Preferential allelic expression of *PIK3CA* mutations is frequent in breast cancer and is prognostically significant

Lizelle Correia [1], Joana M. Xavier[1], Ramiro Magno [1] , Bernardo P. de Almeida [1,#], Filipa Esteves[1,2], Isabel Duarte[1], Matthew Eldridge [3], Chong Sun [4], Astrid Bosma [4], Lorenza Mittempergher [4], Ana Marreiros [1,5], Rene Bernards [4], Carlos Caldas [3,6,7,8], Suet-Feung Chin [3,6,8,¶] and Ana-Teresa Maia[1,5,¶]

¶ Joint last authors/Corresponding authors

Correspondence:

suet-feung.chin@cruk.cam.ac.uk

atmaia@ualg.pt

[1] Centre for Biomedical Research (CBMR), Algarve Biomedical Center (ABC), Universidade do Algarve, Faro, Portugal

[2] ProRegeM-PhD Program in Mechanisms of Disease and Regenerative Medicine, Universidade do Algarve , Faro, Portugal

[3] Cancer Research UK Cambridge Institute, Li Ka Shing Centre, University of Cambridge, Robinson Way, Cambridge, UK

[4] Division of Molecular Carcinogenesis, The Netherlands Cancer Institute, Amsterdam, The Netherlands

[5] Faculty of Medicine and Biomedical Sciences (FMCB), Universidade do Algarve, Faro, Portugal

[6] Department of Oncology, University of Cambridge, Cambridge, UK

[7] Cambridge Breast Unit, Addenbrooke's Hospital, Cambridge University Hospital NHS Foundation Trust and NIHR Cambridge Biomedical Research Centre, Cambridge, UK

[8] Cancer Research UK Cambridge Cancer Centre, Cambridge, UK

Abstract

Genomic allelic imbalances of mutations have been widely associated with tumor evolution and treatment response. However, allelic imbalances generated at the transcriptomic level by germline and somatic regulatory variants have been less studied. We found widespread allelic expression imbalance between *PIK3CA* missense mutant and wild-type alleles in breast cancers, predominantly towards the preferential expression of the mutant allele. Expression imbalance was more frequently due to regulatory variation in *cis*, although contribution of copy number allelic imbalances at the DNA level showed a greater effect. Preferential expression of one allele due to *cis*-regulatory variant effects was associated with poor prognosis, particularly identifying a poorer prognosis subgroup in ER^+, PR^+ and $Her2^-$ tumors that expressed the mutant allele at extremely low levels. This knowledge challenges the benefit of treating these patients with PI3Kα inhibitors. Overall, our work raises the clinical importance of *PIK3CA* mutations by demonstrating that their transcriptional allelic imbalance is prognostic in breast tumor biology and by presenting compelling evidence that allelic expression levels of mutations should be taken into consideration during patient clinical management.

Keywords: breast cancer, *PIK3CA* mutations, *cis*-regulation, gene expression, prognosis, allele-specific expression

Introduction

Activating oncogenic mutations are often characterized by gain-of-function single base alterations or focal DNA copy-number amplifications, where gain of just a single copy of a mutant allele is sufficient for tumorigenesis [1]. These gains change the stoichiometric balance between mutant and wild-type alleles and are selected for in cancers, affecting approximately half of all oncogenic driver mutations [1]. Ultimately, they could dictate prognosis and therapeutic sensitivity.

However, the impact of gene dosage differences of oncogenic mutations generated at the level of gene expression have been largely unexplored. Genetic variation and mutations regulate gene expression in an allele-specific manner – known as *cis*-regulatory variation [2]

– by altering protein and miRNA binding. Normal *cis*-regulatory variation affects most of the human genome in all tissues and generates the wealth of phenotypic variation seen in a species [3–5]. Recently, it has been shown that there is extensive contribution from non-coding variants to RNA alterations observed in tumors [6] including allelic imbalance of somatic mutations [7]. But one unsolved aspect is how much each mechanism contributes to generating allelic imbalances in expression, and whether they do it independently or in synergy.

In breast tissue, germline regulatory variation affects commonly mutated genes [8] and is associated with risk to disease. We and others have shown that variants affecting the expression levels of *BRCA1* and *BRCA2* modify risk of breast cancer in germline mutation carriers [9,10]. We found that carriers of germline nonsense mutations in *BRCA2*, a tumor suppressor gene, had a lower risk of developing breast cancer when the remaining wild-type allele was highly expressed.

Here, we hypothesize that *cis*-regulatory variation also modulates the penetrance of oncogenic coding mutations. We postulate that an oncogenic activating mutation that is present in the tumors of two patients will have a stronger clinical impact in the patient whose mutation is predominantly expressed. We tested this model in the context of heterozygous mutations in *PIK3CA*, the most frequently mutated gene in breast cancer. Firstly, we investigated whether normal *cis*-regulatory variation regulated the expression of *PIK3CA* in normal breast tissue; then we calculated allelic expression ratios between mutant and wild-type copies in tumors from two large breast cancer datasets – METABRIC and TCGA. This approach allows us to distinguish between expression imbalances that were generated from cis-regulatory variation alone, by altered DNA copy-number, or both mechanisms.

Thus, we investigated whether *cis*-regulatory variation operating either in synergy or independently of copy number alterations, could modify the penetrance of *PIK3CA* somatic mutations in breast cancer.

Results

Normal *cis*-regulatory variation affects *PIK3CA* expression in healthy breast tissue

To investigate whether the expression of *PIK3CA* in normal breast tissue was modulated by *cis*-regulatory variation (Figure 1A), we analyzed data from previous allelic expression analysis of normal breast tissue from 64 healthy donors [11]. We calculated the logarithm of the ratio of

expression of one allele by the other in heterozygous SNP positions (aeSNPs) in all samples, and defined differential allelic expression (DAE) when the difference between alleles was greater than or equal to 1.5-fold, i.e., the absolute value of $\log_2$ allelic expression ratio ≥ 0.58 (See Methods). A key advantage of this approach is its robustness to detect *cis*-acting variant effects, as it cancels out the *trans* effects that might be acting on the same gene and influence both alleles equally. We found that six SNPs in *PIK3CA* displayed DAE (daeSNPs), out of 14 tested aeSNPs (Figure 1B). All daeSNPs, except rs3729679, are in strong linkage disequilibrium (LD) with each other (Supplementary Table 1). rs3960984 showed the largest proportion of heterozygotes displaying DAE (57%), while the smallest fraction (14%) was shared by three daeSNPs: rs12488074, rs4855093 and rs9838411.

Table 1 - DAE of normal breast samples in six daeSNPs located at the PIK3CA region

daeSNP	Heterozygotes with DAE *	Preferentially Expressed Allele	Mean DAE *	SD	Min-Max	*P* t.test (one sample)
rs7636454	24% (5/21)	A	0.68	0.04	0.63-0.76	5.70E-06
rs3960984	57% (12/21)	G	1.11	0.46	0.64-2.28	5.91E-06
rs3729679	20% (7/35)	G	0.85	0.21	0.61-1.17	6.07E-05
	3% (1/35)	A	0.63	0	0.63-0.63	-
rs12488074	14% (3/21)	G	0.74	0.09	0.61-0.82	8.06E-03
rs4855093	9% (2/22)	A	0.74	0.04	0.70-0.78	3.23E-02
	5% (1/22)	G	1.1	0	1.10-1.10	-
rs9838411	14% (3/22)	A	0.76	0.2	0.59-1.04	3.34E-02

* DAE calculated as log2(expression of allele 1/expression of allele 2); SD – Standard Deviation; *P* - p-value

The range of allelic expression imbalances detected was between 1.55 and 2.14-fold (DAE ratio between 0.63 and 1.10). In four out of the six daeSNPs (rs7636454, rs3960984, rs12488074, rs9838411), the DAE ratios showed a unilateral distribution, with all samples with DAE preferentially expressing the same allele. For the other two daeSNPs, only one sample for each differed in the preferentially expressed allele. These patterns of DAE distribution suggest that the daeSNPs and the possible functional regulatory SNPs (rSNPs) are in strong, but incomplete, LD with each other [12].

