Comparison of Affymetrix expression array summarization methods for reproducibility and consistency across studies

Xiaoyang Ruan[1], Ourania Kosti[2], Rado Goldman[2], Hongfang Liu[1*]

[1]Department of Health Sciences Research, Mayo Clinic College of Medicine, Rochester, MN

[2]Department of Oncology, Lombardi Comprehensive Cancer Center, Washington DC, USA

ABSTRACT

Affymetrix gene expression microarray is a popularly used platform for differential analysis. The analysis pipeline includes five steps: background correction, normalization, PM-only correction, and summarization, and differential analysis. Using publicly available microarray data, we compared the performance of five summarization methods: Median, Mean, Median Polish, Robust Linear Model, Li-Wong. Our evaluation criterion was reproducibility between studies designed to answer same scientific questions. Our analysis shows that mean value summarization gives smaller number of transcripts with inconsistent fold change direction while maintaining reproducibility comparable to competing complex methods. We conclude that after raw data has been preprocessed by the most popularly used pipeline (Robust Multiple Regression (RMA) background correction, quantile normalization, and PM-only correction), mean value summarization may convey a better representation of the true expression levels of target transcripts. The study suggests that the selection of bioinformatics algorithms needs to be application oriented. Sometimes simple initiative approach is probably better.

1 INTRODUCTION

Microarray technology, based on DNA hybridization to measure expression levels of mRNA or to detect Single Nucleotide Polymorphism (SNP) and copy number, has become an invaluable tool in biomedical research since the mid 1990s [1, 2]. One of the popular gene expression microarray platforms is Affymetrix where a target transcript is typically represented by a probe set consisting of 11-16 pairs of short oligos. Each pair consists of a perfect match (PM) and a mismatch (MM) oligo. The PM probe exactly matches the sequence of a particular standard genotype, while the MM differs in a single substitution in the central (13th base), intended to distinguish noise caused by non-specific hybridization. Transcript expression level is a summarization of the signal of individual probes in the corresponding probeset [3].

Data analysis for Affymetrix microarray generally consists of four preprocessing steps: background correction, normalization, PM correction and summarization. Background correction removes noise signals arising from many sources, such as non-specific binding, processing bias in wash stage or optical noise from the scanner. Normalization rescales intensity from multiple chips to the same level so that gene expression levels on different chips can be comparable. PM correction controls for non-specific binding between probe and non-target sequences. The summarization step estimates the transcript expression level based on intensity measures of probes in the corresponding probe set.

There is a rich source of algorithms available to pre-process raw data from Affymetrix gene expression array. A relatively complete list of currently available preprocessing steps was tabulated by Irizarry R.A. [4] and Harr B. [5]. However, some of these have become obsolete given the accumulating evidence of poor performance. For example, MAS and subtractmm methods for PM correction were shown to consistently yield negative signals, which indicates that use of MM probes for detection of non-specific binding is unreliable [3, 6]. The widely used background correction method, robust multi-array average (RMA), relies solely on PM values [3]. GCRMA [7] was developed to take the effect of GC content on different probes into consideration. Bolstad *et al.* [8] compared several normalization methods and showed that quantile normalization has advantages in both speed and bias. Nowadays, the following pre-processing pipeline, RMA or GCRMA background correction—quantile normalization—pmonly correction—median polish or Li-Wong summarization, has become a standard [4, 5, 9].

The performance of various pre-processing methods is generally evaluated using spike-in and dilution data series [3, 4, 10, 11], MAQC data series [12-15], or based on the classification power of the number of differentially expressed genes obtained [16]. When using spike-in data, the differentially expressed genes are known in advance and assumed to be the true targets. However, this assumption is not safe in biological questions since it is unknown whether a gene expression difference reflects a true biological difference or not. This is especially important in microarray data analysis because of the high background noise and the various sources of variation (including but not limited to differences in probe labeling efficiency, RNA concentration, and hybridization efficiency). Moreover, many known comparison studies are based on a single dataset or specific controlling samples. This is potentially susceptible to the data structure of specific type (or group) of sample, or specific type (or batch) of microarray chip. MAQC [12] project spearheaded by FDA involved multi-platform and cross-lab comparison. However, it is actually based on fixed controlling RNA samples. Several existing publications on MAQC project did not discuss the performance of different summarization

*To whom correspondence should be addressed methods in true data. It is hard to design a single trail that can take all potential confounding factors into consideration. In this paper, we compared the performance of different summarization pipelines by applying competing algorithms on microarray dataset pairs that are publicly available and can be used to answer the same scientific questions. We aimed to identify the method(s) that yield(s) consistent results between the pairs. In the following, we first present experimental design and evaluation metrics. We then discuss and conclude the study.

