AD________________

Award Number:
W81XWH-08-1-0572

TITLE:
RNAi as a Routine Route Toward Breast Cancer Therapy

PRINCIPAL INVESTIGATOR:
Gregory J. Hannon, Ph.D.

CONTRACTING ORGANIZATION: Cold Spring Harbor Laboratory
Cold Spring Harbor, NY 11724

REPORT DATE:
May 2014

TYPE OF REPORT:
Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT:

X Approved for public release; distribution unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

# REPORT DOCUMENTATION PAGE	*Form Approved* *OMB No. 0704-0188*

1. REPORT DATE *(DD-MM-YYYY)* May 2014	**2. REPORT TYPE** Final	**3. DATES COVERED** *(From - To)* 1Sep2008- 28Feb2014

4. TITLE AND SUBTITLE RNAi as a Routine Route Toward Breast Cancer Therapy	**5a. CONTRACT NUMBER**
	5b. GRANT NUMBER W81XWH-08-1-0572
	5c. PROGRAM ELEMENT NUMBER
6. AUTHOR(S) Gregory J. Hannon, PhD.	**5d. PROJECT NUMBER**
	5e. TASK NUMBER
email: hannon@cshl.edu	**5f. WORK UNIT NUMBER**
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Cold Spring Harbor Laboratory One Bungtown Road Cold Spring Harbor, NY 11724	**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) US Army Medical Research and Materiel Command Fort Detrick, MA 21702	**10. SPONSOR/MONITOR'S ACRONYM(S)**
	11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

During the first year of this innovator award, we made significant progress toward two of our aims. We constructed a third generation RNAi library and made that available to the breast cancer community. This resource has nearly 75,000 independent, sequence verified clones targeting ~18,000 human genes. A similar library for the mouse genome is nearing completion. We also scaled up our shRNA screening platform in preparation for lethality surveys of all suitable and available BC cell lines, including matched pairs of lines that have acquired resistance to herceptin in vitro. Relevant to our second aim, we have profiled microRNA from each of the identifiable epithelial cell types in the mouse mammary gland and are undertaking similar efforts in human. The goal is to develop microRNA sensor strategies that will permit visualization of each cell type in vivo and enable their isolation and manipulation in vitro. Finally, we showed that two microRNAs, let-7 and miR-93, can deplete tumor initiating cells from a number of basal breast cancer cell lines.

15. SUBJECT TERMS

RNAi, sequencing

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** USAMRMC
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	UU	112	**19b. TELEPHONE NUMBER** *(include area code)*

Table of Contents

Page

Cover……………………………………………………………………………….1

SF298……………………………………………………………………………….2

Table of Contents…………………………………………………………………3

Introduction…………………………………………………………….………..……..4

Body……………………………………………………………………………..4-17

Key Research Accomplishments………………………………………….18

Reportable Outcomes…………………………………………………………18-19

Conclusion…………………………………………………………………………19-20

References…………………………………………………………………………21

Appendix …………………………………………………………………….…..22-106

INTRODUCTION

The goal of this innovator award is to continue to develop and apply RNAi-based screening methods to discover new routes toward breast cancer therapy. The project has three sets of goals. First is to integrate genomic and genetic information on available breast cancer cell lines to identify tumor-specific vulnerabilities and to understand genetic determinants of therapy resistance. Second is to probe the roles of breast cancer stem cells, with a particular emphasis on microRNAs. The third is to examine regions that determine familial susceptibility to breast cancer by applying novel, focal re-sequencing methods developed in the laboratory.

BODY

Third (V3)/Fourth (V4)/Fifth (V5) generation RNAi libraries

Although our second-generation shRNA library was a major improvement over our first version (1), we knew that even given improved designs, only a fraction of constructs showed highly penetrant knockdown. Based on our previous knowledge from shRNA screening, we also knew that genuine shRNA hits tend to be hairpins of high efficacy. We therefore needed to develop a clever, large-scale, high throughput, shRNA validation approach toward measuring hairpin efficacy. This led to the development of our 'sensor assay' for shRNA potency, a project we undertook in collaboration with Scott Lowe's lab (CSHL & MSKCC) (2). In essence we produced shRNA sensor libraries by using oligonucleotides (185nt), which carry both the shRNA sequence and its corresponding target site, synthesized on DNA microarrays (Agilent Technologies Inc.). After converting the single-stranded oligonucleotides into double-stranded DNAs, the PCR amplified constructs are inserted into a vector and repaired with a second insert to form two independent transcriptional units. One of which expresses the shRNA from an inducible promoter and one of which expresses the target (composed of a Venus fluorescent marker with the target site inserted into the 3'UTR of the reporter) constitutively. In this way, each individual cell reports the activity of a particular shRNA, with the readout of efficiency being a change in expression of the fluorescent reporter. Using this method, we were able to measure the activity of more 22,000 different shRNAs in a single pool. We applied this approach to measure the efficacy of the top 12 designs of shRNAs for every gene in the human genome based on the DSIR siRNA prediction algorithm (3). This generated approximately 250,000 measurements of shRNA efficacy, the largest such dataset ever generated, and provided the training data to devise a predictive algorithm (we termed shERWOOD) that can essentially predict the results of functional, sensor testing of shRNAs *in silico*. The top 6 DSIR designs were used to produce our third-generation shRNA libraries (for human and mouse genomes). This shRNA library was used in all of our genomewide shRNA screens of breast cancer cell lines covering all three disease treatment subtypes. This huge undertaking was done in collaboration with Stephen Elledge (Harvard Medical School).

All algorithms that predict effective RNAi tools tend to choose sequences that being with a U. This is thought to have a structural basis in the interaction between the RNA and Argonaute, the key core of the RNAi effector complex. That 5' residue has been shown to reside in a binding pocket, which favors interaction with U. Therefore, the sequence space available for effective RNAi tools is really restricted to only ¼ of the transcriptome. When the small RNA interacts with Argonaute, its 5' end is not available for pairing to the target RNA. Therefore, even though the 5' U contributes to RISC binding, it is irrelevant to target recognition. We therefore tested the idea that we could expand available sequence space by simply releasing the aforementioned constraint, in essence predicting on every positing in the transcriptome and changing the small RNA guide that would pair to that site so that it contained a 5' U. This produced even higher scores in the algorithm and was especially important for small genes with limited numbers of potential target sequences.

To build our version 4 (V4) libraries, we generated a collection of over 1.3 million oligonucleotides corresponding to shERWOOD shRNA predictions encompassing the human, mouse, rat, and fly genomes. These used both the conventional genomes and the 1U strategy. We confined our fourth generation libraries to REFSEQ genes and employed a series of heuristics to maximize the likelihood that our target sites would fall within constitutive exons. We cloned these into our basic shRNA expression vector and produced arrayed collections of sequence-verified clones targeting the human, mouse, and fly genomes. The completed V4 human shRNA library is composed of 76,861 shRNAs targeting 18,651 genes. The mouse V4 library currently has 58,113 shRNAs targeting 18,769 genes with 11,936 genes represented by three or more shRNAs.

During the first quarter of 2014, we began construction of our fifth-generation shRNA libraries. This V5 design, termed ‘ultramir’ shRNAs, consists of shERWOOD predicted hairpins inserted into an improved microRNA shell, resulting in more than 2-fold increase in mature small RNA production compared to our V4 design. This increase translates to a significant improvement in efficacy. We have recently submitted a manuscript describing the shERWOOD algorithm and the ultramir shRNA design to *Molecular Cell* and it is currently under review. See attached manuscript of the title “ A computational algorithm to predict shRNA potency “.

Progress of Genomewide shRNA screens of breast tumor-derived cell lines

Breast Cancer Cell Lines	Screening Conditions	Status
Her2+ treatment category		
JIMT1	No drug (straight-lethal)	Screen completed & sequenced
JIMT1	Lapatinib IC20	Screen completed & sequenced
MDA-MB-453	No drug (straight-lethal)	Screen completed & sequenced
MDA-MB-453	Lapatinib IC20	Screen completed & sequenced
MDA-MB-361	No drug (straight-lethal)	Screen completed
MDA-MB-361	Lapatinib IC20	Screen completed
EFM-TR	No drug (straight-lethal)	Screen completed
EFM-TR	Trastuzumab (15ug/ml)	Screen completed
EFM192A	No drug (straight-lethal)	Screen completed & sequenced
EFM-192A	Trastuzumab (15ug/ml)	Screen completed & sequenced
SkBr3	No drug (straight-lethal)	Screen completed & sequenced
SkBr3	Trastuzumab (15ug/ml)	Screen completed & sequenced
Sk-TR	No drug (straight-lethal)	Screen completed & sequenced
Sk-TR	Trastuzumab (15ug/ml)	Screen completed & sequenced
HCC1954	No drug (straight-lethal)	Screen completed & sequenced
ER+ treatment category		

ZR75-1 Parental	+ E2	Screen completed & sequenced
ZR75-1 Parental	- E2	Screen completed & sequenced
ZR75-1Parental	- E2 / + Tamoxifen	Screen completed & sequenced
ZR75-1-EDR	+ E2	Screen completed & sequenced
ZR75-1-EDR	- E2	Screen completed & sequenced
ZR75-1-TAMR	+ E2	Screen completed & sequenced
ZR75-1-TAMR	- E2 / + Tamoxifen	Screen completed & sequenced
MCF7 Parental	+ E2	Screen completed
MCF7 Parental	- E2	Screen completed
MCF7 Parental	- E2 / + Tamoxifen	Screen completed
MCF7 -EDR	+ E2	Screen completed
MCF7 –EDR	- E2	Screen completed
MCF7-TAMR	+ E2	Screen completed
MCF7-TAMR	- E2 / + Tamoxifen	Screen completed
T47D	No drug (straight-lethal)	Screen completed/microarray analysis completed

TN/Basal treatment category

Hs578T	No drug (straight-lethal)	Screen completed
MDAMB231	No drug (straight-lethal)	Screen completed
MDAMB468	No drug (straight-lethal)	Screen completed
MDAMB436	No drug (straight-lethal)	Screen completed / microarray analysis completed
HCC1143	No drug (straight-lethal)	Screen completed / microarray analysis completed
HCC1937	No drug (straight-lethal)	Screen completed / microarray analysis completed
SUM149	No drug (straight-lethal)	Screen completed / microarray analysis completed
SUM1315	No drug (straight-lethal)	Screen completed / microarray analysis completed

We have set out to perform 24 genome-wide RNAi screens on breast cancer cell line models of all three treatment subgroups and we have exceeded our goal. The Hannon/Elledge groups has completed 38 genome-wide RNAi screens (in triplicate) using our second-generation (75,905 shRNAs targeting 19,011 genes) and third-generation (74,304 shRNAs and targeting over 19,000 genes) shRNA libraries.

Genome-wide RNAi screens of Her2-positive models

Approximately 20% to 25% of invasive breast cancers exhibit overexpression of the human epidermal growth factor receptor (HER) 2 tyrosine kinase receptor. As elevated HER2 levels are associated with reduced disease-free and overall survival in metastatic breast cancer, therapeutic strategies have been developed to target this oncoprotein. Trastuzumab, a recombinant humanized monoclonal antibody directed against an extracellular region of HER2, was the first HER2-targeted therapy approved for treatment of HER2-overexpressing metastatic breast cancer. This drug is active as a single agent and in combination with adjuvant chemotherapy (either in sequence or in combination) in HER2-positive breast cancers. However, the objective response rates to trastuzumab mono-therapy were low (12% to 35%), and for a median duration of nine months, suggesting a majority of HER2-overexpressing tumors demonstrated *de novo* resistance. Phase III trials revealed that the combination of trastuzumab and paclitaxel or docetaxel could increase response rates, time to disease progression, and overall survival compared to trastuzumab mono-therapy. For HER2-positive patients who had not received prior chemotherapy, the median time to progression in response to trastuzumab as single-agent was less than five months. In patients who received trastuzumab and chemotherapy, the median time to progression was 7.5 months. Thus, the majority of patients who achieve an initial response to trastuzumab-based regimens develop resistance within one year. Elucidating the molecular mechanisms underlying acquired resistance to trastuzumab is essential for improving the survival of HER2-positive, metastatic breast cancer patients. The second-generation drug, lapatinib, targeting both HER1 and HER2 receptors, has demonstrated efficacy in killing trastuzumab-resistant human breast cancer cells. Unfortunately, intrinsic resistance to lapatinib has been observed in a number of metastatic, HER2-positive tumor derived cell lines.

To identify genes conferring secondary (acquired) resistance to trastuzumab from the datasets, we set a cutoff of FDR<0.25 and filtered for genes that depleted only upon trastuzumab treatment in the drug resistant line, SKTR, but not in either of the two drug sensitive lines, SKBR3 and EFM192A. This produced a list of 25 genes (Figure 1: Heatmap of the 25 genes), which included those from PI3K-mTOR signaling (PI4K2A, Raptor, Insulin receptor, and EIF4A), RNA processing (PRPF8, U2AF1, and LSM6), mitotic checkpoint (BUB1B), and genes of relatively less well-known function. Identification of the insulin receptor and members of the PI3K-mTOR signaling pathway fulfills our expectation of finding these genes in this screen since they are known to be functionally associated with trastuzumab resistance. However, TNFSF11/RANKL (ligand of the receptor activator of nuclear factor kappa B), a gene of significant relevance to breast cancer, was also found in the screen as one of two most highly depleted hits.

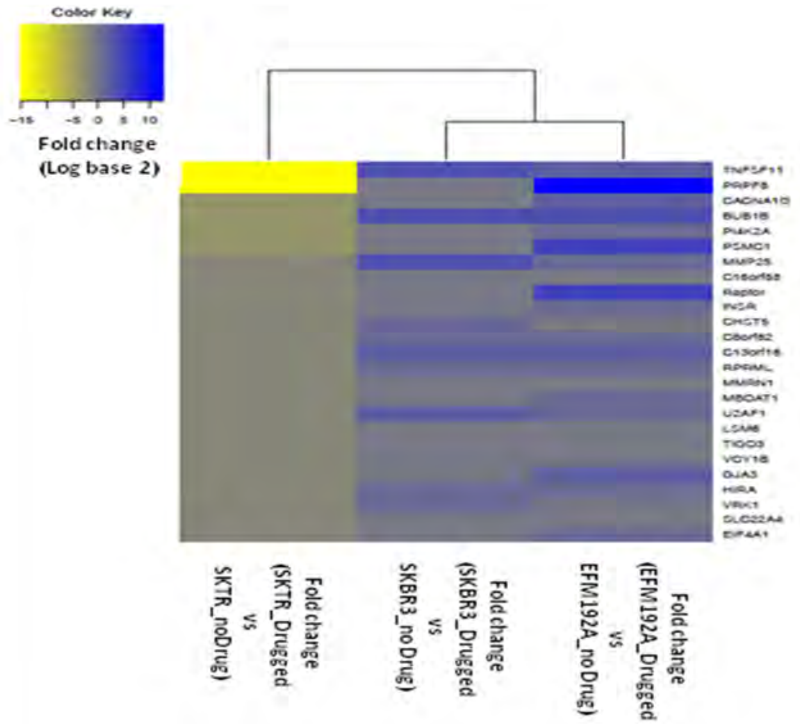


Figure 1. Heatmap of shRNAs (FDR<0.25) for gene targets that sensitize drug resistant (SKTR) and not drug sensitive cell line models (SKBR3 and EFM192A). These are potentially novel candidates for trastuzumab combination therapy.

Figure 2. Competition assays to validate selected hits. Cell line: SKTR (Drug resistant), Drug = Trastuzumab.

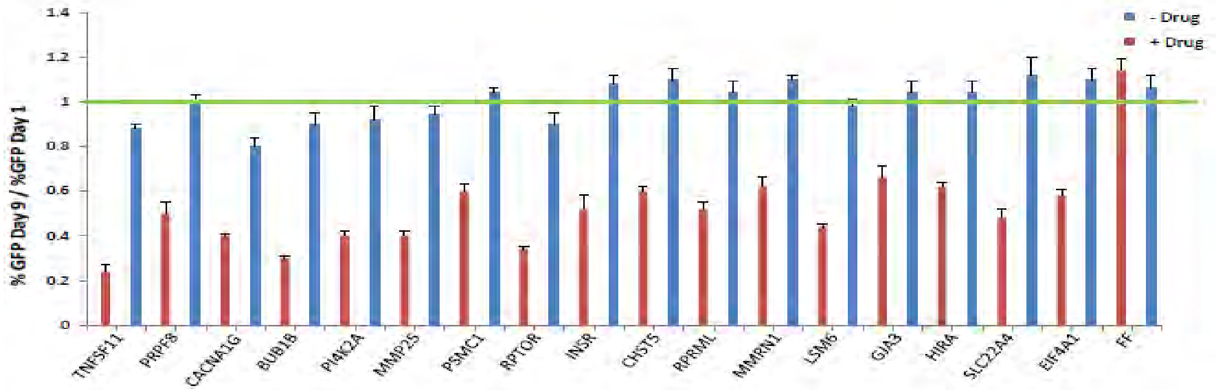
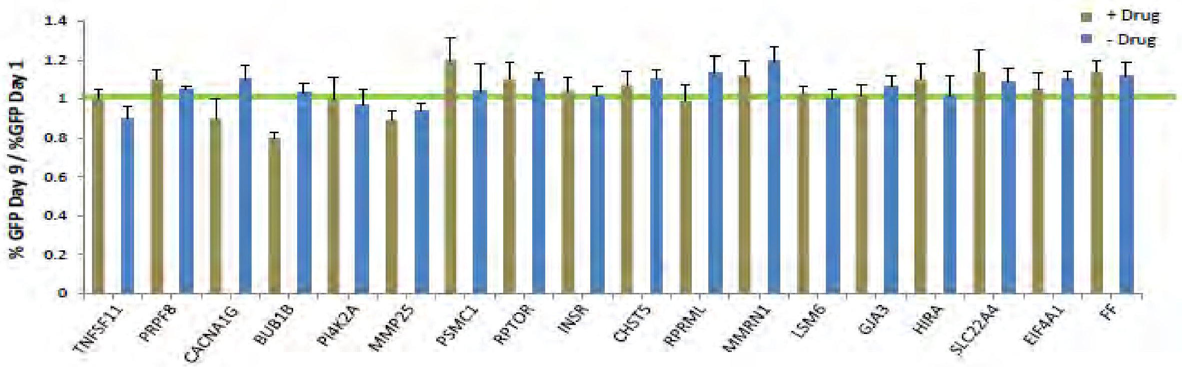


Figure 3. Competition assays to validate selected hits. Cell line: SKBR3 (Drug Sensitive), Drug = Trastuzumab.



We validated 17 of the 25 targets (Figure 1) by using competition assays and the results demonstrated that these target genes specifically sensitize the trastuzumab resistant cells (SKTR) when silenced by RNAi (Figure 2 and 3). In the competition assay, shRNA-expressing cells (GFP+) are mixed with an equal proportion of parental cells (no shRNA). Each cell line mixture is plated in triplicate and either untreated or treated with drug. The percentage of GFP+ cells remaining is then tracked over time. We prioritized our effort by selecting RANKL/TNFSF11 as the first target for further validation due to the existence of a clinically approved inhibitor called Denosumab. Xenografts of these cell lines are currently being tested to validate whether RANKL is a target for transtuzumab sensitization *in vivo*. Ultimately, our goal is to test whether this combination approach will have an impact on reducing tumorigenicity in primary human breast cancer cells from patients that are resistant to trastuzumab therapy. In addition, we will continue to pursue the remaining targets in further validation studies.

Denosumab is a humanized monoclonal antibody designed to inhibit RANKL for treating various bone related conditions. This drug was approved to treat giant cell tumor of the bone, breast cancer patients on adjuvant aromatase inhibitor therapy to increase bone mass, postmenopausal women with risk of osteoporosis, and for the prevention of skeletal-related events in patients with bone metastases from solid tumors. Denosumab is highly specific as it binds human RANKL, but not murine RANKL, human TRAIL, or other human TNF family members. RANKL and its receptor RANK are best known for their essential function in bone remodeling and bone-related pathologies including osteoporosis and arthritis. The dysregulation of the RANKL-RANK system is the major cause of osteoporosis in post-menopausal women. Appropriate RANKL signaling is also required for the formation of a lactating mammary gland, and both RANKL and RANK are expressed under the control of progesterone, prolactin, and the parathyroid hormone protein-related peptide (PTHrP). Recent data also implicate RANKL and RANK in the control of metastasis of breast cancer cells to the bone and sex hormone-driven primary mammary cancer. Unfortunately, synthetic progesterone derivatives (progestins), such as medroxyprogesterone acetate (MPA), used in hormone replacement therapy and contraceptives have been demonstrated to induce the RANKL-RANK system, providing growth and survival advantage to damaged mammary epithelium, a requisite for tumor initiation. In addition, recent evidence links Her2 expression to RANKL-RANK signaling. Her2 expression is increased in luminal tumor cells grown in mouse bone xenografts, as well as in bone metastases from patients with breast cancer as compared to matched primary tumors. The increase in Her2 protein levels was not due to gene amplification, but rather was mediated by RANKL in the bone environment.

We have analyzed the genome-wide RNAi screen data of JIMT1 (no drug) and MDA-MB-453 (no drug) for common genes that are essential for *de novo* lapatinib resistance. This common gene list was filtered against essential genes for the ER-positive cell line ZR75-1 to remove those genes that might also be essential for ER-positive breast cancer cells. This analysis produced a list of candidate genes that is specific for Her2 driven cancer cells. Molecular pathways that are enriched for this set of genes include the cell cycle, protein ubuiquitination, proteasome, organelle biogenesis and organization, and others. Among the candidate genes is LGR5/GPR49, a cell surface marker involved in self-renewal in normal and cancer cells (e.g. colon cancer). Also of note is TOP1 (topoisomerase I), one of the genes that is predicted to be essential for JIMT1 and MDA-MB-453 cells to survive. We will validate a selected list of targets including TOP1 (using both RNAi and small molecule inhibition with irinotecan) and LGR5 for the survival of *de novo* lapatinib resistant cells.

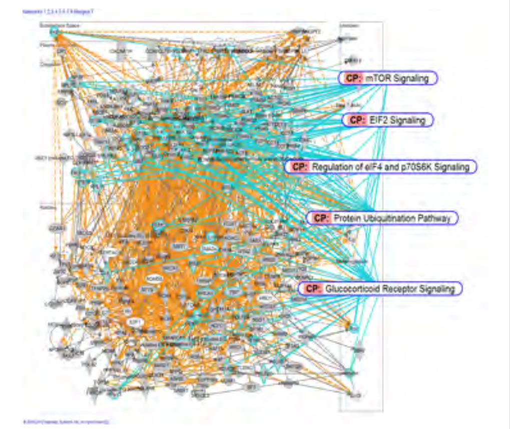

CP: mTOR Signaling
CP: EIF2 Signaling
CP: Regulation of eIF4 and p70S6K Signaling
CP: Protein Ubiquitination Pathway
CP: Glucocorticoid Receptor Signaling

Figure 4. Interaction network of lapatinib sensitizing genes for MDA-MB-453 cell line

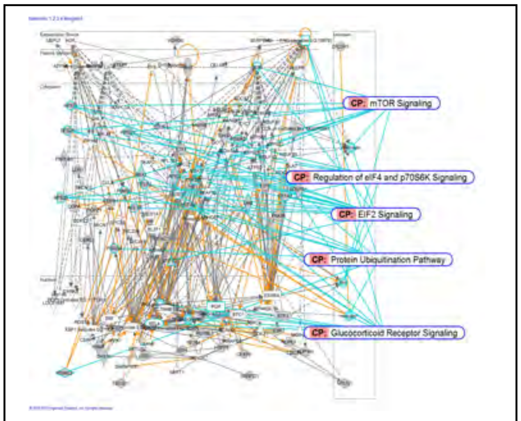

CP: mTOR Signaling
CP: Regulation of eIF4 and p70S6K Signaling
CP: EIF2 Signaling
CP: Protein Ubiquitination Pathway
CP: Glucocorticoid Receptor Signaling

Figure 5. Interaction network of lapatinib sensitizing genes for JIMT1 cell line (Ingenuity).

We have also analyzed the same data in a manner to inform us of potential modifiers of lapatinib resistance, particularly genes that could be targeted to sensitize *de novo* lapatinib resistant cells to the drug. Molecular pathway enrichment analysis of genes common to both JIMT1 and MDA-MB-453 suggests that several molecular complexes could be targeted to sensitize lapatinib resistant cells to the drug, including the APC/C (anaphase promoting complex/cyclosome), proteasome, and coatamer complexes. Other highly enriched cellular pathways include mTOR, EIF2, EIF4/p70S6 kinase, glucocorticoid receptor, and protein ubiquitination (Figures 1 and 2).

Validation will be carried out on a panel of *de novo* lapaitinib resistant cell lines (including JIMT1 and MDA-MB-453), lapatinib sensitive lines, and normal (immortalized) human epithelial cells (HMEC) *in vitro*. Promising candidates will be further tested for their ability to sensitize lapatinib resistance *in vivo*.

Genome-wide RNAi screens of Triple-negative models

Triple-negative (TN) breast cancer accounts for 15-20% of all breast cancer cases, and is associated with poor prognosis and unfavorable clinical outcome. Our goal is to uncover genetic dependencies of survival to define heterogeneities between different TN tumor-derived cell lines and on the identification of pathways/genes that can be candidates for targeted therapy. Cells defective in DNA repair (BRCA1-negative) have been demonstrated to be sensitive to strategies involving PARP inhibition. We have focused our effort to find new targets for BRCA1 mutant, TN, cancer cells. Candidate genes essential for proliferation or survival for BRCA1-negative cell lines were identified by comparing lethal signatures of BRCA1 -/- (HCC1937, SUM149, SUM1315, and MDA-MB-436) and BRCA1 +/+ cells (HCC1143, HCC1954, and T47D). Initial validation of candidates uncovered four potential BRCA1 synthetic lethal genes. This study will be continued beyond the funding period of this grant.

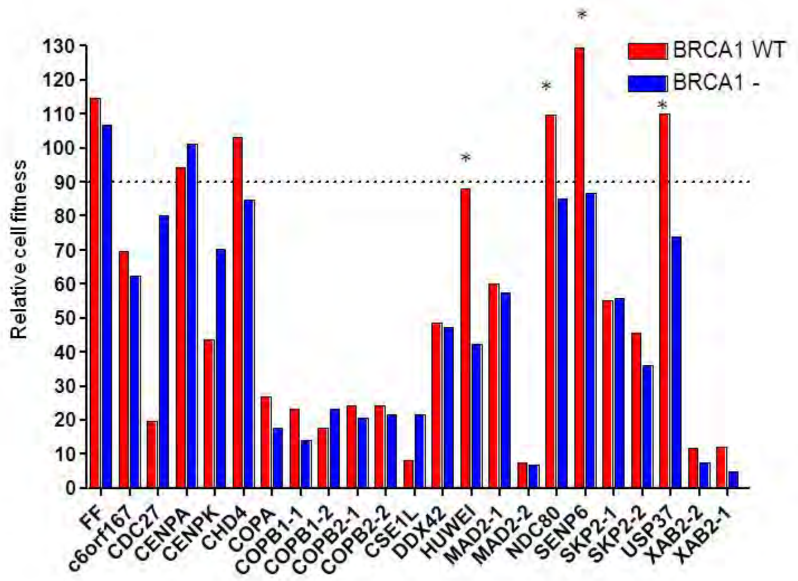


Figure 6. Validation of hits from BRCA1 synthetic lethal screens.

Focused RNAi Screens

It has become increasingly clear that tumor cells, like normal cells, are driven by self-renewing compartments known as tumor-initiating or cancer stem cell populations. These cells exhibit higher resistance to targeted therapy than the rest of the tumor cell population. Thus it is important to understand the essential pathways that drive tumor-initiating cells and how these signatures differ from those that enable normal breast progenitor cells to survive. To mark progenitor cells in normal mouse mammary epithelial cells, we have applied an approach developed by E. Fuchs laboratory (Rockefeller University) to identify/purify label-retaining cells (LRCs) that mark the skin stem cell niche. The system is built upon the premise that stem cells are slow cycling and active for the keratin5 promoter. It utilizes a tetracycline-responsive promoter driving histone 2B-GFP (H2B-GFP) in a transgenic mouse model expressing the tet repressor-VP16 (tTA) transgene from the K5 promoter. In the absence of doxycyline, the expression of GFP is high in epithelial cells. After feeding the animal doxycycline for a period of several weeks only very small populations of epithelial cells retain GFP fluorescence.

Using this system we have isolated LRCs from mammary epithelial cells and have demonstrated their self-renewal potential both *in vitro* (mammosphere formation assay) and *in vivo* (reconstitution of mammary gland). Furthermore, we have profiled the H2B-GFP+ cells for transcriptome, methylome, and miRNA expression analysis. Comparison of the transcriptomes of LRCs and other subpopulations of the mammary gland (luminal ductal cells, luminal alveolar cells, luminal progenitors, myoepithelial progenitors, and myoepithelial cells) has allowed us to identify new a cell surface-specific marker for the H2B-GFP+ cells, CD1. CD1-specific cell populations were also found to be present in human basal breast cancer cell lines. We are currently testing if these human CD1-specific tumor cell populations have self-renewing capacity.

To find essential genes/survival pathways for H2B-GFP+ breast cancer progenitor cells, we have made a focused shRNA library targeting the MaSC (mammary gland stem cell) genes, which are highly expressed in the H2B-GFP+ compartment but not in other normal mammary cell types. This MaSC shRNA library was then used to perform a well-by-well RNAi screen to test the effect of each individual shRNA on the survival of COMMA-D cells (normal mouse mammary epithelial cell line) and 4T-1 cells (mouse mammary basal-like, metastatic cell line). Several candidate genes, including CD1, from these screens are being tested to determine whether they are essential for survival in human triple-negative breast cancer cells.

This study was published in the journal *Proceedings of the National Academy of Sciences* and is titled " Molecular hierarchy of mammary differentiation yields refined markers of mammary stem cells " (see attachment).

To explore the role of epigenetics in cancer cell survival, we have completed another well-by-well RNAi screen in the mouse metastatic, basal-like, 4T1 cells using a collection of 1,100 shRNAs targeting 243 genes involved in chromatin regulation. BRD4, a gene that was recently identified as a therapeutic target in acute myeloid leukemia (C. Vakoc, CSHL), was one of the top hits (three independent shRNAs were identified). After completion of initial candidate validations, we chose to focus on the gene BPTF (involved in chromatin remodeling), which so far has demonstrated its requirement for the survival of human breast cancer cell lines representing all three treatment subgroups. Additional preclinical studies are being carried out to determine whether BPBF could be a new target for clinical testing.

Regulation of breast tumor-initiating cells by miRNAs

Accumulating evidence suggests that cancer stem cells are key to tumor formation, progression, and metastasis. This subset of tumor cells may resist conventional therapies providing a potential reservoir for relapse. We have previously discovered that the stem/progenitor compartment of comma-1D cells fail to express the let-7 and miR-93 miRNAs and their enforced expression can deplete this population of stem cells (4). Further studies of miR-93 have uncovered that it can modulate the fate of breast cancer stem cells by regulating their proliferation and differentiation states. In claudin-low SUM159 cells, expression of miR-93 induced MET (Mesenchymal-Epithelial Transition) and is associated with the downregulation of multiple stem cell regulatory genes, including JAK1, AKT3, SOX4, EZH1, and HMGA2, resulting in cancer stem cell depletion. Enforced expression of miR-93 completely blocked tumor development in mammary fat pads and development of metastases following intracardiac injection in xenograft mouse models. This work was done in collaboration with Max Wicha (University of Michigan) and was published in *PLOS Genetics* (see attachment).

Epigenetic characterization of the mammary epithelial lineage

Using whole-genome shotgun bisulfate sequencing to generate single nucleotide-resolution methylation profiles, a method we developed in 2011(5) (see attachment), we have fully characterized the DNA methylation status and gene expression patterns of mammary gland cells of nulliparous (virgin) and parous (two to three sets of pregnancies) mice.

Having established that H2b-GFP^h (GFP^h=GFP+) MaSCs have mammary gland reconstitution properties, we endeavored to characterize their global mammary gland DNA methylation patterns. Using a combination of cell surface markers, six distinct cell types were isolated via FACS to a purity of >90%: H2b-GFP MaSCs (Lin$^-$CD24$^+$CD29^{h}H2b-GFP^{h}CD61$^-$), myoepithelial progenitor cells (Lin-CD24$^+$CD29^{h}H2b-GFP$^{-/l}$CD61$^+$), myoepithelial differentiated cells (Lin$^-$CD24$^+$CD29^{h}H2b-GFP$^-$CD61$^-$), luminal progenitor cells (Lin$^-$CD24^{h}CD29$^+$CD61$^+$CD133$^-$), luminal ductal cells (Lin$^-$CD24^{h}CD29$^+$CD61$^-$CD133$^+$), and luminal alveolar cells (Lin$^-$CD24^{h}CD29$^+$CD61$^-$CD133) (Figure 7).

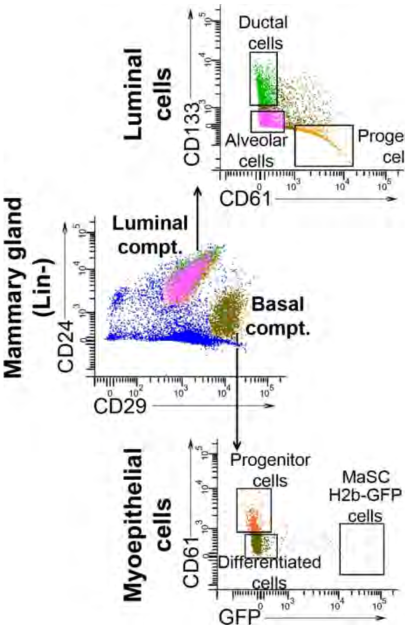


Figure 7 **Mammary gland cells sorting strategy:** We used a combination of 4 cell surface markers (ref), in addition to H2b-GFP expression, to segregate the lineage depleted mammary gland cells into 6 distinct cell types: H2b-GFP^h MaSCs (Lin$^-$CD24$^+$CD29^{h}H2b-GFP^{h}CD61$^-$), myoepithelial progenitors cells (Lin-CD24$^+$CD29^{h}H2b-GFP$^-$CD61$^+$), myoepithelial differentiated cells (Lin$^-$CD24$^+$CD29^{h}H2b-GFP$^-$CD61$^-$), luminal progenitor cells (Lin$^-$CD24^{h}CD29^{l}CD61$^+$CD133$^-$), luminal ductal cells (Lin$^-$CD24^{h}CD29^{l}CD61$^-$CD133$^+$) and luminal alveolar cells (Lin$^-$CD24^{h}CD29^{l}CD61$^-$ CD133$^-$). For each library two biological replicates were analyzed.

In all sequenced samples, we achieved an optimal genome read coverage (with a mean of nine-fold), enabling us to interrogate the status of the majority of CpG sites in the genome. Hierarchical clustering of the methylation levels on promoter-associated CpGs effectively separated the six cell types into two major branches (Figure 8A). The same compartment clustering was demonstrated after pair-wise comparision among all different cell types of the levels of methylation of Differentiated Hypomethylated Regions (DMRs) (Figure 8B). The notion that mammary gland cells are segregated into two compartments was first suggested based on gene expression analysis of murine and human cells.

We next defined luminal-differentiated DMRs (luminal alveolar and luminal ductal cell types) (Figure 8C, far left panel) and myoepithelial differentiated DMRs (Figure 8C, left panel) and plotted the levels of DNA methylation for H2b-GFP+ cells for the same regions. Patterns of DNA methylation of H2b-GFP+ cells greatly overlapped with those of basal differentiated cells, supporting the idea that a MaSC-enriched population is biased towards the basal compartment. Conversely, analysis of H2b-GFP DNA methylation levels in luminal progenitor DMRs (Figure 8C, right panel) and in myoepithelial progenitors DMRs (Figure 8C, far right panel) revealed a more intermediate methylation status, but still basally-biased, at regions where luminal progenitors and basal progenitors showed opposing methylation patterns. This observation could suggest that differentiation from a more stem-like cell type to a more lineage-committed cell type involves both acquisition and loss of DNA methylation. Regulation of epigenetic mechanisms at the mammary gland stem cell level is important in the control of self-renewal and differentiation, since the default condensed methylation levels in stem cells accommodate changes in DNA methylation that would dictate lineage specificity, a hypothesis experimentally

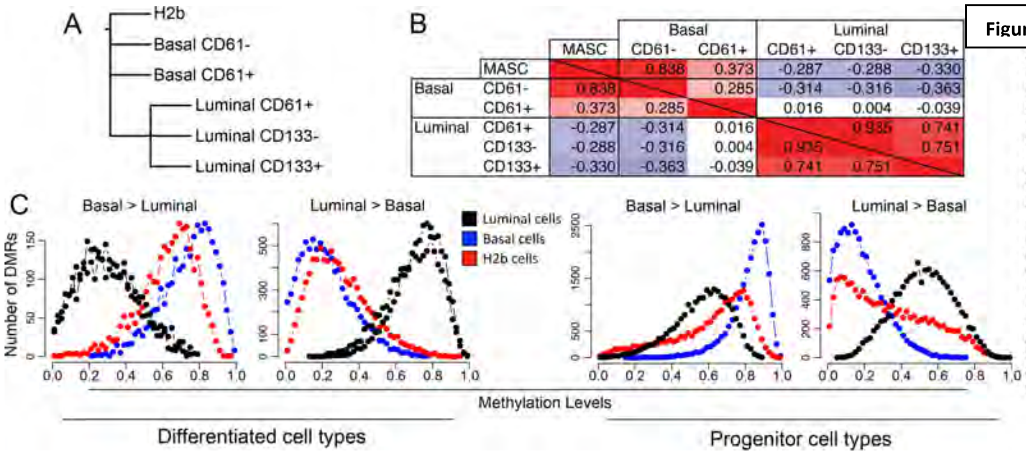


Figure 8: DNA methylation status across mammary gland cell types. (A) Promoter methylation levels. Approximately 15,000 promoters, with minimal 20 CpGs each and covered by at least one read, were analyzed followed by Pearson correlation calculation. (B) HMR Pearson Correlation Heatmap. (C) DMR methylation levels across H2b-GFP positive cells. DMRs were calculated comparing Basal Differentiated cells to Luminal Differentiated cells (left panel) and Basal Progenitor cells to Luminal progenitor cells (right panel) and vice-versa. The average methylation levels were plotted for each DMR. For the same regions, average methylation levels of H2b-GFP positive cells were collected and plotted.

supported in variety of tissues. In order to understand how DNA methylation orchestrates mammary gland cell differentiation or lineage specification we carried out RNAseq for each one of the cell types and computed their RPKM values. We next defined three sets of differentially expressed genes: H2b upregulated genes, basal upregulated genes (all genes commonly upregulated in H2b-GFP cells, myoepithelial progenitor cells and myoepithelial differentiated cells), and luminal upregulated genes (all genes commonly upregulated in luminal progenitor cells, luminal alveolar cells and luminal ductal cells). Each set contained approximately 50 genes (Figure 9A). We next collected data regarding methylation levels surrounding the transcription start site (TSS) of genes upregulated in each one of these gene pools (Figure 9B). In all six mammary gland cell types, genes differentially expressed in H2b cells displayed unchanged DNA methylation levels upstream of the TSS and slightly lower levels downstream of the TSS relative to genes that were not differentially expressed. The landscape of methylation levels across the TSS of upregulated genes displayed a greater degree of differential methylation 1-2kb downstream of the TSS. Collectively, our results contributed to the elaboration of the first mouse mammary methylome and provided important insights about the dynamics of DNA methylation across a spectrum of mammary gland cell types. We are currently analyzing DNA methylation libraries from CD1d-isolated MaSCs to further improve our knowledge of DNA methylation dynamics of mammary gland cells.

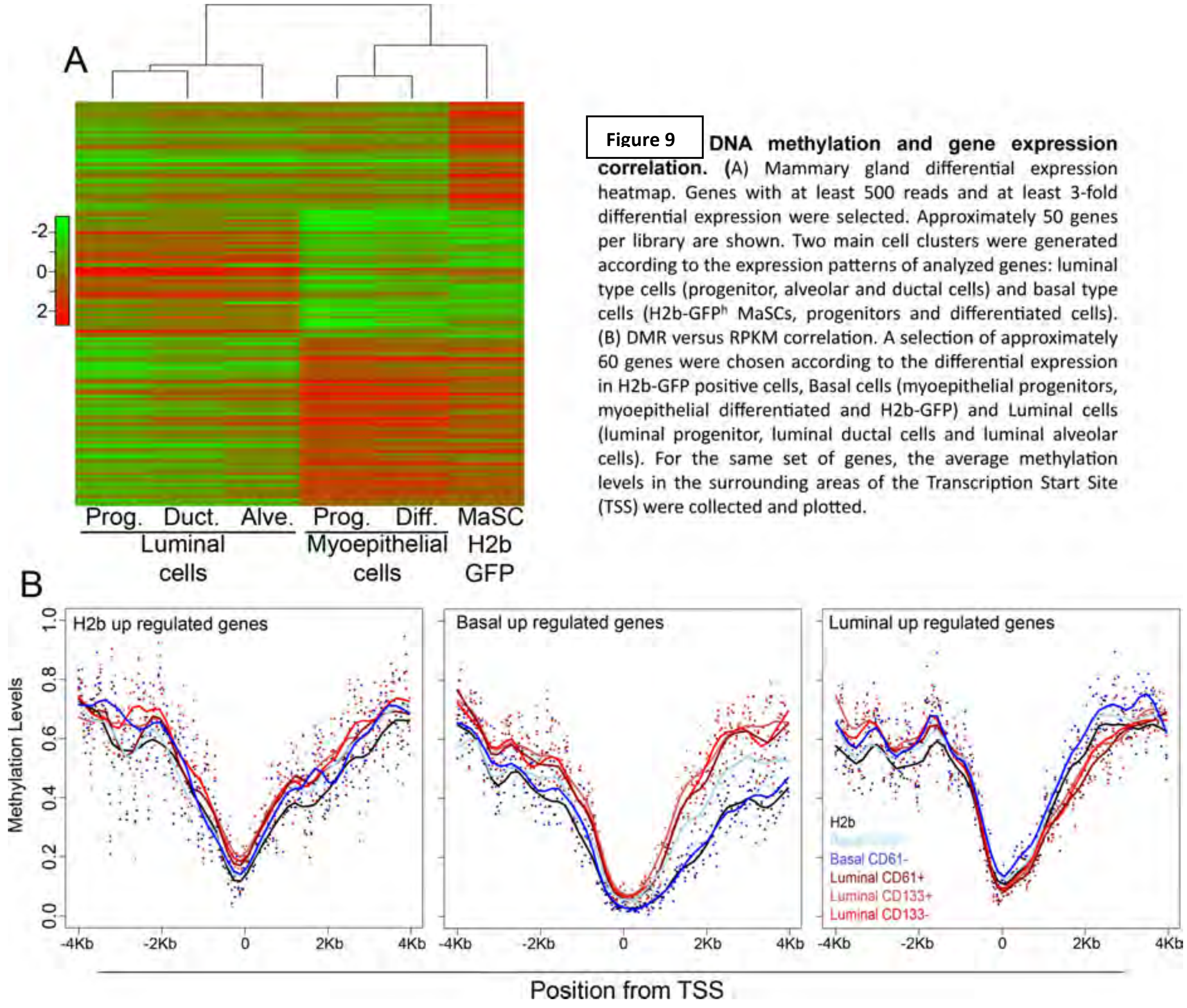
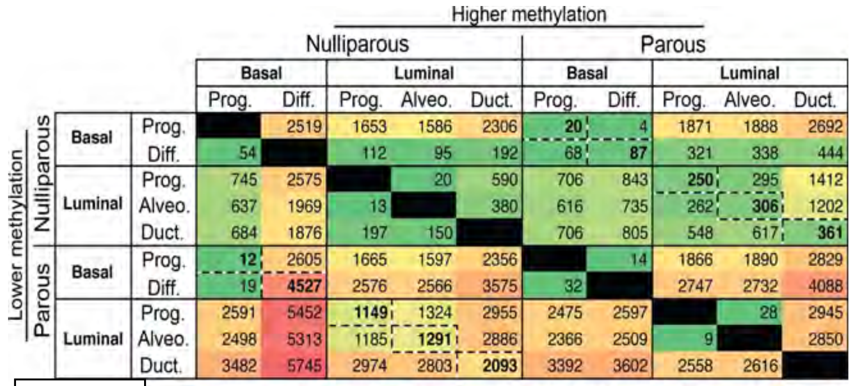


Figure 9 DNA methylation and gene expression correlation. (A) Mammary gland differential expression heatmap. Genes with at least 500 reads and at least 3-fold differential expression were selected. Approximately 50 genes per library are shown. Two main cell clusters were generated according to the expression patterns of analyzed genes: luminal type cells (progenitor, alveolar and ductal cells) and basal type cells (H2b-GFP^{h} MaSCs, progenitors and differentiated cells). (B) DMR versus RPKM correlation. A selection of approximately 60 genes were chosen according to the differential expression in H2b-GFP positive cells, Basal cells (myoepithelial progenitors, myoepithelial differentiated and H2b-GFP) and Luminal cells (luminal progenitor, luminal ductal cells and luminal alveolar cells). For the same set of genes, the average methylation levels in the surrounding areas of the Transcription Start Site (TSS) were collected and plotted.

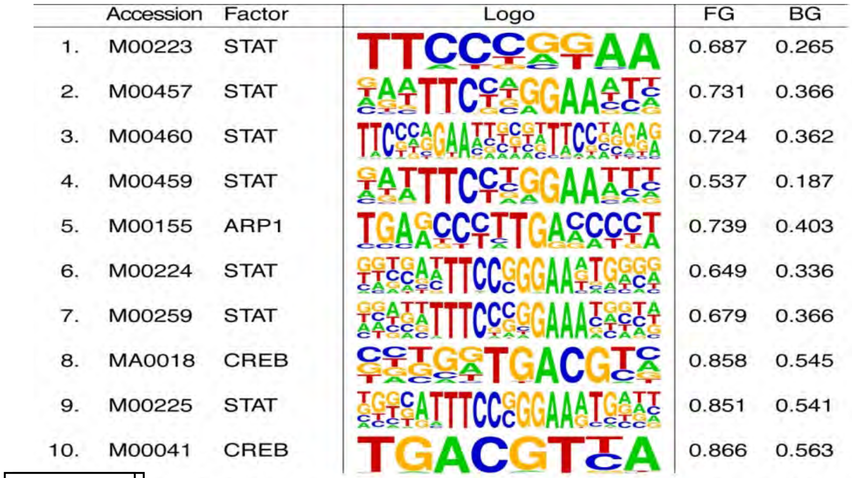
Having documented the DNA methylation signature of all mammary cells from nulliparous (virgin) mammary gland, we next generated parous mammary methylome libraries using the same cell sorting strategy described above. Female mice were allowed two full pregnancy cycles, including birth, nursing and full involution (two months). Due to increased cell division rates, no H2b-GFP+ cells were present in the glands of parous mice. We are currently preparing DNA methylation libraries from CD1d-isolated MaSCs to investigate the effects of pregnancy in the MaSC compartment. Genomic coverage for the parous libraries resembles that achieved for the nulliparous methylome (approximately 9-fold coverage).

In order to map DMRs we analyzed both libraries in a bidirectional pair-wise fashion, by comparing each cell to its corresponding cell type before and after pregnancy (Figure nulliparous DMRs (lower methylation levels before pregnancy) was substantially smaller (dashed line, upper right side) than the number of parous DMRs (lower methylation levels after pregnancy), suggesting a dramatic loss of methylation by most cell types post-pregnancy (dashed line, lower left side). The loss of methylated sites after pregnancy could translate into changes in gene expression, an observation that was previously suggested to be the case for a small subset of genes. We are currently comparing RNAseq libraries of all mouse mammary cell types before and after pregnancy to more precisely identify the changes in gene expression patterns.



Lower methylation			Higher methylation: Nulliparous					Parous				
			Basal		Luminal			Basal		Luminal		
			Prog.	Diff.	Prog.	Alveo.	Duct.	Prog.	Diff.	Prog.	Alveo.	Duct.
Nulliparous	Basal	Prog.		2519	1653	1586	2306	**20**	4	1871	1888	2692
		Diff.	54		112	95	192	68	**87**	321	338	444
	Luminal	Prog.	745	2575		20	590	706	843	**250**	295	1412
		Alveo.	637	1969	13		380	616	735	262	**306**	1202
		Duct.	684	1876	197	150		706	805	548	617	**361**
Parous	Basal	Prog.	**12**	2605	1665	1597	2356		14	1866	1890	2829
		Diff.	19	**4527**	2576	2566	3575	32		2747	2732	4088
	Luminal	Prog.	2591	5452	**1149**	1324	2955	2475	2597		28	2945
		Alveo.	2498	5313	1185	**1291**	2886	2366	2509	9		2850
		Duct.	3482	5745	2974	2803	**2093**	3392	3602	2558	2616	

Figure 10 **Nulliparous x Parous DMR analysis**. Total CpG methylation status were analyzed and hypomethylated regions (HMRs) calculated. Nulliparous or parous HMRs were analyzed according to Parous or nulliparous methylation status, and regions with differential methylated (DMRs, at least 10 significantly differing CpG per DMR) were plotted.



	Accession	Factor	Logo	FG	BG
1.	M00223	STAT		0.687	0.265
2.	M00457	STAT		0.731	0.366
3.	M00460	STAT		0.724	0.362
4.	M00459	STAT		0.537	0.187
5.	M00155	ARP1		0.739	0.403
6.	M00224	STAT		0.649	0.336
7.	M00259	STAT		0.679	0.366
8.	MA0018	CREB		0.858	0.545
9.	M00225	STAT		0.851	0.541
10.	M00041	CREB		0.866	0.563

Figure 11 **Transcription factor enrichment analysis.** Parous luminal DMRs were analyzed according to their enrichment for transcription factor binding sites . Top 5 most abundant motifs are displayed. FG displays the likehood of a specific nucleotide combination to serve as binding site, whereas BG displays the likehood of neighboring nucleotide sequences to serve as binding sites.

The luminal compartment, exhibited the most DNA methylation changes after pregnancy. The extent of these differences was reflected in the number of acquired hypomethylated sites (DMRs) but was most importantly also correlated with DNA methylation loss. Among all luminal cell types (progenitor cells, alveolar cells and ductal cells), a great portion of shared DMRs occurred nearby binding sites recognized by the STAT transcription factor (Figure 11), which might suggest a role for this family of proteins during pregnancy, lactation and involution. Interestingly, the STAT-associated DMRs were further enriched for a class of genes with known roles in apoptosis and potential antitumor activity.

STAT transcription factors have been previously suggested to play important roles in mammary cells. STAT5a and STAT5b have been implicated in the transcriptional activation of milk protein in response to progesterone levels, although STAT5a and STAT5b protein levels only slightly increased during pregnancy and lactation. Lack of STAT5a expression resulted in decreased lobuloalveolar development during normal mammopoiesis and blocked milk production in the first pregnancy, although ductal density and milk production resumed at the onset of a second pregnancy. Conversely, overexpression of full-length STAT5a not only induced lobuloalveolar development but also delayed involution, whereas overexpression of a c-terminally truncated form of STAT5a accelerated apoptosis during involution. Further understanding of how STAT transcription factors regulate gene expression in mammary cells, including how this regulation is susceptible to changes during pregnancy could provide a clear foundation for evaluating the role of STATs in pregnancy-induced breast cancer protection.

Identification of potential new breast tumor suppressor genes

SNP-based linkage analysis in 41 non-BRCA1/2 families identified several candidate regions containing breast cancer susceptibility genes (6). The large size of the candidate regions and the high number of genes described within them, led us to couple linkage analysis with high throughput sequencing-based mutational screening as a new strategy for variation detection. For this study we used hybrid selection, a method developed in my lab (7), of discrete genomic intervals on custom-designed microarrays as an exon-specific enrichment of all the genes located within two complete candidate regions on chromosomes 3 and 6. This study was performed in collaboration with Javier Benitez (Spanish National Cancer Research Center). This work was published in *PLOS One* in 2010 (see attachment).

Identification of structural variations in breast cancer

Cancer is a disease driven by genetic alterations such as single nucleotide variations (SNVs), structural variations (SVs), and anueploidy. Large scale SVs including deletions, insertions, inversions, tandem duplications, translocations, and more complex rearrangements could alter gene function to favor tumorigenesis. We have developed a method to detect structural variations by constructing/Illumina sequencing of mate-pair genomic DNA libraries of 5kb fragment size (see Figure 12).

Figure 12

Mate-Pair libraries scan Structural Variation in low resolution

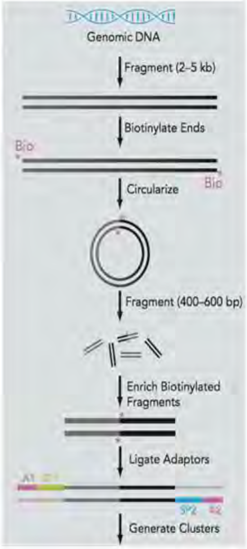
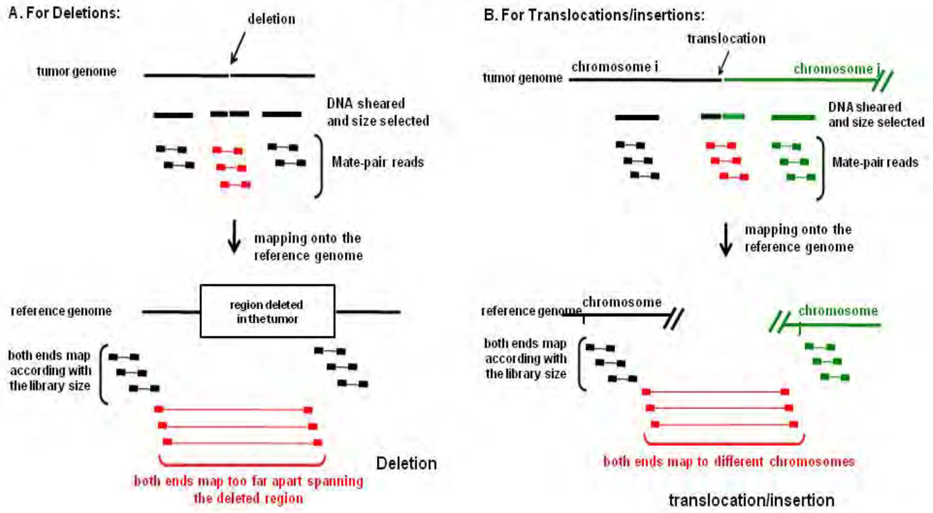
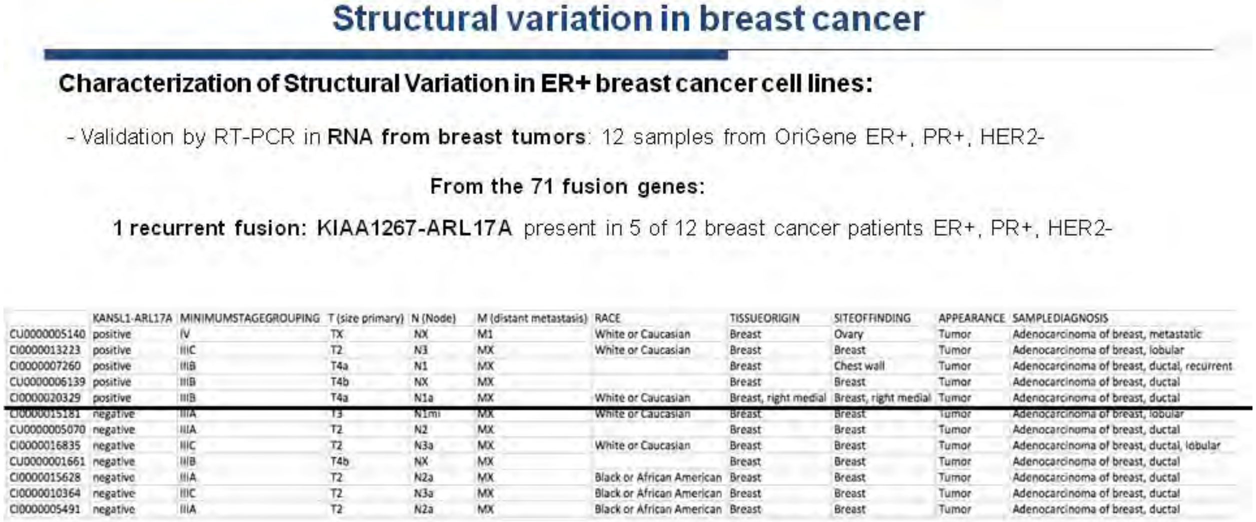


Figure 13


A. For Deletions:
deletion
tumor genome
DNA sheared and size selected
Mate-pair reads
mapping onto the reference genome
reference genome
region deleted in the tumor
both ends map according with the library size
both ends map too far apart spanning the deleted region
Deletion
B. For Translocations/insertions:
translocation
tumor genome
chromosome i
chromosome j
DNA sheared and size selected
Mate-pair reads
mapping onto the reference genome
reference genome
chromosome
chromosome
both ends map according with the library size
both ends map to different chromosomes
translocation/insertion

We set of out to characterize structural variation in a panel of ER-positive human breast cancer cell lines as a proof of principle. We constructed mate pair libraries from seven cell lines (MDA-MB-134IV, MDA-MB-175VI, MDA-MB-415, MCF-7, CAMA-1, T47D, ZR-75-1), and using human mammary epithelial cells (HMECs) as negative control (reference genome). Paired-end mapping detected SV when both ends of the mate-pair map in the reference genome according to fragment size of the library (see Figure 13). We used the Hydra algorithm (8), which allows for identification of SV from mate-pairs with discordant mapping positions such that multiple mappings can be considered for characterization of SV in repetitive regions of the genome. We honed our technique to specifically detect translocations. Our criteria for detecting a translocation requires mate-pairs to be derived from different chromosomes and represented by at least 3 different mate-pairs for each candidate SV. We excluded those SVs that are found in HMECs. We identified 53 translocations in MDA-MB-134IV, 28 in MDA-MB-175VI, 29 in MDA-MB-415, 428 in MCF7, 93 in CAMA-1, 90 in T47D, and 47 in ZR-75-1. Spectral karyotyping (SKY) was performed to confirm whether the putative translocations can be visibly detected. In addition, we performed RNA-seq analysis to detect novel gene fusions using a gene-fusion transcript discovery software called deFuse. Eighty candidate fusion genes were identified at the mRNA level and fifty-two of them were also found at the DNA level by mate-pair analysis. Seventy-one of the candidates were validated by RT-PCR and were present in the breast cancer cell lines and not HMEC. Next we screened RNAs from 12 $ER^+/HER2^-$ human patient samples for the occurrence of the 71 candidate gene fusions by RT-PCR. One fusion gene candidate, KIAA1267(KANSL1)-ARL17A, was present in 5 of 12 breast cancer patient samples (Figure 14). This fusion gene is a product of tandem duplication on chromosome 17 and is composed of the first 3 exons of the KANSL1 gene fused to the last 2 exons of ARL17. Recent data (unpublished) from another group have found that this gene fusion presents itself as a genetic variation in the human genome in one quarter of Caucasion populations. In particular, the gene product was detected in 12% of breast cancers in this group. We will continue to pursue this finding by other funding sources. This study demonstrates that this technology can be applied to detect SVs in the genome.



Structural variation in breast cancer

Characterization of Structural Variation in ER+ breast cancer cell lines:

- Validation by RT-PCR in **RNA from breast tumors**: 12 samples from OriGene ER+, PR+, HER2-

From the 71 fusion genes:

1 recurrent fusion: KIAA1267-ARL17A present in 5 of 12 breast cancer patients ER+, PR+, HER2-

	KANSL1-ARL17A	MINIMUMSTAGEGROUPING	T (size primary)	N (Node)	M (distant metastasis)	RACE	TISSUEORIGIN	SITEOFFINDING	APPEARANCE	SAMPLEDIAGNOSIS
CU0000005140	positive	IV	TX	NX	M1	White or Caucasian	Breast	Ovary	Tumor	Adenocarcinoma of breast, metastatic
CI0000013223	positive	IIIC	T2	N3	MX	White or Caucasian	Breast	Breast	Tumor	Adenocarcinoma of breast, lobular
CI0000007260	positive	IIIB	T4a	N1	MX		Breast	Chest wall	Tumor	Adenocarcinoma of breast, ductal, recurrent
CU0000006139	positive	IIIB	T4b	NX	MX		Breast	Breast	Tumor	Adenocarcinoma of breast, ductal
CI0000020329	positive	IIIB	T4a	N1a	MX	White or Caucasian	Breast, right medial	Breast, right medial	Tumor	Adenocarcinoma of breast, ductal
CI0000015181	negative	IIIA	T3	N1mi	MX	White or Caucasian	Breast	Breast	Tumor	Adenocarcinoma of breast, lobular
CU0000005070	negative	IIIA	T2	N2	MX		Breast	Breast	Tumor	Adenocarcinoma of breast, ductal
CI0000016835	negative	IIIC	T2	N3a	MX	White or Caucasian	Breast	Breast	Tumor	Adenocarcinoma of breast, ductal, lobular
CU0000001661	negative	IIIB	T4b	NX	MX		Breast	Breast	Tumor	Adenocarcinoma of breast, ductal
CI0000015628	negative	IIIA	T2	N2a	MX	Black or African American	Breast	Breast	Tumor	Adenocarcinoma of breast, ductal
CI0000010364	negative	IIIC	T2	N3a	MX	Black or African American	Breast	Breast	Tumor	Adenocarcinoma of breast, ductal
CI0000005491	negative	IIIA	T2	N2a	MX	Black or African American	Breast	Breast	Tumor	Adenocarcinoma of breast, ductal

Figure 14. Validation of 71/80 fusion genes by RT-PCR from RNA of human breast tumor samples. One fusion gene candidate, KIAA1267 (or KANSL)-ARL17A were scored in 5 of 12 human patient samples.

KEY RESEARCH ACCOMPLISHMENTS

- Development of a highly parallel, large scale, high throughput method to measure efficacy of shRNAs.
- Generation of the largest dataset of approximately 250,000 measurements of shRNA efficacy.
- Derivation of a shRNA-based prediction algorithm, called shERWOOD, that essentially predicts the results of functional, sensor testing of shRNAs *in silico*.
- Constructed sequence-verified, V4, shRNA libraries at full genome coverage for human and mouse.
- Development of a fifth-generation shRNA design, called 'ultramir' by optimization of the miR30 miRNA scaffold.
- Construction of the V5 shRNA libraries began during the first quarter of 2014.
- Completed a comprehensive, genome-wide, functional profiling study for genes involved in proliferation and survival in all three treatment subtypes (ER^+, $HER2^+$, Triple-negative) of breast cancer.
- Discovery of potential target genes to sensitize Her2-positive, trastuzumab-resistant cell models of breast cancer.
- Discovery of potential genes to target BRCA-mutant breast cancer cells.
- Isolation/purification of $H2b\text{-}GFP^h$ MaSCs and demonstrated its mammary gland reconstitution potential
- Defining the molecular hierarchy of mammary differentiation yielded refined markers of mammary stem cells.
- Development of a methodology for whole genome, shotgun, bisulfate sequencing to generate single nucleotide-resolution methylation profiles.
- Characterization of DNA methylation and gene expression patterns of mammary gland cells of nulliparous and parous mice uncovered STAT transcription factors as being involved in mammary gland development during pregnancy.
- Development of a method to detect structural variation in the genome.
- Study to uncover structural variations in breast cancer cells led to the identification of a gene fusion called KANSL1-ARL17A.
- Discovery of miR-93 in modulating fate of tumor-initiating cells by regulating the proliferation and diffentiation states.
- Enforced expression of miR-93 competely blocked tumor development in mammary fat pads and development of metastases following intracardiac injection in mouse xenograft models.

REPORTABLE OUTCOMES

- Developed a shRNA-specific prediction algorithm called shERWOOD.
- Completed construction of fourth-generation shRNA libraries targeting human and mouse genomes. These resources have been made available to the scientific community.
- Began construction of our V5 shRNA libraries. These clones will also be made publicly available. Completion is expected to be sometime during 2015.
- Publication: Deep sequencing of Target Linkage Assay identified regions in familial breast cancer: methods, analysis pipeline, and troubleshooting.

