Jeannette Reinartz
is Group leader Production, Lynx Therapeutics, GmbH, Germany.

Eddy Bruyns
is Group leader Functional Genomics, Lynx Therapeutics, GmbH, Germany.

Jing-Zhong Lin
is Senior Scientist, Lynx Therapeutics, Inc., Hayward, California, USA.

Tim Burcham
is Vice President, IT and Bioinformatics, Lynx Therapeutics, Inc., Hayward, California, USA.

Sydney Brenner
is Scientific Advisor for Lynx and Professor, Salk Institute, La Jolla, California, USA.

Ben Bowen
is Vice President, Discovery Research, Lynx Therapeutics, Inc., Hayward, California, USA.

Michael Kramer
is Managing Director, Lynx Therapeutics GmbH, Germany.

Rick Woychik
is Chief Scientific Officer, Lynx Therapeutics, Inc., Hayward, California, USA.

Keywords: *gene expression profiling, massively parallel signature sequencing, MPSS, Megaclone*

Rick Woychik,
Lynx Therapeutics, Inc., 25861
Industrial Blvd., Hayward, CA
94545, USA

Tel: +1 (510) 670 9488
Fax: +1 (510) 670 9303
E-mail: Rwoychik@lynxgen.com

Technique review

Massively parallel signature sequencing (MPSS) as a tool for in-depth quantitative gene expression profiling in all organisms

Jeannette Reinartz, Eddy Bruyns, Jing-Zhong Lin, Tim Burcham, Sydney Brenner, Ben Bowen, Michael Kramer and Rick Woychik

Date received (in revised form): 6th December 2001

Abstract

Massively parallel signature sequencing (MPSS) is one of the newest tools available for conducting in-depth expression profiling. MPSS is an open-ended platform that analyses the level of expression of virtually all genes in a sample by counting the number of individual mRNA molecules produced from each gene. There is no requirement that genes be identified and characterised prior to conducting an experiment. MPSS has a routine sensitivity at a level of a few molecules of mRNA per cell, and the datasets are in a digital format that simplifies the management and analysis of the data. Therefore, of the various microarray and non-microarray technologies currently available, MPSS provides many advantages for generating the type of complete datasets that will help to facilitate hypothesis-driven experiments in the era of digital biology.

INTRODUCTION

Recent developments in the sequencing of many vertebrate, invertebrate, plant and microbial genomes have prompted an increasing interest in using genomics reagents to study patterns of gene expression. Building large relational databases filled with content that includes in-depth gene expression profiles from multiple cell types will undoubtedly make an enormous contribution to any experimental effort in digital biology. Several DNA microarray platforms,[1–6] serial analyses of gene expression (SAGE),[7,8] cDNA sequencing and a variety of other technologies are available for analysing the expression of hundreds to thousands of genes simultaneously. Each of these existing technologies has limitations when it comes to generating complete datasets for building relational databases. In this paper, one of the newest tools for evaluating gene expression is reviewed, called massively parallel signature sequencing (MPSS),[9,10] which overcomes many of the limitations of the current technologies. MPSS is a novel microbead technology that is totally unlike other bead-based operations such as those developed by Illumina[11] or Luminex.[12]

MPSS AS A TOOL FOR QUANTITATIVE GENE EXPRESSION PROFILING

Unlike most microarray technologies that capture data that are analogue in nature, MPSS is one of the few technologies that produces data in a digital format. MPSS captures data by counting virtually all mRNA molecules in a tissue or cell sample. All genes are analysed simultaneously, and bioinformatics tools are used to sort out the number of mRNAs from each gene relative to the total number of molecules in the sample. At least one million molecules are typically

Use of MPSS for expression profiling involves counting mRNA molecules in the sample

counted in any given sample, so even genes that are expressed at low levels can be quantified with high accuracy. The fact that the data are of a digital nature provides an ideal format for building databases where the temporal/spatial expression profiles for all genes can be deduced by comparing results from multiple cell types electronically.

Counting mRNAs with MPSS is based on the ability to identify uniquely every mRNA in a sample. This is done by generating a 17-base sequence for each mRNA at a specific site upstream from its poly(A) tail (first *Dpn*II site in double-stranded (ds) cDNA, see below). The 17-base sequence is then used as an mRNA identification 'signature'. To measure the level of expression of any given gene, the total number of signatures for that gene's mRNA is counted.

CLONING AND SEQUENCING cDNA FRAGMENTS ON BEADS

MPSS signatures for mRNAs in a sample are generated by sequencing ds cDNA fragments cloned onto microbeads using the Lynx Megaclone technology (Figure 1). The Megaclone technology has

