BIOINFORMATICS

Vol. 27 ISMB 2011, pages i383–i391
doi:10.1093/bioinformatics/btr247

Characterization and improvement of RNA-Seq precision in quantitative transcript expression profiling

Paweł P. Łabaj[1], Germán G. Leparc[1,†], Bryan E. Linggi[2], Lye Meng Markillie[2], H. Steven Wiley[2] and David P. Kreil[1,*]

[1]Chair of Bioinformatics, Boku University Vienna, 1190 Muthgasse 18, Vienna, Austria and [2]Environmental Molecular Sciences Laboratory, Pacific Northwest National Laboratory, 902 Battelle Boulevard, Richland, Washington, USA

ABSTRACT

Motivation: Measurement precision determines the power of any analysis to reliably identify significant signals, such as in screens for differential expression, independent of whether the experimental design incorporates replicates or not. With the compilation of large-scale RNA-Seq datasets with technical replicate samples, however, we can now, for the first time, perform a systematic analysis of the precision of expression level estimates from massively parallel sequencing technology. This then allows considerations for its improvement by computational or experimental means.

Results: We report on a comprehensive study of target identification and measurement precision, including their dependence on transcript expression levels, read depth and other parameters. In particular, an impressive recall of 84% of the estimated true transcript population could be achieved with 331 million 50 bp reads, with diminishing returns from longer read lengths and even less gains from increased sequencing depths. Most of the measurement power (75%) is spent on only 7% of the known transcriptome, however, making less strongly expressed transcripts harder to measure. Consequently, <30% of all transcripts could be quantified reliably with a relative error <20%. Based on established tools, we then introduce a new approach for mapping and analysing sequencing reads that yields substantially improved performance in gene expression profiling, increasing the number of transcripts that can reliably be quantified to over 40%. Extrapolations to higher sequencing depths highlight the need for efficient complementary steps. In discussion we outline possible experimental and computational strategies for further improvements in quantification precision.

Contact: rnaseq10@boku.ac.at

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

RNA-Seq is a novel method for gene expression profiling by next-generation sequencing of transcripts. The technology has been applied to gain global views of the complex transcriptomes of mammalian samples, including human embryonic kidney and B-cells (Sultan *et al*., 2008), mouse embryonic stem cells (Cloonan *et al*., 2008), blastomeres (Tang *et al*., 2009), and different mouse tissues (Mortazavi *et al*., 2008). An advantage of RNA-Seq over other profiling technologies is that it allows a comprehensive assay of gene expression that is not reliant on probes for targets that must be specified in advance. It is particularly well suited for the *de novo* detection of splice junctions and allows genome-wide qualitative expression profiling of organisms with unknown genome sequence.

Transcript detection obviously benefits from the digital nature of counting sequence reads. The observed identification rate increases with additional sequencing but is partly determined by the non-random nature of biological sequences and the highly skewed distribution of transcript abundances. We can extrapolate an expected achievable identification rate from the observed dependency on experimental parameters like read depth and read length. In addition, we examine the effects of random read sampling on the identification of low-copy number transcripts, as resulting from the distribution of reads mapped to different spliceforms. With many transcription factors being biologically active in low-copy numbers, this is particularly topical for studies of gene regulation.

Increasingly, there has been an interest in applying RNA-Seq not only for qualitative transcriptome profiling but also for the quantification of gene expression (Blow, 2009; Jiang and Wong, 2009; Shendure, 2008; Trapnell *et al*., 2010; Wilhelm *et al*., 2008). Using raw read counts mapped to individual targets, however, can result in length-dependent bias (Oshlack and Wakefield, 2009, and see Supplementary Material). A common approach for the quantification of gene expression in an RNA-Seq experiment thus computes the number of reads per kilobase of exonic sequence per million mapped reads (RPKM) to produce a gene expression measure which overall correlates well with measurements from microarrays (Mortazavi *et al*., 2008). Such normalization is also necessary to allow the combination or the comparison of RNA-Seq runs.

While earlier work has focused on reads that unambiguously identify a transcript, current developments extend data analysis to complex gene models of alternative splicing, also taking into account the many reads that may come from different spliceforms (Griffith *et al*., 2010; Jiang and Wong, 2009; Lee *et al*., 2011; Mortazavi *et al*., 2008). A popular recently emerging approach is to align reads to the genome, and then use this information to assemble transcripts *de novo* and calculate their abundances, as implemented by the TopHat/Cufflinks tools (Trapnell *et al*., 2009, 2010).

Despite or perhaps even because of the fast pace of development of both the measurement technology and the associated novel analysis tools (Datta *et al*., 2010), the central question of measurement reliability or, of how precisely we can actually quantify transcript expression, has received relatively little attention beyond initial observations of overall good correlation (Marioni *et al*., 2008;

[*] To whom correspondence should be addressed.

[†] Present address: Institute of Molecular Pathology, 1030 Dr Bohr Gasse 7, Vienna, Austria.

Wilhelm *et al*., 2008). Simple correlation coefficients, however, can be misleading, as they are dominated by a small number of very highly expressed genes (see Supplementary Section S8 for discussion/examples). Despite the excellent overall correlation, reproducibility seems to be lower for gene classes that are less strongly expressed (Mortazavi *et al*., 2008). We will examine this in greater detail to show that, as a consequence, some transcripts can be assessed with extremely good precision, whereas a large number of transcripts are hard to measure reliably. It is therefore interesting to consider measurement precision for all targets individually.

Similar to the early microarray data, however, there has been a lack of large RNA-Seq datasets with the necessary technical replicates. Now a comprehensive analysis of the reproducibility of gene expression level measurements by RNA-Seq has become possible and constitutes a necessary complement to characterizations of systemic measurement bias in next-generation sequencing (Bullard *et al*., 2010). Measurement precision in particular determines the power of any analysis to reliably identify relevant signals or changes, such as in screens for differential expression, independent of whether replicates are employed or not (Anders and Huber, 2010). Mortazavi *et al*. (2008) compiled one of the first large RNA-Seq datasets with technical replicates (2×40 million reads per sample), reporting reduced precision for less strongly expressed transcripts. We here provide a systematic study of the reliability of expression level estimates from an extended dataset with technical replicates (3×331 million reads).

Based on our observations, we then introduce a hybrid approach in the analysis of sequencing reads, for which we can demonstrate substantially improved quantification performance.

2 METHODS AND DATA

2.1 Experiments

Cell culture: the human mammary epithelial cell (HMEC) line 184A1 was obtained from Martha Stampfer (Lawrence Berkeley National Laboratory) and maintained in DFCI-1 media as previously described (Band and Sager, 1989).

RNA isolation and processing: a total of 192 μg of RNA was isolated from 5×10^7 cells using the RNeasy kit (Qiagen), including DNase treatment, and followed by mRNA isolation in two rounds of Poly(A) Purist (Ambion), according to the manufacturers' protocols. RNA quantity and quality were measured using a Bioanalzyer chip (Agilent).

