

# Identifying In-Trans Process Associated Genes in Breast Cancer by Integrated Analysis of Copy Number and Expression Data

Miriam Ragle Aure<sup>1,2,9</sup>, Israel Steinfeld<sup>3,9</sup>, Lars Oliver Baumbusch<sup>1,9</sup>, Knut Liestøl<sup>4,5</sup>, Doron Lipson<sup>3</sup>, Sandra Nyberg<sup>1,2</sup>, Bjørn Naume<sup>6</sup>, Kristine Kleivi Sahlberg<sup>1,2</sup>, Vessela N. Kristensen<sup>1,2,7</sup>, Anne-Lise Børresen-Dale<sup>1,2</sup>, Ole Christian Lingjærde<sup>2,4,5\*</sup>, Zohar Yakhini<sup>3,8\*</sup>

**1** Department of Genetics, Institute for Cancer Research, Oslo University Hospital Radiumhospitalet, Oslo, Norway, **2** K. G. Jebsen Centre for Breast Cancer Research, Institute for Clinical Medicine, University of Oslo, Oslo, Norway, **3** Laboratory of Computational Biology, Computer Science Department, Israel Institute of Technology, Haifa, Israel, **4** Biomedical Informatics Lab, Department of Computer Science, University of Oslo, Oslo, Norway, **5** Centre for Cancer Biomedicine, University of Oslo, Oslo, Norway, **6** Division of Cancer Medicine and Radiotherapy, Department of Oncology, Oslo University Hospital Radiumhospitalet, Oslo, Norway, **7** Institute for Clinical Epidemiology and Molecular Biology (EpiGen) Akershus University Hospital, Akershus, Norway, **8** Agilent Laboratories, Tel Aviv, Israel

## Abstract

Genomic copy number alterations are common in cancer. Finding the genes causally implicated in oncogenesis is challenging because the gain or loss of a chromosomal region may affect a few key driver genes and many passengers. Integrative analyses have opened new vistas for addressing this issue. One approach is to identify genes with frequent copy number alterations and corresponding changes in expression. Several methods also analyse effects of transcriptional changes on known pathways. Here, we propose a method that analyses *in-cis* correlated genes for evidence of *in-trans* association to biological processes, with no bias towards processes of a particular type or function. The method aims to identify *cis*-regulated genes for which the expression correlation to other genes provides further evidence of a network-perturbing role in cancer. The proposed unsupervised approach involves a sequence of statistical tests to systematically narrow down the list of relevant genes, based on integrative analysis of copy number and gene expression data. A novel adjustment method handles confounding effects of co-occurring copy number aberrations, potentially a large source of false positives in such studies. Applying the method to whole-genome copy number and expression data from 100 primary breast carcinomas, 6373 genes were identified as commonly aberrant, 578 were highly *in-cis* correlated, and 56 were in addition associated *in-trans* to biological processes. Among these *in-trans process associated and cis-correlated* (iPAC) genes, 28% have previously been reported as breast cancer associated, and 64% as cancer associated. By combining statistical evidence from three separate subanalyses that focus respectively on copy number, gene expression and the combination of the two, the proposed method identifies several known and novel cancer driver candidates. Validation in an independent data set supports the conclusion that the method identifies genes implicated in cancer.

**Citation:** Aure MR, Steinfeld I, Baumbusch LO, Liestøl K, Lipson D, et al. (2013) Identifying In-Trans Process Associated Genes in Breast Cancer by Integrated Analysis of Copy Number and Expression Data. PLoS ONE 8(1): e53014. doi:10.1371/journal.pone.0053014

**Editor:** Paolo Provero, University of Turin, Italy

**Received:** June 11, 2012; **Accepted:** November 22, 2012; **Published:** January 30, 2013

**Copyright:** © 2013 Aure 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:** (1) Norwegian Research Council (Grant no 193387/V50, Understanding breast cancer genomics); (2) Norwegian Cancer Society (Grant no 138296 – PR-2008-0108, Exploring The Systems Biology of Breast Cancer); (3) South-Eastern Norway Regional Health Authority (Project no 2011042, OSBREAC: Towards personalized therapy for breast cancer); (4) K.G. Jebsen Center for Breast Cancer Research. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

**Competing Interests:** One author (Zohar Yakhini) is affiliated with Agilent Laboratories, Tel Aviv, Israel. This does not alter the authors' adherence to all the PLOS ONE policies on sharing data and materials.

\* E-mail: ole@ifi.uio.no (OCL); zohar\_yakhini@agilent.com (ZY)

9 These authors contributed equally to this work.

## Introduction

Genomic copy number alterations resulting from genomic instability are commonly observed in cancer [1,2]. Substantial effort has been invested in identifying aberration events playing a critical role in the disease development. In breast carcinomas, the genomic architectural changes are diverse and involve various events such as loss and gain of whole chromosome arms, inversions, translocations, and more focal gains and losses [3,4]. Several array comparative genomic hybridization (aCGH) studies of breast tumors and breast cancer cell lines point to commonly observed gains and losses on regions of chromosome 8, 13 and 17

– regions known to contain breast cancer associated genes such as *BRCA2*, *ERBB2* and *MYC* [5,6,7,8,9].

Recurring aberrations in tumors may be indications of selection driven by changes in the expression of key genes in the affected regions. Since recurrent segmental gains and losses frequently involve several genes, their relative contribution to increased or decreased cell viability and proliferation cannot be inferred from copy number alone. This problem, often portrayed as distinguishing between ‘drivers’ and ‘passengers’, is a key challenge in the task of linking copy number alterations to genes and processes involved in cancer development and progression. One way to proceed would be to focus on genes for which copy number variation

substantially affects gene expression. Integrated analyses of copy number and gene expression data have revealed that the strength of the *in-cis* correlation between copy number and expression varies extensively between genes [10], and subsets of genes with high correlation have been identified and proposed as candidate driver genes [10,11,12,13,14,15].

It has been suggested that the oncogenic effect of molecular alterations is to cause perturbations at the network level, leading cells to malignant phenotypic states (see, e.g. [16]). Several studies have aimed at identifying pathways and networks perturbed by copy number aberrations, thus establishing associations between genomic profiles and aberrant pathways in cancer [17,18,19], clinical outcome and survival [13,20,21,22,23]. One may ask whether a particular gene through its genomic aberrations has an effect on higher-order phenotypes such as processes, pathways and networks. A natural way to approach this would be to first investigate how other genes are affected by the aberration, and second to study whether any biological processes are overrepresented in the list of affected genes. Following this idea, we propose a workflow for integration of copy number and gene expression data based on the stepwise application of a series of gene selection criteria. The method combines correlation analysis, regression analysis, and gene set enrichment, and to avoid confounding effects, the method adjusts for co-occurring copy number aberrations. A key element of the approach is the direct integration of a statistical enrichment step enabling the assignment of statistical confidence to *in-trans* associations between genes and biological processes. The resulting genes are referred to as *in-trans process associated and cis-correlated* (iPAC) genes.

The purpose of combining *in-cis* and *in-trans* analyses is here to identify genes that are *cis*-regulated and for which the correlation structure in the gene expression data provides further support for a role in the alteration of cell phenotype in cancer. The method was applied to a matched data set of aCGH and mRNA expression from 100 well-characterized human primary breast tumors [24,25,26,27], and subsequent application to a second, indepen-

dent breast cancer cohort showed consistent behavior of the iPAC genes found in the first data set. A small selection of iPAC genes were further studied using siRNA knockdown experiments.

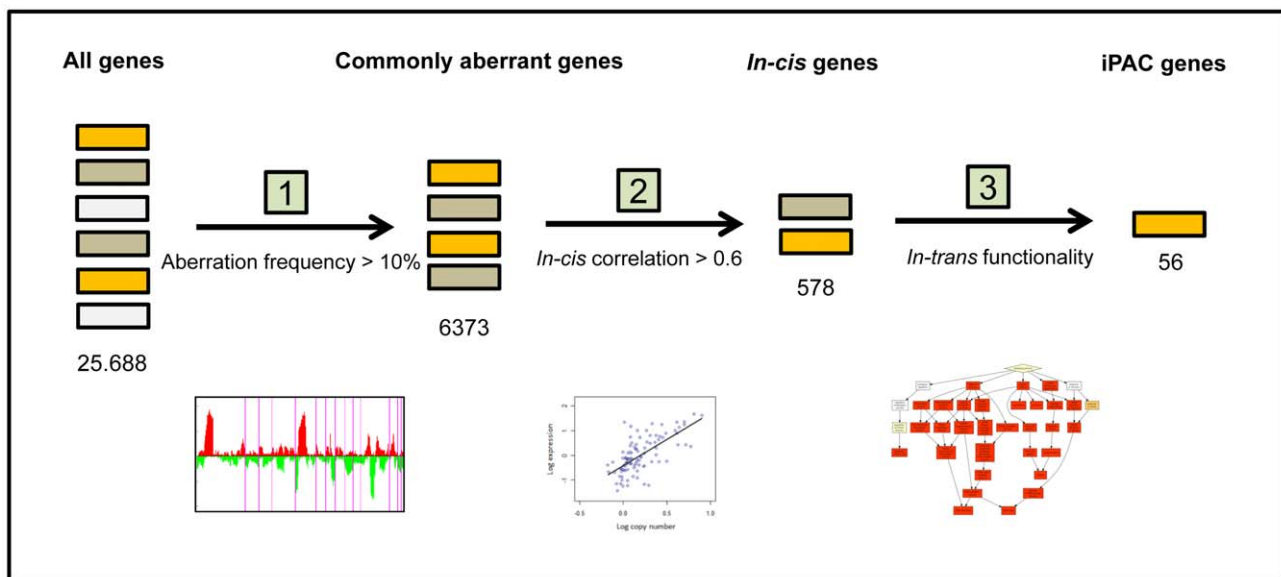
## Materials and Methods

### Ethics statement

The study was approved by the Norwegian regional committee for medical research ethics, Health region II (reference number S-97103), and patients have given written consent for the use of material for research purposes.

