B-CAST

D4.3 - Description of genotyping database

The Breast Cancer Association Consortium (BCAC) has collated genotype data on more than 260,000 breast cancer cases and controls. These data come predominantly from two large array-based genotyping projects – iCOGS and OncoArray, which provide genotypes on polymorphic genetic variants across the genome. The iCOGS and OncoArray comprised 211,000 and 533,631 variants, respectively. Data on a smaller number of variants (388) are also available in the BCAC database. These were genotyped using smaller scale platforms (principally TaqMan, Fluidigm and iPlex).

Genotype Calling

iCOGS and OncoArray genotypes were called using Illumina's Gentrain software, as discussed elsewhere[1,2]

A standard series of QC filters have been applied to the iCOGS and OncoArray datasets. These include exclusions of probable duplicates and close relatives within each study, and probable duplicates among studies, samples with a call rate <95% and samples with extreme heterozygosity (4.89 SD from the mean for the ethnicity). Ancestry was computed for each subject using a principal component analysis: subjects were thus classified as European, East Asian, African, or other/mixed.

The SNP QC involved exclusion of subjects with low call rate (<95%), SNPs not in Hardy-Weinberg equilibrium ($P<10^{-7}$ in controls or $P<10^{-12}$ in cases) and SNPs with concordance <98% among duplicate sample pairs. SNPs where the intensity cluster plot was judged not to be ideal have also been excluded.

Genotype cluster plots have been generated for all variants on both iCOGS and OncoArray. These cluster plots are available to participating groups through the BCAC website and permit manual checking of genotyping quality.

After quality control, the iCOGS and OncoArray datasets include 199,961 and 494,763 SNPs, respectively.

Imputation

The iCOGS and OncoArray genotypes were used to estimate the genotypes at more than 21M variants using imputation, with the 1000 Genomes Project (version 3) as a reference. Imputation was attempted for all variants with a minor frequency >0.1% in either Europeans or East Asians. Imputation was carried out in a two-stage procedure, using SHAPEIT for phasing and IMPUTEv2 for

imputation. The imputation was performed in 5Mb non-overlapping intervals. Genotypes for SNPs with an imputation quality score >0.3 have been retained for analysis.

Data Storage

Genotype calls and intensities, and imputed genotypes, for the iCOGS and OncoArray projects are stored in the BCAC data repository in Cambridge in *netCDF* files (http://www.unidata.ucar.edu/software/netcdf/). This format allows efficient storage and large matrices and efficient retrieval of the data on subsets of subjects and variants. Scripts have been written to extract the data as required for analyses; these datasets are then provided to analysts via secure web download.

Final Dataset

The genotype dataset includes data on 261,614 subjects, summarised below:

Ethnicity	iCOGS*		OncoArray*	
	Controls	Cases	Controls	Cases
European	45,692	56,522	58,377	81,414
Asian	6,624	6,268	6,892	8,452
African	931	1,116	2,068	3,716
Other	-	-	1,218	1,342

*19,018 subjects were genotyped on both platforms.

1. Michailidou K, Hall P, Gonzalez-Neira A, et al: Large-scale genotyping identifies 41 new loci associated with breast cancer risk. Nat Genet 45:353-361, 2013

2. Amos CI, Dennis J, Wang Z, et al: The OncoArray Consortium: A Network for Understanding the Genetic Architecture of Common Cancers. Cancer Epidemiol Biomarkers Prev 26:126-135, 2017