RNA-Seq Analysis in Galaxy (advanced)
answer sheet

by Youri Hoogstrate

June 24, 2014

Contents

1 Introduction

This document provides the answers to the advanced RNA-seq tutorial given at the Galaxy Community Conference 2014 (GCC2014) in Baltimore during the Training day.

2 SNP/SNV detection

Question 1
The name of the gene is *OPRM1*.

Question 2
Only one single nucleotide variant is detected.

Question 3

1. For each position, we have compared every sequenced nucleotide with the most common nucleotide from "thirteen anonymous volunteers from Buffalo, New York" according to the wikipedia page.
2. The reference says barely nothing about the population frequency of this variant because only 13 genomes have been used to construct it.
3. The called SNP can indeed be a very common variant in humans. The only thing you know, is that the SNP was not the most common variant in the majority of the 13 people from New York. Because the reference genome describes only one base per position, you might even call a SNP if the minority of these 13 people do have it. If you want to find SNPs that have been introduced by a mutation, you would like to remove common variants in the human genome population. This can for example be done by sequencing both a diseased an control sample, and subtracting the SNPs from the control from the diseased sample. Another possibility is to compare the detected SNPs to SNP databases.
   Theoretically you could over-come this problem by having a reference genome describing all know genomic variants (instead of 1 base per position), although it will be more complex to show it in programs like IGV (you will get some graph-like structure).

Question 4
The one detected SNP is also present in dbSNP.

Question 5
You can find the SNP at:

`http://snpedia.com/index.php/Rs1799971`

The website says:
"Carriers of at least one rs1799971(G) allele appear to have stronger cravings for alcohol than carriers of two rs1799971(A) alleles, and are thus hypothesized to be more at higher risk for alcoholism."

Question 6
There can be many things going on; polymerase or sequencing errors.
But what we want to point out it that intentional modification in the RNA also take place. Such mechanisms are referred to as "*RNA-editing*": `http://en.wikipedia.org/wiki/RNA_editing`
A quote from the wikipedia page: "*The diversity of RNA editing phenomena includes nucleobase modifications*

such as cytidine (C) to uridine (U) and adenosine (A) to inosine (I) deaminations, as well as non-templated nucleotide additions and insertions."

Question 7
The name of the "aligner" is: `ArtificialAligner: generate_reads (v1.0.0)`. When the BAM file was generated using tools like e.g. TopHat or STAR you can often find all used parameters.

Question 8
The indels are supported by 29.04% and 28.25% of the reads. The reason why the support is not about 100% is because the indels are heterozygous. There can be multiple reasons why the ratio's are also not close to 50%:

- The gene-expression is allele specific; one allele encompasses ∼30% and the other ∼70% of the expression of the gene (RNA-seq specific).
- There are more than 2 genomic copies of the chromosome. E.g. 2 copies (∼66.7%) that lack the indel and 1 copy (∼33.3%) that contains it (both DNA-seq and RNA-seq).

– Combination the answers above are also possible.

Question 9
The indel is causing a frame-shift. This means that all amino acids starting from the insertion may differ and the protein sequence will probably become very different from the original. If the length of an indel is a multitude of 3 ($3, 6, 9$ and so on) and the mutation doesn't introduce a stop-codon, the protein sequence will obtain an inserted amino acid but the remainder of the sequence will stay identical.

3 DGE (edgeR)

Question 1

- 23384799 reads are assigned
- $660652 + 3570909 + 2940537 = 7172098$ reads are unassigned
- $23384799/(23384799 + 7172098) * 100 = 76.53\%$ of the reads are assigned

If, by accident, you have counting using to the **wrong reference genome hg19**, you would have:

- 1786487 reads are assigned
- $30723 + 3570909 + 25168778 = 28770410$ reads are unassigned
- $1786487/(1786487 + 28770410) * 100 = 5.85\%$ of the reads are assigned

Question 2
53 reads have been aligned to the exon regions of gene `SAMD11` as defined in the GTF file. (If you find 288 counted reads, you have chosen hg19 instead of hg18!)

Question 3
The gene-length provided by featureCounts corresponds to the number of nucleotides that were used to count reads. With the used settings, this excludes introns and regions that span multiple genes. Therefore this length is not equal to the length of the entire gene.

Question 4
3: *Control*, *E2* and *Unknown*.

Question 5
The gene-cards website cites the an Entrez entry:
"**Entrez Gene summary for GREB1 Gene:**
This gene is an estrogen-responsive gene that is an early response gene in the estrogen receptor-regulated pathway. It is thought to play an important role in hormone-responsive tissues and cancer. Three alternatively spliced transcript variants encoding distinct isoforms have been found for this gene. (provided by RefSeq, Jul 2008)"

The "*E2*"-group was treated with the estrogen β-estradiol.

Question 6
"**GeneCards Summary for GREB1 Gene:**
GREB1 (growth regulation by estrogen in breast cancer 1) is a protein-coding gene, and is affiliated with the lncRNA class. Diseases associated with GREB1 include breast cancer, and endometriosis. An important paralog of this gene is GREB1L."

MCF-7 cell-lines are breast cancer cell lines and this gene is specifically associated with breast cancer.