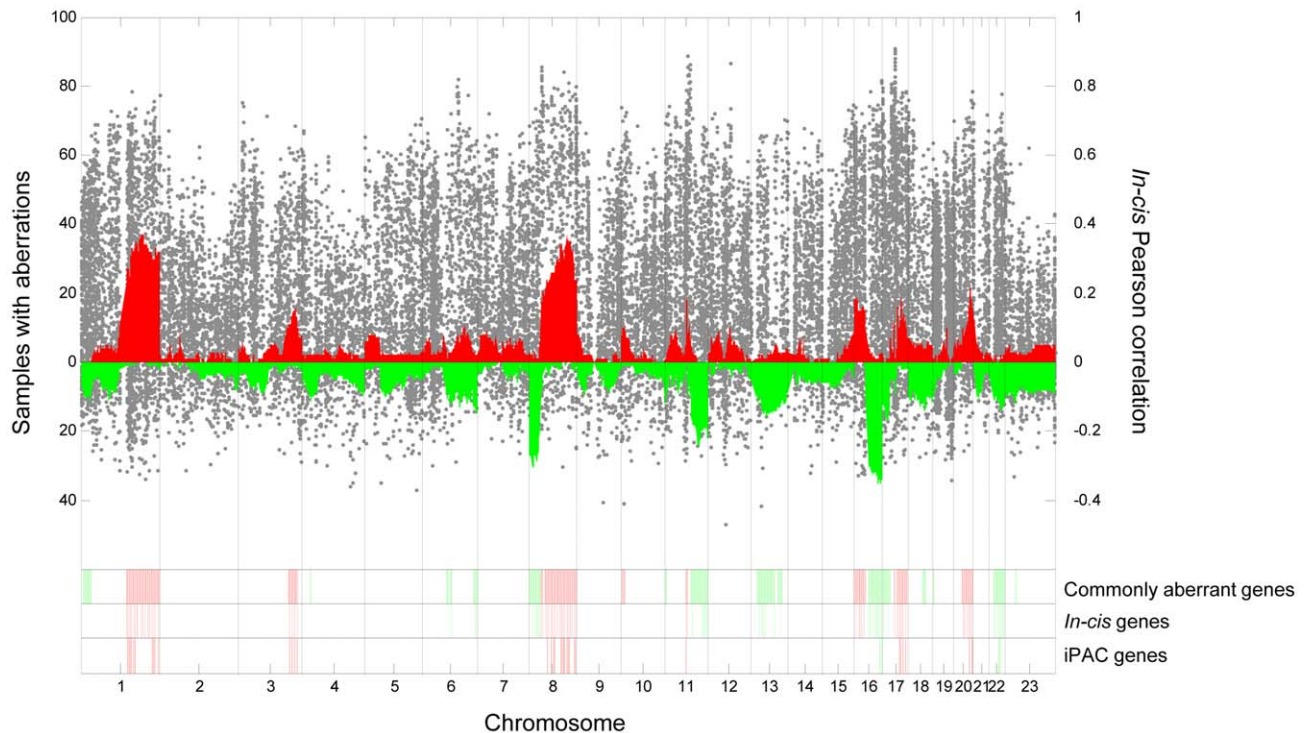
### Patient samples and array experiments

Primary breast carcinoma samples from 100 patients previously described as part of the MicMa cohort were used [24]. All samples were fresh frozen and contained at least 40% tumor cells. The majority of the tumor specimens represent tumor size T1/T2, node status N0/N1 (9/11), and histological grade 2 or 3. Tumor DNA was extracted using an ABI 341 Nucleic Acid Purification System (Applied Biosystems, CA, USA) according to the manufacturer's protocol. Tumor RNA was isolated using TRIzol reagent (Invitrogen, CA, USA) as previously described [28]. The subtype classification deriving from mRNA expression has previously been presented [29]. The aCGH and the mRNA expression data sets have previously been published [26,30]. The expression data (measured using Agilent 4 by 44K one-color oligonucleotide arrays) are available in Gene Expression Omnibus (GEO) with accession number GSE19783 [26], and the copy number data (measured using Illumina Human-1 109K BeadChip SNP arrays) are available on request to OCL. A breast cancer data set from the University of North Carolina, Chapel Hill (UNC), including 73 samples profiled on the same array platforms as described above, was used for validation (see [25,31] for details). Genomic locus annotation for all analyzed data is based on the human genome build hg17.



**Figure 1. Workflow of the proposed method to identify iPAC genes.** (1) Starting with all genes, the commonly aberrant genes are selected as those that have more than 10% gains or losses; (2) Next, those genes which in addition have an *in-cis* Pearson correlation above 0.6 are selected and referred to as *in-cis* genes; (3) Finally, statistical enrichment analysis is performed to assess *in-trans* functionality, leading to identification of the 56 iPAC genes.

doi:10.1371/journal.pone.0053014.g001



**Figure 2. Copy number aberrations and in-cis correlations.** The frequency of samples with gains (red) and losses (green) is shown at the top. Each gray point shows the level of in-cis correlation between copy number and expression for a particular gene. The chromosomal positions of the genes selected in our workflow are shown at the bottom. This includes commonly aberrant genes ( $n=6373$ ; upper band), in-cis genes ( $n=578$ ; middle band), and the iPAC genes ( $n=56$ ; lower band). Colors indicate whether the gene is most frequently amplified (red) or deleted (green). doi:10.1371/journal.pone.0053014.g002

### Overview of analysis

The proposed method is based on the stepwise application of a series of gene selection criteria, and a core element is the use of a rigorous statistical enrichment technique to reveal significant associations in *trans* between the selected genes and biological processes (see Figure 1). This enrichment step is combined with a novel correction method designed to alleviate the problem of co-occurring copy number alterations across the genome.

### Segmentation

Copy number data were  $\log_2$ -transformed, and each sample was segmented by fitting a piecewise constant regression function to the data using the piecewise constant fitting (PCF) algorithm in the R package *copynumber* [3,32,33,34]. A fitted value (“PCF value”) was then obtained for each segment (and was inherited by each probe in the segment) by averaging the log-transformed copy number values for all probes located in that segment. The user controls the trade-off between sensitivity and specificity with a penalty parameter ( $\gamma$ ) and the minimal number of probes per segment ( $k_{\min}$ ). We chose  $\gamma=70$  which is fairly conservative and thus provides robustness against the presence of potential local (spurious) trends which are common in aCGH data due to varying GC-content and other reasons (see [34] for details), and use the default value  $k_{\min}=3$ .

### Matching copy number and expression values

In order to obtain matching copy number and expression data sets, we first identified all expression probes annotated with a gene symbol in the data set. For each such probe, the copy number probe mapping to the nearest location in the genome was identified. Copy number and gene expression data were then

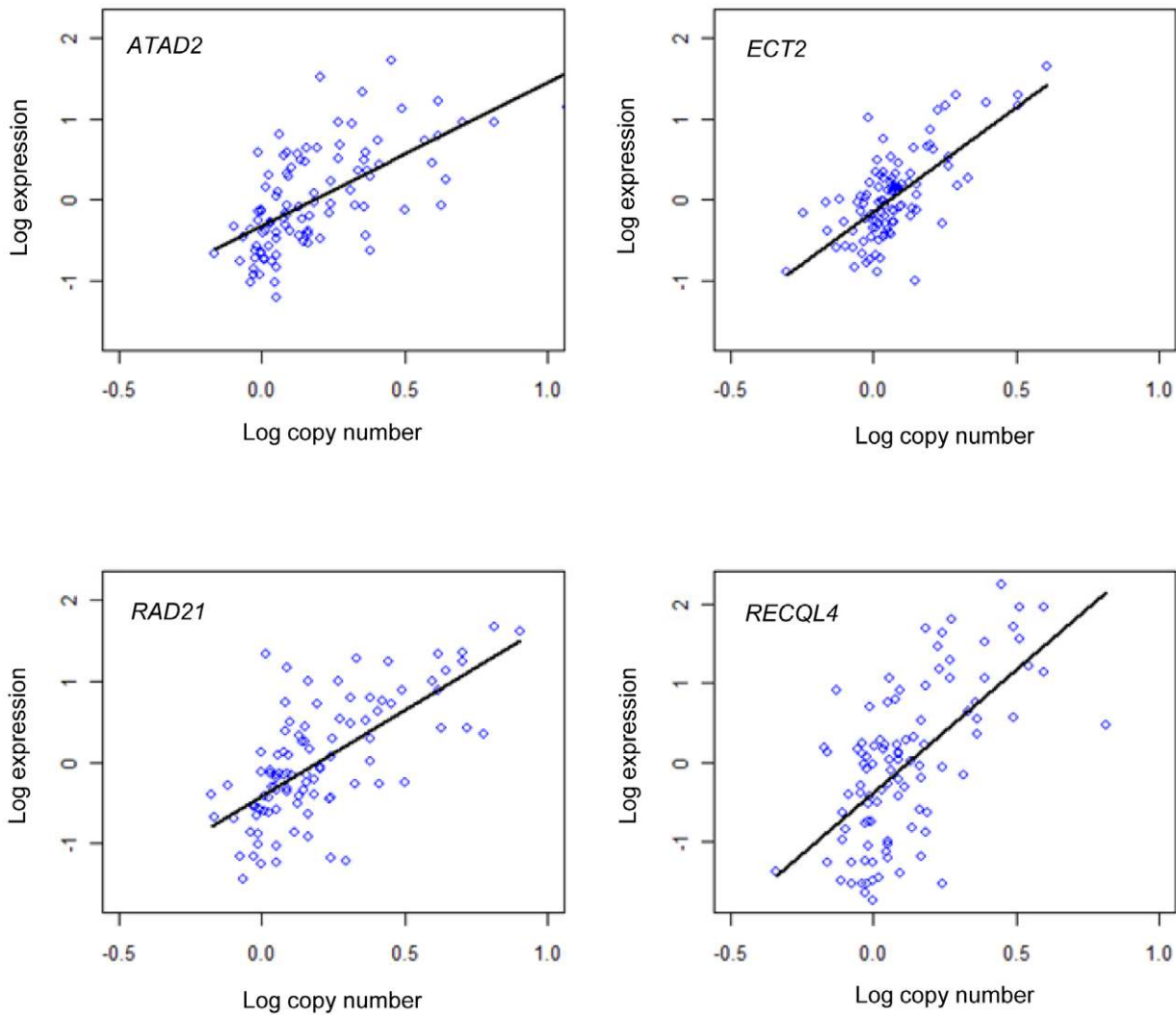
averaged over the corresponding probe values for each gene symbol, resulting in a unique copy number value and expression value for each patient and each gene. The corresponding pair of values was assigned a genomic position by averaging over the locations of the expression probes associated with the gene symbol. Analogous methods are also used in other studies [12,35,36]. This procedure yielded two  $25,688 \times 100$  matrices of copy numbers and corresponding gene expressions, where each row represents a gene and each column a patient sample.

### Aberration calling

To call aberrations, a parameter  $\theta > 0$  determining the sensitivity of the aberration calling (and hence what is considered a significant aberration) was introduced. Probes with a PCF value larger than  $\theta$  were called as gains, and probes with a PCF value less than  $-\theta$  were called as losses. Following the recommended practice for threshold selection in the R package *copynumber* [34], we concluded that  $\theta=0.2$  was an appropriate threshold, which is very similar to the threshold used in [33] where a subset of the copy number data considered in this paper was analyzed.

### Identification of common gains and losses

To identify genomic loci where the copy number events are skewed towards either gain or loss, a sign test was applied. Let  $n$  denote the total number of samples with an aberration in a particular locus, and suppose  $G$  of these aberrations are gains and  $L$  are losses (so that  $n=G+L$ ). Modeling the number of gains as a binomial distribution with  $n$  draws and success probability  $p$ ,  $G \sim \text{Bin}(n, p)$ , we may formally infer whether gains are overrepresented by testing the null hypothesis  $H_0: p=0.5$  against the alternative  $H_1: p > 0.5$ . Using the difference  $D=G-L$  as the test



**Figure 3. Association between expression and copy number.** Linear regression of log-expression as a function of log-copy number for four selected iPAC genes. doi:10.1371/journal.pone.0053014.g003

statistic, we want to determine the rejection region  $D \geq d$  where  $d > 0$  is a given threshold. Since  $D = G - (n - G) = 2G - n$ , we have that

$$\Pr(D \geq d) = \Pr\left(G \geq \frac{n+d}{2}\right) = \sum_{k=r(n)}^n \binom{n}{k} p^k (1-p)^{n-k}$$

where  $r(n) = \lceil (n+d)/2 \rceil$ . Assuming a significance level of  $\alpha = 0.05$ , we seek the least integer  $d > 0$  for which we have  $\Pr(D \geq d) \leq 0.05$  for all  $n$  under the null hypothesis of  $p = 0.5$ . In practice (see Results) the number of aberrations never exceeds 38 in any given locus, and  $n$  may be restrained correspondingly above. A simple calculation then shows that the appropriate threshold is  $D \geq 11$ , and this value was used in the analyses. Thus, all genes for which  $G - L > 10$  were defined as being commonly gained.

