Quality Control with FastQC

Clyde Dapat

WHO Collaborating Centre for Reference and Research on Influenza
VIDRL

Doherty Institute
THE UNIVERSITY OF MELBOURNE
The Royal Melbourne Hospital

A joint venture between The University of Melbourne and The Royal Melbourne Hospital

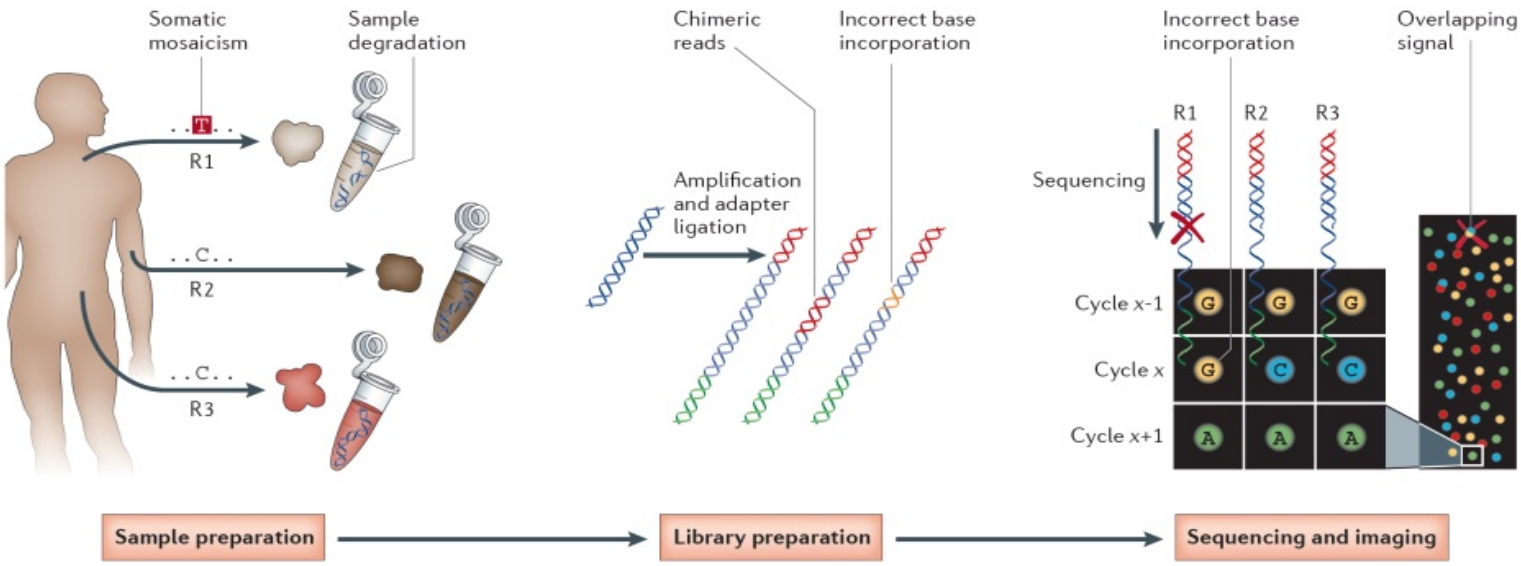
Objective

- To learn how to assess the quality of NGS data

Why is QC important before doing data analysis?

- Failure modes
  - Contamination
  - Library failures
- Artefacts
- Biases
- Mis-interpretations

Sources of sequencing errors



Sequencing errors can stem from any time point throughout the experimental workflow, including initial sequence preparation, library preparation and sequencing.

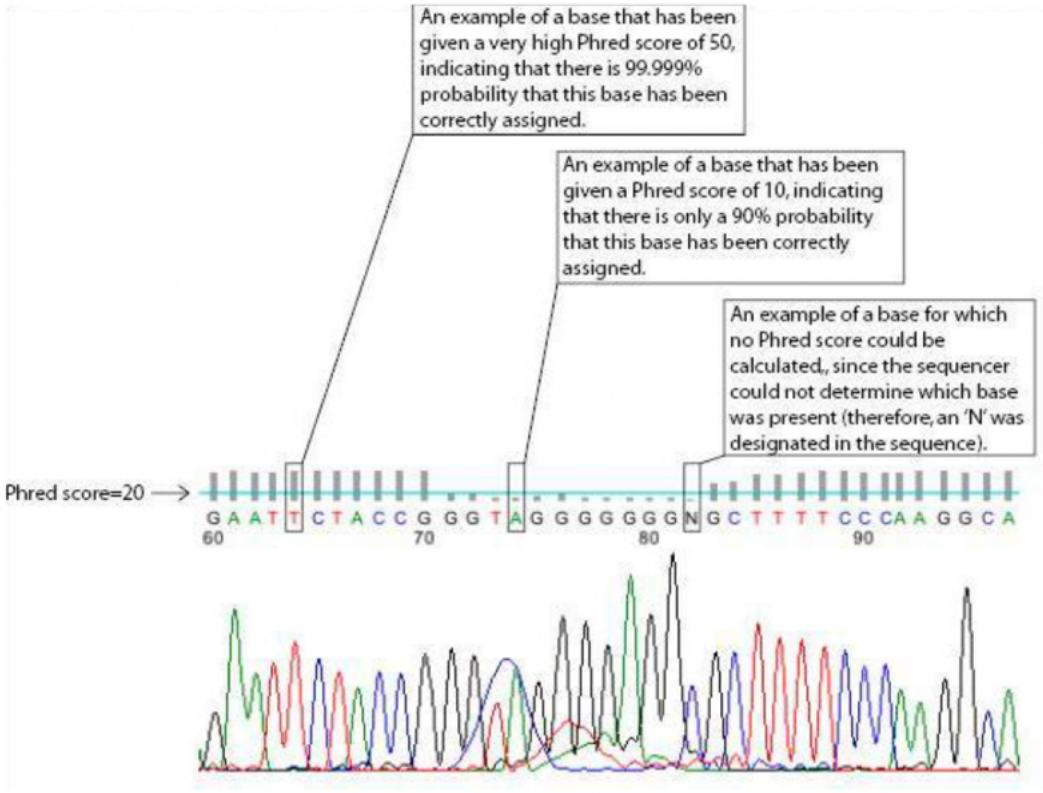
The role of replicates for error mitigation in next-generation sequencing. Robasky et al. Nature Reviews Genetics, 2014, 15, 56-62

Sequencing error rates

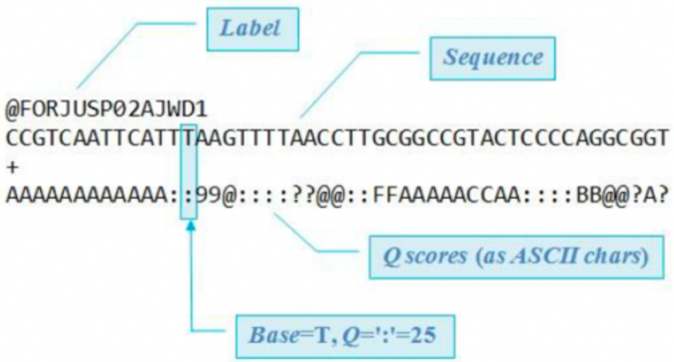
Platform	Frequent error types	Ratio
Capillary sequencing	Single nucleotide substitutions	10^{-1}
Ion Torrent	Deletions	10^{-2}
illumina	Single nucleotide substitutions	10^{-3}

https://www.researchgate.net/profile/Shifu_Chen/publication/323733012/figure/tbl1/AS:688075940302849@1541061496360/A-comparison-of-sequencing-error-ratios-of-different-sequencing-platforms.png

Chromatogram



FASTQ



http://drive5.com/usearch/manual/fastq_files.html

```
@A00180:10:H7NK5DMXX:1:1101:16821:1000 1:N:0:GGACTT+GTCGTTCG
TGCAGCAGCTAATGAGGAACCACTTCCTCCCTCCAGCCGCTCTAAATACCTCAGAACAATAGGATCATCATAATAATCCCCTAGTCTGAACTG
+
8FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@A00180:10:H7NK5DMXX:1:1101:19090:1016 1:N:0:GGACTT+GTCGTTCG
TGCAGAAGAAGAAAGCACAAGTATTTACGCCTATCCTTCATATTTTCCGCAAGGTAACTATCTCGGTTTCATATCGAGATTTATATAGAATCT
+
FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF-FFFFFFFFFF
@A00180:10:H7NK5DMXX:1:1101:19325:1016 1:N:0:GGACTT+GTCGTTCG
TGCAGGAAGTTATGCAGGGGCATCCTGTATTATTAAATAGAGCACCTACTCTTCATAGATTAGGTATACAGGCGTTCCAACCTATTTTAGTGG
+
FF-88F8FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF-F-FFFFFFFF
@A00180:10:H7NK5DMXX:1:1101:30897:1016 1:N:0:GGACTT+GTCGTTCG
TGCAGAGTACATCAACAAAAGAAACCTAACTGCCCTACCGGCAAACCGGTAGAGTACCCTTCCCCAAAAGTATTACTCCCAGTCAATATAAGG
+
8F888FFF-FFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF-FFFFFFFFFFFFFFF-FFFFFFFFFFFFFFF-FFFFFFFFFF
```

https://www.reneshbedre.com/blog/fqqualfmt.html

FASTQ

http://drive5.com/usearch/manual/fastq_files.html

ASCII TABLE

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

https://simple.m.wikipedia.org/wiki/File:ASCII-Table-wide.svg

Base call quality

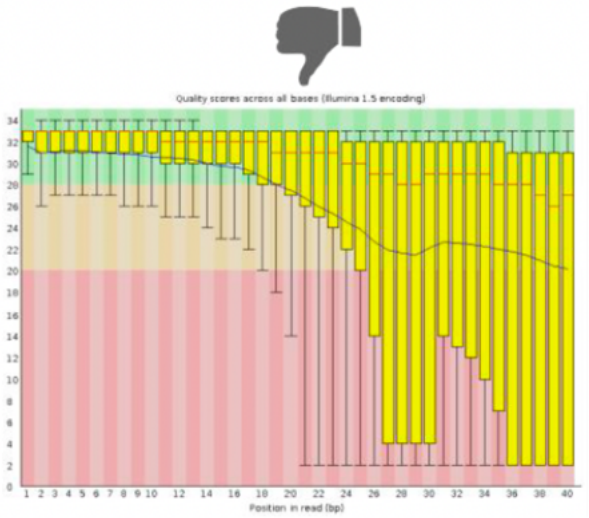
Phred Quality Score	Probability of incorrect base call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

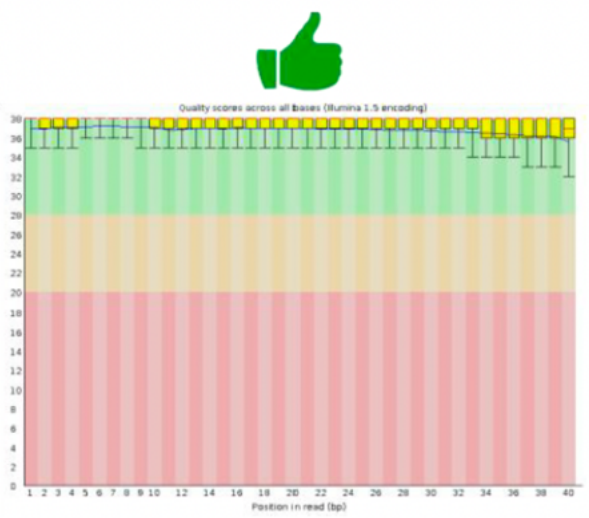
Sequence quality control

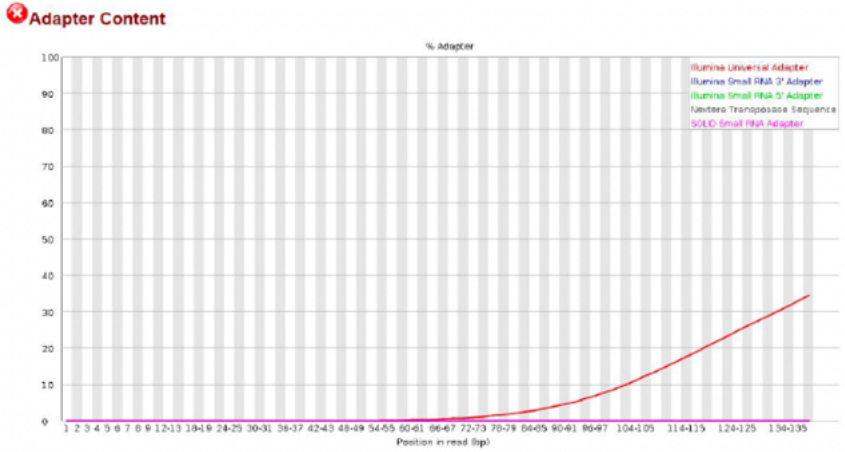
General Steps

1. Initial sequence quality assessment
2. Adapter clipping
3. Trimming of low quality sequence ends
4. Read quality filtering
5. Pairing of reads (for paired-end reads)
6. Final sequence quality assessment

