nature genetics

Noncoding somatic and inherited single-nucleotide variants converge to promote *ESR1* expression in breast cancer

Swneke D Bailey[1,2,11], Kinjal Desai[3,11], Ken J Kron[1,2], Parisa Mazrooei[1,2], Nicholas A Sinnott-Armstrong[4], Aislinn E Treloar[1,2], Mark Dowar[1], Kelsie L Thu[5], David W Cescon[1,5], Jennifer Silvester[5], S Y Cindy Yang[1,2], Xue Wu[1,10], Rossanna C Pezo[1], Benjamin Haibe-Kains[1,2,6], Tak W Mak[2,5], Philippe L Bedard[1,7], Trevor J Pugh[1,2], Richard C Sallari[8] & Mathieu Lupien[1,2,9]

Sustained expression of the estrogen receptor-α (ESR1) drives two-thirds of breast cancer and defines the ESR1-positive subtype. ESR1 engages enhancers upon estrogen stimulation to establish an oncogenic expression program[1]. Somatic copy number alterations involving the *ESR1* gene occur in approximately 1% of ESR1-positive breast cancers[2–5], suggesting that other mechanisms underlie the persistent expression of *ESR1*. We report significant enrichment of somatic mutations within the set of regulatory elements (SRE) regulating *ESR1* in 7% of ESR1-positive breast cancers. These mutations regulate *ESR1* expression by modulating transcription factor binding to the DNA. The SRE includes a recurrently mutated enhancer whose activity is also affected by rs9383590, a functional inherited single-nucleotide variant (SNV) that accounts for several breast cancer risk–associated loci. Our work highlights the importance of considering the combinatorial activity of regulatory elements as a single unit to delineate the impact of noncoding genetic alterations on single genes in cancer.

Noncoding regulatory elements are the primary target of inherited risk variants[6–8], and their functional relevance to cancer is supported by the mutational constraint observed within these elements across tumors[9,10]. Functional noncoding SNVs can underlie 'single gene' diseases[11], confirming their ability to exert large phenotypic effects commonly associated with coding variants. This is highlighted in sporadic and familial melanoma, where somatic and germline genetic alterations in the promoter of *TERT* (which encodes telomerase) favor oncogenesis through an increase in *TERT* expression[12,13].

Genome-wide association studies (GWAS) have identified several SNVs associated with breast cancer risk at the *ESR1* locus among individuals of European and East Asian ancestry[14–18]. The population-specific patterns of linkage disequilibrium (LD) among the different lead SNVs seem consistent with a single underlying causal SNV. GWAS risk loci are enriched in regulatory elements, and they function by altering gene expression[6–8]. To identify the functional SNV(s), we first intersected all SNVs within a 5-Mb window of the original *ESR1* locus lead SNVs with functional annotations generated by the ENCODE project[19] in MCF-7 and T-47D ESR1-positive breast cancer cells. We then calculated the population-specific LD between the European and the East Asian lead SNVs (rs3734805 and rs2046210, respectively) and the neighboring SNVs using the genotype data from the 1000 Genomes Project[20]. We identified nine SNVs common to both Europeans and East Asians that share LD with the original population-specific lead SNVs ($r^2 \geq 0.8$ in both populations). SNVs rs9383590 and rs9397068, which were in perfect LD with each other and located 95 bp apart within the same DNase I hypersensitivity site (DHS), coincided with multiple functional genomic annotations generated by the ENCODE project[19] (**Fig. 1a** and **Supplementary Figs. 1** and **2**). These SNVs were also in strong LD ($r^2 = 0.81$) with the European breast cancer lead SNV rs9383938 (ref. 17). The rs9383590 SNV mapped to the second position of a GATA DNA recognition motif (**Fig. 1b**). The intragenomic replicates (IGR) tool[6] (Online Methods) predicted a decrease in the chromatin binding intensity of GATA3 for the variant allele (**Fig. 1c**), which was supported by allele-specific chromatin immunoprecipitation (ChIP)-qPCR in the heterozygous HCC1419 breast cancer cell line (**Fig. 1d**).

Enhancers regulate gene expression through physical interaction with the promoters of their target genes. Cross-cell-type correlation in DNase I hypersensitivity (C3D)[21,22] ($r \geq 0.7$; Online Methods) identified the enhancer harboring the rs9383590 SNV as potentially

[1]Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada. [2]Department of Medical Biophysics, University of Toronto, Toronto, Ontario, Canada. [3]Department of Genetics, Norris Cotton Cancer Center, Dartmouth Medical School, Lebanon, New Hampshire, USA. [4]Department of Genetics, Stanford University School of Medicine, Stanford, California, USA. [5]Campbell Family Institute for Breast Cancer Research, Princess Margaret Cancer Centre, University Health Network, Toronto, Ontario, Canada. [6]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada. [7]Division of Medical Oncology, Department of Medicine, University of Toronto, Toronto, Ontario, Canada. [8]Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts, USA. [9]Ontario Institute for Cancer Research, Toronto, Ontario, Canada. [10]Present address: Geneseeq Technology, Inc., Toronto, Ontario, Canada. [11]These authors contributed equally to this work. Correspondence should be addressed to M.L. (mlupien@uhnres.utoronto.ca).

Received 29 October 2015; accepted 26 July 2016; published online 29 August 2016; doi:10.1038/ng.3650

