Mapping of *cis*-regulatory variants by differential allelic expression analysis identifies candidate risk variants and target genes of 27 breast cancer risk loci

Joana M. Xavier[1], Ramiro Magno[1], Roslin Russell[2,‡], Bernardo P. de Almeida[3,4,#], Ana Jacinta-Fernandes[3], André Duarte[1], Mark Dunning[2,§], Shamith Samarajiwa[5], Martin O'Reilly[2,±], Cátia L. Rocha[3,*], Nordiana Rosli[3,6,+], Bruce A. J. Ponder[2], Ana Teresa Maia.[1,3,7,¶]

1 – Center for Research in Health Technologies and Information Systems (CINTESIS), Universidade do Algarve, Faro, Portugal;
2 – Cambridge Institute - CRUK/University of Cambridge, Cambridge, United Kingdom;
3 – Faculdade de Medicina e Ciências Biomédicas (FMCB), Universidade do Algarve, Faro, Portugal;
4 – Instituto de Medicina Molecular, Faculdade de Medicina, Universidade de Lisboa, Lisboa, Portugal;
5 – Medical Research Council (MRC) Cancer Unit, University of Cambridge, Hutchison/MRC Research Centre, Cambridge, United Kingdom;
6 – Training Division, Ministry of Health Malaysia, Putrajaya, Malaysia;
7 – Algarve Biomedical Centre, Universidade do Algarve, Faro, Portugal.

¶ Corresponding Author: atmaia@ualg.pt

Current Addresses:
‡ Department of Genetics, University of Cambridge, Cambridge, United Kingdom
# Research Institute of Molecular Pathology (IMP), Vienna Biocenter (VBC), Vienna, Austria
§ Sheffield Institute for Translation Neuroscience (SITraN), University of Sheffield, Sheffield, United Kingdom
± MRC Toxicology Unit, Cambridge, United Kingdom
* Instituto de Saúde Ambiental (ISAMB), Faculty of Medicine, University of Lisbon, Lisbon, Portugal
+ Biometrology Group, Division of Chemical and Biological Metrology, Korea Research Institute of Standards and Science, Daejeon, South Korea

ABSTRACT

Background

Breast cancer (BC) genome-wide association studies (GWAS) have identified hundreds of risk-loci that require novel approaches to reveal the causal variants and target genes within them. As causal variants are most likely regulators of gene expression, we hypothesize that their identification is facilitated by pinpointing the variants with greater regulatory potential within risk-loci.

Methods

We performed genome-wide differential allelic expression (DAE) analysis using microarrays data from 64 normal breast tissue samples. Then, we mapped the variants associated with DAE (daeQTLs) and intersected these with GWAS data to reveal candidate risk regulatory variants. Finally, we validated our approach by functionally analysing the 5q14.1 breast cancer risk-locus.

Results

We found widespread gene expression regulation by *cis*-acting variants in breast tissue, with 80% of coding and non-coding expressed genes displaying DAE (daeGenes). We identified over 23K daeQTLs for 2753 (16%) daeGenes, including at 154 known BC risk-loci. And in 31 of these risk-loci, we found risk-associated variant(s) and daeQTLs in strong linkage disequilibrium suggesting that the risk-causing variants are *cis*-regulatory, and in 27 risk-loci we propose 37 candidate target genes. As validation, we identified five candidate causal variants at the 5q14.1 risk-locus targeting the *ATG10, RPS23,* and *ATP6AP1L* genes, likely via modulation of miRNA binding, alternative transcription, and transcription factor binding.

Conclusion

Our study shows the power of DAE analysis and daeQTL mapping to identify causal regulatory variants and target genes at BC risk loci, including those with complex regulatory landscapes, and provides a genome-wide resource of variants associated with DAE for future functional studies.

Keywords: cis-regulation; polymorphism; cancer predisposition; breast cancer

INTRODUCTION

Genome-wide association studies (GWAS) for breast cancer (BC) have identified hundreds of risk-associated loci and have generated long lists of candidate ones requiring further validation [1]. Still, the identification of the causal variants and their target genes, as well as understanding the underlying biological mechanisms, remain challenging. This is because disease risk-loci often have many variants in high linkage disequilibrium (LD) with the risk-associated variant, harbour multiple genes and mainly fall in non-coding regions of the genome [2]. However, the overrepresentation of potential causal variants located at active gene regulatory regions [3,4] indicates that BC genetic predisposition is strongly influenced by variants regulating gene expression levels, both proximally and at long-range [5–11]. These variants have commonly been mapped by expression quantitative trait loci (eQTL) analysis, but this approach is impacted by the effects of negative feedback control and environmental factors [12]. An increasingly popular alternative approach is to detect imbalances in allelic transcript levels, or differential allelic expression (DAE). By comparing the relative expression of the two alleles in a heterozygous individual, each allele will serve as an internal standard for the other, thus controlling for trans-regulatory and environmental factors affecting both alleles [13,14]. Consequently, this is a direct indicator of regulatory variants acting in *cis* (*cis*-acting regulatory SNPs or rSNPs).

Given the importance of *cis*-regulatory variants for BC susceptibility, a genome-wide map of *cis*-regulatory variants would be key to interpreting GWAS results and identifying causal variants of risk. Studies in various healthy tissues showed that DAE is a relatively common event [13,15–19]. Given that gene expression regulation is tissue-specific, it is important to perform these studies in the tissue from which the disease arises, namely normal breast tissue. Although others have used allelic expression analysis to identify BC risk, this was either carried out in tumour tissue or lymphoblastoid cells [20,21]. This study proposes an integrative approach to identify causal variants of risk that have a *cis*-regulatory role (Figure 1): to combine GWAS results with SNPs associated with DAE levels in normal breast tissue. Hence, we first carried out DAE analysis in normal breast tissue samples, at a genome-wide level, then mapped the candidate risk regulatory variants for GWAS loci and finally functionally unveiled the mechanisms underlying BC risk at a selected locus.

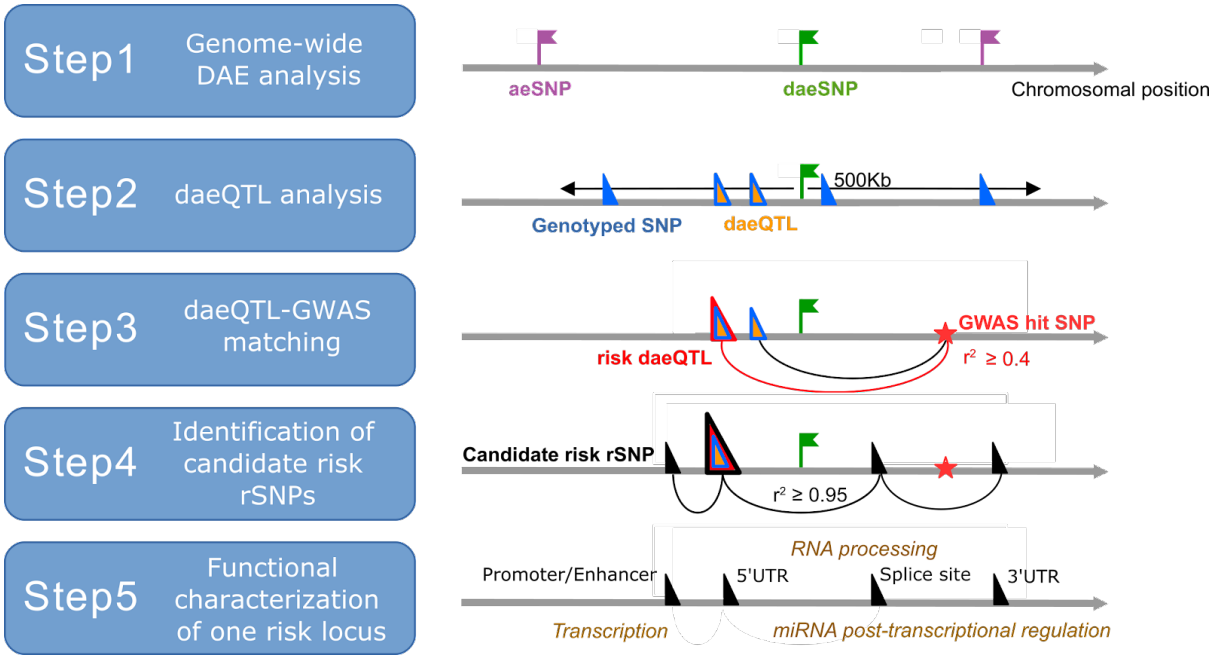

Figure 1 - Framework of the strategy used for the identification of causal variants and target genes associated with breast cancer risk. Legend: *aeSNP* - a SNP that passed quality control and at which allelic expression (AE) was measured; daeSNP - a aeSNP showing differential AE (DAE); Genotyped SNP - a SNP with genotype information (either genotyped in the study or imputed) and tested for association with AE ratios; daeQTL - a SNP associated with AE ratios measured for a daeSNP; risk_daeQTL - a daeQTL with an $r^2 \geq 0.4$ with a GWAS hit variant; candidate risk rSNP - a variant with an $r^2 \geq 0.95$ with the risk daeQTL.

MATERIAL AND METHODS

Data set

Seventy-six samples of normal breast tissue were collected from women submitted to a reduction mastectomy, for reasons unrelated to cancer, at Addenbrooke's Hospital, Cambridge, United Kingdom. Samples were collected with approval from Addenbrooke's Hospital Local Research Ethics Committee (REC reference 06/Q0108/221). DNA and total RNA were extracted as previously described [22].

Genome-wide DAE analysis

Seventy-six DNA and cDNA samples derived from total RNA from a given individual were run on Illumina Infinium Exon510S-Duo arrays generating 304 idat files for both red and green channels [23]. These exon-centric microarrays contain probes for 511,354 SNPs, with more than 60% of the markers located within 10kb of a gene and targeting more than 99.9% of human RefSeq genes. This data set is available from Gene Expression Omnibus (GEO, www.ncbi.nlm.nih.gov/geo/) under accession number GSE35023. The sample filtering and normalization were performed as described previously and 12 samples were removed from further analysis [23]. After normalization and before allelic expression analysis, extensive quality control was performed and only SNP respecting the following filters were further analysed: (1) to remove non-expressed SNPs, a minimum cut-off of average $\log_2$ RNA intensity values ≥ 9.5 for each probe was applied; (2) to verify the efficacy of allelic discrimination at the RNA level by the microarray probes, RNA $\log_2$ ratios of the signal of each SNP's two probes was compared between heterozygous (AB) and homozygous groups (AA and BB) and minimum two-sample t-test p-value ≤ 0.05 was applied; (3) to guarantee high quality genotyping data, a minimum call rate ≥ 90%, a Hardy-Weinberg equilibrium p-value > 1.0E-05 and at least five heterozygotes were requested for each SNP; and finally, (4) only SNPs uniquely mapped in the genome, not flagged as suspected in dbSNP149 GRCh38p7 and located in autosomes were kept.

Allelic expression was measured in the filtered dataset of SNPs and samples, in a varying number of individuals heterozygous (AB) for each transcribed SNP (aeSNP). As the cDNA was prepared from total RNA, without enrichment for poly-A mRNAs, AE was measured for variants located in fully processed mRNAs and in unspliced primary transcripts. Allelic expression ratios (AE ratios) were defined as the $\log_2$ of the ratio between the levels of allele A transcript and the levels of allele B transcript (heterozygote ratio), normalized by the same heterozygote ratio calculated for genomic DNA (gDNA) (Figure S1), to account for copy number variation and correct for technical biases. DAE was defined when AE ratios were greater than 0.58 or less than -0.58 (corresponding to the $\log_2$ of 1.5-

fold difference between alleles), and aeSNPs were considered to display DAE (daeSNPs) when at least 10% and three heterozygous samples displayed DAE (Figure S1, Figure 1 - step1). Genes with at least one daeSNP were henceforth denominated daeGenes.

aeSNPs for which monoallelic expression profiles with two distinct groups of heterozygotes solely expressing one allele or the other were identified, suggesting imprinting or random monoallelic expression [24] were classified as being mono-allelic expressed (maeSNPs). Genes with at least one maeSNP were labelled as maeGenes.