Rosa-Rosa J.M.[1], Gracia-Aznárez F.J., Hodges E., Pita G., Rooks M., Xuan Z., Bhattacharjee A., Brizuela L., Silva J.M., Hannon G.J., Benitez J. (2010). ***PLOS One***, 5(4):e9967.

- Publication: Functional identification of optimized RNAi triggers using a massively parallel sensor assay.
  Fellmann C., Zuber J., McJunkin K., Chang K., Malone C.D., Dickins R.A., Xu Q., Hengartner M.O., Elledge S.J., Hannon G.J., Lowe S.W. (2011). ***Molecular Cell***, 41(6):733-46.
- Publication: Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment.
  Hodges E.[1], Molaro A., Dos Santos C.O., Thekkat P., Song Q., Uren P.J., Park J., Butler J., Rafii S., McCombie W.R., Smith A.D., Hannon G.J. (2011). ***Molecular Cell***, 44(1):17-28.
- Publication: MicroRNA93 regulates proliferation and differentiation of normal and malignant breast stem cells.
  Liu S.[1], Patel S.H., Ginestier C., Ibarra I., Martin-Trevino R., Bai S., McDermott S.P., Shang L., Ke J., Ou S.J., Heath A., Zhang K.J., Korkaya H., Clouthier S.G., Charafe-Jauffret E., Birnbaum D., Hannon G.J., Wicha M.S. (2012). ***PLOS Genetics,*** 8(6):e1002751.
- Publication: Molecular hierarchy of mammary differentiation yields refined markers of mammary stem cells.
  dos Santos C.O., Rebbeck C., Rozhkova E., Valentine A., Samuels A., Kadiri L.R., Osten P., Harris E.Y., Uren P.J., Smith A.D., Hannon G.J. (2013). ***Proc. Natl. Acad. Sci. USA, 110***,(18):7123-30.
- Publication under review (***Molecular Cell***): A computational algorithm to predict shRNA potency.
  Simon R.V. Knott[1], Ashley Maceli[1], Nicolas Erard[1], Kenneth Chang[1], Krista Marran, Xin Zhou, Assaf Gordon, Osama El Demerdash, Elvin Wagenblast, Christof Fellmann&, and Gregory J. Hannon.

CONCLUSIONS

The overall goal of this innovator award is to apply RNAi and other technologies my lab have developed over the past 10 years toward improving breast cancer therapy. This is achieved through three main aims. The first aim is to apply our whole genome shRNA screening approach to screen human breast cancer cell lines representing all three major clinical subgroups of the disease. This unbiased approach identified pathways that support growth and survival of breast tumor cells, both during tumor progression and in the face of therapeutic challenge. The dataset also addresses drug resistance mechanisms of herceptin, lapatinib, tamoxifen, and estrogen-deprivation therapy. This study produced lethality signatures for all types of breast cancer cells and will be a rich resource of data to discover new therapeutic leads for novel treatments of breast cancer. We have embarked on several in-depth investigations on selected candidate genes for targeting the HER2-positive and triple-negative subtypes. This dataset will be made available to the scientific community. Our second aim is to address the role of the tumor-initiating compartment of breast cancer cells in the context of miRNAs. While we have studied how particular miRNAs impact the self-renewing populations of breast cancer cells, we are well aware that this compartment is very heterogeneous and complex. To this end, we set out to identify normal mammary gland stem cells (in parous and nulliparous animals), using a strategy that does not depend on known cell surface markers from other types of stem cells. We demonstrated that this slow-cycling compartment (H2b-GFP^h) has mammary gland regeneration potential and have profiled its transcriptome and methylome. Furthermore, we have identified a new cell surface marker, CD-1, for this self-renewing population. Currently we are researching the function of how H2b-GFP^h populations in breast cancer cells contribute to tumor progression and therapy resistance. Our hope is that this knowledge will lead to new therapeutic approaches to breast cancer. Our last aim is to apply our next-generation re-sequencing approach (hybrid selection) to hone in on candidate chromosomal regions containing breast cancer susceptibility genes that were previously identified by SNP-based linkage analysis on BRCA1/2 families. This method selectively enriched for exons located within those regions to identify new tumor suppressor genes responsible for hereditary breast cancer.

With this award, we have managed to produce some very promising outcomes and have been very productive in terms of scientific achievement (5 published articles and 1 currently under review). While it is beyond the scope of this award to translate our new findings to clinical studies, the data from this grant will pave the way for promising new therapeutic treatments for breast cancer.

REFERENCES

1. Silva J.M., Li M.Z., Chang K., Ge W., Golding M.C., Rickles R.J., Siola D., Hu G., Paddison P.J., Schlabach M.R., Sheth N., Bradshaw J., Burchard J., Kulkarni A., Cavet G., Sachidanandam R., McCombie W.R., Cleary M.A., Elledge S.J., Hannon G.J. (2005), *Nature Genetics*, 37(11):1281-8. Second-generation shRNA libraries covering the mouse and human genomes.

2. Fellmann C., Zuber J., McJunkin K., Chang K., Malone C.D., Dickins R.A., Xu Q., Hengartner M.O., Elledge S.J., Hannon G.J., Lowe S.W., (2011), *Molecular Cell*, 41(6)733- 46. Functional identification of optimized RNAi triggers using a massively parallel sensor assay.

3. Vert J.P., Foveau N., Lajaunie C., Vandenbrouck Y. (2006), *BMC Bioinformatics*, 7:520. An accurate and interpretable model for siRNA efficacy prediction.

4. Ibarra I., Erlich Y., Muthuswamy S.K., Sachidanandam R., Hannon G.J. (2007), *Genes & Development*, 21(24):3238-43. A role for microRNAs in maintenance of mouse mammary epithelial progenitor cells.

5. Hodges E., Molaro A., Dos Santos C.O., Thekkat P., Song Q., Uren P.J., Park J., Butler J., Rafii S., McCombie W.R., Smith A.D., Hannon G.J. (2011). *Molecular Cell*, 44(1):17-28. Directional DNA methylation changes and complex intermediate states accompany lineage specificity in the adult hematopoietic compartment.

6. Rosa-Rosa J.M., Pita G., Urioste M., Llort G., Brunet J., et al. (2009), *Am J Hum Genet*, 84: 115–122. Genome-wide linkage scan reveals three putative breast-cancer-susceptibility loci.

7. Hodges E., Rooks M., Xuan Z., Bhattacharjee A., Benjamin Gordon D., Brizuela L., Richard McCombie W., Hannon G.J. (2009), *Nature Protocols*, 4(6):960-74. Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing.

8. Quinlan A.R., Clark R.A., Sokolova S., Leibowitz M.L., Zhang Y., Hurles M.E., Mell J.C., Hall I.M. (2010), *Genome Research*, 20(5):623-35. Genome-wide mapping and assembly of structural variant breakpoints in the mouse genome.

APPENDICES

Attached are 5 published articles and 1 manuscript currently under review.

Deep Sequencing of Target Linkage Assay-Identified Regions in Familial Breast Cancer: Methods, Analysis Pipeline and Troubleshooting

Juan Manuel Rosa-Rosa[1], Francisco Javier Gracia-Aznárez[1], Emily Hodges[2], Guillermo Pita[3], Michelle Rooks[2], Zhenyu Xuan[4], Arindam Bhattacharjee[5], Leonardo Brizuela[5], José M. Silva[6], Gregory J. Hannon[2], Javier Benitez[1,3]*

1 Human Genetics Group, Spanish National Cancer Research Centre (CNIO), Madrid, Spain, **2** Howard Hughes Medical Institute, Cold Spring Harbor Laboratory (CSHL), Cold Spring Harbor, New York, United States of America, **3** Genotyping Unit (CEGEN), Spanish National Cancer Research Centre, Madrid, Spain, **4** Department of Molecular and Cell Biology, The University of Texas, Richardson, Texas, United States of America, **5** Agilent Technologies, Inc., Santa Clara, California, United States of America, **6** Irving Cancer Research Center, New York, New York, United States of America

Abstract

Background: The classical candidate-gene approach has failed to identify novel breast cancer susceptibility genes. Nowadays, massive parallel sequencing technology allows the development of studies unaffordable a few years ago. However, analysis protocols are not yet sufficiently developed to extract all information from the huge amount of data obtained.

Methodology/Principal Findings: In this study, we performed high throughput sequencing in two regions located on chromosomes 3 and 6, recently identified by linkage studies by our group as candidate regions for harbouring breast cancer susceptibility genes. In order to enrich for the coding regions of all described genes located in both candidate regions, a hybrid-selection method on tiling microarrays was performed.

Conclusions/Significance: We developed an analysis pipeline based on SOAP aligner to identify candidate variants with a high real positive confirmation rate (0.89), with which we identified eight variants considered candidates for functional studies. The results suggest that the present strategy might be a valid second step for identifying high penetrance genes.

Citation: Rosa-Rosa JM, Gracia-Aznárez FJ, Hodges E, Pita G, Rooks M, et al. (2010) Deep Sequencing of Target Linkage Assay-Identified Regions in Familial Breast Cancer: Methods, Analysis Pipeline and Troubleshooting. PLoS ONE 5(4): e9976. doi:10.1371/journal.pone.0009976

Editor: Jörg Hoheisel, Deutsches Krebsforschungszentrum, Germany

Received January 20, 2010; **Accepted** March 9, 2010; **Published** April 2, 2010

Funding: This project was funded by the Spanish Association Against Cancer (AECC) with no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript; and Agilent, Inc. supplying the arrays that can be used for the hybrid selection procedures described in this manuscript. JMRR's short stay at CSHL was supported by funding from the Ministerio de Eduacion y Cultura, the Universidad Autonoma de Madrid, and the European Association for Cancer Research, with no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Competing Interests: The authors declare the following conflicts of interest: Leonardo Brizuela and Arindam Bhattacharjee are employees of Agilent, Inc. The authors confirm that this does not alter their adherence to all the PLoS ONE policies on sharing data and materials.

* E-mail: jbenitez@cnio.es

Introduction

Breast cancer (BC [MIM #114480]) is the most frequent malignancy among women, with approximately one million new cases per year around the world [1]. About five percent of all BC cases are considered to be hereditary, and mutations in either the *BRCA1* [MIM +113705] or the *BRCA2* [MIM +600185] gene account for 25–30% of these cases [2]. Thus, about 70% of BC families remain unexplained, and are known as non-BRCA1/2 families [3]. In this regard, several linkage studies have been performed during the last years on familial BC, and many candidate regions that may contain BC susceptibility genes have been described. However, mutational screenings in linkage assay-identified regions using the classical candidate-gene approach did not identify any clear pathogenic variants [4,5,6]. Therefore, new strategies seem to be necessary.

Massive parallel sequencing technology allows nowadays the development of studies unachievable a few years ago. Despite the fact that the advantages of this technology were evidenced in each of the published studies based on it, the analysis of all data generated by this process remains a hard task to face. The first step in the regular analysis protocol is based on the alignment of millions of short sequences obtained from each run. For that reason, during the past years many computer tools have been developed to improve the accuracy of this process [7,8,9]. One of the main obstacles is the specificity in the analysis of the data to obtain the output required by any given study. For example, the identification of novel variants, chromosomal translocations, or transcription factor target sites are some of the aims of the many studies that can be performed using high throughput sequencing technologies, and each of them would require a specific analysis pipeline.

A major success in the use of this technology was the re-sequencing of the whole human genome, published in several studies in the past years [10,11,12,13,14]. The results showed a higher complexity level of the human genome, with the appearance of many new variants, short length insertions and deletions and, even, small inversions. However, the amount of time required for re-sequencing large genomes and its high economic cost make it impractical as a common laboratory technique. For that reason, the search for specific sequence-enrichment protocols applicable to massive parallel sequencing has been an important goal during these years. One of the approaches developed [15] and improved [16] to obtain this specific enrichment is the technique of selective exon-capture or hybrid selection, based on high-density tiling DNA microarrays.

In a previous study, we performed a SNP-based linkage analysis in 41 non-BRCA1/2 families and obtained several candidate regions for containing BC susceptibility genes [17]. The large size of the candidate regions and the high number of genes described within them, led us to couple linkage analysis with high-throughput sequencing-based mutational screening as a new strategy for variation detection. In the present study, we used hybrid selection of discrete genomic intervals on custom-designed microarrays as an exon-specific enrichment of all the genes located within two complete candidate regions on chromosomes 3 and 6, that presented a suggestive linkage LOD score (LOD>2.2). We also describe the complex analysis pipeline based on SOAP aligner, developed to analyse all the data obtained, and present eight variants that are candidates for playing a role in the development of familial BC.

Results

Regions and capture

We performed a complete mutational screening via single-end massive parallel sequencing technology on the 128 known genes located on two different chromosomal regions, previously described as candidates for containing a breast cancer susceptibility gene [17]. The first region is located on 3q25, extends over 10.8 Mb (from 160,964 to 171,786 Kb), and contains 69 known genes; the second region is located on 6q24 and spans 6.5 Mb (from 146,078 to 152,515 Kb), containing 59 known genes. To selectively sequence the coding region of these genes, we used hybrid selection on tiling microarrays [16], which was validated via qPCR (data not shown).

Reads, coverage and depth

DNA samples from 20 affected individuals, belonging to 9 different non-BRCA1/2 families, and 4 healthy unrelated individuals from the control population were used for the analysis process. More than 102 million reads were obtained from the affected individuals and almost 22.4 million reads were obtained from the control individuals (see Table 1). The average number of reads per affected individual was 5.14 million, and this number was increased to 5.21 million when control individuals were taken into account. We used SOAP v1.0 to align our data set, obtaining an average of 91.58% aligned sequences against the whole genome with ≤2 mismatches. Among these, an average of 39.99% matched our candidate coding regions. Thus, since the total number of base pairs covered by the tiling array is 0.014% of the genome length (434,039 bp), the average enrichment was approximately 2.85 thousand times, calculated as the percentage of the sequences aligned to the genome that successfully matched to the candidate regions (39.99) divided by the percentage of the genome length that candidate regions represent (0.014).

We calculated the coverage (number of bases covered after the alignment) per individual and obtained an average of 98.25% of candidate bases covered for the affected and the control individuals (Table 1). We did not observe significant coverage differences between the candidate coding regions located on chromosomes 3 and 6. Importantly, a lack of correlation between the coverage and the total number of sequences was observed ($r^2 = 0.013$, Figure S1A). However, this lack of correlation turned into a logarithmic trend when the number of sequences aligned to the candidate coding regions was used ($r^2 = 0.69$, Figure S1B).

In order to know the power to confidentially detect possible causal variants, we calculated the global depth (number of sequences that cover a single base) for every base along the candidate coding regions per individual. As expected, a strong correlation ($r^2 = 0.96$) was found between the global depth and the number of reads aligned to the candidate coding regions (Figure S1C). Taking into account only those bases that presented coverage (depth >0), from both affected and controls individuals, the mean and the median of the depth were 37 and 34 respectively, showing a strong correlation between them ($r^2 = 0.98$, Figure S1D).

Data quality control

a) Coverage homogeneity. We calculated the mean and the median of the depth for stretches of 15 bases along the candidate regions, and obtained the log-ratio of the median between each affected individual and the control pool. The mean log-ratio for the global data was −0.06 with a standard deviation of 0.55 (Table 2), showing a homogeneous distribution of the coverage between affected and control individuals. We calculated the upper and lower threshold for each individual (Table 2) and we identified several regions where the coverage differed between the affected individuals and the control pool (data not shown), most of them flanking the candidate coding regions (which are usually low coverage regions). In addition, the strong correlation ($r^2 = 0.99$, Figure S1E) between the median and the mean of the coverage for these regions in the dataset supported that these differences in coverage are not due to extreme values within the same coding region but are due to chance, ruling out potential problems in the capture step.

b) Score. Genotype calling accuracy was demonstrated elsewhere (>90% of the known SNPs) by using a HapMap sample in the exon-capture report [16]. In order to test the suitability of our analysis pipeline when looking for unknown polymorphisms, we used the candidate SNPs from a single family obtained using different Depth Score (DS, see Material and Methods) thresholds for both the affected samples and the control pool data (Table 3). We observed that from DS = 50 onwards for the samples, the DS threshold used for the control pool data had no effect on the number of candidate SNPs, highlighting the specificity of our DS. In order to be as conservative as possible, we considered that the best False Positive/False Negative (FP/FN) relationship was for a DS >50 for samples and a DS >14 for the control pool. Additionally, we performed an analysis using MAQ software in a subgroup of families and we observed that there was no correlation between MAQ variant score and Sanger confirmation (Table 4). These results demonstrated the higher accuracy of our analysis methodology compared to the algorithm used by MAQ.

Variant identification

We included a first filter (see Material and Methods) in our SOAP-based SNP-caller to discard the maximum amount of false positives, obtaining 99% of SNP variants discarded (Figure 1). We selected only those variants located on the chromosome of interest for each individual, resulting in an average of 71 SNPs per individual (Table 5). Subsequent filters (discarding homozygous variants, comparing to controls and comparing within members of

Table 1. Summary of high throughput sequencing data.

			Number of sequences			Depth		
Chromosome[a]	**Family**	**Individual**	**Total**	**Aligned to whole genome (%*)**	**Aligned to candidate regions (%**)**	**Coverage in %**	**Mean**	**Median**
3	**27**	**07S722**	3,123,937	2,956,483 (94.64)	1,186,611 (40.14)	98.04	26	25
		07S723	4,922,157	4,538,392 (92.20)	1,518,625 (33.46)	98.43	29	29
		07S724	4,183,568	3,954,837 (94.53)	1,515,614 (38.32)	97.89	28	26
		07S725	2,952,969	2,839,271 (96.15)	1,168,679 (41.16)	97.11	24	22
	60	**06-240**	2,652,926	2,580,914 (97.29)	882,837 (34.21)	97.96	22	20
		96-652	5,934,453	4,737,175 (79.82)	1,670,157 (35.26)	98.15	28	24
	531	**I-1408**	12,228,047	11,188,204 (91.50)	4,694,871 (41.96)	99.07	57	48
		I-904	4,293,087	3,585,982 (83.53)	1,531,322 (42.70)	97.50	30	22
	713	**07S635**	7,568,672	7,442,938 (98.34)	2,793,056 (37.53)	99.11	45	44
		07S636	7,160,552	6,889,152 (96.21)	2,574,119 (37.36)	98.94	43	42
6	**11**	**04-168**	5,734,052	5,599,100 (97.65)	2,459,740 (43.93)	98.57	43	42
		96-265	6,240,024	6,012,522 (96.35)	2,642,942 (43.96)	98.22	35	32
	40	**07S576**	2,006,661	1,667,648 (83.11)	779,723 (46.76)	97.11	18	17
		07S581	4,016,214	3,618,178 (90.09)	1,568,060 (43.34)	97.66	25	23
	929	**I-1627**	5,811,276	5,665,182 (97.49)	2,311,149 (40.80)	98.52	33	32
		I-3345	2,602,250	2,554,051 (98.15)	1,059,131 (41.47)	98.27	23	23
	990	**I-1927**	8,134,956	7,903,785 (97.16)	3,029,994 (38.34)	98.84	51	50
		I-1928	7,922,500	7,590,406 (95.81)	2,817,358 (37.12)	99.02	49	48
	1125	**I-2033**	2,747,911	2,666,280 (97.03)	1,105,059 (41.45)	97.87	24	23
		I-4347	2,517,619	2,406,505 (95.59)	1,088,350 (45.23)	97.74	24	24
		TOTAL	102.753.831	96,397,005 (93.81)	38,397,397 (39.83)			
		Average Affected	5,137,692	4,819,850 (93.63)	1,919,870 (40.22)	98.20	33	31
		Control pool	22,390,251	18,221,565 (81.38)	7,438,610 (40.82)	99.33	111	98
		Average All	5,214,336	4,775,773 (91.58)	1,909,833 (39.99)	98.25	37	34

The number of sequences and the depth values are shown for a total of 20 individuals from 9 non-BRCA1/2 families and 4 individuals from the control population (Control pool).

[a] chromosome in which linkage signal was found for each of the families.

*with respect to the total number of sequences, ** with respect to the numbers of sequences aligned to the whole genome.

doi:10.1371/journal.pone.0009976.t001

the same family) discarded almost 80% of the remaining SNPs. Then, we used a newly developed Perl script to differentiate between described and previously undescribed SNPs and also to obtain the functional consequences for each undescribed SNP, resulting in 5 undescribed variants per family on average. Since we observed that one variant could have consequences for more than one transcript/gene, subsequent filtering was performed using the consequences of the variants instead of the variants themselves. Next, we discarded intronic consequences and continued the analysis with exonic consequences only (an average of 23% of the original number of consequences). After that, each of the remaining SNPs showed only one consequence, and in the following step we checked if the remaining SNPs could be due to homology between different regions. Finally, an average of 1 (6.25%) of the SNPs shared by all the affected members of a family could be considered a strong candidate (Table 5). Finally, 8 out of 9 candidate SNPs were confirmed by Sanger sequencing, supporting a high real positive confirmation rate (0.89). Information data about variant position, gene, type of change, Alamuth prediction, and gene function from the 8 final candidate SNPs is shown in Table 6. The mean DS for confirmed variants was 115 (74–185), whereas the mean DS for ruled out variants was 56.5 (56–57), which means that our DS score is a suitable variable for the filtering process.

Regarding indel variants, we followed the same analysis pipeline as with SNP variants. Even though 12 bp gaps were allowed (maximum length allowed for an indel variant), an average of only 1.15 indels per individual were identified, and this number decreased to 0.50 after comparison to the control pool of data. Moreover, no indel was found to segregate in our family dataset. In order to rule out the possibility that a putative causal indel could be found in one member of a family but not in the other members, we checked the consequences for the remaining indels after the comparison to the control pool data for all 20 individuals from the 9 non-BRCA1/2 families. We observed a total of 8 undescribed indel variants for the 20 individuals. Three of them were located in intronic regions, two were located in 3' UTR regions, and the remaining three indels were homozygous with a global depth = 1. Thus, no putative truncating indel was found in our individual dataset (Table S1).

Discussion

High-throughput sequencing

The classical candidate-gene approach turned out to be a low-efficiency tool with regard to cost and time when used for the identification of causal genes in genetic diseases, especially when there

Table 2. Index value parameters for coverage study.

Chr	Fam	Ind	Mean	St Dev	Upper	Lower
3	27	07S722	−0.13	0.75	1.12	−1.38
		07S723	−0.04	0.31	0.77	−0.86
		07S724	−0.07	0.35	0.77	−0.92
		07S725	−0.10	0.43	0.82	−1.03
	60	06-240	−0.01	0.40	0.89	−0.91
		69-652	0.00	0.59	1.10	−1.09
	531	I-1408	0.05	0.53	1.08	−0.97
		I-904	−0.31	1.83	2.02	−2.64
	713	07S635	−0.04	0.62	1.08	−1.16
		07S636	−0.04	0.45	0.92	−0.99
6	11	04-168	−0.09	0.35	0.76	−0.94
		96-265	−0.02	0.36	0.83	−0.88
	40	07S576	−0.05	0.73	1.18	−1.29
		07S581	0.00	0.41	0.91	−0.91
	929	I-1627	−0.02	0.36	0.83	−0.88
		I-3345	−0.09	0.36	0.77	−0.95
	990	I-1927	−0.08	0.80	1.22	−1.38
		I-1928	−0.03	0.63	1.10	−1.16
	1125	I-2033	−0.03	0.44	0.91	−0.97
		I-4347	−0.09	0.39	0.80	−0.98
		Global	−0.06	0.55	0.99	−1.11

In order to evaluate the quality of the coverage within the candidate coding regions, we calculated an index value (*Is*, see Material and Methods). Mean, standard deviation, and lower and upper thresholds for *Is* used in the coverage study are shown for each affected individual and for the entire set of individuals (**Global**).

doi:10.1371/journal.pone.0009976.t002

is a large number of candidate genes (dozens to hundreds). For that reason, we decided to explore the new possibilities that massive parallel sequencing brings for the analysis of hundred of genes in the same reaction. Moreover, in the present study we developed a solid analysis pipeline based on SOAP aligner for variant detection in families with a common genetic disease via high throughput sequencing data. Hybrid selection on tiling microarrays [16] was used for the enrichment of exonic sequences within two candidate regions for carrying a breast cancer susceptibility gene. We analysed the data from 20 affected individuals from 9 different non-BRCA1/2 families to perform a mutational screening of 128 known genes and, through an exhaustive filtering process, we obtained 8 variants that are currently under different genetic and functional studies (Table 6). From a technical point of view, we obtained an average of 5.14 million reads per individual, which allowed reaching a mean global depth of 33×. We observed that multiplexing of the samples by using 5-base barcodes did not affect the capture step or the sequencing process, and that it thus represents a valuable tool when the number of sequences required to confidently cover the target region is proportionally lower than the number of sequences obtained per lane.

We performed a CGH-like analysis obtaining the log-ratio between the normalised depth data from every affected individual and the control pool to confirm a homogenous distribution of the coverage along the candidate regions. Although the results showed that the distribution was very homogeneous, some differences in coverage were observed mainly in regions flanking the coding candidate regions (low coverage regions), where small and random differences in depth value may produce bigger differences in the

Table 3. Depth Score threshold optimization assay.

Chr	6																																
Family	990																																
DS_Individuals	0			10			20			30			40			50			60			70			80			90			100		
DS_Control Pool	0	14	50	0	14	50	0	14	50	0	14	50	0	14	50	0	14	50	0	14	50	0	14	50	0	14	50	0	14	50	0	14	50
Candidate SNPs before homology	9768	15791	15880	33	60	101	12	14	27	10	10	15	9	9	12	8	8	8	7	7	7	5	5	5	3	3	3	2	2	2	1	1	1
Candidate SNPs after homology	–	–	–	25	43	71	7	7	16	5	5	8	4	4	6	4	4	4	4	4	4	4	4	4	3	3	3	2	2	2	1	1	1
Confirmed by Sanger	–	–	–	–	–	–	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	4	3	3	3	2	2	2	1	1	1
FPR	–	–	–	–	–	–	43	43	75	20	20	50	0	0	33	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
FNR	–	–	–	–	–	–	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	25	25	25	50	50	50	75	75	75

The filtering results for various DS scores in a sample family (Family 990) are shown below. We observed that the threshold used for the control pool did not affect the total number of variants identified when using a DS threshold of 50 or higher for the cases. Taking into account the false positive rates (FPR) and false negative rates (FNR), and in order to be as conservative as possible, we finally chose a DS threshold of 50 for cases and a DS threshold of 14 for the control pool data.

doi:10.1371/journal.pone.0009976.t003

Table 4. Variant filtering results using MAQ software.

Chr	Family	Position (hg18)	Gen	Reference	Variant	MS[a]	Consequence	Sanger confirmation
3	27	170968060	TERC [MIM:602322]	G	C	255/39	NON_SYNONYMOUS_CODING	No
	531	162443032	NMD3 [MIM:611021]	A	G	16/55	NON_SYNONYMOUS_CODING	No
6	531	150094561	NUP43 [MIM:608141]	C	T	38/11	STOP_GAINED	No[b]
	713	149763387	SUMO4 [MIM:608829]	A	G	50/50	NON_SYNONYMOUS_CODING	No
	990	146761618	GRM1 [MIM:604473]	C	T	93/80	NON_SYNONYMOUS_CODING	Yes
		150205485	LRP11[c]	T	C	55/77	NON_SYNONYMOUS_CODING	Yes
		151202809	PLEKHG1[c]	G	A	63/46	NON_SYNONYMOUS_CODING	Yes[d]

A lack of correlation between MAQ score (MS) and Sanger sequencing confirmation was observed, since variants showing a MS =255 (maximum) were not confirmed whereas others with a MS =55 were validated.

[a] MAQ score for individual 1/individual 2.

[b] variant selected in a non-candidate chromosome for this family because of the truncating effect.

[c] No MIM reference [25].

[d] recently described in the Ensembl database.

doi:10.1371/journal.pone.0009976.t004

index value. These data supported that the capture and sequencing of the candidate regions were successfully fulfilled in the sample set.

Mutational screening

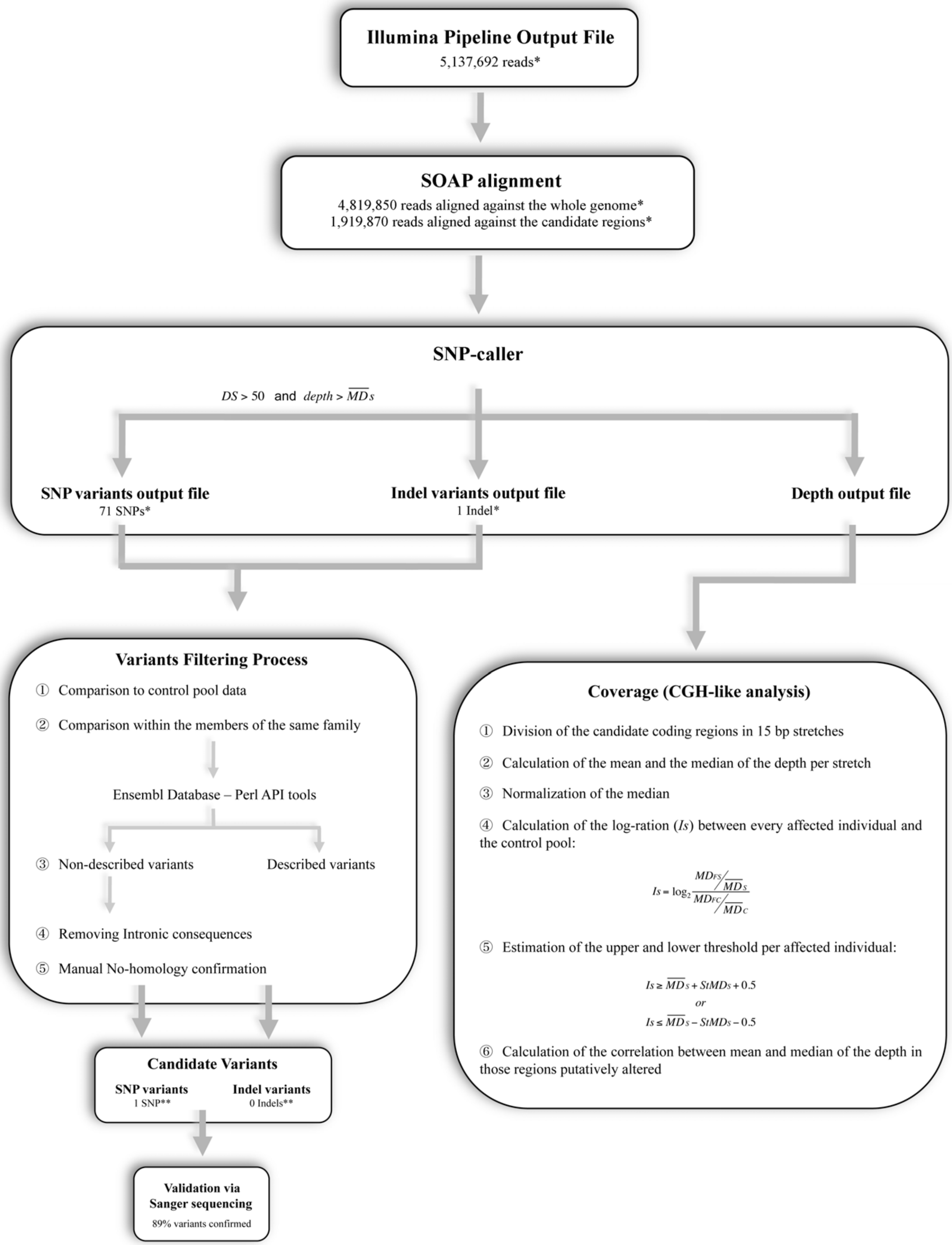
In order to identify de novo mutations, we consider that the best option is to maintain those sequences that match equally to more than one location, even though it could be a source of alignment errors due to homology between regions. For that reason, we tested two different aligners (SOAP and Mosaik), which allow the possibility of maintaining this kind of sequences. We finally decided to use SOAP v1.0 because its output file kept all the information of the input data, while being able to detect SNPs and short (gaps >1 bp) indel variants in single-end data.

The two high throughput sequencing-based mutation-detection studies published in the literature used MAQ software for SNP-calling [11,18]. In the first study, mutation detection was performed comparing data from tumour and skin tissues from a single leukaemia patient (AML [MIM 601626])[11]. The analysis pipeline showed a high false positive rate (87.4%) when trying to confirm final candidate SNPs. This high false positive rate suggested that more stringent conditions were necessary to filter the variants. For that reason, we designed two scores based on intrinsic variant features such as base quality (QS) and depth (DS), which were used as a first filter in our SOAP-based SNP-caller. Although more than 99% of the variants were discarded with this filter, the number of remaining variants suggested the need for further refinement. For that reason, we developed a comparison-based pipeline as previously described (see Material and Methods: Data analysis and Figure 1). Only 1 SNP (less than 1.5% of the variants detected in each individual, and around 6% of the variants shared by the same family) per family on average passed this filtering process, showing that, even after using restrictive thresholds, additional information has to be used to select the most probable variants (Table 5). Candidate SNPs were Sanger sequencing validated, and finally 8 of the 9 SNPs (89%) were confirmed. These variants had a minimum depth of 19× and a maximum depth of 155×, of which between 9 and 70 reads carried the variant. Quality scores were all close to 100, corroborating Solexa's good base-call quality in those positions. The 8 variants, which are currently under study (including functional characterization), are located in different genes (Table 6) and could be considered as the final candidate SNPs. The confirmation rate (0.89) and the lack of putatively causal mutations wrongly discarded in the filtering process (Table S2) suggested that our restrictive analysis pipeline successfully identified real previously undescribed variants.

In the second mutational screening study performed using high-throughput technology and MAQ software, the authors set up a global exome-capture method based on microarrays and a specific analysis pipeline, which was conceptually quite similar to ours, to identify causal variants for a monogenic disease [18]. The analysis pipeline was based on comparisons between data from affected individuals, HapMap individuals, and the dbSNP database, considering as candidates those functional variants that were not present in HapMap sample data nor in the dbSNP database. With a subset of our data, we performed a test using MAQ software and our analysis pipeline and observed no correlation between MAQ variant score and the confirmation rate (Table 4), evidencing a lack of accuracy in MAQ's algorithm.

Regarding indel variants, no indel fulfilled the criteria to be considered a candidate variant, neither in our dataset (Table S1) nor in the previously cited study. This could be due to the fact that indel discovery on single-end data is not as accurate as with the new paired-end technology, affecting the sensitivity of indel detection in both studies. Similarly, we cannot discard the possibility of missing the existence of large rearrangements due to the limitations of single-end data. Recently published studies are starting to demonstrate the efficiency of paired-end sequencing in the identification of genomic rearrangements [19,20].

In another study, a complete mutational screening using Sanger sequencing was performed on 718 genes located on chromosome X in probandi from a set of 208 families diagnosed with X-linked mental retardation [21]. The authors obtained 1858 coding variants, among which 1814 (980 missense, 22 nonsense, 13 splice-site, and 799 synonymous) were SNPs, 3 were double SNPs (missense), and 41 were indels. However, only 18 SNPs (17 nonsense and 1 missense, less than 1% of the initial SNPs) and 15 indels located in 26 different genes resulted to be strong candidates


Illumina Pipeline Output File
5,137,692 reads*
SOAP alignment
4,819,850 reads aligned against the whole genome*
1,919,870 reads aligned against the candidate regions*
SNP-caller
$DS > 50$ and $depth > \overline{MD}s$
SNP variants output file
71 SNPs*
Indel variants output file
1 Indel*
Depth output file
Variants Filtering Process
① Comparison to control pool data
② Comparison within the members of the same family
Ensembl Database – Perl API tools
③ Non-described variants
Described variants
④ Removing Intronic consequences
⑤ Manual No-homology confirmation
Candidate Variants
SNP variants
1 SNP**
Indel variants
0 Indels**
Validation via Sanger sequencing
89% variants confirmed
Coverage (CGH-like analysis)
① Division of the candidate coding regions in 15 bp stretches
② Calculation of the mean and the median of the depth per stretch
③ Normalization of the median
④ Calculation of the log-ration (*Is*) between every affected individual and the control pool:
$Is = \log_2 \frac{MD_{FS} / \overline{MD_S}}{MD_{FC} / \overline{MD_C}}$
⑤ Estimation of the upper and lower threshold per affected individual:
$Is \geq \overline{MD}s + StMDs + 0.5$
or
$Is \leq \overline{MD}s - StMDs - 0.5$
⑥ Calculation of the correlation between mean and median of the depth in those regions putatively altered

* per individual
** per family

Figure 1. Filtering process. Analysis pipeline used in the identification of the candidate variants. Left boxes correspond to processing of variants; right box corresponds to coverage analysis. See text for details.
doi:10.1371/journal.pone.0009976.g001

for being involved in X-linked mental retardation. Similarly, our results showed that around 1% of the initial SNP variants obtained via high throughput sequencing could be considered candidates (Table 6). Because no truncating mutations passed our filters, further functional studies are required to assess whether any of the confirmed variants is ultimately a causal mutation, specifically those variants considered of interest because of their functional implications (missense and 3'UTR) and gene function.

In summary, we designed an integral analysis pipeline for mutational screening via SOAP v1.0 that resulted in a low false positive rate with a low probability of discarding real positive variants, with which we identified 8 candidate variants that are currently under functional characterization. We consider that the present strategy might be a valid second step for identifying high penetrance genes, specifically when the regions of interest show significant evidence of linkage.

Materials and Methods

Ethics Statement

All patients provided written informed consent for the collection of samples and subsequent analysis. We obtained ethics approval for this study from the ethics committees at all institutions/hospitals where participants were recruited [17]. The GEO [22] accesion number (GSE20406) for this study has been approved as well as GEO accession numbers for each of the samples (Table 7).

Samples and candidate regions

As stated earlier, in a previous study we genotyped a total of 132 individuals from 41 non-BRCA1/2 families with almost 6,000 SNP markers, and we observed a linkage profile showing several candidate regions. Suggestive linkage signals (HLOD >2.2) were found in two regions located on chromosomes 3 and 6, which span 10.8 and 6.5 Mb, respectively, and we found 6 and 5 families putatively linked to each region [17].

In the present study, 10 of these 11 families were selected based on the availability of DNA (from at least 2 affected members per family for a total of 22 DNA samples collected) to perform a mutational screening via massive parallel sequencing. One family (Family 21) was excluded from the final analysis because the DNA library preparation of one of the members failed. However, putatively truncating mutations (e.g. new stop codons or modifications within essential splice sites, and indel variants) were analysed for the remaining individual of this family (see Table S1).

We also sequenced DNA of 4 healthy individuals from Spanish control population, which were pooled into a single control data file. This pooled-control design presented several advantages,

Table 5. Summary of the variant filtering process.

Chr	Family	Individual	SNPs	After control	Shared by family	Undescribed	Consequences	Exonic	Candidate SNPs (%)*
3	27	07S722	49	10	0	0	0	0	0 (0.00)
		07S723	39	2					
		07S724	45	18					
		07S725	25	4					
	60	06-240	47	18	15	7	6	3	1 (6.67)
		96-652	61	26					
	531	I-1408	38	15	5	2	1	1	0 (0.00)
		I-904	66	42					
	713	07S635	50	32	8	4	5	2	1 (12.50)
		07S636	46	24					
6	11	96_265	96	36	17	5	12	1	0 (0.00)
		04_168	81	35					
	40	07S581	93	40	26	8	32	7	3 (11.54)
		07S576	96	53					
	929	I-3345	81	28	13	3	10	3	0 (0.00)
		I-1627	75	34					
	990	I-1927	131	63	52	14	40	8	4 (7.69)
		I-1928	119	54					
	1125	I-4347	99	33	11	2	7	0	0 (0.00)
		I-2033	74	26					
		Average	71	30	16	5	13	3	1 (6.25)

The number of variants after each of the filtering steps is shown for the 9 non-BRCA1/2 families. The original SNPs were matched against the control pool as well as with the other member/s of the family. Previously undescribed variants were then selected and consequences obtained using PerlAPI tools. Intronic consequences were discarded and finally the variants were checked for homology.
*with respect to SNPs shared by family.
doi:10.1371/journal.pone.0009976.t005

Table 6. Final candidate SNPs.

Chr	Family	Position (hg18)	Gene	Reference	Variant	QS[a]	DS[b]	Consequence[c]	Alamuth prediction[d]	Gene function
3	60	161301596	AC026118.17[e]	A	T	91/91	56/57	NCG		Pseudogene
	713	**170284589**	**EVI1** [MIM:165215]	**A**	**G**	**95/94**	**128/100**	**3UTR**		**Hematopoietic proliferation protein, related to acute myeloid leukemia**
6	**40**	**152502855**	**SYNE1** [MIM:608441]	**C**	**T**	**105/98**	**90/92**	**NSYN**	**TOL**	**A spectrin repeat containing protein expressed in skeletal and smooth muscle, and in peripheral blood lymphocytes, that localizes to the nuclear membrane**
		151203125	**PLEKHG1**	**C**	**T**	**103/98**	**133/146**	**SYN S1186S**		**Unknown**
		151713613	**AKAP12** [MIM:604698]	**C**	**T**	**98/96**	**185/183**	**SYN P700P**		**Scaffold protein in signal transduction, is a cell growth-related protein**
	990	**146761618**	**GRM1**	**C**	**T**	**101/96**	**101/81**	**NSYN R584C**	**AFF**	**Metabotropic glutamate receptor**
		150087915	**NUP43**	**T**	**C**	**95/93**	**101/97**	**3UTR**		**Part of a nuclear pore complex, mediating bidirectional transport of macromolecules between cytoplasm and nucleus**
		150205485	**LRP11**	**T**	**C**	**101/101**	**74/95**	**NSYN I312V**	**NDB**	**Unknown**
		151564223	**AL451072.14[e]**	**G**	**A**	**93/98**	**119/116**	**NCG**		**Non-coding RNA**

Variants confirmed by direct sequencing are marked in **bold**.

a) Quality Score Ind1/Ind2, b) Depth Score Ind1/Ind2, c) NCG: Non-coding gene, SYN: Synonymous, 3UTR: 3′ UTR, NSYN: Non-synonymous.

d) NDB: Not in database, AFF: predicted to affect the protein, TOL: predicted to be tolerated.

e) No MIM reference.

doi:10.1371/journal.pone.0009976.t006

namely reduced repercussion of differential sample DNA degradation, increased sample heterozygosity, and a balancing effect over the number of reads per sample in the final control data, leading to increased data homogeneity.

Sequencing and exon-capture

We used the Illumina single-end technique to perform the sequencing process and an exon-capture protocol recently published [16], which was divided into two steps:

a) Library construction. Genomic DNA (2.5 µg) was fragmented by a triple sonication step. To optimise the sequencing process, we used 5-base barcodes (#1: 5′-GATCT-3′; #2: 5′-ATCGT-3′; #3: 5′-TGTCT3′; #4: 5′-GTGAT-3′) linked to the ligated oligos, in order to be able to pool 3 or 4 individuals per array and to discriminate the sequences obtained from each individual. After ligating the modified adaptors, 150–300 bp fragments were selected by agarose gel electrophoresis and purified. To obtain a suitable amount of product, multiple parallel PCR reactions were carried out per sample.

b) Exon capture. A total of 142,983 60-mer probes were designed to cover the 787 exonic regions and immobilized on high-density tiling arrays, in order to capture the coding sequences of the 128 genes located in both candidate regions. We used 14 arrays to perform the enrichment step for all the samples. A final amount of 20 µg of PCR product from 3–4 individuals was hybridised onto the array, with several blocking agents, during 65 hours (see Figure 1). After amplification of the eluted DNA, the enrichment was validated by quantitative PCR (qPCR) using the product from 4 different arrays.

Data analysis

a) Alignment. After the identification of the sequences belonging to each individual and before the alignment, we removed the first five bases of every sequence corresponding to each of the barcodes. We selected SOAP v1.0 to analyse our dataset using the following parameters: a seed of 12 positions, gaps allowed up to 12 bases, a maximum of 2 mismatches, and a maximum of 5 repeated regions (reporting every region on which the sequences matched equally), using the whole genome as

Table 7. GEO accession numbers of the raw data from each of the samples.

Family	Individual	Accesion Number
27	07S722	GSM511164
	07S723	GSM511165
	07S724	GSM511166
	07S725	GSM511167
60	06-240	GSM511168
	96-652	GSM511169
531	I-1408	GSM511170
	I-904	GSM511171
713	07S635	GSM511172
	07S636	GSM511173
11	96-265	GSM511175
	04-168	GSM511174
40	07S581	GSM511177
	07S576	GSM511176
929	I-3345	GSM511179
	I-1627	GSM511178
990	I-1927	GSM511180
	I-1928	GSM511181
1125	I-4347	GSM511183
	I-2033	GSM511182
Control pool		GSM511184

doi:10.1371/journal.pone.0009976.t007

reference sequence. One of the main advantages of SOAP v1.0 is its ability to identify short insertions and deletions (indels) in single-end data, while maintaining all input information in the output file. We developed a specific SNP caller written in the Perl programming language to extract the information contained in the SOAP output file. The whole analysis pipeline for re-sequencing data is shown in Figure 1.

b) Coverage. To evaluate the homogeneity of the coverage, we calculated the mean and the median of the depth for stretches of 15 bases along the candidate regions. For each sample, we obtained the index value (*Is*) for each 15-base fragment as:

$$Is = \log_2 \frac{MD_{FS}/\overline{MD}_S}{MD_{FC}/\overline{MD}_C},$$

where MD_{FS} and MD_{FC} are the median of the depth in a 15-base fragment for the sample and the control pool respectively, and $\overline{MD}s$ and $\overline{MD}c$ are the global median of the depth for the sample and the control pool respectively. To select supposedly altered regions, we calculated the upper and the lower thresholds as:

$$Is \geq \overline{MD}s + StMDs + 0.5$$
$$or$$
$$Is \leq \overline{MD}s - StMDs - 0.5$$

where $\overline{MD}s$ is the global median and *StMDs* the standard deviation of the depth for the sample. Finally, for these putatively altered regions, we calculated the correlation between the mean and the median of the depth to evaluate possible coverage gaps within them.

c) Scores. To further select candidate heterozygous variants, we calculated the mean quality (base-calling quality from Illumina Genome Analyser) and the allele depth (number of different sequences in which an allele appeared) for both the variant and the reference alleles, and also the global depth (number of different sequences that cover a single base). We calculated two different scores based on the same mathematical formula:

$$score = \left(\frac{X_V}{X_R}\right) \times 100,$$

where X_V represents the mean quality (for Quality Score calculations) or the allele depth (for Depth Score calculations) of the variant allele, and X_R represents in each case the same parameter but for the reference allele. Scores close to 100 indicate that both alleles are equally represented.

In order to determine the optimal DS threshold for the analysis, we performed a study using the information from a single family (Table 3). We used different DS score values for the affected individuals of Family 990 and for the control pool data. We observed that the number of final candidate variants depended mostly on the threshold used for the sample data, although the false positive rate increased when using a DS threshold of 50 for the control data. Taking into account the optimal False Positive/False Negative detection rate and being as conservative as possible, we finally selected a DS threshold of 50 for the samples and a DS threshold of 14 for the control pool data (Table 3).

d) Variants. In the pipeline analysis (Figure 1), firstly those variants with a global depth $< \overline{MD}s$ and a DS <50 were removed from the sample files as an integral function of the newly-developed SNP-caller. Homozygous variants were also discarded since we expected low-frequency heterozygous variants to be the causal variants. Similarly, variants presenting a DS <14 in the control file were discarded (see above for explanation). Only non-common variants from the previous step were selected for subsequent analysis. The next step was an intrafamilial comparison, with which variants putatively segregating in each family were obtained. Although the filtering process was performed for both SNP and indel variants, from this point onwards subsequent filters were applied to SNP variants only because no indel variants were found that fulfilled the previous conditions (Table S1). We developed a Perl tool to distinguish between described and previously undescribed variants, and also to obtain the consequences of every undescribed variant on the known transcripts, via the *Ensembl* database through the PerlAPI tools [23]. From this point onwards, intronic consequences where filtered out for each affected individual. In order to rule out possible false positives due to homology artefacts, each variant was manually checked for homology using BLAT search [24]. The following step was the confirmation of those variants that passed all filters previously mentioned via Sanger sequencing. As a final step, we used Alamut version 1.5 software to evaluate, *in silico*, how non-synonymous variants would affect the functionality of the respective candidate proteins (Figure 1).

In order to rule out the possibility that a truncating mutation was detected in one member of a family but not in the others, we analysed the truncating consequences (e.g. stop-gains and alterations in essential splice sites) of the remaining SNPs after comparison against the control pool data (Table S2).

Finally, we compared the sensitivity of our scores with a published reference, and for that we followed our analysis pipeline using MAQ software in a subset of families applying no threshold for MAQ variant score (Table 4).

Supporting Information

Figure S1 Correlations. Coverage along the candidate regions was very high (98% on average) and no correlation between coverage and the number of sequences obtained per individual was observed (A), although we observed a logarithmic trend when the number of sequences aligned to the candidate regions was used (B). On the other hand, a strong correlation between the number of sequences aligned to the candidate coding regions and the mean depth was observed in our dataset (C). Failures in the capture step were discarded since high correlations between the global mean and the global median of the depth per individual (D) and between the mean and the median of the depth in putatively altered 15-bp regions for all the individuals (E) were observed (see text for details).
Found at: doi:10.1371/journal.pone.0009976.s001 (0.13 MB DOCX)

Table S1 Indel variant filtering process.
Found at: doi:10.1371/journal.pone.0009976.s002 (0.05 MB DOCX)

Table S2 Putative truncating variants discarded during the filtering process. A. Summary of the whole list of truncating variants (STOP_GAINED and ESSENTIAL_SPLICE_SITE) in some individuals after comparison to the control pool data. The original list has been filtered (Depth Score > 50) for simplicity, given that variants showing a Depth Score below the threshold are likely to be false positives. As evidenced in the table, none of the variants above the threshold has a high Global Depth value, which paired to the fair Depth Score means that in every case the variant allele was detected in a low proportion in relation to the respective reference allele. Additionally, this table shows that no putative candidate truncating variants were discarded during the filtering process, reassuring that filtering using Depth Score and Global Depth is a stringent but adequate filtering step. In conclusion, the low Global Depth and Depth Scores explain why these variants are likely to be false positive results and they were therefore excluded from the final candidate SNP list (Table 6). B. The list of truncating variants for sample 05_980. The other member of this family failed in the library preparation step, but we still performed the analysis of the variants. The most likely variant (in red) was ruled out through Sanger sequencing.
Found at: doi:10.1371/journal.pone.0009976.s003 (0.15 MB DOCX)

Acknowledgments

We would like to thank to Ana Osorio and Fernando Fernandez from the Human Genetics Group in the CNIO, and Gordon Assaf from CSHL for their practical and technical support.

Author Contributions

Conceived and designed the experiments: JMS GJH JB. Performed the experiments: JMRR EH. Analyzed the data: JMRR FJGA EH ZX. Contributed reagents/materials/analysis tools: JMRR FJGA EH GP MR AB LB. Wrote the paper: JMRR FJGA JB.

References

1. Parkin DM, Bray F, Ferlay J, Pisani P (2005) Global cancer statistics, 2002. CA Cancer J Clin 55: 74–108.
2. Nathanson KL, Wooster R, Weber BL (2001) Breast cancer genetics: what we know and what we need. Nat Med 7: 552–556.
3. Diez O, Osorio A, Duran M, Martinez-Ferrandis JI, de la Hoya M, et al. (2003) Analysis of BRCA1 and BRCA2 genes in Spanish breast/ovarian cancer patients: a high proportion of mutations unique to Spain and evidence of founder effects. Hum Mutat 22: 301–312.
4. Bergman A, Abel F, Behboudi A, Yhr M, Mattsson J, et al. (2008) No germline mutations in supposed tumour suppressor genes SAFB1 and SAFB2 in familial breast cancer with linkage to 19p. BMC Med Genet 9: 108.
5. Oldenburg RA, Kroeze-Jansema KH, Houwing-Duistermaat JJ, Bayley JP, Dambrot C, et al. (2008) Genome-wide linkage scan in Dutch hereditary non-BRCA1/2 breast cancer families identifies 9q21-22 as a putative breast cancer susceptibility locus. Genes Chromosomes Cancer 47: 947–956.
6. Rosa-Rosa JM, Pita G, Gonzalez-Neira A, Milne RL, Fernandez V, et al. (2009) A 7 Mb region within 11q13 may contain a high penetrance gene for breast cancer. Breast Cancer Res Treat 118: 151–159.
7. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18: 1851–1858.
8. Li R, Li Y, Kristiansen K, Wang J (2008) SOAP: short oligonucleotide alignment program. Bioinformatics 24: 713–714.
9. Lin H, Zhang Z, Zhang MQ, Ma B, Li M (2008) ZOOM! Zillions of oligos mapped. Bioinformatics 24: 2431–2437.
10. Levy S, Sutton G, Ng PC, Feuk L, Halpern AL, et al. (2007) The diploid genome sequence of an individual human. PLoS Biol 5: e254.
11. Ley TJ, Mardis ER, Ding L, Fulton B, McLellan MD, et al. (2008) DNA sequencing of a cytogenetically normal acute myeloid leukaemia genome. Nature 456: 66–72.
12. Pushkarev D, Neff NF, Quake SR (2009) Single-molecule sequencing of an individual human genome. Nat Biotechnol.
13. Wang J, Wang W, Li R, Li Y, Tian G, et al. (2008) The diploid genome sequence of an Asian individual. Nature 456: 60–65.
14. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, et al. (2008) The complete genome of an individual by massively parallel DNA sequencing. Nature 452: 872–876.
15. Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, et al. (2007) Genome-wide in situ exon capture for selective resequencing. Nat Genet 39: 1522–1527.
16. Hodges E, Rooks M, Xuan Z, Bhattacharjee A, Benjamin Gordon D, et al. (2009) Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. Nat Protoc 4: 960–974.
17. Rosa-Rosa JM, Pita G, Urioste M, Llort G, Brunet J, et al. (2009) Genome-wide linkage scan reveals three putative breast-cancer-susceptibility loci. Am J Hum Genet 84: 115–122.
18. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, et al. (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461: 272–276.
19. Leary RJ, Kinde I, Diehl F, Schmidt J, Clouser C, et al. (2010) Development of personalized tumor biomarkers using massively parallel sequencing. Science Translational Medicine Vol 2.
20. Stephens PJ, McBride DJ, Lin ML, Varela I, Pleasance ED, et al. (2009) Complex landscapes of somatic rearrangement in human breast cancer genomes. Nature 462: 1005–1010.
21. Tarpey PS, Smith R, Pleasance E, Whibley A, Edkins S, et al. (2009) A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation. Nat Genet 41: 535–543.
22. http://www.ncbi.nlm.nih.gov/geo/.
23. http://www.ensembl.org/info/data/api.html.
24. http://genome.ucsc.edu/cgi-bin/hgBlat.
25. http://www.ncbi.nlm.nih.gov/Omim/.

Functional Identification of Optimized RNAi Triggers Using a Massively Parallel Sensor Assay

Christof Fellmann,[1,3,7] Johannes Zuber,[1,7,8] Katherine McJunkin,[1] Kenneth Chang,[1] Colin D. Malone,[1] Ross A. Dickins,[4] Qikai Xu,[5] Michael O. Hengartner,[3] Stephen J. Elledge,[5,6] Gregory J. Hannon,[1,2,*] and Scott W. Lowe[1,2,*]
[1]Cold Spring Harbor Laboratory
[2]Howard Hughes Medical Institute
1 Bungtown Road, Cold Spring Harbor, NY 11724, USA
[3]Institute of Molecular Life Sciences, University of Zurich, Winterthurerstrasse 190, CH-8057 Zurich, Switzerland
[4]The Walter and Eliza Hall Institute of Medical Research, 1G Royal Parade, Parkville, Melbourne, 3050 VIC, Australia
[5]Department of Genetics
[6]Howard Hughes Medical Institute
Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA
[7]These authors contributed equally to this work
[8]Present address: Research Institute of Molecular Pathology, Dr. Bohr-Gasse 7, A-1030 Vienna, Austria
*Correspondence: hannon@cshl.edu (G.J.H.), lowe@cshl.edu (S.W.L.)
DOI 10.1016/j.molcel.2011.02.008

SUMMARY

Short hairpin RNAs (shRNAs) provide powerful experimental tools by enabling stable and regulated gene silencing through programming of endogenous microRNA pathways. Since requirements for efficient shRNA biogenesis and target suppression are largely unknown, many predicted shRNAs fail to efficiently suppress their target. To overcome this barrier, we developed a "Sensor assay" that enables the biological identification of effective shRNAs at large scale. By constructing and evaluating 20,000 RNAi reporters covering every possible target site in nine mammalian transcripts, we show that our assay reliably identifies potent shRNAs that are surprisingly rare and predominantly missed by existing algorithms. Our unbiased analyses reveal that potent shRNAs share various predicted and previously unknown features associated with specific microRNA processing steps, and suggest a model for competitive strand selection. Together, our study establishes a powerful tool for large-scale identification of highly potent shRNAs and provides insights into sequence requirements of effective RNAi.

INTRODUCTION

RNA interference (RNAi) provides a programmable mechanism for targeted suppression of gene expression. Through a highly conserved pathway, the RNAi machinery recognizes and processes double-stranded RNAs into small RNAs that guide the repression of complementary genes (for review, see Bartel, 2004; Hannon, 2002). Experimental RNAi acts by providing exogenous sources of double-stranded RNA that mimic endogenous triggers and has paved the way for rapid loss-of-function studies that range from exploring the function of single genes to large-scale genetic screens. Moreover, RNAi is being developed into new therapies that can, in principle, inhibit any gene product.

In animals, somatic RNAi is mainly programmed by microRNAs (miRNAs), small noncoding RNAs that regulate gene expression (for review, see Bartel, 2004; Filipowicz et al., 2008). Most miRNAs are produced through a coordinated processing program whereby primary miRNA transcripts (pri-miRNAs) are cleaved by the nuclear Drosha/DGCR8 complex, resulting in the formation of precursor miRNAs (pre-miRNAs). These short hairpin-like molecules are actively exported to the cytoplasm, where Dicer excises mature small RNA duplexes that are incorporated into the RNA-induced silencing complex (RISC). Following strand selection, AGO2 discards the passenger (Leuschner et al., 2006; Matranga et al., 2005) and uses the guide for selection of complementary target mRNA substrates, whose expression is suppressed by accelerated mRNA degradation and/or translational inhibition.

Synthetic sources of double-stranded RNA can enter the RNAi pathway at various points. The most basic approach involves transfection of small interfering RNA (siRNA) duplexes (Elbashir et al., 2001) that resemble Dicer products. Although often potent, siRNA effects are transient and limited to transfectable cell types. An alternative approach relies on vectors that express stem-loop short hairpin RNAs (shRNAs), which resemble pre-miRNAs and enable stable and heritable gene silencing (Brummelkamp et al., 2002; Paddison et al., 2002). shRNAs can also be embedded in the context of endogenous miRNA transcripts—a configuration that creates a natural substrate for miRNA pathways (Silva et al., 2005; Zeng et al., 2002), enables stable and regulated expression from polymerase-II promoters (Dickins et al., 2005; Stegmeier et al., 2005), and reduces shRNA associated toxicity (Castanotto et al., 2007; McBride et al., 2008). Such miRNA-mimetics provide a versatile tool for long-term gene suppression in vitro and in vivo, as well as pool-based RNAi screening (see, for example, Dickins et al., 2007; Schlabach et al., 2008; Silva et al., 2008; Zender et al., 2008; Zuber et al., 2011).

While powerful, RNAi technology has some limitations. Besides suppressing the intended target gene, synthetic RNAi triggers can evoke off-target effects by suppressing unintended transcripts due to sequence homologies of either the sense or the antisense strand. Generally, the potential for misinterpreting such false positive results can be minimized through the use of several independent RNAi triggers targeting the same transcript. In addition, high intracellular levels of synthetic small RNAs can result in toxicities related to saturation of the RNAi machinery (Grimm et al., 2006). Such effects can be reduced by the use of miRNA-based RNAi triggers (Castanotto et al., 2007; McBride et al., 2008) and, in principle, would be eliminated through the use of shRNAs that effectively repress gene expression at low concentrations.

Beyond off-target effects, it remains difficult to identify potent shRNAs from among hundreds or thousands of possibilities within a given transcript. Consequently, many shRNAs are ineffective, leading to false-negative results in functional studies and screens. The precise sequence requirements of efficient RNAi remain incompletely understood, hampering the establishment of rational shRNA prediction rules. Studies using siRNA data sets indicate that RISC loading and target repression are dictated by sequence features in both the mature small RNA and the targeted mRNA region (Ameres et al., 2007; Khvorova et al., 2003; Schwarz et al., 2003). These include a preference for thermodynamic asymmetry (Khvorova et al., 2003; Schwarz et al., 2003), low G/C content (Reynolds et al., 2004), and a strong bias for A/U at the 5′ end of the guide strand (Tomari and Zamore, 2005). Nonetheless, these features are not sufficient to accurately distinguish between potent and weak RNAi triggers.

Machine-learning-based applications trained on siRNA data sets have produced algorithms that facilitate prediction of potent siRNAs (Huesken et al., 2005; Vert et al., 2006). However, such analyses have not been applied to shRNAs, which may require more stringent criteria as they rely on transcription and multistep miRNA processing for the production of small RNA duplexes. Indeed, experience indicates that siRNA algorithms are inefficient for predicting potent shRNAs, leaving their identification to laborious testing (Bassik et al., 2009; Li et al., 2007). Moreover, key RNAi applications such as pooled shRNA screening and RNAi transgenics require shRNAs that are effective even when expressed from a single genomic locus (“single copy”). Since most currently available shRNA reagents are not designed or tested to fulfill such stringent criteria, studies using shRNAs often rely on suboptimal reagents, and libraries contain many ineffective shRNAs that complicate the execution and interpretation of genetic screens.

Here we describe a high-throughput assay to evaluate shRNA potency in a massively parallel format. Our approach is based on a single-vector reporter assay that functionally monitors the interaction of shRNAs with their specific target sites, and thereby takes into account all aspects of shRNA biogenesis and target repression. This simple strategy reliably identifies rare potent shRNAs, most of which are not predicted using existing algorithms. By tracking the behavior of 20,000 shRNAs through all steps of miRNA biogenesis, we uncovered sequence preferences that contribute to potent and specific RNAi. Such information will advance the use of RNAi in functional studies and lays the groundwork for validated shRNA libraries.

RESULTS

Single-Vector Sensor Assay for Functional shRNA Evaluation

Synthetic RNAi triggers can be accurately evaluated in functional assays by placing their cognate target site (“Sensor”) in the 3′UTR of a reporter gene and quantifying its RNAi-mediated repression (Du et al., 2004; Kumar et al., 2003). In previous systems, the reporter construct and RNAi trigger were delivered separately and thus had to be assayed in a one-by-one format. We reasoned that physically linking shRNAs and their cognate target sites in a single vector would enable multiplexed analysis of shRNA-target pairs. Therefore, we constructed a reporter vector (pSENSOR; Figure 1A) harboring an shRNA expressed under the control of a Tet-responsive element (TRE^{tight}) (Gossen and Bujard, 1992; Sipo et al., 2006) and its cognate target sequence (Sensor) in the 3′UTR of a constitutively expressed fluorescent reporter (Venus) (Nagai et al., 2002). Since the adjacent context of target sites may affect RNAi potency (Ameres et al., 2007), we designed Sensors as 50 nt fragments of the endogenous mRNA, harboring the 22 nt target in the center (see Figure S1A available online). In reporter cells expressing the reverse Tet-transactivator (rtTA) (see the Experimental Procedures; Gossen et al., 1995), doxycycline (Dox) induces shRNA expression, which in turn represses the Venus reporter to an extent that reflects the potency of the shRNA (Figure 1A).

