NIH Public Access
Author Manuscript
Curr Protoc Hum Genet. Author manuscript; available in PMC 2015 October 01.

Published in final edited form as:
Curr Protoc Hum Genet. ; 83: 11.13.1–11.13.20. doi:10.1002/0471142905.hg1113s83.

RNA-seq Data: Challenges in and Recommendations for Experimental Design and Analysis

Alexander G. Williams[1], **Sean Thomas**[1,2], **Stacia K. Wyman**[1], and **Alisha K. Holloway**[1,2]

[1]Gladstone Institute of Cardiovascular Disease, San Francisco, CA

[2]Department of Epidemiology and Biostatistics, University of California, San Francisco, CA

Abstract

RNA-seq is widely used to determine differential expression of genes or transcripts as well as identify novel transcripts, identify allele-specific expression, and precisely measure translation of transcripts. Thoughtful experimental design and choice of analysis tools are critical to ensure high quality data and interpretable results. Important considerations for experimental design include number of replicates, whether to collect paired-end or single-end reads, sequence length, and sequencing depth. Common analysis steps in all RNA-seq experiments include quality control, read alignment, assigning reads to genes or transcripts, and estimating gene or transcript abundance. Our aims are two-fold: to make recommendations for common components of experimental design and assess tool capabilities for each of these steps. We also test tools designed to detect differential expression since this is the most widespread use of RNA-seq. We hope these analyses will help guide those who are new to RNA-seq and will generate discussion about remaining needs for tool improvement and development.

Keywords

RNA-seq experimental design; biological replicates; sequence length; sequencing depth; splice-aware alignment; paired-end sequencing; transcript abundance; differential expression

Assessing the content and abundance of RNA for the entire transcriptome has revolutionized our ability to understand gene expression over developmental time, between genotypes, drug treatments, or tissues, and even differences between species. Almost 20 years ago, the first microarrays were a breakthrough technology that allowed us to assess RNA content and abundance for all *known* transcripts simultaneously. RNA-seq has extended the scope and depth of investigation to the entire transcriptome of known and novel RNAs. This rich experimental technique has become a standard in many labs and the tools developed for analysis are constantly improving. Tools have been developed for assessing allele-specific expression(Skelly et al., 2011), quantifying alternative transcript usage(Hu et al., 2013; Emig et al., 2010; Sacomoto et al., 2012; Nicolae et al., 2011; Singh et al., 2011; Richard et

Alexander G. Williams: alex.williams@gladstone.ucsf.edu, 415-734-2746
Sean Thomas: sean.thomas@gladstone.ucsf.edu, 415-734-2738
Stacia K. Wyman: stacia.wyman@gladstone.ucsf.edu, 415-734-2000
Alisha K. Holloway: alisha.holloway@gladstone.ucsf.edu, 415-734-2723

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

al., 2010; Trapnell et al., 2010), and discovering novel transcripts(Roberts et al., 2011) or gene fusions(Sakarya et al., 2012). RNA-seq can also be used for ribosomal profiling to precisely measure the translation of each transcript(Ingolia et al., 2009). By far, the most widespread usage of RNA-seq methods is for experiments that are designed to detect differential gene expression between two or more experimental groups and many advances have been made in the tools designed to detect differential expression(Wu et al., 2013; Chung et al., 2013; Soneson and Delorenzi, 2013; Anders et al., 2013; Anders and Huber, 2010; Robinson et al., 2009; Wang et al., 2009; Law et al., 2014).

Regardless of the exact experiment, each RNA-seq experiment consists of several common steps: experimental design, quality control, read alignment, assigning reads to genes or transcripts, and estimating gene or transcript abundance. Several excellent papers have been published over the past few years that critically assess one or more facets of RNA-seq experimental design and analysis (e.g., (Fang and Cui, 2011; Bullard et al., 2010; Auer and Doerge, 2010; McIntyre et al., 2011; Nookaew et al., 2012; Grant et al., 2011; Marioni et al., 2008; Robles et al., 2012; Vijay et al., 2013; Roberts et al., 2011; Soneson and Delorenzi, 2013; Katz et al., 2010; Young et al., 2011)). However, rapid changes in technology that have allowed for longer reads, deeper sequencing, and lower costs, have changed the complexity of experimental designs, which requires us to constantly reevaluate what is the best approach for RNA-seq experiments.

Our aims for this work were to assess experimental design and available tools for the common steps of an RNA-seq experiment. We used real RNA-seq data as well as simulated data to 1) test and suggest parameters for experimental design, and 2) test software for each step of an RNA-seq analysis designed to detect differential expression. The experimental design elements that we assessed include: number of replicates, sequence read length, read depth, and whether to do paired-end (PE) or single-end (SE) sequencing. Laboratory protocols for the RNA extraction method, assessing RNA quality, barcoding/indexing, and library sequencing protocol are critical for getting good sequence data, but we will not discuss those parameters in-depth here. Our hope is that these recommendations on experimental design and tool choice will serve as a guide to the community and generate discussion surrounding the remaining needs for tool development or improvement of existing tools for RNA-seq experiments.

Data Set

Sequence data for the following analyses are from cardiac precursors and cardiomyocytes that were derived by differentiating mouse embryonic stem cells. The embryonic stem cells were generated by culturing blastocysts from Smarcd3 (Baf60c) del/+ intercrosses (Hota et. al, in prep). The wild-type ES cells were derived from the littermate controls. There were a total of 12 samples, with n=3 for each combination of cell type and genotype. RNA-seq libraries were prepared with ovation RNA-seq system v2 kit (NuGEN). The double-stranded DNA was then amplified using single primer isothermal amplification (SPIA). Random hexamers were used to amplify the second-strand cDNA linearly. Finally, libraries from the SPIA amplified cDNA were made using the Ultralow DR library kit (NuGEN). The RNA-seq libraries were analyzed by Bioanalyzer and quantified by qPCR (KAPA). All 12 indexed

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

samples were sequenced in each of three lanes on an Illumina HiSeq 2500. The data are PE 100 bp sequences with an insert size of 300-350 bp. On average, 57 million fragments were sequenced from each sample. Raw sequence data can be downloaded from GEO (accession # TBD). Initially, data were aligned with Tophat2 (v2.0.8b) to mm9 with mostly default settings (except the following: --mate-inner-dist=150 --no-discordant --no-mixed --no-coverage-search --splice-mismatches=2 --microexon-search). The annotation is from Ensembl v66 annotation(Flicek et al., 2014); pseudogene annotations were removed and only standard chromosomes were used. Of 684 million fragments, 67.8% of read pairs mapped concordantly.

Technical Considerations and Quality Control

Minimizing Technical Variation

Technical variation in RNA-seq experiments stems from many sources, such as differences in quality and quantity of RNA recovered during sample preparation, library preparation batch effect, flow cell and lane effects when using Illumina technology, and adapter bias(Bullard et al., 2010; Fang and Cui, 2011; Auer and Doerge, 2010; McIntyre et al., 2011). Bullard et al. (2010) showed that library preparation was the largest source of technical variation, but that technical variation was minimal compared to the level of biological variation between samples from different tissues (Ambion brain vs. the Universal Human Reference). However, other work shows that technical replicates can have high variance and that estimation of expression is most severely compromised for lowly expressed genes(McIntyre et al., 2011). Auer and Doerge (2010) provide an excellent review of experimental design, especially with respect to the importance of replication and technical variation. As Auer and Doerge state, “No amount of statistical sophistication can separate confounded factors after data have been collected”(Auer and Doerge, 2010). Given these concerns, we recommend that samples be randomized during preparation and diluted to the same concentration. Then, to mitigate the effects of flow cell and lane we recommend indexing and multiplexing samples, with all samples included on all lanes/flow cells. However, some experiments have more samples than the number of available barcodes or indexes. In cases where all samples cannot be multiplexed on all lanes, a blocking design can be used that includes some samples from each group on each lane of sequencing(Auer and Doerge, 2010; Fang and Cui, 2011).

Pooling samples vs replicates

Generally, for RNA-seq, each biological replicate within an experimental group is prepared separately. Data from each replicate are then used in statistical analysis, with biological variance estimated from the replicates. An alternative design is to pool biological replicates within a group before library construction and sequencing. A pooled design removes the estimate of biological variance. We created “pooled” samples from our biological replicates by combining counts over samples. For each replicate, the total depth was accounted for before combining counts. We then conducted tests of differential expression for pooled samples using a binomial test vs. biological replicates using a test based on a negative binomial distribution (DESeq2, (Anders and Huber, 2010) and found that FDR-adjusted p-values were correlated (Spearman’s Rho r=0.9). However, our data set has very low

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

biological variance within groups. Genes with high variance in expression may appear to be differentially expressed in a pooled design; this is especially problematic for genes with low expression levels. Additionally, when biological variance is low, replicates add power to the statistical test and subtle changes in gene expression can be identified. Biswas et al., (2013) conducted a thorough analysis of separate vs. pooled replicates. They were careful to ensure that each biological replicate was diluted to the same concentration of RNA before pooling. They concluded that that most differentially expressed genes can be identified with a biologically averaged design and that this design may be a good alternative to reduce the cost of experiments. However, when cost is not a limiting factor, maintaining separate biological replicates is ideal.

PCR duplicates

During library preparation each sample goes through a PCR amplification step where fragments are amplified so that there is sufficient material to load onto the sequencer. It is expected that a single copy or very few exact copies of each fragment would be sampled during sequencing. Therefore, if the PCR amplification step is completely unbiased, it should be very rare for two or more sequenced fragments to have the same start and end positions. However, depending on the sequencing depth, complexity of the library (i.e., number of transcripts expressed), and expression levels of transcripts, some fragments may by chance completely overlap yet would not be PCR duplicates. Under naïve conditions with no bias, we have simulated the expected number of these overlapping fragments with differing levels of expression. We simulated 30 million reads from 15,000 genes with a mean gene length of 2 kb, which would give an average of 67 FPM per gene. Fragment lengths were drawn from a normal distribution with mean 200. Expression level for genes ranged from 0.1 to 10 fold of the average. Our simulations show that the proportion of duplicates should be very low, even for highly expressed genes (about 5% for genes that are expressed 10-fold higher than the mean expression level). In experiments, RNA-seq data vary widely in the total proportion of duplicated fragments and this variance is greater than expected from our simulations. PCR bias is a well-known phenomenon that results in more amplification of some fragments and less amplification of other fragments. One major factor in amplification bias is the base composition of the fragment(Fang and Cui, 2011). Fragments with higher levels of amplification may be sampled more often during sequencing. This leads to oversampling of some PCR amplicons and inaccurate estimates of transcript abundance.

