

RESEARCH

Open Access

In Silico cancer cell versus stroma cellularity index computed from species-specific human and mouse transcriptome of xenograft models: towards accurate stroma targeting therapy assessment

Xinan Yang^{1,6}, Yong Huang¹, Younghee Lee¹, Vincent Gardeux^{2,3,4,11}, Ikbel Achour^{2,4,11}, Kelly Regan^{2,4}, Ellen Rebman^{2,4}, Haiquan Li^{2,4,11}, Yves A Lussier^{1,2,4,5,7,8,9,10,11*}

From The 3rd Annual Translational Bioinformatics Conference (TBC/ISCB-Asia 2013) Seoul, Korea. 2-4 October 2013

Abstract

Background: The current state of the art for measuring stromal response to targeted therapy requires burdensome and rate limiting quantitative histology. Transcriptome measures are increasingly affordable and provide an opportunity for developing a stromal versus cancer ratio in xenograft models. In these models, human cancer cells are transplanted into mouse host tissues (stroma) and together coevolve into a tumour microenvironment. However, profiling the mouse or human component separately remains problematic. Indeed, laser capture microdissection is labour intensive. Moreover, gene expression using commercial microarrays introduces significant and underreported cross-species hybridization errors that are commonly overlooked by biologists.

Method: We developed a customized dual-species array, **H&M array**, and performed cross-species and species-specific hybridization measurements. We validated a new methodology for establishing the stroma vs cancer ratio using transcriptomic data.

Results: In the biological validation of the H&M array, cross-species hybridization of human and mouse probes was significantly reduced (4.5 and 9.4 fold reduction, respectively; $p < 2 \times 10^{-16}$ for both, Mann-Whitney test). We confirmed the capability of the H&M array to determine the stromal to cancer cells ratio based on the estimation of cellularity index of mouse/human mRNA content *in vitro*. This new metrics enable to investigate more efficiently the stroma-cancer cell interactions (e.g. cellularity) bypassing labour intensive requirement and biases of laser capture microdissection.

Conclusion: These results provide the initial evidence of improved and cost-efficient analytics for the investigation of cancer cell microenvironment, using species-specificity arrays specifically designed for xenografts models.

Background

In spite of developments in high throughput molecular assessment biotechnologies, it remains particularly challenging to investigate molecular expression in the tumor

microenvironment containing local tumor stroma, reactive stromal cells and cancer cells. There is increasing evidence that the development and progression of cancer is significantly affected by interactions between the tumor and its microenvironment [1-6], suggesting that the ability to profile gene expression in both stromal and cancer cell compartments is critical. Malignant tumors are composed of cancer cells and their associated host cells, which include

* Correspondence: lussier.y@gmail.com

¹Ctr for Biom. Info. Section of Genetic Medicine, Dept. of Medicine, U. of Chicago, IL, USA

Full list of author information is available at the end of the article

blood and lymph endothelial cells, immune cells, fibroblasts and myofibroblasts [7,8]. The host cells, also termed stromal cells here, make up about half of most malignant tumors [9] and have emerged as targets of anti-tumor therapy in recent years [10].

Currently, the best way to selectively isolate cancer cells from a heterogeneous population of cells in a tumor is Laser Capture Microdissection (LCM) [11]. While LCM is increasingly used in the selection of cancer cells [12-14], the application of LCM to stromal masses presents several important technical challenges: (i) some cell types (e.g. endothelial cells) are too long and thin or intertwined with cancer cells to be isolated due to their infiltrating growth and (ii) the visualization of samples and accurate determination of cell type can be difficult because cellular staining via immunochemistry is not compatible with array analyses [10]. Furthermore, LCM of large numbers of cells from many sections and samples is rate limiting and cumbersome, requiring a considerable amount of time (1 day/sample) and sub-zero temperatures. Therefore, novel and affordable high throughput approaches are required to assess the expression of cancer and stromal cells to determine their molecular interactions. Further, transcriptome measures from RNA-seq or conventional expression arrays present the opportunity to develop *in silico* estimates of cancer cells versus stromal cell ratios, which otherwise require labour intensive quantitative histology measures across hundreds of microscopic fields per tissue sample.

Xenograft models, where human cancer cells are grown in immunodeficient mice, are popular for studying the tumor microenvironment. In such models, the genes of the stromal compartment and cancer cells come from distinct species: mouse and human respectively. While gene expression profiles of human cancer and mouse stroma have been analysed by human microarrays [15], this approach is limited because the human and mouse genomes are highly homologous and commercially available arrays are not designed for simultaneously measuring both mouse and human mRNAs [16]. Thus, the interpretation of transcriptome intensities is confounded by **Cross-Species Hybridization (CSH)**, which we have biologically confirmed to exist between human array probes with universal mouse RNAs [17]. We also predicted that this CSH in xenograft models would preferentially occur among homologous human-mouse genes in human arrays resulting in combined gene expression signals where deregulated mouse stromal genes are jointly measured along with human cancer genes [17]. We previously hypothesized that deregulated probes on human arrays exposed to whole xenograft tissue experiments could be mined to identify those probes most likely harbouring human-mouse CSH (due to homologous genes and alternative non-homology factors), and that we could impute the enrichment of biological processes and

molecular functions of stromal mechanisms via these CSH probes using Gene Ontology (GO) gene set annotations [17]. While many probes of homologous genes can be enriched for a GO signal, a limitation of this previous study is that this *geneset*-based approach is not designed to impute differential expression of *individual* mouse stromal genes. There is thus a need for improved species specificity of probes in newer arrays designed specifically for jointly studying stromal and cancer cell gene expression in xenograft models.

In many tumor biology and pharmacology xenograft studies, the interplay and co-expression patterns of cancer and stromal cell genes need to be studied using the simultaneous analysis of both species in the same array experiment. When tissue expression is detected using separate mouse and human arrays, comparing mouse and human gene expression requires complex and artefact-inducing normalization of the two datasets. Currently, there is no genome-wide dual-species microarray available. While whole transcriptome RNA-sequencing (RNA-seq) of xenograft models provides an elegant alternative, the problem of disambiguating the expression intensities of homologous genes remains (though the solution is quite different: an ambiguous RNA assembly problem). There exist two published dual-species partial arrays, in which one contains a mere 516 human probe-sets and 456 mouse probe-sets [16] and a second contains only protease and inhibitor genes [18]. Previously, we reengineered the process of microbial diagnosis by designing a Panmicrobial array that contained 9,477 species specific probes (SSPs) to address 1,710 distinct vertebrate viruses [19-21]. Drawing from this experience, we hypothesized that we could establish a comprehensive map of CSH in human and mouse arrays. We thus designed the first genome-wide human-mouse dual-species array (H&M array) for comprehensively investigating gene expression of stromal. Furthermore, we also hypothesised that this new dual species array would allow for the development of a novel estimation of the stroma/cancer cell ratio via a cellularity index based on SSPs of homologous housekeeping genes as an alternative to traditional histological LCM assessments. Here we report the first genome-wide dual species array for human and mouse gene expression in xenograft models and identify novel biological functions for CSH probes, that until now have been misattributed to either human cancer cells or mouse stromal cells.