RNA-Seq: a sequencing library was created from 1.6 μg of mRNA using the SOLiD Whole Transcriptome Analysis Kit. Emulsion PCR was performed using SOLiD EZ bead kits. The resulting bead library was divided into three aliquots, loaded in separate flow cells, and sequenced for 50 bp on a SOLiD 3+ system (Applied Biosystems).

Microarrays: as above, RNA was isolated from 10^7 cells. For each array, 100 ng of total RNA was labelled for hybridization to a GeneChip Human Gene 1.0 ST Array (Affymetrix) according to the manufacturer's protocol.

2.2 Handling and characterization of read sequences

Annotation: while the *de novo* identification of splice junctions is a particular strength of RNA-Seq, comprehensive gene models or known full-length cDNA sequences are required to assess the extent to which RNA-Seq reads can identify individual spliceforms (Carninci *et al*., 2003). For an unbiased assessment of transcript identification, we focused on reads alignable to the human transcriptome as annotated by EnsEMBL (release 58, May 2010), which provides the most comprehensive collection of human gene transcripts currently available (Flicek *et al*., 2010). The collection of 140 079 transcripts is based on an automated annotation pipeline combined with manually curated transcripts from the Vega and CCDS projects (Pruitt *et al*., 2009; Wilming *et al*., 2008).

Alignment: while many alignment tools perform well and with high sensitivity (Homer *et al*., 2009; Li *et al*., 2008; Ning *et al*., 2001), the fast increase of generated read volumes has led to algorithms exploiting the Burrows–Wheeler transform, reducing runtime by two orders of magnitude and thus facilitating alignment of very deep read sets (e.g. Li and Durbin, 2009; Li *et al*., 2009). From those available for local installation, we apply the now well established Bowtie program (v0.12.7), giving a satisfactory tradeoff between sensitivity and speed (Langmead *et al*., 2009). Alignment of reads to the human genomic sequence (EnsEMBL r58) was performed by TopHat v1.1.4 (Trapnell *et al*., 2009). TopHat internally uses the Bowtie aligner. This also facilitates subsequent direct comparisons of alignments to genomic or spliceform sequences. Both programs ran with default settings and took advantage of the ABI SOLiD colour-space format for higher quality alignments. Finally, we introduce and test an approach combining alignment of reads to spliceform sequences by Bowtie with subsequent mapping to genomic locations according to EnsEMBL gene models.

Subsampling: investigating the effect of read depth, data were subsampled to between 10 000 and 240 million read alignments per replicate.

2.3 Assessing expression levels and reproducibility

Quantifying expression levels: expression levels from unique reads aligned by Bowtie were calculated as RPKM values (Mortazavi *et al*., 2008). Expression levels for the other, gene model based approaches were calculated using Cufflinks (Trapnell *et al*., 2010). As specified, Cufflinks was either run in *de novo* gene model discovery mode or it was provided the EnsEMBL gene models. Parameters were set for maximal sensitivity (`—min-frags-per-transfrag 1` and `-F 0`). When processing alignments to spliceform sequences option `-A 0` could be set because all parts of a read were known to match the same spliceform; this parameter is normally used to support reliable splice junction discovery through TopHat. For spliceforms supported by less than one read alignment as assigned by Cufflinks, expression levels were set to zero.

Measures of reproducibility: for a systematic assessment of reproducibility, we can either consider the coefficient of variation on the linear scale, or the standard deviation of log expression levels. For a number of reasons, gene expression data is typically analysed on a log scale, on which differences in expression are probed by a *t*-test. Differences on the log-scale then correspond to a fold-change on the linear scale. In this context, the appropriate measure of precision is the standard deviation on the log scale. When referring to a relative error of 20% or less in the manuscript, we threshold the standard deviation $\sigma < \log_2(120\%)$ so that a value $\mu+\sigma$ compared to μ on the $\log_2$-scale corresponds to a relative error of 20% or less on the linear scale. The 20% 'benchmark' value for the relative error was chosen arbitrarily for ease of discussion, whereas results for other values are shown in the figures.

Note that reproducibility also determines the power of statistics that operate without replicates and are instead based on theoretical properties of count distributions, or which estimate a variance–mean relationship from non-replicate samples (Anders and Huber, 2010).

Finally, many analyses considering replicate precision exclude measurements with no signal in any one of the replicates. This creates a methods bias towards a better perceived precision in the assay. In the examined dataset, 14% of all identified transcript targets had zero reads in one or two of the replicates but non-zero counts in the others, with 1–26 reads observed. These transcripts substantially contribute to the measurement noise at low expression levels, and consequently have to be counted towards the fraction of unreliable measurements. That approach is consistent because

Table 1. Statistics of reads, mapping and alignment

Replicate	Reads	Bowtie (transcriptome)			TopHat (genome)			Bowtie (combined)		
		Mapped reads (%)	Align's	Unique reads (%)	Mapped reads (%)	Align's	Junct	Mapped reads (%)	Align's	Junct
1	340	168 (50)	772	36 (11)	172 (51)	241	18	168 (50)	249	45
2	341	167 (49)	776	35 (10)	170 (50)	238	17	167 (49)	247	45
3	311	152 (49)	699	32 (10)	155 (50)	217	16	152 (49)	225	41
Total	993	487 (49)	2237	103 (10)	497 (50)	695	51	487 (49)	721	131
Quantification by				unique reads			alignments			alignments

Each row shows results for one of the three technical replicate samples. Sums are displayed at the bottom of the table. All counts are given in millions, percentages are relative to the number of reads collected. The first group of columns assesses aligning reads to all the known transcript sequences with Bowtie. Every mapped read can have multiple alignments. It is, however, the unique reads, which unambiguously map to a single spliceform that determine the quantification of transcript specific expression levels (RPKM). Note that this uses only a small fraction of reads. The second group of columns assesses mapping reads to the genomic sequence with TopHat. In that approach, gene expression levels can be estimated from gene models explaining the observed alignments by Cufflinks. Alignments are also permitted in gene regions that contribute to different alternative spliceforms. Finally, the last couple of columns shows the result for the combined approach introduced here, where reads are mapped to known transcript sequences using Bowtie but quantification is again based on gene models explaining the alignments of these reads to the genome. While the number of mapped reads and observed alignments is similar to those from TopHat, almost three times as many alignments cover exon junctions, which often are central for identifying the expression of a particular spliceform. The resulting differences in quantification performance are compared in Figure 1.

they have an infinite error on the log-scale (and also the coefficient of variation on the linear scale is always $>80\%$).

