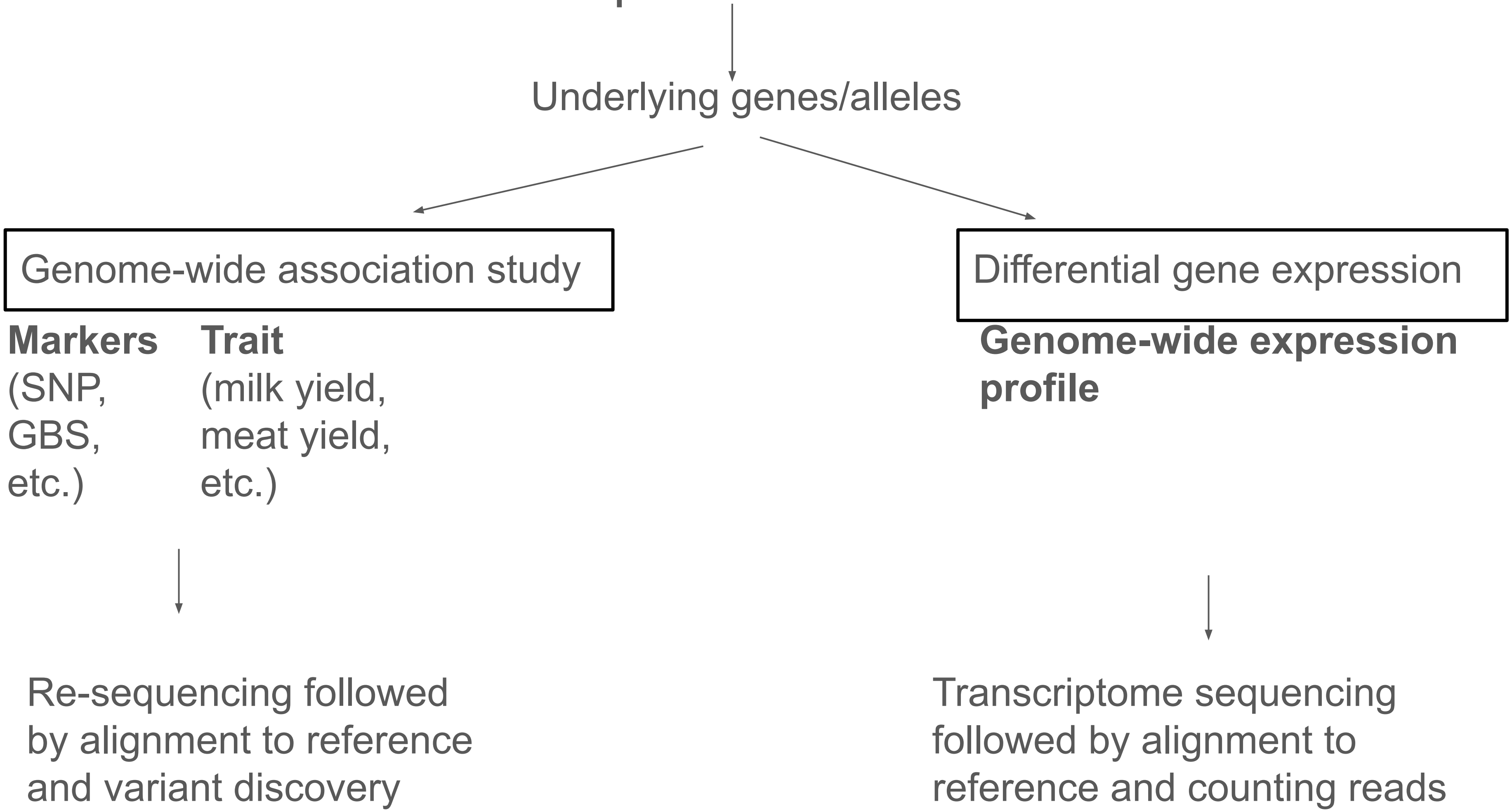
- Read alignment
-Exploration of alignment

Why and how?

NIAB NGS workshop-2024
Day 2 (afternoon session)

A reverse view of the workshop content

Genetic improvement of livestock



Underlying genes/alleles

Genome-wide association study

Markers (SNP, GBS, etc.)

Trait (milk yield, meat yield, etc.)