To determine the dynamic range of the assay, we constructed a set of pSENSOR vectors harboring 17 pre-existing shRNAs of different potency, which were re-evaluated by western blotting and classified into groups of strong, intermediate, and weak shRNAs (Figure 1B, Figures S1B–S1E). Following transduction into rtTA-reporter cells, we quantified changes in Venus expression after Dox treatment for all 17 shRNAs (Figures 1C and 1D and data not shown). Induction of strong shRNAs resulted in dramatic reduction of Venus fluorescence; conversely, intermediate and weak shRNAs induced only a moderate or slight reduction of Venus intensity, respectively. Overall, Venus repression reflected the efficacy of individual shRNAs in suppressing their endogenous target, indicating that the Sensor assay accurately quantifies shRNA potency.

Pooled Evaluation of shRNAs

Since each shRNA and its corresponding Sensor are delivered in a single vector, our assay is adaptable to a pooled format. In such a setting, pooled shRNA-Sensor constructs must be transduced at single copy to ensure that Venus fluorescence of each cell reports the activity of a single shRNA. Upon shRNA induction, cells displaying strong Venus repression can be isolated using fluorescence-activated cell sorting (FACS), and potent shRNAs subsequently identified through sequencing of proviral shRNA cassettes. To evaluate this approach, we transduced a pool of 17 pretested pSENSOR constructs into rtTA reporter cells and sorted equal fractions of low, medium, and high Venus-expressing cells in the absence and presence of Dox (Figure 1E and Table S1). Next, genomic DNA was isolated from

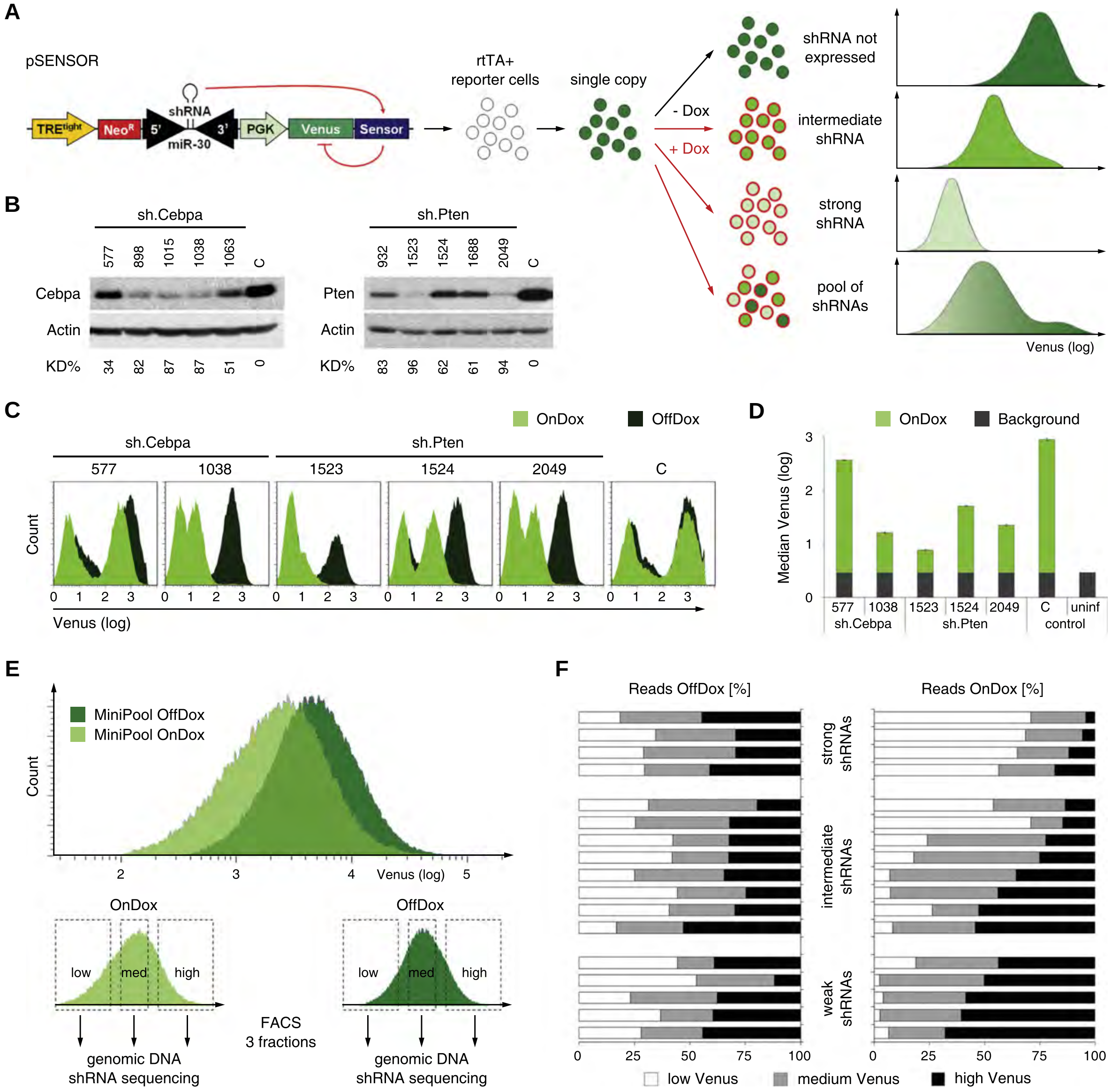


Figure 1. Sensor Assay for Assessment of shRNA Potency

(A) Schematic of the Sensor assay. The pSENSOR vector harbors a Tet-inducible shRNA and its cognate target sequence in the 3′UTR of a PGK-driven Venus reporter. Upon infection of cells expressing rtTA, Dox treatment induces shRNA expression. In turn, the extent of Venus knockdown directly reports shRNA potency. Histograms depict predicted fluorescence intensity distributions for shRNAs of different potencies.

(B) Immunoblotting of Cebpa and Pten in NIH 3T3s transduced with Cebpa or Pten shRNAs of different potencies. C, sh.Luci.1309 control shRNA. KD%, knockdown level relative to C and normalized to actin.

(C) Flow cytometry analysis of ERC reporter cells transduced with pSENSOR carrying indicated shRNA-Sensor cassettes and treated with or without Dox (On/OffDox). The leftmost peaks represent uninfected cells. C, control construct harboring an shRNA with a noncorresponding Sensor.

(D) Quantification of Venus fluorescence intensity of OnDox cells shown in (C) and uninfected reporter cells (uninf) used to define background levels. Error bars represent the standard deviation of triplicates.

(E) Flow cytometry plots of rtTA reporter cells transduced with a MiniPool of 17 shRNA-Sensor constructs and treated ± Dox for 7 days. The lower panels illustrate Venus-based cell sorting of each population into three equal subpopulations.

(F) Quantification of shRNA sequence reads within the sorted populations outlined in (E). For each shRNA, the distribution of reads among low/medium/high Venus fractions is plotted as a percentage of total reads of that shRNA. The shRNAs are clustered according to their preannotated groups (see Figure S1E for details).

sorted cells in each fraction, and the abundance of each shRNA was determined by capillary sequencing (288 reads for each fraction). In the absence of Dox, each shRNA was equally distributed among the three fractions (Figure 1F). Following Dox addition, potent shRNAs were enriched in the low Venus fraction and underrepresented in the high Venus fraction, while weak shRNAs

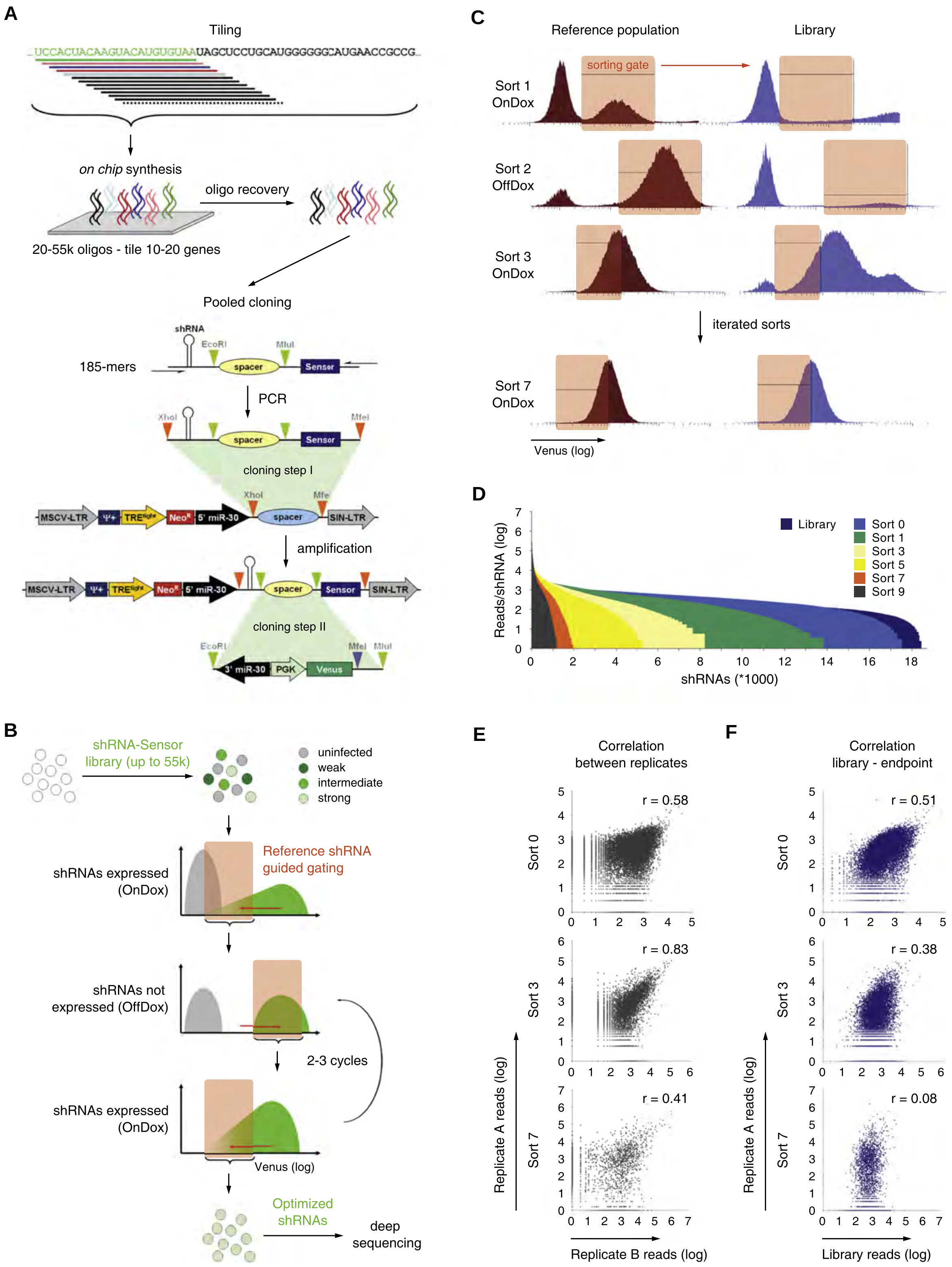


Figure 2. Sensor Ping-Pong Strategy for Deconvolution of Complex shRNA-Sensor Libraries

(A) Schematic of design and cloning of shRNA-Sensor libraries. A library of 20,000 constructs tiling every possible target site in nine mammalian transcripts was designed. 185-mer oligonucleotides containing shRNAs and cognate Sensors were synthesized and cloned into a 5′miR30 recipient vector. In a second step, the 3′miR30-PGK-Venus reporter cassette was cloned between shRNAs and their cognate Sensor to reconstitute complete pSENSOR vectors.

were shifted to the high Venus fraction and almost absent in the low Venus fraction. Thus, the Sensor assay can be used to select shRNAs based on their potency in a pooled format.

Optimization of the Sensor Assay

In pilot experiments, we observed that potent shRNA-Sensor constructs showed decreased viral titers, potentially reducing their representation in the population. We hypothesized that this was due to potent shRNAs targeting their Sensor on proviral transcripts, thereby inducing their degradation. To circumvent this, we transiently suppressed shRNA biogenesis in packaging cells by cotransfecting a potent DGCR8 siRNA. Indeed, this modification normalized the packaging and transduction efficiency of pSENSOR constructs (Figures S2A–S2C).

We also realized that effects of shRNAs on their endogenous target might alter the proliferation and/or viability of reporter cells and thereby bias the assay. For example, potent shRNAs targeting essential genes will deplete reporter cells and thereby escape identification in a pooled setting. Since RNAi utilizes an evolutionarily conserved machinery, we reasoned that an avian reporter cell line would provide an accurate system for evaluating mammalian shRNAs, where biases induced by effects on endogenous targets would be minimized due to divergence at the nucleic acid level. We therefore engineered DF-1 chicken embryonic fibroblasts (Himly et al., 1998) to express the ecotropic retroviral receptor and an improved reverse Tet-transactivator (rtTA3) (Das et al., 2004). When tested using different shRNA-Sensor constructs, ‘‘Eco-rtTA-chicken’’ (ERC) reporter cells accurately reported shRNAs of different potency (Figures S2D–S2F), indicating that shRNA processing is similar between ERC and mammalian cells (see Figure S5I for large-scale confirmation). Therefore, avian ERC cells are accurate reporters for the Sensor assay and less sensitive to the biological effects of mammalian shRNAs.

Generation of a High-Complexity Sensor Library

To evaluate the ability of the Sensor assay to simultaneously analyze the potency of thousands of shRNAs, we constructed and surveyed a library of ~20,000 shRNA-Sensor constructs comprising every possible shRNA for nine mammalian transcripts (Table S2). To ensure that individual shRNAs were cloned together with their specific Sensor, we applied large-scale on-chip oligonucleotide synthesis (Cleary et al., 2004) to produce ~20,000 185-mers, each harboring an shRNA and its target sequence separated by cloning sites, and used them to assemble the Sensor library in a pooled two-step procedure (Figure 2A). Serving as internal controls, all 17 previously characterized shRNAs were included at 15-fold representation to ensure their presence in the final pool. Deep sequencing of the library after cloning revealed that >99% of all designed shRNAs were present (Figure 2D).

Multiplexed Evaluation of shRNA Potency Using Sensor Ping-Pong Sorting

To evaluate shRNA potency in this complex library, we initially applied fractionated sorting paralleling our analysis of small pools (Figure 1E). However, at an increased complexity level this strategy failed to distinguish strong and weak control shRNAs (data not shown). Reasoning that iterative rounds of selection could be used to strongly enrich potent shRNAs and eliminate background noise, we developed a FACS strategy (Sensor Ping-Pong, Figure 2B) that involves sequential cycles of shRNA induction and withdrawal, each followed by sorting for reporter cells displaying Venus levels similar to potent shRNA-Sensor controls. In this approach, OnDox sorts for ‘‘Venus-low’’ reporter cells exclude cells harboring dysfunctional shRNAs (thus maintaining high Venus levels); conversely, OffDox steps for ‘‘Venus-high’’ reporters eliminate cells with constitutively defective reporters, e.g., due to positional effects of the retroviral integration. In each sort, FACS gating was guided by parallel analysis of two small reference pools containing five strong (Top5) and five weak (Bottom5) control shRNAs. By four cycles of enrichment (Sort 7), the OnDox FACS profile of the library became more uniform and resembled that of the Top5 reference population (Figure 2C, Figures S2G and S2H).

To monitor the representation of individual shRNAs throughout the procedure, genomic DNA was extracted after every sort, and shRNA guide strands were amplified and quantified by deep sequencing. While more than 98% of all cloned constructs were initially represented in infected ERC reporter cells, each sort led to a reduction of library complexity such that less than 2000 shRNAs remained after seven sorts (Figure 2D). Importantly, the shRNA composition of independent duplicates correlated throughout the procedure (Figure 2E), while their correlation to the initial population was progressively lost (Figure 2F). Therefore, the decrease in pool complexity that occurred throughout the procedure results from a nonrandom enrichment of specific shRNAs.

Next, we quantified the abundance of our 17 internal control shRNAs throughout the experiment. After the second cycle (Sort 3), strong shRNA controls already showed significant enrichment, and weak shRNAs were depleted (Figure 3A).

(B) Schematic of the Sensor Ping-Pong sorting strategy. Reporter cells infected with an shRNA-Sensor library at single copy are cultured sequentially in presence or absence of Dox. According to reference populations, sorting gates are drawn to include only cells harboring potent shRNAs (see Figures S2G and S2H for details). Through iterative rounds of shRNA induction and FACS-based selection, the initial library is reduced to a pool of functional shRNA-Sensor constructs that can be identified by deep sequencing.

(C) Representative flow cytometry histograms of Top5 reference and library populations at sorting steps 1, 2, 3, and 7. ERC reporter cells were infected with the Library, Top5, or Bottom5 pools; grown repeatedly for 6–7 days On- then OffDox; and sorted according to the indicated gates. Data are presented as Venus intensity histograms; actual sorts were done using Venus/FSC dot plots (see Figure S2H for details).

(D) Histogram of library complexity over sort cycles. Shown are normalized read numbers in one replicate for each shRNA represented within the pool after the indicated sorts.

(E) Correlation of reads per shRNA between two replicates after the indicated sorts. r, Pearson correlation coefficient.

(F) Correlation of reads per shRNA between the initial library pool and the pools after the indicated sorts. r, Pearson correlation coefficient.

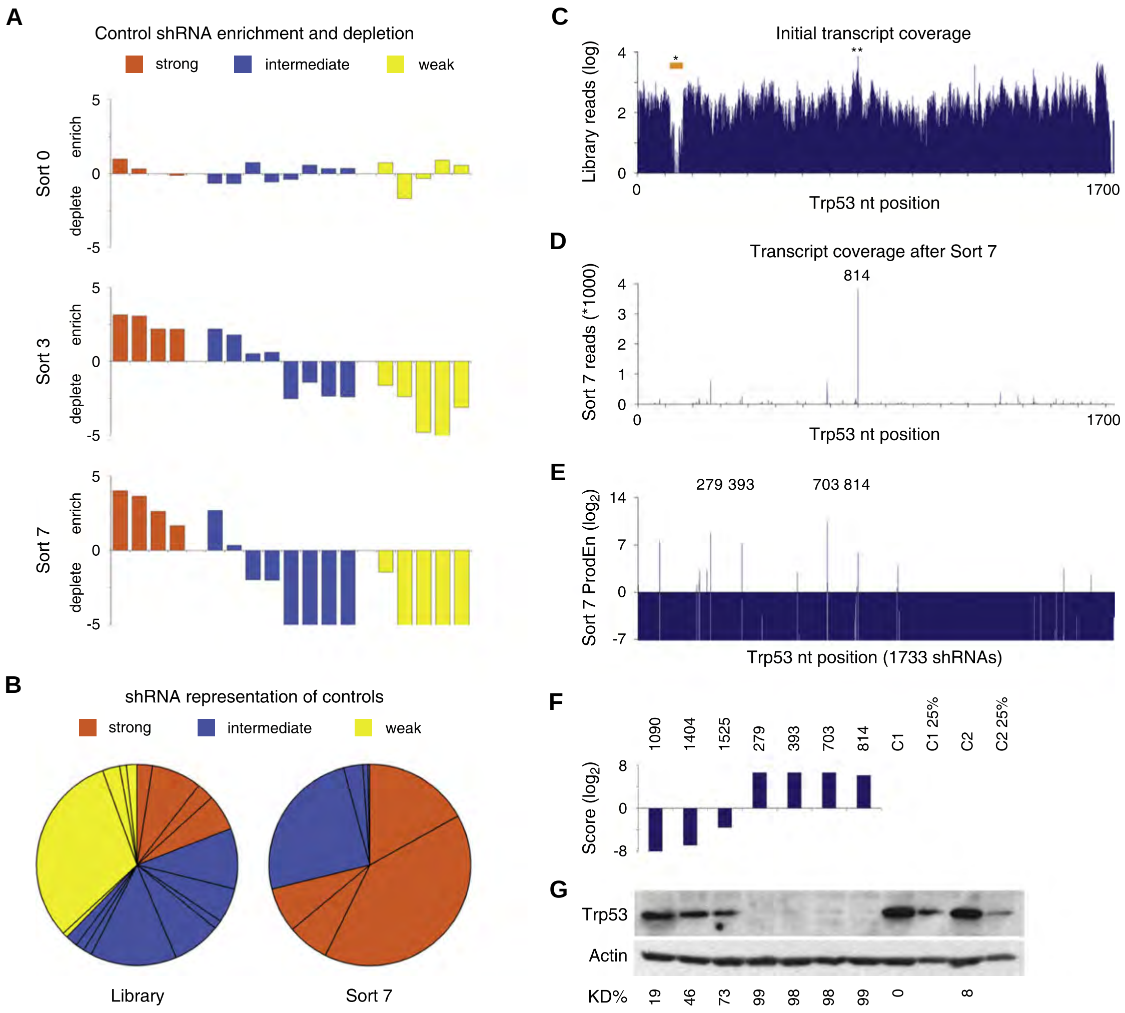


Figure 3. Assay Performance of Control shRNAs and shRNA-Sensor Constructs Tiling *Trp53*

(A) Enrichment or depletion of 17 control shRNA-Sensor constructs in transduced reporter cells before sorting (top), and after Sort 3 (middle) and Sort 7 (bottom). Values denote the $\log_2$ ratio of reads at the indicated stage compared to reads in the initial shRNA-Sensor plasmid library.

(B) Representation of control shRNA-Sensor constructs in the initial plasmid library and after seven sorts. Pie wedges represent mean values of technical (Library) or biological (Sort 7) duplicates.

(C) *Trp53* transcript coverage in the initial library. Shown are absolute reads (mean of technical duplicates) for 1733 Trp53 shRNA-Sensor constructs in the plasmid pool. Asterisk, XhoI restriction site affecting the cloning of 45 shRNAs. Double asterisk, most abundant shRNA, sh.p53.816.

(D) Read numbers (mean of biological duplicates) for Trp53 shRNA-Sensor constructs after four Ping-Pong cycles (Sort 7). 814, most abundant shRNA, sh.p53.814.

(E) Product enrichment scores (ProdEn, representing the product of the relative enrichment in each replicate at Sort 7 compared to the initial library) of all 1733 Trp53 shRNA-Sensor constructs. Numbers (279, 393, 703, 814) pinpoint highly enriched shRNAs.

(F) Integrated Score for selected shRNAs analyzed by western blotting.

(G) Western blot analysis of Trp53 levels in adriamycin-treated NIH 3T3s stably expressing the shRNAs indicated above from a single genomic integration. C1 and C2, control shRNAs (sh.Bcl2.906, sh.Bcl2.1132). C1 25% and C2 25%, 1:4 diluted control samples. KD%, knockdown level relative to C1 and normalized to actin.

By the fourth cycle (Sort 7), all strong shRNAs were robustly enriched, while all weak and most intermediate shRNAs were virtually eliminated (Figures 3A and 3B). The initial overrepresentation of some weak control shRNAs in the library did not prevent their eventual depletion (Figure 3B), suggesting the assay can tolerate imbalances in the initial pool composition. Optimal enrichment was obtained by four cycles, after which we did not observe any additional changes in the overall representation of our control shRNAs (data not shown). We also explored the use of bar-coded shRNA-Sensor libraries in conjunction with microarray-based monitoring of shRNA representation and found that this approach can stratify controls of known potency (Figures S3A and S3B). Collectively, the behavior of our 17 control shRNAs indicates that the Sensor Ping-Pong assay strongly enriches for potent shRNAs while robustly depleting nonfunctional and weak shRNAs.

Validation of Sensor-Identified shRNAs

Sequence analysis indicated that the Sensor assay can identify potent shRNAs from complex libraries de novo. For example, all 1733 possible Trp53 shRNAs were represented at the beginning of the assay, with the exception of 11 shRNAs containing a restriction site used for cloning and five shRNAs in the poly(A) tail (Figure 3C). Conversely, after four cycles most shRNAs were completely absent from the pool (Figure 3D), while only a few were enriched. Strikingly, the most prominent hit based on total reads was sh.p53.814 (formerly annotated as sh.p53.1224)—an shRNA that was previously identified empirically and shown to be extremely potent (Dickins et al., 2005).

To rank and select shRNAs for further validation, we developed two complementary scoring systems. The quantitative product enrichment (ProdEn), defined as the product of enrichment ratios in independent replicates, takes both the initial representation and consistency between replicates into account (Figure 3E). A second semiquantitative score uses a logistic function to integrate the initial representation of each shRNA, the consistency between replicates, and the trend for shRNA enrichment or depletion throughout all sorts (Figure 3F and Table S3). Based on these readouts, we examined the potency of four top-scoring and three nonscoring Trp53 shRNAs by immunoblotting (Figure 3G). All three newly identified Trp53 shRNAs showed similar potency to sh.p53.814, suppressing Trp53 expression to virtually undetectable levels, while the nonscoring shRNAs had no effect. These results validate the Sensor assay's ability to identify potent shRNAs and suggest that these RNAi triggers are very rare and equally distributed over a given transcript.

These observations were confirmed by Sensor results from other tiled transcripts. While the initial transcript coverage was nearly complete (98.1% overall), only a small number of shRNAs were enriched for each transcript after four Sensor Ping-Pong cycles (2.4% of all shRNAs had a score >10; Figures 4A, 4E, 5A, and 5E and Figure S4C). The vast majority of scoring shRNAs examined (85%–90%) showed strong knockdown of their target protein when expressed at single copy (Figures 4C, 4G, 5C, 5G; Figure S4E; and Table S4). Importantly, nonscoring shRNAs that were ineffective at single copy often showed substantial knockdown when transduced under conditions that lead to multiple proviral integrations (Figure 4D and data not shown). Hence, the Sensor assay accurately distinguishes between shRNAs that work at single versus high copy—the latter of which are useless in pool-based shRNA screens or other applications where only single integrations are achievable or desirable.

To functionally validate selected shRNAs, we developed a series of simple biological readouts for several of the genes. Generally, shRNAs that showed potent knockdown by immunoblotting displayed the most pronounced biological effects. For Mcl1, an antiapoptotic protein, we transduced NIH 3T3s at single or multiple copies with sh.Mcl1.1334 or a control shRNA and treated them with various concentrations of ABT-737 (Oltersdorf et al., 2005), an inhibitor of Bcl-2, Bcl-X_L, and Bcl-w that is known to synergize with Mcl1 inactivation to promote cell death (van Delft et al., 2006). As predicted, knockdown of Mcl1 sensitized NIH 3T3s to ABT-737 in a dose-dependent manner (Figure 4H).

For Rpa3 and Myc, proteins involved in DNA replication and cell proliferation, respectively, we examined shRNA potency using competitive proliferation assays. All five tested top-scoring shRNAs targeting mouse Myc rapidly depleted B cell lymphoma cells isolated from diseased Eμ-Myc; *p53*$^{-/-}$ transgenic mice (Figures 5A–5D). Similarly, the most potent human MYC shRNAs displayed deleterious effects in two human leukemia cell lines (Figures 5E–5H). Such potent shRNAs can be readily applied in Tet-regulated expression systems, where Dox titration can be used to generate hypomorphic states (Figure S4A). All strongly scoring Rpa3 shRNAs tested impaired proliferation of fibroblasts, while several randomly selected nonscoring shRNAs were neutral (Figures S4C–S4F). A few functional Rpa3 shRNAs that were previously identified empirically (Zuber et al., 2011) were not identified using the Sensor assay, suggesting it does not identify every potent shRNA. However, all other previously characterized functional and nonfunctional shRNAs reported correctly (data not shown).

Comparison to Existing Design Algorithms

To compare our results to existing siRNA-based design tools, we obtained the top 50 predictions for all nine transcripts using three different algorithms (Huesken et al., 2005; Sachidanandam, 2004; Vert et al., 2006) and compared them to the 50 highest scoring Sensor-derived shRNAs for each gene. Strikingly, >70% of our scoring shRNAs were not identified in the top 50 predictions of any algorithm (Figure S5A). While such false negatives, in principle, may have little practical significance, the majority of algorithm-predicted shRNAs did not score in the Sensor assay (Figure S5B), closely resembling their low validation rate in empirical testing (J.Z. and S.W.L., unpublished data). Together, these results demonstrate that siRNA algorithms are poor at predicting potent shRNAs (see also Bassik et al., 2009; Li et al., 2007) and underscore the value of the Sensor approach.

Global Analysis of shRNA Processing

We noticed that potent shRNAs identified through our unbiased functional assay share common sequence features. Top-scoring shRNAs (Score > 10; 453 shRNAs in total) are predominantly A/U rich (Figure 6A) and exhibit a strong thermodynamic asymmetry (Figure 6B)—two features that have been previously observed in studies of effective siRNAs (Khvorova et al., 2003; Reynolds et al., 2004; Schwarz et al., 2003). In contrast to nonscoring shRNAs and flanking mRNA regions, the nucleotide composition of potent guide strands shows many significant positional biases ($p < 0.01$, Pearson's χ^2 test with Šidák correction) that progressively emerge throughout the assay (Figure 6C and Figure S5C). Overwhelmingly, 88% of all top-scoring shRNAs carry U or A in guide position 1. Other A/U-rich positions include 2, 10, 13, and 14, while positions 20 and 21 are the only ones with a slight G/C bias. Position 20 also shows a remarkable depletion of A. Notably, most of these features have not been observed in siRNA-based studies.

To systematically analyze the interplay between nucleotide composition, shRNA processing, and biologic activity, we transduced the entire Sensor library into human HEK293T and chicken ERC cells. In the absence of cell sorting, we generated and quantified small RNA libraries designed to represent shRNA intermediates after major biogenesis steps (pri-, pre-, and mature miRNAs), and compared their abundance with our functional Sensor data from Ping-Pong sorted cells. At the pri-miRNA

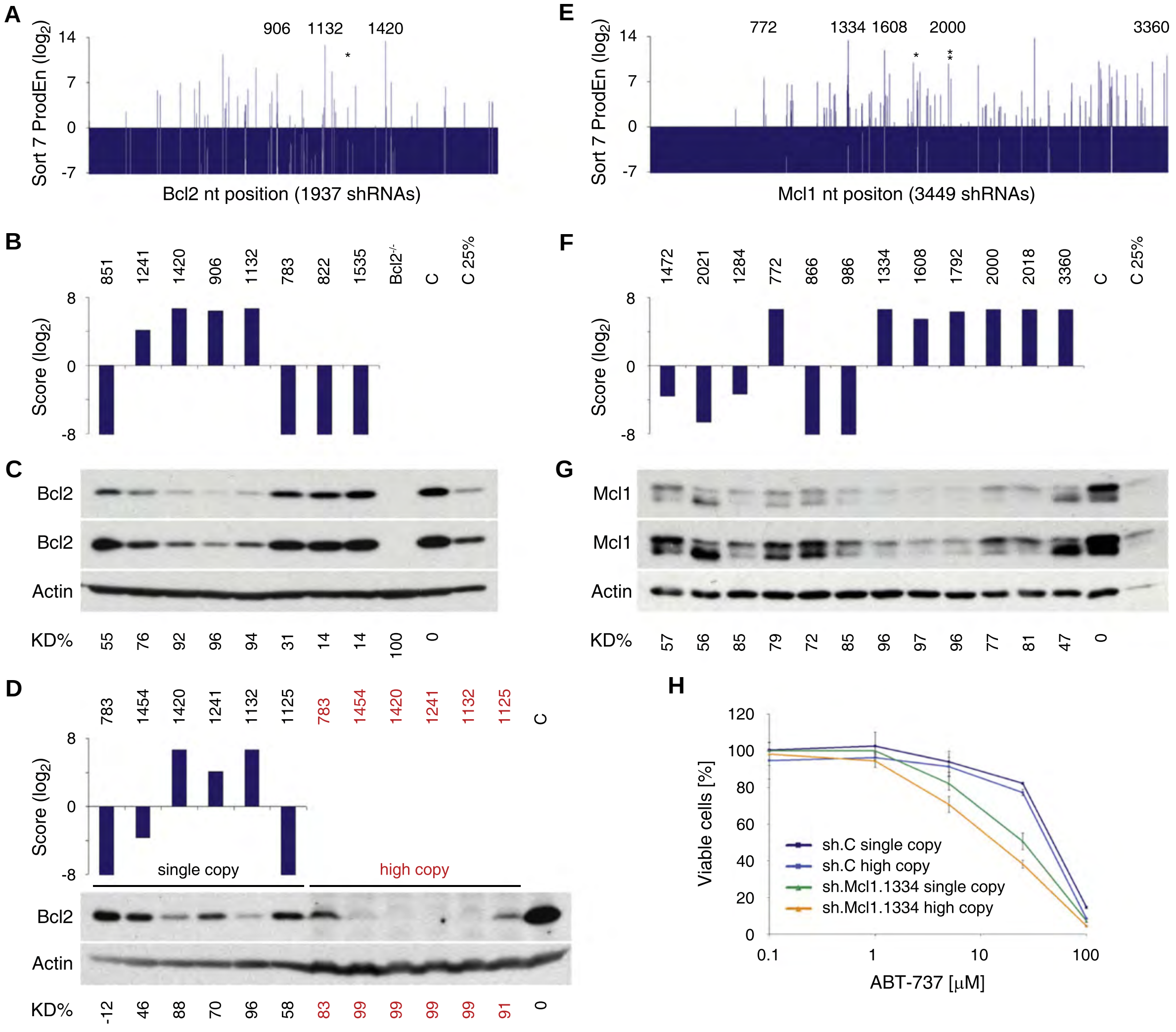


Figure 4. Analysis of Sensor-Identified shRNAs Targeting Bcl2 and Mcl1

(A) Product enrichment scores (ProdEn) of 1937 shRNA-Sensor constructs tiling the common region of both murine *Bcl2* transcripts. Asterisk, sh.Bcl2.1241. Numbers highlight enriched shRNAs that were analyzed by immunoblotting.

(B) Integrated score for selected Bcl2 shRNAs.

(C) Western blot analysis of Bcl2 levels in NIH 3T3s expressing the shRNAs indicated above from a single genomic integration. *Bcl2*$^{-/-}$ MEFs served as control. C, sh.Luci.1309. Two different exposures are shown. KD%, knockdown level relative to C and normalized to actin.

(D) Integrated score for selected Bcl2 shRNAs and western blot analysis of Bcl2 levels in NIH 3T3s expressing the indicated shRNAs from a single (single copy) or multiple (high copy, red) genomic integrations. C, uninfected cells.

(E) Product enrichment scores (ProdEn) of 3449 shRNA-Sensor constructs covering the mouse *Mcl1* transcript. Asterisk, sh.Mcl1.1792. Double asterisk, sh.Mcl1.2018. Numbers highlight enriched shRNAs that were analyzed by immunoblotting.

(F) Integrated score for selected Mcl1 shRNAs.

(G) Western blot analysis of Mcl1 expression in NIH 3T3s expressing the shRNAs indicated above from a single genomic integration. C, sh.Luci.1309. Two different exposures are shown.

(H) Synthetic lethal assay using the BH3-mimetic ABT-737 in combination with a potent Sensor-identified Mcl1 shRNA (sh.Mcl1.1334) or a control shRNA (C, sh.Luci.1309). NIH 3T3s expressing the indicated shRNA from a single or multiple genomic integrations were treated with ABT-737 for 48 hr and subsequently analyzed for viable cell numbers using flow cytometry (FSC/SSC and propidium iodide staining). DMSO (1%)-treated cells were used for normalization. Error bars represent the standard deviation of duplicate experiments.

level, >97% of all 18,720 shRNAs were identified and their abundances strongly correlated with those in the input library (r = 0.83 and 0.89 for ERC and HEK293T cells, respectively), indicating the absence of sequence biases in transduction and transcription. In both cell types, Drosha/DGCR8 cleavage occurred in >70% at the predicted site for most shRNAs (Figures S1A, S5D, and S5E). Miscleaved pre-miRNAs were associated with G/C richness and a particular bias for C at guide position 20 (Figure S5F, $p < 0.01$), suggesting that structural signals in pri-miRNAs guide processing to a specific site.

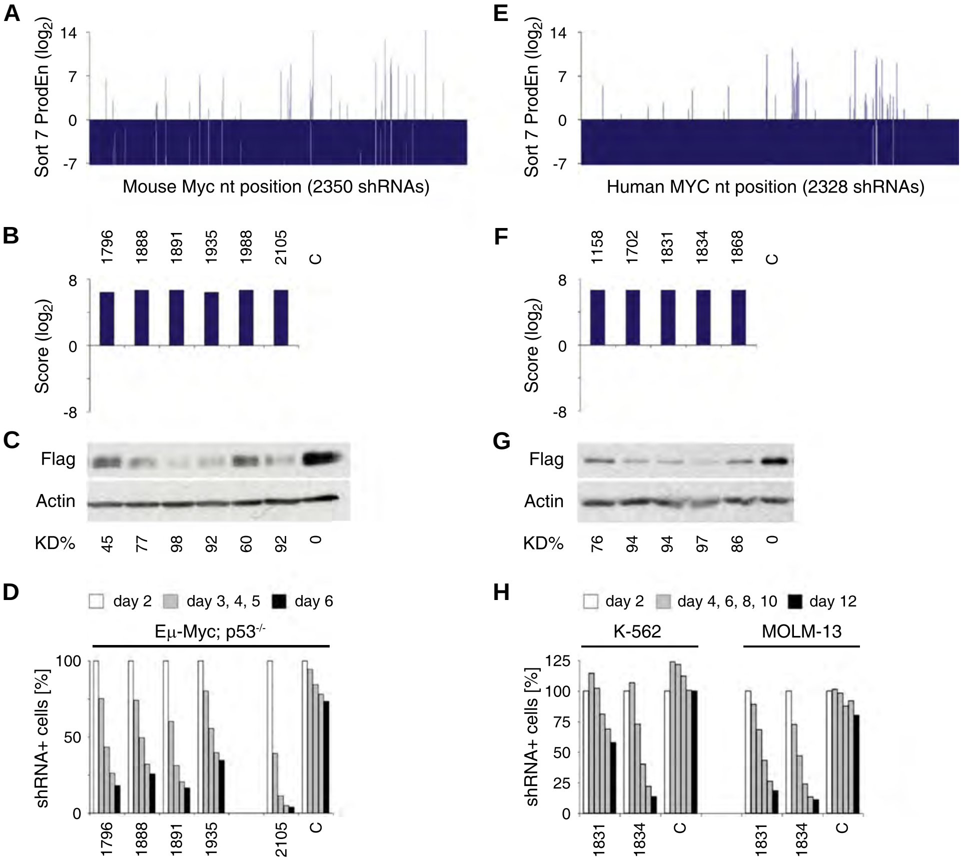


Figure 5. Analysis of Sensor-Identified shRNAs Targeting Mouse and Human MYC

(A) Product enrichment scores (ProdEn) of 2350 shRNA-Sensor constructs tiling the mouse *Myc* transcript.

(B) Integrated score for selected scoring Myc shRNAs.

(C) Western blot analysis of immortalized Rosa26-rtTA-M2 (RRT) MEFs expressing Flag-tagged murine Myc and shRNAs indicated above at single copy. Overexpression of Myc lacking the 3′UTR rescues knockdown by sh.Myc.1988 and 2105 (Figure S4B). C, sh.Luci.1309. KD%, knockdown level relative to C and normalized to actin.

(D) Competitive proliferation assay of Eμ-Myc; *p53*$^{-/-}$ lymphoma cells expressing the indicated shRNAs. The relative percentage of shRNA expressing cells at indicated days following retroviral transduction is shown. C, sh.Luci.1309.

(E) Product enrichment scores (ProdEn) of 2328 shRNA-Sensor constructs tiling the human *MYC* transcript.

(F) Integrated score for selected scoring MYC shRNAs.

(G) Flag-tag western blot analysis in RRT MEFs expressing Flag-tagged human MYC and shRNAs indicated above at single copy. C, sh.Luci.1309.

(H) Competitive proliferation assay of K-562 and MOLM-13 human leukemia cell lines expressing the indicated shRNAs. The relative percentage of shRNA-expressing cells at the indicated days following shRNA induction is shown. C, sh.Luci.1309.

To determine the stage in which dysfunctional shRNAs are eliminated from the biogenesis pathway, we calculated the dropout rate for each processing step. Our data reveal that a substantial fraction of shRNAs fail processing at each level (Figure S5G), while the representation of individual precursors remained highly correlated between ERC and HEK293T cells throughout miRNA biogenesis (Figure S5I). Together, this indicates that each processing step has restrictive and specific requirements. Notably, shRNAs that score in the Sensor assay are enriched at each processing step (Figure S5H), illustrating that efficient shRNA processing is a key determinant of potency.

To explore specific features associated with effective processing, we analyzed the nucleotide composition of shRNAs that were enriched at each step (Figure 6D). Efficient Drosha/DGCR8 cleavage was strongly associated with a prevalence of A/U at position 13/14 and G at position 20 and 21 ($p < 0.01$ for all). The transition from pre- to mature miRNAs, which represents Dicer/TRBP cleavage and likely AGO2 loading, shows biases for A/U in position 1 ($p < 0.01$), while the remaining guide is characterized by a flat profile with a slightly G-rich 3′ side (nt 10–22). To monitor features associated with the terminal pathway steps (AGO2 loading, target recognition, and cleavage) we analyzed shRNAs that showed an increase in their relative abundance from the mature miRNA stage to the endpoint of the Sensor assay (Sort 7). Only at this level, the structural pattern of enriched shRNAs exhibited a strong thermodynamic asymmetry (Figure 6D). Importantly, guide position 1 presented an extreme bias for U ($p < 0.01$)

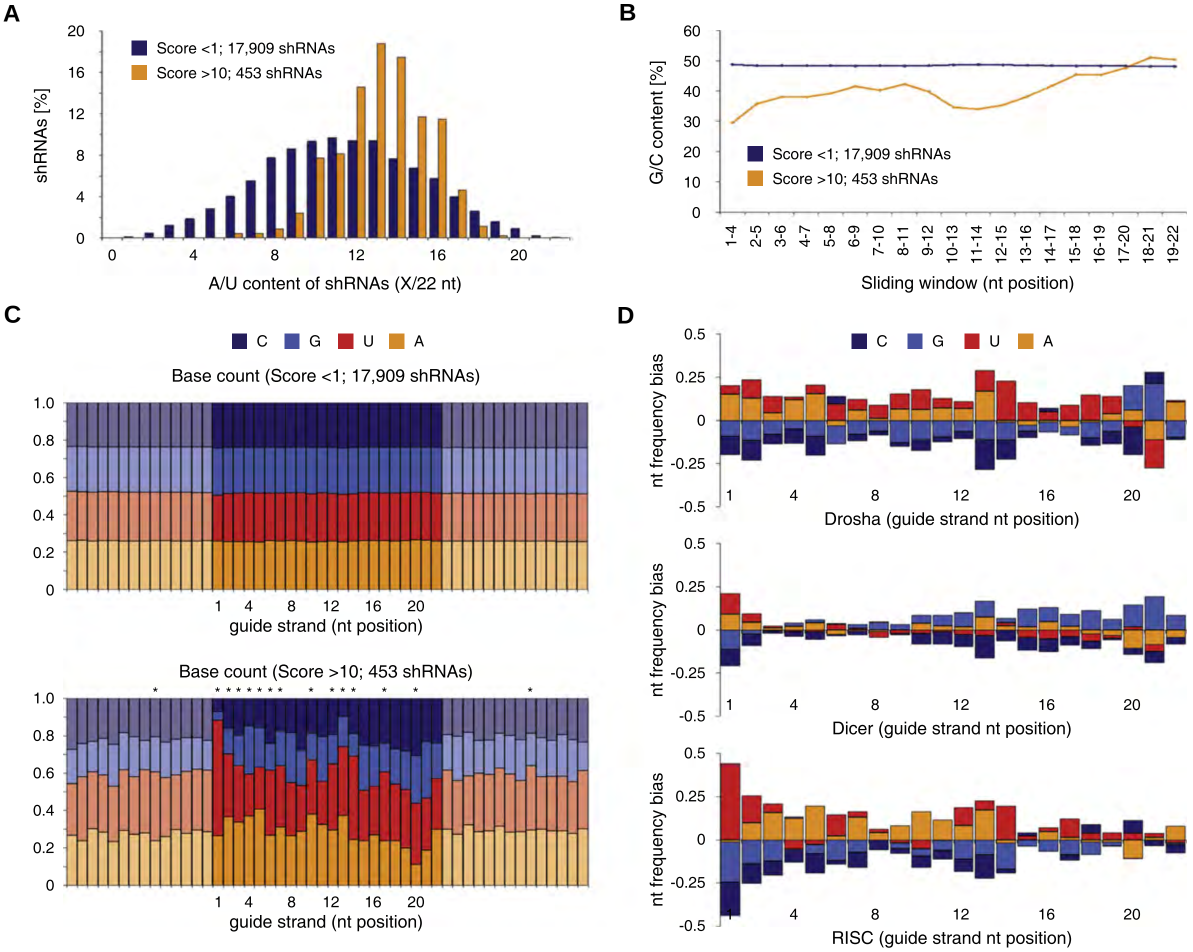


Figure 6. Sequence Features of Sensor-Identified shRNAs and Step-Specific RNAi Requirements

(A) Overall A/U content of nonscoring (Score < 1) and scoring (Score > 10) shRNAs, showing enrichment of relatively A/U-rich shRNAs.

(B) Local G/C content (4 nt sliding window) of nonscoring (Score < 1) and scoring (Score > 10) shRNAs, indicating thermodynamic asymmetry of scoring shRNAs.

(C) Nucleotide frequency in nonscoring (Score < 1, top) and scoring (Score > 10, bottom) shRNA-Sensor constructs. Shown are 22 nt shRNA guide strands (dark colors, reverse complement to 22 nt target site in endogenous transcript) and adjacent mRNA regions flanking the target site (pastel colors, reverse complement to mRNA). Asterisk, $p < 0.01$ (Pearson's χ^2 test with Šidák correction).

(D) Nucleotide bias of shRNAs that were significantly enriched (>5-fold) at the respective step and sufficiently represented (>100 reads) in the previous state. Drosha indicates sequences enriched from pri- to pre-miRNA (733 shRNAs). Dicer indicates sequences enriched from pre- to mature miRNA (931 shRNAs). RISC indicates sequences enriched from mature miRNA to shRNA representation at the genomic level after Sort 7 (root-mean-square value of all four replicates; 216 shRNAs). Data are shown for ERC cells; comparable patterns were observed in HEK293T cells.

and a near absence of G and C. These biases show a remarkable correlation to recently reported nucleotide binding affinities of the MID domain of human AGO2 (Frank et al., 2010) (Figure 7A), suggesting that a strong interaction between AGO2 and the 5′ end of the guide strand is a decisive prerequisite for potent RNAi.

In nucleotide profiles associated with mature miRNA production and function, we also noted an unusual rareness of A at position 20 (Figure 6D, $p < 0.05$). In line with the above, shRNAs harboring A in guide position 20 will yield passenger strands carrying U at their 5′ end such that the passengers may outcompete target-specific guide strands in RISC loading due to their affinity for AGO2 binding (Figure 7B). Indeed, shRNAs showing strong guide selection are biased for U in position 1 ($p < 0.01$) and against A in position 20 ($p < 0.01$), while the key features of shRNAs with passenger strand preference are an absence of A/U in position 1 ($p < 0.01$) and a strong bias for A/U in position 20 ($p < 0.01$, Figure 7C and Figure S6A). Notably, guide:passenger ratios for individual shRNAs were highly correlated between HEK293T and ERC cells (Figure S6B), indicating that preferences in strand selection are due to a conserved and specific process. Overall, potent shRNAs identified in our assay show extreme guide selection biases (39- and 95-fold in HEK293T and ERC cells, respectively, Figure 7D), illustrating that a strong preference for utilizing target-specific guide strands is a hallmark of effective RNAi.

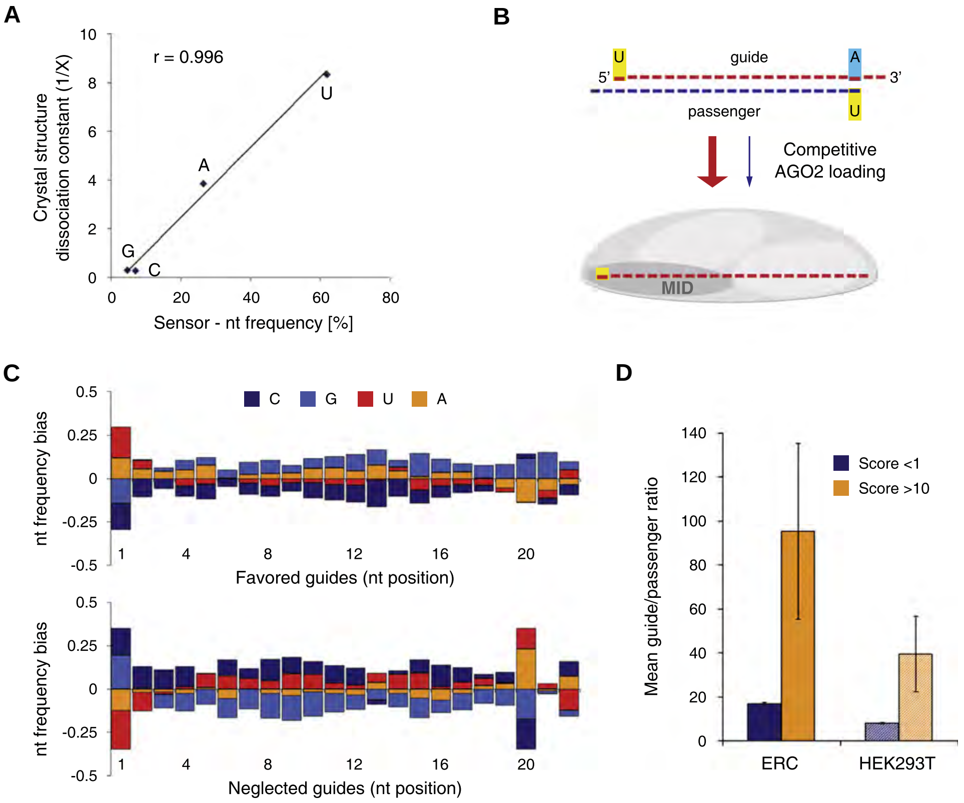


Figure 7. Potent shRNAs Show a Strong Strand Bias Dictated by Guide Positions 1 and 20

(A) Graphical analysis of the nucleotide bias at position 1 of the guide strand, demonstrating specific binding preferences for each nucleotide. The graph shows the correlation between dissociation constants obtained from data on crystal structures of the MID domain of human AGO2 (Frank et al., 2010) and Sensor-derived nucleotide frequencies. A linear regression indicates the trend of the data set. r, Pearson correlation coefficient.

(B) Model for AGO2-mediated competitive guide selection. Specific binding of the 5′ nucleotide to the MID domain of vertebrate AGO2 strongly influences strand selection, thereby defining the RISC-loaded guide strand (see Figure S6C for details).

(C) Nucleotide frequency bias of favored (guide/passenger > 50; 1546 shRNAs) and neglected (guide/passenger < 0.02; 439 shRNAs) guide strands in ERC cells transduced with the Sensor library. Comparable results were obtained with HEK293T cells.

(D) Mean guide versus passenger ratios for scoring (Score > 10) and nonscoring (Score < 1) shRNAs in ERC and HEK293T cells transduced with the Sensor library. Error bars represent the standard error of the mean.

DISCUSSION

Here we describe an unbiased, accurate, and scalable strategy for identifying highly potent shRNAs targeting any gene. Our approach measures the potency of shRNAs by monitoring their interaction with a surrogate target cloned into the 3′UTR of a fluorescent reporter, and thus integrates most aspects of shRNA biogenesis, target recognition, and repression. Combining on-chip synthesis of long oligonucleotides with a two-step cloning procedure, we generated a library of ~20,000 shRNA-Sensor constructs representing almost every target site (>99%) in nine mammalian transcripts. Using genetically distant avian reporter cells, we simultaneously evaluated the potency of every shRNA within this library via iterative cycles of FACS-based enrichment and deep-sequencing-based quantification, and thereby established a straightforward protocol for identifying potent shRNAs in a multiplexed format.

Our Sensor strategy accurately predicts the activity of shRNAs toward their endogenous targets and reliably identifies shRNAs that are effective when expressed from a single genomic integration—a criterion largely neglected in current shRNA libraries and prediction tools. As such, the assay vastly outperforms existing siRNA-based algorithms, which miss >70% of Sensor-derived shRNAs and generally necessitate the testing of many predictions to identify even a single potent shRNA. For example, despite previously testing ~15 top siRNA predictions from state-of-the-art algorithms, we found zero and only one potent shRNA targeting murine Mcl1 and Bcl2, respectively (data not shown). In contrast, the Sensor approach readily identified multiple highly effective shRNAs for both genes (Figures 4C and 4G).

Roughly 10%–15% of scoring shRNAs did not efficiently suppress their endogenous target. These false positives could arise from technical problems linked to our multistep protocol or off-target effects of the tested shRNA on the Venus transcript. Additionally, a subset of target sites could be occluded by long-range RNA interactions or protein binding events that are not reproduced on the abbreviated target site in our system.

Although we presently have no estimate of false-negative rates, the Sensor assay generally allowed us to easily find two or more potent RNAi triggers for every gene tested.

By surveying ~20,000 shRNAs produced in the absence of any design bias, our study describes a systematic analysis of shRNA efficiency and provides the largest data set of functionally annotated RNAi triggers currently available. Our data reveal that potent single-copy shRNAs are surprisingly rare, with frequencies ranging between 0.5% (Trp53) and 4.4% (Pcna) across the surveyed transcripts (2.4% on average). Except for sparing G/C-rich regions, potent shRNAs appear to be evenly distributed throughout transcripts, indicating that there is no preferential targeting of 3′UTRs.

To systematically explore the importance of efficient shRNA biogenesis for RNAi potency, we overlaid our functional data with a deep-sequencing-based analysis of small RNA species at different stages of miRNA maturation. Surprisingly, a substantial fraction of shRNAs failed to be processed at each step, while potent shRNAs were consistently well represented (Figures S5G and S5H). Highly processed shRNAs shared distinct sequence features that are attributable to specific steps in miRNA biogenesis and, mostly, have not been noted previously. For example, efficient pre-shRNA production is associated with A/U in position 13/14 and G in position 20/21, while C in position 20 impairs the accuracy of Drosha/DGCR8 cleavage. Together, our findings illustrate that the multistep process of miRNA biogenesis introduces additional structural constraints, providing an explanation for why siRNA-based algorithms often fail to predict functional shRNAs.

Other determinants of shRNA potency emerge at the end of the RNAi pathway. Strikingly, nucleotide frequencies of potent shRNAs at guide position 1 precisely mirror nucleotide binding affinities of AGO2 (Frank et al., 2010) and resemble Argonaute-loading preferences for 5′ U-containing strands in other organisms (Buhler et al., 2008). Together with biases at position 20, this suggests that the interaction between AGO2 and the 5′ end of both strands plays a decisive role in competitive strand selection (Figure S6C). In turn, most potent shRNAs are characterized by a strong preference for selecting the intended guide, suggesting that accurate strand selection is a key feature of effective RNAi.

Preferentially loaded strands also showed a subtle general bias for G but lacked thermodynamic asymmetry (Figure 7C), which previously has been implicated in RISC assembly (Schwarz et al., 2003). Since well-selected guide strands that potently suppress their target show thermodynamic asymmetry (Figure 6D and Figure S6D), this feature may become relevant only after strand selection, e.g., by facilitating target release after cleavage and enhancing RISC turnover (Haley and Zamore, 2004; Leuschner et al., 2006). Together, our data suggest that RISC loading is based on competitive binding of the 5′ nucleotides of both strands to AGO2, while thermodynamic asymmetry enhances the efficiency of later steps in the RNAi process.

Although the Sensor assay was designed to improve RNAi potency, its implementation may also impact RNAi specificity. First, our assay helps to control for sequence-specific off-target effects by enabling the identification of multiple potent shRNAs against any gene. Second, it will reduce passenger-mediated off-target effects by selecting potent shRNAs with a bias for incorporating the intended guide strand into RISC. Third, the identification of parameters guiding Drosha/DGCR8 processing will help to minimize off-target effects mediated by aberrant guide strands. Finally, by providing shRNAs with single-copy activity, our assay should further reduce off-target toxicities owing to saturation of the RNAi machinery. Indeed, we see that miR30-based shRNAs expressed from a single-copy promoter do not interfere with the processing of endogenous miRNAs (Premsrirut et al., 2011).

We believe that the Sensor assay provides a powerful and efficient method for identifying potent shRNAs. By taking an unbiased approach, our pilot study not only validated the Sensor assay but, unexpectedly, revealed insights into sequence requirements of miRNA biogenesis, strand selection, and efficient target knockdown. Indeed, features deduced from our analyses provide an shRNA-specific criteria framework for rational shRNA design (Table S5). Although these simple rules do not fully recapitulate the accuracy of the assay, they can be used to filter shRNAs prior to their Sensor-based evaluation and thereby dramatically increase the number of genes that can be surveyed in one Sensor experiment. As such, our approach lays out a practical workflow for the rapid generation of functionally validated shRNA libraries as well as the identification of potent RNAi triggers for biological studies and, eventually, RNAi therapeutics.

EXPERIMENTAL PROCEDURES

Vectors and Library Construction

The pSENSOR reporter vector, containing TRE^{tight}-Neo^{R}-miR30-PGK-Venus-Sensor, was assembled in the pQCXIX retroviral backbone (Clontech). We designed ~20,000 185-mer oligonucleotides (each containing a 101 nt miR30-shRNA fragment, an EcoRI/MluI cloning site, the cognate 50 nt Sensor cassette, and an 18 nt primer binding site), which were synthesized alongside controls on a 55,000 features oligonucleotide array (Agilent Technologies). The shRNA-Sensor library was constructed in a two-step procedure, involving cloning of PCR-amplified shRNA-Sensor fragments into a 5′miR30-pSENSOR recipient vector, and inserting the 3′miR30-PGK-Venus cassette between shRNA and Sensor cassette. shRNAs were named according to the position of the 3′ nucleotide of the guide strand on the tiled transcript.

Reporter Cell Lines

RRT MEFs were generated by immortalizing Rosa26-rtTA-M2 MEFs through transduction of lentiviral SV40 large T antigen and subsequent passaging. ERC reporter cells were derived from a single-cell clone of DF-1 chicken embryonic fibroblasts (Himly et al., 1998) transduced with MSCV-rtTA3-PGK-Puro and MSCV-EcoReceptor-PGK-Hygro retroviruses, and grown in DMEM supplemented with 10% FBS, 1 mM sodium pyruvate, 100 U/ml penicillin, and 100 μg/ml streptomycin. Tet-regulatable shRNAs were induced using Dox concentrations of 1.0–2.0 μg/ml in RRT MEFs and 0.5 μg/ml in ERC cells.

Sensor Ping-Pong Assay

FACS procedures were carried out on a FACSAria II (BD Biosciences). ERC reporter cells were infected with pSENSOR libraries at singly copy and sorted in iterative cycles, either after treatment with Dox and G418 (500 μg/ml) for 6–7 days (OnDox) or after Dox and G418 withdrawal for 6–7 days (OffDox). The gating was guided by reference cells transduced with small pools of potent (Top5) and weak (Bottom5) control shRNA-Sensor constructs. In all sorts, a representation of 1000-fold the pool complexity was maintained. Deep sequencing template libraries were generated by PCR amplification of shRNA guide strands from genomic DNA of at least 10 million cells, using primers that tag standard Illumina adapters to the product, and sequenced using a primer reading reverse into the guide strand. Only sequences completely matching the Sensor library were retained.

Small RNA Libraries

Libraries were generated as previously described (Malone et al., 2009). In brief, total RNA from HEK293T or ERC cells transduced with the pSENSOR library was extracted with TRIzol (Invitrogen) and two phenol:chloroform:IAA (Ambion) purification steps. Of total RNA, 40 µg was run on a 12% denaturing polyacrylamide gel and 18–26 nt mature small RNAs or 50–70 nt pre-miRNAs were selected for cloning; pri-miRNA libraries were obtained by direct amplification from total RNA using miR30-specific primers. Following Illumina sequencing, only sequences completely matching the Sensor library were retained for further analysis.

SUPPLEMENTAL INFORMATION

Supplemental Information includes six figures, six tables, Supplemental Experimental Procedures, and Supplemental References and can be found with this article at doi:10.1016/j.molcel.2011.02.008.

ACKNOWLEDGMENTS

We thank A. Rappaport, C. Miething, D. Burgess, M. Spector, A. Vaseva, A. Chicas, L. Dow, M. Saborowski, S. Nuñez, and H. Varmus for providing reagents. We gratefully acknowledge K. Marran, B. Ma, S. He, S. Muller, P. Moody, and T. Spencer for excellent technical assistance. We also thank E. Hodges, M. Rooks, and W.R. McCombie and his team for help with deep sequencing, as well as Y. Erlich, A. Gordon, D. Lewis, M. Hammell, and V. Thapar for bioinformatics support. S.J.E., G.J.H., and S.W.L. are founders of Mirimus and serve on its scientific advisory board. This work was supported by program project grants from the National Cancer Institute and generous gifts from the Don Monti Memorial Research Foundation. J.Z. was supported by the Andrew Seligson Memorial Fellowship, and K.M. was supported by the Robert and Teresa Lindsay Fellowship of the Watson School of Biological Sciences. S.J.E., G.J.H., and S.W.L. are Howard Hughes Medical Institute investigators.

Received: June 10, 2010
Revised: November 29, 2010
Accepted: February 2, 2011
Published online: February 24, 2011

REFERENCES

Ameres, S.L., Martinez, J., and Schroeder, R. (2007). Molecular basis for target RNA recognition and cleavage by human RISC. Cell *130*, 101–112.

Bartel, D.P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. Cell *116*, 281–297.

Bassik, M.C., Lebbink, R.J., Churchman, L.S., Ingolia, N.T., Patena, W., LeProust, E.M., Schuldiner, M., Weissman, J.S., and McManus, M.T. (2009). Rapid creation and quantitative monitoring of high coverage shRNA libraries. Nat. Methods *6*, 443–445.

Brummelkamp, T.R., Bernards, R., and Agami, R. (2002). A system for stable expression of short interfering RNAs in mammalian cells. Science *296*, 550–553.

Buhler, M., Spies, N., Bartel, D.P., and Moazed, D. (2008). TRAMP-mediated RNA surveillance prevents spurious entry of RNAs into the Schizosaccharomyces pombe siRNA pathway. Nat. Struct. Mol. Biol. *15*, 1015–1023.

Castanotto, D., Sakurai, K., Lingeman, R., Li, H., Shively, L., Aagaard, L., Soifer, H., Gatignol, A., Riggs, A., and Rossi, J.J. (2007). Combinatorial delivery of small interfering RNAs reduces RNAi efficacy by selective incorporation into RISC. Nucleic Acids Res. *35*, 5154–5164.

Cleary, M.A., Kilian, K., Wang, Y., Bradshaw, J., Cavet, G., Ge, W., Kulkarni, A., Paddison, P.J., Chang, K., Sheth, N., et al. (2004). Production of complex nucleic acid libraries using highly parallel in situ oligonucleotide synthesis. Nat. Methods *1*, 241–248.

Das, A.T., Zhou, X., Vink, M., Klaver, B., Verhoef, K., Marzio, G., and Berkhout, B. (2004). Viral evolution as a tool to improve the tetracycline-regulated gene expression system. J. Biol. Chem. *279*, 18776–18782.

Dickins, R.A., Hemann, M.T., Zilfou, J.T., Simpson, D.R., Ibarra, I., Hannon, G.J., and Lowe, S.W. (2005). Probing tumor phenotypes using stable and regulated synthetic microRNA precursors. Nat. Genet. *37*, 1289–1295.

Dickins, R.A., McJunkin, K., Hernando, E., Premsrirut, P.K., Krizhanovsky, V., Burgess, D.J., Kim, S.Y., Cordon-Cardo, C., Zender, L., Hannon, G.J., et al. (2007). Tissue-specific and reversible RNA interference in transgenic mice. Nat. Genet. *39*, 914–921.

Du, Q., Thonberg, H., Zhang, H.Y., Wahlestedt, C., and Liang, Z. (2004). Validating siRNA using a reporter made from synthetic DNA oligonucleotides. Biochem. Biophys. Res. Commun. *325*, 243–249.