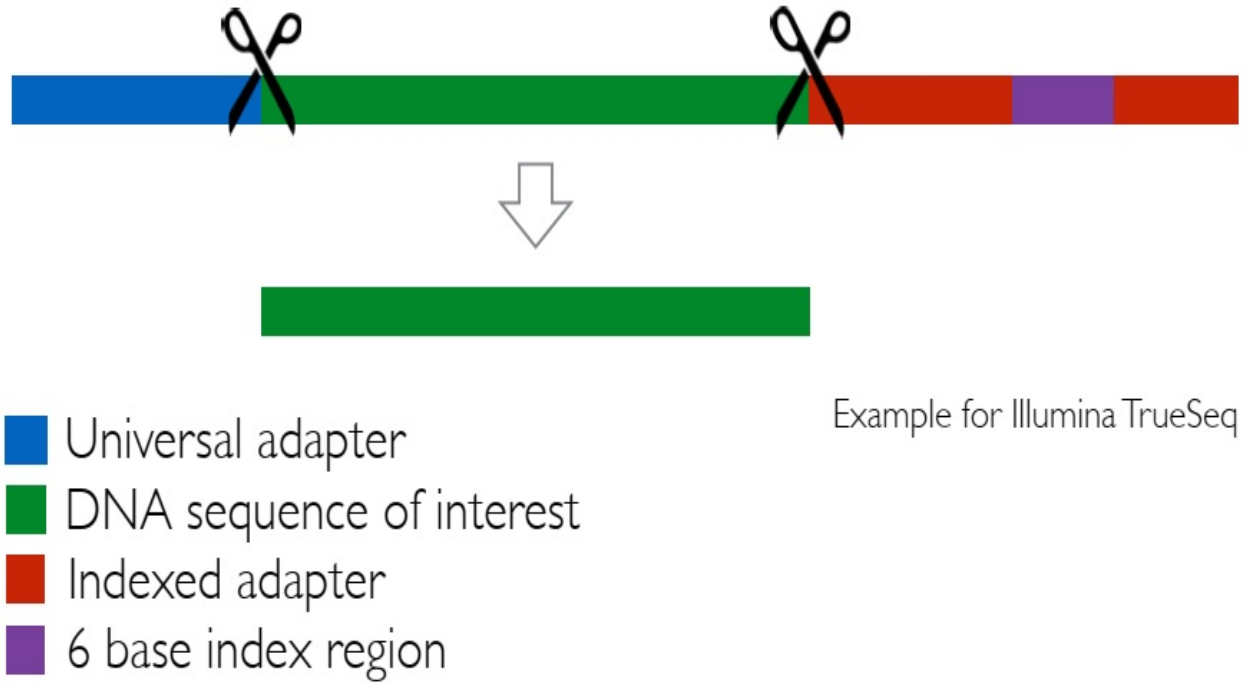
Initial quality assessment



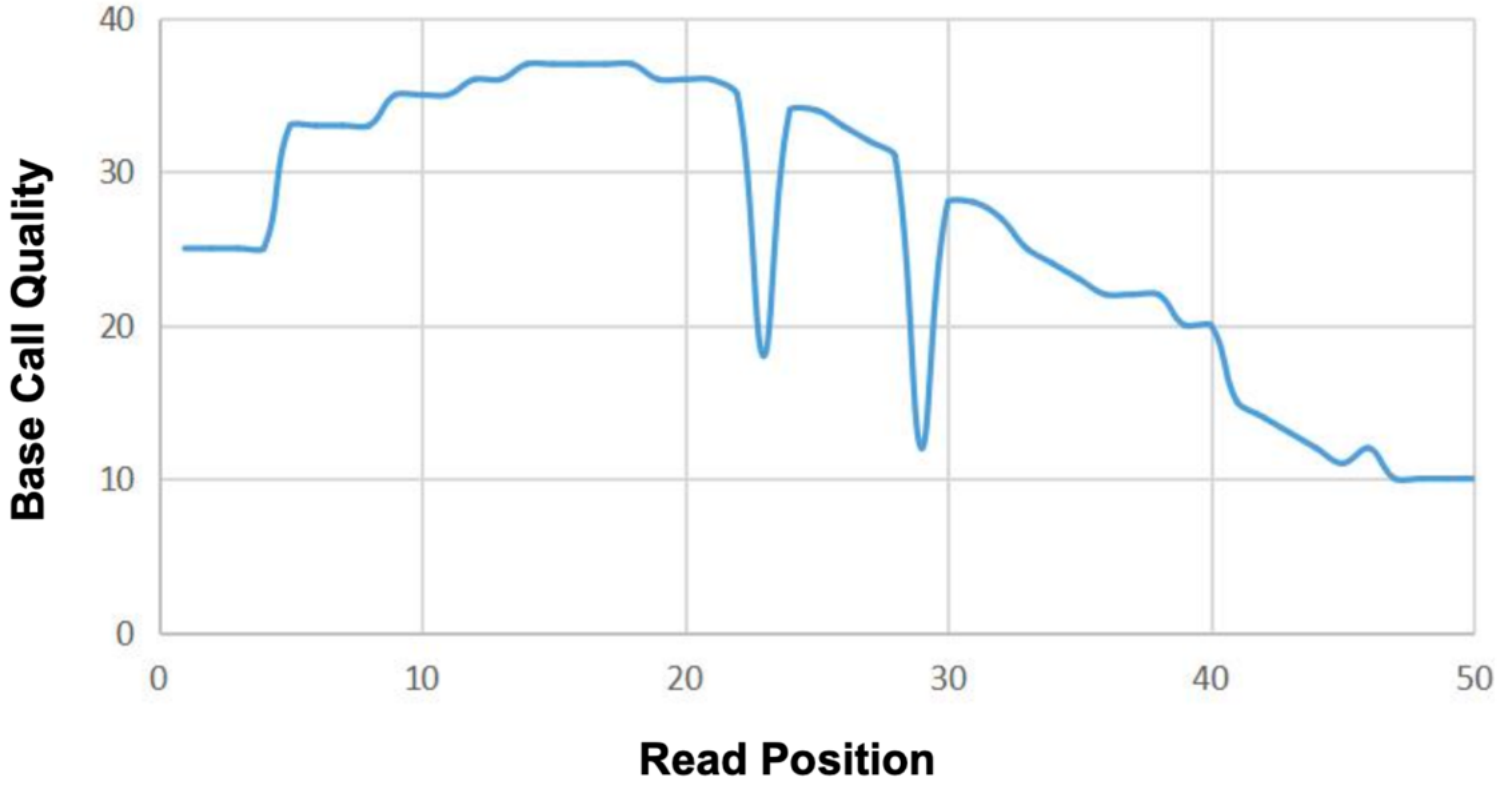




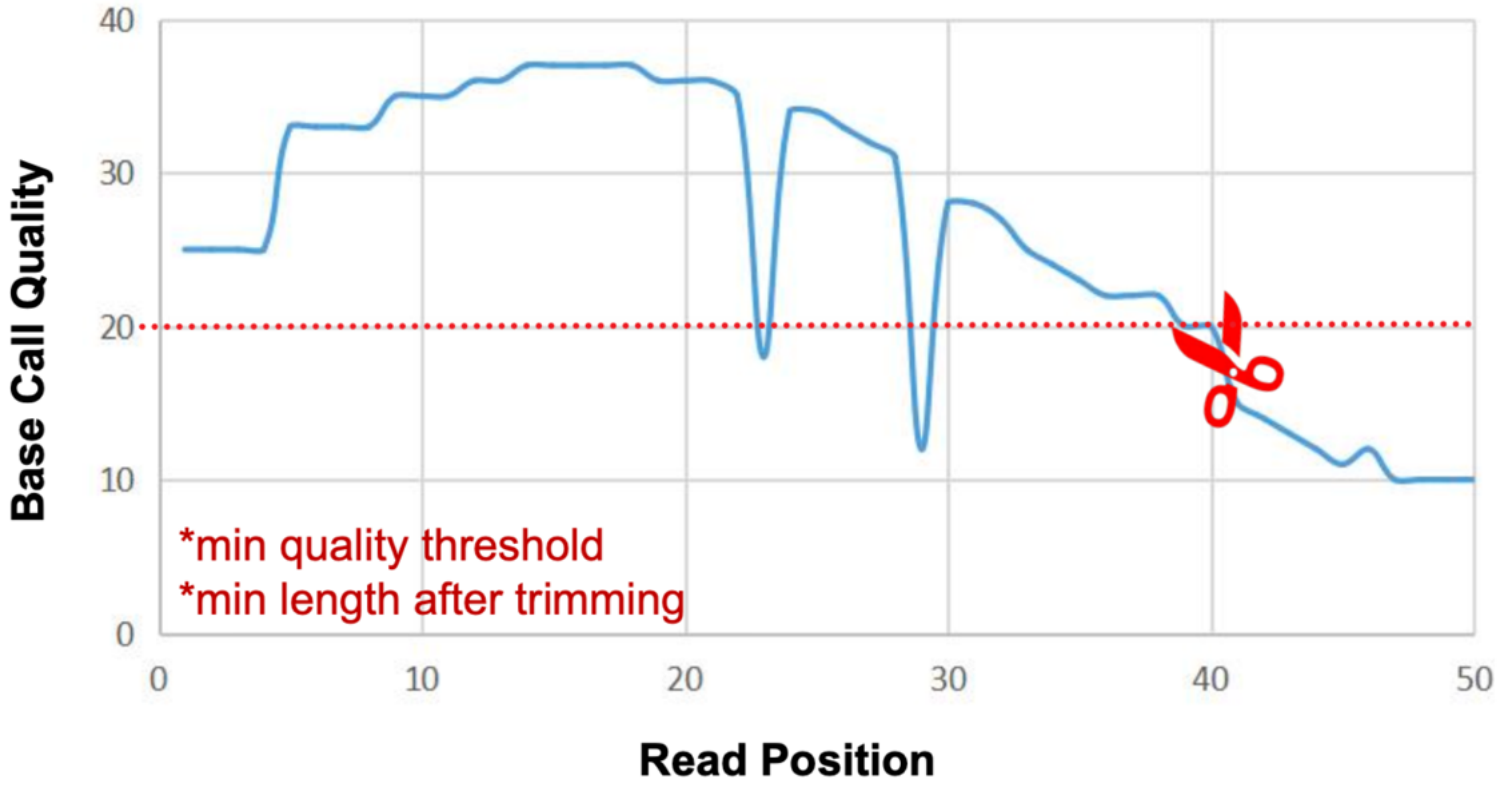
Adapter clipping



Trimming of low sequence ends

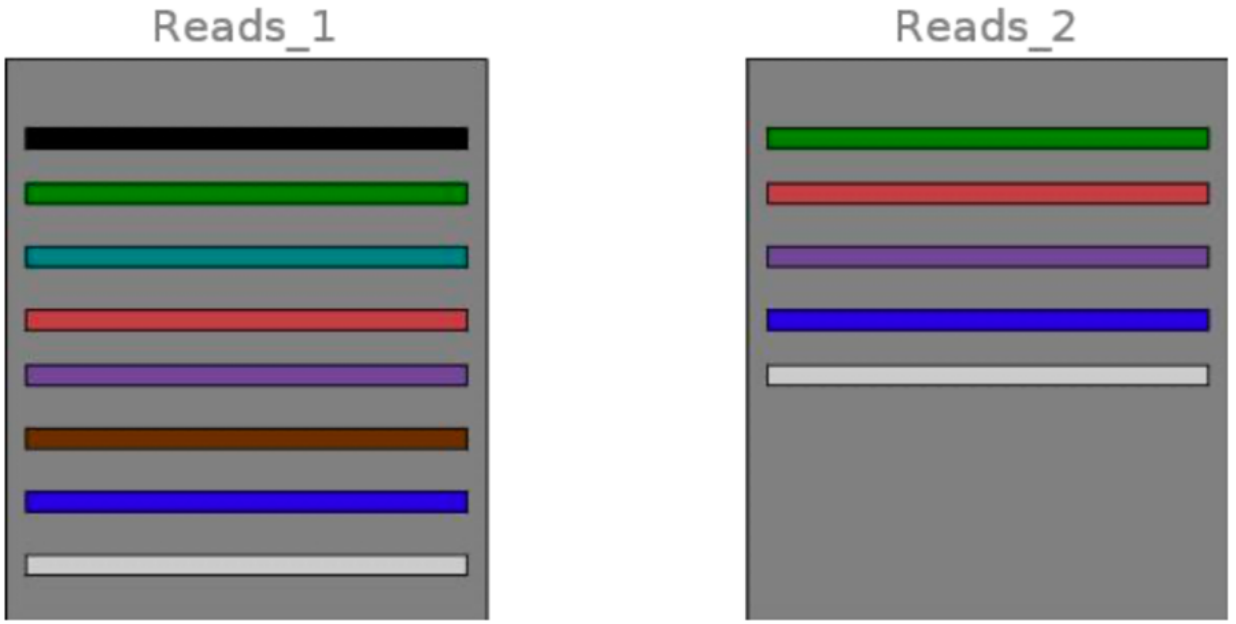


Trimming of low sequence ends

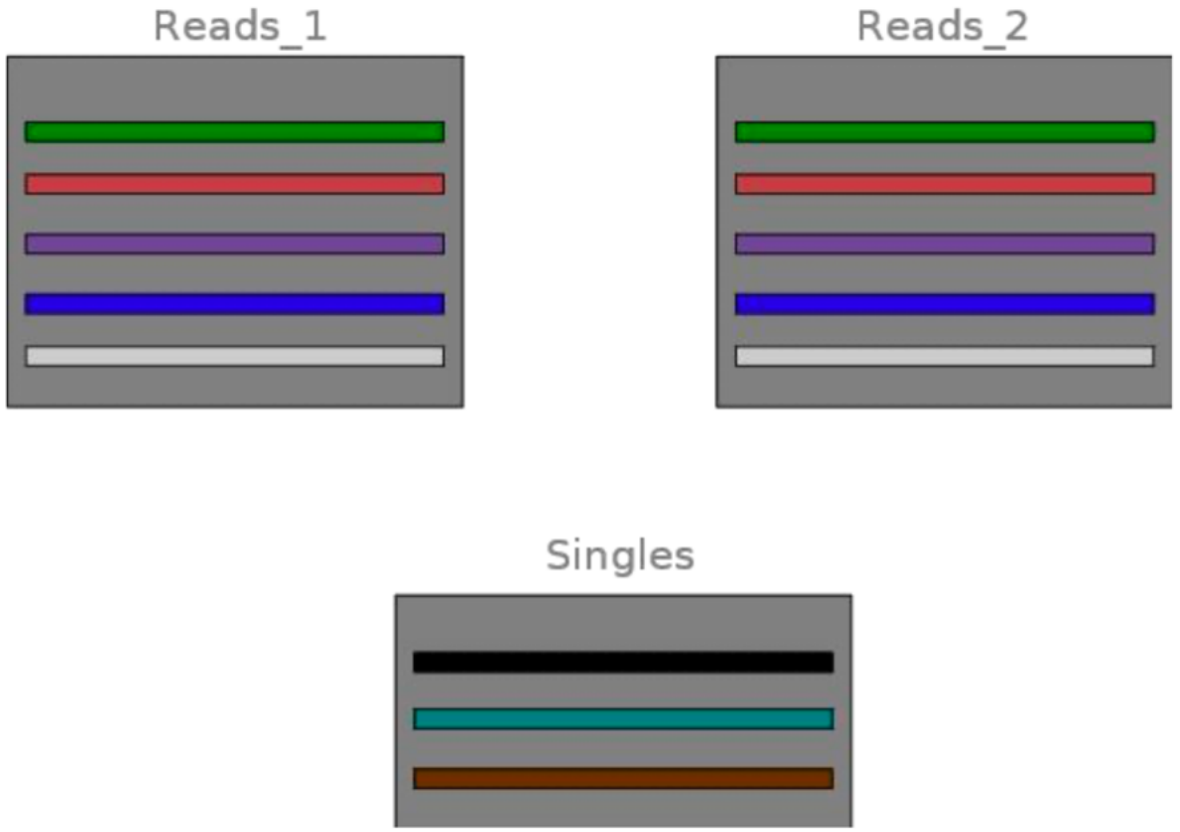


Read quality filtering

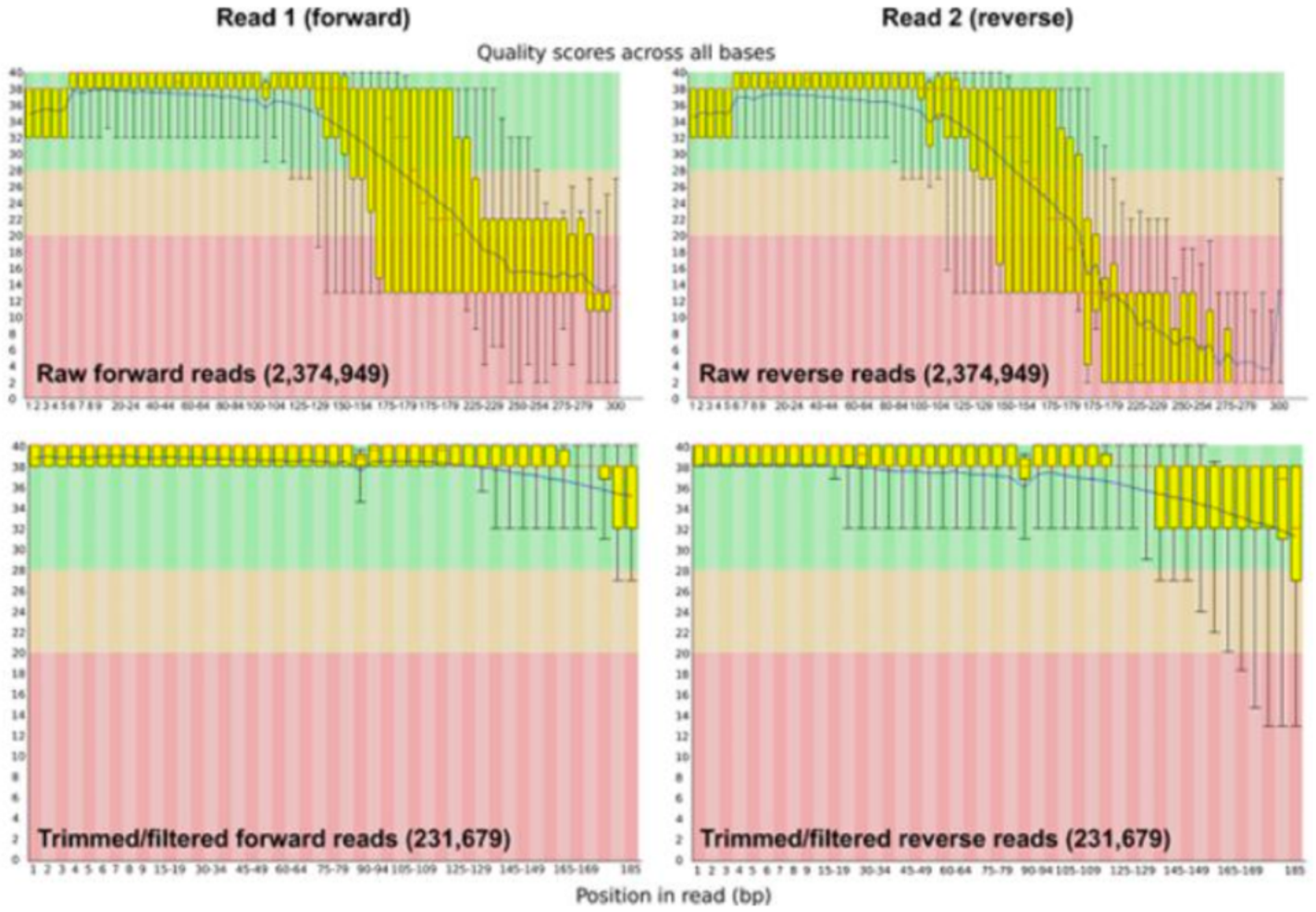
Pairing of reads (for paired-end data)



Pairing of reads (for paired-end data)



Final quality assessment



https://www.researchgate.net/profile/Richard-Tennant/publication/312355161/figure/fig2/AS:450870568591361@1484507328098/Sequence-Quality-Per-base-Before-and-After-Trimming-and-Adapter-Removal-The-per-base.png

FastQC

- Download FastQC at https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
- There are versions for all operating systems
- It is a freeware

Report Evaluation

- Overall, it is easy to locate potential problems in the FASTQ files by looking at the summary column in the report.
- The summary has several modules that show various aspects relevant for sequence quality inspection

Table 1. Various analysis modules incorporated in the FastQC program.

Analysis Modules	Definitions
Basic statistics	General statistics and some background information regarding the input file
Per base sequence quality	Bases' quality values across all the reads of the input FASTQ file
Per sequence quality scores	Average sequence quality scores for the input FASTQ file
Per base sequence content	Percentage of A, C, G, T across the FASTQ reads
Per base **GC** content	GC content across the FASTQ reads, for each base position
Per sequence GC content	Average GC distribution over all sequences, and provided a comparison of it with a normal distribution
Per base N content	Percentage of N base calls at each position across the FASTQ reads
Sequence length distribution	Summary on length distribution for the FASTQ reads, useful after trimming reads
Sequence duplication levels	Summary of the counts for every sequence in the FASTQ file, useful in detecting biased enrichment problems such as **PCR** over amplification
Overrepresented sequences	Frequency summary of sequences, useful in detecting and classifying contaminants in sequencing, for example PCR primers
Kmer content	Frequency summary of nucleotide substrings with the length of K

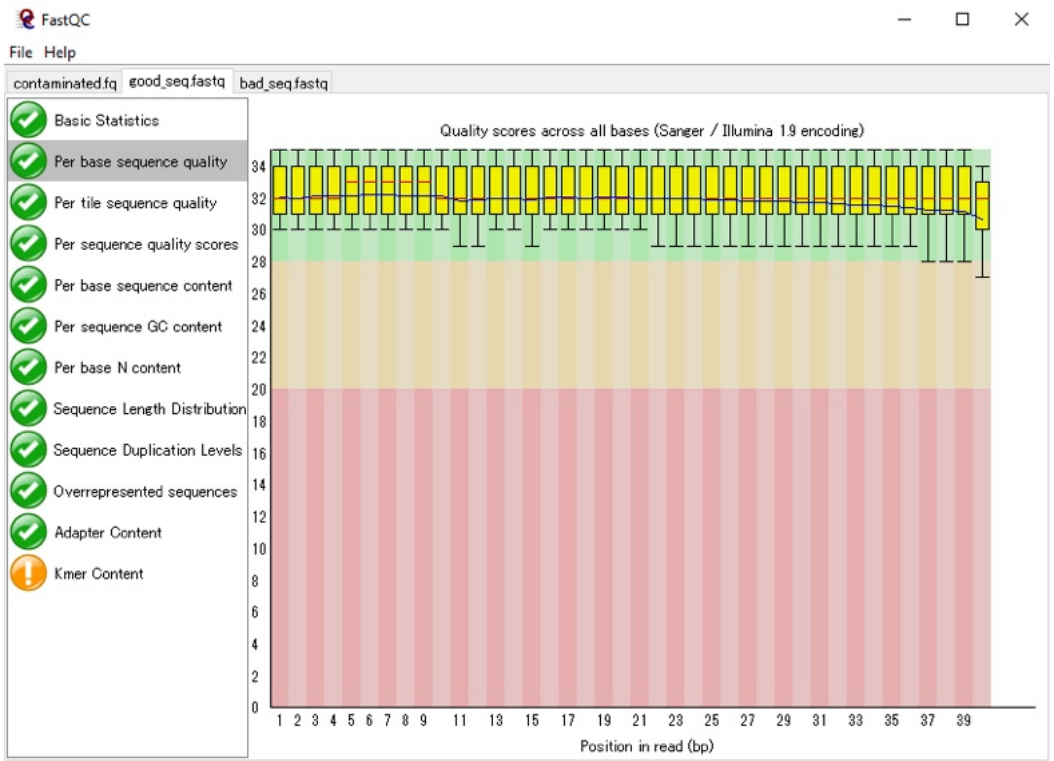
Besides each of the categories, there are symbols that represent their results:

 Represents problem in this category

 Represents acceptable in this category

 Represents warning in this category

Per base sequence quality

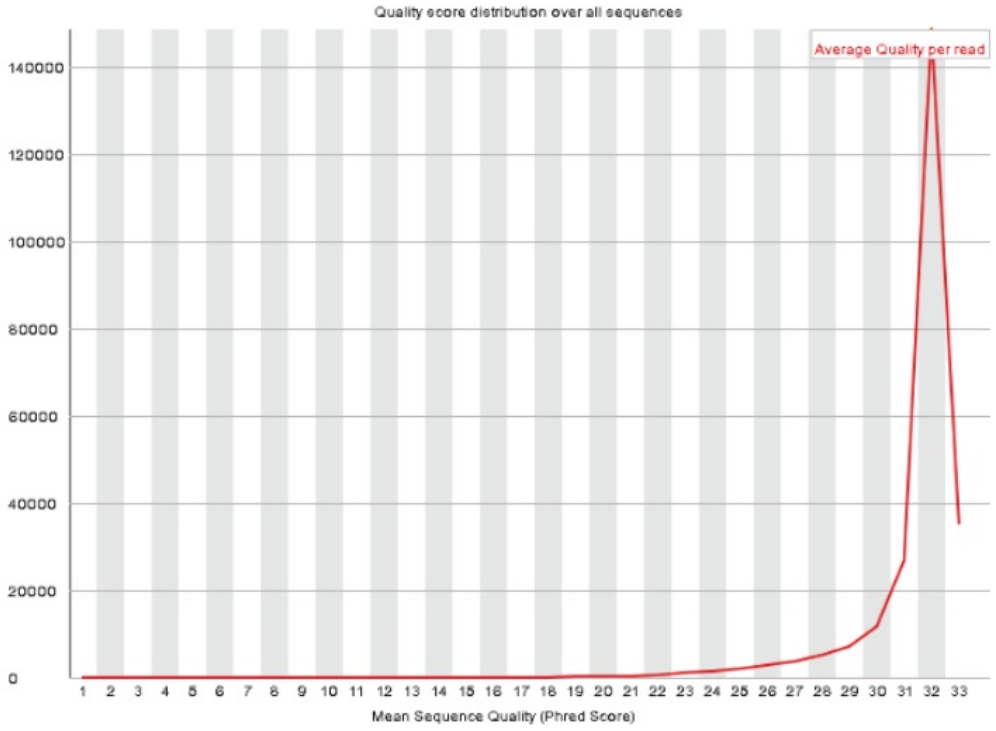


- Overall, the quality of this dataset is defined as good, because the box plots which represent the base quality were all in the green region (score >28). All the bases, on average, has base quality of >30.
- A warning will be issued if the median score <25, and a failure if <20.
- Usually, the quality of the bases deteriorates towards the end of the read, with the forward read showing better quality than the reverse read.

