Cancer Epidemiology, Biomarkers & Prevention

The OncoArray Consortium: A Network for Understanding the Genetic Architecture of Common Cancers

Christopher I. Amos[1], Joe Dennis[2], Zhaoming Wang[3], Jinyoung Byun[1], Fredrick R. Schumacher[4], Simon A. Gayther[5], Graham Casey[6], David J. Hunter[7], Thomas A. Sellers[8], Stephen B. Gruber[6], Alison M. Dunning[2], Kyriaki Michailidou[2], Laura Fachal[2], Kimberly Doheny[9], Amanda B. Spurdle[10], Yafang Li[1], Xiangjun Xiao[1], Jane Romm[9], Elizabeth Pugh[9], Gerhard A. Coetzee[11], Dennis J. Hazelett[12], Stig E. Bojesen[13], Charlisse Caga-Anan[14], Christopher A. Haiman[5], Ahsan Kamal[1], Craig Luccarini[2], Daniel Tessier[15], Daniel Vincent[15], François Bacot[15], David J. Van Den Berg[6], Stefanie Nelson[14], Stephen Demetriades[16], David E. Goldgar[17], Fergus J. Couch[18], Judith L. Forman[1], Graham G. Giles[19,20], David V. Conti[21], Heike Bickeböller[22], Angela Risch[23,54,55], Melanie Waldenberger[24], Irene Brüske-Hohlfeld[25], Belynda D. Hicks[26], Hua Ling[9], Lesley McGuffog[19,20], Andrew Lee[2], Karoline Kuchenbaecker[2], Penny Soucy[27], Judith Manz[24], Julie M. Cunningham[18], Katja Butterbach[28], Zsofia Kote-Jarai[29], Peter Kraft[7], Liesel FitzGerald[19,20], Sara Lindström[7,30], Marcia Adams[9], James D. McKay[31], Catherine M. Phelan[8], Sara Benlloch[2], Linda E. Kelemen[32], Paul Brennan[31], Marjorie Riggan[33], Tracy A. O'Mara[34], Hongbing Shen[35], Yongyong Shi[36], Deborah J. Thompson[2], Marc T. Goodman[12], Sune F. Nielsen[13,37], Andrew Berchuck[33], Sylvie Laboissiere[15], Stephanie L. Schmit[8,38], Tameka Shelford[9], Christopher K. Edlund[6], Jack A. Taylor[39], John K. Field[40], Sue K. Park[41], Kenneth Offit[42,43,44], Mads Thomassen[45], Rita Schmutzler[46], Laura Ottini[47], Rayjean J. Hung[48], Jonathan Marchini[49], Ali Amin Al Olama[2], Ulrike Peters[50], Rosalind A. Eeles[29], Michael F. Seldin[51,52], Elizabeth Gillanders[14], Daniela Seminara[14], Antonis C. Antoniou[2], Paul D.P. Pharoah[2], Georgia Chenevix-Trench[34], Stephen J. Chanock[53], Jacques Simard[27], and Douglas F. Easton[2]

[1]Biomedical Data Science, Geisel School of Medicine at Dartmouth, Hanover, New Hampshire. [2]Centre for Cancer Genetic Epidemiology, University of Cambridge, Cambridge, United Kingdom. [3]Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, Tennessee. [4]Department of Epidemiology and Biostatistics, School of Medicine, Case Western Reserve University, Cleveland, Ohio. [5]The Center for Bioinformatics and Functional Genomics at Cedars Sinai Medical Center, Greater Los Angeles Area, Los Angeles, California. [6]Department of Preventive Medicine, Keck School of Medicine, University of Southern California Norris Comprehensive Cancer Center, Los Angeles, California. [7]Department of Epidemiology, Program in Molecular and Genetic Epidemiology, Harvard School of Public Health, Boston, Massachusetts. [8]Department of Cancer Epidemiology, H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida. [9]Center for Inherited Disease Research, Institute of Genetic Medicine, Johns Hopkins School of Medicine, Baltimore, Maryland. [10]Molecular Cancer Epidemiology, QIMR Berghofer Medical Research Institute, Herston, Queensland, Australia. [11]Van Andel Research Institute, Grand Rapids, Michigan. [12]Cedars-Sinai Medical Center, Los Angeles, California. [13]Department of Clinical Biochemistry, Herlev and Gentofte Hospital, Copenhagen University Hospital, Copenhagen, Denmark. [14]Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, Maryland. [15]Génome Québec Innovation Centre, Montreal, Canada and McGill University, Montreal, Canada. [16]University Health Network- The Princess Margaret Cancer Centre, Toronto, California. [17]Huntsman Cancer Institute, Salt Lake City, Utah. [18]Mayo Clinic, Rochester, Minnesota. [19]Cancer Epidemiology Centre, Cancer Council Victoria, Melbourne, Australia. [20]Cancer, Genetics and Immunology, Menzies Institute for Medical Research, Hobart, Australia. [21]Division of Biostatistics, Department of Preventive Medicine, Zilkha Neurogenetic Institute, University of Southern California, Los Angeles, California. [22]Department of Genetic Epidemiology, University Medical Center, Georg-August-University, Göttingen, Germany. [23]University of Salzburg and Cancer Cluster Salzburg, Salzburg, Austria. [24]Research Unit of Molecular Epidemiology, Institute of Epidemiology II, Helmholtz Zentrum München, German Research Center for Environmental Health, Neuherberg, Germany. [25]Helmholtz Zentrum München, Institut für Epidemiologie I, Neuherberg, Oberschleissheim, Germany. [26]Cancer Genomics Research Laboratory, Frederick National Laboratory for Cancer Research, Frederick, Maryland. [27]Cancer Genomics Laboratory, Centre Hospitalier Universitaire de Québec and Laval University, Québec City, Canada. [28]Cancer Epidemiology, German Cancer Research Center, Heidelberg, Germany. [29]Institute of Cancer Research, London, England. [30]Department of Epidemiology, University of Washington, Seattle, Washington. [31]International Agency for Research on Cancer, World Health Organization, Lyon, France. [32]Department of Public Health Sciences, Medical University of South Carolina, Charleston, South Carolina. [33]Department of Gynecology, Duke University Medical Center, Durham, North Carolina. [34]Cancer Division, QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia. [35]Department of Epidemiology and Biostatistics, Jiangsu Key Lab of Cancer Biomarkers, Prevention and Treatment, Collaborative Innovation Center for Cancer Medicine, School of Public Health, Nanjing Medical University, Nanjing, P.R. China. [36]Key Laboratory for the Genetics of Developmental and Neuropsychiatric Disorders, Ministry of Education, Bio-X Institutes, Shanghai Jiao Tong University, Shanghai, P.R. China. [37]Department of Oncology, Herlev and Gentofte Hospital, Copenhagen University Hospital, Copenhagen, Denmark. [38]Department of Gastrointestinal Oncology, H. Lee Moffitt Cancer Center, Tampa, Florida. [39]Molecular and Genetic Epidemiology Group, National Institute for Environmental Health Sciences, Research Triangle Park, North Carolina. [40]Institute of Translational Medicine, University of Liverpool, Liverpool, United Kingdom. [41]College of Medicine, Seoul National University, Gwanak-gu, Seoul, Korea. [42]Clinical Genetics Service, Memorial Hospital, New York, New York. [43]Cancer Biology and Genetics Program, Sloan Kettering Institute, New York, New York. [44]Department of Medicine, Weill Cornell Medical College, New York, New York. [45]Department of Clinical Genetics, Odense University Hospital, Odense, Denmark. [46]Zentrum Familiärer Brust- und Eierstockkrebs, Universitätsklinikum Köln, Köln, Germany. [47]Department

Abstract

Background: Common cancers develop through a multistep process often including inherited susceptibility. Collaboration among multiple institutions, and funding from multiple sources, has allowed the development of an inexpensive genotyping microarray, the OncoArray. The array includes a genome-wide backbone, comprising 230,000 SNPs tagging most common genetic variants, together with dense mapping of known susceptibility regions, rare variants from sequencing experiments, pharmacogenetic markers, and cancer-related traits.

Methods: The OncoArray can be genotyped using a novel technology developed by Illumina to facilitate efficient genotyping. The consortium developed standard approaches for selecting SNPs for study, for quality control of markers, and for ancestry analysis. The array was genotyped at selected sites and with prespecified replicate samples to permit evaluation of genotyping accuracy among centers and by ethnic background.

