Computational Analysis of RNA-Seq Data: From Quantification to High-Dimensional Analysis

Junhyong Kim, PhD

Department of Biology and Penn Genome Frontiers Institute
University of Pennsylvania
Philadelphia, Pennsylvania

NOTES

Introduction

High-throughput RNA sequencing is providing unparalleled resolution of the transcriptome and has been especially instrumental in revealing the transcriptome's sequence-level complexity (Core et al., 2008; Morin et al., 2008; Trapnell et al., 2009, 2010; Wang et al., 2009; Guttman et al., 2010; Nechaev et al., 2010; Pickrell et al., 2010; Marquez et al., 2012). In this chapter, I will discuss some of the computational and statistical challenges of quantifying the transcriptome from high-throughput RNA sequence data. I will also set forth the principles of high-dimensional data analysis using quantified transcriptomes. RNA sequencing, quantification, and data analysis share many of the same problems and solutions with microarray-based assays. Therefore, I will concentrate on issues more specific to RNA sequencing—especially RNA sequencing from single cells. I will assume familiarity with the overall experimental scheme for massively parallel short-sequence reads provided by instruments such as Illumina HiSeq (Illumina, San Diego, CA) and ABI SOLiD platforms (Applied Biosystems, Carlsbad, CA). It should be noted that many of the specific experimental, statistical, and computational problems are still being actively addressed in the field, so best practices for using massive RNA sequencing data for functional genomics are expected to continue evolving.

Transcript Quantification from RNA Sequencing Reads

Aligning the reads to a reference genome

When processing short reads from RNA sequencing, the key computational step consists of aligning the reads to a reference genome. If the sequence reads are exact copies of contiguous regions of the reference genome, this step is straightforward. However, the sequence reads may differ from the genomic sequence owing to several factors: errors in the sequence chemistry, artifacts created by the library construction step (e.g., concatemers or fusion of separate molecules), or biological RNA processing such as splicing and RNA editing. A more important source of variation is found in the polymorphism of the reference genome, which is likely to contain indels and single nucleotide polymorphisms (SNPs) not present in the sequenced strain. Therefore, the computational alignment of the reads to the genome must take into account such possible variations.

Algorithmic procedures and alignment strategies

The standard algorithmic procedure for dealing with such variations involves finding the best local alignment for subsequences of the reads to the subsequences to the genome, and assembling the matches while respecting the positional constraints. The simplest reasonable algorithm for this procedure goes through a number of steps that are proportional to the product of the read length and the reference genome length. However, while such a computation is feasible for any single read, it becomes computationally impossible when multiplied for tens or hundreds of millions of sequence reads. Therefore, available algorithms try to approximate the best solutions within a reasonable computing time. The main strategies involve indexing the reference genome or the read set with various kinds of k-mer seeds (the so-called filtration strategy) and using special data structures (e.g., suffix arrays) to organize all substrings of the sequences (Li and Homer, 2010).

Obtaining high-quality alignment involves tradeoffs in processing speed versus accuracy. Various strategies center around ways to allow for more sensitive alignments without exacting too high a computational penalty. One important consideration is that algorithms that allow gapped reads can be costly for computation. For genomic sequencing, an alignment algorithm that does not allow indels can generate a large number of false-positive SNPs. However, for RNA sequencing, transcript counts are the desired output, and false SNPs are not as important. Therefore, algorithms that try to increase the sensitivity of the alignment, say by allowing larger deviations, are more important than those that try to increase specificity and accuracy.

Increasing alignment specificity

Increasing alignment specificity may involve a consideration of the specific sequencing experiment. For example, *in vitro* transcription (IVT)–amplified RNA (Van Gelder et al., 1990), used in single-cell transcriptome analysis, tends to create short transcript templates with 5' poly-T leaders in the amplified RNA. Many of the fragments in the library will keep the 5' poly-T sequence, which needs to be trimmed for effective alignment. Aligning across potential splice variants also creates challenges, and typical strategies involve using known splice signals and intron–exon boundaries to increase the reference variants. However, using only known gene models may impede the detection of novel splice variants.