2 METHODS

2.1 Experiment Design

The experiment contains three levels of cross validation. The first level is different datasets pairs extracted from research results, which sheds a light on the possible performance difference caused by data structure of specific samples. The second level is different microarray platforms, which helps to avoid platform specific influence. The last level is the use of two different differential analysis algorithms, which takes the possible impact of algorithms specific effect on the competing methods into consideration.

Specifically, we identified three dataset-pairs (six datasets in total) from respective Affymetrix microarray platforms. The preprocessing pipeline is fixed to RMA—quantiles—pmonly, and only different summarization methods were compared. Five summarization algorithms primarily available in the latest Affymetrix built-in processing method [17], including median (Avgdiff), mean, median polish [10], robust linear model (RLM) [18, 19] and Li-Wong (dChip) [20], were compared for reproducibility between datasets extracted to address the same questions. Two other summarization methods, MAS [21] and playerout [22], were not discussed because they are less common these days (Table 1).

Two differential analysis algorithms (significance analysis of microarray (SAM) and CyberT) were implemented to the processed datasets to get the final result. SAM [23] estimates t statistics by adding a small constant s_0 to denominator to minimize coefficient of variation at low expression level. CyberT [24] uses regularized t-test in the Bayesian probabilistic framework. We also utilized GeneGo webtool to investigate the impact of competing methods on consistency of inferred biological pathways.

2.2 Datasets Pairs

Raw data were downloaded from the NCBI Gene Expression Omnibus (GEO) website. Sample annotations were parsed from the sample description files or the description column contained in each GSM sample. The three dataset pairs used in our analysis are summarized in Table 2 and details are presented below.

Pair a - GSE6956 [25] and GSE17356 [26] were designed to investigate biological factors that predispose African American (AA) men to prostate cancer when compared to European American (EA) men. GSE6956 contained 89 samples from prostate tumor tissue samples (n=69) and non-tumor tissue samples (n=20). We used the array data of 69 tumor samples for our study. Samples in GSE17356 are primary prostate cancer epithelial cell cultures (n=27). Group1 are prostate cancer samples isolated from AA men. Group2 are samples isolated from EA men. Fifteen genes were shown to be differentially expressed between AA an EA prostate cancer patients in both studies (See Table IV in paper reporting GSE17356 [26]). The common scientific question is “Which genes are differentially expressed between AA and EA men with prostate cancer”.

Table 1. Summarization methods

Summarization Method	Author	Year	R Package	Discussed in Paper
Mean	-	-	-	yes
Median (Avgdiff)	Affymetrix	1999	expresso [17]	yes
MAS	Affymetrix	2002	expresso [17]	no
Median Polish	Irizarry RA et al	2003	expresso [17]	yes
Li-Wong	Li C, Wong WH	2001	expresso [17]	yes
playerout	Emmanuel. N.Lazaridis	2002	expresso [17]	no
Robust Linear Model (RLM)	Sboner A et al	2009	threestep (affyPLM)	yes

Pair b - GSE6532 [27] is a series with multiple data sources and platforms. It was designed in an effort to identify a gene classifier for predicting clinical prognosis of Tamoxifen-treated estrogen receptor positive (ER+) breast cancer patients. GSE6532 has a total of 741 samples (Supplementary Table 1). For comparative analysis we used 56 samples tested on U133A platform from the John Radcliffe Hospital (OXFT) and 81 samples from London, United Kingdom, Uppsala University Hospital (KIT). For both datasets, only ER+ breast cancer patients treated with Tamoxifen were used in our analysis. Group1 is defined as individuals with distant metastasis free survival (DMFS) <=3 years, and Group2 are those with DMFS>=5 years. The common scientific question is “In Tamoxifen-treated ER+ breast cancer patients, which genes are differentially expressed between individuals with DMFS <=3 and >=5 years”.