Methods

Datasets and custom human-mouse whole genome arrays used in this study

Microarray data from the two *in vitro* experiments (Table 1) are available in the GEO repository under accession number GSE23377 (initial *in vitro* biological experiment: GSE23054, version 1 GEO platform: GPL10714; *in*

Table 1 Cross-species hybridization experiments for H&M array

Initial experiment**						
Dual channel array	1			2		
Exposed to RNA in red channel	1.5ug Hs* RNA (sample 1)			1.5ug Mm* RNA (sample 2)		
Exposed to RNA in green channel	0.75ug Hs* RNA + 0.75 Mm*			0.75ug Hs* RNA + 0.75 Mm*		
Array design	Agilent 2x105k arrays					
Validation experiment						
Single channel array	1	2	3	4	5	6
Exposed to RNA	human	human	human	mouse	mouse	mouse
Array design	46k Human probes and 46k mouse probes					

*Hs: Homo Sapiens, Mm: Mus Musculus

**In the initial experiment, two human-mouse dual-species arrays were custom printed on Agilent 2x105k arrays. Each array contained only one species of universal mRNAs on the red channel (either human or mouse) and had a control sample containing equal quantities of universal human and mouse mRNAs on the green channel.

in vitro validation experiment for CSH and stroma cell ratio estimation: GSE23364, version 2 GEO platform: GPL10749, **Agilent Custom Array Order #023265**). We used nine genomic datasets to compute human-mouse CSH (Additional file 1- **Supp. Table S1**). The list of housekeeping mouse genes was previously published [22,23] and the list of homologous mouse-human genes was downloaded from NCBI (<ftp://ftp.ncbi.nih.gov/pub/HomoloGene>) on Feb, 2009 (Build 63). By design, these experiments generated **Datasets S1, S2 and S3** that we provide for reuse by other groups at http://lussierlab.org/publications/HsMm_array

H&M array design, annotation, and analysis (Figure 1: panels A-C)

As shown in Figure 1, the experimental design consists first of an assemblage of the two sets of 44k probes based on Agilent's commercial human array and mouse array. Agilent also provided cross hybridizing estimates between mouse probes and human mRNAs and vice versa (7141 human and 6420 mouse probes; **Datasets S2, S3**), which are redesigned to correct for comprehensive CSH. We designed two custom arrays: the second and final one improved on the original design after biological validation. The H&M arrays contain all legacy probes from the commercial human and mouse arrays as well as the newly designed ones (Figure 1, **panel A**). The second and final custom H&M array is identified as "H&M array version 2" (Agilent GEO platform: GPL10749; **Agilent Custom Array Order #023265**). In an initial *in vitro* biological experiment, we cross-hybridize H&M array version 1 with mouse mRNAs, human mRNAs, and a combination of both human and mouse RNAs (Figure 1, Step B1; Table 1). We also conduct a parallel and extensive BLASTn analysis of potential cross species hybridization of probes (Step B2). Taken together the biological and the computationally derived cross hybridizations serve as a model for CSH propensity of probes, and have an extensive optimization method (Step B3). As shown in panel B of Figure 1,

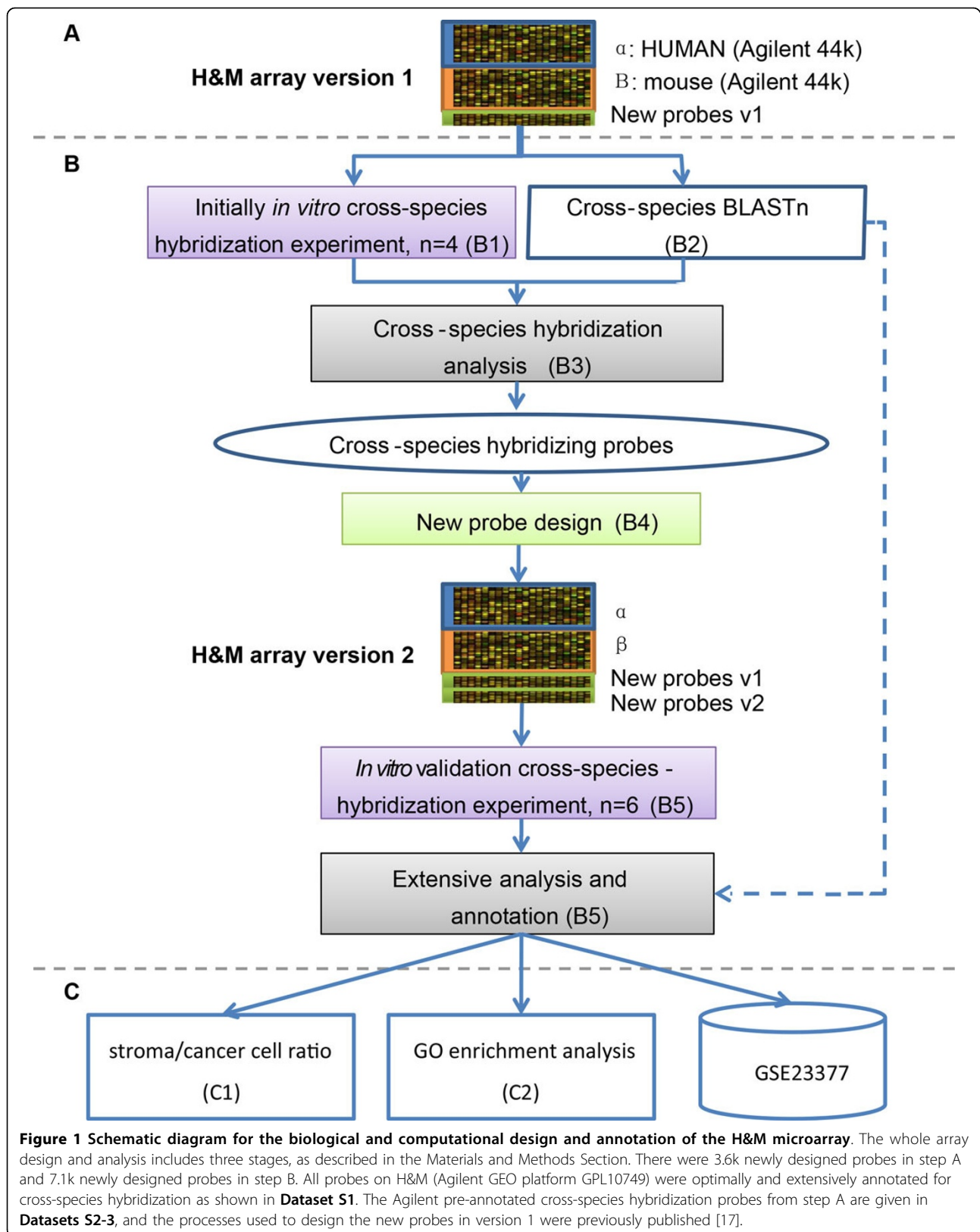
we conduct a second iteration of probe design (Step B4), and produce version 2 of the H&M array. A second and more comprehensive biological experiment that cross hybridizes six arrays (H&M array version 2) with mouse and human mRNAs is used for the final evaluation (Step B5; Table 1) and improved the annotation of the CSH of each probe (**Datasets S2, S3**). Gene Ontology enrichment studies are also performed and show rectification of the cross-hybridization and validation of cellularity ratios (Figure 1, **panel C**).

Step B1: initial cross-hybridization experiment of H&M array version 1

As shown in Table 1, each channel of the two H&M arrays is cross-hybridized producing four readouts. Biological experiments are conducted to produce physical evidence of human-mouse cross-hybridization and are designed to identify CSH by independently arraying mouse mRNAs and human mRNAs (in the red channel) against the joint human and mouse probes of the dual-species array. Using dual channel technology (measuring two groups of mRNA labelled with different colors simultaneously on the same array), we also pool human and mouse mRNA in equal parts as a control on the green channel.

We first obtain human and mouse universal reference RNAs from ArrayIt (Sunnyvale, CA), which were derived from a homogenate of all tissues from the organism, making them somewhat representative of any cell type. This universal RNA is amplified and labelled according to the Agilent Low RNA input Fluorescent Linear Amplification Kit protocol and according to Agilent's two-color quick amp labelling protocol. The two arrays are scanned with 5 µm resolution and 100% power gain at 600PMT for both Cy3 & Cy5 using the GenePix 4000B Axon Scanner.