Results: The OncoArray consortium genotyped 447,705 samples. A total of 494,763 SNPs passed quality control steps with a sample success rate of 97% of the samples. Participating sites performed ancestry analysis using a common set of markers and a scoring algorithm based on principal components analysis.

Conclusions: Results from these analyses will enable researchers to identify new susceptibility loci, perform fine-mapping of new or known loci associated with either single or multiple cancers, assess the degree of overlap in cancer causation and pleiotropic effects of loci that have been identified for disease-specific risk, and jointly model genetic, environmental, and lifestyle-related exposures.

Impact: Ongoing analyses will shed light on etiology and risk assessment for many types of cancer. *Cancer Epidemiol Biomarkers Prev; 26(1); 126–35. ©2016 AACR.*

Introduction

Cancer is one of the leading causes of death worldwide. In 2012, the estimated number of cancer cases around the world was 14.1 million, and this number is estimated to swell to 21 million by 2030 (1). Cancer has a sizable heritable component. A large twin study estimated that heritable factors may explain between 20% and 40% of the variance in cancer risk (2). High-penetrance mutations, including those in *BRCA1* and *BRCA2*, *APC*, and DNA mismatch repair genes, are estimated to account for less than 5% of all cases (3, 4). As for other common complex diseases, it is expected that much of the inherited susceptibility to cancer is likely to be explained by common alleles having low penetrance (2–5). Large consortial efforts may identify effects from additional rarer alleles (6, 7). As pointed out by Ponder (8, 9) and Peto (10), common genetic variants account for a large proportion of cancer incidence, even though they do not individually lead to strong clustering within families. Moreover, the combinations of effects from genetic and environmental factors may account for substantial differences in cancer susceptibility within and among populations (8–13).

Over the past decade, genome-wide association studies (GWAS) of cancer have discovered multiple low-penetrance loci. Given that the effect sizes are generally weak (relative risks per allele of 1.3 or less), increasing the sample size has become crucial in identifying and characterizing true genetic associations. Genetic signatures of cancer etiology indicated novel influences in cancer development, thereby providing new insights into etiologic mechanisms that suggest interventions (14). By identifying many new loci influencing cancer development, genomic research has identified pathways that influence cancer development (15). In addition, Mendelian randomization has emerged as an effective approach for confirming nongenetic etiologic factors identified through epidemiologic studies, removing potential concerns about reverse causality (16).

Once the loci are identified, fine-mapping studies are a critical next step in finding functional variant(s) and in the discovery of nearby, independent, secondary signals, which may increase the heritable fraction explained by each region. More than 90% of risk-alleles lie in nonprotein-coding DNA and there is now unequivocal evidence that risk regions are enriched for regulatory elements, including enhancers, promoters, insulators, and silencers (17). In general, genome-wide estimates in humans indicate about 500,000 enhancers may alter regulation of expression and thus alter risk by controlling expression of target susceptibility genes (17–20). Analyses to date indicate that several regions harbor multiple distinct susceptibility variants for different cancer types, suggesting common mechanisms but tissue-specific regulation (21). Thus, fine-mapping of multiple cancer types using a

of Molecular Medicine, Sapienza, University of Rome, Rome, Italy. [48]Lunenfeld-Tanenbaum Research Institute of Mount Sinai Hospital, University of Toronto, Toronto, Canada. [49]Department of Statistics, Oxford University, Oxford, United Kingdom. [50]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, Washington. [51]Department of Biochemistry and Molecular Medicine, University of California at Davis, Davis, California. [52]Department of Internal Medicine, University of California at Davis, Davis, California. [53]Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, Bethesda, Maryland. [54]Division of Epigenomics and Cancer Risk Factors, German Cancer Research Center, Heidelberg, Germany. [55]Translational Lung Research Center Heidelberg, Member of the German Center for Lung Research, Heidelberg, Germany.

Note: Supplementary data for this article are available at Cancer Epidemiology, Biomarkers & Prevention Online (http://cebp.aacrjournals.org/).

C.I. Amos, D.J. Hunter, F.R. Schumacher, S.B. Gruber, T.A. Sellers on behalf of the GAME-ON Consortium; A.B. Spurdle on behalf of the ECAC Consortium; H. Bickeböller on behalf of the GLC Consortium; D.F. Easton on behalf of the BCAC Consortium; G. Chevenix-Trench on behalf of the CIMBA consortium; J. Simard on behalf of the PERSPECTIVE consortium; P.D.P. Pharoah on behalf of the OCAC consortium; and R.A. Eeles on behalf of the PRACTICAL consortium. Other contributors provided input independent of these consortia.

Corresponding Author: Christopher I. Amos, Department of Biomedical Data Science, Geisel School of Medicine at Dartmouth, 1 Medical Center Drive, Lebanon, NH 03766. Phone: 603-650-1972; Fax: 603-650-1966; E-mail: Christopher.I.Amos@dartmouth.edu

doi: 10.1158/1055-9965.EPI-16-0106

common array is likely to be an effective strategy for finding new alleles influencing common cancers and for unravelling mechanisms in their etiology.

The overall goal of the OncoArray Consortium is to gain new insights into the genetic architecture and mechanisms underlying common cancers by deploying a new genotyping array, the OncoArray, and using it to genotype a large number of cases with cancers of the breast, colon, lung, ovary, prostate, or endometrium as well as genetically susceptible individuals such as *BRCA1* and *BRCA2* mutation carriers along with a large number of cancer-free controls. The collaboration arose, in part, through the efforts of the Genetic Associations and Mechanisms in Oncology (GAME-ON) consortium, which was a multi-year project to characterize SNP associations for common cancers and to understand their mechanistic and functional consequences in disease development. The OncoArray project provides an unprecedented opportunity both to discover new cancer susceptibility variants, common and rare, and to identify the likely causal variants at known loci through fine-mapping and the integration of disease-associated variants with tissue-specific regulatory information. In addition, joint genotyping across cancer sites permits sharing of controls and a more comprehensive assessment of genetic risk among many cohort studies that participated in this study. Moreover, given the evidence that some of the loci influencing cancer risk are shared among cancer sites, the genotyping of a common array across multiple cancer sites provides an excellent opportunity to study the pleiotropic effect of susceptibility loci. However, while there is tremendous value in organizing a genotyping consortium on this scale, there are also substantial challenges in how best to integrate data across this diverse spectrum of cancer sites and genotyping locations. To facilitate the analysis, the consortium developed shared procedures for genotype calling and quality control. This report describes the development of the consortium, the array that was designed, and quality control approaches that have been implemented across the consortium.

Materials and Methods

Principles in sample and SNP selection

The OncoArray Consortium is focused on the discovery of variants influencing common cancers, in particular cancers of the breast, colon, lung, ovary, and prostate. These cancers were chosen for analysis based upon prior observation of some common causal pathways (15) as well as the opportunity provided by common funding through the GAME-ON, a consortium of U19 grants studying genetic etiology of breast, ovarian, prostate, colon, and lung cancers. The existence of an effective, multi-consortium collaboration provided an opportunity primarily because of economies of scale. The potential to utilize common control sets across the consortia gave added value. A description of the sample sets is provided in Supplementary Tables S1A–S1G. Endometrial cancer cases were included because endometrial cancer shares several risk factors with breast cancer and ovarian cancer, such as the genetic locus (*HNF1B*), which has shared variants with prostate (22, 23) and ovarian cancer (24). Finally, there are similarities in tumor phenotype and/or shared tissue of origin between endometrial cancer, the benign gynaecologic condition endometriosis, the endometrioid and clear cell histologies of ovarian cancer, and basal-like breast cancer (25–27). Thus, pooling ovarian and endometrial (23, 28, 29) cases could uncover novel loci.

The array was designed from a final list of approximately 600,000 markers, of which approximately 533,000 were successfully manufactured. Of these, nearly 50% of the markers were selected as a GWAS backbone (Illumina HumanCore). These markers were selected to tag the large majority of known common variants, via imputation. The remaining markers were selected from seven lists: five from the disease consortia representing the main cancer sites, one from the Consortium of Investigators of Modifiers of BRCA1/2 (CIMBA) including potential modifiers of cancer risk in *BRCA1* or *BRCA2* carriers, and a seventh "common" list that included variants of common interest (see below). SNPs were allocated to these disease sites, and to CIMBA, according to the number of samples that each consortium would be contributing. In addition, the array that was configured by Illumina allows flexibility for cancers not originally participating in the design of the array by allowing additional custom content to be added to the array. The general principles for SNP allocation were set by consensus by members of the OncoArray Consortium as presented in Table 1. More detailed descriptions of the SNP selection process for disease sites participating in the OncoArray are also provided in the Supplementary Methods and governance described in Supplementary Information about the Oncoarray Consortium. Below, we present the general approaches that were taken for nominating SNPs for the Array.

