Published in final edited form as:

Brain Res. 2010 June 18; 1338: 146–154. doi:10.1016/j.brainres.2010.03.039.

Deep sequencing of coding and non-coding RNA in the CNS

Marcel van der Brug^{1,2}, Michael A Nalls¹, and Mark R Cookson^{1,CA}

¹Laboratory of Neurogenetics, National Institute on Aging, National Institutes of Health, Bethesda, MD, 20892, USA

²Department of Neuroscience, The Scripps Research Institute, 130 Scripps Way, Jupiter, FL 33458, USA

Abstract

Several methods now exist for identifying and quantifying many biological events in parallel and in a relatively unbiased fashion. For gene expression experiments, cloning approaches have been supplemented with microarray platforms over the past few years. The focus of this review is on deep sequencing, a new set of techniques that can be used to both identify RNA species and quantify them in a massively parallel fashion. Deep sequencing has some advantages over other methods, driven largely by the high depth of coverage for any library of nucleic acids. This allows, for example, estimates of alternative splicing and untranslated region utilization. We will discuss how deep sequencing methods are being applied to characterization of gene expression in the brain and how these technologies might develop over the next few years.

Keywords

Alternative splicing; Deep sequencing; Gene expression; massively parallel signature sequencing; Next generation sequencing; non-coding RNA; small RNA

“Unbiased” approaches to RNA expression

Unbiased approaches to study biological questions have great promise in that they address a problem without a preconceived hypothesis. An example of an unbiased approach is finding genes by linkage analysis where, without any *a priori* knowledge of the gene other than understanding inheritance patterns, causal mutations can be found for any given phenotype, such as a human disease.

The same logic can be applied to other large datasets where an underlying biological effect can be mapped to changes at the genomic, protein or metabolite level. The excitement of this way of thinking is to identify truly novel changes in any given context, which is more powerful than arguing solely by analogy to known changes. This is particularly important in the brain, because although 30–50% of all known protein coding genes are expressed in the brain, about half of them do not have an assigned function (Diaz, 2009) and, therefore, novel hypotheses are always welcome.

^{CA}Address correspondence to: Mark R. Cookson, Cell Biology and Gene Expression Unit, Laboratory of Neurogenetics, National Institute on Aging, 35 Convent Drive, Bethesda MD 20892-3707, USA. Phone : +1 301 451 3870. Fax : +1 301 451 7295. cookson@mail.nih.gov.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

One unbiased approach that has been widely used over the past few years is to identify differences in gene expression profiles at the RNA level. Several cloning based strategies have been used to find genes that are differentially expressed between tissues or between treatments, including differential display, subtractive hybridization or SAGE (serial analysis of gene expression) (Colantuoni et al., 2000). SAGE and the related technique CAGE (cap analysis gene expression) both capture short sequence 'tags' representative of each transcript and by sequencing large numbers of cloned tags an estimate of relative abundance of large numbers of transcripts can be derived. SAGE uses pairs of restriction enzymes to fragment cDNA libraries prior to ligation of adaptors that can then be used to clone the short fragments whereas CAGE uses biotinylation and isolation of the 5' cap of mature mRNAs to capture fragments retaining the most 5' ~20 nucleotides from each mRNA, again then used for sequencing large numbers of clones in parallel.

These methods tend to be quite cumbersome, and have been supplemented in recent years by array based methods where RNA is amplified to make probes that are hybridized to oligonucleotide probes or cDNA. The pervasiveness of microarray technology is illustrated by recent reviews on use of microarrays to address problems in brain development (Diaz, 2009), brain tumors (Johnson et al., 2009), effects of mood stabilizing drugs (Fatemi et al., 2009), sleep (Naidoo, 2009) and diseases such as multiple sclerosis (Kinter et al., 2008), Alzheimer's disease (McShea et al., 2006) or Parkinson's disease (Miller and Federoff, 2006). Demonstrating the potential utility of such studies, some microarray 'hits' show promise for diagnostic applications in neurological disorders (Scherzer, 2009).

However, none of these approaches are genuinely unbiased. Most of these technologies focus on mature poly-adenylated mRNA species and miss non-coding RNAs (ncRNA) as well as alternately spliced and chimeric transcripts (Gingeras, 2009; Gustincich et al., 2006; Mercer et al., 2009). Neglecting these RNA species underestimates the complexity of the entire RNA complement of the cell. For example, there are many natural antisense ncRNAs that are conserved between human and mice (Carninci et al., 2005) but also human specific ncRNAs (e.g. HAR1F) primarily expressed in Cajal-Retzius neurons during development (Benjaminov et al., 2008; Pollard et al., 2006). Furthermore, array technologies use probes (either a complementary DNA [cDNA] or oligonucleotide based) and thus can only capture sequences that are expected, usually in a species-specific manner.

Underpinning the adoption of different types of high throughput methods is the development of technology to support them. The focus of this article is on one of these newer techniques, deep sequencing, where attempts are made to characterize and quantify all RNA species present within the CNS. We will discuss recent studies that have attempted to use deep sequencing in brain tissue, and outline the pros and cons of applying this technology to studies of CNS systems and disorders

Deep Sequencing methodologies, Pros and Cons

The various methodologies and platform specific techniques used in deep sequencing have been reviewed elsewhere (Mardis, 2008; Metzker, 2010; Shendure and Ji, 2008) but are worth revisiting briefly here to discuss how they can be applied to RNA analysis. Although there are technological differences between the available platforms, the key concept that distinguishes deep sequencing from traditional Sanger sequencing is its foundation on the generation and assembly of large numbers of short read sequences (Figure 1). The terminology of these techniques varies, with 'next generation sequencing' and 'massively parallel signature sequencing' (MPSS) (Brenner et al., 2000) also used. We prefer not to use 'next generation' as presumably this will be outdated once the next next generation is developed. MPSS refers

to tagging approaches (discussed below), which is only one of the several deep sequencing methods.

