Molecular Profiling of Breast Cancer Cell Lines Defines Relevant Tumor Models and Provides a Resource for Cancer Gene Discovery

Jessica Kao1,9, Keyan Salari1,2,9, Melanie Bocanegra1, Yoon-La Choi1,3, Luc Girard4, Jeet Gandhi4, Kevin A. Kwei1, Tina Hernandez-Boussard2, Pei Wang5, Adi F. Gazdar4, John D. Minna4, Jonathan R. Pollack1*

1 Department of Pathology, Stanford University, Stanford, California, United States of America, 2 Department of Genetics, Stanford University, Stanford, California, United States of America, 3 Department of Pathology, Samsung Medical Center, Sungkyunkwan University School of Medicine, Seoul, South Korea, 4 Hamon Center for Therapeutic Oncology Research, University of Texas Southwestern Medical Center, Dallas, Texas, United States of America, 5 Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington, United States of America

Abstract

**Background:** Breast cancer cell lines have been used widely to investigate breast cancer pathobiology and new therapies. Breast cancer is a molecularly heterogeneous disease, and it is important to understand how well and which cell lines best model that diversity. In particular, microarray studies have identified molecular subtypes—luminal A, luminal B, ERBB2-associated, basal-like and normal-like—with characteristic gene-expression patterns and underlying DNA copy number alterations (CNAs). Here, we studied a collection of breast cancer cell lines to catalog molecular profiles and to assess their relation to breast cancer subtypes.

**Methods:** Whole-genome DNA microarrays were used to profile gene expression and CNAs in a collection of 52 widely-used breast cancer cell lines, and comparisons were made to existing profiles of primary breast tumors. Hierarchical clustering was used to identify gene-expression subtypes, and Gene Set Enrichment Analysis (GSEA) to discover biological features of those subtypes. Genomic and transcriptional profiles were integrated to discover within high-amplitude CNAs candidate cancer genes with coordinately altered gene copy number and expression.

**Findings:** Transcriptional profiling of breast cancer cell lines identified one luminal and two basal-like (A and B) subtypes. Luminal lines displayed an estrogen receptor (ER) signature and resembled luminal-A/B tumors, basal-A lines were associated with ETS-pathway and BRCA1 signatures and resembled basal-like tumors, and basal-B lines displayed mesenchymal and stem/progenitor-cell characteristics. Compared to tumors, cell lines exhibited similar patterns of CNA, but an overall higher complexity of CNA (genetically simple luminal-A tumors were not represented), and only partial conservation of subtype-specific CNAs. We identified 80 high-level DNA amplifications and 13 multiplicity deletions, and the resident genes with concomitantly altered gene-expression, highlighting known and novel candidate breast cancer genes.

**Conclusions:** Overall, breast cancer cell lines were genetically more complex than tumors, but retained expression patterns with relevance to the luminal-basal subtype distinction. The compendium of molecular profiles defines cell lines suitable for investigations of subtype-specific pathobiology, cancer stem cell biology, biomarkers and therapies, and provides a resource for discovery of new breast cancer genes.


Editor: Mikhail V. Blagosklonny, Roswell Park Cancer Institute, United States of America

Received March 19, 2009; Accepted June 2, 2009; Published July 3, 2009

Copyright: © 2009 Kao et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: This work was supported by grants from the NIH (CA97139; J.R.P.), the California Breast Cancer Research Program, (8KB-0135; J.R.P.), and the Longenbaugh Foundation (J.D.M.). K.S. is a Paul & Daisy Soros Fellow and fellow of the Medical Scientist Training Program. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: pollack1@stanford.edu

**These authors contributed equally to this work.

Introduction

Breast cancer, a leading cause of cancer death in women, is recognized to be a molecularly heterogeneous disease. Markers such as estrogen receptor (ER), progesterone receptor (PR) and ERBB2/HER2 are used for prognostication, and to stratify patients for appropriately targeted therapies [1].

More recently, DNA microarray studies have suggested a refined classification of breast cancer, distinguishing five major subtypes based on different patterns of gene expression, underlying DNA copy number alterations (CNAs), and associated clinical outcomes [2–5]. Luminal subtypes A and B are ER positive and share expression markers with the luminal epithelial layer of cells lining normal breast ducts. Luminal-A tumors are genetically...
simple (1q/16p gain) and are associated with favorable outcome, while luminal-B tumors exhibit high proliferation rates, frequent DNA amplification (e.g. 9q22/MYC), and less favorable prognosis. Basal-like tumors share expression markers with the underlying basal (myoepithelial) layer of normal breast ducts, are ER negative, exhibit frequent chromosome segmental gains/losses, and are associated with poor outcome in most studies. The ERBB2 subtype is associated with expression of genes co-amplified with ERBB2 (encoding HER2) on chromosome cytoband 17q12, and the normal-like subtype shares expression patterns with normal breast tissue.

Breast cancer cell lines have been used widely to investigate breast cancer pathobiology, and to screen and characterize new therapeutics [6,7]. Advantages of cell lines include the relative ease of pharmacologic and genetic manipulation, the variety of available functional assays, and, for some studies, the purity of the cancerous epithelial population (and absence of stromal cell contamination). However, while some investigators choose particular cell lines based on the known ER or HER2 status, many others rely on standard “workhorses” like MCF7 without regard to the particular tumor subtypes being modeled. The recent recognition of microarray molecular subtypes points to the need for additional consideration in cell line selection.

The goal of our study was to profile gene expression and CNAs genome-wide in a collection of 32 publicly-available and commonly-used breast cancer cell lines, in order to assess the relation of these cell lines to the recognized molecular subtypes of breast cancer, and to discover new candidate breast cancer genes and pathways.

Materials and Methods

Breast Cancer Cell Lines

184A1, BT20, BT474, BT549, Hs578T, hTERT-HME1, MCF7, MCF10A, MDA-MB134, MDA-MB157, MDA-MB231, MDA-MB361, MDA-MB436, MDA-MB453, MDA-MB468, SKBR3, T47D, UACC812, UACC933, ZR75-1 and ZR75-30 were obtained from ATCC (Manassas, VA, USA). EFM19 and EFM192A were obtained from DSMZ (Braunschweig, Germany), HCC33, HCC370, HCC292, HCC712, HCC1107, HCC1143, HCC1395, HCC1419, HCC1428, HCC1500, HCC1569, HCC1599, HCC1806, HCC1937, HCC1954, HCC2157, HCC2185, HCC2218, HCC2688 and HCC3153 were obtained from the cell repository of the Hamon Center for Therapeutic Oncology Research, UT Southwestern Medical Center (many are now available from ATTC). CAL51 was a kind gift from J. Giovannetti from the Centre Antoine-Lacassagne, Nice, France. SUM44PE, SUM52PE, SUM102PT, SUM149PT and SUM190PT were kind gifts from Dr. Stephen P. Ethier (now available from Asterand, Detroit, MI). MCF10A was grown in MEGM media (Cambrex, East Rutherford, NJ). SUM52PE and SUM149PT were grown in a Ham’s F12 media with 5% FBS, supplemented with 5 μg/ml insulin and 1 μg/ml hydrocortisone. SUM44PE, SUM102PT and SUM190PT were grown in Ham’s F12 media with 0.1% BSA, supplemented with 5 μg/ml insulin, 1 μg/ml of hydrocortisone, 5 mM ethanolamine, 10 mM HEPES, 5 μg/ml transferrin, 10 nM of Triiodo Thyronin (T3) and 50 nM sodium selenite (10 ng/ml EGF was also included for SUM102PT). All other cell lines were grown in RPMI-1640 with 10% FBS and 1% Pen/Strep. Clinicopathological characteristics of cell lines are summarized in Table 1. A subset of cell lines (focused on the HCC series) was subjected to a more detailed molecular pathological characterization of ESR1, PGR, ERBB2, EGFR and BRCA1, as summarized in Table 2.

RNA and DNA isolation

Cells were grown to 70-80% confluence, then harvested for total RNA and genomic DNA. For HCC lines, RNA was prepared using the Qiagen RNNeasy Midi Kit (Qiagen, Valencia, CA) and DNA by phenol/chloroform extraction. For all other lines, RNA was isolated using Trizol (Invitrogen, Carlsbad, CA) according to the manufacturer’s protocol, and DNA by the Blood Cell Maxi Kit (Qiagen).

ERBB2 copy number assessment by quantitative PCR

ERBB2 copy number was quantified by real-time quantitative PCR (Q-PCR), using the Chromo4 PCR System (Bio-Rad Laboratories, Hercules, CA). GAST, located at 17q21 (on the same chromosomal arm as ERBB2) was used as a reference control. PCR primer sequences for ERBB2 and GAST are as follows (forward and reverse, respectively): ERBB2, 5′-TTGGAGGAGCTGGCATTTTCT-3′ and 5′-AGGTGATGCGTGCCACTCTT-3′; GAST, 5′-GTAGGCACTCCTCCCC-CATT-3′ and 5′-AGCCATGTCGCCGTCTTTT-3′, with PCR product lengths of 59 and 70 base pairs, respectively. Primers were chosen by TaqMan Primer ExpressTM 1.5 (Applied Biosystem, Foster City, CA) and purchased from Invitrogen. PCR reactions were carried out in a final volume of 20 μl containing 20 ng genomic DNA, 300 nM each primer (for both ERBB2 and GAST, in independent reactions) and 1 x Power SYBR Green PCR Master Mix (Applied Biosystems, Foster City, CA). PCR conditions were as follows: one cycle at 95°C for 10 minutes, followed by 40 cycles each at 95°C for 15 seconds and 60°C for 1 minute. Samples were analyzed in triplicate. Each amplification reaction was checked for the absence of nonspecific PCR products by melting curve analysis. ERBB2 copy number calculation was carried out using the comparative Ct method [8] after validating that the efficiencies of PCR reactions of both ERBB2 and GAST were equal. Human Genomic DNA (DNA20) (EMD Biosciences, Darmstadt, Germany), a mixture of pooled human whole blood from 6-8 individual male and female donors, was run in every assay as a calibrator sample. ERBB2 gene copy number in normal human genomic DNA was set as 2 and copy number more than 4 in cell lines was considered to be increased.

