

TRAINER'S MANUAL

Introduction to RNA-Seq Hands-on Workshop

Nathan S. Watson-Haigh (ACPFG)

Ute Baumann (ACPFG)

Rad Susecki (ACPFG)

John Toubia (ACPFG)

Timo Tiirikka (ACPFG)

Bastien Llamas (ACAD)

David Lawrence (ACRF Cancer Genomics Facility)

Katherine Pillman (CCB)

 Follow @BIG_SA1

Adelaide, Australia

18th Apr. 2013

TRAINER'S MANUAL

Licensing

This work is licensed under a Creative Commons Attribution 3.0 Unported License and the below text is a summary of the main terms of the full Legal Code (the full licence) available at <http://creativecommons.org/licenses/by/3.0/legalcode>.

You are free:

- to copy, distribute, display, and perform the work
- to make derivative works
- to make commercial use of the work

Under the following conditions:

Attribution - You must give the original author credit.

With the understanding that:

Waiver - Any of the above conditions can be waived if you get permission from the copyright holder.

Public Domain - Where the work or any of its elements is in the public domain under applicable law, that status is in no way affected by the license.

Other Rights - In no way are any of the following rights affected by the license:

- Your fair dealing or fair use rights, or other applicable copyright exceptions and limitations;
- The author's moral rights;
- Rights other persons may have either in the work itself or in how the work is used, such as publicity or privacy rights.

Notice - For any reuse or distribution, you must make clear to others the licence terms of this work.



Contents

Licensing	3
Contents	4
Workshop Information	5
The Trainers	6
Document Structure	7
Resources Used	8
Data Quality	9
Key Learning Outcomes	10
Resources You'll be Using	10
Useful Links	10
Introduction	11
Prepare the Environment	12
Quality Visualisation	12
Read Trimming	15
Appendix	17
RNA-Seq	21
Key Learning Outcomes	22
Resources You'll be Using	22
Introduction	24
Prepare the Environment	24
Alignment	25
Isoform Expression and Transcriptome Assembly	28
Differential Expression	30
Appendix	33
References	35
Post-Workshop Information	37
Access to Computational Resources	38
Access to Workshop Documents	52
Access to Workshop Data	52
Space for Personal Notes or Feedback	55

Workshop Information

The Trainers



Dr. Nathan S. Watson-Haigh

Research Fellow in Bioinformatics
The Australian Centre for Plant Functional Genomics (ACPFG), SA
nathan.haigh@acpfg.com.au



Dr. Ute Baumann

Bioinformatics Leader
The Australian Centre for Plant Functional Genomics (ACPFG), SA
ute.baumann@acpfg.com.au



Dr. Rad Suchecki

Postdoctoral Fellow in Bioinformatics
The Australian Centre for Plant Functional Genomics (ACPFG), SA
radoslaw.suchecki@acpfg.com.au



Mr. John Toubia

Bioinformatics Officer
The Australian Centre for Plant Functional Genomics (ACPFG), SA
john.toubia@acpfg.com.au



Mr. Timo Tiirikka

Bioinformatics Officer
The Australian Centre for Plant Functional Genomics (ACPFG), SA
timo.tiirikka@acpfg.com.au



Dr. Bastien Llamas

ARC Postdoctoral Fellow
The Australian Centre for Ancient DNA (ACAD), SA
bastien.llamas@adelaide.edu.au



Mr. David Lawrence

Bioinformatics Research Fellow
ACRF Cancer Genomics Facility, Centre for Cancer Biology, SA
David.Lawrence@health.sa.gov.au



Dr. Katherine Pillman

Bioinformatics Research Fellow
Centre for Cancer Biology, SA
kapillman@gmail.com

Document Structure

We have provided you with an electronic copy of the workshop's hands-on tutorial documents. We have done this for two reasons: 1) you will have something to take away with you at the end of the workshop, and 2) you can save time (mis)typing commands on the command line by using copy-and-paste.

We advise you to use Acrobat Reader to view the PDF. This is because it properly supports some features we have implemented to ensure that copy-and-paste of commands works as expected. This includes the appropriate copy-and-paste of special characters like tilde and hyphens as well as skipping line numbers for easy copy-and-past of whole code blocks.



While you could fly through the hands-on sessions doing copy-and-paste you will learn more if you take the time, saved from not having to type all those commands, to understand what each command is doing!

The commands to enter at a terminal look something like this:

```
1 | tophat --solexa-quals -g 2 --library-type fr-unstranded -j \
  annotation/Danio_rerio.Zv9.66.spliceSites -o tophat/ZV9_2cells \
  genome/ZV9 data/2cells_1.fastq data/2cells_2.fastq
```

The following icons are used in the margin, throughout the documentation to help you navigate around the document more easily:



Important



For reference



Follow these steps



Questions to answer



Warning - STOP and read



Bonus exercise for fast learners



Advanced exercise for super-fast learners

Resources Used

We have provided you with an environment which contains all the tools and data you need for the duration of this workshop. However, we also provide details about the tools and data used by each module at the start of the respective module documentation.

Module: Data Quality

Primary Author(s):
Sonika Tyagi sonika.tyagi@agrif.org.au

Contributor(s):
Nathan S. Watson-Haigh nathan.watson-haigh@awri.com.au

Key Learning Outcomes

After completing this practical the trainee should be able to:

- Assess the overall quality of NGS sequence reads
- Visualise the quality, and other associated matrices, of reads to decide on filters and cutoffs for cleaning up data ready for downstream analysis
- Clean up and pre-process the sequences data for further analysis

Resources You'll be Using

Tools Used

FastQC

<http://www.bioinformatics.babraham.ac.uk/projects/fastqc/>

FASTX-Toolkit

http://hannonlab.cshl.edu/fastx_toolkit/

Picard

<http://picard.sourceforge.net/>

Useful Links

FASTQ Encoding

http://en.wikipedia.org/wiki/FASTQ_format#Encoding

Introduction



Going on a blind date with your read set? For a better understanding of the consequences please check the data quality!

For the purpose of this tutorial we are focusing only on Illumina sequencing which uses 'sequence by synthesis' technology in a highly parallel fashion. Although Illumina high throughput sequencing provides highly accurate sequence data, several sequence artifacts, including base calling errors and small insertions/deletions, poor quality reads and primer/adapter contamination are quite common in the high throughput sequencing data. The primary errors are substitution errors. The error rates can vary from 0.5-2.0% with errors mainly rising in frequency at the 3' ends of reads.

One way to investigate sequence data quality is to visualize the quality scores and other metrics in a compact manner to get an idea about the quality of a read data set. Read data sets can be improved by post processing in different ways like trimming off low quality bases, cleaning up any sequencing adapters and removing PCR duplicates. We can also look at other statistics such as, sequence length distribution, base composition, sequence complexity, presence of ambiguous bases etc. to assess the overall quality of the data set.

Highly redundant coverage (>15X) of the genome can be used to correct sequencing errors in the reads before assembly and errors. Various k-mer based error correction methods exist but are beyond the scope of this tutorial.

Quality Value Encoding Schema

In order to use a single character to encode Phred qualities, ASCII characters are used (<http://shop.alterlinks.com/ascii-table/ascii-table-us.php>). All ASCII characters have a decimal number associated with them but the first 32 characters are non-printable (e.g. backspace, shift, return, escape). Therefore, the first printable ASCII character is number 33, the exclamation mark (!). In Phred+33 encoded quality values the exclamation mark takes the Phred quality score of zero.

Early Solexa (now Illumina) sequencing needed to encode negative quality values. Because ASCII characters < 33 and non-printable, using the Phred+33 encoding was not possible. Therefore, they simply moved the offset from 33 to 64 thus inventing the Phred+64 encoded quality values. In this encoding a Phred quality of zero is denoted by the ASCII number 64 (the @ character). Since Illumina 1.8, quality values are now encoded using Phred+33.

FASTQ does not provide a way to describe what quality encoding is used for the quality values. Therefore, you should find this out from your sequencing provider. Alternatively, you may be able to figure this out by determining what ASCII characters are present in the FASTQ file. E.g the presence of numbers in the quality strings, can only mean the quality values are Phred+33 encoded. However, due to the overlapping nature of the Phred+33 and Phred+64 encoding schema it is not always possible to identify what

encoding is in use. For example, if the only characters seen in the quality string are (@ABCDEFGHI), then it is impossible to know if you have really good Phred+33 encoded qualities or really bad Phred+64 encoded qualities.

For a graphical representation of the different ASCII characters used in the two encoding schema see: http://en.wikipedia.org/wiki/FASTQ_format#Encoding.

Prepare the Environment



To investigate sequence data quality we will demonstrate tools called FastQC and FASTX-Toolkit. FastQC will process and present the reports in a visual manner. Based on the results, the sequence data can be processed using the FASTX-Toolkit. We will use one data set in this practical, which can be found in the QC directory on your desktop.



