European Journal of Human Genetics (2017) 25, 384–387

www.nature.com/ejhg

SHORT REPORT

SNP variants at the *MAP3K1/SETD9* locus 5q11.2 associate with somatic *PIK3CA* variants in breast cancers

Roberto Puzone*,[1] and Ulrich Pfeffer[2]

Genome-wide association studies have revealed many breast cancer (BC) risk-associated genetic variants that might functionally interact with other molecular determinants of BC. We analysed the association of 21 known risk-associated single-nucleotide variants (SNVs) with recurrent somatic variants in two cohorts of 77 and 754 oestrogen receptor α-positive BCs. Four SNVs located at 5q11.2 were found to be associated with the somatic PIK3CA variant status in the pilot cohort of 77 cases with odds ratio (OR) up to 6.5 indicating strong effects, and were selected for the validation phase. Two of these SNVs, rs252913 and rs331499, located in the MAP3K1/SETD9 gene boundary, were confirmed to be associated with somatic PIK3CA variants in the large cohort with OR 2.97 (1.17–7.75) and 1.76 (1.11–2.77), respectively, notably higher than their BC risk-associated values, both around 1.1. In the presence of the SNV or of somatic PIK3CA variants, cancers express significantly elevated levels of MAP3K1 and SETD9, with synergy of SNV and PIK3CA variants in *MAP3K1* gene overexpression, consistent with a preferential PIK3CA-dependent regulation of the variant alleles.
European Journal of Human Genetics (2017) **25,** 384–387; doi:10.1038/ejhg.2016.179; published online 28 December 2016

INTRODUCTION

Less than 10% of human breast cancers (BCs) show pronounced familiarity that can be explained by high penetrance germline variants, but sporadic BCs are also co-determined by low penetrance variants.[1–3] Genome-wide association studies in BC cohorts compared with the general population have identified almost a 100 single-nucleotide variants (SNVs) associated with BC with odds ratio (OR) below 1.2 for most single SNVs.[4,5] However, differently from high penetrance variants, how SNVs functionally increase BC risk is difficult to establish and mostly unknown.

In sporadic BC, next-generation sequencing has revealed a mutational landscape characterised by a large number of somatic variants (SM), but few recurrently mutated genes: *PIK3CA* (25–35%), *TP53* (20–30%), and to a lesser extent *MAP3K1*, *GATA3* and *CDH1*, with a certain preference for specific BC (sub)types.[6–8] Recurrent somatic variants drive tumour progression, but little is known about what leads to the development of tumours carrying these variants. We reasoned that risk-associated genetic variants might modify driver gene penetrance, and investigated the issue by analysing the association of a small series of known BC risk-associated SNVs with the occurrence of SM in the frequently mutated genes *PIK3CA*, *TP53* and *MAP3K1*.

METHODS

Twenty-one SNVs were selected (Supplementary Table S1): 11 from O'Brien *et al.*,[9] rs889312, an expression quantitative trait locus (eQTL) SNV,[10,11] and three further SNVs in high linkage disequilibrium (LD) with the former all located at 5q11.2 (Supplementary Table S2), as well as seven further SNVs were selected from Rhie *et al.*[12] We preferred SNVs associated with oestrogen receptor α-positive (ER+) BC risk (often higher than for all BC) and close to coding sequences, to maximise the use of genotyped data without imputations of allelic variant status.

Because no previous OR assessment was available, we started by analysing a small pilot data set powered enough to observe strong association, Ellis *et al.*,[13] whole-genome and exome sequences of samples from 77 ER+ BC patients of CEU ethnicity (genomic data from dbGap[14] repository http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000472.v1.p1, authorisation #7444; SM data from Ellis *et al.*[13]). To confirm the observed associations, we used the 754 ER+ BC patient data from the BRCA data set of the TCGA consortium collection[7] (genomic data http://www.ncbi.nlm.nih.gov/projects/gap/cgi-bin/study.cgi?study_id=phs000178.v9.p8, available from CGHub[15] repository, authorisation #7821; public data at https://tcga-data.nci.nih.gov, and many research portals such as CBIO (http://www.cbioportal.org/)[16]). Clinical and protected SRR/BAM files of normal blood samples, collecting selected sequences that included the 21 SNVs for the Ellis 2012 data set or the four 5q11.2 SNVs for the BRCA data set, were downloaded by using dbGap's SRA-Toolkit service for Elllis 2012 and beta2 version of CGHub's BAM-Slicer service for TCGA-BRCA. Selected sequence downloading allowed a crucial reduction of the multi-terabytes secure file handling required for full data sets; to our knowledge, this is the first report of an analysis entirely performed by using selected downloads. Data were processed by Python and R scripts for realignments, filtering and allele calls, then linked to public (level 3) SM data of BC samples. For the Ellis 2012 pilot data set, ORs were calculated for each allele (dominant model) of the 21 SNVs *vs* the SM status of each of *PIK3CA*, *MAP3k1* and *TP53*. Uncorrected Fisher's exact test $P<0.1$ and OR >3 were used to select the resulting associations. Allele correlations were assessed by the Pearson's r coefficient. For the BRCA data set, ER+ samples were selected and a similar analysis was performed on the four 5q11.2 SNVs *vs PIK3CA* SM only. Fisher's P were reported without multiple test correction owing to the high correlations among the four SNVs.[17] Logistic regression and ethnicity strata calculation with forest plots and Breslow–Day homogeneity test were also

