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

Authors:
Christopher Amos
Joe Dennis
Zhaoming Wang
Jinyoung Byun
Fredrick Schumacher
Simon A. Gayther
David J. Hunter
Brian Henderson*
Thomas A. Sellers
Stephen Gruber
Alison Dunning
Kyriaki Michailidou
Laura Fachal
Kimberly Doheny
Amanda B. Spurdle
Yafang Li
Xiangjun Xiao
Jane Romm
Elizabeth Pugh
Gerhald A. Coetzee
Dennis J. Hazelett
Stig Bojesen
Charlisse Caga-Anan
Christopher Haiman
Ahsan Kamal
Craig Luccarini
Daniel Tessier
Daniel Vincent
François Bacot
David Van Den Berg
Stephen Demetriades
Fergus Couch
Judith L. Forman
Graham Giles
David Conti
Heike Bickeböller
Angela Risch
Melanie Waldenberger
Irene Brüske-Hohlfeld
Belynda Hicks
Hoai-Thu Vo
Hua Ling
Lesley McGuffog
Andrew Lee
Karoline Kuchenbaecker
Penny Soucy
Judith Manz
Julie Cunningham

Katja Butterbach
Zsofia Kote-Jarai
Peter Kraft
Liesel M. Fitzgerald
Sara Lindström
Marcia Adams
James McKay
Catherine Phelan
Sara Benlloch
Paul Brennan
Hongbin Shen
Yongyong Shi
Sune F. Nielsen
Sylvie Laboissiere
Tameka Shelford
Jack Taylor
John K. Field
Sue Park
Mads Thomassen
Ken Offit
Rita Schmutzler
Laura Ottini
Rayjean Hung
Jonathan Marchini
Ulrike Peters
Rosalind Eeles
Michael Seldin
Elizabeth Gillanders
Stefanie Nelson
Daniela Seminara
Antonis C. Antoniou
Paul Pharoah
Georgia Chenevix-Trench
Stephen Chanock
Jacques Simard
Douglas F Easton

Representing the GAME-ON, ECAC, GLC, CIMBA, PERSPECTIVE and PRACTICAL consortia.

The authors of this manuscript do not have any conflicts of interest to disclose and have not received commercial reimbursement for any aspect of the research reported in this manuscript. Funding sources for the research are disclosed in the supplementary funding sources document and include government and nonprofit organizations only.

*Post-humous

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 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 423,029 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.

Introduction

Cancer is one of the leading causes of death world-wide. 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 (http://www.wcrf.org/cancer_statistics/world_cancer_statistics.php). 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.[1] 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.[2,3] 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[4-7], with additional risk due to uncommon alleles that may have higher penetrance remaining to be discovered. By developing large consortia, the effects of these rare alleles can be estimated [8,9]. As pointed out by Ponder [10,11] and Peto,[12] 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 between populations[10-15].

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 and provided new insights into etiologic mechanisms that suggest interventions [16] By identifying many new loci influencing cancer development, genomic research has identified pathways that influence cancer development[17]. In addition, Mendelian randomization has emerged as an effective approach for confirming non-genetic etiologic factors identified through epidemiologic studies, removing potential concerns about reverse causality[18].

Once the loci are identified, fine-mapping studies are a critical next step in narrowing in on the underlying functional variant(s) and in the discovery of nearby, independent, secondary signals, which may increase significantly the heritable fraction explained by each region. Furthermore, fine-mapping studies can aid in functional follow-up, by defining the most likely candidate variant(s). More than 90% of risk-alleles lie in non-protein coding DNA and there is now unequivocal evidence that risk regions are enriched for regulatory elements, including enhancers, promoters, insulators and silencers[19]. In general genome-wide estimates in humans indicate about 500,000 enhancers may alter regulation of expression thus alter risk by controlling expression of target susceptibility genes[19-22]. Analyses to date indicate that several regions harbor multiple distinct susceptibility variants for different cancer types, suggesting common mechanisms but tissue-specific regulation[23]. Thus fine-mapping of multiple cancer types using a 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, in particular through the development 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 endometrial cancer 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, http://epi.grants.cancer.gov/gameon/) 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. By designing a common array for multiple cancers, we were able to achieve economies of scale and hence genotype a large number of study subjects from many countries and ethnic backgrounds. Additionally, 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.

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[17] 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 Table 1. Endometrial cancer cases were also included as a part of the genotyping study for several reasons. First, endometrial cancer shares several risk factors with breast cancer and ovarian cancer. Second, there is at least one genetic locus (*HNF1B*) shared by endometrial cancer[24,25], prostate cancer and ovarian cancer[26], providing a rationale for exploring additional common susceptibility across other cancer sites. Finally, there are similarities in tumor phenotype and/or shared tissue of origin between endometrial cancer, the benign gynaecological condition endometriosis, the endometrioid and clear cell histologies of ovarian cancer, and basal-like breast cancer [27-29]. Thus, pooling ovarian and endometrial[25,30,31] 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. Approximately 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; this set of markers has been incorporated into several other arrays and hence were expected to genotype successfully. The remaining markers were selected from seven lists: five from the disease consortia representing the main cancer sites, one from the CIMBA consortium 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. More detailed descriptions of the SNP selection process for disease sites participating in the Oncoarray are also provided in the Supplementary Methods. Below, we present the general approaches that were taken for nominating SNPs for the Array.

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 1) 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); 2) Fine mapping of risk loci based on 1000 Genomes Project data and resequencing studies; 3) Candidate rare variants from whole genome and whole exome studies, and exome arrays; 4) findings from previously published studies of other cancers 5) other "wild-card" variants, for example variants of potential functional significance. 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 non-coding, 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 1Mb 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 500kb either side of each hit. Common regions were defined as regions including hits within 1Mb 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 (a) all variants correlated with the known hits at $r^2>0.6$ (b) all variants from lists of potentially functional variants, defined through RegulomeDB and (c) a set of SNPs designed to tag all remaining variants at $r^2>0.9$.

Selection of "Common" SNPs

Previous analyses[32-35] 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 1Mb (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 (http://www.ebi.ac.uk/gwas/) iii) had been found to be relevant to cancer associated traits[36] including BMI, height, and waste to hip ratio (in collaboration with the GIANT consortium[37]), smoking, age at menopause or menarche (in collaboration with the REPROGEN consortium[38]), and telomere length in lymphocytes[39]. 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, iii) SNPs nominated by PharmGKB and variants nominated from study of cell lines to affect expression of pharmacogenetically relevant genes[40]. SNPs from the region of chromosome 15q25.1 that associates 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 (non-polymorphic) probes for each exon of *BRCA1*, *BRCA2*, *MLH1* and *MSH2* in order 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's Lymphoma), esophageal, gastric, glioma, kidney, osteosarcoma, pancreas, testis or scan data for non-Caucasian studies including Asian non-smoking 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 Figure 1.

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 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 2.

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:

a. Include the GWAS backbone
b. Include the Common fine-mapping list
c. Choose the remaining SNPs iteratively from the five ranked lists. At each stage chose 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 ensures that the correct proportions will be kept.
d. Include the SNP unless the exact SNP has already been chosen. In either case, augment the count n for that list by 1.
e. Increase the number of beadtypes for chosen SNPs, where necessary because variation could not be captured by a single beadtype.

Based on the merged list of 715,637 unique SNPs (76,290 from lung; 224,074 from from familial and sporadic breast and ovarian; 81,009 from prostate; 50,110 from colorectal; 17,547 from common list), we further performed the LD pruning (r^2>0.95). It 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. Genotyping results were stored in 'top' format because that provided a unique algorithm for SNP genotype labeling, and a strand alignment file was developed to permit realignment to the strand forward direction for imputation and final reporting of results.

Quality control steps

A detailed quality control plan was developed and is included as supplementary material but salient features are presented here. Additionally, an imputation guideline is also presented. Participating sites genotyped a common set of Hapmap samples so that strand alignment and integrity of imputation could be compared among analytical sites. All sites extracted genotypes in top alignment and used a common genotype clustering file that can be downloaded from http://consortia.ccge.medschl.cam.ac.uk/oncoarray/onco_v2c.zip. A list (onco_duplicate_variants_excluded.csv) of 765 was compiled of duplicate probes that should be excluded. The probe with lower call rate was excluded.

Reclustering process
A selection of 56,284 samples with high call rates from across the genotyping centres 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. (Initial analyses found 23,249 variants with a MAF below 0.0005.)

Variants likely to have problematic clusters were selected for manual inspection using these criteria: call rate below 99%, very rare variants (MAF below 0.001), poor Illumina intensity and clustering metrics, deviation from the expected frequency as observed in the 1000 Genomes. We inspected 68,000 cluster plots and 3,964 variants were identified where the cluster positions could be manually improved from the initial cluster file. The final cluster file with the manual adjustments was distributed and applied throughout the consortium. Plots initially scored as failed were inspected by a second analyst and 16,526 variants were excluded from the analysis.

Filtering of samples and genotypes were performed separately by consortium. We excluded samples with call rates <80% then removed SNPs with call rates less than 80%, then excluded samples with call rates <95%. We also excluded unexpected genotypic males/females/males (using X and Y markers). Individuals with identified as XO, XXY, or with low X heterozygosity (<5%) were flagged for exclusion. A list of 300 Y markers confirmed to genotype well in males and to have non-autosomal cluster patterns were used for gender checking (chr_Y_SNPs_for_sex_checking.csv). We exclude from the test chromosome SNPs that show a high level of heterozygous calls in males and/or autosomal cluster patterns (chr_X_SNPs_with_autosomal_clusters.csv).

Ancestry Analysis

Ancestry analysis was performed using a standardized approach in which 2,318 ancestry informative markers (AIMs) with minor allele frequencies of 0.05 or higher were analyzed on data from 66,610 samples including 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, in order to estimate the contribution of European, Asian and African ancestry to each individual. The method is further described in the software package FastPop (http://sourceforge.net/projects/fastpop/) that we developed and distributed to consortium members. Individuals were thus classified into 4 groups for downstream analyses: European (defined as >80% European ancestry), East Asian (>40% Asian ancestry), African (>20% African ancestry) and other (not fulfilling any of the above criteria). Any markers showing deviation from Hardy-Weinberg equilibrium with $P<10^{-7}$ in controls or $P<10^{-12}$ in cases were flagged for exclusion from imputation analyses and for further review of cluster plots. Within ethnic groups, samples with overall heterozygosity <5% or > 40% were excluded.

Additional Quality Control Steps
Duplicate checking was assessed using PLINK[41] or genabel (http://www.genabel.org/). Unexpected duplicates that could not be resolved were removed, while for resolved duplicates, the sample with the higher call rate was

retained for analysis. One individual from any group (usually a pair) with estimated identity by descent values of 0.45 or higher was retained for primary case-control comparisons. Genotypes showing 2% or higher discordance in duplicate samples were removed from consortium-specific analyses and flagged for exclusion from imputations.

Prior to imputation, a reduced set of SNPs was selected to insure the same high quality SNPs were analyzed across all consortia. Variants that had call rates below 98% or MAF <0.01 in Europeans in any consortium were excluded. Strand information was obtained by blasting the Illumina TOP sequences against the 1000 genomes sequences to convert to a consistent forward alignment. Some manifest positions identified by "rs" numbers were updated from dbSNP and the new positions confirmed by sequence matching. The variants on the chip were then matched to the variants from the 1000 Genomes Phase 3 release variant set provided for the Impute software:
https://mathgen.stats.ox.ac.uk/impute/1000GP%20Phase%203%20haplotypes%206%20October%202014.html.

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 (Figure 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, while amplified DNA derived from non-blood samples showed poorer performance. The success rate for genotyping Hapmap derived samples was 100% and the overall genotyping failure rate for lymphoblastoid lines was 0.5%. To evaluate the reliability of genotyping across samples including post-imputation processing we evaluated concordance of imputed genotypes (?for Hapmap samples – needs to be specified) among the centers. Results show > 95% concordance in imputed genotype calls for all Hapmap samples with lower levels of concordance found among non-European descent samples.

Analysis of concordance of sample genotypes

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).

To illustrate the performance of the imputation, we performed genome-wide imputation for 69,900 cases and 51,056 controls of European ancestry from the breast cancer data set. These were imputed using v3 of the 1000 Genomes Project as a reference panel, resulting in the imputation of ~22M SNPs with a minor allele frequency of >0.1% in European samples. Imputation was performed using Impute v2.0 after prephasing of genotypes using SHAPEIT2. Imputation was carried out in ~600 5Mb sections, with the number of contributing haplotypes (k_hap) set to 800. Imputation quality was extremely high for common variants: more than 65% of variants with MAF>5% had an imputation quality score >0.975 and 93% had a quality score >0.8 (Figure 4, panel a). As expected, this proportion was lower for rarer SNPs (9% for variants with a MAF<5% and 4% for variants with a MAF<1% had a quality score >0.975) (Figure 4, panel b). However, even for rarer variants a substantial fraction could be imputed, albeit less reliably. Thus the proportions of SNPs with a quality score

>0.3 were 81% and 76% for SNPs with MAF<5% and <1%, respectively. Supplementary Figure 1 compares the imputation accuracy of the Oncoarray to several other arrays.

Genotyping results

The populations that have been genotyped as a part of the Oncoarray are presented in Supplementary Table 1. This table provides a description of the design of the studies that are participating in the Oncoarray along with the reported ethnic background of the participating studies. Samples that were genotyped at the Center for Inherited Disease Research will be available for analysis through the dbGAP portal in March, 2016. Data from other samples along with more detailed phenotyping data are available through collaborative requests to the participating consortia. Websites that provide details about the process for obtaining genotyping information are available for lung cancer at the Transdisciplinary Research in Cancer of the Lung website (www.u19tricl.org), for prostate cancer through PRACTICAL (http://practical.ccge.medschl.cam.ac.uk/), for Breast Cancer at BCAC (http://apps.ccge.medschl.cam.ac.uk/consortia/bcac/), for Ovarian Cancer at OCAC (http://apps.ccge.medschl.cam.ac.uk/consortia/ocac//), for colon cancer at CORECT (http://epi.grants.cancer.gov/gameon/), for endometrial cancer at (http://apps.ccge.medschl.cam.ac.uk/consortia/ocac/ http://epi.grants.cancer.gov/eecc/), and for *BRCA1* and *BRCA2* mutation carriers at CIMBA (http://apps.ccge.medschl.cam.ac.uk/consortia/cimba/). In total after all quality control exclusions there are 494,763 SNPs that were retained for analysis.

To characterize the continental ancestries of individuals studied by the Oncoarray we applied Fastpop to 66,105 samples genotyped at CIDR, Cambridge and Genome Quebec/McGill Innovation Center (the primary contributing centers) and Hapmap samples with 2,318 intercontinental ancestry informative markers. The 66,105 samples were divided into 70% (46,274) as discovery set and 30% (19,831) as validation set. To compute SNP weights for prediction of scores in Principal Component Analysis, 46,274 discovery samples (70% out of 66,105 TRICL and UK samples) and 505 Hapmap2 samples were combined. The R-scripts for prediction of scores using 2318 SNP weights, PCAScore is available at https://morgan1.dartmouth.edu/~jbyun/Software/PCAScore_R/.

To build a model for intercontinental ancestry analysis, we began with a sample of 51,987 with 95% call rate or higher. Hapmap2 samples include three continental ancestry definitions from CEU, CHB, and YRI as European, Asian, and African-American, respectively. Using the pre-calculated SNP weights with the same 2,318 AIMs as in the discovery samples, we predicted the scores of 51,987 samples. To calculate each inference of individual ancestry membership among three continental ethnicities, first we computed each continental centroid from Hapmap2 samples and then performed a distance-based approach in the triangular region, "InterContinentalDistanceMetrics.R" using the R-package FastPop (http://sourceforge.net/projects/fastpop/). As shown in Supplementary Figure 2, using these definitions led to no samples being assigned to multiple continental origins.