The first step in deep sequencing is the preparation of representative templates for each sample. The two primary methods involve either clonally amplified fragments (454, Illumina, SOLiD3) or direct sequencing of RNA (Helicos). In most of the current technologies, individual molecules are immobilized on a support (a multiwell plate or a flowcell) such that the sequence of bases can be imaged. Thus, each representative molecule in the library gives a specific sequence distinct from its neighbors.

Five major platforms fill the currently deep sequencing space: Roche/454, Illumina/Solexa, Life/SOLiD3, Helicos/HeliScope and the open source Polonator. In each of these deep sequencing methods, each read is relatively short compared to conventional capillary-based methods. Currently, 26–330 nucleotides reads are typical, depending on the technology used (Metzker, 2010). It is also worth noting that the pace of innovation on each platform is rapid, and the fidelity and read length improves with each iteration. The potential limitation of short reads is partially overcome by sheer numbers – from hundreds of thousands to many millions. This yields anywhere from 0.5–50 gigabases per sequencing run, which can take machine run time to over a week. For RNA analysis, sequences are then used to identify known and novel non-coding RNA species, and quantitate the numbers of sequences found across samples.

There are several variations on this technique that give slightly different information on specific parts of the RNA population in a given sample. Small RNA analysis may use polyacrylamide gel electrophoresis (PAGE) fractionation to isolate microRNA (miRNA), Piwi-interacting RNA (piRNA) and other small ncRNA species from other RNA fragments either prior to, or after library preparation. Because these species have a size range of ~19–32 nucleotides, the ideal size for short read sequencing, they can be ligated with adaptors and used directly for reverse transcription and amplification. In mRNA sequencing (RNA-seq), larger polyadenylated mRNA species are isolated, usually using an oligo-dT based capture method prior to entering the sequencing protocol. Importantly, those RNAs (possibly many non-coding RNAs) without a polyA tail will be excluded from these sequencing experiments. This is an unfortunate limitation of the technology as it stands, as the primary reason for both of these selection procedures is to reduce the content of ribosomal RNA in the pool of RNAs to be sequenced, which if not removed, would make up the bulk of all sequenced fragments.

Deep sequencing can also be combined with various approaches that involve generation of libraries of short tags from each RNA species, including SAGE (Hanriot et al., 2008) or CAGE (de Hoon and Hayashizaki, 2008). This is sometimes referred to as digital gene expression or DGE. Although the sequence complexity along the entire RNA is not captured, SAGE and CAGE based approaches should robustly estimate abundance of RNA species as each sequenced tag is of a similar length and thus should correlate with overall abundance of that transcript, irrespective of splicing or editing. Deep sequencing can also be used to interrogate protein-nucleic acid interactions. For example, short DNA fragments isolated using chromatin-immunoprecipitation can be used as to generate libraries of an appropriate size for several sequencing technologies, giving identities of the interacting sequences as well as abundance (Mardis, 2008). There are therefore different methods for sequence and abundance estimation for different pools of RNA in the cell with some alternative approaches for approach to quantitation.

Given the complexity and quantity of the data generated it is perhaps unsurprising that bioinformatics of this type of data is critical. Analysis of deep sequencing data begins with base-calling using proprietary software for each platform. After base calling, reads are typically aggregated into text files of various loosely standardized formats containing strings of bases

as well as quality scores for each base. Initial steps of analyses consist of extracting high quality reads or pairs of reads per sample/experiment, permitting quantification of experimental success by comparing the proportion of reads passing pre-defined quality control (QC) thresholds. In some methods, common contaminating sequences (such as mitochondrial and ribosomal DNAs, as well as sequences associated with adaptor sequences) are removed prior to QC steps. This is particularly useful as these abundant sequences and artifacts can impact the normalization of expression levels.

Base pair reads are then aligned to reference transcriptome sequences using one of several available software packages (Pepke et al., 2009), usually incorporating several quality control steps (Lao et al., 2009). The abundance of reads per exon or gene is most often normalized to number of reads per kilobase per million reads (RPKM (Mortazavi et al., 2008)). This facilitates the ranking of individual transcripts, which then represents the expression 'level' of each transcript. Additionally, multiple software packages capable of mapping paired reads can identify novel spliced RNA species by aligning these reads to reference sets of exon-exon junctions and alignments to regions outside of the canonical reference transcriptome, in principal allowing quantitation of alternate splicing and chimeric transcripts missing from conventional array platforms.

Having outlined how one might perform an experiment, what then are the advantages and disadvantages of deep sequencing compared to other methods such as microarrays? In principle, because the raw data counts individual transcripts, estimates of abundance of RNA are more accurate than for hybridization-based methods, where signal intensity is proportional both to RNA input and hybridization strength. This is particularly important for low abundance transcripts that, in conventional arrays, are close enough to background hybridization values to be discarded. Assuming no cross-contamination between samples, if a single transcript is sequenced in a deep sequencing experiment it can be considered to have been definitively 'present' in the original library. Whether one can definitively exclude a transcript based on non-appearance is less clear and thus comparing low abundance transcripts is still a question of statistical reliability of rare events.

