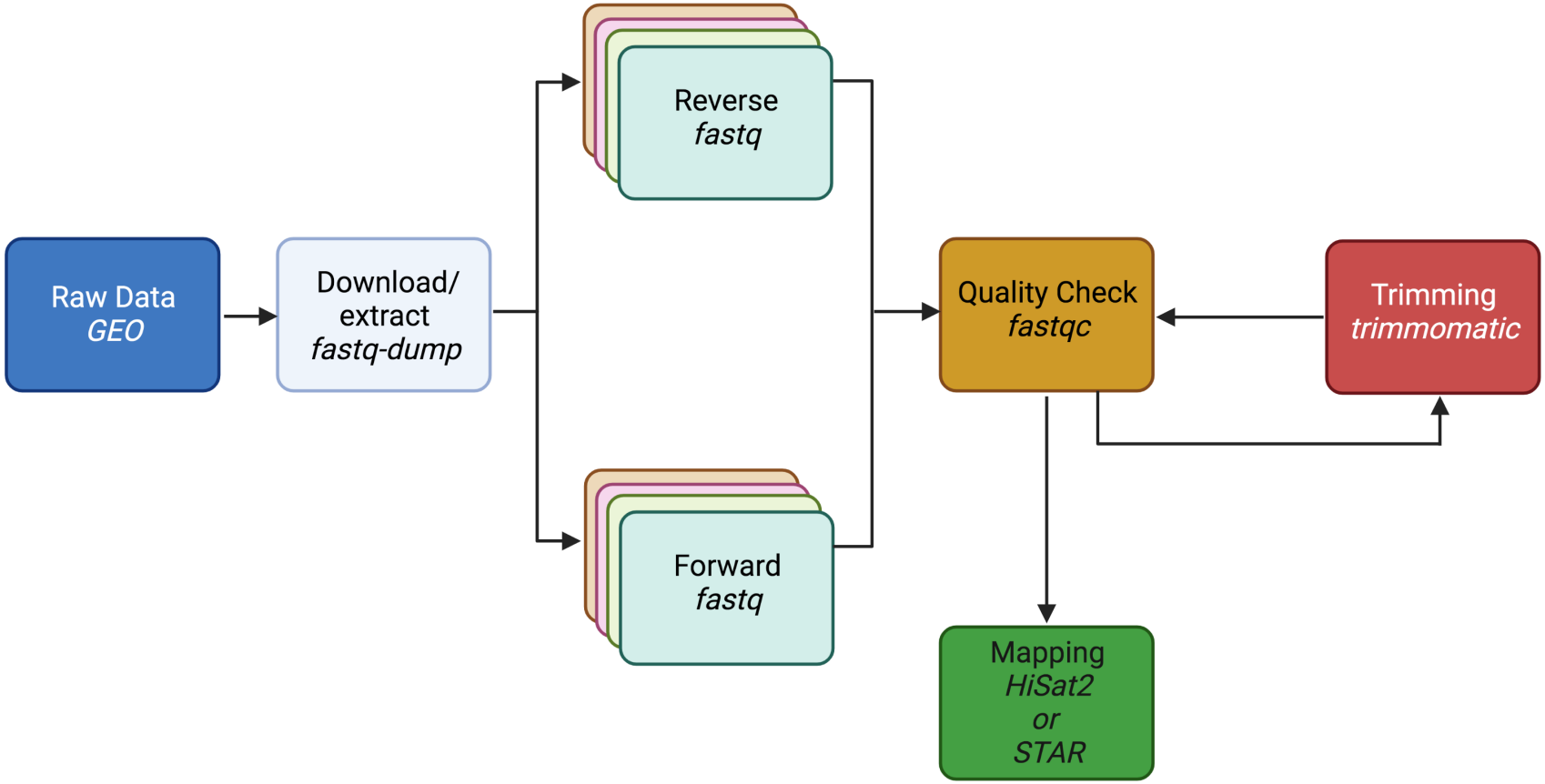
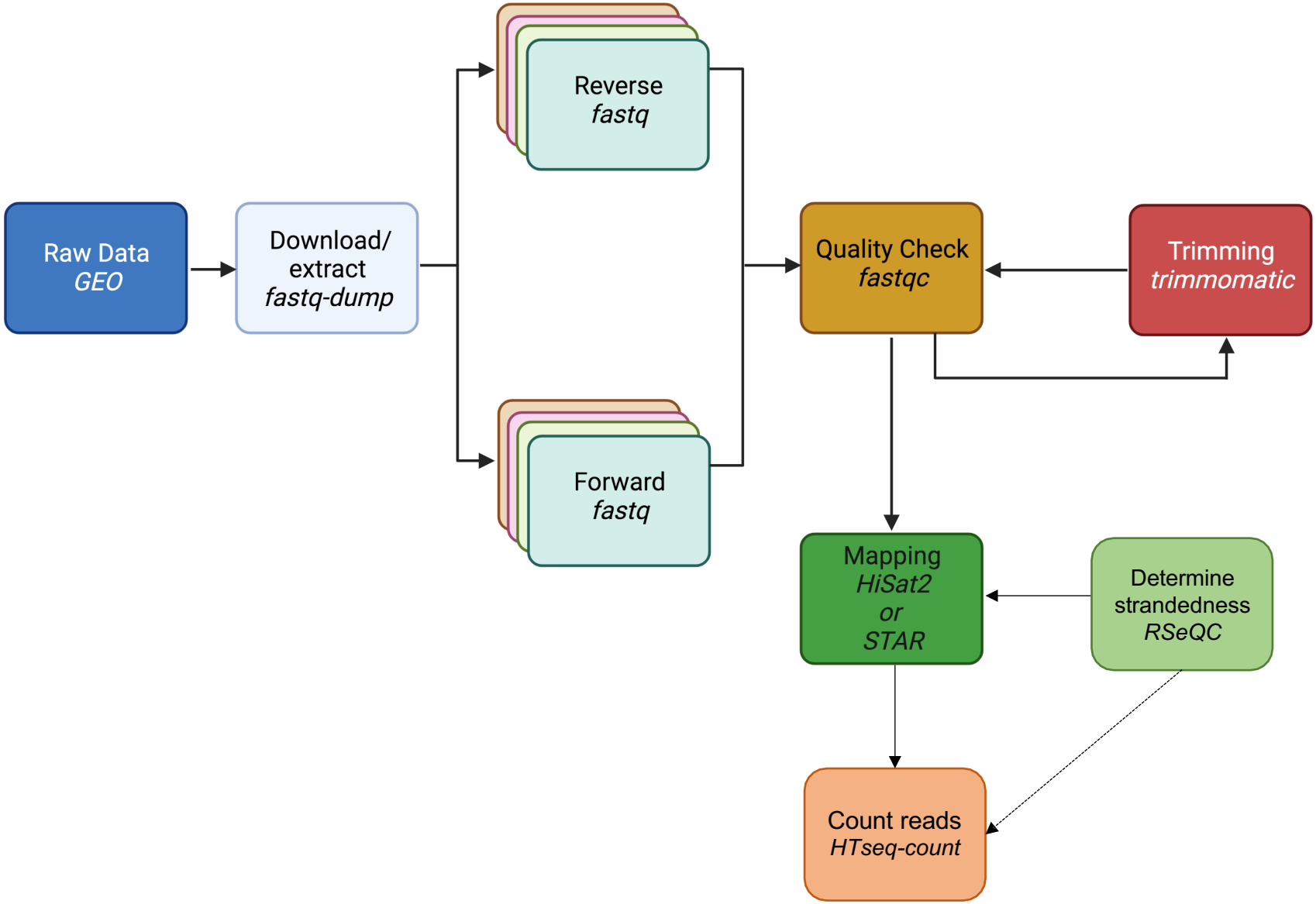
RSeQC & HTSeq