2.4 Microarray data analysis

Probe annotation: depending on the target organism, updated genome annotation can considerably affect differential expression estimates for 30–40% of all the targets of an Affymetrix chip (Dai *et al*., 2005). We therefore used the 584 345 probes of the `HuGene-1.0-ST-v1` chip matching the transcript annotation of EnsEMBL r58 (custom CDF v13). For further stringency, confounding probesets were removed (Supplementary Material), yielding 88 464 sets with a median of 18 probes. To allow principled presence calls, we randomly assembled 500 negative probesets with a matching probeset size distribution from probes provided by Affymetrix not matching the genome.

Low-level analysis, normalization: probe specific effects have been fit using an Empirical Bayes 'affinities' model for removing both probe specific background and adjusting perfect-match signal intensities for probe specific affinities, known to significantly increase accuracy and precision (Wu *et al*., 2004). Subsequently, two variants for normalization and estimation of expression levels were considered: (i) the standard MAS 5.0 protocol by Affymetrix, (ii) a combination of modern alternative algorithms: Bioconductor `vsnMatrix`, normalizing for different backgrounds and overall hybridization intensities of individual chips using an iterative 20%-trimmed least squares fit of a generative model with additive-multiplicative noise (Huber *et al*., 2002). The variance-stabilizing generalized log transform for this model was calibrated for asymptotic equivalence to a $\log_2$ transformation. A robust fit of a linear multi-chip probe-level model was then used to compute transcript expression estimates (Bolstad, 2004).

Presence calls: the constructed negative controls allowed more accurate and precise presence calls (Warren *et al*., 2007). The Bioconductor `panp` package was extended to support the `HuGene-1.0-ST-v1` chip.

3 RESULTS

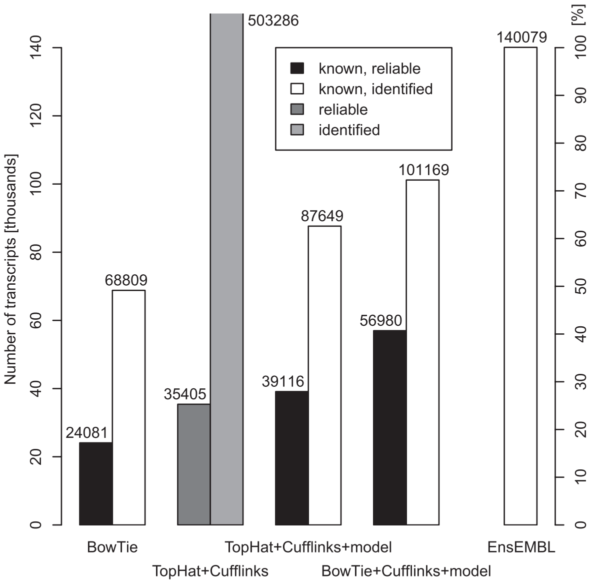
We performed three replicate measurements of mRNA extracted from a human HMEC 184A1 cell line culture. With a total of 993 million 50 bp reads, corresponding to an entire ABI SOLiD-3+ flowcell per measurement sample, this constitutes one of the largest RNA-Seq datasets featuring technical replicates to date.

As a first step for estimating expression levels, different methods for mapping these reads to targets were considered. While Bowtie directly aligned about half of the collected reads to all the known transcript spliceform sequences, only 10% 'unique reads' identified spliceforms unambiguously (Table 1) because for targets with many common sequence regions, only a fraction of aligned reads will be discriminative. These can then be used to estimate gene expression levels. Alternatively, TopHat first aligns reads to the genomic sequence. Estimating expression levels using 'gene models' to explain the observed read alignments arising from different spliceforms then allows the exploitation of non-unique reads, which make up 80% of all the mapped reads. We want to know how reliable these estimates are.

3.1 Reproducibility of quantitative expression profiling

Figure 1 shows the influence of alignment choices and the exploitation of gene models on the number of genes for which transcript expression levels can be estimated reliably (with a relative error of 20% or less). The unique reads mapped by Bowtie identified 68 809 spliceforms (49% of all known transcripts). Of these, 24 081 spliceforms (35%) could be measured reliably. In contrast, TopHat and Cufflinks, set to discover spliceforms *de novo*, identified 503 286 spliceforms. Of these, 35 405 could be measured reliably (7%). It is interesting to consider if providing Cufflinks with a set of known gene models can improve on this result. While one loses the capability of detecting novel spliceforms, assignments to known spliceforms will improve, particularly for spliceforms covered by fewer reads. For the EnsEMBL gene models, 87 640 spliceforms could then be identified (63% of all known transcripts). Of these, 39 116 could be measured reliably (44%). Note that this now also includes spliceforms with non-unique reads.

Considering this marked improvement, we suggest an exploitation of gene models already at the alignment stage by directly mapping reads to the known transcriptome by Bowtie, combined with



Transcripts	Bowtie	TopHat + Cuff-links	TopHat + Cuff-links + model	Bowtie + Cuff-links + model
identified	68,809 (49%)	503,286 (–)	87,649 (63%)	101,169 (**72%**)
reliable	24,081 (17%)	35,405 (–)	39,116 (28%)	56,980 (**41%**)
reliable %	35%	7%	44%	57%

Fig. 1. Statistics of identified and reliably measured transcripts. The plot compares, for different read processing methods, the number of transcripts identified (white and light bars), as well as the number of transcripts that could be measured reliably, i.e.where expression levels could be quantified with an error of 20% or less (black and dark bars). The grey bars are for gene models constructed *de novo*. The black and white bars are for known spliceforms as given by the EnsEMBL gene models. The alternate *y*-axis on the right gives the corresponding fraction of all spliceforms known (EnsEMBL, numbers given in brackets). The last table row gives the ratio of reliably measured transcript relative to the number of identified transcripts. The plot considers four different approaches of computing transcript expression from the observed reads. The first couple of bars assesses read alignments to the transcriptome from Bowtie with subsequent calculation of expression levels from the unambiguously mapping reads. Only 24 081 spliceforms could be measured reliably. The next group shows results for the established programs TopHat and Cufflinks, where an alignment of reads to the genome is followed by expression level estimates from *de novo* constructed gene models. While more transcripts can now be assessed reliably (35 405), also an extremely large number of spliceforms is predicted (off-scale). When Cufflinks is allowed to use the known EnsEMBL gene models, the number of reliably measurable transcripts increases to 39 116 or 28% of all known spliceforms. The approach of read alignment to the transcriptome by Bowtie combined with gene model based expression level estimates by Cufflinks identifies the largest number of known transcripts (72%) and allows the reliable measurement of 56 980, that is 41% of all known transcripts.

