

## **Microarray standard data set and figures of merit for comparing data processing methods and experiment designs: Supplementary**

Yudong D. He, Hongyue Dai, Eric E. Schadt, Guy Cavet, Stephen W. Edwards, Sergey B. Stepaniants, Sven Duenwald, Robert Kleinhanz, Allan R. Jones, Daniel D. Shoemaker, and Roland B. Stoughton\*

Rosetta Inpharmatics LLC.<sup>†</sup>, 12040 115<sup>th</sup> Avenue Northeast, Kirkland, Washington 98034, USA

<sup>†</sup>A wholly owned subsidiary of Merck & Co. Inc.

<sup>†</sup>To whom correspondence should be addressed.

### **SYSTEMS AND METHODS**

#### **Experiment Design and Rationale**

The data set was designed to provide a range of difficulty for pattern matching and detection of expression changes. mRNA samples A through T from 20 different cell lines and tissues listed in Table S1 were combined in different ways to provide condition pairs with both subtle and moderate expected expression differences. Figure S1 summarizes the sample pairings for the two-color hybridizations. Each connecting line segment represents a pair of arrays where each array is hybridized with the two samples but the assignment of fluorescent labels is reversed. This fluor reversal strategy allows removal of most of the biases caused by dye-specific interactions during the process. Each fluor-reversed-pair (FRP) of arrays in which the samples in Cy3 and Cy5 were non-identical was repeated to provide the redundancy crucial to the proposed *Type 2* figure of merit evaluation method. Thus each connector in Figure S1 actually represents four microarray hybridization experiments.

The overall experiment design can be thought of as a ring with spokes and chords. The ring design allows constructing arbitrary pairwise comparisons *in silico* and has some advantages over comparing each condition with the same reference sample. The chords allow comparing direct

experimental pairings with *in silico* estimates of these pairings. An evaluation method for this experiment design is described below utilizing control ‘A vs. A’ pairings and the *Type I* figure of merit.

The near-reference Pool 1 was formed from the 20 samples A through T (see Table S1). The distant far-reference Pool 2 was formed from 10 cell line or tissue samples that were not included in A through T. In order to produce expression differences in a moderate and small amplitude range appropriate for challenging methods development, we formed samples “Pool 1 +  $\alpha X$ ” ( $X = A$  through T) as shown in Figure S1. For samples A through J,  $\alpha = \epsilon = 0.3$ , and for samples K through T,  $\alpha = \delta = 0.01$ -0.1 (see 4th column in Table S1). In the following, we often refer to 2-color hybridization “Pool 1 +  $\alpha A$  vs. Pool 1 +  $\alpha B$ ” as “A vs. B” (profile A against B), “Pool 1 +  $\alpha X$  vs. Pool 1” as “X vs. Pool 1” (profile X against near-reference), and “Pool 1 +  $\alpha X$  vs. Pool 2” as “X vs. Pool 2” (profile A against far-reference). With this construction, the two pools will have substantial expression differences, and each pool will contain a substantial fraction of measurably expressed genes. The spoke pairings of individual samples Pool 1 +  $\alpha A$  through Pool 1 +  $\alpha J$  with Pool 1 in Figure S1 will show subtle and moderate differences of each individual sample with respect to the reference Pool 1. The spoke pairings of individual samples Pool 1 +  $\alpha A$  through Pool 1 +  $\alpha J$  with Pool 2 in Figure S1 will show pronounced systematic expression differences. Modulation of these differences by changes in the intensity-abundance relation for each reported gene in each hybridization will therefore challenge correct pattern matching of duplicates among the spoke experiments. This feature of the experiment design addresses an important issue in data processing and experiment design: using a reference condition that is too different from the conditions being studied, for example referencing many tumor samples to a normal tissue or cell line pool instead of to a pool of the tumors, reduces the ability to discriminate subtle differences between the samples (van ‘t Veer *et al.*, 2002, van de Vijver *et al.*, 2002). However, this approach is sometimes necessary to provide a common reference to diverse experiments, so optimizing data processing to be tolerant of this type of difficulty is a worthwhile goal.

Control experiments involved sample pairs that were nominally the same, but independently derived either from the same amplified cRNA material or from the same labeled ccDNA sample. Labeled ccDNA samples processed independently from the identical cRNA were hybridized to the same array. These control experiments are used to assess false positives in post-amplification steps in microarray profiling. They can also potentially be used to develop gene-specific or probe-specific error models and to support the *Type 1* figure of merit for detection performance. Since we did not derive the mRNA independently, there is no biological variability reflected in the resulting profiles. This confines our tests to assay variability. However, our methodology can be easily generalized to include biological variability in future studies. There are in total 88 experiments (10 duplicated experiments against the near-reference pool, 10 duplicated experiments against the far-reference pool, and the 20 duplicated experiment along the ring, and 4 duplicated experiments on chords) and 14 ‘same vs. same’ controls. Needless to say, the design presented in Figure S1 is a compromise between multiple analysis objectives and resources.

