MDI Biological Laboratory

Comparative Genomics & Data Science Core

Workflows in the Cloud

Heath Fuqua, Bioinformatician I, Comparative Genomics and Data Science Core

April 2024

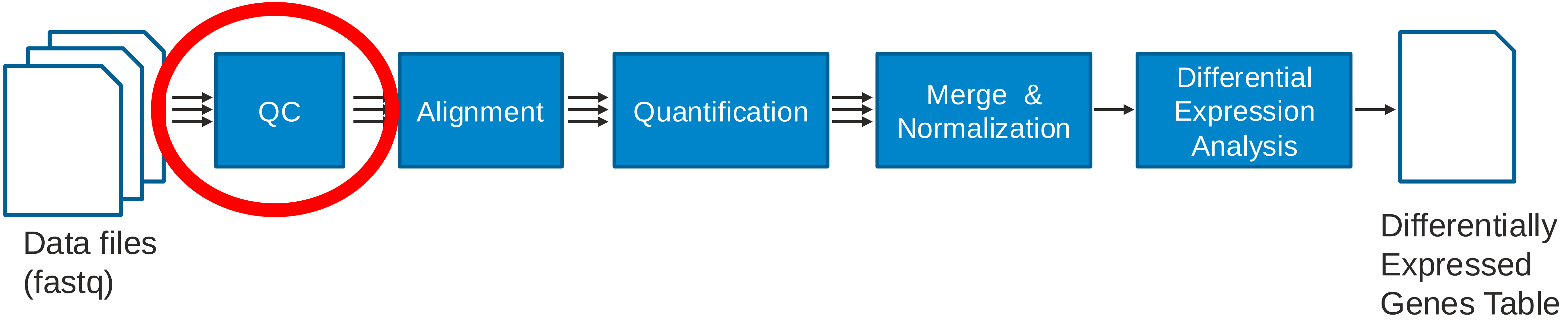
Agenda

- Bioinformatics Pipelines and Approaches
- Amazon Web Services (AWS)
- Nextflow/nf-core & Memverge
- NF-Core RNAseq Pipeline Setup & Launch
- Lunch
- Results
- Next Steps/Questions

What goes into the complete analysis of a genome-scale data set?

(using Bulk RNA-seq as an example)

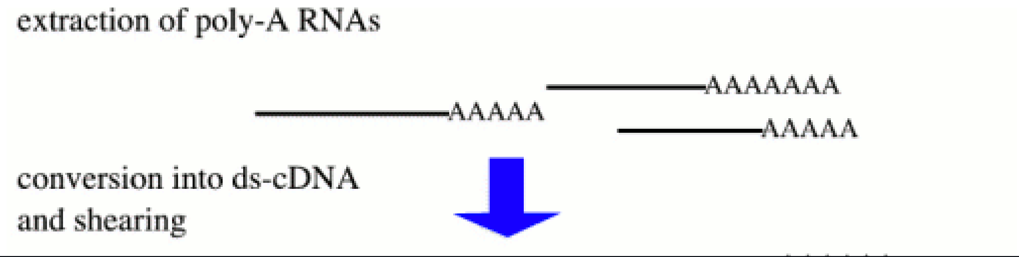
- Most complex data needs multiple steps to go from raw data to ”answers”
- Example: RNAseq data to Differentially expressed genes


QC
Alignment
Quantification
Merge & Normalization
Differential Expression Analysis
Data files (fastq)
Differentially Expressed Genes Table

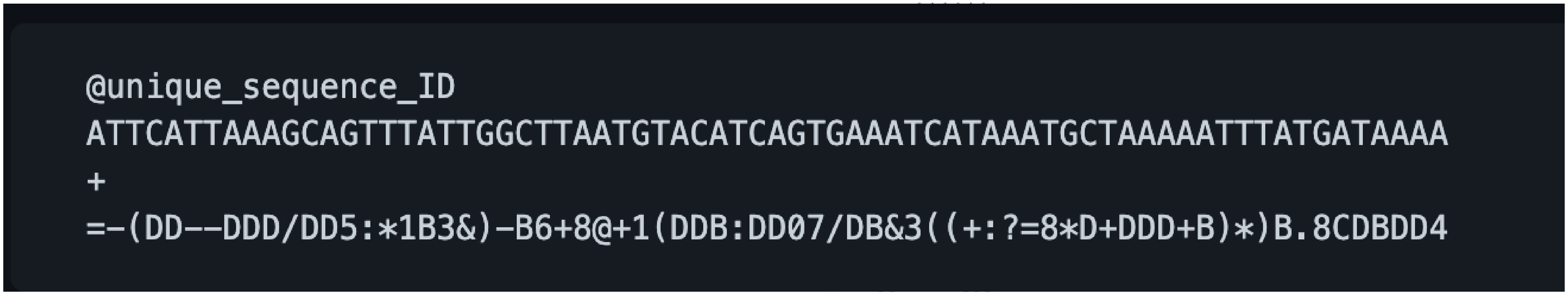
MDI Biological Laboratory

Pioneering new approaches in regenerative medicine

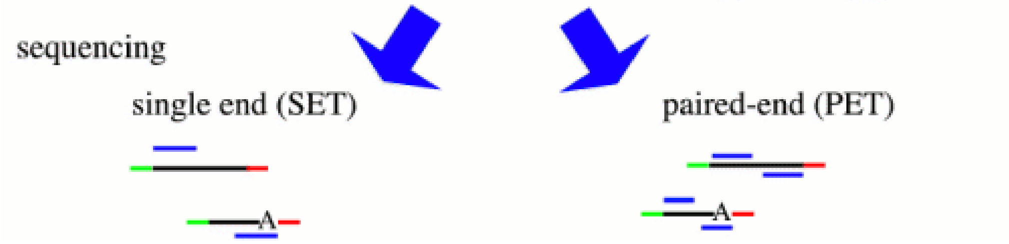
A typical RNA Seq experiment (and why we need QC)



extraction of poly-A RNAs
AAAAA
AAAAAAA
AAAAA
conversion into ds-cDNA and shearing



