Анализ транскриптома клеток, инфицированных вирусом гриппа.

Студент: Баранов Я.А.
Научный руководитель: Васин А.В., к.б.н.
Лаборатория структурной и функциональной протеомики, НИИ Гриппа

Задача

Проанализировать экспрессию генов в норме и после воздействия вируса

Что было

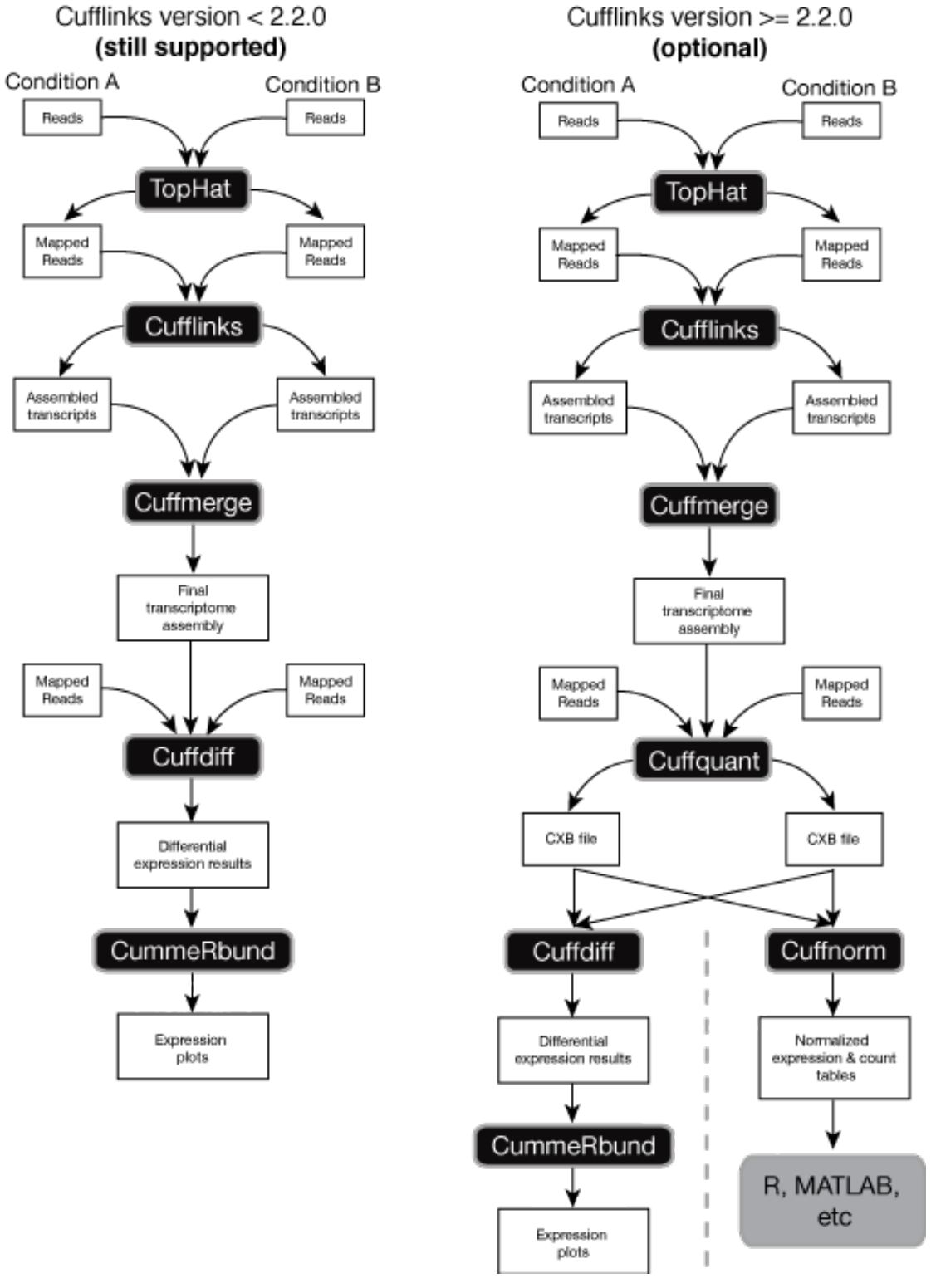
Клеточная линия A549 (human lung carcinoma)

Два образца:

- контроль
- образец был заражен реассортантным вирусом
  6/2 A/Perth/16/09 (H3N2)

Что сделано

- QC сборки
  (отрезали адаптеры и конец рида)
- Собран транскриптом
  (pipeline: tuxedo workflow
  референс и аннотация: hg19)
- Найдены гены, которые изменились значимо (CummeRbund)


