Chapter 6

Genomic Analyses of Neural Stem Cells

Nasir Malik*, Soojung Shin† and Mahendra S. Rao*,**
*National Institutes of Health, National Institute of Arthritis and Musculoskeletal and Skin Diseases, Bethesda, MD; †Invitrogen Corporation, Carlsbad, CA; **National Institutes of Health Center for Regenerative Medicine, Bethesda, MD

GLOSSARY

Neural stem cells  Self-renewing, multipotent cells that can differentiate into neurons, oligodendrocytes and astrocytes.
Lineage restricted precursor cells Cells that are still dividing but have a more restricted developmental potential than stem cells.
Large scale genomic analysis Methods or techniques to analyze a significant fraction of the information coded in RNA or DNA present in cells.
Data mining The use of statistical and graphical tools to compare expression patterns between data sets.
Reference standard A reference sample that allows comparison of data across laboratories and different platforms.

SUMMARY

- Neural stem cells (NSC) can be derived from fetal neuroepithelial cells or pluripotent stem cells and differentiated into all central nervous system neuronal and glial lineages.
- Technologies that can be used to understand NSCs at the genomic and transcriptomic level include microarrays, RNA-seq, DNA-seq, and Chip-seq.
- Because each of these technologies has its own advantages and limitations, researchers must carefully determine which one best suit their needs.
- The amount of data generated is vast, and planning must be performed in advance to ensure that proper analyses are performed in a timely fashion.
- Information gathered from large scale genomic and transcriptomic analyses can provide valuable insights into the biology underlying neural function and how it can be altered in neurological disorders.
6.1 INTRODUCTION

The nervous system is one of the earliest organ systems to differentiate from the blastula stage embryo. Neural stem cells (NSCs) are derived from neuroepithelial cells in the neural tube and give rise to all neuronal and glial cells of the central nervous system (CNS) by symmetric and asymmetric divisions. This process can be mimicked in culture and NSCs can be derived from human pluripotent stem cell (PSC) cultures over a period of 2–3 weeks (Reubinoff et al., 2001; Zhang et al., 2001; Shin et al., 2006; Elonbet al et al 2008; Koch et al., 2009; Reinhardt et al., 2013). PSCs are the \textit{in vitro} counterpart of the inner cell mass of the blastula stage embryo, and have the potential to generate any cell in the body. \textit{In vivo}, the primitive neural tube forms by approximately the fourth week of gestation, and neurogenesis has commenced by the fifth week of development in humans (Kennea et al., 2002). At the time of neurulation, at about the fourth week of development, the neuroectoderm segregates from the ectoderm by a process called neural induction (see Chapter 11). The early neural plate then undergoes a stereotypical set of morphogenetic movements to form a hollow tube, which is comprised primarily of stem cells, by a process called primary neurulation (Rao, 1999). The neural crest, which will form the peripheral nervous system, segregates from the CNS at this stage (see Chapter 18). Stem cells that will generate the CNS reside in the ventricular zone (VZ) throughout the rostrocaudal axis, they appear to be regionally specified by the expression of different positional markers, and proliferate at different rates. The anterior neural tube undergoes a dramatic expansion and can be delineated into three primary vesicles: the forebrain (prosencephalon), midbrain (mesencephalon), and hindbrain (rhombencephalon). Differential growth and further segregation lead to the additional delineation of the prosencephalon into the telencephalon and the diencephalon, and of the rhombencephalon into the metencephalon and the myelencephalon. The caudal neural tube does not undergo a similar expansion, but it does increase in size in a similar proportion to the embryo, and it undergoes further differentiation to form the spinal cord.

The properties of VZ stem cells have been characterized and they appear to be homogeneous, despite the acquisition of rostrocaudal and dorsoventral identity.

As development proceeds, the VZ becomes much reduced in size, and additional zones of mitotically active precursors can be identified. Mitotically active cells that accumulate adjacent to the VZ have been called subventricular zone (SVZ) cells. The VZ is later called the subependymal zone, as the VZ is reduced to a single layer of ependymal cells. The SVZ is prominent in the forebrain, and it can be identified as far back as the fourth ventricle. No SVZ is detectable in the more caudal regions of the brain, and, if it exists, it is likely a very small population of cells. An additional germinal matrix that is derived from the rhombic lip of the fourth ventricle, called the external granule layer, generates the granule cells of the cerebellum. Like the VZ, the SVZ can be divided into subdomains that express different rostrocaudal markers and that generate phenotypically distinct progeny. Discrete SVZ domains identified include the cortical SVZ, the medial ganglionic eminence, the lateral ganglionic eminence, and the caudal ganglionic eminence. The proportion of SVZ stem cells declines with development, and, in the adult, multipotent stem cells are likely present only in regions of ongoing neurogenesis (e.g., the anterior SVZ underlying the hippocampus). At this stage, marker expression is relatively heterogeneous (Bernier et al., 2000; Doetsch et al., 1996; Pevny and Rao, 2003).

Stem cells do not generate differentiated progeny directly, but rather generate the dividing populations of more restricted precursors that are analogous to the blast cells or restricted progenitors described in the hematopoietic lineages (Bedi et al., 1995; Katsuara et al., 2001; Mujitaba et al., 1999). These precursors can divide and self-renew, but they are located in regions that are distinct from stem cell populations and they can be distinguished from the stem cell population by the expression of cell surface and cytoplasmic markers (Cai et al., 2004b; Kalyani et al., 1997; Liu et al., 2004). Investigators have begun to analyze NSC populations (Table 6.1) using a variety of techniques, with the idea that by understanding these populations and identifying the factors that regulate self-renewal and direct differentiation, one will be able to modulate the development and response of stem cells to environmental signals. Many of these approaches depend on large scale analytic tools that rely on comparisons of purified populations of cells that differ with regard to their stage of development or their exposure to factors or carry specific genetic abnormalities. In this chapter, we focus on general principles that should guide such an analysis, and how data mining efforts have provided important insights into the properties of NSCs.

