Title: Integrative genomic analysis predicts causative cis-regulatory mechanisms of the breast cancer-associated genetic variant rs4415084

Authors and affiliations: Yi Zhang[1,2*], Mohith Manjunath[2*], Shilu Zhang[3], Deborah Chasman[3], Sushmita Roy[3,4], and Jun S. Song[2,5]

*Equal contribution

[1]Department of Bioengineering, [2]Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA. [3]Wisconsin Institute for Discovery, [4]Department of Biostatistics and Medical Informatics, University of Wisconsin–Madison, Madison, WI 53792, USA. [5]Department of Physics, University of Illinois at Urbana-Champaign, Urbana, IL 61801, USA.

Running title: Regulatory genomics of GWAS SNPs

Abbreviations: GWAS - genome-wide association study; eQTL - expression quantitative trait loci; ASE - allele-specific expression; TF - transcription factor; LD - linkage disequilibrium; FPKM - fragments per kilobase of transcript per million mapped reads; LCASE - local chromosome allele-specific expression; DHS - DNase I hypersensitive sites; PWM - position weight matrix; TCGA - The Cancer Genome Atlas; ER+ - estrogen receptor positive; TAD - topologically associated domain; MAF - minor allele frequency; RPKM - reads per kilobase of transcript per million mapped reads; SNP - single nucleotide polymorphism; MAPQ - mapping quality; ChIP-seq - chromatin immunoprecipitation sequencing; ASB - allele-specific binding; ncRNA - non-coding RNA; TSS - transcription start site; DNase-seq - DNase I

hypersensitive sites sequencing; RNA-seq - RNA sequencing; Hi-C: high-throughput chromosome conformation capture sequencing.

Corresponding author: Jun S. Song, Carl R. Woese Institute for Genomic Biology, University of Illinois at Urbana-Champaign, 1206 W. Gregory Dr., Urbana, IL 61801. Phone: (217) 244-7750; Fax: (217) 244-2496; Email: songj@illinois.edu

Conflict of interest: The authors declare no potential conflicts of interest.

Abstract

Previous genome-wide association studies (GWAS) have identified several common genetic variants that may significantly modulate cancer susceptibility. However, the precise molecular mechanisms behind these associations remain largely unknown; it is often not clear whether discovered variants are themselves functional or merely genetically linked to other functional variants. Here we provide an integrated method for identifying functional regulatory variants associated with cancer and their target genes by combining analyses of expression quantitative trait loci (eQTL), a modified version of allele-specific expression (ASE) that systematically utilizes haplotype information, transcription factor (TF) binding preference, and epigenetic information. Application of our method to a breast cancer susceptibility region in 5p12 demonstrates that the risk allele rs4415084-T correlates with higher expression levels of the protein-coding gene mitochondrial ribosomal protein S30 (MRPS30) and lncRNA RP11-53O19.1. We propose an intergenic SNP rs4321755, in linkage disequilibrium (LD) with the GWAS SNP rs4415084 (r^2=0.988), to be the predicted functional SNP. The risk allele rs4321755-T, in phase with the GWAS rs4415084-T, created a GATA3 binding motif within an enhancer, resulting in differential GATA3 binding and chromatin accessibility, thereby promoting transcription of MRPS30 and RP11-53O19.1. MRPS30 encodes a member of the mitochondrial ribosomal proteins, implicating the role of risk SNP in modulating mitochondrial activities in breast cancer. Our computational framework provides an effective means to integrate GWAS results with high-throughput genomic and epigenomic data and can be extended to facilitate rapid functional characterization of other genetic variants modulating cancer susceptibility.

Major Findings

We developed a computational framework for integrating GWAS results with heterogeneous cancer genomic data and tissue-specific epigenetic data to facilitate the discovery of causative variants functioning through long-distance gene regulation. Applied to a breast cancer susceptibility region in 5p12, our method provides strong support for a putative causative SNP that is predicted to modulate GATA3 binding and regulate the expression of MRPS30 and nearby lncRNAs.

Quick Guide to Equations and Assumptions

Since the majority of GWAS variants lie in non-coding regions of the human genome where a direct link to gene function is not obvious, we searched for (causative SNP, TF, target gene) triplets under the model of gene regulation by enhancers, in which the SNP interferes with the binding affinity of a key transcription factor (TF). With this assumption, we built a regulation model for a breast cancer susceptibility locus harboring three GWAS SNPs in the 5p12 region. To infer candidate target genes, we first performed expression quantitative trait loci (eQTL) analysis by regressing gene expression levels against two co-variates: genotype status at a given GWAS SNP and copy number of the gene. For each pair of $i \in$ {GWAS SNPs in 5p12} and $j \in$ {genes in 5p12 TAD}, the eQTL model can be expressed as:

$$E_j = \alpha_{ij} + \beta_{ij} G_i + \gamma_{ij} CN_j + \epsilon_{ij} ,$$

where $E_j = log_2(FPKM_j + 1)$ is the expression level of gene j, $G_i \in \{0,1,2\}$ the genotype status of SNP i indicating the number of risk alleles, CN_j the copy number of gene j, α_{ij} the intercept, and ϵ_{ij} the error term. Genes with $\overline{FPKM} \geq 1$ ($\overline{FPKM}$: mean expression among tumor samples) and $p_{ij}^G \leq \alpha$ (p_{ij}^G: p-value of β_{ij} ; $\alpha = 0.05$) were called as significant eQTL target genes. Bonferroni correction for multiple

hypothesis testing was further applied using $p_{ij}^{G} \leq \frac{\alpha}{n}$, where n is the total number of genes tested in the TAD ($n = 22$, thus $\frac{\alpha}{n} = 0.0023$).

To identify *cis*-regulated target genes, we tested local chromosome allele-specific expression (LCASE) using exonic SNPs that were properly phased with the GWAS SNP i. For each exonic SNP m, we obtained a subset of K patients who had heterozygous genotypes at both the GWAS SNP i and the exonic SNP m. For each patient k ($k \in \{1, \ldots, K\}$), we identified the risk allele of SNP m ($risk_m$) and also its protective allele (pro_m) by phasing them to the risk or protective allele of the GWAS SNP. Allelic coverage at the exonic SNP m was determined to obtain $n(risk_m)$ and $n(pro_m)$, the number of reads containing the risk or protective allele, respectively. Depending on the sample size K, two statistics were used to test for transcription imbalance between the two chromosome copies:

$$\begin{cases} n(risk_m) \sim binomial(n(risk_m) + n(pro_m), p_0 = 0.5), & K < 5 \\ n(pro_m) - n(risk_m) \sim Wilcoxon\ signed - rank, & K \geq 5 \end{cases}$$

To discover causative SNPs, we reasoned that functional SNPs should localize within open chromatin regions in enhancers and affect the binding affinity of a TF by changing its recognition motif. We thus obtained a list of candidate SNPs by overlapping LD SNPs of GWAS variants ($r^2 \geq 0.8$) with DNase I hypersensitive sites (DHS) measured in breast cancer cell lines. For each candidate SNP, we scanned the two sequences harboring different alleles of the SNP with a set of position weight matrices (PWM). Suppose that a sequence $\boldsymbol{x}$ of length L matched a PWM, where $\boldsymbol{x} = (x_1, x_2, \ldots, x_k, \ldots, x_{L-1}, x_L)$ harbored one allele of the candidate SNP at position k. Let $\boldsymbol{x^a} = (x_1, x_2, \ldots, x_k^a, \ldots, x_{L-1}, x_L)$ denote the sequence harboring the other allele

x_k^a, and let $p_{l,x}$ denote the probability of nucleotide x at position l in the PWM. To quantify the effect of allele change ($x_k \rightarrow x_k^a$) on the motif, the difference in motif scores was calculated as $D(\boldsymbol{x}, \boldsymbol{x^a}) = score(\boldsymbol{x}) - score(\boldsymbol{x^a})$, where $score(\boldsymbol{x})$ was defined as $\sum_{l=1}^{L} \log p_{l,x_l}$. We devised two approaches to measure the significance of $D(\boldsymbol{x}, \boldsymbol{x^a})$. In the first approach of simulating neutral mutations, we constructed an empirical null distribution of $D(\boldsymbol{x}, \boldsymbol{x^a})$ by introducing random single-nucleotide mutations n times ($n = 5000$), and calculated the empirical p-value. In the second approach of analysing differential k-mer enrichment (with $k = L$), we tested for a difference in the occurrence frequencies of $\boldsymbol{x}$ and $\boldsymbol{x^a}$ between TF ChIP-seq peaks and control regions sampled from breast cancer cell line DHS (Chi-squared test).

Introduction

Genome-wide association studies (GWAS) have identified thousands of common genetic variants associated with various traits and diseases, including cancer (1). However, most of these variants lie in non-coding regions of the genome where a direct link to gene function or regulation is difficult to assess (2). Furthermore, it is often unclear whether the true molecular perturbations associated with carcinogenesis, cancer progression, or therapeutic response indeed lie in the reported GWAS variants themselves or some other linked genetic variants. As a result, discovering the direct functional consequences of genetic variation at GWAS loci has been a critical missing step in utilizing the rich GWAS results to advance cancer research. We here address this important challenge by presenting an integrative computational framework that can facilitate the rapid identification of candidate functional regulatory variants in open chromatin regions and their target genes.

