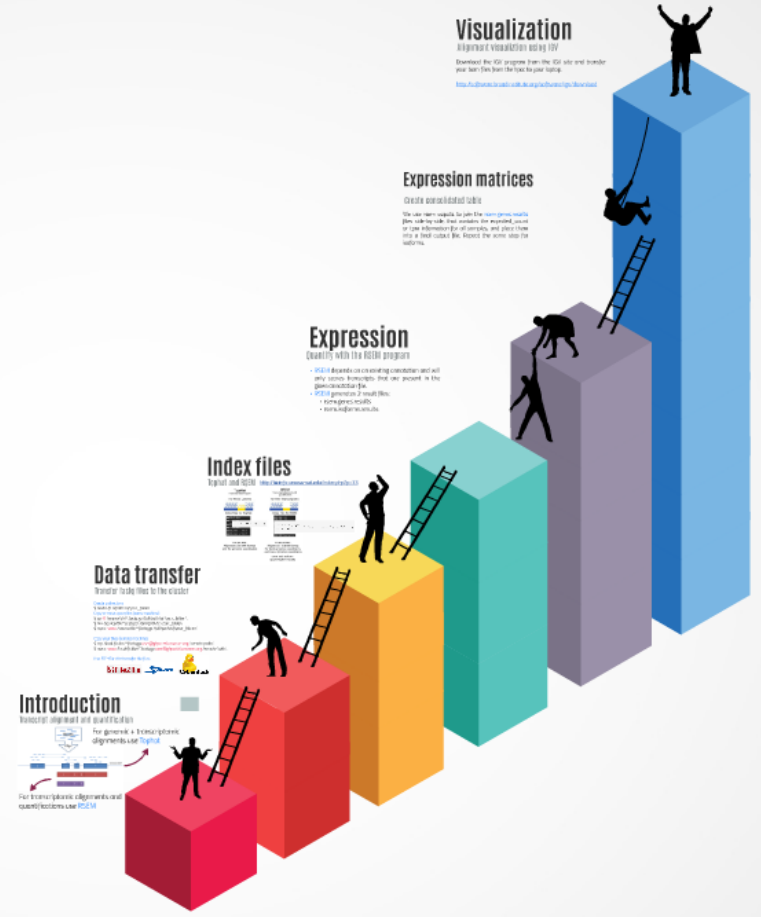
RNA-Seq

Data processing

03/07/2019



University of Massachusetts Medical School
UMASS

Alper Kucukural
& Onur Yukselen

Bioinformatics Core

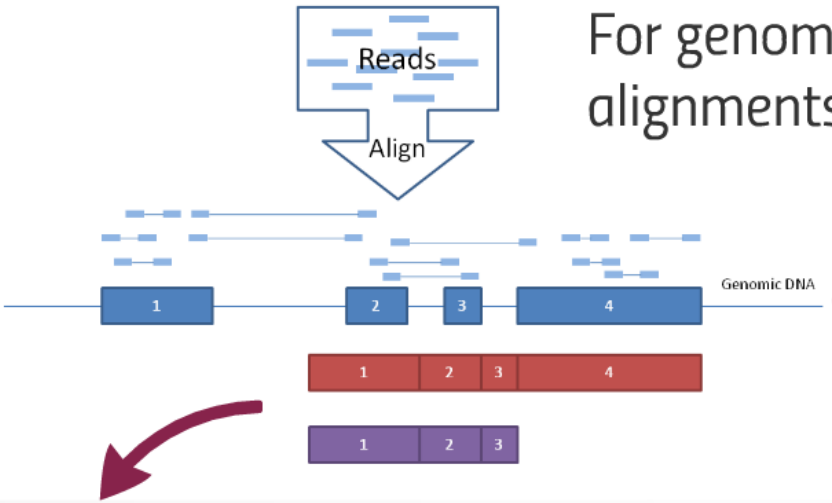
$ rsync -vazu /local/folder/*.fastq.gz user@ghpcc06.umassrc.org:/remote/path/.

