An intergenic enhancer deletion in 2q35 modulates breast cancer risk by deregulating IGFBP5 expression

Asaf Wyszynski[1,2], Chi-Chen Hong[3], Kristin Lam[4], Kyriaki Michailidou[5], Christian Lytle[6], Song Yao[3], Yali Zhang[3], Manjeet K. Bolla[5], Qin Wang[5], Joe Dennis[5], John L. Hopper[7], Melissa C. Southey[8], Marjanka K Schmidt[9], Annegien Broeks[9], Kenneth Muir[10,11], Artitaya Lophatananon[10], Peter A. Fasching[12,13], Matthias W. Beckmann[12], Julian Peto[14], Isabel dos-Santos-Silva[14], Elinor J. Sawyer[15], Ian Tomlinson[16,] Barbara Burwinkel[17,18], Frederik Marme[17,19], Pascal Guénel[20,21], Thérèse Truong[20,21], Stig E. Bojesen[22,23,] Børge G. Nordestgaard[22,23], Anna González-Neira[24], Javier Benitez[25,26,] Susan L. Neuhausen[27], Hermann Brenner[28,29], Aida Karina Dieffenbach[28,29], Alfons Meindl[30], Rita K. Schmutzler[31,31,33,34], Hiltrud.Brauch[35,36], The GENICA Network[35,36,37,38,39,40,41], Heli Nevanlinna[42], Sofia Khan[42], Keitaro Matsuo[43], Hidemi Ito[44], Thilo Dörk[45], Natalia V. Bogdanova[46], Annika Lindblom[47], Sara Margolin[48], Arto Mannermaa[49,50,51], Veli-Matti Kosma[49,50,51], kConFab Investigators[52], Australian Ovarian Cancer Study Group[52,53], Anna H. Wu[54], David Van Den Berg[54], Diether Lambrechts[55,56], Hans Wildiers[57], Jenny Chang-Claude[58], Anja Rudolph[58], Paolo Radice[59], Paolo Peterlongo[60] , Fergus J. Couch[61], Janet E. Olson[62], Graham G. Giles[63,64], Roger L. Milne[63,64,] Christopher A. Haiman[54], Brian E. Henderson[54], Martine Dumont[65], Soo Hwang Teo[66,67], Tien Y. Wong[68], Vessela Kristensen[69,70,71], Wei Zheng[72], Jirong Long[72], Robert Winqvist[73], Katri Pylkäs[73], Irene L. Andrulis[74,75], Julia A. Knight[76,77], Peter Devilee[78], Caroline Seynaeve[79], Montserrat García-Closas[80,81], Jonine Figueroa[82], Daniel Klevebring[83], Kamila Czene[83], Maartje J. Hooning[84], Ans M.W. van den Ouweland[85],

Hatef Darabi[83], Xiao-Ou Shu[86], Yu-Tang Gao[87], Angela Cox[88], William Blot[86,89], Lisa B. Signorello[86,89], Mitul Shah[90], Daehee Kang[91,92,93], Ji-Yeob Choi[92,93], Mikael Hartman[94,95], Hui Miao[94], Ute Hamann[96], Anna Jakubowska[97], Jan Lubinski[97], Suleeporn Sangrajrang[98], James McKay[99], Amanda E. Toland[100], Drakoulis Yannoukakos[101], Chen-Yang Shen[102,103,104], Pei-Ei Wu[102,103], Anthony Swerdlow[80,105], Nick Orr[106], Jacques Simard[65], Paul D.P. Pharoah[5,90], Alison M. Dunning[90], Georgia Chenevix-Trench[107], Per Hall[83], Elisa Bandera[108], Chris Amos[109], Christine Ambrosone[3], Douglas F Easton[110,111], Michael D. Cole[2,112 *]

*To whom correspondence should be addressed:

Michael D. Cole, Geisel School of Medicine at Dartmouth, Hanover, NH USA 03755,

Email: Michael.D.Cole@dartmouth.edu; Tel: (603) 653-9975; Fax: (603) 653-9952

1. Department of Community and Family Medicine, Geisel School of Medicine at Dartmouth, Hanover, NH 03755 USA
2. Department of Pharmacology & Toxicology, Geisel School of Medicine at Dartmouth, Hanover, NH 03755 USA
3. Department of Cancer Prevention and Control, Roswell Park Cancer Institute, Buffalo, NY USA
4. Dartmouth College, Hanover, NH 03755 USA
5. Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, CB1 8RN, UK
6. Molecular Biology Core Facility, Dartmouth College, Hanover, NH 03755 USA

7. Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Victoria 3010, Australia
8. Department of Pathology, The University of Melbourne, Melbourne, Victoria 3010, Australia
9. Netherlands Cancer Institute, Antoni van Leeuwenhoek hospital, 1066 CX Amsterdam, The Netherlands
10. Division of Health Sciences, Warwick Medical school, Warwick University, Coventry, CV4 7AL, UK
11. Institute of Population Health, University of Manchester, Manchester, M13 9PL, UK
12. University Breast Center Franconia, Department of Gynecology and Obstetrics, University Hospital Erlangen, Friedrich-Alexander University Erlangen-Nuremberg, Comprehensive Cancer Center Erlangen-EMN, 91054 Erlangen, Germany
13. David Geffen School of Medicine, Department of Medicine Division of Hematology and Oncology, University of California at Los Angeles, Los Angeles, CA 90095, USA
14. Department of Non-Communicable Disease Epidemiology, London School of Hygiene and Tropical Medicine, London, WC1E 7HT, UK
15. Research Oncology, Division of Cancer Studies, King's College London, Guy's Hospital, London, SE1 9RT, UK
16. Wellcome Trust Centre for Human Genetics and Oxford Biomedical Research Centre, University of Oxford, OX3 7BN, UK

17. Department of Obstetrics and Gynecology, University of Heidelberg, 69120 Heidelberg, Germany
18. Molecular Epidemiology Group, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany
19. National Center for Tumor Diseases, University of Heidelberg, 69120 Heidelberg, Germany
20. Inserm (National Institute of Health and Medical Research), CESP (Center for Research in Epidemiology and Population Health), U1018, Environmental Epidemiology of Cancer, 94807 Villejuif, France
21. University Paris-Sud, UMRS 1018, 94807 Villejuif, France
22. Copenhagen General Population Study, Herlev Hospital, Copenhagen University Hospital, 2730 Herlev, Denmark
23. Department of Clinical Biochemistry, Herlev Hospital, Copenhagen University Hospital, 2730 Herlev, Denmark
24. Human Genotyping-CEGEN Unit, Human Cancer Genetics Program, Spanish National Cancer Research Centre (CNIO), 28029 Madrid, Spain
25. Human Genetics Group, Human Cancer Genetics Program, Spanish National Cancer Research Centre (CNIO), 28029 Madrid, Spain
26. Centro de Investigación en Red de Enfermedades Raras (CIBERER), 46010 Valencia, Spain
27. Beckman Research Institute of City of Hope, Duarte, CA 91010, USA
28. Division of Clinical Epidemiology and Aging Research, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany

29. German Cancer Consortium (DKTK), 69120 Heidelberg, Germany
30. Division of Gynaecology and Obstetrics, Technische Universität München, 81675 Munich, Germany
31. Division of Molecular Gyneco-Oncology, Department of Gynaecology and Obstetrics, University Hospital of Cologne, 50931 Cologne, Germany
32. Center of Familial Breast and Ovarian Cancer, University Hospital of Cologne, 50931 Cologne, Germany
33. Center for Integrated Oncology (CIO), University Hospital of Cologne, 50931 Cologne, Germany
34. Center for Molecular Medicine Cologne (CMMC), University of Cologne, 50931 Cologne, Germany
35. Dr. Margarete Fischer-Bosch-Institute of Clinical Pharmacology, 70376 Stuttgart, Germany
36. University of Tübingen, 72074 Tübingen, Germany
37. Institute for Prevention and Occupational Medicine of the German Social Accident Insurance, Institute of the Ruhr University Bochum (IPA), 44789 Bochum, Germany
38. Department of Internal Medicine, Evangelische Kliniken Bonn gGmbH, Johanniter Krankenhaus, 53113 Bonn, Germany
39. Molecular Genetics of Breast Cancer, Deutsches Krebsforschungszentrum (DKFZ), 69120 Heidelberg, Germany
40. Institute of Pathology, Medical Faculty of the University of Bonn, 53127 Bonn, Germany

41. Institute of Occupational Medicine and Maritime Medicine, University Medical Center Hamburg-Eppendorf, 20246 Hamburg, Germany
42. Department of Obstetrics and Gynecology, University of Helsinki and Helsinki University Central Hospital, Helsinki, FI-00029 HUS, Finland
43. Department of Preventive Medicine, Kyushu University Faculty of Medical Sciences, Fukuoka, Japan
44. Division of Epidemiology and Prevention, Aichi Cancer Center Research Institute, Nagoya, Aichi, 464-8681, Japan
45. Department of Obstetrics and Gynaecology, Hannover Medical School, 30625 Hannover, Germany
46. Department of Radiation Oncology, Hannover Medical School, 30625 Hannover, Germany
47. Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm SE-17177, Sweden
48. Department of Oncology - Pathology, Karolinska Institutet, Stockholm SE-17177, Sweden
49. School of Medicine, Institute of Clinical Medicine, Pathology and Forensic Medicine, University of Eastern Finland, FI-70211 Kuopio, Finland
50. Cancer Center of Eastern Finland, University of Eastern Finland, FI-70211 Kuopio, Finland
51. Imaging Center, Department of Clinical Pathology, Kuopio University Hospital, 70210 Kuopio, Finland
52. Peter MacCallum Cancer Center, Melbourne, Victoria 3002, Australia

53. QIMR Berghofer Medical Research Institute, Brisbane, QLD 4006, Australia
54. Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA
55. Vesalius Research Center (VRC), VIB, 3000 Leuven, Belgium
56. Laboratory for Translational Genetics, Department of Oncology, University of Leuven, 3000 Leuven, Belgium
57. Multidisciplinary Breast Center, Department of General Medical Oncology, University Hospitals Leuven, B-3000 Leuven, Belgium
58. Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany
59. Unit of Molecular Bases of Genetic Risk and Genetic Testing, Department of Preventive and Predictive Medicine, Fondazione IRCCS Istituto Nazionale dei Tumori (INT), 20133 Milan, Italy
60. IFOM, Fondazione Istituto FIRC di Oncologia Molecolare, 20139 Milan, Italy
61. Department of Laboratory Medicine and Pathology, Mayo Clinic, Rochester, MN 55905, USA
62. Department of Health Sciences Research, Mayo Clinic, Rochester, MN 55905, USA
63. Cancer Epidemiology Centre, Cancer Council Victoria, Melbourne, Victoria 3004, Australia
64. Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, Melbourne, Victoria 3010, Australia

