www.sciencetranslationalmedicine.org/cgi/content/full/6/217/217ra2/DC1

Supplementary Materials for

Silencing *HoxA1* by Intraductal Injection of siRNA Lipidoid Nanoparticles Prevents Mammary Tumor Progression in Mice

Amy Brock, Silva Krause, Hu Li, Marek Kowalski, Michael S. Goldberg, James J. Collins, Donald E. Ingber*

*Corresponding author. E-mail: don.ingber@wyss.harvard.edu

Published 1 January 2014, *Sci. Transl. Med.* **6**, 217ra2 (2014)
DOI: 10.1126/scitranslmed.3007048

The PDF file includes:

SUPPLEMENTARY METHODS

Microarray gene expression analysis

Tumor gene expression profiling was carried out using Mouse Genome MG 430 2.0 Affymetrix GeneChip arrays. In brief, total RNA was extracted from tissue with the RNeasy kit including DNase digestion (Qiagen). Biotinlabeled cDNA was obtained from 1 μg total RNA with the GeneChip One-Cycle labeling kit (Affymetrix). According to the manufacturer's instructions, 1.5 μg of cDNA were fragmented and hybridized to Affymetrix MG 430 2.0 GeneChip arrays. Arrays were processed by the Harvard Genotyping and Microarray Center. DNA chips were washed, stained and scanned using an Affymetrix Fluidics device and a GCS3000 scanner, and the images obtained were analyzed using GCOS software. The experiment was performed in 5 replicates for wide type mice at 8 weeks of age and transgenic mice at 8, 12, 16, and 20 weeks of age. Data normalization was calculated with the robust multichip average (RMA) algorithm (*33*) implemented in BioConductor (www.bioconductor.org). The array data have been deposited in the GEO database under the publication.

MNI compendium generation

Approximately 11,000 mouse microarrays data were collected from the following publically available databases: the NIH Gene Expression Omnibus (GEO). All the arrays were processed based on the Affymetrix 430 2.0 chip which contain over 45,000 probe sets representing over 34,000 well-substantiated mouse genes. Sequences used in the design of the array were selected from GenBank, dbEST, and RefSeq. Eleven pairs of oligonucleotide probes measure the

level of transcription of each sequence represented on the GeneChip Mouse Genome 430 2.0 Array. Data normalization on the total compendium was calculated with the robust multichip average (RMA) algorithm implemented in BioConductor (www.bioconductor.org). K-mean clustering method was applied on the total mouse compendium data and approximately 3000 microarrays were selected as final training input data set for MNI.

MNI network inference algorithm

A detailed description of the MNI algorithm and underlying assumptions has been published previously (34). Here we provide a brief summary. The algorithm consists of two phases. In the first (the training phase), a model of regulatory influences in the cell is learned from an NxM data matrix, X, consisting of measurements of steady-state expression ratios of N genes in M experiments. To estimate the network model A, the MNI algorithm uses a recursive strategy. The algorithm begins by using the data matrix, X, and a naive model of the regulatory structure (i.e. no genes regulate any other) to estimate P, an NxM matrix of external influences on the genes. The estimate of P is then used, with A, to determine A by principal components regression analysis. Estimates of A and P are then used to recursively estimate one another until the estimates converge. In the second phase of the algorithm, the A matrix, representing a model of regulatory influences in the cell, is used to estimate the "key mediator genes" of the disease stage of interest. The disease stage then becomes an Nx1 vector, p, of gene-specific influences that result in the log-transformed expression ratios, x, measured for that disease state. The p vector is then calculated directly from the log-linear regulatory model as: P=Ax.

The significance of each element of the p vector is then calculated as a z-score. Genes are ranked according to the z-score of their corresponding element in the p vector and then top-ranked genes and pathways are selected as probably key mediators of the disease state.

