# Nucleotide sequence analysis

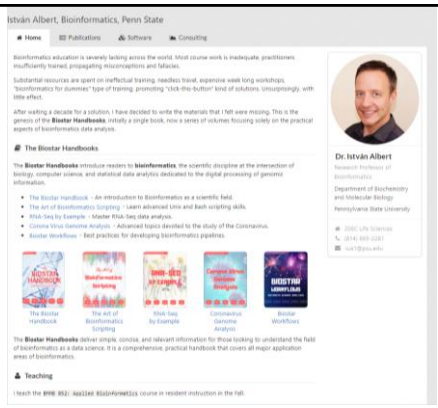
Bioinformatics, 2nd year
UP FAMNIT

## Plan for lectures

- RNA-Seq (transcriptome analysis)
- Genome assembly
- SNP calling
- Metabarcoding
- Genome wide association analysis

## Literature

- Pevsner, J. 2015. Bioinformatics and functional genomics. Wiley Blackwell.
- Material for practicals (e-classroom)
- Korpelainen et. al. 2015. RNA-seq Data Analysis: A Practical Approach.
- István Albert. The Biostar Handbooks https://www.biostarhandbook.com/
- scientific articles




https://www.biostarhandbook.com/

## Obligations and Grading system

Please be advised that attendance at practical sessions is mandatory.

**The final grade** will be composed based on the following criteria:

70% Project work

30% Oral exam

**Student's obligations**:

Each student must analyze their own dataset and prepare a laboratory notebook with a documented workflow.

*Include all commands and explain them, including options (flags) selected with each program. Document your work with screenshots and output files. Interpret the results, including graphs obtained with FastQC. Report the number of reads in your dataset, how many were removed after quality filtering, etc. Answer all questions from the worksheets.*

Laboratory notebooks should be uploaded to the e-classroom.

The report must be submitted at least 1 week before the exam date.

Please note: You cannot attend the oral exam if you haven't submitted the notebook before the deadline!

Oral Defense:

During the oral defense, expect questions related to the project content or general questions about nucleotide sequence analysis, such as "Describe the SAM file! How can it be converted to a BAM file?" and questions related to the lectures.

The oral defense will occur on the date specified in ŠIS.

## Transcriptome analysis (RNA-Seq)

## Low- and high-throughput technologies to study mRNAs

Three techniques for the study of mRNA:

- complementary DNA (cDNA) libraries
- microarrays (e.g. using the Affymetrix platform)
- RNA-seq (Chapter 11)

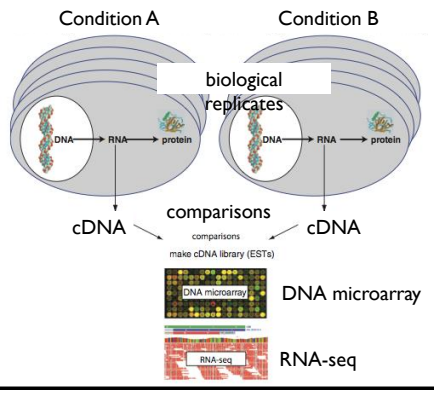
Low throughput techniques (Northern blots, PCR) may seem laborious and able to provide only limited amounts of information.

Yet they also serve as trusted "gold standards" and provide crucial validation of high throughput techniques.

