Review Your Peers: Application Challenge 2

Review by August 2, 11:59 PM PDT

Reviews 5 left to complete

RNAseq gene expression

by Rita Seeböck

July 31, 2017

like Flag this submission

PROMPT

Measuring the size of the BRAIN1 Dataset

We will begin with a series of questions analyzing the STAR+DESeq workflow. To find the results of this RNA Express workflow, click on “Projects,” then click on the “RNA-seq Methods Results” project you joined earlier. There should be a folder called “RNA Express 07/14/2015 1:12:15” (i.e. “RNA Express,” followed by the execution date, followed by the execution time).

Question 1: How many reads are in each of the four brain samples? What are the lengths of “Read 1” and “Read 2”? After learning how genome assembly algorithms work, what do you think that the terms “Read 1” and “Read 2” are referring to?

- Reads per Brain sample:
  1 = 97,730,535
  2 = 94,064,211
  3 = 83,374,339
  4 = 84,897,013
- Lengths of Read 1 and Read 2:
  75
- read 1 and read2 may refer to the forward and reverse sequencing

RUBRIC

Read totals:

- BRAIN1 (mRNA-Brain-C4): 97,730,535 reads.
- BRAIN2 (mRNA-Brain-C6) has 94,064,211 reads.
- UHRR1 has 83,374,339 reads.

Help Center

- UHRR2 has 84,897,013 reads.

Read 1 and Read 2 both have length 75. “Read 1” and “Read 2” are the two reads contributing to a paired-end read (i.e. two reads that are separated by a distance approximately equal to the insert length).

- 0 pts
  Learner provided less than two correct values, and no explanation of Read 1 and Read 2 was provided
- 1 pt
  Learner provided more than two correct values, but no explanation of Read 1 and Read 2 was provided
- 2 pts
  All provided values are correct, as well as the explanation

PROMPT

Quantifying unambiguously aligned reads

Fill in the following information from the STAR+DESeq workflow (found in the Summary Tab). The workflow outputs the following statistics.

“% Unaligned” refers to the percentage of reads that were filtered out by STAR because they were poorly aligned, meaning that they could not be aligned with a small number mismatches/insertions/deletions.

“% Abundant” refers to the percentage of reads that align to “abundant” transcripts, such as transcripts arising in mitochondrial and ribosomal genes.

“Multi-Mapped” refers to the percentage of aligned reads that have more than one equally good alignment position in the genome.

“Reads with Spliced Alignment” refers to the percentage of aligned paired-end reads whose pairs align to multiple positions in the genome, e.g., one half of the read aligns to a position in the genome far away from the other. Such reads often represent reads that align to multiple exons.

Question 2: Fill in the following table.

	%Unaligned	%Abundant	Multi-Mapped	Reads with Spliced Alignment
BRAIN1				
BRAIN2				
UHRR1				
UHRR2				

%Unaligned % Abundant % MulitMapped (% Aligned Reads) Reads with Spliced Alignment (%Aligned Reads)
BRAIN1 6,63 % 17,32 % 4,31 % 18,70 %
BRAIN2 2,68 % 17,18 % 3,55 % 17,60 %
UHRR1 2,77 % 9,51 % 8,44 % 24,20 %
UHRR2 2,06 % 8,44 % 7,65 % 22,50 %

Help Center

RUBRIC

	%Unaligned	%Abundant	Multi-Mapped	Reads with Spliced Alignment
BRAIN1	6.63%	17.32%	4.31%	18.70%
BRAIN2	2.68%	17.18%	3.55%	17.60%
UHRR1	2.77%	9.51%	8.44%	24.20%
UHRR2	2.06%	8.44%	7.65%	22.50%

- 0 pts
  The learner did not fill in the table correctly.
- 1 pt
  The learner filled in all but one or two values correctly.
- 2 pts
  The learner filled in the table completely correctly.

PROMPT

Question 3: What could cause some RNA-seq reads to be mapped to multiple locations (corresponding to the multi-mapped metric)?

genetic motifs can occur in repeats, therefore they will be identified and mapped multiple times. Also when parts of the genome are rearranged within a subset of cells, this could manifest in multi-mapping.

RUBRIC

If a gene is duplicated in the genome and two gene copies are very similar, then RNA-seq reads from that gene would also map to the duplicated gene.