By an analogous argument, all genes for which  $L - G > 10$  were defined as being commonly lost. Whenever  $|G - L| > 10$ , the gene was referred to as being commonly aberrant. Note that the purpose of this step was to filter out the bulk of aberrant genes with no indication of skewness towards either gain or loss, and hence

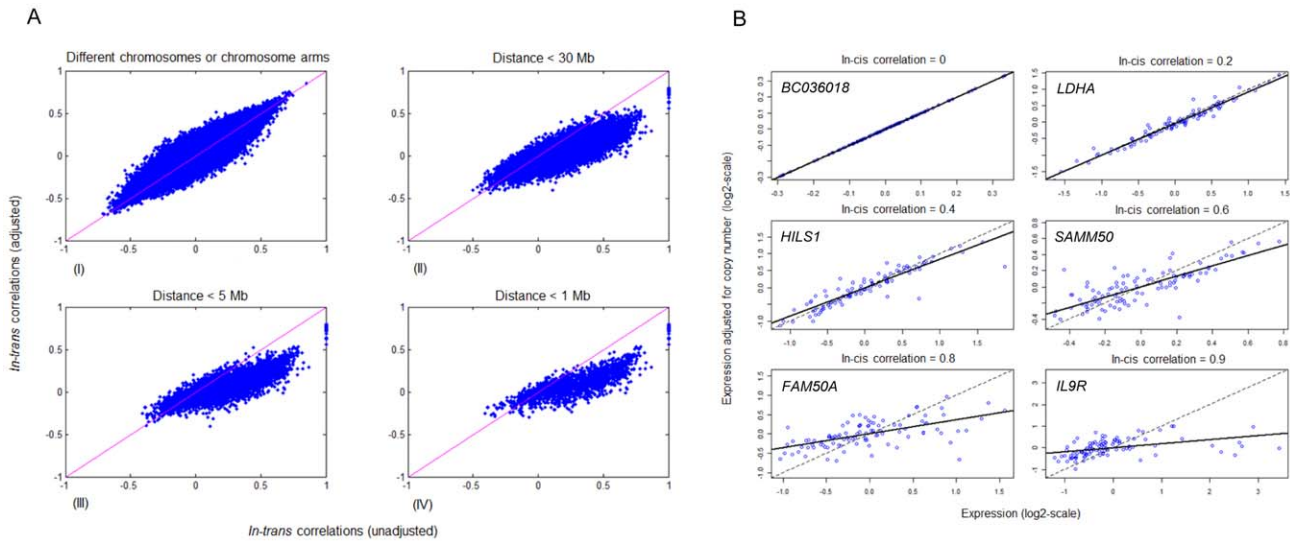
the above significance criterion was designed to be very mild and did not take into account multiple comparisons.

### Identification of in-cis correlated genes

To seek the genes for which the expression is significantly influenced by the copy number, we identified in-cis correlated genes. To identify significant in-cis correlations between log copy number and log gene expression, the in-cis correlations of all the commonly aberrant genes were compared to a background distribution of in-cis correlations. The background distribution was generated by performing 2000 shuffling simulations where in each, only the gene order in the aCGH data set was shuffled and the in-cis correlations were recalculated. By selecting the genes with (Pearson) in-cis correlation  $r > 0.6$  we achieved a false discovery rate (FDR) of less than 2%. This cut-off corresponds to a coefficient of determination of  $(0.6)^2 = 0.36$ , meaning that at least 36% of the variation in log-expression is accounted for by the in-cis variation of log-copy number.

The above procedure corresponds to keeping only the genes  $g$  for which the following log-linear model provides a good fit to the observed copy number and gene expression levels:





**Figure 4. Effect of using copy number-adjusted residual expression.** (A) Comparison of in-trans correlations calculated with and without adjustment for in-cis correlation, i.e. copy number-adjusted-residual expression. In each panel, the x-axis represents the in-trans correlation without adjustment for in-cis correlation, and the y-axis represents the in-trans correlation with adjustment for in-cis correlation. The diagonal lines extend from  $(-1, -1)$  to  $(1, 1)$ . Each point represents one pair of genes among all the  $578 \times 25,688$  gene pairs  $(G, g)$  where  $G$  is an in-cis gene and  $g$  denotes any gene; (I) All pairs for which  $G$  and  $g$  are either on different chromosomes or on the same chromosome but on different arms; (II) All pairs for which  $G$  and  $g$  are within a distance of 30 Mb from each other; (III) All pairs for which  $G$  and  $g$  are within a distance of 5 Mb from each other; (IV) All pairs for which  $G$  and  $g$  are within a distance of 1 Mb from each other. (B) The copy number-adjusted residual expression as a function of the non-adjusted expression, in log space. Shown here are the expression levels for six genes with an in-cis correlation ranging from 0 to 0.9. Each dot represents one breast cancer patient. The effect of copy number-adjusted-residual expression increases with increasing in-cis correlation level. The dotted line is the diagonal, and the solid line is the regression line. doi:10.1371/journal.pone.0053014.g004

$$\log E_{g,i} = a_g + b_g \log C_{g,i} + \varepsilon_{g,i} \quad (1)$$

where  $E_{g,i}$  and  $C_{g,i}$  denote respectively expression and copy number of gene  $g$  in the  $i$ th sample, and  $\varepsilon_{g,i}$  are independent and identically distributed noise terms. Equation (1) implies that log expression is a linear function of log copy number (and noise). Suppressing the gene subscript and ignoring the noise term, equation (1) is equivalent to  $E_i = A \cdot C_i^b$  (where  $A = 2^a$ , assuming logarithms in base 2). Accordingly, model (1) is flexible enough to allow both linear ( $b = 1$ ) and nonlinear ( $b \neq 1$ ) relations between the expression and the copy number of a gene.

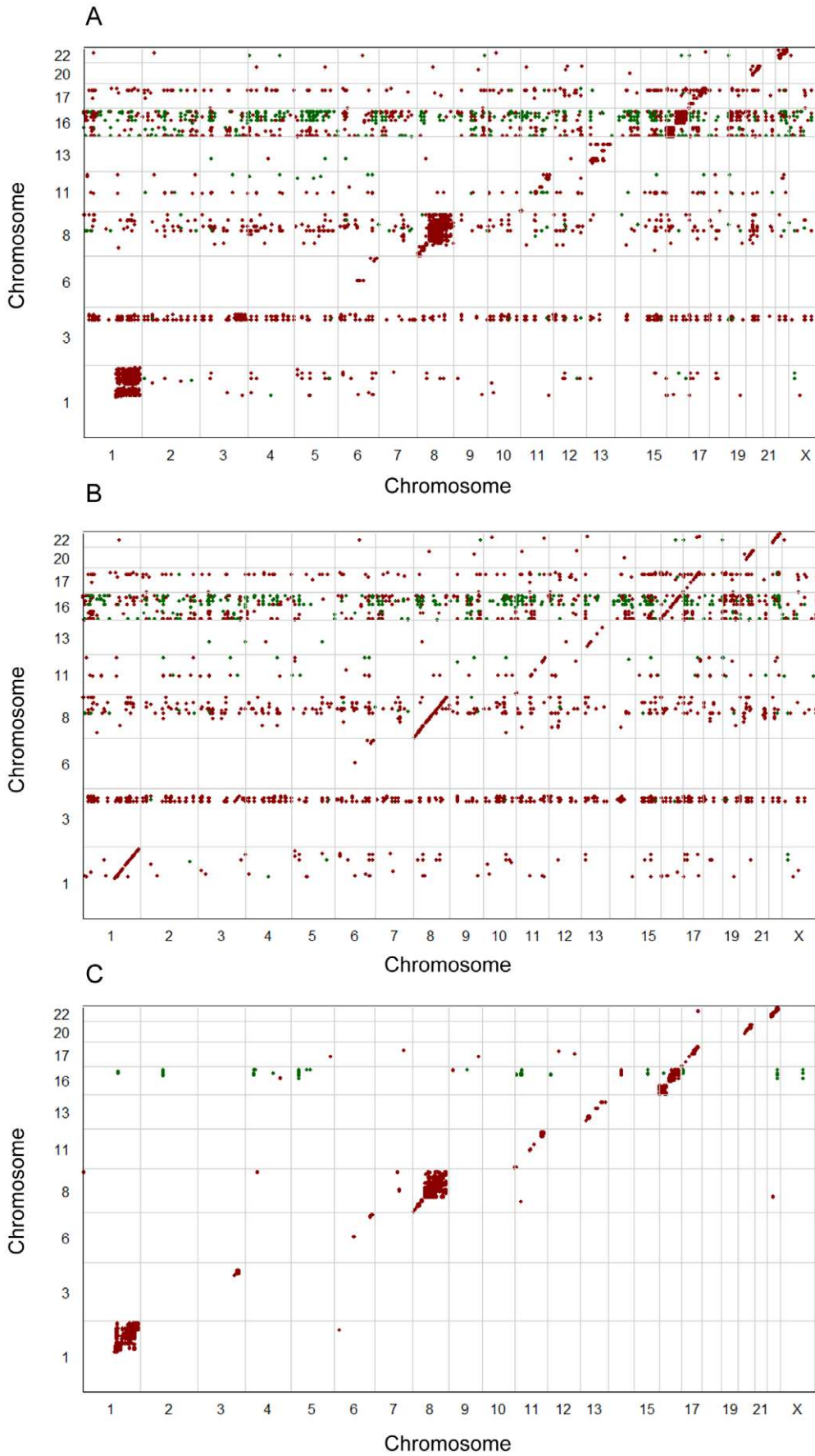
### In-trans correlation analysis

The purpose of this step was to quantify the level of association between the in-cis genes and other genes. To do this, we considered the correlation between the in-cis gene and all other genes. A potential problem in this context is that genes close to each other on the same chromosome may be affected by the same copy number alterations, with inflated correlation between the expression levels of the two genes as a possible result. Thus, co-occurring copy number aberrations can act as a confounding factor, and this should be taken into account when assessing potential expression-mediated effects of one gene on another. To avoid this problem we calculated for each gene  $g$  the residual log expression values  $R_{g,i} = \log E_{g,i} - \hat{a}_g - \hat{b}_g \log C_{g,i}$  over all the samples, where the coefficients  $\hat{a}_g$  and  $\hat{b}_g$  were found by fitting the model in (1), and quantified the in-trans effect of an in-cis gene  $G$  on gene  $g$  by the Pearson correlation between the observed log expressions  $\log E_{G,i}$  of the in-cis gene and the residual log expressions  $R_{g,i}$  of the gene in trans.

### Identification of in-cis genes associated in-trans with processes

In order to identify in-cis genes that were associated with processes in trans, we considered 8284 gene sets defined by Gene Ontology (GO) biological process terms [37]. Using each in-cis gene  $G$  in turn as a pivot, all other genes  $g$  were ranked according to the correlation between  $\log E_{G,i}$  and  $R_{g,i}$  (from high positive correlation to high negative correlation), and an enrichment score was calculated for each GO term in the ranked list of genes. This was done separately for the genes in the top and the bottom of the ranked list. The enrichment score was defined as the p-value from the minimum hypergeometric (mHG) test (see [38,39] for details). Such scores were calculated for each in-cis gene and each GO biological process term. For further analysis, we only considered associations between in-cis genes and GO terms with a p-value score  $p < 0.05$  (after Bonferroni correction).

