Common variation and heritability estimates for breast, ovarian and prostate cancers

The studies from COGS included in this collection nearly double the number of common genetic variants significantly associated with susceptibility to breast, ovarian and prostate cancers. While this set of cancer-associated variants contributes substantially to disease heritability, these studies also predict that an additional large number of common variants not yet associated with disease have the potential to explain the majority of the familial recurrence of these cancers.

Orli Bahcall

The COGS (Collaborative Oncological Gene-environment Study) research included in a *Nature Genetics* collection provides substantial insights into the contribution of common genetic variation to susceptibility to three common hormone-related cancers: breast, prostate and ovarian. The application of the specifically designed Illumina iCOGS array for genotyping of large numbers of individuals in association studies for each of these cancers has resulted in the collective identification of 74 new susceptibility loci for breast, ovarian or prostate cancer, nearly doubling the number of known susceptibility loci. These studies also calculate the magnitude of the overall contribution of common variants in susceptibility to these cancers.

Summary of common variation identified

Turning of COGS moves forward findings for hormonally mediated cancers

Sakoda, L.C., Jorgenson, E. & Witte, J.S.

doi:10.1038/ng.2587

The findings from ten of the coordinated COGS papers are presented in **Table 1**. Of the six published in this issue of *Nature Genetics*, four involved GWAS meta-analysis and validation using iCOGS data[1–4], and one entailed fine mapping and functional analysis of the *TERT* region (5p15.33) in relation to mean telomere length and the risk of breast and ovarian cancers[5]. These studies detected new loci associated with overall risk of breast cancer ($n = 41$)[1], ovarian cancer ($n = 2$)[3] and prostate cancer ($n = 23$)[4]. A few additional loci were identified specifically for risk of estrogen receptor (ER)-negative breast cancer ($n = 4$)[2] and for serous ovarian cancer ($n = 1$)[3]. In the *TERT* region, several distinct SNP associations for breast cancer, ovarian cancer and telomere length were evident, underscoring the complex interplay of common variants across this genomic region in carcinogenesis and telomere maintenance[5].

Table 1 COGS overview—study design, characteristics and published results

View table (PDF)

Full text

*Breast cancer***.** There have been 27 previously published loci associated with breast cancer ($P < 5 \times 10^{-8}$). All but four of these showed clear evidence of association with overall breast cancer risk in the COGS data. Common variants at 41 new loci were associated at genome-wide statistical significance ($P < 5 \times 10^{-8}$) with overall breast cancer risk (*Nat. Genet.* doi:10.1038/ng.2563, 27 March 2013), and SNPs at a further 4 loci were associated

Table 1 | Summary of the number of susceptibility loci identified for each phenotype and the proportion of the familial relative risk explained

Cancer[a]	New loci[b]	Total loci[c]	FRR[d]	FRR(rarer alleles)[e]
Breast cancer[f]	49	76	15%	21%
Prostate cancer	26	78	31%	5%
Ovarian cancer	8	12	4%	36%
ER-negative breast cancer[g]	4	11[i]	9%	
Breast cancer in *BRCA1* mutation carriers[h]	5	10	6%	–
Breast cancer in *BRCA2* mutation carriers[h]	3	15	7%	–
Ovarian cancer in *BRCA1* mutation carriers[h]	2	7	6%	

[a]Cancer or cancer subtype (ER-negative breast cancer or cancer in *BRCA1* or *BRCA2* mutation carriers). [b]Total number of new susceptibility loci identified for each cancer type across the combined current set of COGS publications. Includes four additional independent loci for breast cancer and three for prostate cancer identified through fine mapping of the 11q13 and 5p15 (*TERT*) regions, respectively. [c]Total number of loci now known, including those established in previous publications at genome-wide significance (see h). [d]Percentage of familial relative risk (FRR) due to all known loci in column c. See Heritability estimates explained. [e]Percentage of familial relative risk due to high- or moderate-penetrance alleles. For genes included, see Figure 1. [f]For breast cancer, the estimated contribution to FRR of all SNPs selected for the iCOGS array on the basis of evidence of association in a meta-analysis of nine breast cancer GWAS in women of European ancestry was 28%. [g]Assuming FRR for ER-negative breast cancer is also 2. [h]Includes breast or ovarian cancer susceptibility loci in the general population shown to modify cancer risk in *BRCA1* or *BRCA2* mutation carriers (at $P < 0.05$) (*PLoS Genet.* **9**, e1003212, 2013 and *PLoS Genet.* **9**, e1003173, 2013). [i]At $P < 5 \times 10^{-8}$. Of the known breast cancer susceptibility loci, 43 show evidence of association for ER-negative disease at $P < 0.05$ (*Nat. Genet.* doi:10.1038/ng.2561, 27 March 2013).