Re-sequencing followed by alignment to reference and variant discovery

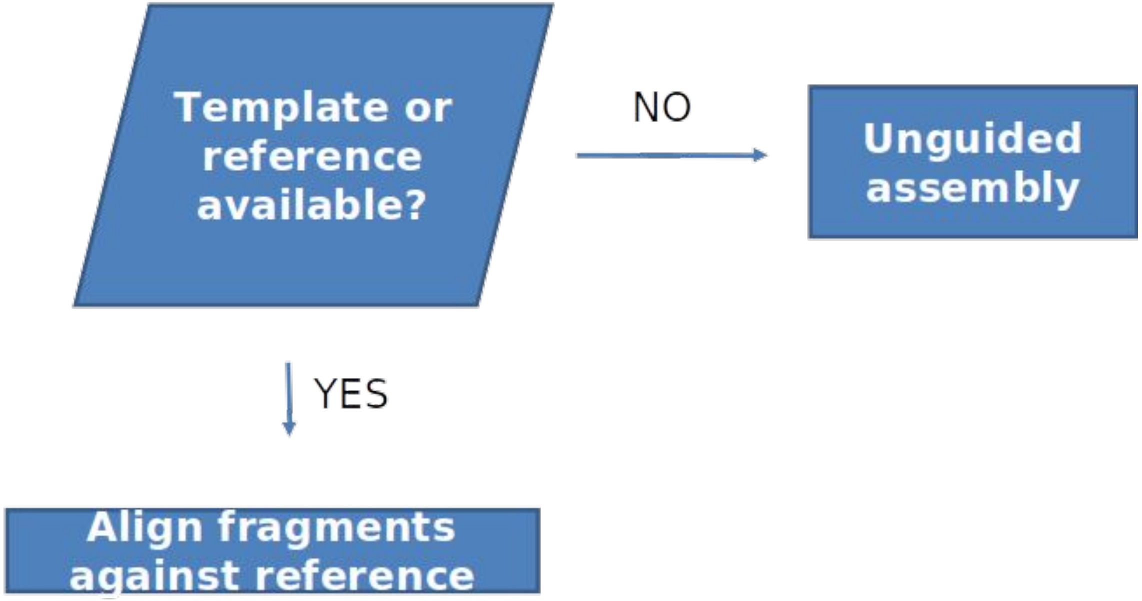
Differential gene expression

Genome-wide expression profile

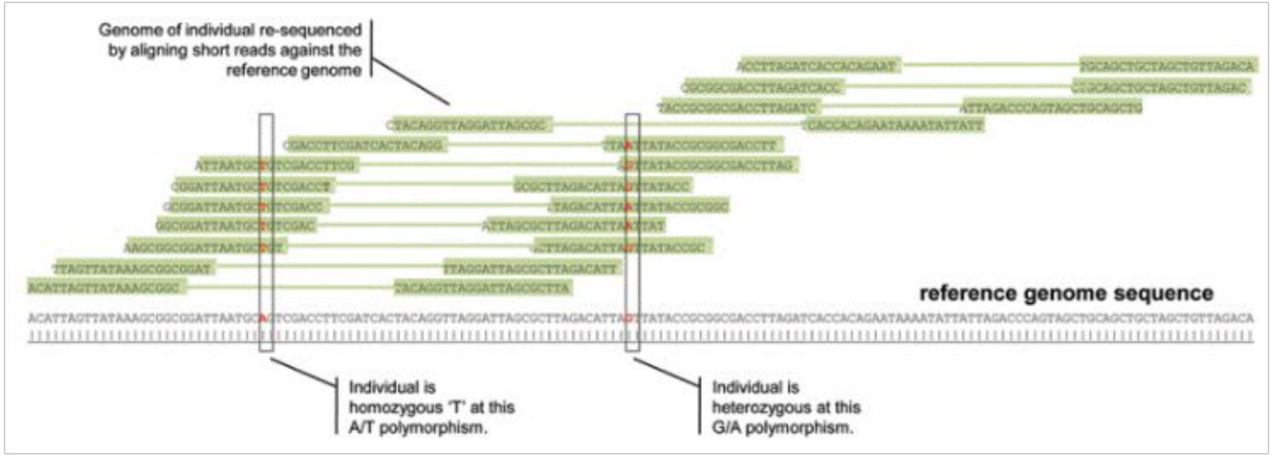
Transcriptome sequencing followed by alignment to reference and counting reads

Getting the original sequence back from fragments

- “Small *milate jao*, Large *banate jao*”


Template or reference available?
NO
Unguided assembly
YES
Align fragments against reference

Alignment against reference


Genome of individual re-sequenced by aligning short reads against the reference genome
reference genome sequence
ACATTAGTTATAAAGCGGCGGATTAATGCATCGACCTTCGATCACTACAGGTTAGGATTAGCGCTTAGACATTAGTTATACCGCGGCGACCTTAGATCACCACAGAATAAAATATTATTAGACCCAGTAGCTGCAGCTGCTAGCTGTTAGACA
Individual is homozygous 'T' at this A/T polymorphism.
Individual is heterozygous at this G/A polymorphism.

- Tools: BWA, bowtie2, star-aligner, soap, etc.

historyofnimr.org.uk

What if I’m interested in reference based assembly?

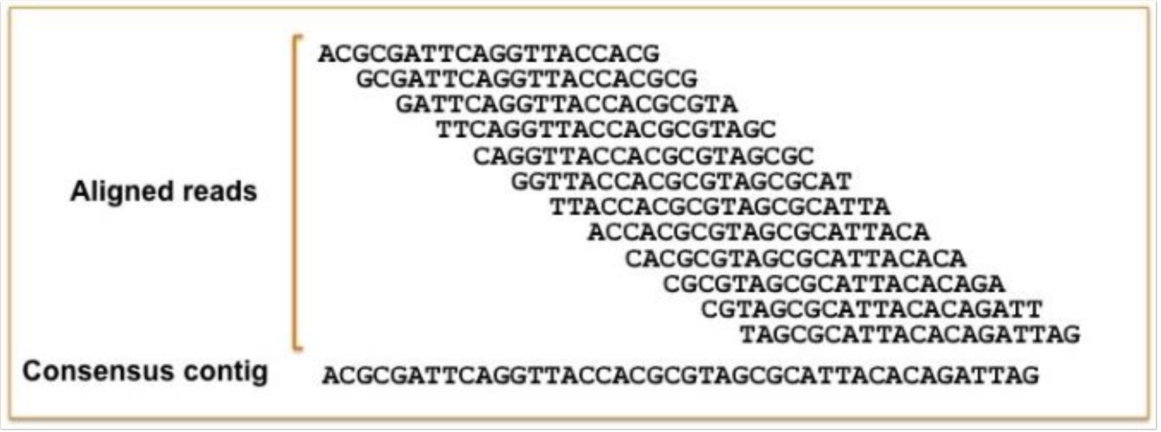
Preparing reference…

```
$ bwa index reference.fasta
```