To obtain empirical p-values for the associations selected above, 100 random simulations were performed. In each simulation instance, we shuffled the order of the samples in the residual expression data set only and recalculated all enrichment scores. This approach preserves existing expression dependencies between genes. Let  $P_{g,s}$  be the enrichment score (mHG p-value) of the association between in-cis gene  $g$  and gene set (GO term)  $s$ , and let  $P_{g,s}^{*k}$  be the enrichment score of this association in the  $k^{th}$  simulation instance ( $k = 1, \dots, 100$ ). We considered an in-cis gene  $g$  to be significantly associated with a gene set  $s$  if  $P_{g,s} < \min_{t,k} (P_{t,s}^{*k})$ , where  $k = 1, \dots, 100$  and  $t$  ranges over all in-cis genes. That is, a relation between an in-cis gene and a gene set was called significant if the observed enrichment score (mHG p-value) was less than the enrichment scores obtained for that gene set for all in-cis genes in all the simulations. This step alleviates differences in attainable p-



**Figure 5. Effect of residual expression.** Correlation plots showing how the level of high-level *in-trans* correlations change across the genome with and without copy number-adjusted residual expression correlation. Red dots signify positive *in-trans* Pearson correlation above 0.6, and green dots signify negative *in-trans* Pearson correlation below  $-0.6$ . The x-axis shows the genomic positions of all 25,688 genes and the y-axis represents the genomic position of the 578 *in-cis* genes. **(A)** High *in-trans* correlations between expression of *in-cis* genes to expression of all genes. **(B)** High *in-trans* correlations between expression of *in-cis* genes to residual expressions of all genes. **(C)** High *in-trans* correlations between copy number of *in-cis* genes to the expression of all genes.  
doi:10.1371/journal.pone.0053014.g005

values due to correlated null hypotheses (the GO gene sets have strong overlaps and genes within a set may be strongly dependent).

### Enrichment analysis using GOrilla

GOrilla (<http://cbl-gorilla.cs.technion.ac.il/>) [38,39] was used with default parameters to investigate and visualize the enrichment of GO biological process in ranked lists of selected gene sets.

### Results

We have presented a computational framework for identification of aberrant genes potentially leading to a substantial shift in transcriptional programs. The proposed method was applied to matched copy number and expression data from a cohort of 100 breast carcinomas. The resulting iPAC genes were further validated in a data set from another breast cancer cohort. The

workflow of our approach is depicted in Figure 1 and details are provided in Materials and Methods.

### Common aberrations

The first step of our workflow was the identification of genes that were commonly aberrant between the patient samples. Among the 25,688 genes profiled, a total of 6373 genes were found to be commonly aberrant, of which 3499 were commonly amplified and 2874 commonly deleted (notice that by the definition of commonly aberrant genes given in Materials and Methods, a gene cannot be both commonly amplified and commonly deleted). These genes are scattered throughout the genome with highest frequency on chromosomes 1, 8, 11, 13, 16, and 17 (Figure 2; Figure S1A). For all genes combined, 7.5% of the variance of the expression values was explained by copy number alterations in *cis*. Considering only commonly aberrant genes, this fraction increased to 11.5%.

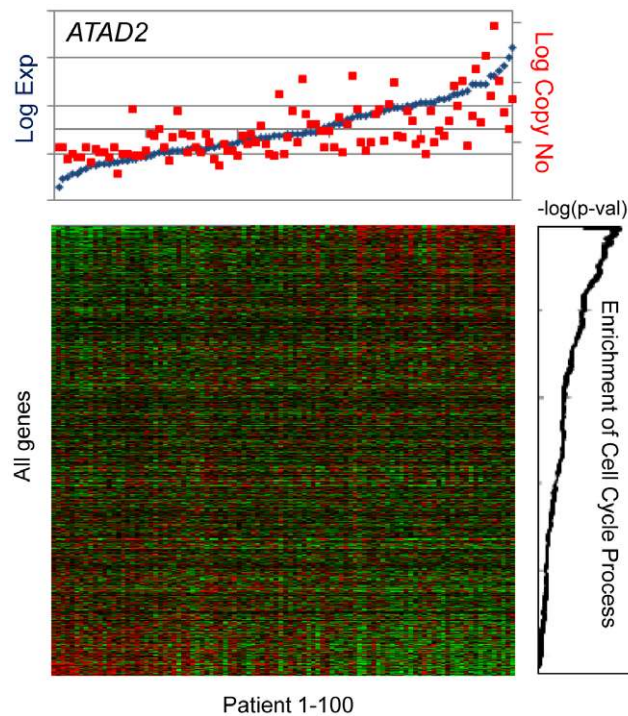
### In-cis associations

The *in-cis* correlation is shown for all genes in Figure 2 and Figure S1B. Ranking the 6373 commonly aberrant genes according to their *in-cis* correlation reveals that the genes with highest correlation are enriched with the GO terms of DNA repair, cell cycle, DNA recombination, and chromatin modification and organization (see Figure S2 and Table S1 for a full list of results). Genes with high *in-cis* correlation (Pearson's  $r > 0.6$ ) were selected among the commonly aberrant genes, resulting in 578 *in-cis* genes (see Figure 2 and 3, and Table S2). These genes were predominantly found on chromosomes 1, 8, 16, and 17. Of these, 423 genes were commonly amplified and 155 commonly deleted (Figure 2).

The *in-cis* genes included known cancer-associated genes such as *ERBB2*, *MAP3K7*, *MDM4*, *FGFR1*, *CCND1* and *FADD*. Further annotation of the 578 genes showed that 19% code for enzymes, 8% regulators of transcription, 7% transporters, 4% kinases, 2% peptidases, and 2% phosphatases (Figure S3A). The remaining genes encode various sorts of proteins, e.g. zinc finger proteins, ribosomal proteins, RNA binding proteins, and mitochondrial proteins (see Table S2 for description). The fraction of the variance in expression explained by copy number alterations increased to 46.6% when considering only the *in-cis* genes. Although the *in-cis* genes exhibit strong correlation between copy number and expression, a substantial proportion of the variability in these genes across samples is also related to other influences. Thus, their expression reflects copy number as well as various other factors.

### In-trans associations to biological processes

The final step of the workflow led to the identification of *in-cis* genes significantly associated with at least one biological process in *trans*. For this purpose, the copy number-adjusted residual expression was calculated for all 25,688 genes. Each *in-cis* gene was taken separately as a pivot and all 25,688 genes were ranked according to the *in-trans* correlation between their copy number-adjusted residual expression and the non-adjusted expression of the pivot gene. The importance of adjusting for copy number is



**Figure 6. Enrichment of the Cell Cycle Process GO term in *ATAD2* correlated genes.** All genes were ranked according to the level of correlation between their copy number-adjusted-residual expression profile and the expression levels of *ATAD2* (pivot for this analysis). The heatmap represents the expression levels of all 25,688 genes after ranking them according to the criteria mentioned above and after sorting the samples according to *ATAD2* expression levels. Top panel in blue and red presents the expression and copy number levels of *ATAD2* across the 100 samples, respectively. The graph shows the significance level in  $-\log(\text{hypergeometric } p\text{-value})$  of cell cycle process genes in the ranked list of genes. Optimal enrichment is attained at the top 189 genes, with 14 times more cell cycle process genes than would be expected by chance (mHG  $p < 4 \times 10^{-83}$ ).  
doi:10.1371/journal.pone.0053014.g006

**Table 1.** Description and properties of the 56 iPAC genes.

Gene	Full gene name	Cytoband	Highest associated GO term (trait)	Score	Annot
DARS2	<i>aspartyl-tRNA synthetase 2, mitochondrial</i>	1q25.1	nucleic acid metabolic proc.	94.31	
ATAD2	<i>ATPase family, AAA domain containing 2</i>	8q24.13	cell cycle	91.53	BC
SMC4	<i>structural maintenance of chromosomes 4</i>	3q25.33	cell cycle	90.42	BC
ACTL6A	<i>actin-like 6A</i>	3q26.33	cell cycle	87.78	C
RECQL4	<i>RecQ protein-like 4</i>	8q24.3	cell cycle	86.78	BC
ECT2	<i>epithelial cell transforming sequence 2 oncogene</i>	3q26.31	cell cycle	82.84	C
POGK	<i>pogo transposable element with KRAB domain</i>	1q24.1	nucleic acid metabolic proc.	82.62	
MTBP	<i>Mdm2, transformed 3T3 cell double minute 2, p53 binding protein</i>	8q24.12	cell cycle	80.67	C
VPS72	<i>vacuolar protein sorting 72 homolog (S. cerevisiae)</i>	1q21.2	nucleic acid metabolic proc.	80.29	
NUDCD1	<i>NudC domain containing 1</i>	8q23.1	nucleic acid metabolic proc.	80.27	C
MTERFD1	<i>MTERF domain containing 1</i>	8q22.1	nucleic acid metabolic proc.	78.77	
WDSOF1	<i>DDB1 and CUL4 associated factor 13</i>	8q22.3	nucleic acid metabolic proc.	78.14	BC
NUP85	<i>nucleoporin 85kDa</i>	17q25.1	cell cycle	75.28	
RAD21	<i>RAD21 homolog (S. pombe)</i>	8q24.11	cell cycle	74.77	BC
KPNA2	<i>karyopherin alpha 2 (RAG cohort 1, importin alpha 1)</i>	17q24.2	cell cycle proc.	74.09	BC
C8orf76	<i>chromosome 8 open reading frame 76</i>	8q24.13	cell cycle proc.	69.94	
POP1	<i>processing of precursor 1, ribonuclease P/MRP subunit</i>	8q22.2	cell division	64.40	C
TATDN1	<i>TatD DNase domain containing 1</i>	8q24.13	cellular macromol. metabolic proc.	61.06	BC
PDCD10	<i>programmed cell death 10</i>	3q26.1	cellular macromol. metabolic proc.	58.81	C
THRAP6	<i>mediator complex subunit 30</i>	8q24.11	cellular nitrogen compound metab. proc.	55.46	BC
RPL30	<i>ribosomal protein L30</i>	8q22.2	cellular macromolecule biosynth. proc.	46.14	C
PRCC	<i>papillary renal cell carcinoma (translocation-associated)</i>	1q23.1	organelle organization	38.66	C
C1orf35	<i>chromosome 1 open reading frame 35</i>	1q42.13	chromosome organization	34.69	C
PARP1	<i>poly (ADP-ribose) polymerase 1</i>	1q42.12	chromosome organization	34.10	BC
MRPS23	<i>mitochondrial ribosomal protein S23</i>	17q23.2	positive regulation of ligase activity	28.47	C
PSMD4	<i>proteasome (prosome, macropain) 26S subunit, non-ATPase, 4</i>	1q21.2	response to DNA damage stimulus	27.44	
SETDB1	<i>SET domain, bifurcated 1</i>	1q21.2	chromatin modification	23.51	C
HNRPU	<i>heterogeneous nuclear ribonucleoprotein U</i>	1q44	chromatin modification	20.52	C
BOP1	<i>block of proliferation 1</i>	8q24.3	DNA conformation change	17.27	C
SIAHBP1	<i>poly-U binding splicing factor 60KDa</i>	8q24.3	mitotic sister chromatid segregation	16.14	C
PRPF3	<i>PRP3 pre-mRNA processing factor 3 homolog (S. cerevisiae)</i>	1q21.2	mRNA transport	15.77	C
PPM1D	<i>protein phosphatase, Mg2+/Mn2+ dependent, 1D</i>	17q23.2	mitotic cell cycle checkpoint	14.29	BC
FAM33A	<i>spindle and kinetochore associated complex subunit 2</i>	17q23.2	mitotic cell cycle checkpoint	14.12	BC
MRPL9	<i>mitochondrial ribosomal protein L9</i>	1q21.3	establishment of organelle localization	13.74	
C22orf28	<i>chromosome 22 open reading frame 28</i>	22q12.3	cellular protein metabolic proc.	13.41	
SLMO2	<i>slowmo homolog 2 (Drosophila)</i>	20q13.32	spindle checkpoint	11.92	
CHRAC1	<i>chromatin accessibility complex 1</i>	8q24.3	mitotic metaphase plate congression	11.91	C
C16orf61	<i>chromosome 16 open reading frame 61</i>	16q23.2	spindle checkpoint	11.28	BC
ISG20L2	<i>interferon stimulated exonuclease gene 20kDa-like 2</i>	1q23.1	DNA-dependent DNA replication init.	11.07	
CSNK1E	<i>casein kinase 1, epsilon</i>	22q13.1	neural tube development	10.25	BC
FAM91A1	<i>family with sequence similarity 91, member A1</i>	8q24.13	establishment of mitotic spindle loc.	10.08	
TOMM20	<i>translocase of outer mitochondrial membrane 20 homolog (yeast)</i>	1q42.3	transcription	10.03	
C20orf20	<i>chromosome 20 open reading frame 20</i>	20q13.33	mitotic cell cycle spindle checkpoint	9.67	C
GALNS	<i>galactosamine (N-acetyl)-6-sulfate sulfatase</i>	16q24.3	carbohydrate catabolic proc.	9.05	BC
AZIN1	<i>antizyme inhibitor 1</i>	8q22.3	histone mRNA metabolic proc.	8.19	C
MTLS	<i>metallothionein-like 5, testis-specific (tesmin)</i>	11q13.2	water-soluble vitamin biosynthetic proc.	-8.37	
TPD52	<i>tumor protein D52</i>	8q21.13	regeneration	-8.57	BC
ARID4B	<i>AT rich interactive domain 4B (RBP1-like)</i>	1q42.3	organ regeneration	-8.80	C
THC2340878	NA	8q13.2	programmed cell death	-9.06	