Some approaches are currently known for investigating candidate target genes and causative SNPs of GWAS variants, but they usually yield many false positives (3,4). For identifying candidate target genes, two popular approaches correlate gene transcription level with the variants across a population of patients: one is expression quantitative trait loci (eQTL) analysis and the second is allele-specific expression (ASE) analysis. Both methods make use of genotype and messenger RNA (mRNA) transcription profiles in large patient cohorts available from various databases such as The Cancer Genome Atlas (TCGA). The first method has been successful for identifying genes that are correlated in overall mRNA level with the GWAS variant genotypes. For example, Li *et al.* (5) performed eQTL analysis on estrogen receptor positive (ER+) breast cancer data and found SNPs correlating with the expression of

essential genes such as ESR1 and c-MYC. However, traditional eQTL analysis is only correlative and cannot distinguish between direct and secondary target genes. The ASE method uses patients who are heterozygous at a given GWAS SNP and tests for direct *cis*-regulated genes showing an imbalance in transcription activity between the two separate chromosome copies harboring different alleles of the variant. Such an imbalance is typically assessed from RNA-seq data by counting the allele-specific coverage of exonic SNPs present in candidate genes within a certain distance from the GWAS SNP. Most studies to date, however, use unphased ASE analysis; that is, the exonic SNPs showing an allelic skew are not phased with the GWAS variant (5–7), thereby losing key information about whether the GWAS variant attenuates or promotes target gene transcription. In limited studies utilizing phased ASE (8), phasing accuracy has not been rigorously evaluated; as a result, testing for the consistency of phased ASE across patients may lose power when incorrectly phased haplotypes are used for some patients.

Our integrative approach improves upon these ideas in several ways. First, we restrict eQTL analysis only to those genes that lie within the same topologically associated domain (TAD) as a given GWAS SNP. Since TADs are thought to represent physical chromatin loops containing regulatory elements and their target genes (9), restricting the analysis to GWAS locus-containing TADs should help remove false positive genes and improve upon the current practice of choosing an arbitrary distance cutoff (5). We then apply ASE analysis on exonic SNPs properly phased with GWAS SNPs by removing patient samples whose reconstructed haplotypes show unstable phasing in our simulation study. This filtering method, which we termed the local-chromosome allele-specific expression (LCASE), effectively analyzes ASE on a local

chromosome segment that is in LD with a GWAS SNP and increases statistical power by removing incorrect haplotypes from consideration. To find the causal regulatory SNP linked to a given GWAS SNP, we search for linked SNPs residing in functional open chromatin regions and computationally assess the SNPs' effect on candidate transcription factor (TF) binding motifs, as we have previously done to show that the two well-known point mutations in the *TERT* promoter create new binding sites of GABP to reactivate *TERT* transcription (10).

We demonstrate the utility of our method by applying it to a breast cancer susceptibility region in 5p12, which is a GWAS hotspot harboring three non-coding GWAS SNPs replicated in previous studies (11–13). First, we show that all three GWAS SNPs may be targeting the same genes, the protein-coding gene MRPS30 and lncRNA RP11-53O19.1, both of which have been implicated in cancer (12,14,15). MRPS30 encodes a member of the mitochondrial ribosomal large subunits (14), suggesting the risk SNP's role in modulating mitochondrial activities. The lncRNA RP11-53O19.1, also known as breast cancer-associated transcript 54 (BRCAT54), is overexpressed in luminal A breast cancer subtype (ER+) (15), suggesting its specific role in ER+ breast cancers. We then propose that an intergenic SNP, in LD with one of the 5p12 GWAS SNPs, is the predicted functional SNP. We provide multiple lines of evidence supporting that the risk allele of the predicted functional SNP increases the binding affinity of GATA3, an important TF known to cooperate with ESR1 and FOXA1 in ER+ breast cancers (16). Although this paper focuses on the ER+ breast cancer risk hotspot in 5p12, the described method for dissecting the functional consequences of GWAS variants can be generalized to other noncoding loci implicated in a variety of cancers or traits.

Materials and Methods

TCGA breast cancer data

Germline and tumor genotypes at tag SNPs from 979 patients were downloaded from the TCGA Data Portal and subsequently used for imputation. The following data were obtained from the NCI Genomic Data Commons (GDC) Legacy Archive (17): 693 patients' tumor copy number segmentation data, 788 patients' processed tumor gene expression data, and 795 ER+ breast cancer patients' tumor RNA-seq raw reads. Data from normal tissue and primary tumor were matched based on patient barcodes. Patients having all four types of data available were selected to obtain a final set of 679 female ER+ breast cancer patients.

Genotype imputation

TCGA germline genotypes of tag SNPs measured by Affymetrix human SNP array 6.0 were mapped to 1000 Genomes Project phase 3 (18) variants. For each patient, genotypes with low quality (5^{th} quantile in confidence score) were excluded. Since none of the three GWAS SNPs (rs4415084, rs10941679, rs7116600) were directly genotyped by the SNP array, imputation was performed in the 5p12 region using IMPUTE2 (19,20) and the 1000 Genomes Project Phase 3 (October 2014 version) (18) as a reference panel. Imputed genotypes were retained if the maximum genotype probability exceeded the threshold 0.9 and the minor allele frequency (MAF) exceeded 0.01 (for high imputation quality (20)). To obtain a more accurate picture of germline genotypes, the genotypes from normal blood rather than primary tumor were used in all analysis to avoid miscalls from potential genotyping errors and somatic mutations.

Chromosome interaction data and TAD

Hi-C data were obtained from GSE66733 (21) and GSE51687 (22) for MCF-7; and from GSE53463 (23) and the Encyclopedia of DNA Elements (ENCODE) (24) ENCSR549MGQ for T-47D. MCF-7 TAD structures were determined based on insulation score from Hi-C data (GSE66733). The TAD information in T-47D was obtained from the ENCODE dataset ENCFF075QYD. We used the MCF-7 TAD information in the analysis, as the two data sets were similar. MCF-7 ChIA-PET interactions mediated by CTCF, ESR1, and POLR2A were obtained from ENCODE (24).

TCGA breast cancer eQTL analysis

We performed TCGA-based eQTL analysis by constructing a multivariate linear model that regressed the expression level of each gene against the copy number of the gene and the genotype status at a given GWAS SNP locus. For the ith GWAS SNP in 5p12, we encoded its germline genotype G_i as the number of risk alleles ($G_i \in \{0,1,2\}$). To determine the copy number of the jth gene residing within the 5p12 TAD region, we took length-weighted average ($\overline{s_j}$) of tumor copy number segment data covering the gene and transformed it into copy number $CN_j = 2 * 2^{\overline{s_j}}$. The RNA-seq RPKM value of gene j was log-transformed as $E_j = log_2(RPKM_j + 1)$. Multivariate linear regression (25), $E_j = \alpha_{ij} + \beta_{ij} G_i + \gamma_{ij} CN_j + \epsilon_{ij}$, was then performed for each pair of $i \in \{GWAS\ SNPs\ in\ 5p12\}$ and $j \in \{genes\ in\ 5p12\ TAD\}$, where α_{ij} is the intercept and ϵ_{ij} the error term. Genes with $\overline{RPKM} \geq 1$ ($\overline{RPKM}$: mean expression among tumor samples) and $p_{ij}^G \leq \alpha$ (p_{ij}^G: p-value of β_{ij} ; $\alpha = 0.05$) were called as significant eQTL target genes. Bonferroni

correction for multiple hypothesis testing was further applied using $p_{ij}^{G} \leq \frac{\alpha}{n}$, where n is the total number of genes tested in the TAD ($n = 22$, thus $\frac{\alpha}{n} = 0.0023$).

LCASE analysis

We tested for within-patient differential transcription between the two chromosome copies harboring different alleles of a given GWAS SNP using a modified version of ASE analysis which we termed LCASE. To distinguish between the transcripts from different chromosome copies, we utilized exonic SNPs (GENCODE v19 (26)) that were heterozygous based on imputation results. For each exonic SNP, we selected a subset of patients with heterozygous genotypes at both the GWAS SNP and the exonic SNP. For each patient in this subset, we used SHAPEIT2 (27) to determine the phase between the GWAS SNP and exonic SNP. We denoted the risk and protective alleles at the GWAS SNP as *risk* and *pro*, respectively, and denoted the reference and alternative alleles of the exonic SNP under consideration as *0* and *1*, respectively. To remove uncertain phasing, we further sampled haplotypes *n* times (n=100) from the haplotype model generated by SHAPEIT2 and computed the occurrence probability $(p_{risk|0}, p_{risk|1})$ of the two phases, thereby removing patient samples with no dominant phase as indicated by the condition $\max(p_{risk|0}, p_{risk|1}) < 0.9$ (**Supplementary Fig. 1**). We then counted the allelic-specific coverage of each exonic SNP from high quality RNA-seq reads ($MAPQ \geq 20$) and tested for transcription imbalance using two statistical tests depending on the sample size K: binomial test (when $K < 5$) and Wilcoxon signed-rank test (when $K \geq 5$).

