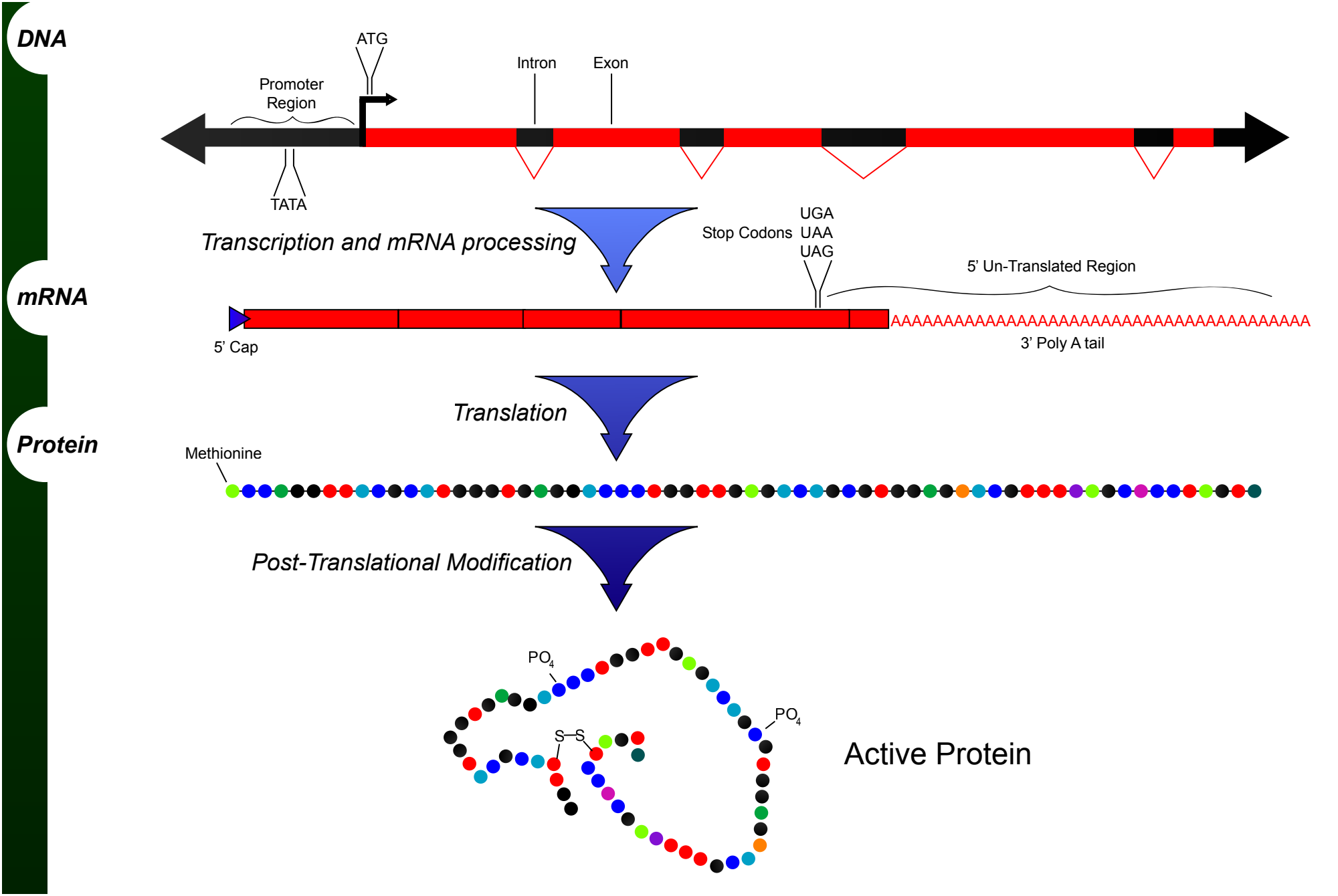
Characterizing transcriptomes using ngs data

T. Källman

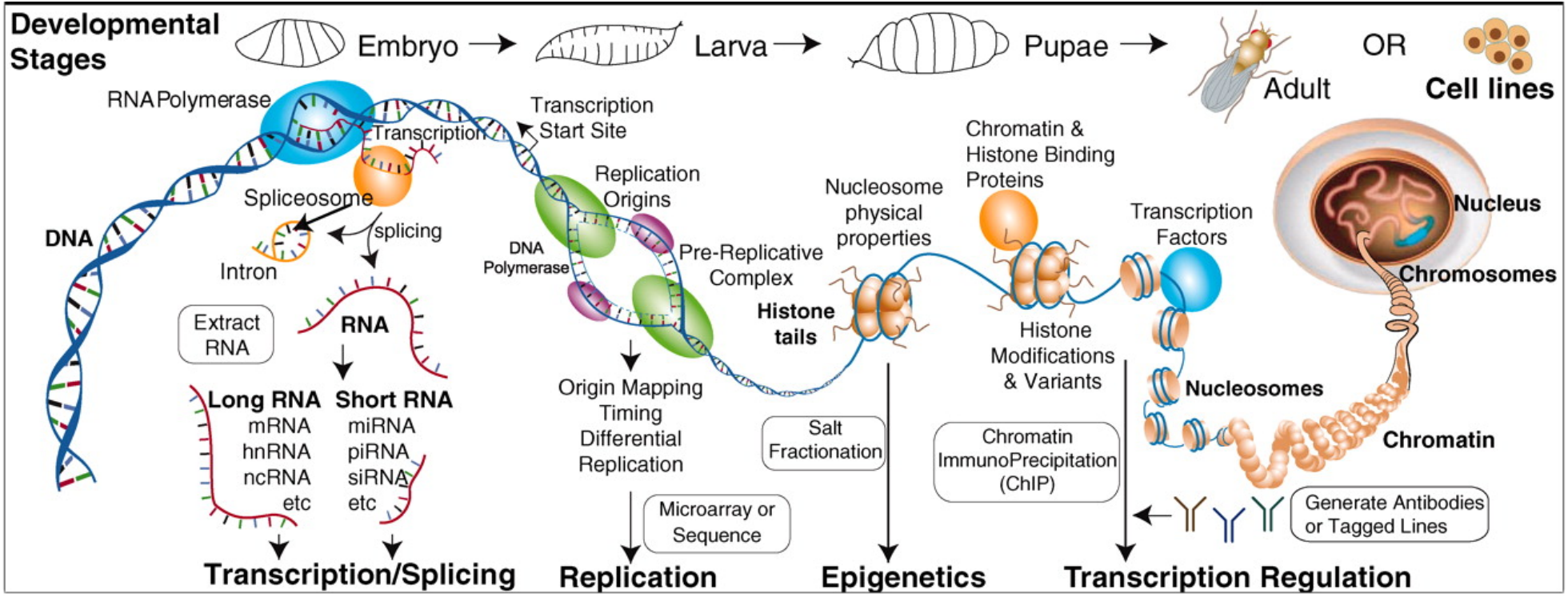
BILS/Scilife Lab/Uppsala University

Sep. 2015

Outline

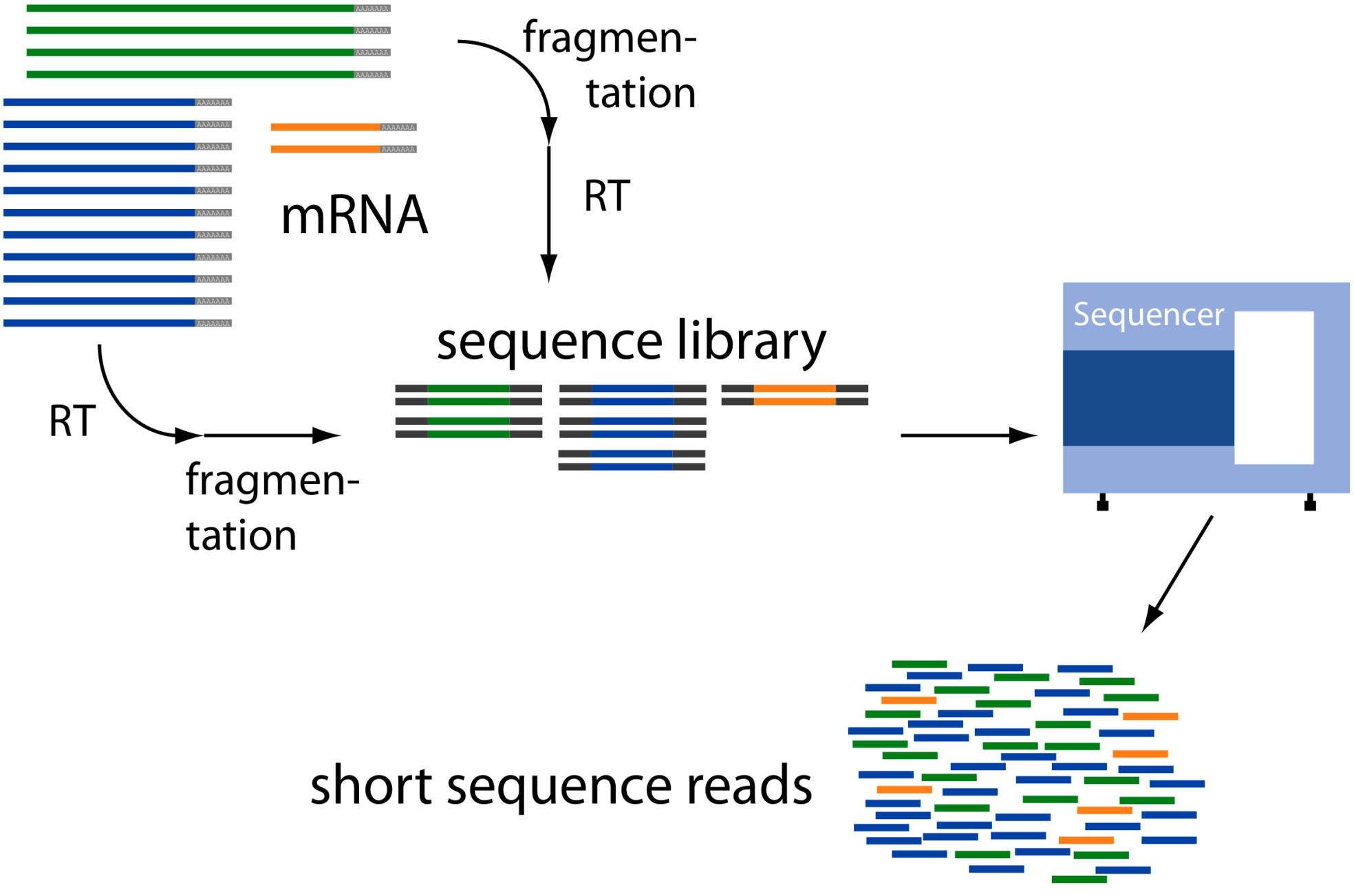

DNA
ATG
Promoter Region
Intron
Exon
TATA
Transcription and mRNA processing
Stop Codons
UGA
UAA
UAG
5' Un-Translated Region
mRNA
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
5' Cap
3' Poly A tail
Translation
Protein
Methionine
Post-Translational Modification
PO4
PO4
S—S
Active Protein