Use FTP Client to transfer the files:

FileZilla WinSCP Cyberduck

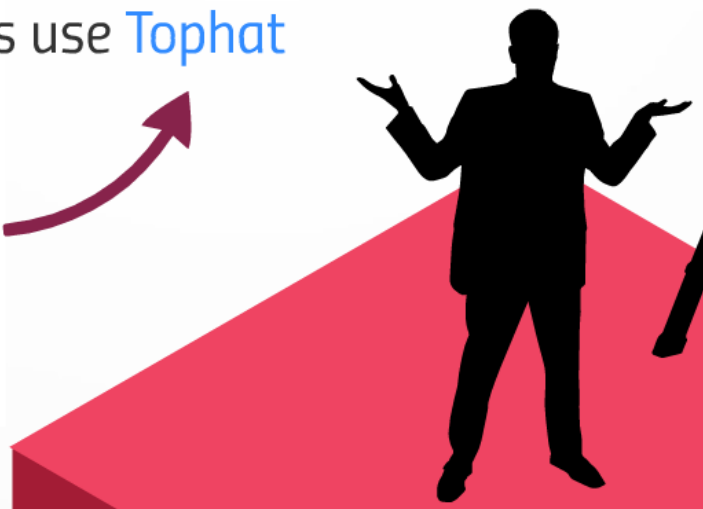
Introduction

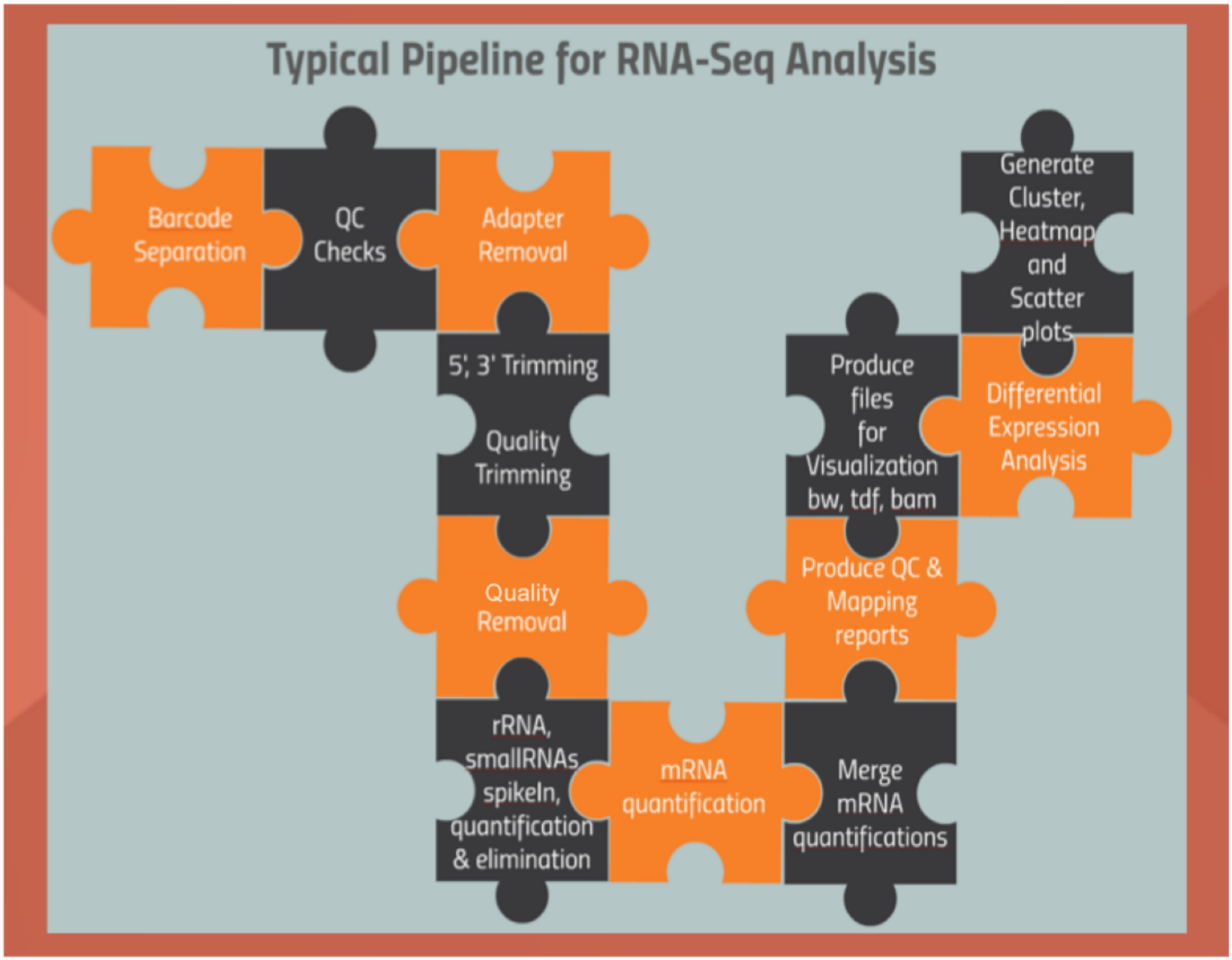
Transcript alignment and quantification



For genomic + transcriptomic alignments use Tophat

For transcriptomic alignments and quantifications use RSEM




Typical Pipeline for RNA-Seq Analysis
Barcode Separation
QC Checks
Adapter Removal
5', 3' Trimming
Quality Trimming
Quality Removal
rRNA, smallRNAs spikeIn, quantification & elimination
mRNA quantification
Merge mRNA quantifications
Produce QC & Mapping reports
Produce files for Visualization bw, tdf, bam
Generate Cluster, Heatmap and Scatter plots
Differential Expression Analysis

```
bowtie2-build -f mm10.fa mm10
```

```
mm10.rev.1.bt2
mm10.rev.2.bt2
```

Output file:
Alignments in BAM format
only for genomic coordinates

```
rsem-prepare-reference \
--gtf ucsc.gtf --transcript-to-gene-map ucsc_into_genesymbol.rsem \
mm10.fa mm10.rsem
```

Output files:
Alignments in BAM format
for both genomic coordinates
and transcriptomic coordinates

Gene and isoform
quantification results

Data transfer

Transfer fastq files to the cluster

Create a directory:
$ mkdir /full/path/of/your_folder
Copy or move your files (same machine):
$ cp -R /source/dir/*.fastq.gz /full/path/of/your_folder/.
$ mv /source/dir/*.fastq.gz /full/path/of/your_folder/.
