



LEIDEN UNIVERSITY MEDICAL CENTER

RNA-seq using Galaxy

Jeroen F. J. Laros

Leiden Genome Technology Center

Department of Human Genetics

Center for Human and Clinical Genetics



Data generation.



Figure 1: HiSeq 2000.

Characteristics:

- Paired end.
 - 2×150 base pairs.
- High Throughput.
 - 150 Giga bases in 8 days.
- Cheap.

General layout of an RNA-seq pipeline.

1. Pre-alignment.
 - QC.
 - Data cleaning.

General layout of an RNA-seq pipeline.

1. Pre-alignment.
 - QC.
 - Data cleaning.
2. Alignment.
 - Use a specialised (RNA) aligner.

General layout of an RNA-seq pipeline.

1. Pre-alignment.
 - QC.
 - Data cleaning.
2. Alignment.
 - Use a specialised (RNA) aligner.
3. Expression (gene, transcripts) analysis.
 - Known transcripts.

General layout of an RNA-seq pipeline.

1. Pre-alignment.
 - QC.
 - Data cleaning.
2. Alignment.
 - Use a specialised (RNA) aligner.
3. Expression (gene, transcripts) analysis.
 - Known transcripts.
4. Transcript assembly.
 - New transcripts, alternative splicing, etc.

FastX / FastQC.

We use the Trimmomatic / FastX toolkit for data cleaning.

- Remove linker sequences.
- Clip low quality reads at the end of the read.
- Judge the read that is left over.

FastX / FastQC.

We use the Trimmomatic / FastX toolkit for data cleaning.

- Remove linker sequences.
- Clip low quality reads at the end of the read.
- Judge the read that is left over.

The FastQC tool kit is used for quality control (both before and after the data cleaning step).

- GC content.
- GC distribution.
- Quality scores distribution.
- ...

FastQC report.

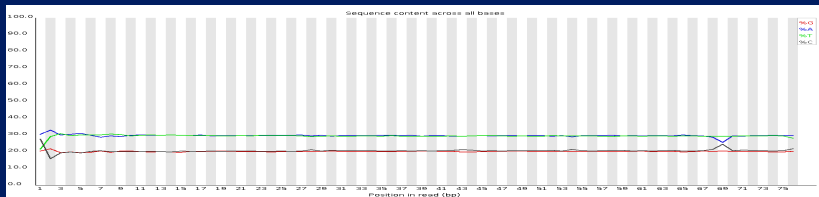


Figure 2: Per base sequence content.

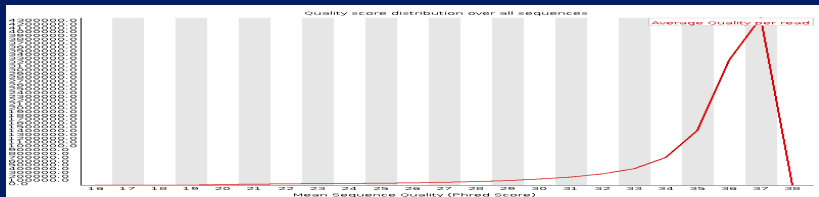


Figure 3: Per sequence quality.

Gmap.

Gmap: A Genomic Mapping and Alignment Program for mRNA and EST Sequences.

Gsnap: Genomic Short-read Nucleotide Alignment Program.

<http://research-pub.gene.com/gmap/>

Gmap.

Gmap: A Genomic Mapping and Alignment Program for mRNA and EST Sequences.

Gsnap: Genomic Short-read Nucleotide Alignment Program.

Some features:

- Split read alignment.
 - Split both ends.
 - Split a read into many pieces.
- Fast.
- Memory efficient.
 - No limit on intron size.

<http://research-pub.gene.com/gmap/>

Cufflinks.

Input:

- Aligned reads.
 - Tophat.
 - Gmap / Gsnap.

<http://cufflinks.cbc.umd.edu/>

Cufflinks.

Input:

- Aligned reads.
 - Tophat.
 - Gmap / Gsnap.

What it can do:

- Assembled transcripts.
- Estimated transcript abundance.

<http://cufflinks.cbc.umd.edu/>

Cufflinks.

Input:

- Aligned reads.
 - Tophat.
 - Gmap / Gsnap.

What it can do:

- Assembled transcripts.
- Estimated transcript abundance.

Differential expression and regulation (**cuffcompare**).

<http://cufflinks.cccb.umd.edu/>

Combining tools in a pipeline.

```
1 bwa aln -t 8 $reference $i > $i.sai
2 bwa samse $reference $i.sai $i > $i.sam
3 samtools view -bt $reference -o $i.bam $i.sam
```

Listing 1: Shell script.

