Feiqiao Brian Yu
2012.03.16

Comparison of Microarray and RNA-seq Analysis Methods for Single Cell Transcriptomics

Introduction

Behavior of single cells can be explained through changes in the transcription level of the genome followed by translation of the resulting mRNA into proteins (1). Changes in gene expression levels of each cell, in turn, are controlled by sensory networks that respond to the external environment. Even though within an organism or tissue all cells have the same genome, diverse phenotypes exist because of varying type and amounts of mRNA transcripts referred to as the transcriptome of the cell (1-3). Because of the existence of such heterogeneity within an isogenic cell population, transcriptome analysis at single cell resolution becomes crucial to identify cell to cell spatiotemporal variations that are often masked by ensemble averages (3, 4).

One example that illustrates the distinction between single cell and bulk measurement is performed by Mathies group which looked at GAPDH expression in individual Jurkat cells (Figure 1). Their data shows that the average expression level of 50 cells is not representative of any individual cell (5). Another example deals with the study of NF-kB expression in mouse fibroblast cells in response to TNF-α using high throughput microfluidics. In contrast to population level studies showing gradual analog changes in expression levels, results show digital activation at the single cell level (6). These experiments show the importance of measuring gene expression at a single cell level.

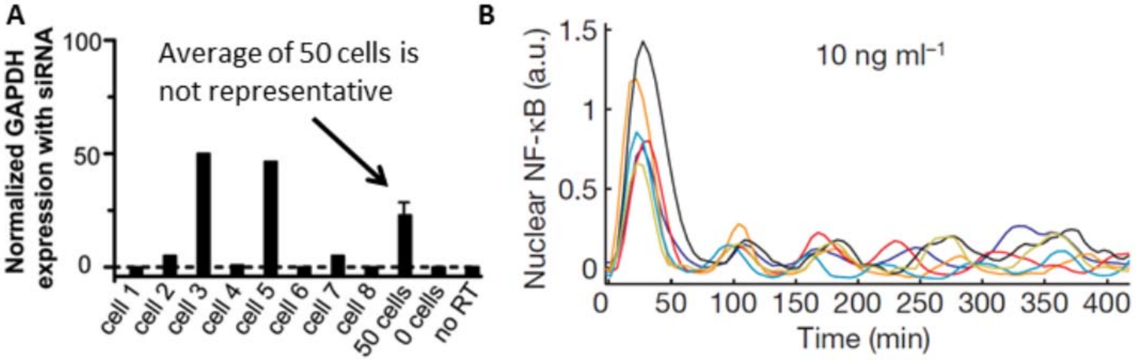


Figure 1: Single cell studies show ensemble average misses individual variations within a population. (A) Expression patters of GAPDH in 50 individual Jurkat cells show that population average is not representative of any individual (5). (B) Contrary to previous studies, NF-kB expression in mouse fibroblast cells occur digitally in response to 10 ng ml^{-1} TNF-α (6).

Currently, gene expression microarrays and RNA-seq are two popular ways of extracting single cell transcriptome data. Both methods allow high throughput analysis of many cells and gene targets. Developed in the 1990s, high density microarrays are more mature than deep sequencing technologies. With a good understanding of the technology biases and costing less per experiment, microarrays can differentiate cell and tissue types and show how expression changes across development, diseases, within and among species (7, 8). However, microarray suffers from background hybridization, limited accuracy of expression for transcripts in low abundance, and cannot be used to detect splice variants or unknown genes (1). RNA-seq, on the other hand, provides direct access to the sequence without a reference genome or pre-designed probes. It provides a larger dynamic range (five-log) and can be used to detect splice variants, isoforms, and new genes. However, RNA-seq experiments are typically high cost, and sources of bias such as coverage and heterogeneity are not well studied (8). RNA-seq experiments often generate orders of magnitude more data than microarrays, requiring much more complex informatics to extract meaningful results.