- 0 pts
  The learner did not answered correctly.
- 2 pts
  The learner answered correctly.

PROMPT

Visualizing the comparison of brain and control samples

Click on the “Output Files” tab on the left side of the page, then navigate to “RNAExpress-AppResult Files” --> “differential” --> “UHRR_vs_Brain”. Download the **heatmap** in the file “UHRR_vs_Brain.deseq.heatmap.pdf”, which is used to visualize the number of reads mapped to each gene in each sample. In particular, the numeric value of the colors (shown in the “color key” on the heatmap) are log-2 normalized values *counts*(*i,j*) holding the number of reads mapped to gene *i* in tissue *j*. For example, a bright red color indicates that on the order of 2^6 reads were mapped to a given gene in a tissue, whereas a bright yellow indicates the number of reads as on the order of 2^{16}. The log-2 normalized counts are produced via DESeq’s “regular log transformation” function, which is described in the DESeq manual (http://rpackages.ianhowson.com/bioc/DESeq2/man/rlog.html).

Help Center

Question 4: What does each row (horizontal line) in the heatmap represent? What does each column in the heatmap represent?

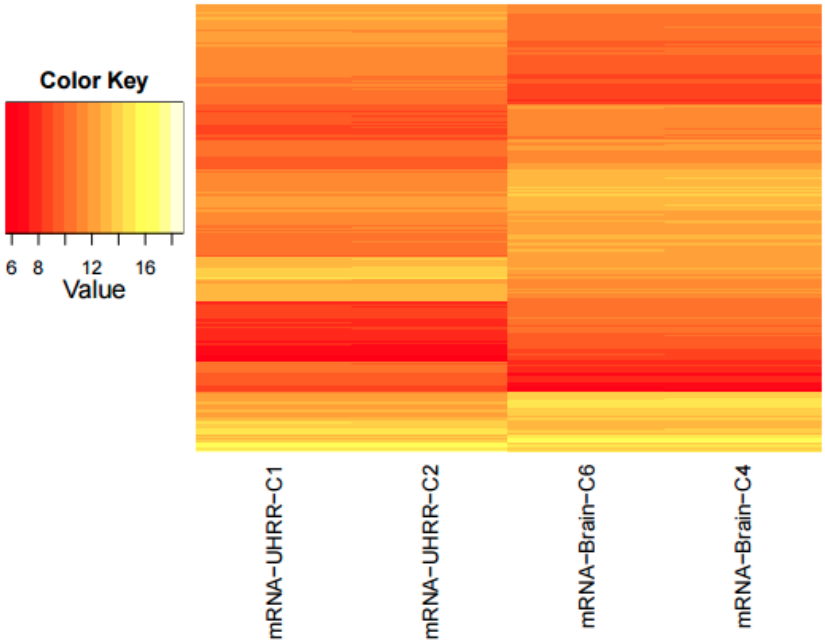
The lines are single genes, columns (2) are grouped for Brain samples and Control samples

RUBRIC

Each horizontal line in the heatmap corresponds to a gene. Each column in the heatmap corresponds to a sample.

- 0 pts
  The learner provided wrong answer for both questions.
- 1 pt
  The learner provided correct answer for one of the questions.
- 2 pts
  The learner provided correct answer for both questions.

PROMPT



Question 5: Why does the 4-column heatmap (reproduced above) look like it only has two columns?

Because Brain samples are very similar to each other as are the UHRR samples, when looking very closely differences can be identified.

Help Center

RUBRIC

The heat maps for the two control samples are almost identical, as is the case for the two experimental samples (it requires zooming in to even see the differences between the two). This is because the two control samples/two experimental samples are very similar to each other.

- 0 pts
  The learner did not answered correctly.
- 2 pts
  The learner answered correctly.

PROMPT

Identifying the most expressed genes in the brain samples

Close the heatmap file, and download the "UHRR_vs_Brain.deseq.counts.csv" file from the output file list, which we will use to answer the following questions about the most expressed genes. We start from the naïve (and often incorrect) assumption that the genes with the highest read counts represent the most expressed genes. (We will revisit this assumption soon.)

Question 6: What are the three genes with the highest read counts in BRAIN1, and what are their read counts? What are the three genes with the highest read counts in UHRR1, and what are their read counts? Do the two sets of genes overlap?