Candidate rSNP for *PIK3CA* affects a NF-YA binding site at its promoter

Next, we identified the rSNPs acting on *PIK3CA* by carrying mapping analysis for all daeSNPs, using genotyping information from the SNPs that were located within a 250kb window from the beginning and end of the *PIK3CA* gene (NM_006218, GRCh38/hg38). Initially we identified a list of 44 candidate SNPs that, together with their 229 proxy SNPs, significantly explained the DAE observed (See Methods, Supplementary Table 2).

Functional analysis was carried for these 273 candidates, starting with *in-silico* analysis to identify those with greater potential for being *cis*-regulatory, followed by *in-vitro* validation, resulting in one strong candidate rSNP, rs2699887. This variant was significantly associated with DAE detected at rs12488074 (permuted $p = 0.03$) (Figure 1C), and was associated with differences in the total expression of *PIK3CA* ($p = 0.012$, Supplementary Figure 1) in tumors from METABRIC[24]. rs2699887 resides in the first intron of *PIK3CA*, in a region rich in epigenetic marks and classified as an active promoter in breast mammary epithelial cells (vHMEC) and breast myoepithelial primary cells (Figure 1D). The motif position weight matrix (PWM) analysis [19,20] of the sequence surrounding rs2699887 suggested that NF-Y proteins could have preferential binding affinity to the minor allele (T allele) (Figure 1E). Electrophoretic mobility shift assays (EMSA) confirmed the preferential protein-DNA binding potential of the minor allele of rs2699887 (T allele), which upon competition and supershift assays was confirmed to bind preferentially to the transcription factor NF-YA (Figure 1F).

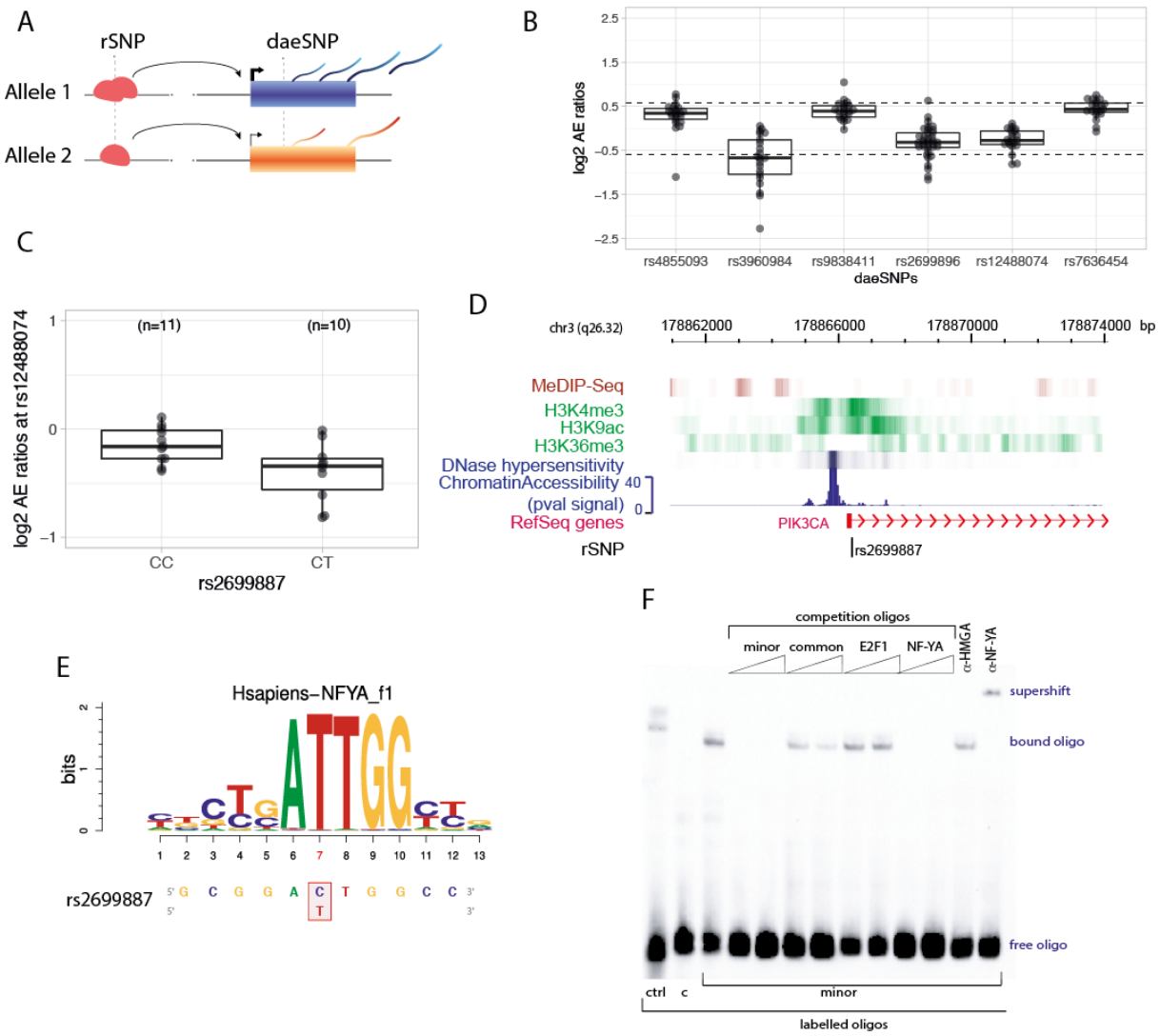

Figure 1 – *Cis*-regulatory variation impacts on *PIK3CA* gene expression in normal breast tissue. **A** – Schematic representation of *cis*-regulatory SNP (rSNP) generating DAE on a target gene. Different alleles of an rSNP may, for example, have differential binding affinities to transcription factors (red), generating differences in the expression of both alleles, the effect of which can be measured in heterozygous individuals for the daeSNP. **B** - DAE ratios for six daeSNPs in the *PIK3CA* gene region, each dot is a heterozygous individual for the corresponding variant indicated in the x-axis, dotted lines delimit the levels of 1.5-fold difference for either allele preferential expression (|DAE|= 0.58); **C** – rs2699887 is significantly associated with DAE for rs12488074. Indicated p-values correspond to Student's *t*-test test, with 1000 permutations; **D** – Genomic view of the location of rs2699887, showing methylation of surrounding CpG sites (MeDIP-seq), typical active promoter chromatin modification marks (H3K4me3, H3K9ac, H3K36me3) and open chromatin status (*DNAse*I hypersensitivity and Chromatin Accessibility p-val), as obtained from the Roadmap of Epigenomics project for breast HMECs; **E** – PWM analysis suggested rs2699887 disrupts a NF-YA binding motif at the four-nucleotide core sequence; **F** – rs2699887 differentially binds transcription factor NF-YA in vitro. Representative EMSA analysis using biotin-labelled oligonucleotides containing either rs2699887-C or rs2699887-T alleles, and a positive control (ctrl), using protein extracts of breast cancer cell line HCC1954. Competition assays were performed with one (100x) or two concentrations (10x, 100X, as indicated by the gradient symbol) of unlabeled oligonucleotides, which included consensus binding sequences for NF-YA and E2F1. Supershift assays were carried out using monoclonal antibodies against NF-YA and a negative control (HGMA).

