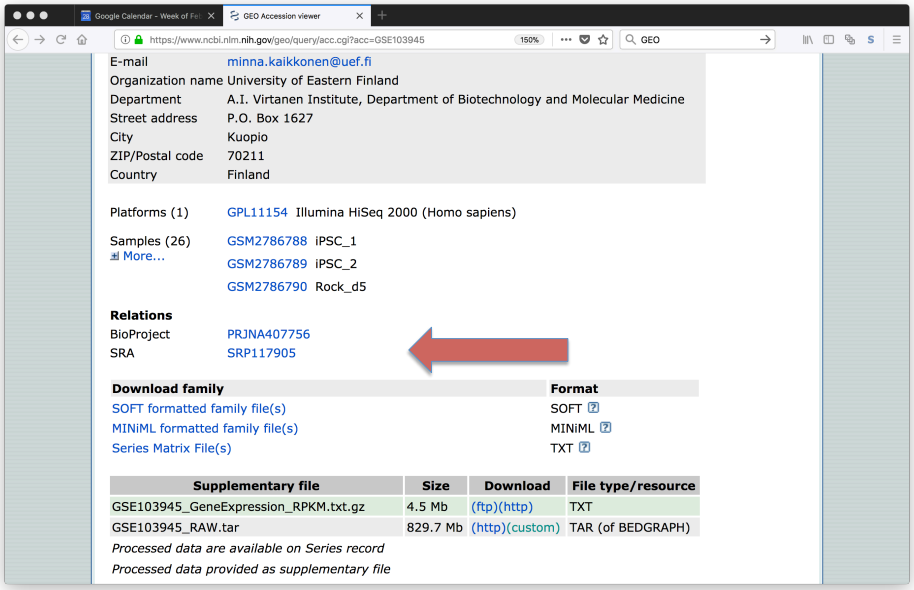
Project Management

Obtain data from GEO or SRA



E-mail	minna.kaikkonen@uef.fi
Organization name	University of Eastern Finland
Department	A.I. Virtanen Institute, Department of Biotechnology and Molecular Medicine
Street address	P.O. Box 1627
City	Kuopio
ZIP/Postal code	70211
Country	Finland

Platforms (1) GPL11154 Illumina HiSeq 2000 (Homo sapiens)

Samples (26) GSM2786788 iPSC_1
More... GSM2786789 iPSC_2
GSM2786790 Rock_d5

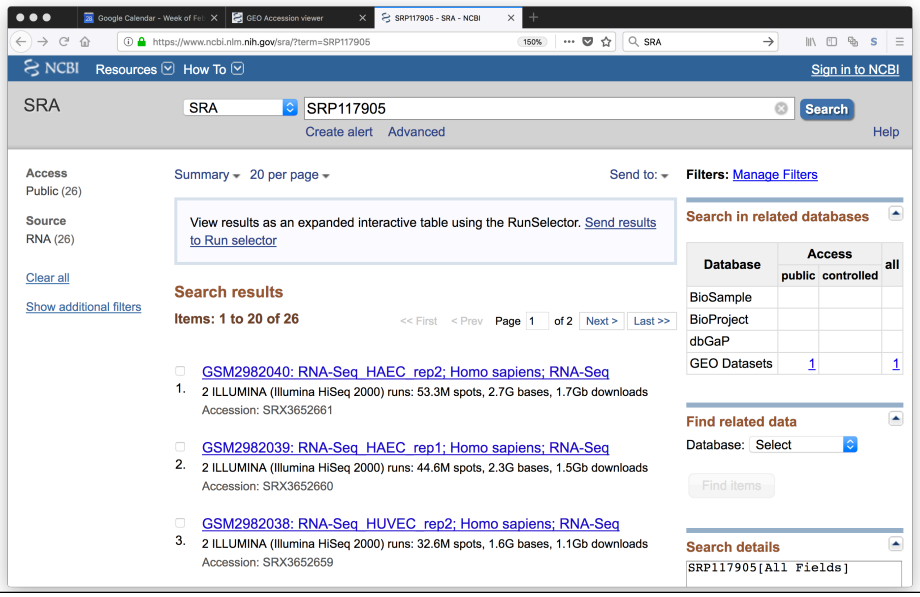
Relations
BioProject PRJNA407756
SRA SRP117905

Download family	Format
SOFT formatted family file(s)	SOFT
MINiML formatted family file(s)	MINiML
Series Matrix File(s)	TXT

