先見基因科技股份有限公司

高通量定序

轉錄體組分析報告

RNAseq

Bioinformatic Analysis Report

Insight Genomics Co., Ltd
Email: service@i-genomics.com.tw
Phone: +886 6 2095869

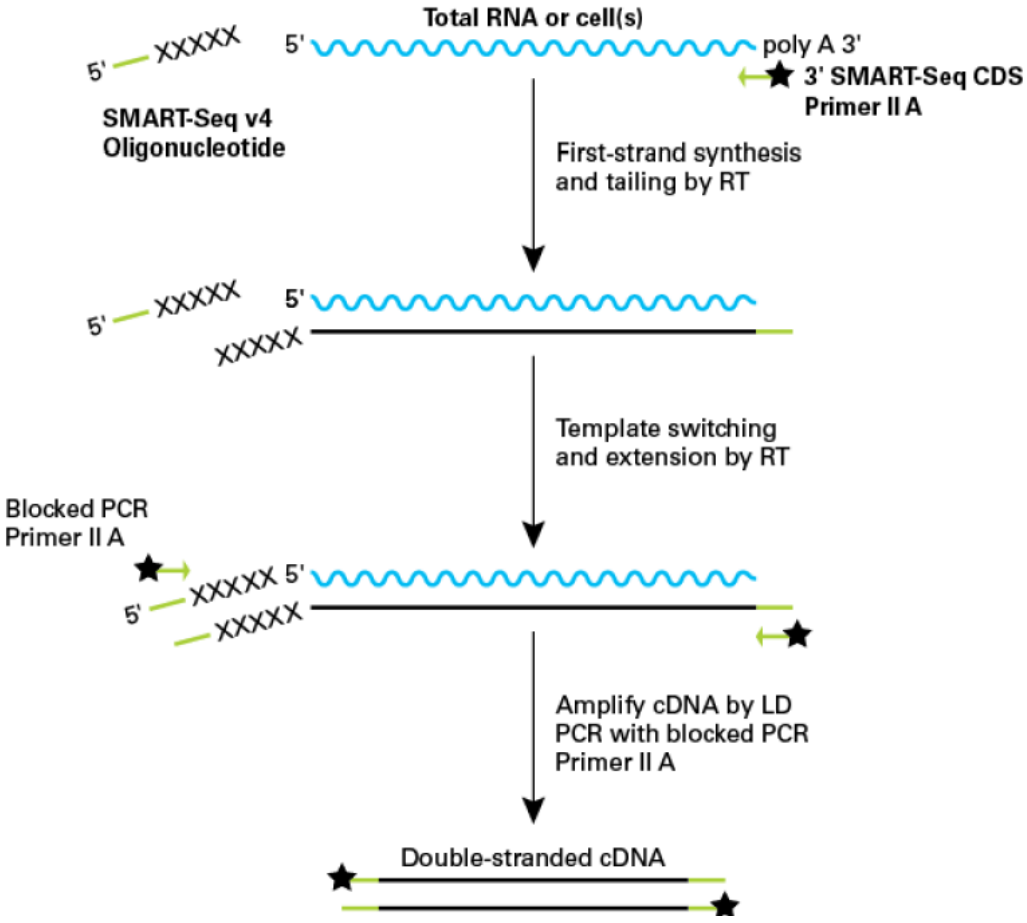
基本資料 (General information)

客戶資料 (Customer information)			
研究單位/部門 (Department)		單位負責人 (PI name)	
電話 (Phone number)		單位聯絡人 (Customer name)	
電子郵件 (E-mail)			
樣本資料 (Sample information)			
樣本編號 (Sample #)	樣本名稱 (Sample name)	定序種類 (Library type)	備註 (Note)
sample_01	Sample 1	CUST_001	Mus musculus
sample_02	Sample 2	CUST_001	
sample_03	Sample 3	CUST_001	
sample_04	Sample 4	CUST_001	
sample_05	Sample 5	CUST_001	
sample_06	Vsample 1	CUST_001	
sample_07	Vsample 2	CUST_001	
sample_08	Vsample 3	CUST_001	
sample_09	Vsample 4	CUST_001	
sample_10	Vsample 5	CUST_001	