specifically with ER-negative breast cancer (Table 1 in *Nat. Genet.* doi:10.1038/ng.2561, 27 March 2013). The variant at one of the ER negative–specific loci was also associated with breast cancer risk in *BRCA1* mutation carriers at $P < 5 \times 10^{-8}$ (*PLoS Genet.* **9**, e1003212, 2013), and a further independent locus was associated specifically with breast cancer risk in *BRCA2* mutation carriers (*PLoS Genet.* **9**, e1003173, 2013). Fine-mapping studies identified three independent loci in the previously reported region at 11q13 (*Am. J. Hum. Genet.* doi:10.1016/j.ajhg.2013.01.002, 27 March 2013) and two additional independent breast cancer susceptibility loci in a region at 5p15 (Table 2 in *Nat. Genet.* doi:10.1038/ng.2566, 27 March 2013). These findings bring the total number of breast cancer susceptibility loci to 76. Of the 27 previously established breast cancer loci, 26 were included on the iCOGS array (rs2284378 at 20q11 was not selected), and consistent evidence of association with overall or ER-negative breast cancer risk ($P < 1.0 \times 10^{-4}$) was observed for all but 2 of these (*Nat. Genet.* doi:10.1038/ng.2563, 27 March 2013 and *Nat. Genet.* doi:10.1038/ng.2561, 27 March 2013). Weaker evidence for association was found for rs1045485 in *CASP8* and rs2380205 at 10p15.

*Ovarian cancer***.** Four loci have previously been reported to be associated with ovarian cancer at genome-wide significance. These were all confirmed by the COGS data, as were two other loci previously reported close to genome-wide significance. Three new ovarian cancer susceptibility loci were identified at genome-wide significance (Table 2 in *Nat. Genet.* doi:10.1038/ng.2564, 27 March 2013); two were associated with overall risk, and one was associated specifically with risk of the serous subtype.

A more detailed analysis of the association at 17p12 reported by Pharoah *et al.* is reported in Shen e*t al.* (Table 1 of *Nat. Commun*. doi:10.1038/ncomms2629, 27 March 2013). Of particular interest was the finding that different loci in the same region were associated with the serous and clear-cell subtypes. Variants at the 17q21.31 locus were associated with ovarian cancer risk in *BRCA1* mutation carriers (*PLoS Genet.* **9**, e1003212, 2013) and also in *BRCA2* mutation carriers (*PLoS Genet.* **9**, e1003173, 2013). Permuth-Wey *et al.* also reported an association between this locus and invasive serous epithelial ovarian cancer (EOC) risk in the general population (Table 1 in *Nat. Commun.* doi:10.1038/ncomms2613, 27 March 2013), and Bojesen *et al.* identified two new loci in the 5p15 region, one associated with overall ovarian cancer risk and the other with risk of serous low-malignant-potential (LMP) disease (Table 2 in *Nat. Genet.* doi:10.1038/ng.2566, 27 March 2013). Finally, an additional locus was associated with ovarian cancer risk for *BRCA1* mutation carriers only (*PLoS Genet.* **9**, e1003212, 2013). These findings bring the total number of ovarian cancer susceptibility loci to 12.