Supplementary file	Size	Download	File type/resource
GSE103945_GeneExpression_RPKM.txt.gz	4.5 Mb	(ftp)(http)	TXT
GSE103945_RAW.tar	829.7 Mb	(http)(custom)	TAR (of BEDGRAPH)

Processed data are available on Series record
Processed data provided as supplementary file

Individual experiment information



NCBI Resources How To Sign in to NCBI

SRA SRA SRP117905 Search
Create alert Advanced Help

Access
Public (26)

Source
RNA (26)

Clear all
Show additional filters

Summary 20 per page Send to: Filters: Manage Filters

View results as an expanded interactive table using the RunSelector. Send results to Run selector

Search results
Items: 1 to 20 of 26 << First < Prev Page 1 of 2 Next > Last >>

1. GSM2982040: RNA-Seq_HAEC_rep2; Homo sapiens; RNA-Seq
2 ILLUMINA (Illumina HiSeq 2000) runs: 53.3M spots, 2.7G bases, 1.7Gb downloads
Accession: SRX3652661

2. GSM2982039: RNA-Seq_HAEC_rep1; Homo sapiens; RNA-Seq
2 ILLUMINA (Illumina HiSeq 2000) runs: 44.6M spots, 2.3G bases, 1.5Gb downloads
Accession: SRX3652660

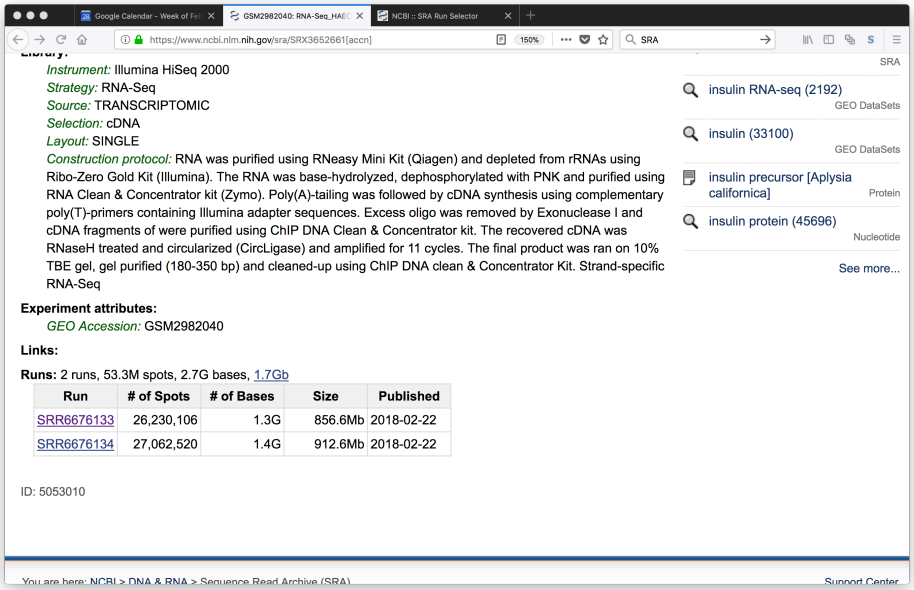
3. GSM2982038: RNA-Seq_HUVEC_rep2; Homo sapiens; RNA-Seq
2 ILLUMINA (Illumina HiSeq 2000) runs: 32.6M spots, 1.6G bases, 1.1Gb downloads
Accession: SRX3652659

Search in related databases

Database	Access public	Access controlled	all
BioSample			
BioProject			
dbGaP			
GEO Datasets	1		1

Find related data
Database: Select
Find items

Search details
SRP117905[All Fields]



Instrument: Illumina HiSeq 2000
Strategy: RNA-Seq
Source: TRANSCRIPTOMIC
Selection: cDNA
Layout: SINGLE
Construction protocol: RNA was purified using RNeasy Mini Kit (Qiagen) and depleted from rRNAs using Ribo-Zero Gold Kit (Illumina). The RNA was base-hydrolyzed, dephosphorylated with PNK and purified using RNA Clean & Concentrator kit (Zymo). Poly(A)-tailing was followed by cDNA synthesis using complementary poly(T)-primers containing Illumina adapter sequences. Excess oligo was removed by Exonuclease I and cDNA fragments of were purified using ChIP DNA Clean & Concentrator kit. The recovered cDNA was RNaseH treated and circularized (CircLigase) and amplified for 11 cycles. The final product was ran on 10% TBE gel, gel purified (180-350 bp) and cleaned-up using ChIP DNA clean & Concentrator Kit. Strand-specific RNA-Seq

insulin RNA-seq (2192) GEO DataSets
insulin (33100) GEO DataSets
insulin precursor [Aplysia californica] Protein
insulin protein (45696) Nucleotide
See more...