B&FG 3e
Page 452

Source: Pevsner. Bioinformatics and Functional Genomics, 3rd Edition

## Gene expression measured with high-throughput technologies

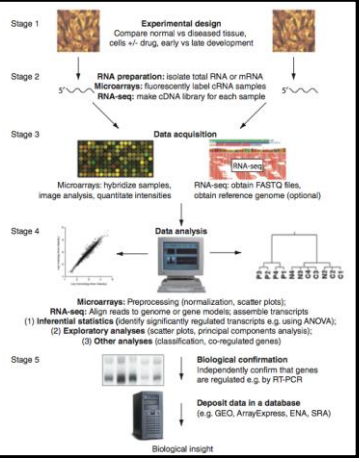



B&FG 3e
Fig. 10.11
Page 454

## Transcriptome analysis (RNA-Seq)

- High throughput RNA sequencing (RNA-Seq) offers possibility to investigate the expression profiles at the transcriptional level and also identifying novel and non-coding transcripts
- The **reference-based** transcriptome analysis method is based on aligning the sequenced reads to a pre-existing reference genome, (followed by assembling overlapping alignments into transcripts).
- In contrast, the reference-free **de novo** transcriptome analysis method allows to directly assemble sequenced reads into transcripts by using high levels of redundancy and overlapping of reads, without using a reference genome.

Overview of the process of generating high-throughput gene expression data using microarrays or RNA-seq




B&FG 3e
Fig. 10.13
Page 462

Alignment based vs classification based RNA-Seq

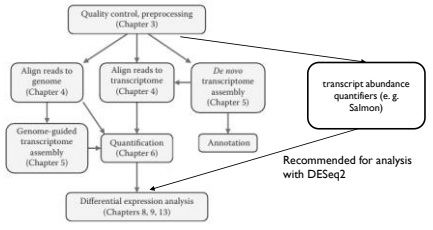



FIGURE 2.1 Possible paths in RNA-seq data analysis. In the beginning, the quality of reads is checked and if necessary, reads are preprocessed to remove low-quality data and artifacts. The read's origin is then identified by aligning them to a reference genome if available. Novel genes and transcripts are detected using genome-guided transcriptome assembly, and gene and transcript expression are quantified. Alternatively, gene and transcript discovery can be skipped and expression quantified only for known genes and transcripts. If the reference genome is not available, reads can be aligned and quantified using a reference transcriptome instead. If a transcriptome is not available, it can be produced from reads using *de novo* transcriptome assembly. When abundance estimates are obtained using one of these paths, expression differences between sample groups can be analyzed using statistical testing. The details of each step can be found in the chapters indicated in parentheses.

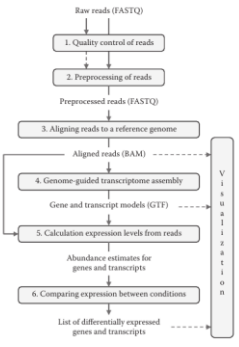



FIGURE 2.3 Differential expression analysis workflow consists of several, interrelated steps. The typical output file formats are indicated in parentheses.

**What are the tradeoffs?**

Genome base methods:
1. Allow us to visualize the data in the context of the entire genome.
2. Enable us to discover/validate new transcripts.

Transcriptome and classification based methods:
1. Are typically more accurate.
2. Require lower computational resources.

## Counting Reads per Genes

Counting Reads per Genes

- The simplest way of estimating expression is to count reads per genes. Several tools are available: HTSeq, Cufflinks, RSubread (Bioconductor package)

Counting Reads per Transcripts

Counting Reads per Exons

What is the final result of an RNA-Seq analysis?

The result of an RNA-Seq analysis is a *quantification matrix*. For our toy example, the file might look like this:

| name | control | shock |
|---|---|---|
| Gene A | 100 | 200 |
| Gene B | 80 | 60 |
| Gene C | 120 | 180 |
| ... | | |

## Article for practical work

Received: 23 February 2022 | Revised: 4 September 2022 | Accepted: 13 September 2022

DOI: 10.1111/mec.16708

ORIGINAL ARTICLE

MOLECULAR ECOLOGY WILEY

**Coral larvae suppress heat stress response during the onset of symbiosis decreasing their odds of survival**

Sheila A. Kitchen[1] | Duo Jiang[2] | Saki Harii[3] | Noriyuki Satoh[4] | Virginia M. Weis[1] | Chuya Shinzato[4]