Aligning short reads onto reference

```
$ bwa mem  reference.fasta
          /home/vivek/6966-1_R1.fastq.gz
          /home/vivek/6966-1_R2.fastq.gz
          >alignment.sam
```

De novo assembly



Aligned reads

```
ACGCGATTCAGGTTACCACG
  GCGATTCAGGTTACCACGCG
    GATTCAGGTTACCACGCGTA
      TTCAGGTTACCACGCGTAGC
        CAGGTTACCACGCGTAGCGC
          GGTTACCACGCGTAGCGCAT
            TTACCACGCGTAGCGCATTA
              ACCACGCGTAGCGCATTACA
                CACGCGTAGCGCATTACACA
                  CGCGTAGCGCATTACACAGA
                    CGTAGCGCATTACACAGATT
                      TAGCGCATTACACAGATTAG
```

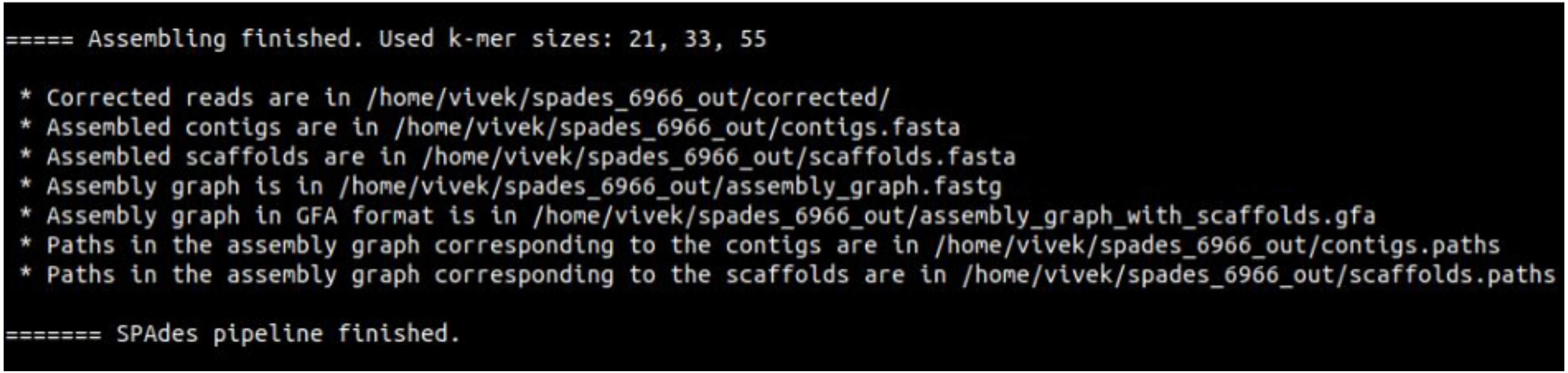
Consensus contig

```
ACGCGATTCAGGTTACCACGCGTAGCGCATTACACAGATTAG
```

