RNA-Seq Analysis in Galaxy (advanced)

by Youri Hoogstrate

June 29, 2014

Contents

1 Introduction

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

2 SNP/SNV detection

2.1 Hands on

SNP detection is a process that consumes quite some resources. Therefore we restricted this manual to a simple (small) example. Load the file `hg19_mutant.bam` from the shared data library into galaxy. Ensure that the corresponding reference genome is hg19. This BAM file consists of reads generated by the computer to target a specific region in the genome. To get an impression of what data we are really looking at, we can view the BAM file in e.g. Trackster[?], IGV[?], Dalliance[?] or UCSC[?]. Press the IGV button in the history item and check out the data by browsing to the following genomic region (also here: make sure you use reference genome hg19):

```
chr6:154358000-154431000
```

Question 1
What is the name of the gene where the reads are aligned to?

To estimate the SNPs in the BAM file, go to the Tools pane (left) and type "varscan" in search. Please select the tool:

> "VarScan2 Call SNPs from BAM VarScan2 SNP/SNV detection; directly reading *.bam file(s) & using parallel mpileup generation, to avoid unnecessary I/O overhead and increase performance".

This tool compares for every genomic position whether the aligned bases are identical to the reference (using VarScan and samtools[?, ?]). This process needs a mpileup file in-between. Because these files are humongous, we designed the wrapper in a way that the output is careful with its resources. As consequence, this wrapper doesn't (yet) allow comparison between two BAM files (e.g. to detect somatic variants). When you have started the wrapper, please select:

VarScan2 Call SNPs from BAM

- Alignment file: "*hg19_mutants.bam*"
- Reference genome: hg19 (Should be selected automatically when you selected the alignment file)
- Output format: *VCF*
- Other setting: default

Press [**Execute**] to start the analysis.

Question 2
How many single nucleotide variants are detected?

Question 3
Given that the reference sequence is a plain FASTA sequence, describing only one base pair per position:

1. Who/what have we compared our BAM file with? (hint: `http://en.wikipedia.org/wiki/Reference_genome`)
2. Does comparing to a reference say anything about the population frequency of this variant?
3. Can the called SNP be a common variant in humans?

To indicate whether a detected SNP has been annotated in a database or has been described in literature before, the detected SNPs in VCF files can be compared with databases like dnSNP[?] etc. using ANNOVAR[?]. If you search for "annovar" in the Tools menu, you can find its wrapper by the name:

ANNOVAR Annotate a file using ANNOVAR

**Note: the ANNOVARs license is proprietary and you have to request access to the binary yourself.*

If the ANNOVAR wrapper is started, select:

ANNOVAR

- Reference genome: "**Human Feb. 2009 (GRCh37/hg19) (hg19)**".
- Select file to annotate: *"VarScan2 Call SNPs from BAM on . . . hg19_mutant.bam"*
- filetype: "*VCF4 file*"
- You should de-select every annotation, except dbSNP: "**130 (hg19 only)**"

Press [**Execute**] to start the analysis and examine the report carefully to answer the following questions.

Question 4
How many annotated SNPs are found in the VCF file?

Question 5
For the SNP located at position `chr6:154360797`, please browse to the snpDB link using the "dbsnp130"-ID given in the last column of the ANNOVAR output:

`http://snpedia.com/index.php/`*dbsnp-ID*

Examine what the expected effect(s) of having this SNP is/are; is it a good or a bad thing to have this SNP?

Question 6
Imagine we have sequenced the DNA from the same person, and `chr6:154360797` is highly covered with reads. The analysis has indicated that there is NO SNP in the DNA at `chr6:154360797`. Could you think of a mechanism that may have caused the discrepancy between SNP detection in RNA and DNA?