The other major advantage of deep sequencing is that it is closer to being truly unbiased by virtue of being agnostic to the underlying sequence. Because there are no sequences against which to hybridize, one could compare libraries from widely divergent organisms on the same platform, something that would result in a great deal of discarded data using conventional arrays. Early data suggests that differential gene expression estimates using deep sequencing seem to be less influenced by GC content (which varies across genes and species) than other methods (t Hoen et al., 2008).

Perhaps more relevant for many studies, deep sequencing is a true discovery platform in that both known and unknown sequences are represented and this has been particularly helpful in discovering new small RNA and ncRNA sequences. Large genomic regions, and perhaps even the majority of the human genome, have no apparent protein-coding potential. Analysis of the ENCODE project, which aims to identify all functional elements in the human genome (Birney et al., 2007), suggests that ncRNA are expressed from many genomic loci (Wu et al., 2008). Such ncRNA molecules are not widely interrogated on conventional microarrays but will be present in RNA-seq libraries if they are polyadenylated.

An obvious significant disadvantage is cost – each library and sequencing run is several times the investment required for conventional arrays. Furthermore, the time investment is significantly higher both in bench time and analysis. One outcome of this is that experiments are often rather limited, with insufficient biological replicates included (a common problem with early microarray studies). This of course is somewhat of a transient concern as the much

publicized decrease in sequencing costs will steer us towards a sub 1000 dollar genome within the next few years. Concomitant with this is the decreasing costs of microarrays, and it seems likely that there will always be a gap in cost between the two technologies. Rather than view deep sequencing as rendering arrays redundant, it is more likely that they will fulfill complementary roles within a laboratory setting as discovery and validation platforms.

The informatics involved also provide significant challenges, as has been discussed elsewhere (Nobuta et al.; Wang et al., 2009). A truly dedicated sequencing pipeline is needed and for many key steps there are not yet well defined standards (e.g. for normalization) and the issue of error rates in sequencing needs to be addressed for most platforms.

Further considerations should also be given to sequencing biases. All the platforms discussed so far with the exception of the Helicos platform use a reverse transcription and amplification step, which like any PCR reaction favors CG poor sequence stretches. Conversely, the single molecule approach is prone to lower throughputs and higher error rates. It is worth noting however that these concerns largely apply to the manufacture provided protocols as they exist today and there are increasing numbers of user derived sequencing protocols that address them, such as an amplification free protocol for the Illumina platform (Mamanova et al., 2010). Apart from platform dependant technological artifacts there are specific problems that arise from short read lengths. For example, ncRNAs often have a mix of exons located within intronic segments that also overlap with exonic regions of protein coding genes on the opposing strand. This ncRNA/mRNA transcriptional complexity may be further compounded if both the mRNA and paired ncRNA may be alternately spliced. Such complexities are not well suited for resolution by short read sequencing. Similar considerations occur with deconvoluting alternate splicing events from RNA-seq data and for unambiguously identifying chimeric transcripts. As libraries are randomly prepared, the proportion of a sequence read that sits neatly and evenly across any exon-exon boundary is relatively low and therefore subject to ambiguity when aligning to a reference genome. Longer read length may help solve some of these problems by providing higher specificity for alignments, perhaps with some decrease in sensitivity, and several deep sequencing platforms are actively developing protocols for longer reads.

One final limitation applies to the brain more so than other tissue types, which is that deep sequencing currently requires a fairly significant amounts of starting template RNA (e.g., 1 µg for mRNA sequencing on the Solexa platform). This contrasts to conventional arrays, which can be performed with nanogram amounts of input RNA. Starting material is not a routine problem for bulk samples or homogenous cell lines but does represent a limitation for a complex organ such as the brain where cellular heterogeneity can be a significant confound in many comparisons and may mask detection of transcripts found only in some cell types. There are some recent publications starting from single cells, specifically blastocysts showing that deep sequencing methods can be adapted down to small numbers of cells (Lao et al., 2009; Tang et al., 2009). There is therefore hope that deep sequencing might eventually be applied even to heterogenous tissues such as the brain.

Deep sequencing the brain

At the time of writing there are only a few studies that have applied deep sequencing technology to neurological tissue and so the major conclusions of each can be discussed here in some depth.

Two papers published in 2008 used tagging approaches to generate libraries, in methods that are derived from LongSAGE, a serial analysis of gene expression variant that captures 17–21 base pair sequences from each RNA transcript. Hanriot et al compared libraries of RNA extracted from the hypothalamus using a SAGE technique with conventional sequencing and a modified SAGE combined with deep sequencing, in this case using a Illumina/Solexa genome

analyzer (Hanriot et al., 2008). The two techniques gave largely similar results, reliably finding hypothalamic transcripts such as pro-melanin converting hormone, prehypocretin and pro-dynorphin and showing that in both cases, the relative numbers of tags of each correlated roughly with estimates of copy number by quantitative real-time PCR (qRT-PCR). However, using deep sequencing provided deeper library coverage and increased numbers of tags compared to conventional sequencing, and thus potentially captures a greater depth of the complexity of the poly-adenylated RNA population in this brain region. There were some tags that were only identified with one technique rather than the other, but comparison of rare sequences were complicated because the libraries generated by the two techniques were from different mouse strains and thus different individuals and different dissections. However, overall, the data suggests that deep sequencing of a tag library performs at least as well as SAGE with conventional sequencing with substantial improvement in throughput.