[1]Department of Integrative Biology, Oregon State University, Corvallis, Oregon, USA
[2]Statistics Department, Oregon State University, Corvallis, Oregon, USA
[3]Tropical Biosphere Research Center, University of the Ryukyus, Okinawa, Japan
[4]Marine Genomics Unit, Okinawa Institute of Science and Technology Graduate University, Okinawa, Japan

Correspondence
Sheila A. Kitchen, Department of Biology and Biological Engineering, California Institute of Technology, Pasadena, CA, USA.
Email: sakitchen@caltech.edu

Present address
Chuya Shinzato, Division of Marine Life Science, Department of Marine Biosciences, Atmosphere and Ocean Research Institute, University of Tokyo, Chiba, Japan

Funding information
Japan Society for the Promotion of Science, Grant/Award Number: GP-18007; Office of International Science and Engineering, Grant/Award Number: OISE-1545803; National Science Foundation, Grant/Award Number: 1122191; Sigma Xi

Handling Editor: Pim Bongaerts

Abstract
The endosymbiosis between most corals and their photosynthetic dinoflagellate partners begins early in the host life history, when corals are larvae or juvenile polyps. The capacity of coral larvae to buffer climate-induced stress while in the process of symbiont acquisition could come with physiological trade-offs that alter behaviour, development, settlement and survivorship. Here we examined the joint effects of thermal stress and symbiosis onset on colonization dynamics, survival, metamorphosis and host gene expression of *Acropora digitifera* larvae. We found that thermal stress decreased symbiont colonization of hosts by 50% and symbiont density by 98.5% over 2 weeks. Temperature and colonization also influenced larval survival and metamorphosis in an additive manner, where colonized larvae fared worse or prematurely metamorphosed more often than noncolonized larvae under thermal stress. Transcriptomic responses to colonization and thermal stress treatments were largely independent, while the interaction of these treatments revealed contrasting expression profiles of genes that function in the stress response, immunity, inflammation and cell cycle regulation. The combined treatment either cancelled or lowered the magnitude of expression of heat-stress responsive genes in the presence of symbionts, revealing a physiological cost to acquiring symbionts at the larval stage with elevated temperatures. In addition, host immune suppression, a hallmark of symbiosis onset under ambient temperature, turned to immune activation under heat stress. Thus, by integrating the physical environment and biotic pressures that mediate prosettlement event in corals, our results suggest that colonization may hinder larval survival and recruitment under projected climate scenarios.

KEYWORDS
cell cycle arrest, cellular senescence, coral larvae, heat stress, inflammation, symbiosis onset

OXFORD

*Briefings in Bioinformatics*, 2022, **23(2)**, 1–30
https://doi.org/10.1093/bib/bbab563
**Review**

## A simple guide to *de novo* transcriptome assembly and annotation

Venket Raghavan[1], Louis Kraft[1], Fantin Mesny[2] and Linda Rigerte[2]

Corresponding authors: Venket Raghavan, Quantitative and Computational Biology, Max Planck Institute for Biophysical Chemistry, 37077 Göttingen, Germany. E-mail: vraghav@mpibpc.mpg.de; Louis Kraft, Quantitative and Computational Biology, Max Planck Institute for Biophysical Chemistry, 37077 Göttingen, Germany. E-mail: louis.kraft@mpibpc.mpg.de
[1]These authors are joint first coauthors.
[2]These authors are joint second coauthors.

**Abstract**

A transcriptome constructed from short-read RNA sequencing (RNA-seq) is an easily attainable proxy catalog of protein-coding genes when genome assembly is unnecessary, expensive or difficult. In the absence of a sequenced genome to guide the reconstruction process, the transcriptome must be assembled *de novo* using only the information available in the RNA-seq reads. Subsequently, the sequences must be annotated in order to identify sequence-intrinsic and evolutionary features in them (for example, protein-coding regions). Although straightforward at first glance, *de novo* transcriptome assembly and annotation can quickly prove to be challenging undertakings. In addition to familiarizing themselves with the conceptual and technical intricacies of the tasks at hand and the numerous pre- and post-processing steps involved, those interested must also grapple with an overwhelmingly large choice of tools. The lack of standardized workflows, fast pace of development of new tools and techniques and paucity of authoritative literature have served to exacerbate the difficulty of the task even further. Here, we present a comprehensive overview of *de novo* transcriptome assembly and annotation. We discuss the procedures involved, including pre- and post-processing steps, and present a compendium of corresponding tools.