Imagine you want to reproduce an experiment you found online of which you were only able to find the BAM files, but you can't find all of the used alignment settings in the corresponding article neither in the supplement. Fortunately, the BAM-format support the possibility to store additional info besides alignment. You can check out the hidden data in a BAM or SAM-file using the `samtools view -H` wrapper[?]. If you search for "samtools" in the Tools menu, you can find the wrapper by the name:

Samtools view header Show only the header of a BAM or SAM file using samtools

To run the analysis, run the wrapper with the following setting:

Samtools view header

- Alignment in BAM or SAM format: *hg19_mutant.bam*

Question 7
What version of which aligner did we use?

In the previous example we have used an artificial dataset, where reads were simulated with a computer without modelling noise. Therefore we will proceed with real data. We took a genomic region (`chr10:8136673-8157170`) as a subset of an original BAM file to save resources. The original BAM file is derived from a TopHat2[?, ?] alignment on the MCF7 breast cancer cell line[?]. Instead of detecting polymorphisms we will focus this time on indels, since VarScan separates these methods. If you search for ”varscan” in the Tools menu, you can find its wrapper by the name:

> VarScan2 Call INDELs from BAM VarScan2 INDEL detection; directly reading *.bam file(s) & using parallel mpileup generation, to avoid unnecessary I/O overhead and increase performance

If the VarScan INDEL wrapper is started, select:

VarScan2 Call INDELs from BAM

- Alignment file: ”*GSM1244822_Control_Rep7.hg18.subset.chr10_8136673-8157170.bam*”
- Reference genome: hg18 (This should go automatically when you selected the alignment file)
- Output format: *VCF*
- Other setting: default

Question 8
What are the ratio’s between reads that do and do not support the detected indels? Can you give an explanation why the ratio’s are far away from 100% but close to each other?

The results should contain an INDEL between `chr10:8151519-8151520` (hg18). Previous research on the MCF7 cell line focussed on the effect of this insertion [?]. The article can be found at the following url:

`http://dx.doi.org/10.1186/1471-2407-14-278`

Take a look at their *Figure 1* and read the legend carefully.

Question 9
What is the effect of the insertion (in terms of translation)? How is this effected related to the length of the indel?

3 DGE (edgeR)

A fundamental part of RNA-seq research is the experimental design. Imagine you have sequenced a batch of samples, but the biological question can not be answered using the experimental set-up. For analysis on differentially expressed genes you measure a genes’ expression and how much the expression varies within the populations. The difficulty of the estimation of variance is that it requires preferably as many replicates as possible (which are expensive). A recent article recalled the importance of adding more replicates by comparing it to the effects of increasing the sequencing depth [?]. They show that after a certain amount of reads, increasing the sequencing depth has only a small effect on the detection of differentially expressed genes while adding more replicates increases the power to detect differentially expressed genes significantly. In this hands on session we want to illustrate the idea of their findings.

3.1 Hands on 01: estimating gene expression

For this hands on we need the RNA-seq alignment `GSM1244809 E2_Rep1.bam`. Ensure that the file is annotated at reference genome ”*Human Mar. 2006 (NCBI36/hg18) (hg18)*”. To estimate expression in RNA-Seq, we simply count the number of reads that are aligned to the exons of a list of candidate genes. This list is provided as a GTF/GFF file. There are a variety of tools available for counting reads, of which featureCounts[?] is exceptionally fast and works directly on BAM files. If you search for ”*featurecounts*” in the Tools menu, you can find the wrapper by the name:

featureCounts Measure gene expression in RNA-Seq experiments from SAM or BAM files

To run the analysis, run the wrapper with the following settings:

featureCounts

- Alignment file: *GSM1244809 E2_Rep1.bam*
- GFF/GTF Source: *Use a built-in index (which fits your reference)*
- Reference Gene Sets used during alignment (GFF/GTF): *UCSC hg18*
- Output format: *Gene-name "\t" gene-count \t" gene-length (tab-delimited)*
- Number of the CPU threads: *2*
- featureCounts parameters: *Default settings*

The counting procedure will take about ∼5 min on an average computer.