(Note: If you are using Microsoft Excel or OpenOffice to view the .csv file, you will need the "Sort" function in order to view each column from largest to smallest.)

Brain C4: MBP (670689)
GFAP (340606)
PLP1 (285541)
UHRR1: EEF1A1 (549587)
GAPDH (387474)
FN1 (323639)
The two setd do not overlap

RUBRIC

The three most expressed genes in BRAIN1 are MBP (670689 reads), GFAP (340606 reads), and PLP1 (285541 reads). The three most expressed genes in UHRR1 are EEF1A1 (549587 reads), GAPDH (387474 reads), and FN1 (323639 reads). They do not overlap.

If you're interested: here is some additional information on the six genes.

- MBP: The protein encoded by the MBP gene is a major constituent of the myelin sheath of oligodendrocytes and Schwann cells in the nervous system. The myelin sheath covers the axon of some nerve cells; myelin helps increase the speed of electrical signals in the nerve cell and (because it is an insulator) prevents the loss of these signals from the axon.
- GFAP: This gene encodes one of the major intermediate filament proteins of mature astrocytes, star-shaped cells that facilitate nerve cell function in the brain and spinal cord. It is used as a marker to distinguish astrocytes from other glial cells during development.

- PLP1: This gene encodes a transmembrane proteolipid protein that is the predominant component of myelin. The encoded protein may play a role in the compaction, stabilization, and maintenance of myelin sheaths, as well as in oligodendrocyte development and axonal survival.
- EEF1A1: This gene encodes an isoform of the alpha subunit of the elongation factor-1 complex, which is responsible for the enzymatic delivery of aminoacyl transfer RNAs to the ribosome. This isoform (alpha 1) is expressed in the brain, placenta, lung, liver, kidney, and pancreas, whereas the other isoform (alpha 2) is expressed in the brain, heart and skeletal muscle.
- GAPDH: This gene encodes a member of the glyceraldehyde-3-phosphate dehydrogenase protein family. The encoded protein has been identified as a moonlighting protein based on its ability to perform mechanistically distinct functions. The product of this gene catalyzes an important energy-yielding step in carbohydrate metabolism.
- FN1: This gene encodes fibronectin, a glycoprotein present in a soluble dimeric form in plasma, and in a dimeric or multimeric form at the cell surface and in extracellular matrix.

○ 0 pts
The learner provided more than two values or gene names correctly.

○ 1 pt
The learner provided all but one or two values or gene names correctly.

○ 2 pts
The learner answered correctly.

PROMPT

Question 7: What are the three most expressed genes in BRAIN2, and what are their read counts? Do they overlap with the three most expressed genes in BRAIN1? Are you surprised?

BRAINC6: MBP (648709)
GFAP (331318)
PLP1 (301100)
These two sets do overlap, which is not surprising, it fits the picture presented by the heatmap

RUBRIC

The three most expressed genes in BRAIN2 are MBP (648709 reads), GFAP (331318 reads), PLP1 (301100 reads). These are the same genes we found for BRAIN1, which is not surprising; they are both human brain samples, and so their expression profiles should be similar.

○ 0 pts
The learner provided more than two values or gene names incorrectly.

○ 1 pt
The learner provided all but one or two values or gene names correctly.

○ 2 pts
The learner answered correctly.

Help Center

PROMPT

Normalizing read counts

Question 8: We mentioned that drawing conclusions directly from read counts per gene is not ideal. What is the problem with such an inference? How can read counts be normalized to provide a more informative statistic?

There are several ways how read counts can be normalized. One could look for a gene that is expressed at constant levels throughout all analyzed cell types (typically genes that are required for the cells architecture of cytoskeleton, house-keeping-genes). One could refer to the ratio of genes compared to the reference gene.

RUBRIC

We expect longer genes to have more reads than shorter genes because there will be more fragments from longer genes after the DNA is fragmented. A simple way to normalize read counts is to divide the read counts for a gene by the length of the gene (other answers may exist).

- 0 pts
  The learner did not answered correctly.
- 2 pts
  The learner answered correctly or provided reasonable justification.