- Tools: Velvet, Allpaths-LG, AbySS, soap-denovo

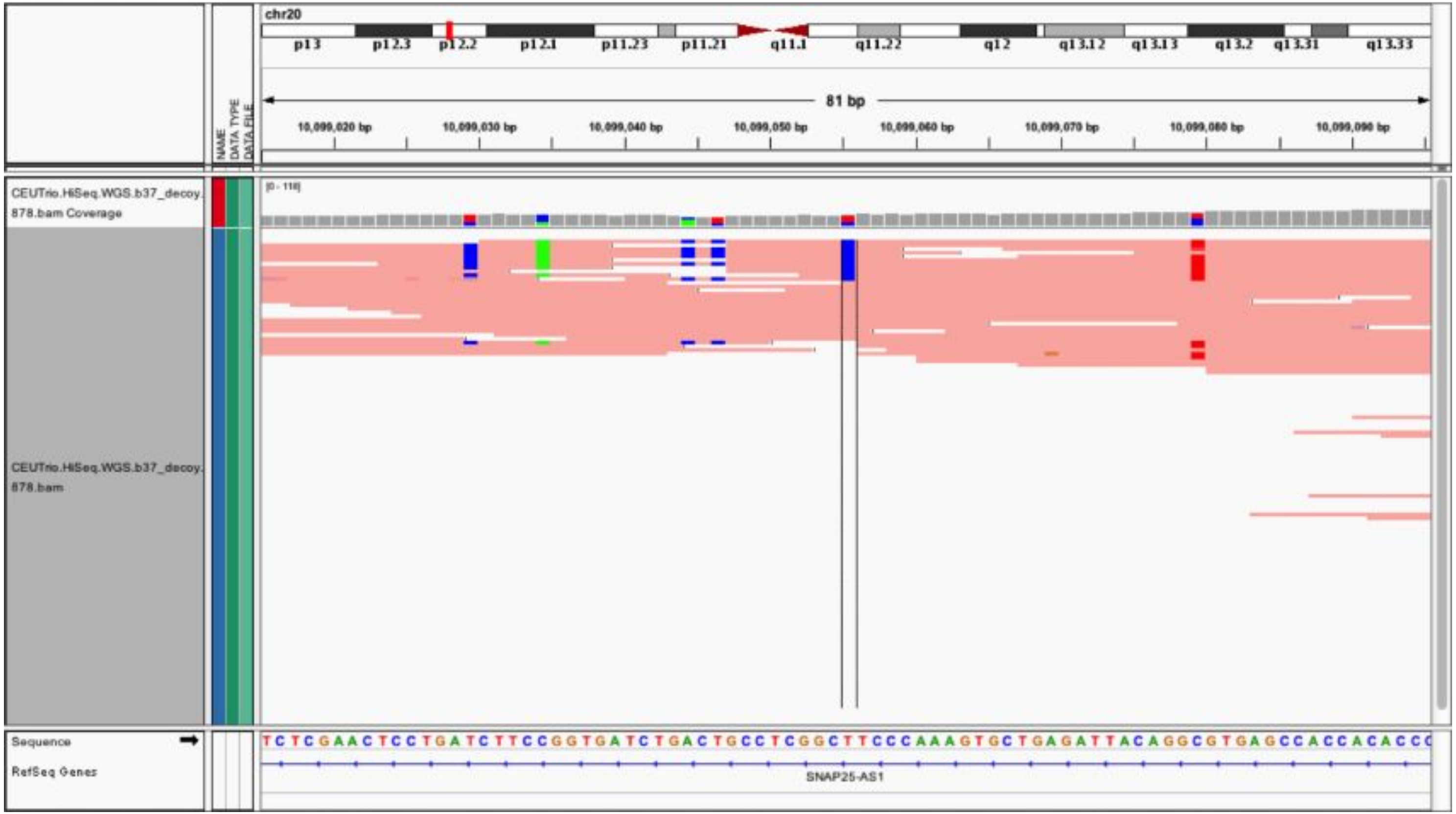
contig,.wordpress.com

```
SPAdes-3.12.0-Linux/bin/spades.py
--pe1-1 /home/vivek/6966-1_R1.fastq.gz
--pe1-2 /home/vivek/6966-1_R2.fastq.gz
-o      /home/vivek/spades_6966_out
-t      23
```

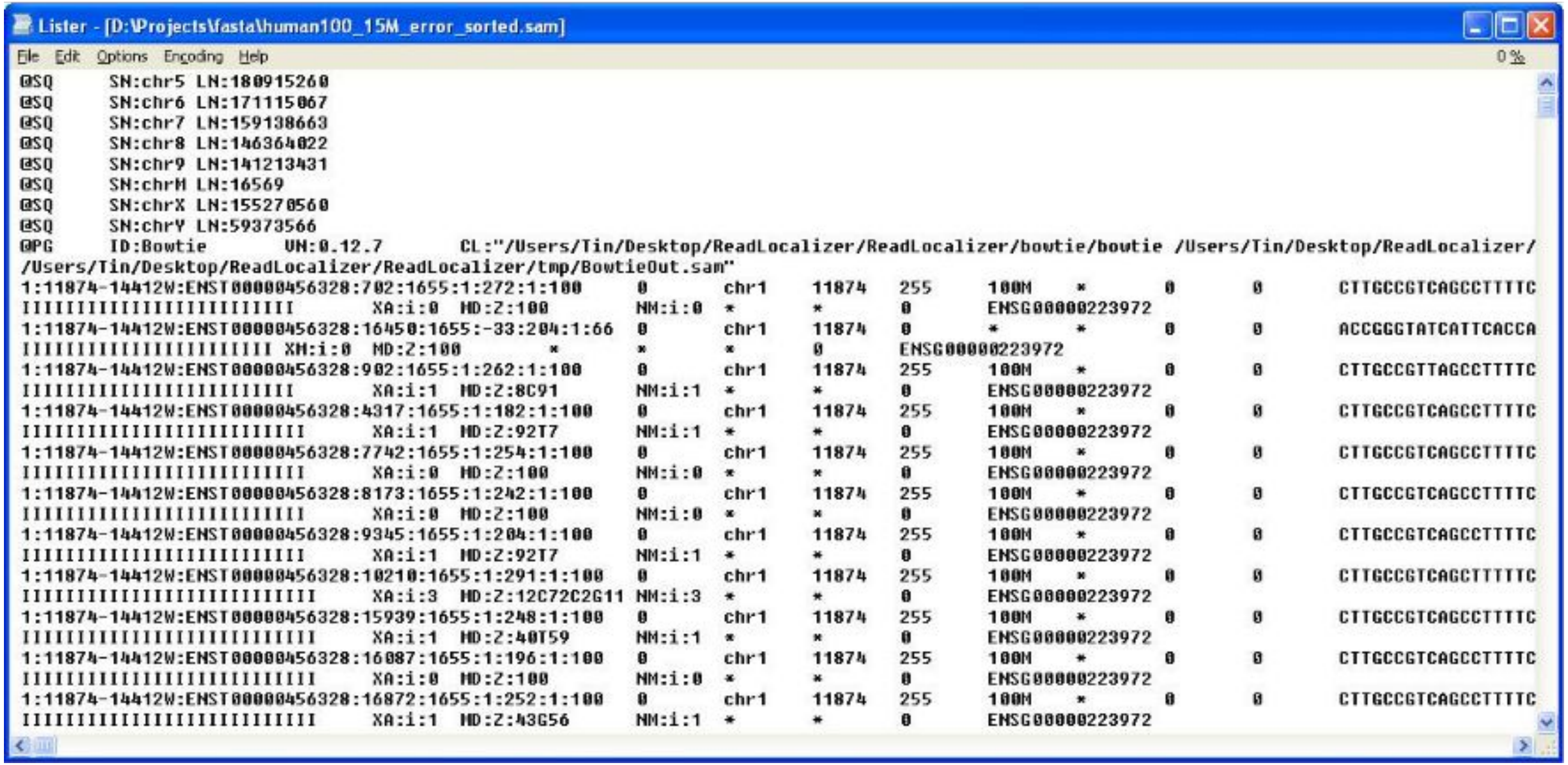


```
===== Assembling finished. Used k-mer sizes: 21, 33, 55

 * Corrected reads are in /home/vivek/spades_6966_out/corrected/
 * Assembled contigs are in /home/vivek/spades_6966_out/contigs.fasta
 * Assembled scaffolds are in /home/vivek/spades_6966_out/scaffolds.fasta
 * Assembly graph is in /home/vivek/spades_6966_out/assembly_graph.fastg
 * Assembly graph in GFA format is in /home/vivek/spades_6966_out/assembly_graph_with_scaffolds.gfa
 * Paths in the assembly graph corresponding to the contigs are in /home/vivek/spades_6966_out/contigs.paths
 * Paths in the assembly graph corresponding to the scaffolds are in /home/vivek/spades_6966_out/scaffolds.paths

======= SPAdes pipeline finished.
```

