Available online at www.sciencedirect.com
ScienceDirect
Genomics 90 (2007) 636–645
GENOMICS
www.elsevier.com/locate/ygeno

Method

Utilizing microarray spot characteristics to improve cross-species hybridization results

Carmiya Bar-Or [a,b], Eugene Novikov [c], Anat Reiner [d], Henryk Czosnek [b], Hinanit Koltai [a,*]

[a] Department of Ornamental Horticulture, ARO Volcani Center, Bet Dagan 50250, Israel
[b] The Robert H. Smith Institute of Plant Sciences and Genetics in Agriculture, Faculty of Agricultural, Food and Environmental Quality Sciences, The Hebrew University of Jerusalem, Rehovot, Israel
[c] Service Bioinformatique, Institut Curie, Paris, France
[d] Department of Physics of Complex Systems, Weizmann Institute of Science, Rehovot, Israel

Received 7 March 2007; accepted 21 June 2007
Available online 20 September 2007

Abstract

Cross-species hybridization (CSH), i.e., the hybridization of a (target) species RNA to a DNA microarray that represents another (reference) species, is often used to study species diversity. However, filtration of CSH data has to be applied to extract valid information. We present a novel approach to filtering the CSH data, which utilizes spot characteristics (SCs) of image-quantification data from scanned spotted cDNA microarrays. Five SCs that were affected by sequence similarity between probe and target sequences were identified (designated as BS-SCs). Filtration by all five BS-SC thresholds demonstrated improved clustering for two of the three examined experiments, suggesting that BS-SCs may serve for filtration of data obtained by CSH, to improve the validity of the results. This CSH data-filtration approach could become a promising tool for studying a variety of species, especially when no genomic information is available for the target species.

Keywords: cDNA microarrays; Gene expression; Image quantification; Species diversity; Transcriptome

Cross-species hybridization (CSH), i.e., the hybridization of a (target) species RNA to a DNA microarray that represents another (reference) species, is widely used in comparative ecological and evolutionary studies, as well as when no representative microarray platform is available for the target species [1]. However, microarrays are designed for species-specific hybridizations (SSHs; [2]), not CSHs. Hence, CSH is still considered a nonstandard application of microarrays, and its results need to be carefully interpreted. Indeed, previous studies have produced contradictory results regarding the ability of CSH to reflect valid biological results (e.g., [3,4]). Rather, filtration of CSH data, to gain SSH-like data, has to be applied to extract valid information [1].

During CSH, a low level of sequence matching between transcripts of the target species and some of the microarray probes is expected [1]. In cases in which sufficient genomic information is available for the target species, its proper utilization for the filtration of data from CSH to spotted cDNA microarrays has been shown to improve the results [5,6]. However, for cases in which genomic information is lacking for a target species (most species), no data-filtration approach for CSH to spotted cDNA microarrays is available.

cDNA microarray spots are composed of multiple probe molecules; during CSH, many of the probes are expected to have a low level of matching to the transcripts of the target species. We therefore hypothesized that the low-level probe–transcript matching during CSH will have cumulative effects on hybridized spots. These cumulative effects might be evident for a spot detected in an image of a scanned spotted cDNA microarray in the form of spot characteristics (SCs; e.g., spot signal uniformity or spot dimensions); SC values would differ between spots bearing probes with high and low probe–transcript matching. If this hypothesis is correct, then detection of SCs for a given spot could indicate the level of matching between a probe and the target transcript. Data filtration that includes those corresponding to spots with high probe–transcript match could then be performed by SC values. Such

* Corresponding author. Fax: +972 3 9669583.
E-mail address: hkoltai@agri.gov.il (H. Koltai).

0888-7543/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.ygeno.2007.06.008

filtration might lead to the extraction of valid data from CSH. Filtration by spot dimensions (i.e., spot diameter) has been performed for SSH [7].

Here, we present several lines of evidence in support of our hypothesis and, as a consequence, a novel data-filtration approach that utilizes SCs to improve CSH results. Since only SCs (and no genomic data) are needed for the data filtration, this approach could become a powerful tool for the extraction of valid data from CSH for a variety of species.

Results

Identification of bit-score-correlated SCs

To test our hypothesis of cumulative effects of low-level probe–transcript matching on a microarray spot, and the suggested data-filtration approach, SCs affected by sequence similarity between probe and target sequences were sought.

In a previous study [5], the level of sequence similarity between probes and transcripts was proven to be key to the derivation of SSH-like knowledge (and hence valid biological knowledge) from CSH data. The level of sequence similarity was determined between reference-species microarray probes and target-species genes: for each cDNA microarray clone, the sequence of the best clone-representative tentative consensus (TC) of the reference species was matched with the target species TCs, using the Basic Local Alignment Search Tool (BLAST). The sequence-similarity level, in terms of BLAST bit-score values, was used to filter the CSH data for those corresponding to probes with a high match to target transcripts, thereby facilitating extraction of SSH-like knowledge from CSH [5].

Therefore, in our search for SCs that are affected by sequence similarity between probe and target sequences, we looked for BLAST bit-score-correlated SCs (BS-SCs), i.e., SCs whose values correlate with high or low values of probe–transcript match: matching was detected by BLAST between TCs of the reference and of the target species and was scored by BLAST bit-score values. These sequence-similarity-affected SCs were then examined as parameters for data filtration that corresponds to probes with a high match to target transcripts.

To identify BS-SCs, publicly available TIGR (The Institute for Genomic Research) scanned microarray images from three experiments, designated 054, 057, and 058 (http://www.tigr.org/tdb/potato/), were examined. These experiments involved time-point profiling of gene expression in plants responding to salt, heat, and cold stress, respectively. In each experiment, RNA from seven Solanaceae plants (potato, tomato, eggplant, pepper, tobacco, *Nicotiana benthamiana,* and petunia) was subjected to SSH and CSH to TIGR potato spotted cDNA microarrays. MAIA software [8] was chosen for quantification of the microarray images, since it provides values for 10 different SCs.

Sufficient genomic information was available for only two (tomato and potato) of the seven Solanaceae species (http://compbio.dfci.harvard.edu/tgi/plant.html). Therefore, bit-score values were determined for sequence matching only between tomato TCs and microarray-represented potato TCs.

Notably, SC analyses were performed on potato clones, represented by the microarray spots. However, TCs were used for bit-score value determinations (see above). Consequently, the BS-SC results refer to both clones and TCs. For simplicity, both clones and TCs will be referred to as "genes."

To find BS-SCs, values of the 10 SCs were determined [8] for each spot of the tomato-potato CSH (i.e., tomato RNA hybridized to potato microarray) for each of the three experiments. Each of the obtained SC values was plotted against BLAST bit-score values, the latter obtained for each spot as the matching value for matching between the spot-represented potato TC and the best BLAST-matched tomato TC.

Five of the 10 SCs were found to be correlated to bit scores (BS-SCs); these were found across all three examined experiments for the tomato-potato CSH data (Fig. 1; Supplementary Data 1), although the correlation to the bit score was incomplete. Four of the BS-SCs were measures of a single spot. These included Det—coefficient of determination between the intensities of the Cy3 and the Cy5 channels, Dia—spot diameter, GSym—spot geometrical symmetry, and CVR—coefficient of variation of two gene-expression ratio estimates: one by a linear regression approach and the other by a segmentation algorithm [8]. The fifth BS-SC was a coefficient of variation (CV) of gene-expression ratios obtained by replicated spots.