**Table 1. Cont.**

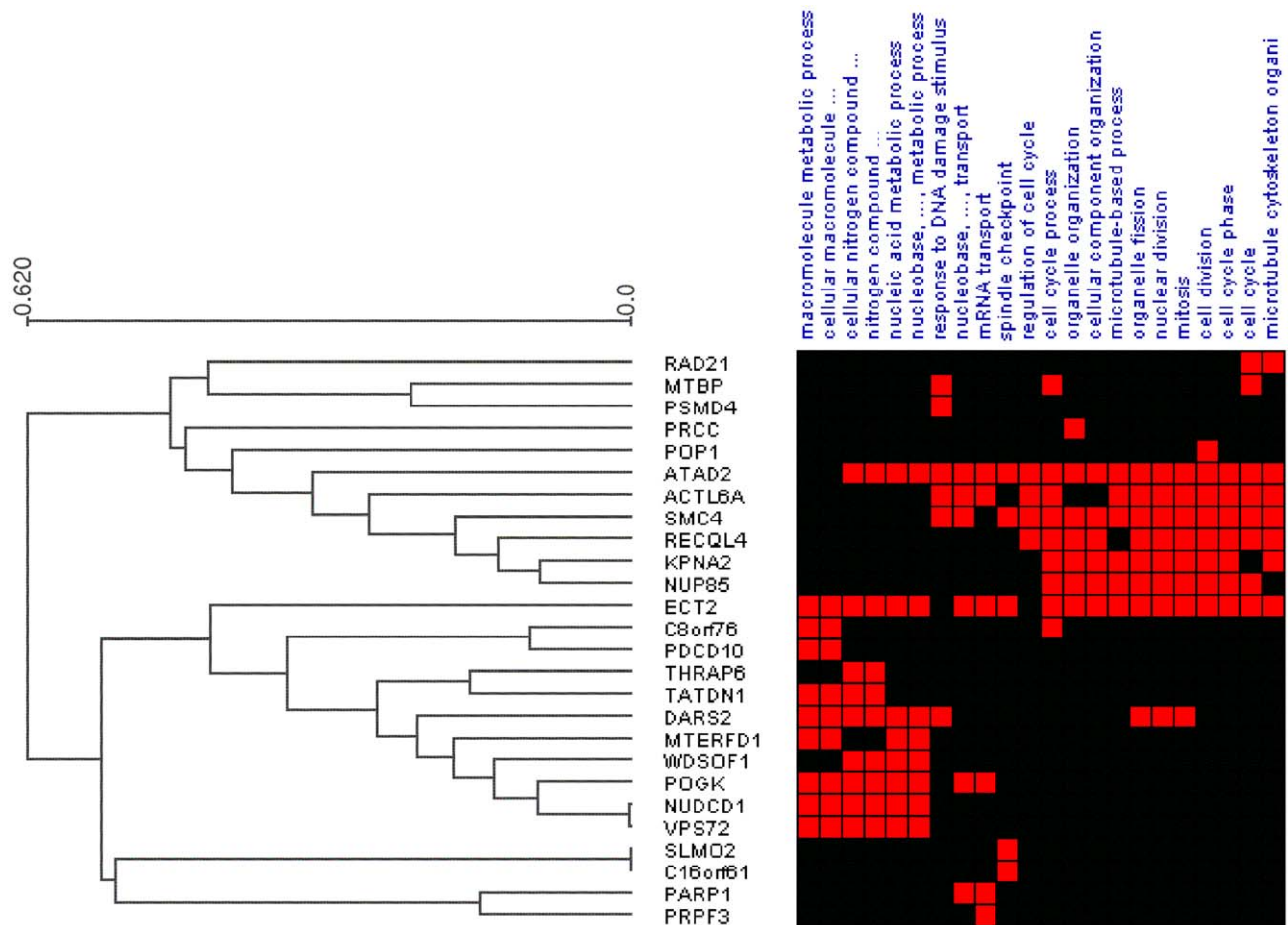
Gene	Full gene name	Cytoband	Highest associated GO term (trait)	Score	Annot
CHTOP	chromatin target of PRMT1	1q21.3	activation of plasma proteins	-11.73	
TMEM70	transmembrane protein 70	8q21.11	regulation of Rho protein signal transd.	-13.02	BC
DPM1	dolichyl-phosphatemannosyltransferase polypeptide 1, cat. subunit	20q13.13	negative regulation of gene expression	-13.36	C
PYCR1	pyrroline-5-carboxylate reductase-like	8q24.3	membrane invagination	-15.15	
IMPAD1	inositol monophosphatase domain containing 1	8q12.1	positive regulation of cell death	-15.88	
STX16	syntaxin 16	20q13.32	cellular protein metabolic process	-15.88	
PIGM	phosphatidylinositol glycan anchor biosynthesis, class M	1q23.2	response to external stimulus	-22.09	

Scores in the table are the negative logarithms of the enrichment scores, the sign indicating whether the association of the trait to the genes is positively or negatively correlated with the iPAC gene. The annotation column indicates genes previously linked with breast cancer (BC) and among those that are not, genes linked to cancer in general (C), based on annotation of the genes obtained with IPA (Ingenuity® Systems, www.ingenuity.com). doi:10.1371/journal.pone.0053014.t001

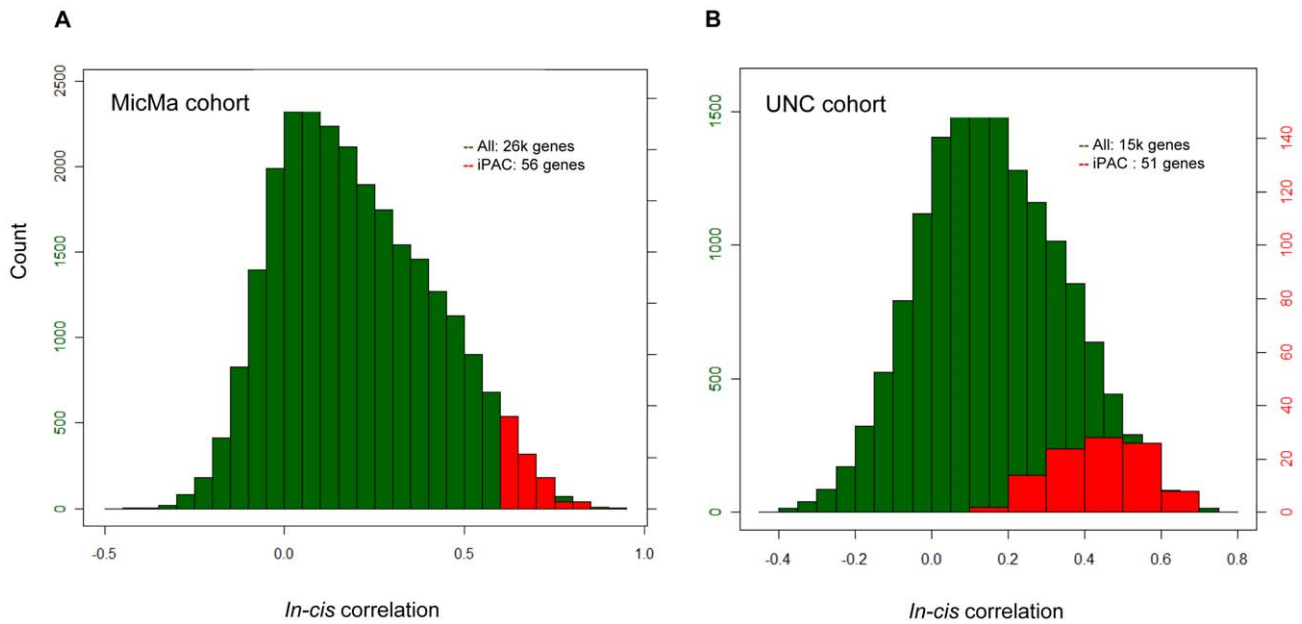
most pronounced for genes in close proximity (see Figure 4A), and the effect of using copy number-adjusted residual expression increases with the in-cis correlation (see Figure 4B and Figure 5).

Overrepresentation of Gene Ontology (GO) biological process terms in the above ranked list of genes was statistically assessed (Figure 6). Out of the  $578 \times 8284 = 4.7 \times 10^6$  potential associations

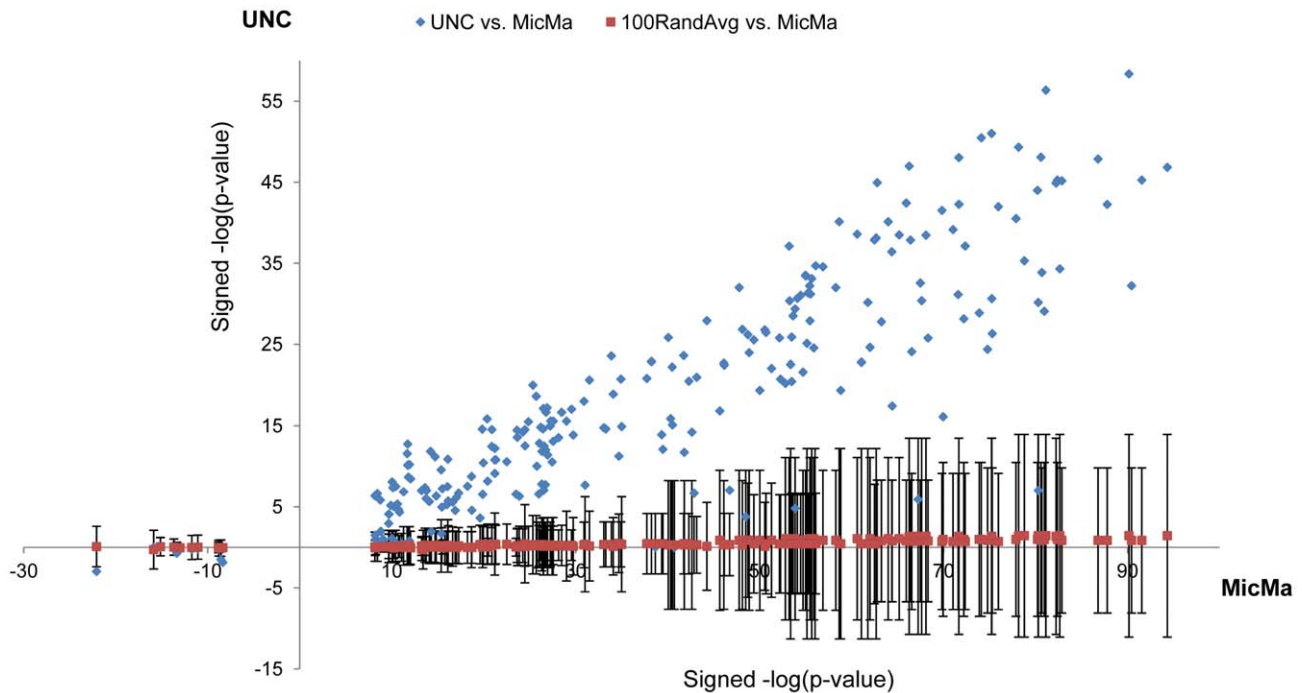
(for every in-cis gene and every GO term tested), we first selected those with enrichment score  $p < 1.04 \times 10^{-8}$  (corresponding to  $p < 0.05$  after Bonferroni correction). This resulted in 19,606 associations covering 467 GO terms and all 578 in-cis genes. Finally, simulations were used to call significant gene-process associations (see Materials and Methods). This yielded a total of



