Research article

Open Access

Somatic sequence alterations in twenty-one genes selected by expression profile analysis of breast carcinomas

Stephen J Chanock[1,2], Laurie Burdett[2,3], Meredith Yeager[2,3], Victor Llaca[3], Anita Langerød[4], Shafaq Presswalla[2,3], Rolf Kaaresen[5], Robert L Strausberg[6], Daniela S Gerhard[7], Vessela Kristensen[1,4,8], Charles M Perou[9] and Anne-Lise Børresen-Dale[4,8]

[1]Section of Genomic Variation, Pediatric Oncology Branch, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892-4605, USA
[2]Core Genotyping Facility, National Cancer Institute, National Institutes of Health, Bethesda, Maryland 20892-4605, USA
[3]Intramural Research Support Program, SAIC-Frederick, NCI-FCRDC, Frederick, Maryland 21702, USA
[4]Department of Genetics, Institute for Cancer Research, Rikshospitalet-Radiumhospitalet Medical Center, Montebello, 0310 Oslo, Norway
[5]Department of Surgery, Ullevål University Hospital, 0407 Oslo, Norway
[6]J Craig Venter Institute, Medical Center Drive, Rockville, Maryland 20850, USA
[7]Office of Cancer Genomics, National Cancer Institute, NIH, Bethesda, Maryland 20892, USA
[8]Medical Faculty, University of Oslo, 0316 Oslo, Norway
[9]Departments of Genetics and Pathology, Laboratory Medicine, Lineberger Comprehensive Cancer Center, University of North Carolina, Chapel Hill, North Carolina 27599, USA

Corresponding author: Stephen J Chanock, chanocks@mail.nih.gov

Received: 29 Aug 2006 Revisions requested: 3 Oct 2006 Revisions received: 15 Dec 2006 Accepted: 15 Jan 2007 Published: 16 Jan 2007

Breast Cancer Research 2007, **9**:R5 (doi:10.1186/bcr1637)
This article is online at: http://breast-cancer-research.com/content/9/1/R5

Abstract

Introduction Genomic alterations have been observed in breast carcinomas that affect the capacity of cells to regulate proliferation, signaling, and metastasis. Re-sequence studies have investigated candidate genes based on prior genetic observations (changes in copy number or regions of genetic instability) or other laboratory observations and have defined critical somatic mutations in genes such as *TP53* and *PIK3CA*.

Methods We have extended the paradigm and analyzed 21 genes primarily identified by expression profiling studies, which are useful for breast cancer subtyping and prognosis. This study conducted a bidirectional re-sequence analysis of all exons and 5', 3', and evolutionarily conserved regions (spanning more than 16 megabases) in 91 breast tumor samples.

Results Eighty-seven unique somatic alterations were identified in 16 genes. Seventy-eight were single base pair alterations, of which 23 were missense mutations; 55 were distributed across conserved intronic regions or the 5' and 3' regions. There were nine insertion/deletions. Because there is no *a priori* way to predict whether any one of the identified synonymous and noncoding somatic alterations disrupt function, analysis unique to each gene will be required to establish whether it is a tumor suppressor gene or whether there is no effect. In five genes, no somatic alterations were observed.

Conclusion The study confirms the value of re-sequence analysis in cancer gene discovery and underscores the importance of characterizing somatic alterations across genes that are related not only by function, or functional pathways, but also based upon expression patterns.

Introduction

Somatic mutations in key genes, such as oncogenes and tumor suppressor genes, have been reported to contribute to the risk for development of many human cancers. Genomic alteration has been shown to confer altered capacity for cell proliferation, metastasis, and responsiveness to either normal cellular signals or therapeutic agents [1]. Advances in sequence technology have lead to renewed efforts in the discovery and characterization of somatic mutations in different cancers. This avenue of investigation has emerged as a promising approach to dissect the profile of genetic alterations of cancer in order to better classify cancer subtypes, identify new mechanisms of carcinogenesis, and characterize possible biomarkers for susceptibility and outcome [2,3]. In fact, it is

bp = base pairs; DWD = distance weighted discrimination; PCR = polymerase chain reaction; SNP = single nucleotide polymorphism; TTGE = temporal temperature gel electrophoresis.

anticipated that re-sequence analysis of complete cancer genomes will be pursued in the future, because of the confluence of two major trends, the availability of complete human genome sequences, and advances in sequencing technology and analysis. Although very promising, this approach is yet to be fully developed but has been fueled in part by the characterization of somatic sequence alterations in candidate gene studies of specific cancers, such as breast and colon cancer.

Because it is still a formidable task to sequence entire genomes, investigators have analyzed individual candidate genes chosen based on results of previous studies in cell lines, animal models, or other primary human tumors. Initial re-sequence studies examined individual candidate genes based on prior genetic observations (changes in copy number or regions of genetic instability) or those identified in animal or *in vitro* laboratory studies and have defined critical somatic mutations in genes such as *TP53* and *PIK3CA* [4-11]. To date, most studies have concentrated on coding regions and the adjacent intronic region, in search of mutations that alter the coding sequence or RNA splicing. Selected studies have extended the choice of candidate genes to include a complete gene family, such as the protein kinase family or tyrosine phosphatome [2,3,12-14]. Concentrated investigation in the protein kinase genes has been conducted because of prior evidence that selected genes, such as *PIK3CA*, are frequently mutated in breast cancer [7-11]. Stephens and coworkers [15] reported on the re-sequence analysis of the coding region of 518 protein kinase genes in breast, lung, and testicular cancer. Recently, Sjoblom and colleagues [16] surveyed somatic alterations in 13,023 genes in 11 breast and 11 colon cancer cell lines or xenografts.