Preferential expression of the *PIK3CA* mutated alleles is frequent in breast tumors

Changes in DNA copy number in tumors have been previously associated with changes in gene expression in *cis* [1,13–16], but whether these differences could also be due to *cis*-regulation of expression has never been analyzed. In view of the DAE we identified for *PIK3CA* in normal breast tissue, we set out to assess whether *PIK3CA* somatic mutations would have their functional effects modified by imbalances in allelic expression. We hypothesized that preferential expression of a gain-of-function mutation would have a stronger clinical impact than those occurring in lowly expressed alleles, thus generating inter-tumor clinical heterogeneity (Figure 2A). To test this, we carried mutant vs. wild-type allelic expression analysis in breast tumor samples carrying somatic *PIK3CA* missense mutations, on two independent sets of data, the METABRIC (n=94) and the TCGA (n=178) projects. A summary description of the two sets can be found in Supplementary Table 3.

Next, we calculated three allelic ratios from DNA-seq and RNA-seq data for each mutation:

$\alpha = \log_2$ (number of mutant RNA-seq reads / number of wild-type RNA-seq reads), i.e., the net mutant allele expression imbalance;

$\beta = \log_2$ (number of mutant DNA-seq reads / number of wild-type DNA-seq reads), i.e., the mutant allele relative copy-number;

$\gamma = \alpha - \beta$, i.e., the net mutant allele expression imbalance normalized for the DNA allelic copy-number imbalances, which corresponds to a putative mutant allele expression imbalance due to *cis*-regulation.

In this way, α reports on the overall allelic expression imbalance, which can be generated by different mechanisms including copy-number aberrations, cellularity differences and *cis*-regulatory variation, while γ reports specifically on the contribution from *cis*-regulatory variation (rVar in Figure 2A), including normal genetic variation, somatic noncoding mutations and allelic epigenetic changes. The distributions of the different ratios are displayed in Figure 2B.

We found that overall mutant allele expression imbalances (α ratio) are frequent in breast tumors, as seen for METABRIC (75.5%) and TCGA (68%) (Table 2). The same is true for γ ratios, at 50.4% for METABRIC and 46% for TCGA, indicating that *cis*-regulatory effects acting on mutations are also frequent in breast tumors. In both sets, we found samples with striking overall preferential allelic expression for the mutant allele (maximum 54-fold and 220-

fold in METABRIC and TCGA, respectively), but not so for the preferential expression of the wild-type allele (fold differences of 5.5 and 28.2 in METABRIC and TCGA, respectively) (Table 2, Figure 2B). Similarly, the trend of more pronounced preferential expression of the mutant allele was found for the γ ratios, 13 and 5.5-fold for METABRIC and TCGA respectively, albeit with smaller fold differences between alleles.

Table 2 - Summary of allelic expression imbalance of somatic *PIK3CA* mutations in breast tumorsc.

Variable	N	**METABRIC**, N = 94[1]	**TCGA**, N = 178[1]
alpha	272	-0.96 (-2.47, 5.75)	-0.76 (-4.82, 7.78)
alpha_group	272		
> wt		63 (67%)	96 (54%)
neutral		23 (24%)	57 (32%)
> mut		8 (8.5%)	25 (14%)
gamma	272	0.48 (-1.22, 3.70)	0.19 (-2.10, 2.46)
gamma_group	272		
> wt		6 (6.4%)	24 (13%)
neutral		47 (50%)	96 (54%)
> mut		41 (44%)	58 (33%)
MADE	272	47 (50%)	82 (46%)

[1] Statistics presented: Median (Range); n (%)

Interestingly, we observed that within the samples with mutant allele differential expression (which we denominated MADE, defined as $|\gamma| \geq 0.58$) there was a significant prevalence of samples that preferentially expressed the mutated allele in both data sets (binomial test Prob = 0.88, CI95% = [0.75 - 0.95] , p = 1.0e-07 for METABRIC and Prob = 0.70, CI95% = [0.58 - 0.80], p = 7.6e-04 for TCGA).

Cis-regulatory variants contribute significantly to imbalances in the expression of mutant alleles

Next, hypothesizing that both copy number and *cis*-regulatory variants are the major contributors to allelic expression, we set out to assess the contribution of each mechanism towards the net mutant allele expression imbalances detected in these tumors. First, we found positive correlations between overall allelic expression and both copy number and *cis*-regulatory variation (Figure 2C), albeit with an effect for the copy number approximately double the size of that found for *cis*-regulatory variation (average Pearson correlation r-square = 0.82 and 0.39, respectively). Next, considering the variance (Var) of overall allelic expression as the

sum of the effects of both mechanisms, plus the covariance (Cov) accounting for predicted non-mutually exclusion of both the mechanisms acting on any given allele:

$$\mathrm{Var}(\alpha) = \mathrm{Var}(\beta) + \mathrm{Var}(\gamma) + 2\,\mathrm{Cov}(\beta,\gamma)$$

we calculated the contribution of *cis*-regulatory variation to the variance of overall allelic expression as: $[\mathrm{Var}(\gamma) + \mathrm{Cov}(\beta,\gamma)] / \mathrm{Var}(\alpha)$. Here, we found that *cis*-regulatory variants explain 17.9% and 16.7% of the variability of overall mutant allelic expression seen in METABRIC and TCGA, respectively.

Finally, assessing how the two mechanisms act simultaneously on each tumor, we found that the majority of samples (71.28% and 54.49% for the METABRIC and TCGA, respectively) had positive γ and negative β values (Figure 2D), meaning that although the mutant allele was in lower genomic quantity, it was nevertheless preferentially expressed compared to the wild-type allele. Only a small fraction of samples displayed co-occurring preferential allelic expression and higher allele copy number of the mutant allele (5.32% and 10.11% for the METABRIC and TCGA, respectively). These results were independent of the effect of tumor cellularity (Supplementary Figure 2).