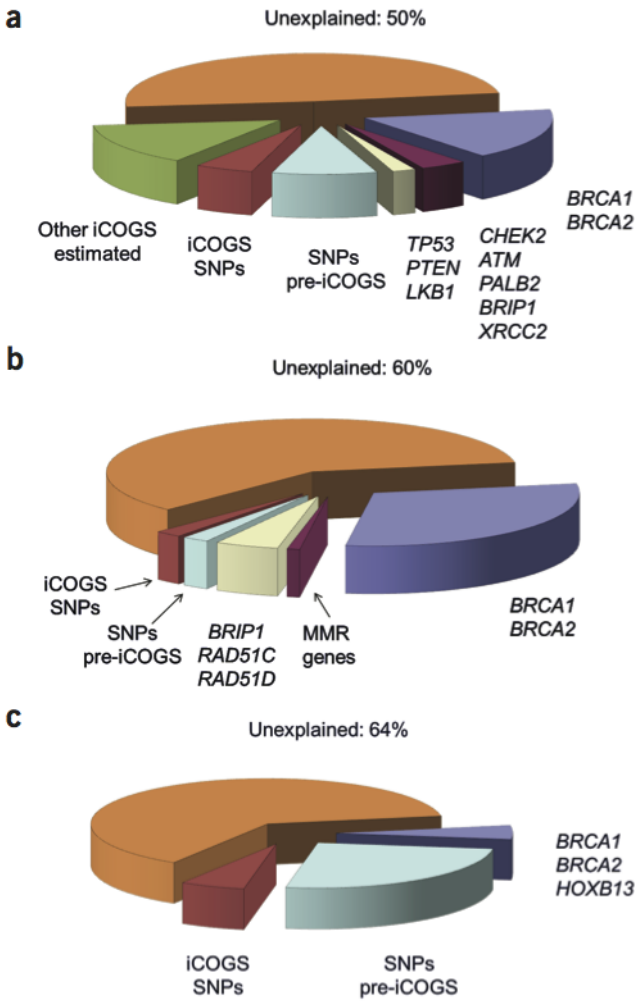

Figure 1 | Estimated proportions of the FRR attributable to different types of genetic variation. (a -c) FRR estimates are shown for breast (a), ovarian (b) and prostate (c) cancers. The listed genes refer to genes with known disease-causing mutations (broadly categorized as those conferring high risk (for example, BRCA1 and BRCA2), other genes conferring high risk (for example, TP53) and genes conferring moderate risk (for example, PALB2). Pre-iCOGS SNPs refers to the contribution of all published SNPs associated with this cancer before the current COGS publications; iCOGS SNPs refers to the contribution of SNPs established in the current COGS publications (these totals are noted in Table 1). For breast cancer only, other iCOGS estimated refers to the estimated contribution of all other SNPs on the iCOGS array that were selected for replication of GWAS2.

*Prostate cancer***.** There have been 53 previously published loci associated with prostate cancer ($P < 5 \times 10^{-8}$). Twenty-three new prostate cancer susceptibility loci were identified at genome-wide statistical significance with overall prostate cancer risk (Table 1 in *Nat. Genet.* doi:10.1038/ng.2560, 27 March 2013). Thirteen SNPs showed clear association when analysis was restricted to aggressive disease (significant at $P < 0.01$), and, for 22 of the 23 SNPs, the estimated odds ratio (OR) was in the same direction for aggressive and non-aggressive disease. Aggressive disease was defined as that having Gleason score ≥ 8, prostate serum antigen (PSA) > 100 ng/ml, disease stage of distant (outside the pelvis) or death from prostate cancer. Fine mapping of a previously reported locus at 5p15 identified multiple independently associated loci in the *TERT* region (*Hum. Mol. Genet.* doi:10.1093/hmg/ddt086, 27 March 2013). This brings the total number of known susceptibility loci for prostate cancer to 78.

Fine-mapping efforts in COGS

One of the major aims of the COGS project was the fine mapping of previously identified susceptibility loci for breast, ovarian and prostate cancers by genotyping a very dense panel of markers, drawn from the 1000 Genomes Project dataset across these regions. The COGS project included the fine mapping of over 50 selected genomic regions to identify variants more strongly associated with the disease than those reported in the original GWAS, as well as additional associated variants that may contribute to heritability. The fine mapping of two regions, at 5p15 and 11q1, has been reported in detail in these publications. These analyses suggest that the fine mapping of susceptibility regions can be productive in identifying further susceptibility variants that contribute to overall genetic risks for these cancers.