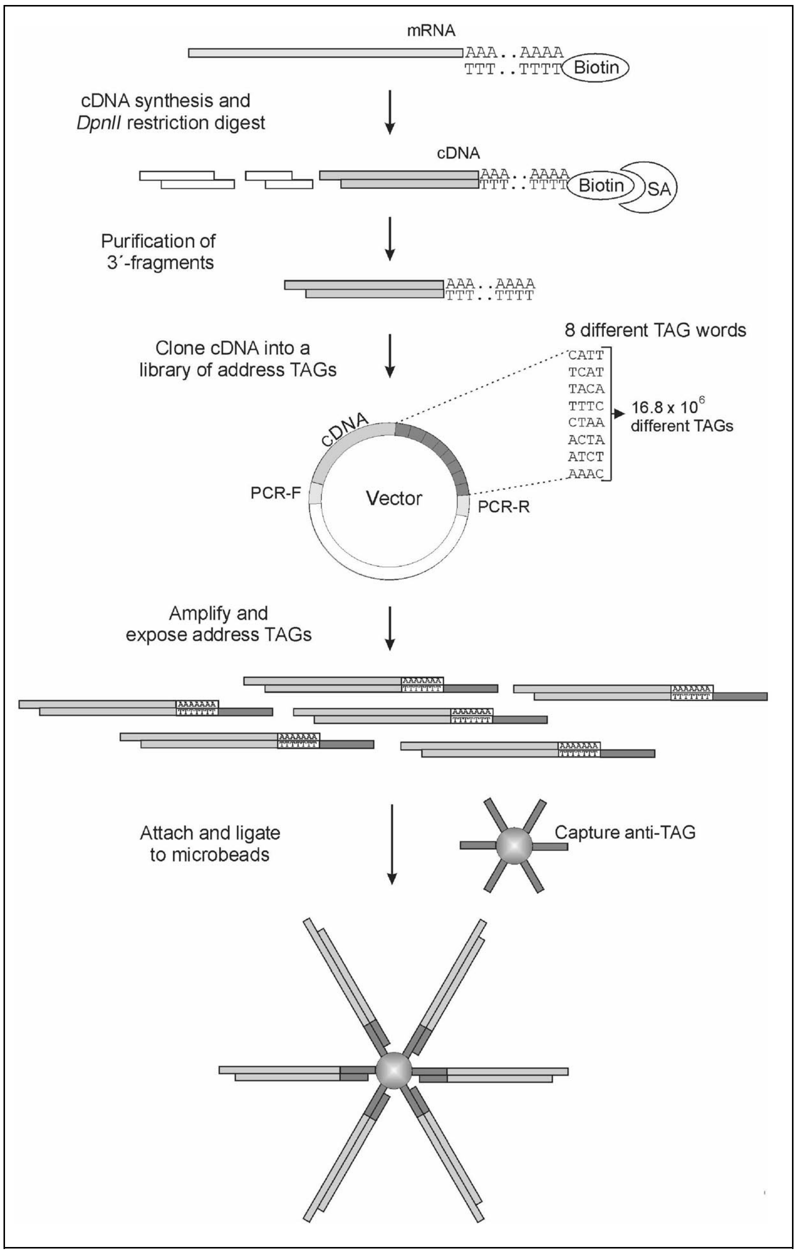

Figure 1: Summary of the Megaclone technology. Poly(A) mRNA molecules are converted into double-stranded cDNA molecules, which are digested with *Dpn*II and cloned into a specially designed plasmid vector containing a 32 base pair (bp) oligonucleotide tag. There are 16.8 × 10^6 million different 32-base sequences available in the reference tag library, and each cDNA clone contains a different sequence.[13] A library of cDNA inserts, along with their adjacent 32 bp oligonucleotide tags, are polymerase chain reaction (PCR) amplified, and the resulting linear molecules are partially treated with an exonuclease to make the 32-base tag single stranded. The 32-base tags at the end of each of the cDNA molecules are hybridised to 32-base complementary tags that are covalently linked to 5 μm microbeads. There are 16.8 million different complementary tags, each of which corresponds to one of the 16.8 million different 32-base tags;[13] therefore, for every tag on a cDNA molecule, there will be one bead with a complementary 32-base tag available for hybridisation. Each bead contains a vast excess of one particular 32-base complementary tag sequence. Once the cDNA molecules are hybridised to the beads, the nicks are sealed enzymatically. The end-product is a microbead with approximately 100,000 identical cDNA molecules covalently attached to the surface

Megaclone is the process of cloning DNA fragments onto microbeads

previously been described in detail.[13] Briefly, complementary DNA (cDNA) is prepared from poly(A) RNA using a biotin-labelled oligo-dT primer. The oligo-dT is designed to prime each mRNA molecule exactly at the poly(A) junction. The cDNA fragments are then digested with *Dpn*II (recognition sequence, GATC), and the 3′-most *Dpn*II-poly(A) fragments are purified utilising the biotin label at the end of each molecule. The fragments are subsequently cloned onto 5 μm diameter microbeads using a set of 32-base tag/anti-tags developed at Lynx (see Brenner *et al.*[13] for details). This process yields a library of beads where one starting mRNA molecule is represented by one microbead, and each microbead contains approximately 100,000 identical cDNA fragments from that mRNA. Therefore, starting with one million mRNA molecules from a particular cell or tissue sample, Megaclone will produce one million beads, each containing 100,000 cloned copies of cDNA from each mRNA molecule. All molecules are covalently attached to the microbeads at their poly(A) ends, so the *Dpn*II end is available for the sequencing reactions.