mRNA levels of ESR1, PGR, ERBB2 and EGFR

Transcript levels of ESR1, PGR, ERBB2 and EGFR were analyzed as a part of RT2 Profiler Custom PCR Array (SuperArray Bioscience, Frederick, MD). After making cDNA from 1.0 μg total RNA using RT2 PCR Array First Strand Kit (SuperArray Bioscience), quantitative PCR was performed with the Chromo4 PCR System (Bio-Rad Laboratories) using RT2 Real-Time SYBR Green PCR Master Mix (SuperArray Bioscience) according to the manufacturer’s protocol. We chose two different housekeeping genes, β-actin (ACTB) and glyceraldehyde-3-phosphate dehydrogenase (GAPDH) as internal controls, using the average of their Ct values. Primers were chosen by Taqman Primer Express™ 1.5 and purchased from Invitrogen, as follows: (forward and reverse, respectively): ESR1, 5′-ATCTGCG-GTTCGCCGATGGAATCATGCC-3′ and 5′-TGCTGGAC-GAAATGTGTAACCCAGCA-3′; PGR, 5′-CGTTCTGG-GACGTTAAGGTCTT-3′ and 5′-GCAGTCTATTTTCTCCAGAGAA-3′; ERBB2, 5′-TGACCTGTCGGGAAAGG-GGAGCG-3′ and 5′-TGCCCTGCGCATGGGAGAATGATCA-G-3′; EGFR, 5′-ATAGTGCGCCAAGATTTCCGTGAGT-3′.
Table 1. Clinicopathological features of breast cancer cell lines.

<table>
<thead>
<tr>
<th>Cell line</th>
<th>Subtype#</th>
<th>ER*</th>
<th>PR*</th>
<th>ERBB2/HER2*</th>
<th>Source#</th>
<th>Tumor type#</th>
</tr>
</thead>
<tbody>
<tr>
<td>18A41</td>
<td>B</td>
<td>—</td>
<td>NA</td>
<td>—</td>
<td>RM</td>
<td>NA</td>
</tr>
<tr>
<td>BT20</td>
<td>A</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>PT</td>
<td>AC</td>
</tr>
<tr>
<td>BT474</td>
<td>L</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>PT</td>
<td>IDC</td>
</tr>
<tr>
<td>BT483</td>
<td>L</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>PT</td>
<td>IDC</td>
</tr>
<tr>
<td>BT549</td>
<td>B</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>PT</td>
<td>IDC</td>
</tr>
<tr>
<td>CAL51</td>
<td>B</td>
<td>—</td>
<td>NA</td>
<td>—</td>
<td>PE</td>
<td>AC</td>
</tr>
<tr>
<td>EFM19</td>
<td>L</td>
<td>+</td>
<td>+</td>
<td>—</td>
<td>PE</td>
<td>IDC</td>
</tr>
<tr>
<td>EFM192A</td>
<td>L</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>PE</td>
<td>AC</td>
</tr>
<tr>
<td>HCC38</td>
<td>B</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>PT</td>
<td>DC</td>
</tr>
<tr>
<td>HCC70</td>
<td>A</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>PT</td>
<td>DC</td>
</tr>
<tr>
<td>HCC202</td>
<td>L</td>
<td>—</td>
<td>+</td>
<td>—</td>
<td>PT</td>
<td>DC</td>
</tr>
<tr>
<td>HCC712</td>
<td>L</td>
<td>+</td>
<td>—</td>
<td>—</td>
<td>PT</td>
<td>DC</td>
</tr>
<tr>
<td>HCC1007</td>
<td>L</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>PT</td>
<td>DC</td>
</tr>
<tr>
<td>HCC1143</td>
<td>A</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>PT</td>
<td>DC</td>
</tr>
<tr>
<td>HCC1187</td>
<td>A</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>PT</td>
<td>DC</td>
</tr>
<tr>
<td>HCC1395</td>
<td>B</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>PT</td>
<td>DC</td>
</tr>
<tr>
<td>HCC1419</td>
<td>L</td>
<td>—</td>
<td>+</td>
<td>+</td>
<td>PT</td>
<td>DC</td>
</tr>
<tr>
<td>HCC1428</td>
<td>L</td>
<td>+</td>
<td>+</td>
<td>—</td>
<td>PE</td>
<td>Met AC</td>
</tr>
<tr>
<td>HCC1500</td>
<td>L</td>
<td>+</td>
<td>+</td>
<td>—</td>
<td>PT</td>
<td>DC</td>
</tr>
<tr>
<td>HCC1569</td>
<td>A</td>
<td>—</td>
<td>—</td>
<td>+</td>
<td>PT</td>
<td>Met C</td>
</tr>
<tr>
<td>HCC1599</td>
<td>A</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>PT</td>
<td>DC</td>
</tr>
<tr>
<td>HCC1806</td>
<td>NA</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>PT</td>
<td>Sq C</td>
</tr>
<tr>
<td>HCC1937</td>
<td>A</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>PT</td>
<td>DC</td>
</tr>
<tr>
<td>HCC1954</td>
<td>A</td>
<td>—</td>
<td>—</td>
<td>+</td>
<td>PT</td>
<td>DC</td>
</tr>
<tr>
<td>HCC2157</td>
<td>A</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>PT</td>
<td>NA</td>
</tr>
<tr>
<td>HCC2185</td>
<td>L</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>PE</td>
<td>Met LC</td>
</tr>
<tr>
<td>HCC2218</td>
<td>L</td>
<td>—</td>
<td>+</td>
<td>—</td>
<td>PT</td>
<td>DC</td>
</tr>
<tr>
<td>HCC2688</td>
<td>L</td>
<td>—</td>
<td>NA</td>
<td>—</td>
<td>PT</td>
<td>DC</td>
</tr>
<tr>
<td>HCC3153</td>
<td>A</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>PT</td>
<td>DC</td>
</tr>
<tr>
<td>HSS57T</td>
<td>B</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>PT</td>
<td>C Sar</td>
</tr>
<tr>
<td>hTERT-HME1 B</td>
<td>—</td>
<td>NA</td>
<td>—</td>
<td>RM</td>
<td>NA</td>
<td>N</td>
</tr>
<tr>
<td>MCF7</td>
<td>L</td>
<td>+</td>
<td>+</td>
<td>—</td>
<td>PE</td>
<td>Met AC</td>
</tr>
<tr>
<td>MCF10A</td>
<td>B</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>RM</td>
<td>F</td>
</tr>
<tr>
<td>MDA134</td>
<td>L</td>
<td>+</td>
<td>—</td>
<td>—</td>
<td>PE</td>
<td>IDC</td>
</tr>
<tr>
<td>MDA157</td>
<td>B</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>PE</td>
<td>Med C</td>
</tr>
<tr>
<td>MDA175</td>
<td>L</td>
<td>+</td>
<td>—</td>
<td>—</td>
<td>PE</td>
<td>IDC</td>
</tr>
<tr>
<td>MDA231</td>
<td>B</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>PE</td>
<td>Met AC</td>
</tr>
<tr>
<td>MDA361</td>
<td>L</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>BR</td>
<td>Met AC</td>
</tr>
<tr>
<td>MDA436</td>
<td>B</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>PE</td>
<td>AC</td>
</tr>
<tr>
<td>MDA453</td>
<td>L</td>
<td>—</td>
<td>+</td>
<td>+</td>
<td>PE</td>
<td>Met C</td>
</tr>
<tr>
<td>MDA468</td>
<td>A</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>PE</td>
<td>Met AC</td>
</tr>
<tr>
<td>SKBR3</td>
<td>L</td>
<td>—</td>
<td>—</td>
<td>+</td>
<td>PE</td>
<td>AC</td>
</tr>
<tr>
<td>SUM44</td>
<td>NA</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>PE</td>
<td>ILC</td>
</tr>
<tr>
<td>SUM52</td>
<td>L</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>PE</td>
<td>Met C</td>
</tr>
<tr>
<td>SUM102</td>
<td>B</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>PE</td>
<td>IDC, apocrine</td>
</tr>
<tr>
<td>SUM149</td>
<td>B</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>PE</td>
<td>Inf</td>
</tr>
<tr>
<td>SUM190</td>
<td>L</td>
<td>—</td>
<td>+</td>
<td>—</td>
<td>PT</td>
<td>Inf</td>
</tr>
</tbody>
</table>

Abbreviations: A = Basal A subtype; AC = adenocarcinoma; AF = ascites fluid; B = Basal B subtype; BR = brain; C Sar = carcinoma sarcoma; DC = ductal carcinoma; F = fibrocystic disease; IDC = invasive ductal carcinoma; Inf = inflammatory carcinoma; ILC = invasive lobular carcinoma; L = Luminal subtype; Med C = medullary carcinoma, Met AC = metastatic adenocarcinoma; Met C = metastatic carcinoma, Met LC = metastatic lobular carcinoma, NA = not available; PE = pleural effusion; PT = primary tumor; RM = reduction mammoplasty; Sq C = Squamous Carcinoma.

* Determined from this study.

** Determined from the ATCC (http://www.atcc.org) and DSMZ (http://www.dsmz.de) websites, and references therein, or from this study.

*ERBB2 amplified but not highly expressed.

doi:10.1371/journal.pone.0006146.t001

Table 1. Cont.

