Short Course I

Transcriptomics: Assessing Genomic Networks in Normal and Diseased Brains

Organized by James Eberwine, PhD

SOCIETY FOR NEUROSCIENCE

Please cite articles using the model:
[AUTHOR'S LAST NAME, AUTHOR'S FIRST & MIDDLE INITIALS] (2012)
[CHAPTER TITLE] In: Transcriptomics: Assessing Genomic Networks in Normal and Diseased Brains.
(Eberwine J, ed) pp. [xx-xx]. Washington, DC: Society for Neuroscience.

NEUROSCIENCE 2012

SHORT COURSE #1

Transcriptomics: Assessing Genomic Networks in Normal and Diseased Brains

Organized by: James Eberwine, PhD
Friday, October 12
8 a.m. – 6 p.m.
Location: Ernest N. Morial Convention Center
Room: La Nouvelle AB
New Orleans

TIME	AGENDA TOPICS	SPEAKER
7:30 – 8 a.m.	CHECK-IN	
8 – 8:15 a.m.	Opening Remarks	James Eberwine, PhD *University of Pennsylvania*
8:15 – 9:10 a.m.	The Missing 50% and Synaptic Transcriptomic's in Aging and Alzheimer's Disease	Paul Coleman, PhD *University of Rochester*
9:10 – 10:05 a.m.	Mapping the Genomic Pathways that Dysregulate Brain Inhibition in Disease	Shelley Russek, PhD *Boston University*
10:05 – 10:20 a.m.	MORNING BREAK	
10:20 – 11:15 a.m.	Enhancing the Interpretation of Genomic Data Using RNA-seq From Ips-Derived Neurons	Ken Kosik, MD *University of California, Santa Barbara*
11:15 a.m. – 12:10 p.m.	Computational Analysis of RNA-seq Data: From Quantification to High-Dimensional Analysis	Junhyong Kim, PhD *University of Pennsylvania*
12:10 – 1:10 p.m.	LUNCH: ROOM 265 – 268	
1:10 – 2:05 p.m.	Insight into the Molecular Basis of Addiction From Chip-Seq and RNA-Seq	Eric Nestler, PhD, MD *Mt. Sinai School of Medicine*
2:05 – 3 p.m.	Single Cell Transcriptomics: Surprises and Insights	James Eberwine, PhD *University of Pennsylvania*
3 – 3:40 p.m.	SUMMARY, DISCUSSION, BREAKOUT GUIDE	
3:40 – 4 p.m.	AFTERNOON BREAK	

SOCIETY FOR NEUROSCIENCE

NEUROSCIENCE 2012

AFTERNOON BREAKOUT SESSIONS

Participants select one discussion group at 4 p.m. and one at 5 p.m.

TIME	THEME	ROOM
4 – 5 p.m.	BREAKOUT SESSIONS	
	GROUP 1 – Inferring Function From RNA-Seq Data Ken Kosik, MD Junhyong Kim, PhD	260
	GROUP 1 – RNA-Seq Insights Into Complex Diseases Shelley Russek, PhD Paul Coleman, PhD	261
	GROUP 1 – Integrating RNA-Seq Data With Higher Order Cell Biologies Eric Nestler, PhD, MD James Eberwine, PhD	262
5 p.m. – 6 p.m.	REPEAT SESSIONS ABOVE. SELECT A SECOND DISCUSSION GROUP.	

SOCIETY FOR NEUROSCIENCE

Table of Contents

Introduction

Most cells in an organism have a very similar genome yet mRNA expression (called the expression profile) can vary dramatically. These expression differences give rise to specialized cellular phenotypes and functioning. As analysis of the proteome is still quite difficult and doesn't provide high sensitivity, analysis of the transcriptome provides a surrogate that can be viewed as the functional potential of the cell/tissue. This is the case as a protein can only be made if the RNA is expressed. Characterization of expression profiles has evolved and matured over the years moving from Northern Analysis, through PCR to microarrays and now the current application of NextGen sequencing (RNA-Seq). RNA-Seq methodologies permit sequence characterization and abundance measurements for all RNAs from a sample even as small as a single cell, in an unbiased manner. The advent of RNA sequencing (RNA-Seq) based transcriptomic's eliminates the requirement to choose sequences for investigation (as with PCR or microarray analysis), as there is no need to choose targets or probes. Sequencing can provide a greater depth of information regarding transcript variants and gives a more complete picture of the transcriptome and in turn cellular phenotype.

Transcriptomic analysis has provided fundamental insights into cell biology including showing the existence of many alternatively and noncanonically spliced mRNAs from multiple genes expressed within a tissue sample. These variant splice forms are not limited to exonic coding region differences as previously undescribed retained introns have been found for a number of cytoplasmic mRNAs in various tissues. It is important to note that these results are not unexpected as much previous data on mRNA transcript sequence is based upon the most easily detectable (most abundant) isoforms present in cells, which then serve to define what we think of as canonical exons, introns, UTRs and gene boundaries. Further, transcriptome analysis has enabled the discovery of large numbers of distinct noncoding and small RNAs within tissues including the CNS. The discovery of variant splice forms of mRNA and other classes of RNA have encouraged extensive and continuing experimentation into the role of these RNAs in cellular function.

With regard to the CNS, transcriptomic analysis has been used to investigate the impact of behavior and drug responsiveness upon normal and disease associated brain and peripheral nerve functioning in a variety of organisms. This short course will highlight advances in understanding of CNS function enabled by transcriptomic analysis while emphasizing the complexities of data and functional analysis.

Synapses and Epigenetics in the Alzheimer's Brain

Diego Mastroeni,[1] Nicole C. Berchtold, PhD,[2]
Carl W. Cotman, PhD,[2] and Paul D. Coleman, PhD[1]

[1]L.J. Roberts Center for Alzheimer's Research
Banner Sun Health Research Institute
Sun City, Arizona

[2]Institute for Brain Aging and Dementia
University of California, Irvine
Irvine, California

NOTES

Correlating Synapse Density with Cognitive Status

It has been generally accepted that synapses are the best correlate of cognitive status in Alzheimer's disease (AD). This is intuitively appealing because synapses provide the mechanisms by which information is transmitted from cell to cell, processed, and stored. The concept that synapses play a vital role in cognition was reinforced by two early quantitative studies of the relationship between synapse density and cognitive status, both of which yielded correlations of approximately +0.7 (DeKosky and Scheff, 1990; Terry et al., 1991). However, this result yields an R^2 of ~0.50, indicating that synapse density accounts for only 50% of the variability in cognitive status. This conclusion then leads to the question: Where is the missing 50%?

More recent study of the relationship between synapses and cognitive scores used unbiased stereological methods to quantify total synaptic numbers in lamina 3 of the inferior temporal gyrus (Scheff et al., 2011). The results showed an even lower correlation of +0.5 between cognitive score (according to the Mini-Mental State Examination [MMSE]) and synapse numbers, yielding an R^2 of 0.25. So in this case, synapse numbers accounted for only 25% of the variance in cognition scores. Where is the missing 50–75%?

There are several potential responses to the question:

(1) The earlier studies did not use methods of unbiased stereology, suggesting that researchers did not account for the potential effects of changes in size of synapses and volume of the brain region studied;
(2) It may be presumptuous to consider that synapse density in only one brain region could account for a behavior as complex as cognitive status; or
(3) The missing percentage can be found in synapses that are structurally present but functionally impaired.

More recent data have emphasized the complexity of relationships between synapses and cognition. For example, they have reported a correlation of 0.97 between delayed nonmatching to sample and size of the spine head in thin spines in prefrontal cortex area 46 of monkey (Morrison and Baxter, 2012).

Gene Expression in Alzheimer's Disease

A wide range of studies has provided data demonstrating impaired expression in AD of genes that play major roles in synaptic function. Studies of specific molecules in AD have, for example, shown reduced expression of dynamin 1 (Yao et al., 2003), which is critical in recycling synaptic vesicles, and losses in cholinergic receptor systems (Parri et al., 2011), a system that has a significant role in memory formation.

Beyond studies of specific molecules in AD, array studies have yielded an appreciation of the wide range of synapse-related genes whose expression is affected in AD (Berchtold et al., 2008; Liang et al., 2007, 2008). These studies indicate that a wide variety of synaptic gene classes are affected in AD, including transmitter receptor systems, transmitter synthesizing enzymes, transport systems, synapse stabilizing genes, postsynaptic structural genes, and ion channels, to name a few. Data such as these show coordinated modulation of expression of a wide range of genes and raise the question—what mechanism is coordinating such a wide variety of changes?

Epigenetic Mechanisms in Alzheimer's Disease

Recent study findings

During early development of an organism, the specification of cells and tissues requires the expression of large numbers of genes, modulated by epigenetic mechanisms in a coordinated fashion to produce specified cell types and tissues (Allis et al., 2009; Olynik and Rastegar, 2012). The ability of epigenetic mechanisms to regulate chromatin structure and, consequently, the coordinated expression of large gene sets, has motivated a number of studies on the role of epigenetic mechanisms in AD.

One of the early demonstrations of a relationship between AD and an epigenetic mechanism came from the demonstration of reduced expression in AD of 10 epigenetic markers in layer II neurons of the entorhinal cortex (Mastroeni et al., 2010). This study also showed that, in the AD brain, neurons bearing neurofibrillary tangles (NFTs) showed greater decrements in DNA methylation than tangle-free neurons.

An opportunity to eliminate genetic contributions from findings of epigenetic differences in AD came with an opportunity to examine epigenetic differences between a pair of identical twins discordant for AD (Mastroeni et al., 2009). The male twins of this study were educated together as chemical engineers and died within three years of each other. The twin with a clinical diagnosis of AD showed profuse NFTs and plaques, whereas the other twin had only extremely sparse NFTs and plaques. Compared with the nondemented twin, the levels of

NOTES

global DNA methylation were significantly reduced in the temporal neocortex of the AD twin, who had spent much of his career working with pesticides. This, as well as other epigenetic studies of identical twins (Fraga et al., 2005), points to the pertinence of environmental factors affecting the epigenome and, consequently, the phenotype.

Evidence from transcript expression

Although the above studies are consistent with a role for epigenetic mechanisms in altering the structure and function of synapses in AD, they are far from proving such a relationship. More detailed information about relationships between epigenetic variables and the expression of synaptic genes come from determining the correlations between expression of synaptic transcripts and the expression of selected transcripts known to regulate DNA methylation and histone acetylations. Selected aspects of these data, taken from analysis of an array study of four brain regions in brains covering the age range 20–99 years old, have already been published (Berchtold et al., 2008).

Figure 1 presents new analysis of data from Berchtold et al., 2008. It graphically represents correlations between a selection of four transcripts (Fig. 1*A–D*) related to synaptic structure and function (dynamin 1, PSD95, $AMPA_{A1}$, and synaptophysin) and ten selected transcripts that play roles in the methylation and acetylation actions of epigenetic mechanisms. These correlations are shown for four conditions: (1) AD in the postcentral gyrus (pcg, a region relatively unaffected in AD); (2) AD in the severely affected hippocampus (hipp); (3) the postcentral gyrus in nondemented, age-matched control cases; and (4) the hippocampus in nondemented, age-matched control cases.

These data show that the relationship between epigenetics and expression of synaptic genes depends on both the brain region examined and the disease state. In AD, quantification of DNA methyltransferase 1 (DNMT1) expression in hippocampus and postcentral gyrus shows a negative or low correlation with expression of all synaptic genes examined. On the other hand, the correlations are positive or low in both brain regions in age-matched controls. DNMT3a consistently yields negative correlations with all four synaptic genes, with the exception of age-matched control hippocampus, for which the correlations were either positive or low (Fig. 1*A–D*). The differences between the correlation patterns of these two DNMTs may relate to their presumptive differential functions in both *de novo* and maintenance DNA methylation.

The data for histone deacetylases (HDACs) 1, 2, 6, and 9 show varying patterns of correlation between expression of synaptic and epigenetic transcripts, depending on brain region and disease state. However, the histone acetyltransferases (HATs) MYST3 and MYST4 are fairly consistent in being negatively related to expression of the synaptic genes shown. In other words, reduced expression of these HATs is associated with increased expression of the synaptic genes shown (Fig. 1*A–D*). This finding is inconsistent with the common association of acetylation with a more open chromatin structure. It also serves to remind us of the potential complexity

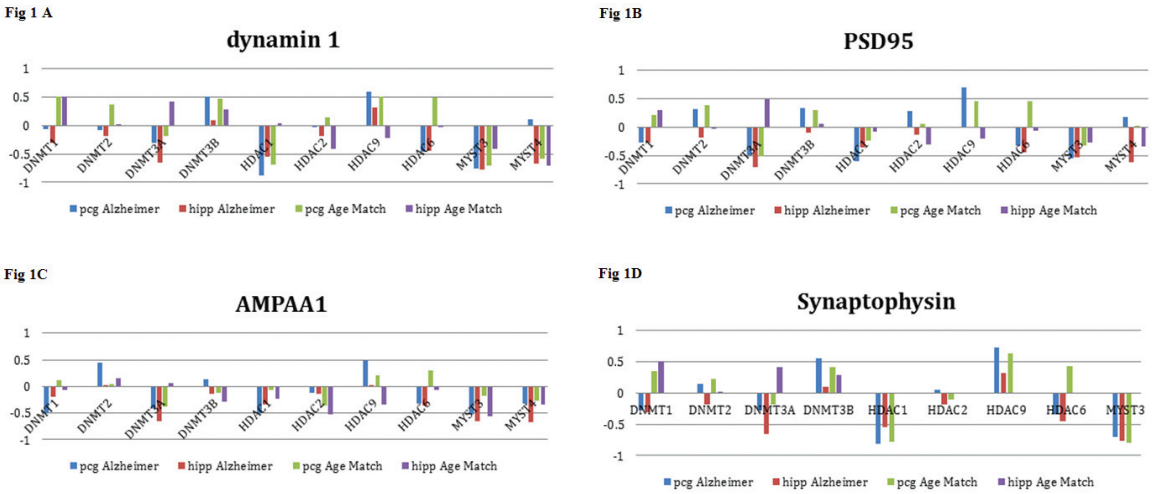

Figure 1. ***A–D***, Four plots that represent correlations (vertical axis) between the expression of 10 epigenetic molecules and 4 transcripts related to synaptic structure and function. Four data sets are represented by 4 different colors as shown: Data from AD postcentral gyrus, from AD hippocampus, from age-matched postcentral gyrus, and from age-matched hippocampus. $AMPA_{A1}$, ligand-gated ion channel, a subclass of glutamate receptor; DNMTs, a family of DNA methyltransferases; dynamin1, functions in the recycling of vesicles, especially at the synapse; HDACs, a family of histone deacetylases; MYSTs, a family of histone acetyltransferases; PSD95, postsynaptic density 95; synaptophysin, major synaptic vesicle protein.

NOTES

of interactions among histone modifications, DNA methylations, and other molecules affecting transcription at specific sites in the genome.

Another caution against generalizing these data is that the initial data came from homogenates of selected brain regions. Thus, these data represent not only neurons but also glia and vascular cells. Also, although more neuron-specific data have been derived from laser capture microdissection of single neurons (Liang et al., 2008), the amplification required for single-cell data may have differentially affected the selected genes of interest.

Implications

The correlations shown here do not, of course, prove causality. They do, however, offer suggestions for further studies, experimental manipulation, and potential therapeutic intervention for AD. At the same time, they providing a cautionary tale about the potential for complex interactions among the many epigenetic molecules, specific sites of synapse-related genes, brain regions, and disease states.

References

Allis CD, Jenuwein T, Reinberg D. (2009) Epigenetics. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press.

Berchtold NC, Cribbs DH, Coleman PD, Rogers J, Head E, Kim R, Beach T, Miller C, Troncoso J, Trojanowski JQ, Zielke HR, Cotman CW (2008) Gene expression changes in the course of normal brain aging are sexually dimorphic. Proc Natl Acad Sci USA 105:15605–15610.

DeKosky ST, Scheff SW (1990) Synapse loss in frontal cortex biopsies in Alzheimer's disease: Correlation with cognitive severity. Ann Neurol 27:457–464.

Fraga MF, Ballestar E, Paz MF, Ropero S, Setien F, Ballestar ML, Heine-Suñer D, Cigudosa JC, Urioste M, Benitez J, Boix-Chornet M, Sanchez-Aguilera A, Ling C, Carlsson E, Poulsen P, Vaag A, Stephan Z, Spector TD, Wu YZ, Plass C, Esteller M (2005) Epigenetic differences arise during the lifetime of monozygotic twins. Proc Natl Acad Sci USA 102:10604–10609.

Liang WS, Dunckley T, Beach TG, Grover A, Mastroeni D, Walker DG, Caselli RJ, Kukull WA, McKeel D, Morris JC, Hulette C, Schmechel D, Alexander GE, Reiman EM, Rogers J, Stephan DA. (2007) Gene expression profiles in anatomically and functionally distinct regions of the normal aged human brain. Physiol Genomics 28:311–322.

Liang WS, Dunckley T, Beach TG, Grover A, Mastroeni D, Ramsey K, Caselli RJ, Kukull WA, McKeel D, Morris JC, Hulette CM, Schmechel D, Reiman EM, Rogers J, Stephan DA (2008) Altered neuronal gene expression in brain regions differentially affected by Alzheimer's disease: A reference data set. Physiol Genomics 33:240–256.

Mastroeni D, McKee A, Grover A, Rogers J, Coleman PD. (2009) Epigenetic differences in cortical neurons from a pair of monozygotic twins discordant for Alzheimer's disease. PLoS One 4(8):e6617.

Mastroeni D, Grover A, Delvaux E, Whiteside C, Coleman PD, Rogers J. (2010) Epigenetic changes in Alzheimer's disease: Decrements in DNA methylation. Neurobiol Aging 31:2025–2037.

Morrison JH, Baxter MG (2012) The aging cortical synapse: Hallmarks and implications for cognitive decline. Nat Rev Neurosci 13:240–250.

Olynik BM, Rastegar M. (2012) The genetic and epigenetic journey of embryonic stem cells into mature neural cells. Front Genet 3:81.

Parri HR, Hernandez CM, Dineley KT. (2011) Research update: Alpha7 nicotinic acetylcholine receptor mechanisms in Alzheimer's disease. Biochem Pharmacol 82:931–942.

Scheff SW, Price DA, Schmitt FA, Scheff MA, Mufson EJ. (2011) Synaptic loss in the inferior temporal gyrus in mild cognitive impairment and Alzheimer's disease. J Alzheimers Dis 24:547–557.

Terry RD, Masliah E, Salmon DP, Butters N, DeTeresa R, Hill R, Hansen LA, Katzman R. (1991) Physical basis of cognitive alterations in Alzheimer's disease: synapse loss is the major correlate of cognitive impairment. Ann Neurol 30:572–580.

Yao PJ, Zhu M, Pyun EI, Brooks AI, Therianos S, Meyers VE, Coleman PD. (2003) Defects in expression of genes related to synaptic vesicle trafficking in frontal cortex of Alzheimer's disease. Neurobiol Dis 12:97–109.