Microarray Technology

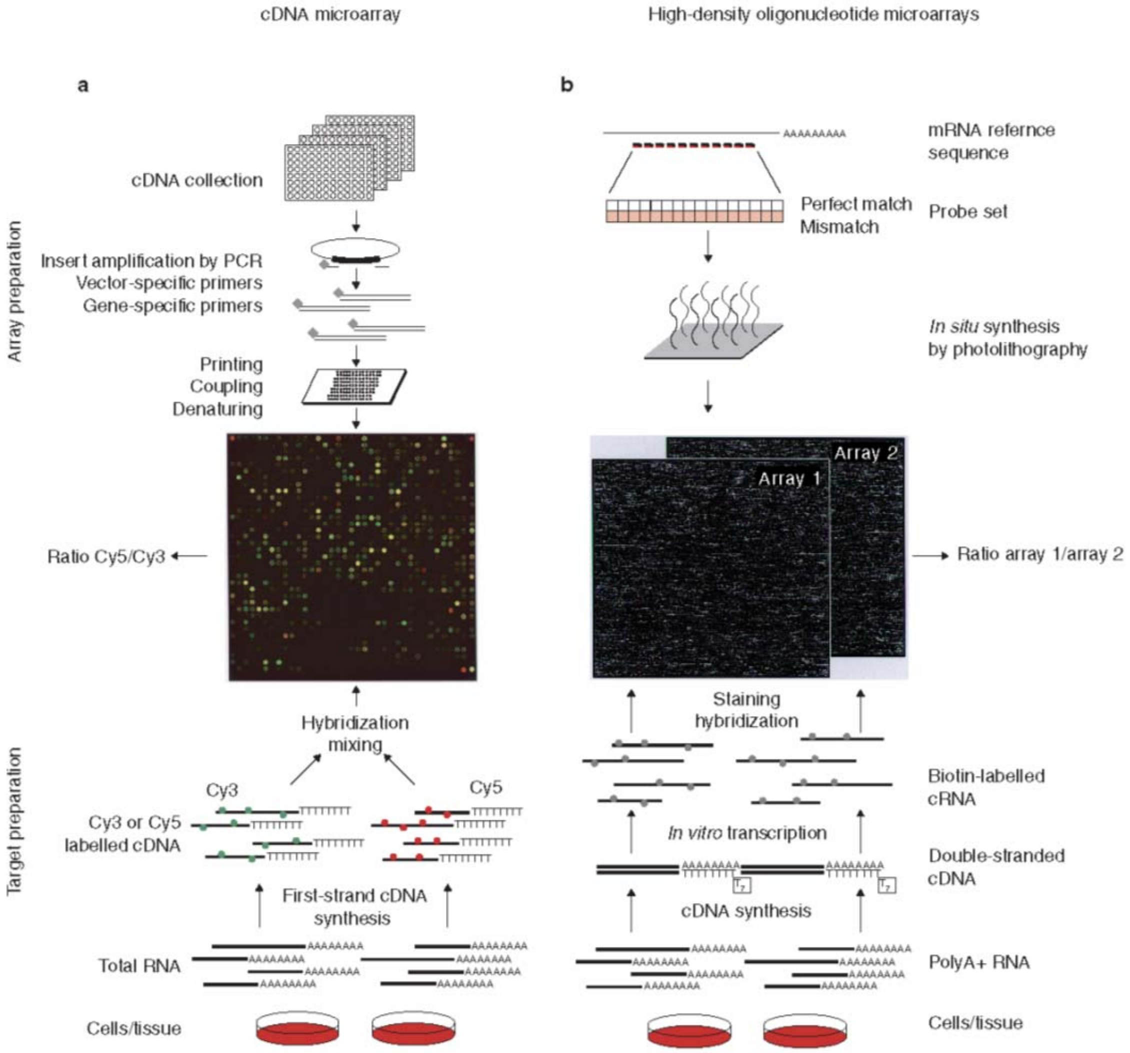


Figure 2: Schematic overview of probe and target preparation for microarray experiments using (A) cDNA and (B) high density oligonucleotides

Microarray experiments include two components: probe and target. Probe could be prepared using cDNA or pre-synthesized oligonucleotides. Probes are either attached to silicon wafers using photolithography (Affymetrix) or printed onto a glass slide (Agilent Technologies). During target preparation, total RNA is usually extracted from the cell sample followed by poly A selection (7, 9). The selected mRNA is then converted to single strand cDNA fragments in the presence of fluorescently labeled nucleotides. The target cDNA is hybridized to immobilized probes on the microarray surface. Finally, unbound material is washed off and data is collected as fluorescent images whose intensity represents mRNA abundance (10). Figure 2 illustrates typical process flows for microarray experiments using cDNA library and high-density oligonucleotides.

Microarray Analysis Methods

The true power of microarray becomes apparent when probing global gene expression patterns. Because such experiments generate large amounts of data, systematic methods are required to organize and extract meaningful expression relations. Since microarray technology is relatively mature, many commercially available packages exist including GeneCluster, Expression Profiler, XCluster, and Cleaver (10, 11). These packages employ a variety of algorithms to perform 3 main functions: normalization, grouping, and feature reduction.

Normalization

Normalization is a technique that removes systematic variations in microarray data (ie. Intensities of fluorescent labels in the final images). Many sources of variation exist including differences in labeling efficiency between different dyes, differences in the power of lasers used to image the microarray, differences in hybridization efficiencies, and spatial biases across the microarray surface (12). In terms of single cell experiments, we are often interested in small differences in expressions patterns among subpopulations (spatial) or same cells at different time points (temporal). In order to distinguish small differential expressions, removing global bias becomes important.

One commonly used normalization method involves preselecting housekeeping genes which are assumed to be constantly expressed under testing conditions. The expression level of these genes is used to generate a normalization factor which makes the geometric mean of this set of genes 1. An alternative method that uses all samples to generate this normalization factor is used when none of the genes in the experiment can be considered “housekeeping”. When applied, this method makes the mean log ratio of all the data equal to zero. In other words, this method shifts the distribution so that it is centered on 0. Intensity dependent normalization is another method that compensates for spot intensities using data from dye swap experiments. However, when there is a spatial bias across microarray surfaces. It is more advantageous to use a different normalization factor on each section of the microarray (12).

Grouping

Normalization techniques remove global bias. The next task is to reveal genes that are co-regulated in different cells or are expressed in different amounts at different time points. This is done through grouping. Grouping methods help data visualization and can be broadly divided into two sets: supervised and unsupervised.

Unsupervised group is also known as *clustering*. These algorithms simplify large gene expression data sets by collecting similar profiles without prior knowledge of the data. Similarity among gene expression profiles is calculated based on distance metrics such as statistical correlation coefficient or Euclidean distance. Most common clustering strategies used include hierarchical clustering, self-organizing maps, and k-means clustering (11).

Hierarchical clustering works similar to the distance method of generating phylogenetic trees. It is an iterative algorithm that

1. Assigns each item to its own cluster.
2. Identifies the closest pair of clusters, joins them together and consequently reducing the total number of clusters by one.
3. Computes distances between the new cluster to all the old clusters.
4. Repeat steps 2 and 3 until there is a single cluster left (13).