The sequencing process is initiated by ligation of an adaptor molecule and digestion with a type IIs restriction enzyme. Approximately one million microbeads are then loaded into a specially designed flow-cell in a way that allows them to stack together along channels and form a tightly packed monolayer in the flow-cell. The flow-cell is connected to a computer-controlled microfluidics network that delivers different reagents for the sequencing reactions. A high-resolution CCD camera is positioned directly over the flow-cell in order to capture fluorescent images from the microbeads at specific stages of the sequencing reactions. Details on the instrumentation used for MPSS is described in detail in Brenner *et al.*[9]

The actual DNA sequencing reaction involves an automated series of adaptor ligations and enzymatic steps and has previously been described in detail[9] (Figure 2). The process is initiated by ligating an adaptor molecule to the GATC (*Dpn*II) ss overhangs, and then digesting the samples with *Bbv*I, which is a type IIs restriction enzyme that cuts the DNA at a position nine to 13 nucleotides away from the recognition sequence. This produces molecules with a four base single-stranded (ss) overhang immediately adjacent to the *Dpn*II recognition

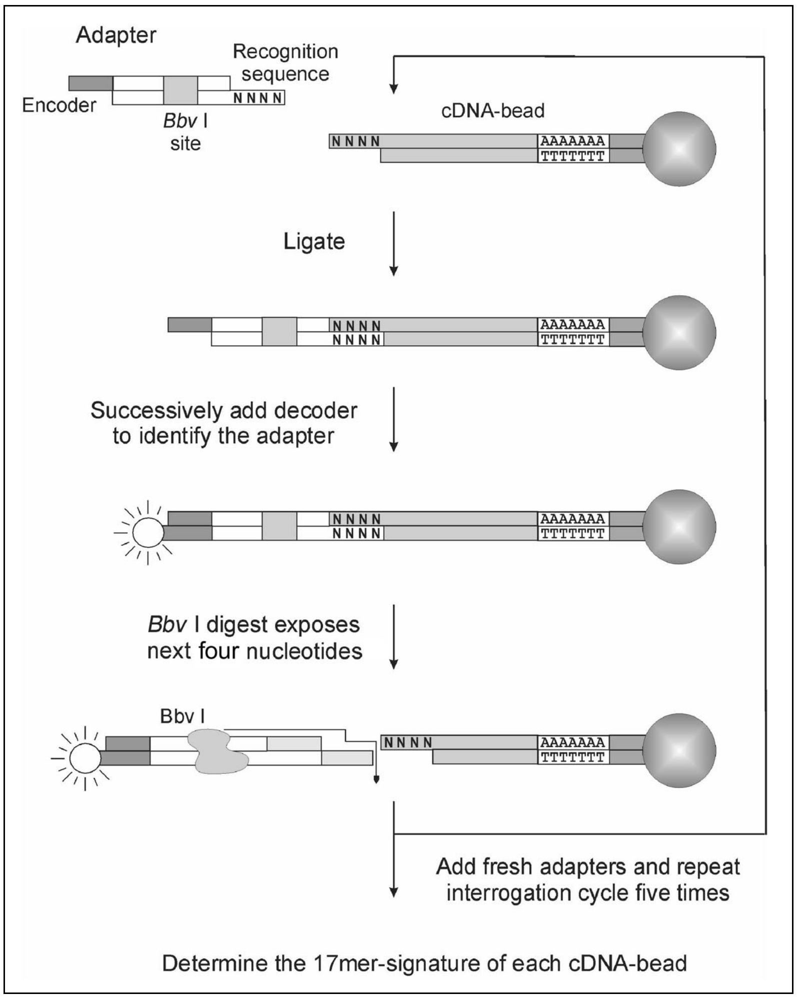
Adapter
Recognition sequence
NNNN
Encoder
Bbv I site
cDNA-bead
NNNN
AAAAAAA
TTTTTTT
Ligate
Successively add decoder to identify the adapter
Bbv I digest exposes next four nucleotides
Bbv I
Add fresh adapters and repeat interrogation cycle five times
Determine the 17mer-signature of each cDNA-bead

Figure 2: Summary of the sequencing reactions for MPSS. Encoded adaptors are ligated to the ends of the cDNA molecules attached to the microbeads. Sixteen different fluorescent-labelled decoder probes are then sequentially hybridised to the encoded adaptor ends in order to deduce the first four nucleotides at the end of each molecule. The encoded adaptor from the first round is then removed by digestion with *Bbv*I, which exposes the next four nucleotides as a four-base single-stranded overhang. The process is repeated several times in order to generate a total of 17 bases of sequence

sequence. Another set of adaptors, called encoded adaptors, are hybridised and ligated to the four base overhangs on each molecule. The encoded adaptors contain a four base ss overhang with all possible nucleotide combinations at one end, and a ss coded sequence at the other end. One member of the encoded adaptor set will find a partner on the DNA molecules attached to the beads in the flow-cell. The exact sequence of the four base ss overhang on each encoded adaptor that hybridises to the DNA on a microbead is decoded through a series of 16 different sequential hybridisation reactions with a set of fluorescent decoder probes. This process yields the first four nucleotides at the end of each molecule. To collect additional sequence, the encoded adaptor from the first round is removed by digestion with *Bbv*I, and the process is repeated several times. In the end, a 17-base signature sequence is generated for each bead in the flow-cell.

Sequencing with MPSS produced a 17-base signature sequence for each cDNA molecule

DATA HANDLING AND CALCULATION OF RNA ABUNDANCE

A typical MPSS experiment with about one million microbeads will yield 250,000–400,000 high quality 17-base signature sequences. Two or more flow-cells with microbeads initiated with either of two different adaptors (to address issues of palindromes, as discussed in Brenner *et al.*[9]) are used for each experiment. Additionally, MPSS datasets are additive in nature, which means that datasets from multiple analyses with the same starting mRNA sample can be combined. In many instances, this has been done at Lynx to produce datasets that involve in excess of one million mRNAs counted per sample (see below). This ability to combine datasets in an additive manner leads to an increased sensitivity for all genes being analysed, particularly those that are expressed at very low levels within the sample.