Open the Terminal and go to the directory where the data are stored:

```
1 cd ~/QC/  
2 pwd
```

At any time, help can be displayed for FastQC using the following command:

```
1 fastqc -h
```

Quality Visualisation



We have a file for a good quality and bad quality statistics. FastQC generates results in the form of a zipped and unzipped directory for each input file.



Execute the following command on the two files:

```
1 fastqc -f fastq bad_example.fastq  
2 fastqc -f fastq good_example.fastq
```

View the FastQC report file of the dab data using a web browser such as firefox.

```
1 firefox bad_example_fastqc/fastqc_report.html &
```



The report file will have a Basic Statistics table and various graphs and tables for different quality statistics. E.g.:

Table 1: FastQC Basic Statistics table

Filename	bad_example.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	40000
Filtered Sequences	0
Sequence length	100
%GC	48

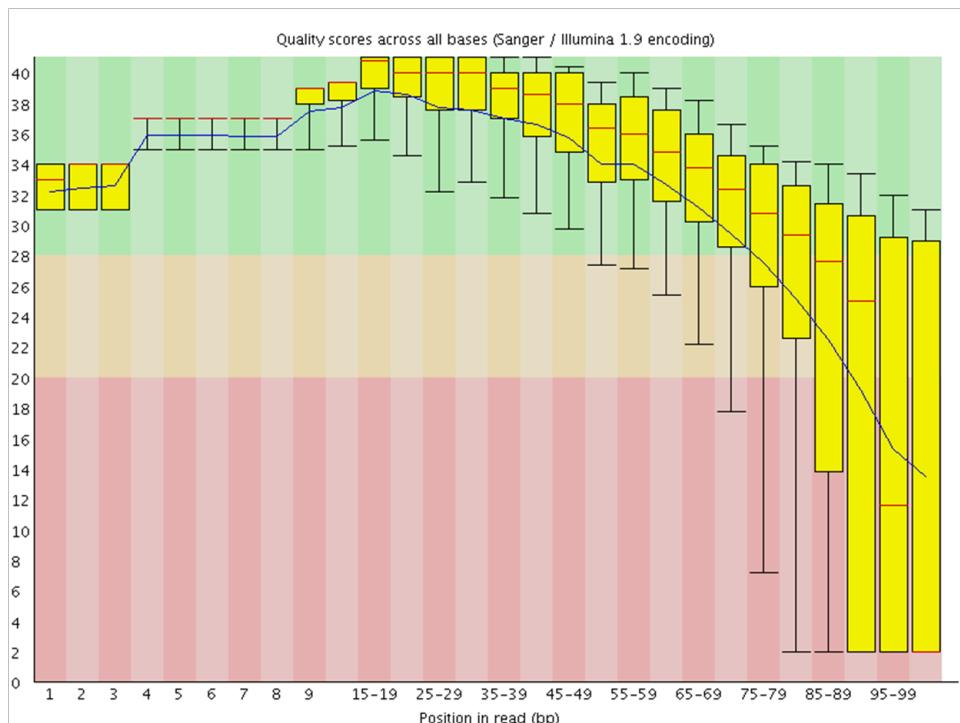


Figure 1: Per base sequence quality plot for bad_example.fastq.



A Phred quality score (or Q-score) expresses an error probability. In particular, it serves as a convenient and compact way to communicate very small error probabilities. The probability that base A is wrong ($P(\sim A)$) is expressed by a quality score, $Q(A)$, according to the relationship:

$$Q(A) = -10 \log_{10}(P(\sim A))$$

The relationship between the quality score and error probability is demonstrated with the following table:

Table 2: Error probabilities associated with various quality (Q) values

Quality score, Q(A)	Error probability, P($\sim A$)	Accuracy of the base call
10	0.1	90%
20	0.01	99%
30	0.001	99.9%
40	0.0001	99.99%
50	0.00001	99.999%



How many sequences were there in your file? What is the read length? **40,000. read length=100bp**

Does the quality score values vary throughout the read length? (hint: look at the 'per base sequence quality plot') **Yes. Quality scores are dropping towards the end of the reads.**

What is the quality score range you see? **2-40**

At around which position do the scores start falling below Q20? **Around 80 bp position**

How can we trim the reads to filter out the low quality data? **By trimming off the bases after a fixed position of the read or by trimming off bases based on the quality score on a read-by-read basis.**



Good Quality Data

View the FastQC report files `fastqc_report.html` to see examples of a good quality data and compare the quality plot with that of the `bad_example_fastqc`.

```
1 | firefox good_example_fastqc/fastqc_report.html &
```



Sequencing errors can complicate the downstream analysis, which normally requires that reads be aligned to each other (for genome assembly) or to a reference genome (for detection of mutations). Sequence reads containing errors may lead to ambiguous paths in the assembly or improper gaps. In variant analysis projects sequence reads are aligned against the reference genome. The errors in the reads may lead to more mismatches than expected from mutations alone. But if these errors can be removed or corrected, the read

alignments and hence the variant detection will improve. The assemblies will also improve after pre-processing the reads with errors.

Read Trimming

Read trimming can be done in a variety of different ways. Choose a method which best suits your data. Here we will use quality-based trimming but a fixed-length trimming example is provided in the appendix.

Quality Based Trimming

Base call quality scores can also be used to dynamically determine the trim points for each read. A quality score threshold and minimum read length following trimming can be used to remove low quality data.

 Run the following commands to quality trim your bad data:

```
1 cd ~/QC
2 fastq_quality_trimmer -h
3 fastq_quality_trimmer -Q 33 -t 20 -l 50 -i bad_example.fastq -o \
    bad_example_quality_trimmed.fastq
```

 Run FastQC on the quality trimmed file and visualise the quality scores.

```
1 fastqc -f fastq bad_example_quality_trimmed.fastq
2 firefox bad_example_quality_trimmed_fastqc/fastqc_report.html &
```

The output should look like:

Table 3: FastQC Basic Statistics table

Filename	bad_example_quality_trimmed.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	38976
Filtered Sequences	0
Sequence length	50-100
%GC	48

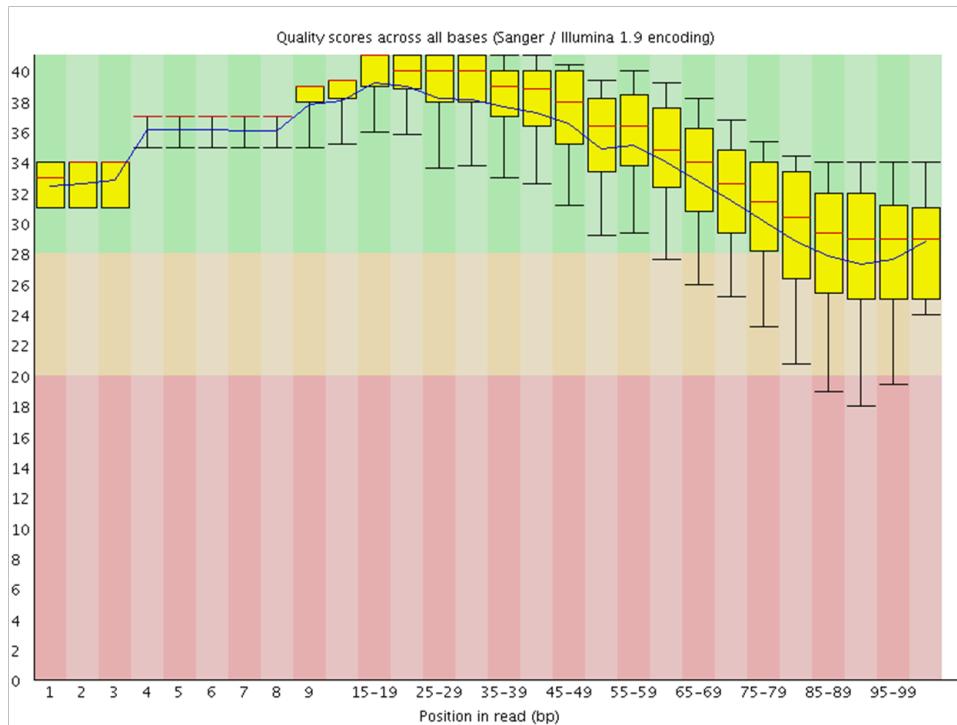


Figure 2: Per base sequence quality plot for the quality-trimmed `bad_example.fastq` reads.



How many sequences did you have following the trimming? 38,976. More than 1,000 reads were discarded for being too short following trimming.

What was the quality score range following trimming? 18-40

How did the read lengths change following trimming? Prior to trimming, all reads were 100bp long. Following trimming, some of the reads were as short as the 50bp minimum read length we specified with the `-l 50` argument in the `fastq_quality_trimmer` command.