Heirarchical clustering can be implemented easily, but is suffers from repeatability as the creation of branching point is often an arbitrary decision.

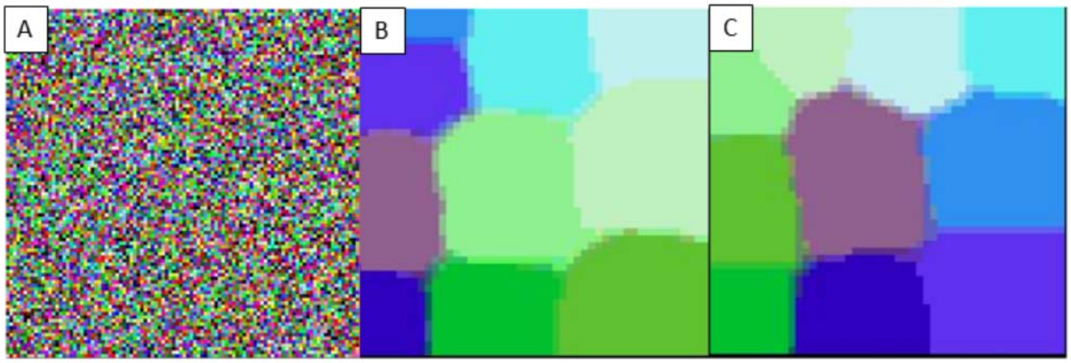


Figure 3: Self-organizing maps. (A) Initial randomized data representing microarray data (B, C) processed data after applying SOM algorithm. Notice that SOM does not always guarantee a unique solution, but it can successfully assign similarities to members inside each group. Reference: http://davis.wpi.edu/~matt/courses/soms/

Self-organizing maps employ a different iterative algorithm where

1. Initialization is performed by generating a random set of weights for each data.
2. Randomly selecting a sample, the algorithm determines the best matched sample in the entire data space based on some distance metric.
3. The weight of this closest sample and its “neighbors” are rewarded by becoming more like the sample vector through scaling.
4. This process is repeated for many randomly selected samples until the weights do not move anymore.

An example shows the results of self-organizing maps on a set of colors, it can be seen that a data set represented by randomly distributed colors can be organized into groups of similar colors (Figure 3). However, the order of these colors is not always deterministic. Therefore, interpreting the meaning and relatedness of clusters could sometimes be a problem.

K-means clustering, on the other hand, involves a predetermined number of desired clusters k.

1. The method first randomly places the center of the k clusters.
2. Then it calculates the distance from each sample to the closest cluster center.
3. Using the distances within a cluster, calculate a new center for the cluster.
4. Repeat the above 2 steps until cluster centers converge.

This algorithm is useful when the expected k is known. However, that is not always the case. In addition, it can sometimes be stuck at local optima. Therefore, running the algorithm multiple times is a requirement.

Unsupervised grouping can find new expression profiles but are not always designed to reliably reproduce groupings. Supervised grouping methods are usually termed *classification* and are extremely well suited for separating a collection of samples into known groups using previous expression knowledge. Classification algorithms are usually based on machine learning techniques such as regression, neural networks, and linear discriminant analysis. All of these methods require a set of known samples and their corresponding expression patterns to “learn”. Regression estimates a predictor function based on a linear log-likelihood model. Neural network creates a multi-layered computational network based on the training samples. It then uses that model to predict categories for each unknown cases. Finally, linear discriminant analysis estimates a probability distribution function for the genes and samples in the training set. Then, given a new sample, it tries to find the closest distribution and assigns the sample to that set.

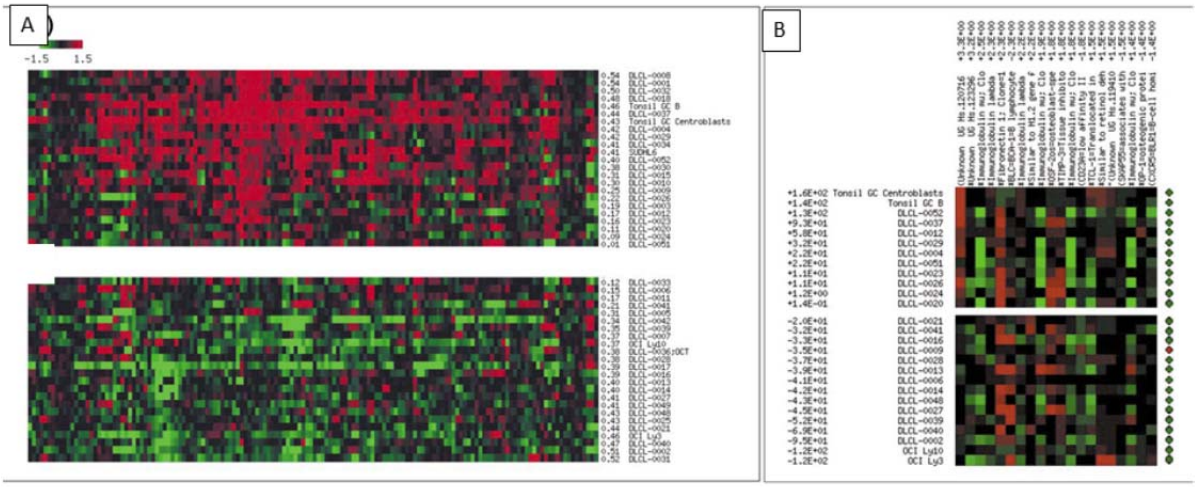


Figure 4: Microarray data on two different types of lymphoma expression patterns grouped using (A) unsupervised and (B) supervised methods. (A) Expressions are successfully clustered into two sets representing activated and germinal center subtypes. (B) Using LDA and a training set of both subtypes, all but one sample was classified incorrectly (marked in red) (11).

