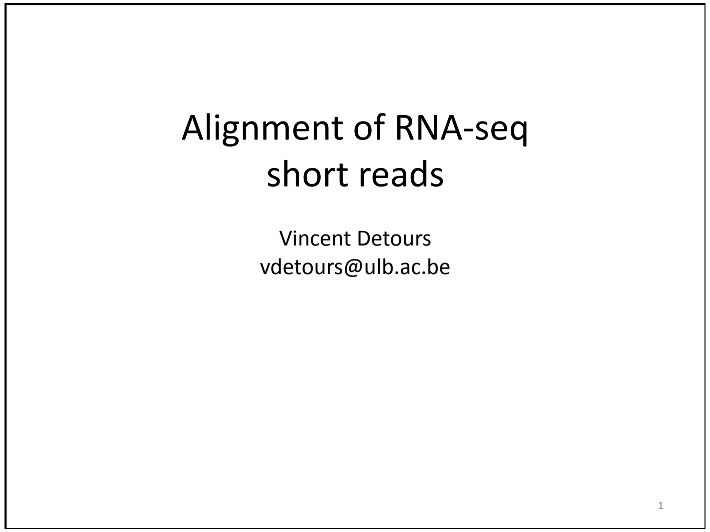
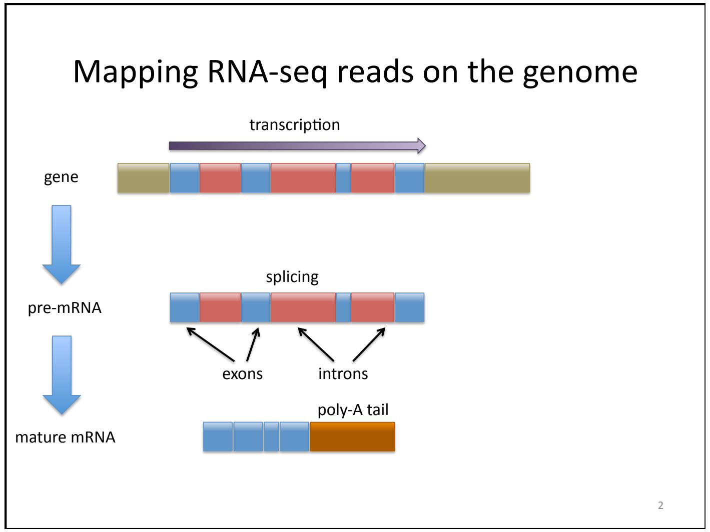
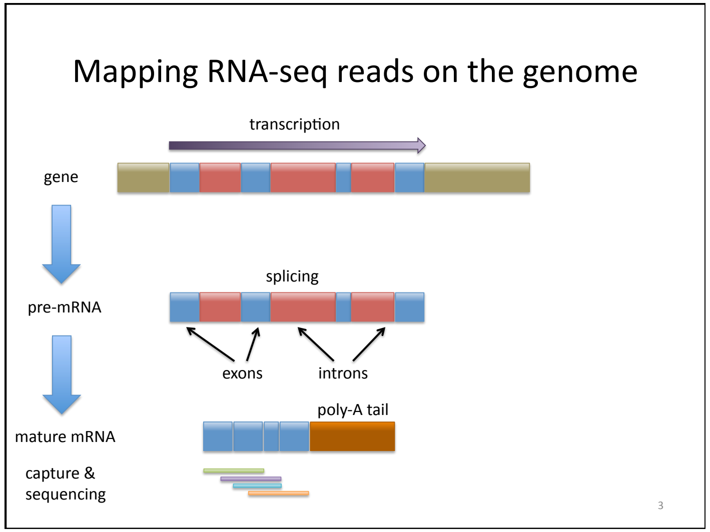