Functional SNPs prioritization

A list of GWAS variants in the 5p12 region were obtained from the NHGRI GWAS Catalog (28), and the risk alleles were extracted from Stacey *et al.* (11) and Thomas *et al.* (13). All common ($MAF \geq 0.01$) SNPs from 1000 Genomes Project Phase 3 in high LD ($r^2 \geq 0.8$) with any of the 5p12 GWAS SNPs were selected for prioritization. All r^2 values were calculated based on the 1000 Genomes Project Phase 3 EUR population. For SNP prioritization, DNase I hypersensitivity sites (DHS) in MCF-7 and T-47D were collected from the ENCODE (24) database, including DHS under estradiol treatment (**Supplementary Table 1**). Only those LD SNPs that overlapped with a narrow peak from at least one of DHS replicate experiments were considered for further investigation. We also used histone modification and TF binding profiles to facilitate functional SNP prioritization. Histone modification data were obtained from ENCODE and the Gene Expression Omnibus (GEO) database (GSE26831, GSE63109 and GSE69112). ChIP-seq data in MCF-7 and T-47D for TFs – including ESR1, PGR, GATA3, and FOXA1 known to be important in breast cancer – were obtained from ENCODE and GEO (**Supplementary Table 2**). Raw ChIP-seq data from GEO were mapped using BWA (29) (default parameters), and peaks were called using MACS2 (30) (TF ChIP-seq: default parameters; Histone modification ChIP-seq: broadpeak mode, 0.1 FDR).

Motif analysis

We collected TF position weight matrices (PWM) from HOCOMOCO Human v10 (31), FACTORBOOK (32,33), TRANSFAC (34), JASPAR vertebrates (35) and Jolma2013 (36) (**Supplementary Table 3**). To identify motifs potentially affected by a SNP, we used the program FIMO (37) (version 4.12.0) to scan the two 50bp sequences centered at the different alleles of each candidate SNP (FIMO threshold

10^{-3}). We then sought for candidate SNP-TF pairs for which the sequence harboring one allele of the SNP matched the TF motif, whereas the sequence harboring the other allele did not. For each SNP-TF pair, the statistical significance of creating or disrupting the motif via the SNP was evaluated with two approaches: mutation simulations and, when available, ChIP-seq k-mer enrichment analysis. The first approach compared the change in motif score caused by the SNP against the changes caused by random single-nucleotide mutations simulated 5000 times, where the motif score was defined as $\sum_{l=1}^{L} \log p_{l,n_l}$ (p_{l,n_l}: the probability of nucleotide n_l at position l in the PWM of length L; **Supplementary Methods**). The second approach tested for the differential enrichment of k-mers (e.g. $k = L$) harboring different alleles by comparing their occurrence frequency in TF ChIP-seq peaks against that in control regions (random open chromatin regions in MCF-7 or T-47D; Chi-squared test; **Supplementary Methods**) (38).

TF-target correlation analysis

The candidate SNP-TF pairs from motif analysis were further filtered based on the correlation structure between each candidate TF and its eQTL target gene. A list of human TFs was obtained from AnimalTFDB (39); and, only the expressed TFs ($\overline{RPKM} \geq 1$ among ER+ breast cancer patients) overlapping with the candidates from motif analysis were further considered and divided into three groups: candidate activators (*Pearson* correlation coefficient $r \geq r_{+2\sigma}$), candidate repressors ($r \leq r_{-2\sigma}$), and uncorrelated TFs ($r_{-2\sigma} < r < r_{+2\sigma}$), where the thresholds $r_{-2\sigma}$ and $r_{+2\sigma}$ were chosen based on the distribution of correlation coefficients between random pairs of TF and non-TF genes ($r_{-2\sigma} = -0.337$; $r_{+2\sigma} = 0.409$; **Supplementary Methods**). We then narrowed the list of candidate TFs based on four models shown in

Supplementary Fig. 2. In this paper, we focused on an example of enhancer-activating risk allele. In this case where the target gene expression was positively correlated with the number of GWAS risk allele, we selected TFs passing the following filters as candidate activators: 1) the risk allele of the candidate enhancer SNP significantly increased the motif score of the TF; 2) the TF expression level was positively correlated ($r \geq r_{+2\sigma}$) with the target gene expression level among patients having two copies of the risk allele; 3) the TF-target gene expression correlation in patients carrying two copies of the risk allele (two copies of the motif) was stronger than that in patients carrying no risk allele (no motif).

TF allele-specific binding (ASB)

For TFs with ChIP-seq data available in MCF-7 and T-47D, we tested for ASB by searching for a skew in the read coverage of candidate heterozygous SNPs located within peaks. The significance of a skew was measured by using the binomial test assuming no bias. To determine the genotype status, tag SNP data for MCF-7 and T-47D were obtained from CCLE project (40), and imputation was performed using the same pipeline described above. To check that the candidate loci have no deletion or amplification events, which might complicate the ASB analysis, copy number data for the cell lines were obtained from CCLE (40) and ENCODE (24) (GSE40698). ChIP-seq reads from different replicates were combined and deduplicated.

Results

Integrative analysis framework for GWAS functional characterization

We integrated multiple genomic analyses to identify systematically the triplets of functional SNP, corresponding TF regulator, and target gene (**Fig. 1a, b**). Our main

hypothesis was that non-coding SNPs modulating cancer risk would likely lie inside putative regulatory elements to promote or attenuate the binding affinity of relevant TFs which would be reflected at the mRNA level of the target gene. A schematic representation of our framework is shown in **Fig. 1a**. Target genes were identified by using within-TAD eQTL and LCASE analyses of GWAS SNPs, and functional SNP-TF candidates were determined by integrating motif analysis, expression correlation analysis, and epigenetic annotation of regulatory regions (**Methods**) for GWAS SNPs and their high LD SNPs.

5p12 GWAS risk alleles correlate with elevated MRPS30/RP11-53O19.1 expression

We applied our computational framework to a breast cancer susceptibility region in 5p12. As there are three ER+ breast cancer GWAS SNPs in the 5p12 risk hotspot (11,13) (**Fig. 2**), we first performed eQTL analysis for individual SNPs to investigate whether they all targeted the same genes. We restricted the analysis to the TAD containing all three GWAS SNPs based on the MCF-7 TAD information (21). The 5p12 TAD region (chr5:43,480,001-45,000,000, hg19) containing all three GWAS SNPs encompassed 7 protein-coding genes and 15 non-coding RNAs (GENCODE v19). Among them, the protein-coding gene MRPS30 and lncRNAs RP11-53O19.1 and RP11-53O19.3 showed significant eQTL correlation (*Genotype* effect $p \leq 0.05$) with all three GWAS SNPs, but RP11-53O19.3 had only borderline significance with rs7716600 (**Supplementary Table 4, 5**). Notably, the GWAS SNP rs4415084 correlated strongest with these genes compared to other GWAS SNPs (MRPS30, $p = 1.39 \times 10^{-5}$; RP11-53O19.1, $p = 6.19 \times 10^{-6}$; RP11-53O10.3, $p = 1.23 \times 10^{-4}$; eQTL analysis in **Methods; Fig. 3a**). Furthermore, rs4415084 was the only

GWAS SNP that remained significant after Bonferroni correction (**Supplementary Table 4**).

It should be noted that the three genes are likely co-expressed, as RP11-53O19.1 and MRPS30 share a divergent promoter and RP11-53O19.3 is contained in several elongated transcripts of MRPS30 (**Supplementary Fig. 3**). As RP11-53O19.3 might be a processed transcript of MRPS30, our subsequent analyses focused only on MRPS30 and RP11-53O19.1. The positive correlation between the number of risk alleles and the expression level of these target genes was also confirmed in both isoform-level eQTL analysis in TCGA data (**Supplementary Table 6**) and gene-level eQTL analysis in GTEx normal breast tissue (MRPS30, $p = 5.46 \times 10^{-3}$; RP11-53O19.1, $p = 4.14 \times 10^{-4}$; **Supplementary Fig. 4**). Together, these results showed that all three GWAS SNPs in the 5p12 region may target the same set of genes within the TAD, with rs4415084 displaying the strongest effect.

LCASE detects higher MRPS30/RP11-53O19.1 expression on the risk chromosome

Since eQTL analysis is only correlative and cannot distinguish between *trans*-regulation and *cis*-regulation, we sought evidence for *cis*-regulation by utilizing heterozygous exonic/UTR SNPs properly phased with the GWAS SNPs (**Fig. 3b; Methods**). We first examined all common exonic variants ($MAF \geq 0.01$, 1000 Genomes Phase 3, EUR) in the protein-coding gene MRPS30 and found three SNPs suitable for LCASE analysis (rs61754779 and rs34522103 in exons; rs79210252 in 3'UTR; **Fig. 3c**). There were five patients with heterozygous genotypes at both the GWAS SNP rs4415084 (C/T) and the SNP rs61754779 (G/C) located in the first exon. Among these five, two were removed because of uncertain phasing. The risk

haplotypes in these three patients were all rs61754779-G|rs4415084-T; equivalently, all of the protective haplotypes were rs61754779-C|rs4415084-C. RNA-seq read coverage at this exonic SNP displayed a significantly higher transcription activity from the risk haplotype in two patients ($p = 1.33 \times 10^{-4}$, $p = 9.72 \times 10^{-17}$; one-sided binomial test; **Fig. 3c**), but the skew in the third patient was not significant ($p = 0.64$). Similar analyses at the exonic SNPs rs34522103 and rs79210252 identified a different set of three patient samples that showed the same direction of transcription imbalance towards the risk haplotype, and the skew was statistically significant in two of these patients (rs34522103: $p = 6.71 \times 10^{-47}$; rs79210252: $p = 1.22 \times 10^{-3}$, $p = 0.28$; one-sided binomial test; **Fig. 3c**).

