Nucleotide sequence analysis

Bioinformatics, 2^{nd} year

UP FAMNIT

1

Plan for lectures

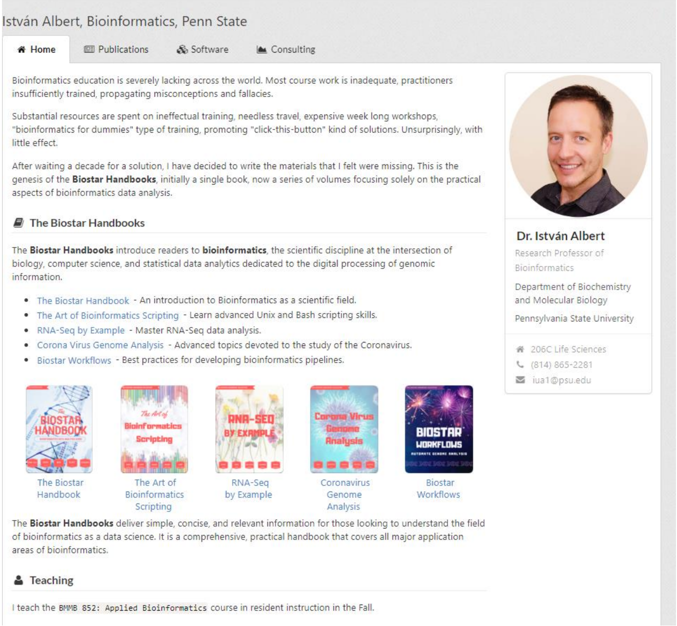
- RNA-Seq (transcriptome analysis)
- Genome assembly
- SNP calling / variant annotation
- Metabarcoding
- Genome wide association analysis

2

Literature

- Pevsner, J. 2015. Bioinformatics and functional genomics. Wiley Blackwell. (Chapter 8, 9, 10, 11 + Part III Genome Analysis)
- 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://training.galaxyproject.org/

3



István Albert, Bioinformatics, Penn State

Bioinformatics education is severely lacking across the world. Most course work is inadequate, practitioners insufficiently trained, propagating misconceptions and fallacies.

Substantial resources are spent on ineffectual training, needless travel, expensive week long workshops, "bioinformatics for dummies" type of training, promoting "click-this-button" kind of solutions. Unsurprisingly, with little effect.

After waiting a decade for a solution, I have decided to write the materials that I felt were missing. This is the genesis of the **Biostar Handbooks**, initially a single book, now a series of volumes focusing solely on the practical aspects of bioinformatics data analysis.

The Biostar Handbooks

The **Biostar Handbooks** introduce readers to **bioinformatics**, the scientific discipline at the intersection of biology, computer science, and statistical data analytics dedicated to the digital processing of genomic information.

- The Biostar Handbook - An introduction to Bioinformatics as a scientific field.
- The Art of Bioinformatics Scripting - Learn advanced Unix and Bash scripting skills.
- RNA-Seq by Example - Master RNA-Seq data analysis.
- Corona Virus Genome Analysis - Advanced topics devoted to the study of the Coronavirus.
- Biostar Workflows - Best practices for developing bioinformatics pipelines.

The Biostar Handbook | The Art of Bioinformatics Scripting | RNA-Seq by Example | Coronavirus Genome Analysis | Biostar Workflows

The **Biostar Handbooks** deliver simple, concise, and relevant information for those looking to understand the field of bioinformatics as a data science. It is a comprehensive, practical handbook that covers all major application areas of bioinformatics.

Teaching

I teach the `BMMB 852: Applied Bioinformatics` course in resident instruction in the Fall.

Dr. István Albert
Research Professor of Bioinformatics
Department of Biochemistry and Molecular Biology
Pennsylvania State University

206C Life Sciences
(814) 865-2281
iua1@psu.edu

https://www.biostarhandbook.com/

4

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.

5

Transcriptome analysis (RNA-Seq)

6

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

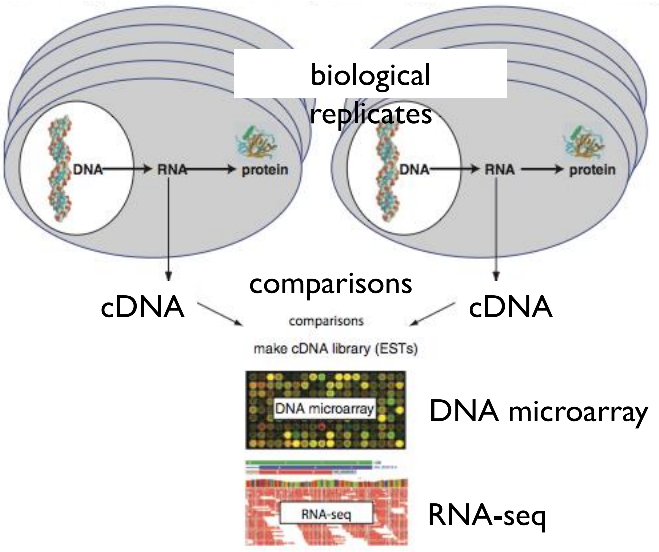
Source: Pevsner. Bioinformatics and Functional Genomics, 3rd Edition

7

Gene expression measured with high-throughput technologies

Condition A

Condition B



biological replicates

cDNA

comparisons

cDNA

DNA microarray

RNA-seq

B&FG 3e
Fig. 10.11
Page 454

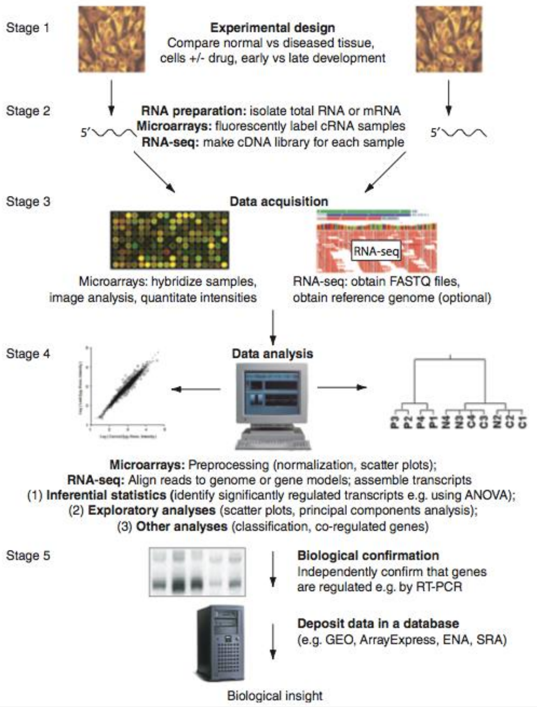
8

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.