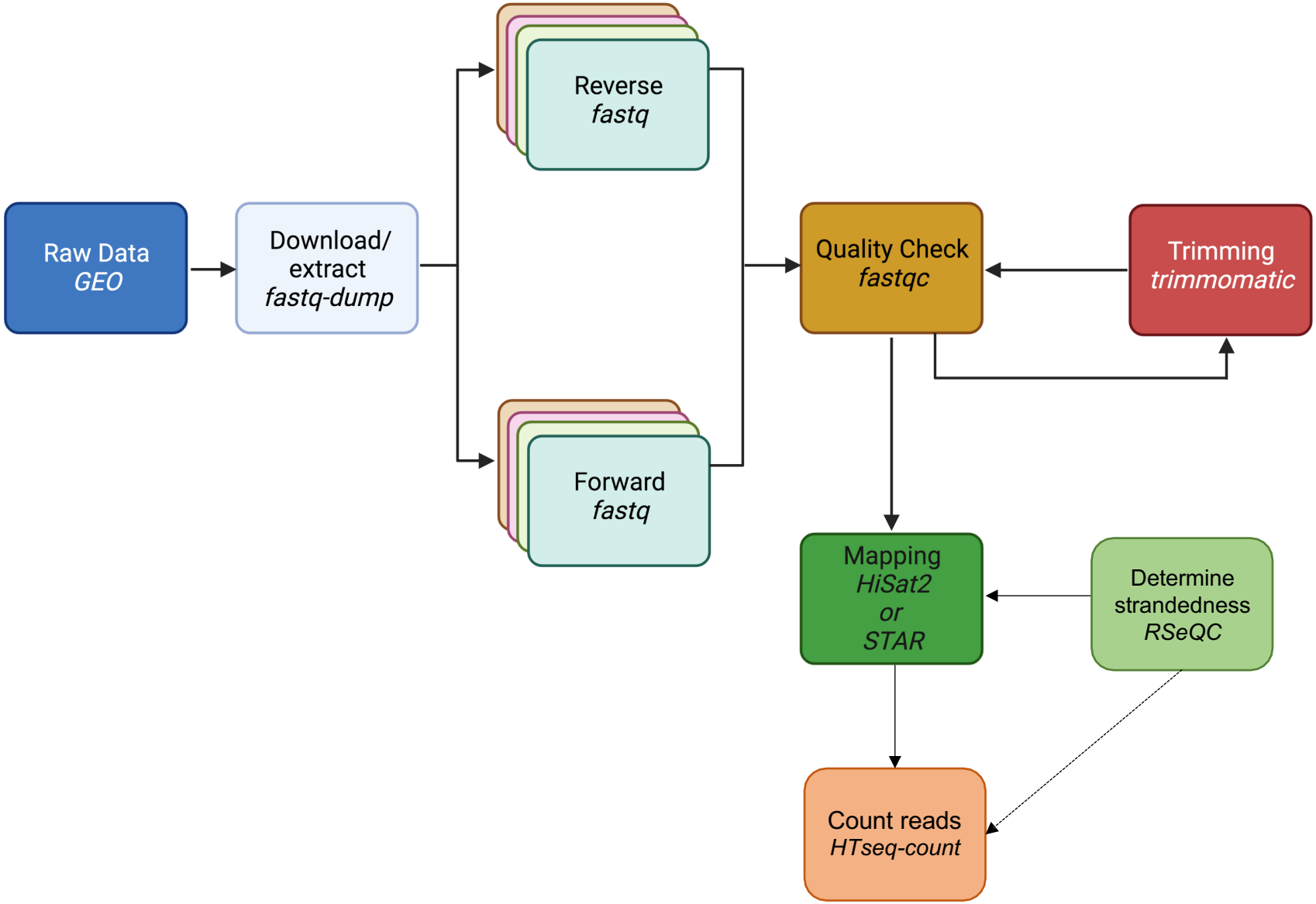
March 7th, 2024


Raw Data
GEO
Download/
extract
fastq-dump
Reverse
fastq
Forward
fastq
Quality Check
fastqc
Trimming
trimmomatic
Mapping
HiSat2
or
STAR


Raw Data
GEO
Download/
extract
fastq-dump
Reverse
fastq
Forward
fastq
Quality Check
fastqc
Trimming
trimmomatic
Mapping
HiSat2
or
STAR
Determine
strandedness
RSeQC
Count reads
HTseq-count

Install RSeQC

- We will install RSeQC using conda
- Conda is an open-source management system
- Conda quickly installs, runs, and updates packages and their dependencies
- For this installation we will be creating a ‘conda environment’ called rseqc
- To use rseqc program in the future, you will need to perform ‘conda activate rseqc’



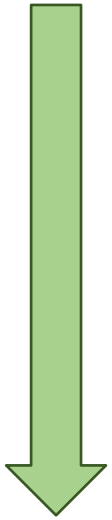
Take a break to install

Pre & post alignment QC

- **raw reads** QC
  - adapter/primer/other contaminating and over-represented sequences
  - sequencing quality
  - GC distributions
  - duplication levels

- **aligned reads** QC
  - % (uniquely) aligned reads
  - % exonic vs. intronic/intergenic
  - gene diversity
  - gene body coverage
  - strandedness