Table 1. Organization of SNP requests within consortia

	Consortium				
Selection of SNP	**TRICL**	**BCAC/DRIVE/CIMBA**	**FOCI/OCAC**	**ELLIPSE/PRACTICAL**	**CORECT**
Fine-mapping	0.437	0.259	0.700	0.359	0.346
Significant SNPs from existing GWAS	0.032	0.465	0	0.379	0.598
Sequencing/rare variants	0.001	0.075	0	0.025	0.012
Other GWAS studies/ethnicities	0.072	0.044	0	0.128	0.005
Candidate SNP and pathways	0.156	0.012	0.105	0.056	0.027
Correlated phenotypes	0.083	0.015	0	0.051	0
GxG or GxE interactions	0.004	0.063	0.115	0	0.012
SNPs from tumor genes	0.053	0	0	0	0
Functional and eQTL	0.161	0.005	0.002	0	0
Survival	0	0.062	0.079	0	0
SNPs within consortium	32,464	88,475	42,921	67,757	37,397

Abbreviations: BCAC, Breast Cancer Association Consortium; CORECT, Colorectal Transdisciplinary Study; DRIVE, Discovery, Biology, and Risk of Inherited Variants in Breast Cancer; Ellipse, Elucidating Loci Involved in Prostate cancer Susceptibility; FOCI, Follow-up of Ovarian Cancer Genetic Association and Interaction Studies; OCAC, Ovarian Cancer Association Study; PRACTICAL, Prostate Cancer Association Group to Investigate Cancer Associated Alterations in the Genome; TRICL, Transdisciplinary Research in Cancer of the Lung.

Selection of SNPs for inclusion within disease site

SNPs to be included in the array were nominated by participating consortia organized into each of the major disease site groups that participated in the primary array development. Each cancer site used its own prioritization scheme. Generally, selection of SNPs were based on (i) candidate SNPs from loci enriched showing some evidence of association (e.g., $P < 10^{-5}$) from previous GWAS of common cancers (breast, ovarian, prostate, colon, and lung; refs. 30–37); (ii) fine-mapping of risk loci based on 1000 Genomes Project data and resequencing studies (38); (iii) candidate rare variants from whole genome and whole exome studies, and exome arrays (39); (iv) findings from previously published studies of other cancers provided by the NHGRI SNP catalog (40) and other online resources; and (v) other "wild-card" variants, for example, variants of potential functional significance (18, 41, 42). The majority of SNP selection was based on regions previously identified from GWAS in European populations, but disease sites also allocated tagging SNPs to capture variability for Asian and African descent populations. In addition to site-specific variants, some of which were nominated by more than one group, candidates were nominated from *in silico* functional analyses that suggested putative mechanistic targets for risk variants based either on their predicted effects on the coding sequence of candidate genes, or their intersection with noncoding, putative regulatory targets (see below). Finally, variants associated with phenotypes that correlate with cancers (such as smoking or BMI) were also selected.

Selection of SNPs for fine-mapping

Similar procedures were followed for each site. We first defined a 1 Mb interval surrounding the known lead signal for each genome-wide signal. Where such regions overlapped, the intervals were amalgamated into a single interval so as to include 500 kb either side of each hit. Common regions were defined as regions including hits within 1 Mb for more than one cancer type, amalgamated as described. We then identified and obtained design scores for all variants in the interval from the 1000 Genomes Project (phase I version 3, March 2012 release). From among designable SNPs, we then selected three sets of variants: (i) all variants correlated with the known hits at $r^2 > 0.6$, (ii) all variants from lists of potentially functional variants, defined through RegulomeDB, and (iii) a set of SNPs designed to tag all remaining variants at $r^2 > 0.9$.

Selection of "Common" SNPs

Previous analyses (30, 32, 43, 44) have demonstrated that association signals for different cancers tend to cluster together, perhaps reflecting common mechanisms. For this reason, we selected a dense set of SNPs within 1 Mb (see above) across all regions in which this occurred for more than one cancer type. Variants were nominated for inclusion if they (i) occurred within genes that have been found to associate with pharmacogenetic traits relevant to cancer, (ii) had previously been associated at genome-wide levels of significance for any other cancer type (not among the five primary cancers sites participating in the OncoArray Consortium) as defined by the GWAS Catalog (45), and (iii) had been found to be relevant to cancer-associated traits (46) including BMI, height, and waist-to-hip ratio [in collaboration with the GIANT consortium (47)], smoking, age at menopause or menarche [in collaboration with the REPROGEN consortium (48)], and telomere length in lymphocytes (31). We also included additional SNPs that showed evidence of association with other cancer types including endometrial, testis, bladder, and pancreatic cancer, Wilms' tumor, and glioma, and SNPs tagging known common eQTLs (i.e., associated with expression across a range of tissues).

Pharmacogenetic variants were nominated by several collaborators based on (i) functional variants in 19 genes nominated by the pharmacogenetics network, (ii) functional variants or tagging SNPs in CYP2A6 and CYP2B6, and (iii) SNPs nominated by PharmGKB and variants nominated from study of cell lines to affect expression of pharmacogenetically relevant genes (49). SNPs from the region of chromosome 15q25.1 that associate with lung cancer and smoking behavior were placed in the common region given the ubiquitous effects of smoking on cancer risks. Of note, *BRCA1* and *BRCA2* were finally released from patent controls two days before the final selection of SNPs so that common functional variants of these loci could be included in the array. We included additional (nonpolymorphic) probes for each exon of *BRCA1*, *BRCA2*, *MLH1* and *MSH2* to capture large deletions in these genes. Finally, we included a panel of Y chromosome and mitochondrial markers to provide data on population ancestry.

The Division of Cancer Epidemiology and Genetics of the National Cancer Institute accumulated GWAS scan data for other cancer sites including bladder, NHL (non-Hodgkin lymphoma), esophageal, gastric, glioma, kidney, osteosarcoma, pancreas, testis, or scan data for non-Caucasian studies including Asian nonsmoking female lung cancer and African American lung cancer. The top 200–400 most significant loci from each scan were selected after ranking by association test *P* value and LD pruning ($r^2 > 0.6$).

Functional characterization and selection. Risk variants at known susceptibility loci for breast, colorectal, lung, ovarian, and prostate cancer were integrated with epigenomic datasets from ENCODE and other published sources, to identify intersections between risk SNPs and tissue-specific regulatory features that define the most likely causal variants and their functional targets. We interrogated associations between SNPs and DNAse Hypersensitivity (DHS) sites generated in the pan-cancer cell line panel from ENCODE, as well the LNCaP cell line (for prostate cancer–specific marks), the HMEC line (for breast), the SAEC line (for lung cancer), the HCT116 line (for colorectal cancer), and the CaOV3 line (for ovarian cancer). The most likely causal SNPs from these analyses were prioritized in the selection of fine-mapping variants described above. In addition, we identified candidate causal SNPs at loci associated with risk of two or more cancers, to identify the putative functional targets that are common across cancer types as well as those that are tissue/cancer specific at these loci. A summary of these analyses are illustrated in Fig. 1. This approach evaluates regions around the significant SNPs common to cancers to identify regional variants that impact chromatin structure, expression levels or transcription factor binding sites to enrich for SNPs directly related to cancer development.

