



MPSS

Massively Parallel Signature Sequencing

Lynx Therapeutics, Inc, Hayward, CA

www.lynxgen.com



Armidale Animal Breeding Summer Course, UNE, Feb. 2006



MPSS

Introduction

- Invented by Sydney Brenner
- Alternative to microarray: Counts all mRNAs in a sample
- Designed to capture the complete transcriptome
- High sensitivity to detect low abundance transcripts -- typical analysis involves about 1 million transcripts
- Digital data that is amenable to developing large relational databases
- Excellent dynamic range in excess of 100X up or down regulation
- Can be applied to any organism

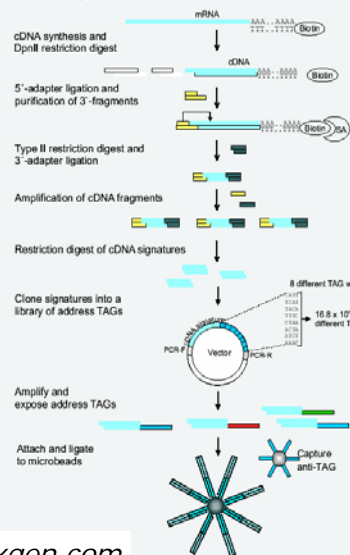
Armidale Animal Breeding Summer Course, UNE, Feb. 2006



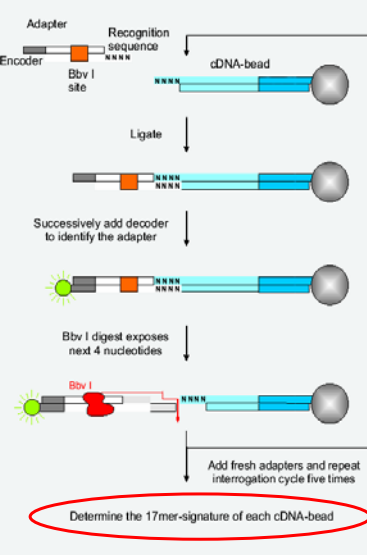
MPSS

Process Overview

Loading cDNA on microbeads



Identification of cDNAs on microbeads



Armidale Animal Breeding Summer Course, UNE, Feb. 2006



MPSS

Comparison to Microarray

- MPSS detects virtually all mRNAs in a sample, while microarrays are limited to gene elements on the array
- MPSS has greater sensitivity for routine detection of low level expressed transcripts; microarray sensitivity influenced by many factors that can be difficult to rigorously control
- “Digital” data output of MPSS makes it possible to readily import data into complex relational databases; microarray data provides a ratio between an experimental and control fluorescence that is difficult to convert into values for quantitative expression levels
- MPSS can be used to conduct quantitative and in-depth expression analysis on any organism, including those with a genome that has not been sequenced or studied in great detail
- Microarrays have the advantage of being a high-throughput technology for analyzing large numbers of samples

Armidale Animal Breeding Summer Course, UNE, Feb. 2006



MPSS

Comparison to SAGE Serial Analysis of Gene Expression

- Signature sequence of SAGE is 14 nucleotides compared with 17 nucleotides with MPSS:
 - Less ambiguity with MPSS when mapping to the mammalian genome
 - Easier to connect MPSS tags with known genes
- Typical SAGE data set is 20,000-60,000 tags compared to over a million signatures sequences for MPSS
 - Lynx cloning and MPSS sequencing done on a miniaturized platform that is amenable to high-through
 - SAGE conducted with standard cloning and sequencing that are expensive, time consuming and labor intensive
 - Larger MPSS data sets provide enhanced depth of analysis



MPSS

My First Impressions

NB: Conservative Figures

Technology	\$\$\$
cDNA	100
Affymetrix	1,000
MPSS	20,000



MPSS

My First Impressions

NB: Conservative Figures

	Microarray	vs	MPSS
N Genes	~10,000		~25,000
% DE Genes	2 – 10		15 – 25
N DE Genes	200 – 1,000		3,750 – 6,250