65. Centre Hospitalier Universitaire de Québec Research Center and Laval University, QC, G1V 4G2, Canada
66. Cancer Research Initiatives Foundation, Sime Darby Medical Centre, 47500 Subang Jaya, Selangor, Malaysia
67. Breast Cancer Research Unit, University Malaya Cancer Research Institute, University Malaya Medical Centre (UMMC), 59100 Kuala Lumpur, Malaysia
68. Singapore Eye Research Institute, National University of Singapore, Singapore 168751
69. Department of Genetics, Institute for Cancer Research, Oslo University Hospital, Radiumhospitalet, N-0310 Oslo, Norway
70. Institute of Clinical Medicine, University of Oslo (UiO), 0450 Oslo, Norway
71. Department of Clinical Molecular Biology (EpiGen), University of Oslo (UiO), 0450 Oslo, Norway
72. Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, TN 37203, USA
73. Laboratory of Cancer Genetics and Tumor Biology, Department of Clinical Chemistry and Biocenter Oulu, University of Oulu, NordLab Oulu/Oulu University Hospital, FI-90220 Oulu, Finland
74. Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, Toronto, ON, M5G 1X5, Canada
75. Department of Molecular Genetics, University of Toronto, Toronto, ON, M5S 1A8, Canada

76. Prosserman Centre for Health Research, Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, Toronto, ON, M5G 1X5, Canada
77. Division of Epidemiology, Dalla Lana School of Public Health, University of Toronto, Toronto, ON, M5S 1A8, Canada
78. Department of Human Genetics & Department of Pathology, Leiden University Medical Center, 2333 ZC Leiden, The Netherlands
79. Family Cancer Clinic, Department of Medical Oncology, Erasmus MC-Daniel den Hoed Cancer Center, 3075 EA Rotterdam, The Netherlands
80. Division of Genetics and Epidemiology, Institute of Cancer Research, Sutton, SM2 5NG, UK
81. Breakthrough Breast Cancer Research Centre, Division of Breast Cancer Research, The Institute of Cancer Research, London, SW3 6JB, UK
82. Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, MD 20850, USA
83. Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm SE-17177, Sweden
84. Department of Medical Oncology, Erasmus University Medical Center, 3075 EA Rotterdam, The Netherlands
85. Department of Clinical Genetics, Erasmus University Medical Center, 3075 EA Rotterdam, The Netherlands
86. Division of Epidemiology, Department of Medicine, Vanderbilt Epidemiology Center, Vanderbilt-Ingram Cancer Center, Vanderbilt University School of Medicine, Nashville, TN 37203, USA

87. Department of Epidemiology, Shanghai Cancer Institute, Xuhui, Shanghai, China
88. CRUK/YCR Sheffield Cancer Research Centre, Department of Oncology, University of Sheffield, Sheffield, S10 2RX, UK
89. International Epidemiology Institute, Rockville, MD 20850, USA
90. Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, CB1 8RN, UK
91. Department of Preventive Medicine, Seoul National University College of Medicine, Seoul 110-799, Korea
92. Department of Biomedical Sciences, Seoul National University Graduate School, Seoul 151-742, Korea
93. Cancer Research Institute, Seoul National University College of Medicine, Seoul 110-799, Korea
94. Saw Swee Hock School of Public Health, National University of Singapore and National University Health System, Singapore 117597
95. Department of Surgery, Yong Loo Lin School of Medicine, National University of Singapore and National University Health System, Singapore 117597
96. Molecular Genetics of Breast Cancer, German Cancer Research Center (DKFZ), 69120 Heidelberg, Germany
97. Department of Genetics and Pathology, Pomeranian Medical University, 70-115 Szczecin, Poland.
98. National Cancer Institute, Bangkok 10400, Thailand
99. International Agency for Research on Cancer, 69372 Lyon, CEDEX 08, France

100. Department of Molecular Virology, Immunology and Medical Genetics, Comprehensive Cancer Center, The Ohio State University, Columbus, OH 43210, USA
101. Molecular Diagnostics Laboratory, IRRP, National Centre for Scientific Research "Demokritos", Aghia Paraskevi Attikis, 153 10 Athens, Greece
102. Taiwan Biobank, Institute of Biomedical Sciences, Academia Sinica, Taipei 115, Taiwan
103. Institute of Biomedical Sciences, Academia Sinica, Taipei 115, Taiwan
104. School of Public Health, China Medical University, Taichung 404, Taiwan
105. Division of Breast Cancer Research, Institute of Cancer Research, Sutton, SM2 5NG, UK
106. Breakthrough Breast Cancer Research Centre, The Institute of Cancer Research, London, SW3 6JB, UK
107. Department of Genetics, QIMR Berghofer Medical Research Institute, Brisbane, QLD 4006, Australia
108. Rutgers Cancer Institute of New Jersey, New Brunswick, NJ 08901 USA
109. Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth, Hanover, NH 03755 USA
110. Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK
111. Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, UK

112. Department of Genetics, Geisel School of Medicine at Dartmouth, Hanover, NH 03755 USA

Abstract

Breast cancer is the most diagnosed malignancy and the second leading cause of cancer mortality in females (1). Previous association studies have identified variants on 2q35 associated with the risk of breast cancer (2-5). To identify functional susceptibility loci for breast cancer, we interrogated the 2q35 gene desert for chromatin architecture and functional variation correlated with gene expression. We report an intergenic enhancer copy number variation (enCNV; deletion) located approximately 400Kb upstream to *IGFBP5*, which overlaps an intergenic ERα-bound enhancer that loops to the *IGFBP5* promoter. The enCNV is correlated with modified ERα binding and monoallelic-repression of *IGFBP5* following estrogen treatment. We investigated the association of enCNV genotype with breast cancer in 1,182 cases and 1,362 controls, and replicate our findings in an independent set of 62,533 cases and 60,966 controls from 41 case control studies and 11 GWAS. We report a dose-dependent inverse association of 2q35 enCNV genotype (percopy OR=0.68 95%CI 0.55‑0.83, P=0.0002; replication OR=0.77 95%CI 0.73-0.82, $P=2.1x10^{-19}$) and identify 13 additional linked variants ($r^2>0.8$) in the 20Kb linkage block containing the enCNV ($P=3.2x10^{-15}$ - $5.6x10^{-17}$). These associations were independent of previously reported 2q35 variants, rs13387042 and rs16857609, and were stronger for ER-positive than ER-negative disease. Together, these results suggest that 2q35 breast cancer risk loci may be mediating their effect through *IGFBP5*.

Introduction

The 2q35 risk locus falls within a 400Kb gene desert bounded by genes *TNP1* (MIM: 190231) and *DIRC3* (MIM: 608262), nearby two members of the insulin growth factor binding protein family, *IGFBP5 (MIM: 146734)* and *IGFBP2 (MIM: 146731).* *IGFBP5* plays a critical role in mammary development (6, 7) and has been consistently implicated in tumorigenesis (6-10).The neighboring intergenic region contains the previously identified breast cancer (MIM: 114480) risk loci, rs13387042 (3) (Genbank: NC_000002 g.217041109A>G), rs16857609 (2)(Genbank: NC_000002 g.217431785C>T), and rs4442975(5) (Genbank: NC_000002 g.217920769G>T) as well as numerous intergenic enhancers, of which many whose function remains elusive.

We sought to identify intergenic variation that may affect the estrogen-mediated transcriptional regulation *IGFBP5* and to contribute to the understanding of functional chromatin architecture at the 2q35 risk locus.

Results

To evaluate the possibility that *IGFBP5* transcription is regulated by a distal enhancer within the 2q35 gene desert, we investigated the chromatin interaction profile across the 2q35 gene desert with the *IGFBP5* promoter using chromosome conformation capture (3C) (11) in the MCF7 breast cancer cell line. Results of this interaction analysis indicated strong physical proximity of the *IGFBP5* promoter with a region containing an estrogen receptor (ERα)-bound enhancer element approximately 400Kb telomeric to the *IGFBP5* promoter (**Figure 1**). Sequence analysis of this intergenic looping enhancer revealed a 1.3 Kb copy number variation (CNV; deletion) spanning the enhancer in MCF7 cells (**Figure S1**); however, the proximal estrogen response element (ERE) was not deleted (**Figure 2A**). We examined the implications of this enCNV on ERα binding activity using chromatin immunoprecipitation coupled with allele-specific qPCR (ChIP-qPCR). Our data revealed enhanced binding activity on the variant allele ($P<0.004$; **Figure 2B**), both before and after treatment with estrogen.

We hypothesized that differential allelic-binding of ERα at the 2q35 enCNV would affect allele-specific *IGFBP5* transcription in response to estrogen signaling. We investigated the effect of the polymorphic enhancer on *IGFBP5* expression by tracking a heterozygous *IGFBP5* intronic SNP (rs7565131; Genbank: NM_000599 c.338A>C) as a marker of allele specific expression (**Figure 2C**). Prior to estrogen treatment, MCF7 cells robustly express *IGFBP5*, although a majority (>95%) of expression is from the A-allele. Following treatment with low dose estrogen, the abundance of *IGFBP5* nuclear RNA (rs7565131-A) is markedly reduced at 1 hour, relative to vehicle treated cells ($P=0.027$).

This pattern of monoallelic repression is sustained at 24 hours of exposure to estrogen (P=0.014).

To resolve the question of ERα binding at this site being repressive of *IGFBP5-A* versus merely upregulating *IGFBP5-C*, we utilized a transactivator-fused nuclease-defective CRISPR system (13) to activate specific genomic sites in 2q35. We hypothesized that if ERα binding is repressive at this locus under normal conditions, when targeting a definitive transactivator molecule to this site, we would see the inverse transcriptional response (i.e. we expect to see an increase in *IGFBP5-A* relative to *IGFBP5-C*). Targeting of this construct to the IGFBP5 promoter showed no significant change in allelic balance (Figure S2; P=0.52 and 0.91). Targeting to the ERE at the 2q35 enCNV shows a significant increase in *IGFBP5-A* expression (Figure S2; P=0.004). Given that the activator *increases* expression of *IGFBP5-A* and, conversely, E2-bound ERα acting here as a repressor *decreases* expression of the same *IGFBP5-A* allele, we conclude that in MCF7s the 2q35 enCNV variant allele is in *cis* with *IGFBP5-A*. Additionally, these findings confirm our assertion that ERα binding at this distal enhancer is repressive of *IGFBP5* expression and suggests a functional mechanism for estrogen-induced regulation of *IGFBP5* transcription through this enhancer.

