Confluence
Uncovering breast cancer genetics

Scientific Steering Committee

Member		Affiliation	Representing/coordinating
Montserrat	Garcia-Closas	Division of Cancer Epidemiology and Genetics, USA	Co-Chair (DCEG)
Doug	Easton	Cambridge University, England	Co-Chair (BCAC)
Jonas	Almeida	Division of Cancer Epidemiology and Genetics, USA	Data Science
Antonis	Antoniou	Cambridge University, England	CIMBA/Statistical Genetics
Jenny	Chang-Claude	German Cancer Research Center DKFZ, Heidelberg, Germany	BCAC risk factor working group
Nilanjan	Chatterjee	Johns Hopkins University, USA	Statistical genetics
Georgia	Chenevix-Trench	QIMR-Berghofer, Australia	CIMBA
Fergus	Couch	Mayo Clinic, USA	ENIGMA
Laura	Fejerman	University of California in San Francisco (UCSF), USA	LAGENO-BC
Judy	Garber	Harvard University, USA	Clinical Trials
Liz	Gillanders	Division of Cancer Control and Prevention Sciences, NCI, USA	DCCPS/NCI extramural
Chris	Haiman	University of South California, USA	AABCGS
Pete	Kraft	Harvard University, USA	Prospective cohorts/ Statistical genetics
Roger	Milne	University of Melbourne, Australia	BCAC DACC Chair
Nick	Orr	Queen's University Belfast, North Ireland	Male breast cancer studies
Julie	Palmer	Boston University, USA	AABCGS
Paul	Pharoah	Cambridge University, England	BCAC pathology/survival Working Group
Marjanka	Schmidt	Netherlands Cancer Institute, The Netherlands	BCAC pathology/survival Working Group
Jacques	Simard	University of Laval, Canada	PERSPECTIVE I&I
Wei	Zhang	Vanderbilt University, USA	ABCC and AABCGS

ABSTRACT

Genome wide association studies (GWAS) have been successful in identifying over 180 common susceptibility loci for breast cancer. However, heritability analyses indicate that breast cancer is a highly polygenic disease with thousands of common genetic variants of small effects, and that increasing sample sizes will generate new discoveries. The Confluence project aims to build a large research resource of over 300,000 cases and 300,000 controls of different ancestries—doubling current sample sizes to study the genetic architecture of breast cancer. This will be accomplished by the confluence of existing and new genome-wide genotyping data to be generated through this project. The specific aims of this project are: (1) to discover susceptibility loci and advance knowledge of etiology of breast cancer overall and by subtypes, (2) to develop polygenic risk scores and integrate them with known risk factors for personalized risk assessment for breast cancer overall and by subtypes, and (3) to discover loci for breast cancer prognosis, long-term survival, response to treatment, and second breast cancer. To be eligible to participate, studies with cases of *in situ* or invasive breast cancer (females or males) must have genome-wide genotyping data and/or germline DNA for genotyping, core phenotype data, and appropriate ethics approval for genetic studies and data sharing. During September-December 2018, we reached out to potential studies through existing GWAS consortia and other means to request interest in participating in this project. We received an excellent response demonstrating the feasibility of reaching the target number of cases and controls. This large increase in sample size and diversity of populations will enable discoveries that will lead to a better understanding of the etiology of distinct breast cancer subtypes and the role of genetic variation in prognosis and treatment response, thus improving risk stratification, prevention, and clinical care of breast cancer across ancestry groups.

BACKGROUND

Genome wide association studies (GWAS) have been successful in identifying over 180 common susceptibility loci for breast cancer through large consortia currently including ~150,000 cases (plus controls), mostly of European and East Asian ancestry[1-3]. These important discoveries, coupled with follow-up functional studies, are providing unprecedented insights into the biological mechanisms linking common genetic variation with breast cancer predisposition. These include understanding the impact of variation in regulatory regions across the genome, enrichment of specific transcription factor binding sites and the overlap between candidate target genes and somatic driver mutations in breast tumors [1, 2]. These studies have provided strong evidence for heterogeneity of breast cancer etiology, with many loci being differentially associated with subtypes of breast cancer [2, 4-8]. Recent analyses show that ER status is a major determinant for heterogeneity of genotype risk associations, followed by grade of differentiation [6]. In addition, this research has shown that most loci predisposing to breast cancer in the general population are associated with risk in high-risk populations, such as *BRCA1* and *BRCA2* mutation carriers [9, 10]. Finally, clinical studies suggest a contribution of germline variants to breast cancer prognosis [11-13], response to treatment and toxicities [14-31].

Most GWAS discoveries to date have been derived from analyses of populations of European ancestry in two highly successful multi-consortia efforts, the Collaborative Oncological Gene-environmental Study (COGS; *PIs: Per Hall and Doug Easton*) and the Oncoarray project (*PIs: Jacques Simard and Doug Easton)* in collaboration with the NCI-funded "Discovery, Biology and Risk of Inherited Variants in Breast Cancer Consortium" (DRIVE; *PI: Pete Kraft*). Studies have participated in these efforts through the well-established Breast Cancer Association Consortium (BCAC) led by *Doug Easton*, and the Consortium of Investigators of Modifiers of BRCA1/2 (CIMBA) led by *Georgia Chenevix-Trench* and *Antonis Antoniou [1, 2].* GWAS analyses in women of East Asian ancestry in BCAC and the Asia Breast Cancer Consortium (ABCC) led by *Wei Zheng* have identified additional loci, and shown associations with most loci identified in women of European ancestry [3, 32, 33]. GWAS in women of African ancestry have been substantially smaller and they have confirmed many associations found in other populations. However, few loci have been specifically identified in women of African ancestry [34-36]. This gap in knowledge is being addressed in an NCI-funded effort by the African-Ancestry Breast Cancer Genetic Study (AABCGS), a multi-consortia GWAS of over 16,000 cases and similar number of controls (*PIs: Wei Zheng; Chris Haiman; Julie Palmer).* GWAS in women of Latin American or US Latina women have been even smaller and thus, new efforts are needed to understand genetic predisposition in highly admixed Latina populations [37-39]. The Latin America Genomics Breast Cancer Consortium (LAGENO-BC) led by *Laura Fejerman* has recently been formed to address this question. GWAS of male breast cancer led by *Nick Orr*, although small in size due to the rarity of this disease in males, have shown differences in associations compared to females, with larger effect sizes for select known breast cancer loci [40]. Multi-ancestry analyses leveraging similarities across populations while accounting for differences, will be critical for discovery of new loci, as well as for fine-mapping and functional follow-up studies.

Previous investigations have provided evidence for the possible roles of germline variation in determining prognosis [11-13], differential responses to adjuvant therapies [14-18]; and for determining therapeutic toxicities associated with treatments such as aromatase inhibitors [18-21], taxanes [22-26], and radiotherapy [27-31]. However, additional clinical studies are needed to further investigate the contribution of germline variants to these clinical outcomes. The Confluence Project will provide an unprecedented opportunity to address these clinical questions by bringing together resources and expertise from international clinical trials and population-based studies.

Individual variants identified by GWAS are associated with small changes in risk; however, combining information on many common variants through the development of polygenic risk scores (PRS) can identify women, both with and without family history of breast cancer, at substantially different levels of genetic risk [41, 42]. Similar arguments apply for women with *BRCA* mutations [9]. For instance, a recently published PRS using information on 313 genetic variants (SNP: single nucleotide polymorphisms) can identify 1% of women of European ancestry with the highest

PRS who are three times more likely of developing breast cancer than women at average risk [41]. The 313-SNP PRS alone can provide levels of risk stratification in the population that are larger than those provided by classical risk factors (i.e. menstrual, reproductive, hormonal and lifestyle factors) or mammographic breast density (Choudhury et al. Under Review). An integrated model with classical risk factors, breast density and PRS would provide the highest level of risk stratification. An extension of the BOADICEA risk model to include the 313-variant PRS can also provide substantial information on risk among carriers of *BRCA1/2*, *PALB2*, *CHEK2*, and *ATM* [43]. Furthermore, PRSs for female breast cancer have been shown to be predictive of male breast cancer risk for *BRCA1* and *BRCA2* mutation carriers [44]. Thus, newly developed PRSs based on GWAS discoveries can substantially improve identification of women at different levels of risk in the population that could translate into improvements in stratified prevention and screening strategies for breast cancer [42, 45-47].

