RNA-Seq I

Measuring Gene Expression

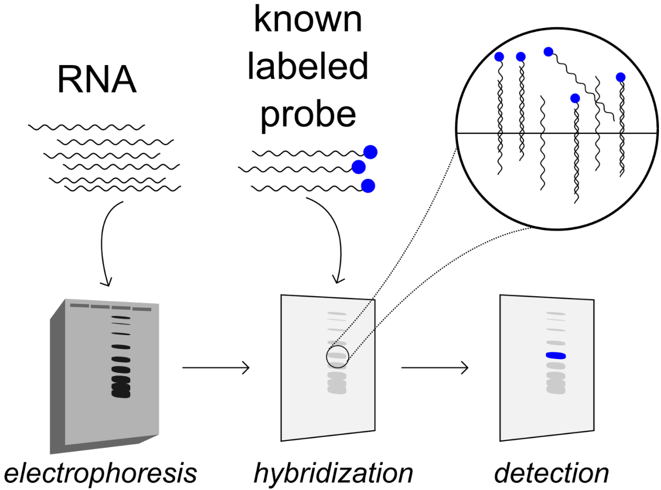
Antonio Marco

School of Biological Sciences
University of Essex

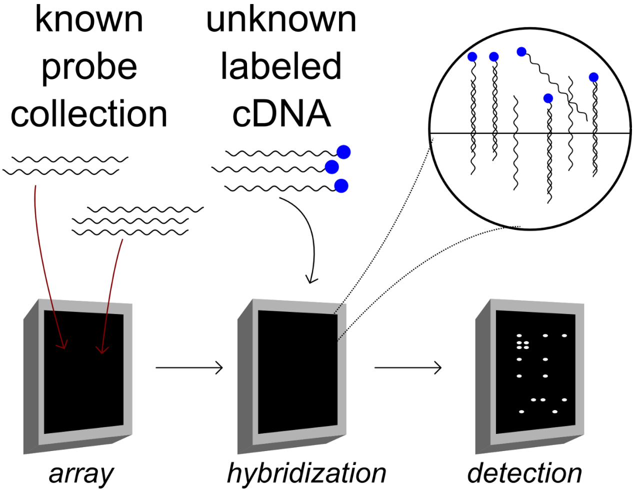
6-Jun-17

Northern blot

The 'IS IT THERE?' approach



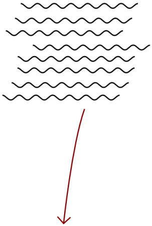
The 'IS ANY OF THESE?' approach

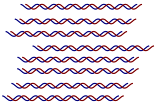


RNA-Seq

The 'WHAT'S IN THERE?' approach

RNA



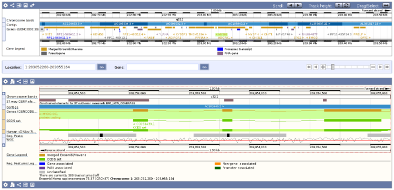


cDNA

acgtcagact
gctagagct
gctcagagtcgcgat
cgtagctacccgatc
cgctagcccatgcac
gctagctctgacc
ccctagtcgatccca

sequencing

analysis

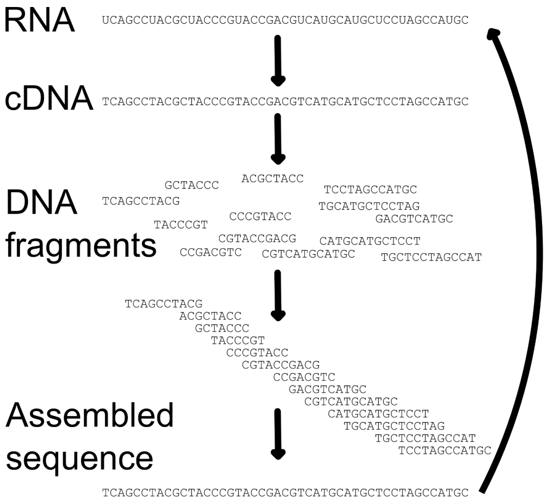


Location: 1:203052260-203055164

There are currently 380 tracks turned off
Ensembl Homo sapiens version 75.37 (GRCh37) Chromosome 1: 203,052,260 - 203,055,164