### **Chip Design and Rationale**

Multiple probes per gene and per exon were desired to provide tests of probe-specific error models and of benefits derived from probe averaging. Also, when two probes are known to come from the same exon, co-regulation across multiple conditions provides a test that fits within the *Type 2* figure of merit framework. A large number of reported genes was desired to provide robustness when assessing detection and pattern matching performance. Also, different kinds of control spots were desired to test correction schemes for background and cross-hybridization. These three goals are in conflict given the constraint of a single ~25,000 reporter array design. The adopted compromise design has 2,466 RefSeq genes represented by 22,657 probes (1 or 2 probes per exon, 5-10 probes per transcript with an average of 8 probes per gene). All probe sequences were 60 nucleotides long (see Methods for details).

The array also includes probes for housekeeping and array manufacturing quality control purposes. In addition, 51 negative control sequences were distributed 6 times over the array. These sequences were chosen to have little or no affinity to any expected sequence in the biological samples or

positive control spike-ins. See Methods for details concerning gene selection strategy, probe selection criteria, control probe sequences, and array layout. Chips were synthesized by Agilent Technologies as described in Hughes *et al.*, 2001.

### **Sequence Selection and Array Design**

5,703 RefSeq sequences were selected as candidates for inclusion on the basis of showing significant regulation in previous microarray experiments (data not shown). A subset of 3,141 sequences that could be effectively mapped to the draft human genome sequence and that contained at least 5 exons were divided into exons and carried forward for probe design. The adopted final design has 2,466 RefSeq genes represented by 22,657 probes (1 or 2 probes per exon, 5-10 probes per transcript with an average of 8 probes per gene).

Oligonucleotide probes were selected to represent specific exons from the candidate transcript sequences. Up to 10 oligonucleotide probes were selected for each of the RefSeq sequences. Probe selection was constrained by the presence of repeat and vector sequences, exon length, probe overlap and predicted probe performance. For each of the selected exons, the probe performance was evaluated using an algorithm that takes into account binding energies, base composition, sequence complexity, cross-hybridization binding energies, and secondary structure (T. Hughes *et al.*, 2001). The resulting array design carries 2,353 control probes and 22,657 probes for 2,466 RefSeq sequences.

### **Sample Preparation and Hybridization Protocols**

Procedures used in our microarray experiments are sketched in Figure S2. mRNA samples were purchased from Clontech (see Table S1 for details). For duplicate hybridizations, the same intermediate cRNA product (Figure S2-B) was split and the two halves were processed independently through the remainder of the protocol.

We describe procedures used in our microarray experiments as sketched in Figure S2. Extraction of mRNA from tissues or cell lines listed in Figure S2-A was not done in house; instead, we purchased mRNA samples from Clontech (see Table S1 for details). A randomly primed amplification protocol was used to generate full-length cDNA for use in the hybridization in procedure B. The second RT reaction in C produces ccDNA samples that are ready for dye coupling and hybridization. Procedures in D are part of data extraction and analysis. After hybridization, slides were washed and scanned using a confocal laser scanner (Agilent Technologies). Fluorescence intensities on scanned images were quantified, corrected for background and normalized.

Two types of control experiments were carried out. cRNA-to-end control experiments include variations resulting from the second reverse transcription reaction, purifying ccDNA, coupling Cy-dyes to ccDNA, cleaning up uncoupled dye, hybridization, microarray synthesis, post-hybridization washing, scanning and image processing. Hybridization-to-end control experiments include only variations resulting from hybridization, microarray synthesis, post-hybridization washing, scanning and image processing.

### **Baseline Image Processing, Background Determination, Normalization, and De-trending**

Fluorescent images obtained with the Agilent scanner were quantitated using an in-house image processing code called Qhyb. Segmentation identifies the most trustworthy regions of each spot and pixels for local background estimation. Automated artifact recognition flags suspicious spots based on a neural net classifier that uses measures such as color ratio uniformity and pixel intensity distribution shape and that was trained on human-selected spot training sets. Typically ~1% of spots are flagged as artifacts in this way.

