RNA-seq with reference genome: What kind of information can be extracted from it?

Coline Billerey, Alexandre Cormier, Matthieu Defrance, Marc Deloger, Justine Guegan, Carl Herrmann, Rachel Legendre, Denis Puthier, Matthias Zytnicki
1. Differential expression analysis

2. Splicing variants detection/reconstruction/quantification

3. SNP/indels calling

4. Fusion transcripts research

5. Others (reannotation, small RNAs etc...)
1. Differential expression analysis
2. Splicing variants detection/reconstruction/quantification
3. SNP/indels calling
4. Fusion transcripts research
5. Others (reannotation, small RNAs etc...).
1. Differential expression analysis

2. Splicing variants detection/reconstruction/quantification

3. SNP/indels calling

4. Fusion transcripts research

5. Others (reannotation, small RNAs etc...)
Technical Recommandations (ENCODE standards)

1. Differential expression analysis (provide biological replicates)
   
   - $\geq 2 \times 15$ millions 75bp reads (paired-end stranded mRNA polyA sequencing protocol) per replicate

2. Splicing variants detection/reconstruction/quantification
   
   - $\geq 2 \times 50$ millions 75bp reads (paired-end stranded mRNA polyA sequencing protocol)

WARNING: asking for a mean read coverage is meaningless in RNA-seq contrary to DNA-seq
Technical Recommandations (ENCODE standards)

1. Differential expression analysis (provide biological replicates)

   ➢ ≥ 2 x 15 millions 75bp reads (paired-end stranded mRNA polyA sequencing protocol) per replicate

2. Splicing variants detection/reconstruction/quantification

   ➢ ≥ 2 x 50 millions 75bp reads (paired-end stranded mRNA polyA sequencing protocol)

WARNING: asking for a mean read coverage is meaningless in RNA-seq contrary to DNA-seq
Technical Recommandations  
(ENCODE standards)

1. Differential expression analysis (provide biological replicates)
   
   - $\geq 2 \times 15\text{ millions } 75\text{bp reads (paired-end stranded mRNA polyA sequencing protocol)}$ per replicate

2. Splicing variants detection/reconstruction/quantification
   
   - $\geq 2 \times 50\text{ millions } 75\text{bp reads (paired-end stranded mRNA polyA sequencing protocol)}$

**WARNING**: asking for a mean read coverage is meaningless in RNA-seq contrary to DNA-seq
Known Bias
(to keep in mind)

1. Be careful about experimental design, e.g.: avoid putting all the replicates in the same lane, using the same barcode for the replicates, putting different number of samples in lanes etc...

2. Non-uniformity of the per base read distribution, e.g.: Illumina Random Hexamer Priming bias visible on the 13 first bases.

3. Bias hierarchy: biological condition >> concentration > run/flowcell > lane (Library preparation ?)

4. At equivalent expression level, a long gene will have more reads than a short one.

5. Non-random coverage along the transcript.

6. Multiple hit for some reads alignments.

NGS statistics presentation by Julie AUBERT ("StatOmique" group) => IMPORTANCE OF NORMALISATION METHODS
Known Bias (to keep in mind)

1. Be careful about experimental design, e.g.: avoid putting all the replicates in the same lane, using the same barcode for the replicates, putting different number of samples in lanes etc...
2. Non-uniformity of the per base read distribution, e.g.: Illumina Random Hexamer Priming bias visible on the 13 first bases.
3. Bias hierarchy: biological condition >> concentration > run/flowcell > lane (Library preparation ?)
4. At equivalent expression level, a long gene will have more reads than a short one.
5. Non-random coverage along the transcript.
6. Multiple hit for some reads alignments.

NGS statistics presentation by Julie AUBERT ("StatOmique" group) => IMPORTANCE OF NORMALISATION METHODS
Known Bias (to keep in mind)

1. Be careful about experimental design, e.g.: avoid putting all the replicates in the same lane, using the same barcode for the replicates, putting different number of samples in lanes etc...
2. Non-uniformity of the per base read distribution, e.g.: Illumina Random Hexamer Priming bias visible on the 13 first bases.
3. Bias hierarchy: biological condition >> concentration > run/flowcell > lane (Library preparation ?)
4. At equivalent expression level, a long gene will have more reads than a short one.
5. Non-random coverage along the transcript.
6. Multiple hit for some reads alignments.

NGS statistics presentation by Julie AUBERT ("StatOmique" group) => IMPORTANCE OF NORMALISATION METHODS
Known Bias
(to keep in mind)

1. Be careful about experimental design, e.g: avoid putting all the replicates in the same lane, using the same barcode for the replicates, putting different number of samples in lanes etc...
2. Non-uniformity of the per base read distribution, e.g: Illumina Random Hexamer Priming bias visible on the 13 first bases.
3. Bias hierarchy: biological condition >> concentration > run/flowcell > lane (Library preparation ?)
4. At equivalent expression level, a long gene will have more reads than a short one.
5. Non-random coverage along the transcript.
6. Multiple hit for some reads alignments.