9

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



Stage 1

Experimental design
Compare normal vs diseased tissue, cells +/- drug, early vs late development

Stage 2

RNA preparation: isolate total RNA or mRNA
Microarrays: fluorescently label cRNA samples
RNA-seq: make cDNA library for each sample

Stage 3

Data acquisition

Microarrays: hybridize samples, image analysis, quantitate intensities

RNA-seq: obtain FASTQ files, obtain reference genome (optional)

Stage 4

Data analysis

Microarrays: Preprocessing (normalization, scatter plots);
RNA-seq: Align reads to genome or gene models; assemble transcripts
(1) **Inferential statistics (**identify significantly regulated transcripts e.g. using ANOVA);
(2) **Exploratory analyses** (scatter plots, principal components analysis);
(3) **Other analyses** (classification, co-regulated genes)

Stage 5

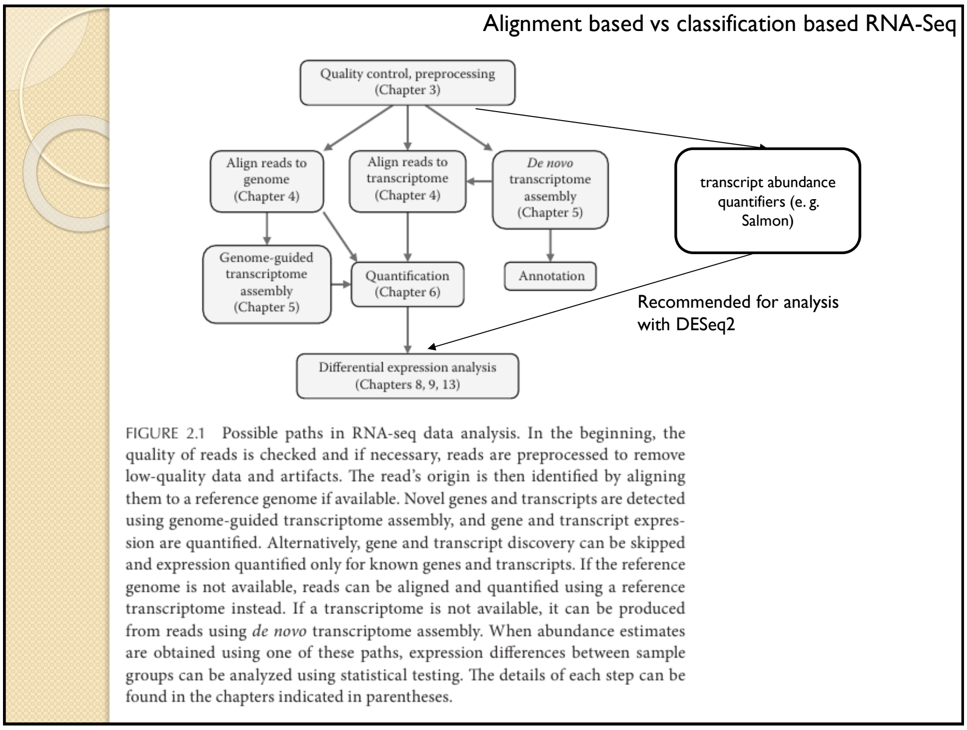
Biological confirmation
Independently confirm that genes are regulated e.g. by RT-PCR

Deposit data in a database
(e.g. GEO, ArrayExpress, ENA, SRA)

Biological insight

B&FG 3e
Fig. 10.13
Page 462

10



Alignment based vs classification based RNA-Seq

Quality control, preprocessing (Chapter 3)

Align reads to genome (Chapter 4)

Align reads to transcriptome (Chapter 4)

De novo transcriptome assembly (Chapter 5)

transcript abundance quantifiers (e. g. Salmon)

Genome-guided transcriptome assembly (Chapter 5)

Quantification (Chapter 6)

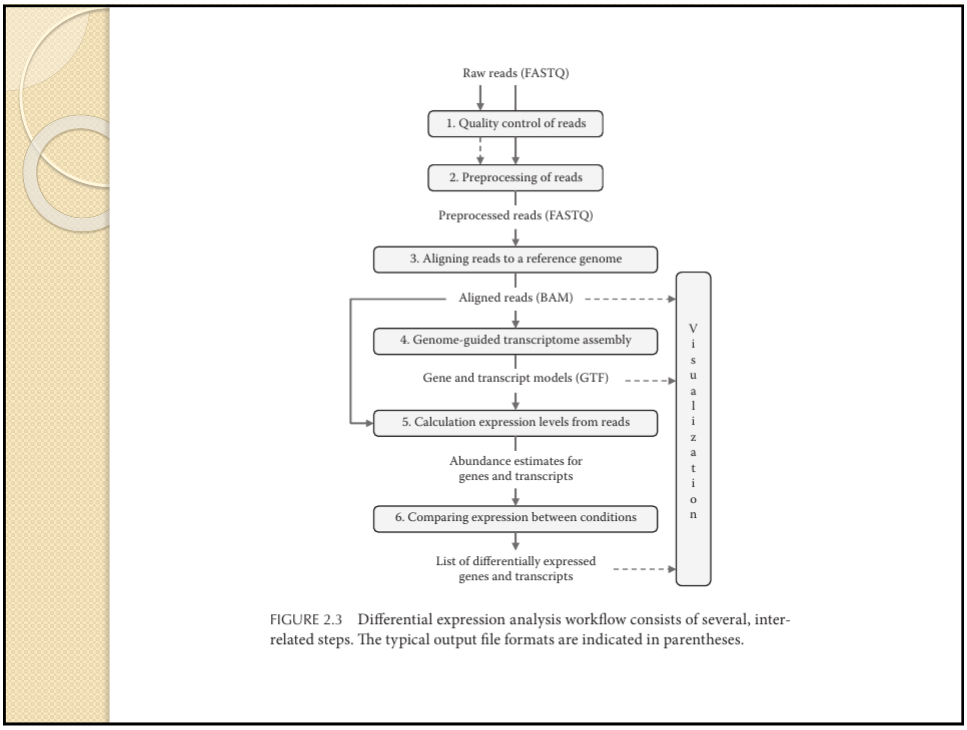
Annotation

Recommended for analysis with DESeq2

Differential expression analysis (Chapters 8, 9, 13)

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.

11



Raw reads (FASTQ)

1. Quality control of reads

2. Preprocessing of reads

Preprocessed reads (FASTQ)

3. Aligning reads to a reference genome

Aligned reads (BAM)

4. Genome-guided transcriptome assembly

Gene and transcript models (GTF)