ral Dogma

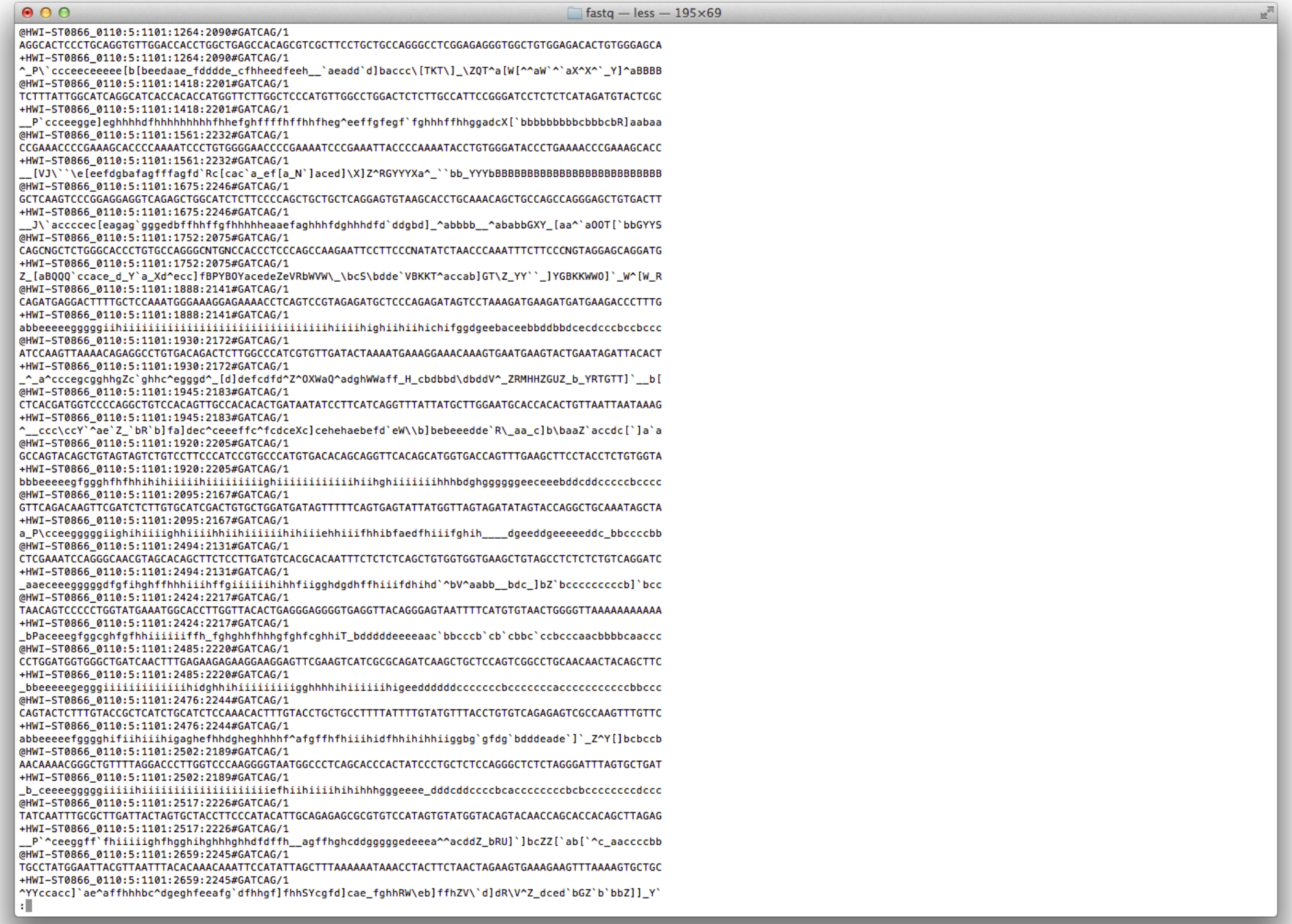

Developmental Stages
Embryo
Larva
Pupae
Adult
OR
Cell lines
RNA Polymerase
Transcription
Transcription Start Site
Spliceosome
splicing
DNA
Intron
Extract RNA
RNA
Long RNA
mRNA
hnRNA
ncRNA
etc
Short RNA
miRNA
piRNA
siRNA
etc
Transcription/Splicing
Replication Origins
DNA Polymerase
Pre-Replicative Complex
Origin Mapping
Timing
Differential Replication
Microarray or Sequence
Replication
Nucleosome physical properties
Histone tails
Salt Fractionation
Epigenetics
Chromatin & Histone Binding Proteins
Histone Modifications & Variants
Chromatin ImmunoPrecipitation (ChIP)
Transcription Factors
Nucleosomes
Chromatin
Nucleus
Chromosomes
Generate Antibodies or Tagged Lines
Transcription Regulation

omplex view

omes vs genomes


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

RNA sequence technologies


fastq — less — 195×69
@HWI-ST0866_0110:5:1101:1264:2090#GATCAG/1
AGGCACTCCCTGCAGGTGTTGGACCACCTGGCTGAGCCACAGCGTCGCTTCCTGCTGCCAGGGCCTCGGAGAGGGTGGCTGTGGAGACACTGTGGGAGCA
+HWI-ST0866_0110:5:1101:1264:2090#GATCAG/1
^_P\`ccceeceeeee[b[beedaae_fdddde_cfhheedfeeh__`aeadd`d]baccc\[TKT\]_\ZQT^a[W[^^aW`^`aX^X^`_Y]^aBBBB
@HWI-ST0866_0110:5:1101:1418:2201#GATCAG/1
TCTTTATTGGCATCAGGCATCACCACACCATGGTTCTTGGCTCCCATGTTGGCCTGGACTCTCTTGCCATTCCGGGATCCTCTCTCATAGATGTACTCGC
+HWI-ST0866_0110:5:1101:1418:2201#GATCAG/1
__P`cceeegge]eghhhhdfhhhhhhhhhfhhefghffffhffhhfheg^eeffgfegf`fghhhffhhggadcX[`bbbbbbbbbcbbbcbR]aabaa
@HWI-ST0866_0110:5:1101:1561:2232#GATCAG/1
CCGAAACCCCGAAAGCACCCCAAAATCCCTGTGGGGAACCCCGAAAATCCCGAAATTACCCCAAAATACCTGTGGGATACCCTGAAAACCCGAAAGCACC
+HWI-ST0866_0110:5:1101:1561:2232#GATCAG/1
__[VJ\``\e[eefdgbafagfffagfd`Rc[cac`a_ef[a_N`]aced]\X]Z^RGYYYXa^_``bb_YYYbBBBBBBBBBBBBBBBBBBBBBBBBBB
@HWI-ST0866_0110:5:1101:1675:2246#GATCAG/1
GCTCAAGTCCCGGAGGAGGTCAGAGCTGGCATCTCTTCCCCAGCTGCTGCTCAGGAGTGTAAGCACCTGCAAACAGCTGCCAGCCAGGGAGCTGTGACTT
+HWI-ST0866_0110:5:1101:1675:2246#GATCAG/1
__J\`accccec[eagag`gggedbffhhffgfhhhhheaaefaghhhfdghhhdfd`ddgbd]_^abbbb__^ababbGXY_[aa^`aOOT[`bbGYYS
@HWI-ST0866_0110:5:1101:1752:2075#GATCAG/1
CAGCNGCTCTGGGCACCCTGTGCCAGGGCNTGNCCACCCTCCCAGCCAAGAATTCCTTCCCNATATCTAACCCAAATTTCTTCCCNGTAGGAGCAGGATG
+HWI-ST0866_0110:5:1101:1752:2075#GATCAG/1
Z_[aBQQQ`ccace_d_Y`a_Xd^ecc]fBPYBOYacedeZeVRbWVW_\bcS\bdde`VBKKT^accab]GT\Z_YY``_]YGBKKWWO]`_W^[W_R
@HWI-ST0866_0110:5:1101:1888:2141#GATCAG/1
CAGATGAGGACTTTTGCTCCAAATGGGAAAGGAGAAAACCTCAGTCCGTAGAGATGCTCCCAGAGATAGTCCTAAAGATGAAGATGATGAAGACCCTTTG
+HWI-ST0866_0110:5:1101:1888:2141#GATCAG/1
abbeeeeegggggiihiiiiiiiiiiiiiiiiiiiiiiiiiiiiiiihiiiihighiihiihichifgggdgeebaceebbddbbdcecdcccbccbccc
@HWI-ST0866_0110:5:1101:1930:2172#GATCAG/1
ATCCAAGTTAAAACAGAGGCCTGTGACAGACTCTTGGCCCATCGTGTTGATACTAAAATGAAAGGAAACAAAGTGAATGAAGTACTGAATAGATTACACT
+HWI-ST0866_0110:5:1101:1930:2172#GATCAG/1
_^_a^cccegcgghhgZc`ghhc^egggd^_[d]defcdfd^Z^OXWaQ^adghWWaff_H_cbdbbd\dbddV^_ZRMHHZGUZ_b_YRTGTT]`__b[
@HWI-ST0866_0110:5:1101:1945:2183#GATCAG/1
CTCACGATGGTCCCCAGGCTGTCCACAGTTGCCACACACTGATAATATCCTTCATCAGGTTTATTATGCTTGGAATGCACCACACTGTTAATTAATAAAG
+HWI-ST0866_0110:5:1101:1945:2183#GATCAG/1
^__ccc\ccY`^ae`Z_`bR`b]fa]dec^ceeeffc^fcdceXc]cehehaebefd`eW\\b]bebeeedde`R_aa_c]b\baaZ`accdc[`]a`a
@HWI-ST0866_0110:5:1101:1920:2205#GATCAG/1
GCCAGTACAGCTGTAGTAGTCTGTCCTTCCCATCCGTGCCCATGTGACACAGCAGGTTCACAGCATGGTGACCAGTTTGAAGCTTCCTACCTCTGTGGTA
+HWI-ST0866_0110:5:1101:1920:2205#GATCAG/1
bbbeeeeegfgggghfhfhhihihiiiiihiiiiiiiiighiiiiiiiiiiiihiihghiiiiiiihhhbdghgggggeeceeebddcddccccbcccc
@HWI-ST0866_0110:5:1101:2095:2167#GATCAG/1
GTTCAGACAAGTTCGATCTCTTGTGCATCGACTGTGCTGGATGATAGTTTTTCAGTGAGTATTATGGTTAGTAGATATAGTACCAGGCTGCAAATAGCTA
+HWI-ST0866_0110:5:1101:2095:2167#GATCAG/1
a_P\cceeggggiighihiiiighhiiihhiihiiiiihihiiiehhiiifhhibfaedfhiiifghih____dgeeddgeeeeeddc_bbccccbb
@HWI-ST0866_0110:5:1101:2494:2131#GATCAG/1
CTCGAAATCCAGGGCAACGTAGCACAGCTTCTCCTTGATGTCACGCACAATTTCTCTCTCAGCTGTGGTGGTGAAGCTGTAGCCTCTCTCTGTCAGGATC
+HWI-ST0866_0110:5:1101:2494:2131#GATCAG/1
_aaeceeeggggdfgfihghffhhhiiihffgiiiiihihhfiigghdgdhffhiiifdhihd`^bV^aabb__bdc_]bZ`bccccccccb]`bcc
@HWI-ST0866_0110:5:1101:2424:2217#GATCAG/1
TAACAGTCCCCCTGGTATGAAATGGCACCTTGGTTACACTGAGGGAGGGGTGAGGTTACAGGGAGTAATTTTCATGTGTAACTGGGGTTAAAAAAAAAAA
+HWI-ST0866_0110:5:1101:2424:2217#GATCAG/1
_bPaceeegfggcghfgfhhiiiiiiffh_fghghhfhhhgfghfcghhiT_bddddeeeeaac`bbcccb`cb`cbbc`ccbcccaacbbbbcaaccc
@HWI-ST0866_0110:5:1101:2485:2220#GATCAG/1
CCTGGATGGTGGGCTGATCAACTTTGAGAAGAGAAGGAAGGAGTTCGAAGTCATCGCGCAGATCAAGCTGCTCCAGTCGGCCTGCAACAACTACAGCTTC
+HWI-ST0866_0110:5:1101:2485:2220#GATCAG/1
_bbeeeeegegggiiiiiiiiiiiihidghhihiiiiiiiiigghhhhihiiiiiihigeeddddddccccccbccccccaccccccccccbbccc
@HWI-ST0866_0110:5:1101:2476:2244#GATCAG/1
CAGTACTCTTTGTACCGCTCATCTGCATCTCCAAACACTTTGTACCTGCTGCCTTTTATTTTGTATGTTTACCTGTGTCAGAGAGTCGCCAAGTTTGTTC
+HWI-ST0866_0110:5:1101:2476:2244#GATCAG/1
abbeeeeefgggghifiihiiihigaghefhhdgheghhhhf^afgffhfhiiihidfhhihihhiiggbg`gfdg`bdddeade`]`_Z^Y[]bcbccb
@HWI-ST0866_0110:5:1101:2502:2189#GATCAG/1
AACAAAACGGGCTGTTTTAGGACCCTTGGTCCCAAGGGGTAATGGCCCTCAGCACCCACTATCCCTGCTCTCCAGGGCTCTCTAGGGATTTAGTGCTGAT
+HWI-ST0866_0110:5:1101:2502:2189#GATCAG/1
_b_ceeeeggggiiiiihiiiiiiiiiiiiiiiiiiefhiihiiiihihihhhgggeeee_dddcddccccbcaccccccccbcbccccccccdccc
@HWI-ST0866_0110:5:1101:2517:2226#GATCAG/1
TATCAATTTGCGCTTGATTACTAGTGCTACCTTCCCATACATTGCAGAGAGCGCGTGTCCATAGTGTATGGTACAGTACAACCAGCACCACAGCTTAGAG
+HWI-ST0866_0110:5:1101:2517:2226#GATCAG/1
__P`^ceeggff`fhiiiighfhgghihghhhghhdfdffh__agffhghcddgggggedeeea^^acddZ_bRU]`]bcZZ[`ab[`^c_aaccccbb
@HWI-ST0866_0110:5:1101:2659:2245#GATCAG/1
TGCCTATGGAATTACGTTAATTTACACAAACAAATTCCATATTAGCTTTAAAAAATAAACCTACTTCTAACTAGAAGTGAAAGAAGTTTAAAAGTGCTGC
+HWI-ST0866_0110:5:1101:2659:2245#GATCAG/1
^YYccacc]`ae^affhhhbc^dgeghfeeafg`dfhhgf]fhhSYcgfd]cae_fghhRW\eb]ffhZV\`d]dR\V^Z_dced`bGZ`b`bbZ]]_Y`
:

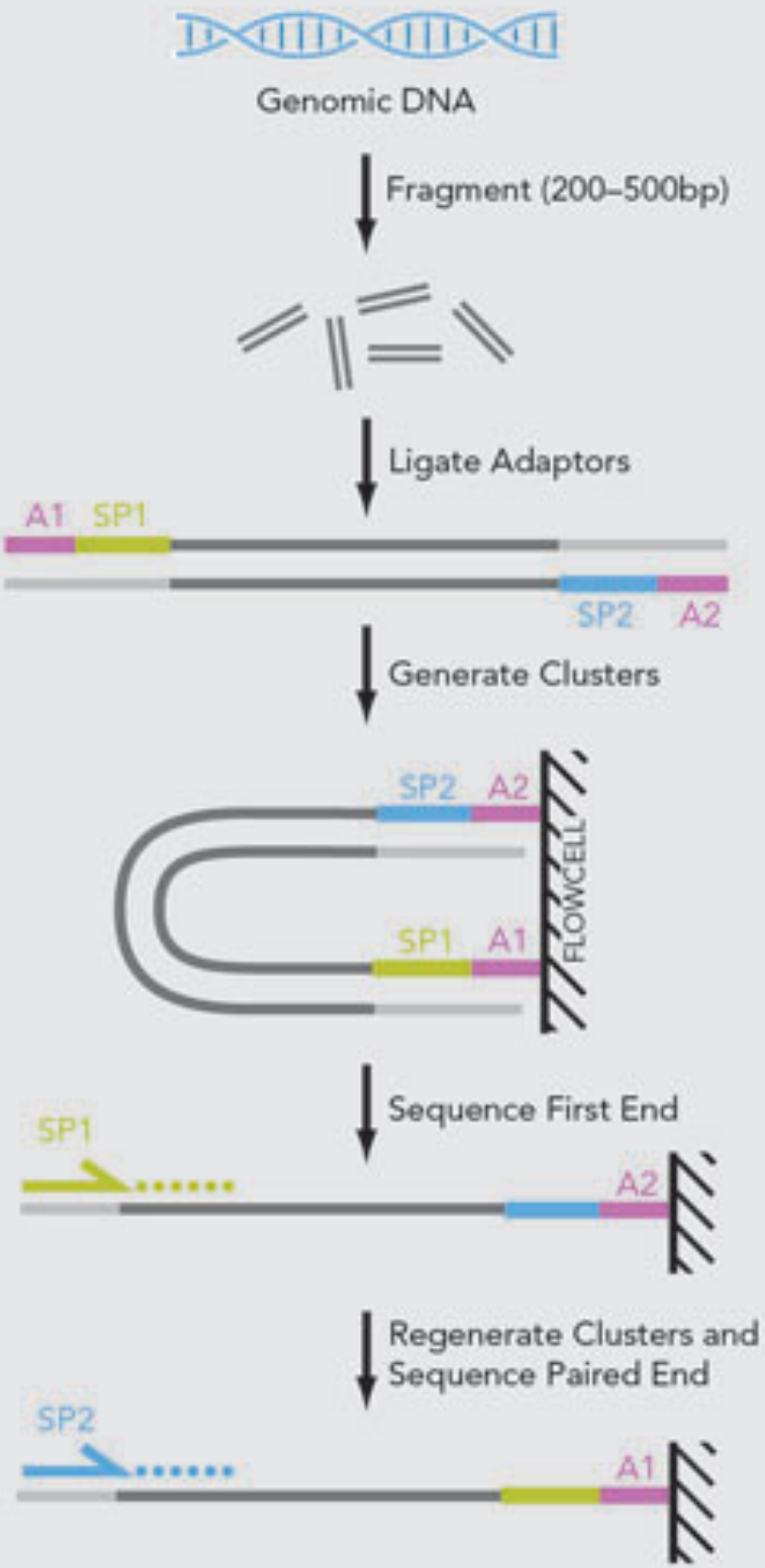
output

```
@SRR038845.3 HWI-EAS038:6:1:0:1938 length=36
CAACGAGTTCACACCTTGGCCGACAGGCCCGGGTAA
+SRR038845.3 HWI-EAS038:6:1:0:1938 length=36
BA@7>B=>:>>7@7@>>9=BAA?;>52;>:9=8.=A
@SRR038845.41 HWI-EAS038:6:1:0:1474 length=36
CCAATGATTTTTTTCCGTGTTTCAGAATACGGTTAA
+SRR038845.41 HWI-EAS038:6:1:0:1474 length=36
BCCBA@BB@BBBBAB@B9B@=BABA@A:@693:@B=
@SRR038845.53 HWI-EAS038:6:1:1:360 length=36
GTTCAAAAAGAACTAAATTGTGTCAATAGAAAACTC
+SRR038845.53 HWI-EAS038:6:1:1:360 length=36
BBCBBBBBB@@BAB?BBBBCBC>BBBAA8>BBBAA@
```

output


SSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSSS.....
.....XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX.....
.....IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII.....
.....JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJ.....
..LLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLLL.....
!"#$%&'()*+,-./0123456789:;<=>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ[\]^_`abcdefghijklmnopqrstuvwxyz{|}~
33 59 64 73 104 126
0.....26...31.....40
-5....0.....9.....40
0.....9.....40
3.....9.....40
0.2.....26...31.....41
S - Sanger Phred+33, raw reads typically (0, 40)
X - Solexa Solexa+64, raw reads typically (-5, 40)
I - Illumina 1.3+ Phred+64, raw reads typically (0, 40)
J - Illumina 1.5+ Phred+64, raw reads typically (3, 40)
with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (bold)
(Note: See discussion above).
L - Illumina 1.8+ Phred+33, raw reads typically (0, 41)

e quality

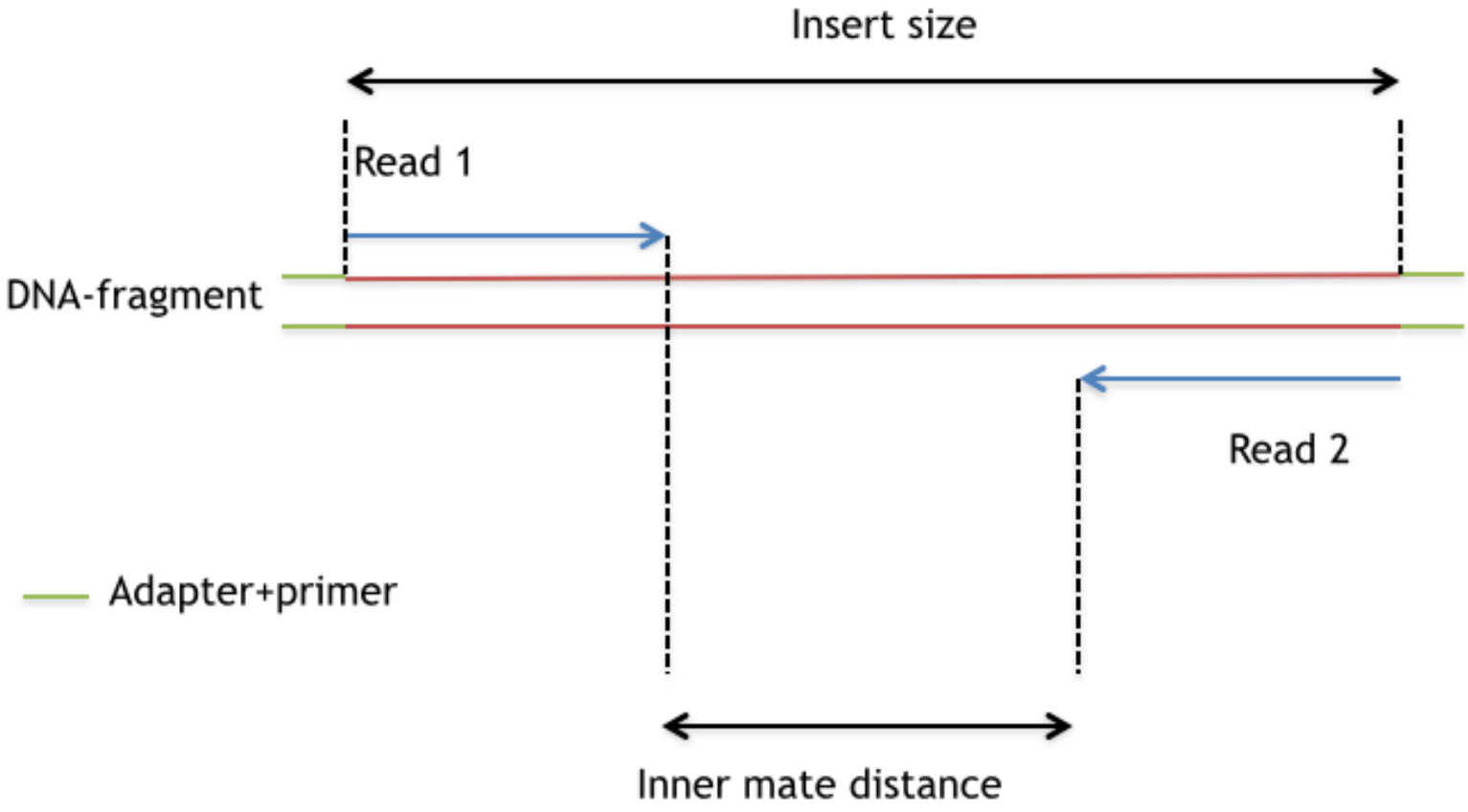

Genomic DNA
Fragment (200–500bp)
Ligate Adaptors
A1
SP1
SP2
A2
Generate Clusters
SP2
A2
SP1
A1
FLOWCELL
Sequence First End
SP1
A2
Regenerate Clusters and
Sequence Paired End
SP2
A1

PE) sequencing

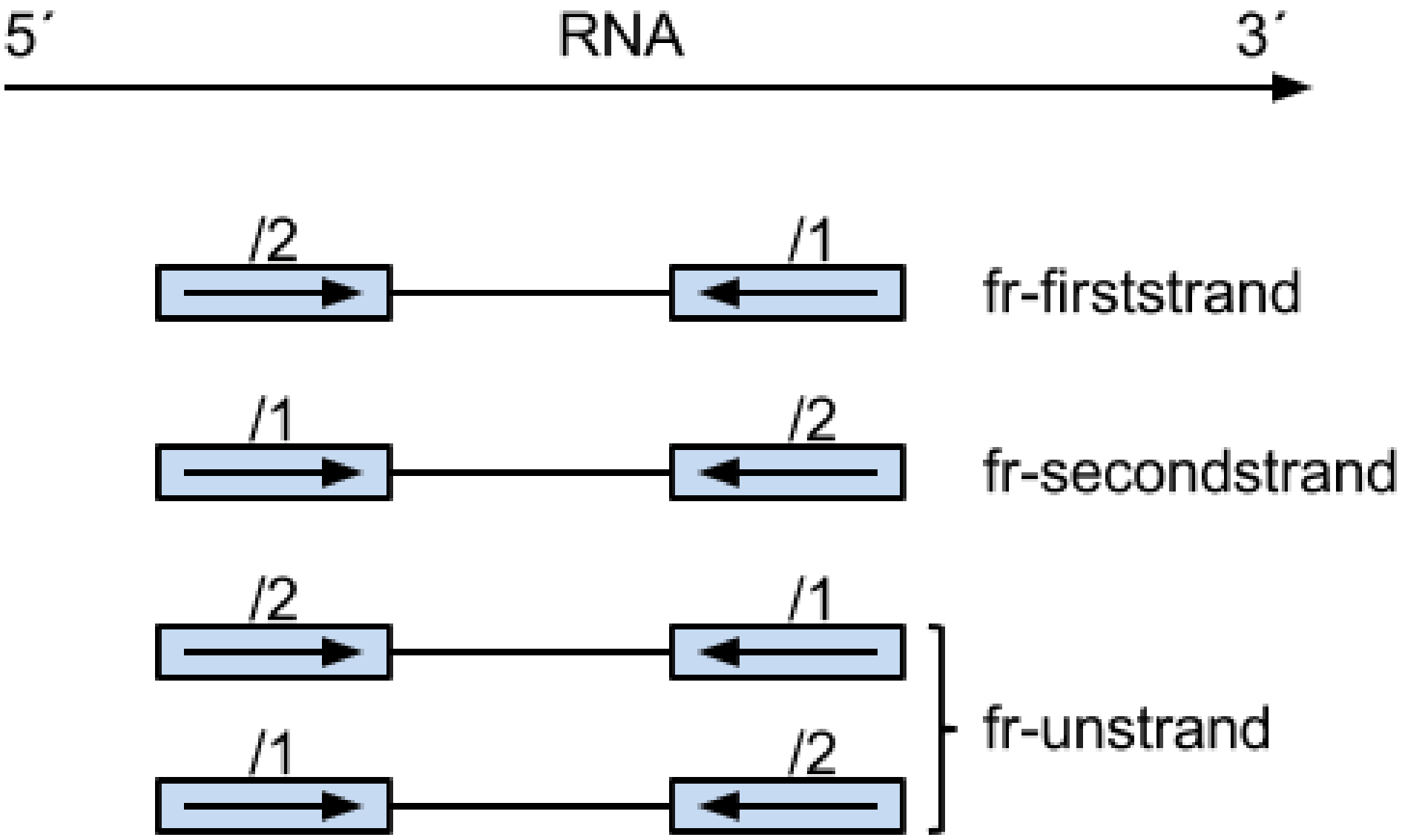
```
@61DFRAAXX100204:1:100:10494:3070/1
AAACAACAGGGCACATTGTCACTCTTGTATTTGAAAAACACTTTCCGGCCAT
+
ACCCCCCCCCCCCCCCCCCCCCCCCCCCCCCBC?CCCCCCCCC@@CACCCCCA
```

```
@61DFRAAXX100204:1:100:10494:3070/2
ATCCAAGTTAAAACAGAGGCCTGTGACAGACTCTTGGCCCATCGTGTTGATA
+
_^_a^ccceggcgghhgZc`ghhc^egggd^_[d]defcdfd^Z^OXWaQ^ad
```

