A computational approach to microarray universal reference sample

Georgia Tsiliki, Sofia Kaforou, Manouela Kapsetaki, George Potamias and Dimitris Kafetzopoulos

Abstract— **The choice of a reference material in two-colour microarray experiments is an important issue of the experimental design. We consider a number of cell lines derived from a variety of primary tissues in order to construct a reference material. The aim of the present study is to understand the correlation structure of the data and develop a common reference that would enable comparison of expression levels in multiple microarray hybridizations with high coverage. We analyze 22 cultured cell line samples using a common reference pool. After estimating the coverage for each cell line or combinations of cell lines based on methods suggested by the literature, we employ stochastic optimization techniques to estimate the optimal set of cell lines in the reference sample. We found that only a subset of cell lines is necessary to achieve coverage as high or higher than that of the original reference sample used. We tested experimentally the performance of the new reference sample suggested and found that its coverage outperforms the coverage achieved by Stratagene's human reference sample for the particular platform.**

I. INTRODUCTION

Comparisons of expression levels across different two-colour microarray experiments and identification of informative patterns of their gene expression are facilitated when a common reference is co-hybridized to every microarray [14]. Then, the power of microarray analysis can be increased depending on the *coverage* of genes spotted on the array, where coverage or *microarray coverage* is defined as the percentage of spots with hybridization signal above a user defined threshold [5]. To provide optimal coverage of genes spotted on the array, common reference samples are often generated from RNA derived from various cell lines.

Initially, one sample originating from one cell line, or time point zero, was used as a common reference [10]. A disadvantage of this approach is that the control sample does not provide signal in all spots. To overcome this drawback, a pool reference sample originating from diverse cell lines was suggested, which would exhibit a more complete gene representation ([6],[14]). Although, cell culturing can be time and space consuming, this approach is the most popular. However, coverage strongly depends on the structure of the sample. A commercially available pool reference sample commonly used in two-color microarray experiments is the Universal Human RNA Reference (UHRR) from *Stratagene* company (http://www.stratagene.com). Another possible reference sample would consists of parts of RNA from all experimental samples [12]. This pool reference is experiment specific and requires large amount of reference pool which might not be feasible in case of limited experimental samples.

According to Sterrenburg et al.(2002) a common reference for DNA microarrays would consist of a mix of the products that were spotted on the array [10]. Their Polymerase Chain Reaction (PCR) reference was generated by pooling a fraction of all amplified probes prior to printing. A very low number of spots could not be analyzed because reference, as opposed to target, did not give a significant signal. Dye-swap experiment were conducted to test reference reproducibility which yielded a reproducible hybridization signal in 99.5% of the microarray platform content.

Yang et al.(2002) suggested the use of a limited number of cell lines, each expressing a large number of diverse genes. Particularly, they constructed two reference pools from those cell lines with the greatest representation of unique genes, which are also easy to grow and yield high quantities of RNA. In their first reference pool they mixed equal amounts of colon cell lines $CaCO2$, $KM12L4A$ and the ovarian cell line $OVCAR3$, whereas in their second pool $OVCAR3$ was replaced with the brain cell line $U118MG$. They found that adding more cell lines to the pool would not necessarily improve the overall gene representation because some genes were diluted below the detection limit. The first reference sample exhibited similar coverage with Statagene's UHRR sample (75%), and the second reference sample had coverage equal to 80%. The observed difference in the coverage percentages may be due to the fact that brain exhibits the greatest diversity of transcripts or that the subset of genes expressed in brain is more disjoint with the genes observed in colon than in ovary. Thus, according to Yang et al.(2002), a simple pool of RNA from diverse cell lines can provide a superior reference.

Human, mouse and rat reference RNA samples were considered by Novoradovskaya et al.(2004), which they referred to as *Universal Reference RNA* (URR) and were prepared from pools of RNA derived from individual cell lines representing different tissues. Specifically, each of the three URR consisted of ten human, eleven mouse and fourteen rat cell lines, respectively. They evaluated microarray coverage based on a pre-specified threshold equal to the background intensity or twice the background intensity of each channel, using different microarray platforms. Probes with intensities above threshold were characterized as *present*. They reported

This work was partially supported by grants to *AKMON* (*AP*.6260 EFA1250/17 – 5 – 2004) co-funded from the European Regional Development Fund by 70% and from the Hellenic Ministry of Development by 30% through the Operational Programmae Competitiveness, *PEP* (*KR*8/16/6/2006) and Crete Innovation Pole (11 – *PPK* – 06).