Although the LCASE in MRPS30 was significant in several patients, the sample size was small, because exonic SNPs were generally rare in the genome. We thus sought for more evidence in MRPS30 ASE by analyzing additional SNPs in the non-coding RNA (ncRNA) located at the 3' end of MRPS30 (GENCODE track, **Supplementary Fig. 3**). *De novo* assembly of transcripts using StringTie (41) on tumor RNA-seq reads revealed multiple MRPS30 elongated transcripts covering both MRPS30 3' ncRNA and RP11-53O19.3 (**Supplementary Fig. 3**), suggesting that these two ncRNAs could arise from abnormal elongation of the protein-coding MRPS30 gene. This elongation hypothesis was also supported by high correlation in expression level between MRPS30 and RP11-53O19.3 ($r = 0.95$).

We therefore tested whether there were additional LCASE SNPs falling in this elongated region and found four transcribed SNPs that could be phased accurately in a large number of patient samples ($n = 23, n = 248, n = 238, n = 240$, respectively).

Consistent with the analysis of exonic SNPs, all of these SNPs in the elongated region displayed significantly higher transcription activity on the risk chromosome ($p = 9.5 \times 10^{-7}, p = 5.5 \times 10^{-13}, p = 7.7 \times 10^{-17}, p = 3.2 \times 10^{-14}$, respectively; Wilcoxon signed-rank test; **Fig. 3d; Supplementary Fig. 5**). Similar analyses in RP11-53O19.3 found that five out of eight transcribed SNPs showed a skew towards the risk chromosome, while two other SNPs close to the 3' end displayed an opposite trend (**Supplementary Fig. 6**). Together, the eQTL and LCASE analyses thus demonstrated that the breast cancer risk allele rs4415084-T was associated with increased MRPS30 and RP11-43O19.1/3 expression among ER+ breast cancer patients in a *cis*-regulating manner (**Fig. 3c, d; Supplementary Fig. 5, 6, 7**).

Prioritization of candidate SNPs

We next prioritized candidate LD SNPs using epigenetic information. We gathered 123 variants in high LD ($r^2 \geq 0.8$, 1000 Genomes Phase 3, EUR) with any of the three GWAS SNPs rs4415084, rs10941679 or rs7716600, including the GWAS SNPs themselves (**Fig. 2**). Using the DNase I hypersensitivity (DHS) data from ENCODE in two breast cancer cell lines (MCF-7 and T-47D) (24,42) (**Fig. 2**), we found that sixteen of these SNPs resided in DHS peaks, all of them being in high LD with rs4415084. Among the sixteen SNPs, three SNPs were located near the center of DHS peaks (**Supplementary Fig. 8**) and were thus prioritized for further investigation in motif analysis and TF correlation analysis.

rs4321755 is a candidate functional SNP creating a GATA3 binding motif

Analyzing the prioritized candidate SNPs provided multiple lines of evidence for rs4321755 (chr5:44646195; about 163 kb upstream of MRPS30 transcription start site

(TSS)) being the functional SNP. The risk allele of SNP rs4321755 could be determined to be T, because this SNP was in high LD with GWAS rs4415084 ($r^2 = 0.988$, 1000 Genomes Phase 3, EUR), and rs4321755-T was the allele in phase with the GWAS risk allele rs4415084-T. First, our motif analysis showed that the risk allele rs4321755-T created a GATA3 binding motif; the 9bp sequence covering rs4321755 matched a GATA3 motif only when it harbored the allele rs4321755-T, and not when it contained the alternative allele rs4321755-C (**Fig. 4a**, FIMO $p = 1.59 \times 10^{-4}$). We further confirmed that the T-to-C conversion at the SNP resulted in a dramatic frequency drop from 93.7% to 0.0% in the GATA3 PWM, which was a significant change based on both mutation simulations ($p = 0.003$; **Supplementary Methods**), and ChIP-seq k-mer enrichment analysis ($p = 0.025$, Chi-squared test; **Supplementary Methods**).

Second, GATA3 showed a consistent pattern in the TF-target gene correlation analysis: among the TCGA patients, GATA3 expression positively correlated with the expression levels of predicted target genes MRPS30, RP11-53O19.1 and RP11-53O19.3, supporting GATA3's role as a transcription activator (**Fig. 4b**). Furthermore, the positive correlation was strongest (and above $r_{+2\sigma}$ for MRPS30 and RP11-53O19.3) in patients carrying the rs4321755-T/T genotype, moderate in patients with the rs4321755-C/T genotype, and weakest in patients with the rs4321755-C/C genotype (**Fig. 4b**), consistent with our model of enhancer-activating risk allele (**Supplementary Fig. 2**). This correlation trend implied that the regulation mediated by GATA3 was strongest when two copies of GATA3 motif existed at the SNP on both paternal and maternal chromosomes, while the regulation was weakest when both GATA3 binding sites at the SNP were disrupted.

We next sought direct experimental evidence of GATA3 binding at the predicted causal SNP. Consistent with our prediction, we found that rs4321755 was located at the center of a peak in the T-47D GATA3 ChIP-seq data from ENCODE (q-value $= 10^{-4}$, 2bp from the summit; **Fig. 4c**). Moreover, FOXA1 and PGR ChIP-seq in T-47D also showed co-binding with GATA3 near rs4321755, supporting that this SNP indeed lies within an active regulatory region in breast cancer, given that FOXA1 is a pioneer factor in ER+ breast cancers (43) and PGR can interact with ER (44) (**Fig. 4c**); ESR1 ChIP-seq showed no direct binding of ER itself near rs4321755 (24) (GSE32465). Furthermore, this regulatory region was evolutionarily conserved (100 vertebrates base-wise conservation by PhyloP (45), **Fig. 4d)**, indicating its important role retained through evolution. To measure GATA3 ASB, we imputed the genotypes and found T-47D to be heterozygous at rs4321755. The copy number of the segment containing the SNP suggested that T-47D had no deletion or amplification at this locus, informing that the null hypothesis should be unbiased binding between the two alleles (**Methods**). We obtained and deduplicated DNase-seq and multiple TF ChIP-seq reads for ASB analysis in T-47D. As shown in **Fig. 4e**, GATA3 ChIP-seq reads contained more of the risk allele rs4321755-T than the protective allele rs4321755-C ($p = 0.019$, one-sided binomial test). DNase-seq and PGR ChIP-seq also showed significantly more reads with T than C (DNase-seq, $p = 0.032$; PGR, $p = 0.002$; one-sided binomial test); FOXA1 had a similar imbalance, although the p-value was not significant due to low coverage at the SNP (**Fig. 4c, e**). Quite interestingly, MCF-7, which was homozygous for the protective allele C at rs4321755, had a closed chromatin configuration at the SNP (**Supplementary Fig. 8**). Thus, the SNP likely contributed to the lack of GATA3 binding in MCF-7, potentially jointly with

GATA3's impaired DNA-binding ability caused by a heterozygous frameshift mutation in MCF-7 (46). Together, these findings provided experimental evidence that the putative causal SNP rs4321755 might not only influence GATA3 ASB, but also modulate differential chromosomal accessibility and, thus, enhancer activity.

We next explored available chromatin interaction data to assess potential long-distance enhancer-promoter interactions, since the candidate SNP rs4321755 was about 163 kb away from the MRPS30 TSS. However, rs4321755 was in DHS only in the T-47D cell line, for which no GATA3 or other ChIA-PET data were available. In MCF-7, data were available for Hi-C (21) and ChIA-PET of breast cancer-related factors, such as ESR1 and CTCF (47). However, no significant chromatin interactions originating from the regulatory region were observed in MCF-7, probably because this enhancer was specific to T-47D and was not accessible in MCF-7 which carried the C/C genotype at rs4321755 (**Supplementary Fig. 8**). Comparing the Hi-C data in MCF-7 versus T-47D at 40 kb resolution using edgeR (48), we found a significantly higher contact count in the heterozygous T-47D than the homozygous MCF-7 between the bin containing rs4321755 and the bins spanning MRPS30 (edgeR $FDR < 1 \times 10^{-2}$) and RP11-53O19.1 (edgeR $FDR < 1 \times 10^{-4}$; **Supplementary Methods** and **Supplementary Table 7**).

Discussion

This study presented a computational framework for systematically investigating the functional consequences of GWAS SNPs. We applied our method to a breast cancer susceptibility region in 5p12, and discovered a causative SNP with multiple lines of evidence supporting its function in modulating GATA3 binding affinity. This

causative SNP potentially explained the molecular mechanism of one 5p12 GWAS SNP, rs4415084, while the other two GWAS SNPs in the 5p12 region had no candidate SNPs found in open chromatin. It is currently unclear whether these three GWAS SNPs function through the same or independent regulatory elements. For instance, the risk allele G of rs10941679 was completely contained in the background of the risk allele rs4415084-T that had higher allele frequency and thus broader impact in population (233 out of 234 patients with rs10941679-G carry rs4415084-T; rs10941679 MAF: 0.23; rs4415084 MAF: 0.41; 1000 Genomes Project Phase 3, EUR (18)). Moreover, Stacey *et al.* (11) showed that the risk in 5p12 could be explained by either rs4415084 or rs10941679, with the significance of rs4415084-T remaining after correcting for rs10941679-G and vice versa. Although a previous study found that the GWAS SNP rs10941679 itself could be causal (12), this particular SNP was not in open chromatin regions of the breast cancer cell lines that we examined (**Supplementary Fig. 8**). Another study reported correlation between our target gene MRPS30 and rs7716600 genotype, but no candidate causative SNPs in TF binding sites were discussed (49). Although the breast cancer susceptibility harbored in 5p12 is not totally understood, we here propose one functional SNP which may directly link rs4415084 to the regulation of predicted target genes.