Mapping the Genomic Pathways That Dysregulate Brain Inhibition in Disease

Shelley J. Russek, PhD

Laboratory of Translational Epilepsy
Department of Pharmacology and Experimental Therapeutics
Boston University School of Medicine
Boston, Massachusetts

NOTES

Introduction: The Role of GABA in the CNS

GABA is the major inhibitory neurotransmitter in the CNS. It activates three different classes of receptors: the ionotropic type A receptor ($GABA_AR$) and type C receptor ($GABA_C$) and the G protein-coupled type B receptor ($GABA_B$). In the retina, the $GABA_C$ receptor regulates fast synaptic inhibition, while in the brain, this function is specific to the $GABA_AR$ (MacDonald and Olsen, 1994; Bormann and Feigenspan, 1995; Rabow et al., 1995; Sieghart and Sperk, 2002). $GABA_B$ receptors are involved in slower, more prolonged inhibitory signaling (Jacob, et al., 2008).

$GABA_A$ receptors

Similar to other members of the ligand-gated ionotropic receptor family, such as the nicotinic acetylcholine receptor, the $GABA_AR$ is defined by the assembly of five subunits, as well as the presence of GABA and benzodiazepine (BZ) binding sites (Choi et al., 1981; MacDonald and Olsen, 1994; Chebib and Johnston, 1999). $GABA_ARs$ mediate fast synaptic inhibition by regulating the flow of Cl^- ions down their concentration gradient into the cell to hyperpolarize the postsynaptic neuronal membrane, hindering the spread of excitability (Costa, 1998). While GABA is an inhibitory neurotransmitter in the adult brain, in embryonic and early postnatal mammalian hippocampal neurons, synaptically released or exogenously applied GABA depolarizes and excites postsynaptic membranes via $GABA_AR$ activation (Cherubini et al., 1991). This excitatory response has been attributed to the presence of an embryonic chloride transporter (NKCC1 [sodium–potassium–chloride cotransporter 1]) that increases intracellular chloride concentration opposed to KCC2, which is expressed in adult neurons and extrudes Cl^- (Ben-Ari, 2002).

Role in normal development and disease

Dynamic changes in NKCC1 expression during brain development have recently been associated with the critical migration of neuroblasts to their targets (Mejia-Gervacio et al., 2011), while misexpressed NKCC1 has been implicated in multiple disorders, including epilepsy (Palma et al., 2006). Moreover, a role for the excitatory function of GABA has been proposed for the development of synaptic connections, as well as the subsequent plasticity and establishment of key neuronal networks dysregulated in developmental disorders such as autism (Kriegstein and Owens, 2001).

GABA plays a highly significant role in both the developing nervous system, the adult brain, and the compromised brain (as reflected in multiple disease states). As a result, our laboratory has had a longstanding interest in identifying specific protein–DNA interactions that regulate the transcription of unique $GABA_AR$ subunit genes (*GABRs*) and critical $GABA_AR$-associated proteins. Our investigations are meant to shed light on the genome response that shapes both present and future affective and cognitive behavior.

The $GABA_AR$ Genome and Its Gene Products

$GABA_AR$ subunits

In mammals, $GABA_AR$ subunits are divided into seven subunit classes based on sequence homology; including α_1-6, β1-3, γ1-3, δ, ε, θ, and π (Rabow et al., 1995). Although the majority of genes coding for multi-subunit receptor families lie scattered in the genome, evolution has preserved the organization of the $GABA_AR$ subunit genes, challenging us to understand the forces behind cluster preservation and expansion. Conservation of gene order and orientation on human chromosomes is shown in Figures 1 and 2. Taken together with the conservation of intron position in the β genes (Russek and Farb, 1994), it demonstrates that the diversity of $GABA_A$ receptor subunit genes originated from the duplication of an ancestral gene and the subsequent translocation of an ancestral gene cluster (Russek, 1999). Head-to-head orientation of the α and β subunit genes also suggests that they may be positively or negatively regulated by the proximity of regulatory elements.

Additional $GABA_AR$ subunit variants are observed from alternative splicing of individual subunit transcripts (Barnard et al., 1998; Jacob et al., 2008). Subunits within groups share 60–80% homology, whereas between different subunit families, homology is only 30% (Costa, 1998). $GABA_ARs$ are assembled from subunits in the endoplasmic reticulum (ER) (Jacob et al., 2008). Their departure from the ER depends on proteins reaching conformation maturity, contributing to a diverse population of $GABA_ARs$ at the cell surface. While many subunit conformations are possible, only a limited number of $GABA_ARs$ actually exit the ER, as less than 25% of translated subunits assemble into $GABA_ARs$ (Gorrie et al., 1997). Misfolded or nonassembled subunits are degraded through the ubiquitin–proteasome pathway (Bedford et al., 2001; Jacob et al., 2008). Once assembled, $GABA_ARs$ are trafficked to the Golgi to be packaged into vesicles for transport to and insertion into cellular membranes (Jacob et al., 2008).

NOTES

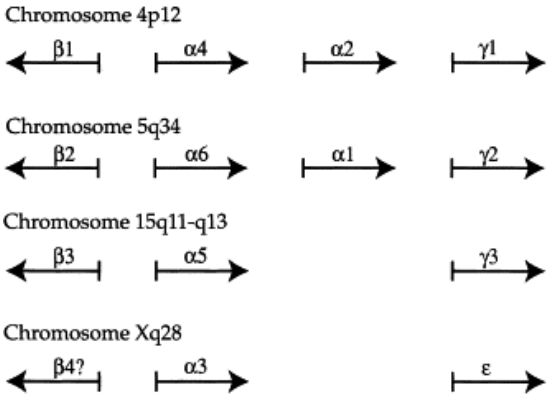

Figure 1. Gene organization is conserved for the $GABA_A$ receptor gene clusters on human chromosomes 4, 5, 15, and X. Orientation of subunit genes are indicated by arrow direction. (The schematic is not drawn to scale.) The most current cytogenetic localization of the gene cluster is indicated next to the chromosome number. Information for chromosomes 4 and 5 is presented in Russek, 1999. Information for chromosomes 15 and X is from Greger et al., 1995; Levin et al., 1996; and Wilke et al., 1997. Note that no β4 (avian) was ever reported in mammals and that sequencing of the genome revealed presence of θ in that position for rodents and humans.) Russek, 1999, Figure 5, reprinted with permission.

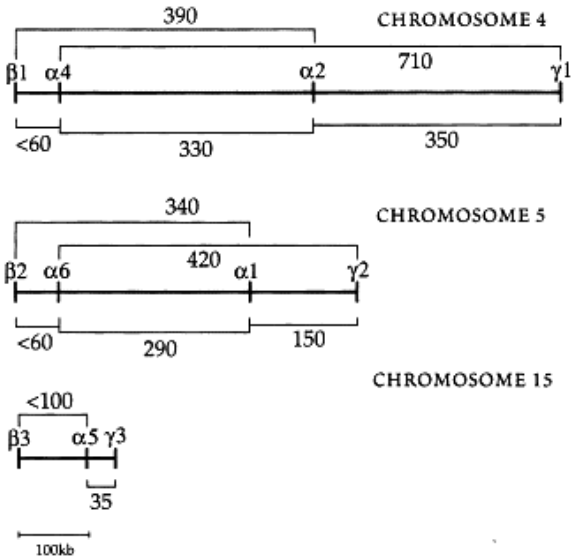

Figure 2. Schematic representation of the genomic organization of $GABA_A$ receptor gene clusters on chromosomes 4, 5, and 15 in the human genome. Estimates of genomic distance between genes on chromosomes 4 and 5 were obtained from interphase mapping. Distance measurements for the genes on chromosome 15 were obtained by restriction fragment-length fingerprinting and interphase fluorescence *in situ* hybridization (FISH) mapping (Greger et al., 1995) and by restriction fragment length analysis with field-inversion gel electrophoresis (Sinnet et al., 1993). Measurements are close to the distance verified by sequencing of the human genome. The diagram has been drawn to scale. Scale bar: 100 kb. Russek, 1999, Figure 3, reprinted with permission.

Pharmacological properties of $GABA_A$Rs

Different $GABA_A$R subtypes, a product of differential subunit composition, confer distinct receptor localization and function. Fully functional $GABA_A$Rs require at least one α, one β, and one other subunit type, allowing for GABA-gated Cl^- flux (Pritchett et al., 1989; Johnston, 1996; Chebib and Johnston, 1999). The most common receptor subtype contains 2α, 2β, and 1γ (or δ) subunit (MacDonald and Olsen, 1994; Jacob et al., 2008).

$GABA_A$Rs are the site of action for many therapeutics, including barbiturates, benzodiazepines (BZs), ethanol, and anesthetic steroids (Vicini, 1991; MacDonald and Olsen, 1994; Brooks-Kayal et al., 1998a). Research has demonstrated that different subunits confer distinct pharmacological properties to $GABA_A$Rs. For example, BZs act as allosteric modulators of $GABA_A$Rs, amplifying GABA signaling with varying levels of efficacy depending on the α and γ subunits present in the complex (Pritchett et al., 1989). Neurosteroids and barbiturates can also amplify GABA-gated current in most $GABA_A$R subtypes, by increasing chloride channel open time (Costa, 1998; Puia et al., 1990). In the adult brain, α_1 is the most abundant $GABA_A$R subunit and is found in 50% of $GABA_A$Rs (Duggan and Stephenson, 1990; McKernan et al., 1991). In general, receptors containing α_1 are mostly synaptic, sensitive to BZ, insensitive to zinc inhibition, and mediate most phasic inhibition in the brain (Pritchett et al., 1989; Puia et al., 1991; MacDonald and Kapur, 1999).

Levels of α_1 subunit expression can be altered by treatment with different mediators of synaptic signaling, suggesting that its expression may be activity-dependent. Treatment with NMDA stimulates α_1 expression in cultured cerebellar granule cells (Harris et al., 1994; Zhu et al., 1995). In contrast, prolonged treatment with GABA or BZ decreases α_1 expression in cortical and hippocampal neurons, respectively (Tietz et al., 1993; Lyons et al., 2000). Additional experiments using immunoprecipitation with subunit-specific antibodies followed by radiolabeled muscimol binding (a ligand that binds to the GABA binding site between α and β subunits) found that α_1 precipitated 70–90% of radiolabeled muscimol binding sites from rat or mouse brain membrane extracts (Sieghart and Sperk, 2002).

Alpha4 $GABA_A$ receptors

$GABA_A$Rs containing α_4 are less abundant, detected mainly in the hippocampus and thalamus (Rabow et al., 1995; Whiting et al., 1995; Benke et al., 1997; Sur et al., 1999). Furthermore, $GABA_A$Rs containing α_4 subunits are predominately extrasynaptic, insensitive

NOTES

to BZs, sensitive to zinc inhibition, and mediate tonic inhibition (Knoflach et al., 1996; Benke et al., 1997; Fisher and MacDonald, 1998; Lagrange et al., 2007). Immunoprecipitation experiments with α_4 subunit–specific antibodies detected α_4 in only 6% of $GABA_A$Rs in the brain (Sieghart and Sperk, 2002).

Altered expression of $GABA_A$Rs

Differences in the number and subunit composition of $GABA_A$Rs contribute to their unique function in discrete brain regions. Any alteration in such expression has been observed in multiple disease states, including alcoholism, Alzheimer's disease, autism, drug abuse, epilepsy, and schizophrenia. Changes in subunit expression are also observed in many of the comorbidities associated with epilepsy, such as anxiety disorders, cognitive deficits, and depression (Jacob et al., 2008).

With GABA and its type A receptors playing such critical roles in brain development and in brain inhibition more generally, they present a unique opportunity for the research community. The goal is to test the power of modern transcriptomics as a means of uncovering basic principles of brain design and function, which may be represented in the structure of the genome.

Gene Regulatory Networks That Control $GABA_A$R Subunit Genes

As discussed above, altered GABAergic function has been associated with multiple brain disorders. An additional feature of these disorders is a marked change in neurotrophic signaling, especially as orchestrated by brain-derived neurotrophic factor (BDNF). Work from our laboratory (in collaboration with Amy Brooks-Kayal and her group, who model temporal lobe epilepsy *in vivo*) has uncovered a unique relationship between these two receptor systems: $GABA_A$Rs and BDNF receptors (trkB and p75 neurotrophin receptor [p75NTR]). These findings suggest the two systems are part of an important gene regulatory network that is active in normal and diseased brain (Brooks-Kayal et al., 2009). Briefly, by activating the trkB receptor and downstream mitogen-activated protein kinase (MAPK) and protein kinase C (PKC) intracellular cascades, BDNF increases levels of early growth response factor 3 (Egr3). Egr3, in turn, is an activator of the *GABRA4* promoter that drives the expression of $GABA_A$R α_4 subunits (Roberts et al., 2005, 2006).

In parallel, working via a novel pathway we have recently shown links p75NTR to the JAK/STAT cascade, BDNF increases levels of inducible cAMP

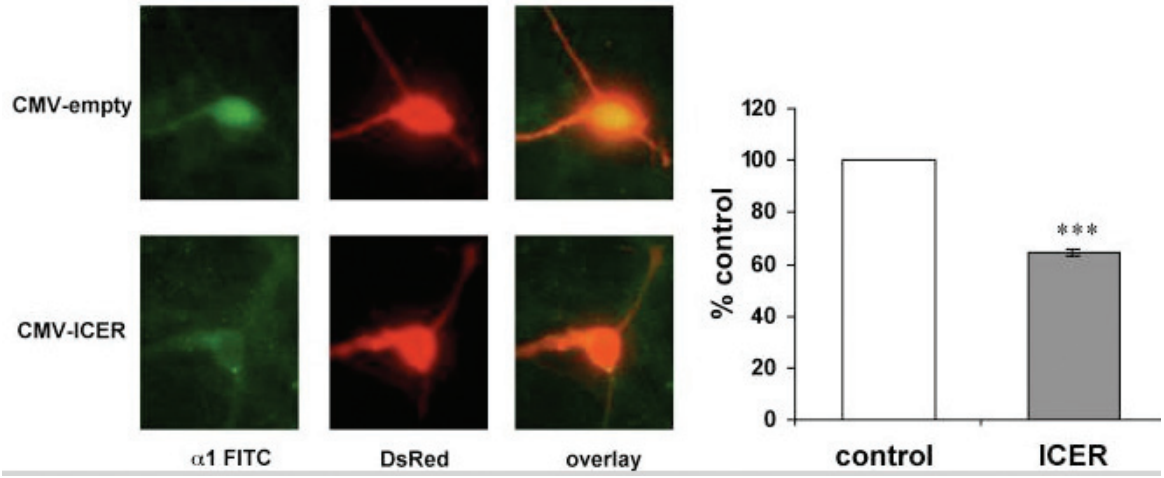

Figure 3. Effects of ICER induction on cell-surface α_1 expression. Overexpression of ICER decreases the endogenous levels of α_1 subunit detected at the cell membrane. Primary cultured neocortical neurons were cotransfected with pDsRed2-Monomer and ICER expression (*CMV-ICER*) or control vectors (*CMV-empty*). At 48 h after transfection, unpermeabilized cells were fixed and stained with an α_1-specific antibody using a standard protocol. The DsRed-transfected cells were viewed by using an Olympus IX71 inverted fluorescence microscope (Olympus America, Center Valley, PA), and the images were analyzed by using IPLab software (Becton Dickinson, Franklin Lakes, NJ). Representative images are shown (empty vector, top panel; ICER construct, bottom panel). Quantitation data are presented in the *histogram* (***, $p < 0.01$; mean ± S.E.; $n = 3$). FITC, fluorescein isothiocyanate. Hu et al., 2008, their Figure 9B, reprinted with permission.

NOTES

early repressor (ICER): a repressor of the core *GABRA1* promoter (Lund et al., 2008). Figure 3 depicts the overexpression of ICER in primary neurons, a process that alters the number of α_1 subunits at the cell surface (Hu et al., 2008). For the first time, the presence of ICER has been demonstrated to be directly relevant to the disappearance of the subunit from a functional compartment. Multiple intracellular signaling pathways regulate *GABRA1* transcription. Figure 4 depicts the process in which activation of PKC enhances transcription while activation of protein kinase A (PKA), like BDNF, represses transcription, dependent on the presence or absence of ICER.

Dynamics of $GABA_AR$ Transcription as Revealed by High-Density ChIP Sequencing

Increased access to new opportunities to probe genome activity at a global level holds great promise of open-ended discovery in GABA biology in the years to come. Recent evidence suggests that paracrine GABA, released from emerging neuroblasts, may participate in a negative feedback mechanism that causes cells to exit the cell cycle, producing fewer progenitors and supporting cellular differentiation (LoTurco et al., 1995, Kreigstein and Owens, 2001; Andäng et al., 2008) . The identification of early GABA as a switch that may control the potential pool of neural progenitors in the developing brain has shed light on the importance of studying this receptor system in the embryo, in addition to the adult brain, where most researchers have concentrated their efforts.

As a first step in this direction, we examined the chromatin state of *GABR* gene clusters in mouse embryonic stem cells (ESs) as compared with ES-derived neural progenitors (NPs), mouse embryonic fibroblasts (MEFs), and whole brain tissue (WB). We used the UCSC Genome Browser (http://genome.ucsc.edu) to analyze the results of chromatin immunoprecipitation (ChIP). ChIP was performed at the Broad Institute, using antibodies to three markers: H3K4me3 (a histone marker of transcriptional activation found at or close to active transcriptional start sites); H3K27me3 (a histone marker associated with genes that are silenced); and H3K36me3 (a histone marker usually found immediately after transcriptional start sites associated with active transcription) (Mikkelsen et al, 2007; Meissner et al., 2008). Genes displaying a sharp peak at both H3K4me3 and H3K27me3 contain a bivalent chromatin mark that has been associated with marks that play key roles in lineage-specific activation or repression.

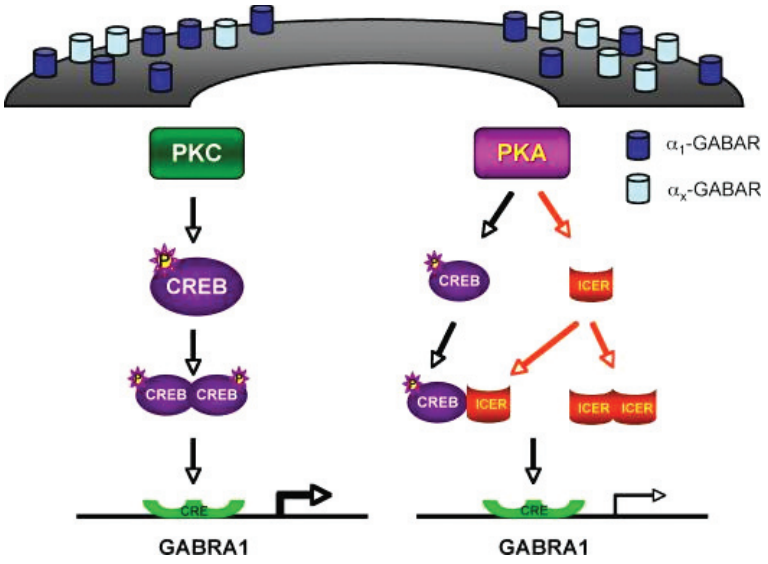

Figure 4. A model for the role of CREB and ICER in the regulation of *GABRA1* transcription. Activation of the PKC pathway leads to phosphorylation of CREB without induction of ICER. Phosphorylated CREB at Ser-133 forms homodimers to increase *GABRA1* expression (left). Activation of the PKA pathway induces synthesis of ICER and phosphorylation of CREB. Homodimers of ICER or ICER and CREB heterodimers repress transcription of *GABRA1* (right) and alter the number of α1-containing $GABA_AR$s (α1-GABAR) at the cell surface. None α_1 subunit–containing receptors are as indicated by "αx" key. CREB, cAMP response element-binding protein. Hu et al., 2008, their Figure 10, reprinted with permission.