$ rsync -vazu /source/dir/*.fastq.gz /full/path/of/your_folder/.

Copy your files (remote machine):
$ scp /local/folder/*.fastq.gz user@ghpcc06.umassrc.org:/remote/path/.
$ rsync -vazu /local/folder/*.fastq.gz user@ghpcc06.umassrc.org:/remote/path/.

Use FTP Client to transfer the files:

FileZilla WinSCP Cyberduck

ction

nd quantification

- rsem.genes.results
- rsem.isoforms.results

Index files

Tophat and RSEM http://bioinfo.umassmed.edu/index.php?p=33

TopHat
A spliced read mapper

For Whole Genome

Index files for TopHat:

```
mm10.1.bt2
mm10.rev.1.bt2
mm10.rev.2.bt2
```

```
bowtie2-build -f mm10.fa mm10
```

Output file:
Alignments in BAM format
only for genomic coordinates

RSEM
Transcript alignment and quantification

For Only Transcriptome

Index files for RSEM:

```
mm10.rsem.1.bt2
mm10.rsem.2.bt2
mm10.rsem.3.bt2
mm10.rsem.transcripts.fa
```

```
rsem-prepare-reference \
--gtf ucsc.gtf --transcript-to-gene-map ucsc_into_genesymbol.rsem \
mm10.fa mm10.rsem
```

Output files:
Alignments in BAM format
for both genomic coordinates
and transcriptomic coordinates

Gene and isoform
quantification results

Expression

Quantify with the RSEM program

- RSEM depends on an existing annotation and will only scores transcripts that are present in the given annotation file.
- RSEM generates 2 result files:
  - rsem.genes.results
  - rsem.isoforms.results