6.2 THE IMPORTANCE OF GLOBAL ANALYSIS AND CAVEATS WHEN COMPARING CELL SAMPLES

The overall disposition of a cell depends on the steady state of a complex of interacting factors. The integration of these instructions occurs in the nucleus through combinations of signal-activated and tissue-restricted transcription factors binding to and controlling related enhancers or \textit{cis}-regulatory modules of co-expressed genes. Additional regulation is provided by previously unappreciated epigenetic mechanisms, such as histone modulation, CpG island methylation, and by
microRNAs. Thus, the response of the cell to any one perturbation is dependent on context, and this in part explains conflicting results reported in the literature. For example, the effect of Sonic hedgehog on NSCs is dependent on the presence or absence of fibroblast growth factor (Wechsler-Reya et al., 1999). Likewise, the response to bone morphogenetic protein depends on the density of the culture and the presence or absence of various regulatory genes (Rajan et al., 2003; Wilson et al., 2001). This context-dependent response suggests that the overall state of a cell needs to be understood before perturbation experiments are initiated so that consistent and meaningful analyses of the results can be obtained.

Several variables remain poorly understood for NSCs. For example, no distinction has been made between long-term and short-term self-renewing populations. Although there is evidence that stem cell age has an effect (Shen et al., 1998; Svendsen, 2000), no analysis so far has taken into account the effects of aging, acquisition of karyotypic abnormalities, the differences as a result of the acquisition of positional identity, or the differences between the types of NSCs that are present during development. For example, radial glia type stem cells, transdifferentiated stem cell populations, VZ stem cells, SVZ-derived stem cells, and neurosphere-forming stem cells fulfill the criteria of NSC, such as self-renewal and the ability to differentiate into neurons and glia, but the comparisons between them have not been rigorous.

Two other observations have suggested that caution needs to be exercised when stem cell populations are analyzed. Stem cells propagated in culture stochastically differentiate, and, as such, they are invariably contaminated by various amounts of differentiated cells. For example, the proportion of stem cells in a neurosphere culture can vary greatly. Notably, the largest contaminating populations are astrocytes and astrocyte precursors which like NSCs divide and express Nestin. A confounding point is that these cells are difficult to distinguish from stem cells using the standard battery of tests. A second important observation that has been made is that there are species differences between stem cells, and, thus, extrapolating from mouse to human can lead to errors (Barker et al., 2003; Ginis et al., 2004). Each of these differences will add variability to the results and make cross-laboratory comparisons difficult unless attention is paid to the quality of the sample and detailed information is available regarding time of isolation, passage number in culture, and number of contaminating cells present. These differences need to be documented and accounted for when comparing data sets, because noise from such variability can skew the final analysis. In addition, when analyzing cells it is desirable to use a reliable and reproducible method that is cost-effective and sensitive enough to detect with high fidelity the global differences among the populations of cells as well as the subtle differences introduced as the cells are propagated in the culture or as they mature. Although several different methodologies have been proposed, none of the cross platform comparisons is very useful unless normalizing algorithms and considerations of technical variables inherent in large scale analyses are carefully considered (Table 6.2).

Nevertheless, as improvements have been made in our ability to obtain pure populations of cells, harvest RNA from single cells or small amounts of tissues, construct libraries, sort cells, and obtain high-quality genomic information, such large scale analyses are now used more commonly. However, it is recommended that analyses and comparisons be limited to one stage of development, in one species, with a single platform. The samples should be carefully examined for the presence of contaminating populations and the degree of contamination should be assessed (Figure 6.1). This initial quality control will be critical in yielding useful results.

### Table 6.1

<table>
<thead>
<tr>
<th>Cells Characterized</th>
<th>Methods Used</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neural stem cells and progenitor cells</td>
<td>Microarray</td>
<td>Luo et al., 2002</td>
</tr>
<tr>
<td>Neuroepithelial cells</td>
<td>Subtractive suppression hybridization</td>
<td>Cai et al., 2004b</td>
</tr>
<tr>
<td>Astrocyte-restricted precursors</td>
<td>Immunohistochemistry</td>
<td>Liu et al., 2004</td>
</tr>
<tr>
<td>Neural stem cells</td>
<td>Microarray</td>
<td>Wright et al., 2003</td>
</tr>
<tr>
<td>SSEA1-positive cells</td>
<td>Microarray</td>
<td>Abramova et al., 2005</td>
</tr>
<tr>
<td>Central nervous system progenitors</td>
<td>Microarray</td>
<td>Geschwind et al., 2001</td>
</tr>
<tr>
<td>Neurosphere-forming cells</td>
<td>Massively parallel sequencing</td>
<td>Cai et al., 2006</td>
</tr>
<tr>
<td>Fetal and ES-derived NSCs</td>
<td>Microarray</td>
<td>Shin et al., 2007</td>
</tr>
<tr>
<td>Spinal cord and brain derived neural stem cells</td>
<td>Microarray</td>
<td>Kelly et al., 2009</td>
</tr>
<tr>
<td>Embryonic stem cell, neural stem cells</td>
<td>RNA-seq</td>
<td>Sun et al., 2011</td>
</tr>
<tr>
<td>Adult neural stem cells</td>
<td>RNA-seq, Chip-Seq</td>
<td>Ramos et al., 2013</td>
</tr>
</tbody>
</table>
THE USE OF A REFERENCE STANDARD

When analyzing large batches of data generated in different laboratories using different techniques or slightly different cell culture protocols, one must determine the best way to make valid comparisons. Several strategies have been proposed including the idea of a reference standard (Dybkaer et al., 2004; Novoradovskaya et al., 2004; Loven et al., 2012). This idea, while not new, appears to be underappreciated in the stem cell field. Most researchers have however found it all but impossible to mine across data sets as there are too many variables that need to be normalized and too many assumptions that need to be made. Furthermore, there are often circumstances in which one simply lacks the data to make any appropriate assumption. In the ESC field, several strategies have been proposed (Loring et al., 2006). These include establishing a publicly available, well curated, data set that can be used as a ready reference, a set of standards that are readily available from a commercial or a not-for-profit provider or a control sample that all investigators can use as a standard. In principle each of these could be applied to the NSC field but to our knowledge no such common database exists as yet.