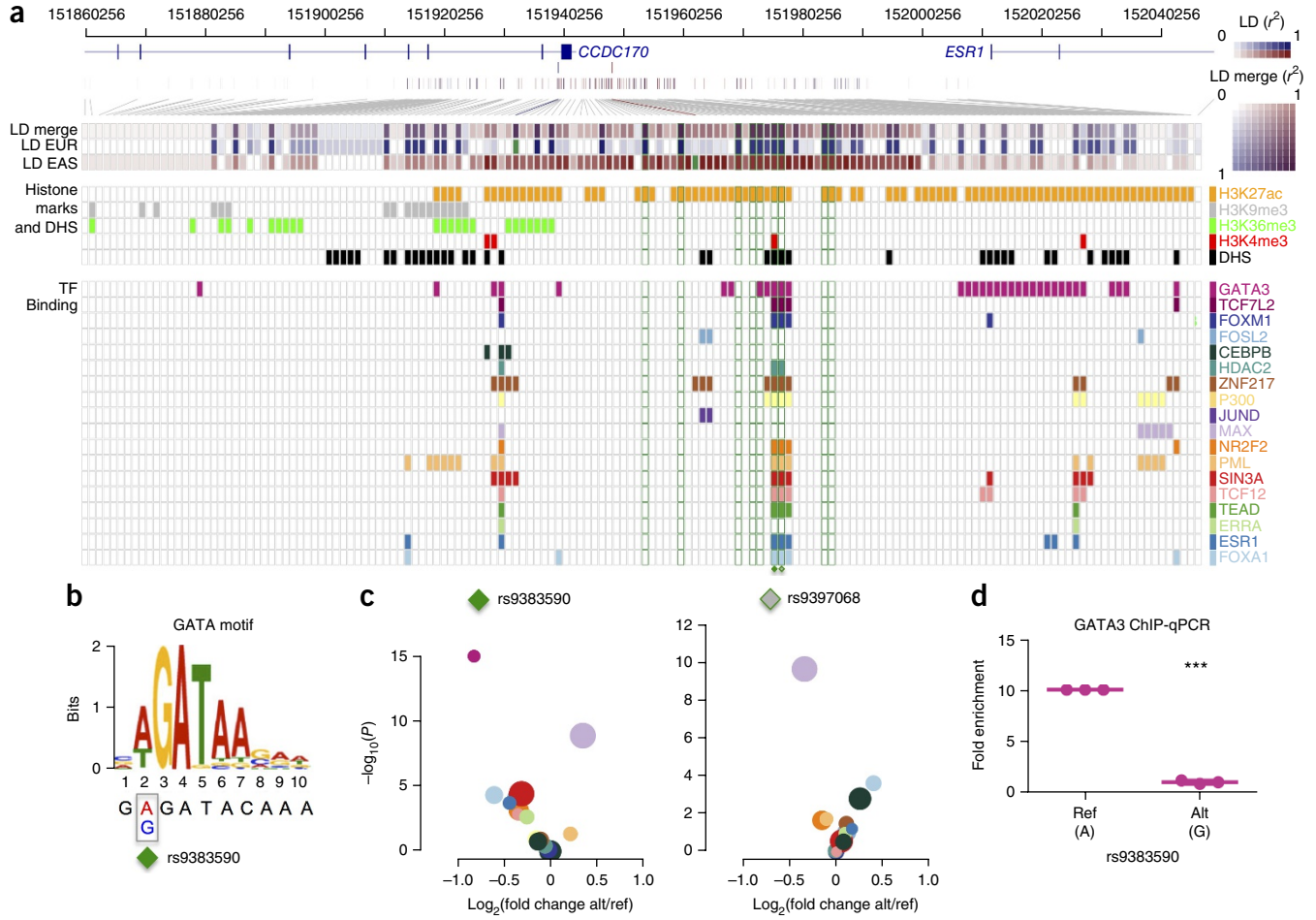

Figure 1 Identification of a functional risk-associated SNV shared between Europeans and East Asians. (**a**) LD shared between the European and East Asian lead SNVs ($n = 504$ and $n = 503$, respectively); LD merge, the composite strength of the LD for the European and East Asian lead SNVs; LD EUR, European LD pattern for SNV rs3734805; LD EAS, the East Asian LD pattern for SNV rs2046210. Green-filled squares correspond to the population-specific lead SNVs (rs3734805 and rs2046210); green outlines indicate the 9 LD SNVs with an $r^2 \geq 0.8$ with both the European and East Asian lead SNVs. Overlapping functional annotations observed in MCF-7 or T-47D cells profiled by the ENCODE project[19] are indicated by color (right). (**b**) Location of the rs9383590 SNV within the GATA3 DNA recognition motif. (**c**) A volcano plot of the IGR results for all transcription factors overlapping rs9383590 and rs9397068 as in **a**. The area of the circle is proportional to the maximum average signal intensity of the two alleles. (**d**) Allele-specific GATA3 ChIP-qPCR for rs9383590. ref, reference allele; alt, alternate allele. Experiments were done in triplicate. ***$P < 0.005$, two-sided one-sample *t*-test. Data are mean ± s.e.m.

interacting with the *ESR1* promoter (**Fig. 2a**). This interaction is corroborated by RNA polymerase II (Pol II) chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) data from MCF-7 cells produced by the ENCODE project[19] (**Fig. 2a**). To determine whether the rs9383590 SNV affects *ESR1* gene expression, we performed an expression quantitative trait locus (eQTL) analysis. We did not observe an additive association between the variant allele of the rs9383590 SNV and *ESR1* expression in ESR1-positive breast tumors profiled by the Cancer Genome Atlas (TCGA) or the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC)[23]. However, a linked SNV, rs9397435 ($r^2 = 0.97$ and $r^2 = 1$ with rs9383590 in Europeans and East Asians, respectively), was previously reported as a recessive eQTL associated with *ESR1* expression in breast tumor samples[18]. Consistent with those results, we observed a weak recessive eQTL among the luminal breast tumors in the larger METABRIC sample, using the rs9397437 SNV as proxy for the rs9383590 SNV ($r^2 = 1$ among Europeans and East Asians) ($n = 970$, $P = 0.039$) (**Fig. 2b**) (Online Methods). This eQTL should be interpreted with caution, as it was not observed within the TCGA samples. However, a luciferase reporter assay showed increased enhancer activity for the rs9383590 SNV variant allele (**Fig. 2c**). In addition, Li *et al.*[24] observed a significant allelic imbalance among TCGA breast tumors heterozygous for the lead East Asian SNV rs2046210. Using SNV rs9397437 as a proxy, we observed a consistent allelic imbalance in *ESR1* expression among heterozygous breast tumors measured with two independent coding marker SNVs (rs2077647 and rs1801132; $P < 0.05$) (**Fig. 2d** and Online Methods) and within the heterozygous HCC1419 breast cancer cell line ($P = 1.13 \times 10^{-4}$) (**Supplementary Fig. 3** and Online Methods). A similar result for the rs9397437 SNV was reported by Dunning *et al.*[25]. The variant allele of the rs9397068 SNV also increased enhancer activity (**Supplementary Fig. 4**). However, the effect of both SNVs did not appear to be additive (**Supplementary Fig. 4**). Together with the reference-allele-biased binding of GATA3, these results suggests that GATA3 may act as repressor, which has been previously reported by others[26].

Convergence of inherited risk variants and acquired somatic mutations on regulatory elements occurs at the *TERT* promoter in melanoma[12]. Using a set of 98 breast cancer samples profiled by whole-genome sequencing (WGS)[27], we found two samples harboring

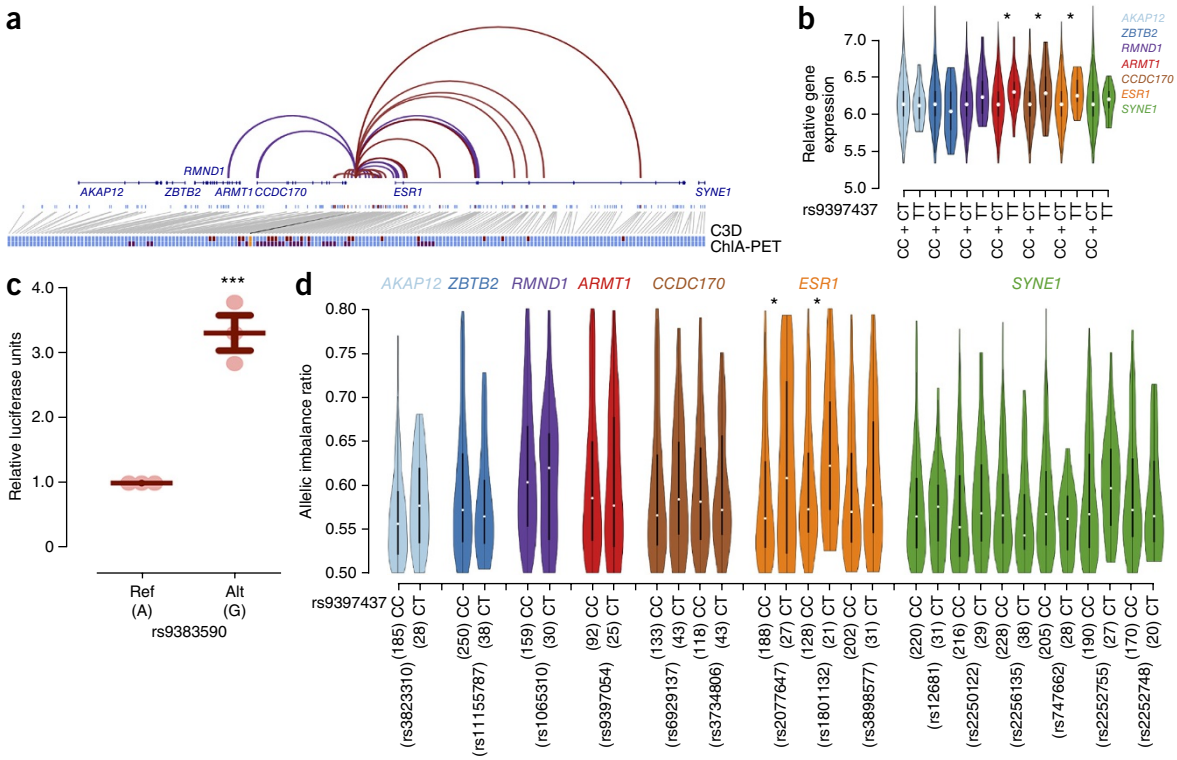

Figure 2 The rs9383590 SNV interacts with the *ESR1* promoter altering gene expression. (**a**) C3D-predicted (red) and Pol II ChIA-PET-determined (purple) chromatin interactions between the breast cancer risk–associated locus enhancer DHS and neighboring DHSs. Single DHS resolution is presented for the C3D approach. All DHSs within a paired-end tag are considered as interacting in ChIA-PET data. DHSs with no evidence of a chromatin interaction are shown in light blue; the position of nearby genes is also shown (top). Predicted (red) and experimentally determined (purple) DHSs interacting with the breast cancer risk–associated locus enhancer (orange) are enlarged to reveal the overlap (bottom). (**b**) Gene expression values for the genes at the *ESR1* locus by rs9397437 genotype ($n = 970$). $^*P < 0.05$, linear regression under a recessive model. (**c**) Reporter assay results for rs9383590. Experiments were performed in triplicate. $^{***}P < 0.005$, two-sided one-sample *t*-test. (**d**) Allelic imbalance of the genes at the *ESR1* locus among TCGA breast tumors profiled by RNA-seq. The allelic imbalance ratio represents the frequency of the most abundant allele within the RNA-seq reads. $^*P < 0.05$, two-sided approximate Fisher-Pitman test using 10,000 permutations. Data are mean ± s.e.m.