Experiment attributes:
GEO Accession: GSM2982040

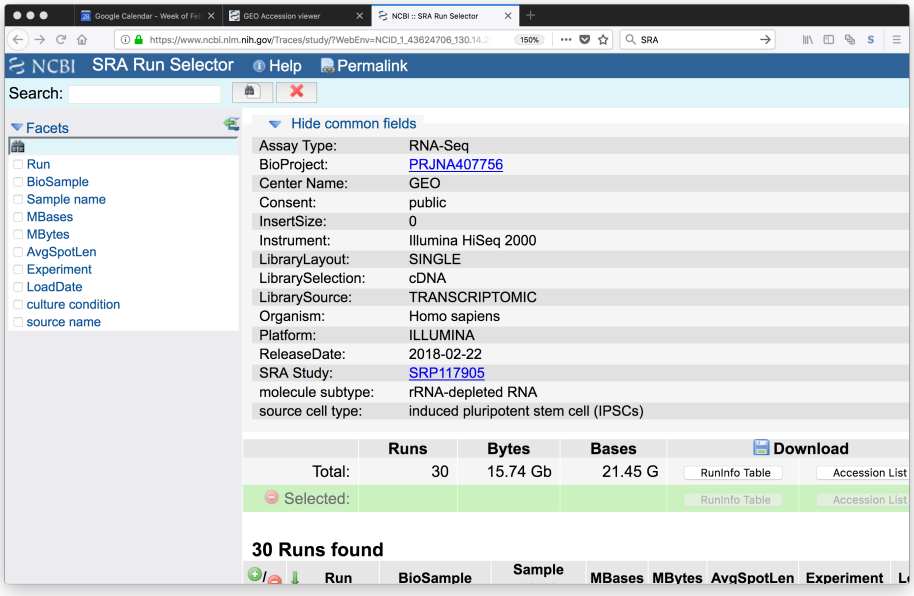
Links:

Runs: 2 runs, 53.3M spots, 2.7G bases, 1.7Gb

Run	# of Spots	# of Bases	Size	Published
SRR6676133	26,230,106	1.3G	856.6Mb	2018-02-22
SRR6676134	27,062,520	1.4G	912.6Mb	2018-02-22

ID: 5053010

OR use SRA Run Selector



NCBI SRA Run Selector Help Permalink

Search:

Facets
- Run
- BioSample
- Sample name
- MBases
- MBytes
- AvgSpotLen
- Experiment
- LoadDate
- culture condition
- source name

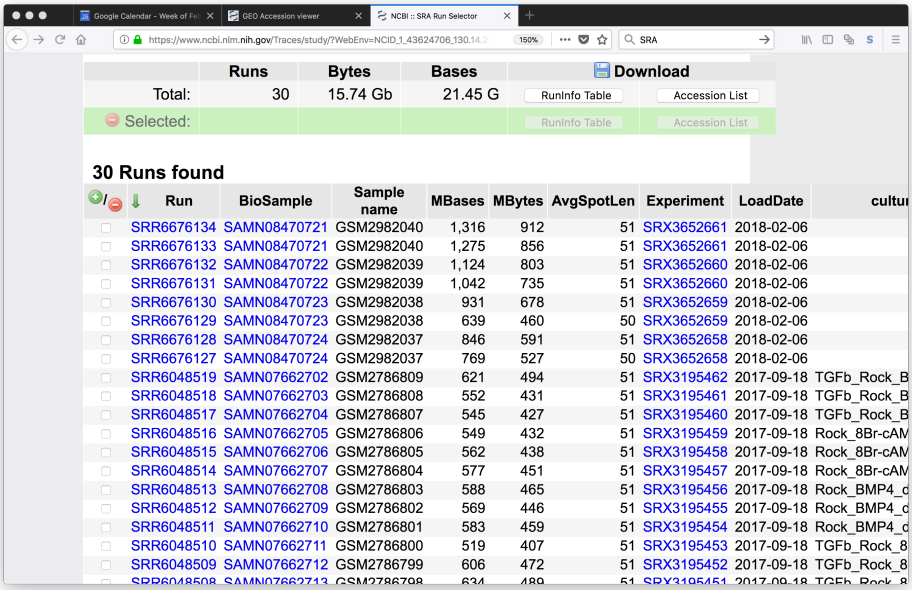
Hide common fields

Assay Type:	RNA-Seq
BioProject:	PRJNA407756
Center Name:	GEO
Consent:	public
InsertSize:	0
Instrument:	Illumina HiSeq 2000
LibraryLayout:	SINGLE
LibrarySelection:	cDNA
LibrarySource:	TRANSCRIPTOMIC
Organism:	Homo sapiens
Platform:	ILLUMINA
ReleaseDate:	2018-02-22
SRA Study:	SRP117905
molecule subtype:	rRNA-depleted RNA
source cell type:	induced pluripotent stem cell (IPSCs)