Analyses of the estimated underlying genetic architecture and effect size distribution of breast cancer susceptibility based on existing GWAS data indicate that this is a highly polygenic disease involving thousands of small risk variants, and that larger efforts should result in new discoveries. However, to make substantial gains in our understanding of breast cancer genetics, a major effort through strong collaborations across consortia and many studies is required. The Confluence project will function as a consortium of consortia to bring together existing GWAS data from about 150,000 cases and 200,000 controls and double it by generating new genotypes in female and male populations of different ancestry backgrounds from at least 150,000 new breast cancer cases and 100,000 controls (total of at least 300,000 cases and 300,000 controls).

The large increase in sample size and diversity of populations that will be attained through the Confluence Project will enable more powerful modeling of the underlying polygenic risk of breast cancer, which along with information on linkage disequilibrium and genomic annotations, can better inform our understanding of the etiology of distinct breast cancer subtypes. In addition, extension of clinical studies with survival data, treatments, toxicities and second breast cancers will substantially improve the power for analyses of the role of genetic variation for these outcomes. Overall, this work will result in improvements in risk stratification, prevention, and clinical care of breast cancer across ancestry groups.

OBJECTIVES

The Confluence project aims to build a large research resource of at least 300,000 breast cancer cases and 300,000 controls to conduct multi-ancestry genome wide association studies (GWAS) of breast cancer risk and prognosis.

Specific aims:

Aim 1: To discover susceptibility loci and advance knowledge of etiology of breast cancer overall and by subtypes.
Aim 2: To develop polygenic risk scores and integrate them with known risk factors for personalized risk assessment for breast cancer overall and by subtypes.
Aim 3: To discover loci for breast cancer prognosis, long-term survival, response to treatment, and second breast cancer.

In addition to the specific aims listed above, this resource will allow us to address a broad range of scientific questions in breast cancer genetics, and will serve as the basis for further studies that will require collection of additional data or materials, for instance:

- Discovery of new insights into biological mechanisms underlying genetic associations through follow-up functional laboratory-based studies.
- Integration with genomic characterization of tumors to understand germline-somatic relationships.
- Integration of validated prognostic loci or PRS in prognostication tools.
- Genetic predisposition to second cancers (other than breast cancer).

- Identification of genetic determinants for known or suspected risk factors and assessment of "causal" relationships through Mendelian Randomization analyses.
- Mosaicism/clonal hematopoiesis analyses.

Ongoing studies by collaborators in the Confluence project are aimed to identify and characterize rare variants and highly penetrant mutations through targeted, exome and whole genome sequencing strategies, e.g. CARRIERS (PIs: *Fergus Couch, Pete Kraft*), BRIDGES (PIs *Peter Devilee, Doug Easton*), PERSPECTIVE I&I (PI: *Jacques Simard*), and BRA-STRAP (PI: *Melissa Southery*). These projects (and CIMBA) are estimating the combined effect of common variants and rare variants (e.g. in *BRCA/PALB2/ATM/CHEK2* carriers). Addition of high-risk variants to the Confluence genotyping array chip, informed by these other ongoing efforts, will allow for a comprehensive assessment of the underlying architecture of breast cancer along the continuum from high to low risk variants.

Findings from the Confluence project will be highly relevant for science, public health, and clinical practice by advancing our understanding of the underlying genetic architecture of breast cancer for different subtypes and ancestry groups; improving breast cancer risk stratification in the general population and for moderate-high risk mutation carriers. Moreover, these results will establish the foundation for translational studies in stratified prevention through comprehensive breast cancer risk models.

PRELIMINARY DATA

After receiving a funding commitment from the NCI Director in September 2018, we approached studies through existing GWAS consortia (BCAC, CIMBA, AABCGS, ABCC), the extramural Division of Cancer Control and Prevention/NCI, leaders of previous GWAS in males and Latina women, leaders of clinical trials and other large breast cancer studies, as well as posted an invitation of participation in the Confluence project on the DCEG website. For planning purposes, we used an online study inventory to collect information from studies interested in participating in the Confluence project. As of the 10th of December 2018, we received expressions of interest from 136 studies with over 150,000 breast cancer cases requiring new genotyping (**Table 1**). Approximately one-half of these cases have DNA already extracted and one-half require new extractions. The female non-*BRCA* mutation carrier breast cancer cases are from 98 studies, of which 53 belong to BCAC, 40 are not part of any existing consortia, and 4 are part of other consortia. Most female and male *BRCA* mutation carrier studies are already part of CIMBA.

Table 1: Number of breast cancer cases and controls requiring new genotyping

Group	Breast cancer cases	Controls
Female non-carriers	125,352	302,310
Female BRCA1/2 carriers (CIMBA)	21,493	17,897
Males (carrier and non-carriers)	4,372	9,034
TOTAL	**151,217**	**329,241**

In addition, we have started to contact leaders of clinical trial studies and presented this project at the NCI Breast Cancer Steering Committee during the San Antonio Breast Cancer Symposium in December 2018. This exercise demonstrated that the goal of more than doubling the size of current GWAS studies by genotyping 150,000 new cases and 100,000 new controls through Confluence is feasible. It is also plausible that we could surpass this goal, which will add additional power to the discovery arm of this study. **Table 2** shows that a substantial proportion of female cases from non-carrier studies are of non-European (White) ancestry, and that new genotyping will

substantially increase the numbers of diverse populations. Efforts to identify studies will continue until reaching the desired numbers, with particular emphasis on identifying understudied populations and tumor types.

Table 2: Race/ethnic distribution of non-carrier female breast cancer cases

Race/ethnic group	New genotyping	%	Existing GWAS	%	TOTAL	%
White	91,834	73.3%	144,195	79.1%	236,029	76.7%
Asian	9,803	7.8%	14,068	7.7%	23,871	7.8%
Black or African American	6,670	5.3%	16,508	9.1%	23,178	7.5%
Hispanic or Latina	11,781	9.4%	7,488	4.1%	19,269	6.3%
Unknown	4,324	3.4%			4,324	1.4%
Other	813	0.6%			813	0.3%
American Indian or Alaska Native	109	0.1%			109	<0.1%
Native Hawaiian / Pacific Islander	18	<0.1%			18	<0.1%
TOTAL	**125,352**		**182,259**		**307,611**	

We did not collect information on pathology characteristics from the cases during this planning phase, but based on previous data, we expect about 70% of cases to be ER-positive and 30% ER-negative.

APPROACH

STUDY POPULATION

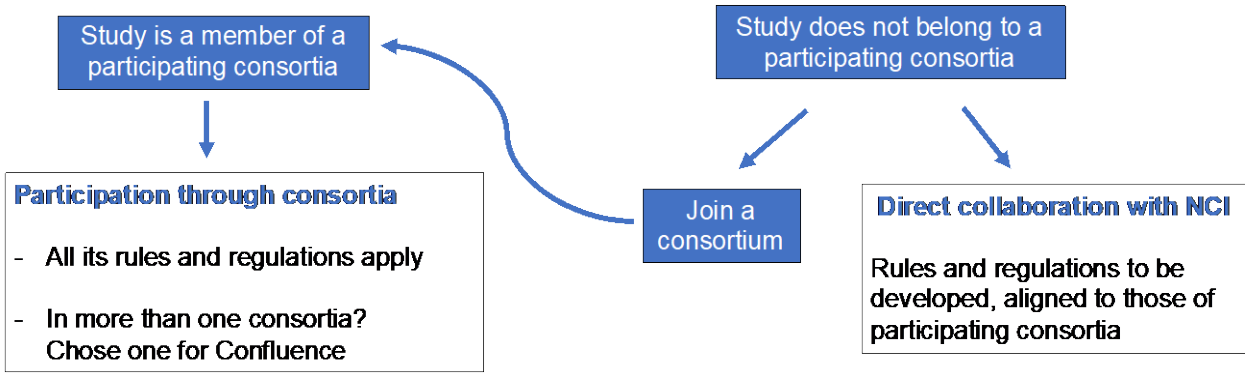
Studies must meet the following criteria to be eligible to participate:

- Studies of *in situ* or invasive breast cancer
  - Female or Male
  - Any subtype of breast cancer
- Genome-wide genotyping data or germline DNA for genotyping, i.e.:
  - existing genome-wide genotyping data, or
  - germline DNA available for new genotyping, or
  - blood/buccal samples for germline DNA isolation and genotyping.
- Core phenotype data (as defined below)
- Ethics approval and consent for genetic studies
- Data sharing plan

Studies can have a wide range of study designs, including case-control studies, prospective cohorts, clinical case series, clinical trials, or special cohorts such retrospective cohorts of *BRCA1/2* mutation carriers, or carriers of mutations in other established breast cancer susceptibility genes (e.g. *ATM*, *CHEK2*, *PALB2*). The design and data available will determine whether studies can participate in all or a subset of the study aims described below. If we identify more studies to be genotyped than required, priority will be given to studies based on study size, understudied populations, availability of extracted high quality DNA, high quality data on risk factors, tumor characteristics, treatment and clinical outcomes.

For studies that are already participating in breast cancer GWAS consortia, they can participate in the Confluence project through an existing consortium (**Figure 1**). Studies not already in consortia can participate by joining an existing consortium, forming a new group/consortia, or through a direct collaboration with DCEG, NCI.

Figure 1: Study participation through consortia or direct collaboration with DCEG/NCI



GENOTYPING

Participating studies will be able to provide: 1) existing individual-level germline genotype data from previously scanned samples, or 2) samples (extracted germline DNA or blood/buccal samples for extraction) not previously scanned to be genotyped through Confluence. Contribution of summary GWAS data from studies unable to provide individual-level data will also be considered.

New genotyping will be performed at two centers, the Cancer Genomics Research Laboratory (CGR) at DCEG/NCI (Stephen Chanock), and Strangways Laboratory at Cambridge University (UCAM, Doug Easton). Contribution of existing genotype data from studies that have been genotyped as part of an existing consortia (e.g. BCAC, CIMBA, AABCGS) can be done through the consortia after approval from individual studies. For studies that require new genotyping, the Confluence Project will cover the costs of sample shipment and materials (plates/tubes), DNA extractions (if needed), DNA quantitation/QC, return of left-over DNA (if requested), and genotyping and return of genotype files to contributing studies. However, Confluence will not be able to cover the costs from sample retrieval, preparation and aliquoting by individual studies.

Existing genotyping data from scanned samples

We will accept existing genotype data generated from eligible study samples using Illumina or Affymetrix chips; however, other methodologies may be considered. The following files/information will be requested:

- Genotyping chip and manifest files
- Genotype files: we can accept a range of data formats, including called genotype files with documentation of clustering/QC process, or pre-QC raw genotyping files.
- Sample sheet (ID mapping)

For studies contributing existing genotyping data through a consortium (e.g. BCAC, CIMBA), we anticipate requesting post-QC data along with the clustering/QC steps and metrics used. For studies willing to participate in genetic mosaicism studies, analysis will necessitate access to the B-allele frequency and Log R data from the scanned chips.

Genotyping chips for new genotyping

We will be using the Illumina Infinium Global Screening Array (GSA) with >665,000 variants in populations of non-African ancestry, and the Multi-Ethnic Genotyping Array (MEGA) with >1.3 million variants in populations of African ancestry because of its improved coverage and imputation accuracy in this population. A custom content of ~100,000 variants will be added to the arrays, and will include known pathogenic variants in breast cancer genes such as *BRCA1, BRCA2, ATM, PALB2* and *CHEK2*, as well as novel variants identified in ongoing efforts to identify and characterize rare variants and highly penetrant mutations through targeted, exome and whole genome sequencing (including CARRIERS, BRIDGES, PERSPECTIVE I&I, BRA-STRAP, AABCGS and ENIGMA). In addition, we will add content to facilitate fine mapping studies, copy number variation calling, and other questions of interest.

Biological sample requirements

For germline isolation we anticipate requesting the following specimen types as a source of germline DNA:

- 300-400uL (150uL minimum) of whole blood or buffy coat
- 1mL of saliva, Oragene™ or mouthwash/oral rinses

Tumor or serum samples *will not* be accepted as a source of germline DNA.

The anticipated DNA requirements for isolated germline DNA from blood or buccal sources are:

- 500-1000ng if dsDNA quantitation, e.g. PicoGreen
- 1.0-1.5ug if spectrophotometric quantitation, e.g. NanoDrop

The requested amounts are larger than the minimum input material for genotyping to ensure receiving adequate DNA for array work, anticipating large variation in quantification across different laboratories, and allowing for residual raw material to use in case of a failure. Reduction in total mass/volume can be requested by studies, with the understanding that the likelihood of sample failure will be higher and the ability for recovery efforts will be limited. Existing library preps from exome or whole genome sequencing might be accepted if native DNA is not available. Upon request, residual material will be returned to study sites at the completion of the work.

Quality Control, genotyping calling and imputation

Standard QC procedures will involve the following steps: 1. Sample and SNP level completion rate check; 2. Sample heterozygosity assessment; 3. Sample duplication/assay concordance check; 4. Sex verification; 5. Relatedness check (with allowances for study designs that include relatives by default, e.g. CIMBA); 6. Ancestry and population structure assessment; 7. Assessment on deviation from Hardy-Weinberg Proportions. Both existing and newly generated genotyping data will be imputed by chip and ancestry group using the most appropriate reference panels available at the time of analyses.

Return of genotyping data

For studies contributing samples for genotyping, the genotyping files/information detailed above will be returned to each participating study after QC procedures (we will consider requests for specific file formats from studies).

PHENOTYPE DATA

The following *core phenotype data will be required* from all participating studies: subject and sample IDs, age, sex, race/ethnicity, family history, and ER status (index tumor). Complete core data is not required, if it has not been collected by the study.

In addition, we will request the following data from studies that have it available (*not mandatory*):

Risk factors:

- Menstrual cycle: age menarche, menopausal status, age menopause
- Pregnancy: number of full-term births, age at first and last full-term birth, breastfeeding
- Height, weight, body mass index
- Oral contraceptives and menopausal therapy
- Alcohol consumption and cigarette smoking
- Benign breast disease and mammographic breast density

Pathology (first and second breast tumors):

- Behavior, Morphology
- Grade, Nodes, Size
- ER (core variable), PR, HER2, KI67 status

Treatment/clinical follow up:

- Treatment and toxicity information (to the extent available)
- Locoregional relapse, years to relapse, distant metastases
- Diagnosis of second breast cancer/s.
- Age at diagnosis, follow up time, vital status, cause of death for survival analyses

DATA MANAGEMENT

For studies contributing to the Confluence project through a consortia of studies, questionnaire, pathology and survival/treatment data the consortia data coordinating center will be the custodians of data, providing data management and harmonization: female breast cancer studies of European or East Asian ancestry will be managed by BCAC (Cambridge University, Netherlands Cancer Institute and German Cancer Research Center); female and male studies of *BRCA1/2* mutation carriers will be managed by CIMBA (Cambridge University); female studies of Hispanic/Latina ancestry will be managed by LAGENO-BC (UCSF); female studies of African ancestry (DCEG/NCI) and unselected male breast cancer studies will be managed by Nick Orr (Queens University). Data management and harmonization for studies contributing directly through NCI will be carried out by DCEG/NCI. The genotyping laboratories at DCEG/NCI and Cambridge University will be responsible for the management, QC and imputation of existing and new genome-wide genotyping data.