The success of the candidate gene approach has provided the impetus for this study to re-sequence 21 genes chosen mainly based upon expression profiling studies. These genes vary in expression across breast carcinoma samples and have been shown to be useful for breast cancer subtyping and prognosis [17-19]. Overall, copy number of the genes was not changed in the breast tumors, as assessed by array-based comparative genomic hybridization analyses [20]. Herein, we report the re-sequence analysis in 91 primary breast tumors of coding and noncoding regions of genes drawn from a novel paradigm, which was to select genes that have an altered expression pattern across breast carcinomas.

Materials and methods

DNA sampling and sequence analysis

Genomic DNA from 91 tumor samples were included in this study, of which 82 were from Norwegian breast cancer cases and nine were from a new breast cancer study in North Carolina. The Norwegian breast cancer samples were selected from a series of 215 previously published primary breast cancer samples, of which 63 of the 82 included in this study have been analyzed using cDNA microarrays (Langerod and coworkers, unpublished data). Patient tissue samples were sequentially collected at Ullevål University Hospital from 1990 to 1994 under an institutional review board approved protocol.

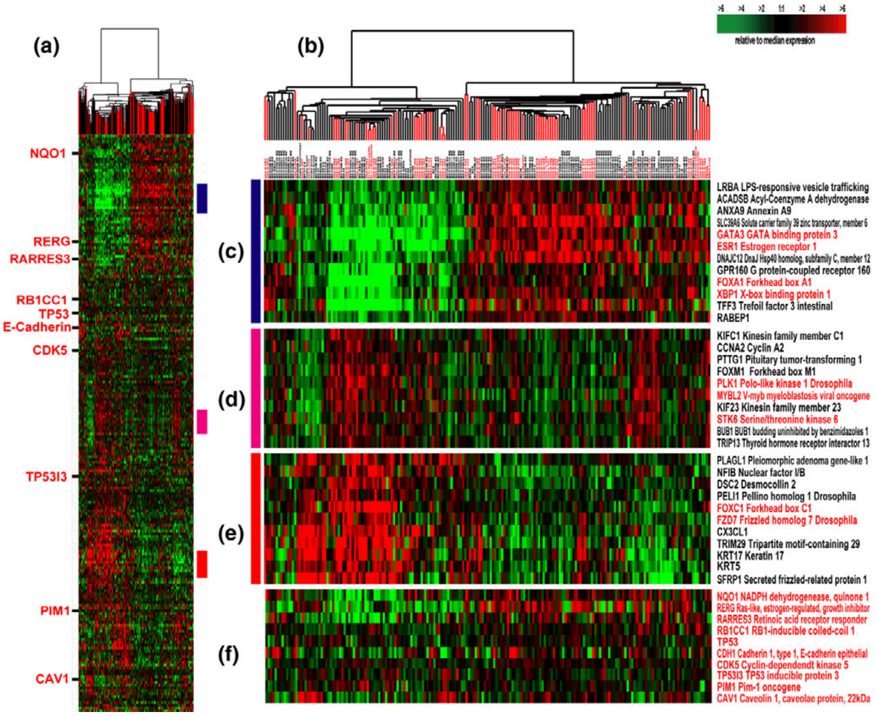
Primary breast carcinoma tissue was snap frozen and stored at -80°C. Frozen sections stained with hematoxylin/eosin were reviewed to confirm tumor content. More than 80% of the samples analyzed had more than 40% tumor cell content. Blood samples were collected in EDTA tubes and frozen at -40°C before DNA was isolated. DNA was extracted from both peripheral blood cells and tumor tissue using a method of chloroform/phenol extraction followed by ethanol precipitation (Nuclear Acid Extractor 340A; Applied Biosystems, Foster City, CA, USA), according to standard procedures. Matched control genomic DNA was available from peripheral blood from 36 of the Norwegian breast cancer cases.

Of the set of 21 genes selected for this re-sequencing analysis, 13 of them (*FZD7*, *NQO1*, *MYBL2*, *PLK1*, *STK6*, *ESR1*, *FOXA1*, *FOXC1*, *GATA3*, *RARRES3*, *RERG*, *XBP1*, and *CDK5*) were selected primarily based on their variation in gene expression patterns from previous studies of breast carcinomas [17,18] and eight were selected based on previous reports that they harbor somatic mutation in breast cancers (*CAV1*, *CDH1*, *FBXW7*, *PIM1*, *PIN1*, *TP53*, *TP53I3*, and *RB1CC1*), although they showed considerable variation in expression patterns (Figure 1 and Additional file 1).

Sequencing primers were designed for bidirectional sequence analysis using Primer3 software [21]. Each oligonucleotide was extended with a universal sequencing primer: M13 forward (TGTAAAACGACGGCCAGT) or M13 reverse (CAGGAAACAGCTATGACC). Primers and conditions are posted on the SNP500 Cancer website [22]. Standard cycle sequence analysis was performed (MJ Research PTC-200 Thermacycler) (MJ Sciences Waltham, MA, USA). Polymerase chain reaction (PCR) products were cleaned up with Exonuclease I/Shrimp Alkaline Phosphatase (USB, Cleveland, OH, USA). PCR products were sequenced using a modified ABI Prism® BigDye Terminator protocol (ABI, Foster City, CA, USA). Pgem®-3Zf(+) (Promega Corp., Madison, WI, USA) was used for controls in all sequencing reactions. The sequencing reactions were cleaned up by either Sephadex G-50 (Sigma, St Louis, MO, USA) spin columns in a MultiScreen®-HV 96-well filter plate (Millipore, Billerica, MA, USA) or Performa® DTR 384-well spin plate (Edge BioSystems, Gaithersburg, MD, USA). The reactions were run on either ABI 3700 or ABI 3730XL (ABI). Sequence traces were reviewed by two independent reviewers.

