Available online at www.sciencedirect.com

SciVerse ScienceDirect

Current Opinion in Chemical Biology

Transcriptomics in the RNA-seq era

Paul A McGettigan

The transcriptomics field has developed rapidly with the advent of next-generation sequencing technologies. RNA-seq has now displaced microarrays as the preferred method for gene expression profiling.The comprehensive nature of the data generated has been a boon in terms of transcript identification but analysis challenges remain. Key among these problems is the development of suitable expression metrics for expression level comparisons and methods for identification of differentially expressed genes (and exons). Several approaches have been developed but as yet no consensus exists on the best pipeline to use.*De novo* transcriptome approaches are increasingly viable for organisms lacking a sequenced genome. The reduction in starting RNA required has enabled the development of new applications such as single cell transcriptomics.The emerging picture of mammalian transcription is complex with further refinement expected with the integration of epigenomic data generated by projects such as ENCODE.

Address
School of Agriculture and Food Science, University College Dublin, Belfield, Dublin 4, Ireland

Corresponding author: McGettigan, Paul A (paul.mcgettigan@ucd.ie)

Current Opinion in Chemical Biology 2013, **17**:4–11

This review comes from a themed issue on **Omics**

Edited by **Matthew Bogyo** and **Pauline M Rudd**

For a complete overview see the Issue and the Editorial

Available online 2nd January 2013

1367-5931/$ – see front matter,

http://dx.doi.org/10.1016/j.cbpa.2012.12.008

Introduction

Transcriptomics — one of the original 'omics

The 'transcriptome' is defined as 'the complete complement of mRNA molecules generated by a cell or population of cells'. The term was first proposed by Charles Auffray in 1996 [1] and first used in a scientific paper in 1997 [2]. Unlike many of the technologies that have acquired the '-ome' appendage the 'Transcriptome' has a long pedigree and certainly meets the requirements of a true 'omics technology [3].

The last couple of years have seen intense development of transcriptomic applications and the supplanting of microarrays by RNA-seq as the technology of choice for gene expression analysis. However the amount of data generated by these technologies has generated problems both of data management and storage as well as posing novel analytical problems.

Although the transcriptome can encompass many species of RNA (miRNA, snoRNA, etc.) this review will focus mainly on mRNAs, specifically mammalian mRNAs. Readers can find good reviews of the advances that have been made in nonmammalian and noneukaryotic transcriptomics in other locations [4,5].

In contemporary multidisciplinary projects global transcription profiling is frequently the first 'omics technology to be applied. It generates information about which genes are expressed, at what level and can also provide information about different transcript isoforms used. A preliminary analysis via microarray or RNA-seq can indicate the appropriateness or usefulness of other 'omics technologies such as proteomics, glycomics or metabolomics. It can be a relatively cheap way of determining the likely interesting subsets of samples that are likely to generate results in other 'omics technologies. It can also be used to indicate modifications of capture protocols which should be for technologies such as proteomics; where the biochemical idiosyncrasies of particular proteins or protein families can make it difficult to isolate proteins or metabolites which the RNA-seq data have indicated to be of potential interest.

One example of this type of multidisciplinary approach can be found in our own work. For the past five years our reproductive biology cluster has been profiling different tissues of the female bovine reproductive tract under different conditions of pregnancy status, stage of estrus cycle or embryo development. In each case the initial RNA-seq experiment is then complemented by additional profiling with proteomics, metabolomics, or glycomics. Each 'omics technology helps to piece together a complex biological picture for example; how the endometrium tissue can support embryo growth and implantation (proteomics analysis of histotroph [6] following RNA-seq of endometrium [7] and embryo [8]), how enzymes expressed in follicular tissue can support the development of oocytes before ovulation (RNA-seq of theca and granulosa cells [9] followed by metabolomic profiling of follicular fluid [10]) or to determine exactly how the modulation of glycosylation enzymes impact on cervical mucus structure and generate a permissive or hostile environment for sperm or bacterial transit (glycomic profiling of cervical mucus following RNA-seq of cervical tissue [11]).

Glossary

cDNA: Complementary DNA is synthesized from mRNA using reverse transcriptase. This is the starting material typically used in nextgen sequencing or gene expression microarray protocols for measuring RNA levels.

De novo **assembly:** Constructing a transcriptome in the absence of an assembled genome sequence for the organism.

DGE: Digital Gene Expression. An alternative protocol for measuring gene expression. It is a version of the SAGE protocol adapted for use with next-generation sequencers.

ENCODE: Encyclopedia of DNA elements. A research consortium whose goal is to identify the functional elements in the human genome.

EST: Expressed Sequence Tag. A subsequence of a cDNA sequence. Generated with earlier generations of DNA sequencers (Sanger sequencing method).

Microarray: Expression microarrays are collection of DNA spots (probes) attached to a solid surface. These probes hybridize to cDNA reverse transcribed from RNA samples. Levels of hybridization are measured using fluorescence and converted into expression measurements.

mRNA: Messenger RNA. An RNA product that is transcribed from the DNA and ultimately transported to a ribosome where it is translated into protein.

miRNA: Micro-RNA, a short RNA (~22 bp when fully processed) which can bind complementary sequences on target mRNAs resulting in translational repression or mRNA degradation.

