Leidos Biomedical Research, Inc.

Workflows for RNA-Seq analysis

Best practices on managing, analyzing and visualizing RNA-Seq data
A perspective from the CCR Bioinformatics Core

Parthav Jailwala
CCR Bioinformatics Core

What is the CCR Informatics core ?

- Collaborative resource providing bioinformatics analysis assistance to CCR investigators

- 138 requests concluded across 91 labs
  - ✓29 requests concluded for RNA-Seq data analysis

- Workflow/pipeline development

- Education and training

Value of working with CCRIFX

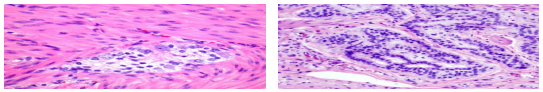
- Understand research question & suggest viable experimental designs
- Guide on use of technologies
- Assist in data tracking and transfer of large datasets
- Perform analysis and explain results
- Deliver figures and methods for manuscripts, presentations, posters
- Maximize research impact of your invested effort

Outline

- Computational aspects of RNA-Seq data analysis
  - Best practices
- Example CCRIFX pipeline for RNA-Seq analysis:
  - Pre-alignment QC
  - Tuxedo RNA-Seq pipeline
  - Post alignment QC
- Visuals of FastQC plots (good vs bad data)
- Pointers

RNA-Seq : A 5-step process

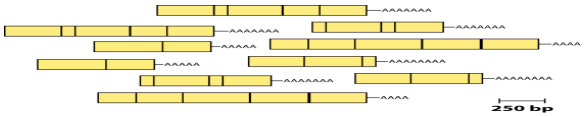
Design Experiment



Condition 1 (normal) | Condition 2 (tumor)

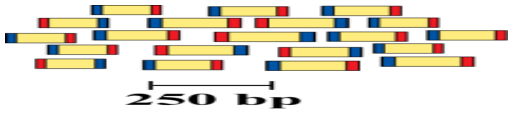
Primary Objective
Replication & Sequencing depth
Samples of interest

Purify RNA



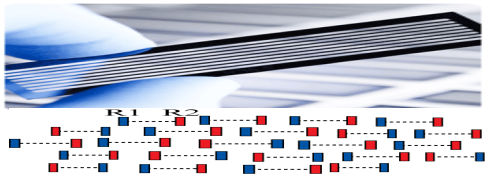
Isolate RNAs
Target enrichment
RNA fragmentation

Prepare Libraries



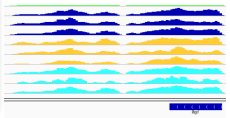
Generate cDNA, fragment, size select, add primers, adapters

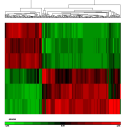
Sequence



Sequence ends

Analysis





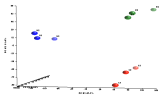


Image analysis, Base calling, QC, Mapping, DGE analysis, visualization and Biological interpretation

Features of RNA-Seq data

- Very large volume of raw and processed data
- Variety of file formats with non-standard naming conventions
- Analysis requires specialized software tools

Computational aspects of RNA-Seq analysis

- **Data Management Plan**
  - Data retrieval and short-term storage
  - Data backup and archival storage
  - Standardized nomenclature (file and directory naming)
- **Quality control of sequence reads & alignments**
  - Source: Low-quality RNA, library construction biases, sequencing chemistry
  - Solution: Trimming and/or filtering reads before alignment
- **Primary data analysis**
  - Pre-alignment QC
  - Alignment to the transcriptome
  - Post-alignment QC
  - Quantifying transcript levels & Differential expression detection
- **Data visualization and Biological interpretation**

Methods Mol Biol. 2012;883:201-19

Why automate an RNA-Seq workflow ?

- **Standardizes and codifies the procedure**
- **Increases consistency and robustness of the analysis across multiple samples or experiments**
- **Facilitates evaluation of changes and simplifies debugging of anomalies and artifacts**
- **Minimizes human error**

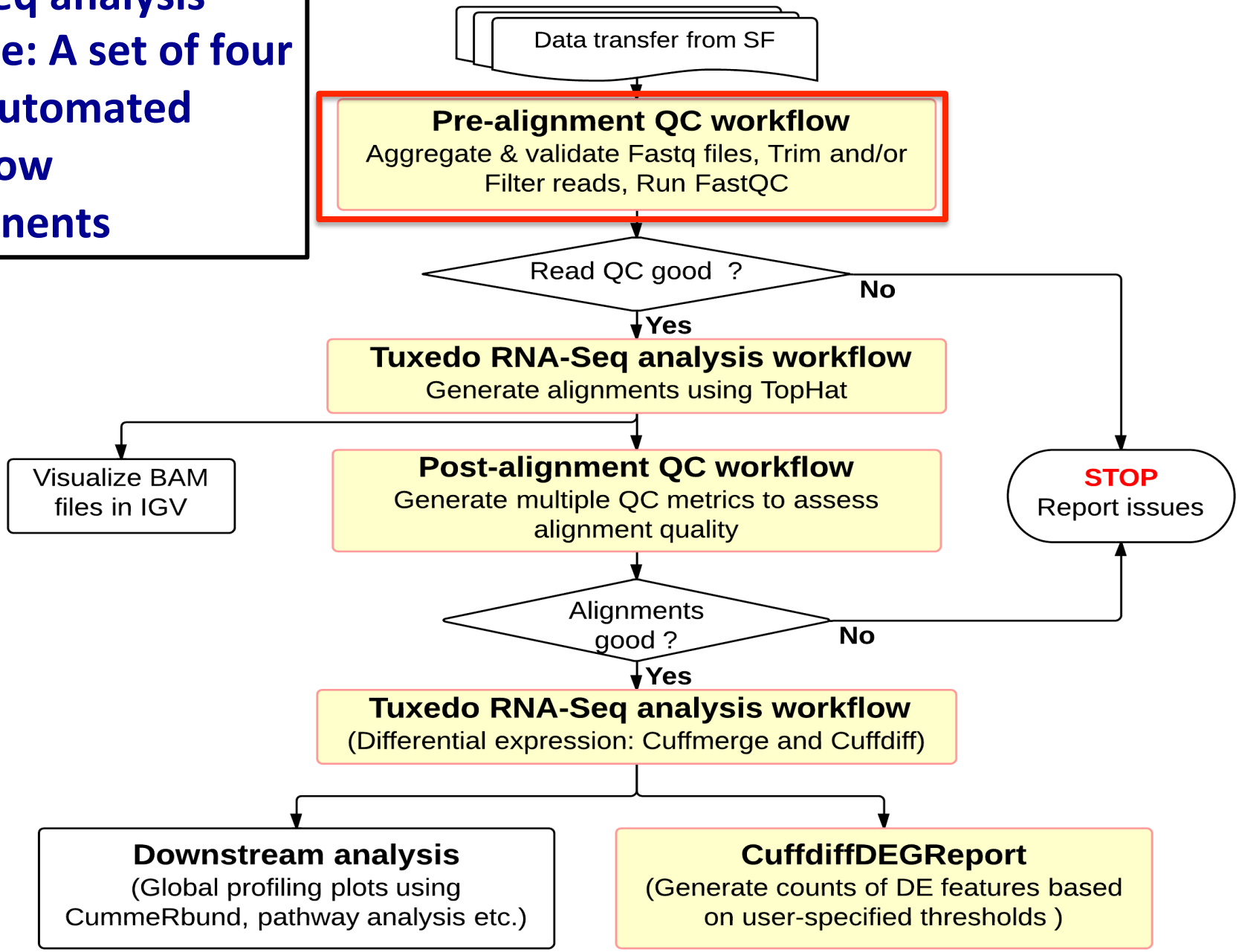
A large RNA-Seq project that makes use of workflows

- Transcriptomes in mouse embryos & embryonic stem cells
- 39 samples: 13 conditions x 3 biological reps/condition
- HiSeq2000, 101bp PE reads, Avg. 161M reads/sample
- 10 differential expression comparisons
- 4 TB of raw fastq files, 11 TB of analysis files

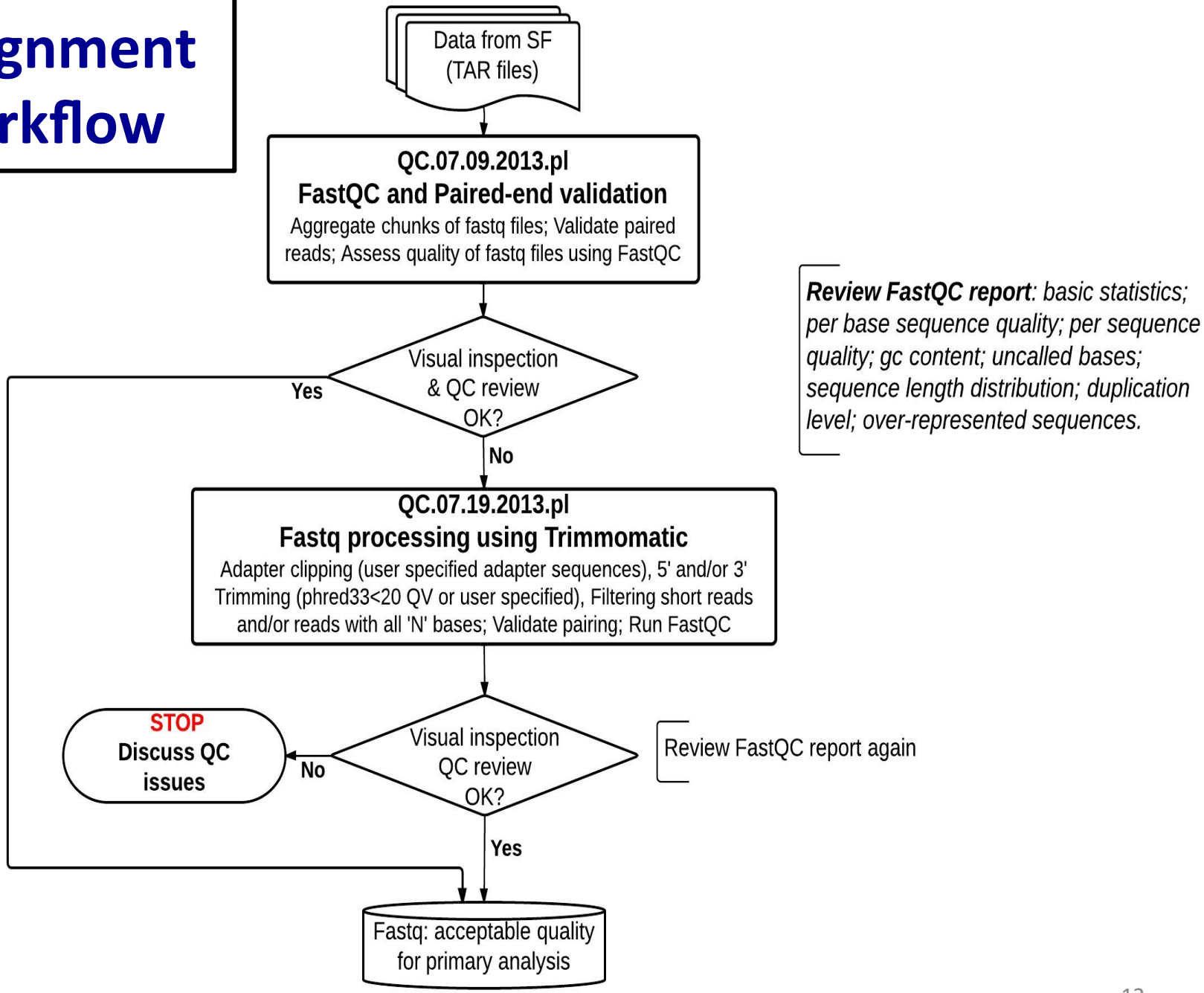
Outline

- Computational aspects of RNA-Seq data analysis
  - Best practices
- Example CCRIFX pipeline for RNA-Seq analysis:
  - Pre-alignment QC
  - Tuxedo RNA-Seq pipeline
  - Post alignment QC
- Visuals of FastQC plots (good vs bad data)
- Pointers

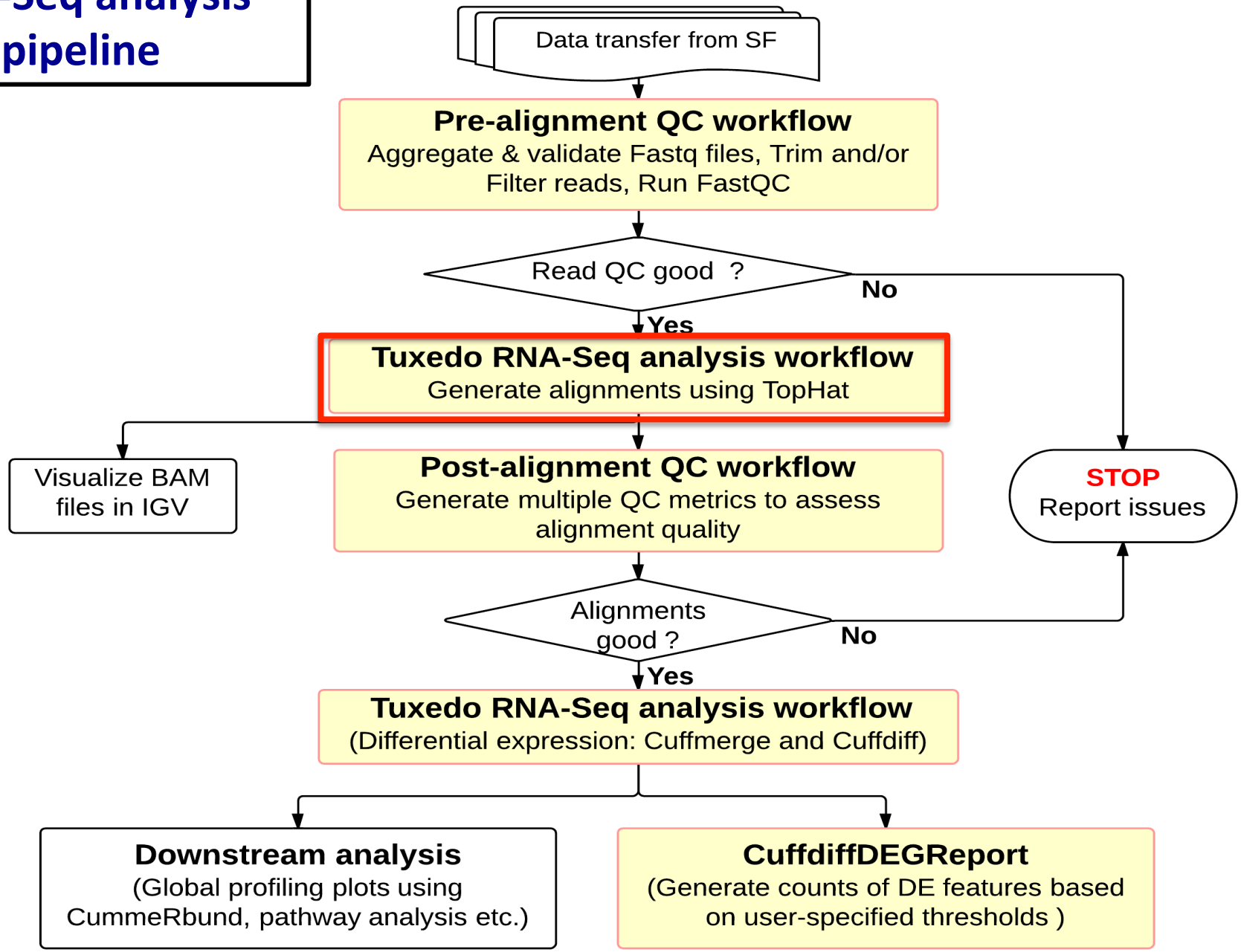
RNA-Seq analysis pipeline: A set of four semi-automated workflow components