NOTES

Increasing algorithm sensitivity

One possible solution is to use a hierarchical processing strategy where the reads from a sample are processed through algorithms of increasing sensitivity. For example, the nearly exact reads might be first mapped using efficient algorithms, and the remaining reads might be processed through increasingly sensitive algorithms. The recently developed program RNA-Seq Unified Mapper (RUM) (Grant et al., 2011) utilizes such a strategy. The downside of this strategy is that computational time may greatly increase, depending on the particular sample. For example, processing 100 million 100 bp reads through fastest aligners (e.g., BOWTIE and BWA [Burrows–Wheeler Aligner]) (Langmead et al., 2009; Li and Durbin, 2009) takes ~30 CPU hours on typical computers, whereas RUM may take up to 1,500 CPU hours. (Note that RUM is just another variation of filtration strategy.) Another problem (in addition to computational time) is that with increasing sensitivity comes the potential increase in false-positive alignments. Because the algorithms involve heuristic tradeoffs between sensitivity and specificity, the researcher has to make a decision between optimizing these two objectives.

Importance for RNA sequencing

The key issue for RNA sequencing is whether different alignment strategies produce biased samples of true transcripts, regardless of their false-negative and false-positive rate. In our experience, there is a considerable variation in read counts mapped to specific transcript models, depending on the alignment algorithm used. Unfortunately, the particular types and degree of biases are still unresolved, and at this time, consistent comparison of datasets requires identical processing of the short read set (as discussed below under Complexities of Quantifying the Transcripts). Lastly, the relatively short length of the sequence reads in next-generation sequencing (100–150 bp) makes it very difficult to consider *de novo* genomes that do not have a reference sequence. The short reads are generally too brief to assemble into a unique transcriptome. However, recent computational approaches have been making progress toward recovering a large fraction of the transcriptome from *de novo* assembly (Grabherr et al., 2011). Also, longer reads from improved chemistry and coupling of paired-end or mate-paired sequencing from multiple insert libraries are expected to lead to effective characterization of novel transcriptomes in the near future.

Benefits of RNA Sequencing

Sequencing RNA provides three major benefits (albeit with caveats to be discussed in the following pages): precision, dynamic range, and the ability to detect novel transcripts.

Precision

Precision of RNA sequencing comes from the ability of a sequence read to uniquely identify the presence of a particular transcribed RNA. If we see a sequence in the high-throughput data that is sufficiently complex that it uniquely maps to the genome, there must be at least one RNA molecule that contains that sequence in the original library preparation. Sequencing chemistry can be surprisingly error-prone at the 3'UTR ends, but if a read maps uniquely to the genome within a prespecified mismatch tolerance, the presence of the molecule can be confirmed with high confidence. The only caveat here is contamination, which is not specific to the instrument, and the possibility of misalignment to a paralogous locus. Alignment to a paralogous locus can be a problem, especially if the study strain has polymorphisms vis-à-vis the available reference genome. Therefore, when considering singular sequence reads as possible evidence of a transcript, it is advisable to carry out additional alignment to the reference genome under less stringent criteria to confirm unique alignment. The numerical precision of the sequencing (i.e., the precision of relative counts of transcript molecules) depends on many factors that will be discussed further below.

Dynamic range

A key advantage of RNA sequencing is that the dynamic range of quantification can be modulated by the sequencing depth. The total number of reads required to recover a rare transcript depends on the cell (tissue) type and the distribution of the frequency of the transcript—that is, the expected frequency of the most highly expressed transcript, the expected frequency of the next most highly expressed transcript, etc. In various single-cell samples, we find a surprising diversity of transcriptome frequency distributions. For example, a mouse brown adipose cell sample recovers ~6,500 distinct transcripts with ~20 million mapped unique reads, whereas a rat cortex cell sample recovers ~17,000 distinct transcripts with ~10 million mapped unique reads.