We recommend optimizing PCR cycles, providing sufficient RNA for library prep protocols, marking duplicates (using the Picard tool, MarkDuplicates(McKenna et al., 2010)) and assessing the level of duplication during QC. Samples with a similar level of complexity of transcripts that are sampled to equivalent depths should have similar levels of duplication that are due to PCR. We do not recommend removing duplicates for analysis of RNA-seq data that pass QC because this will underestimate abundance of highly expressed transcripts since it is not possible to disambiguate PCR duplicates and fragments that overlap precisely in highly expressed genes.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Quality Control

Quality assessment is fundamental both before and after alignment. To assess the quality of RNA-seq data, we use FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) for raw fastq files and then RSeQC(Wang et al., 2012) for aligned (sam/bam) files. We have developed a perl script that aggregates the results of FastQC (http://github.com/staciaw/fastqc_aggregator); the output is an html document that includes each QC metric as a single figure with all samples represented (e.g., Fig. 1). Important QC metrics include: base quality over read length, per sequence quality scores, GC content over read length, and proportion duplicate reads. QC metrics after alignment include proportion mapped, proportion mapped uniquely, proportion duplicate reads, proportion of reads mapping to rRNA, and map location relative to gene annotation (e.g., proportion of reads aligning to exons and other features).

Factors Affecting Alignment and Tools for Splice Junction Mapping

RNA fragments may be sequenced at a single end or at both ends of the fragment (i.e., paired end). Paired-end sequencing provides additional information for alignment because it doubles the amount of sequenced real estate and increases the probability of mapping across splice junctions. Paired-end data can be very useful for estimating alternative transcript usage, identifying novel transcripts, mapping to high-homology regions, and for *de novo* transcript assembly.

Assessment of Sequence Length and the Utility of Paired-End vs. Single-End Reads

We used the gold standard data set of 100 bp PE reads described in the *Splice-Aware Alignment Tools* section to test the effects of mapping with only SE reads as well as mapping fidelity with shorter read lengths. We further filtered the gold standard data set by mapping with Tophat2; the data set includes reads that map unambiguously to a single location (Tophat2 setting –max-multihits=1) and with the pairs mapping concordantly. We also removed duplicate read pairs to allow only unique fragments (MarkDuplicates v. 1.90). For tests of SE vs. PE alignment, we simply aligned the first read of the pair and disregarded the second end. To test reads the effect of read length, we trimmed reads to 35, 50, or 75 bp using fastx trimmer (http://hannonlab.cshl.edu/fastx_toolkit/). Surprisingly, only a few percent of reads from the filtered gold standard data set were not uniquely mapped with the short read or SE read data sets (Fig. 2). With SE, 50 bp reads, over 95% of reads map uniquely. In sum, long PE reads only give marginal improvements over short SE reads in the ability to map reads uniquely.

Splice-aware Alignment Tools

Splice junction mapping is critical for mapping reads across splice junctions and understanding alternative transcript usage. If all exons and splice junctions were known, we could align reads to the exons and splice junctions and then map them back to genomic coordinates. However, new transcripts of known genes and new genes are discovered regularly, necessitating the use of splice-aware mappers. Many tools are now available for aligning reads across splice junctions (Aschoff et al., 2013; Grant et al., 2011; Li et al., 2013; Dobin et al., 2013; Wang et al., 2010; Au et al., 2010; Trapnell et al., 2009; Kim et al.,

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

2013). We compared splice junction mapping in Tophat v2.0.9, MapSplice v2.1.5, and STAR v2.3.1h to assess mappability of splice junction spanning reads. We also subsampled our data to assess the depth of sequencing needed to recover splice junctions.

Gold standard alignments—Alignments to the transcriptome were conducted using the BWA-MEM algorithm in BWA v0.7.5a(Li and Durbin, 2010) in PE mode with default settings and short split hits marked as secondary (-M). We mapped to the transcriptome by creating a transcript index using the Ensembl v66 annotation(Flicek et al., 2014). Each fragment was tagged with a read group ID that maintained sample information. There were a total of 91.8 million fragments with paired reads that mapped concordantly and uniquely. These fragments were used for all subsequent analyses except where noted that simulated data were used. Transcript coordinates were mapped back to the genome using a custom python script developed by N. Salomonis (getCoordinates.py) that is available upon request.

We aligned the complete set of fragments that mapped concordantly, but only assessed splice junction mapping in genes with a single transcript. We used this single transcript set to mitigate any confounding issues with mapping around small alternatively expressed exons. We also created subsets of the full gold data set in order to test the depth of sequencing needed to recover splice junctions. We sampled 10-80% of read pairs without replacement for the subset data sets.

Tophat2—Bowtie2(Langmead and Salzberg, 2012) provides high quality alignments to the genome and Tophat2(Kim et al., 2013) employs Bowtie2 to map reads in two ways. Reads are mapped first to the to the genome, which maps reads that do not span splice junctions and also helps identify exons. Unmapped reads are possible splice junction spanning reads, which cannot be mapped contiguously. Tophat2 then segments these reads and segments are mapped with Bowtie2 to find mapping locations for splice junction spanning reads. We also show results for mapping with Bowtie2 alone, which is a short read mapping program not designed to be a splice-aware aligner. This test was intended to show the importance of using a splice-aware aligner for mRNA-seq data.

Tophat2 (v2.0.9) was run with the mostly default settings (except for: --read-mismatches 2, -- mate-inner-dist 200, --splice-mismatches 1, --microexon-search, and --no-novel-juncs). Transcript annotations were from Ensembl v66 with only single transcript genes. Tophat2 was run with annotation and without annotation included. We tested Tophat2’s ability to find “novel” junctions by removing the second exon from transcripts in the annotation and (i.e., without --no-novel-juncs in the command) and results were similar to those when Tophat2 is provided with an annotation (results not shown). We also ran Tophat2 with just the first read of the PE data set. Finally, we tested splice junction recovery with shorter reads by truncating the reads at 50 bp and ran Tophat2 with PE data as well as with just the first read.

MapSplice—The algorithm for MapSplice has two phases that include tag alignment and splice inference(Wang et al., 2010). First tags are partitioned into short segments and aligned to the genome. Unmapped segments are considered for splice junction spanning beginning with location of neighboring aligned segments. Segment alignments for a tag are

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

then merged and scored; candidates with the highest score based on alignment and junction quality are reported.

MapSplice (v2.1.5) does not use transcript annotation. We used default settings except that we allowed non-canonical splice junctions (i.e., with the --non-canonical setting). We ran MapSplice on the PE data as well as just with the first read from the data set.

STAR—STAR is an extremely fast aligner (approximately 30-40× faster than Tophat2), but requires a significant amount of RAM (~27 GB for a vertebrate genome)(Dobin et al., 2013). STAR aligns non-contiguous sequences to a reference genome in two steps: a seed searching step and then a step that clusters, stitches, and scores possible alignments. In the seed searching step, STAR finds the maximal mappable prefix (MMP), then takes the unmapped portion of the read and finds the MMP for that segment. In the second step, segments from a read are stitched together with any number of mismatches but only a single insertion or deletion. The stitched combination with the highest score is selected as the best alignment of the read.

The splice junction database for STAR (v2.3.1h) was created using the command “genomeGenerate” with the Ensembl v66 annotation of single transcript genes. Default parameters were used except the splice junction database overhang (--sjdbOverhang) parameter was set at 100. STAR alignments were conducted with mostly default parameters with and without annotation. Default settings require 66% match for each read. Therefore, we ran STAR with relaxed criteria for match (--outFilterScoreMinOverLread 0.5 --outFilterMatchNminOverLread 0.5). We also ran STAR with just the first read of the PE data set.

We compared these splice-aware mapping tools using our gold set of concordantly mapped read pairs from the BWA alignments to the transcriptome for single transcript genes. The splice-aware mapping tools we tested recovered 95-99% of splice junctions that are spanned by 5 or more reads (Fig. 3A, Table 1). Junctions with low coverage (<5 fragments) in the BWA alignments (~14% of total) are much less likely to be recovered with splice-aware mapping tools (Fig. 3A, Table 1). It is interesting that 50 bp PE reads have a higher proportion mapped across splice junctions than 100bp PE reads (Table 1). This is possibly due to the likelihood of longer reads including regions with small indels or variants that exceed the setting for allowed mismatches (Tophat2 setting --read-mismatches 2). The vast majority of junctions are spanned by at least some fragments for all analyses (Fig. 3A, Table 1) and the absolute counts are quite similar between algorithms. STAR appears to be a more sensitive splice-aware aligner as it maps a higher proportion of splice-junction spanning reads than the other two mappers (Table 1). As expected, fewer junctions are recovered with smaller subsamples, especially when junctions are covered by fewer reads (Fig. 3B). Counts for both PE and SE analyses were highly correlated with counts from BWA (Fig. 4); STAR and Tophat2 performed extremely well. Each fragment is counted a single time for PE or SE analysis. The BWA alignment with PE data has about 2-fold more splice junction spanning fragments than are captured by the SE mappings with other tools; this is expected since there is approximately 2-fold more genomic real estate covered with the PE data.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Sequencing depth is an important consideration with RNA-seq experiments. With 55 million mapped read pairs, 94% of junctions identified in the full data set are recovered when using Tophat2 and providing an annotation (Fig. 3C).

Overall, the splice-aware aligners performed well. STAR aligned more reads to the same location as the BWA alignments, recovered more splice junctions, and runs at least an order of magnitude faster than Tophat2. However, STAR does require a significant amount of RAM. Tophat2, when provided annotation, has a higher correlation with the bwa alignments in total number of reads assigned to a splice junction. MapSplice, which does not use an annotation, performs as well as Tophat2 without annotation.

Recommendations for sequence length, paired vs. single-end, and splice-aware aligners

Logically, we assumed that long, PE reads would map with the highest fidelity, and provide the most information about splice junctions and alternative transcript abundance. Nevertheless, we find that SE reads recover junctions fairly well and that short reads are sufficient for unique mapping. Over 95% of SE, 50 bp reads align uniquely compared to those that map from a 100 bp PE alignment using Tophat2 (Fig. 2). We recommend 50bp SE reads for most experiments. The exception is when hypotheses rest on identifying alternative transcript usage, chimeric or fusion transcripts, and in cases where *de novo* transcript assembly is necessary. All three splice-aware aligners that we tested perform well, but STAR slightly outperforms the other two in sensitivity and is extremely fast.

