Cancer Research

Discordant Haplotype Sequencing Identifies Functional Variants at the 2q33 Breast Cancer Risk Locus

Nicola J. Camp[1], Wei-Yu Lin[2], Alex Bigelow[1,3], George J. Burghel[2], Timothy L. Mosbruger[4], Marina A. Parry[2], Rosalie G. Waller[1], Sushilaben H. Rigas[2], Pei-Yi Tai[1], Kristofer Berrett[1], Venkatesh Rajamanickam[1], Rachel Cosby[1], Ian W. Brock[2], Brandt Jones[1], Dan Connley[2], Robert Sargent[1], Guoying Wang[1], Rachel E. Factor[1], Philip S. Bernard[1], Lisa Cannon-Albright[1], Stacey Knight[1], Ryan Abo[1], Theresa L. Werner[1], Malcolm W.R. Reed[2], Jason Gertz[1], and Angela Cox[2]

Abstract

The findings from genome-wide association studies hold enormous potential for novel insight into disease mechanisms. A major challenge in the field is to map these low-risk association signals to their underlying functional sequence variants (FSV). Simple sequence study designs are insufficient, as the vast numbers of statistically comparable variants and a limited knowledge of noncoding regulatory elements complicate prioritization. Furthermore, large sample sizes are typically required for adequate power to identify the initial association signals. One important question is whether similar sample sizes need to be sequenced to identify the FSVs. Here, we present a proof-of-principle example of an extreme discordant design to map FSVs within the 2q33 low-risk breast cancer locus. Our approach employed DNA sequencing of a small number of discordant haplotypes to efficiently identify candidate FSVs. Our results were consistent with those from a 2,000-fold larger, traditional imputation-based fine-mapping study. To prioritize further, we used expression-quantitative trait locus analysis of RNA sequencing from breast tissues, gene regulation annotations from the ENCODE consortium, and functional assays for differential enhancer activities. Notably, we implicate three regulatory variants at 2q33 that target CASP8 (rs3769823, rs3769821 in CASP8, and rs10197246 in ALS2CR12) as functionally relevant. We conclude that nested discordant haplotype sequencing is a promising approach to aid mapping of low-risk association loci. The ability to include more efficient sequencing designs into mapping efforts presents an opportunity for the field to capitalize on the potential of association loci and accelerate translation of association signals to their underlying FSVs. *Cancer Res; 76(7); 1916–25. ©2016 AACR.*

Introduction

The large number of compelling disease loci identified from disease-association studies (e.g., 1–4) provides immense potential for novel insight into disease mechanisms and new opportunities for diagnosis, prevention, and treatment. To date, 2,511 genome-wide association studies (GWAS) of 1,353 traits have generated 8,345 hits at $P \leq 5 \times 10^{-8}$ (5). However, the task of identifying the underlying functional sequence variants (FSV) has proven extremely challenging, and, thus far, very few FSVs have been identified for these types of low-risk disease loci (e.g., 6–8). A number of factors contribute to the challenge, including cost and an inadequate functional knowledge base. Performing DNA sequencing (DNA-seq) of the disease loci would capture the FSVs, but this is often financially impractical for large case–control designs (GWAS for common traits can include 100,000s of cases and controls). Furthermore, empirical evidence suggests that the vast majority of FSVs for these loci will be common, noncoding variants influencing gene expression (9). Hence, the identification of the FSV via simple prioritization of the sequence data may be difficult because many variants will be statistically comparable, and our understanding of noncoding regulatory elements is incomplete.

Instead of complete DNAseq of a GWAS region, a standard approach to identify cFSVs is a more detailed statistical

[1]University of Utah School of Medicine, Salt Lake City, Utah. [2]Department of Oncology and Metabolism, University of Sheffield, Sheffield, United Kingdom. [3]University of Utah School of Computing, Salt Lake City, Utah. [4]Bioinformatics Shared Resource, University of Utah, Salt Lake City, Utah.

Note: Supplementary data for this article are available at Cancer Research Online (http://cancerres.aacrjournals.org/).

Current address for W.-Y. Lin: Cardiovascular Epidemiology Unit, Department of Public Health and Primary Care, University of Cambridge, Cambridge, CB1 8RN, United Kingdom; current address for G.J. Burghel: Manchester Centre for Genomic Medicine, Manchester, M13 9WL, United Kingdom; Current address for M.A. Parry: Cancer Research UK Manchester Institute, University of Manchester, Manchester, M20 4BX, United Kingdom; current address for S.H. Rigas: School of Medical Sciences, Faculty of Life Sciences, University of Bradford, Bradford, BD7 1DP, United Kingdom; current address for S. Knight: Intermountain Heart Institute, Intermountain Medical Center, Salt Lake City, Utah 84107; current address for R. Abo: Dana-Farber Cancer Institute, Boston, Massachusetts 02215; and current address for M.W.R. Reed: Brighton and Sussex Medical School, University of Sussex, Brighton, BN1 9RH, United Kingdom.

Corresponding Author: Nicola J. Camp, University of Utah, Huntsman Cancer Institute, Salt Lake City, UT 84112. Phone: 801-587-9351; Fax: 801-585-5357; E-mail: nicola.camp@hci.utah.edu

doi: 10.1158/0008-5472.CAN-15-1629

AACR

interrogation of the region ("fine-mapping"). Often these fine-mapping studies are undertaken in the same large sample sets as the initial analyses, they explore evidence for multiple signals and use statistical imputation to estimate the association evidence for other common and rare variants in the region that were not directly genotyped (10). A caveat of statistical imputation is that it relies on external imputation panels that are not only distinct from the target populations, but are often much smaller and sequenced only at very low depth (e.g., 4×), leading to potential inaccuracies, especially for rare variants. DNAseq of cases and controls would remove the need for imputation, but remains cost prohibitive for large case–control studies. Here, we propose a cost-efficient extreme discordant haplotype DNAseq design that is financially viable and can greatly aid in the identification of cFSVs.

At a predefined risk locus, we statistically phase the genotype data and perform haplotype mining to define risk haplotype/s that optimally distinguish cases from controls. Phasing does not require external data. Under the hypothesis that FSV/s lie on a risk haplotype backbone, selecting cases homozygous for risk haplotype and controls without risk haplotypes results in an extreme discordant design with good power to detect underlying FSVs with small sample sizes conducive to quick and affordable high-throughput sequencing (HTS). To illustrate the gain in power of the discordant design, consider a FSV with population risk allele frequency (RAF) = 0.32 and an allelic OR = 1.09 (i.e., RAF = 0.34 in cases). Even if the associated SNP is in perfect linkage disequilibrium (LD) with the FSV ($r^2 = 1.0$), greater than 85,000 samples are required to identify it with 80% power at a significance of $\alpha = 5 \times 10^{-8}$. Now consider a haplotype discordant design: cases homozygous for the risk haplotype and controls with zero copies. If the risk haplotype were a perfect proxy, the FSV RAF would be enriched to 1.0 in cases and reduced to 0.0 in controls. Although enrichment will not be perfect, it will be greatly enhanced; thus, if RAFs become 0.9 (cases) and 0.1 (controls), a sample size of only 28 has 80% power at $\alpha = 5 \times 10^{-8}$ (see Supplementary Methods for more details).

As proof-of-concept, we present a discordant haplotype HTS DNAseq study in 38 individuals for the 2q33 (*CASP8*/*ALS2CR12*) region, a breast cancer risk locus (11). In order to address the lack of knowledge surrounding gene regulation and better prioritize our sequence variants (SV), we supplemented publicly available annotations with RNAseq in normal and tumor breast tissue, and identified three FSVs affecting enhancer activity.

Materials and Methods

An overview of study design detailing the sample sets used for the DNAseq and RNAseq analyses can be found in the Supplementary Data.