Figures 5–7 display the results of ChIP sequencing through three *GABR* clusters in the mouse genome. Proposed *GABRs* with the highest probability of being expressed, either at the ES-cell level or upon commitment to NP or MEF, are indicated by red lettering. Analysis of histone marks in these different cell populations suggests that the early $GABA_AR$ is composed of *GABRA2*, *GABRB3*, and *GABRG1*. Results of RNA sequencing will confirm or refute this hypothesis and provide the necessary feedback to determine whether unique histone marks can predict the expression of $GABA_AR$ gene clusters in normal and diseased brains.

NOTES

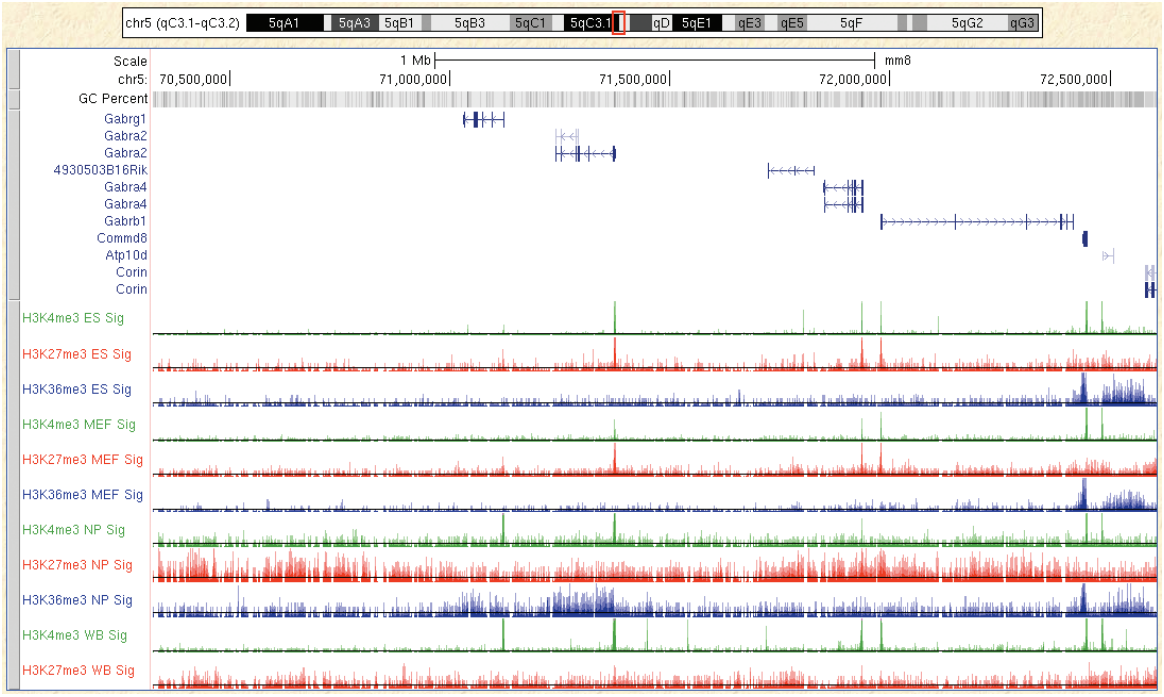

Figure 5. *GABRG1*, *GABRA2*, *GABRA4*, *GABRB1* gene cluster. Note that *GABRG1* is not univalent but is marked for transcription in NP, while *GABRA2* is univalent with high levels of expression in NP and H3K36me3 close to the start site.

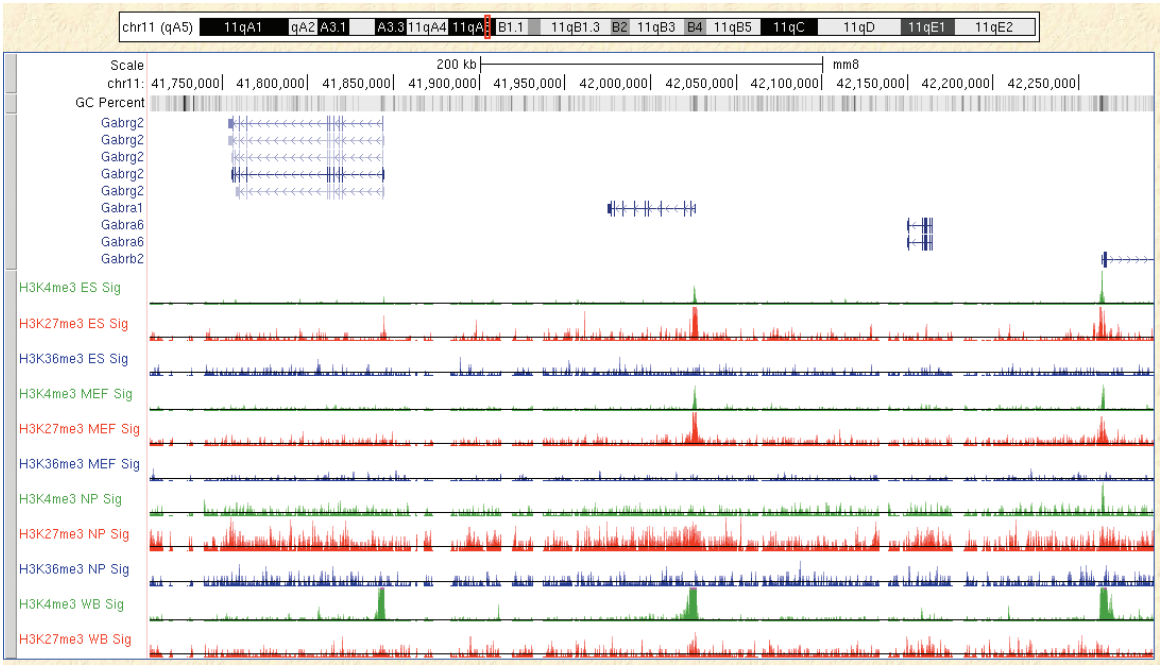

Figure 6. *GABRG2, GABRA1, GABRA6, GABRB2* gene cluster. Note that there are no peaks associated with *GABRG2* or *GABRA6*. However, there are univalent marks for ES and MEF with their loss in NP for *GABRA1*.

NOTES

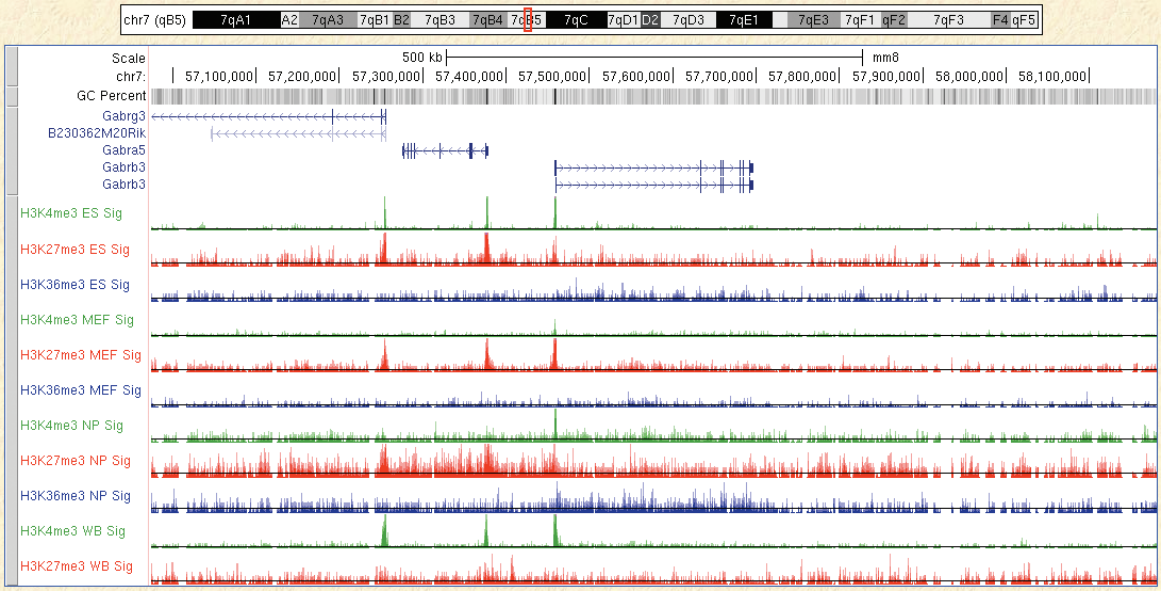

Figure 7. *GABRG3, GABRA5, GABRB3* gene cluster. Note that all genes in this cluster are univalent with evidence for some *GABRB3* in stem cells. There is a loss of a *GABRB3* univalent mark in MEF and strong H3K36me3 in NP, suggestive of a transcribed gene.

Breaking New Ground in a Familiar Landscape

Our lab and others across the country have identified a handful of gene regulatory proteins that are critical to the altered expression of certain *GABRs* in disease models. We have done so through a combination of traditional candidate gene regulatory assays and investigator-driven bioinformatic analysis. These discoveries have opened up new avenues for whole-genome investigations, using the power of ChIP-Seq and RNA-Seq analysis, to determine the transcriptome that is regulated in a coordinate or independent manner. Little is still known about how *GABRs* are coordinately regulated and why they have remained in clusters throughout our evolution. New techniques such as chromosome conformation capture—as used first to describe the beta-globin locus (Tolhuis et al, 2011)—may be powerful tools for exploring this new and complex territory.

Future discoveries in the field of GABA subunit gene regulation will take place in the background of an extensive history of GABA receptor biology that parallels the development of the larger field of neuroscience. Identifying gene duplications and inversions within *GABR* clusters that associate with human diseases will also provide a window onto the relationship between $GABA_AR$ number and kind that is key to maintaining a healthy balance of GABAergic neurotransmission in the young and old. These important questions have perplexed neuroscientists for over two decades; finally, the techniques are powerful enough to provide some answers.

References

Andäng M, Hjerling-Leffler J, Moliner A, Lundgren TK, Castelo-Branco G, Nanou E, Pozas E, Bryja V, Halliez S, Nishimaru H, Wilbertz J, Arenas E, Koltzenburg M, Charnay P, El Manira A, Ibañez CF, Ernfors P (2008) Histone H2AX-dep. GABAR regulation of stem cell proliferation. Nature 451:460–464.

Barnard EA, Skolnick P, Olsen RW, Mohler H, Sieghart W, Biggio G, Braestrup C, et al. (1998) International Union of Pharmacology. XV. Subtypes of gamma-aminobutyric $acid_A$ receptors: classification on the basis of subunit structure and receptor function. Pharmacol Rev 50:291–313.

Bedford FK, Kittler JT, Muller E, Thomas P, Uren JM, Merlo D, Wisden W, Triller A, Smart TG, Moss SJ (2001) $GABA_A$ receptor cell surface number and subunit stability are regulated by the ubiquitin-like protein Plic-1. Nat Neurosci 4:908–916.

Ben-Ari Y (2002) Excitatory actions of GABA during development: The nature of the nurture. Nat Rev Neurosci 3:728–739.

NOTES

Benke D, Michel C, Mohler H (1997) GABA(A) receptors containing the alpha4-subunit: Prevalence, distribution, pharmacology, and subunit architecture *in situ*. J Neurochem 69:806–814.

Bormann J, Feigenspan A (1995) $GABA_C$ receptors. Trends Neurosci 18:515–519.

Brooks-Kayal AR, Jin H, Price M, Dichter MA (1998) Developmental expression of GABA(A) receptor subunit mRNAs in individual hippocampal neurons *in vitro* and *in vivo*. J Neurochem 70:1017–1028.

Brooks-Kayal AR, Raol Y, Russek SJ (2009) Alteration of epileptogenesis genes. Neurotherapeutics 6:312–318.

Chebib M, Johnston GAR (1999) The "ABC" of GABA receptors: A brief review. Clin Exp Pharmacol Physiol 26:937–940.

Cherubini E, Gaiarsa JL, Ben-Ari Y (1991) GABA: An excitatory transmitter in early postnatal life. Trends Neurosci 14:515–519.

Choi DW, Farb DH, Fischbach GD (1981) Chlordiazepoxide selectively potentiates GABA conductance of spinal cord and sensory neurons in culture. J Neurophysiol 45:621–631.

Costa E (1998) From $GABA_A$ receptor diversity emerges a unified vision of GABAergic inhibition. Annu Rev Pharmacol Toxicol 38:321–350.

Duggan MJ, Stephenson FA (1990) Biochemical evidence for the existence of gamma-aminobutyrate$_A$ receptor iso-oligomers. J Biol Chem 265:3831–3835.

Fisher JL, MacDonald RL (1998) The role of an alpha subtype M2-M3 His in regulating inhibition of $GABA_A$ receptor current by zinc and other divalent cations. J Neurosci 18:2944–2953.

Gorrie GH, Vallis Y, Stephenson A, Whitfield J, Browning B, Smart TG, Moss SJ (1997) Assembly of $GABA_A$ receptors composed of alpha1 and beta2 subunits in both cultured neurons and fibroblasts. J Neurosci 17:6587–6596.

Greger V, Knoll JH, Woolf E, Glatt K, Tyndale RF, DeLorey TM, Olsen RW, Tobin AJ, Sikela JM, Nakatsu Y, et al. (1995) The gamma-aminobutyric acid receptor gamma 3 subunit gene (*GABRG3*) is tightly linked to the alpha 5 subunit gene ($GABRA_5$) on human chromosome 15q11–q13 and is transcribed in the same orientation. Genomics 26:258–264.

Harris BT, Charlton ME, Costa E, Grayson DR (1994) Quantitative changes in alpha 1 and alpha 5 gamma-aminobutyric acid type A receptor subunit mRNAs and proteins after a single treatment of cerebellar granule neurons with N-methyl-D-aspartate. Mol Pharmacol 45:637–648.

Hu Y, Lund I, Gravielle M, Farb DH, Brooks-Kayal AR, Russek SJ (2008) Surface expression of GABA(A) receptors is transcriptionally controlled by the interplay of CREB and its binding partner ICER. J Biol Chem 283:9328–9340.

Jacob TC, Moss SJ, Jurd R (2008) GABA(A) receptor trafficking and its role in the dynamic modulation of neuronal inhibition. Nat Rev Neurosci 9:331–343.

Johnston G (1996) GABA(A) receptor pharmacology. Pharmacol Ther 69:173–198.

Knoflach F, Benke D, Wang Y, Scheurer L, Lüddens H, Hamilton BJ, Carter DB, Mohler H, Benson JA (1996) Pharmacological modulation of the diazepam-insensitive recombinant gamma-aminobutyric acid$_A$ receptors alpha 4 beta 2 gamma 2 and alpha 6 beta 2 gamma 2. Mol Pharmacol 50:1253–1261.

Kriegstein AR, Owens DF (2001) GABA may act as a self-limiting trophic factor at developing synapses. Sci STKE 95:PE1.

Lagrange AH, Botzolakis EJ, MacDonald RL (2007) Enhanced macroscopic desensitization shapes the response of alpha4 subtype-containing $GABA_A$ receptors to synaptic and extrasynaptic GABA. J Physiol 578(Pt 3):655–676.

Levin ML, Chatterjee A, Pragliola A, Worley KC, Wehnert M, Zhuchenko O, Smith RF, Lee CC, Herman GE (1996) A comparative transcriptional map of the murine bare patches (Bpa) and striated (Str) critical regions and Xq28. Genome Res 6:465–477.

LoTurco JJ, Owens DF, Heath MJS, Davis MBE, Kriegstein AR (1995) GABA and glutamate depolarize cortical progenitor cells and inhibit DNA synthesis. Neuron 15:1287–1298.

Lund IV, Hu Y, Raol YH, Benham RS, Faris R, Russek SJ, Brooks-Kayal AR (2008) BDNF selectively regulates $GABA_A$ receptor transcription by activation of the JAK/STAT pathway. Sci Signal 1(41):ra9.

NOTES

Lyons HR, Gibbs TT, Farb DH (2000) Turnover and down-regulation of GABA(A) receptor alpha1, beta2S, and gamma1 subunit mRNAs by neurons in culture. J Neurochem 74:1041–1048.

MacDonald RL, Kapur J (1999) Pharmacological properties of recombinant and hippocampal dentate granule cell $GABA_A$ receptors. Adv Neurol 79:979–990.

MacDonald RL, Olsen R (1994) GABA(a) receptor channels. Ann Rev Neurosci 17:569–602.

McKernan RM, Quirk K, Prince R, Cox PA, Gillard NP, Ragan CI, Whiting P (1991) $GABA_A$ receptor subtypes immunopurified from rat brain with alpha subunit-specific antibodies have unique pharmacological properties. Neuron 7:667–676.

Meissner A, Mikkelsen TS, Gu H, Wernig M, Hanna J, Sivachenko A, Zhang X, Bernstein BE, Nusbaum C, Jaffe DB, Gnirke A, Jaenisch R, Lander ES (2008) Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature 454:766–770.

Mejia-Gervacio S, Murray K, Lledo P-M (2011) NKCC1 controls GABAergic signaling and neuroblast migration in the postnatal forebrain. Neural Dev 6:4.

Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, Alvarez P, Brockman W, Kim TK, Koche RP, Lee W, Mendenhall E, O'Donovan A, Presser A, Russ C, Xie X, Meissner A, Wernig M, Jaenisch R, Nusbaum C, et al. (2007) Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature 448:553–560.

Palma E, Amici M, Sprero F, Spinelli G, Di Angelantonio S, Ragozzino D, Mascia A, Scoppetta C, Esposito V, Miledi R, Eusebi F (2006) Anomalous levels of Cl- transporters in the hippocampal subiculum from temporal lobe epilepsy patients make GABA excitatory. Proc Natl Acad Sci USA 103:8465–8468.

Pritchett DB, Sontheimer H, Shivers BD, Ymer S, Kettenmann H, Schofield PR, Seeburg P (1989). Importance of a novel $GABA_A$ receptor subunit for benzodiazepine pharmacology. Nature 338:582–585.

Puia G, Santi MR, Vicini S, Pritchett DB, Purdy RH, Paul SM, Seeburg PH, Costa E (1990) Neurosteroids act on recombinant human $GABA_A$ receptors. Neuron 4:759–765.