NGS statistics presentation by Julie AUBERT ("StatOmique" group) => IMPORTANCE OF NORMALISATION METHODS
Known Bias
(to keep in mind)

1. Be careful about experimental design, e.g.: avoid putting all the replicates in the same lane, using the same barcode for the replicates, putting different number of samples in lanes etc...
2. Non-uniformity of the per base read distribution, e.g.: Illumina Random Hexamer Priming bias visible on the 13 first bases.
3. Bias hierarchy: biological condition >> concentration > run/flowcell > lane (Library preparation ?)
4. At equivalent expression level, a long gene will have more reads than a short one.
5. Non-random coverage along the transcript.
6. Multiple hit for some reads alignments.

NGS statistics presentation by Julie AUBERT ("StatOmique" group) => IMPORTANCE OF NORMALISATION METHODS
Known Bias
(to keep in mind)

1. Be careful about experimental design, e.g.: avoid putting all the replicates in the same lane, using the same barcode for the replicates, putting different number of samples in lanes etc...
2. Non-uniformity of the per base read distribution, e.g.: Illumina Random Hexamer Priming bias visible on the 13 first bases.
3. Bias hierarchy: biological condition >> concentration > run/flowcell > lane (Library preparation ?)
4. At equivalent expression level, a long gene will have more reads than a short one.
5. Non-random coverage along the transcript.
6. Multiple hit for some reads alignments.
Known Bias
(to keep in mind)

1. Be careful about experimental design, e.g: avoid putting all the replicates in the same lane, using the same barcode for the replicates, putting different number of samples in lanes etc...
2. Non-uniformity of the per base read distribution, e.g: Illumina Random Hexamer Priming bias visible on the 13 first bases.
3. Bias hierarchy: biological condition >> concentration > run/flowcell > lane (Library preparation ?)
4. At equivalent expression level, a long gene will have more reads than a short one.
5. Non-random coverage along the transcript.
6. Multiple hit for some reads alignments.

Remember the NGS statistics presentation by Julie AUBERT ("StatOmique" group) => IMPORTANCE OF NORMALISATION METHODS
1. Differential expression analysis

2. Splicing variants detection/reconstruction/quantification

3. SNP/indels calling

4. Fusion transcripts research

5. Others (reannotation, small RNAs etc...
Differential expression analysis: Microarray vs RNA-seq

Differential expression analysis

RNA-Seq reads

Align reads to genome

Assemble transcripts *de novo*

Assemble transcripts from spliced alignments

Genome

More abundant

Less abundant

Jeudi 09/10 : 8h30-10h

Erwan CORRE : Jeudi 09/10 14h-17h30
Objective of the project: Identify the genes differentially expressed in 2 human cell lines

- Data retrieved from the ENCODE project: http://genome.ucsc.edu/cgi-bin/hgFileUi?db=hg19&g=wgEncodeCaltechRnaSeq

- 2 human cell lines:
  - Gm12878 (lymphoblastoid cell line)
  - Hct116 (colorectal carcinoma cell line)

- Each experiment is performed 2 times leading to a total of 4 samples.

- Illumina sequencing, paired-ends 2x75 bp, insert size ~200 bp - the selected genome region is the chromosome 22.
Differential expression analysis : step1

Get Data :

- Import published history called "Diff_Expr_TP_Datasets"

WARNING : ALL THE FOLLOWING DIFFERENT STEPS HAS TO BE DONE ON THE 4 SAMPLES
Differential expression analysis : step2

Quality control of datasets (FastQC) : remember the Olivier INIZAN / Alban LERMINE talk (Monday 06/10 : 14h – 15h45)

Differential expression analysis: step 2

R1 / R2 Quality Control comparison:
Differential expression analysis: step 2

REMEMBER "Known Bias n°2": Non-uniformity of the per base read distribution, e.g. Illumina Random Hexamer Priming bias visible on the 13 first bases.

"Biases in Illumina transcriptome sequencing caused by random hexamer priming", Kasper D. Hansen & al, Nucleic Acid Res. 2010
Differential expression analysis : step2

Sequence duplication level particularity of RNA-seq datasets (due to libraries which has highly variable levels of duplication) :

![Graph showing sequence duplication level particularity](image)
Differential expression analysis: step 2

Over-represented sequences (remove or not? Well... it depends):

- Biological or technical duplication?
- PCR duplicates?
- Total RNA or mRNA?
- Amplicons?
- PCR primers?
- Illumina primers?
- Barcodes?