<table>
<thead>
<tr>
<th>Cell line</th>
<th>Subtype#</th>
<th>ER*</th>
<th>PR*</th>
<th>ERBB2/HER2*</th>
<th>Source#</th>
<th>Tumor type#</th>
</tr>
</thead>
<tbody>
<tr>
<td>T47D</td>
<td>L</td>
<td>+</td>
<td>+</td>
<td>—</td>
<td>PE</td>
<td>IDC</td>
</tr>
<tr>
<td>UACC812</td>
<td>L</td>
<td>+</td>
<td>—</td>
<td>+</td>
<td>PT</td>
<td>IDC</td>
</tr>
<tr>
<td>UACC893</td>
<td>L</td>
<td>—</td>
<td>—</td>
<td>+</td>
<td>PT</td>
<td>IDC</td>
</tr>
<tr>
<td>ZR75-1</td>
<td>L</td>
<td>+</td>
<td>—</td>
<td>—</td>
<td>AF</td>
<td>IDC</td>
</tr>
<tr>
<td>ZR75-30</td>
<td>L</td>
<td>+</td>
<td>—</td>
<td>+</td>
<td>AF</td>
<td>IDC</td>
</tr>
</tbody>
</table>

Abbreviations: A = Basal A subtype; AC = adenocarcinoma; AF = ascites fluid; B = Basal B subtype; BR = brain; C Sar = carcinoma sarcoma; DC = ductal carcinoma; F = fibrocystic disease; IDC = invasive ductal carcinoma; Inf = inflammatory carcinoma; ILC = invasive lobular carcinoma; L = Luminal subtype; Med C = medullary carcinoma, Met AC = metastatic adenocarcinoma; Met C = metastatic carcinoma, Met LC = metastatic lobular carcinoma, NA = not available; PE = pleural effusion; PT = primary tumor; RM = reduction mammoplasty; Sq C = Squamous Carcinoma.

Western blot analysis and immunohistochemistry (IHC)

Preparation of total cell lysates and Western blotting were done as described previously [10]. Primary antibodies used were mouse monoclonal anti-ER-α (Cell Signaling, Beverly, MA), mouse monoclonal PR (6A1) (Cell Signaling), mouse monoclonal anti-HER2 (Cell Signaling), rabbit monoclonal anti-EGFR (Cell Signaling) and mouse monoclonal anti-actin (Sigma-Aldrich). Actin levels were used as a control for protein loading. Peroxidase-labeled anti-mouse or anti-rabbit antibodies (Amer- sham Pharmacia, Piscataway, NJ) were used as secondary antibody. IHC on breast cancer cell lines was described previously [11].

BRCA1 mutation analysis

DNA sequence analysis was performed on the entire BRCA1 gene in available lymphocyte DNA matched to breast cancer cell lines. In the lymphocyte DNA matching HCC3153, a heterozygous duplication of 10 base pairs was detected at position 943 in exon 11 of BRCA1 (943ins10). The region of BRCA1 exon 11 containing the 943ins10 mutation was amplified from genomic DNA in the tumor cell line (HCC3153) using standard PCR conditions. Sequence analysis revealed only the mutant sequence. Absence of the normal allele was also confirmed by single strand conformation analysis as well as gel electrophoresis of the amplified fragment on 5% acrylamide denaturing gels.

Gene expression profiling

Gene expression profiling was performed on Human Exonic Evidence Based oligonucleotide (HEEBO) arrays obtained from the Stanford Functional Genomics Facility and containing 36,192

and 5′-ACCACGTGTCATGCTTCTTCTCA-3′; ACTB (5′ GGCTTGTGCTGGAAGCTAAG-3′) and 5′-ATGATG-GAGTTGAAAGTGATTTGTGAT-3′) [9]. We also analyzed the values of NC11 (normal lymphocyte) cell line for ESR1, PGR, ERBB2 and EGFR mRNA expression, and the tumor cell values were reported relative to NC11. For data analysis, the comparative Ct method [8] was used.
Table 2. Molecular pathological analysis of breast cancer cell line subset.

<table>
<thead>
<tr>
<th>Cell line</th>
<th>Phenotype</th>
<th>BRCA1</th>
<th>Q-PCR&lt;sup&gt;a&lt;/sup&gt; ESR1</th>
<th>Q-PCR&lt;sup&gt;a&lt;/sup&gt; PGR</th>
<th>Q-PCR&lt;sup&gt;a&lt;/sup&gt; ERBB2</th>
<th>Q-PCR&lt;sup&gt;a&lt;/sup&gt; EGFR</th>
<th>IHC&lt;sup&gt;b&lt;/sup&gt; ESR1</th>
<th>IHC&lt;sup&gt;b&lt;/sup&gt; PGR</th>
<th>IHC&lt;sup&gt;b&lt;/sup&gt; ERBB2</th>
<th>IHC&lt;sup&gt;b&lt;/sup&gt; EGFR</th>
<th>Western ESR1</th>
<th>Western PGR</th>
<th>Western ERBB2</th>
<th>Western EGFR</th>
</tr>
</thead>
<tbody>
<tr>
<td>HCC38</td>
<td>Triple neg</td>
<td></td>
<td>1.18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HCC70</td>
<td>Triple neg</td>
<td></td>
<td>0.37</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HCC202</td>
<td>ERBB2 amp</td>
<td></td>
<td>28.88</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HCC712</td>
<td>Hormone+</td>
<td></td>
<td>0.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HCC1143</td>
<td>Triple neg</td>
<td></td>
<td>1.08</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HCC1187</td>
<td>Triple neg</td>
<td></td>
<td>0.42</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HCC1395</td>
<td>Triple neg</td>
<td></td>
<td>0.36</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HCC1419</td>
<td>ERBB2 amp</td>
<td></td>
<td>8.39</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HCC1428</td>
<td>Hormone+</td>
<td></td>
<td>0.20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HCC1500</td>
<td>Hormone+</td>
<td></td>
<td>0.38</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HCC1569</td>
<td>ERBB2 amp</td>
<td></td>
<td>33.75</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HCC1806</td>
<td>Triple neg</td>
<td></td>
<td>0.08</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HCC1937</td>
<td>Triple neg</td>
<td>INS C 5382</td>
<td>0.33</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HCC1954</td>
<td>ERBB2 amp</td>
<td></td>
<td>45.01</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HCC2185</td>
<td>Triple neg</td>
<td></td>
<td>0.63</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HCC3153</td>
<td>Triple neg</td>
<td>943 ins 10</td>
<td>0.64</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCF7</td>
<td>Hormone+</td>
<td></td>
<td>0.56</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BT483</td>
<td>Hormone+</td>
<td></td>
<td>0.19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BT549</td>
<td>Triple neg</td>
<td></td>
<td>0.63</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MDA157</td>
<td>Triple neg</td>
<td></td>
<td>0.76</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MDA231</td>
<td>Triple neg</td>
<td></td>
<td>0.90</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MDA453</td>
<td>Triple neg</td>
<td></td>
<td>3.88</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MDA134</td>
<td>Hormone+</td>
<td></td>
<td>0.76</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MDA175</td>
<td>Triple neg</td>
<td></td>
<td>0.57</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HMEC1585</td>
<td>Control</td>
<td></td>
<td>0.54</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CALU3</td>
<td>Control</td>
<td></td>
<td>12.59</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NC11</td>
<td>Control</td>
<td></td>
<td>1.75</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNA20</td>
<td>Control</td>
<td></td>
<td>2.00</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<sup>a</sup>Gene copy number determined using DNA20 (from normal lymphocytes) as a diploid control; bold values indicate amplification.

<sup>b</sup>mRNA expression quantified in comparison to the immortalized breast line HMEC1585; Calu3 was used a positive control for ERBB2, and MCF7 for ESR1.

doi:10.1371/journal.pone.0006146.t002

Gene expression profiling analysis

Background-subtracted fluorescence log<sub>2</sub> ratios were globally normalized for each array, and then mean-centered for each gene (i.e. reporting relative to the average log ratio across all samples). Unless otherwise specified, we included for subsequent analysis only well-measured genes defined as those with fluorescence intensities in the Cy5 or Cy3 channel at least 1.5-fold above background in at least 60% of samples. For unsupervised hierarchical clustering, we included only the 8,750 well-measured genes whose expression varied at least 3-fold from the mean in at least 5 samples (Table S1). Hierarchical clustering was performed and displayed using Cluster and TreeView software (http://rana.lbl.gov/EisenSoftware.htm). Enrichment for functionally related genes was tested across a collection of 1,687 curated gene sets (C2) using Gene Set Enrichment analysis (GSEA; Release 2.0) [14]. Cell lines were classified according to breast tumor subtype (luminal-A, luminal-B, ERBB2, basal-like and normal-like) using the nearest centroid method applied to the set of “intrinsic genes” (i.e. genes with small within-specimen compared to between-specimen expression variance), as done previously [15], here using oligonucleotides representing 18,141 mapped human genes. 40 μg of sample RNA and 40 μg of “universal” reference RNA (derived from 11 different established human cell lines) were differentially labeled with Cy5 and Cy3, respectively, using an amino-allyl coupling protocol, then cohybridized onto the microarray in a high volume mixing hybridization at 65°C for 40 hrs. Details of the array processing and sample labeling/hybridization methods have been described [12]. Following hybridization, arrays were washed and scanned using a GenePix 4000B Axon scanner (Axon Instruments, Union City, CA). Fluorescence ratios were extracted using Spot Reader software (Niles Scientific, Portola Valley, CA) and uploaded to the Stanford Microarray Database [13] for storage, retrieval, and analysis. For two lines, HCC1806 and SUM44PE, expression profiling array hybridizations did not meet quality-control inspection and were excluded from analysis. The complete microarray expression data are available at the Stanford Microarray Database (SMD) (http://smd.stanford.edu) and at the Gene Expression Omnibus (GEO) (accession GSE13376); all microarray data reported in the manuscript are described in accordance with MIAME guidelines.
Euclidean distance. To classify breast tumors from the Sorlie *et al.* dataset [3] according to cell line subtype (luminal, basal A, or basal B), we first built a classifier by combining the top 100 genes positively and negatively correlating with each of the three “one vs. others” cell line subtype distinctions, using Significance Analysis of Microarrays (SAM) [16]. The cell line subtype classifier, comprising 484 genes, was then applied to classify primary tumors using the nearest centroid method (with Euclidean distance). We also classified each cell line as being associated with a good or bad prognosis signature (70-gene prognostic signature [17]), the presence or absence of a wound healing signature (512-gene wound signature [18]), and the presence or absence of an hypoxia signature (123-gene hypoxia signature [19]). For each signature, we calculated the gene expression centroid of the two groups of breast tumors (as determined in the original publications), and then correlated each centroid with cell line expression of the respective signature genes. Membership was assigned to the group with the highest correlation (Pearson correlation).