MPSS datasets are digital in nature where mRNA abundance is expressed as transcripts per million (TPM)

Each signature sequence in an MPSS data set is analysed, compared with all other signatures and all identical signatures are counted. The level of expression of any single gene is calculated by dividing the number of signatures from that gene by the total number of signatures for all mRNAs present in the dataset. The data for each gene are usually reported as the transcripts per million (TPM). Analysis of a complete MPSS dataset makes it possible to calculate the numbers of genes that are expressed at varying levels within the sample. For example, it is possible to calculate readily the genes that are expressed at greater than 1,000, 100–1,000, 10 to 100 and less than 10 TPM. From this type of analysis of many datasets over the past several years, it appears that most genes are expressed at a level of 1 to 100 TPM (data not shown).

For any specific library of cDNA fragments loaded onto beads, there is a high degree of reproducibility between MPSS runs. To illustrate this point, two MPSS datasets involving a total of 306,884 counted mRNAs were analysed (Figure 3A). A total of 10,799 different signatures was identified in the dataset, and the abundance of over 99 per cent of these signatures in the duplicate runs was not significantly different ($p < 0.001$). In other words, the level of expression for the vast majority of the genes was not statistically different in the two runs. Using independent libraries of cDNA fragments introduces what is likely to be some biological variability in the data, although the overall reproducibility of the process is still very high — usually being greater than 90 per cent similarity (T. Burcham, unpublished data).

MPSS signature sequences can be connected to known genes by comparison with data in the available genomic sequence and expressed sequence tag (EST) databases. This is usually an efficient process, although occasionally it is not possible to find a signature for a gene known to be expressed in a particular sample. This can happen when a gene does not contain a *Dpn*II site, or when there is a sequence polymorphism in the *Dpn*II site. These problems can be easily overcome by digesting the cDNA