We can approximately compute desired sequencing depth using a variation of the coupon collector's problem: Given the need to collect N distinct coupons in a game, what is the expected number of total coupons needed? In the optimal case, in which all distinct transcripts have equal abundance in the transcriptome, we need ~1.8 million mapped reads to recover 10,000 distinct transcripts with

NOTES

95% confidence (using Markov inequality). In our experience, a typical high-throughput sequencing experiment yields only 25% high-quality, unique paired-end RefSeq mapped reads. Therefore, under the optimal scenario, we need ~8 million in total read depth to recover 10,000 distinct transcripts with high probability. However, as mentioned, some transcripts are much more common than others, greatly skewing this computation. Assuming 100,000 total RNA molecules in a cell, and assuming only a single molecule of a rare transcript, similar computations suggest that we need ~100 million total reads to recover all transcripts (including the most rare transcript) with high confidence. Optimal read yield from Illumina HiSeq Systems is on the order of 350 million reads per lane. Therefore, these calculations suggest 3-fold multiplexing per lane to recover the rarest transcripts.

Ability to detect novel transcripts

As mentioned above, many studies using RNA sequencing are reporting novel transcripts. For example, using RNA sequencing from mechanically dissected dendritic samples, we found that up to 56% of the expressed genes in the mouse hippocampal cells and 50% of the expressed genes in the rat hippocampal cells show evidence of intronic sequences in the cytoplasm: cytoplasmic intron-sequence-retaining transcripts, or CIRTs (Bell et al., 2010; Buckley et al., 2011).

One characteristic of Illumina's sequencing chemistry is that, for every double-stranded template insert, reads are obtained from only the 5'UTR ends of the sense and antisense strand. The 3'UTR ends of the insert are read only if the insert size is smaller than the requested read length (see below). This chemistry produces a key asymmetry in the mapped reads. A given nucleotide will be covered by reads from both the sense and antisense directions only if the insert was smaller than the read length or the library fragmentation step induced cleavage randomly around the nucleotide. This means that if a transcript has a definite end (e.g., in the 5'UTR or the 3'UTR), the reads from the ends will be mostly from a single direction.

Figure 1 shows a moving window plot-of-read density for the 3'UTR end of the *Grin2b* gene from the rat hippocampal transcriptome. The red and blue lines show read density in each direction. Clearly visible is a shift in the density owing to the strand directional bias of the Illumina sequencing chemistry. This bias can be exploited by computing the differential of the read densities in the two directions, shown as black lines with blue fill. A sharp peak in the differential curve indicates the presence of a natural 3'UTR end of the transcript. The horizontal blue bar indicates previously annotated coding sequence and 3'UTR for this gene (thick and thin bars, respectively). As can be seen, these RNA sequence data indicate a novel 3'UTR for this gene. We have used this kind of computational procedure to map 3'UTR isoforms for the rat hippocampal transcriptome.

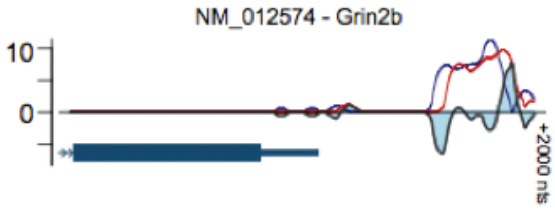

Figure 1. Read-density plot for the *Grin2b* locus. Blue denotes sense direction reads, red denotes antisense direction reads, and blue-filled black curved lines denotes differential in the two directions.

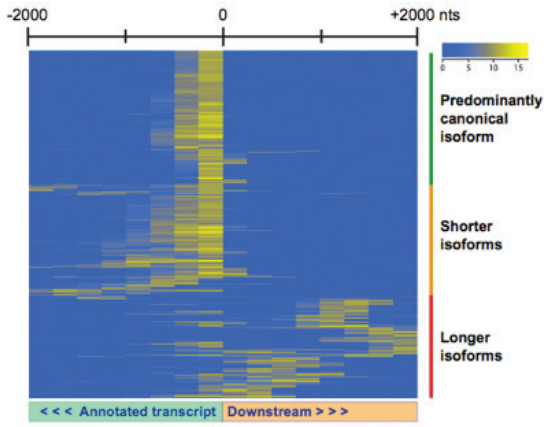

Figure 2. Heat plot of 3'UTR ends from rat hippocampal transcriptome. Rows indicate different genes. Position zero on the *x*-axis indicates canonical 3'UTR annotation.