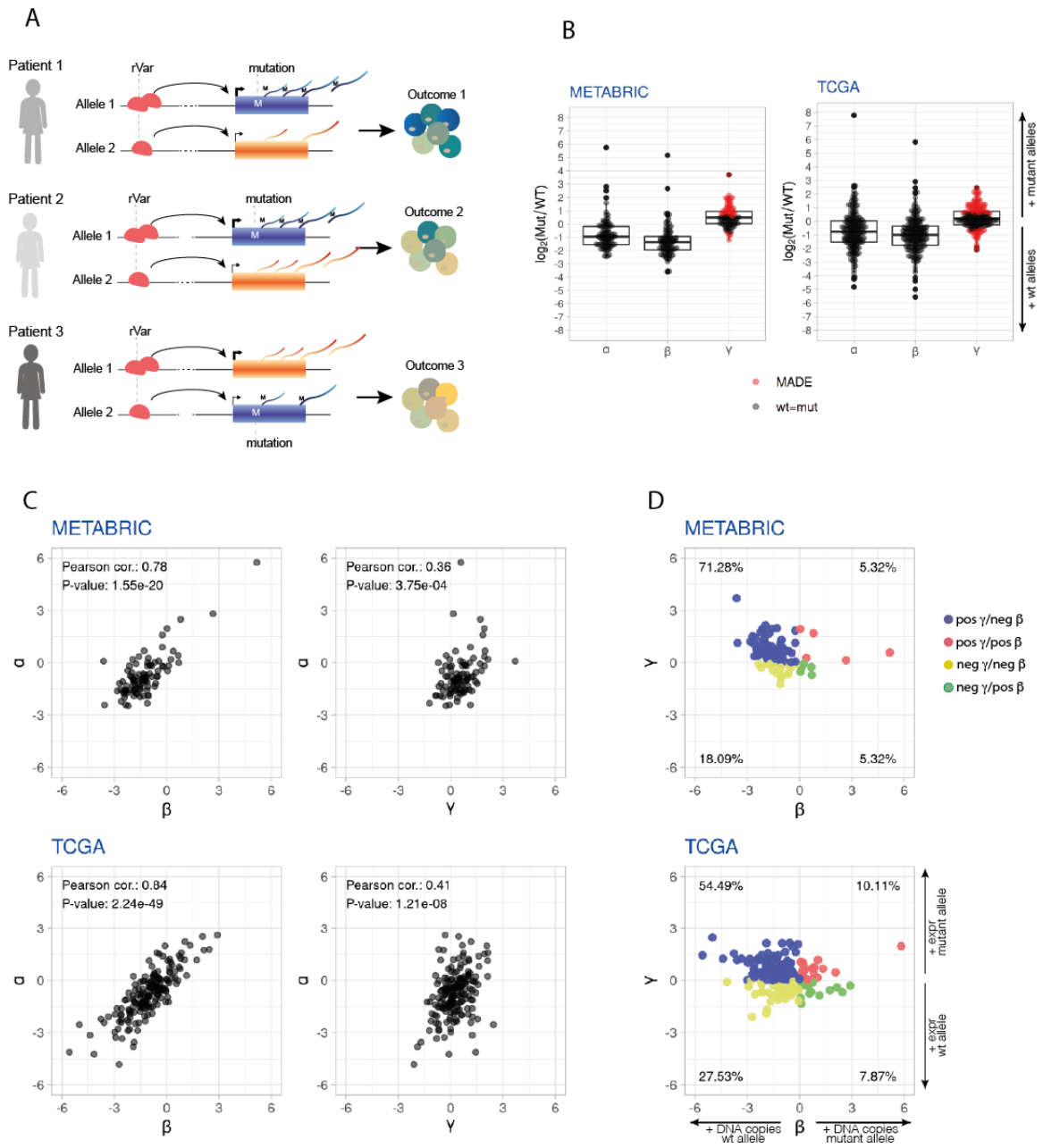

Figure 2: Mutant allelic imbalance in gene expression of somatic missense *PIK3CA* mutations is frequent in breast tumors, particularly for preferential expression of the mutant allele. **A** – Schematic representation of the hypothesis, in which *cis*-acting regulatory variants (rVar), either from germline or somatically acquired, generate different relative allelic expression of mutant and wild-type alleles, resulting in tumors of different prognosis**. B** - Distribution of α, β and γ ratios in breast tumors. Tumors with MADE, $|\gamma| \geq 0.58$, are displayed in red. **C** – Correlation analysis of α vs β and α vs γ, showing that both genomic copy-number dosage and allelic expression regulation contribute to allelic imbalances in the expression of mutations in tumors. **D** – Comparison of matched β and γ values, showing predominance of tumors with preferential allelic expression of the mutated allele, albeit higher number of wild-type allele copies.

MADE defines an aggressive subset of *PIK3CA* mutated tumors

To investigate the impact of differential *cis*-regulation of *PIK3CA*'s mutations on clinical outcome (overall and disease-specific survival), we performed univariate survival analysis with γ ratios categorized in two groups: the MADE group (when $|\gamma| \geq 0.58$) and the *mut*=*wt* group (when $|\gamma| < 0.58$). We uncovered that the MADE group had a poorer overall survival rate (p = 0.03, Supplementary Figure 3) and disease-specific survival rate (p = 0.012, Figure 3A), than the *mut*=*wt* group for METABRIC. The median overall survival for the MADE group was approximately 8 years and for the *mut*=*wt* group was 18.5 years (Figure 3A), whereas in the disease-specific analysis, 40% of MADE patients died during the length of the follow-up, in comparison with 20% deaths in the *mut*=*wt* group (Supplementary Table 4). We further found significantly shorter survival times, in overall and disease-specific survival, for the MADE samples that preferentially expressed the wild-type allele (MADE_wt group, p = 0.0073 and =0.0081, respectively) in METABRIC (Figure 3B, Supplementary Figure 3).

In the multivariate analysis of overall survival, in which a Cox model was fitted while adjusting for age and tumor characteristics, such as size, Scarff-Bloom-Richardson histological grade, clinical stage and ER, PR and Her2 statuses, the categorized γ ratios were not significantly associated with overall survival (Supplementary Figure 4). However, some of the variables that are usually independent prognosis factors, such as PR and Her2 statuses, were not significantly associated with survival either in this analysis. In the TCGA set, there was no significant difference in overall survival for either α or γ (Supplementary Figure 5), although due to the relatively shorter follow-up time of this data set (median ~1 year), and the fact that tumors were mainly Luminal A (~61.2% of samples) [17], the power to detect significant differences was smaller than compared to METABRIC.

PIK3CA MADE associates with clinicopathological variables

Next, we sought to investigate whether *PIK3CA*'s MADE was associated with known prognostic clinicopathological variables, namely hormone receptors (ER, PR) and Her2 amplification (Figure 3C, Table 3), which are directly and indirectly connected to gene expression regulation, respectively.

For METABRIC, the mean α was significantly higher in ER-negative tumors(p = 0.026), PR-negative tumors(p = 0.007) and Her2-positive tumors(p = 0.018) (Supplementary Figure 6). When evaluating the contribution of *cis*-regulatory variation to this association, we also found that higher average γ values associated with lower PR expression (p = 0.040) and Her2-positive tumors (p = 0.025), but did not find a significant association with ER expression (p = 0.129) (Figure 3C). This showed that the predominant abundance of the mutant allele was

associated with worse prognosis variables (ER negativity, PR negativity, Her2positivity), particularly in tumors with preferential mutant allele expression. The analysis of the TCGA set corroborated the significant association between the PR status and γ ratios (p = 0.003) (Figure 3C, Supplementary Figure 6). However, it showed an association between ER status and γ values only (p = 0.014), albeit not with Her2 status.

Table 3 - Statistical analysis between normalized and raw mutant allele expression ratios. P-values indicated correspond to two-sided Wilcoxon rank sum tests with continuity correction, adjusted per study with the Bonferroni method.

	METABRIC			TCGA		
	n (pos/neg)	α	γ	n (pos/neg)	α	γ
Clinical Features						
ERα status	78/16	0.026	0.129	147/16	0.157	0.014
PR status	67/27	0.007	0.040	138/25	0.054	0.003
Her2 status	15/79	0.018	0.025	23/121	1.000	0.453

Legend: ER = estrogen; PR = progesterone; Her2 = human epidermal growth factor 2 receptor

In view of these results, we took MADE into consideration in the survival analysis within the expression subgroups of ER, PR and Her2. Both MADE vs. *mut*=*wt* groups did not reveal significant differences in overall and disease-specific survival, although there was a trend for worse survival for the MADE category in the Her2 negative subgroup (p = 0.056 and 0.027 in overall and disease-specific analyses respectively) (Supplementary Figure 3C). However, when further stratifying the MADE groups, a poorer prognosis for MADE_wt samples was identified in ER positive (p = 0.004), PR positive (p = 0.0013) and Her2 negative tumors(p = 0. 0066) (Figure 3D). Thus, MADE defines an aggressive subgroup within the usually good prognosis ER-positive, PR-positive, and Her2-negative subsets of tumors.

Finally, the association of MADE with other know prognostic variables including lymph node status, tumor size, grade and molecular subtypes, namely PAM50 [18] and IntClust [16], revealed a significant association between γ ratios and PAM50 subtypes only in TCGA (p = 0.002) (Supplementary Figure 7, Supplementary Table 5).