Question 7
$\frac{4460}{22336}$ genes are significantly over expressed (FDR < 0.05)

Question 8
I would expect less differentially expressed genes if the samples were taken from different patients; each patient is unique and has it own individual/personal expression profile (caused by both genetic and environmental factors). In contrast, cell lines have the same genetic profile and are also cultured under the same environmental conditions.

Question 9
$\frac{4962}{22336}$ genes are significantly over expressed (FDR < 0.05) using 10M reads per sample. This is more than using 5M reads and in agreement with the results of the article.

Question 10
$\frac{6164}{22142}$ genes are significantly over expressed (FDR < 0.05) using the full coverage per sample. There many more significantly over-expressed genes using the full coverage instead of a subset of 5M or 10M reads per sample (please be aware that we have used a different gene annotation file, which explains the different number of total genes).

Question 11

There is a very clear separation between the *E2* and *Control*, indicating that it is very easy to distinguish between the groups only based on the expression profiles and that the expression profiles are very specific for the conditions. You could use data with such separation very well for classification. Still, a clinical classification application of this particular dataset is not very useful since you can only use it to predict whether the cell line has had an oestrogen treatment or not.

Question 12

$\frac{4294}{22142}$ It's just a little bit less than 7 samples at 5M reads. The drop in the number of significantly expressed genes caused by removing two replicates has a comparable impact as the removal of $(\sim 30M - 5M)/\sim 30M =$ 80% of the overall reads.

Question 13

$\frac{595}{23368}$ genes are differentially expressed.

Question 14

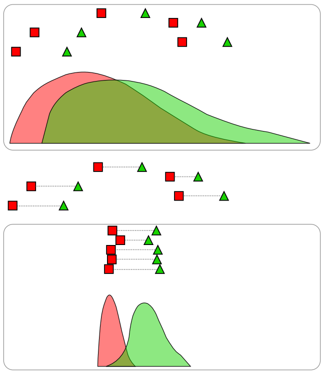
This question is very similar to question 8: the patients are biological replicates with (1) individual expression profiles, and (2) corresponding tumor specific expression profiles. The MCF-7 dataset consists of technical replicates which should all have the same expression profile since they have been cultured from the same genome under the same environmental conditions.

Question 15

$\frac{3117}{23368}$ genes are differentially expressed when pairing/blocking is modelled. By modeling the pairing information, you correct for expression profiles that are specific for a patient/block.

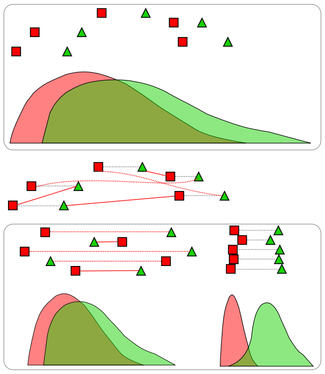
Some genes are not detected as over expressed in the first setting, because for those genes the expression profiles of the two conditions are very similar. When pairing is taken into account, you correct for individual expression profiles. Thus, if genes are considered as differentially expressed only when pairing is modelled, the condition specific expression profiles are more similar to each other than the differences per individual.

The given problem is shown in the figure below. In this figure the expression levels for one gene of 5 tumour (green triangle) and normal (red square) samples are indicated. The horizontal axis represents the expression level; the vertical distance between the squares and triangles has no meaning. As you can see, the tumour samples are slightly shifted to the right, compared to the normale samples. If you would make a distribution of both expression profiles (bulbs in the bottom of the 1st box), you see that they are very similar and overlap for the most part. As a result, there are no significant differences between the groups.
In the spacer, the gray lines in indicate the relation between the samples (the pairs). This shows that for every pair, the expression of the tumour sample is higher than for the normal sample. However, this difference in expression is relatively small, compared to the overall expression differences.
When we normalize for individual expression profiles, as shown in the lower box, the differences between tumour and normal become clearly visible. As result, the corresponding distributions have less overlap.
The genes that in the second but not in the first analysis were found to be differentially expressed, are thus characterized by individual expression profiles that are more different from each other than the differences between tumor and normal.



Question 16 (extra)
The number of significantly expressed genes is dropping to the level of the experiment where no sample pairing was modelled. This is because the patient specific expression is not representing the true nature of the data. Thus, we're correcting for patient specific expressions profiles that do not exist.

This is illustrated in the following figure:



Question 17
The number of replicates is in particularly important for the estimation of the variance in differential gene expression. For the estimation of SNPs it is more convenient to have a high number of reads. However, after a certain amount of sequenced bases per position, the effects of adding more sequences become marginal. Yet on the other hand, because of the expression differences, low expressed regions will be detected more accurately when the sequencing depth increases.

Question 18 (extra)
The plots illustrate two ”problems”:

1. There are no differential expressed genes and
2. The MDS shows 3 subclasses which are different from the actual groups, and therefore doesn't indicate differences between groups 1 and 2.

The sub-classes in the MDS plot indicate that there is more difference between the subclasses than between the biological groups. In other words, the groups are as similar to each other than any of the samples to another (in particularly caused by the 3 subclasses). You can find more obvious expression differences if you would model the subclasses as separate blocks since you correct for the subclass specific expression profiles.