To investigate the hypothesis that variants, which influence *IGFBP5* expression, may be associated with breast cancer risk, we examined the relationship of the 2q35 enCNV with breast cancer in the Women's Circle of Health Study (WCHS) (**Table S1**). We identified 2,134 homozygous wildtype, 368 heterozygous, and 42 homozygous deleted (variant) individuals with an overall genotyping rate of 92%. We observed an inverse association between the 2q35 enCNV and breast cancer risk overall (per copy

OR=0.68 95%CI 0.55 - 0.83, P=0.0002; **Table S2**). The observed association was dose-dependent based on number of deleted alleles in both European American (EA) (P=0.03) and African American (AA) women (P=0.004), with homozygous deletion carriers having approximately 80% decreased breast cancer risk (OR=0.22 95%CI 0.09-0.52, P=0.0005; **Table S2**). The association was consistent in both pre and post-menopausal women combined, however a stronger effect was observed in pre-menopausal women (pre-menopausal per copy OR=0.60 95%CI 0.45-0.80, P=0.001; post-menopausal per copy OR=0.72 95%CI 0.53-0.97, P=0.03; **Table S2**). Among cases with available ER status (74.8%), the protective effect was confined to ER-positive tumors among all women combined (per copy OR=0.74 95%CI 0.58-0.96, P=0.02; **Table S3**).

To evaluate our association results in a larger, independent population, we replicated our findings in data from 46,785 cases and 42,892 controls from 41 case-control studies genotyped with a custom array, participating in the Breast Cancer Association Consortium (iCOGS; http://ccge.medschl.cam.ac.uk/research/consortia/icogs/)(2), together with data from 11 breast cancer GWAS, comprising 15,748 cases and 18,084 controls (2, 26) (http://gameon.dfci.harvard.edu/gameon/). All studies were of predominantly European origin and the 2q35 enCNV was not polymorphic in Asian populations in BCAC or 1000 genomes. The 2q35 enCNV was not genotyped on the iCOGS array or in any of the GWAS, but the variant is present in the 1000 genomes dataset (http://www.1000genomes.org/). We therefore derived imputed genotypes for all variants across a 1Mb interval (Chr 2: 217,731,785-218,796,508; hg19) that encompassed the 2q35 enCNV together with the flanking LD blocks containing the previously reported

2q35 susceptibility loci, rs13387042/rs4442975 and rs16857609. The 2q35 enCNV was reliably imputed in iCOGS (mean r^2=0.74) and in eight of the GWAS (r^2=0.54 to 0.73). The 2q35 enCNV was similarly associated with a reduced breast cancer risk (per copy OR=0.78 95%CI 0.74-0.84, P=6.9x10^{-16} in iCOGS; P=2.1x10^{-19}in iCOGS+GWAS combined). There was weak evidence for heterogeneity (I^2=29.29, *P*=0.04; **Figure S3**) largely driven by one study and the association remained highly significant after removing this study (OR=0.78 95%CI 0.73-0.83, P=4.1x10^{-16}). The OR for homozygous carriers of the deletion (OR=0.88 95%CI 0.56-1.38) did not differ significantly from that in heterozygous carriers (OR=0.77 95%CI 0.72-0.82), but a log-additive model could not be rejected. The association was stronger for ER-positive (OR =0.77 95%CI 0.71-0.82, P=3.1x10^{-13}) than ER-negative disease (OR=0.90 95%CI 0.80-1.01, P=0.09; P-diff=0.0079; **Table S4**), consistent with the effect observed in our initial study and previously for 2q35 loci.

The 2q35 enCNV lies in a linkage disequilibrium (LD) block of ~20Kb and strong sites of recombination separate it from the LD blocks containing the previously reported 2q35 risk loci, rs13387042/rs4442975, rs16857609; the 2q35 enCNV is uncorrelated with either locus (r^2<0.01) (**Figure 3**). In multiple regression analysis based on the iCOGS data, all three loci remain highly significantly associated with disease (**Table S5**). Only one SNP in the LD block containing the 2q35 enCNV, rs16856925 (Genbank: NC_000002 g.217096609A>G), was genotyped on the iCOGS array. This SNP was highly correlated with the 2q35 enCNV (r^2=0.90) and hence largely determined the imputed genotypes; rs16856925 was slightly more strongly associated with disease than the 2q35 enCNV (iCOGS *P*=3.7x10^{-16}; combined *P*=1.2x10^{-20}; **Figure S4 and**

Table S4). The most strongly associated variant in this block was rs34005590 (Genbank: NC_000002 g.217098337C>A; r^2=0.93; iCOGS *P*=5.6x10^{-17}; iCOGS+GWAS combined P=7.4x10^{-22}; **Figure 3, Table S4**). Fourteen variants in this block, including rs16856925 and 2q35 enCNV, were correlated with rs34005590 at r^2>0.8; however, none of these variants could be excluded as being causal at a likelihood ratio of 100:1(27). In conditional analyses, no additional SNPs were associated with disease after adjustment for rs34005590, 2q35 enCNV, or rs16856925; thus, the association results are consistent with a single causal variant within the 20Kb LD block containing the 2q35 enCNV.

Discussion

The understanding of factors affecting breast cancer risk has grown exponentially in recent years. *IGFBP5* and 2q35 have both been consistently implicated in cancer, though little was known about the nature of their interaction. Molecular studies of *IGFBP5* have revealed its essential role in normal mammary epithelial development (6, 7, 28, 29), contributing to the documented involvement of the IGF signaling axis in mammary density as a risk factor for breast cancer (30-32). A recent contemporaneous study describes a neighboring 2q35 breast cancer-associated variant nearby the locus we describe. Their intriguing and independent findings implicate an intergenic SNP in modifying expression of *IGFBP5*, however, their work focused on a narrow genomic region investigated in high resolution on the iCOGS array and excludes our reported risk locus (5). Here we shed light on the complexity of *IGFBP5* transcriptional control by estrogen and identify a polymorphic regulatory region ~400Kb upstream that

differentially regulates *IGFBP5* upon exposure to estrogen. Further, we utilize a transactivator-fused CRISPR system to evaluate 2q35 allele linkage in MCF7 cell line and confirm the repressive nature of ERα binding at the 2q35 enCNV. Targeting the wildtype sequence of the non-deleted enCNV allele results in no significant shift in the allelic balance. When considering the allelic preference of ERα binding at the 2q35 enCNV, these data suggest a model where the wildtype allele performs as a less efficient regulator of *IGFBP5* regulation, and the bulk of expression comes from the efficiently regulated *IGFBP5-A* allele. Our findings are consistent with the current understanding of chromatin architecture (33-35) and suggest that previously under-studied (36) larger CNVs, particularly in intergenic enhancers, may play a striking role in the etiology of disease.

Materials and Methods

Cell Culture and treatments

Cells were maintained according to manufacturer recommendation (ATCC). Briefly, MCF7 cells (passage 14-28) were maintained in complete DMEM (10%FBS, 5mg/mL insulin, 0.4% penicillin-streptomycin) at 37°C in humidified chamber with 5% CO_2. Cells were hormone starved prior to treatment for at least 48 hours in phenol red free media supplemented with 10% charcoal/dextran stripped FBS (Life Technologies, Carlsbad, CA). Cells were treated with vehicle (DMSO) or 17β-estradiol (10nM; Sigma-Aldrich, St. Louis, MO) for the indicated duration.

Chromatin Conformation Capture

Chromatin conformation capture was conducted as previously published with subtle modifications (11). Briefly, nuclei from 5 x 10^6 cells were isolated and crosslinked in 1% formaldehyde for 10 minutes at room temperature. Washed nuclei were resuspended in 1x restriction enzyme buffer and digested overnight with 400U of restriction enzyme (HindIII; New England Biolabs Inc., Ipswich, MA). Digested nuclei were disrupted and diluted to a final volume of 8 mL for ligation for 2-4 hours at 16°C. Ligated DNA was purified and resuspended in TE (Invitrogen Inc., Carlsbad, CA). Site-specific interactions with the "anchor" region (*IGFBP5* promoter) were assayed by realtime quantitative PCR with 100ng 3C DNA per reaction and normalized to a 3C positive control library prepared as previously described (11). All experiments were conducted in biological triplicates and qPCR reactions as technical duplicates. BACs (3096A13, 2565O2, 2505P8; Invitrogen Inc., Carlsbad, CA) were grown according to

manufacturer recommendations and purified (PureLink HiPure; Invitrogen Inc., Carlsbad, CA). Primer sequences are listed in the supplementary data.

Chromatin Immunoprecipitation

Experiments were performed as previously described according to manufacturer recommendation (Upstate Biotechnologies/EMD Millipore, Billerica, MA). Briefly, vehicle or estrogen (10nM in DMSO, 45 minutes) treated cells were crosslinked with 1% formaldehyde and washed. Cells were lysed and chromatin/protein complexes sheared by sonication. IgG or ERα (HC-20; Santa Cruz Biotechnology Inc., Dallas, TX) was immunoprecipitated overnight and complexes collected with protein A/G beads for one hour (Dynabeads; Invitrogen Inc., Carlsbad, CA). Eluted DNAs were decrosslinked and purified by ethanol precipitation. Experiments were conducted in biological triplicate and qPCR reactions in technical duplicate. Binding activity was calculated relative to input. Primer sequences are listed in the supplementary data.