Discussion

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. This is only one of the first steps towards the overarching goal of improving prevention and therapy, but it is a critical step. While most GWAS studies have only been completed within the last 5-10 years, the potential of these findings can already be demonstrated by several examples. In the case of Crohn's disease, GWAS loci pointed to previously unappreciated physiologic processes, such as autophagy, innate immunity, and IL-23R signaling.[42-44] These discoveries have already led to chemical screens for candidate therapeutic agents.[44-46] For age-related macular degeneration, GWAS identified several genes involved in inflammation, a link that had not been established before and has now opened up new treatment approaches and even prevention strategies.[47,48] Identifying the genetic basis of several

Mendelian disorders has led to the development of FDA-approved drugs[49]. Also, GWAS findings are being increasingly used for drug repositioning, whereby existing FDA-approved drugs are shown to influence key pathways influencing disease susceptibility[50-52]. For cancer studies, the availability of results from the Cancer Genome Atlas (TCGA) provides a unique opportunity to begin to explore the relationships between the changes that exist in tumor samples versus the variants that influence cancer susceptibility[53]. The integration of these two sources of data provides an opportunity to identify drivers of cancer development such as APC and RB that play major roles in both the initial development of cancer and also play a role in cancer growth. These and other examples that have led to new therapies and impacted medical practice [54,55] demonstrate the enormous potential of genetic findings.[56,57] However, as drug development takes years (initial findings must be followed by clinical studies testing efficacy and effectiveness) [58,59 60], it is likely that we have barely begun to see the full extent of the impact of the discovery of the 665 novel cancer susceptibility loci that have already been identified (query from the GWAS catalogue of all cancers excluding recurrence or relapse http://www.ebi.ac.uk/gwas accessed 1/6/2016). In summary, providing a more comprehensive list of genes strongly associated with cancer susceptibility will greatly increase opportunities to identify new targets for drug development. Further, the integration of carefully harmonized epidemiologic data with tumor and germline genetic data will allow the investigation of the biological basis of prevention.

The clinical value of genetic testing for SNPs was initially questioned by some commentators because individual variants would have limited power to discriminate cancer outcomes [61,62]. However, theoretical models suggested that polygenic risk scores based on multiple variants would provide sufficient discrimination for risk stratification to improve the efficiency of screening [63] and more recent studies have begun to demonstrate the potential clinical applications of polygenic risk profiling based on known susceptibility variants. For example, Pashayan and colleagues[64] showed that if prostate cancer screening were offered to men with a ten-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[64]. So and colleagues[65] showed how a similar age and polygenic risk could be applied to breast cancer screening [65]. Assuming that eligibility for mammographic screening is based on a ten-year risk of breast cancer of 2.4% - equivalent to the risk of the average 50 year old woman - women at the 90th percentile of a 13-SNP polygenic risk score would be eligible for screening from age 40 whereas those at the 20th percentile would be eligible from age 62. The incorporation of additional genetic variants and other risk factors including family history would improve the discrimination of the polygenic risk models and enhance their clinical applicability. 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 enormous potential of genetic findings[66,67] to impact public health and clinical care through the next several decades of scientific research[68]. Cancer screening tools are available for many cancers, such as mammography, endoscopy or biomarker tests including PSA or CA125 levels, although many currently available biomarkers have limited value in identifying clinically meaningful cancers. Given the expense, limited availability, potential complications, and risks and cost associated with false positive findings identifying those at highest risk will have important public health and cost implications relevant to personalizing cancer prevention. These examples demonstrate the enormous potential of genetic findings[56,57] to impact public health and clinical care through the next several decades of scientific research.[69]

Gene-environment Interactions (GxE)

Several environmental and lifestyle risk factors, many of which are modifiable, such as obesity, physical activity, non-steroidal 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, primarily from GAME-ON investigators, that provides a strong rationale for examining GxE interactions[10,12-15,70-74]. The development of statistical methods for genome-wide GxE with increased power [75,76] 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 gene-environment interaction 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. Additionally, 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[77] – proving causality for risk variants and the genes they regulate - have been described for a tiny fraction of risk associations identified by GWAS[22,78]. This is partly due to our rudimentary knowledge of the non-coding genome and the effects of genetic variation on gene regulation. Integration of GWAS SNP data with methylome data has identified methylation-quantitative trait loci (meQTLs) showing that inherited genetic variation may affect carcinogenesis by regulating the human methylome[79,80]. The ENCODE (**ENC**yclopedia **O**f **D**NA **E**lements) consortium has catalogued genome-wide regulatory elements for many, but by no means all human tissues[81]. Enhancers are often cell type-specific and drive the spatial and temporal diversity of gene expression in and across different cell types. For example, in a study of H3K4 methylation in K562 and HeLa cell lines, each cell line had an estimated 24,000-36,000 enhancers, but only 5,000 of these sites were present in both cell lines[82]. 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 sequence-specific double-stranded breaks (DSBs) into loci of interest followed by homology-directed repair (HDR) to efficiently 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[78,83].

Complementary to genome editing for proving causality of risk SNPs is expression quantitative trait locus (eQTL) analysis to identify the likely target susceptibility gene as susceptibility loci[84,85]. eQTL analyses can interrogate both near or distant regulatory associations between risk genotypes and gene expression on the same chromosome (*cis*-) or across chromosomes *(trans*-) . The role of these genes in neoplastic development can then be evaluated in experimental models of disease[60]. Many groups have applied this concept to identify transcript expression correlated with trait-associated SNPs[86-88]. 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[89-91].

References

1. Lichtenstein P, Holm NV, Verkasalo PK, et al. Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *N.Engl.J Med.* 7/13/2000 2000;343(2):78-85.
2. de la Chapelle A. Genetic predisposition to colorectal cancer. *Nat.Rev.Cancer.* 10/2004 2004;4(10):769-780.
3. Antoniou AC, Easton DF. Risk prediction models for familial breast cancer. *Future oncology.* Apr 2006;2(2):257-274.
4. Chakravarti A. Population genetics--making sense out of sequence. *Nat.Genet.* 1/1999 1999;21(1 Suppl):56-60.
5. Lander ES, Schork NJ. Genetic dissection of complex traits. *Science.* 9/30/1994 1994;265(5181):2037-2048.
6. Reich DE, Lander ES. On the allelic spectrum of human disease. *Trends Genet.* 9/2001 2001;17(9):502-510.
7. Houlston RS, Peto J. The search for low-penetrance cancer susceptibility alleles. *Oncogene.* 8/23/2004 2004;23(38):6471-6476.
8. Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI. Shifting paradigm of association studies: Value of rare single nucleotide polymorphisms. *Genetic Epidemiology.* Sep 2007;31(6):608-608.
9. Zhu QG, D. Maia, J.M.; Petrovski, S.; Dickson, S.P.; Heinzen, E.L., Shianna, K.V.; Goldstein, D.B. A genome-wide comparisonof the functional properties of rare and common genetic variants in humans. *American journal of human genetics.* 2011;88(4):458-468.
10. Ponder BA. Inherited predisposition to cancer. *Trends Genet.* 7/1990 1990;6(7):213-218.
11. Ponder BA. Cancer genetics. *Nature.* 5/17/2001 2001;411(6835):336-341.
12. Peto J. Cancer epidemiology in the last century and the next decade. *Nature.* 5/17/2001 2001;411(6835):390-395.
13. Hunter DJ. Gene-environment interactions in human diseases. *Nat.Rev.Genet.* 4/2005 2005;6(4):287-298.
14. Potter JD. Colorectal cancer: molecules and populations. *J.Natl.Cancer Inst.* 6/2/1999 1999;91(11):916-932.
15. Thomas DC. *Statistical methods in genetic epidemiology.* New York: Oxford University Press; 2004.
16. Hindorff LA, Sethupathy P, Junkins HA, et al. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc.Natl Acad.Sci.U.S.A.* 6/9/2009 2009;106(23):9362-9367.
17. Qian DCB, J.; Han, Y.; Hunter, D.J.; Henderson, B.E.; Eeles, R.; Haiman, C.A.; Easton, D.F; Hung, R.J.; Amos, C.I. . *Identification of genetic factors contributing to development of common cancers through tissue-specific protein interaction analysis* August 1 2015.
18. Smith GD, Ebrahim S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *International journal of epidemiology.* Feb 2003;32(1):1-22.
19. Maurano MT, Humbert R, Rynes E, et al. Systematic localization of common disease-associated variation in regulatory DNA. *Science.* Sep 7 2012;337(6099):1190-1195.
20. Coetzee SG, Shen HC, Hazelett DJ, et al. Cell-type-specific enrichment of risk-associated regulatory elements at ovarian cancer susceptibility loci. *Human molecular genetics.* Jul 1 2015;24(13):3595-3607.
21. Hazelett DJ, Rhie SK, Gaddis M, et al. Comprehensive functional annotation of 77 prostate cancer risk loci. *PLoS genetics.* Jan 2014;10(1):e1004102.
22. Smemo S, Tena JJ, Kim KH, et al. Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature.* Mar 20 2014;507(7492):371-375.
23. Ahmadiyeh N, Pomerantz MM, Grisanzio C, et al. 8q24 prostate, breast, and colon cancer risk loci show tissue-specific long-range interaction with MYC. *Proceedings of the National Academy of Sciences of the United States of America.* May 25 2010;107(21):9742-9746.

24. Spurdle AB, Thompson DJ, Ahmed S, et al. Genome-wide association study identifies a common variant associated with risk of endometrial cancer. *Nature genetics.* May 2011;43(5):451-454.
25. Painter JN, O'Mara TA, Batra J, et al. Fine-mapping of the HNF1B multicancer locus identifies candidate variants that mediate endometrial cancer risk. *Human molecular genetics.* Mar 1 2015;24(5):1478-1492.
26. Shen H, Fridley BL, Song H, et al. Epigenetic analysis leads to identification of HNF1B as a subtype-specific susceptibility gene for ovarian cancer. *Nature communications.* 2013;4:1628.
27. Soslow RA. Histologic subtypes of ovarian carcinoma: an overview. *International journal of gynecological pathology : official journal of the International Society of Gynecological Pathologists.* Apr 2008;27(2):161-174.
28. Pearce CL, Templeman C, Rossing MA, et al. Association between endometriosis and risk of histological subtypes of ovarian cancer: a pooled analysis of case-control studies. *The Lancet. Oncology.* Apr 2012;13(4):385-394.
29. Cancer Genome Atlas Research N, Kandoth C, Schultz N, et al. Integrated genomic characterization of endometrial carcinoma. *Nature.* May 2 2013;497(7447):67-73.
30. Thompson DJ, O'Mara TA, Glubb DM, et al. CYP19A1 fine-mapping and Mendelian randomization: estradiol is causal for endometrial cancer. *Endocrine-related cancer.* Feb 2016;23(2):77-91.
31. O'Mara TA, Glubb DM, Painter JN, et al. Comprehensive genetic assessment of the ESR1 locus identifies a risk region for endometrial cancer. *Endocrine-related cancer.* Oct 2015;22(5):851-861.
32. Eeles RA, Olama AA, Benlloch S, et al. Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. *Nat Genet.* Apr 2013;45(4):385-391, 391e381-382.
33. Pharoah PD, Tsai YY, Ramus SJ, et al. GWAS meta-analysis and replication identifies three new susceptibility loci for ovarian cancer. *Nat Genet.* Apr 2013;45(4):362-370, 370e361-362.
34. Bahcall OG. iCOGS collection provides a collaborative model. Foreword. *Nat Genet.* Apr 2013;45(4):343.
35. Michailidou K, Hall P, Gonzalez-Neira A, et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nat Genet.* Apr 2013;45(4):353-361, 361e351-352.
36. Gudmundsson J, Sulem P, Steinthorsdottir V, et al. Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. *Nat Genet.* Aug 2007;39(8):977-983.
37. Yang JF, T.; Morris, A.P.; Medland, S.E, Genetic Investigation of Anthropometric Traits(GIANT) Consortium; DIAbetes Genetics Replication And Meta-analysis(DIAGRAM) Consortium; Madden, P.A.F.; Heath, A.C.; Martin, N.G.; Montgomery, G.W.; Weedon, M.N.; Loos, R.J.; Frayling, T.M.; McCarthy, M.I.; Hirschhorn, J.N.; Goddard, M.E.; Visscher, P.M. . Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature genetics.* 2013;44:369-375.
38. Cousminer DL, Stergiakouli E, Berry DJ, 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. *Human molecular genetics.* Aug 15 2014;23(16):4452-4464.
39. Bojesen SE, Pooley KA, Johnatty SE, et al. Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. *Nature genetics.* Apr 2013;45(4):371-384, 384e371-372.
40. Wheeler HE, Maitland ML, Dolan ME, Cox NJ, Ratain MJ. Cancer pharmacogenomics: strategies and challenges. *Nature reviews. Genetics.* Jan 2013;14(1):23-34.
41. Purcell S, Neale B, Todd-Brown K, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics.* Sep 2007;81(3):559-575.
42. Van LJ, Wilson DC, Satsangi J. The genetics of Crohn's disease. *Annu.Rev.Genomics Hum.Genet.* 2009 2009;10:89-116.
43. Brest P, Corcelle EA, Cesaro A, et al. Autophagy and Crohn's disease: at the crossroads of infection, inflammation, immunity, and cancer. *Curr.Mol.Med.* 7/2010 2010;10(5):486-502.
44. Green ED, Guyer MS. Charting a course for genomic medicine from base pairs to bedside. *Nature.* 2/10/2011 2011;470(7333):204-213.

45. Rioux JD, Xavier RJ, Taylor KD, et al. Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nat Genet.* 2007 2007;39(5):596-604.
46. Cadwell K, Liu JY, Brown SL, et al. A key role for autophagy and the autophagy gene Atg16l1 in mouse and human intestinal Paneth cells. *Nature.* 11/13/2008 2008;456(7219):259-263.
47. Klein RJ, Zeiss C, Chew EY, et al. Complement factor H polymorphism in age-related macular degeneration. *Science.* 2005 2005;308(5720):385-389.
48. Moshfeghi DM, Blumenkranz MS. Role of genetic factors and inflammation in age-related macular degeneration. *Retina.* 3/2007 2007;27(3):269-275.
49. Dietz HC. New therapeutic approaches to mendelian disorders. *The New England journal of medicine.* Aug 26 2010;363(9):852-863.
50. Okada Y, Wu D, Trynka G, et al. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature.* Feb 20 2014;506(7488):376-381.
51. Cordell HJ, Han Y, Mells GF, et al. International genome-wide meta-analysis identifies new primary biliary cirrhosis risk loci and targetable pathogenic pathways. *Nature communications.* 2015;6:8019.
52. Nelson MR, Tipney H, Painter JL, et al. The support of human genetic evidence for approved drug indications. *Nature genetics.* Aug 2015;47(8):856-860.
53. Seshagiri S. The burden of faulty proofreading in colon cancer. *Nature genetics.* Feb 2013;45(2):121-122.
54. Miller DT, Adam MP, Aradhya S, et al. Consensus statement: chromosomal microarray is a first-tier clinical diagnostic test for individuals with developmental disabilities or congenital anomalies. *Am.J.Hum.Genet.* 5/14/2010 2010;86(5):749-764.
55. Frueh FW, Amur S, Mummaneni P, et al. Pharmacogenomic biomarker information in drug labels approved by the United States food and drug administration: prevalence of related drug use. *Pharmacotherapy.* 8/2008 2008;28(8):992-998.
56. Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. *Science.* 11/7/2008 2008;322(5903):881-888.
57. 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.* 6/26/2008 2008;358(26):2760-2763.
58. Dietz HC. New therapeutic approaches to mendelian disorders. *N.Engl.J.Med.* 8/26/2010 2010;363(9):852-863.
59. Grasemann H, Ratjen F. Emerging therapies for cystic fibrosis lung disease. *Expert.Opin.Emerg.Drugs.* 12/2010 2010;15(4):653-659.
60. Burke W, Laberge AM, Press N. Debating clinical utility. *Public Health Genomics.* 2010 2010;13(4):215-223.
61. Holtzman NA, Marteau TM. Will genetics revolutionize medicine? *The New England journal of medicine.* Jul 13 2000;343(2):141-144.
62. Vineis P, Schulte P, McMichael AJ. Misconceptions about the use of genetic tests in populations. *Lancet.* Mar 3 2001;357(9257):709-712.
63. Pharoah PD, Antoniou A, Bobrow M, Zimmern RL, Easton DF, Ponder BA. Polygenic susceptibility to breast cancer and implications for prevention. *Nature genetics.* May 2002;31(1):33-36.
64. Pashayan N, Duffy SW, Neal DE, et al. Implications of polygenic risk-stratified screening for prostate cancer on overdiagnosis. *Genetics in medicine : official journal of the American College of Medical Genetics.* Jan 8 2015.
65. 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.* May 13 2011;88(5):548-565.
66. Altshuler D, Daly MJ, Lander ES. Genetic mapping in human disease. *Science.* Nov 7 2008;322(5903):881-888.
67. Hunter DJ, Altshuler D, Rader DJ. From Darwin's finches to canaries in the coal mine--mining the genome for new biology. *The New England journal of medicine.* Jun 26 2008;358(26):2760-2763.