Question 1
Before we proceed with the expression levels of the genes, we would like to get a small impression about whether the counting has been performed correctly. Therefore we take a look at featureCounts' output-summary file "*featureCounts on...: GSM1244809 E2_Rep1.bam summary*".

- How many reads are "Assigned"?
- How many reads are "UnAssigned (sum of all)"?
- What is the percentage assigned reads "$(Assigned/(Assigned + Unassigned) * 100)$" − do you think this percentage is acceptable?

Question 2
Take a look at featureCounts' output file "*featureCounts on...: GSM1244809 E2_Rep1.bam*". How many reads are aligned to `SAMD11`?

Question 3
Given that the genomic length of `SAMD11` is ∼8800 nucleotides, why says the gene-length column of the featureCounts output that the length is only ∼2550 nucleotides?

3.2 Hands on 02: low sequencing depth

The authors of the article that recalled the importance of adding many replicates, compared the number of replicates with changing the sequencing depth and have published a part of their RNA-Seq read counts on the MCF7 data as expression matrices. Because aligning and counting reads consumes many resources we proceed with the DGE analysis using these files directly. First we will take a look at the number of genes that are detected as significantly differentially expressed using datasets of 5M and 10M raw reads, `GSE51403_expression_matrix_5M_coverage.txt` and `GSE51403_expression_matrix_10M_coverage.txt` respectively. The experimental design of all samples is given in `GSE51403_design_matrix_subsampled.txt`. Please upload these three files into Galaxy (no reference annotation is required).

Question 4
Please take a look at the file `GSE51403_design_matrix_subsampled.txt`. Given that the first column lists the names of the samples and the second column the samples' corresponding condition, how many conditions does the experiment have?

The relevant conditions to this experiment are *Control* and *E2*. The *Control* group consists of untreated breast cancer cell line replicates and the E2 group is treated with 10nM 17 β-estradiol. To estimate the differentially expressed genes in Galaxy we make use of the tool EdgeR[**?**]. If you search for "edger" in the Tools menu, you can find the wrapper by the name:

edgeR: Differential Gene(Expression) Analysis RNA-Seq gene expression analysis using edgeR (R package)

To run the analysis, run the wrapper with the following settings:

edgeR: Differential Gene(Expression) Analysis

- Expression (read count) matrix: *GSE51403_expression_matrix_5M_coverage.txt*
- Design matrix: *GSE51403_design_matrix_subsampled.txt*
- Contrast (biological question): *Control-E2* (groups from design matrix: Case Sensitive!!)
- False Discovery Rate (FDR): 0.05
- Optional desired outputs:
  - □ Raw counts table
  - □ MDS-plot
  - □ BCV-plot
  - □ MA-plot
  - ✓ P-Value distribution plot
  - □ Hierarchical custering
  - □ Heatmap
  - □ R stdout
  - □ R Data object

After you pressed [**Execute**], you will obtain 3 files:

1. ... edgeR DGE on ...: ...design_matrix_subsampled.txt - differentially expressed genes
2. ... edgeR DGE on ...: ...design_matrix_subsampled.txt - CPM
3. ... edgeR DGE on ...: ...design_matrix_subsampled.txt - P-Value distribution

The first file gives for every counted gene the estimated statistics. The second file contains the normalized read counts and the third file contains a P-value distribution plot. If everything is correct, the gene `GREB1` is located in the top of the file. Please check it's corresponding gene cards page:

`http://www.genecards.org/cgi-bin/carddisp.pl?gene=GREB1`

Question 5
Can you find on the gene cards page a regulatory factor of the gene that relates to the *E2* treatment?

Question 6
Can you find on the gene cards page an association with MCF-7 cells?

