2013

Novel insights into breast cancer genetic variance through RNA sequencing

Anelia Horvath
George Washington University

Suresh B. Pakala
George Washington University

Prakriti Mudvari
George Washington University

Sirigiri Divijendra Natha Reddy
George Washington University

Kazufumi Ohshiro
George Washington University

See next page for additional authors

Recommended Citation

Horvath, A., Pakala, S.B., Mudvari, P., Reddy, S.D.N., Ohshiro, K., Casimiro, S., Pires, R., Fuqua, S.A.W., Toi, M., Costa, L., Nair, S.S., Sukumar, S., Kumar, R. (2013). Novel insights into breast cancer genetic variance through RNA sequencing. Scientific Reports, 3:2256.

Authors

Anelia Horvath, Suresh B. Pakala, Prakriti Mudvari, Sirigiri Divijendra Natha Reddy, Kazufumi Ohshiro, Sandra Casimiro, Ricardo Pires, Suzanne A.W. Fuqua, Masakazu Toi, Luis Costa, Sujit S. Nair, Saraswati Sukumar, and Rakesh Kumar

SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
BREAST CANCER
CANCER GENOMICS
CANCER GENETICS
NEXT-GENERATION SEQUENCING

Received
23 May 2013

Accepted
17 June 2013

Published
25 July 2013

Correspondence and requests for materials should be addressed to R.K. (bcmrxk@gwu.edu)

* These authors contributed equally to this work.

Novel Insights into Breast Cancer Genetic Variance through RNA Sequencing

Anelia Horvath[1,2]*, Suresh Babu Pakala[2]*, Prakriti Mudvari[1]*, Sirigiri Divijendra Natha Reddy[2], Kazufumi Ohshiro[2], Sandra Casimiro[3], Ricardo Pires[3], Suzanne A. W. Fuqua[4], Masakazu Toi[5], Luis Costa[3], Sujit S. Nair[1,2], Saraswati Sukumar[6] & Rakesh Kumar[1,2,7]

[1]McCormick Genomic and Proteomics Center, [2]Department of Biochemistry and Molecular Medicine, The George Washington University, Washington, District of Columbia 20037, USA, [3]Institute of Molecular Medicine and, Hospital de Santa Maria – CHLN, Lisbon, Portugal, [4]Breast Center, Baylor College of Medicine, One Baylor Plaza, Houston, Texas 77030, USA, [5]Department of of Surgery, Kyoto University Graduate School of Medicine, Kyoto, Japan, [6]Sidney Kimmel Comprehensive Cancer Center at Johns Hopkins University, Baltimore, MD 21231, USA, [7]Cancer Research Program, Rajiv Gandhi Centre for Biotechnology, Thiruvananthapuram, India.

Using RNA sequencing of triple-negative breast cancer (TNBC), non-TBNC and HER2-positive breast cancer sub-types, here we report novel expressed variants, allelic prevalence and abundance, and coexpression with other variation, and splicing signatures. To reveal the most prevalent variant alleles, we overlaid our findings with cancer- and population-based datasets and validated a subset of novel variants of cancer-related genes: *ESRP2, GBP1, TPP1, MAD2L1BP, GLUD2* and *SLC30A8.* As a proof-of-principle, we demonstrated that a rare substitution in the splicing coordinator *ESRP2* (R353Q) impairs its ability to bind to its substrate *FGFR2* pre-mRNA. In addition, we describe novel SNPs and INDELs in cancer relevant genes with no prior reported association of point mutations with cancer, such as *MTAP* and *MAGED1.* For the first time, this study illustrates the power of RNA-sequencing in revealing the variation landscape of breast transcriptome and exemplifies analytical strategies to search regulatory interactions among cancer relevant molecules.

Breast cancer is the third most frequent cancer in the world as it affects approximately one in ten women in the western world[1]. The initial knowledge that connected breast cancer to genetic susceptibility originated from the clinical observations that highlighted the clustering of breast cancer cases in families[2,3]. Approximately 5–10% of breast cancers are believed to result from the inheritance of rare genetic components that confer significantly elevated risk[4,5]. For example, mutations in the tumor suppressor genes *BRCA1* and *BRCA2* account for approximately 16% of the familial breast cancer[6–8]. The vast majority of breast cancer cases, however, are derived from a complex interaction between multiple environmental, lifestyle and genetic factors with relatively weak individual risk contribution[9,10].

While the effects of many environmental and lifestyle factors, such as diet, reproductive behavior and radiation are well appreciated, the knowledge on genetically contributing patterns is limited. Association studies have identified *ATM, BRIP1*, *CASP8, CDH1*, *CHEK2*, *PALB2*, *PTEN*, *STK11*, and *TP53* as breast cancer susceptibility genes. Such mutations collectively account for 2.3% of familial risk of breast cancer, and together with *BRCA1*, *BRCA2* and others have been implicated in high risk screening strategies[5,8,11–20]. Nonetheless, significant proportion of the familial and non-familial breast cancer susceptibility remains unknown, suggesting plethora of genetic elements that need to be understood.

Transcriptome sequencing comprises a unique interplay between individual genetic background, reflected in the variation content, and the epigenetic and environmental regulation affecting gene expression levels and splice patterns. Recent transcriptome sequencing efforts have highlighted important somatic events in metastatic triple negative breast cancer (TNBC) and described important for the clinical outcome genotype-phenotype correlations[21]. Further, transcriptome sequencing data have been successfully explored to reveal disrupted pathways in TNBC through genome-wide loss of heterozygosity and mono-allelic expression estimation[22]. As a result of these and other studies, the feasibility of transcriptome sequencing to uncover molecular mechanisms of breast cancer drivers is increasingly appreciated[23].

Here we used whole transcriptome RNA-sequencing to reveal the variation signatures of 17 breast cancer patient tissues, and compared with human normal breast organoids (referred from here on as normal breast tissue, NBT). The 17 samples include six TNBC, lacking expression of therapeutically significant components - estrogen receptor (ER), progesterone-receptor (PR) and the Human Epidermal Growth Factor Receptor 2 (HER2); six Non-TNBC (ER, PR and HER2-positive); and five HER2-positive samples (ER and PR negative). Compared to the extensively performed searches for somatic breast cancer mutations, our RNA-sequencing based approach detects SNPs that are expressed at the mRNA level, and allows estimation of their allelic expression at nucleotide resolution. A set of novel variants were validated through Sanger sequencing. As a proof-of-principle, we have explored the effect of a rare SNP-p.R353Q - in the epithelial splicing regulatory protein ESRP2, on the binding and splicing of its target pre-mRNA. Our study reports a set of novel mutations in essential regulatory molecules in breast cancer and discusses their allelic preferential expression and potential involvement in breast cancer.

Results

Analytical strategies and overall variation landscape. We set out to define the transcribed variation profile of TNBC, Non-TNBC and HER2-positive breast cancer samples. To achieve this, we applied mRNA sequencing on 17 breast cancer samples from unrelated individuals as well as on three NBT samples on the Illumina HiSeq 2000 platform. The raw reads were aligned against Ensembl GRCh37.62 B (hg19) using TopHat[24], and the variants were called using Samtools[25]. Prior to filtering, a total of 1,876,617 SNPs, 331,197 of which were novel, were called across all 17 breast cancer samples, and between 30,294 and 258,465 SNPs (average 110,389) were called in each individual sample (Supplementary Table 1). The overview of the workflow and the filtering strategy is presented in Figure 1. The SNP calls were separated into two groups – either reported in the databases (between 22,914 and 218,411 per sample, average 91,201), or novel. The previously reported SNPs, due to validation by at-least one independent group, were analyzed further without filtering. To increase the confidence in the calls of novel variants, we initially analyzed the SNP calling reads of 1,000 SNPs through Integrative Genomics Viewer (IGV) files, and 96 of the calls were tested by Sanger sequencing. Based on the findings of this pilot validation test, we set up filtering criteria retaining minimum false-positive and false-negative calls as follows: those supported by at least of three bidirectional reads with unique start position, minimum phred quality value of 20, mapping quality value (MQV) > 20, and presence in 3 or less different samples. To ensure that we were not missing any novel high prevalence SNPs among our samples, all the positions at which a novel SNP was called in 4 or more samples were visually examined through IGV before to be assigned as false positives – no novel SNPs called in 4 or more samples were identified. This filtering left us with between 60 and 1143 novel variants per sample (average 285). The transition to transvertion (Ts/Tv) ratio among the novel coding SNPs was 2.8 and aligns with previously reported values for human exome, thus increasing the confidence of our filtering algorithm[26,27].

Prior to filtering, between 1,574 and 11,669 previously reported INDELs were called in each of the studied breast cancer samples and subjected to further analysis (See Supplementary Table 1). The novel INDELs were quality filtered to remove calls with MQV less than 20, phred quality value below 20 and presence in three or more different samples. This left between 18 and 142 novel INDELs (average 59) per sample, which were retained in our further analysis.

Expressed SNP density. To assess the overall expressed variation landscape of the breast cancer samples, we estimated the SNP density by counting the number of SNPs per megabase (MB) genome intervals. The SNP density was calculated individually for each sample and compiled per group (TNBC, non-TNBC and HER-2), and as a whole for the 17 samples (Figure 2). Overall, the SNP density distribution across the three groups was very similar, with a few regions showing group-specific high-density loci. All the TNBC samples presented with high number of SNPs in the region of chr14:10500000–106000000, which was mainly contributed by increased overall SNP number in the large gene encoding nucleoprotein *AHNAK2*. Specifically enriched in all non-TNBC samples was the region on chr19:53000000, mainly due to high number of SNPs in the zinc finger protein (*ZNF*) encoding genes.

We also overlapped the expressed overall SNP density in our samples with somatic genome SNP density calculated from the COSMIC dataset. There was a significant overlap in the overall SNP distribution. However, the regions with highest SNP frequency differed: while in the COSMIC dataset they were chr2:48000000, chr17:20000000 and chr5:72000000, the three top SNP-enriched regions when all of our samples were analyzed together were chr6:31000000–32000000, chr8:144000000 and chr19:53000000. While the high density observed on chromosome 6 was due to a the well-known variability in the histone cluster (*HIST1H1A*) and major histocompatibility complex (*HLA*), the chromosome 8 region was enriched by variants in epidermal antigen Epiplakin1 (*EPPK1*) and lymphocyte antigen (*LY6E*) (see Figure 2).