a somatic mutation in the enhancer modulated by the rs9383590 SNV (**Fig. 3a**). Because the SRE of a gene tightly regulates its expression[28], we hypothesized that mutations within the SRE of *ESR1* could account for its persistent expression in breast cancer. We first delineated the SRE of *ESR1* using the C3D method. This predicted the physical interaction of 24 regulatory elements with the *ESR1* promoter within a 1-Mb window of its transcription start site ($r \geq 0.7$) (**Supplementary Table 1**). Eighteen of these predicted interactions were validated by first- or second-order interactions identified in the Pol II ChIA-PET data sets[19] (**Supplementary Fig. 5**). We then identified mutations in the *ESR1* SRE in approximately 10% of the 98 WGS breast cancer samples (10/98). Nine of these mutations are found in seven enhancers, and one is in the *ESR1* promoter (**Fig. 3b**). We validated the interaction between all mutated enhancers and the *ESR1* promoter by chromatin conformation capture–based assays in MCF-7 cells (**Supplementary Fig. 5**). Of note, each mutated enhancer was flanked by nucleosomes containing histone H3 acetylated on Lys27 (H3K27ac) in breast cancer cells, a feature of active enhancer elements[29] (**Fig. 3c**).

To determine whether the burden of mutations found in the SRE of *ESR1* is significantly more than expected by chance, we designed a conservative analytical approach, termed mutational significance within the regulatory element set (MuSE) (**Fig. 3d** and Online Methods). Briefly, with this approach we consider all regulatory elements, or DHSs, predicted to interact with the *ESR1* promoter as a single unit, which is analogous to splicing together the exons of a gene. We then test for an excess of mutations within the *ESR1* SRE using a binomial probability test given a genome-wide mutation rate (gBMR) and local background mutation rate (lBMR). The gBMR is calculated from all DHSs including the *ESR1* SRE. The lBMR is calculated from the DHSs surrounding the *ESR1* gene that are not connected to its promoter on the basis of C3D (**Fig. 3d**). Each type of mutation is tested separately, and the *P* values are combined using Fisher's method (Online Methods). This approach identified a significant enrichment of noncoding somatic mutations within the *ESR1* SRE (SRE $r \geq 0.7$; size = 20,744 bp; $n = 10$ mutations observed; $P = 8.06 \times 10^{-3}$) (**Fig. 3b** and **Supplementary Table 2**). The number of nucleotides considered exceeds what is typical of coding sequences, which hinders the statistical significance. For example, the median length of a human protein is 375 amino acids[30], which corresponds to 1,125 nucleotides. For comparison, the SRE of *ESR1* is 20,744 nucleotides. Increasing the correlation threshold used for the C3D-predicted promoter–enhancer interactions improved the significance of the measured enrichment of mutations in the *ESR1* SRE, despite the inclusion of fewer mutations ($r \geq 0.9$; 7,746 bp; $n = 6$; $P = 2.57 \times 10^{-4}$). The statistical enrichment was also improved by restricting the analysis to ESR1-positive tumors ($r \geq 0.9$; 7,746 bp; $n = 5$; $P = 7.02 \times 10^{-5}$) (~7% (5/73)) (**Supplementary Table 2**). The mutational significance appeared to be specific to breast cancer mutations, as we did not detect an enrichment of somatic mutations within the *ESR1* SRE defined in breast cancer cells using mutations

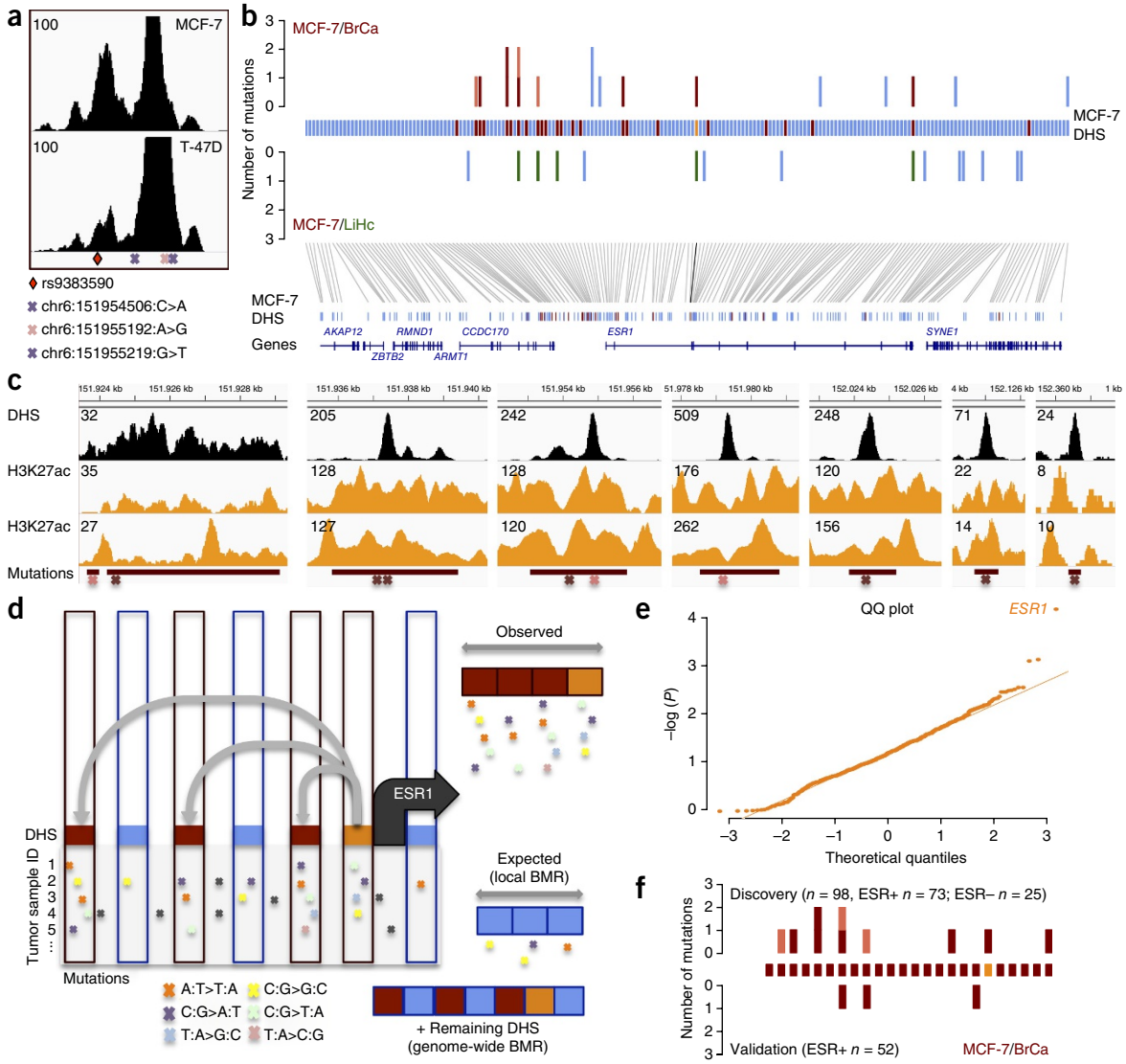

Figure 3 The SRE of *ESR1* is targeted by acquired somatic mutations in breast cancer. (**a**) DNase-seq signal across the enhancer harboring the rs9383590 SNV and three somatic mutations (two in the discovery set of breast tumors and one in the validation set of breast tumors). (**b**) Enrichment of mutations with DHSs that interact (red) or not (blue) with the *ESR1* promoter (orange) in 98 breast tumors (BrCa). The number of mutations identified in ESR1-positive (red) and ESR1-negative (pink) samples are shown. Lack of enrichment within the SRE for mutations from 88 liver tumors (LiHc, green) is also shown. (**c**) DNase-seq and H3K27ac ChIP-seq signal profiles (from ENCODE[19] and Taberlay *et al.*[36]) for regulatory elements harboring a somatic mutation in breast tumors. (**d**) Schematic representation of the mutational significance within the *ESR1* SRE (MuSE) approach. C3D-predicted (gray lines) DHSs (red rectangles) interacting with the *ESR1* gene promoter (orange rectangle) and noninteracting DHSs (blue rectangles) are shown. The mutational load in the interacting versus noninteracting DHSs define the observed versus expected mutational rate in the *ESR1* SRE. (**e**) A QQ plot of the observed −log(*P* values) for the mutational significance of all SREs defined using MCF-7 cells ($r \geq 0.9$). (**f**) Mutational burden within the *ESR1* (± 250 kb) SRE for the discovery (top) and validation (bottom) samples. Enhancers are ranked according to **c**. Red indicates mutations found in ESR1-positive (ESR+) tumors, and pink indicates mutations found in ESR1-negative (ESR–) tumors. Red squares indicate interacting enhancers; orange square indicates the *ESR1* gene promotor.

called in WGS of 88 liver hepatocellular carcinomas[27] (**Fig. 3b**). To determine whether the observed enrichment is greater than expected by chance, we performed a genome-wide MuSE analysis restricted to mutations called in ESR1-positive breast cancer. Focusing on all RefSeq-annotated genes with a promoter DHS in MCF-7 cells connecting to at least one regulatory element (C3D $r \geq 0.9$), we found a significant enrichment of mutations in only the SRE of *ESR1* (**Fig. 3e**) (FDR q-value = 0.045).