**Figure 7. Associations between iPAC genes and traits (biological processes).** A hierarchical clustered heatmap representation of traits associated with at least four iPAC genes. A red entry indicates a significant association between an iPAC gene and the corresponding trait (see Figure S4 for all the significant associations). The Expander suite [66] using average Euclidian distance was used to calculate and visualize the hierarchical clustering analysis. doi:10.1371/journal.pone.0053014.g007



**Figure 8. Distribution of in-cis correlation levels between copy number and expression in the MicMa and UNC cohorts.** Green bins in the histogram show distribution of in-cis correlation levels of all genes in the data set, while red bins show the distribution for only the identified iPAC genes. The left-hand y-axes in each histogram show the count in each bin among all genes, and the right-hand axes show the count for iPAC genes in each bin. **(A)** Distribution of the in-cis correlation levels in the MicMa cohort. **(B)** Distribution of the in-cis correlation levels in the UNC cohort. The iPAC genes were inferred from the MicMa cohort. doi:10.1371/journal.pone.0053014.g008



**Figure 9. Association consistency of iPAC genes in the validation cohort.** Blue dots represent associations between an iPAC gene and a GO term. The blue dots are plotted according to the level of association, as signed  $-\log(p\text{-value})$ , in the MicMa cohort (x-axis) and in the UNC cohort (y-axis), where signed  $-\log(p\text{-value})$  refers to  $-\log(\text{mHG } p\text{-value})$  for positive associations and  $\log(\text{mHG } p\text{-value})$  for negative associations. A monotone relation is observed, supporting the iPAC behavior of the MicMa inferred iPAC genes in the validation cohort. A bar with a red dot in the center is plotted for each blue dot representing 1 standard deviation (SD) of the associations generated by associating 100 random genes from the UNC cohort to the relevant GO term. doi:10.1371/journal.pone.0053014.g009

276 highly significant associations, covering 56 *in-cis* genes (henceforth called iPAC genes) and 97 unique GO terms (called traits) at a false discovery rate (FDR) of less than 1% (see Figure 2, Table 1 and Table S3). Cell cycle related processes commonly occurred as traits of the iPAC genes, consistent with an association to tumor development and progression related processes (see Figure 7 and Figure S4).

### Properties of the identified iPAC genes

Four of the 56 iPAC genes were commonly deleted and 52 commonly amplified. The iPAC genes encode proteins with various biological roles, including enzymes, regulators of transcription and translation, and transporter molecules (Figure S3B). Five of them were themselves members of the biological process(es) they were found to be associated with (*MTBP*, *RAD21*, *RECQL4*, *SETDB1*, and *SMC4*). Comparing the 56 iPAC genes against the background of all other genes using GOrilla, the 56 genes were found to be associated to four biological processes all commonly disrupted in cancer: cell cycle, cell cycle process, nucleic acid metabolic process, and chromosome organization (Table S4 and Figure S5).

Among the iPAC genes, 38 mapped to chromosomes 1 ( $n = 16$ ) and 8 ( $n = 22$ ), while the rest were located on chromosomes 3, 11, 16, 17, 20, and 22 (Figure 2 and Figure S6). There was a tendency for iPAC genes to reside in blocks of commonly aberrant segments. As would be expected by their mutual proximity and their high *in-cis* correlation, the expression levels of iPAC genes residing in the same block were highly correlated (Figure S7). Accordingly, the scope for further narrowing down the list of candidates based on copy number and expression data alone was limited. However, in several cases, iPAC genes in close proximity were found to be associated with different biological processes. For example, *PRPF3* and *SETDB1* are less than 1 Mb apart from each other and were associated with mRNA transport and chromatin modification, respectively.

Figure S8 shows how the patient samples clustered according to the expression of the iPAC genes. Most of the luminal samples clustered together, as did the basal-like samples, with the latter having a tendency towards higher expression of the iPAC genes. We note that the expression levels of the iPAC genes were not found to be significantly associated with survival (data not shown).

### Knockdown experiment with siRNA

To investigate the effect of the selected iPAC genes on cell viability, siRNA knockdown was performed for three iPAC genes (*ECT2*, *PSMD4* and *MTBP*) in two breast cancer cell lines (MCF7 and MDA-MB-231; see the Supporting Information uploaded to the file inventory: File S1.pdf). For one of the siRNAs tested against *ECT2* we observed a ~30% reduction in cell viability ( $p < 0.05$ , Figure S9) in the MCF7 cell line. *ECT2* is a guanine nucleotide exchange factor for Rho family GTPases and was most strongly associated with the cell cycle GO term and amplified in 15% of the samples. The reduced cell viability after knockdown emphasizes the importance of the iPAC gene *ECT2* in the MCF7 cell line. A smaller reduction in cell viability was also observed for *PSMD4* (data not shown).

### Robust iPAC signature in a validation cohort

In order to validate the robustness of the 56 identified genes, we investigated their iPAC characteristics in an independent breast cancer cohort (UNC), consisting of 73 patients [25]. Out of the 56 iPAC genes identified in our study, 51 were among the genes measured in the UNC study. The five remaining genes (*IMPAD1*, *FAM33A*, *FAM91A1*, *PARP1* and *THC2340878*) had been

removed further upstream in the analysis and data preprocessing in the UNC study.

The *in-cis* correlation between copy number and expression for the iPAC genes ranged from 0.16 (*SETDB1*) to 0.69 (*PPM1D*) in the validation cohort, with an average of 0.43 (Figure 8). To assess the *in-trans* associations for the iPAC genes in the validation cohort, all genes were ranked according to the level of correlation between their copy number-adjusted residual expression and the expression of each iPAC gene in the validation cohort. The association between each iPAC gene and each of the 97 GO terms identified in the original analysis was then assessed. The results showed high level of consistency of enrichments between the two cohorts (Figure 9). As a further confirmation of consistency, we compared these results to the association levels (negative logarithms of the enrichment scores) of 100 random genes to each of the above GO terms. It was found that 95% of the iPAC gene/trait pairs tested in the validation cohort had association levels exceeding  $m + SD$ , where  $m$  is the average and  $SD$  is the standard deviation of association levels obtained for the random genes, and 80% of the pairs had association levels exceeding  $m + 2SD$  (Figure 9). This shows that the level of association of the iPAC genes with their relevant biological process (represented by the GO term) in the validation cohort is not random.

## Discussion

Copy number aberrations are common in breast cancer, but to what extent such aberrations affect cancer cell phenotype through alterations of the transcriptional program is not yet known. The methods we propose here aim to identify genes subject to selection in breast cancer by detecting commonly aberrant genes affected on the gene expression level by genomic aberrations. Furthermore, the method requires the identified genes to be correlated with genes collectively enriched with respect to GO biological processes. Thus, it is through the influence on other genes and their associated processes that the iPAC genes are identified.

### Integrative analysis as a tool for inferring causality

Numerous high-throughput profiling expression studies have identified clusters of genes with expression varying in a coordinated manner over time or across disease states. However, such studies generally give no information about the directionality of gene interactions unless additional information is available. Strong association between the mRNA expression levels of two genes may result from one gene regulating the other, both being regulated by a common factor, or a combination of both. The combined use of copy number and expression data allows the distinction between a situation where the expression of one gene influences the expression of another gene and a situation where the expression levels of the two genes are merely correlated [17].

### Relationship to other methods

Several strategies that aim to identify driver genes in cancer exploit the integration of matched copy number and expression data. Woo *et al.* [23] worked with an integrated copy number and expression data set and used the prognostic significance of genes to guide the selection process. Akavia *et al.* [17] also utilized this sort of integration in their CONEXIC algorithm. Their study assumed that a driver mutation would occur more often than by chance in multiple tumors, that the mutation would be correlated with the expression of a group of genes (a module), and that copy number changes often had an effect on expression of the driver that thus further influenced the expression of the module [17]. The CONEXIC approach is founded on the notion that the expression

levels of the driver, rather than the determinants of that expression level, confers a fitness advantage to the tumor. Alteration of copy number is only one way of achieving this, manifested by a high frequency of aberrations in a patient cohort.

The iPAC approach has a similar rationale as CONEXIC. However, our method differs from the approaches described above in several aspects. First, we use residual expression for the *in-trans* correlation analysis, thus bypassing the potential confounder effect of co-occurring copy numbers. Second, we use a robust enrichment analysis approach to identify aberrations that lead to a significant shift in cancer-related transcriptional programs. Using the enrichment framework, we assign statistical significance to gene-process associations. By taking advantage of the residual expression, the modulator properties of the iPAC genes are more robustly captured. Such modulator effects on biological processes interrupted in cancer may go beyond the direct effects on a pathway; transcriptional responses launched by the cell after physiologic alterations may result from various indirect influences and mechanisms [17], and in this respect, the iPAC genes represent a diverse set of candidates.

### Characteristics of the iPAC genes

The list of iPAC genes includes 16 genes previously associated with breast cancer and 20 additional genes associated with cancer in general (Table 1). For example, *ATAD2* was highly associated to the cell cycle process, indicating that the cell cycle module is activated when *ATAD2* is amplified and overexpressed. *ATAD2* is an ATPase and was recently reported to be a cofactor for the *MYC* oncogene [40]. While copy number was a predominant determinant of *ATAD2* expression levels, other factors also probably influence *ATAD2* expression levels and through its expression level, *ATAD2* is proposed to affect its target process. Another example is *TPD52* which was highly associated to regeneration; this gene has previously been suggested as a potential driver gene and reported amplified and overexpressed in various cancer types, including breast cancer [41,42,43,44]. Furthermore, *PPM1D* was strongly associated to the mitotic cell cycle checkpoint; this gene encodes a serine/threonine phosphatase, maps to the 17q23.2 amplicon and has been shown to be involved in the regulation of several tumor suppressor pathways, including the p53 pathway [45,46]. Amplification of this gene has previously been found to be correlated with overexpression in breast cancer [47]. The iPAC gene *KPNA2* was associated with the trait of nuclear division, and is a member of the importin family of proteins involved in nuclear transport. *KPNA2* has been proposed to be a prognostic marker in breast cancer [48], and overexpression of this gene has been associated with poor prognosis, expression signatures of high proliferation, and tumor grade [49,50].