Our integrative approach facilitated the identification of putative target genes by combining eQTL and LCASE analysis. We found breast cancer-specific TADs containing GWAS SNPs and performed intra-TAD eQTL analysis. As TADs are thought to provide physical subdivisions of *cis*-regulation, this approach should help reduce the number of false positives that currently challenge the traditional eQTL analysis. In our LCASE analysis, we controlled for phasing quality by removing

samples that could not be phased confidently upon simulation, enabling a direct measurement of differential transcription activity from two local chromosome copies harboring different GWAS alleles. The target genes identified by our method were already implicated in cancer: the protein-coding gene MRPS30, closely related to mitochondrial activity, was also linked to the 5p12 risk variants (rs10941679 and rs7716600) in earlier studies (12,49). Similarly, the candidate target lncRNA RP11-53O19.1 was previously found to have significantly higher expression in luminal A (ER+) breast cancers compared to other subtypes (12). Even though the expression levels of these target genes may be directly modulated by the identified SNP, cancer is a complex multigenic disease, and understanding how the altered expression levels contribute to cancer predisposition requires further investigation, likely involving a systems-level approach.

Our framework predicted candidate (causative SNP, TF, target gene) triplets to address the challenging problem of discovering functional SNPs that, we hypothesized, might disrupt TF binding activities (**Fig. 5**). We integrated cell type-specific epigenetics profiles, motif analysis, and expression correlation signatures to reduce false positives. Our method actually yielded a ranked list of candidate TFs. Our top candidate GATA3 was validated through GATA3 ChIP-seq data and ASB analysis. Besides GATA3, TCF7L1, NR3C1, and ETS1 also qualified as candidate TFs when the TF-gene correlation thresholds was chosen to be less stringent (**Supplementary Fig. 8**). These predictions would benefit from future ChIP-seq data as they become available and also from imputing binding profiles (50). Currently, a key challenge in annotating functional regulatory elements is the lack of 3D chromatin interaction data that can help validate the physical interaction of predicted

enhancer-promoter pair. For this study, GATA3 ChIA-PET data in T-47D, currently missing, would be most suitable for validating the predicted interaction between the identified enhancer and MRPS30 promoter.

Conclusion and perspectives

We present an integrative computational method using genomic and epigenomic data to identify causative regulatory variants that may directly modulate cancer predisposition. Application of our method to a breast cancer susceptibility region in 5p12 reveals the intergenic SNP rs4321755, in LD with the GWAS SNP rs4415084, as the candidate causative variant. We propose that the risk allele rs4321755-T significantly increases GATA3 binding affinity and therefore results in up-regulation of the predicted target genes MRPS30 and RP11-53O19.1. Our computational framework can be extended to investigate other genetic variants modulating cancer susceptibility, contributing to understanding new pathways in tumorigenesis and developing personalized prevention of cancer.

Acknowledgments

The results appearing here are in part based upon the data generated by the TCGA Research Network (http://cancergenome.nih.gov/, dbGaP accession number phs000424.v6.p1 on 05/06/2016) and the GTEx Project, supported by the Common Fund of the Office of the Director of the National Institutes of Health and by NCI, NHGRI, NHLBI, NIDA, NIMH, and NINDS. We acknowledge the ENCODE consortium that generated the data sets used in the manuscript. Y. Zhang, M. Manjunath, and J.S. Song were supported by the grants 1U54GM114838, awarded by National Institute of General Medical Sciences (NIGMS) through funds provided by

the trans-NIH (National Institutes of Health) Big Data to Knowledge (BD2K) initiative, and R01CA163336. S. Zhang, D. Chasman, and S. Roy were supported by the NIH BD2K grant U54 AI117924. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

1. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014;42.
2. 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.
3. Li Q, Seo J-H, Stranger B, McKenna A, Pe'er I, Laframboise T, et al. Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. Cell. 2013;152:633–41.
4. 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.
5. Li Q, Seo JH, Stranger B, McKenna A, Pe'Er I, Laframboise T, et al. Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. Cell. Elsevier Inc.; 2013;152:633–41.
6. Xiao R, Scott LJ. Detection of cis-acting regulatory SNPs using allelic expression data. Genet Epidemiol. 2011;35:515–25.

7. Sigurdsson MI, Saddic L, Heydarpour M, Chang T-W, Shekar P, Aranki S, et al. Allele-specific expression in the human heart and its application to postoperative atrial fibrillation and myocardial ischemia. Genome Med. 2016;8:127.

8. Conde L, Bracci PM, Richardson R, Montgomery SB, Skibola CF. Integrating GWAS and expression data for functional characterization of disease-associated SNPs: An application to follicular lymphoma. Am J Hum Genet. 2013;92:126–30.

9. Bonev B, Cavalli G. Organization and function of the 3D genome. Nat Rev Genet. 2016;17:772–772.

10. Bell RJA, Rube HT, Kreig A, Mancini A, Fouse SD, Nagarajan RP, et al. The transcription factor GABP selectively binds and activates the mutant TERT promoter in cancer. Science. 2015;348:1036–9.

11. Stacey SN, Manolescu A, Sulem P, Thorlacius S, Gudjonsson SA, Jonsson GF, et al. Common variants on chromosome 5p12 confer susceptibility to estrogen receptor–positive breast cancer. Nat Genet. 2008;40:703–6.

12. 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 Genet. 2016;99:903–11.

13. Thomas G, Jacobs KB, Kraft P, Yeager M, Wacholder S, Cox DG, et al. A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). Nat Genet. 2009;41:579–84.

14. Greber BJ, Boehringer D, Leibundgut M, Bieri P, Leitner A, Schmitz N, et al. The complete structure of the large subunit of the mammalian mitochondrial

ribosome. Nature. 2014;

15. Su X, Malouf GG, Chen Y, Zhang J, Yao H, Valero V, et al. Comprehensive analysis of long non-coding RNAs in human breast cancer clinical subtypes. Oncotarget. 2014;5:9864–76.

16. Theodorou V, Stark R, Menon S, Carroll JS. GATA3 acts upstream of FOXA1 in mediating ESR1 binding by shaping enhancer accessibility. Genome Res. 2013;23:12–22.

17. Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, et al. Toward a Shared Vision for Cancer Genomic Data. N Engl J Med. 2016;375:1109–12.

18. The 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015. page 68–74.

19. Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 2009;5.

20. Howie B, Marchini J, Stephens M. Genotype Imputation with Thousands of Genomes. G3. 2011;1:457–70.

21. Barutcu AR, Lajoie BR, McCord RP, Tye CE, Hong D, Messier TL, et al. Chromatin interaction analysis reveals changes in small chromosome and telomere clustering between epithelial and breast cancer cells. Genome Biol. Genome Biology; 2015;16:214.

22. Mourad R, Hsu P-Y, Juan L, Shen C, Koneru P, Lin H, et al. Estrogen Induces Global Reorganization of Chromatin Structure in Human Breast Cancer Cells. PLoS One. 2014;9:e113354.

23. Le Dily FL, Baù D, Pohl A, Vicent GP, Serra F, Soronellas D, et al. Distinct

structural transitions of chromatin topological domains correlate with coordinated hormone-induced gene regulation. Genes Dev. 2014;28:2151–62.

24. Bernstein BE, Birney E, Dunham I, Green ED, Gunter C, Snyder M. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.

25. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria; 2016.

26. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome Res. 2012;22:1760–74.

27. Delaneau O, Zagury J-F, Marchini J. Improved whole-chromosome phasing for disease and population genetic studies. Nat Methods. 2013;10:5–6.

28. Welter D, MacArthur J, Morales J, Burdett T, Hall P, Junkins H, et al. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. 2014;42:D1001-6.

29. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. 2009;25:1754–60.

30. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based analysis of ChIP-Seq (MACS). Genome Biol. 2008;9:R137.

31. Kulakovskiy I V., Vorontsov IE, Yevshin IS, Soboleva A V., Kasianov AS, Ashoor H, et al. HOCOMOCO: Expansion and enhancement of the collection of transcription factor binding sites models. Nucleic Acids Res. 2016;44:D116–25.

32. Pinello L, Xu J, Orkin SH, Yuan GC. Analysis of chromatin-state plasticity identifies cell-type-specific regulators of H3K27me3 patterns. Proc Natl Acad

Sci U S A. 2014;111:E344-53.

33. Wang J, Zhuang J, Iyer S, Lin X, Whitfield TW, Greven MC, et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. Genome Res. 2012;22:1798–812.

34. Matys V, Kel-Margoulis O V, Fricke E, Liebich I, Land S, Barre-Dirrie A, et al. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. 2006;34:D108-10.

35. Mathelier A, Fornes O, Arenillas DJ, Chen CY, Denay G, Lee J, et al. JASPAR 2016: A major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2016;44:D110–5.

36. Jolma A, Yan J, Whitington T, Toivonen J, Nitta KR, Rastas P, et al. DNA-binding specificities of human transcription factors. Cell. 2013;152:327–39.

37. Grant CE, Bailey TL, Noble WS. FIMO: Scanning for occurrences of a given motif. Bioinformatics. 2011;27:1017–8.

38. Huang D, Ovcharenko I. Identifying causal regulatory SNPs in ChIP-seq enhancers. Nucleic Acids Res. 2015;43:225–36.

39. Zhang HM, Chen H, Liu W, Liu H, Gong J, Wang H, et al. AnimalTFDB: A comprehensive animal transcription factor database. Nucleic Acids Res. 2012;40.

40. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483:603–307.