<table>
<thead>
<tr>
<th>Sequence</th>
<th>Count</th>
<th>Percentage</th>
<th>Possible Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>CAGCGCTCTCGGGACGTTCACCATGGGGCTGGGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTACCTGACACAGGGACACAGGGAC</td>
<td>2118</td>
<td>0.4381555265934338</td>
<td>No Hit</td>
</tr>
<tr>
<td>CAGGGCTCTCGGGACGTTCACCATGGGGCTGGGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCT</td>
<td>1103</td>
<td>0.2488673741699249</td>
<td>No Hit</td>
</tr>
<tr>
<td>CAGGGCTCTCGGGACGTTCACCATGGGGCTGGGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCTGCT</td>
<td>1153</td>
<td>0.23852375928339437</td>
<td>No Hit</td>
</tr>
</tbody>
</table>

Table:

- Homo sapiens cDNA clone MGC:71261 IMAGE:4576612, complete cds
- Human (hybridoma H210) anti-hepatitis A immunoglobulin lambda chain variable region_con
- Homo sapiens cDNA clone MGC:71261 IMAGE:4576612, complete cds
- Human (hybridoma H210) anti-hepatitis A immunoglobulin lambda chain variable region_con
- Homo sapiens ribosomal protein L41 pseudogene 3 (RPL41P3) on chromosome 22
- Homo sapiens ribosomal protein L41 pseudogene 5 (RPL41P5) on chromosome 12
- Homo sapiens ribosomal protein L41 pseudogene 1 (RPL41P1) on chromosome 20

Aviesan
Alliance nationale pour les sciences de la vie et de la santé
Differential expression analysis : step 3

Pre-processing datasets (also known as "cleaning" step):

![FASTQ Quality Trimmer interface](image)

- **Window size:** 8
- **Aggregate action for window:** mean of scores
- **Trim until aggregate score is:** >=
- **Quality Score:** 20.0
Differential expression analysis: step 4

Mapping with Tophat2 => Remember Matthias ZYTNICKI presentation (Monday 06/10 10h30-11h30)
Differential expression analysis: step 4

Other useful parameters? (default parameters optimized for human datasets)

- **Do Fusion Search:**
  - No

- **Number of mismatches allowed in each segment alignment for reads mapped independently:**
  - 2

- **Allow indel search:**
  - Yes

- **Max insertion length:**
  - 3
  - The maximum insertion length.

- **Max deletion length:**
  - 3

Not all accessible via Galaxy web form:
Differential expression analysis : step4

Not all accessible via Galaxy web form (e.g tophat2 = 91 options !!) :

```bash
-v --version
-o --output-dir <string> [ default: /tophat_out ]
--bowtie <string> [ default: bowtie2 ]
--read-nm <int> [ default: 2 ]
--read-gap-length <int> [ default: 2 ]
--read-edit-dist <int> [ default: 2 ]
--read-seed-length <int> [ default: "read-edit-dist" + 1 ]
--a <int> [ default: 8 ]
--s <int> [ default: 8 ]
--g <int> [ default: 8 ]
--min-intron-length <int> [ default: 50 ]
--max-intron-length <int> [ default: 500000 ]
--g <int> [ default: 20 ]
--suppress-hits <bool> [ default: false ]
--transcripts-only <int> [ default: 60 ]
--match-multihits <int> [ default: 3 ]
--match-deletion-length <int> [ default: 3 ]
--solexas <int> [ default: 2 ]
--solexas.3-quals <int> [ default: 2 ]
--phred34-quals <int> [ default: 2 ]
--q <int> [ default: 2 ]
--int <int> [ default: 2 ]
--color <int> [ default: 2 ]
--library-type <string> [ default: fr-unstranded, fr-firststrand, fr-secondstrand ]
--num-threads <int> [ default: 1 ]
--out-dir <string> [ default: output_dir/tmp ]
--gtf <filename> [ default: gtf/off with known transcripts ]
--transcripts-only <string> [ default: map only to the transcriptome ]
--ref-juncs <filename> [ default: ref-juncs ]
--deletions <filename> [ default: deletions ]
--gene-introns <filename> [ default: gene-introns ]
--mate-inner-dist <int> [ default: 50 ]
--mate-id <int> [ default: 20 ]
--cov-search <string> [ default: <dir> ]
--cov-search <string> [ default: <dir> ]
--cov-search <string> [ default: <dir> ]
--cov-search <string> [ default: <dir> ]
--keep-temp <bool> [ default: false ]
--v <int> [ default: 2 ]
```
Differential expression analysis: step 4

Tophat

Map reads on whole genome with Bowtie (FM-index)

De novo assembly from mapped reads:
Identification of potential exons

Generate possible splice sites and exons

Validate exon-exon junction

Map reads against these junctions to confirm them
## Differential expression analysis: step4

Remember Alban LERMINE presentation (Monday 06/10: 14h – 15h45)

### SAM/BAM aligned format

- **SAM Format**: aligned format, human readable

```plaintext
@SQ SN:chr12 LN:133851895

@RG ID:Sample_ID LB:Sample_Library PL:ILLUMINA SM:Sample_Name PU:Platform_Unit
```