Expression analysis

Nuclei from estrogen (10nM, DMSO) or vehicle (DMSO) treated cells were isolated (Nuclear extraction buffer: 100mM Tris, 100mM NaCl, 0.5% NP-40) and nuclear-enriched RNA was extracted with Trizol (Invitrogen Inc., Carlsbad, CA). Residual DNA contaminants were removed by DNAse treatment (Promega Inc., Madison, WI) and cDNA was synthesized per manufacturers recommendation (FirstStrand Synthesis Kit; Invitrogen Inc., Carlsbad, CA). Expression of total *IGFBP5* was quantified by RT-qPCR with primers targeting the 3' UTR and normalized to actin (Integrated DNA Technologies, Coralville, IA). Reactions performed at 95°C, 3min; and cycled 40x at 95°C, 15s; 61°C, 15s; 72°C, 15s, followed by melting curve analysis

(CFX96, Bio-Rad Laboratories, Hercules, CA). Allelic expression of IGFBP5 was determined by 20-cycle pre-amplification of a 700bp fragment surrounding heterozygous intronic rs7565131 A/C (95°C, 5min; cycled 20x 95°C, 30s; 61°C, 30s; 72°C, 30s, followed by a 10 min extension at 72°C). Amplified sequences were column purified (QIAamp PCR cleanup kit, Qiagen Inc., Valencia, CA) and detection was conducted using a modified RT-MAMA-qPCR with allele specific primers (12). All experiments were conducted in biological triplicates and qPCR reactions as technical duplicates. Primer sequences are listed in the supplementary data.

CRISPR-aided analysis of allele linkage

Briefly, MCF7 cells were grown in complete media and transfected with pAC154-dual-dCas9VP160-sgExpression (13) (Addgene, Cambridge, MA) containing appropriate guide RNAs by nucleofection, per manufacturer's recommendation (Nucleofector, Lonza Ltd, Basel, Switzerland). Constructs were validated by sequencing at our core facility. Guide RNAs targeted either *IGFBP5* promoter sites (Promoter site 1: CTACAAACTGGCTGGCAGCC; Promoter site 2: GTTTGTACTGCAAAGCTCCT), the ERE nearest the 2q35 enCNV (ERE: CTGAACTGTCCTCAAGTTCT), or the wildtype sequence within the deleted region (enCNV site 1: TAGATGGATCCCTCAGAAAT; enCNV site 2: CCATAGACAGGTCTTTTTTG). RNA was extracted for expression analysis as described above. Data represent technical and biological duplicates.

Women's Circle of Health Study

Study Population

The study was conducted using samples and data from the Women's Circle of Health Study (WCHS), a case-control study designed to examine risk factors for early/aggressive breast cancer among African American (AA) women compared to American women of European descent (EA). Details of the study design, inclusion criteria, and collection of survey data and biospecimens have been previously described (14, 15). Briefly, incident breast cancer cases were identified in four boroughs of metropolitan New York City using hospital-based case ascertainment, and in seven counties in New Jersey (NJ) using population-based case ascertainment through the NJ State Cancer Registry, a participant of the National Cancer Institute's Surveillance, Epidemiology, and End Results (SEER) program. Cases were women recently diagnosed with primary, histologically confirmed breast cancer with no previous history of cancer except for non-melanoma skin cancer who self-identified as AA or EA, 20-75 years of age, and were English speaking. Controls were frequency matched to cases by self-reported race and 5-year age groups and were recruited from the same target population as cases by using random digit dialing in the same residential area as cases. AA controls in NJ were supplemented by community recruitment efforts to assemble a control sample more representative of the general population (16). A total of 1,369 EAs (680 cases, 689 controls) and 1,403 AAs (628 cases, 775 controls) women were included in the study. The study was approved by the institutional review boards at Roswell Park Cancer Institute (RPCI), the Cancer Institute of New Jersey (CINJ), Mount Sinai School of Medicine (MSSM; now the Icahn School of Medicine at Mount Sinai), and all participating hospitals in New York.

Survey Data, DNA Collection, and Genotyping

Detailed survey data were collected by in-person interviews and included demographic and lifestyle information, family history of cancer, and medical history. Anthropometric measurements and biospecimen collections were obtained by trained interviewers. Pathology data were collected and abstracted by trained study staff from patient medical records and included information on tumor grade and stage, and ER status.

Genomic DNA for study participants was initially extracted from blood samples using the using the FlexiGene^{TM} DNA isolation kits (Qiagen Inc., Valencia, CA) and subsequently from Oragene^{TM} kits following the manufacturer's protocols, with the majority of DNA samples derived from saliva samples collected using Oragene^{TM} kits (DNA Genotek Inc., Kanata, Ontario, Canada). Genomic DNA was evaluated and quantitated by Nanodrop UV-spectrometer (Thermo Fisher Scientific Inc., Wilmington, DE) and PicoGreen-based fluorometric assay (Molecular Probes, Invitrogen Inc., Carlsbad, CA), and stored at -80°C until analysis.

Of 2,772 blinded samples initially included in the study, 228 samples could not be amplified leaving a total N=2,544 (EA: 613 cases, 630 controls; AA: 569 cases, 732 controls) in the study. Blinded samples were genotyped by a custom designed semi-automated multiplex fluorescent-coupled PCR in 96-well format followed by fragment length analysis. PCR reaction conditions were conducted per manufacturer recommendations (HotstarTaq Plus MasterMix, Qiagen Inc., Valencia, CA; 10ng DNA, initial activation of 95°C, 5min; and cycled 30x at 95°C, 30s; 57°C, 90s; 72°C, 30s, followed by a final 10min extension at 72°C). Amplified samples were diluted 4x and loaded for FLA by the Molecular Biology Core Facility at Dartmouth College. Genotypes

were assigned with a peak-calling algorithm in a 4bp window surrounding the expected amplicon size utilizing GeneMapper 4.0 software (Applied Biosystems). Briefly, calls were made by peak calling within 4bp bins centered on predicted sizes of 152 and 292bp. A threshold of 1,000 RFU was used to eliminate rare instances of signal bleed from neighboring overloaded wells (due to initial DNA concentration inconsistencies). Infrequent size calling software abnormalities were resolved manually using the same criteria as above. Quality control was conducted by secondary FLA of entire plates (N=4 x 96-well) and randomly selected individual samples (n=85).

To account for population admixture in the analysis, all samples were also genotyped at the Genomics Core Facility at Roswell Park Cancer Institute using the Illumina GoldenGate Assay (Illumina Inc., San Diego, CA) for a panel of 100 ancestry informative markers (AIMs) that were previously validated in the Black Women's Health Study Ruiz-Narváez, Rosenberg, Wise, Reich and Palmer (17). As a quality control measure, five percent duplicates and two sets of in-house trio samples were included across all plates. Proportions of European and African ancestry for each woman were computed using the Bayesian Markov Chain Monte Carlo clustering algorithm implemented in STRUCTURE (18). Since the sum of two ancestral proportions in each individual is always one, we used only the proportion of European Ancestry in all analyses.

Statistical Analysis

Continuous and categorical descriptive variables were compared between cases and controls using t-tests and chi-square tests for proportion, respectively. Odd ratios (OR) and 95% confidence intervals (CIs) for associations between 2q35 enCNV

genotype and breast cancer risk were estimated using unconditional logistic regression among all women, and stratified by self-reported race. Additional analyses were conducted to examine associations by menopausal and ER status. All analyses were adjusted for age, proportion of European ancestry, attained education, family history of breast cancer, smoking status, parity, use of hormone replacement therapy use, and study site (New York, New Jersey). Women with missing covariate data on smoking history (n=1), use of hormone replacement therapy (n=3), and family history of breast cancer (n=11), were considered to be non-smokers, non-users of hormone replacement therapy, and not to have a family history of breast cancer, respectively. For 4 women without ancestry data, race-specific median values for proportion of European ancestry were used. For analyses with pre- and post-menopausal women combined, menopausal status was also included in the model. For analyses combining EA and AA women together, self-reported race was also included in the model in addition to proportion of European ancestry estimates. Co-dominant models were analyzed and additive genotyping coding based on the number of rare alleles was used as an ordinal variable to determine P-values associated with each copy of the variant allele (p test for linear trend). Case-case unconditional logistic regression analysis was also performed to examine associations between 2q35 enCNV genotype and odds of being diagnosed with ER-negative versus ER-positive tumors. All analyses were conducted using SAS V9.3 (SAS Institute, Cary, CA). All tests were two-sided and considered statistically significant at P=0.05.

Breast Cancer Association Consortium

Genotype data for replication were derived from 11 breast cancer GWAS based on populations of European ancestry, together with 41 additional case-control studies

from populations of European ancestry participating in the Breast Cancer Association Consortium(2). The 11 GWAS were genotyped with using a variety of different platforms, while the 41 additional case-control studies were genotyped using a custom array (iCOGS). After quality control exclusions, data were available for 15,748 cases and 18,084 controls from the GWAS and 46,785 cases and 42,882 controls genotyped using the iCOGS array (after excluding samples overlapping with any GWAS; see Michailidou, 2013 for details). All studies were approved by the relevant local ethics review committee and subjects gave informed consent.

The GWAS genotype data were used to estimate genotypes for other common variants across the region in the study subjects by imputation, with IMPUTE v.2.2 (19) and the March 2012 release of the 1000 Genomes Project as reference panel, after prephasing using SHAPEIT (20) with the exception of three GWAS - BCFR, BPC3 and TNBCC - for which imputation was performed using MACH (21) and Minimac (22). Per-allele odds ratios (ORs) and standard errors for individual studies were generated using SNPTEST (23) and ProbABEL (24). For the iCOGS samples the imputation was performed in one step without pre-phasing using IMPUTE.v2 and the March 2012 release of the 1000 genomes as reference, analysis for the iCOGS samples was done using logistic regression in R. Estimated ORs for the combined analysis were generated using a fixed-effect meta-analysis adjusting for genomic control, using METAL (25). Data for SNPs with an imputation accuracy r2>0.3 in a given study were included in the combined analysis. For the combined analysis of the GWAS and iCOGS, we reanalyzed the iCOGS data to remove samples also included in a GWAS, to generate independent datasets. For the iCOGS data we adjusted for study and used nine principal components to adjust for

potential population stratification. GWAS were adjusted for differing sets of principal components as previously described (2). The iCOGS data were similarly used to estimate per-allele ORs separately for ER-positive and ER-negative disease (27,078 and 7,333 cases, respectively).

To evaluate the evidence for association between the 2q35 enCNV and other association SNPs on 2q35, we performed multiple logistic regression in the iCOGS dataset, including all SNPs together with study and principal component as covariates. The P value for each SNP, after adjustment for all other SNPs, was determined by a Wald test.