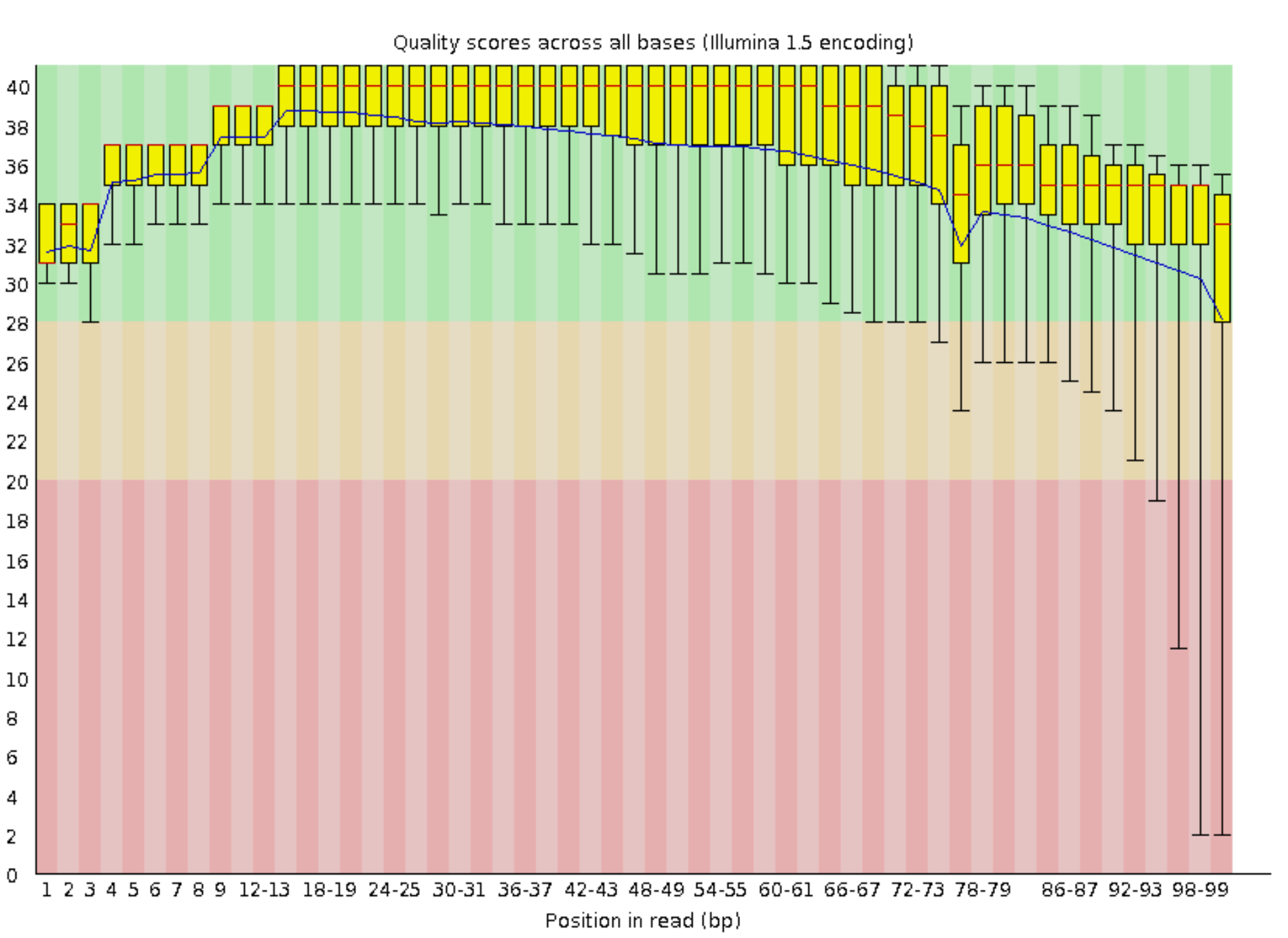
eads


Insert size
Read 1
DNA-fragment
Read 2
Adapter+primer
Inner mate distance

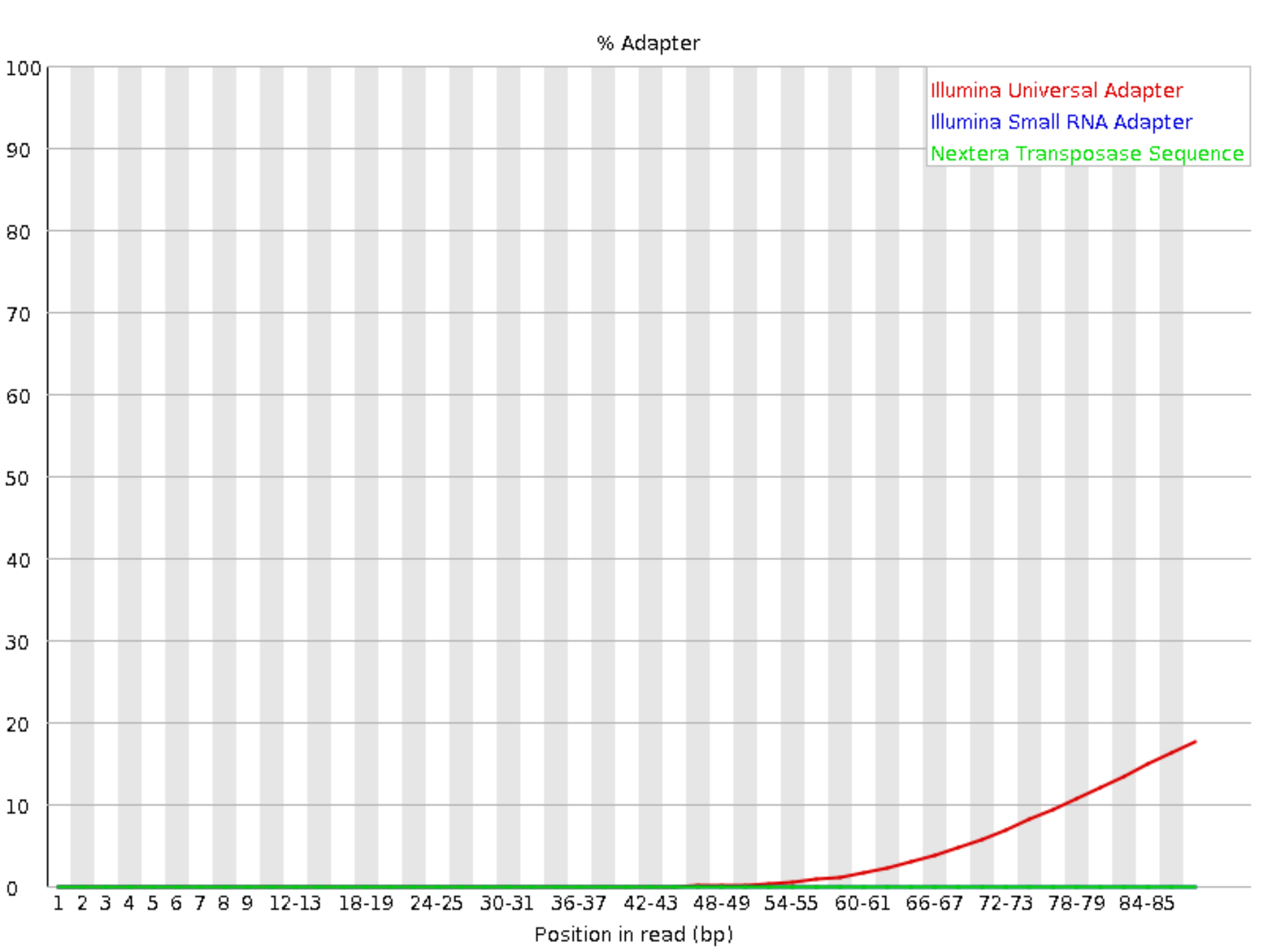
data


5´
RNA
3´
/2
/1
fr-firststrand
/1
/2
fr-secondstrand
/2
/1
/1
/2
fr-unstrand

or not

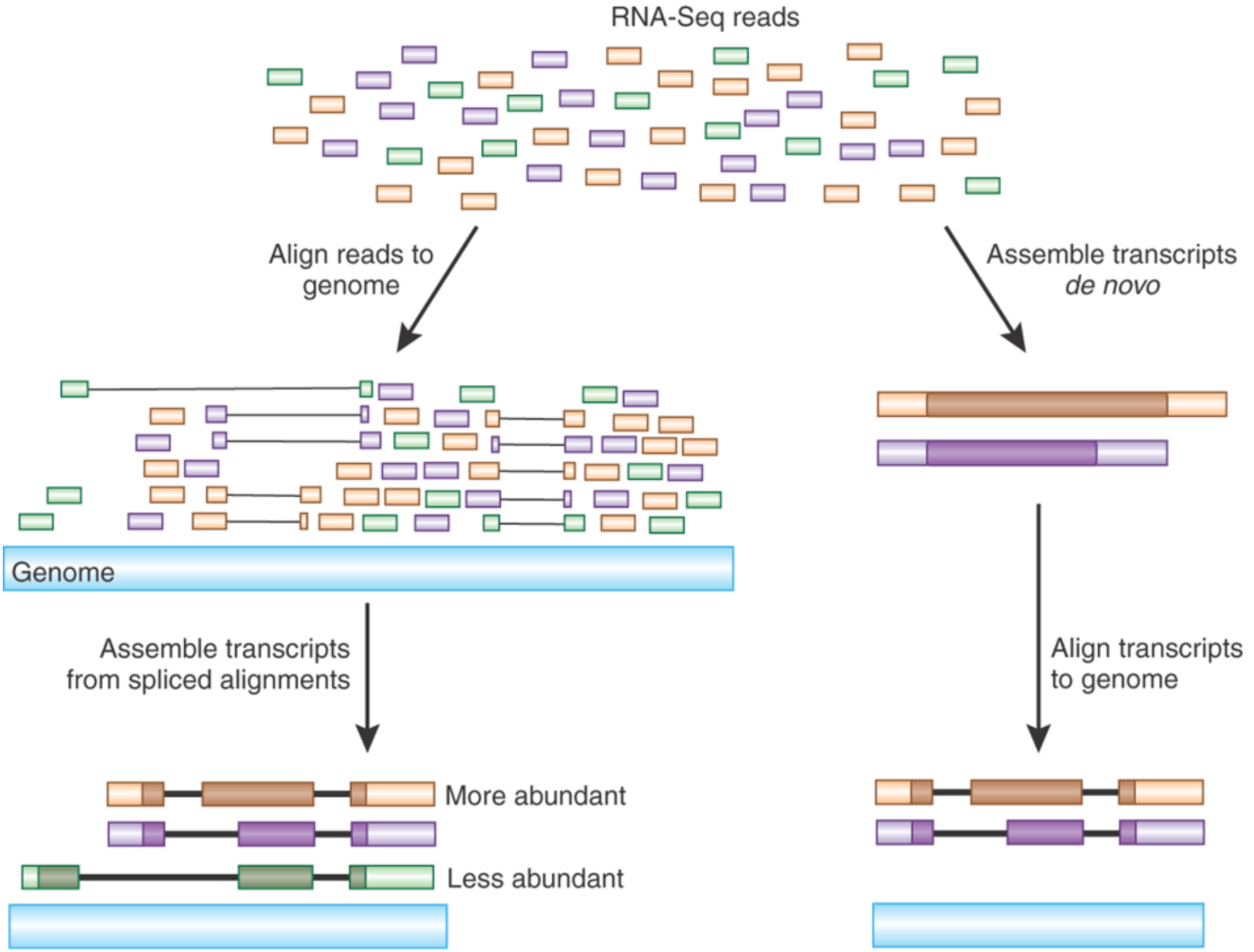

Quality scores across all bases (Illumina 1.5 encoding)
40
38
36
34
32
30
28
26
24
22
20
18
16
14
12
10
8
6
4
2
0
1 2 3 4 5 6 7 8 9 12-13 18-19 24-25 30-31 36-37 42-43 48-49 54-55 60-61 66-67 72-73 78-79 86-87 92-93 98-99
Position in read (bp)