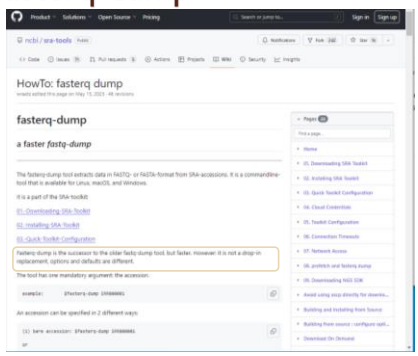
## SRA toolkit: `fastq-dump` to obtain FASTQ formatted data

```
$ fastq-dump -X 3 -Z SRR390728
Read 3 spots for SRR390728
Written 3 spots for SRR390728
@SRR390728.1 1 length=72
CATTCTTCACGTAGTTCTCGAGCCTTGGTTTTCAGCGATGGAGAATGACTTTGACAAGCTGAGAGAAGNTNC
+SRR390728.1 1 length=72
;;;;;;;;;;;;;;;;;;;;;;;;;;;;9;;665142;;;;;;;;;;;;;;;;;;;;;;;;;;;;;96&&&&(
@SRR390728.2 2 length=72
AAGTAGGTCTCGTCTGTGTTTTCTACGAGCTTGTGTTCCAGCTGACCCACTCCCTGGGTGGGGGGACTGGGT
+SRR390728.2 2 length=72
;;;;;;;;;;;;;;;;4;;;;3;393.1+4&&5&&;;;;;;;;;;;;;;;;;;;;;;;<9;<;;;;;464262
@SRR390728.3 3 length=72
CCAGCCTGGCCAACAGAGTGTTACCCCGTTTTTACTTATTTATTATTATTATTTTGAGACAGAGCATTGGTC
+SRR390728.3 3 length=72
-;;;8;;;;;;;,*;;'|p-4,44;,:&,1,4'./&19;;;;;;669;;99;;;;;-;3;2;0;+;744262/
```

NCBI offers the SRA Toolkit to manipulate sequence data. The `fastq-dump` command can pull FASTQ-formatted data from an accession number (such as SRR390728).

B&FG 3e
Page 393

## fasterq-dump




https://github.com/ncbi/sra-tools/wiki/HowTo:-fasterq-dump

## SRA toolkit: `fastq-dump` to obtain FASTA formatted data

```
$ fastq-dump -X 3 -Z SRR390728 -fasta 36
Read 3 spots for SRR390728
Written 3 spots for SRR390728
>SRR390728.1 1 length=72
CATTCTTCACGTAGTTCTCGAGCCTTGGTTTTCAGC
GATGGAGAATGACTTTGACAAGCTGAGAGAAGNTNC
>SRR390728.2 2 length=72
AAGTAGGTCTCGTCTGTGTTTTCTACGAGCTTGTGT
TCCAGCTGACCCACTCCCTGGGTGGGGGGACTGGGT
>SRR390728.3 3 length=72
CCAGCCTGGCCAACAGAGTGTTACCCCGTTTTTACT
TATTTATTATTATTATTTTGAGACAGAGCATTGGTC
```

B&FG 3e
Page 393

## FASTQ format

The FASTQ format stores DNA sequence data as well as associated Phred quality scores of each base.

```
@EAS54_6_R1_2_1_413_324
CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;;;;7;;;;;;;88
@EAS54_6_R1_2_1_540_792
TTGGCAGGCCAAGGCCGATGGATCA
+
;;;;;;;;;;;7;;;;;-;;;3;83
@EAS54_6_R1_2_1_443_348
GTTGCTTCTGGCGTGGGTGGGGGGG
+EAS54_6_R1_2_1_443_348
;;;;;;;;;;;9;7;;.7;393333
```