Pre-alignment: FastQC, fastp



Post-alignment: RSeQC, QoRTs

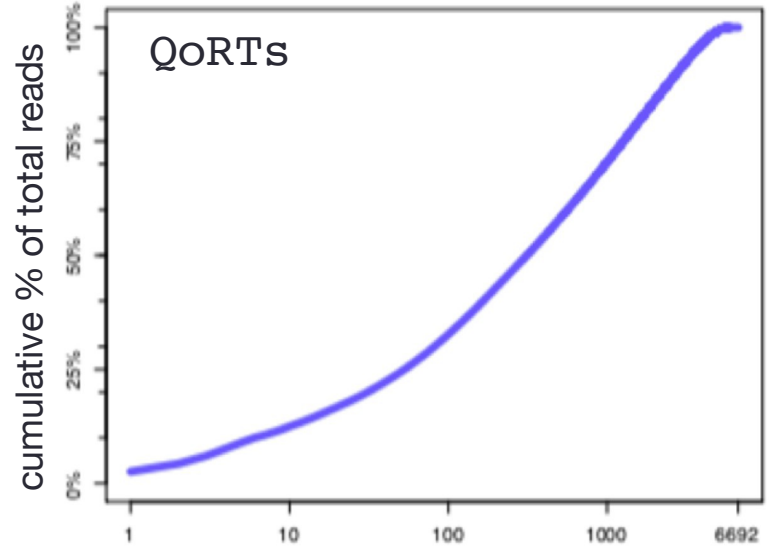
Pre QC Questions

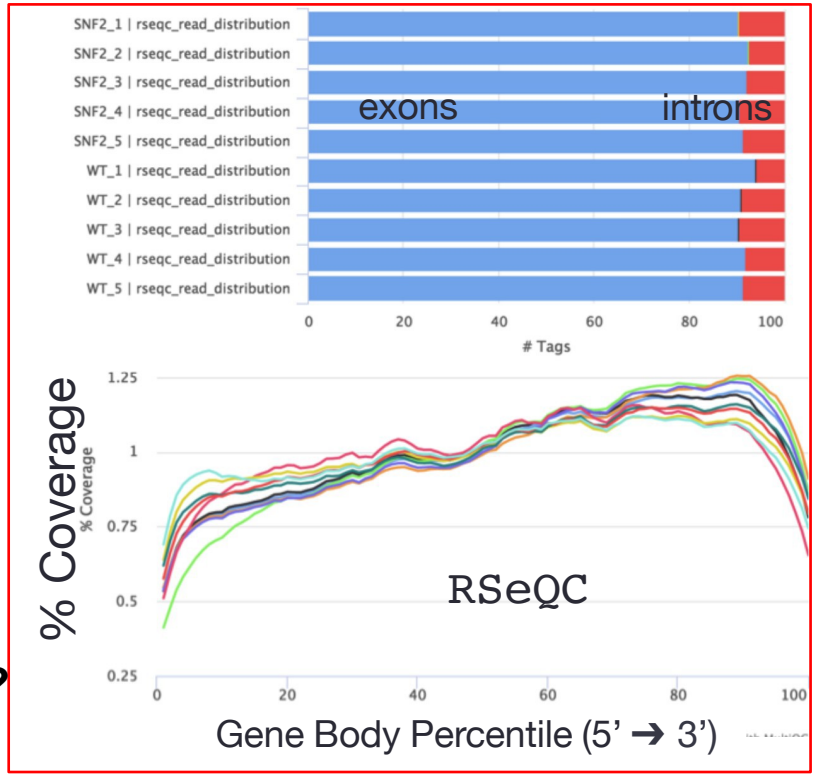
- Before mapping:
  - *How to identify and remove reads with low base calls?*
  - *How to identify and remove reads with linkers/adaptors?*
  - *How to screen for potential species/vector/ribosomal contamination?*
  - *How is your library complexity?*

Post-alignment QC

- lack of **gene diversity**:
  - dominance of rRNAs, tRNAs or other highly abundant transcripts
- **read distribution**
  - high intron coverage: incomplete poly(A) enrichment
  - many intergenic reads: gDNA contamination
- **gene body coverage**
  - 3' bias: RNA degradation + poly(A) enrichment

– ***What is percentage of reads aligned?***
– ***Is your sequencing library stranded or unstranded?***





Stranded libraries

- A major decision to be made during the library preparation step is whether to preserve RNA strand information.
- Unlike DNA molecules, RNA molecules exist as single-stranded threads that could result from the sense or antisense strand.
- The creation of stranded libraries are now standard with Illumina TruSeq ‘stranded’ RNA-Seq kits
- This means that with a great amount of certainty you can identify which strand of DNA the RNA was transcribed from

Why retain stranded information?

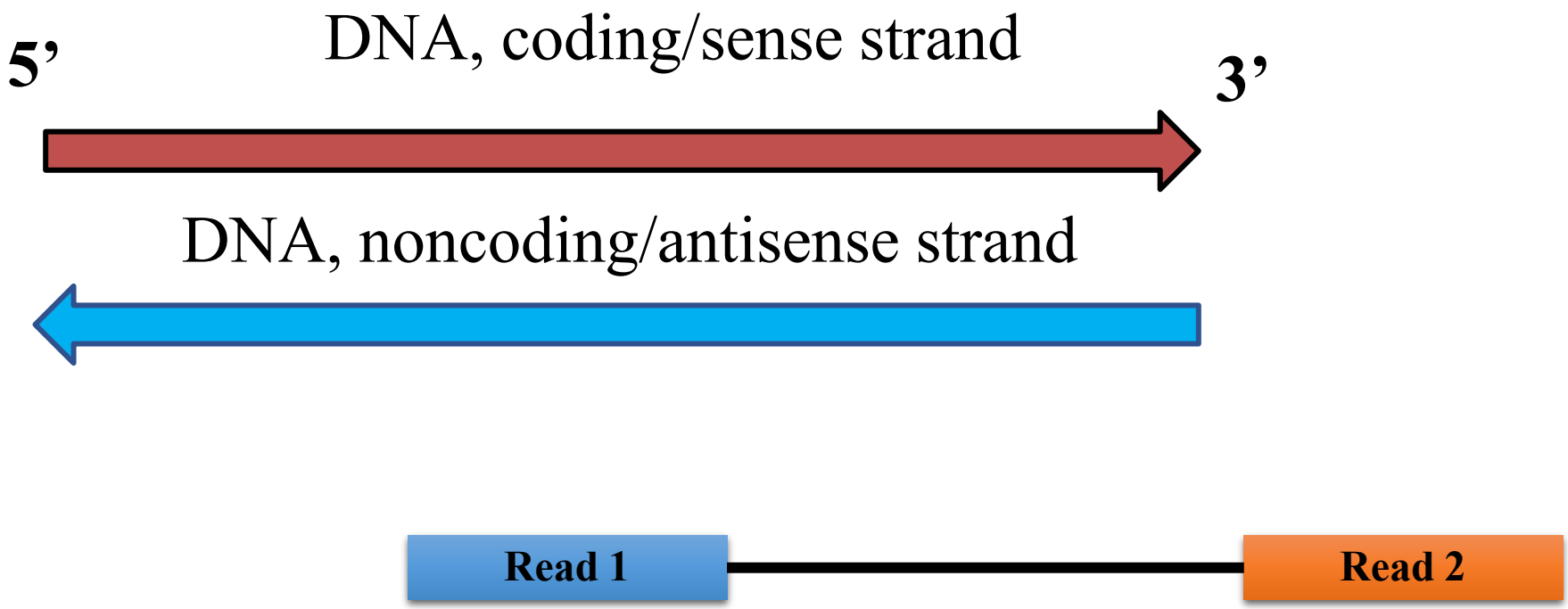
- It makes sense to begin with the most information possible – even if immediately that is not of interest
- Useful for identifying antisense transcripts, mapping splicing events, and detecting overlapping transcripts.
- They are commonly used in studies of transcriptomics, gene expression analysis, and RNA editing, and *de novo* assembly.

Why is this important to determine prior to counting?