To test our hypothesis further, we sought to examine the correlation between the identified BS-SCs and the CSH data of a species that is more phylogenetically distant from potato than tomato. Hence, BS-SC values of a petunia-potato CSH were plotted against the bit-score values obtained by the tomato-potato TC sequence comparison. Among the examined species, petunia is considered the most phylogenetically distant from potato (the reference species) ([9]; http://www.sgn.cornell.edu/about/about_solanaceae.pl). Hence, for the petunia-potato CSH data, we expected a lower correlation between the five BS-SC values and the tomato-potato bit-score values, relative to that obtained for the tomato-potato CSH data. Indeed, a reduced correlation was observed (Fig. 1), suggesting that the BS-SCs reflect a CSH effect (i.e., effect of matching between multiple probes and transcripts on a transcriptomics scale; [1]), which is related to phylogenetic distance. Some degree of correlation still existed between the five BS-SC values and the petunia-potato CSH data due to a degree of sequence similarity between petunia and tomato genomes (both being Solanaceae family members [9]).

Taken together, the results obtained for the tomato-potato and petunia-potato CSH data strengthened the concept of an effect of gene-sequence similarity between reference and target species on the five BS-SCs. The question then became, can BS-SCs, similar to bit-score values, be used as parameters for the filtration of CSH data to improve result validity?

Determination of BS-SC thresholds

Thresholds were determined for the BS-SC-based data filtration. Since bit-score values were incompletely correlated to the BS-SCs, a determination of BS-SC thresholds based on a particular bit-score value [5] was impractical. Hence, on one

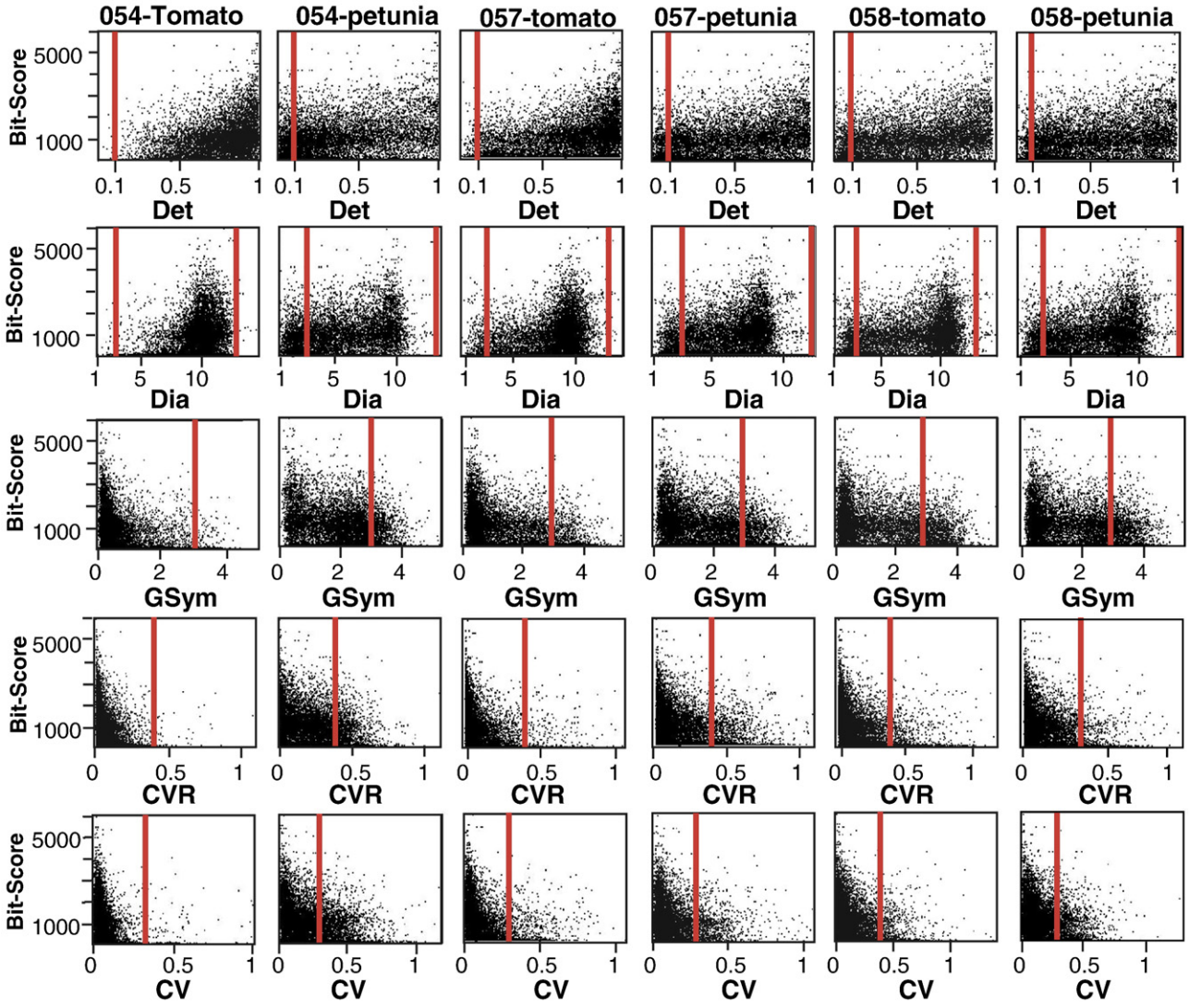

Fig. 1. Scatter plot of BS-SCs from CSH of tomato or petunia RNA to potato microarrays, plotted against bit-score values for tomato-potato sequences. Each dot represents the average value of a BS-SC, for each microarray spot, over all target species and examined hybridizations in an experiment (054, 057, or 058) and is plotted against the highest bit-score value of potato-tomato TC-sequence matches corresponding to that one spot. BS-SC designations: Det—coefficient of determination between the intensities of the Cy3 and the Cy5 channels, Dia—spot diameter, GSym—spot geometrical symmetry, CVR—coefficient of variation of two gene-expression ratio estimates, one by a linear regression approach and the other by a segmentation algorithm, CV—coefficient of variation of gene-expression ratios obtained by replicated spots. Vertical (red) line denotes the BS-SC threshold values used for data filtration ($Det > 0.1$, $3 < Dia < 13$, $GSym < 3$, $CVR < 0.4$, and $CV < 0.3$). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

hand, the CSH data from all examined species was enriched for those corresponding to high bit-score values (of tomato-potato); i.e., BS-SC thresholds were determined based on the scattering of bit-score values for different BS-SC values, such that the BS-SCs that corresponded most closely to low bit-score values, and less to high bit-score values, would be eliminated (Fig. 1). On the other hand, so that most of the CSH data would be included, at least 75% of the petunia data (i.e., the species most distantly related to potato in the examined experiments) was included (Fig. 2).

