Application and evaluation of different read mapping approaches for genomic alignment

Markus Wolfien and Andrea Bagnacani

de.NBI Training – 28th June 2018 Jena www.sbi.uni-rostock.de





One possible workflow for data processing







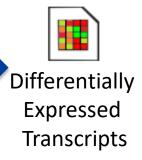
- STAR, Kallisto
- MOSAIK-aligner
 Used by "1000
 Genomes Project"

- Multiple Correction
- Gene fusion
- SNP Calling ready (GATK toolkit)



Pre-Processing (Quality Control, Clipping)

Genomic Alignment Transcript Quantification

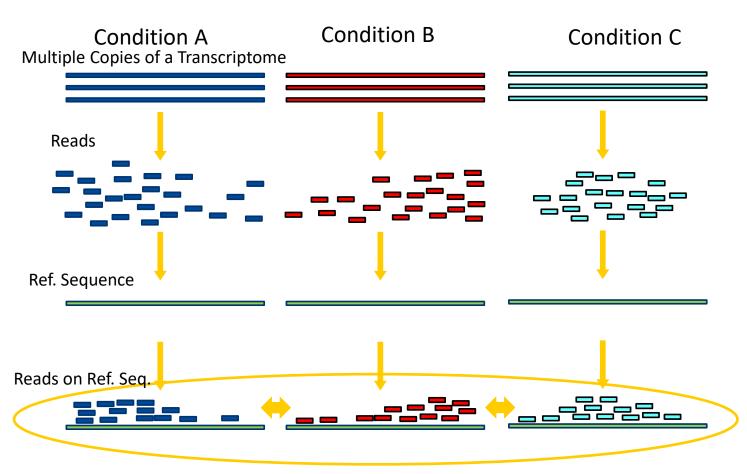


Evaluate Reads
 (e.g. Sequence
 Quality, GC Content,
 Read length)

- Check RPKM Normalization
- Bias Correction

One possible workflow for data processing

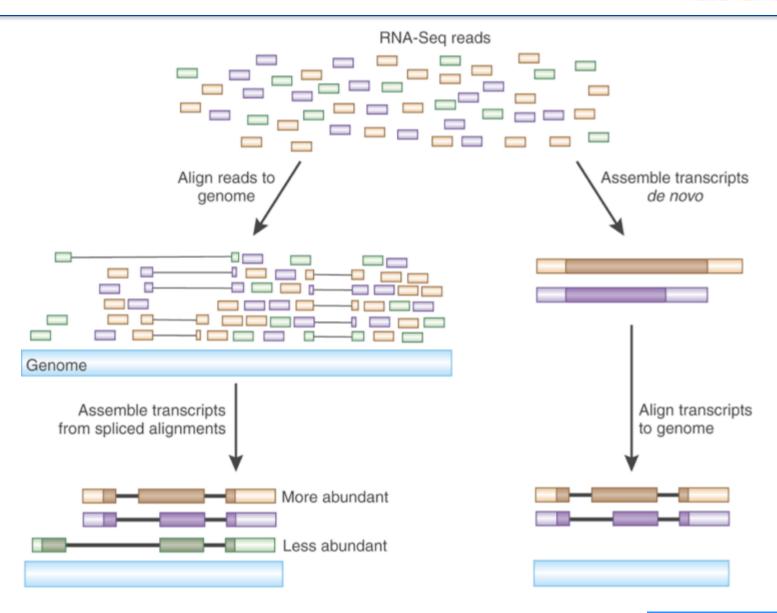




3

What is mapping?

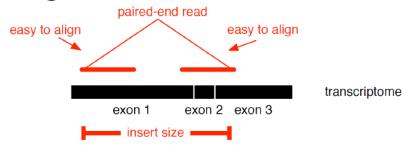




1. Genomic alignment – Splice aware

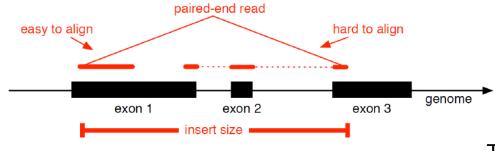


Transcriptome alignment



- reliable gene models required
- no detection of novel genes

Genome alignment (<u>splice-aware</u> read alignment)