Multiple independent variants at the *TERT* locus are associated with telomere length and risks of breast and ovarian cancer

Bojesen, S.E., Pooley, K.A., Johnatty, S.E., Beesley, J., Michailidou, K. *et al.*

doi:10.1038/ng.2566

Resulting from a common interest, members of each of the constituent consortia in the Collaborative Oncological Gene-environment Study (COGS) nominated SNPs surrounding the *TERT* locus for inclusion on a genotyping array. Consequently, the iCOGS array design included a combination of individual *TERT* gene candidate SNPs, as well as a more comprehensive set to fine-scale map the entire locus, for shared use by all consortia. This study had three aims: to assess SNPs across the *TERT* locus for all detectable associations with mean telomere length and breast and ovarian cancer subtypes; to fine-scale map this locus to identify potentially causal variants for the observed associations; and to evaluate the functional effects of the strongest candidate causative variants.

Our comprehensive examination of the *TERT* locus has answered some long-standing questions and raised several new ones. We have identified two independent regions associated with telomere length in leukocyte DNA; these provide definitive evidence for genetic control of telomere length by common *TERT* variants. For rs2736108, the most significant SNP in promoter peak 1, the minor allele is associated with a 1.7% increase in telomere length. This is equal to a telomere length change of ~60 bp, which, because telomere length decreases by approximately 19 bp per year[50], is equivalent in magnitude to an age difference of 3.1 years. We estimate that rs2736108 explains 0.08% of the variance in telomere length in men and 0.06% of the variation in women. SNPs in peak 2 have a stronger effect on telomere length, with each additional A (minor) allele of rs7705526 associated with a 2.6% increase. This is equal to a ~90 bp change in telomere length and, correspondingly, to 4.7 years of age. We estimate that rs7705526 explains 0.31% of the variance in telomere length in men and 0.16% of the variance in women. The only other reported associations with telomere length reaching genome-wide significance involve *TERC*-locus SNP rs1269304 (ref. 51) and *OBFC1*-locus SNP rs4387287 (ref. 52), which have similar effects on telomere length (75 bp and 115 bp per allele, respectively).

Fine-mapping identifies multiple prostate cancer risk loci at 5p15, one of which associates with *TERT* expression

Kote-Jarai, Z. *et al.*

doi:10.1093/hmg/ddt086

Zsofia Kote-Jarai and colleagues report fine mapping of associations to prostate cancer susceptibility at the *TERT* locus, using high-resolution genotyping and imputation (*Hum. Mol. Genet.* doi:10.1093/hmg/ddt086, 27 March 2013). The authors include genotyping of 134 SNPs across the *TERT* locus using the custom Illumina iSelect array (iCOGS) or Sequenom MassArray iPlex, in 22,301 cases and 22,320 matched controls from 23 studies included in the PRACTICAL Consortium. They initially genotyped 114 SNPs across 135 kb of the

SLC6A18–TERT–CLPTM1L region and then narrowed the associated region to a 20-kb interval that included variants with stronger association. They further imputed all iCOGS genotyped samples for variants in 1000 Genome Project Phase 1 data and tested association for the imputed set of 1,094 SNPs. They identified 44 SNPs associated with prostate cancer risk at $P < 1 \times 10^{-5}$. Using stepwise logistic regression analyses, they were able to identify four SNPs showing independent association, suggesting four separate regions influencing susceptibility to prostate cancer.

Research Highlight in *Nature Genetics*, doi:10.1038/ng.2597

Functional variants at the 11q13 breast cancer risk loci regulate cyclin D1 expression through long-range enhancers
French, J.D. *et al.*

doi:10.1016/j.ajhg.2013.01.002

Juliet French and colleagues report fine mapping of the 11q13 breast cancer susceptibility locus (*Am. J. Hum. Genet.* doi:10.1016/j.ajhg.2013.01.002, 27 March 2013). The original genome-wide association study (GWAS)-identified SNP tags a linkage disequilibrium block that spans 683 kb, and the authors selected 731 SNPs from this region to include on the iCOGS array. They genotyped 89,050 individuals of European ancestry and 12,893 individuals of Asian ancestry, all from studies included in BCAC. They identified 204 SNPs associated with overall breast cancer risk, finding that these were all associated with estrogen receptor (ER)-positive but not ER-negative breast cancer. Using stepwise logistic regression, they identified three independently associated SNPs.