Estimating Transcript Abundance

The goal of this set of analyses was to characterize the properties of several tools designed to calculate the relative abundances of transcripts within a given RNA sample and to make recommendations for best practices for future RNA-seq analyses. A recent paper compares several tools for differential splice analysis(Hooper, 2014), but we are focused on estimating transcript abundance. RSEM (v1.2.8, Li and Dewey, 2011), Cufflinks (v2.1.1, Trapnell et al., 2010), MISO (v0.5.2, Katz et al., 2010), and TIGAR (Feb. 2014, Nariai et al., 2013) were used, with default settings, to estimate the transcript abundance within two replicates (WT cardiomyocyte replicates 1 and 2; see Data Set). We also developed a simple counting technique to estimate the abundance of each transcript. The simple counting method takes the sum of all reads that map to the exons of a transcript. This method overestimates the expression of transcripts for genes expressing multiple transcripts since each fragment must originate from a single transcript.

The estimated counts from each tool were first compared in order to determine the relative precision of each method by performing correlation analysis and scatterplots of the results (Table 2, Fig. 5). Then to compare the accuracy of each method, the estimated counts from RSEM were used to generate a simulated RNA-seq dataset of 10 million tags across the transcripts of chromosome 1. The relative abundance of each 50 bp transcript segment compared to the remaining transcriptome was calculated and 10 million transcript segments were chosen at random from that abundance distribution. To assess accuracy, the seeded values that the simulation was based on were compared to the estimated values obtained by each method (Fig. 7).

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Given the high correlations across biological replicates when comparing the estimates of different methods (Fig. 6A), it is clear that the methods mostly agreed with one another, with Cufflinks and RSEM slightly outperforming the rest in terms of technical reproducibility between the replicates we analyzed (Fig. 6B). Compared to a naïve method that doesn't deal with the complexities involved in assigning tags to transcripts, all methods exhibited good collective precision, and certain transcripts that were problematic for this simple counting method (Fig. 6C) exhibited reproducible, precise values when comparing the estimates made by other methods (Fig. 6B). Those genes that were observed to be problematic for all methods exhibited shorter transcript lengths and fewer average transcripts per gene (Fig. 6D). The precision of the methods can be therefore said to be quite high and consistent across methods. The estimates of expression were also consistent across methods, but many of the transcript abundance estimates were quite inaccurate (Fig. 6E,F). Methods are fairly equivalent, but the level of accuracy is still in question. In order to assess "ground truth", we simulated data for this analysis. Given that it is difficult to simulate RNA-seq data in a biologically realistic manner, we hesitate to make strong recommendations about particular tools. We do encourage researchers to generate browser tracks to visualize read densities of exons that are transcript specific in conjunction with estimating transcript abundance. Estimating transcript abundance is a complex problem that will likely be solved as read length increases.

Detecting Differential Expression: Sequencing Depth, Biological Replicates, Normalization and Estimating Variance

One of the primary goals of RNA-seq analysis is to detect differential expression between two or more experimental groups. Unlike microarrays, which measure expression by fluorescence (a continuous measure), RNA-seq data are discrete counts of sequence reads aligning to a particular gene or transcript. Therefore different statistical tests or transformation procedures must be employed to detect differential expression. Early models assumed a Poisson distribution of reads, which sets the variance equal to the mean, and used a straightforward binomial test to detect differential expression between two samples (or groups). However, we now know that there is greater variation than expected under a Poisson distribution; this extra-Poisson variation is termed overdispersion(Wang et al., 2009; Robinson et al., 2009; Auer and Doerge, 2010; Anders and Huber, 2010). Hence, estimates of biological variance must be incorporated into tests for differential expression. Many tools employ a test based on the negative binomial distribution that allows for estimation of variance. Given the small sample sizes of RNA-seq experiments, most methods borrow information over all genes to estimate variation. In this section will discuss methods for estimating variance, testing for differential expression, as well as testing for appropriate sequencing depth and number of replicates.

Several groups have published thorough comparisons of tools for testing differential expression (e.g., (Soneson and Delorenzi, 2013; Kvam et al., 2012)). We tested EdgeR, DESeq (v1 and v2), and VOOM as well as a simple Poisson model. The goals of our tests are two-fold. First, we are interested in sensitivity and specificity of these tools in detecting

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

differential expression. Second, we are also using our simulated data to test sequence depth and number of replicates needed to detect differential expression.

Depth and Replicates

The depth of sequencing and number of replicates are intimately tied together to give power for accurately estimating transcript or gene-level expression, modeling the biological variance, and ascertaining which transcripts or genes are differentially expressed. Some have suggested small pilot projects to better estimate the complexity of the transcriptome in particular samples(Bashir et al., 2010), which would assist with making decisions about experimental design. Alternatively, Busby et al. (2013) developed Scotty, a tool to estimate the necessary sequencing depth and number of replicates for RNA-seq experiments. Others have described methods for estimating sample sizes for technical and for biological replicates(Fang and Cui, 2011; Hart et al., 2013). An additional concern is that regardless of depth, lowly expressed transcripts are difficult to measure accurately. One possible solution is to remove high-abundance RNAs using capture probes(Łabaj et al., 2011) and then to sequence remaining mid- to low-abundance RNAs. Finally, while we are focused on experiments designed to detect differential expression between two or more groups in this work, RNA-seq can also be used to detect allele-specific expression. Heap et al. have developed a tool for power analysis for allele-specific expression(Heap et al., 2010).

Knowledge of the biology is critical for accurately estimating both sample size and sequencing depth. For example, biological variance is relatively minimal in cell lines or in inbred strains of flies or mice. However, with human samples that are possibly collected at different times or even post-mortem there are significant biological, environmental, and technical sources of variation. The recommendations that we make here are from the analysis of differentiated mouse embryonic stem cells that have minimal genetic variation and little technical variation relative to human samples. We tested recovery of splice junctions based on subsampled data sets (see above, *Splice-aware Alignment Tools*) and found that with 55 million reads mapped, we were able to recover 94% of expected junctions. In this section we test the depth of sequencing and number of replicates needed to identify differentially expressed genes using several different tools.

Normalization

Sequencing depth, RNA composition, and GC content of reads may differ between samples. Therefore, samples must be normalized before they can be compared within or between groups (see (Dillies et al., 2013) for review). Library-size (depth) normalization procedures assume that the underlying population of mRNA is similar between samples, which means that samples with dramatically different compositions of RNA will violate this assumption and cannot be compared without accounting for this difference. For example, a few highly expressed genes can bias library-size based normalization (the per million reads mapped as denominator) when those genes are not highly expressed in all samples. The net effect is that these genes consume a large portion of the total number of reads, which causes other genes to be under-sampled. Bullard et al. (2010) showed that by removing the genes with highest expression (e.g., upper quartile) from the denominator, estimation of expression levels was more accurate for less abundant transcripts, but the methods implemented vary by tool. In

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

more extreme cases, ERCC spike-in controls can be used for normalization when samples are likely to have dramatically different RNA compositions(Lovén et al., 2012).

Sample-specific GC-content differences are another source of between sample variation. Two packages, EDASeq(Risso, 2013) and cqn (conditional quantile normalization) (Hansen et al., 2012) estimate correction factors to mitigate sample-specific GC-content. In our experience, dramatic differences in GC content may indicate a problematic sample, which should also be investigated with QC measures.

The tools we use to test for differential expression between groups implement different methods of normalization. edgeR uses a trimmed mean of M-values (TMM), which is a scaling factor for library sizes that minimizes the log-fold change between samples(Robinson and Oshlack, 2010). They then used counts normalized by this effective library size for all downstream analyses. DESeq and DESeq2 take a similar approach and create a virtual reference sample for each gene by taking the geometric mean of counts over all samples and then normalizing each sample to this reference to get a scaling factor per sample(Anders and Huber, 2010). With VOOM, one can either do a simple library size normalization of log of counts per million, a quantile normalization, or use the TMM as in edgeR(Law et al., 2014); we used the TMM in our tests. Our Poisson-based test uses library-size normalization of reads per million.

Modeling the Biological Variation and Testing for Differential Expression

RNA-seq data give a discrete measurement of the fragments mapping to a particular gene or transcript, which is different from the continuous distribution of expression intensities from microarray data. Appropriate statistical models must be used for these count data. It is natural to consider the distribution of read counts as coming from a Poisson process. However, since the days of SAGE (Serial Analysis of Gene Expression), a precursor to high throughput RNA-seq, we have known that biological replicates exhibit higher levels of variance than can be accounted for with the Poisson distribution (Baggerly et al., 2004; Lu et al., 2005), which assumes the mean equals the variance. DESeq, DESeq2 and edgeR employ the negative binomial (NB) distribution to model this overdispersion, which includes a variance parameter that must be estimated from the data. The NB reduces to the Poisson in cases of no overdispersion. edgeR then estimates dispersion using either the quantile-adjusted conditional maximum likelihood for single factor experimental designs or the Cox-Reid profile-adjusted likelihood for more complex designs(Cox and Reid, 1987; McCarthy et al., 2012); DESeq2 uses the latter method and see (Anders and Huber, 2012) for DESeq. For tests of differential expression, DESeq uses the R function nbinomTest and DESeq2 employs a negative binomial generalized linear model fitting β (log2 fold change) with a gene-specific dispersion parameter for each gene and the Wald test to generate p-values. edgeR uses a model similar to DESeq2, but tests for differential expression using a likelihood ratio test. VOOM instead estimates the mean-variance relationship of the log of read counts by fitting a locally weighted regression (LOWESS), which gives a residual standard deviation for each gene. The fitted log counts per million is converted into a predicted count and the variation is based on that count size. The weight for each

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

observation is the inverse of the predicted variance; conveniently, tools such as limma(Smyth, 2005) can then be used to test for differential expression.

