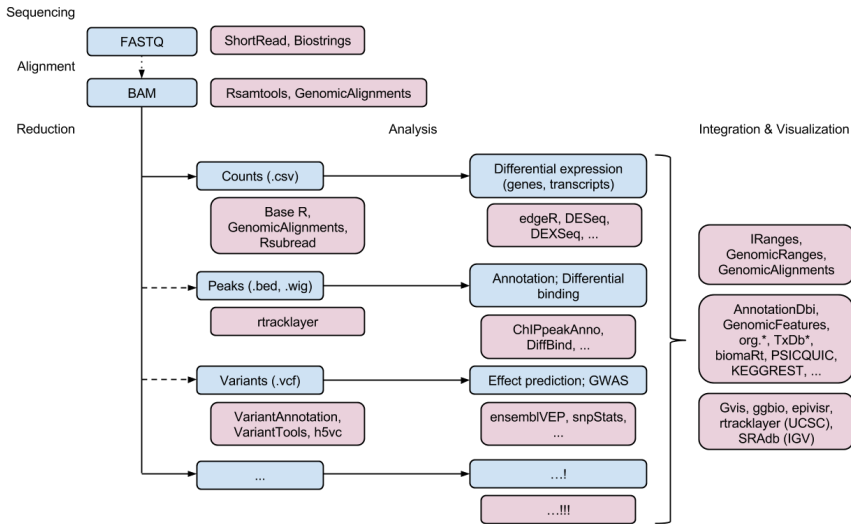


# RNA-seq preprocessing

Mikhail Dozmorov

Spring 2018

# Computational ecosystem of sequencing



<http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0881-8>

# FASTA/FASTQ format

## FASTA: text-based representation of nucleotide sequence

```
>Human mitochondrion
GATCACAGGTCTATCACCCCTATTAACCACTCACGGGAGCTCTCCATGCATTTGGTATTTTCGTCCTGGGGG
GTATGCACGCGATAGCATTGCGAGACGCTGGAGCCGGAGCACCCCTATGTGCGAGTATCTGTCTTTGATTC
CTGCCTCATCCTATTATTTATCGCACCTACGTTCAATATTACAGGCGAACATACTTACTAAAGTGTGTTA
```

## FASTQ: sequence and quality info

```
@M01127:9:000000000-A7LUJ:1:1101:14584:1820 1:N:0:3
CTCAGGTACAAAAGACAGCTGTTTATATTACAGTTTANNNNGTTTCAGAGTTGGACATTTCACTGTAGGATCTAAAACCACTGAGGTTCCAA
NNNNNNNNNNNNNNNNNTTCAACAAATAAGAAGGAAATGATGTAAATTTATTACTGTGCAAGTCCAAATGTGTCAAACNNNNNCAGNNNNNN.
TGAACCATCTG
+
<==<-775<@@@---A-.888A/8///.-/99/####+7777...-99.--9-8AA8.88.8-5--55A----5>+CE---+-87866
A#####321988088@*1*21*10*01*.6.66(/66?<?<66?6;6.(//.6<E=6;
#####-/ /<66<E6(/.<EEE(6(66(66<<6666(
@M01127:9:000000000-A7LUJ:1:1101:16774:1822 1:N:0:3
CGTCACGACATCAGCCATCTGCCAABGCACATCAGNNNNCCGCTGTTCTGACAGCAGCCACATGCCAGTCACCAATATCCACGCTCT
```

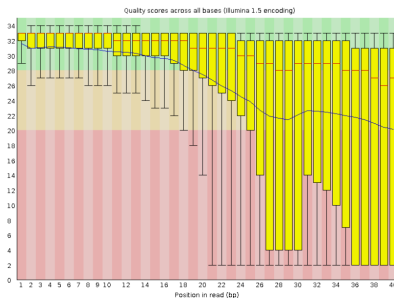
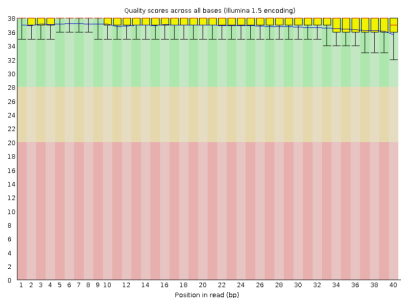
<http://zhanglab.ccmb.med.umich.edu/FASTA/>

# Quality of base calling

- **Phred quality score** is widely used to characterize the quality of base calling
- Phred quality score =  $-10 * \log_{10}(P)$ , where  $P$  is probability that base-calling is wrong
- Phred score of 30 means there is 1/1000 chance that the base-calling is wrong
- The quality of the bases tend to drop at the end of the read, a pattern observed in sequencing-by-synthesis techniques

