Bootstrapped ROC (bROC) for statistical analysis of microarray differential expression experiments

Bootstrapped ROC (bROC) is a resampling-based method for the determination of differentially expressed transcripts in microarray data. The approach uses large sets of simulated expression profiles derived from non-parametric noise estimates in the experimental data. ROC (receiver operating characteristic) analysis is then used for pair-wise comparisons of biological endpoints and for the identification of differentially expressed probes. The method was tested on a published collection of Affymetrix (MOE 430_2) microarray data for normal mouse tissues. Several measures were used to compare bROC performance to three reference parametric and non-parametric methods (MAANOVA, SAM and RankProd). Overall, bROC shows better performance and is especially useful in the analysis of experiments with small number of replicates.

INTRODUCTION

Analysis of differential gene expression remains an area of active research where new methods are still being introduced and no single approach has been accepted as standard. A wide range of statistical methods for the analysis of microarray data have been proposed, studied and compared [e.g., Jeffery 2006, Allison 2005, Kim 2006, Jeanmougin 2010]. In general, lists of differentially expressed genes show poor agreement between results obtained with different approaches. As described below, the bROC algorithm overcomes some of the shortcomings of previously developed methods.

The method was developed and tested using mouse transcriptome data generated at the Genomics Institute of the Novartis Research Foundation (GNF), San Diego, CA (GEO accession number GSE10246). Expression profiles in this database were obtained for 78 normal tissues from C57BL\6J mice using Affymetrix MOE 430_2 arrays, with two experimental replicates per tissue [Su 2004].

IDENTIFICATION OF DIFFERENTIALLY EXPRESSED GENES WITH bROC

The receiver operating characteristic (ROC) is a generally applicable, non-parametric procedure that provides good insight into the discriminatory properties of data features, but requires a large number of experimental samples not usually available in differential expression studies [Pepe 2007].

bROC implements numerical procedures for simulation of large sets of expression data through resampling and for fast analysis of ROC curves (Figure 1). The features (probes) are analyzed individually. For each probe on the array, bROC produces the following discrimination scores:

- Detection parameter CONF = 2(AUC−0.5), where AUC is the area under the curve. AUC = 1 indicates perfect separation of the expression measurements between two states (CONF = 1) and AUC = 0.5 indicates complete overlap of the intensities (CONF = 0). Typically, CONF = 0.95 is used as detection threshold.
- PD (probability of detection) balanced against PFA (probability of false alarm). This value is calculated at the intersection of ROC curve and diagonal of the ROC plot.
- Estimate of standard deviation of PD.
- Rank on the scale from 1 to the total number of probes. RANK = 1 indicates probes that are differentially expressed between the two states with high probability of detection (typically, PD = 1 and CONF = 1). Higher ranking is associated with smaller PD and CONF. Probes considered to have the same discrimination power have the same RANK. This parameter is helpful in sorting the probes by their importance in differentiating the two states.
- Estimate of standard deviation of RANK.