Interestingly, the effect of CSH on the five BS-SCs emerged from their distribution box plots (Fig. 2): the phylogenetic distance between the target and the reference species affected each BS-SC. For example, GSym values were higher for target species that are more closely related to potato (e.g., tomato) than for those more distantly related to potato (e.g., petunia). A statistical analysis confirmed a significant effect of the phylogenetic distance between reference and target species on the five BS-SCs (Table 1; Supplementary Data 2). In addition, despite a separate determination of thresholds (i.e., for 75% of the petunia data to be included, thresholds were determined separately for each experimental system), the resultant thresholds were found similar across all three experimental systems: $Det > 0.1$, $3 < Dia < 13$, $GSym < 3$, $CVR < 0.4$, and $CV < 0.3$. This suggested a fairly steady species effect on the BS-SCs for this particular microarray platform.

Notably, although BS-SC thresholds were set by the petunia BS-SC data distribution, they led to thousands of filtered genes (Table 2). This implies that for each filtered petunia gene, similar-sequence genes of the other Solanaceae species examined (which are phylogenetically related to petunia and hence share a fraction

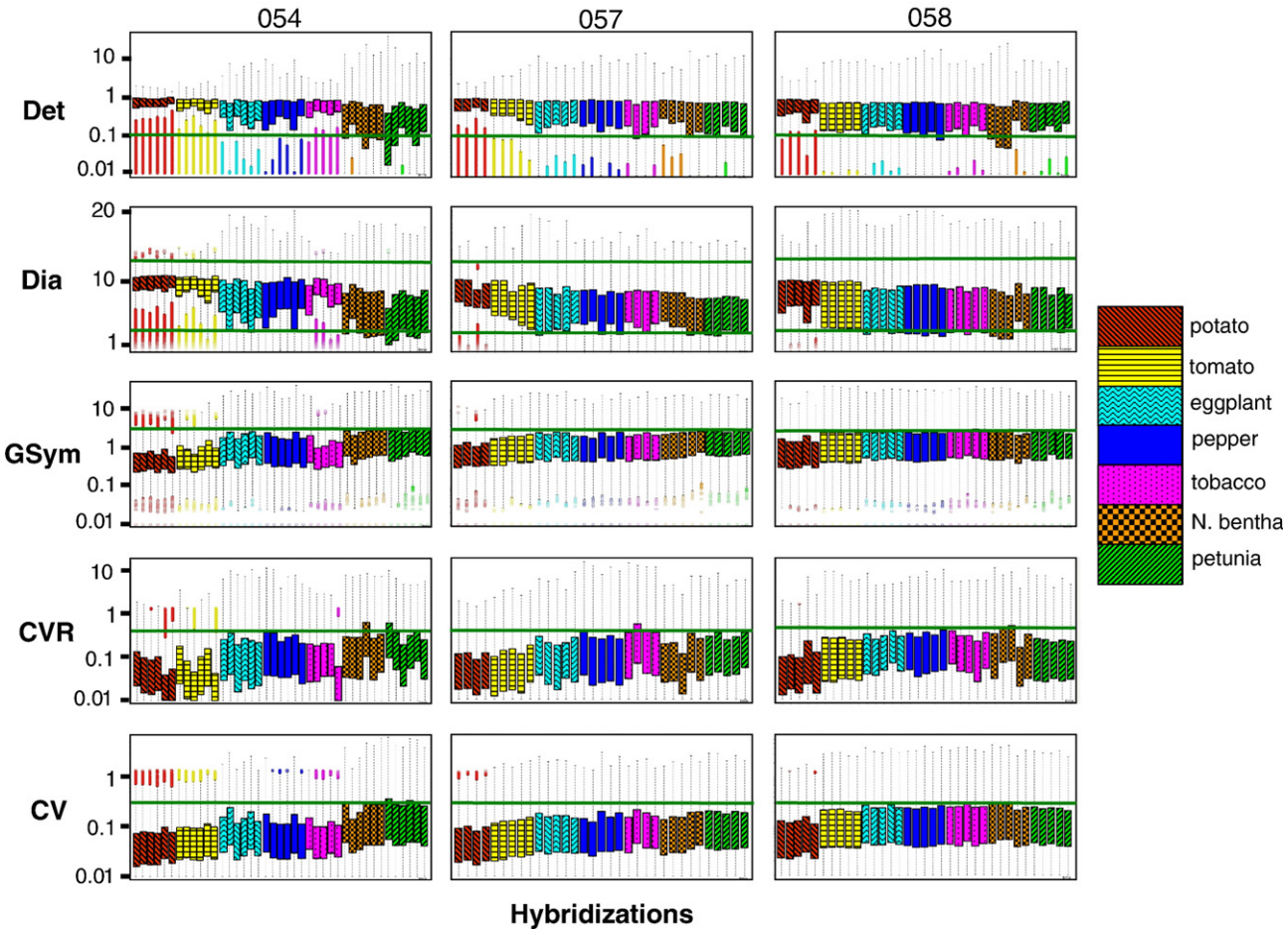

Fig. 2. Box plot of BS-SC value distribution for each hybridization of all examined species and experiments (054, 057, 058). BS-SCs designated as in legend to Fig. 1, pattern and color code for species is presented (N. bentha-*Nicotiana benthamiana*). Horizontal (green) lines denote the BS-SC threshold values used for data filtration (Det > 0.1, 3 < Dia < 13, GSym < 3, CVR < 0.4, and CV < 0.3). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

of similar genes) had similar BS-SC values. This multiplicity of filtered genes further strengthened the notion of effect of a same effector, i.e., CSH, on BS-SCs and thus positions BS-SCs as a potential parameter for CSH data filtration.

Use of BS-SCs for CSH data filtration

The determined BS-SC thresholds were used for data filtration. Hierarchical-clustering results of the filtered versus unfiltered CSH data were used as a validity measure for BS-SC-based CSH data filtration. Hierarchical clustering of the unfiltered data ("all genes") resulted in clustering results (Fig. 3) whose significance, scored as tree distance of the resultant clustering tree, is shown in Table 2. Once filtration of the data by each and every one of the five BS-SC thresholds was performed, improved clustering results (in terms of tree distance) were evident for the subsets of filtered data from two (057 and 058) of the three experimental systems (Fig. 3;

Table 1
Spearman ρ correlation between BS-SC averages and (ranked) phylogenetic distance between reference and target species

	054		057		058	
	Spearman ρ	Prob > |ρ|	Spearman ρ	Prob > |ρ|	Spearman ρ	Prob > |ρ|
Det[a]	−0.8447	<0.0001	−0.7479	<0.0001	−0.7894	0.0097
Dia	−0.8608	<0.0001	−0.8163	<0.0001	−0.6818	<0.0001
GSym	0.8276	<0.0001	0.8082	<0.0001	0.5248	0.0012
CVR	0.7026	<0.0001	0.6259	<0.0001	0.2263	0.1911
CV	0.8148	<0.0001	0.6957	<0.0001	0.5672	<0.0001

Correlation coefficient values (Spearman ρ) and their significance (Prob > |ρ|) are presented.

[a] BS-SC designations: Det—coefficient of determination between the intensities for the Cy3 and the Cy5 channels, Dia—spot diameter, GSym—spot geometrical symmetry, CVR—coefficient of variation of two gene-expression ratio estimates, one by a linear regression approach and the other by a segmentation algorithm, CV—coefficient of variation of gene-expression ratios obtained by replicated spots.