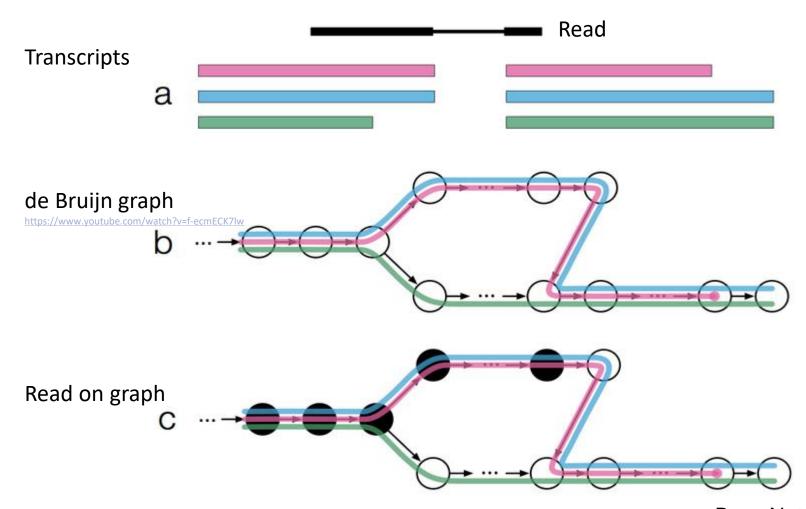
+ detection of novel genes and isoforms

Turro, EMBO, 2012

For clinical usage combination of different algorithms possible

2. Genomic alignment - pseudoalignment

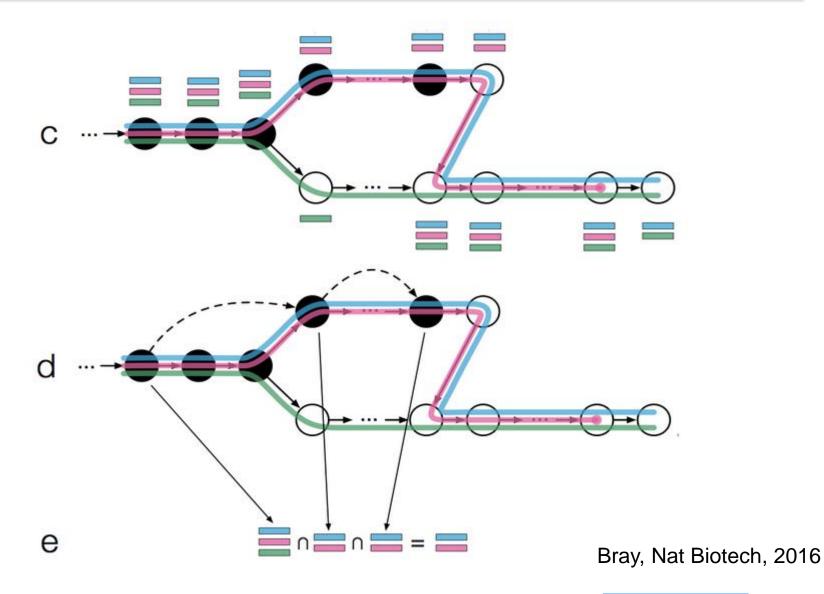




Bray, Nat Biotech, 2016

2. Genomic alignment - pseudoalignment





Transcript quantification



- Read counts
 - Count the reads per feature
 - relatively easy: count the number of reads per gene, exon, ...
 - How to handle multi-mapping reads (i.e. reads with multiple alignments)?
- Normalization aims to make expression levels comparable across:
 - Features (genes, isoforms, ...)
 - RNA libraries (samples)
- Normalization methods:
 - TPM, RPKM / FPKM (Cufflinks /Cuffdiff) (Mortazavi, Nat Meth, 2008)
 - TMM (edgeR) (Robinson & Oshlack, Genome Biol, 2010)
 - DESeq2 (DESeq2) (Love et al., Genome Biol, 2014)





	Read II	D Read	sequence	Rea	d positio	n			
The street of	1	7							
The state of the s	272+13M170; CCTGAGGAGA/		chr1 497	37 0;==-=9;		15 >>>>>>>	100338662 >>>>>> MD:7:37	0 XT:A:U	NM:1:0
	F:275+18M2		X1:1:0	Chr1	X0:1:0 17644	8	37M	=	17919
314	TATGACTGC	TAATAATACC	TACACATGTTA	AGAACCAT	>>>>>>>>	>>>>>>	*****	4::>>:<9	
RG: Z: UM00	98:1	XT:A:R	NM:1:0	SM:1:0	AM:1:8	X0:1:4	X1:1:0	XM:1:8	X0:1:0
XG:1:8	MD:Z:37								
19:20389:	F:275+18M2	D19M	147	chr1	17919	8	18M2D19M		
17644	-314	GTAGTACCA	ACTGTAAGTC	CTTATCTTCAT	ACTITGT	:44999:499	9<8<8<<<8	<><<<>><<	: <<<>>><
XT:A:R 18^CA19	NM:1:2	SM:1:0	AM:1:0	X8:1:4	X1:1:8	XM:1:0	X0:i:1	XG:1:2	MD:Z:
9:21597+1	9M2I25M:R:	-289	83	chr1	21678	0	8M2I27M		
21469	-244	CACCACATO	ACATATACCAZ	AGCCTGGCTGT	GTCTTCT	<:9<<5><<	*******	><>><9>><>	>>9>>><>
XT:A:R	NM:1:2	SM:1:8	AM:1:8	X0:1:5	X1:1:8	XM:1:8	X0:1:1	XG:1:2	MD:Z:35
								-	

Visualization of .bam files





http://software.broadinstitute.org/software/igv/







https://ics.hutton.ac.uk/tablet/



Hands on part 3

10:15 - 11:15

"Application of different read mapping approaches for genomic alignment"

Material: http://galaxyproject.github.io/training-material/topics/sequence-analysis/