Simulated Differential Expression with Varying Numbers of Replicates and Sequence Depth

We simulated replicates and a treatment and control group from the initial Tophat2 alignment of real RNA-seq data described in the Data Set section. Input to the simulator is a bam file and a table of counts for each gene. The tool htseq-count was used to assign reads to genes(Anders et al., 2014). Using real data for generating data sets preserves any technical biases in the data, such as random hexamer bias or base composition bias. Parameter settings for generating simulated data include number of replicates (default=3), proportion of genes to be differentially expressed (DE, default=0.1), minimum read count per gene for inclusion (default=20 reads), differential expression level for each DE gene (range 0-1, default 0.2-0.45), and overdispersion for each treatment-control pair (+/− 0.05). Data set generation involves five steps. First, all aligned reads from the initial experiment were combined and then for each gene, were split into equal-sized bins, with each bin serving as one treatment-control replicate pair. Second, we determine whether a particular gene will be DE by drawing from a uniform distribution; if the random number is less than the proportion to be DE, that gene is set as DE. Third, we determine the level of differential expression for each DE gene by drawing from a uniform distribution with minimum and maximum set by the differential expression level parameter. Fourth, we set the level of overdispersion for each treatment-control pair of each gene and add that to the level of DE. For example, if the level of DE is 0.3 and the overdispersion values for each of three replicates are 0, −0.02, and 0.05, then the total proportion of reads that go to each of the three treatments are 0.3, 0.28, and 0.35. Finally, we use these treatment-control pair values for each gene and for each read, we actually draw from a uniform distribution and would assign the read to treatment 1, replicate 1, if the value is less than or equal to that value (0.3 in the example). Read distribution for treatment-control pairs of non-DE genes is simply based on the overdispersion parameter. The output of the simulator is a table with a gene on each row and counts for each sample in columns. We conducted 3 different simulations, but all three have 24,913 genes, with 500 differentially expressed genes. We conducted two simulations with 6 replicate pairs with ~38 million reads per replicate and different settings for overdispersion +/− 0.05 and +/− 0.15. For subsampling tests to assess sequencing depth, we simulated 3 replicate pairs with 76 million reads per replicate with overdispersion +/− 0.05. Genes with fewer than 10 reads on average are excluded from analysis. We generated ROC curves showing true positive rate and false positive rate for each method and with different numbers of replicates and depth using the ROCR package in R (Fig. 8)(Sing et al., 2007). Results from VOOM and the Poisson-based test were post-processed to set p-value=1 when the average number of reads across all samples was <5 because both methods returned highly significant p-values for genes with small numbers of reads; other methods were not significantly affected by small read counts and were not post-processed.

We show the power to detect differential expression with overdispersion +/− 0.05 for varying levels of replicates (n=2-6) with 38 million fragments per sample (Fig. 8A-D). With 3 replicates and 38 million mapped fragments per sample, we recovered 94% of DE genes at

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

FPR 0.1 with EdgeR and just slightly fewer with both versions of DESeq (Fig. 8B). Increasing the number of replicates to n=4, marginally improves identification of DE genes (Fig. 8C). Decreasing the number of reads to 19 million per sample reduces the power to detect differential expression slightly for both edgeR and DESeq (Fig. 8F). Further reductions in read depth begin to have dramatic effects on the ability to detect differential expression (Fig. 8G,H), and this would be especially difficult with lowly expressed genes. Increasing the overdispersion to +/− 0.15 (results not shown) decreases the ability to detect differential expression, as expected. If biological variance is high, more replicates are needed; with n=6 replicates per group and high overdispersion (+/− 0.15), edgeR, both versions of DESeq, and VOOM recover about 85-86% of DE genes at FPR 0.1. Soneson and Delorenzi (2013) conducted thorough simulations that tested biases when there are large numbers of differentially expressed genes or in cases where one experimental group has a bias in the direction of expression patterns. Their results show that edgeR and DESeq perform quite well except when there are very large numbers of differentially expressed genes and they are highly biased in the direction of expression (i.e., one group always shows higher expression). Given the results from our simulations and others, we recommend 3 or more replicates per group, 30+ million mapped fragments per sample, and using edgeR (or DESeq) to detect differential expression on a gene-level with samples from inbred lines or cell culture.

SUMMARY RECOMMENDATIONS

RNA-seq experiments may take many different forms with a variety of sample types. Herein, we address common steps of designing experiments and analyzing data when the goal of the experiment is to detect differential expression. We make a number of recommendations based on these analyses (Fig. 9). Our first recommendation is to collect at least 3 biological replicates for each experimental group. This recommendation is based on analysis of differentiated embryonic stem cells that have minimal genetic and environmental variation and should be valid for inbred lines of mice or other cell lines. When analyzing samples with high levels of biological variation, such as clinical or post-mortem samples, it may be necessary to have many more replicates. Tools have been developed to help estimate the number of replicates needed based on the level of biological and technical variation(Bashir et al., 2010; Busby et al., 2013; Fang and Cui, 2011; Hart et al., 2013). For gene-based analysis, we recommend 30+ million mapped fragments. Surprisingly, 50 bp single-end reads map with high specificity. However, if the experiment relies on detecting alternative splicing, or chimeric transcripts, paired-end sequencing is recommended. Once sequence data are collected, we recommend thorough QC. Our fastQC aggregator may be of utility for comparing quality across samples (http://github.com/staciaw/fastqc_aggregator). We also recommend trimming adapter sequences before alignment and assessing the level of read duplication before and after alignment.

Many available tools exist for each step of the analysis. In our hands STAR and Tophat2 performed best for splice-aware alignment; we slightly prefer STAR because it is extremely fast and slightly more sensitive. Many tools exist for estimating transcript abundance and we tested a handful of them. Others that we did not test include IsoEM(Nicolae et al., 2011) and FDM(Singh et al., 2011). The estimates of transcript abundance for each method tested were

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

correlated between methods. However, there was significant variation in estimates compared to our ground truth data set. We recommend viewing the data within a genome browser following estimates of transcript abundance. Finally, for detecting differential expression, we found that edgeR (Robinson et al., 2009) performed best, followed closely by DESeq(Anders and Huber, 2010).

Acknowledgments

We thank Katie Pollard for insightful comments and continual support during this project, Nathan Salomonis for providing the script to convert transcript coordinates to genomic locations, Swetansu Hota and Benoit Bruneau for providing early access to their data, the Gladstone Genomics Core for high quality library preps, and David Nix, Jon Seidman, Jonathan Hill, Dennis Kostka, and Charlene Schramm for discussions about important facets of RNA-seq analysis. This study was funded by the Bench to Bassinet Program (B2B) award number U01-HL098188. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

LITERATURE CITED

Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biology. 2010; 11:R106. [PubMed: 20979621]

Anders, S.; Huber, W. Differential expression of RNA-Seq data at the gene level–the DESeq package. 2012. http://watson.nci.nih.gov/bioc_mirror/packages/2.11/bioc/vignettes/DESeq/inst/doc/DESeq.pdf

Anders S, McCarthy DJ, Chen Y, Okoniewski M, Smyth GK, Huber W, Robinson MD. Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. 2013 arXiv: 13023685.

Anders, S.; Pyl, PT.; Huber, W. HTSeq–A Python framework to work with high-throughput sequencing data. bioRxiv. 2014. doi: http://dx.doi.org/10.1101/002824

Aschoff M, Hotz-Wagenblatt A, Glatting KH, Fischer M, Eils R, Konig R. SplicingCompass: differential splicing detection using RNA-Seq data. Bioinformatics. 2013; 29:1141–1148. [PubMed: 23449093]

Au KF, Jiang H, Lin L, Xing Y, Wong WH. Detection of splice junctions from paired-end RNA-seq data by SpliceMap. Nucleic Acids Research. 2010; 38:4570–4578. [PubMed: 20371516]

Auer PL, Doerge RW. Statistical Design and Analysis of RNA Sequencing Data. Genetics. 2010; 185:405–416. [PubMed: 20439781]

Baggerly KA, Deng L, Morris JS, Aldaz CM. Overdispersed logistic regression for SAGE: modelling multiple groups and covariates. BMC Bioinformatics. 2004; 5:144. [PubMed: 15469612]

Bashir A, Bansal V, Bafna V. Designing deep sequencing experiments: detecting structural variation and estimating transcript abundance. BMC genomics. 2010; 11:385. [PubMed: 20565853]

Bullard JH, Purdom E, Hansen KD, Dudoit S. Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics. 2010; 11:94. [PubMed: 20167110]

Busby MA, Stewart C, Miller CA, Grzeda KR, Marth GT. Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression. Bioinformatics. 2013; 29:656–657. [PubMed: 23314327]

Chung LM, Ferguson JP, Zheng W, Qian F, Bruno V, Montgomery RR, Zhao H. Differential expression analysis for paired RNA-seq data. BMC Bioinformatics. 2013; 14:110. [PubMed: 23530607]

Cox DR, Reid N. Parameter Orthogonality and Approximate Conditional Inference. Journal of the Royal Statistical Society. Series B (Methodological). 1987; 49:1–39.

Dillies M-A, Rau A, Aubert J, Hennequet-Antier C, Jeanmougin M, Servant N, Keime C, Marot G, Castel D, Estelle J, et al. A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Briefings in Bioinformatics. 2013; 14:671–683. [PubMed: 22988256]

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013; 29:15–21. [PubMed: 23104886]

Emig D, Salomonis N, Baumbach J, Lengauer T, Conklin BR, Albrecht M. AltAnalyze and DomainGraph: analyzing and visualizing exon expression data. Nucleic Acids Research. 2010; 38:W755–W762. [PubMed: 20513647]

Fang Z, Cui X. Design and validation issues in RNA-seq experiments. Briefings in Bioinformatics. 2011; 12:280–287. [PubMed: 21498551]

Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, et al. Ensembl 2014. Nucleic Acids Research. 2014; 42:D749–55. [PubMed: 24316576]

Grant GR, Farkas MH, Pizarro A, Lahens N, Schug J, Brunk B, Stoeckert CJ, Hogenesch JB, Pierce EA. Comparative Analysis of RNA-Seq Alignment Algorithms and the RNA-Seq Unified Mapper (RUM). Bioinformatics. 2011; 27:2518–2528. [PubMed: 21775302]

Hansen KD, Irizarry RA, Wu Z. Removing technical variability in RNA-seq data using conditional quantile normalization. Biostatistics. 2012; 13:204–216. [PubMed: 22285995]

Hart SN, Therneau TM, Zhang Y, Poland GA, Kocher J-P. Calculating Sample Size Estimates for RNA Sequencing Data. Journal of computational biology: a journal of computational molecular cell biology. 2013; 20:970–978. [PubMed: 23961961]

Heap GA, Yang JHM, Downes K, Healy BC, Hunt KA, Bockett N, Franke L, Dubois PC, Mein CA, Dobson RJ, et al. Genome-wide analysis of allelic expression imbalance in human primary cells by high-throughput transcriptome resequencing. Human Molecular Genetics. 2010; 19:122–134. [PubMed: 19825846]

Hooper JE. A survey of software for genome-wide discovery of differential splicing in RNA-Seq data. Human Genomics. 2014; 8:3. [PubMed: 24447644]

Hu Y, Huang Y, Du Y, Orellana CF, Singh D, Johnson AR, Monroy A, Kuan P-F, Hammond SM, Makowski L, et al. DiffSplice: the genome-wide detection of differential splicing events with RNA-seq. Nucleic Acids Research. 2013; 41:e39. [PubMed: 23155066]

