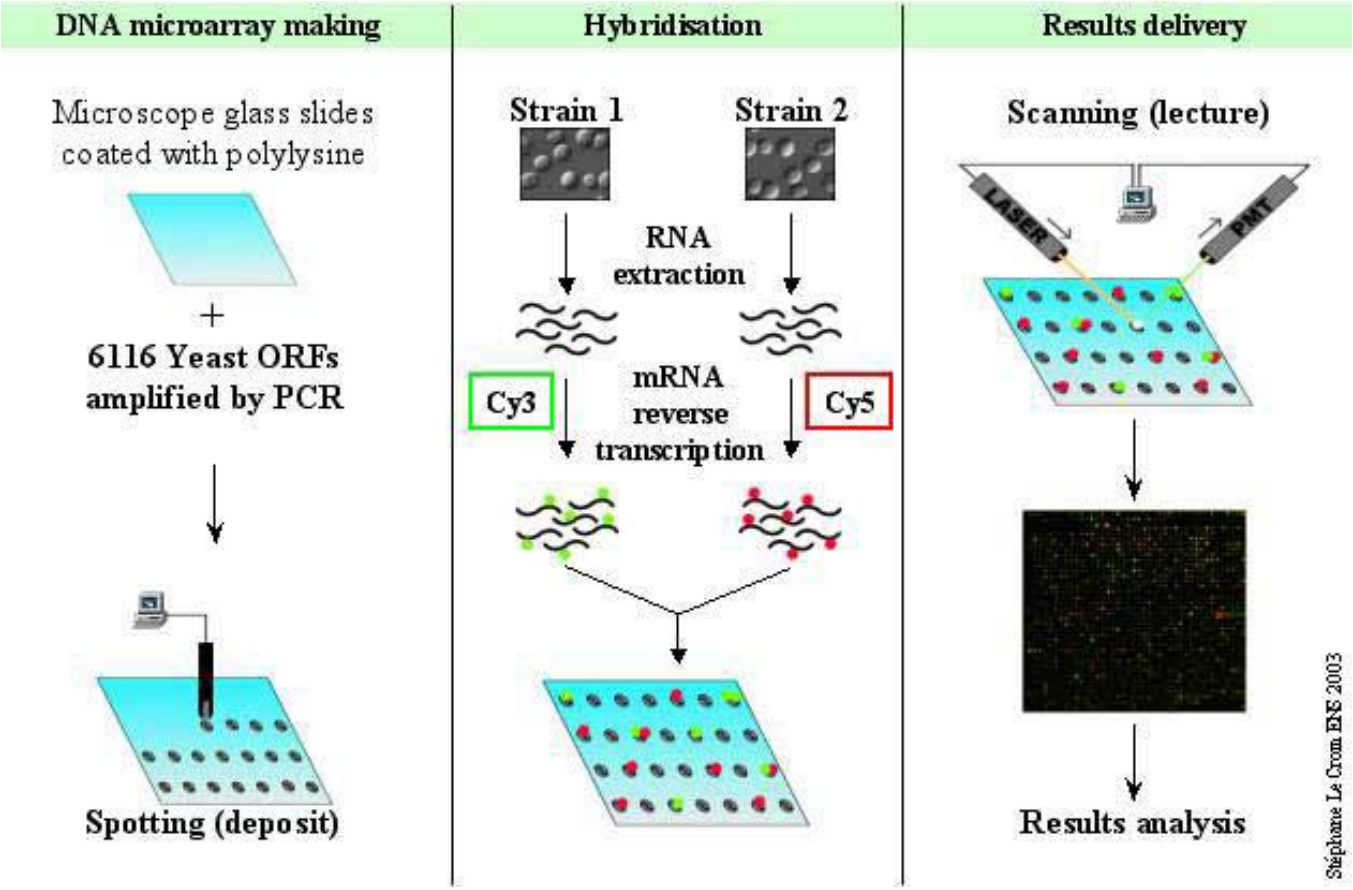
Outline

Sanger sequencing

- Create DNA strands of different lengths by tagged ddNTPs
- Not accurate for long strands
- Cost and time consuming



Microarrays



Microarrays

- (Cross-)hybridzation artifacts
- Dye-based detection issues
- Limited RNA splice detection
- No novel exon or novel gene detection