Elbashir, S.M., Harborth, J., Lendeckel, W., Yalcin, A., Weber, K., and Tuschl, T. (2001). Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. Nature *411*, 494–498.

Filipowicz, W., Bhattacharyya, S.N., and Sonenberg, N. (2008). Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? Nat. Rev. Genet. *9*, 102–114.

Frank, F., Sonenberg, N., and Nagar, B. (2010). Structural basis for 5′-nucleotide base-specific recognition of guide RNA by human AGO2. Nature *465*, 502–506.

Gossen, M., and Bujard, H. (1992). Tight control of gene expression in mammalian cells by tetracycline-responsive promoters. Proc. Natl. Acad. Sci. USA *89*, 5547–5551.

Gossen, M., Freundlieb, S., Bender, G., Müller, G., Hillen, W., and Bujard, H. (1995). Transcriptional activation by tetracyclines in mammalian cells. Science *268*, 1766–1769.

Grimm, D., Streetz, K.L., Jopling, C.L., Storm, T.A., Pandey, K., Davis, C.R., Marion, P., Salazar, F., and Kay, M.A. (2006). Fatality in mice due to oversaturation of cellular microRNA/short hairpin RNA pathways. Nature *441*, 537–541.

Haley, B., and Zamore, P.D. (2004). Kinetic analysis of the RNAi enzyme complex. Nat. Struct. Mol. Biol. *11*, 599–606.

Hannon, G.J. (2002). RNA interference. Nature *418*, 244–251.

Himly, M., Foster, D.N., Bottoli, I., Iacovoni, J.S., and Vogt, P.K. (1998). The DF-1 chicken fibroblast cell line: transformation induced by diverse oncogenes and cell death resulting from infection by avian leukosis viruses. Virology *248*, 295–304.

Huesken, D., Lange, J., Mickanin, C., Weiler, J., Asselbergs, F., Warner, J., Meloon, B., Engel, S., Rosenberg, A., Cohen, D., et al. (2005). Design of a genome-wide siRNA library using an artificial neural network. Nat. Biotechnol. *23*, 995–1001.

Khvorova, A., Reynolds, A., and Jayasena, S.D. (2003). Functional siRNAs and miRNAs exhibit strand bias. Cell *115*, 209–216.

Kumar, R., Conklin, D.S., and Mittal, V. (2003). High-throughput selection of effective RNAi probes for gene silencing. Genome Res. *13*, 2333–2340.

Leuschner, P.J., Ameres, S.L., Kueng, S., and Martinez, J. (2006). Cleavage of the siRNA passenger strand during RISC assembly in human cells. EMBO Rep. *7*, 314–320.

Li, L., Lin, X., Khvorova, A., Fesik, S.W., and Shen, Y. (2007). Defining the optimal parameters for hairpin-based knockdown constructs. RNA *13*, 1765–1774.

Malone, C.D., Brennecke, J., Dus, M., Stark, A., McCombie, W.R., Sachidanandam, R., and Hannon, G.J. (2009). Specialized piRNA pathways act in germline and somatic tissues of the Drosophila ovary. Cell *137*, 522–535.

Matranga, C., Tomari, Y., Shin, C., Bartel, D.P., and Zamore, P.D. (2005). Passenger-strand cleavage facilitates assembly of siRNA into Ago2-containing RNAi enzyme complexes. Cell *123*, 607–620.

McBride, J.L., Boudreau, R.L., Harper, S.Q., Staber, P.D., Monteys, A.M., Martins, I., Gilmore, B.L., Burstein, H., Peluso, R.W., Polisky, B., et al. (2008). Artificial miRNAs mitigate shRNA-mediated toxicity in the brain: implications for the therapeutic development of RNAi. Proc. Natl. Acad. Sci. USA *105*, 5868–5873.

Nagai, T., Ibata, K., Park, E.S., Kubota, M., Mikoshiba, K., and Miyawaki, A. (2002). A variant of yellow fluorescent protein with fast and efficient maturation for cell-biological applications. Nat. Biotechnol. *20*, 87–90.

Oltersdorf, T., Elmore, S.W., Shoemaker, A.R., Armstrong, R.C., Augeri, D.J., Belli, B.A., Bruncko, M., Deckwerth, T.L., Dinges, J., Hajduk, P.J., et al. (2005). An inhibitor of Bcl-2 family proteins induces regression of solid tumours. Nature *435*, 677–681.

Paddison, P.J., Caudy, A.A., Bernstein, E., Hannon, G.J., and Conklin, D.S. (2002). Short hairpin RNAs (shRNAs) induce sequence-specific silencing in mammalian cells. Genes Dev. *16*, 948–958.

Premsrirut, P.K., Dow, L.E., Kim, S.Y., Camiolo, M., Malone, C.D., Miething, C., Scuoppo, C., Zuber, J., Dickins, R.A., Kogan, S.C., et al. (2011). A rapid and scalable system for studying gene function in mice using conditional RNA interference. Cell *145*, in press.

Reynolds, A., Leake, D., Boese, Q., Scaringe, S., Marshall, W.S., and Khvorova, A. (2004). Rational siRNA design for RNA interference. Nat. Biotechnol. *22*, 326–330.

Sachidanandam, R. (2004). RNAi: design and analysis. Curr. Protoc. Bioinformatics, Chapter 12, Unit 12.3.

Schlabach, M.R., Luo, J., Solimini, N.L., Hu, G., Xu, Q., Li, M.Z., Zhao, Z., Smogorzewska, A., Sowa, M.E., Ang, X.L., et al. (2008). Cancer proliferation gene discovery through functional genomics. Science *319*, 620–624.

Schwarz, D.S., Hutvagner, G., Du, T., Xu, Z., Aronin, N., and Zamore, P.D. (2003). Asymmetry in the assembly of the RNAi enzyme complex. Cell *115*, 199–208.

Silva, J.M., Li, M.Z., Chang, K., Ge, W., Golding, M.C., Rickles, R.J., Siolas, D., Hu, G., Paddison, P.J., Schlabach, M.R., et al. (2005). Second-generation shRNA libraries covering the mouse and human genome. Nat. Genet. *37*, 1281–1288.

Silva, J.M., Marran, K., Parker, J.S., Silva, J., Golding, M., Schlabach, M.R., Elledge, S.J., Hannon, G.J., and Chang, K. (2008). Profiling essential genes in human mammary cells by multiplex RNAi screening. Science *319*, 617–620.

Sipo, I., Picó, A.H., Wang, X., Eberle, J., Petersen, I., Weger, S., Poller, W., and Fechner, H. (2006). An improved Tet-On regulatable FasL-adenovirus vector system for lung cancer therapy. J. Mol. Med. *84*, 215–225.

Stegmeier, F., Hu, G., Rickles, R.J., Hannon, G.J., and Elledge, S.J. (2005). A lentiviral microRNA-based system for single-copy polymerase II-regulated RNA interference in mammalian cells. Proc. Natl. Acad. Sci. USA *102*, 13212–13217.

Tomari, Y., and Zamore, P.D. (2005). Perspective: machines for RNAi. Genes Dev. *19*, 517–529.

van Delft, M.F., Wei, A.H., Mason, K.D., Vandenberg, C.J., Chen, L., Czabotar, P.E., Willis, S.N., Scott, C.L., Day, C.L., Cory, S., et al. (2006). The BH3 mimetic ABT-737 targets selective Bcl-2 proteins and efficiently induces apoptosis via Bak/Bax if Mcl-1 is neutralized. Cancer Cell *10*, 389–399.

Vert, J.P., Foveau, N., Lajaunie, C., and Vandenbrouck, Y. (2006). An accurate and interpretable model for siRNA efficacy prediction. BMC Bioinformatics *7*, 520.

Zender, L., Xue, W., Zuber, J., Semighini, C.P., Krasnitz, A., Ma, B., Zender, P., Kubicka, S., Luk, J.M., Schirmacher, P., et al. (2008). An oncogenomics-based in vivo RNAi screen identifies tumor suppressors in liver cancer. Cell *135*, 852–864.

Zeng, Y., Wagner, E.J., and Cullen, B.R. (2002). Both natural and designed micro RNAs can inhibit the expression of cognate mRNAs when expressed in human cells. Mol. Cell *9*, 1327–1333.

Zuber, J., McJunkin, K., Fellmann, C., Dow, L.E., Taylor, M.J., Hannon, G.J., and Lowe, S.W. (2011). Toolkit for evaluating genes required for proliferation and survival using tetracycline-regulated RNAi. Nat. Biotechnol. *29*, 79–83.

Please cite this article in press as: Hodges et al., Directional DNA Methylation Changes and Complex Intermediate States Accompany Lineage Specificity in the Adult Hematopoietic Compartment, Molecular Cell (2011), doi:10.1016/j.molcel.2011.08.026

Molecular Cell
Article

Directional DNA Methylation Changes and Complex Intermediate States Accompany Lineage Specificity in the Adult Hematopoietic Compartment

Emily Hodges,[1,2] Antoine Molaro,[1,2] Camila O. Dos Santos,[1,2] Pramod Thekkat,[1,2] Qiang Song,[3] Philip J. Uren,[3] Jin Park,[3] Jason Butler,[2,4] Shahin Rafii,[2,4] W. Richard McCombie,[1] Andrew D. Smith,[3,*] and Gregory J. Hannon[1,2,*]
[1]Watson School of Biological Sciences, Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA
[2]Howard Hughes Medical Institute
[3]Molecular and Computational Biology, University of Southern California, Los Angeles, CA 90089, USA
[4]Department of Genetic Medicine and Ansary Stem Cell Institute, Weill Cornell Medical College, New York, NY 10065, USA
*Correspondence: andrewds@usc.edu (A.D.S.), hannon@cshl.edu (G.J.H.)
DOI 10.1016/j.molcel.2011.08.026

SUMMARY

DNA methylation has been implicated as an epigenetic component of mechanisms that stabilize cell-fate decisions. Here, we have characterized the methylomes of human female hematopoietic stem/progenitor cells (HSPCs) and mature cells from the myeloid and lymphoid lineages. Hypomethylated regions (HMRs) associated with lineage-specific genes were often methylated in the opposing lineage. In HSPCs, these sites tended to show intermediate, complex patterns that resolve to uniformity upon differentiation, by increased or decreased methylation. Promoter HMRs shared across diverse cell types typically display a constitutive core that expands and contracts in a lineage-specific manner to fine-tune the expression of associated genes. Many newly identified intergenic HMRs, both constitutive and lineage specific, were enriched for factor binding sites with an implied role in genome organization and regulation of gene expression, respectively. Overall, our studies represent an important reference data set and provide insights into directional changes in DNA methylation as cells adopt terminal fates.

INTRODUCTION

Development and tissue homeostasis rely on the balance between faithful stem-cell self-renewal and the ordered, sequential execution of programs essential for lineage commitment. Under normal circumstances, commitment is thought to be unidirectional with repressive epigenetic marks stabilizing loss of plasticity (De Carvalho et al., 2010). However, certain differentiated mammalian cells can be reverted to an induced pluripotent state (iPSCs) through exogenous transduction of specific transcription factors (Takahashi and Yamanaka, 2006). Yet, even these reprogrammed cells retain a residual “memory” of their former fate, displaying DNA methylation signatures specific to their tissue of origin (Kim et al., 2010).

DNA methylation is critical for the self-renewal and normal differentiation of somatic stem cells. For example, within the hematopoietic compartment, impaired DNA methyltransferase function disrupts stem cell maintenance (Maunakea et al., 2010; Trowbridge and Orkin, 2010), and loss of DNMT1 leads to defective differentiation and unbalanced commitment to the myeloid and lymphoid lineages (Bröske et al., 2009; Trowbridge et al., 2009). These studies highlight the well-characterized hematopoietic compartment as a context in which to study the link between DNA methylation patterns and cell-fate specification.

Toward this end, DNA methylation profiles of murine hematopoietic progenitors through early stages of lineage commitment were recently compared with CHARM (Irizarry et al., 2008; Ji et al., 2010), which profiles a predefined set of CpG-dense intervals. Overall, CHARM revealed that early lymphopoeisis involves more global acquisition of DNA methylation than myelopoiesis and that DNMT1 inhibition skews progenitors toward the myeloid state. These data support earlier reports that DNMT1 hypomorphic hematopoietic stem and progenitor cells (HSPCs) show reduced lymphoid differentiation potential (Bröske et al., 2009). Importantly, regions identified to have differential methylation through sequential stages of differentiation most often did not correspond to CpG islands (CGIs) but instead lay adjacent in areas referred to as “shores.”

Higher-resolution maps of DNA methylation with shotgun bisulfite sequencing have mainly been produced from cultured cells (Laurent et al., 2010; Lister et al., 2009) or mixed cell types (Li et al., 2010). Several unexpected findings emerged from these early studies including significant frequencies of cytosines methylated in a non-CpG context in human embryonic stem cells (ESCs), a characteristic previously thought to be restricted to plants. Other genome-wide studies have implicated DNA methylation in the regulation of alternative promoters and even RNA splicing patterns (Maunakea et al., 2010). These observations emphasize the need for complete, unbiased, and quantitative assessment of cytosine methylation and the establishment of reference methylomes from purified populations of primary cells.

Here, we performed whole-genome shotgun bisulfite sequencing on female human HSPCs, B cells, and neutrophils to

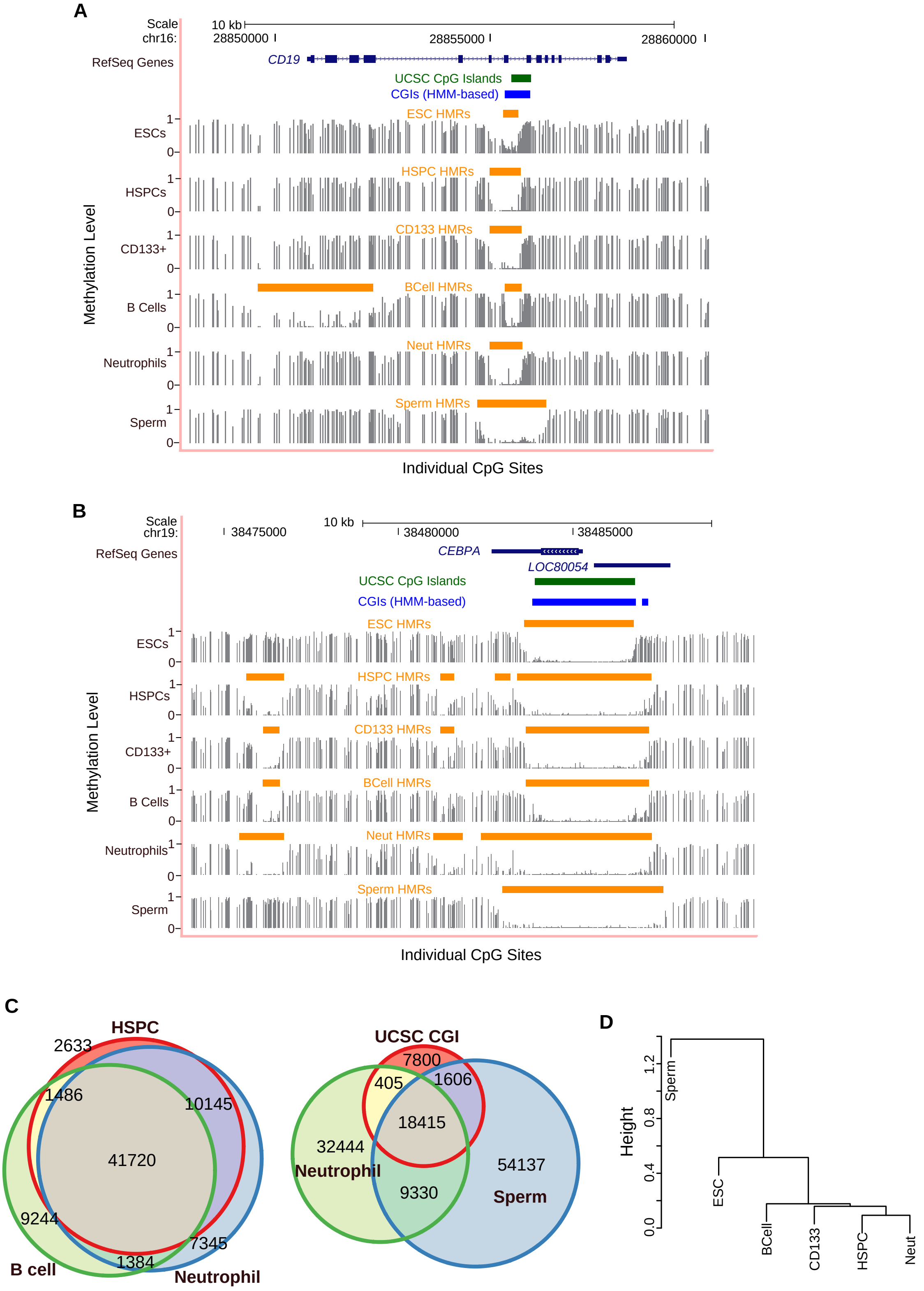


Figure 1. Features of Methylomes in Hematopoietic Cells

(A and B) Genome browser tracks depict methylation profiles across a lymphoid (A) and myeloid (B) specific locus in blood cells, ESCs, and sperm. Methylation frequencies, ranging between 0 and 1, of unique reads covering individual CpG sites are shown in gray with identified hypomethylated regions (HMRs) indicated

Please cite this article in press as: Hodges et al., Directional DNA Methylation Changes and Complex Intermediate States Accompany Lineage Specificity in the Adult Hematopoietic Compartment, Molecular Cell (2011), doi:10.1016/j.molcel.2011.08.026

examine the relationships between the methylation states of multipotent blood-forming stem cells and two divergent derived lineages. This enabled us to probe directional changes in DNA methylation associated with cell-fate specification. Comparison of the three reference methylomes revealed a number of important principles of epigenetic regulation, in addition to providing insights into the dynamics of epigenetic changes during development.

RESULTS AND DISCUSSION

Lineage-Specific Hypomethylated Regions Extend beyond Annotated CGIs

We sought to generate reference, single nucleotide-resolution methylation profiles for several nodes within the human hematopoietic lineage using whole-genome bisulfite sequencing (see the Experimental Procedures). Therefore, we examined CD34+ CD38–Lin– HSPCs, CD19+ B cells, and granulocytic neutrophils from peripheral blood of pooled human female donors. These cell types represent one of the earliest self-renewing, multipotent populations, and two derived, mature cell types from the lymphoid and myeloid lineages, respectively. For comparison, we generated methylomes from HSPCs from male umbilical cord blood (CD133+CD34+CD38–Lin–) and compared to data sets created from primate sperm (Molaro et al., 2011) and embryonic stem cells (Laurent et al., 2010). In all cases, we achieved a median of 10× independent sequence coverage, sufficient to interrogate 96% of genomic CpG sites (Figure S1A and Table S1A available online). While this level of coverage is still subject to sampling error at individual sites (see discussion in Hodges et al., 2009), features such as transitions from high to low levels of methylation can still be identified with a resolution of the boundaries to within a few CpG sites.

In the genome as a whole, CpG dinucleotides have a strong tendency to be methylated (70%–80%) (Lister et al., 2009). Coincidently, CpGs are also underrepresented, perhaps because of their vulnerability to methylation-induced deamination and consequent loss over evolutionary time (Cooper and Krawczak, 1989; Gardiner-Garden and Frommer, 1987). Areas of increased CpG density, called CpG islands (CGIs) have a lower probability of being methylated and these or their adjacent regions (CGI shores) have been implicated as potential regulatory domains (Gardiner-Garden and Frommer, 1987; Irizarry et al., 2009a; Wu et al., 2010). Though CGIs have been defined computationally (Irizarry et al., 2009b), we developed an algorithm to identify hypomethylated regions (HMRs) empirically in bisulfite sequencing data sets, based on their methylation state alone (see Figures 1A and 1B).

Between 50,000 and 60,000 HMRs were identified from each hematopoietic profile (Table S1B), with neutrophils displaying the greatest number (~60,000), followed by HSPCs (~55,000) and B lymphocytes (~53,000) (Figure 1C). Interestingly, this was lower than the number in male germ cells (~80,000), perhaps because of the extensive repeat hypomethylation observed in sperm as compared to somatic cells.

Certainly, many annotated CGIs were contained within our set of functionally defined HMRs; however, CGIs appeared to fall short as a benchmark by which to define all HMRs with probable regulatory significance. Annotated CGIs accounted for fewer than half of the HMRs identified in any cell type (Figure 1C and Figure S1B). Moreover, many HMRs whose biological relevance is supported by lineage-specific methylation failed to meet the conservative CGI criteria.

Sequence tracks showing methylation levels for a lymphoid- (Figure 1A) or myeloid- (Figure 1B) specific gene illustrate several characteristics of HMRs. The locus for the B cell marker *CD19* displays a broad, cell type-specific HMR at its transcriptional start site (TSS), which does not overlap a predicted CGI. In contrast, "tidal" methylation at CGI shores characterizes several HMRs surrounding the myeloid transcription factor, *CEBPA*. The cores of these HMRs are shared among blood forming cells, but their widths differ, with neutrophils demonstrating the most expansive hypomethylation. In fact, shared HMRs often show variable widths, suggesting that the boundaries of HMRs fluctuate in a cell type-dependent manner. Due to the dynamic behavior of the HMRs, we were motivated to seek further validation of these characteristics as biological phenomena, rather than as technical artifacts of the methodology. Therefore, we focused on an independent dataset derived from chimpanzee. We reasoned that genic relationships to methylation dynamics should be preserved in closely related species. Indeed, HMRs show significant overlap between human and chimp, with chimp HMRs following very similar patterns of boundary fluctuations (Table S1C and Figure S2).

While a high proportion of identified HMRs (≥70%) intersected all blood cell types studied, ~10-fold more HMRs were shared only between HSPCs and neutrophils than exclusively between HSPCs and B cells (Figure 1C). In contrast, ~45%–50% of HMRs identified in blood cells overlap sperm HMRs. Interestingly, the diversity of differentially expressed genes within the hematopoietic lineage has been reported to be similar to the complexity observed across human tissues (Novershtern et al., 2011). However, at the epigenetic level, HMR profiles easily distinguished closely related cell types (blood forming) from distantly related ones (Figure 1D), indicating that patterns of DNA methylation are strongly correlated within a lineage.

HMR Expansion Correlates with Differential Expression

Differentially methylated regions (DMRs) at promoters have been ascribed regulatory roles, with differential methylation being

by orange bars. UCSC predicted/annotated CpG islands (green bars) and HMM-based CpG islands (blue bars) (Irizarry et al., 2009b) are also displayed. Numbers (top) indicate base position along the chromosome.

(C) Venn diagrams depict the intersection between HMRs identified in blood as well as the overlap between blood-derived cells, sperm, and UCSC CpG islands. The size of the circles and the proportion of circle overlap reflect the relative number of HMRs identified as well as the degree of intersection between each set of HMRs.

(D) Dendrogram clusters cell-types according to their pearson correlations of individual CpG methylation levels within HMRs, both overlapping and nonoverlapping, across all tissues examined.

See also Figures S1 and S2 and Table S1.

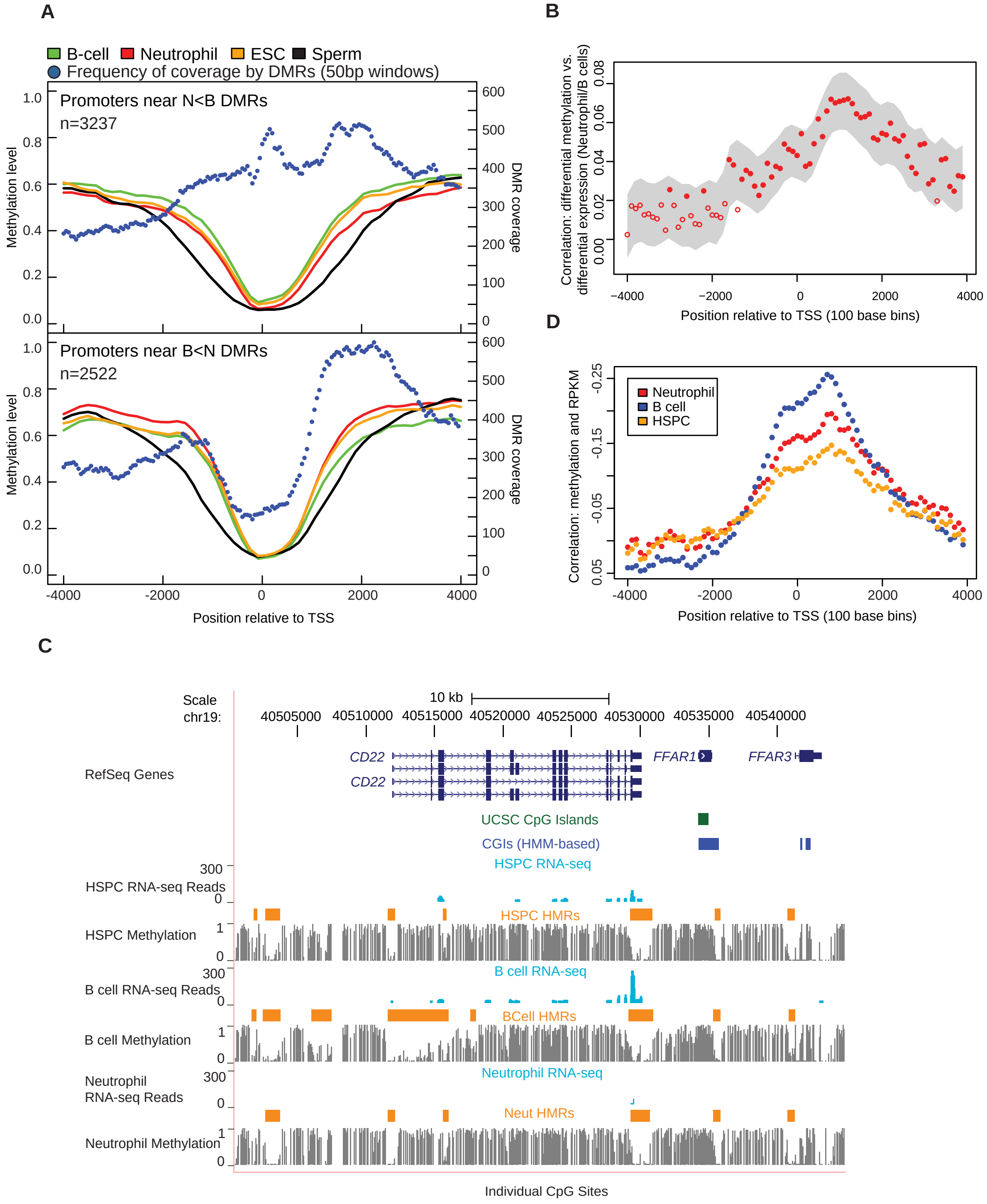


Figure 2. Promoter Differential Methylation and Gene Expression

(A) Average methylation levels across promoters of genes having a DMR within 4 kb of the TSS are shown. Two separate graphs display neutrophil hypomethylated promoter DMRs relative to B cells (N < B, top) and B cell hypomethylated promoter DMRs relative to neutrophils (B < N, bottom). The number of DMRs covering nonoverlapping 50 bp windows across the promoter is also shown.

(B) Correlations between differential methylation and differential expression between neutrophils and B cells as a function of position relative to the TSS are shown. The correlations were obtained by comparing log odds of differential methylation and log of RPKM. The probability for differential methylation at a given CpG is described in the Supplemental Experimental Procedures. The gray area displays the smoothed 95% confidence interval. The closed circles indicate correlation coefficients that are significantly different from 0.

linked to tissue-specific expression. Yet, HSPCs, B cells, and neutrophils mainly share promoter-associated HMRs at differentially expressed genes. Prior studies have associated changes in gene expression with changes in methylation states adjacent to constitutively hypomethylated CGIs, in so-called ''CGI shores'' (Irizarry et al., 2009a). Therefore, we examined correlations between the geography of promoter HMRs and changes in lineage-specific expression, focusing on a comparison of B cells and neutrophils.

Differential methylation often manifested as a broadening of TSS-associated HMRs in a specific lineage (Table S2A). The changes were asymmetric, with the greatest loss of methylation on the gene-ward side (Wilcoxon ranks sum: $p < 5e-60$, both DMR sets). Globally, these HMRs were broadest in sperm and constricted in ESCs (Figure 2A) (see also Molaro et al., 2011), widening again in a tissue-specific fashion. Thus, our analyses provide global support for ''tidal'' methylation changes at CGI shores.

For deeper analysis of these tidal patterns, we measured differential methylation in 50 base windows surrounding TSSs (Figure 2A). Moving 3′ toward B cell hypomethylated promoters (B < N), coverage by DMRs peaked between 1.5 Kbp and 2 Kbp downstream of the TSS. A slightly different pattern was observed for neutrophil hypomethylated promoters (N < B), with DMRs rising to a peak directly at the TSS. In both data sets, the greatest concentration of differential methylation occurred ~1–2 Kb downstream of the TSS, consistent with overall methylation being selectively reduced in the transcribed regions of genes with tissue-specific DMRs.

We next asked whether any element of DMR geography correlated with tissue-specific gene expression. We carried out RNA-seq and computed RPKM values for each cell type (Table S2B). We then computed the correlation between differential expression and differential methylation in 100 base windows surrounding the TSS (see the Experimental Procedures). This correlation was strongly asymmetric, peaking ~1,000 bases downstream of the TSS. Notably, this corresponded with the expansion of HMRs that contributes to tissue-specific promoter hypomethylation (Figure 2B).

CD22 provides a specific example of the general phenomena that we observed (Figure 2C). *CD22* is expressed in B cells, but not neutrophils. In each cell type its TSS is covered by an HMR, which in HSPCs and neutrophils extends ~500 bp and centered on the TSS. In B cells, the HMR begins at the same position upstream of the *CD22* TSS, but extends more than 4,300 bp into the transcribed region.

The properties noted for differentially expressed genes were extensible to the entire set of REFSEQ genes. Though hypomethylation was largely symmetric around REFSEQ TSSs, a strong correlation could be seen between RPKM and lower methylation levels peaking ~1.0 Kb downstream of the TSS (Figure 2D). This was true of all cell types examined, though the magnitude of the effect was lowest in HSPCs.

Our results are in accord with a recent study that revealed a unique chromatin signature surrounding the TSS of tissue-specific loci. Spreading of H3K4me2 into the 5′ untranslated region (UTR) was observed at tissue-specific genes, whereas it remained as a discrete peak at the TSS of ubiquitously expressed genes (Pekowska et al., 2010). To look for similar relationships between histone profiles and expanding promoter HMRs, we analyzed chromatin immunoprecipitation sequencing (ChIP-seq) data for H3K4me3, H3K4me1, and H3K27ac enrichment across eight different ENCODE cell lines (Bernstein et al., 2005; Birney et al., 2007). The ENCODE cell lines are derived from a variety of tissues and include GM12878, which is a lymphoblastoid cell line. First, we observe a strong enrichment for these histone marks at B cell promoters containing expanded HMRs. In addition, the greatest difference between the lymphoid cell line and the other cell lines appears upstream and downstream of the TSS compared to all promoters. Interestingly, the H3K4me3 differential enrichment is biased on the 3′ side of the TSS (Figure 3).

It has also been noted that for a subset of CGI-associated promoters, high CpG density extends downstream of the TSS and hypomethylation of the extended region is required for RNA polymerase II binding (Appanah et al., 2007). In fact, analysis of existing lymphoid ChIP-seq data of RNA polymerase II revealed a 3× enrichment in B cell expanded HMR regions compared to neutrophil-expanded regions (Table S2C) (Barski et al., 2010). This suggests that while core CGI promoters remain hypomethylated by default, expansion downstream of the TSS may be important for productive transcription.

Features of Shared and Lineage-Specific Intergenic HMRs

While REFSEQ gene promoters were often associated with an HMR, the majority of HMRs were not found at promoters (Figure S3). Nearly half of all identified HMRs were located in gene bodies. An additional quarter lay >10 Kb from the nearest annotated genes, and we defined this class as ''intergenic HMRs.''

Like promoter-associated HMRs, intergenic HMRs showed sequence conservation, suggesting that these are functional elements (Figure 4A). In fact, genome-wide comparisons of methylation states of orthologous sites in the corresponding cell types of chimpanzee supported concomitant conservation of constitutive and cell type-specific patterns of intergenic methylation (data not shown). Intergenic HMRs tended to be narrower than those found at promoters and were less likely to be shared among cell types. When they were shared, they displayed patterns of expansion and contraction very similar to what was observed for promoter-associated regions (Figure 4A), with their overall extent being widest in sperm.

(C) The browser image shows gene expression for CD22 in the form of mapped read profiles from RNA-seq data. Methylation profiles are also shown (as in Figure 1A) along with HMRs.

(D) Correlations between methylation levels and expression levels represented by RPKM values are shown as a function of position relative to the TSS. Correlation coefficients were averaged in 100 bp bins across regions between 4 kb upstream and downstream of the TSS. Y axis labels were reversed.

See also Figure S3 and Table S2.

Please cite this article in press as: Hodges et al., Directional DNA Methylation Changes and Complex Intermediate States Accompany Lineage Specificity in the Adult Hematopoietic Compartment, Molecular Cell (2011), doi:10.1016/j.molcel.2011.08.026

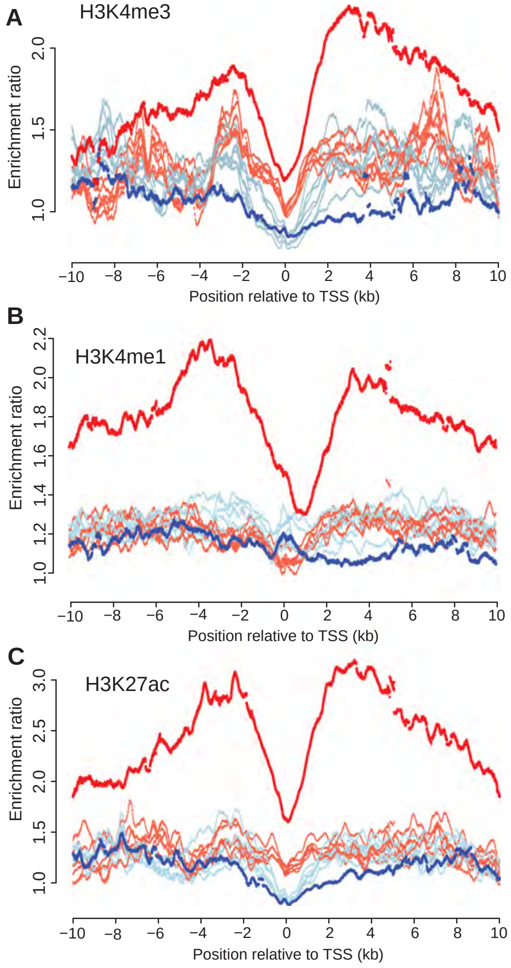


Figure 3. Histone Enrichment across Expanded HMRs

Read count enrichment ratios per 25 bp bins located 10 kb upstream and 10 kb downstream of the TSS were calculated for promoters overlapping HMRs included in Figure 2A for B cell HMRs (red lines) or neutrophil HMRs (blue lines) for H3K4me3 (A), H3K4me1 (B), and H3K27ac (C) by comparison of read counts across all REFSEQ annotated promoters. Data were obtained from ENCODE and include histone profiles for eight different cell lines. The lymphoblastoid cell line GM12878 is highlighted in darker shaded colors.

An early, pervasive view of DNA methylation proposed that germ cell profiles should represent a default state of hypomethylation in all potential regulatory regions (Gardiner-Garden and Frommer, 1987). This was based on the idea that hypomethylation in germ cells would prevent CpG erosion over evolutionary time spans. The high number of nonoverlapping HMRs in the adult somatic cell strongly argues against both of these notions (Figure 1C). However, the width of both genic and intergenic HMRs in sperm compared to somatic cells suggests that germ cells can define the ultimate boundaries of somatic HMRs.

Guided by the strong general enrichment for potential transcription factor binding sites in all HMRs (see Table 1), we searched for motifs in intergenic DMRs specific to neutrophils or B cells (Figure 4B). The strongest scoring motifs in the neutrophil-specific intergenic DMRs included those associated with C/EBP and ETS families, along with HLF and STAT motifs. This striking enrichment for C/EBP and ETS family binding sites is consistent with the functions of ETS factor PU.1 and several C/EBP factors as multipotent progenitors commit to become myeloblasts, which ultimately give rise to neutrophils (Nerlov and Graf, 1998). Because the ETS family contains a large number of transcription factors, we sought experimental support for their binding at HMRs. Therefore we probed existing ChIP-seq data of PU.1 from human HSPCs (Novershtern et al., 2011). We find numerous examples PU.1 enrichment in HMRs, several of which are provided in Figure S4. In contrast, the strongest scoring motifs in B cell-specific intergenic DMRs included the EBF motif, POU family motifs, E-boxes, a PAX motif, and those associated with NFκB and IRF. The simultaneous enrichment of EBF, E-box, and PAX motifs is consistent with the interacting roles of EBF, E2A (which binds E-boxes) and PAX5 as common lymphoid progenitors progress along the B cell lineage (Lin et al., 2010; Medina et al., 2004; Sigvardsson et al., 2002). The enrichment of NFκB and IRF motifs is consistent with the known roles for these factors in both activation and differentiation of lymphocytes (Hayden et al., 2006). Considered together, these analyses strongly suggest that at least a subset of intergenic DMRs can be engaged by tissue-specific transcription factors, leading to changes in chromatin organization that might have long-distance impacts on annotated genes or more local impacts on as yet unidentified ncRNAs. In fact, we do find evidence of transcriptional activity surrounding intergenic DMRs in our RNA-seq data sets, but we have not yet pursued this observation further (data not shown). Irrespective of the model, our results strongly support the biological relevance of tissue-specific intergenic HMRs.

We also probed the possible functions of shared intergenic HMRs. Prior studies had experimentally identified binding sites for the insulator protein, CTCF, by chromatin immunoprecipitation (Kim et al., 2007). These sites are strongly enriched (155-fold) in nonrepeat intergenic HMRs that are common to all cell types examined. In fact, ~90% (>500) of the nonrepeat, shared intergenic HMRs contain a CTCF site. This correlates with the known propensity of CTCF to bind unmethylated regions and suggests that many of the shared intergenic HMRs that we detect may function in the structural organization of chromosomes and nuclear domains.

Myeloid-Biased, Poised Methylation States Characterize HSPC Methylomes

For loci whose differential expression characterizes the lymphoid and myeloid lineages, we set out with a simple general expectation. Low methylation levels in stem and progenitor cells would be permissive for expression in either lineage, and an accumulation of methylation during differentiation would correlate with silencing of loci in the lineage in which they are not expressed.

To test this hypothesis, we selected lineage-specific HMRs arising from a comparison of neutrophils and B cells and examined their status in HSPCs. Both at the level of individual CpGs (Figure 5A) and at the level of overall methylation (Figure 5B), HSPCs showed intermediate methylation states at sites where B cells and neutrophils show opposing methylation patterns.

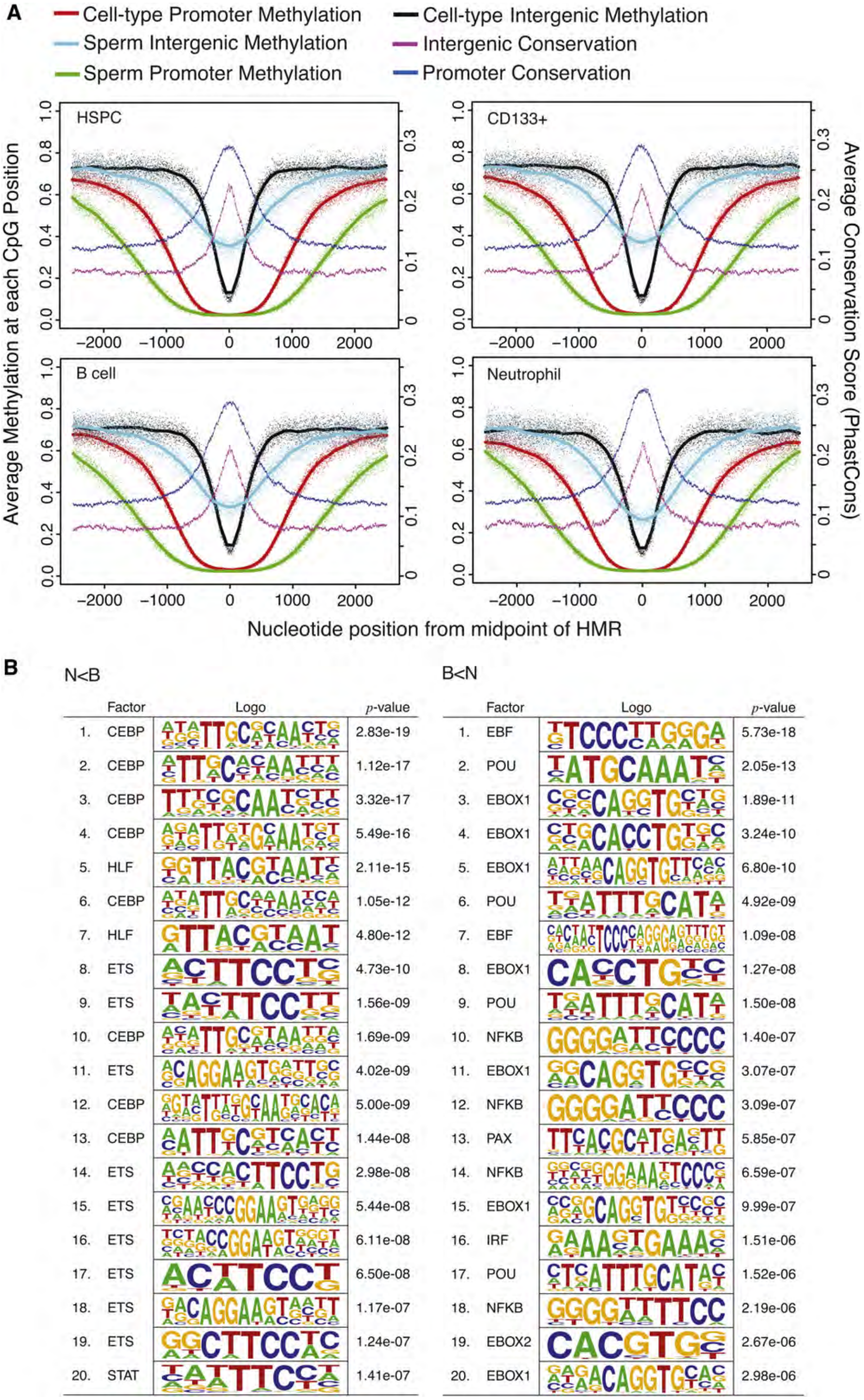


Figure 4. Features of Intergenic HMRs and DMRs

(A) Composite methylation profiles are plotted for individual CpG sites within HMRs. The x axes of the plots indicate genomic position centered on the midpoint of HMRs in the reference cell type labeled for each plot. Methylation profiles are given for the reference cell and sperm, separately for regions where the reference

This suggests that differentiation involves both gains and losses of DNA methylation at lineage-specific HMRs, an observation consistent with recent studies using other methodologies (Attema et al., 2007; Claus et al., 2005; Ji et al., 2010).

At the level of individual CpGs, HSPC patterns correlated better with those seen in neutrophils at myeloid HMRs than they did with B cell methylation patterns at nonoverlapping lymphoid HMRs (Figure 5A). Moreover, the median methylation level for B cells at B cell DMRs was more than twice as high as the median level at neutrophil specific DMRs (Figure 5B). This finding, along with the fact that B cells exhibited fewer total HMRs than either HSPCs or neutrophils, supported an earlier observation that lymphoid commitment in mice involves globally increased DNA methylation (Ji et al., 2010). As a whole, our results indicate that the HSPC methylome has more myeloid than lymphoid character. Many fewer DMRs were identified in comparisons of HSPC and neutrophil methylation profiles than of HSPCs and B cells (Figure S3). Such a myeloid bias is also consistent with prior studies, which point to the myeloid lineage as a default differentiation path for HSPCs (Månsson et al., 2007).

Regions that exhibit intermediate methylation occurred in two forms. The well-documented mode is allelic methylation that is characteristic of dosage compensated and imprinted genes. We detected such loci abundantly in our data sets, and these encompassed both known monoallelic genes and new candidates for monoallelic expression (data not shown). More prevalent were regions of intermediate methylation wherein each chromosome displayed different patterns of CpG modification with little correlation between the states of adjacent CpGs. Partially methylated regions were previously noted in ESCs (Lister et al., 2009), though they did not investigate whether these presented allelic versus stochastic and complex patterns.

To discriminate between allelic and complex patterns, we performed targeted conventional bisulfite PCR sequencing of individual clones from HSPCs across a selected set of myeloid loci and a known locus with allele-specific methylation (Figure 5C, Figure S5, and Table S3). This allowed detailed analysis of adjacent CpG methylation on individual molecules. As expected, for the allelic *XIST* locus on chromosome X, we observed uniform methylation profiles of adjacent CpG sites within individual clones representing two states that contributed nearly equally to the partial methylation observed. In contrast, the myeloid *AZU1* locus exemplified a stochastic pattern of methylation in HSPC. We cannot determine whether the complex states that we observed were in dynamic equilibrium or whether they were fixed in each chromosome that contributed to our analysis.

While the mechanisms underlying complex, partial methylation patterns in HSPCs are unclear, they are reminiscent of bivalent promoters that contain both repressive and active histone marks (Bernstein et al., 2006). Both during embryonic development and during stem cell differentiation, such poised promoters are converted to a determinate chromatin state by shifting the balance of histone marks. This has already been noted for lineage-specific genes in HSPCs (Attema et al., 2007), and our data indicate that this well-established property of chromatin may also extend to DNA methylation patterns.

Alternative explanations for our results must also be considered. Since we have used pooled individuals, each of the observed patterns could be specific to one donor, giving rise to a complex pool of clones; however, this seems unlikely as we also detect lower correlations between neighboring CpGs within single clones. Alternatively, complex states could represent heterogeneity within the isolated HSPC population (see Figure S6), with our data coming from a mixture of self-renewing and more committed cell types. To investigate this possibility, we searched within our RNA-seq data for expression patterns characteristic of each purified cell population. Transcriptional profiles revealed the top differentially expressed genes within the HSPC compartment to be highly enriched for signature gene markers associated with self-renewing hematopoietic stem cells (Figure 5D) and depleted for genes associated with committed progenitors. Collectively, these data suggest that the observed methylation patterns are likely derived from a highly enriched stem cell population, and indicate that those populations may naturally adopt complex, potentially dynamic, methylation patterns at lineage-specific HMRs.

Both the general trends of methylation loss along a lineage and the possibility of dynamic poised methylation states imply that demethylation, either passive or active, is a common event. In mammals, factors capable of promoting active demethylation have remained somewhat elusive (Ooi and Bestor, 2008). In vitro studies have demonstrated that MBD2, a methyl-CpG binding protein, can specifically demethylate cytosines, and components of the elongator complex and the cytidine deaminase, AID, have been implicated in demethylation during early development (Bhattacharya et al., 1999; Okada et al., 2010; Popp et al., 2010). Furthermore, in zebrafish, the coordinated activities of glycosylases, deaminases, and DNA repair proteins have been reported to cause differentiation defects when disrupted, and this has been posited as an effect of improper DNA methylation (Rai et al., 2010). Alternatively, demethylation could potentially be achieved through the action of hydroxymethylases (e.g., TET1-3), which have been proposed to execute an intermediate step toward methylation loss (Ito et al., 2010; Tahiliani et al., 2009; Zhang et al., 2010). Additional information will be necessary to resolve the relevance of any of these pathways to the transition in methylation states between HSPCs and mature neutrophils and B cells.

As a whole, our data not only provide insights into the global behavior of DNA methylation, both in individual cell types and along a well-characterized lineage, but also provide a critical

cell HMR spans a TSS and intergenic region (>10 Kbp from any RefSeq transcript; not overlapping a repeat). Average cross-species conservation scores from PhyloP probabilities derived from 44-way multiple alignments are plotted separately for promoter and intergenic HMRs.

(B) Transcription factor binding site motifs enriched in DMRs between neutrophils and B cells are shown. The top 20 most enriched motifs are shown separately for N < B and B < N DMRs, based on the motifclass tool in the CREAD package. See the Supplemental Experimental Procedures for details of enrichment calculations.

See also Figures S3 and S4.

reference data set to enable detailed future studies of both the mechanisms that set somatic DNA methylation patterns and the consequences of those patterns for gene expression and genome organization.

Table 1. TFBS Enrichment in HMRs across Intergenic and Promoter Regions

Cell	Region	CGI?	HMR[a]	TFBS	Expected	Enrichment
N/A	promoter		34,257	244,998	91,570.8	2.7
	promoter	cgi	24,601	191,452	65,760.9	2.9
	promoter	nocgi	9,656	53,852	25,810	2.1
	intergenic	cgi	10,630	13,608	4,603.76	3.0
B Cell	all		53,834	339,943	76,196.1	4.5
	intergenic		5,849	16,150	3,779	4.3
	intergenic	cgi	1,670	4,802	1,194.97	4.0
	intergenic	nocgi	4,179	11,348	2,584.01	4.4
	promoter		13,650	212,644	36,548.3	5.8
	promoter	cgi	12,828	206,556	35,080	5.9
	promoter	nocgi	822	6,088	1,468.27	4.1
CD133	all		49,593	339,191	67,778.2	5.0
	intergenic		6,494	17,708	3,816.73	4.6
	intergenic	cgi	1,630	4,817	1,207.45	4.0
	intergenic	nocgi	4,864	12,891	2,609.26	4.9
	promoter		13,745	224,955	37,395.1	6.0
	promoter	cgi	12,965	219,407	36,309.9	6.0
	promoter	nocgi	780	5,548	1,085.18	5.1
ESC	all		40,476	318,377	65,062.3	4.9
	intergenic		3,768	11,220	2,404.28	4.7
	intergenic	cgi	1,151	3,295	882.802	3.7
	intergenic	nocgi	2,617	7,925	1,521.45	5.2
	promoter		13,098	222,654	36,332.4	6.1
	promoter	cgi	12,661	218,765	35,769.4	6.1
	promoter	nocgi	437	3,889	562.951	6.9
HSPC	all		55,984	352,574	77,671.2	4.5
	intergenic		6,154	17,619	3,972.1	4.4
	intergenic	cgi	1,663	4,775	1,222.27	3.9
	intergenic	nocgi	4,491	12,844	2,749.81	4.7
	promoter		13,820	222,635	37,830.8	5.9
	promoter	cgi	12,948	216,433	36,461.3	5.9
	promoter	nocgi	872	6,202	1,369.4	4.5
Neut.	all		60,594	362,074	82,427.7	4.4
	intergenic		6,422	18,515	4,212.75	4.4
	intergenic	cgi	1,626	4,760	1,243.88	3.8
	intergenic	nocgi	4,796	13,755	2,968.85	4.6
	promoter		13,862	224,621	38,503.6	5.8
	promoter	cgi	12,950	218,281	37,060.6	5.9
	promoter	nocgi	912	6,340	1,442.93	4.4
Sperm	all		81,446	440,856	201,006	2.2
	intergenic		2,616	14,903	3,158.15	4.7
	intergenic	cgi	865	6,181	1,307.11	4.7
	intergenic	nocgi	1,751	8,722	1,851.02	4.7
	promoter		14,051	270,798	63,641.3	4.3
	promoter	cgi	13,588	266,658	62,357.8	4.3
	promoter	nocgi	463	4,140	1,283.49	3.2

Enrichment of predicted transcription factor binding sites (TFBSs) in intergenic HMRs and HMRs that overlap promoters. For each set of HMRs, corresponding to a cell type, the TFBS enrichment (observed/expected site counts) is given for all HMRs, those overlapping promoters, those that are intergenic, separately according to whether the HMRs overlap CGIs. Data are presented for each of the following cell types: B cells, CD133 cord blood, HSPCs, ESCs, neutrophils, and sperm. For comparison, the TFBS enrichment in the full set of promoters (including those overlapping CGIs) is given, along with enrichment in the full set of intergenic CGIs.

[a] For the “N/A” group, the HMRs are simply the number of promoters or CGIs.

EXPERIMENTAL PROCEDURES

Flow Cytometry and DNA Extraction

Peripheral blood was collected from six healthy female donors ages 25–35 and pooled. After isolation by Ficoll gradient, mononuclear cells were fixed in 1% paraformaldehyde (PFA) and stained with antibodies against the following human cell surface markers (eBiosciences): anti-CD34 (mucosialin) conjugated to PE-Cy7, anti-CD38 conjugated to APC, anti-CD45 conjugated to PE, anti-CD19 conjugated to PE, and anti-CD235a (Glycophorin) conjugated to PE. For lineage depletion, either a combination of PE-conjugated antibodies against CD45, CD19, and CD235a or a commercially available human hematopoietic lineage cocktail was used. CD34+CD38–Lin– hematopoietic stem cells and CD19+ B cells were purified with the FACSAriaII (Becton Dickinson). Neutrophils were purified according to their forward and side-scatter profile. FACS profiles are provided in Figure S6. Umbilical cord blood was collected from a single donor, and CD133+ cells were selected via magnetic separation on CD133+ microbeads (Milteny Biotec) according to instructions supplied by the manufacturer. Two column separations were performed for additional purity. All cells were collected in cell lysis buffer (50 mM Tris, 10 mM EDTA and 1% SDS), and PFA induced crosslinks were reversed with RNase A and a 65°C incubation overnight, after which residual proteins were digested with Proteinase K for 3 hr at 42°C. DNA was extracted with an equal volume of phenol:chloroform, followed by a single extraction with chloroform and ethanol precipitation. Human sperm was purified and sequenced according to methods described in Molaro et al. (2011).

Illumina Library Preparation for Bisulfite Sequencing

Bisulfite sequencing libraries were generated by previously described methods (Hodges et al., 2009) and on the manufacturer’s instructions (Illumina) but with several additional modifications. In brief, after each enzymatic step, genomic DNA was recovered by phenol:chloroform extraction and ethanol precipitation. Adenylated fragments were ligated to Illumina-compatible paired-end adaptors synthesized with 5′-methyl-cytosine, and, when necessary, adaptors were diluted 100×–1000× to compensate for low-input libraries and maintain an approximate 10-fold excess of adaptor oligonucleotides. After ligation, DNA fragments were purified and concentrated on MinElute columns (QIAGEN). The standard gel purification step for size selection was excluded from the protocol. Fragments were denatured and treated with sodium bisulfite with the EZ DNA Methylation Gold kit according to the manufacturer’s instructions (Zymo). Lastly, the sample was desulfonated and the converted, adaptor-ligated fragments were PCR enriched with paired-end adaptor-compatible primers 1.0 and 2.0 (Illumina) and the Expand High Fidelity Plus PCR system (Roche). Paired-end Illumina sequencing was performed on bisulfite converted libraries for 76–100 cycles each end.

RNA-Seq

For isolation of RNA from target cell populations, unfixed (live) cells were sorted as described above into Trizol-LS (Invitrogen), and RNA was purified

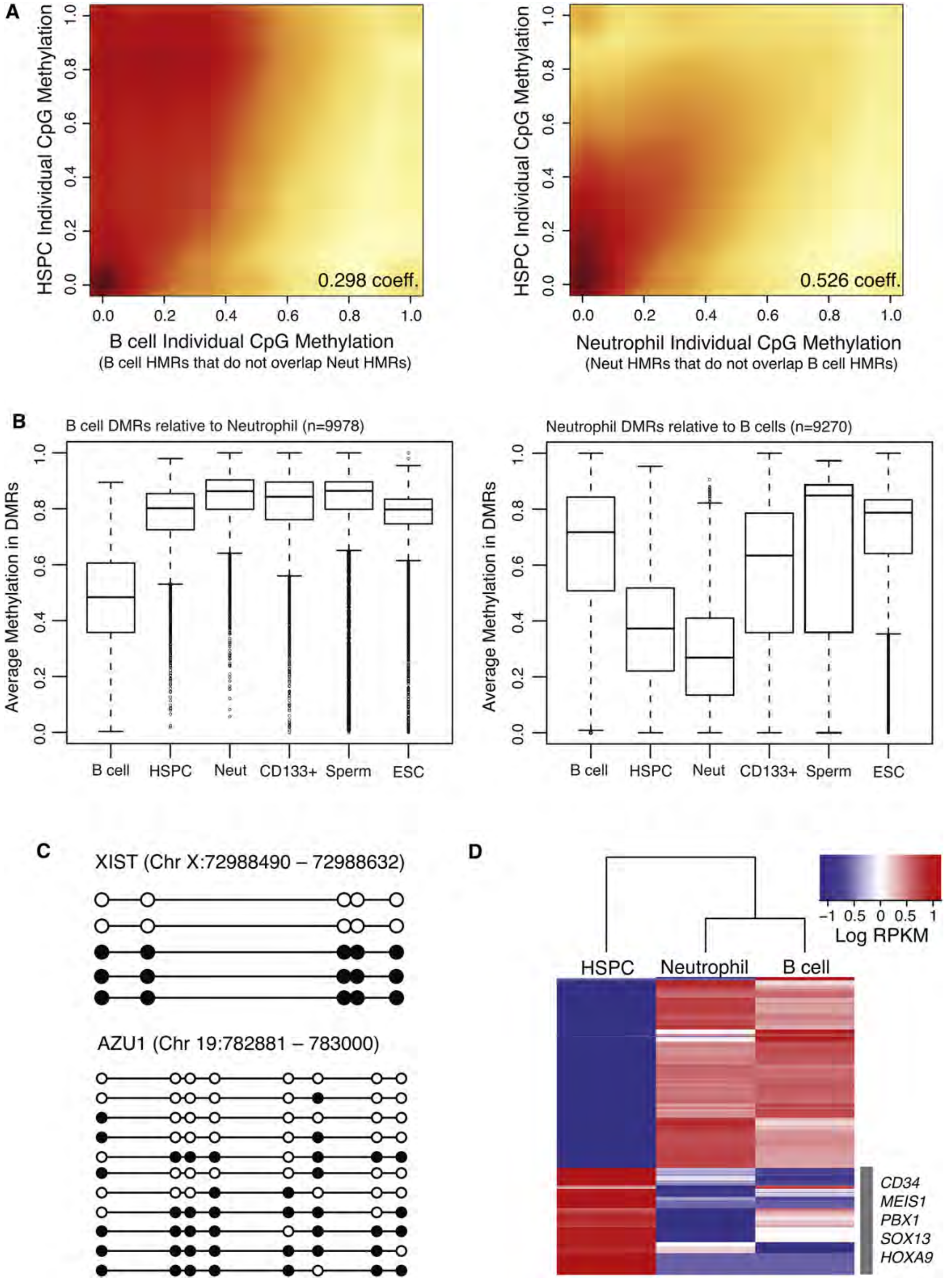


Figure 5. Methylation Dynamics during Lineage Selection

(A) Smoothed scatter plot heat maps showing the correlation between individual CpG methylation levels in HSPCs versus B cells (left) and HSPCs versus neutrophils (right) within B cell- and neutrophil-specific HMRs, respectively. Darker shading (red) indicates greater density of data points, while lighter (yellow) shading reflects lower density. Positive correlations between HSPCs and both B cells and neutrophils indicate an intermediate state for HSPCs.

Please cite this article in press as: Hodges et al., Directional DNA Methylation Changes and Complex Intermediate States Accompany Lineage Specificity in the Adult Hematopoietic Compartment, Molecular Cell (2011), doi:10.1016/j.molcel.2011.08.026

according to the manufacturer's recommendations. Double-stranded complementary DNA (cDNA) libraries were generated with the Ovation RNA-seq system (Nugen). After reverse transcription and cDNA amplification, double-stranded cDNA fragments were phosphorylated, adenylated, and ligated to Illumina paired-end adaptors followed by 15 cycles of PCR amplification with Phusion HF PCR master mix (Finnzymes) according to the standard Illumina protocol for genomic libraries. Single-end sequencing was performed for 36 cycles.

Conventional Bisulfite Cloning and Sanger Sequencing

Genomic DNA isolated from pooled human HSPCs was bisulfite converted with the EZ DNA Methylation Gold kit (Zymo). For selection of specific regions for amplification, forward and reverse primers were designed with Methprimer (Li and Dahiya, 2002). Primer sequences are provided in the Table S3. The following PCR reaction components were combined in a total volume of 25 μl: 5 μl 5× Expand High Fidelity Plus buffer without $MgCl_2$, 1 μl 10 mM dNTPs, 1 μl 10 mM each forward and reverse primers, 2.5 μl 25 mM $MgCl_2$, 2 μl DNA template, and 11.5 μl nuclease-free water. Thermal cycling was performed as follows: 35 cycles each of denaturation at 94°C for 2 min, annealing at 60°C or 53°C for 1 min, and extension at 72°C for 30 s followed by 7 min at 72°C. The PCR products were purified on columns with a PCR purification kit (QIAGEN). PCR products were adenylated with Klenow exo– and purified. Purified amplicons were cloned and sequenced according to previously described methods (Hodges et al., 2009).

Computational Methods Summary

The Supplemental Experimental Procedures contain a detailed description of computational methods. Mapping bisulfite treated reads was done with methods described by Smith et al. (2009) with tools from the RMAP package (Smith et al., 2009). Hypomethylated regions (HMRs) were identified with a hidden Markov model as described in Molaro et al. (2011). DMRs were identified by (1) computation of probabilities of differential methylation at individual CpGs based on number of reads and frequencies of methylation, and (2) identification of peaks in these profiles after kernel smoothing. Cross-species conservation information was taken from UCSC MULTIZ 44-way vertebrate alignments and PhyloP profiles from these alignments.

ACCESSION NUMBERS

Data analyzed herein have been deposited in GEO with accession number GSE31971.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, six figures, and three tables and can be found with this article online at doi:10.1016/j.molcel.2011.08.026.

ACKNOWLEDGMENTS

We thank members of the McCombie lab and Michelle Rooks for help with experimental procedures, and Assaf Gordon, Luigi Manna, and the Cold Spring Harbor Laboratory and University of Southern California High Performance Computing Centers for computational support. Chimp blood was supplied by the New Iberia Research Center and the Southwest National Primate Center. This work was supported in part by grants from the National Institutes of Health and by a kind gift from Kathryn W. Davis (A.S., G.J.H.). The ENCODE ChIP-seq data were generated at the Broad Institute and in the Bradley E. Bernstein lab at the Massachusetts General Hospital/Harvard Medical School. Data generation and analysis was supported by funds from the National Human Genome Research Institute, the Burroughs Wellcome Fund, Massachusetts General Hospital, and the Broad Institute.

Received: May 20, 2011
Revised: July 19, 2011
Accepted: August 26, 2011
Published online: September 15, 2011

REFERENCES

Appanah, R., Dickerson, D.R., Goyal, P., Groudine, M., and Lorincz, M.C. (2007). An unmethylated 3′ promoter-proximal region is required for efficient transcription initiation. PLoS Genet. *3*, e27.

Attema, J.L., Papathanasiou, P., Forsberg, E.C., Xu, J., Smale, S.T., and Weissman, I.L. (2007). Epigenetic characterization of hematopoietic stem cell differentiation using miniChIP and bisulfite sequencing analysis. Proc. Natl. Acad. Sci. USA *104*, 12371–12376.

Barski, A., Chepelev, I., Liko, D., Cuddapah, S., Fleming, A.B., Birch, J., Cui, K., White, R.J., and Zhao, K. (2010). Pol II and its associated epigenetic marks are present at Pol III-transcribed noncoding RNA genes. Nat. Struct. Mol. Biol. *17*, 629–634.

Bernstein, B.E., Kamal, M., Lindblad-Toh, K., Bekiranov, S., Bailey, D.K., Huebert, D.J., McMahon, S., Karlsson, E.K., Kulbokas, E.J., 3rd, Gingeras, T.R., et al. (2005). Genomic maps and comparative analysis of histone modifications in human and mouse. Cell *120*, 169–181.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell *125*, 315–326.

Bhattacharya, S.K., Ramchandani, S., Cervoni, N., and Szyf, M. (1999). A mammalian protein with specific demethylase activity for mCpG DNA. Nature *397*, 579–583.

Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigó, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T., Thurman, R.E., et al; ENCODE Project Consortium; NISC Comparative Sequencing Program; Baylor College of Medicine Human Genome Sequencing Center; Washington University Genome Sequencing Center; Broad Institute; Children's Hospital Oakland Research Institute. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447*, 799–816.

Bröske, A.M., Vockentanz, L., Kharazi, S., Huska, M.R., Mancini, E., Scheller, M., Kuhl, C., Enns, A., Prinz, M., Jaenisch, R., et al. (2009). DNA methylation protects hematopoietic stem cell multipotency from myeloerythroid restriction. Nat. Genet. *41*, 1207–1215.

Claus, R., Almstedt, M., and Lübbert, M. (2005). Epigenetic treatment of hematopoietic malignancies: in vivo targets of demethylating agents. Semin. Oncol. *32*, 511–520.

Cooper, D.N., and Krawczak, M. (1989). Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. Hum. Genet. *83*, 181–188.

(B) Box plots show the distribution of average methylation levels in regions of differential methylation (DMRs) between B cells and neutrophils. Whiskers represent minimum and maximum values, while boxes depict the interquartile range, with horizontal lines indicating the median value. Outliers are shown as open circles.
(C) Lollipop diagrams display the methylation status of HSPC-derived clones sequenced by conventional methods following bisulfite conversion and site-specific PCR amplification across an interval near the XIST gene (top) and the AZU1 gene (bottom). Filled and open circles represent methylated and unmethylated CpG sites, respectively.
(D) Heat map of log RPKM values show expression levels for the top 100 differentially expressed genes (rows), selected for high expression in one cell type compared to the other, in each cell population (columns). Signature marker genes found within the HSPC cluster are listed.
See also Figures S3, S5, and S6 and Table S3.

De Carvalho, D.D., You, J.S., and Jones, P.A. (2010). DNA methylation and cellular reprogramming. Trends Cell Biol. *20*, 609–617.

Gardiner-Garden, M., and Frommer, M. (1987). CpG islands in vertebrate genomes. J. Mol. Biol. *196*, 261–282.

Hayden, M.S., West, A.P., and Ghosh, S. (2006). NF-kappaB and the immune response. Oncogene *25*, 6758–6780.

Hodges, E., Smith, A.D., Kendall, J., Xuan, Z., Ravi, K., Rooks, M., Zhang, M.Q., Ye, K., Bhattacharjee, A., Brizuela, L., et al. (2009). High definition profiling of mammalian DNA methylation by array capture and single molecule bisulfite sequencing. Genome Res. *19*, 1593–1605.

Irizarry, R.A., Ladd-Acosta, C., Carvalho, B., Wu, H., Brandenburg, S.A., Jeddeloh, J.A., Wen, B., and Feinberg, A.P. (2008). Comprehensive high-throughput arrays for relative methylation (CHARM). Genome Res. *18*, 780–790.

Irizarry, R.A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., Cui, H., Gabo, K., Rongione, M., Webster, M., et al. (2009a). The human colon cancer methylome shows similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat. Genet. *41*, 178–186.

Irizarry, R.A., Wu, H., and Feinberg, A.P. (2009b). A species-generalized probabilistic model-based definition of CpG islands. Mamm. Genome *20*, 674–680.

Ito, S., D'Alessio, A.C., Taranova, O.V., Hong, K., Sowers, L.C., and Zhang, Y. (2010). Role of Tet proteins in 5mC to 5hmC conversion, ES-cell self-renewal and inner cell mass specification. Nature *466*, 1129–1133.

Ji, H., Ehrlich, L.I., Seita, J., Murakami, P., Doi, A., Lindau, P., Lee, H., Aryee, M.J., Irizarry, R.A., Kim, K., et al. (2010). Comprehensive methylome map of lineage commitment from haematopoietic progenitors. Nature *467*, 338–342.

Kim, T.H., Abdullaev, Z.K., Smith, A.D., Ching, K.A., Loukinov, D.I., Green, R.D., Zhang, M.Q., Lobanenkov, V.V., and Ren, B. (2007). Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell *128*, 1231–1245.

Kim, K., Doi, A., Wen, B., Ng, K., Zhao, R., Cahan, P., Kim, J., Aryee, M.J., Ji, H., Ehrlich, L.I., et al. (2010). Epigenetic memory in induced pluripotent stem cells. Nature *467*, 285–290.

Laurent, L., Wong, E., Li, G., Huynh, T., Tsirigos, A., Ong, C.T., Low, H.M., Kin Sung, K.W., Rigoutsos, I., Loring, J., and Wei, C.L. (2010). Dynamic changes in the human methylome during differentiation. Genome Res. *20*, 320–331.

Li, L.C., and Dahiya, R. (2002). MethPrimer: designing primers for methylation PCRs. Bioinformatics *18*, 1427–1431.

Li, Y., Zhu, J., Tian, G., Li, N., Li, Q., Ye, M., Zheng, H., Yu, J., Wu, H., Sun, J., et al. (2010). The DNA methylome of human peripheral blood mononuclear cells. PLoS Biol. *8*, e1000533.

Lin, Y.C., Jhunjhunwala, S., Benner, C., Heinz, S., Welinder, E., Mansson, R., Sigvardsson, M., Hagman, J., Espinoza, C.A., Dutkowski, J., et al. (2010). A global network of transcription factors, involving E2A, EBF1 and Foxo1, that orchestrates B cell fate. Nat. Immunol. *11*, 635–643.

Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.M., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature *462*, 315–322.

Månsson, R., Hultquist, A., Luc, S., Yang, L., Anderson, K., Kharazi, S., Al-Hashmi, S., Liuba, K., Thorén, L., Adolfsson, J., et al. (2007). Molecular evidence for hierarchical transcriptional lineage priming in fetal and adult stem cells and multipotent progenitors. Immunity *26*, 407–419.

Maunakea, A.K., Nagarajan, R.P., Bilenky, M., Ballinger, T.J., D'Souza, C., Fouse, S.D., Johnson, B.E., Hong, C., Nielsen, C., Zhao, Y., et al. (2010). Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature *466*, 253–257.

Medina, K.L., Pongubala, J.M., Reddy, K.L., Lancki, D.W., Dekoter, R., Kieslinger, M., Grosschedl, R., and Singh, H. (2004). Assembling a gene regulatory network for specification of the B cell fate. Dev. Cell *7*, 607–617.

Molaro, A., Hodges, E., Fang, F., Song, Q., McCombie, W.R., Hannon, G.J., and Smith, A.D. (2011). Sperm methylation profiles reveal features of epigenetic inheritance and evolution in primates. Cell *146*, 1029–1041.

Nerlov, C., and Graf, T. (1998). PU.1 induces myeloid lineage commitment in multipotent hematopoietic progenitors. Genes Dev. *12*, 2403–2412.

Novershtern, N., Subramanian, A., Lawton, L.N., Mak, R.H., Haining, W.N., McConkey, M.E., Habib, N., Yosef, N., Chang, C.Y., Shay, T., et al. (2011). Densely interconnected transcriptional circuits control cell states in human hematopoiesis. Cell *144*, 296–309.

Okada, Y., Yamagata, K., Hong, K., Wakayama, T., and Zhang, Y. (2010). A role for the elongator complex in zygotic paternal genome demethylation. Nature *463*, 554–558.

Ooi, S.K., and Bestor, T.H. (2008). The colorful history of active DNA demethylation. Cell *133*, 1145–1148.

Pekowska, A., Benoukraf, T., Ferrier, P., and Spicuglia, S. (2010). A unique H3K4me2 profile marks tissue-specific gene regulation. Genome Res. *20*, 1493–1502.

Popp, C., Dean, W., Feng, S., Cokus, S.J., Andrews, S., Pellegrini, M., Jacobsen, S.E., and Reik, W. (2010). Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. Nature *463*, 1101–1105.

Rai, K., Sarkar, S., Broadbent, T.J., Voas, M., Grossmann, K.F., Nadauld, L.D., Dehghanizadeh, S., Hagos, F.T., Li, Y., Toth, R.K., et al. (2010). DNA demethylase activity maintains intestinal cells in an undifferentiated state following loss of APC. Cell *142*, 930–942.

Sigvardsson, M., Clark, D.R., Fitzsimmons, D., Doyle, M., Akerblad, P., Breslin, T., Bilke, S., Li, R., Yeamans, C., Zhang, G., and Hagman, J. (2002). Early B-cell factor, E2A, and Pax-5 cooperate to activate the early B cell-specific mb-1 promoter. Mol. Cell. Biol. *22*, 8539–8551.

Smith, A.D., Chung, W.Y., Hodges, E., Kendall, J., Hannon, G., Hicks, J., Xuan, Z., and Zhang, M.Q. (2009). Updates to the RMAP short-read mapping software. Bioinformatics *25*, 2841–2842.

Tahiliani, M., Koh, K.P., Shen, Y., Pastor, W.A., Bandukwala, H., Brudno, Y., Agarwal, S., Iyer, L.M., Liu, D.R., Aravind, L., and Rao, A. (2009). Conversion of 5-methylcytosine to 5-hydroxymethylcytosine in mammalian DNA by MLL partner TET1. Science *324*, 930–935.

Takahashi, K., and Yamanaka, S. (2006). Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. Cell *126*, 663–676.

Trowbridge, J.J., and Orkin, S.H. (2010). DNA methylation in adult stem cells: New insights into self-renewal. Epigenetics *5*, 189–193.

Trowbridge, J.J., Snow, J.W., Kim, J., and Orkin, S.H. (2009). DNA methyltransferase 1 is essential for and uniquely regulates hematopoietic stem and progenitor cells. Cell Stem Cell *5*, 442–449.

Wu, H., Caffo, B., Jaffee, H.A., Irizarry, R.A., and Feinberg, A.P. (2010). Redefining CpG islands using hidden Markov models. Biostatistics *11*, 499–514.

Zhang, H., Zhang, X., Clark, E., Mulcahey, M., Huang, S., and Shi, Y.G. (2010). TET1 is a DNA-binding protein that modulates DNA methylation and gene transcription via hydroxylation of 5-methylcytosine. Cell Res. *20*, 1390–1393.

MicroRNA93 Regulates Proliferation and Differentiation of Normal and Malignant Breast Stem Cells

Suling Liu[1]*, Shivani H. Patel[1], Christophe Ginestier[2], Ingrid Ibarra[3], Rachel Martin-Trevino[1], Shoumin Bai[1], Sean P. McDermott[1], Li Shang[1], Jia Ke[1], Sing J. Ou[1], Amber Heath[1], Kevin J. Zhang[1], Hasan Korkaya[1], Shawn G. Clouthier[1], Emmanuelle Charafe-Jauffret[2], Daniel Birnbaum[2], Gregory J. Hannon[3], Max S. Wicha[1]*

[1] Comprehensive Cancer Center, Department of Internal Medicine, University of Michigan, Ann Arbor, Michigan, United States of America, [2] Centre de Recherche en Cancérologie de Marseille, Laboratoire d'Oncologie Moléculaire, UMR891 Inserm/Institut Paoli-Calmettes, Université de la Méditerranée, Marseille, France, [3] Program in Genetics and Bioinformatics, Cold Spring Harbor Laboratory, Cold Spring Harbor, New York, United States of America

Abstract

MicroRNAs (miRNAs) play important roles in normal cellular differentiation and oncogenesis. microRNA93 (mir-93), a member of the mir106b-25 cluster, located in intron 13 of the MCM7 gene, although frequently overexpressed in human malignancies may also function as a tumor suppressor gene. Using a series of breast cancer cell lines representing different stages of differentiation and mouse xenograft models, we demonstrate that mir-93 modulates the fate of breast cancer stem cells (BCSCs) by regulating their proliferation and differentiation states. In "claudin^{low}" SUM159 cells, expression of mir-93 induces Mesenchymal-Epithelial Transition (MET) associated with downregulation of TGFβ signaling and downregulates multiple stem cell regulatory genes, including JAK1, STAT3, AKT3, SOX4, EZH1, and HMGA2, resulting in cancer stem cell (CSC) depletion. Enforced expression of mir-93 completely blocks tumor development in mammary fat pads and development of metastases following intracardiac injection in mouse xenografts. The effect of mir-93 on the CSC population is dependent on the cellular differentiation state, with mir-93 expression increasing the CSC population in MCF7 cells that display a more differentiated "luminal" phenotype. mir-93 also regulates the proliferation and differentiation of normal breast stem cells isolated from reduction mammoplasties. These studies demonstrate that miRNAs can regulate the states and fates of normal and malignant mammary stem cells, findings which have important biological and clinical implications.

Citation: Liu S, Patel SH, Ginestier C, Ibarra I, Martin-Trevino R, et al. (2012) MicroRNA93 Regulates Proliferation and Differentiation of Normal and Malignant Breast Stem Cells. PLoS Genet 8(6): e1002751. doi:10.1371/journal.pgen.1002751

Editor: Jeffrey M. Rosen, Baylor College of Medicine, United States of America

Received April 20, 2012; **Accepted** April 24, 2012; **Published** June 7, 2012

Funding: This work was supported by Stand Up 2 Cancer (MSW and GJH), NIH grants CA66233 and CA101860 (MSW), DOD Grant (MSW and GJH), University of Michigan Cancer Center NIH support grant 5 P 30 CA46592 (MSW), the Taubman Research Institute (MSW), NIH and HHMI grants (GJH), and INCa PL 2008 and Label LNCC (CG, EC-J, DB). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: MSW is a scientific advisor and has financial holdings in OncoMed Pharmaceuticals.

* E-mail: sulingl@med.umich.edu (SL); mwicha@umich.edu (MSW)

Introduction

miRNAs serve vital functions in many of normal developmental processes, as well as in carcinogenesis. A number of these miRNAs have been shown to function as oncogenes with increased expression in lung cancer, prostate cancer and colorectal cancer [1,2,3,4,5,6,7,8]. In contrast, other miRNAs such as Let7 are frequently downregulated in malignancies including breast cancer and lung cancer in these contexts functioning as a tumor suppressor gene [9,10,11]. The mir106b-25 cluster is composed of the highly conserved miRNA106b (mir-106b), miRNA93 (mir-93) and miRNA25 (mir-25) that have been reported to be overexpressed in a number of cancers including gastric, prostate and pancreatic neural endocrine tumors, neuroblastoma and multiple myeloma [1,2,3]. These miRNAs are located in a 515-base region on chromosome band 7q22 in intron13 of the host MCM7 gene where they are co-transcribed in the context of MCM7 primary transcripts [1]. MCM7 is a DNA licensing factor obligate for cellular replication. Studies have suggested that the mir-106b-25 miRNA cluster functions as a proto oncogene. Several studies suggest that a primary mechanism of oncogenesis involves targeting of PTEN which cooperates with MCM7 to drive cellular proliferation [12]. Despite evidence for this miRNA cluster functioning as a proto oncogene, in some contexts it has been reported to function as a tumor suppressor inhibiting tumor growth [13]. The molecular mechanisms accounting for this discrepancy have not been determined.

Studies associating miRNA expression with oncogenesis have largely been performed in bulk tumor populations. However, there is substantial evidence supporting the CSC hypothesis which suggests that tumors are hierarchically organized and that many tumors, including those of the breast, are maintained by a subpopulation of cells that displays stem cell properties [14,15,16]. These cells may mediate invasion and metastasis and contribute to treatment resistance [17]. miRNAs have also been found to play important roles in normal and malignant stem cell function. Silber et al, reported that mir-124 and mir-137 induced differentiation of neural and glioblastoma stem cells, a state associated with cell cycle arrest [18]. Furthermore, recent studies have shown that the miRNAs Let7 and mir-200c regulate self-renewal of BCSCs [10,19]. Stem cell regulatory genes such as BMI-1 and HMGA2 may mediate this process [10,19]. We have previously demon-

Author Summary

Recent evidence suggests that many cancers, including those of the breast, are maintained by a population of cancer cells that display stem cell properties. These "cancer stem cells" may also contribute to tumor metastasis, treatment resistance, and relapse. Recently, miRNAs (small non-coding RNAs) have been reported to be capable of functioning as oncogenes or tumor suppressors. miRNA93 (mir-93) is frequently overexpressed in human cancer but, paradoxically, has been found to function as a tumor suppressor in some contexts. Using a series of breast cancer cell lines representing different stages of differentiation and mouse xenograft models, we demonstrate that mir-93 modulates the fate of breast cancer stem cells by regulating their proliferation and differentiation states. In less differentiated tumors, enforced expression of mir-93 completely blocks tumor development in mammary fat pads and development of metastases following intracardiac injection in mouse xenografts by reducing breast cancer stem cells. However, the effect of mir-93 on the cancer stem cell population is dependent on the cellular differentiation state, with mir-93 expression increasing the cancer stem cell population in more differentiated breast tumors. These studies demonstrate that miRNAs can regulate breast stem cell proliferation and differentiation, an observation with important biological and clinical implications.

strated that normal breast tissue, primary breast cancers and breast cancer cell lines contain subpopulations with stem cell properties that can be enriched by virtue of their expression of aldehyde dehydrogenase (ALDH) as assessed by the Aldefluor assay (Stem Cell Technologies, Inc., Vancouver, British Columbia) or by tumor initiation in NOD/SCID mice [20]. Recently, Ibarra, et al, showed that Let7, as well as mir-93 are highly depleted in mouse mammary stem/progenitor cells isolated with the stem cell marker ALDH [21]. We have utilized breast cancer cell lines representing different molecular subtypes of breast cancer as well as primary xenografts of breast cancer and normal mammary cells to examine the role of mir-93 in the regulation of normal and malignant breast stem cells. We demonstrate that this miRNA is able to regulate stem cell fate including cellular proliferation and differentiation. These studies suggest that miRNAs regulate the transition between CSC states findings which have important biological and clinical implications.

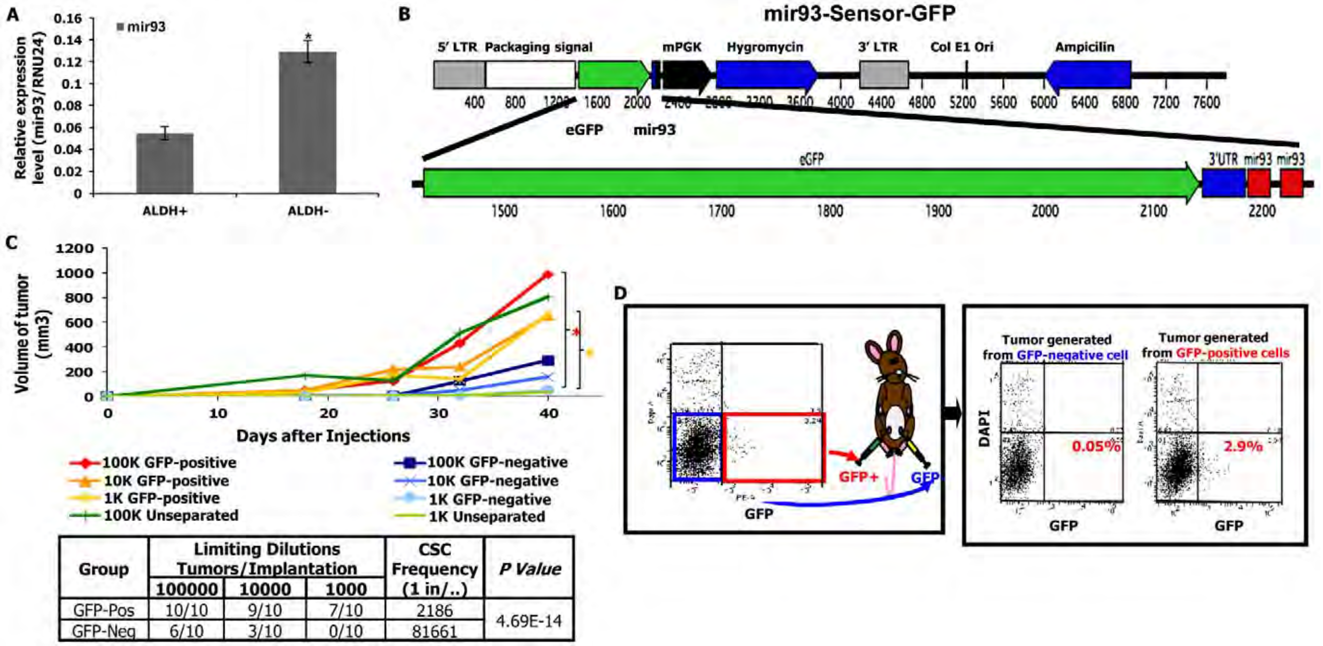
Results

Tumor initiating capacity is associated with low mir-93 expression

We have previously demonstrated that primary human breast cancers and established breast cancer cell lines contain subpopulations with stem cell properties that can be isolated by virtue of their expression of ALDH as assessed by the Aldefluor assay. These cells displayed increased tumor initiating capacity and metastatic potential compared to corresponding Aldefluor-negative cells [22]. mir-93 was shown as one of the most abundant miRNAs in ALDH$^-$ cells [21]. As assessed by qRT-PCR, mir-93 expression was significantly increased in the ALDH$^-$ compared to ALDH$^+$ populations in SUM159 claudin^{low} and HCC1954 basal subtype of human breast cancer (Figure 1A and Figure S1A). As shown in Figure S3, mir-93 expression was lower in CSCs which were characterized by their expression of the CSC markers: ALDH$^+$ or CD24$^-$CD44$^+$. To determine the relationship between mir-93 expression and tumor initiating capacity, we constructed a mir-93 sensor tagged with GFP (mir-93-sensor-GFP) containing a mir-93 target UTR coupled to GFP. In cells transfected with this vector, mir-93 expression results in degradation of GFP mRNA (sensor-negative), whereas mir-93-negative cells express GFP (sensor-positive) (Figure 1B). mir-93 expression was significantly higher in GFP-negative cells than GFP-positive cells (Figure S4) and the ALDH1A1 was much lower in GFP-negative cells than GFP-positive cells as accessed by western blot or immunohistochemical staining (Figure S5). Furthermore, GFP was significantly reduced by overexpression of mir-93 (Figure S6), demonstrating that the sensor reports mir-93 function. The relationship between mir-93 expression and tumor initiation was determined by introducing serial dilutions of sensor-positive (mir-93-negative) and sensor-negative (mir-93-positive) SUM159 cells into the mammary fatpads of NOD/SCID mice. As shown in Figure 1C, sensor-positive (mir-93-negative) cells had significantly higher tumor initiating capacity and CSC frequency than sensor-negative (mir-93-positive) cells. Moreover, mir-93-negative cells gave rise to tumors containing both mir-93-negative and mir-93-positive populations, whereas mir-93-positive cells gave rise only to small, slow growing tumors containing exclusively mir-93-positive populations (Figure 1D). Similar findings were seen using HCC1954 cells (data not shown). These studies demonstrated that in these breast cancer cell lines low mir-93 expression is associated with the CSC phenotype characterized by increased aldehyde dehydrogenase expression, tumor initiating capacity and the ability to generate heterogeneous tumors containing both stem cell and non-stem cell populations.

mir-93 overexpression decreases CSCs *in vitro*

We utilized a tetracycline (TET) inducible mir-93 construct tagged with RFP (pTRIPZ-mir-93-RFP) to determine the functional role of mir-93 in CSCs. mir-93 levels were significantly increased by ten hours following tetracycline induction in these cells (Figure 2A). Induction of mir-93 was associated with a significant decrease in the CSC population as assessed by the Aldefluor assay (Figure 2B), which were also seen in two basal breast cancer cell lines HCC1954 and SUM149 (Figure S1B and Figure S7). Furthermore, this decrease did not result from induction of apoptosis in these cells as assessed by Annexin V staining (Figure 2B). Our group and others have previously shown that CSCs were relatively resistant to cytotoxic chemotherapy. Consistent with this, addition of the cytotoxic agent docetaxel resulted in a relative increase in the percentage of Aldefluor-positive cells (Figure 2B), an increase associated with induction of apoptosis in the bulk cell population (42.8% versus 1.1% control) (Figure 2B). The relative increase in the Aldefluor-positive population seen with docetaxel treatment was abrogated by simultaneous mir-93 expression (Figure 2B). These experiments suggested that unlike cytotoxic agents which primarily target the bulk cell population, mir-93 overexpression was able to reduce the CSC population. Moreover, this did not appear to result from increased CSC apoptosis suggesting a potential role for mir-93 in promoting differentiation of CSCs. Furthermore, since the TET-inducible mir-93 system allows for the controlled regulation of CSC populations, it provides a valuable tool for assessing the role of CSCs in tumor growth in mouse xenograft models. Furthermore, the ability to regulate the CSC population during different phases of tumor growth allows for the assessment of the role of these cells in tumor initiation and maintenance.



Group	Limiting Dilutions Tumors/Implantation			CSC Frequency (1 in/..)	*P Value*
	100000	10000	1000		
GFP-Pos	10/10	9/10	7/10	2186	4.69E-14
GFP-Neg	6/10	3/10	0/10	81661	

Figure 1. mir-93-negative SUM159 cells have increased tumor-initiating capacity. A. $ALDH^+$ cells from SUM159 cells shows lower mir-93 expression level in comparison to $ALDH^-$ cells as accessed by qRT-PCR. $P<0.05$; Error bars represent mean ± STDEV. B. A schematic of mir-93-Sensor-GFP lentiviral construct; C. SUM159 cells were transduced with the mir-93-sensor-GFP lentivirus and selected with hygromycin B, and cells were sorted based on the GFP expression. A serial dilution of mir-93-negative (sensor/GFP-positive) SUM159 cells and mir-93-positive (sensor/GFP-negative) SUM159 cells were injected into the 4th fatpads of NOD/SCID mouse. $^*p<0.05$. D. mir-93-negative cells gave rise to tumors containing both mir-93-negative and mir-93-positive cell populations, but mir-93-positive cells only gave rise to tumors containing mir-93-positive cell populations.
doi:10.1371/journal.pgen.1002751.g001

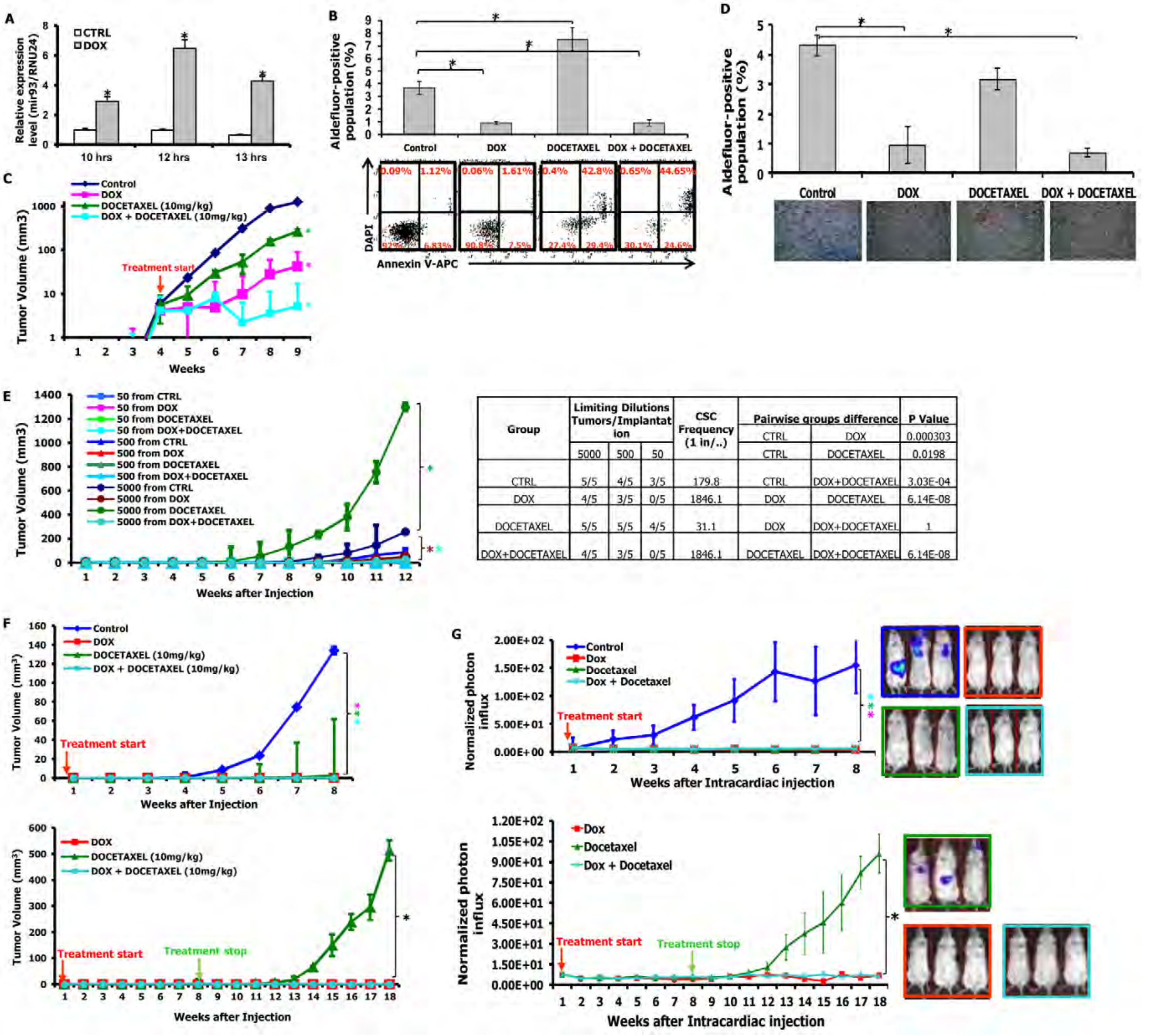
mir-93 overexpression inhibits the growth of established tumor xenografts

We first determined the effect of mir-93 induction on the growth of established tumors and compared these effects to those of cytotoxic chemotherapy. When tumors reached 0.2–0.3 cm in diameter, we induced mir-93 (with doxycycline treatment, hereafter DOX) or initiated cytotoxic chemotherapy with docetaxel or the combination. Induction of mir-93 significantly inhibited the growth of SUM159 and HCC1954 xenografts (Figure 2C and Figure S1C). Furthermore, induction of mir-93 further reduced tumor growth when added to the docetaxel chemotherapy (Figure 2C and Figure S1C). Following five weeks of treatment, animals were sacrificed and CSC populations were assessed by the Aldefluor assay and ALDH1 immunohistochemistry. Induction of mir-93 alone or in combination with docetaxel reduced the Aldefluor-positive population by more than 60% compared to control or docetaxel alone (Figure 2D and Figure S1D). These observations were confirmed by immunohistochemistry of ALDH1 expression (Figure 2D and Figure S1D). mir-93 expression was significantly higher in DOX group compared to the control group at the end of treatment (Figure S2). To provide a more definitive assessment of CSCs, we determined the ability of serial dilutions of cells obtained from primary tumors to form tumors in secondary NOD/SCID mice. Tumor cells isolated from docetaxel treated mice, initiated tumors at lower concentrations with accelerated growth compared to control animals (Figure 2E, Figure S1E). This was consistent with previous studies demonstrating a relative increase in CSCs following chemotherapy [17]. In contrast, cells isolated from tumors with mir-93 induction with or without docetaxel chemotherapy had markedly reduced tumor initiating capacity in secondary mice with no tumors observed from introduction of fifty cells from the mir-93 docetaxel treated group (Figure 2E, Figure S1E). The CSC frequency was lower in the groups of DOX alone and DOX+docetaxel, and was significantly increased in the docetaxel group (Figure 2E and Figure S1E). These studies demonstrated that mir-93 induction reduced the CSC population reducing growth of established tumor xenografts.

In order to determine whether down-regulation of mir-93 promoted tumorigenesis, we utilized a mirZip anti-sense miRNA in SUM159 cells. qRT-PCR was utilized to confirm the efficient knock-down of mir-93 (Figure S8A). $ALDH^+$ cells were significantly increased after mir-93 was knocked down (mirZip93-DsRed) (Figure S8B). As shown in Figure S8C, knockdown of mir-93 significantly promoted the growth of SUM159 cells in tumor xenografts and increased the CSC frequency. Furthermore, the proportion of $ALDH^+$ cells were significantly increased after mir-93 was knocked down (mirZip93-DsRed) (Figure S8D).

mir-93 expression in the adjuvant setting prevents tumor growth

Preclinical models have suggested that CSCs play a role in tumor recurrence and metastasis following adjuvant therapy [23]. This suggests that targeting of CSCs may have more dramatic effects in the adjuvant than in the advanced tumor settings. To simulate the adjuvant setting we induced mir-93 and/or administered docetaxel immediately after tumor cell implantation. Although tumors grew after four to five weeks in control animals, there was no observed tumor growth following mir-93 induction and/or docetaxel treatments for eight weeks (Figure 2F, Figure S1F). After eight weeks, treatments were stopped and animals observed for an additional ten weeks. In SUM159 xenografts, tumors developed in all mice who received eight weeks of docetaxel alone. In contrast, no tumors developed in mice



Group	Limiting Dilutions Tumors/Implantation			CSC Frequency (1 in/..)	Pairwise groups difference		P Value
					CTRL	DOX	0.000303
	5000	500	50		CTRL	DOCETAXEL	0.0198
CTRL	5/5	4/5	3/5	179.8	CTRL	DOX+DOCETAXEL	3.03E-04
DOX	4/5	3/5	0/5	1846.1	DOX	DOCETAXEL	6.14E-08
DOCETAXEL	5/5	5/5	4/5	31.1	DOX	DOX+DOCETAXEL	1
DOX+DOCETAXEL	4/5	3/5	0/5	1846.1	DOCETAXEL	DOX+DOCETAXEL	6.14E-08

Figure 2. mir-93 inhibits tumor growth and metastasis by decreasing CSCs in SUM159 cells. A. SUM159 cells were transduced with the pTRIPZ-mir-93 lentivirus and selected with Puromycin for 7 days. Tetracycline (DOX) induces mir-93 expression in suspension-cultured SUM159 cells by 10 hours; B. 1×10^6 SUM159 cells or pTRIPZ-SUM159 -mir-93 cells were plated in T75 flasks and, after overnight, the cells were treated with Vehicle control, with (DOX) or without (CTRL) DOX (1 ug/ml), docetaxel (10 nM) or the combination for 7 days. Cells were utilized for Aldefluor assay and stained for Annexin V-APC and DAPI for apoptosis assay. C. 100 k pTRIPZ-SUM159-mir-93 cells were injected into the 4th fatpads of NOD/SCID mice. The treatment started as indicated by the red arrow. DOX alone (1 mg/ml in drinking water), or docetaxel (10 mg/kg i.p. once weekly) alone, or the combination inhibits SUM159 tumor growth in vivo (note: The Y-axis is on a logarithmic scale). D. Tumors from each group were collected. ALDH was accessed by the Aldefluor assay on viable dissociated cells and by ALDH1 immunohistochemistry on fixed sections. E. Serial dilutions of cells obtained from these xenografts were implanted in the 4th fatpads of secondary mice, which received no further treatment. F. 10k pTRIPZ-SUM159-mir-93 cells were injected into the 4th fatpads of NOD/SCID mice. The treatment started immediately after injection as indicated by the red arrow and stopped as indicated by the green arrow. G. 200k pTRIPZ-SUM159-mir-93-Luc cells in 100 ul of PBS were injected into the left ventricle of NOD/SCID mice. The treatment started immediately after injection as indicated by the red arrow and stopped as indicated by the green arrow. Metastasis formation was monitored using bioluminescence imaging. Quantification of the normalized photon flux, measured at weekly intervals following inoculation. *$p<0.05$; Error bars represent mean ± STDEV. The colored "*" on the side of the tumor growth curve indicates that the tumor growth or metastasis is significantly different between the control group and the group with the same colored curve.
doi:10.1371/journal.pgen.1002751.g002

following mir-93 induction with or without docetaxel (Figure 2F, Figure S1F). In order to extend these observations to primary breast tumors, we examined the effect of mir-93 induction on three primary breast xenografts, MC1 (Figure S10A), UM2 (Figure S10B) and UM1 (Figure S10C) which were directly established from patient tumors and not passaged *in vitro*. MC1 and UM1 were derived from claudin^{low} and UM2 from a basal breast carcinoma. Induction of mir-93 upon cell implantation completely prevented tumor growth in this model. Together these studies suggested that mir-93 regulated the CSC population and that this population mediates tumor growth following adjuvant therapy.

mir-93 prevents tumor metastasis in the adjuvant setting

Previous studies have demonstrated that CSCs mediate tumor invasion and metastasis. To determine the effect of mir-93 expression on tumor invasion, we examined the effect of mir-93 induction and/or downregulation on invasion of SUM159 cells using a matrigel invasion assay. Overexpression of mir-93 significantly inhibited the ability of SUM159 cells to invade in this assay (Figure S8E). In contrast, knockdown of mir-93 utilizing the mirZip93-DsRed promoted tumor invasion (Figure S8F).

To determine whether the expression of mir-93 affect the growth of tumor metastasis *in vivo*, SUM159 (Figure 2G) and HCC1954 (Figure S1G) cells co-transfected with the inducible mir-93 vector and luciferase were introduced into NOD/SCID mice by intracardiac injection and metastasis formation monitored by bioluminescence imaging. DOX and/or docetaxel treatments were initiated following intracardiac injection. As shown in Figure 2G, mir-93 induction completely suppressed whereas docetaxel partially suppressed metastasis formation. Metastasis was confirmed by histologic examination with pan-cytokeratin staining (Figures S9, S1H). Treatments were stopped at eight weeks and animals were observed for an additional ten weeks for development of metastasis. In animals receiving docetaxel alone, metastasis rapidly developed following cessation of therapy. In contrast, no metastases developed in mice following mir-93 induction with or without docetaxel chemotherapy (Figure 2G). In animals injected with HCC1954 cells, animals from all groups developed metastasis following cessation of therapy. However, development of metastasis were delayed and reduced in mice following mir-93 induction with or without docetaxel chemotherapy (Figure S1G).

mir-93 overexpression increases the CSC population and accelerates tumor growth in luminal subtype MCF7 cells

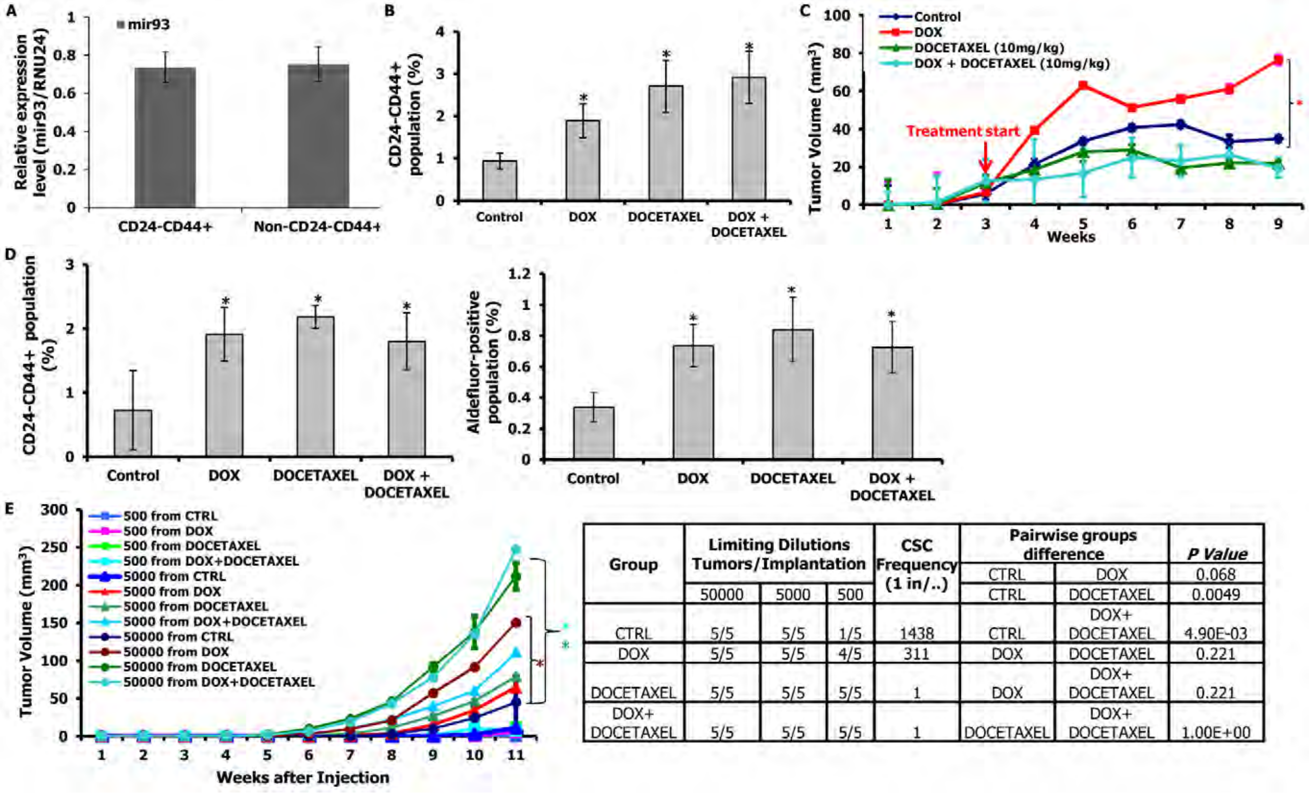
Human breast cancer represents a heterogeneous set of diseases with distinct molecular profiles and clinical behaviors [24]. These subtypes may represent different cells of origin and/or differentiation state. It has been proposed that the most undifferentiated "claudin^{low}" tumors originate from and resemble normal mammary stem cells, whereas the triple-negative basal tumors arise from a more differentiated luminal progenitor cell and the most differentiated luminal tumors which express estrogen and progesterone receptors originate from and are composed of the most differentiated cells [24]. To determine the relationship between mir-93 expression and level of cellular differentiation, we compared the expression of mir-93 in claudin^{low} (SUM159), basal (HCC1954) and luminal (MCF7) cells. As shown in Figure S11, mir-93 levels correlate with postulated differentiation state of these cell lines. Furthermore, in the claudin^{low} SUM159 cells and basal HCC1954 cells, mir-93 expression is significantly lower in Aldefluor-positive as compared to Aldefluor-negative populations (Figure S11). In contrast, the CSC population in MCF7 cells characterized by the phenotype CD24^{-}CD44^{+} [25] expressed the same high level of mir-93 as did the other (non-stem) cells constituting the bulk population (Figure 3A, Figure S11). This suggests that mir-93 may play a different role in more differentiated luminal breast cancer than in the more undifferentiated claudin^{low} and basal subtype. Consistent with this, induction of mir-93 in MCF7 cells increased the CD24^{-}CD44^{+} population (Figure 3B). Docetaxel also increased this population, as did the combination of mir-93 plus docetaxel (Figure 3B). In xenografts, induction of mir-93 accelerated the growth of MCF7 xenografts compared to control (Figure 3C), findings which were confirmed using two additional luminal cell lines MDA-MB-453 and T47D (Figures S12A and S13A). In contrast, docetaxel reduced tumor growth (Figure 3C). Analysis of treated MCF7 tumors confirmed that mir-93 induction increased the proportion of CD24^{-}CD44^{+} cells and ALDH^{+} cells in tumors as did docetaxel or DOX plus docetaxel (Figure 3D). mir-93 expression level was significantly higher in DOX group compared to the control group at the end of treatment (Figure S14). mir-93 induction increased the proportion of ALDH^{+} cells from 1.01% to 9.5% in MDA-MB-453 tumors (Figure S12B) and from 1.26% to 3.84% in T47D tumors (Figure S13B). Furthermore, the calculated tumor initiating frequency was significantly increased after mir-93 induction (Figure 3E, Figures S12C and S13C). These results were confirmed and extended by demonstrating that mir-93 induction in primary tumors increased their tumor-initiating capacity when implanted into secondary recipients (Figure 3E, Figures S12C, S13C). Together, these experiments suggested that the effects of mir-93 on the CSC population differed in different molecular subtypes of breast cancer, an observation consistent with the hypothesis that miRNA effects might be differentiation state dependent.

mir-93 downregulates stem cell regulatory genes in BCSCs

In order to determine the cellular targets of mir-93 in BCSCs, ALDH^{+} and ALDH^{-} populations of SUM159 cells were separated and cultured in suspension in the presence or absence of DOX for ten hours. Gene expression profiles in the four populations were determined utilizing Affymetrix oligonucleotide microarrays (Figure 4A). Of the 2,000 genes downregulated at least two-fold upon DOX treatment in the ALDH^{+} population (Table S1), 127 overlapped with the predicted target sequences of mir-93 including twenty-four genes known to be involved in stem cell regulation (Figure 4A and Table S2) including JAK1, SOX4, STAT3, AKT, E2H1 and HMGAZ. The downregulation of these genes in pTRIPZ-SUM159-mir-93, pTRIPZ-HCC1954-mir-93 cell lines and pTRIPZ-MC1-mir-93 were confirmed with customized PCR array plates (Figures S15, S16, S17). In contrast, only 352 genes were significantly downregulated by DOX in the ALDH^{-} population (Table S3) with twelve of these genes (no stem cell genes) overlapping with the predicted mir-93 targets. These studies suggest that mir-93 regulates the CSC population by simultaneously targeting a number of stem cell regulatory genes. To confirm this, we utilized a luceriferase reporter assay to determine the effect of mir-93 on the expression of the stem cell regulatory genes AKT3, SOX4 and STAT3 selected from the expression profiling data. Expression of mir-93 reduced the level of these stem cell regulatory genes in SUM159 (Figure 4B) and HCC1954 cells (Figure S18) but not in luminal MCF7 and MDA-MB-453 cells (Figure S19). Furthermore, knockdown of STAT3 or SOX4 but not AKT3 decreases the proportion of ALDH^{+} SUM159 cells suggesting these genes play a role in the regulation of CSC self-renewal (Figure S20). The 127 genes in pTRIPZ-MCF7-mir-93 were also tested with customized PCR array plates, and interestingly, most of the stem cell genes were not knocked-down by mir-93 induction in the ALDH^{+} proportion of MCF7 (Figure S21).

mir-93 regulates cell proliferation

To determine the relationship between mir-93 expression and cell cycle kinetics, we assessed mir-93 expression in quiescent and cycling ALDH^{+} and ALDH^{-} populations. Cycling (S/G2/M) cells expressed significantly higher levels of mir-93 compared to



Group	Limiting Dilutions Tumors/Implantation			CSC Frequency (1 in/..)	Pairwise groups difference		*P Value*
					CTRL	DOX	0.068
	50000	5000	500		CTRL	DOCETAXEL	0.0049
CTRL	5/5	5/5	1/5	1438	CTRL	DOX+ DOCETAXEL	4.90E-03
DOX	5/5	5/5	4/5	311	DOX	DOCETAXEL	0.221
DOCETAXEL	5/5	5/5	5/5	1	DOX	DOX+ DOCETAXEL	0.221
DOX+ DOCETAXEL	5/5	5/5	5/5	1	DOCETAXEL	DOX+ DOCETAXEL	1.00E+00

Figure 3. mir-93 promotes tumor growth by increasing CSCs in MCF7 cells. A. Mir-93 is expressed equally in $CD24^{-}CD44^{+}$ and bulk (non-$CD24^{-}CD44^{+}$) populations of MCF7 cells. B. 1×10^{6} pTRIPZ-MCF7-mir-93 cells were plated in T75 flasks and, after overnight, the cells were treated with Vehicle control, DOX (1 ug/ml), docetaxel (10 nM) or the combination for 7 days. DOX alone, docetaxel alone or the combination increased the $CD24^{-}CD44^{+}$ population *in vitro*. C. 1000k pTRIPZ-MCF7-mir-93 cells were injected into the 4th fatpads of NOD/SCID mice. Treatment was initiated as indicated by the red arrow. DOX alone (1 mg/ml in drinking water) promoted MCF7 tumor growth in vivo; docetaxel (10 mg/kg i.p. once weekly) alone or the combination inhibits MCF7 tumor growth in vivo. D. Tumors from each group were collected. Analysis for CD24 and CD44 was performed on dissociated cells. DOX alone, docetaxel alone, or the combination increased the $CD24^{-}CD44^{+}$ populations in MCF7. E. Serial dilutions of cells obtained from these xenografts were implanted in the 4th fatpads of secondary mice, which received no further treatment. Cells from DOX-, docetaxel-, or combination-treated tumors formed secondary tumors at all dilutions (50000, 5000, 500), whereas only higher numbers of cells (50000, 5000) obtained from control xenografts were able to generate tumors. *$p<0.05$; Error bars represent mean ± STDEV. The colored "*" on the side of the tumor growth curve indicates that tumor growth is significantly different between the control group and the group with the same colored curve.
doi:10.1371/journal.pgen.1002751.g003

quiescent (G0/G1) cells in both the $ALDH^{-}$ and $ALDH^{+}$ compartments (Figure 5A). To determine whether mir-93 induces or is a consequence of cellular proliferation, we utilized the DOX inducible mir-93 construct to determine the effect of mir-93 induction on cell cycle distribution. Induction of mir-93 reduced the quiescent cell population from 64% to 42% suggesting that this miRNA has the capacity to directly regulate the cell cycle (Figure 5B). Furthermore, induction of mir-93 increased the proliferation of SUM159 by 29% (Figure S22). Although mir-93 induction had similar effects on the basal HCC1954 cell lines it had no significant effect on the cell cycle of the luminal MCF7 cells (Figure S22, Figure S23).

To determine the relationship between the stem cell phenotype and cell cycle kinetics, we determined the cell cycle distribution of $ALDH^{+}$ and $ALDH^{-}$ populations. The $ALDH^{+}$ population in SUM159 cells had a higher fraction of non-cycling cells compared to $ALDH^{-}$ cells (Figure 5A). This finding was confirmed by Ki67 and MCM7 staining (Figure S24).

mir-93 promotes Mesenchymal-Epithelial Transition (MET) in SUM159 cells

SUM159 cells are derived from a "$\text{claudin}^{\text{low}}$" subtype of breast cancer which is characterized as having a high proportion of cells displaying "epithelial-mesenchymal transition (EMT)". This state is characterized by loss of epithelial characteristics such as apical basal polarity and E-Cadherin expression and acquisition of mesenchymal characteristics, including loss of cell polarity and expression of Vimentin. We determined the effects of mir-93 expression on MET of SUM159 cells by assessing markers of these states at the protein and mRNA levels. SUM159 cells have a mesenchymal morphology and express Vimentin, but not the epithelial marker E-Cadherin, an effect not dependent on cell density (Figure 6A). Expression of mir-93 in these cells caused them to assume a more epithelial appearance associated with a decrease in Vimentin and an increase in E-Cadherin expression (Figure 6A). Similar effects were seen in the basal HCC1954 cell line (Figure S25) although these were less pronounced. To confirm and extend these results we determined the effect of mir-93 expression on mRNA expression of a wider panel of epithelial and mesenchymal markers. We also determined the time course of expression of epithelial and mesenchymal marker mRNAs expressed in $ALDH^{+}$ stem cells and $ALDH^{-}$ non-stem cell populations. Expression of mir-93 in SUM159 cells resulted in a time dependent decrease in expression of mesenchymal markers, Vimentin, N-cadherin and Twist, and an increase in the epithelial markers E-Cadherin and Claudin (Figure 6B). Furthermore,

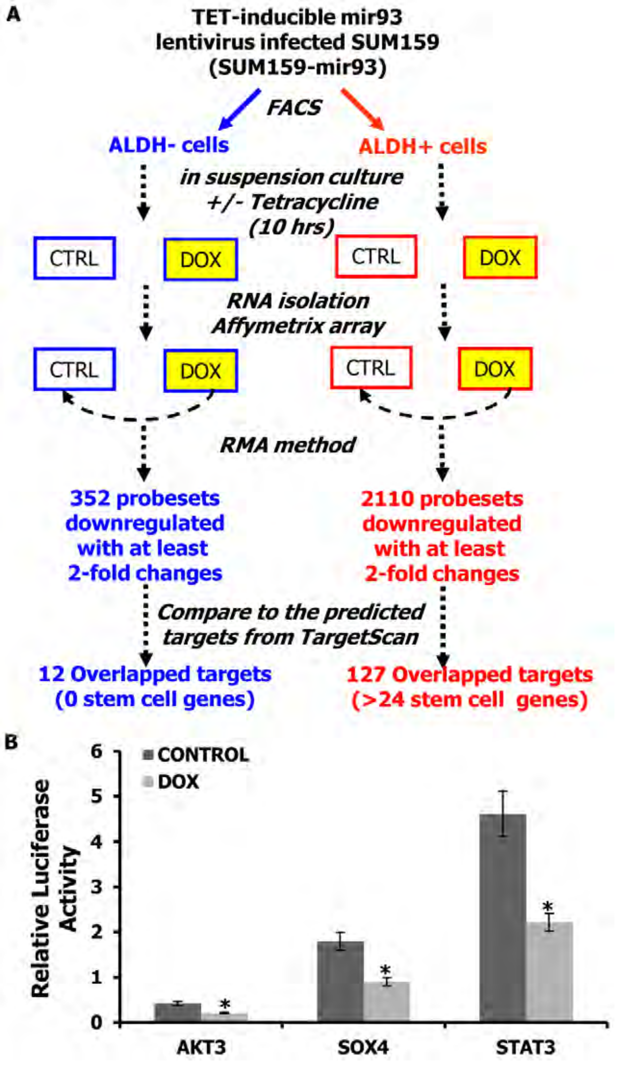

A
TET-inducible mir93
lentivirus infected SUM159
(SUM159-mir93)
FACS
ALDH- cells
ALDH+ cells
in suspension culture
+/- Tetracycline
(10 hrs)
CTRL
DOX
CTRL
DOX
RNA isolation
Affymetrix array
CTRL
DOX
CTRL
DOX
RMA method
352 probesets downregulated with at least 2-fold changes
2110 probesets downregulated with at least 2-fold changes
Compare to the predicted targets from TargetScan
12 Overlapped targets (0 stem cell genes)
127 Overlapped targets (>24 stem cell genes)
B
CONTROL
DOX
Relative Luciferase Activity
0
1
2
3
4
5
6
*
AKT3
SOX4
STAT3

Figure 4. mir-93 targets stem cell regulatory genes. A. Schematic representation of the experimental design to identify the direct targets of mir-93 in SUM159 cells. B. Activity of the luciferase gene linked to the 3′UTR of AKT3, SOX4, or STAT3. The pMIR-REPORT firefly luciferase reporter plasmids with the wild-type 3′UTR sequences of AKT3, SOX4, or STAT3 were transiently transfected into pTRIPZ-mir-93-SUM159 cells and an internal control ACTB luciferase reporter was co-transfected for normalization. The cells were treated with or without DOX. Luciferase activities were measured after 48 hr. The relative luciferase activity was calculated as the ratio of (the results from the cells transfected by individual reporter)/(the results from the cells transfected by the internal control in the same cell group). The data are mean and standard deviation (SD) of separate transfections (n = 4). *p<0.05; Error bars represent mean ± STDEV.
doi:10.1371/journal.pgen.1002751.g004

although these effects were seen in ALDH$^-$ populations, they were even more pronounced in the ALDH$^+$ stem cell compartment. Since TGFβ is a major inducer of the EMT [26,27], we examined the effect of mir-93 expression on components of this pathway. Interestingly, expression of mir-93 significantly reduced expression of the mRNA for TGFβR2 in both ALDH$^+$ and ALDH$^-$ SUM159 cells. This effect was seen as early as twelve hours, suggesting a potential role for down regulation of TGFβ signaling in inducing the MET.

mir-93 maintains normal breast stem cells in an epithelial state

In addition to breast cancer cells, we also determined the effects of mir-93 expression on normal breast cell differentiation. We utilized flow cytometry to access expression of EpCAM and CD49f in breast epithelial cells obtained from reduction mammoplasties. It has previously been shown that mammary stem cells are contained within the EpCAM$^-$CD49f$^+$ population while double positive (EpCAM$^+$CD49f$^+$) cells are luminal progenitors, EpCAM$^+$CD49f$^-$ more differentiated Luminal cells, while EpCAM$^-$CD49f$^-$ constitute stromal cells [28]. We compared mir-93 expression levels in these four populations. Interestingly, we found that the highest level of mir-93 is expressed in the EpCAM$^+$CD49f$^+$ population (Figure 7A), which suggested mir-93 was required to maintain the cells as EpCAM$^+$CD49f$^+$. Furthermore, overexpression of mir-93 in freshly isolated normal breast cells or in immortalized non-transformed MCF-10A cells increased the proportion of cells expressing EpCAM (Figure 7B, 7C). These studies suggested that mir-93 played a role in maintaining normal breast cells in an epithelial (EpCAM$^+$) state.

Discussion

In these studies, we demonstrate that mir-93 is capable of modulating breast CSC populations by regulating their proliferation and differentiation states. To examine this, we utilized breast cancer cell lines representing different states of differentiation. The levels of endogenous mir-93 expression paralleled cellular differentiation states with mir-93 levels lowest in the most primitive "claudin^{low}" SUM159 cells, highest in the "luminal" MCF7 cells and intermediate in the "basal" HCC1915 cells. We utilized a DOX inducible system to determine the effects of enforced mir-93 expression on the CSC populations assessed by expression of the stem cell markers ALDH and CD24$^-$CD44$^+$ as well as by mouse xenograft assays [14,22]. Enforced mir-93 expression in claudin^{low} and basal breast cancer cell lines significantly reduced the CSC populations as assessed by the Aldefluor assay. To assess the functional relevance of this, we determined the effect of mir-93 induction in SUM159 and HCC1954 cells on tumor growth in NOD/SCID mouse xenografts. The effects of mir-93 expression on tumor initiating capacity was confirmed using two primary breast xenografts generated without *in vitro* culture. mir-93 expression decreased the CSC in these claudin^{low} primary xenografts. In contrast, overexpresson of mir-93 in the luminal MCF7 cells line resulted in an increase in CD24$^-$CD44$^+$ CSC resulting in increased tumor growth. This demonstrates that the effect of mir-93 on CSC populations is dependent on the cellular differentiation state. This model allowed us to simulate potential clinical scenarios involving CSC targeting agents. To simulate the effects of CSC targeting agents in advanced disease, tumors were inoculated into mammary fatpads and when the tumors were palpable mir-93 was induced by addition of doxycycline to the mouse drinking water. In this setting, mir-93 induction had only a modest effect in reducing tumor growth. Addition of the chemotherapeutic agent docetaxel resulted in a more significant reduction in tumor size, an effect that was accentuated by mir-93 induction. CSC models predict that the efficacy of CSC targeting agents should be most pronounced in the adjuvant setting where tumor growth from micrometastasis is dependent on stem cell self-renewal [29]. Consistent with this model, induction of mir-93 immediately after fatpad implantation or after development of micrometastasis by intracardiac injection completely blocked tumor recurrence. Furthermore, when treatment was discontinued at eight weeks, animals that received chemotherapy alone

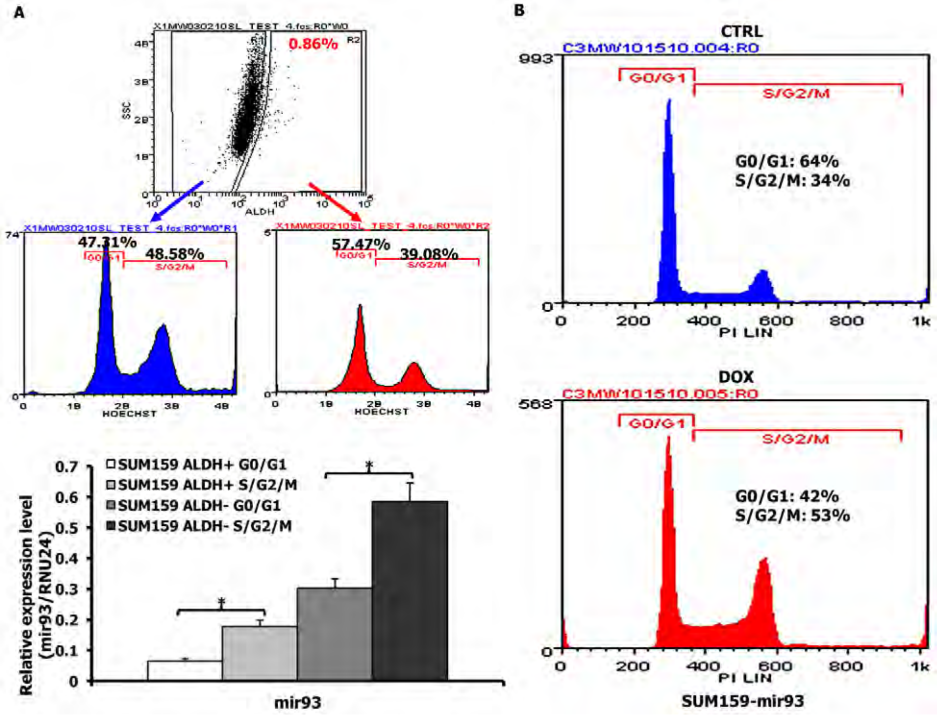


Figure 5. mir-93 regulates the cell cycle in SUM159 cells. A. SUM159 cells were stained with Aldefluor and Hoechst33342 and dead cells excluded by 7-AAD staining. Cells from G0/G1 and S/G2/M were sorted from ALDH^{+} or ALDH^{-} populations and mir-93 expression was measured with qRT-PCR. B. Cell cycle analysis of pTRIPZ-SUM159-mir-93 cells in the presence or absence of DOX for 7 days. Propidium iodide staining followed by flow cytometry was used to analyze cell cycle distribution. mir-93 induction with DOX resulted in a decreased proportion of cells in G0/G1 and an increased proportion of cells in S/G2/M. *$p<0.05$; Error bars represent mean ± STDEV.
doi:10.1371/journal.pgen.1002751.g005

developed local tumor growth and metastasis while those with mir-93 induction with or without chemotherapy showed no recurrence when animals were sacrificed after four months. These studies provide strong support for the CSC hypothesis and provide a valuable animal model for clinical trial design using CSC targeting agents.

To determine the molecular mechanisms of mir-93 CSC regulation, we employed an unbiased approach assessing the effect of mir-93 expression on early changes in global gene expression profile coupled with prediction of miRNA target sequences. Interestingly, this analysis revealed that twenty-four genes known to be involved in stem cell self-renewal including JAK1, SOX4, STAT3, AKT, EZH1, HMGA2 are targeted by mir-93. In addition, this miRNA targets two important regulators of TGFβ signaling, TGFβR2 and SMAD5.

mir-93 expression was also associated with and in turn regulates cellular proliferation. Quiescent G0/G1 cells expressed lower levels of mir-93 than proliferating cells in S/G2/M phase. Furthermore, enforced expression of mir-93 increased the fraction of cycling cells.

We demonstrate that induction of mir-93 in mesenchymal-like SUM159 cells induces an MET in the ALDH^{+} CSC population characterized by increased expression of E-Cadherin and Claudin, with concomitant downreguation of mesenchymal genes, such as Vimentin, N-Cadherin and Twist. mir-93 also inhibits TGFβ signaling by targeting TGFβR2, an effect seen within twelve hours of mir-93 induction. This was followed by an MET in the Aldefluor-positive CSC population. Since TGFβ is a major regulator of EMT, abrogation of this signaling pathway may facilitate MET. Of interest, it has been recently reported that the mir-106b-25 cluster including mir-93 is induced in the early stages of nuclear reprogramming of fibroblasts into IPS cells [30]. This is accompanied by a mesenchymal to epithelial conversion in these cells which is obligatory for reprogramming to recur. This suggests that this miRNA cluster may regulate MET in multiple biological contexts.