Figure 2 shows a heatmap of estimated 3'UTR ends, where the coordinate 0 indicates the previously annotated 3'UTR for these transcripts. We found evidence that some genes have more than seven different end-isoforms, and two-thirds of the transcriptome show novel, previously unannotated 3' UTRs.

Complexities of Quantifying the Transcript

Once the short read set has been mapped to the reference genome, quantifying the transcript numbers has several complexities. We first assume that the RNA sample has been prepared to satisfactory quality

NOTES

(i.e., we assume that quality issues not specific to RNA sequencing are not part of the problem). The RNA pool is typically fragmented, cDNA is created to an appropriate size class, and adaptors are ligated for library amplification and sequencing.

Bias correction

Fragmentation and cDNA creation bias

Many authors have noted biases in the library resulting from both the fragmentation and cDNA creation step (Bullard et al., 2010; Hansen et al., 2010). Even without the bias, however, longer transcript molecules will be sampled more frequently during fragmentation and thus be more accurately measured, leading to greater statistical power for detecting differential expression (Oshlack and Wakefield, 2009). Several ad hoc bias correction methods have been suggested, but the optimal procedure is still uncertain at this point. In our experience, a pile-up visualization of the RNA sequencing reads on the genome shows clear heterogeneities. These include a large amount of reads that locate to a focal region or regions, with complete absence of reads despite high coverage in other adjacent regions. These kinds of variations are difficult to completely control and are likely to lead to artifactual theories of the transcriptome.

PCR bias

The PCR step in library construction can also lead to counts that are nonlinear in terms of input molecules and to a tendency to inflate the counts of more frequent molecules. The PCR bias can be modeled by noting the reads that map to nearly identical locations of the genome.

Associating read counts and normalizing read depth

The more critical problem is associating read counts to transcript models and normalizing the read counts to quantities that are comparable across different sequencing libraries. Different RNA preps and library preps yield different numbers of total reads and mapping reads. Initial attempts at quantification divided the reads mapping to a transcript model (e.g., RefSeq annotations) by the total number of mapping reads and the length of the transcript model. These calculations resulted in quantities such as reads per kilobase of exon model per million mapped reads (RPKM), which is still commonly used. Model-based methods have been proposed wherein the read coverage at any given base pair is assumed to be a Poisson sample with an unknown intensity parameter that represents the biological transcription level. Several variations of the model-based approach take into account possible intensity variation across a putative transcript molecule owing to such factors as fragmentation during library construction and convolution of biological variation from different samples.

Normalizing for read depth is also not so simple because the total mapped reads can be dominated by a small number of highly expressed genes. In such a case, there will be loss of sampling of more moderately expressed genes, distorting the estimate of relative expression levels. One simple corrective approach that has been suggested is to normalize the counts by a quantile of the read counts, such as the 75% quantile (i.e., every library is normalized such that the 75th percentile read count of a gene is 1).

Nonunique mapping reads and isoforms

The two largest problems with quantification are how to handle nonunique mapping reads and how to handle multiple isoforms of a given transcribed region of the genome. Nonunique maps can result either from redundant sequences of the genome or from overlapping transcriptional units. The former may be resolved with increasing sequence read length, but the latter has a biological origin and thus will be difficult to resolve without full-length sequencing of the transcript.

Isoforms of a transcript result from alternative splicing and lead to dependencies between reads and genomic regions: That is, the same read may result from multiple transcript molecules. Approaches to the isoform problem involve fitting the read data as samples from multiple transcript models. The models might involve using existing annotations of possible transcripts or estimating splice variants *de novo* by generating the best fitting models.