Because the "*...differentially expressed genes*"-file is very large, it is difficult to grasp its content at a glance. To get a quick and dirty impression of the (relative) amount of significantly expressed genes, take a look at file "*edgeR DGE on ...: ...design_matrix_subsampled.txt - P-Value distribution*". It shows the distribution of FDR corrected P-values (< 0.99, since genes that have no counts are marked as 1) as a histogram. When there are few differentially expressed genes, the distribution tends to be uniform and the bars should be equally high. The more significant genes are found, the higher the spike(s) near FDR $= 0$. The $e =$ indicates the root squared error of the spikes compared to the horizontal red line. This is a measurement of how much the P-value distribution deviates from a uniform distribution.

Question 7
How much genes are differentially expressed (FDR < 0.05) between the *Control* and *E2* using 5M raw reads per sample? (rough estimate is enough)

Question 8
Would you expect a similar amount of differentially expressed genes if the samples were taken from different patients instead of the cell line replicates?

Repeat the experiment but change the following setting:

edgeR: Differential Gene(Expression) Analysis

- Expression (read count) matrix: *GSE51403_expression_matrix_10M_coverage.txt*

Question 9
How much genes are differentially expressed (FDR < 0.05) between the *Control* and *E2* using 10M raw reads per sample? (rough estimate is enough) – Is this more or less than the dataset with 5M reads?

3.3 Hands on 03: full sequencing depth

Since we are curious about the effects of down-sampling, we would like to apply the same analysis also on the full dataset (without sub-sampling of reads). The corresponding expression matrices are not available online, so therefore we reproduced the experiment. We are aware that we have used a slightly different GTF file, resulting in a different number of genes. Also, we have used featureCounts[**?**] instead of HTSeq-count[**?**]. To run the analysis, run the wrapper with the following settings:

edgeR: Differential Gene(Expression) Analysis

- Expression (read count) matrix: *GSE51403_expression_matrix_full.txt*
- Design matrix: *GSE51403_design_matrix_full_depth.txt*
- Contrast (biological question): *Control-E2*   (groups from design matrix: Case Sensitive!!)
- False Discovery Rate (FDR): 0.05
- Optional desired outputs:

  ☐ Raw counts table

✓ MDS-plot (This was not selected in the previous analysis)

☐ BCV-plot

☐ MA-plot

✓ P-Value distribution plot

☐ Hierarchical clustering

☐ Heatmap

☐ R stdout

☐ R Data object

Question 10
How much genes are differentially expressed (FDR < 0.05) between the *Control* and *E2* using the full data set? How does this compare to the sub-sampled data sets?

Take a look at the multi dimensional scaling (MDS) plot. It draws the euclidean distance between the samples, based on their log normalized expression profiles. If the expression profiles of samples are similar to each other, they will be close to each other in the MDS space, and vice versa.

Question 11
Do the conditions separate from each other? Can you think of a clinical application where separation is useful?

3.4 Hands on 04: full sequencing depth, low sequencing depth

To lower the number of replicates we can simply make a sub-selection of the columns. If you search for ”edger” in the Tools menu, you can find the wrapper by the name:

edgeR: Concatenate Expression Matrices Create a full expression matrix . . . from specific count tables

To run the analysis, run the wrapper with the following settings:

edgeR: Concatenate Expression Matrices

- Add a gene-IDs column at the end of the file: *yes*
- Select Read-count dataset that contains a column for GeneIDs: *GSE51403_expression_matrix_full.txt*
- Contrast (biological question): *Control-E2* (groups from design matrix: Case Sensitive!!)
- Select GeneID column: *c1: Geneid*
- **[Add new Expression Table]**
  - Read-count dataset that belongs to a pair: *GSE51403_expression_matrix_full.txt*

☐ c1: Geneid

✓ c2: GSM1244816: Control_Rep1

✓ c3: GSM1244817: Control_Rep2

✓ c4: GSM1244818: Control_Rep3

✓ c5: GSM1244819: Control_Rep4

☐ c6: GSM1244820: Control_Rep5 (excluded)

☐ c7: GSM1244821: Control_Rep6 (excluded)

✓ c8: GSM1244822: Control_Rep7

✓ c9: GSM1244809: E2_Rep1

✓ c10: GSM1244810: E2_Rep2

✓ c11: GSM1244811: E2_Rep3

✓ c12: GSM1244812: E2_Rep4

□ c13: GSM1244813: E2_Rep5 (excluded)

□ c14: GSM1244814: E2_Rep6 (excluded)

✓ c15: GSM1244815: E2_Rep7

□ c16: Length

If you have created the new expression matrix with less replicates, use it to repeat the differential gene expression analysis. Make sure you use the new expression matrix (5 + 5 samples) instead of the old (7 + 7 samples). Take a look at the output.