A similar, but more extensive study, of the hippocampus also supported the concept that deep sequencing based DGE is robust (t Hoen et al., 2008). Hippocampal RNA libraries were prepared from mice transgenic for the short isoform of doublecortin-like kinase-1 (δ C-DLCK1) and wild type controls from the same C57/Bl6 background (male mice were used). DGE using deep sequencing showed good inter-laboratory reliability ($r^2 > 0.99$), had a dynamic range of 4–5 orders of magnitude and revealed several differentially expressed genes between the two groups of mice even with relatively low numbers of biological replicates ($n=4/\text{group}$). These authors also noted how efficient the two approaches were, estimating that this scale of library would take about a year using conventional techniques compared to 3 days for deep sequencing which also gave a 60-fold improvement in library coverage, all driven by the massively parallel sequencing achieved.

Perhaps more interestingly, t Hoen et al also compared deep sequencing with several widely used microarray platforms. A major advantage of deep sequencing was the identification of alternate poly-adenylated species and natural antisense transcripts, which were not interrogated on arrays. The important observation here, though, was the number of such species – 51% of all genes had a natural antisense transcript and in a few cases the relative abundance was greater than that of the coding transcript. Deep sequencing also performed well for low abundance transcripts, which are notoriously difficult to validate in (and often discarded from) conventional microarrays as they are close to the noise background in hybridization based techniques. DGE using deep sequencing also showed higher inter-laboratory reliability than conventional arrays. This is a little surprising for a new technology where *a priori* one might expect more variation due to learning new processes. t Hoen et al propose that hybridization protocols are difficult to standardize from one laboratory to another, but it will be interesting to see if this experience generalizes to other studies and across deep sequencing platforms.

There were some ways in which hybridization based microarrays had better apparent performance than DGE and deep sequencing. In particular, qRT-PCR, which is widely used as a validation technique, correlated better with arrays than with deep sequencing. de Hoon et al suggest that this is because hybridization arrays captures groups of transcripts with common sequences (e.g. all splice forms) as does qRT-PCR with a single primer set whereas DGE will split these sequences into individual forms. These types of issues need to be interrogated further in other datasets and indicate that for many current applications microarrays will still provide an important measurement for some time to come.

The issue of alternate splicing, and utilization of different poly-adenylation signals in a switch-like manner, is highlighted by RNA-seq studies from various tissues, including brain (Mortazavi et al., 2008; Mudge et al., 2008; Wang et al., 2008). A computational analysis of the transcriptional diversity in different tissues estimated using RNA-seq suggests that brain tissue has a particularly complex set of RNA species (Ramskold et al., 2009). Most genes are

expressed at detectable levels in the brain and ubiquitously expressed genes account for a lower proportion of all expression than in other tissues. Another difference between brain and other tissues is that brain-enriched genes, including those for axon guidance, tend to have longer 3'-UTR sequences than other groups of genes. Inclusion or exclusion of 3'-UTR sequences, and hence potential alternate poly-adenylation signals, changes estimates of transcript abundance. Interestingly, inclusion of only coding exons gives more stable estimates of abundance that correlate with qRT-PCR measures, perhaps consistent with the work of de Hoon et al. Collectively, these results and analyses suggest that the brain has a complex transcribed genome and that RNA-seq might therefore be helpful in examining brain gene expression in different circumstances.

Both small RNAs (miRNAs etc) and long non-coding RNAs are also expressed in the brain and are hypothesized to have important functional roles (Kuss and Chen, 2008; Ponjavic et al., 2009; St Laurent et al., 2009). Furthermore, the small RNA complement of brain seems to be quite different from other tissues, based on libraries of miRNAs (Landgraf et al., 2007).

To date, there have been relatively few reports of using deep sequencing for small RNA or other ncRNAs in brain. One recent study used small RNA libraries prepared from a single rat brain to compare sequencing platforms including conventional and deep sequencing approaches (Linsen et al., 2009). An interesting observation was that library preparation methods can introduce bias into the estimates of absolute miRNA quantitation. This was largely platform independent, suggesting that this is not a problem with sequencing *per se*, but reflects inherent difficulties in isolating and tagging small RNA molecules. However, estimates of differences in miRNA abundance between samples in deep sequenced libraries matched well with qRT-PCR, suggesting that these biases are not a fatal problem with using deep sequencing for relative quantification of small RNA. Also, and as for RNA-seq, this method allows discovery of all sequences, thus allowing one to find novel small RNA sequences.

An important extension to deep sequencing techniques is to be able to interrogate protein-RNA interactions. High throughput cross-linking immunoprecipitation has been used to identify RNA bound to the neuronal splicing factor Nova (Licatalosi et al., 2008). Many short pieces of RNA were recovered in this set of experiments that were then mapped to the mouse genome. The ability of deep sequencing to provide quantitation of the numbers of tags recovered was important in this study as it allowed Licatalosi et al to show that large numbers of Nova tags were recovered in the 3' UTR of genes, which then lead to the biological insight that Nova regulates alternative polyadenylation in the brain, which had not previously been noted. A similar approach was used to find the miRNA species bound to Argonaute 1 in the mouse brain (Chi et al., 2009). Because of the high specificity of sequenced miRNA tags, Chi et al were able to produce a detailed mRNA:miRNA map that could be applied to cell lines or to brain. Another excellent examination of the non-coding RNA landscape and the proteins that interact with them is the study by Granneman et al. Here they examined the Nop1, Nop56, Nop58, and Rrp9 proteins by UV crosslinking and deep sequencing of the interacting RNA to reveal binding sites within U3 snoRNA and pre-rRNA species (Granneman et al., 2009).