5. Calculation expression levels from reads

Abundance estimates for genes and transcripts

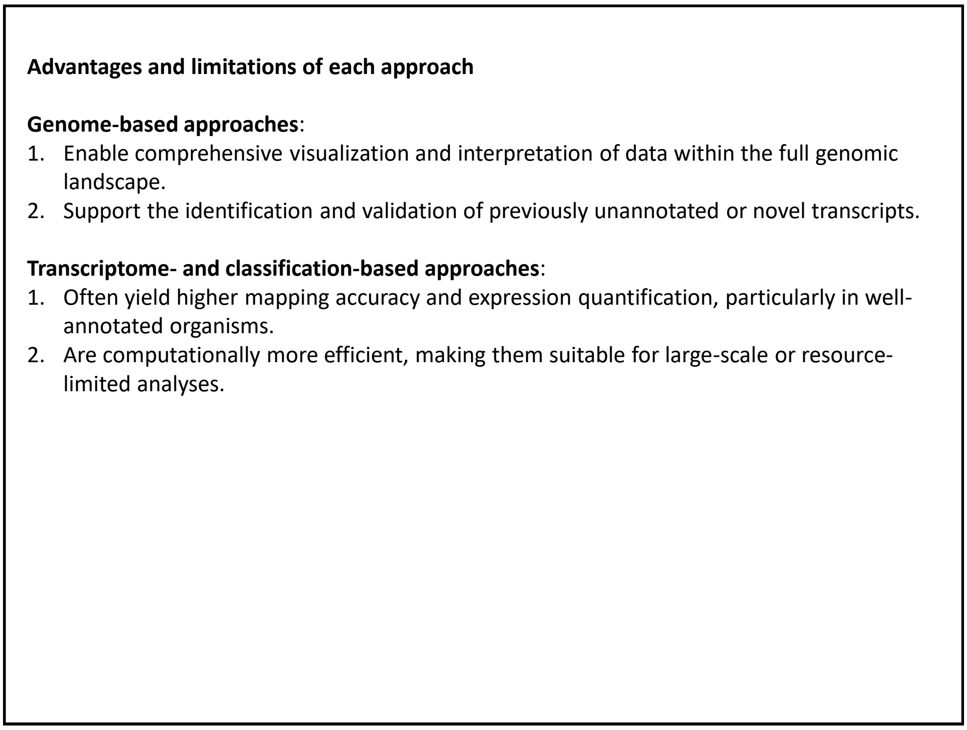
6. Comparing expression between conditions

List of differentially expressed genes and transcripts

Visualization

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

12



Advantages and limitations of each approach

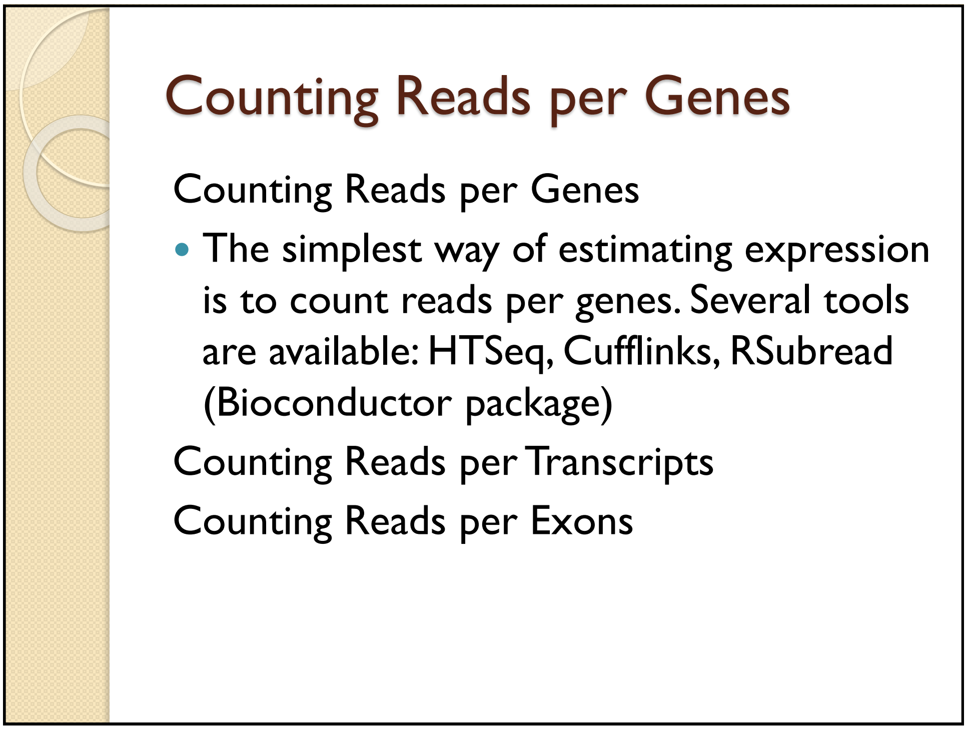
Genome-based approaches:

1. Enable comprehensive visualization and interpretation of data within the full genomic landscape.
2. Support the identification and validation of previously unannotated or novel transcripts.

Transcriptome- and classification-based approaches:

1. Often yield higher mapping accuracy and expression quantification, particularly in well-annotated organisms.
2. Are computationally more efficient, making them suitable for large-scale or resource-limited analyses.

13



Counting Reads per Genes

Counting Reads per Genes

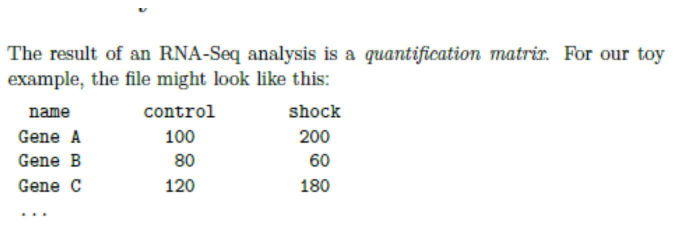
- The simplest way of estimating expression is to count reads per genes. Several tools are available: HTSeq, Cufflinks, RSubread (Bioconductor package)

Counting Reads per Transcripts

Counting Reads per Exons

14

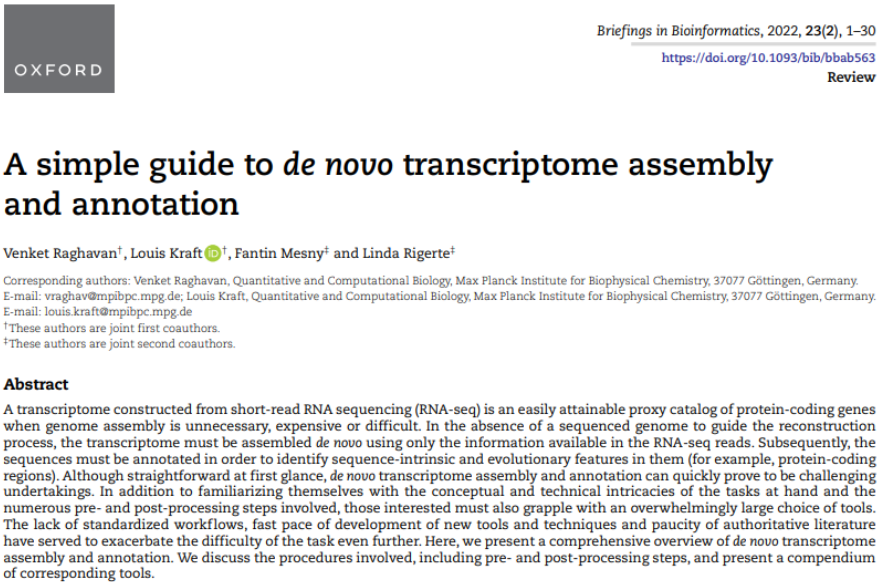
What is the final result of an RNA-Seq analysis?