Pruning and merging procedures

As a starting point, we "forced-in" all SNPs in the GWAS backbone (260,660) and the common fine-mapping list (32,548). All other lists include SNPs that passed design at Illumina and were rank ordered with the most important SNPs first, and were pruned

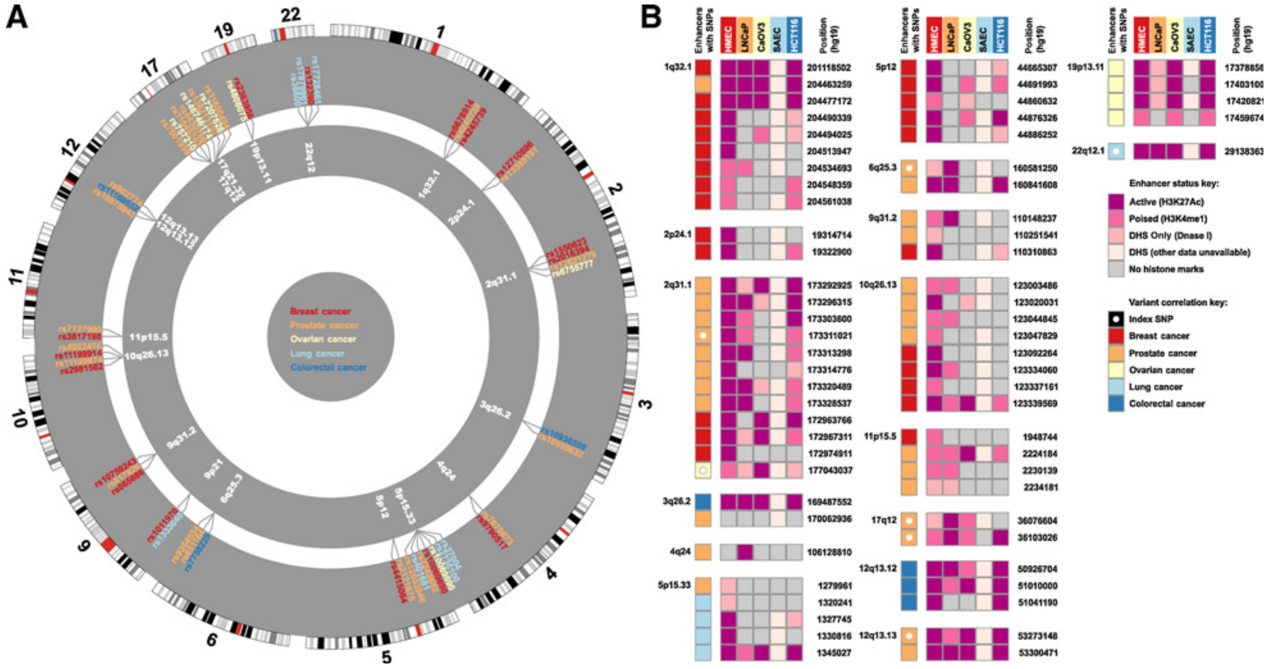

Figure 1.

Twenty risk regions analyzed as part of the GAME-ON OncoArray, including 17 pleiotropic regions conferring risks to two or more common cancers (breast, colorectal, lung, ovarian, or prostate cancers). **A,** Circos plot illustrating the 24 different regions ordered by chromosome and cytoband. The index SNP(s) at each locus are color coded by cancer type. **B,** Integration of correlated risk SNPs at each locus with regional catalogs of regulatory marks for related tissue types for common cancers to identify SNPs intersecting tissue-specific regulatory targets. Publicly available genome-wide regulatory profiling data were available for the HMECs (human mammary epithelial cells; specific to breast cancer), LNCap cancer cells (for prostate cancer), CaOV3 cancers (for ovarian cancer), SAEC cells (for lung cancer). The first column indicates a risk-associated SNP that intersects a regulatory mark, color coded by cancer type. For other columns, colored squares represent an intersection between a risk associated SNP and a regulatory mark, and in which tissue type, indicating which marks are common across tissues and which are tissue specific. White squares indicate the most strongly associated SNPs (index SNP) in a region and a dot within the square indicates an intersection between a regulatory mark and an index. The position of each regulatory mark is indicated relative to hg19 coordinates. In **B,** only SNPs with regulatory marks are shown, thus excluding 24 of the regional associations shown in **A**.

to exclude redundant SNPs in LD (r^2 > 0.9) with other SNPs in the same list or the "force-in" set described above.

The proportions allocated to each disease site are listed in the Supplementary Table S2.

The final merging took the lists of SNPs generated by the disease sites and for common mapping and generated a single list in the following order:

(i) Include the GWAS backbone

(ii) Include the Common fine-mapping list

(iii) Choose the remaining SNPs iteratively from the five ranked lists. At each stage, choose the next SNP from the list with the smallest value of n/p, where n is the number of SNPs already chosen from that list and p is the proportional allocation of that list, as given in the above table. This ensured that the correct proportions were kept.

(iv) Include the SNP unless the exact SNP has already been chosen. In either case, augment the count n for that list by 1.

(v) Increase the number of beadtypes for chosen SNPs, where necessary because variation could not be captured by a single beadtype.

On the basis of the merged list of 715,637 unique SNPs (76,290 from lung; 224,074 from familial and sporadic breast and ovarian; 81,009 from prostate; 50,110 from colorectal; 17,547 from common list), we further performed LD pruning ($r^2 > 0.95$). This process resulted in a total of 651,216 SNPs. A set of obligatory SNPs provided by each contributing lists was not allowed to be "pruned."

After this process, we submitted 568,712 SNPs (reaching the total number of ~600,000 beadtypes) from the priority lists to Illumina for manufacturing. Of these, a total of 533,631 (93.8%) passed quality control procedures and were included as valid markers on the array.

Genotyping

To minimize variability that might result from genotyping among sites and to improve efficiency, the large majority of genotyping was performed at just 8 sites CIDR (n = 211,638), Cambridge (n = 98,770), Genome Quebec/McGill Innovation Center (n = 55,121), the National Cancer Institute (26,803), the Mayo Clinic (n = 22,023), Denmark (n = 5,961), and Shanghai (n = 3,840). To ensure comparability among centers, selected HapMap samples were analyzed by all groups.

Quality control steps

A detailed quality control plan was developed and is included as Supplementary Material, but the salient features are presented here. Participating sites genotyped a common set of HapMap samples so that strand alignment and integrity of imputation could be compared among analytic sites. All sites used a common genotype clustering file that can be downloaded from http://consortia.ccge.medschl.cam.ac.uk/oncoarray/onco_v2c.zip and removed 765 duplicated probes (onco_duplicate_variants_excluded.csv).

Clustering process

A selection of 56,284 samples with high call rates from across the genotyping centers were combined into a single Illumina Genome Studio project and automatic clustering performed using the GenTrain 2 clustering algorithm. This included 3,687 African American, 5,590 Asian, and 2,608 Hispanic samples. A large number of samples was used to increase the chances of including heterozygotes for the many rare variants on the array (23,249 variants have a MAF below 0.0005). Variants showing poor clustering (57,673) were manually evaluated and revised, which reduced the number to 16,526 variants excluded from the analysis.

Ancestry analysis

Ancestry analysis was performed using a standardized approach in which 2,318 ancestry informative markers (AIM) with minor allele frequencies of 0.05 or higher were used on 66,105 samples genotyped at CIDR, Cambridge, and Genome Quebec/McGill Innovation Center (the primary contributing centers) and 505 HapMap 2 samples. We noted that among those individuals not clearly aligning into one of the major continental ancestry groups there are clines connecting ancestral groups along axes connecting the centroids of the ancestral populations. We mapped ancestry to regions of a triangle connecting the three regions, to estimate the contribution of European, Asian, and African ancestry to each individual. The method is further described in the software package FastPopc (50) distributed to consortium members. Individuals were thus classified into 4 groups for downstream analyses: European (defined as >80% European ancestry), Asian (>40% Asian ancestry), African (>20% African ancestry), and other (not fulfilling any of the above criteria; ref. 50; see Supplementary Methods).

Results

Genotyping quality

Samples passed genotyping quality control steps if more than 95% of SNPs had valid calls. After manual review of cluster plots for SNPs failing to achieve 95% call rates a total of 494,763 SNPs were retained for analysis. The call rate varied according to tissue source and DNA processing steps (Fig. 2). Overall, 97% of samples had call rates of 95% or higher. However, the efficiency in genotyping varied markedly among sources of DNA. In particular, genotyping of samples derived from peripheral blood provided excellent performance with a 98% success rate, while amplified DNA derived from nonblood samples show poor performance (18% overall failure rate for amplified buccal or saliva). The success rate for genotyping HapMap-derived samples was 100% and the overall genotyping success rate for lymphoblastoid lines was 99.5%.

Analysis of concordance of sample genotypes

To evaluate the reliability of genotyping across samples including postimputation processing, we evaluated concordance of imputed SNP genotype probabilities among the centers. Figure 3 depicts average squared correlations among 19,367,932 variants imputed from v3 of the 1000 Genomes Project for HapMap samples genotyped and imputed in Cambridge versus the same samples genotyped by CIDR and imputed at Dartmouth using the same imputation protocol (Supplementary Methods). The integral values along the x-axis depict results for the same individual, with multiple replicate samples having been genotyped for individuals 1, 4, 5, 6 and 8. Samples 1–8 derive from European descent individuals, samples 9–10 are Chinese, sample 11 is Japanese, and samples 12–14 are Yoruban. Correlations in genotypes performed at different centers were high but were slightly higher for European descent samples (average $R^2 = 0.985$) versus Chinese (average $R^2 = 0.958$), Japanese (average $R^2 = 0.961$) or Yorubans (average $R^2 = 0.975$). Supplementary Figure S1 compares the imputation accuracy of the OncoArray to several other arrays.