Biochemist



MPSS

My First Impressions

- Distribution
- Sensitivity
- Analysis

SIGNATURE	Sample 1		Sample2		p_value
	TPM	stdev	TPM	stdev	
GATCAAATTCATCTCTA	0	0	15	2	0.01014036
GATCAAATTGACCGCTT	8	6	23	9	0.07050718
GATCAAATTGGTGGGGG	11	18	2	2	0.08729055
GATCAAATTGTACTAGT	2	3	10	7	0.09091700
GATCAAATTGTGCAGTA	15	11	35	4	0.05020690
GATCCCGGTGTGAGGTA	124	1.2	125	1.2	0.58218485
GATCTGCCGGTGAGGTA	163	0	165	0	0.62550128
...					



MPSS

My First Impressions

Empirical Distribution of Tags

MPSS Paper PNAS 03, 100:4702				MPSS Test Data No Tags = 25,503		cDNA Noise Paper PNAS 02, 99:14031
tpm		N Tags	%	S 1	S 2	$f(x) = \exp\left(-\frac{2x^2}{1+x}\right)$
> 1	(0.0)	27,965	100.00	100.00	100.00	100.00
5	(0.7)	15,145	54.16	57.14	49.87	56.19
10	(1.0)	10,519	37.61	36.11	33.66	36.79
50	(1.7)	3,261	11.66	10.89	10.74	11.76
100	(2.0)	1,719	6.15	5.73	5.67	6.95
500	(2.7)	298	1.07	1.21	1.13	1.94
1,000	(3.0)	154	0.55	0.57	0.55	1.11
5,000	(3.7)	26	0.09	0.15	0.11	0.29
10,000	(4.0)	7	0.02	0.05	0.05	0.16

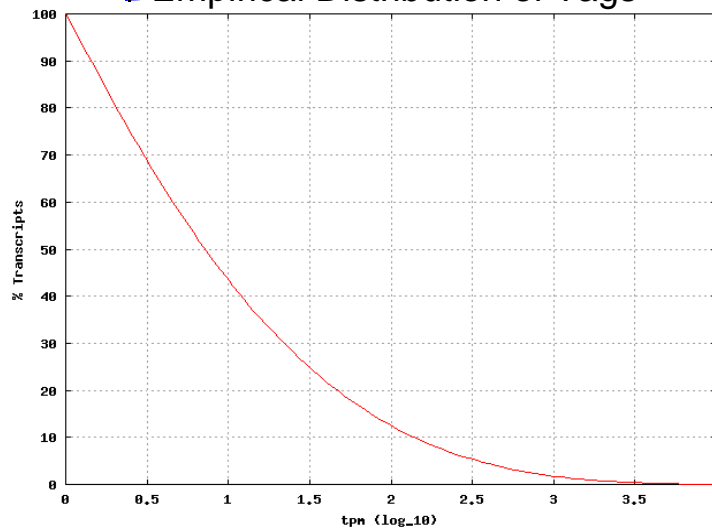
Armidale Animal Breeding Summer Course, UNE, Feb. 2006