# Quality control

- **FASTQC** - Quality of raw and aligned sequencing data
  - Base quality per position
  - Nucleotide per position
  - GC content
  - K-mer enrichment



<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>, Video tutorial how to interpret,  
<https://www.youtube.com/watch?v=bz93ReOv87Y>

# RNA-seq-specific quality control

- **RNASeQC** - quality of mapped (aligned) data
- **RSeQC** - Python-based table and graph QC reports
- **MultiQC** - Summarization and visualization QC results for multiple samples in one report. Recognizes multiple QC tools

<http://www.broadinstitute.org/cancer/cga/rna-seqc>, Deluca DS, Levin JZ, Sivachenko A, Fennell T, Nazaire MD, Williams C, Reich M, Winckler W, Getz G. (2012) RNA- SeQC: RNA-seq metrics for quality control and process optimization. Bioinformatics

<http://rseqc.sourceforge.net/>, Wang L, Wang S, Li W\* RSeQC: quality control of RNA- seq experiments Bioinformatics (2012) 28 (16): 2184-2185. doi: 10.1093/bioinformatics/bts356

<http://multiqc.info/>

# Adapter trimming

- **Cutadapt** - full control over adapter trimming
- **FASTX-Toolkit** - set of tools for low-level sequence trimming/cutting
- **Trimmomatic** - well-documented and easy-to-use adapter trimmer using multiple algorithms. Handles single- and paired-end reads, accountss for read quality
- **Flexbar**: similar to Trimmomatic by functionality

<https://cutadapt.readthedocs.io/en/latest/guide.html>

[http://hannonlab.cshl.edu/fastx\\_toolkit/](http://hannonlab.cshl.edu/fastx_toolkit/)

<http://www.usadellab.org/cms/?page=trimmomatic>

<https://github.com/seqan/flexbar/wiki/Manual>

# Duplicates removal

- Duplicates may correspond to biased PCR amplification of particular fragments
- For highly expressed, short genes, duplicates are expected even if there is no amplification bias
- Removing them may reduce the dynamic range of expression estimates

Generally, do not remove duplicates from RNA-seq data

- If you ultimately want to remove duplicates, use Picard tools' `MarkDuplicates` command

<https://broadinstitute.github.io/picard/command-line-overview.html#MarkDuplicates>



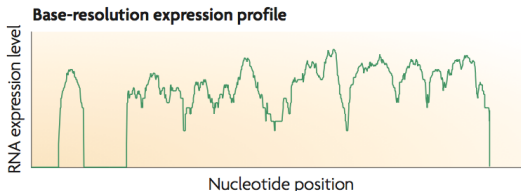
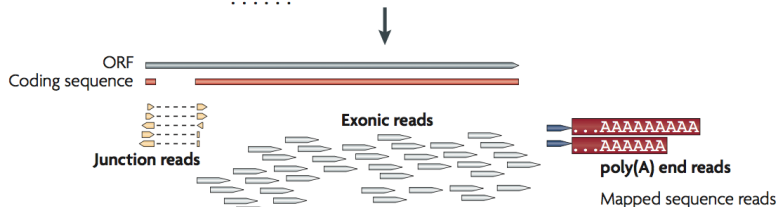
# Alignment

- RNA-seq aligners face an additional problem, not encountered in DNA-only alignment: many RNA-seq reads will span introns
- The average human intron length is  $>6,000$  bp (some are  $>1$  Mbp in length)
- In a typical human RNA-seq experiment using 100-bp reads,  $>35\%$  of the reads will span multiple exons - align over splice junctions
- Aligners must be splice-aware, especially when aligning longer ( $>50$ bp) reads

# Three types of reads: exonic-, junction- and poly(A) end-reads.

```
ATCACAGTGGGACTCCATAAATTTTCT
CGAAGGACCAGCAGAAACGAGAGAAAAA
GGACAGAGTCCCCAGCGGGCTGAAGGGG
ATGAAACATTAAAGTCAAACAATATGAA
.....
```

Short sequence reads



# Strategies for gapped alignments of RNA-seq reads

## Exon-first method

- Map full, unspliced reads (reads originating from a single exon) to exons
- Divide the remaining reads into smaller pieces and map them to the genome
- An extension process extends mapped smaller pieces to find candidate splice sites to support a spliced alignment.

