Genome **Medicine**

RESEARCH HIGHLIGHT

RNA-Seq and find: entering the RNA deep field

Adam Roberts[1] and Lior Pachter[1,2]*

Abstract

Initial high-throughput RNA sequencing (RNA-Seq) experiments have revealed a complex and dynamic transcriptome, but because it samples transcripts in proportion to their abundances, assessing the extent and nature of low-level transcription using this technique has been difficult. A new assay, RNA CaptureSeq, addresses this limitation of RNA-Seq by enriching for low-level transcripts with cDNA tiling arrays prior to high-throughput sequencing. This approach reveals a plethora of transcripts that have been previously dismissed as 'noise', and hints at single-cell transcription fingerprints that may be crucial in defining cellular function in normal and disease states.

The deep field

Techniques for directly assessing and quantifying RNA by high-throughput sequencing, collectively known as RNA-Seq [1], have revealed unexpected complexity and diversity in human transcriptomes [2]. Many previously unknown transcripts that are being detected are low-abundance long intergenic non-coding RNAs (lncRNAs), which seem to be crucial for function and in disease [3]. A key advantage of RNA-Seq over previous microarray-based methods for assessing transcription is the ability to query all transcripts on a genome-wide scale without prior knowledge about the locations and structures of genes. However, this advantage of RNA-Seq has also been its Achilles heel: transcriptomes are dominated by few highly abundant transcripts, and the frequent sampling of such transcripts in proportion to their abundances and lengths reduces the power to detect transcripts that are rare or short [4].

An apt analogy for a genome biologist attempting to measure transcription of a short, low-abundance gene from genome-wide RNA-Seq data is the difficulty encountered by an astronomer attempting to detect a low magnitude star from images collected in a low resolution sky survey. The tradeoff between breadth and depth is one that astronomers have grappled with for a long time and have ultimately resolved with the development of telescopes that can limit the scope of a detector to areas of interest, along with guiding technology enabling deep sampling of a region over long exposures. This approach was the design principle for the Hubble Deep Field, which focused a powerful detector on the darkest portions of the sky. With the bright 'foreground' of nearby objects removed, an immense number of galaxies were discovered in what was previously thought to be empty space. Mercer *et al.* [5] have designed an analogous focused experiment for probing the transcriptome, RNA CaptureSeq, and describe a similar outcome: regions with only scattered coverage in genome-wide experiments are revealed to be loci with transcription of low-abundance RNAs. A key aspect of CaptureSeq is that the integrity of the transcriptome in the 'deep field' is preserved: the relative proportions of transcripts sampled with CaptureSeq are shown to be equivalent to the relative proportions in conventional RNA-Seq.

Seq and find

Mercer *et al.* [5] first performed conventional RNA-Seq focusing on a primary human foot fibroblast cell line. *De novo* assembled transcripts from conventional RNA-Seq were combined with 'dark' intergenic regions that seemed not to be transcribed to design capture arrays. The targeted regions were then pulled down using the array, followed by sequencing. Mercer *et al.* [5] provide numerous controls to show that the approach maintains library diversity without introducing PCR amplification bias or other biases.

The enrichment provided by CaptureSeq is estimated to be 380-fold more than conventional RNA-Seq in the targeted regions. Therefore, the resolution achievable with CaptureSeq (in the targeted regions) is approximately equivalent to what could be obtained with 10 billion conventional RNA-Seq reads. As Mercer *et al.* [5] point out, such depth is necessary for finding very-low-abundance transcripts and for accurately quantifying abundances. We reinforce the latter observation in Figure 1, which gives an example showing that extreme

*Correspondence: lpachter@math.berkeley.edu
[2]Departments of Mathematics and Molecular and Cell Biology, UC Berkeley, Berkeley, CA 94720, USA
Full list of author information is available at the end of the article