MPSS

My First Impressions

Empirical Distribution of Tags



Armidale Animal Breeding Summer Course, UNE, Feb. 2006

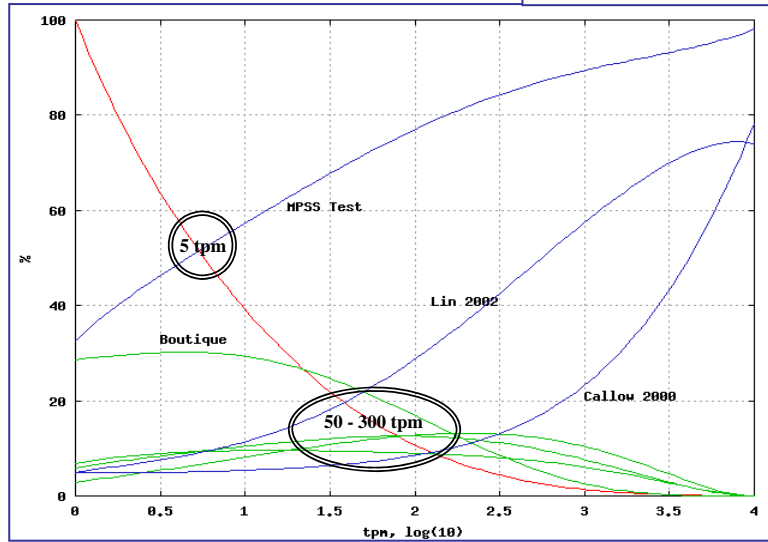


MPSS

My First Impressions

Sensitivity

Adapted from Reverter et al., 2004



Armidale Animal Breeding Summer Course, UNE, Feb. 2006



MPSS

Annotation Analysis

Annotation of Signatures No Tags: 25,503

Informatives

- ✚ Genes hit by 1 Tag: 7,492
- ✚ Genes hit by 2 Tags: 2,232
- ✚ Genes hit by >= 3 Tags: 1,775

Tags

Non-Informatives

- ✚ Hitting a Chromosome: 1,470
- ✚ "REPEAT" (>100 hits): 1,468
- ✚ Multi Genome: 1,488
- ✚ No Hits at all: 1,439

P Values for genes >= 20 tags
 $\sigma^2_w/\sigma^2_b = 0.92$

23%

Armidale Animal Breeding Summer Course, UNE, Feb. 2006



Statistical Analysis

- Categorical Data → Normal approximation for Binomial proportions

$$p_1 = \frac{x_1}{n_1} \quad p_2 = \frac{x_2}{n_2}$$

$$N\left(p_1 - p_2, \sqrt{pq\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}\right) \quad \hat{p} = \frac{x_1 + x_2}{n_1 + n_2} \quad \hat{q} = 1 - \hat{p}$$

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



Statistical Analysis

	Sample 1	Sample 2	
Gene 1	n_{11}	n_{12}	$N_{1.}$
Others	n_{21}	n_{22}	$N_{2.}$
	$N_{.1}$	$N_{.2}$	$N_{..}$

$$p_1 = n_{11}/N_{.1} \quad p_2 = n_{12}/N_{.2} \quad p_0 = 0.5(p_1 + p_2)$$

- Categorical Data → Normal approximation for Binomial proportions

$$\chi^2 = \frac{N_{..}(n_{11}n_{22} - n_{12}n_{21})^2}{N_{1.}N_{2.}N_{.1}N_{.2}}$$

$$Z = \frac{p_1 - p_2}{\sqrt{p_0(1-p_0)\left(1/N_{.1} + 1/N_{.2}\right)}}$$

- Fisher's exact (?) test:

$$P = \frac{N_{1.}!N_{2.}!N_{.1}!N_{.2}!}{N_{..}!n_{11}!n_{12}!n_{21}!n_{22}!}$$

- Audic & Claverie's test:

$$P(n_{12} | n_{11}) = \left(\frac{N_{.2}}{N_{.1}}\right)^{n_{12}} \frac{(n_{11} + n_{12})!}{n_{11}!n_{12}!(1 + N_{.2}/N_{.1})^{(n_{11} + n_{12} + 1)}}$$

$$P = \min \left\{ \sum_{k=0}^{k \leq n_{12}} P(k | n_{11}), \sum_{k=n_{12}}^{\infty} P(k | n_{11}) \right\}$$

a la SAGE data:

- Man et al., 2000
- Vencio et al., 2003

No Hypothesis testing ☹



Statistical Analysisan alternative

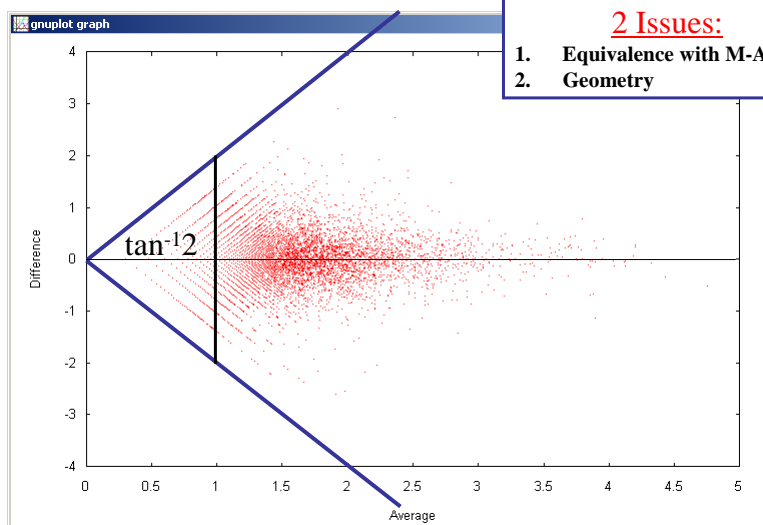
Differential Gene Expression
from MPSS data based on
Bootstrap Percentile Confidence Intervals