DNA sequencing and variant prioritization

Haplotype mining was performed previously using *hapConstructor* (12) based on 45 tagging-SNPs across 220 kb spanning *CFLAR*, *CASP10*, *CASP8*, and *ALS2CR12* in 3,888 individuals from the Utah and Sheffield Breast Cancer Studies (1,882 breast cancer cases and 1,896 controls). A six-SNP risk haplotype was defined ($P < 5 \times 10^{-6}$) that optimally extracted the association evidence across the region (13). Specifically, the alleles on the haplotype were ins-A-G-G-T-del across SNPs rs3834129-rs6723097-rs3817578-rs7571586-rs36043647-rs35010052. These six SNPs spanned *CASP8* and the intergenic region between *CASP8* and *ALS2CR12*. *HapMC* (14) was used to impute missing genotypes and infer phase on all 3,888 individuals. Female breast cancer cases were selected who were homozygous for the risk haplotype, and female cancer-free controls who had zero copies (with probability > 0.95). For these individuals, germ-line DNA (200 ng) from peripheral blood leukocytes was selected for the sequencing panel based on DNA quantity and quality requirements (3–5 μg of high quality, high molecular weight genomic DNA, free of RNA contamination, and at a minimum concentration of 100 ng/μL with A_{260}/A_{280} and A_{260}/A_{230} ratios ≥1.8 and ≥1.9, respectively; confirmed by Bioanalyzer). The final panel comprised 38 individuals (21 cases and 17 controls).

In order to capture any underlying FSVs tagged through long-range LD with the risk haplotype, we sequenced across a broader region. We thus defined the chromosome 2q33.1 low-risk breast cancer locus to be from 201,566,128 to 202,566,128 bp (hg19); the minimal region required to contain all variants in the HapMap or 1000G projects with a $r^2 \geq 0.1$ with the previously reported variants associated with breast cancer (rs1045485 and rs10931936; refs. 15,16). The region is gene-rich with 18 genes: *AOX2P*, *LOC100507140*, *BZW1*, *CLK1*, *PPIL3*, *NIF3L1*, *ORC2*, *FAM126B*, *NDUFB3*, *CFLAR*, *CASP10*, *CASP8*, *ALS2CR12*, *TRAK2*, *STRADB*, *ALS2CR11*, *TMEM237*, and *MPP4* (Fig. 1).

The SureSelect Target Enrichment system was used to generate libraries for next-generation sequencing based on all nonrepetitive genic and intergenic sequence in the 1 Mb region. Capture bait design was carried out using Agilent E-array and Windowsmasker (NCBI C++ Toolkit), resulting in 8,227 baits that mapped to 567,411 bp. Samples were multiplexed 12 per lane for ABI SOLiD sequencing to generate 50 bp single-end reads. Raw sequence reads were aligned using NovoalignCS (http://www.novocraft.com), developed specifically for color space SOLiD technology data. On average, 69.2% of reads accurately mapped to the 1 Mb region, and 609,165 bases had ≥10× coverage (mean, 76×). Alignment was followed by duplicate marking using Picard (http://broadinstitute.github.io/picard/) and GATK (17) best practice V2 for variant calling, including local realignment, recalibration, joint calling with UnifiedGenotyper, and variant filtration. A total of 1,197 positions were heterozygous in at least one individual in the sequencing panel, and passed quality control variant filtration with good variant quality score (Q > 30), 63 SVs were not acknowledged by the 1000G project and were removed from consideration, resulting in 1,134 SVs for comparisons between cases and controls.

In the homozygous discordant risk haplotype design, the FSV should reside on all case chromosomes and no control chromosomes. To prioritize variants based on this expectation, the allele frequency (AF) was calculated for both REF (the allele designated as the reference allele) and ALT (the alternate allele) alleles at each of the 1,134 SVs in both cases and controls. An SV was considered a cFSV if an allele had AF ≥ 0.9 in cases and AF ≤ 0.1 in controls. If so, that allele was considered the risk allele. Nine hundred and fifty-five SVs had an allele with AF ≥ 0.9 in cases, but only 18 of these also had AF ≤ 0.1 in controls. For these 18, the RAF varied from 0.905 to 1.0 in cases, and from 0.0 to 0.094 in controls. One variant aligned perfectly with expectations (rs3769821, RAFs of 1.0 and 0.0 in cases and controls, respectively). Chi-squared tests of allele counts by case/control group were calculated to help interpret these observed discrepant RAF frequencies. All were highly significant ($6.9 \times 10^{-13} \leq P \leq 2.8 \times 10^{-18}$; Table 1). Background frequencies for the risk alleles for the CEU (Utah residents)/GBR (Great Britain) 1000G samples are shown in Table 1.

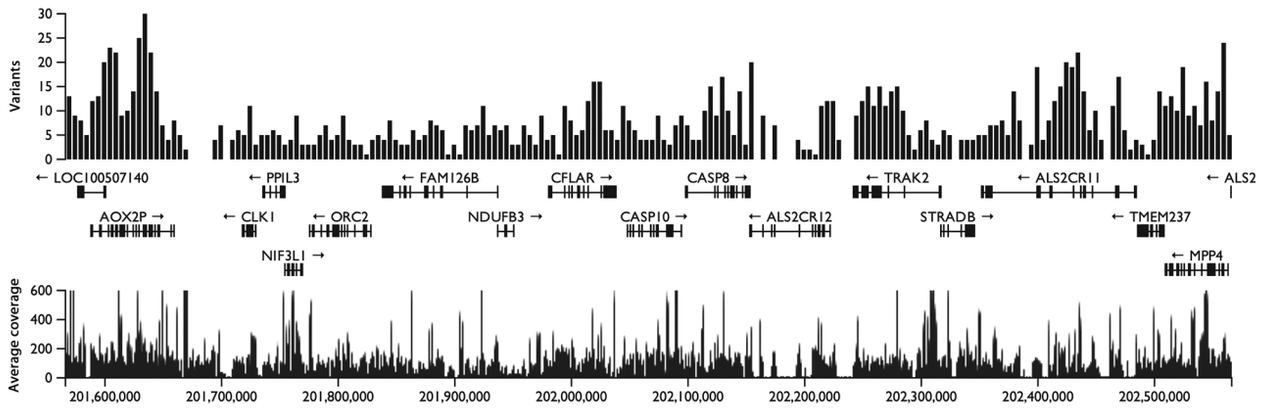

Figure 1.
DNA sequencing coverage and variants in the 1 Mb region. Eighteen genes reside in the region defined by chromosome 2 201,566,128 to 202,566,128 bp (hg19). Top, a graph of the number of variant positions identified in the DNA sequencing across the region. Bottom, a graph of the average depth of sequencing coverage in the 38 sequenced individuals.

RNA sequencing and differential expression

A panel of fresh-frozen breast tissue samples was assembled from 88 women who had surgery at the Huntsman Cancer Hospital from 2009 to 2012. This included paired tumor and normal (adjacent grossly uninvolved) tissues for 69 breast cancer patients and normal tissues from an additional 19 women undergoing breast reduction surgery. RNA was extracted from all tissue samples. For germ-line genotypes, DNA extracted from peripheral blood ($N = 15$), or from the normal tissue ($N = 73$) was used. For genotyping, we required 10 μL of DNA at a concentration of 50 ng/μL measured by PicoGreen. An Illumina BeadExpress assay was designed for the 18 cFSVs. One variant failed design (rs10635401) and another failed genotyping QC (rs3769820), leaving 16 cFSVs with good quality genotype data for differential expression analyses (carriage of risk allele vs. no copies). For RNAseq, we required 50 μL of RNA at a concentration of 25 ng/μL, and RIN $\geq$ 6.0. Bioanalyzer results indicated that quality was generally high (average RIN = 7.9), although one tumor sample yielded poor quality RNA (RIN=2.5) and was removed from consideration, resulting in a panel of 156 breast samples (88 normal, 68 tumor). RNAseq was carried out using Illumina TruSeq Stranded mRNA sample preparation with oligo dT selection, and samples were multiplexed 8 per lane on the Illumina HiSeq2000 to generate 50 bp single-end reads. Raw sequence reads were aligned using Novoalign (V2.08.01) against the reference genome plus extended splice junctions (hg19 annotations for gene model). We used USeq, an extensive package of open-source RNAseq workflow written and benchmarked by the University of Utah Bioinformatics Shared Resource (18), to quality control, wrap programs, manage, and prepare files. Two samples failed RNAseq QC and were removed (degraded signatures by Picard) leaving 154 samples (86 normal, 68 tumor), with an average of 19.1 M reads per sample for analysis. We performed differential expression analyses based on carriage of risk alleles at cFSVs using DESeq2 (19), which employs a negative binomial distribution to test for differences. The RNAseq panel provided