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
...
```

15



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[†], Louis Kraft[†], Fantin Mesny[‡] and Linda Rigerte[‡]

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

[†]These authors are joint first coauthors.

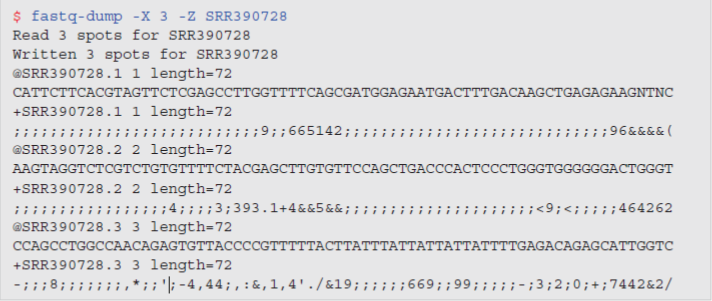
[‡]These authors are joint second coauthors.

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.

16

SRA toolkit:
fastq-dump to obtain FASTQ formatted data



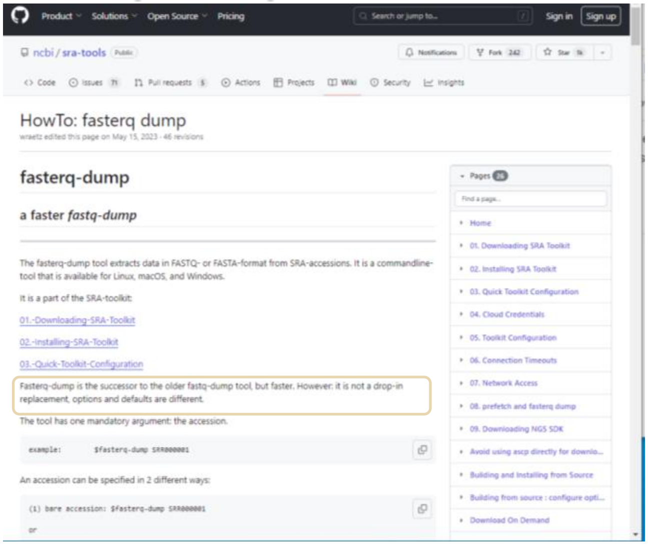
```
$ fastq-dump -X 3 -Z SRR390728
Read 3 spots for SRR390728
Written 3 spots for SRR390728
@SRR390728.1 1 length=72
CATTCTTCACGTAGTTCTCGAGCCTTGGTTTTCAGCGATGGAGAATGACTTTGACAAGCTGAGAGAAGNTNC
+SRR390728.1 1 length=72
;;;;;;;;;;;;;;;;;;;;;;;;;;;9;;665142;;;;;;;;;;;;;;;;;;;;;;;;;;;;;96&&&&(
@SRR390728.2 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;;;;;;;,*;;';-4,44;,:&,1,4'./&19;;;;;;669;;99;;;;;-;3;2;0;+;7442&2/
```

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

17

fasterq-dump



HowTo: fasterq dump

wraetz edited this page on May 15, 2023 · 46 revisions

fasterq-dump

a faster *fastq-dump*

The fasterq-dump tool extracts data in FASTQ- or FASTA-format from SRA-accessions. It is a commandline-tool that is available for Linux, macOS, and Windows.

It is a part of the SRA-toolkit:

01.-Downloading-SRA-Toolkit

02.-Installing-SRA-Toolkit

03.-Quick-Toolkit-Configuration

Fasterq-dump is the successor to the older fastq-dump tool, but faster. However: it is not a drop-in replacement, options and defaults are different.

The tool has one mandatory argument: the accession.

```
example:      $fasterq-dump SRR000001
```

An accession can be specified in 2 different ways:

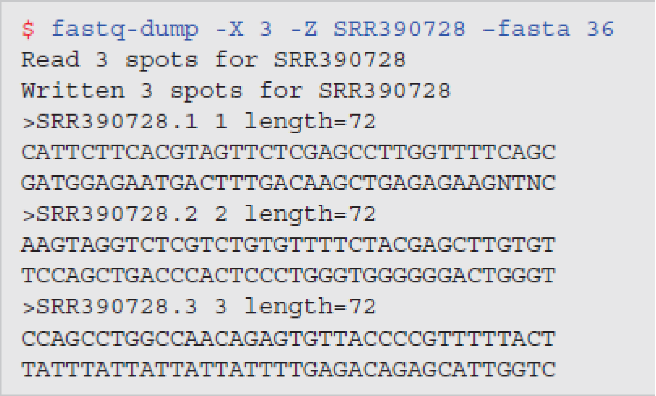
```
(1) bare accession: $fasterq-dump SRR000001

or
```

Pages 26
Find a page...
Home
01. Downloading SRA Toolkit
02. Installing SRA Toolkit
03. Quick Toolkit Configuration
04. Cloud Credentials
05. Toolkit Configuration
06. Connection Timeouts
07. Network Access
08. prefetch and fasterq dump
09. Downloading NGS SDK
Avoid using ascp directly for downlo...
Building and Installing from Source
Building from source : configure opti...
Download On Demand

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

18

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



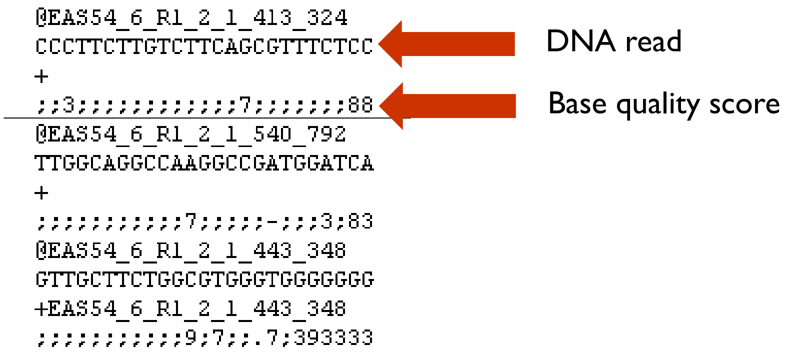
```
$ fastq-dump -X 3 -Z SRR390728 -fasta 36
Read 3 spots for SRR390728
Written 3 spots for SRR390728
>SRR390728.1 1 length=72
CATTCTTCACGTAGTTCTCGAGCCTTGGTTTTCAGC
GATGGAGAATGACTTTGACAAGCTGAGAGAAGNTNC
>SRR390728.2 2 length=72
AAGTAGGTCTCGTCTGTGTTTTCTACGAGCTTGTGT
TCCAGCTGACCCACTCCCTGGGTGGGGGGACTGGGT
>SRR390728.3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