Validation of nine daeSNPs was performed by Taqman® PCR technology, as described before [22], in 25 independent normal breast tissue samples heterozygous for a variable number of individual per SNP, using the following Taqman® Genotyping Assays predesigned by Applied Biosystems: C___8354687_10; C__29939330_20; C__31232634_10; C___3133316_10; C__11844169_10; C___2627792_10; C___1517694_1_; C____787630_20; C___3108259_10.

Annotation of variants

Variants were annotated according to hg38/GRCh38 with biomaRt v 2.40.5, and Ensembl IDs without the HGCN symbol were excluded from the tables but were considered for statistics reporting the number of genes. To test whether classes of consequence type and gene biotype were over-represented (i. e. enriched) in the list of daeSNPs we applied a one-tailed Fisher's exact test (alternative = greater).

Information from imprinted genes was retrieved from a comprehensive study of genomic imprinted in breast [25] and from geneimprint database (http://www.geneimprint.com) searching for Imprinted Genes: by Species: Human.

Genotype imputation

Imputation was run on the Illumina Exon 510 Duo germline genotype data from the 64 samples that passed microarrays quality control filters. Before imputation data, 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-value < 1.0E-05 were excluded from the analysis. Imputation was performed using MACH1.0 [26] and the phased haplotypes for HapMap3 release (HapMap3 NCBI Build 36, CEU panel - Utah residents with Northern and Western European ancestry) as reference panel. We applied the recommended two-step imputation process: model parameters (crossover and error rates) were estimated before imputation using all haplotypes from the study subjects and running 100 iterations of the Hidden Markov Model (HMM) with the command option - greedy and -r 100. Genotype imputation was then carried out using the model parameter estimates from the previous round with command options of -greedy, -mle, and -mldetails specified.

Imputation results were assessed by the platform-specific measures of imputation uncertainty for each SNP (rq Score) and filtered for an rq-score≥0.3, as suggested in the author webpage (http://csg.sph.umich.edu/abecasis/mach/tour/) and MAF ≥ 0.01.

Candidate rSNPs mapping

Mapping of candidate rSNPs associated with the DAE observed - henceforth designated as daeQTLs (differential allelic expression quantitative trait loci) (Figure S1, Figure 1 - Step2) - took into consideration the pattern of AE ratios distribution displayed by each daeSNPs, as this is highly dependent on the LD between the daeSNP and the rSNP acting upon the gene [27]. Therefore, to identify daeQTLs we followed two mapping approaches:

- Mapping Approach 1 - applied to daeSNPs for which all heterozygous samples displayed DAE - all genotyped/imputed SNPs located within ±500Kb of the daeSNP equally heterozygous for all samples were considered daeQTLs.
- Mapping Approach 2 - applied to daeSNPs for which only some heterozygous samples displayed DAE. The absolute values of AE ratios at the daeSNP were used in an asymptotic one-way Fisher-Pitman test to compare heterozygotes and the combined homozygotes for the candidate rSNP (AB vs AA+BB, Figure S1). Genotyped/imputed SNPs located within 500Kb of the daeSNP which had at least 2 heterozygous and 2 homozygous samples were selected for analysis. The null hypothesis tested was that the mean of the heterozygotes group was equal or smaller than that of the combined heterozygotes. The premise for this was that only samples that are heterozygous for the candidate rSNP will show differences between the expression of the two alleles at the daeSNP, i.e., higher absolute AE ratios. We combined the two groups of homozygotes as no phased data was available for these samples, and therefore a logistic regression could not be applied to three genotype groups separately. Tests were applied using a one way-test implemented in the coin package [28]. P-values were adjusted with the Benjamini-Hochberg method [29], using all 18572521 daeSNP/tested SNP pairs, with the distance between them as a covariate (package ihw, R) [30] and reported as significant when the false discovery rate was below 10%.

Breast cancer GWAS data retrieval

Nine hundred and sixty-eight GWAS-significant risk-associated SNPs for BC published until April 2018 were retrieved from the NHGRI-EBI Catalog of published genome-

wide association studies (GWAS Catalog) [31] using the gwasrapidd R package [32]. Filters included a significance level cut-off of p-value ≤ 1.0E-05 and the reported traits: “Breast cancer", "Breast cancer (early onset)", "Breast cancer (estrogen-receptor negative)", "Breast cancer (male)", "Breast Cancer in BRCA1 mutation carriers", "Breast cancer in BRCA2 mutation carriers", "breast cancer male", "Breast cancer and/or colorectal cancer). The full list of SNPs is presented in Table S1.

Proxy SNPs retrieval

Variants in LD with index SNPs were retrieved from Ensembl [33] using the function get_ld_variants_by_window() from the ensemblr R package (https://github.com/ramiromagno/ensemblr) using the 1000 GENOMES project data (phase_3) for the CEU population, and a genomic window size of 500kb (250kb upstream and downstream of the queried variant). The r^2 cut-off used varied between 0.2 and 0.95 depending on the analysis and is indicated in each analysis description**.**

Retrieval of previously suggested BC targets genes

Genes previously suggested as targets of *cis*-acting regulatory variation in post-GWAS studies for BC, with extensive fine-scale mapping and *in-silico* prediction or functional analysis, and those classified as *Inquisit 1* by Fachal and colleagues [4] are indicated in Table S2.

GTEx eQTL and gene expression data retrieval

The Genotype-Tissue Expression (GTEx) project identified expression quantitative trait loci (eQTL) using normal mammary tissue samples [34]. eGenes (genes with at least one SNP in cis significantly associated, at a false discovery rate (FDR) of ≤0.05, with expression differences of that gene) and significant variant-gene associations based on permutations were downloaded from GTEx Analysis V8 (dbGaP Accession phs000424.v8.p2, available on 18/07/2019).

All SNP-gene associations tested for breast mammary tissue, including non-significant, together with genes expression levels (TPM) were downloaded from GTEx Analysis V7 (available on 2016-01-15).

Comparison of daeGenes, eGenes and gwasGenes

First, the list of publicly available eGenes was compared with the daeGenes identified in our study, restricting this comparison to genes analysed in both datasets. Then, we investigated the percentage of gwasGenes, defined as genes containing variants that are

in moderate to strong LD (r2≥0.4) with GWAS index SNPs, displaying evidence of *cis*-regulation by either DAE or eQTL analysis.

Functional characterization of candidate risk SNPs

Candidate risk rSNPs at the 5q14.1 locus were examined for potential regulatory potential. Overlap of variants location with epigenetic marks derived from the ENCODE [35] and NIH Roadmap Epigenomics projects data [36], was carried out using the UCSC Genome Browser [37,38], HaploReg v4.1 [39] and RegulomeDB v1.1 [40] tools. Emphasis was given to overlapping with DNase I hypersensitivity sites, H3K4me1, H3K4me3, and H3K27ac histone modifications, and transcription factor (TF) binding identified in normal human mammary epithelial cells (HMECs), normal human mammary fibroblasts (HMFs), two BC cell lines (MCF-7 and T47D) and two normal breast cell lines (BR.MYO and BR.H35).

Allele-specific epigenetic modifications (H3k4me3 and DNAse I), and RNA polymerase II (POL2) and transcription factors (TF) binding with alignment data available in HMEC, MCF-7 and MCF-10A breast cancer cell lines from ENCODE were retrieved and visualized using the Integrative Genomics Viewer (IGV Version 2.3.71) tool [41], to analyse protein-DNA interactions and allelic preferential binding. Differential allelic binding was analysed in heterozygous candidate risk rSNPs located within TF binding peaks in experiments with a read coverage at the SNP site higher than 20. We applied a two-tailed binomial test with the null hypothesis assuming no bias (balanced binding of the protein to the two alleles of the variant). P-value was corrected for multiple testing using the R package qvalue [42]. When multiple tracks for the same SNP, trait and Cell line existed, it was only reported in the main manuscript the p-value for the experiment with higher total read counts.

Prediction of SNP-allele miRNA binding was performed for five SNPs located in *RPS23* 3'UTR using a modified version of TargetScan – a miRNA-binding prediction algorithm, that performs allele-specific queries [43].

Analysis related to alternative transcription was carried out in three ways. Firstly, sQTLseekeR (v1.4) [44] was used to test the association of genetic variants with alternative isoform expression, in both normal breast and tumour tissue, using total read counts derived from RNA-seq data from the TCGA (TCGA-BRCA, hg19) and GTEx (phs000424.v6.p1, hg38) projects. Only *ATG10* displayed sufficient alternative transcription dispersion to allow the sQTL analysis. Additionally, all SNPs within 5kb upstream or downstream of *ATG10* were included in the analysis and not only the 92 candidate risk rSNPs, to increase the stringency of the association exercise. P-values for all SNPs tested for ATG10 sQTL analysis were controlled for multiple testing using a 5% FDR. Correlation analyses between -log10 (FDR q-value) and LD (r^2) with rs7707921 were performed using a Pearson's test.

Then, overlapping of variant location with RNA processing-associated proteins was assessed using CLIP data retrieved from POSTAR2 (http://lulab.life.tsinghua.edu.cn/postar/) [45] and from RBP-Var (http://www.rbp-var.biols.ac.cn/) [46], which additionally informed on riboSNitch potential [47]. Finally, allele-specific RBP binding predictions were performed with RBPmap [48] using the analysed variant flanking sequence (30 nucleotides on each side, with the variant at index 31) using all available human RBP motifs.

Haplotype analysis

Haplotypes in the 5q14.1-14.2 region were analysed on Haploview 4.2 using the imputed genotypes from the 64 normal breast tissue samples [49]. For candidate risk SNPs whose genotype was not possible to determine (because it was neither genotyped nor imputed), a proxy SNP in strong LD ($r^2 \geq 0.95$) was used instead. Haplotype blocks were generated using the default algorithm.

TCGA-BRCA gene expression analysis

Processed gene expression and isoform expression from RNA-Seq data for 113 normal solid tissues and 1102 primary solid tumours from the TCGA-BRCA project, together with corresponding clinical data, were retrieved from the Genomic Data Commons archive using the R package TCGAbiolinks [50] accessed in October 2018. Isoform expression was annotated according to the genome assembly hg19 and total gene expression annotated according to hg38. To compare the mean of the expression of *ATG10* isoforms between normal-solid tissues (normal-matched) and breast tumours we applied two-sample Wilcoxon tests, correcting for multiple testing with the Benjamini and Hochberg (BH) procedure. To correlate gene expression among *ATG10*, *RPS23*, *ATP6AP1L* we applied a Pearson's test. To correlate *ATG10*, *RPS23, and ATP6AP1L* with *MYC* and *MAX* gene expression a Spearman's test was applied instead.