We quantify gene expression values using Agilent Feature Extraction Software and subtract the adjusted background for each probe from the mean expression value. Finally, we use an inclusion criterion to identify the probes that show some RNA binding where the minimum expression



intensity has to be larger than a value of “1” when exposed to the pooled human and mouse RNA in both arrays. This threshold was set because when the expressed intensity was less than 1, probes vary widely across the two experimental repeats using the same quantity of pooled human and mouse RNAs.

Step B2: BLASTn analysis of cross species hybridization in array probes

The Basic Local Alignment Search Tool (BLAST) algorithm [24] is run with default parameters to theoretically predict the possibility of CSH for each probe by comparing human array probes with mouse RNAs, and mouse array probes with human RNAs. BLASTn was used because of its ability to find “fuzzy” alignments to related nucleotide sequences from other organisms. The hits are ranked by their BLASTn score of which the highest scoring hit is recorded, together with (i) the *length of the alignments* and (ii) the number of *mismatches* (Details are given in Table 2 of this manuscript and Figure D in the Suppl. Methods of our previous publication [17]). Further evaluation of the set of parameters used in the optimal prediction model was also conducted (Table 3, Additional file 1- **Supp. Figure S1**).

Step B3: modelling cross species hybridization with biological experiments and sequence alignments

To predict human-mouse cross-hybridizations and further redesign the probes, we optimize both the statistical models for biological measurement (biological Gold Standard - **GS**) and the parameters used in the CSH theoretical prediction models which are compared with the candidate biological GSs in an iteratively optimal way [17]. Each comparison results in an F-score (**Equation 1** described below after the **Gold Standard**) that balances the recall

(sensitivity) and precision and the CSH theoretical prediction model with the largest F-score is selected. Here we present, as an example, our procedure for identifying mouse probes that are most likely to hybridize with human RNAs, the converse would follow similar steps.

- **CSH theoretical prediction models:** Based on our previously selected theoretical prediction models of cross-hybridization [17] (Table 2, Model 1), we further construct the BLAST prediction model based on the total length of the sequence alignment and the admissible maximum mismatches within the sequence, as shown in Table 3. In order to balance recall and precision while identifying cross-hybridizing probes for which the target gene requires a substitute SSP on the H&M Array, the optimal set of BLAST predictions (*length of the alignments* and the number of *mismatches*), are determined using the ideal F-Score while systematically varying each parameter independently (Additional file 1- **Supp. Figure S2**).

- **Gold standards derived from the biological experiments:** We stratify the mouse probes exposed to human RNAs according to their level of expression in the biological array experiment (Table 1). For H&M array version 1, using a dual channel experimental design, the background adjusted absolute CSH expression of probes (i.e. the raw, unmodified expression of human probes when exposed to mouse RNAs or the expression of mouse probes when exposed to human RNAs), was stratified and alternate GSs are produced accordingly. The optimal biological GS is determined using relative high F-scores[17]. Since the single channel repeats of H&M Array version 2 allow for the use of a relative ratio of average expression of cross species hybridization over

Table 2 Previously established definitions of four biological gold standards (GS) and the parameters of BLASTn results used in the theoretical predictions for identification of CSH probes (mouse probes used as an example).

Gold Standard	Parameters
GS1.x	Top x ratio of relative expression (red/green) when exposed to erroneous RNA vs to correct RNA:
GS2.x*	Top x% absolute expression when exposed to human RNA with mouse probes
GS3.x	Top x% relative expression when exposed to human RNA with mouse probes (the expression of mouse probes when exposed to human RNA divided by the expression when exposed to mouse RNA)
GS4.x	The same as GS2 but absolute log2 based expression on both green channels of human and mouse RNA expression of mouse probes > 4
Models	Conditions and parameters
Model1*	{18<Align≤20 and mis<5} U {20<Align≤ 30 and mis<6} U {30<Align≤40 and mis<7} U {40<Align≤50 and mis<8} U {50<Align and mis<9}
Model2	{15 < Align ≤ 30 and mis < 2} U {30 < Align ≤ 40 and mis < 3} U {40 < Align ≤ 50 and mis < 5} U {50< Align and mis < 7}
Model3	Align > 50 & mismatches < 6
Model4	score > 70 & Align > 58 & mis < 6

Legend: biological GS: biological gold standard; CSH: cross-species hybridization; Align: alignment; mis: mismatch; U: union.

The Gold Standards used in the final design are annotated with an asterisks (*). These were selected by comparing the initial biological experimental data and BLAST predictions using H&M array version 1.0 and details of this section are provided in our previous publication [17].

species-specific expression after Variance Stabilization Normalization (VSN) normalization[25], the optimal biological GS was determined using both the relatively higher F-scores and the absolute highest recall.

• **Accuracy calculation:** True positive CSH probes were those that cross-species hybridized in the biological GS and were also predicted by BLAST to do so. False positives were those that were predicted by BLAST but were not found in the GS. The precisions, recalls (sensitivities) and F-scores (*F*) were calculated for each theoretical model of cross-hybridization against each biological GS at different thresholds of biologically cross-hybridizing probes (**Equation 1**).

$$F = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (1)$$

• **Identification of cross-hybridizing probes** (Table 3): Once the optimal set of BLAST prediction parameters and biological GSs were determined, all probes from the GS that were considered as CSH were tagged and new SSPs were designed - regardless of whether they were predicted as cross-hybridizing by BLAST. Additionally, probes were tagged for future design when they did not have any expression on the array and were predicted as cross-hybridising using BLASTn program (Table 3: Model 1-1).

Step B4: probe redesign (green boxes in Figure 1)

I. Gene targets associated with cross species hybridizing probes were identified for redesign. Transcripts (RefSeq

IDs) associated with each probe were identified from the Agilent microarray annotation file (Additional file 1- **Supp. Table S1**, resources 1-2). For those probes with a retired Refseq or Unigene ID, the batch BLASTn program [24] was used to retrieve the associated antisense Refseq or Unigene sequences. The transcripts targeted by these probes were downloaded from the NCBI GEO database via accession numbers GPL6840 and GPL7202.

II. New probes for the identified gene targets were designed using the “GE Probe Design” program in the Agilent eArray (Agilent, Santa Clara, CA) software. The eArray parameters, that we used, include probe length (60nt), probes per target (1), probe orientation (sense) and design options (“best probe methodology” and “design with 3’ bias”). These probes were further verified computationally for cross-hybridization.

III. Newly designed probes were evaluated by the eArray “Probe Check” program using the built-in transcriptome. The new candidate probes were retained if they did not hybridize within their species with other targets than the intended one nor across species with any other transcripts. Otherwise, these candidate probes were tagged as inappropriate and new probes were designed for the intended target iteratively until they met these criteria.

IV. All probes of the commercial arrays were retained, while new probes were also added to the H&M array. Annotation tables of cross species hybridization were used as filters to identify the most SSP for each gene, while the legacy probes were intended to allow for comparison of results with studies using the commercial platforms.

Table 3 Definition of two sets of parameters used in the optimal CSH theoretical predictions for identification of CSH probes.

	Conditions	Predicted # of CSH probes for vers.1 mouse probes	Predicted # of CSH probes for vers.1 human probes
Model 1-1	{18<Alignment<=20 and mismatches<6} UNION {20<Alignment<=30 and mismatches<7} UNION {30<Alignment<=40 and mismatches<8} UNION {40<Alignment<=45 and mismatches<9} UNION {45<Alignment<=50 and mismatches<10} UNION {50<Alignment and mismatches<11}	3819	4755
Model 1-2	{30<Alignment<=40 and mismatches<8} UNION {40<Alignment<=45 and mismatches<9} UNION {45<Alignment<=50 and mismatches<10} UNION {50<Alignment and mismatches<11}	3525	4370

Computational optimization results shown in Additional file 1 - Supp. Figure S1.