Ingolia NT, Ghaemmaghami S, Newman JRS, Weissman JS. Genome-Wide Analysis in Vivo of Translation with Nucleotide Resolution Using Ribosome Profiling. Science. 2009; 324:218–223. [PubMed: 19213877]

Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. Nature Methods. 2010; 7:1009–1015. [PubMed: 21057496]

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biology. 2013; 14:R36. [PubMed: 23618408]

Kvam VM, Liu P, Si Y. A comparison of statistical methods for detecting differentially expressed genes from RNA-seq data. American Journal of Botany. 2012; 99:248–256. [PubMed: 22268221]

Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature Methods. 2012; 9:357–359. [PubMed: 22388286]

Law CW, Chen Y, Shi W, Smyth GK. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology. 2014; 15:R29. [PubMed: 24485249]

Li B, Dewey CN. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics. 2011; 12:323. [PubMed: 21816040]

Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics. 2010; 26:589–595. [PubMed: 20080505]

Li Y, Li-Byarlay H, Burns P, Borodovsky M, Robinson GE, Ma J. TrueSight: a new algorithm for splice junction detection using RNA-seq. Nucleic Acids Research. 2013; 41:e51. [PubMed: 23254332]

Lovén J, Orlando DA, Sigova AA, Lin CY, Rahl PB, Burge CB, Levens DL, Lee TI, Young RA. Revisiting global gene expression analysis. Cell. 2012; 151:476–482. [PubMed: 23101621]

Lu J, Tomfohr JK, Kepler TB. Identifying differential expression in multiple SAGE libraries: an overdispersed log-linear model approach. BMC Bioinformatics. 2005; 6:165. [PubMed: 15987513]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y. RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome Research. 2008; 18:1509–1517. [PubMed: 18550803]

McCarthy DJ, Chen Y, Smyth GK. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Research. 2012; 40:4288–4297. [PubMed: 22287627]

McIntyre LM, Lopiano KK, Morse AM, Amin V, Oberg AL, Young LJ, Nuzhdin SV. RNA-seq: technical variability and sampling. BMC genomics. 2011; 12:293. [PubMed: 21645359]

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, et al. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research. 2010; 20:1297–1303. [PubMed: 20644199]

Nariai N, Hirose O, Kojima K, Nagasaki M. TIGAR: transcript isoform abundance estimation method with gapped alignment of RNA-Seq data by variational Bayesian inference. Bioinformatics. 2013; 29:2292–2299. [PubMed: 23821651]

Nicolae M, Mangul S, Măndoiu II, Zelikovsky A. Estimation of alternative splicing isoform frequencies from RNA-Seq data. Algorithms for molecular biology: AMB. 2011; 6:9. [PubMed: 21504602]

Nookaew I, Papini M, Pornputtapong N, Scalcinati G, Fagerberg L, Uhlén M, Nielsen J. A comprehensive comparison of RNA-Seq-based transcriptome analysis from reads to differential gene expression and cross-comparison with microarrays: a case study in Saccharomyces cerevisiae. Nucleic Acids Research. 2012; 40:10084–10097. [PubMed: 22965124]

Richard H, Schulz MH, Sultan M, Nurnberger A, Schrinner S, Balzereit D, Dagand E, Rasche A, Lehrach H, Vingron M, et al. Prediction of alternative isoforms from exon expression levels in RNA-Seq experiments. Nucleic Acids Research. 2010; 38:e112–e112. [PubMed: 20150413]

Risso, D. EDASeq: Exploratory Data Analysis and Normalization for RNA-Seq. 2013. http://140.107.3.20/packages/release/bioc/vignettes/EDASeq/inst/doc/EDASeq.pdf

Roberts A, Pimentel H, Trapnell C, Pachter L. Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics. 2011; 27:2325–2329. [PubMed: 21697122]

Robinson MD, Oshlack A. A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biology. 2010; 11:R25. [PubMed: 20196867]

Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2009; 26:139–140. [PubMed: 19910308]

Robles JA, Qureshi SE, Stephen SJ, Wilson SR, Burden CJ, Taylor JM. Efficient experimental design and analysis strategies for the detection of differential expression using RNA-Sequencing. BMC genomics. 2012; 13:484. [PubMed: 22985019]

Sacomoto GAT, Kielbassa J, Chikhi R, Uricaru R, Antoniou P, Sagot M-F, Peterlongo P, Lacroix V. KISSPLICE: de-novo calling alternative splicing events from RNA-seq data. BMC Bioinformatics. 2012; 13:55. [PubMed: 22480135]

Sakarya O, Breu H, Radovich M, Chen Y, Wang YN, Barbacioru C, Utiramerur S, Whitley PP, Brockman JP, Vatta P, et al. RNA-Seq Mapping and Detection of Gene Fusions with a Suffix Array Algorithm. PLoS Computational Biology. 2012; 8:e1002464. [PubMed: 22496636]

Sing T, Sander O, Beerenwinkel N, Lengauer T. ROCR: visualizing classifier performance in R. Bioinformatics. 2005; 21:3940–3941. [PubMed: 16096348]

Singh D, Orellana CF, Hu Y, Jones CD, Liu Y, Chiang DY, Liu J, Prins JF. FDM: a graph-based statistical method to detect differential transcription using RNA-seq data. Bioinformatics. 2011; 27:2633–2640. [PubMed: 21824971]

Skelly DA, Johansson M, Madeoy J, Wakefield J, Akey JM. A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from RNA-seq data. Genome Research. 2011; 21:1728–1737. [PubMed: 21873452]

Smyth, GK. limma: Linear Models for Microarray Data. *In* Bioinformatics and Computational Biology Solutions Using R and Bioconductor. Springer-Verlag; New York: 2005. p. 397-420.

Soneson C, Delorenzi M. A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics. 2013; 14:91. [PubMed: 23497356]

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009; 25:1105–1111. [PubMed: 19289445]

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

Vijay N, Poelstra JW, Kuenstner A, Wolf JBW. Challenges and strategies in transcriptome assembly and differential gene expression quantification. A comprehensive in silico assessment of RNA-seq experiments. Molecular Ecology. 2013; 22:620–634. [PubMed: 22998089]

Wang K, Singh D, Zeng Z, Coleman SJ, Huang Y, Savich GL, He X, Mieczkowski P, Grimm SA, Perou CM, et al. MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Research. 2010; 38:e178. [PubMed: 20802226]

Wang L, Feng Z, Wang X, Wang X, Zhang X. DEGseq: an R package for identifying differentially expressed genes from RNA-seq data. Bioinformatics. 2009; 26:136–138. [PubMed: 19855105]

Wang L, Wang S, Li W. RSeQC: quality control of RNA-seq experiments. Bioinformatics. 2012; 28:2184–2185. [PubMed: 22743226]

Wu H, Wang C, Wu Z. A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. Biostatistics. 2013; 14:232–243. [PubMed: 23001152]

Young, MD.; McCarthy, DJ.; Wakefield, MJ.; Smyth, GK.; Oshlack, A.; Robinson, MD. Differential Expression for RNA Sequencing (RNA-Seq) Data: Mapping, Summarization, Statistical Analysis, and Experimental Design. Springer New York; New York, NY: 2011. p. 169-190.

Łabaj PP, Leparc GG, Linggi BE, Markillie LM, Wiley HS, Kreil DP. Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling. Bioinformatics. 2011; 27:i383–i391. [PubMed: 21685096]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

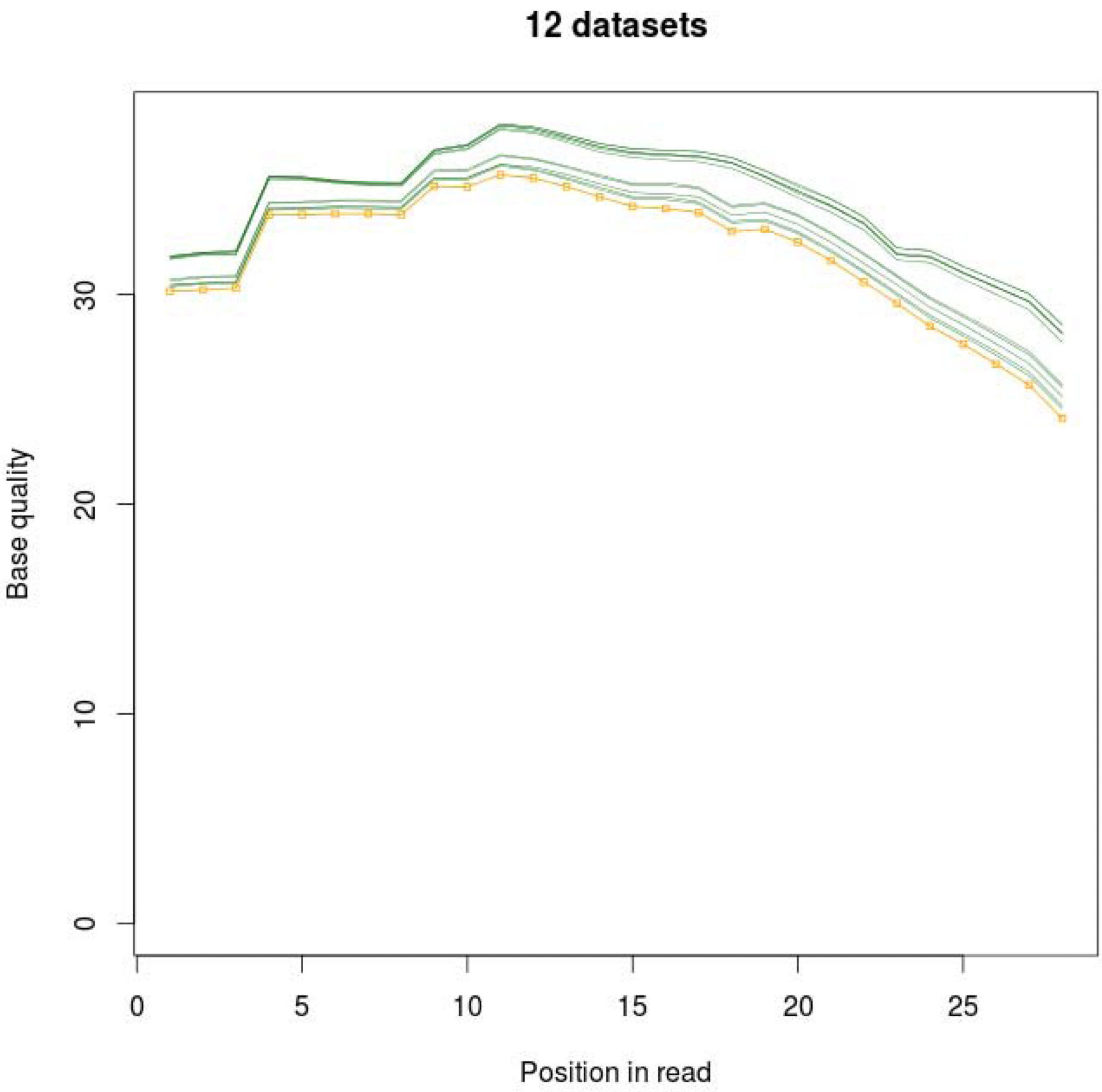


Figure 1.