### Array-based comparative genomic hybridization (aCGH)

Arrays for CGH were obtained from the Stanford Functional Genomics Facility. aCGH was performed using cDNA arrays containing 39,632 cDNAs, representing 22,279 mapped human genes (18,049 UniGene clusters [20], together with 4,230 additional mapped ESTs not assigned to UniGene IDs), according to previously published protocols [21,22]. Briefly, 4 μg of genomic DNA from cell lines was random-primer labeled with Cy3 and co-hybridized onto a microarray along with 4 μg of Cy3 labeled normal leukocyte female reference DNA. Following overnight hybridization, the arrays were washed and scanned as above. The complete aCGH data are available at SMD and at GEO (accession GSE15376).

### aCGH analysis

Background-subtracted log2 fluorescence ratios were normalized for each array by mean centering. Well-measured genes used for subsequent analysis were those with fluorescence intensities in the Cy3 reference channel at least 1.4 fold above background. Map positions for arrayed cDNA clones were assigned using the NCBI genome assembly, accessed through the UCSC genome browser database (NCBI Build 36.1). For genes represented by multiple arrayed cDNAs, the average log ratio was used. The complete processed aCGH dataset is available as Table S2. DNA gains and losses were identified using the cghFLasso (R package for Fused Lasso) method [23], which controls the false discovery rate (FDR) by using normal-normal hybridization arrays to approximate the null distribution of the test statistics (see [23] for more details). A FDR<1% was used to call gains and losses. The fraction of the genome altered was determined by calculating the fraction of genes with fluorescence ratios ≥3 (for amplifications) or with significant non-zero fused lasso calls (for gains and losses). Some analyses (where indicated) were carried out on cytobands (boundaries defined by NCBI Build 36.1) rather than individual genes. For each cell line, cytobands exhibiting CNA were defined as those with at least two genes called by cghFLasso, and the magnitude of the CNA defined as the average log2 ratio of genes within the cytoband. We defined high-level DNA amplifications and multi-copy deletions as continuous regions identified by cghFLasso with at least 50% of genes having fluorescence ratios ≥3 or ≤0.25 respectively. These sites were also checked against known copy number variants (CNVs) reported in the Database of Genomic Variants (http://projects.tcag.ca/variation). Significant associations between cytobands and gene-expression subtypes were identified using SAM with a FDR<5%.

### Integrating genomic and transcriptional profiles

To integrate DNA copy number data (generated using cDNA microarrays) and gene-expression data (HEEBO oligonucleotide arrays), each gene expression measurement was first assigned a DNA copy number from either a probe interrogating the same named gene, or the average copy number of the nearest 5’ and 3’ probes (NCBI Build 36.1). Identification of genes with correlated copy number and expression was carried out using the DR-Correlate application of DR-Integrator (K. Salari, manuscript in preparation). Briefly, for each gene a modified Student’s t-test was performed comparing gene expression levels in cell lines from the lowest and the highest deciles of all cell lines’ copy number for the same gene; random permutations of sample labels were used to estimate a FDR.

### Results

#### Transcriptional profiling identifies three breast cancer cell line subtypes

To catalog molecular variation in a collection of 52 widely-used breast cancer cell lines, we first profiled gene expression using whole genome oligonucleotide microarrays. Unsupervised hierarchical clustering of the 8,750 most variably expressed genes stratified cell lines into two main groups (see dendrogram, Fig. 1B). One group, designated “luminal” (blue dendrogram branches), contained all the ER-positive cell lines (Fig. 2A), and was characterized by the expression of ERα-regulated genes (e.g. *MIB1, RET, EGR3, TFF1*; Fig. 1H, and not shown) [24–27], as well as genes associated with luminal epithelial differentiation (e.g. *GATA3* and *FOXA1*, Fig. 1I) [28]. The other group, designated “basal”, contained only ER-negative cell lines (Fig. 2A) and was characterized by the expression of basal epithelial gene markers including MSN, ETS1, CAV1 and EGFR (Fig. 1E, and not shown) [29–32]. Basal cell lines were further stratified into two subgroups, designated A and B (in line with Neve *et al.* [33], discussed further below). The basal-A subtype (red dendrogram branches) contained many of the “HCC” lines established at UT Southwestern, including two known *BRCA1* mutant lines (HCC1937, HCC3153) ([34], and this study). Basal-A lines were characterized by expression of PROM1 (aka CD133), a marker of various cancer stem cells [35], as well as other genes like *GABRP* and *VTCN1* (Fig. 1F and 2C). Some of the basal-A lines also shared expression of luminal epithelial markers like *KRT8* and *KRT18* (Fig. 1G).

The basal-B subtype (orange dendrogram branches) included non-tumorigenic lines (MCF10A, hTERT-HME1, 184A1) as well as several highly invasive lines exhibiting features of epithelial-mesenchymal transition (EMT) (MDA-MB231, MDA-MB436, MDA-MB157, Hs578t) [36]. Basal-B lines were characterized by markers associated with aggressive tumor features, including *PLAT* (plasminogen activator) [37] and *TGFB1* [38] (Fig. 1G), as well as marker phenotypes associated with normal breast and breast cancer progenitor/stem cells (*MUC7/CALLA*; *CD44*/*CD24−/+/−*; and *ITGB3/CD61*) ([39–41]. In contrast to other basal lines, the subset of mesenchymal-like basal-B lines lacked expression of basal cytokeratin markers KRT5 and KRT17 (Fig. 1D, and not shown).

Subtype-specific differences in gene expression could also be identified by pathway analysis, using Gene Set Enrichment Analysis (GSEA) [14]. Included among the top signature associations (Table 3), the luminal cell line subtype was
Figure 1. Clustering of expression profiles defines breast cancer cell line subtypes. (A) Thumbnail “heatmap” of two-way hierarchical clustering of 50 breast cancer cell lines (columns) and 8,750 variably expressed genes (rows) (data available as Table S1). Gene expression ratios are depicted by log2 pseudocolor scale shown; gray represents poorly measured data. (B) Enlarged view of the sample dendrogram. Clustering stratifies cell lines into two main groups, luminal (blue dendrogram branches) and basal, the latter further subdivided into two subgroups, basal A (red) and basal B (orange). (C–I) Selected gene expression patterns extracted from the cluster; corresponding locations in the thumbnail are indicated by the vertical colored bars. (C) Basal-B; (D) Basal cytokeratins; (E) Basal; (F) Basal-A; (G) Luminal cytokeratins; (H) ER-associated; (I) Luminal differentiation. doi:10.1371/journal.pone.0006146.g001
characterized by enriched expression of ER and good prognosis signatures, basal-A by ETS pathway and BRCA1 signatures, and basal-B by EMT and epidermal growth factor (EGF) signatures.

In regard to molecular markers and gene mutations (Fig. 2A), the luminal subtype included all the ER-positive cancer lines \((P = 0.001, \text{ 2-tailed Fisher’s exact test})\), and all but two of the ERBB2-positive lines \((P = 0.002)\), half of which were also ER-positive. \(P\)-value inactivating mutations and \(PIK3CA\) activating mutations, functioning on the same pathway, were mutually exclusive in all but one sample. Interestingly, \(P\)-value mutations were more common in the combined basal-like cell lines \((P = 0.020)\), while \(PIK3CA\) mutations were more frequent in luminal lines \((P = 0.022)\). \(TP53\) mutations occurred more often in basal-like lines \((P = 0.038)\).

Relationship of breast cancer cell line and tumor subtypes

To determine the relation between breast cancer cell line subtypes (luminal, basal-A, basal-B) and breast tumor subtypes (luminal-A, luminal-B, ERBB2, basal-like, and normal-like), we first classified cell lines according to tumor subtype using a nearest centroid approach applied to the set of “intrinsic genes” used originally to define the tumor subtypes \([2,3]\) (see Methods) (Fig. 2B). By expression patterns, most of the luminal lines most
closely resembled either luminal-A or luminal-B tumors. Most basal-A lines resembled basal-like tumors, and most basal-B lines resembled either basal-like or ERBB2 tumors (despite that none were ERBB2-positive).

We also carried out the reverse analysis, building a cell line subtype classifier to classify 96 breast tumors (from the original Stanford/Norway study defining the five tumor subtypes [3]) according to cell line subtype (see Methods) (Fig. 2D). Notably, all basal-like tumors most resembled basal-A cell lines. Luminal-A and -B tumors most resembled luminal cell lines, while ERBB2 subgroup tumors most resembled either luminal or basal-A cell lines. A similar analysis of breast tumors arising in carriers of BRCA1 mutation, analyzed from a different dataset (The Netherlands Cancer Institute) [17], revealed highest resemblance in 17 of 18 cases to basal-A lines (not shown), while two BRCA2 mutation associated cases most resembled luminal cell lines.

In addition to the above cluster-derived luminal/basal tumor subtypes, alternative breast tumor subtype classifiers have been proposed, including a 70-gene prognostic signature supervised on the metastatic/non-metastatic distinction [17], a “wound” signature trained on the serum response of cultured fibroblasts [18], and a hypoxia signature derived from the hypoxic response of cultured mammary and renal tubular epithelial cells [19]. Each of the three signatures predicts unfavorable clinical outcome. Interestingly, the basal-like lines (considered together) were those predominantly expressing the 70-gene (P = 0.001, Fisher’s exact test) wound (P = 0.004), and hypoxia (P < 0.001) signatures (Fig. 2B).