The future

At the time of writing, several large-scale projects are underway to characterize the total RNA component of various tissues including the brain using deep sequencing. For example, gene expression omnibus (<http://www.ncbi.nlm.nih.gov/geo>) contains several tagging approaches including human and mouse brain that can be downloaded and analyzed. Our expectation is that many similar datasets will be generated over the next few years, giving a much deeper view of the complexities of RNAs found in neuronal tissue. This is evidenced by the Genotype-Tissue Expression (GTEx) project (<http://nihroadmap.nih.gov/GTEx/>), which will no doubt

feature quantitation by RNA sequencing as the database grows. The lessons learnt from these programs and projects of course, also apply to deep sequencing of RNA - quality control of tissues and RNAs is critical to generating data that is both accurate and valid for comparison to thousands of other samples. As costs decrease, one would imagine that throughput will increase, analogously to the large scale sequencing of the 1000 genomes project (<http://www.1000genomes.org>) compared to the few individuals sequenced in the original human genome projects.

As sequencing technology becomes further established, one will have to ask where it is best applied. Clearly, the list of places where microarrays have added data and new hypotheses is an obvious start, and therefore should include studies of normal development of the brain, brain diseases and pharmacological and genetic manipulations. In some cases, the extra information gained by deep sequencing may be helpful, for example in interrogating alternate splicing or RNA editing although caution will be needed to ensure that sufficient numbers of biological replicates are included and that validation of the specific transcripts. Areas where RNA sequencing clearly shines in the discovery of new transcripts, not represented on arrays or in cDNA databases. This of course, true for much of the transcriptome represented by non-coding RNAs.

One continuing problem with working with brain tissue is the issue of cellularity. If the RNA complement of a tissue is more complex than previously realized, one would predict that the RNA complement of a terminally differentiated neuron would be more complex still. Deep sequencing may help somewhat in being able to identify rare molecules characteristic of small numbers of cells in a complex tissue but probably won't answer quantitative issues especially where samples vary in their cellular composition. Improvements in library preparation methods from smaller amounts of inputs are probably needed to address this problem.

One emerging technology that may be especially helpful for complex tissue is single molecule sequencing (Ozsolak et al., 2009). Although this has not yet been applied to brain, there are two potential advantages over current approaches. First, there is no cDNA synthesis step, which may limit some of the inevitable biases that occur in library generation discussed above. Second, the amount of starting material needed is in the femtomole (~2ng) range, which may be feasible for laser captured samples from small groups of neurons.

Overall, these considerations suggest that deep sequencing, in various different forms and flavors, will likely be an important addition to the toolkit used for interrogating RNA abundance and identity in the brain over the coming years. Further technological and analytic developments are probably needed to fully exploit sequencing technologies but these are active areas of research.

Abbreviations

CAGE	cap analysis gene expression
cDNA	complementary DNA
miRNA	microRNA
piRNA	Piwi-interacting RNA
MPSS	massively parallel signature sequencing
ncRNA	non-coding RNA
PAGE	polyacrylamide gel electrophoresis
qRT-PCR	quantitative real-time PCR

RPKM	reads per kilobase per million reads
SAGE	serial analysis of gene expression
UTR	untranslated region

Acknowledgments

This research was supported in part by the Intramural Research Program of the NIH, National Institute on Aging.