DNA read

Base quality score

| Dec | Char | Dec | Char | Sanger FASTQ | Dec | Char | Sanger FASTQ | Dec | Char | Sanger FASTQ |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | Non-printing | 32 | Space | | 64 | @ | 31 | 96 | ` | 63 |
| 1 | Non-printing | 33 | ! | 0 | 65 | A | 32 | 97 | a | 64 |
| 2 | Non-printing | 34 | " | 1 | 66 | B | 33 | 98 | b | 65 |
| 3 | Non-printing | 35 | # | 2 | 67 | C | 34 | 99 | c | 66 |
| 4 | Non-printing | 36 | $ | 3 | 68 | D | 35 | 100 | d | 67 |
| 5 | Non-printing | 37 | % | 4 | 69 | E | 36 | 101 | e | 68 |
| 6 | Non-printing | 38 | & | 5 | 70 | F | 37 | 102 | f | 69 |
| 7 | Non-printing | 39 | ' | 6 | 71 | G | 38 | 103 | g | 70 |
| 8 | Non-printing | 40 | ( | 7 | 72 | H | 39 | 104 | h | 71 |
| 9 | Non-printing | 41 | ) | 8 | 73 | I | 40 | 105 | i | 72 |
| 10 | Non-printing | 42 | * | 9 | 74 | J | 41 | 106 | j | 73 |
| 11 | Non-printing | 43 | + | 10 | 75 | K | 42 | 107 | k | 74 |
| 12 | Non-printing | 44 | , | 11 | 76 | L | 43 | 108 | l | 75 |
| 13 | Non-printing | 45 | - | 12 | 77 | M | 44 | 109 | m | 76 |
| 14 | Non-printing | 46 | . | 13 | 78 | N | 45 | 110 | n | 77 |
| 15 | Non-printing | 47 | / | 14 | 79 | O | 46 | 111 | o | 78 |
| 16 | Non-printing | 48 | 0 | 15 | 80 | P | 47 | 112 | p | 79 |
| 17 | Non-printing | 49 | 1 | 16 | 81 | Q | 48 | 113 | q | 80 |
| 18 | Non-printing | 50 | 2 | 17 | 82 | R | 49 | 114 | r | 81 |
| 19 | Non-printing | 51 | 3 | 18 | 83 | S | 50 | 115 | s | 82 |
| 20 | Non-printing | 52 | 4 | 19 | 84 | T | 51 | 116 | t | 83 |
| 21 | Non-printing | 53 | 5 | 20 | 85 | U | 52 | 117 | u | 84 |
| 22 | Non-printing | 54 | 6 | 21 | 86 | V | 53 | 118 | v | 85 |
| 23 | Non-printing | 55 | 7 | 22 | 87 | W | 54 | 119 | w | 86 |
| 24 | Non-printing | 56 | 8 | 23 | 88 | X | 55 | 120 | x | 87 |
| 25 | Non-printing | 57 | 9 | 24 | 89 | Y | 56 | 121 | y | 88 |
| 26 | Non-printing | 58 | : | 25 | 90 | Z | 57 | 122 | z | 89 |
| 27 | Non-printing | 59 | ; | 26 | 91 | [ | 58 | 123 | { | 90 |
| 28 | Non-printing | 60 | < | 27 | 92 | \ | 59 | 124 | \| | 91 |
| 29 | Non-printing | 61 | = | 28 | 93 | ] | 60 | 125 | } | 92 |
| 30 | Non-printing | 62 | > | 29 | 94 | ^ | 61 | 126 | ~ | 93 |
| 31 | Non-printing | 63 | ? | 30 | 95 | _ | 62 | 127 | DEL | |

FASTQ quality scores use ASCII characters

…relating quality scores (e.g. Q30 for 1 in $10^{-3}$ error rate) to a compact, one character symbol

B&FG 3e Fig.9-8 Page 392

You do not need to learn the one character symbols, but you should know the importance of base quality scores in sequence analysis.