Immortalized or cancer stem cell lines such as C17.2, RT-4 or more recently identified cancer stem cell lines harvested from the appropriate species of interest have been proposed as possible standards (Imada et al., 1978; Snyder et al., 1992; Steindler, 2002). It is important however, when using such lines as a reference, to carefully assess the subclone that is being used. C17 subclones, for example, have shown remarkable variability and diametrically opposite results have been reported depending on the subclone used. The karyotype of this line is unstable, which may account for some of the differences seen. Nevertheless, since it has been so widely used, it could serve as a reference provided sufficient care was taken to use the same passage sample; banked at ATCC or some other responsible cell banking facility.

Fetal tissue samples from which pure populations of stem cells can be harvested at a defined stage in development in rodents may be an alternative choice for a reference standard. Many commercial entities provide such samples and these could therefore become a de facto standard. The equivalent stages of development are not readily accessible in humans however, and as such, an alternative control will need to be considered. Sorting or negative selection strategies have been

---

**TABLE 6.2 Comparison of Methods for Large Scale Genomic Analysis**

<table>
<thead>
<tr>
<th></th>
<th>RNA-Seq</th>
<th>Chip-Seq</th>
<th>Microarray</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample required</td>
<td>150 ng – 2 μg</td>
<td>2–10 ng</td>
<td>500 ng – 1 μg</td>
</tr>
<tr>
<td>Data presentation</td>
<td>Reads per kb of exon model per million mapped reads</td>
<td>Peaks aligned to reference genome</td>
<td>Fluorescence hybridization intensity</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>Very high (can increase or decrease by varying read number)</td>
<td>Very high (can increase or decrease by varying read number)</td>
<td>Moderate</td>
</tr>
<tr>
<td>Technical biases</td>
<td>Sequencing errors</td>
<td>Antibody quality</td>
<td>Noise at low signal intensity</td>
</tr>
<tr>
<td></td>
<td>Correctly mapping reads</td>
<td>Sequencing errors</td>
<td>Saturation at high signal intensity</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Correctly mapping reads</td>
<td></td>
</tr>
<tr>
<td>Cost</td>
<td>High</td>
<td>High</td>
<td>Low</td>
</tr>
<tr>
<td>Turn-around time</td>
<td>2–4 weeks</td>
<td>2–4 weeks</td>
<td>1–2 weeks</td>
</tr>
</tbody>
</table>

**FIGURE 6.1 Flow Chart of Techniques Used to Characterize Cell Populations Using Large Scale Analysis.** Samples for global analysis require verification in advance to assure dependable data production. Once the quality of the sample is controlled, one needs to decide upon a method of large scale analysis that is most appropriate for one’s purpose. Once the generated dataset is processed and analyzed, then an independent method is selected to confirm the acquired results.

6.3 THE USE OF A REFERENCE STANDARD

When analyzing large batches of data generated in different laboratories using different techniques or slightly different cell culture protocols, one must determine the best way to make valid comparisons. Several strategies have been proposed including the idea of a reference standard (Dybkaer et al., 2004; Novoradovskaya et al., 2004; Loven et al., 2012). This idea, while not new, appears to be underappreciated in the stem cell field. Most researchers have however found it all but impossible to mine across data sets as there are too many variables that need to be normalized and too many assumptions that need to be made. Furthermore, there are often circumstances in which one simply lacks the data to make any appropriate assumption. In the ESC field, several strategies have been proposed (Loring et al., 2006). These include establishing a publicly available, well curated, data set that can be used as a ready reference, a set of standards that are readily available from a commercial or a not-for-profit provider or a control sample that all investigators can use as a standard. In principle each of these could be applied to the NSC field but to our knowledge no such common database exists as yet.

Immortalized or cancer stem cell lines such as C17.2, RT-4 or more recently identified cancer stem cell lines harvested from the appropriate species of interest have been proposed as possible standards (Imada et al., 1978; Snyder et al., 1992; Steindler, 2002). It is important however, when using such lines as a reference, to carefully assess the subclone that is being used. C17 subclones, for example, have shown remarkable variability and diametrically opposite results have been reported depending on the subclone used. The karyotype of this line is unstable, which may account for some of the differences seen. Nevertheless, since it has been so widely used, it could serve as a reference provided sufficient care was taken to use the same passage sample; banked at ATCC or some other responsible cell banking facility.

Fetal tissue samples from which pure populations of stem cells can be harvested at a defined stage in development in rodents may be an alternative choice for a reference standard. Many commercial entities provide such samples and these could therefore become a de facto standard. The equivalent stages of development are not readily accessible in humans however, and as such, an alternative control will need to be considered. Sorting or negative selection strategies have been
shown to enrich for stem cell populations and markers that define the stem cell stage are available, thus a reference standard for human stem cell analysis could be considered. As the ability to derive NSCs from PSCs has improved, stable NSC lines that have long-term renewal potential are now available. Lines which have generated high-quality data that is reproducible at both the global expression scale and in regards to specific markers known to be present in NSCs can serve as excellent reference standards that can be used every time an array is hybridized. Although this takes away one lane from the array that could have been used for an uncharacterized sample, the ability of the reference to make it possible to compare across different arrays makes this a desirable option. Alternatively, a publicly available, well curated data set could be generated from expression data gathered from NSCs from various sources as a digital reference standard.

In the absence of any of the above, we recommend obtaining RNA, genomic DNA, and miRNA from NTERA2. NTERA2 is a pluripotent line derived from a human embryonal testicular carcinoma that readily differentiates into neurons and glia. It has been carefully analyzed by several groups and labeled and unlabeled clones are commercially available as well as being available from ATCC. The line is often used as a comparator for ESC work and thus significant data on several different platforms are already available. These cells however are not optimal when detailed high resolution comparisons are required (Schwartz et al., 2005).