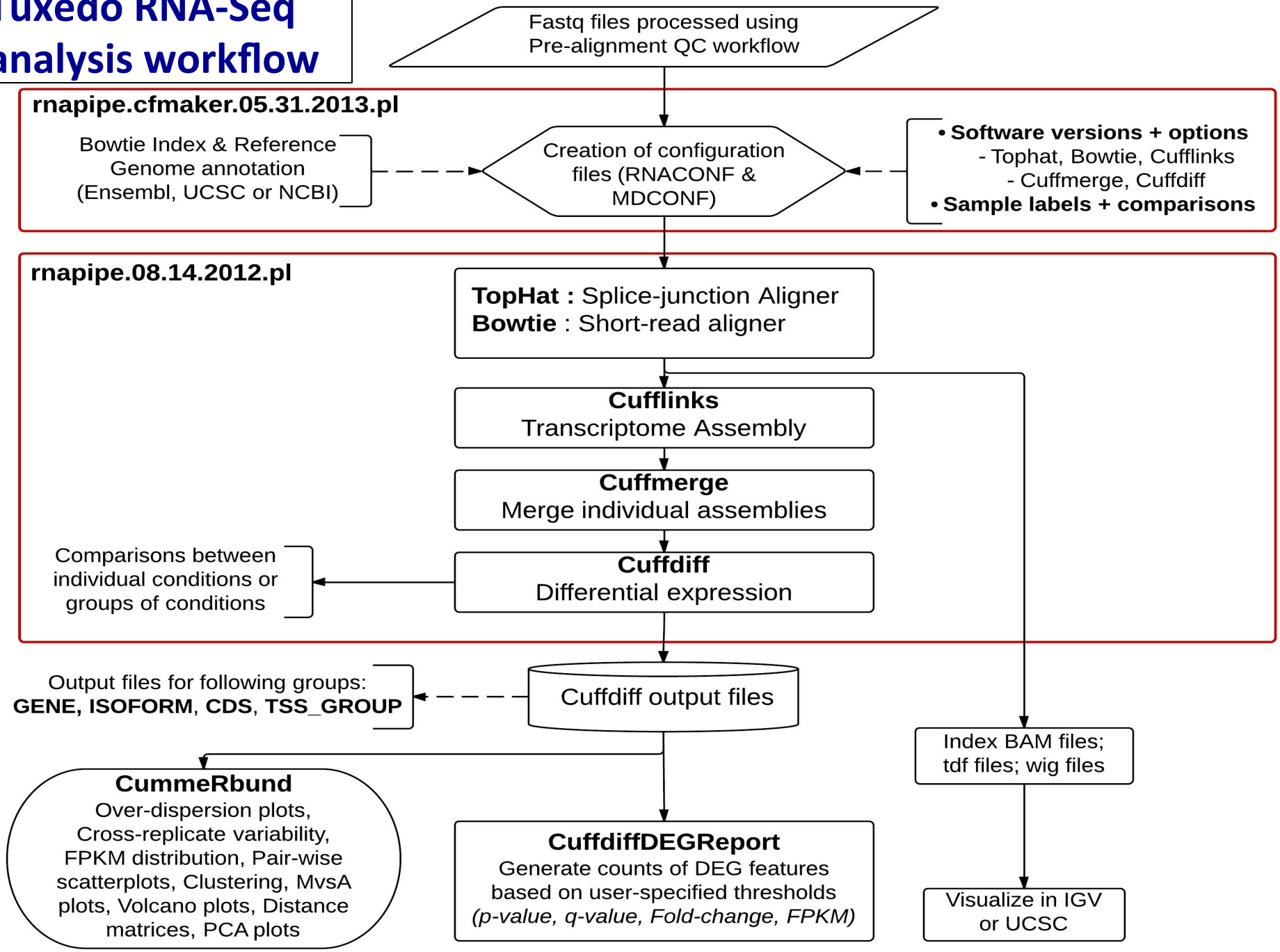
Pre-alignment QC workflow



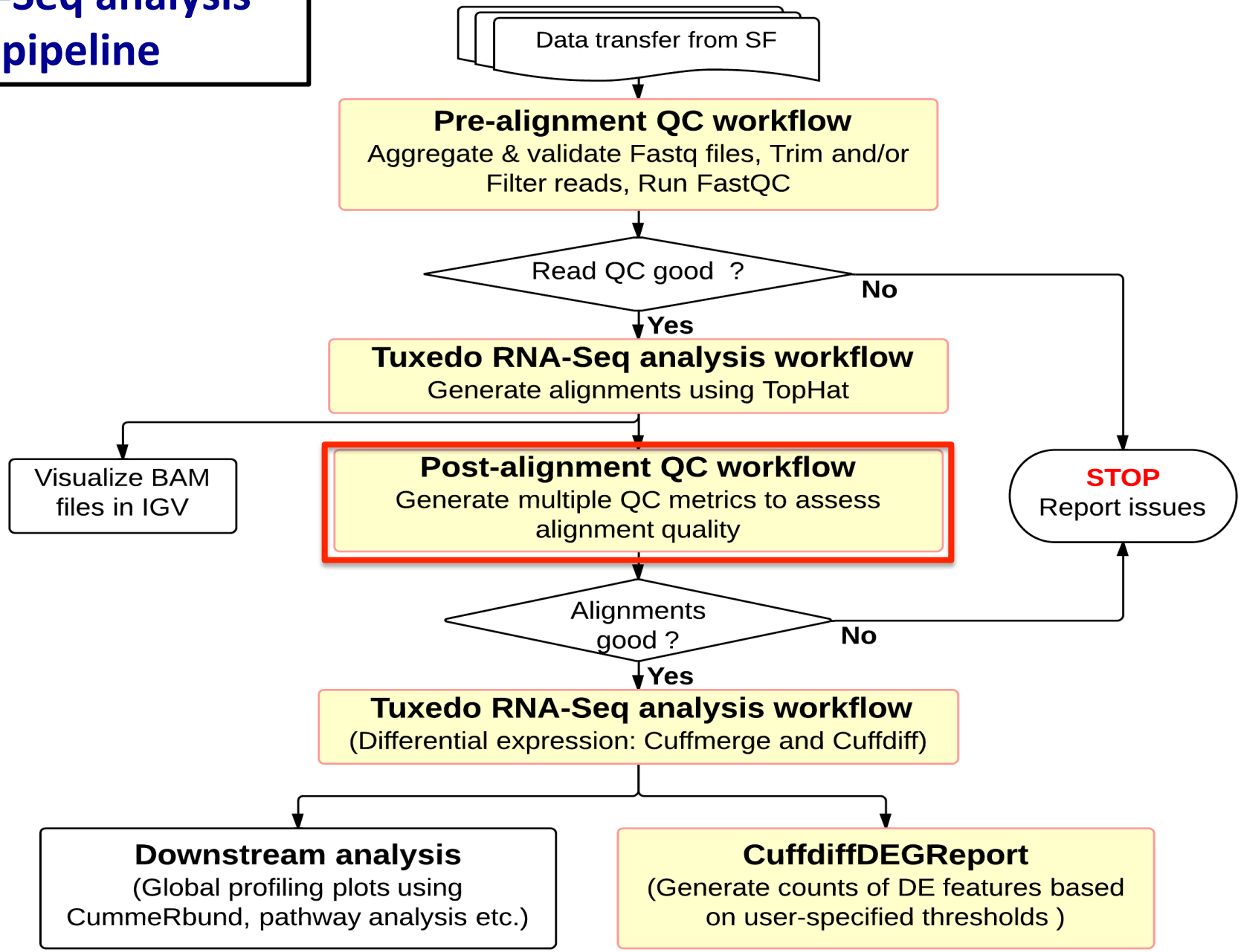
RNA-Seq analysis pipeline



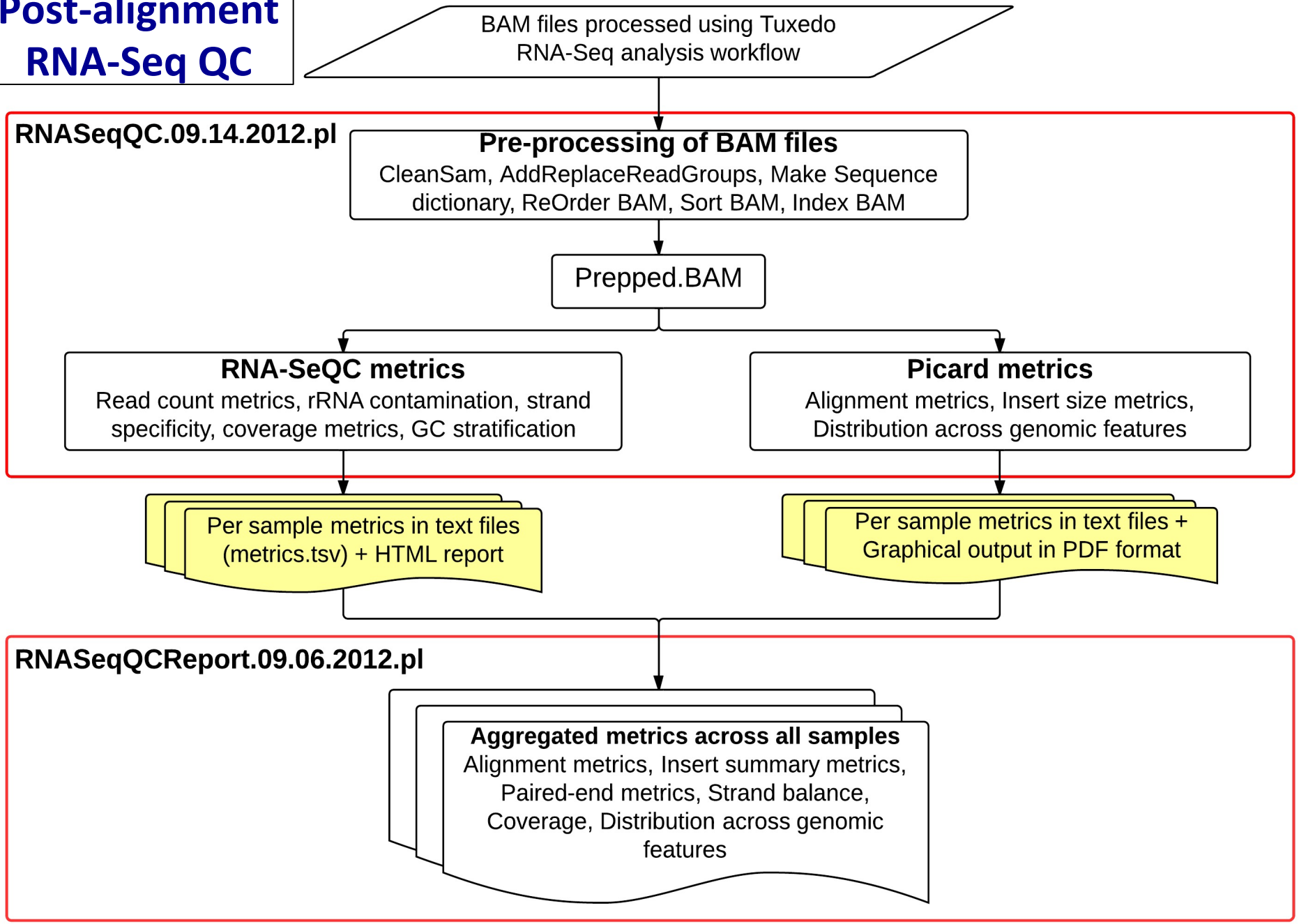
Tuxedo RNA-Seq analysis workflow



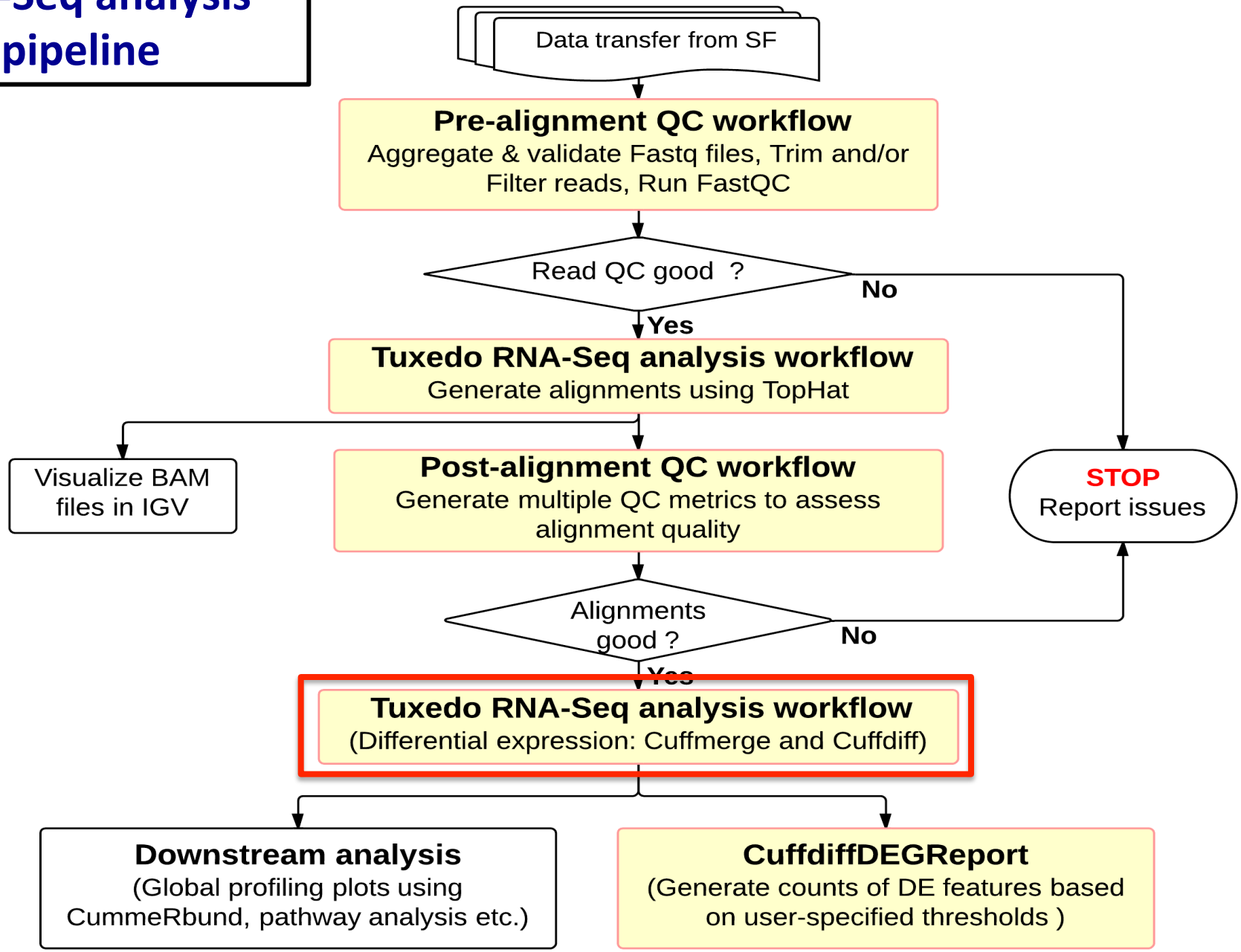
RNA-Seq analysis pipeline



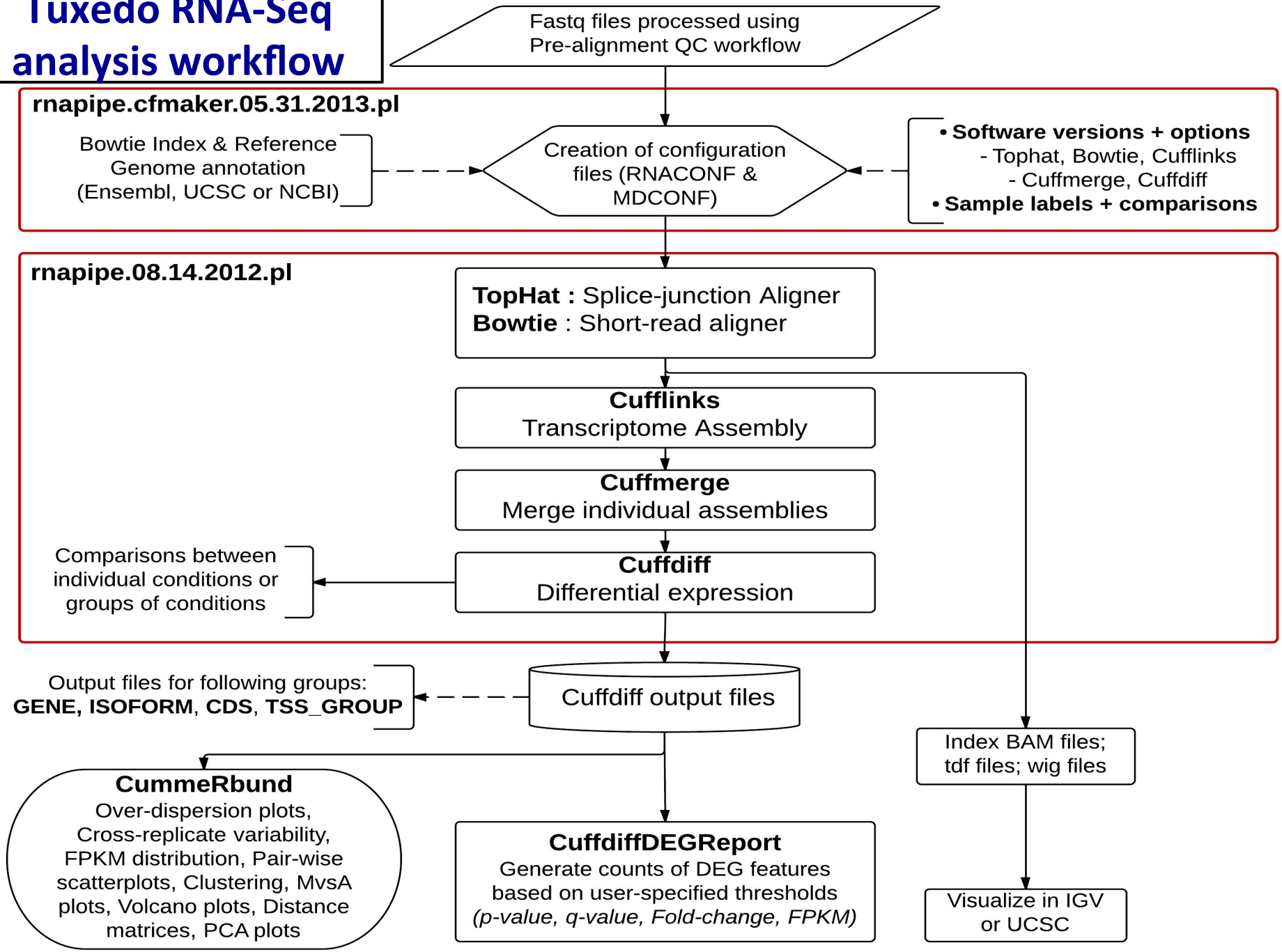
Post-alignment RNA-Seq QC



RNA-Seq analysis pipeline