The aggregation of individual participant data on patient characteristics, treatment and follow up data on clinical outcomes and events (including toxicities) from clinical trials is a critical component for combined analyses to identify novel genetic determinants of clinical outcomes. We plan to accomplish this by establishing collaborations with existing clinical trial collaborative groups.

The Confluence project will cover the costs of study and data management by consortium data coordinating centers through contracts. Although the project will not cover costs from data preparation by individual studies according to the data dictionary, it will be able to assist studies and accept raw data for centralized data coding, if this work cannot be carried out by individual studies. Data management and stewardship will follow FAIR (Findability, Accessibility, Interoperability, and Reusability) principles [48].

ANALYTIC PLAN

Below is a description of the analytical plan to address the main aims of the Confluence project. However, it is anticipated that methodologies and functional annotations will continue to evolve, thus we will use the most appropriate methods available at the time of analyses.

Aim 1. *To discover susceptibility loci and advance knowledge of etiology of breast cancer overall and by subtypes*

The primary discovery analysis will involve standard single-SNP association testing across genome-wide panel of SNPs. As we would expect heterogeneity in associations across ethnic groups and subtypes, we plan to run additional association tests to maximize power in the presence of heterogeneous association, while borrowing strength across groups when associations are homogeneous (e.g. [49-51]). The primary goal would be to identify novel loci through association analysis combining all the data and then characterizing subtype and ethnicity-specific associations. Association analyses for *BRCA1/2* mutation carriers will be based on modeling the retrospective likelihood of the observed genotypes conditional on breast cancer phenotypes [52], using adjusted test statistic to allow for non-independence among related individuals and account for correlation in genotypes [53]. Analyses will be done by genotyping chip and combined using fixed-effects meta-analysis [54].

Transcriptome-wide Association Study (TWAS): In addition to single-SNP association test, we plan to carry out TWAS [55] by exploiting information on quantitative trait loci associated with gene-expression and other genomic characteristics, such as methylation. TWAS can improve power of discovery of genetic loci where multiple underlying variants affect disease mediated through an underlying common mechanism such as regulation of gene-expression. Recent studies have suggested that combining information from multiple tissues can improve the power for discovery of association analysis even for diseases that are very tissue specific [56-58]. Thus, we will consider cross-tissue TWAS analysis using the latest version of the Genotype-Tissue Expression (GTEx) and other genomic datasets.

Enrichment analyses: Based on the association statistics generated from GWAS, we plan to conduct enrichment analysis of association signals in relationship to functional genomic and population genetic characteristics (e.g. LD) of the genome characterized by ENCODE and other databases that may be available in the future. We anticipate to use stratified LD-score regressions [59] and related extensions for characterizing enrichment of association in multivariate models that can adjust for correlated annotations for each other. We will carry out the analyses with respect to both broad and cell-type specific annotations.

Fine mapping and functional analyses of identified signals using functional annotation data: We will conduct fine mapping analysis, informed with external functional data, around each locus identified through discovery stage of the analysis. We anticipate using Bayesian methods[60] that can integrate information on associations, local linkage disequilibrium pattern, and external functional information, such as eQTL characteristics, to compute posterior probabilities for each SNP within a fine-mapping region to be a causal variant.

Heritability analyses for breast cancer overall and by subtype across ancestry groups: Availability of large GWAS across subtypes and multiple ancestry groups will provide us the opportunity to explore the variation in genetic architecture of breast cancer in a more powerful way than it has been possible before. We will use state-of-the-art methods [61] for estimation of GWAS heritability to characterize how much of breast cancer risk variation can be explained by common variants across the different subtypes/ancestry groups. We will estimate degree of polygenicity and underlying effect-size distribution across different groups using methods we have recently developed [62]. We also plan to conduct GWAS co-heritability analysis to explore the overlap in genetic architecture across the different subtypes/mutation carriers/ancestry groups. Characterization of global genetic architecture of breast cancer using heritability, effect-size distribution, and genetic correlations will provide insight into similarity and differences in the genetic basis of breast cancer by different subtypes/ancestry groups and will allow us to explore the potential for genetic risk prediction at our current and future studies.

Aim 2: *To develop polygenic risk scores and integrate them with known risk factors for personalized risk assessment for breast cancer overall and by subtypes*

We plan to develop optimal PRS for predicting breast cancer across different ancestry groups and subtypes. Similar to association testing, our general strategy would be to utilize state of the art methods that can borrow strength across the different groups while allowing for potential heterogeneity in associations. Based on the experience of developing subtype-specific PRS for European ancestry GWAS data, we have found that a strategy of selecting SNPs based on global association analysis across groups and then estimating association coefficients of selected SNPs in group-specific manner leads to a robust strategy for building PRS. We will also explore alternative, more advanced, methods, such as Bayes/Empirical-Bayes techniques, that allows estimation of association coefficient for SNPs across different groups under a "prior" model that will account for suitable degree of heterogeneity. We will develop PRS for the general population and also assess whether PRS specific to mutation carriers (e.g. *BRCA1*, *BRCA2*, *ATM*, *CHEK2* or *PALB2*) are required for optimal risk prediction in these high-risk populations.

Genotypes will be aggregated with data on risk factors (reproductive and hormonal factors, anthropometry, alcohol consumption and other lifestyle factors, family history and breast features including benign breast disease and mammographic density). This will enable evaluations of gene-environment interactions, and development of population-specific risk models for overall and subtype risk predictions. We will use the iCARE and BOADICEA risk models [43, 63] to combine information on PRS, classical risk factors/family history and population incidence rates to develop integrated models for predicting absolute risk of breast cancer by subtypes and ancestry groups. Data from a parallel collaboration with the NCI Cohort Consortium to build risk prediction models will be used to obtain precise risk estimates of associations for classical risk factors by breast cancer subtypes and ancestry groups.

Aim 3: *To discover loci for breast cancer prognosis, long-term survival, response to treatment, and second breast cancer*

Analyses for aim 3 will be limited to cases with information on clinical prognosis, treatment, toxicities and second breast cancers. We plan to conduct standard genome-wide survival analysis using a Cox proportional hazard model framework for breast cancer outcomes (e.g. breast cancer specific mortality, total mortality, and diagnosis of a second breast cancer following the index breast cancer). Time-to-event will be calculated from the date of diagnosis with left truncation to account for cases enrolled into studies after diagnosis (prevalent cases). Analyses will be stratified by tumor characteristics and treatment to evaluate treatment response. We will also evaluate heterogeneity in associations by factors such as tumor characteristics, ancestry group and treatment using standard interaction analyses, as well as newer methods for combined association analysis (see Aim 1). For studies with information on treatment-related toxicities (primarily clinical trials), we will conduct candidate and genome-wide survival analyses to identify genetic determinants of toxicities that could range from radiation exposure (e.g. fibrosis), aromatase inhibitors (e.g. musculoskeletal adverse events) and chemotherapy (e.g. anemia, febrile neutropenia, peripheral neuropathy).

The lead statisticians for the Confluence SSC are *Nilanjan Chatterjee*, *Doug Easton*, *Antonis Antoniou* and *Pete Kraft*. They will provide oversight and expertise for the statistical analyses plans to address the main aims of the project as outlined above. However, they will not bear sole reasonability for data analyses. It is anticipated that primary statistical analyses will be performed in collaboration across analytical teams led by different members of the SSC. Other investigators will be able to propose and lead additional analyses through the submission of study concepts via the Confluence Data Platform (see below).

PROJECTED DISCOVERIES AND IMPROVEMENT IN RISK STRATIFICATION

Using GENESIS, a novel method to characterize the effect size distribution of common variants based on existing summary-level GWAS data [62], we estimated that there are over 5,000 common susceptibility variants for breast cancer (MAF>5%), most of them with very small (OR<1.01) effect sizes [50]. The proposed sample size of at least 300,000 cases and 300,000 controls was chosen to increase the percentage polygenic variance for overall breast cancer explained by genome-wide significant variants from about 40% to nearly 60% [1], (Zhang et al. In preparation). This effort should identify virtually all variants with OR>=1.02, and about half of variants with OR ~1.01. Identification of additional variants will require much large sample sizes due to their very small effect sizes**.** Substantial improvements in risk stratification are also expected by the addition of improved polygenic risk scores to breast cancer risk models [64].