In our laboratory, having an internal reference standard has proven invaluable in allowing us to compare NSCs to each other, and to samples run at different times and to compare our results with those of several other colleagues without the necessity of repeating all the experiments. These standards have also allowed us to check the quality of markers, our Fluorescence Activated Cell Sorting (FACS) sorting efficiencies, the quality of our antibodies and have provided a basis for comparisons across different laboratories. By running a reference sample in one laboratory and sending results to another laboratory, our experiments can be easily compared and over time cross platform comparisons also become possible (Figure 6.2).

Once the concept of a standard is accepted widely, commercial providers can provide RNA, DNA and genomic material from such a reference that can be used as a comparator for all types of large scale studies. In the ESC field and in the microarray field, groups have been established to determine such standards; such uniformity has yielded useful results (Brazma et al., 2001; Husser et al., 2006; Wei et al., 2005) and we fully expect that a similar effort in the NSC field will be equally useful.

Methods of Analysis

The past several years have seen dramatic advances in technology and equally dramatic reductions in costs. Both genomic and proteomic methods are now available at costs that allow an average small laboratory to begin performing such experiments. The amount of material required for such an analysis has also become much smaller than was previously necessary,
making these experimental approaches feasible even when the number of stem cells available is limited. For example, a gene expression profile using an Illumina chip costs about $150 per sample when calculating only the cost of the chip, and most institutions have core facilities that will perform all of the work associated with hybridizing the array at a reasonable cost. Even more importantly, the requirement for material has been dramatically reduced. We estimate that one can perform the entire battery of tests excluding proteomic analysis on about two million stem cells in any species, and newer technologies are bringing this figure down by another order of magnitude and also further reducing costs. Approximately ten million cells are enough for a mass spectrometry based analysis with equivalent amounts of material being required for Stable Isotopic Labeling using Amino acids in cell Culture (SILAC) and other similar methods. Indeed, it is expected that whole genome sequencing will soon cost less than US $1,000 and that profiling services will require even less material than is currently needed. These and other technical advances have facilitated the large scale gene expression analysis.

6.4 EPIGENETIC MODULATION

Over the past few years the importance of heritable epigenetic remodeling has been highlighted in regulating stem cell proliferation, cell fate determination, and carcinogenesis (Beaujean et al., 2004; Huntriss et al., 2004; Meehan, 2003; Ohgane et al., 2004; Vignon et al., 2002). Until, recently it has been difficult to study these events on a global scale; however, the ability to grow large numbers of cells and to differentiate them along specific pathways, coupled with the ability to perform such studies in a high throughput fashion has now made this possible. The concept of a bivalent chromatin state in PSCs, which facilitates expression of pluripotent genes but maintains lineage specific genes in a poised state ready to be activated when necessary, is now well established (Bernstein, 2006). Studies have begun to describe how this bivalent state is resolved during neurogenesis by derepression of poised genes required for neural commitment (Mikkelsen et al., 2007). Global methylation studies can be performed using a microarray or by whole genome bisulfite sequencing (Maitra et al., 2005). Illumina has described a bead array strategy to look at methylation patterns at 1500 loci encompassing regulatory elements in almost 400 genes. These include most genes known to be regulated during early embryonic development and those altered in tumorigenesis. These arrays have been used to examine methylation profiles in cancer stem cells and ESCs (Bibikova et al., 2006). Whole genome bisulfite sequencing of human ESCs differentiated to the three germ layers has been combined with Chip-seq and RNA-seq to understand the dynamic changes that take place at both the epigenetic and transcriptional levels (Gifford et al., 2013). The ENCODE consortium has also generated a vast dataset of genomic protein binding sites and epigenetic modifications in a host of cell types including human embryonic stem cells and neurons derived from them (ENCODE Project Consortium, 2012). These data are publicly available and can be viewed on the UCSC Genome Browser. Assessing the epigenetic profile of NSCs will be important prior to using them for transplant therapy, as maintenance of a particular epigenetic profile is probably critical for the appropriate function of cells.

6.5 MICRORNA

Micro (mi)RNAs are small non-coding RNA genes found in most eukaryotic genomes that are involved in the post-transcriptional regulation of gene expression. miRNAs are transcribed in the cell nucleus where they are processed into pre-miRNAs. Further processing occurs in the cytoplasm, where the pre-miRNAs are cleaved into their final ~22-nucleotide-long form which appears to regulate gene expression via transcriptional, translational or protein degradation regulation (Bartel, 2004; Szymanski et al., 2003). Recent reports have identified global strategies for identifying miRNAs, and over 450 such untranslated RNAs have been identified in humans, mice, and other species (Houbaviy et al., 2003; Lewis et al., 2003; Rajewsky et al., 2004). These approaches include computational analysis using sophisticated algorithms that recognize potential miRNA coding sequences and potential binding sites. Current estimates predict that there are ~2,000 miRNAs in the human genome. Other strategies have included making miRNA chips (Krichevsky et al., 2003). In addition, sequencing protocols analogous to massively parallel signature sequencing (MPSS) have been developed by Lynx Therapeutics that can be used to obtain quantitative data on miRNA made by a particular cell, which in turn predicts the overall state of the stem cell. Initial studies of the role of miRNAs in NSCs indicate that miR-9* and miR-124 repress the BAF53b and nBAF chromatin remodeling complexes in mice, resulting in the differentiation of NSCs to post-mitotic neurons in mice (Yoo et al., 2009). A recent paper reported that miR-17–92 and its paralogs are required for the proliferation of NSCs in mice, with knockout of these miRNAs resulting in transition of NSCs to more mature intermediate progenitors (Bian et al., 2013). Next generation sequencing (NGS) technologies such as RNA-seq, which can be used to identify all miRNAs in a sample, will be very useful for understanding the global role of miRNAs in NSCs.
6.6 MITOCHONDRIAL SEQUENCING

Structural and functional abnormalities in the mitochondria lead to functional defects in the nervous system, muscles and other organ systems. Somatic mitochondrial mutations are common in human cancers, aging cells and cells maintained in culture for prolonged periods. Mitochondrial DNA is also relevant to nuclear transfer, and estimating its stability is important in assessing the response of cells to stress and their ability to propagate in culture. Techniques to examine mitochondrial DNA mutations have been under development for some time. A recent description of a Polymerase Chain Reaction (PCR)-based approach for sequencing vertebrate mitochondrial genomes has attracted much attention, since it is faster and more economical than the traditional methods, which use cloned mtDNA and primer walking. Maitra and colleagues have developed a mitochondrial Custom RefSeq microarray as an array-based sequencing platform for the rapid and high-throughput analysis of mitochondrial DNA (Maitra et al., 2004). The MitoChip contains oligonucleotide probes synthesized using standard photolithography and solid-phase synthesis, and can sequence >29kb of double-stranded DNA in a single assay.