Table 1. Candidate functional SVs from DNA sequencing

	Position (hg19)	Gene	Variant type	Risk allele	Other allele	1000G background (CEU/GBR) RAF	Correlation with risk_haplotype in BG	Case (homozygous risk haplotype) RAF	Control (homozygous nonrisk) RAF	DNAseq haplotype discordant *P* value[a]	RNAseq panel genotyping
rs12990906	202114624	CASP8	Intronic	T	C	0.397	0.85	1	0.088	8.8×10^{-16}	Y
rs10931934	202119789	CASP8	Intronic	T	C	0.397	0.85	1	0.029	2.1×10^{-17}	Y
rs3769823	202122995	CASP8	Exonic nonsyn	A	G	0.296	0.89	0.929	0	8.1×10^{-16}	Y
rs3769821	202123430	CASP8	Intronic	C	T	0.333	0.97	1	0	2.8×10^{-18}	Y
rs10635401	202123717	CASP8	Intronic	C	CAAGT	0.313	0.90	0.976	0.029	1.5×10^{-16}	N
rs6735656	202124502	CASP8	Intronic	G	T	0.270	0.81	0.929	0.094	6.9×10^{-13}	Y
rs6754084	202124997	CASP8	Intronic	T	C	0.270	0.84	0.905	0.059	2.1×10^{-13}	Y
rs1861270	202126615	CASP8	Intronic	A	G	0.270	0.84	0.905	0	4.4×10^{-15}	Y
rs6751053	202127863	CASP8	Intronic	G	A	0.374	0.89	1	0.029	2.1×10^{-17}	Y
rs6435074	202127947	CASP8	Intronic	A	C	0.270	0.84	0.905	0	4.4×10^{-15}	Y
rs6723097	202128618	CASP8	Intronic	A	C	0.374	0.89	1	0.029	2.1×10^{-17}	Y
rs3769820	202130133	CASP8	Intronic	T	C	0.374	0.89	1	0.029	5.7×10^{-17}	N
rs2349070	202130308	CASP8	Intronic	A	C	0.270	0.84	0.905	0	4.4×10^{-15}	Y
rs10931936	202143928	CASP8	Intronic	T	C	0.282	0.80	0.905	0	4.4×10^{-15}	Y
rs3769818	202151163	CASP8	Intronic	A	G	0.282	0.80	0.905	0	4.4×10^{-15}	Y
rs700635	202153225	ALS2CR12	UTR3	C	A	0.282	0.80	0.905	0	4.4×10^{-15}	Y
rs6714430	202153684	ALS2CR12	Intronic	C	T	0.282	0.80	0.905	0	4.4×10^{-15}	Y
rs6743068	202153920	ALS2CR12	Intronic	A	G	0.282	0.80	0.905	0	4.4×10^{-15}	Y

Abbreviation: Exonic nonsyn, exonic nonsynonymous variant.
[a]Allele-based χ^2 test for independence.

80% power to find ratios $\geq$1.19 or $\leq$0.84 (equivalent to fold differences in expression of $\sim$20%), at a nominal significance level ($\alpha = 0.05$; ref. 20).

We selected *CFLAR*, *CASP10*, *CASP8*, and *ALS2CR12*, which span the central $\sim$0.25 Mb of the 1 Mb region, as potential target genes for our cFSVs. The transcriptome-normalized counts, averaged across all samples, were 79.9, 27.1, 38.4, and <1.0 per million mapped reads, respectively. This indicated that *CFLAR*, *CASP10*, and *CASP8* were expressed sufficiently well in the breast to pursue expression-quantitative trait locus (eQTL) analyses, but that *ALS2CR12* was not well expressed and was not considered further. Consistent with our data, Encyclopedia of DNA Elements (ENCODE) RNAseq data also indicated that *CFLAR*, *CASP10*, and *CASP8* were expressed in breast tissues. All were expressed in the human mammary epithelial (HMEC) normal breast cell line, although the level of *CASP10* expression was lower than the other genes; *CASP8* and *CFLAR* were also expressed in the breast cancer cell line MCF-7. These three genes are also good functional candidates: *CASP8* and *CASP10* are proapoptotic, whereas *CFLAR* (caspase-8 and FADD-like apoptosis regulator) is antiapototic, structurally similar to caspase-8 but lacking the proapoptotic caspase death domains. Because it has been suggested that more than 40% of enhancers do not target the nearest gene, we tested for association between each cFSV and the all three genes (21).

The Cancer Genome Atlas RNAseq data for replication

Genome-wide germ-line SNP genotyping and whole-transcriptome RNAseq data for 97 normal (adjacent grossly uninvolved) tissues and 753 tumor breast tissue samples are available as part of The Cancer Genome Atlas (TCGA) Research Network (http://cancergenome.nih.gov/). Genotypes were available for only three SNPs from our list of 18 cFSVs. Transcriptome-normalized counts (with correction for differences in transcriptome length and specific gene annotations from hg18 to hg19) indicated expression levels of 149.1, 16.0, and 89.9 per million mapped reads for *CFLAR*, *CASP10*, *CASP8*, respectively, and differential analyses were performed on all three genes in both normal and tumor tissues. We note that the average standardized counts for *CASP10* fell below 20 counts per million mapped reads, and more caution should be exercised in interpreting findings for that gene in these data.

Comparison to traditional fine-mapping: meta association evidence from Breast Cancer Association Consortium and GWAS

The R package *metaphor* (http://cran.r-project.org/web/packages/metafor) was used to carry out a meta-analysis for the cFSVs, based on the ORs and 95% confidence intervals provided for 46,450 cases and 42,600 controls (Breast Cancer Association Consortium, BCAC) and 22,627 cases and 10,052 controls (9 GWAS) in Lin and colleagues (11).

Potential functional relevance: luciferase assays for enhancer activity

Variants rs3769823, rs3769821, and rs10197246 were chosen to investigate for ability to affect enhancer activity. To create the enhancer constructs, we performed PCR using genomic DNA from T-47D, which is homozygous for risk alleles at all three SNPs, and MCF10A, which is homozygous for neutral alleles, using the following primer pairs:

rs3769821/rs3769823 forward (1,531 bp product), TACCTGAGCTCGCTAGCCGGATCAATGCTACAAAGACACG

rs3769821/rs3769823 reverse, GGCCAGATCTTGATATCCCAGTCAC-CTCTGGAGGCATT

rs10197246 forward (758 bp), TACCTGAGCTCGCTAGCCGCTGTTTAATTTCCATGCGTTT

rs10197246 reverse, GGCCAGATCTTGATATCCTCTTTAGCAGTAGCACAACACAAA

Phusion HF master mix (NEB) and 10 ng genomic DNA were used for PCR, and PCR products were purified with AMPure XP beads (Beckman Coulter). We then digested pGL4.23 (Promega) with XhoI (NEB) and combined the digest with the purified PCR products to perform Gibson Assembly (NEB) according to the manufacturer's instructions. Clones were verified by Sanger sequencing, and no differences between the haplotypes, except for the expected alleles, were identified.

T-47D (ATCC HTB-133) and MCF10A (ATCC CRL-10317) were obtained from the ATCC, where short tandem repeat profiling analysis was used to authenticate the cell lines, and were grown within 6 months of resuscitation. Each line was cultured according to ATCC recommendations with the following modification: no insulin was added to the media for T-47D. Cells were plated at a density of 10,000 cells/well in 96-well plates, and after 24 hours, Lipofectamine 3000 (Life Technologies) was used to transfect the enhancer constructs as well as pGL4.23 as a negative control. Forty-eight hours after transfection, each well was assayed for luminescence using Steady-Glo Luciferase Assay System (Promega) and read on a GloMax luminometer (Promega). Luminescence was background subtracted using nontransfected wells and normalized by pGL4.23 levels. T tests were used to determine significance of expression differences between alleles.