To independently investigate whether the *ESR1* SRE is recurrently altered in breast cancer, we sequenced the *ESR1* SRE in a set of 52 primary ESR1-positive breast tumors from the IMPACT (NCT01505400) and COMPACT trials. We identified three (~6%) somatic point mutations (chr6:151955219:G>T; chr6:151979547:A>G; chr6:152075097:G>C) within enhancers interacting with the *ESR1* promoter (C3D; $r \geq 0.7$) (**Fig. 3f**). These mutations had a tumor fraction of 0.42, 0.32 and 0.03, respectively (**Supplementary Table 3**). The chr6:151955219:G>T falls within the enhancer altered by the rs9383590 SNV (**Fig. 3a**) and is located 27 bp away from the previously characterized chr6:151955192:A>G mutation.

Similarly to inherited risk variants, noncoding somatic mutations can affect transcription factor activity[31]. All somatic mutations found in the *ESR1* SRE fell within or mapped near relevant transcription factor DNA recognition motifs (**Fig. 4a** and **Supplementary Fig. 6**). In addition, all mutations were predicted by IGR[6] to modulate the

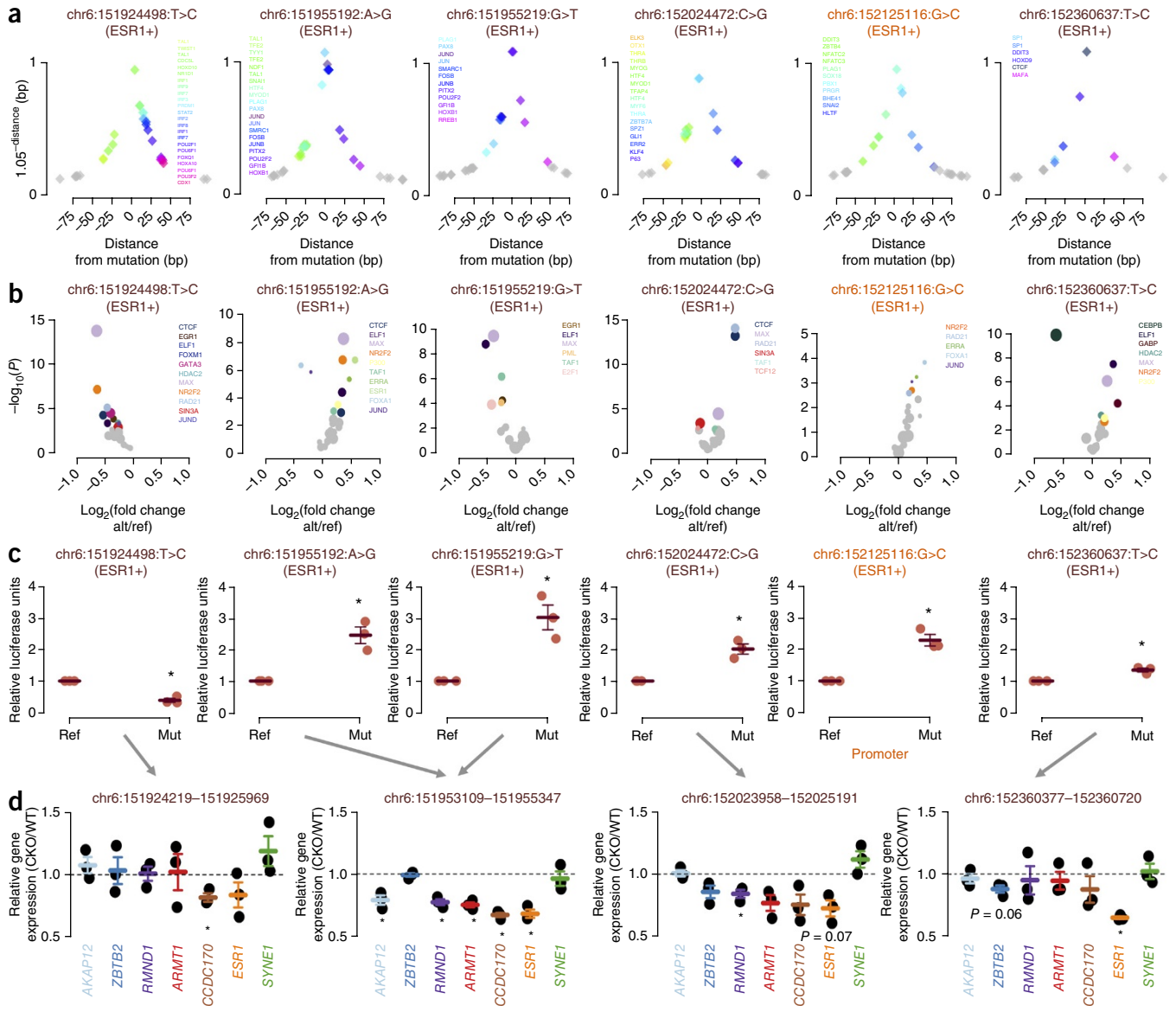

Figure 4 Noncoding somatic mutations targeting *ESR1* increase gene expression. (**a**) Distance between transcription factor DNA recognition motifs and identified mutations. The *y*-axis shows a function ($1.05^{-\text{distance}}$) of the distance to each mutation to emphasize the closest motifs. This function has a range of 0–1 within 100 bp of the mutation. Data points represent the location of a transcription factor DNA recognition motif. (**b**) *P* value versus fold change in chromatin binding intensity predicted by IGR analysis of transcription factors for each mutation in the *ESR1* SRE. All transcription factors profiled by ChIP-seq in MCF-7 or T-47D by ENCODE[19] were tested for each mutation. Those whose binding intensity was predicted to be modulated by the mutations ($P < 0.005$) are indicated in color. (**c**) Reporter assays to assess the impact of six mutations targeting the *ESR1* SRE in ESR1-positive breast tumors on gene expression. $^*P < 0.05$, one-sample *t*-test. Error bars indicate mean ± s.e.m. (**d**) Gene expression levels assessed by RT-qPCR in wild-type T-47D (WT) and T-47D cells with CRISPR–Cas9-based deletion of the respective enhancer (CKO) region. alt, alternate allele; ref, reference allele; mut, mutant allele. All experiments were performed in triplicate. All *P* values are two-sided. $^*P < 0.05$, one-sample *t*-test.

chromatin binding of known regulators of *ESR1* expression, including GATA3, cohesin, SIN3A and ESR1 (**Fig. 4b** and **Supplementary Fig. 7**). Ten out of eleven tested mutations significantly altered the transactivation potential of their regulatory elements (**Fig. 4c** and **Supplementary Fig. 8**). Next, we focused on the mutations within the four enhancers with the strongest predicted interaction with the *ESR1* promoter ($r \geq 0.9$) and the promoter itself. These elements correspond to the MuSE analysis that passed multiple testing correction (FDR < 0.05). All six mutations affecting these regulatory elements identified in ESR1-positive tumors had a significant impact on their transactivation potential ($P < 0.05$), including the chr6:151955219:G>T mutation from the validation set (**Fig. 4c**). The chr6:151924498:T>C mutant allele decreased enhancer activity compared to the wild-type sequence, but the remaining five (83%) mutant alleles significantly increased reporter gene expression compared to the wild-type sequence ($P < 0.05$). We confirmed the regulatory role of these enhancers on *ESR1* gene expression by deleting each of the affected enhancers using the CRISPR–Cas9 system in T-47D cells stably expressing Cas9 (Online Methods). The deletion of two of the enhancers significantly decreased *ESR1* expression, and a trend was observed for the remaining enhancers (**Fig. 4d**). Although the deletions are relatively large, they correspond to a small fragment of the SRE, suggesting a substantial contribution from single elements to *ESR1* expression.