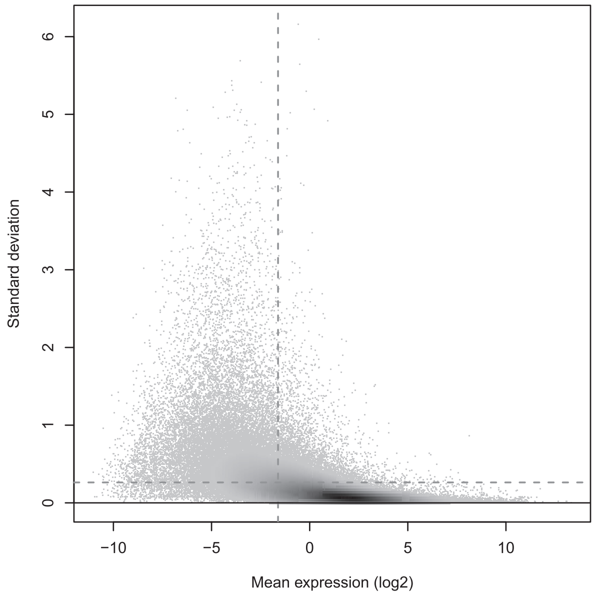


Fig. 2. Standard deviation versus expression level. The plot shows the variation across three technical replicate measurements (standard deviation, *y*-axis), with each discernible dot representing a transcript target. In shaded areas, the grey level represents density, with dark shading indicating higher densities. The standard deviation is in general larger for transcripts with lower mean expression level (*x*-axis). More strongly expressed transcripts could often be measured reliably, with a relative error of 20% or less. Interestingly, just 41% of all transcript targets could be measured that precisely (below the horizontal dashed line). Of the 41% most strongly expressed transcripts (to the right of the vertical dashed line), on the other hand, 84% could be measured reliably (below the horizontal dashed line). This is reflected by the high density of targets on the right (dark shading) falling largely below the horizontal line, which is not the case to the left of the vertical dashed line.

Cufflinks analysis employing the underlying gene models that also permits the use of non-unique reads. Interestingly, this requires less disk space and also runs about an order of magnitude faster, which can be a relevant concern for ultra-deep sequencing sets. Nevertheless, the total mapped read counts are comparable.

While a genome level alignment by TopHat detects additional transcripts *de novo*, the Bowtie alignment of reads to the given spliceform sequences is much more sensitive in the identification of known junctions (almost threefold better; see Table 1). These junctions, however, often play a key role in identifying the expression of a particular spliceform. We could thus identify 101 169 spliceforms (72% of all known transcripts), of which 56 980 could be measured reliably (57%). That means we could assess 41% of all known spliceforms with a relative error of ≤20%. These fall below the horizontal dashed line of Figure 2, which plots the measurement standard deviation versus transcript expression level. The scatter clearly decreases with higher transcript abundance. In view of the 41% of all spliceforms achieving good reproducibility, we can also consider the 41% of targets with the highest expression level. They are found to the right of the vertical mark. Of these, the vast majority (84%) could be measured reliably (below the horizontal line). This is a direct consequence of the sampling nature of RNA-Seq, as discussed below.

3.2 Effects of highly expressed transcripts

Assessing expression levels by randomly sequencing reads from the transcriptome, one expects that some high abundance transcripts can dominate results, such as certain housekeeping genes (e.g. actin, ubiquitin, etc.), or genes abundant in specific cell types or tissues such as secretory proteins or myosin. The difficulty of reliably measuring the expression levels of low abundance spliceforms can be understood from a study of the distribution of sequence reads

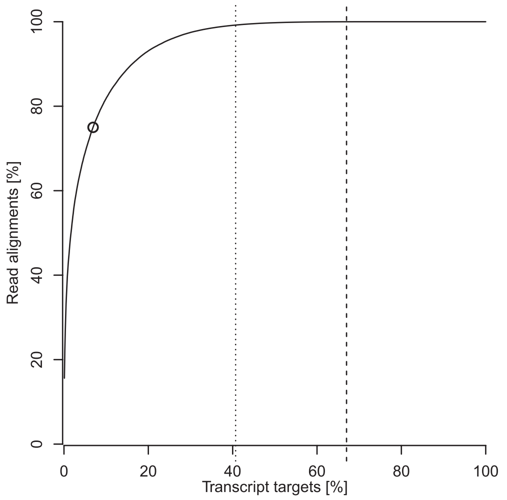


Fig. 3. Cumulative distribution of read alignments across transcript targets. The fraction of read alignments is plotted (*y*-axis) that has been mapped to a certain percentage of transcript targets (*x*-axis). Over 75% of all read alignments cover less than 7% of the known transcriptome (circle symbol). Two particular positions are marked by vertical lines in the figure: The 41% of targets with the highest expression are to the left of the first line (dotted). The vast majority of read alignments (99.5%) has been assigned to these targets, supporting a reliable measurement of their expression levels. Consequently, most of them (84%) could be determined with an error of 20% or less. On average, 67% of all transcript targets were identified in a measurement and this is marked by the second line (dashed). A substantial number of transcript targets falls between the two lines, receiving as few as only one read alignment. Consequently, most of these targets could not be quantified reliably. The remaining 33% of transcript targets falling to the right of the second line (dashed) were either undetected or not expressed.

across transcripts (Fig. 3). On average, 67% of all targets were identified in a measurement (dashed vertical line). Reflecting the complexity of the transcriptome and a highly skewed distribution of expression levels, over 75% of the collected read alignments hit just 7% of all the known spliceforms (circle symbol). Indeed, the vast majority of read alignments (99.5%) has been assigned to the 41% most abundant targets (to the left of the dotted vertical line). Consequently, the expression levels for most of these targets could be determined reliably with an error $\leq 20\%$. In contrast, many targets fall between the two vertical lines, receiving as few as only one read alignment. As a result, most of those could not be quantified with such precision. It is thus interesting to examine how the read depth of an RNA-Seq experiment affects the distribution of genes that can be measured reliably.

3.3 Impact of read depth

In Figure 4 we show the fraction of transcripts with relative measurement error $\leq 20\%$ (*y*-axis) when subsampling 10 000 to 240 million read alignments (*x*-axis). On a log-linear plot this gives a sigmoid relation. The circle symbol indicates 41%, as achieved with an entire flowcell per replicate (331 million reads). For comparison, the plus symbol in the plot marks the results for the standard TopHat+Cufflinks+model pipeline (28%). Remarkably, there is no saturation even at these high read depths, with a doubling now still gaining a further 5% of all known transcripts. Diminishing returns are only seen at much higher read depths.