Alignment of RNA-seq short reads

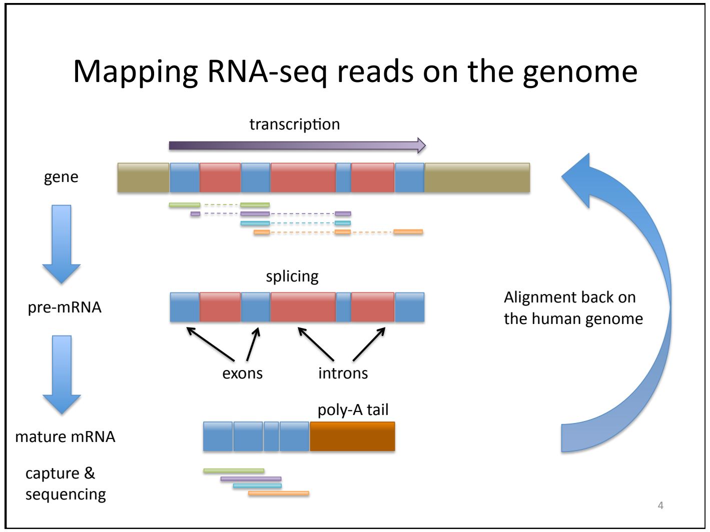
Vincent Detours
vdetours@ulb.ac.be



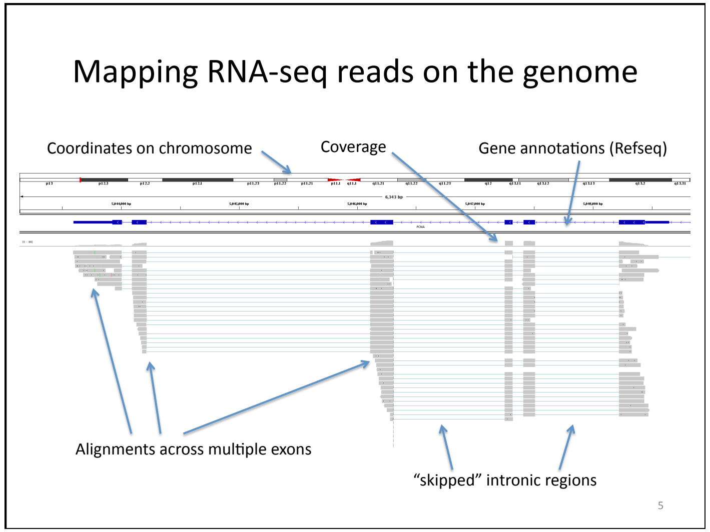
Mapping RNA-seq reads on the genome



Mapping RNA-seq reads on the genome



Mapping RNA-seq reads on the genome



Mapping RNA-seq reads on the genome

Coordinates on chromosome

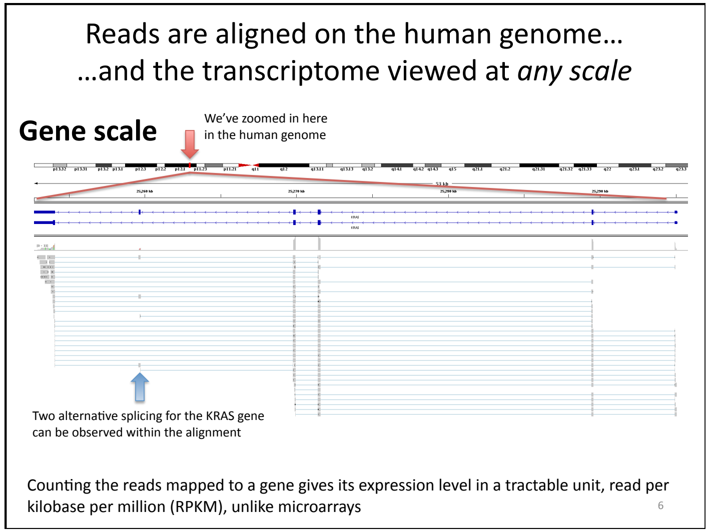
Coverage

Gene annotations (Refseq)

Alignments across multiple exons

“skipped” intronic regions

5



Reads are aligned on the human genome… …and the transcriptome viewed at *any scale*

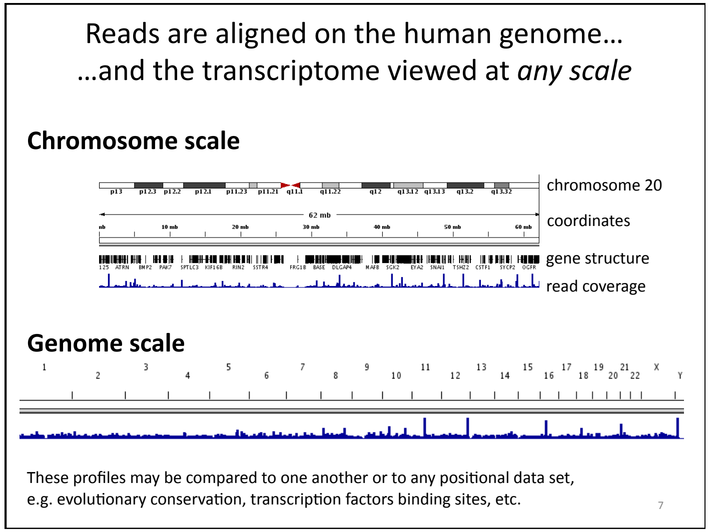
Gene scale

We’ve zoomed in here in the human genome

Two alternative splicing for the KRAS gene can be observed within the alignment

Counting the reads mapped to a gene gives its expression level in a tractable unit, read per kilobase per million (RPKM), unlike microarrays

6



Reads are aligned on the human genome… …and the transcriptome viewed at *any scale*

Chromosome scale

chromosome 20

coordinates

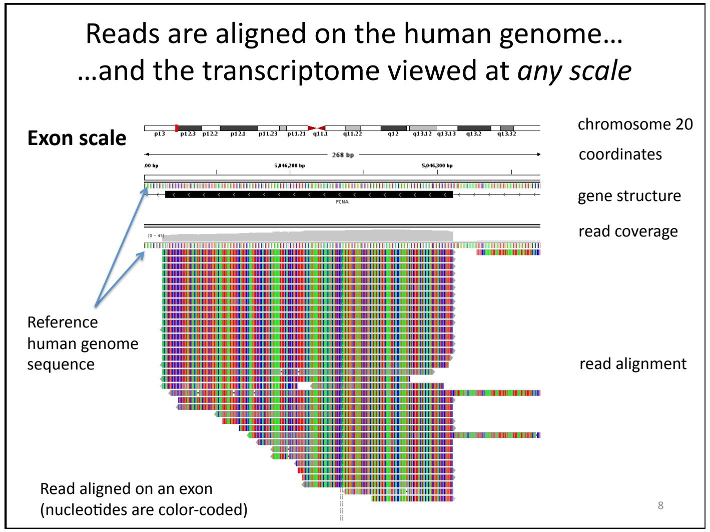
gene structure

read coverage

Genome scale

These profiles may be compared to one another or to any positional data set, e.g. evolutionary conservation, transcription factors binding sites, etc.

7



Reads are aligned on the human genome… …and the transcriptome viewed at *any scale*

Exon scale

chromosome 20

coordinates

gene structure

read coverage

read alignment

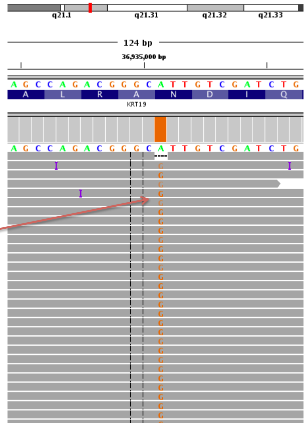
Reference human genome sequence

Read aligned on an exon (nucleotides are color-coded)

8

Reads are aligned on the human genome… …and the transcriptome viewed at *any scale*

Base pair scale



chromosome 17

coordinates

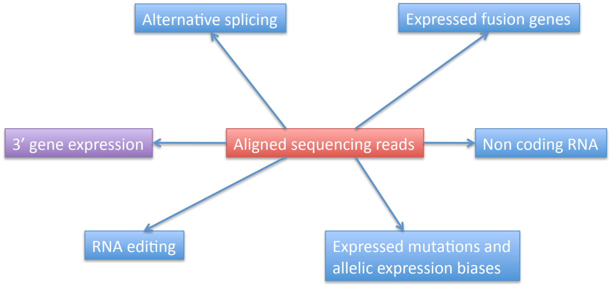
gene structure

read alignment

Reference human genome sequence

All reads from this luminal A tumor have a A>G substitution at chr17:36,935,001 (probably inherited, It is known as rs4601 in dbSNP)