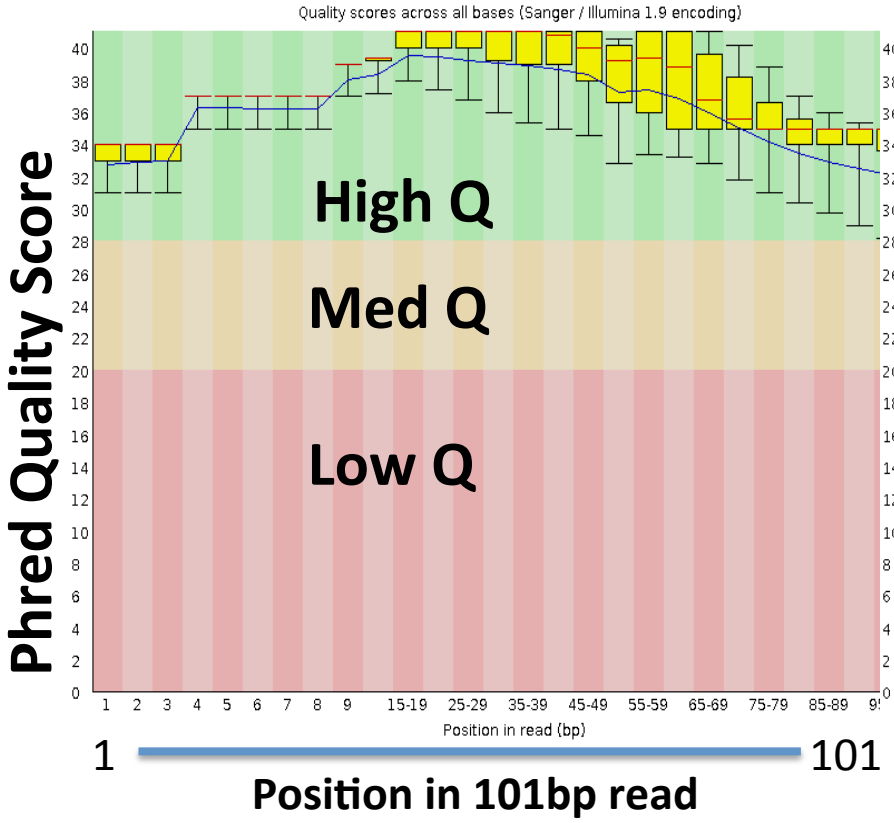
Tuxedo RNA-Seq analysis workflow

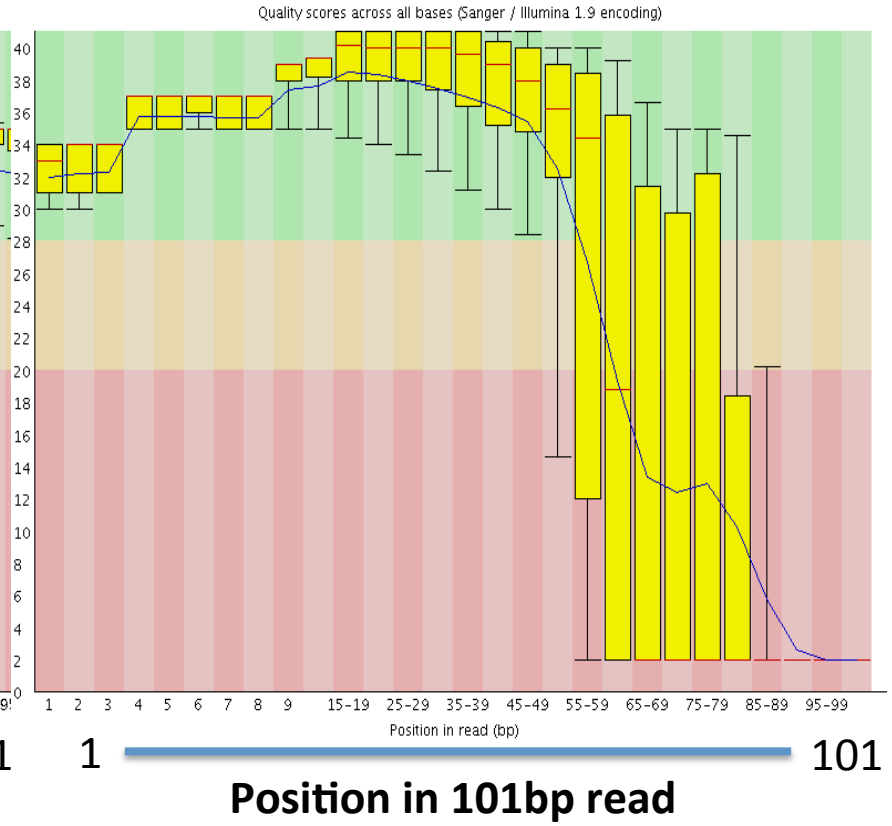


Outline

- Computational aspects of RNA-Seq data analysis
  - Best practices
- Example CCRIFX pipeline for RNA-Seq analysis:
  - Pre-alignment QC
  - Tuxedo RNA-Seq pipeline
  - Post alignment QC
- Visuals of FastQC plots (good vs bad data)
- Pointers