Variations among programs

Even when the algorithms do not try to deconvolute the read data into distinct isoforms, considerable variations can be found in the quantification because different programs handle the multiple reads and transcript models (i.e., the unit of quantification) differently. An important confounding factor is that these problems are sequence-specific and therefore affect different genes in different ways. A computational analysis of the mouse genome suggests that there are fewer than 1,000 possible transcripts without problems associated with transcript variations and overlapping transcript units.

NOTES

Evolving procedures to address complexity

A growing body of literature is addressing these quantification complexities, and we expect the procedures to evolve (Marioni et al., 2008; Bullard et al., 2010; Li et al., 2010; Trapnell et al., 2012). Some experimental protocols, such as ABI Solid SAGE (Applied Biosystems), attempt to characterize only 3'UTR tags, but we have found that the resulting sequences still contain potential artifacts that must be postprocessed. RNA sequences from IVT-amplified single cells have additional characteristics that modulate the quantification process. The IVT protocol involves transcript selection (using 3' poly-A or other A-rich sequences) and template-shortening due to multiple rounds of random hexamer priming. The template-shortening makes it less important to correct for length of the transcript model, but the template selection based on poly-A sequence requires one to consider the relationship of any other A-rich regions *cis* to the putative transcripts.

While these complexities may make RNA sequencing data seem hopelessly difficult to obtain, two facts should be recognized:

(1) Early microarray data required considerable research to arrive at uniform protocols for its usage; and
(2) Many of the complexities affect bias in transcript quantification, which may not be critical for most analyses.

Bias in the estimate of transcript levels can affect absolute quantification but will not affect analysis of differential expression or variational analysis (e.g., the variation associated with single cells).

There are two important caveats to consider going forward:

(1) If we find a significant difference between two samples, the difference may be the result of reads from overlapping maps. In this situation, the biological genesis of the difference may require further dissection that takes into account possibilities of splice isoforms, independent overlapping transcript units, and other sources of variation; and
(2) All quantitative comparisons across different samples need to be processed through the same computational pipeline; thus, it will be important to make the primary short-read data available for independent analyses.

Characterizing Transcriptome Variation

Jointly with the laboratory of Jim Eberwine, we have been characterizing transcriptome variations across individual cells of various cell types, especially CNS cells in rat and mouse. We typically collect RNA through mechanical isolation from dispersed primary cell culture. It is then amplified by IVT protocols, sequenced using the HiSeq platform (Illumina), mapped with the RUM pipeline, and quantified using custom programs. Once the transcriptome is quantified, the resulting data consist of a vector of numbers, representing the normalized read counts. The number of different transcripts depends on the experiment, but for the single cells we have assayed, the transcriptome ranges from ~6,000 to 14,000 different quantified units. We typically analyze the log transform of the read counts both because the RNA library is PCR amplified and because the RNA samples represent relative densities of RNA rather than absolute numbers. From here on, I assume that the data from each sample are represented by log-normalized read counts, which are equated to a vector in high-dimensional space (i.e., the dimensions correspond to distinct transcripts). Therefore, a dataset of multiple transcriptomes comprises a set of points in this high-dimensional space, which I will call the RNA state space (Kim and Eberwine, 2010).

Clustering analysis

It is now routine to perform clustering analysis of transcriptome data from multiple samples, typically with an accompanying heatmap representation of gene expression levels. Clustering analysis generally falls into the class of machine learning algorithms called "unsupervised learning." That is, the algorithms assume no prior information about the points but instead try to use the spatial distribution of the points to group them into clusters. The general idea is that biologically natural groups (such as distinct cell types and functionally coherent tissues) form spatial clumps in the high-dimensional space.