B&FG 3e
Page 393

19

FASTQ format

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

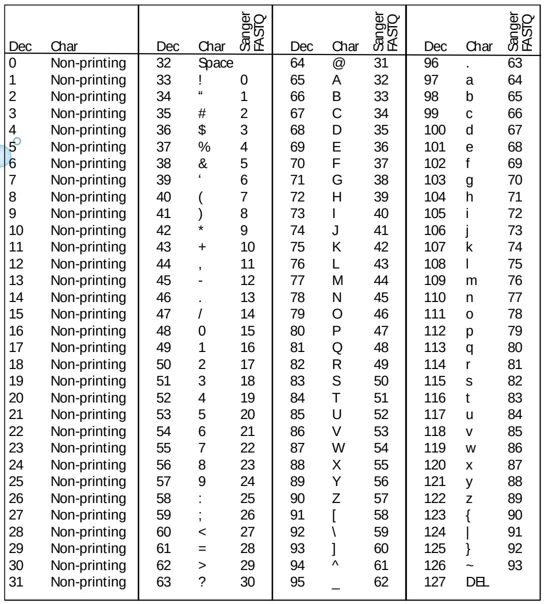


```
@EAS54_6_R1_2_1_413_324
CCCTTCTTGTCTTCAGCGTTTCTCC
+
;;3;;;;;;;;;;;;7;;;;;;;88
@EAS54_6_R1_2_1_540_792
TTGGCAGGCCAAGGCCGATGGATCA
+
;;;;;;;;;;;7;;;;;-;;;3;83
@EAS54_6_R1_2_1_443_348
GTTGCTTCTGGCGTGGGTGGGGGGG
+EAS54_6_R1_2_1_443_348
;;;;;;;;;;;9;7;;.7;393333
```

DNA read

Base quality score

20



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.

21

FASTQ format: Phred scores define quality

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

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

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%

22

Visualizing sequencing quality data

- FastQC tool

https://www.bioinformatics.babraham.ac.uk/projects/fastqc/

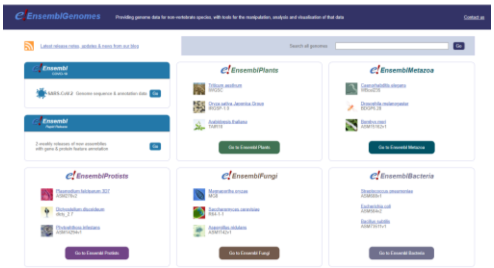
Quality control (QC) – improving data with removing identifiable errors

- Fastp
- Trimmomatic
- Cutadapt
- AdapterRemoval
- BBDuk
- …

Sequencing adapter trimming (same tools as previously mention)

Where genomes are available?

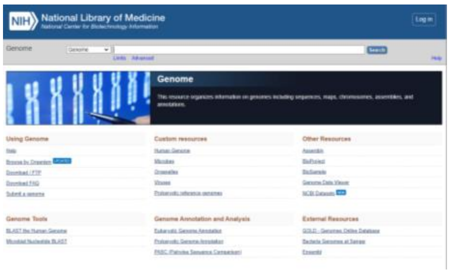
International Nucleotide Sequence Database Collaboration (INSDC) members [National Center for Biotechnology Information (NCBI), European Bioinformatics Institute (EBI), and DNA Data Bank of Japan (DDBJ)]



EnsemblGenomes Providing genome data for non-vertebrate species, with tools for the manipulation, analysis and visualisation of that data

SARS-CoV-2 Genome sequence & annotation data

2 weekly releases of new assemblies with gene & protein feature annotation



Genome

This resource organizes information on genomes including sequences, maps, chromosomes, assemblies, and annotations.

Where genomes are available?

Genomics Proteomics Bioinformatics 19 (2021) 584–589

Genomics Proteomics Bioinformatics

www.elsevier.com/locate/gpb
www.sciencedirect.com

ELSEVIER

DATABASE

Genome Warehouse: A Public Repository Housing Genome-scale Data

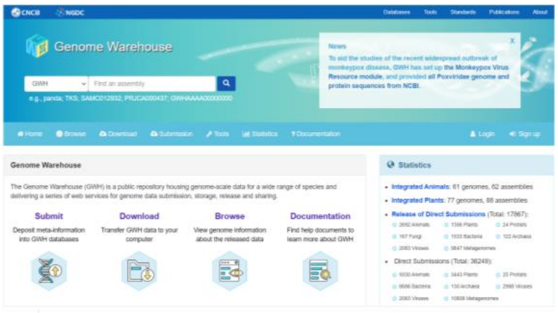
Meili Chen[1,2,#], Yingke Ma[1,2,#], Song Wu[1,2,3], Xinchang Zheng[1,2], Hongen Kang[1,2,3], Jian Sang[1,2,3,4], Xingjian Xu[1,2,3,*], Lili Hao[1,2], Zhaohua Li[1,2,3], Zheng Gong[1,2,3], Jingfa Xiao[1,2,3], Zhang Zhang[1,2,3], Wenming Zhao[1,2,3], Yiming Bao[1,2,3,*]

[1] National Genomics Data Center, Beijing Institute of Genomics, Chinese Academy of Sciences / China National Center for Bioinformation, Beijing 100101, China
[2] CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China
[3] University of Chinese Academy of Sciences, Beijing 100049, China

Received 13 January 2021; revised 19 February 2021; accepted 26 April 2021
Available online 24 June 2021

Handled by Fangqing Zhao

Abstract The Genome Warehouse (GWH) is a public repository housing genome assembly data for a wide range of species and delivering a series of web services for genome data submission, storage, release, and sharing. As one of the core resources in the National Genomics Data Center (NGDC), part of the China National Center for Bioinformation (CNCB, https://ngdc.cncb.ac.cn), GWH accepts both full and partial (chloroplast, mitochondrion, and plasmid) genome sequences with different assembly levels, as well as an update of existing genome assemblies. For each assembly, GWH collects detailed genome-related metadata of biological project, biological sample, and genome assembly, in addition to genome sequence and annotation. To archive high-quality genome sequences and annotations, GWH is equipped with a uniform and standardized procedure for quality control. Besides basic browse and search functionalities, all released genome sequences and annotations can be visualized with JBrowse. By May 21, 2021, GWH has received 19,124 direct submissions covering a diversity of 1108 species and has released 8772 of them. Collectively, GWH serves as an important resource for genome-scale data management and provides free and publicly accessible data to support research activities throughout the world. GWH is publicly accessible at https://ngdc.cncb.ac.cn/gwh.



e.g., panda; TKS; SAMC012932; PRJCA000437; GWHAAAA00000000

News

To aid the studies of the recent widespread outbreak of monkeypox disease, GWH has set up the Monkeypox Virus Resource module, and provided all Poxviridae genome and protein sequences from NCBI.

Genome Warehouse

The Genome Warehouse (GWH) is a public repository housing genome-scale data for a wide range of species and delivering a series of web services for genome data submission, storage, release and sharing.