MNI auto-exec protocol document

The MNI algorithm requires two sets of data: a compendium data set consisting of thousands of mouse genome expression profiles to construct a statistical model of gene-regulatory networks in mouse cells, and a testing data set from breast cancer transgenic mouse cancer progression model. The algorithm operates in two phases. In phase one, the algorithm applies a machine-learning approach based on multiple regression to the training data to construct a statistical model of regulatory influences of genes on one another. In phase two, the algorithm filters the test expression profiles to distinguish the mediators of the treatment response from hundreds of additional genes that also exhibit expression changes. In this step, the algorithm looks for genes with expression changes that are not explained by the regulatory model. The algorithm assumes that such 'outlier' genes are influenced externally by the cancer stage. The significance of this external effect is quantified with a *Z*-score. The high-score genes are identified as the putative MOA of the experimental condition.

A detailed description of the MNI Auto-exec procedure has been published previously (*21*). Here, we give a brief overview of the protocol to apply MNI tools on our breast cancer transgenic mouse cancer progression model.

Step 1: Collection and preparation of the training data set: A compendium data set were collected from the NIH Gene Expression Omnibus (GEO) database consisting of thousands of mouse genome expression profiles on Affymetrix MG430 Plus 2.0 Arrays to construct a statistical model of gene-regulatory networks in mouse cells.

Step 2: Preparation of the testing data set: Breast cancer transgenic mouse cancer progression model microarray profiles were processed on the Affymetrix MG430 Plus 2.0 Arrays.

Step 3: Formatting the data: Raw CEL (cell intensity) files were processed for all the mouse training and testing data sets using RMA normalization method.

Step 4: Running the MNI program with default parameters: The default parameters to run MNI on the breast cancer transgenic mouse cancer progression model are:

```
TopN = 300;                 %Top genes selecting for analysis
ITER = 5;                   %Number of iterations in the MNI core algorithm
thP = 0.25;                 %Threshold for selection of "perturbed" genes
Q =183;                     %Number of reduced dimensions.
NROUNDS = 3;                %Round of tournament
KEEPFRAC = (1/3);           %Number of genes to keep in each round of tournament
CHK_CONVERGE = 1;           %Check the convergence
```

Step 5: Prediction of key mediators by MNI method: The MNI output the top 300 key mediators genes for each test data set of the breast cancer transgenic mouse cancer progression model, and ranking them based on the modified Z-score.

Timing

Depending on the number of genes on the microarray, the number of experiments in the training data set and the computer hardware, it typically takes from 2 to 4 h to finish each run of the MNI program on the mouse genome training and testing data set. Longer run times are possible.

Software available

The MNI algorithm source code is available to academic users for free download (http://dibernardo.tigem.it/softwares/mode-of-action-by-network-inference-mni). The software is available as both an executable and a Matlab script. To run the MNI source code, the Matlab (www.mathworks.com/products/matlab/) interactive environment is required.

MNI reproducibility through K-fold cross-validation

The K-fold cross-validation is a good way to overcome the possibility of multiple local minima and to assess the reproducibility of a model. To improve the specificity of the MNI algorithm target prediction, we adopted a 3-fold cross-validation "tournament" approach. For a given testing expression profile obtained for a specific perturbation, we applied the MNI algorithm in five successive iterations. In the first iteration, we ranked all genes measured in the testing expression profile. We then selected one-third of the genes ranked highest (which is approximately 7300 genes out of approximately 22,000 genes in the mouse training data set). Once again, we reapplied the MNI algorithm to the remaining genes and selected the one-third

highest-ranked genes. We then applied the MNI algorithm three more times in this manner to obtain a final ranking.