SAGE

Serial analysis of gene expression

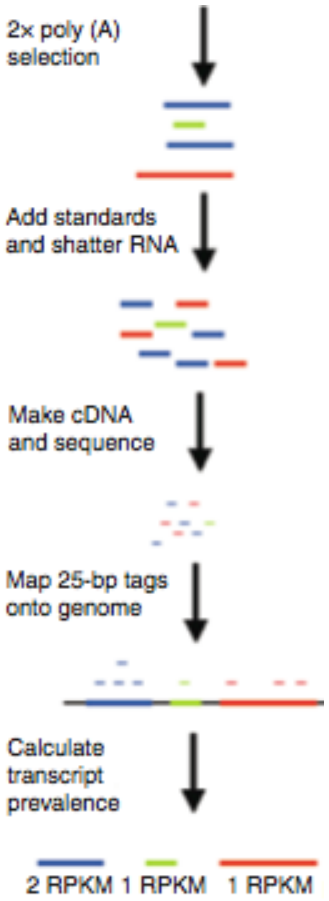
- Uses DNA sequencing of cloned tags 17-25bp from terminal ends
- Maps tags to mRNA database or source genome
- New or low abundant transcripts cannot be found

RNA sequencing

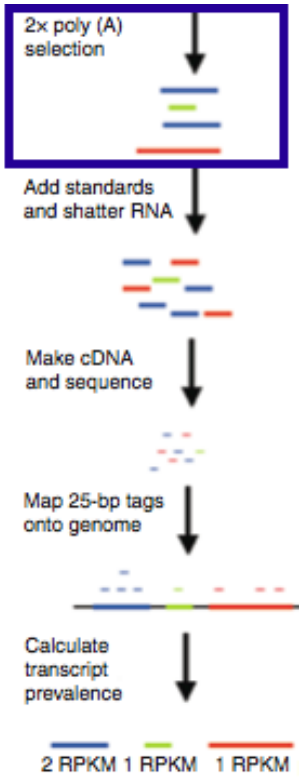
- Ultra-high-throughput sequencing
- No bacterial cloning of cDNA needed (no bacterial cloning constraints)
- Accurate due to short reads
- Can discover new exons, genes
- Can handle RNA splicing

Outline

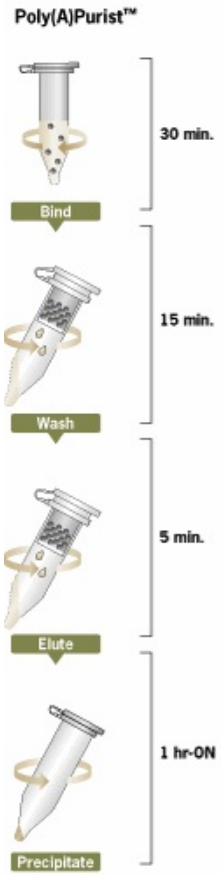
Pipeline - Outline



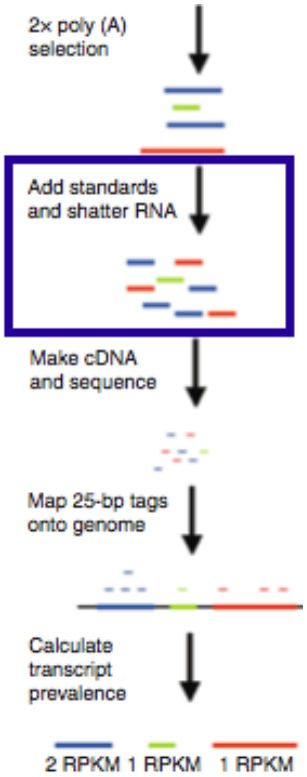
Poly(A)-Selection



- Input: Cell lysate
- Output: RNA

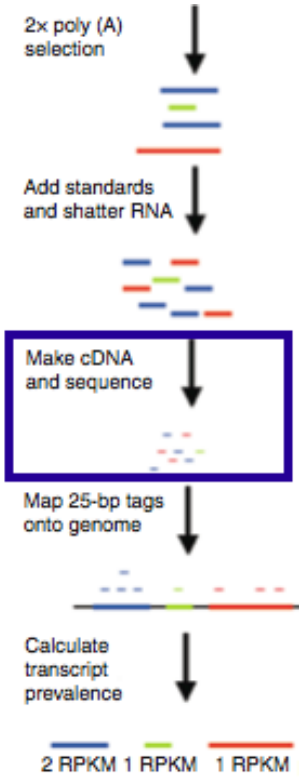


Fragmentation of RNA

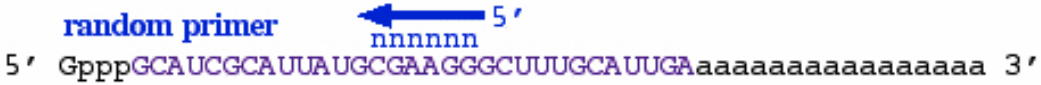


- Important for the uniform sequence coverage
- Why do we shatter before cDNA conversion?
  - Over-representation of 5'-ends of transcripts
  - Occurrence of strongly favored sites