A whole constellation of algorithms exists, and these algorithms differ mainly as to how they interpret the spatial distribution (e.g., whether they consider certain directions more important than others) and how they impose prior ideas about the structure of spatial distribution (e.g., whether the distribution has a hierarchical organization). In terms of analyzing variation, clustering algorithms are useful for revealing distinct spacings or gaps between points and summarizing high-dimensional relationships that might be difficult to intuitively understand. Their downside is that different algorithms and measures of space within the RNA-state

NOTES

space can result in very different clusters, and there is very little guidance on the "correct" procedure.* Nonetheless, clustering the points gives important information on the degree of data heterogeneity, and we typically use the technique to complement other kinds of high-dimensional analysis.

Dimension-reduction techniques

A major problem with high-dimensional data is the number of dimensions itself. This is especially exacerbated in transcriptome data, where the number of variables (i.e., the different transcripts) vastly outnumbers the number of observations (e.g., cells, tissues, and experiments). This mismatch potentially leads to greatly overfitting complex models to sparse data. For example, given enough dimensions, one could easily come up with diagnostic markers for any reasonable classification of the input data.

Several important techniques have been developed to mediate this problem which typically involve either dimension reduction or methods to limit model complexity. Dimension reduction usually involves projecting the original high-dimensional data to lower dimensions. Projections involve taking the original high-dimensional points and projecting their positions onto some geometric object within that space, for example, a line. In fact, each individual coordinate can be seen as a particular projection onto a particular set of orthogonal lines.

Singular value decomposition and principle component analysis

Singular Value Decomposition (SVD) and the related Principle Component Analysis (PCA) have been used extensively in transcriptome analysis. In these techniques, an orthogonal set of linear projections are constructed in which each projection is, in effect, the line closest to the current distribution of points. These techniques transform coordinates into orthogonal coordinate axes, where each dimension can be ordered in terms of how much of the original dispersal pattern is captured on respective projections. This allows both visualization and dimensional reduction. For example, with the assumption that biologically meaningful transcriptome variation is found only in a small number of dimensions, the original data can be reduced to the projection in the PCA directions, and all subsequent analysis can be limited to the reduced dimensions. The caveat is that PCA directions typically tend to involve a very large number of genes, and therefore, the interpretation can become strained in terms of individual genes.

*It is important to note here that algorithmic "learning" from high-dimensional data is generally a difficult problem because it involves inferring potentially complex models of the data *ab initio* rather than fitting the data into simple models such as differential gene expression. Thus, statistics and mathematics used in many approaches have considerable degrees of freedom in determining the significance of any result.

Linear discriminant analysis

A useful dimension-reduction technique is Linear Discriminant Analysis (LDA), in which the projections to lines maximize the separation between a priori classes of points. Figure 3 shows a three-dimensional projection of a single-cell transcriptome from eight different cell types (shown in different colors) using PCA projections (Fig. 3A) and LDA projections (Fig. 3B). As can be seen in this picture, the LDA projections emphasize the separation of the different a priori classified cell types. In effect, each dimension in the LDA projections is a weighted combination of the expression level of genes that best separate the cell types.

Partial Least Squares

Another projection technique is Partial Least Squares (PLS). PLS projection is useful if there is another continuous response variable that is assumed to be a function of the transcriptome, e.g., cell size, cell physiology, or signaling output. The projection tries to find a set of orthogonal lines (directions) in the RNA-state space that best explains the response variable.

Nonlinear projections

Lastly, projections do not have to be linear (i.e., project to lines). For example, we might imagine that, given enough data points, the transcriptome from single hippocampal cells forms nonlinear curves in the transcriptome space (say, because the RNA products have to form dimers and satisfy quadratic stoichiometric relationships). Techniques such as Locally Linear Embedding (LLE) (Roweis and Saul, 2000) aim to detect and characterize such nonlinear geometric distributions.