RNA-Seq

DNA sequencing is done in small fragments



RNA

UCAGCCUACGCUACCCGUACCGACGUCAUGCAUGCUCCUAGCCAUGC

cDNA

TCAGCCTACGCTACCCGTACCGACGTCATGCATGCTCCTAGCCATGC

DNA fragments

GCTACCC ACGCTACC TCCTAGCCATGC
TCAGCCTACG TGCATGCTCCTAG
TACCCGT CCCGTACC GACGTCATGC
CGTACCGACG CATGCATGCTCCT
CCGACGTC CGTCATGCATGC TGCTCCTAGCCAT

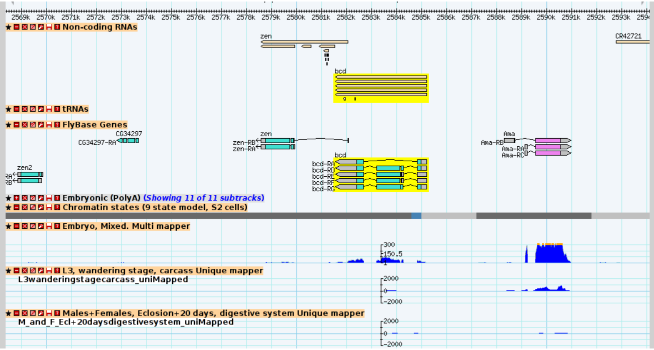
Assembled sequence

TCAGCCTACG
ACGCTACC
GCTACCC
TACCCGT
CCCGTACC
CGTACCGACG
CCGACGTC
GACGTCATGC
CGTCATGCATGC
CATGCATGCTCCT
TGCATGCTCCTAG
TGCTCCTAGCCAT
TCCTAGCCATGC

TCAGCCTACGCTACCCGTACCGACGTCATGCATGCTCCTAGCCATGC

RNA-Seq

RNA-Seq is a quantitative technique


2569k 2570k 2571k 2572k 2573k 2574k 2575k 2576k 2577k 2578k 2579k 2580k 2581k 2582k 2583k 2584k 2585k 2586k 2587k 2588k 2589k 2590k 2591k 2592k 2593k 259
Non-coding RNAs
zen
CR42721
bcd
tRNAs
FlyBase Genes
CG34297
CG34297-RA
zen
zen-RB
zen-RA
Ama
Ama-RB
Ama-RA
Ama-RC
bcd
bcd-RA
bcd-RD
bcd-RE
bcd-RF
bcd-RG
zen2
RA
RB
Embryonic (PolyA) (Showing 11 of 11 subtracks)
Chromatin states (9 state model, S2 cells)
Embryo, Mixed. Multi mapper
300
150.5
L3, wandering stage, carcass Unique mapper
L3wanderingstagecarcass_uniMapped
2000
0
-2000
Males+Females, Eclosion+20 days, digestive system Unique mapper
M_and_F_Ecl+20daysdigestivesystem_uniMapped
2000
0
-2000

Assessing data quality

- Ideally, sequencers always give the actual reads

ACTUAL SEQUENCE	CACCGTGCCATCGCATCCGATCGCATCGCATCGCATC
MACHINE READS	CACCGTGCCATCGCATCCGATCGCATCGCATCGCATC

Assessing data quality

- Ideally, sequencers always give the actual reads

ACTUAL SEQUENCE CACCGTGCCATCGCATCCGATCGCATCGCATCGCATC

MACHINE READS CACCGTGCCATCGCATCCGATCGCATCGCATCGCATC

- In reality, they often contain errors

ACTUAL SEQUENCE CACCGTGCCATCGCATCCGATCGCATCGCATCGCATC

MACHINE READS CACC**T**TGCCATC**C**CATCCGATCGCATCGCA**AAC**CA**CT**

Assessing data quality

- Ideally, sequencers always give the actual reads

ACTUAL SEQUENCE CACCGTGCCATCGCATCCGATCGCATCGCATCGCATC

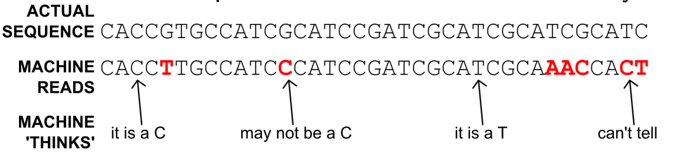
MACHINE READS CACCGTGCCATCGCATCCGATCGCATCGCATCGCATC