Submit
Deposit meta-information into GWH databases

Download
Transfer GWH data to your computer

Browse
View genome information about the released data

Documentation
Find help documents to learn more about GWH

Statistics

- Integrated Animals: 61 genomes, 62 assemblies
- Integrated Plants: 77 genomes, 88 assemblies
- Release of Direct Submissions (Total: 17867):
  - 2692 Animals
  - 1166 Plants
  - 24 Protists
  - 167 Fungi
  - 1533 Bacteria
  - 122 Archaea
  - 2083 Viruses
  - 9847 Metagenomes
- Direct Submissions (Total: 36249):
  - 9330 Animals
  - 3443 Plants
  - 25 Protists
  - 9686 Bacteria
  - 130 Archaea
  - 2988 Viruses
  - 2083 Viruses
  - 10808 Metagenomes

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

Stranded vs non-stranded RNA-Seq

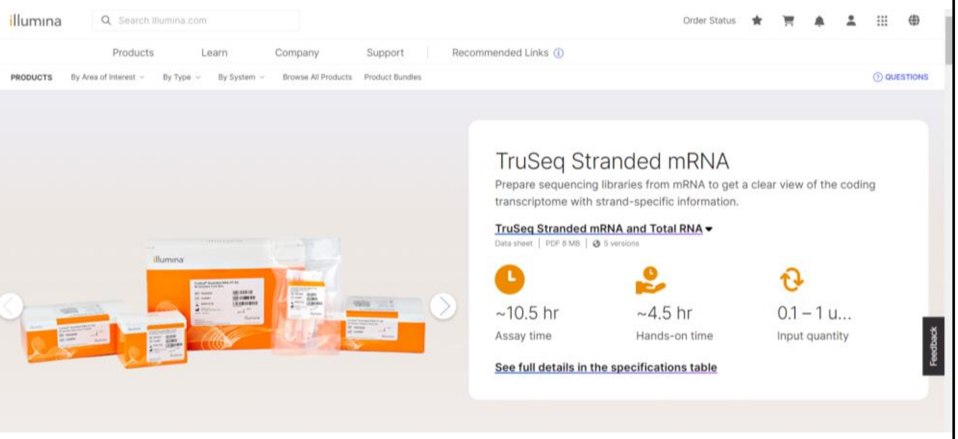
- Stranded (strand-specific) RNA-Seq

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

27



TruSeq Stranded mRNA

Prepare sequencing libraries from mRNA to get a clear view of the coding transcriptome with strand-specific information.

TruSeq Stranded mRNA and Total RNA

Data sheet | PDF 8 MB | 5 versions

~10.5 hr Assay time

~4.5 hr Hands-on time

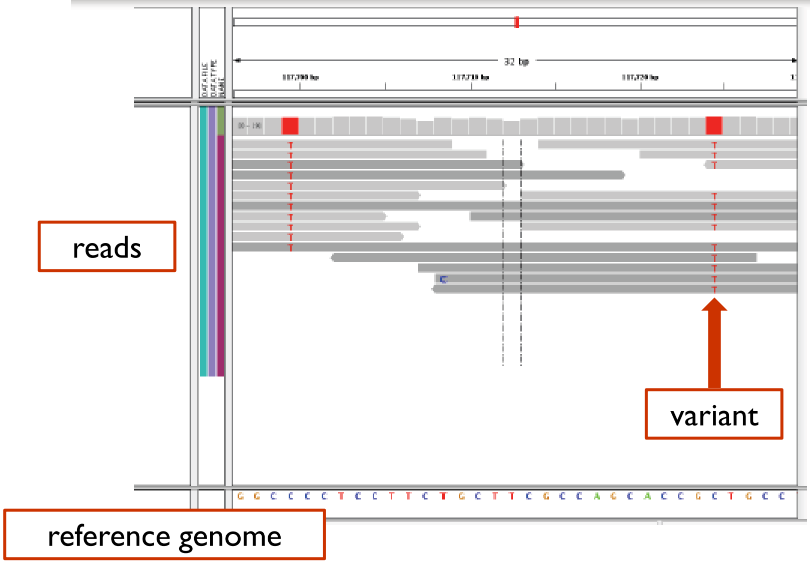
0.1 – 1 u... Input quantity

See full details in the specifications table

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

28

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

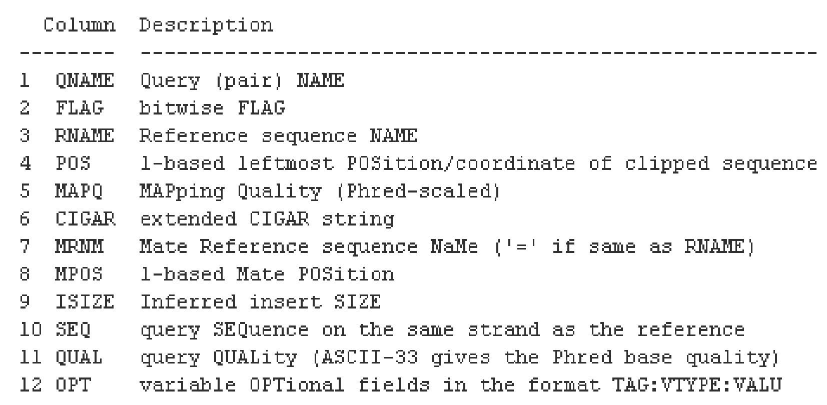

32 bp
117,700 bp
117,710 bp
117,720 bp
0 - 100
G G C C C C T C C T T C T G C T T C G C C A G C A C C G C T G C C

reads

variant

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
https://samformat.pages.dev/sam-format-flag

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/

Mate Pair Library v2 Sample Preparation Guide
For 2–5 kb Libraries

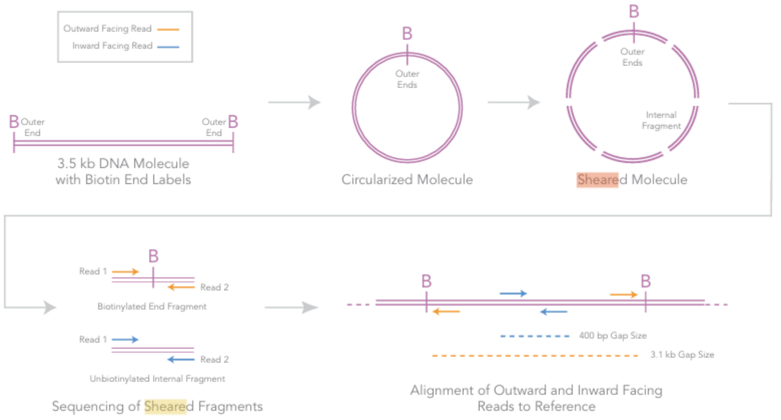
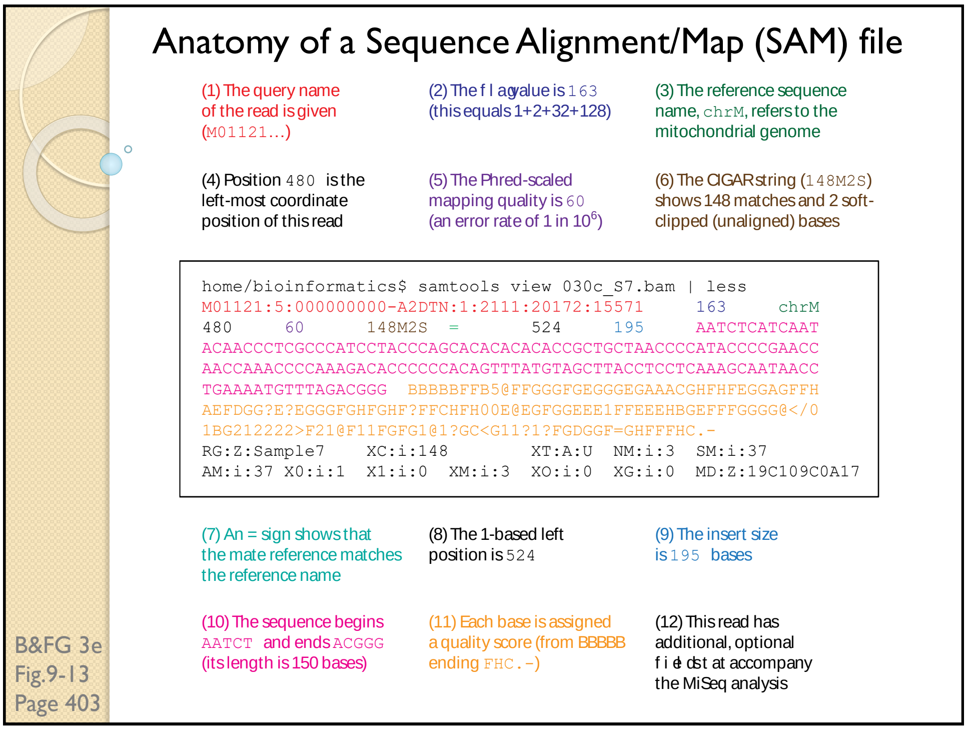