Although negative control spots are included on the array, the baseline processing uses local out-of-spot background pixels averaged over regions ~ 10 by 10 spots in size to estimate the background signal. Correction for sequence specific cross-hybridizations was not attempted. Out-of-spot background

often provides insufficient background subtraction; however, this bias works primarily to move the operating point of the detector *along* the ROC performance curve, trading false positives against positives, and only secondarily hurts the overall detection performance. The background-subtracted intensities from two channels are initially normalized by the mean intensity of each channel, excluding control spots. The second step in background subtraction balances the background bias between two color channels. This step is carried out by selecting low intensity spots and performing a linear fit between intensities from the two colors. The offset is subtracted from the channel with extra background residual. The two background-balanced channels are further re-normalized and de-trended by first binning the data according to logarithmic intensity (geometrical mean of both channels), and then fitting a linear relation between two channels using mean logarithmic intensity from each bin. The offset is due to the gain difference between two channels, and the coefficient to the logarithmic intensity could be due to the slight non-linearity of the scanner used. Both effects are corrected in the red channel. This correction is essential to avoid artifactual pattern similarities when looking for patterns composed of subtle expression changes.

### **Baseline Error Model for a Single Microarray**

The repeatability of measured expression change for each gene across multiple nominally identical independent experiments is the strongest input to the assignment of confidence value. However, in order to average repeated experiments with appropriate relative weights, and to avoid underestimates of variance due to small sample sizes with consequent false positives, a model for the uncertainties in individual array experiments is required. We assign statistical significance to observed change via the statistic (Roberts *et al.*, 2000):

$$X = (a_2 - a_1) / [\sigma_1^2 + \sigma_2^2 + f^2 (a_1^2 + a_2^2)]^{1/2} \quad (1)$$

where  $a_{1,2}$  are the mean-pixel intensities measured in the two channels for each spot,  $\sigma_{1,2}$  are the (additive) uncertainties due to background subtraction, and  $f$  is a fractional (multiplicative) error

such as would come from hybridization non-uniformities, fluctuations in the dye incorporation efficiency, scanner gain fluctuations, etc. Note that while we often refer to ratio and log(ratio) profiles in the text, the underlying error model begins with expression differences. It tends to approximate log(ratio) behavior at the larger intensities as well as for small to moderate expression changes. Equation (1) is empirically motivated in the sense that  $X$  is observed to be approximately normal for many different array technologies and experimental conditions. Ideally,  $\sigma$  and  $f$  are chosen so that  $X$  has unit variance and tracks the increase in fractional error toward lower-intensity spots.  $\sigma$  tends to vary from array to array and so is derived for each array based on the background fluctuation level. The  $\sigma$  of background was first estimated in each panel by averaging over 4 pixels excluding outlier pixels whose intensities are 3 times or more brighter than the median intensity, and then taking the standard deviation of the remaining pixels. For a single slide, we employ conservative adjustments to the mean  $\sigma$  by scaling these values by a factor of 2.5 for all panels and taking these scaled values as the background fluctuation for the slide. We have found  $f$  to be fairly constant from array to array across a number of control experiments where nominally the same sample is hybridized in both channels (data not shown). We opted for a very conservative value  $f=0.2$  for our baseline processing.

The two-sided probability for an observed regulation of magnitude  $|X|$  is then

$$P\text{-Value} = 2(1 - \text{Erf}(|X|)) \quad (2)$$

Because of its conservative nature, it should be pointed out that our P-Value was designed for the purpose of ranking the significance of the differential regulation, not for accurately estimating the false positives.

For display and array averaging purposes it is useful to express measurements and errors in terms of  $\log_{10}(a_2/a_1)$ . This requires somewhat arbitrarily repairing negative values to a small positive value equal to a scanner raw count; we also cap the resulting ratios to lie between 0.01 and 100. The uncertainty in the log(Ratio) is then defined as

$$\sigma_{\log_{10}(a_2/a_1)} = \log_{10}(a_2/a_1) / X \quad (3)$$

and this allows weighted combining of multiple observations of  $\log(\text{Ratio})$ . Measurements that come from faint spots near the background will have small values of  $X$ , large errors, and will be given low weight when combining repeated measurements. This kind of conversion from linear to log is only approximately right when  $X$  is small. However, for the purposes of ranking significances, even large values of  $X$  fit well within this framework.

### Baseline Combining Replicates and Redundant Probes

Linear averaging of intensities, rather than averaging of the  $\log(\text{Ratio})$ , is possible, but leads to poorer accuracy in technologies where the absolute intensity calibration is uncertain from array to array. We use the minimum-variance weighted average to compute the mean  $\log_{10}(a_2/a_1)$  of each reported gene:

$$w_i = 1 / \sigma_i^2 \quad (4)$$

$$\bar{x} = \sum_{i=1,n} w_i x_i / \sum w_i \quad (5)$$

Here  $\sigma_i$  is the error of  $\log_{10}(a_2/a_1)$ ,  $x_i$  stands for  $i$ -th measurement of  $\log_{10}(a_2/a_1)$ , and  $n$  is the number of repeats.