	Runs	Bytes	Bases	Download	
Total:	30	15.74 Gb	21.45 G	RunInfo Table	Accession List
Selected:				RunInfo Table	Accession List

30 Runs found

Run	BioSample	Sample	MBases	MBytes	AvgSpotLen	Experiment



	Runs	Bytes	Bases	Download	
Total:	30	15.74 Gb	21.45 G	RunInfo Table	Accession List
Selected:				RunInfo Table	Accession List

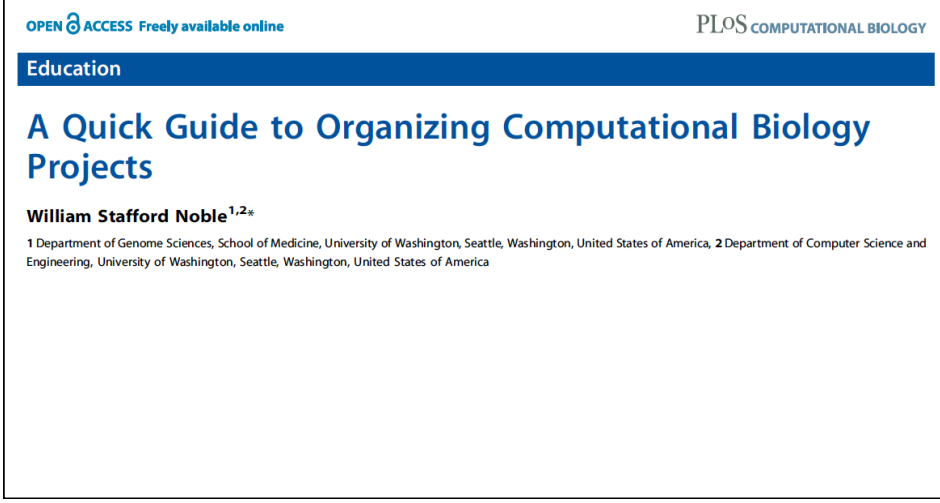
30 Runs found

Run	BioSample	Sample name	MBases	MBytes	AvgSpotLen	Experiment	LoadDate	cultu
SRR6676134	SAMN08470721	GSM2982040	1,316	912	51	SRX3652661	2018-02-06	
SRR6676133	SAMN08470721	GSM2982040	1,275	856	51	SRX3652661	2018-02-06	
SRR6676132	SAMN08470722	GSM2982039	1,124	803	51	SRX3652660	2018-02-06	
SRR6676131	SAMN08470722	GSM2982039	1,042	735	51	SRX3652660	2018-02-06	
SRR6676130	SAMN08470723	GSM2982038	931	678	51	SRX3652659	2018-02-06	
SRR6676129	SAMN08470723	GSM2982038	639	460	50	SRX3652659	2018-02-06	
SRR6676128	SAMN08470724	GSM2982037	846	591	51	SRX3652658	2018-02-06	
SRR6676127	SAMN08470724	GSM2982037	769	527	50	SRX3652658	2018-02-06	
SRR6048519	SAMN07662702	GSM2786809	621	494	51	SRX3195462	2017-09-18	TGFb_Rock_B
SRR6048518	SAMN07662703	GSM2786808	552	431	51	SRX3195461	2017-09-18	TGFb_Rock_B
SRR6048517	SAMN07662704	GSM2786807	545	427	51	SRX3195460	2017-09-18	TGFb_Rock_B
SRR6048516	SAMN07662705	GSM2786806	549	432	51	SRX3195459	2017-09-18	Rock_8Br-cAM
SRR6048515	SAMN07662706	GSM2786805	562	438	51	SRX3195458	2017-09-18	Rock_8Br-cAM
SRR6048514	SAMN07662707	GSM2786804	577	451	51	SRX3195457	2017-09-18	Rock_8Br-cAM
SRR6048513	SAMN07662708	GSM2786803	588	465	51	SRX3195456	2017-09-18	Rock_BMP4_c
SRR6048512	SAMN07662709	GSM2786802	569	446	51	SRX3195455	2017-09-18	Rock_BMP4_c
SRR6048511	SAMN07662710	GSM2786801	583	459	51	SRX3195454	2017-09-18	Rock_BMP4_c
SRR6048510	SAMN07662711	GSM2786800	519	407	51	SRX3195453	2017-09-18	TGFb_Rock_8
SRR6048509	SAMN07662712	GSM2786799	606	472	51	SRX3195452	2017-09-18	TGFb_Rock_8
SRR6048508	SAMN07662713	GSM2786798	634	489	51	SRX3195451	2017-09-18	TGFb_Rock_8