Presented at the International Conference on Bioinformatics
2004 – Auckland, NZ

Armidale Animal Breeding Summer Course, UNE, Feb. 2006



MPSS Test Data



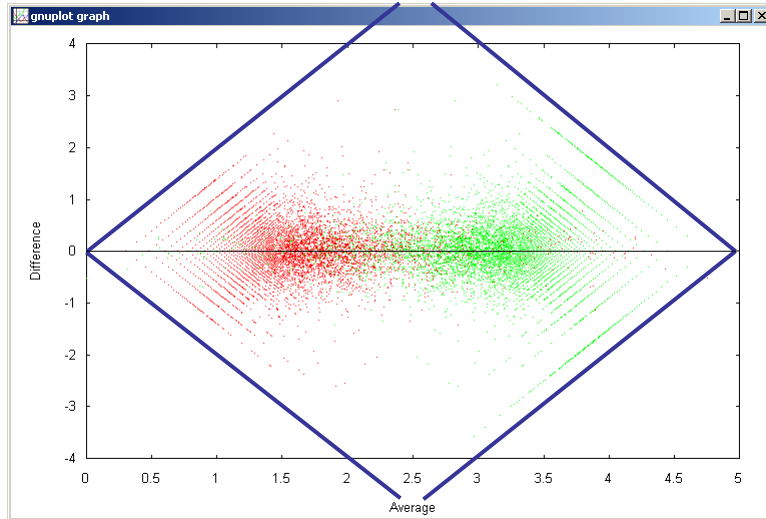
- 2 Issues:
1. Equivalence with M-A plots
 2. Geometry