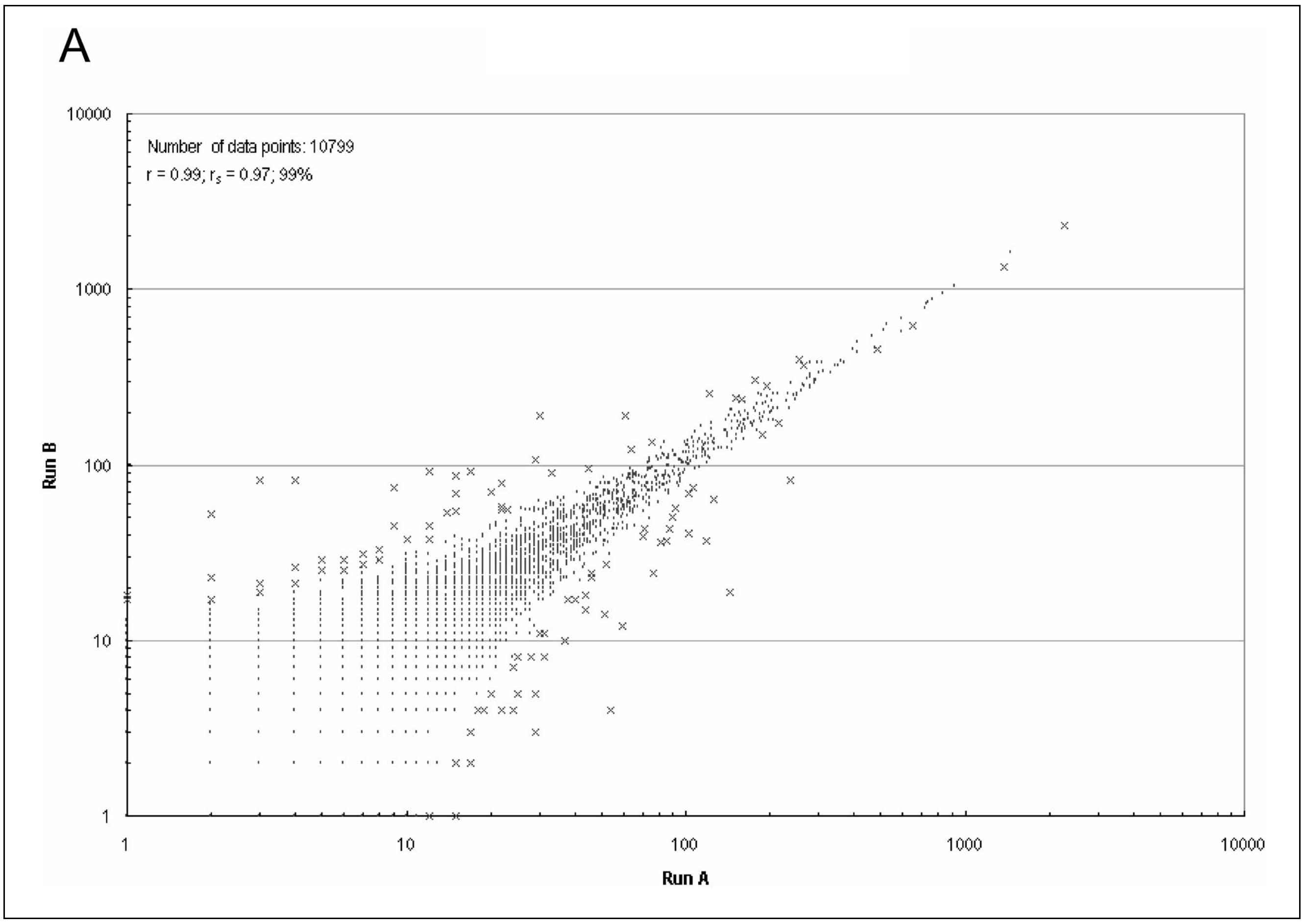

Figure 3: Reproducibility and statistical analysis of MPSS data. (A) Dot plot showing the reproducibility between MPSS runs. Each dot or × represents one of a total of 10,799 signatures that were generated from two independent MPSS runs from the same starting cDNA library loaded on to beads. The X and Y coordinates represent the number of each signature found in each of the two MPSS runs. Each signature represented by a dot occurs within a 0.99 confidence interval, while those represented with an × occur outside this interval. (B) Plot of the level of gene expression (MPSS signatures/million) versus the fold difference that can be detected between two independent samples. Each curve corresponds to a different p-value for the fold difference on the X axis that relates to a signature abundance on the Y axis

with an alternative enzyme. Incomplete sequence representation of a particular gene in the current EST and cDNA clone databases can also complicate the process of assigning a signature sequence to a gene. Not all cDNA clones that have been sequenced extend all the way to the poly(A) addition site, so the sequence that corresponds to an MPSS signature for a specific gene may not be represented in an EST sequence database. For example, the signature sequence for the T-cell transcription factor NFATc did not appear to be included at any significant level in the Lynx Human T-cell-related MPSS datasets using the RefSeq database entry NM_006162. When the RefSeq entry is extended towards the 3′ end with another independent cDNA clone in GenBank (eg with entries like U80917), the signature 'GATCCAATAAAGCCGTA' was identified, which was found in all of the appropriate MPSS T-cell datasets. Therefore, careful and thoughtful analysis of the available data may be necessary during the process of assigning an MPSS signature to a gene.

Connecting MPSS signatures with genes is usually a straightforward process

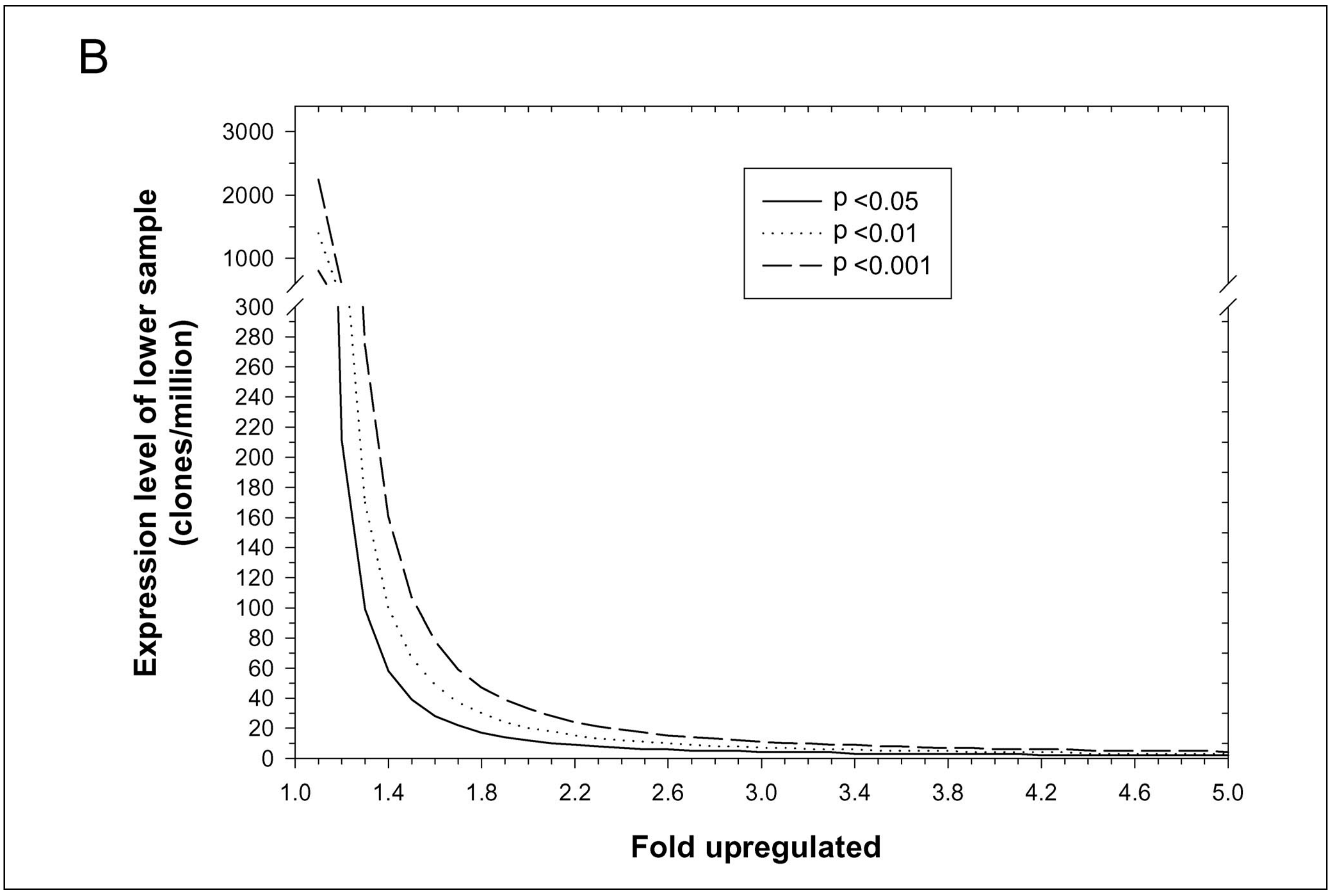

Figure 3: (*Continued*)