It is useful to note that many mutations arise in the D-loop regions, and a simple PCR amplification and sequencing process can capture a large amount of information. No published data on baseline mitochondrial sequence and changes in it after culture of NSCs are currently available, but this information is likely to yield valuable insights about the role of the mitochondrial genome in NSC aging.

6.7 TRANSCRIPTOME MAPPING

Efforts have begun to identify the complete DNA binding sites and corresponding genes targeted by the transcriptional factors. One approach is to use an in silico computational strategy to retrieve putative genes with such binding sites. A direct and physiological approach is to perform chromatin immunoprecipitation (ChIP) of factors crosslinked in vivo to DNA targets, followed by identification of the specific DNA binding sites. Promoter chips, ranging from a focused selection of genes to complete gene sets, are now being made. NGS technologies have now matured such that it is relatively straightforward to determine all the genomic binding of a DNA binding protein with Chip-Seq. The findings from such a project will be of immense biological value in providing a description and understanding of the hierarchal relationships between groups of transcription factors and their target genes as they perform their tasks in embryonic development and specification of lineage fates and terminal differentiation.

Nuclear Run-On Assays

One analogous approach to identifying regulatory elements is to perform labeling of newly processed RNA to examine the genes induced only after a specific stimulus. Such hybridizations, while requiring larger amounts of material, are feasible with cell lines and with ESCs and can provide a global overview of the network of the transcriptional responses to a specific stimulus. More importantly, they provide an element of temporal control allowing one to better place individual genes in a transcriptional network. While such arrays have not been run with NSCs, experiments in other systems have yielded exciting results (Li et al., 2006) and we expect similar results from NSCs in the near future.

Proteomic Analysis, Glycosylation Maps and Other Protein Mapping Strategies

Most of our discussion has focused on methods for assessing genomic differences between cells. However, post-translational modifications play a crucial role in modifying genomic information and increasing the complexity of information that can be processed by a cell. The very complexity of the proteome has made it difficult to study on a large scale. Recently, however, multiple technical breakthroughs have begun to allow large scale analyses. These include advances in sensitivity of mass spectrometry, SILAC, developing variations on 2D gels, and labeling techniques to identify key proteins that are altered under different conditions, as well as the development of methods for isolating and sequencing small quantities of proteins (Elliott et al., 2004; Freeze, 2003; Ong et al., 2003). Proteomic analyses of NSCs have not been reported as yet, despite the medium-term availability of these cell lines.

6.8 DATA MINING: CHROMOSOME MAPPING, PATHWAY ANALYSIS, DATA REPRESENTATION

A major problem with large scale analysis has been knowing how to interpret the data, how to compare it, and how to extract meaningful biologically relevant information. Biologists, as a rule, cannot simply look at long lists of genes to identify critical information and examining the most abundant gene may not be of biological significance. For instance, changes in
Notch or β-catenin signals of two-fold difference or less may be biologically significant, whereas ten or even one hundred fold differences in the expression of some genes may be irrelevant to the biological state of the cell. A particularly telling example of the relevancy of some differential gene expression data comes from the ESC literature, in which digital differential display identified TEX15 and DPPA5 as genes that are highly expressed in ESC but low or absent in all other populations examined (Adjaye et al., 2005; Kim et al., 2005; Lagarkova et al., 2006). These results were verified by PCR and immunocytochemistry, and yet data from knockout mice showed that these genes are dispensable for all assessed functions (Amano et al., 2006). On the other hand, a similar strategy identified Nanog, a previously unknown key regulator of ESC differentiation (Chambers et al., 2003; Mitsui et al., 2003).

In the stem cell field this has led to multiple attempts to consider how one should analyze data sets that are generated. The entire process is outlined in Figure 6.3. In our laboratory we have made the following assumptions: before pooling data or subjecting it to analysis, the quality of the sample used was tested. For example in NSCs, the harvested sample was assessed for its expression of known NSC markers and the absence of markers of differentiation. The presence or absence of these markers on the array is determined and the ability of the array to detect such differences is assessed by running the same sample on an array. If an array hybridization or MPSS analysis fails to detect an expected result, then that data are not used as all subsequent predictions are too uncertain. This analysis allows one to determine the expected sensitivity of the result and provides a rough idea of the sampling space (how much one will miss). We then examine carefully the intensity distribution of the expression levels of the genes present on the array. We have noted that in most cell types the distribution is quite similar and therefore any alteration suggests technical errors. While normalization algorithms can be employed in an attempt to use a particular anomalous data set, we generally red flag it, as most normalization algorithms tend to skew results.

Empirically we have determined that comparison across platforms is fraught with peril. Only positive results can be considered and negative results are generally not interpretable. For example, ESC-derived NSC samples were taken and examined by MPSS and Illumina bead arrays and a concordance of around 50% was shown, whereas the concordance rate for the sample run on a second Illumina array was close to 95% (unpublished observations). Even in such a comparison, only the presence or absence of an expression pattern can be considered and no attempts to compare expression levels

**FIGURE 6.3 Analysis Pipeline for a Microarray Experiment.** Once the hybridized array has been scanned the image files are run through an analysis file and intensities exported to Microsoft Excel worksheets. In these worksheets the data is interrogated for expression of known markers in the samples to ensure sample integrity. The data can then be analyzed at the global level for differentially expressed genes and also in a more focused manner to find important pathways that may be active. Once the analysis is complete interesting hits must be functionally validated.
between different methods should be made. We have also determined that amplification tends to provide a different pattern than using unamplified RNA even if the same sample is used. This appears not to be operator error, as amplification of different biological replicates performed at different times were closer than comparisons between amplified and unamplified samples. Therefore only samples that have been processed identically were compared as far as it is practical.