The deletion of the enhancer harboring the rs9383590 SNV, on chr6:151953109–151955347, led to a significant decrease in *ESR1* expression (**Fig. 4d**), further indicating a direct impact of this SNV on *ESR1* expression. This enhancer also harbors three mutations. Mutations chr6:151955192:A>G and chr6:151955219:G>T were discovered in ESR1-positive tumors; however, the chr6:151954506:C>T mutation was found within an ESR1-negative tumor. Notably, this recapitulates the observed association between the GWAS lead SNVs at the *ESR1* locus and both ESR1-positive and ESR1-negative breast cancers and suggests the presence of additional oncogenes coregulated with *ESR1* at this locus. Of note, deletion of this enhancer in the T-47D cells stably expressing Cas9 also led to a significant reduction in the expression of *ARMT1*, *CCDC170* and *RMND1* (**Fig. 4d**). Coexpression[32] and coregulation[25] of *ESR1*, *CCDC170* and *RMND1* has been reported, and their allele-specific expression may account for the association between variants at the *ESR1* locus and ESR1-negative breast cancer[33]. We found that silencing *RMND1* significantly ($P < 0.05$) reduced the proliferation of ESR1-positive and ESR1-negative breast cancer cells (MCF-7 and MDAMB436, respectively) (**Supplementary Fig. 9**).

By demonstrating that the inherited risk variant and somatic point mutations that populate the SRE of *ESR1* behave as gain-of-function genetic alterations, our results provide a mechanism that could explain the sustained expression of *ESR1* in approximately 7% of ESR1-positive breast cancer patients. This finding contrasts with gain-of-function coding mutations that typically present as mutations recurrently targeting a single codon[34]. Hence, noncoding mutational hot spots may be rare. Instead, mutations affecting distinct regulatory elements converging on the same gene, such as those reported here, may represent the mutational pattern of noncoding driver mutations. These do not need to directly target DNA recognition motifs. Indeed, recent work has found that noncoding mutations can influence transcription factor activity despite falling outside DNA recognition motifs[35]. Taken together, our results support the idea that noncoding mutations relevant to cancer development and the genes they target can be identified by an SRE-focused approach that is inclusive of mutations outside of DNA recognition motifs.

URLs. The Cancer Genome Atlas (TCGA), http://cancergenome.nih.gov/; the Princess Margaret Genomics Centre, http://www.pmgenomics.ca; the European Genome–phenome Archive, https://www.ebi.ac.uk/ega/home; TCGA data portal, https://tcga-data.nci.nih.gov/tcga/tcgaDownload.jsp; The 1000 Genomes Project, http://www.1000genomes.org; the Cancer Genomics Hub, https://browser.cghub.ucsc.edu/; R, http://www.r-project.org; Picard, http://broadinstitute.github.io/picard/; Hotspot algorithm, https://github.com/rthurman/hotspot; Integrated Molecular Profiling in Advanced Cancer Trial (IMPACT) and Community Oncology Molecular Profiling in Advanced Cancer Trial (COMPACT), Gene Expression Omnibus, http://www.ncbi.nlm.nih.gov/geo/; MACS2, https://github.com/taoliu/MACS. dbSNP human build 142, ftp://ftp.ncbi.nlm.nih.gov/snp/organisms/human_9606/.

METHODS

Methods and any associated references are available in the online version of the paper.

Accession codes. Gene Expression Omnibus: data have been deposited under accession codes GSE74718.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank A. Razak, C. Elser, D. Cescon, D. Warr, E. Amir, L. Siu, N. Leighl and S. Sridhar for their involvement in recruiting the IMPACT and COMPACT samples used in this study. We also thank M. Lemaire for helpful discussions. We thank R. Rottapel and O. Kent for use of and help with the Glomax Multi-Detection system. We acknowledge the ENCODE consortium and the ENCODE production laboratories that generated the data sets provided by the ENCODE Data Coordination Center used in the manuscript. We also acknowledge the Cancer Genome Project, for making all the breast cancer and liver cancer called mutations publicly available, and the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC), for making the genotyping and expression data from primary breast tumors data available. We acknowledge the Princess Margaret Genomics Centre and the Bioinformatics group for providing the infrastructure assisting us with the targeted sequencing and analysis of the *ESR1* SRE. Supported by the National Cancer Institute (NCI) at the National Institute of Health (NIH) (R01CA155004 to M.L.), the Princess Margaret Cancer Foundation (T.J.P. and M.L.), The Canadian Cancer Society (CCSRI702922 to M.L.), the Susan G. Komen Foundation (CCR15332792 to T.J.P.) and the Gattuso-Slaight Personalized Cancer Medicine Fund/PMCF (B.H.-K.). M.L. is funded by a young investigator award from the Ontario Institute for Cancer Research (OICR), a new investigator salary award from the Canadian Institute of Health Research (CIHR) and a Movember Rising Star award from Prostate Cancer Canada (PCC) (RS2014-04). K.J.K. and R.C.P. are supported by Canadian Breast Cancer Foundation (CBCF) postdoctoral fellowships. S.D.B. is supported by a Knudson and CIHR postdoctoral fellowship.

AUTHOR CONTRIBUTIONS

The concept of interrogating the mutational load in regulatory elements converging on single genes arose through discussions between S.D.B., N.A.S.-A., R.C.S. and M.L. S.D.B. designed and/or implemented all the computational and statistical approaches except for IGR and analyzed the results under the supervision of M.L. Experimental assessment of the effect of SNVs on enhancer activity, transcription factor binding and gene expression was designed by K.D., S.D.B. and M.L. and conducted by K.D. with assistance from K.J.K., A.E.T. and X.W. The CRISPR–Cas9-based enhancer deletion was conducted by K.D., K.J.K., K.L.T., J.S. and D.W.C. under the supervision of T.W.M. and M.L. P.M. and N.A.S.-A. implemented the IGR approach to predict allele-bias binding of transcription factors on SNVs after improvements to IGR by N.A.S.-A. and R.C.S. R.C.P. and P.L.B. assessed the ESR1, PR and HER2 expression status on primary breast tumors included in our validation cohort. S.Y.C.Y. performed the alignment and gene expression quantification of the TCGA RNA-seq data. M.D. assisted in DNA capture sequencing of the primary breast tumor validation cohort under T.J.P.'s supervision. B.H.-K. oversaw the expression analysis of the METABRIC data set. M.L. oversaw the project. Figures were designed and prepared by S.D.B. and K.D. The manuscript was written by S.D.B., K.D. and M.L. with assistance from all other authors.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Green, K.A. & Carroll, J.S. Oestrogen-receptor-mediated transcription and the influence of co-factors and chromatin state. *Nat. Rev. Cancer* **7**, 713–722 (2007).
2. Vincent-Salomon, A., Raynal, V., Lucchesi, C., Gruel, N. & Delattre, O. *ESR1* gene amplification in breast cancer: a common phenomenon? *Nat. Genet.* **40**, 809, author reply 810–812 (2008).
3. Brown, L.A. *et al. ESR1* gene amplification in breast cancer: a common phenomenon? *Nat. Genet.* **40**, 806–807, author reply 810–812 (2008).
4. Horlings, H.M. *et al. ESR1* gene amplification in breast cancer: a common phenomenon? *Nat. Genet.* **40**, 807–808, author reply 810–812 (2008).
5. Reis-Filho, J.S. *et al. ESR1* gene amplification in breast cancer: a common phenomenon? *Nat. Genet.* **40**, 809–810, author reply 810–812 (2008).
6. Cowper-Sallari, R. *et al.* Breast cancer risk-associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nat. Genet.* **44**, 1191–1198 (2012).
7. Maurano, M.T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
8. Schaub, M.A., Boyle, A.P., Kundaje, A., Batzoglou, S. & Snyder, M. Linking disease associations with regulatory information in the human genome. *Genome Res.* **22**, 1748–1759 (2012).
9. Polak, P. *et al.* Reduced local mutation density in regulatory DNA of cancer genomes is linked to DNA repair. *Nat. Biotechnol.* **32**, 71–75 (2014).
10. Polak, P. *et al.* Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518**, 360–364 (2015).
11. Weedon, M.N. *et al.* Recessive mutations in a distal *PTF1A* enhancer cause isolated pancreatic agenesis. *Nat. Genet.* **46**, 61–64 (2014).