定序與分析流程 (Sequencing and analyzing pipeline)

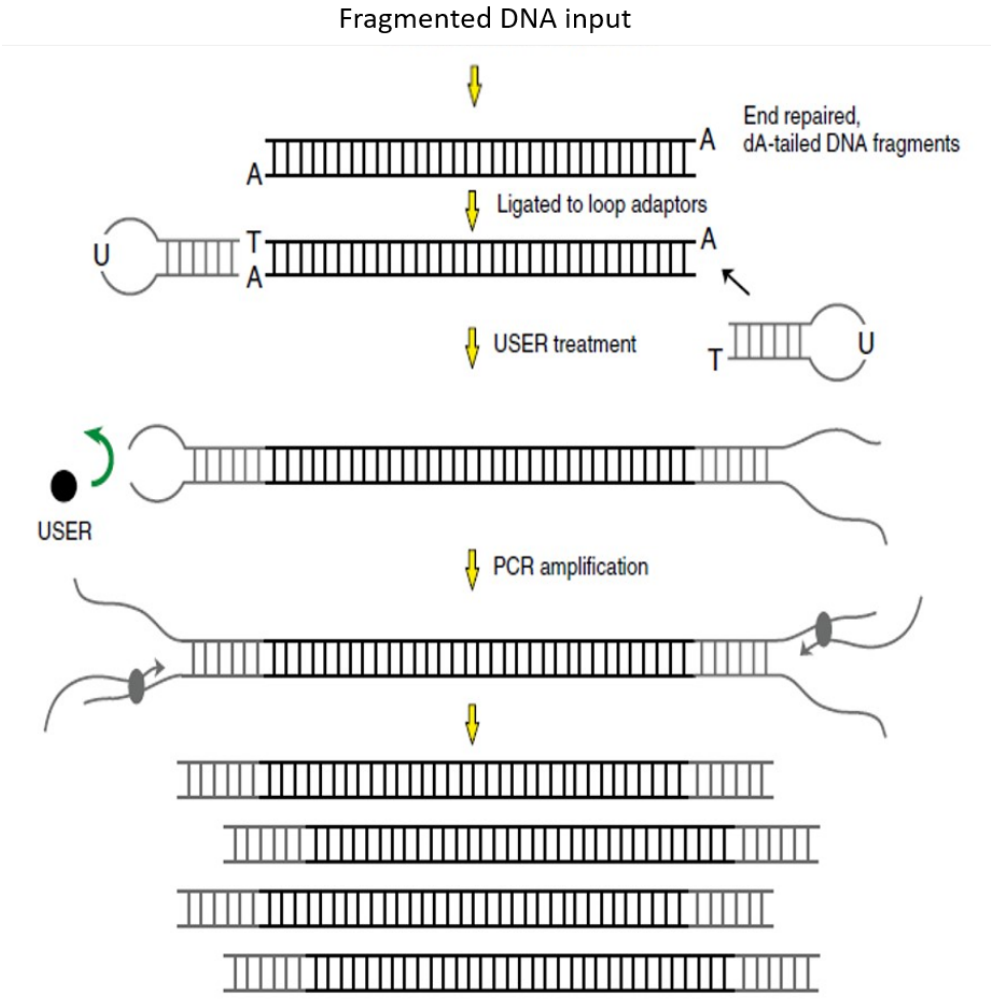
A. cDNA Synthesis

The overall processes of RNA isolation are illustrated as bellow. In brief, the samples were lysis and the double strand cDNA was constructed by using SMART-Seq® v4 Ultra® Low Input RNA Kit for Sequencing (Takara Bio USA, Inc, Cat.: #634888). The experimental process is listed as follows:

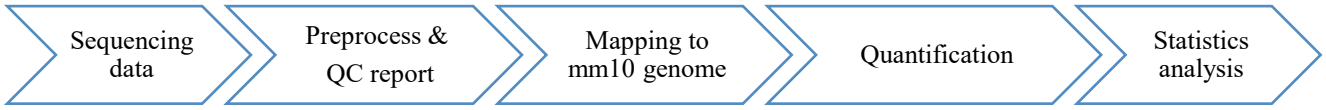


B. RNA library preparation

Library preparation was performed with the NEBNext ® Ultra ™ II DNA Library Prep Kit for Illumina ® (NEW ENGLAND BioLabs, Cat.: # E7645). In brief, cDNA was mechanically sheared followed by reactions of end-repairing, size-selection, 3'A-tailing and adaptor-ligation to generate indexed libraries. The library is then ready for amplification and size-selection by Purification Module with Agencourt AMPure XP Beads (Beckman coulter, Cat.: #A63881). The qualified libraries were analyzed by FRAGMENT ANALYZER™ Automated CE System (Advanced Analytical Technologies, Inc) and quantified by Qubit Fluorometer (Thermo Fisher). The libraries were sequenced with Illumina sequencing platform following the manufacturer's instruction. The experimental processes of library construction are as follows:



C.Bioinformatics analysis



This analysis pipeline is developed by CLC Genomics workbench v20.0.3 software package. First, All genes are extracted from the reference genome. Next, all annotated transcripts are extracted. If there are several annotated splice variants, they are all extracted. Then, the reads are mapped against all the transcripts plus the entire gene and optionally to the whole genome. From this mapping, the reads categorized and assigned to the genes, and expression values for each gene and each transcript are calculated.

The subsequent sequence reads were aligned onto the reference sequence of human genome (mouse, mm10) using the TopHat2 splice-junction mapper (Kim et al. Genome Biol. 2013) and calculated expression value (Fragments Per Kilobase of exon per Million fragments mapped, FPKM = total fragments / mapped reads (M) * exon length (KB)) of each gene at either gene or isoform level (Trapnell et al. Nat Biotechnol. 2010, Roberts et al. Genome Biol. 2011).

For differential expression analysis, we used Cuffdiff (Trapnell et al. Nat Biotechnol. 2013), an algorithm that robustly estimates expression at transcript-level resolution and controls for variability evident across replicate (if present) libraries to identify differential expressed gene (deg) between groups.

References

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008 Jul;5(7):621-8.

Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14, R36 (2013)

Trapnell, C. et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010 May;28(5):511-5.

Roberts, A. et al. Improving RNA-Seq expression estimates by correcting for fragment bias. Genome Biol. 2011;12(3):R22.

Trapnell, C. et al. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nat Biotechnol. 2013 Jan;31(1):46-53.

D. Other information for reference

Sequencing platform & reagent kit：Illumina NextSeq & high Output v2.5 kit (75 cycles)

Reference genome and Annotation：NCBI mouse mm10

定序與生資分析結果 (Sequencing results)

1. 定序樣本定量資料 (Sample QC table)

*QC check point：cDNA amount ≧10ng

樣本名稱	cDNA concentration (ng/μl) [amount]*	cDNA input for library construction (ng)
Sample 1	1.84 [23.92ng]	11.8
Sample 2	1.18 [15.34 ng]	8.4
Sample 3	1.67 [21.71 ng]	9.5
Sample 4	0.858 [11.154 ng]	6.45
Sample 5	1.6 [20.8 ng]	10.95
Vsample 1	1.14 [14.82 ng]	6.2
Vsample 2	1.12 [14.56 ng]	6.55
Vsample 3	1.48 [19.24 ng]	9.9
Vsample 4	1.93 [25.09 ng]	13.4
Vsample 5	1.73 [22.49 ng]	11.85