Read: A sequence of DNA bases generated by a sequencer. Early Illumina/Solexa sequencing generated single-reads 100's of millions of 18 bp in length. Current GA2 and HiSeq machines can generate paired reads of up to 150 bp in length. Reads can be generated either from one end of a DNA fragment (single-read) or from both ends (paired-end reads). The paired-end reads can have advantages for splice isoform reconstruction in RNA-seq.

RNA editing: Molecular process where sequence of an RNA molecule is altered after transcription. In mammals A-I (adenosine to inosine) is the most common form of editing.

RNA-seq (or mRNA-seq): The most popular protocol for measuring RNA levels using nextgen sequencing. Typical steps involve poly-A selection, reverse transcription into cDNA, fragmentation to desired size followed by ligation of the sequencing primers. The protocol can be either strand-specific (retaining information about the directionality of the original transcript) or nonstrand-specific depending on the protocol used.

SAGE: Serial Analysis of Gene Expression. A technique used to measure gene expression. Involves isolation of poly-A RNA, digestion of the RNAs using a restriction enzyme and sequencing of the resulting short tags.

snoRNA: Small nucleolar RNAs. A class of small RNAs that chemically modify other RNAs.

Transcriptome: The complete complement of mRNA molecules generated by an organism or cell type.

Main text

Brief history of transcriptomics

The first efforts at profiling mammalian transcriptomes started in 1991 with the publication of a human EST database compiled by a group from the NIH led by J. Craig Venter [12]. This database consisted of just 609 cDNA clones with an average length of 397 ± 99 bases. It represented one of the earliest applications of the then newly developed automated Sanger sequencing technology. This technology enabled methods such as SAGE (Serial Analysis of Gene Expression) which were one of the first attempts to attempt to quantify gene expression on a global basis [13]. The first SAGE publication contained just 1000 manually sequenced 9 base tags (or 13 bases with 4 inferred from the restriction enzyme cleavage site). At almost the same time the first microarray publications emerged [14]. Microarrays (single color and dual color) utilizing complementary probe hybridization, quickly became the technologies of choice for transcription profiling and came to dominate the field for the next decade.

The recent revolution in sequencing represented by short-read (otherwise known as 2nd generation or next-generation) technologies has enabled the sequencing approach to leap ahead of the microarray approach once again. In 2006 the first RNA-seq paper was published utilizing 454/Roche technology [15]. The data generated comprised just 200,000 reads of length 110 bp for a total of 20 Mbases of data. However for sequencing approaches to be successful a lot of sequencing is needed (the more the better) and it was not until the advent of greater throughput that RNA-seq was able to compete with microarrays. The era of RNA-seq dominance began in earnest in 2008 with a trio of papers utilizing the new short-read technology developed by Solexa (now Illumina) [16–18]. From the outset the Illumina/Solexa technology has generated gigabases of data per run (initially 1 GB per run for the Genome Analyzer when it was initially released in 2006 rising to 600 GB per run for the HiSeq 2500 as of 2012). While the Roche/454 technology has always generated reads long enough for RNA-seq it has been hampered by the relatively low throughput and high cost of the libraries compared to the more popular Illumina technology.

Early transcriptomic projects utilized scaled up versions of SAGE, that is, Digital Gene Expression (DGE), which made a virtue out of the necessarily short (18 bp) reads available when the Solexa machines were first introduced. The DGE protocol was quickly abandoned for mRNA-seq as soon as longer read lengths (≥25 bp) enabled unique mapping of randomly fragmented cDNA reads to a mammalian genome.

Recent technological advances

The most popular technology for RNA-seq has been the Illumina Genome Analyzer and Hi-Seq (http://www.illumina.com/technology/sequencing_technology.ilmn).
Illumina sequencing technology has steadily increased read length and overall number of reads generated per run since its introduction in 2007. The first cohort of mRNA-seq papers used short single reads of 25–40 bp whereas a contemporary project will typically utilize long paired-end strand-specific reads [19]. These longer paired-end reads enable higher levels of mappability. They also enable better identification and mapping of spliced reads [20] as well as enabling the assembly of transcriptomes in

the absence of a reference genome using *de novo* assembly approaches (see below).

Some concern has existed over the impact of PCR amplification on the accuracy of gene expression quantitation via RNA-seq. The Helicos sequencer [21] used an amplification free technology and some of the 3rd generation sequencers are also amplification free (PacBio and Ion-Torrent). There are also methods (such as FRT-Seq [22]) that can be used with the Illumina sequencer that also avoid amplification. The failure of the protocols like FRT-Seq or technologies such as the Helicos sequencer to gain widespread adoption can probably be attributed to the development of strand-specific paired-end protocols which seem to have effectively nullified the problem of amplification bias [22].

Researchers looking to keep abreast of developments in sequencing technology and related bioinformatics analysis problems should consult the excellent Seqanswers website [23].