Conversion to cDNA



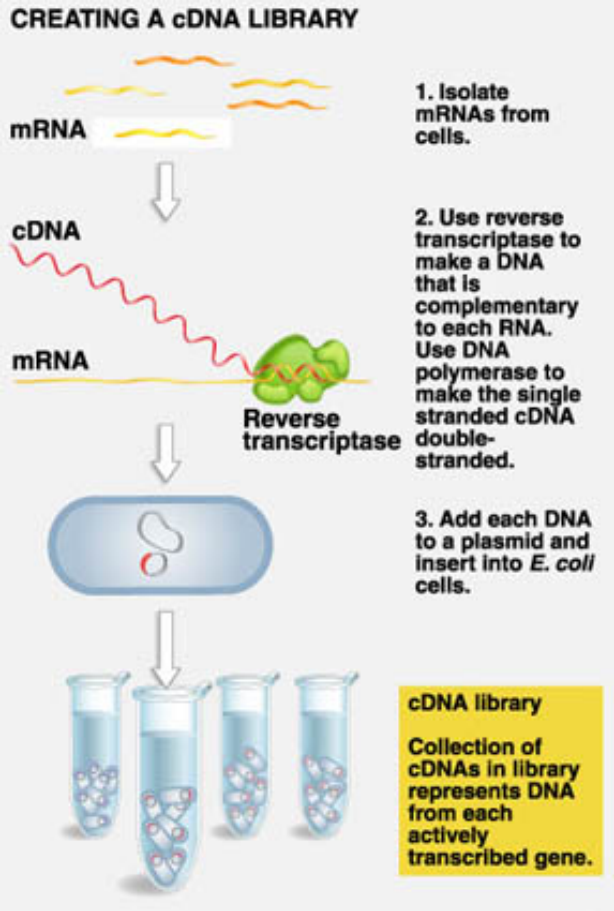
- Conversion by random priming

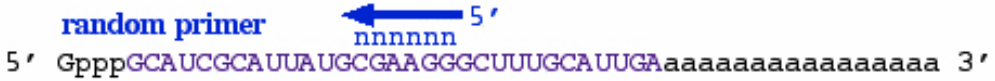


- Why do we use cDNA?
  - RNA: unstable, fast degradation
  - RNA forms secondary and tertiary structures
  - RNA: uridine instead of thymidine

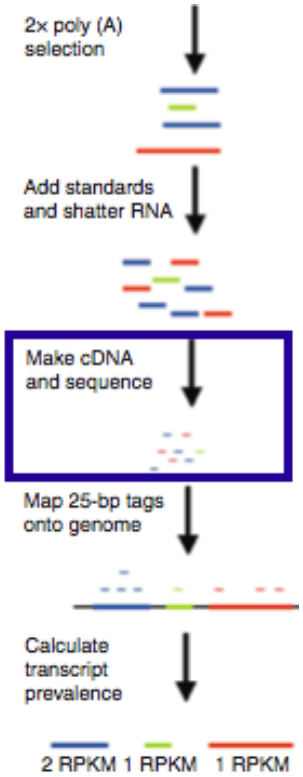
  $\longrightarrow$ More difficult to map