[1]Integrated Oncology Therapies Department, Clinical Epidemiology, IRCCS AOU San Martino – IST Istituto Nazionale per la Ricerca sul Cancro, Genova, Italy; [2]Integrated Oncology Therapies Department, Molecular Pathology, IRCCS AOU San Martino – IST Istituto Nazionale per la Ricerca sul Cancro, Genova, Italy
*Correspondence: Dr R Puzone, Integrated Oncology Therapies Department, Clinical Epidemiology, IRCCS AOU San Martino – IST Istituto Nazionale per la Ricerca sul Cancro, 16132 Genova, Italy. Tel: +39 010 5558429; Fax: +39 010 354103; E-mail: roberto.puzone@hsanmartino.it
Received 11 January 2016; revised 12 October 2016; accepted 22 November 2016; published online 28 December 2016

performed. Public (level 3) log2 normalised gene expression data for the BRCA BC samples were merged with the SNV/SM/clinical data and difference of expression significances were assessed by using the Student's *t*-test. Statistical analyses were performed by using R(x64) 3.1.3 (http://www.R-project.org). Result data were submitted to the GWAS Central repository (http://www.gwascentral.org/study/HGVST1843).

RESULTS

In the Ellis 2012 pilot data set, we found strong indication that four SNVs in the 5q11.2 locus were associated with the mutational status of *PIK3CA* with high OR (Table 1). As expected by their LD, high correlation was found among these variants. Most of the other high OR SNV/SM associations had very low significance levels (full results in Supplementary Table S3). We focused on the associations of the four 5q11.2 SNVs with *PIK3CA* SM and verified them in the 754 ER+ BC patient data from the TCGA-BRCA data set. Two variants, rs331499 (hg19.chr5: g.56210923A>G) and rs252913 (hg19.chr5:g.56195846G>A), located in the boundary of *MAP3K1* and *SETD9* genes, were confirmed to be correlated with *PIK3CA* SM with high OR; a third, rs832552 (hg19.chr5:g.56113850T>G) inside *MAP3K1*, had few valid samples but a similar OR trend (Table 1). High correlation (Pearson's *r* range: 0.73–0.94) among the variants was confirmed (Supplementary Table S4). In logistic regression, no evidence of significant heterogeneity for ethnicity was found (Supplementary Tables S4 and S5).

The three 5q11.2 variants were found to be associated with the overexpression of one or both of their nearest genes *MAP3K1* and *SETD9*; for *MAP3K1*, associations were stronger in *PI3KCA* SM than in wild type (WT; Table 2 and Supplementary Tables S6–S8). Furthermore, we found a direct association between both *MAP3K1* and *SETD9* overexpression, and *PIK3CA* SM status – *MAP3K1* expression at *PIK3CA* SM/WT, difference of means: 0.38 (95% CI: 0.19, 0.57), $P=4e-4$; *SETD9* expression at *PIK3CA* SM/WT, difference of means: 1.44 (95% CI: 1.24, 1.63), $P=2e-16$.

DISCUSSION

In this short report, we show that germline SNVs located near the *MAP3K1/SETD9* genes associate with *PIK3CA* SM in ER+ BC with OR values (1.75 and 2.97 for rs331499 and rs252913, respectively) much higher than their OR of association with BC or BC subtypes (OR about 1.1,[18] as the OR of most cancer-risk SNV[4]). SNV data are coherent with gene expression data: the SNV associations with *MAP3K1/SETD9* overexpression are increased when the distance from the target gene is reduced, and, for *MAP3K1*, are stronger in *PIK3CA* SM BC samples. The overall picture is compatible with a *MAP3K1/SETD9* variant-dependent overexpression affecting *PIK3CA* SM penetrance. Moreover, we found a clear direct association of *PIK3CA* SM with *MAP3K1* and *SETD9* overexpression. Indeed, inter-regulation between PI3K and MAP-kinase pathways has been described in *in vitro* experiments and computer simulation,[19] and combination of drugs targeting both pathways is under clinical investigation.[20] A possible *SETD9* involvement is suggested by the strong SNV associations with *SETD9* overexpression; moreover, 5q11.2 SNV eQTL to *SETD9* has been reported also in normal blood.[18] However, we found a synergy of *PI3KCA* SM and SNV only for *MAP3K1* overexpression.