Armidale Animal Breeding Summer Course, UNE, Feb. 2006



MPSS

MPSS Test Data

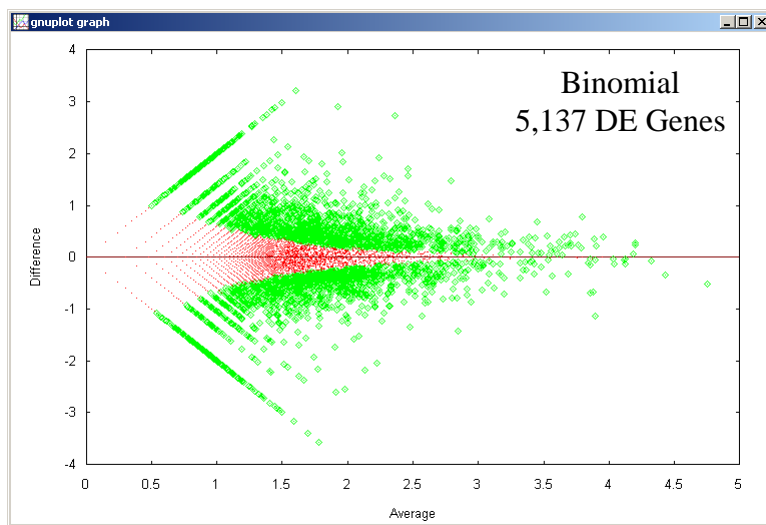


Armidale Animal Breeding Summer Course, UNE, Feb. 2006



MPSS

MPSS Test Data



Armidale Animal Breeding Summer Course, UNE, Feb. 2006



Algorithm for Bootstrap

1. Read transcripts for the i -th signature:

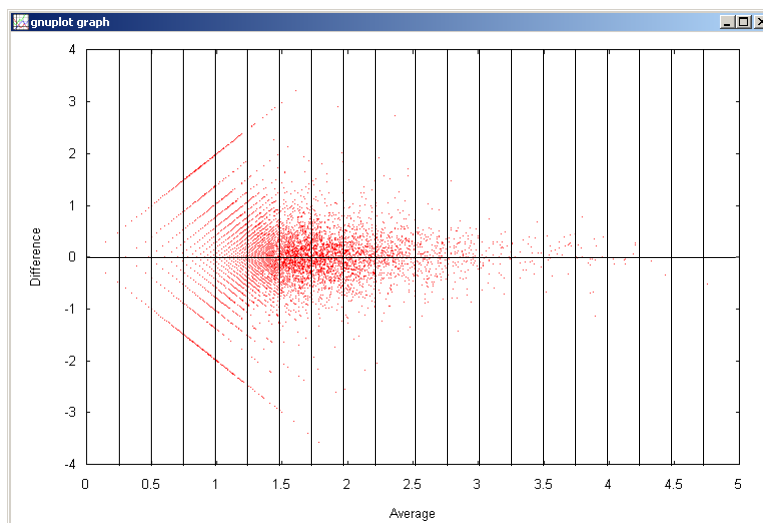
$$t_i = \{x_i, y_i\} \quad \forall i = 1, 2, \dots, n \quad \Rightarrow \quad MA_i = \{m_i, a_i\}; \quad m_i = x_i - y_i, \quad a_i = 0.5(x_i + y_i)$$

2. Sort MA_i by a_i (x-axis)
3. Define b Bins: $B_j \quad \forall j = 1, 2, \dots, b$ (Same width or Same size)
4. Define r BR (Bootstrap Replicates), ...enough for CI (eg. $r = 200$)
Define α (Significance)
5. For each B_j collect $BR_k \quad \forall k = 1, 2, \dots, r$
 - 5.1. Compute: $CI_{j,1-\alpha} = \{LB_{j,\alpha/2}, UB_{j,1-\alpha/2}\}$
 - 5.2. Identify: $MA_i(B_j) \notin CI_{j,1-\alpha}$
6. Stop