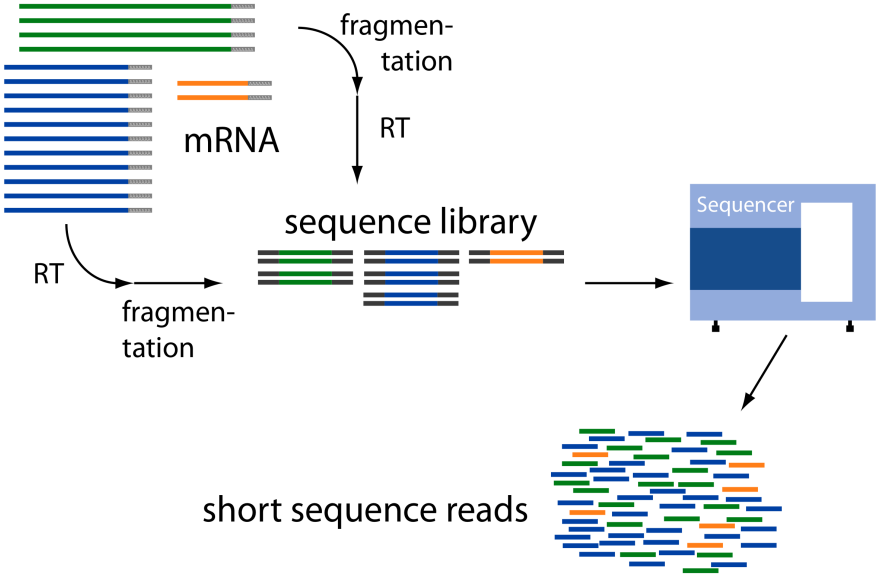
Conversion to cDNA



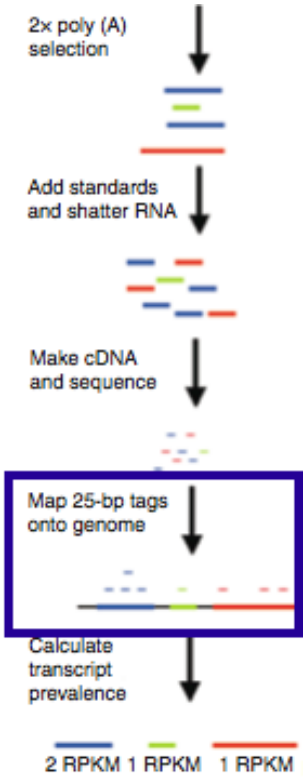


Sequencing

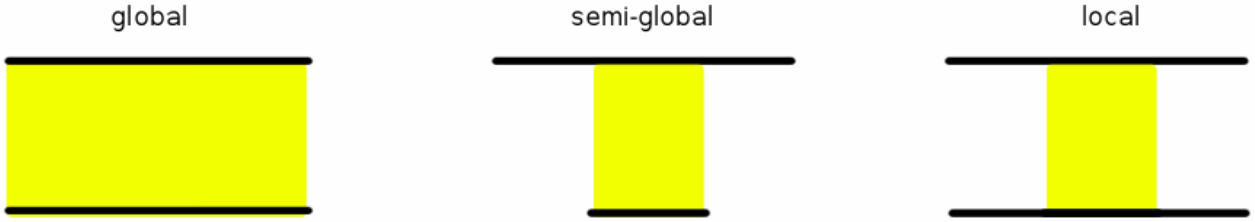

2x poly (A) selection
Add standards and shatter RNA
Make cDNA and sequence
Map 25-bp tags onto genome
Calculate transcript prevalence
2 RPKM
1 RPKM
1 RPKM


fragmen-
tation
RT
mRNA
sequence library
RT
fragmen-
tation
Sequencer
short sequence reads

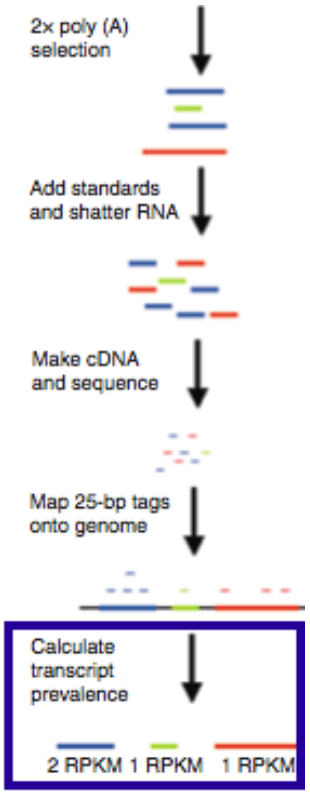
Mapping



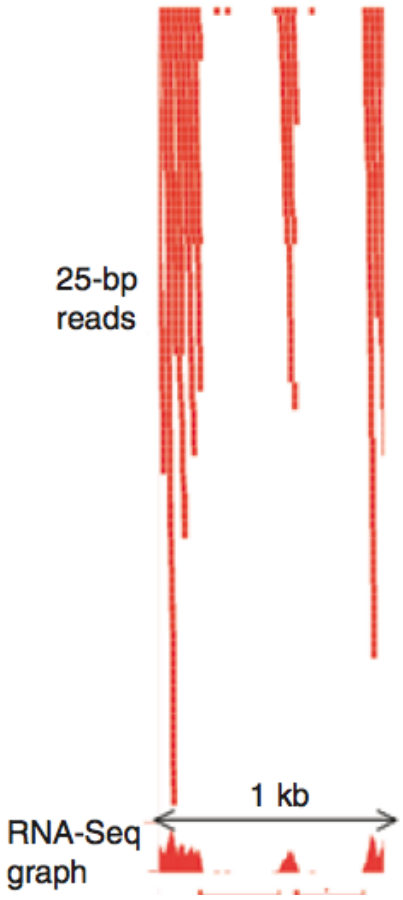
- *Read-mapping* is the semi-global alignment of many (very) short sequences (reads) to a long sequence (the genome)
- Alignment: approximate string matching



Quantification



- Gene coverage
- RPKM
  Reads per kilobase of exon model per million mapped reads



Outline

Overview

Two main phases

1. Filtration
2. Verification

$\Rightarrow$ different (combinations of) algorithms available

Overview

Two main phases

1. Filtration
2. Verification

$\Rightarrow$ different (combinations of) algorithms available

Filtration

For every read:

- identify candidate regions on the genome by an heuristic(!) approach
- discard the rest of genome

Trade-off: Sensitivity vs Speed

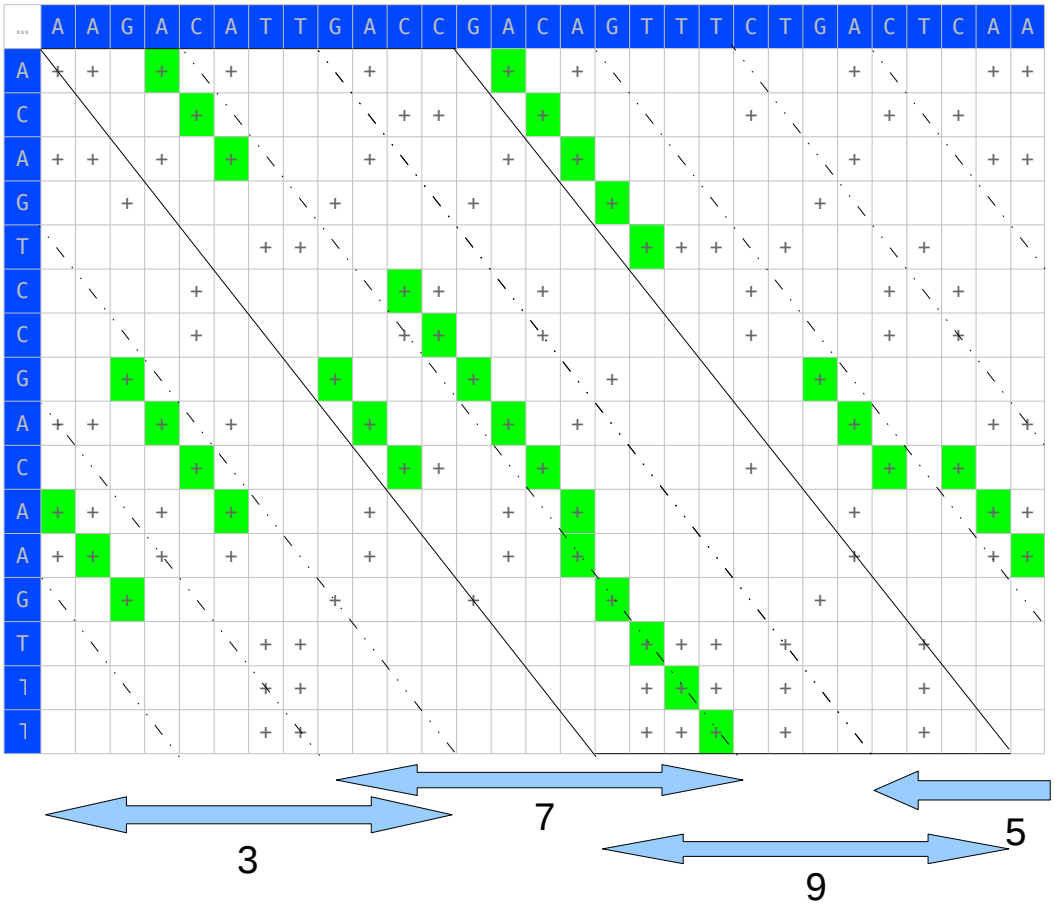
Filtration

For every read:

- identify candidate regions on the genome by an heuristic(!) approach
- discard the rest of genome

Trade-off: Sensitivity vs Speed

Filtration, Q-Gram-Counting



Verification

For every candidate region identified by the filter:

1. calculate alignment with read
2. depending on score, accept or reject

Verification

For every candidate region identified by the filter:

1. calculate alignment with read
2. depending on score, accept or reject

Verification, Alignment Algorithms

- “typical” algorithms can be used
- e.g. just scoring the diagonal with hamming distance (no gaps) [Maq]
- or exact DP-Algorithms like Gotoh (with gaps)
- or heuristic searches like local BLAST [QUASAR] or Banded alignments
- bit-parallel DP-algorithms exist also (Myer’s Bitvector algorithm) [RazerS]

Questions regarding choice of algorithm and implementation:

input how long are the reads? how many are there?
how big is/are the genome(s)?
what technology was used?
known error-profile or quality values?

output gapped or ungapped alignments?
what kind of scoring / distance measurement?