Cufflinks version < 2.2.0
(still supported)
Condition A
Condition B
Reads
Reads
TopHat
Mapped Reads
Mapped Reads
Cufflinks
Assembled transcripts
Assembled transcripts
Cuffmerge
Final transcriptome assembly
Mapped Reads
Mapped Reads
Cuffdiff
Differential expression results
CummeRbund
Expression plots
Cufflinks version >= 2.2.0
(optional)
Condition A
Condition B
Reads
Reads
TopHat
Mapped Reads
Mapped Reads
Cufflinks
Assembled transcripts
Assembled transcripts
Cuffmerge
Final transcriptome assembly
Mapped Reads
Mapped Reads
Cuffquant
CXB file
CXB file
Cuffdiff
Cuffnorm
Differential expression results
Normalized expression & count tables
CummeRbund
R, MATLAB, etc
Expression plots

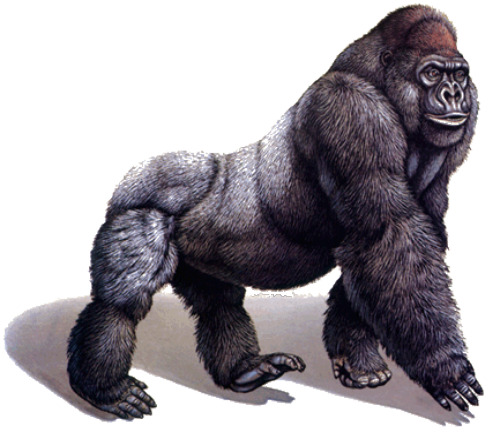
Single replicate RNA-seq data с которой tuxedo protocol не работает

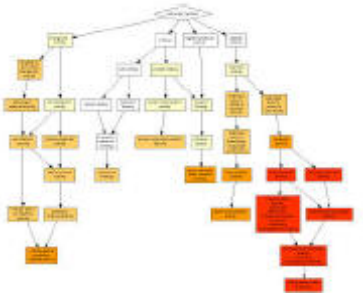
Новый пайплайн

1) GFOLD (generalized fold change) algorithm

2) topGO:
   Enrichment analysis for Gene Ontology
   (R package)

3) Gorilla
   Gene **O**ntology en**RI**chment ana**L**ysis and visua**L**iz**A**tion tool





Что сделано сейчас

- Получены результаты подсчета дифф. экспрессии
- Обработка с помощью R в процессе освоения

JOB count:

GeneSymbol:

For BED file, this is the 4'th column. For GPF file, this is the first column. For GTF format, this corresponds to 'gene_id' if it exists, 'NA' otherwise.

GeneName:

For BED file, this is always 'NA'. For GPF file, this is the 12'th column. For GTF format, this corresponds to 'gene_name' if it exists, 'NA' otherwise.

Read Count:

The number of reads mapped to this gene.

Gene exon length:

The length sum of all the exons of this gene.

RPKM:

The expression level of this gene in RPKM.

JOB diff:

GeneSymbol:

Gene symbols. The order of gene symbol is the same as what appearing in the read count file.

GeneName:

Gene name. The order of gene name is the same as what appearing in the read count file.

GFOLD:

GFOLD value for every gene.
The GFOLD value could be considered as a reliable log2 fold change.
It is positive/negative if the gene is up/down regulated.
The main usefulness of GFOLD is to provide a biological meanlingful ranking of the genes.
The GFOLD value is zero if the gene doesn't show differential expression.

E-FDR:

Empirical FDR(False Discovery Rate) based on replicates. It is always 1 when no replicates are available.

log2fdc:

log2 fold change. If no replicate is available, and -acc is T, log2 fold change is based on read counts and normalization constants. Otherwise, log2 fold change is based on the sampled expression level from the posterior distribution.

1stRPKM:

The RPKM for the first condition. It is available only if gene length is available. If multiple replicates are available, the RPKM is calculated simply by summing over replicates. Because RPKM is acturally using sequencing depth as the normalization constant, log2 fold change based on RPKM could be different from the log2fdc field.

2ndRPKM:

The RPKM for the second condition. It is available only if gene length is available. Please refer to 1stRPKM for more information.

Дальнейшие планы

- Оценить биологическй смысл изменения экспрессии этих генов
- Проверить экспрессию с помощью ПЦР
- Исходя из предыдущих пунктов, внести поправки в дизайн wet-lab эксперимента
- Сделать больше образцов

Спасибо за внимание !

Вопросы, комментарии, советы?