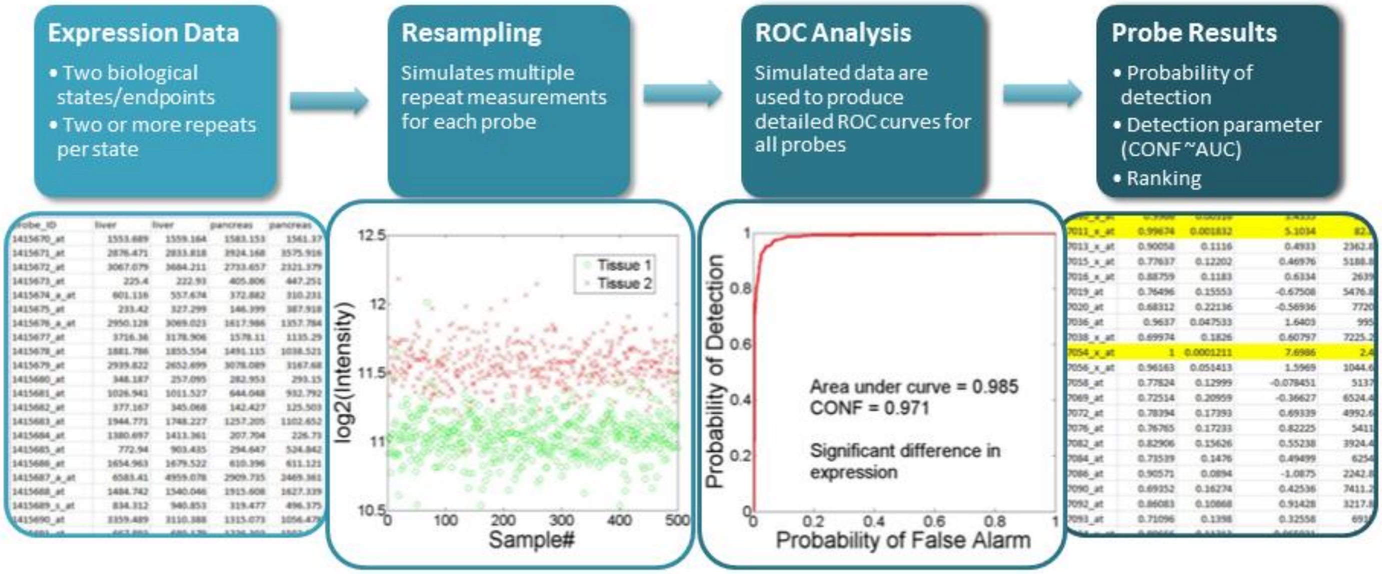
Expression Data
• Two biological states/endpoints
• Two or more repeats per state
Resampling
Simulates multiple repeat measurements for each probe
ROC Analysis
Simulated data are used to produce detailed ROC curves for all probes
Probe Results
• Probability of detection
• Detection parameter (CONF ~AUC)
• Ranking
Tissue 1
Tissue 2
log2(Intensity)
Sample#
Probability of Detection
Probability of False Alarm
Area under curve = 0.985
CONF = 0.971
Significant difference in expression

Figure 1. *Processing flow for bROC algorithm. In this example, the ROC analysis is performed on simulated sets of 500 replicate measurements for a single probe: group 1 (tissue/endpoint 1) – green circles and group 2 (tissue/endpoint 2) – red crosses. Detection parameters are derived from ROC curve shown on the 'ROC Analysis' panel.*

Comparison with other methods

We compared the lists of probes declared to be tissue-specific (in pairwise tissue comparisons) by bROC and the 'reference' methods: (SAM [Tusher 2001], MAANOVA [Kerr 2000] and RankProd [Breitling 2005, Hong 2006]). The reference methods represent different approaches currently available for analysis of microarray data. SAM (Significance Analysis of Microarrays) is a non-parametric method (no assumptions regarding the data probability distribution) that uses a moderated t-statistic and permutation of replicate measurements to estimate the false discovery rate. MAANOVA (Microarray Analysis of Variance) implements the standard analysis of variance using F-test with permutation-based adjustment for false discovery rate. RankProd uses a non-parametric statistic that detects items that are consistently highly ranked in a number of lists. It has been considered as especially useful for analysis of experiments with small number of replicates and for integration of experiments from heterogeneous platforms.

Several aspects of performance were compared as summarized below.

In contrast to other methods, the number of biological replicates has little effect on the number of tissue-specific probes found by bROC

The detection threshold in bROC is defined in terms of sensitivity and specificity for each probe on the array, which provides a clear, intuitive interpretation. The cut-off thresholds have somewhat different meanings for each of the reference methods. However, with the thresholds typically used for each method, the number of probes declared as significantly over/under expressed varies considerably between the methods (Figure 2). For data sets with two replicates, RankProd and SAM produce very short lists. The number of probes detected by bROC is larger (but reasonable), while MAANOVA appears to detect an exceedingly large number of probes (up to 1/3 of all probes).

For all the 'benchmark' methods p-value depends strongly on the number of replicates. Thus, when the same detection thresholds are used, the number of probes declared as up/down regulated increases significantly with the number of experimental replicates. On the other hand, bROC detects a similar number of probes independent of the number of replicates – Figure 2. Similar results were obtained for comparisons of tissues with different levels of similarity.