As a demonstration of grouping algorithms for single cell application, lymphoma samples using 148 genes were clustered using k-mean clustering. The algorithm successfully separated the samples into 2 groups, representing the activated and germinal center subtypes. The separation results are 98% consistent with a correct reference data (Figure 4). Similarly, in the same review, supervised grouping methods are also applied to the same lymphoma set. A subset (10 samples) of the data is correctly identified and used to train the algorithm while the rest of the data becomes the test set. The result shows that only one sample was classified incorrectly (11) (Figure 4).

Feature Reduction

Sometime during microarray analysis it is beneficial to reduce the dimensionality of the data in order to identify salient features. Similar to grouping, feature reduction methods can also be classified into supervised and unsupervised (11).

Supervised feature reduction is called *feature selection* and has to do with selecting the important expression profiles in a data set while removing the less important profiles. This dimensional reduction helps to keep the analysis more focused. One way of performing supervised feature reduction is to iteratively perform supervised grouping and then removing the lease important expression measurement from all expression profiles.

On the other hand, unsupervised feature reduction (*data pruning*) employs algorithms such as Singular Value Decomposition (SVD) and Independent Component Analysis (ICA). SVD (or sometimes known as Principal Component Analysis PCA) attempts to find a set of orthogonal eigenvectors of the expression data that essentially represent gene expression profiles of interest. SVD relies heavily on the theorem from linear algebra that any M by N matrix A (M > N) can be written as the product of a column orthogonal matrix U (M by N), a diagonal matrix W (N by N), and the transpose of an orthogonal matrix V (N by N) (12, 14).

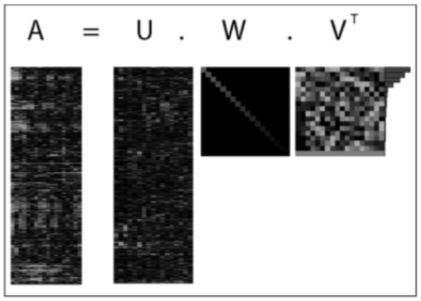


Figure 5: Singular value decomposition. Matrix A is the input expression data. The diagonal matrix W contains eigenvalues that assign importance of each vectors in V to the changes observed in A. Matrix V^T contains the eigenvectors, and U contains the coefficients for the genes in those vectors. The horizontal bars on the right indicate how much information each eigenvector captures (12).

In microarray data, M often represents the number of genes and N the number of samples (experiments). Then, U contains coefficients for each eigenvector, which indicates the amount of information contributed by each gene's expression vector to the final data matrix A. Furthermore, since the diagonal entries of W contain decreasing weights, the corresponding rows of V^T are also arranged in the order of descending importance to expression variations in A (Figure 5). Therefore, SVD is a powerful method to extract the most important gene expression vectors that contribute to the expression variations in the final data.

Independent component analysis (ICA) is another method to extract biological significant dimensions from microarray data. Compared to SVD, ICA assumes non-Gaussian expression variations and models the microarray observations as a linear combination of its components, which are chosen to be as independent as possible. Because it involves higher order statistics, ICA does not suffer as much from noise and artifacts introduced by the fact that gene expression data usually do not have Gaussian distributions (14). The fundamental equation that ICA tries to solve is $A = M \times S$, where A is still the microarray data, S is a matrix containing independent components, and M is a mixing matrix. The goal of the algorithm is to find another matrix W, called the unmixing matrix, such that we can recover the independent components $S = W \times A$ that contain information about the de-convolved gene expression profiles (15, 16). Table 1 summarizes the various microarray related informatics methods.

Table 1: Summary of Techniques Used in Microarray Data Analysis

	Supervised	Unsupervised
Normalization	• Preselecting housekeeping genes • Global mean normalization • Intensity dependent normalization	
Grouping	• Regression • Neural networks • Linear discriminant analysis	• Hierarchical clustering • Self-organizing maps • k-means clustering
Feature Reduction	• Iteratively remove least salient groups	• Singular value decomposition • Independent component analysis

RNA-seq Technology

Compared to microarray, RNA-seq based on deep sequencing technology is a newer method to interrogate single cell transcriptome. Because it allows direct access to sequences of mRNA, bias and variation due to hybridization and labeling efficiencies are avoided. In addition, RNA-seq allows the detection of RNA editing events, such as alternative splicing, the most important source of phenotypic diversity in eukaryotes (7). In an RNA-seq experiment, total RNA is first extracted from cell samples. Since rRNA represents the vast majority of total cellular RNA, to maximize diversity of the sequences retrieved from sequencing, it is of interest to reduce the quantity of rRNA present in the sample either by enrichment of polyadenylated RNA or by deletion of rRNA quantity (17). The enriched RNA sample is treated with DNase to remove any remaining DNA followed by cDNA synthesis. Depending on the sequencing platform and analysis techniques use, cDNA is labeled using a variety of library creation procedure before loading (17, 18). Currently, there are many sequencing platforms available for single cell transcriptome sequencing including Illumina, Complete Genomics, Pacific Biosciences, and Helicose. Lam et al provides a performance comparison using two of the platforms (19).