References

- Benjaminov A, Westhof E, Krol A. Distinctive structures between chimpanzee and human in a brain noncoding RNA. *RNA* 2008;14:1270–1275. [PubMed: 18511501]
- Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P, Boyle PJ, Cao H, Carter NP, Clelland GK, Davis S, Day N, Dhami P, Dillon SC, Dorschner MO, Fiegler H, Giresi PG, Goldy J, Hawrylycz M, Haydock A, Humbert R, James KD, Johnson BE, Johnson EM, Frum TT, Rosenzweig ER, Karnani N, Lee K, Lefebvre GC, Navas PA, Neri F, Parker SC, Sabo PJ, Sandstrom R, Shafer A, Vetric D, Weaver M, Wilcox S, Yu M, Collins FS, Dekker J, Lieb JD, Tullius TD, Crawford GE, Sunyaev S, Noble WS, Dunham I, Denoeud F, Reymond A, Kapranov P, Rozowsky J, Zheng D, Castelo R, Frankish A, Harrow J, Ghosh S, Sandelin A, Hofacker IL, Baertsch R, Keefe D, Dike S, Cheng J, Hirsch HA, Sekinger EA, Lagarde J, Abril JF, Shahab A, Flamm C, Fried C, Hackermuller J, Hertel J, Lindemeyer M, Missal K, Tanzer A, Washietl S, Korbel J, Emanuelsson O, Pedersen JS, Holroyd N, Taylor R, Swarbreck D, Matthews N, Dickson MC, Thomas DJ, Weirauch MT, Gilbert J, Drenkow J, Bell I, Zhao X, Srinivasan KG, Sung WK, Ooi HS, Chiu KP, Foissac S, Alioto T, Brent M, Pachter L, Tress ML, Valencia A, Choo SW, Choo CY, Ucla C, Manzano C, Wyss C, Cheung E, Clark TG, Brown JB, Ganesh M, Patel S, Tammana H, Chrast J, Henrichsen CN, Kai C, Kawai J, Nagalakshmi U, Wu J, Lian Z, Lian J, Newburger P, Zhang X, Bickel P, Mattick JS, Carninci P, Hayashizaki Y, Weissman S, Hubbard T, Myers RM, Rogers J, Stadler PF, Lowe TM, Wei CL, Ruan Y, Struhl K, Gerstein M, Antonarakis SE, Fu Y, Green ED, Karaoz U, Siepel A, Taylor J, Liefer LA, Wetterstrand KA, Good PJ, Feingold EA, Guyer MS, Cooper GM, Asimenos G, Dewey CN, Hou M, Nikolaev S, Montoya-Burgos JI, Loytynoja A, Whelan S, Pardi F, Massingham T, Huang H, Zhang NR, Holmes I, Mullikin JC, Ureta-Vidal A, Paten B, Sereinghaus M, Church D, Rosenbloom K, Kent WJ, Stone EA, Batzoglou S, Goldman N, Hardison RC, Haussler D, Miller W, Sidow A, Trinklein ND, Zhang ZD, Barrera L, Stuart R, King DC, Ameur A, Enroth S, Bieda MC, Kim J, Bhinge AA, Jiang N, Liu J, Yao F, Vega VB, Lee CW, Ng P, Yang A, Moqtaderi Z, Zhu Z, Xu X, Squazzo S, Oberley MJ, Inman D, Singer MA, Richmond TA, Munn KJ, Rada-Iglesias A, Wallerman O, Komorowski J, Fowler JC, Couttet P, Bruce AW, Dovey OM, Ellis PD, Langford CF, Nix DA, Euskirchen G, Hartman S, Urban AE, Kraus P, Van Calcar S, Heintzman N, Kim TH, Wang K, Qu C, Hon G, Luna R, Glass CK, Rosenfeld MG, Aldred SF, Cooper SJ, Halees A, Lin JM, Shulha HP, Xu M, Haidar JN, Yu Y, Iyer VR, Green RD, Wadelius C, Farnham PJ, Ren B, Harte RA, Hinrichs AS, Trumbower H, Clawson H, Hillman-Jackson J, Zweig AS, Smith K, Thakkapallayil A, Barber G, Kuhn RM, Karolchik D, Armengol L, Bird CP, de Bakker PI, Kern AD, Lopez-Bigas N, Martin JD, Stranger BE, Woodroffe A, Davydov E, Dimas A, Eyraas E, Hallgrimsdottir IB, Huppert J, Zody MC, Abecasis GR, Estivill X, Bouffard GG, Guan X, Hansen NF, Idol JR, Maduro VV, Maskeri B, McDowell JC, Park M, Thomas PJ, Young AC, Blakesley RW, Muzny DM, Sodergren E, Wheeler DA, Worley KC, Jiang H, Weinstock GM, Gibbs RA, Graves T, Fulton R, Mardis ER, Wilson RK, Clamp M, Cuff J, Gnerre S, Jaffe DB, Chang JL, Lindblad-Toh K, Lander ES, Koriabine M, Nefedov M, Osoegawa K, Yoshinaga Y, Zhu B, de Jong PJ. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* 2007;447:799–816. [PubMed: 17571346]
- Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M, Roth R, George D, Eletr S, Albrecht G, Vermaas E, Williams SR, Moon K, Burcham T, Pallas M, DuBridge RB, Kirchner J, Fearon K, Mao J, Corcoran K. Gene expression analysis by massively

parallel signature sequencing (MPSS) on microbead arrays. *Nat Biotechnol* 2000;18:630–634. [PubMed: 10835600]