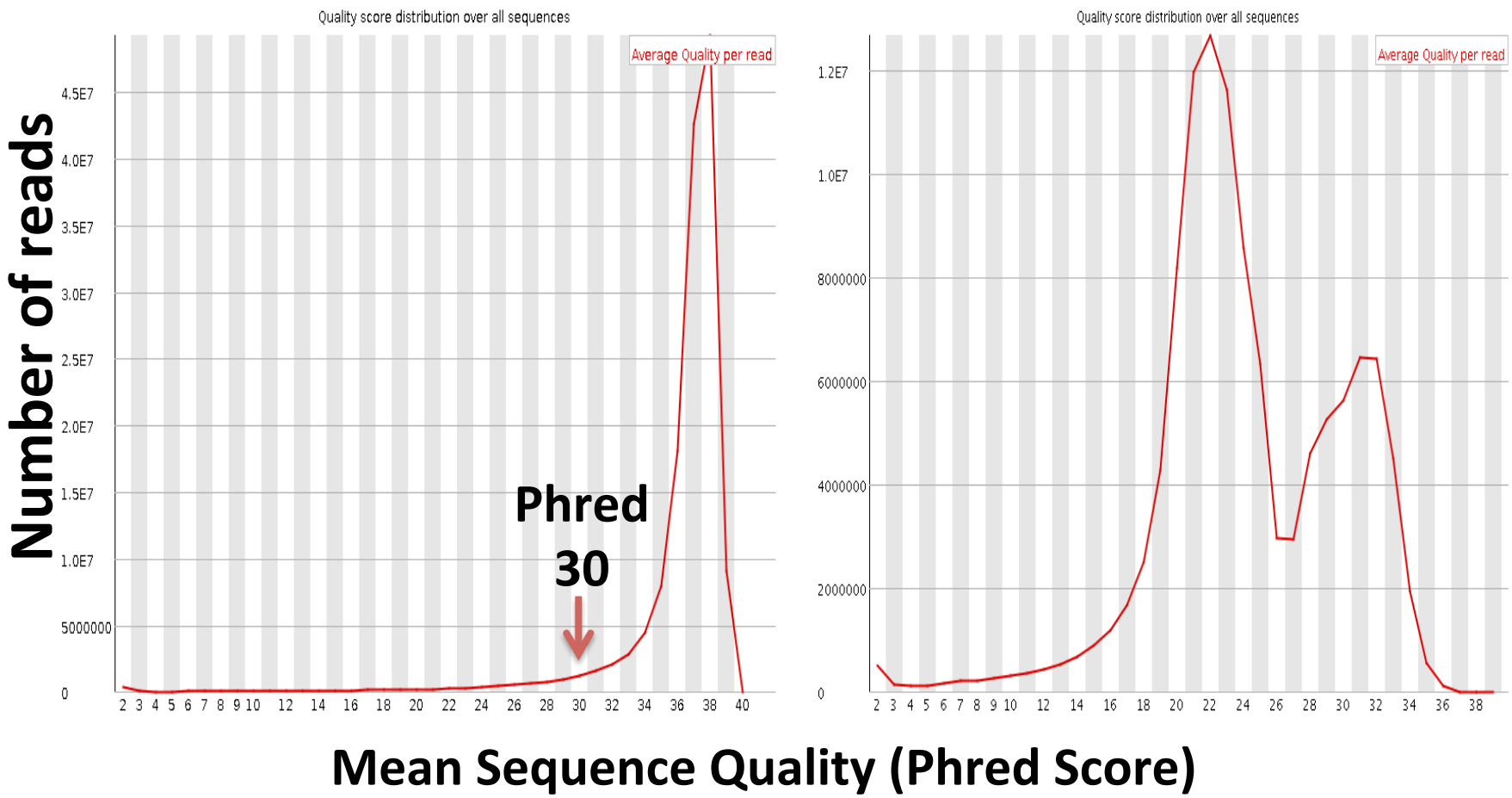
Per base sequence quality





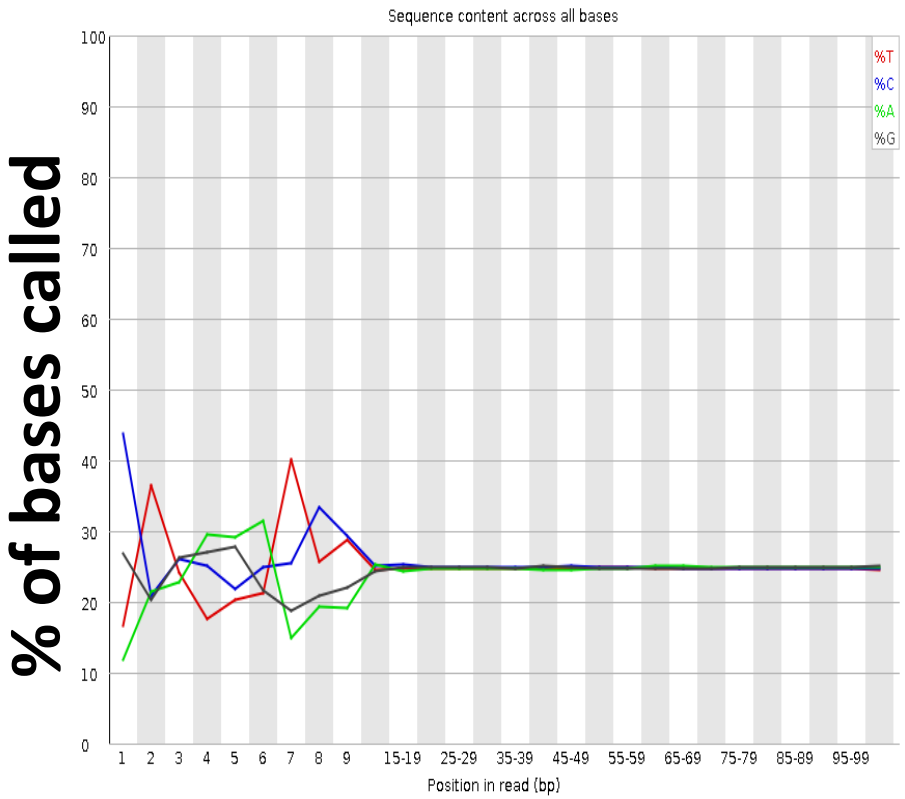
Output from Pre-alignment QC workflow

Per sequence quality scores

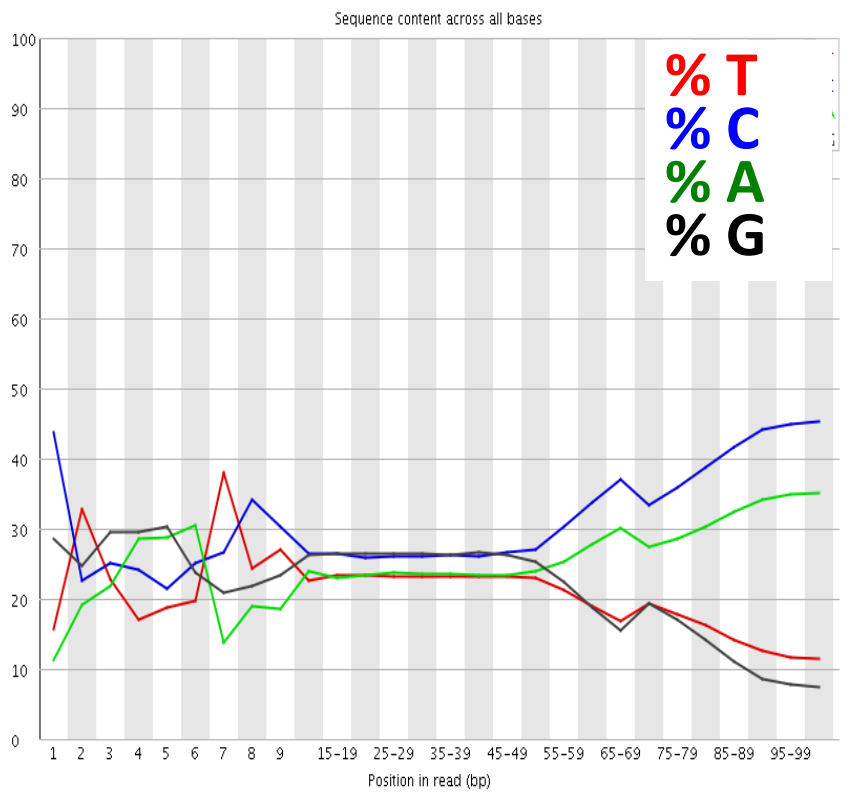


Output from Pre-alignment QC workflow

Per base sequence content



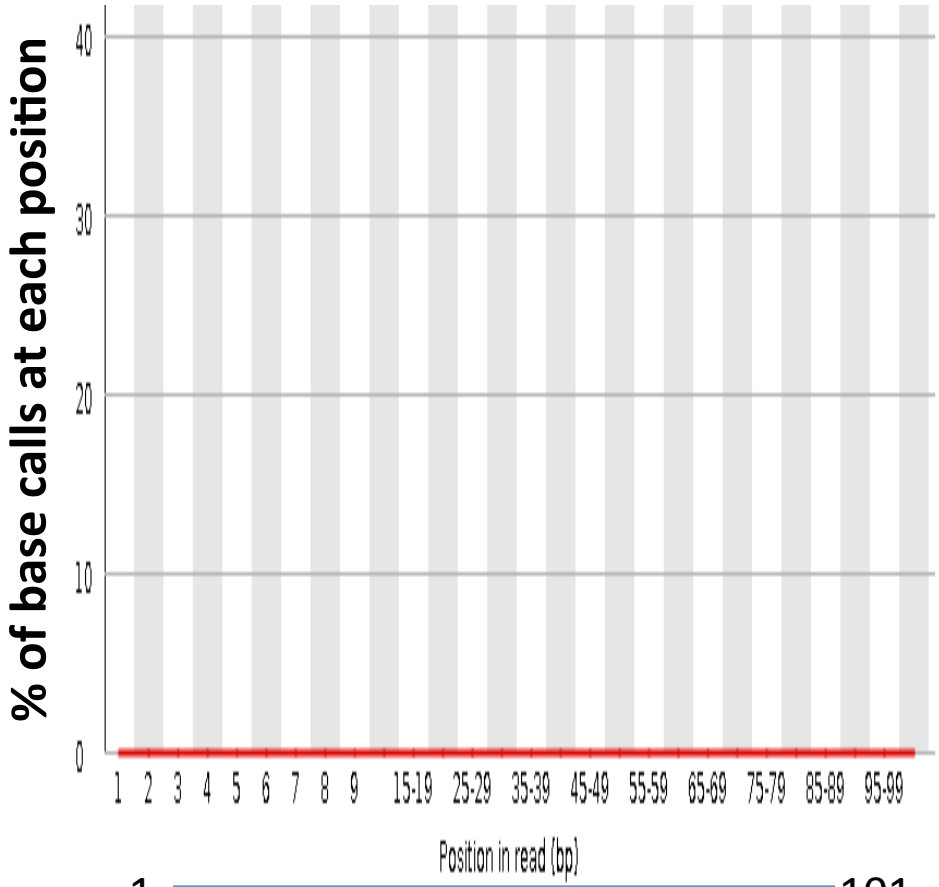
Position in 101bp read



Position in 101bp read

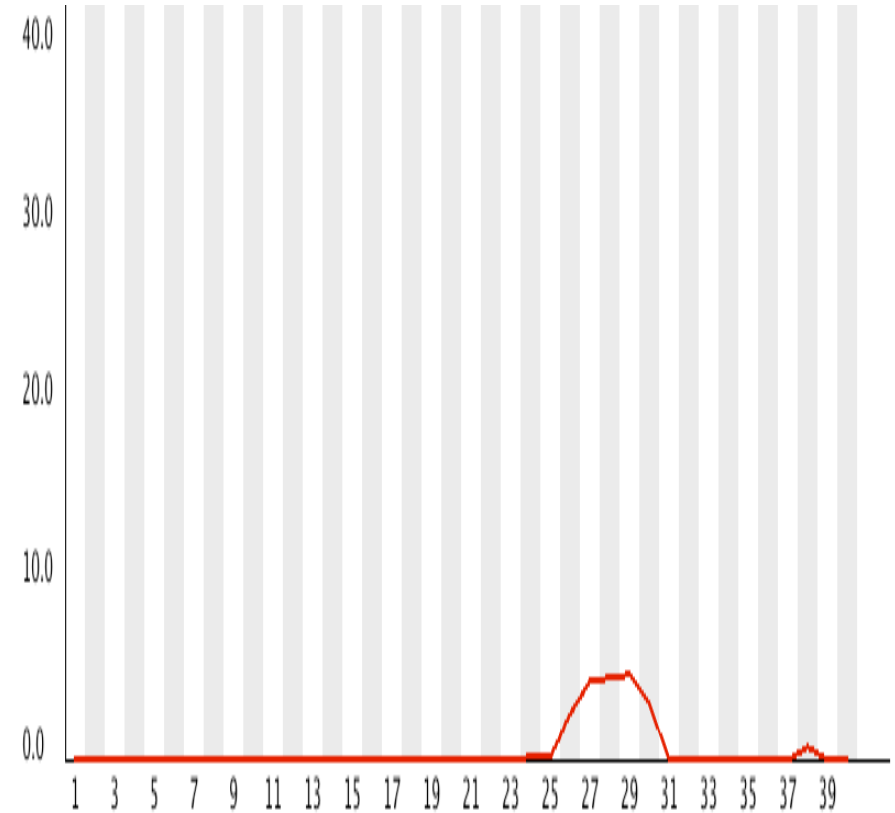
Output from Pre-alignment QC workflow

Per base N content



1 101

Position in 101bp read



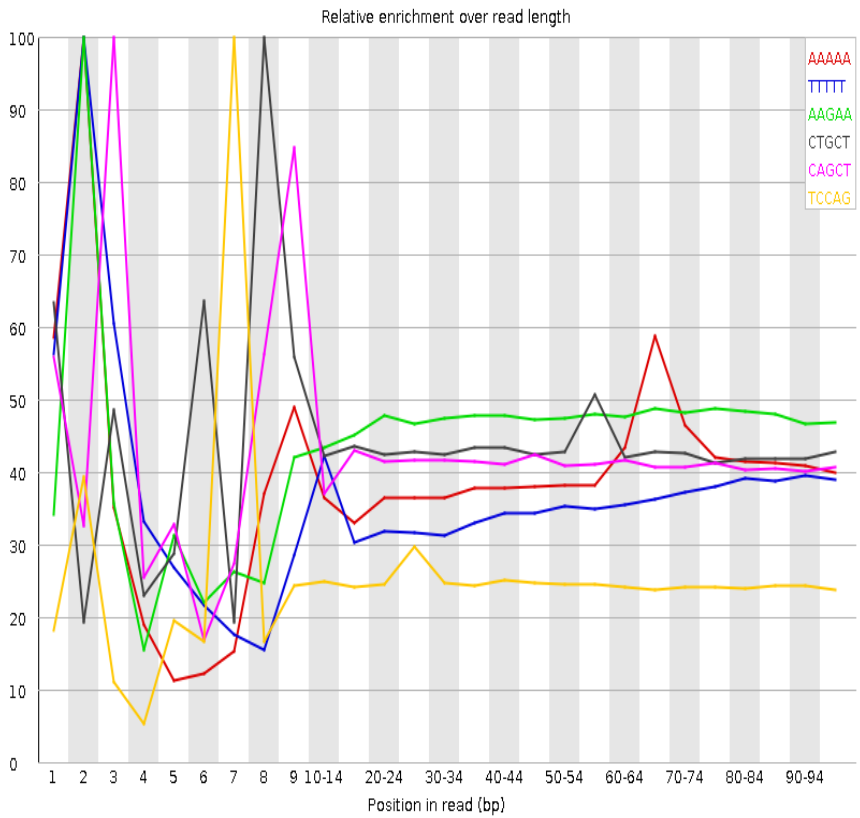
1 40

Position in 40bp read

Output from Pre-alignment QC workflow

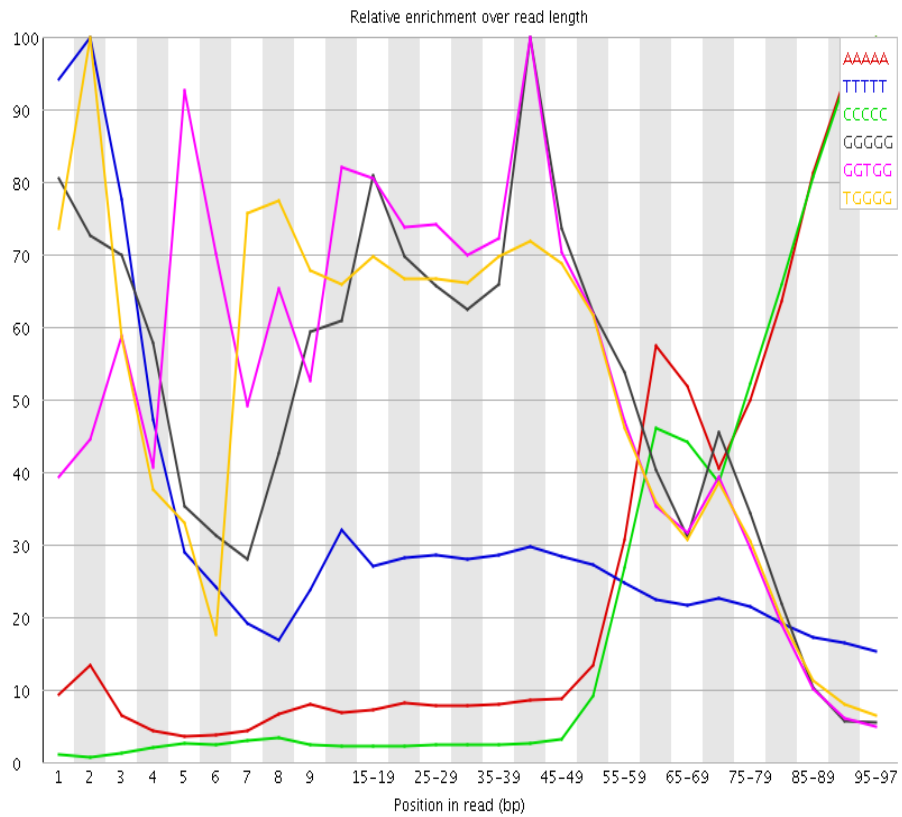
Kmer content

Relative enrichment



1 101

Position in 101bp read



1 101

Position in 101bp read

Output from Pre-alignment QC workflow

FastQC summary: untrimmed Fastq files

Sample	Per base sequence quality	Per sequence quality scores	Per base sequence content	Per base GC content	Per sequence GC content	Per base N content	Sequence Length Distribution	Overrepresented sequences	Kmer Content
10WG:R1	PASS	PASS	FAIL	FAIL	PASS	PASS	PASS	WARN	FAIL
10WG:R2	PASS	PASS	FAIL	FAIL	WARN	PASS	PASS	WARN	FAIL
12WG:R1	PASS	PASS	FAIL	FAIL	WARN	PASS	PASS	WARN	FAIL
12WG:R2	PASS	PASS	FAIL	FAIL	WARN	PASS	PASS	WARN	FAIL
16Msgn1:R1	PASS	PASS	FAIL	FAIL	WARN	PASS	PASS	WARN	FAIL
16Msgn1:R2	PASS	PASS	FAIL	FAIL	WARN	PASS	PASS	WARN	FAIL
23MG:R1	PASS	PASS	FAIL	FAIL	WARN	PASS	PASS	WARN	WARN
23MG:R2	PASS	PASS	FAIL	FAIL	WARN	PASS	PASS	PASS	WARN
24MG:R1	PASS	PASS	FAIL	FAIL	PASS	PASS	PASS	PASS	WARN
24MG:R2	PASS	PASS	FAIL	FAIL	PASS	PASS	PASS	PASS	WARN
26Msgn1:R1	PASS	PASS	FAIL	FAIL	WARN	PASS	PASS	WARN	WARN
26Msgn1:R2	PASS	PASS	FAIL	FAIL	WARN	PASS	PASS	PASS	WARN
38MG:R1	PASS	PASS	FAIL	FAIL	PASS	PASS	PASS	PASS	WARN
38MG:R2	PASS	PASS	FAIL	FAIL	PASS	PASS	PASS	PASS	WARN
39Msgn1:R1	PASS	PASS	FAIL	FAIL	PASS	PASS	PASS	WARN	WARN
39Msgn1:R2	PASS	PASS	FAIL	FAIL	PASS	PASS	PASS	PASS	WARN
3FLef1minusDox1:R1	FAIL	PASS	FAIL	FAIL	PASS	WARN	PASS	PASS	WARN
3FLef1minusDox1:R2	FAIL	PASS	FAIL	FAIL	PASS	PASS	PASS	PASS	WARN
3FLef1minusDox3:R1	PASS	PASS	FAIL	FAIL	PASS	PASS	PASS	PASS	WARN
3FLef1minusDox3:R2	PASS	PASS	FAIL	FAIL	PASS	PASS	PASS	PASS	WARN
3FLef1minusDox4:R1	FAIL	PASS	FAIL	FAIL	PASS	PASS	PASS	PASS	WARN
3FLef1minusDox4:R2	FAIL	PASS	FAIL	FAIL	PASS	PASS	PASS	PASS	WARN
3FLef1plusDox1:R1	FAIL	PASS	FAIL	FAIL	PASS	PASS	PASS	WARN	FAIL
3FLef1plusDox1:R2	FAIL	PASS	FAIL	FAIL	PASS	PASS	PASS	WARN	FAIL
3FLef1plusDox2:R1	FAIL	PASS	FAIL	FAIL	PASS	PASS	PASS	PASS	WARN
3FLef1plusDox2:R2	FAIL	PASS	FAIL	FAIL	PASS	PASS	PASS	PASS	WARN
3FLef1plusDox3:R1	FAIL	PASS	FAIL	FAIL	PASS	PASS	PASS	PASS	WARN
3FLef1plusDox3:R2	FAIL	PASS	FAIL	FAIL	PASS	PASS	PASS	PASS	WARN
6WG:R1	PASS	PASS	FAIL	FAIL	WARN	PASS	PASS	WARN	WARN