41. Pertea M, Pertea GM, Antonescu CM, Chang T-C, Mendell JT, Salzberg SL. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015;33:290–5.

42. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. Nature. 2012;489:75–82.

43. Zaret KS, Carroll JS. Pioneer transcription factors: Establishing competence for gene expression. Genes Dev. 2011. page 2227–41.

44. Mohammed H, Russell IA, Stark R, Rueda OM, Hickey TE, Tarulli GA, et al. Progesterone receptor modulates ER[agr] action in breast cancer. Nature. 2015;523:313–7.

45. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res. 2010;20:110–21.

46. Adomas AB, Grimm SA, Malone C, Takaku M, Sims JK, Wade PA. Breast tumor specific mutation in GATA3 affects physiological mechanisms regulating transcription factor turnover. BMC Cancer. 2014;14:278.

47. Chan CS, Song JS. CCCTC-binding factor confines the distal action of estrogen receptor. Cancer Res. 2008;68:9041–9.

48. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139–40.

49. Quigley DA, Fiorito E, Nord S, Van Loo P, Alnæs GG, Fleischer T, et al. The 5p12 breast cancer susceptibility locus affects MRPS30 expression in estrogen-receptor positive tumors. Mol Oncol. Elsevier B.V; 2014;8:273–84.

50. Qin Q, Feng J. Imputation for transcription factor binding predictions based on deep learning. PLOS Comput Biol. 2017;13.

Figure Legends

Figure 1. (**a**) Schematic representation of the integrated analysis workflow for identifying (causative SNP, TF, target gene) triplets. For inferring target genes (left part), eQTL analysis and a modified version of allele-specific expression analysis using the TCGA data are combined. For identifying causative SNPs and corresponding TFs (right part), epigenetics information, motif analysis and TF-target expression correlation analysis are used to filter the list of candidate causative variants. ChIP-seq data, allele-specific binding events and 3D chromatin interaction data are analysed when available. SNP: single-nucleotide polymorphism; eQTL: expression quantitative trait loci; LCASE: local chromosome allele-specific expression; LD: linkage disequilibrium; DHS: DNase I hypersensitive sites; TF: transcription factor; ASB: allele-specific binding; ChIA-PET: Chromatin Interaction Analysis by Paired-End Tag Sequencing; Hi-C: High-throughput chromosome conformation capture. (**b**) Visual illustration of the genomic analysis pipeline. Candidate SNPs are selected among the SNPs in strong LD with a GWAS SNP (yellow block) by overlapping with DHS (top track). The entire analysis is restricted to the topologically associated domain (TAD) containing the GWAS SNP.

Figure 2. Linkage structure and epigenetic annotation in the 5p12 region. Top triangle shows the linkage (color-coded by r^2 value) among 5p12 SNPs ordered according to their genomic locations. Middle track shows genes annotated by GENCODE v19. In the lower tracks, three GWAS SNPs in the 5p12 region are shown, followed by ChromHMM enhancer annotations in the breast cancer cell line MCF-7 and human mammary epithelial cells (HMEC). DNase I hypersensitive sites in T-47D and MCF-7 are also shown to represent open chromatin regions.

Figure 3. The risk allele of the GWAS SNP rs4415084 correlates with elevated MRPS30/RP11-53O19.1 expression. (**a**) Violin plots of MRPS30 and RP11-53O19.1 expression levels divided into the imputed genotypes at rs4415084, using the TCGA ER+ breast cancer patient data. The p-values are for the multivariate linear regression coefficients of *genotype*. See **Supplementary Table 4** for a full list of eQTL genes and GWAS SNPs in 5p12. (**b**) A schematic representation of local chromosome allele-specific expression (LCASE) analysis. For a certain exonic SNP of interest, we obtain all patients who have heterozygous genotypes both at the GWAS SNP and at the exonic SNP. Haplotype phasing is performed for the chromosome segment covering the GWAS SNP, the exonic SNP and all intermediate SNPs (**Methods**). The reference and alternative alleles of a biallelic SNP are denoted as *0* and *1*, respectively. In this figure, patient 1 and patient 2 have the *1* allele of the exonic SNP phased with the GWAS risk allele, whereas patient *K* has the *0* allele. RNA-seq read coverage is then counted in each patient to measure differential transcription activity between the risk chromosome (red) and the protective chromosome (blue). (**c**) LCASE analysis of exonic SNPs in the protein-coding gene MRPS30. The proportion of reads containing the protective alleles are plotted with the confidence intervals. Four of the six patient samples show significantly fewer reads emanating from the chromosome harboring the protective allele of rs4415084 (one-sided binomial test; $p = 1.3 \times 10^{-4}$, $p = 9.7 \times 10^{-17}$ for patient 1 and patient 2 at rs61754779, respectively; $p = 6.7 \times 10^{-47}$ for patient 4 at rs34522103; $p = 1.2 \times 10^{-3}$ for patient 5 at rs79210252), while patient 3 and patient 6 have non-significant p-values. (**d**) The

genomic locations of LCASE SNPs in the protein-coding MRPS30 and MRPS30 3' non-coding transcript. The p-values are from Wilcoxon signed-rank test with the red color showing transcription preference towards the risk chromosome.

Figure 4. The predicted causal SNP rs4321755 in LD with the GWAS SNP rs4415084 may regulate GATA3 binding. (**a**) Subsequence containing the risk allele T of rs4321755 matches the GATA3 motif, while the protective allele C disrupts the motif. The risk and protective alleles are determined by phasing with the alleles of GWAS SNP rs4415084 ($r^2 = 0.988$). (**b**) GATA3 expression positively correlates with predicted target gene expression. The correlation structure depends on the rs4321755 genotype status; i.e., as the number of risk allele increases, the correlation also increases. (**c**) ChIP-seq and DNase-seq data in T-47D show that rs4321755 is at the center of GATA3, FOXA1, and DNase I peaks (two replicate experiments of DNase-seq are shown: ENCODE accessions ENCFF001EGW and ENCFF001EHA). Shown for each experiment are the read coverage and raw aligned reads (positive strand: yellow; negative strand: cyan). In the read coverage figure, the range of y-axis values is indicated on top right, and the coverage of the putative causative SNP is color-coded based on the risk (red) and protective (blue) allele counts. (**d**) Zoomed-in view of ENCODE TF binding and PhyloP conservation track near rs4321755. (**e**) GATA3 ChIP-seq, PGR ChIP-seq and DNase-seq data show a significant skew towards the rs4321755-T risk allele. Replicates are pooled together and reads are deduplicated; the p-values are calculated by one-sided binomial test.

Figure 5. An illustration of the regulation model for MRPS30/RP11-53O19.1. The top chromosome carrying the protective allele C of the causal SNP rs4321755 has a disrupted GATA3 binding motif, thereby weakening the association between MRPS30/RP11-53O19.1 divergent promoter and the enhancer harboring the SNP. By contrast, the bottom chromosome carrying the risk allele rs4321755-T acquires a strong GATA3 motif, resulting in stronger binding of GATA3 and recruitment of other cofactors like FOXA1 and PGR, which together make this enhancer more active in regulating its target genes MRPS30 and RP11-53O19.1 via chromatin looping.