Bioinformatics challenges

The first major bioinformatics problem posed by the emergence of RNA-seq was the alignment of the reads to a reference genome. Given that the number of reads in a RNA-seq sample can be of the order of millions (even tens of millions) alignment speed has been the primary performance metric by which these tools have been judged. This has led to the displacement of the original cohort of aligners by tools based on the Burrows Wheeler Transform such as Bowtie [24] and SOAP [25].

The early years of microarray analysis were dogged by the analytical problems of high dimensionality, probe cross hybridization and difficulties determining appropriate normalization and differential expression strategies. However consensus analysis approaches eventually emerged [26]. Next-generation sequencing has solved some of the problems associated with microarrays while at the same time posing new ones. Probe hybridization is not a feature of RNA-seq, but the misalignment of reads to closely related genes can lead to a phenomenon called 'transcript shadowing' [27,28]. Similarly the requirement for sample normalization and bias correction has not been eliminated [29–31] although it is arguably much reduced compared to microarrays. At a minimum some normalization must be carried out to account for the different sequencing depths in each library. To this end existing metrics from SAGE analysis such as Tags/Transcripts per million (TPM) and new metrics such as RPKM [16], FPKM [27], per-lane upper quartile correction metric (UQUA) [32], and trimmed mean of M values (TMM) [33] have been developed to compare expression levels both between and within samples. In many cases the simpler TPM metric is sufficient [34].

Batch processing problems previously associated with microarrays were detected among some of the early RNA-seq datasets [35]. However the prudent application of indexing/multiplexing strategies can mitigate many of these problems. Usually RNA-seq technical replicates (same library different lane/flowcell) are so similar as to be capable of being combined via summation [32].

The count nature of the next-generation data has necessitated the development of new algorithms (or rediscovery of SAGE analysis techniques) to accurately estimate differential expression. The early RNA-seq papers frequently used the Poisson model to identify differentially expressed genes. This approach has increasingly been recognized as inappropriate. The most commonly used methods have been parametric methods utilizing variants of the negative binomial distribution such as edgeR [36], HTseq [37], bayseq [38], and NBPseq [39]. Nonparametric methods such as NOISeq [40] and Samseq [41] and expectation–maximization methods such as RSEM [42] have also been applied to this problem. The Fisher Exact Test (FET) also performs well in some comparisons. No consensus has yet emerged as to the best algorithm or pipeline to use [32].

One factor that can confound the identification of optimum analysis strategies is the rapid development of the sequencing chemistries. Most of the reference datasets (e.g. Marioni [43] and MAQC datasets [32]) were generated using early versions of the Illumina chemistry. It is likely that some of the biases identified in these early datasets are no longer present — potentially replaced by new ones.

The last two years has seen the standardization of the SAM/BAM format for short-read alignment data [44]. This standardization has eased the adoption of genome browsers such as the Integrative Genomics Viewer (IGV) from the Broad Institute which enable the rapid graphical navigation of raw RNA-seq data in their genomic context [45]. These tools can be very helpful for interpretation of results and visual identification of potential artifacts.

Splicing

Although the earliest RNA-seq papers identified novel splicing variants the difficulties of alignment, transcript assembly, and annotation have meant that analysis of differential splicing has not yet become routine. Nevertheless it has been shown that different tissues are a key source of differential splicing/differential exon usage [46]. Longer, paired-end, strand-specific [47] reads have enabled easier mapping of reads overlapping spliced junctions and enabled easier linkage of novel exons to known gene models. More robust methods for determining differential splicing building on differential transcript expression analysis have also recently emerged [48,49].

Data repositories

The orders of magnitude difference in size between the raw data from a RNA-seq experiment compared to a microarray experiment present technical and economic challenges particularly for data repositories (which notably passed the 1 million dataset milestone this year [50]). One method to reduce this problem is the application of 'lossy' data compression strategies for short-read data such as the CRAM file format [51]. Other approaches such as reference-based compression [51] as well as the utilization of probabilistic data structures have also been demonstrated [52]. As the expense of data storage mounts it is likely that these formats will see higher levels of adoption.

Distinguishing novel biology from technical artifacts

The sheer size of RNA-seq datasets coupled with systemic read errors can pose problems for naïve analysts exploring the frontiers of this type of data. The highest profile example of this type of problem was seen in the recent controversy about the scope of RNA editing in mammalian cells [53] but it has also been cited in relation to the identification of gene imprinting [54] using RNA-seq. This category of problem has even resulted in the coining of a new 'law' viz MacArthur's Law: 'interesting results are more likely to be artifacts — even after accounting for MacArthur's Law', named after the researcher who has been the most prominent in publicizing this problem [55]. Just as with RNA editing and gene imprinting, RNA-seq also has the capability to identify eQTLs as well as allele specific expression (ASE) via the use of expressed SNPs (eSNPs) [56]. As with RNA editing these types of analysis will require careful filtering of sequencing artifacts.

A problem of potentially much wider impact is the discovery that global transcriptional amplification by c-MYC can increase the overall RNA pool in a cell thus undercutting one of the unspoken assumptions of global transcriptome analysis [57]. Alternative protocols that spike in RNA calibrated to cell number or DNA content have been proposed to correct for this effect. The extent to which this will force reconsideration of previous expression studies is as yet unclear.

