Single Cell Transcriptomics Reveals Heterogeneity of Gene Expression in Mouse Cells

Nupur Banerjee, Tanmay Ghai, Dr. Rizi Ai and Dr. Wei Wang

Department of Chemistry and Biochemistry, University of California, San Diego, 9500 Gilman Drive, La Jolla, California

Introduction

What exactly is RNA-seq?

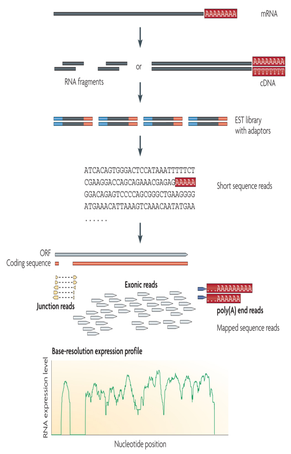
- RNA-seq: next-generation sequencing to extract information about the content of RNA sequences.
- 3 main steps: quantify the abundance of mRNA, determine the structure of genes (their different sized ends, and splicing patterns), and to quantify the varying expression levels of each transcript.
- A typical RNA-seq project involves
  - converting long strands of RNA into many short parts of cDNA (complimentary DNA).
  - sequencing adaptors attach to the each short cDNA fragment and subsequently, a sequence is produced with the use of high-throughput sequencing technology.
  - Sequences produced are then compared and aligned with a transcriptome (also known as a reference genome).
  - categorized aligned sequence into three groups: exonic reads, junction reads, and poly(A) end-reads. These different reads can then be used to view the basic expression profile of a gene.

Why we conduct RNA-seq and why is it important?

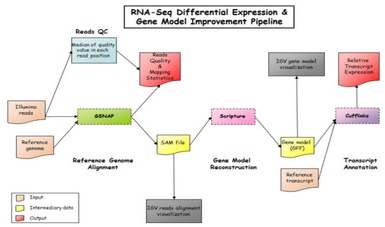
- Recently RNA-seq has revolutionized the world of transcriptomics. Using RNA-seq, we now have much more knowledge on differential expression of genes and their levels of expression, post-transcriptional mutations and even gene fusions.
- Furthermore, RNA-seq helps us understand what exactly a transcriptome is. A basic definition of a transcriptome is a complete set of all the transcripts in a cell and their quantity. However, RNA-seq takes this to a whole new level by allowing us to understand the nuances within a transcriptome.
- This is essential because we can then interpret the functional elements of a genome, reveal the molecular structure of cells and tissues, and understand disease, which could lead to potential cures or prevention information.

Why single-cell RNA-seq?

- Although derived from homogenous population, individual cells exhibit substantial information about gene expression, protein levels, and phenotypic output and functional consequences.
- This gives us valuable insight on how cellular heterogeneity within gene expression can lead to different traits and diseases such as cancer.

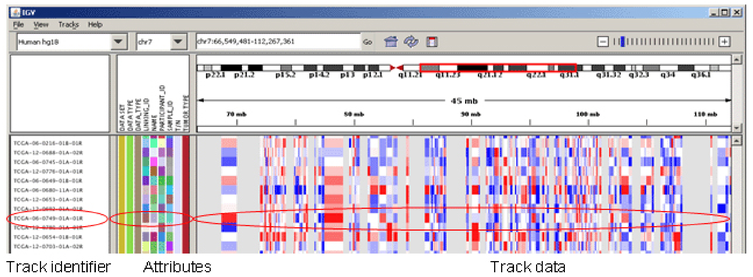


A typical RNA-seq experiment. First, long RNA strands are converted into cDNA fragments. Sequencing adaptors are added to each cDNA fragment, and the sequence of the cDNA is obtained. These reads are aligned with the reference genome, and are classified as exonic reads, junction reads, or poly(A) end-reads. Finally, these types are used to generate a base-resolution expression profile for each gene[2].



The following diagram on the left shows the relationship between GSNAP and Cufflinks, and how they end up producing short reads, which then leads to analyzing the gene expression levels of various cells.
Picture: http://gingerplum.files.wordpress.com/2011/08/8-20-20111.jpg?w=640&h=448

Visualization Through IGV Tools



The Integrative Genomics Viewer (IGV) is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets. Simply put, IGV tools is a tool that converts the reads produced by Cufflinks, in the RNA-seq process, into a visualization that helps us understand what these strands, sequences, and reads really mean.

Above is an example of a visualization of a read that has been tested with a human cell (not the same as our process, which involves mice cells). The image records each track identifier with the correlating attribute and the data that is produced with that read.

Methodology and Procedures

Software

1. *Gsnap* (version 2013-02-05) - GSNAP aligns the single and paired-end reads of the transcripts and produces short reads of RNA sequences.
Command - gsnap -t 4 -A sam -N 0 –D dir -d mm10 -s mm10.splicesites.iit --force-xs-dir --split-output=s *_1.fastq *_2.fastq
Input - .fastq files and Output - .sam files

2. *Samtools* (version 0.1.18) – SAM Tools provide various utilities for manipulating alignments in the SAM format, including sorting, merging, indexing and generating alignments in a per-position format.
Command - samtools-0.1.18/samtools view -Sb s.concordant_uniq > s.concordant_uniq.bam
Input - .sam file and Output - .bam file

3. *Cufflinks* (version 2.1.1) - Cufflinks assembles the transcripts and looks for differential expression and regulation of the genes in the RNA-seq samples. Also, it constructs sets of transcripts that give information about the reads that were observed in the experiment.
Command - cufflinks-2.1.1.Linux_x86_64/cufflinks -p 4 -G genes.gtf merge_concordant.bam
Input - .sam file and Output – transcripts.gtf, transcripts.expr, and genes.expr.