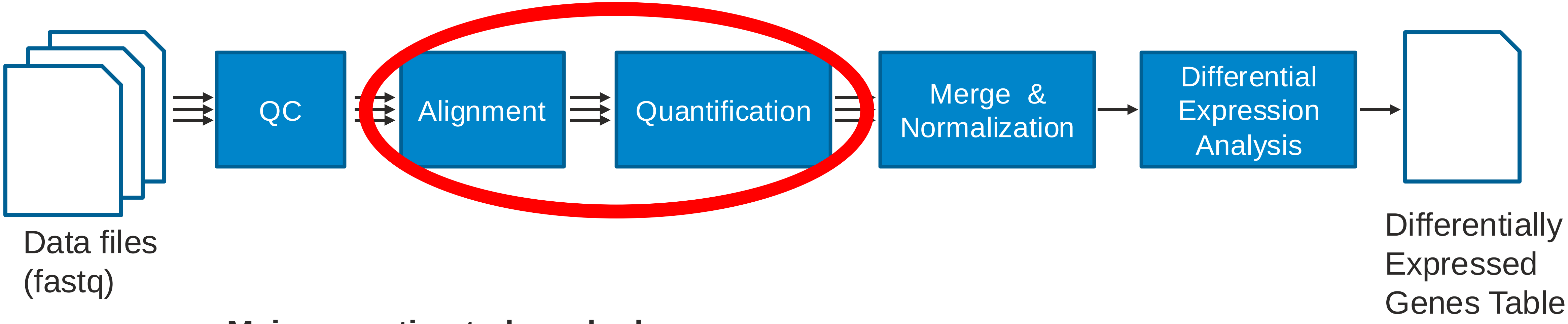
```
@unique_sequence_ID
ATTCATTAAAGCAGTTTATTGGCTTAATGTACATCAGTGAAATCATAAATGCTAAAAATTTATGATAAAA
+
=-(DD--DDD/DD5:*1B3&)-B6+8@+1(DDB:DD07/DB&3((+:?=8*D+DDD+B)*)B.8CDBDD4
```



http://cmb.molgen.mpg.de/2ndGenerationSequencing/Solas/RNA-seq.html

MDI Biological Laboratory
Pioneering new approaches in regenerative medicine

A workflow for RNA-seq Differential Expression



Major question to be asked

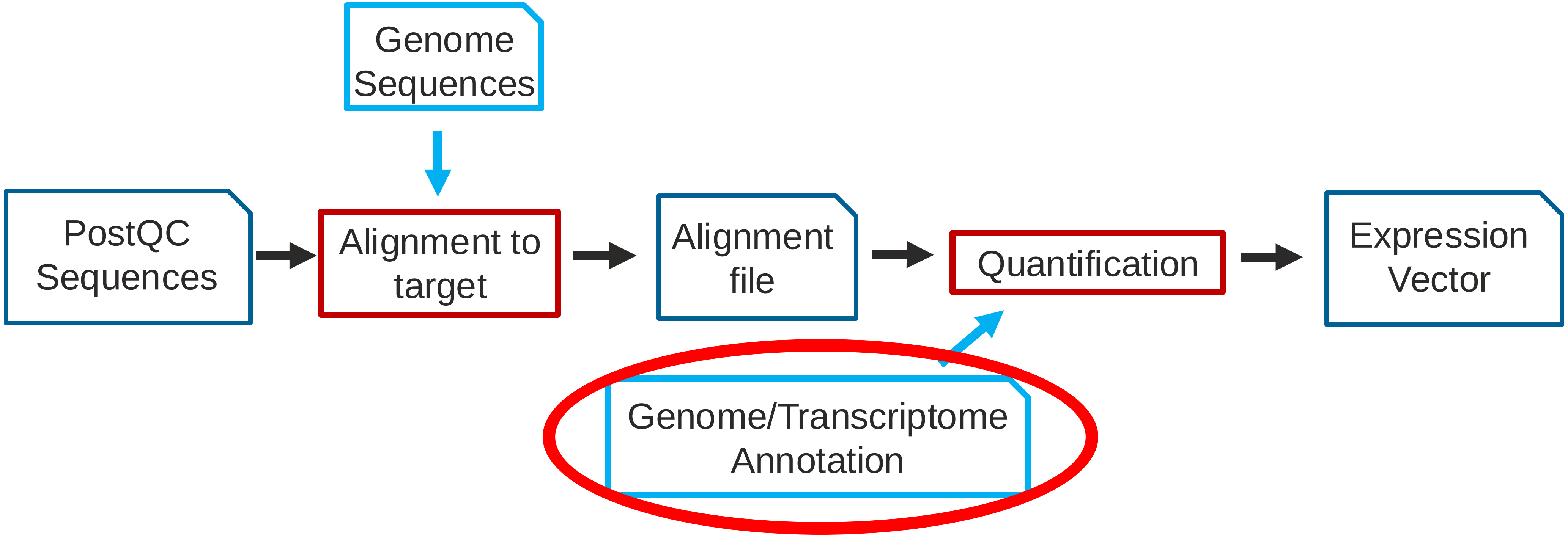
Do I need to search for new genes/transcripts?

1. “Alignment => Assignment” or
2. “Alignment Free”