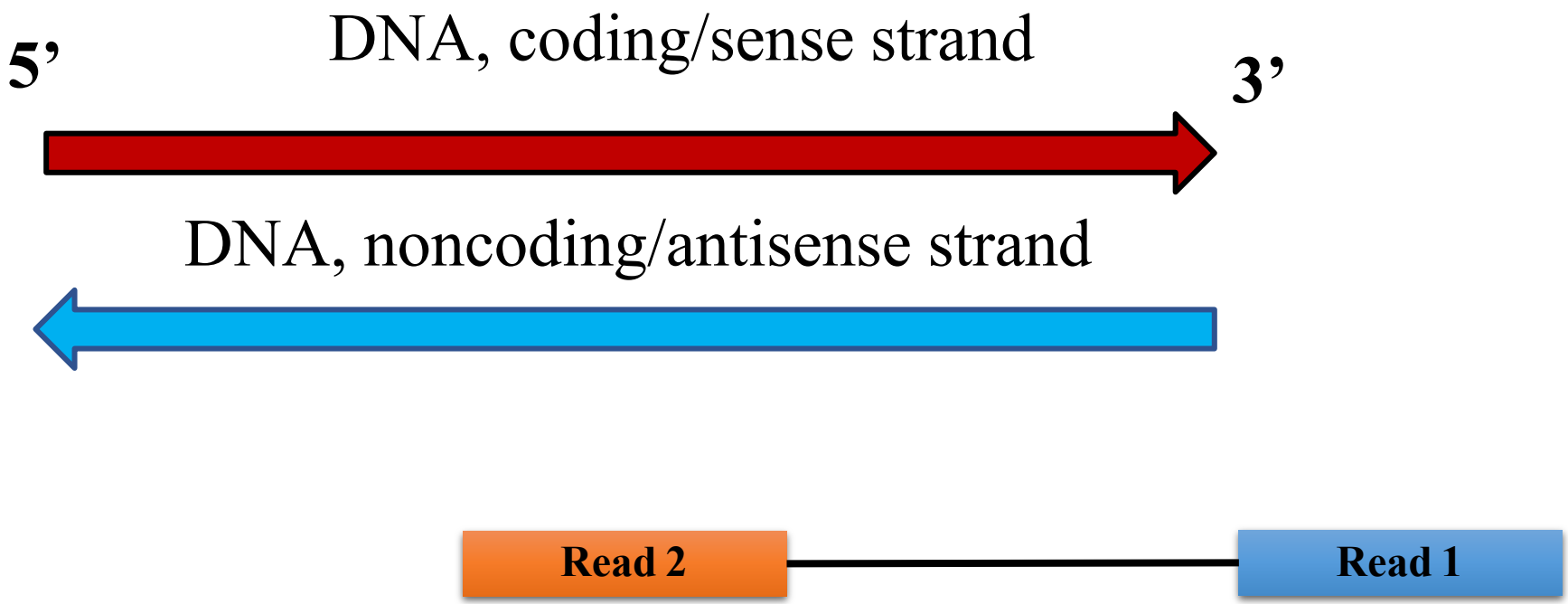
- If you use wrong directionality parameter in the read counting step with HTSeq, the reads are considered to be from the wrong strand.
- This means you won't get any counts, and if there is a gene in the same location on the other strand, your reads are counted for the ***wrong gene***.
- So its important to check, if you are unsure, using tools!

Three scenarios when it comes to stranded libraries

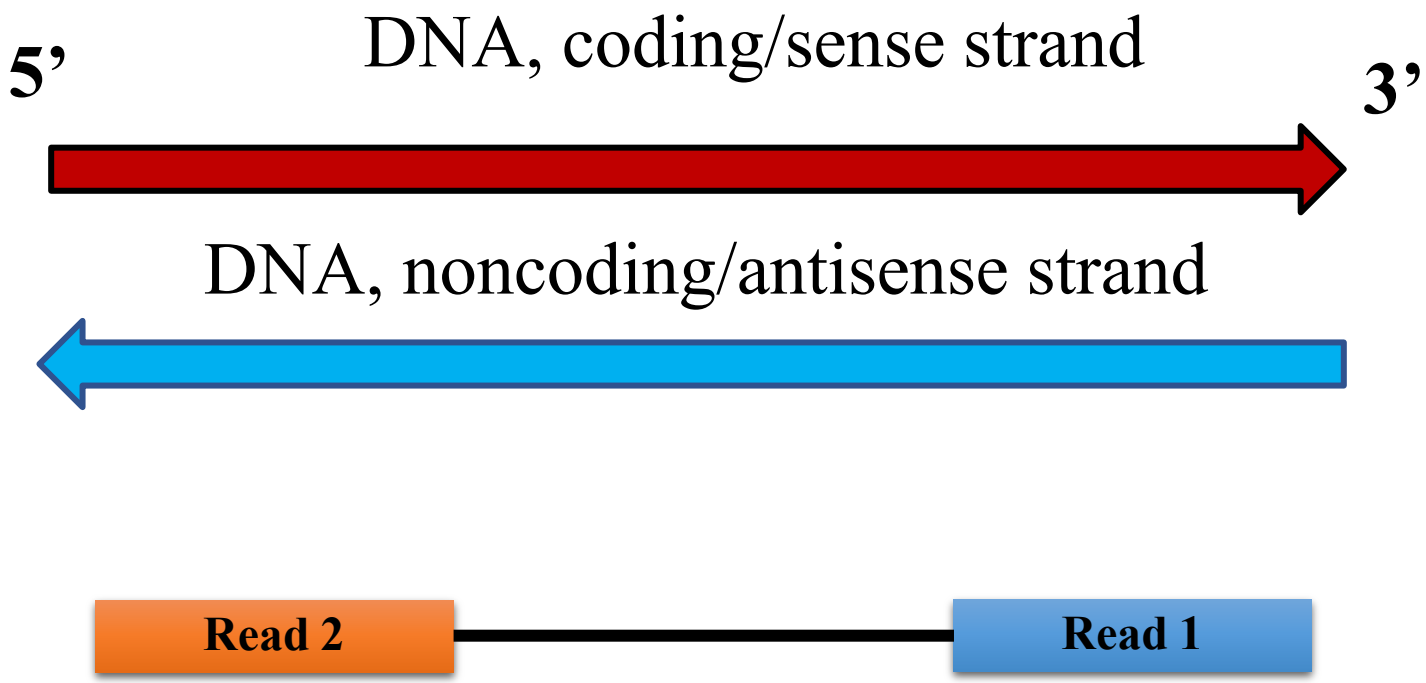
- Forward (secondstrand) – reads resemble the gene sequence
- Reverse (firststrand) – reads resemble the complementary sequence
- Unstranded



If sequences of Read 1 align to the coding, sense strand – the library is "stranded"



If sequences of Read 2 align to the coding, sense strand – the library is " reverse stranded"



If both Read 1 and Read 2 align to the coding, sense strand – the library is "unstranded"