Bidirectional sequence analysis included 166 exons (62,000 bp) and an additional 120,000 bp of noncoding sequence in tumor samples from the 91 cases of breast cancer. In each gene, sequence analysis targeted at least 2 to 3 kb upstream of the first exon and 2 kb downstream of the 3'-untranslated

Figure 1



Hierarchical clustering analysis of 194 breast tumor samples analyzed using the 'SAM264' patient survival associated gene set augmented with nine additional genes included in the resequencing analysis. **(a)** Hierarchical clustering overview that shows the overall context for the 21 genes. **(b)** Close up of the sample associated dendrogram, which identifies the tumor samples that were re-sequenced in red. **(c)** Luminal/ER$^+$ epithelial gene set showing coordinated expression of *ESR1*, *GATA3*, *FOXA1*, and *XBP1*. **(d)** Proliferation gene set showing expression of *STK6*, *MYBL2*, and *PLK1*. **(e)** Basal epithelial gene set showing the expression of *FOXC1* and *FZD7*. **(f)** The expression profiles of the additional genes that were re-sequenced but that did not fall into the previous expression patterns are shown, and their position in the larger cluster is also shown in panel a. All genes identified in red text were analyzed by re-sequencing in this study, and only *FBXW7* and *PIN1* were not included in this cluster analysis because their average expression levels did not meet the gene filtering criteria. ER, estrogen receptor.

region, as well as evolutionarily conserved intronic regions (defined as 75% or greater sequence similarity over a 200 bp fragment for alignment of mouse and human sequence) [23]. For each amplicon in which a sequence variant was observed in a tumor sample, sequence analysis was also performed in the SNP500 Cancer reference set of 102 individuals drawn from the four major ethnic groups of the USA and a set of 94 anonymized Norwegian women who were older than 55 years and had a history of two negative mammograms [24].

We estimated the mutation rate based on the number of somatic events observed divided by the total number of base pairs sequenced in the analysis. This calculation combines potentially functional mutations (for example, driver mutations) with passenger somatic alterations [15].

Hierarchical clustering analysis

A total of 194 breast tumors were analyzed by clustering analysis using a modified version of the 'SAM264' gene list [18]; the 'SAM264' gene set is the set of genes that were associated with survival as identified using a Significance Analysis of Microarray analysis and contained 10 of the 21 genes sequenced. We added the 11 remaining re-sequenced genes to the SAM264 list for clustering analysis. Initially, we created a single sample set that was a combined dataset of the previous 122 samples [18,19], and 63 tumors from Langerod and coworkers (unpublished data) and nine tumors from North Carolina that included most of the samples used for the resequencing analysis presented here. This combined sample set was used to guide gene selection for resequencing analyses.

Because these three sets of samples were assayed using different microarrays, the two-color cDNA microarray datasets [18,19] (Langerod and coworkers; unpublished data) and the nine Agilent A1 microarray experiments performed at University of North Carolina, they were pre-processed similarly and systematic array biases removed using distance weighted discrimination (DWD) [25]. First, gene annotation from each dataset was translated to UniGene Cluster IDs (Build #185) using the SOURCE database [26]. The pre-processing included an initial selection for genes that exhibited a signal intensity of greater than 30 units in both the Cy3 and Cy5 channels across at least 70% of the experiments, which caused *FBXW7* and *PIN1* to be removed from further analyses because of very low signal intensities. Next, we log^2 transformed the R/G ratio and then Lowess normalized the data

[27]. Missing values were imputed using the k-NN imputation algorithm ($k = 10$) described by Troyanskaya and coworkers [28]. The expression values for duplicated probes with the same Unigene cluster ID were collapsed using the median expression value. DWD was performed in a pairwise manner by first combining the dataset reported by Sørlie and coworkers [18] with that by Langerd and coworkers (unpublished data), and subsequently combining this with the University of North Carolina data. In the final step of pre-processing, each individual experiment (microarray) was normalized by setting the mean to zero and its standard deviation to one, and each gene was median centered. The DWD corrected data for the SAM264 genes plus nine additional genes was finally used in a two-way average linkage hierarchical cluster analysis using centered correlation across the 194 microarrays.

Results

In total, more than 16.2 megabases were sequenced and more than 95% of the targeted amplicons were analyzed. Eighty-seven unique somatic nucleotide variants were identified in 16 genes (*TP53*, *GATA3*, *CAV1*, *CDH1*, *ESR1*, *FBXW7*, *FOXA1*, *FOXC1*, *FZD7*, *MYBL2*, *PIN1*, *RB1CC1*, *RERG*, *STK6*, *TP53I3*, and *XBP1*; see Table 1 and Additional file 2 for detailed results for each sample). In five genes (*CDK5*, *NQO1*, *PLK1*, *PIM1*, and *RARRES3*) no somatic sequence alterations were observed. The majority of sequence alterations were observed once, although there were three missense mutations that were observed more than once. The distribution of the somatic alterations per tumor sample is shown in Figure 2. Fifty-three tumors had one or more somatic alteration, and in 38 tumor samples (42%) no somatic alterations were noted. The largest number of somatic alterations observed was seven in one sample, but these were distributed over four separate genes. The overall distribution of the single base pair somatic alterations favored transitions over transversions (50 versus 28).

Of the 87 total somatic alterations observed, 78 single base pair somatic alterations were distributed across 16 of the 21 genes analyzed (Table 1). For the purposes of this analysis, a single base pair somatic alteration was defined on the basis that it was observed in a tumor specimen but not in the constitutional DNA of 102 controls from the SNP500 Cancer set, or matched blood DNA of 36 of the Norwegian breast cancer cases. Of the 78 single pair somatic alterations, we observed 34 alterations in coding regions; 23 were missense alterations and 11 were predicted to be synonymous changes (for example, no alteration of the predicted amino acid). In noncoding regions, 44 single base pair somatic alterations were observed, of which 27 were in evolutionarily conserved intronic regions, 12 in the analyzed 5' region, and five in the analyzed 3' region.