FastQC Report: bad

- The quality of the reads dropped drastically after base number 15. Approximately 20–40% of the bases were in the red zone, and these bases need to be trimmed.
- This is not an absolute rule, but it is fairly common to trim bases below quality of 20 or 25.

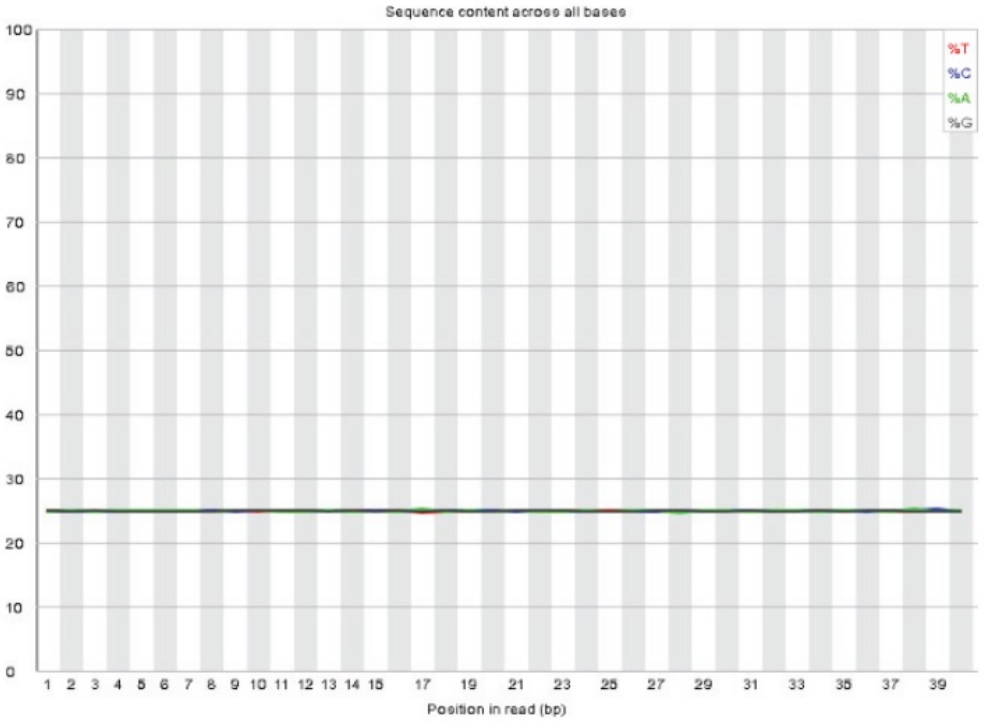
Per sequence quality scores



Example results for average quality score for all sequences.

- This figure shows the quality score distribution over all sequences. The average quality per read is actually very high, at 32.
- A good dataset will have a single peak located around score 30. Warning is given when the mean quality is <27, failure at <20.

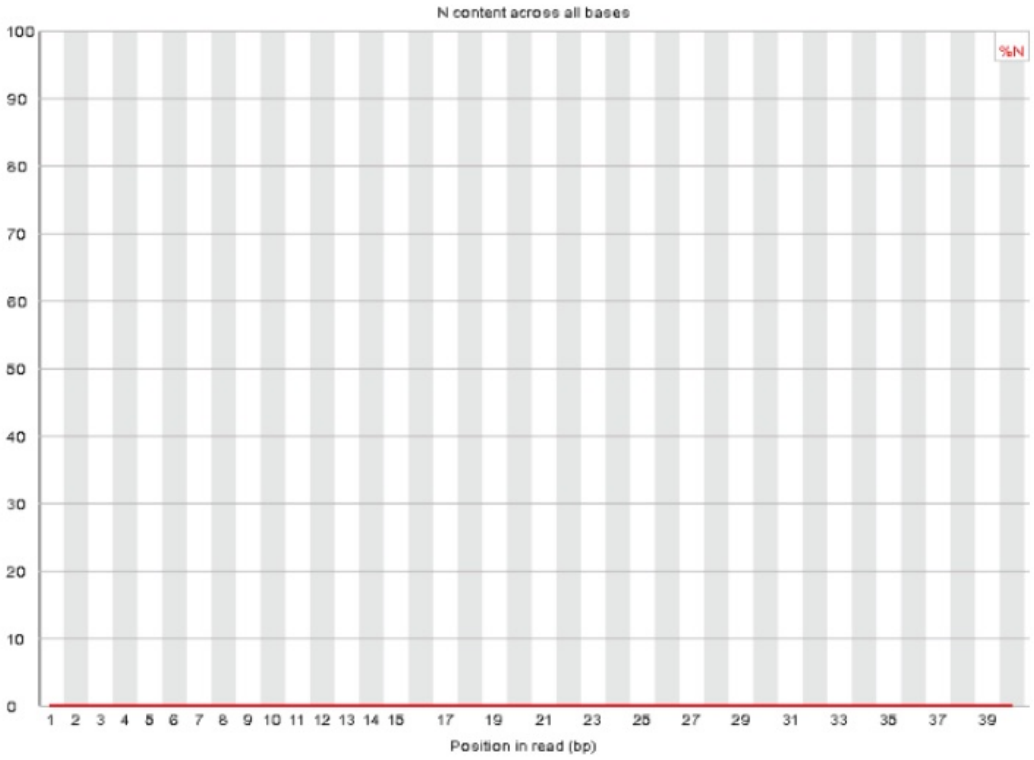
Per base sequence content



Example result for bases content of all sequences.

- In a random library you would expect that there would be little to no difference between the different bases of a sequence run, so the lines in this plot should run parallel with each other.
- At the beginning of the reads, there might be some base bias and fluctuations.
- Warning will be issued if the difference between any of the bases to be >10%, failure when this difference reaches 20% at any position. Even for a dataset with good quality, this test might not necessary pass.

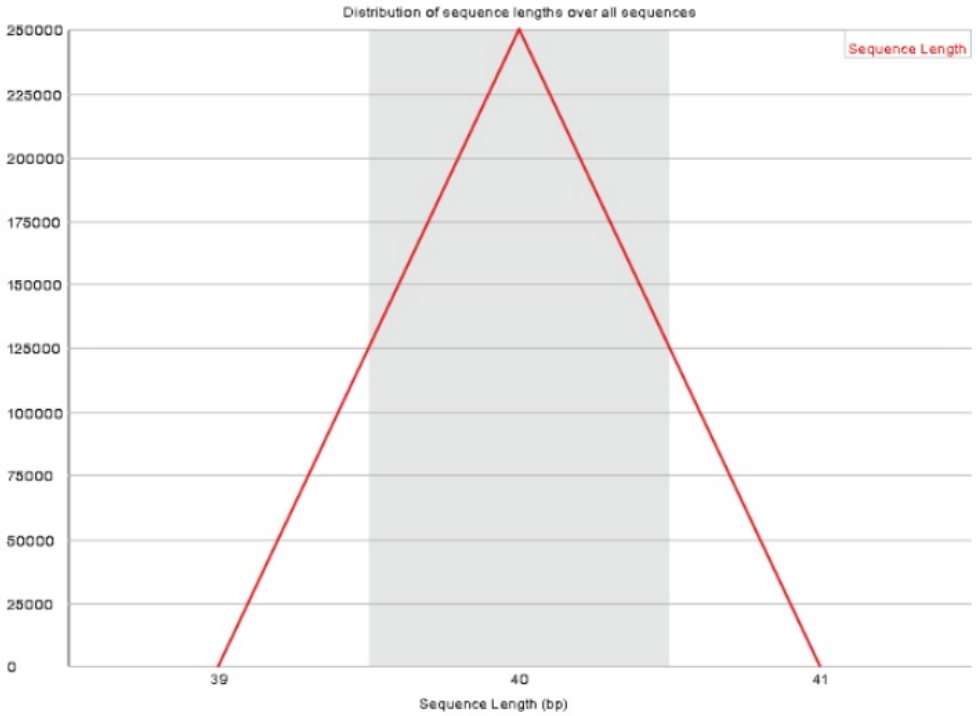
Per base N content



Example result for N (ambiguous) content across all sequences.

- This figure summarizes the number of ambiguous bases, represented as N across the entire raw reads (obtained from a random library).
- If the number of N is >5%, a warning is issued, at >20%, failure. In this case, no N is found in the dataset.

Sequence Length Distribution



Example result for distribution of sequence length over all sequences.

- This figure represents the distribution of sequence lengths. In this case, the sequence length is 40 bp.
- There are other figures generated by FastQC, but above are the figures that we find to be of most importance with regards to the base/sequence quality problem.

Sequence contamination

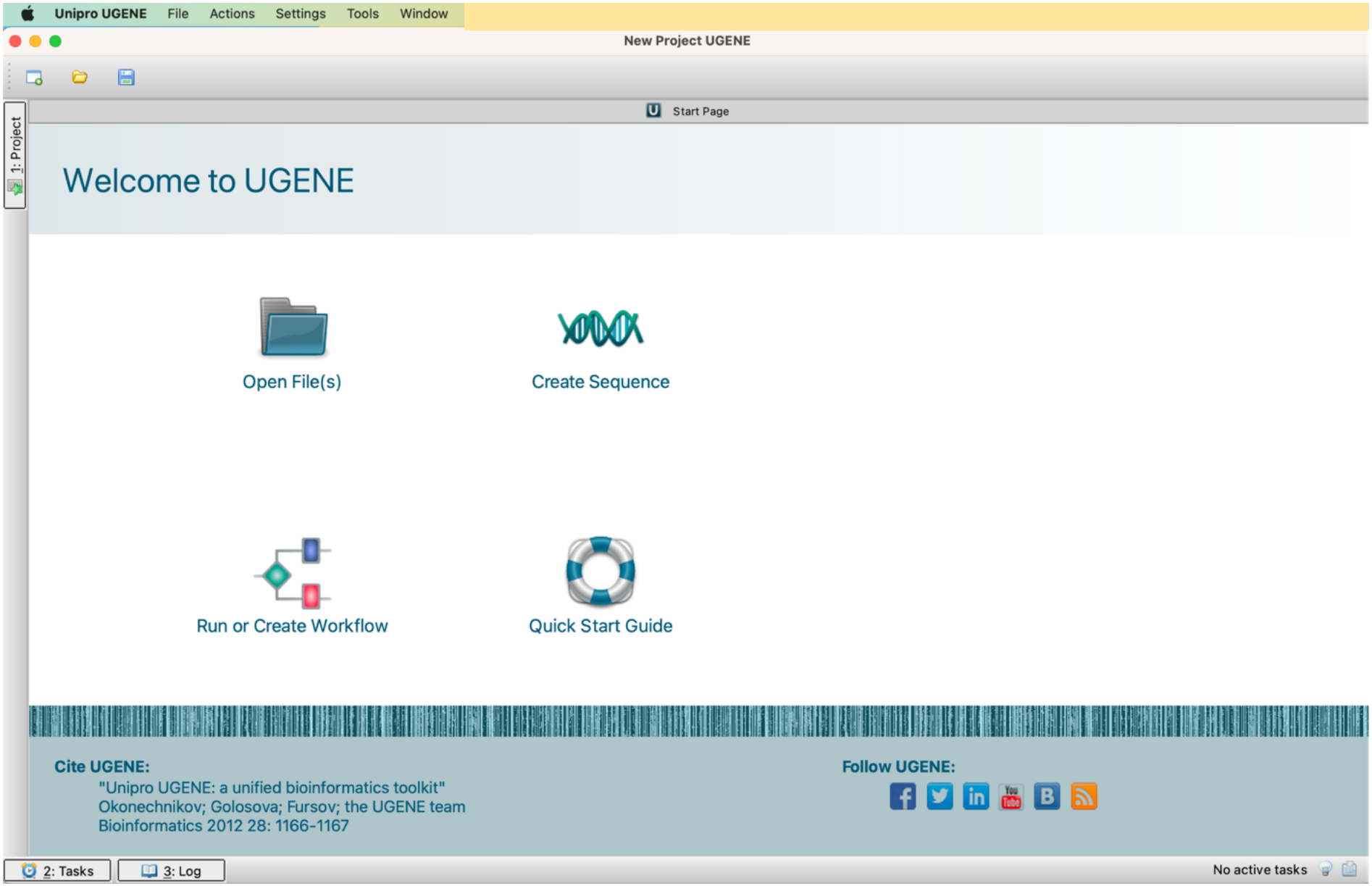
Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACATCACGA	6276	6.276	TruSeq Adapter, Index 1 (100% over 40bp)
AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTAC	6274	6.274	Illumina Single End PCR Primer 1 (100% over 40bp)
CAAGCAGAAGACGGCATACGAGATCGTGATGTGACTGGAG	6252	6.252000000000001	Illumina PCR Primer Index 1 (100% over 40bp)
CAAGCAGAAGACGGCATACGAGATACATCGGTGACTGGAG	6192	6.192	Illumina PCR Primer Index 2 (100% over 40bp)
GATCGGAAGAGCGGTTCAGCAGGAATGCCGAGACCGATCT	6142	6.142	Illumina Paired End PCR Primer 2 (100% over 40bp)

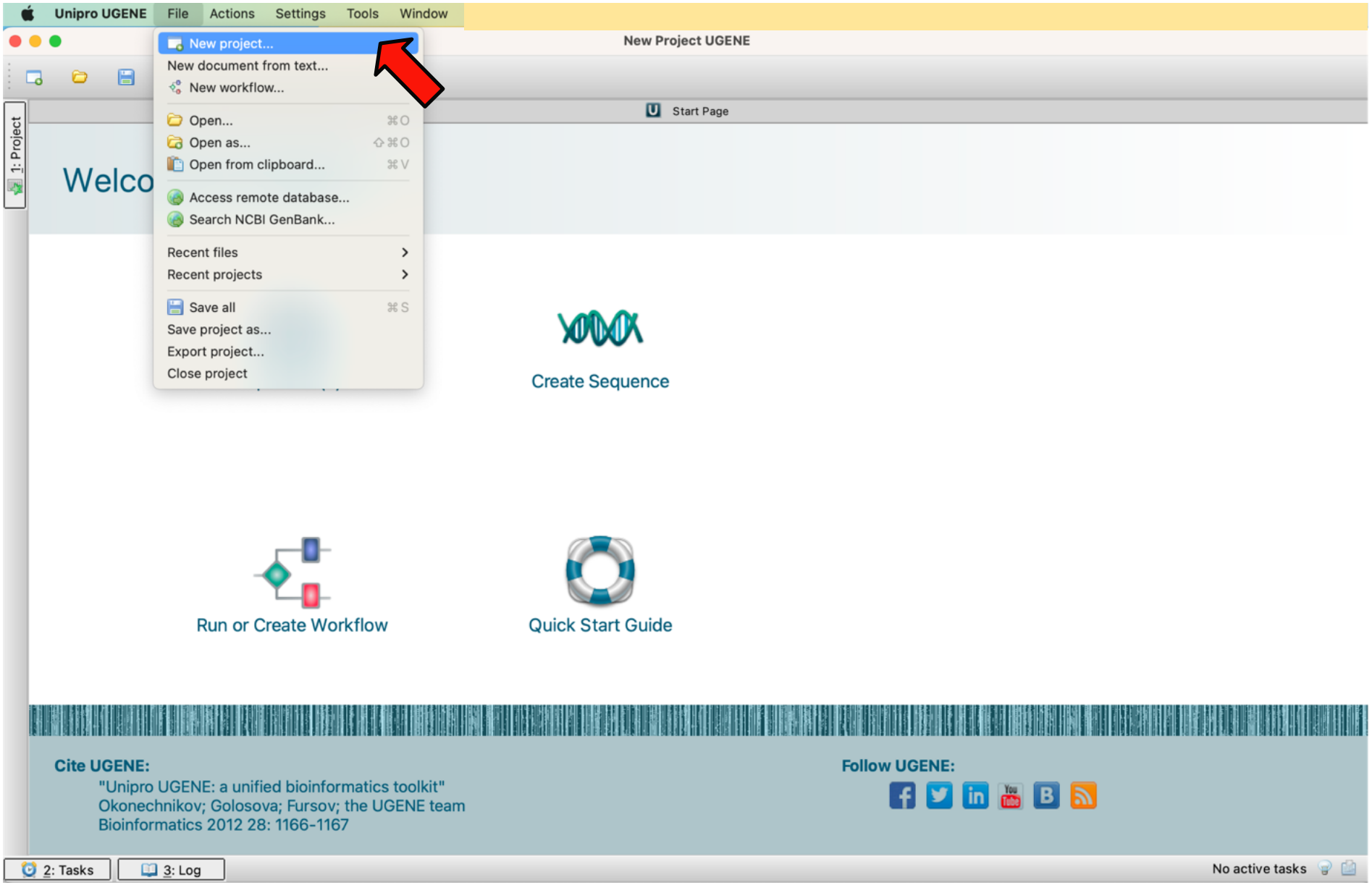
Figure 10. Example result for contamination problem.