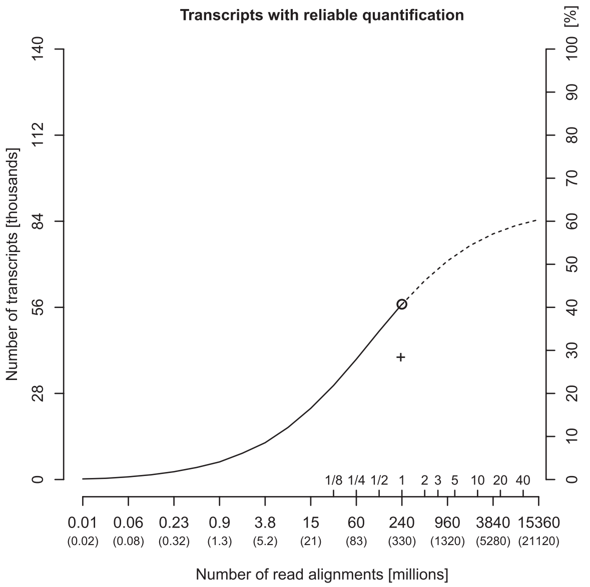


Fig. 4. Transcripts with reliable quantification versus read depth. This graph plots the number of of transcript targets that could be measured reliably with a relative error of 20% or less versus the number of read alignments (*x*-axis). The total number of generated reads is given in parentheses below. Additional tick marks indicate proportions of a flowcell or the number of flowcells worth of sequencing. The alternate *y*-axis on the right shows the percentage of all known transcripts measured reliably. The solid line shows the results for the introduced combined quantification approach (Bowtie+Cufflinks+model). The dependency of the number of transcripts with reliable quantification on the number of read alignments can be described as a function with a sigmoid shape (regression $P < 10^{-15}$). The circle symbol indicates 41%, as achieved with an entire flowcell per replicate (331 million reads). Extrapolation of the fitted sigmoid suggests that about 60% can be reached at 10 billion reads, highlighting the need for efficient complementary steps. See text for discussion. In comparison, the plus sign shows the corresponding result for an established standard approach (TopHat+Cufflinks+model), 28% of all known transcripts. The data shown is for the total, pooled set of reads.

This is in marked contrast to the corresponding analysis of the recall of identified known spliceforms (Fig. 5), which already indicates a point of diminishing returns. The entire set of 993 million reads found 72% targets. The remaining 28% were either undetected or not expressed. This is similar to results from longer read technologies (Mane *et al.*, 2009). An analysis and discussion of read length effects is provided in the Supplementary Section 5.2.

The identification rate as a function of read depth is sigmoid: Growth initially starts slowly (i.e. when few reads are sampled, only highly abundant transcripts will be represented), whereas in the next phase the function increases approximately linearly, with an additional 7% of spliceforms gained for each doubling of the read depth. Beyond 1/4 flowcell (65%), first saturation effects set in, as low-abundance transcripts are already being sampled. This point roughly corresponds to the 40 million mapped reads identified in earlier studies as probably sufficient for the detection of most moderately abundant spliceforms (Mortazavi *et al.*, 2008). It is noteworthy though that the total number of reads one needs to actually collect is higher (as shown on the figure axis in brackets).

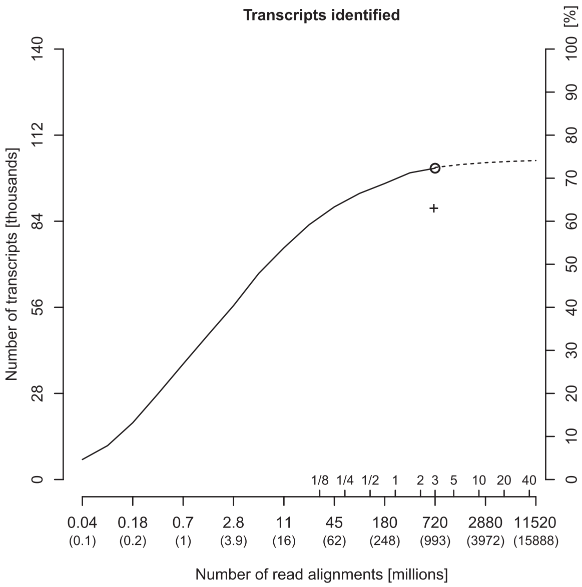


Fig. 5. Transcript identification versus read depth. This plot shows the number of detected transcript targets versus the number of read alignments (x-axis). The total number of generated reads is given in parentheses below. Additional tick marks indicate proportions of a flowcell or the number of flowcells worth of sequencing. The alternate y-axis on the right shows the percentage of all known transcripts detected. The solid line shows the results for the introduced combined quantification approach (Bowtie+Cufflinks+model). The dependency of the number of transcripts identified on the number of read alignments can be described as a function with a sigmoid shape (regression $P < 10^{-12}$). The circle symbol indicates 72%, as obtained with the entire set of 993 million reads. The remaining 28% were either undetected or not expressed in the studied sample. Extrapolation of the sigmoid fit suggests that even with an infinite number of reads only marginally more transcripts would be expected to be identified, yielding an estimate of 20% of transcript targets that are actually not expressed. As a consequence, this experiment reached a target recall of 90% of the estimated true transcript population of the sample. A single flowcell already achieved 84%. In comparison, the plus sign plots the corresponding result for an established standard approach (TopHat+Cufflinks+model), which identified 63% of all known transcripts, which is 79% of the estimated true transcript population. Results are shown for the entire dataset, pooling reads from all replicates.

Extrapolation of the sigmoid shape suggests that, even with an unlimited number of reads, only marginally more transcripts could be identified, yielding an estimate of over 20% of transcript targets actually not expressed, in line with independent other estimates; cf. Supplementary Material and Ghaemmaghami *et al.* (2003).