The error of  $\bar{x}$  can be computed in two ways. One is to propagate the errors  $\sigma_i$ , and another is from the scatter of  $x_i$ :

$$\sigma_p^2 = 1 / \sum w_i, \quad (6)$$

$$\sigma_s^2 = \frac{1}{(n-1) \sum w_i} \sum w_i (x_i - \bar{x})^2. \quad (7)$$

The propagation error  $\sigma_p$  relies totally on the error estimation of each individual slide, and therefore is subject to bias or systematic uncertainties. Whereas the error from the scatter of the data,  $\sigma_s$ , is an unbiased measure, but has large fluctuations when the number of repeats is small. Ideally, we would like to use  $\sigma_p$  when there is only one measurement, and gradually shift to  $\sigma_s$  when the number of repeats is

large. We can accomplish this by doing a weighted mean of error. Statistically, the error of  $\sigma_s$  is expressed as:

$$\sigma_{\sigma_s} = \frac{\sigma_s}{\sqrt{2(n-1)}} . \quad (8)$$

From Eqn. (8) we can conclude that the weight for combining is proportional to  $n$ . We therefore combine the two errors as:

$$\sigma_x = \frac{\sigma_p + (n-1)\sigma_s}{n} . \quad (9)$$

The above procedure was also used to combine multiple probes representing the same exon or gene.

### Virtual Combining of Profiles

For samples  $S_i$ ,  $i = 1, \dots, N$ , one can form the ratio of  $S_N / S_1$  either by direct sample pairings or using the intermediate contiguous sample pairs. A simple approach is to form the estimate

$$\frac{S_N}{S_1} = \prod_{i=1}^{N-1} \frac{S_{i+1}}{S_i} . \quad (10)$$

Taking the log of both sides of this expression transforms it into the sum

$$\log_{10}\left(\frac{S_N}{S_1}\right) = \sum_{i=1}^{N-1} \log_{10}\left(\frac{S_{i+1}}{S_i}\right) . \quad (11)$$

We denote the intermediate log-ratios as  $LR_k = \log_{10}(S_{i+1}/S_i)$  and the associated error bars as  $\sigma_k$ , where  $k = 1, \dots, N-1$ . Assuming independence between the samples, the error bar of the overall  $LR = \log_{10}(S_N / S_1)$  can be computed as the sum of squares of individual error bars

$$\sigma^2 = \sum_{k=1}^{N-1} \sigma_k^2 . \quad (12)$$

Although the errors are in fact correlated over the index  $k$ , the error model for  $\sigma_k$  attempts to estimate the uncorrelated component. The statistic  $X$  and corresponding p-value can then be computed as above using LR and  $\sigma$ , with

$$X = \frac{LR}{\sigma} \cdot \quad (13)$$

### Cluster Analysis Based on Similarities

We used several clustering algorithms including agglomerative hierarchical clustering and k-means and k-median clustering (Hartigan, 1975). To drive clustering we used the correlation based similarity metric  $D_{ij} = 1 - \rho_{ij}$  in which

$$\rho_{ij} = \frac{\sum_{k=1}^n \left( x_{ik} / \sigma_{ik} \right) \left( x_{jk} / \sigma_{jk} \right)}{\sqrt{\sum_{k=1}^n \left( x_{ik} / \sigma_{ik} \right)^2 \sum_{k=1}^n \left( x_{jk} / \sigma_{jk} \right)^2}} \quad (14)$$

is the error-weighted correlation coefficient between experiments  $i$  and  $j$  calculated using the  $\log_{10}$ (expression ratio) values  $x_{ik}$  and their associated errors  $\sigma_{ik}$ . We also used conventional Euclidean distance.

## RESULTS

### Ring vs. Spoke Sample Referencing Schemes

For the purpose of comparing ring-type and spoke-type sample pairing schemes, we conducted four direct hybridization experiments on chords. They are Pool 1 + 0.3A vs. Pool 1 + 0.3C (A vs. C), Pool 1 + 0.3A vs. Pool 1 + 0.3D (A vs. D), Pool 1 + 0.3A vs. Pool 1 + 0.3F (A vs. F), and Pool 1 + 0.3A vs. Pool 1 + 0.3I (A vs. I). One can always derive virtual experiments A vs. C, A vs. D, A vs. F, and A vs. I *in silico* based on experimental data either from ring-type design or from spoke-type design without performing the

corresponding actual hybridizations. For example, one can derive virtual hybridization experiment A vs. D by combining A vs. Pool 1 and Pool 1 vs. D in the spoke-type design (see Methods). One can also achieve the same goal by combining experiments along the ring either clockwise or counter-clockwise, or both. The number of experiments involved in the derivation of virtual hybridization data for the four chord experiments A vs. C, A vs. D, A vs. F, and A vs. I is *always* 2 via Pool 1; 2, 3, 5, and 8 via ring clockwise; and 18, 17, 15, and 12 via ring counter-clockwise. As the number of experiments in the derivation increases, variance accumulates, and therefore the experimental uncertainty becomes large in ring-type designs. Figure S3-A shows the normalized number of positives as a function of the normalized number of false positives for four real chord experiments, together with those derived via Pool 1, via ring clockwise, and via ring counter-clockwise (see Figure S1). We note that for A vs. C, two steps are required via either Pool 1 (use A vs. Pool 1 and C vs. Pool 1 to derive A vs. C) or ring clockwise (use A vs. B and B vs. C to derive A vs. C). Figure S3-A shows the same performance level for these two routes.