Deriving transcriptomes from single cell types

The distribution of transcriptomes for single cells or tissues within the RNA-state space may have complex structures. One way to think about single-cell transcriptomes is that particular levels of RNA expression are maintained for a given cell because certain RNA molecules are required to satisfy the stoichiometric relationship of functional reactions involved in the cell's phenotypic function. For example, a neuron might require the maintenance of certain ratios of different glutamate receptors. The collective effect of such stoichiometric constraints limits the viable points in the RNA-state space for a particular cell type. If there are 10,000 different transcripts in the

NOTES

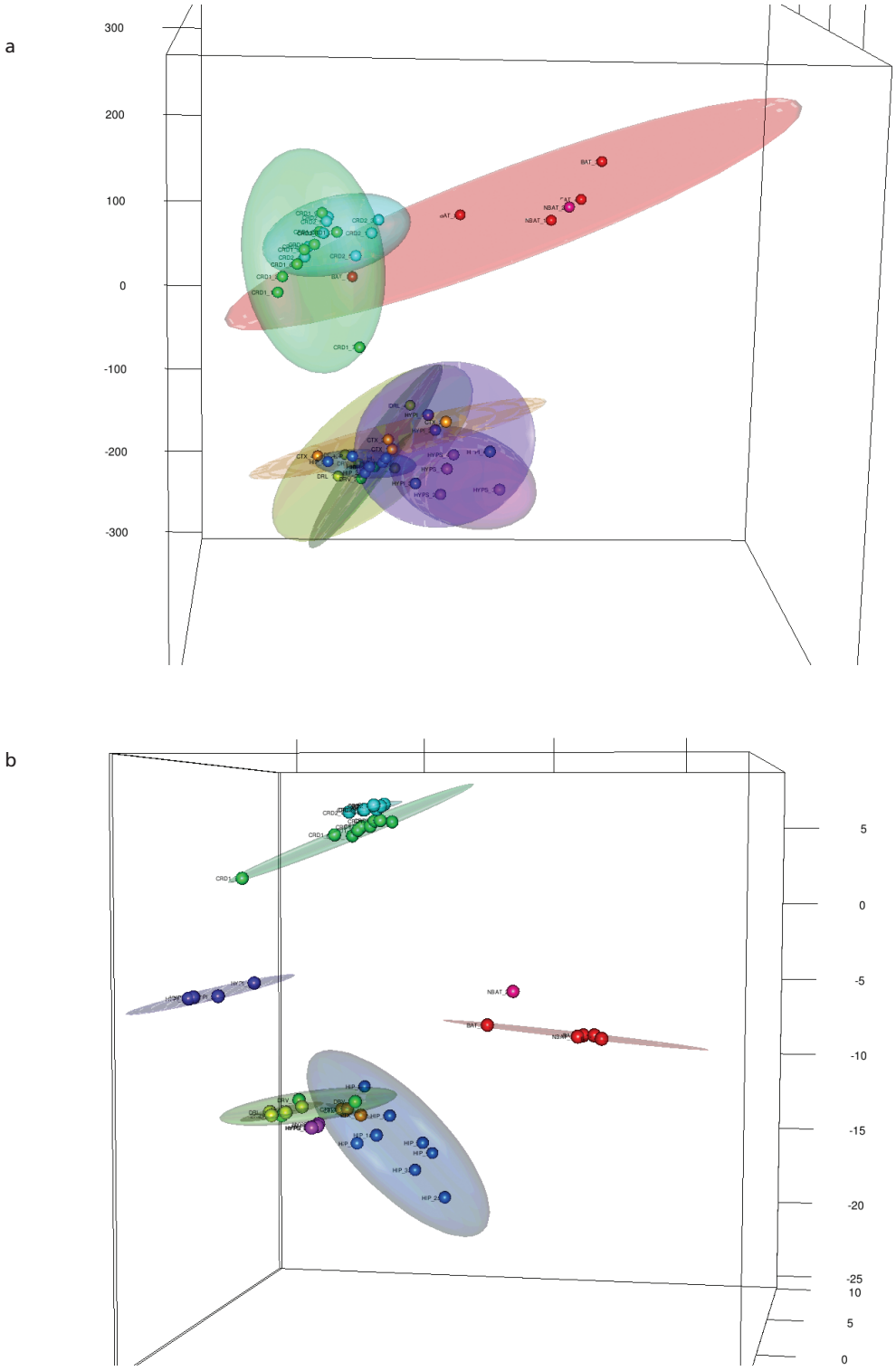

Figure 3. Three-dimensional projection of high-dimensional single-cell transcriptome data from 8 different cell types. ***a***, PCA axis projection; ***b***, LDA axis projection. The axes of both figures are abstract, and the numerical values represent linear combinations of the original variables. The values have no direct interpretation in terms of original data values. The PCA axes are meant to emphasize the overall variation, while the LDA axes emphasize the distinction between groups.