Example plot of one QC metric from FastQC, per base sequence quality, aggregated over samples into a single figure. Plotted line will be orange to indicate warning if the lower quartile for any base is less than 10, or if the median for any base is less than 25. Plotted line will be red to indicate failure if the lower quartile for any base is less than 5 or if the median for any base is less than 20. All aggregated plots can be viewed as separate images and are also combined into an html document.

NIH-PA Author Manuscript NIH-PA Author Manuscript NIH-PA Author Manuscript

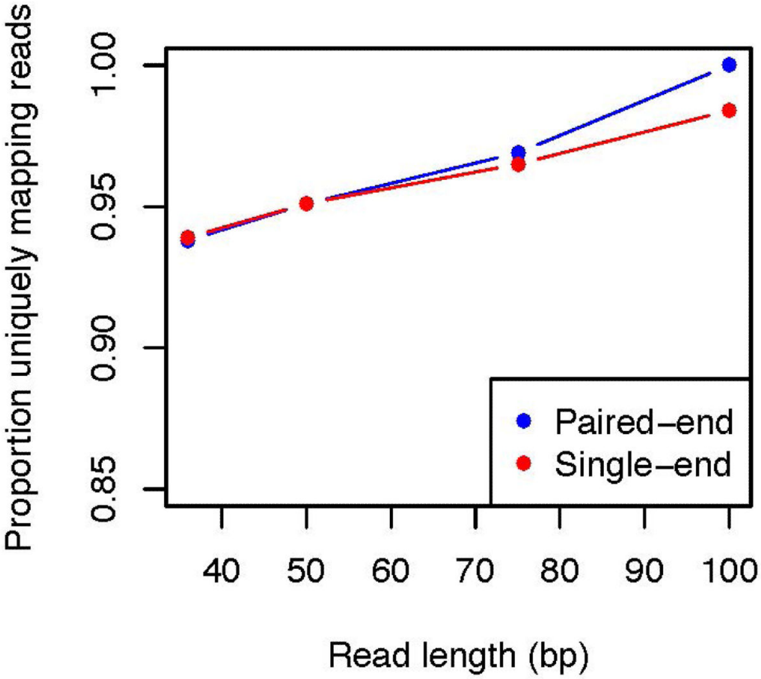


Figure 2.

Proportion of reads mapping uniquely from PE and SE alignments relative to 100 bp PE results using Tophat2.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

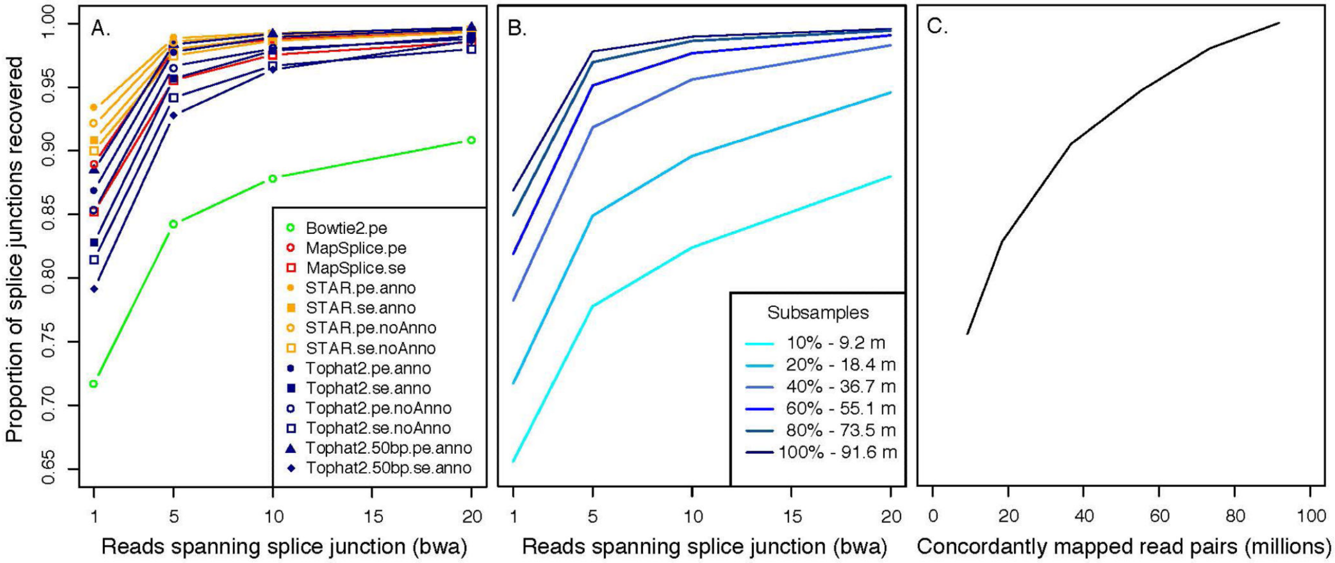


Figure 3.

Splice junction mapping with Tophat2, MapSplice, and STAR. Proportion of junctions recovered with splice-aware mappers based on a minimum number of fragments spanning the splice junction from the BWA gold set alignment (A). Filled symbols indicate an annotation was provided whereas open symbols indicate no annotation was provided to the aligner. Results from Bowtie2 in green, MapSplice in red, STAR in dark yellow, and Tophat2 in blue. Proportion of junctions recovered from subsampled data sets using Tophat2 based on a minimum number of reads spanning splice junction in BWA alignment (B). Proportion of junctions recovered from subsampled data sets relative to the full data set (91.8 million read pairs) using Tophat2 provided with annotation (C).

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

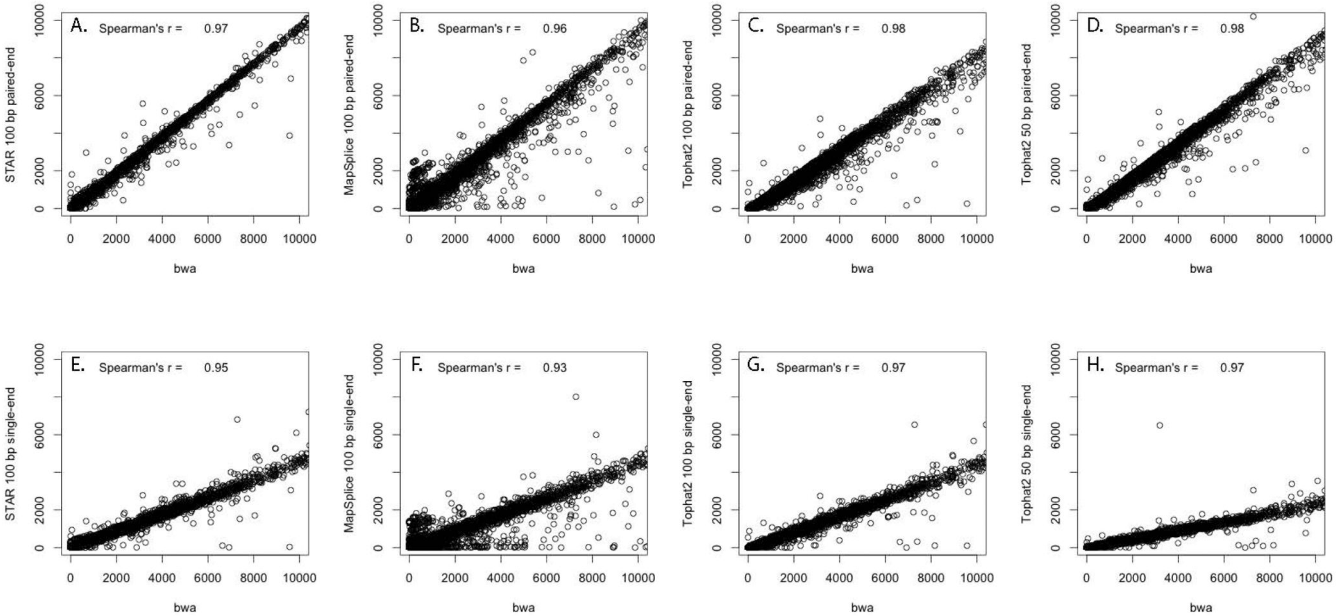


Figure 4.

Correlation between counts of fragments spanning each splice junction for BWA compared to splice-aware aligners. PE (A-D) and SE (E-H) analyses using STAR (A,E), MapSplice (B,F) and Tophat2 (C,D,G,H). Panels D and H are for 50 bp reads; all other panels are 100 bp reads.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