Question 12
How much differentially expressed genes do you find if you have 5 instead of 7 replicates using full sequencing depth? Are that more or less differentially expressed genes than using 7 replicates with 5M reads?

3.5 Hands on 05: paired samples

For the next assignment we will make use of public experiment `SRA020176` with GEO alias `GSE22260`. Within this experiment, several normal prostates and prostate cancers are sequenced of which 10 samples are paired [**?**]. Here **paired means that the tumour and the normal are taken from the same patient** and has nothing to do with paired-end sequencing. We have analysed the paired samples of the data already, using TopHat 2 on hg19. The read counts in the expression matrix were calculated using featureCounts[**?**].
For this experiment we don't have the design matrix ready. However, we can create one very simple by opening the wrapper to generate design matrices. If you search for ”edger” in the Tools menu, you can find the wrapper by the name:

edgeR: Design- from Expression matrix Create design- from an expression matrix

First, in this analysis we will apply a straight forward two group test, as we did before. Then, we will model the paired samples to normalize for patient specific expression profiles, and compare the results. For the unpaired analysis, we define two groups: *tumor* and *normal*. Create a design matrix with the following settings:

edgeR: Design- from Expression matrix

- Expression matrix (read counts): ”*GSE22260_expression_matrix_hg19_subset.txt*”
- Specify a name for the factor / condition: *Condition*
- Specify a condition: *Tumor* (Case sensitive!)
- Select columns that are associated with this factor level:

□ c1: Gene-id

✓ c2: carcinoma_C02 (SRR057629:GSM554076)

✓ c3: carcinoma_C03 (SRR057630:GSM554078)

✓ c4: carcinoma_C06 (SRR057632:GSM554082)

✓ c5: carcinoma_C08 (SRR057634:GSM554086)

✓ c6: carcinoma_C09 (SRR057635:GSM554088)

✓ c7: carcinoma_C11 (SRR057636:GSM554090)

✓ c8: carcinoma_C13 (SRR057637:GSM554092)

✓ c9: carcinoma_C15 (SRR057638:GSM554094)

✓ c10: carcinoma_C19 (SRR057641:GSM554100)

✓ c11: carcinoma_C23 (SRR057642:GSM554102)

☐ c12: normal_N02 (SRR057649:GSM554116)

☐ c13: normal_N03 (SRR057650:GSM554118)

☐ c14: normal_N06 (SRR057651:GSM554120)

☐ c15: normal_N08 (SRR057652:GSM554122)

☐ c16: normal_N09 (SRR057653:GSM554124)

☐ c17: normal_N11 (SRR057654:GSM554126)

☐ c18: normal_N13 (SRR057655:GSM554128)

☐ c19: normal_N15 (SRR057656:GSM554130)

☐ c20: normal_N19 (SRR057657:GSM554132)

☐ c21: normal_N23 (SRR057658:GSM554134)

To add a selection field for the normals click: [**Add new Factor level**].
For this factor, select:

- Specify a condition: *Normal* (Case sensitive!)
- Select columns that are associated with this factor level:

  ☐ c1: Gene-id

  ☐ c2: carcinoma_C02 (SRR057629:GSM554076)

  ☐ c3: carcinoma_C03 (SRR057630:GSM554078)

  ☐ c4: carcinoma_C06 (SRR057632:GSM554082)

  ☐ c5: carcinoma_C08 (SRR057634:GSM554086)