lity control of raw reads


% Adapter
Illumina Universal Adapter
Illumina Small RNA Adapter
Nextera Transposase Sequence
100
90
80
70
60
50
40
30
20
10
0
1 2 3 4 5 6 7 8 9 12-13 18-19 24-25 30-31 36-37 42-43 48-49 54-55 60-61 66-67 72-73 78-79 84-85
Position in read (bp)

lity control of raw reads

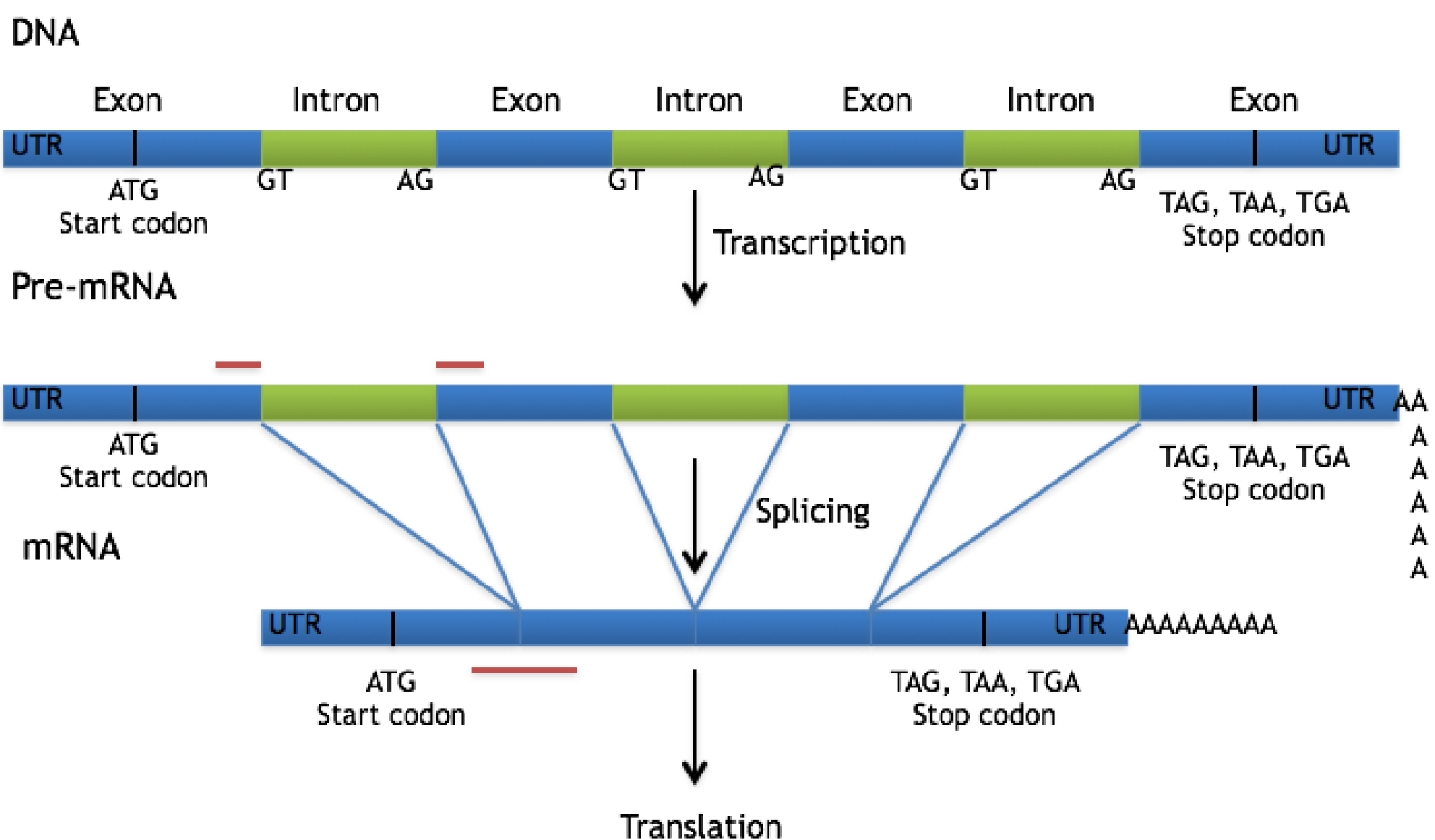
lity control of raw reads


RNA-Seq reads
Align reads to genome
Assemble transcripts de novo
Genome
Assemble transcripts from spliced alignments
Align transcripts to genome
More abundant
Less abundant

routes for analysis

Seqname	Source	Feature	Start	End	Score	Strand	Frame	Attributes
chr12	unknown	exon	87984	88017	.	+	.	gene_id "LOC100288778"; gene_name "LOC100288778"; transcript_id "NR_028269"; tss_id "TSS8200";
chr12	unknown	exon	88257	88392	.	+	.	gene_id "LOC100288778"; gene_name "LOC100288778"; transcript_id "NR_028269"; tss_id "TSS8200";
chr12	unknown	exon	88570	88771	.	+	.	gene_id "LOC100288778"; gene_name "LOC100288778"; transcript_id "NR_028269"; tss_id "TSS8200";
chr12	unknown	exon	88860	89018	.	+	.	gene_id "LOC100288778"; gene_name "LOC100288778"; transcript_id "NR_028269"; tss_id "TSS8200";
chr12	unknown	exon	89675	89827	.	+	.	gene_id "LOC100288778"; gene_name "LOC100288778"; transcript_id "NR_028269"; tss_id "TSS8200";
chr12	unknown	exon	90587	90655	.	+	.	gene_id "LOC100288778"; gene_name "LOC100288778"; transcript_id "NR_028269"; tss_id "TSS8200";
chr12	unknown	exon	90796	91263	.	+	.	gene_id "LOC100288778"; gene_name "LOC100288778"; transcript_id "NR_028269"; tss_id "TSS8200";
chr12	unknown	exon	147946	148509	.	-	.	gene_id "FAM138D"; gene_name "FAM138D"; transcript_id "NR_026823"; tss_id "TSS11862";
chr12	unknown	exon	148612	148814	.	-	.	gene_id "FAM138D"; gene_name "FAM138D"; transcript_id "NR_026823"; tss_id "TSS11862";
chr12	unknown	exon	149052	149412	.	-	.	gene_id "FAM138D"; gene_name "FAM138D"; transcript_id "NR_026823"; tss_id "TSS11862";
chr12	unknown	CDS	176049	176602	.	+	0	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P5442"; transcript_id "NM_001170738"; tss_id "TSS17433";
chr12	unknown	exon	176049	176602	.	+	.	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P5442"; transcript_id "NM_001170738"; tss_id "TSS17433";
chr12	unknown	start_codon	176049	176051	.	+	.	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P5442"; transcript_id "NM_001170738"; tss_id "TSS17433";
chr12	unknown	exon	186542	186878	.	+	.	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P13619"; transcript_id "NM_015232"; tss_id "TSS12565";
chr12	unknown	CDS	208312	208380	.	+	1	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P5442"; transcript_id "NM_001170738"; tss_id "TSS17433";
chr12	unknown	exon	208312	208380	.	+	.	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P13619"; transcript_id "NM_015232"; tss_id "TSS12565";
chr12	unknown	exon	208312	208380	.	+	.	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P5442"; transcript_id "NM_001170738"; tss_id "TSS17433";
chr12	unknown	CDS	234799	235078	.	+	1	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P5442"; transcript_id "NM_001170738"; tss_id "TSS17433";
chr12	unknown	exon	234799	235078	.	+	.	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P5442"; transcript_id "NM_001170738"; tss_id "TSS17433";
chr12	unknown	exon	246577	246793	.	-	.	gene_id "LOC574538"; gene_name "LOC574538"; transcript_id "NR_033859"; tss_id "TSS17153";
chr12	unknown	CDS	247433	248520	.	+	0	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P5442"; transcript_id "NM_001170738"; tss_id "TSS17433";
chr12	unknown	exon	247433	248520	.	+	.	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P13619"; transcript_id "NM_015232"; tss_id "TSS12565";
chr12	unknown	exon	247433	248520	.	+	.	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P5442"; transcript_id "NM_001170738"; tss_id "TSS17433";
chr12	unknown	CDS	247439	248520	.	+	0	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P13619"; transcript_id "NM_015232"; tss_id "TSS12565";
chr12	unknown	start_codon	247439	247441	.	+	.	gene_id "IQSEC3"; gene_name "IQSEC3"; p_id "P13619"; transcript_id "NM_015232"; tss_id "TSS12565";

hort reads from RNA to genomes


DNA
Exon
Intron
Exon
Intron
Exon
Intron
Exon
UTR
UTR
ATG
Start codon
GT
AG
GT
AG
GT
AG
TAG, TAA, TGA
Stop codon
Transcription
Pre-mRNA
UTR
UTR
AA
A
A
A
A
A
ATG
Start codon
TAG, TAA, TGA
Stop codon
Splicing
mRNA
UTR
UTR
AAAAAAAAA
ATG
Start codon
TAG, TAA, TGA
Stop codon
Translation

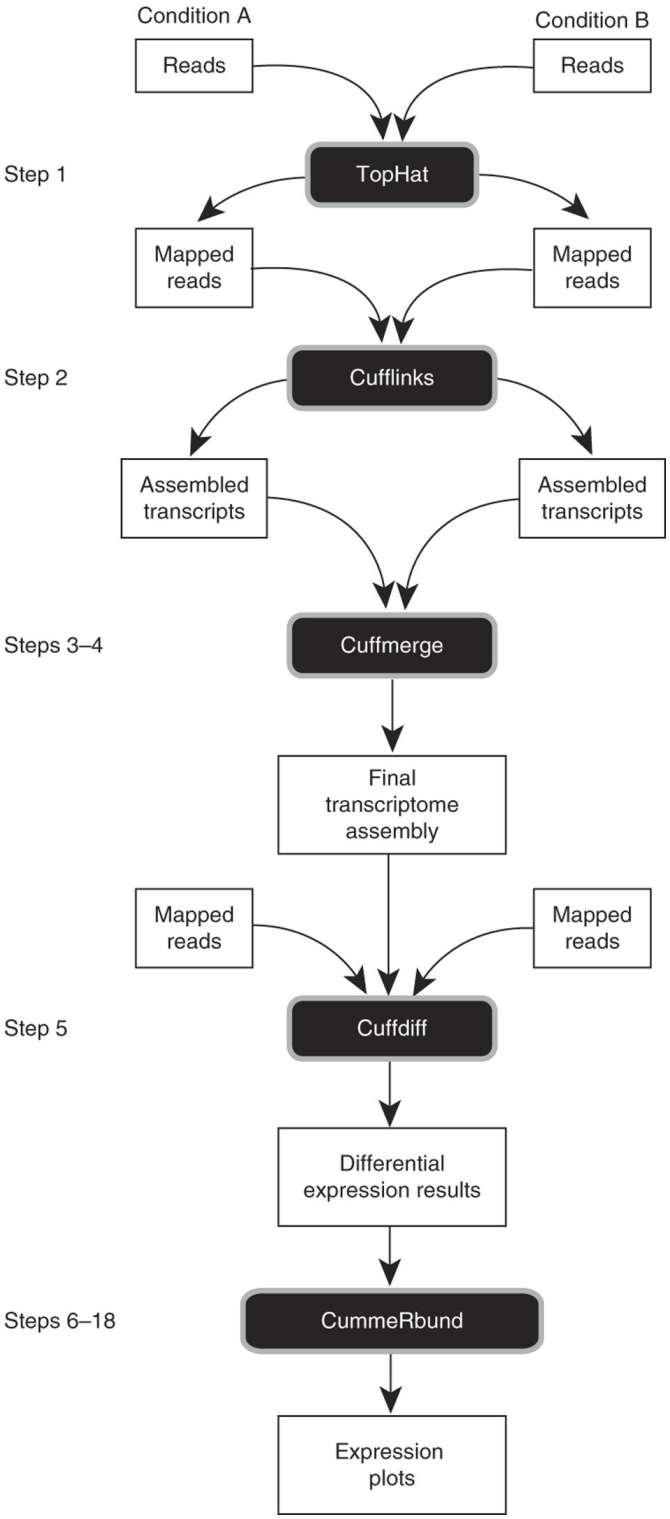
hort reads from RNA to genomes

```
read_distribution.py  -i Pairend_StrandSpecific_51mer_Human_hg19.bam -r hg19.refseq.bed12
```

Output:

Group	Total_bases	Tag_count	Tags/Kb
CDS_Exons	33302033	20002271	600.63
5'UTR_Exons	21717577	4408991	203.01
3'UTR_Exons	15347845	3643326	237.38
Introns	1132597354	6325392	5.58
TSS_up_1kb	17957047	215331	11.99
TSS_up_5kb	81621382	392296	4.81
TSS_up_10kb	149730983	769231	5.14
TES_down_1kb	18298543	266161	14.55
TES_down_5kb	78900674	729997	9.25
TES_down_10kb	140361190	896882	6.39

hort reads from RNA to genomes


Condition A
Condition B
Reads
Reads
Step 1
TopHat
Mapped reads
Mapped reads
Step 2
Cufflinks
Assembled transcripts
Assembled transcripts
Steps 3–4
Cuffmerge
Final transcriptome assembly
Mapped reads
Mapped reads
Step 5
Cuffdiff
Differential expression results
Steps 6–18
CummeRbund
Expression plots

workflow

: Aligns reads to

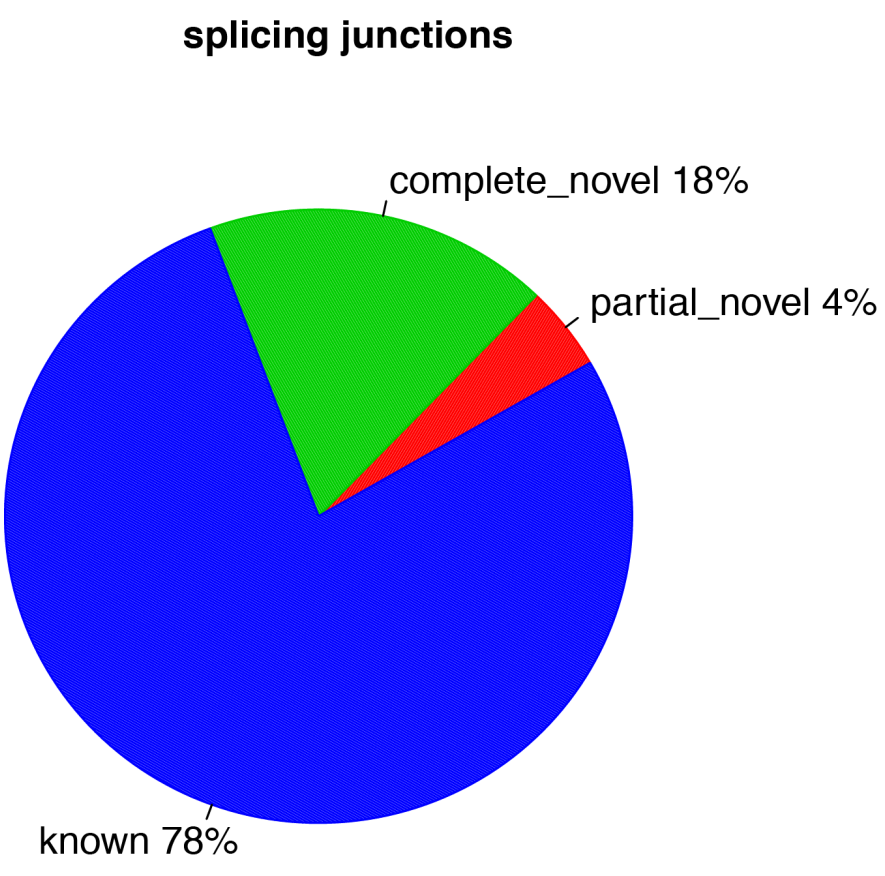
RNA-seq

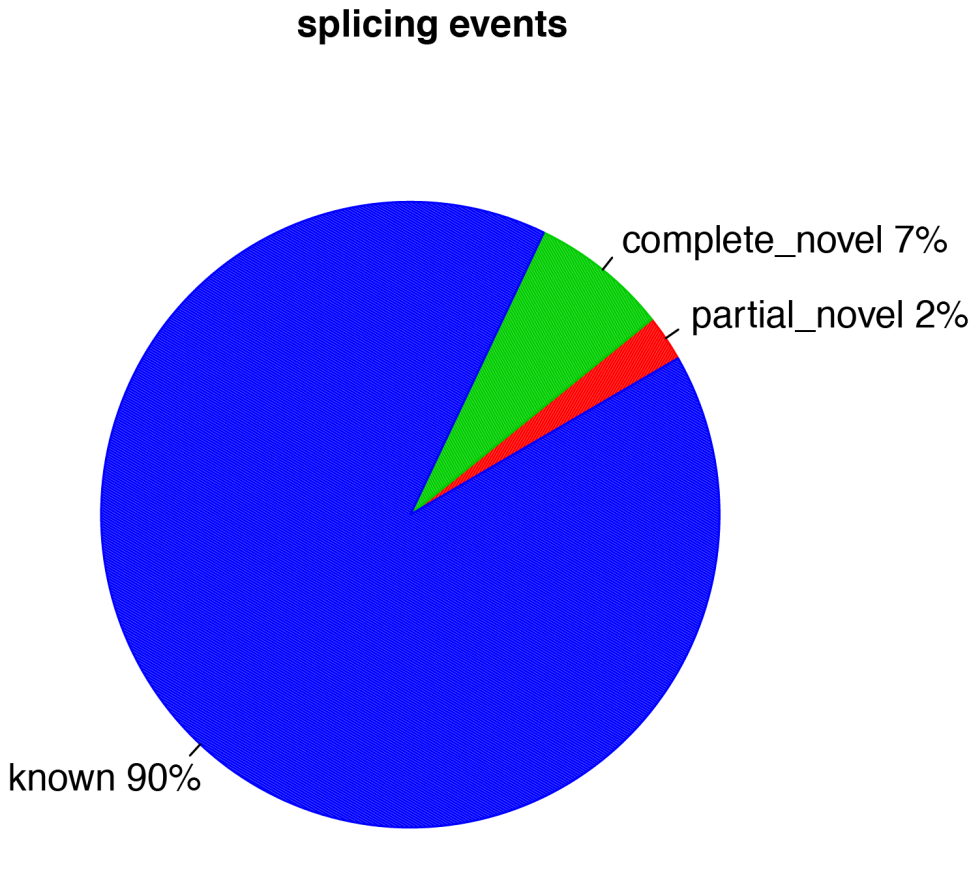
```
read_distribution.py  -i Pairend_StrandSpecific_51mer_Human_hg19.bam -r hg19.refseq.bed12
```

Output:

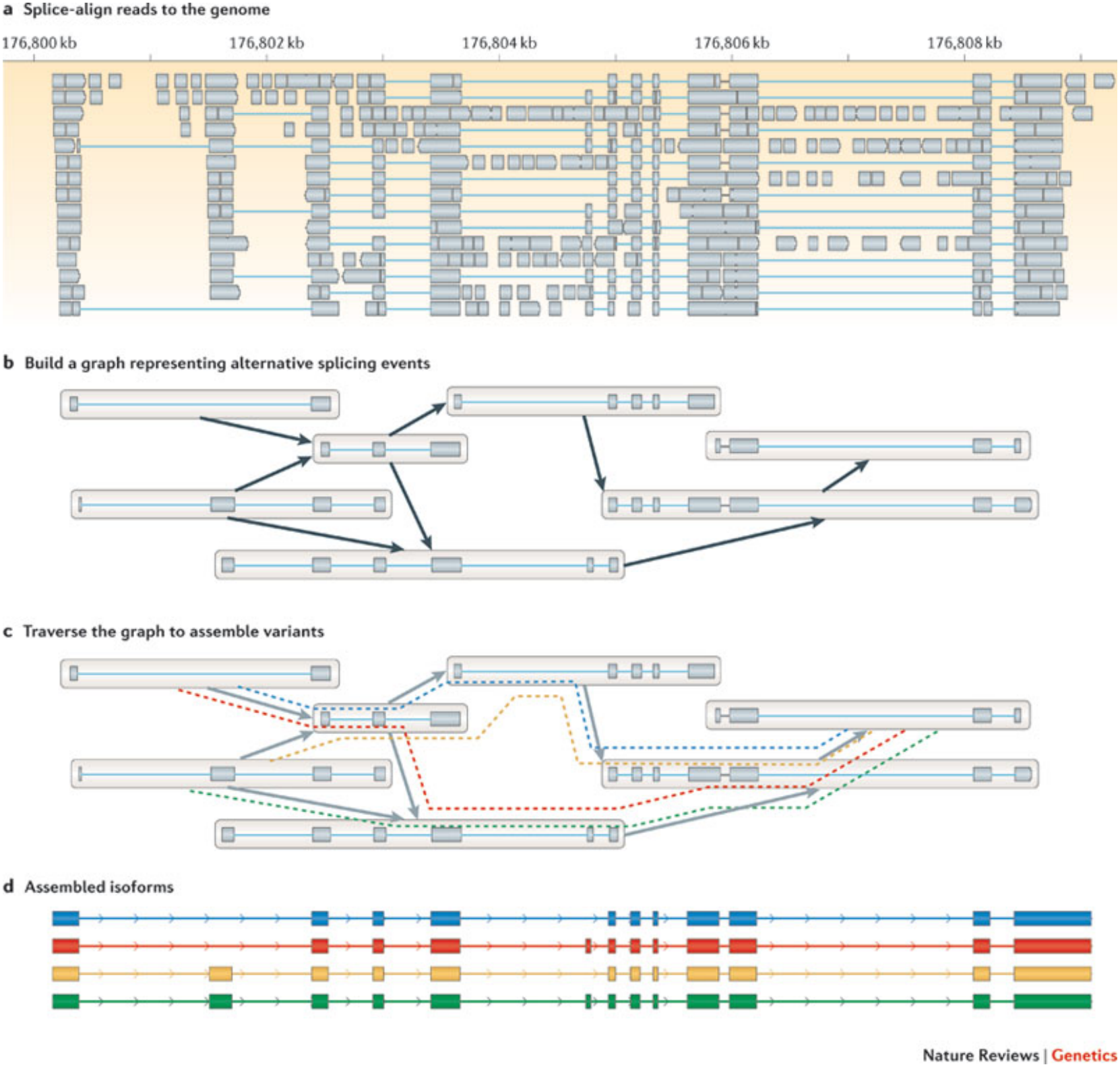
Group	Total_bases	Tag_count	Tags/Kb
CDS_Exons	33302033	20002271	600.63
5'UTR_Exons	21717577	4408991	203.01
3'UTR_Exons	15347845	3643326	237.38
Introns	1132597354	6325392	5.58
TSS_up_1kb	17957047	215331	11.99
TSS_up_5kb	81621382	392296	4.81
TSS_up_10kb	149730983	769231	5.14
TES_down_1kb	18298543	266161	14.55
TES_down_5kb	78900674	729997	9.25
TES_down_10kb	140361190	896882	6.39