12. Horn, S. *et al. TERT* promoter mutations in familial and sporadic melanoma. *Science* **339**, 959–961 (2013).
13. Huang, F.W. *et al.* Highly recurrent *TERT* promoter mutations in human melanoma. *Science* **339**, 957–959 (2013).
14. Zheng, W. *et al.* Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. *Nat. Genet.* **41**, 324–328 (2009).
15. Fletcher, O. *et al.* Novel breast cancer susceptibility locus at 9q31.2: results of a genome-wide association study. *J. Natl. Cancer Inst.* **103**, 425–435 (2011).
16. Turnbull, C. *et al.* Genome-wide association study identifies five new breast cancer susceptibility loci. *Nat. Genet.* **42**, 504–507 (2010).
17. Siddiq, A. *et al.* A meta-analysis of genome-wide association studies of breast cancer identifies two novel susceptibility loci at 6q14 and 20q11. *Hum. Mol. Genet.* **21**, 5373–5384 (2012).
18. Stacey, S.N. *et al.* Ancestry-shift refinement mapping of the *C6orf97–ESR1* breast cancer susceptibility locus. *PLoS Genet.* **6**, e1001029 (2010).
19. ENCODE Project Consortium. A user's guide to the Encyclopedia of DNA Elements (ENCODE). *PLoS Biol.* **9**, e1001046 (2011).
20. 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
21. Thurman, R.E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
22. Bailey, S.D. *et al.* ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters. *Nat. Commun.* **2**, 6186 (2015).
23. Curtis, C. *et al.* The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature* **486**, 346–352 (2012).
24. Li, Q. *et al.* Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell* **152**, 633–641 (2013).
25. Dunning, A.M. *et al.* Breast cancer risk variants at 6q25 display different phenotype associations and regulate *ESR1*, *RMND1* and *CCDC170*. *Nat. Genet.* **48**, 374–386 (2016).
26. Frietze, S. *et al.* Cell type–specific binding patterns reveal that TCF7L2 can be tethered to the genome by association with GATA3. *Genome Biol.* **13**, R52 (2012).
27. Alexandrov, L.B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
28. Sanyal, A., Lajoie, B.R., Jain, G. & Dekker, J. The long-range interaction landscape of gene promoters. *Nature* **489**, 109–113 (2012).
29. Ernst, J. & Kellis, M. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nat. Biotechnol.* **28**, 817–825 (2010).
30. Brocchieri, L. & Karlin, S. Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Res.* **33**, 3390–3400 (2005).
31. Bell, R.J. *et al.* The transcription factor GABP selectively binds and activates the mutant *TERT* promoter in cancer. *Science* **348**, 1036–1039 (2015).
32. Dunbier, A.K. *et al. ESR1* is co-expressed with closely adjacent uncharacterised genes spanning a breast cancer susceptibility locus at 6q25.1. *PLoS Genet.* **7**, e1001382 (2011).
33. Yamamoto-Ibusuki, M. *et al. C6ORF97–ESR1* breast cancer susceptibility locus: influence on progression and survival in breast cancer patients. *Eur. J. Hum. Genet.* **23**, 949–956 (2015).
34. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).
35. Katainen, R. *et al.* CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat. Genet.* **47**, 818–821 (2015).
36. Taberlay, P.C., Statham, A.L., Kelly, T.K., Clark, S.J. & Jones, P.A. Reconfiguration of nucleosome-depleted regions at distal regulatory elements accompanies DNA methylation of enhancers and insulators in cancer. *Genome Res.* **24**, 1421–1432 (2014).

ONLINE METHODS

Genotype calling, linkage disequilibrium and multidimensional scaling. The raw genotype data of the METABRIC samples[23] were downloaded from the European Genome–phenome Archive (EGAS00000000083). The raw genotype data of the TCGA samples were downloaded from the TCGA data portal. The genotypes of the METABRIC and TCGA samples were called using Birdseed[37]. The phase 3 genotype data for the 1000 Genomes Project samples[20] were downloaded from the 1000 Genomes website. Linkage disequilibrium (LD) was calculated using VCFtools[38] within the continental European (CEU, TSI, FIN, GBR and IBS; $n = 503$) and East Asian (CHB, JPT, CHS, CDX and KHV; $n = 504$) groups. The SNVs in LD with GWAS lead SNVs are presented in **Supplementary Table 4**. Ancestry was determined by merging the genotype data with the 1000 Genomes samples and performing multidimensional scaling (MDS) of the genotype data using PLINK[39] (**Supplementary Figs. 10** and **11**).

Intragenomic replicates (IGR). The functional impact of SNVs on transcription factor binding was predicted using the IGR tool as previously described[6]. Briefly, we compare the average ChIP-seq signal intensity across genomic loci that contain short DNA sequences 7 nt in length (7-mers) that match the reference allele and its surrounding DNA sequence against the average signal intensity at genomic loci that contain 7-mers that differ only by the variant allele of each SNV. All 7-mers from a sliding window surrounding each SNV are considered. The 7-mer with the highest average intensity matching the reference allele is tested against the 7-mer with the highest average intensity that matches the variant allele. The genomic locations of all 7-mers are filtered to include only sites within open chromatin. The wgEncodeUwDnaseMcf7PkRep1.narrowPeak and wgEncodeUwDnaseT47dPkRep1.narrowPeak called DHSs produced as part of the ENCODE project[19] were used as filters for the MCF-7 and T-47D cells, respectively. The aligned ChIP-seq files were downloaded from the ENCODE website. The complete list of files used in the analysis is available in **Supplementary Table 5**. Signal files were generated using MACS[40]. All analyses were performed with hg19.

eQTL analysis. We used the sample of breast tumors profiled by METABRIC[23] and TCGA. The expression data for the METABRIC samples were downloaded from the European Genome–phenome Archive. The RNA-seq data for the TCGA breast cancer samples were downloaded from the Cancer Genomics Hub.

The reads were aligned to human reference GRCh37 with Gencode version 15 human transcript annotation using STAR[41] in two-pass mode. Gene level expression values for each sample were quantified using Cufflinks[42]. The expression of *ESR1* is bimodal in METABRIC and TCGA and is explained by ESR1-positive and ESR1-negative tumors (**Supplementary Fig. 12**). Consistent with the METABRIC[23] analysis, we determined the expression status of TCGA tumors for *ESR1*, *PGR* and *ERBB2* among TCGA samples using MClust in R. We fitted a Gaussian finite mixture model with two components. We restricted the analysis to luminal-type tumors, those that express both *ESR1* and *PGR*, but do not overexpress *ERBB2*. TCGA tumor samples with low expression of *ERBB2* were also identified as belonging to a separate distribution by MClust and were removed. We merged the identified luminal METABRIC discovery and validation samples and performed a quantile normalization of the merged samples using the preprocessCore library[43] in R. Statistical significance was determined using linear regression under an additive and recessive model. The reported *P* values are two-sided. To control for potential population stratification we included the first three components of the MDS analysis as covariates. The rs9397437 SNV was used as a proxy for the rs9383590 SNV ($r^2 = 1.0$ and $r^2 = 1.0$ among Europeans and East Asians, respectively). The gene expression values stratified by SNV genotype are presented in **Supplementary Figure 13**.

Allelic imbalance. We analyzed the TCGA breast tumors profiled by RNA-seq. Duplicate reads were removed using Picard. The number of aligned reads containing either the reference or variant alleles of coding marker SNVs was determined using the ABC tool[44]. The default settings of ABC were used. Marker SNVs were identified by intersecting the common SNV database (dbSNP human build 142) with refSeq exon annotations using bedTools[45]. We calculated the allelic imbalance (AI) ratio as the number of reads containing either the reference or variant allele, whichever was larger, divided by the total number of reads. We removed samples with an AI ratio greater than 0.8, as they could represent sequencing error within homozygous individuals[24]. Individuals heterozygous for the rs9397437 SNV, a proxy of the rs9383590 SNV, were compared to individuals homozygous for the common allele using an approximate Fisher-Pitman test with 10,000 permutations implemented in the coin library in R. We included markers SNVs with at least 20 samples heterozygous for the rs9397437 SNV.