<table>
<thead>
<tr>
<th>Read name</th>
<th>Flag</th>
<th>Chr</th>
<th>5' pos</th>
<th>MAPQ</th>
<th>Cigar</th>
<th>paired 5' pos of the mate</th>
<th>Insert size</th>
</tr>
</thead>
<tbody>
<tr>
<td>ERR166338.1</td>
<td>99</td>
<td>chr12</td>
<td>82670685</td>
<td>23</td>
<td>101M</td>
<td>= 82670850</td>
<td>266</td>
</tr>
</tbody>
</table>

`GCCCCTGGGGATGTTTTGCACCAAGCCACTGTCTCCAGCTGG` sequence

`BBC@GIHGCFCEHEAIEIFFGGEONDNJFINIONHNGJNNNNKNJN` Base quality

**RG:Z:Sample_ID**

XT:A:U NM:i:0 X0:i:1 X1:i:1 XM:i:0 XO:i:0 XG:i:0 MD:Z:100 XA:Z tags

**Group affiliation**

- **BAM Format**: Binary SAM Format (not human readable but compressed = smaller)
Differential expression analysis : step 4

SAM "FLAG" and "CIGAR" (Filter SAM ? Website explain flags ?) :

FLAG: bitwise FLAG. Each bit is explained in the following table:

<table>
<thead>
<tr>
<th>Bit</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>0x1</td>
<td>template having multiple segments in sequencing</td>
</tr>
<tr>
<td>0x2</td>
<td>each segment properly aligned according to the aligner</td>
</tr>
<tr>
<td>0x4</td>
<td>segment unmapped</td>
</tr>
<tr>
<td>0x8</td>
<td>next segment in the template unmapped</td>
</tr>
<tr>
<td>0x10</td>
<td>SEQ being reverse complemented</td>
</tr>
<tr>
<td>0x20</td>
<td>SEQ of the next segment in the template being reversed</td>
</tr>
<tr>
<td>0x40</td>
<td>the first segment in the template</td>
</tr>
<tr>
<td>0x80</td>
<td>the last segment in the template</td>
</tr>
<tr>
<td>0x100</td>
<td>secondary alignment</td>
</tr>
<tr>
<td>0x200</td>
<td>not passing quality controls</td>
</tr>
<tr>
<td>0x400</td>
<td>PCR or optical duplicate</td>
</tr>
<tr>
<td>0x800</td>
<td>supplementary alignment</td>
</tr>
</tbody>
</table>
### Differential expression analysis: step4

**SAM "FLAG"**

This utility explains SAM flags in plain English.

#### Explanation:
- **73** read paired
- **133** read unmapped
- **89** read mapped in proper pair
- **121** mate unmapped
- **165** read reverse strand
- **181** mate reverse strand
- **101** first in pair
- **137** second in pair
- **77** not primary alignment
- **141** read fails platform/vendor quality checks
- **69** read is PCR or optical duplicate
- **130** supplementary alignment

#### Summary:
- read paired
- mate unmapped
- first in pair

---

<table>
<thead>
<tr>
<th>Flag</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>73</td>
<td>read paired</td>
</tr>
<tr>
<td>133</td>
<td>read unmapped</td>
</tr>
<tr>
<td>89</td>
<td>read mapped in proper pair</td>
</tr>
<tr>
<td>121</td>
<td>mate unmapped</td>
</tr>
<tr>
<td>165</td>
<td>read reverse strand</td>
</tr>
<tr>
<td>181</td>
<td>mate reverse strand</td>
</tr>
<tr>
<td>101</td>
<td>first in pair</td>
</tr>
<tr>
<td>137</td>
<td>second in pair</td>
</tr>
<tr>
<td>77</td>
<td>not primary alignment</td>
</tr>
<tr>
<td>141</td>
<td>read fails platform/vendor quality checks</td>
</tr>
<tr>
<td>69</td>
<td>read is PCR or optical duplicate</td>
</tr>
<tr>
<td>130</td>
<td>supplementary alignment</td>
</tr>
</tbody>
</table>

---

**Link:** [broadinstitute.github.io/picard/explain-flags.html](http://broadinstitute.github.io/picard/explain-flags.html)
Differential expression analysis : step4

SAM "CIGAR" :

CIGAR: CIGAR string. The CIGAR operations are given in the following table (set ‘*’ if unavailable):

<table>
<thead>
<tr>
<th>Op</th>
<th>BAM</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>M</td>
<td>0</td>
<td>alignment match (can be a sequence match or mismatch)</td>
</tr>
<tr>
<td>I</td>
<td>1</td>
<td>insertion to the reference</td>
</tr>
<tr>
<td>D</td>
<td>2</td>
<td>deletion from the reference</td>
</tr>
<tr>
<td>N</td>
<td>3</td>
<td>skipped region from the reference</td>
</tr>
<tr>
<td>S</td>
<td>4</td>
<td>soft clipping (clipped sequences present in SEQ)</td>
</tr>
<tr>
<td>H</td>
<td>5</td>
<td>hard clipping (clipped sequences NOT present in SEQ)</td>
</tr>
<tr>
<td>P</td>
<td>6</td>
<td>padding (silent deletion from padded reference)</td>
</tr>
<tr>
<td>=</td>
<td>7</td>
<td>sequence match</td>
</tr>
<tr>
<td>X</td>
<td>8</td>
<td>sequence mismatch</td>
</tr>
</tbody>
</table>

e.g : spliced read CIGAR string =

616L7AAXX_HWUSI-EAS627_0005:2:69:6860:19812  81  chr22  16258252  50  52M4600N23M   =  1
Differential expression analysis: step 4