Table 2
Hierarchical-clustering results for unfiltered (all) genes and genes filtered by BS-SC thresholds, for each experiment (054, 057, 058)

Parameter for filtration	BS-SC threshold	054			057			058		
		No. of genes post-BS-SC filtration	Tree distance	Rec[a]	No. of genes post-BS-SC filtration	Tree distance	Rec	No. of genes post-BS-SC filtration	Tree distance	Rec
All genes	NA	15,264	0.931	NA	15,264	0.884	NA	15,264	0.664	NA
Det[b]	>0.1	5891	0.985	NA	8441	0.806	Yes[c]	7738	0.274	Yes
Dia	3 to 13	4916	0.915	NA	7340	0.837	Yes	6126	0.311	Yes[c]
GSym	<3	3901	0.916	NA	5575	0.833	Yes	4496	0.256	Yes[c]
CVR	<0.4	4747	0.933	NA	6803	0.807	Yes	7040	0.269	Yes
CV	<0.3	3957	0.998	NA	5385	0.833	Yes	3961	0.229	Yes
All BS-SCs	All	2855	0.992	NA	4219	0.839	Yes	3031	0.207	Yes

[a] Reconstruction of clustering-separation results obtained by filtering the data for all five BS-SC thresholds, not applicable (NA) for 054 (due to lack of clustering separation), by time for 057, and by species for 058.
[b] BS-SC designations: Det—coefficient of determination between the intensities for the Cy3 and the Cy5 channels, Dia—spot diameter, GSym—spot geometrical symmetry, CVR—coefficient of variation of two gene-expression ratio estimates, one by a linear regression approach and the other by a segmentation algorithm, CV—coefficient of variation between gene-expression ratios obtained by replicated spots.
[c] Imperfect reconstruction.

Table 2). In addition, an improved order of clustering by time and species was observed for 057 and 058, respectively (Fig. 3): for the 057 data subset, which was filtered by all BS-SCs, all experimental hybridizations, except those of potato, clustered by time. For the 058 data subset, which was filtered by all BS-SCs, all experimental hybridizations except one clustered by

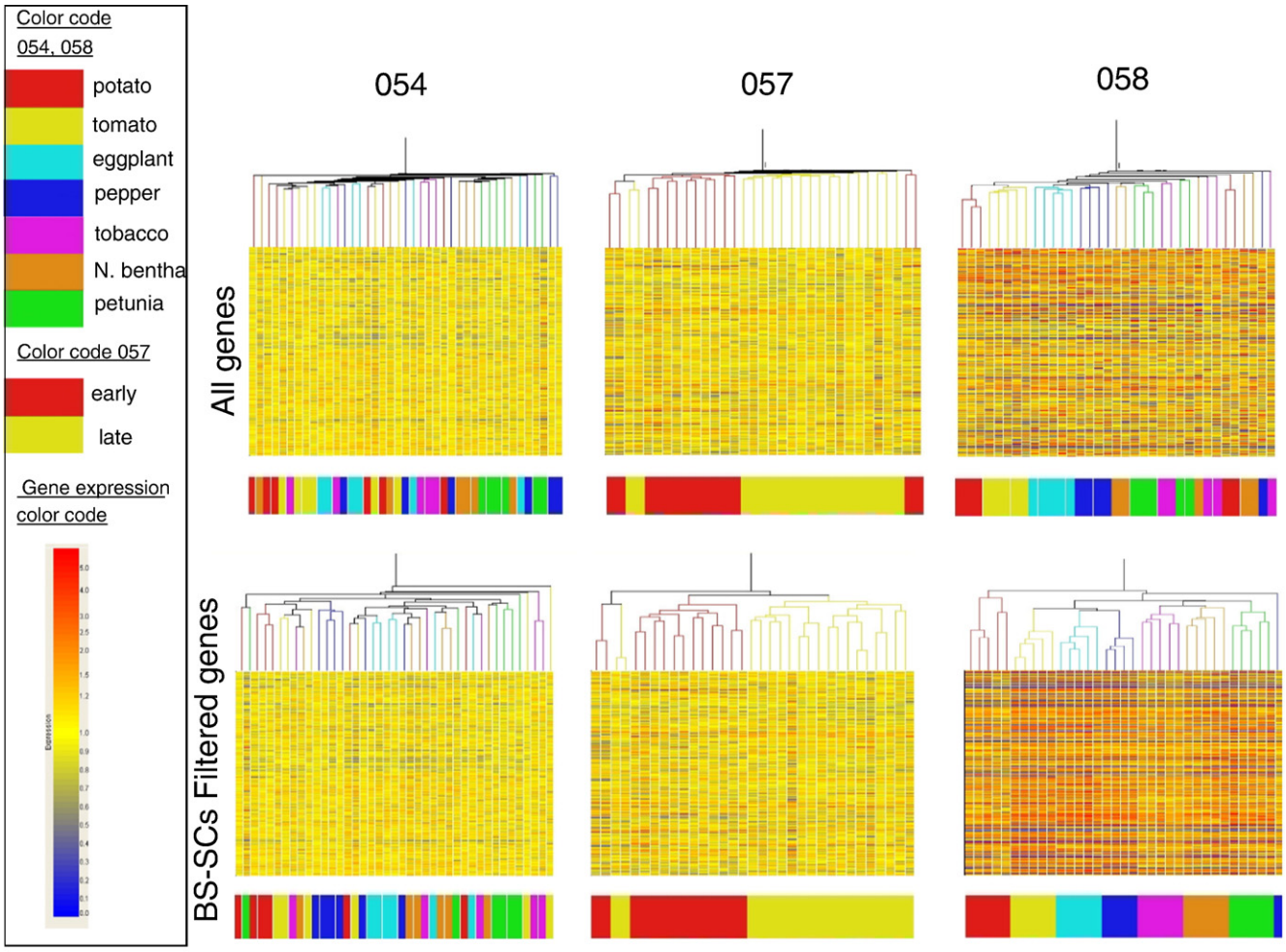

Fig. 3. Hierarchical clustering results for all genes and genes filtered by BS-SC, for all examined experiments (054, 057, 058). Color code for species, for the 054 and 058 bars, is presented. Color code for the 057 bar is by time: early time points ("early"; 6, 12 h) in red; late time points ("late"; 24, 48, 96 h) in yellow. Color code for gene-expression ratios is presented.

species. Taken together, these clustering results for CSH data, filtered by BS-SC thresholds, suggest that BS-SCs can serve, at least in some cases, as parameters for CSH data filtration for improved validity.

Discussion

Correlation between bit-score values and BS-SCs

Here, we demonstrate five SCs that are correlated to bit-score values; this correlation exists despite other factors, such as pin size for spot printing or spot position on the microarray platform, which may affect the BS-SCs. Nevertheless, bit-score values and BS-SCs were incompletely correlated: low bit-score values were correlated with both high-quality and low-quality BS-SC values. However, high bit-score values were mostly correlated with high-quality BS-SC values.

The incomplete correlation between low bit-score values and BS-SCs may result from the imperfect nature of the dataset of bit-score values: neither the tomato nor the potato genomic database is complete. This may be reflected in the multiplicity of low bit-score values, accounting in particular for the unknown-as well as the normal-distribution of the tomato-potato bit-score values (Supplementary Data 3). In a complete genomic dataset, some of these low bit-score values might increase due to improved match between tomato and potato genes. This, in turn, may improve the correlation between bit-score values and BS-SCs.