Comparative analysis with cancer genome variations (COSMIC). We compared the SNPs identified in our samples with the COSMIC cancer genome somatic mutation database (http://www.sanger.ac.uk/cosmic)[28]. A total of 2,169 SNPs from the COSMIC database were found among our samples, 129 of which were present in more than 10 of the 17 samples, and 6 were called in all breast cancer samples. Only one SNP – the relatively common variant R1322X in the ABC transporter gene *ABCA10* was nonsense, 515 were missense and 20 were located within a splice site. Of note, only two of the SNPs in our dataset, both UTR located, overlapped with COSMIC variants found in breast cancer: 1) the promoter T $>$ C substitution in the proto-oncogene binding Yes-associated protein (*YAP1*) was seen in 8 of our 17 patients, and 2) the 3′-UTR C $>$ T substitution in peptidylprolyl isomerase F (*PPIF*) was found in 5 of our samples. Among these comparisons, highly represented in our datasets were COSMIC missense variants in the DNA-repair encoding probable helicase senataxin *SETX* and Ewing's tumor-associated antigen *ETAA1*.

GWAS associated SNPs in the breast cancer transcriptome. To outline SNPs that have been previously associated with breast cancer phenotypes, we overlaid our datasets with the publically available genome wide association studies (GWAS); the results are summarized in Supplementary Table 2. The pre-B-cell leukemia homeobox 1 (*PBX1*) intronic SNP rs1387389 that has been reported to strongly associate with early onset breast cancer[29] was present in 4/17 samples, two of which were homozygous. Similarly, two breast cancer associated SNPs in the fibroblast growth factor receptor (*FGFR2*)[30], rs2420946 and rs2981582, were present each in two of our samples, (one patient was a carrier of both), again, in a homozygous state. Of note, the mitogen-activated protein kinase kinase kinase (*MAP3K1*) SNP reported by the same study was not present among our samples, however, we found a higher prevalence of the closely positioned D860N and V906I missense substitutions in *MAP3K1*; they were called in 13 (9 homozygous) and 16 (11 homozygous) of our samples, respectively. Similar high homozygocity prevalence was seen for the rs704010 rs8170, rs2180341, rs13281615, rs3817198 and rs4973768, but was not observed for the intergenic rs4415084. Other SNPs reported to be in strong association with breast cancer from recent meta-analyses[29–33] were not seen in our samples.

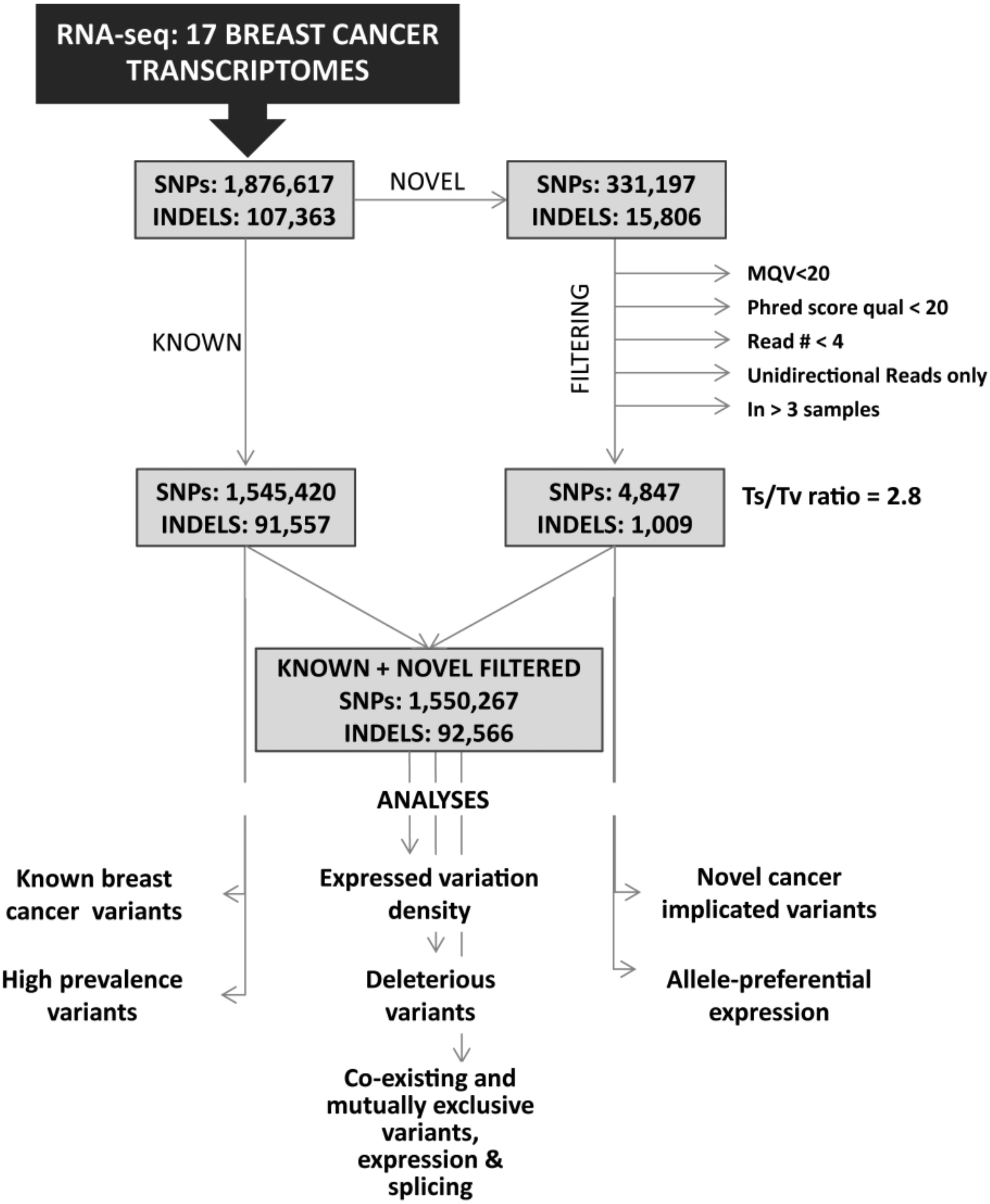

Figure 1 | Workflow and filtering overview. Different filtering strategies were applied to the novel and the previously reported variants. The previously reported variants, due to validation by at-least one independent group, were analyzed further without filtering. The filtering criteria for the novel variants were set as follows: those supported by a minimum of three bidirectional reads with unique start position, minimum phred quality value of 20, mapping quality value (MQV) > 20, and presence in 3 or less different samples. All the positions at which a novel SNP was called in 4 or more samples were visually examined through IGV before assignment as false positives.

Variations in genes previously implicated in hereditary breast cancer. To search for known predisposing breast cancer variants among our samples, we extracted SNPs and INDELS in genes that have been previously associated with hereditary breast cancer. Among all 17 samples, 80 SNP calls (38 unique SNPs) and 66 INDEL calls (38 unique INDELs) mapped within *ATM*, *BRCA1*, *BRCA2*, *BRIP1*, *CASP8*, *CDH1*, *CHEK1*, *PTEN*, *STK11* or *TP53*. While the majority of the SNPs called in those genes variants were common or have no known effect on the protein, several variants have been previously linked to breast cancer predisposition (Table 1). In *BRCA1* and *BRCA2* collectively, twelve different missense substitutions were identified in a total of nine patients. Five of the missense substitutions (p.Q356R, p.R496H, and p.D693N in *BRCA1*, and p.N289H and p.D1420Y in *BRCA2*) have been previously associated with breast cancer through either family or case-control studies[34,35]. Three patients from the non-TNBC group were carriers

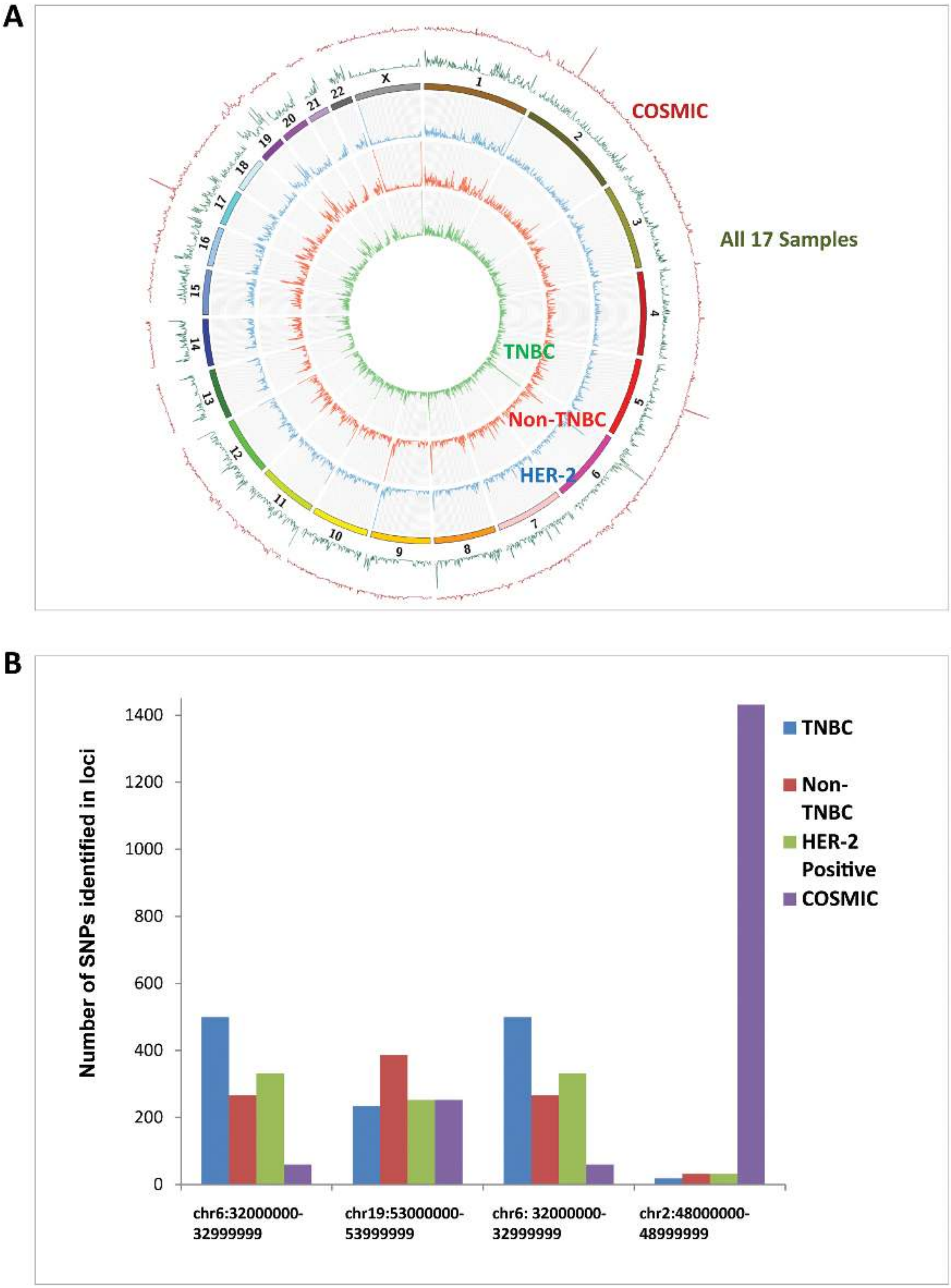

Figure 2 | Expressed SNP density expressed as number of SNPs per megabase (MB) genome intervals. The SNP density was calculated individually for each sample and compiled per group (TNBC, non-TNBC and HER-2), and as a whole for the 17 samples. Overall, similar SNP density distribution is observed across the three groups. (A) Circos plot representing the high density expressed SNPs in TNBC, Non-TNBC and HER2 positive Breast Cancer Samples compared to cancer genome SNP data from COSMIC. (B) The highest SNP density loci for TNBC, Non-TNBC, HER2 and COSMIC, compared to the SNP density for the same locus for the other three groups. The highest SNP density for the COSMIC was observed in the interval chr2:48000000–48999999, containing the genes *MSH6*, *FBX011*, *FOXN2*, *PPP1R21*, *STON1*, *GTF2A1L* and *LHCGR*, while very low expressed SNP density for this region was measured in all three breast cancer subtypes.