Outline

“Not-so-short” reads

Read lengths in NGS

- used to be 20-40bp
- now industry standard is 70-100bp
- read length increases (already up to few hundreds)

Problems

- not possible to use hamming distance
- chance of spanning an intron increases

“Not-so-short” reads

Read lengths in NGS

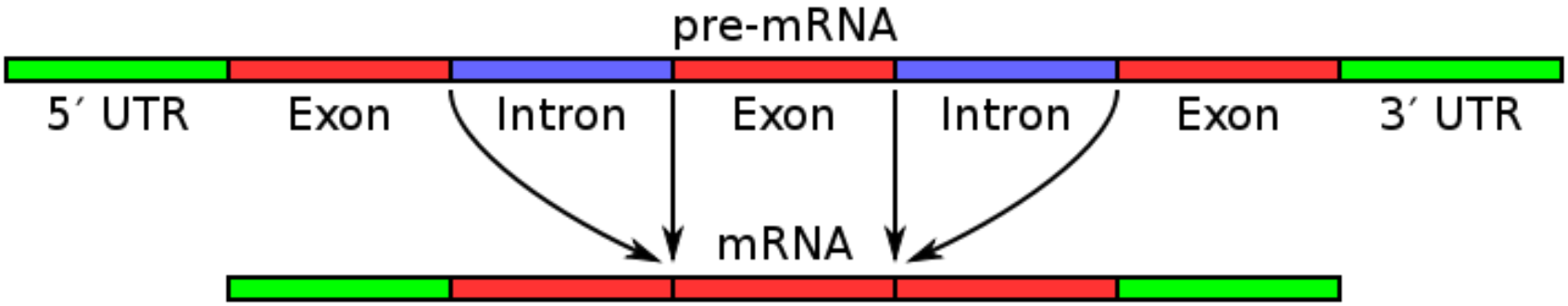
- used to be 20-40bp
- now industry standard is 70-100bp
- read length increases (already up to few hundreds)