SAM "TAG":

<table>
<thead>
<tr>
<th>Tag</th>
<th>Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>XT</td>
<td>i</td>
<td>Reserved fields for end users (together with Y7 and Z7)</td>
</tr>
<tr>
<td>AP</td>
<td>i</td>
<td>The smallest template-independent mapping quality of segments in the read</td>
</tr>
<tr>
<td>AS</td>
<td>i</td>
<td>Alignment score generated by aligner</td>
</tr>
<tr>
<td>BC</td>
<td>Z</td>
<td>Barcode sequence, with any quality scores stored in the QT tag.</td>
</tr>
<tr>
<td>BQ</td>
<td>Z</td>
<td>Offset to base alignment quality (BAQ), of the same length as the read sequence. At the i-th read base, BAQ = Qi − (BQ, − 64) where Qi is the i-th base quality.</td>
</tr>
<tr>
<td>C</td>
<td>Z</td>
<td>Reference name of the next hit; ‘@’ for the same chromosome</td>
</tr>
<tr>
<td>CM</td>
<td>i</td>
<td>Edit distance between the color sequence and the color reference (see also CM)</td>
</tr>
<tr>
<td>CQ</td>
<td>Z</td>
<td>Color read quality on the original strand of the read. Same encoding as QUAL; same length as CS.</td>
</tr>
<tr>
<td>CS</td>
<td>Z</td>
<td>Color read sequence on the original strand of the read. The primer base must be included.</td>
</tr>
<tr>
<td>CT</td>
<td>Z</td>
<td>Complete read annotation tag, used for consensus annotation dummy features.5</td>
</tr>
<tr>
<td>E2</td>
<td>Z</td>
<td>The 2nd most likely base calls. Same encoding and same length as QUAL.</td>
</tr>
<tr>
<td>FI</td>
<td>i</td>
<td>The index of segment in the template.</td>
</tr>
<tr>
<td>FS</td>
<td>Z</td>
<td>Segment suffix.</td>
</tr>
<tr>
<td>FZ</td>
<td>Z, S</td>
<td>Flow signal intensities on the original strand of the read; stored as (uint16_t) round(value * 100.0).</td>
</tr>
<tr>
<td>LB</td>
<td>Z</td>
<td>Library. Value to be consistent with the header RG-LB tag if @RG is present.</td>
</tr>
<tr>
<td>H0</td>
<td>i</td>
<td>Number of perfect hits</td>
</tr>
<tr>
<td>H1</td>
<td>i</td>
<td>Number of 1-difference hits (see also HM)</td>
</tr>
<tr>
<td>H2</td>
<td>i</td>
<td>Number of 2-difference hits</td>
</tr>
<tr>
<td>H3</td>
<td>i</td>
<td>Query hit index, indicating the alignment record is the i-th one stored in SAM</td>
</tr>
<tr>
<td>IH</td>
<td>i</td>
<td>Number of stored alignments in SAM that contains the query in the current record</td>
</tr>
<tr>
<td>MC</td>
<td>Z</td>
<td>CIGAR string for mate/next segment</td>
</tr>
<tr>
<td>MQ</td>
<td>i</td>
<td>Mapping quality of the mate/next segment</td>
</tr>
<tr>
<td>NH</td>
<td>i</td>
<td>Number of reported alignments that contains the query in the current record</td>
</tr>
<tr>
<td>NM</td>
<td>i</td>
<td>Edit distance to the reference, including ambiguous bases but excluding clipping</td>
</tr>
<tr>
<td>NQ</td>
<td>Z</td>
<td>Original base quality (usually before recalibration). Same encoding as QUAL.</td>
</tr>
<tr>
<td>OP</td>
<td>i</td>
<td>Original mapping position (usually before realignment)</td>
</tr>
<tr>
<td>OC</td>
<td>Z</td>
<td>Original CIGAR (usually before realignment)</td>
</tr>
<tr>
<td>PG</td>
<td>Z</td>
<td>Program. Value matches the header PG-ID tag if @PG is present.</td>
</tr>
</tbody>
</table>
| P    | Z    | Phred likelihood of the template, conditional on both the mapping being correct and set FLAG bit 0x10. The \\k\

1For example, a byte array [0x1a,0xe3,0x10] corresponds to a Hex string ‘1AE301’.  
2Explicit typing easy to format parsing and helps reduce the file size when SAM is converted to BAM.  