In summary, our experiments suggest that CSCs can exist in two alternative epithelial and mesenchymal states, the balance of which is regulated by miRNAs including mir-93 (Figure 8). The mesenchymal state associated with an invasive phenotype characterized by quiescence and low mir-93 expression is maintained by growth factors such as TGFβ. Upon activation of cellular proliferation, MYC and E2F are induced leading to expression of MCM7, a licensing factor required for DNA synthesis. Concomitantly, mir-93 and its related miRNA cluster is co-synthesized which promotes further proliferation while simultaneously downregulating TGFβ signaling. This facilitates a mesenchymal to epithelial transition in the CSC population characterized by decreased invasiveness and increased proliferation. Continued expression of mir-93 simultaneously downregulates a number of stem cell self-renewal pathways including JAK/STAT, AKT, EZH1 and HMGH2, promoting cellular differentiation and depleting the CSC population. The model depicted in Figure 8 is consistent with our observation that mir-93 level is highest in the EpCAM^{+}CD49f^{+} normal mammary cells and decreased with terminal differentiation. In contrast, the effects of mir-93 depend on the cellular differentiation state accounting for differences we observed in claudin^{low}, basal and luminal breast cancers, with mir-93 level highest in the luminal MCF7 cell line compared to basal HCC1954 and claudin^{low} SUM159 cell lines. MCF7 cells are highly proliferative although unlike normal mammary cells incapable of terminal differentiation (Figure 8).

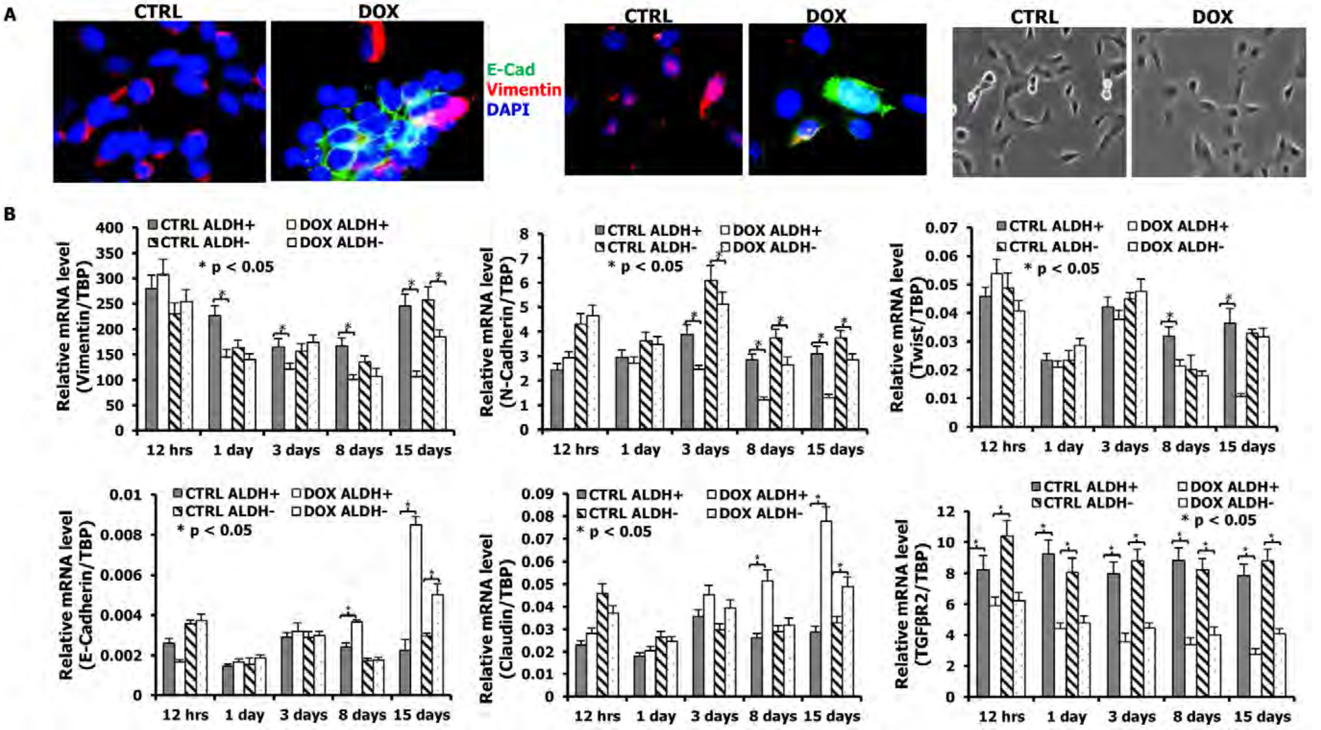


Figure 6. mir-93 initiates MET in SUM159 cells. A. pTRIPZ-SUM159-mir-93 cells were plated in 2-well chamber slides with (DOX) or without (CTRL) Doxycycline for 7 days. E-Cadherin and Vimentin were deleted by immunofluorescence staining. Expression of mir-93 in SUM159 cells causes them to assume a more epithelial appearance associated with a decrease in Vimentin and an increase in membrane localized E-Cadherin expression. The phase micrographs for CTRL and DOX are also shown. E-Cadherin, Green; Vimentin, Red; DAPI, Blue. A representative sample from 3 independent samples is shown. B. The effect of mir-93 expression on a panel of epithelial and mesenchymal markers at the mRNA level as accessed by qPCR. pTRIPZ-SUM159-mir-93 cells were plated with or without DOX, and $ALDH^+$ and $ALDH^-$ cells were sorted at different times (12 hours, 1 day, 3 days, 8 days, 15 days) by Aldefluor assay. qRT-PCR was utilized to access the effects of mir-93 on mRNA expression of mesenchymal markers (Vimentin, N-Cadherin and Twist), epithelial markers (E-Cadherin and Claudin), and TGFβR2. *$p<0.05$; Error bars represent mean ± STDEV.
doi:10.1371/journal.pgen.1002751.g006

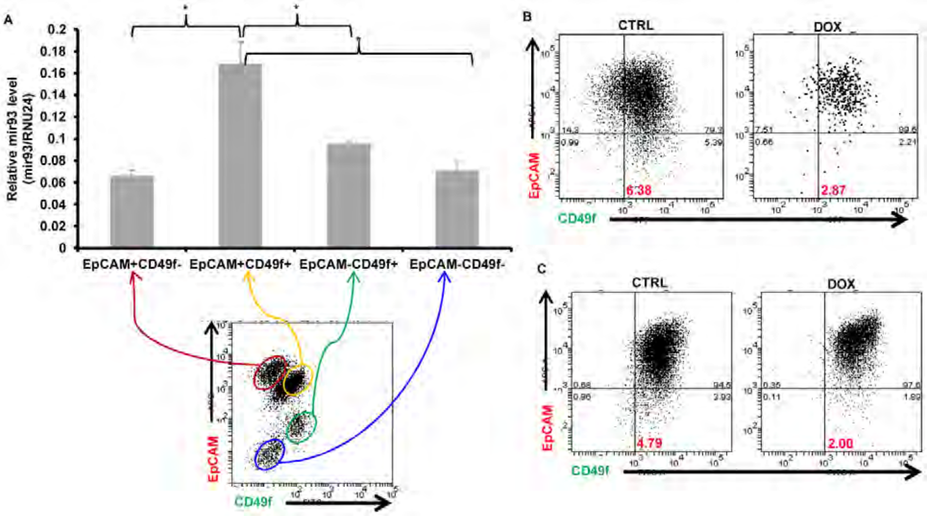


Figure 7. mir93 promotes MET in normal breast epithelial cells. A. Single cells were isolated from normal breast tissues and stained with EpCAM-APC and CD49f-FITC for FACS sorting. After sorting, total RNA were isolated from different sorted groups ($EpCAM^+CD49f^-$, $EpCAM^+CD49f^+$, $EpCAM^-CD49f^+$, $EpCAM^-CD49f^-$) and mir-93 expression were measured by qRT-PCR. $p<0.05$; Error bars represent mean ± STDEV. B. Single cells were isolated from normal breast tissues and infected with mir-93-expressing lentiviruses in suspension. After one week, mammospheres were dissociated into single cells and plated in adherent culture in the absence (CTRL) or presence (DOX) of DOX for two weeks. Then, cells were dissociated and stained with EpCAM-APC and CD49f-FITC for FACS analysis. C. MCF10A cells were cultured in the absence (CTRL) or presence (DOX) of DOX for two weeks. Cells were then dissociated and stained with EpCAM-APC and CD49f-FITC for FACS analysis.
doi:10.1371/journal.pgen.1002751.g007

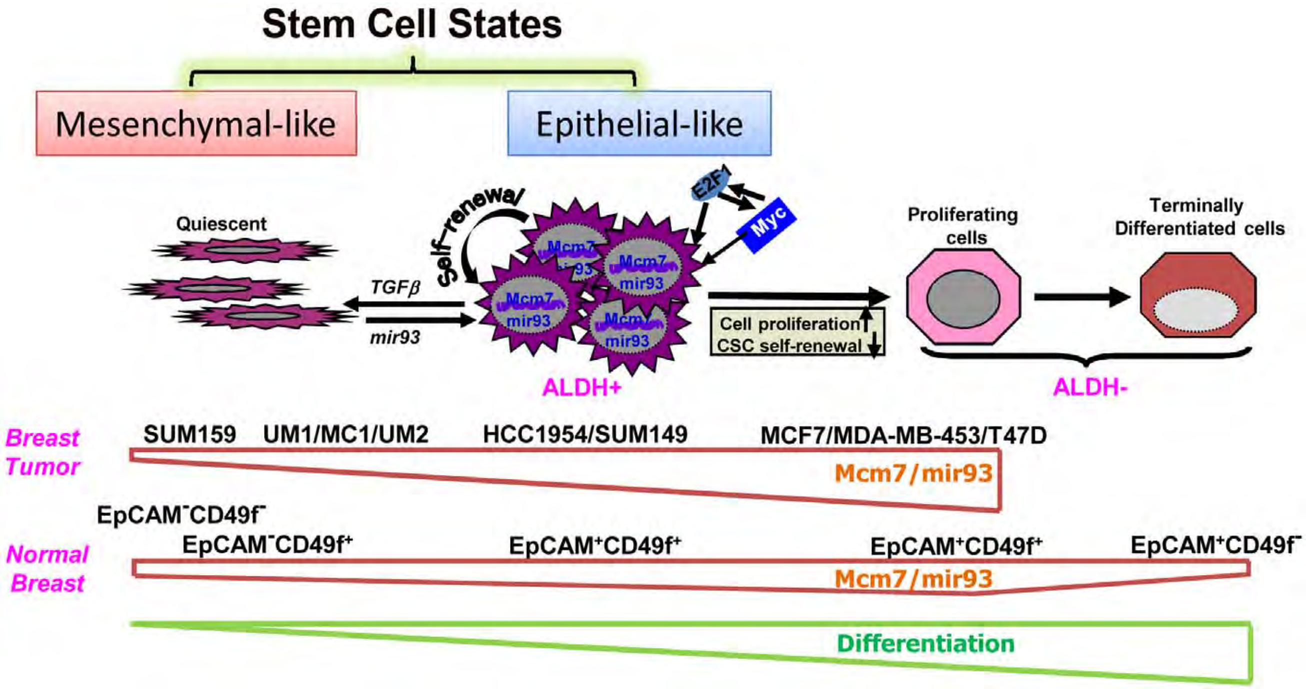


Figure 8. A hypothetic model illustrating regulation of normal and malignant mammary stem cell states and fates by mir-93.
doi:10.1371/journal.pgen.1002751.g008

The existence of alternative CSC states, associated with expression of different protein markers has important implications for understanding the plasticity of CSCs. For example, it has been claimed that CSCs may be generated from non-CSC tumor populations through induction of EMT [31]. However, the existence of alternative CSC state suggests that the acquisition of stem cell markers may reflect transition of CSC states rather than generation of CSCs from non-CSC populations. In addition, the existence of multiple stem cell states suggests the necessity of developing of therapeutic strategies capable of effectively targeting CSCs in all of these states.

Materials and Methods

Cell culture

Breast cancer cell line SUM159 and SUM149 have been extensively characterized (http://www.asterand.com) [32]. HCC1954, MCF-7, MDA-MB-453 and MCF10A were purchased from ATCC. The cell lines were grown using the recommended culture conditions. Briefly, the culture medium for SUM159 and SUM149 is Ham's F-12 (Invitrogen) supplemented with 5% FBS, 5 ug/mL insulin, and 1 ug/mL hydrocortisone (both from Sigma, St. Louis, MO). MCF7, MDA-MB-453 and HCC1954 cells were maintained in RPMI1640 medium (Invitrogen, Carlsbad, CA) supplemented with 10% fetal bovine serum (ThermoFisher Scientific, Pittsburgh, PA), 1% antibiotic-antimycotic (Invitrogen, Carlsbad, CA), and 5 µg/ml insulin (Sigma-Aldrich, St Louis, MO).

Dissociation of mammary tissue

100–200 g of normal breast tissue from reduction mammoplasties was minced with scalpels, dissociated enzymatically, and single cells were isolated as described previously [33,34]. The single cells were utilized for FACS sorting or were cultured in suspension as described previously [33,34]. Mammospheres were dissociated into single cells enzymatically and mechanically, and then cultured in regular cell culture plates [33,34].

Constructs and virus infection

For construction of the mir-93 sensor, miRNA-complementary oligonucleotides were annealed and cloned into a Marx vector that directs GFP expression. The mir-93 miRNA target sequence was engineered into the 3′ untranslated region (UTR) of the cDNA encoding for the GFP fluorescent protein. Expression of this construct in cells that express the mir-93 miRNA results in a RNAi pathway-dependent degradation of GFP mRNA and thus no green fluorescence. In contrast, in cells with repressed mir-93 miRNA, the GFP mRNA is not degraded and resulting in expression of the fluorescent GFP. shRNA oligos for STAT3, AKT3 or SOX4 were inserted to PlentiLox3.7-DsRed lentiviral vector. A highly efficient lentiviral expression system (TRIPZ lentivral vector;www.openbiosystems.com/RNAi) was used to generate mir-93-expressing lentiviruses; and mirZIP-lentivector (SBI, Mountain View, CA) was used to generate mir-93-knockdown lentiviruses in UM Vector Core Facility. The cell lines were infected with the lentiviruses as described previously [34].

Aldefluor assay and flow cytometry

The Aldefluor kit (StemCell Technologies, Inc, Vancouver, BC, Canada) was used to isolate cells with high ALDH enzymatic activity as illustrated in the manufacturer's instructions. Briefly, single cells were suspended in buffer containing ALDH substrate – BAAA (1 µmol/l per 1×10^6 cells) and incubated at 37°C for 40 minutes. In each experiment, the specific ALDH inhibitor diethylaminobenzaldehyde (DEAB) was used as negative control at 50 mmol/L. A FACStarPLUS (Becton Dickinson) was used for FACS. Aldefluor fluorescence was excited at 488 nm and fluorescence emission was detected using a standard fluorescein isothiocyanate (FITC) 530/30 band pass filter. The sorting gates were established based on negative controls. CD44/CD24 staining was performed as previously described [14]. Briefly, cells were stained with primary antibodies anti-CD44 labeled APC (dilution 1:10, BD Pharmingen), and anti-CD24 labeled FITC (dilution 1:10, BD Pharmingen). In all *in vivo* experiments, mouse cells were eliminated by excluding H2Kd$^+$ (mouse histocompatibility class I, BD Pharmagen) cells during flow cytometry. 0.5 µg/ml 4′,6-diamidino-2-phenylindole (DAPI) (Sigma) was used to access cell viability.

Cell cycle analysis

Cells (1×10^6) were harvested and washed in cold PBS followed by fixation in 70% alcohol for thirty minutes on ice. After washing in cold PBS three times, cells were resuspended in 0.8 mL of PBS solution with 40 µg of propidium iodide and 0.1 µg of RNase A for thirty minutes at 37°C. Samples were analyzed for DNA content using aFACSCalibur cytometer (Becton Dickinson, San Jose, CA).

MTT assay

The effect of mir-93 on cell proliferation was measured using an MTT assay. Briefly, 200–500 cells from Control and DOX-treated groups were seeded in 96-well culture plates and were cultured in the absence (CTRL) and presence (DOX) of DOX for 7 days. Subsequently, 0.025 ml of MTT solution (5 mg/ml) was added to each well, and the cells were incubated for 2 h. After centrifugation, the supernatant was removed from each well. The colored formazan crystal produced from MTT was dissolved in 0.15 ml of isopropanol with 4 mM HCl and 0.1% NP40, and the optical density (OD) value was measured at 590 nm.

RNA extraction

Total RNA was isolated using RNeasy Micro Kit (Qiagen, Valencia, CA), and total RNA with enriched miRNA was isolated using miRNeasy mini Kit, according to the manufacturer's instructions.

Gene expression profiling with DNA microarrays

Gene expression analyses used Affymetrix U133 Plus 2.0 human oligonucleotide microarrays containing over 47,000 transcripts and variants including 38,500 well-characterized human genes. Preparation of cRNA, hybridizations, washes and detection were done as recommended by the supplier (http://www.affymetrix.com/index.affx). Expression data were analyzed by the Robust-Multichip Average method in R using Bioconductor and associated packages [35].

Real-time quantitative PCR (qRT–PCR)

MiRNA expression level was measured utilizing TaqMan qRT-PCR (Applied Biosystems, Carlsbad, CA). Single-stranded cDNA was synthesized from 10 ng of miRNA enriched total RNA using specific miRNA primers (TaqMan MiRNA Assay, PN 4427975, Applied Biosystems) and the TaqMan MiRNA Reverse Transcription Kit (PN 4366596, Applied Biosystems). Two ul of cDNA was used as a template in a 20 ul PCR reaction. PCR products were amplified using specific primers (TaqMan MiRNA Assay) and the Taq-Man Universal PCR Master Mix (PN 4324018, Applied Biosystems), and PCR was performed in a ABI PRISM 7900HT sequence detection system with 384-Well block module and automation accessory (Applied Biosystems) by incubation at 50°C for two min and then 95°C for ten min followed by forty amplification cycles (fifteen seconds of denaturation at 95°C and one min of hybridization and elongation at 60°C). PCR reactions for each sample were run in triplicate. The number of cycles required for amplification to reach the threshold limit, the Ct-value was used for quantification. RNU24 was used as an endogenous control for miRNA data normalization, and TBP was used as an endogenous control for other gene normalization. All TaqMan miRNA assays used in this study were obtained from Applied Biosystems.

Tumorigenicity in NOD/SCID mice

All mice were housed in the AAALAC-accredited specific pathogen-free rodent facilities at the University of Michigan. Mice were housed on sterilized, ventilated racks and supplied with commercial chow and sterile water both previously autoclaved. All experimentation involving live mice were conducted in accordance with standard operating procedures approved by the University Committee on the Use and Care of Animals at the University of Michigan. Six-week old female NOD/SCID mice were purchased from Jackson Laboratories (Bar Harbor, ME) and housed in SPF microisolator cages in the animal facility of University of Michigan. Tumorigenicity of 10,000 (Adjuvant setting) cells or 100,000 (Advanced setting) cells in the mamary fatpads of NOD/SCID mice was accessed. Six mice were included in each cohort. The animals were euthanized when the tumors were 1.0–1.5 cm in diameter, in compliance with regulations for use of vertebrate animal in research. A portion of each fat pad was fixed in formalin and embedded in paraffin for histological analysis. Another portion was analyzed by the ALDH or CD24/CD44 cytometric staining.

Primary xenografts

Human breast tumors were obtained as biopsy cores or pieces of tumors after surgery and implanted in humanized cleared fat pads of NOD/SCID mice for establishing xenotransplants. The success of xenotransplantation was approximately 20%, similar to previous reports in the literature. Three xenotransplants were used: an ER$^-$PR$^-$ERBB2$^-$ tumor at the 20th passage in animals (MC1), an ER$^-$PR$^-$ERBB2$^-$ tumor at the 5th passage (UM1), and an ER$^+$PR$^+$ERBB2$^-$ tumor at the 8th passage (UM2).

Intra-cardiac injection

All procedures were approved by the University Committee for the Use and Care of Animals (UCUCA) of the University of Michigan. The intracardiac injection was carried out according to previously published methods [36]. Briefly, six-week-old NOD/SCID mice were anesthetized with isoflurane gas (a 2% isofluorane/air mixture) and injected in the left ventricle of the heart with 100,000 cells in 100 µl of sterile Dulbecco's PBS lacking Ca^{2+} and Mg^{2+}. For each of the cell lines (SUM159-luc, HCC1954-luc), five animals were injected.

Bioluminescence imaging

Baseline bioluminescence was assessed before inoculation and each week thereafter. Mice were anesthetized with isoflurane gas and given a single i.p. dose of 150 mg/kg D-luciferin (Promega) in PBS. For photon flux counting, we used a charge-coupled device camera system (Xenogen) with a nose-cone isofluorane delivery system and heated stage for maintaining body temperature. Results were analyzed after six min of exposure using Living Image software provided with the Xenogen IVIS imaging system.

Immunostaining

For ALDH1 staining, paraffin-embedded sections of breast tumors from xenografts were deparaffinized in xylene and rehydrated in graded alcohol. Antigen enhancement was done by incubating the sections in citrate buffer pH 6.0 (Dakocytomation, Copenhagen, Denmark) as recommended. Slides were stained using Peroxidase histostain-Plus Kit (Zymed) according to the manufacturer's protocol. ALDH1 antibody (BD biosciences) was used at a 1:50 dilution. AEC (Zymed) was used as substrate for peroxidase. Slides were counter-stained with hematoxylin and coverslipped using glycerin. For E-Cadehrin, Vimentin, MCM7, Ki67 and DAPI fluorescent staining, cells were fixed in ice-cold methanol and permeablized with 0.15% triton X-100. E-Cadherin antibody (Santa Cruz, 1:100 dilution), Vimentin antibody (Santa Cruz, 1:200 dilution), MCM7 antibody (Cell signaling, 1:100 dilution), p21 antibody (Cell Signaling, 1:400 dilution) and Ki67 antibody (Dako, 1:150 dilution) were used and incubated for 1 hour at room temperature. PE and FITC labeled secondary antibodies (Jackson Labs) were used at the dilution 1:200 and incubated for twenty min. Nuclei were counterstained with DAPI/antifade (INVITROGEN) and cover slipped. Sections were examined with a Leica fluorescent microscope.

3′ UTR luciferase reporter assay

The pMIR-REPORT luciferase reporter plasimds with the 3′ UTR sequence of AKT3, SOX4, STAT3 or the control ACTB were transfected into the cell lines using Fugene HD tansfection reagent (Roche Applied Science) according to the manufacturer's instruction. After transfection, cells were dissociated and cultured with or without DOX. Luciferase activity was assayed by luciferase assay kit (Promega). Luciferase activities were measured after forty-eight hrs utilizing a luminometer. The results were presented as the luciferase activity of cells transfected with 3′ UTR sequence of AKT3, SOX4, or STAT3 normalized to cells transfected with the luciferase activity of cells transfected with 3′ UTR sequence of ACTB.

Statistical analysis

Results are presented as the mean ± standard deviation (STDEV) for at least three repeated individual experiments for each group using Microsoft Excel. Statistical differences were determined by using ANOVA and student's t-test for independent samples. A p-value of less than 0.05 was considered statistically significant.

Supporting Information

Figure S1 mir93 inhibits tumor growth and metastasis by decreasing CSCs in HCC1954 cells. A. ALDH-positive cells from HCC1954 cells shows lower mir93 expression level in comparison to ALDH-negative cells by qRT-PCR. B. 1×10^6 pTRIPZ-HCC1954-mir93 cells were plated in T75 flasks and, after overnight, the cells were treated with Vehicle control, DOX (1 ug/ml), docetaxel (10 nM) or the combination for 3–7days. Cells were utilized for Aldefluor assay and stained for Annexin V-APC and DAPI for apoptosis assay. C. 100k pTRIPZ-HCC1954-mir93 cells were injected into the 4th fatpads of NOD/SCID mice. The treatment started as indicated by the red arrow. DOX alone (1 mg/ml in drinking water), or docetaxel (10 mg/kg i.p. once weekly) alone, or the combination inhibits SUM159 tumor growth in vivo. D. Tumors from each group were collected. ALDH was accessed by the Aldefluor assay on viable dissociated cells and by ALDH1 immunohistochemistry on fixed sections. E. Serial dilutions of cells obtained from these xenografts were implanted in the 4th fatpads of secondary mice, which received no further treatment. F. 10k pTRIPZ-HCC1954-mir93 cells were injected into the 4th fatpads of NOD/SCID mice. The treatment started immediately after injection as indicated by the red arrow and stopped as indicated by the green arrow. G. 200k pTRIPZ-HCC1954-mir93-luc cells in 100 ul of PBS were injected into the left ventricle of NOD/SCID mice. The treatment started immediately after injection as indicated by the red arrow and stopped as indicated by the green arrow. Metastasis formation was monitored using bioluminescence imaging. Quantification of the normalized photon flux, measured at weekly intervals following inoculation. H. Histologic confirmation, by H&E staining, of metastasis in soft tissues resulting from mice with different treatments in G. *$p<0.05$; Error bars represent mean ± STDEV. The colored "*" on the side of the tumor growth or metastasis curve represents the tumor growth is significantly different between Control group and the group with the same colored curve.
(PDF)

Figure S2 mir-93 is induced in primary tumors with DOX. 100k pTRIPZ-SUM159-mir-93 cells were injected into the 4th fatpads of NOD/SCID mice. Different treatments were initiated (Vehicle Control (Control), DOX alone (DOX), docetaxel alone, or the combination). At the end of treatment, cells from Control and DOX groups were isolated from the tumors and mir-93 expression level was measured by qRT-PCR.
(PDF)

Figure S3 Induction of mir93 decreases both ALDH$^+$ and CD24$^-$CD44$^+$ cells in the primary breast tumor xenografts. Cells isolated from primary human breast xenografts UM2 (A), MC1 (B) and UM1 (C) were sorted for ALDH$^+$ and ALDH$^-$ or CD24$^-$CD44$^+$ and the rest (CD24$^-$CD44$^-$, CD24$^+$CD44$^+$, CD24$^+$CD44$^-$). RNA was isolated from each group of sorted cells and the expression level of mir-93 or RNU24 level was measured by qRT-PCR. *$p<0.05$; Error bars represent mean ± STDEV.
(PDF)

Figure S4 mir-93 expression in Sensor-GFP-positive and Sensor-GFP-negative SUM159 cells. Mir-93-sensor-GFP SUM159 cells were sorted for GFP-positive and GFP-negative cells by flow cytometry and RNA was isolated from both group of sorted cells and the expression level of mir-93 or RNU24 level was measured by qRT-PCR. *$p<0.05$; Error bars represent mean ± STDEV.
(PDF)

Figure S5 ALDH1A1 protein level in Sensor-GFP-positive and Sensor-GFP-negative SUM159 cells. mir-93-sensor-GFP SUM159 cells were sorted for GFP-positive and GFP-negative cells by flow cytometry. A portion of cells were utilized for western blot, and some cells were cytospun down and stained with ALDH1A1 by immunohistochemical staining. *$p<0.05$; Error bars represent mean ±STDEV.
(PDF)

Figure S6 mir-93 induction reduces GFP in mir93-sensor-GFP-SUM159 cells. mir93-sensor-GFP-SUM159 cells were infected with pTRIPZ-mir93 lentivirus and grow in T75 flasks. DOX were added to the culture medium in the DOX group. Images were taken with fluorescence microscope.
(PDF)

Figure S7 1×10^6 pTRIPZ-SUM149-mir93 cells were plated in T75 flasks and, after overnight, the cells were treated with Vehicle control or DOX (1 ug/ml) for 3–7 days. Induction of mir93 expression by DOX decreased the ALDH-positive population. *$p<0.05$; Error bars represent mean ± STDEV.
(PDF)

Figure S8 Modulation of mir93 level in SUM159 cells altered cell invasion in vitro and knockdown increases tumor growth in vivo. SUM159 cells were infected with no virus (non-infection), DsRed control lentivirus (SUM159-Neg-DsRed) or mirZip antisense mir93 lentivirus (SUM159-mirZip93-DsRed). A. microRNA RT-PCR demonstrated SUM159-mirZip93-DsRed reduced mir93 expression by more than 90%. B. SUM159-non-infection, SUM159-Neg-DsRed, SUM159-mirZip93-DsRed cells were grown in T75 flasks and Aldefluor assay was utilized to measure the percentage of $ALDH^+$ cells. C. Serial dilutions of cells were implanted in the 4th fatpads of NOD/SCID mice. SUM159-mirZip93-DsRed cells initiated tumors sooner and accelerated growth compared to equivalent number of control cells. D. Cells were isolated from the tumors in C and Aldefluor assay was utilized to measure the percentage of $ALDH^+$ cells. E. Invasive capacity was assessed by a Matrigel invasion assay using serum as attractant. pTRIPZ-SUM159-mir93 cells are less invasive than the control cells in vitro accesses at 27 hours. F. The invasion was assessed by a Matrigel invasion assay using serum as attractant. SUM159-mirZip93-DsRed cells are more invasive than the control cells in vitro accesses at 27 hours. *$p<0.05$; Error bars represent mean ± STDEV. The colored "*" on the side of the tumor growth curve represents the tumor growth is significantly different between Control group and the group with the same colored curve.
(PDF)

Figure S9 mir-93 inhibits tumor metastasis in SUM159 cells. 200k pTRIPZ-SUM159-mir-93-Luc cells in 100 ul of PBS were injected into the left ventricle of NOD/SCID mice. Different treatments were initiated (Vehicle Control (Control), DOX alone (DOX), docetaxel alone, or the combination). At the end of treatment, H&E staining and Pan-cytokeratin (AE1/AE3) staining (in Brown) were performed to confirm the metastasis in bone and soft tissues resulting from mice with different treatments.
(PDF)

Figure S10 mir93 inhibits tumor growth in primary human breast xenografts MC1, UM2, and UM1. Cells isolated from primary xenografts MC1 (A) or UM2 (B) or UM1 (C) were transduced with the pTRIPZ-mir93 lentivirus in suspension. 10k pTRIPZ-MC1-mir93 or pTRIPZ-UM2-mir93 cells were injected into the 4th fatpads of NOD/SCID mice. The treatment started right after injection as indicated by the red arrow. DOX alone, docetaxel alone or the combination prevented tumor growth. *$p<0.05$; Error bars represent mean ± STDEV. The colored "*" on the side of the tumor growth curve represents the tumor growth is significantly different between Control group and the group with the same colored curve.
(PDF)

Figure S11 Endogenous mir93 expression levels parallel cell differentiation state. $ALDH^+$ population and $ALDH^-$ population were separated from SUM159 cells and HCC1954 cells. $CD24^-CD44^+$ population and the remaining cell populations were separated from MCF7 cells. mir93 level was analyzed by microRNA qRT-PCR. Among these three cell lines, mir93 expression level is highest in MCF7 cells and lowest in the SM159 cells. In both SUM159 cells and HCC1954 cells, $ALDH^+$ cells have lower mir93 expression compared to $ALDH^-$ cells. In contrast, $CD24^-CD44^+$ in MCF7 cells showed no difference for mir93 expression level in comparison to the bulk population. *$p<0.05$; Error bars represent mean ± STDEV.
(PDF)

Figure S12 mir93 promotes tumor growth by increasing CSCs in MDA-MB-453 cells. A. 200k pTRIPZ-MDA-MB-453-mir93 cells were injected into the 4th fatpads of NOD/SCID mice. Treatment was initiated as indicated by the red arrow. DOX (1 mg/ml in drinking water) promoted MDA-MB-453 tumor growth in vivo. B. Tumors from each group were collected. Aldefluor assay was performed on dissociated cells. DOX increased the $ALDH^+$ populations in MDA-MB-453. C. Serial dilutions of cells obtained from these xenografts were implanted in the 4th fatpads of secondary mice, which received no further treatment. Cells from DOX-treated tumors formed secondary tumors at all dilutions (1k, 10k, 32k), whereas only higher numbers of cells (32k) obtained from control xenografts were able to generate tumors. *$p<0.05$; Error bars represent mean ± STDEV.
(PDF)

Figure S13 mir93 promotes tumor growth by increasing CSCs in T47D cells. A. 500k pTRIPZ-T47D-mir93 cells were injected into the 4th fatpads of NOD/SCID mice. Treatment was initiated as indicated by the red arrow. DOX (1 mg/ml in drinking water) promoted T47D tumor growth in vivo. B. Tumors from each group were collected. Aldefluor assay was performed on dissociated cells. DOX increased the $ALDH^+$ populations in T47D. C. Serial dilutions of cells obtained from these xenografts were implanted in the 4th fatpads of secondary mice, which received no further treatment. Cells from DOX-treated tumors formed secondary tumors at all dilutions (5k, 50k, 500k). *$p<0.05$; Error bars represent mean ± STDEV.
(PDF)

Figure S14 mir-93 is induced in primary tumors with DOX. 1000k pTRIPZ-MCF7-mir-93 cells were injected into the 4th fatpads of NOD/SCID mice. Different treatments were initiated (Vehicle Control (Control), DOX alone (DOX), docetaxel alone, or the combination). At the end of treatment, cells from Control and DOX groups were isolated from the tumors and mir-93 expression level was measured by qRT-PCR.
(PDF)

Figure S15 Validation of the 127 overlapped gene expression with customerized StellArray PCR array plate in pTRIPZ-SUM159-mir93.
(PDF)

Figure S16 Validation of the 127 overlapped gene expression with customerized StellArray PCR array plate in pTRIPZ-HCC1954-mir93.
(PDF)

Figure S17 Validation of the 127 overlapped gene expression with customerized StellArray PCR array plate in pTRIPZ-MC1-mir93.
(PDF)

Figure S18 Luciferase assay confirming mir93 targets. The 3′UTR of AKT3, SOX4, and STAT3 pMIR-REPORT firefly luciferase reporter plasmids with the wild-type 3′UTR sequences of AKT3, SOX4, or STAT3 were transiently transfected into pTRIPZ-HCC1954-mir93 cells and an internal control ACTB luciferase reporter was co-transfected for normalization. The cells were treated with or without DOX. Luciferase activities were measured after 48 hr. The relative luciferase activity is shown as the ratio of (the results from the cells transfected by individual reporter)/(the results from the cells transfected by the internal control in the same cell group). *$p<0.05$; Error bars represent mean ± STDEV.
(PDF)

Figure S19 Luciferase assay testing mir93 targets. The 3′UTR of AKT3, SOX4, and STAT3 pMIR-REPORT firefly luciferase reporter plasmids with the wild-type 3′UTR sequences of AKT3, SOX4, or STAT3 were transiently transfected into pTRIPZ-MCF7-mir93 (A) or pTRIPZ-MDA-MB-453-mir93 (B) cells and an internal control ACTB luciferase reporter was co-transfected for normalization. The cells were treated with or without DOX. Luciferase activities were measured after 48 hr. The relative luciferase activity is shown as the ratio of (the results from the cells transfected by individual reporter)/(the results from the cells transfected by the internal control in the same cell group). Error bars represent mean ± STDEV.
(PDF)

Figure S20 Knockdown of STAT3 (A), AKT3 (B) or SOX4 (C) decreases ALDH$^+$ cells in SUM159 cells. SUM159 cells were transfected with PlentiLox3.7-shRNA-DsRed viruses and accessed for the ALDH$^+$ population by Aldeflour assay. *$p<0.05$; Error bars represent mean ± STDEV.
(PDF)

Figure S21 Validation of the 127 overlapped gene expression with customerized StellArray PCR array plate in pTRIPZ-MCF7-mir93.
(PDF)

Figure S22 The effects of mir93 on cell proliferation. Cell proliferation was measured with the MTT assay. 200–500 cells from Control and DOX-treated groups were seeded in 96-well culture plates and were cultured in the absence (CTRL) and presence (DOX) of DOX for 7days.Data represents means SEM, $n=5$. *$p<0.05$; Error bars represent mean ± STDEV.
(PDF)

Figure S23 Analysis of cell cycle for pTRIPZ-HCC1954-mir93 cells and pTRIPZ-MCF7-mir93cells. Cell cycle analysis of pTRIPZ-HCC1954-mir93 cells and pTRIPZ-MCF7-mir93 cells in the presence or absence of DOX. Propidium iodide staining followed by flow cytometry was used to analyze cell cycle distribution. Mir93 induced by DOX treatment resulted in a decreased proportion of cells in the G0/G1 phase and an increased proportion of cells in the S/G2/M phase for pTRIPZ-HCC1954-mir93 cells. In contrast, DOX treatment has no effects on the cell cycle for pTRIPZ-MCF7-mir93 cells.
(PDF)

Figure S24 MCM7 and Ki67 expression is increased in ALDH- compared to ALDH+ SUM159 cells. ALDH + and − cells were separated by Aldefluor assay and expression of Ki67 and MCM7 accessed by immunofluorescence. Ki67, Red; MCM7, Green; DAPI, Blue. One representative sample from 3 independent samples is shown.
(PDF)

Figure S25 mir93 expression induces MET in HCC1954 cells. pTRIPZ-HCC1954-mir93 cells were plated in 2-well chamber slides with (DOX) or without (CTRL) Doxycycline for 7 days. E-Cadherin and Vimentin were stained with immunofluorescence staining. E-Cadherin, Green; Vimentin, Red; DAPI, Blue. One representative sample from 3 independent samples is shown.
(PDF)

Table S1 Downregualted probe set in ALDH$^+$ population from DOX vs. ALDH$^+$ population from CTRL.
(PDF)

Table S2 mir93 direct targets in SUM159 cells. Overlap between mir93 predicted targets from TargetScan 5.1 and profiling data from DOX-treated cells (DOX) to non-DOX-treated cells (CTRL) in the ALDH$^-$ population (12 genes) or in the ALDH$^+$ population (352 genes). Known stem cell regulatory genes highlighted in red. Genes underlined and bolded were analyzed utilizing the luciferase reporter assay.
(PDF)

Table S3 Downregualted probe set in ALDH$^-$ population from DOX vs. ALDH$^-$ population from CTRL.
(PDF)

Acknowledgments

Thanks to Dr. Stephen Ethier for generously providing the breast cancer cell lines SUM159. We thank the University of Michigan Comprehensive Cancer Center Flow Cytometry Core, Microarray core (Joe Washburn, Craig Johnson), Vector Core, Molecular Imaging Core and Pathology Core for invaluable assistance.

Author Contributions

Conceived and designed the experiments: SL CG DB GJH MSW. Performed the experiments: SL SHP CG II RM-T SB SPM LS JK SJO AH KJZ. Analyzed the data: SL CG II HK SGC EC-J MSW. Contributed reagents/materials/analysis tools: SL DB GJH MSW. Wrote the paper: SL MSW.

References

1. Petrocca F, Vecchione A, Croce CM (2008) Emerging role of miR-106b-25/miR-17-92 clusters in the control of transforming growth factor beta signaling. Cancer Res 68: 8191–8194.
2. Hayashita Y, Osada H, Tatematsu Y, Yamada H, Yanagisawa K, et al. (2005) A polycistronic microRNA cluster, miR-17-92, is overexpressed in human lung cancers and enhances cell proliferation. Cancer Res 65: 9628–9632.
3. Mendell JT (2008) miRiad roles for the miR-17-92 cluster in development and disease. Cell 133: 217–222.
4. Bandres E, Agirre X, Ramirez N, Zarate R, Garcia-Foncillas J (2007) MicroRNAs as cancer players: potential clinical and biological effects. DNA Cell Biol 26: 273–282.
5. Hernando E (2007) microRNAs and cancer: role in tumorigenesis, patient classification and therapy. Clin Transl Oncol 9: 155–160.
6. Negrini M, Ferracin M, Sabbioni S, Croce CM (2007) MicroRNAs in human cancer: from research to therapy. J Cell Sci 120: 1833–1840.
7. Wiemer EA (2007) The role of microRNAs in cancer: no small matter. Eur J Cancer 43: 1529–1544.
8. Wijnhoven BP, Michael MZ, Watson DI (2007) MicroRNAs and cancer. Br J Surg 94: 23–30.
9. Yu F, Yao H, Zhu P, Zhang X, Pan Q, et al. (2007) let-7 regulates self renewal and tumorigenicity of breast cancer cells. Cell 131: 1109–1123.
10. Shimono Y, Zabala M, Cho RW, Lobo N, Dalerba P, et al. (2009) Downregulation of miRNA-200c links breast cancer stem cells with normal stem cells. Cell 138: 592–603.
11. Esquela-Kerscher A, Trang P, Wiggins JF, Patrawala L, Cheng A, et al. (2008) The let-7 microRNA reduces tumor growth in mouse models of lung cancer. Cell Cycle 7: 759–764.
12. Poliseno L, Salmena L, Riccardi L, Fornari A, Song MS, et al. (2010) Identification of the miR-106b~25 microRNA cluster as a proto-oncogenic PTEN-targeting intron that cooperates with its host gene MCM7 in transformation. Sci Signal 3: ra29.
13. Xu Y, Fang F, Zhang J, Josson S, St Clair WH, et al. (2010) miR-17* suppresses tumorigenicity of prostate cancer by inhibiting mitochondrial antioxidant enzymes. PLoS ONE 5: e14356. doi:10.1371/journal.pone.0014356.
14. Al-Hajj M, Wicha MS, Benito-Hernandez A, Morrison SJ, Clarke MF (2003) Prospective identification of tumorigenic breast cancer cells. Proc Natl Acad Sci U S A 100: 3983–3988.
15. Li C, Heidt DG, Dalerba P, Burant CF, Zhang L, et al. (2007) Identification of pancreatic cancer stem cells. Cancer Res 67(3): 1030–1037.
16. Singh SK, Clarke ID, Terasaki M, Bonn VE, Hawkins C, et al. (2003) Identification of a cancer stem cell in human brain tumors. Cancer research 63(18): 5821–5828.

17. Li X, Lewis MT, Huang J, Gutierrez C, Osborne CK, et al. (2008) Therapeutic resistance and tumor-initiation: Molecular pathways involved in breast cancer stem cell self-renewal. J Natl Cancer Inst in press.
18. Silber J, Lim DA, Petritsch C, Persson AI, Maunakea AK, et al. (2008) miR-124 and miR-137 inhibit proliferation of glioblastoma multiforme cells and induce differentiation of brain tumor stem cells. BMC Med 6: 14.
19. Yu F, Yao H, Zhu P, Zhang X, Pan Q, Gong C, Huang Y, Hu X, Su F, Lieberman J, Song E (2007) let-7 regulates self-renewal and tumorigenicity of breast cancer cells. Cell 131: 1109–1123.
20. Charafe-Jauffret E, Ginestier C, Iovino F, Wicinski J, Cervera N, et al. (2009) Breast cancer cell lines contain functional cancer stem cells with metastatic capacity and a distinct molecular signature. Cancer Res 69: 1302–1313.
21. Ibarra I, Erlich Y, Muthuswamy SK, Sachidanandam R, Hannon GJ (2007) A role for microRNAs in maintenance of mouse mammary epithelial progenitor cells. Genes Dev 21: 3238–3243.
22. Ginestier C, Hur MH, Charafe-Jauffret E, Monville F, Dutcher J, et al. (2007) ALDH1 is a marker of normal and malignant human mammary stem cells and a predictor of poor clinical outcome. Cell Stem Cell 1: 555–567.
23. Liu S, Wicha MS (2010) Targeting breast cancer stem cells. J Clin Oncol 28: 4006–4012.
24. Prat A, Parker JS, Karginova O, Fan C, Livasy C, et al. (2010) Phenotypic and molecular characterization of the claudin-low intrinsic subtype of breast cancer. Breast Cancer Res 12: R68.
25. Fillmore CM, Kuperwasser C (2008) Human breast cancer cell lines contain stem-like cells with the capacity to self-renew, give rise to phenotypically diverse progeny and survive chemotherapy. Breast Cancer Res 10: R25.
26. Akhurst RJ, Balmain A (1999) Genetic events and the role of TGF beta in epithelial tumour progression. J Pathol 187: 82–90.
27. Barrack ER (1997) TGF beta in prostate cancer: a growth inhibitor that can enhance tumorigenicity. Prostate 31: 61–70.
28. Lim E, Vaillant F, Wu D, Forrest NC, Pal B, et al. (2009) Aberrant luminal progenitors as the candidate target population for basal tumor development in BRCA1 mutation carriers. Nat Med 15: 907–913.
29. Wicha MS, Liu S, Dontu G (2006) Cancer Stem Cells: An Old Idea - A Paradigm Shift. Cancer Research 66: 1883–1890.
30. Li Z, Yang CS, Nakashima K, Rana TM (2011) Small RNA-mediated regulation of iPS cell generation. EMBO J 30: 823–834.
31. Mani SA, Guo W, Liao MJ, Eaton EN, Ayyanan A, et al. (2008) The epithelial-mesenchymal transition generates cells with properties of stem cells. Cell 133: 704–715.
32. Neve RM, Chin K, Fridlyand J, Yeh J, Baehner FL, et al. (2006) A collection of breast cancer cell lines for the study of functionally distinct cancer subtypes. Cancer Cell 10: 515–527.
33. Dontu G, Jackson KW, McNicholas E, Kawamura MJ, Abdallah WM, et al. (2004) Role of Notch signaling in cell-fate determination of human mammary stem/progenitor cells. Breast cancer research : BCR 6(6.
34. Liu S, Dontu G, Mantle ID, Patel S, Ahn NS, et al. (2006) Hedgehog signaling and Bmi-1 regulate self-renewal of normal and malignant human mammary stem cells. Cancer Res 66(12): 6063–6071.
35. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, et al. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics (Oxford, England) 4: 249–264.
36. Mizutani K, Sud S, McGregor NA, Martinovski G, Rice BT, et al. (2009) The chemokine CCL2 increases prostate tumor growth and bone metastasis through macrophage and osteoclast recruitment. Neoplasia 11: 1235–1242.

Molecular hierarchy of mammary differentiation yields refined markers of mammary stem cells

Camila O. dos Santos[a,1], Clare Rebbeck[a], Elena Rozhkova[a], Amy Valentine[a], Abigail Samuels[a,b], Lolahon R. Kadiri[c], Pavel Osten[c], Elena Y. Harris[d,2], Philip J. Uren[d], Andrew D. Smith[d], and Gregory J. Hannon[a,1]

[a]Howard Hughes Medical Institute, Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724; [b]Arts and Science Undergraduate Program, Vanderbilt University, Nashville, TN 37212; [c]Cold Spring Harbor Laboratory, Cold Spring Harbor, NY 11724; and [d]Molecular and Computational Biology, University of Southern California, Los Angeles, CA 90089

This contribution is part of the special series of Inaugural Articles by members of the National Academy of Sciences elected in 2012.

Contributed by Gregory J. Hannon, March 11, 2013 (sent for review December 28, 2012)

The partial purification of mouse mammary gland stem cells (MaSCs) using combinatorial cell surface markers (Lin$^-$CD24$^+$CD29^{h}CD49f^h) has improved our understanding of their role in normal development and breast tumorigenesis. Despite the significant improvement in MaSC enrichment, there is presently no methodology that adequately isolates pure MaSCs. Seeking new markers of MaSCs, we characterized the stem-like properties and expression signature of label-retaining cells from the mammary gland of mice expressing a controllable H2b-GFP transgene. In this system, the transgene expression can be repressed in a doxycycline-dependent fashion, allowing isolation of slowly dividing cells with retained nuclear GFP signal. Here, we show that H2b-GFP^h cells reside within the predicted MaSC compartment and display greater mammary reconstitution unit frequency compared with H2b-GFP^{neg} MaSCs. According to their transcriptome profile, H2b-GFP^h MaSCs are enriched for pathways thought to play important roles in adult stem cells. We found Cd1d, a glycoprotein expressed on the surface of antigen-presenting cells, to be highly expressed by H2b-GFP^h MaSCs, and isolation of Cd1d$^+$ MaSCs further improved the mammary reconstitution unit enrichment frequency to nearly a single-cell level. Additionally, we functionally characterized a set of MaSC-enriched genes, discovering factors controlling MaSC survival. Collectively, our data provide tools for isolating a more precisely defined population of MaSCs and point to potentially critical factors for MaSC maintenance.

FACS sorting | mammary gland transplant | shRNA screen

The murine mammary gland resembles, to some extent, the human mammary gland in development, milk production, and progression to carcinogenesis, making it an ideal system to develop methodologies and form hypotheses of relevance to women. The use of cell surface markers to isolate selected cell types from mice has greatly enhanced our understanding of development and our knowledge of molecular pathways and interactions that influence it. Mammary gland stem cells (MaSCs) have commanded attention because of not only their roles in the cycles of gland morphogenesis but also their potential contribution in tumor initiation. Full characterization of MaSCs, however, has been hampered by their scarcity. Enrichment of the MaSC compartment has, until now, been achieved by using a combination of cell surface markers (Lin$^-$CD24$^+$CD29^{h}CD49f^h) (1, 2). Thus far, these cells have been enriched to 1 MaSC per every 64 cells stained Lin$^-$CD24$^+$CD29^h (1). This is sufficient to test for MaSC repopulation capacity and to some extent, roles in tumorigenesis, but this level of purity is less suitable for more complex molecular analyses that define MaSCs and their properties.

Additional characterization of MaSCs has been achieved using a transgenic mouse model expressing GFP under the control of the *s-ship* promoter (3). This gene is expressed in embryonic and hematopoietic stem cells but not differentiated cells (4). GFP$^+$ cells in this mouse model were shown to reside at the tips of the terminal end buds, where MaSCs are believed to be located in these developing mammary gland structures (3, 5). Transplantation of the MaSC-enriched GFP$^+$CD49f^h cells improved the mammary reconstitution unit (MRU) frequency to 1/48 cells, an increase over the previous shown frequency for CD24$^+$CD29^{h}CD49f^h cells. Although being very elegantly performed and enhancing our understanding of MaSC localization, studies with this mouse model did not achieve a greater enrichment for MaSCs using more conveniently accessible markers, such as cell surface proteins.

Given the limitations in accurately purifying MaSCs, we sought to devise a method better suited for identifying this population. Here, we describe the use of long-term label retention to increase the MRU frequency within MaSC-enriched CD24$^+$CD29^h cells. This approach, previously applied to the isolation of skin stem cells (6), enables the identification of slowly dividing cells, a characteristic of adult stem cells. To mark slowly dividing cells, expression of the H2b histone, linked to GFP, is regulated by a tetracycline responsive element (TRE) and a tet-controlled transcription activator (tTA) under the endogenous keratin K5 promoter (K5tTA-H2b-GFP). In the absence of tetracycline or its analog doxycycline (DOX), the tTA binds to TRE and activates transcription of H2b-GFP. Treatment with DOX prevents the tTA binding to TRE, and transcription of H2b-GFP is terminated (6). As the cell divides, newly synthesized, unlabeled H2b replaces the H2b-GFP; therefore, the more slowly dividing cells will retain GFP expression for an extended period.

We were able to improve the MaSC enrichment by isolating GFP-retaining cells after a long-term inhibition of transgene expression. We refer to these cells as H2b-GFP^h MaSCs (CD24$^+$CD29^{h}H2b-GFP^h). Comparisons between expression profiles of all mammary gland cell types suggested that H2b-GFP^h MaSCs differentially expressed several genes involved in pathways previously described as playing roles in other adult stem cells. Additional analysis of the H2b-GFP^h MaSC expression signature led to the identification of a cell surface marker that, combined

Author contributions: C.O.d.S. and G.J.H. designed research; C.O.d.S., E.R., A.V., A.S., L.K., and P.O. performed research; C.O.d.S., E.Y.H., P.J.U., and A.D.S. analyzed data; and C.O.d.S., C.R., and G.J.H. wrote the paper.

The authors declare no conflict of interest.

Freely available online through the PNAS open access option.

Data deposition: The sequences reported in this paper have been deposited at the National Center for Biotechnology Information Short Read Archive (NCBI SRA) [submission study SRP018968 regarding the following samples: Myo differentiated SRX246984 (SRR768419 and SRR768420); Myo progenitor SRX246985 (SRR768421 and SRR768422); luminal alveolar SRX246986 (SRR768423 and SRR768424); luminal progenitor SRX246987 (SRR768425 and SRR768426); luminal ductal SRX246988 (SRR768427 and SRR768428); H2b GFP MaSC SRX246989 (SRR768429, SRR768430, and SRR768431); CD59 MaSC SRX246990 (SRR768432); and CD1d MaSC SRX246991 (SRR768433)].

[1]To whom correspondence may be addressed. E-mail: dossanto@cshl.edu or hannon@cshl.edu.

[2]Present address: Computer Science Department, California State Polytechnic University, Pomona, CA 91768.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1303919110/-/DCSupplemental.

with conventional markers, resulted in the isolation of an MaSC population with an elevated proportion of MRUs. In addition, we performed a focused shRNA screen, targeting genes that were differentially expressed in our newly characterized MaSC-enriched cell population, revealing potential regulators of mammary gland biogenesis. Overall, this work improves our ability to purify MaSCs and provides valuable insights into their role in mammary gland development and perhaps, even tumor initiation.

Results

H2b-GFP Label-Retaining Cells Enrich for MaSCs. To better enrich for the MaSC population, we assessed the feasibility of using mammary gland label-retaining cells to select for MaSCs, given that a slower division rate is an excepted characteristic of adult stem cells. We adopted a system wherein expression of the H2b histone, linked to GFP, is regulated by a TRE and a tTA under the endogenous keratin K5 promoter K5tTA-H2b-GFP (a gift from Elaine Fuchs, Rockefeller University, New York, NY). Keratin K5 is expressed in cells of the basal compartment, the region considered to be home to MaSCs (7). This system displays some advantages over the previous gene reporter-based methods used to isolate MaSCs, because it takes advantage of one of the more general properties of stem cells: their relative quiescence. In support of the use of this mouse model, there were previous hints that MaSC-enriched CD24^{+}CD29^{h} cells display BrdU label-retaining properties (1), although label-retaining populations were not functionally characterized.

Initial experiments using the H2b-GFP mice assessed the expression and distribution of GFP-positive cells in the adult mammary gland (Fig. 1*A*). Histological sections revealed the presence of several GFP^{+} cells located within structures resembling the mammary gland ductal epithelium (Fig. 1*B* and Fig. S1*A*, *Upper*). Treatment of H2b-GFP mice with DOX over a 12-wk period, thus ceasing transcription of H2b-GFP transgene, dramatically reduced the number of cells expressing GFP. Notably, those cells that remained GFP^{+} were located at the tips of the terminal end buds. These distinct sites in the ductal epithelium are the areas currently believed to be resident by MaSCs (8) (Fig. 1*C* and Fig. S1*A*, *Lower*).

To compliment this observation, under the hypothesis that mammary gland label-retaining cells comprise a population of potential MaSCs, we investigated the correlation between GFP retention and expression of previously defined MaSC-enriched cell surface markers, CD24 and CD29. Using FACS analysis, we were able to subdivide the mammary gland (after depletion of endothelial and hematopoietic cells as shown in Fig. S1*B*) into three distinct cell compartments: luminal (CD24^{h}CD29^{+}), occupied by luminal cells; basal (CD24^{+}CD29^{h}), occupied by myoepithelial cells and MaSCs; and stromal (CD24^{-}CD29^{+}) (1) (Fig. 1*D*, *Upper Left*). The majority of GFP^{+} cells from a transgenic H2b-GFP mouse off DOX could be categorized into either basal or stromal compartments, with far fewer GFP^{+} cells occupying the luminal compartment (Fig. 1*D*, *Upper Right* and Fig. S1*C*, *Left*). After a 12-wk DOX chase, the overall proportion of GFP^{+} cells decreased by more than one-half, and the presence of a GFP^{+} luminal compartment was all but eliminated (Fig. 1*D*, *Lower Left* and Fig. S1*C*, *Center*). Focusing on GFP intensity (a measure that directly relates to the rate of cell division), selection of only the brightest GFP^{+} cells (GFP^{h}) resulted in a greater proportion remaining in the CD24^{+}CD29^{h} basal compartment, whereas the stromal compartment was significantly reduced after GFP^{dim} cells were removed (Fig. 1*D*, *Lower Right* and Fig. S1*C*, *Right*). This result suggests that the most label-retaining cells reside within the basal compartment and may represent the MaSCs population.

The benefit of using GFP to test for label retention, as opposed to BrdU, is that its detection does not require fixation and staining. We were then able to test the biological differences, using mammary gland transplants, between GFP^{h} cells (H2b-GFP^{h} MaSCs) and GFP^{-} cells (H2b-GFP^{-} MaSCs) within the MaSC-enriched compartment. Transplantation assays are a fundamental criterion to evaluate stemness and have been used previously for several tissues, including the mammary gland (1, 2, 9). For these experiments, the inguinal glands were removed from the endogenous tissue of prepubescent females before injection of donor cells. Donor cells were harvested from mammary glands of H2b-GFP mice after a 12-wk DOX chase, dissociated, lineage-depleted, and sorted according to GFP intensity (Fig. S1*D*). Cells (GFP^{h} and GFP^{-}) were then injected, and outgrowths from donor cells were compared (by visualization of GFP^{+} epithelium) 12 wk posttransplantation. Given that the recipient animals are not treated with DOX, all cells derived from the donor mice will resume expression of the H2b-GFP transgene and give rise to GFP^{+} outgrows. MRU frequency was estimated according to the previously described algorithm (10). Transplantation of 500 H2b-GFP^{h} MaSCs ($n = 5$) gave rise to GFP^{+} epithelium in all injected glands. This ability to reconstitute was still retained when only 50 cells were transplanted (Fig. 1*E*). In contrast, only one-half of the glands injected with 500 H2b-GFP^{-} MaSCs displayed fluorescent outgrowths, decreasing to just 29% with injection of 50 cells (Dataset S1). These results represent an increase in the estimated frequency of MRUs from 1/70 cells, when MaSC selection was performed using CD24^{+}CD29^{h} alone (1), to 1/33 cells, with restriction to H2b-GFP^{h} cells to further define MaSCs. Comparatively, the MRU frequency among H2b-GFP^{-} MaSCs was estimated to be 1/149 (Dataset S1). Colony-forming ability was also twofold greater for H2b-GFP^{h} MaSCs when 500 of these cells were seeded in a Matrigel (BD Bioscience) and cultured for 7 d (Fig. S1*E*).

Considered together, these data suggest that mammary gland H2b-GFP^{h} label-retaining cells represent a subset, if not an entire population, of the MaSCs. Our experiments using a repressible H2b-GFP transgene have built on previous knowledge regarding the label-retaining properties of stem cells in the mammary gland and confirmed that MaSC CD24^{+}CD29^{h} cells reside mainly within the H2b-GFP^{h} label-retaining cell population. In addition to these experiments, we also found that hormone-dependent activation of MaSC proliferation and differentiation, triggered by one complete cycle of pregnancy and involution in transgenic H2b-GFP mice treated with DOX, completely depleted GFP^{+} cells, validating that H2b-GFP^{h} cells truly represent a population of slowly dividing cells rather than being a transgenic artifact.

It has been proposed that MaSCs comprise less than 5% of the total basal compartment. Our findings support this notion given that we find label-retaining H2b-GFP^{h} cells to account for ~0.2% of the total CD24^{+}CD29^{h} population (Fig. S1*D*, *Upper Right*). We also compared the distribution of H2b-GFP^{h}–retaining cells with expression of a recently identified marker for myoepithelial progenitor-like cells, CD61. This marker was expressed by most of the H2b-GFP^{dim} population, whereas virtually all H2b-GFP^{h} cells were negative for CD61 staining, suggesting perhaps a unique mammary gland cell differentiation pattern, where H2b-GFP^{h} label-retaining cells might occupy the top of hierarchy.

H2b-GFP Cells Display a Stem Cell-Like Expression Signature. Having established that H2b-GFP^{h} MaSCs have reconstitution properties, we next sought to determine where these cells fall in the mammary differentiation hierarchy with regard to their gene expression patterns. Using a combination of cell surface markers (1, 11), six distinct cell types were isolated by FACS to a purity of >90%: H2b-GFP MaSCs (Lin^{-}CD24^{+}CD29^{h}H2b-GFP^{h}CD61^{-}), myoepithelial progenitor-like cells (Lin-CD24^{+}CD29^{h}H2b-GFP$^{-/l}$CD61^{+}), myoepithelial differentiated cells (Lin^{-}CD24^{+}CD29^{h}H2b-GFP^{-} CD61^{-}), luminal progenitor cells (Lin^{-}CD24^{h}CD29^{+}CD61^{+} CD133^{-}), luminal ductal cells (Lin^{-}CD24^{h}CD29^{+}CD61^{-}CD133^{+}), and luminal alveolar cells (Lin^{-}CD24^{h}CD29^{+}CD61^{-} CD133^{-})

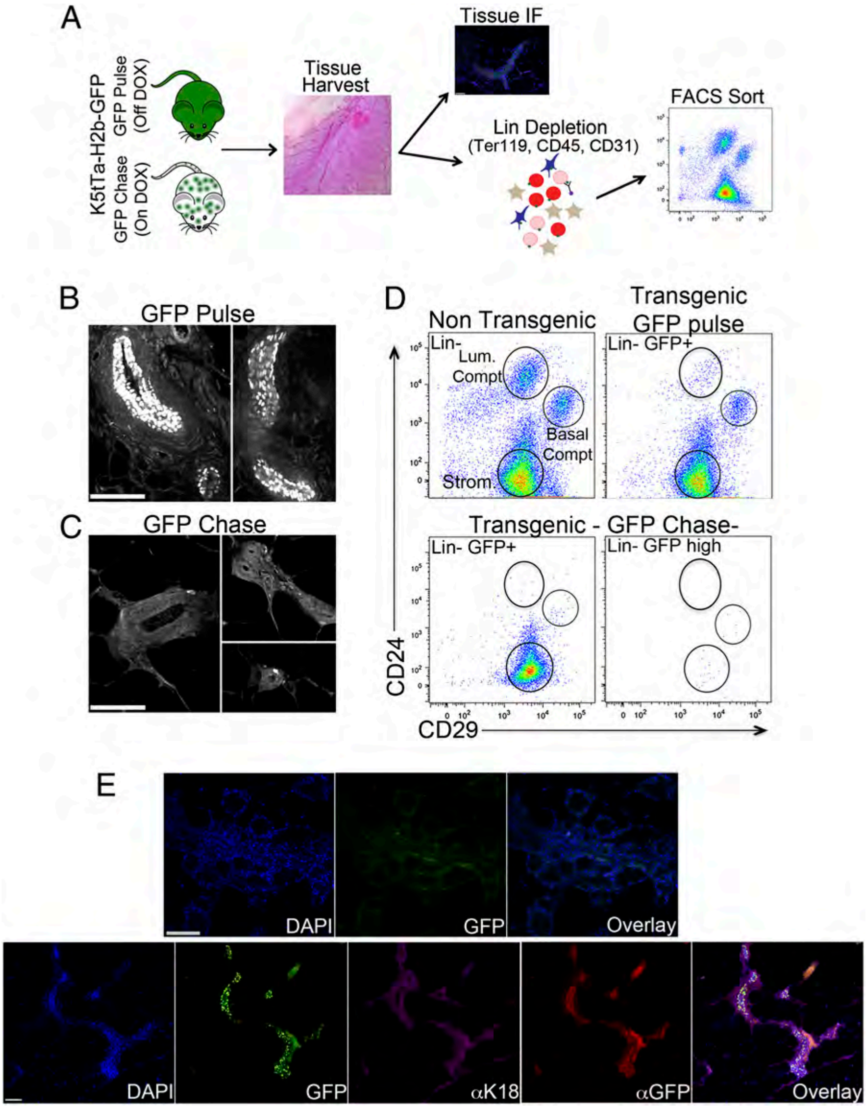


Fig. 1. H2b-GFP label-retaining cells represent a population of MaSCs. (*A*) Experimental scheme. Mammary glands were harvested from K5tTa-H2b-GFP transgenic mice either off (GFP pulse) or on DOX diet (GFP chase) and further processed for immunological staining or single-cell suspension FACS sorting. (*B* and *C*) Tissue histology H2b-GFP^{+} cells distribution. Mammary glands from transgenic mice off and on DOX diet were harvested, defatted, embedded in agarose, and imaged with two-photon microscopy. (*B*) Mice off DOX diet (GFP pulse) showing a broad distribution of GFP^{+} cells in mammary gland ductal structures. (*C*) After a 12-wk DOX chase, H2b-GFP^{+} label-retaining cells became restricted to the edges of the ductal structures. (*D*) Flow cytometry profile of H2B-GFP^{+} cells. *Upper Left* shows the profile of a lineage-depleted (CD45^{-}, Ter119^{-}, and CD31^{-}) nontransgenic mammary gland according to CD24 and CD29 staining and highlights the three cell compartments: luminal (CD24^{h}CD29^{+}; comprising luminal progenitor cells, luminal alveolar cells, and luminal ductal cells); basal (CD24^{+}CD29^{h}; comprising myoepithelial progenitor cells, myoepithelial differentiated cells, and MaSCs); and stromal. Total GFP^{+} cells from H2b-GFP transgenic mice off DOX diet (GFP pulse mice; *Upper Right*) displayed a similar cellular compartmental distribution with fewer luminal-type cells. The CD24CD29 cell profile of H2b-GFP^{+} cells from GFP chase mice (on DOX) were analyzed using two strategies to define GFP-expressing cells. *Lower Left* displays CD24CD29 staining of total H2b-GFP^{+} cells, whereas *Lower Right* shows the CD24CD29 staining of H2b-GFP^{h} cells. The focus on GFP^{h} cells, the most label-retaining cells, drastically decreased the cellular content of all mammary gland compartments and retained a greater proportion of cells inside of the basal compartment, potentially representing MaSCs. (*E*) Histological analysis of mammary gland H2b-GFP^{h} MaSCs outgrowths. Cleared fat pads from prepubescent female mice were injected with either total H2b-GFP^{-} MaSCs (CD24^{+}CD29^{h}GFP^{-} cells) or H2b-GFP^{h} MaSCs (CD24^{+}CD29^{h}GFP^{h}cells), harvested 12 wk after transplantation, embedded in agarose, stained with antibodies, and imaged on a Zeiss 710 LSM (Zeiss) confocal microscope. Images display outgrowths of two distinct glands injected with H2b-GFP^{h} MaSCs.