MDI Biological Laboratory

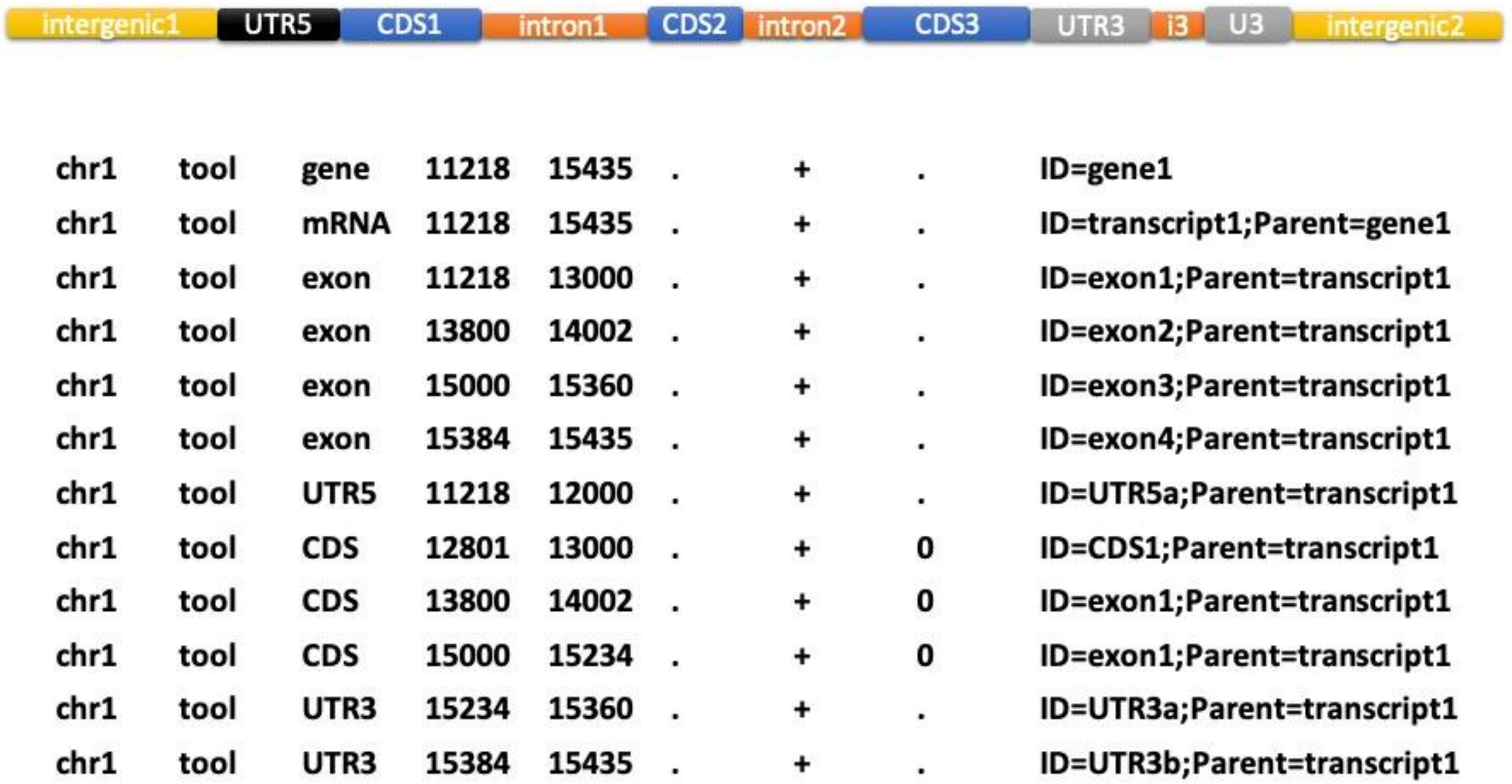
Pioneering new approaches in regenerative medicine

Alignment Approach to Quantification



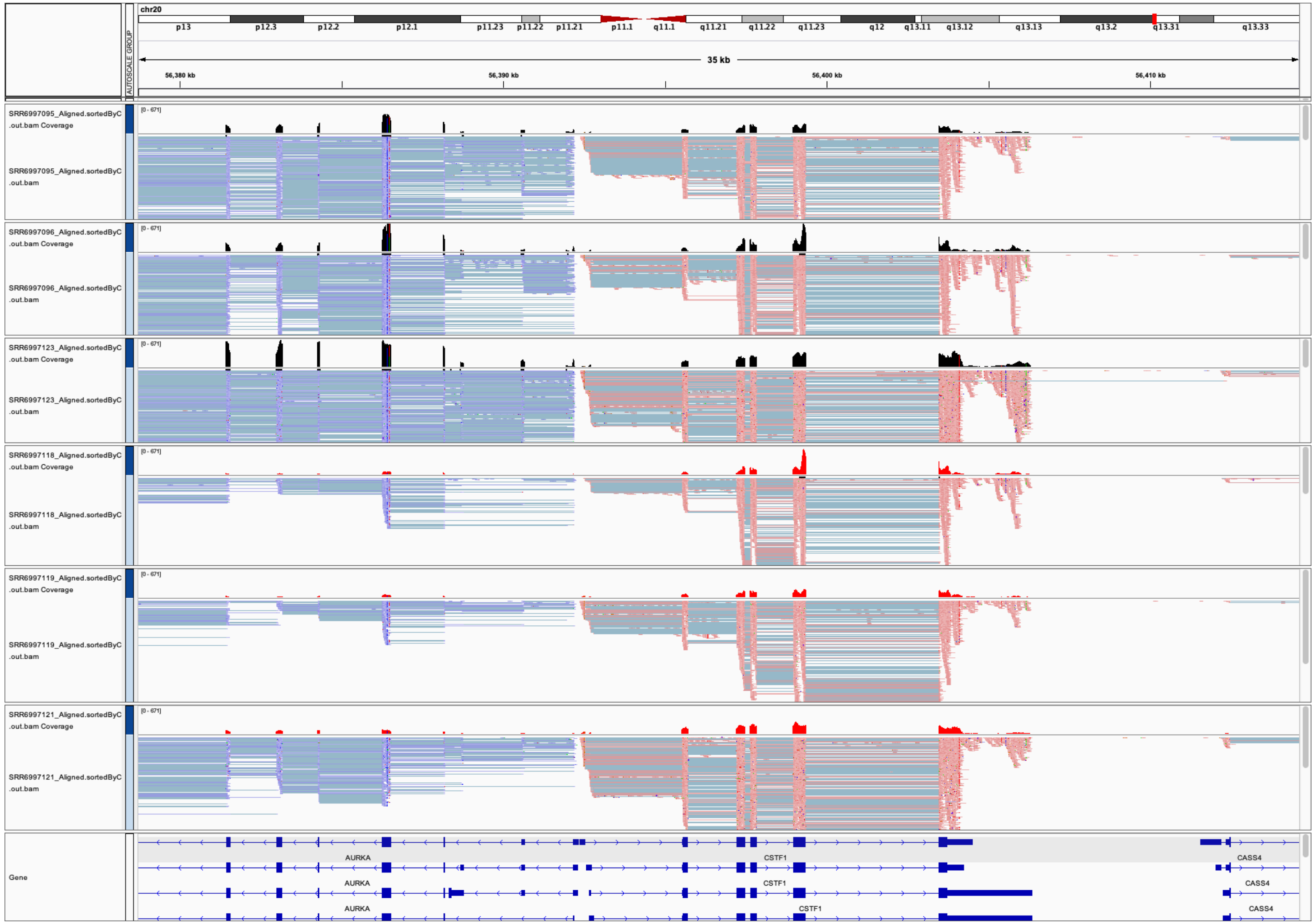
Gene information must be provided (e.g., GFF)