Acknowledgements

The authors thank the many who contributed to the inception and execution of this work. This work was funded primarily by the National Cancer Institute (CA080320). The BCAC is funded by Cancer Research UK [C1287/A10118, C1287/A12014] and by the European Community´s Seventh Framework Programme under grant agreement number 223175 (grant number HEALTH-F2-2009-223175) (COGS). Funding for the iCOGS infrastructure came from: the European Community's Seventh Framework Programme under grant agreement n° 223175 (HEALTH-F2-2009-223175) (COGS), Cancer Research UK (C1287/A10118, C1287/A 10710, C12292/A11174, C1281/A12014, C5047/A8384, C5047/A15007, C5047/A10692, C8197/A16565), the National Institutes of Health (CA128978) and Post-Cancer GWAS initiative (1U19 CA148537, 1U19 CA148065 and 1U19 CA148112 - the GAME-ON initiative), the Department of Defence (W81XWH-10-1-0341), the Canadian Institutes of Health Research (CIHR) for the CIHR

Team in Familial Risks of Breast Cancer, Komen Foundation for the Cure, the Breast Cancer Research Foundation, and the Ovarian Cancer Research Fund. The Australian Breast Cancer Family Study (ABCFS) was supported by grant UM1 CA164920 from the National Cancer Institute (USA). The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the Breast Cancer Family Registry (BCFR), nor does mention of trade names, commercial products, or organizations imply endorsement by the USA Government or the BCFR. The ABCFS was also supported by the National Health and Medical Research Council of Australia, the New South Wales Cancer Council, the Victorian Health Promotion Foundation (Australia) and the Victorian Breast Cancer Research Consortium. J.L.H. is a National Health and Medical Research Council (NHMRC) Australia Fellow and a Victorian Breast Cancer Research Consortium Group Leader. M.C.S. is a NHMRC Senior Research Fellow and a Victorian Breast Cancer Research Consortium Group Leader. The ABCS study was supported by the Dutch Cancer Society [grants NKI 2007-3839; 2009 4363]; BBMRI-NL, which is a Research Infrastructure financed by the Dutch government (NWO 184.021.007); and the Dutch National Genomics Initiative. The Australian Breast Cancer Tissue Bank is generously supported by the National Health and Medical Research Council of Australia, The Cancer Institute NSW and the National Breast Cancer Foundation. The ACP study is funded by the Breast Cancer Research Trust, UK. The work of the BBCC was partly funded by ELAN-Fond of the University Hospital of Erlangen. The BBCS is funded by Cancer Research UK and Breakthrough Breast Cancer and acknowledges NHS funding to the NIHR Biomedical Research Centre, and the National Cancer Research Network (NCRN). ES is supported

by NIHR Comprehensive Biomedical Research Centre, Guy's & St. Thomas' NHS Foundation Trust in partnership with King's College London, United Kingdom. IT is supported by the Oxford Biomedical Research Centre. BOCS is supported by funds from Cancer Research UK (C8620/A8372 and C8620/A8857), a US Military Acquisition (ACQ) Activity, Era of Hope Award (W81XWH-05-1-0204) and the Institute of Cancer Research (UK). C.T. is funded by a Medical Research Council (UK) Clinical Research Fellowship. BOCS acknowledges NHS funding to the Royal Marsden / Institute of Cancer Research NIHR Specialist Cancer Biomedical Research Centre. The BSUCH study was supported by the Dietmar-Hopp Foundation, the Helmholtz Society and the German Cancer Research Center (DKFZ). The CECILE study was funded by Fondation de France, Institut National du Cancer (INCa), Ligue Nationale contre le Cancer, Ligue contre le Cancer Grand Ouest, Agence Nationale de Sécurité Sanitaire (ANSES), Agence Nationale de la Recherche (ANR). The CGPS was supported by the Chief Physician Johan Boserup and Lise Boserup Fund, the Danish Medical Research Council and Herlev Hospital. The CNIO-BCS was supported by the Instituto de Salud Carlos III, the Red Temática de Investigación Cooperativa en Cáncer and grants from the Asociación Española Contra el Cáncer and the Fondo de Investigación Sanitario (PI11/00923 and PI12/00070). The CTS was initially supported by the California Breast Cancer Act of 1993 and the California Breast Cancer Research Fund (contract 97-10500) and is currently funded through the National Institutes of Health (R01 CA77398). Collection of cancer incidence data was supported by the California Department of Public Health as part of the statewide cancer reporting program mandated by California Health and Safety Code Section 103885. HAC receives support from the Lon V Smith Foundation

(LVS39420). The University of Westminster curates the DietCompLyf database created by and funded by Against Breast Cancer Registered Charity No. 1121258
The ESTHER study was supportd by a grant from the Baden Württemberg Ministry of Science, Research and Arts. Additional cases were recruited in the context of the VERDI study, which was supported by a grant from the German Cancer Aid (Deutsche Krebshilfe). The GC-HBOC (German Consortium of Hereditary Breast and Ovarian Cancer) is supported by the German Cancer Aid (grant no 110837, coordinator: Rita K. Schmutzler). The GENICA was funded by the Federal Ministry of Education and Research (BMBF) Germany grants 01KW9975/5, 01KW9976/8, 01KW9977/0 and 01KW0114, the Robert Bosch Foundation, Stuttgart, Deutsches Krebsforschungszentrum (DKFZ), Heidelberg, the Institute for Prevention and Occupational Medicine of the German Social Accident Insurance, Institute of the Ruhr University Bochum (IPA), Bochum, as well as the Department of Internal Medicine, Evangelische Kliniken Bonn gGmbH, Johanniter Krankenhaus, Bonn, Germany. The GESBC was supported by the Deutsche Krebshilfe e. V. [70492] and the German Cancer Research Center (DKFZ).
The HABCS study was supported by an intramural grant from Hannover Medical School
The HEBCS was financilly supported by the Helsinki University Central Hospital Research Fund, Academy of Finland (266528), the Finnish Cancer Society, The Nordic Cancer Union and the Sigrid Juselius Foundation. The HERPACC was supported by a Grant-in-Aid for Scientific Research on Priority Areas from the Ministry of Education, Science, Sports, Culture and Technology of Japan, by a Grant-in-Aid for the Third Term Comprehensive 10-Year Strategy for Cancer Control from Ministry Health, Labour and Welfare of Japan, by Health and Labour Sciences Research Grants for Research on

Applying Health Technology from Ministry Health, Labour and Welfare of Japan, National Cancer Center Research and Development Fund and Grant form Takeda Health Foundation. The HMBCS was supported by a grant from the Friends of Hannover Medical School and by the Rudolf Bartling Foundation. The HUBCS was supported by a grant from the German Federal Ministry of Research and Education (RUS08/017). "Financial support for KARBAC was provided through the regional agreement on medical training and clinical research (ALF) between Stockholm County Council and Karolinska Institutet, the Swedish Cancer Society, The Gustav V Jubilee foundation and and Bert von Kantzows foundation. The KBCP was financially supported by the special Government Funding (EVO) of Kuopio University Hospital grants, Cancer Fund of North Savo, the Finnish Cancer Organizations, and by the strategic funding of the University of Eastern Finland. kConFab is supported by a grant from the National Breast Cancer Foundation, and previously by the National Health and Medical Research Council (NHMRC), the Queensland Cancer Fund, the Cancer Councils of New South Wales, Victoria, Tasmania and South Australia, and the Cancer Foundation of Western Australia. Financial support for the AOCS was provided by the United States Army Medical Research and Materiel Command [DAMD17-01-1-0729], Cancer Council Victoria, Queensland Cancer Fund, Cancer Council New South Wales, Cancer Council South Australia, The Cancer Foundation of Western Australia, Cancer Council Tasmania and the National Health and Medical Research Council of Australia (NHMRC; 400413, 400281, 199600). G.C.T. and P.W. are supported by the NHMRC. RB was a Cancer Institute NSW Clinical Research Fellow. LAABC is supported by grants (1RB-0287, 3PB-0102, 5PB-0018, 10PB-0098) from the California Breast Cancer Research Program.

Incident breast cancer cases were collected by the USC Cancer Surveillance Program (CSP) which is supported under subcontract by the California Department of Health. The CSP is also part of the National Cancer Institute's Division of Cancer Prevention and Control Surveillance, Epidemiology, and End Results Program, under contract number N01CN25403. LMBC is supported by the 'Stichting tegen Kanker' (232-2008 and 196-2010). Diether Lambrechts is supported by the FWO and the KULPFV/10/016-SymBioSysII. The MARIE study was supported by the Deutsche Krebshilfe e.V. [70-2892-BR I, 106332, 108253, 108419], the Hamburg Cancer Society, the German Cancer Research Center (DKFZ) and the Federal Ministry of Education and Research (BMBF) Germany [01KH0402]. MBCSG is supported by grants from the Italian Association for Cancer Research (AIRC) and by funds from the Italian citizens who allocated the 5/1000 share of their tax payment in support of the Fondazione IRCCS Istituto Nazionale Tumori, according to Italian laws (INT-Institutional strategic projects “5x1000”). The MCBCS was supported by the NIH grants CA128978, CA116167, CA176785 an NIH Specialized Program of Research Excellence (SPORE) in Breast Cancer [CA116201], and the Breast Cancer Research Foundation and a generous gift from the David F. and Margaret T. Grohne Family Foundation and the Ting Tsung and Wei Fong Chao Foundation. MCCS cohort recruitment was funded by VicHealth and Cancer Council Victoria. The MCCS was further supported by Australian NHMRC grants 209057, 251553 and 504711 and by infrastructure provided by Cancer Council Victoria. Cases and their vital status were ascertained through the Victorian Cancer Registry (VCR). The MEC was support by NIH grants CA63464, CA54281, CA098758 and CA132839. MSKCC is supported by grants from the Breast Cancer Research Foundation