Puia G, Vicini S, Seeburg PH, Costa E (1991) Influence of recombinant gamma-aminobutyric acid-A receptor subunit composition on the action of allosteric modulators of gamma-aminobutyric acid-gated Cl^- currents. Mol Pharmacol 39:691–696.

Rabow LE, Russek SJ, Farb DH (1995) From ion currents to genomic analysis: recent advances in $GABA_A$ receptor research. Synapse 21:189–274.

Roberts DS, Raol YH, Bandyopadhyay S, Lund IV, Budreck EC, Passini MA, Wolfe JH, Brooks-Kayal AR, Russek SJ (2005) Egr3 stimulation of *GABRA4* promoter activity as a mechanism for seizure-induced up-regulation of GABA(A) receptor alpha4 subunit expression. Proc Natl Acad Sci USA 102:11894–11899.

Roberts DS, Hu Y, Lund IV, Brooks-Kayal AR, Russek SJ (2006) Brain-derived neurotrophic factor (BDNF)–induced synthesis of early growth response factor 3 (Egr3) controls the levels of type A GABA receptor alpha 4 subunits in hippocampal neurons. J Biol Chem 281:29431–29435.

Russek SJ (1999) Evolution of $GABA_A$ receptor diversity in the human genome. Gene 227:213–222.

Russek SJ, Farb DH (1994) Mapping of the beta2 subunit gene (*GABRB2*) to microdissected human chromosome 5q34–q35 defines a gene cluster for the most abundant $GABA_A$ receptor isoform. Genomics 23:528–533.

Sieghart W, Sperk G (2002) Subunit composition, distribution and function of GABA(A) receptor subtypes. Curr Top Med Chem 2:795–816.

Sinnet D, Wagstaff J, Glatt K, Woolf E, Kirkness EJ, Lalande M (1993) High-resolution mapping of the γ-aminobutyric acid receptor β3 and α5 gene cluster on chromosome 15q11–q13, and localization of breakpoints in two Angelman syndrome patients. Am J Hum Genet 52:1216–1229.

Sur C, Farrar SJ, Kerby J, Whiting PJ, Atack JR, McKernan RM (1999) Preferential coassembly of alpha4 and delta subunits of the gamma-aminobutyric $acid_A$ receptor in rat thalamus. Mol Pharmacol 56:110–115.

Tietz EI, Huang X, Weng X, Rosenberg HC, Chiu TH (1993) Expression of alpha 1, alpha 5, and gamma 2 $GABA_A$ receptor subunit mRNAs measured *in situ* in rat hippocampus and cortex following chronic flurazepam administration. J Mol Neurosci 4:277–292.

Tolhuis B, Blom M, Kerkhoven RM, Pagie L, Teunissen H, Nieuwland M, Simonis M, de Laat W, van Lohuizen M, van Steensel B (2011) Interactions among polycomb domains are guided by chromosome architecture. PLoS Genet 7:e1001343.

Vicini S (1991) Pharmacologic significance of the structural heterogeneity of the $GABA_A$ receptor-chloride ion channel complex. Neuropsychopharmacology 4:9–15.

Whiting PJ, McKernan RM, Wafford KA (1995) Structure and pharmacology of vertebrate $GABA_A$ receptor subtypes. Intl Rev Neurobiol 38:95–138.

Wilke K, Gaul R, Klauck SM, Poutska A (1997) A gene in human chromosome band Xq28 (*GABRE*) defines a putative new subunit class of the $GABA_A$ neurotransmitter receptor. Genomics 45:1–10.

Zhu WJ, Vicini S, Harris BT, Grayson DR (1995) NMDA-mediated modulation of gamma-aminobutyric acid type A receptor function in cerebellar granule neurons. J Neurosci 15:7692–7701.

NOTES

Enhancing the Interpretation of Genomic Data Using RNA-Seq from iPS-Derived Neurons

Kenneth S. Kosik, MD, Matthew Lalli, Hongjun Zhou, PhD, Mary Luz Arcila, and Israel Hernandez

Neuroscience Research Institute and
Department of Molecular Cellular and Developmental Biology
University of California Santa Barbara
Santa Barbara, California

NOTES

Introduction

Several disruptive technologies promise a windfall of insights and potential approaches to therapeutic interventions for neurodegenerative diseases. In the realm of genetics, the contributing technologies comprise deep sequencing of both genomes and transcriptomes, along with the many bioinformatic and statistical tools for data analysis. The reprogramming of somatic cells, which potentially allows one to determine the transcriptome for any cell type on a specific individual's genetic background, is further enhancing these approaches.

The Genome

The principal focus of human genomic sequencing is genetic variation. Genomic variation arises from mutations, e.g., single nucleotide polymorphisms (SNPs), insertions and deletions, and copy number variation, as well as from recombination and gene flow (the movement of genes from one population to another). Full genome analyses can track all these parameters using a variety of databases and statistical tools. It is our opinion that large sequencing centers or commercial entities dedicated to genome sequencing can most efficiently and economically obtain full genomes. In contrast (as will be discussed later on), local sequencing in the lab is a preferable method for obtaining complete transcriptomes, which often have specialized requirements and methods.

A core resource for genome analysis is the Reference Sequence (RefSeq) Database. RefSeq is the National Center for Biotechnology Information (NCBI) database of curated, nonredundant genomic DNA contigs; mRNAs and proteins for known genes; and entire chromosomes. RefSeq provides a foundation for uniting sequence data with genetic and functional information. The collection includes sequences from more than 12,000 distinct taxonomic identifiers, ranging from viruses to bacteria to eukaryotes, and represents chromosomes, organelles, plasmids, viruses, transcripts, and more than 12.6 million proteins. RefSeq is available without restriction and is updated daily by NCBI staff and collaborating groups.

Genetic variation is classified as either rare or common when compared against all known variation in the genome. This distinction is of particular interest when considering susceptibility to complex diseases, which is a subset of the larger category of traits that are inherited by multiple genetic variants. Unlike Mendelian traits, which are controlled by genes of large effect size and show simple patterns of inheritance, the transmission of complex phenotypes is governed by multiple factors that lead to complicated patterns of familial inheritance. In fact, a defining feature of complex phenotypes is that no single locus contains alleles that are necessary or sufficient for the disease to develop.

How multiple variants affect a phenotype or epistasis is an unsolved problem in human genetics. Environmental and stochastic factors may also contribute to whether or not a particular phenotype appears, given the same genetic background. Some complex traits, including susceptibility to many neurodegenerative diseases (e.g., Alzheimer's disease, amyotrophic lateral sclerosis, and frontotemporal dementia), also have rare Mendelian forms. This phenomenon may guide one's thinking about the functional role of genetic variants with small effect size.

Genetic factors in complex diseases

Because complex diseases are often prevalent in the population, an early hypothesis proposed that the genetic factors underlying common diseases would be alleles that are quite common in the population at large (Lander, 1996; Chakravarti, 1999). However, allele frequencies are not smoothly distributed across the world's populations. Instead, their frequencies are related more closely to evolutionary processes that include selection, mutation, and genetic drift. Further, selection in diseases with onset beyond the reproductive years must be weighted differently. However, even mutations whose primary effect manifests late in life may have a weak deleterious effect early in life. For example, a mutation that predisposes individuals to Alzheimer's disease might also cause subtle changes in brain function earlier on. Indeed, some data have suggested this is the case for the risk of cognitive decline associated with ApoE4 (Caselli et al., 2009). Subtle early changes may have a very small selection coefficient; nevertheless, even a selection coefficient on the order of 10^{-4} can affect the frequency distribution of an allele (Pritchard, 2001).

The observed allele frequency depends on the population under study and arises from that population's evolutionary history. Alleles that have been in the population for a long time are more likely to have escaped genetic drift, to have been less subject to purifying selection, and even possibly to confer some weak selective advantage. Nevertheless, gene flow may have restricted the presence of an ancient allele in isolated populations, and the allele may not confer the same advantages or disadvantages on every genetic background.

Common variants have been associated with some complex traits. Recent examples include hippocampal

NOTES

volume (which is associated with incipient Alzheimer's disease but reduced in schizophrenia), major depression, and mesial temporal lobe epilepsy (Stein et al., 2012). The associated variant, rs7294919 (12q24.22; N = 21,151; $p = 6.70 \times 10^{-16}$), is intergenic, which raises a frequently encountered problem: how to interpret the mechanism by which a noncoding variant contributes to the phenotype. Another example does implicate a coding gene—glycerophosphocholine phosphodiesterase (*GPCPD1*) in the highly heritable trait of visual cortical surface area (Bakken et al., 2012), which correlates with visual acuity and visual perception. The significantly associated SNP (rs238295; $p = 6.5 \times 10^{-9}$) is located within 4 kb of the 5'UTR of *GPCPD1*, which in humans is more highly expressed in occipital cortex, compared with the remainder of cortex, than are 99.9% of genes genomewide. Late-onset Alzheimer's disease has been associated with common variants at *MS4A4/MS4A6E*, *CD2AP*, *CD33*, and *EPHA1* (Naj et al., 2011).

Alleles that have been in the population a relatively short time are more likely to be rare, to have a more limited distribution, and in the absence of sufficient time for purifying selection, to be deleterious. Owing to explosive population growth in certain geographic locales, the balance between rare and common mutations in many human populations has shifted toward an excess of rare genetic variants (Keinan and Clark, 2012). For example, rare copy-number variants and rare single nucleotide variants occurring as *de novo* mutations are important contributors to the autism phenotype (Sanders et al., 2012). Interestingly, multiple independent *de novo* single nucleotide variants in the same gene (e.g., sodium channel, voltage-gated, type II, α subunit) among unrelated probands were able to reliably identify risk alleles. In another study (O'Roak et al., 2012), *de novo* point mutations in autism spectrum disorder were found to be overwhelmingly paternal in origin (4:1 bias) and positively correlated with paternal age. In this study, 39% (49 of 126) of the most severe or disruptive *de novo* mutations mapped to a highly interconnected β-catenin/chromatin remodeling protein network with recurrent protein-altering mutations observed in two genes: *CHD8* and *NTNG1*. These instructive examples of both rare and common alleles contributing to genetic conditions point out that, when undertaking genomic analyses, dividing SNPs simply into the categories rare or common may be an oversimplification.

The genotype–phenotype interface

The transcriptome is the first phenotypic expression of the genome. Although RNA sequence space maps directly onto DNA sequence space, each genotype corresponds to multiple RNA secondary structures. Thus, RNA folding can be regarded as a minimal model of a genotype–phenotype relation (Fontana and Schuster, 1998b) and represents an enormous expansion of genotypic space. At this level, even neutral change in the genome will alter the "statistical topology" of the set of minimum free energy secondary RNA structures. These RNA structures exist as kinetic minima across a Waddingtonian landscape (Waddington, 1957).

In C.H. Waddington's well-known metaphor, one imagines marbles rolling down a hill and competing for grooves on the slope, in which they come to rest at the lowest points. Although Waddington used this imagery to represent developmental cell fates, it applies to many phenomena, including the variety of possible RNA secondary structures. RNA structures that arise from genotypic variation have been treated as evolutionary trajectories in which some transformations (including those that arise from neutral drift) are irreducibly discontinuous and likely play a key role in evolutionary optimization (Fontana and Schuster, 1998a; Stadler et al., 2001). A more in-depth understanding of RNA topologies may explain how SNPs in noncoding transcripts contribute to phenotypes.

High-Throughput Mapping of the Transcriptome

No high-throughput mechanism exists for determining RNA topologies. However, the technologies to determine RNA transcripts in a high-throughput manner, called RNA-Seq or Whole Transcriptome Shotgun Sequencing, are flourishing. The major platforms, providing what has been called deep sequencing or next-generation sequencing, are the Illumina Genome Analyzer (Illumina, San Diego, CA), ABI Solid Sequencing (Applied Biosystems, Carlsbad, CA), and 454 Life Sciences' Sequencing (454 Life Sciences, Branford, CT). These technologies offer deep coverage and base-level resolution from which one can perform several tasks: infer differential expression of genes, quantify allelic expression, determine differentially expressed spliced transcripts, detect noncoding RNAs, editing and gene fusions. Although these parameters do not capture the entire shape repertoire of RNA, they do represent an expansion of genotypic information because the deeply sequenced transcriptome is the product of both the genome and the epigenome. The epigenome controls transcript levels by way of histone modifications, DNA methylation, and chromatin accessibility as well as translation regulation through noncoding RNAs. Therefore, to fully understand

NOTES

the significance of the transcriptome, one requires knowledge of the underlying genome and epigenome.

The most informative expression sequencing techniques use the following:

- RNA-Seq libraries prepared with poly(A) primers to obtain mRNAs;
- RNA-Seq libraries prepared with random primers to obtain both mRNAs and noncoding transcripts;
- RNA-Seq libraries prepared from gel-purified small RNAs to obtain small noncoding transcripts (which in the brain are primarily the microRNAs); and
- RNA immunoprecipitation followed by sequencing (RIP-Seq), to obtain those RNAs that are immunoprecipitated with an antibody to an RNA-binding protein.

Because ribosomal RNA represents more than 90% of the RNA within cells, its removal increases the capacity to retrieve data from the remaining portion of the transcriptome. However, for samples with extremely small amounts of RNA (~100 ng) in which we could not risk further sample loss with a ribosomal removal step, we have sequenced very deeply and bioinformatically removed ribosomal sequences.

Aligning transcriptomes to genomic databases

Aligning transcriptomes to genomic reference databases faces the challenge of aligning transcripts, which are usually short reads, when they cross exon boundaries. To accomplish this task, specialized algorithms for transcriptome alignment have been developed, including TopHat (Trapnell et al., 2009) and Cufflinks (Trapnell et al., 2010). TopHat is a fast-splice junction mapper that aligns RNA-Seq reads to mammalian-sized genomes using the short-read aligner Bowtie. It then analyzes the mapping results to identify splice junctions between exons. Cufflinks assembles transcripts, estimates their abundances, and tests for differential expression and regulation in RNA-Seq samples. The program accepts aligned RNA-Seq reads, assembles the alignments into a parsimonious set of transcripts, and estimates the relative abundances of these transcripts based on how many reads support each one. To assess relative abundance, sequencing reads are often expressed as reads per kilobase (RPKMs) of exon model per million mapped reads (Mortazavi et al., 2008). These units reduce the error associated with unequal reads over all the exons of an mRNA. If one is doing paired end-reads, then the two fragments are counted together.

MicroRNAs are among the various categories of transcripts obtained by deep-sequencing techniques and are of particular interest to our group. We have published a comprehensive deeply annotated set of miRNAs from mouse hippocampus and staged sets of mouse cells that underwent reprogramming to induced pluripotent stem cells (iPSs) (Zhou et al., 2012). Using a dataset of more than 600 million deeply sequenced small RNAs, we annotated the stem–loop precursors of the known miRNAs in order to identify isomoRs (miRNA-offset RNAs), loops, nonpreferred strands, and guide strands. Products from both strands were readily detectable for most miRNAs. Changes in the dominant isomiR occurred among the cell types, as did switches of the preferred strand. The terminal nucleotide of the dominant isomiR aligned well with the dominant offset sequence, suggesting that Drosha cleavage generates most miRNA reads without terminal modification. Among the terminal modifications detected, most were nontemplated mononucleotide or dinucleotide additions to the 3'-end.

In addition to these descriptive features, the interpretation of the data was enhanced by performing RIP-Seq on an Ago-IP fraction. Ago or Argonaut proteins are key members of the RNA inhibitory silencing complex (RISC), which houses miRNAs as they form duplexes with mRNA targets. The binding between miRNAs and Ago proteins is very tight; therefore, Ago-IP can reveal a set of miRNAs associated with the RISC. This approach allowed us to predict which miRNA modifications—either isomiR modifications or nontemplated additions—might affect RISC loading. Furthermore, sequence variation of the two strands at their cleavage sites suggested higher fidelity of Drosha than Dicer.

Preparing iPS-derived neurons

iPSs offer the possibility of analyzing the complete transcriptome of any cell type against a specific individual's genetic background. The most common approach begins with harvesting skin fibroblasts from an individual of interest. A variety of reprogramming procedures have been described. These techniques are best divided into (1) direct reprogramming to the desired cell type and (2) reprogramming first to an embryonic stem cell, followed by subsequent differentiation to the desired cell type. Although some compelling approaches to direct reprogramming to neurons have been reported recently (Ring et al., 2012), we have generally transitioned cells through the embryonic stem-cell stage before differentiating them to neurons.

NOTES

Although complete control over neuronal fate in the dish still requires much further experimental work, differentiation procedures are selected based on the type of neurons one would like to grow. For example, techniques for growing motor neurons from stem cells are quite well developed (Soundararajan et al., 2006). The procedure involves treatment with a sonic hedgehog (Shh) agonist and retinoic acid (RA). To obtain a broad distribution of cortical neurons, we begin by withdrawing the β-FGF and add NT3, BDNF, or GDNF solutions. The cells pass through a neurosphere stage as neural precursors and get dissociated and plated to undergo neuronal differentiation on a laminin-coated surface. We have verified the neuronal identity of the cells by immunostaining with the following markers: MAP2, tau, synapsin, PSD95, as well as GFAP (to detect glial cells) and nestin (to detect neuronal precursors). We analyzed cultures for the colocalization of the presynaptic and postsynaptic markers (synapsin and PSD95) and for the polarization of the axonal and dendritic markers (tau and MAP2). In addition to immunocytochemical validation of neuronal identity, we have labeled cells with green fluorescent protein (GFP) or dye I to examine spine morphology and loaded cells with FM dye to analyze synaptic vesicle uptake and release.

Use of iPS-derived neurons to enhance the interpretation of genomes and transcriptomes

In collaboration with Fen Gao at the University of Massachusetts and Yadong Huang at the Gladstone Institute, we have prepared and analyzed human iPS cells that harbor tau mutations that are associated with neurodegenerative diseases. The iPS cells were first shown to be bona fide iPS cells based on the expression of pluripotency markers. Next, they were transformed into neurons (Wilson and Stice, 2006). Once the cells were well differentiated, a complete transcriptome was obtained. Full genomes were obtained on the same individuals.

The analysis of these data sets allows us to make several tentative conclusions:

(1) A broad range of neuronal types was present in the cultures, based on the expression of the various neurotransmitter receptor types, and their transcript levels were comparable to that in deeply sequenced brain tissue;
(2) Markers for glial cells were detectable but below the levels found in comparably analyzed brain tissue;
(3) Very low levels of transcripts related to neuronal precursors remained present in the culture;
(4) Potentially deleterious genetic variants in the genome, such as SNPs that alter splice sites, could be analyzed for their effect on transcription; and
(5) The distribution of a mutant allele could be determined as a function of the total reads for the transcript containing the variant. In this way, we could detect allele bias.

Conclusion

In summary, support for determining the significance of genomic variation can come from the transcriptome. The difficulty of obtaining tissue-specific gene expression in poorly accessible tissues such as the brain can be circumvented by using iPS technology and by differentiating iPS to specific cell types.

Acknowledgments

Support for this work was generously provided by the California Institute for Regenerative Medicine, The Tau Consortium, and the Errett Fisher Foundation.