PROMPT

Comparing expression levels between control and experimental samples

After we have collected the expression levels of a variety of genes in an experimental sample and a control sample, we need some way of determining whether a given gene's difference in expression between the two samples is statistically significant. As we saw when analyzing *T. rex* peptides, biologists often measure statistical significance by asking, "What is the probability that the result of this experiment arose purely by chance?" Roughly speaking, this probability is called a **p-value**; for example, a p-value of 0.01 translates to the statement, "There is a 1% chance that the results occurred due to random chance" (see FAQ: What is a p-value?).

To compute p-values for a gene expression experiment, biologists often compare control and experimental samples. Given a gene *g* with expression levels across multiple samples in the control and experimental groups, biologists merge these values into two values: a control value, denoted *control*(*g*), and an experimental value, denoted *exp*(*g*). (See FAQ: How do Biologists Merge Expression Values from Different Samples?) We could therefore take the difference $|exp(g) - control(g)|$, and then ask how likely this measured difference in expression would be for the control gene measured against itself. However, this approach suffers from a flaw of scale; we should draw much different conclusions for the values $exp(g) = 10.001$ and $control(g) = 0.001$ than we would for the values $exp(g) = 1010$ and $control(g) = 1000$, even though $|exp(g) - control(g)| = 10$ in both cases.

As a result, instead of considering the difference $|exp(g) - control(g)|$, biologists often consider the ratio $exp(g)/control(g)$, which is called the **fold-change** of *exp*(*g*) to *control*(*g*). When *exp*(*g*) is larger than *control*(*g*), the fold-change is greater than 1; when *exp*(*g*) is smaller than *control*(*g*), the fold-change is less than 1 (but still greater than or equal to 0). Biologists then transform fold-changes into p-values, although the details of this transformation are beyond the scope of this Application Challenge.

Another popular metric is the base-2 logarithm of the fold-change, denoted *Log2FoldChange*. This way, when *exp*(*g*) is larger than *control*(*g*), *Log2FoldChange* is positive, and when *exp*(*g*) is smaller than *control*(*g*), *Log2FoldChange* is negative. We use 2 as the base of the logarithm so that *Log2FoldChange* can be easily interpreted as "doubles" or "halves"; that is, a *Log2FoldChange* of 1 implies a doubling of expression between the control and experimental samples, and a *Log2FoldChange* of -1 implies a halving of expression.

Help Center

Aggregating RNA-seq results over *n* genes results in an *n* x 4 matrix *C*, where *C*(*i*,*j*) is the number of reads mapped to gene *i* in sample *j* (effectively a *gene expression matrix*). We would like to use this matrix to derive *log2FoldChange* values for each of *n* genes. Of course, we have *four* samples (two brain and two control), yet *log2FoldChange* is the log-ratio of *two* numbers. To learn about the methods behind generating *log2FoldChange*, see Anders and Huber, 2010).

To find the *Log2FoldChange* values and p-values for the genes in our experiment, download the "UHRR_vs_Brain.deseq.res.csv" file in the "Output Files".

Question 9: Which gene had the most negative *Log2FoldChange* value? Search for this gene using your favorite search engine; what does it do in the cell, i.e., what is its **functional annotation**? Does it play a special role in the brain?

the most negative Log2FoldChange is identified with -10,9010262581618 for TYR; The enzyme encoded by this gene catalyzes the first 2 steps, and at least 1 subsequent step, in the conversion of tyrosine to melanin. The enzyme has both tyrosine hydroxylase and dopa oxidase catalytic activities, and requires copper for function. Mutations in this gene result in oculocutaneous albinism, and nonpathologic polymorphisms result in skin pigmentation variation. The human genome contains a pseudogene similar to the 3' half of this gene. So there is no special role of it within the brain.

RUBRIC

TYR had the smallest *log2FoldChange* value. This gene encodes an enzyme that catalyzes the conversion of tyrosine to melanin. It is related to various metabolic pathways, so it is unlikely that the gene plays a special role in the brain.

- 0 pts
  The learner provided wrong answer for all questions.
- 1 pt
  The learner provided correct answer for one of the questions.
- 2 pts
  The learner provided correct answer for all questions.

PROMPT

Note that many genes have a p-value of zero, which represents computational artifacts (rather than genes with abnormally high/low expression in the brain samples as compared to control) and should be ignored.

Question 10: What is the name and function of the gene with the lowest non-zero (and non-"NaN") p-value in this dataset? Does it make sense that this gene has such a low p-value?

Note: "NaN" stands for "Not a Number". For more details, see FAQ: What is NaN?