The advantage of this approach is that the dimensional reduction via singular value decomposition (SVD) in the MNI algorithm preserves more of the differences between genes as the number of genes processed becomes closer to the number of experiments. Thus, with each successive application of the MNI algorithm, differences between similarly regulated genes are more clearly identified and the specificity of the target prediction is improved. We also added a convergence check that stops the iterations if the SVD error increases compared to the previous iteration. After each iteration, the error should decrease as the number of genes processed becomes closer to the number of experiments. If this does not happen, it means that the MNI algorithm did not choose the proper genes, therefore the recursion is stopped and the final ranking is computed on the last "good" iteration. To overcome the potential of improperly eliminating true target genes in early iterations of the MNI algorithm (when the specificity is still low), we also implemented the modified z-score approach to rank genes prior to the selection of the highest-ranked one-third of the genes. The modified z-score was designed to boost the likelihood of including genes with significant changes in the testing expression profile. From our simulations, we observed that using the modified z-score to rank the genes improves the prediction of target genes compared with using the standard z-score.

MAPK signaling pathway analysis

The core components of the mouse MAKP signaling pathway were extracted from the KEGG Database (www.genome.jp/kegg/kegg2.html). Transcriptional gene expression data for these MAKP signaling pathway component was extracted from our tumor gene expression profiling data set from the transgenic mice at 8, 12, 16, and 20 weeks of age. Unsupervised clustering and heatmap of these MAKP signaling pathway was performed and generated by using R Bioconductor Pachage (www.bioconductor.org). The values in the heatmap represent the log ratio intensity fold change of the transgenic mice at different age versus the wild-type mice at 8 weeks of age. The yellow color represents upregulation, the blue color represents down-regulation.

The Oncomine database (www.oncomine.org/resource/login.html) was utilized to probe the MAPK signaling pathway components across large number of human breast cancer dataset.