### Genomic profiles of breast cancer cell lines

To survey DNA copy number alterations in the panel of 52 breast cancer cell lines, we carried out CGH on cDNA microarrays with validated performance characteristics [21] and covering 22,000 genes with an average mapping resolution (inter-probe distance) of <70 Kb. Across the sample set, the most frequent CNAs (called by cghFLasso—see Methods) were gains on 1q, 3q, 5p, 7q, 8q, 11q, 17q, and 20q, and losses on 1p, 9p, 11q, 13q, and 18p. Of these, gains on 1q, 8q, and 20q, and losses on 1p, 8p, and 13q (asterisked in Fig. 3B) represented among the luminal cell lines. Basal-A and basal-B cell lines also exhibited characteristic gains/losses (Fig. 2B), but none also selectively characteristic of basal-like tumors.

Luminal cell lines displayed overall higher frequencies of high-level DNA amplification (i.e. fluorescence ratios ≥3, corresponding to at least 5-fold amplification [21]) (Fig. 4A), a characteristic shared with basal-B tumors [4]. Luminal and basal-A lines both exhibited overall higher frequencies of gain/loss (a characteristic feature of basal-like tumors [4]), compared to basal-B lines (Fig. 4B).

### Integrated analysis for cancer gene discovery

The molecular profiles generated provide opportunities to identify breast cancer cell lines with an altered copy number and expression of known cancer genes, useful to model pathogenesis and therapy, and to discovery new breast cancer genes. For the latter, high-amplitude CNAs, i.e. high-level DNA amplifications and homozygous deletions, are particularly informative in pinpointing new cancer genes. Within the aCGH dataset we identified 80 loci of high-amplitude CNAs, i.e. high-level DNA amplifications and losses (9p, 4q, 10p, 11q, 13q, 17q12, 20p) with luminal-B tumors [4]. Luminal and basal-A lines both exhibited overall higher frequencies of gain/loss (a characteristic feature of basal-like tumors [4]), compared to basal-B lines (Fig. 4B).

### Table 3. GSEA of breast cancer cell line subtypes.

<table>
<thead>
<tr>
<th>Subtype</th>
<th>Gene Set</th>
<th>Description</th>
<th>Source</th>
<th>FDR*</th>
</tr>
</thead>
<tbody>
<tr>
<td>Luminal</td>
<td>BRCA_ER_POS</td>
<td>Correlated with ER+ in breast cancer</td>
<td>[17]</td>
<td>0.017</td>
</tr>
<tr>
<td>BRCA_PROGNOSIS_POS</td>
<td>Correlated with good prognosis in breast cancer</td>
<td></td>
<td>0.094</td>
<td></td>
</tr>
<tr>
<td>Basal-A</td>
<td>ETSPATHWAY</td>
<td>ETS transcription factor pathway</td>
<td>BioCarta</td>
<td>0.063</td>
</tr>
<tr>
<td>BRCA_BRC1A1_POS</td>
<td>Correlated with BRCA1 (germline) in breast cancer</td>
<td>[17]</td>
<td>0.063</td>
<td></td>
</tr>
<tr>
<td>IFN_ALL_UP</td>
<td>Upregulated with interferon-α, β, γ treatment</td>
<td>[96]</td>
<td>0.071</td>
<td></td>
</tr>
<tr>
<td>IFNALLPHA_HCC_UP</td>
<td>Upregulated with interferon-α treatment</td>
<td>[97]</td>
<td>0.076</td>
<td></td>
</tr>
<tr>
<td>GLYCOGEN</td>
<td>Glycogen processing</td>
<td>Broad Institute</td>
<td>0.078</td>
<td></td>
</tr>
<tr>
<td>Basal-B</td>
<td>JECCHLINGER_EMT_UP</td>
<td>Upregulated in EMT</td>
<td>[98]</td>
<td>0.040</td>
</tr>
<tr>
<td>EGF_HDMEC_UP</td>
<td>Upregulated with EGF treatment</td>
<td>[99]</td>
<td>0.042</td>
<td></td>
</tr>
<tr>
<td>DOIRSEY_DOXYCYCLINE_UP</td>
<td>Upregulated with GAB2 expression</td>
<td>[100]</td>
<td>0.047</td>
<td></td>
</tr>
<tr>
<td>HTERT_DN</td>
<td>Downregulated with hTERT-immortalization</td>
<td>[101]</td>
<td>0.048</td>
<td></td>
</tr>
<tr>
<td>HINATA_NFKB_UP</td>
<td>Upregulated by NFκB</td>
<td>[102]</td>
<td>0.049</td>
<td></td>
</tr>
</tbody>
</table>

*Only top five significant gene sets shown.

doi:10.1371/journal.pone.0006146.t003
Figure 3. Genomic profiles define spectra of CNAs in cell line subtypes. (A) Spectra of gains (red) and losses (green) across the genome, plotted as average log2 ratio, for 89 breast tumors [4], above, compared to the set of 50 cell lines (profiled for both expression and CNAs), below. (B) Spectra of gains and losses for the cell line subtypes: luminal (above), basal A (middle) and basal B (below). Statistically significant subtype-specific CNAs, called by SAM (FDR<5%), are marked by a black bar. The subset of those loci that also characterize the corresponding primary breast tumor subtype is marked by an asterisk.

doi:10.1371/journal.pone.0006146.g003
expression correlated significantly (FDR<0.05) with altered gene copy number (Table S3). Of these, 487 resided within loci of high-amplitude CNA (Table 4). This subset included known breast cancer genes, like EGFR (7p11), FGFR1 (8p12), ERBB2 (17q12), PPMID (17q23) and ZNF217 (20q13). This subset is likely also enriched for novel breast cancer genes, and as such represents a rich source for cancer gene discovery. Notably, among the larger group of amplified/overexpressed genes are several with known functions relevant to oncogenesis, like cell proliferation (e.g. EIF3H, HES1, MELK, GAP2, CDC6, GRB2 [42–47], survival (e.g. HIPK1, MCL1, MAPKAPK2, VCP, VDAC2, APP, MAP3K5 [48–54]), migration/invasion (e.g. MUC1, ADAM9, SHPDX2A, CD44, PAK6, GIT1, PTPN1) [55–61], ER-signaling (e.g. BCAS2, MUC1, NCOA3, TFAP2C) [62–65], and maintenance of genome integrity (e.g. NBN, RAD21, FANC, BUB3, RAD9A, TAOK1, RAD51C, RAD7) [66–73]. Also represented are several “druggable” classes [74], like kinases (e.g. HIPK1, MAPKAPK2, MELK, RPS6KB2, PAK1, TAOK1, IP4K2B, RPS6KB1, TLR2, MAP3K3), phosphatases (e.g. PTPN1), proteases (e.g. ADAM9), G protein-coupled receptors (e.g. GPRC5C) and ion channels (e.g. VDAC2).

**Discussion**

Using whole-genome DNA microarrays, we collected transcriptional and genomic profiles across a set of 52 widely used breast cancer cell lines, with the primary goals to establish their suitability in modeling known breast tumor heterogeneity, and to create a resource for cancer gene discovery. Cluster analysis of transcriptional profiles defined three cell line subtypes, one luminal and two basal (A and B), consistent with other recent studies of breast cancer cell lines [31,33,75]. The luminal subtype included all ER-positive cell lines, and associated gene expression patterns reflected both ER and luminal differentiation pathways, the latter including GATA3 and FOXA1, key transcriptional mediators of luminal differentiation [28,76]. The basal-like cell lines were ER-negative and exhibited more frequent mutations of TP53 and PTEN, consistent with findings in basal-like tumors [3,77]. The basal-A subtype exhibited enriched expression of ETS pathway genes, a pathway linked to diverse tumor phenotypes including invasion and metastasis [78]. The basal-B subtype, which included the three non-tumorigenic lines (consistent with prior studies [75]), as well as five highly invasive/metastatic lines with features of EMT, exhibited enriched expression of EMT and EGF regulated genes, the latter pathway also previously linked to basal-like tumors [79].

Recently, Neve et al. [33] profiled 51 breast cancer cell lines (though using a lower-resolution (~1 Mb) CGH platform), 38 of which (~3/4th) overlapped with the 52 we profiled. All the overlapping lines except for one clustered into the same corresponding gene-expression subtype in both their and our study. The exception was HCC1500, which we classified as luminal while Neve et al. labeled it as basal B. The discrepancy may reflect a cell line identification error. We note that ATCC describes the line as ER-positive, more consistent with a luminal classification.

Our comparisons of expression profiles between breast cancer cell line subtypes and breast tumor subtypes provided valuable information relevant to the suitability of cell lines in modeling known breast tumor heterogeneity. Luminal-A/B tumors best matched luminal cell lines. Notably, basal-like tumors most corresponded to basal-A cell lines. Consistent with this finding, two breast cancer cell lines from BRCA1 mutation carriers also clustered in basal-A (and basal-A lines exhibited enrichment of a BRCA1 signature), where it has been established that BRCA1-associated tumors share many features with sporadic basal-like tumors [80]. Interestingly, ERBB2-associated tumors matched both luminal and basal-A lines. While ERBB2 represents a distinct expression tumor subtype in multiple independent cohorts [3,15,81], it is noteworthy that most ERBB2 (HER2+) cell lines clustered in the luminal subtype. The basis for the discrepant ERBB2 grouping in cell lines and tumors is unclear but warrants further investigation.

It has been suggested that the origin of the luminal vs. basal breast cancer distinction reflects the transformation of different breast epithelial progenitor cell compartments [82,83]. Breast epithelial stem/progenitor cells support mammary gland development during puberty and subsequent growth and remodeling during pregnancy [84]. A prevailing view is that breast epithelial stem cells give rise to bipotent basal/luminal progenitors, which then give rise to basal and luminal restricted progenitors, and from there to differentiated basal/myoepithelial and luminal epithelial tumors [80].

**Figure 4. Cell line subtypes exhibit distinct genomic instabilities.** Fraction of genome comprising (A) high-level DNA amplification; or (B) low-level gain/loss, stratified by cell line subtype (luminal, basal-A, basal-B). Box plots show 25th, 50th and 75th percentiles; P-values (Students t-test) for pairwise comparisons are shown. doi:10.1371/journal.pone.0006146.g004
## Table 4. High-amplitude amplifications and deletions.