RESULTS

Cis-regulatory variation is common in normal breast tissue

Genome-wide allelic expression (AE) analysis was performed using microarrays data from 64 samples of normal breast tissue. Normalized allelic expression ratios were calculated for SNPs located in coding and non-coding regions, upon filtering for expression level and allelic discrimination potential of the cDNA signals. Overall, we identified 91,467 autosomal allelic-expressed SNPs (aeSNPs) located in 21,527 annotated Ensembl genes (median of three aeSNPs per gene) (Figure S2). 70,489 of the aeSNPs were annotated uniquely to one gene, resulting in 12,331 genes to which we could directly attribute the allelic expression measurements. Unsurprisingly, the number of aeSNPs analysed per gene correlated with the annotated gene length (rho = 0.60, p-value < 2.2e-16, Figure S3).

We found that almost half of the aeSNPs (44,267 out of 91,467) displayed bi-allelic differential expression (daeSNPs, ≥ 1.5-fold difference between alleles) (Table 1, Figure S4, Table S3), while 84 SNPs displayed mono-allelic expression (maeSNPs). Taqman PCR validated seven out of nine daeSNPs (Figure S5) that showed DAE and concordant preferential expression of the same allele (Fisher's exact test p-value > 0.05).

Table 1. Summary of the genome-wide breast tissue allelic expression analysis results

Set of SNPs	n	Ensembl Gene IDs	Mean number of heterozygous samples	Mean number of DAE samples
All aeSNPs	91467	21527	22.1	-
maeSNPs	85	44	22.8	22.8
daeSNPs	44267	17135	24.4	8.7

The daeSNPs distributed uniformly across the genome, with low inter-chromosomal variability (Figure S6), and overlapped 17,134 (80%) annotated genes (daeGenes), of which 6,525 (30%) harboured three or more daeSNPs (Figure 2a, Table 1, Table S3). When considering daeSNPs mapping exclusively to one gene, we could pinpoint 9,841 daeGenes that showed evidence of being under the control of allele-specific cis-acting factors, either genetic or epigenetic. In terms of consistency of DAE detection across length of these genes, we found that in the majority of daeGenes, daeSNPs represented two thirds of aeSNPs (7740 in 17134), with 626 daeGenes presenting imbalances in all the analysed aeSNPs (Figure 2b). Regarding the number of heterozygotes, we found 8.7 samples on average displaying DAE out of the 24.4 analysed per daeSNP (Table1, Table S4). The aeSNPs showed a large distribution of average |AE ratios| centred at 0.48, with 1% aeSNPs

showing average allelic fold changes ranging from 4.7 to 34 (Figure 2c, Table S5). The amplitude of the imbalances measured at aeSNPs correlated negatively with the average expression level of both alleles (rho = -0.4, p-value < 2.2e-16) (Figure 2d), but not with the standard deviation across individuals (Figure S7).The aeSNPs located mainly in intronic and non-coding transcripts regions, but non-daeSNPs and daeSNPs showed differences in class distribution for consequence type, with daeSNPs enriched at unannotated and intronic regions and at non-coding transcripts (p-value < 0.01, Figure 2e). Although the majority of the aeSNPs analysed located in protein coding genes, daeSNPs were relatively more common at non-coding genes and pseudogenes when compared to non-daeSNPs (p-value < 0.01, Figure 2f, Table S3).

Monoallelic expression in breast tissue

Regarding monoallelic expression, maeSNPs were annotated to 44 Ensembl genes (Table 1**,** Table S6, Figure S8), the majority of which were previously reported as imprinted in breast tissue (e.g., *IGF2* or *ZDBF2*), or in other tissues (e.g., *KCNQ1*, *KCNQ1OT1*, *RTL1*, *NAA60*, *ZIM2,* and *L3MBTL1*), validating our AE analysis. Interestingly, we detected maeSNPs in a region containing the lncRNA *MEG9* and a cluster of miRNAs genes that had only previously been reported as imprinted in non-human species [51–53]. Additionally, we found unreported monoallelic expression at an intergenic region (22q11.23), suggesting the existence of unannotated transcripts at this region. Notably, we observed two separate groups of heterozygotes preferentially expressing opposite alleles of rs17122278, an intronic variant of *ARCN1*, suggesting the latter as a candidate novel mono-allelic expressed protein-coding gene in breast tissue.

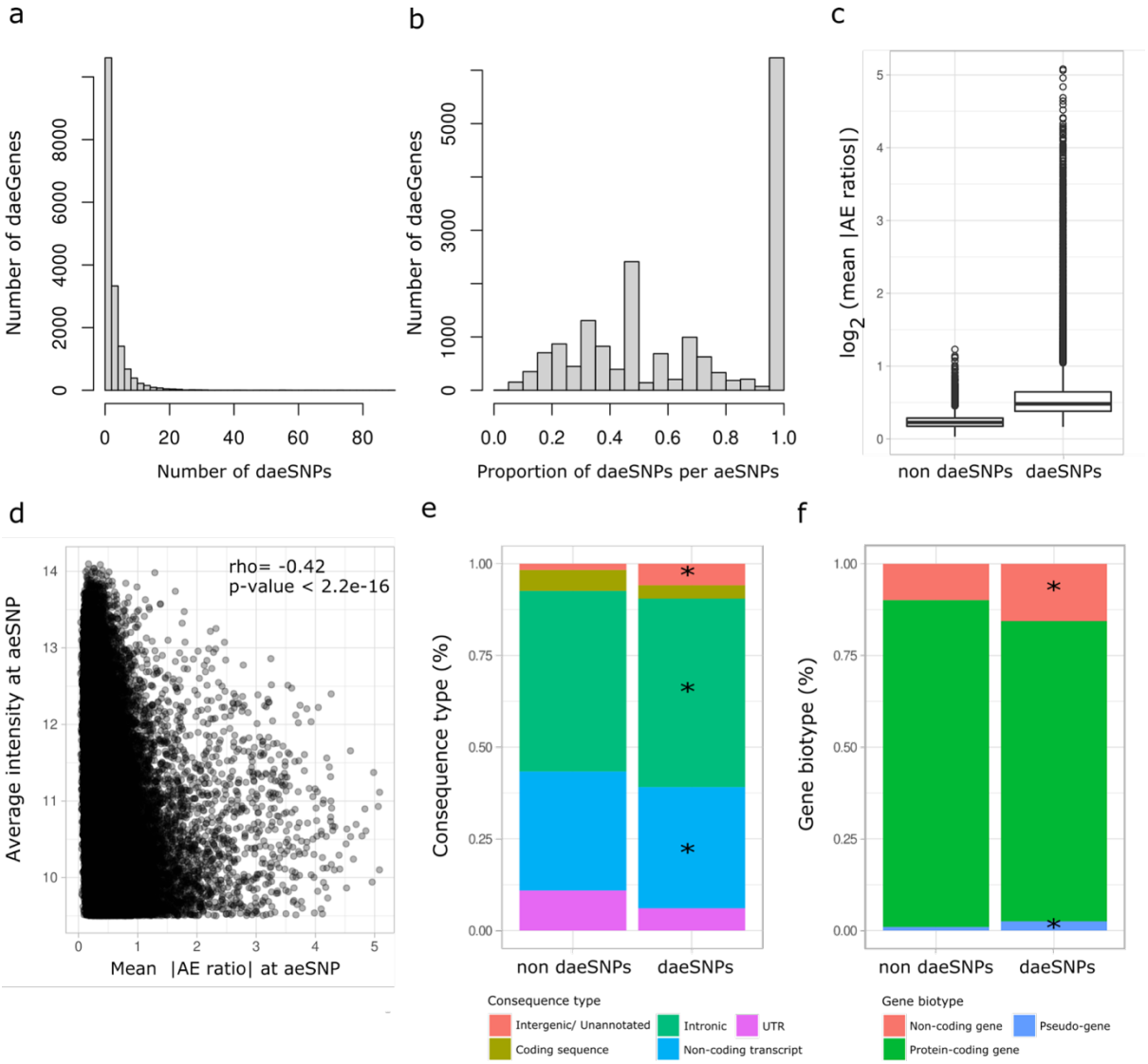

Figure 2 - Characterisation of aeSNPs. a) Histogram of the rank number of daeSNPs identified per gene across 17135 annotated genes. b) Histogram of the rank proportion of daeSNPs per aeSNPs identified per gene. c) Box plot with the distribution of the mean of the absolute values of AE ratios across heterozygous individuals measured at non-daeSNPs and daeSNPs. d) Distribution of the mean absolute values of AE ratios at aeSNPs according to the average intensity of both alleles at aeSNPs, in the microarrays data. It is shown the results of a Spearman's correlation test. e) and f) Relative frequency of aeSNPs and daeSNPs according to consequence type and gene biotype, respectively. (*) denotes the classes for which daeSNPs were enriched ($p<0.01$).

Mapping of daeQTLs in normal breast tissue

Evidence of DAE supports that a gene is under the control of *cis*-regulatory variation (rSNPs), which can be mapped using AE ratios as a quantitative trait – in what we termed DAE quantitative trait loci (daeQTL) analysis. Furthermore, the pattern of AE ratios distribution displayed by each daeSNPs is highly dependent on the LD between the daeSNP and the rSNP acting upon the gene [27].

Here, we found a minority of daeSNPs (1198 out of 44,267) for which all the heterozygotes preferentially expressed the same allele. This is indicative of strong linkage disequilibrium ($r^2 \sim 1$) between the daeSNP and the rSNPs acting on it [27]. Hence, for these daeSNPs, we identified 19316 daeQTLs (Approach 1, Table 2). From the mapping of all other daeSNPs (Approach 2), we identified 5049 daeQTLs (FDR of 0.1, Figure 3a) for 1295 daeGenes. Given that 617 daeQTLs were commonly identified by both approaches, in total we identified 23748 unique daeQTLs for 2753 (16%) daeGenes (Table 2, Table S7). daeQTLs are located mostly within 65Kb from the corresponding daeSNP, but as far as the 500kb window used for the analysis. However, those identified by Approach 2 were more proximal because the distance was used as a covariate in the analysis (95% located within 105Kb) (Figure 3b). A daeQTL for *MROH8,* a coding gene located on chromosome 20, was the most significant one found (nominal p-value = 3.2e-09), but other highly significant daeQTLs (nominal p-values smaller than 5.0e-08) located at nine other loci (Figure 3a, 3c, Table S7).

Table 2. Summary of daeQTLs identified in breast tissue

daeQTLs	Number	Ensembl Target Gene IDs
Total	23748	2753
identified by mapping approach 1	19316	1640
identified by mapping approach 2	5049	1295
common to both approaches	617	182

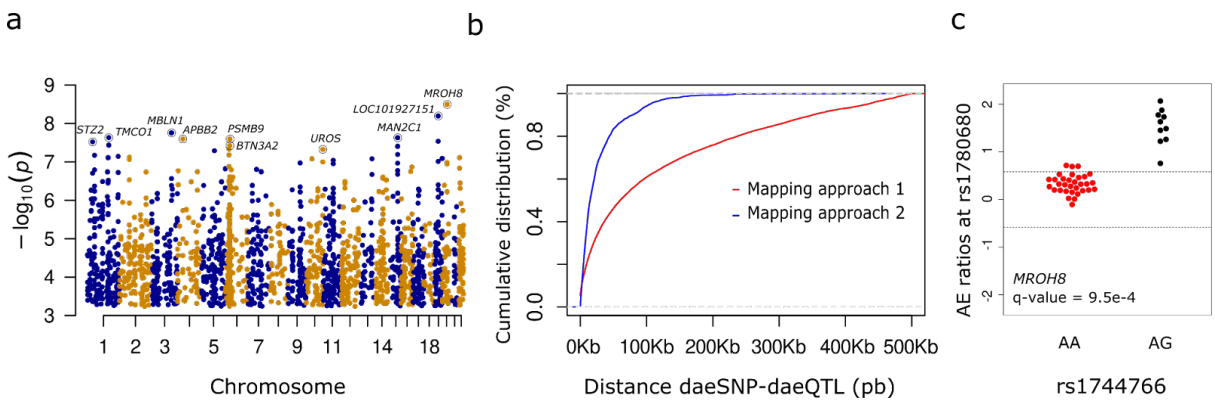

Figure 3 - Mapping of variants associated with differential allelic expression. a) Manhattan plot for the daeQTLs (nominal p-value) identified by mapping approach 2 using an FDR of 10%. aeGenes with daeQTLs with p-values lower than 5.0e-08 are annotated. b) Empirical cumulative distribution for distance between the daeSNP and corresponding mapped daeQTL. Results from mapping approach 1 are shown in red and those from mapping approach 2 are shown in blue. c) daeQTL mapping result for the most significant daeQTL identified for *MROH8*. The AE ratios calculated at the daeSNP rs1780680 are represented on the y-axis in all panels, and are stratified according to genotype at rs1744766 (heterozygous individuals are shown in black and homozygous individuals in red).