New applications of RNA-seq

De novo transcriptomics

Of particular interest to researchers working on organisms where no reference genome exists has been the development of *de novo* methods to characterize the transcriptome. Popular assembly methods include Oases [58], Trinity [59], and trans-ABySS [60]. Most of these methods rely on the de Bruijn graph data structure. This data structure lends itself naturally to genomic and transcriptomic data and saw one of its' first bioinformatics applications in EST assembly [61]. While the *de novo* tools are still outperformed by genome-guided methods they perform well for highly expressed transcripts and will benefit from the use of longer reads and increased sequencing depth as sequencing costs reduce (Figure 1).

One limitation of current *de novo* assemblers is lack of robustness to base miscalls. Crucial read preprocessing steps have been identified for accurate transcriptome assembly. These steps typically involve quality score-based trimming and filtering using tools such as EA-utils [62] and the FASTX-toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) and filtering or trimming reads containing low frequency kmers using tools such as khmer [63].

All of the *de novo* assembly algorithms are highly memory intensive. New algorithmic approaches will be required to reduce reliance on large shared memory machines when using these tools.

Single cell transcriptomics

RNA-seq has enabled the profiling of samples from very small amounts of starting material. This has been a boon for applications such as single cell transcriptomics [64] enabling detailed profiling of: the early embryo transcriptome [65], single neuron transcriptome [66], and peripheral circulating tumor cells [67].

Beyond mRNA-seq — new applications in the transcriptomic ecosystem

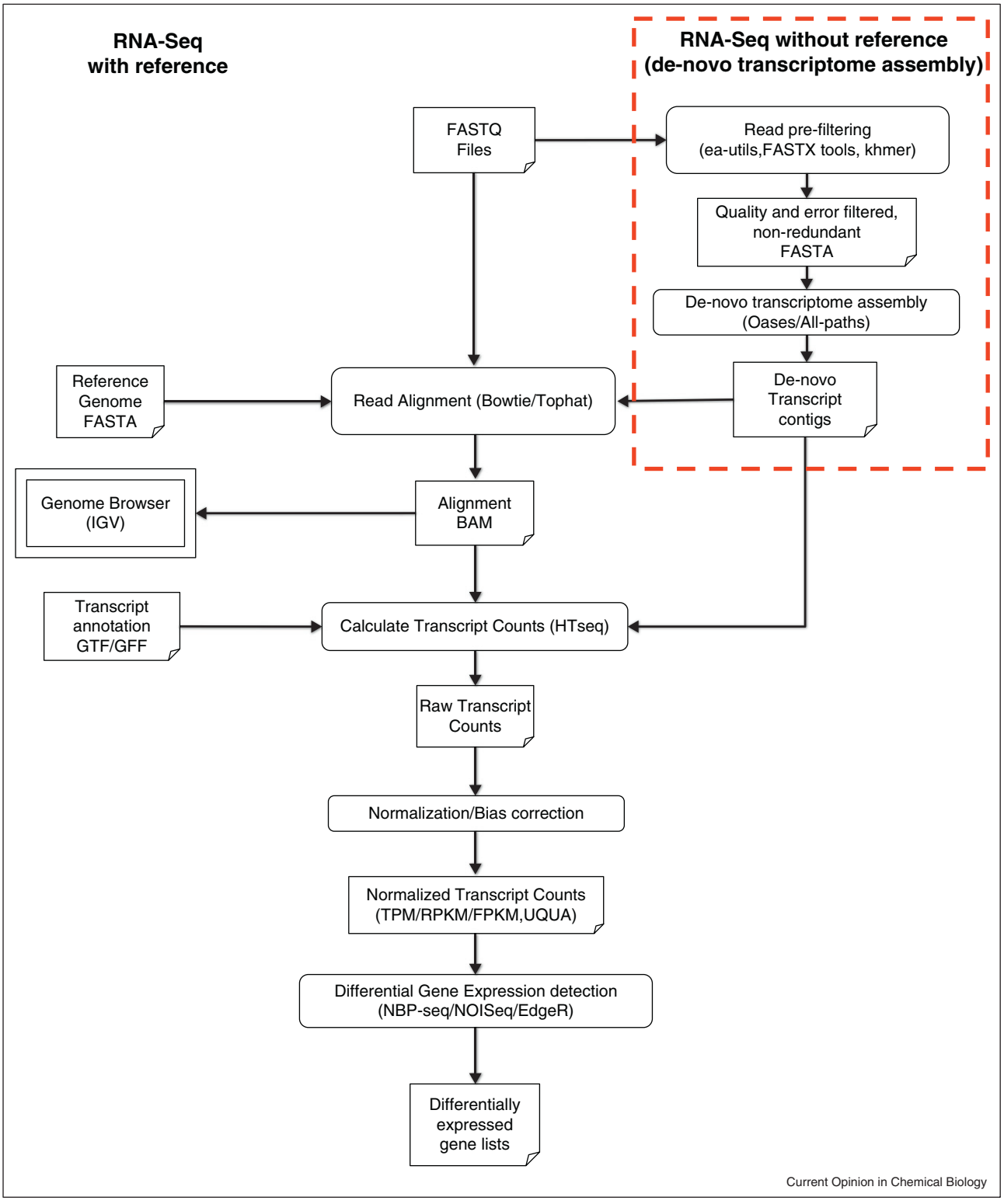
The cellular biology surrounding the transcriptome has seen a similar proliferation in -seq and 'omic approaches. These have been concentrated in the area of the epigenome — or how DNA and histones are modified to control gene expression patterns (see ENCODE below) as well as in the layers of regulation between transcription and translation. Applications such as CLIP-seq [68] profile RNA binding profiles. Ribosome profiling or Riboseq [69] has enabled the first genome wide surveys of exotic phenomena such as dually decoded regions where alternate protein products are produced based on translation frameshifts [70]. Specialized protocols have also been developed to sequence the small RNA population of the cell (miRNAs, snoRNAs, etc.) [71].