On Fiji:

- module load sra/2.8.0
- fastq-dump -O $outdir -split-3 $var
  - spit-3 : 3-file splitting for mate-pairs: First biological reads satisfying dumping conditions are placed in files *_1.fastq and *_2.fastq If only one biological read is present it is placed in *.fastq

Organizing and documenting a project (i.e. structured directories)



OPEN ACCESS Freely available online

PLOS COMPUTATIONAL BIOLOGY

Education

A Quick Guide to Organizing Computational Biology Projects

William Stafford Noble[1,2]*

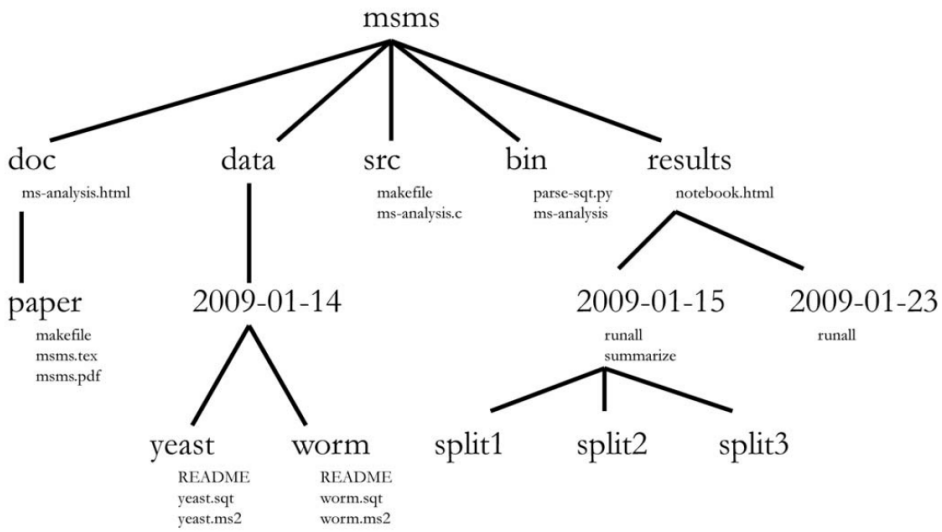
1 Department of Genome Sciences, School of Medicine, University of Washington, Seattle, Washington, United States of America, 2 Department of Computer Science and Engineering, University of Washington, Seattle, Washington, United States of America

Basic rules

- Never edit primary data
- Always backup primary data (home directory)
- Use version control whenever possible
- Use standardized file formats
- Do not reinvent the wheel

- Ideally you document for the you of years from now. Practically – you are documenting so that I can understand your files.

Have an organized directory structure



Digital lab notebooks

- Chronological diary of what you were doing with motivation, results, interpretation, etc.
- README files (what did I do)
- *Avoid editing intermediate files*
- Cross check results
- Avoid command line – use scripts, but make them self documenting.

Jupyter Notebooks, other online tools

Version Control

https://betterexplained.com/articles/a-visual-guide-to-version-control/

Table 1.1. Three Generations of Version Control

Generation	Networking	Operations	Concurrency	Examples
First	None	One file at a time	Locks	RCS, SCCS
Second	Centralized	Multi-file	Merge before commit	CVS, SourceSafe, Subversion, Team Foundation Server
Third	Distributed	Changesets	Commit before merge	Bazaar, Git, Mercurial

Git Hub, Sourceforge

Creating the project proposal …

- DUE March 14th
- You will outline the planned project:
  - what questions are you asking?,
  - How will you answer them?,
  - What software will you use?
  - You will PROVE the data you will use is of sufficient quality.