Annotation of DAE and daeQTLs at BC risk loci

To pinpoint the most likely candidate target genes within BC risk loci, a main post-GWASes challenge, we identified the genes displaying the strongest evidence of being under the control of *cis*-regulatory variation, either by DAE (daeGenes) or eQTL (eGenes) analysis, in GWAS reported loci (gwasGenes). Most gwasGenes (879 out of 957) showed evidence of being cis-regulated, with 51% identified solely by DAE analysis and 43% by both analyses. These percentages reflect a significant enrichment (Fisher's exact test = 1.0e-04) when compared to the initial set of genes analysed for DAE and eQTL (13771 daeGenes or eGenes in 15,706 total). Next, we sought to verify our ability to identify a set of 178 previously proposed target genes. We found that 52% of these were exclusively daeGenes (e.g. *ELL, TOX3, ISYNA1*), 33% were identified as both daeGenes and eGenes (e.g., *CASP8, CCND1, STXBP4*), 4% were exclusively eGenes (e.g. *RMND1, HELQ, PRKRIP1*) and the remainder 11% were not supported by either approach (*CITED4, IGFBP5, MYC*) (Table S2). This predominance of daeGenes compared to eGenes was not biased by the total gene expression levels, as those genes identified solely as eGenes showed higher median levels of expression than common genes (identified by both approaches) and solely daeGenes. Additionally, less than 5% of genes altogether were lowly expressed (median <0.1 TPM), although slightly higher among solely daeGenes (4.2% compared to 0.6% in solely eGenes and 1.9% in common genes) (Figure S9).

Another post-GWAS challenge is the identification of the causal variants within risk loci. To find the most likely variants regulating gwasGenes in *cis*, we successfully mapped daeQTLs for 154 gwasGenes (17.5%) (Figure 4, Table S8). To further identify the most plausible daeQTLs associated with risk (risk-daeQTLs), we identified those in moderate to strong LD ($r^2 \geq 0.4$) with GWAS index SNPs (Figure 1 – Step 3) (GWAS P-value < 1.0e-05). This revealed 404 risk-daeQTLs distributed across 31 different loci in 17 chromosomes, mostly located in intronic regions, followed by non-coding transcripts (Table S9, Figure S10). Furthermore, thirty-seven novel candidate target genes were detected in 22 loci with no previous report of target genes (Table 3). The five remaining loci contained previously proposed target genes, yet we identified additional novel ones, namely *STRADB* and *TRAK2* in 2q33.1 (Table 4). Importantly, a useful resource output from this analysis is the list of 1756 daeGenes with mapped daeQTLs that do not co-localize with currently known BC risk loci but could be mined to study future risk loci (Figure 4, Table S8). Additionally, the daeQTLs in lower LD with GWAS hits ($0.2 \leq r^2 < 0.4$) represent another valuable dataset warranting further exploration (Table S10).

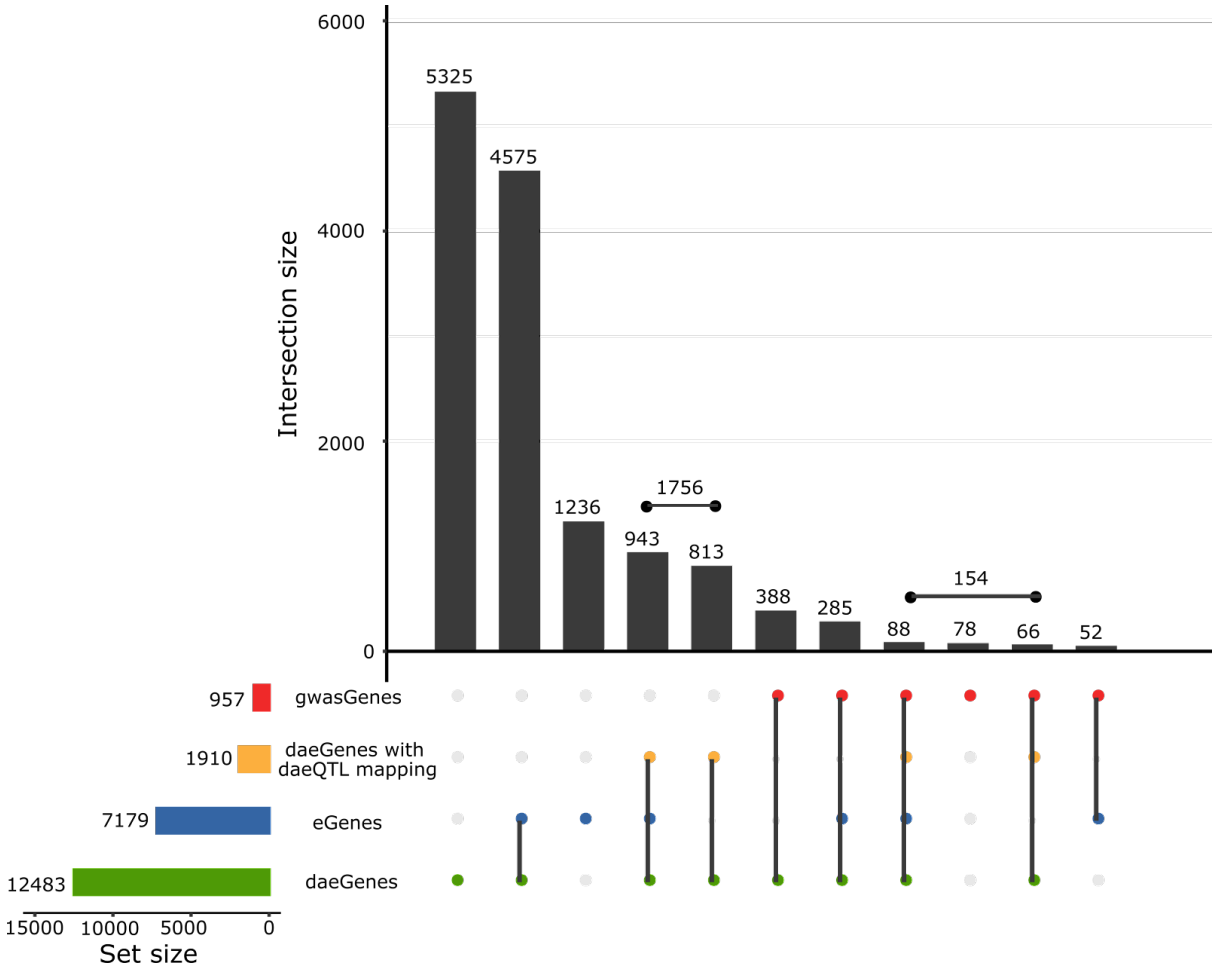

Figure 4 - Intersection of DAE and eQTL data from normal breast tissue and BC GWAS data. Upset plot for 15,706 genes tested for both DAE and eQTL (GTEx breast mammary tissue). Legend: daeGenes - genes identified as having differential allelic express in normal breast tissue; eGenes - genes reported as being eQTL genes in GTEx mammary tissue data (q-value ≤ 0.05); gwasGenes - genes where GWAS index SNPs or proxies ($r^2 \geq 0.4$) are located; daeGenes with daeQTL mapping - daeGenes for which daeQTLs were identified.

Table 3. Loci with candidate risk rSNP and novel suggested target genes.

Locus	GWAS SNP	GWAS nearest Gene	GWAS Study	N candidate risk rSNP	LD Interval[#]	Proposed Target Gene
1p36.23	rs7535752	*SLC45A1*	Song et al. 2013	5	0.49-0.56	*RERE*
	rs9628987	*SLC45A1*	Song et al. 2013	3	0.42-0.45	*RERE*
1q32.1	rs2810650	*SNRPE*	Michailidou 2017	7	0.45-0.62	*ZBED6, ZC3H11A*
	rs4951011	*ZBED6, ZC3H11A*	Cai 2014	10	0.51-1	*ZBED6, ZC3H11A*
	rs59867004	*ZBED6, ZC3H11A*	Michailidou 2017	6	0.40-1	*ZBED6, ZC3H11A*
3p24.1	rs4973768	*SLC4A7*	Ahmed 2009; Turnbull 2010; Fletcher 2011; Michailidou 2013; Michailidou 2015; Michailidou 2017	10	0.44-0.77	*NEK10*
	rs7619833	*SLC4A7*	Michailidou 2017	5	0.52-0.63	*NEK10*
	rs1357245	*NEK10*	Ahmed 2009; Michailidou 2015	10	0.43-0.86	*NEK10*
	rs60936670	*NEK10*	Michailidou 2017	10	0.61-1	*NEK10*
	rs653465	*NEK10*	Ahsan 2014	9	0.61-1	*NEK10*
3p21.31	rs56387622	*MYL3*	Michailidou 2017	1	0.47	*MYL3*
	rs6796502	*PRSS42P*	Michailidou 2015; 2017	6	0.44-0.47	*MYL3*
3q25.31	rs2136690	*TIPARP, TIPARP-AS1*	Michailidou 2017	4	0.41-0.86	*TIPARP, TIPARP-AS1*
	rs7637701	*LEKR1*	Michailidou 2017	4	0.52-0.71	*TIPARP, TIPARP-AS1*
4q22.1	rs2725207	*PKD2*	Michailidou 2017	1	1	*ABCG2*
5p15.33	rs62641919 [*]	*AHRR*	Michailidou 2017	33	0.44-1	*PLEKHG4B*
5p15.2	rs1092913	*ROPN1L*	Sehrawat 2011	9	0.56-1	*MARCH6*
6p22.3	rs3819405	*ATXN1*	Michailidou 2017	1	1	*GMPR*
6p22.2	rs71557345	*ZNF322*	Michailidou 2017	13	0.42-0.88	*BTN3A2*
	rs13195401	*BTN2A1*	Michailidou 2017	35	0.48-1	*BTN3A2*
	rs17598658	*HIST1H2BE*	Michailidou 2017	21	0.69-1	*BTN3A2*
	rs2523992	*AL645929.3*	Michailidou 2017	29	0.44-0.49	*HLA-L, HCG17*
	rs3094146	*ZNRD1ASP*	Michailidou 2017	18	0.41-0.70	*HLA-L, HCG17*
	rs3094054	*UBQLN1P1*	Michailidou 2017	21	0.74	*HLA-L, HCG17*