(Fig. 2*A*). The myoepithelial progenitor-like cells were defined by expression of CD61 as a positive cell surface marker and their positioning as the second most label-retaining cell population.

Hierarchical clustering of combined RNAseq replicates split mammary gland cells into two main branches: the basal compartment, comprising myoepithelial progenitor cells, myoepithelial differentiated cells, and H2b-GFP MaSCs, and the luminal compartment, with luminal progenitor cells and differentiated cells (Fig. 2*B*). As predicted by prior characterization of MaSCs (1), we found the expression profile of H2b-GFP MaSCs to be more closely related to the expression profile of myoepithelial cells than luminal cells; however, H2b-GFP MaSCs were still an outgroup compared with other cells in this cluster. Analysis over all mammary gland cell types yielded several hundred genes differentially expressed among all cell types (Fig. 2*B*), spanning diverse gene ontology groups and pathways (Dataset S2). More specifically, genes differentially expressed in H2b-GFP MaSCs were enriched in G protein-coupled receptors and pathways involving Wnt/B-catenin signaling, areas previously described to play fundamental roles in other adult stem cells (12). Differential

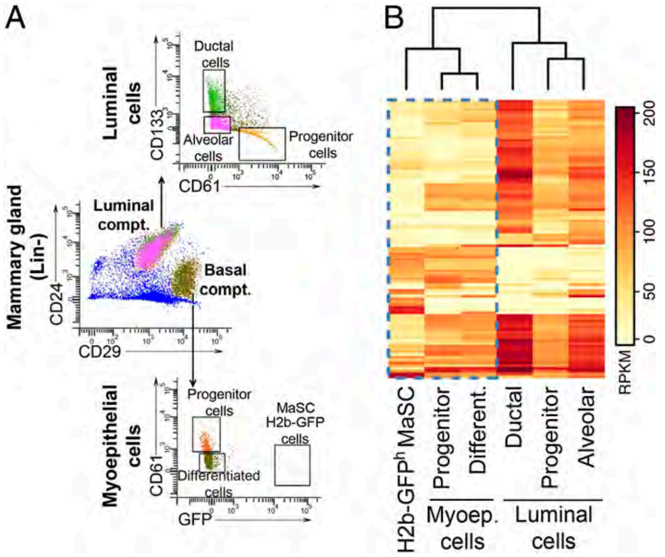


Fig. 2. H2b-GFP^h MaSCs display a stem-like expression signature. (*A*) Sorting strategy. We used a combination of four cell surface markers in addition to H2b-GFP expression to segregate the lineage-depleted mammary gland cells into six distinct cell types: H2b-GFP^h MaSCs (Lin$^-$CD24$^+$CD29^{h}H2b-GFP^{h}CD61$^-$), myoepithelial progenitors cells (Lin-CD24$^+$CD29^{h}H2b-GFP$^-$CD61$^+$), myoepithelial differentiated cells (Lin$^-$CD24$^+$CD29^{h}H2b-GFP$^-$CD61$^-$), luminal progenitor cells (Lin$^-$CD24^{h}CD29^{l}CD61$^+$CD133$^-$), luminal ductal cells (Lin$^-$CD24^{h}CD29^{l}CD61$^-$CD133$^+$), and luminal alveolar cells (Lin$^-$CD24^{h}CD29^{l}CD61$^-$CD133$^-$). For each library, two biological replicates were analyzed. (*B*) Mammary gland differential expression heat map. Clustering of RPKM profiles for the 100 genes with highest variance across all samples. Two main cell clusters were generated according to the expression patterns of analyzed genes: luminal- (progenitor, alveolar, and ductal cells) and basal-type cells (H2b-GFP^h MaSCs, progenitors, and differentiated cells). Note that H2b-GFP^h MaSCs cluster with other basal compartment cells but have an expression signature distinct from the other two cell types in this cluster.

expression patterns were confirmed on four genes by performing quantitative RT-PCR on H2b-GFP MaSCs (n = 4 individually sorted samples) and myoepithelial progenitor cells (n = 3 individually sorted samples) (Fig. S2). mRNA for the *Cd24* and *Cd29* genes was quantified as control, because all myoepithelial cells displayed similar levels of expression for these genes.

These results further confirmed that mammary cell types could be differentiated based on their gene expression profiles, allowing us to use these profiles to select cell type-specific genetic identifiers.

Additional Cell Surface Marker to Improve MaSCs Purification. Because of limitations on the ability to purify MaSCs to homogeneity based on currently used cell surface markers, we searched for new surface markers that might identify MaSCs using the RNAseq data. We first generated a list of ∼500 genes that encode for cell surface markers according to their gene ontology term function (e.g., basolateral membrane, cell surface, membrane protein, or basement membrane). This list was further reduced to genes with high expression levels for the MaSC H2b-GFP cells. Five candidate cell surface proteins came out of this analysis (Fig. S3*A*): CD1d, a glycoprotein expressed on the surface of various mouse and human antigen-presenting cells (13); Cd59a, a regulator of the membrane attack complex (14); CD22, a regulatory lectin involved in repressing hyperactivation of the immune system (15); CD93, a C-type lectin involved in cell–cell adhesion processes (16); and CD74, an HLA class II protein, part of the major histocompatibility complex (17). Antibodies against CD1d, CD59a, and CD22 positively stained a distinct population of cells contained within the Lin-MaSC CD24$^+$CD29^h cells (Fig. 3*A*), whereas antibodies against the proteins, CD93 and CD74, failed to stain any mammary gland cells. We further tested CD1d MaSCs (CD24$^+$CD29^{h}CD1d$^+$), CD59a MaSCs (CD24$^+$CD29^{h}CD59a^h), and CD22 MaSCs (CD24$^+$CD29^{h}CD22$^+$) for their ability to grow colonies in Matrigel culture. Two populations, CD1d MaSCs and CD59a^h MaSCs [representing 1% and 4%, respectively, of the total MaSC (CD24$^+$CD29^h) population], displayed an approximately twofold increase in colony-forming ability compared with the total MaSCs population (Fig. S3*B*). However, we found CD1d MaSCs to have a greater colony-forming ability compared with CD59a^h MaSCs, with one-half as many cells needed to produce the same number of colonies (200 and 500 cells, respectively, seeded on Matrigel). Additional analysis showed that all CD1d$^+$ cells from the MaSC-enriched CD24$^+$CD29^h population were also CD59a^h, whereas the remaining majority of CD59a$^+$ cells from the MaSC-enriched CD24$^+$CD29^h population was negative for CD1d expression (Fig. 3*B*). Based on the enhanced colony-forming abilities of CD1d MaSCs over CD59a^h MaSCs and the overlap of the two markers within the CD1d$^+$ populations, we decided to pursue the experiments using CD1d as an MaSC marker.

We next sorted CD1d$^+$ MaSCs for RNAseq and compared their gene expression profile with those cell populations described in Fig. 2. Cluster analysis of all RNAseq libraries suggests that the CD1d MaSC expression signature is closer to the expression pattern found for H2b-GFP MaSCs than for any other cell type (Fig. S3*C*). These results could be suggestive that the common expression signature between CD1d$^+$ MaSCs and H2b-GFP^h MaSCs defines the stem cell state of mammary gland cells.

To ask whether CD1d$^+$ MaSCs are slowly dividing cells, we performed BrdU label retention experiments. We injected BrdU into eight prepubescence female mice (3 wk old) over 5 consecutive d. Cells were harvested on the day of the last injection (week 0) from one-half of the mice and after 12 wk from the remaining mice (Fig. 3*C*). FACS analysis showed that ∼20% of the total MaSC population retained BrdU, and up to 60% of CD1d MaSCs were BrdU-retentive. This result adds confidence to the use of CD1d as a cell surface marker to represent the H2b-GFP MaSCs, because it is the most label-retaining cells and perhaps, therefore, the most enriched for stem-like cells within the mammary gland.

We then went on to repeat the mammary gland reconstruction assays, but this time, we compared CD1d$^+$ MaSC transplantation efficiency with the transplantation efficiency displayed by the total MaSC (Lin$^-$CD24$^+$CD29^h) population using cells from the H2b-GFP mice off DOX. Comparing donor-derived outgrowths (identifiable by GFP expression) between injection with CD1d$^+$ MaSCs and injection with total MaSCs, we found that, despite bringing the injected cell number down to single digits, CD1d$^+$ MaSCs effectively gave rise to GFP outgrowths in the majority of graft recipients (Fig. 3*D* and Dataset S3). This result gave a predicted MRU frequency of ∼1/8 CD1d MaSCs compared with the 1/44 MRU frequency from total MaSCs (Dataset S3). FACS collection of CD1d$^+$ MaSCs from a reconstructed gland also effectively gave rise to a gland when serially transplanted into another mouse, showing that these cells also have the capacity to self-renew in addition to regenerating the gland (Fig. 3*E*).

MaSC-Focused shRNA Screen. To identify genes and pathways necessary for the maintenance of MaSC reconstitution potential, we selected a set of abundantly and differentially expressed genes from RNAseq libraries of H2b-GFP^h MaSCs and CD1d$^+$ MaSCs and targeted them in shRNA-mediated knockdown experiments. We used shRNAs identified by a prediction algorithm developed in our laboratory, taking, on average, two hairpins per gene. Hairpins targeting nondifferentially expressed genes were also included as well as depletion control hairpins targeting *Rpa3* and *Polr2b* and neutral control hairpins targeting Firefly luciferase and Renilla luciferase. All genes were targeted in a one-by-one approach in an assay lasting ∼3 wk (Fig. S4*A*).

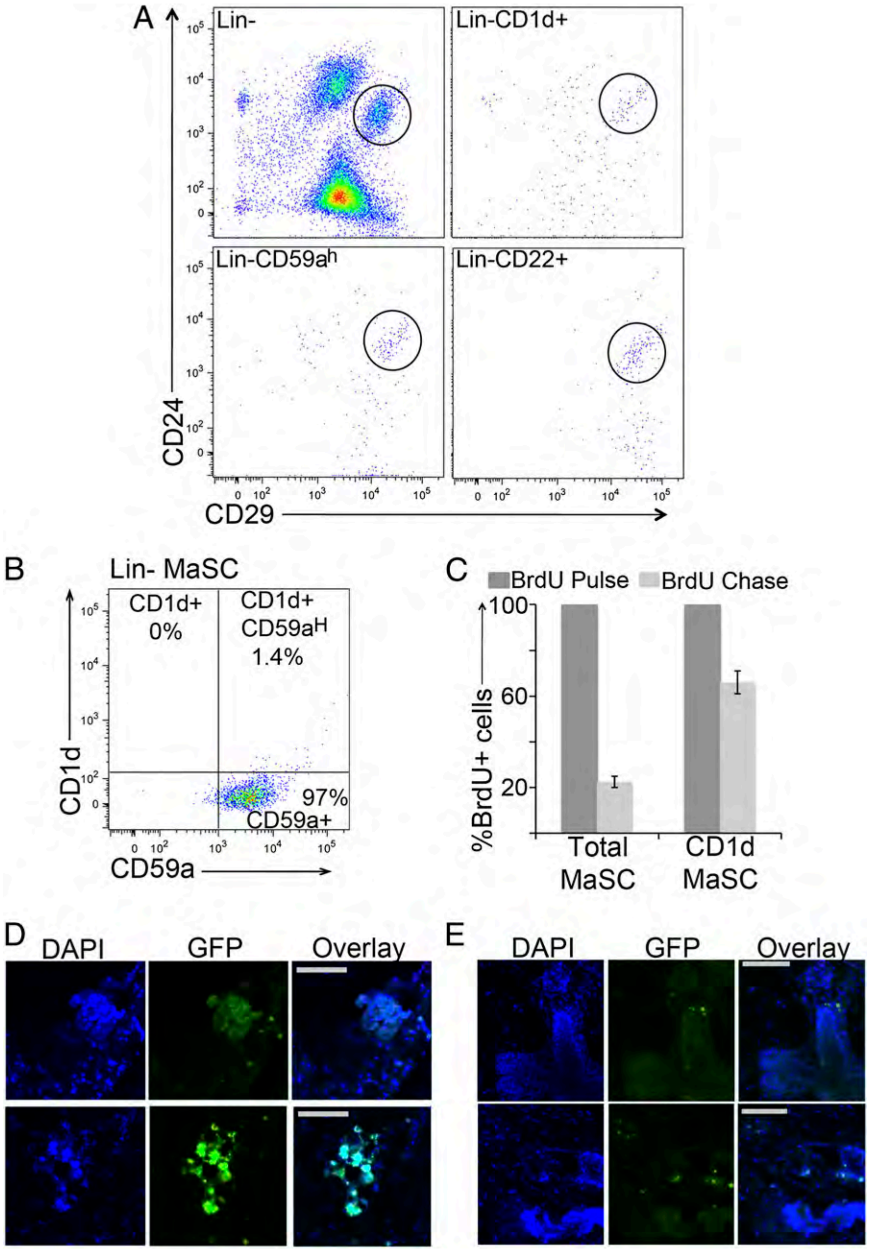


Fig. 3. Cd1d is an additional cell surface marker for purification of MaSCs. (*A*) FACS analysis of MaSC cell surface markers. Total MaSCs (CD24$^+$CD29^h cells) were additionally segregated according to the expression of the cell surface markers CD1d, CD59a^h, or CD22. (*B*) CD1d is expressed by a subset of CD59a^h cells. Lin$^-$ mammary gland cells were stained with antibodies against CD24, CD29, CD1d, and CD59a and further analyzed on an LSRII Cell Analyzer (BD Bioscience). The entire basal compartment (CD24$^+$CD29^h) was selected and analyzed according to CD1d and CD59a expression. The majority of cells within the basal compartment stained positive for CD59a, and CD1d$^+$ cells fell mainly in the CD59a^h area. (*C*) CD1d MaSCs are the most label-retaining cells within the mammary gland. Prepubescence mice were injected with BrdU (50 mg/kg body weight) for 5 d. Glands were either harvested from mice on the last day of BrdU injection to evaluate the total BrdU incorporation (week 0) or harvested after a 12 wk BrdU chase. Single-cell suspensions were stained with antibodies against CD24, CD29, and CD1d and analyzed on an LSRII Cell Analyzer (BD Bioscience). BrdU incorporation was measured in total MaSC (CD24$^+$CD29^h) and CD1d MaSC (CD24$^+$CD29^{h}CD1d$^+$) populations. (*D* and *E*) Histological analysis of mammary gland CD1d MaSCs outgrowths. Cleared fat pads from prepubescent female mice were injected with (*D*) either total MaSCs or CD1d MaSCs and (*E*) 25 CD1d MaSCs cells harvested from glands pretransplanted with CD1d MaSCs. Glands were harvested 12 wk after transplantation and embedded in agarose, and endogenous GFP signal was imaged. Images display outgrowths from two distinct glands injected with CD1d MaSCs cells (*D*) or secondary transplanted CD1d MaSCs (*E*).

shRNAs were introduced into the immortalized mammary gland cell line, Comma-Dβ (18). These cells give rise to both luminal and myoepithelial compartments in colony-forming and transplantation assays, independent of the method of MaSC enrichment (19–21). In addition, ~50% of Comma-Dβ cells stain positive for Cd1d, placing them in our improved MaSCs isolation profile (Fig. S4*B*).

Cells were monitored for GFP expression (as a proxy for shRNA expression), and changes in the proportion of GFP-expressing cells would be indicative of a relevant gene function. The majority of screened shRNAs did not alter GFP levels during the 3-wk screening period (Fig. 4*A*), which could suggest that the correspondent genes were not essential for growth maintenance of Comma-Dβ cells. However, a distinct set of shRNAs altered

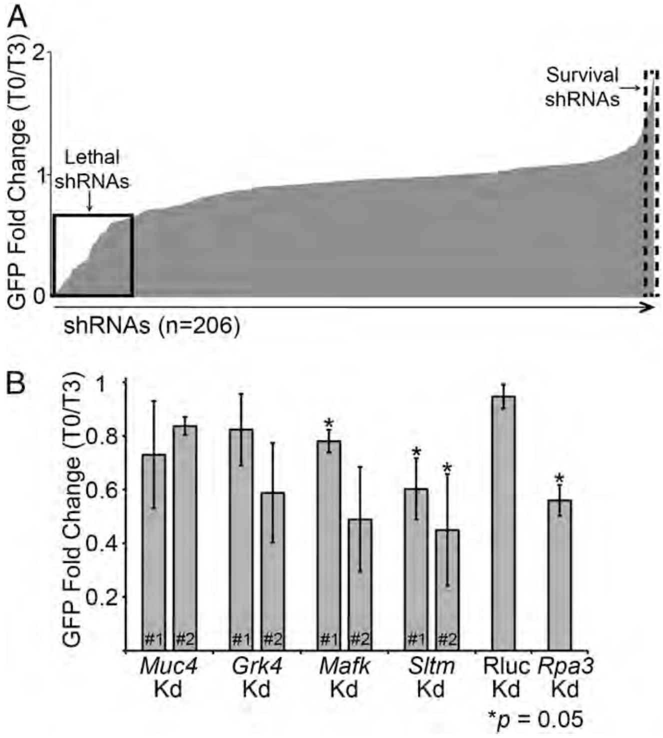


Fig. 4. Mammary gland focused screen. (*A*) One by one screen; 206 shRNAs, covering ~56 genes, were tested. The solid line square shows the fold change of shRNAs considered to be lethal, because GFP percent for these cells decreased overtime, whereas the dashed line square highlights data from shRNAs considered to show survival preferences in cells, because GFP percent increased overtime. (*B*) Screen hits validation. Two new hairpins targeting four genes selected as lethal hits from the first screen. The chart represents results of two independent experiments. *$P = 0.05$ by the *t* test.

the maintenance of GFP-expressing cells by either depleting GFP$^+$ cells (Fig. 4*A*, lethal shRNAs) or promoting expansion of GFP$^+$ cells (Fig. 4*A*, survival shRNAs) over time.

We decided to further investigate a subset of genes that interfered with Comma-Dβ growth, because our focus was to understand the spectrum of genes that might block normal mammary gland biogenesis. Among the selected genes were mucin-like gene (*Muc4*), G protein-coupled receptor gene family member (*Grk4*), and transcription factors (*MafK* and *Sltm*). An additional set of hairpins for these genes was rescreened in Comma-Dβ cells with GFP levels followed for 10 d. No clear effect on the percentage of GFP-positive cells was observed when cells expressed the new shRNAs against *Muc4* and Renilla luciferase control, whereas an shRNA-dependent response was observed, according to GFP frequency, when the genes *Grk4* and *Mafk* were targeted (Fig. 4*B*). In addition, both new shRNAs against the gene *Sltm* consistently decreased GFP-expressing cells to levels comparable with the depletion achieved by *Rpa3*, the lethal control. Interestingly, *Sltm* encodes a transcription factor-like protein that binds both DNA (scaffold attachment factor-box DNA binding motif) and RNA (RNA binding domain) in response to estrogen levels (22). We are currently investigating the implications caused by loss of *Sltm* expression during normal mammary gland development and tumorigenesis.

Discussion

The ongoing interest in stem cells and more recently, cancer stem cells highlights the need for improvements in purification and analysis of this rare but important population. Our previous understanding of MaSCs has been clouded by the limited capability to obtain a pure population devoid of contaminating, more differentiated cells. Here, we took advantage of a previously used system to identify relatively quiescent cells (6) in the mammary gland. We propose that the label-retaining cells from the K5tTa-H2b-GFP mouse represent a subset of active MaSCs, displaying increased mammary gland reconstitution ability over previously published cell populations identified as MaSCs.

Unlike previous methods, where cell selection is based on the presence of constitutive fluorescence in cells (3, 23), the use of a cell state-dependent GFP system allows for a more biological relevant fluorescence reliability. The extended time between halting GFP expression and analysis and also, selection of only the brightest cells decrease the possibility that the GFP protein might be detected in a cell cycling at a normal rate, despite the fact that the GFP expression is switched off. This system allowed for the possibility of a much more stringent selection process; however, it is acknowledged that there are limitations with using this mouse model on a routine basis for enriching for MaSCs. The need to use a transgenic mouse and however reduced, the level of heterogeneity within cells selected—evident by <100% regrowth efficiency—illustrate the need for the cell surface marker, CD1d, identified in our study.

Cd1d is known to be expressed as a cell surface marker on a variety of antigen-presenting cells belonging to a cluster of glycoproteins involved in T-cell antigen presentation (13). Because we physically remove all hematopoietic cells using magnetic beads before FACS, we are confident that this differentially expressed marker does not simply reflect contaminating cells. This statement is supported by the presence of CD1d$^+$ cells within the normal-like mammary gland cell line, Comma-1Dβ. In fact, 50% of these cells, isolated during midpregnancy, were positive for CD1d when stained with two distinct antibodies (Fig. S4*B*).

We, therefore, propose that CD1d is a genuine marker for MaSCs and when used combined with the cell surface markers CD24 and CD29, greatly enhances the purification of reconstituting cells above and beyond those cells selected based on label retention alone and those selected based on previously published markers. We perhaps did not exhaust all of the possibilities presented by our RNAseq data for the description of novel MaSC markers, but our findings do support CD1d as being a valuable component for purifying true MaSCs.

We found the proportion of CD1d$^+$ cells (1%) within the basal compartment to be greater than the proportion of H2b-GFP^h cells (0.2%) in this same compartment. This observation draws to light another drawback of relying solely on this particular label-retaining mouse model and in the same context, relying on GFP expression of a gene reporter mouse to identify MaSCs. The cytokeratin K5 (*Krt5*), for example, although shown to be expressed by basal-type cells, may not be expressed by all cells in this compartment. In addition, GFP expression may also be disrupted in some cells, perhaps by suppression of the transgene promoter; alternatively, some cells could fail to shut down GFP expression on DOX treatment. Had we only selected cells based on GFP expression from the K5 promoter, we would not be selecting all—or solely—those cells capable of self-renewal. This exclusion of MaSCs has been illustrated previously, where cells negative for a reporter GFP were able to still proliferate and regenerate into a new gland (3), something that we also see to a small degree with the K5tTa-H2b-GFP mouse model.

The identification of CD1d as a unique marker for this MaSC population and the distinct transcriptome of the Cd1d MaSCs suggest that these cells perform a function distinct from progenitors and more differentiated cells. Despite their gene expression profile clustering more closely with myoepithelial cells, they are still able to produce a new gland. It is unclear, however, if all of the CD1d MaSCs are multipotent stem cells or if they represent a combination of the recently described luminal and myoepithelial unipotent MaSCs (23). Because we are not using lineage tracing here, we cannot say for certain if all of the injected CD1d MaSCs would give rise to both compartments when allowed to repopulate the gland and if they are themselves the precursors to the cells that are largely responsible for tissue maintenance.

Identifying the genes involved in maintaining a stem cell and their self-renewing capabilities is vital to furthering our understanding of how these genes might be involved in abnormal gland development and tumorigenesis. Our current knowledge on this hypothesis, however, is, at best, limited; until now, it has been difficult to segregate myoepithelial cells and MaSCs, because they share common cell surface markers and very similar gene expression profiles (1, 24). The large number of shared genes expressed among cells identified by standard markers would mask any true differential patterns expressed by those cells with self-renewing properties. CD1d MaSCs cells only become divergent when the expression patterns of a relatively small number of genes are considered, a fact that would be overlooked if not using a more refined selection process. In addition to improving gene profiling as a whole for this minority population, the use of CD1d to isolate single cells for profiling could provide clues to gene expression changes between hypothesized MaSC states. For example, the complete loss of label-retaining cells after pregnancy suggests that these cells have undergone a more extensive process of cell division than in a virgin gland. However, CD1d$^+$ cells are still present, unaltered to some extent, and using this marker, it would be possible to monitor gene expression changes during pregnancy and involution.

It has been suggested that stem cells within the mammary gland contribute in some way to the proposed notion of a cancer stem cell. In mouse mammary tumor virus (MMTV)-Wnt1 and p53$^{-/-}$ mice, for example, a preneoplastic mammary gland was seen to have an increased number of functional MaSCs, and ectopic expression of wnt-1 enhanced the self-renewing capabilities of cells, leading to cancers (24). CD1d itself has even been linked to breast cancer. In one study, antibodies against CD1d, combined with anti death receptor 5 (anti-DR5), a TNF-related apoptosis inducing ligand (TRAIL) receptor, led to rejection of tumor growth after injection of 4T1 tumor cells, a mouse breast cancer cell line (25), into a syngeneic mouse fat pad (26). Whether this observation was a result of the proposed interaction with natural killer cells or a disruption of the ability of the cancer to self-renew remains to be seen. The latter could be possible, because we show that the 4T1 mouse breast cancer cell line (Fig. S4*C*) and primary mouse breast cells (27) (Fig. S4*D*) display a population of cells that is positive for Cd1d. In humans as well, CD1d plays some unknown role. Down-regulation of CD1d expression has been shown to correlate with increasing metastasis in a mouse breast cancer model (28) and disease progression in multiple myeloma (29). Our own studies have shown that CD1d is expressed as a cell surface marker in some but not all of the human breast cancer cell lines tested (Fig. S4*E*). Those cell lines that showed CD1d$^+$ cells were from basal-like breast cancers; luminal-like cancer cell lines, however, showed no CD1d$^+$ cells.

With the ability to now purify a more homogeneous self-renewing population of MaSCs, it is possible to delve more deeply into the biology of these cells. We have not only appointed CD1d as a marker of MaSCs but also used this information to draw out gene targets for disrupting mammary gland development and possibly, malignancies. It is also unknown yet if these specific cell markers are a cause or effect of the ability of the cell to retain stemness; interrupting their expression and studying the effect on gland development and cancer are critical topics of future study.

Materials and Methods

Mice. K5tTa-H2b-GFP heterozygote mice (6) were bred, and 20-d-old pups were checked for GFP expression using the IVIS100 in vivo imaging system (Caliper). CD-1 female mice were purchased from Charles River. Basal-like mouse mammary gland tumors were obtained from the transgenic mouse mammary tumor model C3-tag (27) (a gift from Mikala Egeblad, Cold Spring Harbor Laboratory, New York). All experiments were performed in agreement with and approved by the Cold Spring Harbor Laboratory Institutional Animal Care and Use Committee.

Two-Photon Microscopy. Mammary glands were harvested and defatted by three rounds of acetone treatment (20 min each). Defatted mammary glands were embedded and imaged according to previously published methods (30). In short, experiments were performed on a high-speed multiphoton microscope with integrated vibratome sectioning (TissueCyte 1000; TissueVision, Inc.). 3D scanning of 5-mm Z-volume stacks was achieved with a microscope objective piezo (PI E-665 LVPZT amplifier and P-725 PIFOC long-travel objective scanner), which translated the microscope objective with respect to the sample. Each optical section was imaged as a mosaic of individual fields of view equal to 0.83 × 0.83 mm and reconstructed posthoc using Fiji and custom-written Matlab software.

Antibodies. Antibodies for flow cytometry were purchased from eBioscience unless otherwise specified, and they include anti-CD24 eFluor@ 450, biotinylated and PE-conjugated anti-CD45, biotinylated and phycoerythrin (PE)-conjugated anti-CD31, biotinylated and PE-conjugated anti-Ter119, PE-Cy7–conjugated anti-CD29, FITC- and PE-conjugated anti-CD61, antigen-presenting cell-conjugated anti-CD133, PerCP-CY5.5– and PE-conjugated anti-Cd1d (clones 1B1 and K253, respectively; BioLegend), PE-conjugated anti-CD22, monoclonal CD59a (Hycult Biotech), PE-conjugated human anti-Cd1d, 7-AAD viability staining solution (BioLegend), FITC-conjugated mouse IgG, and PE-conjugated rabbit IgG. Antibodies for immunostaining were chicken anti-GFP (Invitrogen), mouse monoclonal Cytokeratin 18 (SCTB), anti-chicken–IgG-Alexa Fluor 647 (Invitrogen), and anti-mouse–IgG Alexa Fluor 568 (Invitrogen).

Mammary Gland Preparation. Mammary glands were harvested from young female mice (6–10 wk) and dissociated according to previously published protocol (1). After dissociation, cells were resuspended in 1 mL MACS Buffer (Myltenyi Biotech) and incubated with biotinylated anti-CD45, anti-Ter119, and anti-CD31 antibodies for 20 min. Cells were washed with 10 volumes MACS Buffer and further incubated with antibiotin magnetic microbeads (Myltenyi Biotech). Labeled cells were loaded into a magnetic column attached to a magnetic field (Myltenyi Biotech), and lineage-depleted flow-through cells were collected and further stained.

Flow Cytometry. Cells were stained for 30 min at 4 °C with antibody mix in PBS supplemented with 1% (vol/vol) FBS followed by wash with 10× volume PBS. Cells were resuspended in PBS plus 1% (vol/vol) FBS and further stained with 7-AAD immediately before sorting or analysis. Cells were sorted using a FACS ARIAII SORP (BD Bioscience). For cell analysis, LRSII (BD Bioscience) cell analyzer or MACSQuant (Myltenyi Biotech) were used. Data analysis was performed using either FloJo (Tree Star) or Diva (BD Bioscience).

Matrigel Colony Assay. Cells were sorted into 96-well plates containing 100 μL chilled 50% (vol/vol) Matrigel Matrix (BD Bioscience), further transferred to 100% Matrigel Precoated Chamber Slides (Lab-Tek), and incubated at 37 °C for 5 min. Complete Growth Media (1) was added to the chamber and renewed every other day for 10 d. Colonies were counted using Nikon Eclipse T1 microscope (Nikon).

Mammary Gland Transplant. Cells were sorted into 96-well plates containing 30 μL 50% (vol/vol) Growth Factor Reduced Matrigel (BD Bioscience) and 0.01% (vol/vol) Tripan Blue (Sigma). Cells were injected into inguinal glands of 3-wk-old females that had been cleared of endogenous epithelium. Recipient glands were removed for evaluation 8–12 wk after cell injection.

Mammary Transplant Analysis. Frozen sections and/or agarose-embedded sections were fixed with 4% paraformaldehyde (Sigma) for 20 min followed by tissue permeabilization and blocking using 10% (vol/vol) goat serum (Sigma). Paraffin-embedded sections were dewaxed and subjected to antigen retrieval for 15 min in Trilogy buffer (Cell Marque) before blocking as described above. Primary antibody staining was performed overnight at 4 °C with constant agitation followed by three washes with 0.1% (vol/vol) Tween 20. Secondary antibody staining was carried out for 45 min at room temperature with constant agitation followed by three washes with 0.1% (vol/vol) Tween 20. Slides were mounted with ProLong Gold supplemented with

DAPI (Invitrogen). For immunohistochemistry detection of GFP-positive outgrowths, the kit Ace IHC Detection Kit (Epitomics) was used according to the manufacturer's instructions. Tissue sections were analyzed using either the Nikon Eclipse T1 microscope (Nikon) or Zeiss LSM 710 confocal microscope. For whole-mount images, glands were harvested, spread atop a glass slide, defatted, and stained with Carmine Aluminum solution prior image analysis. MRU frequency was estimated using the ELDA algorithm (10). Mammary gland reconstitution was considered successful if, by the time of analysis, at least one-third of the fat pad was repopulated with GFP^+ structures.

RNAseq Library Preparation. Cells were sorted into Eppendorf tubes filled with TRIzol LS (Invitrogen), and RNA purification was performed according to the manufacturer's instructions. DNase-free RNA samples were used for the preparation of double-strand cDNA libraries using the Version 1 Ovation RNAseq System (Nugen). cDNA libraries were phosphorylated, adenylated, and ligated to Illumina adapters followed by PCR enrichment. Single-ended sequencing was performed for 36 cycles in Illumina GAII instruments (Illumina).

RNAseq Mapping and Analysis. We used the Refseq transcriptome (mm9 mouse assembly) downloaded through the University of California, Santa Cruz (USCS) Table Browser (31). Reads were mapped in two stages: first, they were mapped to sequences constructed using all annotated Refseq exons with overlapping exons collapsed, and second, they were mapped to all possible junctions formed from all pairs of exons for the same gene. Mapping was done with RMAP (32) and allowed up to three mismatches in 36 bases. Reads mapping ambiguously (including mapping to an exon and a junction) were discarded. For each Refseq transcript, we counted the number of reads with mapping location that was inside the transcript's exons (allowing a given read to be counted for two distinct transcripts as long as the location is unique) or through one of the transcript's junctions. Reads per kilobase per million (RPKM) calculations discarded duplicate reads and corrected gene size for the portion of the gene that cannot be uniquely mapped. Differential expression between two RNAseq experiments was computed using a 2 × 2 contingency table and either a χ^2 statistic or Fisher exact test to obtain a *P* value for differential expression. Briefly, the contingency tables contained, for each gene, the counts of reads mapping into the gene and the counts of reads mapping outside the gene for both experiments. The *P* values were corrected for multiple testing using the Bonferroni correction. The genes that remained were called as differentially expressed (corrected $P > 0.01$), and rankings for differentially expressed genes were based on ratios of RPKM values.

Quantitative PCR. Cells were sorted into 96-well plates containing 30 μL Cell-To-Ct lysis buffer (Ambion). cDNA synthesis was performed according to the manufacturer's instruction. Real-time PCR was performed using specific Taqman probes (Applied Biosystems) for each gene and *Gapdh* mRNA as an endogenous control. Samples were run on a 7900 Real-Time PCR System (Applied Biosystems).

BrdU Experiment. BrdU label-retaining experiments were performed using the BrdU-APC Flow Kit (BD Bioscience); 3-wk-old female mice were injected with BrdU (one time per day for 5 consecutive d, 50 mg/kg body weight), and mammary glands were harvest at specified time points. Mammary gland cells were prepared according to the BrdU manufacturer's recommendations. Cells were analyzed with an LSRII Cell Analyzer (BD Bioscience), and 1 million cells were recorded per sample. For each experiment ($n = 2$), three glands were analyzed at week 0 (last day of BrdU injection), and three glands were analyzed at week 12 after BrdU injection.

Knockdown Experiment. shRNAs against 56 selected genes were pulled and transferred from pGIPz (LMN vector) lentiviral backbone (Open Biosystems) to MSCV-miR30-PGK-NEO-IRES-GFP retroviral backbone (a gift from Christopher R. Vakoc, Cold Spring Harbor Laboratory, New York). Plasmid was transfected into Plat-E cells (33) using Lipofectamine 2000 and vesicular stomatitis virus g-protein (VSVG), and virus was collected 24 and 36 h posttransfection. Cells were infected by spin infection and allowed 2 d for recovery. GFP levels were measured using MACSQuant Cell analyzer (Miltenyi Biotech) from 10,000 cells. Hairpins used on validation experiments were ordered as oligonucleotides from Integrated DNA Technologies (IDT) and used as the template for PCR reactions using KOD hot-start polymerase (EMD Milipore) and the primers 5MIR (5′-CAGAAGGCTCGAGAAGGTATATTGCTGTTGACAGTGAGCG-3′) and 3MIR (5′-CTAAAGTAGCCCCTTGAATTCCGAGGCAGTAGGCA-3′). PCR products were column-purified (Qiagen), digested with EcoRI and XhoI enzymes, and cloned into predigested LMN vectors using T4 rapid ligase (Promega).

ACKNOWLEDGMENTS. We thank Elaine Fuchs, Mikala Egeblad, and Christopher R. Vakoc for valuable mouse stains and reagents. This work was supported by National Institutes of Health Grand Opportunity Award #1 RC2 CA148507 and P01 Award 2P01CA013106.

1. Shackleton M, et al. (2006) Generation of a functional mammary gland from a single stem cell. *Nature* 439(7072):84–88.
2. Stingl J, et al. (2006) Purification and unique properties of mammary epithelial stem cells. *Nature* 439(7079):993–997.
3. Bai L, Rohrschneider LR (2010) s-SHIP promoter expression marks activated stem cells in developing mouse mammary tissue. *Genes Dev* 24(17):1882–1892.
4. Tu Z, et al. (2001) Embryonic and hematopoietic stem cells express a novel SH2-containing inositol 5′-phosphatase isoform that partners with the Grb2 adapter protein. *Blood* 98(7):2028–2038.
5. Visvader JE, Lindeman GJ (2006) Mammary stem cells and mammopoiesis. *Cancer Res* 66(20):9798–9801.
6. Tumbar T, et al. (2004) Defining the epithelial stem cell niche in skin. *Science* 303(5656):359–363.
7. Mikaelian I, et al. (2006) Expression of terminal differentiation proteins defines stages of mouse mammary gland development. *Vet Pathol* 43(1):36–49.
8. Smalley M, Ashworth A (2003) Stem cells and breast cancer: A field in transit. *Nat Rev Cancer* 3(11):832–844.
9. Neville MC (2009) Introduction: Transplantation of the normal mammary gland: Early evidence for a mammary stem cell. *J Mammary Gland Biol Neoplasia* 14(3):353–354.
10. Hu Y, Smyth GK (2009) ELDA: Extreme limiting dilution analysis for comparing depleted and enriched populations in stem cell and other assays. *J Immunol Methods* 347(1–2):70–78.
11. Asselin-Labat ML, et al. (2008) Delineating the epithelial hierarchy in the mouse mammary gland. *Cold Spring Harb Symp Quant Biol* 73:469–478.
12. Molofsky AV, Pardal R, Morrison SJ (2004) Diverse mechanisms regulate stem cell self-renewal. *Curr Opin Cell Biol* 16(6):700–707.
13. Adams EJ, López-Sagaseta J (2011) The immutable recognition of CD1d. *Immunity* 34(3):281–283.
14. Qin X, et al. (2001) Genomic structure, functional comparison, and tissue distribution of mouse Cd59a and Cd59b. *Mamm Genome* 12(8):582–589.
15. Haas KM, et al. (2006) CD22 ligand binding regulates normal and malignant B lymphocyte survival in vivo. *J Immunol* 177(5):3063–3073.
16. Greenlee MC, Sullivan SA, Bohlson SS (2008) CD93 and related family members: Their role in innate immunity. *Curr Drug Targets* 9(2):130–138.
17. Becker-Herman S, Arie G, Medvedovsky H, Kerem A, Shachar I (2005) CD74 is a member of the regulated intramembrane proteolysis-processed protein family. *Mol Biol Cell* 16(11):5061–5069.
18. Medina D, Oborn CJ, Kittrell FS, Ullrich RL (1986) Properties of mouse mammary epithelial cell lines characterized by in vivo transplantation and in vitro immunocytochemical methods. *J Natl Cancer Inst* 76(6):1143–1156.
19. Danielson KG, et al. (1989) Clonal populations of the mouse mammary cell line, COMMA-D, which retain capability of morphogenesis in vivo. *In Vitro Cell Dev Biol* 25(6):535–543.
20. Deugnier MA, et al. (2006) Isolation of mouse mammary epithelial progenitor cells with basal characteristics from the Comma-Dbeta cell line. *Dev Biol* 293(2):414–425.
21. Ibarra I, Erlich Y, Muthuswamy SK, Sachidanandam R, Hannon GJ (2007) A role for microRNAs in maintenance of mouse mammary epithelial progenitor cells. *Genes Dev* 21(24):3238–3243.
22. Chan CW, et al. (2007) A novel member of the SAF (scaffold attachment factor)-box protein family inhibits gene expression and induces apoptosis. *Biochem J* 407(3):355–362.
23. Van Keymeulen A, et al. (2011) Distinct stem cells contribute to mammary gland development and maintenance. *Nature* 479(7372):189–193.
24. Lim E, et al. (2010) Transcriptome analyses of mouse and human mammary cell subpopulations reveal multiple conserved genes and pathways. *Breast Cancer Res* 12(2):R21.
25. Dexter DL, et al. (1978) Heterogeneity of tumor cells from a single mouse mammary tumor. *Cancer Res* 38(10):3174–3181.
26. Teng MW, et al. (2007) Combined natural killer T-cell based immunotherapy eradicates established tumors in mice. *Cancer Res* 67(15):7495–7504.
27. Green JE, et al. (2000) The C3(1)/SV40 T-antigen transgenic mouse model of mammary cancer: Ductal epithelial cell targeting with multistage progression to carcinoma. *Oncogene* 19(8):1020–1027.
28. Hix LM, et al. (2011) CD1d-expressing breast cancer cells modulate NKT cell-mediated antitumor immunity in a murine model of breast cancer metastasis. *PLoS One* 6(6):e20702.
29. Spanoudakis E, et al. (2009) Regulation of multiple myeloma survival and progression by CD1d. *Blood* 113(11):2498–2507.
30. Ragan T, et al. (2012) Serial two-photon tomography for automated ex vivo mouse brain imaging. *Nat Methods* 9(3):255–258.
31. Karolchik D, et al. (2004) The UCSC Table Browser data retrieval tool. *Nucleic Acids Res* 32(Database issue):D493–D496.
32. Smith AD, et al. (2009) Updates to the RMAP short-read mapping software. *Bioinformatics* 25(21):2841–2842.
33. Morita S, Kojima T, Kitamura T (2000) Plat-E: An efficient and stable system for transient packaging of retroviruses. *Gene Ther* 7(12):1063–1066.

Supporting Information

dos Santos et al. 10.1073/pnas.1303919110

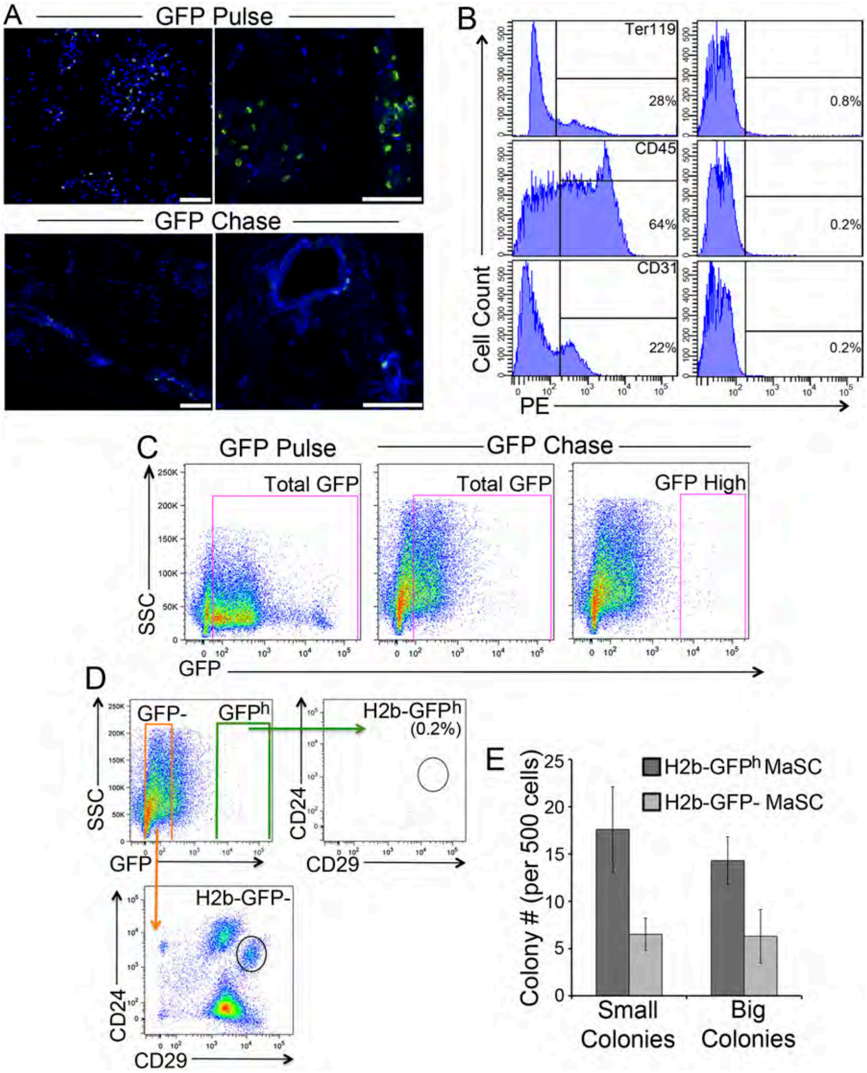


Fig. S1. Characterization of H2b-GFP mammary gland label-retaining cells. Effects of doxycycline diet on K5tTa-H2b-GFP transgenic mouse mammary glands. (*A*) Paraffin-embedded sections with DAPI nuclear staining and anti-GFP antibody. (*B*) Lineage depletion strategy. FACS analysis showing removal of PE-stained red blood cells (Ter119$^+$ cells), white blood cells (CD45$^+$ cells), and endothelial cells (CD31$^+$ cells) after magnetic bead lineage depletion. (*C*) H2b-GFP$^+$ cells gating strategy. Lin$^-$ mammary gland cells were first selected according to GFP expression (GFP$^-$ and GFP$^+$) and further analyzed according to anti-CD24 and anti-CD29 staining as displayed in Fig. 2. (*D*) FACS sorting strategy for transplantation assays. Lin$^-$ GFP chase cells, stained with 7-ADD for dead cell exclusion, were divided based on GFP expression, H2b-GFP$^-$ mammary gland stem cells (MaSCs; CD24$^+$CD29^{h}GFP$^-$), and H2b-GFP^h MaSCs (CD24$^+$CD29^{h}GFP^h) and either transplanted into cleared fat pads of prepubescent female mice or (*E*) carried through to colony-forming assays.

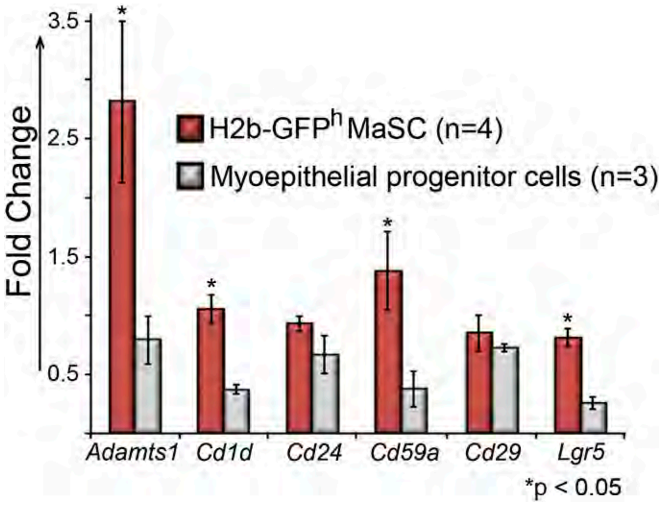


Fig. S2. Quantitative RT-PCR validation of mammary gland transcriptome. Lin^- mammary gland cells (n = 10) were sorted into lyses buffer for quantitative PCR using Taqman probes. *Cd24* and *Cd29* mRNAs were included as controls. Samples were normalized to the levels of *Gapdh* expression. Error bars represent SD among replicated. $^*P < 0.05$ by the *t* test.

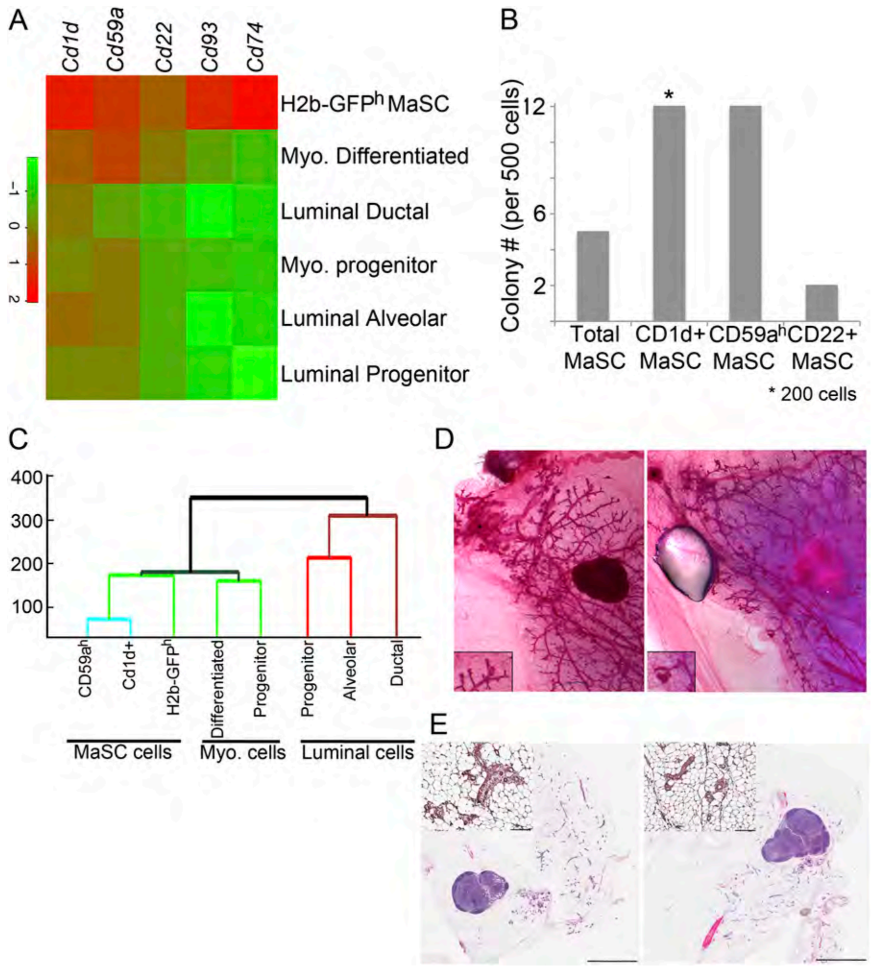


Fig. S3. Identification of MaSC cell surface markers. (*A*) Heat map of cell surface markers expression across all mammary gland cell types profiled. Those cell surface markers shown are the most abundantly expressed within the H2b-GFP^h MaSCs. (*B*) MaSC markers colony-forming assay. Cells were sorted using Cd24^+ Cd29^h alone (total MaSC) or $\text{Cd24}^+\text{Cd29}^\text{h}$ plus one of three markers: CD1d (CD1d MaSC), CD59a (CD59a^h MaSC), or CD22 (CD22 MaSC). *Two hundred cells. (*C*) Expression dendogram for the top most abundantly expressed genes across all mammary gland cells, including CD1d MaSCs and CD59a^h MaSCs. (*D*) Representative whole-mount images from mammary glands injected with Cd1d^+ MaSCs. *Scar tissue from cell injection. (*E*) Representative images from paraffin-embedded sections of Cd1d^+ MaSC-injected glands stained with H&E and anti-GFP immunohistochemistry (IHC) (*Left Inset*). (Scale bar: 2 mm; *Left Inset*, 100 μm.)

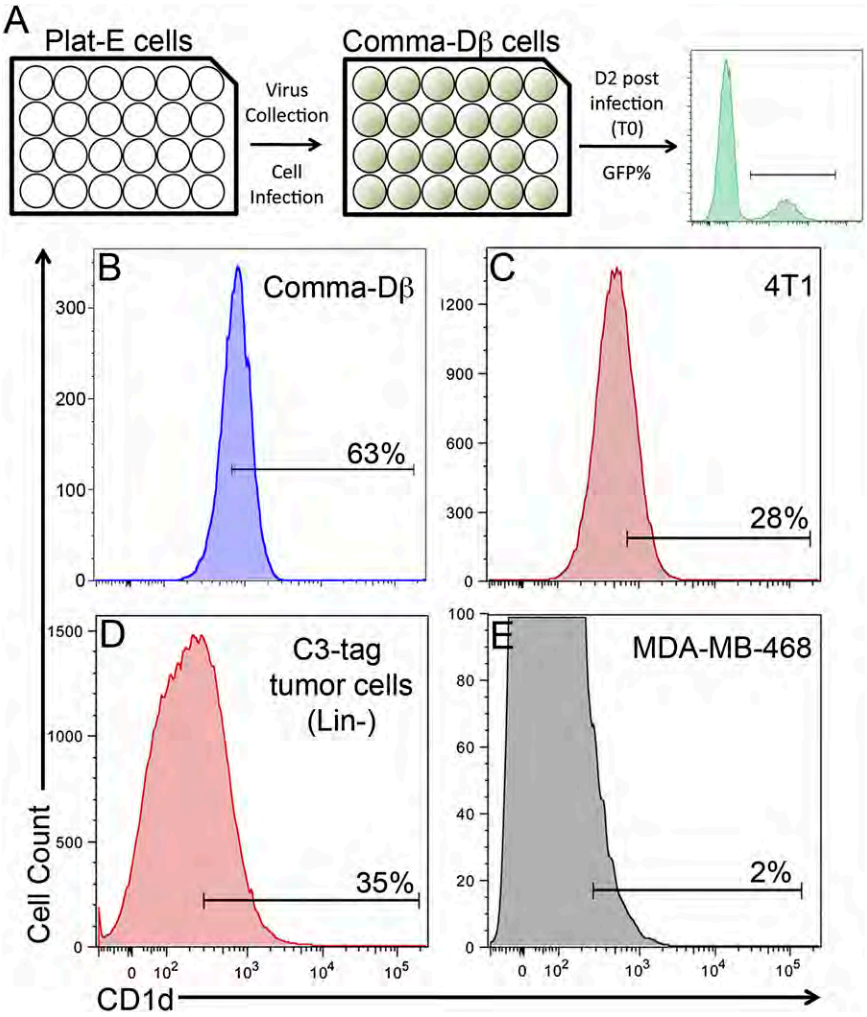


Fig. S4. Mammary gland-focused screen. (*A*) Screen strategy scheme. Plat-E cells were transfected with shRNAs as described in *Materials and Methods*. Comma-Dβ cells were infected with the virus supernatant for 20 h; 48 h postinfection, GFP percent was quantified using the MACSQuant Cell Analyzer (Miltenyi Biotech). T0 represents the GFP percent on day 2 postinfection, and T3 represents the GFP percent on day 12 d postinfection. CD1d$^+$ cells in mouse cell lines (*B*) Comma-Dβ cells, (*C*) 4T1 mouse breast cancer cells, and (*D*) C3-tag breast cancer model primary cells. (*E*) CD1d$^+$ in the human cell line MDA-MB-468.

Dataset S1. Mammary reconstitution unit (MRU) frequency in H2b-GFP^h MaSC

Dataset S1

Reconstituted mammary glands harvested 12 wk postinjection of either H2b-GFP^h MaSCs or H2b-GFP$^-$ MaSCs. A minimum of 25 outgrowths is required to be considered a reconstituted gland. MRU frequency was estimated using the ELDA algorithm.

Dataset S2. Mammary gland pathway analysis

Dataset S2

Top differentially expressed genes of all mammary gland cell types were analyzed for pathway enrichment and molecular functions using Ingenuity Pathways Analysis (Ingenuity Systems). A minimum of 50 genes per cell type was analyzed.

Dataset S3. MRU frequency in CD1d+ MaSC

Dataset S3

Total MaSC cells ($CD24^+CD29^h$) and CD1d MaSC cells ($CD24^+CD29^hCD1d^+$) were isolated from the H2b-GFP transgenic mouse off doxycycline diet (GFP pulse). Reconstituted mammary glands harvested 12 wk postcell injection. A minimum of 25 outgrowths is required to be considered a reconstituted gland. MRU frequency was estimated using the ELDA algorithm.

A computational algorithm to predict shRNA potency

Simon R.V. Knott[1], Ashley Maceli[1], Nicolas Erard[1], Kenneth Chang[1], Krista Marran, Xin Zhou, Assaf Gordon, Osama El Demerdash, Elvin Wagenblast, Christof Fellmann[&], and Gregory J. Hannon[*]

Watson School of Biological Sciences,
Howard Hughes Medical Institute
Cold Spring Harbor Laboratory
1 Bungtown Road
Cold Spring Harbor, New York 11724, USA

1 these authors contributed equally to this work
* to whom correspondence should be addressed (hannon@cshl.edu)
& present address: Mirimus, Inc., 1 Bungtown Road, Cold Spring Harbor, NY 11724

The strength of conclusions drawn from RNAi-based studies is heavily influenced by the quality of tools used to elicit knockdown. Prior studies have developed algorithms for the design of siRNAs. However, to date, no established method has emerged to identify effective shRNAs, which have lower intracellular abundance and undergo additional processing steps. We recently developed a multiplexed assay for identifying potent shRNAs and have used this method to generate ~250,000 shRNA efficacy data points. Using these data, we developed shERWOOD, an algorithm capable of predicting, for any shRNA, the likelihood that it will elicit potent target knockdown. Combined with additional shRNA design strategies, shERWOOD allows the *ab initio* identification of potent shRNAs that target, specifically, the majority of each gene's multiple transcripts. We have validated the performance of our shRNA designs using several orthogonal strategies and have constructed genome-wide collections of shRNAs for humans and mice based upon our approach.

Highlights

- An in-depth analysis of 250,000 shRNA sequence/efficacy data-points identifies key sequence characteristics for predicting shRNA potency

- An shRNA specific algorithm allows for significant increases in shRNA loss-of-function screen quality

- Structure-guided strategies allow for an expanded shRNA search space

- An alternative miR scaffold increases shRNA processing and potency

Introduction

The discovery of RNAi promised a new era in which the power of genetics could be applied to model organisms for which large-scale studies of gene function were previously inconvenient or impossible (Berns et al., 2004; Brummelkamp, Bernards, & Agami, 2002; Chuang & Meyerowitz, 2000; Fire et al., 1998; Gupta, Schoer, Egan, Hannon, & Mittal, 2004; Hannon, 2002; Kamath et al., 2003; Kambris et al., 2006; Paddison et al., 2004; Sanchez Alvarado & Newmark, 1999; Svoboda, Stein, Hayashi, & Schultz, 2000; Timmons & Fire, 1998; Tuschl, Zamore, Lehmann, Bartel, & Sharp, 1999; Zender et al., 2008). Yet, it quickly became clear that implementing RNAi, especially on a genome-wide scale, could be challenging. This was particularly true for applications in mammalian cells wherein discrete sequences, in the form of siRNAs or shRNAs, were used as silencing triggers (Brummelkamp et al., 2002; Elbashir et al., 2001; Paddison, Caudy, Bernstein, Hannon, & Conklin, 2002). The overall degree of knockdown achieved was found to vary tremendously, depending upon the precise sequence of the small RNA that is loaded into the RNAi effector complex (RISC) (Chiu & Rana, 2002; Khvorova, Reynolds, & Jayasena, 2003; Schwarz et al., 2003). Yet, the nature of sequence and structural motifs that favor RISC loading and high turnover target cleavage has yet to be fully revealed (Ameres & Zamore, 2013).

Early studies aimed at optimizing RNAi in mammals used endogenous microRNAs as a guide to the design of effective artificial RNAi triggers (Khvorova et al., 2003; Reynolds et al., 2004; Schwarz et al., 2003; Ui-Tei et al., 2004; Zeng & Cullen, 2003). Canonical microRNAs are processed by a two-step, nucleolytic mechanism (Seitz & Zamore, 2006). The initial cleavage of the primary miRNA transcript in the nucleus by the Microprocessor yields a short, often imperfect, hairpin loop, the pre-miRNA (Denli, Tops, Plasterk, Ketting, & Hannon, 2004; Lee et al., 2003). This is exported to the cytoplasm where a second cleavage by Dicer and its associated cofactors yields a short duplex of ~19-20 nucleotides with 2 nucleotide 3’ overhangs (Bernstein, Caudy, Hammond, & Hannon, 2001; Grishok et al., 2001; Hutvagner et al., 2001; Ketting et al., 2001; Lund, Guttinger, Calado, Dahlberg, & Kutay, 2004; Yi, Qin, Macara, & Cullen, 2003). This duplex serves as a substrate for preferential loading of one strand into Argonaute proteins in the context of RISC (Hammond, Boettcher, Caudy, Kobayashi, & Hannon, 2001; Hutvagner & Zamore, 2002; Khvorova et al., 2003; Martinez, Patkaniowska, Urlaub, Luhrmann, & Tuschl, 2002; Schwarz et al., 2003).

An examination of the sequences of endogenous miRNAs indicated that thermodynamic asymmetry between the two ends of the short duplex was a strong predictor of which strand would be accepted by Argonaute as the “guide” (Khvorova et al., 2003; Schwarz et al., 2003). The first few base pairs of native guide strands were generally less stable than were the terminal few base pairs. Note that the latter can also be considered as the more stable 5’ base pairs of the “passenger” strand that is not effectively loaded. Applying this insight to artificial triggers, initially in the form of siRNAs, validated the generality of this observation, and thermodynamic asymmetry became a key guiding principle of both siRNA and shRNA design (Reynolds et al., 2004; Silva et al., 2005). Subsequent studies of the structure of the Ago-small RNA complex have also indicated a sequence preference for a 5' terminal U that fits into a binding pocket in the mid domain of the Argonaute protein (Seitz, Ghildiyal, & Zamore, 2008; Wang, Sheng, Juranek, Tuschl, & Patel, 2008).

In many ways, siRNAs gain entry into RISC by simulating the end product of the two-step miRNA processing pathway. shRNAs, which mimic either the primary miRNA or pre-miRNA must be nucleolytically processed prior to RISC loading (Brummelkamp et al., 2002; Cullen, 2006; Paddison et al., 2002). Therefore, shRNAs are likely subject to additional constraints that lead to efficient recognition by Drosha and Dicer. Many transcripts can adopt structures that superficially resemble the short hairpin loops that give rise to miRNAs, yet only a tiny fraction of these efficiently give rise to small RNAs. We do not yet understand the selection rules for effective flux through the miRNA biogenesis pathway and therefore cannot predict *ab initio* what transcripts will produce small RNAs. However, studies of Drosha, in particular, have implicated patterns of conservation and base pairing in the basal stem, those regions adjacent to the Drosha cleavage site, as determinants of efficient pri-miRNA cleavage (Auyeung, Ulitsky, McGeary, & Bartel, 2013; Chen, Li, Lodish, & Bartel, 2004; Han et al., 2006; Seitz & Zamore, 2006). Elements within the hairpin loop have also been shown to have an impact both on Drosha efficiently and its site preference (Han et al., 2006; Zhang & Zeng, 2010).

Though many factors affect the efficiency with which a small RNA will be generated from a longer precursor and loaded into RISC, small RNA abundance, *per se*, is clearly not the only determinant of effective silencing. In fact, the overall rate of substrate cleavage by mature RISC seems to have a substantial impact. Initial biochemical studies indicated that RISC might be limited by the product release step (Hutvagner & Zamore, 2002). In fact, *Drosophila* Argonautes have been optimized by evolution for rapid product release (Ago2) or slow product release (Ago1) in a manner that correlates with their acting by an siRNA-based cleavage mode or an miRNA-based, cleavage-independent repression of protein synthesis (Ameres et al., 2010; Meister, 2013; Okamura, Ishizuka, Siomi, & Siomi, 2004). Moreover, in mammals, large-scale studies of silencing potency also point to the efficiency of mature RISC as the major determinant of effective repression (Fellmann et al., 2011; Reynolds et al., 2004). As yet, biochemical studies have not provided detailed guides to the design of small RNAs, which produce high RISC turnover rates.

