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