	rs3132610	*ABCF1*	Michailidou 2017	2	0.77	*HLA-L, HCG17*
	rs3132615	*SUCLA2P1*	Michailidou 2017	21	0.74	*HLA-L, HCG17*
7q11.23	rs377629160	*STAG3L2*	Michailidou 2017	3	0.53-0.89	*GTF2IRD2*
	rs74910854	*GTF2I*	Michailidou 2017	1	0.58	*GTF2IRD2*
9q31.1	rs11290052	*SMC2*	Michailidou 2017	2	0.41	*SMC2*
	rs4742903	*SMC2*	Michailidou 2017	26	0.84-1	*SMC2*
9q34.2	rs3124765	*CACFD1*	Michailidou 2017	1	0.50	*ADAMTS13*
10p12.1	rs7918232	*ANKRD26*	Michailidou 2017	26	0.40-0.93	*MASTL, PTCHD3, YME1L1*
13q32.1	rs1926657	*ABCC4*	Murabito 2007	1	1	*ABCC4*
14q13.2	rs58327846	*PRORP*	Michailidou 2017	2	0.40-0.44	*PRORP, SEPTIN7P1*[§]
14q32.33	rs60226654	*KLC1, COA8*	Michailidou 2017	20	0.41-0.52	*EIF5, KLC1, XRCC3*
	rs10623258	*ADSSL1*	Michailidou 2017	1	0.41	*SIVA1*
15q24.2	rs8027365	*PTPN9*	Michailidou 2017	27	0.80-0.83	*MAN2C1*
15q26.1	rs2290203	*PRC1, PRC1-AS1, AC068831.7*	Cai 2017; Michailidou 2017	1	0.56	*PRC1, PRC1-AS1*
	rs77554484	*PRC1*	Michailidou 2017	1	1	*PRC1, PRC1-AS1*
16q22.2	rs71695136	*IST1*	Michailidou 2017	9	0.48-0.55	*AP1G1*
17q21.31	rs2732699	*ARL17B*	Michailidou 2017	32	0.64-0.72	*ARL17B, LINC02210-CRHR1*[§]*, CRHR1, MAPT-AS1, KANSL1*
	rs4763	*ARHGAP27*	Michailidou 2017	4	0.68	*LINC02210-CRHR1§, CRHR1, MAPT-AS1, KANSL1*
19p13.3	rs3815308	*DOT1L*	Michailidou 2017	3	0.58	*AP3D1*

Legend:
*reported as rs116095464 in the original GWAS
[#]Linkage disequilibrium (LD) values r^2 between the daeQTL and the GWAS risk variant in the European population
[§]Gene not expressed in breast mammary tissue or without expression information in GTEx

Table 4. Loci with candidate risk rSNP and novel and previously suggested target genes.

Locus	GWAS SNP	GWAS nearest Gene	GWAS Study	N candidate risk rSNP	LD Interval #	Confirmed Previously Proposed Target Gene	Novel Proposed Target Gene
2q31.1	rs1550623	*CDCA7*	Michailidou 2013; 2015; 2017	3	0.83-1	*CDCA7*	
	rs200725404	*CDCA7*	Michailidou 2017	4	0.60-0.76	*CDCA7*	
2q33.1	rs182731523	*CFLAR*	Michailidou 2017	5	0.42-0.43	*CASP8*	
	rs1830298	*FLACC1*	Michailidou 2017	9	0.69-1	*CASP8*	
	rs3769821	*CASP8*	Michailidou 2017	9	0.69-0.80	*CASP8*	
	rs2714486	*TRAK2*	Michailidou 2017	31	0.43-0.53	*CASP8*	*STRADB, TRAK2*
	rs77688320	*C2CD6*	Michailidou 2017	31	0.41-0.51	*CASP8*	*STRADB, TRAK2*
5q14.1-14.2	rs7707921	*ATG10*	Michailidou 2015; 2017	21	0.62-1	*ATG10*	
	rs111549985	*ATG10*	Michailidou 2017	9	0.92-1	*ATG10*	
	rs146817970	*ATG10*	Michailidou 2017	21	0.62-1	*ATG10*	
	rs2407156	*ATP6AP1L*	Michailidou 2017	18	0.53-0.65	*ATG10*	
10q21.2	rs10822013	*ZNF365*	Cai 2011	2	0.80	*ZNF365*	
17q22	rs2787486	*STXBP4*	Michailidou 2017	1	0.74	*COX11*	*TOM1L1*
	rs6504950	*STXBP4*	Ahmed 2009; Michailidou 2013; 2015	1	0.85	*COX11*	*TOM1L1*

Legend:
#Linkage disequilibrium (LD) values r^2 between the daeQTL and the GWAS risk variant in the European population

Mapping of *cis*-regulatory risk variants at the 5q14.1-14.2 locus

To further show the potential use of our integrated approach, we focused our follow-up studies on the BC risk locus 5q14.1-14.2, where some of the most significant daeQTLs are strongly associated with BC risk variants. In this locus, rs7707921 was previously associated with BC risk in two meta-analyses (OR for alternative A allele = 1.07, 95% CI = [1.05-1.1], p=5e-11) [9,54]. The region containing this intronic variant of the *ATG10* gene*,* its proxy variants ($r^2 \geq 0.4$) and other risk-associated variants reported in this locus spans through three genes (*ATG10, RPS23,* and *ATP6AP1L*) hindering the identification of the causal variant(s) and their target gene(s) in this locus.

Firstly, all three genes showed DAE supporting their regulation by *cis*-acting variation (rSNPs): 21 daeSNPs out of 37 aeSNPs at *ATG10*, one daeSNP out of two aeSNPs at *RPS23* and four daeSNPs out of five aeSNPs at *ATP6AP1L*. The larger allelic expression imbalances detected at daeSNPs in these genes were 9.2-fold at *ATG10*, 4-fold at *RPS23* and 6-fold at *ATP6AP1L* (Figure 5a, Figure S11).

Next, we found 22 daeQTLs for four daeSNPs in *ATG10*, spreading along the *ATG10-RPS23* region, but none for any of the daeSNPs in *RPS23* or *ATP6AP1L* genes. All but one *ATG10* daeQTLs were in moderate to strong LD ($r^2 \geq 0.4$) with the risk-associated variants (Table S10) and were defined as risk-daeQTLs. Furthermore, the daeQTL *q*-values strongly correlated with the corresponding LD with the GWAS lead-SNP rs7707921 (Figure 5a, Figure S12), further supporting the association of these daeQTLs with risk.

Finally, we identified as candidate risk-rSNPs the 92 variants in higher LD ($r^2 \geq 0.95$) with the 21 risk-daeQTLs (Table S11). These SNPs span over 400Kb from *ATG10* to *ATP6AP1L* genes and were subjected to functional analysis.

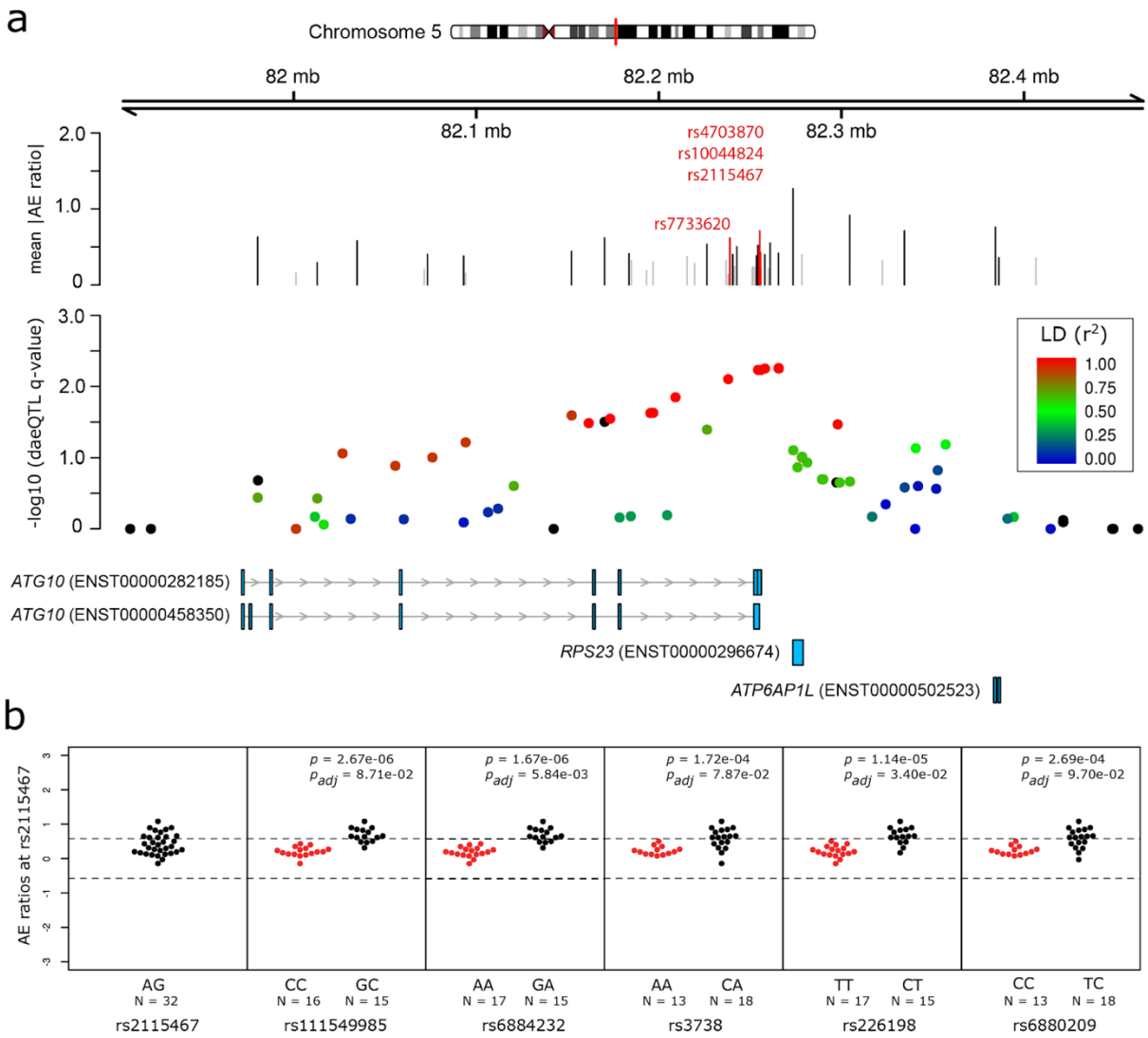

Figure 5 - Evidence of DAE and daeQTL analysis at the 5q14.1 BC risk locus. a) Top to bottom tracks show: 1 - the chromosomal region of variants according to the GRCh38/hg38 assembly; 2 - the mean absolute values of AE ratios measured for each aeSNP, with daeSNPs shown in black, daeSNPs with daeQTLs (including rsIDs) are shown in red and non-daeSNPs are shown in grey; 3 - the daeQTL analysis q-values colored according to the LD (r^2) with the GWAS lead SNP in the region rs7707921 (variants without LD information are shown in black); 4 - location of two *ATG10* transcripts, a *RPS23* transcript and a *ATP6AP1L* transcript. b) *ATG10* daeQTL mapping results for candidate-risk regulatory variants. The AE ratios calculated at the daeSNP rs21115467 are represented on the y-axis in all panels, and are stratified according to genotype at each candidate causal variants in individual panels (heterozygous individuals are shown in black and homozygous individuals in red).

Cis-regulatory risk variants act via three different mechanisms on genes in the 5q14 locus

The subsequent analysis revealed five risk-rSNPs (Figure 5b) with functional evidence supporting their role as risk-causing variants via the control of miRNA binding, alternative transcription, and transcription factor binding, as described below.