Figure 1

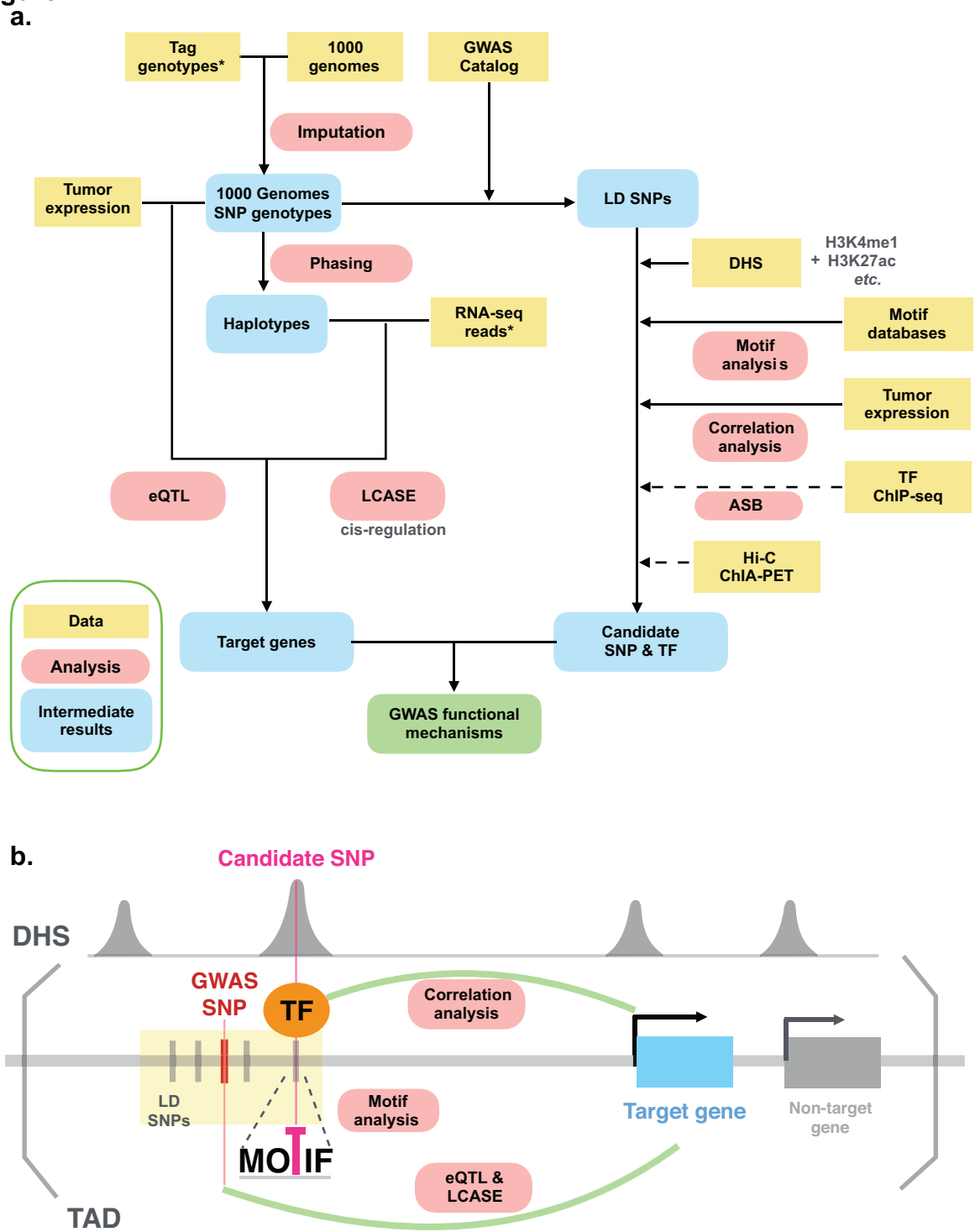


Figure 2

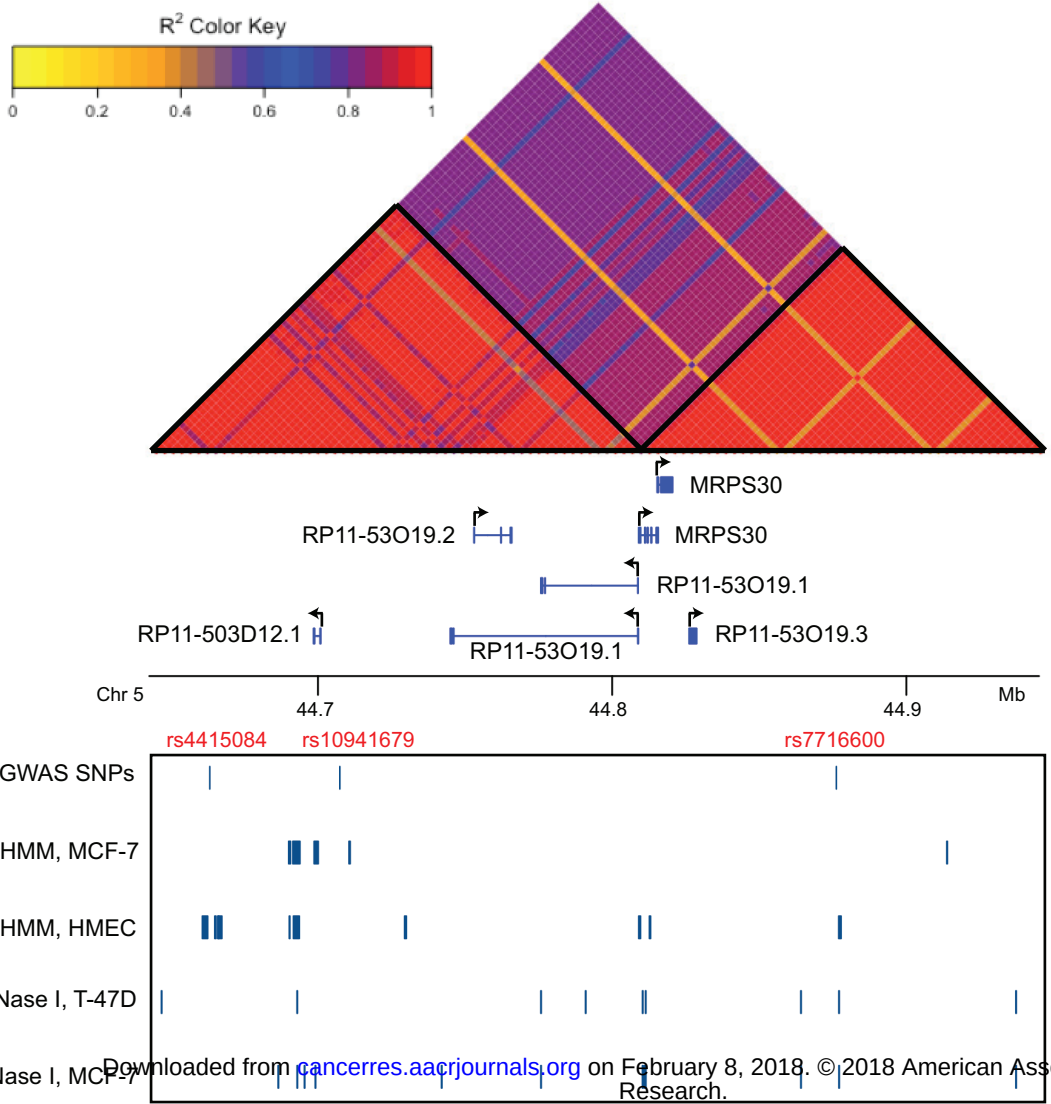
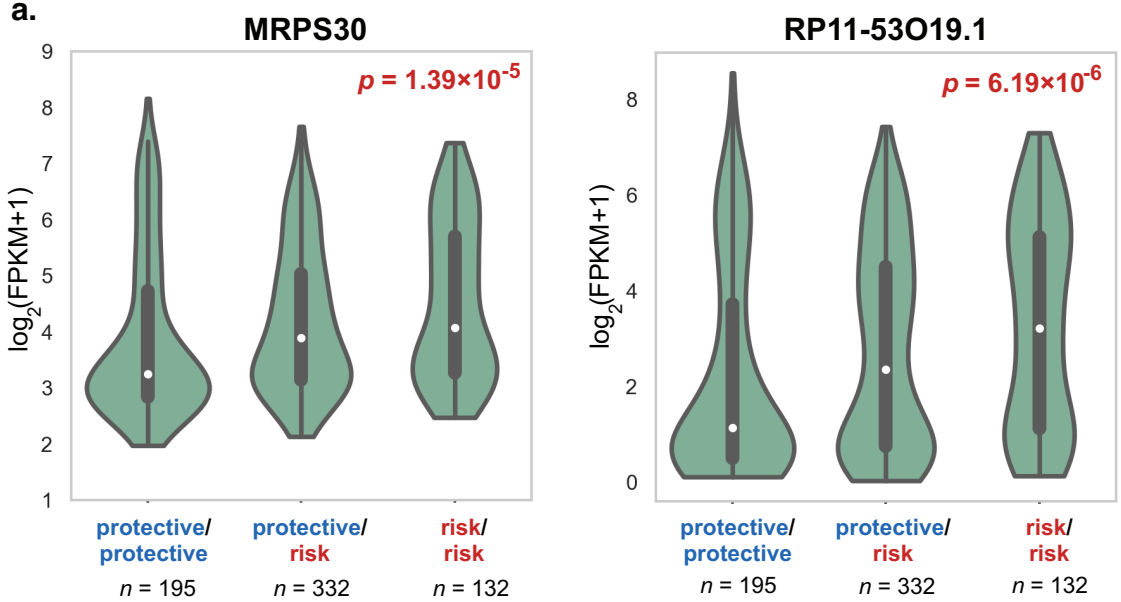
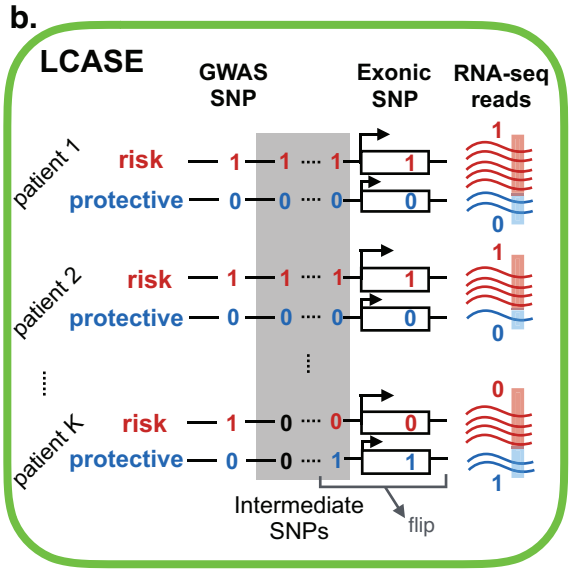
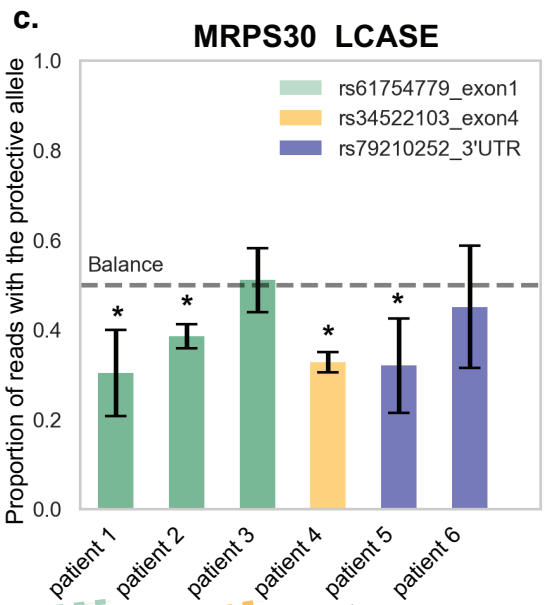