<table>
<thead>
<tr>
<th>Cytoband</th>
<th>P-Border (nt)</th>
<th>Q-Border (nt)</th>
<th>Size (kB)</th>
<th>Cell Lines</th>
<th>Significant DNA-RNA Correlations</th>
<th>Other notable genes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>AMPLIFICATION</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1p22.1-1p21.3</td>
<td>93549298</td>
<td>97052934</td>
<td>3504</td>
<td>SUM44*</td>
<td>DR1, FNBPI, ARHGAP29, ALG14</td>
<td></td>
</tr>
<tr>
<td>1p13.3</td>
<td>10773670</td>
<td>109306637</td>
<td>1568</td>
<td>HCC2688</td>
<td>C1orf59, PRPF38B, STXBp3, GSM2, LCC1</td>
<td>VAV3</td>
</tr>
<tr>
<td>1p13.2</td>
<td>114220960</td>
<td>115183599</td>
<td>963</td>
<td>MCF7, UACC812</td>
<td>AP4B1, DLCRE1B, HIF1K, TRIM33, BCA52, CSDE1, NRAS</td>
<td></td>
</tr>
<tr>
<td>1q21.2</td>
<td>148738080</td>
<td>148858763</td>
<td>148</td>
<td>HCC1143</td>
<td>TARS2, MCL1, ENSA, GOLPH3L</td>
<td></td>
</tr>
<tr>
<td>1q21.2-q21.3</td>
<td>149460307</td>
<td>150130540</td>
<td>670</td>
<td>HCC712, UACC812</td>
<td>PIP5K1A, PSMD4, ZNF687, PIK4B, PSMB4, POG2, SNX27, MRPL9</td>
<td></td>
</tr>
<tr>
<td>1q21.3</td>
<td>151000411</td>
<td>151885402</td>
<td>885</td>
<td>HCC712</td>
<td></td>
<td></td>
</tr>
<tr>
<td>1q22</td>
<td>153424958</td>
<td>153999982</td>
<td>575</td>
<td>UACC812</td>
<td>MUC1, C1orf2, CLK2, HCN3, PKLr, C1orf104, RUSC1, ASH1L, YY1AP1</td>
<td></td>
</tr>
<tr>
<td>1q23.3</td>
<td>159283361</td>
<td>159357995</td>
<td>75</td>
<td>SUM190</td>
<td>KLHDC9</td>
<td></td>
</tr>
<tr>
<td>1q22</td>
<td>204736293</td>
<td>205144756</td>
<td>408</td>
<td>UACC812</td>
<td>MAPKAPK2</td>
<td>IKBKE</td>
</tr>
<tr>
<td>3p14.2-p14.1</td>
<td>61765808</td>
<td>64574645</td>
<td>2809</td>
<td>MCF7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3q26.32</td>
<td>178223920</td>
<td>180535525</td>
<td>2312</td>
<td>HCC2185</td>
<td>TBL1XR1, ZNF639</td>
<td>PIK3CA</td>
</tr>
<tr>
<td>3q29</td>
<td>194971434</td>
<td>195513283</td>
<td>542</td>
<td>HCC1937</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3q29</td>
<td>196888266</td>
<td>196931777</td>
<td>49</td>
<td>HCC1937</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4q12</td>
<td>53304442</td>
<td>54084198</td>
<td>780</td>
<td>HCC1007</td>
<td>SCFD2, FIP1L1</td>
<td></td>
</tr>
<tr>
<td>5p15.33</td>
<td>712977</td>
<td>2811691</td>
<td>2099</td>
<td>HCC1954</td>
<td>ZDHHC11, PDCD6, MRPL36, NDUFS6</td>
<td>TERT</td>
</tr>
<tr>
<td>6p12.1</td>
<td>53538212</td>
<td>57236103</td>
<td>1878</td>
<td>HCC1007</td>
<td>KIAA11586, ZNF451, BAG2</td>
<td></td>
</tr>
<tr>
<td>6q16.3-q21</td>
<td>104858272</td>
<td>109112665</td>
<td>4254</td>
<td>HCC2185</td>
<td>HACE1, ATG5, C6orf203, PDS52, SEC63, OSTM1, SNX3, FOXO3A</td>
<td></td>
</tr>
<tr>
<td>6q21-q22.31</td>
<td>111961945</td>
<td>123089199</td>
<td>11127</td>
<td>HCC2185</td>
<td>C6orf225, HDAC2, DSE, GOPC, NUS1, ASF1A, HSF2, SERINC1</td>
<td></td>
</tr>
<tr>
<td>7p15.2</td>
<td>26557965</td>
<td>27107611</td>
<td>550</td>
<td>HCC1007</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7p11.2</td>
<td>54595526</td>
<td>55931398</td>
<td>1336</td>
<td>BT20, MDA468</td>
<td>EGFR</td>
<td></td>
</tr>
<tr>
<td>7q21.3-q21.2</td>
<td>90779687</td>
<td>91868629</td>
<td>1089</td>
<td>SUM52</td>
<td>MTERF, AKAP9, CYP51A1, KRT1, ANKIB1</td>
<td></td>
</tr>
<tr>
<td>7q21.3</td>
<td>95239813</td>
<td>96499191</td>
<td>1250</td>
<td>SUM52</td>
<td>SLC25A13, SHFM1</td>
<td></td>
</tr>
<tr>
<td>7q22.1</td>
<td>100294293</td>
<td>100421513</td>
<td>127</td>
<td>SUM52</td>
<td>SLC12A9</td>
<td></td>
</tr>
<tr>
<td>8p12.1</td>
<td>21593811</td>
<td>21964633</td>
<td>373</td>
<td>MDA134</td>
<td>XPO7</td>
<td></td>
</tr>
<tr>
<td>8p12.1-p11.21</td>
<td>32328805</td>
<td>41907423</td>
<td>9579</td>
<td>BT483, HCC1500, HCC1599, MDA134, SUM44*, SUM52</td>
<td>FUT10, C8orf41, MAK16, ZNF703, ERLIN2, PROSC, BRF2, RAB11F1P1, EIF4EBP1, ASH2L, LSM1, BAG4, DHD2, WHSC1L1, LETM2, FGFR1, TACC1, PLEKHA2, TM2D2, ADAM9, GOLEGA7, AGPAT</td>
<td>8q12.2-q12.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8q13.3</td>
<td>71707355</td>
<td>72299610</td>
<td>1292</td>
<td>SKBR3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8q21.11-q21.13</td>
<td>79781799</td>
<td>85260376</td>
<td>5479</td>
<td>EFM192A, HCC1419, HCC1599, SKBR3</td>
<td>HEY1, TPD52, ZBTB10</td>
<td></td>
</tr>
<tr>
<td>8q21.3-q22.1</td>
<td>89113344</td>
<td>95233478</td>
<td>6120</td>
<td>EFM192A, HCC1419, SKBR3</td>
<td>OSGIN2, NBN, DECR1, OTUD6B, RBM12B, TMEM57</td>
<td></td>
</tr>
<tr>
<td>8q22.2-q23.2</td>
<td>100879473</td>
<td>101995283</td>
<td>1116</td>
<td>HCC1419, HCC2185</td>
<td>COX5C, POLR2K</td>
<td></td>
</tr>
<tr>
<td>8q22.3</td>
<td>104311423</td>
<td>104550566</td>
<td>239</td>
<td>HCC1419</td>
<td>FZD6</td>
<td></td>
</tr>
<tr>
<td>8q23.1-q24.2</td>
<td>108267427</td>
<td>131134620</td>
<td>22867</td>
<td>EFM192A, HCC1419, HCC1599, HCC2185, SKBR3, ZR75-30</td>
<td>EIF3E, TRPS1, EIF3H, C8orf53, RAD21, TAF2, DSCC1, MRPL13, MTBP, DLR1, WDR67, C8orf76, ZHX1, ATAD2, C8orf52, FAM91A1, TMEM65, TRMT12, RNF139, TATON1, NDUBF9, SQLE, KIAA0196, NSMCE2, FAM84B</td>
<td>MYC</td>
</tr>
<tr>
<td>8q24.22</td>
<td>133917771</td>
<td>134337653</td>
<td>420</td>
<td>ZR75-30</td>
<td>PHF20L1</td>
<td></td>
</tr>
<tr>
<td>8q24.3</td>
<td>141658961</td>
<td>143348731</td>
<td>1690</td>
<td>HCC1419, MDA436, ZR75-30</td>
<td>GP2, FLJ43860</td>
<td></td>
</tr>
<tr>
<td>8q24.3</td>
<td>144310706</td>
<td>144753628</td>
<td>443</td>
<td>MDA436, ZR75-30</td>
<td>ZFP41, GLJ4, ZNF696, C8orf51, RHPN1, MAFA</td>
<td></td>
</tr>
<tr>
<td>Cytoband</td>
<td>P-Border (nt)</td>
<td>Q-Border (nt)</td>
<td>Size (kB)</td>
<td>Cell Lines</td>
<td>Significant DNA-RNA Correlations</td>
<td>Other notable genes</td>
</tr>
<tr>
<td>----------</td>
<td>--------------</td>
<td>--------------</td>
<td>-----------</td>