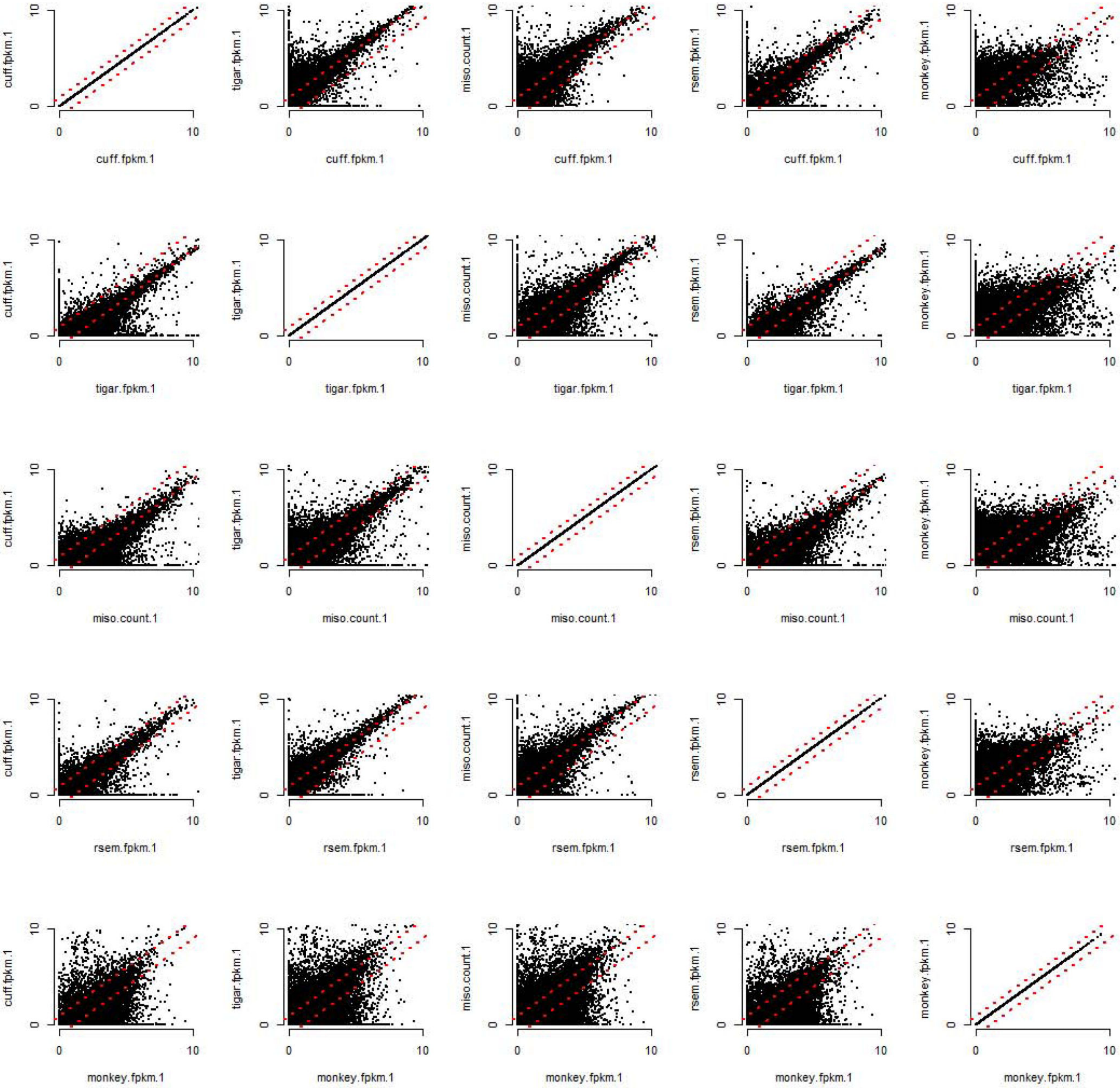
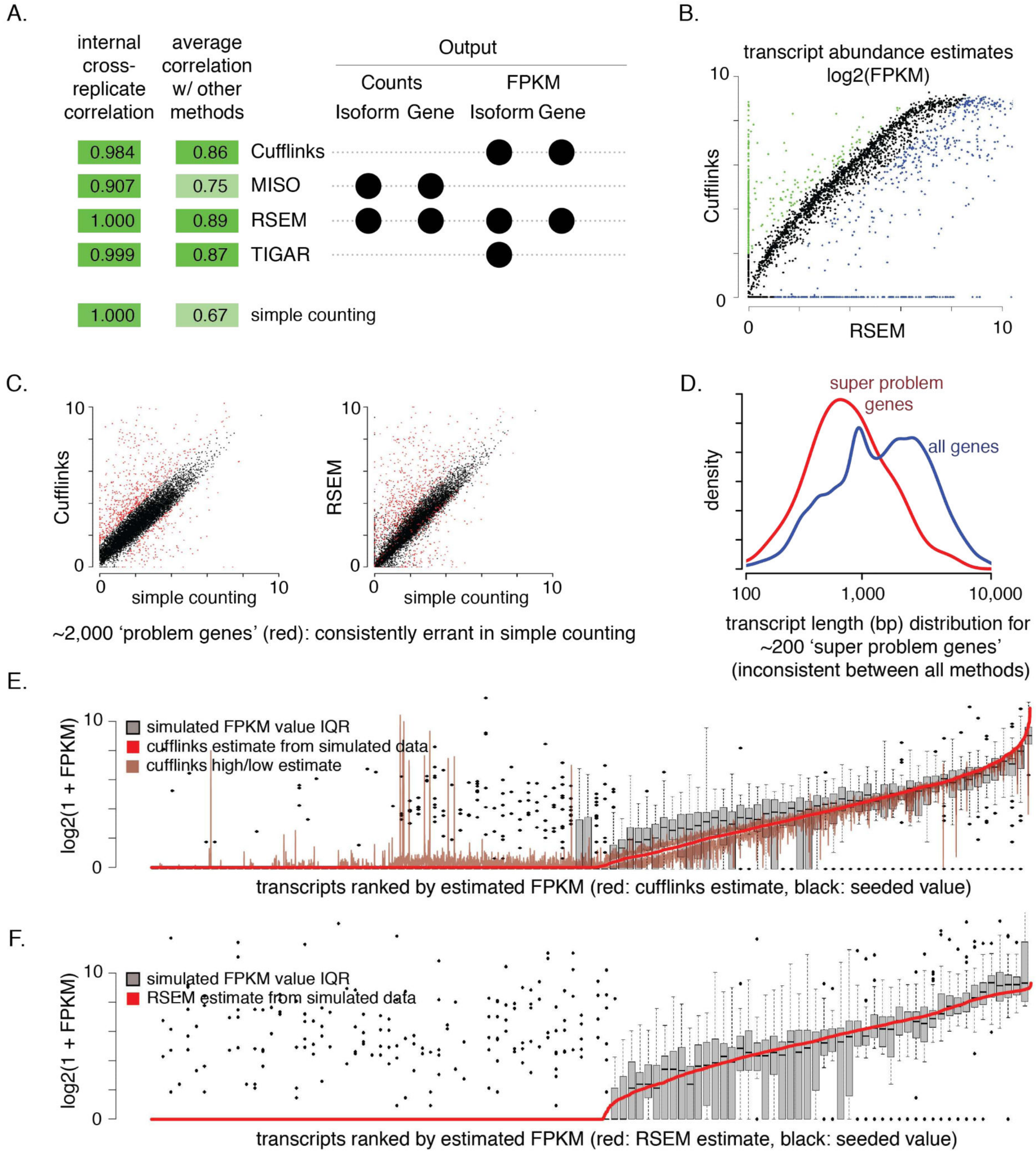


Figure 5.

Scatterplots comparing WT cardiomyocytes replicate 1 and replicate 2 of real RNA-seq data for each transcript abundance tool.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript



internal cross-replicate correlation	average correlation w/ other methods		Output			
			Counts		FPKM	
			Isoform	Gene	Isoform	Gene
0.984	0.86	Cufflinks			●	●
0.907	0.75	MISO	●	●		
1.000	0.89	RSEM	●	●	●	●
0.999	0.87	TIGAR			●	
1.000	0.67	simple counting				

Figure 6.

Comparison of transcript abundance estimating methods. General properties of RNA abundance estimating methods (A). The average correlation between each method and other methods’ abundance estimates of real data are indicated (green=high Spearman correlation) at the far left, followed by a description of the outputs generated by each tool. On the right is indicated the correlation of the abundance estimates with gene length for each method. High correlation between most methods (B). The estimated log2 FPKM values of transcript abundance are shown for Cufflinks and RSEM. Blue and green points indicate genes with

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

large difference in FPKM estimates given simulated data. Most methods are more precise than a simple counting method (C). Shown for Cufflinks and RSEM are comparisons of abundance estimates from each method with a naïve abundance estimate that does no correction for any of the issues that pertain to RNA-seq data including the possibility that a given tag may map to multiple different transcripts. Black points indicate high correlation between Cufflinks and this simple counting technique, red points indicate the ~2,000 genes that exhibit large differences between Cufflinks and this method. In the other comparisons (e.g. RSEM versus simple counting) those red points remain outliers, suggesting that these genes are problematic for simple counting, but whose estimates are more precisely estimated using any of the specified tools, given the much higher correlation in (B). Problematic genes are shorter than average genes (D). The intersection of genes with low precision across a round robin of estimates was labeled as the set of ‘super problem genes’ (e.g., red points in (B)). These ~200 genes had a different distribution of length (red) than the average transcript length of all genes (blue). Methods are accurate on average, but with higher variance than estimates might indicate (E,F). Differences between simulated data with known FPKM values feeding the simulation (black) were compared to estimated FPKMs using Cufflinks (E) and RSEM (F). Transcripts were first ranked by estimated FPKM (red), then binned into groups of 100 genes. Interquartile ranges and outliers for the ‘truth’ simulation (black/grey) are shown for each bin. Both methods work well on average, but many genes have incorrect estimates.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

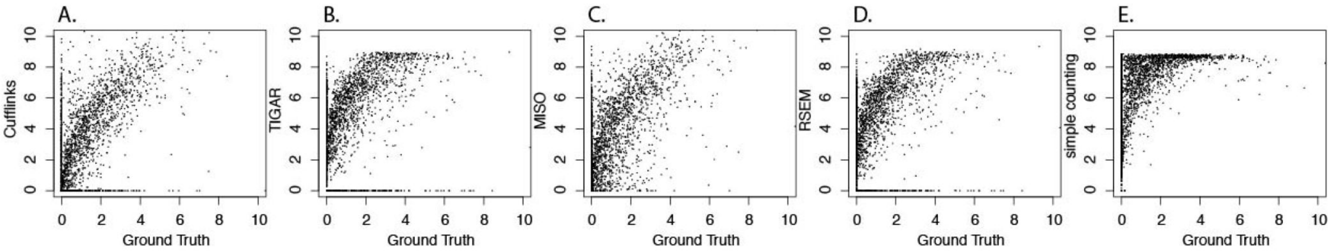


Figure 7.