Appendix

Fixed-Length Read Trimming

Low quality read ends can be trimmed using a fixed-length trimming. We will use the `fastx_trimmer` from the FASTX-Toolkit. Usage message to find out various options you can use with this tool. Type `fastx_trimmer -h` at anytime to display help.



We will now do fixed-length trimming of the `bad_example.fastq` file using the following command.

```
1 cd ~/QC
2 fastx_trimmer -h
3 fastx_trimmer -Q 33 -f 1 -l 80 -i bad_example.fastq -o \
    bad_example_trimmed01.fastq
```



We used the following options in the command above:

- Q 33** Indicates the input quality scores are Phred+33 encoded
- f 1** First base to be retained in the output
- l 80** Last base to be retained in the output
- i** Input FASTQ file name
- o** Output file name



Run FastQC on the trimmed file and visualise the quality scores of the trimmed file.

```
1 fastqc -f fastq bad_example_trimmed01.fastq
2 firefox bad_example_trimmed01_fastqc/fastqc_report.html &
```

The output should look like:

Table 4: FastQC Basic Statistics table

Filename	bad_example_trimmed01.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	40000
Filtered Sequences	0
Sequence length	80
%GC	48

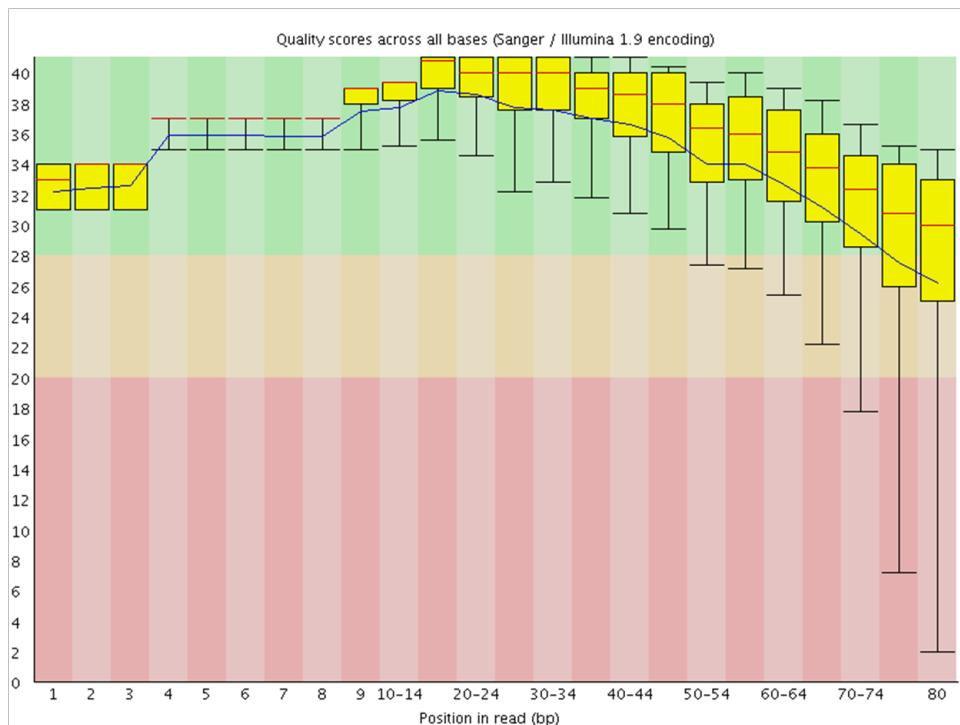


Figure 3: Per base sequence quality plot for the fixed-length trimmed `bad_example.fastq` reads.



What values would you use for `-f` if you wanted to trim off 10 bases at the 5' end of the reads? `-f 11`

Adapter Clipping

Sometimes sequence reads may end up getting the leftover of adapters and primers used in the sequencing process. It's good practice to screen your data for these possible contamination for more sensitive alignment and assembly based analysis.



This is particularly important when read lengths can be longer than the molecules being sequenced. For example when sequencing miRNAs.

Various QC tools are available to screen and/or clip these adapter/primer sequences from your data. (e.g. FastQC, FASTX-Toolkit, cutadapt).



Here we are demonstrating `fastx_clipper` to trim a given adapter sequence.

```
1 cd ~/QC
2 fastx_clipper -h
3 fastx_clipper -v -Q 33 -l 20 -M 15 -a GATCGGAAGAGCGGTTCAGCAGGAATGCCGAG \
```

```
-i bad_example.fastq -o bad_example_clipped.fastq
```



An alternative tool, not installed on this system, for adapter clipping is `fastq-mcf`. A list of adapters is provided in a text file. For more information, see FastqMcf at <http://code.google.com/p/ea-utils/wiki/FastqMcf>.

Removing Duplicates

Duplicate reads are the ones having the same start and end coordinates. This may be the result of technical duplication (too many PCR cycles), or over-sequencing (very high fold coverage). It is very important to put the duplication level in context of your experiment. For example, duplication level in targeted or re-sequencing projects may mean something different in RNA-seq experiments. In RNA-seq experiments oversequencing is usually necessary when detecting low abundance transcripts.



The duplication level computed by FastQC is based on sequence identity at the end of reads. Another tool, Picard, determines duplicates based on identical start and end positions in SAM/BAM alignment files.

We will not cover Picard but provide the following for your information.

Picard is a suite of tools for performing many common tasks with SAM/BAM format files. For more information see the Picard website and information about the various command-line tools available:

<http://picard.sourceforge.net/command-line-overview.shtml>



Picard 1.69 is installed on this system in `/usr/share/java/picard-1.69/`

One of the Picard tools (`MarkDuplicates`) can be used to analyse and remove duplicates from the raw sequence data. The input for Picard is a sorted alignment file in BAM format. Short read aligners such as, bowtie, BWA and tophat can be used to align FASTQ files against a reference genome to generate SAM/BAM alignment format.



Interested users can use the following general command to run the `MarkDuplicates` tool at their leisure. You only need to provide a BAM file for the INPUT argument (not provided):

```
1 cd ~/QC
2 java -jar /usr/share/java/picard/MarkDuplicates.jar \
    INPUT=<alignment_file.bam> VALIDATION_STRINGENCY=LENIENT \
    OUTPUT=alignment_file.dup METRICS_FILE=alignment_file.metrics \
    ASSUME_SORTED=true REMOVE_DUPLICATES=true
```

Module: RNA-Seq

Primary Author(s):

Myrto Kostadima, EMBL-EBI kostadmi@ebi.ac.uk
Remco Loos, EMBL-EBI remco@ebi.ac.uk

Contributor(s):

Nathan S. Watson-Haigh nathan.watson-haigh@awri.com.au

Key Learning Outcomes

After completing this practical the trainee should be able to:

- Understand and perform a simple RNA-Seq analysis workflow.
- Perform gapped alignments to an indexed reference genome using TopHat.
- Perform transcript assembly using Cufflinks.
- Visualize transcript alignments and annotation in a genome browser such as IGV.
- Be able to identify differential gene expression between two experimental conditions.

Resources You'll be Using

Tools Used

Tophat

<http://tophat.cbcn.um.edu/>

Cufflinks

<http://cufflinks.cbcn.um.edu/>

Samtools

<http://samtools.sourceforge.net/>

BEDTools

<http://code.google.com/p/bedtools/>

UCSC tools

<http://hgdownload.cse.ucsc.edu/admin/exe/>

IGV

<http://www.broadinstitute.org/igv/>

DAVID Functional Analysis

<http://david.abcc.ncifcrf.gov/>

Sources of Data

<http://www.ebi.ac.uk/ena/data/view/ERR022484>

<http://www.ebi.ac.uk/ena/data/view/ERR022485>

Introduction

The goal of this hands-on session is to perform some basic tasks in the downstream analysis of RNA-seq data. We will start from RNA-seq data aligned to the zebrafish genome using Tophat.

We will perform transcriptome reconstruction using Cufflinks and we will compare the gene expression between two different conditions in order to identify differentially expressed genes.

Prepare the Environment

We will use a dataset derived from sequencing of mRNA from *Danio rerio* embryos in two different developmental stages. Sequencing was performed on the Illumina platform and generated 76bp paired-end sequence data using polyA selected RNA. Due to the time constraints of the practical we will only use a subset of the reads.

The data files are contained in the subdirectory called **data** and are the following:

2cells_1.fastq and **2cells_2.fastq**

These files are based on RNA-seq data of a 2-cell zebrafish embryo

6h_1.fastq and **6h_2.fastq**

These files are based on RNA-seq data of zebrafish embryos 6h post fertilization



Open the Terminal and go to the **RNA-seq** working directory:

```
1 | cd ~/RNA-seq/
```



All commands entered into the terminal for this tutorial should be from within the **~/RNA-seq** directory.