Figure S4 indicates that for a given FP rate, the detection efficiency varies as a function of the number of experiments involved in the derivation of virtual data. The effect becomes even more striking when one compares the FP rates at a fixed (FP + TP) rate. For example, when one declares that ~5,000 reporters are significantly regulated (normalized number of declared positives = 0.2 indicated by the horizontal black dotted line in Figure S3), one would declare ~15 false positives (normalized number of false positives = 0.0006) for real experiments. However, one would declare ~125, ~300, ~750, ~1,250, ~1,750, ~2,250, ~2,500, and ~2,500 false positives in virtual experiments involving 2, 3, 5, 8, 12, 15, 17, and 18 steps, respectively. The degradation is much more severe according to this measure than it is based on mean-squared error.

Among the potential advantages of the ring-type experiment design are that one may use the minimal number of chips and that one can always sequentially add more experiments as a project expands. These advantages may be more than compensated by the loss of data quality when all virtual pair-wise comparisons are needed among profiled samples. In our case the unique signatures associated with each tissue (C, D, F, I vs. A) were strong enough that we were still able to achieve the same level of

pattern matching (FOM2 = 1 for either route). Forming *in silico* ratios *via* the ring can also result in an accumulation of data drop-outs. Similar performance evaluations for various designs can be carried out based on the data set presented here. The results are not shown here because of space limitations.

## References

Hartigan, J. 1975. *Clustering Algorithms*, John Wiley & Sons, New York.

Hughes, T.R., Marton, M.J., Jones, A.R., Roberts, C.J., Stoughton, R., Armour, C.D., Bennett, H.A., Coffey, E., Dai, H., He, Y.D., Kidd, M.J., King, A.M., Meyer, M.R., Slade, D., Lum, P.Y., Stepaniants, S.B., Shoemaker, D.D., Gachotte, D., Chakraburty, K., Simon, J., Bard, M., Friend, S.H. 2000. Functional discovery via a compendium of expression profiles. *Cell* **102**:109-126.

Roberts, C.J., Nelson, B., Marton, M.J., Stoughton, R., Meyer, M.R., Bennett, H.A., He, Y.D., Dai, H., Walker, W.L., Hughes, T.R., Tyers, M., Boone, C., Friend, S.H. 2000. Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles. *Science* **287**:873-880.

**Table S1. mRNA samples from human tissues or cell lines that were profiled in the standard data set.** Pool 1 is made of equal portion of 20 samples A through T. Pool 2 was made of equal portions of 10 samples a through j, disjoint from A through T.  $\alpha$  is the mixture factor  $\epsilon$  or  $\delta$  in Figure S1. The last column contains the number of differentially expressed genes with P-value  $< 0.01$  and  $|\log(\text{ratio})| > 0.3$  using the Rosetta 25k human chip (Hughes *et al.*, 2001) in a previous study (Shoemaker *et al.*, 2001).

Symbol	Tissue / Cell line Sample	Pool	$\alpha$	# Signature
A	Testes	1	0.30	3358
B	Fetal Kidney	1	0.30	3470
C	Thymus	1	0.30	2506
D	Spleen	1	0.30	3543
E	Brain - Thalamus	1	0.30	2824
F	Uterus	1	0.30	3073
G	Brain - Caudate Nucleus	1	0.30	2672
H	Burkitts Lymphoma (Raji)	1	0.30	2286
I	Brain - Amygdala	1	0.30	1949
J	Leukemia Promyelocytic (HL-60)	1	0.30	2053
K	Fetal Lung	1	0.01	2989
L	Brain - Cerebellum	1	0.02	1463
M	Pancreas	1	0.03	1617
N	Leukemia Lymphoblastic (MOLT-4)	1	0.04	1716
O	Burkitts Lymphoma (Daudi)	1	0.05	1821
P	Adrenal Gland	1	0.06	981
Q	Thyroid	1	0.07	481
R	Bone Marrow	1	0.08	1498
S	Small Intestine	1	0.09	1713
T	Placenta	1	0.10	897
a	Stomach	2		411
b	Heart	2		1059
c	Spinal Chord	2		1580
d	Prostate	2		288
e	Fetal Liver	2		848
f	Trachea	2		86
g	Salivary Gland	2		491
h	Lymph Node	2		475
i	Brain - corpus callosum	2		558
j	Skeletal Muscle	2		545