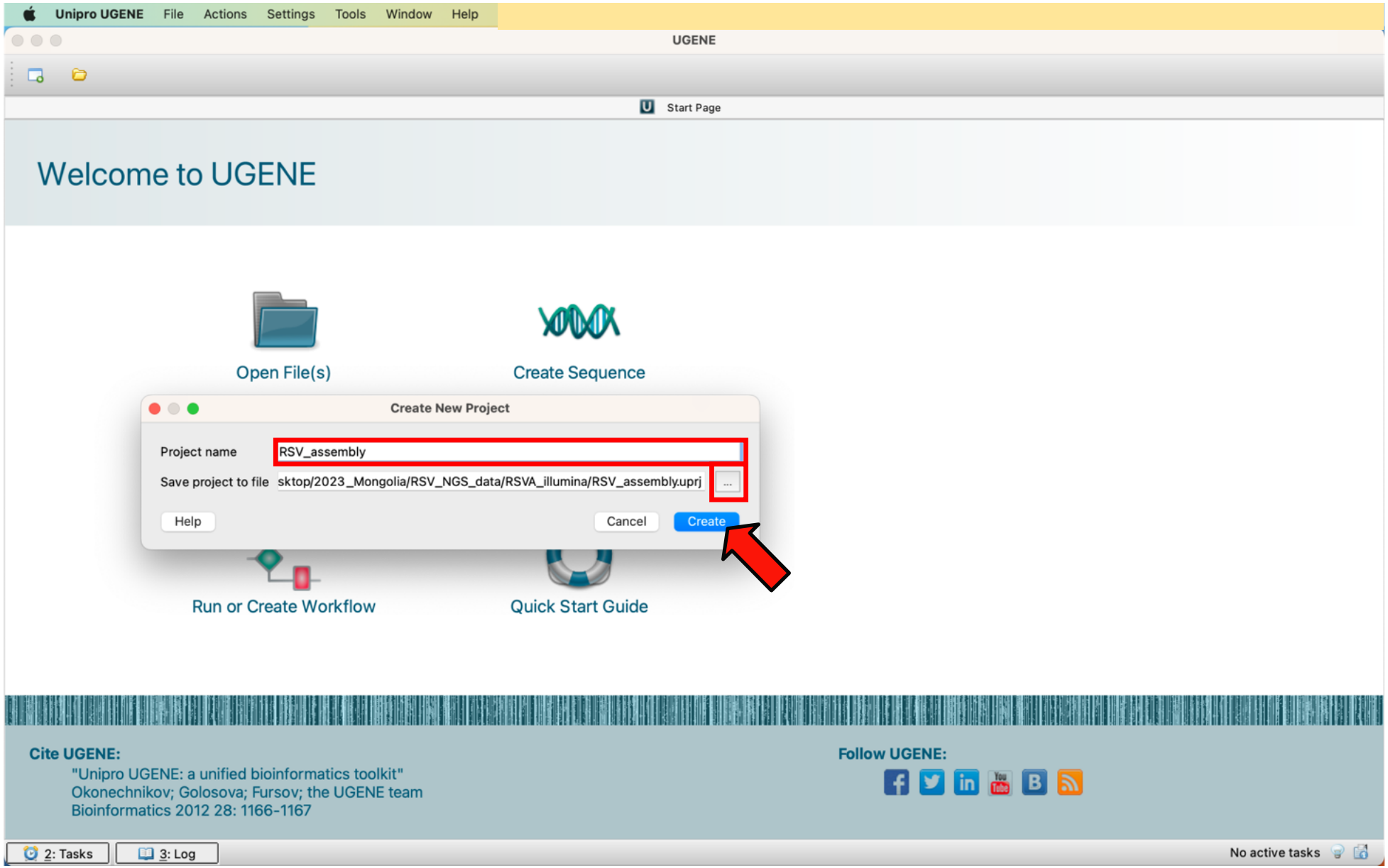
- These contaminated sequences need to be dealt with properly in the case of assembling a genome or else a lot of false genes may be generated.
- Take a look at the result under the “over represented sequences” category. Most of the time, overrepresented sequences are either primers or adapters. BLAST them!

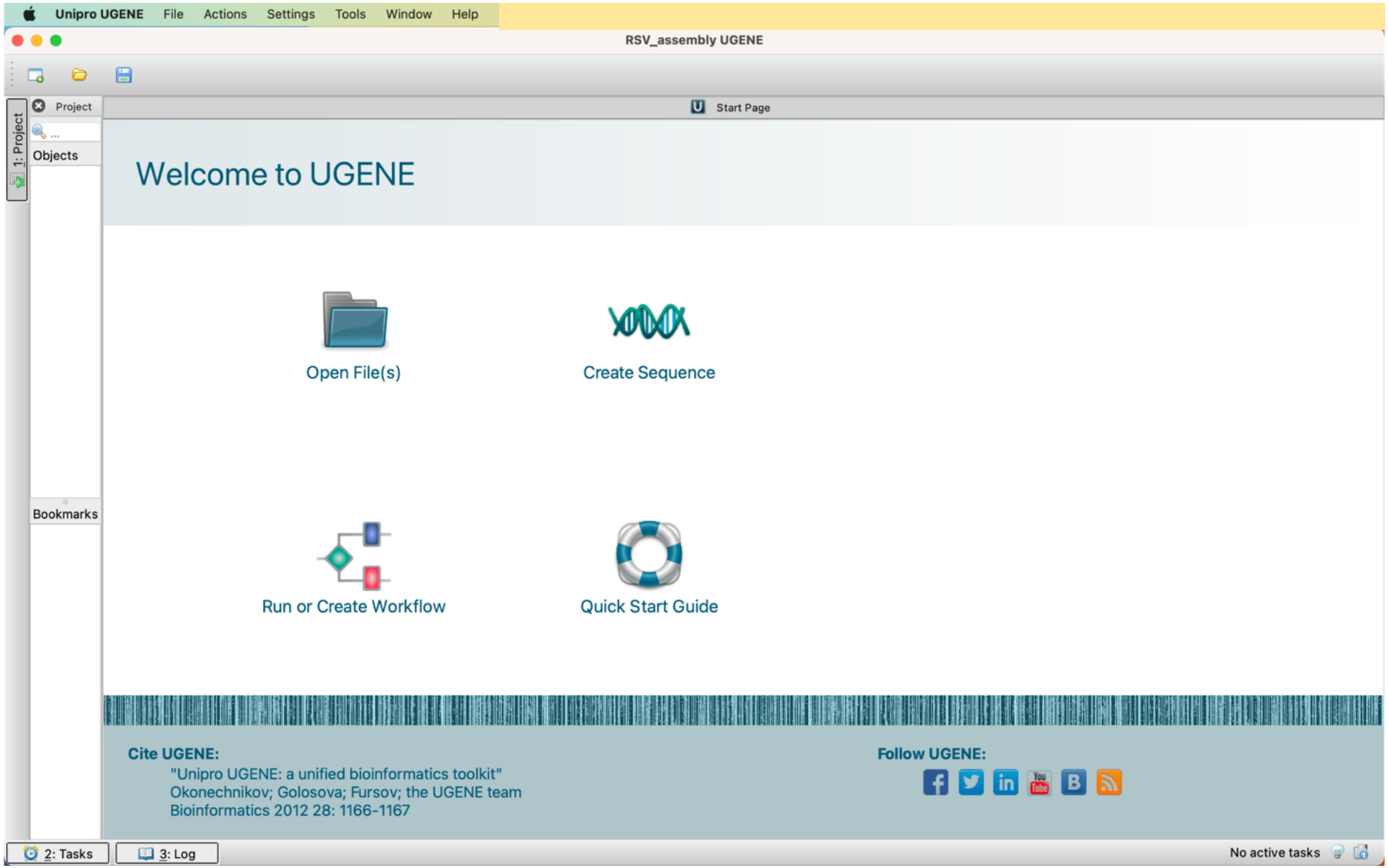
Practical: Quality assessment using FastQC

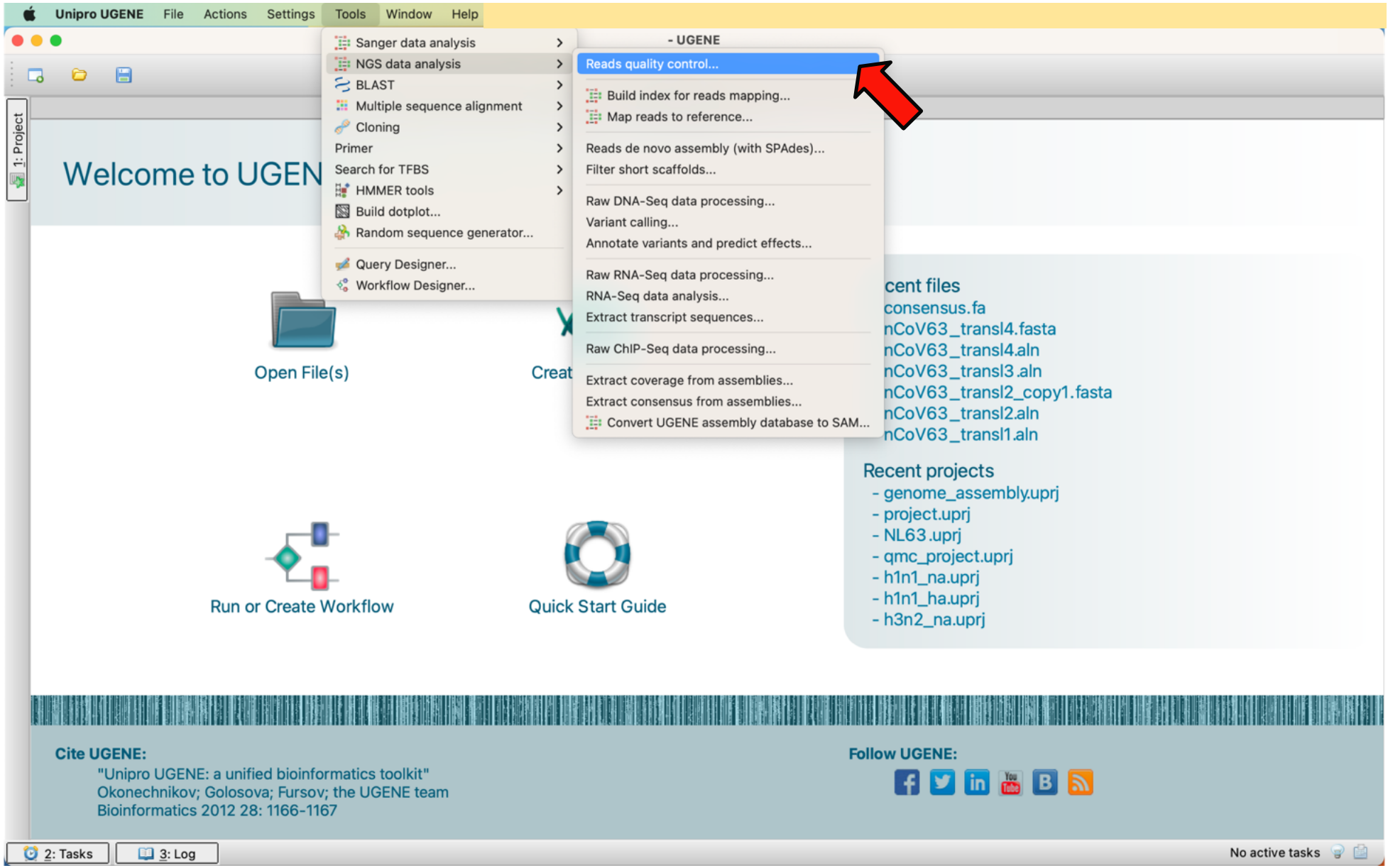
Create a project

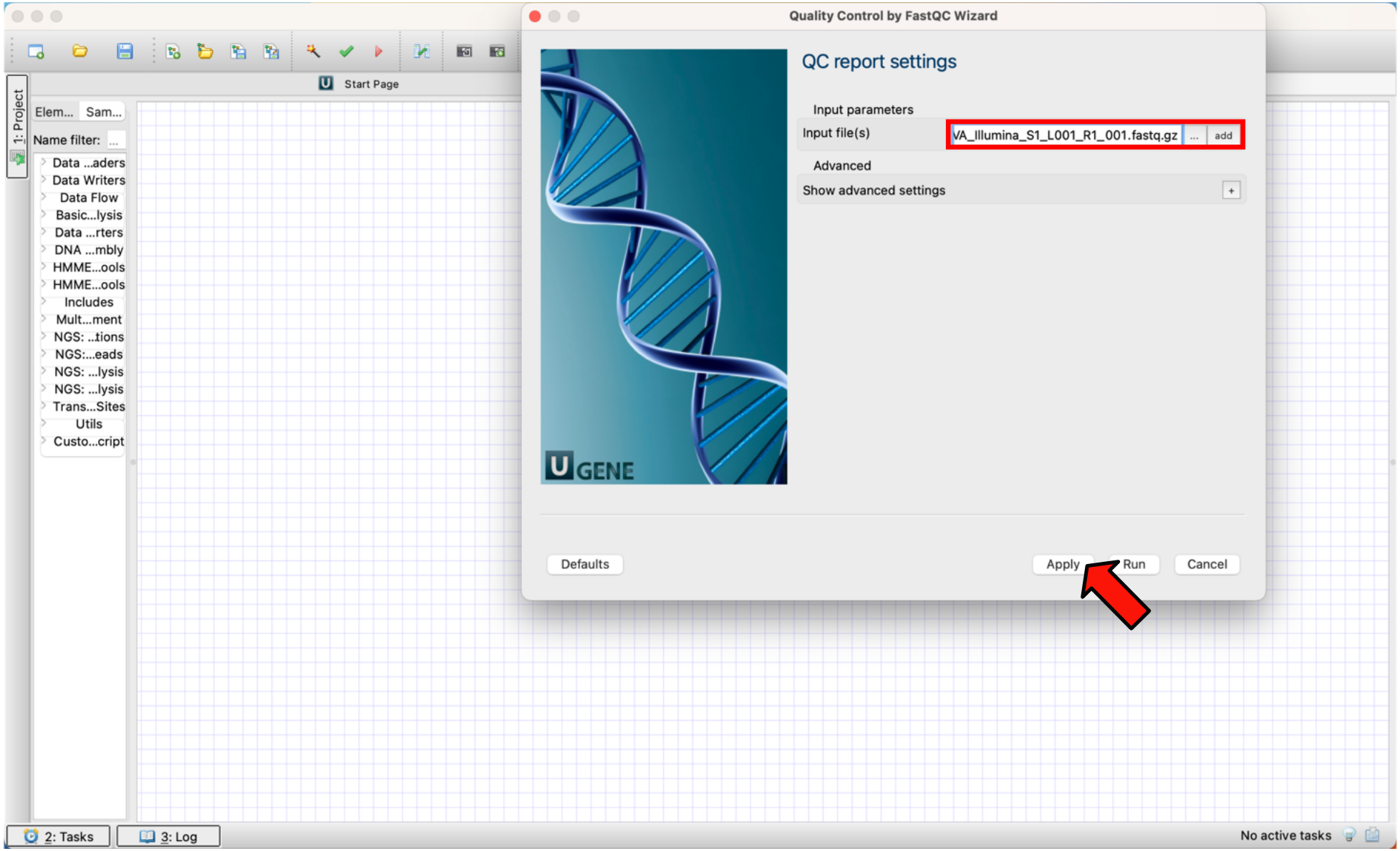

Unipro UGENE
File
Actions
Settings
Tools
Window
New Project UGENE
Start Page
1: Project
Welcome to UGENE
Open File(s)
Create Sequence
Run or Create Workflow
Quick Start Guide
Cite UGENE:
"Unipro UGENE: a unified bioinformatics toolkit"
Okonechnikov; Golosova; Fursov; the UGENE team
Bioinformatics 2012 28: 1166-1167
Follow UGENE:
2: Tasks
3: Log
No active tasks

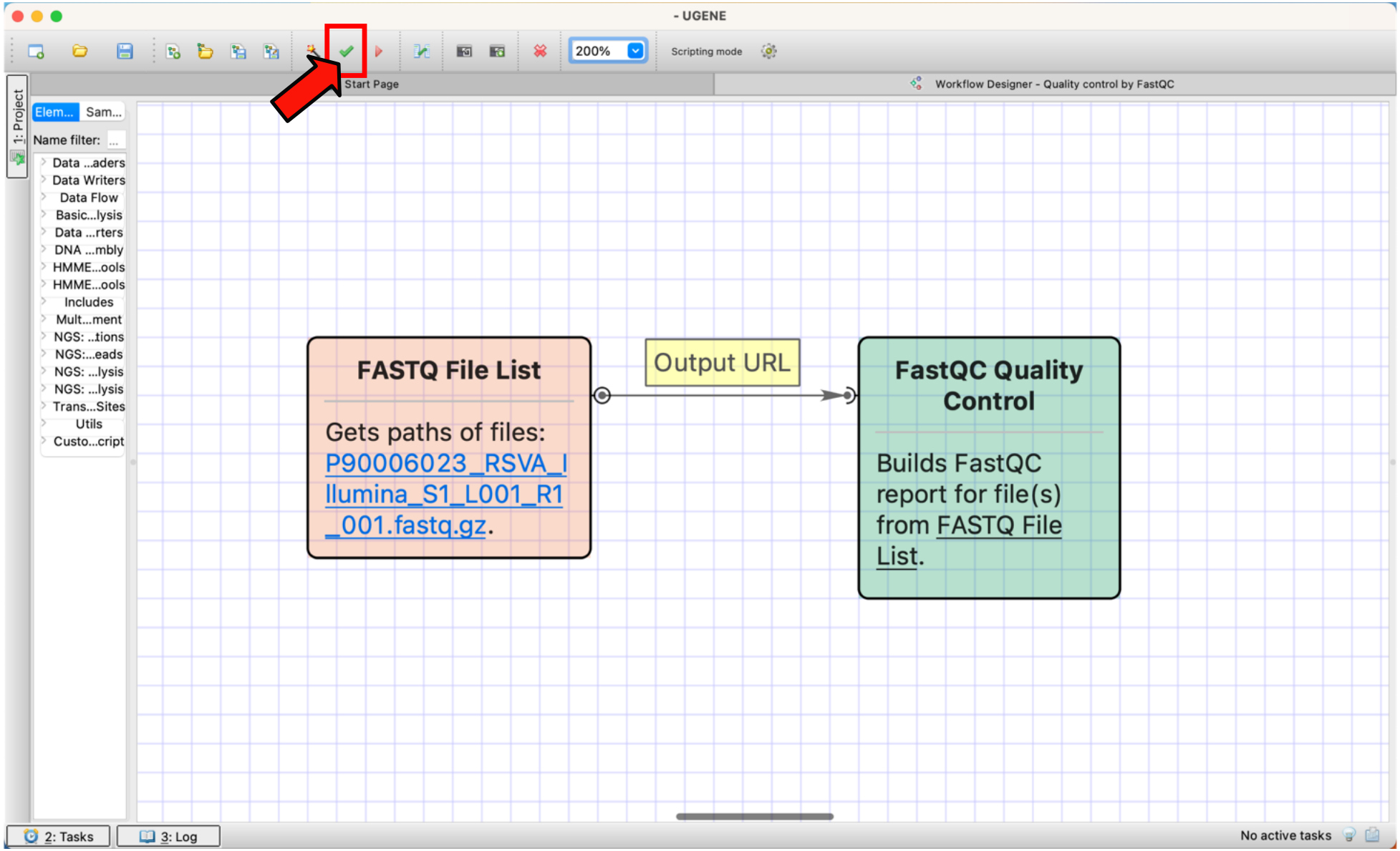

Unipro UGENE
File
Actions
Settings
Tools
Window
New Project UGENE
New project...
New document from text...
New workflow...
Open... ⌘O
Open as... ⇧⌘O
Open from clipboard... ⌘V
Access remote database...
Search NCBI GenBank...
Recent files
Recent projects
Save all ⌘S
Save project as...
Export project...
Close project
1: Project
Start Page
Welco
Create Sequence
Run or Create Workflow
Quick Start Guide
Cite UGENE:
"Unipro UGENE: a unified bioinformatics toolkit"
Okonechnikov; Golosova; Fursov; the UGENE team
Bioinformatics 2012 28: 1166-1167
Follow UGENE:
2: Tasks
3: Log
No active tasks