GOVERNANCE, SCIENTIFIC REVIEW AND DATA SHARING

Governance

The organizational structure has been designed to ensure close involvement of participating studies and consortia in the governance, oversight and operations of the Confluence Project:

- *Scientific Steering Committee (SSC)* co-chaired by the DCEG Deputy Director (Montse Garcia-Closas) and the BCAC lead (Doug Easton) includes representatives from all participating consortia (see full membership on cover page), and other large contributing studies or groups of studies. The mission of this committee is to bring together representatives of different collaborative groups, provide scientific expertise, contribute to the development of the research plan and provide oversight of the research resource for use by the wider scientific community. The SSC reports to the director of DCEG (Stephen Chanock), source of funding for Confluence
- *External Advisory Group* will be formed by international experts in GWAS and advocates to provide logistical and scientific advice to the Confluence Project.

DCEG will be responsible for the overall coordination of the Confluence project, including management, integration and analyses by participating groups and consortia. However, each consortium will be responsible of the management and governance of data from their member studies, according to their rules and regulations.

Scientific Review of the Confluence Project

The Confluence Project has been reviewed by the NCI intramural review process used by all projects funded by the NCI Intramural Research Program. This involves an initial internal review by the *DCEG Senior Advisory Group (SAG)* that includes DCEG senior leadership and two or more *ad hoc* expert reviewers. DCEG SAG is advisory to the DCEG Director. The *NCI Board of Scientific Counselors for Clinical Sciences and Epidemiology* will provide external review of the progress and scientific output of the Confluence Project.

Data Sharing Plan

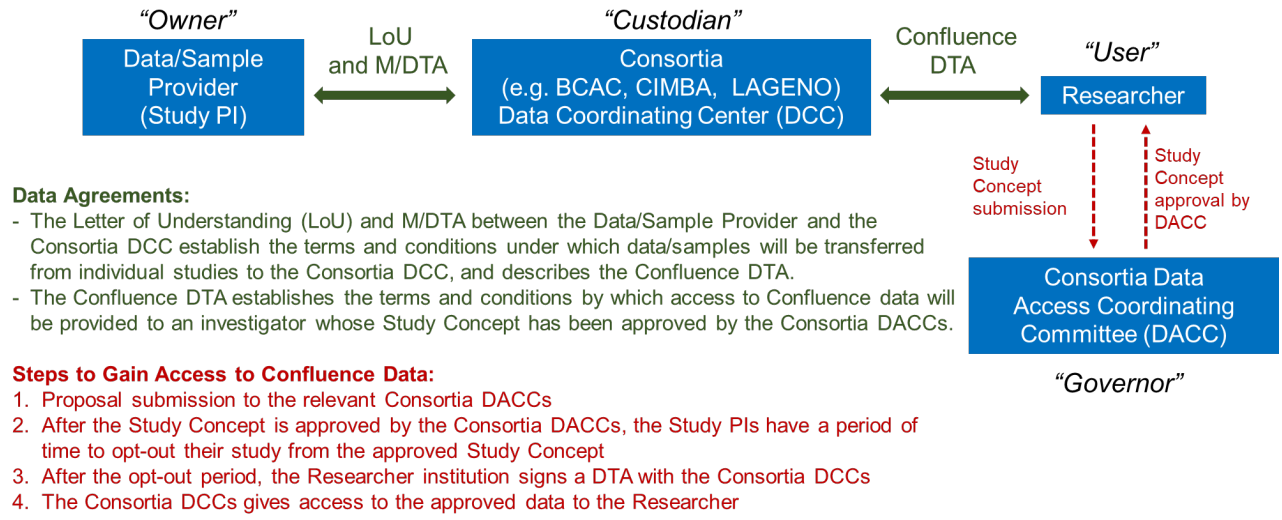
Summary-level data from analyses performed under the Confluence project will be broadly available to the scientific community at the time of manuscript publication reporting the main findings from the project. In addition, individual-level data will be accessible through two mechanisms:

A. Controlled data access through the Confluence Data Platform by eligible researchers:

Eligible researchers will be able to request access to individual-level data for specific analyses through the *Confluence Data Platform* that will be securely hosted in a Cloud environment (**Figure 2**). This platform will be designed to

manage and facilitate data intake, access, governance, visualization and analyses of data following FAIR principles [48], compatible with individual study IRB's and consortia policies. This approach will greatly facilitate collaborative analyses across multiple groups in a shared analytical space. Data will only be shared for academic research, not for commercial use.

Figure 2: Anticipated steps to gain access to Confluence Data



The ownership of the data will stay with the individual studies (i.e. the data/sample provider). For studies contributing data to the Confluence project through consortia, the custodian of the data will be the consortium data coordinating center (DCC; e.g. Cambridge University for BCAC or CIMBA), and data access will be governed by the consortium Data Access Coordinating Committee (DACC). A Letter of Understanding (LoU) and Material/Data Transfer Agreement (M/DTA) between the Data/Sample Provider and the Consortia DCC establish the terms and conditions under which data/samples will be transferred from individual studies to the Consortia DCC and describes the Confluence DTA. The Confluence DTA establishes the terms and conditions by which access to Confluence data will be provided to a researcher whose Study Concept has been approved by the Consortia DACCs. Researchers will then be able to visualize and analyze the data in the Cloud without data downloads (i.e. *code travels to the data*). Exceptions might be possible depending on analytical requirements.

The anticipated process to get access to Confluence Data by eligible researchers is:

- Researcher submits a study concept describing the project, including variables of interest, via the Confluence Data Platform to the consortia DACCs that govern the requested data.
- After approval by the relevant consortia DACCs, individual studies contributing data are notified and given a time period to opt-out their study from the approved project.
- After the opt-out period has elapsed, the researcher's institution signs a DTA for the study concept with the consortium data coordinating center(s) governing the data.
- Upon DTA signature, the data coordinating center(s) will be able to provide access of the approved data to researchers through the Confluence Data Platform. Requests will be digitally linked to specific variables so that following all required approvals access to the requested data from studies that do not opt-out will be "automatically" provided.

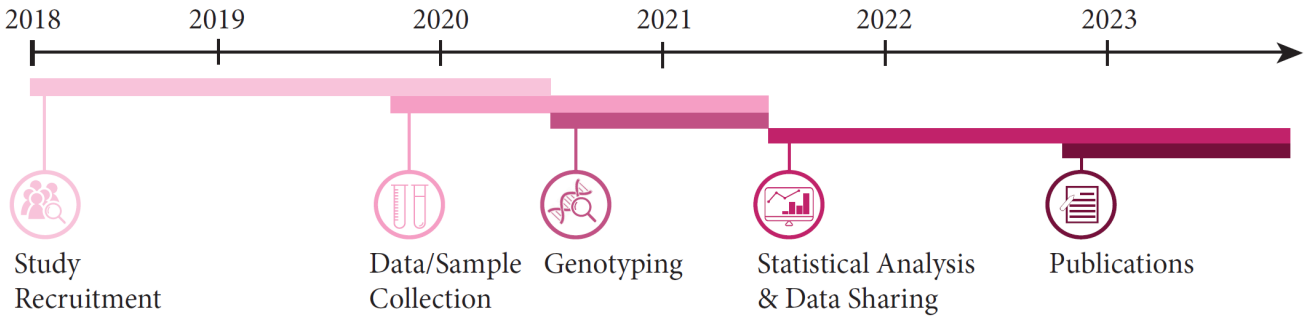
DCEG will work with the consortia DACCs to develop procedures/policies that facilitate data access/sharing across multiple consortia that are consistent with individual consortium data sharing policies and the NIH Genomic Data Sharing policy (see below), and to develop authorship guidelines for Confluence publications. To improve data

traceability and reproducibility of analyses/results, the data access policy for the Confluence project will be to provide data access for analyses through the Confluence Data Platform without downloading the data. Special requests for downloading the data will be considered if required analysis tools are not available on the data platform, and if data cannot be read remotely by the analysis tools.