and Robert and Kate Niehaus Clinical Cancer Genetics Initiative. The work of MTLGEBCS was supported by the Quebec Breast Cancer Foundation, the Canadian Institutes of Health Research for the "CIHR Team in Familial Risks of Breast Cancer" program – grant # CRN-87521 and the Ministry of Economic Development, Innovation and Export Trade – grant # PSR-SIIRI-701. MYBRCA is funded by research grants from the Malaysian Ministry of Science, Technology and Innovation (MOSTI), Malaysian Ministry of Higher Education (UM.C/HlR/MOHE/06) and Cancer Research Initiatives Foundation (CARIF). Additional controls were recruited by the Singapore Eye Research Institute, which was supported by a grant from the Biomedical Research Council (BMRC08/1/35/19/550), Singapore and the National medical Research Council, Singapore (NMRC/CG/SERI/2010). The NBCS has received funding from the K.G. Jebsen Centre for Breast Cancer Research; the Research Council of Norway grant 193387/V50 (to A-L Børresen-Dale and V.N. Kristensen) and grant 193387/H10 (to A-L Børresen-Dale and V.N. Kristensen), South Eastern Norway Health Authority (grant 39346 to A-L Børresen-Dale) and the Norwegian Cancer Society (to A-L Børresen-Dale and V.N. Kristensen). The NBHS was supported by NIH grant R01CA100374. Biological sample preparation was conducted the Survey and Biospecimen Shared Resource, which is supported by P30 CA68485. The Northern California Breast Cancer Family Registry (NC-BCFR) was supported by grant UM1 CA164920 from the National Cancer Institute (USA). The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the Breast Cancer Family Registry (BCFR), nor does mention of trade names, commercial products, or organizations imply endorsement by the USA Government or

the BCFR. The NHS was funded by NIH grant CA87969. The OBCS was supported by research grants from the Finnish Cancer Foundation, the Academy of Finland (grant number 250083, 122715 and Center of Excellence grant number 251314), the Finnish Cancer Foundation, the Sigrid Juselius Foundation, the University of Oulu, the University of Oulu Support Foundation and the special Governmental EVO funds for Oulu University Hospital-based research activities. The Ontario Familial Breast Cancer Registry (OFBCR) was supported by grant UM1 CA164920 from the National Cancer Institute (USA). The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the Breast Cancer Family Registry (BCFR), nor does mention of trade names, commercial products, or organizations imply endorsement by the USA Government or the BCFR. The ORIGO study was supported by the Dutch Cancer Society (RUL 1997-1505) and the Biobanking and Biomolecular Resources Research Infrastructure (BBMRI-NL CP16). The PBCS was funded by Intramural Research Funds of the National Cancer Institute, Department of Health and Human Services, USA. The pKARMA study was supported by Märit and Hans Rausings Initiative Against Breast Cancer. The RBCS was funded by the Dutch Cancer Society (DDHK 2004-3124, DDHK 2009-4318). The SASBAC study was supported by funding from the Agency for Science, Technology and Research of Singapore (A*STAR), the US National Institute of Health (NIH) and the Susan G. Komen Breast Cancer Foundation. The SBCGS was supported primarily by NIH grants R01CA64277, R01CA148667, and R37CA70867. Biological sample preparation was conducted the Survey and Biospecimen Shared Resource, which is supported by P30 CA68485. The scientific development and funding of this project were, in part, supported

by the Genetic Associations and Mechanisms in Oncology (GAME-ON) Network U19 CA148065. The SBCS was supported by Yorkshire Cancer Research S295, S299, S305PA and Sheffield Experimental Cancer Medicine Centre. The SCCS is supported by a grant from the National Institutes of Health (R01 CA092447). Data on SCCS cancer cases used in this publication were provided by the Alabama Statewide Cancer Registry; Kentucky Cancer Registry, Lexington, KY; Tennessee Department of Health, Office of Cancer Surveillance; Florida Cancer Data System; North Carolina Central Cancer Registry, North Carolina Division of Public Health; Georgia Comprehensive Cancer Registry; Louisiana Tumor Registry; Mississippi Cancer Registry; South Carolina Central Cancer Registry; Virginia Department of Health, Virginia Cancer Registry; Arkansas Department of Health, Cancer Registry, 4815 W. Markham, Little Rock, AR 72205. The Arkansas Central Cancer Registry is fully funded by a grant from National Program of Cancer Registries, Centers for Disease Control and Prevention (CDC). Data on SCCS cancer cases from Mississippi were collected by the Mississippi Cancer Registry which participates in the National Program of Cancer Registries (NPCR) of the Centers for Disease Control and Prevention (CDC). The contents of this publication are solely the responsibility of the authors and do not necessarily represent the official views of the CDC or the Mississippi Cancer Registry. SEARCH is funded by a programme grant from Cancer Research UK [C490/A10124] and supported by the UK National Institute for Health Research Biomedical Research Centre at the University of Cambridge. SEBCS was supported by the BRL (Basic Research Laboratory) program through the National Research Foundation of Korea funded by the Ministry of Education, Science and Technology (2012-0000347). SGBCC is funded by the NUS start-up Grant,

National University Cancer Institute Singapore (NCIS) Centre Grant and the NMRC Clinician Scientist Award. Additional controls were recruited by the Singapore Consortium of Cohort Studies-Multi-ethnic cohort (SCCS-MEC), which was funded by the Biomedical Research Council, grant number: 05/1/21/19/425. SKKDKFZS is supported by the DKFZ. The SZBCS was supported by Grant PBZ_KBN_122/P05/2004 The TBCS was funded by The National Cancer Institute Thailand. The TNBCC was supported by: a Specialized Program of Research Excellence (SPORE) in Breast Cancer (CA116201), a grant from the Breast Cancer Research Foundation, a generous gift from the David F. and Margaret T. Grohne Family Foundation, the Stefanie Spielman Breast Cancer fund and the OSU Comprehensive Cancer Center, the Hellenic Cooperative Oncology Group research grant (HR R_BG/04) and the Greek General Secretary for Research and Technology (GSRT) Program, Research Excellence II, the European Union (European Social Fund – ESF), and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - ARISTEIA. The TWBCS is supported by the Taiwan Biobank project of the Institute of Biomedical Sciences, Academia Sinica, Taiwan. The UCIBCS component of this research was supported by the NIH [CA58860, CA92044] and the Lon V Smith Foundation [LVS39420]. The UKBGS is funded by Breakthrough Breast Cancer and the Institute of Cancer Research (ICR), London. ICR acknowledges NHS funding to the NIHR Biomedical Research Centre. The US3SS study was supported by Massachusetts (K.M.E., R01CA47305), Wisconsin (P.A.N., R01 CA47147) and New Hampshire (L.T.-E., R01CA69664) centers, and Intramural Research Funds of the National Cancer Institute, Department of Health and Human Services, USA. The USRT

Study was funded by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, U.S. Department of Health and Human Services.

Conflict of Interest Statement

AW is a founder and shareholder of Genextropy Inc.; the remaining authors declare no potential conflicts of interest.

References

1 Siegel, R., Naishadham, D. and Jemal, A. (2013) Cancer statistics, 2013. *CA Cancer J Clin*, **63**, 11-30.

2 Michailidou, K., Hall, P., Gonzalez-Neira, A., Ghoussaini, M., Dennis, J., Milne, R.L., Schmidt, M.K., Chang-Claude, J., Bojesen, S.E., Bolla, M.K. *et al.* (2013) Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat. Genet.*, **45**, 353-361, 361e351-352.

3 Milne, R.L., Benítez, J., Nevanlinna, H., Heikkinen, T., Aittomäki, K., Blomqvist, C., Arias, J.I., Zamora, M.P., Burwinkel, B., Bartram, C.R. *et al.* (2009) Risk of estrogen receptor-positive and -negative breast cancer and single-nucleotide polymorphism 2q35-rs13387042. *J. Natl. Cancer Inst.*, **101**, 1012-1018.

4 Stacey, S.N., Manolescu, A., Sulem, P., Rafnar, T., Gudmundsson, J., Gudjonsson, S.A., Masson, G., Jakobsdottir, M., Thorlacius, S., Helgason, A. *et al.* (2007) Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat. Genet.*, **39**, 865-869.

5 Ghoussaini, M., Edwards, S.L., Michailidou, K., Nord, S., Cowper-Sal Lari, R., Desai, K., Kar, S., Hillman, K.M., Kaufmann, S., Glubb, D.M. *et al.* (2014) Evidence that breast cancer risk at the 2q35 locus is mediated through IGFBP5 regulation. *Nature communications*, **4**, 4999.

6 Ning, Y., Hoang, B., Schuller, A.G.P., Cominski, T.P., Hsu, M.-S., Wood, T.L. and Pintar, J.E. (2007) Delayed mammary gland involution in mice with mutation of the insulin-like growth factor binding protein 5 gene. *Endocrinology*, **148**, 2138-2147.

7 Tonner, E., Barber, M.C., Allan, G.J., Beattie, J., Webster, J., Whitelaw, C.B.A. and Flint, D.J. (2002) Insulin-like growth factor binding protein-5 (IGFBP-5) induces premature cell death in the mammary glands of transgenic mice. *Development*, **129**, 4547-4557.

8 Becker, M.A., Hou, X., Harrington, S.C., Weroha, S.J., Gonzalez, S.E., Jacob, K.A., Carboni, J.M., Gottardis, M.M. and Haluska, P. (2012) IGFBP ratio confers resistance to IGF targeting and correlates with increased invasion and poor outcome in breast tumors. *Clin. Cancer Res.*, **18**, 1808-1817.

9 Butt, A.J., Dickson, K.A., McDougall, F. and Baxter, R.C. (2003) Insulin-like growth factor-binding protein-5 inhibits the growth of human breast cancer cells in vitro and in vivo. *J. Biol. Chem.*, **278**, 29676-29685.

10 Hermani, A., Shukla, A., Medunjanin, S., Werner, H. and Mayer, D. (2013) Insulin-like growth factor binding protein-4 and -5 modulate ligand-dependent estrogen receptor-α activation in breast cancer cells in an IGF-independent manner. *Cell. Signal.*, **25**, 1395-1402.

11 Miele, A., Gheldof, N., Tabuchi, T.M., Dostie, J. and Dekker, J. (2006) Mapping chromatin interactions by chromosome conformation capture. *Curr Protoc Mol Biol*, **Chapter 21**, Unit-21.11.

12 Cha, R.S., Zarbl, H., Keohavong, P. and Thilly, W.G. (1992) Mismatch amplification mutation assay (MAMA): application to the c-H-ras gene. *Genome Res.*, **2**, 14-20.

13 Cheng, A.W., Wang, H., Yang, H., Shi, L., Katz, Y., Theunissen, T.W., Rangarajan, S., Shivalila, C.S., Dadon, D.B. and Jaenisch, R. (2013) Multiplexed

activation of endogenous genes by CRISPR-on, an RNA-guided transcriptional activator system. *Cell research*, **23**, 1163-1171.