Ethics approvals

The women whose samples were involved in the DNAseq and RNAseq experiments described here were enrolled in studies approved by the University of Utah Institutional Review Board or South Yorkshire Research Ethics Committee. Informed consent was obtained from all research participants.

Results

Using HTS technology, we performed targeted DNAseq of all nonrepetitive sequence across 1 Mb at 2q33.1 (201.57–202.57 Mb, hg19) in 38 women selected for extreme discordance of a previously defined risk haplotype (21 breast cancer cases homozygous for the risk haplotype and 17 female, cancer-free controls with zero copies; ref. 13). We identified 1,134 high quality SVs that were also confirmed as SVs in the 1000 Genomes (1000G, http://www.1000genomes.org/) project data (Fig. 1).

In the optimal situation, the FSV would reside on all 42 case chromosomes and no control chromosomes. In selecting cFSVs, we allowed 10% discrepancy, such that an SV was considered a candidate if RAF $\geq$ 0.9 in cases and RAF $\leq$ 0.1 in controls. This resulted in 18 cFSVs (15 in *CASP8* and 3 in the adjacent gene *ALS2CR12*): one was exonic (*CASP8*), 16 intronic, and one in the 3′ untranslated region of *ALS2CR12* (Table 1). As expected, AF differences between the cases and controls were all highly significant due to the discordant design (all $P < 10^{-12}$). Risk alleles were all very common in the general population (1000G RAFs = 0.27–0.40).

Table 2. eQTL results for *CFLAR*, *CASP10*, and *CASP8*

		CFLAR							CASP10							CASP8						
		Normal[a]		Tumor		TCGA normal		TCGA tumor	Normal[a]		Tumor		TCGA normal		TCGA tumor	Normal[a]		Tumor		TCGA normal		TCGA tumor	
	Position (hg19)	Ratio	*P* value	Ratio	*P* value	Ratio	*P* value	Ratio *P* value	Ratio	*P* value	Ratio	*P* value	Ratio	*P* value	Ratio *P* value	Ratio	*P* value	Ratio	*P* value	Ratio	*P* value	Ratio *P* value
rs12990906	202114624	1.06	ns	1.02	ns				1.00	ns	1.14	ns				0.93	0.11	1.01	ns			
rs10931934	202119789	0.99	ns	0.95	ns				0.99	ns	1.14	ns				**0.91**	**0.043**	0.94	ns			
rs3769823	202122995	1.01	ns	0.94	ns				1.00	ns	1.11	ns				**0.89**	**0.0086**	0.94	ns			
rs3769821	202123430	0.98	ns	0.96	ns				0.97	ns	1.12	ns				**0.90**	**0.023**	0.92	ns			
rs10635401[b]	202123717																					
rs6735656	202124502	0.99	ns	0.97	ns				0.94	ns	1.06	ns				**0.88**	**0.0032**	0.87	0.0703			
rs6754084	202124997	0.99	ns	0.97	ns	0.99	ns	1.02 ns	0.94	ns	1.06	ns	1.01	ns	0.99 ns	**0.88**	**0.0032**	0.87	0.0703	**0.86**	**0.0011**	0.97 ns
rs1861270	202126615	0.99	ns	0.97	ns				0.94	ns	1.06	ns				**0.88**	**0.0032**	0.87	0.0703			
rs6751053	202127863	1.04	ns	1.05	ns				0.98	ns	1.11	ns				0.92	0.11	0.99	ns			
rs6435074	202127947	0.99	ns	0.97	ns				0.94	ns	1.06	ns				**0.88**	**0.0032**	0.87	0.0703			
rs6723097	202128618	1.04	ns	1.05	ns	1.01	ns	1.00 ns	0.98	ns	1.11	ns	1.09	ns	0.94 ns	0.92	0.11	0.99	ns	**0.91**	**0.011**	1.00 ns
rs3769820[b]	202130133																					
rs2349070	202130308	0.99	ns	0.97	ns				0.94	ns	1.06	ns				**0.88**	**0.0032**	0.87	0.0703			
rs10931936	202143928	1.01	ns	0.97	ns				0.98	ns	1.08	ns				**0.86**	**0.00036**	0.87	0.058			
rs3769818	202151163	1.01	ns	0.97	ns				0.98	ns	1.08	ns				**0.86**	**0.00036**	0.87	0.058			
rs700635	202153225	1.01	ns	0.97	ns				0.98	ns	1.08	ns				**0.86**	**0.00036**	0.87	0.058			
rs6714430	202153684	1.01	ns	0.97	ns				0.98	ns	1.08	ns				**0.86**	**0.00036**	0.87	0.058			
rs6743068	202153920	1.01	ns	0.97	ns	0.99	ns	1.02 ns	0.98	ns	1.08	ns	0.99	ns	0.96 ns	**0.86**	**0.00036**	0.87	0.058	**0.86**	**2.6E−04**	0.97 ns

NOTE: ns, a *P* value > 0.2; bold indicates a nominal significant expression ratio ($P \leq 0.05$).

[a]Normal includes both adjacent grossly uninvolved tissue from cancer patients and breast reduction tissue from cancer-free patients.

[b]The cFSV failed to genotype on the Illumina BeadExpress platform.

To assess the cFSV for potential functional consequences, we performed eQTL analysis using high quality RNAseq data for *CASP8*, *CASP10*, and *CFLAR* in 86 normal tissue and 68 tumor samples, examining expression differences based on germ-line DNA genotype data for 16 of 18 cFSVs (Table 1). The expression of *CASP8* was significantly decreased for carriers of 12 of 16 cFSVs within normal breast tissue (Table 2), most significant for a group of five cFSVs at the 3′ ends of *CASP8* and *ALS2CR12*, between positions 202,143,928 and 202,153,920 bp (expression ratio = 0.86, $P = 3.6 \times 10^{-4}$; Figs. 2 and 3). Reduction in expression with carriage of these cFSVs was consistent in the subsets of adjacent grossly uninvolved and breast reductions (data not shown). No significant differences based on carriage of risk alleles were found for *CASP8* in tumor (although ratios were in the same direction as for the normal tissue, but less extreme). No differential expression was apparent based on cFSVs in *CASP10* or *CFLAR* in normal or tumor tissues. We replicated eQTL results using data from TCGA Research Network (http://cancergenome.nih.gov/) project, where RNAseq data were available for 97 normal breast and 753 tumor tissue samples, and germ-line genotypes for 3 of 18 cFSVs. Similarly, decreased *CASP8* expression was found for risk alleles at three cFSVs in normal tissue, with the most significant association for rs6743068 (ratio = 0.86, $P = 2.6 \times 10^{-4}$), which resides in the cluster of five most significant cFSVs. As before, there was no evidence for *CASP8* expression differences based on cFSVs in tumor tissues, or for *CFLAR* or *CASP10* expressions in either tissue type.

This small but highly focused haplotype discordant study demonstrates that cFSVs can be identified that are associated with gene expression. However, equally important is how this information aligns and adds value to results from a traditional "fine-mapping" study. To achieve this, we aligned our cFSVs with those suggested by the fine-mapping study by the BCAC that used 89,050 samples and 1,733 imputation fine-mapping SNPs across the same 1 Mb region (11). The BCAC study identified one region that achieved genome-wide significance (referred to as iCHAV1, $P = 1.1 \times 10^{-9}$; ref. 11). The BCAC iCHAV1 SNP set corresponds closely with the set of cFSVs identified here. Nine SVs were present in both sets, seven present only in the BCAC set, and nine present only in the discordant haplotype set (Table 3). The seven SVs present only in the BCAC results were in genomic regions 1000G designated as inaccessible to HTS technologies and were not captured by our sequencing baits. This highlights a limitation of HTS technology, but also suggests caution with imputation in this region, which is based on low coverage sequence data (6 of the 7 missed SNPs were imputed). We genotyped the remaining SV, rs10197246, which indicated it would have passed our cFSV criteria (Table 3). Of the nine cFSVs present only in our discordant haplotype study, several only narrowly missed BCAC inclusion. For example, our set included both rs6735656 and rs10635401, but only the latter was captured in the BCAC set due to slight superiority in significance in that data. Interestingly, in a combined BCAC + 9 GWAS meta-analysis, the order of significance is reversed for these two SNPs (Table 3). Hence, the discordant haplotype DNAseq study not only aligned with, but also added value to, the very large BCAC fine-mapping study.