## FASTQ format: Phred scores define quality

The FASTQ format stores DNA sequence data as well as associated Phred quality scores of each base.

$$Q_{\text{PHRED}} = -10 \times \log_{10}(P_e)$$

| Phred quality score | Probability of incorrect base call | Base call accuracy |
|---|---|---|
| 10 | 1 in 10 | 90% |
| 20 | 1 in 100 | 99% |
| 30 | 1 in 1,000 | 99.9% |
| 40 | 1 in 10,000 | 99.99% |
| 50 | 1 in 100,000 | 99.999% |

**Visualizing sequencing quality data**

- FastQC tool
https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

**Quality control (QC)** – improving data with removing identifiable errors
- Fastp
- Trimmomatic
- Cutadapt
- AdapterRemoval
- BBDuk
- …

**Sequencing adapter trimming** (same tools as previously mention)

## Where genomes are available?

International Nucleotide Sequence Database Collaboration (INSDC) members [National Center for Biotechnology Information (NCBI), European Bioinformatics Institute (EBI), and DNA Data Bank of Japan (DDBJ)]


## Where genomes are available?



https://ngdc.cncb.ac.cn/gwh/

## Alignment based vs classification based RNA-Seq

Aligners

- HISAT2
- STAR
- Bowtie (a splice unaware aligner)
- BWA (a splice unaware aligner)
- …

A classification-based RNA-Seq works via so-called “pseudo-alignments”.

Pseudo-alignment tools

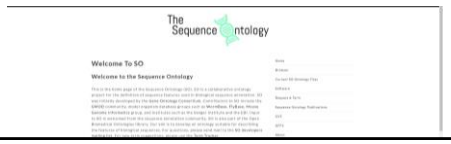
- Kallisto
- Salmon

## Indexing reference genome/transcriptome

- Indexing reference genome
  - Incorporating the annotation into the index

Annotation file


The formal specification of GFF3 is on the Sequence Ontology web site.



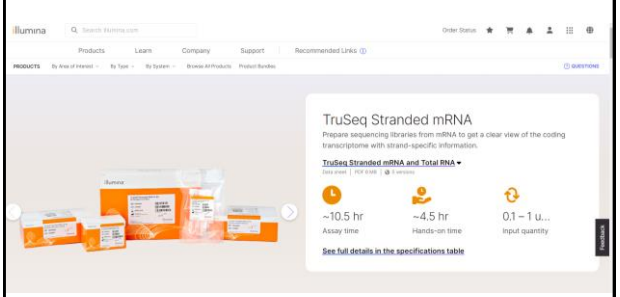

## Stranded vs non-stranded RNA-Seq

- Stranded (strand-specific) RNA-Seq

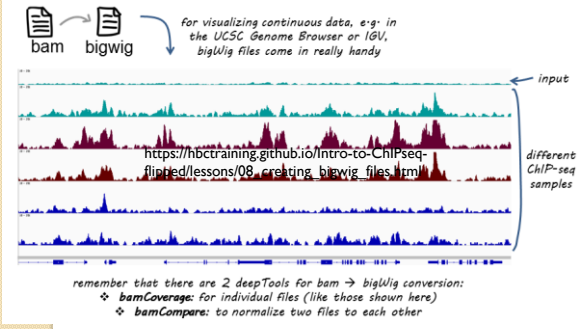
Preserves strand information—all mapped reads are aligned in the direction of transcription relative to the chromosomal strand

Ion Total RNA-Seq Kit v2

https://www.thermofisher.com/document-connect/document-connect.html?url=https://assets.thermofisher.com/TFS-Assets%2FLSG%2Fmanuals%2FMAN00010654_IonTotalRNASeqKit_v2_UG.pdf