of the missense variant p.Q356R, and the other cancer-associated variants were present in one patient each – from either non-TNBC or HER2 positive groups. The two *BRCA2* missense substitutions were seen in HER2-positive patients. In addition, one TNBC patient carried a small *BRCA1* deletion (chr17:41246251delC, rs80357794) that leads to a frame-shift and premature stop codon expected to completely abolish protein function. In the *ATM* gene, we identified the non-synonymous substitutions p.F858L and

Table 1 | Variants in *ATM*, *BRCA1*, *BRCA2* and *STK11* identified in the 17 breast cancer samples (from HGMD, http://www.hgmd.cf.ac.uk/ac/index.php. The homo- or heterozygosity and the number of the unique variant and reference reads are also shown

Gene	Chromosomal Location	cDNA	Protein	rsID	Function	Cancer Associated	ID	Cancer subtype	Zygosity	Var/ref calls
ATM	chr11:108138003	c.2572 T > C	p.F858L	1000056	missense	YES	IP2-71	Non-TNBC	heterozygote	3/2
	chr11:108160350	c.4258 C > T	p.L1420F	1000058	missense	YES	171	HER2	heterozygote	9/10
	chr11:108175462	c.5557 G > A	p.D1853N	1801516	missense	NO	IP2-42	Non-TNBC	heterozygote	6/2
							IP2-49	Non-TNBC	homozygote	4/0
							56	HER2	heterozygote	10/3
							83	HER2	heterozygote	6/7
BRCA1	chr17:41246481	c.1067 T > C	p.Q356R	1799950	missense	YES	IP2-42	Non-TNBC	homozygote	2/0
							IP2-49	Non-TNBC	heterozygote	4/7
							IP2-66	Non-TNBC	heterozygote	3/6
	chr17:41246061	c.1487 C > T	p.R496H	28897677	missense	YES	IP2-71	Non-TNBC	heterozygote	4/4
	chr17:41245471	c.2077 C > T	p.D693N	4986850	missense	YES	83	HER2	homozygote	4/1
	chr17:41246251	c.1156delG	p.A386Pfs	80357794	indel	YES	IP2-78	TNBC	homozygote	4/0
	chr17:41244936	c.2612 G > A	p.P871E	799917	missense	NO	IP2-48	Non-TNBC	heterozygote	3/1
							IP2-49	Non-TNBC	homozygote	6/0
							IP2-66	Non-TNBC	heterozygote	6/3
							26	HER2	homozygote	20/0
							83	HER2	heterozygote	4/1
							171	HER2	homozygote	5/0
							IP2-78	TNBC	heterozygote	3/2
	chr17:41244435	c.3113 T > C	p.E1038G	1799966	missense	NO	IP2-48	Non-TNBC	homozygote	2/0
	chr17:41244000	c.3548 T > C	p.K1183R	16942	missense	NO	IP2-48	Non-TNBC	homozygote	3/0
							83	HER2	heterozygote	4/1
							171	HER2	homozygote	5/0
	chr17:41223094	c.4900 T > C	p.S1634G	1799966	missense	NO	IP2-48	Non-TNBC	homozygote	3/0
							IP2-66	Non-TNBC	homozygote	4/0
							171	HER2	homozygote	4/0
BRCA2	chr13:32906480	c.865 A > C	p.N289H	766173	missense	YES	56	HER2	heterozygote	3/2
	chr13:32912750	c.4258 G > T	p.D1420Y	766173	missense	YES	26	HER2	heterozygote	6/5
	chr13:32911463	c.2971 A > G	p.N991D	1799944	missense	NO	56	HER2	heterozygote	4/4
	chr13:32929387	c.7397 T > C	p.V2466A	169547	missense	NO	IP2-50	TNBC	homozygote	2/0
							26	HER2	homozygote	3/0
							56	HER2	homozygote	3/0
							83	HER2	homozygote	2/0
							171	HER2	homozygote	5/0
	chr13:32930673	c.7544 C > T	p.T2515I	28897744	missense	NO	IP2-66	Non-TNBC	homozygote	2/0
STK11	chr19:1220427	c.520 C > T	p.H174Y	0	missense	YES/lung	IP2-42	Non-TNBC	heterozygote	8/18

p.L1420F, which have been previously associated with increased risk for breast cancer[36]. The *ATM* missense p.F858L is known to impact the interaction of *ATM* with beta-adaptin, which is necessary for clatherin mediated receptor endocytosis and is proposed to contribute to the hereditary radio sensitivity and breast cancer[37,38]. In addition to breast cancer-associated variants, a missense substitution in *STK11*, p.H175Y, previously reported in a lung carcinoma[39] was been found in one patient. One non-TNBC patient carried simultaneously pathogenic variants in *BRCA1* and *ATM*. Of note, overall a higher number of variant versus reference reads was assessed across *BRCA1*, *BRCA2* and *ATM* variations.

Prevalence of rare variants. To reveal variants that might be overrepresented in our samples compared to the general population, we compared the allele frequency of coding SNPs called in our samples against 11,666 alleles from the Exome Sequencing Project dataset (http://evs.gs.washington.edu/EVS/). To minimize error due to different variant calling platforms and to increase the statistical significance of the findings, we excluded from this analysis SNPs called in less than 10 alleles from the ESP dataset and in less than 3 individuals among our 17 samples. SNPs called in all 17 of our samples were also excluded. For the purposes of allele frequencies comparison, we assigned two alleles for every homozygote call in our dataset; and Yates corrected chi-square was calculated for each distribution. The top 50 most prevalent missense SNPs among the 17 samples are presented in Table 2. The highest difference in the allele distribution between our dataset and ESP was estimated for rs2305376 in the gene *HOOK2*, encoding a component of the FTS/Hook/FHIP (FHF) complex that has a role in vesicle trafficking and maintenance of centrosome function and is known to interact with the *JUN* proto-oncogene[40,41]. Interestingly, the variant is predicted to be damaging change due to glycine to arginine substitution (p.G10R), which in addition to its low prevalence in the ESP datasets, is rare to absent in the European population datasets (see Table 2). This variant was called in 3 of our 17 samples, and all of them were called homozygous due to the high abundance of variant over wild type reads. Another overrepresented SNP in our dataset was the missense substitution p.T573A in the protein tyrosine phosphatase *PTPN12*, whose activity is lost in a large proportion of breast cancer cases[42]. Of note, while *PTPN12* loss is most strongly associated with the TNBC phenotype[43], we found this variant equally prevalent in all three breast cancer subtypes; one TNBC and one HER2-positive samples carried it in homozygote state. Other breast cancer implicated genes with prevalent variants amongst our samples were *PLEC*, *PRCP*, *DSG2* and *ERBB2IP*, all harboring predicted to be damaging aminoacid changes[44–47]. Potential contribution of such variants to the phenotype in these patients is worth investigation.

Deleterious protein mutations. A selected subgroup of potentially deleterious SNPs consisted of mutations predicted to generate premature stop (PMS) codons through either nonsense substitution or a splice-site aberration leading to a frame-shift due to out-of-frame exon skip or intron retention. In this group we also retained SNPs

Table 2 | Comparison between the observed prevalence of known SNPs in the 17 studied breast cancer samples and Exome Sequencing Project – presented are the top 50 variants showing higher prevalence among the 17 samples. The variants are sorted according descending chi-square value. High Polyphen score indicates high probability of the variant to alter the protein function