RNA-seq Analysis Methods

In addition to its cost, RNA-seq also generates order of magnitude more data per experiment. Therefore, computational power and data analysis techniques become crucial. Many packages such as TopHat, Cufflinks, and Scripture exist that perform annotation and quantification of transcriptome. These computational methods generally fall into 3 categories: sequence mapping, transcriptome reconstruction, and expression quantification (20). Figure 7 shows a table of computational packages and their respective applications.

Sequence Mapping

During sequence mapping, the short length of the RNS-seq reads and high error rates are often challenges for alignment methods. Two major classes of algorithms used to perform sequence mapping are unspliced aligners, which do not allow gaps, and spliced aligners that do allow gaps. Unspliced aligners are ideal for mapping reads against a reference cDNA database for quantification. These approaches sometimes use a ‘seed method’, where matches are found for small seed sequences that are assumed to match the reference (21). Alternatively, ‘Burrows-Wheeler transform’ can be used which manipulates the data structure for fast searching of perfect matches (22).

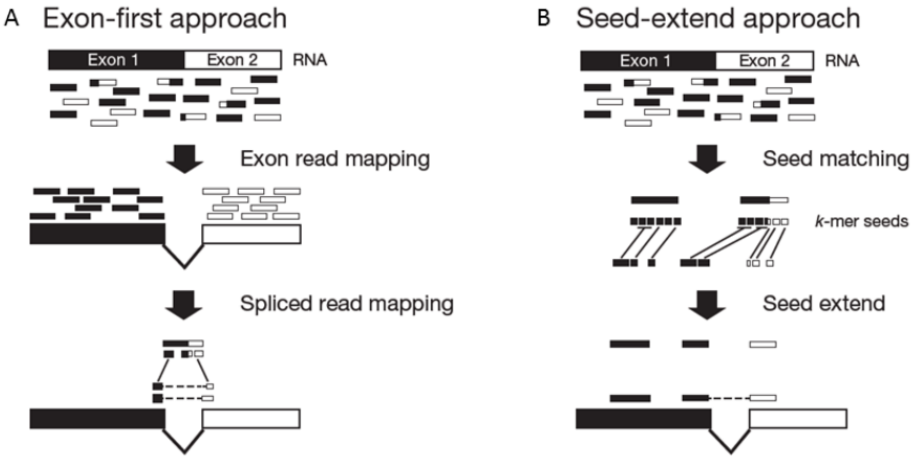


Figure 6: Schematic overview of two common spliced read aligner methods. (A) Exon-first (B) Seed-Extend (20)

When only a distant reference cDNA database is available, unspliced alignment algorithms can be employed. These include the ‘exon first’ or ‘seed first’ approaches (Figure 6). These methods continuously map reads to the reference using unspliced methods. Those sequences that do not match are further broken down into shorter segments and aligned again. Finally, regions around the mapped reads are searched for possible spliced events using more sensitive methods such as Smith-Waterman. Exon-first methods are faster and require less computational power. However, they can miss spliced alignments for reads that also maps to the genomic region. Seed-first methods evaluate spliced and unspliced events together to reduce bias (20).

Transcriptome Reconstruction

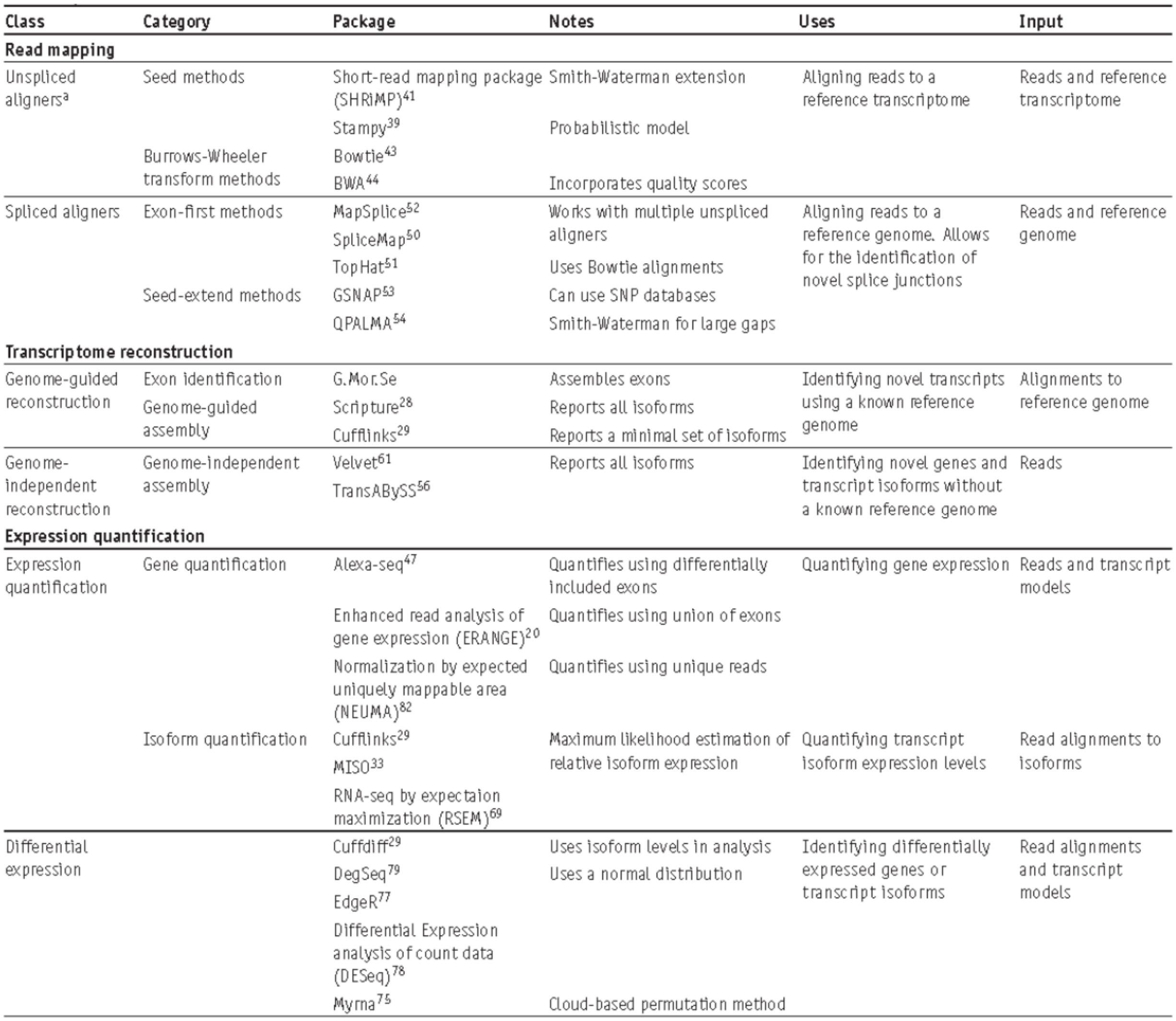
Transcriptome reconstruction involves defining a precise map of all the transcripts and isoforms expressed in a particular sample. This process may be challenging, again, for 3 reasons. The dynamic range of gene expression is high, the reads could be generated by mature as well as precursor RNA, and the reads are short with possibly many isoforms. To combat these difficulties, two approaches have been developed referred to as the genome-guided reconstruction and genome-independent reconstruction.