G.Tsiliki, S.Kaforou, M.Kapsetaki and D.Kafetzopoulos are with FORTH, Institute of Molecular Biology & Biotechnology, P.O. Box 1385, 711 10, Heraklion, Greece
tsiliki@imbb.forth.gr, kaforou@imbb.forth.gr

G. Potamias is with FORTH, Institute of Computer Science, P.O. Box 1385, 711 10, Heraklion, Greece

microarray coverage greater than 80% for all arrays tested when threshold was equal to background, and greater than 60% when threshold was equal to twice the background. Consequently, they agreed with Yang et al.(2002) that pools of RNA derived from a limited but diverse set of cell lines resulted in an optimal reference sample.

In summary, some important prerequisites for common reference samples are: significant signal for each spot, large quantities sufficient to satisfy longterm requirements of many researchers and reproducibility in a way that different batches would be indistinguishable from one another.

A common reference sample accomplishing the above criteria would provide an internal standard against which gene expressions of each experimental sample can be compared [6]. In this paper we study the structure of the data and examine whether small number of samples with a differential representation of expressed genes can outperform more complex mixtures. A computational approach to estimate the optimal cell line set is described.

II. Material and Methods

A. Cell line hybridizations

A Qiagen human library of 70mer probes (www1.qiagen.com) was used, which consists of $34,772$ probes representing $24,650$ genes and $37,123$ gene transcripts. The oligos were printed in duplicates onto aminosilane glass slides activated with PDITC. The number of cells per millilitre (ml) was determined before freezing, and aliquots of five million cells were completely lysed by passing the lysate five times through a 20-gauge needle (0.9 millimetre (mm) diameter) fitted to a sterile syringe. Total RNA was purified on Rneasy columns according to manufacture's instructions. $Cy5$ labelled mRNA from each cell line was co-hybridized with equal amount of $Cy3$ labelled reference mRNA mix, where the latter consists of equimolar quantities of RNA from a variety of cell lines.

We tested 22 assays specific to 22 human carcinoma cell lines, where their abbreviations and origin are shown in Table I. The third column of the table includes abbreviations specific to the medium used for cell line culturing, which was supplemented with 50μg/ml gentamycin in all cases. Those 22 cell lines were also included in the reference sample considered here. Three mixes with different combinations of external RNA controls were spiked into RNA samples. Labelling efficiency and quantity of labelled RNA was determined with the spectrophotometer Nanodrop 3.0.1. Arrays were then scanned using a GSI Lumonics ScanArray5000.

Fluorescent intensities of Cy5, Cy3 channels on each slide were subjected to spot filtering and normalization. Particularly, we employed *print-tip lowess* normalization method to remove intensity dependent dye-specific and spatial effects [9]. While lowess normalization greatly reduces dye-specific artifacts that often appear for low or high (saturated) intensity data points, the data exhibit additional structure that can be used to evaluate patterns of gene expression.

TABLE I
Cell line list

Cell lines	Derivation	Culturing Conditions
1. HL60	Bone marrow	RPMI 1640, 10% FBS
2. Hs578T	Mammary Gland	DMEM, 15% FBS
3. McF7	Mammary Gland	DMEM, 10% FBS
4. OVCAR3	Ovary	RPMI 1640, 10% FBS
5. Panc1	Pancreas	RPMI 1640, 10% FBS
6. SKMEL3	Skin	McCoy's 5A, 10% FBS
7. SKMM2	Bone Marrow	RPMI 1640, 10% FBS
8. T47D	Mammary Gland	RPMI 1640, 10% FBS
9. TERA1	Testis	McCoy's 5A, 10% FBS
10. U87MG	Brain	DMEM, 10% FBS
11. Raji	B-lymphoblasts	RPMI 1640, 10% FBS
12. JAR	Genital	DMEM, 10% FBS
13. Saos2	Bone	DMEM, 10% FBS
14. SW872	Liposarcoma	Leibovitch L-15, 10% FBS
15. THP1	Peripheral Blood	RPMI 1640, 10% FBS
16. HCT116	Colon	McCoy's 5A, 10% FBS
17. HUVEC	Umbelical Vein	EBM-2, 10% FBS
18. HepG2	Liver	DMEM, 10% FBS
19. HeLa	Cervix	DMEM, 10% FBS
20. LNCap	Prostate	RPMI 1640, 15% FBS
21. Molt4	T-lymphoblasts	RPMI 1640, 10% FBS
22. WERI1	Retina	RPMI 1640, 10% FBS

B. Structure in the data set

Only for this section and in order to estimate the significant genes and study the correlation structure between cell lines, we consider the M values for the probes, where $M = \log_2(Cy5) - \log_2(Cy3)$. We first impute missing values in a similar fashion as the weighted K-nearest neighbors (KNN) algorithm suggested by Troyanskaya et al.(2001), but instead of the Euclidean distance between the 10 nearest neighbours of the missing value, we compute Pearson's correlation coefficient between the sample with the missing value and its nearest 10 neighbours. For the rest of the analysis we consider the data with the imputed values.