References

Bakken TE, Roddey JC, Djurovic S, Akshoomoff N, Amaral DG, Bloss CS, Casey BJ, Chang L, Ernst TM, Gruen JR, Jernigan TL, Kaufmann WE, Kenet T, Kennedy DN, Kuperman JM, Murray SS, Sowell ER, Rimol LM, Mattingsdal M, Melle I, et al. (2012) Association of common genetic variants in GPCPD1 with scaling of visual cortical surface area in humans. Proc Natl Acad Sci USA 109:3985–3990.

Caselli RJ, Dueck AC, Osborne D, Sabbagh MN, Connor DJ, Ahern GL, Baxter LC, Rapcsak SZ, Shi J, Woodruff BK, Locke DE, Snyder CH, Alexander GE, Rademakers R, Reiman EM (2009) Longitudinal modeling of age-related memory decline and the APOE epsilon4 effect. N Engl J Med 361:255–263.

Chakravarti A (1999) Population genetics—making sense out of sequence. Nat Genet 21:56–60.

Fontana W, Schuster P (1998a) Continuity in evolution: On the nature of transitions. Science 280:1451–1455.

Fontana W, Schuster P (1998b) Shaping space: The possible and the attainable in RNA genotype–phenotype mapping. J Theor Biol 194:491–515.

Keinan A, Clark AG (2012) Recent explosive human population growth has resulted in an excess of rare genetic variants. Science 336:740–743.

Lander ES (1996) The new genomics: global views of biology. Science 274:536–539.

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

Naj AC, Jun G, Beecham GW, Wang LS, Vardarajan BN, Buros J, Gallins PJ, Buxbaum JD, Jarvik GP, Crane PK, Larson EB, Bird TD, Boeve BF, Graff-Radford NR, De Jager PL, Evans D, Schneider JA, Carrasquillo MM, Ertekin-Taner N, Younkin SG, et al. (2011) Common variants at *MS4A4/MS4A6E*, *CD2AP*, *CD33* and *EPHA1* are associated with late-onset Alzheimer's disease. Nat Genet 43:436–441.

O'Roak BJ, Vives L, Girirajan S, Karakoc E, Krumm N, Coe BP, Levy R, Ko A, Lee C, Smith JD, Turner EH, Stanaway IB, Vernot B, Malig M, Baker C, Reilly B, Akey JM, Borenstein E, Rieder MJ, Nickerson DA, et al. (2012) Sporadic autism exomes reveal a highly interconnected protein network of *de novo* mutations. Nature 485:246–250.

Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69:124–137.

Ring KL, Tong LM, Balestra ME, Javier R, Andrews-Zwilling Y, Li G, Walker D, Zhang WR, Kreitzer AC, Huang Y (2012) Direct reprogramming of mouse and human fibroblasts into multipotent neural stem cells with a single factor. Cell Stem Cell 11:100–109.

Sanders SJ, Murtha MT, Gupta AR, Murdoch JD, Raubeson MJ, Willsey AJ, Ercan-Sencicek AG, DiLullo NM, Parikshak NN, Stein JL, Walker MF, Ober GT, Teran NA, Song Y, El-Fishawy P, Murtha RC, Choi M, Overton JD, Bjornson RD, Carriero NJ, et al. (2012) *De novo* mutations revealed by whole-exome sequencing are strongly associated with autism. Nature 485:237–241.

Soundararajan P, Miles GB, Rubin LL, Brownstone RM, Rafuse VF (2006) Motoneurons derived from embryonic stem cells express transcription factors and develop phenotypes characteristic of medial motor column neurons. J Neurosci 26:3256–3268.

Stadler BM, Stadler PF, Wagner GP, Fontana W (2001) The topology of the possible: Formal spaces underlying patterns of evolutionary change. J Theor Biol 213:241–274.

Stein JL, Medland SE, Vasquez AA, Hibar DP, Senstad RE, Winkler AM, Toro R, Appel K, Bartecek R, Bergmann O, Bernard M, Brown AA, Cannon DM, Chakravarty MM, Christoforou A, Domin M, Grimm O, Hollinshead M, Holmes AJ, Homuth G, et al. (2012) Identification of common variants associated with human hippocampal and intracranial volumes. Nat Genet 44:552–561.

Trapnell C, Pachter L, Salzberg SL (2009) TopHat: Discovering splice junctions with RNA-Seq. Bioinformatics 25:1105–1111.

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

Waddington CH (1957) The strategy of the genes: A discussion of some aspects of theoretical biology. London: Allen & Unwin.

Wilson PG, Stice SS (2006) Development and differentiation of neural rosettes derived from human embryonic stem cells. Stem Cell Rev 2:67–77.

Zhou H, Arcila ML, Li Z, Lee EJ, Henzler C, Liu J, Rana TM, Kosik KS (2012) Deep annotation of mouse iso-miR and iso-moR variation. Nucleic Acids Res 40:5864–5875.

NOTES

Computational Analysis of RNA-Seq Data: From Quantification to High-Dimensional Analysis

Junhyong Kim, PhD

Department of Biology and Penn Genome Frontiers Institute
University of Pennsylvania
Philadelphia, Pennsylvania

NOTES

Introduction

High-throughput RNA sequencing is providing unparalleled resolution of the transcriptome and has been especially instrumental in revealing the transcriptome's sequence-level complexity (Core et al., 2008; Morin et al., 2008; Trapnell et al., 2009, 2010; Wang et al., 2009; Guttman et al., 2010; Nechaev et al., 2010; Pickrell et al., 2010; Marquez et al., 2012). In this chapter, I will discuss some of the computational and statistical challenges of quantifying the transcriptome from high-throughput RNA sequence data. I will also set forth the principles of high-dimensional data analysis using quantified transcriptomes. RNA sequencing, quantification, and data analysis share many of the same problems and solutions with microarray-based assays. Therefore, I will concentrate on issues more specific to RNA sequencing—especially RNA sequencing from single cells. I will assume familiarity with the overall experimental scheme for massively parallel short-sequence reads provided by instruments such as Illumina HiSeq (Illumina, San Diego, CA) and ABI SOLiD platforms (Applied Biosystems, Carlsbad, CA). It should be noted that many of the specific experimental, statistical, and computational problems are still being actively addressed in the field, so best practices for using massive RNA sequencing data for functional genomics are expected to continue evolving.

Transcript Quantification from RNA Sequencing Reads

Aligning the reads to a reference genome

When processing short reads from RNA sequencing, the key computational step consists of aligning the reads to a reference genome. If the sequence reads are exact copies of contiguous regions of the reference genome, this step is straightforward. However, the sequence reads may differ from the genomic sequence owing to several factors: errors in the sequence chemistry, artifacts created by the library construction step (e.g., concatemers or fusion of separate molecules), or biological RNA processing such as splicing and RNA editing. A more important source of variation is found in the polymorphism of the reference genome, which is likely to contain indels and single nucleotide polymorphisms (SNPs) not present in the sequenced strain. Therefore, the computational alignment of the reads to the genome must take into account such possible variations.

Algorithmic procedures and alignment strategies

The standard algorithmic procedure for dealing with such variations involves finding the best local alignment for subsequences of the reads to the subsequences to the genome, and assembling the matches while respecting the positional constraints. The simplest reasonable algorithm for this procedure goes through a number of steps that are proportional to the product of the read length and the reference genome length. However, while such a computation is feasible for any single read, it becomes computationally impossible when multiplied for tens or hundreds of millions of sequence reads. Therefore, available algorithms try to approximate the best solutions within a reasonable computing time. The main strategies involve indexing the reference genome or the read set with various kinds of k-mer seeds (the so-called filtration strategy) and using special data structures (e.g., suffix arrays) to organize all substrings of the sequences (Li and Homer, 2010).

Obtaining high-quality alignment involves tradeoffs in processing speed versus accuracy. Various strategies center around ways to allow for more sensitive alignments without exacting too high a computational penalty. One important consideration is that algorithms that allow gapped reads can be costly for computation. For genomic sequencing, an alignment algorithm that does not allow indels can generate a large number of false-positive SNPs. However, for RNA sequencing, transcript counts are the desired output, and false SNPs are not as important. Therefore, algorithms that try to increase the sensitivity of the alignment, say by allowing larger deviations, are more important than those that try to increase specificity and accuracy.

Increasing alignment specificity

Increasing alignment specificity may involve a consideration of the specific sequencing experiment. For example, *in vitro* transcription (IVT)–amplified RNA (Van Gelder et al., 1990), used in single-cell transcriptome analysis, tends to create short transcript templates with 5' poly-T leaders in the amplified RNA. Many of the fragments in the library will keep the 5' poly-T sequence, which needs to be trimmed for effective alignment. Aligning across potential splice variants also creates challenges, and typical strategies involve using known splice signals and intron–exon boundaries to increase the reference variants. However, using only known gene models may impede the detection of novel splice variants.

NOTES

Increasing algorithm sensitivity

One possible solution is to use a hierarchical processing strategy where the reads from a sample are processed through algorithms of increasing sensitivity. For example, the nearly exact reads might be first mapped using efficient algorithms, and the remaining reads might be processed through increasingly sensitive algorithms. The recently developed program RNA-Seq Unified Mapper (RUM) (Grant et al., 2011) utilizes such a strategy. The downside of this strategy is that computational time may greatly increase, depending on the particular sample. For example, processing 100 million 100 bp reads through fastest aligners (e.g., BOWTIE and BWA [Burrows–Wheeler Aligner]) (Langmead et al., 2009; Li and Durbin, 2009) takes ~30 CPU hours on typical computers, whereas RUM may take up to 1,500 CPU hours. (Note that RUM is just another variation of filtration strategy.) Another problem (in addition to computational time) is that with increasing sensitivity comes the potential increase in false-positive alignments. Because the algorithms involve heuristic tradeoffs between sensitivity and specificity, the researcher has to make a decision between optimizing these two objectives.

Importance for RNA sequencing

The key issue for RNA sequencing is whether different alignment strategies produce biased samples of true transcripts, regardless of their false-negative and false-positive rate. In our experience, there is a considerable variation in read counts mapped to specific transcript models, depending on the alignment algorithm used. Unfortunately, the particular types and degree of biases are still unresolved, and at this time, consistent comparison of datasets requires identical processing of the short read set (as discussed below under Complexities of Quantifying the Transcripts). Lastly, the relatively short length of the sequence reads in next-generation sequencing (100–150 bp) makes it very difficult to consider *de novo* genomes that do not have a reference sequence. The short reads are generally too brief to assemble into a unique transcriptome. However, recent computational approaches have been making progress toward recovering a large fraction of the transcriptome from *de novo* assembly (Grabherr et al., 2011). Also, longer reads from improved chemistry and coupling of paired-end or mate-paired sequencing from multiple insert libraries are expected to lead to effective characterization of novel transcriptomes in the near future.

Benefits of RNA Sequencing

Sequencing RNA provides three major benefits (albeit with caveats to be discussed in the following pages): precision, dynamic range, and the ability to detect novel transcripts.

Precision

Precision of RNA sequencing comes from the ability of a sequence read to uniquely identify the presence of a particular transcribed RNA. If we see a sequence in the high-throughput data that is sufficiently complex that it uniquely maps to the genome, there must be at least one RNA molecule that contains that sequence in the original library preparation. Sequencing chemistry can be surprisingly error-prone at the 3'UTR ends, but if a read maps uniquely to the genome within a prespecified mismatch tolerance, the presence of the molecule can be confirmed with high confidence. The only caveat here is contamination, which is not specific to the instrument, and the possibility of misalignment to a paralogous locus. Alignment to a paralogous locus can be a problem, especially if the study strain has polymorphisms vis-à-vis the available reference genome. Therefore, when considering singular sequence reads as possible evidence of a transcript, it is advisable to carry out additional alignment to the reference genome under less stringent criteria to confirm unique alignment. The numerical precision of the sequencing (i.e., the precision of relative counts of transcript molecules) depends on many factors that will be discussed further below.

Dynamic range

A key advantage of RNA sequencing is that the dynamic range of quantification can be modulated by the sequencing depth. The total number of reads required to recover a rare transcript depends on the cell (tissue) type and the distribution of the frequency of the transcript—that is, the expected frequency of the most highly expressed transcript, the expected frequency of the next most highly expressed transcript, etc. In various single-cell samples, we find a surprising diversity of transcriptome frequency distributions. For example, a mouse brown adipose cell sample recovers ~6,500 distinct transcripts with ~20 million mapped unique reads, whereas a rat cortex cell sample recovers ~17,000 distinct transcripts with ~10 million mapped unique reads.

We can approximately compute desired sequencing depth using a variation of the coupon collector's problem: Given the need to collect N distinct coupons in a game, what is the expected number of total coupons needed? In the optimal case, in which all distinct transcripts have equal abundance in the transcriptome, we need ~1.8 million mapped reads to recover 10,000 distinct transcripts with

NOTES

95% confidence (using Markov inequality). In our experience, a typical high-throughput sequencing experiment yields only 25% high-quality, unique paired-end RefSeq mapped reads. Therefore, under the optimal scenario, we need ~8 million in total read depth to recover 10,000 distinct transcripts with high probability. However, as mentioned, some transcripts are much more common than others, greatly skewing this computation. Assuming 100,000 total RNA molecules in a cell, and assuming only a single molecule of a rare transcript, similar computations suggest that we need ~100 million total reads to recover all transcripts (including the most rare transcript) with high confidence. Optimal read yield from Illumina HiSeq Systems is on the order of 350 million reads per lane. Therefore, these calculations suggest 3-fold multiplexing per lane to recover the rarest transcripts.

Ability to detect novel transcripts

As mentioned above, many studies using RNA sequencing are reporting novel transcripts. For example, using RNA sequencing from mechanically dissected dendritic samples, we found that up to 56% of the expressed genes in the mouse hippocampal cells and 50% of the expressed genes in the rat hippocampal cells show evidence of intronic sequences in the cytoplasm: cytoplasmic intron-sequence-retaining transcripts, or CIRTs (Bell et al., 2010; Buckley et al., 2011).

One characteristic of Illumina's sequencing chemistry is that, for every double-stranded template insert, reads are obtained from only the 5'UTR ends of the sense and antisense strand. The 3'UTR ends of the insert are read only if the insert size is smaller than the requested read length (see below). This chemistry produces a key asymmetry in the mapped reads. A given nucleotide will be covered by reads from both the sense and antisense directions only if the insert was smaller than the read length or the library fragmentation step induced cleavage randomly around the nucleotide. This means that if a transcript has a definite end (e.g., in the 5'UTR or the 3'UTR), the reads from the ends will be mostly from a single direction.

Figure 1 shows a moving window plot-of-read density for the 3'UTR end of the *Grin2b* gene from the rat hippocampal transcriptome. The red and blue lines show read density in each direction. Clearly visible is a shift in the density owing to the strand directional bias of the Illumina sequencing chemistry. This bias can be exploited by computing the differential of the read densities in the two directions, shown as black lines with blue fill. A sharp peak in the differential curve indicates the presence of a natural 3'UTR end of the transcript. The horizontal blue bar indicates previously annotated coding sequence and 3'UTR for this gene (thick and thin bars, respectively). As can be seen, these RNA sequence data indicate a novel 3'UTR for this gene. We have used this kind of computational procedure to map 3'UTR isoforms for the rat hippocampal transcriptome.

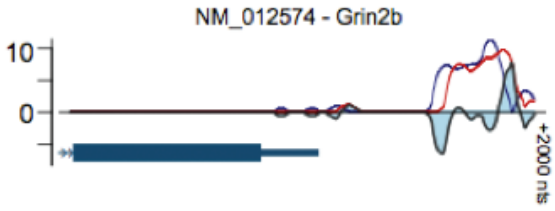

Figure 1. Read-density plot for the *Grin2b* locus. Blue denotes sense direction reads, red denotes antisense direction reads, and blue-filled black curved lines denotes differential in the two directions.

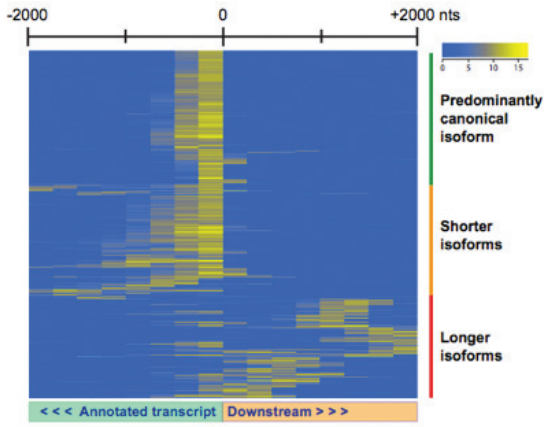

Figure 2. Heat plot of 3'UTR ends from rat hippocampal transcriptome. Rows indicate different genes. Position zero on the *x*-axis indicates canonical 3'UTR annotation.

Figure 2 shows a heatmap of estimated 3'UTR ends, where the coordinate 0 indicates the previously annotated 3'UTR for these transcripts. We found evidence that some genes have more than seven different end-isoforms, and two-thirds of the transcriptome show novel, previously unannotated 3' UTRs.

Complexities of Quantifying the Transcript

Once the short read set has been mapped to the reference genome, quantifying the transcript numbers has several complexities. We first assume that the RNA sample has been prepared to satisfactory quality

NOTES

(i.e., we assume that quality issues not specific to RNA sequencing are not part of the problem). The RNA pool is typically fragmented, cDNA is created to an appropriate size class, and adaptors are ligated for library amplification and sequencing.

Bias correction

Fragmentation and cDNA creation bias

Many authors have noted biases in the library resulting from both the fragmentation and cDNA creation step (Bullard et al., 2010; Hansen et al., 2010). Even without the bias, however, longer transcript molecules will be sampled more frequently during fragmentation and thus be more accurately measured, leading to greater statistical power for detecting differential expression (Oshlack and Wakefield, 2009). Several ad hoc bias correction methods have been suggested, but the optimal procedure is still uncertain at this point. In our experience, a pile-up visualization of the RNA sequencing reads on the genome shows clear heterogeneities. These include a large amount of reads that locate to a focal region or regions, with complete absence of reads despite high coverage in other adjacent regions. These kinds of variations are difficult to completely control and are likely to lead to artifactual theories of the transcriptome.

PCR bias

The PCR step in library construction can also lead to counts that are nonlinear in terms of input molecules and to a tendency to inflate the counts of more frequent molecules. The PCR bias can be modeled by noting the reads that map to nearly identical locations of the genome.

Associating read counts and normalizing read depth

The more critical problem is associating read counts to transcript models and normalizing the read counts to quantities that are comparable across different sequencing libraries. Different RNA preps and library preps yield different numbers of total reads and mapping reads. Initial attempts at quantification divided the reads mapping to a transcript model (e.g., RefSeq annotations) by the total number of mapping reads and the length of the transcript model. These calculations resulted in quantities such as reads per kilobase of exon model per million mapped reads (RPKM), which is still commonly used. Model-based methods have been proposed wherein the read coverage at any given base pair is assumed to be a Poisson sample with an unknown intensity parameter that represents the biological transcription level. Several variations of the model-based approach take into account possible intensity variation across a putative transcript molecule owing to such factors as fragmentation during library construction and convolution of biological variation from different samples.