	17 Breast cancer samples				ESP				Chi-Square Test				HapMap					
Gene	#Samples	#Var Alleles	#Ref allelels	Var/Ref alleles	#Var Alleles	#Ref allelels	Var/Ref alleles	FREQ 17/ FREQ ESP	value	*P*	Yates value	Yates *P*	Afr	Eur	Asia	rsID	AA change	Polyphen Score
HOOK2	3	6	28	0.21	8	12136	0.001	325.1	912	0	766	0	0	0	7.3	2305376	G10R	0.99
ANTXR1	3	3	31	0.10	21	12903	0.002	59.5	138	0	95	0	NA	NA	NA	28365986	R7K	0.009
MCOLN1	3	4	30	0.13	46	12960	0.004	37.6	115	0	88	0	NA	NA	NA	73003348	T261M	0.614
C3orf17	3	4	30	0.13	65	12925	0.005	26.5	82	0	62	0	NA	NA	NA	115971253	V297I	0.117
TNKS1BP1	5	6	28	0.21	182	12808	0.014	15.1	63	0	52	0	NA	NA	NA	34448143	A100P	0.194
PLEC	4	6	28	0.21	185	12391	0.015	14.4	59	0	49	0	NA	NA	NA	3135103	R569Q	0.993
DDX18	3	4	30	0.13	93	12913	0.007	18.5	56	0	42	0	NA	NA	NA	61755349	V371I	0.011
AP1M2	3	3	31	0.10	52	12238	0.004	22.8	54	0	37	0	NA	NA	NA	34276903	Y85C	0.998
GEMIN4	3	6	28	0.21	297	12201	0.024	8.8	33	0	27	0	NA	NA	NA	191778127	H873Q	0
LY75	5	8	26	0.31	658	12348	0.053	5.8	24	0	20	0	NA	NA	NA	35941588	T1393I	0.435
TXNDC5	4	5	29	0.17	289	10583	0.027	6.3	18.8	0.00001	14.4	0.0001	NA	NA	NA	183777097	P49S	0
HIST1H1B	6	6	28	0.21	512	12494	0.041	5.2	16.7	0.00004	13.3	0.0003	0	7.5	0	34144478	A211T	0.404
IMPACT	8	11	23	0.48	1602	11404	0.140	3.4	12.5	0.0004	10.8	0.001	4.3	16.9	2.2	582234	D125E	0.001
AKAP9	7	9	25	0.36	1229	11777	0.104	3.4	11.4	0.0007	9.5	0.002	1.7	15	0.6	35759833	K2476R	0.001
QRSL1	3	5	29	0.17	456	12550	0.036	4.7	12.5	0.0004	9.4	0.002	0	5	5.6	34221917	N263S	0.593
DSG2	6	7	27	0.26	759	11175	0.068	3.8	11.5	0.0007	9.2	0.002	0	9.9	0.2	2230234	I293V	0.99
HLA-DRB5	6	10	24	0.42	1288	10016	0.129	3.2	10.9	0.001	9.2	0.002	NA	NA	NA	112872773	V209L	0
JRK	4	5	29	0.17	459	12219	0.038	4.6	11.9	0.0006	8.9	0.002	NA	NA	NA	34288113	T30M	0.988
LAPTM5	11	14	20	0.70	2541	10465	0.243	2.9	10.1	0.001	8.8	0.003	2.5	25.8	7.2	35351292	R226K	0.002
DDRGK1	8	11	23	0.48	1762	11244	0.157	3.1	10.2	0.001	8.7	0.003	0	19.2	0	11591	A303T	0.004
KIF20B	7	12	22	0.55	2044	10952	0.187	2.9	9.8	0.001	8.4	0.004	2.5	26.9	37.9	1129777	A50G	0.918
GLB1	4	6	28	0.21	637	11695	0.054	3.9	10.7	0.001	8.3	0.004	NA	NA	NA	73826339	S401G	0.001
WDR55	5	7	27	0.26	873	12133	0.072	3.6	10.4	0.001	8.3	0.004	NA	NA	NA	35983033	Y235C	0.998
THUMPD3	8	10	24	0.42	1564	11442	0.137	3.0	9.7	0.001	8.1	0.004	4.3	18.9	7.8	1129174	R459Q	0.056
NLRP2	8	14	20	0.70	2629	10377	0.253	2.8	9.2	0.002	8	0.005	NA	NA	NA	34804158	T529A	0
XPO5	4	6	28	0.21	645	11467	0.056	3.8	10.1	0.001	7.9	0.005	0	14.2	0	34324334	S241N	0.002
RRS1	3	4	30	0.13	374	12560	0.030	4.5	9.4	0.002	6.6	0.01	NA	4.1	20.4	3739336	R191L	0.999
PLSCR1	6	6	28	0.21	770	12236	0.063	3.4	8.3	0.004	6.4	0.01	0	8.3	3.1	343320	H262Y	0.945
PTPN12	10	12	22	0.55	2252	10754	0.209	2.6	7.6	0.005	6.4	0.01	19.7	11.3	29.8	3750050	T573A	0.001
REV3L	4	6	28	0.21	789	12215	0.065	3.3	7.9	0.004	6	0.01	3.8	4.1	0	458017	Y1156C	0.002
ERBB2IP	7	9	25	0.36	1498	11506	0.130	2.8	7.4	0.006	6	0.01	0	17.8	8.1	3213837	S274L	0.024
SHARPIN	4	4	30	0.13	401	12255	0.033	4.1	8.1	0.004	5.6	0.02	NA	NA	NA	34674752	P294S	0.447
KIAA1755	10	19	15	1.27	4605	8401	0.548	2.3	6.2	0.01	5.4	0.02	10.8	47.5	38.8	3746471	R1045W	0.003
SEC63	3	4	30	0.13	427	12579	0.034	3.9	7.6	0.006	5.2	0.02	NA	NA	NA	17854547	V556I	0.203
LAMC2	4	6	28	0.21	848	12158	0.070	3.1	6.9	0.009	5.2	0.02	0	9.9	0.2	11586699	T124M	0.999
ZNF880	4	7	27	0.26	376	4190	0.090	2.9	6.7	0.009	5.2	0.02	NA	NA	NA	14048	V12M	0.999
IGFBP7	7	8	26	0.31	640	5676	0.113	2.7	6.6	0.01	5.2	0.02	NA	NA	NA	11573021	L11F	0.005
PHF3	3	3	31	0.10	255	12751	0.020	4.8	8.2	0.04	5.1	0.02	NA	NA	NA	34288820	V525I	0.972
KIAA0232	3	4	30	0.13	395	11489	0.034	3.9	7.5	0.006	5.1	0.02	NA	NA	NA	116439703	P1138S	0.22
PJA1	7	11	23	0.48	1742	8821	0.197	2.4	6.2	0.01	5.1	0.02	12.5	24.6	12.5	11539157	E606D	0.992
PRCP	3	3	31	0.10	265	12741	0.021	4.7	7.8	0.005	4.8	0.03	NA	NA	NA	2228312	T465S	0.021
MXRA5	14	23	11	2.09	5006	5557	0.901	2.3	5.6	0.02	4.8	0.03	22	27.6	34.9	1974522	P1665S	0.689
PARP14	7	7	27	0.26	1022	10826	0.094	2.7	6.1	0.01	4.7	0.03	4	7.6	10.9	13093808	A561E	0.972
EML4	12	15	19	0.79	3414	9592	0.356	2.2	6.6	0.02	4.7	0.03	11.7	26.7	43.9	28651764	K398R	0.003

Table 2 | Cont.

Gene	17 Breast cancer samples #Samples	 #Var Alleles	 #Ref allelels	 Var/Ref alleles	ESP #Var Alleles	 #Ref allelels	 Var/Ref alleles	FREQ 17/ FREQ ESP	Chi-Square Test value	 *P*	 Yates value	 Yates *P*	HapMap Afr	 Eur	 Asia	rsID	AA change	Polyphen Score
ACOX1	13	24	10	2.40	6616	6390	1.035	2.3	5.3	0.02	4.5	0.03	13.6	34.9	29	1135640	I312M	0.007
SP110	3	6	28	0.21	928	12078	0.077	2.8	5.6	0.01	4.16	0.04	0	9	4	11556887	A128V	0.999
PPL	12	20	14	1.43	5261	7733	0.680	2.1	4.7	0.03	4	0.05	0	47.5	30.4	2037912	Q1573E	0.994
TBL2	3	4	30	0.13	489	12517	0.039	3.4	5.9	0.01	3.9	0.05	NA	3.1	NA	35607697	V345I	0.969
ZBTB45	3	4	30	0.13	491	12511	0.039	3.4	5.9	0.01	3.9	0.05	NA	NA	NA	35430780	D293E	0
TMEM106C	4	6	28	0.21	962	12044	0.080	2.7	5.1	0.02	3.8	0.05	0	11.2	40.4	2286025	S175F	0.985

removing a stop codon, because of their known severe biological consequences. A total of 1,593 variants of this type were called across our 17 samples, from which 77 were different nonsense SNPs and 16 were INDELs leading to a stop codon generation (Table 3). Among the stop-codon mutations, p.R93X in Steroid Receptor activator *SRA1* was present in 10 of the 17 samples, and in 8 of them it was called homozygous. Given that SRA1 is involved in regulating the activity of steroid receptors and is deregulated in breast cancer[48], the high expression rate of a nonsense variant might indicate functional implications in our breast cancer samples. In addition, two HER2 and two TNBC samples were positive for the p.Q281X in the zinc-activated ligand-gated ion channel (*ZACN*). Interestingly, the nucleotide change causing this mutation - chr17: 74077797 C > T - also resides in the 5′UTR of the gene encoding exocyst complex component (*EXOC7*), whose deregulated expression was reported to be a strong predictor for metastatic outcome in early stage TNBC[49]. The two TNBC samples positive for the variant did not show presence of reference reads in this position. When all the genes affected by a deleterious mutation were analyzed through Ingenuity Pathway Analysis (IPA), the top affected molecular networks were cell death and survival, cellular development, and cellular growth and proliferation, and the top affected canonical pathway was estrogen receptor signaling (Supplementary Figure 1).

Novel expressed variations in breast cancer, and allele specific expression. The statistics on the filtered novel SNPs and INDELs are summarized in Supplementary Table 3; a complete list of the novel exonic annotated variations is available upon request. As expected, majority novel variants mapped within gene regions (70% of the SNPs and 66% INDELs). Filtering out of the intronic calls significantly reduced both SNP and INDEL numbers to between 43 and 186 SNPs per sample (average 76) and between none and 17 INDELs (average 8). Overall, 8% of the novel intergenic SNPs and 4% of the novel INDELs mapped within exons. Across the 17 samples, the total number of genes with coding and regulatory sequences affected by at least one novel SNP was 2103, and the genes with at least one novel INDEL were 566. A selected set of exonic variants were confirmed by Sanger sequencing (Figure 3).

From the novel exonic SNPs, 285 unique SNPs were predicted to alter the protein sequence. Based on position and function, three of these SNPs were annotated to generate a novel stop codon, 114 were located within 2 bp of a splice junction, and 174 were missense, from which 70 were predicted to significantly affect the protein function. Six novel SNPs had dual annotation: missense substitutions located at a splice site. A total of 121 novel coding SNPs affected highly conservative nucleotide positions. Three novel coding SNPs - one missense and two synonymous substitutions - were called in two samples, and one - a stop codon in the solute carrier *SLC30A8* (p.Q28X) - in three different breast cancer samples (Table 4). Interestingly, p.Q28X affects only one (NM_173851.2) of the five protein coding *SLC30A8* splice isoforms; this isoform possesses an alternative 5′-end and is present in all three samples from our set that expressed this isoform. The stop codon is located early in the protein chain and likely leads to complete abolishment of the protein expression. Since this *SLC30A8* isoform was not expressed in the remaining 14 breast cancer samples, this early stop codon may indicate regulatory mechanism preventing the expression of this particular isoform in breast tissue.