6WG:R2	PASS	PASS	FAIL	FAIL	WARN	PASS	PASS	PASS	WARN
FSp5minusDox2:R1	FAIL	PASS	FAIL	FAIL	PASS	PASS	PASS	PASS	WARN
FSp5minusDox2:R2	FAIL	PASS	FAIL	FAIL	PASS	PASS	PASS	PASS	WARN
FSp5minusDox3:R1	PASS	PASS	FAIL	FAIL	PASS	PASS	PASS	WARN	WARN
FSp5minusDox3:R2	PASS	PASS	FAIL	FAIL	PASS	PASS	PASS	PASS	WARN
FSp5minusDox4:R1	PASS	PASS	FAIL	FAIL	PASS	PASS	PASS	PASS	WARN
FSp5minusDox4:R2	PASS	PASS	FAIL	FAIL	PASS	PASS	PASS	PASS	WARN
FSp5plusDox1:R1	PASS	PASS	FAIL	FAIL	PASS	PASS	PASS	WARN	WARN
FSp5plusDox1:R2	PASS	PASS	FAIL	FAIL	PASS	PASS	PASS	PASS	WARN
FSp5plusDox3:R1	PASS	PASS	FAIL	FAIL	PASS	PASS	PASS	PASS	WARN
FSp5plusDox3:R2	PASS	PASS	FAIL	FAIL	PASS	PASS	PASS	PASS	WARN
FSp5plusDox4:R1	PASS	PASS	FAIL	FAIL	PASS	PASS	PASS	PASS	WARN
FSp5plusDox4:R2	PASS	PASS	FAIL	FAIL	PASS	PASS	PASS	PASS	WARN
Sp512:R1	PASS	PASS	FAIL	FAIL	WARN	PASS	PASS	WARN	WARN
Sp512:R2	PASS	PASS	FAIL	FAIL	WARN	PASS	PASS	PASS	WARN
SP55:R1	PASS	PASS	FAIL	FAIL	WARN	PASS	PASS	PASS	WARN
SP55:R2	PASS	PASS	FAIL	FAIL	WARN	PASS	PASS	PASS	WARN
Sp58:R1	PASS	PASS	FAIL	FAIL	PASS	PASS	PASS	WARN	WARN
Sp58:R2	PASS	PASS	FAIL	FAIL	PASS	PASS	PASS	WARN	FAIL
Sp5Sp824:R1	PASS	PASS	FAIL	FAIL	PASS	PASS	PASS	PASS	WARN
Sp5Sp824:R2	PASS	PASS	FAIL	FAIL	PASS	PASS	PASS	PASS	WARN
Sp5Sp828:R1	PASS	PASS	FAIL	FAIL	WARN	WARN	PASS	WARN	WARN
Sp5Sp828:R2	PASS	PASS	FAIL	FAIL	PASS	PASS	PASS	PASS	WARN
Sp5Sp89:R1	PASS	PASS	FAIL	FAIL	PASS	PASS	PASS	WARN	WARN
Sp5Sp89:R2	PASS	PASS	FAIL	FAIL	PASS	PASS	PASS	PASS	WARN
Sp82:R1	PASS	PASS	WARN	FAIL	PASS	PASS	PASS	FAIL	FAIL
Sp82:R2	PASS	PASS	WARN	FAIL	WARN	PASS	PASS	WARN	FAIL
Sp823:R1	FAIL	PASS	FAIL	FAIL	PASS	PASS	PASS	WARN	WARN
Sp823:R2	FAIL	PASS	FAIL	FAIL	WARN	PASS	PASS	PASS	WARN
Sp830:R1	PASS	PASS	FAIL	FAIL	WARN	PASS	PASS	PASS	WARN
Sp830:R2	PASS	PASS	FAIL	FAIL	WARN	PASS	PASS	PASS	WARN
Sp834:R1	PASS	PASS	FAIL	FAIL	WARN	PASS	PASS	PASS	WARN
Sp834:R2	PASS	PASS	FAIL	FAIL	WARN	PASS	PASS	PASS	WARN
Sp84:R1	PASS	PASS	FAIL	FAIL	PASS	PASS	PASS	WARN	WARN
Sp84:R2	PASS	PASS	FAIL	FAIL	PASS	PASS	PASS	PASS	WARN
Sp88:R1	PASS	PASS	FAIL	FAIL	PASS	PASS	PASS	WARN	WARN
Sp88:R2	PASS	PASS	FAIL	FAIL	PASS	PASS	PASS	WARN	WARN
Wnt3a22:R1	PASS	PASS	FAIL	FAIL	WARN	PASS	PASS	WARN	WARN
Wnt3a22:R2	PASS	PASS	FAIL	FAIL	WARN	PASS	PASS	PASS	WARN
Wnt3a26:R1	PASS	PASS	FAIL	FAIL	PASS	PASS	PASS	WARN	WARN
Wnt3a26:R2	PASS	PASS	FAIL	FAIL	PASS	PASS	PASS	PASS	WARN
Wnt3a3:R1	FAIL	PASS	FAIL	FAIL	PASS	PASS	PASS	PASS	WARN
Wnt3a3:R2	PASS	PASS	FAIL	FAIL	PASS	PASS	PASS	PASS	WARN
Wnt3a31:R1	PASS	PASS	FAIL	FAIL	PASS	PASS	PASS	WARN	WARN
Wnt3a31:R2	PASS	PASS	FAIL	FAIL	PASS	PASS	PASS	PASS	WARN
Wnt3a32:R1	PASS	PASS	FAIL	FAIL	PASS	PASS	PASS	PASS	WARN
Wnt3a32:R2	PASS	PASS	FAIL	FAIL	PASS	PASS	PASS	PASS	WARN
Wnt3a39:R1	PASS	PASS	FAIL	FAIL	WARN	PASS	PASS	WARN	WARN
Wnt3a39:R2	PASS	PASS	FAIL	FAIL	WARN	PASS	PASS	PASS	WARN