68. Lander ES. Initial impact of the sequencing of the human genome. *Nature.* Feb 10 2011;470(7333):187-197.
69. Lander ES. Initial impact of the sequencing of the human genome. *Nature.* 2/10/2011 2011;470(7333):187-197.
70. Collins FS, Manolio TA. Merging and emerging cohorts: necessary but not sufficient. *Nature.* 1/18/2007 2007;445(7125):259.
71. Khoury MJ, Wacholder S. Invited commentary: from genome-wide association studies to gene-environment-wide interaction studies--challenges and opportunities. *Am J Epidemiol.* 1/15/2009 2009;169(2):227-230.
72. Manolio TA, Collins FS. Genes, environment, health, and disease: facing up to complexity. *Hum.Hered.* 2007 2007;63(2):63-66.
73. Thomas D. Gene-environment-wide association studies: emerging approaches. *Nat.Rev.Genet.* 3/9/2010 2010;11(4):259-272.
74. Thomas D. Methods for investigating gene-environment interactions in candidate pathway and genome-wide association studies. *Annu.Rev.Public Health.* 4/21/2010 2010;31:21-36.
75. Evans DM, Marchini J, Morris AP, Cardon LR. Two-stage two-locus models in genome-wide association. *PLoS.Genet.* 9/22/2006 2006;2(9):e157.
76. Marchini J, Donnelly P, Cardon LR. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet.* 4/2005 2005;37(4):413-417.
77. Freedman ML, Monteiro AN, Gayther SA, et al. Principles for the post-GWAS functional characterization of cancer risk loci. *Nature genetics.* Jun 2011;43(6):513-518.
78. Spisak S, Lawrenson K, Fu Y, et al. CAUSEL: an epigenome- and genome-editing pipeline for establishing function of noncoding GWAS variants. *Nature medicine.* Sep 23 2015.
79. Shi J, Marconett CN, Duan J, et al. Characterizing the genetic basis of methylome diversity in histologically normal human lung tissue. *Nature communications.* 2014;5:3365.
80. Scherf DB, Sarkisyan N, Jacobsson H, et al. Epigenetic screen identifies genotype-specific promoter DNA methylation and oncogenic potential of CHRNB4. *Oncogene.* Jul 11 2013;32(28):3329-3338.
81. Thurman RE, Rynes E, Humbert R, et al. The accessible chromatin landscape of the human genome. *Nature.* Sep 6 2012;489(7414):75-82.
82. Heintzman ND, Hon GC, Hawkins RD, et al. Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature.* May 7 2009;459(7243):108-112.
83. Bauer DE, Kamran SC, Lessard S, et al. An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. *Science.* Oct 11 2013;342(6155):253-257.
84. Veyrieras JB, Kudaravalli S, Kim SY, et al. High-resolution mapping of expression-QTLs yields insight into human gene regulation. *PLoS genetics.* Oct 2008;4(10):e1000214.
85. Stranger BE, Montgomery SB, Dimas AS, et al. Patterns of cis regulatory variation in diverse human populations. *PLoS genetics.* 2012;8(4):e1002639.
86. Kwan T, Grundberg E, Koka V, et al. Tissue effect on genetic control of transcript isoform variation. *PLoS genetics.* Aug 2009;5(8):e1000608.
87. Lalonde E, Ha KC, Wang Z, et al. RNA sequencing reveals the role of splicing polymorphisms in regulating human gene expression. *Genome research.* Apr 2011;21(4):545-554.
88. Cookson W, Liang L, Abecasis G, Moffatt M, Lathrop M. Mapping complex disease traits with global gene expression. *Nature reviews. Genetics.* Mar 2009;10(3):184-194.
89. Li Q, Seo JH, Stranger B, et al. Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell.* Jan 31 2013;152(3):633-641.
90. Grisanzio C, Werner L, Takeda D, et al. Genetic and functional analyses implicate the NUDT11, HNF1B, and SLC22A3 genes in prostate cancer pathogenesis. *Proceedings of the National Academy of Sciences of the United States of America.* Jul 10 2012;109(28):11252-11257.
91. Lawrenson K, Li Q, Kar S, et al. Cis-eQTL analysis and functional validation of candidate susceptibility genes for high-grade serous ovarian cancer. *Nature communications.* 2015;6:8234.

Figure 1. Twenty risk regions analyzed as part of the OncoArray, including 17 pleiotropic regions conferring risks to two or more common cancers (breast, colorectal, lung, ovarian or prostate cancers). Panel (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 catalouges of regulatory marks for related tissue types for common cancers to identify SNPs intersecting tissue specific regulatory targets. Publically available genome wide regulatory profiling data were available for the HMEC 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 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 panel b, only SNPs with regulatory marks are shown, thus excluding 24 of the regional associations shown in panel a.

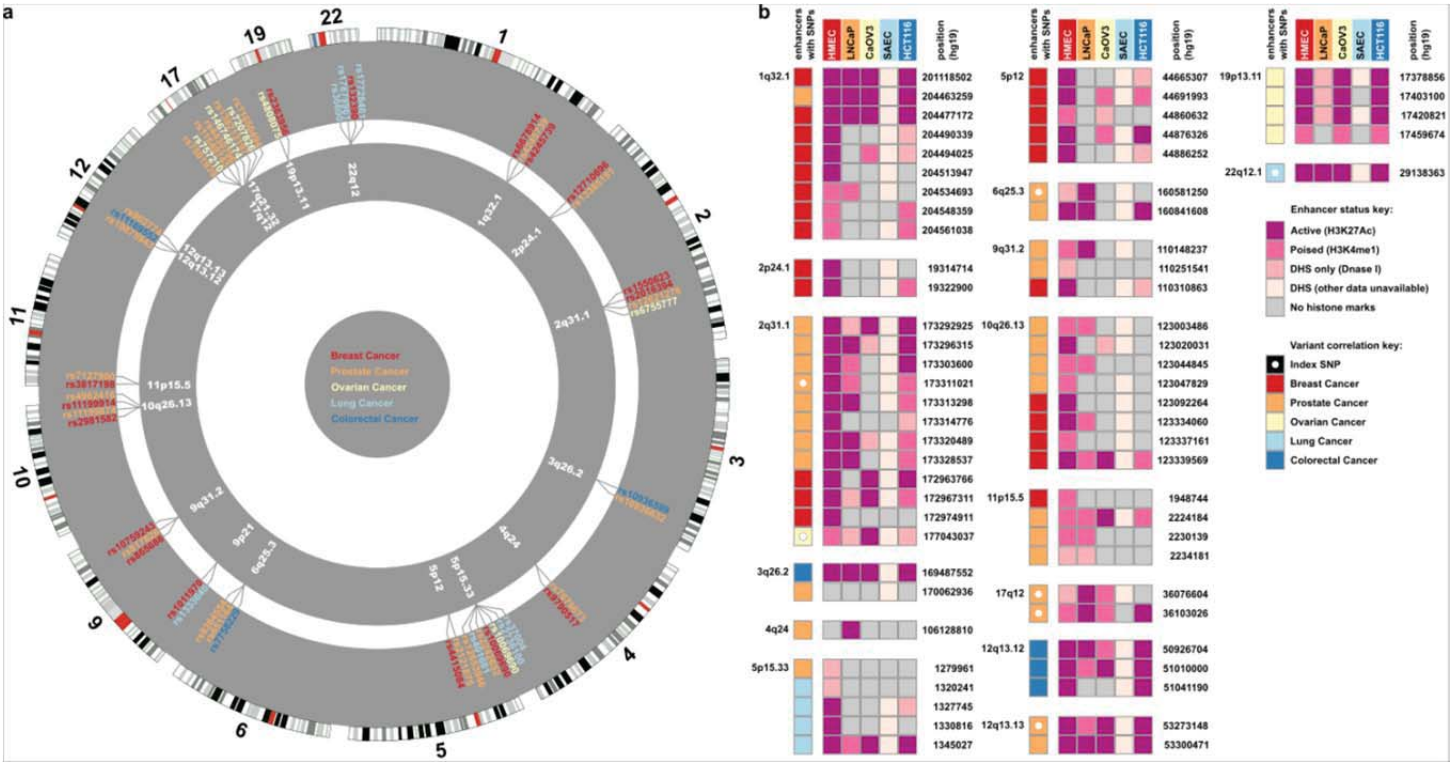


Figure 2. Failure rates (<95% of SNPs called) for 211,638 samples genotyped by CIDR across multiple tissue types. The overall failure rate was 3.17%.

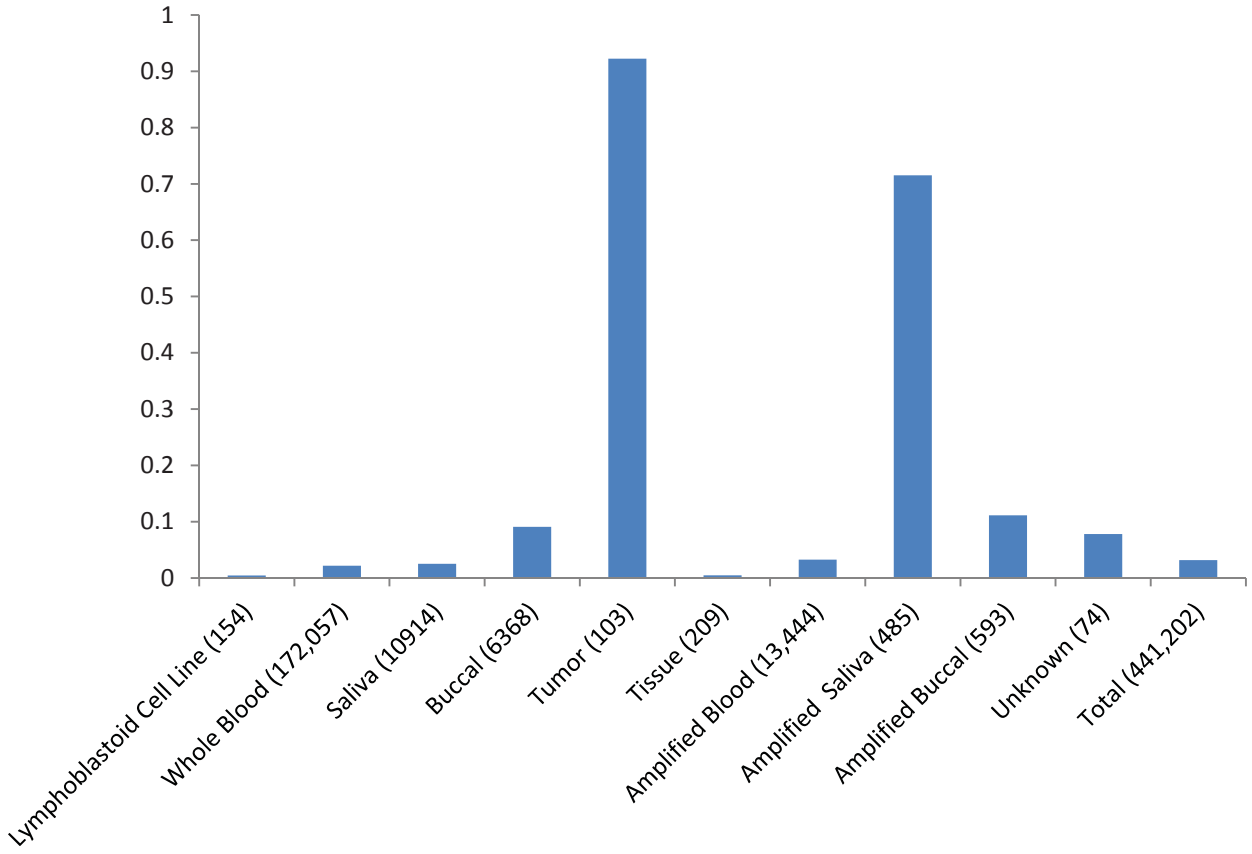


Figure 3: Correlation between replicate Hapmap samples genotyped at Cambridge versus the Center for Inherited Disease Research. Samples 1-8 are of European origin while samples 9-14 are Asian or African. There are multiple replicates of samples 1, 4, 5, 6 and 8. Samples 1-8 are European, 9-10 are Chinese,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

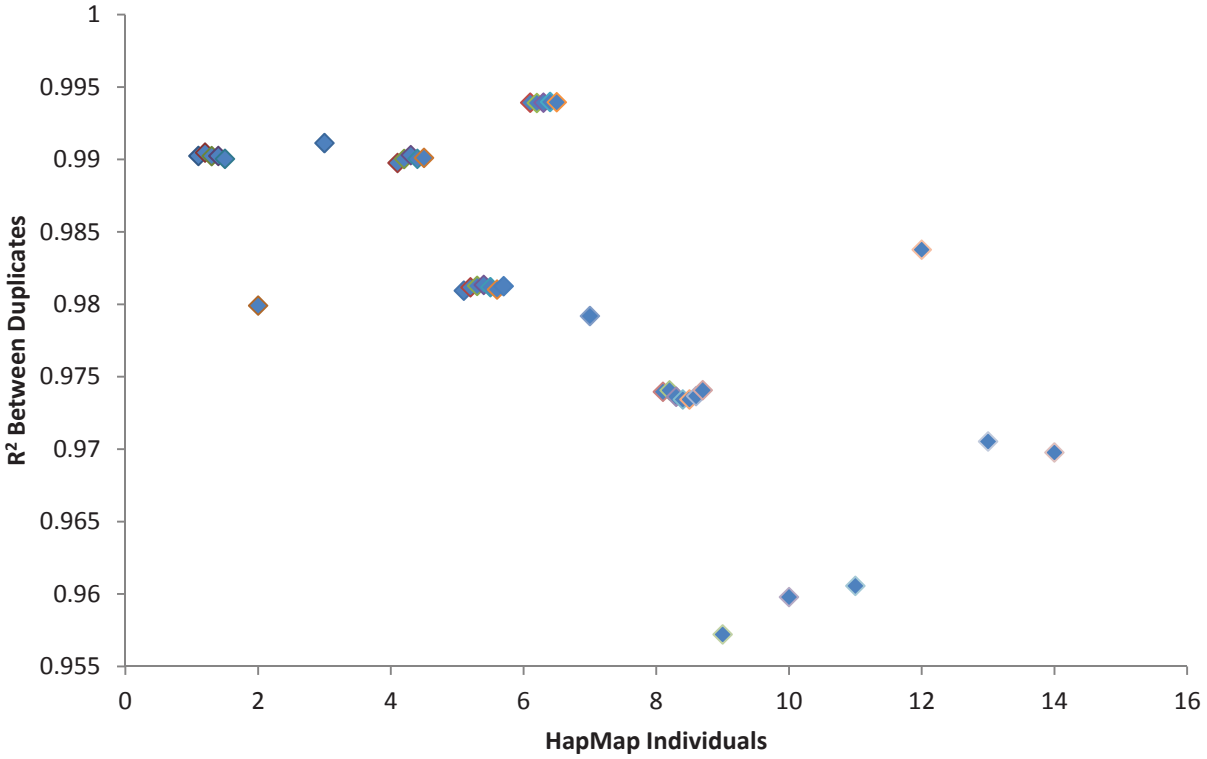
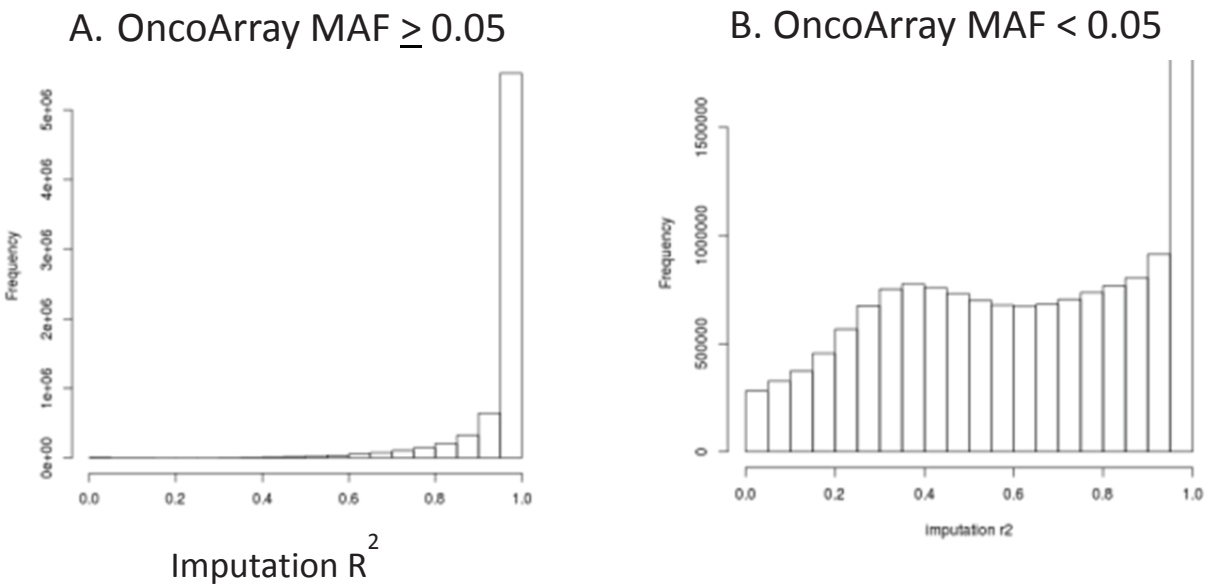