Of the candidate risk-rSNPs mapping to the 3'UTR of *RPS23* (Table S11), rs3738 is a strong candidate to *cis*-regulate *RPS23:* the C allele (minor) is predicted to promote a specific stable pair binding between hsa-miR-6813-5 and a highly expressed *RPS23* transcript (ENST00000296674) (context score of -0.57) (Table S12, Figure 6a), which requires *in vivo* validation. This is consistent with the lower expression of the C allele at the rs3738 in most of the samples in the AE analysis (Figure S7). In previous work, we had not found significant allelic-specific binding of miRNAs for any of the candidate risk-rSNPs mapped at the 3'UTR of *ATG10 (*Table S12*)* [43].

We also found evidence that two candidate risk-rSNPs affect alternative transcription of *ATG10.* To increase the stringency of the sQTL analysis, we tested not only the 92 candidate risk rSNPs but also all SNPs located within 5Kb upstream and downstream of *ATG10.* We identified six sQTLs (FDR ≤ 5%) in tumour data, whose minor alleles were associated with changes in the expression of two protein-coding isoforms: decreased expression of ENST00000458350 (one extra exon) and increase expression of ENST00000282185 (longer 3'UTR) (Figure S13a, Table S13). Interestingly, ENST00000282185 uniquely harbours three of the four daeSNPs (rs4703870, rs10044824 and rs2115467) for which we identified daeQTLs (Table 3) and is significantly less expressed in tumours than in normal-matched tissue, in line with the reported oncogenic effect of UTR length [55], although with a small effect size (fold-change = 1.20) (Figure S14). The strong correlation between sQTL q-values and LD with the lead GWAS SNP rs7707921 (r=0.94, p-value = 3.15E-12, Figure S13b) supports the contribution of alternative transcription of *ATG10* to BC risk. Although no sQTL was detected for *ATG10* in normal breast data (Table S13), sQTL nominal p-values and LD with rs7707921 still correlated in normal matched breast samples (r = 0.59, p-value = 0.002) (Figure S15). *RPS23* and *ATP6AP1L* did not display sufficient alternative transcription dispersion to allow the sQTL analysis. Subsequent functional analysis of *ATG10's* sQTLs, and their proxy SNPs (LD $r^2 \geq 0.95$), revealed the prediction of rs111549985 (5'UTR) and rs6884232 (3'UTR) to cause a riboSNitch (a functional RNA structure disrupted by a SNP [47]). Although RBP binding data for breast tissue does not exist, these variants are known to disrupt the binding of Xrn2 (involved in termination by RNA polymerase II) and of Igf2bp1 (a translation regulator) in K562 cells (Table S14, Table S15), which would require confirmation in breast cells.

Lastly, we investigated candidate risk-rSNPs overlaying active promoter/enhancers and *DNase*I hypersensitive sites (DHSs) in normal mammary cells and/or breast tumour cell lines for allelic differences in transcription factor binding (Table S16, Figure S16). One of the these, rs111549985, also identified as an sQTL for *ATG10* (see above), overlays the active promoter of *ATG10* (Figure S12), and its minor G-allele preferentially associated with the H3K4Me3 modification in HMEC cells (2.7-fold, p = 3.7e-03) and shows a strong preferential

binding by POL2 in MCF7 cells (9-fold, p = 4.0e-04). However, DHS was more significantly associated with the major/reference C allele in T47D cells (0.5-fold, p = 4.6e-05) (Figure 6C, Table S17). Another two risk-rSNPs, rs226198 (intronic to *RPS23*) and rs6880209 (located at *RPS23* 5'UTR), overlay the shared promoter of *RPS23* and *ATP6AP1L* and a predicted enhancer interacting with the *ATG10* promoter (Figure S17). The minor C-allele of rs226198 showed preferential binding by MYC and MAX transcriptions factors, known to cooperate in cancer [56], (12.6-fold and 7.9-fold difference, respectively, p < 2.2e-16) and preferential H3K4me3 marking (2.7-fold, p = 1.4e-02) in MCF-7 cells (Figure 6c, Table S17). It would be interesting to elucidate further whether rs226198 impacts the binding of both factors and the H3K4me3 deposition or this epigenetic mark is a consequence of altered transcription, as previously suggested [57,58]. The minor T-allele of rs6880209 also showed preferential binding by MYC (4.8-fold, p < 2.2e-16) and MAX (2.4-fold, p = 2.7e-03), with smaller fold-change differences than rs226198, and additional preferential binding by POL2 (2.6-fold, p = 1.27e-06) in MCF7 cells. However, like rs111549985, DHS preferentially occurred in the major/reference C-allele in T47D cells (5.3-fold, p = 9.1e-04) (Figure 6C, Table S17). Interestingly, the expression of *MAX* was correlated with all three candidate target genes and the expression of *MYC* was correlated with the expression of *ATG10* (Figure S18). Furthermore, the expression levels of *ATG10* and *ATP6AP1L* were positively correlated in breast tissue from healthy women (top 2.5% quantile of 500,000 pairwise tests) and in normal-matched tissue from patients with BC (Figure S19). Interestingly, *ATG10* and *ATP6AP1L* are in different topologically associating domains (TAD) and the risk-rSNPs rs226198 and rs6880209 fall on the boundary between them (Figure S20). This suggests that the correlated gene expression is not driven by a shared pattern of chromatin condensation but instead by a shared *cis*-regulatory sequence.

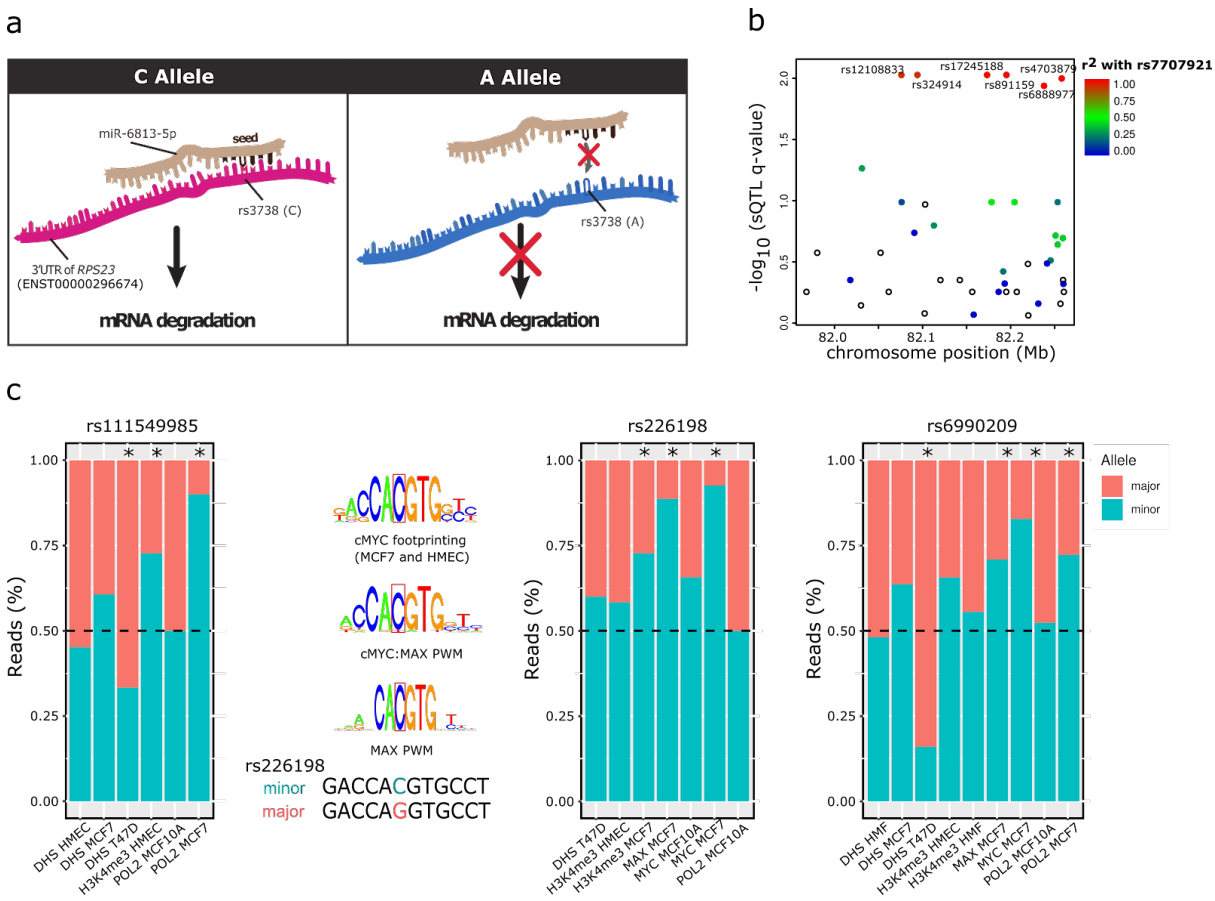

Figure 6 - Variants at the 5q14.1 risk locus associate with altered miRNA binding, alternative transcription, and transcription factor binding. a) Scheme of the predicted allelic effect of rs3728 on the binding of miR-6813-5p to the 3'UTR of transcript ENST00000296674. Image was adapted from Kelvinsong and is licensed under CC BY 3.0. **b**) Six sQTLs in high LD with rs7707921 were identified for *ATG10*. -log10(q-value) for the sQTL analysis (y-axis) are shown for the 5q14.1-14.2 region (hg38). Colour intensity represents the LD (r^2) between the analysed variants and the GWAS lead SNP rs7707921. **c**) Allele specific analysis of the effect of three candidate risk rSNPs — rs111549985, rs226198 and rs6880209 — on RNA polymerase II (POL2) and transcription factor (TF) binding, DNase I targeting (DHS) and H3K4me3 modification in different heterozygous cell lines. Statistically significant imbalances (two-sided binomial-test, p-value ≤ 0.05) are indicated by an asterisk. Legend: HMEC - Human Mammary Epithelial Cells; MCF7 - human breast (adenocarcinoma) cell line; T47D - human breast tumour cell line; MCF10A - human breast epithelial cell line.

Risk model for 5q14.1 links higher expression of *ATG10* and *ATP6AP1L* and lower expression of *RPS23* with protection against BC

Haplotype analysis of the samples included herein revealed two common haplotypes: one harbouring the major alleles of all proposed risk-rSNPs and the GWAS lead SNP

rs7707921 (frequency of 71.1%), and another with the corresponding minor alleles (frequency of 21.9%) (Figure S21). Additionally, the proposed risk-rSNPs are among the most significant eQTLs for the three genes: rs111549985 for *ATG10*, rs3738 for *RPS23* and rs6880209 for *ATP6AP1L* (Figure S18) [59]. Therefore, the most common haplotype is associated with increased risk for BC, lower expression of *ATG10* and *ATP6AP1L* and higher expression of *RPS23* (Figure S22).

Our proposed model for risk at 5q14.1 (Figure 7) establishes that the minor alleles of rs111549985, rs226198, and rs6880209 confer protection against BC by (1) increasing the binding of POL2 II to the promoter of *ATG10* (driven by rs111549985), (2) the binding of POL2 to the shared promoter of *RPS23/ATP6AP1L* (driven by rs6880209), and (3) the binding of cMYC and MAX to a regulatory region (possible enhancer) (driven by rs226198), increasing the expression of *ATG10* and *ATP6AP1L.* We propose that the minor allele of rs3738 confers protection to BC risk by a post-transcriptional regulatory mechanism independent of the regulation of *ATG10* or *ATP6AP1L*, in which allele-specific binding of the miRNA hsa-miR-6813-5p to the 3'UTR of an *RPS23* transcript results in its decreased expression. Taken together, these results reveal a complex regulatory landscape at the 5q14.1-14.2 locus, with multiple independent causal variants.