The lowest pvalue (4,75904073965853E-308) is found for RYR1, ryanodine receptor 1, which encodes a ryanodine receptor found in skeletal muscle. The encoded protein functions as a calcium release channel in the sarcoplasmic reticulum but also serves to connect the sarcoplasmic reticulum and transverse tubule. Mutations in this gene are associated with malignant hyperthermia susceptibility, central core disease, and minicore myopathy with external ophthalmoplegia. Alternatively spliced transcripts encoding different isoforms have been described.
Therefore RYR1 could be used as a housekeeping gene, and it makes sense to find such a low p-value. When visiting the human protein atlas it can also be seen that RYR1 protein levels are high in brain, whereas RNA levels are rather low. Therefore one can conclude that RYR1 is very stable, again supporting the low pvalue

RUBRIC

RYR1 had the smallest p-value ($4.76 * 10^{-308}$). This gene encodes the skeletal muscle ryanodine receptor, which serves as a calcium release channel of the sarcoplasmic reticulum as well as a bridging structure connecting the sarcoplasmic reticulum and transverse tubule. It can also mediate the release of Ca(2+) from intracellular stores in neurons, and may thereby promote prolonged Ca(2+) signaling in the brain.

The gene's function in the brain explains why, with a positive *Log2FoldChange* (implying that this gene is up-regulated in brain with respect to UHRR), RYR1 has a significant p-value.

- 0 pts
  The learner provided wrong answer for both questions.
- 1 pt
  The learner provided correct answer for one of the questions.
- 2 pts
  The learner provided correct answer for both questions.

PROMPT

Question 11: What is the function of the ten genes with the largest values of *Log2FoldChange*? How many of them do you think are related to brain functions?

Gene Log2foldChange Brain involvement?
GFAP 11,631294563494 yes This gene encodes one of the major intermediate filament proteins of mature astrocytes
STMN2 11,232949811443 yes This gene encodes a member of the stathmin family of phosphoproteins. Stathmin proteins function in microtubule dynamics and signal transduction. The encoded protein plays a regulatory role in neuronal growth and is also thought to be involved in osteogenesis
GABRA6 11,193170820769 yes GABA, of which GABRA5 is a buildingblock, is the major inhibitory neurotransmitter in the mammalian brain
OPALIN 11,040349845123 yes oligodendrocytic myelin paranodal and inner loop protein
MOBP 11,018164148283 yes myelin-associated oligodendrocyte basic protein
SYT4 11,007689124262 yes synaptotagmin 4
NEUROD2 10,898338393466 yes neuronal differentiation 2; This gene encodes a member of the neuroD family of neurogenic basic helix-loop-helix (bHLH) proteins
TTC9B 10,861648400115 Yes tetratricopeptide repeat domain 9B, expressed almost exclusively in brain
HMP19 10,851388948057 yes a.k.a. neuronal vesicle trafficking associated 2
GPR12 10,803816269844 yes G protein-coupled receptor 12, expressed in high levels in brain

RUBRIC

Eight of the ten genes listed below have some function related to the brain (marked as *BRAIN FUNCTION*).

- GFAP encodes one of the major intermediate filament proteins of mature astrocytes. Astrocytes are cells in the central nervous system, so the gene very likely plays a special role in the brain.. *BRAIN FUNCTION*
- STMN2 encodes a member of the stathmin family of phosphoproteins. (*Learner may or may not identify this as having brain function without penalty*)
- GABRA6 is the inhibitory neurotransmitter in the mammalian brain. *BRAIN FUNCTION*

Help Center

- OPALIN is a transmembrane protein associated with fucosidosis. *BRAIN FUNCTION*
- MOBP is a gene whose GO annotations include *Rab GTPase binding* and *structural constituent of myelin sheath*. Diseases associated with MOBP include schizophrenia. *BRAIN FUNCTION*
- SYT4 is a gene whose GO annotations include calcium ion binding and clathrin binding. Diseases associated with SYT4 include pheochromocytoma. *BRAIN FUNCTION*
- NEUROD2 encodes a member of the neuroD family of neurogenic basic helix-loop-helix (bHLH) proteins. Expression of this gene can induce transcription from neuron-specific promoters. The encoded protein can induce neurogenic differentiation in non-neuronal cells in Xenopus embryos, and play a role in the determination and maintenance of neuronal cell fates. *BRAIN FUNCTION*
- TTC9B encodes Tetratricopeptide Repeat Domain 9B. *BRAIN FUNCTION*
- HMP19 encodes a protein with GO annotations that include clathrin light chain binding. *BRAIN FUNCTION*
- GPR12 promotes neurite outgrowth and blocks myelin inhibition in neurons (by similarity). Receptor with constitutive G(s) signaling activity that stimulates cyclic AMP production. *BRAIN FUNCTION*