Supplementary Figures

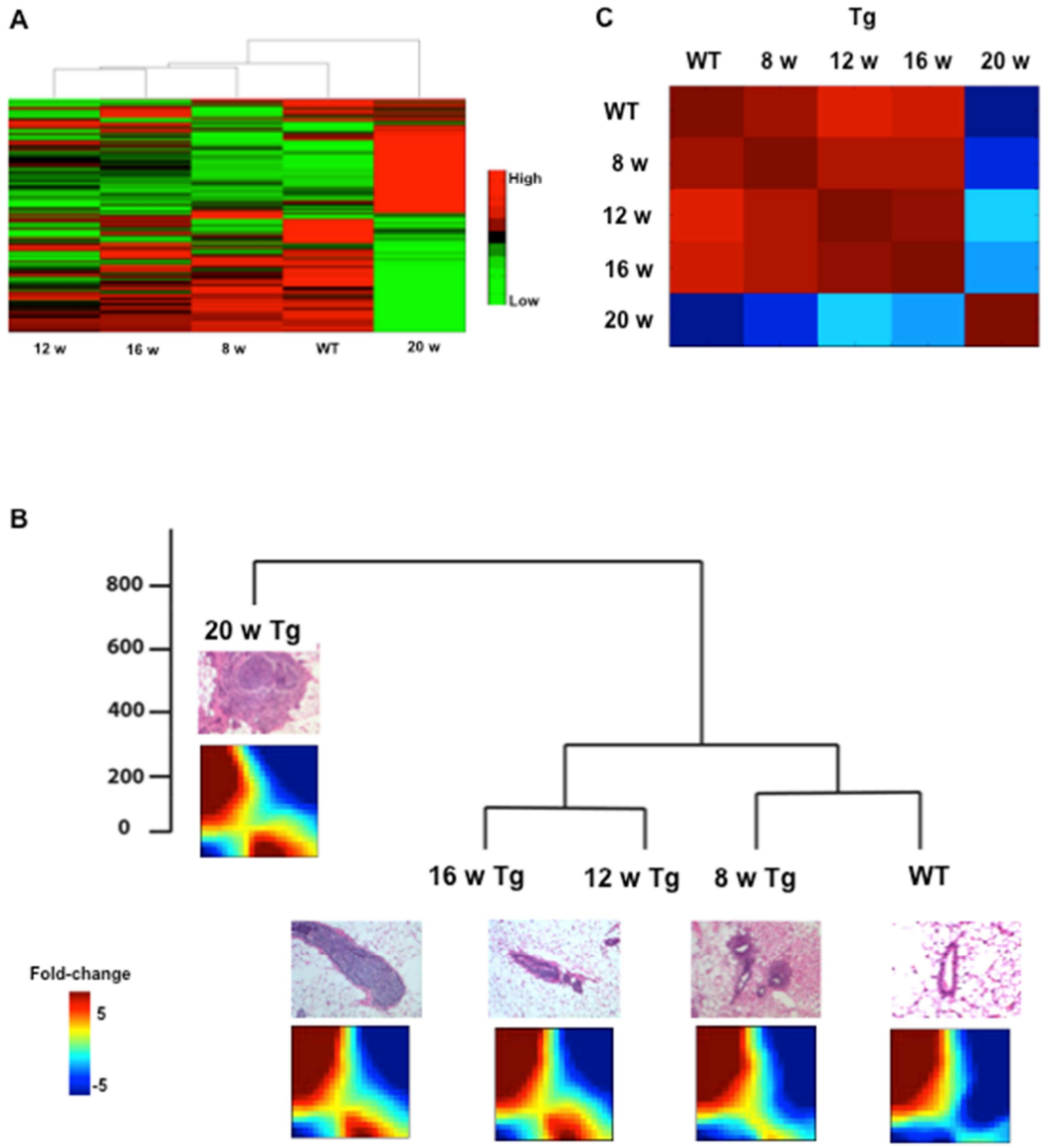

Figure S1. Conventional bioinformatics analysis of mammary gland gene expression changes in C3(1)-SV40TAg–driven tumorigenesis. Differential gene expression analysis of tumorigenesis in C3(1)-SV40TAg mammary glands. (**A** and **B**) Dendrogram visualization (A) and self-organized maps (B) illustrate clustering by Euclidean distance between gene expression profiles by disease stage. The expression of each gene was normalized to a mean expression of zero across all the samples. Tg, transgenic; w, weeks; WT, wild-type. (**C**) Heat map of *n* = 5 tissue samples at each time point from C3(1)-SV40TAg glands and WT FVB control glands. Pearson's linear correlation coefficient was calculated between each pair of averaged tumor sample stages, and the heat map was constructed based on the correlation coefficient value accordingly.