Figure 4. Imputation accuracy of SNPs from v3 of the 1000 Genomes project in individuals of European Descent



Supplementary Table 1a. OncoArray – Participating cites and numbers of participants from Prostate Cancer Studies

					Cases		Controls		White		African American		Asian		Other	
Genotyping Center	Acronym	PIs	Country	Design	Males	Females	Males	Females	Case No.	Control No.	Case No.	Control No.	Case No.	Control No.	Case No.	Control No.
Cambridge	Aarhus	Karina Dalsgaard Sorensen, Torben Falck Ørntoft	Denmark	Hospital-based, Retrospective, Observational	1140	0	570	0	1130	558	1	0	4	7	5	5
NCI	AHS	Laura E. Beane-Freeman, Michael Alavanja, Stella Koutros	USA	Nested case-control study within prospective cohort	514	0	1314	0	514	1314	0	0	0	0	0	0
NCI	ATBC	Demetrius Albanes	USA	Prospective, nested case-control	1474	0	2205	0	1449	2189	0	0	5	4	20	12
USC	BioVu	Melinda Aldrich, Dana C. Crawford	USA	Cases identified in a biobank linked to electronic health records	213	0	0	0	0	0	213	0	0	0	0	0
CIDR	CCI	Matthew Parliament, Nawaid Usmani	Canada	Case series, Hospital-based	285	0	0	0	275	0	4	0	4	0	2	0
NCI	ProGene (CeRePP)	Olivier Cussenot, Géraldine Cancel-Tassin	France	Case-Control, Prospective, Observational, Hospital-based	1064	0	881	0	952	771	109	107	1	2	2	1
BGI	CHIPGECS	Yong-Jie Lu, Guangwen Cao, Hong-Wei Zhang , Ninghan Feng, Xin Guo, Guomin Wang, Zan Sun	China	Case-control	533	0	666	0	0	0	0	0	532	666	1	0
CIDR	COH	Susan L. Neuhausen	USA	hospital-based cases and controls from outside	263	0	269	0	259	269	0	0	3	0	1	0
CIDR	COSM	Alicja Wolk	Sweden	Population-based cohort	2406	0	1204	0	2389	1193	0	0	11	6	6	5
Copenhagen	CPCS1	Børge G. Nordestgaard	Denmark	Case-control - Denmark	552	0	269	0	551	269	0	0	1	0	0	0
Copenhagen	CPCS2		Denmark		461	0	238	0	461	238	0	0	0	0	0	0
USC	CPDR	Shiv Srivastava, Jennifer C. Cullen, George Petrovics	USA	Retrospective cohort	145	0	44	0	0	0	145	44	0	0	0	0
NCI	CPS-II	Susan M. Gapstur, Victoria L. Stevens	USA	Nested case-control derived from a prospective cohort study	4743	0	4508	0	4688	4453	11	14	33	28	11	13
CIDR	EPIC	Tim J. Key, Ruth C. Travis, Elio Riboli	Multi Center in EU	Case-control - Germany, Greece, Italy, Netherlands, Spain, Sweden, UK	697	0	739	0	681	723	4	0	3	8	9	8
Centre National de Genotypage (CNG)	EPICAP	Florence Menegaux	France	Case-control, Population-based, ages less than 75 years at diagonosis, Hérault, France	64	0	63	0	0	0	64	63	0	0	0	0
Cambridge	ERSPC	Christopher Bangma, Monique J. Roobol		Population-based randomised trial	75	0	75	0	73	73	2	0	0	1	0	1
Cambridge	ESTHER	Hermann Brenner	Germany	Case-control, Prospective, Observational, Population-based	341	0	333	0	339	333	0	0	1	0	1	0
CIDR	FHCRC	Janet L. Stanford	USA	Population-based, case-control, ages 35-74 years at diagnosis, King County, WA, USA	434	0	421	0	418	403	2	1	6	12	8	5
CIDR	Gene-PARE	Barry Rosenstein, Harry Ostrer		Hospital-based	1330	0	0	0	274	0	48	0	996	0	12	0
CIDR	Hamburg-Zagreb	Marija Gamulin, Davor Lessel	Croatia	Hospital-based, Prospective	154	0	154	0	154	154	0	0	0	0	0	0
CIDR	HPFS	Sara Lindstrom, Edward Giovannucci, Kathryn L. Penney, Lorelei Mucci	USA	Nested case-control	1233	0	1095	0	1212	1081	14	4	4	8	3	2
Cambridge	IMPACT	Rosalind A. Eeles	UK	Observational	60	0	993	0	58	975	0	3	2	10	0	5
Cambridge	IPO-Porto	Manuel R. Teixeira	Portugal	Hospital-based	386	0	190	0	384	190	0	0	0	0	2	0
USC	Karuprostate	Laurent Brureau, Luc Multigner, Pascal Blanchet	W. Indies	Case-control, Retrospective, Population-based	384	0	411	0	0	0	384	411	0	0	0	0
Genome Quebec	KULEUVEN	Frank Claessens, Thomas Van den Broeck, Steven Joniau	Belgium	Hospital-based, Prospective, Observational	175	0	103	0	174	103	0	0	1	0	0	0
USC	LAAPC	Sue Ann Ingles	USA	Population-based, Case-control	789	0	621	0	456	283	2	0	7	6	324	332
USC	Malaysia	Azad Razack, Jasmine Lim, Soo-Hwang Teo, Meng H. Tan, Aik T. Ong	Malaysia	Case-control	210	0	210	0	1	0	0	0	208	209	1	1
CIDR	MCC-Spain	Manolis Kogevinas, Gemma Castaño-Vinyals, Javier Llorca Diaz	Spain	Case-control	542	0	443	0	534	425	1	1	4	11	3	6
CIDR	MCCS	Graham G. Giles, Melissa C. Southey	Australia	Nested case-control, Melbourne, Victoria	780	0	334	0	776	334	0	0	3	0	1	0
USC	MD Anderson	Sara S. Strom	USA		1139	0	316	0	532	0	47	0	39	0	521	316
CIDR	MDACC_AS	Christopher J. Logothetis, Jeri Kim	USA	A prospective cohort study	633	0	0	0	532	0	47	0	39	0	15	0
USC	MEC	Christopher A. Haiman, Brian E. Henderson, Fredrick Schumacher	USA	Population-based	1310	0	1396	0	625	664	490	530	32	30	163	172
USC	WFPCS	Jennifer J. Hu	USA		59	0	66	0	0	0	59	66	0	0	0	0
USC	MOFFITT	Jong Y. Park	USA	Hospital-based	602	0	346	0	429	226	129	101	7	3	37	16
USC	NMHS	Jay Fowke	USA	Case-control, clinic based, Nashville TN	188	0	201	0	0	0	188	201	0	0	0	0

(Continued)

Supplementary Table 1a. OncoArray – Participating cites and numbers of participants from Prostate Cancer Studies

					Cases		Controls		White		African American		Asian		Other	
Genotyping Center	Acronym	PIs	Country	Design	Males	Females	Males	Females	Case No.	Control No.	Case No.	Control No.	Case No.	Control No.	Case No.	Control No.
Cambridge	CONOR (Oslo)	Lovise Maehle, Eli Marie Grindedal, Johanna Schleutker, Fredrik Wiklund	Norway	Population-based, Retrospective, Observational	1513	0	0	0	1487	0	2	0	11	0	13	0
CIDR	Canary PASS	Daniel W. Lin	USA	Prospective, Multi-site, Observational Acive Surveillance Study	380	0	0	0	369	0	0	0	9	0	2	0
USC	PCaP	Jeannette T. Bensen, James Mohler, Elizabeth T.H. Fontham, Gary J. Smith	USA	Population-based , case only	1022	0	0	0	0	0	1022	0	0	0	0	0
Cambridge	PCMUS	Radka Kaneva, Vanio Mitev, Chavdar Slavov	Bulgaria	Case-control - Sofia, Bulgaria	195	0	90	0	195	90	0	0	0	0	0	0
CIDR	PHS	Meir Stampfer, Sara Lindstrom, Peter Kraft, Kathryn L. Penney	USA	Nested case-control	664	0	286	0	642	271	4	2	11	11	7	2
NCI	PLCO	Sonja I. Berndt, Stephen Chanock, Gerald Andriole	USA	Nested case-control	1010	0	1275	0	999	1187	0	58	1	1	10	29
CIDR	Poland	Cezary Cybulski	Poland	Case-control	510	0	345	0	509	344	0	0	1	1	0	0
USC?	PRAGGA	Manuela Gago Dominguez, Jose Esteban Castelao	Spain	Case-control	133	0	104	0	132	102	0	0	1	2	0	0
CIDR	PROCAP	Henrik Gronberg, Fredrik Wiklund	Sweden	Population-based, Retrospective, Observational	677	0	339	0	675	332	0	1	1	6	1	0
Cambridge	PROFILE	Rosalind A. Eeles	UK	Hospital-based, Prospective, Observational	32	0	88	0	30	85	1	2	1	1	0	0
CIDR	PROGReSS	Ana Vega	Spain	Hospital-based, Prospective, Observational	696	0	349	0	692	348	0	0	2	1	2	0
Cambridge	ProMPT	David E. Neal	UK	A study to collect samples and data from subjects with and without prostate cancer. Retrospective, Experimental	1002	0	12	0	975	12	11	0	7	0	9	0
Cambridge	ProtecT	Jenny L. Donovan, Freddie C. Hamdy, David E. Neal, Richard Martin	UK	Trial of treatment. Samples taken from subjects invited for PSA testing from the community at nine centres across United Kingdom	4	0	1448	0	4	1429	0	2	0	11	0	6
USC	PROtEuS	Marie-Elise Parent	Canada	Case-control, population-based	72	0	58	0	0	0	72	58	0	0	0	0
CIDR	QLD	Jyotsna Batra, Suzanne Chambers, Amanda Spurdle	Australia	Case-control	3489	0	1356	0	3425	1336	3	0	52	15	9	5
Cambridge	RAPPER	Alison Dunning, Catharine West, Neil Burnet		Multi-centre, hosptial based blood sample collection study in patients enrolled in clinical trials with prospective collection of radiotherapy toxicity data	2350	0	0	0	2255	0	52	0	29	0	14	0
USC	SABOR	Ian M. Thompson Jr.	USA	Prostate Cancer Screening Cohort	366	0	366	0	0	0	106	106	0	0	260	260
USC	SCCS	William J. Blot, Wei Zheng	USA	Case-control in cohort, Southeastern USA. Prospective, Observational, Population-based	257	0	1601	0	0	0	257	1601	0	0	0	0
USC	SCPCS	Maureen Sanderson	USA	Population-based , Retrospective, Observational	64	0	39	0	0	0	64	39	0	0	0	0
Cambridge/CIDR	SEARCH	Paul Pharoah, Nora Pashayan, Alison Dunning	UK	Case-control - East Anglia, UK	2932	0	1520	0	2852	1504	30	3	30	9	20	4
USC	SFPCS	Esther M. John	USA	Population-based case-control study, Retrospective, Observational	378	0	249	0	290	212	86	37	1	0	1	0
Cambridge	SNP_Prostate_Ghent	Kim De Ruyck, Piet Ost	Belgium	Hospital-based, Retrospective, Observational	334	0	141	0	325	141	1	0	3	0	5	0
Cambridge	SPAG	Claire Aukim-Hastie, Samantha Larkin, Paul A. Townsend	UK	Hospital-based, Retrospective, Observational	47	0	192	0	47	189	0	0	0	1	0	2
CIDR	STHM2	Henrik Gronberg, Fredrik Wiklund	Sweden	Population-based, Retrospective, Observational	3148	0	1576	0	3104	1557	12	0	18	11	14	8
CIDR	SWOG-PCPT	Catherine M. Tangen, Ian M. Thompson		Case-control from a randomized clinical trial	1211	0	1424	0	1097	1080	88	239	22	71	4	34
CIDR	SWOG-SELECT	Catherine M. Tangen, Ian M. Thompson		Case-cohort from a randomized clinical trial	1847	0	3122	0	1507	2215	263	697	41	85	36	125
Cambridge	TAMPERE	Johanna Schleutker	Finland	Case-control - Finland, Retrospective, Observational, Population-based	2544	0	1226	0	2534	1215	0	0	4	8	6	3
Ontario Cancer Institute Genomics Center	Toronto	Robert J. Hamilton, Neil E. Fleshner, Antonio Finelli		Prospective hospital-based biopsy cohort	821	0	599	0	677	464	60	28	65	89	19	18
USC	UGANDA	Stephen Watya	Uganda		567	0	489	0	0	0	567	489	0	0	0	0
Cambridge, USC, CIDR	UKGPCS	Kenneth Muir, Rosalind A. Eeles, ZSofia Kote-Jarai	UK	ICR, UK	14107	0	7601	0	13168	7494	708	6	145	49	86	52
Cambridge	ULM	Christiane Maier	Germany	Case-control - Germany	475	0	190	0	471	188	1	0	1	1	2	1
CIDR / USC	WUGS	Bettina Drake, Adam S. Kibel, Aleksandra Klim, Graeme Colditz	USA	Cases Series, USA	930	0	153	0	704	0	211	153	7	0	8	0

Supplementary Table 1b. OncoArray – Participating cites and numbers of participants from Breast Cancer Studies