GFF example of a gene and its graphical representation



```
chr1    tool    gene    11218   15435   .       +       .       ID=gene1
chr1    tool    mRNA    11218   15435   .       +       .       ID=transcript1;Parent=gene1
chr1    tool    exon    11218   13000   .       +       .       ID=exon1;Parent=transcript1
chr1    tool    exon    13800   14002   .       +       .       ID=exon2;Parent=transcript1
chr1    tool    exon    15000   15360   .       +       .       ID=exon3;Parent=transcript1
chr1    tool    exon    15384   15435   .       +       .       ID=exon4;Parent=transcript1
chr1    tool    UTR5    11218   12000   .       +       .       ID=UTR5a;Parent=transcript1
chr1    tool    CDS     12801   13000   .       +       0       ID=CDS1;Parent=transcript1
chr1    tool    CDS     13800   14002   .       +       0       ID=exon1;Parent=transcript1
chr1    tool    CDS     15000   15234   .       +       0       ID=exon1;Parent=transcript1
chr1    tool    UTR3    15234   15360   .       +       .       ID=UTR3a;Parent=transcript1
chr1    tool    UTR3    15384   15435   .       +       .       ID=UTR3b;Parent=transcript1
```

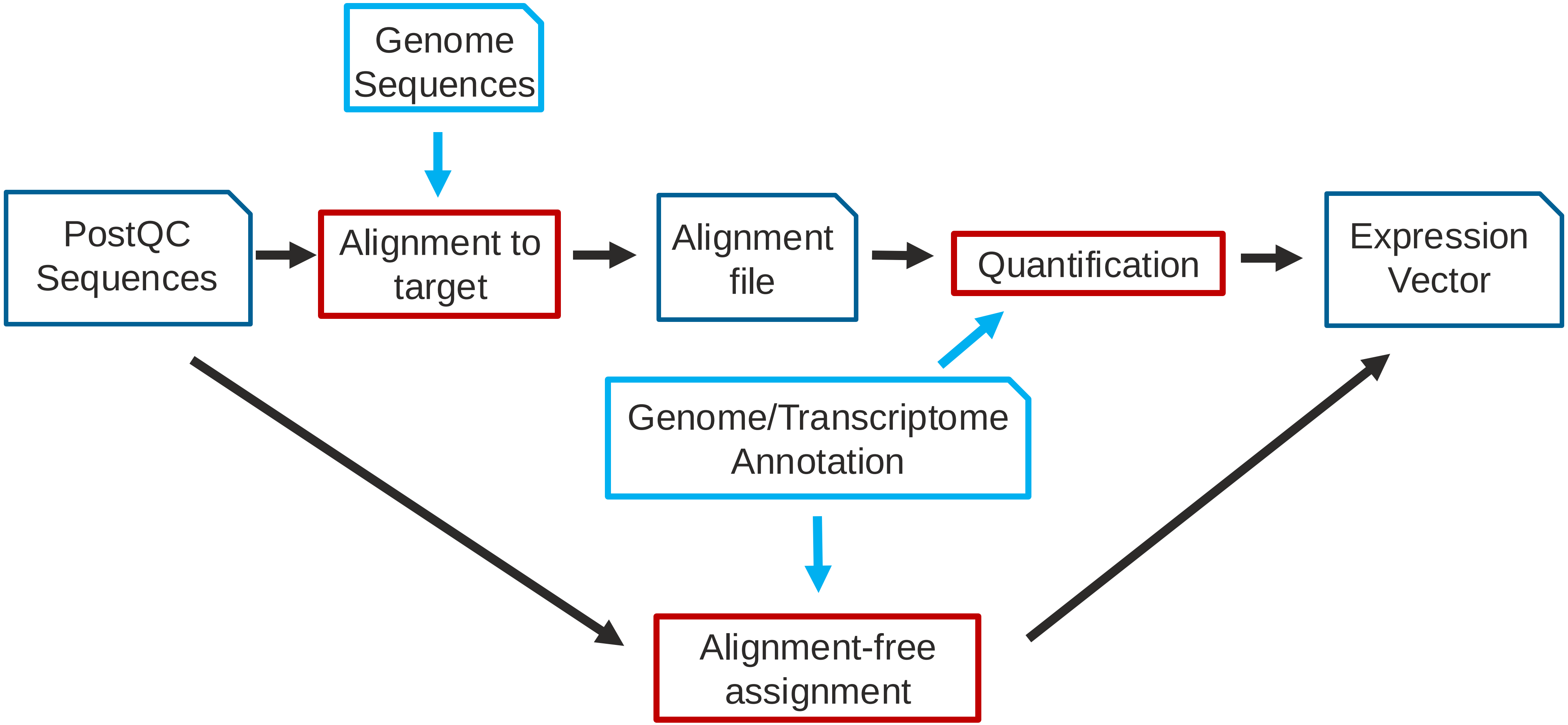
RNA-seq analysis: alignment/quantification


chr20
p13
p12.3
p12.2
p12.1
p11.23
p11.22
p11.21
p11.1
q11.1
q11.21
q11.22
q11.23
q12
q13.11
q13.12
q13.13
q13.2
q13.31
q13.33
AUTOSCALE GROUP
35 kb
56,380 kb
56,390 kb
56,400 kb
56,410 kb
SRR6997095_Aligned.sortedByC .out.bam Coverage
[0 - 671]
SRR6997095_Aligned.sortedByC .out.bam
SRR6997096_Aligned.sortedByC .out.bam Coverage
[0 - 671]
SRR6997096_Aligned.sortedByC .out.bam
SRR6997123_Aligned.sortedByC .out.bam Coverage
[0 - 671]
SRR6997123_Aligned.sortedByC .out.bam
SRR6997118_Aligned.sortedByC .out.bam Coverage
[0 - 671]
SRR6997118_Aligned.sortedByC .out.bam
SRR6997119_Aligned.sortedByC .out.bam Coverage
[0 - 671]
SRR6997119_Aligned.sortedByC .out.bam
SRR6997121_Aligned.sortedByC .out.bam Coverage
[0 - 671]
SRR6997121_Aligned.sortedByC .out.bam
Gene
AURKA
CSTF1
CASS4
AURKA
CSTF1
CASS4
AURKA
CSTF1
CASS4