Viewing the alignment: IGV, Tablet

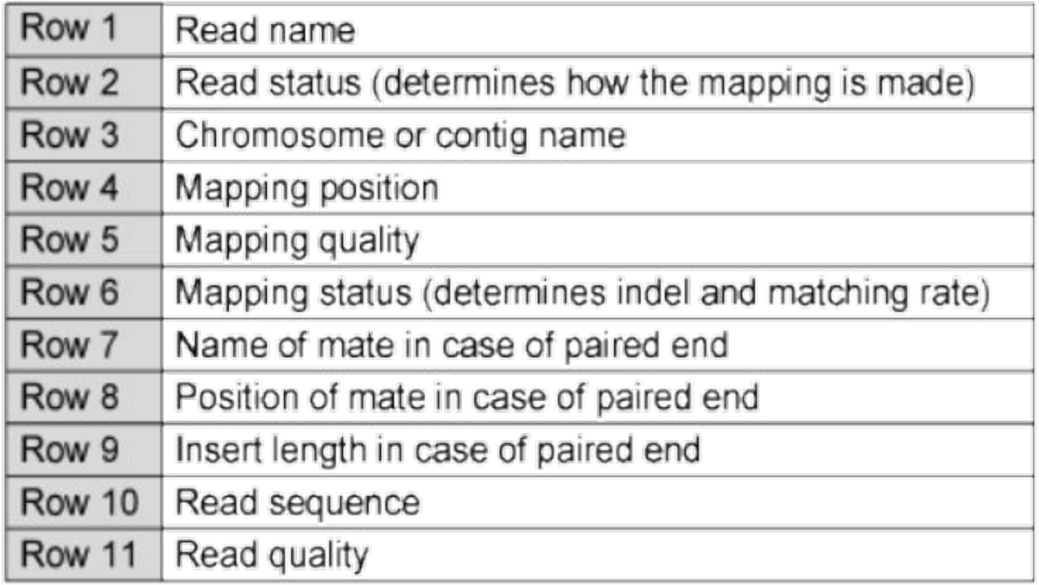


Alignments are saved in a particular format (SAM)



```
@SQ     SN:chr5 LN:180915260
@SQ     SN:chr6 LN:171115067
@SQ     SN:chr7 LN:159138663
@SQ     SN:chr8 LN:146364022
@SQ     SN:chr9 LN:141213431
@SQ     SN:chrM LN:16569
@SQ     SN:chrX LN:155270560
@SQ     SN:chrY LN:59373566
@PG     ID:Bowtie       VN:0.12.7       CL:"/Users/Tin/Desktop/ReadLocalizer/ReadLocalizer/bowtie/bowtie /Users/Tin/Desktop/ReadLocalizer/
/Users/Tin/Desktop/ReadLocalizer/ReadLocalizer/tmp/BowtieOut.sam"
1:11874-14412W:ENST00000456328:702:1655:1:272:1:100     0       chr1    11874   255     100M    *       0       0       CTTGCCGTCAGCCTTTTC
IIIIIIIIIIIIIIIIIIIIIIIII       XA:i:0  MD:Z:100        NM:i:0  *       *       0       ENSG00000223972
1:11874-14412W:ENST00000456328:16450:1655:-33:204:1:66  0       chr1    11874   0       *       *       0       0       ACCGGGTATCATTCACCA
IIIIIIIIIIIIIIIIIIIIIII XM:i:0  MD:Z:100        *       *       *       0       ENSG00000223972
1:11874-14412W:ENST00000456328:902:1655:1:262:1:100     0       chr1    11874   255     100M    *       0       0       CTTGCCGTTAGCCTTTTC
IIIIIIIIIIIIIIIIIIIIIIIII       XA:i:1  MD:Z:8C91       NM:i:1  *       *       0       ENSG00000223972
1:11874-14412W:ENST00000456328:4317:1655:1:182:1:100    0       chr1    11874   255     100M    *       0       0       CTTGCCGTCAGCCTTTTC
IIIIIIIIIIIIIIIIIIIIIIIIII      XA:i:1  MD:Z:92T7       NM:i:1  *       *       0       ENSG00000223972
1:11874-14412W:ENST00000456328:7742:1655:1:254:1:100    0       chr1    11874   255     100M    *       0       0       CTTGCCGTCAGCCTTTTC
IIIIIIIIIIIIIIIIIIIIIIIIII      XA:i:0  MD:Z:100        NM:i:0  *       *       0       ENSG00000223972