- Carninci P, Kasukawa T, Katayama S, Gough J, Frith MC, Maeda N, Oyama R, Ravasi T, Lenhard B, Wells C, Kodzius R, Shimokawa K, Bajic VB, Brenner SE, Batalov S, Forrest AR, Zavolan M, Davis MJ, Wilming LG, Aidinis V, Allen JE, Ambesi-Impimbato A, Apweiler R, Aturaliya RN, Bailey TL, Bansal M, Baxter L, Beisel KW, Bersano T, Bono H, Chalk AM, Chiu KP, Choudhary V, Christoffels A, Clutterbuck DR, Crowe ML, Dalla E, Dalrymple BP, de Bono B, Della Gatta G, di Bernardo D, Down T, Engstrom P, Fagiolini M, Faulkner G, Fletcher CF, Fukushima T, Furuno M, Futaki S, Gariboldi M, Georgii-Hemming P, Gingeras TR, Gojobori T, Green RE, Gustincich S, Harbers M, Hayashi Y, Hensch TK, Hirokawa N, Hill D, Huminiecki L, Iacono M, Ikeo K, Iwama A, Ishikawa T, Jakt M, Kanapin A, Katoh M, Kawasawa Y, Kelso J, Kitamura H, Kitano H, Kollias G, Krishnan SP, Kruger A, Kummerfeld SK, Kurochkin IV, Lareau LF, Lazarevic D, Lipovich L, Liu J, Liuni S, McWilliam S, Madan Babu M, Madera M, Marchionni L, Matsuda H, Matsuzawa S, Miki H, Mignone F, Miyake S, Morris K, Mottagui-Tabar S, Mulder N, Nakano N, Nakauchi H, Ng P, Nilsson R, Nishiguchi S, Nishikawa S, Nori F, Ohara O, Okazaki Y, Orlando V, Pang KC, Pavan WJ, Pavesi G, Pesole G, Petrovsky N, Piazza S, Reed J, Reid JF, Ring BZ, Ringwald M, Rost B, Ruan Y, Salzberg SL, Sandelin A, Schneider C, Schonbach C, Sekiguchi K, Semple CA, Seno S, Sessa L, Sheng Y, Shibata Y, Shimada H, Shimada K, Silva D, Sinclair B, Sperling S, Stupka E, Sugiura K, Sultana R, Takenaka Y, Taki K, Tammoja K, Tan SL, Tang S, Taylor MS, Tegner J, Teichmann SA, Ueda HR, van Nimwegen E, Verardo R, Wei CL, Yagi K, Yamanishi H, Zabarovsky E, Zhu S, Zimmer A, Hide W, Bult C, Grimmond SM, Teasdale RD, Liu ET, Brusic V, Quackenbush J, Wahlestedt C, Mattick JS, Hume DA, Kai C, Sasaki D, Tomaru Y, Fukuda S, Kanamori-Katayama M, Suzuki M, Aoki J, Arakawa T, Iida J, Imamura K, Itoh M, Kato T, Kawaji H, Kawagashira N, Kawashima T, Kojima M, Kondo S, Konno H, Nakano K, Ninomiya N, Nishio T, Okada M, Plessy C, Shibata K, Shiraki T, Suzuki S, Tagami M, Waki K, Watahiki A, Okamura-Oho Y, Suzuki H, Kawai J, Hayashizaki Y. The transcriptional landscape of the mammalian genome. *Science* 2005;309:1559–1563. [PubMed: 16141072]
- Chi SW, Zang JB, Mele A, Darnell RB. Argonaute HITS-CLIP decodes microRNA-mRNA interaction maps. *Nature* 2009;460:479–486. [PubMed: 19536157]
- Colantuoni C, Purcell AE, Bouton CM, Pevsner J. High throughput analysis of gene expression in the human brain. *J Neurosci Res* 2000;59:1–10. [PubMed: 10658179]
- de Hoon M, Hayashizaki Y. Deep cap analysis gene expression (CAGE): genome-wide identification of promoters, quantification of their expression, and network inference. *Biotechniques* 2008;44:627–628. 630, 632. [PubMed: 18474037]
- Diaz E. From microarrays to mechanisms of brain development and function. *Biochem Biophys Res Commun* 2009;385:129–131. [PubMed: 19460360]
- Fatemi SH, Reutiman TJ, Folsom TD. The role of lithium in modulation of brain genes: relevance for aetiology and treatment of bipolar disorder. *Biochem Soc Trans* 2009;37:1090–1095. [PubMed: 19754458]
- Gingeras TR. Implications of chimaeric non-co-linear transcripts. *Nature* 2009;461:206–211. [PubMed: 19741701]
- Granneman S, Kudla G, Petfalski E, Tollervey D. Identification of protein binding sites on U3 snoRNA and pre-rRNA by UV cross-linking and high-throughput analysis of cDNAs. *Proc Natl Acad Sci U S A* 2009;106:9613–9618. [PubMed: 19482942]
- Gustincich S, Sandelin A, Plessy C, Katayama S, Simone R, Lazarevic D, Hayashizaki Y, Carninci P. The complexity of the mammalian transcriptome. *J Physiol* 2006;575:321–332. [PubMed: 16857706]
- Hanriot L, Keime C, Gay N, Faure C, Dossat C, Wincker P, Scote-Blachon C, Peyron C, Gandrillon O. A combination of LongSAGE with Solexa sequencing is well suited to explore the depth and the complexity of transcriptome. *BMC Genomics* 2008;9:418. [PubMed: 18796152]
- Johnson R, Wright KD, Gilbertson RJ. Molecular profiling of pediatric brain tumors: insight into biology and treatment. *Curr Oncol Rep* 2009;11:68–72. [PubMed: 19080744]
- Kinter J, Zeis T, Schaeren-Wiemers N. RNA profiling of MS brain tissues. *Int MS J* 2008;15:51–58. [PubMed: 18782500]
- Kuss AW, Chen W. MicroRNAs in brain function and disease. *Curr Neurol Neurosci Rep* 2008;8:190–197. [PubMed: 18541114]