Sequence analysis of the tumor samples identified 252 single nucleotide polymorphisms (SNPs), all of which were confirmed in blood samples drawn from 94 Norwegian women with no history of breast cancer and the reference SNP500 Cancer set [22].

We observed 23 missense mutations in eight of the 21 genes studied. *TP53* and *GATA3* were notable because of the large number of sequence alterations observed, which included missense mutations and insertion/deletions previously reported [4-6,29]; in total, there were 14 distinct missense mutations, one pre-terminal stop codon, four insertion/deletions, and five noncoding alterations. For both of these genes, the majority of sequence variants have been shown to be functionally significant somatic mutations, and thus could be considered as 'driver' mutations for oncogenesis [15]. Eight novel missense alterations were found in six additional genes (*CDH1*, *FBWX7*, *ESR1*, *RB1CC1*, *TP53I3*, and *XBP1*). Of the eight missense alterations, two were observed in *CDH1* and two in *ESR1*, and four overall resulted in significant amino acid shifts by Miyata criteria [30]. In *RB1CC1*, a significant amino acid shift is predicted, namely R1514C, with a high Miyata score of 3.06; this results in a positively charged residue being changed to a hydrophilic residue. The mutation M180K in *TP53I3* has a Miyata score of 2.63 and predicts a change from a hydrophobic to a positively charged residue. In *ESR1*, a H6Y with a Miyata score of 2.27 predicts a shift of a positive charge to a hydrophilic charge. In the *FBXW7* gene, the E117K substitution with a Miyata score of 1.14 results in a shift from negative to positive charge. There are several conservative substitutions that have low Miyata scores: in the *CDH1* gene the M282I variant was observed twice and gave a Miyata score of 0.29, and the D777N variant has a Miyata score of 0.65; in *ESR1* the M264I variant has a score of 0.29; and in *XBP1* the variant R232K has a Miyata score of 0.4.

In *FZD7*, we observed two synonymous variants, L23L and L26L, which are both in close proximity to a common SNP in codon 24 that results in a conservative shift from glycine to arginine. Notably, a second SNP that also affects codon 24, namely G24S, was seen in the Norwegian population, which also results in a conservative shift with a Miyata score of 0.85. These data suggest that this could be a region of increased mutational activity, but further work on breast tumors and cell lines is needed to characterize the functional implications of the changes. The distribution of alterations did not differ from that of the SNPs across the same regions for both coding and noncoding regions.

Insertion/deletion somatic alterations were observed in four genes, and there were a total of nine. Six were insertion/deletion alternations within the coding region of the gene and one, a 4 bp insertion, occurred at the splice site junction in *CDH1*. The gene most frequently observed with insertion/deletion (four) was *CDH1*, which has previously been reported to have altered copy number (loss of heterozygosity), and can undergo somatic mutation and silencing by methylation [31-33]. Muta-

Table 1

Somatic alterations by region in breast cancer re-sequence analysis

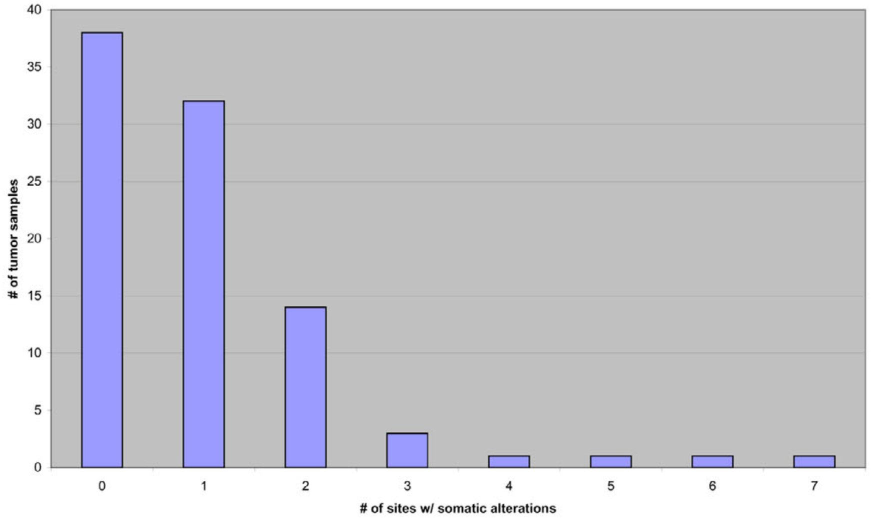
Gene	Nonsynonymous	Synonymous	5'	IVS	3'	Indel
TP53[a]	R110P, F113V, A138V (2), Y163C, Y163H, H193T, I195T, V216M, S241A, R249M, I251S, D259Y, R273C (3)			IVS1+75A>G IVS6-2A>G IVS7-1G>T		7 bp @ G286 22 bp @ L189
GATA3[a]	R366X, R366L			IVS2-281C>G	Ex5+311C>G	CA @ IVS3-3 A @ T315
CAV1		D143D	-2768A>C -1446T>A -596A>G		Ex3-3G>C	
CDH1	D777N M282I (2)	A563A		IVS3+128T>C IVS3+260A>G IVS7+47T>C IVS7+2049T>A IVS8-175G>A		CCGG @ Ex3+19 A @ Ex345 G @ Ex7+15 AAGT @ IVS13+3
ESR1	H6Y M264I		Ex1+139G>C		Ex8+2144T>A	
FBXW7	E117K			IVS1-1641A>C		
FOXA1			-3717G>A			
FOXC1			-2713G>T -1770C>T		+936G>A +940G>A	
FZD7		L23L L26L G409G	-1387G>C			
MYBL2				IVS7+15A>G IVS8-14C>G IVS12+28G>C		
PIN1		S38S		IVS2+3447G>A IVS2+3419G>A IVS2+3370T>C		
RB1CC1	R1514C	S1424S L1511L	-31467C>A -30224C>G	IVS1+5248T>G IVS1+6068A>G IVS11+97T>C IVS11-36C>T IVS15+1535C>G IVS21-16G>A		
RERG			Ex1+63G>T	IVS2+27438T>G IVS2-30415C>T		CTTdel @ IVS2-7163
STK6[b]		A172A E175E		IVS4-35A>G IVS9-33A>G		
TP53I3	M180K	P102P				
XBP1	R232K		-2339G>C	IVS4-11G>A		