- 0 pts
  The learner provided more than two wrong or missing genes.
- 1 pt
  Learner provided all but one or two correct genes.
- 2 pts
  The learner provided correct list of genes.

PROMPT

Applying transcript-level expression analysis

As mentioned previously, the TopHat+Cufflinks+CuffDiff pipeline has the ability to distinguish between gene-level expression and transcript-level expression. To find the results of this workflow, go to “Projects,” then the “RNA-seq Methods Results” project you joined earlier. There should be a folder called “Cufflinks Assembly & DE 07/16/2015 10:55:19” (i.e. “Cufflinks Assembly & DE,” followed by the execution date, followed by the execution time”). Open this folder.

Click on the “Output Files” tab, open the “Cufflinks-Report” folder, open the “differential” folder, then open the “cuffdiff” folder (the “metrics” folder contains alignment data, and the “differential” folder contains differential gene expression data). In this folder, there are four files that end in “exp.diff”.

- UHRR_vs_Brain.cds_exp.diff: Coding sequence differential expression
- UHRR_vs_Brain.gene_exp.diff: Gene-level differential expression
- UHRR_vs_Brain.isoform_exp.diff: Transcript-level differential expression
- UHRR_vs_Brain.tss_group_exp.diff: Primary transcript (i.e. transcribed mature RNA that has not yet been processed into mRNA/tRNA/rRNA/etc.) differential expression

Refer to the CuffDiff manual to assist you (http://cole-trapnell-lab.github.io/cufflinks/cuffdiff/.

Question 12: Using a *p*-value threshold of 0.01, how many (if any) isoforms are significantly differentially expressed at the isoform level (i.e., $p <= 0.01$ at the isoform level) but not significantly differentially expressed at the gene level (i.e., $p > 0.01$ at the gene level)?

Note: In Excel, to test whether element A1 is less than *x* and B1 is greater than or equal to *x*, we can use the following formatting:

Help Center

```
AND(A1 < x, B1 >= x)
```

The AND function works in OpenOffice Calc as well, but the syntax requires the use of a semicolon instead of a comma:

```
AND(A1 < x; B1 >= x)
```

This will be TRUE if both A1 < *x* and B1 >= *x* (and FALSE otherwise). We can then use the Excel syntax

```
C1 = IF(AND(A1 < x, B1 >= x), 1, 0)
```

which will set element C1 equal to 1 if AND(A1 < *x*, B1 >= *x*) is TRUE and set C1 equal to 0 otherwise. Again, in OpenOffice Calc, replace commas with semicolons. We can then double click the bottom right corner of C1 to extend this formula all the way down column C. Then, click the first empty element below column C, and then click “AutoSum”, which will compute the sum of all the “1” elements in column C, which is exactly the number of values in columns A and B satisfying the above specified condition.

There are 9661 significant gene differential expressions found, and 9596 isoforms
of these 9325 overlap

RUBRIC

479

- 0 pts
  The learner's answer was not within 50 of the correct response.
- 1 pt
  The learner's answer was within 50 of the correct response but not within 5.
- 2 pts
  The learner's answer was between 474 and 484, inclusively.

PROMPT

Question 13: Say that a given gene has roughly equal expression levels in samples *i* and *j* when looking at *gene*-level resolution. Explain how it is possible for this gene to be differentially expressed between samples *i* and *j* at the *transcript*-level by providing a hypothetical example of gene isoform expression.