**FIGURE CAPTION**

**Figure S1. Experiment design.** Each connecting line or arc represents a redundant pair of fluor-reversal array pairs (total 4 hybridizations). Virtual ratios generated by traversing multiple arcs can be compared with actual two-color ratio profiles done as indicated by the chords. Pool 1 is an average of human tissue or cell line samples A + B + ... + T and so provides a near reference to the condition pairs on the ring. Pool 2 is an average of 10 tissue or cell line samples other than A through T and so provides a distant far-reference, such as would occur if cancer profiles were compared to normal rather than to a common pool of affected individuals. Loops indicate “same vs. same” control experiments (cRNA-to-end in blue color and hybridization-to-end in brown) in which can be used to generate estimates of false positive rates and more detailed error models. For details of sample information and  $\epsilon$  and  $\delta$ , see Table S1.

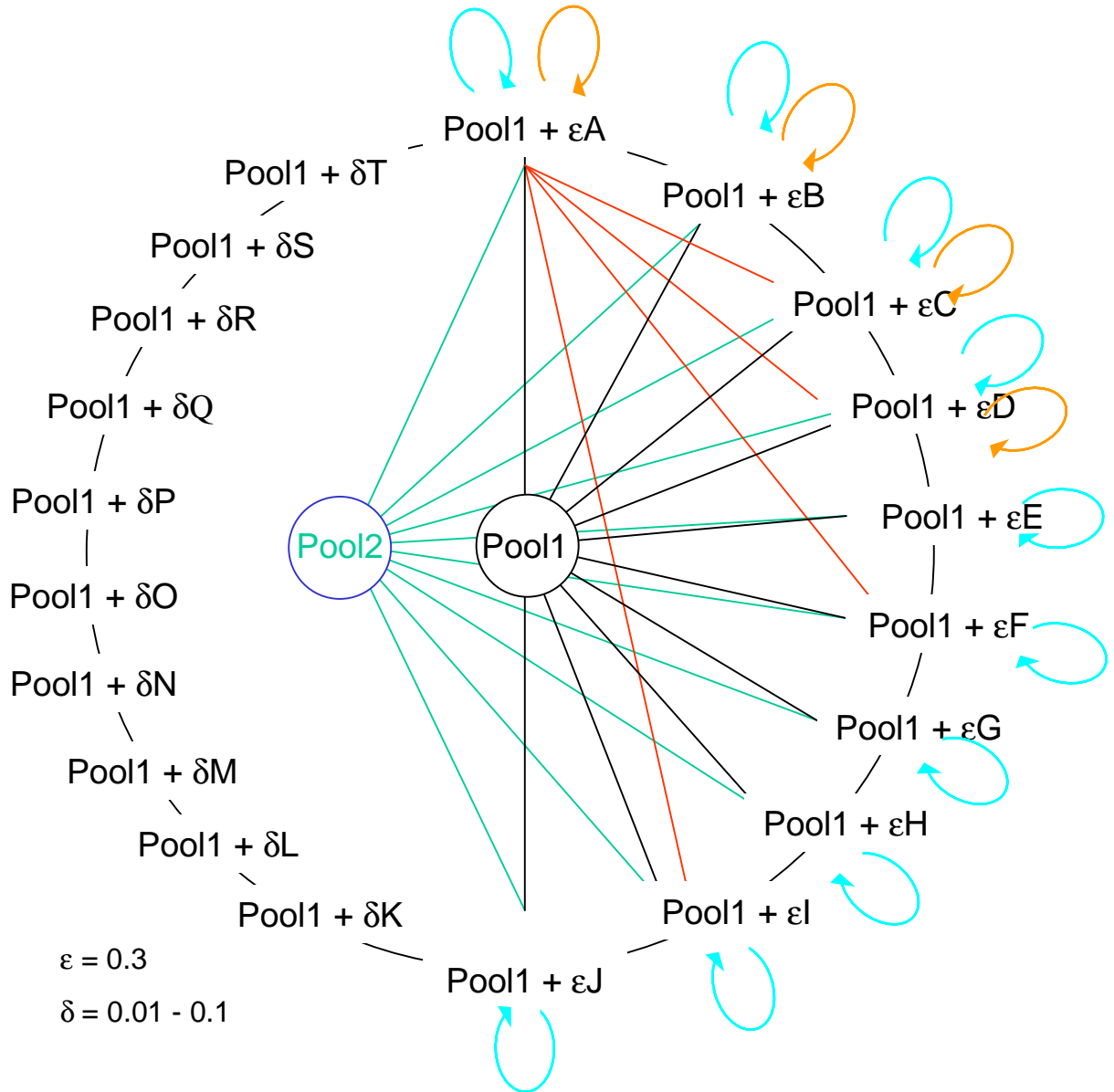
**Figure S2. Experimental procedure.** Steps are described in detail in text.