Alignment is the prerequisite for a wide range of investigations



Two alignment strategies

De novo assembly:

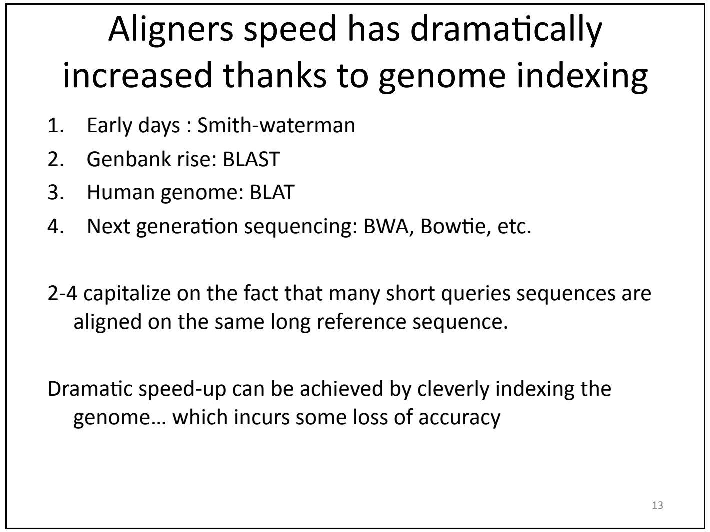
- The genome structure is not known beforehand
- Short 'shot gun' sequences are assembled together like the piece of a puzzle
- Used for new organism, transcripts, etc.

Alignment on a reference genome

- The genome of the organism has been sequenced, new sequences are placed on the reference sequence
- Use for resequencing (e.g. human cancer, human genetics, etc.)

Alignment is challenging

- A typical sequencing run generate millions or billions of reads that need to be aligned on the 3 billions base pair human genome. It's a lot of data and a lot of computations.
- The genome is highly repetitive
  - It evolved as a gigantic copy/paste game through duplications and losses of large and small fragments of DNA. Gene duplication followed by divergence is a major route for genetic innovation.
  - The genome is cluttered with parasitic elements, eg. Alu = 8%, endogenous retroviruses another 8%, etc.
- Sequencing reads are short (2*100bp) and noisy, locating from where they originate in the genome can be error prone, or impossible

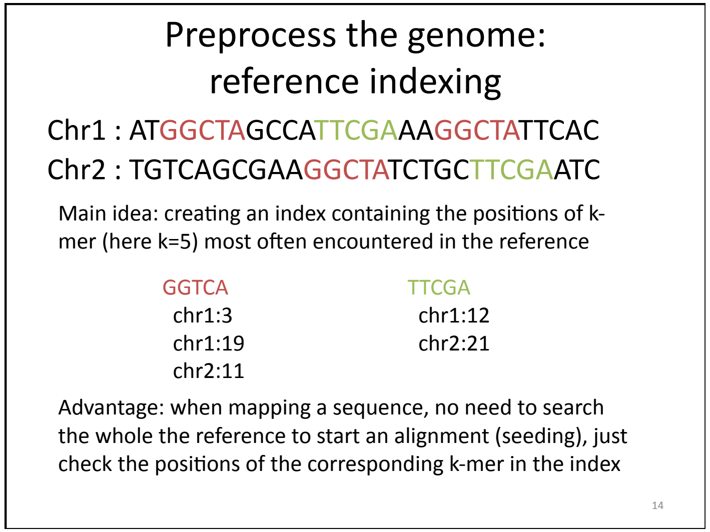


Aligners speed has dramatically increased thanks to genome indexing

1. Early days : Smith-waterman
2. Genbank rise: BLAST
3. Human genome: BLAT
4. Next generation sequencing: BWA, Bowtie, etc.

2-4 capitalize on the fact that many short queries sequences are aligned on the same long reference sequence.

Dramatic speed-up can be achieved by cleverly indexing the genome… which incurs some loss of accuracy



Preprocess the genome: reference indexing

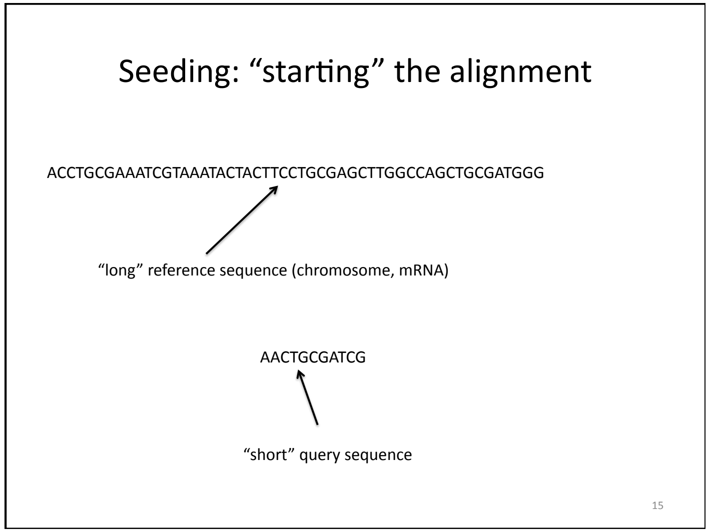
Chr1 : ATGGCTAGCCATTCGAAAGGCTATTCAC

Chr2 : TGTCAGCGAAGGCTATCTGCTTCGAATC

Main idea: creating an index containing the positions of k-mer (here k=5) most often encountered in the reference

GGTCA	TTCGA
chr1:3	chr1:12
chr1:19	chr2:21
chr2:11	

Advantage: when mapping a sequence, no need to search the whole the reference to start an alignment (seeding), just check the positions of the corresponding k-mer in the index