Eighty-seven unique somatic alterations were identified by sequence analysis in 16 of 21 genes analyzed (*TP53*, *GATA3*, *CAV1*, *CDH1*, *ESR1*, *FBXW7*, *FOXA1*, *FOXC1*, *FZD7*, *MYBL2*, *PIN1*, *RB1CC1*, *RERG*, *STK6*, *TP53I3*, and *XBP1*). No somatic alterations were detected in five genes (*CDK5*, *NQO1*, *PIM1*, *PLK1*, and *RARRES3*). All alterations were observed singly in bidirectional sequence analysis; numbers in parentheses indicates the number of unique tumor samples with somatic alterations. Intronic analysis restricted to regions including 100 bp on either side of exonic junctions and evolutionarily conserved regions between mouse and human (>75% similarity over 200 bp). [a]Most of the nonsynonymous mutations and deletions in *TP53* and *GATA3* were previously reported [4,19]. [b]In the analysis of *STK6* (also known as *STK15*), additional alterations were observed at six sites within the target 5' region but lie in the adjacent gene *CSTF1* (-7698C>T, -7648A>G, -7105C>T, -5992A>G, -4868A>G, and -4221G>A). None of the variants result in nonsynonymous alterations in the coding region of *CSTF*.

tions in *CDH1*, particularly frameshift mutations, are seen more frequently in the lobular histologic subtype [34], and in our series two out of three with frameshift somatic alterations were observed in tumors classified as lobular.

Of the 21 genes included in the re-sequencing analysis, 10 of the 21 were contained within the 'SAM264' set of genes, which represents genes that were associated with breast cancer patient survival times [17]. In order to visualize the expression patterns of all the re-sequenced genes, we added the 10 missing genes (*CDH1*, *CAV1*, *FBXW7*, *PIM1*, *PIN1*, *TP53*, *TP53I3*, *RB1CC1*, *FZD7*, and *CDK5*) to the SAM264 gene set and performed a hierarchical clustering analysis using a dataset of 194 tumors, which was the combined data on the tumors used to select genes for re-sequencing analysis [18,19] and 72 of the tumors that were actually re-sequenced (Figure 1). After standard data quality gene filtering methods were employed, 19 out of 21 genes were present (*FBXW7* and *PIN1* gave very low signal intensities) in the clustering analysis and clearly exhibited significant variation in expression

Figure 2



Distribution of samples with observed number of somatic alterations. Total number of somatic alterations per sample are included underneath the bars and the total number of samples in each category is represented on the vertical axis.

across the 194 breast tumors (Figure 1). A clustering analysis using the modified SAM264 list and just the 63 Norwegian samples that were included in the re-sequencing was also performed (Additional file 1), and in this analysis the differential expression of the genes over this set of tumor samples recapitulated our previous findings and showed that many of the re-sequenced genes have expression patterns that define the breast tumor subtypes [18,19].

Discussion

We report the results of a re-sequence analysis of 21 candidate genes in 91 primary breast tumors. The candidate genes were chosen mainly based upon previous expression profiling studies and the target sequencing regions were extended beyond coding regions to include evolutionarily conserved regions and the 5' and 3' regions. This latter point is essential because it underscores the importance of examining genetic regions that could alter the expression or stability of a gene. Our results identified a spectrum of single base sequence alterations in 16 of the 21 genes selected for targeted re-sequence analysis.

Unlike the report by Stephens and coworkers [15], we did not observe clustering of sequence alterations in a single tumor. The maximum number of alterations in any of the 21 genes observed in a single tumor was seven, and we observed no somatic alteration in nearly 40% of the samples analyzed. We can exclude the likelihood that loss of heterozygosity could account for this, because the density of common SNPs observed across the 21 genes was comparable to the density observed in the SNP500 Cancer set and the International HapMap study [35]. We observed a comparable ratio for transitions to transversions to that reported by Stephens and colleagues [15].

In our analysis we observed 27 total somatic alterations mutations in eight of the 21 genes studied, and of these 23 were unique missense alterations. In contrast, there were 11 synonymous alterations. The difference in the observed number of nonsynonymous changes relative to synonymous SNPs did not deviate significantly from expected [36]. Unlike the study conducted by Stephens and coworkers [15], we did not observe enrichment of nonsynonymous changes relative to synonymous changes in our set of 21 genes chosen on the basis of expression profiles.

Wide variance in the number of somatic alterations per gene was observed, which did not always correlate with previous reports. For instance, in the *RB1CC1* gene, which was previously reported to undergo truncating mutations in breast tumors [37], our bidirectional sequence analysis revealed eight noncoding alterations, two synonymous alterations, and a single nonsynonymous change, namely R1514C (Miyata score of 3.06), which results in a positively charged residue shift to a hydrophilic residue. In *PIN1*, a gene previously reported to be mutated in breast cancer [38], we observed only one synonymous nucleotide change. Interestingly, two alterations were observed in *FBXW7*, namely a nonsynonymous E117K with a significant amino acid shift and an intronic alteration; this is in contrast to previous reports [39,40], which found a higher rate of mutation in *FBXW7* in breast cancer cell lines.