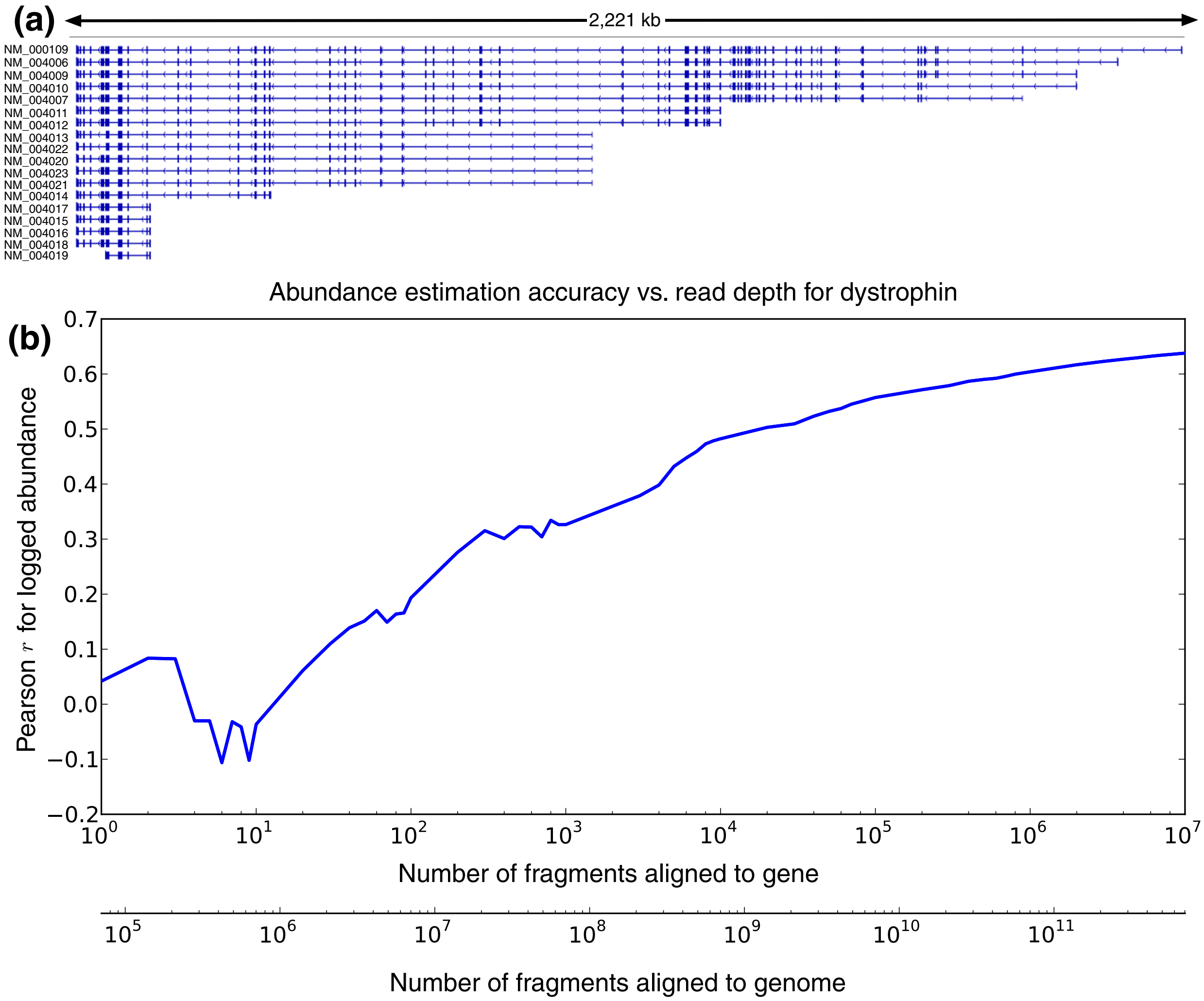

Figure 1. To demonstrate the difficulty of accurate isoform-level abundance estimation on low-abundance genes, we simulated an RNA CaptureSeq experiment on the 18 isoforms of the dystrophin gene as annotated in RefSeq (hg19), shown in (a). For each number of 76 bp paired-end fragments that aligned to the gene, we estimated abundances of each isoform using the online EM algorithm (for details of the model used see [8] and for details of the implementation see [11]). **(b)** The accuracy of isoform abundance estimation measured as the Pearson correlation coefficient (*r*) of the logged relative abundance estimates compared with the true abundance used to generate the simulated data. The results are averaged over four simulations from different random abundance distributions. Because of the similarity of the isoforms, only 2.5% of fragments aligned uniquely to a single isoform on average, making the deconvolution particularly difficult. The bottom *x*-axis shows how many alignable paired-end fragments would be required to achieve the same *r* in a genome-wide experiment as in the CaptureSeq simulation. Here we assume 3.17 fragments per kilobase per million mapped reads (FPKM) for the gene, which is what we estimated from a sample ENCODE dataset [accession SRR065495].

depth is necessary for accurate isoform abundance estimation for the dystrophin gene, a complex multi-isoform gene that can harbor mutations causing muscular dystrophy. The need for deep sequencing in the example arises from the overlap between the multiple isoforms of the gene and the ambiguity that this causes in read mapping. Large gene families are equally difficult to resolve for the same reason. It has been estimated that only 60% of transcripts can be accurately quantified with 10 billion conventional RNA-Seq reads [6].

Increased accuracy is only one advantage of CaptureSeq. Another striking result of Mercer *et al.* [5] is the number of previously unknown transcripts discovered and the corollary that current sequencing experiments are very far from saturation. The message is clearly 'seq and find' and this is exactly what is happening in RNA-Seq. The experiments surveyed in [1] are an order of magnitude smaller than the norm today, and it is reasonable to extrapolate that as the costs of sequencing drop precipitously, the average depth of sequencing in

RNA-Seq experiments will increase by another order of magnitude in the next 3 years.

Breaking the curse of deep sequencing