4. *Cuffdiffs* (version 2.1.1) - Cuffdiffs can be used to find significant changes in transcript expression, splicing, and promoter use.
Command – cuffdiff ../s1/genes.gtf ../s1/merge_concordant.bam ../s2/merge_concordant.bam
Input - .gtf file and .sam files and Output - .fpkm_tracking files among others

Single Cell Samples:

To conduct this RNA-Seq experiment, we used data from 18 bone-marrow dendritic cells (BMDCs) from a mouse (mus musculus).[1]

Gene Ontology:

From the data that we used, we selected ten genes to compare the expression levels between the 18 single BMDCs.

Results

Selected Genes with Different Expression Levels

Gene	Expression level (FPKM)	Function
Zf12	0	Gene not expressed
Ints8	7.10652e-317	Component of the Integrator complex, a complex involved in the small nuclear RNAs (snRNA) U1 and U2 transcription and in their 3'-box-dependent processing.
Foxp1	0.0682131	Expressed in developing lung, neural, intestinal and cardiovascular tissues.
Hook1	3.48482	Defects in Hook1 are the cause of the azh (abnormal spermatozoon head shape) mutant phenotype,
Timm22	10.531	Essential core component of the TIM22 complex, a complex that mediates the import and insertion of multi-pass transmembrane proteins into the mitochondrial inner membrane.
Phyh	16.3463	Defects in Phyh are the cause of lupus nephritis, a severe autoimmune disease.
Capza1	21.6231	F-actin-capping proteins bind in a Ca(2+)-independent manner to the fast growing ends of actin filaments thereby blocking the exchange of subunits at these ends.
Snrpe	52.9471	Belongs to the snRNP Sm proteins family.,
1600029D21Rik	1066.27	May participate in the wound response during the healing process, and promote wound repair
Lyz2	23255.5	Lysozymes have primarily a bacteriolytic function; those in tissues and body fluids are associated with the monocyte-macrophage system and enhance the activity of immunoagents.

The table above represents the gene's expressed throughout the RNA-seq process and the level at which they are expressed in the units FPKM, which stands for "for end pair sequencing." To the right, each gene's function is explained. This information can be found at http://david.abcc.ncifcrf.gov.

Mapping and Assembly

	Mapping Quality	Total Reads	Proper Paired	% Proper Paired	Unique Mapper	% Unique mapper	Genes	Isoforms
Cell_1	Very good!	9421315	7859185	83.4	7430034	78.9	3182	3446
Cell_2	Very good!	7813112	6312444	80.8	5881004	75.3	4751	5165
Cell_3	Very good!	7252243	5968877	82.3	5509608	76.0	4150	4517
Cell_4	Very good!	8762922	7243616	82.7	6827052	77.9	3337	3596
Cell_5	Very good!	9619357	7713570	80.2	7211808	75.0	3821	4154
Cell_6	Very good!	8035212	6537042	81.4	6197844	77.1	3514	3789
Cell_7	Very good!	9166757	7588491	82.8	7141259	77.9	3877	4215
Cell_8	Very good!	8119591	6598845	81.3	6196639	76.3	3716	4027
Cell_9	Very good!	7946851	6526587	82.1	6142598	77.3	4019	4371
Cell_10	Very good!	9672123	7979027	82.5	7552492	78.1	3204	3429
Cell_11	Very good!	6814517	5653627	83.0	5403171	79.3	2737	2896
Cell_12	Very good!	9024335	7363060	81.6	6917248	76.7	3239	3498
Cell_13	Very good!	8224518	6855884	83.4	6472480	78.7	4488	4867
Cell 14	Very good!	8168995	6761990	82.8	6395450	78.3	3215	3475
Cell 15	Very good!	8266431	6838516	82.7	6522522	78.9	2369	2519
Cell 16	Very good!	8525597	7079661	83.0	6586434	77.3	4576	4955
Cell_17	Very good!	8477944	6883249	81.2	6421991	75.7	4012	4375
Cell 18	Good	7065766	5565973	78.8	5239209	74.1	5069	5506

Differentially Expressed Genes

We used Cuffdiff to compare the gene expression levels from two different cells, in order to verify the accuracy of our results. We found that **982** genes were **differentially expressed** between cells 1 and 2, while **22,378** genes were **not differentially expressed** between the two. Since the large majority of the expression levels were in a similar range, and there was no significant difference between the two cells, we concluded that our results were precise.

Conclusions and Future Work

- RNA-seq has now become a vital part of bioinformatics and will continue to help us gain new knowledge and insight about genes, their expression levels, and what these expression levels mean.
- In this study, 18 mouse single-cell RNA-seq were analyzed, including mapping, assembly, quantification and functional analysis
- Differential expressed genes were compared between cell 1 and cell2
- In the future, the functions of 982 differentially expressed genes will be further analyzed to determine the heterogeneity between cells. With this information, we can identify what certain genes lead to various diseases and with this knowledge we can further learn how to prevent those genes from being expressed

References

1. Shalek AK, Satija R, Adiconis X, Gertner RS, Gaublomme JT, Raychowdhury R, Schwartz S, Yosef N, Malboeuf C, Lu D, Trombetta JJ, Gennert D, Gnirke A, Goren A, Hacohen N, Levin JZ, Park H, Regev A. Single-cell transcriptomics reveals bimodality in expression and splicing in immune cells. Nature. 2013 Jun 13;498(7453):236-40
2. Wang Z, Gerstein M, Snyder M. Nat Rev Genet. RNA-Seq: a revolutionary tool for transcriptomics. 2009 Jan;10(1):57-63.
3. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L. Nat Protoc. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. 2012 Mar 1;7(3):562-78.