Problems

- not possible to use hamming distance
- chance of spanning an intron increases

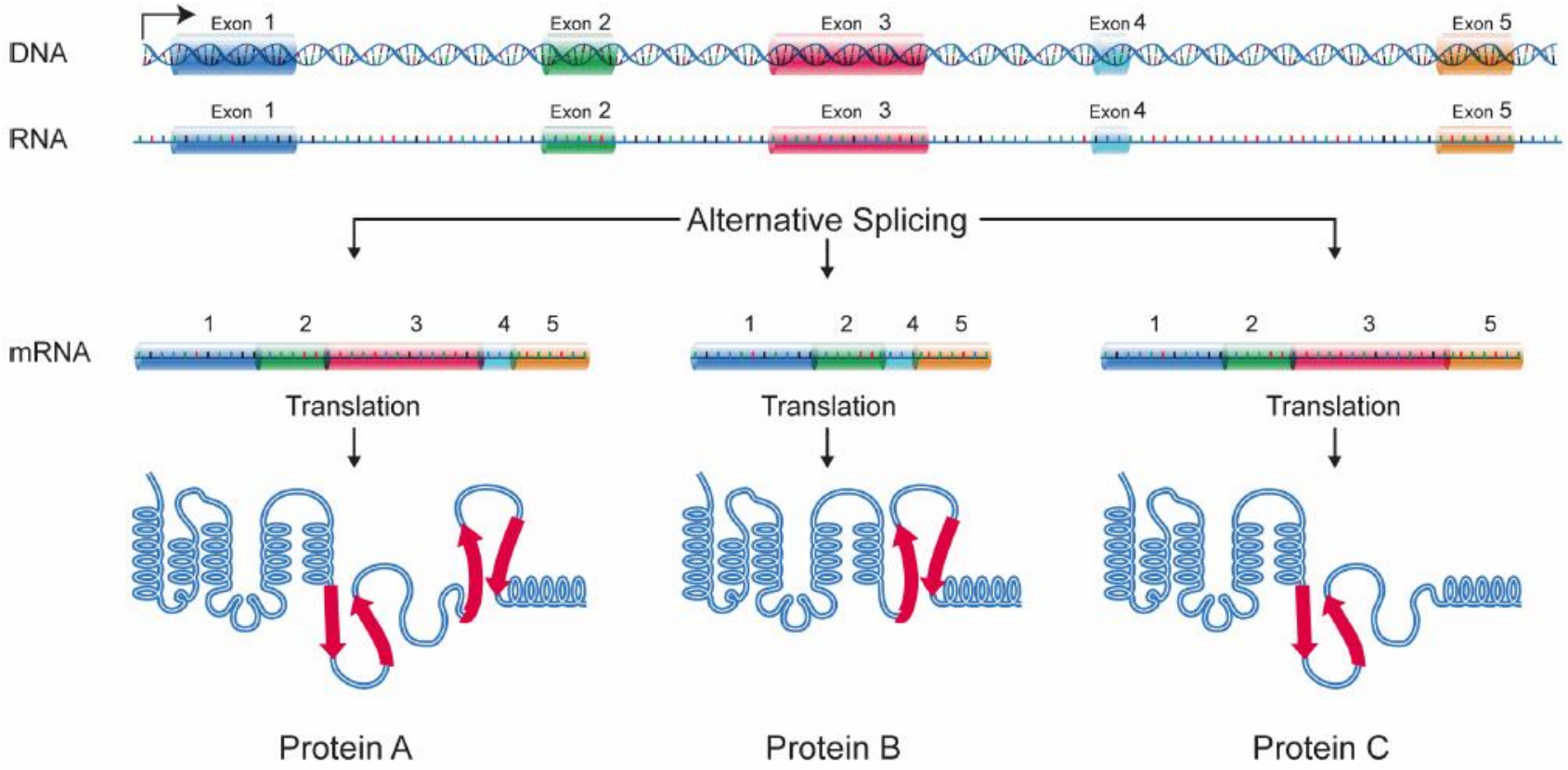
Outline

Brief repetition of splicing



Problem: How to find splicing junctions?

Alternative splicing



Problem: Which isoforms exist?

Spliced mapping

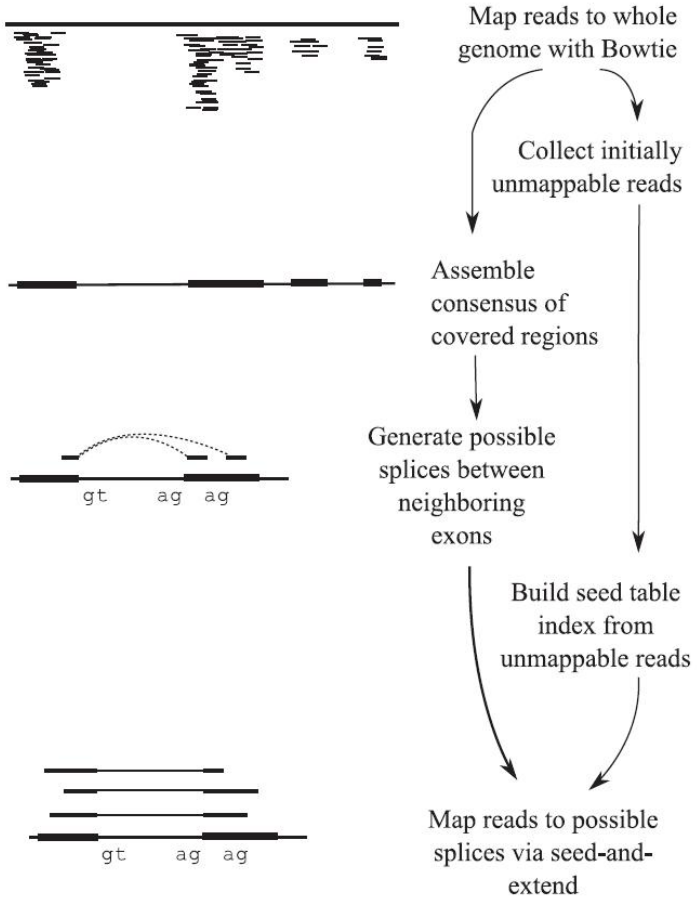
Solutions 1:

- Map reads to known splicing patterns
- Good results for (moderately) abundant transcripts, but not appropiate for rare or unknown transcripts!

Spliced Mapping

Solution 2: TopHat

- Finds novel splice junctions fast
- Synthesizes possible splicing from mapped reads and aligns them with unmapped ones
- Junctions missed when sequencing coverage too low
- Introns could contain non-canonical splicing

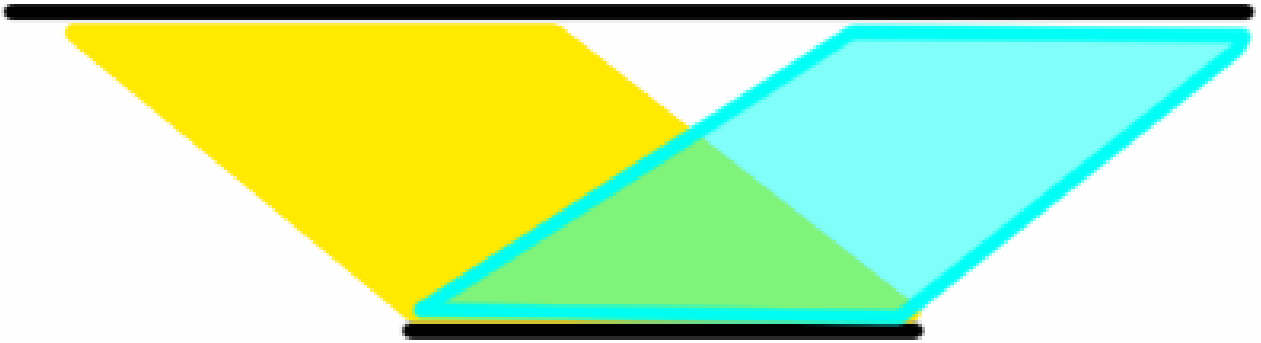


Outline

Multireads

Definition: Reads that have high-scoring alignments to multiple positions in a reference genome