  ☐ c6: carcinoma_C09 (SRR057635:GSM554088)

  ☐ c7: carcinoma_C11 (SRR057636:GSM554090)

  ☐ c8: carcinoma_C13 (SRR057637:GSM554092)

  ☐ c9: carcinoma_C15 (SRR057638:GSM554094)

  ☐ c10: carcinoma_C19 (SRR057641:GSM554100)

  ☐ c11: carcinoma_C23 (SRR057642:GSM554102)

  ✓ c12: normal_N02 (SRR057649:GSM554116)

  ✓ c13: normal_N03 (SRR057650:GSM554118)

  ✓ c14: normal_N06 (SRR057651:GSM554120)

  ✓ c15: normal_N08 (SRR057652:GSM554122)

  ✓ c16: normal_N09 (SRR057653:GSM554124)

  ✓ c17: normal_N11 (SRR057654:GSM554126)

  ✓ c18: normal_N13 (SRR057655:GSM554128)

  ✓ c19: normal_N15 (SRR057656:GSM554130)

  ✓ c20: normal_N19 (SRR057657:GSM554132)

  ✓ c21: normal_N23 (SRR057658:GSM554134)

- Define blocking (paired or grouped samples): *No*

In the history a new item "*...: Design matrix*" will appear. Within Galaxy, please rename it to "*...: Design matrix (unpaired)*" to avoid confusion later on. Now we have both our design- and expression matrix and we can continue the differential gene expression analysis:

edgeR: Differential Gene(Expression) Analysis

- Expression (read count) matrix: ”*GSE22260_expression_matrix_hg19_subset.txt*”
- Design matrix: *Design matrix (unpaired)*
- Contrast (biological question): *Tumor-Normal*
- False Discovery Rate (FDR): 0.05
- Optional desired outputs:

  ☐ Raw counts table
  ☐ MDS-plot (This was not selected in the previous analysis)
  ☐ BCV-plot
  ☐ MA-plot
  ✓ P-Value distribution plot
  ☐ Hierarchical custering
  ✓ Heatmap
  ☐ R stdout
  ☐ R Data object

Question 13
How many significantly differentially expressed genes do we find between *Tumor* and *Normal* (rough estimate is enough)

Question 14
We have used even 10 samples per group, but found less significant genes than the MCF-7 dataset with 7 replicates. Why do you think this happens? (Hint: other than sequencing depth)

Go to the history menu item ”*... Design matrix (unpaired)*” and click the ↻-button to repeat the experiment, but we're going to model the pairing in a separate design matrix by selecting:

edgeR: Design- from Expression matrix

- Define blocking (paired or grouped samples): *Yes*
- Specify a name for a blocking condition: *Patients*

For every of the ten pairs, add a block (yes it's a lot of clicking), and for every block select the two corresponding samples. The following samples correspond to each other:

blocks		
c2: carcinoma_C**02** (SRR057629:GSM554076)	↔	c12: normal_N**02** (SRR057649:GSM554116)
c3: carcinoma_C**03** (SRR057630:GSM554078)	↔	c13: normal_N**03** (SRR057650:GSM554118)
c4: carcinoma_C**06** (SRR057632:GSM554082)	↔	c14: normal_N**06** (SRR057651:GSM554120)
c5: carcinoma_C**08** (SRR057634:GSM554086)	↔	c15: normal_N**08** (SRR057652:GSM554122)
c6: carcinoma_C**09** (SRR057635:GSM554088)	↔	c16: normal_N**09** (SRR057653:GSM554124)
c7: carcinoma_C**11** (SRR057636:GSM554090)	↔	c17: normal_N**11** (SRR057654:GSM554126)
c8: carcinoma_C**13** (SRR057637:GSM554092)	↔	c18: normal_N**13** (SRR057655:GSM554128)
c9: carcinoma_C**15** (SRR057638:GSM554094)	↔	c19: normal_N**15** (SRR057656:GSM554130)
c10: carcinoma_C**19** (SRR057641:GSM554100)	↔	c20: normal_N**19** (SRR057657:GSM554132)
c11: carcinoma_C**23** (SRR057642:GSM554102)	↔	c21: normal_N**23** (SRR057658:GSM554134)

Now you have created the design matrix that takes sample pairing into account, rename the new history item to "*... Design matrix (paired)*" to avoid confusion. Use the new "paired" design matrix to repeat the differential gene expression analysis.