We employ a bootstrapping method to select the genes which have M values significantly different from zero across all 22 cell lines. The bootstrap method assigns measures of accuracy to sample estimates [3], which allow us to have a value of variability for the data, without destroying the observed structure. Particularly, we generated $1,000$ bootstrap samples each one of which consists of $N = 22$ cell lines sampled with replacement from the original population, and $n = 34,771$ probes (one probe which had only one non-missing value was excluded). For each gene we use one-sample Student t-test to test the null hypothesis of $E(M) = 0$ and keep only those genes with mean M value significantly different from zero at a 1% significance level. The distribution of the M values is well approximated by the standard normal distribution, hence the assumption of normality for the parametric t-test is not violated. We select 124 probes that appear to be significant in at least 50% of the bootstrap samples considered, after controlling *False Discovery Rate* (FDR) with *Benjamini and Yekutieli* (BY) correction [1].

As a first attempt to understand variation in the sample,

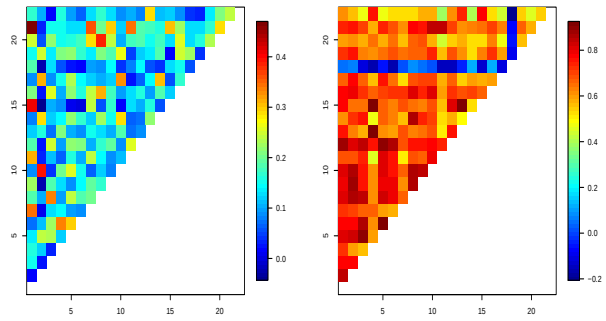

Fig. 1. **Pearson's pairwise correlation coefficients between the** 22 **cell lines.** All probes (left) or only significant probes (right) are considered. Each of the 231 coloured squares in the plots correspond to a pairwise coefficient.

we test for association between the 22 cell lines considered. Fig. 1 shows Pearson's pairwise correlation coefficients between cell lines. The left hand-side plot shows correlation structure between cell lines for all $34,771$ probes, whereas the right hand-side plot for only the 124 probes found significant. In the first plot the pairwise coefficients exhibit small values which vary in $[0,0.5]$, while in the second plot more evident correlation patterns appear. Particularly, there are some positive correlation coefficients at the bottom of the plot which correspond to the pairwise coefficients of the first six or even seven cell lines of Table I. Also, row 18, which corresponds to liver $HepG2$ cell line, appears to be uncorrelated with the remaining cell lines.

We apply hierarchical clustering algorithm to expression ratios using Euclidean distance metric and Ward clustering algorithm [13]. A clustered image map which relates genes and cell lines is shown in the first graph of Fig.2. The graph is accompanied by two dendograms specific to genes and cell lines. We can observe some pattern between groups of cell lines for the set of significant probes which reflects the connection of those probes with the specific cell lines. Particularly, cell lines appear to group into approximately three clusters.

In the second plot of Fig.2 we test the robustness of the hierarchical relationship between cell lines, as this was suggested by the first plot of Fig.2. For this reason, we apply the multiclass bootstrap resampling technique suggested by [8] and check how strongly the estimated clusters are supported by the data. In brief, their algorithm generates $10,000$ bootstrap samples, performs hierarchical cluster analysis for each sample and reports empirical p-values (%) for each of the initially estimated clusters which suggest how likely we are to observe those clusters. Red values correspond to *approximate unbiased* (AU) p-values, and green values to ordinary bootstrap p-values (BP). Bias corrected p-values are calculated from the slope of the regression curve applied to bootstrap probabilities. Clusters with $AU \geq 0.95$ are strongly supported by the data and are highlighted by rectangles in Fig.2. We can observe two significant clusters according to bootstrapping

TABLE II

CATEGORIZATION OF PROBES

Category	Number of probes (%)
common	11,540 (33.19)
unique	2,679 (7.76)
shared	11,721 (33.71)
non-expressed	8,813 (25.34)

analysis, which include all cell lines except for $HepG2$ and $WERI1$. Particularly, $A_1 = \{OVCAR3, Saos2, THP1\}$ cluster was observed in all bootstrap samples $(AUp-value = 1)$ and $A_2 = \{HL60, Hs578T, McF7, Panc1, SKMEL3, SKMM2, T47D, TERA1, U87MG, Raji, JAR, SW872, HCT116, HUVEC, HeLa, LNCap, Molt4\}$ was observed in 97 out of 100 bootstrap samples. A_2 cluster contains all three mammarian gland cell lines.