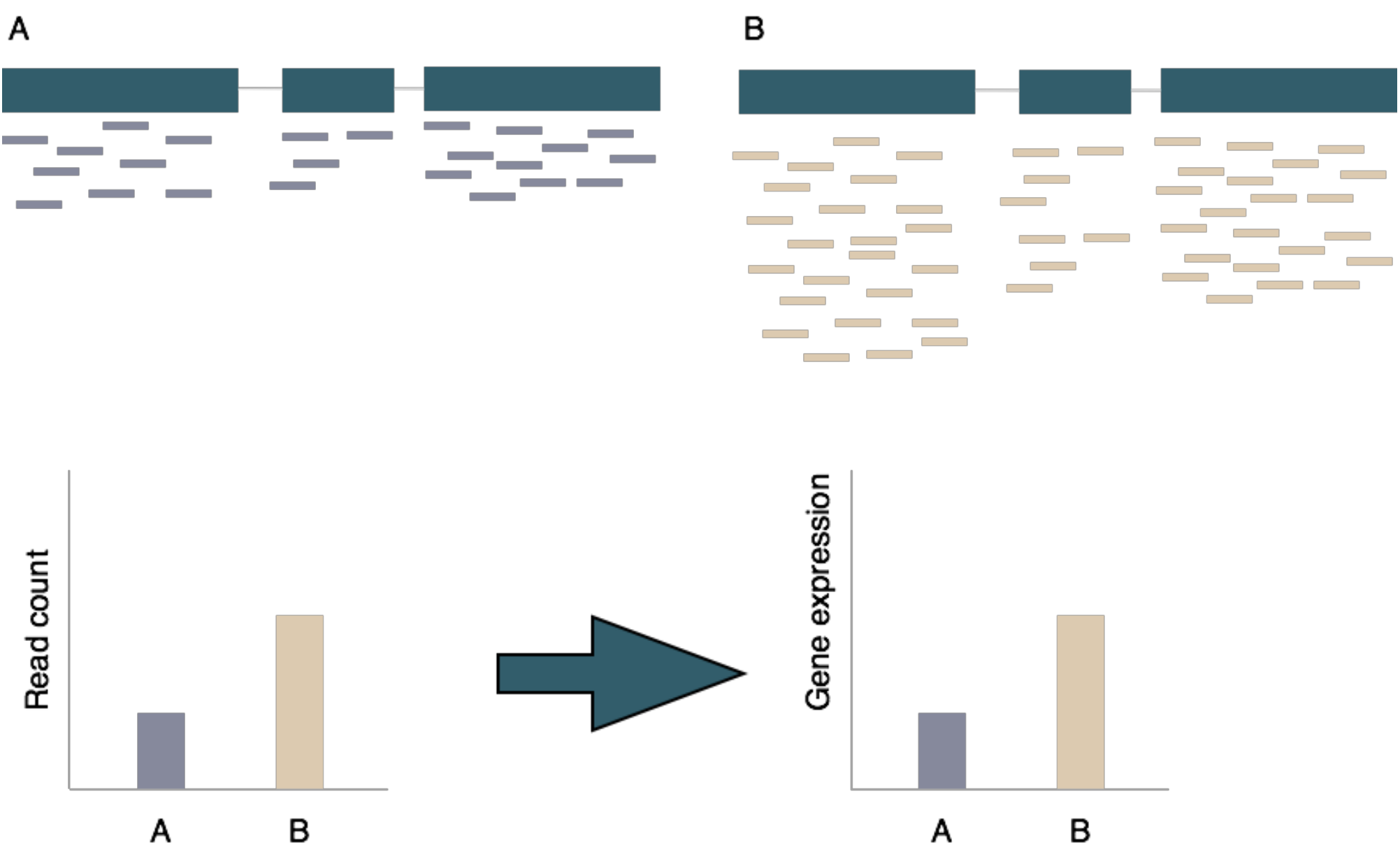
pped reads


splicing junctions
complete_novel 18%
partial_novel 4%
known 78%

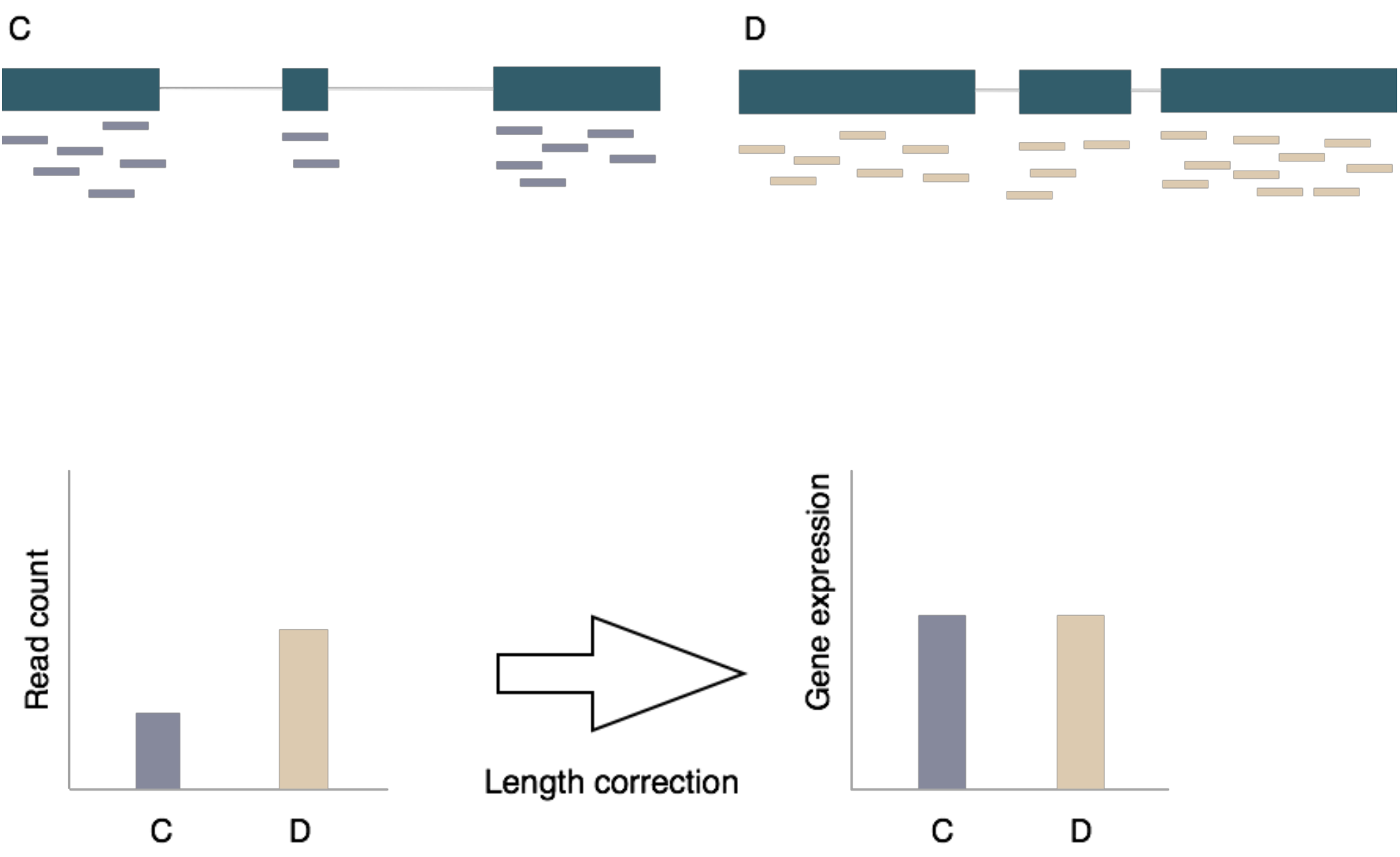

splicing events
complete_novel 7%
partial_novel 2%
known 90%

pped reads

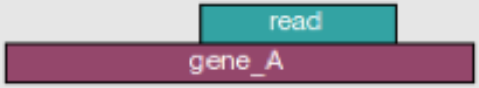
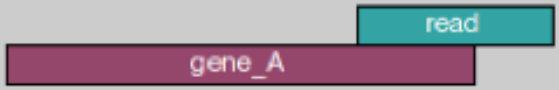
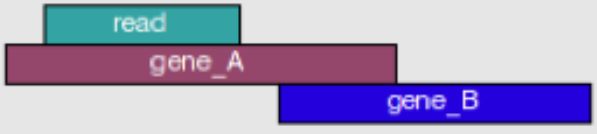
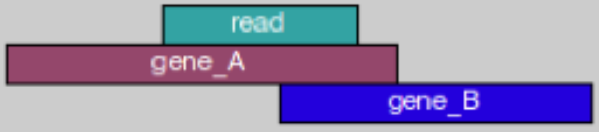
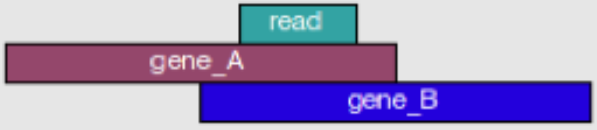

a Splice-align reads to the genome
176,800 kb
176,802 kb
176,804 kb
176,806 kb
176,808 kb
b Build a graph representing alternative splicing events
c Traverse the graph to assemble variants
d Assembled isoforms


A
B
Read count
A
B
Gene expression
A
B

nts to gene expression

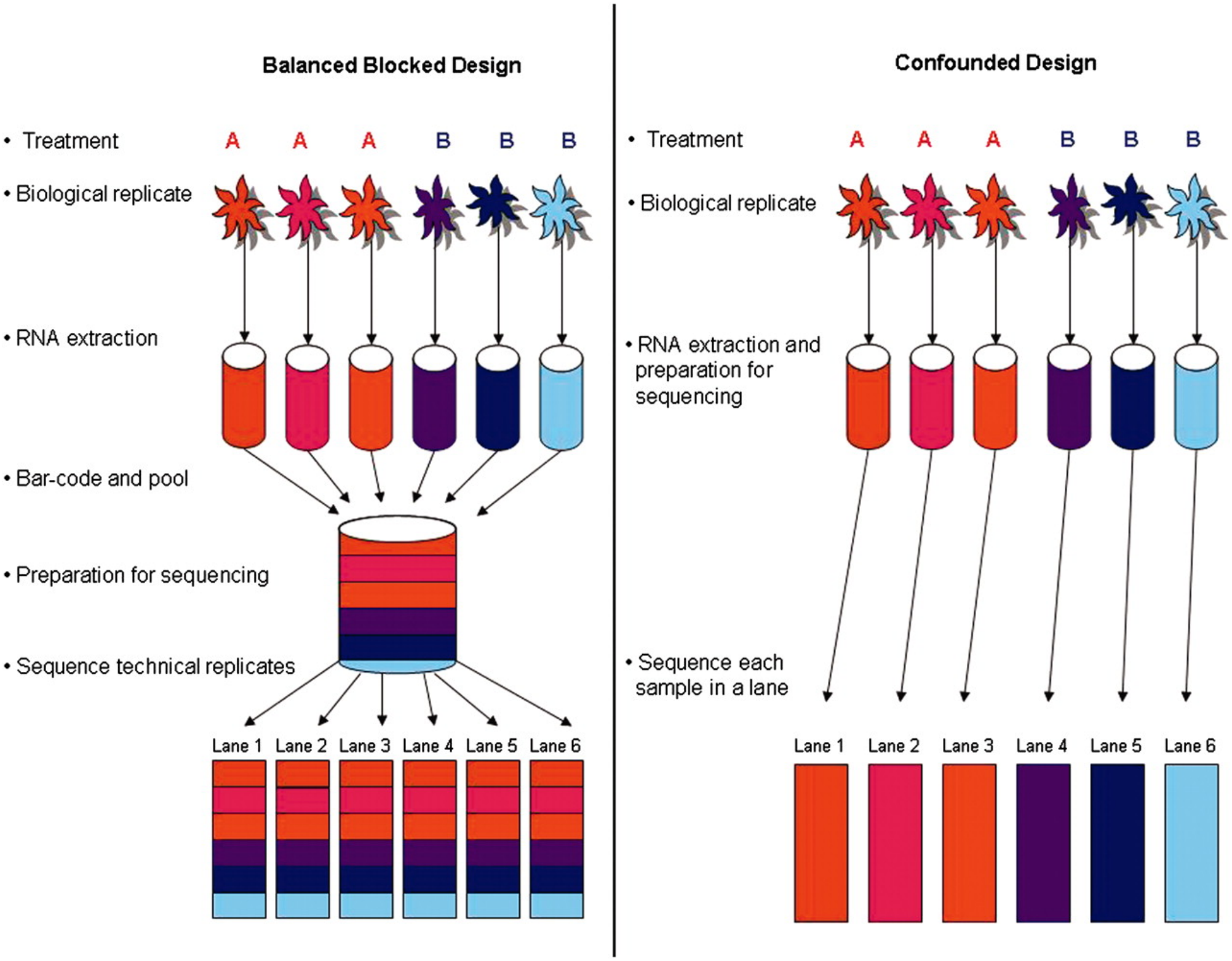

C
D
Read count
C
D
Length correction
Gene expression
C
D

nts to gene expression

	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

ads are the same

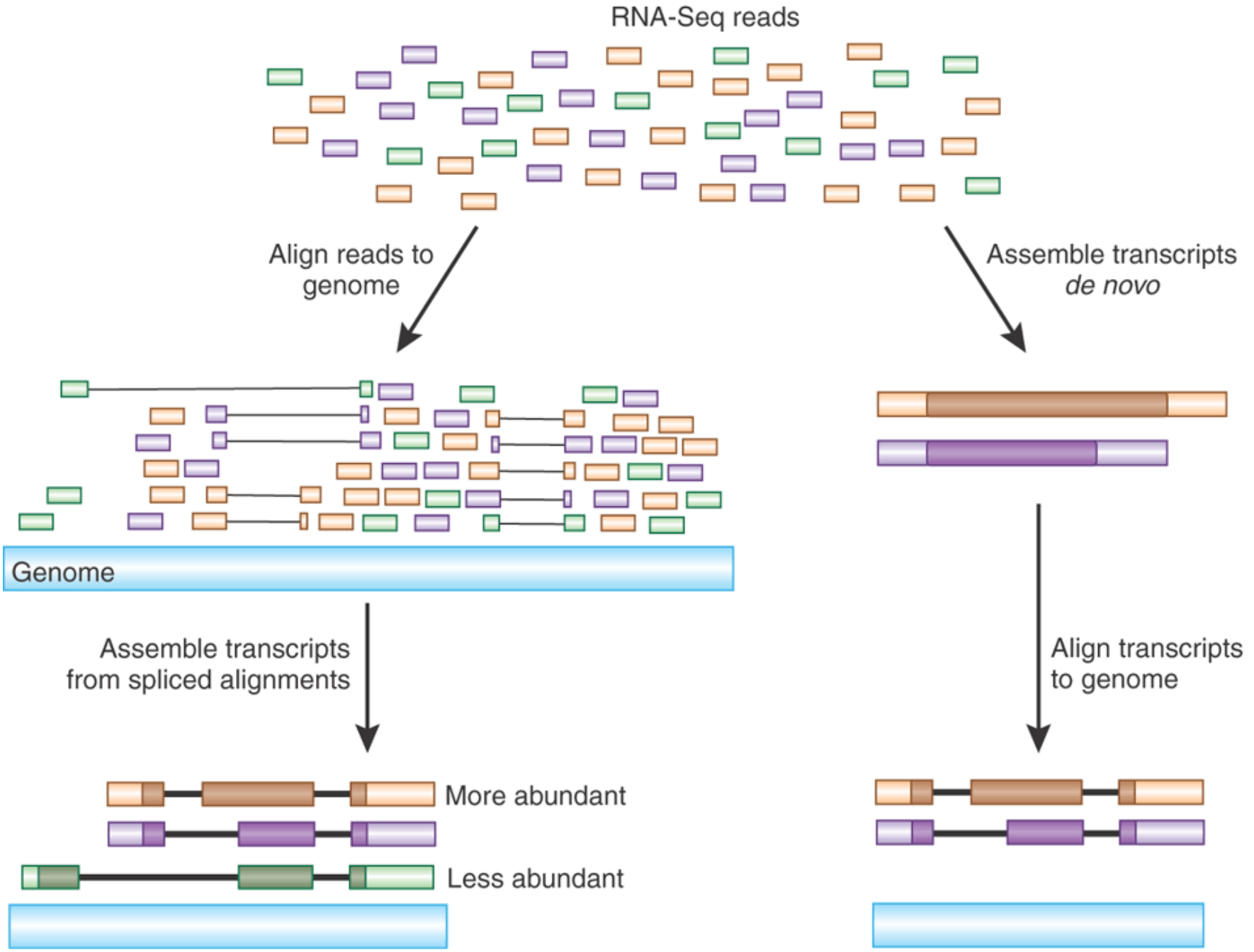
ed expression Values


Balanced Blocked Design
• Treatment
A A A B B B
• Biological replicate
• RNA extraction
• Bar-code and pool
• Preparation for sequencing
• Sequence technical replicates
Lane 1 Lane 2 Lane 3 Lane 4 Lane 5 Lane 6
Confounded Design
• Treatment
A A A B B B
• Biological replicate
• RNA extraction and preparation for sequencing
• Sequence each sample in a lane
Lane 1 Lane 2 Lane 3 Lane 4 Lane 5 Lane 6

ntal design

	Condition 1	Condition 2	Fold_Change	Significant?
Gene A	1	2	2-fold	No
Gene B	100	200	2-fold	Yes