Based on the combined 25 cFSVs, we used publicly available annotations and RNAseq eQTL results to prioritize for functional follow-up (Table 3). The three most compelling cFSVs are rs3769823, rs3769821, and rs10197246. All three yield highly significant results in both the discordant haplotype and

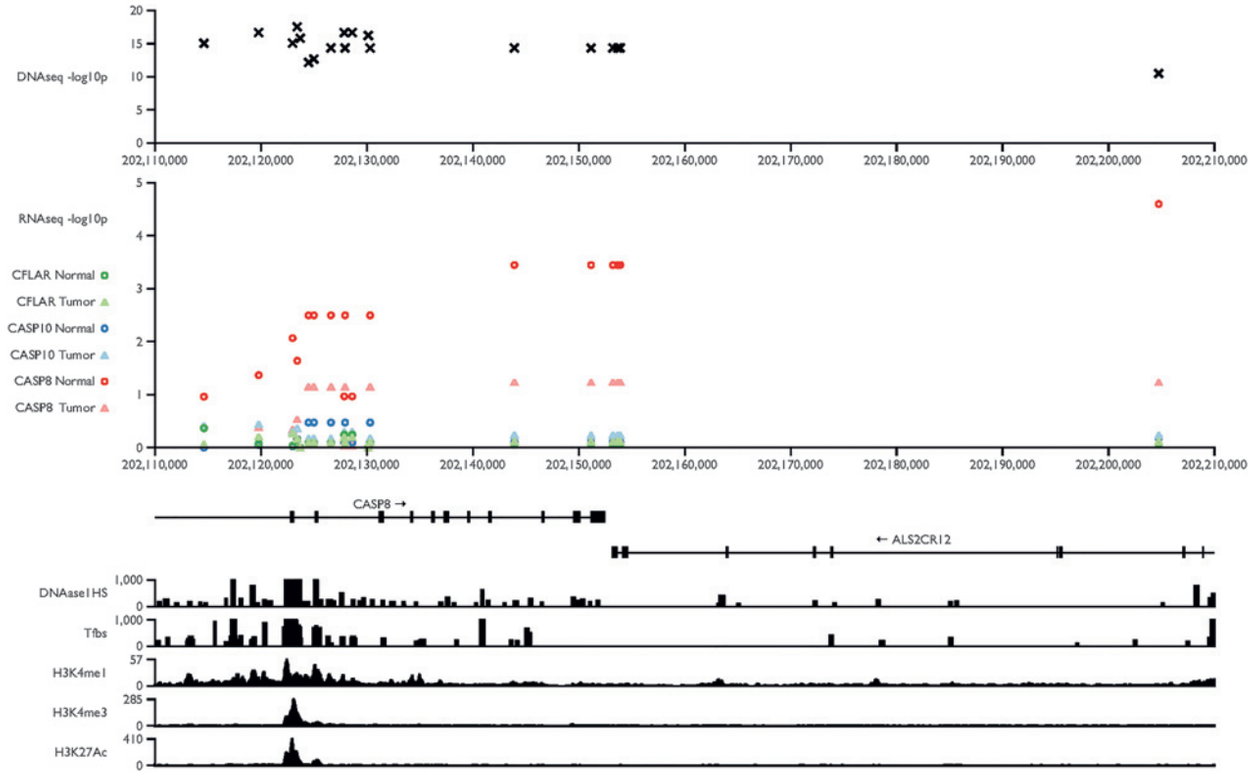

Figure 2.
Association and eQTL evidence and regulatory annotations for 17 candidate functional SVs. Seventeen cFSVs are illustrated (16 selected from DNA sequencing, plus rs10197246). These reside within 100 kb region at chromosome 2 202,110,000 to 202,210,000 bp (hg19). Two genes, *CASP8* and *ALS2CR12*, reside in this genomic region. Top, a graph showing the association evidence ($-\log_{10}p$) for each of the 17 cFSVs in the discordant haplotype design. Middle, a graph showing eQTL evidence ($-\log_{10}p$) based on RNA sequencing data in the local tissue panel for expression of three genes: *CASP8* (red), *CASP10* (blue), and *CFLAR* (green) in normal (circle) and tumor (triangle) breast tissues. Bottom, five regulatory annotation tracks from UCSC Genome Browser using ENCODE data (22,29). In order, from top to bottom: (i) DNAseIHS, DNAse I hypersensitivity clusters in 125 cell types; (ii) Tfbs, transcription factor ChIP-seq (161 factors); (iii) H3K4Me1, layered H3K4Me1 histone modification marks (often found near regulatory elements) on 7 cell lines; (iv) H3K4Me3, layered H3K4Me3 histone modification marks (often found near promoters) on 7 cell lines; (v) H3K27Ac, layered H3K27Ac histone modification marks (often found near active regulatory elements) on 7 cell lines.

traditional fine-mapping approaches and are associated with *CASP8* expression differences in normal breast tissue. The first two reside 435 bp apart in a promoter proximal region that exhibits DNAse I hypersensitivity in mammary epithelial and breast cancer cell lines, and binds multiple transcription factors (STAT3, MYC, Pol2, and CTCF) based on ENCODE data (22). Also notable are results from an eQTL study in blood (23), indicating highly significant decreased *CASP8* expression associated with these cFSVs ($Z = -5.44$, $P = 5.3 \times 10^{-8}$) and suggesting the expression signature is also readily observed in blood (Table 3). The latter cFSV, rs10197246, resides in a potential distal regulatory element, 80 kb from a *CASP8* promoter, that is bound by MYC in the breast cancer cell line MCF-7. Hence, MYC-triggered apoptosis may be one possible mechanism for the association, worthy of investigation.

We tested for functional differences between allelic variants of rs3769823, rs3769821, and rs10197246, by performing enhancer assays based on a luciferase reporter gene in cell lines derived from MCF10A (normal breast tissue) and T-47D (breast tumor). Figure 4 shows the enhancer activity of the risk and neutral alleles in each cell line. We found that the region containing the risk alleles of rs3769821 and rs3769823 drove significantly lower gene expression in both the normal and tumor cell lines compared with the same region harboring the neutral alleles. The region surrounding the risk allele of rs10197246 produced approximately 3-fold lower expression compared with the neutral allele in the normal breast cell line, but there was no difference between alleles in tumor cells. Overall, the functional results are consistent with our eQTL study and suggest that these cFSVs cause reproducible reduction of *CASP8* expression in normal breast tissue.

Discussion

In our proof-of-principle analysis of the 2q33 locus for breast cancer, we considered our focused extreme discordant haplotype study together with the much larger BCAC traditional fine-mapping study, in addition to tissue-specific RNAseq and public annotations, to select rs3769823, rs3769821, and rs10197246 as the three most compelling cFSVs for breast cancer risk. The first two SVs are 435 bp apart in *CASP8*, whereas rs10197246 resides 81 kb telomeric within the adjacent *ALS2CR12* gene. Despite the 81 kb physical separation, all three cFSVs target *CASP8*, based on the eQTL data. Although all three lie on the risk haplotype, LD between rs3769823/rs3769821 and rs10197246 is not

particularly strong ($r^2 = 0.66$), suggesting that there may be two functional hits on the haplotype. This hypothesis is consistent with the luciferase enhancer activity results that illustrate that risk alleles at both rs3769823/rs3769821 and rs10197246 independently and significantly decreased enhancer activity. Of particular interest was that *CASP8* expression differences were consistently stronger in normal breast tissue compared with breast tumor material, both in the frozen tissue samples (RNA-seq data) and the representative cell lines (enhancer assays). These observations support the hypothesis that these three variants are functionally relevant breast cancer risk variants that create an environment more conducive to tumorigenesis by reducing *CASP8* gene expression (and hence levels of apoptosis) in the normal breast. Moving forward, MYC-triggered apoptosis may be one possible mechanism worthy of investigation, given rs3769823 and rs10197246 appear good functional candidates and both reside in regions bound by MYC in the breast cancer cell line MCF-7. The importance of *CASP8* in breast cancer tumorigenesis was also recently underscored by its identification as one of 32 genes significantly somatically mutated in breast cancer tumors (24), and one of only 22 genes identified as significantly mutated in at least three different common tumor types (25), suggesting that other common cancers may follow the pattern for susceptibility that we see here.