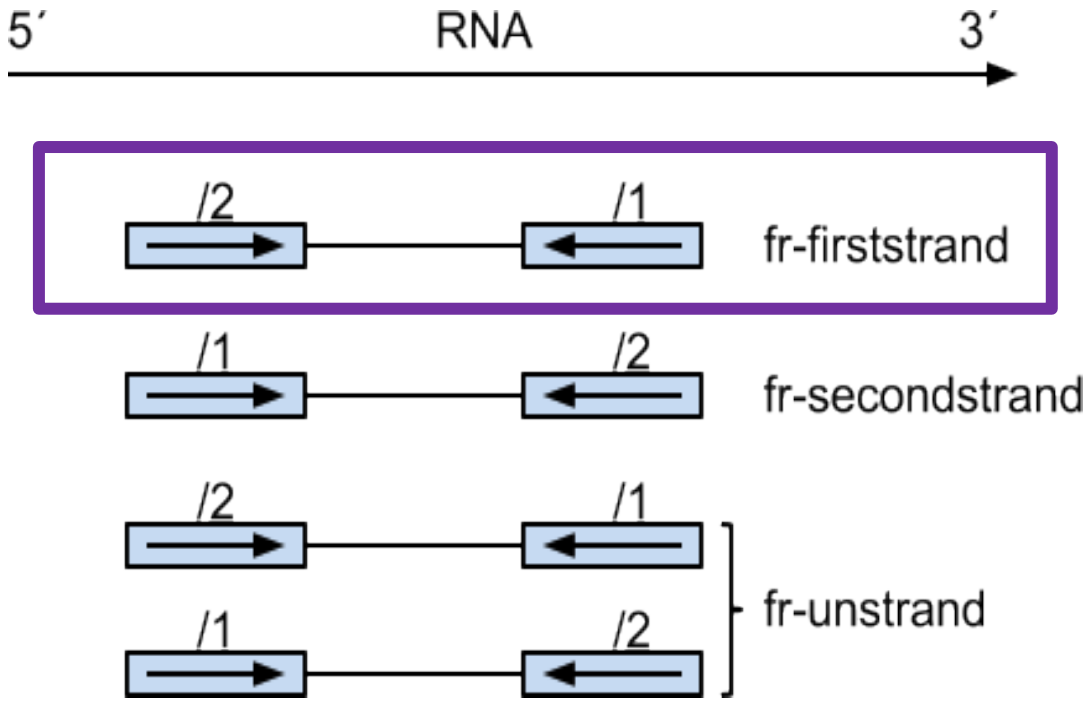
Different tools have different names for stranded settings:

	Option 1 **RF/fr-firststrand** **Reverse**	**Option 2** **FR/fr-secondstrand** **Stranded**	**Option 3** **Unstranded**
HISAT2 (--rna-strandedness)	RF (for PE) R (for SE)	FR (for PE) F (for SE)	Default
STAR	n/a	n/a	n/a
SALMON	-I ISR	-I ISR	-I IU
HTSeq	stranded=reverse	stranded=yes	stranded=no
Methods or Kits	dUTP Illumina TruSeq NEBNext Ultra II	Ligation Standard SOLID, NuGEN, 10X	Standard Illumina NuGEN, SMARTer

Option 1
RF/fr-firststrand

	Option 1 RF/fr-firststrand	Option 2 FR/fr-secondstrand	Option 3 Unstranded
HISAT2 (--rna-strandedness)	RF (for PE) R (for SE)	FR (for PE) F (for SE)	Default
STAR	n/a	n/a	n/a
SALMON	-I ISR	-I ISR	-I IU
HTSeq	**stranded=reverse**	stranded=yes	stranded=no
Methods or Kits	dUTP Illumina TruSeq NEBNext Ultra II	Ligation Standard SOLID, NuGEN, 10X	Standard Illumina NuGEN, SMARTer

Infer_experiment.py pair-end RNA-seq



The second read (read 2) is from the original RNA strand/template, first read (read 1) is from the opposite strand.

Fraction of reads explained by "1++,1--,2+-,2-+": 0.0169

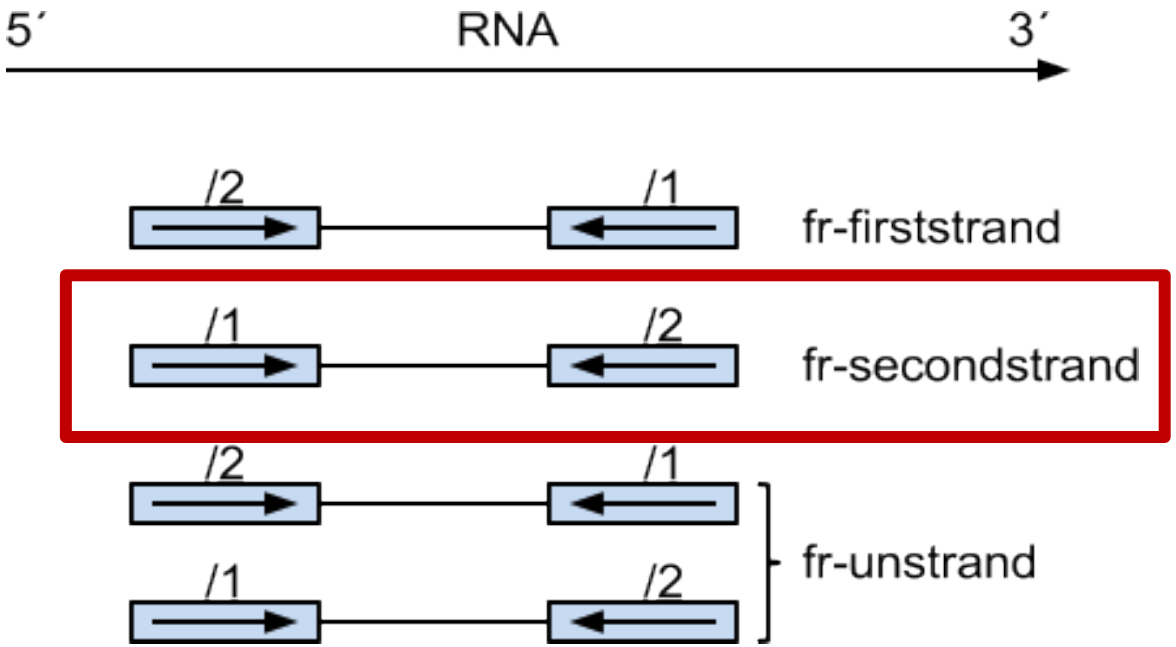
Fraction of reads explained by "1+-,1-+,2++,2-- ": 0.8827

Strand-specific pair-end RNA-seq data using dUTP protocol

Option 2
FR/fr-secondstrand

	Option 1 RF/fr-firststrand	Option 2 FR/fr-secondstrand	Option 3 Unstranded
HISAT2 (--rna-strandedness)	RF (for PE) R (for SE)	FR (for PE) F (for SE)	Default
STAR	n/a	n/a	n/a
SALMON	-I ISR	-I ISR	-I IU
HTSeq	stranded=reverse	**stranded=yes**	stranded=no
Methods or Kits	dUTP Illumina TruSeq NEBNext Ultra II	Ligation Standard SOLID, NuGEN, 10X	Standard Illumina NuGEN, SMARTer

Infer_experiment.py pair-end RNA-seq



The first read (read 1) is from the original RNA strand/template, second read (read 2) is from the opposite strand.