Thus, most of the cell lines are grouped together in A_2 cluster indicating their similarity in terms of ratio intensities, which might also indicate their ability to produce signal in the same group of probes spotted on the array. For this reason we examine reducing the number of cell lines involved in the pool reference sample and compare their coverage with that of the reference sample used. In part of our analysis we use information from cell line clustering.

III. REFERENCE POOL SELECTION

A. Setting an absolute intensity cutoff value

For the following analysis we focus on $Cy5$ intensities of the probes in order to estimate the array coverage of the reference sample which involves only one of the channels. We can categorize probes based on a pre-specified threshold C_0 for the $Cy5$ intensities, after adjusting their distribution across arrays ([14], [5]). If we choose this threshold to be equal to the mean $Cy5$ intensity across cell lines and in particular $C_0 = 350$, then probes can be categorized as shown in Table II. With the term common we refer to those probes with $Cy5 \geq C_0$ in all cell lines, the term unique refers to probes with $Cy5 \geq C_0$ in just one cell line, the term shared refers to probes with $Cy5 \geq C_0$ for more than one, but not all, cell lines, and the rest of the probes are denoted as non-expressed. The percentage of missing values per cell line varies from 0.5% to 3.7%, however, here we categorize absolute intensities after we imputed missing values as explained in section II. The large percentage of non-expressed probes can be partly explained from the strict C_0 used, for example Novoradovskaya et al.(2002) used a less strict threshold. Also the percentage of commonly expressed genes is very similar to that of shared genes, which means that we observe signal in most of the cases. We exclude common probes from the analysis presented in section III-B, because, since we are only interested in whether a probe exceeds C_0 and not by what extent, thus they would not affect our analysis. However, all probes are considered in section III-C.

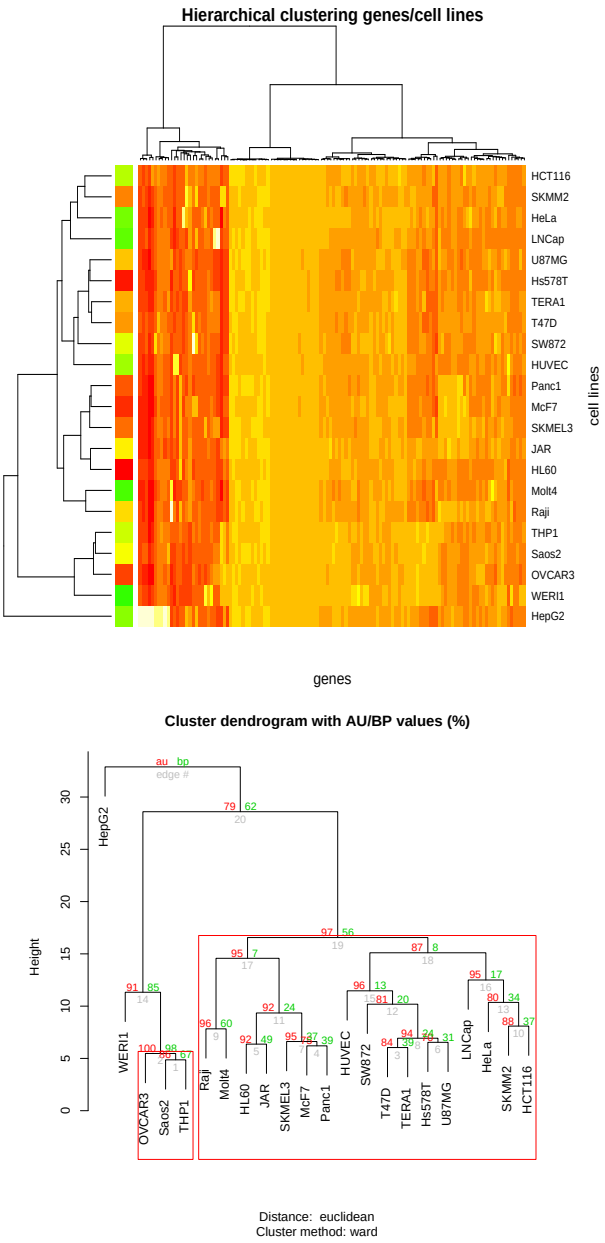

Fig. 2. The first plot presents a coloured representation of the data table (heatmap), with the rows (probes) and columns (cell lines) in cluster order. The colour in each cell of this table reflects the mean-adjusted expression level of the gene and cell line. Red corresponds to low values and yellow to high values. In the second plot the dendrogram represents the hierarchical relationships between clusters and how those clusters are supported by the data. Red and green numbers in the edges correspond to AU and BP p-values, respectively.