Step B5: cross-hybridization experiment and extensive annotation of the H&M array version 2

To validate the newly designed probes and extensively annotate the H&M array version 2, a second biological experiment was conducted by hybridizing all the probes on the H&M array version 2 to three human melanoma cell lines (MDA-MB-435) and three universal mouse mRNAs. Six Samples (three with human RNA and three with mouse RNA) were hybridized to the H&M array version 2.0 (Table 1). The subsequent array scanning and image processing followed the same procedure as the initial biological experiment. Additionally, the background-subtracted expression measurements of six arrays were normalized using VSN method [25] and were log₂ transformed. Human and mouse probe sets were separately normalized to control for variation in the proportion of human and mouse RNA across xenograft replicates [16]. For each probe, the mean signal intensity of the three biological replicates exposed to the RNAs of the same species was used as the expression value for that species. GS and BLAST parameter optimizations were performed as described in Step 4 and in our previous publication [17], and the propensity for CSH of each probe was annotated (**Dataset S1**). This extensive annotation of CSH was conducted by systematically extending the threshold of the biological GS in 10% increments from 0% to 100% for all probes.

Stromal cell to cancer cell mRNA ratio estimation (Figure 1, step C-1)

Housekeeping genes are typically constitutively expressed genes required for the maintenance of basal cellular functions, and are thus assumed to be stably expressed under the same experimental conditions as well as between human tissues and mouse tissues [22,23]. To estimate the RNA expression ratio between human and mouse using RNA quantity controlled H&M array experiments, the expression value of human SSPs targeting human housekeeping genes was divided by the expression value of mouse SSPs targeting homologous housekeeping genes.

The SSPs satisfied the following four criteria: 1) they were positively expressed (expressed intensity larger than 1, see Method: Step B1) across all samples when exposed to the intended RNA, 2) among all probes for one species, they expressed below a certain percentage of all expression values when exposed to the RNA of the other species, 3) they were not identified as CSH probes, and 4) were not theoretically predicted as CSH probes by the BLASTn algorithm. **For simplicity**, a set of probes classified as “species-specific” in the 2nd criteria that passed the threshold x are **hereafter named “SSP x ”**. Different thresholds were tested, and the optimal threshold was selected according to the lowest variance of resulting RNA ratios.

Stromal cell to cancer cell mRNA ratio and its validation (Figure 1, step C-1)

Evaluation of the stromal cell to cancer cell mRNA ratio, hereafter named as S/C ratio for simplicity, was independently carried in two steps. The samples of the *in vitro* biological experiments data for which the proportion of mouse to human RNA is designed and known (GSE23377) were used to validate the mouse to human RNA ratio metric. This metric then serves as a proxy for stroma to cancer cell RNA ratio in xenograft models consisting of human cancer cell tumors harvested from mice hosts.

Gene ontology enrichment of cross-species hybridizing probes (Figure 1, step C-2) and visualization

After masking the top 5% of the most likely probes to exhibit CSH, KEGG/GO enrichment analysis was performed among the genes targeted by probes on the custom array using the conditional hypergeometric test in the Bioconductor *GOSTats* package[26]. For comparison, the same pathway/GO enrichment assessment was performed on the commercial CSH probes, for the human and mouse species separately. The resulting p-values were adjusted for multiple testing by FDR[27] and pathways/GO terms with less than 500 gene members were counted [28]. Hexagon plots were conducted using the R package *hexbin* with default parameters.

Results

Human tumor and mouse xenograft cross-species hybridization of commercial arrays

Our previous study reported interspecies genetic differences and elucidated stromal microenvironment signals from probes on human arrays unintentionally cross-hybridizing with mouse homologous genes in xenograft tumor models (Figures 1, 3, 4 in the publication [17]). By identifying CSH probes from sequence alignment and CSH experiment for the human whole-genome arrays, deregulated stromal genes can be identified and then their biological significance confirmed by the laser capture microdissection of stromal cells from tumor specimens [17]. In this study, within Agilent 44k whole genome human arrays and mouse arrays, 4,259 commercial probes (2,104 human, 2,155 mouse) pertaining to 3,550 distinct genes cross-hybridized above our optimal threshold (highest 5% CSH) were annotated as cross-species cross-hybridizing probes.

Design of the dual species H&M array

Using our computational cross-species hybridization algorithms and the initial biologic experiment (**Methods**, 5% threshold CSH), we identified 4,259 commercial probes as cross-hybridizing (2,104 human, 2,155 mouse) corresponding to 3,550 distinct genes. Therefore, we designed a

custom 2x105k array (H&M array, Agilent custom platform) combining all probes from both the human and mouse Agilent 44k commercial whole genome arrays (human: 41,000 probes; mouse: 41,174 probes). In addition, the H&M array contains new species-specific probes we designed: 5.5k human and 5.2k mouse (Additional file 1- **Supp. Table S2**), covering 30k distinct human targets and 32k mouse targets (Refseq IDs), respectively.

Analysis and annotation of the H&M arrays

A total of 87k probes were included on the H&M platform version 1, containing all commercial human and mouse probes, and 3,608 newly custom designed probes (Additional file 1- **Supp. Table S2**). Annotation for version 1 of the array was conducted by comparing the results from the initial *in vitro* biological experiment with the CSH theoretical predictions of probes on the H&M array. The optimal biological GS of CSH as observed in the *in vitro* experiments was defined by comparing the optimal CSH theoretical prediction model with the most ideal parameters using a threshold of 18% for mouse probes [17] and 20% for human probes. Meanwhile, the optimal parameters for BLASTn result interpretation were selected as previously published (Figure D in Supplementary Methods of our previous publication [17]), where the sequence alignment derived from BLAST depended on the length of the alignment (varying from 1 to 54) and the BLAST mismatch number (varying from 1 to 12).

We further sought to incorporate theoretically predicted and experimentally verified CSH probes to the array. As illustrated in Figure 1 and Additional file 1- **Supp. Table S2**, H&M array version 2 included 7.1k new custom probes, containing 5k probes designed from experimentally annotated CSH probes in version 1 of the platform, in addition to 2.1k probes that the Agilent protocol suggested as acceptable but whose theoretical prediction resulted in CSH. Together with the 3.6k custom probes designed in version 1 array, the total 10.7k custom probes covered 16% (5.3k) of human genes and 15% (5.1k) of mouse genes (Additional file 1- **Supp. Table S2**). The set of BLAST prediction parameters derived from the initial *in vitro* biological experiment (**Methods, step B3**) performed better in predicting CSH probes than a similar set of parameters with shorter mismatch lengths (**Methods, step B2**). Two GSs for CSH (Table 2: GS2 and GS3) performed similarly in the initial *in vitro* experiment [17], making it necessary to validate the initial experiment. We found that biological GS3 that yielded a higher F-score than biological GS2. The former used relative CSH expression values (i.e. the ratio of the absolute CSH expression when exposed to the RNA of the other species compared with the expression when exposed to the RNA of the same species), while the latter used absolute CSH expression values due to the larger number of genes with

multiple probes in version 2 of the H&M array. A total of 4,644 probes on the H&M array (version 2) were annotated as CSH, which comprised the top 5% relatively erroneously expressed (relative CSH) probes achieving the highest recall when compared with CSH theoretical prediction in Model 1-1 (Additional file 1- **Supp. Table S2**). This threshold of 5% was derived from H&M array version 1 probes, as version 2 included all version 1 probes. The estimated recall for mouse CSH probes was 25.6% (i.e. 25.6% probes were truly CSH among all annotated "CSH" mouse probes) and precision was 13% (13% of all theoretically predicted CSH mouse probes are annotated by our biological GS), while the annotated recall and precision was 27% and 11%, respectively, for human CSH probes. Additionally, our extensive annotation stratified all probes on H&M array version 2 into ten levels according to their probability of CSH, thus providing array users with more flexibility to disregard CSH probes in their analyses.