In addition, bit-score values in our study were generated to include those corresponding to the best-matched tomato TC for each potato TC, represented by each spot. However, a different gene, with reduced sequence similarity to the particular spot, may be expressed and hybridized (by cross-hybridization [1]) under the examined conditions, rather than the best-matched tomato TC (for which the bit-score value was included). This phenomenon, given the constructed bit-score values, may be another reason for the reduced correlation between low bit-score values and BS-SCs.

How does CSH affect the BS-SCs?

During both SSH and CSH, in addition to perfect matches between target transcripts and reference probes, low and no matches are expected [1]. A low match for SSH may be the result of, for example, expression and hybridization of gene family members, whereas for CSH it is mostly the result of reference-species probe hybridization to target-species gene transcripts with reduced sequence similarity. Finding no match for SSH indicates a lack of transcripts, whereas for CSH, in addition to lack of transcripts, this may indicate a lack of a similar-sequence gene(s). Generally, more low probe-transcript matches and less perfect matches are expected during CSH than during SSH.

Nevertheless, once hybridization occurs, low or perfectly matched transcripts are potentially equally fluorescent. Hence, the detection of low and perfect probe-transcript matches, by means of image processing for scanned microarray images, should be similar. Therefore, results detected from scanned images of SSH should not differ from those of CSH; yet, the BS-SC values resulting from image processing are different between SSH and CSH scanned microarray images (Fig. 2). What are the underlying mechanisms that might affect SCs differently for CSH and SSH?

It might be that during post-hybridization washing, low-matched transcripts dissociate (Fig. 4A; [10]). This leaves clearings of probes with no bound transcripts (Fig. 4B). Since low-matched transcripts are relatively abundant in CSH compared to SSH, the former is expected to have more of these clearings (Fig. 4B). Are these suggested differences between SSH and CSH evident in a scanned microarray image? Spotted microarray is composed of spots; a detected spot of a microarray image is composed of multiple imaged pixels. Since the signal intensity detected for a pixel measures the sum of the bound transcripts, probe–transcript clearings might be detected as a low pixel signal. As a consequence, pixels with low signals might be falsely classified as background (Fig. 4C).

Low-signal pixels and those falsely classified as background can potentially have three effects on a detected spot (Fig. 4C): The first effect is detection of a reduced signal for a spot; low signal has been observed for CSH [3–5,11,12]. The second effect is observation of a nonuniform signal for a spot. The third effect is distortions in the spot contour (Fig. 4C). These three effects may be visible in a scanned microarray image (Fig. 4D).

Evidence for these effects is suggested in the form of BS-SCs: Cy3 and Cy5 signals are nonuniformly and indeterministically distributed and hence the coefficient of determination is lower (lower Det). Consequently, linear regression of the Cy5/Cy3 ratio is less robust; hence, the CVR is larger. A smaller number of pixels is counted per spot, hence spot diameter is smaller (smaller Dia). The circular shape of the spot is distorted, hence GSym is higher. As a result, ratios obtained by the CVR segmentation algorithm become uncertain, leading to higher CVR. A nonuniform signal, introducing uncertainties as to the Cy5/Cy3 ratios, may decrease reproducibility between replicated spots (higher CV). High correlations were observed between the BS-SC values (Table 3). These high correlations might support the suggested associations, further strengthening the notion of a CSH effect on the BS-SCs.

Notably, the larger apparent CV values for CSH of distantly related species, and the smaller apparent CV values for CSH of closely related species, suggested nonreproducible results for CSH. This is in contrast to other studies that have found CSH results to be reproducible (e.g., [5,13,14]). Nonreproducibility of CSH data might result from the instability of hybridizations of low-matched transcripts to microarray probes.

Can other (non-bit-score-correlated) SCs be used to improve CSH results?

One could argue that BS-SCs are merely quality SCs. Hence, can other (non-bit-score-correlated) SCs be used for CSH result improvement? The use of SCs, regardless of the gene-sequence similarity between reference and target species, considerably decreased the number of filtered genes (not shown). In contrast, using the five BS-SCs for filtration facilitated a rational usage of

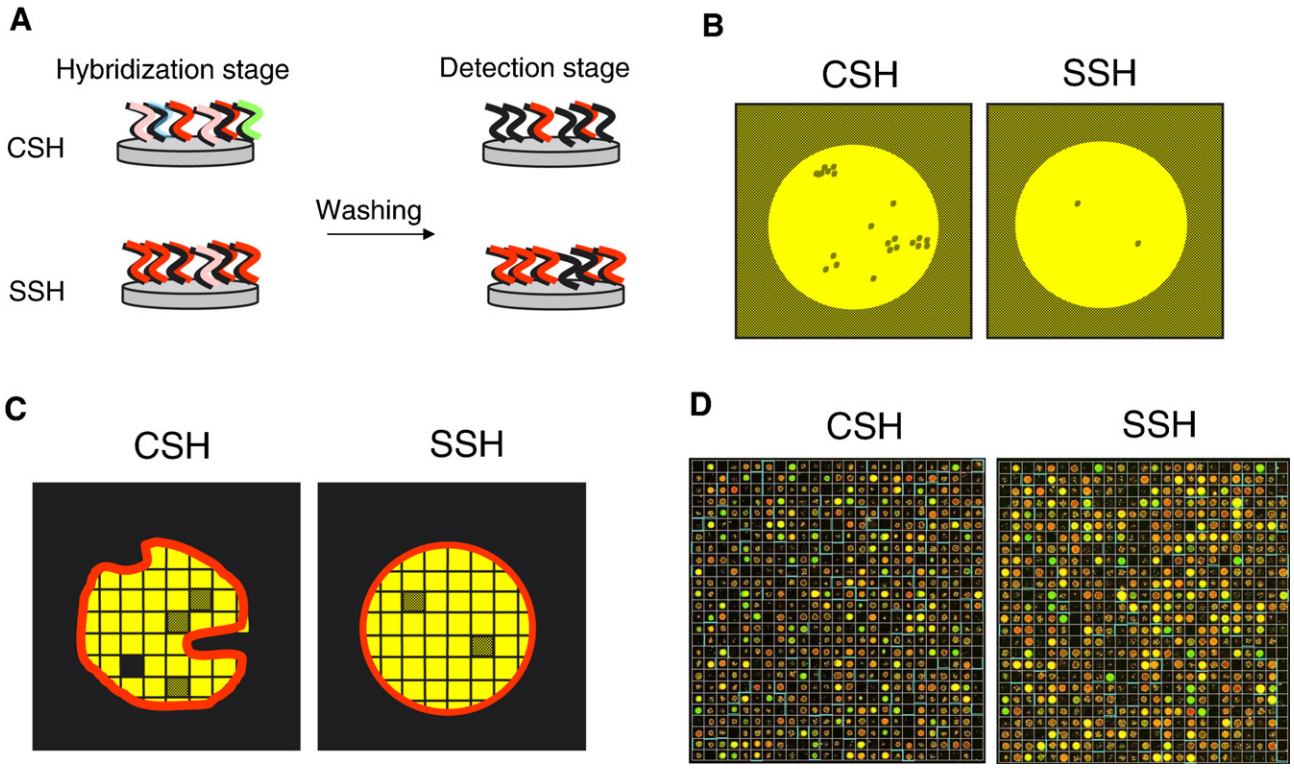

Fig. 4. Illustration of suggested mechanisms that affect SCs differently for CSH versus SSH. (A) Loss of transcripts for CSH during post-hybridization washing stage. A gray cylinder denotes a microarray spot, black lines denote probe molecules, dark (red) lines denote transcripts of a single gene that perfectly matches the spot probes, and lighter (pink, blue, or green) lines denote transcripts of genes with lower sequence similarity to the spot probes. (B) A microarray spot following SSH or CSH. Dots denote clearings on probes with no bound transcripts. (C) Processed scanned images of the microarray spots illustrated in (B). Squares denote pixels. Light (yellow) squares denote pixels with high signal, shaded (gray) squares denote pixels with low signal, black squares denote pixels falsely classified as background. Curved lines denote spot contour. (D) Example images to demonstrate effects of CSH and SSH on SCs of a single microarray block. The block images are of petunia-potato CSH and potato-potato SSH, randomly chosen from experiment 058. Grids and spot contours were plotted by MAIA post-image processing. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