Question 15
How many significant genes do you find between *Tumor* and *Normal* if you use sample pairing? Why would this additional information provide more statistical power?

Question 16 (extra)
"Randomly" assign the pairing information to the samples such that none obeys the true blocking as defined in the table using wrapper "*edgeR: Design- from Expression matrix*". Repeat the differential gene expression experiment and explain the differences with the previous experiment.

Question 17
What would be more desirable for the estimation of SNPs in one single sample: high sequencing depth or more replicates?

3.6 Hands on 06: bonus question

For this question we make use of confidential but *real* data. Use the expression matrix "`expression_matrix_hg19_bonus_question.txt`" in the differentially gene expression analysis. Create a corresponding design matrix using the tool "*edgeR: Design- from Expression matrix Create design- from an expression matrix*", where you simply divide the samples based on their names into "*group1*" and "*group2*". Do the differential gene expression analysis by comparing the following contrast "*group1-group2*".

Question 18 (extra)
What is going on with this dataset? (Hint: examine the *P-value distribution plot*, *multi dimensional scaling(MDS) plot* carefully and yes, the class labels are correct)

* If you still have no clue, please create a second design matrix and instead of group1 and group2, add to *groupX* the samples from groups `A`, `F`, `D`, `E` and `B`. Add to *groupY* the samples from the remaining groups `G`, `H`, `I` and `C`. Run the differential gene expression again, using the new design matrix and the contrast "*groupX-groupY*" and try to understand what has changed.

4 Appendix: requirements

4.1 SNP/SNV detection

For users who want to do the tutorial on e.g. their own machine, the following Galaxy toolshed repository must be installed:

- `https://testtoolshed.g2.bx.psu.edu/view/yhoogstrate/varscan_mpileup2snp_from_bam`
  - This package requires the FASTA file of the human reference genome: hg19 (configured in "`tool-data/all_fasta.loc`").
  - This package requires a "`{FASTA file's full path}.fai`" file that corresponds to this FASTA reference. If the file is not found, VarScan[?] will use samtools[?] to create it. Therefore samtools must be installed and accessible in `$PATH` and the directory with the reference must be writeable for the user that runs galaxy.
- `https://testtoolshed.g2.bx.psu.edu/view/yhoogstrate/varscan_mpileup2indel_from_bam`

  - This package requires the FASTA file of the human reference genome: hg18 (configured in ”`tool-data/all_fasta.loc`”).
  - This package requires a ”`{FASTA file’s full path}.fai`” file that corresponds to this FASTA reference. If the file is not found, VarScan[?] will use samtools[?] to create it. Therefore samtools must be installed and accessible in `$PATH` and the directory with the reference must be writeable for the user that runs galaxy.

- `https://testtoolshed.g2.bx.psu.edu/view/yhoogstrate/samtools_view_header`
- `http://toolshed.g2.bx.psu.edu/view/saskia-hiltemann/annovar`
  - Because this program is proprietary, please follow the packagers instructions to get access to the binary.
  - Make sure references ”`hg19_snp130.txt`” and ”`hg19_snp130.txt.idx`” are configured properly.
- Genome browser IGV[?] configured in galaxy.