Defining sets of regulatory elements (SREs). We took advantage of the known relationship between the cross-cell-type correlation in DNase I hypersensitivity signals (C3D) and chromatin interactions[21] to predict connections between regulatory elements. We used the uniformly processed DNase I hypersensitivity sequencing signal files for 79 cell lines available from the ENCODE project[19]. We performed the correlation of DNase I signal intensities in a cell type–specific manner by interrogating only DHSs identified in the MCF-7 cell line[22]. The DHSs used in our study were identified by the Hotspot algorithm[46] and produced as part of the ENCODE project[19]. We validated the predicted interactions called for breast cancer with a Pol II ChIA-PET data, profiled in MCF-7 cells, created by the ENCODE project[19]. We combined all four replicates for our analyses.

Calculating mutational significance within the regulatory element set (MuSE). DHSs predicted to interact with the gene promoter at a given correlation threshold ($r \geq 0.7$–0.9) are combined to create the test region or SRE. We use the binomial test implement in R to assess whether the observed number of mutations within the test region is greater than expected given both a genome-wide and local background mutation rate (lBMR and gBMR, respectively). The approach is comparable to that employed by MutSig[47] and MuSiC[48] but is applied to noncoding regions and mutations. We calculate the lBMR using the remaining DHSs that are below the correlation threshold but within the specified window surrounding the anchor DHS. This approach is thought to be conservative, as it is possible that mutations included in the lBMR are functional. It is important that the lBMR and gBMR be calculated from DHSs and that these DHSs be cell type–specific, because somatic mutations have been shown to preferentially fall in heterochromatic noncoding regions in a cell type–specific manner[9,10]. To control for different rates of mutations, a separate binomial test is performed for each of the six mutation types (n), and a final combined P value is calculated using Fisher's method from a χ^2 distribution with $2n$ degrees of freedom in R. Only one mutation within the test region is counted per tumor. However, all mutations contribute to the BMR calculation, which again should be conservative. If we are unable to calculate the lBMR for a given mutation type, because we do not observe a mutation within the window, we use the gBMR for that mutation type. We excluded tumors profiled by whole-exome sequencing, because they are typically sequenced to a greater depth and regulatory elements co-occur with coding sequencings[49].

Mutation data (discovery). The breast and liver cancer–associated mutations used in the MuSE calculations were reported by Alexandrov *et al.*[27]. We used the cleaned mutation data set in all analyses. We included only those samples with known ESR1 status ($n = 98$) in our analysis[50]. The identifiers of ESR1 positivity and triple-negative breast cancer (TNBC) tumors are listed in **Supplementary Table 6.**

Targeted sequencing of the *ESR1* SRE (validation). We validated the enrichment of mutations within the *ESR1* SRE in an independent set of 52 primary ESR1-positive breast tumors and matched normal blood samples from the Integrated Molecular Profiling in Advanced Cancer Trial (IMPACT) and the Community Oncology Molecular Profiling in Advanced Cancer Trial (COMPACT) trials conducted at the Princess Margaret Cancer Centre (PMCC). The research ethics board of the University Health Network (UHN) approved the retrospective analysis of the breast cancer samples. Informed consent was obtained from all study participants. We used hybrid capture to isolate the *ESR1* SRE elements using a custom panel of xGen Lockdown Probes (Integrated DNA Technology Inc). The 120-bp probes were spaced 60 bp apart. The probe sequences and the targeted regions are available in **Supplementary Tables 1** and **7**. Captured fragments were sequenced using 150-bp paired-end reads from a NextSeq 500 sequencer (Illumina) at the Princess

Margaret Genomics Centre. Tumors and normal samples were sequenced to median >600× coverage.

Calling somatic point mutations (validation). Reads were aligned to the human reference genome, hg19, using BWA[51]. Base recalibration and realignment around insertions and deletions (indels) was performed with GATK[52]. Duplicate reads were marked with Picard. Somatic point mutations were called from tumor–normal pairs using muTect[53]. The identified mutations are listed in **Supplementary Tables 3** and **8**.

Localization of transcription factor motifs surrounding mutations. We searched the flanking sequences (±100 bp) of each somatic mutation for human transcription factor DNA recognition motifs using position weight matrices compiled in the *Homo sapiens* Comprehensive Model Collection (HOCOMOCO)[54] with FIMO[55].

Annotation of LD SNVs and identification of active enhancers. We downloaded all available called peaks and signal files for the two breast cancer model cell lines, MCF-7 and T-47D, produced as part of the ENCODE project[19]. To verify the identified enhancers, we downloaded additional H3K27ac ChIP-seq data profiled by Taberlay *et al.*[36] from the Gene Expression Omnibus (GEO GSM1383859). The reads were aligned to the reference genome using BWA[51] and signal files were generated using MACS2 (ref. 40).

Cell culture. HCC1419 cells (ATCC) were grown to 95% confluence in RPMI 1640 medium supplemented with 10% FBS. MCF-7 (in house), T-47D (in house) and MDAMB436 (in house) cell lines were grown in DMEM supplemented with 10% FBS. The cell lines used in this study have not been listed as cross-contaminated or as commonly misidentified by the International Cell Line Authentication Committee (ICLAC). All cell lines were determined to be mycoplasma-free using the EZ-PCR mycoplasma test kit (Biological Industries).

Western blotting. We verified the ESR1 protein expression status of the HCC1419 and MDAMB436 by western blot. Cells were lysed for 5 min on ice in lysis buffer (1% SDS, 10 mM EDTA and 50 mM Tris-HCl, pH 8.1) supplemented with a protease inhibitor cocktail (Roche), followed by sonication. The protein concentration of the lysates was determined using a Pierce Micro BCA Protein Assay Kit (Thermo Scientific). Equal amounts of protein (25 µg) were electrophoresed on TGX Protein Gels (Bio-Rad), and then transferred to a PVDF membrane (Bio-Rad). Membranes were blocked with 5% BSA (AMRESCO) in Tris-buffered saline (TBS) with 0.1% Tween-20 (Fisher Scientific) for 2 h at room temperature. Primary antibodies were incubated overnight at 4 °C. Western blots for ESR1 were performed with the rabbit anti-ESR1 antibody (sc-543X, Santa Cruz), and blots for β-actin were performed using the rabbit anti-β-Actin (4970, Cell Signaling). Membranes were washed then probed with anti-rabbit IgG, HRP-linked antibody (7074, Cell Signaling) at room temperature for 1 h. Bands were visualized with ECL Western Blotting Detection Reagent (GE Healthcare) and scanned using the FluorChemQ (Alpha Innotech).

Cell viability assays after silencing of *RMND1*. Two sets of small interfering RNA (siRNA) against *RMND1* were designed through Thermo Fisher Scientific (RMND1 Silencer Select #s29968 and s29969, catalog #4392420). A scrambled siRNA was used as a negative control. RNA was isolated from MCF-7 and MDAMB436 cells using the RNeasy Mini kit (Qiagen) according to the manufacturer's protocol. Reverse transcription (RT) was performed to convert RNA into cDNA (iScript cDNA Synthesis Kit, Bio-Rad). The resulting cDNA was subjected to qPCR using SensiFAST SYBR (Bioline) to confirm the silencing effect of the siRNAs against *RMND1*. Expression levels were quantified using the ΔΔCt method with actin as the calibrator[56] and then normalized to *RMND1* RNA levels in cells that were treated with negative control siRNA. The primers are as listed in **Supplementary Table 9**.

The relative proportion of viable MCF-7 and MDAMB436 breast cancer cells were measured using AlamarBlue reagent (Thermo Fisher Scientific) 72 h after silencing of *RMND1*. AlamarBlue reagent was added to the wells and the cells were incubated at 37 °C for 4 h. Fluorescence was read at excitation and emission wavelengths of 550 nm and 590 nm respectively.

RNA-seq and allele-specific expression within HCC1419 cells. Total RNA was extracted from HCC1419 cells using RNeasy Mini Kit (Qiagen) according to manufacturer's instructions. Two RNA-seq libraries were prepared using the Truseq Stranded mRNA kit (Illumina). RNA was sequenced using 75-bp paired-end reads. Reads were mapped to the reference genome, hg19, using TopHat[42]. Allele-biased expression was called using a binomial test with the ABC tool[44].