Pair c - GSE5460 [28] was designed to investigate the ability of global gene expression in primary breast tumors to predict receptor status, histological and other characteristics of the tumors. It contains 129 breast cancer samples from PLUS2 platform. GSE2109 is from expression project for oncology (expO) contributed by the International Genomics Consortium (IGC). A total of 2158 samples from roughly 100 tumor tissues are represented, of which 360 samples are from female breast cancer tissue. Since detailed phenotypic information is available for the two studies, we arbitrarily narrowed down sample phenotype to grade III ductal carcinoma to minimize the difference between pairing datasets. In the remaining part, ER+ samples were set as Group1 and estrogen receptor negative (ER-) samples as Group2. The common scientific question is “In grade III ductal carcinoma, which genes are differentially expressed between ER+ and ER- individuals”.

Table 2. Construction of comparing datasets

Datasets Pair	GSE number	Microarray Platform	Probe set Number	Group1 status	Group2 status	Group1 number	Group2 number
Pair a	GSE6956 GSE17356	HG-U133A 2.0	22277	AA[a]	EA[a]	34	35
						14	13
Pair b	GSE6532KIT GSE6532OXFT	HG-U133A	22283	ER+ & TAM DMFS<=3 [b]	ER+ & TAM DMFS>=5 [b]	21	35
						24	57
Pair c	GSE2109 GSE5460	HG-U133PLUS2 [c]	22283	ER-	ER+	65	48
						45	18

[a] African American and European American men with prostate cancer

[b] Tamoxifen (TAM) treated estrogen positive (ER+) breast cancer with distant metastasis free survival (DMFS) <=3 and >=5 years

[c] Plus2 is basically a combination of HG-U133A and HG-U133B. Only HG-U133A probe sets were extracted out from Plus2 for the analysis due to a large number of non-gene targeting probe sets in HG-U133B part.

2.3 Summarization Algorithms

Five summarization algorithms (mean, median, median polish, robust linear model (RLM), and Li-Wong) were compared in the R environment. A complete list of the processing steps is listed in Table 3. Two differential analysis methods (SAM and CyberT) were used to get p value (use default option). FDR was obtained by applying *q-value* [29] function with default options. The relevant software package was downloaded from the BioConductor website.

Median

The median value of probes in a probe set was used to represent summary expression level. The median method gives result same as the result by avgdiff approach provided in affymetrix built-in processing method [17] .

Mean

The mean value of probes in a probe set was used to represent summary expression level.

Median Polish

The model of median polish can be written as $T(PM_{ij}) = e_i + a_j + \varepsilon_{ij}$, where $T(PM_{ij})$ represents the measure after background correction, normalization, and log2 transformation of the PM intensity, e_i represents the log2 scale expression value found on array i, a_j represents the log scale affinity effects for probes j , and ε_{ij} represents random error. Implementation of median polish method is available in *expresso* function of R package *affy*.

Li-Wong

Li-Wong method has the following model: $MM_{ij} = v_j + \theta_i\alpha_j + \varepsilon_{ij}$, and $PM_{ij} = v_j + \theta_i\alpha_j + \theta_i\phi_j + \varepsilon_{ij}$. Here PM_{ij} and MM_{ij} denote the PM and MM intensity values for array i and probe pair $j$for this gene, v_j is the baseline response of probe pair j due to nonspecific hybridization, θ_i is expression index for the gene in array i, α_j is the rate of increase of the MM response of probe pair j, ϕ_j is the additional rate of increase in the corresponding PM response, and ε_{ij} represents random error. The rates of increase are assumed to be nonnegative. The model for individual probe responses can be written as $y_{ij=}PM_{ij} - MM_{ij} = \theta_i\phi_j + \varepsilon_{ij}$ .In the case of PM-only correction, $PM_{ij} - MM_{ij}$ is simply replaced by PM_{ij}. Implementation of Li-Wong method is available in *expresso* function of R package *affy*.

Robust Linear Model

The RLM method was developed by Hampel F.R.[19]. Use of RLM as summarization method was provided in *threestep* function of *affyPLM* R package (an extension of the base affy package).