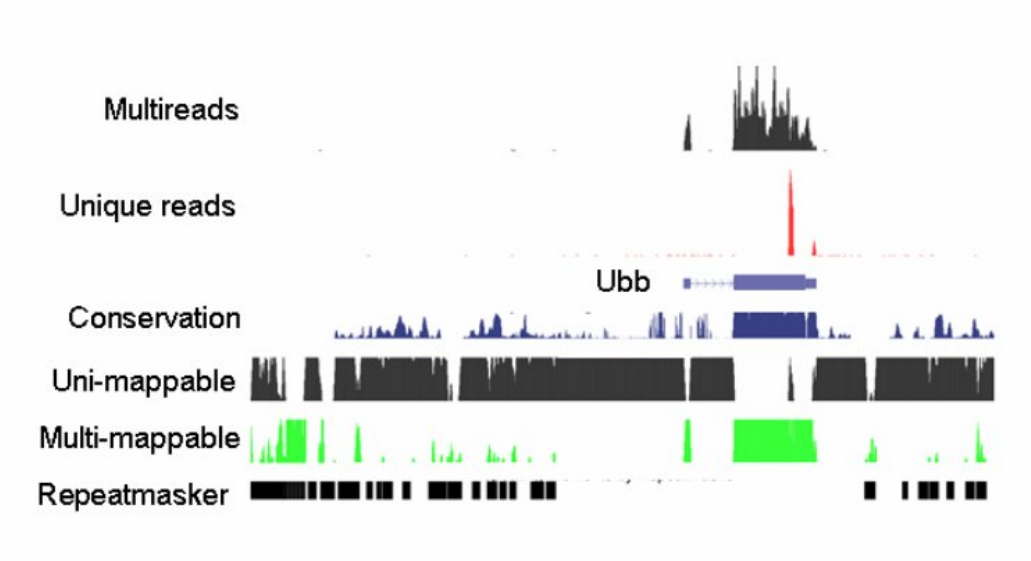
- Gene multiread: map to multiple genes
- Isoform multiread: map to isoforms of one gene



Multireads

Strategies of handling multireads:

- Discard of the multireads
- Allocation of multiread in proportion to the number of unique and splice reads recorded at similiar loci



Multialignment

What happend there?

- Sequence erros (Multiple genes)
- Mapping of reads form other isoforms of that gene
- Repetitive Sequences
  - Identification of columns where differences occur
    Algorithm of Tammi
  - Separation and Identification of the different instances
    Algorithm of Kececioglu

```
...AGCCGTCAGA...
...AGCCGTCAGA...
...AGCCCTCTGA...
...TGTCGTCTGA...
...AGTCGTCTCA...
...AGTCGTCTGA...
```

Sources

- Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer & Barbara Wold.
  Mapping and quantifying mammalian transcriptomes by RNA-Seq.
  Nature Methods, Vol.5 No.7, July 2008, 621–628.
- Bo Li, Victor Ruotti, Ron M. Stewart, James A. Thomson & Colin N. Dewey.
  RNA-Seq gene expression estimation with read mapping uncertainty.
  Bioinformatics, Vol. 26 No. 4 2010, pages 493–500.
- Ben Langmead, Cole Trapnell, Mihai Pop & Steven L Salzberg.
  Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.
  Genome Biology 2009, 10:R25.

Thank you

Questions?

Google | What is Read Mapping | Search | Advanced Search

Web | Show options... | Results 1 - 10 of about 417,000,000 for What is Read Mapping.

How to read a map
Learning to read maps is easy and intuitive. A person can use the map skills discussed in this article to solve navigation problems, plan future activities, ...
www.pacificislandtravel.com/nature.../howtoreadamap.htm - Cached - Similar

How to Read Topographical Maps
Knowing how to read a USGS topographical map is essential to successfully finding a ghost town. USGS topographical maps are useful because they show the ...
www.ghosttowns.com/topotmaps.html - Cached - Similar

How to read a map
How to read a map and a compass. Knowing how to read a map and a compass is a required skill for all wilderness travelers.
www.wilderness-backpacking.com/how-to-read-a-map.html - Cached - Similar

Map Reading - A Free e-book on how to read topographic maps and ...
A free online e-book on map reading and land navigation. It covers how to read a map and how to use a compass. Based on US Army training manuals.
www.map-reading.com/ - Cached - Similar