Discussion

Comparison with other large-scale genotyping efforts

The OncoArray is a scientific community-derived effort from many worldwide investigators to understand common causes

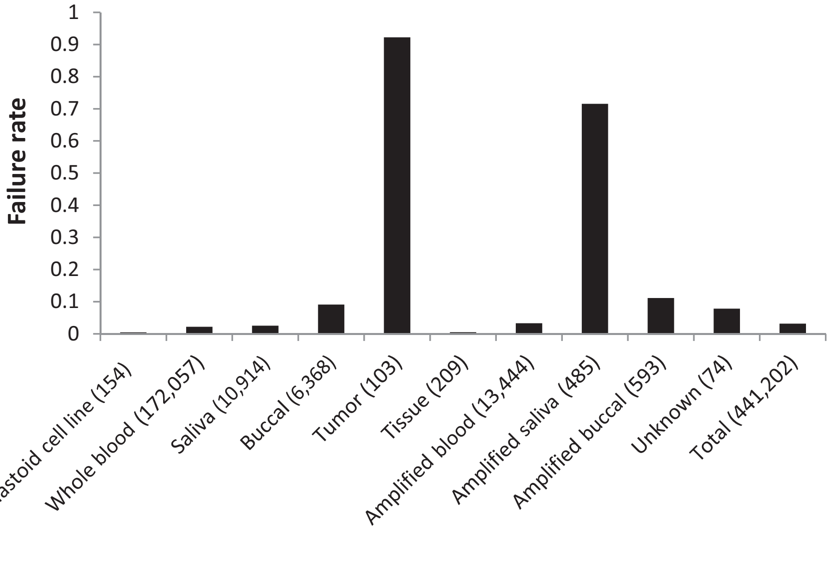

Figure 2.
Failure rates (<95% of SNPs called) for 211,594 samples genotyped by CIDR across multiple tissue types. The overall failure rate was 2.97%. All tumor samples were prepared from formalin-fixed paraffin embedded ovarian cancers.

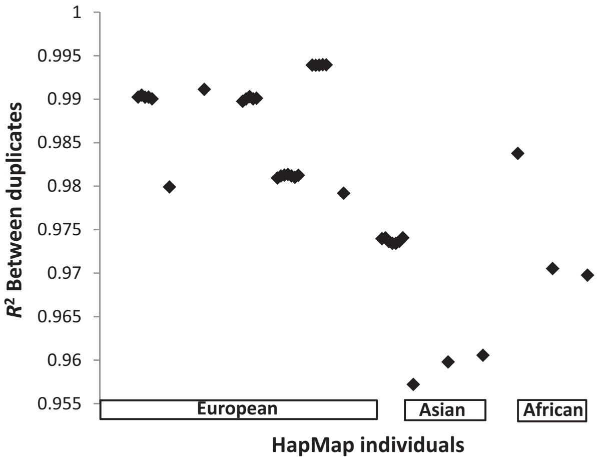

Figure 3.
Correlation between replicate HapMap samples genotyped at Cambridge versus the Center for Inherited Disease Research. Samples 1–8 are of European origin while samples 9–14 are Asian or African. There are multiple replicates of samples 1, 4, 5, 6, and 8. Samples 1–8 are European, 9–10 are Chinese, sample 11 is Japanese, and samples 12–14 are Yoruban.

of cancer susceptibility and progression. The array that was configured balanced several needs. First, each of the contributing groups had specific interest in fine-mapping and validation of previously suggested loci. This element of the OncoArray is similar to prior large-scale consortia such as the Metabochip (51) and the Immunochip (52), which are highly targeted arrays. Balanced against the fine-mapping element, we also allocated about 50% of the array to permit further discovery of novel variants. The array balances the needs for new discovery with validation and fine-mapping; it is unlike prior arrays such as the Metabochip or Immunochip which did not include a GWAS array backbone. More generic platforms such as the Biobank array can be applied for a broad range of diseases but did not include content specific for known cancer loci. The OncoArray, thus, has broad value for studying cancer or related conditions. In addition, the platform allows additional content to be added so that other scientists or consortia such as Gliogene or Pancan could add content specific to their cancer types with minimal additional cost.

Impact of findings on prevention and treatment

We expect the discovery of novel genetic risk factors for cancer to provide insight into the genetic architecture of cancer and help elucidate its underlying biology. Providing a more comprehensive list of loci strongly associated with cancer susceptibility will greatly increase our knowledge of the pathophysiology of early stages in cancer development.

The clinical value of genetic testing for SNPs was questioned by some commentators because individual variants have limited power to discriminate cancer risk (53, 54). However, modeling studies show analysis with multiple variants provides discrimination in risk stratification sufficient to improve the efficiency of screening (55), as borne out by recent studies. For example, Pashayan and colleagues (56) showed that if prostate cancer screening were offered to men with a 10-year absolute risk of greater than 2% then risk stratification based on age and a 31-SNP polygenic risk score would result in 16% fewer men being eligible for screening than risk stratification based on age alone, but only 3% fewer cases would be detected (56). So and colleagues (57) showed that a polygenic risk model allows more precise enrollment of women according to age reducing the cost and burden of mammography. Given the expense and potential harms associated with prevention and early diagnosis (e.g., overdiagnosis and false positive findings) identifying those at highest risk might have important public health implications. These examples demonstrate the potential of genetic findings (58, 59) to impact public health and clinical care through the next several decades (60).

Gene–environment interactions

Several environmental and lifestyle risk factors, many of which are modifiable, such as obesity, physical activity, nonsteroidal anti-inflammatory drug (NSAID) use, hormone use, diet, smoking, and alcohol have been associated with various cancers. To fully understand the impact on the etiology of cancer, it is important to examine whether the genetic factors modify the effect of environmental factors. Recently, there has been extensive methodologic and applied work that provides a strong rationale for examining gene–environment interactions (GxE) interactions (8, 10–13, 61–65). The development of statistical methods for genome-wide GxE with increased power (66, 67) has led to detection of genetic variants whose effects are modified by environmental factors; and identification of variants that would have been missed through searches of marginal effects alone. As genetic profiles are fixed, modifying environmental exposures to alter deleterious effects of alleles remains the most viable preventive strategy. Importantly, even in the absence of GxE on the multiplicative scale, the absolute reduction in risk due to a change to a lower risk lifestyle is greater in those at higher genetic risk, making the development of tools to predict genetic risk a critical component of advice on lifestyle risks. In addition, the application of large-scale genetic testing of the same platform on a very large number of individuals permits an unprecedented opportunity for studying the impact that epistasis, interaction among loci, has upon risk for cancer development.

Functional characterization of risk loci

Perhaps the greatest challenge facing large collaborative genotyping projects such as the OncoArray is to understand of the functional mechanisms underlying disease development at each susceptibility locus. The pace of discovery of genetic risk associations for cancer and other traits and diseases continues to accelerate, creating an increasing bottleneck between discovery and functional validation. The basic tenets of functional characterization (68), proving causality for risk variants and the genes they regulate, have been described for a tiny fraction of risk associations identified by GWAS (20, 69). This is partly due to our rudimentary knowledge of the noncoding genome and the effects of genetic variation on gene regulation. Integration of GWAS SNP data with methylome data has identified methylation-quantitative trait loci (meQTL) showing that inherited genetic variation may affect carcinogenesis by regulating the human methylome (70, 71). The ENCODE (encyclopedia of DNA elements) consortium has cataloged genome-wide regulatory elements for many, but by no means all human tissues (72). Enhancers are often cell type–specific and drive the spatial and temporal diversity of gene expression in and across different cell types (73). One of the main challenges will therefore be to define the regulatory landscape for the relevant cell type for each trait-associated locus, followed by integration with genetic fine-mapping data to identify the most likely regulatory targets.

The ability to test the function of specific risk alleles has been enhanced by recent developments in genome editing, a powerful and highly efficient methodology for introducing DNA sequence alterations in human cells. Engineered nucleases (e.g., the CRISPR-Cas9 system) with customizable cleavage specificities can be used to introduce induce precise DNA base substitutions at the site of risk SNPs. The molecular and phenotypic effects of the different alleles of each risk SNP can then be evaluated *in vitro* or *in vivo*. The success of genome editing has been recently demonstrated for GWAS risk variants associated with fetal hemoglobin and prostate cancer (69, 74).