2.4 Performance Metrics

Assume datasets A and B have N_a and N_b probe sets differentially expressed at significance level p_x . They share N_{both} probe sets in common. Among N_{both} probe sets, N_{diff} values have different fold change (FC) direction (i.e., the probe set is up-regulated in one dataset and down-regulated in another), and N_{same} have the same direction. The inconsistent FC proportion (IFP) and reproducibility are defined as following

Table 3. Processing flow for raw CEL file

	Background correction	Normalization	PM correction	Summarization	Differential Analysis Tool
Analysis Method	RMA[a]	Quantiles	PM-only	Median (avgdiff)	SAM CyberT
				Mean	
				Median Polish	
				Robust Linear Model (RLM)	
				Li-Wong (lw)	

[a] For RLM method, RMA2 background correction method is used (RMA is not available in *threestep* function and it is not easy to reproduce).

Table 4. Number of consistent pathways with p<0.05

Dataset pair	Platform	SAM					CyberT				
		median	mean	mp^{a}	RLM^{b}	lw^{c}	median	mean	mp^{a}	RLM^{b}	lw^{c}
GSE6956 VS. GSE17356	HG-U133A2	54	84	50	67	100	60	65	68	66	67
GSE6532OXFT VS. GSE6532KIT	HG-U133A	52	103	69	57	86	51	96	52	48	70
GSE2109 VS. GSE5460	HG-U133PLUS2^{d}	96	77	90	55	68	90	78	82	48	65

[a] Median Polish [b] Robust Linear Model [c] Li-Wong [d] Only probe sets from HG-U133A used

$$IFP = \frac{N_{diff}}{N_{both}} \qquad Reproducibility = \frac{N_{same}}{N_a + N_b - N_{same}}$$

2.5 Pathway Consistency and TAP-k Score Ranking

We fetched the top 1000 significant probe sets from each dataset and conducting pathway analysis using the GeneGo web tool (GeneGo Inc.). Pathways with P value less than 0.05 from the two comparing datasets were used for pathway consistency analysis.

Threshold Average Precision (TAP-k) [30], a metric used in bioinformatics area for comparing retrieval efficacy of different search engines, is used to measure pathway level consistency. To use TAP-k, a reference pathway database was constructed to represent the “true” pathways. In our study, a reference pathway is defined as those appeared >=3 times among the pathway consistency analysis results by using the five summarization methods. TAP-k score is used to rank summarization method based on concordance rate with the reference pathways.

3 RESULTS

3.1 Reproducibility and Inconsistent Fold Change Direction Proportion

Figure 1 shows the comparison result of the five summarization methods using two differential analysis tools. There are five plots for each pair to show the trend: IFP vs. N_{same} (the number of consistent probe sets), Reproducibility vs. N_{same} , IFP vs. p-value, Reproducibility vs. p-value, and N_{same} vs. p-value. In datasets pairs a and b, where HG-U133A2 and HG-U133A were respectively used, mean value summarization showed a constantly lower inconsistent fold change proportion (IFP) than competing methods (red line in Figure 1 a-1, a-3, b-1, b-3). The same tendency is observed when using either SAM (solid red line) or CyberT (dashed red line) as differential analysis tool. The reproducibility of mean strategy is comparable to other methods at different significance levels (Figure 1 a-4, b-4). Li-Wong summarization method produced more consistent probe sets when SAM is used (cyan line in Figure 1 a-5, b-5), but at the cost of high IFP (cyan line in Figure 1 a-3, b-3) and hence poor performance in the plot of IFP versus N_{same} (cyan line in Figure 1 a-1, b-1). Moreover, the performance of Li-Wong method is more sensitive to the two differential analysis strategies currently used. As indicated in Figures 1 a-5 and b-5 (cyan color), Li-Wong identified more consistent probe sets when SAM (solid line) is used, but this is not reproduced when applying CyberT (dashed line) method. Median summarization strategy performs worse in all the three dataset pairs we considered here.

Pair c has an overall low IFP (near zero when p<0.05) and high reproducibility. In Figure 1 c-5, RLM (Blue) and Li-Wong (Cyan) methods identified more consistent probe sets than other methods when same p value cutoff standard is used. However, when plotting reproducibility vs. N_{same}, we see slightly better performance of median polish (Green) and mean (Red) methods (Figure 1 c-2). All summarization methods have IFP near to zero when N_{same} is less than 1000 (Figure 1 c-1).

3.2 Genego Pathway Consistency Analysis

The reference pathways constructed in the TAP-k score ranking test of each dataset pairs were provided as supplementary materials. The performance of five summarization methods ranked by TAP-k score is illustrated in Figure 2 a, b, c.