Once we are comfortable with the quality of each sample processed, we examine differential gene expression by examining pairwise comparisons rather than pooling the data or compare to a reference standard of baseline data that have been generated. This allows one to generate larger data sets and to compare across laboratories. For example, with NSCs we suggest using a well-characterized NSC sample as a potential reference standard. It is widely available, standardized, and has been run across multiple platforms and such data sets are publicly available. This method allows one to readily determine whether hybridization results are within the normal range, and whether the data are usable and comparable to results from other laboratories.

Once we have determined that we have a reasonable set of data, we then determine an appropriate cut-off for sensitivity that we are comfortable with. This ranges from array phenotype to phenotype and is an important criterion in any assessment. In MPSS for example, a theoretical sensitivity when three million tags are sequenced is three transcripts per million (tpm) (Miura et al., 2004a). However, we have empirically determined that testing or validating expression at such low levels is difficult and as such may or may not be useful to focus on. In our hands, an expression level of 50 tpm is readily verifiable and is a cut-off we use routinely (Cai et al., 2006); this does mean that we are potentially discarding useful information (which can be substantial as a majority of genes are expressed at low levels). Likewise, with Illumina bead arrays we use a cut-off of 50 to 100 arbitrary intensity units, even though the theoretical sensitivity of the arrays is much lower. It is however important that this be made clear in any publication, or alternatively the raw data be made available for independent analysis. This is particularly important as the genomic databases are constantly being curated and EST tags assigned to a particular locus are being reassigned as better data becomes available. This curation sometimes means that a gene tag on an array may not represent the gene that it was originally thought to identify. Various groups have estimated this frequency and have recognized it is an important source of error in these types of analysis. We therefore strongly recommend that such curation and updating be a regular part of the data analysis.

Once we have determined a cut-off, established a set of phenotypes for analysis, and collected the curated data we then can begin to consider how best to analyze the data. This analysis is dependent on the biological questions one wishes to pose and while each strategy will be different, we can perhaps highlight some of the simple strategies that can be used. Currently we feel that the most complete genomic data available are human and mouse. We therefore have focused on large scale analysis in these two species. We find that mapping the expressed genes onto a genome browser (chromosome mapping) provides a very useful overview of global patterns of gene expression. It allows one to study regulation of genes on a global scale, identify genomic hotspots or correlate with known chromosomal breakpoints or SNP data sets or data developed by groups working in different disciplines. We would recommend the UCSF (University of California San Francisco) golden path browser for this purpose.

A second important strategy that we use routinely is to perform what we term a pathway analysis. Here, rather than looking at the expression of an individual gene, we examine an entire signaling pathway by mapping the expression of all genes detected for that pathway. A visual representation of the “on” and “off” state of the pathway can be easily obtained, and observation of the pathway as cells differentiate can provide a much clearer idea of whether the genes in that pathway are important in the process of differentiation. Such a pathway analysis, for example, clearly identified the LIF/GP130 pathway as being important in NSCs in humans but not in mice. It also helps identify key genes that must be tested in verification studies. Multiple commercial programs to perform such a pathway analysis exist and we recommend using any one of these.

Overall, our experience has been that if effort is made to develop well curated data sets, verification of observations is generally greater than 50% and one can glean important and reliable information. While a hit rate of 50% may sound low, it is useful to consider that this is much better than the one in 30,000 chance of finding a functionally useful gene that one started with prior to analysis. Using NSC biomarkers as an example, currently there are about ten markers that have been identified. By a large scale analysis one could readily identify perhaps two hundred of which perhaps 25% will be novel or unexpected genes (50); with a hit rate of 50% one could identify 25 new markers in a three month experiment, more than doubling the number of known NSC markers. In the next section we will discuss some general observations made about NSCs.

### 6.9 General Observations About the Properties of Neural Stem Cells

We find that NSCs appear similar to other cells in synthesizing about 10,000–12,000 genes of an estimated 35,000 or 40,000 genes annotated in the RefSeq database. Average total RNA per cell tends to be higher in stem cells than in other cells (average of 5–10 pg) but is similar to levels seen in metabolically active cells. The distribution of transcript frequency...
suggested that most genes are transcribed at relatively low levels (less than 50 transcripts per cell) as assessed by MPSS. Mitochondrial, ribosomal and housekeeping genes tend to be more abundant, whereas transcription factors, growth factors and other cell type specific molecules are expressed at much lower levels. This pattern of gene expression is similar to that seen in most other cell types. Comparison of data with that for other cell types suggests that on average, the great majority of genes are shared or in common, while approximately 20% are different (by greater than ten-fold) in any two samples. Mapping of all expressed genes does not show a chromosomal bias as has been suggested in other stem cell populations. An example from an array experiment performed in our laboratory compared NSCs, astrocyte precursor cells (APCs), and oligodendrocyte precursor cells (OPCs) (Campanelli et al., 2008). The average correlation rate among NSC populations grown under different conditions was 0.90 (0.79–0.99) while the score dropped to 0.65 (0.57–0.71) and 0.77 (0.76–0.80) when NSCs were compared to APCs and OPCs, respectively. The correlation is much closer when identical samples are compared in two independent sequencing runs to 0.99 (0.96–0.99). In contrast, when identical samples are compared between two methods (Serial Analysis of Gene Expression (SAGE) and MPSS or Expressed Sequence Tag and MPSS) then the concordance rates are much lower (0.7), suggesting that those genes that are common between two methods are likely to be important. However, the lack of a high concordance when identical samples are analyzed by different methodologies suggests caution should be exercised in assuming that the failure to detect expression by any one method means that result is valid.