Check that the **data** directory contains the above-mentioned files by typing:

```
1 | ls data
```

Alignment

There are numerous tools for performing short read alignment and the choice of aligner should be carefully made according to the analysis goals/requirements. Here we will use Tophat, a widely used ultrafast aligner that performs spliced alignments.

Tophat is based on the Bowtie aligner and uses an indexed genome for the alignment to speed up the alignment and keep its memory footprint small. Created the index for the *Danio rerio* genome:

```
1 | cd ~/RNA-seq  
2 | bowtie-build genome/Danio_rerio.Zv9.66.dna.fa genome/ZV9
```



Tophat has a number of parameters in order to perform the alignment. To view them all type:

```
1 | tophat --help
```



The general format of the tophat command is:

```
1 | tophat [options]* <index_base> <reads_1> <reads_2>
```



Where the last two arguments are the .fastq files of the paired end reads, and the argument before is the basename of the indexed genome.



The quality values in the FASTQ files used in this hands-on session are Phred+33 encoded. We explicitly tell tophat of this fact by using the command line argument **--solexa-quals**. See the appendix for further information about FASTQ encoding.



You can look at the first few reads in the file **data/2cells_1.fastq** with:

```
1 | head -n 20 data/2cells_1.fastq
```



Some other parameters that we are going to use to run Tophat are listed below:

-g Maximum number of multihits allowed. Short reads are likely to map to more than one location in the genome even though these reads can have originated from only one of these regions. In RNA-seq we allow for a limited number of multihits, and in this case we ask Tophat to report only reads that map at most onto 2 different loci.

--library-type Before performing any type of RNA-seq analysis you need to know a few things about the library preparation. Was it done using a strand-specific protocol or not? If yes, which strand? In our data

the protocol was NOT strand specific.

- j Improve spliced alignment by providing Tophat with annotated splice junctions. Pre-existing genome annotation is an advantage when analysing RNA-seq data. This file contains the coordinates of annotated splice junctions from Ensembl. These are stored under the sub-directory `annotation` in a file called `ZV9.spliceSites`.
- o This specifies in which subdirectory Tophat should save the output files. Given that for every run the name of the output files is the same, we specify different directories for each run.

It takes some time (approx. 20 min) to perform tophat spliced alignments, even for this subset of reads. Therefore, we have pre-aligned the `2cells` data for you using the following command:



You DO NOT need to run this command yourself - we have done this for you.

```
1 tophat --solexa-quals -g 2 --library-type fr-unstranded -j \
    annotation/Danio_rerio.Zv9.66.spliceSites -o tophat/ZV9_2cells \
    genome/ZV9 data/2cells_1.fastq data/2cells_2.fastq
```



Align the `6h` data yourself using the following command:

```
1 # Takes approx. 20mins
2 tophat --solexa-quals -g 2 --library-type fr-unstranded -j \
    annotation/Danio_rerio.Zv9.66.spliceSites -o tophat/ZV9_6h \
    genome/ZV9 data/6h_1.fastq data/6h_2.fastq
```

The `6h` read alignment will take approx. 20 min to complete. Therefore, we'll take a look at some of the files, generated by tophat, for the pre-computed `2cells` data.

Alignment Visualisation in IGV

The Integrative Genomics Viewer (IGV) is able to provide a visualisation of read alignments given a reference sequence and a BAM file. We'll visualise the information contained in the `accepted_hits.bam` and `junctions.bed` files for the pre-computed `2cells` data. The former, contains the tophat sliced alignments of the reads to the reference while the latter stores the coordinates of the splice junctions present in the data set.



Open the RNA-seq directory on your Desktop and double-click the `tophat` subdirectory and then the `ZV9_2cells` directory.

1. Launch IGV by double-clicking the “IGV 2.1” icon on the Desktop (ignore any warnings that you may get as it opens). *NOTE: IGV may take several minutes to load for the first time, please be patient.*
2. Choose “Zebrafish (Zv9)” from the drop-down box in the top left of the IGV window.
3. Load the `accepted_hits.sorted.bam` file by clicking the “File” menu, selecting “Load from File” and navigating to the Desktop/RNA-seq/tophat/ZV9_2cells directory.
4. Rename the track by right-clicking on its name and choosing “Rename Track”. Give it a meaningful name like “2cells BAM”.
5. Load the `junctions.bed` from the same directory and rename the track “2cells Junctions BED”.
6. Load the Ensembl annotations file `Danio_rerio.Zv9.66.gtf` stored in the RNA-seq/annotation directory.
7. Navigate to a region on chromosome 12 by typing `chr12:20,270,921-20,300,943` into the search box at the top of the IGV window.



Can you identify the splice junctions from the BAM file? **Slice junctions can be identified in the alignment BAM files. These are the aligned RNA-Seq reads that have skipped-bases from the reference genome (most likely introns).**

Are the junctions annotated for CBY1 consistent with the annotation? **Read alignment supports an extended length in exon 5 to the gene model (cby1-001)**

Are all annotated genes, from both RefSeq and Ensembl, expressed? **No BX000473.1-201 is not expressed**



Once tophat finishes aligning the 6h data you will need to sort the alignments found in the BAM file and then index the sorted BAM file.

```
1 samtools sort tophat/ZV9_6h/accepted_hits.bam \
    tophat/ZV9_6h/accepted_hits.sorted
2 samtools index tophat/ZV9_6h/accepted_hits.sorted.bam
```

Load the sorted BAM file into IGV, as described previously, and rename the track appropriately.

Isoform Expression and Transcriptome Assembly

There are a number of tools that perform reconstruction of the transcriptome and for this workshop we are going to use Cufflinks. Cufflinks can do transcriptome assembly either *ab initio* or using a reference annotation. It also quantifies the isoform expression in Fragments Per Kilobase of exon per Million fragments mapped (FPKM).



Cufflinks has a number of parameters in order to perform transcriptome assembly and quantification. To view them all type:

```
1 | cufflinks --help
```

We aim to reconstruct the transcriptome for both samples by using the Ensembl annotation both strictly and as a guide. In the first case Cufflinks will only report isoforms that are included in the annotation, while in the latter case it will report novel isoforms as well.

The Ensembl annotation for *Danio rerio* is available in `annotation/Danio_rerio.Zv9.66.gtf`.



The general format of the `cufflinks` command is:

```
1 | cufflinks [options]* <aligned_reads.(sam|bam)>
```

Where the input is the aligned reads (either in SAM or BAM format).



Some of the available parameters for Cufflinks that we are going to use to run Cufflinks are listed below:

- o Output directory.
 - G Tells Cufflinks to use the supplied GTF annotations strictly in order to estimate isoform annotation.
 - b Instructs Cufflinks to run a bias detection and correction algorithm which can significantly improve accuracy of transcript abundance estimates. To do this Cufflinks requires a multi-fasta file with the genomic sequences against which we have aligned the reads.
 - u Tells Cufflinks to do an initial estimation procedure to more accurately weight reads mapping to multiple locations in the genome (multi-hits).
- library-type** Before performing any type of RNA-seq analysis you need to know a few things about the library preparation. Was it done using a strand-specific

protocol or not? If yes, which strand? In our data the protocol was NOT strand specific.



Perform transcriptome assembly, strictly using the supplied GTF annotations, for the 2cells and 6h data using cufflinks:

```
1 # 2cells data (takes approx. 5mins):
2 cufflinks -o cufflinks/ZV9_2cells_gtf -G \
    annotation/Danio_rerio.Zv9.66.gtf -b \
    genome/Danio_rerio.Zv9.66.dna.fa -u --library-type fr-unstranded \
    tophat/ZV9_2cells/accepted_hits.bam
3 # 6h data (takes approx. 5mins):
4 cufflinks -o cufflinks/ZV9_6h_gtf -G annotation/Danio_rerio.Zv9.66.gtf \
    -b genome/Danio_rerio.Zv9.66.dna.fa -u --library-type fr-unstranded \
    tophat/ZV9_6h/accepted_hits.bam
```



Cufflinks generates several files in the specified output directory. Here's a short description of these files:

genes.fpkm_tracking Contains the estimated gene-level expression values.

isoforms.fpkm_tracking Contains the estimated isoform-level expression values.

skipped.gtf Contains loci skipped as a result of exceeding the maximum number of fragments.

transcripts.gtf This GTF file contains Cufflinks' assembled isoforms.

The complete documentation can be found at: http://cufflinks.cbcu.umd.edu/manual.html#cufflinks_output



So far we have forced cufflinks, by using the **-G** option, to strictly use the GTF annotations provided and thus novel transcripts will not be reported. We can get cufflinks to perform a GTF-guided transcriptome assembly by using the **-g** option instead. Thus, novel transcripts will be reported.