Genome-guided reconstruction relies on a reference genome to which all reads are mapped. For short reads, exon identification methods can be used to define boundaries and establish connections between exons. For longer reads, methods such as Cufflinks and Scripture use genome-guided assembly methods to directly reconstruct the transcriptome from spliced reads (23, 24). In genome-independent reconstruction, consensus transcripts are first built from the reads and then mapped to a genome for annotation (25). A popular algorithm is the de Bruijn graph, which uses k-mers to reduce the complexity with handling millions of read. When comparing the 2 methods of reconstruction, the best method depends on the particular application. If a reference sequence is not present, genome-independent methods are the obvious choices. However, genome-guided methods can provide higher sensitivity for better annotation.

Estimating Expression Levels

The type of RNA-seq computational analysis that is most relevant to data from different time points of single cell samples has to do to estimating differential expression levels. However, in order to quantitatively estimate gene expression, read counts must be normalized. Instead of fluorescent intensity or labeling efficiency which cause bias in microarray experiments, sources of read variation for RNA-seq experiments result from two factors: RNA fragmentation during library construction causes longer transcripts to generate more reads and the variability in the number of read produced from each run causes fluctuations in the number of fragments mapped across samples (8).

To normalize read counts, one can use the reads per kilobase of transcript per million mapped reads (RPKM) metric which normalizes transcript's read count by both length and total number of mapped read. The fragments per kilobase of transcript per million mapped reads (FPKM) metric, on the other hand, is the analogous method to normalize paired-end reads. In addition to normalizing read counts, an understanding of expression variations across conditions is required to correctly identify differential expression (20). As a result, many statistically based methods have been used to assign significance to expression levels.



Class	Category	Package	Notes	Uses	Input
Read mapping					
Unspliced aligners^{a}	Seed methods	Short-read mapping package (SHRiMP)^{41}	Smith-Waterman extension	Aligning reads to a reference transcriptome	Reads and reference transcriptome
		Stampy^{39}	Probabilistic model		
	Burrows-Wheeler transform methods	Bowtie^{43}			
		BWA^{44}	Incorporates quality scores		
Spliced aligners	Exon-first methods	MapSplice^{52}	Works with multiple unspliced aligners	Aligning reads to a reference genome. Allows for the identification of novel splice junctions	Reads and reference genome
		SpliceMap^{50}			
		TopHat^{51}	Uses Bowtie alignments		
	Seed-extend methods	GSNAP^{53}	Can use SNP databases		
		QPALMA^{54}	Smith-Waterman for large gaps		
Transcriptome reconstruction					
Genome-guided reconstruction	Exon identification	G.Mor.Se	Assembles exons	Identifying novel transcripts using a known reference genome	Alignments to reference genome
	Genome-guided assembly	Scripture^{28}	Reports all isoforms		
		Cufflinks^{29}	Reports a minimal set of isoforms		
Genome-independent reconstruction	Genome-independent assembly	Velvet^{61}	Reports all isoforms	Identifying novel genes and transcript isoforms without a known reference genome	Reads
		TransABySS^{56}			
Expression quantification					
Expression quantification	Gene quantification	Alexa-seq^{47}	Quantifies using differentially included exons	Quantifying gene expression	Reads and transcript models
		Enhanced read analysis of gene expression (ERANGE)^{20}	Quantifies using union of exons		
		Normalization by expected uniquely mappable area (NEUMA)^{82}	Quantifies using unique reads		
	Isoform quantification	Cufflinks^{29}	Maximum likelihood estimation of relative isoform expression	Quantifying transcript isoform expression levels	Read alignments to isoforms
		MISO^{33}			
		RNA-seq by expectaion maximization (RSEM)^{69}			
Differential expression		Cuffdiff^{29}	Uses isoform levels in analysis	Identifying differentially expressed genes or transcript isoforms	Read alignments and transcript models
		DegSeq^{79}	Uses a normal distribution		
		EdgeR^{77}			
		Differential Expression analysis of count data (DESeq)^{78}			
		Myrna^{75}	Cloud-based permutation method		