---

2For example, a byte array [0x1a,0xe3,0x10] corresponds to a Hex string ‘1AE301’.  
3Explicit typing easy to format parsing and helps reduce the file size when SAM is converted to BAM.  
4For example, a byte array [0x1a,0xe3,0x10] corresponds to a Hex string ‘1AE301’.  
5Explicit typing easy to format parsing and helps reduce the file size when SAM is converted to BAM.  
6For example, a byte array [0x1a,0xe3,0x10] corresponds to a Hex string ‘1AE301’.  
7Explicit typing easy to format parsing and helps reduce the file size when SAM is converted to BAM.  
8For example, a byte array [0x1a,0xe3,0x10] corresponds to a Hex string ‘1AE301’.  
9Explicit typing easy to format parsing and helps reduce the file size when SAM is converted to BAM.
Differential expression analysis : step5

Mapping statistics with SAMtools flagstat :

![Image of SAMtools flagstat tool with BAM file input]

- **flagstat** version 1.0.1
- **BAM File to Convert:** 16: Tophat2 on Gm12878_rep1 accepted_hits

Note: The image provides an example of how to use the SAMtools flagstat tool to generate mapping statistics from a BAM file.
Differential expression analysis : step 5

Mapping statistics with SAMtools flagstat:

890094 + 0 in total (QC-passed reads + QC-failed reads)
0 + 0 duplicates
890094 + 0 mapped (100.00%:nan%)
890094 + 0 paired in sequencing
445247 + 0 read1
444847 + 0 read2
646290 + 0 properly paired (72.61%:nan%)
860352 + 0 with itself and mate mapped
29742 + 0 singletons (3.34%:nan%)
0 + 0 with mate mapped to a different chr
0 + 0 with mate mapped to a different chr (mapQ>=5)

Properly paired reads =

- Insert size \leq \text{max\_insert\_size}
- R1 / R2 mapped on same chromosome
- R1 \Rightarrow \text{R2}
Differential expression analysis: step 6

Visualizing mapped reads with IGV: Remember Elodie GIRARD presentation (Tuesday 07/10 10h30 -12h)
Differential expression analysis: step 6

Visualizing mapped reads with IGV: Remember Elodie GIRARD presentation (Tuesday 07/10 10h30 -12h)
Differential expression analysis: step 6

Color by tag (e.g.: NM)
Differential expression analysis: step 7

Counting reads on each gene with HTSeq-count:

http://www-huber.embl.de/users/anders/HTSeq/doc/count.html
Differential expression analysis : step7

Mode ? :
Differential expression analysis: step 7

Stranded? : In which case the read will be counted for this gene?

- Stranded = "No" => Easiest way => All the reads that are "congruent" with the "Mode" selected previously are counted for the gene.

- Stranded = "Yes" => "congruent" with the "Mode" AND "congruent" with the gene orientation => If the gene is annotated as belonging on the "+" strand, the read has to map on the "+" strand too.

- Stranded = "Reverse" => as the term seem to signify => "congruent" with the "Mode" BUT NOT "congruent" with the gene orientation => If the gene is annotated as belonging on the "+" strand, the read has to map on the "-" strand.
Differential expression analysis : step 7

Sometimes, 1 figure can tell more than 450 characters (551 with spaces):