To assess potential allele preferential expression, we analyzed the ratio of reference and variant reads at all coding positions for the novel SNPs called by 6 or more reads harboring the variant nucleotide (Figure 4). Fifty seven of these novel SNPs were called by variant reads only (i.e. no reference call was present at the corresponding position), and additional 53 showed higher than 5-fold number of variant calling reads over the wild type (Supplementary Table 4). Among the most preferentially expressed novel SNPs were missense

Table 3 | Deleterious mutations identified among the 17 Breast Cancer Samples

Gene	Sample ID	# Samples	Function	Location	REF allele	Var allel	Change	Zygosity
ABCA10	56	1	stop-gained	chr17:67149973	G	A	p.R1322X	homozygote
C17orf77	IP2-76	2	stop-gained	chr17:72588806	C	A	p.C207X	homozygote
C17orf77	IP2-71	2	stop-gained	chr17:72588806	C	A	p.C207X	homozygote
C5orf20	56	1	stop-gained	chr5:134782450	T	A	p.R117X	homozygote
CARM1	IP2-50	1	stop-gained	chr19:11019790	C	A	p.Y155X	heterozygote
CCDC25	IP2-90	1	stop-gained	chr8:27610077	C	A	p.E66X	homozygote
DCAF11	IP2-50	1	stop-gained	chr14:24592413	C	A	p.Y486X	heterozygote
EFHA1	IP2-69	1	stop-gained	chr13:22077082	T	A	p.K306X	heterozygote
EGFL6	171	1	stop-gained	chrX:13624542	C	T	p.R189X	homozygote
ERV3-1	IP2-50	1	stop-gained	chr7:64452738	G	A	p.R223X	homozygote
EXOC7,ZACN	56	4	stop-gained	chr17:74077797	C	T	p.Q281X	heterozygote
EXOC7,ZACN	IP2-76	4	stop-gained	chr17:74077797	C	T	p.Q281X	homozygote
EXOC7,ZACN	IP2-78	4	stop-gained	chr17:74077797	C	T	p.Q281X	homozygote
EXOC7,ZACN	171	4	stop-gained	chr17:74077797	C	T	p.Q281X	heterozygote
FCGR2A	IP2-65	1	stop-gained	chr1:161476204	C	T	p.Q62X	homozygote
GAB4	IP2-90	1	stop-gained	chr22:17469049	C	A	p.G163X	homozygote
GET4	IP2-83	1	stop-gained	chr7:931966	C	A	p.Y219X	homozygote
HNRNPR	IP2-76	1	stop-gained	chr1:23637469	G	T	p.Y460X	heterozygote
IL17RB	IP2-83	1	stop-gained	chr3:53899276	C	T	p.Q484X	heterozygote
LAIR2	IP2-76	1	stop-gained	chr19:55019261	C	T	p.R76X	heterozygote
LOC1009964	26	1	stop-gained	chr6:57398270	C	T	p.Q325X	heterozygote
MAD2L1BP	IP2-49	1	stop-gained	chr6:43608124	C	T	p.R227X	heterozygote
MADD	IP2-83	1	stop-gained	chr11:47306630	C	T	p.R766X	homozygote
MAGEB16	IP2-66	1	stop-gained	chrX:35821127	C	T	p.R272X	homozygote
METAP1	IP2-50	1	stop-gained	chr4:99982427	C	T	p.R374X	heterozygote
MTA2	IP2-83	1	stop-gained	chr11:62364262	G	T	p.Y243X	homozygote
NHLRC2	IP2-76	1	stop-gained	chr10:115618327	C	A	p.Y73X	heterozygote
PDE4DIP	171	1	stop-gained	chr1:144915561	G	A	p.R622X	heterozygote
PDE4DIP	IP2-65	1	stop-gained	chr1:145075683	C	T	p.W60X	homozygote
PDE4DIP	IP2-49	2	stop-gained	chr1:144916676	C	T	p.W560X	heterozygote
PDE4DIP	26	2	stop-gained	chr1:144916676	C	T	p.W560X	heterozygote
PELI3	IP2-50	1	stop-gained	chr11:66235714	G	T	p.E39X	homozygote
PKD1L2	IP2-65	1	stop-gained	chr16:81242198	G	A	p.Q220X	homozygote
PRB4	IP2-49	1	stop-gained	chr12:11461802	G	A	p.R39X	homozygote
PRM3	IP2-76	1	stop-gained	chr16:11367143	G	A	p.R104X	homozygote
RHBDD3	IP2-49	1	stop-gained	chr22:29656431	C	T	p.W289X	homozygote
SKIV2L	IP2-42	1	stop-gained	chr6:31936654	C	T	p.R1063X	homozygote
SYNE2	26	1	stop-gained	chr14:64560092	G	A	p.W4001X	homozygote
TMEM134	IP2-42	1	stop-gained	chr11:67235051	G	A	p.R84X	heterozygote
VPS13B	171	1	stop-gained	chr8:100133706	T	G	p.Y413X	homozygote
ZSWIM3	171	1	stop-gained	chr20:44505411	G	T	p.E72X	homozygote
ANKS1A	IP2-69	1	INDEL	chr6:34738008	A	AA		homozygote
ANKS1A	IP2-42	1	INDEL	chr6:34902473	G	GT		heterozygote
CABIN1	171	1	INDEL	chr22:24455826	G	GAAAA		homozygote
CABIN1	83	1	INDEL	chr22:24448944	T	TT		homozygote
CANX	IP2-69	2	INDEL	chr5:179140762	A	AA		homozygote
CMYA5	56	1	INDEL	chr5:78982956	GCTT	GCTTCTT		homozygote
EME1	56	1	INDEL	chr17:48276005	C	CC		homozygote
LAMA3	IP2-49	1	INDEL	chr18:21434967	ATAAA	A		homozygote
MGST2	IP2-42	3	INDEL	chr4:140619265	T	TT		homozygote
MRPS15	56	1	INDEL	chr1:36921785	TA	TGGAAAA		homozygote
SLC17A5	IP2-78	1	INDEL	chr6:74351412	AC	ACACC		homozygote
SLC5A8	IP2-66	1	INDEL	chr12:101550975	CACA	CACACA		heterozygote
SMARCA5	171	1	INDEL	chr4:144340520	AAGAA	AA		heterozygote
TRAPPC9	IP2-83	1	INDEL	chr8:141413543	A	AA		homozygote
ZNF100	IP2-50	1	INDEL	chr19:21908799	CACA	CA		homozygote
ZNFX1	IP2-42	1	INDEL	chr20:47871283	GACCCTTGGA			homozygote

variants in previously linked breast cancer genes, such as methylthioadenosine phosphorylase *MTAP* (p.K71R), and melanoma antigen *MAGED1* (p.G87A)[50].

Studies revealing impaired interaction of splicing coordinator ESRP2 bearing a R353Q substitution. Among the novel and rare SNPs predicted to be protein-altering in our breast cancer samples, we selected to study the functional effect of the R353Q substitution in *ESRP2*, based on the established connection of ESRP2 to cancer through its role in epithelial-to-mesenchymal transition (EMT)[51–53]. Arginine 353 is located in the second RNA recognition motif (RRM) domain of the ESRP2, which is known to interact with specific pre-mRNAs sequences. There are three RRM domains in ESRP2, and they are implicated in regulating the expression of specific splice variants of *FGFR2*, *CTNND1* and *ENAH* that are involved in EMT. We applied site-directed mutagenesis to generate *ESRP2*^{R353Q} harboring expression vector and transfected MDA-MB-231 human breast cancer cells in parallel with expression constructs containing wild type *ESPR2*. After transfection, wild type and the mutant ESRP2^{R353Q} proteins were purified and compared for their ability

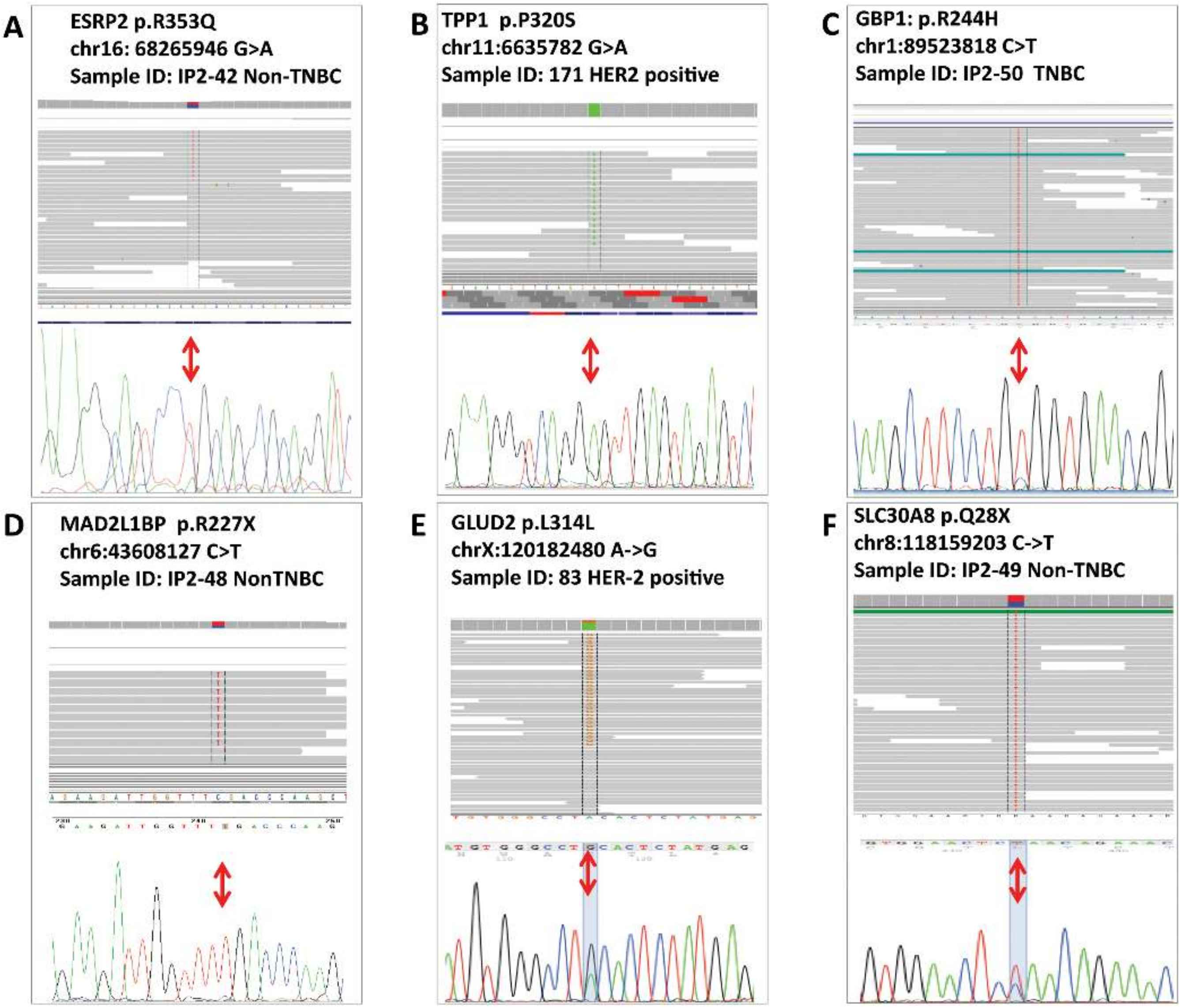

Figure 3 | Sanger Sequencing validation of selected variants; IGV is also presented. (A) *ESRP2*, p.R353Q; (B) *TPP1*, p.P320S; (C) *GBP1*, p.R244H; (D) *MAD2BP1*, p.R227X. For *TPP1*, *GBP1* and *MAD* both IGV and chromatogram show prevalence of the variant allele.

to bind the *FGFR2* pre-mRNA region through Electrophoretic Mobility Shift Assay (EMSA); the results are shown in Figure 5. We observed strong interactions between the wild type ESRP2 and the *FGFR2* pre-mRNA as previously reported[52]. However, this interaction was significantly impaired for the mutant ESRP2^{R353Q} (compare lanes 6 and 7 with Lanes 2 and 3, Panels A–C), suggesting that the R353Q substitution compromises ESPR2 binding, and potentially, splice regulation of the *FGFR2* pre-mRNA. This effect was observed in all three tested breast cancer cell lines: MCF-7 (Figure 5A), MDA-MB-231 (Figure 5B), and BT-549 (Figure 5C). Further, in line with previous observations[52], RT-PCR showed increased expression of the epithelial FGFR2 isoform IIIb after transfection with wild type ESRP2 in the mesenchymal FGFR2 IIIc-expressing cell lines MDA-MB-231 and BT-549 (Figure 5D). This increase in FGFR2 IIIb expression was lower (BT-549) to completely abolished (MDA-MB-231) after the transfection with the mutant ESRP2^{R353Q} (Figure 5D).