					Cases		Control		White	African	Asian	Other
Genotyping Center	**Acronym**	**Principal Investigator**	**Country**	**Design**	**Female**	**Male**	**Female**	**Male**	**Cases + Control**	**Cases + Control**	**Cases + Control**	**Cases + Control**
CIDR	2SISTER	Jack Taylor	USA	Family-based study	1130		0		1082	48	0	0
Genome Quebec	ABCFS	John Hopper	Australia	Population-based case-control study	1120		189		1309	0	0	0
Cambridge	ABCS	Marjanka Schmidt	Netherlands	Hospital-based consecutive cases; population-based controls	270		187		459	0	0	0
Genome Quebec	ABCTB	Jane Carpenter	Australia	Hospital based multi site newly diagnosed breast cancer case	687		189		1062	0	0	0
Cambridge	ACP	Kenneth Muir	Thailand	Hospital based case-control study	753		375		0	0	1395	0
Genome Quebec	BBCC	Peter Fasching	Germany	Hospital based cases; population based controls	442		642		695	0	0	0
Cambridge	BBCS	Julian Peto	UK	Cancer registry and National Cancer Research network (NCRN) based cases; population based controls	122		253		564	0	0	0
Cambridge	BCEES	Jennifer Stone	Australia	Population-based case-control study	785		442		1620	0	0	0
Genome Quebec	BCFR-NY	Mary-Beth Terry	USA	Clinic-based family study	465		835		493	0	0	0
Genome Quebec	BCFR-PA	Mary Daly	USA	Clinic-based family study	141		28		141	0	0	0
Genome Quebec	BCFR-Utah	David Goldgar	USA	Clinic-based family study	103		0		103	0	0	0
Genome Quebec	BCINIS	Gadi Rennert	Israel	Population-based case-control	1439		0		2163	0	0	0
Cambridge	BREOGAN	Manuela Gago	Spain	Population-based case-control	1379		724		2108	0	0	0
Genome Quebec	BSUCH	Barbara Burwinkel	Germany	Hospital based cases;healthy blood donator controls	269		729		437	0	0	0
Genome Quebec	CAMA	Elad Ziv	Mexico	Population-based case-control study	709		168		0	0	0	1366
Genome Quebec	CBCS	Kristan Aranson	Canada	Population-based case-control study	1025		657		1514	0	506	0
Cambridge	CCGP	Manolis Saloustros	Greece	Hospital-based case-control study	683		995		1015	0	0	0
Genome Quebec	CECILE	Pascal Guenel	France	Population-based case-control study	306		332		465	0	0	0
Copenhagen	CGPS	Stig Bojesen	Denmark	Population-based case-control study	1411		159		2127	0	0	0
Genome Quebec	COLBCCC	Diana Torres	Colombia	Hospital-based case-control study	633		716		0	0	0	1194
CIDR	CPSII	Mia Gaudet	USA	Nested case-control study	3055		561		6087	0	0	0
Genome Quebec	CTS	Hoda Anton-Culver	USA	Nested case-control study	1075		3032		1616	0	0	0
Cambridge	DIETCOMPLYF	Miriam Dwek	UK	Multi-centre prospective case-only study	711		541		711	0	0	0
CIDR	EPIC	Elio Riboli	MULTIPLE	Nested case-control study	3850		0		7497	0	0	0
Genome Quebec	ESTHER	Hermann Brenner	Germany	Population-based case-control study	296		3647		483	0	0	0
Cambridge	GC-HBOC	Alfons Meindl	Germany	Clinic-based family study	3647		187		5240	0	0	0
Genome Quebec	GENICA	Hiltrud Brauch	Germany	Population-based case-control study	460		1593		744	0	0	0
Genome Quebec	GeparSixto	Peter Fasching	Germany	Clincal Trial	390		284		390	0	0	0
Cambridge	GESBC	Jenny Chang-Claude	Germany	Population-based case-control study	360		0		541	0	0	0
Cambridge	HaBCS	Thilo Doerk	Germany	Hospital-based case-control study	934		181		1800	0	0	0
Genome Quebec	HCSC	Atocha Romero	Spain	Hospital-based case-control study	427		866		427	0	0	0
Genome Quebec	HEBCS	Heli Nevanlinna	Finland	Hospital-based case-control study + additional familial cases	281		0		458	0	0	0
Genome Quebec	HERPACC	Keitaro Matsuo	Japan	Hospital-based case-control study	282		177		0	0	565	0
Genome Quebec	HKBCS	Ava Kwong	Hong Kong	Hospital-based case-control study	564		283		0	0	1018	0
Cambridge	HMBCS	Thilo Doerk	Belarus	Hospital based cases; population based controls	212		454		463	0	0	0
Cambridge	HUBCS	Thilo Doerk	Russia	Hospital based cases; population based controls	312		251		538	0	0	0
Genome Quebec	KARBAC	Annika Lindblom	Sweden	Population and hospital-based cases; geographically matched controls	506		226		506	0	0	0
Cambridge	KBCP	Arto Mannermaa	Finland	Population-based case-control study	557		0		802	0	0	0
Genome Quebec	KOHBRA	Sue Park	Korea	Hospital-based case-control study	1464		245		0	0	2129	0
Genome Quebec	LMBC	Diether Lambrechts	Belgium	Hospital-based case-control study	806		665		2075	0	0	0
Cambridge	MaBCS	Thilo Doerk	Macedonia	Hospital-based case-control study	90		1269		183	0	0	0
Genome Quebec	MARIE	Jenny Chang-Claude	Germany	Population-based case-control study	512		93		801	0	0	0
Genome Quebec	MBCSG	Paolo Radice	Italy	Clinic-based recruitment of familial cases; population-based controls	788		289		1154	0	0	0
Genome Quebec	MCBCS	Fergus Couch	USA	Hospital-based case-control study	370		366		551	0	0	0

(Continued)

Supplementary Table 1b. OncoArray – Participating cites and numbers of participants from Breast Cancer Studies

					Cases		Control		White	African	Asian	Other
Genotyping Center	Acronym	Principal Investigator	Country	Design	Female	Male	Female	Male	Cases + Control	Cases + Control	Cases + Control	Cases + Control
Mayo	MCBCS (Mayo)	Fergus Couch	USA	Hospital-based case-control study	558		181		598	0	0	0
Cambridge	MCCS	Graham Giles	Australia	Nested case-control study	862		40		1677	0	0	0
CIDR	MCCS (CIDR)	Graham Giles	Australia	Nested case-control study	190		815		355	0	0	0
CIDR	MEC	Chris Haiman	USA	Nested case-control study	674		725		1399	0	0	0
Cambridge	MISS	Hakan Olsson	Sweden	Nested case-control study	703		731		2248	0	0	0
Genome Quebec	MMHS	Celine Vachon	USA	Nested case-control study	306		1545		541	0	0	0
Mayo	MMHS (Mayo)	Celine Vachon	USA	Nested case-control study	78		235		1481	0	0	0
Genome Quebec	MTLGEBCS	Mark Goldberg	Canada	Hospital-based case-control study	343		1403		513	0	0	0
Genome Quebec	MYBRCA	Soo Hwang-Teo	Malaysia	Hospital-based case-control study	845		170		0	0	2103	0
Genome Quebec	NBCS	Vessela Kristensen	Norway	Hospital-based case-control study	1285		1258		1286	0	0	0
CIDR	NBHS	Wei Zheng	USA	Population-based case-control study	887		1		1329	354	0	0
Genome Quebec	NC-BCFR	Esther John	USA	Population-based familial case-control study	1264		796		960	254	503	0
Genome Quebec	NGOBCS	Motoki Iwasaki	Japan	Hospital-based case-control study	369		199		0	0	735	0
CIDR	NHS	Peter Kraft	USA	Nested case-control study	1594		366		3402	0	0	0
CIDR	NHS2	Peter Kraft	USA	Nested case-control study	1609		1808		3517	0	0	0
Genome Quebec	OFBCR	Irene Andrulis	Canada	Population-based familial case-control study	1669		1908		2045	0	0	0
Cambridge	ORIGO	Peter Devilee	Netherlands	Hospital-based case-control study	1059		376		1721	0	0	0
NCI	PBCS	Montse Garcia-Closas	Poland	Population-based case-control study	1931		662		3976	0	0	0
Cambridge	pKARMA	Kamila Czene	Sweden	Population-based case-control study	2993		2045		9080	0	0	0
NCI	PLCO	Robert Hoover	USA	Nested case-control study	869		6087		1727	0	0	0
Cambridge	POSH	Diana Eccles	UK	Clinic-based case-only study	1091		858		1091	0	0	0
Genome Quebec	PreFace	Peter Fasching	Germany	Clinical Trial	991		0		991	0	0	0
Cambridge	RBCS	Maartje Hooning	Netherlands	Hospital-based case-control study	475		0		717	0	0	0
Genome Quebec	SBCGS	Wei Zheng	China	Population-based case-control study	840		242		0	0	1775	0
Cambridge	SEARCH	Paul Pharoah	UK	Population-based case-control study	4062		1828		6746	0	0	0
Genome Quebec	SEBCS	Daehee Kang	Korea	Hospital-based case-control study	1103		1791		0	0	2210	0
Genome Quebec	SGBCC	Mikael Hartman	Singapore	Hospital-based cases, population based controls	927		1107		0	0	1665	0
CIDR	SISTER	Jack Taylor	USA	Population-based family study	2187		738		3609	325	0	0
Genome Quebec	SKKDKFZS	Uta Hamann	Germany	Hospital-based case-only study	1097		1747		1097	0	0	0
Cambridge	SMC	Alicja Wolk	Sweden	Nested case-control study	1504		0		2213	0	0	0
Genome Quebec	SuccessB	Peter Fasching	Germany	Clinical Trial	440		709		440	0	0	0
Genome Quebec	SuccessC	Peter Fasching	Germany	Clinical Trial	1343		0		1343	0	0	0
Genome Quebec	SZBCS	Jan Lubinski	Poland	Hospital-based case-control study	387		0		561	0	0	0
Mayo	TNBCC	Fergus Couch	MULTIPLE	Case series from multiple countries	1439		69		1508	0	0	0
Genome Quebec	TWBCS	Chen-Yang Shen	Taiwan	Hospital-based case-control study	551		0		0	0	807	0
Genome Quebec	UCIBCS	Hoda Anton-Culver	USA	Population-based case-control study	507		256		767	0	0	0
Cambridge	UKBGS	Anthony Swerdlow	UK	Nested case-control study	1632		260		2337	0	0	0
UKOPS	UKOPS	Usha Menon	UK	Population-based cohort			705		976	0	0	0
CIDR	WAABCS	Fumni Olopade	MULTIPLE	Hospital-based case-control study	315		976		0	626	0	0
CIDR	WHI	Ross Prentice	USA	Nested case-control study	4937		311		9555	0	0	0

Supplementary Table 1c. OncoArray – Participating cites and numbers of participants from CIMBA Studies

				Females			Males			Female and Male Case + Control			
Genotyping Center	Acronym	PI	Country	Case + Control	BRCA1 *	BRCA2	Case + Control	BRCA1	BRCA2	White ^	Black	Asian	Other
MAYO	BCFR-AU	Melissa Southey	AUSTRALIA	81	40	41	1	0	1	82	0	0	0
CIDR	BCFR-NC	Esther John	USA	18	7	11	8	3	5	24	0	0	2
CIDR	BCFR-NY	Mary Beth Terry	USA	124	67	57	0	0	0	124	0	0	0
GQ	BCFR-ON	Irene Andrulis	CANADA	219	126	93	18	7	11	237	0	0	0
CIDR	BCFR-PA	Mary Daly	USA	52	45	7	0	0	0	52	0	0	0
CIDR	BCFR-UT	David Goldgar	USA	253	130	123	47	17	30	298	0	0	2
CIDR	BFBOCC	Ramunas Janavicius/Liene Nikitinia-Zake	LITHUANIA/LATVIA	267	249	18	4	4	0	271	0	0	0
CIDR	BIDMC	Nadine Tung	USA	140	86	54	0	0	0	140	0	0	0
MAYO	BMBSA	Lizette Jansen van Rensburg	SOUTH AFRICA	203	59	144	0	0	0	203	0	0	0
GQ	BRICOH	Susan Neuhausen	USA	320	181	139	82	19	63	400	0	0	2
MAYO	CBCS	Hansen	DENMARK	343	203	140	1	1	0	344	0	0	0
CIDR	CNIO	Javier Benitez/Ana Osorio	SPAIN	128	66	62	5	1	4	133	0	0	0
CIDR	COH	Jeffrey Weitzel	USA	652	431	221	5	3	2	388	0	0	269
CAM	TEAM	Paolo Radice	ITALY	870	550	320	183	83	100	1053	0	0	0
CIDR	S	Koulis Yannoukakos	GREECE	271	235	36	7	4	3	278	0	0	0
CIDR	DFCI	Judy Garber	USA	283	150	133	0	0	0	283	0	0	0
CAM	DKFZ	Ute Hamann	GERMANY	85	60	25	7	5	2	92	0	0	0
CAM	EMBRACE	Douglas Easton	UK/IRELAND	3441	1749	1692	305	78	227	3744	0	0	2
CIDR	FCCC	Andrew Godwin	USA	123	78	45	18	3	15	141	0	0	0
CAM	FPGMX	Ana Vega	SPAIN	190	112	78	0	0	0	189	1	0	0
CAM	GC-HBOC	Rita Schmutzler	GERMANY	3039	1928	1111	162	44	118	3201	0	0	0
GQ	GEMO	Sylvie Mazoyer/Dominique Stoppa-Lyonnet/Fabienne Lesueur	FRANCE/USA	2459	1501	958	69	10	59	2528	0	0	0
CIDR	WN	Claudine Isaacs	USA	15	15	0	0	0	0	15	0	0	0
CAM	G-FAST	Kathleen Claes	BELGIUM	360	195	165	31	0	31	391	0	0	0
CIDR	HCSC	Trinidad Caldes	SPAIN	305	146	159	37	0	37	342	0	0	0
CAM	HEBCS	Heli Nevannlina	FINLAND	259	126	133	33	8	25	292	0	0	0
CAM	HEBON	Matti Rookus	NETHERLANDS	1528	901	627	15	8	7	1543	0	0	0
CIDR	HRBCP	Ava Kwong	HONG KONG	120	51	69	0	0	0	0	0	120	0
MAYO	HUNBOCS	Edith Olah	HUNGARY	398	282	116	26	8	18	424	0	0	0
MAYO	HVH	Orland Diez	SPAIN	256	120	136	20	2	18	276	0	0	0
CIDR	ICO	Conxi Lazaro	SPAIN	648	288	360	73	8	65	721	0	0	0
GQ	IHCC	Jakabowska	POLAND	205	205	0	0	0	0	205	0	0	0
CAM	ILUH	Rosa Barkardottir	ICELAND	147	0	147	43	0	43	190	0	0	0
GQ	INHERIT	Jacques Simard	(QUEBEC)	183	96	87	0	0	0	183	0	0	0
CAM	IOVHBOCS	Marco Montagna	ITALY	374	206	168	21	1	20	395	0	0	0
CAM	IPOBCS	Manuel Teixeira	PORTUGAL	281	117	164	12	0	12	293	0	0	0
CIDR/MAYO	KCONFAB	Georgia Chenevix-Trench	AUSTRALIA	1594	892	702	272	68	204	1866	0	0	0
GQ	KOHBRA	Sue Park	KOREA	502	194	308	65	20	45	1	0	566	0
CIDR	KUMC	Priyanka Sharma	USA	44	29	15	0	0	0	44	0	0	0
CIDR	MAYO	Fergus Couch	USA	387	258	129	4	2	2	391	0	0	0
GQ	MCGILL	Mark Tischkowitz	CANADA (QUEBEC)	88	54	34	0	0	0	88	0	0	0
CIDR	MSKCC	Ken Offit	USA	772	396	376	52	14	38	824	0	0	0
CIDR/MAYO	MUV	Christian Singer	AUSTRIA	806	541	265	22	4	18	828	0	0	0
CIDR	NAROD	Steven Narod	CANADA	380	286	94	0	0	0	301	0	32	47
CIDR	NCI	Mark Greene	USA	236	153	83	18	10	8	254	0	0	0
CIDR	NNPIO	Evgeny Imyanitov	RUSSIA	75	73	2	0	0	0	75	0	0	0
CIDR	NORTHSHORE	Peter Hulick	USA	139	82	57	0	0	0	139	0	0	0
CIDR	NRG_ONCOLOGY	Mark Greene	USA/AUSTRALIA	628	332	296	0	0	0	628	0	0	0
GQ	OCGN	Irene Andrulis	CANADA	382	208	174	19	7	12	401	0	0	0
CIDR	OSU CCG	Amanda Toland	USA	197	93	104	14	4	10	211	0	0	0
MAYO	OUH	Mads Thomassen	DENMARK	1000	568	432	105	27	78	1105	0	0	0
CIDR	PBCS	Maria Caligo	ITALY	98	91	7	4	0	4	102	0	0	0
CAM/GQ	SEABASS	Soo Hwang-Teo	MALAYSIA	111	61	50	12	9	3	0	0	123	0
CIDR	SMC	Eitan Friedman	ISRAEL	254	171	83	0	0	0	254	0	0	0
CIDR	SWE-BRCA	Ake Borg/Johanna Rantala	SWEDEN	498	434	64	11	8	3	509	0	0	0
CIDR	UCHICAGO	Funmi Olopade	USA	156	98	58	23	8	15	179	0	0	0
CIDR	UCSF	Robert Nussbaum	USA	170	100	70	0	0	0	170	0	0	0
CAM	UKGRFOCR	Susan Ramus	UK	74	57	17	0	0	0	74	0	0	0
CIDR	UPENN	Kate Nathanson	USA	850	489	361	90	44	46	927	7	0	6
CIDR	UPITT	Darcy Thull	USA	265	158	107	13	1	12	278	0	0	0
CIDR	UTMDACC	Banu Arun	USA	122	48	74	0	0	0	122	0	0	0
CIDR/MAYO	VFCTG	Gillian Mitchell	AUSTRALIA	469	244	225	32	13	19	500	0	1	0
CIDR	WCP	Beth Karlan	USA	213	157	56	0	0	0	213	0	0	0

* 16 individuals carried both a BRCA1 and a BRCA2 mutation
^ includes Ashkenazi Jewish carriers

Supplementary Table 1d. OncoArray – Participating cites and numbers of participants from CORECT Studies