The availability of orthogonal transcriptomic and proteomic datasets has enabled the measurement of the linkages between RNA and protein levels in individual samples. The results of these analyses show that the transcript and protein levels of an individual gene vary quite dramatically. One estimate was that only 40% correlation exists between the transcript and protein level of genes overall [72]. These results suggest caution when attempting to project results from the transcriptomic domain into the proteomic.

Epigenetic control of the transcriptome — the ENCODE project

The complexity of mammalian transcription is dictated by the differential expression of different genes or isoforms in different tissues. These patterns of expression

Figure 1



Bioinformatics pipeline showing typical tasks involved in RNA-seq analysis. Additional steps required for *de novo* transcriptome assembly are shown in box at top right.

are reflected in the chromatin state (epigenetic state) of the cell and the accessibility of the DNA control regions of different genes to the various transcription factors that are present.

Until recently most of these regulatory elements have been unidentified. A project in the forefront of identifying them has been the ENCODE project. This project aims to catalog and dissect the functional noncoding elements of the human genome [73,74]. The initial phase of the project focused on just 1% of the genome and utilized primarily array-based technologies. The full-scale project aims to profile these elements across the entire genome and in multiple cell lines and has pioneered the use of various -seq approaches (ChIP-seq, DNase-seq, FAIRE-seq, etc.). While the data generated by ENCODE are recognized as being of great utility, their claim that 80% of the human genome has a specific biological function [75] has been greeted with much skepticism [76].

Already the results of the initial phase of the ENCODE project have led to an evolution in the definition of a gene [77]. It also showed that the majority of the nonrepetitive genome is transcribed under some circumstances [74]. Several novel types of transcription occurring in the vicinity of mRNAs were identified, that is, PASR/TASR, TSSa-RNA, tiRNA [78] corresponding to promoter/terminator associated RNAs, transcription start site associated and transcription initiation RNAs respectively. Other types of noncoding RNAs (such as lincRNAs [79]) have also been identified. These have the potential to complicate RNA-seq analyses and frustrate automated annotation algorithms. However the functionality of much of these novel transcripts, especially low abundance intergenic transcripts, have been called into question [80] especially with the well known lack of specificity of the Pol2 enzyme [81].

The recently produced atlases of regulatory elements in different tissues and cell lines for the mouse [82] and human [83] are good examples of the use of ENCODE type regulatory datasets. It is likely that similar atlases will be developed for other organisms in the years ahead and they will be important for our understanding of transcription control — that is, not just what genes are expressed but why they are expressed.

The future — application of 3rd generation sequencing technologies

While the Illumina GA and HiSeq sequencers currently dominate the transcriptomic landscape, the last two years has seen the development of the basic sequencer in several directions: that is, speed of data generation and length of reads generated. These improvements have come about both by modification of existing sequencing technologies (i.e. MiSeq) and new sequencing technologies (PacBio and IonTorrent/PGM). The long reads generated by PacBio could be very useful in hybrid approaches to *de novo* transcriptomics, similar to how it can be used in hybrid approaches to *de novo* genome assembly [84].

Conclusions

Five years into the next-generation sequencing revolution RNA-seq has been widely adopted and has effectively displaced microarrays for gene expression analysis. Unfortunately RNA-seq has not been the panacea to the problems of gene expression analysis that some may have hoped: artifacts and biases exist that still need to be identified and controlled for.

The last two years has seen an explosion of RNA-seq analysis approaches. The next few years will hopefully see consensus emerge on the best analysis pipeline.

However the real advantages to RNA-seq has been in the new applications it has enabled: opening up the transcriptome of nonmodel organisms and exposing the full complexity of the mammalian transcriptome much of which was hidden from microarrays.

The future will see the further integration of transcriptomics with other –omic technologies providing a more complete understanding of how individual cells and different tissue types are organized and controlled.

One thing that will not change is the fact that the analysis of the transcriptome by any technology is a challenging high dimensional biological problem and will remain so.

Conflict of interest

None declared.

Acknowledgements

PM is funded through a grant from Science Foundation Ireland (07/SRC/B1156). The author would like to thank Professor Alex Evans for very constructive criticism during the drafting of this review.

References

1. Pietu G, Mariage-Samson R, Fayein NA, Matingou C, Eveno E, Houlgatte R, Decraene C, Vandenbrouck Y, Tahi F, Devignes MD *et al.*: **The Genexpress IMAGE knowledge base of the human brain transcriptome: a prototype integrated resource for functional and computational genomics**. *Genome Res* 1999, **9**:195-209.