B. Cell lines coverage using a cutoff value

We define the coverage of cell line k which we denote by Ω_k, $k \in S = \{1,2,...,22\}$, as

$$\Omega_k = \frac{\sum_{i=1}^{n} I_{ik}}{n} \qquad (1)$$

where I_{ik} is a binary variable, with $I_{ik} = 1$, if probe i intensity of the kth cell line exceeds C_0, and 0 otherwise. We found coverages to vary in the closed interval $[50.84, 53.67]$, where the minimum coverage corresponds to $Molt4$ cell line and $WERI1$ achieves the maximum coverage.

The estimated coverage is not adequately high given that in a microarray experiment the researcher expects to receive a detectable signal for most probes of the array. For that reason we test whether a subgroup of the available cell lines would exhibit better coverage percentage with respect to their between arrays adjusted intensities. In this case all possible combinations of the 22 cell lines are considered. As an example, if $k = 2$ we consider all 231 pairwise combinations of the 22 cell lines. Let us denote by $\mathbf{s}_m^k = \{s_{m(1)}^k, ..., s_{m(k)}^k\}$ a vector which consists of indexes for the k cell lines sampled without replacement from S and $m \in [1, ..., \frac{k(k-1)}{2}]$. Here, the estimate coverage $\Omega'_{\mathbf{s}_m^k}$ is given by

$$\Omega'_{\mathbf{s}_m^k} = \frac{\sum_{j=s_{m(1)}^k}^{s_{m(k)}^k} \sum_{i=1}^{n} I_{ij}}{kn} \qquad (2)$$

where I_{ij} is the binary variable defined above. Our aim is to find the optimal number k and the particular content of $\mathbf{s}_m^k$ which would achieve coverage percentage higher than 53.63%. In Fig.3 we can observe the maximum estimated coverage percentage per combination, i.e. for each $k \in S$ and all possible $\mathbf{s}_m^k$ groups. As can be seen from the graph, coverage varies from 55.1% to 57.12% and achieves its maximum when $k = 11$ (green vertical line), whereas, when $k = 22$ the estimated coverage equals 55.85%. The optimal pool sample for $k = 11$ is $\{Hs578T, OVCAR3, Panc1, TERA1, Raji, Saos2, THP1, HUVEC, HepG2, LNCap, WERI1\}$. Fig. 4 shows the individual coverage percentages per cell line participant in that optimal mix. Although, those vary in $[51.68, 53.67]$, when considered together they succeed a coverage of 57.12%.

C. Estimation of optimal cell line sample without employing cutoff value

In the previous section we considered an exhaustive search along all possible combinations of cell lines in the data. Here we consider algorithms to minimize the computational cost of section III-B. An important advantage of the analysis in this section is that we use absolute $Cy5$ intensities without imposing a cutoff value.

In particular, one approach would be to order genes in descending order for each cell line based on the magnitude of their absolute expression values, and assign a significance level to each cell line given the number of genes that achieve their maximum value in that cell line. Then,

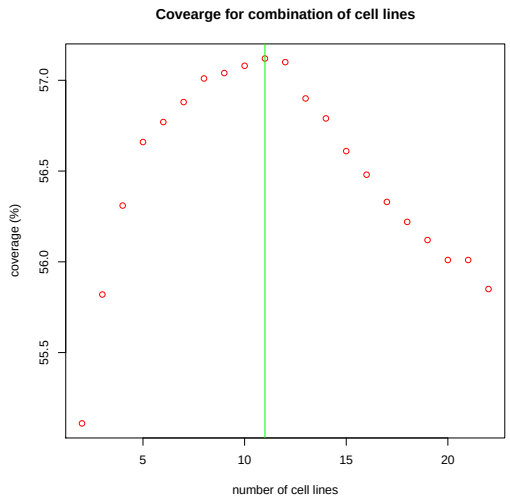

Fig. 3. Maximum coverage percentages for $k \in S$. The green vertical line corresponds to the peak of the curve for the estimated optimal mix of $k = 11$.