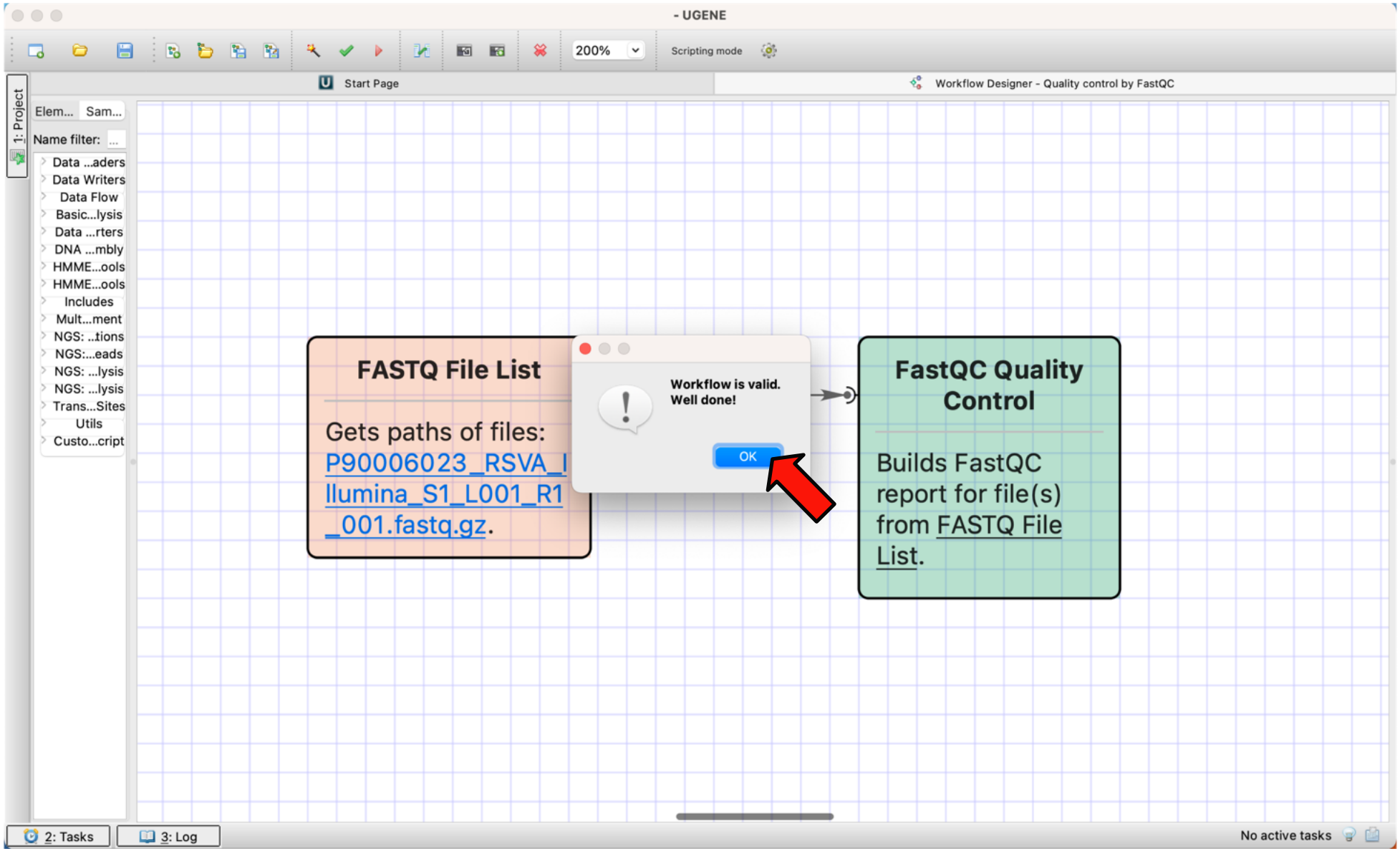

Unipro UGENE File Actions Settings Tools Window Help
UGENE
Start Page
Welcome to UGENE
Open File(s)
Create Sequence
Create New Project
Project name
RSV_assembly
Save project to file
sktop/2023_Mongolia/RSV_NGS_data/RSVA_illumina/RSV_assembly.uprj
...
Help
Cancel
Create
Run or Create Workflow
Quick Start Guide
Cite UGENE:
"Unipro UGENE: a unified bioinformatics toolkit"
Okonechnikov; Golosova; Fursov; the UGENE team
Bioinformatics 2012 28: 1166-1167
Follow UGENE:
2: Tasks
3: Log
No active tasks

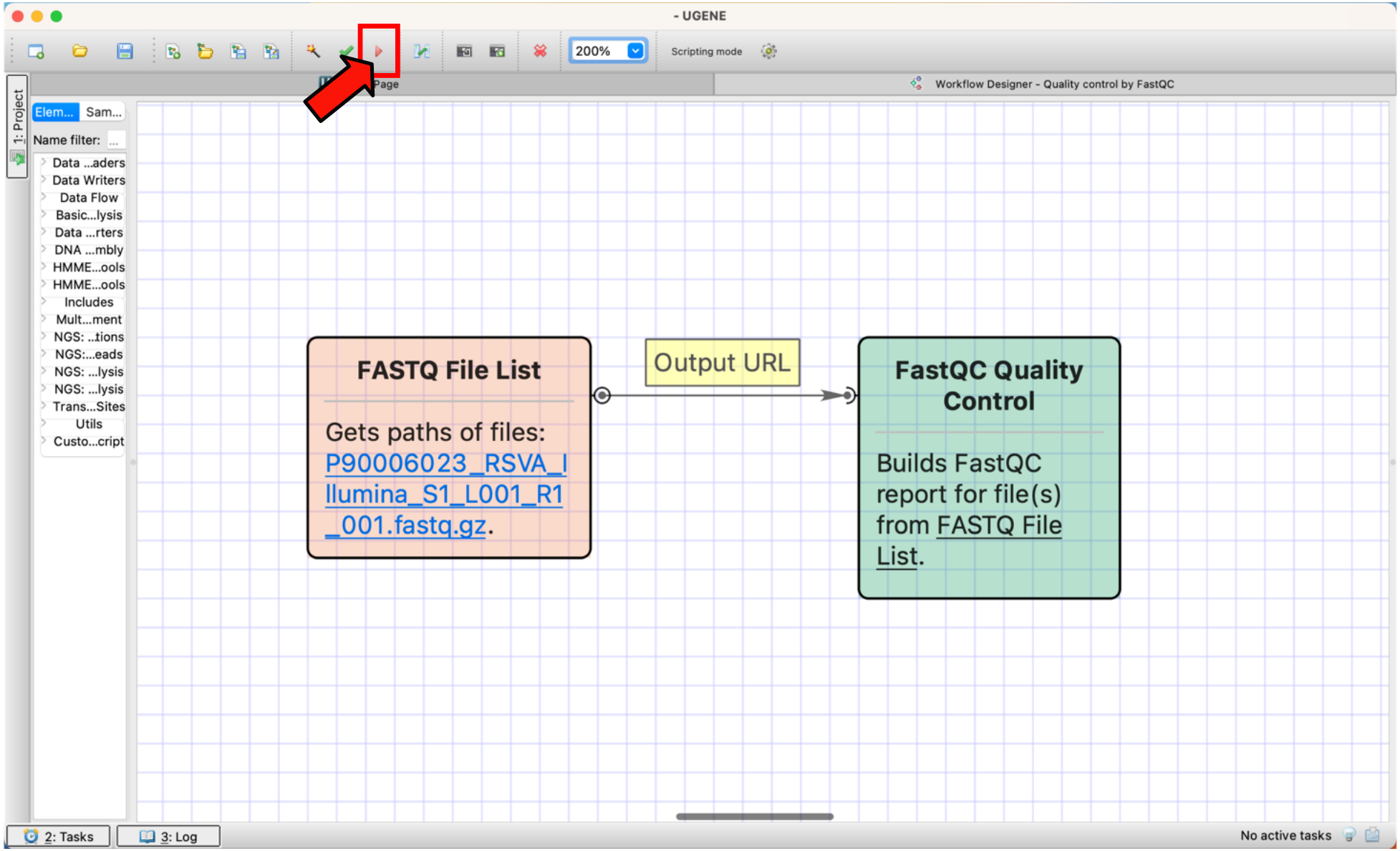

Unipro UGENE
File
Actions
Settings
Tools
Window
Help
RSV_assembly UGENE
Project
Start Page
1: Project
Objects
Bookmarks
Welcome to UGENE
Open File(s)
Create Sequence
Run or Create Workflow
Quick Start Guide
Cite UGENE:
"Unipro UGENE: a unified bioinformatics toolkit"
Okonechnikov; Golosova; Fursov; the UGENE team
Bioinformatics 2012 28: 1166-1167
Follow UGENE:
2: Tasks
3: Log
No active tasks

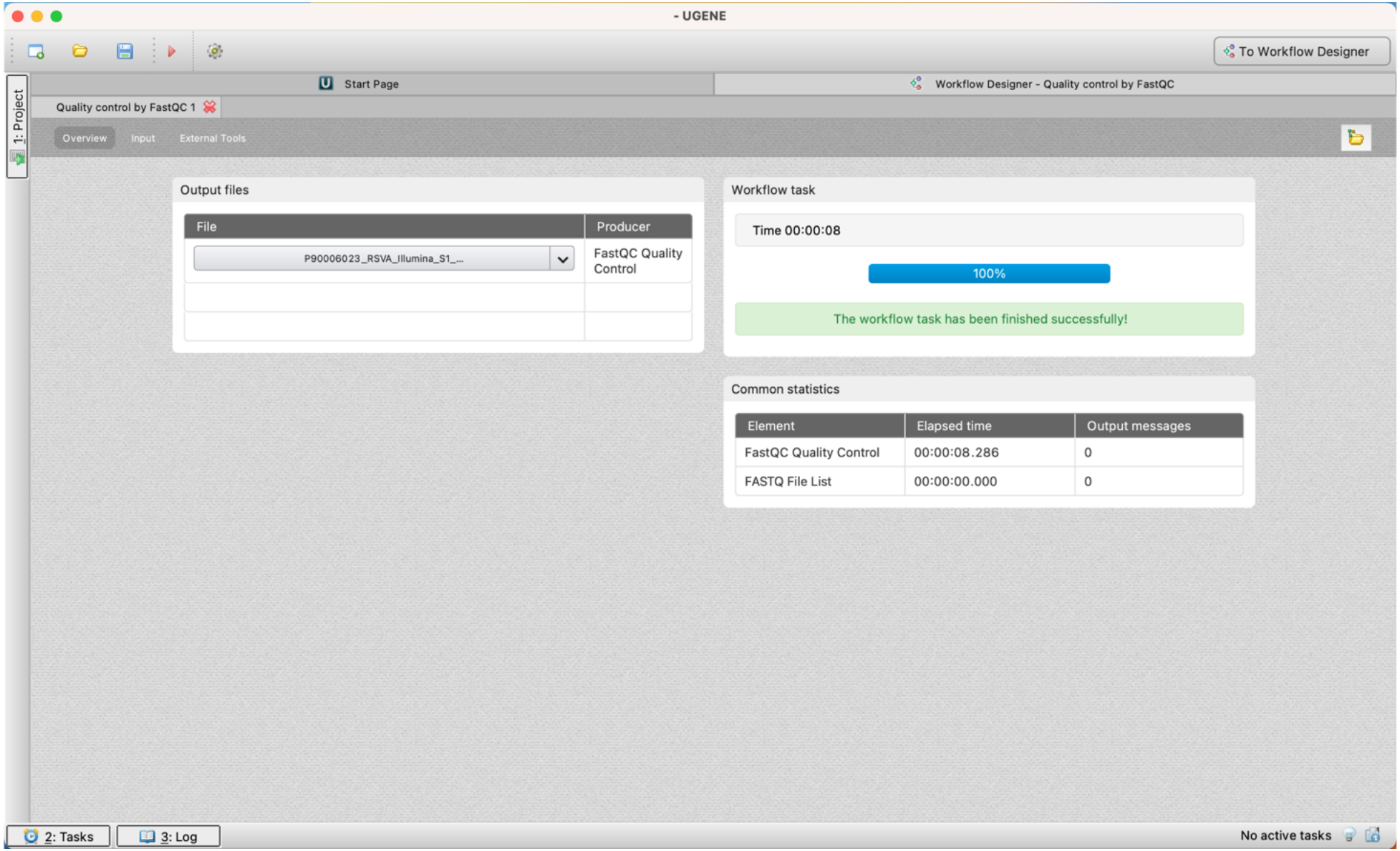

Unipro UGENE
File
Actions
Settings
Tools
Window
Help
- UGENE
Sanger data analysis
NGS data analysis
BLAST
Multiple sequence alignment
Cloning
Primer
Search for TFBS
HMMER tools
Build dotplot...
Random sequence generator...
Query Designer...
Workflow Designer...
Reads quality control...
Build index for reads mapping...
Map reads to reference...
Reads de novo assembly (with SPAdes)...
Filter short scaffolds...
Raw DNA-Seq data processing...
Variant calling...
Annotate variants and predict effects...
Raw RNA-Seq data processing...
RNA-Seq data analysis...
Extract transcript sequences...
Raw ChIP-Seq data processing...
Extract coverage from assemblies...
Extract consensus from assemblies...
Convert UGENE assembly database to SAM...
1: Project
Welcome to UGEN
Open File(s)
Creat
Run or Create Workflow
Quick Start Guide
cent files
consensus.fa
nCoV63_transl4.fasta
nCoV63_transl4.aln
nCoV63_transl3.aln
nCoV63_transl2_copy1.fasta
nCoV63_transl2.aln
nCoV63_transl1.aln
Recent projects
- genome_assembly.uprj
- project.uprj
- NL63.uprj
- qmc_project.uprj
- h1n1_na.uprj
- h1n1_ha.uprj
- h3n2_na.uprj
Cite UGENE:
"Unipro UGENE: a unified bioinformatics toolkit"
Okonechnikov; Golosova; Fursov; the UGENE team
Bioinformatics 2012 28: 1166-1167
Follow UGENE:
2: Tasks
3: Log
No active tasks

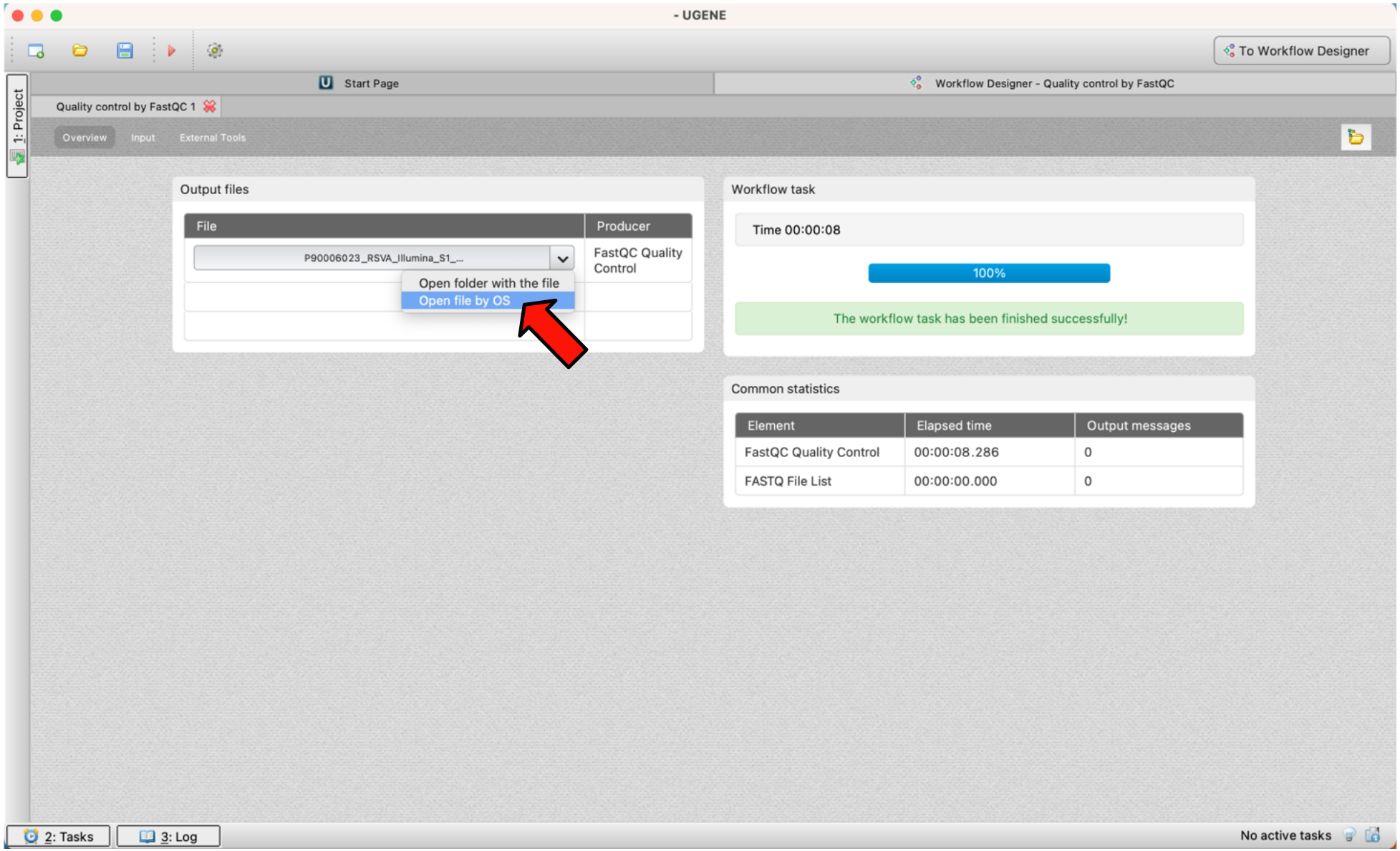

Quality Control by FastQC Wizard
QC report settings
Input parameters
Input file(s)
VA_Illumina_S1_L001_R1_001.fastq.gz
...
add
Advanced
Show advanced settings
+
UGENE
Defaults
Apply
Run
Cancel
Start Page
1: Project
Elem...
Sam...
Name filter:
Data ...aders
Data Writers
Data Flow
Basic...lysis
Data ...rters
DNA ...mbly
HMME...ools
HMME...ools
Includes
Mult...ment
NGS: ...tions
NGS:...eads
NGS: ...lysis
NGS: ...lysis
Trans...Sites
Utils
Custo...cript
2: Tasks
3: Log
No active tasks