MDI Biological Laboratory

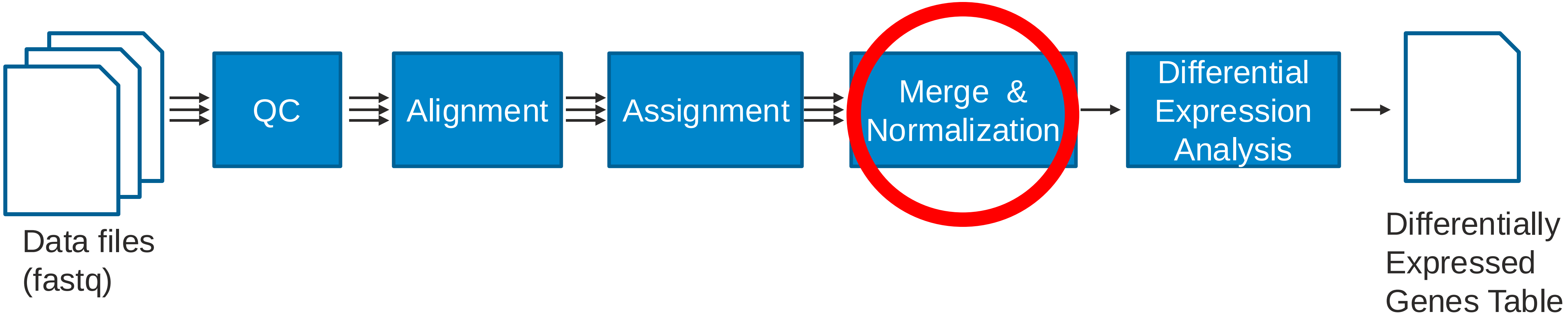
Pioneering new approaches in regenerative medicine

Alternative approaches to Quantification



MDI Biological Laboratory
Pioneering new approaches in regenerative medicine

A workflow for RNA-seq Differential Expression



Data files (fastq)

QC

Alignment

Assignment

Merge & Normalization

Differential Expression Analysis

Differentially Expressed Genes Table

MDI Biological Laboratory

Pioneering new approaches in regenerative medicine

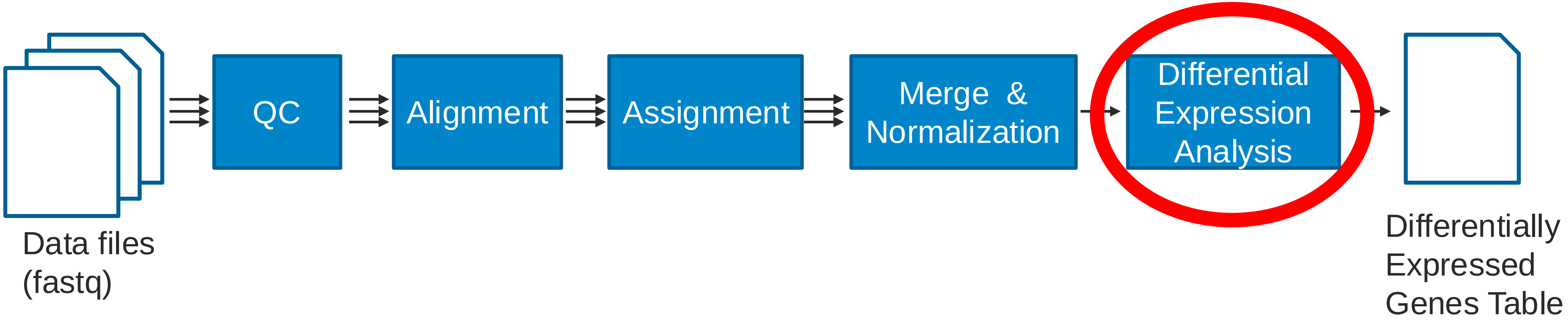
After expression is assessed in each sample, they are merged into a “count matrix”

gene_name	AL_TO_rep01	AL_TO_rep02	AL_TO_rep03	DR_TO_rep01	DR_TO_rep02	DR_TO_rep03
aap-1	753	747	743	940	947	982
aat-1	27	24	14	15	28	14
aat-2	30	33	24	60	65	68
aat-3	134	137	127	78	67	93
aat-4	23	45	35	22	30	27
aat-5	38	33	29	123	84	105
aat-6	40	39	28	41	46	55
aat-7	1	1	0	2	4	6
aat-8	1	1	2	14	3	10
aat-9	362	399	374	370	328	370

Computational normalization is critical for transcriptome analysis

- Three standard approaches to computational normalization
  - Internal normalization (Quantile, VST, FPKM, TPM, etc)
    - Assume all samples are roughly the “same,” and force equal distributions
    - Insensitive to global changes
  - Internal standard normalization
    - Identify a relatively small number of “unchanging” targets and scale all values so that these values are equal in all samples
  - External standard normalization
    - Add a known control (“Spike-in”) and then scale values such that the values for the controls are the same

A workflow for RNA-seq Differential Expression


Data files
(fastq)
QC
Alignment
Assignment
Merge & Normalization
Differential Expression Analysis
Differentially Expressed Genes Table