Normalizing for read depth is also not so simple because the total mapped reads can be dominated by a small number of highly expressed genes. In such a case, there will be loss of sampling of more moderately expressed genes, distorting the estimate of relative expression levels. One simple corrective approach that has been suggested is to normalize the counts by a quantile of the read counts, such as the 75% quantile (i.e., every library is normalized such that the 75th percentile read count of a gene is 1).

Nonunique mapping reads and isoforms

The two largest problems with quantification are how to handle nonunique mapping reads and how to handle multiple isoforms of a given transcribed region of the genome. Nonunique maps can result either from redundant sequences of the genome or from overlapping transcriptional units. The former may be resolved with increasing sequence read length, but the latter has a biological origin and thus will be difficult to resolve without full-length sequencing of the transcript.

Isoforms of a transcript result from alternative splicing and lead to dependencies between reads and genomic regions: That is, the same read may result from multiple transcript molecules. Approaches to the isoform problem involve fitting the read data as samples from multiple transcript models. The models might involve using existing annotations of possible transcripts or estimating splice variants *de novo* by generating the best fitting models.

Variations among programs

Even when the algorithms do not try to deconvolute the read data into distinct isoforms, considerable variations can be found in the quantification because different programs handle the multiple reads and transcript models (i.e., the unit of quantification) differently. An important confounding factor is that these problems are sequence-specific and therefore affect different genes in different ways. A computational analysis of the mouse genome suggests that there are fewer than 1,000 possible transcripts without problems associated with transcript variations and overlapping transcript units.

NOTES

Evolving procedures to address complexity

A growing body of literature is addressing these quantification complexities, and we expect the procedures to evolve (Marioni et al., 2008; Bullard et al., 2010; Li et al., 2010; Trapnell et al., 2012). Some experimental protocols, such as ABI Solid SAGE (Applied Biosystems), attempt to characterize only 3'UTR tags, but we have found that the resulting sequences still contain potential artifacts that must be postprocessed. RNA sequences from IVT-amplified single cells have additional characteristics that modulate the quantification process. The IVT protocol involves transcript selection (using 3' poly-A or other A-rich sequences) and template-shortening due to multiple rounds of random hexamer priming. The template-shortening makes it less important to correct for length of the transcript model, but the template selection based on poly-A sequence requires one to consider the relationship of any other A-rich regions *cis* to the putative transcripts.

While these complexities may make RNA sequencing data seem hopelessly difficult to obtain, two facts should be recognized:

(1) Early microarray data required considerable research to arrive at uniform protocols for its usage; and
(2) Many of the complexities affect bias in transcript quantification, which may not be critical for most analyses.

Bias in the estimate of transcript levels can affect absolute quantification but will not affect analysis of differential expression or variational analysis (e.g., the variation associated with single cells).

There are two important caveats to consider going forward:

(1) If we find a significant difference between two samples, the difference may be the result of reads from overlapping maps. In this situation, the biological genesis of the difference may require further dissection that takes into account possibilities of splice isoforms, independent overlapping transcript units, and other sources of variation; and
(2) All quantitative comparisons across different samples need to be processed through the same computational pipeline; thus, it will be important to make the primary short-read data available for independent analyses.

Characterizing Transcriptome Variation

Jointly with the laboratory of Jim Eberwine, we have been characterizing transcriptome variations across individual cells of various cell types, especially CNS cells in rat and mouse. We typically collect RNA through mechanical isolation from dispersed primary cell culture. It is then amplified by IVT protocols, sequenced using the HiSeq platform (Illumina), mapped with the RUM pipeline, and quantified using custom programs. Once the transcriptome is quantified, the resulting data consist of a vector of numbers, representing the normalized read counts. The number of different transcripts depends on the experiment, but for the single cells we have assayed, the transcriptome ranges from ~6,000 to 14,000 different quantified units. We typically analyze the log transform of the read counts both because the RNA library is PCR amplified and because the RNA samples represent relative densities of RNA rather than absolute numbers. From here on, I assume that the data from each sample are represented by log-normalized read counts, which are equated to a vector in high-dimensional space (i.e., the dimensions correspond to distinct transcripts). Therefore, a dataset of multiple transcriptomes comprises a set of points in this high-dimensional space, which I will call the RNA state space (Kim and Eberwine, 2010).

Clustering analysis

It is now routine to perform clustering analysis of transcriptome data from multiple samples, typically with an accompanying heatmap representation of gene expression levels. Clustering analysis generally falls into the class of machine learning algorithms called "unsupervised learning." That is, the algorithms assume no prior information about the points but instead try to use the spatial distribution of the points to group them into clusters. The general idea is that biologically natural groups (such as distinct cell types and functionally coherent tissues) form spatial clumps in the high-dimensional space.

A whole constellation of algorithms exists, and these algorithms differ mainly as to how they interpret the spatial distribution (e.g., whether they consider certain directions more important than others) and how they impose prior ideas about the structure of spatial distribution (e.g., whether the distribution has a hierarchical organization). In terms of analyzing variation, clustering algorithms are useful for revealing distinct spacings or gaps between points and summarizing high-dimensional relationships that might be difficult to intuitively understand. Their downside is that different algorithms and measures of space within the RNA-state

NOTES

space can result in very different clusters, and there is very little guidance on the "correct" procedure.* Nonetheless, clustering the points gives important information on the degree of data heterogeneity, and we typically use the technique to complement other kinds of high-dimensional analysis.

Dimension-reduction techniques

A major problem with high-dimensional data is the number of dimensions itself. This is especially exacerbated in transcriptome data, where the number of variables (i.e., the different transcripts) vastly outnumbers the number of observations (e.g., cells, tissues, and experiments). This mismatch potentially leads to greatly overfitting complex models to sparse data. For example, given enough dimensions, one could easily come up with diagnostic markers for any reasonable classification of the input data.

Several important techniques have been developed to mediate this problem which typically involve either dimension reduction or methods to limit model complexity. Dimension reduction usually involves projecting the original high-dimensional data to lower dimensions. Projections involve taking the original high-dimensional points and projecting their positions onto some geometric object within that space, for example, a line. In fact, each individual coordinate can be seen as a particular projection onto a particular set of orthogonal lines.

Singular value decomposition and principle component analysis

Singular Value Decomposition (SVD) and the related Principle Component Analysis (PCA) have been used extensively in transcriptome analysis. In these techniques, an orthogonal set of linear projections are constructed in which each projection is, in effect, the line closest to the current distribution of points. These techniques transform coordinates into orthogonal coordinate axes, where each dimension can be ordered in terms of how much of the original dispersal pattern is captured on respective projections. This allows both visualization and dimensional reduction. For example, with the assumption that biologically meaningful transcriptome variation is found only in a small number of dimensions, the original data can be reduced to the projection in the PCA directions, and all subsequent analysis can be limited to the reduced dimensions. The caveat is that PCA directions typically tend to involve a very large number of genes, and therefore, the interpretation can become strained in terms of individual genes.

*It is important to note here that algorithmic "learning" from high-dimensional data is generally a difficult problem because it involves inferring potentially complex models of the data *ab initio* rather than fitting the data into simple models such as differential gene expression. Thus, statistics and mathematics used in many approaches have considerable degrees of freedom in determining the significance of any result.

Linear discriminant analysis

A useful dimension-reduction technique is Linear Discriminant Analysis (LDA), in which the projections to lines maximize the separation between a priori classes of points. Figure 3 shows a three-dimensional projection of a single-cell transcriptome from eight different cell types (shown in different colors) using PCA projections (Fig. 3A) and LDA projections (Fig. 3B). As can be seen in this picture, the LDA projections emphasize the separation of the different a priori classified cell types. In effect, each dimension in the LDA projections is a weighted combination of the expression level of genes that best separate the cell types.

Partial Least Squares

Another projection technique is Partial Least Squares (PLS). PLS projection is useful if there is another continuous response variable that is assumed to be a function of the transcriptome, e.g., cell size, cell physiology, or signaling output. The projection tries to find a set of orthogonal lines (directions) in the RNA-state space that best explains the response variable.

Nonlinear projections

Lastly, projections do not have to be linear (i.e., project to lines). For example, we might imagine that, given enough data points, the transcriptome from single hippocampal cells forms nonlinear curves in the transcriptome space (say, because the RNA products have to form dimers and satisfy quadratic stoichiometric relationships). Techniques such as Locally Linear Embedding (LLE) (Roweis and Saul, 2000) aim to detect and characterize such nonlinear geometric distributions.

Deriving transcriptomes from single cell types

The distribution of transcriptomes for single cells or tissues within the RNA-state space may have complex structures. One way to think about single-cell transcriptomes is that particular levels of RNA expression are maintained for a given cell because certain RNA molecules are required to satisfy the stoichiometric relationship of functional reactions involved in the cell's phenotypic function. For example, a neuron might require the maintenance of certain ratios of different glutamate receptors. The collective effect of such stoichiometric constraints limits the viable points in the RNA-state space for a particular cell type. If there are 10,000 different transcripts in the

NOTES

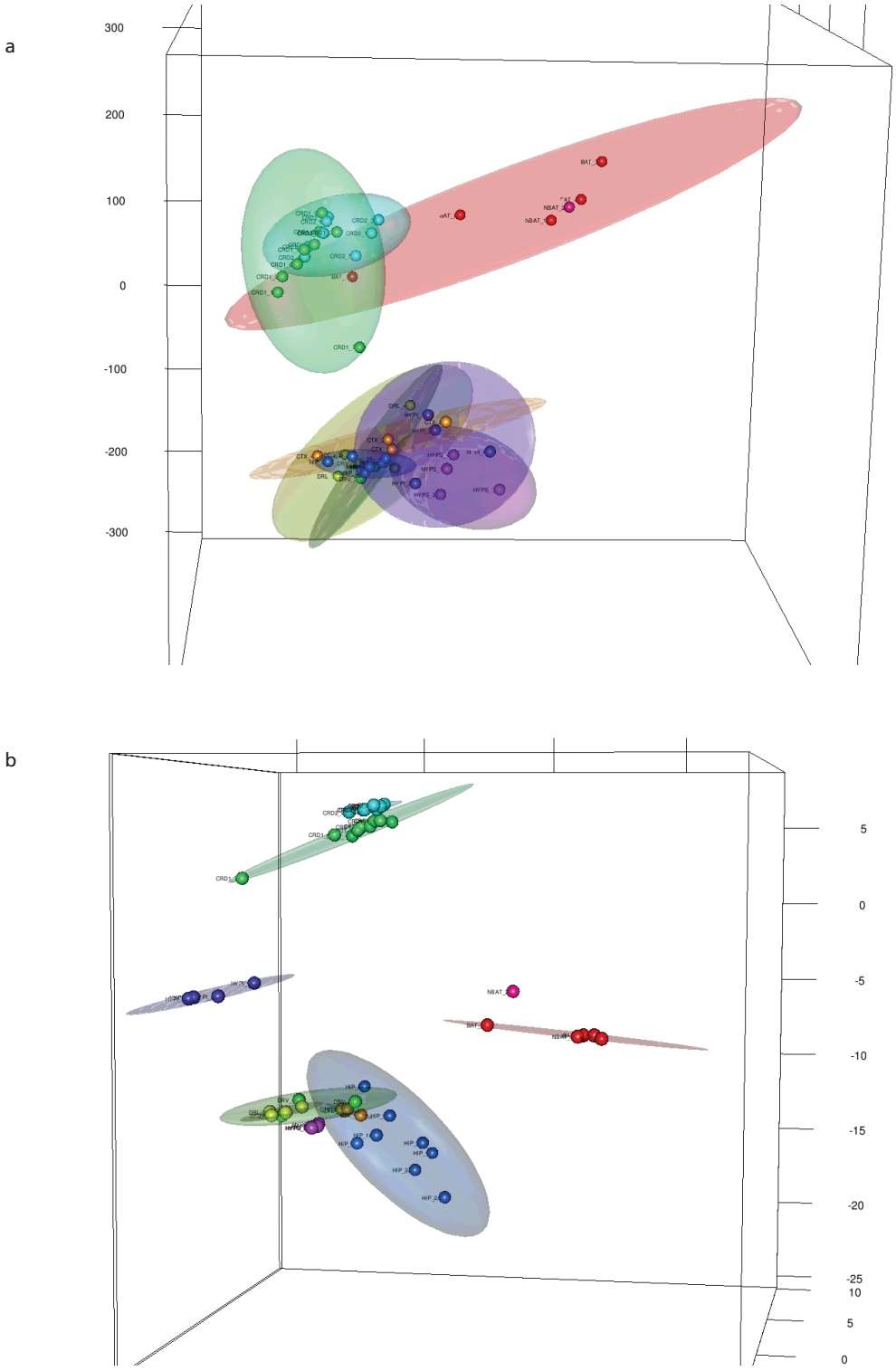

Figure 3. Three-dimensional projection of high-dimensional single-cell transcriptome data from 8 different cell types. ***a***, PCA axis projection; ***b***, LDA axis projection. The axes of both figures are abstract, and the numerical values represent linear combinations of the original variables. The values have no direct interpretation in terms of original data values. The PCA axes are meant to emphasize the overall variation, while the LDA axes emphasize the distinction between groups.

NOTES

cell, then each constraint reduces the viable dimension by one. If there are 10,000 constraints, then we might expect the transcriptome to maintain a particular set of expressions, i.e., be concentrated around a single point. If there are fewer than 10,000 constraints, then the transcriptome has multiple degrees of freedom and the single-cell transcriptomes might form a broad distribution, as seen in Figure 3.

Given enough data (i.e., transcriptomes from multiple single cells of the same type), it might be feasible to characterize the viable functional transcriptome states of a particular cell type using these projection techniques. It may also be possible to identify the physiological constraints for these cells' function. In the last part of the talk, I will present some potential models for analyzing such single-cell variation data.

References

Bell TJ, Miyashiro KY, Sul JY, Buckley PT, Lee MT, McCullough R, Jochems J, Kim J, Cantor CR, Parsons TD, Eberwine JH (2010) Intron retention facilitates splice variant diversity in calcium-activated big potassium channel populations. Proc Natl Acad Sci USA 107:21152–21157.

Buckley PT, Lee MT, Sul JY, Miyashiro KY, Bell TJ, Fisher SA, Kim J, Eberwine J (2011) Cytoplasmic intron sequence-retaining transcripts can be dendritically targeted via ID element retrotransposons. Neuron 69:877–884.

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

Core LJ, Waterfall JJ, Lis JT (2008) Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. Science 322:1845–1848.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol 29:644–652.

Grant GR, Farkas MH, Pizarro AD, Lahens NF, Schug J, Brunk BP, Stoeckert CJ, Hogenesch JB, Pierce EA (2011) Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq Unified Mapper (RUM). Bioinformatics 27:2518–2528.

Guttman M, Garber M, Levin JZ, Donaghey J, Robinson J, Adiconis X, Fan L, Koziol MJ, Gnirke A, Nusbaum C, Rinn JL, Lander ES, Regev A (2010) Ab initio reconstruction of cell type–specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. Nat Biotechnol 28:503–510.

Hansen KD, Brenner SE, Dudoit S (2010) Biases in Illumina transcriptome sequencing caused by random hexamer priming. Nucleic Acids Res 38:e131.

Kim J, Eberwine J (2010) RNA: State memory and mediator of cellular phenotype. Trends Cell Biol 20:311–318.

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

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

Li H, Homer N (2010) A survey of sequence alignment algorithms for next-generation sequencing. Brief Bioinform 11:473–483.

Li J, Jiang H, Wong WH (2010) Modeling non-uniformity in short-read rates in RNA-Seq data. Genome Biol 11:R50.

Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y (2008) RNA-Seq: An assessment of technical reproducibility and comparison with gene expression arrays. Genome Res 18:1509–1517.

Marquez Y, Brown JW, Simpson C, Barta A, Kalyna M (2012) Transcriptome survey reveals increased complexity of the alternative splicing landscape in *Arabidopsis*. Genome Res 22:1184–1195.

Morin RD, O'Connor MD., Griffith M, Kuchenbauer F, Delaney A, Prabhu AL, Zhao Y, McDonald H, Zeng T, Hirst M, Eaves CJ, Marra MA (2008). Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. Genome Res 18, 610–621.

Nechaev S, Fargo DC, dos Santos G, Liu L, Gao Y, Adelman K (2010) Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in *Drosophila*. Science 327:335–338.

Oshlack A, Wakefield MJ (2009) Transcript length bias in RNA-Seq data confounds systems biology. Biol Direct 4:14.

Pickrell JK, Marioni JC, Pai AA, Degner JF, Engelhardt BE, Nkadori E, Veyrieras JB, Stephens M, Gilad Y, Pritchard JK (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. Nature 464:768–772.

Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. Science 290:2323–2326.

Trapnell C, Pachter L, Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. Bioinformatics 25:1105–1111.

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

Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL, Pachter L (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7:562–578.

Van Gelder RN, von Zastrow ME, Yool A, Dement WC, Barchas JD, Eberwine JH (1990) Amplified RNA synthesized from limited quantities of heterogeneous cDNA. Proc Natl Acad Sci USA 87:1663–1667.

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63.

NOTES

NOTES

Insight into the Molecular Basis of Addiction from ChIP-Seq and RNA-Seq

Eric J. Nestler, MD, PhD

Fishberg Department of Neuroscience and Friedman Brain Institute
Mount Sinai School of Medicine
New York, New York

NOTES

Introduction: Models of Addiction

Repeated use of addictive drugs such as cocaine causes long-lasting changes in the brain's reward circuitry, a key component of which is the nucleus accumbens. Accordingly, a major goal in the field has been to uncover the molecular mechanisms underlying addiction-associated neuroadaptations in this brain region. It has been hypothesized that one such mechanism of drug-induced neuroadaptations is the regulation of gene expression (Nestler, 2001). Since then, numerous studies have documented altered expression of genes, through candidate gene approaches or through gene expression microarrays, in the nucleus accumbens, among them Freeman et al., 2001; McClung and Nestler, 2003; Yao et al., 2004; and Yuferov et al., 2005. In addition, several transcription factors have been shown to be altered in this brain region after chronic cocaine exposure. These factors include ΔFosB (a Fos family protein) (Nestler, 2008) and cAMP-response element binding protein (CREB) (Carlezon et al., 2005), both of which have been related directly to the behavioral abnormalities that characterize an addicted state.

Limitations

Investigations of cocaine-induced changes in gene expression to date have focused by necessity on measures of steady-state mRNA levels, which may not reflect the transcriptional regulation of the encoding gene. Rather, such measures provide a static snapshot of mRNA levels without yielding insight into how the genes are regulated by subsequent stimuli. Another limitation, regardless of the platform used, is the very high rates of false-positive and false-negative findings, even within the same laboratory. This variability has proven a major hindrance in gene discovery efforts and has made longitudinal studies (i.e., identifying in animal models long-lasting changes in gene expression that contribute to addiction vulnerability over a lifetime) very challenging.