Two of our findings indicate that a complex BC risk SNV structure is present in the 5q11.2 region. First, only the SNV in the boundary of *MAP3K1/SETD9* genes (but not the reference risk SNP rs889312, which they are in high LD with) were found associated with *PI3KCA* SM. Second, phasing data showed that the SNV alleles associated with increased *PIK3CA* SM (and *MAP3k1/SETD9* overexpression) are actually correlated with the reduced BC risk allele (A) of rs889312[11]

Table 1 Association of SNVs close to the *MAP3K1* gene with somatic PIK3CA variants, two cohorts

SNV			Ellis 2012 pilot data set				TCGA-BRCA (ER+) data set			
Name/allele	*HGVS ID*	*Distance (kbases) from MAP3K1*	*OR (95% CI)*	*P-value*	*Patients*	*Allele freq*	*OR (95% CI)*	*P-value*	*Patients*	*Allele freq*
rs889312 A	hg19.chr5:g.56031884A>C	−79.0	Inf (0.69, Inf)	0.09	45	0.64	0.90 (0.33, 2.51)	1	112	0.66
rs832552 T	hg19.chr5:g.56113850T>G	<>	6.48 (1.18, 48.3)	0.03	45	0.5	1.55 (0.58, 4.24)	0.48	91	0.5
rs252913 G	hg19.chr5:g.56195846G>A	3.9	5.29 (0.88, 40.0)	0.06	45	0.51	2.97 (1.17, 7.75)	0.01	177	0.57
rs331499 A	hg19.chr5:g.56210923A>G	18.9	3.40 (0.70, 17.7)	0.15	45	0.51	1.75 (1.11, 2.77)	0.02	541	0.56

Abbreviations: <>, inside; Distance, SNV distance from gene start or end (the shortest); OR, odds ratio; CI, confidence interval for OR; *P*-value, *P*-value from Fisher's exact test.

Table 2 5q11.2 SNV association with MAP3K1 and SETD9 gene expression in TCGA-BRCA (ER+) data set

	SNV		*MAP3K1 gene expression*				*SETD9 gene expression*			
			PIK3CA SM			*PIK3CA WT*	*PIK3CA SM*			*PIK3CA WT*
Name/allele	*Distance (kbases) from*									
	MAP3K1	*SETD9*	*Diff. allele +/− (95% CI)*	*P-value*	*Diff. allele +/− (95% CI)*	*P-value*	*Diff. allele +/− (95% CI)*	*P-value*	*Diff. allele +/− (95% CI)*	*P-value*
rs832552 T	< >	−91.2	0.90 (0.46, 1.33)	4.0E−04	0.20 (−0.56, 0.96)	0.57	0.51 (−0.05, 1.08)	0.07	0.06 (−0.91, 0.79)	0.87
rs252913 G	3.9	−9.2	0.77 (0.36, 1.18)	1.0E−03	0.51 (−0.23, 1.25)	0.17	0.64 (0.05, 1.23)	0.04	0.72 (0.19, 1.25)	0.01
rs331499 A	18.9	< >	0.09 (−0.45, 0.61)	0.74	0.30 (−0.06, 0.66)	0.1	0.69 (0.30, 1.08)	0.02	0.81 (0.50, 1.13)	3.0E−06

Abbreviations: < >, inside; Distance, SNV distance from gene start or end (the shortest); PIK3CA SM/WT, in samples with PIK3CA somatic variants/wild type; Diff. allele +/− (95% confidence interval (CI)), difference of the means of gene expression in samples with presence/samples with absence of the allele, with 95% CI; *P*-value, *P*-value from Student's *t*-test.

(Supplementary Table S10). Hence, their opposite alleles should be associated with BCs in which PIK3CA is not mutated to build up to their overall increased BC risk.[18] This 'reverse' phase should not surprise because the SNV/*PIK3CA* SM associations found have an allele unbalancing effect one order of magnitude stronger than the reported SNV/BC risk. However, it predicts the presence of multiple classes of SNV BC risk in the 5q11.2 segment that split when probed for PIK3CA somatic variants.

This multiplicity could be a consequence of MAP3K1 ubiquitinase activity in addition to its kinase activity, which can therefore both activate and destabilise MAP-kinases.[21,22] The complex BC risk SNV structure has been confirmed by a recent fine scale analysis of 5q11.2 region in a large cohort of patients (not available when we started our investigation) that identified, by logistic regression, four BC risk-associated haplo-blocks.[17] By analysing in the BRCA data set, four genotyped SNV representative of the haplo-blocks, we found that only one SNV allele correlated with enriched *PIK3CA* SM, and it was associated with a reduced BC risk (Supplementary Table S9).