- UGENE
200%
Scripting mode
Start Page
Workflow Designer - Quality control by FastQC
1: Project
Elem...
Sam...
Name filter:
Data ...aders
Data Writers
Data Flow
Basic...lysis
Data ...rters
DNA ...mbly
HMME...ools
HMME...ools
Includes
Mult...ment
NGS: ...tions
NGS:...eads
NGS: ...lysis
NGS: ...lysis
Trans...Sites
Utils
Custo...cript
FASTQ File List
Gets paths of files: P90006023_RSVA_Illumina_S1_L001_R1_001.fastq.gz.
Output URL
FastQC Quality Control
Builds FastQC report for file(s) from FASTQ File List.
2: Tasks
3: Log
No active tasks


- UGENE
200%
Scripting mode
Start Page
Workflow Designer - Quality control by FastQC
1: Project
Elem...
Sam...
Name filter:
Data ...aders
Data Writers
Data Flow
Basic...lysis
Data ...rters
DNA ...mbly
HMME...ools
HMME...ools
Includes
Mult...ment
NGS: ...tions
NGS:...eads
NGS: ...lysis
NGS: ...lysis
Trans...Sites
Utils
Custo...cript
FASTQ File List
Gets paths of files: P90006023_RSVA_Illumina_S1_L001_R1_001.fastq.gz.
Workflow is valid.
Well done!
OK
FastQC Quality Control
Builds FastQC report for file(s) from FASTQ File List.
2: Tasks
3: Log
No active tasks


- UGENE
200%
Scripting mode
Page
Workflow Designer - Quality control by FastQC
1: Project
Elem...
Sam...
Name filter:
Data ...aders
Data Writers
Data Flow
Basic...lysis
Data ...rters
DNA ...mbly
HMME...ools
HMME...ools
Includes
Mult...ment
NGS: ...tions
NGS:...eads
NGS: ...lysis
NGS: ...lysis
Trans...Sites
Utils
Custo...cript
FASTQ File List
Gets paths of files: P90006023_RSVA_Illumina_S1_L001_R1_001.fastq.gz.
Output URL
FastQC Quality Control
Builds FastQC report for file(s) from FASTQ File List.
2: Tasks
3: Log
No active tasks


- UGENE
To Workflow Designer
Start Page
Workflow Designer - Quality control by FastQC
1: Project
Quality control by FastQC 1
Overview
Input
External Tools
Output files
File
Producer
P90006023_RSVA_Illumina_S1_...
FastQC Quality Control
Workflow task
Time 00:00:08
100%
The workflow task has been finished successfully!
Common statistics
Element
Elapsed time
Output messages
FastQC Quality Control
00:00:08.286
0
FASTQ File List
00:00:00.000
0
2: Tasks
3: Log
No active tasks


- UGENE
To Workflow Designer
Start Page
Workflow Designer - Quality control by FastQC
1: Project
Quality control by FastQC 1
Overview
Input
External Tools
Output files
File
Producer
P90006023_RSVA_Illumina_S1_...
FastQC Quality Control
Open folder with the file
Open file by OS
Workflow task
Time 00:00:08
100%
The workflow task has been finished successfully!
Common statistics
Element
Elapsed time
Output messages
FastQC Quality Control
00:00:08.286
0
FASTQ File List
00:00:00.000
0
2: Tasks
3: Log
No active tasks

FastQC Report

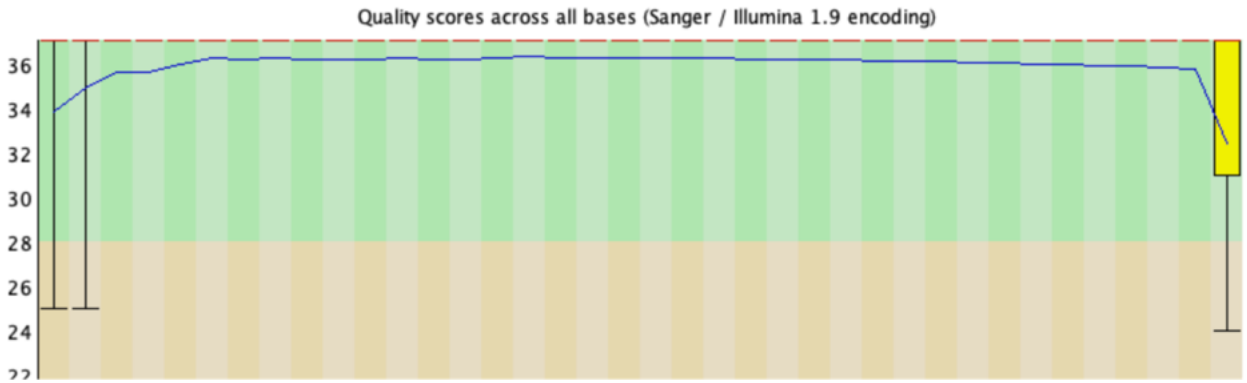
Thu. 5 Oct. 2023
P90006023_RSVA_Illumina_S1_L001_R1_001.fastq.gz

Summary

Basic Statistics

Measure	Value
Filename	P90006023_RSVA_Illumina_S1_L001_R1_001.fastq.gz
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	94076
Sequences flagged as poor quality	0
Sequence length	35-151
%GC	36