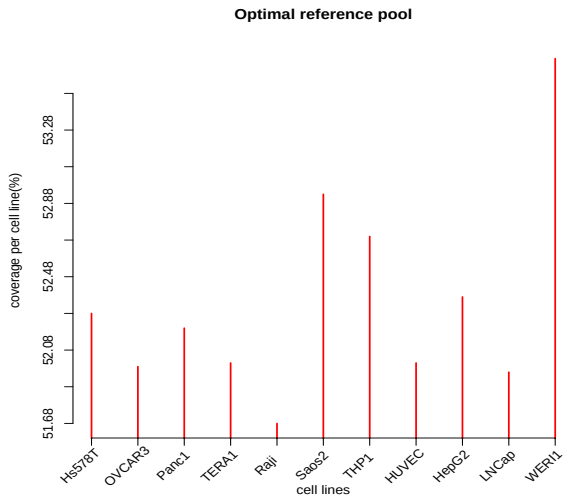

Fig. 4. The individual coverage percentages for the 11 cell lines which were selected for the optimal reference pool.

adding one cell line at a time from their ordered list and calculating $\Omega'_{\mathbf{s}_m^k}$, where $k \in S$ now refers to the ordered list of cell lines, would give the coverage for each specific set of cell lines. We found the ordered cell line list to be $\{WERI1, THP1, Raji, TERA1, HepG2, Hs578T, OVCAR3, HUVEC, Panc1, U87MG, LNCap, JAR, McF7, SKMM2, Saos2, Molt4, T47D, HL60, SW872, HeLa, HCT116, SKMEL3\}$. The optimal mix in this case is the subset of the first nine cell lines of the previous ordering with coverage percentage 57.03%, which is very close to 57.12% found from the exhaustive search.

Additionally, we consider pool reference sample selection as a variable selection problem. Our aim is to identify a subset of the original variables that can approximate the whole data set without using a threshold value on the absolute intensities, but rather rely on their underlying correlation. We then compare our findings with those of section III-B. Specifically, we employ *Simulated Annealing* (SA) algorithm [4] on the correlation matrix of the absolute intensity values when all 34,771 genes and 22 cell lines are considered. SA is an iterative stochastic optimization method which is often employed when the function h, we wish to maximize has many local maxima. SA performs random walks along the problem space $S' = \{\mathbf{s}^k_{m(k)}, m \in [1,...,\frac{k(k-1)}{2}], k \in [1,22]\}$, with probability depending on a *temperature function* T. In this setup, let $\theta \in S'$ be a subset of k cell lines, known as the current state, and $\theta' \in S'$ the proposed subset or state, which differs from θ by a single variable. SA accepts the proposed state with probability α equal to the Metropolis function

$$\alpha = \min\{1, exp(\frac{h(\theta') - h(\theta)}{T})\} \quad (3)$$

We consider three alternatives for h and in particular three of the criteria suggested by Cadima et al.(2004) in the content of *Principal component analysis* (PCA), i.e. *RM coefficient*, *RV coefficient* and *Yanai's generalised coefficient of determination* (GCD), which are implemented in *subselect* package of R software (http://www.r-project.org/). All three criteria measure the similarity between the correlation matrix of the data and a sub-matrix defined by θ' say, and vary in the closed interval $[0,1]$. Briefly, they are defined as

$$RM = \{\frac{\sum_{i=1}^N \lambda_i r_i^2}{\sum_{i=1}^N \lambda_i}\}^{\frac{1}{2}} \quad (4)$$

$$RV = \{\frac{\sum_{i=1}^N \lambda_i^2}{trace(V^2)}\} \quad (5)$$

$$GCD = \frac{1}{k}\sum_{i=1}^k r_i^2 \quad (6)$$

where V is the variance-covariance matrix of the data, λ_i are the principal components variances and r_i^2 are the correlations between the data matrix and the sub-matrix defined by θ'.

We set the number of iterations for SA equal to 10^6, the initial temperature is $T = 1$ and decreases in each iteration with a rate of 0.05, which allows for faster moves in the surface of function h. The algorithm finds the best set of cell lines for each k, and we then select a specific configuration with coefficient 0.9 to control for parsimonious models.

Additionally, we consider grouping cell lines based on results from the bootstrapping analysis as this is shown in Fig. 2. Namely, we focus our search on estimating the cell lines that can best describe the wider A_2 cluster and reduce its dimension. Cell lines outside A_2 are considered important.

The results of our analysis are shown in Table III, where by *Cov* we denote coverage analysis presented in section III-A, by *Ord* the analysis based on the ordering of cell lines given their absolute intensity values, and by $2g$ we denote results of the SA algorithm when it was only employed to cluster A_2 with RM coefficient. The rest of the columns correspond to SA results when the criterion specified by the column's name is used. We can observe that all methods except from *RV* suggest similar sets, and they all agree in the importance of