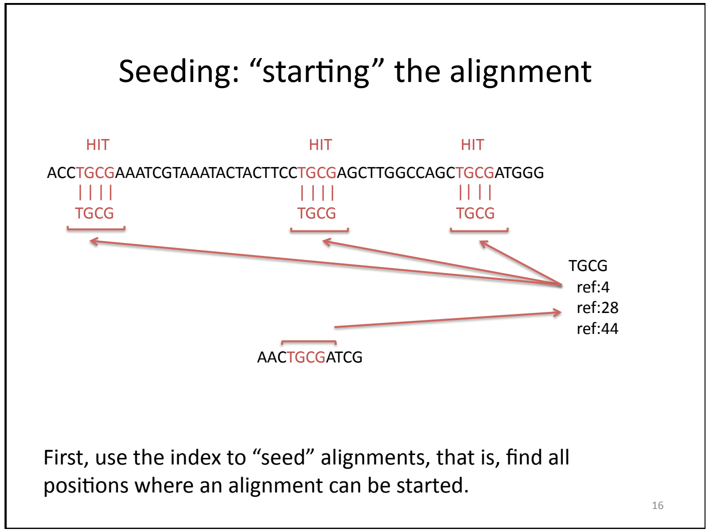
Seeding: "starting" the alignment

ACCTGCGAAATCGTAAATACTACTTCCTGCGAGCTTGGCCAGCTGCGATGGG

"long" reference sequence (chromosome, mRNA)

AACTGCGATCG

"short" query sequence



Seeding: "starting" the alignment

HIT HIT HIT

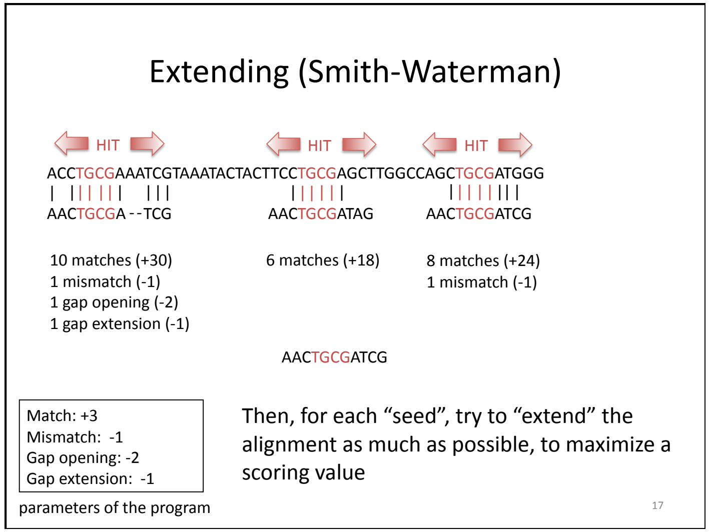
ACCTGCGAAATCGTAAATACTACTTCCTGCGAGCTTGGCCAGCTGCGATGGG

TGCG TGCG TGCG

TGCG
ref:4
ref:28
ref:44

AACTGCGATCG

First, use the index to "seed" alignments, that is, find all positions where an alignment can be started.



Extending (Smith-Waterman)

HIT HIT HIT

```
ACCTGCGAAATCGTAAATACTACTTCCTGCGAGCTTGGCCAGCTGCGATGGG
|  ||||||    |||          ||||||       ||||||||
AACTGCGA--TCG             AACTGCGATAG  AACTGCGATCG
```

10 matches (+30)
1 mismatch (-1)
1 gap opening (-2)
1 gap extension (-1)

6 matches (+18)

8 matches (+24)
1 mismatch (-1)

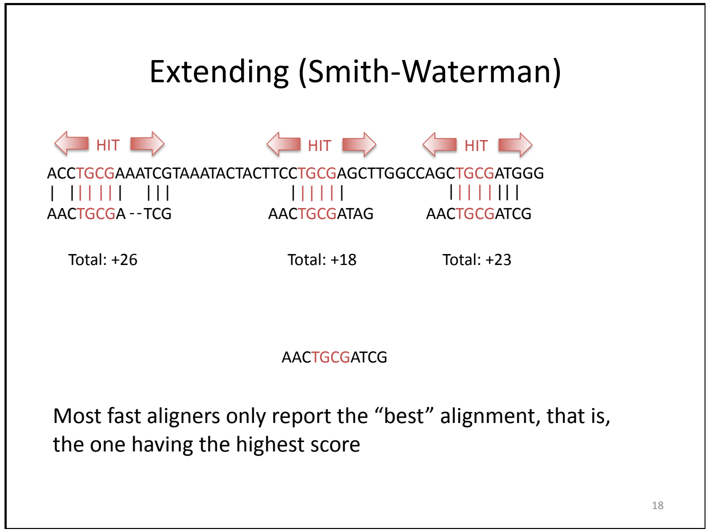
AACTGCGATCG

Match: +3
Mismatch: -1
Gap opening: -2
Gap extension: -1

parameters of the program

Then, for each “seed”, try to “extend” the alignment as much as possible, to maximize a scoring value

17



Extending (Smith-Waterman)

HIT HIT HIT

```
ACCTGCGAAATCGTAAATACTACTTCCTGCGAGCTTGGCCAGCTGCGATGGG
|  ||||||    |||          ||||||       ||||||||
AACTGCGA--TCG             AACTGCGATAG  AACTGCGATCG
```

Total: +26

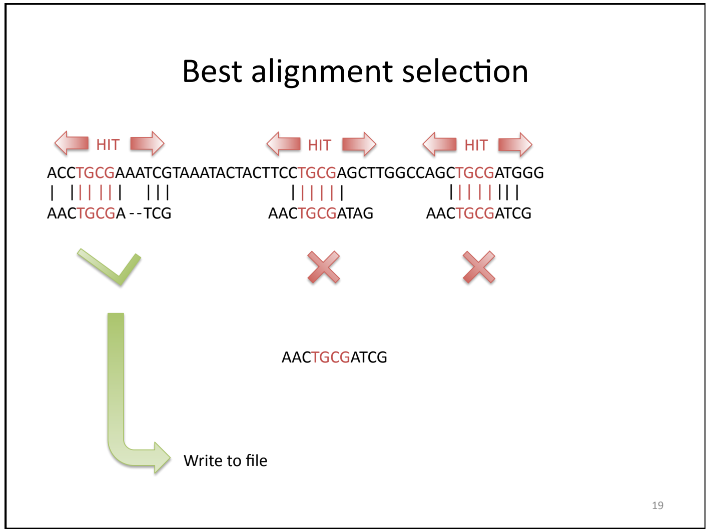
Total: +18

Total: +23

AACTGCGATCG

Most fast aligners only report the “best” alignment, that is, the one having the highest score