MPSS data can be treated as categorical data for statistical analyses

STATISTICAL ANALYSIS OF MPSS DATA

One of the most important attributes of MPSS data over those produced by other technologies like microarrays relates to the fact that MPSS data can be treated as 'categorical' from a statistical point of view. This makes it possible to capitalise on the large number of measurements of a given signature in the dataset (typically ten to 1,000 or more, depending on the gene) as well as the size of the entire dataset (typically over one million) to calculate whether the gene giving rise to this signature is differentially expressed in multiple different samples. This compares with data from microarrays, where only two or three measurements, generated from the number of replicate microarrays performed for any given experiment, would be factored into the calculation.

To test whether a gene is differentially expressed between two samples, a normal approximation test for difference in binomial proportions is used (also described as the Z-test, which has been independently employed for analysis of SAGE data sets[14]). If x_1 and x_2 represent the abundance of a specific signature in samples 1 and 2, respectively, and n_1 and n_2 represent the total number of signatures generated for all mRNAs in samples 1 and 2, the proportions $p_1 = x_1/n_1$ and $p_2 = x_2/n_2$ each follow a binomial distribution. Since n_1 and n_2 are large in MPSS (typically in the order of 10^6), the difference $(p_1 - p_2)$ follows an approximate normal distribution defined by formula (1)

$$N\left((p_1 - p_2), \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right) \qquad (1)$$

where the unknown parameters p and q

can be estimated as $\hat{p} = (x_1 + x_2)/(n_1 + n_2)$ and $\hat{q} = 1 - \hat{p}$, respectively.

The test statistic defined by equation (2)

$$\lambda = \frac{p_1 - p_2}{\sqrt{\hat{p}\hat{q}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \qquad (2)$$

MPSS datasets are ideal for building relational databases for systems biology applications

follows a standard normal distribution and can be used to test whether expression of the gene bearing the signature between the two samples could be due to chance alone. Standard statistical tables can be employed to determine the p-value of this influence based on values of λ.

Theoretical calculations based on equation (2) show an inverse relationship between the level of expression and size of the difference that can be evaluated between samples (Figure 3B). For example, for $p < 0.001$, it is possible to detect a two-fold change for a gene that is expressed at a level of 30–40 copies per million. For genes that are expressed at a higher abundance, it is possible to detect a substantially smaller difference. A 40 per cent difference can be ascertained for genes that are expressed at about 200 copies per million. These characteristics are in contrast to the analyses for gene expression data generated by hybridisation-based methods, such as microarrays, where a significance test is possible only if the experiment is replicated several times,[15] and where differential expression can usually be detected only for genes with relatively high levels of expression[1] and with a large difference between samples.[3]

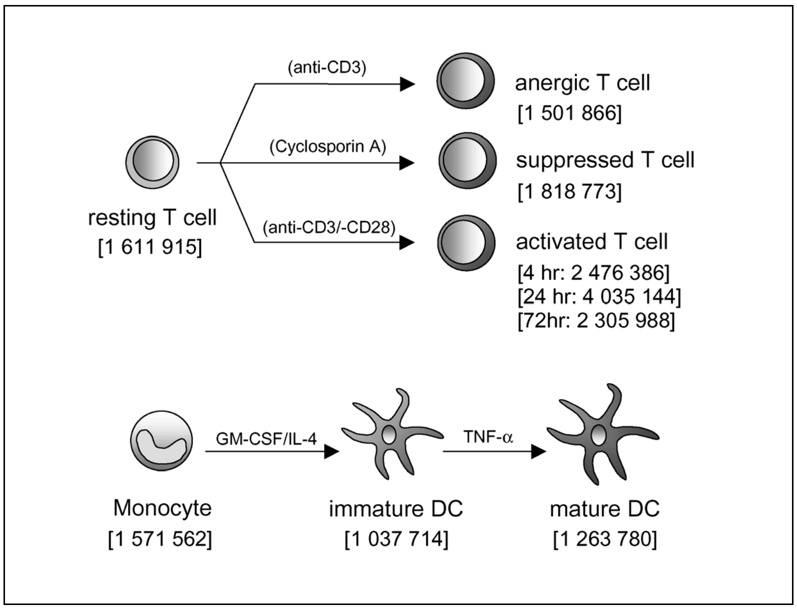

Figure 4: The immunology focus programme at Lynx in Heidelberg, Germany is dedicated to the analysis of important immunological and haematopoietic cell populations using MPSS. Several immunologically important cell populations were analysed, including resting, anergic, immunosuppressed and activated T cells. Additionally, monocytes, immature dendritic cells and mature dendritic cells (DC) were prepared. All cell types were generated from blood-derived monocytes or resting T cells by *in vitro* treatments summarised in the figure. Given in squared brackets is the total number of MPSS signature sequences that were generated for each cell type. The MPSS datasets are formatted into a relational database, termed GeneCat

MPSS DATASETS AND DEVELOPMENT OF RELATIONAL DATABASES

Many of the experiments that have used gene expression profiling tools were designed to simply compare one sample with another. This often produces a list of differentially expressed genes which, by itself, is less than complete for extracting biologically useful information. Given the right experimental design, in-depth expression profiling has the potential to revolutionise experiments in systems biology where the object is to study and ultimately understand complex biological processes at the molecular level. For example, investigators at Lynx GmbH (Heidelberg, Germany) have collected gene expression data by MPSS from many different purified cell samples from the human immune system (Figure 4). This substantial dataset forms truly relational database content, where meaningful comparisons and biological questions and hypotheses can be addressed at the keyboard.