Per base sequence quality

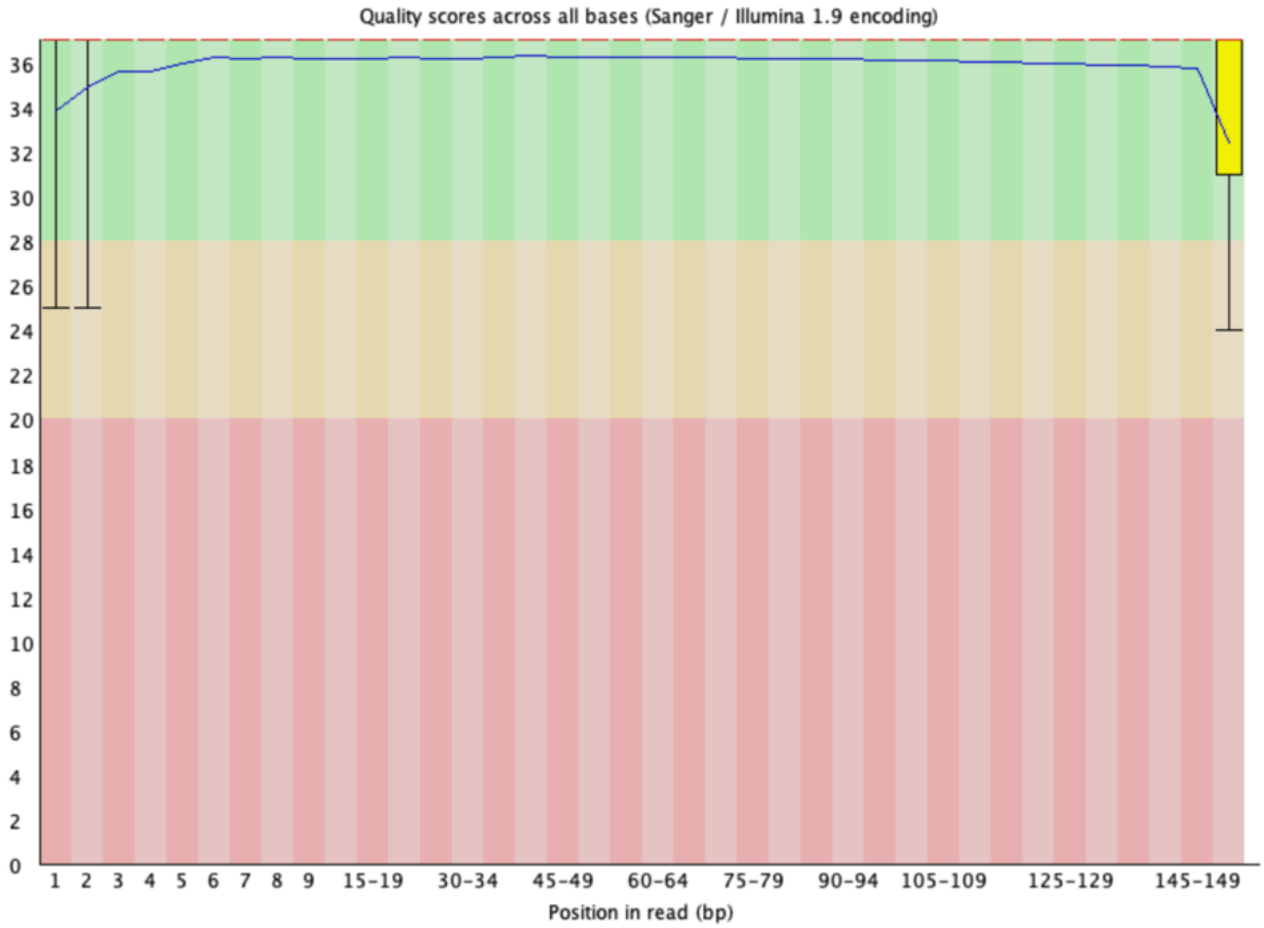


FastQC Report

Thu. 5 Oct. 2023
P90006023_RSVA_Illumina_S1_L001_R1_001.fastq.gz

Summary

Per base sequence quality

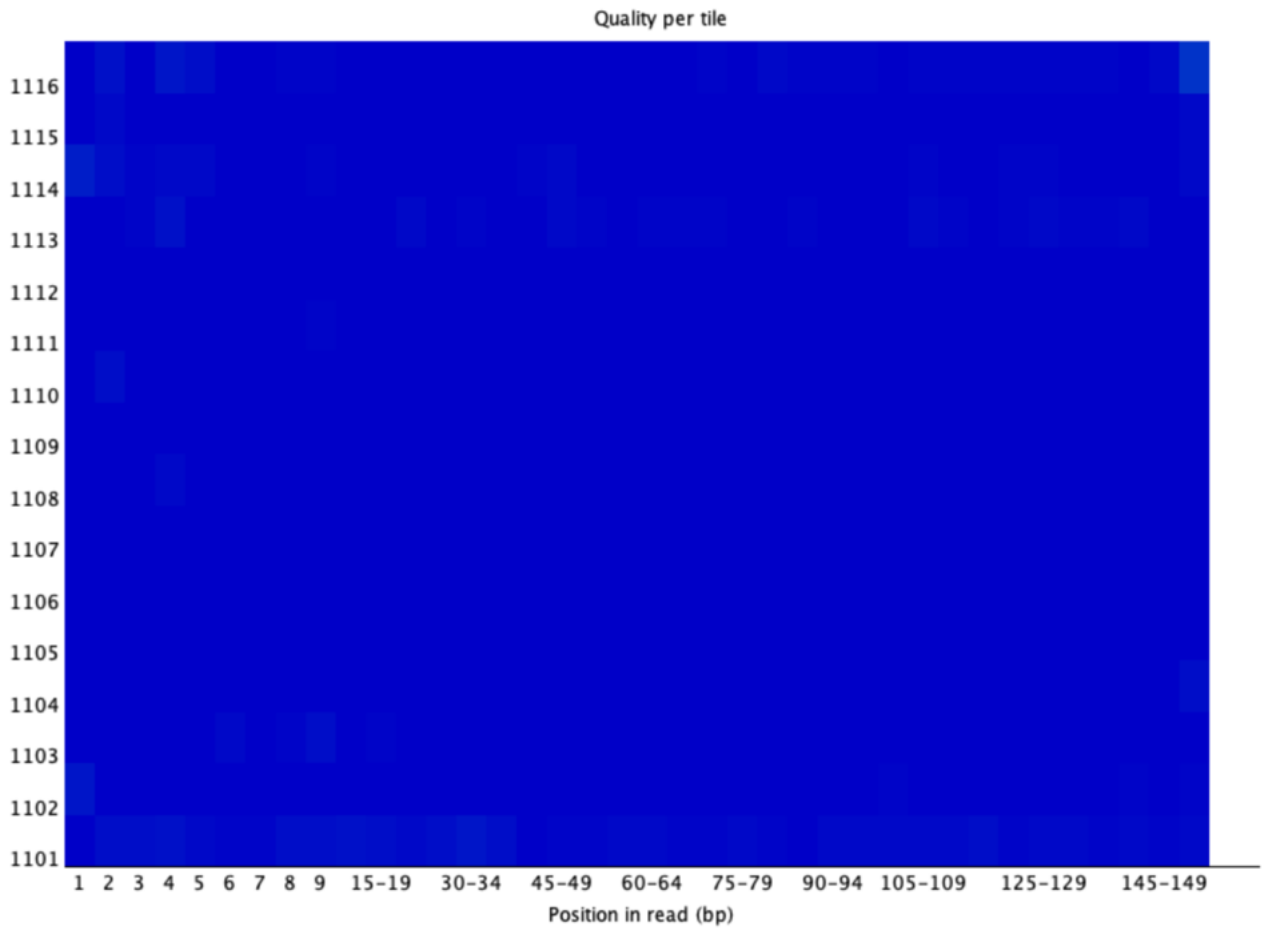


FastQC Report

Thu. 5 Oct. 2023
P90006023_RSVA_Illumina_S1_L001_R1_001.fastq.gz

Summary

Per tile sequence quality

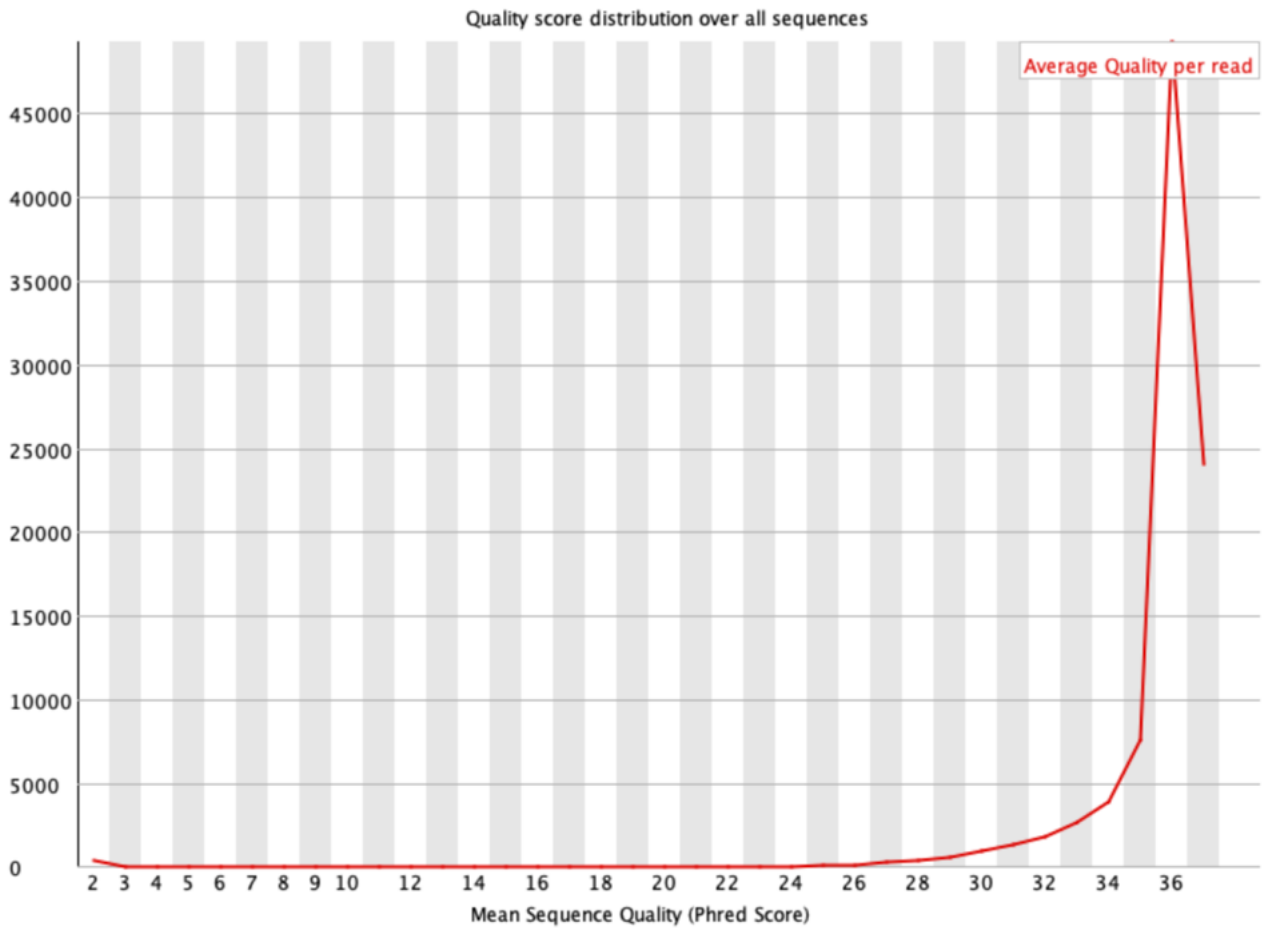


FastQC Report

Thu. 5 Oct. 2023
P90006023_RSVA_Illumina_S1_L001_R1_001.fastq.gz

Summary

Per sequence quality scores

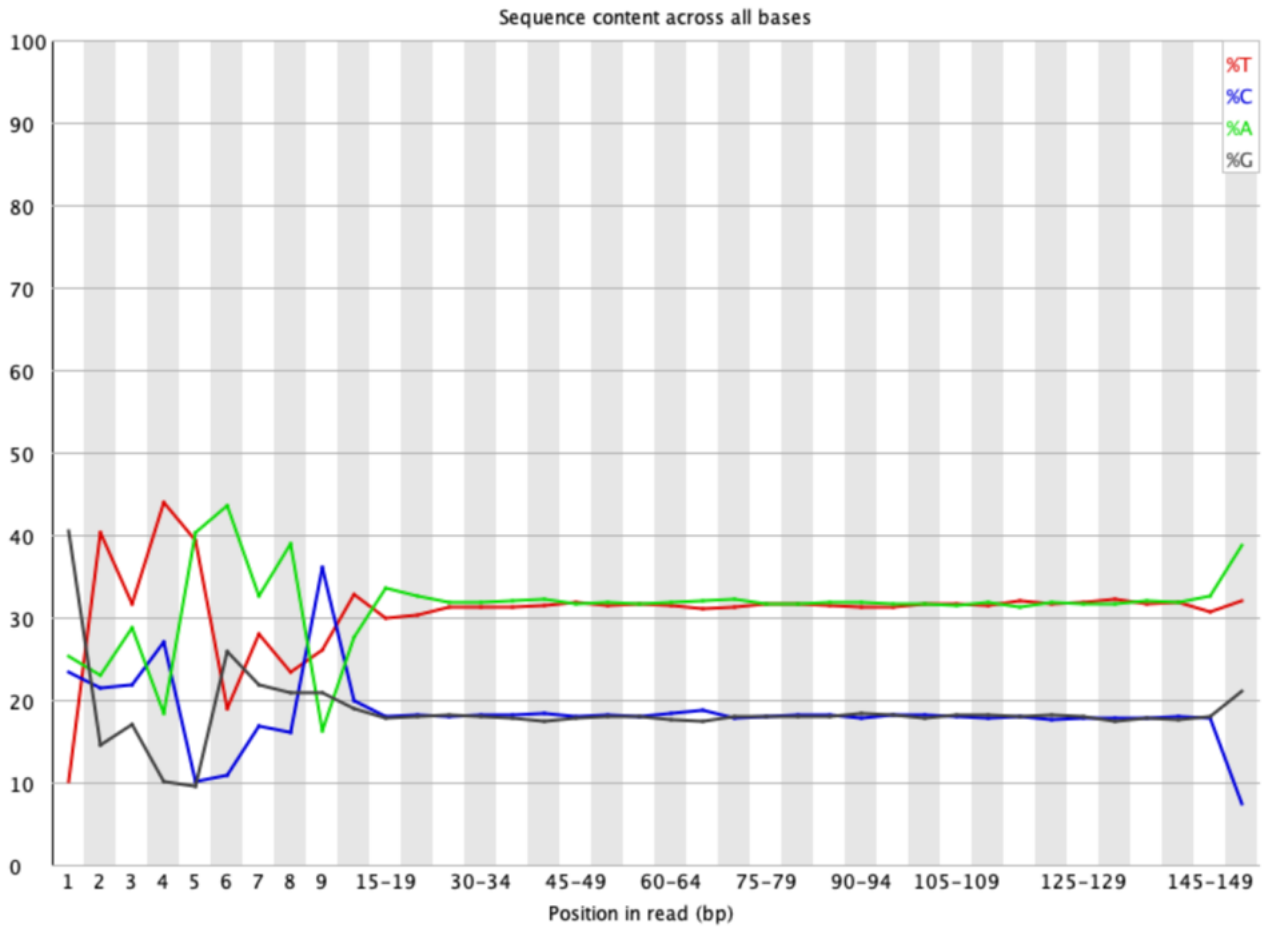


FastQC Report

Thu. 5 Oct. 2023
P90006023_RSVA_Illumina_S1_L001_R1_001.fastq.gz

Summary

Per base sequence content

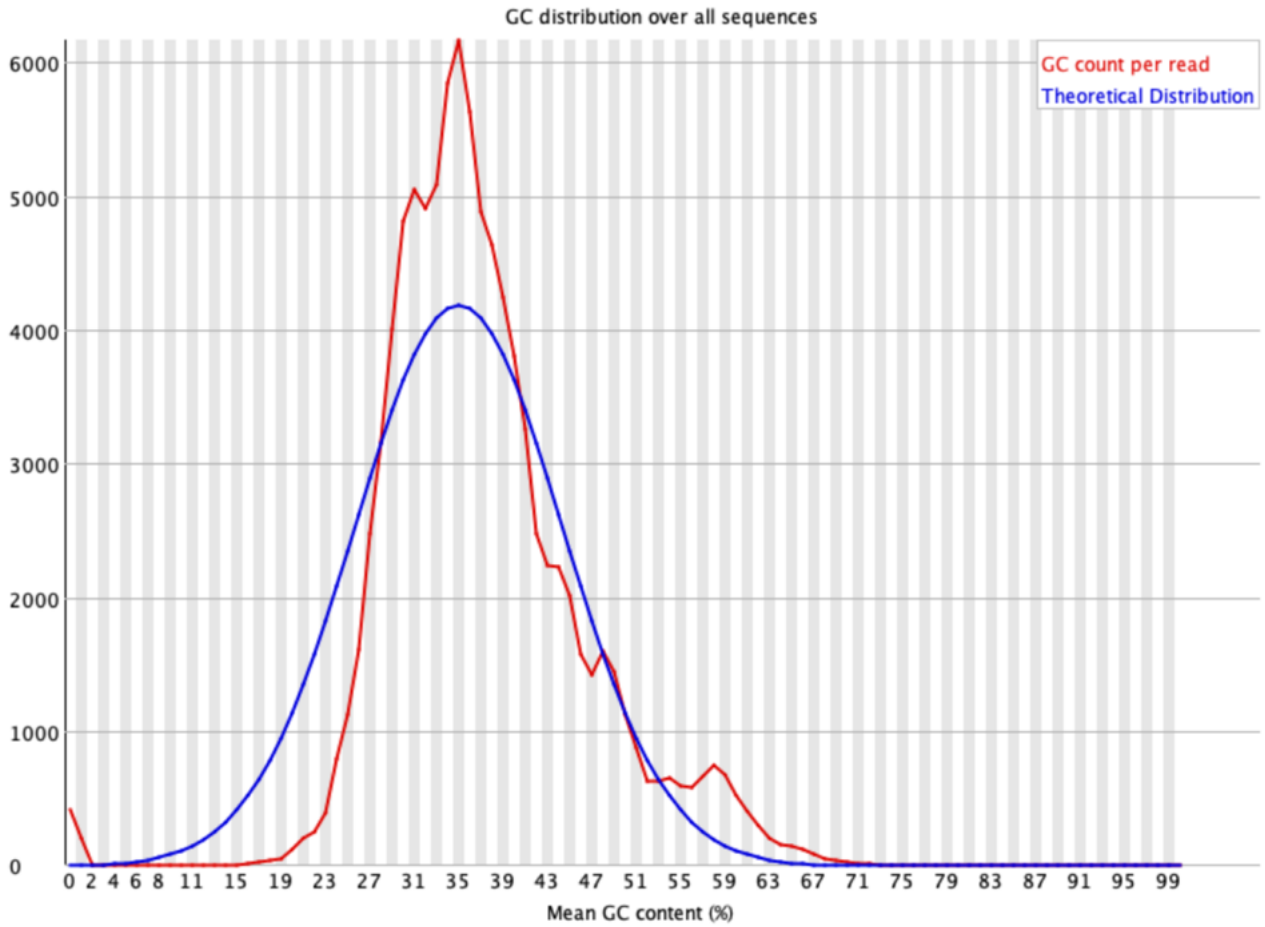


FastQC Report

Thu. 5 Oct. 2023
P90006023_RSVA_Illumina_S1_L001_R1_001.fastq.gz

Summary

Per sequence GC content

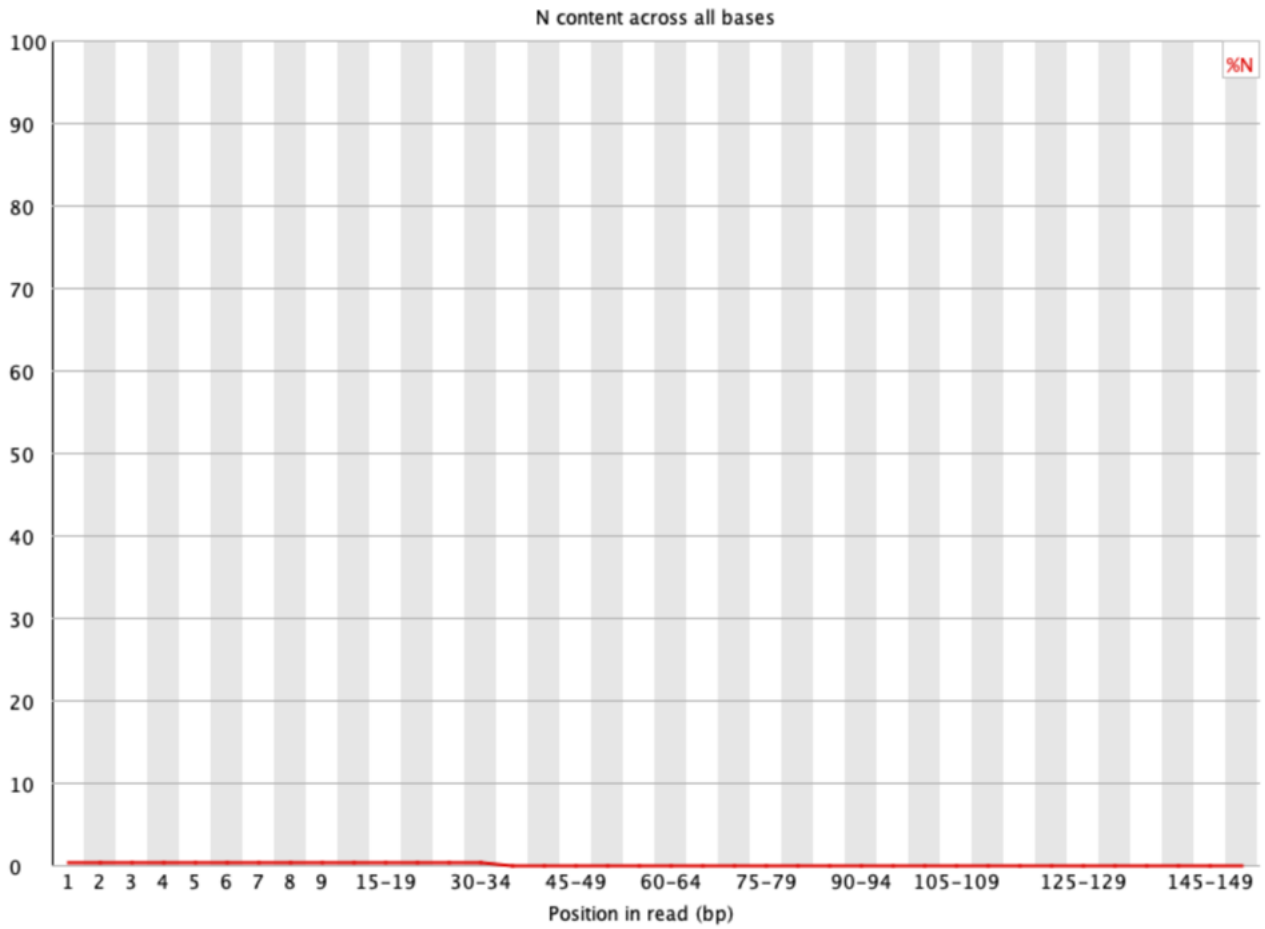


FastQC Report

Thu. 5 Oct. 2023
P90006023_RSVA_Illumina_S1_L001_R1_001.fastq.gz

Summary

Per base N content

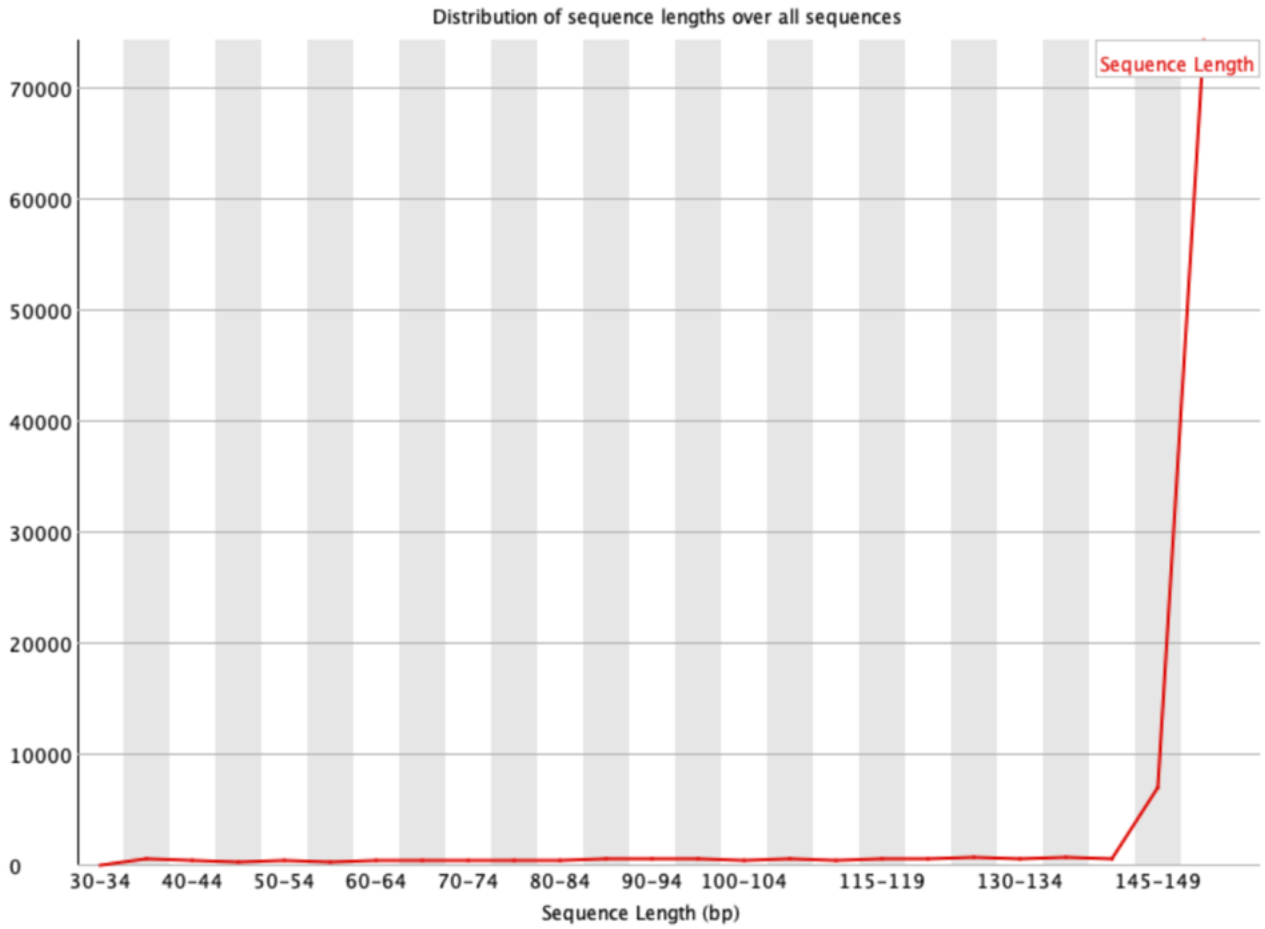


FastQC Report

Thu. 5 Oct. 2023
P90006023_RSVA_Illumina_S1_L001_R1_001.fastq.gz

Summary

Sequence Length Distribution

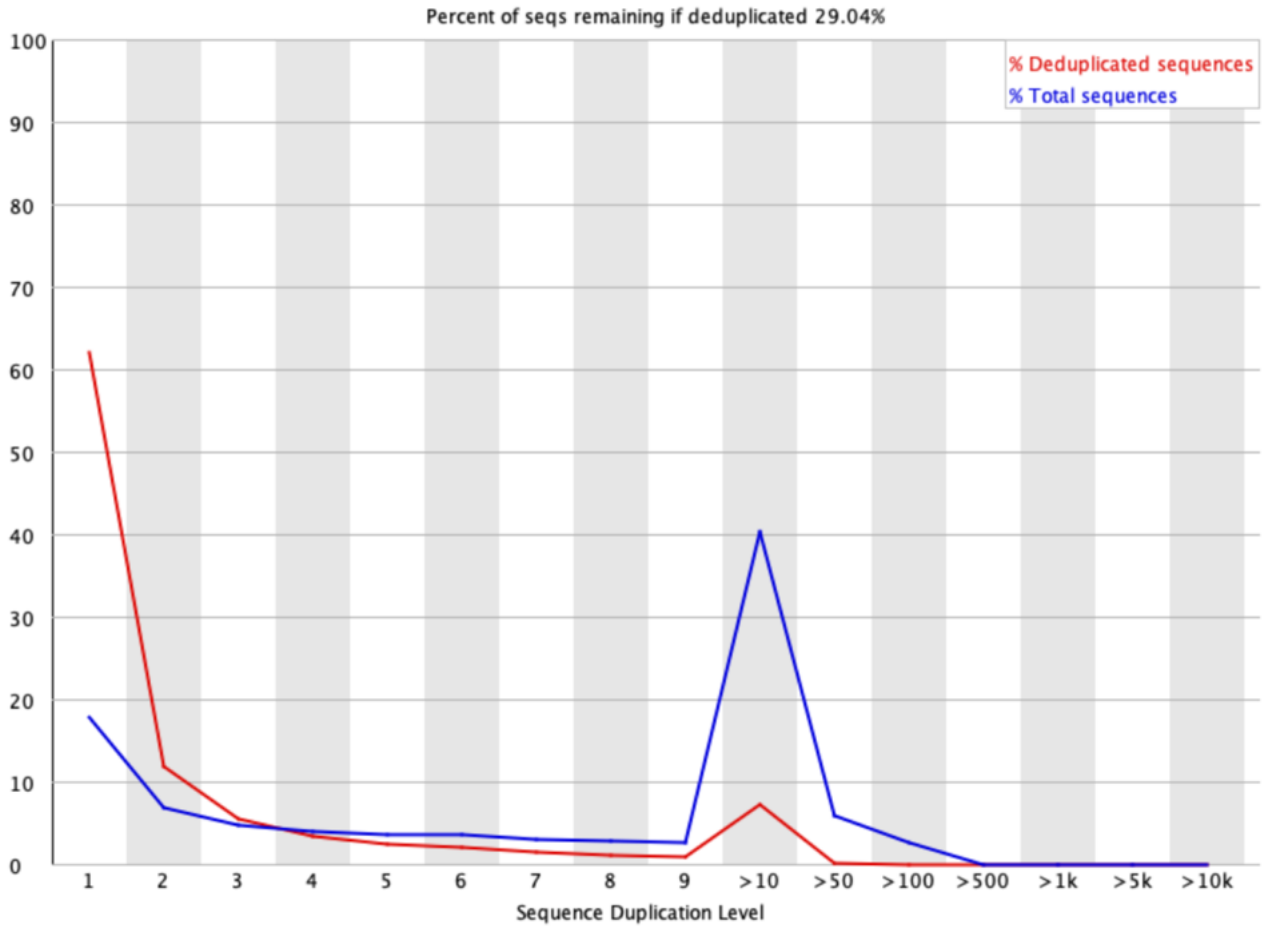


FastQC Report

Thu. 5 Oct. 2023
P90006023_RSVA_Illumina_S1_L001_R1_001.fastq.gz

Summary

Sequence Duplication Levels



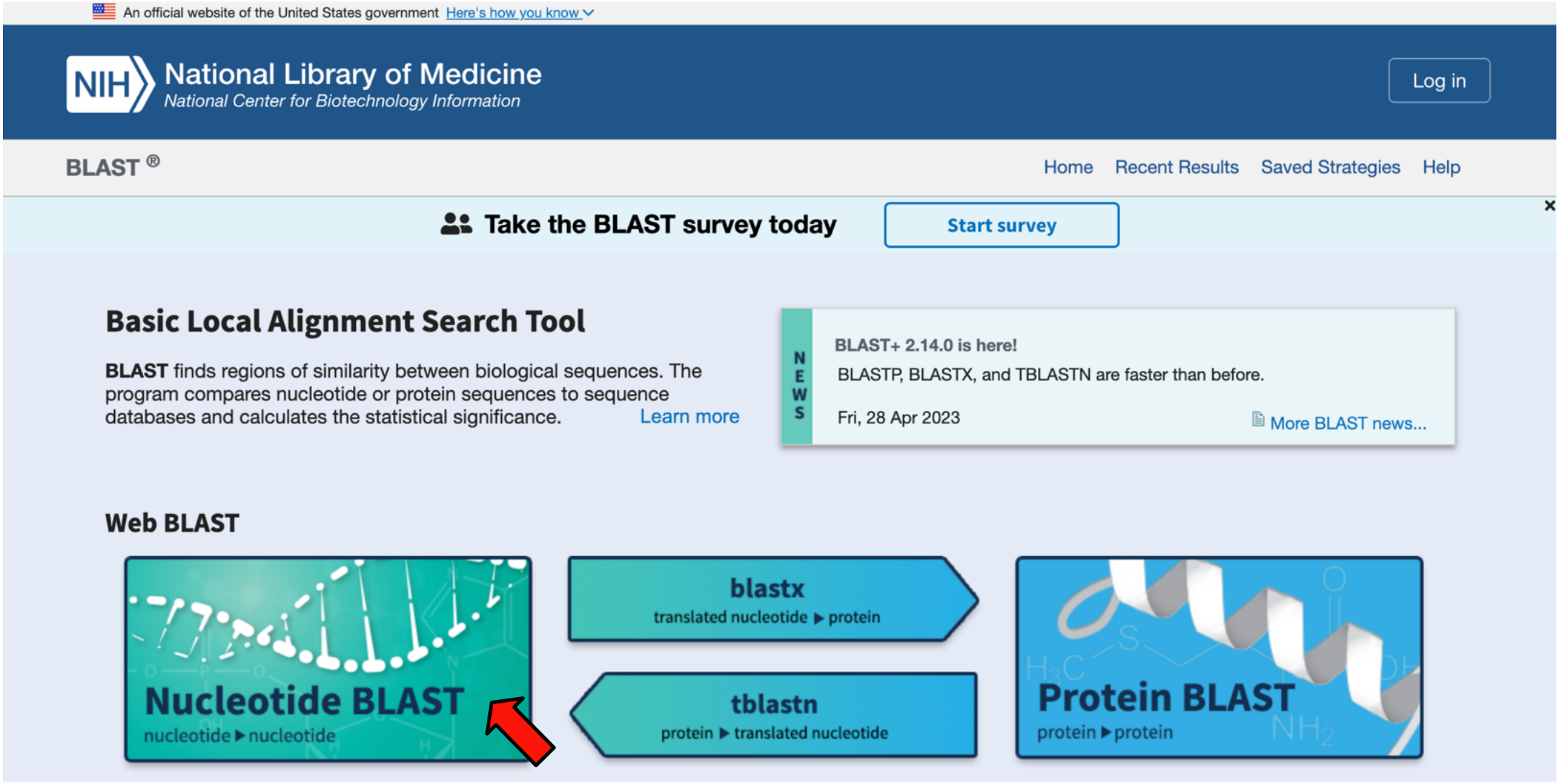
FastQC Report

Thu. 5 Oct. 2023
P90006023_RSVA_Illumina_S1_L001_R1_001.fastq.gz

Summary

Overrepresented sequences

Sequence	Count	Percentage	Possible Source
NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN	422	0.44857349377099365	No Hit
TTCATAAACAATCCTGCAAAAATCCCTTCAACTCTACTGCCACCTCTGGT	193	0.2051532803265445	No Hit
GTGCATGGGGTGGCCATTCAAAATAATTGTGACTAAAATGACACCTTTTA	184	0.19558654704706835	No Hit
CTCTTTATACAGTAATTTAAGACTATTTAATGCTAGCGTATAATCAGCTT	160	0.17007525830179854	No Hit
CTACAAACTTGCGTAAACCAAAAAAATGGGGCAAATAAGAATTTGATAAG	155	0.16476040647986734	No Hit
GTTATATGCTATGTCTAGATTAGGAAGAGAAGACACCATAAAAATACTCA	148	0.15731961392916366	No Hit
GTGTAGTTCTGATGTTTGTTTTGGTGGTGTTGATGGTTGGCTTTCCTGTA	145	0.15413070283600494	No Hit
GAACTACACTGCTCACCTTCAACACCAAAGGAAATCCAGAACAAACAAGT	135	0.14350099919214254	No Hit
CTTCTATACTGCAAGGCCAGAAGCACACCAGTCACATTAAGTAAGGATCA	134	0.1424380288277563	No Hit
GCATATAACATACCTATTAACCCAGTGAATTTATGATTAGCATCTTCTGT	126	0.13393426591266636	No Hit
GACTCAGGAATGAAGAAAGTGAAAAGATGGCAAAAGACACATCAGATGAA	126	0.13393426591266636	No Hit
CTAGTATACCAACCAGTTCTTAGAGCACTAAGATAGCCTTTGCTAACTGC	115	0.1222415919044177	No Hit
GGTTAAGACACTGACTCCATTAGATAAGCTGACTACAGCCTTGTTTGTGG	114	0.12117862154003146	No Hit
GTATAGAAGCAGTCCAACTGCAATTAATGCTAACAATATTACTATAATTA	108	0.11480079935371401	No Hit
GTATACTAGTGTTATAACTATAGAATTAAGTAATATCAAGGAAAATAAGT	106	0.11267485862494155	No Hit
ATCTAAGCCCATCACAAATCTATACAACATCCGAGTACCTATCACAATCT	105	0.1116118882605553	No Hit
GGTTAGAGATGGGCAGCAACTCATTGAGTATGATAAAAGTTAGATTGCAA	104	0.11054891789616907	No Hit
GCTCTCAACCTAATGGCCTAATAGATGACAATTGTGAAATTAAATTCTCC	100	0.10629703643862409	No Hit
GTATTTGTGCATGTTATTACAAGTAGTGATATTTGCCCTAATAATAATAT	99	0.10523406607423785	No Hit


An official website of the United States government Here's how you know
NIH National Library of Medicine
National Center for Biotechnology Information
Log in
BLAST ®
Home Recent Results Saved Strategies Help
Take the BLAST survey today
Start survey
Basic Local Alignment Search Tool
BLAST finds regions of similarity between biological sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance. Learn more
NEWS
BLAST+ 2.14.0 is here!
BLASTP, BLASTX, and TBLASTN are faster than before.
Fri, 28 Apr 2023
More BLAST news...
Web BLAST
Nucleotide BLAST
nucleotide ▶ nucleotide
blastx
translated nucleotide ▶ protein
tblastn
protein ▶ translated nucleotide
Protein BLAST
protein ▶ protein

An official website of the United States government Here's how you know

National Library of Medicine
National Center for Biotechnology Information

Log in

BLAST ® » blastn suite

Home Recent Results Saved Strategies Help

Check out the ClusteredNR database on BLAST+ Learn more Give us feedback

Standard Nucleotide BLAST

blastn blastp blastx tblastn tblastx

BLASTN programs search nucleotide databases using a nucleotide query. more...

Reset page Bookmark

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

TTCATAAACAATCCTGCAAAAATCCCTTCAACTCTACTGCCACCTCTGGT

Query subrange

From

To

Or, upload file Choose file No file chosen

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database: Standard databases (nr etc.): rRNA/ITS databases Genomic + transcript databases Betacoronavirus

New Experimental databases

Try experimental taxonomic nt databases Download

For more info see What are taxonomic nt databases?

Nucleotide collection (nr/nt)

Organism Optional

Enter organism name or id--completions will be suggested exclude Add organism

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Feedback

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

TTCATAAACAATCCTGCAAAAATCCCTTCAACTCTACTGCCACCTCTGGT

Query subrange

From

To

Or, upload file Choose file No file chosen