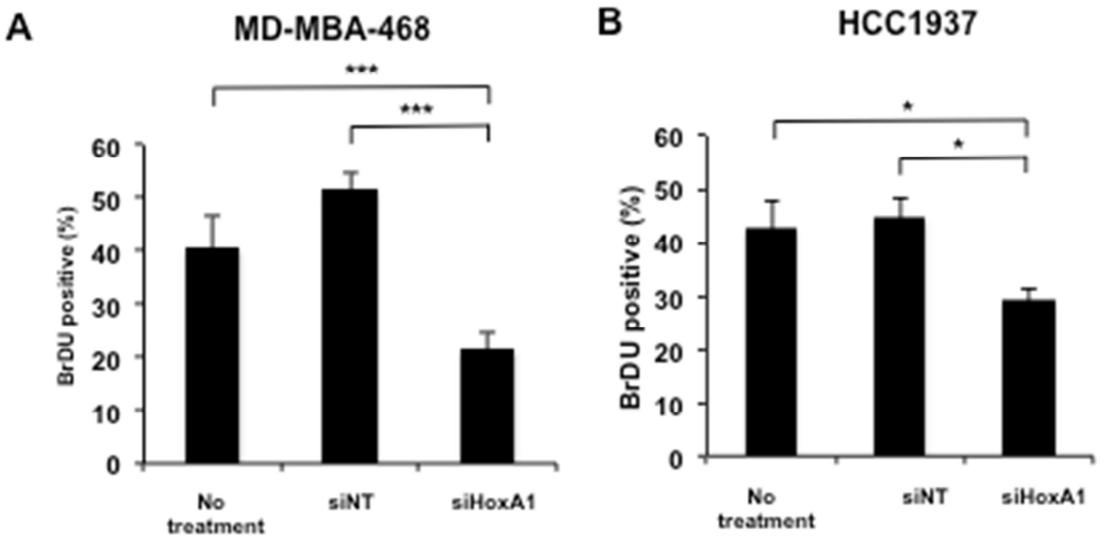

Figure S2. Silencing of *HoxA1* in human cell lines in vitro. (**A** and **B**) Proliferation of human breast cancer cells MDA-MB-468 (A) and HCC1937 (B) cultured on tissue culture plates after treatment with siHoxA1 or nontargeted, scrambled control (siNT). Data are means ± SD (*n* = 3; 500-750 cells scored per condition per cell type). **P* = 0.2, ****P*= 0.001, Fisher's exact t-test.

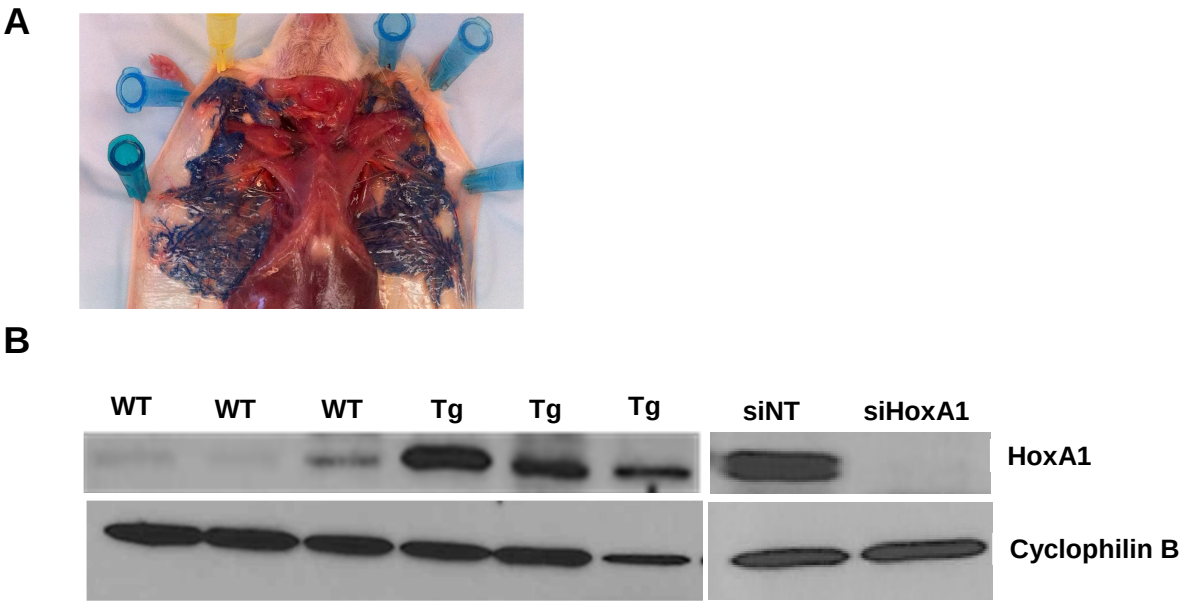

Figure S3. In vivo delivery of siRNA to the mammary ductal tree. (**A**) Evans blue dye was used to visualize the ducts and determine appropriate intraductal injection volumes. Following injection into the intact nipple of a euthanized mouse, a midline incision was performed to expose the mammary glands and the staining pattern. Pilot studies in *n* = 6 animals (6 glands per animal) determined that 10-20 µl volume was sufficient to fill the ductal tree of each mouse thoracic gland (shown here is 20 µl). (**B**) In vivo knockdown of *HoxA1* in the mammary gland of transgenic mice following intraductal injection of siHoxA1 formulated with lipidoid nanoparticles. siRNA was delivered biweekly from 12 weeks of age through 20 weeks of age. At 21 weeks of age, protein lysates were prepared from siRNA-treated glands and untreated wild type (WT) (*n* = 3 animals per condition) for analysis by Western blot. Cyclophilin B served as loading control.