Discussion

Here we present the first mRNA sequencing based study that reports expressed variations in TNBC, Non-TNBC and HER2-positive breast cancer transcriptome. Several molecular mechanisms, such as RNA editing and allele preferential expression, could cause a discrepancy between the variations found at mRNA and DNA levels. Compared to exome and genome sequencing, RNA-seq provides essential insights into the functionality of the variants through estimation of the absolute and relative abundance of variant reads and the co-existence or mutual exclusion of variations, expression and splicing patterns. In addition to outlining the general landscape of the breast cancer variation transcriptome, our study reports novel variants in an allele-specific expression context, aligns our findings with the existing knowledge on breast cancer genetics, and exemplifies efficient extraction of information from the transcriptome through extensive analyses.

Mutations in previously associated breast cancer genes, *BRCA1*, *BRCA2* and *ATM*, were called in 9/17 (53%) samples which is a higher than the previously reported mutation prevalence among breast cancer patients[5, 12–16,18–20]. While only one patient was a carrier of known pathological variants in both *BRCA1* and *ATM*, five other individuals carried missense substitutions in at least two different breast cancer associated genes (see Table 1). Whether the disease in these patients could be contributed to cumulative impaired functioning of these genes is a subject of further investigation; nevertheless, the relatively frequent co-occurrence of protein altering variations in

Table 4 | Novel Exonic Variants that are seen in two and three out of the 17 breast Cancer Samples

Gene	Chromosomal Location	cDNA Annotation	Protein Annotation	Function	Samples	Cancer subtype
SLC30A8	chr8:118159203 C > T	c.82 C > T	p.Q28X	stop-codon	IP2-48	Non-TNBC
					IP2-49	Non-TNBC
					IP2-66	Non-TNBC
AGL	chr1:100387140 G > T	c.4484 G > T	p.C1495F	missense	IP2-49	Non-TNBC
					IP2-66	Non-TNBC
GLUD2	chrX:120182480 A > G	c.942 A > G	p.L314L	synonymous	IP2-50	TNBC
					83	TNBC
GPN1	chr2:27873001 A > G	c.1101 A > G	p.E367E	synonymous	IP2-49	Non-TNBC
					IP2-66	Non-TNBC

known breast cancer-associated genes in different cancers raises the necessity to examine larger series of patients and controls for combinatorial genetic risk. An interesting observation is the high prevalence of homozygote vs. heterozygote calls in *BRCA1*, *BRCA2* and *ATM* for both breast cancer-associated genes, and those not known to be pathogenic variants, suggesting potential allelic loss in those genes.

Among the essential findings of our study is a subset of novel SNPs and INDELs, some affecting genes previously implicated in breast cancer, in which, however, no predisposing or causative point variants have been reported so far. An example is p.K71R in *MTAP*, frequently seen co-deleted with the *CDKN2A* and *CDKN2B* tumor suppressor genes in a large cohort of 2000 breast tumors[50]. While the biological significance of p.K71R in *MTAP* and other novel variations in cancer-associated genes is currently unclear, overexpression of novel variant over reference alleles points to a possible contribution to tumor initiation or progression. Since these variants have not been previously reported, they are not likely to be present in a homozygote state at the genomic level, and their allelic dominance may indicate expression or growth advantage, as well as potential loss of heterozygosity. Because such events may drive or contribute to cancer, a systematic investigation of allelic dominance of novel variants across larger expression sets is needed.

In addition, we also identified a higher frequency (compared to non-breast cancer populations) of previously reported variants in many genes, including breast cancer-associated genes, such as *PTPN12*, *PRCP*, *PLEC, DSG2* and *ERBB2IP*. Estimation of the prevalence of such variants in larger breast cancer cohorts is needed as it

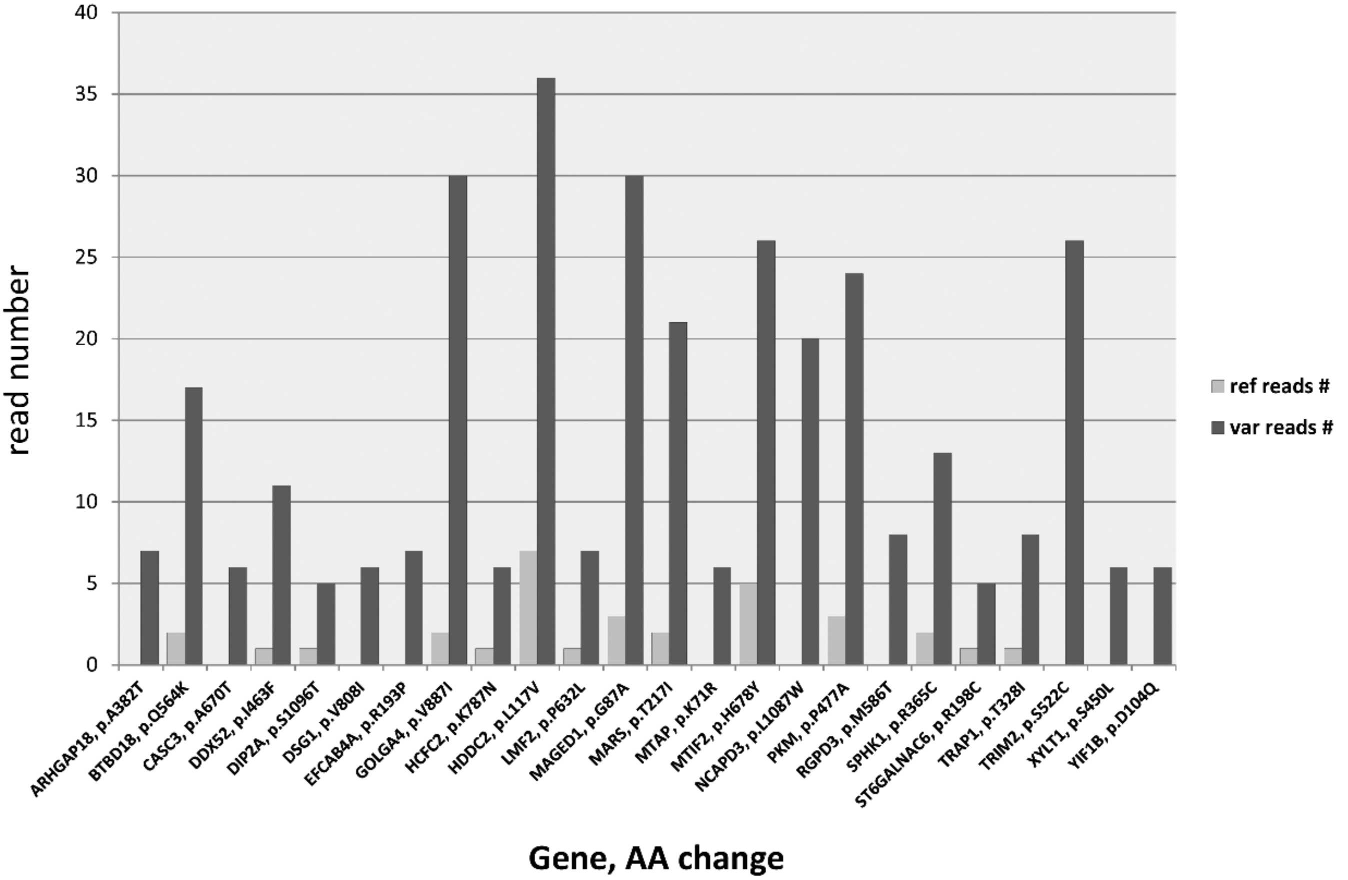

Figure 4 | Allele preferentially expressed novel missense variants through estimation of the ratio of reference and variant reads. The Variant-to wild type allele ration was estimated for all the novel SNPs called by 6 or more reads harboring the variant nucleotide. Fifty seven novel SNPs were called by variant reads only (i.e. no reference call was present at the corresponding position), and additional 53 showed higher than 5-fold number of variant calling reads over the wild type. Among the most preferentially expressed novel SNPs were novel missense variants in previously linked to breast cancer genes such as methyl-thioadenosine phosphorylase *MTAP* (p.K71R), and melanoma antigen *MAGED1* (p.G87A).