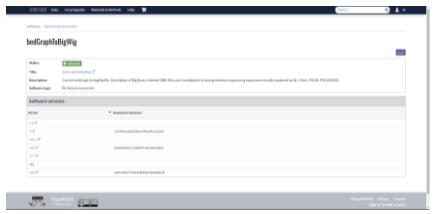



https://www.illumina.com/products/by-type/sequencing-kits/library-prep-kits/truseq-stranded-mrna.html




https://hbctraining.github.io/Intro-to-ChIPseq-flipped/lessons/08_creating_bigwig_files.html

https://hbctraining.github.io/Intro-to-ChIPseq-flipped/lessons/08_creating_bigwig_files.html




```
# Turn each BAM file into bedGraph coverage. The files will have the .bg extension
cat ids.txt | parallel "bedtools genomecov -ibam bam/{}.bam -split -bg > bam/{}.bg

# Convert each bedGraph coverage into bigWig coverage. The files will have the .bw
cat ids.txt | parallel "bedGraphToBigWig bam/{}.bg $(IDX).fai bam/{}.bw"
```

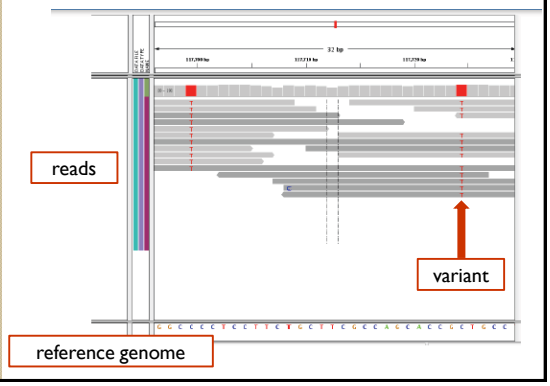
The resulting bedGraph and bigWig files will have *.bg and *.bw extensions and are placed in the bam directory.

You can drag your bigWig files in the IGV[7] panels, and the coverage information will load up much-much faster. Below I have loaded all samples, re-sized the tracks, and colored them by samples. I have also turned on logarithmic and automatic scaling. The resulting browser track is quite informative:

Biostar handbook

https://www.encodeproject.org/software/bedgraphtobigwig/

## Next-generation sequence data: visualizing of short reads aligned to a reference genome



## BWA and other aligners produce output in the SAM format

| Column | Description |
|---|---|
| 1 QNAME | Query (pair) NAME |
| 2 FLAG | bitwise FLAG |
| 3 RNAME | Reference sequence NAME |
| 4 POS | 1-based leftmost POSition/coordinate of clipped sequence |
| 5 MAPQ | MAPping Quality (Phred-scaled) |
| 6 CIGAR | extended CIGAR string |
| 7 MRNM | Mate Reference sequence NaMe ('=' if same as RNAME) |
| 8 MPOS | 1-based Mate POSition |
| 9 ISIZE | Inferred insert SIZE |
| 10 SEQ | query SEQuence on the same strand as the reference |
| 11 QUAL | query QUALity (ASCII-33 gives the Phred base quality) |
| 12 OPT | variable OPTional fields in the format TAG:VTYPE:VALU |

https://www.samformat.info/sam-format-flag
https://broadinstitute.github.io/picard/explain-flags.html

## Sequence alignment/map format (SAM) and BAM

- SAM is a common format having sequence reads and their alignment to a reference genome.
- BAM is the binary form of a SAM file.
- Aligned BAM files are available at repositories (Sequence Read Archive at NCBI, ENA at Ensembl)
- SAMTools is a software package commonly used to analyze SAM/BAM files.
- Visit http://samtools.sourceforge.net/

## Anatomy of a Sequence Alignment/Map (SAM) file

(1) The query name of the read is given (M01121…)