Every cell has two copies of every gene in its DNA, therefore a leveled out gene-level is a good sign, that identical amounts of genetic material were analyzed.
RNA is only transcribed when needed and can be present in multiple copies, therefore a different transcript level indicates that a certain gene is activated more (or less) in two analyzed samples

RUBRIC

Help Center

As mentioned above, if gene *X* has roughly equal expression levels in two samples when looking at gene-level resolution, then it could still be differentially expressed at the transcript-level. Say *X* has two isoforms *X*1 and *X*2. If, in sample *i*, *X*1 has high expression and *X*2 has low expression, but in sample *j*, *X*2 has high expression and *X*1 has low expression. (At the gene-level, our conclusion would be "Gene *X* is not differentially expressed between samples *i* and *j*." However, at the transcript-level, our conclusion would be "Transcripts *X*1 and *X*2 are both differentially expressed between *i* and *j*.")

- 0 pts
  The learner did not answered correctly.
- 2 pts
  The learner answered correctly.

PROMPT

Identifying the most differentially expressed genes reported by TopHat+Cufflinks+CuffDiff

Question 14: According to the file "UHRR_vs_Brain.gene_exp.diff", which ten genes are the most differentially expressed between UHRR and Brain (i.e., have the greatest values of *Log2FoldChange*)? Ignore all entries with a *Log2FoldChange* of "inf" or "-inf".

SNAP25 9.98606
HBG2 -9.98402
LY86-AS1 9.97666
LINC00261 -9.97607
OPCML 9.97372
CA10 9.97368
HBE1 -9.95812
C4BPB -9.95555
NXF3 -9.9427
LOC150568 9.93553

RUBRIC

C1QB, ELAVL4, TTC9B, SST, STMN2, ZP2, SLC39A12, NEUROD2, AMER3, GFAP

- 0 pts
  The learner provided more than two wrong or missing genes.
- 1 pt
  Learner provided all but one or two correct genes.
- 2 pts
  The learner provided correct list of genes.

Help Center

PROMPT

Question 15: Do these ten most differentially expressed genes resulting from TopHat+Cufflinks+CuffDiff analysis overlap with the ten most differentially expressed genes resulting from the STAR+DESeq analysis?

None of the genes I found are identical using the two approaches

RUBRIC

There are **four** genes shared by the two sets (TTC9B, STMN2, NEUROD2 and GFAP).

- 0 pts
  The learner did not answered correctly.
- 2 pts
  The learner answered correctly.

PROMPT

Comparing the two RNA-seq analysis workflows

You now have two lists of genes with their expression levels, one from each of the two workflows. However, these lists also contain genes that are not significantly differentially expressed.

Question 16: How many genes are there having p-values at most equal to 0.001 in both lists? Exclude genes with a p-value of 0 in either list, as these represent computational artifacts.

Note: In Excel and OpenOffice Calc, we can test that element A1 and B1 are both between values *x* and *y* by simply adding arguments to AND(), which can take any number of parameters. That is, we can use the expression

```
1  AND(A1 > x, A1 < y, B1 > x, B1 < y)
```

which will return TRUE if all four conditions are true and FALSE otherwise. Again, in OpenOffice Calc, replace commas with semicolons. We can then use the previous note in order to determine the number of elements in a column satisfying these properties.

68 genes are significantly in both approaches

RUBRIC

There are 9108 genes that have p-values of less than or equal to 0.001 in both lists.

- 0 pts
  The learner's answer was not within 50 of the correct response.
- 1 pt
  The learner's answer was within 50 of the correct response but not within 5.
- 2 pts
  The learner's answer was between 9103 and 9113, inclusively.

Help Center

PROMPT

Identifying Genes with Disease Correlation

Each workflow provides a list of genes with their differential expression levels between the two groups, but the lists are not human-readable. To annotate these gene lists, we ran the NextBio Annotates RNA-seq app on the two lists of differentially expressed genes. When you run this application, the most differentially expressed genes from the experiment are annotated via the NextBio Research API. The output report shows you how those genes relate to the information in previously published papers.

Open the “Projects” tab of BaseSpace, open the “RNA-seq Methods Results” project you created previously, and open the “NextBio Annotates RNA-seq (TopHat+Cufflinks+CuffDiff)” folder. When the summary page opens, the content of the page depends on the gene that is selected in the dropdown menu of differentially expressed genes.

Question 17: What are the top three most differentially expressed genes (i.e., the first three genes in the dropdown menu)? What are the five “Most Correlated Tissues” of the most expressed gene? What are the five “Most Correlated Diseases” of the most expressed gene? Do the “Most Correlated Tissues” and “Most Correlated Diseases” seem to relate to each other?