**Single end:**

- stranded = no
- stranded = yes
- stranded = reverse

**Paired-end:**

- stranded = no
- stranded = yes
- stranded = reverse

Read1

Read2
Differential expression analysis: step 7

<table>
<thead>
<tr>
<th>Segname</th>
<th>Source</th>
<th>Feature</th>
<th>Start</th>
<th>End</th>
<th>Score</th>
<th>Strand</th>
<th>Frame</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td>chr22</td>
<td>hg19_refGene</td>
<td>exon</td>
<td>16162066</td>
<td>16162388</td>
<td>0.000000</td>
<td>+</td>
<td></td>
<td>gene_id &quot;NR_073459&quot;; transcript_id &quot;NR_073459&quot;;</td>
</tr>
<tr>
<td>chr22</td>
<td>hg19_refGene</td>
<td>exon</td>
<td>16164482</td>
<td>16164569</td>
<td>0.000000</td>
<td>+</td>
<td></td>
<td>gene_id &quot;NR_073459&quot;; transcript_id &quot;NR_073459&quot;;</td>
</tr>
<tr>
<td>chr22</td>
<td>hg19_refGene</td>
<td>exon</td>
<td>16171952</td>
<td>16172826</td>
<td>0.000000</td>
<td>+</td>
<td></td>
<td>gene_id &quot;NR_073459&quot;; transcript_id &quot;NR_073459&quot;;</td>
</tr>
<tr>
<td>chr22</td>
<td>hg19_refGene</td>
<td>exon</td>
<td>16162066</td>
<td>16162388</td>
<td>0.000000</td>
<td>+</td>
<td></td>
<td>gene_id &quot;NR_073460&quot;; transcript_id &quot;NR_073460&quot;;</td>
</tr>
<tr>
<td>chr22</td>
<td>hg19_refGene</td>
<td>exon</td>
<td>16164482</td>
<td>16164569</td>
<td>0.000000</td>
<td>+</td>
<td></td>
<td>gene_id &quot;NR_073460&quot;; transcript_id &quot;NR_073460&quot;;</td>
</tr>
<tr>
<td>chr22</td>
<td>hg19_refGene</td>
<td>exon</td>
<td>16171952</td>
<td>16172826</td>
<td>0.000000</td>
<td>+</td>
<td></td>
<td>gene_id &quot;NR_073460&quot;; transcript_id &quot;NR_073460&quot;;</td>
</tr>
<tr>
<td>chr22</td>
<td>hg19_refGene</td>
<td>exon</td>
<td>16256332</td>
<td>16256677</td>
<td>0.000000</td>
<td>-</td>
<td></td>
<td>gene_id &quot;NM_001136213&quot;; transcript_id &quot;NM_001136213&quot;;</td>
</tr>
<tr>
<td>chr22</td>
<td>hg19_refGene</td>
<td>stop_codon</td>
<td>16258186</td>
<td>16258188</td>
<td>0.000000</td>
<td>-</td>
<td></td>
<td>gene_id &quot;NM_001136213&quot;; transcript_id &quot;NM_001136213&quot;;</td>
</tr>
<tr>
<td>chr22</td>
<td>hg19_refGene</td>
<td>CDS</td>
<td>16258189</td>
<td>16258303</td>
<td>0.000000</td>
<td>- 1</td>
<td></td>
<td>gene_id &quot;NM_001136213&quot;; transcript_id &quot;NM_001136213&quot;;</td>
</tr>
<tr>
<td>chr22</td>
<td>hg19_refGene</td>
<td>exon</td>
<td>16258185</td>
<td>16258303</td>
<td>0.000000</td>
<td>-</td>
<td></td>
<td>gene_id &quot;NM_001136213&quot;; transcript_id &quot;NM_001136213&quot;;</td>
</tr>
<tr>
<td>chr22</td>
<td>hg19_refGene</td>
<td>CDS</td>
<td>16266792</td>
<td>16267095</td>
<td>0.000000</td>
<td>- 0</td>
<td></td>
<td>gene_id &quot;NM_001136213&quot;; transcript_id &quot;NM_001136213&quot;;</td>
</tr>
<tr>
<td>chr22</td>
<td>hg19_refGene</td>
<td>exon</td>
<td>16266929</td>
<td>16267095</td>
<td>0.000000</td>
<td>-</td>
<td></td>
<td>gene_id &quot;NM_001136213&quot;; transcript_id &quot;NM_001136213&quot;;</td>
</tr>
<tr>
<td>chr22</td>
<td>hg19_refGene</td>
<td>CDS</td>
<td>16268137</td>
<td>16268181</td>
<td>0.000000</td>
<td>- 0</td>
<td></td>
<td>gene_id &quot;NM_001136213&quot;; transcript_id &quot;NM_001136213&quot;;</td>
</tr>
<tr>
<td>chr22</td>
<td>hg19_refGene</td>
<td>exon</td>
<td>16268137</td>
<td>16268181</td>
<td>0.000000</td>
<td>- 2</td>
<td></td>
<td>gene_id &quot;NM_001136213&quot;; transcript_id &quot;NM_001136213&quot;;</td>
</tr>
<tr>
<td>chr22</td>
<td>hg19_refGene</td>
<td>CDS</td>
<td>16269873</td>
<td>16259943</td>
<td>0.000000</td>
<td>-</td>
<td></td>
<td>gene_id &quot;NM_001136213&quot;; transcript_id &quot;NM_001136213&quot;;</td>
</tr>
</tbody>
</table>
Differential expression analysis : step 8

Differential expression analysis with DESeq:

Differential expression analysis: step 8

Differential expression analysis with DESeq:

- **HTSeq-count results:**
  - Replicate 1: hseq-count on Gm12878_rep1
  - Replicate 2: hseq-count on Gm12878_rep2

- **Replicate condition:**
  - Condition one

- **Replicate label name:**
  - C_rep1
  - C_rep2

You can specify a label for your replicates.

- **Remove replicate read count 1**
- **Remove replicate read count 2**

[DESeq](http://bioconductor.org/packages/release/bioc/html/DESeq.html)
Differential expression analysis : step 8

Differential expression analysis with DESeq:

Differential expression analysis : step 8

DESeq advanced mode ? :

DESeq advanced mode:
Full parameter list
Use default settings or set custom values for any DESeq parameter.

DESeq mode:
maximum
How variance estimate is computed with respect to the fitted line. See DESeq manual for more explain.

DESeq method:
pooled
How samples are pooled to estimate dispersion. If no replicates use 'blind'.