To interpret our count matrix, we need an Experimental Design File

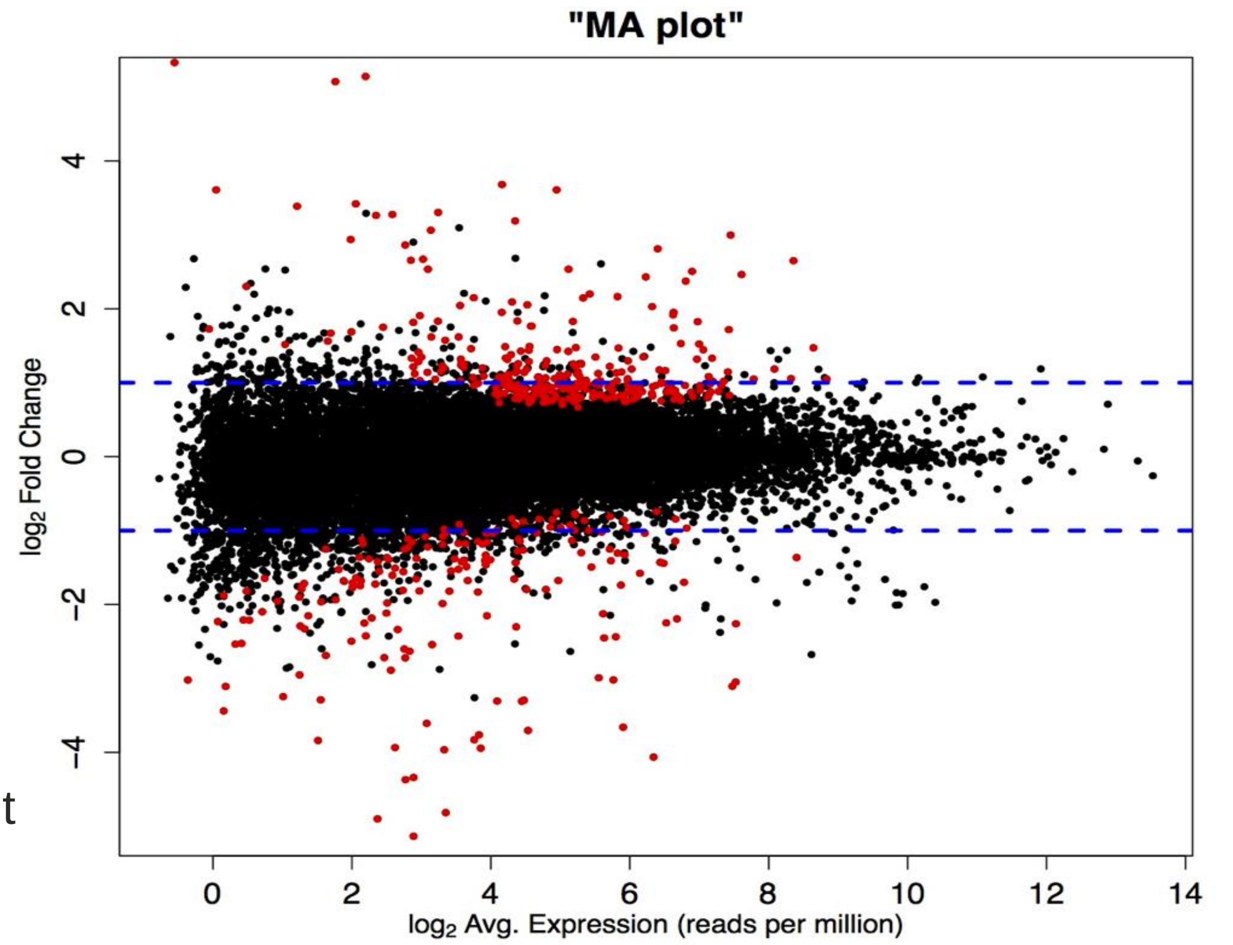
- At minimum, the Design File must contain
  - Identifiers for each sample (ideally matched to a data filename)
  - Assignment of all experimental parameters under consideration to each sample
- Ideally- ANY feature/variable that might vary between samples

sample	treatment	rep
AL_TO_rep01	AL	rep01
AL_TO_rep02	AL	rep02
AL_TO_rep03	AL	rep03
DR_TO_rep01	DR	rep01
DR_TO_rep02	DR	rep02
DR_TO_rep03	DR	rep03

In the end, a table of DE Gene Scores (*e.g.*, with DESeq2)

id	baseMean	log2FoldChange	lfcSE	stat	pvalue	padj
aagr-1	269.129364535602	-1.7442675672456	0.117789943380256	-14.8082893767481	1.29494023689411E-49	4.77497892947039E-48
aagr-3	2008.77205021688	-0.150425067741619	0.0418534931952695	-3.59408632965959	0.0003255318965062	0.00115773135585242
aak-2	243.639422569596	0.278051661358966	0.118395785760709	2.34849289248301	0.0188495589454439	0.0458599158655762
aakb-2	415.838439463941	0.561118701249279	0.100734483891487	5.57027424544835	2.54338675055636E-08	1.56247902940004E-07
aakg-1	365.852541550914	0.50046549824763	0.0971820567244253	5.149772654707	2.6080239032197E-07	1.40999077665866E-06
aakg-3	14.7626365586319	1.32753612196484	0.538581116196545	2.46487684406741	0.0137060352076937	0.034635107180592
aakg-4	72.0407425048923	1.73861272918138	0.251164464315594	6.92220825871598	4.44656882831695E-12	3.78784449623833E-11
aakg-5	736.490245516047	-0.171877365357521	0.063262738817093	-2.71688150989571	0.00659001957329092	0.018076559672254
aap-1	846.749244306947	0.216032066870877	0.0694242604499855	3.11176619629258	0.00185971722699245	0.00572240163660642
aars-2	2065.39673387659	0.132015428540962	0.0446828104046726	2.9545014591821	0.0031317467246952	0.00916240085776832
aat-2	45.7630124225589	1.01639572482027	0.300017225171763	3.38779123178136	0.000704578716201275	0.00236338649524766

The end result for all genes (in visual form)



Increased by treatment

Decreased by treatment

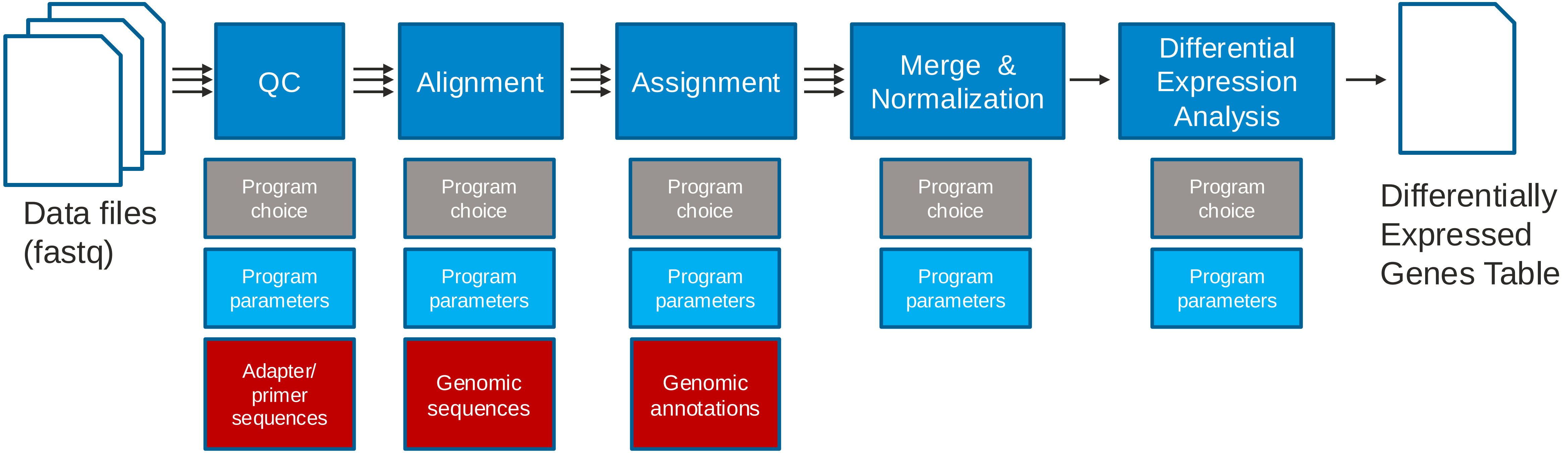
Increased average expression across all samples

MDI Biological Laboratory

Pioneering new approaches in regenerative medicine

What goes into the complete analysis of a genome-scale data set?