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database Standard databases (nr etc.): rRNA/ITS databases Genomic + transcript databases Betacoronavirus

New Experimental databases **Try experimental taxonomic nt databases** Download

For more info see What are taxonomic nt databases?

Nucleotide collection (nr/nt)

Organism Optional

Enter organism name or id--completions will be suggested exclude Add organism

Enter organism common name, binomial, or tax id. Only 20 top taxa will be shown

Exclude Optional Models (XM/XP) Uncultured/environmental sample sequences

Limit to Optional Sequences from type material

Entrez Query Optional YouTube Create custom database

Enter an Entrez query to limit search

Program Selection

Optimize for

- Highly similar sequences (megablast)
- More dissimilar sequences (discontiguous megablast)
- Somewhat similar sequences (blastn)

Choose a BLAST algorithm

Feedback

BLAST Search database Nucleotide collection (nr/nt) using Megablast (Optimize for highly similar sequences)

Descriptions | Graphic Summary | Alignments | Taxonomy

Sequences producing significant alignments

Download | Select columns | Show 100

select all *100 sequences selected*

GenBank | Graphics | Distance tree of results | MSA Viewer

Description	Scientific Name	Max Score	Total Score	Query Cover	E value	Per. Ident	Acc. Len	Accession
Human respiratory syncytial virus A strain RSVA/Australia/VR26K_S102/1956, partial genome	Human respiratory syncytial virus A	93.5	93.5	100%	2e-15	100.00%	14670	OQ941776.1
Human respiratory syncytial virus A strain RSVA/Australia/648_16R_S97/2016, partial genome	Human respiratory syncytial virus A	93.5	93.5	100%	2e-15	100.00%	14806	OQ941775.1
Human respiratory syncytial virus A strain RSVA/Australia/7545B_S6/2018, partial genome	Human respiratory syncytial virus A	93.5	93.5	100%	2e-15	100.00%	14740	OQ941773.1
Human respiratory syncytial virus A strain RSVA/Australia/VR26G_S114/1956, partial genome	Human respiratory syncytial virus A	93.5	93.5	100%	2e-15	100.00%	14790	OQ941771.1
Human respiratory syncytial virus A strain RSVA/Australia/VR26E_S110/1956, partial genome	Human respiratory syncytial virus A	93.5	93.5	100%	2e-15	100.00%	14773	OQ941770.1
Human respiratory syncytial virus A strain RSVA/Australia/98_S113/2018, partial genome	Human respiratory syncytial virus A	93.5	93.5	100%	2e-15	100.00%	14768	OQ941768.1
Human respiratory syncytial virus A strain RSVA/Australia/97_S124/2018, partial genome	Human respiratory syncytial virus A	93.5	93.5	100%	2e-15	100.00%	14769	OQ941767.1
Human respiratory syncytial virus A strain RSVA/Australia/94_S52/2018, partial genome	Human respiratory syncytial virus A	93.5	93.5	100%	2e-15	100.00%	14768	OQ941765.1
Human respiratory syncytial virus A strain RSVA/Australia/93_S34/2018, partial genome	Human respiratory syncytial virus A	93.5	93.5	100%	2e-15	100.00%	14777	OQ941764.1
Human respiratory syncytial virus A strain RSVA/Australia/92_S57/2018, partial genome	Human respiratory syncytial virus A	93.5	93.5	100%	2e-15	100.00%	14775	OQ941763.1
Human respiratory syncytial virus A strain RSVA/Australia/8_S85/2018, partial genome	Human respiratory syncytial virus A	93.5	93.5	100%	2e-15	100.00%	14777	OQ941762.1
Human respiratory syncytial virus A strain RSVA/Australia/87_S94/2018, partial genome	Human respiratory syncytial virus A	93.5	93.5	100%	2e-15	100.00%	14700	OQ941761.1
Human respiratory syncytial virus A strain RSVA/Australia/82_S12/2018, partial genome	Human respiratory syncytial virus A	93.5	93.5	100%	2e-15	100.00%	14777	OQ941759.1
Human respiratory syncytial virus A strain RSVA/Australia/7545C_S47/2018, partial genome	Human respiratory syncytial virus A	93.5	93.5	100%	2e-15	100.00%	14779	OQ941758.1
Human respiratory syncytial virus A strain RSVA/Australia/68_S46/2017, partial genome	Human respiratory syncytial virus A	93.5	93.5	100%	2e-15	100.00%	14945	OQ941756.1
Human respiratory syncytial virus A strain RSVA/Australia/55_S17/2016, partial genome	Human respiratory syncytial virus A	93.5	93.5	100%	2e-15	100.00%	14782	OQ941752.1
Human respiratory syncytial virus A strain RSVA/Australia/544_18R_S126/2018, partial genome	Human respiratory syncytial virus A	93.5	93.5	100%	2e-15	100.00%	14822	OQ941751.1
Human respiratory syncytial virus A strain RSVA/Australia/215_S100/2018, partial genome	Human respiratory syncytial virus A	93.5	93.5	100%	2e-15	100.00%	14771	OQ941750.1
Human respiratory syncytial virus A strain RSVA/Australia/214_S71/2017, partial genome	Human respiratory syncytial virus A	93.5	93.5	100%	2e-15	100.00%	14844	OQ941749.1

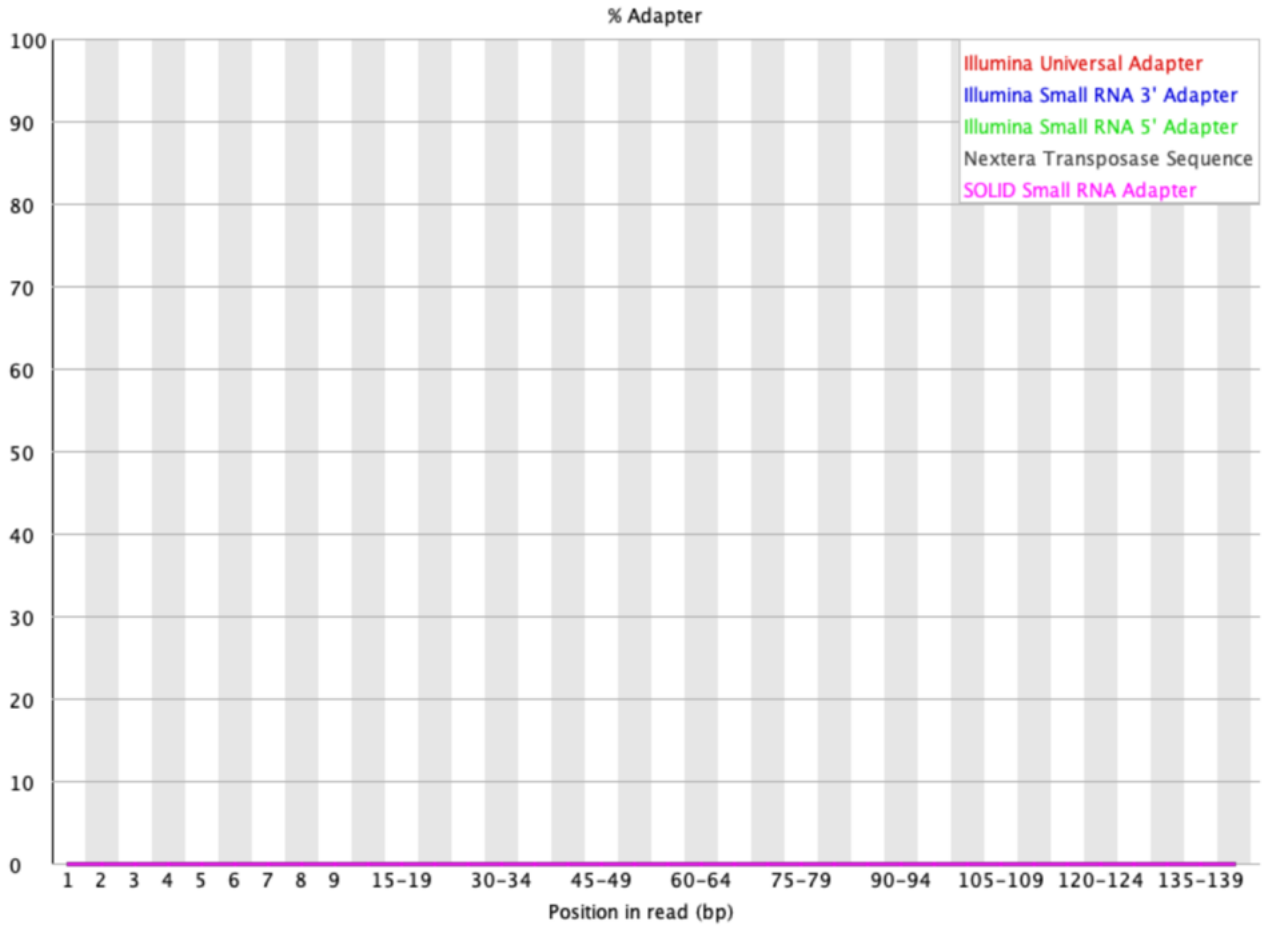
Feedback

FastQC Report

Thu. 5 Oct. 2023
P90006023_RSVA_Illumina_S1_L001_R1_001.fastq.gz

Summary

Adapter Content



Thank you !

clyde.dapat@influenzacentre.org

WHO Collaborating Centre for Reference and Research on Influenza
VIDRL

Doherty Institute
THE UNIVERSITY OF MELBOURNE
The Royal Melbourne Hospital

A joint venture between The University of Melbourne and The Royal Melbourne Hospital