We have used the Illumina microarray platform to investigate the gene expression of normal and disease NSCs, neurons derived from these NSCs, astrocytes derived from a normal NSC line, and an immortalized astrocyte line. All the samples displayed similar intensity distributions, indicating that valid comparisons about gene expression could be made across this data set (Figure 6.4). As shown in Figure 6.5, at the global level, the normal NSCs and disease NSCs cluster together (r² = 0.95–0.96) and the neurons from these samples also group together (r² = 0.94–0.95). The NSC-derived astrocytes show greatest similarity to the NSC-derived neurons (r² = 0.91–0.92) while being more divergent from immortalized astrocytes (r² = 0.89), and with even less similarity to the parental NSC line from which they were derived (r² = 0.83). The data for the normal NSC lines are consistent with what we have seen in other array experiments as we generally see a correlation of around 0.95 for the various NSC lines in our laboratory, indicating that the lines differ very little in their gene expression profiles. It is not unexpected that the NSC-derived astrocytes and neurons are more similar to each other than the NSCs, as they primarily represent a population of differentiated cell types. It is perhaps surprising that the NSC-derived astrocytes are more similar to neurons than the immortalized NSC line, but this may be due to the possibility that the immortalization process has affected gene expression in the astrocytes. We also were able to identify a subset of ~20 genes that appeared to be unique to either NSCs, astrocytes, and neurons and via a pathway analysis find that Notch, TGF-β, and calcium signaling play important roles in these cell types (data not shown).

Examining NSC-enriched genes from previous studies suggests that several major pathways are active in NSCs (Abramova et al., 2005; Cai et al., 2006). These include the LIF/gp130 (in humans only), the FGF signaling pathway, the cell cycle regulatory pathways including myc and DNA repair and anti-apoptotic pathways. Intriguingly, the genes regulating the timing of differentiation, antisense RNA, siRNA, specific methylases and chromatin remodeling enzymes appear to be present at high levels, suggesting that epigenetic remodeling is an important aspect of NSC cell biology; as has also been shown for ESC cells.

6.10 SPECIES DIFFERENCES

An important finding that has been made clear from the availability of data sets from both mouse and human ESCs is the variability between species. While many key pathways are conserved, many differences have been highlighted as well. For example:

1) Oct3/4 homologs likely do not exist in chicken embryos (Soodeen-Karamath et al., 2001);
2) LIF signaling, which is critical for ESC self-renewal in rodents, does not appear to be critical or even required for human ESCs (Daheron et al., 2004; Ginis et al., 2004; Niwa, 2001);
3) No paralogs of E-Hox have been identified in humans and ES cell expressed Ras (Eras) appears to be a pseudogene in humans (Bhattacharya et al., 2004).

The low overall concordance rate between human and rodent ESCs (in one comparison around 40%) relative to that seen in human-to-human cell comparisons (90% between human ESC samples) provides additional support for these findings (Wei et al., personal communication). Although no analogous studies have been performed comparing mouse and human NSCs we expect similar species differences for this cell type.
FIGURE 6.4 Intensity Distributions for NSC, Neuronal, and Astrocyte Samples. NSCs, NSC-derived neurons and astrocytes and an immortalized astrocyte line were assessed for intensity distribution across the entire range of values at an intensity above 50 and the intensity distributions were graphed. All samples had similar distributions both within and across subsets.

FIGURE 6.5 Dendrogram Displaying the Relationship Between NSCs, Neurons, and Astrocytes at the Global Level. The Gene Expression module of Illumina’s Genome Studio software was used to generate a dendrogram to show the similarity of gene expression at the global level between several NSC lines, neurons and astrocytes derived from these lines and an immortalized astrocyte line.
6.11 LACK OF A “STEMNESS” PHENOTYPE

There is an operating bias in the literature that most stem cells will be similar to each other and indeed several manuscripts simply use the term “stem cell” rather than specifying which tissue they arise from and the stage of development used. This bias, however, is not supported by the large scale analytical data. Comparison of NSCs with other populations of stem cells has not identified any common stemness pathway (Cai et al., 2004a). Another important finding that has become obvious from such data set examination is the lack of any common stemness genes between ESCs and other somatic stem cells. Comparing data sets with expression in NSCs does not identify a common subset of genes and suggests that NSCs are quite different from other somatic stem cells (Fortunel et al., 2003; Ivanova et al., 2002; Ramalho-Santos et al., 2002). Our recent findings from a comparison of ESC-derived NSCs and fetal-derived NSCs (Shin et al., 2007) and more limited comparisons (Pevny and Rao, 2003) have shown clearly that stem cells can be readily distinguished from each other, and that stem cells or progenitor cells harvested at different stages of development behave differently.

6.12 ALLELIC VARIABILITY

It is clear that while individual stem cells share properties such as the ability to self-renew in culture and to differentiate into various phenotypes, there will nevertheless be differences between individual cell lines based on the genetic profile of the individual from whom they were derived. These allelic differences are not unexpected and merely reflect the diversity of phenotypes seen in the human population. Perhaps what has not been appreciated as much is how much variability this might impose on stem cell behavior, even if care is taken to isolate from the same region at the same developmental time using methods as similar as possible. The consequences of such allelic variability have been described in allergic responses, the ability to smell different odors, to metabolize alcohol, to digest milk, or to respond to toxins (Sultatos et al., 2004; Usuku et al., 1992).

When global pairwise comparisons are made between multiple stem cell lines using the same platform, one sees a high reproducibility when the same sample is run repeatedly (correlation coefficients in the range of 0.98–0.99). However, when two samples isolated by the same group at around the same time are run, the differences are much larger (correlation coefficient 0.92–0.94) indicating that a significant number of genes are differentially expressed between two stem cell lines. This sort of difference is consistently seen in all stem cell populations and is generally less than the difference between a cell line and its differentiated progeny and is less than the difference between a cell line from one species compared to another (Ginis et al., 2004). Nevertheless, at the individual gene level such changes can be quite dramatic (Bhattacharya et al., 2005). Such allelic variability likely underlies the differences in propagation, self-renewal and ability to differentiate into specific phenotypes that have been reported.