PASS

WARN

FAIL

FastQC summary: trimmed Fastq files

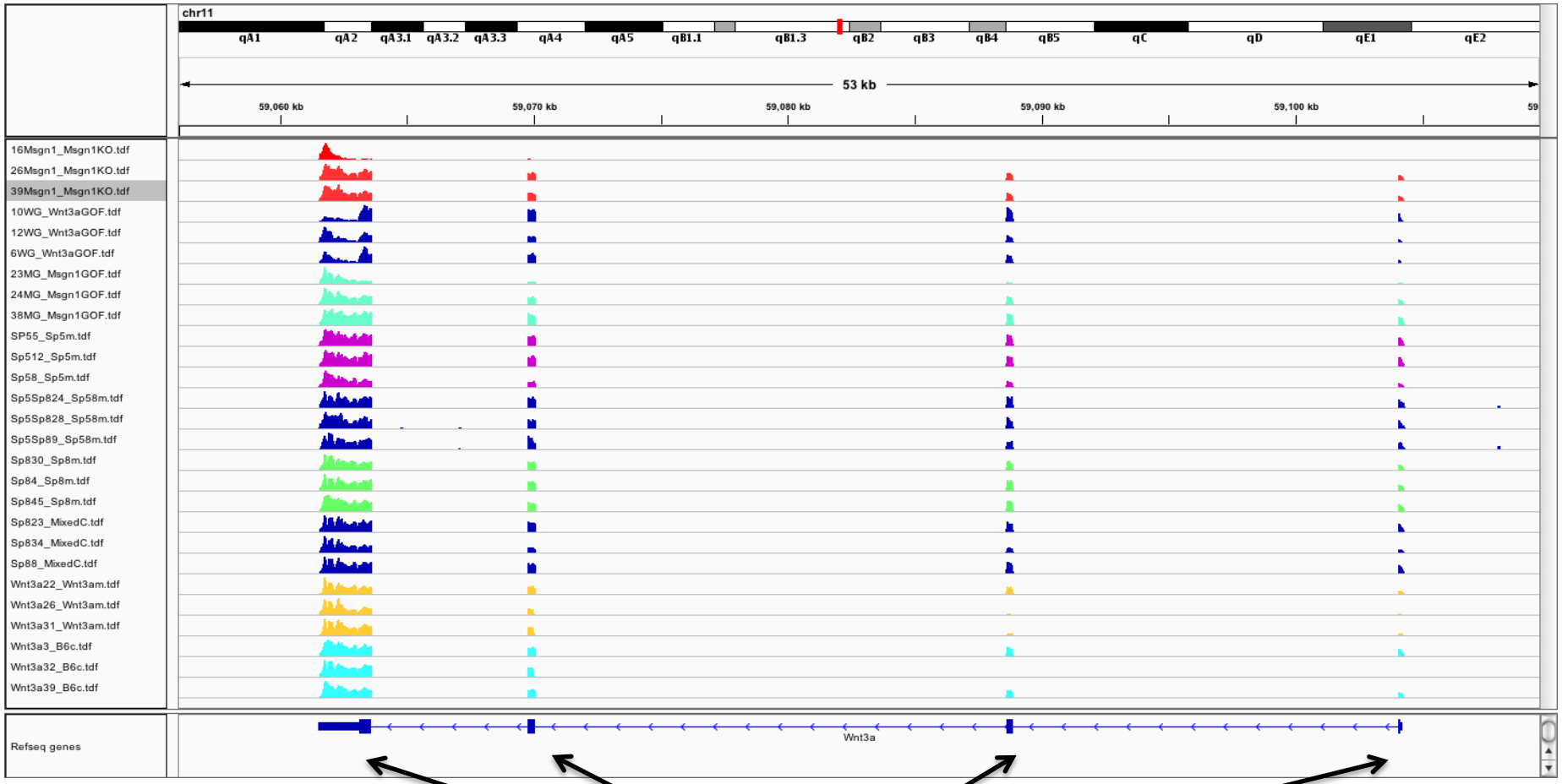
Sample	Per base sequence quality	Per sequence quality scores	Per base sequence content	Per base GC content	Per sequence GC content	Per base N content	Sequence Length Distribution	Overrepresented sequences	Kmer Content
10WG:Read #1	PASS	PASS	PASS	WARN	WARN	PASS	PASS	WARN	FAIL
10WG:Read #2	PASS	PASS	PASS	WARN	WARN	PASS	PASS	WARN	FAIL
12WG:Read #1	PASS	PASS	PASS	WARN	WARN	PASS	PASS	WARN	FAIL
12WG:Read #2	PASS	PASS	PASS	WARN	WARN	PASS	PASS	WARN	FAIL
16Msgn1:Read #1	PASS	PASS	PASS	WARN	WARN	PASS	PASS	WARN	FAIL
16Msgn1:Read #2	PASS	PASS	PASS	WARN	WARN	PASS	PASS	WARN	FAIL
23MG:Read #1	PASS	PASS	PASS	WARN	WARN	PASS	PASS	WARN	WARN
23MG:Read #2	PASS	PASS	PASS	WARN	WARN	PASS	PASS	PASS	WARN
24MG:Read #1	PASS	PASS	PASS	WARN	WARN	PASS	PASS	PASS	WARN
24MG:Read #2	PASS	PASS	PASS	WARN	WARN	PASS	PASS	PASS	WARN
26Msgn1:Read #1	PASS	PASS	PASS	WARN	WARN	PASS	PASS	WARN	WARN
26Msgn1:Read #2	PASS	PASS	PASS	WARN	WARN	PASS	PASS	PASS	WARN
38MG:Read #1	PASS	PASS	PASS	WARN	WARN	PASS	PASS	PASS	WARN
38MG:Read #2	PASS	PASS	PASS	WARN	WARN	PASS	PASS	PASS	WARN
39Msgn1:Read #1	PASS	PASS	PASS	WARN	WARN	PASS	PASS	WARN	WARN
39Msgn1:Read #2	PASS	PASS	PASS	WARN	WARN	PASS	PASS	PASS	WARN
3FLef1minusDox1:Read #1	PASS	PASS	PASS	WARN	PASS	PASS	PASS	PASS	PASS
3FLef1minusDox1:Read #2	PASS	PASS	PASS	WARN	WARN	PASS	PASS	PASS	PASS
3FLef1minusDox3:Read #1	PASS	PASS	PASS	WARN	PASS	PASS	PASS	PASS	PASS
3FLef1minusDox3:Read #2	PASS	PASS	PASS	WARN	PASS	PASS	PASS	PASS	WARN
3FLef1minusDox4:Read #1	PASS	PASS	PASS	WARN	WARN	PASS	PASS	PASS	PASS
3FLef1minusDox4:Read #2	PASS	PASS	PASS	WARN	PASS	PASS	PASS	PASS	PASS
3FLef1plusDox1:Read #1	PASS	PASS	PASS	WARN	PASS	PASS	PASS	WARN	FAIL
3FLef1plusDox1:Read #2	PASS	PASS	PASS	WARN	PASS	PASS	PASS	WARN	FAIL
3FLef1plusDox2:Read #1	PASS	PASS	PASS	WARN	PASS	PASS	PASS	PASS	PASS
3FLef1plusDox2:Read #2	PASS	PASS	PASS	WARN	PASS	PASS	PASS	PASS	PASS
3FLef1plusDox3:Read #1	PASS	PASS	PASS	WARN	PASS	PASS	PASS	PASS	PASS
3FLef1plusDox3:Read #2	PASS	PASS	PASS	WARN	PASS	PASS	PASS	PASS	PASS
6WG:Read #1	PASS	PASS	PASS	WARN	WARN	PASS	PASS	WARN	WARN
6WG:Read #2	PASS	PASS	PASS	WARN	WARN	PASS	PASS	PASS	WARN
FSp5minusDox2:Read #1	PASS	PASS	PASS	WARN	PASS	PASS	PASS	PASS	PASS
FSp5minusDox2:Read #2	PASS	PASS	PASS	WARN	PASS	PASS	PASS	PASS	PASS
FSp5minusDox3:Read #1	PASS	PASS	PASS	WARN	PASS	PASS	PASS	WARN	PASS
FSp5minusDox3:Read #2	PASS	PASS	PASS	WARN	WARN	PASS	PASS	PASS	WARN
FSp5minusDox4:Read #1	PASS	PASS	PASS	WARN	PASS	WARN	PASS	PASS	PASS
FSp5minusDox4:Read #2	PASS	PASS	PASS	WARN	PASS	PASS	PASS	PASS	PASS
FSp5plusDox1:Read #1	PASS	PASS	PASS	WARN	PASS	PASS	PASS	WARN	WARN
FSp5plusDox1:Read #2	PASS	PASS	PASS	WARN	PASS	PASS	PASS	WARN	WARN
FSp5plusDox3:Read #1	PASS	PASS	PASS	WARN	PASS	PASS	PASS	PASS	PASS
FSp5plusDox3:Read #2	PASS	PASS	PASS	WARN	WARN	PASS	PASS	WARN	PASS
FSp5plusDox4:Read #1	PASS	PASS	PASS	WARN	PASS	PASS	PASS	PASS	PASS
FSp5plusDox4:Read #2	PASS	PASS	PASS	WARN	WARN	PASS	PASS	WARN	PASS
Sp512:Read #1	PASS	PASS	PASS	WARN	WARN	PASS	PASS	WARN	PASS
Sp512:Read #2	PASS	PASS	PASS	WARN	WARN	PASS	PASS	PASS	WARN
SP55:Read #1	PASS	PASS	PASS	WARN	WARN	PASS	PASS	PASS	PASS
SP55:Read #2	PASS	PASS	PASS	WARN	WARN	PASS	PASS	WARN	WARN
Sp58:Read #1	PASS	PASS	PASS	WARN	WARN	PASS	PASS	WARN	WARN
Sp58:Read #2	PASS	PASS	PASS	WARN	WARN	PASS	PASS	WARN	FAIL
Sp5Sp824:Read #1	PASS	PASS	PASS	WARN	WARN	PASS	PASS	PASS	WARN
Sp5Sp824:Read #2	PASS	PASS	PASS	WARN	WARN	PASS	PASS	PASS	WARN
Sp5Sp828:Read #1	PASS	PASS	PASS	WARN	WARN	PASS	PASS	WARN	WARN
Sp5Sp828:Read #2	PASS	PASS	PASS	WARN	WARN	PASS	PASS	PASS	WARN
Sp5Sp89:Read #1	PASS	PASS	PASS	WARN	WARN	PASS	PASS	WARN	WARN
Sp5Sp89:Read #2	PASS	PASS	PASS	WARN	WARN	PASS	PASS	PASS	WARN
Sp82:Read #1	PASS	PASS	WARN	PASS	PASS	PASS	PASS	FAIL	FAIL
Sp82:Read #2	PASS	PASS	WARN	PASS	PASS	PASS	PASS	WARN	FAIL
Sp823:Read #1	PASS	PASS	PASS	WARN	WARN	PASS	PASS	WARN	PASS
Sp823:Read #2	PASS	PASS	PASS	WARN	PASS	PASS	PASS	PASS	WARN
Sp830:Read #1	PASS	PASS	PASS	WARN	WARN	PASS	PASS	PASS	PASS
Sp830:Read #2	PASS	PASS	PASS	WARN	WARN	PASS	PASS	PASS	WARN
Sp834:Read #1	PASS	PASS	PASS	WARN	WARN	PASS	PASS	PASS	PASS
Sp834:Read #2	PASS	PASS	PASS	WARN	WARN	PASS	PASS	PASS	WARN
Sp84:Read #1	PASS	PASS	PASS	WARN	WARN	PASS	PASS	WARN	WARN
Sp84:Read #2	PASS	PASS	PASS	WARN	WARN	PASS	PASS	PASS	WARN
Sp88:Read #1	PASS	PASS	PASS	WARN	WARN	PASS	PASS	WARN	WARN
Sp88:Read #2	PASS	PASS	PASS	WARN	WARN	PASS	PASS	WARN	WARN
Wnt3a22:Read #1	PASS	PASS	PASS	WARN	WARN	PASS	PASS	WARN	WARN
Wnt3a22:Read #2	PASS	PASS	PASS	WARN	WARN	PASS	PASS	PASS	WARN
Wnt3a26:Read #1	PASS	PASS	PASS	WARN	PASS	PASS	PASS	WARN	WARN
Wnt3a26:Read #2	PASS	PASS	PASS	WARN	PASS	PASS	PASS	PASS	WARN
Wnt3a3:Read #1	PASS	PASS	PASS	WARN	PASS	PASS	PASS	PASS	PASS
Wnt3a3:Read #2	PASS	PASS	PASS	WARN	PASS	PASS	PASS	PASS	WARN
Wnt3a31:Read #1	PASS	PASS	PASS	WARN	WARN	PASS	PASS	WARN	PASS
Wnt3a31:Read #2	PASS	PASS	PASS	WARN	WARN	PASS	PASS	PASS	WARN
Wnt3a32:Read #1	PASS	PASS	PASS	WARN	PASS	PASS	PASS	PASS	WARN
Wnt3a32:Read #2	PASS	PASS	PASS	WARN	WARN	PASS	PASS	PASS	WARN
Wnt3a39:Read #1	PASS	PASS	PASS	WARN	WARN	PASS	PASS	WARN	WARN
Wnt3a39:Read #2	PASS	PASS	PASS	WARN	WARN	PASS	PASS	PASS	WARN

PASS

WARN

FAIL

Visualization of BAM files



Exons

Post-alignment QC

Post-alignment QC: Picard metrics

- **Alignment summary metrics**

SAMPLE	TOTAL READS	ALIGNED READS	MISMATCH RATE	INDEL RATE	READ LENGTH	% READS ALIGNED IN PAIRS	STRAND BALANCE	% CHIMERAS
SAMPLE 1 READ 1	121,264,839	113,207,834	0.0024	0.0001	80	0.97	0.50	0.002
SAMPLE 1 READ 2	120,622,996	112,611,702	0.0019	0.0001	80	0.98	0.50	0.002
SAMPLE 1 PAIR (READ1 + READ2)	241,887,835	225,819,536	0.0022	0.0001	80	0.97	0.50	0.002

Post-alignment QC: Picard metrics

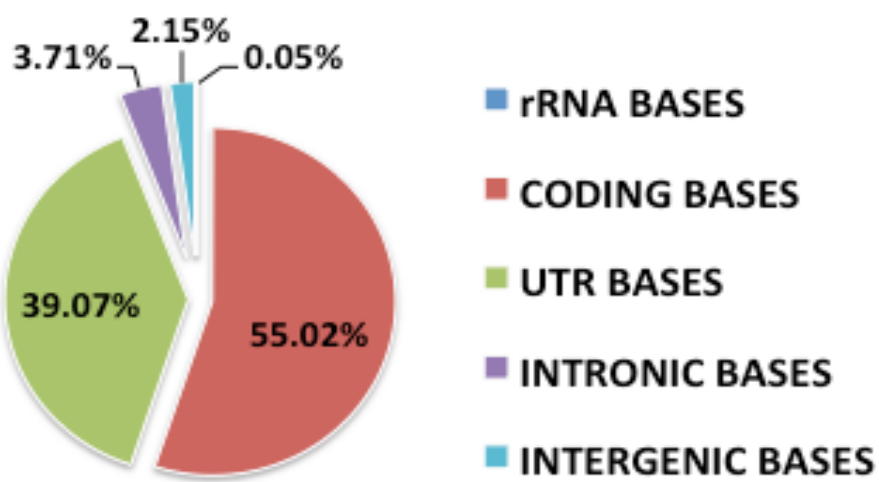
- **Insert size metrics**

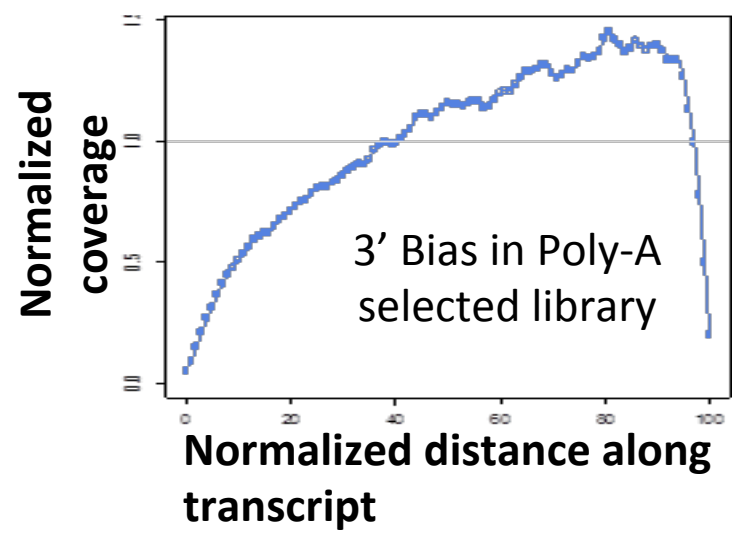
SAMPLE	MEDIAN INSERT SIZE	MEDIAN ABSOLUTE DEVIATION	MEAN INSERT SIZE	STANDARD DEVIATION	TOTAL READ PAIRS	PAIR ORIENTATION
SAMPLE 1	189	80	229.66	189.05	116,986,895	FR
SAMPLE 2	161	55	184.31	120.28	146,691,871	FR
SAMPLE 3	159	52	181.34	112.64	131,356,931	FR