(2) The FLAG value is 163 (this equals 1+2+32+128)

(3) The reference sequence name, chrM, refers to the mitochondrial genome

(4) Position 480 is the left-most coordinate position of this read

(5) The Phred-scaled mapping quality is 60 (an error rate of 1 in $10^6$)

(6) The CIGAR string (148M2S) shows 148 matches and 2 soft-clipped (unaligned) bases

```
home/bioinformatics$ samtools view 030c_S7.bam | less
M01121:5:000000000-A2DTN:1:2111:20172:15571     163     chrM
480     60      148M2S  =       524     195     AATCTCATCAAT
ACAACCCTCGCCCATCCTACCCAGCACACACACACCGCTGCTAACCCCATACCCCGAACC
AACCAAACCCCAAAGACACCCCCCACAGTTTATGTAGCTTACCTCCTCAAAGCAATAACC
TGAAAATGTTTAGACGGG  BBBBBFFB5@FFGGGFGEGGGEGAAACGHFHFEGGAGFFH
AEFDGG?E?EGGGFGHFGHF?FFCHFH00E@EGFGGEEE1FFEEEHBGEFFFGGGG@</0
1BG212222>F21@F11FGFG1@1?GC<G11?1?FGDGGF=GHFFFHC.-
RG:Z:Sample7    XC:i:148        XT:A:U  NM:i:3  SM:i:37
AM:i:37 X0:i:1  X1:i:0  XM:i:3  XO:i:0  XG:i:0  MD:Z:19C109C0A17
```

(7) An = sign shows that the mate reference matches the reference name

(8) The 1-based left position is 524

(9) The insert size is 195 bases

(10) The sequence begins AATCT and ends ACGGG (its length is 150 bases)

(11) Each base is assigned a quality score (from BBBBB ending FHC.-)

(12) This read has additional, optional fields at accompany the MiSeq analysis

B&FG 3e
Fig.9-13
Page 403

## Anatomy of a Sequence Alignment/Map (SAM) file

(1) The query name of the read is given (M01121...)

In this example we'll look at a file called `030c_s7.bam`. It is a BAM file (the binary of a SAM). Most software manipulates BAM files rather than SAM.

The $ symbol indicates a command prompt in Unix

```
home/bioinformatics$ samtools view 030c_S7.bam | less
M01121:5:000000000-[illegible]:1:2111:[illegible]:15571   163   chrM
480   60   148M2S   524   195   AATCTCATCAAT
ACAACCCTCGCCCATCCTACCCAGCACACACACACCGCTGCTAACCCCATACCCCGAACC
AACCAAACCCCAAAGACACCCCCCACAGTTTATGTAGCTTACCTCCTCAAAGCAATAACC
TGAAAATCCTTAGACGGG   BBBBBFFB50FF[illegible]
```

Type `samtools` to run that program, and it includes a series of tools (such as `view`) to accomplish particular tasks—here, to view the contents of a file

The | symbol (called "pipe") indicates to send the results to another program—in this case to the utility called `less` that displays one page at a time on your terminal.

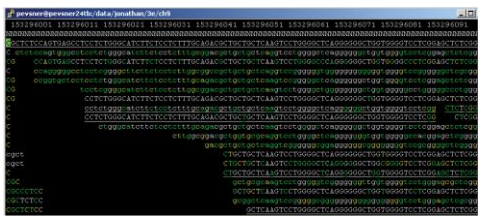
(10) The sequence begins AATCT and ends ACGGG (its length is 150 bases)

(11) Each base has a quality score (from BBBBB ending FHC.~)

(12) Many other fields are additional, optional fields that accompany the MiSeq analysis

B&FG 3e
Fig.9-13
Page 403

## SAMTools tview visualization of reads from a BAM file



There are many tools to view SAM/BAM files. A popular software package (SAMTools, used in Linux) includes `tview` visualization of reads from a BAM file

B&FG 3e
Fig.9-14
Page 404