GTF-guided transcriptome assembly is more computationally intensive than strictly using the GTF annotations. Therefore, we have pre-computed these GTF-guided assemblies for you and have placed the results under subdirectories: `cufflinks/ZV9_2cells_gtf_guided` and `cufflinks/ZV9_6h_gtf_guided`.

You DO NOT need to run these commands. We provide them so you know how we generated the the GTF-guided transcriptome assemblies:

```
1 # 2cells guided transcriptome assembly (takes approx. 30mins):
2 cufflinks -o cufflinks/ZV9_2cells_gtf_guided -g \
    annotation/Danio_rerio.Zv9.66.gtf -b \
    genome/Danio_rerio.Zv9.66.dna.fa -u --library-type fr-unstranded \
    tophat/ZV9_2cells/accepted_hits.bam
3 # 6h guided transcriptome assembly (takes approx. 30mins):
4 cufflinks -o cufflinks/ZV9_6h_gtf_guided -g \
    annotation/Danio_rerio.Zv9.66.gtf -b \
    genome/Danio_rerio.Zv9.66.dna.fa -u --library-type fr-unstranded \
    tophat/ZV9_6h/accepted_hits.bam
```



1. Go back to IGV and load the pre-computed, GTF-guided transcriptome assembly for the `2cells` data (`cufflinks/ZV9_2cells_gtf_guided/transcripts.gtf`).
2. Rename the track as “`2cells GTF-Guided Transcripts`”.
3. In the search box type `ENSDART00000082297` in order for the browser to zoom in to the gene of interest.



Do you observe any difference between the Ensembl GTF annotations and the GTF-guided transcripts assembled by cufflinks (the “`2cells GTF-Guided Transcripts`” track)?

Yes. It appears that the Ensembl annotations may have truncated the last exon. However, our data also doesn't contain reads that span between the last two exons.

Differential Expression

One of the stand-alone tools that perform differential expression analysis is Cuffdiff. We use this tool to compare between two conditions; for example different conditions could be control and disease, or wild-type and mutant, or various developmental stages.

In our case we want to identify genes that are differentially expressed between two developmental stages; a `2cells` embryo and `6h` post fertilization.

 The general format of the cuffdiff command is:

```
1 | cuffdiff [options]* <transcripts.gtf> \
    <sample1_replicate1.sam[,...,sample1_replicateM]> \
    <sample2_replicate1.sam[,...,sample2_replicateM.sam]>
```

Where the input includes a `transcripts.gtf` file, which is an annotation file of the genome of interest, and the aligned reads (either in SAM or BAM format) for the conditions. Some of the Cufflinks options that we will use to run the program are:

- o Output directory.
- L Labels for the different conditions
- T Tells Cuffdiff that the reads are from a time series experiment.
- b Instructs Cufflinks to run a bias detection and correction algorithm which can significantly improve accuracy of transcript abundance estimates. To do this Cufflinks requires a multi-fasta file with the genomic sequences against which we have aligned the reads.
- u Tells Cufflinks to do an initial estimation procedure to more accurately weight reads mapping to multiple locations in the genome (multi-hits).
- library-type** Before performing any type of RNA-seq analysis you need to know a few things about the library preparation. Was it done using a strand-specific protocol or not? If yes, which strand? In our data the protocol was NOT strand specific.

 Run cuffdiff on the cufflinks generated BAM files for the 2cells vs. 6h data sets:

```
1 | cuffdiff -o cuffdiff/ -L ZV9_2cells,ZV9_6h -T -b \
    genome/Danio_rerio.Zv9.66.dna.fa -u --library-type fr-unstranded \
    annotation/Danio_rerio.Zv9.66.gtf \
    tophat/ZV9_2cells/accepted_hits.bam tophat/ZV9_6h/accepted_hits.bam
```

 We are interested in the differential expression at the gene level. The results are reported by Cuffdiff in the file `cuffdiff/gene_exp.diff`. Look at the first few lines of the file using the following command:

```
1 | head -n 20 cuffdiff/gene_exp.diff
```

We would like to see which are the most significantly differentially expressed genes. Therefore we will sort the above file according to the q value (corrected p value for multiple testing). The result will be stored in a different file called `gene_exp_qval.sorted.diff`.

```
1 | sort -t$'\t' -g -k 13 cuffdiff/gene_exp.diff > \
cuffdiff/gene_exp_qval.sorted.diff
```

Look again at the first few lines of the sorted file by typing:

```
1 | head -n 20 cuffdiff/gene_exp_qval.sorted.diff
```

Copy an Ensembl transcript identifier from the first two columns for one of these genes (e.g. `ENSDARG00000077178`). Now go back to the IGV browser and paste it in the search box.



Do you see any difference in the read coverage between the `2cells` and `6h` conditions that might have given rise to this transcript being called as differentially expressed?



Note that the coverage on the Ensembl browser is based on raw reads and no normalisation has taken place contrary to the FPKM values.

The read coverage of this transcript (`ENSDARG00000077178`) in the `2cells` data set is much higher than in the `6h` data set.

Appendix

Quality Value Encoding Schema

In order to use a single character to encode Phred qualities, ASCII characters are used (<http://www.ascitable.com/>). All ASCII characters have a decimal number associated with them but the first 32 characters are non-printable (e.g. backspace, shift, return, escape). Therefore, the first printable ASCII character is number 33, the exclamation mark (!). In Phred+33 encoded quality values the exclamation mark takes the Phred quality score of zero.

Early Solexa (now Illumina) sequencing needed to encode negative quality values. Because ASCII characters < 33 and non-printable, using the Phred+33 encoding was not possible. Therefore, they simply moved the offset from 33 to 64 thus inventing the Phred+64 encoded quality values. In this encoding a Phred quality of zero is denoted by the ASCII number 64 (the @ character). Since Illumina 1.8, quality values are now encoded using Phred+33.

FASTQ does not provide a way to describe what quality encoding is used for the quality values. Therefore, you should find this out from your sequencing provider. Alternatively, you may be able to figure this out by determining what ASCII characters are present in the FASTQ file. E.g the presence of numbers in the quality strings, can only mean the quality values are Phred+33 encoded. However, due to the overlapping nature of the Phred+33 and Phred+64 encoding schema it is not always possible to identify what encoding is in use. For example, if the only characters seen in the quality string are (@ABCDEFGHI), then it is impossible to know if you have really good Phred+33 encoded qualities or really bad Phred+64 encoded qualities.

For a graphical representation of the different ASCII characters used in the two encoding schema see: http://en.wikipedia.org/wiki/FASTQ_format#Encoding.

Functional Annotation of Differentially Expressed Genes

After you have performed the differential expression analysis you are interested in identifying if there is any functionality enrichment for your differentially expressed genes. On your Desktop click:

 1 | Applications >> Internet >> Firefox Web Browser

And go to the following URL: <http://david.abcc.ncifcrf.gov/> On the left side click on Functional Annotation. Then click on the Upload tab. Under the section Choose from File, click Choose File and navigate to the `cuffdiff` directory. Select the file called `globalDiffExprs_Genes_qval.01_top100.tab`. Under Step 2 select ENSEMBL_GENE_ID from the drop-down menu. Finally select Gene list and then press Submit List. Click on Gene Ontology and then click on the CHART button of the GOTERM_BP_ALL item.



Do these categories make sense given the samples we're studying? **Developmental Biology**

Browse around DAVID website and check what other information are available.
Cellular component, Molecular function, Biological Processes, Tissue expression, Pathways, Literature, Protein domains

References

1. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25, 1105-1111 (2009).
2. Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* 28, 511-515 (2010).
3. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 10, R25 (2009).
4. Roberts, A., Pimentel, H., Trapnell, C. & Pachter, L. Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics* 27, 2325-2329 (2011).
5. Roberts, A., Trapnell, C., Donaghey, J., Rinn, J. L. & Pachter, L. Improving RNA-Seq expression estimates by correcting for fragment bias. *Genome Biol.* 12, R22 (2011).

Post-Workshop Information

Access to Computational Resources

By the end of the workshop we hope you're thinking one or more of the following:

- I'm interested in dabbling some more during my day-job!
- How do I access a Linux box like the one I've been using in the workshop - I *really* don't want the hassle of setting this all up myself!
- I'm hooked! I *really* want to get down and dirty with NGS data! What computational resources do I need, what do I have access to and how do I access them?

We're ecstatic you're thinking this way and want to help guide you! However, lets take this one step at a time.

The quickest way to dabble is to use a clone of the operating system (OS) you've been using during this workshop. That means you'll have hassle-free access to a plethora of pre-installed, pre-configured bioinformatics tools. You could even set it up to contain a copy of all the workshop data and handouts etc to go through the hands-on practicals in your own time!