Complementary to genome editing for proving causality of risk SNPs is expression quantitative trait locus (eQTL) analysis to identify the likely target susceptibility gene (75, 76). eQTL analyses can interrogate both near and distant regulatory associations between risk genotypes and gene expression on the same chromos "ome (*cis-*) or across chromosomes *(trans-*). The role of these genes in neoplastic development can then be evaluated in experimental models of disease (77). Many groups have applied this concept to identify transcript expression correlated with trait-associated SNPs (78–80). For example, GAME-ON investigators have successfully used eQTL analysis to identify susceptibility genes at several breast, prostate, and ovarian cancer loci, and confirmed the significance of these genes through their functional analysis in disease models (42, 81, 82).

Disclosure of Potential Conflicts of Interest

M.T. Goodman is a consultant/advisory board member for Johnson and Johnson. R. Schmutzler has ownership interest (including patents). R.A. Eeles has received speakers bureau honoraria for ASCO-GU ASCO Meeting. No potential conflicts of interest were disclosed by the other authors.

Authors' Contributions

Conception and design: C.I. Amos, Z. Wang, F.R. Schumacher, S.A. Gayther, S.B. Gruber, A.M. Dunning, G.A. Coetzee, D.E. Goldgar, F.J. Couch, G.G. Giles, H. Bickeböller, Z. Kote-Jarai, P. Kraft, L.E. Kelemen, A. Berchuck, J.K. Field, A. Amin Al Olama, U. Peters, R.A. Eeles, E. Gillanders, D. Seminara, A.C. Antoniou, P.D.P. Pharoah, G. Chenevix-Trench, S.J. Chanock, J. Simard, D.F. Easton, K. Offit

Development of methodology: C.I. Amos, J. Dennis, Z. Wang, D.J. Hunter, T.A. Sellers, S.B. Gruber, A.M. Dunning, A. Lee, J.D. McKay, C.K. Edlund, M.F. Seldin, D. Seminara, A.C. Antoniou, J. Simard, K. Offit

Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): C.I. Amos, J. Dennis, F.R. Schumacher, S.A. Gayther, G. Casey, S.B. Gruber, A.M. Dunning, K. Doheny, A.B. Spurdle, J. Romm, E. Pugh, S.E. Bojesen, C. Luccarini, D. Vincent, D.J. Van Den Berg, D.E. Goldgar, F.J. Couch, G.G. Giles, H. Bickeböller, A. Risch, M. Waldenberger, I. Brüske-Hohlfeld, B.D. Hicks, H. Ling, P. Soucy, J. Manz, J.M. Cunningham, Z. Kote-Jarai, P. Kraft, L. FitzGerald, S. Lindström, M. Adams, C.M. Phelan, L.E. Kelemen, P. Brennan, Y.-Y. Shi, M.T. Goodman, S.F. Nielsen, A. Berchuck, S. Laboissiere, T. Shelford, J.A. Taylor, J.K. Field, S.K. Park, M. Thomassen, R. Schmutzler, L. Ottini, R.J. Hung, U. Peters, R.A. Eeles, A.C. Antoniou, S.J. Chanock, J. Simard, D.F. Easton, K. Offit

Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): C.I. Amos, J. Dennis, Z. Wang, J. Byun, F.R. Schumacher, S.A. Gayther, K. Michailidou, L. Fachal, Y. Li, X. Xiao, J. Romm, D.J. Hazelett, D.V. Conti, B.D. Hicks, H. Ling, A. Lee, K. Kuchenbaecker, Z. Kote-Jarai, P. Brennan, Y.-Y. Shi, D.J. Thompson, S.F. Nielsen, S.L. Schmit, C.K. Edlund, S.K. Park, J. Marchini, A. Amin Al Olama, R.A. Eeles, M.F. Seldin, A.C. Antoniou, J. Simard, D.F. Easton

Writing, review, and/or revision of the manuscript: C.I. Amos, J. Dennis, Z. Wang, F.R. Schumacher, S.A. Gayther, T.A. Sellers, S.B. Gruber, A.M. Dunning, K. Michailidou, L. Fachal, K. Doheny, A.B. Spurdle, E. Pugh, D.J. Hazelett, S.E. Bojesen, C.A. Haiman, A. Kamal, S. Nelson, D.E. Goldgar, F.J. Couch, G.G. Giles, H. Bickeböller, A. Risch, B.D. Hicks, P. Soucy, J.M. Cunningham, Z. Kote-Jarai, P. Kraft, L. FitzGerald, S. Lindström, C.M. Phelan, L.E. Kelemen, T.A. O'Mara, M.T. Goodman, S.F. Nielsen, A. Berchuck, S.L. Schmit, C.K. Edlund, J.A. Taylor, J.K. Field, M. Thomassen, R. Schmutzler, L. Ottini, R.J. Hung, A. Amin Al Olama, U. Peters, R.A. Eeles, M.F. Seldin, E. Gillanders, D. Seminara, P.D.P. Pharoah, G. Chenevix-Trench, J. Simard, D.F. Easton

Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): C.I. Amos, J. Dennis, D.J. Hazelett, C. Caga-Anan, A. Kamal, D. Tessier, F. Bacot, S. Nelson, S. Demetriades, F.J. Couch, J.L. Forman, B.D. Hicks, L. McGuffog, K. Butterbach, S. Benlloch, M. Riggan, T.A. O'Mara, S.F. Nielsen, A. Berchuck, C.K. Edlund, J.A. Taylor, R.J. Hung, R.A. Eeles, E. Gillanders, D. Seminara, A.C. Antoniou, J. Simard, K. Offit

Study supervision: C.I. Amos, Z. Kote-Jarai, P. Kraft, A. Berchuck, J.A. Taylor, R.J. Hung, R.A. Eeles, A.C. Antoniou, J. Simard

Other (PI of one of the centers): T.A. Sellers

Other (coordination of the Endometrial Cancer Consortium): A.B. Spurdle

Other (performed genotyping): J. Manz

Acknowledgments

The authors are very grateful to have worked with Brian E. Henderson, M.D. The development and completion of the OncoArray and the organization of this worldwide resource was a vision that we are proud to have completed, but sadly after his death on June 20, 2015. We thank Lamar D. Moss for his help in organizing this manuscript submission.

Grant Support

This work was supported by TRICL (Transdisciplinary Research for Cancer of Lung) and International Lung Cancer Consortium (ILCCO) by NIH grant U19 CA148127-01 (to C.I. Amos, J. Byun, Y. Li, X. Xiao, J. L. Forman, A. Risch, H. Bickeböller, M. Waldenberger, I. Brüske, J. Manz, P. Brennan, R. Hung, H. Shen, Y. Shi, A. Kamal, J.K. Field), DRIVE (Discovery, Biology, and Risk of Inherited Variants in Breast Cancer) by NIH grant U19 CA148065 (to D.J. Hazelett, D. F. Easton, S. Lindström, P. Kraft, J. Dennis, A. Dunning, K. Michailidou, L. Fachal, S. Benlloch, J. Cunningham, K. Butterbach, K. Offit, R. Schmutzler, L. Ottini, D. Vincent), CORECT (ColoRectal Transdisciplinary Study) by NIH grants U19 CA148107, R01 CA81488, P30 CA014089 (to S. Gruber, U. Peters, G. Casey), ELLIPSE (ELLIPSE, Elucidating Loci in Prostate Cancer Susceptibility) supported by U19 CA148537 (to F. Schumacher, S.A. Gayther, S.L. Schmit, C.K. Edlund, D.J. Hazelett, G.A. Coetzee, C. Haiman, S. Demetriades, D. Van Den Berg); FOCI (Follow-up of Ovarian Cancer Genetic Association & Interaction Studies) by NIH grant U19-CA148112 (to T.A. Sellers, P. Pharoah, S. Gayther, A. Monteiro, J. Schildkraut, E. Iversen) Canadian Cancer Society Research Institute (grant no. 020214; to R. Hung), R01-CA122443 (to B. Hicks) P50-CA116201, CA192393 (to F. Couch), P50-CA136393 (to J. Cunningham, F. Couch), OCAC-CA-149429 (to C. Phelan), P30-CA15083 (to B. Hicks), R01 CA 136924 and R01 CA 190182 (to G.A. Coetzee); Cancer Research UK (C490/A8339; to A. Antoniou, P. Pharoah, A. Dunning), (C1287/A16563 and C1287/A10118) to D.F. Easton, C490/A16561 (to P. Pharoah, A. Dunning), C490/A10119 (to A. Dunning, P. Pharoah), C490/A10124 (to A. Dunning, P. Pharoah); ASTERISK: a Hospital Clinical Research Program (PHRC) and supported by the Regional Council of Pays de la Loire, the Groupement des Entreprises Françaises dans la Lutte contre le Cancer (GEFLUC), the Association Anne de Bretagne Génétique, and the Ligue Régionale Contre le Cancer (LRCC; to F. Bacot); DACHS: German Research Council (Deutsche Forschungsgemeinschaft, BR 1704/6-1, BR 1704/6-3, BR 1704/6-4, and CH 117/1-1; to C.K. Edlund), and the German Federal Ministry of Education and Research (01KH0404 and 01ER0814). Prostate Lung Colon Ovarian Screening trial: Intramural Research Program of the Division of Cancer Epidemiology and Genetics and supported by contracts from the Division of Cancer Prevention, National Cancer Institute, NIH, DHHS. The GC-HBOC (German Consortium of Hereditary Breast and Ovarian Cancer) is supported by the German Cancer Aid (grant no 110837, coordinator: to R. Schmutzler, Cologne). The PERSPECTIVE (J. Simard) project was supported by the Government of Canada through Genome Canada and the Canadian Institutes of Health Research (MOP-86727, to L.E. Kelemen), the Ministère de l'Économie, Innovation et Exportation du Québec through Genome Québec, and the Quebec Breast Cancer Foundation, including additional funding from the Canadian Breast Cancer Foundation and the National Cancer Institute through Breast Cancer Family Registry Cohort (#1UM1CA164920-01A1); European Union Framework Programme for Research and Innovation (MSCA-IF-2014-EF-656144; to L. Fachal); The Breast Cancer Research Foundation, P30 CA08748, Andrew Sabin Foundation (to K. Offit). CIDR genotyping for the Oncoarray was conducted under contract 268201200008I (to K. Doheny) and through grant 1×01HG007491-01 (to C.I. Amos). The Norris Cotton Cancer Center (P30CA023108), The Quantitative Biology Research Institute (P20GM103534), and the Coordinating Center for Screen Detected Lesions (U01CA196386) also supported efforts of C.I. Amos.