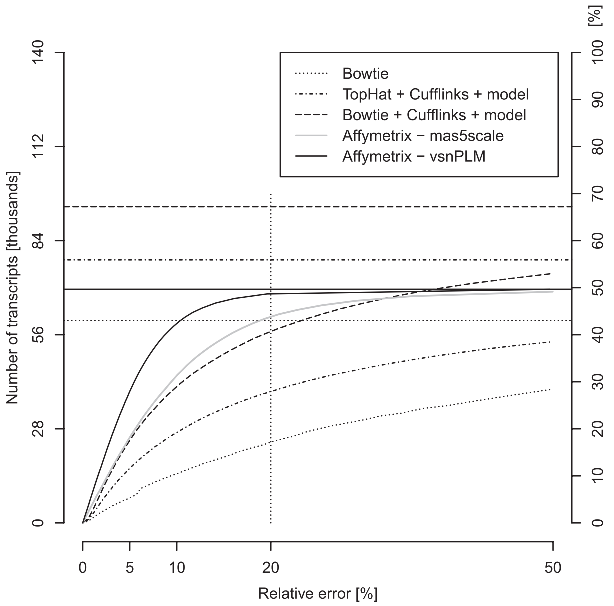


Fig. 6. Comparison of measurement variation. The graph compares the rescaled cumulative distributions of the standard deviation for alternative technologies and data processing protocols. On the y-axis, the number of transcripts is shown for which the quantification error was not more than a given value (x-axis). Technologies and data processing protocols are represented by different line styles and grey shades. The matching horizontal limit marks indicate the respective total number of transcripts identified. The alternate y-axis on the right gives the corresponding fraction of all spliceforms known. Black dotted lines refer to RNA-Seq with the standard Bowtie protocol—mapping reads to known spliceform sequences and quantification based on unique reads. Requiring reliable quantification with a relative error of no more than 20% leaves only 24 081 transcripts of the total 60 266 identified. That is 17% and 43% of all known transcripts, respectively. Dashed and dot-dashed lines refer to RNA-Seq protocols estimating gene expression levels from the read alignments to given gene models. The standard 'TopHat+Cufflinks+model' protocol allowed the reliable quantification of 39 116 spliceforms (28%, dot-dashed line). In contrast, the combined approach introduced in this article yielded 56 980 such spliceforms (41%, dashed line), providing an extension by almost 50%. The horizontal limits mark the total number of transcripts identified, respectively, 78 320 (56%) and 94 157 (67%). Solid lines, finally, represent results from Affymetrix microarrays. With the default processing (MAS5 and scale normalization, grey solid line), 61 394 or 44% of all known transcripts were identified as significantly expressed ('present') and had an error of 20% or less. With more modern processing algorithms (vsn and probe level model expression estimates, black solid line), even 68 278 (49%) could be assessed reliably. The corresponding horizontal mark indicates the 69 586 transcripts (50%) identified as significantly expressed on the array ('presence calls').

3.4 Versus alternative expression profiling platforms

As an independent reference point, we also examined the measurement precision of Affymetrix GeneChips, a well established microarray platform. Figure 6 compares the distributions of the measurement errors for RNA-Seq and chips for different data processing protocols (line styles and shades). On the y-axis, the number of transcripts is shown for which the quantification error was not more than a given value (x-axis). For the standard Bowtie protocol (black dotted lines), only 17% of all known transcripts could be assessed reliably with an error $\leq 20\%$. The 'TopHat+Cufflinks+model' protocol (dot-dashed) yielded 39 116 reliably measured spliceforms (28%). In contrast, the combined approach introduced in this article yielded 56 980 such spliceforms (41%), providing an extension by almost 50%.

Interestingly, the number of transcripts measured precisely by the chip is even higher: 61 394 or 44% of all known transcripts were identified as significantly expressed ('present') and had a relative error of 20% or less for the default Affymetrix protocol (MAS 5.0, grey solid line). With the more modern processing algorithms, 68 278 (49%) could be assessed reliably (black solid line).

By extrapolation from Figure 4, a similar number of reliably measurable transcripts would require a read depth of 10^9 reads per sample. It is worthwhile emphasizing that the observed differences between chips and RNA-Seq are directly due to its uniform sampling of mRNA pools, which is dominated by a small fraction of transcripts. While this apparent dominance was independent of the sequencing technology and the examined sample type (Supplementary Material), we have further verified that such distributions of expression levels were also measured on microarrays, ruling out a sequencing specific artefact (Supplementary Fig. S7).

4 DISCUSSION

Many modern applications in the life-sciences rely on accurate profiles of gene expression, supporting as diverse areas as functional genomics and systems biology. Gene expression profiling by next-generation sequencing protocols like RNA-Seq now promises to discriminate alternative spliceforms, assess allele specific expression (Thas *et al*., 2010), and detect transcript fusion (Levin *et al*., 2009). With a growing interest in applying RNA-Seq for the quantitative assessment of expression levels, the questions of systemic bias and random noise become particularly topical. While systemic deviations due to length bias, lane effects, and processing artefacts are increasingly being investigated (Bullard *et al*., 2010; Marioni *et al*., 2008; Oshlack and Wakefield, 2009), reproducibility has in general received much less attention. Measurement precision, however, determines the power of most analyses, including screens for differential expression, whether they exploit replicates or not (Anders and Huber, 2010).

As illustrated, however, high correlation coefficients and the perceived tightness of scatter plots alone can be misleading (Supplementary Section S8). Indeed, despite overall good correlation between replicates, in one of the first large RNA-Seq studies with technical replicates (2 × 40 million reads per sample), Mortazavi *et al*. (2008) observed reduced precision for less strongly expressed transcripts. We here report on a comprehensive systematic study of the reliability of expression level estimates from an extended dataset with technical replicates (3 × 331 million reads). The observations and trends apparent from this large ABI SOLiD dataset are the direct result of the uniform sampling approach of RNA-Seq. They are therefore of generic nature and independent of a particular sequencing platform or analysis particulars; see supporting complementary results for Illumina Genome Analyser data (Mortazavi *et al*., 2008) and an experiment with paired-end reads from a second generation Illumina device (Supplementary Sections S5 and S6, respectively).

Considering the first step in estimating expression levels, we examined alternative ways of read alignment. Similar to other recent RNA-Seq analyses (Cloonan *et al*., 2008; Sultan *et al*., 2008; Tang *et al*., 2009), only a proportion of the collected reads could ever be mapped to known transcript sequences, even allowing for multiple mismatches (Supplementary Tables S5 and S6). It is noteworthy that at present there is no explanation for the remaining unmapped reads. These constitute a substantial proportion of reads that cannot be identified even by modern alignment algorithms (Homer *et al*., 2009; Langmead *et al*., 2009; Li and Durbin, 2009; Li *et al*., 2009) and that are also not due to contamination with sequences from other organisms (data not shown). The unmappable reads may reflect errors specific to certain sequencing kits (Mortazavi, personal communication, 2009) as well as artefacts from the processing of RNA for massive parallel sequencing (Blow, 2009).

For the simple identification of transcripts, a further increase of read depth may show diminishing returns, as saturation effects were already apparent in the detection rate for higher read depths (Fig. 5). Pooling all 993 million reads yielded a target recall of 90% of the estimated true transcript population in the profiled sample. A single flowcell already achieved 84%. A considerable extension beyond that can still be expected from longer read lengths (Supplementary Fig. S3), especially up to lengths of 300bp per fragment, or from the application of advanced sequencing strategies like paired-end reads (Supplementary Sections S5.2 and S6). While, on one hand, the reads of massively parallel sequencing technologies are getting longer (Wall *et al*., 2009), on the other hand, the sequencing depth achieved by longer-read technologies is continuously being improved (Mane *et al*., 2009). With some approaches claiming full-length reads (Eid *et al*., 2009), a comprehensive identification of collected reads may become feasible in the future.

Depending on the dataset, at present only 10–33% of the collected reads could be mapped to known spliceforms unambiguously. Modern algorithms therefore exploit gene models of alternative spliceforms to infer expression levels that explain both unique and ambiguous read alignments (Griffith *et al*., 2010; Jiang and Wong, 2009; Trapnell *et al*., 2009, 2010).