We have created an image file (approx. 10 GBytes in size) of the NGS Training OS for you to use as you wish:

https://swift.rc.nectar.org.au:8888/v1/AUTH_33065ff5c34a4652aa2fefb292b3195a/VMs/NGSTrainingV1.2.1.vdi

We would advise one of the following two approaches for making use of it:

- Import it into VirtualBox to setup a virtual machine (VM) on your own computer.
- Instantiate a VM on the NeCTAR Research Cloud.

Setting up a VM using VirtualBox

This approach requires the least amount of mind-bending to get up and running. However, you will need to install some software. If you do not have administrator access or your system administrator is slow or unwilling to install the software, you may find using the NeCTAR Research Cloud to be viable alternative.

This approach will use, at most, the computational resources available on your own computer. If you are analysing non-microbial organisms or performing *de novo* assemblies, you may find these resources are insufficient. If this is the case, you really should speak to someone from IT support at your institution or get in touch with a bioinformatician for advise.

The software you need is VirtualBox, a freely available, Open Source virtualisation product from Oracle (<https://www.virtualbox.org/>). This software essentially allows you to

run an operating system (the guest OS) within another (the host OS). VirtualBox is available for several different host OSes including MS Windows, OS X, Linux and Solaris (<https://www.virtualbox.org/wiki/Downloads>). Once VirtualBox is installed on your host OS, you can then install a guest OS inside VirtualBox. VirtualBox supports a lot of different OSes (https://www.virtualbox.org/wiki/Guest_OSes).

Here are the steps to setting up a VM in VirtualBox with our image file:

1. Download and install VirtualBox for your OS: <https://www.virtualbox.org/wiki/Downloads>
2. Start VirtualBox and click New to start the Create New Virtual Machine wizard
3. Give the VM a useful name like “NGS Training” and choose Linux and either Ubuntu or Ubuntu (64-bit) as the OS Type
4. Give the VM access to a reasonable amount of the host Oses memory. i.e. somewhere near the top of the green. If this value is < 2000 MB, you are likely to have insufficient memory for your NGS data analysis needs.
5. For the virtual hard disk, select “Use existing hard disk” and browse to and select the NGSTrainingV1.2.1.vdi file you downloaded.
6. Confirm remaining settings
7. Select the “NGS Training” VM and click Start to boot he machine.
8. Once booted, log into the VM as either `ubuntu` (a sudoer user; i.e. has admin rights) or as `ngstrainee` (a regular unprivileged user). See table below for passwords.

Setting up a VM using the NeCTAR Research Cloud

All Australian researchers, who are members of an institution which subscribes to the Australian Access Federation (AAF; <http://www.aaf.edu.au/>), have access to a small amount of computing resources (2 CPU's and 8 GBytes RAM) on the NeCTAR Research Cloud (<http://nectar.org.au/research-cloud>).

Login to the NeCTAR Research Cloud Dashboard

The online dashboard is a graphical interface for administering (creating, deleting, rebooting etc) your virtual machines (VMs) on the NeCTAR research cloud.

1. Go to the dashboard: <http://dashboard.rc.nectar.org.au>
2. When you see the following page, click the “Log In” button:

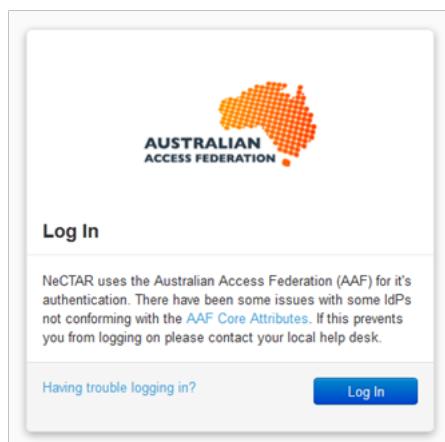


Figure 4:

3. At the following screen, simply choose your institution from the dropdown box and click “Select”. Now follow the on screen prompts and enter your regular institutional login details.

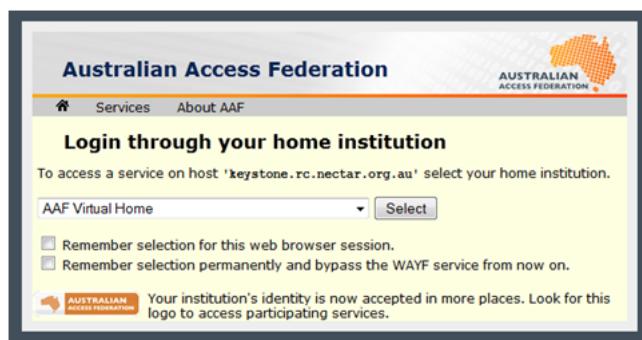


Figure 5:

4. If you see the following screen, congratulations, you have successfully logged into the NeCTAR Research Cloud dashboard!

The screenshot shows the NeCTAR Research Cloud dashboard. On the left, there's a sidebar with a logo and navigation links for 'Project' (selected), 'Manage Compute', 'Overview', 'Instances', 'Images & Snapshots' (selected), 'Access & Security', 'Object Store', 'Containers', 'Allocations', and 'New Request'. The main content area is titled 'Overview' and shows a message 'Select a month to query its usage:' with dropdown menus for 'July' and '2012', and a 'Submit' button. Below this, it displays 'Active Instances: 1 Active Memory: 8GB This Month's VCPU-Hours: 334.69 This Month's GB-Hours: 19129.48'. A 'Usage Summary' table has columns for 'Instance Name', 'VCPU', 'Disk', 'RAM', and 'Uptime'. It lists one item: 'ngstraining-1' with 1 VCPU, 8GB RAM, 100GB Disk, and 2 days, 17 hours uptime. There's also a 'Download CSV Summary' button.

Figure 6:

Instantiating Your Own VM

We will now show you how to instantiate the “NGS Training” image using your own personal cloud allocation.

1. In the NeCTAR Research Cloud dashboard, click “Images & Snapshots” to list all the publicly available images from which you can instantiate a VM. Under “Snapshots” Click the “Launch” button for the latest version of the “NGSTraining” snapshot:

The screenshot shows the Nectar cloud interface. On the left, there's a sidebar with various project and management links. The 'PROJECT pt-175' section is active. Under 'Instances & Volumes', the 'Images & Snapshots' link is highlighted with a red box. The main content area is titled 'Images & Snapshots' and contains two tables: 'Images' and 'Instance Snapshots'. The 'Images' table lists several AMI and BARE images. The 'Instance Snapshots' table lists two items, with 'NGSTrainingV1.1' highlighted by a red box.

Image Name	Type	Status	Public	Container Format	Actions
VU-ANDS-base20121022	Image	Active	Yes	AMI	<button>Launch</button>
eRes2012-c3_images	Image	Active	Yes	BARE	<button>Launch</button>
Fedora17-x86_64	Image	Active	Yes	BARE	<button>Launch</button>
Scientific Linux 6	Image	Active	Yes	OVF	<button>Launch</button>
Ubuntu 11.10 (Oneiric) amd64 UEC	Image	Active	Yes	AMI	<button>Launch</button>

Image Name	Type	Status	Public	Container Format	Actions
VU-ANDS-base20121022	Image	Active	Yes	AMI	<button>Launch</button>
NGSTrainingV1.1	Snapshot	Active	Yes	BARE	<button>Launch</button>

Figure 7:

2. You will now see a “Launch Instances” window where you are required to enter some details about how you want the VM to be setup before clicking “Launch Instance”. In the “Launch Instances” pop-up frame choose the following settings:

Server Name A human readable name for your convenience. e.g. “My NGS VM”

Flavor The resources you want to allocate to this VM. I suggest a Medium sized VM (2 CPUs and 8 GBytes RAM). This will use all your personal allocation, but anything less will probably be insufficient. You could request a new allocation of resources if you want to instantiate a larger VM with more memory.

Security Groups Select SSH.

3. Click the “Launch Instance” button

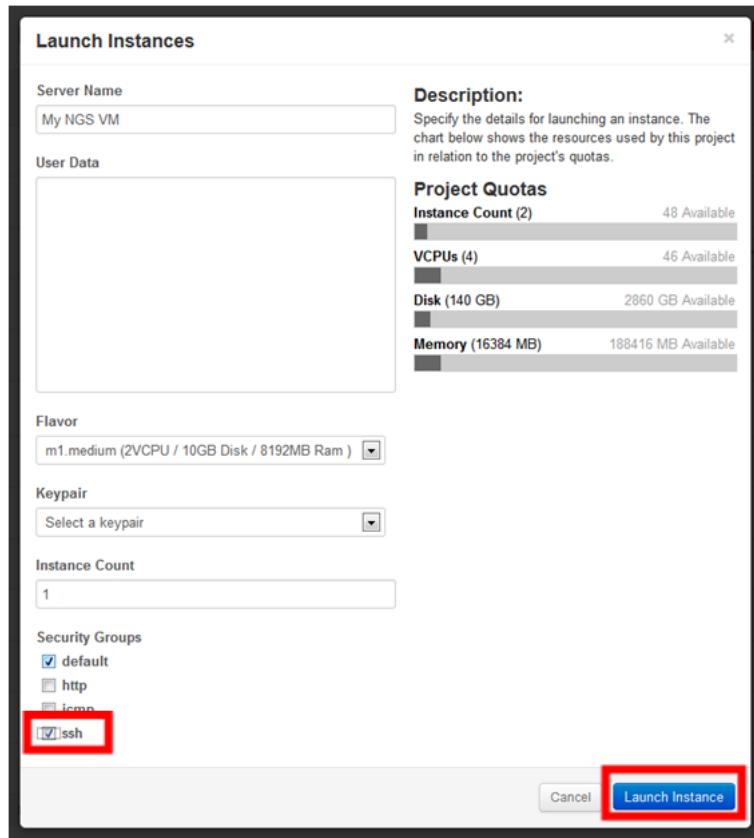


Figure 8:

4. You will be taken to the “Instances” page and you will see the “Status” and “Task” column for your new VM is “Building” and “Spawning”. Once the “IP Address” cell is populated, take a note of it as you will need it for configuring the NX Client later on.