1:11874-14412W:ENST00000456328:8173:1655:1:242:1:100    0       chr1    11874   255     100M    *       0       0       CTTGCCGTCAGCCTTTTC
IIIIIIIIIIIIIIIIIIIIIIIIII      XA:i:0  MD:Z:100        NM:i:0  *       *       0       ENSG00000223972
1:11874-14412W:ENST00000456328:9345:1655:1:204:1:100    0       chr1    11874   255     100M    *       0       0       CTTGCCGTCAGCCTTTTC
IIIIIIIIIIIIIIIIIIIIIIIIII      XA:i:1  MD:Z:92T7       NM:i:1  *       *       0       ENSG00000223972
1:11874-14412W:ENST00000456328:10210:1655:1:291:1:100   0       chr1    11874   255     100M    *       0       0       CTTGCCGTCAGCTTTTTC
IIIIIIIIIIIIIIIIIIIIIIIIIII     XA:i:3  MD:Z:12C72C2G11 NM:i:3  *       *       0       ENSG00000223972
1:11874-14412W:ENST00000456328:15939:1655:1:248:1:100   0       chr1    11874   255     100M    *       0       0       CTTGCCGTCAGCCTTTTC
IIIIIIIIIIIIIIIIIIIIIIIIIII     XA:i:1  MD:Z:40T59      NM:i:1  *       *       0       ENSG00000223972
1:11874-14412W:ENST00000456328:16087:1655:1:196:1:100   0       chr1    11874   255     100M    *       0       0       CTTGCCGTCAGCCTTTTC
IIIIIIIIIIIIIIIIIIIIIIIIIII     XA:i:0  MD:Z:100        NM:i:0  *       *       0       ENSG00000223972
1:11874-14412W:ENST00000456328:16872:1655:1:252:1:100   0       chr1    11874   255     100M    *       0       0       CTTGCCGTCAGCCTTTTC
IIIIIIIIIIIIIIIIIIIIIIIIIII     XA:i:1  MD:Z:43G56      NM:i:1  *       *       0       ENSG00000223972
```

What individual columns mean?



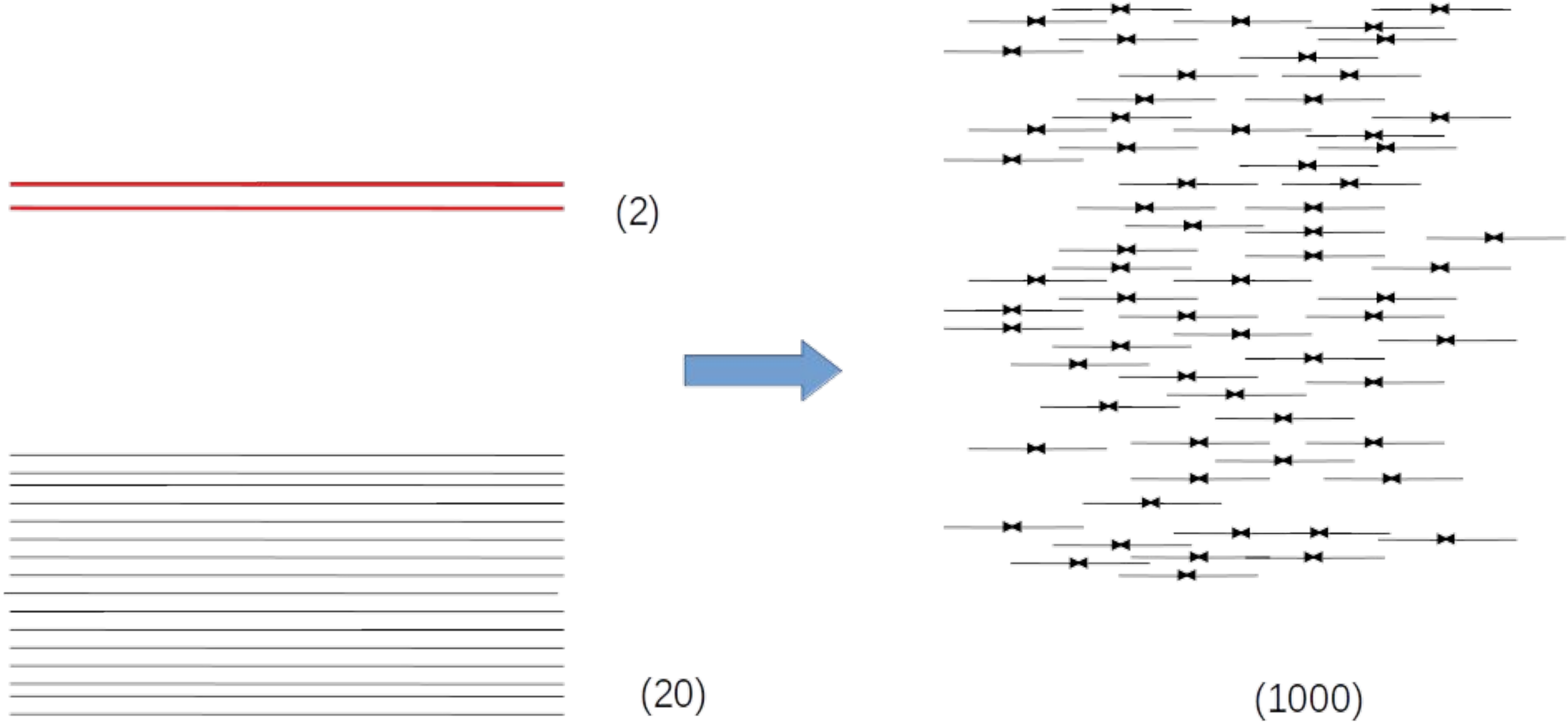
Row 1	Read name
Row 2	Read status (determines how the mapping is made)
Row 3	Chromosome or contig name
Row 4	Mapping position
Row 5	Mapping quality
Row 6	Mapping status (determines indel and matching rate)
Row 7	Name of mate in case of paired end
Row 8	Position of mate in case of paired end
Row 9	Insert length in case of paired end
Row 10	Read sequence
Row 11	Read quality