Our interrogation of the 2q33 breast cancer risk locus implicates extremely common FSVs (RAF∼0.28) with small effect sizes (∼17% decreased gene expression), consistent with the common disease common variant model. The largely noncoding candidates identified and the consistent stronger significance in normal tissues suggest that greater availability of resources for gene expression and other epigenetic data in normal tissues may prove critical in the prioritization and identification of common, low-risk variants. Initiatives, such as the Roadmap Epigenomics Project and ENCODE, will certainly help in this regard as they come to fruition (26). Here, we were successful in illustrating differences in 88 local samples that could be replicated based on 97 samples from TCGA. However, in general, much larger sample sizes may be needed to generate robust findings.

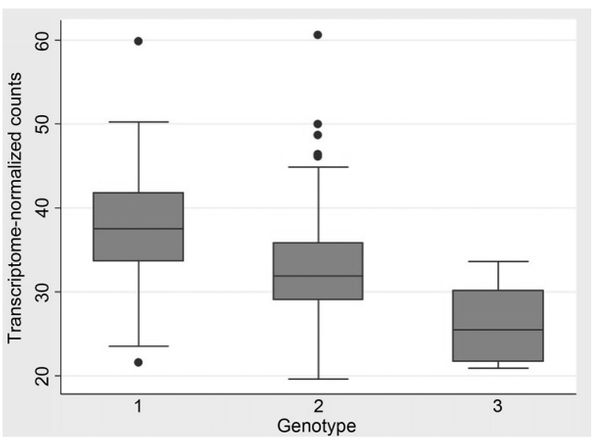

Figure 3.
Expression levels of *CASP8* in normal breast tissue. The three categories on the *x*-axis indicate genotype at rs10931936, rs3769818, rs700635, rs6714430, or rs6743068 (all five of these cFSV are identical for genotype in our RNAseq panel of women). These cFSVs exhibited the most significant evidence for association with *CASP8* expression in normal breast tissue (ratio = 0.85, $P = 5 \times 10^{-4}$). Genotype = 1 includes women homozygous for the common allele; genotype = 2 includes heterozygous women; and genotype = 3 includes women homozygous for the risk allele.

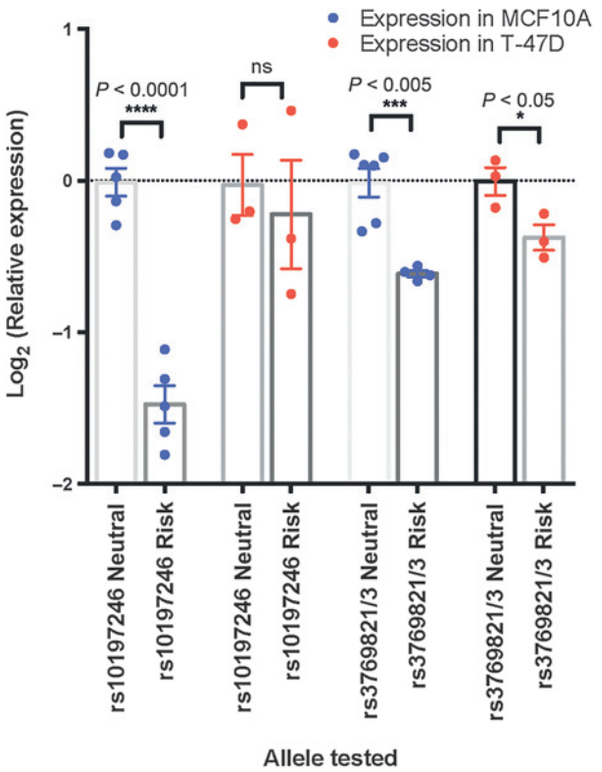

Figure 4.
Allele-specific enhancer activity for rs3769821/rs3769823 and rs10197246. The bar graph shows the $\log_2$ of the relative expression of the risk allele to the neutral allele, within a cell type, as measured by luciferase output in an enhancer assay. Expression of the neutral allele was used to normalize each pair of experiments. Expression measured in MCF10A is shown in blue, and expression measured in T-47D is shown in red. Significance level is indicated above each pair. ns, not significant. Neutral/risk alleles are as follows: rs3769821, A/G; rs3769823, C/T; and rs10197246, C/T.

Our small, focused, discordant haplotype DNAseq study identified 18 cFSVs for the *CASP8/ALS2CR12* region. These results aligned with and extended the 16 cFSV suggested from a traditional fine-mapping association study almost four orders of magnitude larger, leading to a combined set of 25 cFSVs, and resulting in three variants (rs3769823, rs3769821, and rs10197246) implicated as functionally relevant breast cancer risk alleles influencing *CASP8* expression. The discordant haplotype sequencing and traditional fine-mapping approaches have distinctly different strengths and weaknesses: statistically, technologically, and economically. Large population-based association studies attain good power to identify common, low-risk loci with accurate risk estimates, but fine-mapping efforts remain largely constrained to imputation because sequencing costs are prohibitive. The discordant haplotype design attains excellent power specifically for the locus for which they are optimized and is therefore cost effective for DNAseq, but needs to be carried out separately for each association signal if there is more than one in a region. A noteworthy limitation of both approaches is the challenge posed by repetitive DNA sequences, with around 50% of the human genome consisting of repetitive elements (27). For

Table 3. Summary of association and eQTL findings and breast-specific regulatory annotations for 25 cFSVs