Recent advances

Recent advances in chromatin biology have made it possible for the first time to extend this level of knowledge, based on gene expression microarrays, to direct examination of transcriptional mechanisms. Thus, we now know, largely from studies of nonneural tissue, that the state of activation or repression of a gene is typically reflected in the covalent modifications of histone proteins in the gene's vicinity as well as in a host of other chromatin modifications (Borelli et al., 2008). Recent work has demonstrated robust regulation of epigenetic mechanisms by drugs of abuse (McQuown and Wood, 2010; Robison and Nestler, 2011). We believe that analysis of the epigenetic landscape of genes will assist significantly in defining drug-induced changes in gene expression within the brain's reward circuitry.

Experimental Approaches

Next-generation sequencing technologies

Our approach to defining the drug "transcriptome" is to overlay several types of information from the coordinated use of RNA-Seq, ChIP-Seq, and related methods. RNA-Seq has many advantages over older microarray technology because it provides a far more complete and quantitative analysis of expressed RNAs within a microdissected brain region. It better captures and distinguishes between multiple splice variants of a gene, and it enables the analysis of several forms of noncoding RNAs, both long and short species. The limitation of RNA-Seq is that it results in extremely large and complicated datasets whose analysis is not as established as for microarrays.

Likewise, ChIP-Seq is superior to earlier methodologies, such as ChIP-chip (the analysis of immunoprecipitated DNA on promoter chips). ChIP-Seq provides far more precise information concerning the binding site of a particular transcription factor or modified histone. It is also more quantitative than ChIP-chip and far more comprehensive: ChIP-chip allows the analysis of only promoter regions, whereas ChIP-Seq provides a full genomewide view of chromatin modifications. This advantage is particularly important, since we know that nonpromoter regions are crucial for gene regulation and that regulation of nongenic regions likely contributes to the genomic effects of drugs of abuse.

There is particular interest in defining changes in DNA methylation genomewide in drug addiction models, based on the view that this regulatory mechanism likely contributes to the long-lasting effects of drug exposure on gene expression. Indeed, DNA methyltransferases and methyl-DNA binding proteins have been implicated in drug action (Deng et al., 2010; Im et al., 2010; LaPlant et al., 2010). However, genomewide approaches to analyze the specific genes affected by methylation in drug abuse models have not yet been reported.

Our published work to date validates this general approach to defining the drug transcriptome, although it has relied thus far on ChIP-chip technology. An example of our approach is shown in Figure 1. We first used ChIP-chip to define gene promoters in nucleus

NOTES

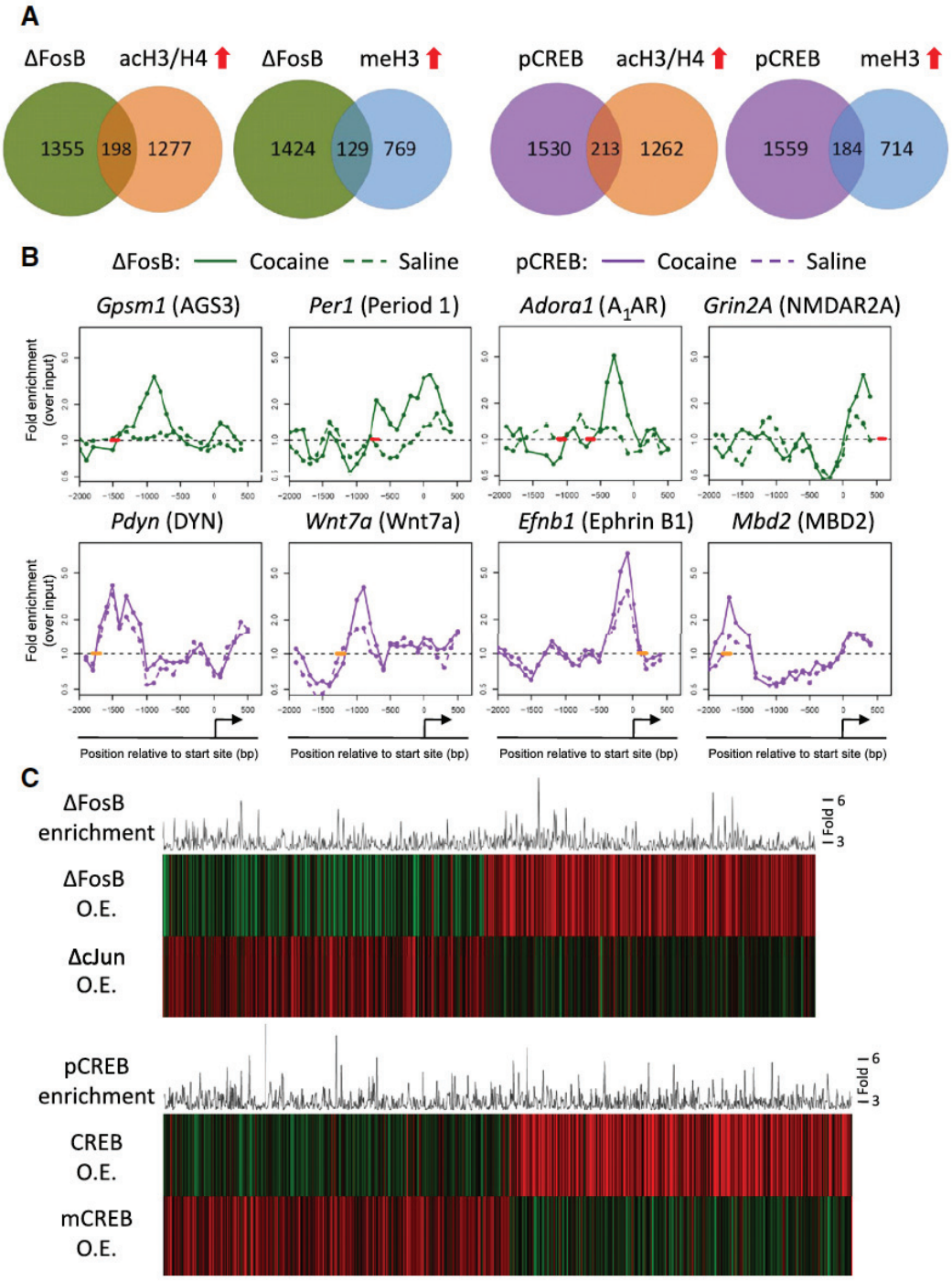

Figure 1. Regulation of ΔFosB and phospho-CREB binding at gene promoters in nucleus accumbens by chronic cocaine administration. ***A***, Venn diagrams of genes that show significant levels of ΔFosB or phospho-CREB binding, or of H3/H4 acetylation or H3 methylation (dimethylation of Lys 9 or 27), after 7 days of cocaine. ***B***, Patterns of ΔFosB (green) and phospho-CREB (purple) binding at representative gene promoters after chronic cocaine (solid line) or saline (dotted line) treatment. Short bold lines on the *x*-axes indicate positions of consensus or near-consensus AP1 (red) or CRE (orange) sites. ***C***, The top panel illustrates significant ΔFosB target genes from ChIP-chip (histogram) after chronic cocaine exposure and how expression of the encoded mRNAs are regulated upon inducible overexpression of either ΔFosB or its dominant negative antagonist ΔcJun in nucleus accumbens (heatmaps) ($\rho = -0.09$, $p = 0.005$). The bottom panel illustrates significant phospho-CREB target genes from ChIP-chip (histogram) after chronic cocaine exposure and how expression of the encoded mRNAs is regulated upon inducible overexpression of either CREB or its dominant negative antagonist mCREB (heatmaps) in the nucleus accumbens ($\rho = -0.3$; $p < 1E\text{-}16$). Renthal et al. (2009), their Figure 2, reprinted with permission.

NOTES

accumbens that show enrichment of ΔFosB or phospho-CREB (the active form of CREB) binding as well as a change in histone H3 or H4 acetylation (a major mark of gene activation) or a change in repressive histone methylation (dimethylation of Lys 9 or 27 on H3) in response to chronic exposure to cocaine (Fig. 1*A*) (Renthal et al., 2009). Examples of specific genes are shown in Figure 1*B*. We then asked how these patterns of chromatin modifications relate to earlier studies where we defined the mRNAs that are regulated in nucleus accumbens upon the inducible overexpression of ΔFosB or CREB, or a dominant negative antagonist of these transcription factors, in this brain region (McClung and Nestler, 2003). As Figure 1C shows, there is considerable overlap across these various levels of analysis, which provides important validation of this experimental approach and reveals a subset of genes that shows robust regulation by cocaine. Indeed, genes that display regulation across these several platforms of analysis can be viewed as bona fide cocaine targets, since ~90% are validated upon analysis of independent tissue samples.

Role of sirtuins

These gene discovery efforts are providing novel understanding into the molecular basis of addiction. An example is provided by the sirtuin class of proteins (categorized as Class III histone deacetylases), which regulates numerous cell functions through the deacetylation of histones and many other proteins. Sirtuins had not been implicated in drug abuse until our genomewide studies, which found significant induction of H3 acetylation at the SIRT1 and SIRT2 gene promoters, along with enrichment of ΔFosB at SIRT2, in nucleus accumbens after chronic cocaine administration (Renthal et al., 2009). Based on these findings, we validated the hypothesis that chronic cocaine use indeed induces SIRT1 and SIRT2 mRNA expression as well as catalytic activity in this brain region. We subsequently demonstrated that local activation of SIRTs increases the firing rate of nucleus accumbens neurons and enhances behavioral responses to cocaine, including increased cocaine self-administration, whereas inhibition of SIRTs in this region exerts the opposite effects (Renthal et al., 2009). These findings illustrate how the genomewide approaches described here can lead to fundamental new insights into how cocaine changes the brain to cause addiction.

Chromatin studies

Chromatin studies also make it possible to go well beyond a static view of steady-state mRNA levels. These studies are able to identify those genes primed (sensitized) for greater induction, or desensitized for diminished induction, as a consequence of prior drug exposure, providing a far more dynamic view of gene regulation (Maze et al., 2010; Robison and Nestler, 2011). For example, we are finding that certain chromatin marks (e.g., the binding of certain phosphorylated forms of RNA polymerase II to gene promoters) provide a mark of genes that are poised for sensitized induction in response to subsequent exposure to cocaine (Damez-Werno et al., 2012).

Importantly, chromatin studies provide the first ever insight into the molecular mechanisms underlying gene regulation in the brain *in vivo* (Robison and Nestler, 2011). By contrast, all prior studies have relied by necessity on examining the mechanisms underlying a change in mRNA levels by working in cell culture, even though we know that what happens in cultured cells—even cultured neurons—does not accurately reflect what occurs in the intact brain. Epigenetic studies thus reveal truly unique insight into the molecular mechanisms underlying addiction. Our work on ΔFosB and CREB, and associated histone modifications, represents early efforts to define the precise transcriptional steps through which cocaine alters the epigenetic status at specific genes in concert with their regulation within the nucleus accumbens *in vivo*.

Finally, work at the chromatin level is beginning to define the effects of drug exposure at nongenic regions. In a recent study, we found that chronic administration of cocaine decreases total levels of H3K9me3 (trimethylation of Lys9 of H3) in nucleus accumbens, as measured by Western blotting and immunohistochemistry (Maze et al., 2011). This was a surprising finding, since H3K9me3 has been shown in other, nonneural systems to be concentrated in heterochromatic regions of the genome where it would not be expected to regulate gene expression (Barski et al., 2007). This finding was also confirmed for nucleus accumbens by ChIP-Seq, where we found the H3K9me3 mark to be enriched almost exclusively at nongenic regions and yet to be dynamically regulated by cocaine (Maze et al., 2011). In fact, we found that chronic cocaine use reduces H3K9me3 levels at many repetitive genomic sequences, including at LINE-1 (long interspersed nuclear element-1) repeats, and increases expression of LINE-1 retrotransposons in this brain region. Although the functional consequences of this regulation remain unknown, these observations further illustrate the complex genomic regulation that a drug of abuse induces within specific regions of brain: information accessible only through advanced sequencing methodologies.

NOTES

Future Directions

Our ultimate goal is to define codes of chromatin modifications that can be used to predict altered steady-state RNA expression levels or active priming or desensitization of genes in response to subsequent stimulation. A related goal is to identify codes of chromatin modifications that can predict highly stable modifications (as opposed to the majority of chromatin changes observed to date, which are highly labile) and thereby mark genes that are good candidates for mediating the very persistant (in some cases, life-long) nature of addiction. Whether such codes in fact exist remains unknown, as work to date has revealed an extraordinary complexity of chromatin mechanisms in the brain. Nevertheless, delineating such mechanisms of gene regulation will facilitate the identification of the genes and biochemical pathways involved in distinct aspects of the addiction syndrome and should provide new pathways forward in the development of novel therapeutics for drug addiction.

References

Barski A, Cuddapah S, Cui K, Roh TY, Schones DE, Wang Z, Wei G, Chepelev I, Zhao K (2007) High-resolution profiling of histone methylations in the human genome. Cell 129:823–837.

Borrelli E, Nestler EJ, Allis CD, Sassone-Corsi P (2008) Decoding the epigenetic language of neuronal plasticity. Neuron 60:961–974.

Carlezon WA Jr, Duman RS, Nestler EJ (2005) The many faces of CREB. Trends Neurosci 28:436–445.

Damez-Werno D, LaPlant Q, Dietz DM, Sun SH, Scobie KN, Walker I, Koo JW, Mouzon E, Russo SJ, Nestler EJ (2012) Drug experience epigenetically primes Fosb gene inducibility in rat nucleus accumbens and caudate putamen. J Neurosci, 32:10267-10272.

Deng JV, Rodriguiz RM, Hutchinson AN, Kim IH, Wetsel WC, West AE (2010) MeCP2 in the nucleus accumbens contributes to neural and behavioral responses to psychostimulants. Nat Neurosci 13:1128–1136.

Freeman WM, Nader MA, Nader SH, Robertson DJ, Gioia L, Mitchell SM, Daunais JB, Porrino LJ, Friedman DP, Vrana KE (2001) Chronic cocaine-mediated changes in non-human primate nucleus accumbens gene expression. J Neurochem 77:542–549.

Im HI, Hollander JA, Bali P, Kenny PJ (2010) MeCP2 controls BDNF expression and cocaine intake through homeostatic interactions with microRNA-212. Nat Neurosci 13:1120–1127.

LaPlant Q, Vialou V, Covington HE 3rd, Dumitriu D, Feng J, Warren BL, Maze I, Dietz DM, Watts EL, Iñiguez SD, Koo JW, Mouzon E, Renthal W, Hollis F, Wang H, Noonan MA, Ren Y, Eisch AJ, Bolaños CA, Kabbaj M, Xiao G, Neve RL, Hurd YL, Oosting RS, Fan G, Morrison JH, Nestler EJ (2010) Dnmt3a regulates emotional behavior and spine plasticity in the nucleus accumbens. Nat Neurosci 13:1137–1143.

Maze I, Covington HE III, Dietz DM, LaPlant Q, Renthal W, Russo SJ, Mechanic M, Mouzon E, Neve RL, Haggarty SJ, Ren YH, Sampath SC, Hurd YL, Greengard P, Tarakovsky A, Schaefer A, Nestler EJ (2010) Essential role of the histone methyltransferase G9a in cocaine-induced plasticity. Science 327:213–216.

Maze I, Feng J, Wilkinson MB, Sun H, Shen L, Nestler EJ (2011) Cocaine dynamically regulates heterochromatin and repetitive element unsilencing in nucleus accumbens. Proc Natl Acad Sci USA 108:3035–3040.

McClung CA, Nestler EJ (2003) Regulation of gene expression and cocaine reward by CREB and ΔFosB. Nat Neurosci 11:1208–1215.

McQuown SC, Wood MA (2010) Epigenetic regulation in substance use disorders. Curr Psychiatry Rep 12:145–153.

Nestler EJ (2001) Molecular basis of long-term plasticity underlying addiction. Nat Rev Neurosci 2:119–128.

Nestler EJ (2008) Transcriptional mechanisms of addiction: role of deltaFosB. Philos Trans R Soc London B Biol Sci 363:3245–3255.

Renthal W, Kumar A, Xiao GH, Wilkinson M, Convington HE III, Maze I, Sikder D, Robison AJ, LaPlant Q, Dietz DM, Russo SJ, Vialou V, Chakravarty S, Kodadek TJ, Stack A, Kabbaj M, Nestler EJ (2009) Genome wide analysis of chromatin regulation by cocaine reveals a novel role for sirtuins. Neuron 62:335–348.

Robison AJ, Nestler EJ (2011) Transcriptional and epigenetic mechanisms of addiction. Nat Rev Neurosci 12:623–637.

Yao WD, Gainetdinov RR, Arbuckle MI, Sotnikova TD, Cyr M, Beaulieu JM, Torres GE, Grant SG, Caron MG (2004) Identification of PSD-95 as a regulator of dopamine-mediated synaptic and behavioral plasticity. Neuron 41:625–638.

Yuferov V, Nielsen D, Butelman E, Kreek MJ (2005) Microarray studies of psychostimulant-induced changes in gene expression. Addict Biol 10:101–118.

Single-Cell Transcriptomics in the Brain

Ditte Lovatt, PhD, Tae Kyung Kim, PhD, Peter Buckley, PhD, Jennifer M. Singh, PhD, and James Eberwine, PhD

Department of Pharmacology
University of Pennsylvania
Philadelphia, Pennsylvania

NOTES

Introduction to Transcriptomics Analysis

Transcriptomics analysis provides valuable qualitative and quantitative information about the global set of messenger RNAs (mRNAs) in a given sample. From such studies, tens of thousands of transcripts can be investigated simultaneously, from which information can be inferred about the sample's biochemical and functional properties. The most frequently used methods today to sample the transcriptome are cDNA microarrays and next-generation RNA sequencing (RNA-Seq) (Lockhart and Winzeler, 2000; Mortazavi et al., 2008).

Microarrays

Microarrays have been used extensively over the past decade in order to investigate the relative expression of specific mRNAs among different cell samples (Table 1). The major drawback of microarrays is that detection is based on the hybridization signal between an oligonucleotide anchored onto the chip and the fluorescently tagged nucleotide sample. This detection principle requires prior knowledge about the nucleotide sequences to be investigated and cannot lead to discoveries about, for instance, novel transcripts, splice variants, and retained introns. Also, the inherently high background noise on most commercial microarrays makes distinguishing between low-abundant RNAs and false-positives difficult, so such information must be validated using other methods, e.g., *in situ* hybridization. However, microarrays do provide a robust method for investigating sequence-specific mRNA abundances and thus, they remain a powerful quantitative method for most transcripts.

RNA sequencing

In contrast, the recent development of RNA-Seq has made unbiased mRNA sequence examination possible and eliminated concern about low-abundant transcripts, false-positives, and prior knowledge about sequence information. As in the examples given below, RNA-Seq makes unbiased sequence discoveries possible and has been applied to solving a variety of problems and discoveries, e.g., retained-introns, alternative splicing, and microRNAs (miRNAs). Although the algorithms for comparative quantification of specific transcripts are still being developed for RNA-Seq, this method is far superior to microarrays and provides a vast amount of detailed sequence information (Wang et al., 2009).