NOTES

cell, then each constraint reduces the viable dimension by one. If there are 10,000 constraints, then we might expect the transcriptome to maintain a particular set of expressions, i.e., be concentrated around a single point. If there are fewer than 10,000 constraints, then the transcriptome has multiple degrees of freedom and the single-cell transcriptomes might form a broad distribution, as seen in Figure 3.

Given enough data (i.e., transcriptomes from multiple single cells of the same type), it might be feasible to characterize the viable functional transcriptome states of a particular cell type using these projection techniques. It may also be possible to identify the physiological constraints for these cells' function. In the last part of the talk, I will present some potential models for analyzing such single-cell variation data.

References

Bell TJ, Miyashiro KY, Sul JY, Buckley PT, Lee MT, McCullough R, Jochems J, Kim J, Cantor CR, Parsons TD, Eberwine JH (2010) Intron retention facilitates splice variant diversity in calcium-activated big potassium channel populations. Proc Natl Acad Sci USA 107:21152–21157.

Buckley PT, Lee MT, Sul JY, Miyashiro KY, Bell TJ, Fisher SA, Kim J, Eberwine J (2011) Cytoplasmic intron sequence-retaining transcripts can be dendritically targeted via ID element retrotransposons. Neuron 69:877–884.

Bullard JH, Purdom E, Hansen KD, Dudoit S (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics 11:94.

Core LJ, Waterfall JJ, Lis JT (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science 322:1845–1848.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644–652.

Grant GR, Farkas MH, Pizarro AD, Lahens NF, Schug J, Brunk BP, Stoeckert CJ, Hogenesch JB, Pierce EA (2011) Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq Unified Mapper (RUM). Bioinformatics 27:2518–2528.

Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, Rinn JL, Lander ES, Regev A (2010) Ab initio reconstruction of cell type–specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol 28:503–510.

Hansen KD, Brenner SE, Dudoit S (2010) Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res 38:e131.

Kim J, Eberwine J (2010) RNA: State memory and mediator of cellular phenotype. Trends Cell Biol 20:311–318.

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25.

Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics 25:1754–1760.

Li H, Homer N (2010) A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform 11:473–483.

Li J, Jiang H, Wong WH (2010) Modeling non-uniformity in short-read rates in RNA-Seq data. Genome Biol 11:R50.

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 18:1509–1517.

Marquez Y, Brown JW, Simpson C, Barta A, Kalyna M (2012) Transcriptome survey reveals increased complexity of the alternative splicing landscape in *Arabidopsis*. Genome Res 22:1184–1195.

Morin RD, O'Connor MD., Griffith M, Kuchenbauer F, Delaney A, Prabhu AL, Zhao Y, McDonald H, Zeng T, Hirst M, Eaves CJ, Marra MA (2008). Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. Genome Res 18, 610–621.

Nechaev S, Fargo DC, dos Santos G, Liu L, Gao Y, Adelman K (2010) Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in *Drosophila*. Science 327:335–338.

Oshlack A, Wakefield MJ (2009) Transcript length bias in RNA-Seq data confounds systems biology. Biol Direct 4:14.

Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature 464:768–772.

Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. Science 290:2323–2326.

Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25:1105–1111.

Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol 28:511–515.

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7:562–578.

Van Gelder RN, von Zastrow ME, Yool A, Dement WC, Barchas JD, Eberwine JH (1990) Amplified RNA synthesized from limited quantities of heterogeneous cDNA. Proc Natl Acad Sci USA 87:1663–1667.

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63.

NOTES