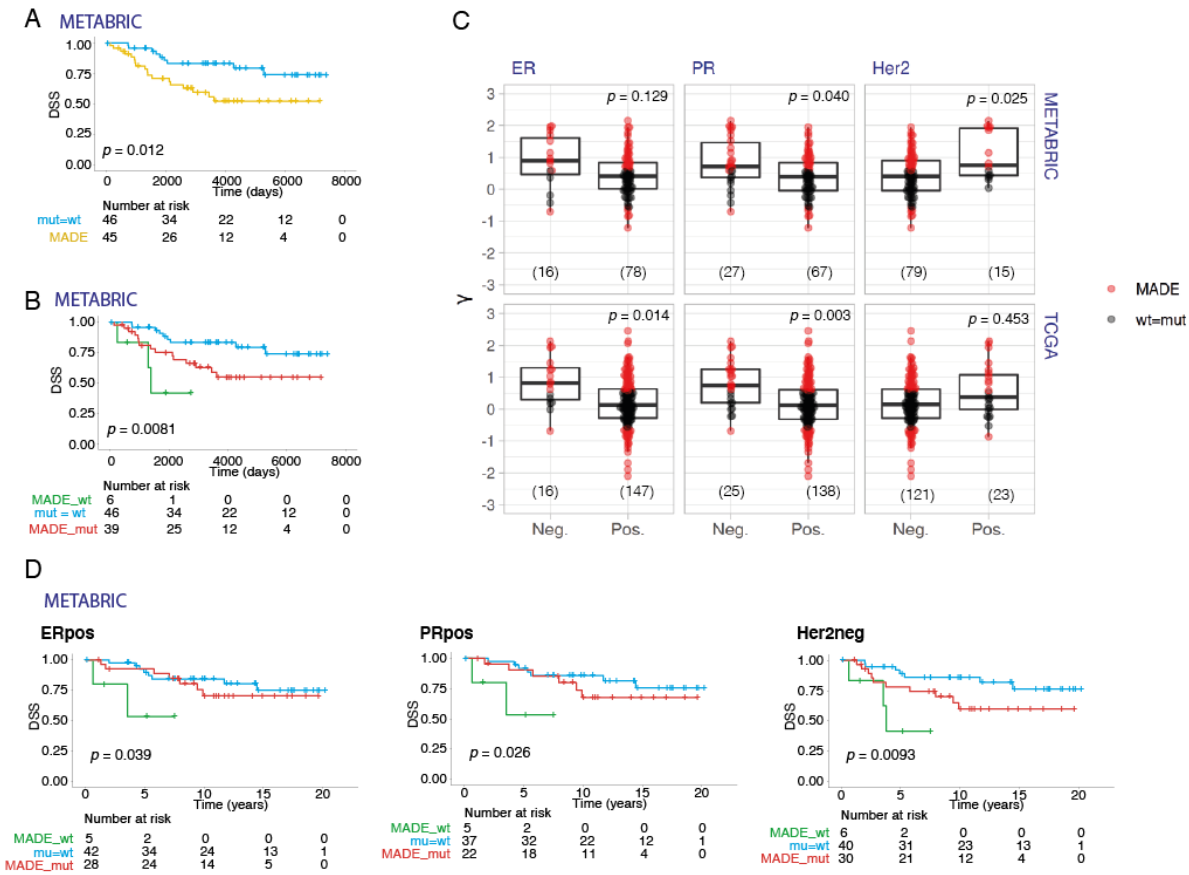

Figure 3: Allelic preferential expression of PIK3CA mutations is associated with survival and clinicopathological parameters in breast cancer. **A –** Kaplan–Meier curves of disease-specific survival showing the worse prognosis of patients with differential expression of the *PIK3CA* mutation (MADE group, shown in yellow) compared to those expressing equimolar levels of mutation and wild-type alleles (*mut=wt* group, shown in blue), in METABRIC. Shown below the graph are the number of patients at risk per group throughout time. **B –** Disease-Specific survival Kaplan-Meier curves with the MADE group subdivided into those with preferential expression of the mutant allele (MADE_mut, shown in red) and those with preferential expression of the wild-type allele (MADE_wt, shown in green), confirming worse survival compared to the *mut=wt* group (shown in blue), in METABRIC. Shown below the graph are the number of patients at risk per group throughout time. **C** – Preferential expression of the mutated allele is associated with ER negative, PR negative and Her2 positive breast tumors. In all graphs, samples were colored according to the MADE classification. P-values indicated correspond to Wilcoxon rank sum test with continuity correction, corrected for multiple testing using the Bonferroni correction. **D –** Kaplan-Meier curves showing the worse disease-specific survival of the MADE groups compared with the *mut=wt* group, in the ER positive, PR positive and Her2 negative subtypes (ER = estrogen; PR = progesterone; Her2 = human epidermal growth factor 2 receptor) in METABRIC.

Discussion

Our work reveals the role of *cis*-regulatory variation acting on *PIK3CA* somatic mutations as modifiers of mutation penetrance. As far as we know, we show for the first time that allelic expression imbalance between *PIK3CA's* mutant and wild-type alleles is common and prognostic in breast cancer. Particularly, preferential expression of the mutant allele is significantly more common than expected, and whether there is positive selection for it in tumor evolution should be further investigated. Yet, genomic allelic imbalance remains the determinant of allelic expression dosage with larger effect, with *cis*-regulatory variation being responsible for ~18% of the variability in overall allelic expression in tumors. Our findings are supported by analysis of RNA-seq data from two independent cohorts of tumor samples, the METABRIC and TCGA projects.

Moreover, we show that preferential expression of the mutant allele is associated with poor prognosis variables, such as PR [19] negative and Her2-positive [20] statuses. In the METABRIC dataset, we also found that patients with tumors with mutant allelic expression imbalance had worse overall survival than those without it, particularly in the ER positive, PR positive and Her2 negative subsets. Interestingly, the worst outcome is observed in the group with preferential expression of the wild-type allele. In these tumors, *PIK3CA* mutations are lowly expressed and are possibly passenger events, hence the poor prognosis might be due to other mutations, or combinations thereof. Albeit the significant findings, one caveat of our study is that this subgroup of samples is very small, and identification of what other events might be contributing to the worse survival requires larger sample number and combinatorial analysis with additional data on other mutations.

Besides the potential use of our findings as a prognosis biomarker in the clinic, these results may also have therapeutic implications. Some of the major clinical challenges in cancer treatment are the identification of biomarkers of prognosis and defining which patients will benefit from a given therapy. In particular, the ability to identify patients unlikely to respond to specific therapies is crucial to prevent unnecessary drug cytotoxicity without any therapeutic benefits. Our results have revealed the importance of considering allelic expression in the assessment of somatic mutations and how it can help in these two aspects of patient management. In spite of the high frequency of *PIK3CA* mutation in breast cancers, the response to PI3K inhibitor therapy has been unsuccessful as expected, and the prognostic significance of detecting somatic *PIK3CA* mutations in breast tumor is unclear [21]. Relevant to this discussion, we have previously shown that presence of *PIK3CA* mutations confer a poorer prognosis in patients with ER positive breast cancer only when stratified into copy number driven subgroups (IntClust 1+,2+,9+) [22]. In the present study we provide new evidence for prognostic significance of these mutations at the level of expression in unstratified

ER positive tumors. This has potential impact on therapy response and clinical management, since one may hypothesize that in the cases that do not express the targetable mutation little to no benefit would come from treatment. This should be further evaluated by assessing the allelic expression of mutant oncogenes in the tumors of patients enrolled in molecular-driven trials.

More challenging is determining which *cis*-regulatory mechanisms are promoting allelic expression imbalances. It is known that both inherited [10,23] and acquired variants [6,24–26] can affect gene expression in an allelic manner, and also that epigenetics can play a role [27,28]. Here we show that *PIK3CA's* expression is controlled by normal *cis*-regulatory variation in normal breast tissue, and our data suggests that the SNP rs2699887 is possibly the main regulatory variant. We believe the mechanism includes modification of an NF-YA transcription factor binding site at the promoter of *PIK3CA*, supported by *in-silico* and *in-vitro* data. We also found the heterozygous genotype to be associated with higher expression of *PIK3CA* gene, compared to the common homozygotes. Although there is published data supporting the clinical association of rs2699887 with poor prognosis in other cancers [29,30], linked to an increase in PI3K signaling, there is still some data supporting the opposite association [31,32]. We did not find an association between rs2699887 and survival, which suggests that other mechanisms besides normal *cis*-regulatory variation are contributing to the preferential allelic expression in these tumors. Double *PIK3CA* mutations in the same allele are frequent in breast tumors [33], and the impact of non-coding mutations in cancer is just starting to be explored [6]. So, a strong possibility is that the combination of non-coding and coding mutations in a same gene might be underlying the allelic expression imbalances we are detecting.