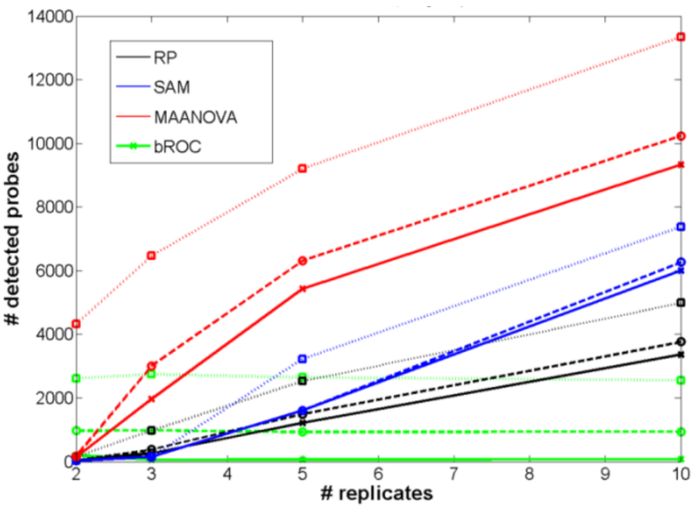

Figure 2. *Example of the dependence of the number of probes declared as differentially expressed on the number of (biological) replicates. Probes over-expressed in cerebellum in comparison with cerebral cortex. P-values of 0.005 (solid line), 0.01 (dashed line) and 0.05 (dotted line) are used as cutoff thresholds, except for bROC where CONF parameter is employed (CONF = 0.995, 0.990 and 0.950, repectively). Only bROC produces a consistent number of differentially expressed probes, essentially independent of the number of replicates. Total number of mouse-specific probe sets on the MOE 430_2 array is 42,712.*

bROC results are consistent for experiments with different number of replicates

To determine the 'consistency', we calculate the fraction (percentage) of the probes detected in all experiments with a different number of replicates (with respect to the total number of detected probes). RankProd, another non-parametric method, is most consistent in selection of differentially expressed probes, at over 80% for similar and over 90% for dissimilar tissues. This result may be related to the method favoring the fold change ranking (see below). bROC is about 80-90% consistent, while MAANOVA is less consistent (<70%) and SAM is very inconsistent(< 40%).

bROC shows desirable moderate correlation between fold change and ranking

The Pearson correlation between fold change and rank (or, method-specific detection parameter) strongly distinguishes the tested methods – Table 1.

Table 1. *Average Pearson correlation between rank (bROC, RankProd), p-value (SAM, MAANOVA – F1 and FS statistcs) for top 2000 probes detected by tested methods in 10 experiments comparing tissues with different level of similarity. STD ≅ 0.07.*

bROC	RankProd	SAM	MAANOVA
0.60	0.88	0.25	0.33/0.42

Fold change is calculated as $\log_2(I2/I1)$, where I2 and I1 are the average measured intensities for the tissues being compared. For SAM and MAANOVA, the correlation coefficient increases with the number of experimental replicates. For RankProd, there is a small decrease but the ranking is strongly correlated with measured fold change regardless of the number of available replicates and tissue similarity. The results are largely independent of the number of top probes selected. The moderate correlation observed for bROC conforms to recommendations produced by previous studies.

bROC excels in the ranking of tissue specific probes

We compared the ability of the four methods to classify as differentially expressed those probes/genes that are known to be up-regulated in specific tissues. Over 50 pairwise comparisons were performed for each method using tissues of different levels of similarity. In order to avoid ambiguities associated with detection thresholds, for the results illustrated here 2000 top-ranked probes were considered – relative performance of the algorithms is rather independent of this arbitrary number. Three different sources of publically available data were used to support the analysis: the DAVID tool [Dennis 2003], Broad Institute GSEA data [Subramanian 2005], and the Allen Brain Atlas [Lein 2007]. Where appropriate, gene symbols were converted to mouse probes on the MOE430.2 array.

DAVID Bioinformatics Resources tool integrates tissue expression data from multiple resources. Its functional annotation and classification tools help to identify tissues in which particular transcripts are preferentially, but not exclusively, expressed. A sample of results is shown in Figures 3 and 4.