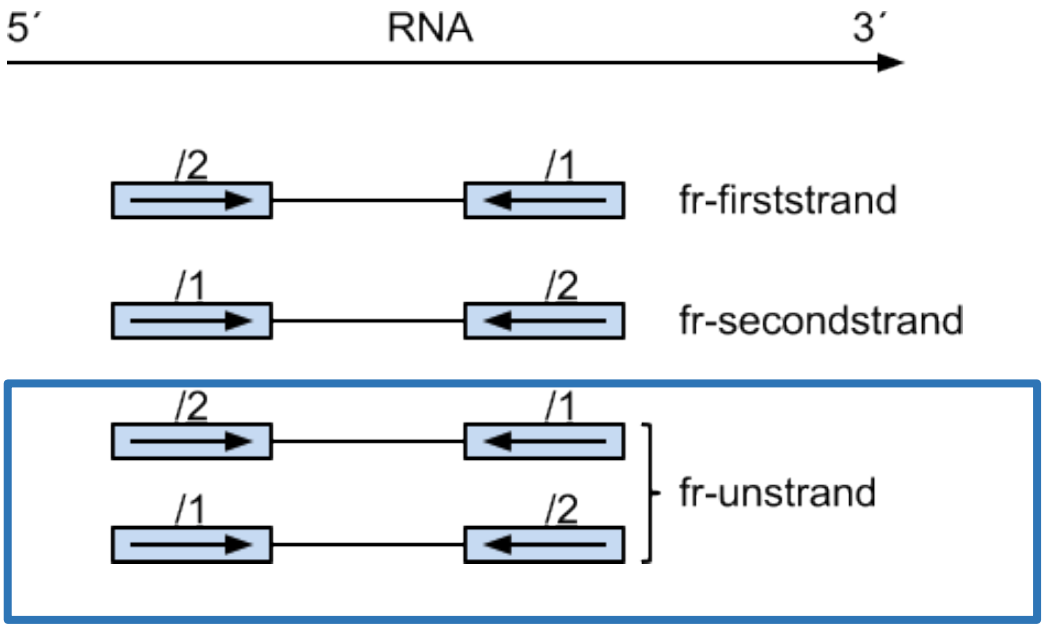
Fraction of reads explained by "1++,1–,2+-,2-+": 0.9807
Fraction of reads explained by "1+-,1-+,2++,2-- ": 0.0193

Strand-specific pair-end RNA-seq data using Ligation protocol

Option 3
Unstranded

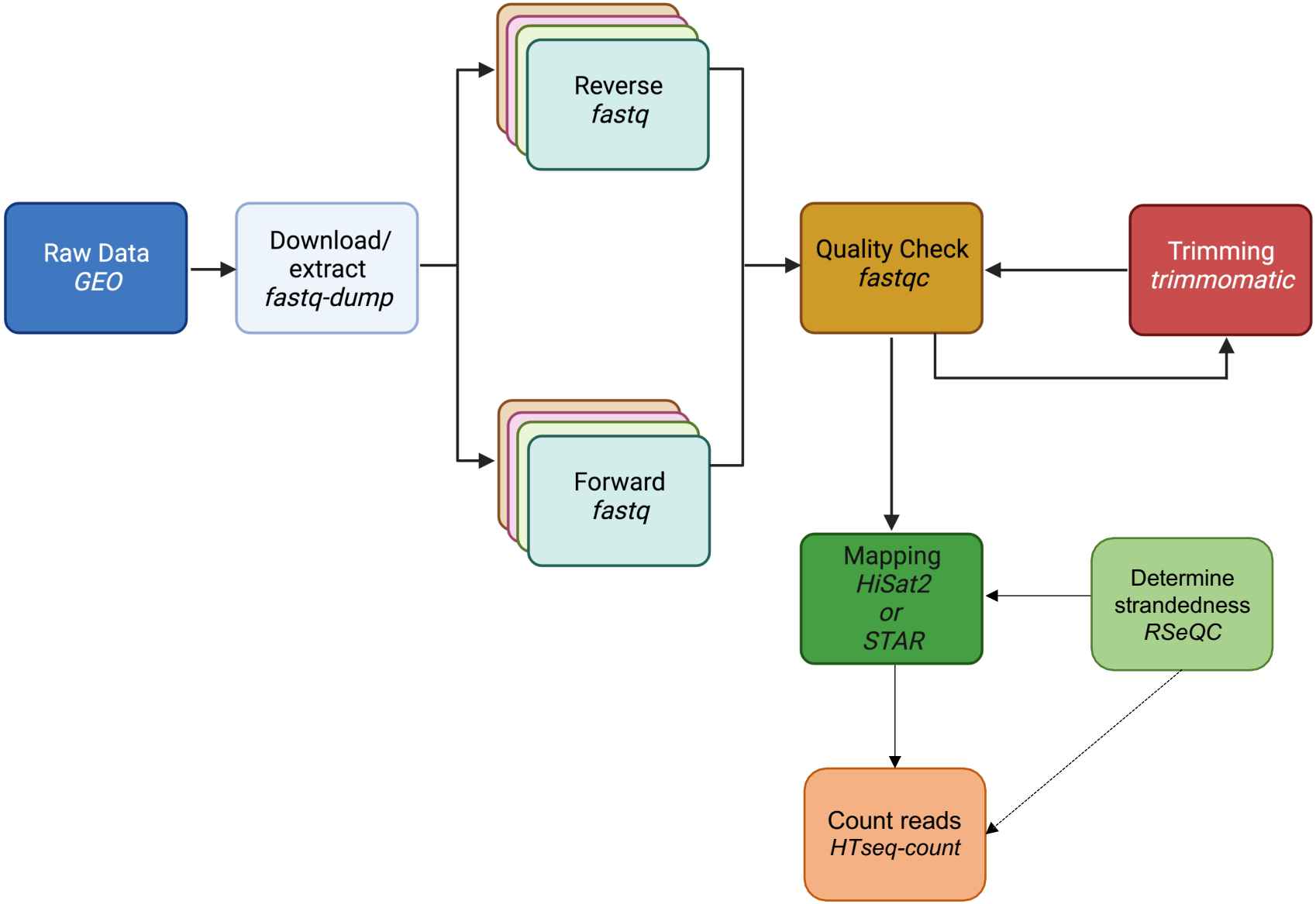
	Option 1 RF/fr-firststrand	Option 2 FR/fr-secondstrand	Option 3 Unstranded
HISAT2	RF (for PE) R (for SE)	FR (for PE) F (for SE)	Default
STAR	n/a	n/a	n/a
SALMON	-I ISR	-I ISR	-I IU
HTSeq	stranded=reverse	stranded=yes	**stranded=no**
Methods or Kits	dUTP Illumina TruSeq NEBNext Ultra II	Ligation Standard SOLID, NuGEN, 10X	Standard Illumina NuGEN, SMARTer

Infer_experiment.py pair-end RNA-seq



Fraction of reads failed to determine: 0.0648
Fraction of reads explained by "1++,1--,2+-,2-+": 0.4590
Fraction of reads explained by "1+-,1-+,2++,2--": 0.4763

Information regarding the strand is not conserved
(it is lost during the amplification of the mRNA fragments).



Take a break to run RSeQC to infer strandedness

Is your library stranded or not stranded?