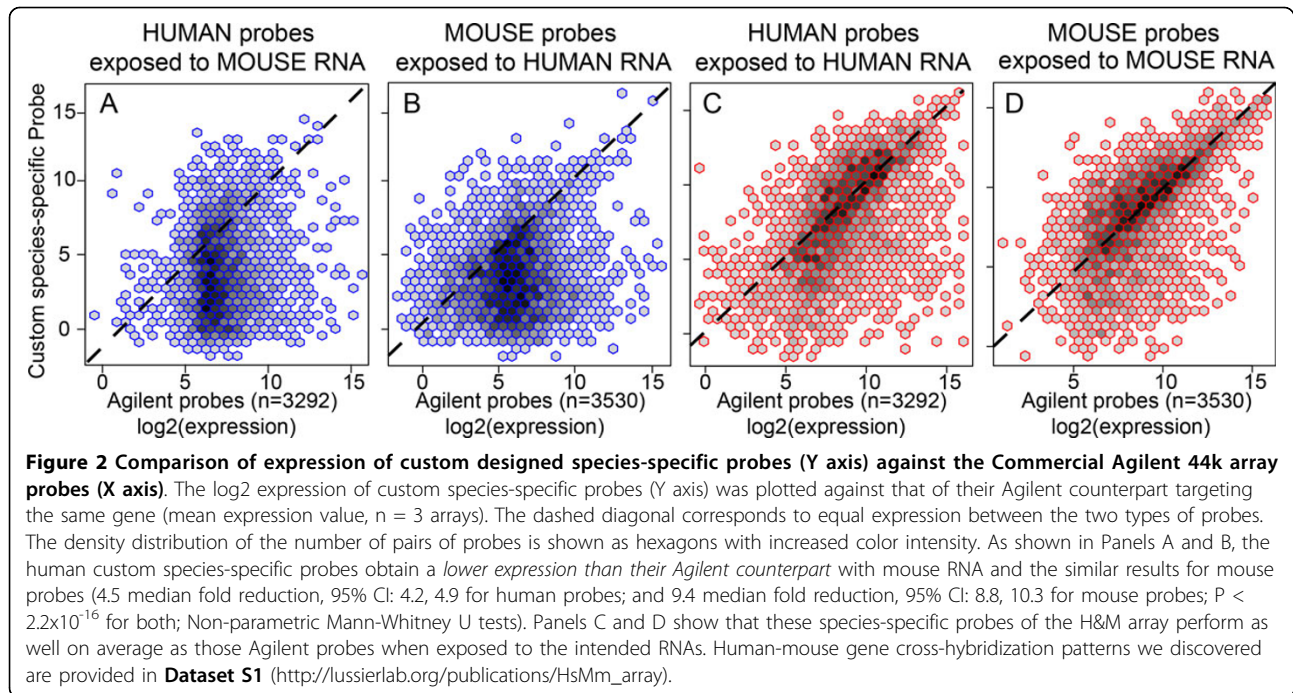
Validation of H&M array version 1 annotation and reduction of CSH

The reduction in CSH was further shown by an analyses illustrated in Figure 2, which shows the comparison of version 1 custom designed 3,292 human probes and 3,530 mouse probes (y-axis) with the pair-wise probes on the Agilent commercial array (x-axis) for cross-species expression using the validating biological experiment. In contrast, when exposed to the RNA of the correct species, there was no significant difference between custom newly designed probes and the corresponding probes on the Agilent 44k whole genome array (Figure 2, **panels C-D**). Additionally, version 1 annotated CSH probes were significantly over-represented among the version 2 annotated CSH probes ($p < 2 \times 10^{-16}$, OR = 3.3). Here, the overlap between the annotated CSH probes demonstrates the robustness of our design to identify the CSH probes in a xenograft model.

Estimation of stromal cell to cancer cell mRNA ratio

In an *in vitro* validation experiment for stromal to cancer cell ratio estimation, H&M arrays were separately exposed to human and mouse RNAs (Table 1). The ratios of the normalized expression between human SSPs and mouse SSPs targeting homologous housekeeping genes listed in Figure 3 were calculated.

As shown in Figure 3, the expression values of human probes exposed to only human RNA were larger than when exposed to only mouse RNA (Figure 3, **panel A**, median ratio between human and mouse probes is 10.4), and as shown in **Panel B**, the expression of mouse probes exposed to only mouse RNA was larger than that when exposed to only human RNA (median ratio between mouse and human probes is 9.7). Changing the thresholds to define an individual probe as species-specific



(**Method**, Step C-1) to include from 10 to 50 pairs of human and mouse SSPs provided a similar result, however the expression ratio between human and mouse probes tended to be unchanged when all non-SSP were included (data not shown). These results suggested that the proportion of each species RNA can be estimated by the species-specific human and mouse probes of the pair-wise homologous housekeeping genes.

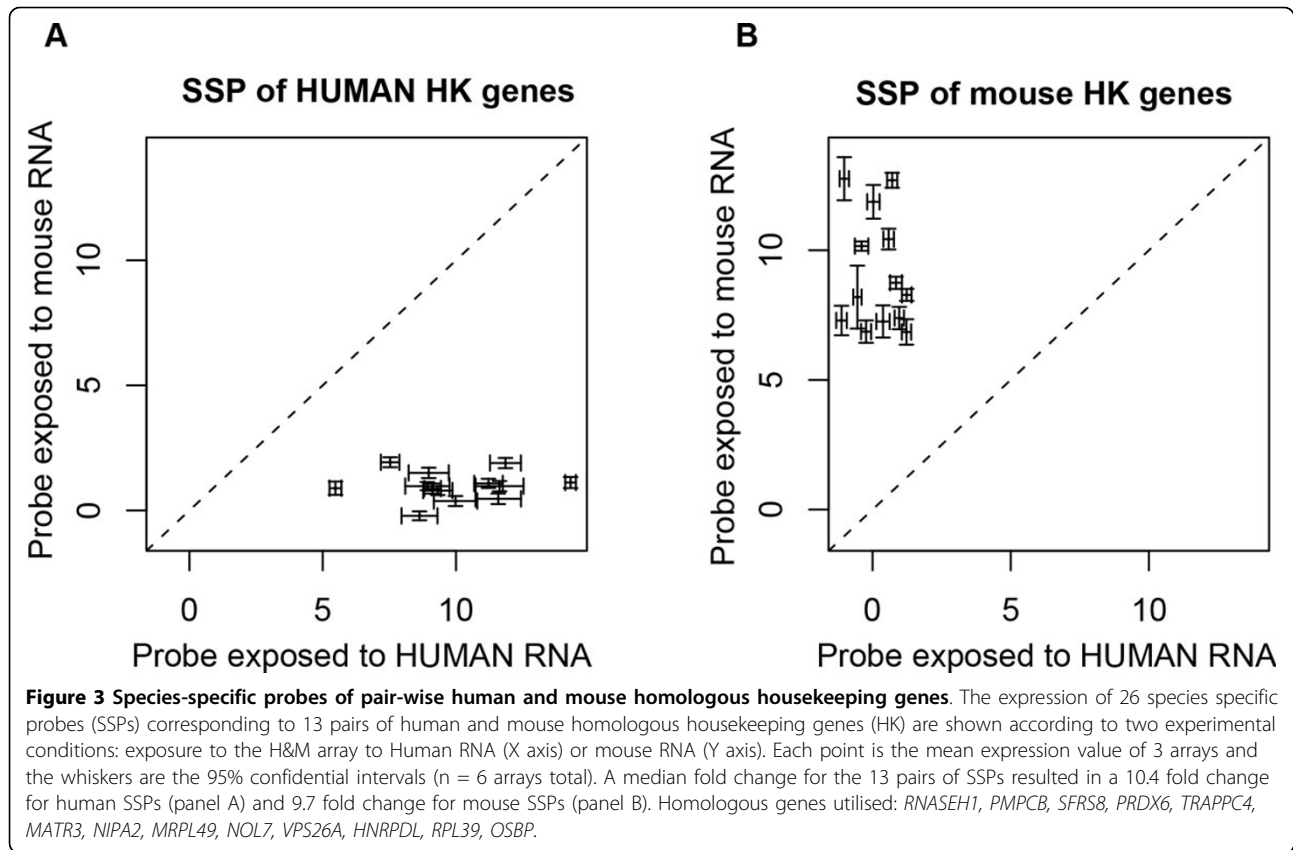
Stromal cells to cancer cell mRNA ratio (S/C ratio) validation