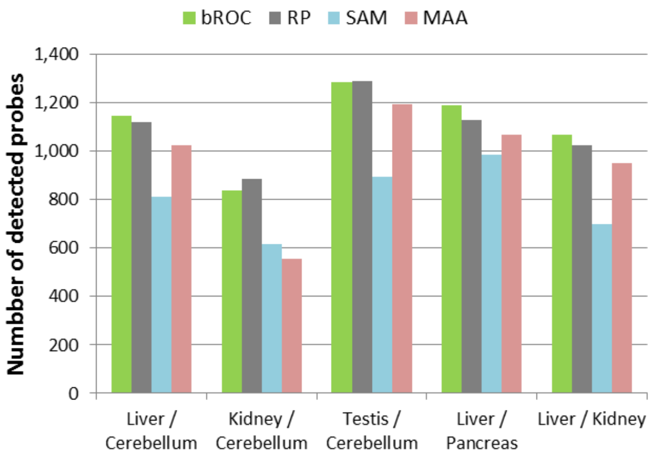

Figure 3. *Number of known tissue-specific probes detected by different methods in pair-wise comparisons (probes expected to be up-regulated in the first listed tissue). The total number of tissue–specific probes on MOE 430_2 array: liver – 8,349; kidney – 4,960; testis – 7,317 (out of 42,712 mouse probes). Data from DAVID Bioinformatics Resources tool version 6.7.*

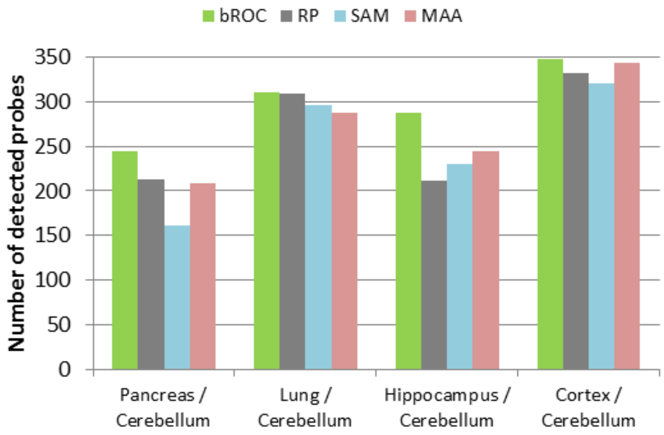

Figure 4. *Number of known tissue-specific probes detected in pair-wise comparisons of similar and dissimilar tissues (probes expected to be up-regulated in the first listed tissue). The total number of tissue–specific probes on MOE 430_2 array: pancreas – 1,445, lung – 3,177, hippocampus – 2,815, cerebral cortex – 1,428. Expression data from DAVID Bioinformatics Resources tool version 6.7.*

Distantly-related tissues were analyzed by pairwise comparisons between gene expression in liver and four different brain regions (Figure 5). A set of 337 liver specific probes was based on the Broad Institute's curated set of liver selective genes associated with key biological processes in this organ. Out of 337 liver specific probes bROC detected, on average, 75%, RankProd 70%, MAANOVA 61% and SAM only 41% of probes.

For analysis of closely related tissues we produced pair-wise comparisons of six separate brain regions. Lists of brain region-specific probes were compiled based on the data in the Allen Mouse Brain Atlas. The Atlas contains images of gene expression maps of various anatomical structures of the mouse brain obtained using high throughput in situ hybridization (which gives an independent method of verification). Depending on regions compared, 139-408 probes were analyzed in 8 separate experiments, in pairwise comparisons between 6 regions of the brain. The averaged results are summarized in Figure 6. Figure 6 also shows summary results of comparison between cerebellum and four unrelated tissues (kidney, lung, testis and liver).

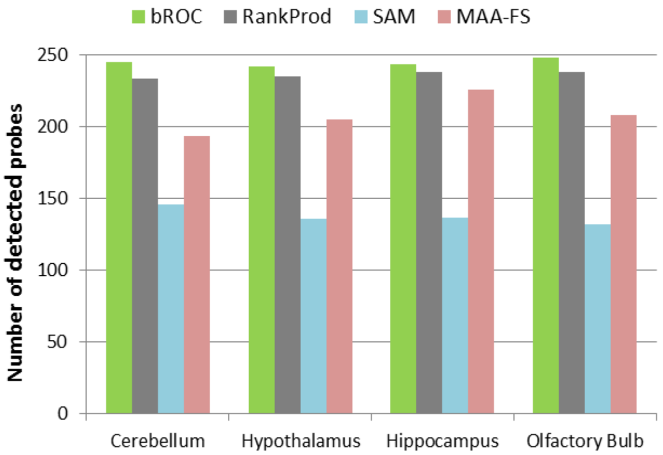

Figure 5. *Performance comparison for dissimilar tissues. Number of liver specific probes detected by different methods in top 2000 up-regulated probes in pairwise comparisons of liver with different brain tissues.*