18



Best alignment selection

HIT HIT HIT

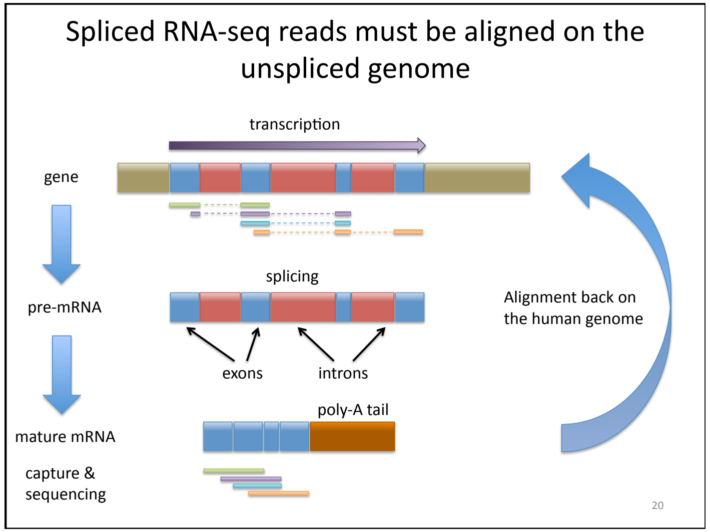
```
ACCTGCGAAATCGTAAATACTACTTCCTGCGAGCTTGGCCAGCTGCGATGGG
|  ||||||    |||          ||||||       ||||||||
AACTGCGA--TCG             AACTGCGATAG  AACTGCGATCG
```