Datasets

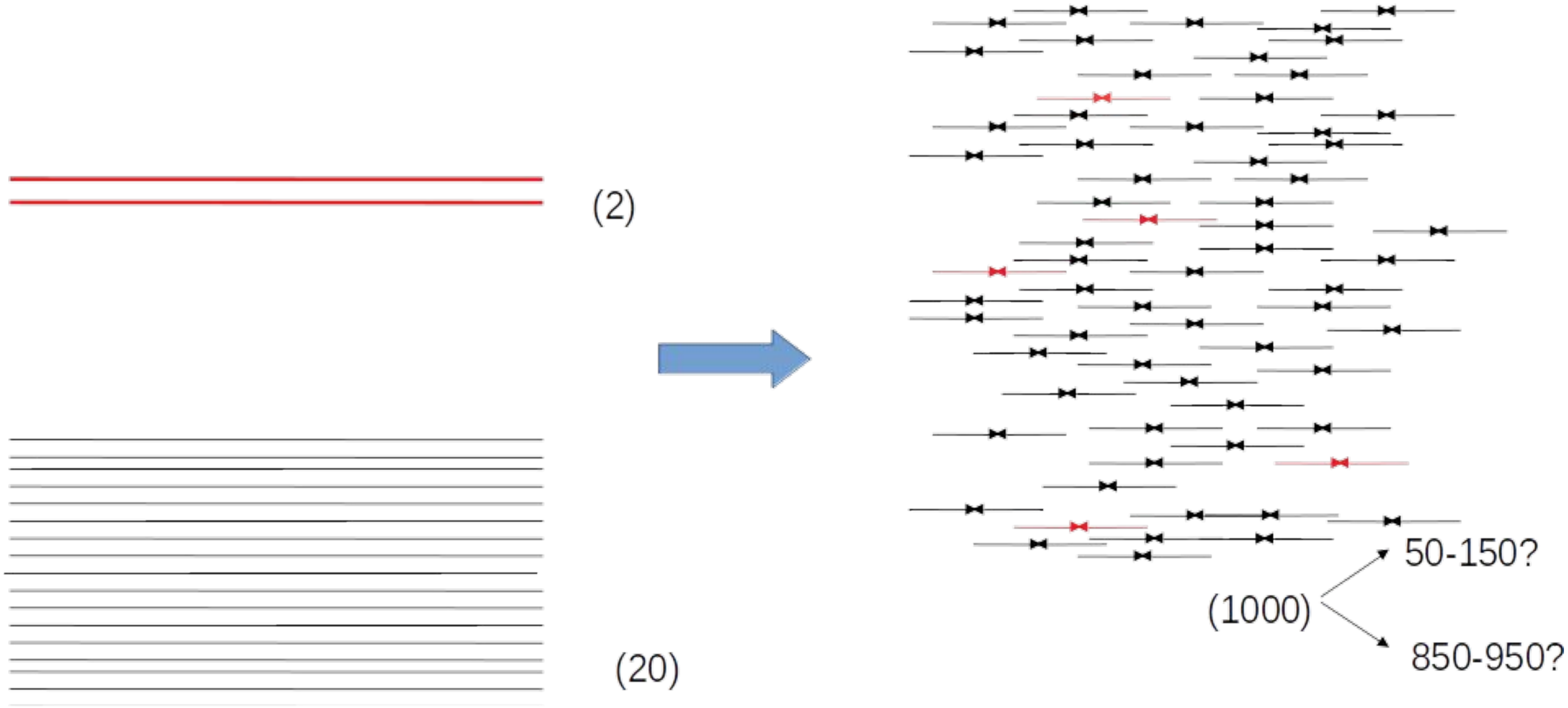
1. Whole genome sequencing
    1. Study: A Bioproject, which represent 565 whole genome sequences from Bos taurus dairy cattle. The collection represents both male and female animals, primarily of New Zealand Holstein-Friesian and Jersey ancestry, and crosses thereof.
    2. Data download: From the above study, only one sample. which was of size ~3.5 Gb.
2. Reference sequences: Whole genome and transcripts of either *Bos taurus indicus* or *Bos taurus taurus*, whichever is appropriate or available.