The key to building relational databases for useful experiments in systems biology is to ensure that all genes in a sample are represented within a dataset, and that the level of expression for each gene is digitised in a manner that truly reflects its relative expression in the sample. This is not an issue with MPSS, since virtually all genes are represented in a dataset and the level of expression of each gene is represented by a number that reflects the number of transcripts per million in the sample.

In contrast, there are many noteworthy issues relating to the use of microarray data for building expression databases. First of all, and probably most critical, is the fact that not all genes are represented on any given microarray. The Affymetrix human U95 set contains elements to study the expression of more than 60,000 genes and ESTs. It is not clear whether these sets contain elements for all human genes. Also, high-density microarrays are not available for many organisms that are of interest to the biological community, which limits the use of microarrays for experiments in many non-human systems. Secondly, strong claims have been made about the detection sensitivity of some arrays, eg detection sensitivities in the range of one in 100,000 or better.[17] Many variables associated with the manufacture and use of microarrays must be optimised to reach this level of sensitivity routinely. Whether it is possible to reproduce these detection sensitivities on a consistent basis is an issue that needs to be addressed by each investigator using a microarray. What is clear is that the vast majority of genes are expressed at levels which are less than 1:10,000 to 1:100,000; therefore, achieving the highest level of sensitivity is critical for all of the genes in a sample. Thirdly, Aach *et al.*[18] nicely articulated the issues related to building databases with data generated with non-Affymetrix microarrays. Experiments with these microarrays report data as a ratio of the fold change in an experimental, relative to a control, condition in order to compensate for several sources of bias and noise in the intensity results (outlined well in Zhou *et al.*[5]). Converting this ratio into a value for quantitative expression is complicated by many variables and is not straightforward to achieve.[17] Developing strategies to overcome these limitations will be essential if gene expression data produced with the microarray formats are to be used for applications in systems biology.

MPSS has many advantages over other technologies for analysing gene expression

COMPARISON OF MPSS WITH cDNA SEQUENCING, SAGE AND MICROARRAY TECHNOLOGIES

Several other technologies are available for conducting gene expression experiments. Most are based primarily on an analogue format where the relative level of gene expression is established by quantifying the hybridisation of a labelled probe to a solid support. Some other (cDNA, SAGE) technologies are similar to MPSS in that they are digital in nature and count mRNA molecules in the sample. This paper will compare MPSS with cDNA sequencing, serial analysis of gene expression (SAGE) and the microarray chip technologies. These comparisons capture most of the advantages and disadvantages of MPSS over the other gene expression profiling technologies that are currently in use.

Direct sequencing of cDNAs was the first digital technology for measuring gene expression. Both MPSS and direct cDNA sequencing involve the generation of a cDNA library as the first step of analysis. Once the cDNA library is made, sequencing of cDNA clones involves the purification and sequencing of DNA using standard procedures that are both costly and time consuming. With Megaclone, at least one million cDNA molecules are cloned onto beads, and, with MPSS, over one million clones are sequenced simultaneously. The time, effort and cost to generate data from one million mRNAs in a sample with MPSS is a small fraction of that required to sequence the same number of clones using conventional technologies.

With respect to SAGE,[7,8] MPSS has two noteworthy advantages. First, SAGE is also a transcript counting technique that generates a tag sequence for each mRNA. The length of the SAGE tag is 14 nucleotides for the current SAGE procedure, which compares with a 17-nucleotide signature with MPSS. From a theoretical standpoint, signature lengths of 14 nucleotides are 80 per cent unique, while the 17-nucleotide signature lengths generated with MPSS are approximately 95 per cent unique on the human genome. Investigators at Lynx recently mapped a large number of MPSS and SAGE tags on the available human genome sequence and demonstrated that a much higher percentage of MPSS signature sequences map to unique locations on the genome compared with SAGE tags (T. Burcham, unpublished data). Therefore, assignment of a signature sequence or tag to a specific gene on the genome is much less ambiguous with MPSS than with SAGE.

Secondly, the automated nature of MPSS makes it possible to produce efficiently a very large dataset of signature sequences. Many SAGE tag sets are comprised of only 20,000–60,000 sequenced mRNAs. Knowing that a large percentage of genes are expressed at a level of 0.01 per cent or less, it is not clear whether a dataset of 20,000–60,000 sequenced mRNAs has enough depth to allow the quantitation and analysis of all genes within a sample, particularly those that are biologically important and expressed at very low levels in the cell. An MPSS dataset of one million or more signature sequences is more likely to provide a depth of analysis that will allow low-level expressed genes to be accurately quantitated.

MPSS is complementary with microarrays for biological experiments

In comparison with the various microarray platforms, all of the issues described above for database development apply. Again, MPSS is most notable in that it is a technology that has the potential to capture virtually all genes present within the sample, and not just those that have been placed on the microarray. No prior knowledge of a gene's sequence is required for MPSS. While this is most relevant to non-human species whose genomes have not been sequenced, it also applies to genes on the human genome that have not been identified and annotated. Also microarrays have the limitation that homologous genes can cross-hybridise, which makes it impossible to detect individual members of highly homologous gene family members. With MPSS, the signature sequence, which is often in the 3′ untranslated region, can be different for individual family members. Therefore, it is possible, in many cases, to differentiate highly homologous genes from each other.