The iPAC genes also include several genes not previously associated with cancer. One interesting example is the gene *MTL5* which was negatively correlated with the water-soluble vitamin biosynthetic process and encodes a protein with homology to the metal-binding motif of the metallothionein (MT) family [51]. *MTL5* is located on chromosome 11q13.2 and was found amplified in 17% of our investigated breast cancer samples. Through their ability to bind metal, MT proteins can affect the activity of several proteins and enzymes dependent on metals as co-factors. In this respect, MT proteins play important roles in apoptosis and proliferation [52]. Furthermore, elevated expression of MT proteins has been reported in various cancer types, including breast cancer [52,53,54] and was also linked to modulation of p53 activity through zinc exchange [55,56]. Dividing our samples according to p53 mutational status, *MTL5* was one of the top 2% most down regulated genes in mutated p53

( $p < 10^{-5}$ ) (data not shown). As *MTL5* was found to be amplified in a significant proportion of the samples in our cohort, and because of its iPAC properties, our results indicate that the gene may have an important role in breast cancer, similar to the homologous MT proteins. Many homologs of *MTL5* exist both in animals and plants, suggesting that the function of this gene is conserved [57].

### Proof-of-concept knockdown experiments

We selected three iPAC genes for siRNA knockdown experiments. Out of these, silencing of *ECT2* led to significant decrease in cell viability. By using our approach, this gene was found to be most highly associated with cell cycle related traits. The protein ECT2 has been shown to regulate cytokinesis [58], which can explain the effect on cell viability after knockdown. *ECT2* has been found to be up-regulated during transition to malignancy in a mouse model [59], to be amplified and overexpressed in non-small cell lung cancer [60], and to have an elevated expression in colorectal cancer [61].

In another study, siRNA-mediated knockdown of the iPAC gene *RAD21* was found to decrease cell growth and enhance cytotoxicity in MCF7 and T47D breast cancer cell lines [62]. *RAD21* encodes a phosphoprotein and is a component of the cohesin complex essential for chromosome segregation during mitosis/meiosis and DNA repair [63,64]. In our breast cancer cohort, *RAD21* was found to be amplified in 36% of the tumor samples and to be highly associated with the cell cycle trait. Strong association to cell cycle has been shown to correlate with cell proliferation for the same patient samples [26], and enhanced expression of this protein has been associated with poor prognosis and resistance to chemotherapy in breast cancers [65].

### Conclusion

Whole-genome integrative analyses of copy number and gene expression data is a useful tool in genome-wide searches for candidate driver genes in cancer. The first phase of analysis is typically to detect genes with frequent aberrations in copy number and strong *in-cis* correlation to gene expression. For example, in our study, the gene *ERBB2* was ranked 7 out of 6373 genes with respect to the *in-cis* correlation level, indicating a direct link between copy number and expression. However, even among those genes that satisfy these criteria there are potentially many passengers with no direct oncogenic role. In the opposite direction, there may be genes that manifest moderate *in-cis* expression but still drive cancer-related processes through their expression level. Regulation of these expression levels may be selected for in the cancer through copy number changes as well as other mechanisms (e.g. altered methylation). Our aim has been to detect genes for which the gene-gene correlation structure of the expression data reveals additional evidence to support a link to a phenotype. The iPAC gene *ATAD2*, which was ranked only 450 in the *in-cis* correlated genes and hence would easily have been missed by *in-cis* focused methods, illustrates this point. Several similar examples are described above, indicating that the iPAC procedure does indeed capture biologically relevant genes not found on the top of the list of *in-cis* correlated genes.

Validation in an independent cohort of the proposed methodology and of the observation regarding the 56 iPAC genes found in our initial analysis supports method robustness and justify focus on the identified genes with respect to their tumorigenic role. In this study, we have selected GO biological process terms as they represent a comprehensive view of functional traits. It is clearly possible to select other annotation approaches for this purpose. For example, one could assess the enrichment of molecular pathways or transcription factor networks



among the *in-trans* correlated genes. We provide cell line based experimental data for the effect of *ECT2* on cell viability; however, further functional validation is still needed to firmly establish the role of the 56 iPAC genes in breast cancer.

The framework for the identification of *in-trans* regulatory mechanisms, as exemplified here in human breast cancer, is applicable to any kind of data with existing comparable aCGH, expression profiles and a collection of gene sets representing transcriptional programs. We propose this method as an unbiased and robust approach for the identification of genes of relevance to tumorigenesis.

## Supporting Information

**Figure S1 Copy number and expression correlations.** (A) Pearson correlation of copy number data for all the 25,688×25,688 genes. (B) Pearson correlation of copy number and expression of all 25,688×25,688 genes, with *in-cis* correlation along the diagonal. Color map represents the Pearson correlation coefficient. (TIF)

**Figure S2 GO terms enriched among the *in-cis* correlated genes.** The GO biological process statistical enrichment analysis was performed by GOrilla. The input for GOrilla in this analysis was the list of 6373 commonly aberrant genes ranked according to their *in-cis* correlation. (TIF)

**Figure S3 Functional annotation of genes.** (A) The 578 *in-cis* genes; (B) The 56 iPAC genes. The genes were annotated using IPA (Ingenuity® Systems, www.ingenuity.com). (TIF)

**Figure S4 Associations between iPAC genes and their traits (GO terms).** Extension of Figure 7. A hierarchical clustered heatmap representation of all significant associations between iPAC genes and biological processes. A red entry indicates a significant association between an iPAC gene and the corresponding traits. The Expander suite [66] using average Euclidian distance was used to calculate and visualize the hierarchical clustering analysis. (TIF)

**Figure S5 Statistical enrichment analysis of the 56 iPAC genes for GO biological processes.** Performed by GOrilla, on the list of 56 iPAC genes, compared to a background gene list consisting of all the remaining genes. (TIF)

**Figure S6 Sample-wise genomic copy number aberrations.** Copy number aberrations are shown for chromosomes harboring at least one iPAC gene. The x-axis represents chromosomal location and the y-axis represents sample no (1–100). Green lines are regions of loss ( $\theta < -0.2$ ), and red lines are regions of gain ( $\theta > 0.2$ ). The vertical black lines indicate the locations of the 56 iPAC genes. (TIF)

## References

- Hanahan D, Weinberg RA (2000) The Hallmarks of Cancer. *Cell* 100: 57–70.
- Hanahan D, Weinberg Robert A (2011) Hallmarks of Cancer: The Next Generation. *Cell* 144: 646–674.
- Russnes HG, Vollan HKM, Lingjærde OC, Krasnitz A, Lundin P, et al. (2010) Genomic Architecture Characterizes Tumor Progression Paths and Fate in Breast Cancer Patients. *Science Translational Medicine* 2: 38ra47.
- Inaki K, Hillmer AM, Ukil L, Yao F, Woo XY, et al. (2011) Transcriptional consequences of genomic structural aberrations in breast cancer. *Genome Research* 21: 676–687.
- Hyman E, Kauraniemi Pi, Hautaniemi S, Wolf M, Mousses S, et al. (2002) Impact of DNA Amplification on Gene Expression Patterns in Breast Cancer. *Cancer Research* 62: 6240–6245.
- Bergamaschi A, Kim YH, Wang P, Sorlie T, Hernandez-Boussard T, et al. (2006) Distinct patterns of DNA copy number alteration are associated with different clinicopathological features and gene-expression subtypes of breast cancer. *Genes, Chromosomes and Cancer* 45: 1033–1040.
- Fridlyand J, Snijders A, Ylstra B, Li H, Olshen A, et al. (2006) Breast tumor copy number aberration phenotypes and genomic instability. *BMC Cancer* 6: 96.

**Figure S7 Correlation plots.** (A) Pairwise correlations of log copy number of the 56 iPAC genes. (B) Pairwise correlations of log expression levels of the 56 iPAC genes. Chromosomes are indicated with numbers. (TIF)

**Figure S8 Hierarchical clustering of the expression levels of the 56 iPAC genes.** Samples are color-coded according to gene expression subtype. The clustering was made with Pearson correlation using Ward linkage. Three samples could not be subtyped and were omitted from the analysis. Color map represents log expression values. (TIF)

**Figure S9 siRNA knockdown of the iPAC gene *ECT2*.** (A) Effect of siRNA knockdown of *ECT2* on cell viability in the MCF7 cell line. Four various siRNAs against *ECT2* were tested in addition to controls (bars show SD from eight replicates). The *ECT2\_5* siRNA shows a statistically significant reduction in cell viability compared to the non-transfected cells (asterisk; Student's t-test,  $p < 0.05$ ). (B) Relative quantification (RQ) of *ECT2* mRNA after siRNA transfections (9 replicates), showing the specificity of the knockdown in the MCF7 cell line. The data were normalized to the control (cells + transfection lipid). (TIF)

**Table S1 GOrilla results from ranking the 6373 commonly aberrant genes.** (XLSX)

**Table S2 The 578 *in-cis* genes.** (XLSX)

**Table S3 Description of the 56 iPAC genes.** (XLSX)

**Table S4 GOrilla results from 56 iPAC genes versus background gene set.** (XLSX)

**File S1 Cell culture and siRNA transfection.** (PDF)

## Acknowledgments

We thank Silje Nord and Grethe I. G. Alnæs for valuable discussions and assistance in data normalization and technical support. We thank Roy Navon and Amir Ben-Dor for useful discussions.

## Author Contributions

Designed the software used in the analysis: MRA IS OCL ZY. Data analysis: MRA IS LOB KL OCL ZY. Interpretation of results: MRA IS LOB KL KKS ALBD OCL ZY. Conceived and designed the experiments: MRA SN BN KKS VNK ALBD OCL ZY. Contributed reagents/materials/analysis tools: IS KL DL BN KKS VNK ALBD OCL ZY. Wrote the paper: MRA OCL IS LOB ALBD OCL ZY.