Our approach differed from the prior reported studies in that the re-sequence analysis also targeted regions of possible regulatory importance. In fact, our study targeted sequence in contiguous noncoding regions, which could be enriched for regulatory regions to be defined functionally. Because there is no *a priori* way to predict whether any one of the identified synonymous and noncoding somatic alterations disrupts function, analysis unique to each gene will be required to establish whether the change is functional. At this time, we interpret the majority of these changes to be nonfunctional and perhaps related to an increase in background mutation rate in cancer, specifically in breast cancer. In this regard, we confirmed the findings of Stephens and coworkers [15] that the majority of observed somatic variants are probably passenger or hitchhiking mutations and not necessarily subject to selection. Further laboratory work is needed to determine which sequence alterations bear functional consequences for the development of breast cancers.

It is notable that we observed somatic alterations in genes not observed in the study conducted by Sjoblom and coworkers [16] (for instance, *ESR1*, *GATA3*, and *CDH1*). This is not surprising because our study included a large number of estrogen receptor positive and estrogen receptor negative samples in our SNP discovery phase. Moreover, *GATA3* and *ESR1* mutations appear to be mutated primarily in estrogen receptor positive tumors, which were not included in the discovery cell line set used in the study conducted by Sjoblom and colleagues [29].

The prevalence of somatic mutations probably varies between different cancers and possibly by populations [2,3,14,15]. Previous surveys of the coding regions in colon and breast cancer were biased toward genes of the protein kinase family and reported a rate of approximately one somatic alteration per megabase of sequence. We estimate that the rate of somatic mutation in genes altered in expression pattern in breast cancer is slightly higher than that reported for colon cancer and breast cancer [2,3,12]. Based on this survey of coding and noncoding sequence (at a ratio of 1:3) for 21 genes, we estimate the rate of somatic alteration could be as high as 5.3 per megabase, but this assumes that all variants are indeed somatic variants. It is possible that a subset could be rare germline variants. Thus, our estimate for the somatic mutation rate in breast cancer is slightly higher than the previous reports of approximately 1.2 nonsynonymous somatic changes per megabase in colon cancer [2,3,15]. We also note that two-thirds of the sequence analyzed in the present study is noncoding. An estimate of the rate of somatic alterations did not differ between noncoding and coding regions, suggesting that the majority could be hitchhiking mutations. In an analysis of 11 breast cancer cell lines and colon tumor xenografts, Sjoblom and coworkers [16] also observed more somatic alterations, approximately 2.5 more, than in the earlier studies [2,3,15]. Together with our results, these data suggest that somatic alterations could arise more frequently than was originally reported. It is also notable that our study also targeted noncoding regions, where somatic alterations rates might differ from those in coding regions. Further studies are required to address this important point.

It is plausible that our study might also underestimate the rate of somatic alteration because we previously identified five additional mutations in the *TP53* gene [6] (Langerod and coworkers, unpublished data) in the Norwegian tumor samples using a highly sensitive screening technique, namely temporal temperature gel electrophoresis (TTGE), prior to sequencing (these mutations are marked in Additional file 2). This prescreening allowed us to detect mutations in a heterogeneous tumor sample with as low as 1% mutated cells [41], and aberrant migrating band on the TTGE gel can be sequenced directly to define the sequence alteration. Microdissection before sequencing will not fully avert this problem because of tumor heterogeneity. To use TTGE as pre-screening is impractical because it is labor intensive to establish and not easily amenable to high throughput analyses. Because we had previously performed TTGE for *TP53* analyses, we were able to assess the sensitivity of the different techniques; both techniques failed to identify all mutations.

Conclusion

Systematic re-sequence analysis of a sufficiently large set of tumor samples drawn from well designed clinical and epidemiologic studies promises to identify new cancer associated genes and somatic mutations that are linked to response to cancer therapy [42-44]. The present study confirms the value of re-sequence analysis in cancer gene discovery and underscores the importance of characterizing somatic alterations across genes related not only by function, or functional pathways, but also based upon expression patterns. Advances in sequencing technology will certainly accelerate the characterization of somatic alterations in the cancer genome, but the task of defining the importance of observed somatic changes will continue to rest on the shoulders of future laboratory investigators.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SJC conceived the project, analyzed data, and wrote the paper. LB analyzed sequence tracings and managed the dataset. MY analyzed data. VL analyzed sequence tracings and managed the dataset. AL handled samples and analyzed data. SP performed experiments. RK collected clinical samples and analyzed data. RLS analyzed data and revised the paper. DSG analyzed data and revised the paper. VK collected samples, analyzed the data, and revised the paper. CMP conceived project, analyzed the data, and revised the paper. ALBD conceived the project, collected samples, analyzed the data, and revised the paper.

Additional files

The following Additional files are available online:

Additional file 1

A pdf file showing a hierarchical clustering analysis based on the 63 breast tumor samples from Norway that were used for the re-sequencing analysis, which was clustered using the augmented 'SAM264' patient survival associated gene set. (a) Hierarchical clustering overview that shows the overall context for the 19 genes. (b) Close up of the sample associated dendrogram. (c) Basal epithelial gene set showing the expression of FZD7. (d) Proliferation gene set showing expression of STK6, MYBL2, and PLK1. (e) Luminal/ER$^+$ epithelial gene set showing coordinated expression of ESR1, GATA3, FOXA1, and XBP1. (f) The expression profiles of the additional genes that were re-sequenced but that did not fall into the previous three expression patterns are shown, and their position in the larger cluster is also shown in panel A. All genes identified by red text were analyzed by re-sequencing in this study, and only FBXW7 and PIN1 were not included in this cluster analysis because their average expression levels did not meet the gene filtering criteria.
See http://www.biomedcentral.com/content/supplementary/bcr1637-S1.pdf

Additional file 2

A doc file in which observed somatic alterations are reported by individual breast cancer tissue sample (n = 53). Somatic alterations previously reported in TP53 and GATA3 are highlighted [4-6,29] (Langerod and coworkers, unpublished data).
See http://www.biomedcentral.com/content/supplementary/bcr1637-S2.doc

Acknowledgements

We acknowledge Drs Phil Bernard and Juan Palazzo for samples that were used in this study. CMP was supported by funds from the NCI Breast SPORE program to UNC-CH (P50-CA58223-09A1) and by RO1-CA-101227-01. This work was also supported by grants D-99061 from The Norwegian Cancer Society, 155218/300 from the Research Council of Norway (NFR) and the SalusAnsvar Medical award to ALBD. VNK was visiting scientist at NCI, NIH supported by grant 152004/150 Functional Genomics, FUGE, NFR. This research was supported by the Intramural Research Program of the NIH, National Cancer Institute, Center for Cancer Research, Division of Cancer Epidemiology and Genetics and Office of Cancer Genomics.