				DNAseq study ($N = 38$)			BCAC–9 GWAS meta ($N = 109,772$)				Publicly available annotations											
	Position (hg19)	Gene	Variant type	DNAseq *P* value	RNAseq *P* value (normal tissue, CASP8)	DNAseq candidate FSV	BCAC CHAVs	BCAC data type	OR and CI (BCAC–9 GWAS meta assoc)	*P* value (BCAC–9 GWAS meta assoc)	r^2 with best BCAC signal (rs1830298)	Coding annotation by dbSNP	PolyPhen score (1 = deleterious)	SIFT scores (0 = deleterious)	HMEC H3K4me1 histone modi-fications	HMEC H3K4me3 histone modi-fications	HMEC H3K27ac histone modi-fications	DNase I HS (#cell lines)	DNase Clusters in breast cell lines (0–1000)	Txn Factor BS (#factors)	Txn Factor Chip in breast cell lines (0–1,000)	eQTL in blood for CASP8 (Z-score)	eQTL in blood for CASP8 (*P* value)
rs12990906	202114624	CASP8	Intronic	8.8×10^{-16}	0.11	Y		Typed	1.03 (1.01–1.05)	9.2E−04	0.46				2	0	1						
rs10931934	202119789	CASP8	Intronic	2.1×10^{-17}	0.043	Y		Typed	1.03 (1.01–1.05)	6.1E−04	0.48				8[b]		2	2					
rs3769823	202122995	CASP8	Exonic nonsyn	8.1×10^{-16}	0.0086	Y	Y	Typed	1.05 (1.03–1.07)	1.3E−07	0.74	R (A**G**A) → K (A**A**A)	0.37	0.76	11	76[b]	174[b]	122	39 (MCF-7); 35 (HMEC); 17 (T-47D)	20	1,000 (STAT3 MCF-10A-Er-Src); 852 (c-Myc MCF-7); 267 (Pol2 MCF-10A-Er-Src)	−5.441	5E−08
rs3769821	202123430	CASP8	Intronic	2.8×10^{-18}	0.023	Y	Y	Typed	1.05 (1.03–1.07)	7.3E−08	0.66				7	86[b]	132[b]	122	39 (MCF-7); 35 (HMEC); 17 (T-47D)	7	267 (Pol2 MCF-10A-Er-Src); 67 (CTCF MCF-7);	−3.905	9E−05
rs10635401	202123717	CASP8	Intronic	1.5×10^{-16}	nd	Y	Y	Imputed	1.05 (1.03–1.07)	4.8E−06	0.74				15[b]	63[b]	70[b]	122	39 (MCF-7); 35 (HMEC); 17 (T-47D)				
rs6735656	202124502	CASP8	Intronic	6.9×10^{-13}	0.0032	Y		Typed	1.05 (1.03–1.07)	2.6E−06	0.88				10	4[b]	9[b]	8				−7.167	8E−13
rs6754084	202124997	CASP8	Intronic	2.1×10^{-13}	0.0032	Y	Y	Typed	1.05 (1.03–1.07)	2.3E−06	0.89				27[b]	3[b]	6[b]					−7.167	8E−13
rs1861270	202126615	CASP8	Intronic	4.4×10^{-15}	0.0032	Y		Typed	1.05 (1.03–1.07)	5.3E−06	0.93				12[b]	4	2	17				−7.169	8E−13
rs6751053	202127863	CASP8	Intronic	2.1×10^{-17}	0.11	Y		Typed	1.03 (1.01–1.05)	4.9E−04	0.53				4	32							
rs6435074	202127947	CASP8	Intronic	4.4×10^{-15}	0.0032	Y		Typed	1.05 (1.03–1.07)	7.7E−06	0.92				2	2	1						
rs6723097	202128618	CASP8	Intronic	2.1×10^{-17}	0.11	Y		Imputed	1.03 (1.01–1.05)	3.6E−04	0.51				3	0	2	12		1			
rs3769820	202130133	CASP8	Intronic	5.7×10^{-17}	nd	Y		Imputed	1.03 (1.01–1.05)	3.8E−04	0.52				1	0							
rs2349070	202130308	CASP8	Intronic	4.4×10^{-15}	0.0032	Y		Typed	1.05 (1.03–1.07)	3.6E−06	0.94					1							
rs10931936	202143928	CASP8	Intronic	4.4×10^{-15}	0.00036	Y	Y	Typed	1.05 (1.03–1.07)	1.0E−07	0.98					2	1	4				−6.96	3E−12
rs3769818	202151163	CASP8	Intronic	4.4×10^{-15}	0.00036	Y	Y	Typed	1.05 (1.03–1.07)	6.6E−08	0.98				3	2	2	2					
rs700635	202153225	ALS2CR12	Utr3	4.4×10^{-15}	0.00036	Y	Y	Typed	1.05 (1.03–1.07)	6.5E−08	0.98				4	1	3					−6.928	4E−12
rs6714430	202153684	ALS2CR12	Intronic	4.4×10^{-15}	0.00036	Y	Y	Typed	1.05 (1.03–1.08)	4.1E−08	0.99				1								
rs6743068	202153920	ALS2CR12	Intronic	4.4×10^{-15}	0.00036	Y	Y	Typed	1.05 (1.03–1.07)	6.5E−08	0.98				2	4						−6.915	5E−12
rs10200279	202170655	ALS2CR12	Intronic	na	na	na	Y	Imputed	1.06 (1.04–1.08)	4.0E−08	0.96						2						
rs6719014	202176024	ALS2CR12	Intronic	na	na	na	Y	Imputed	1.06 (1.04–1.08)	4.6E−08	0.98				3	1	1						
rs7582362	202176294	ALS2CR12	Intronic	na	na	na	Y	Imputed	1.06 (1.04–1.08)	4.6E−08	0.99					1	1						
rs1830298	202181247	ALS2CR12	Intronic	na	na	na	Y	Imputed	1.06 (1.04–1.08)	1.1E−09	1.00						1						
rs9677180	202184331	ALS2CR12	Intronic	na	na	na	Y	Imputed	1.05 (1.03–1.08)	5.1E−08	0.98					1							
rs2349073	202186986	ALS2CR12	Intronic	na	na	na	Y	Imputed	1.06 (1.04–1.08)	5.7E−08	0.89				1								
rs10197246	202204741	ALS2CR12	Intronic	3.1×10^{-11}[a]	0.000025	na	Y	Typed	1.06 (1.04–1.08)	1.7E−08	0.90				3	1	4			1	256 (c-Myc MCF-7)		

NOTE: BCAC indicates results from the recent Breast Cancer Association Consortium fine-mapping paper (11). SIFT is an algorithm to predict the effects of coding nonsynonymous variants on protein function. MCF-7, mammary gland, adenocarcinoma (estrogen-positive breast tumor cell line; ref. 28); MCF-10A-Er-Src, MCF-10A parent cells (mammary gland, nontumorigenic epithelial, inducible cell line), but containing ER-SRC, a derivative of the SRC kinase oncoprotein (v-SRC) that is fused to the ligand-binding domain of the estrogen receptor (estrogen-positive breast tumor–like cell line); T-47D, human ductal breast epithelial tumor (triple-negative breast tumor cell line).

Abbreviations: nd, no data (experiment not performed); na, not applicable (the base positions were not contained in the DNAseq bait set due to sequence inaccessible, hence there was no coverage for these variants); HMEC, human mammary epithelial cells ("normal" breast cell line).

[a]Association results are gained from genotyping data.

[b]Statistically significant regulatory finding (as per ENCODE "peak" tracks). Important note: not all experiments in ENCODE have been performed on all cell lines. A blank indicates no data.

targeted capture, it is not possible to design uniquely mapping capture baits for these regions, and for any HTS technology, alignment is difficult across repetitive elements. In our discordant haplotype study, for example, we were able to capture only 61% of the region at a read depth of $\geq 10\times$ (average $76\times$ depth across the full capture). In particular, we did not identify 7 SVs that were available in the BCAC imputation study. Highly repetitive sequence affects imputation studies differently. Repetitive regions are likely to harbor increased rates of misalignment, leading to nonrandom genotype errors in the reference panel, from which imputations are made. Quality scores may not easily identify such errors. An example of this in the exome design is the appearance of highly variable "promiscuous" genes within gene families with a high degree of identity at the nucleotide level. In addition, the 1000G phase III data, frequently used as a whole-genome imputation reference panel, have only an approximately 4 to $6\times$ average read depth across the majority of the noncoding regions of the genome, leading to low quality genotypes and an overabundance of homozygotes. To investigate this phenomenon, we downloaded all Caucasian 1000G whole genome .bam files that were indicated to have passed quality control. We performed a best practice haplotypeCaller GATK variant calling pipeline on these 191 genomes and inspected the data for the 7 SVs imputed in the BCAC study but not captured by baits in the current DNAseq study. As expected, the 1000G sequencing coverage for these 7 SVs was low (median DP = 5–$7\times$). Call rates ranged from 83% to 99%. Of those called, individual-level genotype qualities were relatively poor (median GQ = 15–24; GQ < 20 often considered unreliable, GQ $\geq$ 30 often used as a quality filter). Hence, both designs are affected by repetitive genome sequences, albeit in slightly different ways, adding further value to a joint-interpretation approach. We conclude that these two approaches are extremely complementary and suggest that DNAseq in a nested discordant haplotype design within larger case–control studies could play an important role in identifying comprehensive short lists for functional studies.

Disclosure of Potential Conflicts of Interest

L. Cannon-Albright has expert testimony in Myriad Genetics. No potential conflicts of interest were disclosed by the other authors.