Figure 9 *Origin and Alignment of Inward and Outward-Facing Reads*

chrome-extension://efaidnbmnnnibpcajpcglclefindmkaj/https://support.illumina.com.cn/content/dam/illumina-support/documents/documentation/chemistry_documentation/samplepreps_legacy/MatePair_v2_2-5kb_SamplePrep_Guide_15008135_A.pdf



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?FFCHFH0OE@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

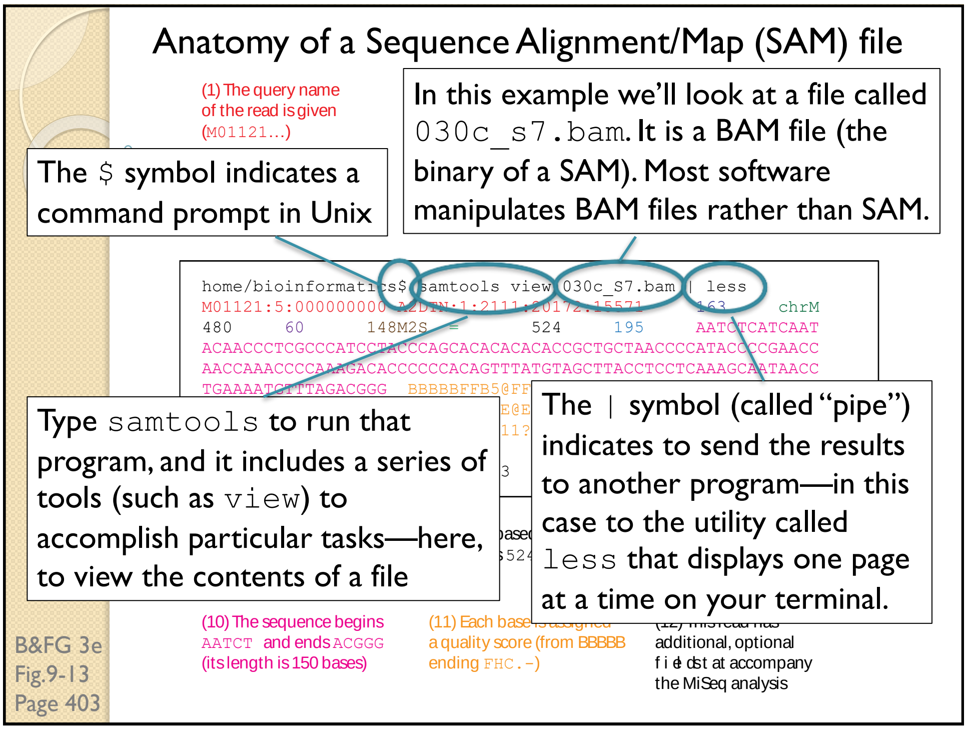
(10) The sequence begins AATCT and ends ACGGG (its length is 150 bases)

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

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

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

33



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-A2DTN:1:2111:20172:15571      163      chrM
480      60       148M2S   =        524      195      AATCTCATCAAT
ACAACCCTCGCCCATCCTACCCAGCACACACACACCGCTGCTAACCCCATACCCCGAACC
AACCAAACCCCAAAGACACCCCCCACAGTTTATGTAGCTTACCTCCTCAAAGCAATAACC
TGAAAATGTTTAGACGGG  BBBBBFFB5@FF
```

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)

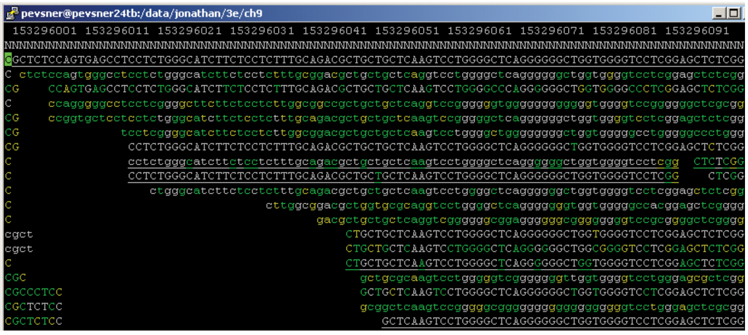
(11) Each base is assigned a quality score (from BBBBB ending FHC.-)

additional, optional fields that accompany the MiSeq analysis

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

34

SAMTools tview visualization of reads from a BAM file



pevsner@pevsner24tb:/data/jonathan/3e/ch9