Post-alignment QC: Picard metrics

- **RNASeq metrics**

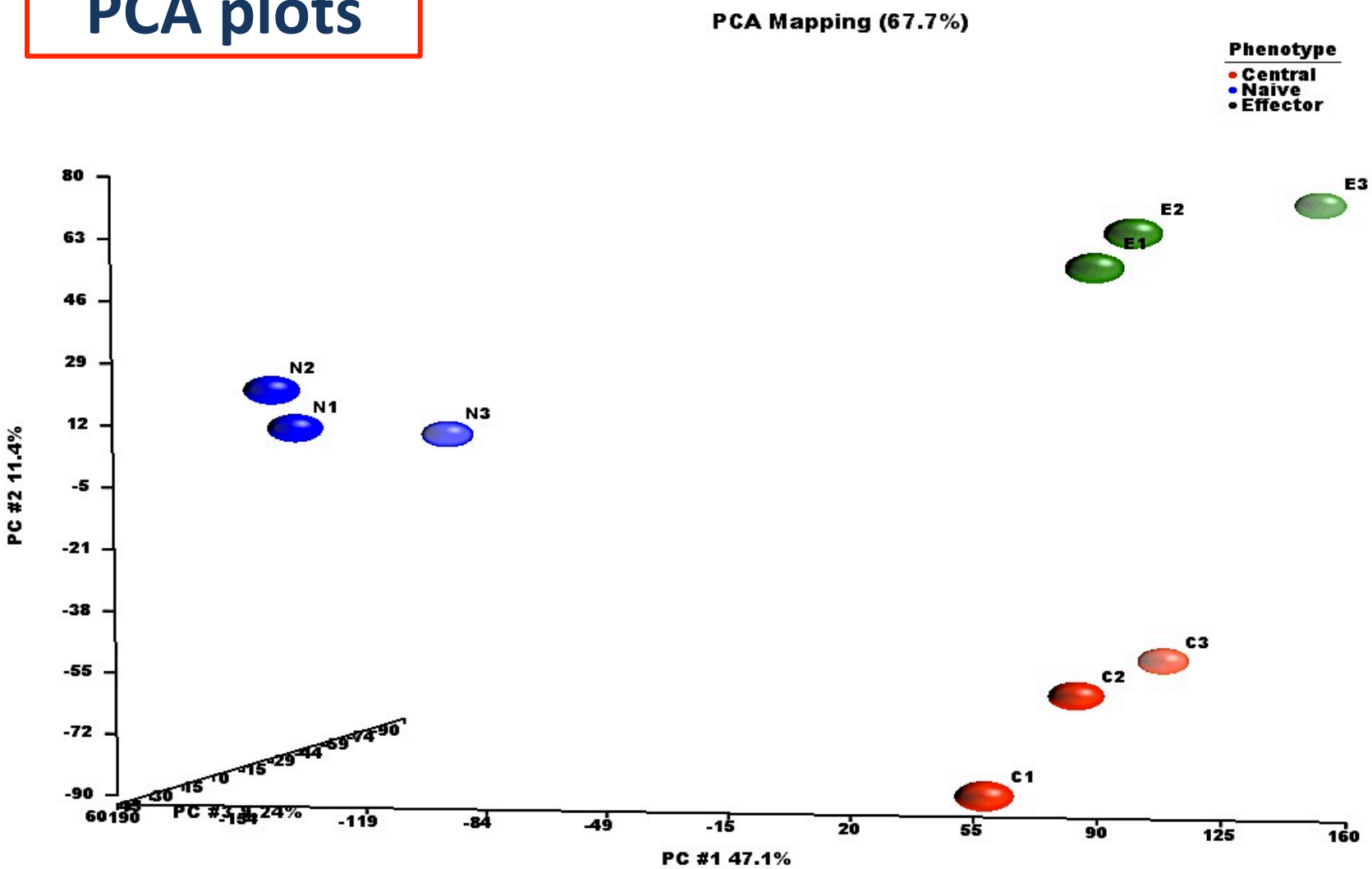
SAMPLE	rRNA BASES	CODING BASES	UTR BASES	INTRONIC BASES	INTERGENIC BASES	mRNA BASES	MEDIAN CV COVERAGE	5' BIAS	3' BIAS	5' TO 3' BIAS
SAMPLE 1	0.05%	55.02%	39.07%	3.71%	2.15%	94.09%	0.509	0.319	0.373	0.696
SAMPLE 2	0.11%	48.41%	47.16%	2.19%	2.14%	95.57%	0.689	0.169	0.600	0.275
SAMPLE 3	0.27%	45.47%	49.79%	2.12%	2.35%	95.26%	0.799	0.112	0.726	0.165