Further studies on allelic expression imbalances of activating mutations, and even inactivating ones, should further reveal the contribution of *cis*-regulatory mechanisms in tumor development and progression. Particularly interesting to study are the questions of whether the mutation is acquired in an allele predisposed with higher expression, or whether there is a sequence of somatic events that introduce the coding activating mutation and additional *cis*-regulatory non-coding mutations. The answers could have important repercussions in our understanding of tumor evolution.

Conclusion

In summary, we show that differential expression between the mutant and wild-type alleles of *PIK3CA* is common in breast cancer, and mostly due to allele-specific cis-regulatory

effects. We further show that mutant allele differential expression is associated with clinical parameters such as ER, PR and Her-2 statuses, and is prognostically significant. Collectively, our work establishes the prognostic relevance of transcriptional allelic regulation of somatic mutations. As such, it supports a shift in the mutation testing in patient management, where the level of expression of mutations should be considered, besides the detection at the DNA level.

Methods

Differential allelic expression analysis

DNA and total RNA from 64 samples of normal breast tissue, were hybridized onto Illumina Exon510S-Duo arrays (humanexon510s-duo) and data were analyzed as described before [11]. In short, after sample filtering and normalization, SNPs with RNA log2 allelic intensity values on average greater than 9.5 and heterozygous in five or more samples. were kept for further analysis. Allelic log-ratios were calculated for RNA and DNA intensity data: log-ratio = $\log_2 A - \log_2 B$, for alleles A and B. A two-sample Student's t-test was applied to compare RNA log-ratios between heterozygous (AB) and homozygous groups (AA and BB). Only SNPs with p-values lower than 0.05 for all comparisons were further analyzed. Allelic expression (AE) ratios were normalized for allelic DNA content (RNA log-ratio – DNA log-ratio), and differential allelic expression (DAE) was called when |AE ratio| ≥ 0.58 (1.5-fold or greater between alleles). This threshold was established based on the sensitivity and specificity of the applied DAE detection method. Linkage disequilibrium (LD) between daeSNPs was evaluated using the genetic variant-centered annotation browser SNiPA [34].

Genotype imputation analysis on normal breast tissue samples

Genotype imputation was run on the Illumina Exon 510 Duo germline genotype data from the 64 samples that passed microarrays quality control filters. Prior to imputation, a quality control was applied to the genotyping data and SNPs with call rates < 85%, minor allele frequency < 0.01, and Hardy-Weinberg equilibrium with $p = < 1.0e^{-05}$ were excluded from the analysis. Genotype data from the chromosome 3 Illumina SNPs that passed quality control was used to impute genotypes at all additional known SNPs in the chromosome using MACH1.0 [35]. The phased haplotypes for HapMap3 release (HapMap3 NCBI Build 36, CEU panel - Utah residents with Northern and Western European ancestry) were used as the reference panel. For imputation with MaCH1.0, we applied the recommended two step

imputation process: model parameters (crossover and error rates) were estimated prior to imputation using all haplotypes from the study subjects and running 100 Hidden Markov Model (HMM) iterations with the command options: - greedy and -r 100. Genotype imputation was then carried out using the model parameter estimates from the previous round with options: -greedy, -mle, and -mldetails. Imputation results were assessed by the platform-specific measures of imputation uncertainty for each SNP, given by the rq score. Results were filtered for an rq score higher than 0.3 [35].

Differential allelic expression (DAE) mapping analysis on normal breast tissue samples

Differential allelic expression mapping analysis was performed by stratifying AE ratios at each *PIK3CA* daeSNP according to the genotype at the genotyped/imputed SNPs located within ±250Kb of the daeSNP. A two-sample t-test was applied to assess differences between the mean AE ratio between the heterozygous group samples and the combined homozygous groups. P-values were corrected for multiple testing using 1000 permutations. Permutation-corrected p-values were considered significant below 0.05 and when on average the heterozygous samples displayed larger fold differences between alleles when compared to homozygous samples.

Proxy SNPs were identified using HaploReg v4.1 available at https://pubs.broadinstitute.org/mammals/haploreg/haploreg.php, using genotype data from the Phase 1 of the 1000G project, for the CEU population, with a LD threshold of $r^2 \geq 0.8$ [36,37].

Functional annotation of associated variants

To predict the most likely functional variants, the ones associated with AE levels were mapped to epigenetic marks derived from the Encyclopedia of DNA Elements (ENCODE) and NIH Roadmap Epigenomics projects, such as chromatin states (chromHMM) annotation, regions of *DNase* I hypersensitivity, transcription factor binding sites and histone modifications of epigenetic markers (H3K4Me1, H3K4Me3 and H3K27Ac) (http://genome.ucsc.edu/ENCODE/) for normal human mammary epithelial cells (HMECs), human mammary fibroblasts (HMFs), BR.MYO (breast myoepithelial cells) and BR.H35 (breast vHMEC) and for two breast cancer cell lines MCF-7 and T47D. We prioritized variants that were located in either active promoter or enhancer regions in mammary cell lines, and for which ChIP-Seq data indicated protein binding or position weight matrix (PWM) scores predicted differential protein binding for different alleles. Two publicly available tools, RegulomeDB and HaploReg v4.1 and the MotifBreakR Bioconductor package were also used to evaluate those candidate functional variants [36–38].

Electrophoretic mobility shift assay (EMSA)

MCF-7 (ER positive) and HCC1954 (ER-negative) breast cancer cell lines were cultured in DMEM and RPMI culture media, respectively, supplemented with 10% FBS and 1% PS (penicillin and streptomycin). Nuclear protein extracts were prepared using the Thermo Scientific Pierce™ NER kit, according to the manufacturer's instructions. Oligonucleotide sequences corresponding to the C (common) and T (minor) alleles of rs2699887 (5'-AGCGTGAGTAGAGCGCGGA[C/T]TGGCCGGTAGCGGGTGCGGTG-3') were labelled using the Thermo Scientific Pierce Biotin 3' End DNA Labelling Kit according to the manufacturer's instructions. Antibodies used for super-shift competition assays were NF-YA (H-209) (Santa Cruz Bioechnology, SC-10779X), and HMGA1a/HMGA1b (Abcam, ab4078). EMSA experiments were performed using the Thermo Scientific LightShift™ Chemiluminescent EMSA Kit, using the buffer and binding reactions conditions previously described [39]. Each EMSA was repeated at least twice for all combinations of cell extract and oligonucleotide, which were also tested in serial dilution amounts.

Breast tumor samples

The METABRIC dataset of tumor samples included 2433 samples from the METABRIC project [16] with DNA sequencing data, among which 480 were subjected to a capture-based RNA sequencing study. Sequencing libraries were generated as previously described. In brief, sequencing libraries using total RNA generated from frozen tissues with a TruSeq mRNA Library Preparation Kit using poly-A-enriched RNA (Illumina, San Diego, CA, USA) and enriched with the human kinome DNA capture baits (Agilent Technologies, Santa Clara, CA, USA). Six libraries were pooled for each capture reaction, with 100 ng of each library and sequenced (paired-end 51bp) on an Illumina HiSeq2000 platform. A subset of samples that have both DNA and RNA sequencing data and harbor *PIK3CA* mutations was selected for further analysis.

The TCGA dataset comprised of 695 samples from TCGA breast cancers [40], from which a subset of 289 samples with *PIK3CA* mutations was selected for further analysis. Demographic features and disease characteristics of the two datasets are summarized in Supplementary Table 2.