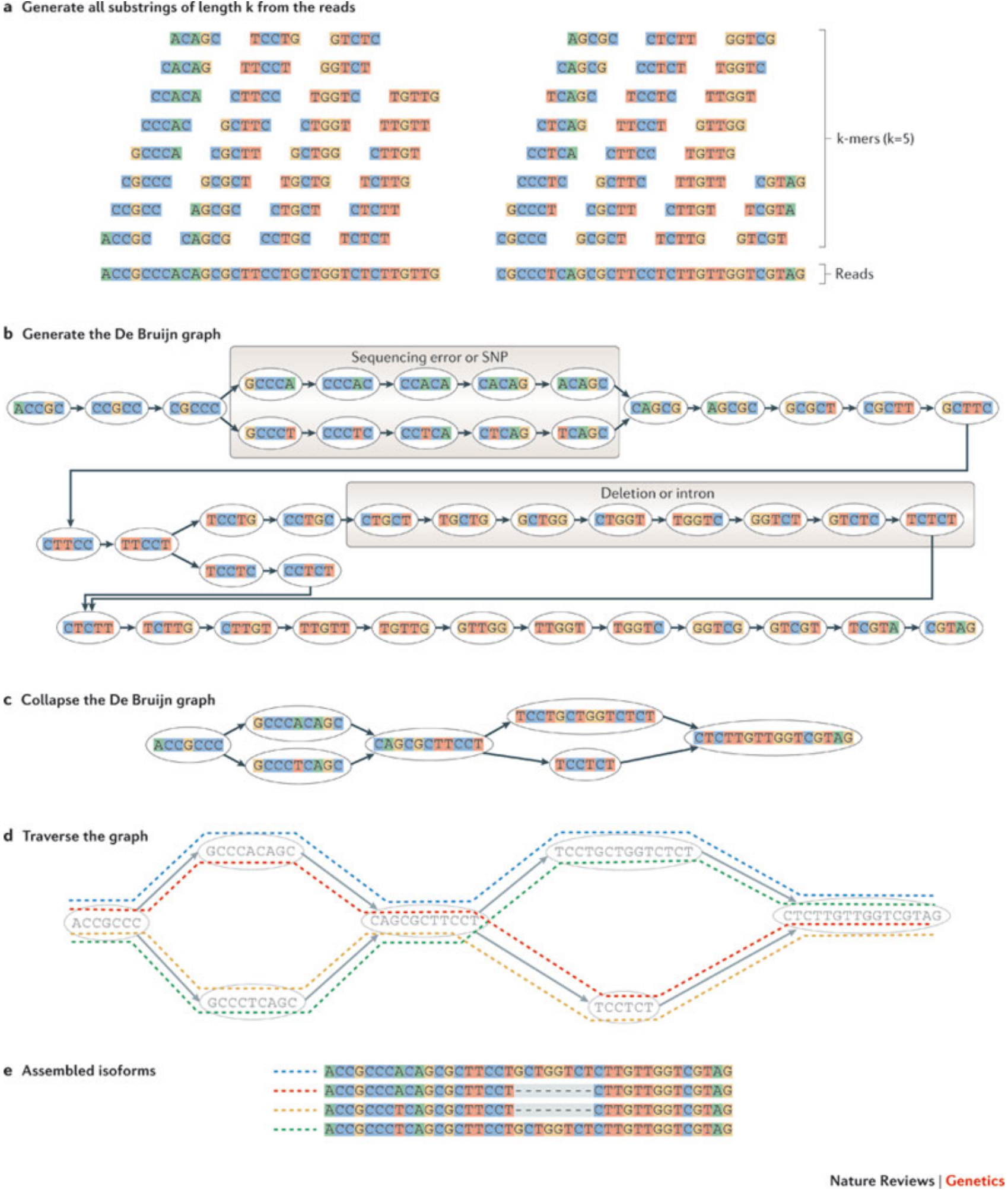
ital design


RNA-Seq reads
Align reads to genome
Assemble transcripts de novo
Genome
Assemble transcripts from spliced alignments
Align transcripts to genome
More abundant
Less abundant

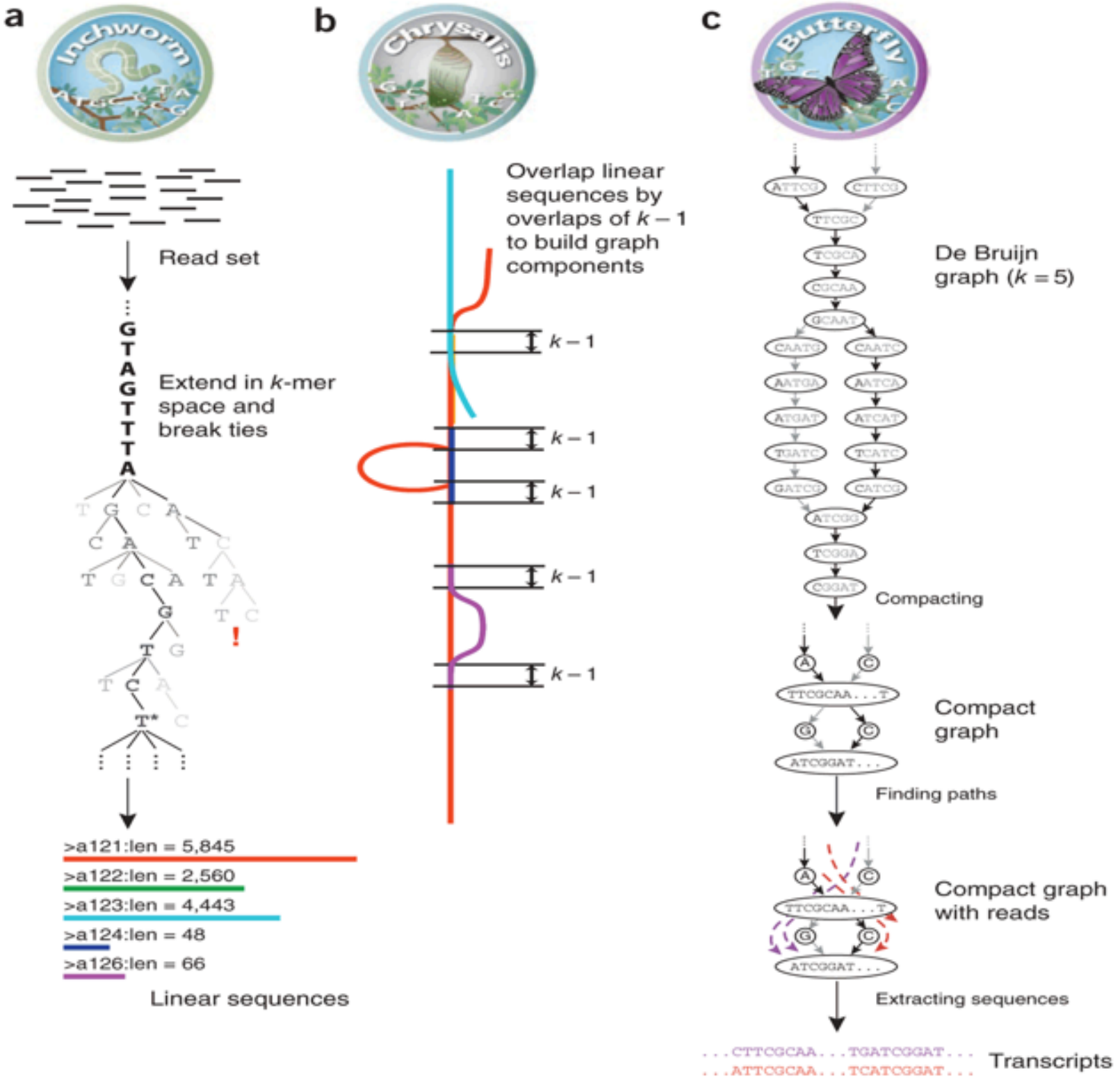
routes for analysis

llenges in relation to genome assem

rograms available


a Generate all substrings of length k from the reads
ACAGC TCCTG GTCTC
CACAG TTCCT GGTCT
CCACA CTTCC TGGTC TGTTG
CCCAC GCTTC CTGGT TTGTT
GCCCA CGCTT GCTGG CTTGT
CGCCC GCGCT TGCTG TCTTG
CCGCC AGCGC CTGCT CTCTT
ACCGC CAGCG CCTGC TCTCT
AGCGC CTCTT GGTCG
CAGCG CCTCT TGGTC
TCAGC TCCTC TTGGT
CTCAG TTCCT GTTGG
CCTCA CTTCC TGTTG
CCCTC GCTTC TTGTT CGTAG
GCCCT CGCTT CTTGT TCGTA
CGCCC GCGCT TCTTG GTCGT
k-mers (k=5)
ACCGCCCACAGCGCTTCCTGCTGGTCTCTTGTTG
CGCCCTCAGCGCTTCCTCTTGTTGGTCGTAG
Reads
b Generate the De Bruijn graph
Sequencing error or SNP
ACCGC CCGCC CGCCC GCCCA CCCAC CCACA CACAG ACAGC GCCCT CCCTC CCTCA CTCAG TCAGC CAGCG AGCGC GCGCT CGCTT GCTTC
Deletion or intron
CTTCC TTCCT TCCTG CCTGC CTGCT TGCTG GCTGG CTGGT TGGTC GGTCT GTCTC TCTCT TCCTC CCTCT
CTCTT TCTTG CTTGT TTGTT TGTTG GTTGG TTGGT TGGTC GGTCG GTCGT TCGTA CGTAG
c Collapse the De Bruijn graph
ACCGCCC GCCCACAGC GCCCTCAGC CAGCGCTTCCT TCCTGCTGGTCTCT TCCTCT CTCTTGTTGGTCGTAG
d Traverse the graph
ACCGCCC GCCCACAGC GCCCTCAGC CAGCGCTTCCT TCCTGCTGGTCTCT TCCTCT CTCTTGTTGGTCGTAG
e Assembled isoforms
ACCGCCCACAGCGCTTCCTGCTGGTCTCTTGTTGGTCGTAG
ACCGCCCACAGCGCTTCCT--------CTTGTTGGTCGTAG
ACCGCCCTCAGCGCTTCCT--------CTTGTTGGTCGTAG
ACCGCCCTCAGCGCTTCCTGCTGGTCTCTTGTTGGTCGTAG

RNA-seq analysis de-novo assembly


a
Inchworm
Read set
GTAGTTTA
Extend in k-mer space and break ties
>a121:len = 5,845
>a122:len = 2,560
>a123:len = 4,443
>a124:len = 48
>a126:len = 66
Linear sequences
b
Chrysalis
Overlap linear sequences by overlaps of k – 1 to build graph components
k – 1
k – 1
k – 1
k – 1
k – 1
c
Butterfly
ATTCG
CTTCG
TTCGC
TCGCA
CGCAA
GCAAT
CAATG
CAATC
AATGA
AATCA
ATGAT
ATCAT
TGATC
TCATC
GATCG
CATCG
ATCGG
TCGGA
CGGAT
De Bruijn graph (k = 5)
Compacting
TTCGCAA...T
ATCGGAT...
Compact graph
Finding paths
Compact graph with reads
Extracting sequences
...CTTCGCAA...TGATCGGAT...
...ATTCGCAA...TCATCGGAT...
Transcripts

RNA-seq analysis de-novo assembly

- with ref.

- without ref.