14 Ambrosone, C.B., Ciupak, G.L., Bandera, E.V., Jandorf, L., Bovbjerg, D.H., Zirpoli, G., Pawlish, K., Godbold, J., Furberg, H., Fatone, A. *et al.* (2009) Conducting Molecular Epidemiological Research in the Age of HIPAA: A Multi-Institutional Case-Control Study of Breast Cancer in African-American and European-American Women. *J Oncol*, **2009**.

15 Yao, S., Zirpoli, G., Bovbjerg, D.H., Jandorf, L., Hong, C.C., Zhao, H., Sucheston, L.E., Tang, L., Roberts, M., Ciupak, G. *et al.* (2012) Variants in the vitamin D pathway, serum levels of vitamin D, and estrogen receptor negative breast cancer among African-American women: a case-control study. *Breast Cancer Res.*, **14**.

16 Bandera, E.V., Chandran, U., Zirpoli, G., McCann, S.E., Ciupak, G. and Ambrosone, C.B. (2013) Rethinking sources of representative controls for the conduct of case-control studies in minority populations. *BMC Med Res Methodol*, **13**.

17 Ruiz-Narváez, E.A., Rosenberg, L., Wise, L.A., Reich, D. and Palmer, J.R. (2011) Validation of a small set of ancestral informative markers for control of population admixture in African Americans. *Am. J. Epidemiol.*, **173**, 587-592.

18 Pritchard, J.K., Stephens, M., Rosenberg, N.A. and Donnelly, P. (2000) Association mapping in structured populations. *Am. J. Hum. Genet.*, **67**, 170-181.

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

20 Delaneau, O., Zagury, J.-F. and Marchini, J. (2013) Improved whole-chromosome phasing for disease and population genetic studies. *Nat. Methods,* **10**, 5-6.

21 Li, Y., Willer, C.J., Ding, J., Scheet, P. and Abecasis, G.R. (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet. Epidemiol.*, **34**, 816-834.

22 Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. and Abecasis, G.R. (2012) Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat. Genet.*, **44**, 955-959.

23 Marchini, J. and Howie, B. (2010) Genotype imputation for genome-wide association studies. *Nat Rev Genet,* **11**, 499-511.

24 Aulchenko, Y.S., Struchalin, M.V. and van Duijn, C.M. (2010) ProbABEL package for genome-wide association analysis of imputed data. *BMC Bioinformatics,* **11**.

25 Willer, C.J., Li, Y. and Abecasis, G.R. (2010) METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics,* **26**, 2190-2191.

26 Garcia-Closas, M., Couch, F.J., Lindstrom, S., Michailidou, K., Schmidt, M.K., Brook, M.N., Orr, N., Rhie, S.K., Riboli, E., Feigelson, H.S. *et al.* (2013) Genome-wide association studies identify four ER negative-specific breast cancer risk loci. *Nat. Genet.*, **45**, 392-398, 398e391-392.

27 Edwards, S.L., Beesley, J., French, J.D. and Dunning, A.M. (2013) Beyond GWASs: illuminating the dark road from association to function. *Am. J. Hum. Genet.*, **93**, 779-797.

28 Li, B.D., Khosravi, M.J., Berkel, H.J., Diamandi, A., Dayton, M.A., Smith, M. and Yu, H. (2001) Free insulin-like growth factor-I and breast cancer risk. *Int. J. Cancer*, **91**, 736-739.

29 Marshman, E., Green, K.A., Flint, D.J., White, A., Streuli, C.H. and Westwood, M. (2003) Insulin-like growth factor binding protein 5 and apoptosis in mammary epithelial cells. *J. Cell. Sci.*, **116**, 675-682.

30 Biong, M., Gram, I.T., Brill, I., Johansen, F., Solvang, H.K., Alnaes, G.I.G., Fagerheim, T., Bremnes, Y., Chanock, S.J., Burdett, L. *et al.* (2010) Genotypes and haplotypes in the insulin-like growth factors, their receptors and binding proteins in relation to plasma metabolic levels and mammographic density. *BMC Medical Genomics*, **3**.

31 Milanese, T.R., Hartmann, L.C., Sellers, T.A., Frost, M.H., Vierkant, R.A., Maloney, S.D., Pankratz, V.S., Degnim, A.C., Vachon, C.M., Reynolds, C.A. *et al.* (2006) Age-related lobular involution and risk of breast cancer. *J. Natl. Cancer Inst.*, **98**, 1600-1607.

32 Woolcott, C.G., Courneya, K.S., Boyd, N.F., Yaffe, M.J., McTiernan, A., Brant, R., Jones, C.A., Stanczyk, F.Z., Terry, T., Cook, L.S. *et al.* (2013) Longitudinal changes in IGF-I and IGFBP-3, and mammographic density among postmenopausal women. *Cancer Epidemiol. Biomarkers Prev.*, **22**, 2116-2120.

33 Consortium, T.E.P. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57-74.

34 Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Xu, H., Mohamed, Y.B., Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H. *et al.* (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, **462**, 58-64.

35 Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289-293.

36 Mills, R.E., Pittard, W.S., Mullaney, J.M., Farooq, U., Creasy, T.H., Mahurkar, A.A., Kemeza, D.M., Strassler, D.S., Ponting, C.P., Webber, C. *et al.* (2011) Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Res.*, **21**, 830-839.

37 Joseph, R., Orlov, Y.L., Huss, M., Sun, W., Kong, S.L., Ukil, L., Pan, Y.F., Li, G., Lim, M., Thomsen, J.S. *et al.* (2010) Integrative model of genomic factors for determining binding site selection by estrogen receptor-α. *Mol. Syst. Biol.*, **6**.

38 Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., Haussler and David. (2002) The Human Genome Browser at UCSC. *Genome Res.*, **12**, 996-1006.

Legends to Figures

Figure 1: Epigenetic and chromatin interaction profiles of the 2q35 gene desert upstream of IGFBP5. ChIP-seq read density was plotted for estrogen receptor (ERα) (37), H3K27Ac, H3K27me3, and ENCODE layered H3K27Ac (33) for breast cancer cell line MCF7 (upper panels, as labeled). Relative interaction frequency was investigated with Chromosome Conformation Capture (3C) (11) for the *IGFBP5* promoter (Anchor) in breast cancer cell line MCF7 (lower panel). Primer locations for 3C are indicated, and average profile (red line) and standard deviation (shaded region) for biological triplicates are plotted. The browser graphic was modified from the UCSC genome browser (http://genome.ucsc.edu/index.html) (38).

Figure 2: Analysis of allelic binding and effects on allelic expression of IGFBP5. ChIP-seq read density for a 3Kb region overlapping the ERα-bound looping enhancer was plotted for ERα (37), H3K27Ac, H3K27me3, and ENCODE layered H3K27Ac (33) (panels as labeled). The blue bar indicates the location of the intergenic enhancer copy number variation. ERα binding activity at the ERE (orange bar) was assayed by Chromatin Immunoprecipitation (ChIP)-qPCR for the variant (red) and wildtype (blue) alleles, and a negative control region (in *ACTB*, purple), in heterozygous MCF7 cells with estrogen treatment (vehicle and estrogen indicated in light and dark shades for each site, respectively). Allelic detection primers were designed as indicated on inset map. Error bars represent SD of biological triplicates. *P<0.004; **P<0.002. Investigation of allele-specific expression of *IGFBP5* was conducted by allelic amplification of intronic marker SNP, rs7565131. Briefly, nuclear RNA from estrogen or vehicle treated cells was

isolated. Total *IGFBP5* nuclear RNA was determined by detection of 3'UTR sequence (total bar height; error bars represent SD of biological triplicates). Allelic expression was evaluated by detection of allele-specific products by a modified MAMA(12)-qPCR. Relative abundance (total signal %) indicated by color (rs7565131-A and C as red and blue, respectively). Error bars with hats represent SD of biological triplicates. ***P=0.027; ****P=0.014

Figure 3: Regional plots of the three independent 2q35 breast cancer risk loci in 41 case control studies and 11 GWAS (n=123,499). For imputed variation within a 500Kb region including the 2q35 enCNV, $-\log_{10}$ P-values are plotted against genomic position (human reference sequence, hg19). The most strongly associated SNP in the 20Kb linkage block containing the enCNV, rs34005590, is represented by a purple diamond. The 13 additional variants in high LD (r^2>0.8) cluster tightly around ~218,000,000 (**Table S4**). Previously identified independent loci, rs13387042/rs4442975 and rs16857609 lie in centromeric and telomeric peaks, respectively. Image drawn with LocusZoom (http://csg.sph.umich.edu/locuszoom/).