Given the observations above, it is natural to speculate that conventional RNA-Seq with 10 billion reads will be commonplace in the near future and to ask whether technologies such as CaptureSeq are truly necessary. It certainly seems plausible that exome sequencing, which is to genome sequencing as CaptureSeq is to conventional RNA-Seq, will eventually be replaced by routine whole genome sequencing. However, in addition to technological challenges that must be overcome to allow routine sequencing of 10 billion reads, there are also bioinformatics problems that must be solved if such data are to be useful. In particular, increased numbers of reads in RNA-Seq lead to the 'curse of deep sequencing', in which extra sequence actually reduces performance and accuracy in current processing pipelines. This happens for two reasons. Firstly, most existing algorithms for RNA-Seq quantification require loading a substantial fraction of the total number of sequenced reads into memory. This is already a challenging prospect at 100 million reads. Algorithms whose running times are not linear in the number of reads are also likely to fail with large amounts of sequence data. Secondly, reads have errors and are prone to various biases [7], which, while appearing at a fixed frequency, occur at greater numbers with increased reads. For example, if a sufficiently large amount of sequence is obtained, a recurring error in a highly abundant gene may appear to be a (false) novel isoform. It therefore becomes imperative with large amounts of sequence to correct for sequencing artifacts, and this can be computationally prohibitive.

CaptureSeq breaks the curse of deep sequencing by providing increased transcript resolution at fixed sequencing depth. This means that existing methods can be readily applied to the analysis of CaptureSeq data, and indeed the authors [5] show that the Cufflinks suite of tools can be used for both assembly [8] and quantification [7] of CaptureSeq data.

The single cell transcriptome

Mercer *et al.* [5] emphasize the discovery of novel lncRNAs in the deep field. They found 163 novel neighboring and antisense lncRNAs around protein coding genes. In general, they found that captured lncRNAs have very low expression of only 0.011 FPKM (fragments per kilobase per million mapped reads). These findings follow on the heels of a recent comprehensive annotation of human lncRNAs [9] and together suggest that very rare transcripts may bestow individual transcriptional 'fingerprints' on cells. In fact, Mercer *et al.* [5] estimate that the newly discovered lncRNAs are present at an average copy number of 0.0006 transcripts per cell. This precise quantification together with evidence from CaptureSeq that the sequenced fragments are samples from complete transcripts (and not just 'noise') points towards the presence of very rare transcripts, possibly even unique to individual cells.