GeneGo pathway consistency analysis showed largely variable performance of competing methods depending on both the comparing dataset and differential analysis method used. In general, our analysis shows mean and Li-Wong methods have better performance in identifying more consistent pathways on pairs a and b. In pair c, median and median polish has the best performance (Table 4).

Mean method (Red) ranked first or second in three dataset pairs and its performance is more stable than competing methods. RLM (Blue) and Li-Wong (Cyan) have high TAP-k score in pair a, but the performance is not reproduced in pair c. No obvious alteration in ranking was observed between SAM and CyberT.

4 DISCUSSION

The comparison study of Kerby Shedden [16] based on one ovary tumor dataset and one colon tumor dataset (both used HG-U133A platform) showed that Trimmed mean and Li-Wong methods are more sensitive---detect more genes at a given FDR level. However, the number of significantly differentially expressed genes detected at given FDR level highly depends on the differential analysis algorithm used. Li-Wong strategy by SAM returned nearly double number of probe sets at a given significance level (same when FDR is used) than by CyberT (Supplementary Table 2). Moreover, certain truncation (in the manner recommended by the developers of each method) was implemented in Kerby Shedden’s

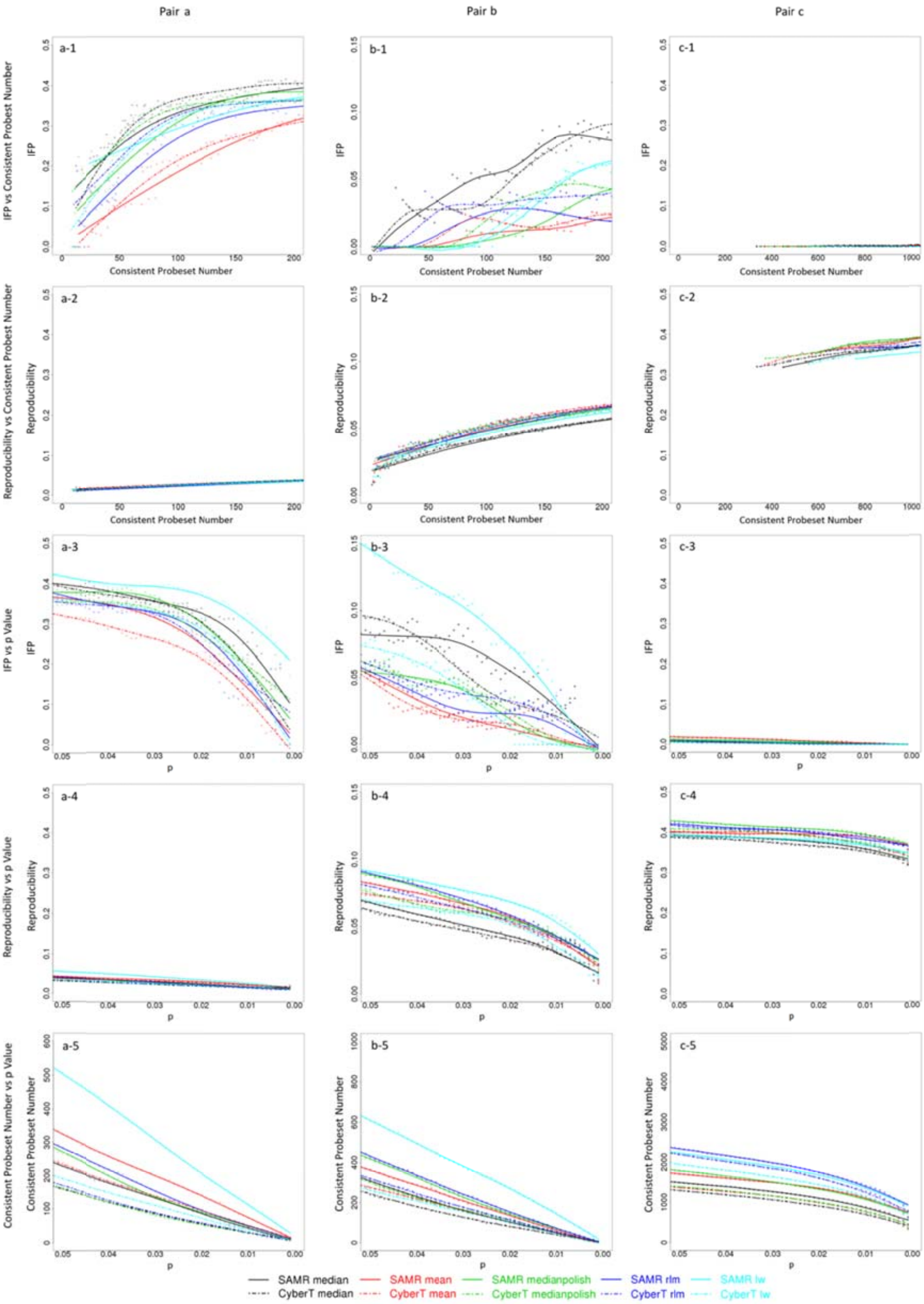

Fig.1. Performance plots for competing summarization methods on three dataset pairs: a) GSE6956 and GSE17356, b) GSE6532KIT and GSE6532OXFT, c) GSE2109 and GSE4560. Solid and dashed lines are results from SAM and CyberT algorithms respectively. Black, red, green, blue, cyan colors are results from median, mean, median polish, RLM, Li-Wong summarization methods respectively.