Transcriptome Data: One of a Kind or Just Average?

While transcriptome-generating methods can produce a vast amount of expression data, the interpretation of such data depends entirely on the type of sample. Studying the transcriptome of pools of cells provides a unique window into their biochemistry and function; however, information about cell-to-cell variability is lost. This becomes especially significant if the pool of cells is very heterogeneous, such as in intact brain tissue. The advantage of performing single-cell transcriptomics can easily be appreciated when considering the effect of averaging over the entire pools of cells. Furthermore, several single-cell transcriptome studies have concluded that single cells, even of the same type, are unique, and their subtleties of expression differences can have important biological functions. Limited information exists about the transcriptional differences among single neurons *in vivo*. Even so, one can easily speculate about how single neurons provide an especially unique system with inherent single-cell variability that may account for the differences in functional properties of those neurons and permit plasticity-associated changes.

The transcriptome of mRNA extracted from bulk tissue will give insight into the types of mRNAs species in the tissue. However, information about

Table 1. Comparison of transcriptome analysis methods[a]

Property	Microarrays	RNA-Seq
Quantitative	**	***
Qualitative	**	***
Low-abundant mRNA detection	*	***
Generation of false-positives	***	*
Cost[b]	*	***

[a] Increasing numbers of asterisks signify increased ability to generate the itemized data, e.g., RNA-Seq.
[b] More asterisks signify increased cost.
***, generates more quantitative data than microarrays, **.

NOTES

individual cells where these mRNAs originated from will be lost (Fig. 1). The result of this "averaging effect" will mask information about mRNA species that are present only in a subset of cells in the tissue, as their impact will be diluted. If microarray profiling is used, such subset-specific transcripts may be diluted out even beyond the detection limit; consequently, they will not be detected at all. In addition, the use of transcriptome profiling on mRNA samples from bulk tissue or pools of cells prevents us from distinguishing whether two mRNA species, *X* and Y, that function in the same signaling pathway are coexpressed in the same cell or expressed by different cells. Inferring information about the regulation of existing pathways is therefore also compromised. Although network analysis has been applied to transcriptome data from bulk tissue as a pathway-mapping tool in individual cell types, for the reasons stated above, this method can mistakenly conclude the existence of pathways. That is, *X* may be expressed exclusively by cell *A*, and *Y* may be expressed exclusively by cell *B*, so although they both are detected in the bulk mRNA, they actually cannot interact (Fig. 1).

The averaging effect will also mask information about cell-to-cell variability in the expression level of mRNAs that are expressed by the majority of cells in the tissue (Fig. 1). For instance, the majority of cells may express transcript *Z*, but at the single-cell level, *Z* can be expressed at either a high level, a low level, or not at all. However, such information is masked by the averaging effect, another argument for why single-cell transcriptomics is crucial to employ for questions related to how single cells function and interact with one another.

Single-Cell mRNA Isolation Methods

Clearly, the averaging effect has an important impact on the interpretation of transcriptome data if the tissue contains several unique cells or cell types. This is particularly true for brain tissue, which contains neurons, glia, and vascular cell types. In order to investigate how these cell types differ and how cell-to-cell variability characterizes single cells, one must apply transcriptomics to single cells instead of bulk tissue. The use of dispersed cell cultures could be an option to easily isolate single cells and perform transcriptomics on such samples. Nevertheless, the ultimate capture of a transcriptome is to sample cells from live intact tissue, in which all the synaptic architecture and cell-to-cell interactions are still in place.

Live intact tissue sampling can be accomplished by using acutely cut live brain slices or sampling cells in live animals through cranial windows. However, several technical obstacles prevent single-cell mRNA isolation in intact tissue. Brain tissue is very heterogeneous, and most cell types within the brain

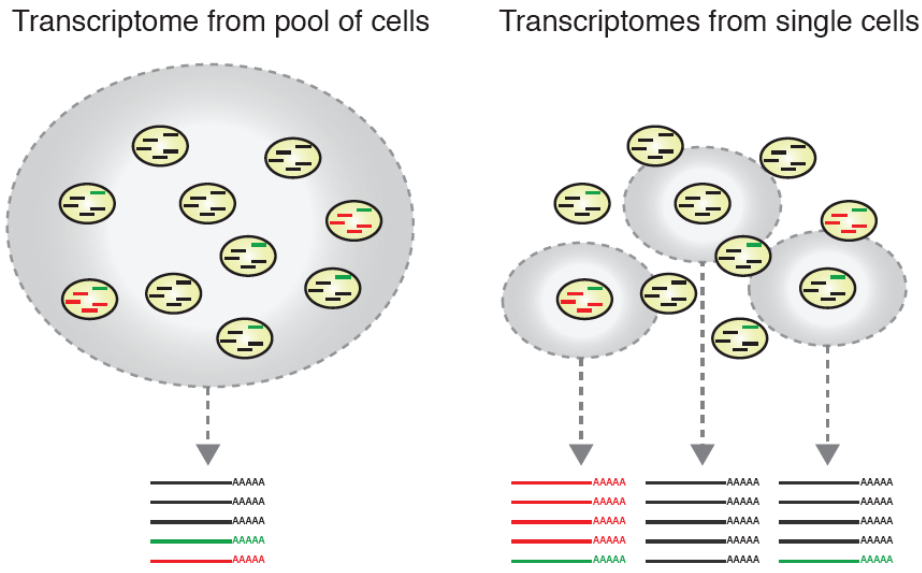

Figure 1. The "averaging effect." Transcriptomes of mRNA from bulk tissue (left) that comprises many cell types is subjected to an averaging effect in which mRNA data from each is averaged. This effect results in the dilution of mRNA species that are only present in a subset of cell transcripts (green and red). In contrast, transcriptomes from single cells (right) precisely report the abundances of each mRNA specie relative to other mRNA species in that particular cell.

NOTES

have polarized and highly branched morphologies that intermingle. This anatomical feature compromises our ability to isolate single cells because the degree of contamination from neighboring cells is significant. Although laser capture microdissection (LCM) is capable of isolating single cells from frozen and fixed tissue (Espina et al., 2006; Tang et al., 2009), these procedures adversely affect RNA quality. Also, LCM adds a significant degree of RNA contamination from neighboring structures to the dissected sample compared with other intact-tissue RNA isolation methods, such as fluorescence-activated cell sorting (FACS), immunopanning, and manual sorting (Okaty et al., 2011).

Perhaps the most successful single-cell RNA isolation method used in intact tissue to date is the micropipette approach, which isolates cytosolic mRNA from whole-cell patched cells by aspirating the cytosol (Surmeier et al., 1996; Martina et al., 1998). This method has been used in a variety of cell types including neurons of the preoptic area of the mouse hypothalamus, pyramidal neurons of the hippocampus, and serotonergnic neurons of the raphe. Indeed, it was using this approach that researchers first demonstrated that hundreds of G-protein coupled receptor (GPCR) genes can be expressed in a single cell. Data such as these offer a rationale for choosing receptor agonists and antagonists, to be used alone or in combination with other drugs (e.g., 5-HT_1 agonists), for physiological testing in selected cells or to study specific behavioral responses.

However, it should be noted that the use of a patch pipette to harvest cells from a live slice will cause mechanical damage to the slice. Thus, the development of an RNA isolation method that could isolate mRNA from single cells (or even subcellular structures, like dendrites) without contamination or induction of injury-related pathways would provide a valuable tool for studying single-cell transcriptomics.

Once mRNA has been isolated from a single cell, it has to be processed to prepare it for the downstream transcriptome method. First the mRNA has to be amplified, since the amount of mRNA from a single cell falls in the hundreds of femptograms-to-picograms range—far below the detection limit of most transcriptome methods, including RNA-Seq. In order to perform quantitative transcriptome analysis, it is crucial to use linear amplification (as opposed to PCR amplification) of the mRNA to maintain the stoichiometry among the different mRNA species (Morris et al., 2011). Linear amplification techniques are well developed. Following amplification of the mRNA to micrograms of amplified RNA (aRNA), either the aRNA need to be processed for microarray or RNA-Seq libraries need to be constructed. Altogether, these processing procedures take approximately one week before the prepped sample can be submitted to a microarray or RNA-Seq facility.

Single-Cell Transcriptomics to Distinguish TIPeR Transcriptome Transfer

Transcriptome profiling of single cells can be used to address a variety of scientific problems. In our lab, we previously used single-cell transcriptomics to validate transcriptome-induced phenotype remodeling (TIPeR)–mediated cell-to-cell transcriptome transfers (Sul et al., 2009). TIPeR is the process by which RNA populations are transferred into single cells to alter or remodel their phenotype. A successfully remodeled TIPeR cell will gradually change its transcriptome through activation and suppression of host-cell transcriptional pathways from the host cell toward that of the desired cell type. This process eventually gives rise to new cellular phenotypes in the TIPeR cells and, potentially, may be used in cell-replacement therapies.

To validate the transfer of the TIPeR cells, poly-A+ tailed mRNA from single TIPeR cells is linearly amplified (Morris et al., 2011) and processed for microarray or RNA-Seq analysis. Microarray data from TIPeR cells can be analyzed using conventional analysis software with modified algorithms that account for the 3' end amplification bias. To deal with the bias, these algorithms extract the second highest intensity values from each probe set and use them for quantitative expression analysis. Once the program obtains expression values, it selects probe sets based on their ability to distinguish the donor from the recipient TIPeR cells. Most often, such probe sets are cell-type–specific transcripts. The analysis results are presented in the form of clustering, differential gene expression profiles, and gene ontology tables to show the difference between TIPeR cells and non-TIPeR cells.

In a previous study, we used the TIPeR approach to transfer the transcriptome of cardiomyocytes into mouse fibroblasts, which converted the phenotype of the fibroblasts into cardiomyocyte-like cells (tCardiomyocyte) (Kim et al., 2011). Besides examining phenotypic signs of successful conversion, the TIPeR process was validated using single-cell transcriptomics, as described above. To this end, we isolated poly-A^+ RNAs from single adult cardiomyocytes, tCardiomyocytes, control cells, and

NOTES

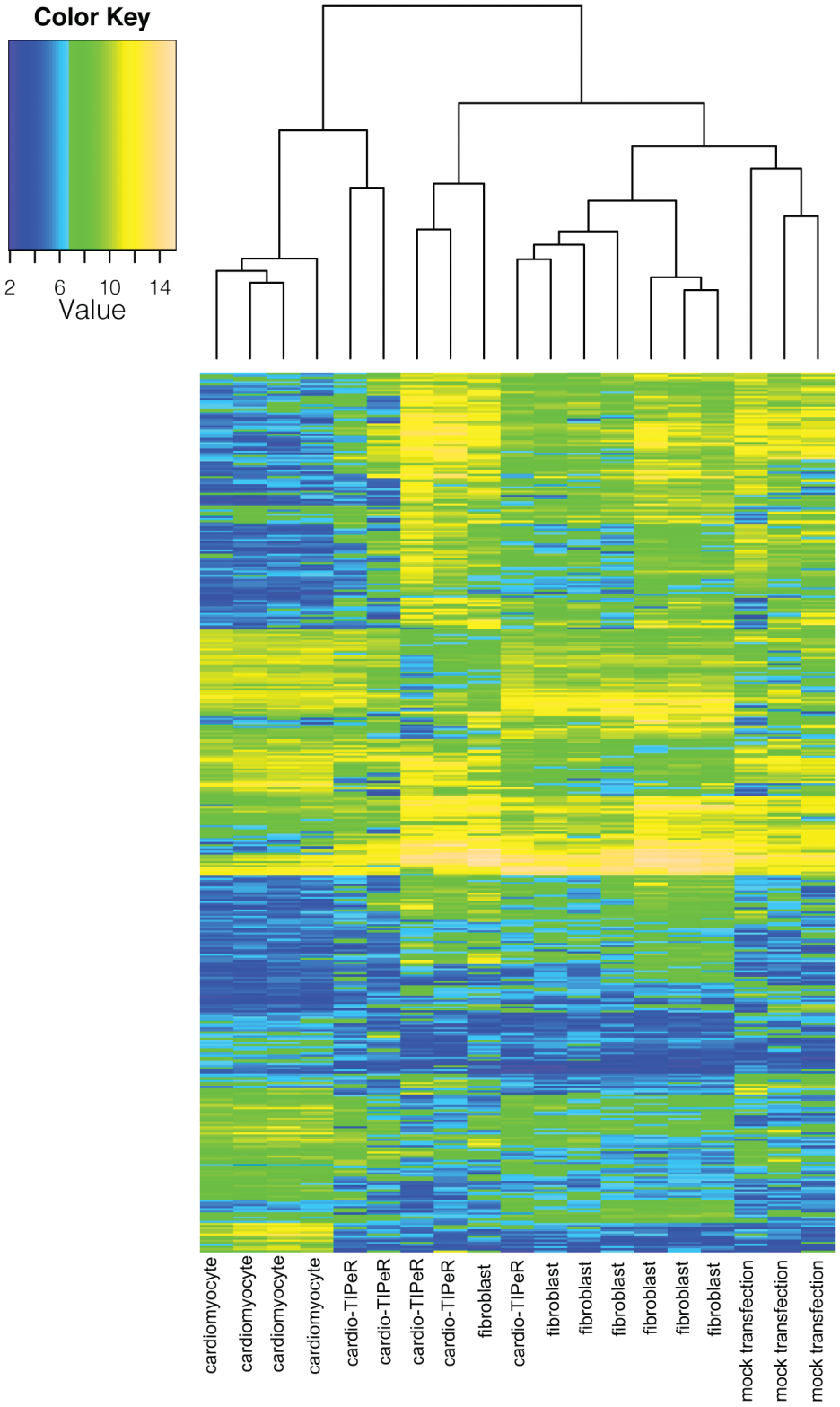

Figure 2. Transcriptomics used to validate TIPeR-mediated phenotype conversion. Global gene expression of tCardiomyocytes is reprogrammed toward adult cardiomyocytes. Dendrogram and heatmap show hierarchical clustering (Euclidean distance, complete linkage) of single cardiomyocytes, fibroblasts, cardio-TIPeR, and mock transfection using the expression values of 418 informative genes.

NOTES

fibroblasts 3–4 weeks after transfection and then amplified and processed the mRNA for microarray and transcriptome analysis. A comparison of the transcriptomes showed that tCardiomyocytes clustered closely with adult cardiomyocytes and far from fibroblasts, as expected for successful TIPeRing. However, not all TIPeR cells clustered with fibroblasts, suggesting that some TIPeR cells are not remodeled completely (Fig. 2). Global gene expression profiles also show that the expression pattern of differentially regulated genes was similar between tCardiomyocytes and adult cardiomyocytes but differed from TIPeR control cells or fibroblasts (Fig. 2). In conclusion, single-cell transcriptomics is a powerful method to validate TIPeR cells.

Single-Cell Transcriptomics for Discovering Novel Transcripts

RNA-Seq has allowed for the unbiased discovery of functionally important, low abundance transcript variants that would have been missed using conventional approaches. For example, a broad class of cytoplasmic intron-retaining transcripts (CIRTs) has been described in the dendrites of primary rat neurons (Buckley et al., 2011). Sequencing libraries constructed from the mRNA of mechanically isolated dendrites revealed not only coding region sequences for dendritically localized transcripts but also a subset of intronic regions located across the genomic organization of their respective genes. Although feasible, using microarray or PCR techniques to screen for these retained introns represents a significant challenge because it requires a priori knowledge of sequences that may be retained and decisions regarding which intronic sequences to target.

Further, these retained intron sequences have been demonstrated as functionally relevant for normal cellular function in neurons. Introducing small interfering RNA (siRNA) that targets a retained intron in *KCNMA1* leads to alterations in the protein distribution of the channel as well as changes to the intrinsic excitability of cells (Bell et al., 2008). Additionally, intron definition (ID) element sequences harbored within retained introns of the *CAMK2B* and *FMR1* transcripts are capable of competing for endogenous targeting machinery for those transcripts and impacting both RNA and protein distribution throughout the cell (Buckley et al., 2011). These results have linked relatively rare transcripts directly to observable endogenous functions. The identification of these sequences would not have been possible without current single-cell techniques like RNA-Seq.

Conclusions

Single-cell biology has undergone dramatic developments over the past decade. Performing single-cell transcriptomics from live cells in complex tissues is still difficult. Nevertheless, the development of novel methods that can isolate RNA from single cells with little resultant tissue damage promises to yield new insights into gene regulation of individual cells and how single cells in multicellular organisms work in concert. As more quantitative single-cell methods are being developed for sampling other "omes" (e.g., the proteome or metabolome), such large-scale data can be correlated to elucidate the link between gene expression and a cell's functional properties. It is through such correlations at the level of the single cell that the complexities of gene-product interactions will be identified. The goal of such research is to rationally modify these biological processes to produce predicted outcomes, including disease therapies.

References

Bell TJ, Miyashiro KY, Sul JY, McCullough R, Buckley PT, Jochems J, Meaney DF, Haydon P, Cantor C, Parsons TD, Eberwine J (2008) Cytoplasmic BK(Ca) channel intron-containing mRNAs contribute to the intrinsic excitability of hippocampal neurons. Proc Natl Acad Sci USA 105:1901–1906.

Buckley PT, Lee MT, Sul JY, Miyashiro KY, Bell TJ, Fisher SA, Kim J, Eberwine J (2011) Cytoplasmic intron sequence-retaining transcripts can be dendritically targeted via ID element retrotransposons. Neuron 69:877–884.

Espina V, Wulfkuhle JD, Calvert VS, VanMeter A, Zhou W, Coukos G, Geho DH, Petricoin EF 3rd, Liotta LA (2006) Laser-capture microdissection. Nat Protoc 1:586–603.

Kim TK, Sul JY, Peternko NB, Lee JH, Lee M, Patel VV, Kim J, Eberwine JH (2011) Transcriptome transfer provides a model for understanding the phenotype of cardiomyocytes. Proc Natl Acad Sci USA 108:11918–11923.

Lockhart DJ, Winzeler EA (2000) Genomics, gene expression and DNA arrays. Nature 405:827–836.

Martina M, Schultz JH, Ehmke H, Monyer H, Jonas P (1998) Functional and molecular differences between voltage-gated K^+ channels of fast-spiking interneurons and pyramidal neurons of rat hippocampus. J Neurosci 18:8111–8125.

Morris J, Singh JM, Eberwine JH (2011) Transcriptome analysis of single cells. J Vis Exp 50:2634.

NOTES

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

Okaty BW, Sugino K, Nelson SB (2011) A quantitative comparison of cell-type-specific microarray gene expression profiling methods in the mouse brain. PLoS One 6:e16493.

Surmeier DJ, Song WJ, Yan Z (1996) Coordinated expression of dopamine receptors in neostriatal medium spiny neurons. J Neurosci 16:6579–6591.

Tang F, Barbacioru C, Wang Y, Nordman E, Lee C, Xu N, Wang X, Bodeau J, Tuch BB, Siddiqui A, Lao K, Surani MA (2009) mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods 6:377–382.

Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: A revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63.