					Cases		Controls		White		African American		Asian		Other		
Genotyping Center	**Acronym**	**PI**	**Country**	**Design**	**Males**	**Females**	**Males**	**Females**	**Case No.**	**Control No.**	**Case No.**	**Control No.**	**Case No.**	**Control No.**	**Case No.**	**Control No.**	
CIDR	MECC	Gad Rennert, Steve Gruber	Israel	Case-Control	1735	1532	1192	1055	3267	2247							
USC	MECC	Gad Rennert, Steve Gruber	Israel	Case-Control	480	431	326	329	911	655							
CIDR	MSKCC	Kenneth Offit	USA	Case-Control	53	73	0	0	126	0							
CIDR	SEARCH	Paul Pharoah	United Kingdom	Case-Control	2516	1907	137	144	4423	281							
USC	SEARCH	Paul Pharoah	United Kingdom	Case-Control	166	117	0	0	283	0							
CIDR	SPAIN	Victor Moreno	Spain	Case-Control	628	353	499	442	981	941							
USC	SPAIN	Victor Moreno	Spain	Case-Control	0	0	45	88		117						12	4 race na
USC	FIRE3	Heinz-Josef Lenz	Germany	Clinical Trial Phase 3	184	72	0	0	262	0							6 sex na
USC	TRIBE	Heinz-Josef Lenz	Italy	Clinical Trial Phase 3	270	177	0	0	447	0							
CIDR	MEC	Loic Le Marchand	USA	Cohort	210	192	218	206							402	424	
CIDR	CFR-Hawaii	Loic LeMarchand	USA	Case-Control	30	19	0	0	49	0							
CIDR	CFR-Australia	John Hopper / Mark Jenkins	Australia	Case-Control	139	120	10	15	259	25							
CIDR	CFR-Seattle	Polly Newcomb	USA	Case-Control	529	484	300	298	1013	598							
CIDR	CRF-Mayo	Noralane Lindor	USA	Case-Control	149	146	0	0	295	0							
CIDR	CFR-Ontario	Steve Gallinger	Canada	Case-Control	201	174	129	86	375	215							
CIDR	CFR-USC	Graham Casey	USA	Case-Control	85	107	0	0	192	0							
CIDR	Esther/Verdi	Hermann Brenner	Germany	Case-Control	300	166	313	166	466	479							
CIDR	Kiel	Jochen Hampe / Clemens Schafmeyer	Germany	Case-Control	635	499	0	0	1134	0							
CIDR	ColoCare - Germany + Seattle	Cornelia (Neli) Ulrich / Bill Grady	Germany and USA	Case-Series	292	179	18	32	471	50							
USC	Moffitt - Total Cancer Care and ColoCare	Erin Siegel	USA	Case-Series	230	197	0	0	356		8		3		5		55 race NA
CIDR	NHS2	Andrew Chan	USA	Cohort	0	127	0	123	127	123							
CIDR	MCCS	Graham Giles	Australia	Cohort	119	109	116	103	228	219							
CIDR	Korea	Wei Zheng	Korea	Case-Control	2241	1292	1516	2076					3533	3592			
CIDR	Shanghai-Mens	Wei Zheng	China	Case-Control	125		125						125	125			
CIDR	Shanghai-Womens	Wei Zheng	China	Case-Control		125		125					125	125			
CIDR	Sweden-Lindblom	Annika Lindblom	Sweden	Cohort	1768	1477	1485	1371	3245	2856							
CIDR	Sweden-Wolk	Alicja Wolk	Sweden	Cohort	346	250	506	370	596	876							
CIDR	USC-HRT-CRC	Anna Wu	USA	Case-Control	0	524	0	518	524	518							
CIDR	ATBC	Jarmo Virtamo - Finland & Dimetrius Albanes - NIH	Finland	Cohort	154	0	32	0	186	0							
CIDR	Hispanic	Jane Figueiredo	USA	Case-Control	55	33	0	0							88		
USC	Puerto Rico	Marcia Cruz Correa	Puerto Rico	Case-Control	170	176	105	228							346	333	3 sex na
USC	MAVERICC	Heinz-Josef Lenz	USA	Clinical Trial Phase 2	212	125	0	0	279		29		15		6		8 race NA
USC	Taiwan	Yun Yen	Taiwan	Case-Control	264	236	0	0					500				
USC	PPS3-Folate Aspirin	John Baron & Elizabeth Barry	USA and Canada	Clinical Trial	8	2	580	328									
CIDR	PPS4-Calcium Vit D	John Baron & Elizabeth Barry		Clinical Trial	6	3	1185	682	9	1867							
USC	Cancer Genetics Registry	Stephen Gruber	USA	Case-Series	112	109	2	9									
		Total			14412	11533	8839	8794	20504	12067	37	0	4301	3842	847	769	
		Total cases		25945													
		Total controls		17633													
		Grant Total		43578													

Supplementary Table 1e. OncoArray – Participating cites and numbers of participants from Lung Cancer Studies.

					Cases		Controls		White		African American		Asian		Other	
Genotyping	**Acronym**	**PI**	**Country**	**Design**	**Males**	**Females**	**Males**	**Females**	**Case No.**	**Control No.**	**Case No.**	**Control**	**Case No**	**Control**	**Case No.**	**Control**
CIDR	Norway	Aage Haugen	Norway	Hosp CC	239	100	293	134	339	427						
CIDR	MDACC	Xifeng Wu	US	Hosp CC	518	507	515	502	1005	990	4	3	1	1	15	23
CIDR	HSPH	David Christiani	US	Hosp CC	1461	1632	331	464	3020	745	42	5	10	7	22	38
CIDR	Liverpool_2008	John Field	UK	nested CC	62	46	70	48	108	118						
CIDR	Liverpool_2013	John Field	UK	nested CC	193	157	225	177	342	390	1	3		2	5	5
CIDR	CARET	Chu Chen, Jen Dohtery	US	nested CC	421	191	421	192	578	579	22	22	5	5	7	7
CIDR	NELCS	Angeline Andrew	US	Pop CC	86	104	80	104	176	179			1		13	5
CIDR	Tampa	Philip Lazarus	US	Hosp CC	234	174	233	171	390	365	8	34			10	5
CIDR	Resolucent	Penella J Woll , Dawn Teare	UK	family, Pop C	343	344	173	270	591	390	2		3		93	53
CIDR	ISRAEL	Gad Rennert	Israel	Pop CC	467	264	349	209	731	557		1				
CIDR	Nijmegen	Lambertus A. Kiemeney	The Netherlands	Pop CC	266	173	278	179	387	457			2		3	
CIDR	EAGLE	Maria Theresa Landi	Italy	Hosp CC	1465	380	1423	441	1845	1864						
CIDR	CAPUA	Adonina Tardon	Spain	Hosp CC	713	89	678	104	800	780					2	
CIDR	EPIC	Mattias Johansson	Europe	nested CC	761	453	765	470	1214	1235						
CIDR	MEC	Loic Le Marchand	US	nested CC	551	380	567	402	231	240	147	144	304	316	249	269
CIDR	MSH-PMH	Rayjean Hung, Geoffrey Liu	Canada	Clinic CC	729	723	501	507	1446	1006					5	
CIDR	PLCO	Neil Caporaso	US	nested CC	1040	660	917	654	1550	1114	86	395	31	33	33	29
CIDR	MLD	Paul Brennan	Russia	Hosp CC	833	308	712	426	1025	1084					116	55
CIDR	Seoul	Yun-Chul Hong	Korea	Hosp CC	206	96	199	291					302	490		
CIDR	ATBC	Demetrius Albanes	US	nested CC	1040		721		1040	721						
CIDR	LCRI-DOD	Susanne Arnold	US	Pop CC	50	50	65	72	98	133	1	1			1	3
CIDR	MDCS	Jonas Manjer	Sweden	nested CC	70	95	79	96	165	175						
CIDR	TLC	Matthew B. Schabath	US	case only	212	247			432		11				16	
CIDR	Vanderbilt2	Melinda Aldirch	US	Hosp CC	428	370	429	370	740	735	58	56		1		7
CIDR	SCHC	Jian-Min Yuan	Singapore	nested CC	292	126	291	127					418	418		
CIDR	SCS	Jian-Min Yuan	China	nested CC	178		325						178	325		
CIDR	Canadian screening	Stephen Lam, Ming-Sound Tsao, Geoff I	Canada	nested CC	117	152	202	267	263	455	2	4	4	9		
CIDR	NSHDC	Mattias Johansson	Sweden	nested CC	123	121	136	133	244	269						
Heidelberg	GLC	Angela Risch	Germany	Family CC	686	343	514	171								
					13098	7942	11492	6981								
Caret samples removed where there were overlaps with TRICL meta-analysis																

Supplementary Table 1f. OncoArray – Participating cites and numbers of participants from OCAC Studies

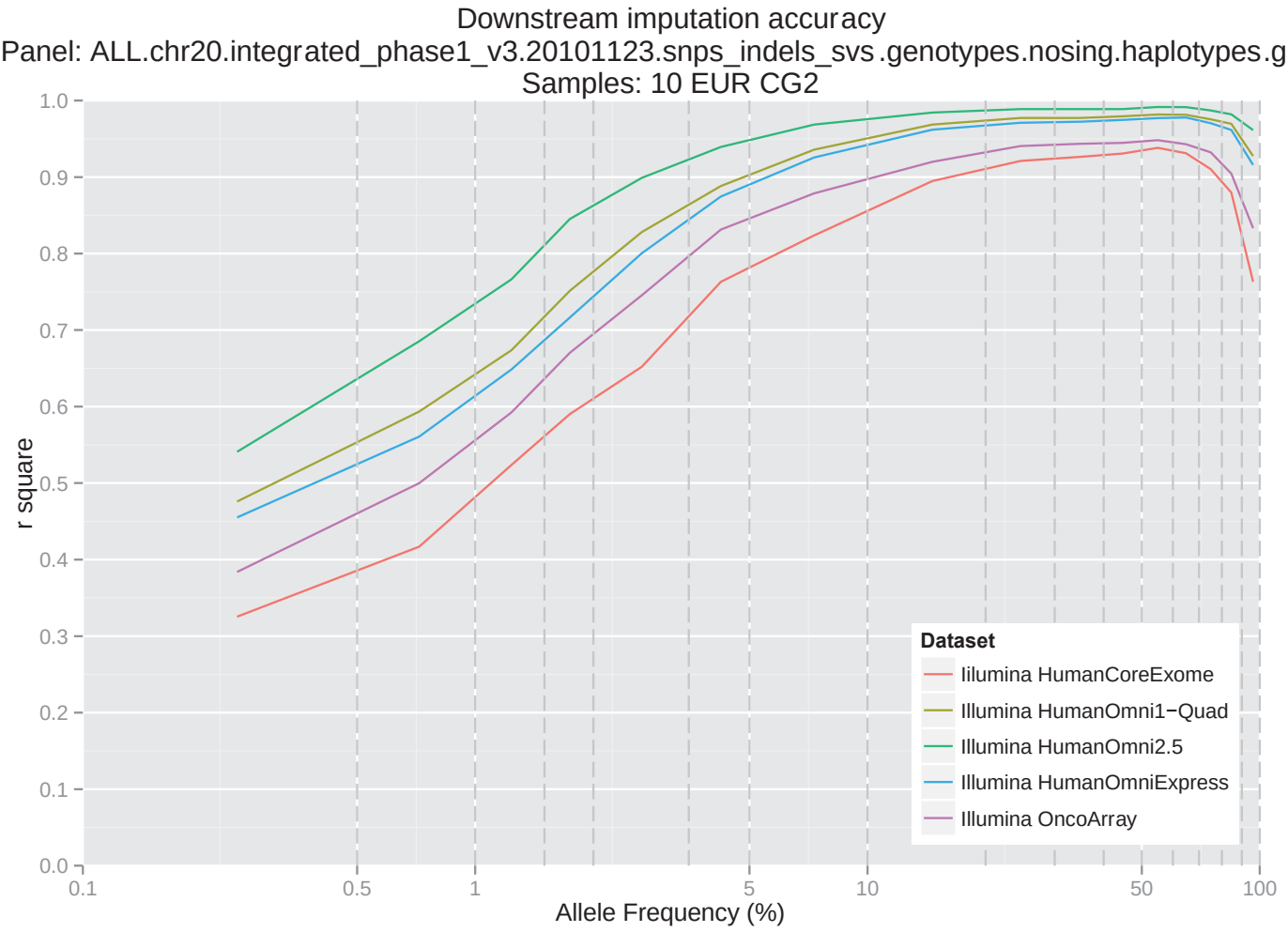
					Cases		Controls		White		African American		Asian		Other	
Genotyping Center	**Acronym**	**PI**	**Country**	**Design**	**Males**	**Females**	**Males**	**Females**	**Case No.**	**Control No.**	**Case No.**	**Control No.**	**Case No.**	**Control No.**	**Case No.**	**Control No.**
CIDR	AAS	Joellen Schildkraut, Patricia Moorman	USA	Case-control	0	296	0	475	1	0	295	475	0	0	0	0
MAYO	AOCS/ACS	Georgia Chenevix-Trench, Penelope Webb	Australia	Case-control	0	1,504	0	1,206	1420	1167	4	1	56	18	24	20
MAYO	AUS	Georgia Chenevix-Trench, Penelope Webb	Australia	Case-control	0	112	0	0	106	0	1	0	5	0	0	0
MAYO	BAV	Peter Fasching	Germany	Case-control	0	293	0	287	292	284	0	1	1	1	0	1
MAYO/CIDR	BEL	Diether Lambrechts	Belgium	Case-control	0	799	0	1,306	789	1297	4	5	2	2	4	2
CAM	BGS	Anthony Swerdlow	UK	Cohort	0	226	0	0	226	0	0	0	0	0	0	0
MAYO	BVU	Digna Velez Edwards	USA	Case-control	0	149	0	496	135	391	9	102	1	1	4	2
CAM	CAM	James Brenton	UK	Case-only	0	231	0	0	228	0	1	0	1	0	1	0
SHANGHAI	CHA	Kexin Chen, Fengju Song	China	Case-control	0	1,244	0	2,072	1	0	0	0	1243	2072	0	0
SHANGHAI	CHN	Li Yan, Kang Shan	China	Case-only	0	390	0	0	0	0	0	0	390	0	0	0
MAYO	CNI	Javier Benitez, María J. García, Cristina Rodriguez-Antona	Spain	Case-control	0	83	0	179	81	176	0	0	2	2	0	1
CIDR	DKE	Joellen Schildkraut, Andrew Berchuck	USA	Case-only	0	93	0	0	80	0	10	0	2	0	1	0
CIDR	DOV	Mary Anne Rossing	USA	Case-control	0	1,346	0	1,568	1246	1460	12	36	62	45	26	27
CIDR	EPC	Charlotte Oniand-Moret, Elio Riboli	Europe	Nested case-control	0	437	0	876	431	872	0	2	3	1	3	1
CIDR	GER	Jenny Chang-Claude	Germany	Case-control	0	205	0	376	203	376	0	0	2	0	0	0
MAYO	GRC	Drakoulis Yannoukakos	Greece	Case-only	0	327	0	0	325	0	0	0	2	0	0	0
CAM	GRR	Kirsten Moysich	USA	Case-only	0	22	0	0	22	0	0	0	0	0	0	0
CIDR	HAW	Marc Goodman	USA	Case-control	0	397	0	626	105	172	6	9	275	412	11	33
CAM	HJO	Thilo Doerk-Bousset, Matthias Duerst	Germany	Case-control	0	244	0	0	242	0	0	0	0	0	2	0
CAM	HMO	Thilo Doerk-Bousset, Natalia Bogdanova	Germany	Case-control	0	66	0	287	65	283	0	0	1	1	0	3
CAM	HOC	Ralf Butzow	Finland	Case-control	0	265	0	280	264	280	0	0	1	0	0	0
CIDR	HOP	Francesmary Modugno, Kirsten Moysich, Roberta Ness	USA	Case-control	0	549	0	1,217	524	1189	21	23	2	1	2	4
MAYO	HSK	Florian Heitz	Germany	Case-only	0	123	0	0	122	0	0	0	1	0	0	0
CAM	HUO	Thilo Doerk-Bousset	Germany	Case-control	0	73	0	235	49	126	0	0	2	16	22	93
MAYO	ICN	Florian Heitz	UK	Case-only	0	415	0	0	390	0	4	0	9	0	12	0
CIDR	JPN	Keitaro Matsuo	Japan	Case-control	0	150	0	232	0	1	0	0	150	231	0	0
MAYO	KRA	Sue Park	Korea	Case-control	0	310	0	688	0	0	0	0	310	688	0	0
CIDR	LAX	Beth Karlan	USA	Case-only	0	476	0	0	384	0	27	0	34	0	31	0
CAM	LUN	Håkan Olsson	Sweden	Case-control	0	41	0	1,577	41	1576	0	0	0	0	0	1
MAYO	MAC	Ellen Goode	USA	Case-only	0	213	0	0	205	0	2	0	2	0	4	0
CIDR	MAL	Susanne Kruger Kjaer	Denmark	Case-control	0	384	0	649	384	649	0	0	0	0	0	0
CIDR	MAS	Soo-Hwang Teo, Yin Ling Woo	Malaysia	Case-control	0	179	0	181	0	0	0	0	152	158	27	23
MAYO	MAY	Ellen Goode	USA	Case-control	0	1,170	0	1,146	1145	1135	6	5	7	4	12	2
MAYO	MCC	Graham Giles, Laura Baglietto, Gianluca Severi	Australia	Nested case-control	0	136	0	141	135	141	0	0	1	0	0	0
MAYO	MDA	Karen Lu, Michelle Hildebrandt	USA	Case-control	0	313	0	298	307	297	1	0	1	0	4	1
MAYO	MEC	Wendy Setiawan	USA	Case-control	0	67	0	79	6	6	14	15	19	28	28	30
CIDR	MOF	Thomas Sellers, Jennifer Permuth Wey, Catherine Phelan, Alvaro Monteiro	USA	Case-control	0	414	0	459	371	412	19	22	9	14	15	11
CIDR	MSK	Douglas Levine	USA	Case-control	0	238	0	245	201	205	13	26	15	6	9	8
CIDR	NCO	Joellen Schildkraut	USA	Case-control	0	994	0	925	837	734	142	179	12	3	3	9
CIDR	NEC	Daniel Cramer, Kathryn Terry	USA	Case-control	0	532	0	586	502	569	13	7	12	6	5	4
CIDR	NHS	Shelley Tworoger, Meir Stampfer, Walter Willett	USA	Nested case-control	0	342	0	316	337	314	3	0	2	2	0	0
CIDR	NOR	Helga B Salvesen	Norway	Case-control	0	186	0	342	184	342	1	0	1	0	0	0
CIDR	NTH	Lambertus Kiemeney, Leon Massuger	Netherlands	Case-control	0	263	0	588	255	588	1	0	2	0	5	0
MAYO	OPL	Penelope Webb	Australia	Case-only	0	510	0	0	484	0	0	0	13	0	13	0
MAYO	ORE	Tanja Pejovic	USA	Case-only	0	92	0	0	84	0	2	0	5	0	1	0
CIDR	OVA	Linda Cook, Nhu Le	Canada	Case-control	0	756	0	797	669	734	2	0	63	46	22	17
NCI	PLC	Nicolas Wentzensen	USA	Cohort	0	277	0	1,257	263	1119	7	94	5	43	2	1
CIDR	POC	Jacek Gronwald	Poland	Case-control	0	183	0	0	183	0	0	0	0	0	0	0
NCI	POL	Nicolas Wentzensen	Poland	Case-control	0	272	0	0	272	0	0	0	0	0	0	0
CIDR	PVD	Estrid Hogdall, Claus Hogdall	Denmark	Case-only	0	197	0	0	193	0	1	0	1	0	2	0
CAM	RBH	Georgia Chenevix-Trench	Australia	Case-only	0	141	0	0	139	0	0	0	1	0	1	0
CIDR	RMH	Paul Pharoah	UK	Case-only	0	182	0	0	174	0	2	0	1	0	5	0
CAM	RPC	Kirsten Moysich	USA	Case-only	0	106	0	0	99	0	5	0	1	0	1	0
MAYO/CIDR	SEA	Paul Pharoah	UK	Case-control	0	2,180	0	1,869	2154	1844	7	4	4	7	15	14
CIDR	SIS	Dale Sandler	USA	Cohort	0	131	0	1,507	118	1306	9	150	2	15	2	36
MAYO	SMC	Alicja Wolk	Sweden	Cohort	0	83	0	93	83	93	0	0	0	0	0	0
CIDR	SOC	Ian Campbell, Diana Eccles	UK	Case-only	0	301	0	0	297	0	1	0	0	0	3	0
MAYO	SRO	Jim Paul, Nadeem Siddiqui, Ros Glasspool, Iain McNeish, Susana Banerjee	UK	Case-only	0	3	0	0	3	0	0	0	0	0	0	0
CIDR	STA	Alice Whittemore, Weiva Sieh	USA	Case-control	0	424	0	464	282	307	16	51	77	60	49	46
CAM	SWH	Wei Zheng	China	Case-control	0	135	0	135	0	0	0	0	135	135	0	0
CIDR	SZB	Jacek Gronwald	Poland	Controls	0	0	0	181	0	180	0	0	0	0	0	1
MAYO	TBO	Rebecca Sutphen, Catherine Phelan	USA	Case-control	0	176	0	138	176	138	0	0	0	0	0	0
CIDR	TOR	Catherine Phelan, Steven Narod, Harvey Risch	Canada	Case-control	0	474	0	486	445	477	1	1	17	3	11	5
CIDR	UCI	Hoda Anton-Culver	USA	Case-control	0	311	0	348	258	295	2	3	22	21	29	29
MAYO	UHN	Marcus Bernardini	Canada	Case-only	0	211	0	0	177	0	7	0	13	0	14	0
CIDR	UKO	Usha Menon, Simon Gayther	UK	Case-control	0	755	0	998	737	979	6	11	4	2	8	6
CIDR	UKR	Paul Pharoah	UK	Case-only	0	49	0	0	48	0	0	0	0	0	1	0
CIDR	USC	Leigh Pearce, Anna Wu	USA	Case-control	0	926	0	1,026	607	787	40	34	113	93	166	112
CAM	VAN	David Huntsman	Canada	Case-only	0	221	0	0	171	0	1	0	34	0	15	0
MAYO	WMH	Anna deFazio	Australia	Case-only	0	176	0	0	145	0	1	0	16	0	14	0
MAYO	WOC	Jolanta Kupryjanczyk	Poland	Case-control	0	200	0	207	200	207	0	0	0	0	0	0