Given this constraint, several attempts have been made to extract predictive rules for the design of effective small RNAs from endpoint silencing data. The first serious attempt applied Artificial Neural Networks (ANNs) to a set of ~2,000 paired data points associating the sequence of siRNA guides with a corresponding knockdown measurement (established using fluorescent reporters) (Huesken et al., 2005). Experience in the field supported the effectiveness of BIOPREDSi; however, access to the algorithm eventually became impossible. The same dataset was subsequently used to produce a second algorithm, DSIR, which included additional input variables (the frequency of each nucleotide, each 2mer and each 3mer within the guide) (Vert, Foveau, Lajaunie, & Vandenbrouck, 2006). To accommodate this large number of parameters, linear modeling was performed using Lasso Regression (a form of linear regression that iteratively decreases the use of non-predictive variables in the linear model) (Tibshirani, 1995).

siRNA design algorithms could be applied for the design of shRNAs, and these did inform the design of genome-wide shRNA collections (Berns et al., 2004; Paddison et al., 2004). However, the prognostic power of siRNA design algorithms is compromised for shRNA design. shRNAs, expressed from RNA polII or polIII promoters, reach lower intracellular concentrations than do transfected, synthetic siRNAs (Berns et al., 2004; Paddison et al., 2004). Moreover, shRNAs have additional constraints for effective processing. Therefore, it was imperative that shRNA-specific algorithms be developed.

The generation of accurate siRNA design algorithms was only made possible with the creation of large training datasets. Thus far, a corresponding shRNA dataset has been lacking. Recently, we developed a “sensor” method that allows for the parallel assessment of shRNA potencies on a massive scale (Fellmann et al., 2011). Using the sensor approach, we interrogated ~250,000 shRNAs for their effectiveness in the reporter setting. We have used this dataset to train a machine-learning algorithm for potent shRNA prediction. We have tested this algorithm, which we term, shERWOOD, both at the level of individual shRNAs and at the level of optimized shRNA mini-libraries. We have demonstrated that by applying computational shRNA selection in combination with novel target selection heuristics and with an optimized microRNA scaffold, we are able to create highly potent shRNAs. We have build upon this result to design and construct next-generation shRNA libraries targeting the constitutive exomes of mice and humans. Predictions for other organisms and custom shRNA designs are also made available via a web-based version of shERWOOD.

Results

Neighboring Positions of the Target Sequence are Predictive of ShRNA Strength

As a prelude to creating an shRNA design algorithm, it was necessary to generate a large “sensor” dataset in which shRNA potency was measured and associated with sequence information. To perform the assay, we synthesized constructs that include a doxycycline inducible shRNA and a GFP-tagged shRNA target sequence located downstream of a constitutive promoter (Fellmann et al., 2011). Libraries of ~25K constructs were packaged and infected (at single copy) into a reporter cell line. In the absence of doxycycline, GFP was detectable in each cell. However, in the presence of doxycycline the shRNAs became expressed and the resultant GFP signal was reduced in a manner proportional to shRNA potency. Using Florescence Activated Cell Sorting (FACS), cells with low GFP levels, in the presence of drug, were gathered and analyzed via NGS to determine which shRNAs became enriched (i.e. which shRNAs have high potency). Operating iterative cycles of this assay has been shown to identify extremely potent constructs (Fellmann et al., 2011).

We next wished to extract what variables (sequence characteristics) were most predictive of shRNA efficacy. This subset of characteristics could then be employed as inputs during machine learning. First, we calculated, for each of the ~250, 000 shRNAs tested via the sensor, a potency measurement. To define this measurement we constructed a matrix wherein rows correspond to shRNAs and columns represents the enrichment level of each shRNA at each iteration of the sensor (with respect to the initially infected shRNA population). From this matrix, the first principal components were extracted and averaged across replicates to give a final potency measurement. These values accurately capture the enrichment pattern of individual iterations of the sensor in one single value, thus allowing downstream machine learning to proceed more easily (Fig. 1A). Analysis of the coefficients used for principal component extraction shows that information from the final sensor iteration contributes the most to the final potency value, however information from the second iteration is also included, (Fig. S1A).

To distinguish discretely between strong and weak shRNAs, we applied an Empirical-Bayes Moderated T-Test to the replicate potency measurements for each shRNA. This allowed us to classify shRNAs as being either potent or weak (FDR < 0.05), which, in turn, allowed for standard statistical techniques to be employed to assess how various sequence characteristics stratified the two shRNA classes. To test individual nucleotide positions for their predictive capacity, we compared, at each position in the target sequence, each nucleotide’s enrichment and or depletion levels in the potent shRNAs with respect to the weak shRNAs (binomial-test p-value < 0.05). This task was performed separately for twelve non-overlapping sets of ~25K shRNAs that were analyzed separately via the sensor assay. The results from a representative analysis is shown in Figure 1B, and the reproducibility of results across all sets can be seen in Figure S1B. In general, low GC content in the nucleotides flanking the shRNA target site is predictive of high efficacy. This is also the case within the target region, with the exception of the third nucleotide, which shows a strong selection for cytosine. Also of note is a lack of enrichment for thymidine at the 22nd position of the target (corresponding to the first position of the guide). This bias arose because the majority of our input datasets were derived from shRNAs pre-selected by DSIR. Examination of unbiased sets of tiled shRNAs does not show a similar bias.

We next tested whether any pairs of positions within the target sequence had predictive capacity for shRNA strength, beyond that achieved by summing their individual predictive power. To calculate a measurement for each position pair, we applied linear regression to identify nucleotide combinations with synergistic predictive capacity (p-value < 0.05; see Supplemental Methods). Following this, each position-pair was assigned a value equal to the sum of nucleotide combinations that were predictive of shRNA potency when assessed at the two positions. Figure 1C shows the relative predictive power of position-pairs when all data is combined (for a corresponding analysis of the twelve shRNA subsets, see Figure S1C). For a given position within the target, the most predictive partner is the neighboring nucleotide. An exception to this trend is observed in the positions corresponding to the shRNA guide seed, where predictive position-pairs are also observed in nucleotides separated by up to four bases.

Finally, we wished to determine if triplets of positions showed a similar trend to that observed in the pair-wise analysis. For this, we performed a modified version of the linear regression tests described above, where triplets instead of pairs of nucleotides were assessed for synergistic predictive capacity. Figure 1D demonstrates that, as with the pairwise analysis, neighboring triplets of positions within the target show strong predictive power as compared to triplets of non-neighboring positions. Also, as with the pairwise analysis, the distance between predictive triplets is extended slightly in the guide seed region of the shRNA.

A Sensor-Based Computational Algorithm to Predict shRNA Efficacy

Since sequence-based characteristics correlated with shRNA efficiency, we sought to apply machine learning to the sensor-derived efficacy measurements. The goal was to develop a computational algorithm that would predict, for any target sequence, the potency of a corresponding shRNA. We reasoned that the best machine-learning tool to apply to this task was Random Forest Regression Analysis. The reasons for this decision were two-fold. First, there is no decrease in the accuracy of Random Forests when the number of input variables is large. Second, the architecture of the algorithm takes into account increases in accuracy that can be achieved by analyzing combinations of input variables.

Our training dataset was of two distinct types. One comprised an unbiased set of shRNAs that tiled every nucleotide of 9 genes (Fellmann et al., 2011). A second comprised a larger set of shRNAs corresponding to ~18,000 genes and representing the top 12 DSIR predictions for those genes. We therefore chose to separate data corresponding to each input class and to train separate forests. We also chose to separate data based upon the 5' nucleotide of the guide. This was done for two reasons. First, previous studies, supported by structural insights, had suggested that the 5' nucleotide of the guide was a prominent determinant of small RNA potency (Fellmann et al., 2011; Frank, Sonenberg, & Nagar, 2010; Khvorova et al., 2003; Reynolds et al., 2004). Therefore training forests individually for shRNAs initiating with each base focused the prediction process on additional determinants. Moreover, the DSIR-based predictions were already heavily biased toward U and A at the 5' position. In fact, the bias was so strong that we did not have sufficient data to train 5'C and 5'G forests for these datasets. This meant that, in the first pass, we trained six independent modules.

In each module, input data were composed of individual base information as well as all neighboring pairs of bases throughout the guide sequence. In addition, the set of triplet-position/nucleotide-combinations found to be predictive, as assessed by linear regression, were also included (Figure 1D). To consolidate these modules, a second-tier random forest was trained using the first tier outputs, the corresponding shRNA-guide 5' base information, and a set of thermodynamic properties extracted from each shRNA (e.g. enthalpy, entropy). See Figure S2A for a schematic representation of the algorithm and Table S1 for the list of thermodynamic properties included during training. We name the compiled algorithm, shERWOOD.

To test the prognostic power of shERWOOD, we took advantage of the unbiased nature of the tiled shRNA sensor data (Fellmann et al., 2011) (Fig. 2A). For each of the 9 genes represented, we independently trained a shERWOOD algorithm without the data corresponding to that gene. We could then test shERWOOD performance against experimental data in a manner that was not skewed by the use of that data for training. We saw an overall Pearson correlation of 0.72 between experimentally derived potency measurements and computational predictions, with the prediction separated according to the first base of the guide. When we consider only shRNAs with a 5' U, the correlation rises to 0.78, likely due to the greater number of data point available for training that algorithm. For comparison, DSIR achieves a correlation of 0.4 and a prior shRNA prediction algorithm trained on a subset of the sensor data used in this study achieves 0.56 (Matveeva, Nazipova, Ogurtsov, & Shabalina, 2012; Vert et al., 2006). This indicates that shERWOOD achieves a roughly 180% increase in performance over currently existing siRNA prediction algorithms and a 126% increase in efficacy over existing shRNA specific prediction algorithms.

We have supplemented shERWOOD with additional heuristics to maximize the probability of successfully reducing protein levels in most cell and tissue types.

The complex nature of alternative splicing patterns provided a strong motivation for directing shRNAs against constitutive exons. We therefore developed a strategy that iteratively searches for regions within a gene that are shared by at least 80% of transcripts (see Supplemental Methods). This algorithm also tests whether high potency shRNAs have the potential to co-suppress paralogous genes, in order to minimize such off target effects. Considered together, these strategies have the potential to maximize the probability of biologically meaningful results from studies using shRNAs.

Benchmarking shERWOOD

To assess the performance of the shERWOOD algorithm, we felt that it was necessary to test a large number of shRNAs for their biological effects, as one can find anecdotal evidence for excellent performance for nearly any algorithm or strategy. We therefore chose ~2,200 genes based upon their enrichment in gene ontology (GO) categories likely to impact the growth and survival of cells in culture (Fig. 2B). As controls, particularly for the likelihood of off-target effects, we included 400 olfactory receptor genes. Olfactory receptors are expressed only in olfactory neurons, and even then, they display allelic choice so that only one paralog is expressed per cell. Thus, shRNAs targeting olfactory receptors are highly unlikely to have relevant, on-target biological effects in any cell line screened in vitro. To benchmark the performance of shERWOOD, we compared a focused, mini-library predicted with this algorithm to two widely used genome-wide collections, namely the TRC collection distributed by Sigma-Genosys and the so-called Hannon-Elledge v3 library distributed presently by Thermo-Fisher (Chang et al., unpublished). To produce the shERWOOD-based library and a deeper simulation of the v3 library, we used either shERWOOD or DSIR to predict their top 10 scoring shRNAs for our test genes. The sequences of TRC shRNAs are listed on a public web portal and we selected all listed shRNAs for each gene. In the case of TRC shRNAs, it was necessary to adapt them to a 22bp stem for placement into the miR-30 context.

For each test library, we synthesized 27,000 oligonucleotides in solid phase on microarrays (Cleary et al., 2004). These were cleaved, amplified, and cloned directly into a miR-30 scaffold within an MSCV-based retroviral vector without sequence validation. In this arrangement, the primary shRNA was transcribed from the LTR promoter while GFP and Neomycin resistance were separately expressed as a bicistronic transcription unit from the Phosphoglycerate Kinase promoter (PGK; Fig. S2D). Pilot sequencing showed that each library was of similar quality and representation.

Each library was infected separately into the pancreatic ductal adenocarcinoma cell line, A385. Two days after infection, cells were collected for a reference time-point, and after ~12 doublings cells were again harvested for a final time-point (see Supplemental Methods). shRNA representation was determined following amplification of hairpin inserts from genomic DNA (Sims et al., 2011), and after processing, shRNA read counts were compared between the initial and final time-points. Corresponding log-fold changes were normalized to remove GC bias (see Supplemental Figure S2E,F and Supplemental Methods). Normalized log-ratios were then analyzed using an Empirical-Bayes Moderated-T-Test to call significantly enriched and depleted shRNAs (FDR<0.05; Figure S2G).

To enable direct comparisons between libraries, we censored the shERWOOD and DSIR-based libraries on a per gene basis to contain the same number of hairpins as were available in the TRC library, keeping those with the best algorithmic scores. We then selected the consensus set of "essential" genes, accepting only those where at least two hairpins in each library passed the statistical threshold (FDR<0.05). As expected, the resulting set of genes that were important for the growth and survival of A385 were strongly depleted of olfactory receptor shRNAs (Fig. 2C). An inspection of the olfactory shRNAs that depleted found them to be common to all libraries. None mapped to any alternative genomic locations. In contrast, the set of consensus essential genes was enriched for GO terms associated with translation and replication.

To benchmark shRNA selection strategies against each other, we determined the percentage of shRNAs in each mini-library that scored for each consensus essential gene. For the TRC library 22% of shRNAs achieved significant depletion, whereas 26% of DSIR-predicted sequences and 31% of shERWOOD-based hairpins scored (Fig. 2D). We also considered performance from the perspective of median log-fold depletion. For the TRC collection the average log-fold change was -0.08; for DSIR this rose to -0.11, and it increased further to -0.14 for shERWOOD shRNAs (Fig. 2E). We note that this type of analysis favors slightly the library with the weakest overall shRNAs, since it will be this collection that sets entry criteria for the consensus essential gene set.

To assess whether shERWOOD scores were a proxy for shRNA potency, we examined the relationship between shERWOOD score and the probability of being significantly depleted for each consensus essential gene. For this, we analyzed all 10 shERWOOD predictions using a sliding scale of shERWOOD score cut-offs (Fig. 2F). As examples, considering shRNAs with a score greater than 0.5, the likelihood that an shRNA will be depleted if it targets one of our consensus essential genes is 33%. This rises to 39% for shRNAs with a score greater than 1. Again, this underestimates the information content of shERWOOD scores since in the cumulative plot shown, the minimum number of scoring hairpins for a given gene irrespective of scores is 2 (i.e., 20%).

Considered together, these data demonstrate that the use of shERWOOD can maximize the probability of obtaining potent shRNAs, increasing the number of hairpins that are called significant for any given gene and increasing also the penetrance of resulting phenotypes. By selecting hairpins with the highest shERWOOD scores, one can both increase confidence in the results of large-scale screens and maximize the probability that a given hairpin will effectively silence its target and produce a phenotype in validation studies.

Structure-guided insights expand the shRNA prediction space

Regardless of the accuracy of predictive models, we sometimes found it difficult to identify potent shRNAs due to search space restrictions imposed by sequence constraints (e.g. GC content), gene length, or the complexity of alternative splicing patterns. We therefore sought ways to expand the sequence space to which we could apply the shERWOOD approach. Analysis of miRNA seed sequences as well as other data have suggested that the first base of the small RNA guide does not pair with its target (Lai, 2002; Lewis, Burge, & Bartel, 2005; Yuan, Pei, Chen, Tuschl, & Patel, 2006). Structural studies have supported this hypothesis by showing that the first base of the guide is tightly bound within a pocket in the mid domain of Ago proteins (Fig. S3A) (Elkayam et al., 2012; Frank et al., 2010; Nakanishi, Weinberg, Bartel, & Patel, 2012; Wang et al., 2008). Moreover, there is a clear structural basis for the U preference at this site. Since the first base of the guide is a strong contributor to shRNA efficacy, we reasoned that we could expand the range of possible effective shRNAs by simply changing the first base of all potential guides to a U, promoting their binding to RISC and theoretically not altering target site choice. We will henceforth refer to this as the 1U-strategy. A simulated construction of a human genome-wide shRNA library demonstrates that, when this strategy is implemented, predicted shRNA-potencies increase dramatically, particularly for short GC rich genes (Fig. S3B).

To test the 1U-strategy in a high-throughput manner, we constructed a sensor library where the top 15 shRNAs targeting a set of ~2000 “druggable” genes were predicted using the canonical genome or the 1U-strategy. The constructs were designed such that the shRNAs contained the 1U-conversion and the target sites contained the endogenous base. shRNA potencies were extracted as described for Figure 1, and these are plotted in Figure 3A along with the mean potencies of a set of weak, fair, good and very good shRNA controls (with mean knockdown efficiencies of 25%, 50%, 75% and >90%, respectively). The distribution indicates that ~50% of the shRNAs were strong or very strong (knockdown efficiency >75%). When shRNAs were separated into native and artificial 1U sets and the score distributions were plotted, we were surprised to see a significant reduction in the efficacy of the non-native-1U shRNAs (p-value < 0.01). This was strongly suggestive that RISC interacts not only with the 1U of the guide but also with the first base of the target site.

We therefore stratified 1U shRNAs into four sets based on their endogenous 5’ nucleotide (Fig. 3C). This analysis indicated that only a subset of shRNAs perform well when a 1U-switch is made (based on the bi-modal distributions for endogenous 1A, 1C and 1G shRNAs), but the subset that do perform well are predicted to be quite efficacious by the sensor assay. This bimodal distribution is not observed for shERWOOD-selected endogenous 1U shRNAs and we see that the majority of this shRNA class are efficient (75% have sensor scores higher than the “good” shRNA controls).

Given these results, we sought to determine whether we could predict those sequences for which a 1U conversion would result in a highly effective shRNA. We fit a Gaussian-mixture model to the sensor scores (Fig. S3C) and applied this model to assign shRNAs into one of the two resultant populations (Fig. S3D). Following clustering, we applied a binomial test separately for shRNAs where the endogenous base was 1A, 1C, 1G and 1U to determine if any nucleotides were enriched/depleted in the strong shRNAs with respect to weak shRNAs. As can be seen in Figure 3F, all sets show a strong enrichment for U in the target region corresponding to the shRNA guide positions 3, 7 and 8. There is also a strong selection for Cs in the target region corresponding position 19 of the endogenous 1A, 1C and 1G shRNA guides.

These results prompted us to develop a computational algorithm that could both select the strongest endogenous 1U shRNAs and identify which endogenous 1C, 1G and 1A shRNAs were likely to yield potent 1U-converted molecules. Data points for which the mixed-Gaussian clustering resulting in less than a 70% confidence group assignment were censored (Fig. S3E). We trained a random forest using the 22 nucleotides of the endogenous base as well as all neighboring pairs of nucleotides as input and the corresponding 1U-conversion sensor scores as output. The algorithm was able to achieve 80% specificity while maintaining 50% sensitivity. Notably, we were able to increase the specificity to 85% through the supplemental application of previously reported rules for shRNA selection (Fig. 3E)(Fellmann et al., 2011; Matveeva et al., 2012).

To validate this addition to the shERWOOD algorithm, we performed an shRNA screen as described above, wherein shRNAs were selected with the 1U-strategy with or without applying the additional filter. We also applied the new variant of the algorithm to shRNA screen described for Figure 2. We found that when additional filters were applied to the 1U strategy, shRNAs targeting our set of consensus essential genes showed a significantly higher percentage of depleted shRNAs per gene (p<0.01) and a stronger mean depletion as measured by log ratio (p<0.01; Fig. 3F).

A variant miRNA scaffold increases shRNA potency

Studies of evolutionarily conserved determinants of Drosha processing raised the possibility that the placement of the EcoRI site in the standard miR-30 scaffold might have reduced the efficiency of pri-miRNA cleavage (Auyeung et al., 2013). Others have reported that alternatively positioning the EcoRI site within the scaffold increases small RNA levels, presumably by improving biogenesis. This led to overall more potent knockdown (Fellmann et al., 2013). We therefore chose to create shRNAs by Gibson assembly, thus removing restriction sites altogether from the shRNA scaffold. We felt that this was the safest way to avoid any unanticipated impacts of altering processing signals (Fig. 4A). In addition, based on the rates at which intermediate substrates are transformed into E. coli, library transfer via the Gibson assembly is roughly 10-fold more efficient than with traditional cloning methods. We termed this scaffold, ultramiR.

To test ultramiR performance, we inserted two shRNAs, targeting luciferase or mouse RPA3, into the standard scaffold and into ultramiR. These constructs were packaged and infected in duplicate (MOI < 0.3) into the modified DF1 fibroblast cell line (Gallus gallus) that we employ as a reporter line for the sensor screen (Fellmann et al., 2011). Following selection for singly infected cells, we analyzed levels of mature shRNAs by small RNA sequencing (Malone, Brennecke, Czech, Aravin, & Hannon, 2012). shRNA guide counts were normalized across libraries by determining their log-fold enrichment relative to the median count of the ten most highly expressed microRNAs. A comparison of the normalized shRNA values indicated that, when shRNAs were placed into the ultramiR scaffold, mature small RNA levels were significantly increased relative to levels observed using the standard miR-30 scaffold (roughly two-fold; Fig. 4B). Notably, the performance of ultramiR and the previously described alternate scaffold, miR-E, were indistinguishable (not shown).

To provide a more rigorous test of ultramiR performance, we created a variant of shERWOOD-selected 1U-strategy shRNA library, as described above, and compared its performance to that of the same library in the standard scaffold. Considering the consensus essential gene set, nearly half of all shRNAs in the library were significantly depleted (Fig. 4C). This substantial improvement (from 38% to 47%, p<0.01) was accompanied by a greater degree of mean log-fold depletion for each construct (from -0.13 to -0.18, p<0.01).

We also tested a limited number of individual shRNAs for their potency by target knockdown. We selected the four shRNAs with the highest shERWOOD scores for mouse Mgp, Serpine2 SerpinE2 and Slpi. These were cloned into an MSCV-based ultramiR vector wherein hygromycin resistance and mCherry were also expressed as a bicistronic transcript from the PGK promoter. Mouse 4T1 cells were infected at single copy and knockdown was tested following selection of infected cells. Every one of the shRNAs tested reduced target mRNA levels by over 80%, and the vast majority reduced target mRNA levels by more than 90% (Fig. 4D). Considered together, our data indicate that the combined use of shERWOOD and the ultramiR scaffold consistently produces highly potent shRNAs.

Discussion

The application of RNAi in mammalian cells promised a revolution in understanding gene function and in the discovery and validation of therapeutic targets. While the impact of RNAi has been enormous, there have also been substantial frustrations in attempts to fully realize the potential of this technology. Many different sequences often need to be tested in order to obtain one that potently suppresses expression, a problem that is particularly acute with shRNAs expressed from single-copy transgenes. This, and the resulting variability in the quality of publicly available genome-wide shRNA collections, has caused consternation, particularly when very similar shRNA screens carried out by different investigators yield largely non-overlapping results (Babij et al., 2011; Luo et al., 2009; Scholl et al., 2009). We have tried to address problems with current shRNA technologies both by optimizing target sequence choice and by optimizing small RNA production.

We have leveraged our prior development of a high-throughput assay for testing shRNA potency to develop a computational algorithm capable of accurately predicting the outcome of the sensor screen and in turn predicting potentially potent shRNAs. Though iterative cycles of training and refinement, we have produced a tool that permits highly efficacious shRNAs to be generated for nearly any gene.

We have validated the performance of our approach and benchmarked it against current tools using non-sequence verified, focused shRNA libraries. Based upon our analyses, we can now generate shRNA libraries where nearly 60% of all hairpins targeting essential genes are strongly depleted in multiplexed screens. This means that for any library containing on average 4 hairpins per gene, most bona fide hits will be identified by multiple hairpins, greatly reducing the probability of false-positive calls. Since our libraries were used in their raw form, we feel that this is a lower boundary of performance, since sequence-validated and arrayed collections will not contain a mixture of shRNA variants generated by synthesis and PCR errors.

Given the promise of our approach, we have undertaken the construction of fourth- and fifth-generation, sequence-verified shRNA libraries targeting the mouse and human genomes. The fourth generation toolkit takes advantage of shERWOOD in a canonical miR-30 scaffold and currently comprises over 75,000 shRNAs targeting human genes and 40,000 shRNAs targeting mouse genes. The fifth generation toolkit places shERWOOD shRNAs in the ultramiR scaffold and is presently ~50% complete.

We have predicted shERWOOD shRNAs targeting constitutive exons of annotated mouse and human protein coding genes, and these are available via a web portal. We have also broadened predictions to rat, *Drosophila*, and canine genes. We have additionally made shERWOOD available as a web-based tool for custom shRNA prediction, for example for the design of shRNAs for other model organisms or for specific mRNA isoforms or non-coding RNAs.

Overall, we feel that the combination of improvements to shRNA technologies described herein creates a next-generation RNAi toolkit that will produce more reliable outcomes for investigators, whether applied on a gene-by-gene basis or in the context of unbiased, genome-wide screens.

Experimental Procedures

Vectors and Library Construction

The vector used for the sensor assay was the same as reported in (Fellmann et al., 2011). All RNAi screens and small RNA cloning experiments were performed with an MSCV-based retroviral vector harboring a bi-cistronic transcript (eGFP-IRES-Neomycin) downstream of the PGK promoter. Single target knockdown experiments were performed with a similar vector where Neomycin is replaced with Hygromycin and eGFP is replaced with mCHERRY.

To ensure high complexity end products, all shRNA libraries were amplified from 16 separate 1 ul 100 uM aliquots of input material using 22 PCR cycles. All transformations of were performed with Invitrogen’s MegaX DH10B T1 Electrocomp cells using a Biorad Gene Pulser Xcell and Biorad Gene Pulser 1mm cuvettes for electroporation. For each library a minimum of 5M successfully transformed cells were obtained.

Cell Lines

The sensor algorithm was performed using ERC cells (derived from DF-1 chicken embryonic fibroblasts (Fellmann et al., 2011). All shRNA screens were performed in the pancreatic adenocarcinoma cell line A385 (Cui et al., 2012). All small RNA cloning was performed in the ERC cell line. Individual shRNA knockdown experiments were performed in the 4T1 murine mammary cancer cell line (Dexter et al., 1978).

Library Preparation for High-Throughput Sequencing.

Sensor assays and RNAi screen preparation involved a two-step PCR process. For the first step (to maximize library complexity) ~200 ug of genomic DNA was PCRed using primers that were fully complementary to regions flanking the shRNA. These PCRs were performed in 96 well plates with 2 ug input material in each well. The second PCR step involved addition of illumina adapter sequences. One primer was composed of illumina’s P5 sequence followed by a small insert and then the shRNA loop. The small insert was designed such that it in combination with the shRNA loop had the same GC content and melting temperature as Illuminas SBS3 sequencing primer. The other primer contained

Illumina’s P7 sequence and then a region complementary to the PGK promoter. For each PCR step 25 cycles were performed.
All small RNA cloning libraries were constructed using Illumina’s small RNA cloning kit.

The twelve DSIR based sensor experiments were sequenced on Illumina’s GAII High-Throughput sequencing system. The shERWOOD sensor assay and all RNAi screens were sequenced on an Illumina Hi-Seq High Throughput sequencing system. Small RNA sequencing was performed on an Illumina Mi-Seq High Throughput sequencing system.

Author Contributions

S.R.V.K and G.J.H designed the experiments and wrote the manuscript. S.R.V.K designed the algorithm. A.M performed all shRNA screens. A.M. and X.Z. performed the 1U-sensor. N.E. performed all small RNA cloning. S.R.V.K, A.M. and N.E. constructed the sequence verified libraries. K.C. and K.M. performed the DSIR-sensors. S.R.V.K and A.G. developed and implemented the exom-inclusion and off-target minimization strategies. A.G. and O.E.D designed the shERWOOD website. E.W. performed the individual knockdown experiments.

Acknowledgments

This work was supported by the Howard Hughes Medical Institute as well as grants from the NIH (G.J.H.) and a kind gift from Kathryn W. Davis. S.R.V.K is supported by a fellowship from The Hope Funds For Cancer Research. E.W. is supported by a fellowship from the Boehringer Ingelheim Foundation. We thank all members of the Hannon lab for help with initial library construction. We also like to thank Blake Simmons (Transomics Technologies) for help in constructing sequence verified versions of the libraries. Finally, we thank Vasily Vagin for helpful comments on this manuscript.

References

Ameres, S. L., Horwich, M. D., Hung, J. H., Xu, J., Ghildiyal, M., Weng, Z., & Zamore, P. D. (2010). Target RNA-directed trimming and tailing of small silencing RNAs. *Science, 328*(5985), 1534-1539. doi: 10.1126/science.1187058

Ameres, S. L., & Zamore, P. D. (2013). Diversifying microRNA sequence and function. *Nat Rev Mol Cell Biol, 14*(8), 475-488. doi: 10.1038/nrm3611

Auyeung, V. C., Ulitsky, I., McGeary, S. E., & Bartel, D. P. (2013). Beyond secondary structure: primary-sequence determinants license pri-miRNA hairpins for processing. *Cell, 152*(4), 844-858. doi: 10.1016/j.cell.2013.01.031

Babij, C., Zhang, Y., Kurzeja, R. J., Munzli, A., Shehabeldin, A., Fernando, M., . . . Dussault, I. (2011). STK33 kinase activity is nonessential in KRAS-dependent cancer cells. *Cancer Res, 71*(17), 5818-5826. doi: 10.1158/0008-5472.can-11-0778

Berns, K., Hijmans, E. M., Mullenders, J., Brummelkamp, T. R., Velds, A., Heimerikx, M., . . . Bernards, R. (2004). A large-scale RNAi screen in human cells identifies new components of the p53 pathway. *Nature, 428*(6981), 431-437. doi: 10.1038/nature02371

Bernstein, E., Caudy, A. A., Hammond, S. M., & Hannon, G. J. (2001). Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature, 409*(6818), 363-366. doi: 10.1038/35053110

Brummelkamp, T. R., Bernards, R., & Agami, R. (2002). A system for stable expression of short interfering RNAs in mammalian cells. *Science, 296*(5567), 550-553. doi: 10.1126/science.1068999

Chen, C. Z., Li, L., Lodish, H. F., & Bartel, D. P. (2004). MicroRNAs modulate hematopoietic lineage differentiation. *Science, 303*(5654), 83-86. doi: 10.1126/science.1091903

Chiu, Y. L., & Rana, T. M. (2002). RNAi in human cells: basic structural and functional features of small interfering RNA. *Mol Cell, 10*(3), 549-561.

Chuang, C. F., & Meyerowitz, E. M. (2000). Specific and heritable genetic interference by double-stranded RNA in Arabidopsis thaliana. *Proc Natl Acad Sci U S A, 97*(9), 4985-4990. doi: 10.1073/pnas.060034297

Cleary, M. A., Kilian, K., Wang, Y., Bradshaw, J., Cavet, G., Ge, W., . . . Hannon, G. J. (2004). Production of complex nucleic acid libraries using highly parallel in situ oligonucleotide synthesis. *Nat Methods, 1*(3), 241-248. doi: 10.1038/nmeth724

Cui, Y., Brosnan, J. A., Blackford, A. L., Sur, S., Hruban, R. H., Kinzler, K. W., . . . Eshleman, J. R. (2012). Genetically defined subsets of human pancreatic cancer show unique in vitro chemosensitivity. *Clin Cancer Res, 18*(23), 6519-6530. doi: 10.1158/1078-0432.ccr-12-0827

Cullen, B. R. (2006). Induction of stable RNA interference in mammalian cells. *Gene Ther, 13*(6), 503-508. doi: 10.1038/sj.gt.3302656

Denli, A. M., Tops, B. B., Plasterk, R. H., Ketting, R. F., & Hannon, G. J. (2004). Processing of primary microRNAs by the Microprocessor complex. *Nature, 432*(7014), 231-235. doi: 10.1038/nature03049

Dexter, D. L., Kowalski, H. M., Blazar, B. A., Fligiel, Z., Vogel, R., & Heppner, G. H. (1978). Heterogeneity of tumor cells from a single mouse mammary tumor. *Cancer Res, 38*(10), 3174-3181.

Elbashir, S. M., Harborth, J., Lendeckel, W., Yalcin, A., Weber, K., & Tuschl, T. (2001). Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature, 411*(6836), 494-498. doi: 10.1038/35078107

Elkayam, E., Kuhn, C. D., Tocilj, A., Haase, A. D., Greene, E. M., Hannon, G. J., & Joshua-Tor, L. (2012). The structure of human argonaute-2 in complex with miR-20a. *Cell, 150*(1), 100-110. doi: 10.1016/j.cell.2012.05.017

Fellmann, C., Hoffmann, T., Sridhar, V., Hopfgartner, B., Muhar, M., Roth, M., . . . Zuber, J. (2013). An optimized microRNA backbone for effective single-copy RNAi. *Cell Rep, 5*(6), 1704-1713. doi: 10.1016/j.celrep.2013.11.020

Fellmann, C., Zuber, J., McJunkin, K., Chang, K., Malone, C. D., Dickins, R. A., . . . Lowe, S. W. (2011). Functional identification of optimized RNAi triggers using a massively parallel sensor assay. *Mol Cell, 41*(6), 733-746. doi: 10.1016/j.molcel.2011.02.008

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., & Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature, 391*(6669), 806-811. doi: 10.1038/35888

Frank, F., Sonenberg, N., & Nagar, B. (2010). Structural basis for 5'-nucleotide base-specific recognition of guide RNA by human AGO2. *Nature, 465*(7299), 818-822. doi: 10.1038/nature09039

Grishok, A., Pasquinelli, A. E., Conte, D., Li, N., Parrish, S., Ha, I., . . . Mello, C. C. (2001). Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control C. elegans developmental timing. *Cell, 106*(1), 23-34.

Gupta, S., Schoer, R. A., Egan, J. E., Hannon, G. J., & Mittal, V. (2004). Inducible, reversible, and stable RNA interference in mammalian cells. *Proc Natl Acad Sci U S A, 101*(7), 1927-1932. doi: 10.1073/pnas.0306111101

Hammond, S. M., Boettcher, S., Caudy, A. A., Kobayashi, R., & Hannon, G. J. (2001). Argonaute2, a link between genetic and biochemical analyses of RNAi. *Science, 293*(5532), 1146-1150. doi: 10.1126/science.1064023

Han, J., Lee, Y., Yeom, K. H., Nam, J. W., Heo, I., Rhee, J. K., . . . Kim, V. N. (2006). Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. *Cell, 125*(5), 887-901. doi: 10.1016/j.cell.2006.03.043

Hannon, G. J. (2002). RNA interference. *Nature, 418*(6894), 244-251. doi: 10.1038/418244a

Huesken, D., Lange, J., Mickanin, C., Weiler, J., Asselbergs, F., Warner, J., . . . Hall, J. (2005). Design of a genome-wide siRNA library using an artificial neural network. *Nat Biotechnol, 23*(8), 995-1001. doi: 10.1038/nbt1118

Hutvagner, G., McLachlan, J., Pasquinelli, A. E., Balint, E., Tuschl, T., & Zamore, P. D. (2001). A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science, 293*(5531), 834-838. doi: 10.1126/science.1062961

Hutvagner, G., & Zamore, P. D. (2002). A microRNA in a multiple-turnover RNAi enzyme complex. *Science, 297*(5589), 2056-2060. doi: 10.1126/science.1073827

Kamath, R. S., Fraser, A. G., Dong, Y., Poulin, G., Durbin, R., Gotta, M., . . . Ahringer, J. (2003). Systematic functional analysis of the Caenorhabditis elegans genome using RNAi. *Nature, 421*(6920), 231-237. doi: 10.1038/nature01278

Kambris, Z., Brun, S., Jang, I. H., Nam, H. J., Romeo, Y., Takahashi, K., . . . Lemaitre, B. (2006). Drosophila immunity: a large-scale in vivo RNAi screen identifies five serine proteases required for Toll activation. *Curr Biol, 16*(8), 808-813. doi: 10.1016/j.cub.2006.03.020

Ketting, R. F., Fischer, S. E., Bernstein, E., Sijen, T., Hannon, G. J., & Plasterk, R. H. (2001). Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C. elegans. *Genes Dev, 15*(20), 2654-2659. doi: 10.1101/gad.927801

Khvorova, A., Reynolds, A., & Jayasena, S. D. (2003). Functional siRNAs and miRNAs exhibit strand bias. *Cell, 115*(2), 209-216.

Lai, E. C. (2002). Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation. *Nat Genet, 30*(4), 363-364. doi: 10.1038/ng865

Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., . . . Kim, V. N. (2003). The nuclear RNase III Drosha initiates microRNA processing. *Nature, 425*(6956), 415-419. doi: 10.1038/nature01957

Lewis, B. P., Burge, C. B., & Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell, 120*(1), 15-20. doi: 10.1016/j.cell.2004.12.035

Lund, E., Guttinger, S., Calado, A., Dahlberg, J. E., & Kutay, U. (2004). Nuclear export of microRNA precursors. *Science, 303*(5654), 95-98. doi: 10.1126/science.1090599

Luo, J., Emanuele, M. J., Li, D., Creighton, C. J., Schlabach, M. R., Westbrook, T. F., . . . Elledge, S. J. (2009). A genome-wide RNAi screen identifies multiple synthetic lethal interactions with the Ras oncogene. *Cell, 137*(5), 835-848. doi: 10.1016/j.cell.2009.05.006

Malone, C., Brennecke, J., Czech, B., Aravin, A., & Hannon, G. J. (2012). Preparation of small RNA libraries for high-throughput sequencing. *Cold Spring Harb Protoc, 2012*(10), 1067-1077. doi: 10.1101/pdb.prot071431

Martinez, J., Patkaniowska, A., Urlaub, H., Luhrmann, R., & Tuschl, T. (2002). Single-stranded antisense siRNAs guide target RNA cleavage in RNAi. *Cell, 110*(5), 563-574.

Matveeva, O. V., Nazipova, N. N., Ogurtsov, A. Y., & Shabalina, S. A. (2012). Optimized models for design of efficient miR30-based shRNAs. *Front Genet, 3*, 163. doi: 10.3389/fgene.2012.00163

Meister, G. (2013). Argonaute proteins: functional insights and emerging roles. *Nat Rev Genet, 14*(7), 447-459. doi: 10.1038/nrg3462

Nakanishi, K., Weinberg, D. E., Bartel, D. P., & Patel, D. J. (2012). Structure of yeast Argonaute with guide RNA. *Nature, 486*(7403), 368-374. doi: 10.1038/nature11211

Okamura, K., Ishizuka, A., Siomi, H., & Siomi, M. C. (2004). Distinct roles for Argonaute proteins in small RNA-directed RNA cleavage pathways. *Genes Dev, 18*(14), 1655-1666. doi: 10.1101/gad.1210204

Paddison, P. J., Caudy, A. A., Bernstein, E., Hannon, G. J., & Conklin, D. S. (2002). Short hairpin RNAs (shRNAs) induce sequence-specific silencing in mammalian cells. *Genes Dev, 16*(8), 948-958. doi: 10.1101/gad.981002

Paddison, P. J., Silva, J. M., Conklin, D. S., Schlabach, M., Li, M., Aruleba, S., . . . Hannon, G. J. (2004). A resource for large-scale RNA-interference-based screens in mammals. *Nature, 428*(6981), 427-431. doi: 10.1038/nature02370

Reynolds, A., Leake, D., Boese, Q., Scaringe, S., Marshall, W. S., & Khvorova, A. (2004). Rational siRNA design for RNA interference. *Nat Biotechnol, 22*(3), 326-330. doi: 10.1038/nbt936

Sanchez Alvarado, A., & Newmark, P. A. (1999). Double-stranded RNA specifically disrupts gene expression during planarian regeneration. *Proc Natl Acad Sci U S A, 96*(9), 5049-5054.

Scholl, C., Frohling, S., Dunn, I. F., Schinzel, A. C., Barbie, D. A., Kim, S. Y., . . . Gilliland, D. G. (2009). Synthetic lethal interaction between oncogenic KRAS dependency and STK33 suppression in human cancer cells. *Cell, 137*(5), 821-834. doi: 10.1016/j.cell.2009.03.017

Schwarz, D. S., Hutvagner, G., Du, T., Xu, Z., Aronin, N., & Zamore, P. D. (2003). Asymmetry in the assembly of the RNAi enzyme complex. *Cell, 115*(2), 199-208.

Seitz, H., Ghildiyal, M., & Zamore, P. D. (2008). Argonaute loading improves the 5' precision of both MicroRNAs and their miRNA* strands in flies. *Curr Biol, 18*(2), 147-151. doi: 10.1016/j.cub.2007.12.049

Seitz, H., & Zamore, P. D. (2006). Rethinking the microprocessor. *Cell, 125*(5), 827-829. doi: 10.1016/j.cell.2006.05.018

Silva, J. M., Li, M. Z., Chang, K., Ge, W., Golding, M. C., Rickles, R. J., . . . Hannon, G. J. (2005). Second-generation shRNA libraries covering the mouse and human genomes. *Nat Genet, 37*(11), 1281-1288. doi: 10.1038/ng1650

Sims, D., Mendes-Pereira, A. M., Frankum, J., Burgess, D., Cerone, M. A., Lombardelli, C., . . . Lord, C. J. (2011). High-throughput RNA interference screening using pooled shRNA libraries and next generation sequencing. *Genome Biol, 12*(10), R104. doi: 10.1186/gb-2011-12-10-r104

Svoboda, P., Stein, P., Hayashi, H., & Schultz, R. M. (2000). Selective reduction of dormant maternal mRNAs in mouse oocytes by RNA interference. *Development, 127*(19), 4147-4156.

Tibshirani, R. (1995). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, 58*(1), 267-288.

Timmons, L., & Fire, A. (1998). Specific interference by ingested dsRNA. *Nature, 395*(6705), 854. doi: 10.1038/27579

Tuschl, T., Zamore, P. D., Lehmann, R., Bartel, D. P., & Sharp, P. A. (1999). Targeted mRNA degradation by double-stranded RNA in vitro. *Genes Dev, 13*(24), 3191-3197.

Ui-Tei, K., Naito, Y., Takahashi, F., Haraguchi, T., Ohki-Hamazaki, H., Juni, A., . . . Saigo, K. (2004). Guidelines for the selection of highly effective siRNA sequences for mammalian and chick RNA interference. *Nucleic Acids Res, 32*(3), 936-948. doi: 10.1093/nar/gkh247

Vert, J. P., Foveau, N., Lajaunie, C., & Vandenbrouck, Y. (2006). An accurate and interpretable model for siRNA efficacy prediction. *BMC Bioinformatics, 7*, 520. doi: 10.1186/1471-2105-7-520

Wang, Y., Sheng, G., Juranek, S., Tuschl, T., & Patel, D. J. (2008). Structure of the guide-strand-containing argonaute silencing complex. *Nature, 456*(7219), 209-213. doi: 10.1038/nature07315

Yi, R., Qin, Y., Macara, I. G., & Cullen, B. R. (2003). Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes Dev, 17*(24), 3011-3016. doi: 10.1101/gad.1158803

Yuan, Y. R., Pei, Y., Chen, H. Y., Tuschl, T., & Patel, D. J. (2006). A potential protein-RNA recognition event along the RISC-loading pathway from the structure of A. aeolicus Argonaute with externally bound siRNA. *Structure, 14*(10), 1557-1565. doi: 10.1016/j.str.2006.08.009

Zender, L., Xue, W., Zuber, J., Semighini, C. P., Krasnitz, A., Ma, B., . . . Lowe, S. W. (2008). An oncogenomics-based in vivo RNAi screen identifies tumor suppressors in liver cancer. *Cell, 135*(5), 852-864. doi: 10.1016/j.cell.2008.09.061

Zeng, Y., & Cullen, B. R. (2003). Sequence requirements for micro RNA processing and function in human cells. *RNA, 9*(1), 112-123.

Zhang, X., & Zeng, Y. (2010). The terminal loop region controls microRNA processing by Drosha and Dicer. *Nucleic Acids Res, 38*(21), 7689-7697. doi: 10.1093/nar/gkq645

Figure Legends

Figure 1. Identification of Sequence Characteristics Predictive of shRNA Efficacy A) shRNA score determination via sensor NGS data. On the left is a heatmap representation of normalized shRNA read counts for each on-dox sensor sort. The right panel represents shRNA potencies, calculated by extracting the first principal component of the left panel matrix. **B**) A nucleotide logo representing enriched (top) and depleted (bottom) nucleotides (p-value < 0.05) in potent shRNAs. **C**) A heatmap demonstrating the predictive capacity (with respect to shRNA potency) of each pair of positions within the target region. Heatmap cells are colored to represent the number of nucleotide combinations that were significantly predictive (p-value <0.05), at each position-pair. **D**) The predictive capacity of each triplet of positions within the target region. Data-point colors and sizes represent the number of nucleotide triplets that were significantly predictive (p-value <0.05) at each position-triplet.

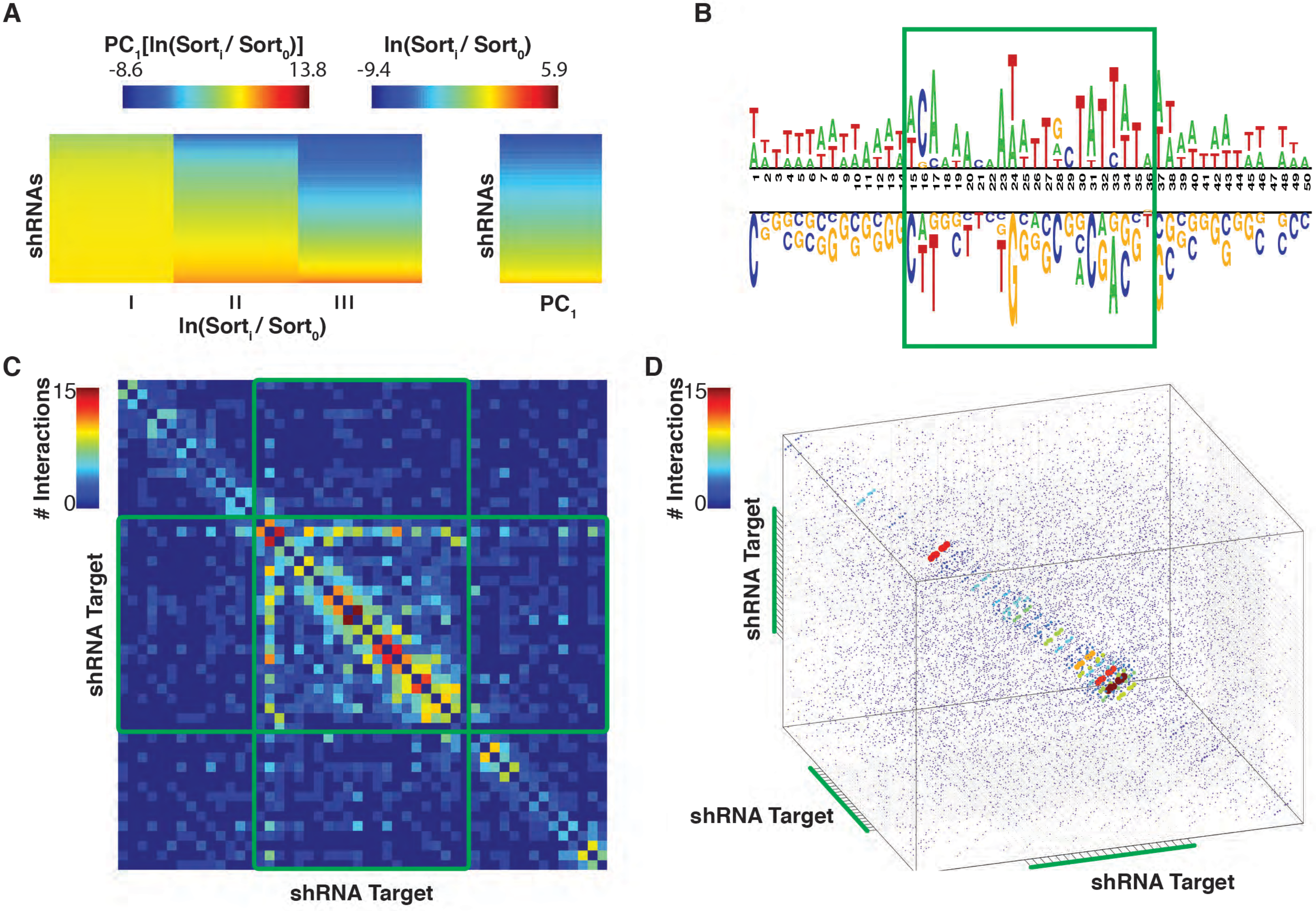
Figure 2. Construction and Validation of an shRNA-specific Predictive Algorithm A) Consolidated cross validation of predictions vs. sensor-scores for all shRNAs in the Fellmann et al. dataset (shRNAs are separated by the guide 5' nucleotide). **B**) GO-term instances associated with the targeted gene set selected for shRNA validation screens. **C**) GO-term instances associated with genes for which at least two hairpins significantly depleted in each of the TRC, Hannon-Elledge (HE) and shERWOOD (SW) validation screens **D**) The percentage of shRNAs targeting consensus essential genes that depleted in each of the TRC, HE and shERWOOD shRNA screens. **E**) Average log-fold change for shRNAs targeting consensus essential genes (per gene) for each of the TRC, EH and shERWOOD validation screens. **F**) The percentage of shRNAs corresponding to consensus essential genes that, for any given shERWOOD score, depleted in the shERWOOD validation screen. e.g., on average, ~34% of shERWOOD selected shRNAs with a score of 0.5 or greater depleted.

Figure 3. Structure-guided Maximization of shRNA-Prediction Space A) Histogram of sensor scores for the top fifteen shRNAs, as identified by the shERWOOD-1U strategy, targeting ~2000 "druggable" genes. Overlaid are the mean sensor scores for control shRNAs representing poor, medium, potent and very potent shRNAs (with mean knockdown efficiencies of 25%, 50%, 75% and >90%, respectively). **B**) The distribution of shERWOOD-1U prediction scores for shRNAs where endogenous 1U-shRNAs are separated from endogenous non-1U-shRNAs. Sensor scores for endogenous 1U- and non-1U-shRNAS are displayed on the left. **C**) Distribution of sensor scores for shERWOOD-1U-selected shRNAs, separated by endogenous guide 5' nucleotides. **D**) A nucleotide logo representing enriched (top) and depleted (bottom) nucleotides (p-value < 0.05) in potent shERWOOD-1U-selected shRNAs (separated by endogenous guide 5' nucleotides). E) The distribution of sensor scores for shRNAs classified as weak and potent by a random forest classifier trained on the shERWOO-1U sensor data. F) The distributions of the percentage of

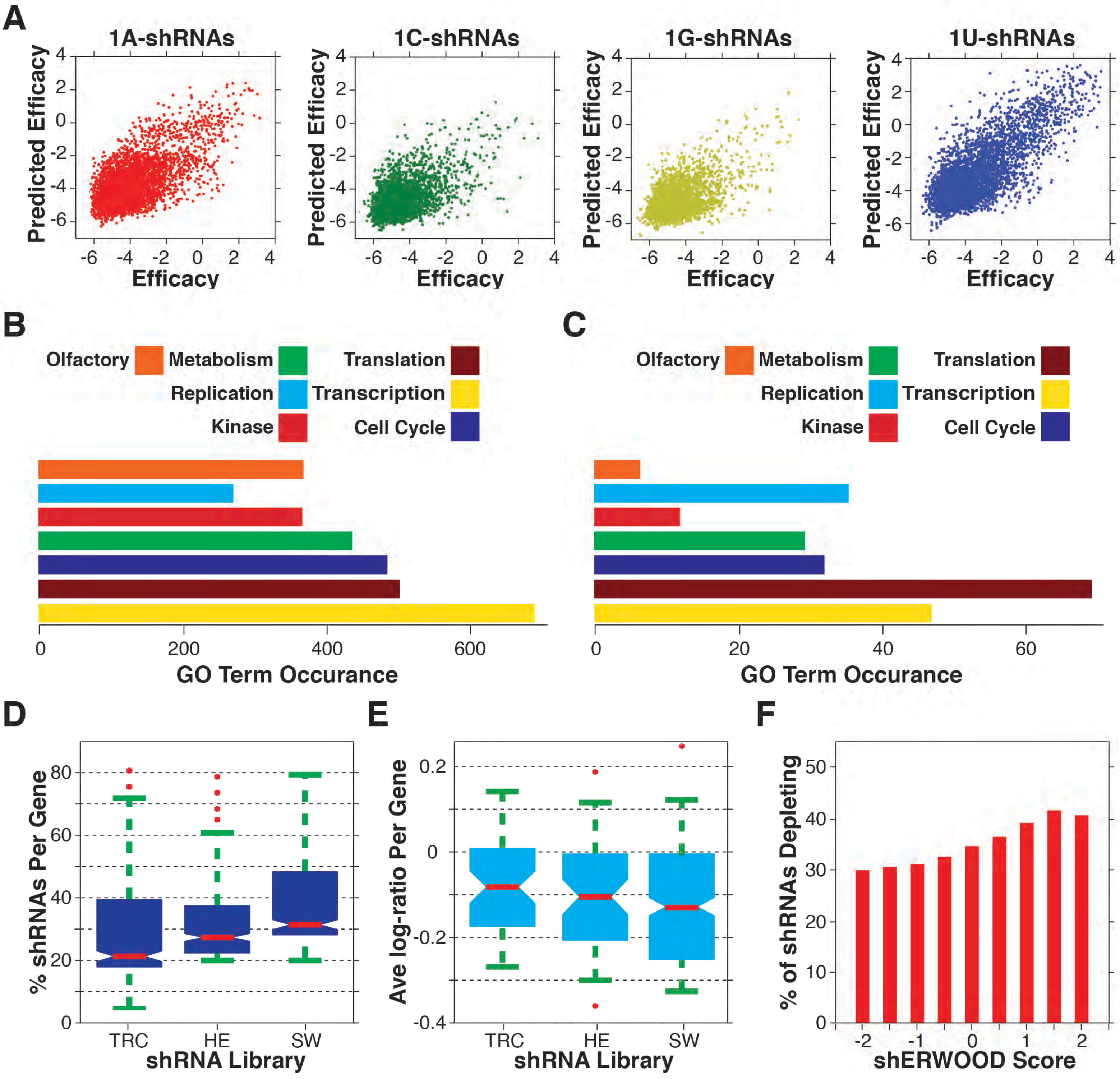
shERWOOD- and shERWOOD-1U-selected shRNAs targeting consensus essential genes that depleted in validation screens (left). In addition normalized log-fold changes of shRNAs, identified under each selection scheme, are displayed (right).

Figure 4. Validation of an Alternative Mir Scaffold A) Schematic representation of the cloning schemes for traditional miR30 and ultramiR shRNA scaffolds. **B**) Relative abundances of processed guide sequences for two shRNAs (as determined via small RNA cloning + NGS analysis) when cloned into traditional miR30 and ultramiR scaffolds. Values represent the log-fold enrichment of shRNA guides with respect to sequences corresponding to the ten most abundant microRNAs. **C**) Distributions of the percentage of shHERWOOD-1U-selected shRNAs targeting consensus essential genes that depleted in validation screens when shRNAs were placed into miR30 and ultramiR scaffolds. Log-fold changes for the same constructs are displayed on the left. D) Knockdown efficiencies for twelve shERWOOD-1U selected shRNAs cloned into the ultramiR scaffold. Four shRNAs were selected for three genes, as indicated.

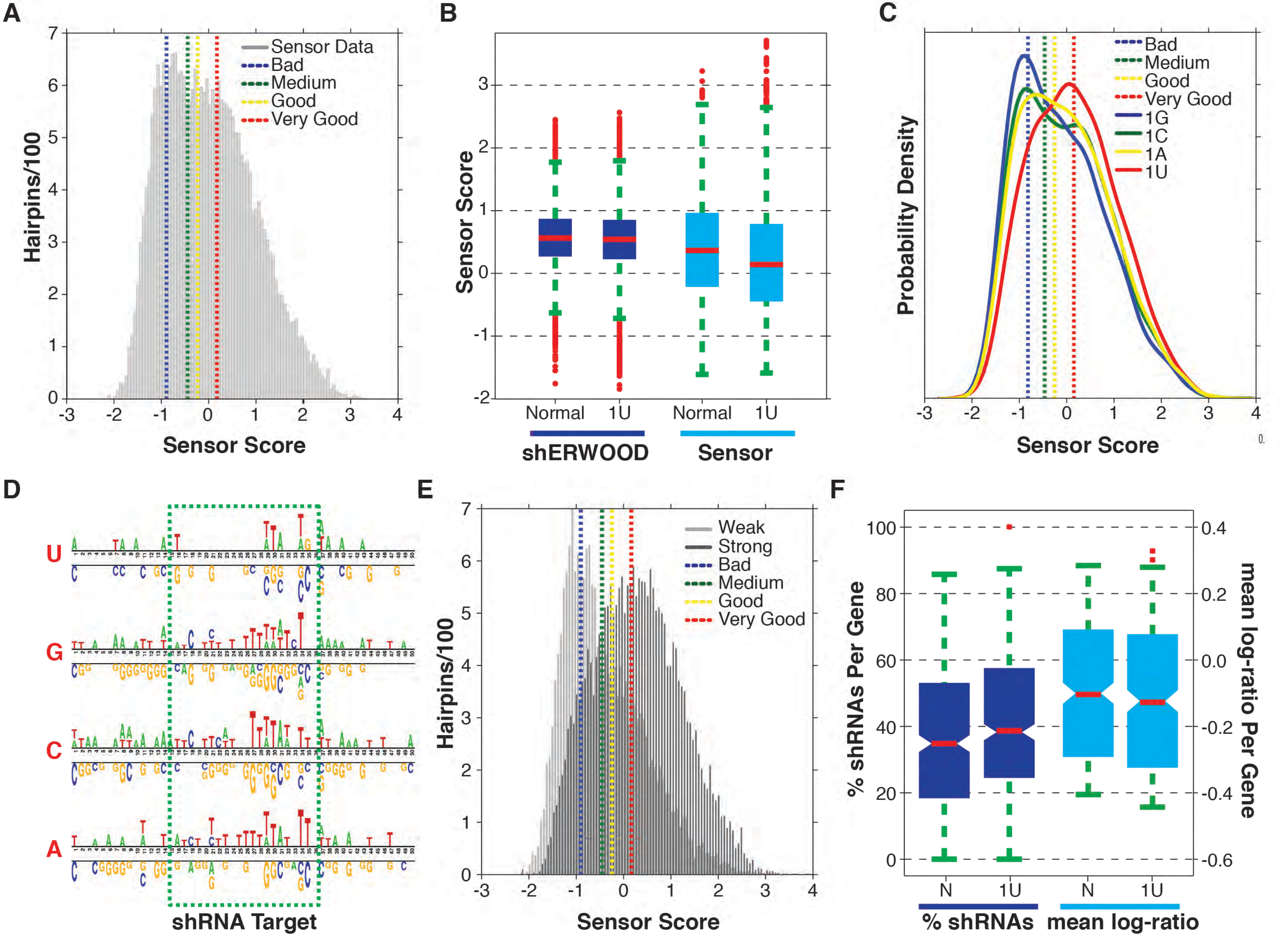
Knott et al. Figure 1



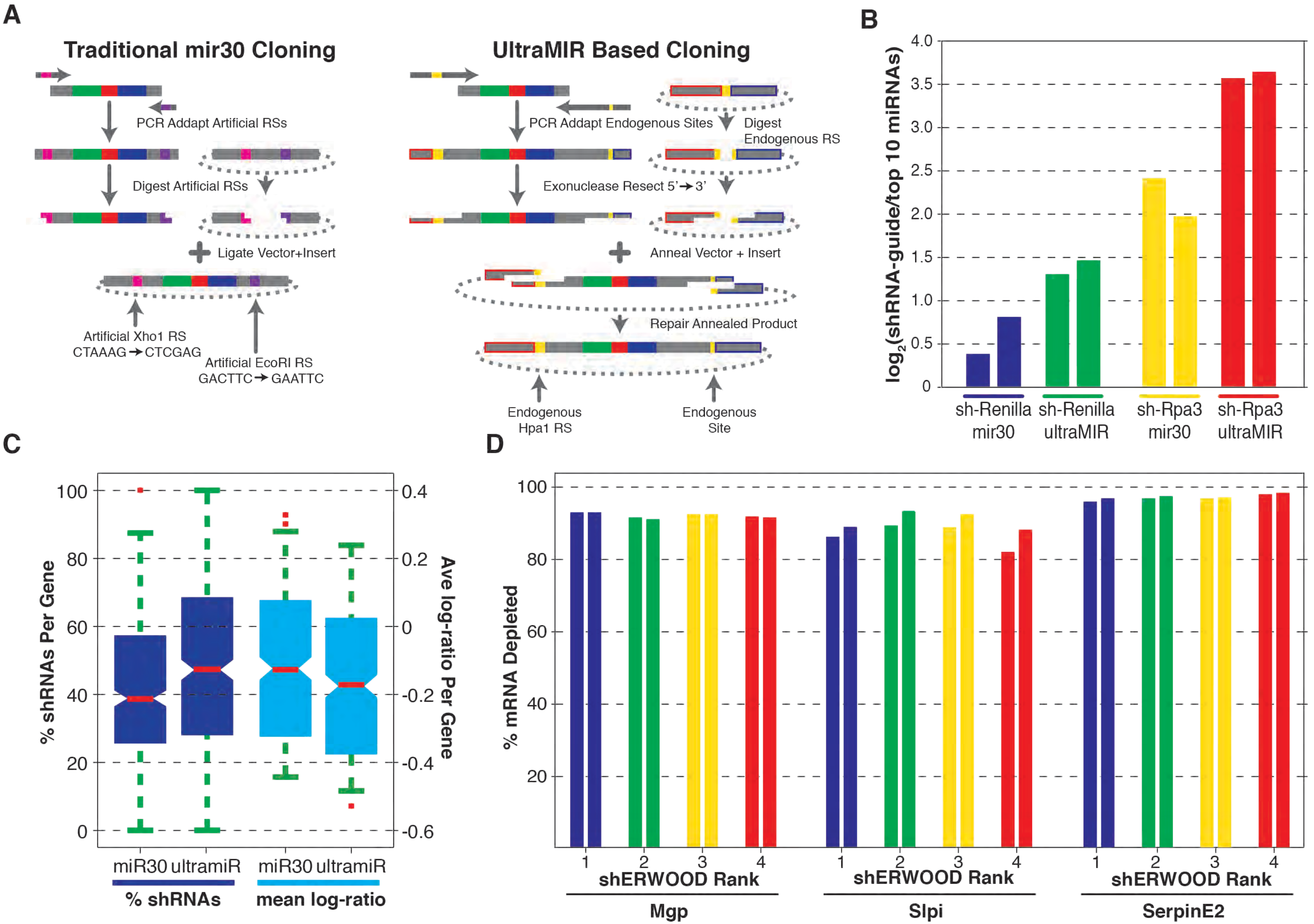
Knott et al. Figure 2



Knott et al. Figure 3



Knott et al. Figure 4


A
Traditional mir30 Cloning
PCR Addapt Artificial RSs
Digest Artificial RSs
Ligate Vector+Insert
Artificial Xho1 RS
CTAAAG → CTCGAG
Artificial EcoRI RS
GACTTC → GAATTC
UltraMIR Based Cloning
PCR Addapt Endogenous Sites
Digest
Endogenous RS
Exonuclease Resect 5' → 3'
Anneal Vector + Insert
Repair Annealed Product
Endogenous
Hpa1 RS
Endogenous
Site
B
log$_2$(shRNA-guide/top 10 miRNAs)
sh-Renilla
mir30
sh-Renilla
ultraMIR
sh-Rpa3
mir30
sh-Rpa3
ultraMIR
C
% shRNAs Per Gene
Ave log-ratio Per Gene
miR30 ultramiR
miR30 ultramiR
% shRNAs
mean log-ratio
D
% mRNA Depleted
shERWOOD Rank
Mgp
shERWOOD Rank
Slpi
shERWOOD Rank
SerpinE2