✓ ✗ ✗

AACTGCGATCG

Write to file

19



Spliced RNA-seq reads must be aligned on the unspliced genome

transcription

gene

splicing

pre-mRNA

exons

introns

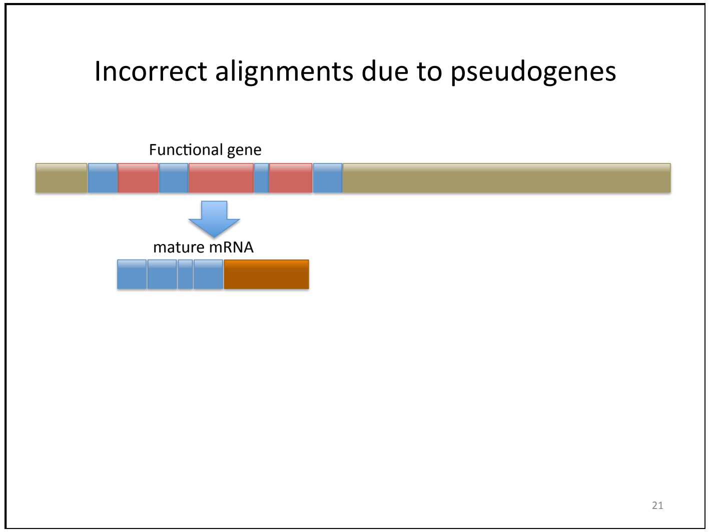
poly-A tail

mature mRNA

capture & sequencing

Alignment back on the human genome

20

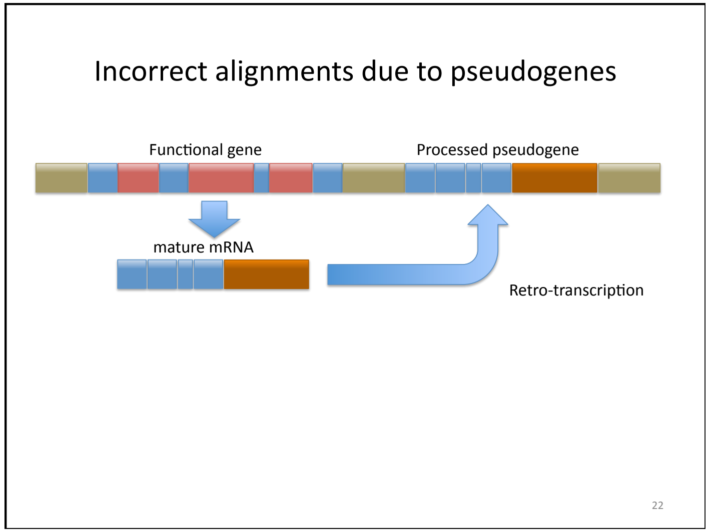


Incorrect alignments due to pseudogenes

Functional gene

mature mRNA

21



Incorrect alignments due to pseudogenes

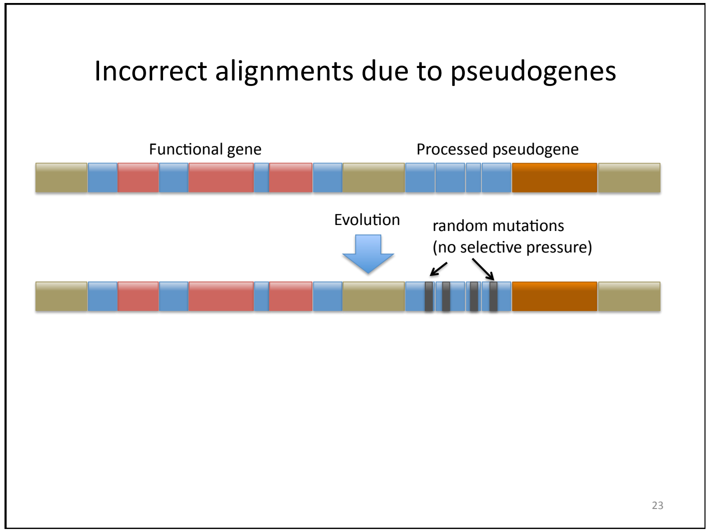
Functional gene

Processed pseudogene

mature mRNA

Retro-transcription

22



Incorrect alignments due to pseudogenes

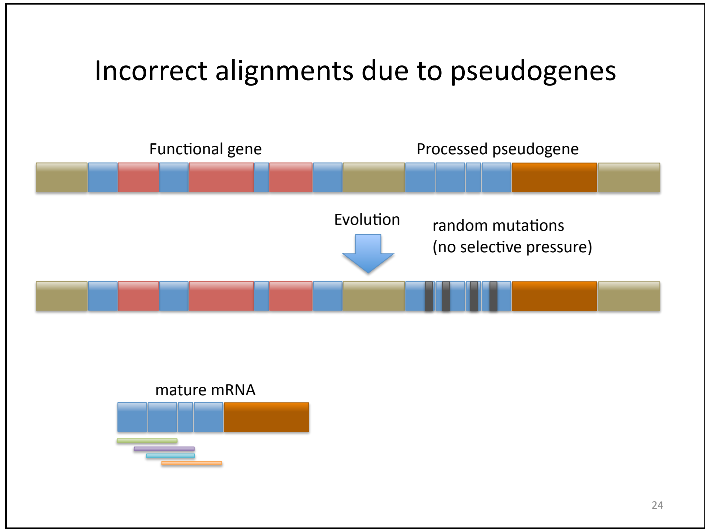
Functional gene

Processed pseudogene

Evolution

random mutations
(no selective pressure)

23



Incorrect alignments due to pseudogenes

Functional gene

Processed pseudogene

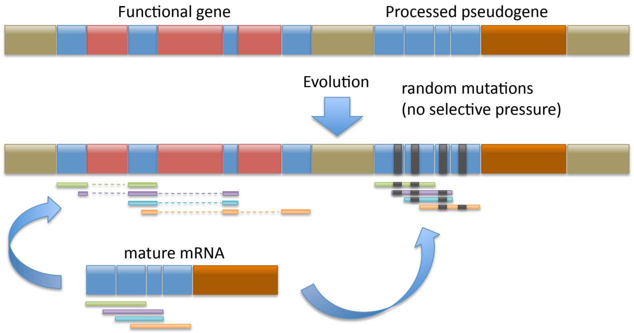
Evolution

random mutations
(no selective pressure)

mature mRNA

24

Incorrect alignments due to pseudogenes



Functional gene

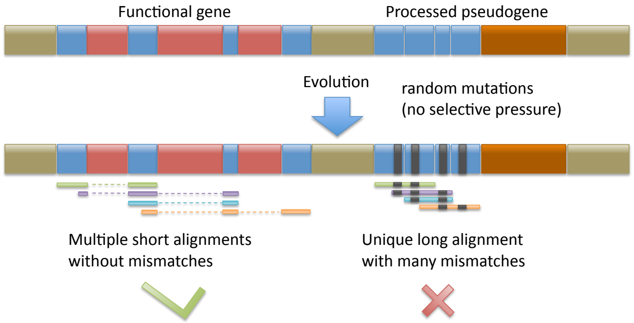
Processed pseudogene

Evolution

random mutations
(no selective pressure)

mature mRNA

Incorrect alignments due to pseudogenes



Functional gene

Processed pseudogene

Evolution

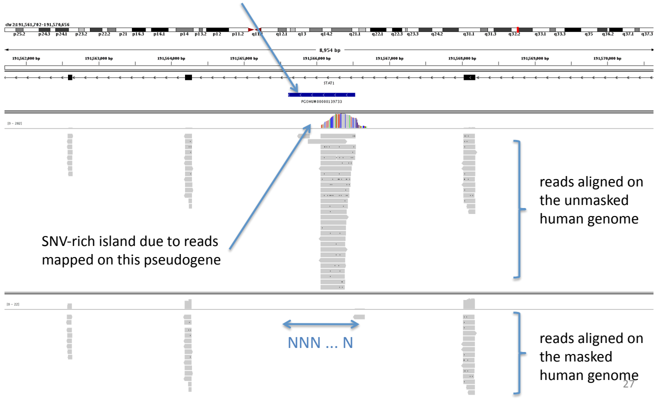
random mutations
(no selective pressure)

Multiple short alignments
without mismatches

Unique long alignment
with many mismatches

Example of induced alignment artifact

pseudogene PGOHUM00000139733, located within a STAT1 intron (chr2:191,565,613-191,566,538)



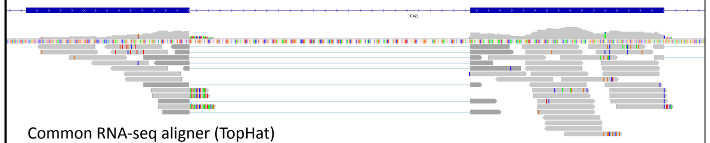
SNV-rich island due to reads
mapped on this pseudogene

reads aligned on
the unmasked
human genome

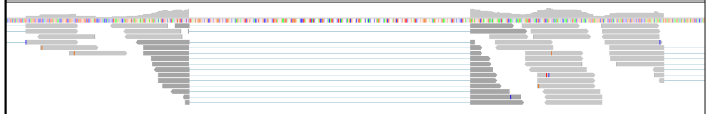
NNN ... N

reads aligned on
the masked
human genome

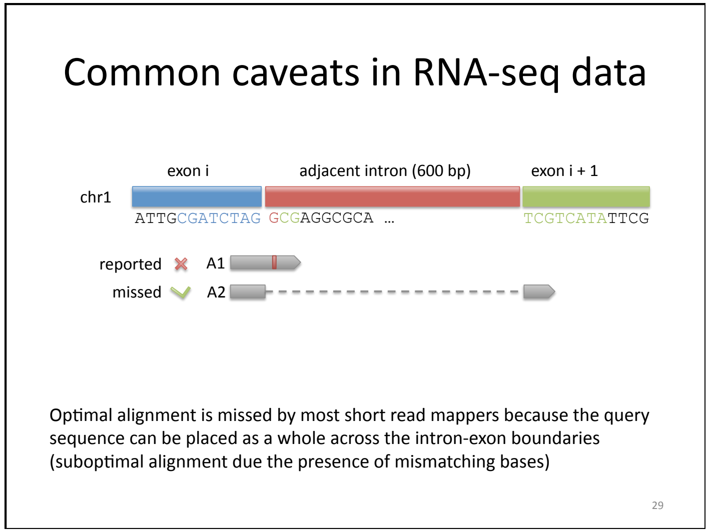
Common caveats in RNA-seq data



Common RNA-seq aligner (TopHat)