Figure 3







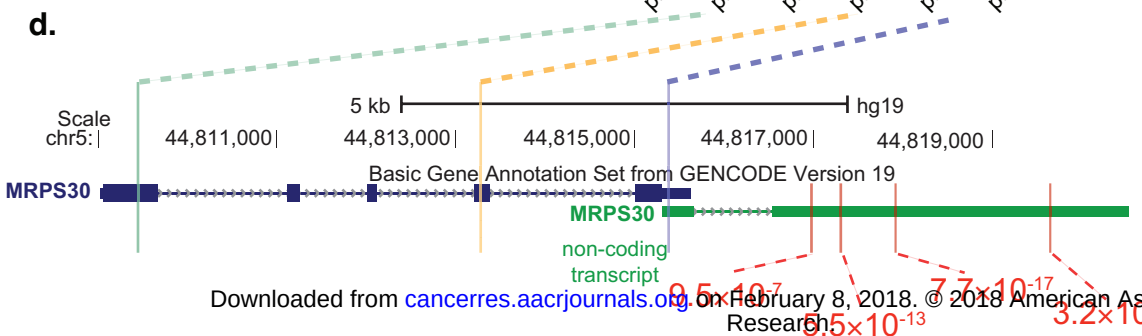


Figure 4

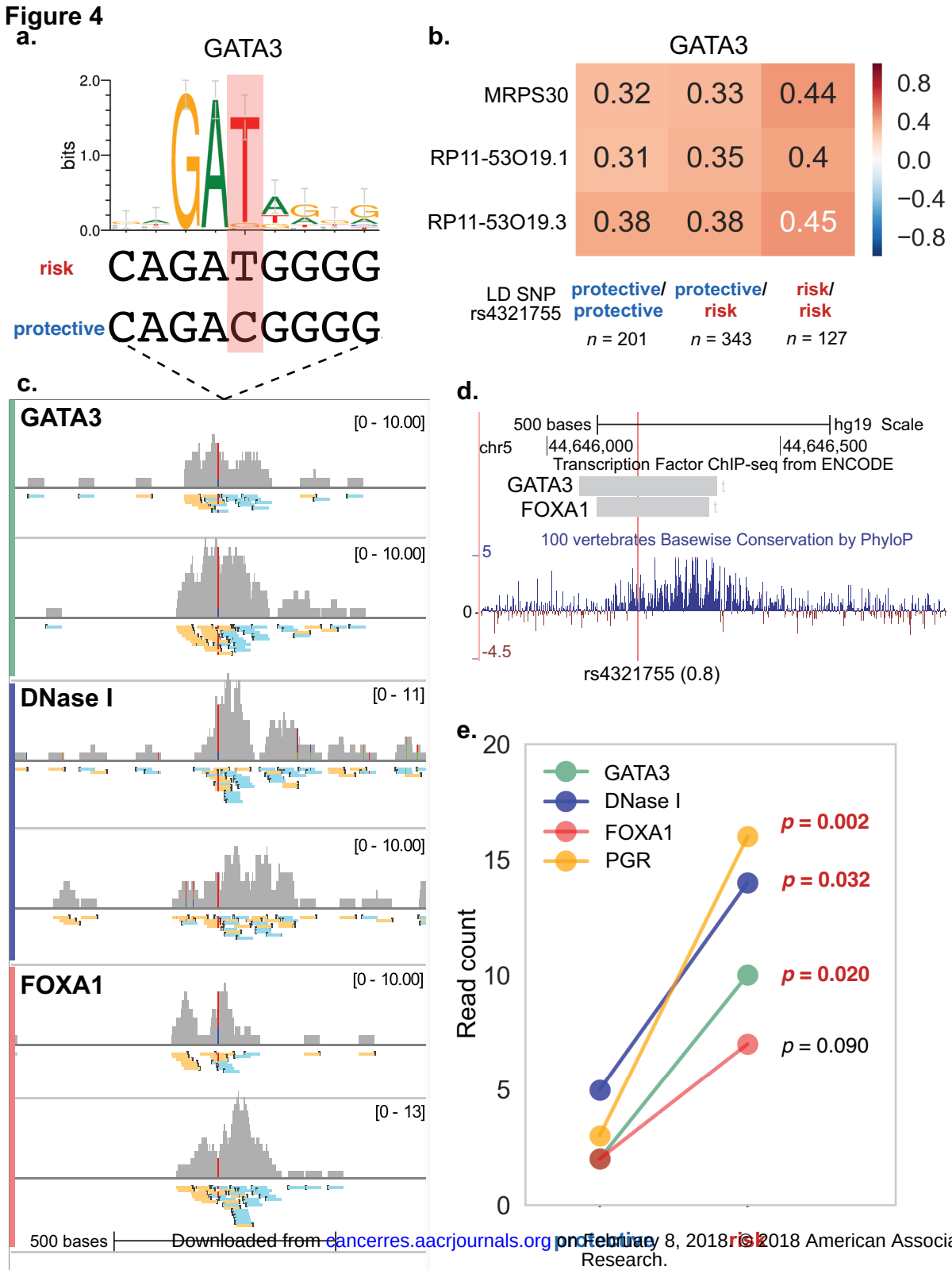
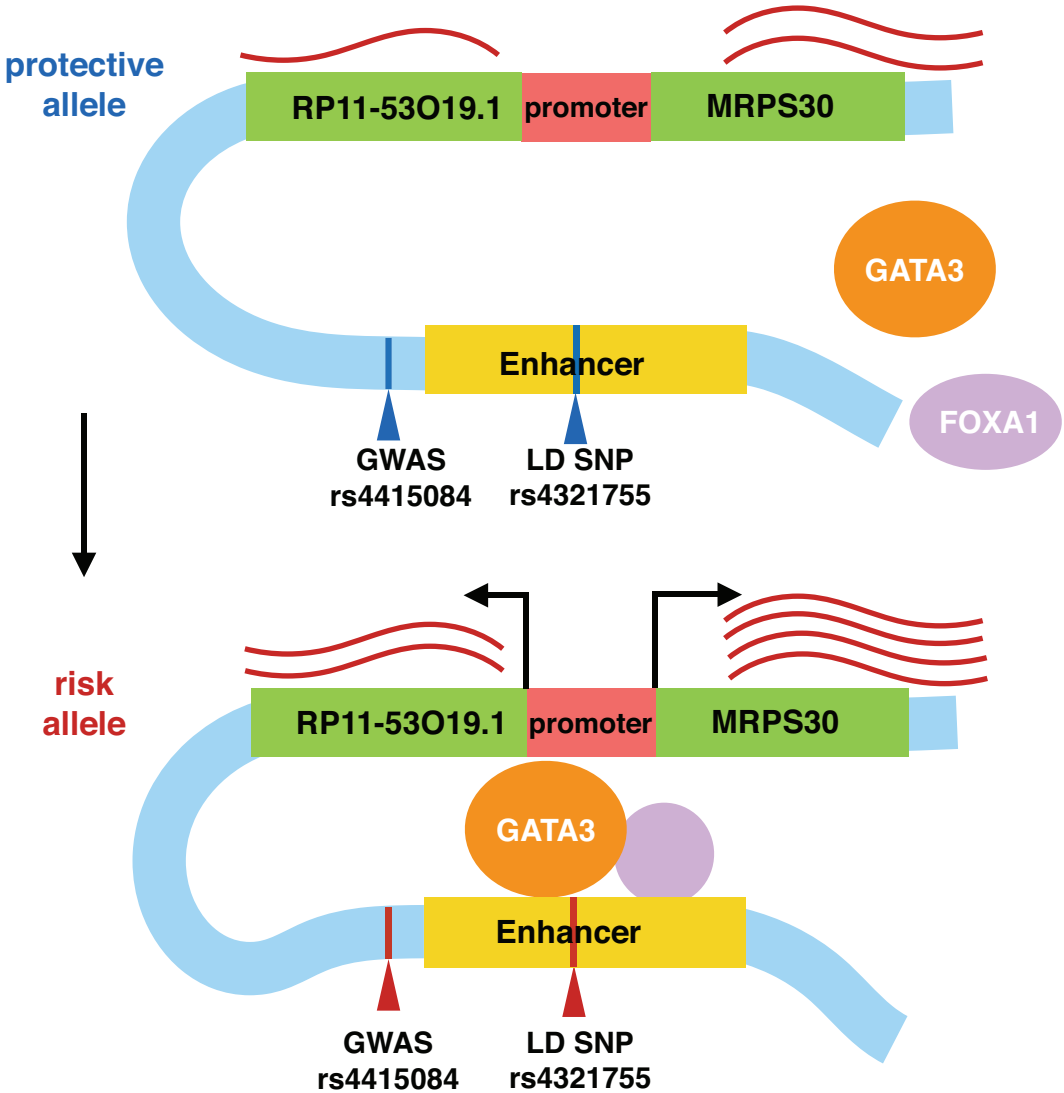


Figure 5



AACR American Association for Cancer Research

Cancer Research

The Journal of Cancer Research (1916–1930) | The American Journal of Cancer (1931–1940)

Integrative genomic analysis predicts causative cis-regulatory mechanisms of the breast cancer-associated genetic variant rs4415084

Yi Zhang, Mohith Manjunath, Shilu Zhang, et al.

Cancer Res Published OnlineFirst January 19, 2018.