Research Highlight in *Nature Genetics*, doi:10.1038/ng.2596

Heritability estimates in COGS studies

Additional heritability

As shown in **Table 1**, the estimated proportion of the familial risk attributable to the total currently known set of susceptibility loci for these three cancers (including previous associations and those identified in the current COGS publications), ranges from ~4% for ovarian cancer to ~31% for prostate cancer. However, there are many more SNPs that show nominally significant associations to each cancer but that do not reach the established genome-wide significance threshold for association in these publications.

This is shown most clearly for breast cancer, with Michailidou *et al.* finding that there is a clear excess of significant associations among all SNPs genotyped on the iCOGS array that were selected from GWAS, even at levels of significance below the conventional genome-wide significance threshold. Further support for the existence of a much larger number of breast cancer susceptibility loci is provided by the finding that the associations in the iCOGS replication data sets tend to go in the same direction as those in the GWAS from which the SNPs were selected, even for SNPs with associations below the genome-wide significance threshold.

Large-scale genotyping identifies 41 new loci associated with breast cancer risk
Michailidou, K., Hall, P., Gonzalez-Neira, A., Ghoussaini, M., Dennis, J. *et al.*

doi:10.1038/ng.2563

However, the overall excess of significant associations for SNPs selected from the breast cancer GWAS for genotyping in the iCOGS stage suggests that a much larger number of loci contribute to susceptibility, although they did not have associations reaching genome-wide levels of significance in the current study.

To assess this hypothesis more formally, we identified a set of 10,668 SNPs selected from the GWAS that were uncorrelated ($r^2 < 0.1$ between any pair). Of these, the estimated OR was in the same direction as in the combined GWAS for 5,918 SNPs and in the opposite direction for 4,750 SNPs. Assuming that SNPs with effects in opposite directions are not associated with risk, an estimated 1,168 loci selected from the GWAS are associated with risk. However, this is an underestimate because weakly associated SNPs might have effects in opposite directions in the two stages.

A similar finding was seen for prostate cancer, in comparing the direction of association for SNPs on the iCOGS array and those for the same SNPs in the GWAS dataset. Once again, the direction of effect in the iCOGS replication data set was in the same direction for over half of the selected SNPs, suggesting that there may be a much larger number of prostate cancer susceptibility loci.

Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array
Eeles, R.A., Al Olama, A.A., Benlloch, S., Saunders, E.J., Leongamornlert, D.A. *et al.*

doi:10.1038/ng.2560

The overall inflation in the test statistics for those SNPs selected for GWAS replication suggests that the number of susceptibility loci may be much larger. To address this possibility more formally, we identified 22,662 SNPs selected for replication of the prostate cancer GWAS that were uncorrelated ($r^2 < 0.1$ for any pair) and examined the directions of the estimated ORs in the iCOGS replication data set. The estimated effects were in the same direction as in the GWAS for 12,278 SNPs and in the opposite direction for 10,384 SNPs. On the basis of this analysis, 1,894 (95% CI = 1,600–2,188) selected SNPs reflect true associations with disease.

Large-scale genotyping identifies 41 new loci associated with breast cancer risk
Michailidou, K., Hall, P., Gonzalez-Neira, A., Ghoussaini, M., Dennis, J. *et al.*

doi:10.1038/ng.2563

As an alternative approach, we fitted the distribution of *z* scores for the iCOGS stage, aligned to the direction of the effect in the GWAS, as a mixture of two normal distributions representing those SNPs that were or were not associated with disease (**Fig. 2** and Online Methods)[58]. On the basis of the posterior probabilities from this analysis, an estimated 92% of loci ($n = 9{,}815$) were associated with breast cancer risk (95% CI = 85–100%), and these contributed approximately 18% of the familial risk of breast cancer. It should be noted, however, that the large majority of the loci had very small individual effects on risk: for example, the estimated OR was >1.05 for only 10 loci, and 920 loci had an estimated OR of >1.02. When taking into account effects from the previously known loci, these analyses suggest that ~28% of familial risk is explained by common variants selected for iCOGS, of which ~14% can be explained by the 67 established loci (with a further ~20% due to higher penetrance loci).