quality characteristics: it led to exhaustion of the CSH data, i.e., it improved the results for a rational number of genes.

BS-SC thresholds

Due to a fairly steady species effect on BS-SCs, the BS-SC thresholds identified in our study may be applicable only to the examined experiments, or perhaps to the potato microarray platform and the experimental "hands" (i.e., TIGR labs). Thresholds for other CSH experiments should be determined, based on their BS-SC distributions, such that they include most of the data (e.g., 75%) from the most phylogenetically distant species examined. However, as in the present study, when SSH data are available, the obtained thresholds are suggested to be within the corresponding SSH distributions: due to the CSH effect on BS-SCs, the corresponding SSH distribution might give an estimate of compatibility between the reference platform and the target species.

Improved clustering results

Filtration by all five BS-SC thresholds in our study demonstrated improved clustering for two (057 and 058) of the three experiments. The improved clustering for 057 and 058 (by time and species, respectively) suggested that BS-SCs can serve as parameters for CSH data filtration for improved results validity.

The obtained data-clustering by species, for all (except one) of the hybridizations in experiment 058, suggested a species-unique response to cold stress. By filtering for BS-SCs, we forcibly selected for probes that represent conserved genes. Hence, one might argue that differences between species may be masked. Nevertheless, the 058 clustering results of a species-unique response to stress suggested that using BS-SCs for data filtration is unlikely to mask differences in gene expression between species. Rensink et al. [7] suggested a robust response to cold in potato, by identifying the highest number of differentially expressed clones for cold response relative to other stress responses. We suggest that cold response not only is robust, but also differs between the examined Solanaceae species.

The obtained data clustering by time for experiment 057 suggested similarity between species in stress perception (early time points), followed by similarity between species in stress response (late time points). Similar results, involving the clustering of gene-expression data by time, have been demonstrated by Rensink et al. [7] for salt- and cold-stress responses in potato. Nevertheless, in our study the SSH data clustered separately from the CSH data. This could signify some limitation in the CSH's ability to reflect SSH-like results completely.

Table 3
An example of correlation values between BS-SCs, for tomato-potato or petunia-potato CSHs for each experiment (054, 057, 058)

Det[b]	Dia	GSym	CVR	CV
054_09[a]*_tomato*				
Det	0.397[c]	−0.6466	−0.6702	−0.5277
Dia		−0.5134	−0.2625	−0.2143
GSym			0.4754	0.2589
CVR				0.5113
CV				
054_38_petunia				
Det	0.9163	−0.4436	−0.6666	−0.5548
Dia		−0.4023	−0.7388	−0.6295
GSym			0.4234	0.2739
CVR				0.4711
CV				
057_10_tomato				
Det	0.8369	−0.6913	−0.7013	−0.5775
Dia		−0.6839	−0.6371	−0.5465
GSym			0.5288	0.406
CVR				0.5069
CV				
057_32_petunia				
Det	0.9072	−0.6129	−0.7423	−0.5611
Dia		−0.6134	−0.7622	−0.6041
GSym			0.5311	0.3777
CVR				0.5339
CV				
058_06_tomato				
Det	0.875	−0.7255	−0.6524	−0.5451
Dia		−0.764	−0.6209	−0.5595
GSym			0.5103	0.4232
CVR				0.4646
CV				
058_33_petunia				
Det	0.8955	−0.7244	−0.6212	−0.5109
Dia		−0.7363	−0.6496	−0.5277
GSym			0.504	0.3955
CVR				0.4559
CV				

[a] Within-experiment hybridization number.
[b] BS-SC designations: Det—coefficient of determination between the intensities for the Cy3 and the Cy5 channels, Dia—spot diameter, GSym—spot geometrical symmetry, CVR—coefficient of variation of two gene-expression ratio estimates, one by a linear regression approach and the other by a segmentation algorithm, CV—coefficient of variation of gene-expression ratios obtained by replicated spots.
[c] Pearson correlation was used for Det, CVR, and CV; Spearman ρ correlation for Dia and GSym. For all values $p < 0.0001$.

For experiment 054, filtration did not improve the clustering results. This might be due to the fact that the changes in gene expression were minor in this experiment: in comparison with experimental systems 057 and 058, 054 possessed the lowest amplitude of gene expression (Fig. 3), implying only a minor transcriptional response of Solanaceae plants to salt stress. Hence, clustering of the relatively "silent" gene-expression data of 054 would inhibit manifestation of robust clustering structures. Rather, nonstructured results for both unfiltered and BS-SC-filtered data are expected for experimental system 054.

Conclusions and future prospects

Several lines of evidence in our study suggest that five SCs are affected by CSH and that these five SCs may serve for CSH data filtration. First, the five SCs correlated with BLAST bit-score values for matching between reference- and target-species sequences, for the tomato-potato CSH data. Second, we demonstrated that for petunia (a more distantly related species than tomato to potato), the corresponding correlation is reduced, suggesting that the BS-SCs reflect a CSH effect, which is related to phylogenetic distance. Third, the distribution of the five BS-SCs, plotted for each experimental system and for each of the hybridizations performed in each of the seven examined species, demonstrated a species effect on the BS-SCs; this observation was supported by statistical analysis. Fourth, setting filtration thresholds for each of the BS-SCs, based on data derived from each experimental system, resulted in similar BS-SC thresholds across all three experimental systems, suggesting a similar CSH effect on BS-SCs for a given microarray platform. Fifth, filtration by all five BS-SCs (at their determined thresholds) resulted in numerous genes for each of the Solanaceae species. This multiplicity of genes suggests that all BS-SCs are affected by the same effector (i.e., CSH). A high correlation between the five BS-SCs further supported this notion. Sixth, improved clustering results were obtained for two of the three experimental systems, and a reason for the lack of improvement with the third experimental system was suggested. In addition, a model for CSH was suggested, which supports the study results and hence the notion of a CSH effect on the BS-SCs.

Consequently, we conclude that the five BS-SCs may serve for filtration of data obtained by CSH to spotted cDNA microarrays to improve the validity of the biological results; this novel filtration approach does not require any additional data (e.g., genomic) or hybridizations (e.g., of gDNA [15]). Hence, its future development and usage may promote the study of complex species, particularly those lacking genomic data, by means of CSH.