Authors' Contributions

Conception and design: N.J. Camp, G.J. Burghel, S. Knight, J. Gertz, A. Cox
Development of methodology: N.J. Camp, S.H. Rigas, G. Wang, S. Knight, R. Abo
Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): N.J. Camp, M.A. Parry, S.H. Rigas, P.-Y. Tai, K. Berrett, I.W. Brock, B. Jones, P.S. Bernard, L. Cannon-Albright, M.W.R. Reed, J. Gertz, A. Cox
Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): N.J. Camp, W.-Y. Lin, G.J. Burghel, T.L. Mosbruger, M.A. Parry, R.G. Waller, V. Rajamanickam, R. Cosby, B. Jones, S. Knight, J. Gertz
Writing, review, and/or revision of the manuscript: N.J. Camp, G.J. Burghel, R.G. Waller, R. Cosby, , R.E. Factor, P.S. Bernard, L. Cannon-Albright, S. Knight, R. Abo, T.L. Werner, J. Gertz, A. Cox
Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): N.J. Camp, A. Bigelow, R.G. Waller, I.W. Brock, B. Jones, D. Connley, R. Sargent, L. Cannon-Albright
Study supervision: N.J. Camp, B. Jones

Acknowledgments

The authors gratefully acknowledge funding support. They thank Neil Hall, Lisa Olohan, and Xuan Liu of the Liverpool MRC Genomics Hub for DNA sequencing. At the University of Utah, they thank Derek Warner and Michael Kline of the Genomics Core for genotyping; Brian Dalley of the High Throughput Genomics Core for RNA sequencing; David Nix and Brett Milash of the Bioinformatics Shared Resource; the Biorepository and Molecular Pathology core; Roger Edwards and the Research Informatics Core; and the Women's Cancer Disease-Oriented Team at the Huntsman Cancer Institute. At the University of Sheffield, they acknowledge Helen Cramp, Sabapathy Balasubramanian, Simon S. Cross, and Sue Higham for subject recruitment and data collection. The results described are in part based upon data generated by the TCGA Research Network: http://cancergenome.nih.gov/. Finally, they thank all the participants in this multicenter study for making this research possible.

Grant Support

This work was supported by The Avon Foundation (02-2009-080 to N.J. Camp); the NCI (NCI R01 CA163353 to N.J. Camp), and a pilot grant to N.J. Camp/J. Gertz from the Cancer Control and Population Sciences and Nuclear Control Programs of the NCI Huntsman Cancer Center (P30 CA42014); The Susan G. Komen Foundation (BCTR0706911 to N.J. Camp); Yorkshire Cancer Research (YCR S305PA, S295 to A. Cox and SPP060 to S.H. Rigas); and Cancer Research UK, CRUK (C9528/A11292 to A. Cox). Ascertainment in Utah was made possible in part by the Utah Cancer Registry (UCR) and the Utah Population Database (UPDB). The UCR is funded by Contract No. HHSN261201300017I from the NCI, with additional support from the Utah State Department of Health and the University of Utah. Partial support for all datasets within the UPDB is provided by the HCI and its NCI Cancer Center Support grant, P30 CA42014.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Received July 6, 2015; revised December 22, 2015; accepted December 31, 2015; published OnlineFirst January 21, 2016.

References

1. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, et al. Defining the role of common variation in the genomic and biological architecture of adult human height. Nat Genet 2014;46:1173–86.
2. Perry JR, Day F, Elks CE, Sulem P, Thompson DJ, Ferreira T, et al. Parent-of-origin-specific allelic associations among 106 genomic loci for age at menarche. Nature 2014;514:92–7.
3. Al Olama AA, Kote-Jarai Z, Berndt SI, Conti DV, Schumacher F, Han Y, et al. A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. Nat Genet 2014;46:1103–9.
4. Nalls MA, Pankratz N, Lill CM, Do CB, Hernandez DG, Saad M, et al. Large-scale meta-analysis of genome-wide association data identifies six new risk loci for Parkinson's disease. Nat Genet 2014;46:989–93.
5. Hindorff LA, MacArthur J (European Bioinformatics Institute), Morales J (European Bioinformatics Institute), Junkins HA, Hall PN, Klemm AK, et al. A catalog of published genome-wide association studies. Available from: www.genome.gov/gwastudies. Accessed October 6, 2014.
6. Smemo S, Tena JJ, Kim KH, Gamazon ER, Sakabe NJ, Gómez-Marín C, et al. Obesity-associated variants within FTO form long-range functional connections with IRX3. Nature 2014;507:371–5.
7. Musunuru K, Strong A, Frank-Kamenetsky M, Lee NE, Ahfeldt T, Sachs KV, et al. From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. Nature 2010;466:714–9.
8. Gregory AP, Dendrou CA, Attfield KE, Haghikia A, Xifara DK, Butter F, et al. TNF receptor 1 genetic risk mirrors outcome of anti-TNF therapy in multiple sclerosis. Nature 2012;488:508–11.
9. Farh KK, Marson A, Zhu J, Kleinewietfeld M, Housley WJ, Beik S, et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature 2015;518:337–43.

10. Barrett JH, Taylor JC, Bright C, Harland M, Dunning AM, Akslen LA, et al. Fine mapping of genetic susceptibility loci for melanoma reveals a mixture of single variant and multiple variant regions. Int J Cancer 2015;136:1351–60.
11. Lin WY, Camp NJ, Ghoussaini M, Beesley J, Michailidou K, Hopper JL, et al. Identification and characterization of novel associations in the CASP8/ALS2CR12 region on chromosome 2 with breast cancer risk. Hum Mol Genet 2015;24:285–98.
12. Abo R, Knight S, Wong J, Cox A, Camp NJ. hapConstructor: automatic construction and testing of haplotypes in a Monte Carlo framework. Bioinformatics 2008;24:2105–7.
13. Camp NJ, Parry M, Knight S, Abo R, Elliott G, Rigas SH, et al. Fine-mapping CASP8 risk variants in breast cancer. Cancer Epidemiol Biomarkers Prev 2012;21:176–81.
14. Abo R, Wong J, Thomas A, Camp NJ. Haplotype association analyses in resources of mixed structure using Monte Carlo testing. BMC Bioinformatics 2010;11:592.
15. Cox A, Dunning AM, Garcia-Closas M, Balasubramanian S, Reed MW, Pooley KA, et al. A common coding variant in CASP8 is associated with breast cancer risk. Nat Genet 2007;39:352–8.
16. Turnbull C, Ahmed S, Morrison J, Pernet D, Renwick A, Maranian M, et al. Genome-wide association study identifies five new breast cancer susceptibility loci. Nat Genet 2010;42:504–7.
17. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 2011;43:491–8.
18. Nix DA, Courdy SJ, Boucher KM. Empirical methods for controlling false positives and estimating confidence in ChIP-Seq peaks. BMC Bioinformatics 2008;9:523.
19. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. Genome Biol 2014;15:550.
20. Hart SN, Therneau TM, Zhang Y, Poland GA, Kocher JP. Calculating sample size estimates for RNA sequencing data. J Comput Biol 2013;20:970–8.
21. Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, et al. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell 2012;148:84–98.
22. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature 2012;489:57–74.
23. Westra HJ, Peters MJ, Esko T, Yaghootkar H, Schurmann C, Kettunen J, et al. Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet 2013;45:1238–43.
24. Stephens PJ, Tarpey PS, Davies H, Van Loo P, Greenman C, Wedge DC, et al. The landscape of cancer genes and mutational processes in breast cancer. Nature 2012;486:400–4.
25. Lawrence MS, Stojanov P, Mermel CH, Robinson JT, Garraway LA, Golub TR, et al. Discovery and saturation analysis of cancer genes across 21 tumour types. Nature 2014;505:495–501.
26. Bernstein BE, Stamatoyannopoulos JA, Costello JF, Ren B, Milosavljevic A, Meissner A, et al. The NIH Roadmap Epigenomics Mapping Consortium. Nat Biotechnol 2010;28:1045–8.
27. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, et al. Initial sequencing and analysis of the human genome. Nature 2001;409:860–921.
28. Soule HD, Vazguez J, Long A, Albert S, Brennan M. A human cell line from a pleural effusion derived from a breast carcinoma. J Natl Cancer Inst 1973;51:1409–16.
29. Rosenbloom KR, Sloan CA, Malladi VS, Dreszer TR, Learned K, Kirkup VM, et al. ENCODE data in the UCSC Genome Browser: year 5 update. Nucleic Acids Res 2013;41(Database issue):D56–63.