Figure 1

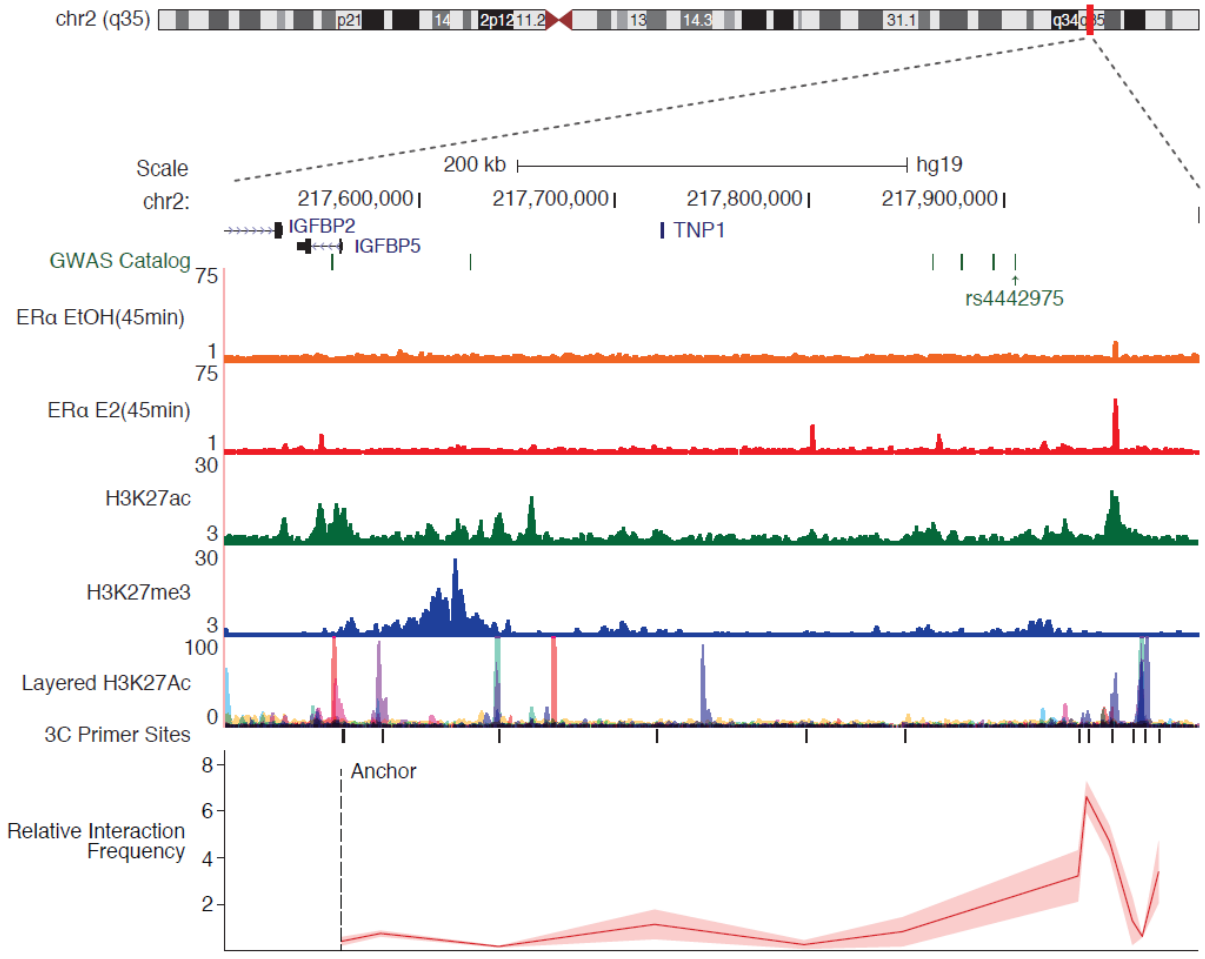

chr2 (q35)
Scale
200 kb
hg19
chr2:
217,600,000
217,700,000
217,800,000
217,900,000
IGFBP2
IGFBP5
TNP1
GWAS Catalog
rs4442975
ERα EtOH(45min)
ERα E2(45min)
H3K27ac
H3K27me3
Layered H3K27Ac
3C Primer Sites
Anchor
Relative Interaction Frequency

Figure 2

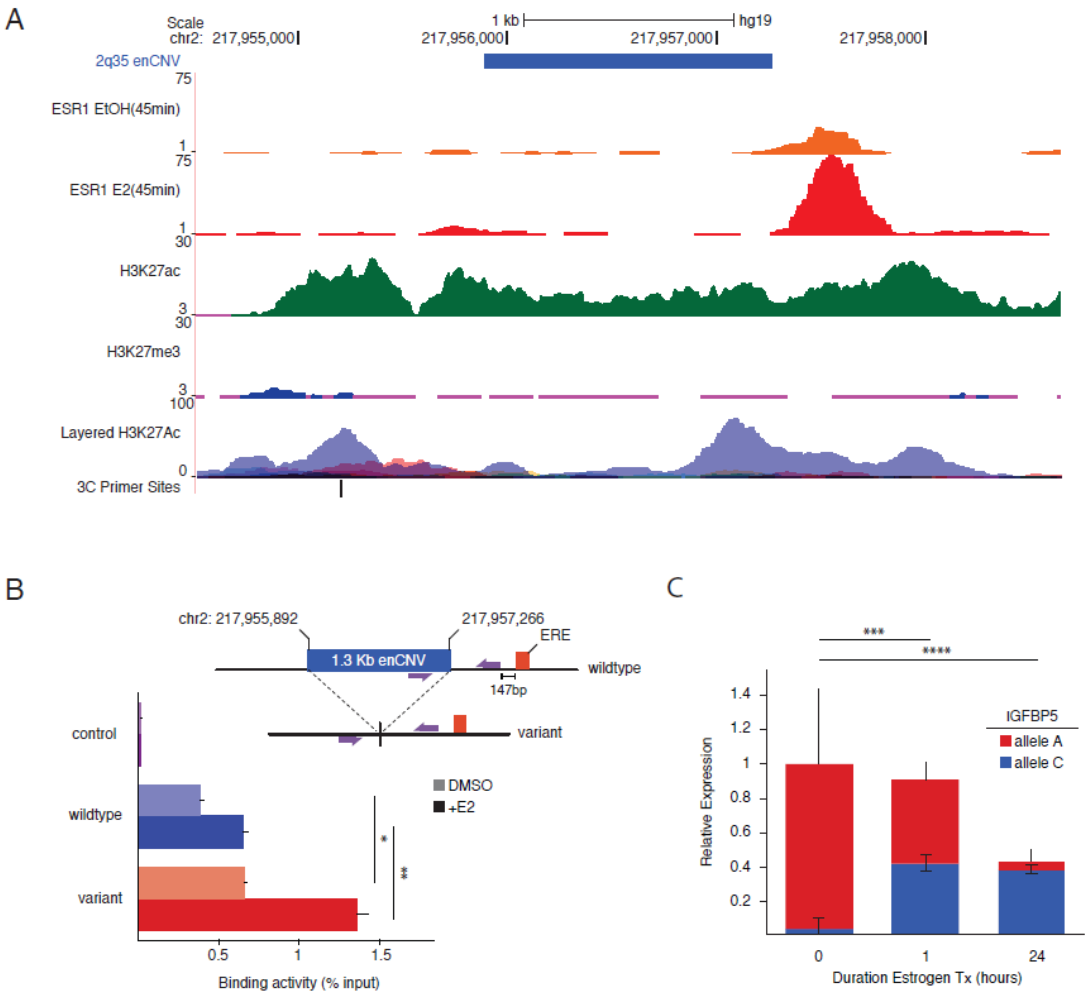
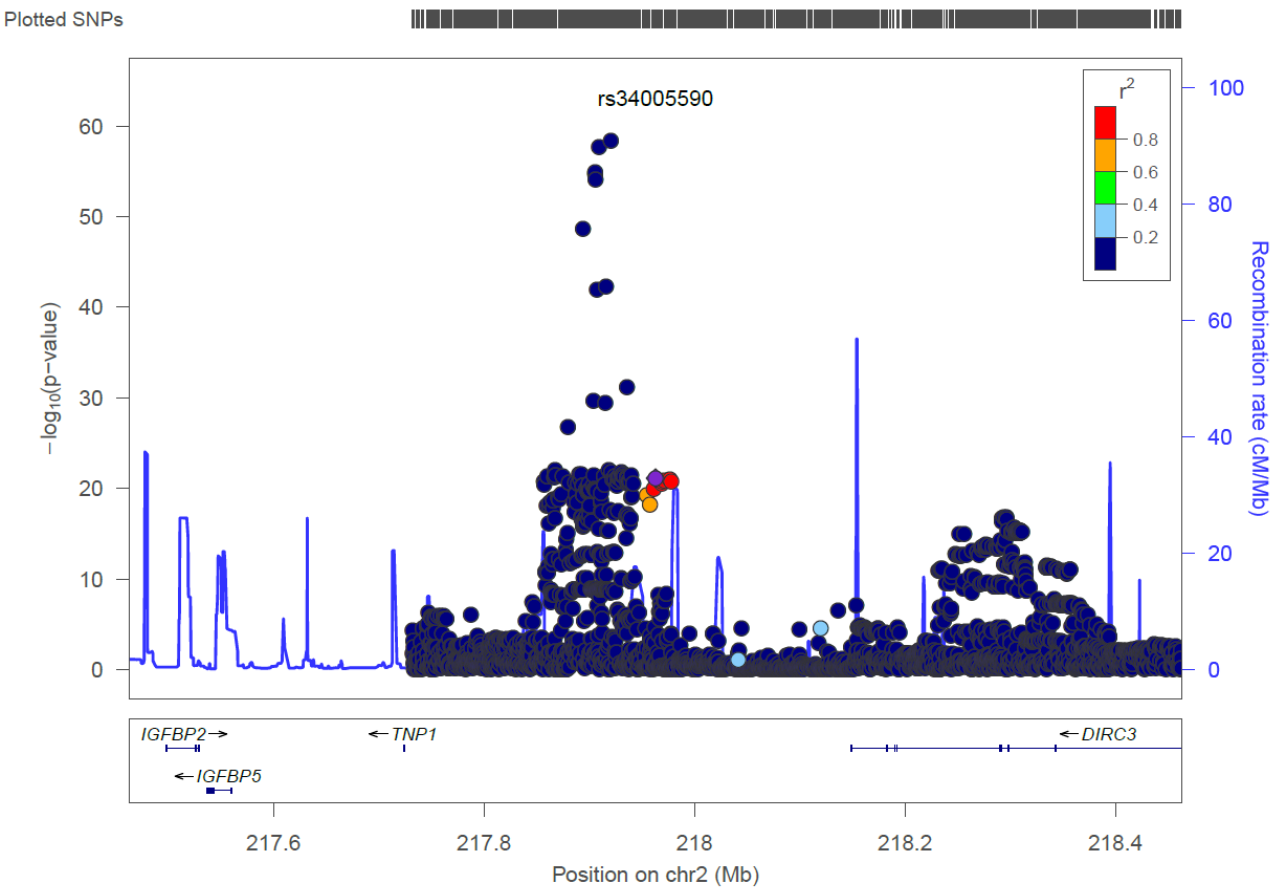

A
Scale
chr2: 217,955,000
1 kb
hg19
217,956,000
217,957,000
217,958,000
2q35 enCNV
ESR1 EtOH(45min)
ESR1 E2(45min)
H3K27ac
H3K27me3
Layered H3K27Ac
3C Primer Sites
B
chr2: 217,955,892
217,957,266
ERE
1.3 Kb enCNV
wildtype
147bp
variant
control
wildtype
variant
DMSO
+E2
Binding activity (% input)
C
IGFBP5
allele A
allele C
Relative Expression
Duration Estrogen Tx (hours)

Figure 3


Plotted SNPs
rs34005590
r2
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
IGFBP2
TNP1
DIRC3
IGFBP5
Position on chr2 (Mb)