2. Velculescu VE, Zhang L, Zhou W, Vogelstein J, Basrai MA, Bassett DE Jr, Hieter P, Vogelstein B, Kinzler KW: **Characterization of the yeast transcriptome**. *Cell* 1997, **88**:243-251.

3. Eisen J: **Badomics words and the power and peril of the ome-meme**. *Gigascience* 2012, **1**:6.

4. Croucher NJ, Thomson NR: **Studying bacterial transcriptomes using RNA-seq**. *Curr Opin Microbiol* 2010, **13**:619-624.

5. Mader U, Nicolas P, Richard H, Bessieres P, Aymerich S: **Comprehensive identification and quantification of microbial transcriptomes by genome-wide unbiased methods**. *Curr Opin Biotechnol* 2011, **22**:32-41.

6. Mullen MP, Elia G, Hilliard M, Parr MH, Diskin MG, Evans AC, Crowe MA: **Proteomic characterization of histotroph during the preimplantation phase of the estrous cycle in cattle**. *J Proteome Res* 2012, **11**:3004-3018.

7. Forde N, Duffy GB, McGettigan PA, Browne JA, Mehta JP, Kelly AK, Mansouri-Attia N, Sandra O, Loftus BJ, Crowe MA *et al.*: **Evidence for an early endometrial response to pregnancy in cattle: both dependent upon and independent of interferon tau**. *Physiol Genomics* 2012, **44**:799-810.

8. Mamo S, Mehta JP, McGettigan P, Fair T, Spencer TE, Bazer FW, Lonergan P: **RNA sequencing reveals novel gene clusters in bovine conceptuses associated with maternal recognition of pregnancy and implantation**. *Biol Reprod* 2011, **85**:1143-1151.

9. Walsh SW, Mehta JP, McGettigan PA, Browne JA, Forde N, Alibrahim RM, Mulligan FJ, Loftus B, Crowe MA, Matthews D *et al.*: **Effect of the metabolic environment at key stages of follicle development in cattle: focus on steroid biosynthesis**. *Physiol Genomics* 2012, **44**:504-517.

10. Bender K, Walsh S, Evans AC, Fair T, Brennan L: **Metabolite concentrations in follicular fluid may explain differences in fertility between heifers and lactating cows**. *Reproduction* 2010, **139**:1047-1055.

11. Pluta K, McGettigan PA, Reid CJ, Browne JA, Irwin JA, Tharmalingam T, Corfield A, Baird AW, Loftus BJ, Evans AC *et al.*: **Molecular aspects of mucin biosynthesis and mucus formation in the bovine cervix during the periestrous period**. *Physiol Genomics* 2012, **44**:1165-1178.

12. Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF *et al.*: **Complementary DNA sequencing: expressed sequence tags and human genome project**. *Science* 1991, **252**:1651-1656.

13. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW: **Serial analysis of gene expression**. *Science* 1995, **270**:484-487.

14. Schena M, Shalon D, Davis RW, Brown PO: **Quantitative monitoring of gene expression patterns with a complementary DNA microarray**. *Science* 1995, **270**:467-470.

15. Bainbridge MN, Warren RL, Hirst M, Romanuik T, Zeng T, Go A, Delaney A, Griffith M, Hickenbotham M, Magrini V *et al.*: **Analysis of the prostate cancer cell line LNCaP transcriptome using a sequencing-by-synthesis approach**. *BMC Genomics* 2006, **7**:246.

16. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq**. *Nat Methods* 2008, **5**:621-628.

17. Sultan M, Schulz MH, Richard H, Magen A, Klingenhoff A, Scherf M, Seifert M, Borodina T, Soldatov A, Parkhomchuk D *et al.*: **A global view of gene activity and alternative splicing by deep sequencing of the human transcriptome**. *Science* 2008, **321**:956-960.

18. Wilhelm BT, Marguerat S, Watt S, Schubert F, Wood V, Goodhead I, Penkett CJ, Rogers J, Bahler J: **Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution**. *Nature* 2008, **453**:1239-1243.

19. Parkhomchuk D, Borodina T, Amstislavskiy V, Banaru M, Hallen L, Krobitsch S, Lehrach H, Soldatov A: **Transcriptome analysis by strand-specific sequencing of complementary DNA**. *Nucleic Acids Res* 2009, **37**:e123.

20. Katz Y, Wang ET, Airoldi EM, Burge CB: **Analysis and design of RNA sequencing experiments for identifying isoform regulation**. *Nat Methods* 2010, **7**:1009-1015.

21. Ozsolak F, Platt AR, Jones DR, Reifenberger JG, Sass LE, McInerney P, Thompson JF, Bowers J, Jarosz M, Milos PM: **Direct RNA sequencing**. *Nature* 2009, **461**:814-818.

22. Mamanova L, Andrews RM, James KD, Sheridan EM, Ellis PD, Langford CF, Ost TW, Collins JE, Turner DJ: **FRT-seq: amplification-free, strand-specific transcriptome sequencing**. *Nat Methods* 2010, **7**:130-132.