While microarrays have limitations for in-depth gene expression analyses, they have the advantage of being very useful for the high throughput analysis of multiple samples. Therefore, it may be useful to think of the microarray and MPSS technologies as being complementary in nature — different tools for different types of experiments. For example, to generate in-depth and quantitative gene expression data for building complex relational databases, MPSS may be the technology of choice. After these databases are mined for interesting biological information, it may be necessary to test whether sets of genes are differentially expressed in a large number of samples (eg tumours of a specific type); here, the microarray platform would clearly be the technology of choice. Having access to both MPSS and at least one of the microarray technologies would seem to be ideal for most investigators.

Overall, compared with several existing technologies, MPSS has the advantage that it provides in-depth quantitation of virtually all genes that are expressed in a sample. Since there is no requirement for prior knowledge of any gene or genome, it is possible to generate quantitative gene expression datasets from any organism. Additionally, since an MPSS dataset typically involves one million or more signature sequences, it has the sensitivity to quantitate accurately genes that are expressed at very low levels within a cell. No other single technology has these performance characteristics.

Acknowledgments

Thanks to Norrie Russell for carefully reading and making suggestions on the manuscript, and to Kevin Corcoran, Jen-i Mao, Keith Moon, Larry DeDionisio and Yvonne Bauer for their assistance in all aspects of the work described in this paper.

References

1. Duggan, D. J., Bittner, M., Chen, Y., Meltzer, P. and Trent, J. (1999), 'Expression profiling using cDNA microarrays', *Nat. Genet.*, Vol. 21, pp. 10–14.
2. Lipshutz, R. J., Fodor, S. P. A., Gingeras, T. R. and Lockhart, D. J. (1999), 'High density synthetic oligonucleotide arrays', *Nat. Genet. Suppl.*, Vol. 21, pp. 20–24.
3. Evertsz, E., Starink, P., Gupta, R. and Watson, D. (2000), 'Technology and applications of gene expression microarrays', in 'Microarray Biochip Technology', Schena, M., Ed., BioTechniques Books, Natick, MA, pp. 149–166.
4. Lockhart, D. J. and Winzeler, E. A. (2000), 'Genomics, gene expression and DNA arrays', *Nature*, Vol. 405, pp. 827–836.
5. Zhou, Y., Kalocsai, P., Chen, J. and Shams, S. (2000), 'Information processing issues and solutions associated with microarray technology', in 'Microarray Biochip Technology', Schena, M., Ed., BioTechniques Books, Natick, MA, pp. 167–200.
6. Hughes, T. R., Mao, M., Jones, A. R. *et al.* (2001), 'Expression profiling using microarrays fabricated by an ink-jet oliogonucleotide synthesizer', *Nat. Biotech.*, Vol. 19, pp. 342–347.
7. Velculescu V. E., Zhang, L., Vogelstein, B. and Kinzler, K. W. (1995), 'Serial analysis of gene expression', *Science*, Vol. 270, pp. 484–487.
8. Zhang, L., Zhou, W., Velculescu, V. E. *et al.* (1997), 'Gene expression profiles in normal and cancer cells', *Science*, Vol. 276, pp. 1268–1272.
9. Brenner, S., Johnson, M., Bridgham, J. *et al.* (2000), 'Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays', *Nat. Biotech.*, Vol. 18, pp. 630–634.
10. Tyagi, S. (2000), 'Taking a census of mRNA populations with microbeads', *Nat. Biotech.*, Vol. 18, pp. 597–598.
11. Walt, D. R. (2000), 'Techview: molecular biology. Bead-based fiber-optic arrays', *Science*, Vol. 287, pp. 451–452.
12. URL: www.luminexcorp.com
13. Brenner, S., Williams, S., Vermass, E. H. *et al.* (2000), '*In vitro* cloning of complex mixtures of DNA on microbeads: physical separation of differentially expressed cDNAs', *Proc. Natl. Acad. Sci. USA*, Vol. 97, pp. 1665–1670.
14. Kal, A. J., van Zonneveld, A. J., Benes, V. *et al.* (1999), 'Dynamics of gene expression revealed by comparison of serial analysis of gene expression transcript profiles from yeast grown on two different carbon sources', *Mol. Biol. Cell*, Vol. 10, pp. 1859–1872.
15. Kerr, M. K. and Churchill, G. A. (2001), 'Statistical design and the analysis of gene expression microarray data', *Biostatistics*, Vol. 2, pp.183–201.
16. Wodicka, L., Dong, H., Mittmann, H., Ho, M. and Lockhart, D. J. (1997), 'Genome wide expression monitoring in *Saccharomyces cerevisiae*', *Nat. Biotech.*, Vol. 15, pp. 1359–1367.
17. Lockhart, D. J., Dong, H., Byrne, M. C. *et al.* (1996), 'Expression monitoring by hybridization to high-density oligonucleotide arrays', *Nat. Biotech.*, Vol. 14, pp. 1675–1680.
18. Aach, J., Rindone, W. and Church, G. M. (2000), 'Systematic management and analysis of yeast gene expression data', *Genome Res.*, Vol. 10, pp. 431–445.