Figure 7: Comprehensive summary of a list of RNA-seq analysis programs used for 1 Read mapping 2 Transcriptome reconstruction and 3 Expression quantification (20)

Comparison of Microarray and RNA-seq Data

Several groups have compared results from single cell transcriptome experiments using both microarray and RNA-seq data. Through these experiments, one can develop an understanding of the pros and cons of each technique and their associated informatics requirements. This part of the review will look at two specific studies that assess single cell expression results from microarray and RNA-seq on the same samples.

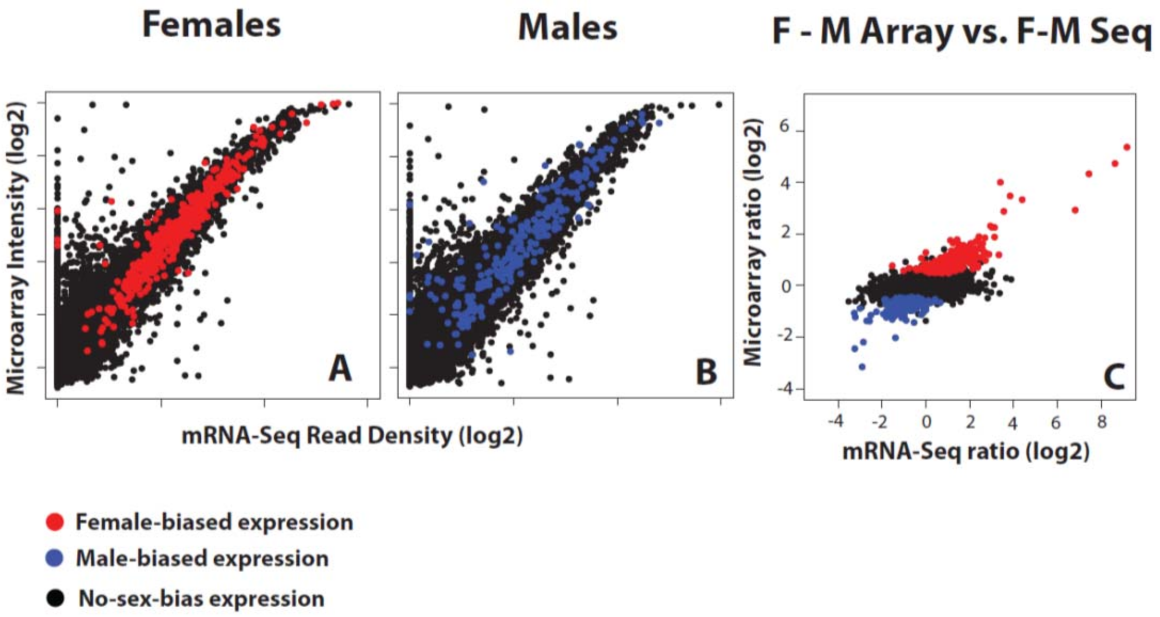


Figure 8: Comparison of microarray and RNA-seq expression data on cells from male and female *Drosophila* head. (A, B) Both male and female expression results across platforms show agreement until low expressed genes. (C) Reasonable measurement congruency for the entire data set but high in high in fold change measurements (7).

Malone et al. looked at gene expression differences in *Drosophila* head cells between male and female (7). Using Illumina Genome Analyzer and Affymetrix microarray, the group saw strong congruence in relative array intensities compared to RNA-seq read counts (Figure 8). The group used biological replicates and moderate t-test to detect differentially expressed genes between female and males. Results show agreement between microarrays and RNA-seq for sex-biased expression. However, at low expression values, both platforms suffer from background noise causing scattering.

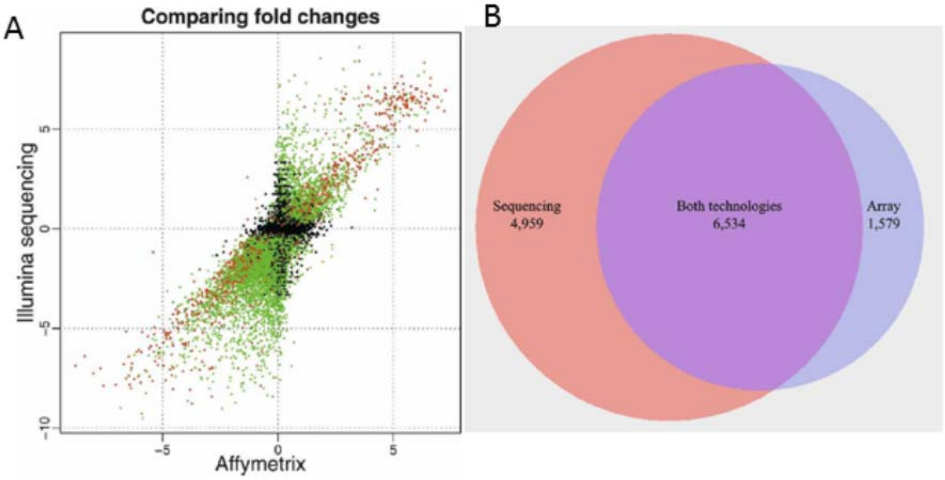


Figure 9: Cross platform study on differential expression between liver and kidney cells. (A) Comparison of estimated log2 fold changes from microarray and sequencing data show strongest correlation when genes are mapped to by many read. Correlation is weaker for genes mapped to by fewer reads. (B) Venn diagram summarizing the overlap between genes called as differentially expressed (8).