We next investigated how well the expression levels of individual identified transcripts could be quantified. In general, the more strongly expressed transcripts could be measured more reproducibly (Fig. 2). It is interesting to consider the mechanism behind the larger technical scatter for the less strongly expressed transcripts. Earlier SAGE studies have already shown that a small proportion of genes can be responsible for the majority of a cell's mRNA mass, with just 14% of measured genes contributing 75% of the expressed mRNA (Zhang *et al*., 1997). While it is thus recognized that abundant transcripts can dominate collected samples of expressed mRNAs, the consequences of this effect for genome-wide studies are remarkable, even for deep sequencing. Here we find that over 75% of all read alignments concentrate on just 7% of the known transcriptome. Similarly, the more abundant transcripts in the remaining transcriptome received most of the remaining reads, and so on (Fig. 3). Most of the measurement power was thus spent on a small number of highly abundant transcripts, thus explaining the higher sampling noise for the remaining targets.

With new technologies promising ever higher read depth, we examined its effect on measurement precision. Although the achievable dynamic range of RNA-Seq increases linearly with higher sequencing depths, most of the additional reads will again hit already extensively sampled transcripts. As a result, the number of spliceforms that can be measured reliably follows a sigmoid shape, indicating that even at higher read depths transcripts with low to moderate expression levels are difficult to quantify at good precision with current RNA-Seq protocol (Fig. 4).

As the cost of next-generation sequencing drops and new advanced platforms and protocols emerge, it will be interesting to see the results of full-scale comparisons with sufficiently elaborate replication structures in their experimental designs, similar to recent efforts for a number of microarray technologies (CAMDA, 2008; Editorial, 2008; Tilstone, 2003). In *lieu* of such an elaborate comparison, was can still consider the measurement precision that can be achieved today with a typical microarray. Testing measurement precision on a standard Affymetrix chip, 68 278 transcripts had a signal significantly above background and could be measured reliably with a relative error of 20% or less. This keeps a clear 20% lead on the 56 980 that could be quantified reliably by RNA-Seq. Such a performance should be considered typical, however, rather than optimal for a microarray because spliceform discrimination was not the original goal for this chip design, even though the exon specific probes could identify 69 586 expressed transcript variants (Supplementary Material). It is noteworthy that the performance of microarray technology has improved considerably over the years and in addition owes much to the development of better data analysis algorithms (Huber *et al.*, 2002; Wu and Irizarry, 2005).

We have demonstrated here that newer approaches to read-processing can considerably improve RNA-Seq measurement precision. In particular, exploiting gene models of alternative spliceforms to explain both unique and ambiguous read alignments was clearly beneficial. We then introduced an approach that combines these models with the consideration of known spliceforms already at the alignment stage, which extended the number of transcripts that could be measured reliably by almost 50% from 39 116 to 56 980, reaching 41% of all known transcripts (Fig. 1). Additional targets that could be measured reliably can still come from *de novo* gene models—adding 11 288 novel spliceforms for this dataset (Supplementary Section S1). Such a combined approach thus allows reliable quantitative profiling for a much extended range of existing transcripts while also adding many novel spliceforms. Future tools may adapt this strategy or explore alternatives for maximizing the number of spliceforms that can be profiled reliably.

Nevertheless, the uniform sampling of transcripts by deep sequencing considerably limits the precision achievable by RNA-Seq for low abundance targets. Highly expressed transcripts dominate mRNA samples both independent of sample type (Supplementary Figures S3 and S5) and measurement technology (Fig. S7). This issue is not easily overcome by a further increase in sequencing depths (Fig. 4). It is not just the required increase of read generation but also their processing cost that renders a brute force solution inefficient. Recent technological developments, however, like an extension of the Oligo Library Synthesis kit for RNA libraries by Agilent, offer to draw down subsets of targets through custom mixtures of up to 60 000 different 150–200 bp oligonucleotides, allowing targeted sequencing. This can be used to enrich for low-abundance transcripts when their identities are known (Levin *et al.*, 2009). The removal of high-abundance targets identified early by initial RNA-Seq runs could be an efficient alternative. By using solution reactions and careful design of the capture probes, this approach might not adversely affect the quantitative nature of the profiling protocol while not requiring *a priori* knowledge of low abundance transcripts. An application of several RNA-Seq runs where high-abundance targets are iteratively removed then promises more precise quantitative profiling from the combined results (Fig. 7a). Calibration experiments are, however, required to establish whether such a multi-stage approach can robustly deliver quantitative measurements. Also, manufacture of custom oligo capture kits is still too slow and expensive. Considering the easy availability of custom arrays (Agilent, NimbleGen), normalized mRNA libraries can however already be exploited to identify comprehensive target sets using RNA-Seq. For these targets, custom microarrays can then easily be designed (Leparc *et al.*, 2009) and applied (Fig. 7b). This makes the most of (1) RNA-Seq with its unique *de novo* sequence discovery capabilities and (2) established microarray platforms with their efficient and reliable quantitative assessment of low-abundance targets, combining complementary approaches for state-of-the-art quantitative gene expression profiling.

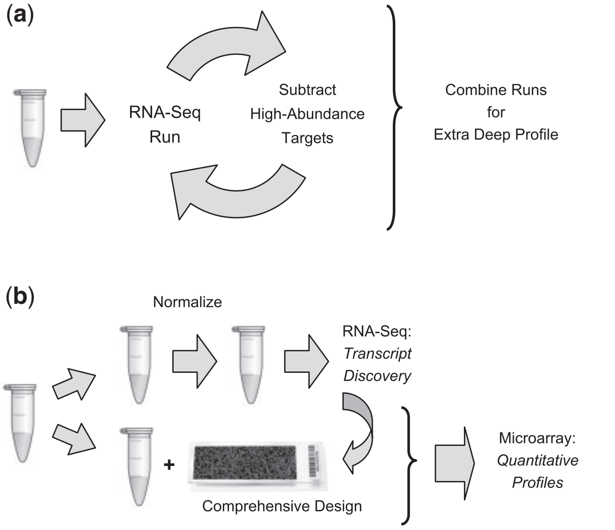

(a)
RNA-Seq Run
Subtract High-Abundance Targets
Combine Runs for Extra Deep Profile
(b)
Normalize
RNA-Seq: Transcript Discovery
+
Comprehensive Design
Microarray: Quantitative Profiles

Fig. 7. Alternative RNA-Seq application schemas. (**a**) In an iterative approach, high-abundance transcripts can be identified in low-read sequencing runs, followed by iterative subtraction of the sequences dominating each sample. A profile from the combined runs promises higher measurement precision of expression levels for weakly to moderately expressed transcripts. (**b**) After normalization of an aliquot (top row), the strength of RNA-Seq in *de novo* sequence discovery can be exploited for the compilation of a comprehensive target library, against which a custom microarray can then be designed easily (Leparc *et al.*, 2009). The remaining aliquot can then be quantitatively profiled on this optimized array (bottom row). The performance of both approaches of course depends on the quality of the subtraction or normalization step, respectively.