- In reality, they often contain errors

ACTUAL SEQUENCE CACCGTGCCATCGCATCCGATCGCATCGCATCGCATC

MACHINE READS CACC**T**TGCCATC**C**CATCCGATCGCATCGCA**AAC**CA**CT**

- Good news is, sequencers tell us how confident they are



ACTUAL SEQUENCE CACCGTGCCATCGCATCCGATCGCATCGCATCGCATC

MACHINE READS CACC**T**TGCCATC**C**CATCCGATCGCATCGCA**AAC**CA**CT**

MACHINE 'THINKS' it is a C | may not be a C | it is a T | can't tell

Assessing data quality

- Ideally, sequencers always give the actual reads

ACTUAL SEQUENCE CACCGTGCCATCGCATCCGATCGCATCGCATCGCATC

MACHINE READS CACCGTGCCATCGCATCCGATCGCATCGCATCGCATC

- In reality, they often contain errors

ACTUAL SEQUENCE CACCGTGCCATCGCATCCGATCGCATCGCATCGCATC

MACHINE READS CACC**T**TGCCATC**C**CATCCGATCGCATCGCA**AAC**CA**CT**

- Good news is, sequencers tell us how confident they are

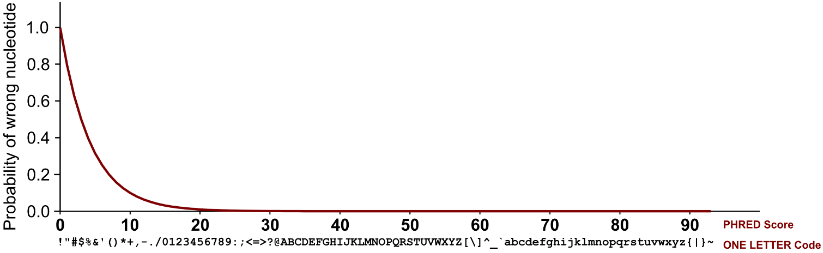
ACTUAL SEQUENCE CACCGTGCCATCGCATCCGATCGCATCGCATCGCATC

MACHINE READS CACC**T**TGCCATC**C**CATCCGATCGCATCGCA**AAC**CA**CT**

```
QUALITY SCORE ~~~~?~~~~~~~?~~~~~~~~~~~~~~~~~!!!~~!!
```

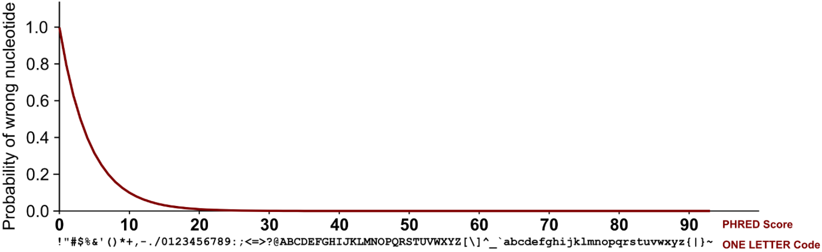
Phred Score and FASTQ format

- Phred score measures the probability of a sequencing error



Phred Score and FASTQ format

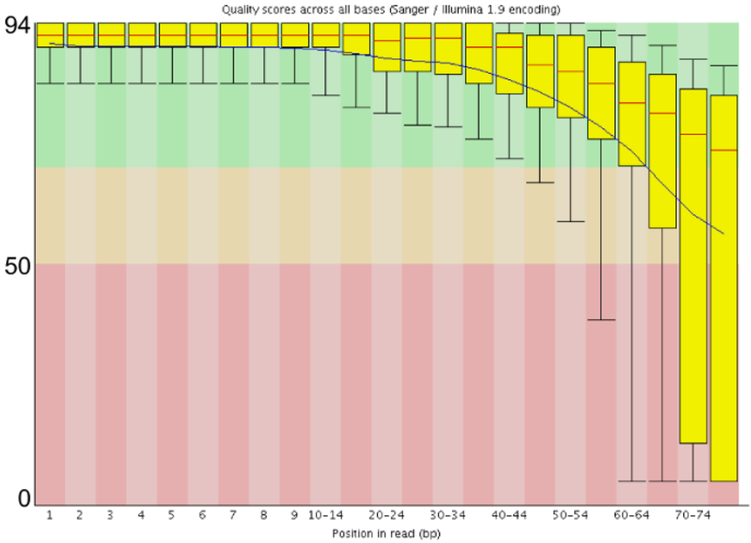
- Phred score measures the probability of a sequencing error