```
 153296001 153296011 153296021 153296031 153296041 153296051 153296061 153296071 153296081 153296091
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN
CGCTCTCCAGTGAGCCTCCTCTGGGCATCTTCTCCTCTTTGCAGACGCTGCTGCTCAAGTCCTGGGGCTCAGGGGGGCTGGTGGGGTCCTCGGAGCTCTCGG
C ctctccagtgggcctcctctgggcatcttctcctctttgcggacgctgctgctcaggtcctggggctcagggggggctggtggggtcctcggagctctcgg
CG    CCAGTGAGCCTCCTCTGGGCATCTTCTCCTCTTTGCAGACGCTGCTGCTCAAGTCCTGGGGCCCAGGGGGGCTGGTGGGGCCCTCGGAGCTCTCGG
C     ccaggggggcctcctcggggcttcttctcctcttggcggccgctgctgctcaggtccgggggtgggggggggggtggggtccggggggctcgcgg
CG    ccggtgctcctcctctgggcatcttctcctctttgcagacgctgctgctcaagtccggggctcaggggggctggtggggtcctcggagctctcgg
CG              tcctcggggcatcttctcctcttggcggacgctgctgctcaagtcctggggctgggggggctggtgggggcctgggggccctggg
CG               CCTCTGGGCATCTTCTCCTCTTTGCAGACGCTGCTGCTCAAGTCCTGGGGCTCAGGGGGGCTGGTGGGGTCCTCGGAGCTCTCGG
C                cctctgggcatcttctcctctttgcagacgctgctgctcaagtcctggggctcaggggggctggtggggtcctcgg  CTCTCGG
C                CCTCTGGGCATCTTCTCCTCTTTGCAGACGCTGCTGCTCAAGTCCTGGGGCTCAGGGGGGCTGGTGGGGTCCTCGG    CTCGG
C                   ctgggcatcttctcctctttgcagacgctgctgctcaagtcctggggctcaggggggctggtggggtcctcggagctctcgg
C                                   cttggcggacgctggtgcgcaggtcctggggctcaggggggggtggtgggggccacggagctcgggg
C                                          gacgctgctgctcaggtcgggggcggagggggggcgggggggtccgcggggctcgggg
cgct                                           CTGCTGCTCAAGTCCTGGGGCTCAGGGGGGCTGGTGGGGTCCTCGGAGCTCTCGG
cgct                                           CTGCTGCTCAAGTCCTGGGGCTCAGGGGGGCTGGCGGGGTCCTCGGAGCTCTCGG
C                                              CTGCTGCTCAAGTCCTGGGGCTCAGGGGGGCTGGTGGGGTCCTCGGAGCTCTCGG
CGC                                              gctgcgcaagtcctggggtcgggggggttggtggggtcctgggagcgctcgg
CGCCCTCC                                         GCTGCTCAAGTCCTGGGGCTCAGGGGGGCTGGTGGGGTCCTCGGAGCTCTCGG
CGCTCTCC                                         gcggctcaagtccggggcggggggggggggggggggtcctgggagctcgcgg
CGCTCTCC                                            GCTCAAGTCCTGGGGCTCAGGGGGGCTGGTGGGGTCCTCGGAGCTCTCGG
```

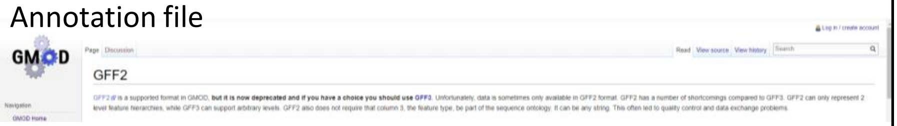
There are many tools to view SAM/BAM files. A popular software package (SAMTools, used in Linux) includes `tview` visualization of reads from a BAM file

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

Indexing reference genome/transcriptome

- Indexing reference genome
  - Incorporating the annotation into the index

Annotation file



GFF2

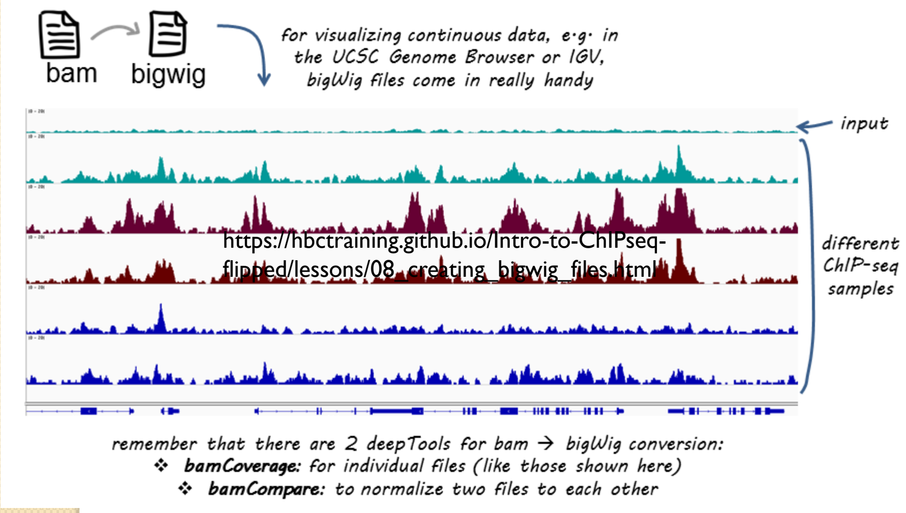
GFF2 is a supported format in GMOD, **but it is now deprecated and if you have a choice you should use GFF3.** Unfortunately, data is sometimes only available in GFF2 format. GFF2 has a number of shortcomings compared to GFF3. GFF2 can only represent 2 level feature hierarchies, while GFF3 can support arbitrary levels. GFF2 also does not require that column 3, the feature type, be part of the sequence ontology. It can be any string. This often led to quality control and data exchange problems.

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

Welcome To SO

Welcome to the Sequence Ontology

This is the home page of the Sequence Ontology (SO). SO is a collaborative ontology project for the definition of sequence features used in biological sequence annotation. SO was initially developed by the **Gene Ontology Consortium**. Contributors to SO include the **GMOD** community, model organism database groups such as **WormBase, FlyBase, Mouse Genome Informatics** group, and institutes such as the Sanger Institute and the EBI. Input to SO is welcomed from the sequence annotation community. SO is also part of the Open Biomedical Ontologies library. Our aim is to develop an ontology suitable for describing the features of biological sequences. For questions, please send mail to the **SO developers mailing list**. For new term suggestions, please use the **Term Tracker**.



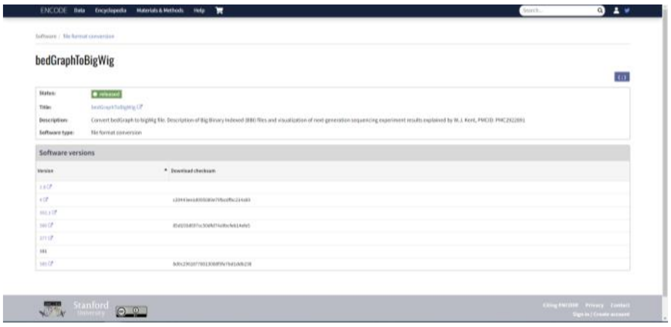
for visualizing continuous data, e.g. in the UCSC Genome Browser or IGV, bigWig files come in really handy

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

remember that there are 2 deepTools for bam → bigWig conversion:

- **bamCoverage:** for individual files (like those shown here)
- **bamCompare:** to normalize two files to each other

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



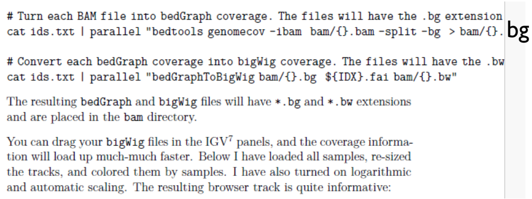
Software / File format conversion

bedGraphToBigWig

Status: released
Title: bedGraphToBigWig
Description: Convert bedGraph to bigWig file. Description of Big Binary Indexed (BBI) files and visualization of next generation sequencing experiment results explained by W.J. Kent, PMCID: PMC2922891
Software type: File format conversion

Software versions

Version	Download checksum
2.8	
4	[illegible]
302.1	
369	[illegible]
377	
381	
385	[illegible]



```
# 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/