ACKNOWLEDGEMENTS

The authors are grateful to Petri Auvinen, Ali Mortazavi, Mitchell Rawlins and Peter Sykacek for helpful discussions.

Funding: Vienna Science and Technology Fund (WWTF); Baxter AG, the Austrian Institute of Technology; Austrian Centre of Biopharmaceutical Technology. A portion of this research was performed using the EMSL, a national scientific user facility sponsored by the Office of Biological and Environmental Research of the Department of Energy of the USA, located at Pacific Northwest National Laboratory; LDRD funding from Pacific Northwest National Laboratory (to B.E.L., in part).

Conflict of Interest: none declared.

REFERENCES

Anders,S. and Huber,W. (2010) Differential expression analysis for sequence count data. *Genome Biol.*, **11**, R106.

Band,V. and Sager,R. (1989) Distinctive traits of normal and tumor-derived human mammary epithelial cells expressed in a medium that supports long-term growth of both cell types. *Proc. Natl Acad. Sci. USA*, **86**, 1249–1253.

Blow,N. (2009) Transcriptomics: the digital generation. *Nature*, **458**, 239–242.

Bolstad,B. (2004) Low level analysis of high-density oligonucleotide array data: background, normalization and summarization. PhD Thesis, University of California, Berkeley, USA.

Bullard,J.H. *et al*. (2010) Evaluation of statistical methods for normalization and differential expression in mrna-seq experiments. *BMC Bioinformatics*, **11**, 94.

CAMDA. (2008) Critical assessment of microarray data analysis conference. Available at http://camda.bioinfo.cipf.es/camda08 (last accessed date May 5, 2011).

Carninci,P. *et al.* (2003) Targeting a complex transcriptome: the construction of the mouse full-length cDNA encyclopedia. *Genome Res.*, **13**, 1273–1289.

Cloonan,N. *et al.* (2008) Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat. Meth.*, **5**, 613–619.

Dai,M. *et al.* (2005) Evolving gene/transcript definitions significantly alter the interpretation of GeneChip data. *Nucleic Acids Res.*, **33**, e175.

Datta,S. *et al*. (2010) Statistical analyses of next generation sequence data: a partial overview. *J. Proteomics Bioinformatics*, **3**, 511–515.

Eid,J. *et al.* (2009) Real-time DNA sequencing from single polymerase molecules. *Science*, **323**, 133–138.

Flicek,P. *et al.* (2010) Ensembl's 10th year. *Nucleic Acids Res.*, **38**, D557–D562.

Ghaemmaghami,S. *et al*. (2003) Global analysis of protein expression in yeast. *Nature*, **425**, 737–741.

Griffith,M. *et al.* (2010) Alternative expression analysis by rna sequencing. *Nat. Methods*, **7**, 843–847.

Homer,N. *et al*. (2009) BFAST: An alignment tool for large scale genome resequencing. *PLoS ONE*, **4**, e7767.

Huber,W. *et al*. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics*, **18** (Suppl. 1), S96–S104.

Jiang,H. and Wong,W.H. (2009) Statistical inferences for isoform expression in RNA-Seq. *Bioinformatics*, **25**, 1026–1032.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Lee,S. *et al.* (2011) Accurate quantification of transcriptome from RNA-Seq data by effective length normalization. *Nucleic Acids Res.*, **39**, e9.

Leparc,G.G. *et al*. (2009) Model-based probe set optimization for high-performance microarrays. *Nucleic Acids Res.*, **37**, e18.

Levin,J.Z. *et al*. (2009) Targeted next-generation sequencing of a cancer transcriptome enhances detection of sequence variants and novel fusion transcripts. *Genome Biol.*, **10**, R115.

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

Li,H. *et al*. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, **18**, 1851–1858.

Li,R. *et al.* (2009) SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, **25**, 1966–1967.

Mane,S.P. *et al.* (2009) Transcriptome sequencing of the microarray quality control (MAQC) RNA reference samples using next generation sequencing. *BMC Genomics*, **10**, 264.

Marioni,J.C. *et al.* (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res.*, **18**, 1509–1517.

Mortazavi,A. *et al.* (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat. Meth.*, **5**, 621–628.

Editorial. (2008) Going for algorithm gold. *Nat. Meth.*, **5**, 659.

Ning,Z. *et al*. (2001) SSAHA: a fast search method for large DNA databases. *Genome Res.*, **11**, 1725–1729.

Oshlack,A. and Wakefield,M. (2009) Transcript length bias in RNA-seq data confounds systems biology. *Biol. Direct*, **4**, 14.

Pruitt,K.D. *et al.* (2009) The consensus coding sequence (CCDS) project: identifying a common protein-coding gene set for the human and mouse genomes. *Genome Res.*, **19**, 1316–1323.

Shendure,J. (2008) The beginning of the end for microarrays? *Nat. Meth.*, **5**, 585–587.

Sultan,M. *et al.* (2008) A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome. *Science*, **321**, 956–960.

Tang,F. *et al.* (2009) mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Meth.*, **6**, 377–382.

Thas,O. *et al*. (2010) Probabilistic allelic read calling: a quasi-Poisson mixed model for the analysis of allelic read counts. *2nd StatSeq Workshop*, May 2010, Ghent, Belgium.

Tilstone,C. (2003) DNA microarrays: vital statistics. *Nature*, **424**, 610–612.

Trapnell,C.*et al*. (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, **25**, 1105–1111.

Trapnell,C. *et al.* (2010) Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.*, **28**, 511–515.

Wall,P.K. *et al.* (2009) Comparison of next generation sequencing technologies for transcriptome characterization. *BMC Genomics*, **10**, 347.

Warren,P. *et al.* (2007) PANP–a new method of gene detection on oligonucleotide expression arrays. *Bioinformatics and Bioengineering, 2007 BIBE 2007. Proceedings of the 7th IEEE International Conference in Boston on 14-17 Oct 2007*, pp. 108–115.

Wilhelm,B.T. *et al*. (2008) Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution. *Nature*, **453**, 1239–1243.

Wilming,L.G. *et al*. (2008) The vertebrate genome annotation (Vega) database. *Nucleic Acids Res.*, **36**, D753–D760.

Wu,Z. and Irizarry,R.A. (2005) Stochastic models inspired by hybridization theory for short oligonucleotide arrays. *J. Comput. Biol.*, **12**, 882–893.

Wu,Z. *et al*. (2004) A model-based background adjustment for oligonucleotide expression arrays. *J. Am. Stat. Assoc.*, **99**, 909–917.

Zhang,L. *et al*. (1997) Gene expression profiles in normal and cancer cells. *Science*, **276**, 1268–1272.