Scatterplots comparing estimated transcript abundance compared to ground truth for each tool. Log2 FPKM values for Cufflinks (A), TIGAR (B), MISO (C), RSEM (D), simple counting (E). Each tool is correlated with ground truth, but there is a considerable amount of scatter.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

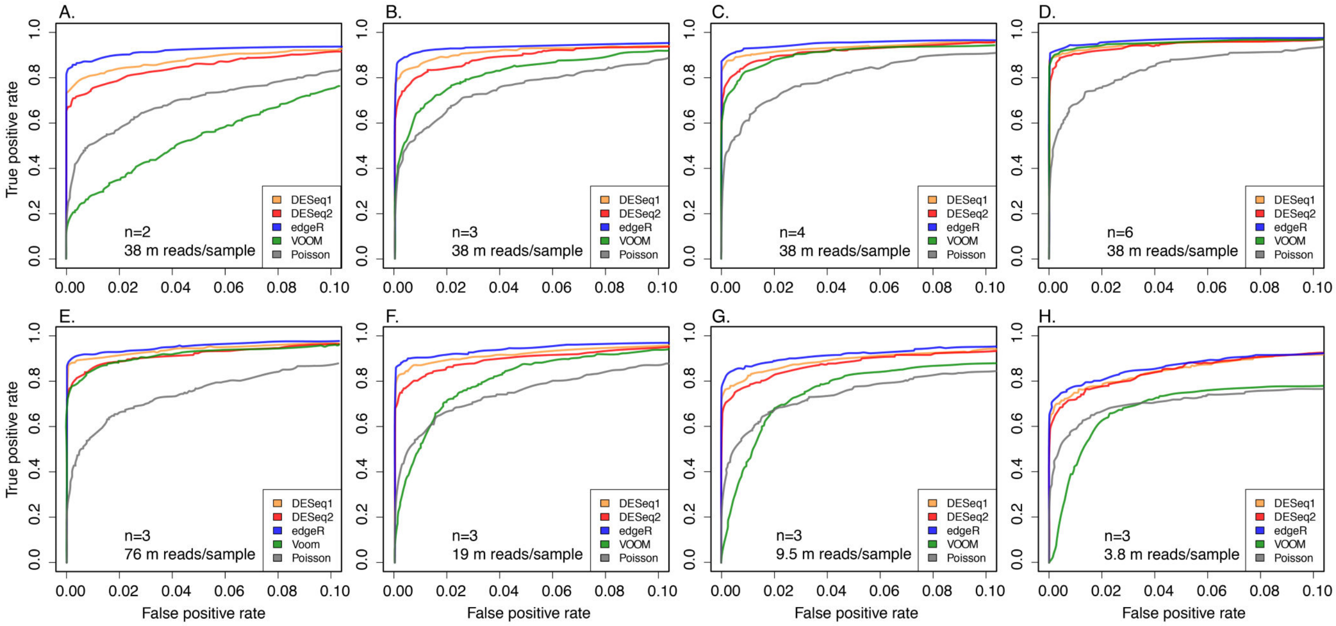


Figure 8.

Assessment of power with various sequencing depth and number of replicates for tools that test for differential expression. Comparison of sensitivity with varying numbers of replicates (n=2-6) and 38 million mapped reads per sample (A-D). DESeq and EdgeR perform well with 3 or more replicates. VOOM makes dramatic improvements with increasing sample size. Comparison of sensitivity with differing sequencing depths: 76, 38, 19, 9.5, 3.8 million reads per sample (E,B,F,G,H, respectively). Each group has n=3 replicates for testing depth. Power to detect differentially expressed genes drops off considerably with only 19 million reads per sample. Note that the false positive rate on the x-axis of each figure ranges from 0 – 0.10; this zoomed-in view shows subtle differences between methods, number of replicates, or depth of sequencing.

NIH-PA Author Manuscript
NIH-PA Author Manuscript
NIH-PA Author Manuscript

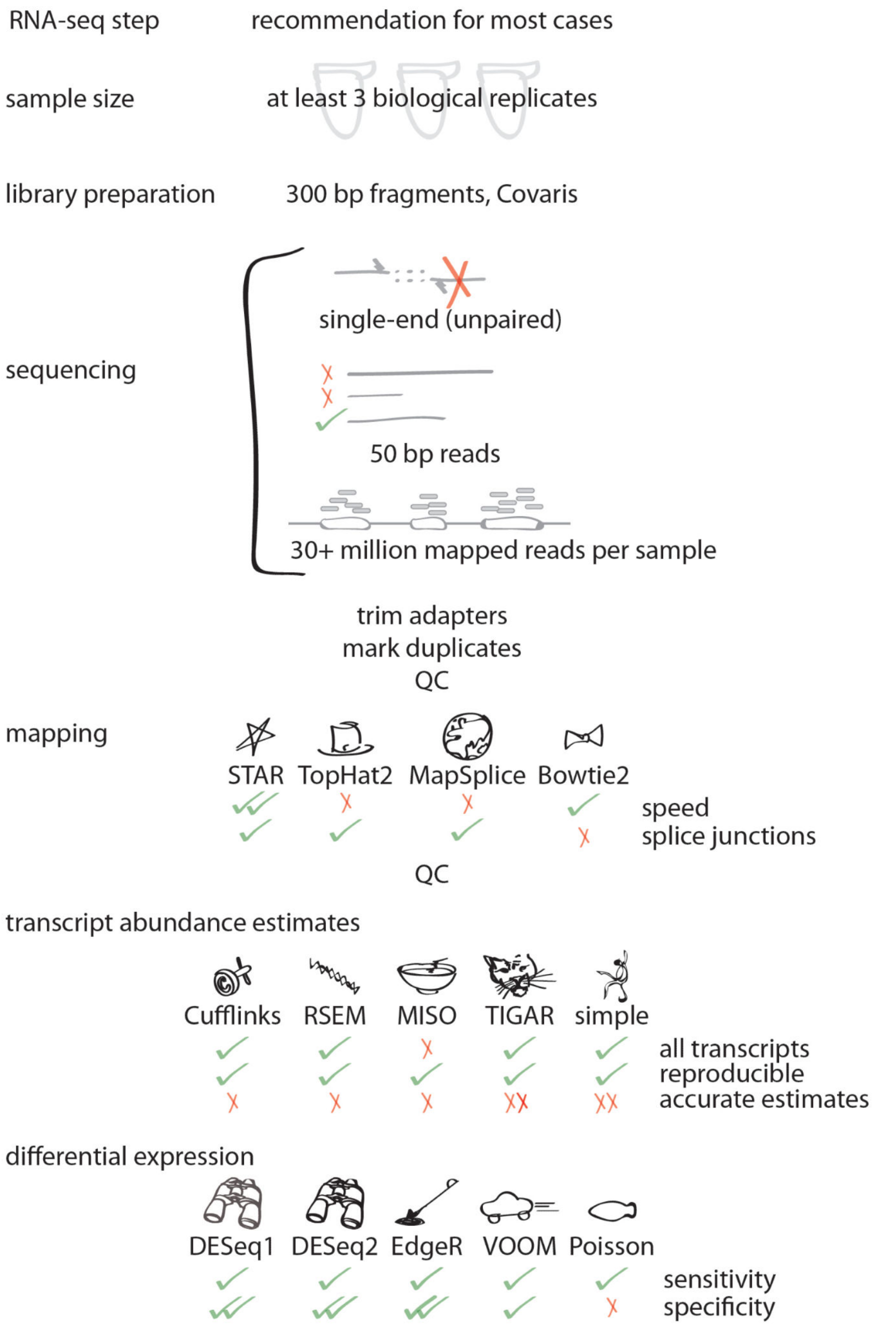


Figure 9.

Workflow and recommendations for RNA-seq analysis. We recommend 3 replicates per group when biological variance is expected to be low (e.g., for inbred lines of mice or other cell lines) and using tools to estimate sample size when biological and/or technical variance is expected to be higher (e.g., with clinical or post-mortem samples). For gene-based analysis, we recommend 30+ million mapped fragments with 50 bp single-end reads However, if the experiment relies on detecting alternative splicing, or chimeric transcripts, much deeper 50 bp paired-end sequencing is recommended. Once sequence data are

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

collected, we recommend QC with FastQC and our aggregator for visualization, trimming adapter sequences with fastx trimmer (http://hannonlab.cshl.edu/fastx_toolkit/) before alignment. We also recommend QC and marking duplicates following alignment. We recommend STAR for splice-aware alignment, and carefully viewing the data within a genome browser following estimates of transcript abundance. Finally, for detecting differential expression, we recommend edgeR or DESeq. The green check marks and red Xs are meant to give a qualitative assessment of tool capabilities.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 1

Proportion of Junctions Recovered Based on Number of Fragments Spanning Splice Junction in BWA Alignments

				Prop. SJS Relative to	Percent Splice Junctions Recovered*				
SJM	Length	SE/PE	Annotation	BWA	1+	5+	10+	20+	50+
Bowtie2	100	PE	No	0.10	71.7%	84.2%	87.9%	90.8%	93.1%
Mapsplice	100	PE	No	0.89	89.0%	98.0%	98.8%	99.3%	99.7%
Mapsplice	100	SE	No	0.43	85.2%	95.6%	97.6%	98.5%	99.2%
STAR	100	PE	Yes	0.94	93.2%	98.9%	99.3%	99.6%	99.9%
STAR	100	SE	Yes	0.44	90.8%	98.0%	98.9%	99.5%	99.8%
STAR	100	PE	No	0.83	92.0%	98.6%	99.1%	99.5%	99.8%
STAR	100	SE	No	0.35	90.0%	97.5%	98.7%	99.3%	99.7%
Tophat2	100	PE	Yes	0.73	86.9%	97.8%	98.9%	99.6%	99.9%
Tophat2	100	SE	Yes	0.41	82.8%	95.7%	97.9%	99.0%	99.7%
Tophat2	100	PE	No	0.64	85.3%	96.7%	98.0%	98.8%	99.5%
Tophat2	100	SE	No	0.37	81.5%	94.2%	96.7%	98.0%	98.9%
Tophat2	50	PE	Yes	0.82	88.5%	98.4%	99.2%	99.7%	99.9%
Tophat2	50	SE	Yes	0.22	79.1%	92.8%	96.4%	98.7%	99.7%

*Binned by minimum number of fragments spanning junction in BWA alignment.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 2

Transcript abundance estimates for each of two replicates were calculated using Cufflinks, TIGAR, MISO, RSEM, and a simple counting method.

	Cufflinks.1	TIGAR.1	MISO.1	RSEM.1	count.1
Cufflinks.1	1.00	0.96	0.84	0.96	0.70
Cufflinks.2	0.99	0.96	0.80	0.95	0.74
TIGAR.1	0.96	1.00	0.81	0.98	0.74
TIGAR.2	0.93	0.99	0.77	0.96	0.76
MISO.1	0.84	0.81	1.00	0.87	0.45
MISO.2	0.85	0.83	0.97	0.89	0.51
RSEM.1	0.96	0.98	0.87	1.00	0.70
RSEM.2	0.94	0.98	0.84	0.99	0.74
count.1	0.70	0.74	0.45	0.70	1.00
count.2	0.69	0.73	0.43	0.69	1.00

Correlation coefficients are Spearman’s Rho.