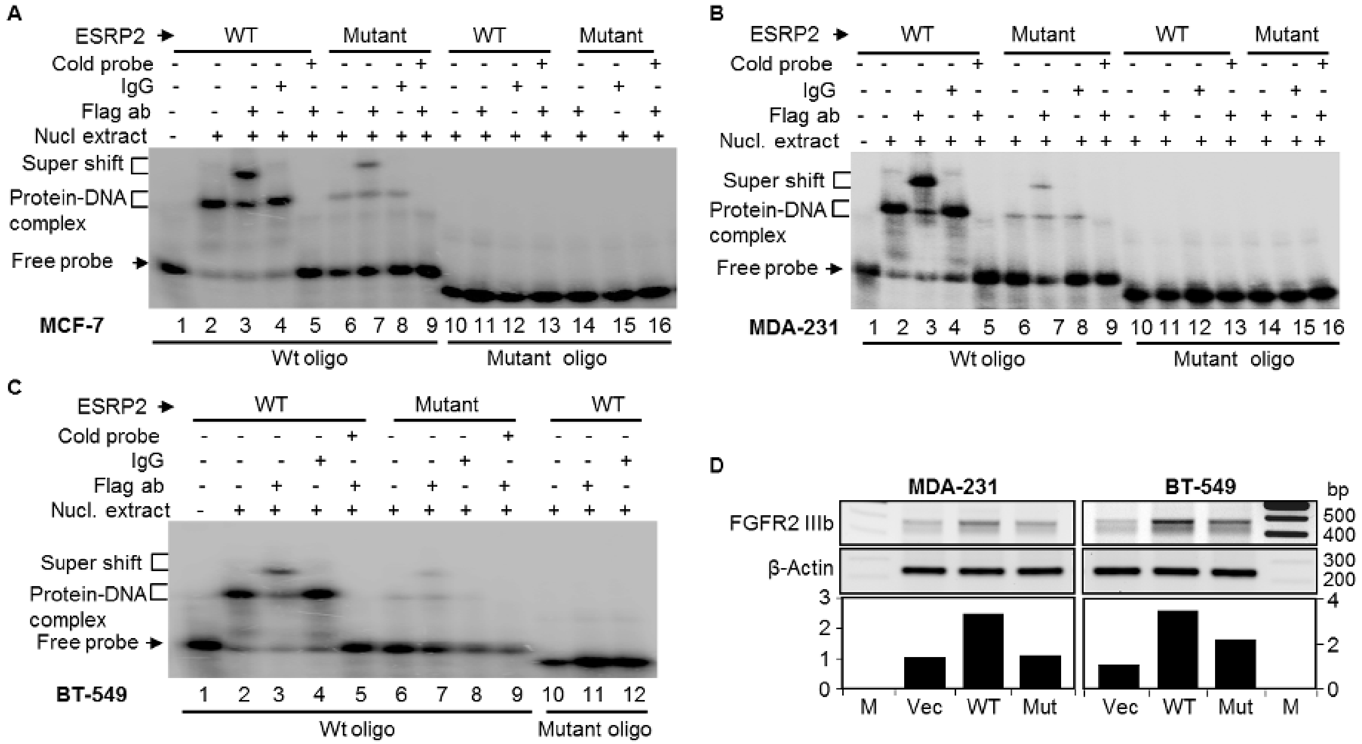

Figure 5 | **EMSA of ESRP2 interaction with *FGFR2* ISE/ISS-3 cis-regulatory motif pre-RNA.** The R353Q mutation in ESRP2 compromises FGFR2-IIIb expression. Vector (PIBX-CFF-B), ESRP2^{wt} or ESRP2^{R353Q} were transiently transfected in breast cancer cell lines: MCF-7 (A), MDA-MB-231 (B) and BT-549 (C). RNA binding of ESRP2^{wt} or ESRP2^{R353Q} is shown. The incubated samples were resolved on 6% native-PAGE gel and detected by Phosphor imager. D) R353Q mutation in ESRP2 compromises FGFR2-IIIb expression. Vector (PIBX-CFF-B), ESRP2-Wt or ESRP2-Mut (R353Q) were transiently transfected in mesenchymal breast cancer cell lines, MDA-MB-231 and BT-549. FGFR-IIIb was detected by RT-PCR. The bands were quantified and normalized by the actin band intensities.

may indicate contribution to genetic risk or co-existence with causative mutations. Although this analysis holds promising potential to identify overrepresented alleles, it is important to take into account that transcriptome sequencing variant calls differs from the exome sequencing in allelic representation of homo- and heterozygote state (i.e. number of alleles). While homo vs. heterozygosity on transcriptome level provides an additional layer of information on the potential functionality of these variants, the results should be used only after confirmation by independent studies. Nevertheless, statistical confidence may be increased for SNPs in which the difference is achieved through the analysis of high number of samples rather than homozygote appearance, such as *PTPN12* and *DSG2* (see Table 2). Such prevalent variants in genes implicated in breast cancer are worthy of investigation in independent breast cancer datasets.

Similarly to the above discussed prevalence of mutant reads in breast cancer-associated variants, GWAS associated SNPs in our set also showed high prevalence of homozygous vs. heterozygous calls. This overall prevalence of variant over reference reads for variant positions in cancer implicated genes, needs to be further investigated as potential indicator of mechanistic implications, such as loss of heterozygosity or preferential allelic expression. As the information content of the transcriptome as a common denominator combining frequency and expression data is emerging, large scale studies are expected to enlighten the feasibility and the information value of these types of analyses[32].

Finally, we selected a rare, predicted to be protein damaging missense substitution from our dataset – p.R353Q in the splicing coordinator *ESRP2* - to demonstrate *in vitro* the effect of the p.R353Q substitution on the *ESRP2* protein function. We were able to show that the replacement of the arginine 353 with the variant glutamine leads to a significant reduction of the binding ability of ESRP2 to *FGFR2* pre-mRNA. Thus, this could potentially affect epithelial-to-mesenchymal transition programs.

Overall, our analysis identified enrichment of variants known to be implicated in breast cancer as well as novel and rare variants in genes associated with breast cancer in our set of 17 breast cancer samples. Further, the within-individual exploration of the variance showed multiple disease associating variants in most of the individuals, and points to the need for estimation of cumulative action of genetic alterations. This study reports an initial collection of variants that are expressed across the breast cancer transcriptome, including novel and reported mutations in their allelic abundance and copresence with other variants. In addition to providing an overall variation landscape of the breast cancer transcriptome, such as expressed SNP density and deleterious variants scaffold, we exemplify different analytical strategies to search for molecular interactions and regulatory networks potentially implicated in breast tumorigenesis. Compared to exome and genome studies, transcriptome exploration provides higher information content through the estimation of the expression abundance, in the immediate context of allelic prevalence and co-existence with expression and splicing features[54,55]. It is essential to keep in mind however that the transcriptome only captures a snap shot and further functional characterization of the observed molecular features is needed to prove disease-causative relationships. Nevertheless, our study provides an important breast cancer transcriptome dataset for further explorations on either high-throughput or individual gene/protein scale.

Methods

Human patient samples. The human breast cancer tissue RNA samples were provided by Dr. Suzanne Fuqua (Baylor College of Medicine). All of the human samples were used in accordance with the IRB procedures of Baylor College of

Medicine and Dana-Farber Cancer Institute and Harvard Medical School, respectively. The breast tumor types, TNBC, Non-TNBC and HER2-positive, were classified on the basis of RNA sequencing FPKM abundance and immunohistochemical and qRT-PCR classification (data not shown) as previously described[54,55]. All breast cancer patients were from European descent.

Illumina genome sequencing RNA sequencing library preparation. Large and small ribosomal RNA (rRNA) was removed from total RNA using RiboMinus Eukaryote Kit (Invitrogen, Carlsbad, CA). Five micrograms of total RNA were hybridized to rRNA-specific biotin labeled probes at 70°C for 5 minutes. The rRNA-probe complexes were then removed by streptavidin-coated magnetic beads. The rRNA-free transcriptome RNA was concentrated by ethanol precipitation. The cDNA synthesis and DNA library construction for all the seventeen samples were performed as described[54,55].

Read alignment and transcript assembly. The paired end raw reads were aligned using the TopHat version 1.2.0 that allows two mismatches in the alignment. The aligned reads were assembled into transcripts using cufflinks version 2.0.0. The alignment quality and distribution of the reads were estimated using SAM tools. From the aligned reads, *de novo* transcript assembly was performed to capture the major splice rearrangements and novel variations that occur in the transcriptomes of TNBC, Non-TNBC and HER2- positive breast cancers in comparison to NBT using cufflinks version 1.3.036. The cuffcompare program was used to identify transcripts that are identical to the reference human genome (the Ensembl GRCh37.62 B (hg19) reference genome). Further analysis and novel isoform call was performed through the reconstructed transfrags that comprise novel splice junctions and share at least one splice junction with a reference transcript. The very low abundant transcripts were identified by binning the transcripts according to their FPKM and the transcripts with FPKM below 0.3 were eliminated from further analysis. All the analyses presented in this manuscript are performed using two categories of transcripts: transcripts that are identical to reference and transcripts that comprise novel junctions. The global statistics, which includes the distributions of FPKM scores across samples and the dendogram that shows the relationship between the samples based on the reconstructed transcripts, were analyzed using cummeRbund package of cufflinks suite of programs. The average exon number was in the reassembled transcripts is comparable to the human genome reference average. To annotate novel splice events, we used Multivariate Analysis of Transcript Splicing (MATS). Additionally, for consistency checking and independent validation we used an in-house built program (http://ccb.jhu.edu/software/ASprofile/) to compare the exon models between isoforms assembled with the program cufflinks for the normal and cancer samples. As mentioned earlier, only the isoforms that are similar to reference and isoforms that comprise novel splice junctions were considered. We determined the splicing differences indicative of exon inclusion, exclusion, alternative 5′, 3′, and intron retention events.

Variants call and annotation. Variants calls were obtained using Mpileup utility of SAMTools (http://samtools.sourceforge.net/mpileup.shtml). Base Alignment Quality was used to score the variant call. Consensus calling is done using bcftools. Maximum depth call was set at 100000. The variants were annotated using SeattleSeq Annotation Tools version 8.01, dbSNP build 137 (http://snp.gs.washington.edu/SeattleSeqAnnotation137/).

Sanger sequencing. First-strand cDNA was synthesized with SuperScript III reverse transcriptase (Invitrogen, Inc) using 1 μg of total RNA and mixture of oligo dT primer and random hexamers. For selected variants, cDNA primers flanking the variant position were designed using Primer3[56] and in RT-PCR to amplify the region of interest. The products were separated on 1% agarose gel, excited and purified using QIAquick Gel Extraction Kit (Qiagen, Inc.) according to the manufacturer instructions. The purified fragments were subjected to bi-directional Sanger sequencing with the forward and the reverse primer used for the amplification.

Statistics. To test if the distribution of variant alleles differed between our group and non-breast cancer populations, we applied chi-square test (2×2 tables). All the values were subjected to Yates correction for contingency to prevent overestimation of significance; p values below 0.05 were considered significant.

EMSA. To determine if the R353Q substitution affects the ability of ESRP2 to bind its substrate we used wild type and mutant FLAG-tagged ESRP2 ORF introduced in PBIX as previously described[52]. Three cell lines MCF-7, MDA-231 and BT-549 were transfected cell lines using FuGENE® Transfection Reagents (Promega, Inc.) according to the manufacturer recommendations. Nuclear extracts were prepared using a Nonidet P-40 lysis method. RNA oligos of ISE/ISS-3 were end labeled with using the annealed [γ-^{32}P] ATP in a 20 μl reaction mixture for 15 min at room temperature. RNA probes were incubated with respective nuclear extracts. Samples were run on a non-denaturing 6% polyacrylamide gel and imaged by autoradiography. Specific competitions were performed by adding a 100-molar excess of unlabeled probe to the incubation mixture and supershift Electrophoretic mobility shift assay (EMSA) were performed using FLAG antibody (Sigma-Aldrich).