Large scale analysis of the kind we have reported allows one to map the underlying basis of such variability and make predictions about the behavior of cell lines (Lo et al., 2003; Ross et al., 2000; Yan et al., 2002a; Yan et al., 2002b). For example Lo et al. showed reliable measurement of allele specific gene expression based on large scale analysis. Selected genes were not only in the imprinted regions but also were distributed throughout the genome (Lo et al., 2003), which may be responsible for phenotypic variation.

Thus, as one begins considering stem cells for therapy one will need to carefully assess whether a lot of NSC prepared from one individual will behave in similar way to another lot of cells prepared from a different donor, even if identical protocols were used and the experiments were performed in a similar fashion. This inherent variability is of major concern when evaluating cells as therapy.

6.13 AGE DEPENDENT CHANGES IN NSCS

Examination of human cells in culture has shed some light on the cellular basis of aging. When grown in culture, normal human cells will undergo a limited number of divisions before entering a state of replicative senescence, in which they remain viable but are unable to divide further (Miura et al., 2004b). This change has been attributed to changes in mitochondria, protein expression, loss of genomic integrity, oxidative damage and progressive loss of DNA repair ability or erosion of telomere ends. Results from our laboratory and others have shown that ESCs appear to be immortal and can be propagated in continuous culture for a period of at least two years. Examination of the expression of immortality-associated genes has suggested that the expression of key regulators is important (Miura et al., 2004b). These include the expression of telomere-associated proteins, expression of some immortality-associated genes, high levels of DNA repair enzymes, the inhibition of p53, and the absence of Rb; thus altering cell cycle regulation. Comparing the phenotype of such genes between NSCs and ESCs shows significant differences. NSCs have shown active expression of Rb and their pattern of
expression of telomere related genes and immortality-associated genes differs from that of ESCs. Differences are also apparent between adult stem cell populations and ESCs, which suggests that NSCs are not easily propagated indefinitely in culture. Indeed, Whittemore and colleagues have shown that long-term propagation of NSCs in culture decreased their ability to differentiate (Quinn et al., 1999). Pruitt and colleagues have shown that NSCs undergo karyotypic changes as the animal ages, providing further confirmation of this hypothesis (Bailey et al., 2004). Our recent results examining the karyotypic stability of NSCs in culture have indicated that NSCs are less stable in culture than ESCs and can acquire karyotypic changes in as few as ten passages. Overall, large scale analysis suggests that there are fundamental differences in cell cycle and immortality-associated genes and these differences predict that NSCs will age in culture and senesce as do other adult stem cell populations.

### 6.14 CANCER STEM CELLS

The discovery that many cancers may be propagated by a small number of stem cells present in the tumor mass is an exciting finding (Al-Hajj et al., 2003; Huntly et al., 2005; Reya et al., 2001; Singh et al., 2004). This was first described in breast cancers and subsequently in a variety of solid tumors. With respect to the nervous system, several reports have suggested that cancer stem cells can be identified and that these cells bear a remarkable similarity to NSCs present in early development (Hemmati et al., 2003; Singh et al., 2004). Likewise, cells resembling glial progenitors have been isolated from some glial tumors (Kondo et al., 2004; Noble et al., 1995). In the case of glioblastoma, a subset of tumor cells that resemble neural stem cells were found to be responsible for the long-term growth of the tumor in mice (Chen et al., 2012). We and others have suggested that tumors can perhaps be phenotypically classified, and this classification correlates well with a cell of origin identified based on findings from lineage relationships that exist between stem and progenitor cells. The advantage of having a detailed database of gene expression from multiple normal cell lines and progenitor cells and the ability to compare their gene expression profiles with their transformed counterparts is obvious. In recent years, numerous gene expression studies have found that although brain cancers can be grouped into molecular subclasses, the signaling networks driving these cancers are remarkably similar, and normally function during early development (Brennan et al., 2009). Results such as these validate our belief that expression profiling at the global level can provide valuable insights regarding pathways that may be important for the development of a particular cell type or progression to a disease state.

### 6.15 CONCLUSIONS

Overall, the initial efforts at profiling NSCs from rodents and human have yielded useful insights, and have allowed the identification of key regulatory pathways and novel genes that are likely to play a role in regulating NSC self-renewal and differentiation. The data sets developed to date are an important resource, and can be mined with readily available tools. Better accessibility and curation of these data for the research community would greatly facilitate more advances in our knowledge of NSC biology. While profiling information has already yielded fruit in terms of understanding the basic biology of the cells, we believe that the availability of this information will have a utility far beyond current applications, including important future roles in the clinical setting in assessing the value of therapeutic interventions relating to advances in stem cell transplantation and treatment of tumors affecting the brain.

### 6.16 CLINICAL RELEVANCE

- Neural stem cells (NSCs) may have therapeutic potential as cellular replacement therapies in neurodegenerative disorders.
- Many tumors of the nervous system are derived from a small number of NSC-like cells that undergo unregulated proliferation.
- As the cost of large scale next generation sequencing technologies decreases, it will become a common practice to sequence tumors from patients to identify the best therapeutic options.
- Likewise, if NSCs become part of a treatment strategy, the ability to sequence the genomes of these cells before transplantation will make it possible to determine whether they carry any mutations that could be dangerous if the cells are transplanted into a patient.
- The ability to generate NSCs and differentiate them on a large scale will facilitate drug screens and the pathways modulated by positive hits from these screen can be identified by the use of large scale genomic technologies to further drug development for neurological disorders.
ACKNOWLEDGMENT

This work was supported by the NIH Common Fund.

RECOMMENDED RESOURCES

Reviews

Websites
Illumina’s Next Generation Sequencing TecnoLOGY:
Invitrogen’s Next Generation Sequencing Technology:
NHGRI DNA Microarray Technology Overview:

REFERENCES