B. Public data access through an NCI-approved data archive (e.g. dbGAP, EGA):

In accordance to the NIH Genomic Data Sharing policy, individual–level genotyping data generated using funds from the Confluence Project must be submitted for public access to an NCI-approved data archive such as the NIH database of Genotypes and Phenotypes (dbGaP), or the European Genome-phenome Archive (EGA), along with associated core phenotype data. To ensure institutional commitment to this policy, an NCI/NIH Genomics Data Sharing Plan Form detailing the required level of commitment on data sharing will need to be signed by studies *prior to* genotyping of samples.

This requirement ***does not apply*** to genotyping data generated using non-NIH funding.



APPROXIMATE TIMELINE

Interested in participating or have questions?

ConfluenceProject@mail.nih.gov

ABBREVIATIONS

AABCGS: African-Ancestry Breast Cancer Genetic Study

ABCC: Asia Breast Cancer Consortium

BCAC: Breast Cancer Association Consortium (http://bcac.ccge.medschl.cam.ac.uk/)

BOADICEA: Breast and Ovarian Analysis of Disease Incidence and Carrier Estimation Algorithm

BRASTRAP: BRCA Refined Analysis of Sequence Tests: Risk And Penetrance (https://www.southeylab.org/bra-strap-project)

BRIDGES: Breast Cancer Risk after Diagnostic Gene Sequencing (https://bridges-research.eu/)

CARRIERS: Cancer Risk Estimates Related to Susceptibility Genes

CIMBA: Consortium of Investigators of Modifiers of BRCA1/2 (http://cimba.ccge.medschl.cam.ac.uk/)

DCC: Data Coordinating Center

DACC: Data Access Coordinating Committee

DCEG: Division of Cancer Epidemiology and Genetics (https://dceg.cancer.gov/)

SAG: Senior Advisory Group

DNA: Deoxyribonucleic acid

DTA: Data Transfer Agreement

ENIGMA: Evidence-based Network for the Interpretation of Germline Mutant Alleles (https://enigmaconsortium.org)

FAIR: Findability, Accessibility, Interoperability, and Reusability

GWAS: Genome Wide Association Studies

HER2: Human Epidermal Growth Factor Receptor 2

LAGENO-BC: Latin America Genomics Breast Cancer Consortium

LoU: Letter of Understanding

M/DTA: Material/Data Transfer Agreement

NCI: National Cancer Institute (https://www.cancer.gov/)

PERSPECTIVE I&I: Personalized risk assessment for prevention and early detection of breast cancer: Integration and Implementation (https://www.genomecanada.ca/en/personalized-risk-assessment-prevention-and-early-detection-breast-cancer-integration-and)

PRS: Polygenic Risk Scores

References

1. Michailidou, K, Lindstrom S, Dennis J, Beesley J, Hui S, Kar S, Lemacon A, Soucy P, Glubb D, Rostamianfar A, et al., *Association analysis identifies 65 new breast cancer risk loci.* Nature, 2017. **551**(7678): p. 92-94.
2. Milne, RL, Kuchenbaecker KB, Michailidou K, Beesley J, Kar S, Lindstrom S, Hui S, Lemacon A, Soucy P, Dennis J, et al., *Identification of ten variants associated with risk of estrogen-receptor-negative breast cancer.* Nat Genet, 2017. **49**(12): p. 1767-1778.
3. Cai, Q, Zhang B, Sung H, Low SK, Kweon SS, Lu W, Shi J, Long J, Wen W, Choi JY, et al., *Genome-wide association analysis in East Asians identifies breast cancer susceptibility loci at 1q32.1, 5q14.3 and 15q26.1.* Nat Genet, 2014. **46**(8): p. 886-90.
4. Garcia-Closas, M, Couch FJ, Lindstrom S, Michailidou K, Schmidt MK, Brook MN, Orr N, Rhie SK, Riboli E, Feigelson HS, et al., *Genome-wide association studies identify four ER negative-specific breast cancer risk loci.* Nat Genet, 2013. **45**(4): p. 392-8, 398e1-2.
5. Sawyer, E, Roylance R, Petridis C, Brook MN, Nowinski S, Papouli E, Fletcher O, Pinder S, Hanby A, Kohut K, et al., *Genetic predisposition to in situ and invasive lobular carcinoma of the breast.* PLoS Genet, 2014. **10**(4): p. e1004285.
6. Ahearn, TU, Zhang H, Lecarpentier J, Michailidou K, Milne RL, Couch FJ, Simard J, Kraft P, Easton D, Pharoah P, et al., *Novel analysis incorporating multiple tumor characteristics provide evidence of highly heterogeneous associations for known breast cancer risk loci*, in *The American Society of Human Genetics Annual Meeting*. 2017: Orlando, FL. p. 657.
7. Couch, FJ, Hart SN, Sharma P, Toland AE, Wang X, Miron P, Olson JE, Godwin AK, Pankratz VS, Olswold C, et al., *Inherited mutations in 17 breast cancer susceptibility genes among a large triple-negative breast cancer cohort unselected for family history of breast cancer.* J Clin Oncol, 2015. **33**(4): p. 304-11.
8. Zhang, H, Lecarpentier J, Ahearn TU, Michailidou K, Milne RL, Kraft P, Simard J, Pharoah P, Schmidt MK, Easton D, et al., *Genome-wide association study (GWAS) identifies 9 novel breast cancer loci from analyses accounting for subtype heterogeneity*, in *The American Society of Human Genetics Annual Meeting*. 2017: Orlando, FL.
9. Kuchenbaecker, KB, McGuffog L, Barrowdale D, Lee A, Soucy P, Dennis J, Domchek SM, Robson M, Spurdle AB, Ramus SJ, et al., *Evaluation of Polygenic Risk Scores for Breast and Ovarian Cancer Risk Prediction in BRCA1 and BRCA2 Mutation Carriers.* J Natl Cancer Inst, 2017. **109**(7).
10. Kuchenbaecker, KB, Neuhausen SL, Robson M, Barrowdale D, McGuffog L, Mulligan AM, Andrulis IL, Spurdle AB, Schmidt MK, Schmutzler RK, et al., *Associations of common breast cancer susceptibility alleles with risk of breast cancer subtypes in BRCA1 and BRCA2 mutation carriers.* Breast Cancer Res, 2014. **16**(6): p. 3416.
11. Garcia, ME, Guo Q, Wessels L, Bader G, Pharoah P, Chenevix-Trench G, Easton D, Canisius S, and Schmidt M, *Abstract 2271: Pathway analysis suggests biological processes driven by germline genetic associations with breast cancer prognosis.* Cancer Research, 2018. **78**(13 Supplement): p. 2271-2271.
12. Guo, Q, Schmidt MK, Kraft P, Canisius S, Chen C, Khan S, Tyrer J, Bolla MK, Wang Q, Dennis J, et al., *Identification of novel genetic markers of breast cancer survival.* J Natl Cancer Inst, 2015. **107**(5).
13. Pirie, A, Guo Q, Kraft P, Canisius S, Eccles DM, Rahman N, Nevanlinna H, Chen C, Khan S, Tyrer J, et al., *Common germline polymorphisms associated with breast cancer-specific survival.* Breast Cancer Res, 2015. **17**: p. 58.
14. Gavin, PG, Song N, Kim SR, Lipchik C, Johnson NL, Bandos H, Finnigan M, Rastogi P, Fehrenbacher L, Mamounas EP, et al., *Association of Polymorphisms in FCGR2A and FCGR3A With Degree of Trastuzumab Benefit in the Adjuvant Treatment of ERBB2/HER2-Positive Breast Cancer: Analysis of the NSABP B-31 Trial.* JAMA Oncol, 2017. **3**(3): p. 335-341.