Validations for our S/C ratio estimation was performed with an *in vitro* biological experiment, where the stromal cell proportions were controlled using known quantity of RNAs as shown in Table 1. Using the *in vitro* CSH experimental data, where two arrays containing both human and mouse probes were exposed to equal quantities of human and mouse RNAs (**GSE23377**), the ratios of 13 pairs of homologous housekeeping gene expression were larger than 10-fold when exposed to the intended RNA of one species only (Figure 4, **panels A,B**), while was between 0 and 10 when exposed to an equal mix of human and mouse RNAs simultaneously (Figure 4, **panel C**, Table 1). The higher expression of mouse SSPs in panel B indicated a better mouse RNA quality than human RNA.

GO and KEGG pathway enrichment analysis of cross-species hybridizing probes

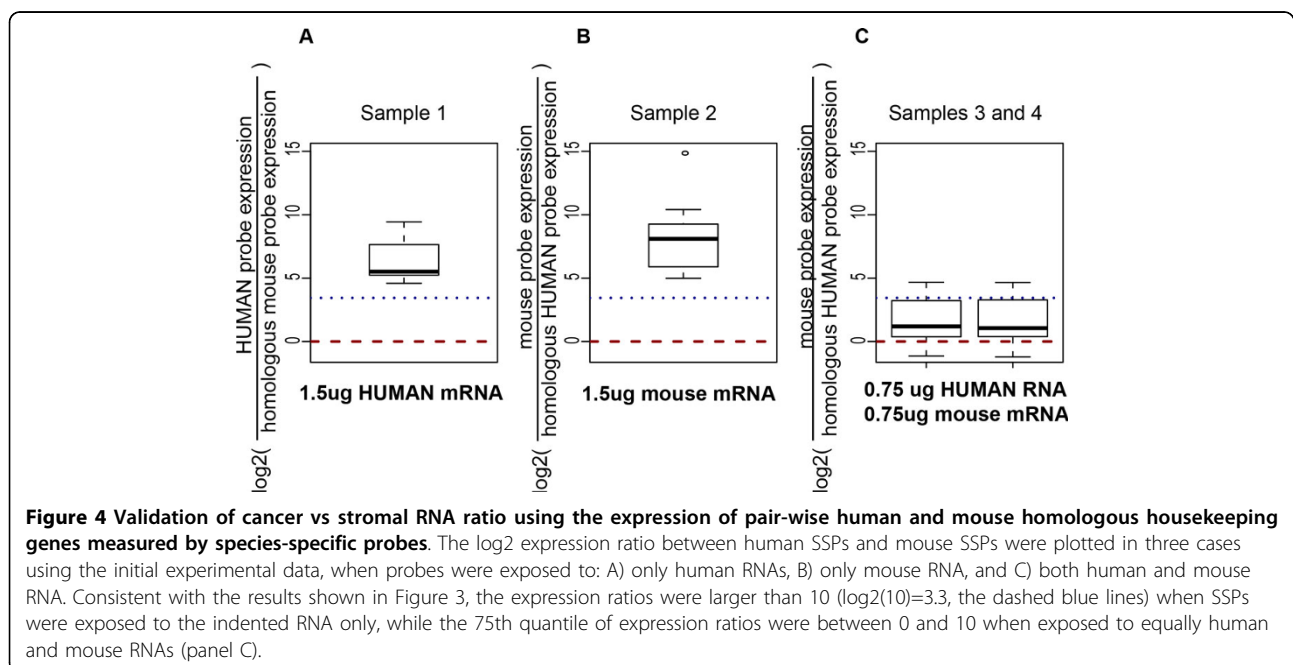
To control for CSH of probes in GO and KEGG pathway enrichment analyses, we masked the 2,324 CSH

human probes on the custom H&M array, and observed no KEGG pathways and only 1 GO molecular function (*RNA binding* including more than 600 gene members) significantly (FDR < 5%) over-represented among the 44,144 human SSPs (targeting 30k Refseq IDs, 97% coverage of Agilent whole human genome array). We found no bias in GO or KEGG for the mouse probes of the custom H&M array. Notably, gene ontology and pathway biases exist among the CSH probes on commercial single-species arrays, and should be accounted for on-going and future xenograft tumor studies. Here we are the first to report comprehensive analysis of CSH probes between human cancer cells avoiding these biases. We identified eight significantly enriched (FDR < 5%) KEGG pathways (seven human genes and one mouse gene targeted only by CSH commercial probes), eight GO molecular functions (all among human genes), and conclusively, stromal functions detected when using a dual species array for xenograft tissues due to gene expression in mouse stromal tissues, rather than in human cancer cells (data not shown). Further, CSH probes confound the results of microarray analyses of the metabolism pathway of xenobiotics by cytochrome P450 when xenograft tumor models were used due to the highly conserved nature of many cytochrome P450 genes (data not shown). Using Agilent 44k human arrays, Sugawara *et al.* reported 28 genes demonstrating higher induction in xenograft tumors compared to the cells in culture medium[15]. Our independent analysis revealed that 10 of these 28 previously



reported genes are targeted by commercial CSH probes (data not shown). Additionally, we observed that the cytochrome P450 pathway involved in xenobiotic metabolism is enriched (cumulative hypergeometric pvalue = 1×10^{-5})

in the CSH probes found in the list of 28 genes (data not shown), as would be expected for our xenograft model. Other significant pathways and functions listed that suggest relevant stromal functions include inflammatory



signalling through the complement system, focal adhesion and tight junctions involving transmembrane proteins mediating intra-membrane and paracellular diffusion, as well as motor activity and actin binding, and thus possibly mediating invasion and metastasis processes. These results indicated that the design of dual-species arrays and the identification of CSH probes are critically important for studies involving highly conserved genes and pathways in xenograft models.

Discussion

We established that overlooking CSH probes in xenograft models using commercial human arrays confound incorrect attribution of deregulated stromal genes to cancer cells. To increase the capability of analysing molecular changes in the stromal compartment as well as the interactions between cancer cells and reactive tumor stroma in xenograft modelling, we designed the first custom genome-wide human-mouse microarray, and biologically validated it by observing an overall reduction of the CSH (4.5 median fold reduction for human probes; and 9.4 median fold reduction, for mouse probes; $p < 2 \times 10^{-16}$, Mann-Whitney test). By design, we select a framework of well-established computational and biological methods to identify cross-hybridizing probes, construct a cross-hybridizing map, and identify new SSPs. Here, we provide a combined description of the interplay between the computational predictions of CSH of probes using BLASTn and observation of biological CSH that were utilized to develop an approach stratifying probes in terms of their propensities for cross-hybridization. We also implemented a rigorous rationale for the selection of a threshold needed to redesign the cross-hybridizing probes and offer the foundational datasets used for conducting analyses over a dual-species array: a comprehensive map of cross-hybridizations annotated with both sequence alignment and biological data intended to increase the accuracy and efficiency of measurements in tumor xenograft models. Additionally, we developed a novel method to estimate the mRNA ratio between mouse and human mRNA as a cellularity ratio, thus providing an alternative to conventional histological counts of stroma vs. cancer cells in xenograft models. Of note, the H&M array could also be utilized in single species experiments, as long as the probes of that species are selected. This information is clearly available in the “species” column of **Dataset S1**. Further, for comparison of H&M array results against previously published dataset using single species commercial Agilent arrays, the H&M probes corresponding to the commercial platforms are annotated in the column “designer” of **Dataset S1**.