DNA-seq and RNA-seq variant calling in tumors

<u>Alignment & Preprocessing</u>

Sequence data (FASTQ) mapped to reference genome (hg19) were aligned using STAR v2.4.1 [41]. A two-pass alignment was carried out: splice junctions detected in the first

alignment run are used to guide the final alignment. Duplicates were marked with Picard v1.131 (http://picard.sourceforge.net). Indel realignment and base quality score recalibration steps were performed using the Genome Analysis Toolkit (GATK).

Variant Calling and Annotations

SNV and indel variants were called using GATK Haplotype Caller. Variants were filtered by applying hard filters using GATK VariantFiltration [42]. Variants were annotated with Ensembl Variant Effect Predictor (VEP) [43]

Heterozygous genotypes were called from DNA data to avoid RNA editing, and other RNA related variants, as true allelic imbalance can lead to heterozygous sites being called homozygous in RNA-based genotype calling.

Statistical analysis of allelic expression imbalances in tumors

Prior to the analysis, a set of filtering steps were performed to select samples which: 1) passed quality control; 2) harbored missense mutations; and 3) and had a minimum of 30 reads for both RNA and DNA data. Clinical data for METABRIC was updated from the original studies with the latest available records. Clinical data for TCGA were imported from https://portal.gdc.cancer.gov/ in 26/11/2018.

For all samples three parameters were calculated: 1) The DNA mutant allele ratio β = Log_2[(mutant allele read count in DNA)/(wild-type allele read count in DNA)], which served to control for sequencing artefacts from heterozygous genotypes and to account for differences in variant frequencies in DNA; 2) α = Log_2[(mutant allele read count in RNA)/(wild-type allele read count in RNA)], that served as a measure of the net allelic expression imbalance in tumors; and 3) the normalized mutant allele expression ratio $\gamma = \alpha - \beta$, which is a measure of mutant allelic expression imbalance due to *cis*-regulation alone.

Association between allelic expression imbalance ratios and clinical data was achieved by bivariate analysis Wilcoxon rank sum test with continuity correction or Kruskal-Wallis rank sum test, as indicated in tables and figures. P-values were adjusted per study using the Bonferroni correction and were considered significant when ≤ 0.05. Correlation analysis “alpha vs. beta” and “alpha vs. gamma” ratios for both sets of samples were performed using a Pearson’s test. All statistical analysis and data visualization were performed using R.

Additionally, we considered that there was mutant allele differential expression (MADE) when $|\gamma| \geq 0.58$ (1.5-fold or greater difference). In the survival analysis, we further separated the MADE samples into MADE_wt ($\gamma \leq -0.58$, or preferential expression of wild-type allele) and MADE_mut ($\gamma \geq 0.58$, or preferential expression of mutant allele).

Survival Analyses

Kaplan-Meier plots and multivariate Cox proportional hazard models were used to examine the association between alpha and gamma allelic expression ratios and survival using the *survival* package from R [44,45]. Death due to all causes was used as the endpoint, and all subjects still alive were censored at the date of the last contact. Kaplan-Meier survival curves were compared using the log-rank test.

For the multivariate analysis, Cox proportional hazard model was used to assess the effect of γ on the overall survival. Hazard ratios (HRs) and 95% confidence intervals (CI) were estimated by fitting the Cox model while adjusting for age and tumor characteristics, such as size, Scarff-Bloom-Richardson histological grade, clinical stage and estrogen receptor (ER), progesterone (PR) and human epidermal growth factor 2 (Her2) statuses. All t-tests were two-sided, and $p \leq 0.05$ were considered statistically significant.

Declarations

Ethics approval and consent to participate

Normal breast and tumor samples were obtained with the consent from donors and appropriate approval from local ethical committees, for which detailed information can be found in the respective original publications: normal tissue [8], METABRIC [16], TCGA [40].

Availability of Data and Code

Microarray raw data is deposited in the Gene Expression Omnibus under accession number GSE35023. Primary data (BAM files) for DNAseq are deposited at the European Genome-phenome Archive (EGA) under study accession number EGAS00001001753, and may be downloaded upon request and authorization by the METABRIC Data Access Committee. Primary data (BAM files) for RNAseq is currently being deposited in EGA (accession number to be determined). Primary data (BAM files) for DNAseq and RNAseq from TCGA are deposited in the database of Genotypes and Phenotypes (dbGaP) under the study accession number phs000178. The filtered data and code for the analysis of mutant allele expression imbalances and for the survival analysis can be publicly accessed at https://github.com/maialab/pik3ca.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported by national Portuguese funding through FCT-Fundação para a Ciência e a Tecnologia, and CRESC ALGARVE 2020, institutional support ALG-01-0145-FEDER-31477 – “DevoCancer”, CBMR - UID/BIM/04773/2013, POCI-01-0145-FEDER-022184 - ”GenomePT”, , the contract DL 57/2016/CP1361/CT0042 (J.M.X.) and individual fellowships SFRH/BPD/99502/2014 (J.M.X.) and PD/BD/114252/2016 (F.E.). Funding was also received from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme FP7/2007-2013/ under REA grant agreement n° 303745 (ATM), and a Maratona da Saúde Award (ATM).

The METABRIC project was funded by Cancer Research UK, the British Columbia Cancer Foundation and Canadian Breast Cancer Foundation BC/Yukon. This sequencing project was funded by CRUK grant C507/A16278 and CRUK core grant A16942. The authors also acknowledge the support of the University of Cambridge, Hutchinson Whampoa, the NIHR Cambridge Biomedical Research Centre, the Cambridge Experimental Cancer Medicine Centre, the Centre for Translational Genomics (CTAG) Vancouver and the BCCA Breast Cancer Outcomes Unit. We thank the Genomics, Histopathology, and Biorepository Core Facilities at the Cancer Research UK Cambridge Institute, and the Addenbrooke’s Human Research Tissue Bank (supported by the National Institute for Health Research Cambridge Biomedical Research Centre).

Authors' Contributions

L.C., J.M.X., S.F.C. and A.T.M. wrote the manuscript.

J.M.X., S.F.C., A.T.M. contributed to the overall design of this study.

Data were collected by F.E., A.B., L.M., R.B., C.C., S.F.C., A.T.M.

Data were analyzed and interpreted by L.C., J.M.X., R.M., B.P.A., F.E., C.S., I.D., M.E, A.M., I.A.S., J.S. and A.T.M.

All authors have read and approved the final version of the manuscript.

Acknowledgements

We thank all the patients who donated tissue and the associated pseudo-anonymised clinical data for this project. The authors would also like to thank the members of the Functional

Genomics of Cancer group at CBMR for helpful discussions, and Vitor Morais at UAIC for administrative support.

References

1. Bielski CM, Donoghue MTA, Gadiya M, Hanrahan AJ, Won HH, Chang MT, et al. Widespread Selection for Oncogenic Mutant Allele Imbalance in Cancer. Cancer Cell. 2018;34:852-862.e4.

2. Pastinen T, Hudson TJ. Cis-Acting Regulatory Variation in the Human Genome. Science. 2004;306:647–50.

3. Ge B, Pokholok DK, Kwan T, Grundberg E, Morcos L, Verlaan DJ, et al. Global patterns of cis variation in human cells revealed by high-density allelic expression analysis. Nat Genet. 2009;41:1216–22.

4. Pastinen T, Sladek R, Gurd S, Sammak A, Ge B, Lepage P, et al. A survey of genetic and epigenetic variation affecting human gene expression. Physiol Genomics. 2004;16:184–93.

5. Morley M, Molony CM, Weber TM, Devlin JL, Ewens KG, Spielman RS, et al. Genetic analysis of genome-wide variation in human gene expression. Nature. 2004;430:743–7.