Another experiment that compares Affymetrix microarray and Illumina RNA-seq is performed by Marioni et al. to compare liver and kidney RNA expression profiles (8). They used a set of 17,708 array probes from annotated genes in the Ensembl database. For RNA-seq, 7 biological replicates are used. Data processing is done using ELAND. Out of all reads, 40% mapped uniquely to genomic locations, and 65% out of those mapped to autosomal or sex chromosomes. Comparing differentially expressed genes across technologies, they identified 6534 differentially expressed genes called by both platforms (Figure 9). Sequencing data seems to call a larger number of differentially expressed genes.

Conclusion

With improvement in technologies and analysis algorithms, microarray and RNA-seq combined holds great promises to reveal deeper insights into the fundamentals of gene expression variations within and among single cells. Next generation sequencing technologies that offer higher read and throughput have made single cell transcriptomics applicable to studying subpopulations of tumor, differentiated embryonic stem cells, bacteria in a biofilm, and a plethora of other heterogeneous subpopulations within a community. Yet, most single cell RNA experiments are still limited to exploring spatial differences in gene expression. The true power of transcription analysis will come when we can track cell lineage and combine spatial with temporal expression patterns to elucidate spatiotemporal genetic regulatory networks of life.

References

1. Tang F, Lao K, & Surani MA (2011) Development and applications of single-cell transcriptome analysis. *Nature methods* 8(4 Suppl):S6-11.

2. Kalisky T, Blainey P, & Quake SR (2011) Genomic analysis at the single-cell level. *Annual review of genetics* 45:431-445.

3. Wang D & Bodovitz S (2010) Single cell analysis: the new frontier in 'omics'. *Trends in biotechnology* 28(6):281-290.

4. Kalisky T & Quake SR (2011) Single-cell genomics. *Nature methods* 8(4):311-314.

5. Toriello NM*, et al.* (2008) Integrated microfluidic bioprocessor for single-cell gene expression analysis. *Proceedings of the National Academy of Sciences of the United States of America* 105(51):20173-20178.

6. Tay S*, et al.* (2010) Single-cell NF-kappaB dynamics reveal digital activation and analogue information processing. *Nature* 466(7303):267-271.

7. Malone JH & Oliver B (2011) Microarrays, deep sequencing and the true measure of the transcriptome. *BMC biology* 9:34.

8. Marioni JC, Mason CE, Mane SM, Stephens M, & Gilad Y (2008) RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res* 18(9):1509-1517.

9. Kawasaki ES (2004) Microarrays and the gene expression profile of a single cell. *Annals of the New York Academy of Sciences* 1020:92-100.

10. Schulze A & Downward J (2001) Navigating gene expression using microarrays--a technology review. *Nature cell biology* 3(8):E190-195.

11. Raychaudhuri S, Sutphin PD, Chang JT, & Altman RB (2001) Basic microarray analysis: grouping and feature reduction. *Trends in biotechnology* 19(5):189-193.

12. Sherlock G (2001) Analysis of large-scale gene expression data. *Briefings in bioinformatics* 2(4):350-362.

13. Johnson SC (1967) Hierarchical Clustering Schemes. *Psychometrika* 32(3):241-254.

14. Yao F, Coquery J, & Le Cao KA (2012) Independent Principal Component Analysis for biologically meaningful dimension reduction of large biological data sets. *BMC bioinformatics* 13(1):24.

15. Kong W, Vanderburg CR, Gunshin H, Rogers JT, & Huang X (2008) A review of independent component analysis application to microarray gene expression data. *BioTechniques* 45(5):501-520.

16. Hyvarinen A & Oja E (2000) Independent component analysis: algorithms and applications. *Neural networks : the official journal of the International Neural Network Society* 13(4-5):411-430.

17. Wilhelm BT, Marguerat S, Goodhead I, & Bahler J (2010) Defining transcribed regions using RNA-seq. *Nature protocols* 5(2):255-266.

18. Tang F*, et al.* (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nature methods* 6(5):377-382.

19. Lam HY*, et al.* (2012) Performance comparison of whole-genome sequencing platforms. *Nature biotechnology* 30(1):78-82.

20. Garber M, Grabherr MG, Guttman M, & Trapnell C (2011) Computational methods for transcriptome annotation and quantification using RNA-seq. *Nature methods* 8(6):469-477.

21. Li H, Ruan J, & Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 18(11):1851-1858.

22. Langmead B, Trapnell C, Pop M, & Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology* 10(3).

23. Guttman M*, et al.* (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs (vol 28, pg 503, 2010). *Nature biotechnology* 28(7):756-756.

24. Trapnell C*, et al.* (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature biotechnology* 28(5):511-515.

25. Robertson G*, et al.* (2010) De novo assembly and analysis of RNA-seq data. *Nature methods* 7(11):909-U962.