Programs we have developed to obtain accurate RNA-seq alignments



Common caveats in RNA-seq data

exon i | adjacent intron (600 bp) | exon i + 1

chr1

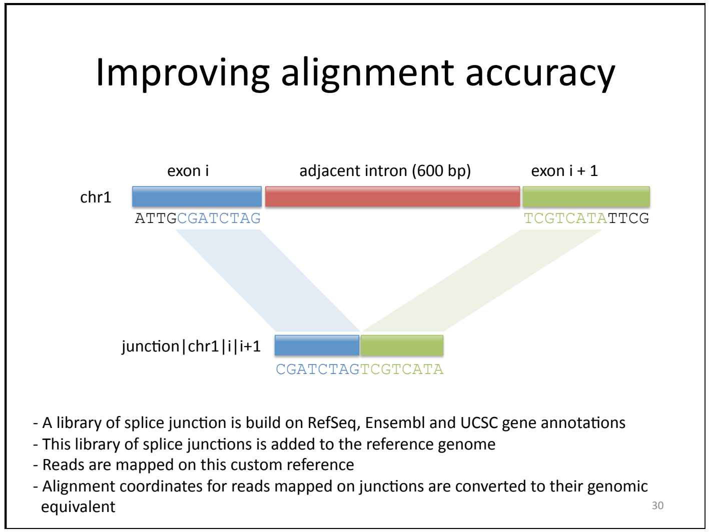
ATTGCGATCTAG GCGAGGCGCA … TCGTCATATTCG

reported ✗ A1

missed ✓ A2

Optimal alignment is missed by most short read mappers because the query sequence can be placed as a whole across the intron-exon boundaries (suboptimal alignment due the presence of mismatching bases)

29



Improving alignment accuracy

exon i | adjacent intron (600 bp) | exon i + 1

chr1

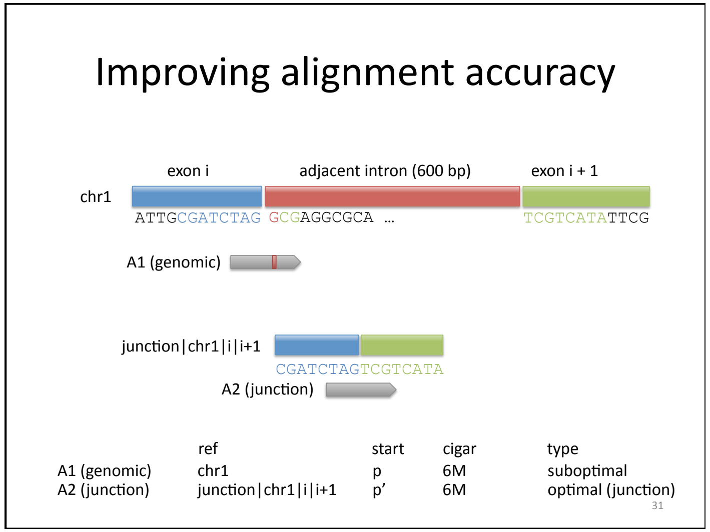
ATTGCGATCTAG TCGTCATATTCG

junction|chr1|i|i+1

CGATCTAGTCGTCATA

- A library of splice junction is build on RefSeq, Ensembl and UCSC gene annotations
- This library of splice junctions is added to the reference genome
- Reads are mapped on this custom reference
- Alignment coordinates for reads mapped on junctions are converted to their genomic equivalent

30



Improving alignment accuracy

exon i | adjacent intron (600 bp) | exon i + 1

chr1

ATTGCGATCTAG GCGAGGCGCA … TCGTCATATTCG

A1 (genomic)

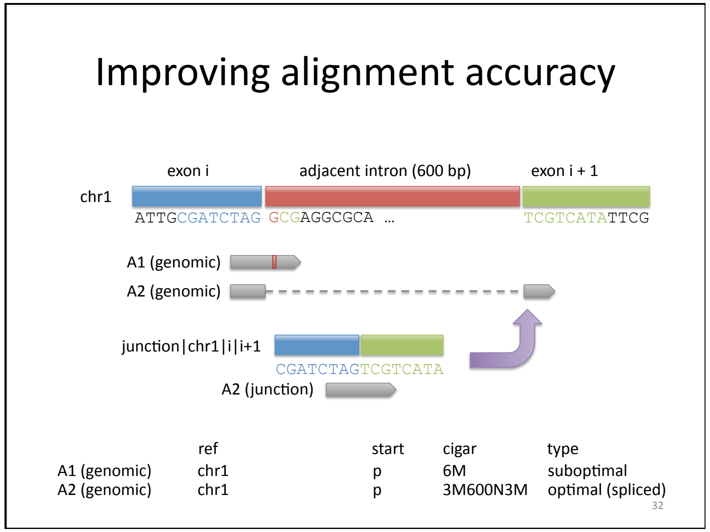
junction|chr1|i|i+1

CGATCTAGTCGTCATA

A2 (junction)

	ref	start	cigar	type
A1 (genomic)	chr1	p	6M	suboptimal
A2 (junction)	junction|chr1|i|i+1	p’	6M	optimal (junction)

31



Improving alignment accuracy

exon i | adjacent intron (600 bp) | exon i + 1

chr1

ATTGCGATCTAG GCGAGGCGCA … TCGTCATATTCG

A1 (genomic)

A2 (genomic)