Supplementary Table 2: Transmitting institutions for organization of SNPs on the Oncoarray along with the proportion of the array allocated to specific cancers, areas of overlapping effects among cancers and for fine mapping among cancers.

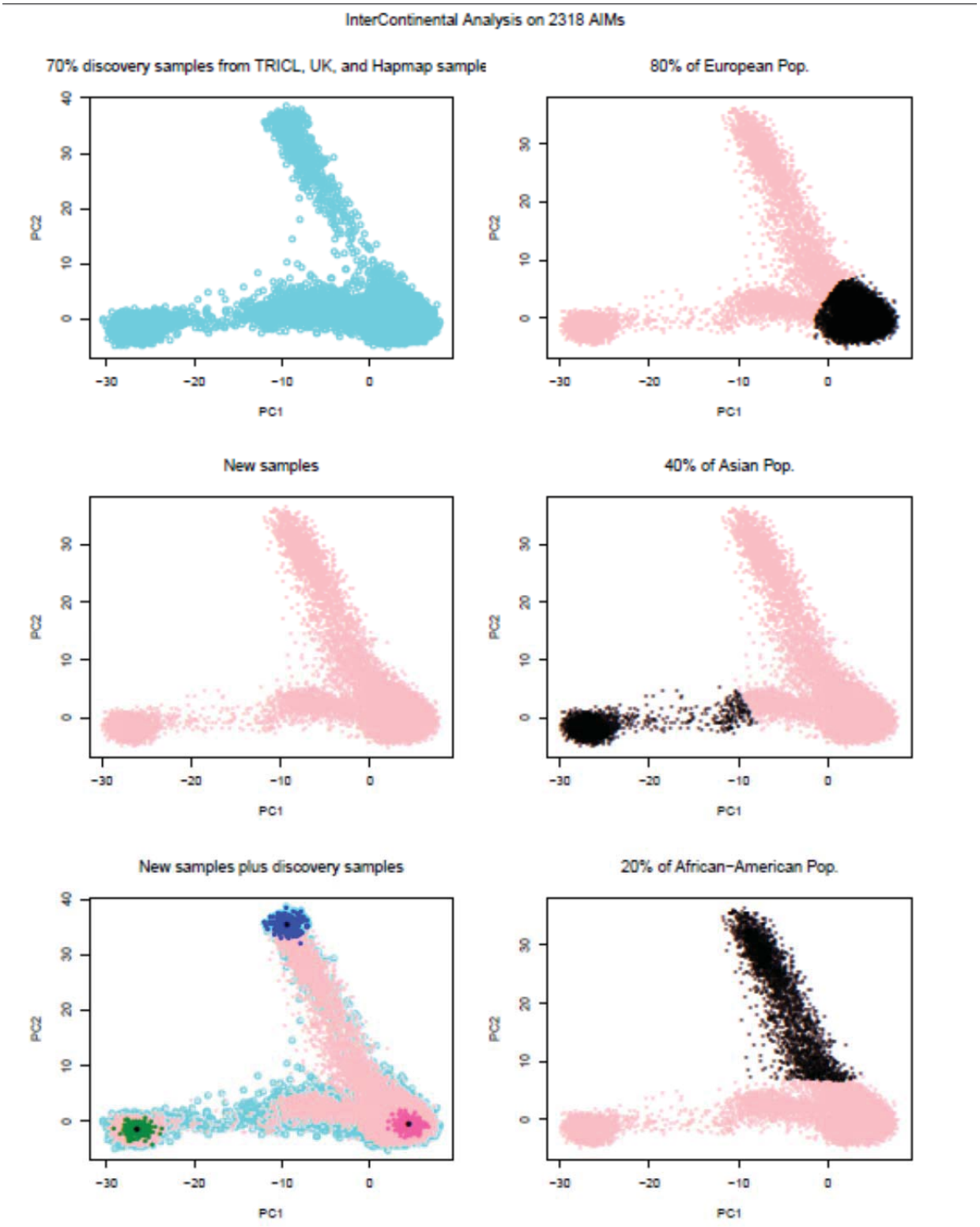
Site	Submitting Center	Proportional allocation
Lung	Dartmouth	13.5%
Ovary*	Cambridge	13.8%
Colorectal	USC	13.1%
Breast*	Cambridge	25.0%
Prostate	USC/ICR	24.5%
*BRCA1/2**	Cambridge	6.0%
Common (non fine-mapping)	Cambridge, Dartmouth, CGR, USC	4.1%
Common fine-mapping	Cambridge	Included in cancer-specific loci

*Breast and *BRCA1/2*, and ovary have shared lists based on meta-analyses. To simplify the final merging process, Cambridge assembled a single list from all three groups with a total allocation of 44.8% (split in the above proportions). The top 1800 SNPs identified by ECAC were included among those submitted as common non fine-mapping SNPs.
USC- University of Southern California; ICR=Institute for Cancer Research

Supplementary Figure 1. Comparison of the Oncoarray to several other Illumina arrays by imputing genotypes to the 1000 genomes release 3.3 or the Haplotype consortium for chromosome 22 using ShapeIt version 2 and Beagle, version 3.3.



Supplementary Figure 2. Scores of discovery set in blue, the predicted scores from SNP weights in discovery set in pink. Three populations in Hapmap2 display CEU in hotpink, CHB in green, and YRI in blue. Three plots on the right side indicate 80% European, 40% Asian, and 20% African-American proportions of population memberships.



Supplementary information about the OncoArray Consortium

The Consortium was formed to develop and genotype a new custom genotyping array (the "OncoArray"). The Oncoarray consortium brings together multiple disease-based consortia, including partnerships between the NCI-funded Genetic Associations and Mechanisms in Oncology (GAME-ON) initiative consortia (TRICL, FOCI, DRIVE, ELLIPSE and CORECT), the Breast Cancer Association Consortium (BCAC) and the Consortium of Investigators of Modifiers of BRCA1/2 (CIMBA). The project has been funded through substantial grants from the NCI to the GAME-ON initiative and the Division of Cancer Epidemiology and Genetics (DCEG), Genome Canada/Genome Quebec/CIHR through the Personalised Risk Stratification for Prevention and Early Detection of Breast Cancer (PERSPECTIVE) international project, Cancer Research UK (University of Cambridge) and a EU FP7 grant ("COGS"), together with many other grants.

The OncoArray Consortium has assembled more than 400,000 samples from existing studies and several biobanks. The OncoArray, which includes approximately 530K SNP markers, is a custom array that was manufactured by Illumina. Genotyping began in October 2013. The array includes a backbone of approximately 260,000 single nucleotide polymorphisms (SNPs) that provide genome-wide coverage of most common variants, together with markers of interest for each of the five diseases identified through genome-wide association studies (GWAS), fine-mapping of known susceptibility regions, sequencing studies, and other approaches. The array also includes loci of interest identified through studies of other cancer types, and other loci of interest to multiple cancer types (including loci associated with cancer related phenotypes, drug metabolism and radiation response). Additionally, SNPs relating to quantitative phenotypes such as BMI, height, and breast density that correlate with common cancer risks are also included.

OncoArray Steering Committee:

- Transdisciplinary Research in Cancer of the Lung (TRICL)
  - Christopher Amos, Ph.D., Dartmouth College
  - Loic Le Marchand, M.D., M.P.H., Ph.D., Cancer Research Center of Hawaii, University of Hawaii

- Follow-up of Ovarian Cancer Genetic Association and Interaction Studies (FOCI)
  - Thomas Sellers, Ph.D., M.P.H., H. Lee Moffitt Cancer Center & Research Institute
  - Georgia Chenevix-Trench, Ph.D., QIMR Berghofer
  - Paul Pharoah, Ph.D., University of Cambridge

- ColoRectal Transdisciplinary Study (CORECT)
  - Stephen Gruber, M.D., Ph.D., M.P.H., University of Southern California

- Elucidating Loci Involved in Prostate Cancer Susceptibility (ELLIPSE)
  - Stephen Chanock, M.D., DCEG, NCI
  - Alison Dunning, Ph.D., University of Cambridge
  - Douglas Easton, Ph.D., University of Cambridge
  - Rosalind Eeles, Ph.D., F.C.R.P., F.R.C.R., The Institute of Cancer Research

- Discovery, Biology, and Risk of Inherited Variants in Breast Cancer (DRIVE)
  - David Hunter, M.B.B.S., Sc.D., Harvard University
  - Douglas Easton, Ph.D., University of Cambridge
  - Stephen Chanock, M.D., DCEG, NCI

- Breast Cancer Association Consortium

- Genome Canada/Genome Quebec/CIHR funded Personalised Risk Stratification for Prevention and Early Detection of Breast Cancer (PERSPECTIVE) international project

- Jacques Simard, Ph.D., Laval University
- Douglas Easton, Ph.D., University of Cambridge

- Cancer Research UK
- Douglas Easton, Ph.D., University of Cambridge
- Alison Dunning, Ph.D., University of Cambridge
- Paul Pharoah, Ph.D.
- Georgia Chenevix-Trench, Ph.D., QIMR Berghofer

- CIMBA
- Antonis Antoniou, Ph.D., University of Cambridge
- Georgia Chenevix-Trench, Ph.D., QIMR Berghofer
- Jacques Simard, Ph.D., Laval University

- NCI EGRP
- Stefanie Nelson, Ph.D. (liason)

- NCI DCEG
- Stephen Chanock, M.D.

- Daniela Seminara, Ph.D.
- NCI DCCPS Office of the Director

SNP Selection for the Oncoarray

General Principles

- SNP selection should be decided in collaboration between all the collaborating groups, i.e. all U19s plus any other groups providing resources (funding or datasets) for the initial project.

- The SNP content should made publicly available.

- The array will be made freely available for purchase by other groups.

- The OncoChip will include 600,000 beadtypes (somewhat less than 600,000 SNPs, because ambiguous C/T or A/T SNPs require two beadtypes).

- The content should be divided between the disease groups. As an initial proposal, these should be divided as follows:

    - Common content 60,000 (10% of the content)
    - The remaining cost to be divided in proportion to the total samples/budget (both CIDR and non-CIDR). Additionally, we decided to allocate 260,000 SNPs for a GWAS backbone so the remaining allocations were made proportional to the proportion of samples that were genotyped.

- Each disease group decided how to select SNPs however a guideline was adopted to ensure a level of consistency.

- The disease-specific components include:

    - Follow-up of combined GWAS/replication
    - Fine-mapping of known hits
    - Follow-up of rare variant/sequencing experiments
    - Ad-hoc candidates

- The relative contributions of each list were up to each disease group to decide.