15. Hurvitz, SA, Betting DJ, Stern HM, Quinaux E, Stinson J, Seshagiri S, Zhao Y, Buyse M, Mackey J, Driga A, et al., *Analysis of Fcgamma receptor IIIa and IIa polymorphisms: lack of correlation with outcome in trastuzumab-treated breast cancer patients.* Clin Cancer Res, 2012. **18**(12): p. 3478-86.
16. Goetz, MP, Sangkuhl K, Guchelaar HJ, Schwab M, Province M, Whirl-Carrillo M, Symmans WF, McLeod HL, Ratain MJ, Zembutsu H, et al., *Clinical Pharmacogenetics Implementation Consortium (CPIC) Guideline for CYP2D6 and Tamoxifen Therapy.* Clin Pharmacol Ther, 2018. **103**(5): p. 770-777.
17. Khan, S, Fagerholm R, Kadalayil L, Tapper W, Aittomaki K, Liu J, Blomqvist C, Eccles D, and Nevanlinna H, *Meta-analysis of three genome-wide association studies identifies two loci that predict survival and treatment outcome in breast cancer.* Oncotarget, 2018. **9**(3): p. 4249-4257.
18. Leyland-Jones, B, Gray KP, Abramovitz M, Bouzyk M, Young B, Long B, Kammler R, Dell'Orto P, Biasi MO, Thürlimann B, et al., *CYP19A1 polymorphisms and clinical outcomes in postmenopausal women with hormone receptor-positive breast cancer in the BIG 1–98 trial.* Breast Cancer Research and Treatment, 2015. **151**(2): p. 373-384.
19. Ingle, JN, Schaid DJ, Goss PE, Liu M, Mushiroda T, Chapman JA, Kubo M, Jenkins GD, Batzler A, Shepherd L, et al., *Genome-wide associations and functional genomic studies of musculoskeletal adverse events in women receiving aromatase inhibitors.* J Clin Oncol, 2010. **28**(31): p. 4674-82.
20. Johansson, H, Gray KP, Pagani O, Regan MM, Viale G, Aristarco V, Macis D, Puccio A, Roux S, Maibach R, et al., *Impact of CYP19A1 and ESR1 variants on early-onset side effects during combined endocrine therapy in the TEXT trial.* Breast Cancer Res, 2016. **18**(1): p. 110.
21. Leyland-Jones, B, Gray KP, Abramovitz M, Bouzyk M, Young B, Long B, Kammler R, Dell'Orto P, Biasi MO, Thürlimann B, et al., *ESR1 and ESR2 polymorphisms in the BIG 1-98 trial comparing adjuvant letrozole versus tamoxifen or their sequence for early breast cancer.* Breast Cancer Research and Treatment, 2015. **154**(3): p. 543-555.
22. Abraham, JE, Guo Q, Dorling L, Tyrer J, Ingle S, Hardy R, Vallier AL, Hiller L, Burns R, Jones L, et al., *Replication of genetic polymorphisms reported to be associated with taxane-related sensory neuropathy in patients with early breast cancer treated with Paclitaxel.* Clin Cancer Res, 2014. **20**(9): p. 2466-75.
23. Baldwin, RM, Owzar K, Zembutsu H, Chhibber A, Kubo M, Jiang C, Watson D, Eclov RJ, Mefford J, McLeod HL, et al., *A genome-wide association study identifies novel loci for paclitaxel-induced sensory peripheral neuropathy in CALGB 40101.* Clin Cancer Res, 2012. **18**(18): p. 5099-109.
24. Schneider, BP, Li L, Radovich M, Shen F, Miller KD, Flockhart DA, Jiang G, Vance G, Gardner L, Vatta M, et al., *Genome-Wide Association Studies for Taxane-Induced Peripheral Neuropathy in ECOG-5103 and ECOG-1199.* Clin Cancer Res, 2015. **21**(22): p. 5082-5091.
25. Sucheston-Campbell, LE, Clay-Gilmour AI, Barlow WE, Budd GT, Stram DO, Haiman CA, Sheng X, Yan L, Zirpoli G, Yao S, et al., *Genome-wide meta-analyses identifies novel taxane-induced peripheral neuropathy-associated loci.* Pharmacogenet Genomics, 2018. **28**(2): p. 49-55.
26. Dorling, L, Kar S, Michailidou K, Hiller L, Vallier AL, Ingle S, Hardy R, Bowden SJ, Dunn JA, Twelves C, et al., *The Relationship between Common Genetic Markers of Breast Cancer Risk and Chemotherapy-Induced Toxicity: A Case-Control Study.* PLoS One, 2016. **11**(7): p. e0158984.
27. Andreassen, CN, Rosenstein BS, Kerns SL, Ostrer H, De Ruysscher D, Cesaretti JA, Barnett GC, Dunning AM, Dorling L, West CML, et al., *Individual patient data meta-analysis shows a significant association between the ATM rs1801516 SNP and toxicity after radiotherapy in 5456 breast and prostate cancer patients.* Radiother Oncol, 2016. **121**(3): p. 431-439.
28. Barnett, GC, Thompson D, Fachal L, Kerns S, Talbot C, Elliott RM, Dorling L, Coles CE, Dearnaley DP, Rosenstein BS, et al., *A genome wide association study (GWAS) providing evidence of an association between common genetic variants and late radiotherapy toxicity.* Radiother Oncol, 2014. **111**(2): p. 178-85.

29. Grossberg, AJ, Lei X, Xu T, Shaitelman SF, Hoffman KE, Bloom ES, Stauder MC, Tereffe W, Schlembach PJ, Woodward WA, et al., *Association of Transforming Growth Factor beta Polymorphism C-509T With Radiation-Induced Fibrosis Among Patients With Early-Stage Breast Cancer: A Secondary Analysis of a Randomized Clinical Trial.* JAMA Oncol, 2018. **4**(12): p. 1751-1757.
30. Zhao, J, Zhi Z, Zhang M, Li Q, Li J, Wang X, and Ma C, *Predictive value of single nucleotide polymorphisms in XRCC1 for radiation-induced normal tissue toxicity.* Onco Targets Ther, 2018. **11**: p. 3901-3918.
31. Dorling, L, Barnett GC, Michailidou K, Coles CE, Burnet NG, Yarnold J, Elliott RM, Dunning AM, Pharoah PD, and West CM, *Patients with a High Polygenic Risk of Breast Cancer do not have An Increased Risk of Radiotherapy Toxicity.* Clin Cancer Res, 2016. **22**(6): p. 1413-20.
32. Zheng, W, Long J, Gao YT, Li C, Zheng Y, Xiang YB, Wen W, Levy S, Deming SL, Haines JL, et al., *Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1.* Nat Genet, 2009. **41**(3): p. 324-8.
33. Han, MR, Long J, Choi JY, Low SK, Kweon SS, Zheng Y, Cai Q, Shi J, Guo X, Matsuo K, et al., *Genome-wide association study in East Asians identifies two novel breast cancer susceptibility loci.* Hum Mol Genet, 2016. **25**(15): p. 3361-3371.
34. Feng, Y, Rhie SK, Huo D, Ruiz-Narvaez EA, Haddad SA, Ambrosone CB, John EM, Bernstein L, Zheng W, Hu JJ, et al., *Characterizing Genetic Susceptibility to Breast Cancer in Women of African Ancestry.* Cancer Epidemiol Biomarkers Prev, 2017. **26**(7): p. 1016-1026.
35. Huo, D, Feng Y, Haddad S, Zheng Y, Yao S, Han YJ, Ogundiran TO, Adebamowo C, Ojengbede O, Falusi AG, et al., *Genome-wide association studies in women of African ancestry identified 3q26.21 as a novel susceptibility locus for oestrogen receptor negative breast cancer.* Hum Mol Genet, 2016. **25**(21): p. 4835-4846.
36. Feng, Y, Stram DO, Rhie SK, Millikan RC, Ambrosone CB, John EM, Bernstein L, Zheng W, Olshan AF, Hu JJ, et al., *A comprehensive examination of breast cancer risk loci in African American women.* Hum Mol Genet, 2014. **23**(20): p. 5518-26.
37. Fejerman, L, Ahmadiyeh N, Hu D, Huntsman S, Beckman KB, Caswell JL, Tsung K, John EM, Torres-Mejia G, Carvajal-Carmona L, et al., *Genome-wide association study of breast cancer in Latinas identifies novel protective variants on 6q25.* Nat Commun, 2014. **5**: p. 5260.
38. Fejerman, L, Chen GK, Eng C, Huntsman S, Hu D, Williams A, Pasaniuc B, John EM, Via M, Gignoux C, et al., *Admixture mapping identifies a locus on 6q25 associated with breast cancer risk in US Latinas.* Hum Mol Genet, 2012. **21**(8): p. 1907-17.
39. Hoffman, J, Fejerman L, Hu D, Hunstman S, Li M, John E, Torres-Mejia G, Kushi L, Ding YC, Weitzel J, et al., *Identification of Novel Common Breast Cancer Risk Variants in Latinas at the 6q25 Locus.* bioRxiv, 2018: p. 343806.
40. Orr, N, Lemnrau A, Cooke R, Fletcher O, Tomczyk K, Jones M, Johnson N, Lord CJ, Mitsopoulos C, Zvelebil M, et al., *Genome-wide association study identifies a common variant in RAD51B associated with male breast cancer risk.* Nat Genet, 2012. **44**(11): p. 1182-4.
41. Mavaddat, N, Michailidou K, Dennis J, Lush M, Fachal L, Lee A, Tyrer JP, Chen TH, Wang Q, Bolla MK, et al., *Polygenic Risk Scores for Prediction of Breast Cancer and Breast Cancer Subtypes.* Am J Hum Genet, 2018.
42. Mavaddat, N, Pharoah PD, Michailidou K, Tyrer J, Brook MN, Bolla MK, Wang Q, Dennis J, Dunning AM, Shah M, et al., *Prediction of breast cancer risk based on profiling with common genetic variants.* J Natl Cancer Inst, 2015. **107**(5).
43. Lee, A, Mavaddat N, Wilcox AN, Cunningham AP, Carver T, Hartley S, Babb de Villiers C, Izquierdo A, Simard J, Schmidt MK, et al., *BOADICEA: a comprehensive breast cancer risk prediction model incorporating genetic and nongenetic risk factors.* Genet Med, 2019.