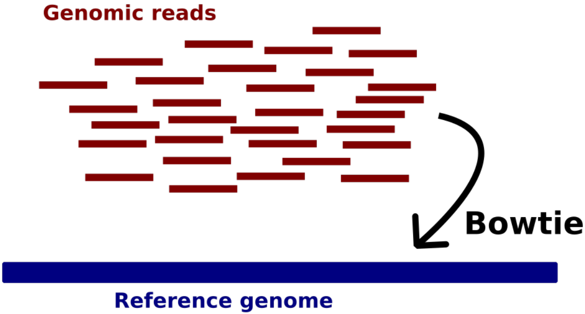
- The FASTQ format includes Phred scores in a one-letter code

```
@SEQUENCE_NAME
CATGGCTAGCTGCTAGCTAGCTAGACATTCATCGAAATCGCTAGCCTAGCTACGA
+
!’’*((((***+))%%%++)(%%%%).1***-+*’’))**55CCF>>>>>>C%%%
```

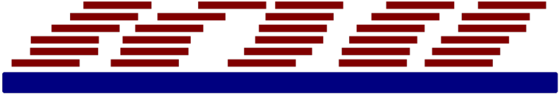
Read quality



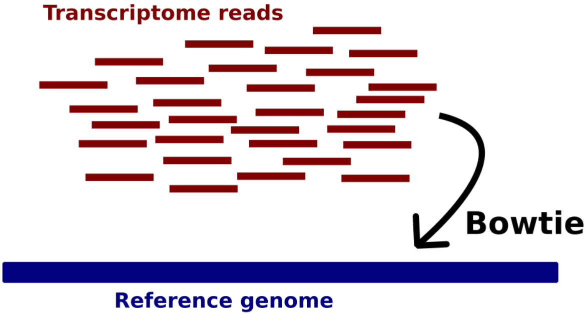
Mapping reads to a genome: Bowtie



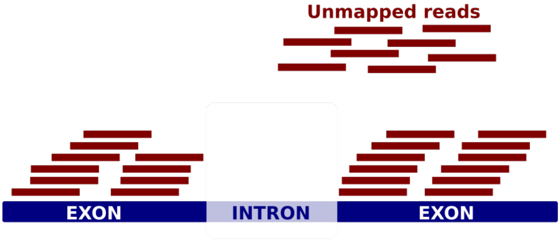
Mapping reads to a genome: Bowtie



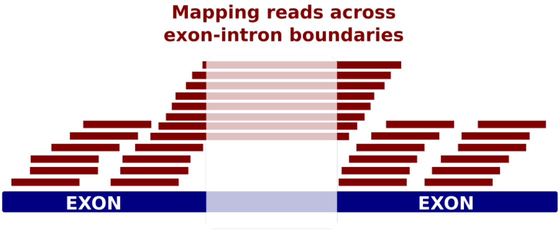
Mapping reads to a transcriptome



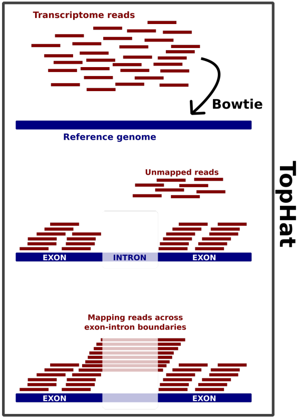
Mapping reads to a transcriptome



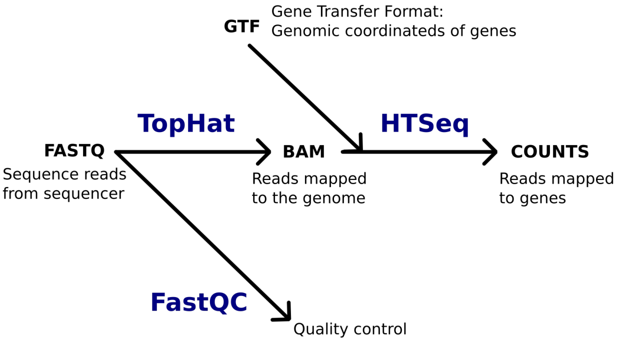
Mapping reads to a transcriptome



Mapping reads to a transcriptome: TopHat



Measuring Gene Expression Workflow



GTF Gene Transfer Format:
Genomic coordinateds of genes

FASTQ
Sequence reads
from sequencer

TopHat

BAM
Reads mapped
to the genome

HTSeq

COUNTS
Reads mapped
to genes

FastQC

Quality control