One limitation of the current study is that grouping all of the stromal cell types into one category may be too

strong of a simplification of the complex cellularity of the host tissue. For instance, changes in the presence of certain types of stromal cells have profound effects on the status of the tumor, and stromal cells are thought to genomically “co-evolve” along with cancer cells during tumor progression [29]. Also, the xenograft model involves immune deficient mice, and therefore cannot be used to understand much about the stromal immune cells and associated pathways that participate in regulating the tumor in the xenograft models. However, xenograft models re-constituting bone marrow cells from syngeneic donors may alleviate this limitation in future studies. Further, with the advent of RNA-sequencing, the utility of the present study may reside in our supplemental datasets that allow to reanalyze the species-specific probes of commercial arrays for comparison with RNA-seq experiments or to inform assembly algorithms of RNA-seq in xenograft models. Another limitation is that we have not considered non-coding RNAs in this study.

Conclusion

There is ample indication that the issue of CSH is overlooked in xenograft modelling, as the majority of tumor xenograft studies utilize gene expression over human arrays indiscriminately with no correction for homologous probes [30-35], leading to the incorrect attribution of deregulated stromal genes to cancer cells. Furthermore, commercial human and mouse arrays are shown to contain a large number CSH probes, suggesting that commercial designs did not consider the potential reuse of these arrays for xenograft modelling. While the computational methods and biological validations are straightforward and rely on well-established methods, the originality of this study resides in (i) appropriately computationally modelling the deceptively simple source of CSH (homologous genes) - a problem that has nonetheless been overlooked by microarray corporations and by many xenograft biologists in their array-related publications, in (ii) providing well-validated and novel work products: a detailed cross-hybridization map of human and mouse probes, as well as a cost-effective and superior array for expression analyses of human xenograft models, and in (iii) offering a new *in silico* metric for estimating the cellularity ratio of cancer over stroma that should in principle be applicable to RNAseq transcriptomes as well. This study's findings will improve our capability of investigating and interpreting the interactions between human cancer and mouse stromal cells in xenograft modelling. Together, these technological improvements will allow effective interrogation of cancer cell-reactive stroma interactions during cancer initiation, progression and response to therapeutic interventions.

Additional material

Additional file 1: Supplement Tables and Figures. This document contains the two supplement tables (Tables S1, S2) and two supplement figures (Figures S1, S2). • Table S1 - Data resources used in this study. • Table S2 - Summary of probes on H&M array version 2 • Figure S1 - Optimization of Models identifying cross-species hybridizing probes. • Figure S2 - BLAST parameter selection based on higher F-scores.

Abbreviations

LCM, laser capture microdissection; CSH, Cross-Species Hybridization; GO, Gene Ontology; H&M array, Human-Mouse dual-species array; KEGG, Kyoto Encyclopedia of Genes and Genomes; GS, Gold Standard; SSP, Species Specific Probes; HK, housekeeping genes; FDR: False Discovery Rate; CI, Confidence Interval; BLAST, Basic Local Alignment Search Tool.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

YAL supervised the entire work; XY designed and conducted the analyses and annotation of data; YH conducted the online custom Agilent array design; YL performed the BLASTn programming; YAL, YH and ER contributed to the biological interpretation of the results; YAL, XY, HY, ER, VG and AI wrote and revised the manuscript.

Acknowledgements

We thank Dr. Scott J. Vacha from Agilent providing the initial list of potential CSH probes.

Declarations

Publication for this article has been funded by grants 1U54CA121852, UL1 RR024999-03, K22 LM008308, R21 CA167305-01A1. This work was supported in part by the UA Cancer Center grant (NCI P30CA023074), the UA BIO5 Institute, and the UA Clinical & Translational Science Institute of the University of Arizona. This article has been published as part of *BMC Medical Genomics* Volume 7 Supplement 1, 2014: Selected articles from the 3rd Translational Bioinformatics Conference (TBC/ISCB-Asia 2013). The full contents of the supplement are available online at <http://www.biomedcentral.com/bmcmedgenomics/supplements/7/S1>.

Authors' details

¹Ctr for Biom. Info. Section of Genetic Medicine, Dept. of Medicine, U. of Chicago, IL, USA. ²Institute for Translational Health Informatics, U. of Illinois at Chicago, IL, USA. ³Department of Informatics, EISTI engineering school, Cergy, France. ⁴Dept. of Medicine, U. of Illinois at Chicago, IL, USA. ⁵Comprehensive Cancer Ctr and Ludwig Ctr for Metastasis Research, U. of Chicago, IL, USA. ⁶Section of Hematology/Oncology, Department of Pediatrics, U. of Chicago Medicine Comer Children's Hospital, IL, USA. ⁷Depts of Bioengineering & of Pharmaceutical Science, U. of Illinois at Chicago, IL, USA. ⁸Comp. Inst. & Inst. for Genomics & Syst. Bio., Argonne Nat. Lab. and U. of Chicago, IL, USA. ⁹Inst. for Personalized Respiratory Medicine, U. of Illinois at Chicago, IL, USA. ¹⁰This work was conducted in part while at The University of Chicago, IL, USA. ¹¹Cancer Ctr, BIO5 Inst., Clinical & Translational Science Inst., U. of Arizona, AZ, USA.

Published: 8 May 2014

References

- Egeblad M, E. LL: **The fibroblastic coconspirator in cancer progression.** *Cold Spring Harb Symp Quant Biol* 2005, **70**:383-388.
- Jodele S, Blavier L, Yoon JM, DeClerck YA: **Modifying the soil to affect the seed: role of stromal-derived matrix metalloproteinases in cancer progression.** *Cancer metastasis reviews* 2006, **25**(1):35-43.
- Sloane BF, Yan S, Podgorski I, Linebaugh BE, Cher ML, Mai J, Cavallo-Medved D, Sameni M, Dosescu J, Moin K: **Cathepsin B and tumor**