(using Bulk RNA-seq as an example)



Data files (fastq)

QC

Alignment

Assignment

Merge & Normalization

Differential Expression Analysis

Differentially Expressed Genes Table

Program choice

Program parameters

Adapter/ primer sequences

Program choice

Program parameters

Genomic sequences

Program choice

Program parameters

Genomic annotations

Program choice

Program parameters

Program choice

Program parameters

MDI Biological Laboratory

Pioneering new approaches in regenerative medicine

“Rigor and Reproducibility”

- Every choice outlined in the last slide can impact results of analysis
- Recording, monitoring, and sharing these factors is now recognized as critical in genomics analysis
    - A required aspect of all NIH grant proposals
    - Also required by many journals
- Resource: Karl Broman (Wisconsin)
    - http://kbroman.org/steps2rr/
    - http://kbroman.org/dataorg/pages/resources.html

Pipelines solve many issues

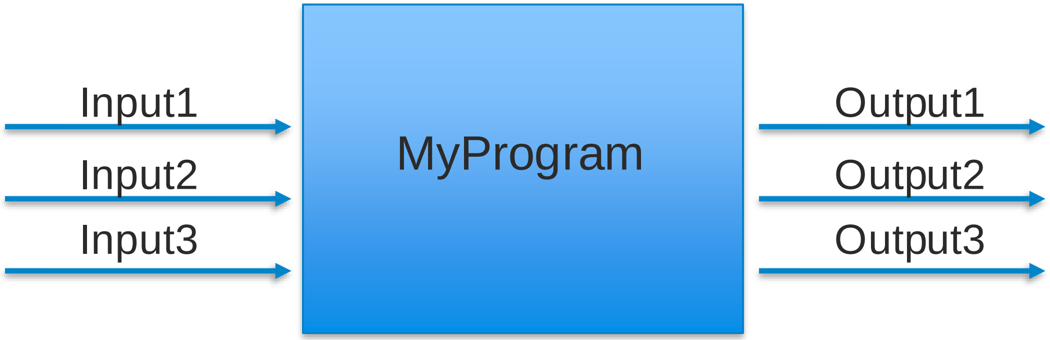
- Easy reproducibility of results
- Reduction in human error
- Organization of output
- Reduced work in program installation, maintenance, and troubleshooting

Basics of Workflow Systems

- A workflow system consists of
  - A language capable of describing the process that captures dependencies and computational complexities
  - A program (“engine”) capable of
    - Reading and executing the workflow description
    - Requesting/allocating the necessary computational resources to carry out the work
- The power of these systems is that workflows
  - Can be run on any system for which an engine has been programmed and set up
  - Can be rerun for new data sets and/or analysis by changing a simple text-formatted parameter file

Basics of Workflow definition languages

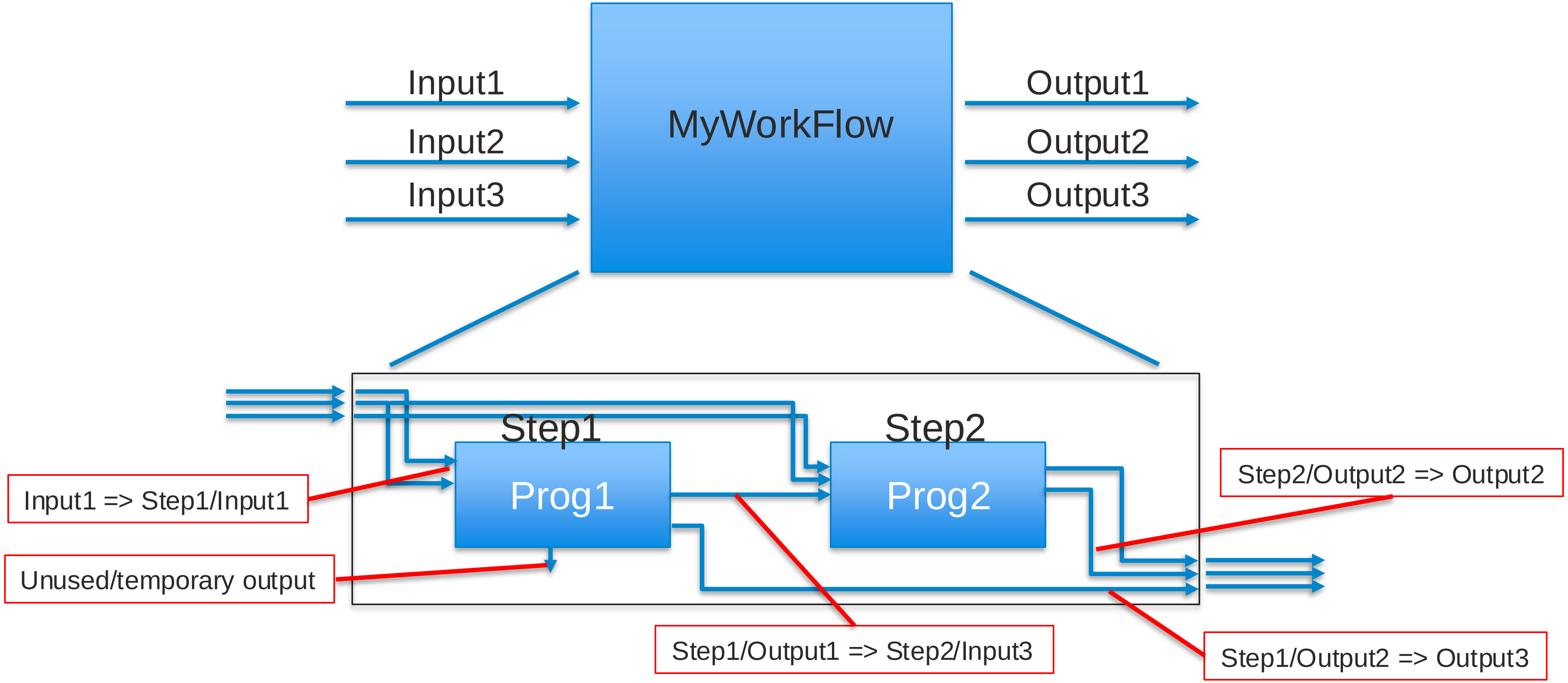
- At the base level is a single command, wrapped to accept input and generate output