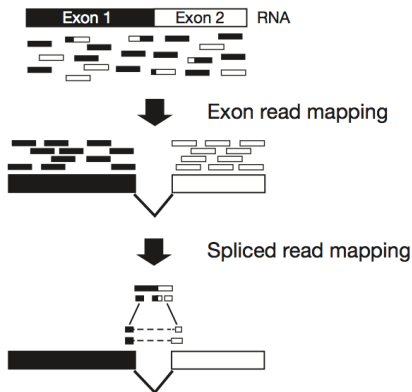
# Strategies for gapped alignments of RNA-seq reads

## Seed-and-extend methods

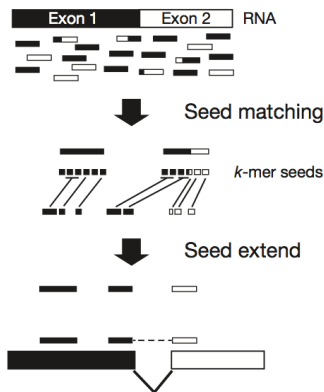
- Divide each RNA-seq read in small words (k-mers) of similar size
- Store a map of all k-mers in the genome in an efficient lookup data structure
- Map k-mers to the genome via the lookup structure
- Mapped k-mers are extended into larger alignments, which may include gaps flanked by splice sites.

# Strategies for gapped alignments of RNA-seq reads

## a Exon-first approach

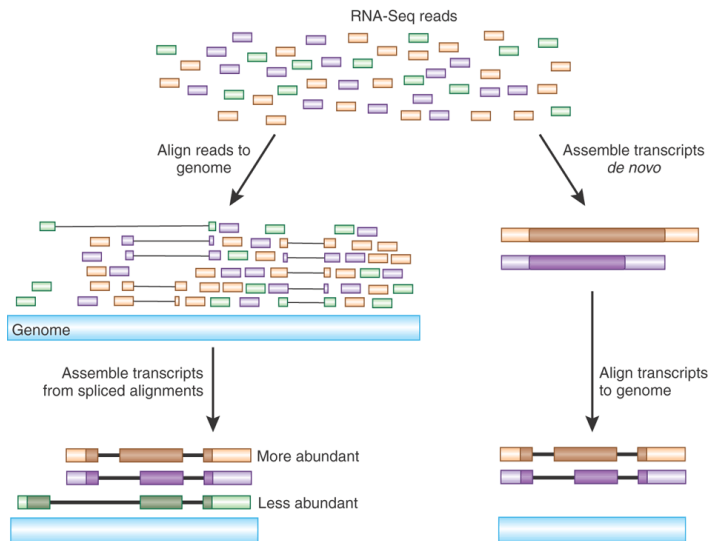


## b Seed-extend approach



<https://www.nature.com/articles/nmeth.1613>

# Alignment to the reference genome is the most frequently used for transcript quantification



# Alignment - Mapping RNA-seq reads to the genome

- **BWA** - general purpose algorithms based on Burrouws-Wheeler Transform
- **STAR** - fast and accurate aligner
- **HISAT**: (hierarchical indexing for spliced alignment of transcripts) uses two types of indexes for alignment: a global, whole-genome index and tens of thousands of small local indexes. Can detect novel splice sites, transcription initiation and termination sites. A part of the new “Tuxedo suite”, including StringTie and Ballgown
- **subread**: a fast and accurate aligner, R and command line. The whole package includes subjunc for junction detection, and featureCounts for extracting read counts per gene from aligned SAM/BAM files

<http://bio-bwa.sourceforge.net/>

<https://github.com/alexdobin/STAR>

<http://ccb.jhu.edu/software/hisat2/index.shtml>

<http://subread.sourceforge.net/>

Timeline and extensive comparison of aligners: [https://www.ebi.ac.uk/~nf/hts\\_mappers/](https://www.ebi.ac.uk/~nf/hts_mappers/)

# De novo assembly

- **Trans-ABYSS** - De novo assembly of RNA-Seq data
- **Velvet-Oases** - De novo transcriptome assembler for very short reads
- **SOAPdenovo-trans** - De novo transcriptome assembler accounting for alternative splicing and different expression level among transcripts
- **Trinity** - RNA-Seq De novo Assembly Using Trinity set of tools

<http://www.bcgsc.ca/platform/bioinfo/software/trans-abyss>

<https://www.ebi.ac.uk/~zerbino/oases/>

<http://soap.genomics.org.cn/SOAPdenovo-Trans.html>

<https://github.com/trinityrnaseq/trinityrnaseq/wiki>



# Trinity *de novo* assembly

- Trinity combines three independent software modules: **Inchworm**, **Chrysalis**, and **Butterfly**
- Trinity partitions the sequence data into many individual de Bruijn graphs, each representing the transcriptional complexity at a given gene or locus, and then processes each graph independently to extract full-length splicing isoforms and to tease apart transcripts derived from paralogous genes.



Grabherr, Manfred G., Brian J. Haas, Moran Yassour, Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis, et al. "Full-Length Transcriptome Assembly from RNA-Seq Data without a Reference Genome." *Nature Biotechnology* 29, no. 7 (May 15, 2011): 644–52. doi:10.1038/nbt.1883.

<https://www.youtube.com/playlist?list=PLfFNmoa-yUIY3-BcTTXsAMMjk-jSVTgwk>