<td>------------</td>
<td>----------------------------------</td>
<td>--------------------</td>
</tr>
<tr>
<td>8q24.3</td>
<td>145137850</td>
<td>146252219</td>
<td>1114</td>
<td>BT483, HCC1419, MDA436, ZR7-50</td>
<td>GRINA, OPLAH, SHARPIN, KIAA1833, FBXL6, CPSF1, VPS28, KIFC2, ZNF252</td>
<td></td>
</tr>
<tr>
<td>9p13.3-p13.2</td>
<td>33876876</td>
<td>38058023</td>
<td>4181</td>
<td>HCC2185</td>
<td>UBE2R2, UBAP2, WD40R40, KIF24, KIAA1161, DCTN3, GALT, IL11RA, VCP, FANCG, PIGO, STOML2, RUSC2, TESK1, CD72, C9orf100, TN1, CREB3, RGP1, HINT2, CLTA, RNF38, MELK, ZCHC7, GRHPR, ZBTB5, POLR1E, FBXO10, R9MTD3, WDR32, MCART1</td>
<td></td>
</tr>
<tr>
<td>9q33.3</td>
<td>128307884</td>
<td>129195638</td>
<td>888</td>
<td>SUM44*</td>
<td>RALGFP1</td>
<td></td>
</tr>
<tr>
<td>10q21.1-q21.2</td>
<td>72507196</td>
<td>73792767</td>
<td>1290</td>
<td>HCC2157</td>
<td>DNAJB12</td>
<td></td>
</tr>
<tr>
<td>10q22.2-q22.3</td>
<td>76461776</td>
<td>82106491</td>
<td>5645</td>
<td>EFM19, HCC2157</td>
<td>SAMDH, VDAC2, DLG5, POLR3A, RPS24, LOC283080, ZM12, PP1F, SFTP1A, FAM22E, C10orf57, ANXA11</td>
<td></td>
</tr>
<tr>
<td>10q24.33-q25.1</td>
<td>105307581</td>
<td>106054698</td>
<td>747</td>
<td>EFM19</td>
<td>SH3PD2A</td>
<td></td>
</tr>
<tr>
<td>10q26.13</td>
<td>124598599</td>
<td>124962466</td>
<td>364</td>
<td>SUM52</td>
<td>IKZF5, BUB3</td>
<td></td>
</tr>
<tr>
<td>9p13.3-q13.4</td>
<td>68427956</td>
<td>70812048</td>
<td>2384</td>
<td>HCC1143, HCC1500, HCC1954, MDA134, MDA175, MDA361, SUM44*, MDA190,</td>
<td>IGHMBP2, FADD, PPFIA1, CTNN, SHANK2, CCND1</td>
<td></td>
</tr>
<tr>
<td>11p13</td>
<td>33062705</td>
<td>35600197</td>
<td>2537</td>
<td>HCC1806*</td>
<td>HIPK3, FBXO3, CAPRIN1, NAT10, ABT2B, CAT, APIP, PDHX, CD44</td>
<td></td>
</tr>
<tr>
<td>11q13.2</td>
<td>66874536</td>
<td>67198753</td>
<td>324</td>
<td>MDA134, ZR7-51</td>
<td>RAD9A, RPS6KB2, CORO1B, TME1M34</td>
<td></td>
</tr>
<tr>
<td>11q13.3-q13.4</td>
<td>68427956</td>
<td>70812048</td>
<td>2384</td>
<td>HCC1143, HCC1500, HCC1954, MDA134, MDA175, MDA361, SUM44*, MDA190,</td>
<td>IGHMBP2, FADD, PPFIA1, CTNN, SHANK2, CCND1</td>
<td></td>
</tr>
<tr>
<td>11q13.4</td>
<td>73316198</td>
<td>73649077</td>
<td>333</td>
<td>BT474, MDA134, SUM190</td>
<td>UCP2, C2CD3, PPM1E</td>
<td></td>
</tr>
<tr>
<td>11q13.4-q14.1</td>
<td>74648813</td>
<td>77963474</td>
<td>3315</td>
<td>MDA134, SUM44*, SUM52, SUM190</td>
<td>ARRB1, PRRRIR, EMSY, PHCA, PAK1, AQP11, CLINS1A, C11orf67, INT5A, NDUF2C, ALG8, GAB2, NAR52</td>
<td></td>
</tr>
<tr>
<td>12p12.3</td>
<td>18727378</td>
<td>19246201</td>
<td>519</td>
<td>HCC1500</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12q21.31-q21.33</td>
<td>88265969</td>
<td>88443930</td>
<td>178</td>
<td>SUM52</td>
<td>WDR51B, GALNT4</td>
<td></td>
</tr>
<tr>
<td>13q22.2-q31.1</td>
<td>74756931</td>
<td>78096263</td>
<td>3339</td>
<td>UACC812</td>
<td>UCHL3</td>
<td></td>
</tr>
<tr>
<td>13q31.3-q32.1</td>
<td>90798074</td>
<td>93942902</td>
<td>3145</td>
<td>UACC812</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16q12.2</td>
<td>51800892</td>
<td>53524601</td>
<td>1724</td>
<td>EFM19, SUM44*</td>
<td>CHD9, FTO</td>
<td></td>
</tr>
<tr>
<td>17p12</td>
<td>12611513</td>
<td>13636592</td>
<td>1025</td>
<td>EFM192A</td>
<td>ELAC2</td>
<td></td>
</tr>
<tr>
<td>17q11.2</td>
<td>23686912</td>
<td>24013273</td>
<td>326</td>
<td>ZR7-50</td>
<td>POLDIP2, TREM199, SLC46A1, PIGS, SPAG5, FLJ25006, KIAA1000, SDF2</td>
<td></td>
</tr>
<tr>
<td>17q11.2</td>
<td>24894649</td>
<td>25818484</td>
<td>924</td>
<td>HCC202</td>
<td>TAO91, LOC116236, GIT1, ANKRD13B, CPD</td>
<td></td>
</tr>
<tr>
<td>17q11.2</td>
<td>27727543</td>
<td>28293356</td>
<td>566</td>
<td>SUM190</td>
<td>ZNF207</td>
<td></td>
</tr>
<tr>
<td>17q12</td>
<td>31206068</td>
<td>31649844</td>
<td>444</td>
<td>MDA361</td>
<td>FLJ12120</td>
<td></td>
</tr>
<tr>
<td>17q12-q12.2</td>
<td>32627885</td>
<td>36209712</td>
<td>3582</td>
<td>BT474, EFM192A, HCC102, HCC1419, HCC1500, HCC1954, HCC2218, MDA361, SKB3, SUM190, UACC812, UACC893, ZR7-50</td>
<td>ACACA, TADA2L, DDX52, SOCS7, MALT7, CID3, PIF2F2, PSM38, PP6K2, CDDC49, RPL23, LPS1P, CAPN1B, FAM153C, RP19, LOC90101, FBXL20, MED1, PPP1R1B, STARD3, TCAP, PERLD1, ERBB2, C17orf37, GRB7, IKZF3, GSDML, ORMID3, PDM3D, MED24, MSL-1, CASC3, CDC6, RARA, SMARCE1</td>
<td></td>
</tr>
<tr>
<td>17q21.31</td>
<td>38419019</td>
<td>38738864</td>
<td>320</td>
<td>SUM190</td>
<td>RND2</td>
<td></td>
</tr>
<tr>
<td>17q21.32-q25.1</td>
<td>43329972</td>
<td>50826668</td>
<td>7497</td>
<td>BT474, EFM192A, HCC102, HCC1419, HCC1500, HCC1954, HCC2218, MDA361, SKB3, SUM190, UACC812, UACC893, ZR7-50</td>
<td>SP2, PNP0, CDK5RAP3, SNX11, HOXB13, CALCOCO2, AP3S1G1, UBE2Z, SNF8, ZNF652, PHB, SPOP, SLC35B1, FAM117A, MYST2, PDQ2, XLYT2, MRPL27, LURCS9, EME1, ACSF2, RSAD1, EPN3, SPAT20, ABCC3, ANKHMD40, CROP, TOB1, NEM1, TOM11L, COX11, STXBP4</td>
<td></td>
</tr>
<tr>
<td>17q23.2-q24.2</td>
<td>53282667</td>
<td>63106134</td>
<td>9823</td>
<td>BT474, HCC1721, HCC2218, MCF7, MDA361, ZR7-50</td>
<td>SFR51, DYNLL2, MK51, SUPHTH1, MTM4, RAD51C, TRIM37, FAM33A, C17orf71, YPEL2, DHX40, CLTC, PTRH2, TME49, TUBD1, RPS6KB1, RNFT1, HEATR6, USP32, APPB8P2, PPMT1D, BRIP1, INT52, MED13, METTL2A, TLR2, TAN2C, CYB561, WDR68, CDDC44, MAP3K3, LK5, CCDC47, DDX42, PSM5C, SMARCD2, DDX5, CCDC45, SMURF2, GNA13, HELZ</td>
<td></td>
</tr>
</tbody>
</table>
cells [84,85]. Bipotent human breast epithelial stem/progenitors have been characterized with the cell surface phenotype MUC[80,85]/low/CALCA[39]/low/+ [39]. Separately, breast cancer stem cells, identified prospectively as tumor initiating cells when transplanted into immunodeficient mice, have been characterized by the surface expression phenotype CD44+/CD24−/low [40], also a presumed phenotype of normal breast epithelial stem or early progenitor cells [84].