**Figure S3. FOM1 for different sample referencing schemes.** Similar to Figures 4 and 5 in the paper but comparing the performance of ring and spoke sample referencing schemes – see Figure S1. Dotted lines: virtual ratio profiles constructed by transitivity via Pool 1; for example, constructing A vs. C from the actual experiments A vs. Pool 1 and C vs. Pool 1. Solid and dash-dot lines: virtual ratio profiles constructed by transitivity along the ring; for example, constructing A vs. C from the actual experiments A vs. B and B vs. C. Dashed lines: actual experiments, as indicated by the chords in Figure S1. Numerals indicate number of segments transitioned along the ring. Vertical dashed line indicates constant false positive rate  $10^{-4}$ .

**Figure S4. Comparison of different sample referencing schemes.** (A) Standard deviation of logarithmic ratios averaged over all probes in the four real condition pairs indicated by the chords in Figure S1 (marked as “Real”), in comparison with corresponding virtual ratio profiles derived from

relevant profiles via Pool 1 (marked as “via Pool 1”), along the ring clockwise (marked as “via Ring +”) and counter-clockwise (marked as “via Ring -”). (B) Similar to (A), but estimated errors from our baseline processing error model. The number of experiments involved in the deviation for each virtual pair is marked.

Figure S1



**Figure S2**

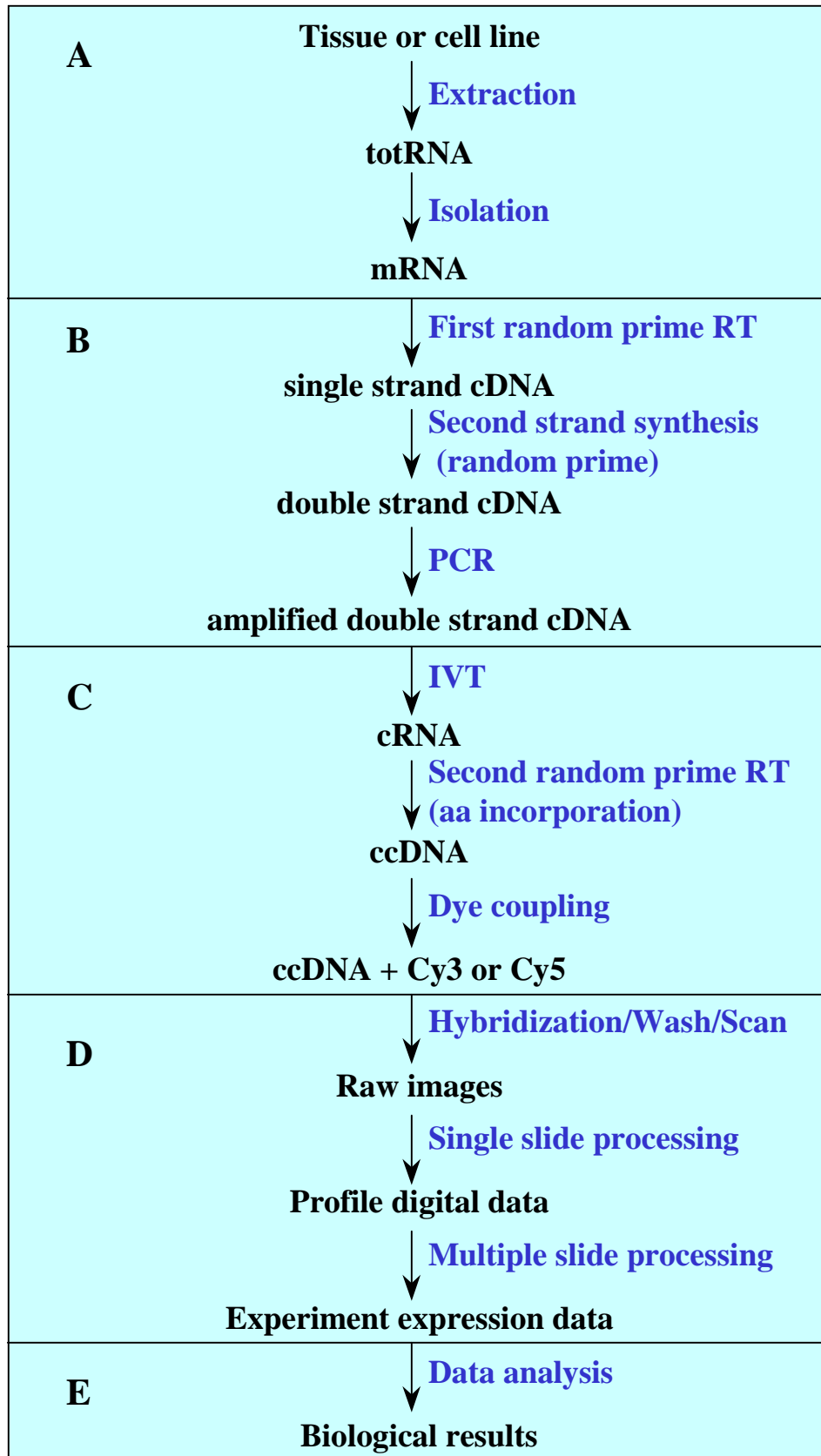


Figure S3

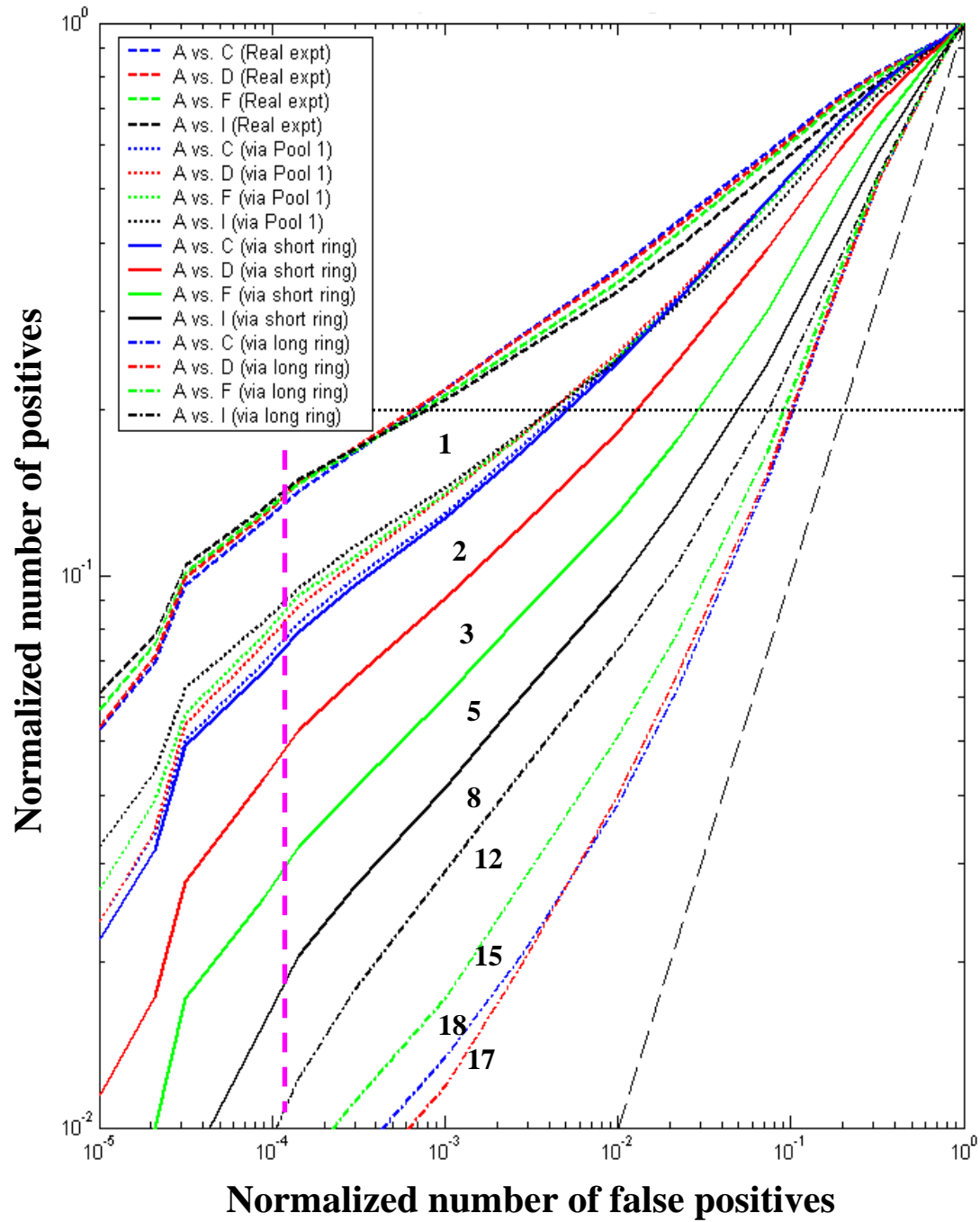


Figure S4