In all cases, all methods correctly assigned the up or down regulated probes to the appropriate tissue, but differed in the numbers of detected tissue-specific probes. Overall, in most of the analyses bROC and RankProd detect significantly more tissue-specific probes than SAM and MAANOVA and bROC performed overall better than RankProd, although in some cases only incrementally better or comparably. The performance of SAM was consistently the worst. With larger number of experimental replicates (simulated data), bROC and RankProd detect about the same number of tissue specific probes, while the performance of SAM improves somewhat. In some cases with ten replicates, SAM performed comparably to bROC and RankProd. Interestingly, MAANOVA performance seemed to initially improve with the number of replicates and then to deteriorate when the number of replicates approached ten.

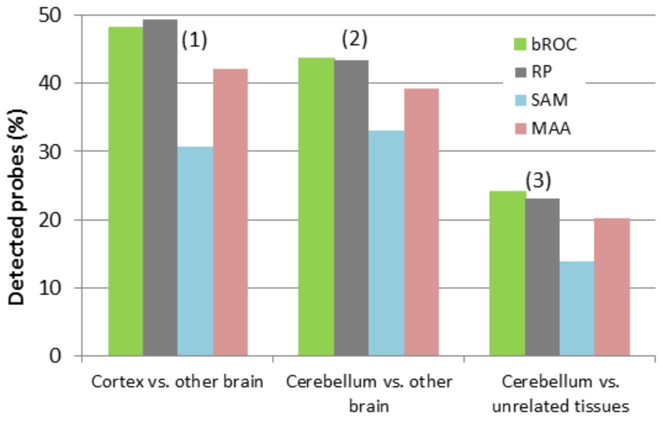

Figure 6. *Percent of known tissue-specific probes in top 2000 up-regulated probes determined by various methods. Allen Mouse Brain Atlas in situ hybridization data. Pairwise comparisons we made between (1) cerebral cortex and other parts of the brain: cerebellum (171 cortex specific probes), hippocampus (139 probes), olfactory bulb (171 probes), hypothalamus (173 probes); (2) cerebellum and other parts of the brain: cerebral cortex (395 cerebellum specific probes), hypothalamus (408 probes), hippocampus (365 probes), olfactory bulb (386 probes); (3) cerebellum and kidney, lung, testis and liver (420 cerebellum specific probes).*

Summary and Conclusions

We have considered only a representative sample of 'benchmark' statistical methods that have been proposed for the analysis of transcription microarray data. Consistent with previous studies, different methods produce vastly divergent lists of differentially expressed genes. Typically, out of the top 1000-2000 probes detected by each method, only 40-60% was also detected by other methods. Agreement between bROC and RankProd was usually higher, in the 70-90% range. All methods detected only a relatively small fraction of known tissue-specific probes/genes in the top 2000 differentially expressed probes, which indicates a large population of probes/genes that differentiate the tissues in pair-wise comparisons but apparently are not expressed uniquely in any of the tissues being compared (and thus, are not considered as tissue-specific in the source databases).

One significant result of the current study is that for Affymetrix GeneChip microarray data, the noise distributions derived through resampling procedure (for log2 probe intensities) are distinctly not normal (non-Gaussian).

Overall, bROC algorithm performs better than the considered benchmark methods in several respects:

- It is effective in discovering differentially expressed probes in experiments with small number of experimental replicates. It may be argued that a fewer number of replicates will fail to produce an adequate picture of experimental and biological noise. However, the capability to extract informative data from a low number of replicates becomes valuable in situations where specimen availability and amounts are limited (e.g., tissue samples in clinical studies).
- Detection criteria are formulated in intuitive terms related to sensitivity and specificity, with CONF (related to the area under the curve) serving as convenient summary parameter for the ROC curve.
- bROC finds similar number of up/down regulated probes independent of the number of experimental replicates. For other methods, the number of detected probes is lower for experiments with only two replicates, and increases significantly with the number of replicates when the same detection criteria are used.
- bROC probe lists show high consistency for experiments with different number of replicates (as measured by the fraction of probes declared differentially expressed in the runs with different number of replicates).
- bROC produces reasonable (0.7-0.8) and consistent correlation between fold change and ranking.
- bROC shows very good ability to detect probes/genes that are known to be expressed in certain tissues, with performance overall somewhat better than RankProd. Moreover, bROC appears to have some advantages over RankProd:
  - The correlation between detection parameter and fold change is smaller for bROC and thus, probes with small (but presumably statistically significant) fold change are more likely to be put at the top of expressed probe list,
  - RankProd lists tend to be similar to those obtained by ranking the fold change between two experimental conditions, while the moderate fold-change correlation of bROC is

- consistent with recommendations produced by previous studies [Rosenfeld 2007],
- Similar to other methods, RankProd appears to run into difficulties when only two replicates per condition are considered and the number of probes identified as differentially expressed depends strongly an on the number of replicates.

As any other statistical method, bROC can make inferences only with regard to data that is presented to the algorithm. From a practical point of view, the analysis may be better served by selection of a smaller number of high quality samples (as determined, for example, by data reproducibility) than by incorporation of a larger number of lower quality replicates. As bROC is not limited by the number of samples, the question of how to select the sample size relates more to the experimental design and data quality than to the number of replicates. In fact, larger datasets that include one or two 'noisy' arrays will produce results that are less accurate than those produced with smaller but not as noisy data sets. Note that the reference methods (particularly SAM and MAANOVA) require large number of samples per condition to be able to assess the data distributions, which may force an investigator to use low-quality data at the expense of reduced overall accuracy.

References

Allison DA, Cui X, Page GP and Sabripour M. Microarray data analysis: from disarray to consolidation and consensus. Nature Reviews Genetics (2005) 7:55-65.

Breitling R and Herzyk P. Rank-based methods as a non-parametric alternative of the t-statistic for the analysis of biological microarray data. J Bioinform Comput Biology (2005) 3:1171-1189.

Dennis G Jr, Sherman BT, Hosack DA, Yang J, Gao W et al. DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol. (2003) 4(5):P3.

Hong F, Breitling R, McEntee CW, Wittner BS, Nemhauser JL, and Chory J. RankProd: a bioconductor package for detecting differentially expressed genes in meta-analysis. Bioinformatics (2006) 22:2825–2827.

Jeanmougin M, de Reynies A, Marisa L, Paccard C, Nuel G, and Guedj M. Should we abandon the t-Test in the analysis of gene expression microarray data: a comparison of variance modeling strategies. PLoS One (2010) 5(9):e12336.

Jeffery IB, Higgins DG, and Culhane AC. Comparison and evaluation of methods for generating differentially expressed gene lists from microarray data. BMC Bioinformatics (2006) 7:359

Kerr MK, Martin M, and Churchill GA. Analysis of variance for gene expression microarray data. Journal of Computational Biology (2000) 7:819-837.

Kim SY, Lee JW, Sohn IS. Comparison of various statistical methods for identifying differential gene expression in replicated microarray data. Stat Methods Med Res. (2006) 15(1):3-20.

Lein ES, Hawrylycz MJ, Ao N, Ayres M, Bensinger A et al. Genome-wide atlas of gene expression in the adult mouse brain, Nature (2007) 445:168-176.

Pepe MS, Longton G, Anderson GL, Schummer M. Selecting differentially expressed genes from microarray experiments. *Biometrics* (2003), 59(1):133-142.

Rosenfeld S. Detection of Differentially Expressed Genes In Small Sets of cDNA Microarrays. Journal of Data Science (2007) 5:451.

Su AI, Wiltshre T, Batalov S, Lapp H, Ching A et al. A gene atlas of the mouse and human protein-encoding transcriptomes. Proc. Natl. Acad. Sci. (2004) 101:6062-6067.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS (2005) 102:15545-15550.

Tusher VG, Tibshirani R, and Chu G. Significance analysis of microarrays applied to the ionizing radiation response. PNAS (2001) 98:5116-5121.

BioFormatix

12396 World Trade Drive, Suite 315
San Diego, CA 92128
USA

info@bioformatix.com
www.bioformatix.com