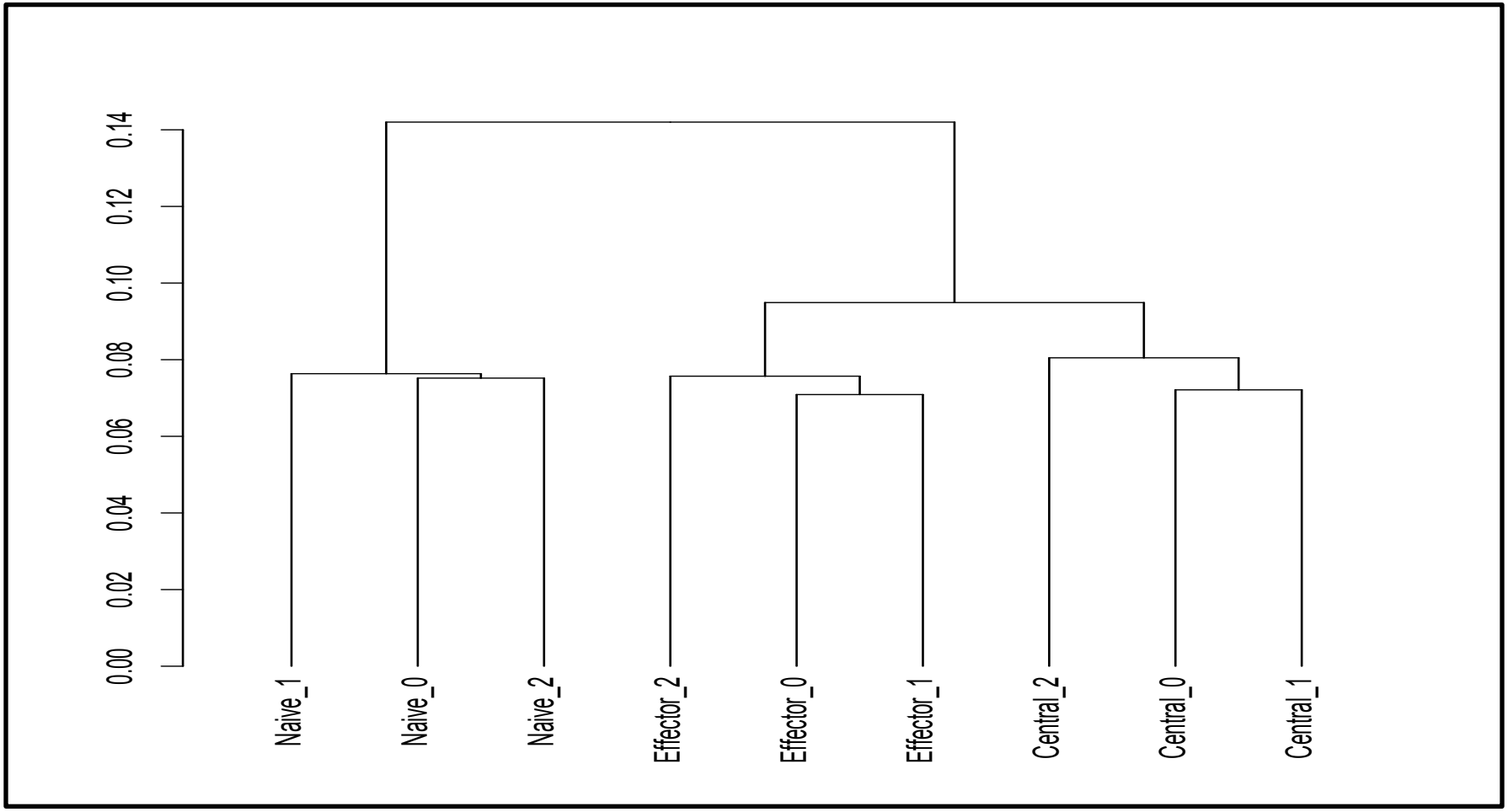
Visuals of downstream analysis

PCA plots



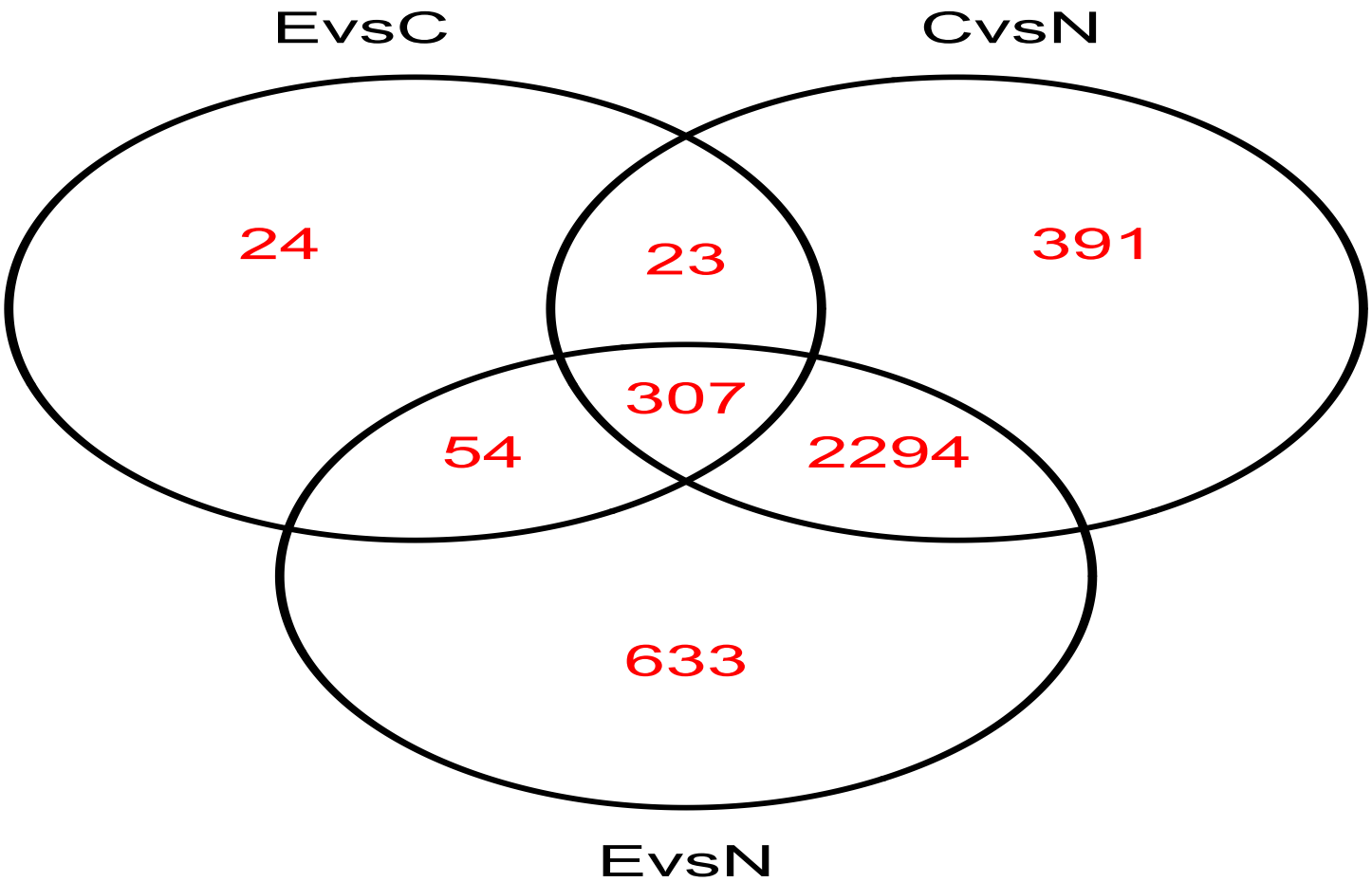
Visuals of downstream analysis

Dendrograms



Visuals of downstream analysis

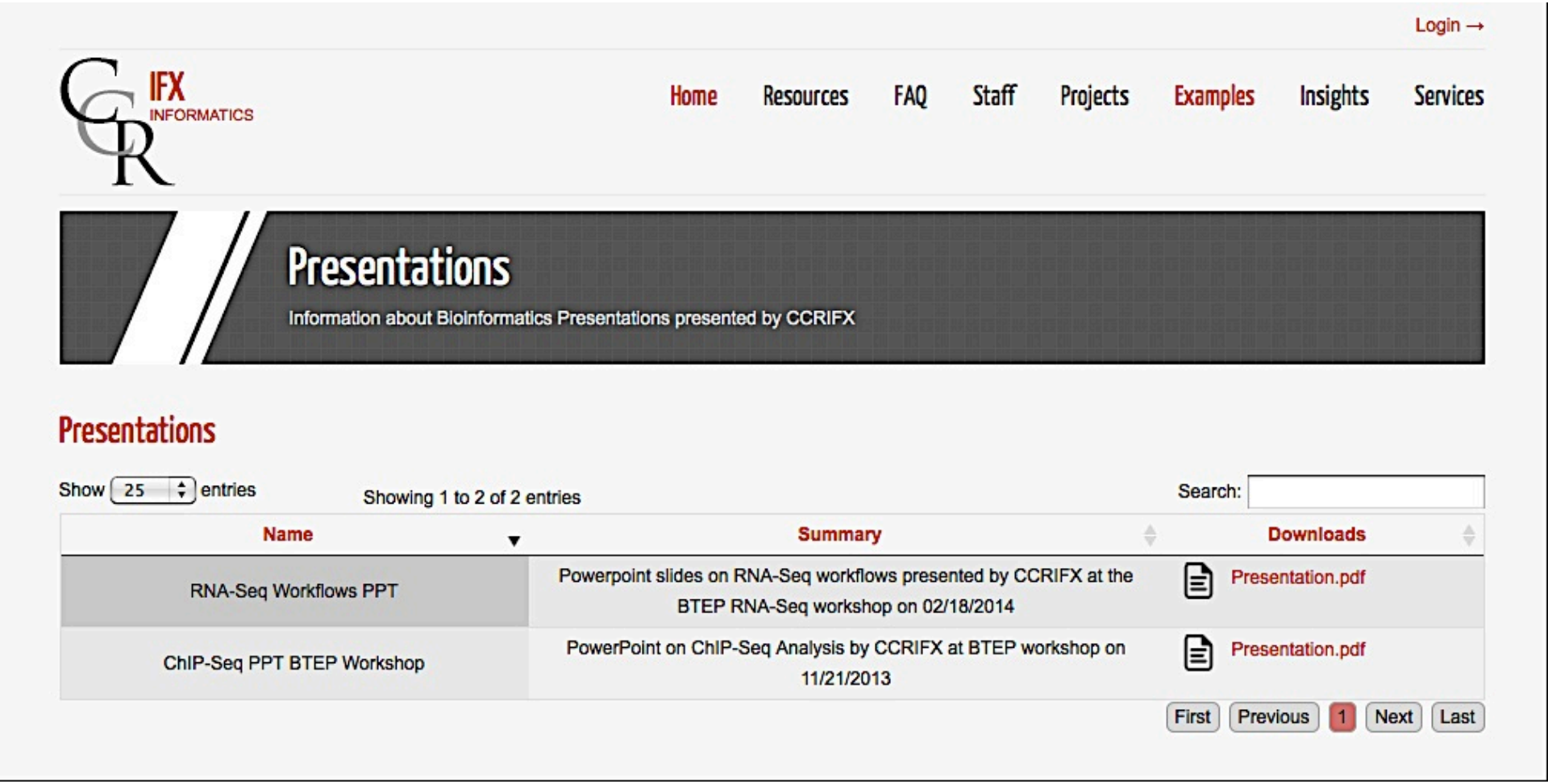
Venn diagrams



Outline

- Computational aspects of RNA-Seq data analysis
  - Best practices
- Example CCRIFX pipeline for RNA-Seq analysis:
  - Pre-alignment QC
  - Tuxedo RNA-Seq pipeline
  - Post alignment QC
- Visuals of FastQC plots (good vs bad data)
- Pointers

Download this PPT



http://ccrifx.cancer.gov/apps/site/presentations

Further Information on Workflows

Login →

IFX INFORMATICS

Home | Resources | FAQ | Staff | Projects | Examples | Insights | Services

Workflows for Bioinformatics Analysis

Information about Bioinformatics Workflows used by CCRIFX

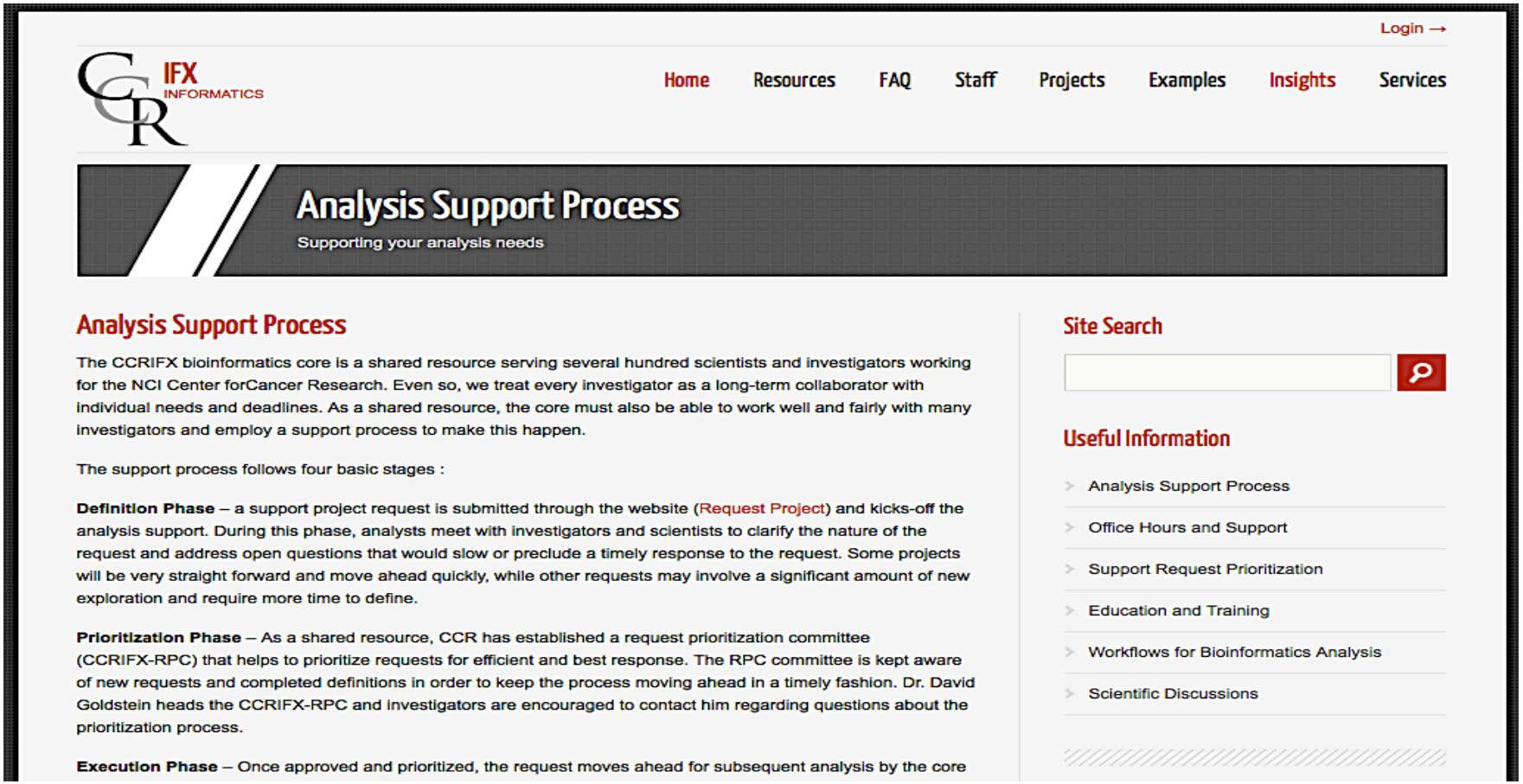
Workflows for Bioinformatics Analysis

Show 25 entries Showing 1 to 3 of 3 entries Search:

Name	Summary	Downloads
Tuxedo RNA-Seq analysis	A set of two scripts that automate the execution of all steps of the Tuxedo pipeline (mapping, transcriptome assembly, differential expression estimates) using configuration files	Flowchart.pdf userDocumentation.pdf
Post-alignment RNA-Seq Quality Control	Automated generation of a series of quality control metrics for RNA-Seq data using BAM files as input. Various Picard and RNA-SeQC alignment metrics are computed and user-friendly aggregated reports are compiled from generated metrics. Supports integrated and automated pre-processing of BAM files to make them suitable for QC modules to work correctly	Flowchart.pdf userDocumentation.pdf
Pre-alignment Quality Control	Workflow to merge Fastq files based on Illumina naming convention, followed by automated job submission for running FastQC on merged FastQ files. Optionaly, trimming and/or adapter filtered can be carried out using Trimmomatic.	Flowchart.pdf userDocumentation.pdf

ccrifx.cancer.gov/apps/site/workflows_for_bioinformatics_analysis

Visit our website for more information



Login →

IFX INFORMATICS

Home | Resources | FAQ | Staff | Projects | Examples | Insights | Services

Analysis Support Process

Supporting your analysis needs

Analysis Support Process

The CCRIFX bioinformatics core is a shared resource serving several hundred scientists and investigators working for the NCI Center forCancer Research. Even so, we treat every investigator as a long-term collaborator with individual needs and deadlines. As a shared resource, the core must also be able to work well and fairly with many investigators and employ a support process to make this happen.

The support process follows four basic stages :

Definition Phase – a support project request is submitted through the website (Request Project) and kicks-off the analysis support. During this phase, analysts meet with investigators and scientists to clarify the nature of the request and address open questions that would slow or preclude a timely response to the request. Some projects will be very straight forward and move ahead quickly, while other requests may involve a significant amount of new exploration and require more time to define.

Prioritization Phase – As a shared resource, CCR has established a request prioritization committee (CCRIFX-RPC) that helps to prioritize requests for efficient and best response. The RPC committee is kept aware of new requests and completed definitions in order to keep the process moving ahead in a timely fashion. Dr. David Goldstein heads the CCRIFX-RPC and investigators are encouraged to contact him regarding questions about the prioritization process.

Execution Phase – Once approved and prioritized, the request moves ahead for subsequent analysis by the core

Site Search

Useful Information

- Analysis Support Process
- Office Hours and Support
- Support Request Prioritization
- Education and Training
- Workflows for Bioinformatics Analysis
- Scientific Discussions

ccrifx.cancer.gov/apps/site/analysis_support_process

List of RNA-Seq Bioinformatics tools

Tool category	Tools
QC, filtering, pre-processing	**FastQC**, **Trimmomatic**, **RNA-SeQC**, **Picard**, Cutadapt, FASTX, Flexbar, htSeqTools, PRINSEQ, qrqc, RSeQC, SAMstat, SEECER, ShortRead
Alignment	**Unspliced:** BFAST, **Bowtie**, BWA, SOAP, Maq, Mosaik, NovoAlign, SHRiMP, Stampy **Spliced (known)**: **TopHat**, Erange, RUM, RNASEQR, SpliceSeq **Spliced (de novo)**: **TopHat**, GSNAP, HMMSplicer, MapSplice, QPALMA, SpliceMap
Quantitative analysis & differential expression	**Cufflinks, DESeq, EdgeR**, DEGSeq, DEXSeq, DiffSplice, BaySeq, Alexa-Seq, BitSeq,, ERANGE, eXpress, RSEM, SpliceTrap
Workbench (integrated pipeline solutions)	**Commercial**: **Partek, Genomatics**, Avadis NGS, **CLC**, DNASTAR, GeneSpring GX **Open Source**: **Galaxy**, Taverna, GenePattern, ArrayExpressHTS, easyRNASeq, MeV
Alternative splicing analysis	MISO, **Cufflinks/Cuffdiff**, DEXseq, SAJR
Fusion genes/chimeras/ translocation finders/ structural variations	BreakDancer, ChimeraScan, **FusionMap**, FusionSeq, SOAPFuse, SOAPfusion, **Tophat-Fusion**, **DeFuse**
Visualization tools	**CummeRbund, IGV,** IGB, MapView, Tablet, Savant, **SeqMonk**, Gbrowse, EagleView
Transcriptome assembly	**Genome-guided:** Cufflinks, iReckon, IsoInfer, RNAeXpress, Scripture **Genome-independent:** Oases, Rannotator, SOAPdenovo, Trans-ABySS, Trinity, Velvet

http://en.wikipedia.org/wiki/List_of_RNA-Seq_bioinformatics_tools

To learn more, visit…

- **Technology, issues & challenges**
  - Canadian Bioinformatics Workshops: http://bioinformatics.ca//files/public/BiCG_2012_Module7.pdf
  - GenomeWeb: http://www.genomeweb.com/node/926779
  - RNA-Seq for everyone: http://rnaseq.uoregon.edu/analysis.html
- **Integrated analysis guides and tutorials**
  - Wikibooks: http://en.wikibooks.org/wiki/Next_Generation_Sequencing_%28NGS%29/RNA
  - Utrecht University hands-on tutorial: http://testweb.science.uu.nl/pmi/publications/PDF/2013/TiPS-VanVerk-Hickman-2013-Hands_on_Tutorial.pdf
  - GenomeQuest: http://wiki.genomequest.com/index.php/RNA_Seq
- **Blogs**
  - SEQanswers, BioStars, RNA-SeqBlog

Acknowledgements

This presentation was developed by the CCRIFX core of the Advanced Biomedical Computing Center operated by Leidos Biomedical Research. Funding for CCRIFX comes from the CCR Bioinformatics Initiative (YT-10-060).

CCRIFX
Natalie Abrams, PhD
Yvonne Edwards, PhD
Parthav Jailwala, MS
Li Jia, MS
Manjula Kasoji, MS
Anand S Merchant, MD, PhD
Fathi Elloumi, PhD
Eddie Salinas, MS
Han Si, PhD
Nathan Starner, BS

Eric Stahlberg, PhD

ABCC
Brian Luke, PhD
Uma Mudunuri, MS
Bob Stephens, PhD
Ming Yi, PhD

Jack Collins, PhD

CCR experts
Maggie Cam, PhD
Sean Davis, MD, PhD
Max Lee, PhD
Peter FitzGerald, PhD
David Goldstein, PhD

Sequencing Facility
Yongmei Zhao, MS
Bao Tran, MS

Get in touch with us!

Submit a request: http://ccrifx.cancer.gov

E-mail us: ccrifx_support@mail.nih.gov

Visit us: **Bethesda: Mon-Wed & Friday**

Building 37, room 1123

Building 41, room B620

Frederick: Thursday & Friday

Building 430, First floor

Office hours: Fridays 9:30am -11:30am

Thank You !

Extra slides

RNA-Seq: Advantages over microarrays

- Higher resolution, more sensitivity
- High signal-to-noise ratio
- Potentially unlimited dynamic range of expression with absolute rather than relative values.
- Requires less RNA sample & results highly reproducible for both technical & biological replicates
- No a-priori knowledge of genome content required
- Investigation of both known and novel transcripts
- Alternative splicing patterns, RNA editing events

Applications of RNA-Seq platform

- Relative expression analyses
- Characterization of alternative splicing patterns
- Discovery of novel transcripts and isoforms
- Small RNA profiling
- Allele-specific & strand-specific expression
- Exploration of non-model-organism transcriptomes
- Mapping transcription start sites
- Gene fusion detection
- Profiling low-quantity RNA samples
- RNA editing