Instances

Success: Instance "My NGS VM" launched.

Instance Name	IP Address	Size	Status	Task	Power State	Actions
My NGS VM	[REDACTED]	8GB RAM 2 VCPU 10.0GB Disk	Build	Spawning	No State	<button>Edit Instance</button>

Displaying 1 item

Project: pt-175

Manage Compute

Overview

Instances

Images & Snapshots

Access & Security

Object Store

Containers

Allocations

New Request

Figure 9:

- Once the Status and Task for the VM change to “Active and “None” respectively, your VM is powered up and is configuring itself. Congratulations, you have now instantiated a Virtual Machine! If you try to connect to the VM too quickly, you might not be successful. The OS may still be configuring itself, so give it a few minutes to finish before continuing.

	Instance Name	IP Address	Size	Status	Task	Power State	Actions
	My NGS VM	[REDACTED]	8GB RAM 2 VCPU 10.0GB Disk	Active	None	Running	<button>Edit Instance</button>

Displaying 1 item

Figure 10:

VM Stuck Building and Spawning

Sometimes, the cloud experiences a “hiccup” and a newly instantiated VM will get stuck in the “Build” and “Spawning” state (step 3) for more than a few minutes. This can be rectified by terminating the instance and creating a new VM from scratch:

1. Selecting “Terminate Instance” under the “Edit Instance” dropdown box:

	Instance Name	IP Address	Size	Status	Task	Power State	Actions
<input checked="" type="checkbox"/>	My NGS VM		8GB RAM 2 VCPU 10.0GB Disk	Build	Scheduling	No State	<input type="button" value="Edit Instance"/> <input type="button" value="Terminate Instance"/>

Figure 11:

2. Go back to step 1 of “Instantiating Your Own VM” and create the VM from scratch:

Remote Desktop with the NoMachine NX Client

During the workshop you were using the free NX client from NoMachine (<http://www.nomachine.com/>) to provide a remote desktop-like connection to VMs running on the NeCTAR Research Cloud. Therefore, we provide information on how to setup your local computer to connect to the VM you just instantiated in the steps above.

We assume that:

- You have administrator rights on your local computer for installing software.

NoMachine NX Client Installing

We show you instructions below for the MS Windows version of the NX Client, but procedures for other supported OSes (Linux, Mac OSX and Solaris) should be very similar.

1. Go to the NoMachine download page: <http://www.nomachine.com/download.php>
2. Click the download icon next to the NX Client for Windows:



Figure 12:

3. On the "NX Client for Windows" page, click the "Download package" button:

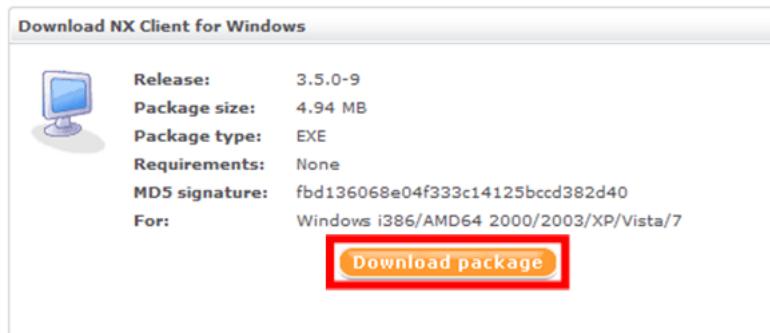


Figure 13:

4. Run the file you just downloaded (accepting defaults is fine)
5. Congratulations, you just installed the NoMachine NX Client!

NoMachine NX Client Configuration

Now we have the NoMachine NX Client installed, we need to configure a new NX "session" which will point to the VM we instantiated in the NeCTAR Research Cloud.

We assume that:

- You know the IP address of the VM you want to remote desktop into.
1. Start the NX Connection Wizard and click "Next" to advance to the "Session" settings page.

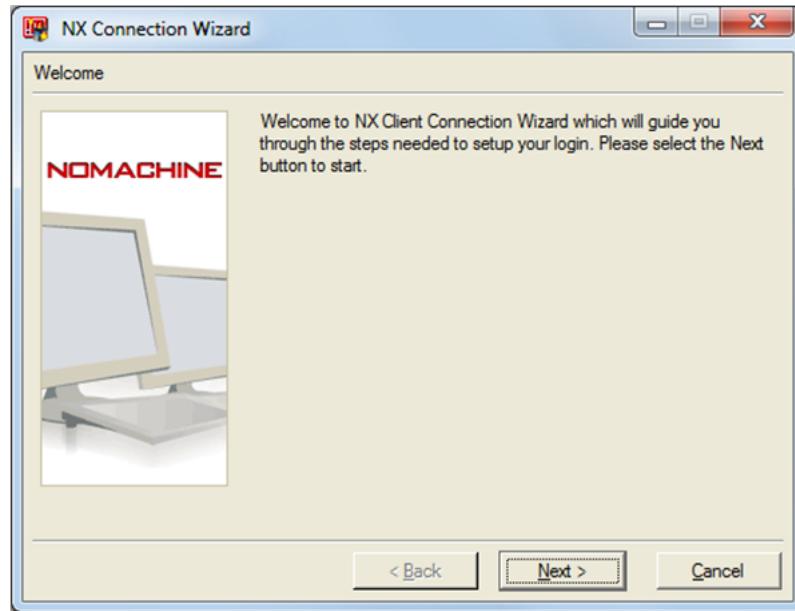


Figure 14:

2. On the "Sessions" settings page enter the following details:

Session A memorable name so you know which VM this session is pointing at. You could use the same name you chose for the VM you instantiated earlier e.g. "NGS Training".

Host Enter the IP address of the VM you instantiated on the NeCTAR Research Cloud.

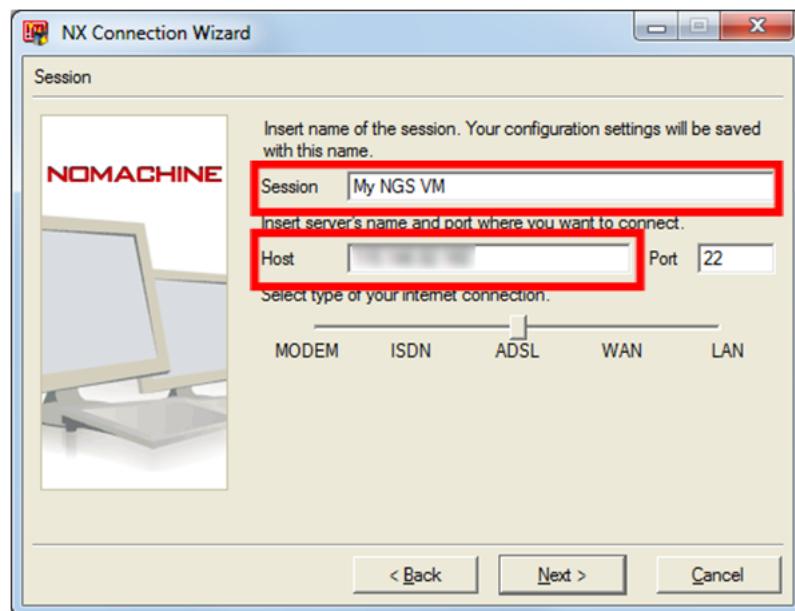


Figure 15:

3. Click "Next" to advance to the "Desktop" settings page. You should use the "Unix GNOME" setting.

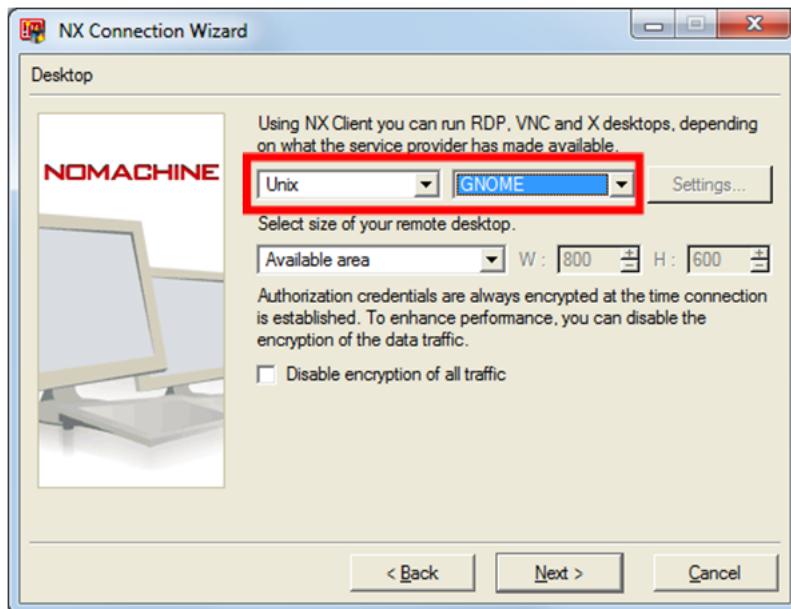


Figure 16:

4. Click "Next" and "Finish" to complete the wizard.

Connecting to a VM

If you just completed the NX Connection Wizard described above, the wizard should have opened the NX Client window. If not, run the "NX Client". You will be presented with a window like this:



Figure 17:

The "Login" and "Password" boxes in the NX Client are for user accounts setup on the VM. By default our image, from which you instantiated your VM, has two preconfigured users:

Username	Password	Notes
ngstrainee	trainee	The user you used during the workshop
ubuntu	bioubuntu	Has "sudo" privileges for performing admin tasks

Figure 18:

Unless you know what you are doing, we suggest you use the `ngstrainee` user account details to initiate an NX connection to your VM. In less than a minute, you should see an NX Window showing the desktop of your VM:

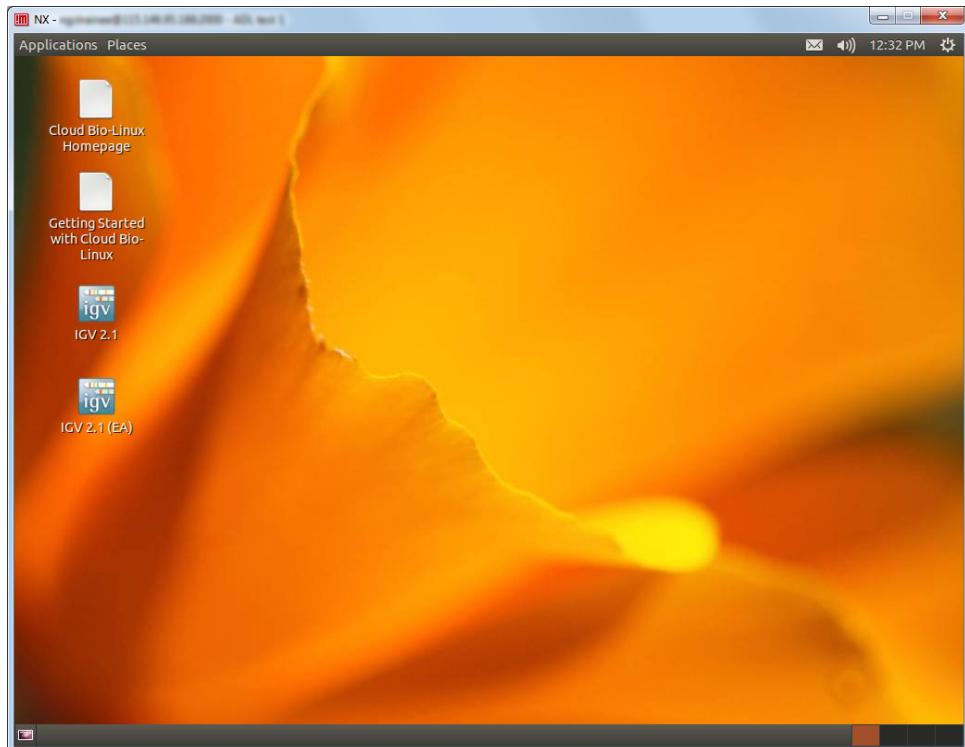


Figure 19:

NX Connection Failure

In the event that you don't get the NX Window with your VM's desktop displaying inside it. The most common errors are:

- You failed to select the "ssh" security group when instantiating the VM. You'll need to terminate the instance and create a new VM from scratch
- You failed to select "Unix GNOME" when you configured the NX Client session. You'll need to reconfigure the session using the NX Client
- Your institution's firewall blocks TCP port 22. You may need to request this port to be opened by your local network team or configure the NX client to use a proxy server.

Advanced Configuration

In the session configuration, you can configure the size of the NX Window in which the desktop of the VM is drawn:

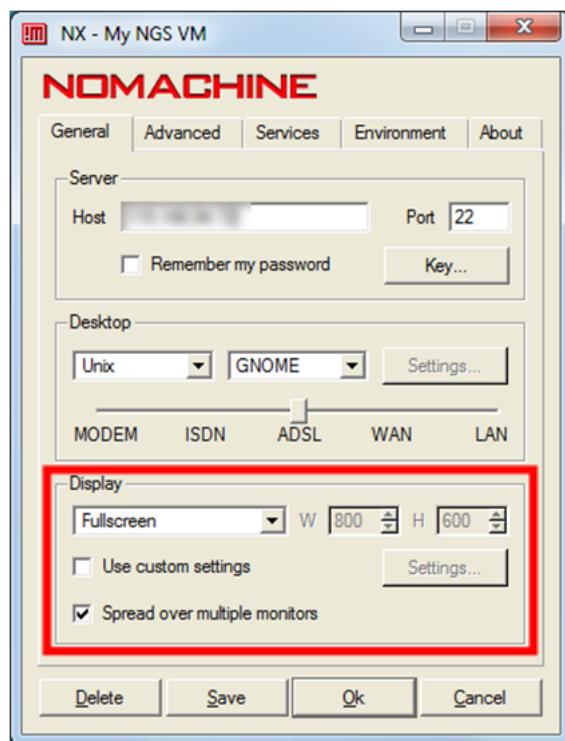


Figure 20:

This can be useful if you want to:

- Have the NX Window occupy the entire screen, without window decorations. This is often desirable if you wish to "hide" the host OS from the person sitting at the computer running the NX Client.
- Have the NX Window spread over multiple monitors.

Access to Workshop Documents

This document has been written in L^AT_EX and deposited in a public github repository (https://github.com/nathanhaigh/ngs_workshop). The documentation has been released under a Creative Commons Attribution 3.0 Unported License (see the Licence page at the beginning of this handout).

For convenience, you can access up-to-date PDF versions of the L^AT_EX documents at:

Trainee Handout

https://github.com/nathanhaigh/ngs_workshop/blob/Intro_to_Linux_and_RNA-Seq/trainee_handout.pdf

Trainer Handout

https://github.com/nathanhaigh/ngs_workshop/blob/Intro_to_Linux_and_RNA-Seq/trainer_handout.pdf

Access to Workshop Data

Once you have created a VM from our image file, either locally using VirtualBox or on the NeCTAR Research Cloud, you can configure the system with the workshop documents and data. This way you can revisit and work through this workshop in your own time.

In order to do this, we have provided you with access to a shell script which should be executed on your NGS Training VM by the `ubuntu` user. This pulls approx. 3.3 GBytes of data from the NeCTAR Cloud storage and configures the system for running this workshop:

```
1 # As the ubuntu user run the following commands:  
2 cd /tmp  
3 wget  
4 https://github.com/nathanhaigh/ngs_workshop/raw/\  
5 Intro_to_Linux_and_RNA-Seq/workshop_setup/\  
6 setup_workshop.sh bash  
7 setup_workshop.sh
```

While you're at it, you may also like to change the timezone of your VM to match that of your own. To do this simply run the following commands as the `ubuntu` user:

```
1 echo "Australia/Adelaide" | sudo tee /etc/timezone  
2 sudo dpkg-reconfigure --frontend noninteractive tzdata
```

For further information about what this script does and possible command line arguments, see the script's help:

```
1 bash setup_workshop.sh -h
```

For further information about setting up the VM for the workshop, please see:

https://github.com/nathanhaigh/ngs_workshop/blob/Intro_to_Linux_and_RNA-Seq/workshop_setup/README.md

Space for Personal Notes or Feedback

Space for Personal Notes or Feedback

Space for Personal Notes or Feedback

Space for Personal Notes or Feedback

Space for Personal Notes or Feedback