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 February 1, 2016; revised June 30, 2016; accepted July 29, 2016; published OnlineFirst October 3, 2016.

References

1. International WCRF. Lung cancer statistics; 2012 [cited 2015 May 2]. Available from: http://www.wcrf.org/int/cancer-facts-figures/data-specific-cancers/lung-cancer-statistics.
2. Chakravarti A. Population genetics–making sense out of sequence. Nat Genet 1999;21:56–60.
3. Lander ES, Schork NJ. Genetic dissection of complex traits. Science 1994; 265:2037–48.
4. Reich DE, Lander ES. On the allelic spectrum of human disease. Trends Genet 2001;17:502–10.
5. Houlston RS, Peto J. The search for low-penetrance cancer susceptibility alleles. Oncogene 2004;23:6471–6.
6. Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI. Shifting paradigm of association studies: value of rare single nucleotide polymorphisms. Genet Epidemiol 2007;31:608.
7. Zhu Q, Ge D, Maia JM, Zhu M, Petrovski S, Dickson SP, et al. A genome-wide comparisonof the functional properties of rare and common genetic variants in humans. Am J Hum Genet 2011;88:458–68.
8. Ponder BA. Inherited predisposition to cancer. Trends Genet 1990; 6:213–8.
9. Ponder BA. Cancer genetics. Nature 2001;411:336–41.
10. Peto J. Cancer epidemiology in the last century and the next decade. Nature 2001;411:390–5.
11. Hunter DJ. Gene-environment interactions in human diseases. Nat Rev Genet 2005;6:287–98.
12. Potter JD. Colorectal cancer: molecules and populations. J Natl Cancer Inst 1999;91:916–32.
13. Thomas DC. Statistical methods in genetic epidemiology. New York, NY: Oxford University Press; 2004.
14. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 2009;106:9362–7.
15. Qian DC, Byun J, Han Y, Hunter DJ, Henderson BE, Eeles R, et al. Identification of shared and unique susceptibility pathways among cancers of the lung, breast, and prostate from genome-wide association studies and tissue-specific protein interactions. Hum Mol Genet 2015;24:7406–20.
16. Smith GD, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? Int J Epidemiol 2003;32:1–22.
17. Maurano MT, Humbert R, Rynes E, Thurman RE, Haugen E, Wang H, et al. Systematic localization of common disease-associated variation in regulatory DNA. Science 2012;337:1190–5.
18. Coetzee SG, Shen HC, Hazelett DJ, Lawrenson K, Kuchenbaecker K, Tyrer J, et al. Cell-type-specific enrichment of risk-associated regulatory elements at ovarian cancer susceptibility loci. Hum Mol Genet 2015;24:3595–607.
19. Hazelett DJ, Rhie SK, Gaddis M, Yan C, Lakeland DL, Coetzee SG, et al. Comprehensive functional annotation of 77 prostate cancer risk loci. PLoS Genet 2014;10:e1004102.
20. Smemo S, Tena JJ, Kim KH, Gamazon ER, Sakabe NJ, Gomez-Marin C, et al. Obesity-associated variants within FTO form long-range functional connections with IRX3. Nature 2014;507:371–5.
21. Ahmadiyeh N, Pomerantz MM, Grisanzio C, Herman P, Jia L, Almendro V, et al. 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC. Proc Natl Acad Sci USA 2010;107:9742–6.
22. Spurdle AB, Thompson DJ, Ahmed S, Ferguson K, Healey CS, O'Mara T, et al. Genome-wide association study identifies a common variant associated with risk of endometrial cancer. Nat Genet 2011;43:451–4.
23. Painter JN, O'Mara TA, Batra J, Cheng T, Lose FA, Dennis J, et al. Fine-mapping of the HNF1B multicancer locus identifies candidate variants that mediate endometrial cancer risk. Hum Mol Genet 2015;24:1478–92.
24. Shen H, Fridley BL, Song H, Lawrenson K, Cunningham JM, Ramus SJ, et al. Epigenetic analysis leads to identification of HNF1B as a subtype-specific susceptibility gene for ovarian cancer. Nat Commun 2013; 4:1628.
25. Soslow RA. Histologic subtypes of ovarian carcinoma: an overview. Int J Gynecol Pathol 2008;27:161–74.
26. Pearce CL, Templeman C, Rossing MA, Lee A, Near AM, Webb PM, et al. Association between endometriosis and risk of histological subtypes of ovarian cancer: a pooled analysis of case-control studies. Lancet Oncol 2012;13:385–94.
27. The Cancer Genome Atlas Research Network, Kandoth C, Schultz N, Cherniack AD, Akbani R, Liu Y, et al. Integrated genomic characterization of endometrial carcinoma. Nature 2013;497:67–73.
28. Thompson DJ, O'Mara TA, Glubb DM, Painter JN, Cheng T, Folkerd E, et al. CYP19A1 fine-mapping and Mendelian randomization: estradiol is causal for endometrial cancer. Endocr Relat Cancer 2016;23:77–91.
29. O'Mara TA, Glubb DM, Painter JN, Cheng T, Dennis J, Australian National Endometrial Cancer Study Group, et al. Comprehensive genetic assessment of the ESR1 locus identifies a risk region for endometrial cancer. Endocr Relat Cancer 2015;22:851–61.
30. Bahcall OG. iCOGS collection provides a collaborative model. Foreword. Nat Genet 2013;45:343.
31. Bojesen SE, Pooley KA, Johnatty SE, Beesley J, Michailidou K, Tyrer JP, et al. Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. Nat Genet 2013;45:371–84.
32. Eeles RA, Olama AA, Benlloch S, Saunders EJ, Leongamornlert DA, Tymrakiewicz M, et al. Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. Nat Genet 2013;45:385–91.
33. Meeks HD, Song H, Michailidou K, Bolla MK, Dennis J, Wang Q, et al. BRCA2 polymorphic stop codon K3326X and the risk of breast, prostate, and ovarian cancers. J Natl Cancer Inst. 2015;108(2): pii: djv315.
34. Timofeeva MN, Hung RJ, Rafnar T, Christiani DC, Field JK, Bickeboller H, et al. Influence of common genetic variation on lung cancer risk: meta-analysis of 14 900 cases and 29 485 controls. Hum Mol Genet 2012;21:4980–95.
35. Wang Y, McKay JD, Rafnar T, Wang Z, Timofeeva MN, Broderick P, et al. Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. Nat Genet 2014;46:736–41.
36. Wang Y, Wei Y, Gaborieau V, Shi J, Han Y, Timofeeva MN, et al. Deciphering associations for lung cancer risk through imputation and analysis of 12 316 cases and 16 831 controls. Eur J Hum Genet 2015;23:1723–8.
37. 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.
38. Kachuri L, Amos CI, McKay JD, Johansson M, Vineis P, Bueno-de-Mesquita HB, et al. Fine mapping of chromosome 5p15.33 based on a targeted deep sequencing and high density genotyping identifies novel lung cancer susceptibility loci. Carcinogenesis 2016;37:96–105.