Top Three: NeuroD6 (13,93); LINC00320 (13,83); MIR-3HG (13,36)
Most Correlated Tissues of NeuroD6: Fetal brain; Cingulate cortex; Superior cervical ganglion; Whole brain; Globus pallidus
Most Correlated Disease of NeuroD6: Spinocerebellar ataxia; Mood disorder; Brain cancer; Alzheimer's disease; Dementia
Yes, The two categories relate, with both having a great correlation to brain, neurons and mental health

RUBRIC

NEUROD6, LINC00320, MIR7-3HG

Most Correlated Tissues

1. Fetal brain
2. Cingulate cortex
3. Superior cervical ganglion
4. Whole brain
5. Globus pallidus

Most Correlated Diseases

1. Spinocerebellar ataxia
2. Mood disorder
3. Brain cancer
4. Alzheimer’s disease
5. Dementia

The most correlated tissues and diseases do seem to relate to each other because they all relate to the brain and nervous system.

Help Center

- 0 pts
  The learner provided wrong answer/list for more than two questions.
- 1 pt
  The learner provided correct answer/list for two or three questions.
- 2 pts
  The learner provided correct answer for all four questions.

PROMPT

Revisiting the ENCODE Controversy: How similar are human and mouse brains?

We now ask you to take a look at the ENCODE paper (Lin et al., 2014) and try to reproduce the results in this paper. Note that this portion of the project requires having the Java Runtime Environment (JRE) installed on your computer, so be sure to install Java if you haven't done so already:

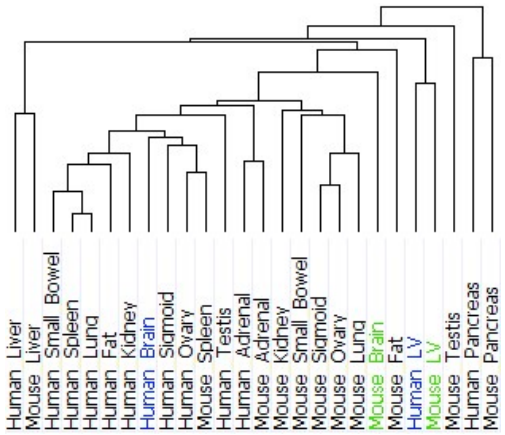
https://java.com/inc/BrowserRedirect1.jsp?locale=en

To answer the following question, which requires clustering see Walkthrough: How to perform hierarchical clustering with GenePattern.

Question 18: Upload the tree that you obtain from running hierarchical clustering on the resulting tree from the hierarchical clustering performed on the ENCODE paper's data.

RUBRIC

The uploaded tree should resemble the image below.



- 0 pts
  The learner's tree has more than 3 incorrect nodes.
- 2 pts
  The learner's tree is similar to the rubric.

Help Center

PROMPT

Question 19: Are there more similarities at the organism level (i.e., implying that the human brain is closer to human lung than to mouse brain) or at the tissue level (i.e., implying that the human brain is closer to mouse brain than human lung)?

There is more difference between species than between organs

RUBRIC

According to this tree, it seems as though there are more similarities at the organism level than at the tissue level (i.e., the human brain is closer to human lung than to mouse brain).

- 0 pts
  The learner did not answered correctly.
- 2 pts
  The learner answered correctly.

PROMPT

Now, read the rebuttal of the ENCODE paper (Gilad and Mizrahi-Man, 2015).

Question 20: Summarize the key arguments in Gilad and Mizrahi-Man's paper against those of Lin et al., 2014.

Gilad and Mizrahi-Man account for the batch effect, and so the corrected comparative gene expression data from human and mouse tend to cluster by tissue, not by species
In the reanalysis lowly expressed genes were excluded
they standardized by the total count of reads that mapped to the ortholog gene pairs
they took into consideration the GC content bias, and the method and sequencing design batch effect
Mitochondrial genes from both species were excluded
correction for the sequencing design batch effect had a drastic impact on the results
Through the reanalysis the conclusions of the Mouse ENCODE Consortium papers pertaining to the clustering of the comparative gene expression data are unwarranted.

RUBRIC

Did the learner provide a thoughtful summary of the provided paper that was well reasoned and in their own words?

- 1 pt
  Yes
- 0 pts
  No

Help Center

PROMPT