- proteolysis: contribution of the tumor microenvironment.** *Seminars in cancer biology* 2005, **15**(2):149-157.
- Overall CM, Kleifeld O: **Tumour microenvironment - opinion: validating matrix metalloproteinases as drug targets and anti-targets for cancer therapy.** *Nature reviews Cancer* 2006, **6**(3):227-239.
- Finak G, Sadekova S, Pepin F, Hallett M, Meterissian S, Halwani F, Khetani K, Souleimanova M, Zabolotny B, Omeroglu A, Park M: **Gene expression signatures of morphologically normal breast tissue identify basal-like tumors.** *Breast cancer research: BCR* 2006, **8**(5):R58.
- Finak G, Bertos N, Pepin F, Sadekova S, Souleimanova M, Zhao H, Chen H, Omeroglu G, Meterissian S, Omeroglu A, Hallett M, Park M: **Stromal gene expression predicts clinical outcome in breast cancer.** *Nature medicine* 2008, **14**(5):518-527.
- Albini A, Sporn MB: **The tumour microenvironment as a target for chemoprevention.** *Nature reviews Cancer* 2007, **7**(2):139-147.
- Barth PJ, Westhoff CC: **CD34+ fibrocytes: morphology, histogenesis and function.** *Current stem cell research & therapy* 2007, **2**(3):221-227.
- Amatangelo MD, Bassi DE, Klein-Szanto AJ, Cukierman E: **Stroma-derived three-dimensional matrices are necessary and sufficient to promote desmoplastic differentiation of normal fibroblasts.** *The American journal of pathology* 2005, **167**(2):475-488.
- Micke P, Ostman A: **Exploring the tumour environment: cancer-associated fibroblasts as targets in cancer therapy.** *Expert opinion on therapeutic targets* 2005, **9**(6):1217-1233.
- Emmert-Buck MR, Bonner RF, Smith PD, Chuaqui RF, Zhuang Z, Goldstein SR, Weiss RA, Liotta LA: **Laser capture microdissection.** *Science* 1996, **274**(5289):998-1001.
- Krop IE, Sgroi D, Porter DA, Lunetta KL, LeVangie R, Seth P, Kaelin CM, Rhei E, Bosenberg M, Schnitt S, Marks JR, Pagon Z, Belina D, Razumovic J, Polyak K: **HIN-1, a putative cytokine highly expressed in normal but not cancerous mammary epithelial cells.** *Proc Natl Acad Sci USA* 2001, **98**(17):9796-9801.
- Ge H, Gong X, Tang CK: **Evidence of high incidence of EGFRvIII expression and coexpression with EGFR in human invasive breast cancer by laser capture microdissection and immunohistochemical analysis.** *Int J Cancer* 2002, **98**(3):357-361.
- Glockner S, Lehmann U, Wilke N, Kleeberger W, Langer F, Kreipe H: **Detection of gene amplification in intraductal and infiltrating breast cancer by laser-assisted microdissection and quantitative real-time PCR.** *Pathobiology: journal of immunopathology, molecular and cellular biology* 2000, **68**(4-5):173-179.
- Sugawara M, Okamoto K, Kadowaki T, Kusano K, Fukamizu A, Yoshimura T: **Expressions of cytochrome P450, UDP-glucuronosyltransferase, and transporter genes in monolayer carcinoma cells change in subcutaneous tumors grown as xenografts in immunodeficient nude mice.** *Drug metabolism and disposition: the biological fate of chemicals* 2010, **38**(3):526-533.
- Schwartz DR, Moin K, Yao B, Matrisian LM, Coussens LM, Bugge T, Fingleton B, Acuff H, Sinnamon M, Nassar H, Platts AE, Krawetz SA, Linebaugh BE, Sloane BF: **Hu/Mu Protn oligonucleotide microarray: dual-species array for profiling protease and protease inhibitor gene expression in tumors and their microenvironment.** *Mol Cancer Res* 2007, **5**(5):443-454.
- Yang X, Lee Y, Huang Y, Chen JL, Xing RH, Lussier YA: **Stromal microenvironment processes unveiled by biological component analysis of gene expression in xenograft tumor models.** *BMC bioinformatics* 2010, **11**(Suppl 9):S11.
- Schwartz DR, Moin K, Weber E, Sloane BF: **The CLIP-CHIP™: A Focused Oligonucleotide Microarray Platform for Transcriptome Analysis of the Complete Human and Murine Cancer Degradomes.** *The Cancer Degradome*. D E, G H-H, F B, BF S; 2008, 17-35.
- Jabado OJ, Liu Y, Conlan S, Quan PL, Hegyi H, Lussier Y, Briese T, Palacios G, Lipkin WI: **Comprehensive viral oligonucleotide probe design using conserved protein regions.** *Nucleic acids research* 2008, **36**(1):e3.
- Liu Y, Sam L, Li J, Lussier YA: **Robust methods for accurate diagnosis using pan-microbiological oligonucleotide microarrays.** *BMC bioinformatics* 2009, **10**(Suppl 2):S11.
- Palacios G, Quan PL, Jabado OJ, Conlan S, Hirschberg DL, Liu Y, Zhai J, Renwick N, Hui J, Hegyi H, Grolla A, Strong JE, Towner JS, Geisbert TW, Jahrling PB, Buchen-Osmond C, Ellerbrok H, Sanchez-Seco MP, Lussier Y, Formenty P, Nichol MS, Feldmann H, Briese T, Lipkin WI: **Panmicrobial**

- oligonucleotide array for diagnosis of infectious diseases. *Emerg Infect Dis* 2007, **13**(1):73-81.
22. de Jonge HJ, Fehrmann RS, de Bont ES, Hofstra RM, Gerbens F, Kamps WA, de Vries EG, van der Zee AG, te Meerman GJ, ter Elst A: **Evidence based selection of housekeeping genes.** *PLoS one* 2007, **2**(9):e898.
 23. Robinson MD, Oshlack A: **A scaling normalization method for differential expression analysis of RNA-seq data.** *Genome Biol* 2010, **11**(3):R25.
 24. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *Journal of molecular biology* 1990, **215**(3):403-410.
 25. Huber W, von Heydebreck A, Sultmann H, Poustka A, Vingron M: **Variance stabilization applied to microarray data calibration and to the quantification of differential expression.** *Bioinformatics* 2002, **18**(Suppl 1): S96-104.
 26. Falcon S, Gentleman R: **Using GOSTats to test gene lists for GO term association.** *Bioinformatics* 2007, **23**(2):257-258.
 27. Benjamini Y, Hochberg Y: **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.** *Journal of the Royal Statistical Society Series B (Methodological)* 2005, **57**:289-300.
 28. Dinu I, Potter JD, Mueller T, Liu Q, Adewale AJ, Jhangri GS, Einecke G, Famulski KS, Halloran P, Yasui Y: **Improving gene set analysis of microarray data by SAM-GS.** *BMC bioinformatics* 2007, **8**:242.
 29. Weinberg R: **Stromal cells are active contributors to tumorigenesis.** *The Biology of Cancer*. 2007, 548-551.
 30. Johannsson OT, Staff S, Vallon-Christersson J, Kytola S, Gudjonsson T, Rennstam K, Hedenfalk IA, Adeyinka A, Kjellen E, Wennerberg J, Baldetorp B, Petersen OW, Olsson H, Oredsson S, Isola J, Borg A: **Characterization of a novel breast carcinoma xenograft and cell line derived from a BRCA1 germ-line mutation carrier.** *Laboratory investigation; a journal of technical methods and pathology* 2003, **83**(3):387-396.
 31. Morimoto AM, Tan N, West K, McArthur G, Toner GC, Manning WC, Smolich BD, Cherrington JM: **Gene expression profiling of human colon xenograft tumors following treatment with SU11248, a multitargeted tyrosine kinase inhibitor.** *Oncogene* 2004, **23**(8):1618-1626.
 32. Khodarev NN, Roach P, Pitroda SP, Golden DW, Bhayani M, Shao MY, Darga TE, Beveridge MG, Sood RF, Sutton HG, Beckett MA, Mauceri HJ, Posner MC, Weichselbaum RR: **STAT1 pathway mediates amplification of metastatic potential and resistance to therapy.** *PLoS one* 2009, **4**(6):e5821.
 33. Hendriksen PJ, Dits NF, Kokame K, Veldhoven A, van Weerden WM, Bangma CH, Trapman J, Jenster G: **Evolution of the androgen receptor pathway during progression of prostate cancer.** *Cancer research* 2006, **66**(10):5012-5020.
 34. Chen CD, Welsbie DS, Tran C, Baek SH, Chen R, Vessella R, Rosenfeld MG, Sawyers CL: **Molecular determinants of resistance to antiandrogen therapy.** *Nature medicine* 2004, **10**(1):33-39.
 35. Morisot S, Wayne AS, Bohana-Kashtan O, Kaplan IM, Gocke CD, Hildreth R, Stetler-Stevenson M, Walker RL, Davis S, Meltzer PS, Wheelan SJ, Brown P, Jones RJ, Shultz LD, Civin CI: **High frequencies of leukemia stem cells in poor-outcome childhood precursor-B acute lymphoblastic leukemias.** *Leukemia* 2010.

doi:10.1186/1755-8794-7-S1-S2

Cite this article as: Yang et al.: *In Silico* cancer cell versus stroma cellularity index computed from species-specific human and mouse transcriptome of xenograft models: towards accurate stroma targeting therapy assessment. *BMC Medical Genomics* 2014 **7**(Suppl 1):S2.

**Submit your next manuscript to BioMed Central
and take full advantage of:**

- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