1. DeSantis, C., Siegel, R., Bandi, P. & Jemal, A. Breast cancer statistics, 2011. *CA Cancer. J. Clin.* **61**, 409–418 (2011).
2. Claus, E. B., Risch, N. J. & Thompson, W. D. Age at onset as an indicator of familial risk of breast cancer. *Am. J. Epidemiol.* **131**, 961–972 (1990).
3. Peto, J. & Mack, T. M. High constant incidence in twins and other relatives of women with breast cancer. *Nat. Genet.* **26**, 411–414 (2000).
4. Lynch, H. T. *et al.* Genetic predisposition to breast cancer. *Cancer* **53**, 612–622 (1984).
5. Turnbull, C. & Rahman, N. Genetic predisposition to breast cancer: past, present, and future. *Annu. Rev. Genomics Hum. Genet.* **9**, 321–345 (2008).
6. Wooster, R. *et al.* Identification of the breast cancer susceptibility gene BRCA2. *Nature* **378**, 789–792 (1995).
7. Stratton, M. R. & Rahman, N. The emerging landscape of breast cancer susceptibility. *Nat. Genet.* **40**, 17–22 (2008).
8. Thompson, D., Easton, D. F. & Breast Cancer Linkage Consortium. Cancer Incidence in BRCA1 mutation carriers. *J. Natl. Cancer Inst.* **94**, 1358–1365 (2002).
9. Hopper, J. L. & Carlin, J. B. Familial aggregation of a disease consequent upon correlation between relatives in a risk factor measured on a continuous scale. *Am. J. Epidemiol.* **136**, 1138–1147 (1992).
10. Parkin, D. M., Pisani, P. & Ferlay, J. Global cancer statistics. *CA Cancer. J. Clin.* **49**, 33–64, 1 (1999).
11. Guilford, P. *et al.* E-cadherin germline mutations in familial gastric cancer. *Nature* **392**, 402–405 (1998).
12. Hemminki, A. *et al.* A serine/threonine kinase gene defective in Peutz-Jeghers syndrome. *Nature* **391**, 184–187 (1998).
13. Jenne, D. E. *et al.* Peutz-Jeghers syndrome is caused by mutations in a novel serine threonine kinase. *Nat. Genet.* **18**, 38–43 (1998).
14. Malkin, D. *et al.* Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. *Science* **250**, 1233–1238 (1990).
15. Meijers-Heijboer, H. *et al.* Low-penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in noncarriers of BRCA1 or BRCA2 mutations. *Nat. Genet.* **31**, 55–59 (2002).
16. Nelen, M. R. *et al.* Localization of the gene for Cowden disease to chromosome 10q22–23. *Nat. Genet.* **13**, 114–116 (1996).
17. Nelen, M. R. *et al.* Germline mutations in the PTEN/MMAC1 gene in patients with Cowden disease. *Hum. Mol. Genet.* **6**, 1383–1387 (1997).
18. Rahman, N. *et al.* PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nat. Genet.* **39**, 165–167 (2007).
19. Renwick, A. *et al.* ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. *Nat. Genet.* **38**, 873–875 (2006).
20. Seal, S. *et al.* Truncating mutations in the Fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. *Nat. Genet.* **38**, 1239–1241 (2006).
21. Craig, D. W. *et al.* Genome and transcriptome sequencing in prospective metastatic triple-negative breast cancer uncovers therapeutic vulnerabilities. *Mol. Cancer. Ther.* **12**, 104–116 (2013).
22. Ha, G. *et al.* Integrative analysis of genome-wide loss of heterozygosity and monoallelic expression at nucleotide resolution reveals disrupted pathways in triple-negative breast cancer. *Genome Res.* **22**, 1995–2007 (2012).
23. Hartmaier, R. J., Priedigkeit, N. & Lee, A. V. Who's driving anyway? Herculean efforts to identify the drivers of breast cancer. *Breast Cancer Res.* **14**, 323 (2012).
24. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–1111 (2009).
25. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
26. DePristo, M. A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
27. Le, S. Q. & Durbin, R. SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Res.* **21**, 952–960 (2011).
28. Shepherd, R. *et al.* Data mining using the Catalogue of Somatic Mutations in Cancer BioMart. *Database (Oxford)* **2011**, bar018 (2011).
29. Rafiq, S. *et al.* Identification of inherited genetic variations influencing prognosis in early-onset breast cancer. *Cancer Res.* **73**, 1883–1891 (2013).
30. Jara, L. *et al.* Genetic variants in FGFR2 and MAP3K1 are associated with the risk of familial and early-onset breast cancer in a South-American population. *Breast Cancer Res. Treat.* **137**, 559–569 (2013).
31. Guo, H. *et al.* A common polymorphism near the ESR1 gene is associated with risk of breast cancer: evidence from a case-control study and a meta-analysis. *PLoS One* **7**, e52445 (2012).
32. Li, Q. *et al.* Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell* **152**, 633–641 (2013).
33. Palmer, J. R. *et al.* Genetic susceptibility loci for subtypes of breast cancer in an African American population. *Cancer Epidemiol. Biomarkers Prev.* **22**, 127–134 (2013).
34. Durocher, F. *et al.* Comparison of BRCA1 polymorphisms, rare sequence variants and/or missense mutations in unaffected and breast/ovarian cancer populations. *Hum. Mol. Genet.* **5**, 835–842 (1996).
35. Schoumacher, F. *et al.* BRCA1/2 mutations in Swiss patients with familial or early-onset breast and ovarian cancer. *Swiss Med. Wkly.* **131**, 223–226 (2001).
36. Fletcher, O. *et al.* Missense variants in ATM in 26,101 breast cancer cases and 29,842 controls. *Cancer Epidemiol. Biomarkers Prev.* **19**, 2143–2151 (2010).
37. Gutierrez-Enriquez, S. *et al.* Functional consequences of ATM sequence variants for chromosomal radiosensitivity. *Genes Chromosomes Cancer* **40**, 109–119 (2004).

38. Stredrick, D. L. *et al*. The ATM missense mutation p.Ser49Cys (c.146C > G) and the risk of breast cancer. *Hum. Mutat.* **27**, 538–544 (2006).
39. Blons, H. *et al*. Genome wide SNP comparative analysis between EGFR and KRAS mutated NSCLC and characterization of two models of oncogenic cooperation in non-small cell lung carcinoma. *BMC Med. Genomics* **1**, 25-8794-1-25 (2008).
40. Baron Gaillard, C. L. *et al*. Hook2 is involved in the morphogenesis of the primary cilium. *Mol. Biol. Cell* **22**, 4549–4562 (2011).
41. Miyamoto-Sato, E. *et al*. A comprehensive resource of interacting protein regions for refining human transcription factor networks. *PLoS One* **5**, e9289 (2010).
42. Sun, T. *et al*. Activation of multiple proto-oncogenic tyrosine kinases in breast cancer via loss of the PTPN12 phosphatase. *Cell* **144**, 703–718 (2011).
43. Albeck, J. G. & Brugge, J. S. Uncovering a tumor suppressor for triple-negative breast cancers. *Cell* **144**, 638–640 (2011).
44. Duan, L. *et al*. Prolylcarboxypeptidase regulates proliferation, autophagy, and resistance to 4-hydroxytamoxifen-induced cytotoxicity in estrogen receptor-positive breast cancer cells. *J. Biol. Chem.* **286**, 2864–2876 (2011).
45. Niwa, T. *et al*. BRCA2 interacts with the cytoskeletal linker protein plectin to form a complex controlling centrosome localization. *Cancer. Sci.* **100**, 2115–2125 (2009).
46. Song, C., Wang, W., Li, M., Liu, Y. & Zheng, D. Tax1 enhances cancer cell proliferation via Ras-Raf-MEK-ERK signaling pathway. *IUBMB Life* **61**, 685–692 (2009).
47. Davies, E. *et al*. The role of desmoglein 2 and E-cadherin in the invasion and motility of human breast cancer cells. *Int. J. Oncol.* **11**, 415–419 (1997).
48. Cooper, C. *et al*. Increasing the relative expression of endogenous non-coding Steroid Receptor RNA Activator (SRA) in human breast cancer cells using modified oligonucleotides. *Nucleic Acids Res.* **37**, 4518–4531 (2009).
49. Yau, C. *et al*. A multigene predictor of metastatic outcome in early stage hormone receptor-negative and triple-negative breast cancer. *Breast Cancer Res.* **12**, R85 (2010).
50. Curtis, C. *et al*. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
51. Tavanez, J. P. & Valcarcel, J. A splicing mastermind for EMT. *EMBO J.* **29**, 3217–3218 (2010).
52. Warzecha, C. C., Sato, T. K., Nabet, B., Hogenesch, J. B. & Carstens, R. P. ESRP1 and ESRP2 are epithelial cell-type-specific regulators of FGFR2 splicing. *Mol. Cell* **33**, 591–601 (2009).
53. Warzecha, Z. *et al*. Therapeutic effect of ghrelin in the course of cerulein-induced acute pancreatitis in rats. *J. Physiol. Pharmacol.* **61**, 419–427 (2010).
54. Eswaran, J. *et al*. RNA sequencing of cancer reveals novel splicing alterations. *Sci. Rep.* **3**, 1689 (2013).
55. Eswaran, J. *et al*. RNA sequencing of cancer reveals novel splicing alterations. *Sci. Rep.* **3**, 1689 (2013).
56. Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.* **132**, 365–386 (2000).

Acknowledgements

This work is supported by the McCormick Genomic and Proteomics Center. The authors wish to thank Sucheta Godbole, Jeyanthy Eswaran and Dinesh Cyanam for their technical assistance in this project.

Author contributions

R.K. directed the project and designed the experiments. A.H. and R.K. designed the analyses, analyzed the data, and wrote the manuscript. P.M. and S.B.P. carried out dry-lab and validation studies, respectively. S.D.R. and K.O. performed EMSA and RT-PCR experiments. S.C., R.P., L.C., S.A.W.F., M.T. and S. S. provided reagents and biological insights. S.S.N. carried out the initial RNA-sequencing experiments and editing.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Horvath, A. *et al*. Novel Insights into Breast Cancer Genetic Variance through RNA Sequencing. *Sci. Rep.* **3**, 2256; DOI:10.1038/srep02256 (2013).