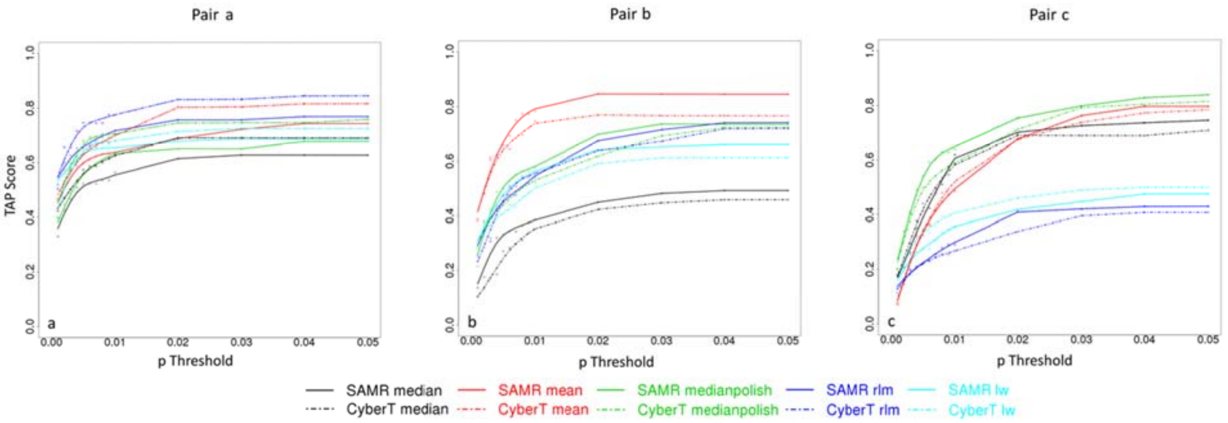

Fig.2. TAP-k score ranking of five summarization methods at different p value threshold: a) GSE6956 and GSE17356, b) GSE6532KIT and GSE6532OXFT, c) GSE2109 and GSE4560. Solid and dashed lines are results from SAM and CyberT algorithms respectively. Black, red, green, blue, cyan colors are results from median, mean, median polish, RLM, Li-Wong summarization methods respectively.

comparison. It is unknown how the truncation may affect the outcome. Rafael A. Irizarry [4] studied the performance of a panel of pre-processing algorithms for Affymetrix using spike-in data. However, it is not clear how those algorithms perform for real biological question. With these potential pitfalls in view, we based the comparison on microarray dataset pairs from research results, and evaluate the performance by checking IFP and reproducibility. One advantage of our approach is that the confounding factors causing a transcript to be falsely called significant in one dataset are unlikely to entirely reappear on the same transcript at another dataset. Partial reappearance of the confounding factors might not be strong enough to constitute a false positive call. Accordingly, results that can be verified in two (or more) datasets with the same study aims are more likely to be true positive. We consider that algorithms generate better consistency among real datasets may convey a better representation of the true expression level. On the other hand, we did not truncate any data in the whole preprocessing flow to avoid unwanted bias towards specific methods. After taking into account these sources of variation, the present study is more likely to reflect the true performance of the competing methods.