Combining tools in a pipeline.

```
1 bwa aln -t 8 $reference $i > $i.sai
2 bwa samse $reference $i.sai $i > $i.sam
3 samtools view -bt $reference -o $i.bam $i.sam
```

Listing 1: Shell script.

```
1 %.sai: %.fq
2 $(BWA) aln -t $(THREADS) $(call MKREF, $@) $< > $@
3
4 %.sam: %.sai %.fq
5 $(BWA) samse $(call MKREF, $@) $^ > $@
6
7 %.bam: %.sam
8 $(SAMTOOLS) view -bt $(call MKREF, $@) -o $@ $<
```

Listing 2: Makefile.

Overview.

Data intensive biology for everyone.

- Open source.
- Web based.
 - No installation required.

<http://galaxy.psu.edu/>

<http://galaxy.nbic.nl/>

Overview.

Data intensive biology for everyone.

- Open source.
- Web based.
 - No installation required.
- Wrapper for command line utilities.
- User friendly.
- Point and click.
- Workflows.
 - Save all the steps you did in your analysis.
 - Rerun the entire analysis on a new dataset.
 - Share your workflow with other people.
 - ...

<http://galaxy.psu.edu/>

<http://galaxy.nbic.nl/>

The Galaxy GUI.

The screenshot displays the Galaxy web interface. At the top, the header reads "Galaxy / Netherlands Bioinformatics Centre" with navigation links for "Analyze Data", "Workflow", "Shared Data", "Visualization", "Help", and "User". A "Using 1%" indicator is visible in the top right.

The main area is the "GMAP (version 2.0.0)" tool configuration panel. It contains the following sections:

- Will you map to a reference genome from your history or use a built-in index?**
 - Use a built-in index (selected)
 - Built-ins were indexed using default options
- Select a reference genome:**
 - Human_UCSC_hg19_complete (selected)
 - If your genome of interest is not listed - contact Galaxy team
- Index size:**
 - Auto (selected)
 - Defaults to highest available index size
- Look for splicing involving known sites or known introns:**
 - Auto (selected)
- Select an mRNA or EST dataset to map:**
 - Auto (selected)
 - Additional mRNA or EST dataset to map: Add new additional mRNA or EST dataset to map
- Protocol for input quality scores:**
 - No quality scores (selected)
- Select the output format:**
 - SAM format (selected)
- SAM paired reads:**
 -
- Do not print headers beginning with '@':**
 -
- Print non-canonical genomic gaps greater than 20 nt in CIGAR using an STRING:**
 - Use default (selected)
- Value to put into read-group id (RG-ID) field:**
 -
- Value to put into read-group name (RG-SM) field:**
 -
- Value to put into read-group library (RG-LB) field:**
 -
- Value to put into read-group library platform (RG-PL) field:**
 -

The left sidebar lists various tool categories such as "Tools", "Jvarkit Analysis", "Graph Display Data", "Regional Variation", "Multiple regression", "Multivariate Analysis", "Evolution", "Multi Tools", "Metagenomic analyses", "FASTA manipulation", "NGS: QC and manipulation", "NGS: Mapping", "NGS: Indel Analysis", "NGS: SV/CRV Analysis", "NGS: RNA Analysis", and "NGS: eRNA Tools".

The right sidebar shows the "History" panel with a list of recent jobs, including "12: SAM to BAM on data 2", "11: Cufflinks on data 8 and data 1: assembled transcripts", "10: Cufflinks on data 8 and data 1: transcript expression", "9: Cufflinks on data 8 and data 1: gene expression", "8: SAM to BAM on data 2: converted BAM", "7: GSNAP on data 3 sam", "6: FastQC_read2.html", "5: FastQC_read1.html", "4: FastQC_read1.html", "3: reads_2.txt", "2: reads_1.txt", and "1: genes.gtf".

Figure 4: Galaxy panels.




Galaxy icons.

Figure 5: Collapsed history item.

- Eye: view.
- Pencil: edit (rename).
- Cross: delete.




- Click on the title for a more detailed view.

Galaxy icons.

34: Mpileup on data 33:   

Output

~1,100,000 genomic coordinates
format: pileup, database: hgtest

1. Chrom	2. Start	3. Base	4	5	6
chr1	25620470	N	1	^	A =
chr1	25620471	N	1	G	=
chr1	25620472	N	1	A	>
chr1	25620473	N	1	T	>
chr1	25620474	N	1	A	>
chr1	25620475	N	1	T	>

Figure 6: History item.

- Diskette: save.
- Blue looping arrow: rerun.

Outline of the practical

1. Do a typical RNA-seq analysis.
2. Workflows.
 - Rerun the analysis with no effort.



Acknowledgements:

Peter-Bram 't Hoen
Johan den Dunnen