Allele-biased chromatin immunoprecipitation. HCC1419 cells were cross-linked with 1% formaldehyde at 37 °C for 10 min. Cells were rinsed with ice-cold PBS plus 5% BSA followed by PBS and harvested with PBS plus 1× protease inhibitor cocktail (Roche Molecular Biochemicals). Harvested cells were centrifuged for 2 min at 3,000 r.p.m. Cells were lysed in 0.35 mL of lysis buffer (1% SDS, 10 mM EDTA, 50 mM Tris-HCl, pH 8.1, 1× protease inhibitor cocktail) by sonication (Diagenode Biorupter 300). The lysed cells were subjected to centrifugation at maximum speed for 15 min. Supernatants were collected and diluted in dilution buffer (1% Triton X-100, 2 mM EDTA, 150 mM NaCl, 20 mM Tris-HCl, pH 8.1).

12.5 µg GATA3-specific antibody (Santa Cruz, sc268x) was prebound for 6 h to protein A and protein G Dynal magnetic beads (Dynal Biotech) and washed three times with ice-cold PBS plus 5% BSA and then added to the diluted chromatin for overnight immunoprecipitation. The magnetic bead–chromatin complexes were collected and washed six times in RIPA buffer (50 mM HEPES, pH 7.6, 1 mM EDTA, 0.7% sodium deoxycholate, 1% NP-40, 0.5 M LiCl) and then washed twice with TE buffer. Cross-linking was reversed with de-cross-linking buffer (1% SDS, 0.1 M NaHCO3) overnight at 65 °C. DNA fragments were purified with a QIAquick Spin Kit (Qiagen, CA). Allele-biased binding was assessed using MAMA primer–based qPCR[57] and verified by Sanger sequencing (**Supplementary Fig. 14**). Fold enrichment was calculated over input. Significance of the differential enrichment was calculated using unpaired *t*-test. A complete list of primers is available in **Supplementary Table 9**.

Luciferase reporter assays. Each enhancer was PCR amplified using PfuUltraII fusion polymerase from Human DNA (Roche Molecular Biochemicals) and cloned at the BamHI (BamHI-HF, NEB) restriction site into the pGL3 and pGL4.23 promoter vector (Promega) in the antisense and/or sense direction. Site-directed mutagenesis was performed using QuickChange XLII kit (Agilent) according to manufacturer's instructions to generate the mutant alleles. The results of luciferase assay in the sense orientation are presented in **Supplementary Figure 15**. All sequences were verified by Sanger sequencing (**Supplementary Fig. 16**). Wild-type and mutant enhancer constructs were independently transfected in T-47D cells grown in estrogen-depleted medium together with a *Renilla* reporter plasmid at a 1:100 ratio. 48 h after transfection, the cells were stimulated with full medium, and luciferase activity was measured using the Dual-Luciferase Reporter Assay System (Promega). Final readings are reported as firefly luciferase normalized to *Renilla* luciferase activity per well. Fold change in luciferase activity of the mutant allele compared to the reference allele was calculated. The values were log transformed, and significance was tested using a one-sample *t*-test. Two-sided *P* values are reported.

Chromatin conformation capture. Chromosome conformation capture (3C) coupled with qPCR was performed according to a published protocol[58]. Briefly, 7.5 million MCF-7 cells were cross-linked using formaldehyde treatment (1%, 10 min at room temperature), followed by HindIII-HF digestion (400 units, overnight at 37 °C) and ligation (T4 DNA ligase 4,000 units, 4 h at 16 °C). A phenol–chloroform extraction was performed on the DNA fragments, followed by ethanol precipitation. The ligated fragments were quantified by qPCR. Genomic DNA amplicons of 60 primer pairs spanning the *ESR1* SRE region (**Supplementary Table 9**) were mixed in equimolar amounts, digested and ligated to generate a randomly ligated control template. This was used to verify primer efficiency and to normalize the 3C interaction frequency. To normalize our 3C data analysis, we generated the Ct value standard curve using the control template for each tested ligation. Then we quantified the ligation products between the *ESR1* promoter and each of the tested 3C sites using the parameters from each corresponding standard curve (*b* = intercept; *a*, slope): value = 10(Ct − *b*)/*a* (ref. 58).

doi:10.1038/ng.3650

CRISPR–Cas9 enhancer deletion in stable Cas9-expressing T-47D cells. Lentivirus carrying the human codon–optimized *Cas9* gene was generated using 293FT cells transfected with Lipofectamine 3000, the viral plasmids VSVG and PAX2, and a lentiviral *Cas9* vector (Addgene plasmid #52962). T-47D cells were infected with the *Cas9* lentivirus and selected for 14 d with blasticidin. Cas9 protein expression was verified by western blotting using an antibody from Diagenode (#C15200203) (**Supplementary Fig. 17a**).

Stable Cas9-expressing T-47D cells were transfected with a pool of four enhancer targeting CRISPR RNA–transactivating crRNA (crRNA–tracrRNA) complexes (Integrated DNA Technologies) or two nontargeting crRNA–tracrRNA complexes. Briefly, 125 pmol of each of four different enhancer targeting crRNA–tracrRNA complexes (or 250 pmol of two nontargeting complexes) were combined with 6 µL Lipofectamine RNAiMAX and 200 µL OptiMEM, incubated for 20 min at room temperature, and added to each well before plating cells. 150,000 cells were then plated per well for each crRNA–tracrRNA pool targeting different enhancers or nontargeting controls. Transfected cells were incubated for 72 h followed by extraction of RNA and genomic DNA using the Qiagen AllPrep DNA/RNA mini kit according the manufacturer's recommended protocol. RNA was reverse transcribed into cDNA using the Bioline SensiFAST cDNA synthesis kit and RT-qPCR performed using the Bioline SensiFAST SYBR mix (No-Rox) on a Bio-Rad CFX96 Real-Time PCR instrument. Threshold qPCR values for each gene were first normalized to Actin mRNA and were then normalized to the nontargeting crRNA–tracrRNA treated cell values for each of three independent replicates. The fold changes were log transformed, and significance was tested using a one-sample *t*-test. Two-sided *P* values are reported. Deletions were verified through gel electrophoresis (**Supplementary Fig. 17b**).

Data availability. ChIP-seq files for ENCODE, http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeHaibTfbs/ and http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeSydhTfbs/; uniformly processed DNase I hypersensitivity sequencing signal files for 79 cell lines, http://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011/byDataType/signal/jan2011/bigwig/; MCF-7 DHSs identified by the Hotspots algorithm, http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeUwDnase/wgEncodeUwDnaseMcf7HotspotsRep1.broadPeak.gz; MCF-7 Pol II ChIA-PET data created by the ENCODE Project, http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeGisChiaPet/; the common SNV database, ftp://hgdownload.soe.ucsc.edu/goldenPath/hg19/database/snp142Common.txt.gz.

37. Korn, J.M. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat. Genet.* **40**, 1253–1260 (2008).
38. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
39. Purcell, S. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
40. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
41. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
42. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
43. Bolstad, B.M., Irizarry, R.A., Astrand, M. & Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* **19**, 185–193 (2003).
44. Bailey, S.D., Virtanen, C., Haibe-Kains, B. & Lupien, M. ABC: a tool to identify SNVs causing allele-specific transcription factor binding from ChIP-Seq experiments. *Bioinformatics* **31**, 3057–3059 (2015).
45. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
46. John, S. *et al.* Chromatin accessibility pre-determines glucocorticoid receptor binding patterns. *Nat. Genet.* **43**, 264–268 (2011).
47. Lawrence, M.S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
48. Dees, N.D. *et al.* MuSiC: identifying mutational significance in cancer genomes. *Genome Res.* **22**, 1589–1598 (2012).
49. Stergachis, A.B. *et al.* Exonic transcription factor binding directs codon choice and affects protein evolution. *Science* **342**, 1367–1372 (2013).
50. Ju, Y.S. *et al.* Frequent somatic transfer of mitochondrial DNA into the nuclear genome of human cancer cells. *Genome Res.* **25**, 814–824 (2015).
51. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
52. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
53. Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
54. Kulakovskiy, I.V. *et al.* HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Res.* **41**, D195–D202 (2013).
55. Grant, C.E., Bailey, T.L. & Noble, W.S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
56. Livak, K.J. & Schmittgen, T.D. Analysis of relative gene expression data using real-time quantitative PCR and the $2(-\Delta\Delta C_T)$ method. *Methods* **25**, 402–408 (2001).
57. Li, B., Kadura, I., Fu, D.J. & Watson, D.E. Genotyping with TaqMAMA. *Genomics* **83**, 311–320 (2004).
58. Hagège, H. *et al.* Quantitative analysis of chromosome conformation capture assays (3C-qPCR). *Nat. Protoc.* **2**, 1722–1733 (2007).