8. Naylor TL, Greshock J, Wang Y, Colligon T, Yu QC, et al. (2005) High resolution genomic analysis of sporadic breast cancer using array-based comparative genomic hybridization. *Breast Cancer Res* 7: R1186–R1198.
9. Pollack JR, Perou CM, Alizadeh AA, Eisen MB, Pergamenschikov A, et al. (1999) Genome-wide analysis of DNA copy-number changes using cDNA microarrays. *Nat Genet* 23: 41–46.
10. Pollack JR, Sorlie T, Perou CM, Rees CA, Jeffrey SS, et al. (2002) Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proceedings of the National Academy of Sciences* 99: 12963–12968.
11. Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, et al. (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* 486: 346–352.
12. Dehan E, Ben-Dor A, Liao W, Lipson D, Frimer H, et al. (2007) Chromosomal aberrations and gene expression profiles in non-small cell lung cancer. *Lung Cancer* 56: 175–184.
13. Goh XY, Rees JRE, Paterson AL, Chin SF, Marioni JC, et al. (2011) Integrative analysis of array-comparative genomic hybridisation and matched gene expression profiling data reveals novel genes with prognostic significance in oesophageal adenocarcinoma. *Gut*.
14. Taylor BS, Schultz N, Hieronymus H, Gopalan A, Xiao Y, et al. (2010) Integrative Genomic Profiling of Human Prostate Cancer. *Cancer Cell* 18: 11–22.
15. Parris TZ, Danielsson A, Nemes S, Kovács A, Delle U, et al. (2010) Clinical Implications of Gene Dosage and Gene Expression Patterns in Diploid Breast Carcinoma. *Clinical Cancer Research* 16: 3860–3874.
16. Creixell P, Schoof EM, Erler JT, Lindling R (2012) Navigating cancer network attractors for tumor-specific therapy. *Nat Biotech* 30: 842–848.
17. Akavia UD, Litvin O, Kim J, Sanchez-Garcia F, Kotliar D, et al. (2010) An Integrated Approach to Uncover Drivers of Cancer. *Cell* 143: 1005–1017.
18. Chari R, Coe B, Vucic E, Lockwood W, Lam W (2010) An integrative multi-dimensional genetic and epigenetic strategy to identify aberrant genes and pathways in cancer. *BMC Systems Biology* 4: 67.
19. Natrajan R, Weigelt B, Mackay A, Geyer F, Grigoriadis A, et al. (2010) An integrative genomic and transcriptomic analysis reveals molecular pathways and networks regulated by copy number aberrations in basal-like, HER2 and luminal cancers. *Breast Cancer Research and Treatment* 121: 575–589.
20. Xu C, Liu Y, Wang P, Fan W, Rue T, et al. (2010) Integrative analysis of DNA copy number and gene expression in metastatic oral squamous cell carcinoma identifies genes associated with poor survival. *Molecular Cancer* 9: 143.
21. Cho Y-J, Tsherniak A, Tamayo P, Santagata S, Ligon A, et al. (2011) Integrative Genomic Analysis of Medulloblastoma Identifies a Molecular Subgroup That Drives Poor Clinical Outcome. *Journal of Clinical Oncology* 29: 1424–1430.
22. Kim H, Huang W, Jiang X, Pennicooke B, Park PJ, et al. (2010) Integrative genome analysis reveals an oncomir/oncogene cluster regulating glioblastoma survivorship. *Proceedings of the National Academy of Sciences* 107: 2183–2188.
23. Woo HG, Park ES, Lee J-S, Lee Y-H, Ishikawa T, et al. (2009) Identification of Potential Driver Genes in Human Liver Carcinoma by Genomewide Screening. *Cancer Research* 69: 4059–4066.
24. Naume B, Borgen E, Kvalheim G, Kåresen R, Qvist H, et al. (2001) Detection of Isolated Tumor Cells in Bone Marrow in Early-Stage Breast Carcinoma Patients. *Clinical Cancer Research* 7: 4122–4129.
25. Kristensen VN, Vaske CJ, Ursini-Siegel J, Van Loo P, Nordgard SH, et al. (2011) Integrated molecular profiles of invasive breast tumors and ductal carcinoma in situ (DCIS) reveal differential vascular and interleukin signaling. *Proceedings of the National Academy of Sciences*.
26. Enerly E, Steinfeld I, Kleivi K, Leivonen S-K, Aure MR, et al. (2011) miRNA-mRNA Integrated Analysis Reveals Roles for miRNAs in Primary Breast Tumors. *PLoS ONE* 6: e16915.
27. Wiedswang G, Borgen E, Kåresen R, Kvalheim G, Nesland JM, et al. (2003) Detection of Isolated Tumor Cells in Bone Marrow Is an Independent Prognostic Factor in Breast Cancer. *Journal of Clinical Oncology* 21: 3469–3478.
28. Sorlie T, Wang Y, Xiao C, Johnsen H, Naume B, et al. (2006) Distinct molecular mechanisms underlying clinically relevant subtypes of breast cancer: gene expression analyses across three different platforms. *BMC Genomics* 7: 127.
29. Naume B, Zhao X, Synnæstvedt M, Borgen E, Russnes HG, et al. (2007) Presence of bone marrow micrometastasis is associated with different recurrence risk within molecular subtypes of breast cancer. *Molecular oncology* 1: 160–171.
30. Nordgard SH, Johansen FE, Alnæs GIG, Bucher E, Syvänen A-C, et al. (2008) Genome-wide analysis identifies 16q deletion associated with survival, molecular subtypes, mRNA expression, and germline haplotypes in breast cancer patients. *Genes, Chromosomes and Cancer* 47: 680–696.
31. Weigman V, Chao H-H, Shabalin A, He X, Parker J, et al. (2011) Basal-like Breast cancer DNA copy number losses identify genes involved in genomic instability, response to therapy, and patient survival. *Breast Cancer Research and Treatment*: 1–16.
32. Lingjærde OC, Baumbusch LO, Liestol K, Glad IK, Borresen-Dale A-L (2005) CGH-Explorer: a program for analysis of array-CGH data. *Bioinformatics* 21: 821–822.
33. Baumbusch LO, Aaroe J, Johansen FE, Hicks J, Sun H, et al. (2008) Comparison of the Agilent, ROMA/NimbleGen and Illumina platforms for classification of copy number alterations in human breast tumors. *BMC Genomics* 9: 379.
34. Nilsen G, Liestol K, Van Loo P, Vollen HKM, Eide MB, et al. (2012) Copynumber: Efficient algorithms for single- and multi-track copy number segmentation. *BMC Genomics* 13: 591.
35. Lipson D, Ben-Dor A, Dehan E, Yakhini Z (2004) Joint analysis of DNA copy numbers and gene expression levels. *Proceedings of Workshop on Algorithms in Bioinformatics (WABI)* in Bergen, Norway.
36. Lipson D, Aumann Y, Ben-Dor A, Linnal N, Yakhini Z (2006) Efficient calculation of interval scores for DNA copy number data analysis. *J Comput Biol* 13: 215–228.
37. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene Ontology: tool for the unification of biology. *Nat Genet* 25: 25–29.
38. Eden E, Lipson D, Yogev S, Yakhini Z (2007) Discovering Motifs in Ranked Lists of DNA Sequences. *PLoS Comput Biol* 3: e39.
39. Eden E, Navon R, Steinfeld I, Lipson D, Yakhini Z (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics* 10: 48.
40. Ciró M, Prosperini E, Quarto M, Grazini U, Walfridsson J, et al. (2009) ATAD2 Is a Novel Cofactor for MYC, Overexpressed and Amplified in Aggressive Tumors. *Cancer Research* 69: 8491–8498.
41. Shehata M, Bièche I, Boutros R, Weidenhofer J, Fanany S, et al. (2008) Nonredundant Functions for Tumor Protein D52-Like Proteins Support Specific Targeting of TPD52. *Clinical Cancer Research* 14: 5050–5060.
42. Balleine RL, Fejzo MS, Sathasivam P, Basset P, Clarke CL, et al. (2000) The hD52 (TPD52) gene is a candidate target gene for events resulting in increased 8q21 copy number in human breast carcinoma. *Genes, Chromosomes and Cancer* 29: 48–57.
43. Choschizk M, Lassen P, Lebeau A, Marx AH, Terracciano L, et al. (2010) Amplification of 8q21 in breast cancer is independent of MYC and associated with poor patient outcome. *Mod Pathol* 23: 603–610.
44. Byrne JA, Balleine RL, Fejzo MS, Mercieca J, Chiew Y-E, et al. (2005) Tumor protein D52 (TPD52) is overexpressed and a gene amplification target in ovarian cancer. *International Journal of Cancer* 117: 1049–1054.
45. Lu X, Ma O, Nguyen T-A, Jones SN, Oren M, et al. (2007) The Wip1 Phosphatase Acts as a Gatekeeper in the p53-Mdm2 Autoregulatory Loop. *Cancer Cell* 12: 342–354.
46. Fujimoto H, Onishi N, Kato N, Takekawa M, Xu XZ, et al. (2005) Regulation of the antioncogenic Chk2 kinase by the oncogenic Wip1 phosphatase. *Cell Death Differ* 13: 1170–1180.
47. Lambros MB, Natrajan R, Geyer FC, Lopez-Garcia MA, Dedes KJ, et al. PPM1D gene amplification and overexpression in breast cancer: a qRT-PCR and chromogenic in situ hybridization study. *Mod Pathol* 23: 1334–1345.
48. Dahl E, Kristiansen G, Gottlob K, Klamann I, Ebner E, et al. (2006) Molecular Profiling of Laser-Microdissected Matched Tumor and Normal Breast Tissue Identifies Karyopherin Alpha2 as a Potential Novel Prognostic Marker in Breast Cancer. *Clinical Cancer Research* 12: 3950–3960.
49. Wang Y, Kljajn JGM, Zhang Y, Siewewerts AM, Look MP, et al. (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *The Lancet* 365: 671–679.
50. Sotiriou C, Wirapati P, Loi S, Harris A, Fox S, et al. (2006) Gene Expression Profiling in Breast Cancer: Understanding the Molecular Basis of Histologic Grade To Improve Prognosis. *Journal of the National Cancer Institute* 98: 262–272.
51. Olesen C, Møller M, Byskov AG (2004) Tesmin transcription is regulated differently during male and female meiosis. *Molecular Reproduction and Development* 67: 116–126.
52. Gomulkiewicz A, Podhorska-Okolow M, Szulc R, Smorag Z, Wojnar A, et al. Correlation between metallothionein (MT) expression and selected prognostic factors in ductal breast cancers. *Folia Histochemica et Cytobiologica* 48: 242–248.
53. Kim HG, Kim JY, Han EH, Hwang YP, Choi JH, et al. Metallothionein-2A overexpression increases the expression of matrix metalloproteinase-9 and invasion of breast cancer cells. *FEBS letters* 585: 421–428.
54. Cherian MG, Jayasurya A, Bay B-H (2003) Metallothioneins in human tumors and potential roles in carcinogenesis. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 533: 201–209.
55. Ostrakhovitch EA, Olsson P-E, Jiang S, Cherian MG (2006) Interaction of metallothionein with tumor suppressor p53 protein. *FEBS letters* 580: 1235–1238.
56. Ostrakhovitch EA, Olsson P-E, von Hofsten J, Cherian MG (2007) P53 mediated regulation of metallothionein transcription in breast cancer cells. *Journal of Cellular Biochemistry* 102: 1571–1583.
57. Matsuura T, Kawasaki Y, Miwa K, Sutou S, Ohinata Y, et al. (2002) Germ cell-specific nucleocytoplasmic shuttling protein, tesmin, responsive to heavy metal stress in mouse testes. *Journal of Inorganic Biochemistry* 88: 183–191.
58. Tatsumoto T, Xie X, Blumenthal R, Okamoto I, Miki T (1999) Human Ect2 Is an Exchange Factor for Rho Gtpases, Phosphorylated in G2/M Phases, and Involved in Cytokinesis. *The Journal of Cell Biology* 147: 921–928.
59. Namba R, Maglione JE, Young IJT, Borowsky AD, Cardiff RD, et al. (2004) Molecular Characterization of the Transition to Malignancy in a Genetically Engineered Mouse-Based Model of Ductal Carcinoma In situ. *Molecular Cancer Research* 2: 453–463.
60. Justilien V, Fields AP (2009) Ect2 links the PKC[ $\iota$ ]-Par6[ $\alpha$ ] complex to Rac1 activation and cellular transformation. *Oncogene* 28: 3597–3607.

61. Jung Y, Lee S, Choi H-S, Kim S-N, Lee E, et al. (2011) Clinical Validation of Colorectal Cancer Biomarkers Identified from Bioinformatics Analysis of Public Expression Data. *Clinical Cancer Research* 17: 700–709.
62. Atienza JM, Roth RB, Rosette C, Smylie KJ, Kammerer S, et al. (2005) Suppression of RAD21 gene expression decreases cell growth and enhances cytotoxicity of etoposide and bleomycin in human breast cancer cells. *Molecular Cancer Therapeutics* 4: 361–368.
63. Peters J-M, Tedeschi A, Schmitz J (2008) The cohesin complex and its roles in chromosome biology. *Genes & Development* 22: 3089–3114.
64. Birkenbihl RP, Subramani S (1992) Cloning and characterization of rad21 an essential gene of *Schizosaccharomyces pombe* involved in DNA double-strand-break repair. *Nucleic Acids Research* 20: 6605–6611.
65. Xu H, Yan M, Patra J, Natrajan R, Yan Y, et al. (2011) Enhanced RAD21 cohesin expression confers poor prognosis and resistance to chemotherapy in high grade luminal, basal and HER2 breast cancers. *Breast Cancer Research* 13: R9.
66. Shamir R, Maron-Katz A, Tanay A, Linhart C, Steinfeld I, et al. (2005) EXPANDER – an integrative program suite for microarray data analysis. *BMC Bioinformatics* 6.