23. Li JW, Schmieder R, Ward RM, Delenick J, Olivares EC, Mittelman D: **SEQanswers: an open access community for collaboratively decoding genomes**. *Bioinformatics* 2012, **28**:1272-1273.

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

25. Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J: **SOAP2: an improved ultrafast tool for short read alignment**. *Bioinformatics* 2009, **25**:1966-1967.

26. Allison DB, Cui X, Page GP, Sabripour M: **Microarray data analysis: from disarray to consolidation and consensus**. *Nat Rev Genet* 2006, **7**:55-65.

27. 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**. *Nat Biotechnol* 2010, **28**:511-515.

28. Pepke S, Wold B, Mortazavi A: **Computation for ChIP-seq and RNA-seq studies**. *Nat Methods* 2009, **6**:S22-S32.

29. Risso D, Schwartz K, Sherlock G, Dudoit S: **GC-content normalization for RNA-Seq data**. *BMC Bioinformatics* 2011, **12**:480.

30. Hansen KD, Irizarry RA, Wu Z: **Removing technical variability in RNA-seq data using conditional quantile normalization**. *Biostatistics* 2012, **13**:204-216.

31. Zheng W, Chung LM, Zhao H: **Bias detection and correction in RNA-Sequencing data**. *BMC Bioinformatics* 2011, **12**:290.

32. 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.

33. Robinson MD, Oshlack A: **A scaling normalization method for differential expression analysis of RNA-seq data**. *Genome Biol* 2010, **11**:R25.

34. Li B, Ruotti V, Stewart RM, Thomson JA, Dewey CN: **RNA-Seq gene expression estimation with read mapping uncertainty**. *Bioinformatics* 2010, **26**:493-500.

35. Auer PL, Doerge RW: **Statistical design and analysis of RNA sequencing data**. *Genetics* 2010, **185**:405-416.

36. Robinson MD, McCarthy DJ, Smyth GK: **edgeR: a Bioconductor package for differential expression analysis of digital gene expression data**. *Bioinformatics* 2010, **26**:139-140.

37. Anders S, Huber W: **Differential expression analysis for sequence count data**. *Genome Biol* 2010, **11**:R106.

38. Hardcastle TJ, Kelly KA: **baySeq: empirical Bayesian methods for identifying differential expression in sequence count data**. *BMC Bioinformatics* 2010, **11**:422.

39. Di Y, Schafer DW, Cumbie JS, Chang JH: **The NBP negative binomial model for assessing differential gene expression from RNA-seq**. *Stat Appl Genet Mol Biol* 2011, **10**.

40. Tarazona S, Garcia-Alcalde F, Dopazo J, Ferrer A, Conesa A: **Differential expression in RNA-seq: a matter of depth**. *Genome Res* 2011, **21**:2213-2223.

41. Li J, Tibshirani R: **Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data**. *Stat Methods Med Res* 2011, **0**:1-18 http://dx.doi.org/10.1177/0962280211428386.

42. Li B, Dewey CN: **RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome**. *BMC Bioinformatics* 2011, **12**:323.

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

44. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**:2078-2079.

45. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP: **Integrative genomics viewer**. *Nat Biotechnol* 2011, **29**:24-26.

46. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: **Alternative isoform regulation in human tissue transcriptomes**. *Nature* 2008, **456**:470-476.

47. Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, Gnirke A, Regev A: **Comprehensive comparative analysis of strand-specific RNA sequencing methods**. *Nat Methods* 2010, **7**:709-715.

48. Anders S, Reyes A, Huber W: **Detecting differential usage of exons from RNA-seq data**. *Genome Res* 2012, **22**:2008-2017.

49. Blekhman R, Marioni JC, Zumbo P, Stephens M, Gilad Y: **Sex-specific and lineage-specific alternative splicing in primates**. *Genome Res* 2010, **20**:180-189.

50. Baker M: **Gene data to hit milestone**. *Nature* 2012, **487**:282-283.

51. Hsi-Yang Fritz M, Leinonen R, Cochrane G, Birney E: **Efficient storage of high throughput DNA sequencing data using reference-based compression**. *Genome Res* 2011, **21**:734-740.

52. Jones DC, Ruzzo WL, Peng X, Katze MG: **Compression of next-generation sequencing reads aided by highly efficient de novo assembly**. *Nucleic Acids Res* 2012, **40**:e171.

53. Li M, Wang IX, Li Y, Bruzel A, Richards AL, Toung JM, Cheung VG: **Widespread RNA and DNA sequence differences in the human transcriptome**. *Science* 2011, **333**:53-58.

54. Hayden EC: **RNA studies under fire**. *Nature* 2012, **484**:428.

55. Macarthur D: **Methods: Face up to false positives**. *Nature* 2012, **487**:427-428.

56. 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 Res* 2011, **21**:1728-1737.

57. Loven 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.

58. Schulz MH, Zerbino DR, Vingron M, Birney E: **Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels**. *Bioinformatics* 2012, **28**:1086-1092.

59. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q *et al.*: **Full-length transcriptome assembly from RNA-Seq data without a reference genome**. *Nat Biotechnol* 2011, **29**:644-652.