References

1. Balmain A, Gray J, Ponder B: **The genetics and genomics of cancer.** *Nat Genet* 2003, **33:**238-244.
2. Wang TL, Rago C, Silliman N, Ptak J, Markowitz S, Willson JK, Parmigiani G, Kinzler KW, Vogelstein B, Velculescu VE: **Prevalence of somatic alterations in the colorectal cancer cell genome.** *Proc Natl Acad Sci USA* 2002, **99:**3076-3080.
3. Weir B, Zhao X, Meyerson M: **Somatic alterations in the human cancer genome.** *Cancer Cell* 2004, **6:**433-438.
4. Langerod A, Bukholm IR, Bregard A, Lonning PE, Andersen TI, Rognum TO, Meling GI, Lothe RA, Borresen-Dale AL: **The TP53 codon 72 polymorphism may affect the function of TP53 mutations in breast carcinomas but not in colorectal carcinomas.** *Cancer Epidemiol Biomarkers Prev* 2002, **11:**1684-1688.
5. Børresen-Dale A-L: **TP53 and breast cancer.** *Hum Mutat* 2003, **21:**292-300.
6. Olivier M, Langerød A, Carrieri P, Bergh J, Klaar S, Eyfjord J, Theillet C, Rodriguez C, Lidereau R, Bieche I, *et al.*: **The clinical value of somatic TP53 gene mutations in 1,794 patients with breast cancer.** *Clin Cancer Res* 2006, **12:**1157-1167.
7. Samuels Y, Wang Z, Bardelli A, Silliman N, Ptak J, Szabo S, Yan H, Gazdar A, Powell SM, Riggins GJ, *et al.*: **High frequency of mutations of the PIK3CA gene in human cancers.** *Science* 2004, **304:**554.
8. Wu G, Xing M, Mambo E, Huang X, Liu J, Guo Z, Chatterjee A, Goldenberg D, Gollin SM, Sukumar S, *et al.*: **Somatic mutation and gain of copy number of *PIK3CA* in human breast cancer.** *Breast Cancer Res* 2005, **7:**R609-R616.
9. Kurose K, Gilley K, Matsumoto S, Watson PH, Zhou XP, Eng C: **Frequent somatic mutations in *PTEN* and *TP53* are mutually exclusive in the stroma of breast carcinomas.** *Nat Genet* 2002, **32:**355-357.
10. Bachman KE, Argani P, Samuels Y, Silliman N, Ptak J, Szabo S, Konishi H, Karakas B, Blair BG, Lin C, *et al.*: **The *PIK3CA* gene is mutated with high frequency in human breast cancers.** *Cancer Biol Ther* 2004, **3:**772-775.
11. Campbell IG, Russell SE, Choong DY, Montgomery KG, Ciavarella ML, Hooi CS, Cristiano BE, Pearson RB, Phillips WA: **Mutation of the PIK3CA gene in ovarian and breast cancer.** *Cancer Res* 2004, **64:**7678-7681.
12. Wang Z, Shen D, Parsons DW, Bardelli A, Sager J, Szabo S, Ptak J, Silliman N, Peters BA, van der Heijden MS, *et al.*: **Mutational analysis of the tyrosine phosphatome in colorectal cancers.** *Science* 2004, **304:**1164-1166.
13. Davies H, Bignell GR, Cox C, Stephens P, Edkins S, Clegg S, Teague J, Woffendin H, Garnett MJ, Bottomley W, *et al.*: **Mutations of the *BRAF* gene in human cancer.** *Nature* 2002, **417:**949-954.
14. Parsons DW, Wang TL, Samuels Y, Bardelli A, Cummins JM, DeLong L, Silliman N, Ptak J, Szabo S, Willson JK, *et al.*: **Colorectal cancer: mutations in a signalling pathway.** *Nature* 2005, **436:**792.
15. Stephens P, Edkins S, Davies H, Greenman C, Cox C, Hunter C, Bignell G, Teague J, Smith R, Stevens C, *et al.*: **A screen of the complete protein kinase gene family identifies diverse patterns of somatic mutations in human breast cancer.** *Nat Genet* 2005, **37:**590-592.
16. Sjoblom T, Jones S, Wood LD, Parsons DW, Lin J, Barber T, Mandelker D, Leary RJ, Ptak J, Silliman N, *et al.*: **The consensus coding sequences of human breast and colorectal cancers.** *Science* 2006, **314:**268-274.
17. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, *et al.*: **Molecular portraits of human breast tumours.** *Nature* 2000, **406:**747-752.
18. Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van den Rijn M, Jeffrey SS, *et al.*: **Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications.** *Proc Natl Acad Sci USA* 2001, **98:**10869-10874.
19. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S, *et al.*: **Repeated observation of breast tumor subtypes in independent gene expression data sets.** *Proc Natl Acad Sci USA* 2003, **100:**8418-8423.
20. Pollack JR, Sorlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, Tibshirani R, Botstein D, Borresen-Dale AL, Brown PO: **Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors.** *Proc Natl Acad Sci USA* 2002, **99:**12963-12968.
21. Rozen S, Skaletsky HJ: *Bioinformatics Methods and Protocols: Methods in Molecular Biology* Totowa, NJ: Humana Press; 2000:365-386.