2. 定序總量與長度大小和比對結果

請參考在檔案夾"定序總量與比對結果統計"下的檔案 SAMPLE_01, SAMPLE_02, SAMPLE_03, SAMPLE_04, SAMPLE_05, Vsample_01, Vsample_02, Vsample_03, Vsample_04, Vsample_05，另外檔案 combined report 是整合所有的個別檔案的統計結果。

3. 樣本比對基因的表現結果

請參考在檔案夾” 樣本 mRNA 表現量”下的檔案 SAMPLE_01 ~SAMPLE_05 及 Vsample_01 ~ Vsample_05。檔名後標示 (GE)代表基因的表現量，標示(TE)代表同一個基因不同 transcript (isoform)的表現量。以下為檔案內參數的註解。

The Expression value parameter describes how expression per gene or transcript can be defined in different ways on both levels:

Total counts. When the reference is annotated with genes only, this value is the total number of reads mapped to the gene. For un-annotated references, this value is the total number of reads mapped to the reference sequence. For references annotated with transcripts and genes, the value reported for each gene is the number of reads that map to the exons of that gene. The value reported per transcript is the total number of reads mapped to the transcript.

Definition of RPKM RPKM, Reads Per Kilobase of exon model per Million mapped reads, is defined in this way [Mortazavi et al., 2008]:

$$RPKM = \frac{total\ exon\ reads}{mapped\ reads(millions) \times exon\ length\ (KB)}.$$

For prokaryotic genes and other non-exon based regions, the calculation is performed in this way:

$$RPKM = \frac{total\ gene\ reads}{mapped\ reads(millions) \times gene\ length\ (KB)}.$$

Total exon reads This value can be found in the column with header **Total exon reads** in the expression track. This is the number of reads that have been mapped to exons (either within an exon or at the exon junction). When the reference genome is annotated with gene and transcript annotations, the mRNA track defines the exons, and the total exon reads are the reads mapped to all transcripts for that gene. When only genes are used, each gene in the gene track is considered an exon. When an un-annotated sequence list is used, each sequence is considered an exon.

Exon length This is the number in the column with the header **Exon length** in the expression track, divided by 1000. This is calculated as the sum of the lengths of all exons (see definition of exon above). Each exon is included only once in this sum, even if it is present in more annotated transcripts for the gene. Partly overlapping exons will count with their full length, even though they share the same region.

Mapped reads The sum of all mapped reads as listed in the RNA-Seq analysis report. If paired reads were used in the mapping, mapped fragments are counted here instead of reads, unless the **Count paired reads as two** option was selected. For more information on how expression is calculated in this case, see section 30.2.4.

- **TPM (Transcripts per million)**. This is computed as $\frac{RPKM \cdot 10^6}{\sum RPKM}$, where the sum is over the RPKM values of all genes/transcripts (see http://bioinformatics.oxfordjournals.org/content/26/4/493.long).

- **Relative RPKM**. The RPKM for the transcript divided by the maximum of the RPKM values among all transcripts of the same gene. This value describes the relative expression of alternative transcripts for the gene.

4. 組間差異比較結果

請參考在檔案夾”樣本組間差異分析”下的檔案 SAMPLE vs Vsample GE expression: 比較基因在 SAMPLE 與 Vsample 組別間的差異。檔案 SAMPLE vs Vsample TE Expression 則顯示那些 transcript(isoform)在兩組間有顯著不同的表現情形。

5. 比較分析圖形

請參考在檔案夾” 樣本間分析圖形”下的檔案 Heat Map for RNA-Seq (GE)是顯示兩組間基因表現比對其 fold change >2 且 p value<0.001 的基因，Heat Map for RNA-Seq (TE)則顯示不同的 isoforms。檔案 PCA for RNA-Seq (GE)是利用基因的表現量來將樣本分群，PCA for RNA-Seq (TE)則是利用不同 isoform 的表現量來分群。檔案 VPlot SAMPLE vs. Vsample (GE)是畫出兩組間基因表現與 p value 的 volcano plot。VPlot SAMPLE vs. Vsample (TE)是畫出兩組間 isoform 表現與 p value 的 volcano plot。