What we have learned

The current collection of COGS publications together demonstrate that (i) the contribution of common SNPs to the heritability of breast, ovarian and prostate cancers is substantial, (ii) the number of SNPs associated with each disease is very large, at least several thousand, and (iii) the contribution of the SNPs not yet definitively

associated with disease is probably much greater than those that have been identified so far and may explain the majority of the familial aggregation of these diseases. The lower contribution of common SNPs to familial risk identified for ovarian cancer may be a reflection of the smaller sample size available, as the effect sizes for the known SNPs are comparable to those for breast cancer.

The above analyses are based on an assumption that the disease-associated variants combine multiplicatively. If there are interactions between loci or between genetic loci and environmental or lifestyle risk factors, the contribution of common variants could be greater. In addition, rare variants that are not captured on the iCOGS array may confer higher risks and explain additional heritability. Such rare susceptibility variants have been identified previously for all these cancers (**Fig. 1**). These variants are located in genes such as *BRCA1* and *BRCA2* that are associated with high risk, identified through linkage analysis and positional cloning, in genes that confer more moderate risk such as *PALB2*, identified through sequencing of candidate genes in case-control studies. The overall contribution of rare variants to cancer susceptibility remains an open question that will only be resolved by large-scale sequencing experiments.

It is notable that (*Lancet* 358, 1389–1399, 2001 and *Am. J. Med. Genet. C. Semin. Med. Genet.* 129C, 65–73, 2004) for breast and prostate cancers, observed familial relative risks in epidemiological studies decrease markedly with increasing age[14,15]. In contrast, the relative risks conferred by the SNPs and, hence, also the overall familial relative risk explicable by the SNPs are not strongly related to age (*Nat. Genet.* doi:10.1038/ng.2563, 27 March 2013 and *Nat. Genet.* doi:10.1038/ng.2560, 27 March 2013). These results suggest that the contribution of common SNPs to the heritability of these cancers, although substantial overall, is smaller at younger ages and, hence, that the search for rarer disease-causing variants should focus on younger cases.

Heritability estimates explained

The heritability of a trait is defined as the proportion of the phenotypic variance that can be attributed to genotype. For a disease trait, the analysis is usually conducted using an underlying continuous liability, whereby individuals are assumed to be affected if they exceed a certain liability threshold. The heritability then refers to the heritability on the liability scale, rather than the heritability of the observed trait values (*Nat. Rev. Genet.* **9**, 255–266, 2008).

Often a more direct and useful measure, however, is the proportion of the observed FRR (denoted λ) that can be attributable to SNPs or other genetic variants.

For a locus with a log-additive association with risk, the FRR to the offspring of a case is given by $\lambda_k = (p_k r_k^2 + q_k)/(p_k r_k + q_k)^2$ where p_k is the frequency of the risk allele, $q_k = 1 - p_k$ and r_k is the per-allele relative risk.

Assuming that the loci combine multiplicatively and are not in linkage disequilibrium, the combined effect of all loci is given by $\lambda_T = \prod_k \lambda_k$ where the product is across all loci. The proportion of the familial relative risk attributable to the SNPs, on a log scale, is then given by $\log(\lambda_T)/\log(\lambda_P)$, where λ_P is the familial relative risk observed in epidemiological studies. λ_P is 2–3-fold for breast, ovarian and prostate cancers.

Another way of expressing this is that, under a simple polygenic model, the observed FRR $\lambda_P = \exp(\sigma_P^2/2)$, where σ_P^2 is the variance of the underlying polygenic component. The proportion of the familial risk explained is then given by $2\sum_k \log(\lambda_k)/\sigma_P^2 = \sigma_T^2/\sigma_P^2 = \sum_k \sigma_k^2/\sigma_P^2$ where $\sigma_k^2 = 2\log(\lambda_k) \approx 2p_k q_k \beta_k^2$ is the proportion of the polygenic variance explained by SNP k and $b_k = \log(l_k)$ is the per-allele log(relative risk).