TABLE III

REFERENCE CELL LINE POOL

Cell line	*Cov*	*Ord*	RM	RV	GCD	2*g*
1. HL60					✓	
2. Hs578T	✓	✓	✓			✓
3. McF7			✓		✓	✓
4. OVCAR3	✓	✓	✓		✓	✓
5. Panc1	✓	✓	✓		✓	✓
6. SKMEL3						
7. SKMM2			✓		✓	
8. T47D						
9. TERA1	✓	✓	✓		✓	✓
10. U87MG						
11. Raji	✓	✓	✓		✓	✓
12. JAR				✓	✓	
13. Saos2	✓				✓	✓
14. SW872				✓	✓	
15. THP1	✓	✓		✓		✓
16. HCT116				✓		
17. HUVEC	✓	✓	✓		✓	✓
18. HepG2	✓	✓	✓		✓	✓
19. HeLa				✓		
20. LNCap	✓		✓		✓	✓
21. Molt4					✓	
22. WERI1	✓	✓	✓	✓	✓	✓

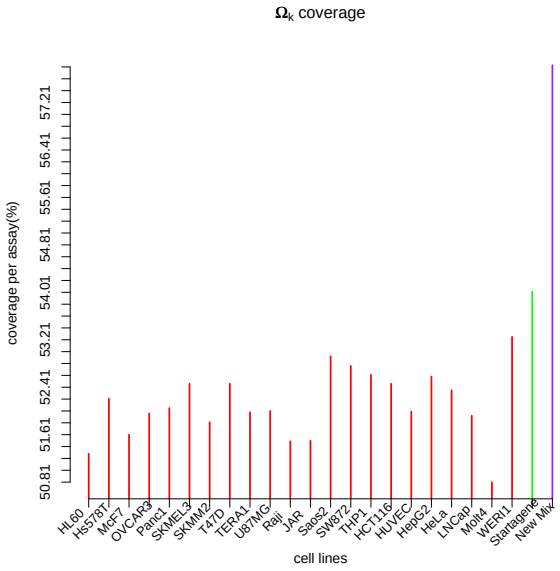

Fig. 5. The individual coverage percentages for the 22 cell lines (coloured red), UHRR Stratagene (coloured green) and the optimal New Mix (coloured purple) selected from *Cov* method.

*WERI*1 cell line. *Ord* method suggests a subset of *Cov*'s optimal mix, excluding *Saos* and *LNCap*. *Cov* and *RM* criteria mostly agree in the content of the optimal set which consists of 11 cell lines in both cases. 2*g* method suggests the same cell line set with *Cov* adding *McF*7 cell line. *GCD* criterion finds 15 cell lines. *RV* criterion suggests the smallest optimal set which consists of 6 cell lines, and mostly disagrees with the other criteria. The remaining five methods agree for $\{OVCAR3, Panc1, TERA1, Raji, HUVEC, HepG2\}$.

D. Verification for Cov optimal reference sample

We carried out two more assays to compare the coverage of optimal reference mix suggested by the *Cov* approach, which we call *New Mix*, and the UHRR Stratagene reference sample. New Mix consists of equal quantities from the 11 cell lines of *Cov* analysis. Each of the the two reference samples were co-hybridized with our original reference sample. Fig. 5 shows the coverage percentages Ω_k for all cell lines, Stratagene's UHRR and New Mix. We can observe that the theoretical value of 57.12% estimated for New Mix, is very similar to the experimental one 57.84%. Furthermore, the UHRR reference sample exhibits a lower coverage of 54.02%.

IV. CONCLUSIONS AND FUTURE WORK

A. Conclusions

We showed that analysis of cell-line samples can identify systematic structure in measured gene expression levels, which was also suggested by Ross et al.(2000) and Yang et al.(2002). Thus, estimation of pooled reference samples could aim not only on the expression of individual probes in each cell line but also on the expression levels of probes within cell lines. Based on the correlation structure found among cell lines, we considered whether the array coverage could benefit from the exclusion of some of the cell lines.

Particularly, when coverage was defined as the percentage of probe intensities above a pre-specified threshold, we found the maximum coverage per cell line to be 53.67. This percentage was increased to 57.12% for an optimal reference pool of 11 cell lines. When that particular reference mix was experimentally tested, it achieved coverage equal to 57.84%. For the same platform Stratagene's UHRR achieved 54.02% coverage.

We also considered alternative methods to estimate the optimal cell line set and test whether our results from different types of analysis were similar. In this case a cutoff value was not used. Firstly, we considered ordering the cell lines according to the number of genes that achieve their maximum in each cell line. We succeeded reducing the computational burden of the exhaustive search and found an optimal mix very similar to the one suggested by *Cov* approach. Secondly, we estimated the set of those cell lines that can best describe the correlation matrix of the data. Particularly, we employed SA algorithm to maximize three PCA relevant criteria given the correlation matrix of the data and found very similar results with *Cov* approach for two of the three criteria used. Thus, SA algorithm would be preferable since it can potentially handle higher dimensional problems.