We envision two routes for further development of the BS-SC-based CSH data-filtration approach. First, to refine the bit-score values better (see above), high-quality and complete genomic databases need to be studied. In addition, the simplified one probe–one transcript model that was used in this study (by associating only one bit-score value to each clone) could be enhanced to include the one probe–multiple transcript (i.e., cross-hybridization [1]) situation that probably prevails during CSH. Such enhancement may be accomplished by associating multiple BLAST results to a clone.

Second, the SC list can be refined. Here, physical modeling of probe–transcript hybridization could be performed for CSH, which would begin by modeling the matching of probe and transcript molecules, their hybridization stability, and the consequent distribution of signal over the microarray-spot surface (further supporting the model described above). A

physical (or statistical) model could then predict distortions in the obtained CSH signal at the subpixel, pixel, or spot level and could perhaps be expanded to include signal distortions at the spot-set and microarray levels. A verified model might lead to a refined SC list, to be used for CSH data filtration.

Eventually, the refined bit-score values and SC list should be subjected to an examination of their association, to create a multivariate model that determines the relationship between a complete list of BS-SCs on the one hand and a complete sequence-similarity dataset between reference and target species on the other. This modeling could then lead to filtration of CSH data with the ability to control carefully, via BS-SCs, the obtained probe–transcript sequence similarity desired for a given CSH study.

These suggested improvements of the BS-SC-based CSH data filtration approach could potentially position CSH as a standard microarray-based technology for studying complex species and their diversity.

Methods

Microarray description

The potato spotted cDNA microarray version 2 was developed and printed by TIGR. The potato microarray contains 15,264 cDNA clones, selected from the potato stolon, root, microtuber, dormant tuber, germinating eye, healthy leaf, and *Phytophthora infestans*-challenged libraries (http://www.tigr.org/tdb/potato/microarray_comp.shtml). Each clone is duplicated (i.e., there are two replicates of each clone). A description of the potato microarrays is also available at GEO (http://www.ncbi.nlm.nih.gov/geo; Accession No. GPL1901).

Microarray experiments

Publicly available data from three experiments were used. These experiments, designated 054, 057, and 058, profiled gene expression in plants responding to salt, heat, and cold stress, respectively. In each experiment, RNA of seven Solanaceae plants [i.e., potato (*Solanum tubersosum*), tomato (*Lycopersicum esculentum*), eggplant (*Solanum melangena*), pepper (*Capsicum annuum*), tobacco (*Nicotiana tabacum*), petunia (*Petunia hybrida*), and *Nicotiana benthamiana*] was subjected to SSH and CSH to the potato microarray. All CSHs were noncompetitive hybridizations, i.e., the hybridized RNA samples in each hybridization were of the same species. For each experiment and for each species, approximately five time points were examined. A complete description of the experiments, including plant growth and stress challenge, RNA extraction, labeling and hybridization, and microarray-scanning protocol, is available at http://www.tigr.org/tigr-scripts/tdb/potato/study/potato_study_hybs.pl?study=54&user=&pass=&sort=id&order=asc for 054; http://www.tigr.org/tigr-scripts/tdb/potato/study/potato_study_hybs.pl?study=57&user=&pass=&sort=id&order=asc for 057; http://www.tigr.org/tigr-scripts/tdb/potato/study/potato_study_hybs.pl?study=58&user=&pass=&sort=id&order=asc (for 058). All resultant images, available online from January 2007, were examined in this study and are available at this link.

Matching potato microarray-represented genes to tomato genes

Since neither potato nor tomato is fully sequenced, their TC sequences (available via ftp://occams.dfci.harvard.edu/pub/bio/tgi/data/Solanum_tuberosum/STGI.release_10.zip and ftp://occams.dfci.harvard.edu/pub/bio/tgi/data/Lycopersicon_esculentum/LGI.release_10-1.zip databases for potato and tomato, respectively) were used for matching. For this purpose, potato microarray TC sequences served as queries for a local BLAST search against the tomato TC database. All potato TC IDs associated with the microarray-printed clones were retrieved from the microarray manufacturer. These are available in the GAL file supplied in Supplementary Data 4, in the column "THC#." For the creation of a tomato database index file, the NCBI FORMATDB utility was used over the tomato TC database. Then, NCBI local BLASTN was used to match tomato-potato TCs. The best match bit-score result was associated with each of the microarray clones. All clone and bit-score pairs are available in the GAL file supplied in Supplementary Data 4, in the columns "Clone name" and "TUID," respectively. Last, all microarray-associated bit-score values were log-transformed and their distribution was plotted by using JMP IN version 5.1 (SAS Institute).

Image quantification

To facilitate a fully automated quantification, the GAL file was further modified to exclude a description of the 26th column for all of the microarray blocks (Supplementary Data 4). This modification eliminated the need for constant manual grid readjustment.

Batch quantification was performed using MAIA version 2.7 ([8]; MAIA available via http://bioinfo.curie.fr/projects/maia/). For quantification, the modified GAL file (Supplementary Data 4) and all of the images of the three experiments served as input. MAIA's output file format was set to include spot ID, all 10 SCs, and Cy3 and Cy5 signal and background values. The batch run included spot localization (i.e., grid-fitting), image alignment (between the two color images), and spot quantification. Quantification results of all three experiments, including only the five BS-SCs and Cy5 and Cy3 foreground and background mean values, are available at ftp://ftp.weizmann.ac.il/incoming/CB2NM_FTP.zip.

Experimental procedures

All of the following procedures were performed separately for each of the three experiments (i.e., 054, 057, and 058).

Search for BS-SCs

Tomato-potato and petunia-potato CSH quantification data were used to search for BS-SCs. The following steps were taken for each of the target species and for each of the examined SCs: SC values, obtained for each spot, were averaged across the target-species hybridizations (e.g., for a given spot X, Det results were averaged over all of the tomato hybridizations of experiment 054). Then, coefficient of variance was calculated for each clone over the clone SC results obtained from its two replicates. All of the resultant CV values were plotted using JMP IN. These plots were used for a qualitative determination of the CV threshold to be used for filtration that would exclude the distribution tail. Clones bearing CV values greater than the determined threshold were filtered out by Microsoft Excel. Last, using the clone bit-score association (Supplementary Data 4), all of the clones' SC results were plotted against the clone-associated bit-score. BS-SC CV distributions and thresholds and the resultant filtered data, along with clone bit-score associations, are available for both the tomato and the petunia CSH data (Supplementary Data 1).

Examination of BS-SC distributions

To examine the BS-SCs further, GeneSpring (GeneSpring 7.3; Silicon Genetics) was used (in a nonstandard fashion) to generate box-plot charts. For this purpose, all of the MAIA output files were uploaded to GeneSpring five times. Each time, a different BS-SC was considered to be the "signal"; no normalization was performed. Then, all of the microarray-represented clones were used to draw the BS-SC box plots.

Correlation between phylogenetic distance (between target and reference species) and BS-SC values

The following steps were taken for each BS-SC: BS-SC values, obtained for all clones and for all hybridizations, were exported from GeneSpring to an output file, which was uploaded to Excel. Then, average BS-SC values were calculated for each hybridization, over all of the microarray clones.