Example:

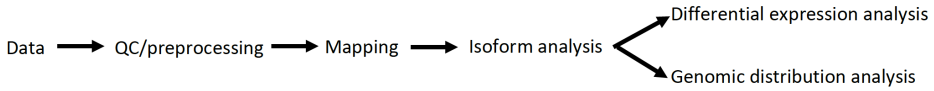
Project Question

I will investigate major differences in the data obtained with three different methods for measuring active transcription: GRO-Seq, NET-Seq, and ChrRNA-Seq. I will address at least two specific questions:

1) Is there differential expression between different methods at the same genes/genomic regions?
2) Is there a similar proportion of reads in coding regions vs. intronic regions vs. intergenic regions? Is there a pattern to differences?

My ultimate goal in this project is to critically examine the information content and biases present in these techniques in order to determine what might be the best technique to use in my future research. If the data seem to present additional methods of comparison than the above two questions, I intend to pursue them.

Analysis Workflow
The workflow is as follows:



QC/preprocessing steps will be performed with fastQC and Cutadapt.
Mapping will be performed with Tophat.
Isoform analysis will be attempted with Cufflinks, although I'm not sure that all of this data will be compatible with isoform analysis.
Differential expression will be done with Cufflinks, Cuffdiff, or DESeq.
Comparison of mapped reads to annotation will be done with HTSeq or BEDtools.

Datasets
I have six datasets for use in my analysis (biological duplicates for three different experiments). All are 2x50 paired-end sequencing runs from untreated HeLa or HeLa S3 cells. All were sequenced by Illumina machines.
Further details are below:

GRO-Seq (Laitem *et al. Nat. Struct. Mol. Biol.* **22,** 396–403 (2015))
ArrayExpress accession #E-MTAB-3360 (includes pair-separated .fastq files for two untreated replicates)
2x50 base sequencing runs from a HiSeq

I ran these two datasets through fastQC. The data are all of good quality, with heavy adapter contamination present throughout the reads. I used Cutadapt to trim reads for replicate 1, read direction 1 and obtained a much cleaner dataset, although I may still need to optimize trimming (Figure 3).

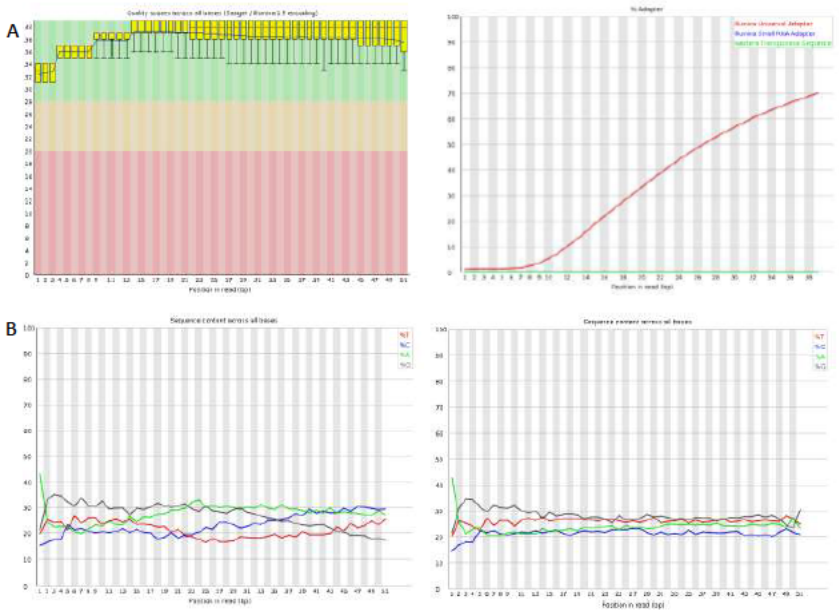


Figure 3: FastQC metrics for GRO-Seq data. A) Graphs showing base quality distributions and distribution of adapter contamination in sample. B) Per-position base composition of reads before (left) and after (right) read trimming for one adapter sequence.

Project proposal

- You will describe the questions you intend to answer using your dataset.
- You will describe the typical analysis workflow that is necessary to address the questions you propose. At each step of the workflow, you will list the software packages you intend to utilize at each step.
- You will describe the dataset you will be working with (where did you get it, what protocol produced it).
- You will provide adequate evidence that the dataset you intend to use is of good quality.

A few words about R (in preparation for next week)

- R studio:

  https://fiji-viz.colorado.edu/rstudio/auth-sign-in

- Command line R:

  > Module load R/3.3.0

- Commonly used bioinformatics packages in R:

  source("https://bioconductor.org/biocLite.R")

  biocLite()