CaptureSeq therefore motivates the development of other approaches to the enrichment of low-abundance transcripts. One complementary possibility is the further development of depletion approaches, which could selectively filter the highest-abundance transcripts before sequencing; an example is removal of ribosomal RNA already performed in many experiments [1]. Although the depletion approach may inadvertently remove lower-abundance RNAs because of cross-hybridization, it offers a genome-wide approach to enrichment. Furthermore, CaptureSeq may have a similar bias in the opposite direction: repetitive sequence in the transcriptome might lead to captured RNA from outside a targeted genomic region.

RNA CaptureSeq and beyond

Regardless of how low-abundance transcripts are detected, Mercer *et al.* [5] have demonstrated the extent of discovery possible in the deep field. The functional relevance of ultra low-abundance transcripts is currently debated [9,10], and the question of whether rare transcripts regulate biologically important processes or are artifacts of stochastic transcription is a key open problem. However, there is increasing recognition that antisense lncRNAs are present at many protein coding genes, including numerous proto-oncogenes, and that they regulate their associated genes via epigenetic modifications [3]. The ability to see farther into the RNA deep field with CaptureSeq is therefore likely to lead to many exciting developments in genomic medicine thanks to better understanding of the aberrant transcription underlying human disease.

Abbreviations

FPKM, fragments per kilobase per million mapped reads; lncRNA, long intergenic non-coding RNA; RNA-Seq, RNA sequencing.

Competing interests

The authors declare that they have no competing interests.

Acknowledgments

AR is supported by a graduate research fellowship from the National Science Foundation. LP is supported in part by grant NIH R01 HG006129-01. We thank Meromit Singer for comments on the curse of deep sequencing.

Author details

[1]Department of Computer Science, UC Berkeley, Berkeley, CA 94720, USA.
[2]Departments of Mathematics and Molecular and Cell Biology, UC Berkeley, Berkeley, CA 94720, USA.

Published: 22 November 2011

References

1. Wilhelm BT, Landry JR: RNA-Seq--quantitative measurement of expression through massively parallel RNA-sequencing. *Methods* 2009, **48:**249-257.
2. Wang Z, Gerstein M, Synder M: RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009, **10:**57-63.
3. Taft RJ, Pang KC, Mercer TR, Dinger M, Mattick JS: Non-coding RNAs: regulators of disease. *J Pathol* 2010, **2:**126-139.
4. Oshlack A, Wakefield MJ: Transcript length bias in RNA-seq data confounds systems biology. *Biol Direct* 2009, **4:**14.
5. Mercer TR, Gerhardt DJ, Dinger ME, Crawford J, Trapnell C, Jeddeloh JA, Mattick JS, Rinn JL: Targeted RNA sequencing reveals the deep complexity of the human transcriptome. *Nat Biotechnol* 2011, Epub ahead of print.
6. Łabaj PP, Leparc GG, Linggi BE, Markillie LM, Wiley HS, Kreil DP: Characterization and improvement of RNA-Seq precision in quantitative expression profiling. *Bioinformatics* 2011, **27:**i383-i391.
7. Roberts A, Trapnell C, Danghey J, Rinn JL, Pachter L: Improving RNA-Seq expression estimated by correcting for fragment bias. *Genome Biol* 2011, **12:**R22.
8. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold B, Pachter L: Transcript assembly and quantification by RNA-Seq reveals unannotated transcript and isoform switching during cell differentiation. *Nat Biotechnol* 2010, **28:**511-515.
9. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL: Integrative annotation of human intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* 2011, **25:**1915-1927.
10. Van Bakel H, Nislow C, Blencowe BJ, Hughes TR: Most "dark matter" transcripts are associated with known genes. *PLoS Biol* 2010, **8:**e1000371.
11. eXpress: Streaming RNA-Seq Analysis [http://bio.math.berkeley.edu/eXpress/]

doi:10.1186/gm290
Cite this article as: Roberts A, Pachter L: RNA-Seq and find: entering the RNA deep field. *Genome Medicine* 2011, **3:**74.