In conclusion, the germline 5q11.2 variants, rs331499 allele A and rs252913 allele G, are associated with *MAP3K1* and *SETD9* over-expression, and correlate with increased *PIK3CA* SM frequency in ER+ BC. Genome-wide analysis of SNV/SM associations can increase our understanding of tumour biology with relevant information for precision medicine.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

The results shown here are in part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/. We thank Paola Ghiorzo, Genova, for helpful comments.

1 Foulkes WD: Inherited susceptibility to common cancers. *N Engl J Med* 2008; **359**: 2143–2153.
2 Pharoah PD, Antoniou A, Bobrow M, Zimmern RL, Easton DF, Ponder BA: Polygenic susceptibility to breast cancer and implications for prevention. *Nat Genet* 2002; **31**: 33–36.
3 Mucci LA, Hjelmborg JB, Harris JR *et al*: Familial risk and heritability of cancer among twins in Nordic countries. *JAMA* 2016; **315**: 68–76.
4 Michailidou K, Beesley J, Lindstrom S *et al*: Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat Genet* 2015; **47**: 373–380.
5 Harlid S, Ivarsson MI, Butt S *et al*: Combined effect of low-penetrant SNVs on breast cancer risk. *Br J Cancer* 2012; **106**: 389–396.
6 Stephens PJ, Tarpey PS, Davies H *et al*: The landscape of cancer genes and mutational processes in breast cancer. *Nature* 2012; **486**: 400–404.
7 Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature* 2012; **490**: 61–70.
8 Stratton MR, Campbell PJ, Futreal PA: The cancer genome. *Nature* 2009; **458**: 719–724.
9 O'Brien KM, Cole SR, Engel LS *et al*: Breast cancer subtypes and previously established genetic risk factors: a bayesian approach. *Cancer Epidemiol Biomarkers Prev* 2014; **23**: 84–97.
10 Li Q, Seo JH, Stranger B *et al*: Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell* 2013; **152**: 633–641.
11 Lu PH, Yang J, Li C *et al*: Association between mitogen-activated protein kinase kinase kinase 1 rs889312 polymorphism and breast cancer risk: evidence from 59 977 subjects. *Breast Cancer Res Treat* 2011; **126**: 663–670.
12 Rhie SK, Coetzee SG, Noushmehr H *et al*: Comprehensive functional annotation of seventy-one breast cancer risk Loci. *PLoS One* 2013; **8**: e63925.
13 Ellis MJ, Ding L, Shen D *et al*: Whole-genome analysis informs breast cancer response to aromatase inhibition. *Nature* 2012; **486**: 353–360.
14 Tryka KA, Hao L, Sturcke A *et al*: NCBI's Database of Genotypes and Phenotypes: dbGaP. *Nucleic Acids Res* 2014; **42**: D975–D979.
15 Wilks C, Cline MS, Weiler E *et al*: The Cancer Genomics Hub (CGHub): overcoming cancer through the power of torrential data. *Database* 2014; e-pub ahead of print 29 September 2014; doi:10.1093/database/bau093.

16 Cerami, Gao J, Dogrusoz U *et al*: The cBio Cancer Genomics Portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov* 2012; **2**: 401.
17 Nyholt DR: A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am J Hum Genet* 2004; **74**: 765–769.
18 Glubb DM, Maranian MJ, Michailidou K *et al*: Fine-scale mapping of the 5q11.2 breast cancer locus reveals at least three independent risk variants regulating MAP3K1. *Am J Hum Genet* 2015; **96**: 5–20.
19 Aksamitiene E, Kiyatkin A, Kholodenko BN: Cross-talk between mitogenic Ras/MAPK and survival PI3K/Akt pathways: a fine balance. *Biochem Soc Trans* 2012; **40**: 139–146.
20 Britten CD: PI3K and MEK inhibitor combinations: examining the evidence in selected tumor types. *Cancer Chemother Pharmacol* 2013; **71**: 1395–1409.
21 Lu Z, Xu S, Joazeiro C, Cobb MH, Hunter T: The PHD domain of MEKK1 acts as an E3 ubiquitin ligase and mediates ubiquitination and degradation of ERK1/2. *Mol Cell* 2002; **9**: 945–956.
22 Xia Y, Wang J, Xu S, Johnson GL, Hunter T, Lu Z: MEKK1 mediates the ubiquitination and degradation of c-Jun in response to osmotic stress. *Mol Cell Biol* 2007; **27**: 510–517.

Supplementary Information accompanies this paper on European Journal of Human Genetics website (http://www.nature.com/ejhg)