Gene expression profiling by sequencing

Era of digital expression: RNA-seq



Era of digital expression: RNA-seq



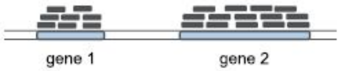
(2)

(20)

(1000)

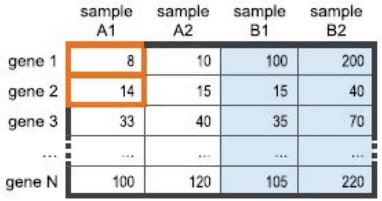
50-150?

850-950?



READ MAPPING

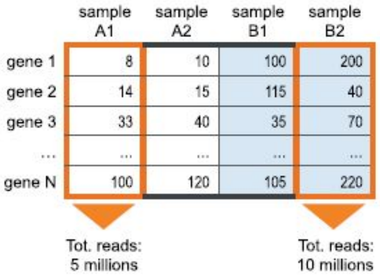
Find original read source within the reference genome or transcriptome.



	sample A1	sample A2	sample B1	sample B2
gene 1	8	10	100	200
gene 2	14	15	15	40
gene 3	33	40	35	70
...	...	...	...	...
gene N	100	120	105	220

COUNTS COMPUTATION

Estimate gene expression with "counts", i.e. with the number of reads mapped on each gene.

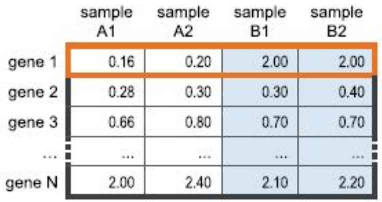


	sample A1	sample A2	sample B1	sample B2
gene 1	8	10	100	200
gene 2	14	15	115	40
gene 3	33	40	35	70
...	...	...	...	...
gene N	100	120	105	220

Tot. reads: 5 millions
Tot. reads: 10 millions

COUNTS NORMALIZATION

Eliminate biases to make expression levels comparable between samples (e.g. different sequencing depths of samples A1 and B2) and within samples (e.g. different lengths of gene 1 and gene 2).



	sample A1	sample A2	sample B1	sample B2
gene 1	0.16	0.20	2.00	2.00
gene 2	0.28	0.30	0.30	0.40
gene 3	0.66	0.80	0.70	0.70
...	...	...	...	...
gene N	2.00	2.40	2.10	2.20

DIFFERENTIAL EXPRESSION ANALYSIS

Identify genes with statistically different expression levels in the compared conditions (e.g. A and B).

Finotello and Di Camillo, BRIEFINGS IN FUNCTIONAL GENOMICS. VOL 14. NO 2. 130^142

Datasets

1. Transcriptome sequencing
    1. Study: A study which to find out early pregnancy markers identified by transcriptomic analysis in peripheral blood immune cells in beef heifers (GEO, Bioproject, SRA collection)
    2. Data download: There are 12 samples, 6 from the pregnant group and the remaining are non-pregnant. OUt of 12, at least two from each should be downloaded and analyzed.
2. Reference sequences: Whole genome and transcripts of either *Bos taurus indicus* or *Bos taurus taurus*, whichever is appropriate or available.

To build RSEM references:

```
rsem-prepare-reference --gff3 GCF_000001405.31_GRCh38.p5_genomic.gff \
             --trusted-sources BestRefSeq,Curated\ Genomic \
             --bowtie \
             GCF_000001405.31_GRCh38.p5_genomic.primary_assembly.fna \
             ref/human_refseq
```

SYNOPSIS: Calculate expression

```
rsem-calculate-expression [options] upstream_read_file(s) reference_name
sample_name
```

```
 rsem-calculate-expression [options] --paired-end upstream_read_file(s)
downstream_read_file(s) reference_name sample_name
```

```
rsem-calculate-expression [options] --alignments [--paired-end] input
reference_name sample_name
```

Calculate expression: with alignment data

```
rsem-calculate-expression --paired-end \
                          --alignments \
                          -p 8 \
                          /data/mmliver_paired_end_quals.bam \
                          /ref/mouse_125 \
                          mmliver_paired_end_quals
```

Calculate expression: with reads data

```
rsem-calculate-expression --paired-end \
                          --star \
                          --star-path /sw/STAR \
                          --gzipped-read-file \
                          -p 8 \
                          /data/mmliver_1.fq.gz \
                          /data/mmliver_2.fq.gz \
                          /ref/mouse_125 \
                          mmliver_paired_end_quals
```