S

nfo.umassmed.edu/index.php?p=33

Expression matrices

Create consolidated table

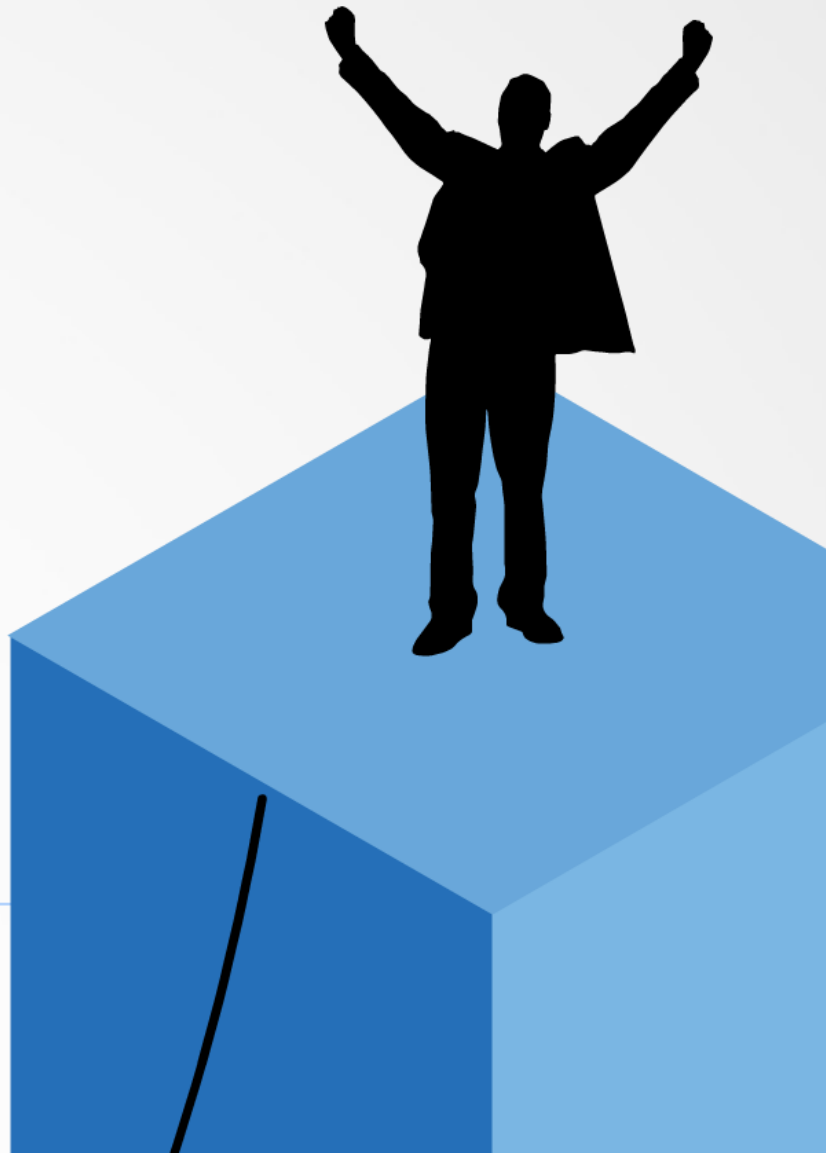
We use rsem ouputs to join the rsem.genes.results files side-by-side, that contains the expected_count or tpm information for all samples, and place them into a final output file. Repeat the same step for isoforms.

Visualization

Alignment visualiztion using IGV

Download the IGV program from the IGV site and transfer your bam files from the hpcc to your laptop.

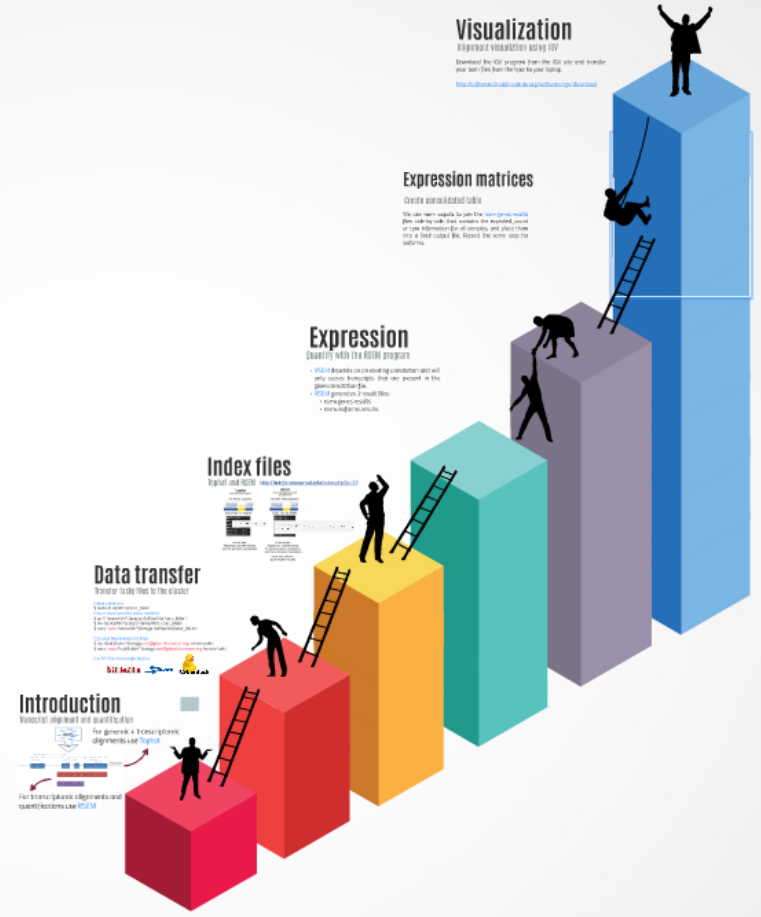
http://software.broadinstitute.org/software/igv/download



RNA-Seq

Data processing

03/07/2019



University of Massachusetts Medical School
UMASS

Alper Kucukural
& Onur Yukselen

Bioinformatics Core