44. Lecarpentier, J, Silvestri V, Kuchenbaecker KB, Barrowdale D, Dennis J, McGuffog L, Soucy P, Leslie G, Rizzolo P, Navazio AS, et al., *Prediction of Breast and Prostate Cancer Risks in Male BRCA1 and BRCA2 Mutation Carriers Using Polygenic Risk Scores.* J Clin Oncol, 2017. **35**(20): p. 2240-2250.
45. Maas, P, Barrdahl M, Joshi AD, Auer PL, Gaudet MM, Milne RL, Schumacher FR, Anderson WF, Check D, Chattopadhyay S, et al., *Breast Cancer Risk From Modifiable and Nonmodifiable Risk Factors Among White Women in the United States.* JAMA Oncol, 2016. **2**(10): p. 1295-1302.
46. Pashayan, N, Morris S, Gilbert FJ, and Pharoah PDP, *Cost-effectiveness and Benefit-to-Harm Ratio of Risk-Stratified Screening for Breast Cancer: A Life-Table Model.* JAMA Oncol, 2018. **4**(11): p. 1504-1510.
47. Garcia-Closas, M, Gunsoy NB, and Chatterjee N, *Combined associations of genetic and environmental risk factors: implications for prevention of breast cancer.* J Natl Cancer Inst, 2014. **106**(11).
48. Wilkinson, MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, et al., *The FAIR Guiding Principles for scientific data management and stewardship.* Scientific Data, 2016. **3**: p. 160018.
49. Bhattacharjee, S, Rajaraman P, Jacobs KB, Wheeler WA, Melin BS, Hartge P, GliomaScan C, Yeager M, Chung CC, Chanock SJ, et al., *A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits.* Am J Hum Genet, 2012. **90**(5): p. 821-35.
50. Zhang, H, Zhao N, Ahearn TU, Wheeler WA, Garcia-Closas M, and Chatterjee N, *A mixed-model approach for powerful testing of genetic associations with cancer risk incorporating tumor characteristics.* bioRxiv, 2018: p. 446039.
51. Yu, K, Zhang H, Wheeler W, Horne HN, Chen J, and Figueroa JD, *A robust association test for detecting genetic variants with heterogeneous effects.* Biostatistics, 2015. **16**(1): p. 5-16.
52. Barnes, DR, Lee A, Investigators E, kConFab I, Easton DF, and Antoniou AC, *Evaluation of association methods for analysing modifiers of disease risk in carriers of high-risk mutations.* Genet Epidemiol, 2012. **36**(3): p. 274-91.
53. Antoniou, AC, Wang X, Fredericksen ZS, McGuffog L, Tarrell R, Sinilnikova OM, Healey S, Morrison J, Kartsonaki C, Lesnick T, et al., *A locus on 19p13 modifies risk of breast cancer in BRCA1 mutation carriers and is associated with hormone receptor-negative breast cancer in the general population.* Nat Genet, 2010. **42**(10): p. 885-92.
54. Willer, CJ, Li Y, and Abecasis GR, *METAL: fast and efficient meta-analysis of genomewide association scans.* Bioinformatics, 2010. **26**(17): p. 2190-1.
55. Gusev, A, Ko A, Shi H, Bhatia G, Chung W, Penninx BW, Jansen R, de Geus EJ, Boomsma DI, Wright FA, et al., *Integrative approaches for large-scale transcriptome-wide association studies.* Nat Genet, 2016. **48**(3): p. 245-52.
56. Finucane, HK, Reshef YA, Anttila V, Slowikowski K, Gusev A, Byrnes A, Gazal S, Loh PR, Lareau C, Shoresh N, et al., *Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types.* Nat Genet, 2018. **50**(4): p. 621-629.
57. Gamazon, ER, Segre AV, van de Bunt M, Wen X, Xi HS, Hormozdiari F, Ongen H, Konkashbaev A, Derks EM, Aguet F, et al., *Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation.* Nat Genet, 2018. **50**(7): p. 956-967.
58. Ferreira, MA, Gamazon ER, Al-Ejeh F, Aittomaki K, Andrulis IL, Anton-Culver H, Arason A, Arndt V, Aronson KJ, Arun BK, et al., *Genome-wide association and transcriptome studies identify target genes and risk loci for breast cancer.* Nat Commun, 2019. **10**(1): p. 1741.
59. Finucane, HK, Bulik-Sullivan B, Gusev A, Trynka G, Reshef Y, Loh PR, Anttila V, Xu H, Zang C, Farh K, et al., *Partitioning heritability by functional annotation using genome-wide association summary statistics.* Nat Genet, 2015. **47**(11): p. 1228-35.

60. Hormozdiari, F, van de Bunt M, Segre AV, Li X, Joo JWJ, Bilow M, Sul JH, Sankararaman S, Pasaniuc B, and Eskin E, *Colocalization of GWAS and eQTL Signals Detects Target Genes.* Am J Hum Genet, 2016. **99**(6): p. 1245-1260.
61. Yang, J, Zeng J, Goddard ME, Wray NR, and Visscher PM, *Concepts, estimation and interpretation of SNP-based heritability.* Nat Genet, 2017. **49**(9): p. 1304-1310.
62. Zhang, Y, Qi G, Park JH, and Chatterjee N, *Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits.* Nat Genet, 2018. **50**(9): p. 1318-1326.
63. Pal Choudhury, P, Maas P, Wilcox A, Wheeler W, Brook M, Check D, Garcia-Closas M, and Chatterjee N, *iCARE: An R Package to Build, Validate and Apply Absolute Risk Models.* bioRxiv, 2018: p. 079954.
64. Choudhury, PP, Wilcox AN, Brook MN, Zhang Y, Ahearn T, Orr N, Coulson P, Schoemaker MJ, Jones ME, Gail MH, et al., *Comparative validation of breast cancer risk prediction models and projections for future risk stratification.* J Natl Cancer Inst, 2019.