B. Future Work

The definition of a broadly accepted universal microarray reference sample would further allow to utilize data from different studies and eventually facilitate the integration of data

obtained from different platforms. Although the estimated coverage of the cell lines mix suggested here is not very high, using the optimization techniques applied we could investigate a wider group of cell lines given the coverage criterion. Future work includes expanding our search to other publicly available data sets, such as the *NCI*60 set [7], and explore whether cell lines with different origin from the examined cell lines, would achieve higher array coverage.

V. ACKNOWLEDGMENTS

The authors thank E.Christodoulou and M.Ioannou for discussions and useful comments.

REFERENCES

[1] Y. Benjamini and D. Yekutieli, The control of the false discovery rate in multiple testing under dependency, *The Annals of Statistics*, vol. 29, 2001, pp 1165-1188.

[2] J. Cadima , J.O. Cerdeira and M. Minhotob, Computational aspects of algorithms for variable selection in the context of principal components, *Computational Statistics and Data Analysis*, vol. 47, 2004, pp 225-236.

[3] B. Efron and R.J. Tibshirani, An introduction to Bootstrap, *Boca Raton, FL: CRC Press*, 1994.

[4] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth and A.H. Teller, Equations of state calculations by fast computing machines, *Journal of Chemical Physics*, vol. 21, 1953, pp 1087-1092.

[5] N. Novoradovskaya, M.L. Whitfield, L.S. Basehore, A. Novoradovsky, R. Pesich, J. Usary, M. Karaca, W.K. Wong, O. Aprelikova, M. Fero, C.M. Perou, D. Botstein and J. Braman1, Universal Reference RNA as a standard for microarray experiments, *BMC Genomics*, vol. 5, 2004, pp 227-235.

[6] C.M. Perou, T. Surlie, M.B. Eisen, M. van de Rijn, S.S. Jeffreyk, C.A. Rees, J.R. Pollack, D.T. Ross, H. Johnsen, L.A. Akslen, E. Fluge, A. Pergamenschikov, C. Williams, S.X. Zhu, P.E. Lunning, A.L. Burresen-Dale, P.O. Brown and David Botstein, Molecular portraits of human breast tumors, *Letters to Nature*, vol. 406, 2000, pp 747-752.

[7] D.T. Ross, U. Scherf, M.B. Eisen, C.M. Perou, C. Rees, P. Spellman, V. Iyer, S.S. Jeffrey, M. Van de Rijn, M. Waltham, A. Pergamenschikov, J.C.F. Lee, D. Lashkari, D. Shalon, T.G. Myers, J.N. Weinstein, D. Botstein and P.O. Brown, Systematic variation in gene expression patterns in human cancer cell lines, *Nature Genetics*, vol. 24, 2000, pp 227-235.

[8] H. Shimodaira, Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling, *The Annals of Statistics*, vol. 32, 2004, pp 2616-2641.

[9] G.K. Smyth, Y.H. Yang and T. Speed, Statistical Issues in cDNA Microarray Data Analysis, *Methods in Molecular Biology*, vol. 224, 2003, pp 111-136.

[10] E. Sterrenburg, R. Turk, J.M. Boer, G.B. van Ommen and J.T. den Dunnen, A common reference for cDNA microarray hybridizations, *Nucleic Acids Research*, vol. 30, 2002, pp e116.

[11] O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein and R.B. Altman, Missing value estimation methods for DNA microarrays, *Bioinformatics*, vol. 17, 2001, pp 520-525.

[12] L.J. van 't Veer, H. Dai, M.J. van de Vijver, Y.D. He, A.A.M. Hart, M. Mao, H.L. Peterse, K. van der Kooy, M.J. Marton, A.T. Witteveen, G.J. Schreiber, R.M. Kerkhoven, C. Roberts, P.S. Linsley, R. Bernards and S.H. Friend, Gene expression profiling predicts clinical outcome of breast cancer, *Letters to Nature*, vol. 415, 2002, pp 530-536.

[13] E. Wit and J. McClure, *Statistics for Microarrays: Design, Analysis and Inference*, John Wiley & Sons, NJ; 2004.

[14] I.V. Yang, E. Chen, J.P. Hasseman, W. Liang, B.C. Frank, S. Wang, V. Sharov, A.I. Saeed, J. White, J. Li, N.H. Lee, T.J. Yeatman and J. Quackenbush, Within the fold: assessing differential expression measures and reproducibility in microarray assays, *Genome Biology*, vol. 3, no 11, 2002, pp 0062.1-0062.13.