- Landgraf P, Rusu M, Sheridan R, Sewer A, Iovino N, Aravin A, Pfeffer S, Rice A, Kamphorst AO, Landthaler M, Lin C, Socci ND, Hermida L, Fulci V, Chiaretti S, Foa R, Schliwka J, Fuchs U, Novosel A, Muller RU, Schermer B, Bissels U, Inman J, Phan Q, Chien M, Weir DB, Choksi R, De Vita G, Frezzetti D, Trompeter HI, Hornung V, Teng G, Hartmann G, Palkovits M, Di Lauro R, Wernet P, Macino G, Rogler CE, Nagle JW, Ju J, Papavasiliou FN, Benzing T, Lichter P, Tam W, Brownstein MJ, Bosio A, Borkhardt A, Russo JJ, Sander C, Zavolan M, Tuschl T. A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell* 2007;129:1401–1414. [PubMed: 17604727]
- Lao KQ, Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Tuch B, Bodeau J, Siddiqui A, Surani MA. mRNA-sequencing whole transcriptome analysis of a single cell on the SOLiD system. *J Biomol Tech* 2009;20:266–271. [PubMed: 19949700]
- Licatalosi DD, Mele A, Fak JJ, Ule J, Kayikci M, Chi SW, Clark TA, Schweitzer AC, Blume JE, Wang X, Darnell JC, Darnell RB. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* 2008;456:464–469. [PubMed: 18978773]
- Linsen SE, de Wit E, Janssens G, Heater S, Chapman L, Parkin RK, Fritz B, Wyman SK, de Bruijn E, Voest EE, Kuersten S, Tewari M, Cuppen E. Limitations and possibilities of small RNA digital gene expression profiling. *Nat Methods* 2009;6:474–476. [PubMed: 19564845]
- 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. [PubMed: 20081834]
- Mardis ER. Next-generation DNA sequencing methods. *Annu Rev Genomics Hum Genet* 2008;9:387–402. [PubMed: 18576944]
- McShea A, Marlatt MW, Lee HG, Tarkowsky SM, Smit M, Smith MA. The application of microarray technology to neuropathology: cutting edge tool with clinical diagnostics potential or too much information? *J Neuropathol Exp Neurol* 2006;65:1031–1039. [PubMed: 17086099]
- Mercer TR, Dinger ME, Mattick JS. Long non-coding RNAs: insights into functions. *Nat Rev Genet* 2009;10:155–159. [PubMed: 19188922]
- Metzker ML. Sequencing technologies - the next generation. *Nat Rev Genet* 2010;11:31–46. [PubMed: 19997069]
- Miller RM, Federoff HJ. Microarrays in Parkinson's disease: a systematic approach. *NeuroRx* 2006;3:319–326. [PubMed: 16815215]
- Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods* 2008;5:621–628. [PubMed: 18516045]
- Mudge J, Miller NA, Khrebtukova I, Lindquist IE, May GD, Huntley JJ, Luo S, Zhang L, van Velkinburgh JC, Farmer AD, Lewis S, Beavis WD, Schilkey FD, Virk SM, Black CF, Myers MK, Mader LC, Langley RJ, Utsey JP, Kim RW, Roberts RC, Khalsa SK, Garcia M, Ambriz-Griffith V, Harlan R, Czika W, Martin S, Wolfinger RD, Perrone-Bizzozero NI, Schroth GP, Kingsmore SF. Genomic convergence analysis of schizophrenia: mRNA sequencing reveals altered synaptic vesicular transport in post-mortem cerebellum. *PLoS One* 2008;3:e3625. [PubMed: 18985160]
- Naidoo N. Cellular stress/the unfolded protein response: relevance to sleep and sleep disorders. *Sleep Med Rev* 2009;13:195–204. [PubMed: 19329340]
- Nobuta K, McCormick K, Nakano M, Meyers BC. Bioinformatics analysis of small RNAs in plants using next generation sequencing technologies. *Methods Mol Biol* 592:89–106.
- Ozsolak F, Platt AR, Jones DR, Reifenger JG, Sass LE, McInerney P, Thompson JF, Bowers J, Jarosz M, Milos PM. Direct RNA sequencing. *Nature* 2009;461:814–818. [PubMed: 19776739]
- Pepke S, Wold B, Mortazavi A. Computation for ChIP-seq and RNA-seq studies. *Nat Methods* 2009;6:S22–S32. [PubMed: 19844228]
- Pollard KS, Salama SR, Lambert N, Lambot MA, Coppens S, Pedersen JS, Katzman S, King B, Onodera C, Siepel A, Kern AD, Dehay C, Igel H, Ares M Jr, Vanderhaeghen P, Haussler D. An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* 2006;443:167–172. [PubMed: 16915236]
- Ponjavic J, Oliver PL, Lunter G, Ponting CP. Genomic and transcriptional co-localization of protein-coding and long non-coding RNA pairs in the developing brain. *PLoS Genet* 2009;5:e1000617. [PubMed: 19696892]

- Ramskold D, Wang ET, Burge CB, Sandberg R. An abundance of ubiquitously expressed genes revealed by tissue transcriptome sequence data. *PLoS Comput Biol* 2009;5:e1000598. [PubMed: 20011106]
- Scherzer CR. Chipping away at diagnostics for neurodegenerative diseases. *Neurobiol Dis* 2009;35:148–156. [PubMed: 19285134]
- Shendure J, Ji H. Next-generation DNA sequencing. *Nat Biotechnol* 2008;26:1135–1145. [PubMed: 18846087]
- St Laurent G 3rd, Faghihi MA, Wahlestedt C. Non-coding RNA transcripts: sensors of neuronal stress, modulators of synaptic plasticity, and agents of change in the onset of Alzheimer's disease. *Neurosci Lett* 2009;466:81–88. [PubMed: 19699259]
- t Hoen PA, Ariyurek Y, Thygesen HH, Vreugdenhil E, Vossen RH, de Menezes RX, Boer JM, van Ommen GJ, den Dunnen JT. Deep sequencing-based expression analysis shows major advances in robustness, resolution and inter-lab portability over five microarray platforms. *Nucleic Acids Res* 2008;36:e141. [PubMed: 18927111]
- Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, Lao K, Surani MA. mRNA-Seq whole-transcriptome analysis of a single cell. *Nat Methods* 2009;6:377–382. [PubMed: 19349980]
- Wang ET, Sandberg R, Luo S, Khrebtkova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB. Alternative isoform regulation in human tissue transcriptomes. *Nature* 2008;456:470–476. [PubMed: 18978772]
- Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 2009;10:57–63. [PubMed: 19015660]
- Wu JQ, Du J, Rozowsky J, Zhang Z, Urban AE, Euskirchen G, Weissman S, Gerstein M, Snyder M. Systematic analysis of transcribed loci in ENCODE regions using RACE sequencing reveals extensive transcription in the human genome. *Genome Biol* 2008;9:R3. [PubMed: 18173853]

Figure 1. An outline of a deep sequencing experiment

To demonstrate the general utility of deep sequencing we outline a simple gene expression profiling using RNA-sequencing from two samples that have differential expression of a small number of genes (red in sample A versus green in sample B). The first set of steps are generation of a library of small fragments; RNA is isolated, fragmented, cDNA synthesized and adaptors are ligated to those fragments. Amplification of millions of small individual fragments takes place in parallel usually in an array or flowcell. Each amplification product is then sequenced using platform-dependent chemistry approaches to build up a base-by-base sequence. Computational approaches are critical; a bioinformatics pipeline is needed to take each of the sequences and align them to a reference genome. The output can either be lists of sequences

or, in this example, counts of identified transcripts showing the representative differentially expressed material initially inputted (red and green dots in the graph; black dots are the many transcripts found at similar abundances in both samples).