There is still no consensus on accurate phylogenetic distances in the Solanaceae family. However, rough estimates are available (e.g., at http://www.sgn.cornell.edu/about/about_solanaceae.pl [9]). Hence, these estimates were used to rank the phylogenetic distances between the target and the reference

species (i.e., ranking from 1 to 7, for potato, tomato, eggplant, pepper, tobacco, *N. benthamiana,* and petunia, respectively). The average BS-SC values associated with the target species and rank are available as Supplementary Data 2.

Together, these data (average BS-SC values and species rank) were uploaded to JMP IN. There, Spearman ρ nonparametric correlation was used to trace the correlation between the BS-SC values and the phylogenetic distance between target and reference species.

Notably, since hybridization IDs follow the phylogenetic distance between the target species and potato (the reference species), we confirmed that there was no batch effect on the above results: hybridization parameters (e.g., barcode numbers of the microarray platforms) were checked and were found to be random.

Normalization of gene-expression data

Gene-expression results obtained by MAIA quantification (i.e., Cy3 and Cy5 foreground and background mean values) were subjected to GeneSpring. Lowess normalization [16] was applied; 35% of the data served for the smoothing. In addition, control channel (i.e., Cy3) values of less than 10 were set to 10.

The normalized Cy5/Cy3 ratios were distributed similarly between all hybridizations (Supplementary Data 5). Hence, no further normalization steps were taken (for any of the experiments). Normalized data are available online (see Appendix A). Scanned image files are available at ftp://ftp.tigr.org/pub/data/s_tuberosum/SGED/054_TIGR_salt1/054_images/; ftp://ftp.tigr.org/pub/data/s_tuberosum/SGED/057_TIGR_heat1/057_images/; ftp://ftp.tigr.org/pub/data/s_tuberosum/SGED/058_TIGR_cold1/058_images/.

Data filtration and creation of gene lists

BS-SC thresholds were determined and used to create gene lists in GeneSpring, by filtering for microarray clones that possess BS-SC values above (or below) the determined BS-SC threshold. This was performed for each of the BS-SCs separately, over all examined species. In addition, an intersection gene list was created to include clones whose BS-SC values do not exceed any of the BS-SC thresholds.

Hierarchical clustering

For each of the gene lists (i.e., "all genes," filtered gene lists according to the BS-SC thresholds, and the intersection gene list; see earlier), Pearson average-linkage hierarchical clustering was performed, followed by a bootstrapping confidence check, with 100 iterations. For each clustering result, the Euclidean distance tree was examined and colored according to either time or species.

Within-BS-SC correlation

To examine the correlation between BS-SCs, BS-SC-quantification results of both tomato-potato and petunia-potato CSH data were randomly sampled. The sampled hybridizations were determined to be 054_09_tomato, 054_38_petunia, 057_10_tomato, 057_32_petunia, 058_06_tomato, and 058_33_petunia. For each of these, all five BS-SC quantification results were subjected to JMP IN multivariate analysis. Examination of the five BS-SCs revealed discrete distributions for Dia and GSym and continuous distributions for Det, CVR, and CV (not shown). Thus, to calculate correlations between BS-SCs we used Spearman ρ correlation, for cases in which one of the examined BS-SCs was GSym or Dia, and Pearson correlation otherwise.

Acknowledgments

H.K. was supported by Grant 522/02-1 from The Israeli Science Foundation. H.C. was supported by Grant 593/02-1 from The Israel Science Foundation and by Grant IS-3479-03 from the United States-Israel Binational Agricultural Research and Development Fund (BARD). TIGR hybridization data were obtained through the support of the Institute for Genomic Research, Potato Functional Genomics Expression Profiling Service (NSF Potato Functional Genomics) funded through the U.S. National Science Foundation (DBI-0218166).

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ygeno.2007.06.008.

References

[1] C. Bar-Or, H. Czosnek, H. Koltai, Cross-species microarray hybridizations: a developing tool for studying diversity, Trends Genet. 23 (2007) 200–207.
[2] S. Tomiuk, K. Hofmann, Microarray probe selection strategies, Brief Bioinformatics 2 (2001) 329–340.
[3] Y. Gilad, S.A. Rifkin, P. Bertone, M. Gerstein, K.P. White, Multi-species microarrays reveal the effect of sequence divergence on gene expression profiles, Genome Res. 15 (2005) 674–680.
[4] S.C. Renn, N. Aubin-Horth, H.A. Hofmann, Biologically meaningful expression profiling across species using heterologous hybridization to a cDNA microarray, BMC Genomics 5 (2004) 42.
[5] C. Bar-Or, M. Bar-Eyal, T.Z. Gal, Y. Kapulnik, H. Czosnek, H. Koltai, Derivation of species-specific hybridization-like knowledge out of cross-species hybridization results, BMC Genomics 7 (2006) 110.
[6] P. Khaitovich, G. Weiss, M. Lachmann, I. Hellmann, W. Enard, B. Muetzel, U. Wirkner, W. Ansorge, S. Paabo, A neutral model of transcriptome evolution, PLoS Biol. 2 (2004) E132.
[7] W.A. Rensink, S. Iobst, A. Hart, S. Stegalkina, J. Liu, C.R. Buell, Gene expression profiling of potato responses to cold, heat, and salt stress, Funct. Integr. Genomics 5 (2005) 201–207.
[8] E. Novikov, E. Barillot, An algorithm for automatic evaluation of the spot quality in two-color DNA microarray experiments, BMC Bioinformatics 6 (2005) 293.
[9] W.A. Rensink, Y. Lee, J. Liu, S. Iobst, S. Ouyang, C.R. Buell, Comparative analyses of six solanaceous transcriptomes reveal a high degree of sequence conservation and species-specific transcripts, BMC Genomics 6 (2005) 124.
[10] G.A. Held, G. Grinstein, Y. Tu, Relationship between gene expression and observed intensities in DNA microarrays-a modeling study, Nucleic Acids Res. 34 (2006) e70.
[11] J. Adjaye, R. Herwig, D. Herrmann, W. Wruck, A. Benkahla, T.C. Brink, M. Nowak, J.W. Carnwath, C. Hultschig, H. Niemann, H. Lehrach, Cross-species hybridisation of human and bovine orthologous genes on high density cDNA microarrays, BMC Genomics 5 (2004) 83.
[12] S. Moore, P. Payton, M. Wright, S. Tanksley, J. Giovannoni, Utilization of tomato microarrays for comparative gene expression analysis in the Solanaceae, J. Exp. Bot. 56 (2005) 2885–2895.
[13] L. Donaldson, T. Vuocolo, C. Gray, Y. Strandberg, A. Reverter, S. McWilliam, Y. Wang, K. Byrne, R. Tellam, Construction and validation of a bovine innate immune microarray, BMC Genomics 6 (2005) 135.
[14] K.R. von Schalburg, M.L. Rise, G.A. Cooper, G.D. Brown, A.R. Gibbs, C.C. Nelson, W.S. Davidson, B.F. Koop, Fish and chips: various methodologies demonstrate utility of a 16,006-gene salmonid microarray, BMC Genomics 6 (2005) 126.
[15] J.M. Ranz, C.I. Castillo-Davis, C.D. Meiklejohn, D.L. Hartl, Sex-dependent gene expression and evolution of the Drosophila transcriptome, Science 300 (2003) 1742–1745.
[16] G.K. Smyth, T. Speed, Normalization of cDNA microarray data, Methods 31 (2003) 265–273.