Common Content

- SNP selection from meta-analysis across diseases (either overall or using mixture model)
- Lists for other cancers (say up to 1000 each, depending on availability)
- Fine-mapping of regions that are hits for more than one cancer type
  (*TERT*, 8q24 (proximal and distal to *MYC*), *HNF1B, TET2, RAD51B*, 11q13)

- QTLs:
    - Menarche
    - Menopause
    - Anthropometric (height, weight, BMI, WHR) *Try to include longer list this time from GIANT*
    - Telomere length

- Confirmed GW significant hits for all cancers and cancer-related phenotypes (e.g. smoking)

- Nominations from cross-site pathway analyses.

- Y and MT

GWAS replication

- Generally, best to base on full available data, i.e. combined GWAS+replication, imputed to 1KG.
- Remove highly correlated SNPs (r^2>0.8), but include additional surrogates for the most strongly associated markers.

    - Overall disease
    - Subtypes
    - Ethnicity specific analysis (lengths of these lists for Asian/African ancestry will depend on how many samples are likely to be genotyped, but we should try to include some).
    - Survival (were there is available data)

- Attempt to include all SNPs, at least for overall disease, that appear to be measurably predictive of risk in a predictive risk score analysis (for prostate, initial analyses suggest at least 10,000)

Fine mapping

Define regions to map, based on both LD and relevant genomic features (e.g. to cover regulatory regions pertinent to genes of interest, if known)

Define complete catalogue of variants (from 1KG, augmented with other sequence data if available)

Attempt to include all variants correlated with best hit, plus dense tagging set of remaining variants

Parameters will depend on number of regions to map and size. For iCOGS, r^2>0.1 was used.

Technical/organizational issues

- NCI DCEG was designated as responsible for the final list and its submission to Illumina. Each disease group was responsible for generating its own list.
- DCEG provided a shared space to exchange lists
- Admissible design score (0.8 was used, but a lower score was allowed for fine-mapping and candidates)

Merging process

Merging performed as a tree (scripts already available to implement this).

- For each disease, make separate lists for each category, ranked by importance
- Merge all the replication lists (choosing surrogates as necessary)
- Merge the replication, fine-mapping, rare variant and candidate lists, to make a final ranked list for each disease (*these lists can be 50-100% larger than the allocation, to allow for overlaps*).

- Final merging (across disease sites), in proportion to the SNP allocation (no surrogates chosen at this stage, only exact duplicates removed).

- GWAS framework included as an additional list, discarding SNPs selected for replication (not from the other lists) if an adequate GWAS SNP existed.

Carefully checked for errors (wrong alleles, position etc.)

SNP Selection - prostate

- Known index signals
- SNPs from COGS
- SNPs from meta of all cases in EAs
- SNPs from meta of adv cases in EAs
- SNPs from meta of all cases in AAs
- SNPs from meta of adv cases in AAs
- SNPs from meta of all cases in all groups
- SNPs from meta of adv cases in all groups
- Fine-mapping of known regions in EAs/AAs
- Top SNPs from Exome chip
- Rare variants from ICPCG (~1000)
- Candidates (~2000)
- PSA GWAS

SNP Selection – breast

Fine-mapping of known regions

Replication: combined analysis from GWAS+iCOGS (imputed to 1KG):

Overall disease (1df and 2df tests)
Disease <40
ER-negative
Grade
Breast density
Survival

Asian ancestry
African ancestry

Exome chip (~5,000SNPS)
Rare variants from COMPLEXO, other consortial nominations (allocate ~1,000)
Variants from whole genome sequencing
Candidates (allocate ~2,000)

SNP Selection – Lung 43,206 variants were nominated

GWAS and GWAS Meta-analyses replication

- Meta-analysis of 16 individual GWAS
- HapMap 2 based meta-analysis
- 1000Genome based meta-analysis
- GWAS in Asian and African-American

Tagging and Fine-mapping

- confirmed loci (5p15; 6p21-11; 9p21.3; 15q15.1; 15q25; 12p13.33, 22q12.2)

Individual Group Variants

- Candidate genes including IPF, asthma, COPD
- Rare variants from sequencing projects – TCGA data on lung adenocarcinoma, squamous carcinoma and head and neck cancers

- Lung eQTL variants
- Inflammation variants
- Histology pathway analysis
- COPD variants
- Tobacco metabolism and smoking phenotypes variants (placed in common area)

Oncoarray QC Guidelines

(All lists referred to should be available on the Oncoarray wiki:

http://consortia.ccge.medschl.cam.ac.uk/oncoarray)

1. Genotype Calling

Call all genotypes with the v2c cluster file. (Download from http://consortia.ccge.medschl.cam.ac.uk/oncoarray/onco_v2c.zip).

Export Illumina TOP alleles from Genome Studio.

2. Sample QC

2.1 Initial call rate filtering (by consortium)

Exclude samples with call rate <80%

Exclude SNPs with call rate <80%

Exclude samples with call rate <95%

Exclude SNPs with call rate <95%

2.2 Ancestry

Define set of uncorrelated markers (~3,000) including all AIMS.

Use to define individuals of European/East Asian/African American ancestry, or other, using Structure, MDS or LAMP. The Dartmouth group has defined principal components for identifying Continental ancestry and will send a procedure out using R. The Dartmouth group will is also deriving principal components using a panel of about 20,000 markers for deriving intraEuropean ancestry.

Consortium specific: for some groups, most studies will be (almost) single ethnicity (European or Asian) and best to exclude minority ancestry from these studies.

2.3 Heterozygosity

Exclude samples with heterozygosity<5% or > 40% and heterozygosity if $p<10^{-6}$, (|Z|>4.892) (GenABEL perid.summary). Test Asian and Europeans separately.

2.4 Sex checks

Exclude unexpected genotypic males/females/males (using X and Y markers). Also exclude XO, XXY, low X heterozygosity (<5%). Use list of 300 Y markers confirmed to work in males and to have non-autosomal cluster patterns (chr_Y_SNPs_for_sex_checking.csv). Exclude from the test chromosome X SNPs that show a high level of heterozygous calls in males and/or autosomal cluster patterns. (chr_X_SNPs_with_autosomal_clusters.csv.)

2.5 Duplicate concordance

Identify duplicates within study.

Check expected duplicates – if consistent exclude the sample with lower call rate.

Identify unexpected duplicates within studies. Liaise with study data-managers to attempt to resolve any discrepancies, remove both if not resolved.

Check with previous iCOGS or pre-iCOGS/GWAS genotyping

Exclude individuals discordant with previous consortium genotyping (*if study co-ordinator cannot resolve*).

2.6 Relatives

Relatives: Identify relatives. Individuals with estimated 0.55>ibd>0.45 were evaluated as likely first degree relatives.

These may be excluded by some of the consortia. For case-control pairs of relatives, exclude the control. Otherwise exclude the lower call rate sample.

2.7 Cross study/consortium duplicates

Check for duplicates across studies within the consortium - mark for exclusion from one study for main analyses except for study specific files.

Between Oncoarray and iCOGS/previous GWAS

Consortium specific – for BCAC/PRACTICAL, mark Oncoarray samples for exclusion in main analysis, but need 2nd version keeping all Oncoarray samples and excluding from iCOGS (for fine-mapping/rare variants).

TRICL retained Oncoarray samples and removed prior genotyping from previous meta-analyses then reperformed meta-analyses.

Across consortia

Generally only exclude for meta-analysis.

3. SNP QC by Consortia

3.1 Call rate

Exclude SNPs zeroed by the cluster file with no genotypes.

Exclude samples with call rate <80%

Exclude SNPs with call rate <80%

Exclude samples with call rate <95%

Exclude SNPs with call rate <95%

3.2 Hardy-Weinberg

Check Hardy-Weinberg: exclude SNP if $P<10^{-7}$ in controls or $P<10^{-12}$ in cases.

(In CIMBA, all subjects treated as controls.) Need to adjust for study (or country), and perform stratified score test. Test separately for Europeans/Asians/Africans. BCAC, OCAC and Practical excluded any SNP that failed in Europeans OR Asians.

4. SNP QC Exclusions Combined Across Consortia

4.1 Combine list of failures

All consortia to exclude SNPS that fail for call rate or HWE in any other consortium. (As at 1st April breast, ovarian, Cimba, prostate (Cambridge) exclusions have been combined, plus call rate exclusions for Lung.)

Chromosome Y exclusions were taken only from Practical. Practical used chromosome X HWE exclusions from BCAC.

4.2 Duplicate calling concordance

If the genotypes for pairs of duplicates differ >2% for any SNP, then exclude that SNP as unreliable. (Do not include differences between a no-call and called genotype.)

Duplicates concordance figures were combined from up to 5,250 duplicates from BCAC, OCAC, Practical, Cimba.

4.3 Duplicate probes

There are a number of variants on the chip with the same probe in the same position (or a few with the same alleles but the sequence from the opposite strand.)

A list (onco_duplicate_variants_excluded.csv) of 765 was compiled of duplicate probes that should be excluded. The probe with the worse QC scores and call rate was chosen for exclusion.

4.4 Cluster Plot Checking

Exclude SNPs where the cluster plot has been confirmed as “Failed” by two independent checks.

5. Additional Steps Before Imputation

5.1 Rare SNPs with poor call rate

Exclude SNPs with call rate below 98% and MAF <0.01 (Europeans) in any consortium from the imputation input files. (The genotyped calls for these SNPs can still be analysed.)

5.2 Non-ideal cluster plots

SNPs with cluster plots that were scored as Possible (P) or Subset interference (S) in the second round of checking should be excluded. These are either rare SNPs where there is no clear heterozygote cluster or SNPs with more than three clouds because of interference from other SNPs or possible copy number variation.

5.3 Variants unmatched to a 1000 Genomes variant

Strand information was obtained by blasting the Illumina TOP sequences against the 1000 genomes sequences. Some manifest positions identified by "rs" numbers were updated from dbSNP and the new positions confirmed by sequence matching.

The variants on the chip were then matched to the variants from Phase 3 variant set provided for the Impute software. (https://mathgen.stats.ox.ac.uk/impute/1000GP%20Phase%203%20haplotypes%206%20October%202014.html)

Variants were matched by position and alleles. Genotypes for variants not matched to a 1000G variant will be included in the imputation input files but marked so as not to be used by Impute.

5.4 Frequency Comparison to 1000 Genomes variants

Allele frequencies for controls from BCAC, OCAC and Practical were combined into a single frequency for Europeans (from 108,000 samples) and Asians (11,000 samples). These were tested against the expected frequency from 1000G using a test provided by Jon Tyrer.

A difference statistic is calculated by the formula:
$(|p1-p2|-0.01)_{+}^{\wedge 2}/((p1+p2)(2-p1-p2))$

where p1 and p2 are the frequencies our dataset and in the 1000 genomes respectively.

A cutoff of 0.008 in Europeans and 0.012 in Asians is needed to pass. Very rare SNPs are less likely to be rejected.

SNPs where the frequency would match if the alleles were flipped were excluded.

A list of strands and matched 1000G variants is provided.

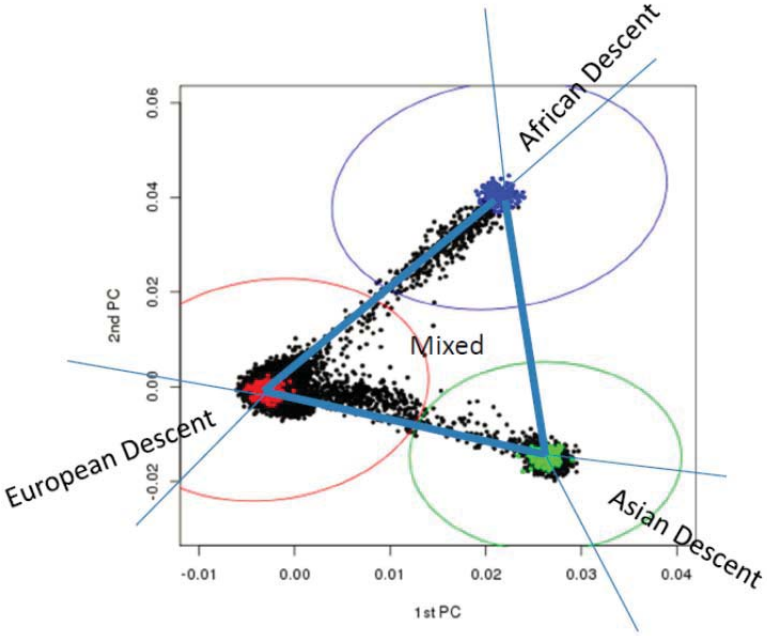
A list of SNP where the Illumina TOP alleles need flipping in order to match the 1000 Genomes alleles is provided.

6. Principal Components

Define Oncoarray consortium PCs and validate against some consortium specific PC definitions. Define a set of PCs for the European and Asian subsets, which could serve as covariates for them plus a global set to use for those of mixed ethnicity. It may also be important where possible to look at inflation in individual studies where specific PCs might be required (e.g. Finland, HMBCS).

The figure below describes either i) using PCA to classify according to ancestry (shown by most likely descent ellipses) or ii) assign continental origin to individuals according to the closest location on the continental ancestry triangle. We prefer the latter approach as the ancestry can then be used as a covariate in analyses or for subsequent selection.

Cluster in PCA



Red, CEU; green, CHB; blue, YRI.

CEU & YRI, ellipses cover samples with 7sd from the mean. CHB, ellipse cover samples with 6sd from the mean.

OncoArray Imputation

We used as reference Dataset the 1000 Genomes Project (GP) Phase 3 (Haplotype release date October 2014) for chromosomes 1 to 22. The 1000 Genomes Project Phase 1 (Haplotype ChrX release date Aug 2012) was used for chromosome X, since the phased data for Chr X from 1000GP Phase 3 is not available.

The OncoArray whole genome data were imputed in a two-stage procedure using SHAPEIT (shapeit.v2.r790.Ubuntu_12.04.4.static) to derive phased genotypes, and IMPUTEv2 (impute_v2.3.2_x86_64_static) to perform imputation of the phased data.

We used the default parameters used to derive phased genotypes with SHAPEIT, increasing:

- the number of burn-in iterations used by the algorithm to reach a good starting point to 10 ("--burn 10"),

- the number of pruning iterations used by the algorithm to find a parsimonious graph for each individual to 10 ("--prune 10"),

- and the number of iterations used by the algorithm to compute transition probabilities in the haplotype graphs to 50 ("--main 50")

We performed imputation with IMPUTEv2 using ~5Mb non-overlapping intervals for the whole genome. The flag "-use_prephased_g" was provided to indicate that pre-phased haplotypes were being used. In addition we excluded from imputation the 1000 GP variants whose minor allele frequency in Europeans and East Asians was lower than 0.001. The missing genotypes at typed SNPs were replaced with imputed genotypes using the option "-pgs_miss". The number of reference haplotypes to use as templates when imputing missing genotypes was increased to 800 ("-k_hap 800"), and the buffer region was increased to 500kb ("-buffer 500").

For the fine mapping regions we also imputed the non-genotyped data with IMPUTEv2 but without prephasing in SHAPEIT in order to improve imputation accuracy. For this we also increased:

- the default number of Markov chain Monte Carlo (MCMC) iterations (including burn-in) to 50 ("-iter 50"),

- the number of MCMC iterations to discard as burn-in to 15 ("-burnin 15"),

- and the number of haplotypes to use as templates when phasing observed genotypes to 100 ("-k 100").

Duplicated position issues

SHAPEIT cannot handle duplicated variants (same position, and same alleles). The program stops when these variants are detected.

IMPUTEv2 cannot handle duplicated positions (different genotyped variants at the same position). Thus, when the genotyped data includes that kind of variants (same position, different alleles), IMPUTEv2 gives the following warning:

“Position XXXX occurs multiple times in Panel 2. The first instance of this SNP will be used for inference, while all subsequent instances will be ignored and omitted form output files”.

Sometimes, when IMPUTEv2 identify more than one variant at the same position, and the alleles of one of these variants cannot be matched with the reference panel (noted in the summary file as "N of these replace existing SNPs with incompatible alleles"), a corrupted warning file (in binary format) is created indicating:

"The –known_haps_g alleles XX do not match the –g alleles XX".

On the contrary, if only one genotyped variant is located at the same position that a variant in the reference panel, and its alleles do not match the reference panel, IMPUTEv2 will consider this genotyped variant as a type 3 variant (present in the genotyped panel but not in the reference panel), and no warnings will be generated.

Therefore, we included for imputation only one of the variants that match the same position.