Our transcriptional profiles of breast cancer cell lines are consistent with an origin in (or at least a likeness of the bulk cell population to) the various stem/progenitor cell compartments. Basal-B lines predominantly express CD44+/CD24−/low and

### Table 4. Cont.

<table>
<thead>
<tr>
<th>Cytoband</th>
<th>P-Border (nt)</th>
<th>Q-Border (nt)</th>
<th>Size (kB)</th>
<th>Cell Lines*</th>
<th>Significant DNA-RNA Correlations*</th>
<th>Other notable genes*</th>
</tr>
</thead>
<tbody>
<tr>
<td>17q25.1</td>
<td>6975691</td>
<td>71418122</td>
<td>1662</td>
<td>HCC2218, MDA361, MDA453, UACC939</td>
<td>GPRC3C, SLC9A3R1, NAT9, TMEM104D, FDXR, C17orf28, CDR2L, ICT1, KCTD2, SUMO2, NUP95, GGA3, MRPS7, MIF4GD, SLC25A19, GRB2, CASK2, TSEN54, MYO1SB, SAP30BP, HIF3B, UNK, WBP2</td>
<td></td>
</tr>
<tr>
<td>18q21.32-q21.33</td>
<td>55178911</td>
<td>57628085</td>
<td>2449</td>
<td>HCC1500</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19p13.2</td>
<td>14932742</td>
<td>15602448</td>
<td>670</td>
<td>HCC1143</td>
<td>ILVBL, BRD4, AKAP8L</td>
<td></td>
</tr>
<tr>
<td>19q12-q13.11</td>
<td>3396349</td>
<td>38052482</td>
<td>4086</td>
<td>HCC1569, HCC1599</td>
<td>UQCRFS1, POP4, PLEKH1, C19orf2, DYP19L3, ANKRD27</td>
<td></td>
</tr>
<tr>
<td>19q13.11</td>
<td>39866832</td>
<td>40146793</td>
<td>280</td>
<td>HCC1599</td>
<td></td>
<td></td>
</tr>
<tr>
<td>19q13.42</td>
<td>6051045</td>
<td>60898029</td>
<td>347</td>
<td>EFM19</td>
<td>FIZ1, ZNF784, CCDC106</td>
<td></td>
</tr>
<tr>
<td>19q13.43</td>
<td>63208125</td>
<td>63774724</td>
<td>567</td>
<td>HCC1806*</td>
<td>ZNF329, ZNF274, ZNF8, ZSCAN22, ZNF324, TRIM28, CHMP2A, UBE2M</td>
<td></td>
</tr>
<tr>
<td>20p12.2</td>
<td>10224083</td>
<td>10433564</td>
<td>209</td>
<td>HCC2185</td>
<td>MKK5</td>
<td></td>
</tr>
<tr>
<td>20q11.22</td>
<td>32363269</td>
<td>33563203</td>
<td>1200</td>
<td>BT474</td>
<td>DYNLRL1, NCOA6, UQCC</td>
<td></td>
</tr>
<tr>
<td>20q13.12</td>
<td>42493067</td>
<td>43286511</td>
<td>793</td>
<td>BT474, SUM52</td>
<td>SERINC3</td>
<td></td>
</tr>
<tr>
<td>20q13.12-q13.13</td>
<td>45234836</td>
<td>48636574</td>
<td>3402</td>
<td>BT474, HCC1419, MCF7</td>
<td>NCOA3, PREX1, ARFGF2, STAUI, DDX27, ZNFX1, SLC9A8, SPATA2, PTPN1</td>
<td></td>
</tr>
<tr>
<td>20q13.13-q13.32</td>
<td>49139330</td>
<td>57334442</td>
<td>8195</td>
<td>BT474, HCC1419, MCF7, SKBR3</td>
<td>ZFPM6, ZNF217, BCA51, PDN4, C20orf108, CSTF1, C20orf43, TAFAP2C, BMP7, RAE1, RBM38, RAB22A, VAPA, STX16, NPEPL1, GNAT, TH1L, ATP5E, SLMO2</td>
<td></td>
</tr>
<tr>
<td>20q13.33</td>
<td>61801252</td>
<td>62370522</td>
<td>569</td>
<td>HCC1419</td>
<td>PRR17, OPRL1</td>
<td></td>
</tr>
<tr>
<td>22q11.21</td>
<td>18256420</td>
<td>19686015</td>
<td>1430</td>
<td>SUM190</td>
<td>COMT, HTFGC, PHKA</td>
<td></td>
</tr>
<tr>
<td>22q12.1</td>
<td>24895479</td>
<td>25885840</td>
<td>990</td>
<td>HCC202</td>
<td>HPS4</td>
<td></td>
</tr>
<tr>
<td>Xp11.23-p11.22</td>
<td>48635684</td>
<td>51225253</td>
<td>2590</td>
<td>HCC712</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Xp11.22</td>
<td>52255712</td>
<td>54236019</td>
<td>1980</td>
<td>HCC202</td>
<td>TMEM29, PHF8</td>
<td></td>
</tr>
<tr>
<td>Xq28</td>
<td>148368959</td>
<td>149592006</td>
<td>1223</td>
<td>HCC202</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**For aberrations spanning multiple lines, inclusive interval indicated.**

*DNA but not RNA profiled.

*Only named genes listed, ordered by genome position; bold text indicates select known cancer genes.

*Within or immediately flanking interval.

doi:10.1371/journal.pone.0006146.t004
MUC+/CALLA+ phenotypes characteristic of stem or bipotent progenitor cells, as well as ITGB3 (CD61), also recently characterized as a cancer stem cell marker in MMTV-wnt-1 induced murine breast cancer [41]. In contrast, basal-A lines appear mainly CD44+/CD24-, but express PROM1 (aka CD133), a marker of luminal progenitors in mice [86] also more recently characterized as a stem cell marker in BRCA1-associated breast cancer [87], while luminal lines express markers of luminal lineage restriction like GATA3 and FOXA1 [28]. Conspicuously absent from our analysis is a breast tumor subtype corresponding to the stem-cell like (and sometimes mesenchymal-like) basal-B lines. Whether basal-B lines reflect an uncommon tumor subtype not yet characterized, or else a stem/progenitor subpopulation of tumor cells enriched in culture, or even an artifact of cell culture, remains to be determined. Regardless, breast cancer cell lines are likely to prove useful for discovering new stem cell markers, and for studying stem/progenitor cell biology.

Our genomic profiles of breast cancer cell lines indicate that overall the spectra of CNAs is reflective of breast tumors, consistent with prior findings from loss of heterozygosity (LOH) analysis [11]. Overall, however, cell lines exhibited higher frequencies and greater complexities of CNAs, and seemingly more than might be explained by a higher sensitivity of detecting CNAs in stromal-free tumor cell populations. Notably absent among the luminal subtype were the “simple” karyotypes characteristic of luminal-A tumors (i.e. 1q+, 16p+/16q–). By genomic profiles, luminal cell lines shared features characteristic of luminal-B tumors, including certain subtype-specific CNAs and overall higher levels of DNA amplification. Likewise, basal-A cell lines and basal-like tumors shared the feature of high levels of chromosome segment gain/loss. However, overall only a subset of subtype-specific CNAs was preserved. Therefore, at the genomic level it is uncertain how well cell line subtypes faithfully represent tumor subtype counterparts.

Taken together, the transcriptional and genomic profiles support the conclusion that luminal and basal-A cell lines are the most appropriate cell line models of luminal-B and basal-like tumors, respectively. Further, the basal lines are likely useful models for biological studies of the 70-gene, wound and hypoxia signatures. Despite incongruent expression results, luminal lines with amplification/overexpression of ERBB2 are likely appropriate models of ERBB2-associated tumors. Our findings indicate that new cell lines are needed to more faithfully model luminal-A tumors. Currently available cell lines likely reflect certain biases in the specimen source of cell line, and/or in the culturing methods, as suggested by the predominance of HCC lines (from UT Southwestern) among the basal-A group. Different culturing methods (e.g. ref. [88]) might support the establishment of cell lines from luminal-A tumors.

Our genomic profiles also identified numerous high-level DNA amplifications and multi-copy deletions, pinpointing known and novel cancer genes. Further, by integrating the genomic and transcriptional datasets, we could define a set of candidate cancer genes residing at these loci and exhibiting both altered copy number and expression. The larger set of amplified/overexpressed genes included several known breast cancer oncogenes, as well as many plausible candidates including genes with known functions relevant to carcinogenesis, like cell proliferation, survival and motility/invasion, and genome integrity (e.g. DNA damage response). Though genes maintaining genome integrity are more typically considered candidate tumor suppressors, the overexpression of such genes has been linked to genome instability [67,89].

The set of amplified/overexpressed genes also included many druggable targets [74], most notably several kinases. Importantly, the same cell lines used for discovery can also be used to functionally examine cancer gene candidates, for example using RNA interference to knockdown the expression of amplified oncogene candidates, and then assaying loss of tumorigenic phenotypes in cultured cells or in vivo (e.g. refs.[90,91]). Indeed, high-throughput RNA interference approaches [92,93] might be used to evaluate many or all of the candidate cancer genes simultaneously.

In summary, transcriptional and genomic profiling of 52 commonly used breast cancer cell lines identifies cell line subtypes, and defines the cell line subtypes that most faithfully capture the known heterogeneity of breast tumors. Specifically, luminal and basal-A lines appear to best model the features of luminal-B and basal-like tumors, while basal-B lines might inform stem cell biology. In addition, our integrated analysis of genomic and transcriptional profiles pinpoints loci and genes with altered copy number and expression, providing a rich source for discovery and future characterization of new breast cancer genes.

Supporting Information

**Table S1** 8,750 variably expressed genes (log2 ratios)
Found at: doi:10.1371/journal.pone.0006146.s001 (1.11 MB ZIP)
**Table S2** Processed aCGH data (log2 ratios)
Found at: doi:10.1371/journal.pone.0006146.s002 (2.82 MB ZIP)
**Table S3** Genes with significantly correlated copy number and expression
Found at: doi:10.1371/journal.pone.0006146.s003 (0.13 MB TXT)

Acknowledgments

We wish to thank the SFGF for microarray manufacture, SMD for database support, and members of the Pollack lab for helpful discussion.

Author Contributions

Conceived and designed the experiments: JK KS LG JKG AFG JDM JRP. Performed the experiments: JK KS MB YLC LG JKG KAK. Analyzed the data: JK KS MB JKG THB PW AFG JDM JRP. Contributed reagents/materials/analysis tools: KS LG THB AFG JDM. Wrote the paper: JK KS JRP.

References


gene expression programs in response to hypoxia: cell type specificity and prognostic
47. Borlado LR, Mendez J (2008) CDC6: from DNA replication to cell cycle
46. Borlado LR, Mendez J (2008) CDC6: from DNA replication to cell cycle
42. Zhang L, Smit-McBride Z, Pan X, Chen J, Rees CA, Jeffrey SS, et al. (2002) Role
40. Schuler GD (1997) Pieces of the puzzle: expressed sequence tags and the
30. Schuler GD (1997) Pieces of the puzzle: expressed sequence tags and the