Armidale Animal Breeding Summer Course, UNE, Feb. 2006



Bins of equal width

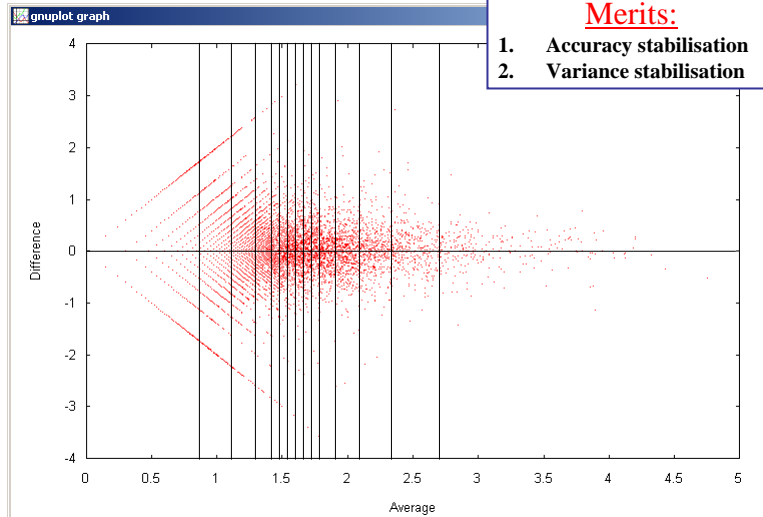


Armidale Animal Breeding Summer Course, UNE, Feb. 2006



MPSS

Bins of equal size

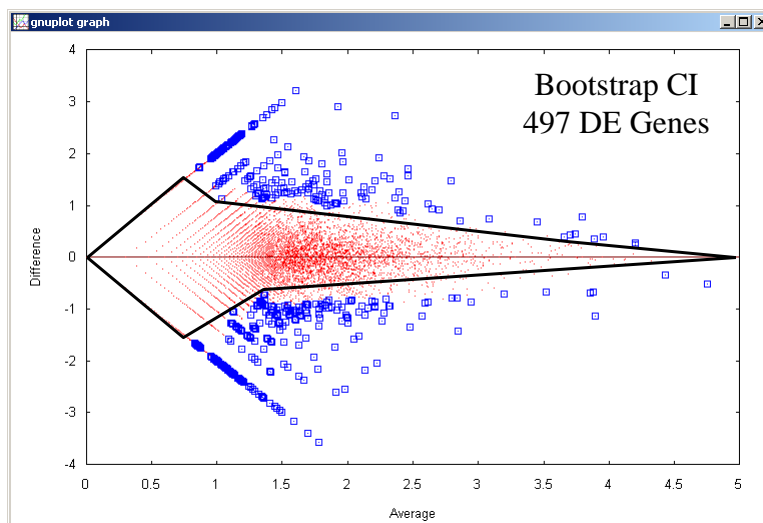


Armidale Animal Breeding Summer Course, UNE, Feb. 2006



MPSS

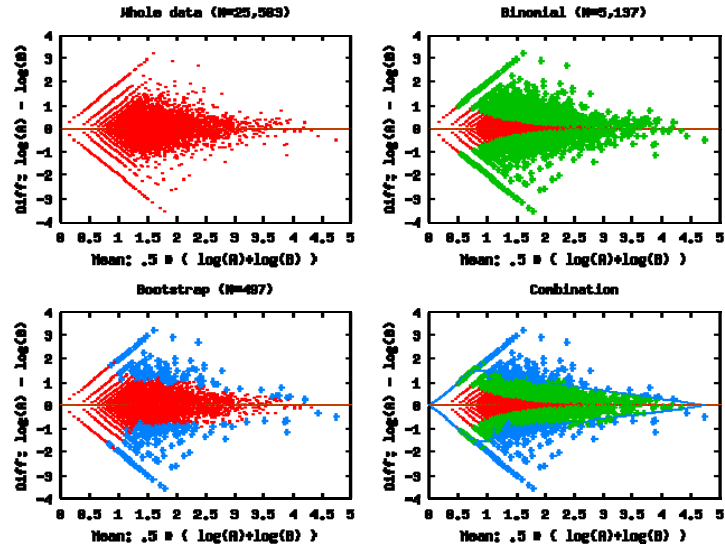
MPSS Test Data



Armidale Animal Breeding Summer Course, UNE, Feb. 2006



MPSS Test Data



Armidale Animal Breeding Summer Course, UNE, Feb. 2006



Conclusions

- Compared to microarray, the analysis of MPSS should be trivial
- Standard parametric (binomial) methods likely to generate a large number of differentially expressed elements.
 - Trade-off: Biological vs Statistically significant
- The proposed method possesses a number of advantages:
 - Very easy to implement
 - Very fast to generate
 - Operates on total transcripts as opposed to proportions
 - Accommodates the inherent heteroskedasticity
- More research (\$) is needed to assess:
 - The impact of MPSS in expression studies
 - The (possible) annotation gap (non-sequenced species)

Armidale Animal Breeding Summer Course, UNE, Feb. 2006



References

Brenner, S., M. Johnson, J. Bridgham, et al. (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. *Nature Biotechnology* 18:630-634.

Jongeneel, C.V., C. Iseli, B.J. Stevenson, et al. (2003) Comprehensive sampling of gene expression in human cell lines with massively parallel signature sequencing. *PNAS, USA*, 100:4702-4705.

Man M.Z., X. Wang, and Y. Wang (2000) POWER_SAGE: comparing statistical tests for SAGE experiments. *Bioinformatics*, 16:953-959.

Reverter, A., S. McWilliam, W. Barris, and B. Dalrymple (2004) A rapid method for computationally inferring transcriptome coverage and microarray sensitivity. *Bioinformatics* (in press).

Tu. Y., G. Stolovitzky, and U. Klein (2002) Quantitative noise analysis for gene expression microarray experiments. *PNAS, USA*, 99:14031-14036.

Vencio, R.Z.N., H. Brentani, and C.A.B. Pereira (2003) Using credibility intervals instead of hypothesis tests in SAGE analysis.