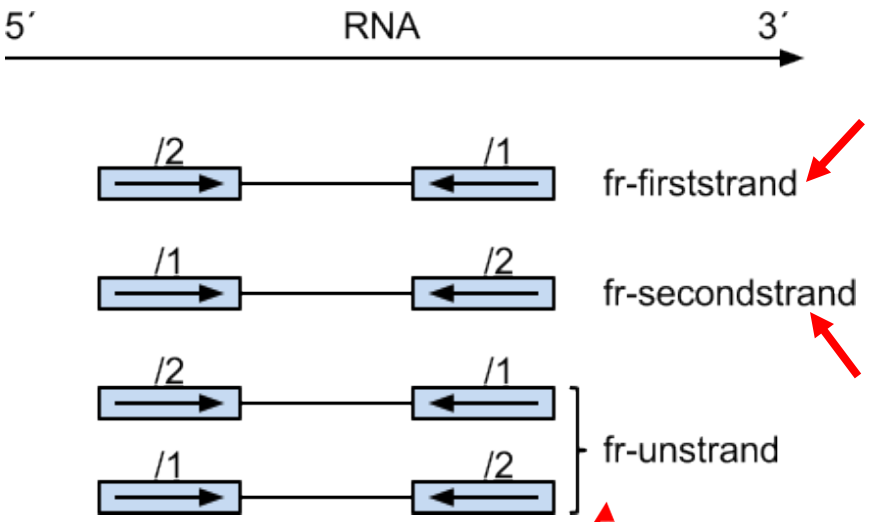
- RSeQC (http://rseqc.sourceforge.net/)

- `infer_experiment.py -i sample.bam -r gene_model.bed`

What would you choose for the unknown?

	Option 1 RF/fr-firststrand	Option 2 FR/fr-secondstrand	Option 3 Unstranded
HISAT2	RF (for PE) R (for SE)	FR (for PE) F (for SE)	Default
STAR	n/a	n/a	n/a
SALMON	-I ISR	-I ISR	-I IU
HTSeq	stranded=reverse	stranded=yes	stranded=no
Methods or Kits	dUTP Illumina TruSeq NEBNext Ultra II	Ligation Standard SOLID, NuGEN, 10X	Standard Illumina NuGEN, SMARTer

Summary

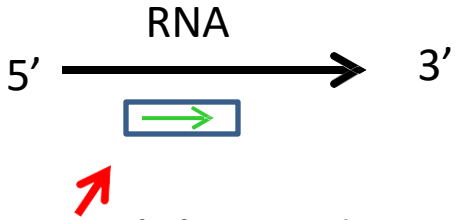


Fraction of reads explained by "1++,1--,2+-,2-+": 0.0193
Fraction of reads explained by "1+-,1-+,2++,2--": 0.8827

Fraction of reads explained by "1++,1--,2+-,2-+": 0.9807
Fraction of reads explained by "1+-,1-+,2++,2--": 0.0193

Fraction of reads failed to determine: 0.0648
Fraction of reads explained by "1++,1--,2+-,2-+":0.4590
Fraction of reads explained by "1+-,1-+,2++,2--": 0.4763

Infer_experiment.py
single-end RNA-seq



Two different ways to strand reads:

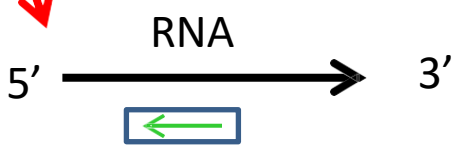
i) ++,--

read mapped to '+' strand indicates parental gene on '+' strand
read mapped to '-' strand indicates parental gene on '-' strand

ii) +-,-+

read mapped to '+' strand indicates parental gene on '-' strand
read mapped to '-' strand indicates parental gene on '+' strand

Strand-specific example:

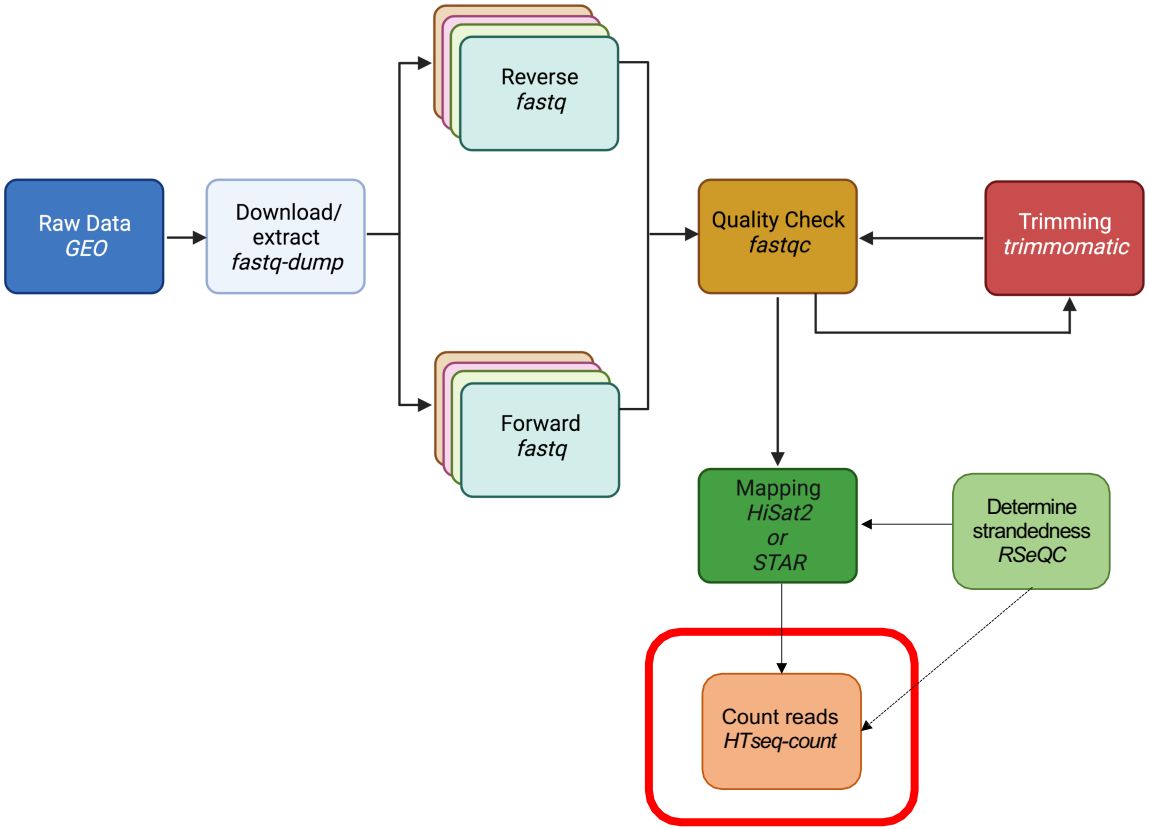


Fraction of reads failed to determine: 0.0170
Fraction of reads explained by "++,--": 0.9669
Fraction of reads explained by "+-,-+": 0.0161