Among the five methods studied here, the median method performs the worst in reproducibility. There is no clear winner with respect to reproducibility. The Li-Wong method indeed shows slightly higher reproducibility in pairs a and b (Figure 1 a-4 and b-4) than others under the same p cutoff value. It is comparable to other methods in the plot of reproducibility versus consistent probe set number. This means in one dataset, in order to get an equal number of probe sets reproduced as other competing methods, Li-Wong method needs to call the same amount of probe sets significant in the pairing dataset. Moreover, the Li-Wong method is among the highest inconsistent fold change under the same p cutoff value. The consequence is that among a large number of calls with significance, only a small proportion can be reproduced with consistent FC direction. This on the other hand indicates that evaluating the performance of one method by merely checking the number of significant probe sets calls is not appropriate. Interestingly, different from our initial speculation that mean value may suffer from high IFP due to its vulnerability to outliers, it outperformed competing methods mainly by lowering IFP on datasets pairs a and b. It is possible that RMA background correction and quantile normalization steps have already excluded potential outliers, and thus further outlier-oriented adjustment is not necessary. Implication is that certain probe sets with the potential to give inconsistent FC direction were not called significant under this strategy. At present, we cannot safely say this controlled FDR without biological evidence that they were not significantly altered by disease status. This strategy, however, indeed renders researches with same study aim more consistency. As indicated in the GeneGo pathway consistency analysis, the mean strategy gives a consistent pathway number ranged first or second in pairs a and b. Its stable performance in TAP-k score ranking indicates its potential to give estimated pathways closer to the reference set. Interestingly, when we compare these five strategies on spike-in dataset proposed by Leslie M. Cope [31] on HG-U133A platform, the performance of median polish and RLM ranked 1st while mean and Li-Wong ranked 15 and 19, which is not in agreement with their performance in pathway consistency analysis. This implies that spike in study may not provide an accurate view of how methods may perform in reality. It might be affected by sources of systematic variation and it is not clear how this might affect evaluation of different data extraction methods.

Plus2 is a combination of the probe sets from HG-U133A and HG-U133B and have probe sets number about twice the size of single A or B platforms. Considering the fact that HG-U133B has nearly eight thousands probe sets with no corresponding gene target, and that a considerable number of the remaining part target the same genes as platform HG-U133A, we only used the probe sets that were covered by HG-U133A when analyzing pair c. This also helps to make the comparison with the other two pairs consistent. The datasets in pair c showed excellent performance in both lower IFP and higher reproducibility than the other two pairs. We also observed much more number of consistent probe sets in pair c. This might be resulted from the large biological difference between ER- and ER+ breast cancers [28]. Thus the differentially expressed transcripts are more easily identifiable. Mean strategy only has slightly better performance (comparable to median polish) in plot of reproducibility versus N_{same}. Its performance in other plot and pathway consistency analysis is not superior to competing methods. A possible explanation is that in situations where obvious biological differences exist, the consistency is less affected by the summarization methods used.

It is intriguing that mean summarization, a remarkably simple algorithm with the lowest time complexity, outperform (dataset pairs a and b) or comparable to (pair c) several competing algorithms. Similar argument can be found in the 70-gene signature for breast cancer prognosis classification developed by Van't Veer *et al.* [32]. The group sorted the differentially expressed genes between relapsed and relapse free breast cancer patients by p value and picked the top 70 most significant genes, and used the mean expression levels of these 70 genes in relapse free group as the signature. This simple strategy has not yet been outperformed by other more sophisticated strategies [33]. A possible explanation is that complex algorithms with too specific kinds of adjustment result in “fit to noise” under circumstances where high background noise exists. Methods such as Li-Wong iteratively fit a model to the probe data from multiple microarrays to exclude outliers. These iterations may cause signal distortion. It might help to increase the reproducibility of “disease-caused” differentially expressed transcripts, but at the cost of high proportion of inconsistent results.

Note that we used p value rather than FDR as cutoff standard because different datasets generate very different number of probe sets at the same FDR level. Pair b has actually no common probe sets when set FDR to <0.1. Additionally, the p values obtained from SAM and CyberT are based on regularized t test (by using adjusted variance). We thus use p value to do the comparison while similar results were obtained when using FDR as cutoff standard (obtained by *qvalue* algorithm [29]).