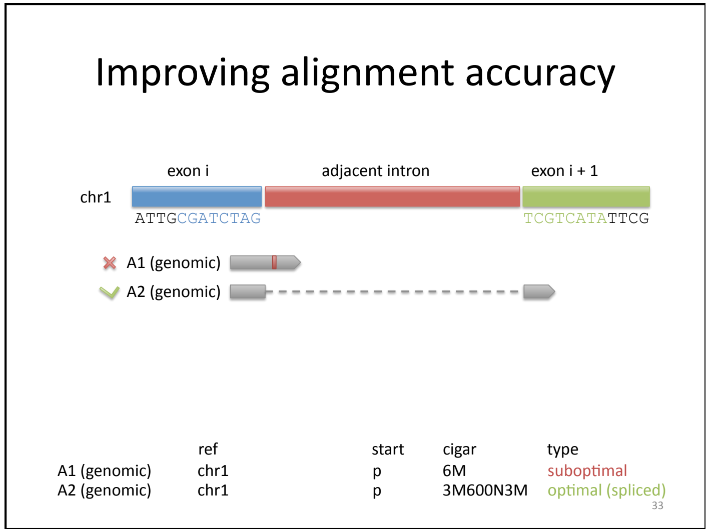
junction|chr1|i|i+1

CGATCTAGTCGTCATA

A2 (junction)

	ref	start	cigar	type
A1 (genomic)	chr1	p	6M	suboptimal
A2 (genomic)	chr1	p	3M600N3M	optimal (spliced)

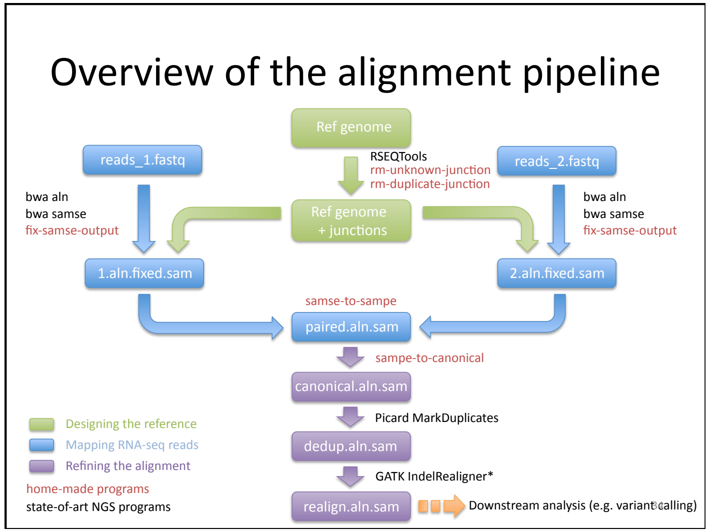
32



Improving alignment accuracy

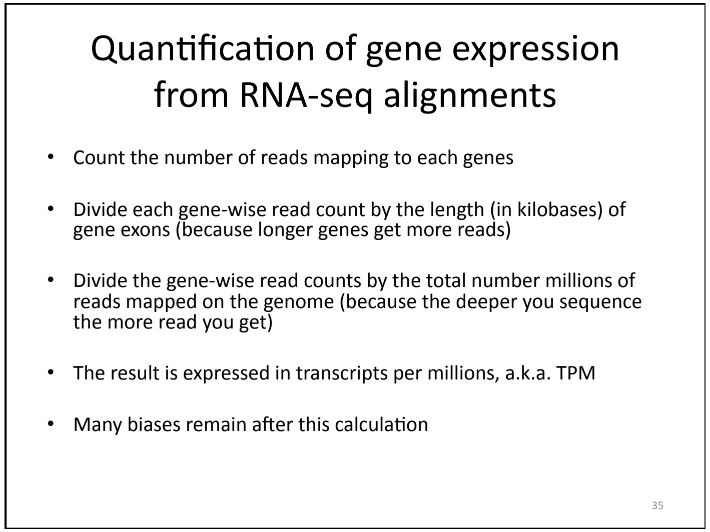
	ref	start	cigar	type
A1 (genomic)	chr1	p	6M	suboptimal
A2 (genomic)	chr1	p	3M600N3M	optimal (spliced)

33



Overview of the alignment pipeline

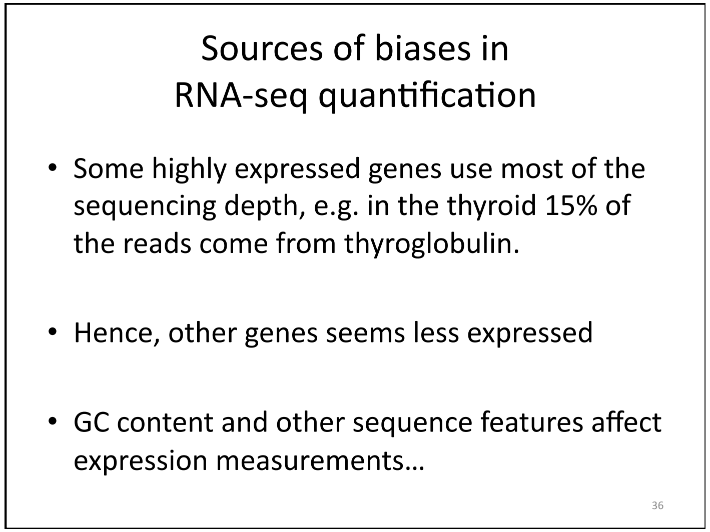
34



Quantification of gene expression from RNA-seq alignments

- Count the number of reads mapping to each genes
- Divide each gene-wise read count by the length (in kilobases) of gene exons (because longer genes get more reads)
- Divide the gene-wise read counts by the total number millions of reads mapped on the genome (because the deeper you sequence the more read you get)
- The result is expressed in transcripts per millions, a.k.a. TPM
- Many biases remain after this calculation

35



Sources of biases in RNA-seq quantification

- Some highly expressed genes use most of the sequencing depth, e.g. in the thyroid 15% of the reads come from thyroglobulin.
- Hence, other genes seems less expressed
- GC content and other sequence features affect expression measurements…

36