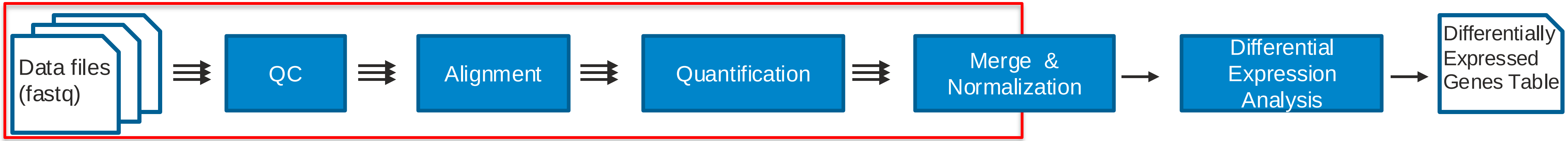
- This structure (including dependencies) is captured in the workflow

Basics of workflow definition languages

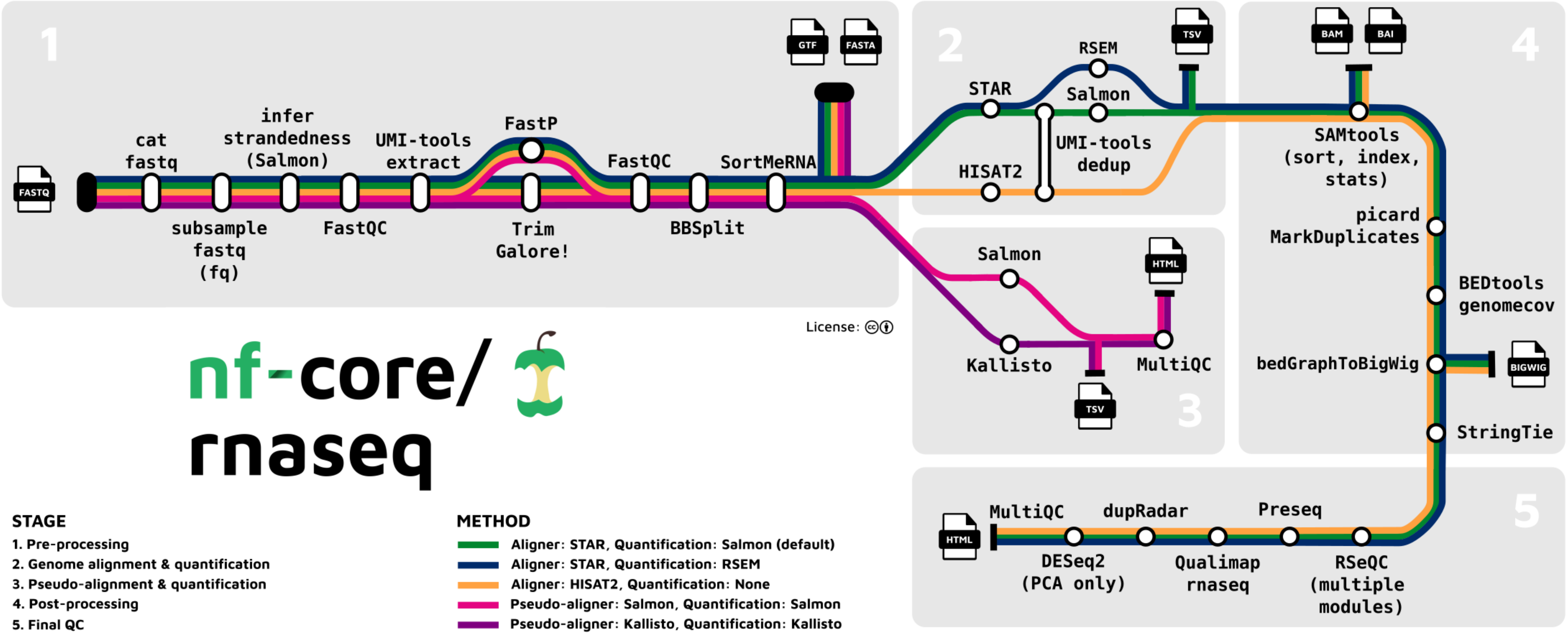
- Workflows are built from multiple steps, encapsulated in “modules”



MDI Biological Laboratory
Pioneering new approaches in regenerative medicine



Data files (fastq)
QC
Alignment
Quantification
Merge & Normalization
Differential Expression Analysis
Differentially Expressed Genes Table



nf-core/ rnaseq

STAGE

1. Pre-processing
2. Genome alignment & quantification
3. Pseudo-alignment & quantification
4. Post-processing
5. Final QC

METHOD

- 🟩 Aligner: STAR, Quantification: Salmon (default)
- 🟦 Aligner: STAR, Quantification: RSEM
- 🟧 Aligner: HISAT2, Quantification: None
- 🟥 Pseudo-aligner: Salmon, Quantification: Salmon
- 🟪 Pseudo-aligner: Kallisto, Quantification: Kallisto

MDI Biological Laboratory

Pioneering new approaches in regenerative medicine

Community supported workflows: NextFlow/NF-core

- https://nf-co.re/
- Nf-core Pipelines are
  - (Mostly) focused on specific data type
  - Supported by teams of volunteers
  - A systematic way to get systematic execution, logging, and organized output
  - Generally “best-practice” accepted steps

After the NF-core: working with your output

- NF-core pipelines generally focus on the standard common analysis step
- Many summary output files are available
- Output tables can become input to other tools
    - RNA-seq analysis with Sequin
    - https://sequin.ncats.io/app/

Summary and concluding thoughts

- Workflows allow for systematic and reproducible execution of complex, multi-step analysis of genome-scale data
- Community-supported workflows let you
  - Carry out best-in-practice analysis plans
  - Reduce effort and potential error
  - Keep track of analysis steps and output for subsequent downstream analysis and reporting/publication
- The learning curve is still not trivial
  - We can help

Thank you!

MDI Biological Laboratory

Pioneering new approaches in regenerative medicine