If you are doing this tutorial outside the GCC2014, you can obtain the data by running the following scripts:

- Artificial alignment file (*hg19*):
  - `hg19_mutant.bam`
- Real alignment file (*hg18*):
  - `GSM1244822_Control_Rep7.hg18.subset.chr10_8136673-8157170.bam`

```
wget http://testtoolshed.g2.bx.psu.edu/repos/yhoogstrate/varscan_mpileup2snp_from_bam/raw-file
/tip/test-data/hg19_mutant.bam

wget http://testtoolshed.g2.bx.psu.edu/repos/yhoogstrate/varscan_mpileup2indel_from_bam
/raw-file/tip/test-data/GSM1244822_Control_Rep7.hg18.subset.chr10_8136673-8157170.bam
```

*Note: don’t use a NEWLINE in the wget command.

4.2 DGE (edgeR)

For users who want to do the tutorial on e.g. their own machine, the following Galaxy toolshed repository must be installed:

- `https://testtoolshed.g2.bx.psu.edu/view/yhoogstrate/featurecounts`
  - Configure the UCSC (hg18) GTF/GFF file in the `./tool-data/gene_sets.loc` file.
- `https://testtoolshed.g2.bx.psu.edu/view/yhoogstrate/edger_with_design_matrix`
  - Dependencies will be automatically installed with it

If you are doing this tutorial outside the GCC2014, you can obtain the data by running the following scripts:

- ”Sub-sampled” datasets 5M and 10M:
  - `GSE51403_design_matrix_subsampled.txt` (design matrix)
  - `GSE51403_expression_matrix_5M_coverage.txt` (expression matrix 5M)
  - `GSE51403_expression_matrix_10M_coverage.txt` (expression matrix 10M)

```
wget http://home.uchicago.edu/~jiezhou/replication/count.matrix.byGene.round1.txt

echo -n -e "gene-id\t" > counts.txt

cat count.matrix.byGene.round1.txt >> counts.txt

subset="2012.568\|2012.563\|2012.565\|2012.562\|2012.569\|2012.564\|2012.566\|2012.575\|
2012.577\|2012.576\|2012.574\|2012.571\|2012.572\|2012.570"

cut -f 1,$(head -1 counts.txt | tr "\t" "\n" | grep -n $subset | grep "p.5M" | cut -f1 -d
 ":" | paste -sd ",") counts.txt > GSE51403_expression_matrix_5M_coverage.txt

cut -f 1,$(head -1 counts.txt | tr "\t" "\n" | grep -n $subset | grep "10M" | cut -f1 -d
 ":" | paste -sd ",") counts.txt > GSE51403_expression_matrix_10M_coverage.txt

wget http://testtoolshed.g2.bx.psu.edu/repos/yhoogstrate/edger_with_design_matrix/raw-file/
tip/test-data/GSE51403/GSE51403_design_matrix_subsampled.txt

rm counts.txt count.matrix.byGene.round1.txt
```

- ”Full” dataset:
  - `GSE51403_design_matrix_full_depth.txt` (design matrix)
  - `GSE51403_expression_matrix_full.txt` (expression matrix)

```
wget http://testtoolshed.g2.bx.psu.edu/repos/yhoogstrate/edger_with_design_matrix/raw-file
/tip/test-data/GSE51403/GSE51403_design_matrix_full_depth.txt

wget http://testtoolshed.g2.bx.psu.edu/repos/yhoogstrate/edger_with_design_matrix/raw-file
/tip/test-data/GSE51403/GSE51403_expression_matrix_full.txt
```

- Paired prostate cancer:
  - `GSE22260_expression_matrix_hg19_subset.txt` (expression matrix only)

```
wget http://testtoolshed.g2.bx.psu.edu/repos/yhoogstrate/edger_with_design_matrix/raw-file
/tip/test-data/GSE22260/GSE22260_expression_matrix_hg19_subset.txt
```

- Bonus question:
  - `expression_matrix_hg19_bonus_question.txt` (expression matrix only)

```
wget http://testtoolshed.g2.bx.psu.edu/repos/yhoogstrate/edger_with_design_matrix/raw-file
/tip/test-data/expression_matrix_hg19_bonus_question.txt
```

5 References