39. Liu Y, Kheradmand F, Davis CF, Scheurer ME, Wheeler D, Tsavachidis S, et al. Focused analysis of exome sequencing data for rare germline mutations in familial and sporadic lung cancer. J Thorac Oncol 2016;11:52–61.
40. Hindorff LA, MacArthur J, Morales J, Junkins HA, Hall PN, Klemm AK, et al. A catalog of published genome-wide association studies [cited 2016 May 2]. Available from: https://www.genome.gov/26525384/catalog-of-published-genomewide-association-studies/.
41. Earp M, Winham SJ, Larson N, Permuth JB, Sicotte H, Chien J, et al. A targeted genetic association study of epithelial ovarian cancer susceptibility. Oncotarget 2016;7:7381–9.
42. Lawrenson K, Li Q, Kar S, Seo JH, Tyrer J, Spindler TJ, et al. Cis-eQTL analysis and functional validation of candidate susceptibility genes for high-grade serous ovarian cancer. Nat Commun 2015;6:8234.
43. Pharoah PD, Tsai YY, Ramus SJ, Phelan CM, Goode EL, Lawrenson K, et al. GWAS meta-analysis and replication identifies three new susceptibility loci for ovarian cancer. Nat Genet 2013;45:362–70.
44. Michailidou K, Hall P, Gonzalez-Neira A, Ghoussaini M, Dennis J, Milne RL, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nat Genet 2013;45:353–61.
45. Burdett T, Hall PN, Hastings E, Hindorff LA, Junkins HA, Klemm AK, et al. The NHGRI-EBI Catalog of published genome-wide association studies. Available from: www.ebi.ac.uk/gwas.
46. Gudmundsson J, Sulem P, Steinthorsdottir V, Bergthorsson JT, Thorleifsson G, Manolescu A, et al. Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. Nat Genet 2007;39:977–83.
47. Yang J, Ferreira T, Morris AP, Medland SE, Genetic Investigation of Anthropometric Traits (GIANT) Consortium; DIAbetes Genetics Replication And Meta-analysis (DIAGRAM) Consortium, et al. Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet 2013;44:369–75.
48. Cousminer DL, Stergiakouli E, Berry DJ, Ang W, Groen-Blokhuis MM, Korner A, et al. Genome-wide association study of sexual maturation in males and females highlights a role for body mass and menarche loci in male puberty. Hum Mol Genet 2014;23:4452–64.
49. Wheeler HE, Maitland ML, Dolan ME, Cox NJ, Ratain MJ. Cancer pharmacogenomics: strategies and challenges. Nat Rev Genet 2013;14:23–34.
50. Li Y, Byun J, Cai G, Xiao X, Han Y, Cornelis O, et al. FastPop: a rapid principal component derived method to infer intercontinental ancestry using genetic data. BMC Bioinformatics 2016;17:122.
51. Voight BF, Kang HM, Ding J, Palmer CD, Sidore C, Chines PS, et al. The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. PLoS Genet 2012;8:e1002793.
52. Cortes A, Brown MA. Promise and pitfalls of the Immunochip. Arthritis Res Ther 2011;13:101.
53. Holtzman NA, Marteau TM. Will genetics revolutionize medicine? N Engl J Med 2000;343:141–4.
54. Vineis P, Schulte P, McMichael AJ. Misconceptions about the use of genetic tests in populations. Lancet 2001;357:709–12.
55. Pharoah PD, Antoniou A, Bobrow M, Zimmern RL, Easton DF, Ponder BA. Polygenic susceptibility to breast cancer and implications for prevention. Nat Genet 2002;31:33–6.
56. Pashayan N, Duffy SW, Neal DE, Hamdy FC, Donovan JL, Martin RM, et al. Implications of polygenic risk-stratified screening for prostate cancer on overdiagnosis. Genet Med 2015;17:789–95.
57. So HC, Kwan JS, Cherny SS, Sham PC. Risk prediction of complex diseases from family history and known susceptibility loci, with applications for cancer screening. Am J Hum Genet 2011;88:548–65.
58. Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. Science 2008;322:881–8.
59. Hunter DJ, Altshuler D, Rader DJ. From Darwin's finches to canaries in the coal mine–mining the genome for new biology. N Engl J Med 2008; 358:2760–3.
60. Lander ES. Initial impact of the sequencing of the human genome. Nature 2011;470:187–97.
61. Collins FS, Manolio TA. Merging and emerging cohorts: necessary but not sufficient. Nature 2007;445:259.
62. Khoury MJ, Wacholder S. Invited commentary: from genome-wide association studies to gene-environment-wide interaction studies–challenges and opportunities. Am J Epidemiol 2009;169:227–30.
63. Manolio TA, Collins FS. Genes, environment, health, and disease: facing up to complexity. Hum Hered 2007;63:63–6.
64. Thomas D. Gene-environment-wide association studies: emerging approaches. Nat Rev Genet 2010;11:259–72.
65. Thomas D. Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies. Annu Rev Public Health 2010;31:21–36.
66. Evans DM, Marchini J, Morris AP, Cardon LR. Two-stage two-locus models in genome-wide association. PLoS Genet 2006;2:e157.
67. Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. Nat Genet 2005; 37:413–7.
68. Freedman ML, Monteiro AN, Gayther SA, Coetzee GA, Risch A, Plass C, et al. Principles for the post-GWAS functional characterization of cancer risk loci. Nat Genet 2011;43:513–8.
69. Spisak S, Lawrenson K, Fu Y, Csabai I, Cottman RT, Seo JH, et al. CAUSEL: an epigenome- and genome-editing pipeline for establishing function of noncoding GWAS variants. Nat Med 2015;21:1357–63.
70. Shi J, Marconett CN, Duan J, Hyland PL, Li P, Wang Z, et al. Characterizing the genetic basis of methylome diversity in histologically normal human lung tissue. Nat Commun 2014;5:3365.
71. Scherf DB, Sarkisyan N, Jacobsson H, Claus R, Bermejo JL, Peil B, et al. Epigenetic screen identifies genotype-specific promoter DNA methylation and oncogenic potential of CHRNB4. Oncogene 2013;32:3329–38.
72. Thurman RE, Rynes E, Humbert R, Vierstra J, Maurano MT, Haugen E, et al. The accessible chromatin landscape of the human genome. Nature 2012;489:75–82.
73. Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature 2009;459:108–12.
74. Bauer DE, Kamran SC, Lessard S, Xu J, Fujiwara Y, Lin C, et al. An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. Science 2013;342:253–7.
75. Veyrieras JB, Kudaravalli S, Kim SY, Dermitzakis ET, Gilad Y, Stephens M, et al. High-resolution mapping of expression-QTLs yields insight into human gene regulation. PLoS Genet 2008;4:e1000214.
76. Stranger BE, Montgomery SB, Dimas AS, Parts L, Stegle O, Ingle CE, et al. Patterns of cis regulatory variation in diverse human populations. PLoS Genet 2012;8:e1002639.
77. Burke W, Laberge AM, Press N. Debating clinical utility. Public Health Genomics 2010;13:215–23.
78. Kwan T, Grundberg E, Koka V, Ge B, Lam KC, Dias C, et al. Tissue effect on genetic control of transcript isoform variation. PLoS Genet 2009;5: e1000608.
79. Lalonde E, Ha KC, Wang Z, Bemmo A, Kleinman CL, Kwan T, et al. RNA sequencing reveals the role of splicing polymorphisms in regulating human gene expression. Genome Res 2011;21:545–54.
80. Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. Mapping complex disease traits with global gene expression. Nat Rev Genet 2009;10:184–94.
81. Li Q, Seo JH, Stranger B, McKenna A, Pe'er I, Laframboise T, et al. Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. Cell 2013;152:633–41.
82. Grisanzio C, Werner L, Takeda D, Awoyemi BC, Pomerantz MM, Yamada H, et al. Genetic and functional analyses implicate the NUDT11, HNF1B, and SLC22A3 genes in prostate cancer pathogenesis. Proc Natl Acad Sci USA 2012;109:11252–7.