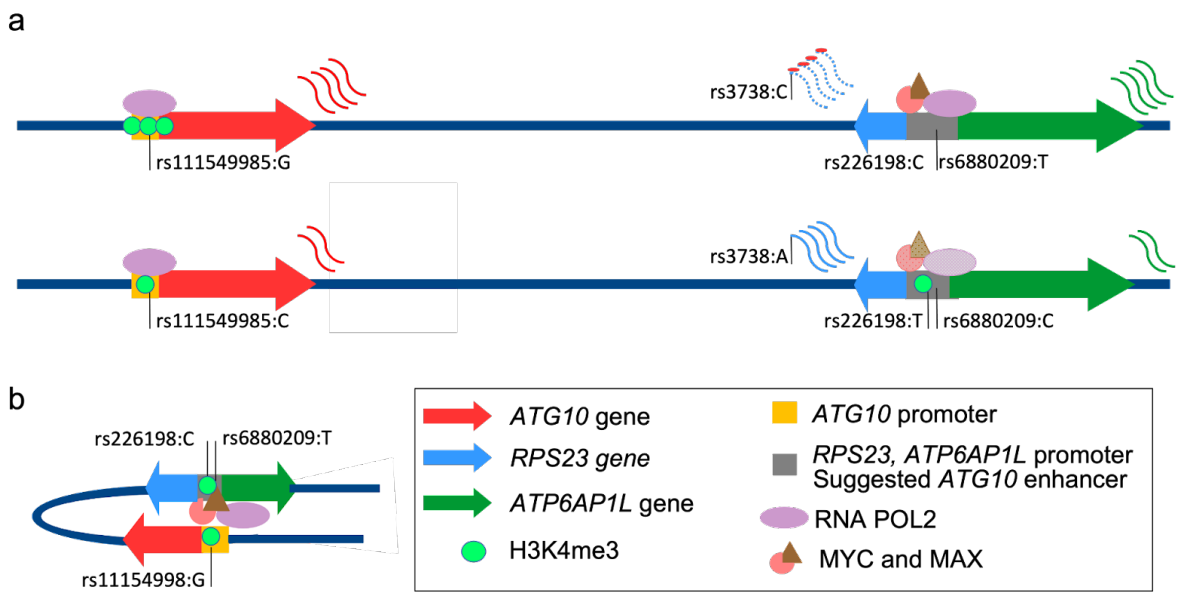

Figure 7 - Complex risk regulatory landscape of the 5q14.1 locus. a) Levels of expression of *ATG10*, *RPS23* and *ATP6AP1L* genes differ between the haplotypes containing either the minor alleles of rs111549985, rs3738, rs226198 and rs6880209 (above) or the major ones (bellow). Coloured arrows indicate the direction of transcription of the individual genes, the strength of protein binding is indicated by the saturation of the corresponding colours, the level of H3K4me3 is indicated by the number of green circles, the relative levels of transcript produced are indicated by the coloured lines above the haplotypes, and the *RPS23* allele targeted by hsa-miR-6813-5p is shown as a dashed

line. b) Schematic representation of the proposed model for the positive correlation between *ATG10* and *ATP6AP1L* via a shared regulatory region.

5

DISCUSSION

Here, we present the first genome-wide map of differential allelic expressed genes (daeGenes) in normal breast tissue and their genetic determinants (daeQTLs). We found widespread differential allelic expression (DAE) across the genome and identified daeQTLs for 16% of daeGenes. By intersecting this map with GWAS data, we identified novel candidate causal variants (risk-daeQTLs) and target genes for 31 BC risk loci, besides confirming others. Our results represent a practical and useful resource for prioritising loci for follow-up GWAS studies. As proof of concept, we characterised the 5q14.1 BC risk locus and proposed five causal regulatory variants targeting the genes *ATG10*, *RPS23,* and *ATP6AP1L* acting via multiple allele-specific mechanisms. Our results suggest a complex regulatory landscape underlying BC aetiology.

We show that *cis*-acting variants regulate the expression of 80% of genes in normal breast tissue, with some genes displaying extreme allelic differences of up to 32-fold. Notably, we identified a novel mono-allelic expressed gene, *ARCN1*, which warrants further inspection to confirm imprinting status. An enrichment of daeSNPs at intergenic and intronic regions, as well as non-coding transcripts, non-coding genes and pseudogenes, concurs with previous reports of predominant allelic imbalances of expression at gene-depleted regions and genes under less evolutionary constraints [60,61].

To overcome the lack of phasing information, we applied two different approaches in the daeQTL mapping that led to the identification of 23748 variants associated with AE ratios for 2753 genes, both coding and non-coding for proteins. The stringent statistical correction and the use of distance as a covariate in the second mapping approach increased its level of confidence but limited the statistical power to identify regulatory variants in lower LD with the daeSNP or located more distally. daeQTLs were identified in common for 22% of the genes by both approaches, increasing further the confidence in the mapping exercise.

We found evidence of expression regulation by *cis*-acting variants for the majority of reported GWAS loci and believe that alternative mechanisms are at play in the remainder. Notably, we identified risk-daeQTLs at 31 different loci, including 22 loci with novel candidate risk target genes (including *NEK10* at 3p24.1 and *ZBED6* and *ZC3H11A* at 1q32.1). Moreover, the initial daeQTLs map in normal breast tissue can be further mined whenever new risk variants are identified through GWAS. These results offer a resource platform for functional studies of causal variants and target genes and can help uncover the role of cis-regulatory variation in BC risk.

Finally, we conducted an *in-silico* functional analysis of the 5q14.1-14.2 BC risk locus, and identified three strong candidates causal variants (rs111549985, rs226198, and

rs6880209) that are risk-daeQTLs for *ATG10*. These variants are predicted to functionally impact TF binding, chromatin state, and gene expression levels of *ATG10, RPS23,* and *ATP6AP1L*. A similar involvement of diverse regulatory mechanisms has been suggested previously for other BC risk loci [4,62,63] Both *ATG10* (involved in autophagy) and *RPS23* (encodes for a ribosomal component of the 40S subunit) have been suggested to have roles in cancer [64–66], but the pseudogene *ATP6AP1L* is less studied. A variant at *ATG10* (rs7313473) was previously associated with BC risk by regulating promoter activity and *ATG10* was suggested to act as a tumour suppressor gene in breast tissue [67]. Although we did not find supporting evidence for the same variant, our results show an indirect association between the lower expression of *ATG10* and BC risk, suggesting that *ATG10* down-regulation may contribute to tumorigenesis.

The advantages of our analysis compared to previous reports of AE in normal breast and tumour tissue [16,18,19,68] include the use of the largest number of normal breast tissue samples, the genome-wide approach, and the mapping of candidate regulatory variants. We found a similar frequency of daeSNPs to previous reports in other tissues/cell lines, but a higher frequency of daeGenes [13,17,18,69]. This higher frequency of daeGenes could be due to our ability to identify genes regulated by common *cis*-acting variants with weak to large effect sizes [19], consequence of the imposed conditions to call DAE (allelic fold-change difference of 1.5-fold and the minimum number of heterozygotes). Also, we did not integrate the AE ratios of multiple daeSNPs in the same gene due to the absence of phase data and to maximise the information withdrawn from daeSNPs that might locate in different LD blocks. This is supported by the complex regulatory landscape we identified at 5q14.1 locus, with multiple cis-acting variants located in the same haplotypes and AE likely resulting from the sum of the effects of each variant. Furthermore, a global measure of the AE imbalance at each gene would impair the mapping of daeQTLs at individual daeSNPs, as we propose, and would restrict the analysis to genes with multiple daeSNPs, which would have prevented the role of *RPS23* to be revealed. Finally, we analysed non-coding genes and pseudogenes, like *ATP6AP1L*, besides the more commonly studied protein-coding genes. Moreover, our results confirm the advantage of using DAE analysis to detect the effect of rSNPs compared to eQTL analysis, as shown by the higher number of daeGenes, than eGenes, amongst gwasGenes [70–72]. As a minority of gwasGenes were exclusively eGenes, we believe that DAE and eQTL analyses are complementary and should be used in parallel when possible.

However, the main limitation of our study design is the use of microarray data, which has a smaller transcriptome coverage than RNA-Seq and less accuracy for quantifying more extreme allelic imbalances. Nevertheless, this is a widely used and precise technology for measuring AE [13,16,23], as proven by our validation of mono-allelic expression in known

imprinted genes and by independent PCR analysis of seven out of nine daeGenes. Another limitation, transversal to all AE analysis approaches, is that only heterozygous individuals are informative.

Here, we provide a genome-wide list of variants with strong regulating potential for normal breast tissue, a valuable resource for researchers looking into prioritizing GWAS results for functional characterization, as well as those interested in other BC related traits. The extensive characterisation of the regulatory landscape at the 5q14.1 BC risk locus identified candidate causal variants and revealed the multiple mechanisms involved. Further studies of this locus will elucidate the mechanisms involved and the relative contributions of each variant and target gene to the genetic risk. Overall, our results reinforce the importance of *cis*-regulatory variation as a major player in BC susceptibility and the power of identifying these variants in the disease's tissue of origin - normal breast tissue. They also show that multiple causal variants may co-occur and act via independent *cis*-regulatory mechanisms at BC risk loci, supporting a broader approach to functional studies.

Declarations

Ethics approval and consent to participate

Not Applicable.

Consent for publication

Not Applicable.

Availability of Data and Materials

All datasets analysed in this study were obtained from publicly available databases and websites. SNPs associated with BC risk were obtained from the GWAS Catalog website at www.ebi.ac.uk/gwas. SNP data were obtained from the Ensembl database (annotated according to GRCh38.p13) available at www. ensembl.org. eQTL data for breast tissue from the GTEx Project (v7 and v8) were retrieved from the GTEx Portal at www. gtexportal.org. Our microarray data is available from GEO (www.ncbi.nlm.nih.gov/geo/) under accession number GSE35023. DAE and daeQTL mapping analyses were carried out using R, the code of which is available on GitHub (https://github.com/maialab/DaeBreastMicroarrays).

Competing Interests

The authors declare 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 CBMR-UID/BIM/04773/2013, POCI-01-0145-FEDER-022184 “GenomePT”, the contract DL 57/2016/CP1361/CT0042 (J.M.X.) and individual postdoctoral fellowship SFRH/BPD/99502/2014 (J.M.X.). The research leading to these results has received funding from the People Programme (Marie Curie Actions) of the European Union’s Seventh Framework Programme FP7/2007-2013/under REA grant agreement no. 303745 (A.T.M.), a Maratona da Saúde Award (A.T.M.) and a BCRF project Grant.

Authors’ Contributions

Writing team: Xavier JM, Maia AT

Analysis team: Xavier JM, Magno Ramiro, de Almeida BP, Jacinta-Fernandes A, Russell R, Samarajiwa S, Dunning M, Maia AT

Lab team: de Almeida BP, Rocha CL, Rosli N, O’Reilly M, Maia AT

Experimental design team: Russell, R., Ponder BAJ, Maia AT

Acknowledgements

The authors would also like to thank Dr Nuno Barbosa-Morais at IMM for the excellent scientific discussions and the support given by the Unidade de Apoio à Investigação (UAIC) at Universidade do Algarve (UAlg), particularly Mr. Vitor Morais, and the Informatics Services of UAlg.

REFERENCES

1. Wendt C, Margolin S. Identifying breast cancer susceptibility genes – a review of the genetic background in familial breast cancer. Acta Oncol. 2019;58:1–12.