60. Robertson G, Schein J, Chiu R, Corbett R, Field M, Jackman SD, Mungall K, Lee S, Okada HM, Qian JQ *et al.*: **De novo assembly and analysis of RNA-seq data**. *Nat Methods* 2010, **7**:909-912.

61. Heber S, Alekseyev M, Sze SH, Tang H, Pevzner PA: **Splicing graphs and EST assembly problem**. *Bioinformatics* 2002, **18(Suppl 1)**:S181-S188.

62. Aronesty E: *ea-utils: ''Command-line Tools for Processing Biological Sequencing Data''*. 2011 http://code.google.com/p/ea-utils.

63. Brown CT, Howe A, Zhang Q, Pyrkosz AB, Brom TH: *A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data*. 2012:. arXiv:1203.4802.

64. Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A *et al.*: **mRNA-Seq whole-transcriptome analysis of a single cell**. *Nat Methods* 2009, **6**:377-382.

65. Tang F, Barbacioru C, Bao S, Lee C, Nordman E, Wang X, Lao K, Surani MA: **Tracing the derivation of embryonic stem cells from the inner cell mass by single-cell RNA-Seq analysis**. *Cell Stem Cell* 2010, **6**:468-478.

66. Qiu S, Luo S, Evgrafov O, Li R, Schroth GP, Levitt P, Knowles JA, Wang K: **Single-neuron RNA-Seq: technical feasibility and reproducibility**. *Front Genet* 2012, **3**.

67. Ramskold D, Luo S, Wang YC, Li R, Deng Q, Faridani OR, Daniels GA, Khrebtukova I, Loring JF, Laurent LC *et al.*: **Full-length mRNA-Seq from single-cell levels of RNA and individual circulating tumor cells**. *Nat Biotechnol* 2012, **30**:777-782.

68. Sanford JR, Wang X, Mort M, Vanduyn N, Cooper DN, Mooney SD, Edenberg HJ, Liu Y: **Splicing factor SFRS1 recognizes a functionally diverse landscape of RNA transcripts**. *Genome Res* 2009, **19**:381-394.

69. Ingolia NT, Ghaemmaghami S, Newman JR, Weissman JS: **Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling**. *Science* 2009, **324**:218-223.

70. Michel AM, Roy Choudhury K, Firth AE, Ingolia NT, Atkins JF, Baranov PV: **Observation of dually decoded regions of the human genome using ribosome profiling data**. *Genome Res* 2012, **22**:2219-2229.

71. Creighton CJ, Benham AL, Zhu H, Khan MF, Reid JG, Nagaraja AK, Fountain MD, Dziadek O, Han D, Ma L *et al.*: **Discovery of novel microRNAs in female reproductive tract using next generation sequencing**. *PLoS One* 2010, **5**:e9637.

72. Vogel C, Marcotte EM: **Insights into the regulation of protein abundance from proteomic and transcriptomic analyses**. *Nat Rev Genet* 2012, **13**:227-232.

73. **A user's guide to the encyclopedia of DNA elements (ENCODE)**. *PLoS Biol* 2011, **9**:e1001046.

74. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE *et al.*: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project**. *Nature* 2007, **447**:799-816.

75. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, Epstein CB, Frietze S, Harrow J, Kaul R *et al.*: **An integrated encyclopedia of DNA elements in the human genome**. *Nature* 2012, **489**:57-74.

76. Eddy S: **The C-value paradox, junk DNA, and ENCODE**. *Curr Biol* 2012, **22**:898-899.

77. Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S, Snyder M: **What is a gene, post-ENCODE? History and updated definition**. *Genome Res* 2007, **17**:669-681.

78. Jacquier A: **The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs**. *Nat Rev Genet* 2009, **10**:833-844.

79. Guttman M, Donaghey J, Carey BW, Garber M, Grenier JK, Munson G, Young G, Lucas AB, Ach R, Bruhn L *et al.*: **lincRNAs act in the circuitry controlling pluripotency and differentiation**. *Nature* 2011, **477**:295-300.

80. van Bakel H, Nislow C, Blencowe BJ, Hughes TR: **Most ''dark matter'' transcripts are associated with known genes**. *PLoS Biol* 2010, **8**:e1000371.

81. Struhl K: **Transcriptional noise and the fidelity of initiation by RNA polymerase II**. *Nat Struct Mol Biol* 2007, **14**:103-105.

82. Shen Y, Yue F, McCleary DF, Ye Z, Edsall L, Kuan S, Wagner U, Dixon J, Lee L, Lobanenkov VV *et al.*: **A map of the cis-regulatory sequences in the mouse genome**. *Nature* 2012, **488**:116-120.

83. Ernst J, Kheradpour P, Mikkelsen TS, Shoresh N, Ward LD, Epstein CB, Zhang X, Wang L, Issner R, Coyne M *et al.*: **Mapping and analysis of chromatin state dynamics in nine human cell types**. *Nature* 2011, **473**:43-49.

84. Koren S, Schatz MC, Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED *et al.*: **Hybrid error correction and de novo assembly of single-molecule sequencing reads**. *Nat Biotechnol* 2012, **30**:693-700.