5 CONCLUSION

In the present work, we compared the performance of five summarization algorithms on their ability to lower IFP and improve reproducibility. While maintaining comparable reproducibility, mean summarization strategy gives smaller proportion of probe sets with inconsistent FC direction in two datasets pairs than several currently widely used summarization approaches. Its performance is weakened in the paired datasets where high biological difference may exist between comparison groups.

ACKNOWLEDGEMENT

The research is supported by the following grant: NSF 0845523, NIH R01LM009959A1, and DOD PC081609.

Reference

1. M. Schena, D. Shalon, R. W. Davis, P. O. Brown, *Science* **270**, 467 (1995).
2. E. M. Southern, *Methods Mol. Biol.* **170**, 1 (2001).
3. R. A. Irizarry *et al.*, *Biostatistics.* **4**, 249 (2003).
4. R. A. Irizarry, Z. Wu, H. A. Jaffee, *Bioinformatics.* **22**, 789 2006).
5. B. Harr, C. Schlotterer, *Nucleic Acids Res.* **34**, e8 (2006).
6. F. Naef, C. R. Hacker, N. Patil, M. Magnasco, *Genome Biol.* **3**, RESEARCH0018 (2002).
7. Z. Wu, R. A. Irizarry, *Nat. Biotechnol.* **22**, 656 (2004).
8. B. M. Bolstad, R. A. Irizarry, M. Astrand, T. P. Speed, *Bioinformatics.* **19**, 185 (2003).
9. N. Jiang *et al.*, *BMC. Bioinformatics.* **9**, 284 (2008).
10. R. A. Irizarry *et al.*, *Nucleic Acids Res.* **31**, e15 (2003).
11. D. Rajagopalan, *Bioinformatics.* **19**, 1469 (2003).
12. *Nat. Biotechnol.* **24**, 1039 (2006).
13. R. D. Canales *et al.*, *Nat. Biotechnol.* **24**, 1115 (2006).
14. L. Shi *et al.*, *Nat. Biotechnol.* **24**, 1151 (2006).
15. R. Shippy *et al.*, *Nat. Biotechnol.* **24**, 1123 (2006).
16. K. Shedden *et al.*, *BMC. Bioinformatics.* **6**, 26 (2005).
17. Ben Bolstad, *Affymetrix* (2010).
18. A. Sboner *et al.*, *J. Proteome. Res.* **8**, 5451 (2009).
19. F. R. Hampel, E. M. Onchetti, P. J. Rousseeuw, W. A. Stahel, *Robust Statistics:The Approach Based on Influence Functions* (John Wiley and Sons, New York, NY, 1986).
20. C. Li, W. H. Wong, *Proc. Natl. Acad. Sci. U. S. A* **98**, 31 (2001).
21. Affymetrix, *Technical report, Affymetrix* (2002).
22. E. N. Lazaridis, D. Sinibaldi, G. Bloom, S. Mane, R. Jove, *Math. Biosci.* **176**, 53 (2002).
23. V. G. Tusher, R. Tibshirani, G. Chu, *Proc. Natl. Acad. Sci. U. S. A* **98**, 5116 (2001).
24. P. Baldi, A. D. Long, *Bioinformatics.* **17**, 509 (2001).
25. T. A. Wallace *et al.*, *Cancer Res.* **68**, 927 (2008).
26. O. A. Timofeeva *et al.*, *Int. J. Oncol.* **35**, 751 (2009).
27. S. Loi *et al.*, *BMC. Genomics* **9**, 239 (2008).
28. X. Lu *et al.*, *Breast Cancer Res. Treat.* **108**, 191 (2008).
29. J. D. Storey, R. Tibshirani, *Proc. Natl. Acad. Sci. U. S. A* **100**, 9440 (2003).
30. H. D. Carroll, M. G. Kann, S. L. Sheetlin, J. L. Spouge, *Bioinformatics.* **26**, 1708 (2010).
31. L. M. Cope, R. A. Irizarry, H. A. Jaffee, Z. Wu, T. P. Speed, *Bioinformatics.* **20**, 323 (2004).
32. Van, V *et al.*, *Nature* **415**, 530 (2002).
33. B. Haibe-Kains, C. Desmedt, C. Sotiriou, G. Bontempi, *Bioinformatics.* **24**, 2200 (2008).