2. Gallagher MD, Chen-Plotkin AS. The Post-GWAS Era: From Association to Function. Am J Hum Genetics. 2018;102:717–30.

3. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science. 2012;337:1190–5.

4. Fachal L, Aschard H, Beesley J, Barnes DR, Allen J, Kar S, et al. Fine-mapping of 150 breast cancer risk regions identifies 191 likely target genes. Nat Genet. 2020;52:56–73.

5. 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.

6. Udler MS, Ahmed S, Healey CS, Meyer K, Struewing J, Maranian M, et al. Fine scale mapping of the breast cancer 16q12 locus. Hum Mol Genet. 2010;19:2507–15.

7. Meyer KB, Maia A-T, O'Reilly M, Ghoussaini M, Prathalingam R, Porter-Gill P, et al. A Functional Variant at a Prostate Cancer Predisposition Locus at 8q24 Is Associated with PVT1 Expression. Plos Genet. 2011;7:e1002165.

8. Darabi H, McCue K, Beesley J, Michailidou K, Nord S, Kar S, et al. Polymorphisms in a Putative Enhancer at the 10q21.2 Breast Cancer Risk Locus Regulate NRBF2 Expression. Am J Hum Genetics. 2015;97:22–34.

9. Michailidou K, Lindström S, Dennis J, Beesley J, Hui S, Kar S, et al. Association analysis identifies 65 new breast cancer risk loci. Nature. 2017;551:92–4.

10. Dunning AM, Michailidou K, Kuchenbaecker KB, Thompson D, French JD, Beesley J, et al. Breast cancer risk variants at 6q25 display different phenotype associations and regulate ESR1, RMND1 and CCDC170. Nat Genet. 2016;48:374–86.

11. Ghoussaini M, French JD, Michailidou K, Nord S, Beesley J, Canisus S, et al. Evidence that the 5p12 Variant rs10941679 Confers Susceptibility to Estrogen-Receptor-Positive Breast Cancer through FGF10 and MRPS30 Regulation. Am J Hum Genetics. 2016;99:903–11.

12. Pastinen T, Ge B, Hudson TJ. Influence of human genome polymorphism on gene expression. Hum Mol Genet. 2006;15 Spec No 1:R9-16.

13. 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.

14. Forton JT, Udalova IA, Campino S, Rockett KA, Hull J, Kwiatkowski DP. Localization of a long-range cis-regulatory element of IL13 by allelic transcript ratio mapping. Genome Res. 2007;17:82–7.

15. Bjornsson HT, Albert TJ, Ladd-Acosta CM, Green RD, Rongione MA, Middle CM, et al. SNP-specific array-based allele-specific expression analysis. Genome Res. 2008;18:771–9.

16. Gao C, Devarajan K, Zhou Y, Slater CM, Daly MB, Chen X. Identifying breast cancer risk loci by global differential allele-specific expression (DASE) analysis in mammary epithelial transcriptome. Bmc Genomics. 2012;13:570–570.

17. Romanel A, Lago S, Prandi D, Sboner A, Demichelis F. ASEQ: fast allele-specific studies from next-generation sequencing data. Bmc Med Genomics. 2015;8:9.

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

19. Aguet F, Brown AA, Castel SE, Davis JR, He Y, Jo B, et al. Genetic effects on gene expression across human tissues. Nature. 2017;550:204–13.

20. Hamdi Y, Soucy P, Adoue V, Michailidou K, Canisius S, Lemaçon A, et al. Association of breast cancer risk with genetic variants showing differential allelic expression: Identification of a novel breast cancer susceptibility locus at 4q21. Oncotarget. 2014;5:80140–63.

21. Zhang Y, Manjunath M, Zhang S, Chasman D, Roy S, Song JS. Integrative genomic analysis predicts causative cis-regulatory mechanisms of the breast cancer-associated genetic variant rs4415084. Cancer Res. 2018;78:1579–91.

22. 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.

23. 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.

24. Gimelbrant A, Hutchinson JN, Thompson BR, Chess A. Widespread Monoallelic Expression on Human Autosomes. Science. 2007;318:1136–40.

25. Goovaerts T, Steyaert S, Vandenbussche CA, Galle J, Thas O, Criekinge WV, et al. A comprehensive overview of genomic imprinting in breast and its deregulation in cancer. Nat Commun. 2018;9:4120.

26. 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.

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

28. Hothorn T, Hornik K, Wiel MA van de, Zeileis A. A Lego System for Conditional Inference. Am Statistician. 2006;60:257–63.

29. Hochberg Y, Benjamini Y. More powerful procedures for multiple significance testing. Stat Med. 1990;9:811–8.

30. Ignatiadis N, Klaus B, Zaugg J, Huber W. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. Nat Methods. 2016;13:577–80.

31. MacArthur J, Bowler E, Cerezo M, Gil L, Hall P, Hastings E, et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017;45:D896–901.

32. Magno R, Maia A-T. gwasrapidd: an R package to query, download and wrangle GWAS catalog data. Bioinformatics. 2019;

33. Yates A, Beal K, Keenan S, McLaren W, Pignatelli M, Ritchie GRS, et al. The Ensembl REST API: Ensembl Data for Any Language. Bioinformatics. 2015;31:143–5.

34. Consortium TGte, Ardlie KG, Deluca DS, Segre AV, Sullivan TJ, Young TR, et al. The Genotype-Tissue Expression (GTEx) pilot analysis: Multitissue gene regulation in humans. Science. 2015;348:648–60.

35. Dunham I, Kundaje A, Aldred SF, Collins PJ, Davis CA, Doyle F, et al. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.

36. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317–30.

37. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The Human Genome Browser at UCSC. Genome Res. 2002;12:996–1006.

38. Gonzalez JN, Zweig AS, Speir ML, Schmelter D, Rosenbloom KR, Raney BJ, et al. The UCSC Genome Browser database: 2021 update. Nucleic Acids Res. 2021;49:D1046–57.

39. 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.

40. Boyle AP, Hong EL, Hariharan M, Cheng Y, Schaub MA, Kasowski M, et al. Annotation of functional variation in personal genomes using RegulomeDB. Genome Res. 2012;22:1790–7.

41. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013;14:178–92.

42. JD S, AJ B, A D, D R. qvalue: Q-value estimation for false discovery rate control. [Internet]. 2021. Available from: http://github.com/jdstorey/qvalue

43. Jacinta-Fernandes A, Xavier JM, Magno R, Lage JG, Maia A-T. Allele-specific miRNA-binding analysis identifies candidate target genes for breast cancer risk. Npj Genom Medicine. 2020;5:4.

44. Monlong J, Calvo M, Ferreira PG, Guigó R. Identification of genetic variants associated with alternative splicing using sQTLseekeR. Nat Commun. 2014;5:4698.

45. Zhu Y, Xu G, Yang YT, Xu Z, Chen X, Shi B, et al. POSTAR2: deciphering the post-transcriptional regulatory logics. Nucleic Acids Res. 2019;47:D203–11.

46. Mao F, Xiao L, Li X, Liang J, Teng H, Cai W, et al. RBP-Var: a database of functional variants involved in regulation mediated by RNA-binding proteins. Nucleic Acids Res. 2016;44:D154–63.

47. Corley M, Solem A, Qu K, Chang HY, Laederach A. Detecting riboSNitches with RNA folding algorithms: a genome-wide benchmark. Nucleic Acids Res. 2015;43:1859–68.

48. Paz I, Kosti I, Ares M, Cline M, Mandel-Gutfreund Y. RBPmap: a web server for mapping binding sites of RNA-binding proteins. Nucleic Acids Res. 2014;42:W361–7.

49. Barrett JC, Fry B, Maller J, Daly MJ. Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics. 2005;21:263–5.

50. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. 2016;44:e71–e71.

51. Seitz H, Youngson N, Lin S-P, Dalbert S, Paulsen M, Bachellerie J-P, et al. Imprinted microRNA genes transcribed antisense to a reciprocally imprinted retrotransposon-like gene. Nat Genet. 2003;34:261–2.

52. Tierling S, Dalbert S, Schoppenhorst S, Tsai C-E, Oliger S, Ferguson-Smith AC, et al. High-resolution map and imprinting analysis of the Gtl2–Dnchc1 domain on mouse chromosome 12. Genomics. 2006;87:225–35.

53. Hagan JP, O'Neill BL, Stewart CL, Kozlov SV, Croce CM. At Least Ten Genes Define the Imprinted Dlk1-Dio3 Cluster on Mouse Chromosome 12qF1. Plos One. 2009;4:e4352.

54. Michailidou K, Beesley J, Lindstrom S, Canisius S, Dennis J, Lush MJ, et al. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. Nat Genet. 2015;47:373–80.

55. Mayr C, Bartel DP. Widespread Shortening of 3′UTRs by Alternative Cleavage and Polyadenylation Activates Oncogenes in Cancer Cells. Cell. 2009;138:673–84.

56. Dang CV. MYC on the path to cancer. Cell. 2012;149:22–35.

57. Floc'hlay S, Wong E, Zhao B, Viales RR, Thomas-Chollier M, Thieffry D, et al. Cis-acting variation is common across regulatory layers but is often buffered during embryonic development. Genome Res. 2020;31:gr.266338.120.

58. Howe FS, Fischl H, Murray SC, Mellor J. Is H3K4me3 instructive for transcription activation? Bioessays. 2017;39:1–12.

59. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45:580–5.

60. Campbell CD, Kirby A, Nemesh J, Daly MJ, Hirschhorn JN. A survey of allelic imbalance in F1 mice. Genome Res. 2008;18:555–63.

61. Tung J, Fédrigo O, Haygood R, Mukherjee S, Wray GA. Genomic features that predict allelic imbalance in humans suggest patterns of constraint on gene expression variation. Mol Biol Evol. 2009;26:2047–59.

62. 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.

63. 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.

64. Jo YK, Kim SC, Park IJ, Park SJ, Jin D-H, Hong S-W, et al. Increased expression of ATG10 in colorectal cancer is associated with lymphovascular invasion and lymph node metastasis. Plos One. 2012;7:e52705.

65. Jo YK, Roh SA, Lee H, Park NY, Choi ES, Oh J-H, et al. Polypyrimidine tract-binding protein 1-mediated down-regulation of ATG10 facilitates metastasis of colorectal cancer cells. Cancer Lett. 2017;385:21–7.

66. Wang Y, Huang J-W, Castella M, Huntsman DG, Taniguchi T. p53 Is Positively Regulated by miR-542-3p. Cancer Res. 2014;74:3218–27.

67. Guo X, Lin W, Bao J, Cai Q, Pan X, Bai M, et al. A Comprehensive cis-eQTL Analysis Revealed Target Genes in Breast Cancer Susceptibility Loci Identified in Genome-wide Association Studies. Am J Hum Genet. 2018;102:890–903.

68. Zhang K, Li JB, Gao Y, Egli D, Xie B, Deng J, et al. Digital RNA Allelotyping Reveals Tissue-specific and Allele-specific Gene Expression in Human. Nat Methods. 2009;6:613–8.

69. Ma X, Liu Y, Liu Y, Alexandrov LB, Edmonson MN, Gawad C, et al. Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. Nature. 2018;555:371–6.

70. Adoue V, Schiavi A, Light N, Almlöf JC, Lundmark P, Ge B, et al. Allelic expression mapping across cellular lineages to establish impact of non-coding SNPs. Mol Syst Biol. 2014;10:754.

71. Almlöf JC, Lundmark P, Lundmark A, Ge B, Maouche S, Göring HHH, et al. Powerful identification of cis-regulatory SNPs in human primary monocytes using allele-specific gene expression. Plos One. 2012;7:e52260.

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