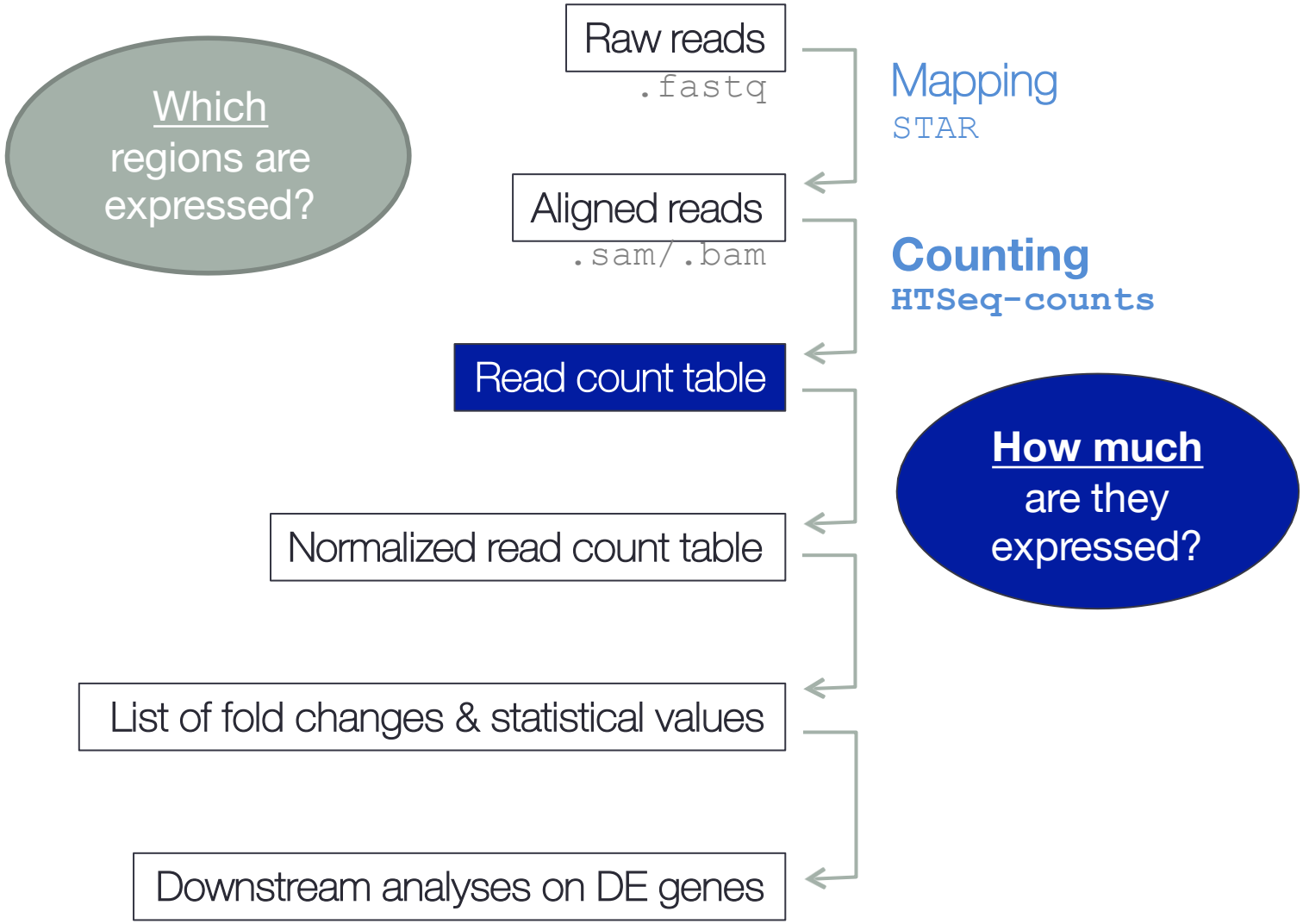
FR/fr-secondstrand
stranded=yes

CLASS ACTIVITY #2

COUNTING READS



Bioinformatics workflow of RNA-seq analysis

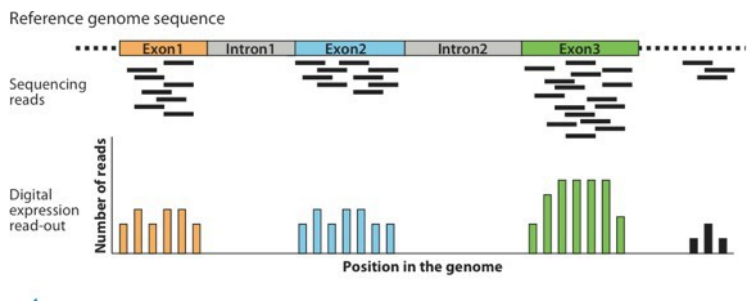


Gene counting programs

- HTSeq (Anders et al.2015, Bioinformatics 31:2)
- Cufflinks (Trapnell et al, 2010, Nat Biotech 28:5)
- StringTie (Pertea et al. 2015, Nat Biotech 33:3)
- featureCounts

We are using HTSeq as this approach will obtain gene-level quantification by directly overlapping with gene loci

Counting per-gene alignments



	sample1	sample2	sample3	sample4	...
gene1	999	701	616	595	
gene2	532	520	41	26	
gene3	14	36	305	322	
...					

- **HTSeq** package
  - Anders, Pyl & Huber, 2015, *Bioinformatics 31:*2
  - Hompage at https://htseq.readthedocs.io/
  - Allows *per-exon* counts
  - Designed for *differential gene expression testing*
  - Includes the **`htseq-count`** command

Each column is a sample

Each row is a gene

GENE ID	KD.2	KD.3	OE.1	OE.2	OE.3	IR.1	IR.2	IR.3
1/2-SBSRNA4	57	41	64	55	38	45	31	39
A1BG	71	40	100	81	41	77	58	40
A1BG-AS1	256	177	220	189	107	213	172	126
A1CF	0	1	1	0	0	0	0	0
A2LD1	146	81	138	125	52	91	80	50
A2M	10	9	2	5	2	9	8	4
A2ML1	3	2	6	5	2	2	1	0
A2MP1	0	0	2	1	3	0	2	1
A4GALT	56	37	107	118	65	49	52	37
A4GNT	0	0	0	0	1	0	0	0
AA06	0	0	0	0	0	0	0	0
AAA1	0	0	1	0	0	0	0	0
AAAS	2288	1363	1753	1727	835	1672	1389	1121
AACS	1586	923	951	967	484	938	771	635
AACSP1	1	1	3	0	1	1	1	3
AADAC	0	0	0	0	0	0	0	0
AADACL2	0	0	0	0	0	0	0	0
AADACL3	0	0	0	0	0	0	0	0
AADACL4	0	0	1	1	0	0	0	0
AADAT	856	539	593	576	359	567	521	416
AAGAB	4648	2550	2648	2356	1481	3265	2790	2118
AAK1	2310	1384	1869	1602	980	1675	1614	1108
AAMP	5198	3081	3179	3137	1721	4061	3304	2623
AANAT	7	7	12	12	4	6	2	7
AARS	5570	3323	4782	4580	2473	3953	3339	2666

Counting features with htseq-count

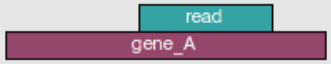
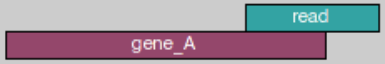
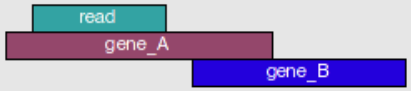
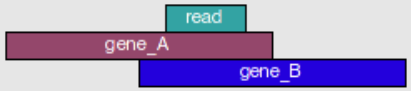
- What features are of interest? Gene, transcript, and/or exon counts?

  type=exon

- What happens if a read overlaps with multiple features?

  mode=union

- Is the RNA stranded, reversed strand, or unstranded?

	union	intersection _strict	intersection _nonempty
	gene_A	gene_A	gene_A
	gene_A	no_feature	gene_A
	gene_A	no_feature	gene_A
	gene_A	gene_A	gene_A
	gene_A	gene_A	gene_A
	ambiguous	gene_A	gene_A
	ambiguous	ambiguous	ambiguous

CLASS ACTIVITY #3