6. Calabrese C, Davidson NR, Demircioğlu D, Fonseca NA, He Y, Kahles A, et al. Genomic basis for RNA alterations in cancer. Nature. 2020;578:129–36.

7. Rhee J-K, Lee S, Park W-Y, Kim Y-H, Kim T-M. Allelic imbalance of somatic mutations in cancer genomes and transcriptomes. Sci Rep-uk. 2017;7:1653.

8. Maia A-T, Spiteri I, Lee AJ, O'Reilly M, Jones L, Caldas C, et al. Extent of differential allelic expression of candidate breast cancer genes is similar in blood and breast. Breast Cancer Res. 2009;11:R88.

9. Cox DG, Simard J, Sinnett D, Hamdi Y, Soucy P, Ouimet M, et al. Common variants of the BRCA1 wild-type allele modify the risk of breast cancer in BRCA1 mutation carriers. Hum Mol Genet. 2011;20:4732–47.

10. Maia A-T, Antoniou AC, O'Reilly M, Samarajiwa S, Dunning M, Kartsonaki C, et al. Effects of BRCA2 cis-regulation in normal breast and cancer risk amongst BRCA2 mutation carriers. Breast Cancer Res. 2012;14:R63.

11. Liu R, Maia A-T, Russell R, Caldas C, Ponder BA, Ritchie ME. Allele-specific expression analysis methods for high-density SNP microarray data. Bioinformatics. 2012;28:1102–8.

12. Xiao R, Scott LJ. Detection of cis-acting regulatory SNPs using allelic expression data. Genet Epidemiol. 2011;35:515–25.

13. Hartman DJ, Davison JM, Foxwell TJ, Nikiforova MN, Chiosea SI. Mutant allele-specific imbalance modulates prognostic impact of KRAS mutations in colorectal adenocarcinoma and is associated with worse overall survival. Int J Cancer. 2012;131:1810–7.

14. Soh J, Okumura N, Lockwood WW, Yamamoto H, Shigematsu H, Zhang W, et al. Oncogene Mutations, Copy Number Gains and Mutant Allele Specific Imbalance (MASI) Frequently Occur Together in Tumor Cells. Plos One. 2009;4:e7464.

15. Krasinskas AM, Moser AJ, Saka B, Adsay NV, Chiosea SI. KRAS mutant allele-specific imbalance is associated with worse prognosis in pancreatic cancer and progression to undifferentiated carcinoma of the pancreas. Modern Pathol. 2013;26:1346–54.

16. Curtis C, Shah SP, Chin S-F, Turashvili G, Rueda OM, Dunning MJ, et al. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. Nature. 2012;486:346–52.

17. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD, et al. An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. Cell. 2018;173:400-416.e11.

18. Sørlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, et al. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc National Acad Sci. 2001;98:10869–74.

19. Dunnwald LK, Rossing MA, Li CI. Hormone receptor status, tumor characteristics, and prognosis: a prospective cohort of breast cancer patients. Breast Cancer Res. 2007;9:R6.

20. Chia S, Norris B, Speers C, Cheang M, Gilks B, Gown AM, et al. Human Epidermal Growth Factor Receptor 2 Overexpression As a Prognostic Factor in a Large Tissue Microarray Series of Node-Negative Breast Cancers. J Clin Oncol. 2008;26:5697–704.

21. Keegan NM, Gleeson JP, Hennessy BT, Morris PG. PI3K inhibition to overcome endocrine resistance in breast cancer. Expert Opin Inv Drug. 2018;27:1–15.

22. Pereira B, Chin S-F, Rueda OM, Vollan H-KM, Provenzano E, Bardwell HA, et al. The somatic mutation profiles of 2,433 breast cancers refine their genomic and transcriptomic landscapes. Nat Commun. 2016;7:11479.

23. Yan H, Yuan W, Velculescu VE, Vogelstein B, Kinzler KW. Allelic Variation in Human Gene Expression. Science. 2002;297:1143–1143.

24. Przytycki PF, Singh M. Differential Allele-Specific Expression Uncovers Breast Cancer Genes Dysregulated by Cis Noncoding Mutations. Cell Syst. 2020;10:193-203.e4.

25. Huang FW, Hodis E, Xu MJ, Kryukov GV, Chin L, Garraway LA. Highly Recurrent *TERT* Promoter Mutations in Human Melanoma. Science. 2013;339:957–9.

26. Mansour MR, Abraham BJ, Anders L, Berezovskaya A, Gutierrez A, Durbin AD, et al. An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. Science. 2014;346:1373–7.

27. Shoemaker R, Deng J, Wang W, Zhang K. Allele-specific methylation is prevalent and is contributed by CpG-SNPs in the human genome. Genome Res. 2010;20:883–9.

28. Ongen H, Andersen CL, Bramsen JB, Oster B, Rasmussen MH, Ferreira PG, et al. Putative cis-regulatory drivers in colorectal cancer. Nature. 2014;512:87–90.

29. Morgese F, Soldato D, Pagliaretta S, Giampieri R, Brancorsini D, Torniai M, et al. Impact of phosphoinositide-3-kinase and vitamin D3 nuclear receptor single-nucleotide polymorphisms on the outcome of malignant melanoma patients. Oncotarget. 2014;5:75914–23.

30. Li Q, Yang J, Yu Q, Wu H, Liu B, Xiong H, et al. Associations between Single-Nucleotide Polymorphisms in the PI3K–PTEN–AKT–mTOR Pathway and Increased Risk of Brain Metastasis in Patients with Non–Small Cell Lung Cancer. Clin Cancer Res. 2013;19:6252–60.

31. Wang L-E, Ma H, Hale KS, Yin M, Meyer LA, Liu H, et al. Roles of genetic variants in the PI3K and RAS/RAF pathways in susceptibility to endometrial cancer and clinical outcomes. J Cancer Res Clin. 2012;138:377–85.

32. Pu X, Hildebrandt MAT, Lu C, Lin J, Stewart DJ, Ye Y, et al. PI3K/PTEN/AKT/mTOR pathway genetic variation predicts toxicity and distant progression in lung cancer patients receiving platinum-based chemotherapy. Lung Cancer. 2011;71:82–8.

33. Vasan N, Razavi P, Johnson JL, Shao H, Shah H, Antoine A, et al. Double PIK3CA mutations in cis increase oncogenicity and sensitivity to PI3Kα inhibitors. Science. 2019;366:714–23.

34. Arnold M, Raffler J, Pfeufer A, Suhre K, Kastenmüller G. SNiPA: an interactive, genetic variant-centered annotation browser. Bioinformatics. 2015;31:1334–6.

35. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR. MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. Genet Epidemiol. 2010;34:816–34.

36. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. Nucleic Acids Res. 2012;40:D930–4.

37. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. A global reference for human genetic variation. Nature. 2015;526:68–74.

38. Coetzee SG, Coetzee GA, Hazelett DJ. motifbreakR: an R/Bioconductor package for predicting variant effects at transcription factor binding sites. Bioinformatics. 2015;31:3847–9.

39. Meyer KB, Maia A-T, O'Reilly M, Teschendorff AE, Chin S-F, Caldas C, et al. Allele-Specific Up-Regulation of FGFR2 Increases Susceptibility to Breast Cancer. Plos Biol. 2008;6:e108.

40. Wilkerson MD, Cabanski CR, Sun W, Hoadley KA, Walter V, Mose LE, et al. Integrated RNA and DNA sequencing improves mutation detection in low purity tumors. Nucleic Acids Res. 2014;42:e107–e107.

41. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29:15–21.

42. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20:1297–303.

43. McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, et al. The Ensembl Variant Effect Predictor. Genome Biol. 2016;17:122.

44. Therneau TM, Grambsch PM. Modeling Survival Data: Extending the Cox Model. Statistics Biology Heal. 2000;

45. Therneau. Package "survial" [Internet]. 2014. Available from: tbrieder.org