DESeq fit type:
parametric
refers to the model. 'Local' is the published model, 'parametric' is glm–based (may not converge).
Differential expression analysis: step 8

Dispersion =

- Data variance estimated from the mean:
  - (Biological variation coefficient)^2
  - E.g.: if a gene expression variation is mean +/- 20% between all samples then dispersion for this gene is equal to 0.2^2 = 0.04

- Correspond to inter-samples variation + measure uncertainty
Differential expression analysis: step 8

DESeq parameters (2013 default, 2010 default):

- **Mode**: underlying hypothesis choice explaining dispersion values
  - *Fit-only*: variations come from sampling effect
  - *Maximum*: variations come from sampling + "real" variations of genes expression between 2 conditions
  - *Gene-est-only*: same as "Maximum" but for \( \geq 7 \) replicates

- **Method**: Experiment design impact choice (e.g.: which condition samples are related to) for dispersion estimation
  - *Blind*: do not take into account the experimental design
  - *Pooled*: take into account all the conditions
  - *Pooled-CR*: use a Cox-Reid likelihood estimator (when samples are paired)
  - *Per-condition*: mean dispersion estimation for each condition

- **Type**: dispersion estimation function choice
  - *Parametric*: estimation by a binomial negative law
  - *Local*: estimation by a local regression, when "Parametric" seems not working (e.g.: high intra-condition variability). DO NOT USE WITH LOW COUNTS
  - *Mean*: estimation by mean (Try to avoid this option)
Differential expression analysis : step 8

3 outputs:

- DE_up_genes (tabulate file)
- DE_down_genes (tabulate file)
- Deseq_result (HTML)
### Differential expression analysis: step 8

**DE_up_genes & DE_down_genes tables:**

<table>
<thead>
<tr>
<th>id</th>
<th>baseMean</th>
<th>baseMean_Gm12878</th>
<th>baseMean_Hct116</th>
<th>foldChange_Hct116/Gm12878</th>
<th>log10FoldChange</th>
<th>pval</th>
<th>padj</th>
</tr>
</thead>
<tbody>
<tr>
<td>NM_152243</td>
<td>2685.30</td>
<td>1,463</td>
<td>5369.1</td>
<td>3670.87</td>
<td>3.565</td>
<td>0.00e+00</td>
<td>0.00e+00</td>
</tr>
<tr>
<td>NR_039988</td>
<td>107.63</td>
<td>0.667</td>
<td>214.6</td>
<td>321.58</td>
<td>2.507</td>
<td>3.37e-53</td>
<td>3.85e-52</td>
</tr>
<tr>
<td>NM_019601</td>
<td>209.51</td>
<td>2.002</td>
<td>417.0</td>
<td>208.32</td>
<td>2.319</td>
<td>1.84e-98</td>
<td>3.01e-97</td>
</tr>
<tr>
<td>NM_052906</td>
<td>1758.64</td>
<td>17.963</td>
<td>3499.3</td>
<td>194.81</td>
<td>2.290</td>
<td>0.00e+00</td>
<td>0.00e+00</td>
</tr>
<tr>
<td>NM_138435</td>
<td>208.21</td>
<td>2.258</td>
<td>414.2</td>
<td>183.42</td>
<td>2.263</td>
<td>2.44e-99</td>
<td>4.13e-98</td>
</tr>
<tr>
<td>NM_000854</td>
<td>43.83</td>
<td>0.667</td>
<td>87.0</td>
<td>130.37</td>
<td>2.115</td>
<td>2.67e-21</td>
<td>1.53e-20</td>
</tr>
<tr>
<td>NM_014351</td>
<td>80.09</td>
<td>1.591</td>
<td>158.6</td>
<td>99.70</td>
<td>1.999</td>
<td>7.51e-38</td>
<td>6.59e-37</td>
</tr>
</tbody>
</table>

- **id** = gene identification
- **baseMean** = normalized mean counts
- **foldChange** = mean ratios
- **pval** = p-value
- **padj** = adjusted p-value
Differential expression analysis : step 8

Total read counts for each sample = depth of coverage variations (better when minimized)
Differential expression analysis : step 8

Proportion of null counts for each condition = genes proportion that will not be analyzed (better when low)
Differential expression analysis: step 8

Violinplot (boxplot + density plot) of raw/normalized counts for each sample:
Differential expression analysis : step 8

Heatmap of euclidean distances between samples
Differential expression analysis: step 8

Heatmap of count data of the 30 most highly expressed genes
Differential expression analysis: step 8

Distribution of p-values

NOT differentially expressed genes

Low count genes
Differential expression analysis: step 8

Control plot of estimated (red line) / observed dispersion (black dots)
Differential expression analysis : step 8

MAplot showing normalised mean versus log2 (fold change) for condition two versus condition one
Differential expression analysis : step8

Enriched volcano plot showing differentially expressed genes satisfying p-value and foldchange thresholds provided:
Upper and right panels show, respectively, the projection of log fold change density and the log p-value density.
Differential expression analysis : DONE

YAHOO !!!

What’s next ?