22. Packer BR, Yeager M, Burdett L, Welch R, Beerman M, Qi L, Sicotte H, Staats B, Acharya M, Crenshaw A, *et al.*: **SNP500Cancer: a public resource for sequence validation, assay development, and frequency analysis for genetic variation in candidate genes.** *Nucleic Acids Res* 2006, **34:**D617-D621.
23. Mayor C, Brudno M, Schwartz JR, Poliakov A, Rubin EM, Frazer KA, Pachter LS, Dubchak I: **VISTA: visualizing global DNA sequence alignments of arbitrary length.** *Bioinformatics* 2000, **16:**1046-1047.
24. Helle SI, Ekse D, Holly JMP, Lonning PE: **The IGF-system in healthy pre- and postmenopausal women: relations to demographic variables and sex-steroids.** *J Steroid Biochem Mol Biol* 2002, **81:**95-102.
25. Benito M, Parker J, Du Q, Wu J, Xiang D, Perou CM, Marron JS: **Adjustment of systematic microarray data biases.** *Bioinformatics* 2004, **20:**105-114.
26. Diehn M, Sherlock G, Binkley G, Jin H, Matese JC, Hernandez-Boussard T, Rees CA, Cherry JM, Botstein D, Brown PO, *et al.*: **SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data.** *Nucleic Acids Res* 2003, **31:**219-223.
27. Yang YH, Dudoit S, Luu P, Lin DM, Peng V, Ngai J, Speed TP: **Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation.** *Nucleic Acids Res* 2002, **30:**e15.
28. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB: **Missing value estimation methods for DNA microarrays.** *Bioinformatics* 2001, **17:**520-525.
29. Usary J, Llaca V, Karaca G, Presswala S, Karaca M, He X, Langerod A, Karesen R, Oh DS, Dressler LG, *et al.*: **Mutation of GATA3 in human breast tumors.** *Oncogene* 2004, **23:**7669-7678.
30. Miyata T, Miyazawa S, Yasunaga T: **Two types of amino acid substitutions in protein evolution.** *J Mol Evol* 1979, **12:**219-236.
31. Rieger-Christ KM, Pezza JA, Dugan JM, Braasch JW, Hughes KS, Summerhayes IC: **Disparate E-cadherin mutations in LCIS and associated invasive breast carcinomas.** *Mol Pathol* 2001, **54:**91-97.
32. Berx G, Becker KF, Hofler H, van Roy F: **Mutations of the human E-cadherin (CDH1) gene.** *Hum Mutat* 1998, **12:**226-237.
33. Graff JR, Herman JG, Lapidus RG, Chopra H, Xu R, Jarrard DF, Isaacs WB, Pitha PM, Davidson NE, Baylin SB: **E-cadherin expression is silenced by DNA hypermethylation in human breast and prostate carcinomas.** *Cancer Res* 1995, **55:**5195-5199.
34. Berx G, Cleton-Jansen AM, Strumane K, de Leeuw WJ, Nollet F, van Roy F, Cornelisse C: **E-cadherin is inactivated in a majority of invasive human lobular breast cancers by truncation mutations throughout its extracellular domain.** *Oncogene* 1996, **13:**1919-1925.
35. The International HapMap Consortium: **The International HapMap Project.** *Nature* 2005, **437:**1299-1320.
36. Gerhard DS, Wagner L, Feingold EA, Shenmen CM, Grouse LH, Schuler G, Klein SL, Old S, Rasooly R, Good P, *et al.*: **The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC).** *Genome Res* 2004, **14:**2121-2127.
37. Chano T, Kontani K, Teramoto K, Okabe H, Ikegawa S: **Truncating mutations of RB1CC1 in human breast cancer.** *Nat Genet* 2002, **31:**285-288.
38. Wulf G, Ryo A, Liou YC, Lu KP: **The prolyl isomerase Pin1 in breast development and cancer.** *Breast Cancer Res* 2003, **5:**76-82.
39. Ekholm-Reed S, Spruck CH, Sangfelt O, van Drogen F, Mueller-Holzner E, Widschwendter M, Zetterberg A, Reed SI: **Mutation of hCDC4 leads to cell cycle deregulation of cyclin E in cancer.** *Cancer Res* 2004, **64:**795-800.
40. Strohmaier H, Spruck CH, Kaiser P, Won KA, Sangfelt O, Reed SI: **Human F-box protein hCdc4 targets cyclin E for proteolysis and is mutated in a breast cancer cell line.** *Nature* 2001, **413:**316-322.
41. Sørlie T, Vu P, Johnsen H, Lind GE, Lothe RA, Børresen-Dale A-L: **Mutation screening of the TP53 gene by temporal temperature gel electrophoresis (TTGE).** In *Methods in Molecular Biology. Molecular Toxicology Protocols Volume 291*. Edited by: Keohavong P, Grant SG. Totowa, NJ: Humana Press Inc; 2004:207-216.
42. Geisler S, Borresen-Dale AL, Johnsen H, Aas T, Geisler J, Akslen LA, Anker G, Lonning PE: ***TP53* gene mutations predict the response to neoadjuvant treatment with 5-fluorouracil and mitomycin in locally advanced breast cancer.** *Clin Cancer Res* 2003, **9:**5582-5588.
43. Lynch TJ, Bell DW, Sordella R, Gurubhagavatula S, Okimoto RA, Brannigan BW, Harris PL, Haserlat SM, Supko JG, Haluska FG, *et al.*: **Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to gefitinib.** *N Engl J Med* 2004, **350:**2129-2139.
44. Paez JG, Janne PA, Lee JC, Tracy S, Greulich H, Gabriel H, Herman P, Kaye FJ, Lindeman N, Boggon TJ, *et al.*: **EGFR mutations in lung cancer: correlation with clinical response to gefitinib therapy.** *Science* 2004, **304:**1497-1500.