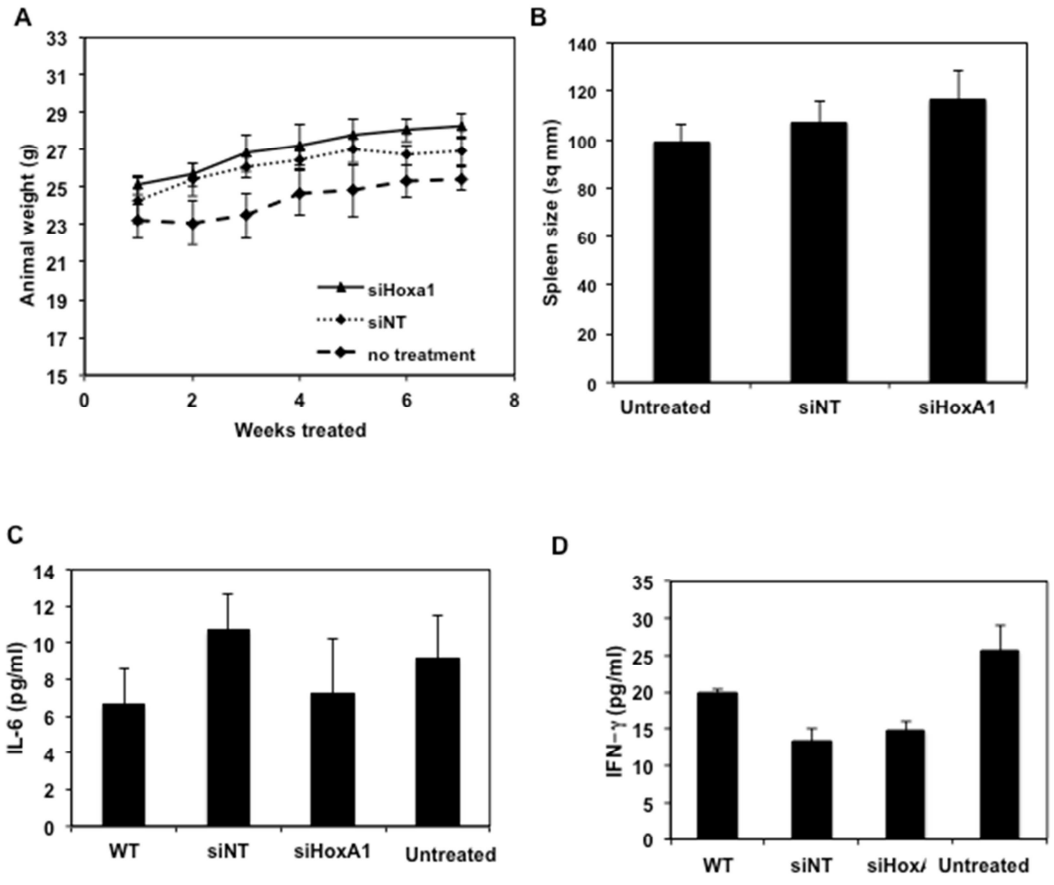

Figure S4. Monitoring systemic response to in vivo RNAi. FVB C3(1)-SV40TAg mice receiving intraductal injections of siRNA lipid nanoparticles were monitored closely for signs of poor health or inflammatory response. (**A**) Body weight over the course of the 9-week study. (**B**) Upon sacrifice at 21 weeks of age, the spleen was removed and measured. (**C** and **D**) Systemic immune activation was assessed by measuring cytokines IL-6 and IFN-γ in serum from FVB C3(1)-SV40TAg transgenic and FVB wild type (WT) animals 24 h post-injection. Data are means ± SD (*n* = 3 animals, cytokine assays were performed in duplicate).