Genome-wide association studies of cancer: current insights and future perspectives

Amit Sud[1], Ben Kinnersley[1], Richard S Houlston[1,2]

1. Division of Genetics and Epidemiology, The Institute of Cancer Research, Sutton, London. UK

2. Division of Molecular Pathology, The Institute of Cancer Research, Sutton, London. UK

Correspondence to:

Richard S Houlston; Tel ++44 (0) 208 722 4175; E-mail: richard.houlston@icr.ac.uk

Key Points

- The architecture of inherited genetic susceptibility to cancer is defined by a spectrum of predisposition alleles which have differing frequency and impact.
- Genome-wide association studies (GWAS) provide an agnostic approach to identify genetic variation influencing cancer risk. GWAS of most cancers have been performed and hundreds of risk alleles have been identified, most of which are common and individually confer a modest increase in risk.
- Most cancer risk loci identified through GWAS locate to non-coding regions of the genome and influence gene expression through diverse mechanisms.
- As well as improving our understanding of cancer, information from GWAS has direct clinical relevance in identifying non-genetic aetiological risk factors, optimising population screening, identifying therapeutic targets, drug repositioning and prognostication.
- Although challenging, deciphering the biological basis of associations is necessary to fully realise the potential of GWAS.

ABSTRACT

Genome-wide association studies (GWAS) provide an agnostic approach for investigating the genetic basis of complex diseases. In oncology, GWAS of nearly all common malignancies have been performed and over 700 genetic variants associated with increased risks identified. As well as revealing novel pathways important in carcinogenesis, these studies have shown that common genetic variation contributes significantly to the heritable risk of many common cancers. The clinical application of GWAS is starting to provide opportunities for drug discovery and repositioning, as well as cancer prevention. Deciphering the functional and biological basis of associations is, however challenging and is in part a barrier to fully unlock the potential of GWAS.

[H1] Introduction

Epidemiological studies provide strong support for a hereditary component to the aetiology of common cancers[1]. Many cancers show a higher concordance in monozygotic twins as compared with dizygotic twins or siblings[2]. While this concordance is compatible with inherited genetic variation rather than lifestyle or environmental risk factors, it does not exclude non-genetic mechanisms as a basis of apparent heritability **[G].** For example, the high concordance of acute leukaemia in monozygotic twins has an *in utero* explanation[3]. The pattern of relative risk (RR) **[G]** for most common cancers is that familial RRs are greatest in relatives of early-onset cancer patients, which is compatible with tumours developing in these genetically susceptible individuals at an earlier age[4]. For most common cancers, risks in first-degree relatives of patients are increased two- to three-fold for the same cancer. Notable exceptions are chronic lymphocytic leukaemia, and thyroid and testicular cancers, for which risks are increased four- to eight-fold[1]. The genetic architecture underscoring these familial risks is now known to reflect a range of alleles with varying frequencies and impact.

More than 40 years ago, Anderson stated that the two- to three-fold excess risks of cancer seen in first-degree relatives of cancer patients, "are not indicative of a strong genetic effect. They are more suggestive of a polygenic mechanism, *i.e.* the involvement of many genes with small effects acting in concert with environmental or non-genetic factors with larger and more important effects"[5]. This conclusion is incorrect, as the RRs in relatives of patients compared with the population will usually be more than one order of magnitude lower than the RRs in susceptible compared with non-susceptible individuals. The observed RRs are diluted by three factors. Firstly, not all cancer patients are susceptible. Secondly, even fewer of the relatives are susceptible. Thirdly, the general population is composed of both susceptible and non-susceptible individuals. However, such modest excess familial risks are entirely compatible with Mendelian predisposition **[G]**, provided that the genetic effect is substantial[4]. Indeed genetic linkage and positional cloning studies performed in the 1980s and 1990s led to the identification of many high penetrance cancer susceptibility genes (CSGs) **[G]**, for example those for breast and ovarian cancers (*BRCA1* and *BRCA2*)[6-8], colorectal cancer (CRC; *APC* and the mismatch repair (MMR) genes *MLH1* and *MSH2*[9-13]) and melanoma *CDKN2A*)[14,15] within certain families.

To date mutations in more than 70 CSGs associated with high-penetrance **[G]** cancer susceptibility syndromes have been identified, which confer RRs of 5-100[16] (**Fig. 1**). However, these syndromes

only account for a small fraction of the familial risks of the respective cancers, leaving much of the heritability unexplained. For example, high-penetrance mutations are responsible for most breast cancer and CRC patients in families with more than three patients (*i.e.* indicative of Mendelian inheritance), but are responsible for only a minority of those with two patients[17-19]. Mutations in known predisposition genes, including *BRCA1* and *BRCA2*, account for less than 25% of the two-fold excess risk in the relatives of patients with breast cancer[18,20]. Similarly, more than 60% of the excess familial risk of CRC remains unaccounted for[19,21].

Over the past 20 years, extensive efforts to discover additional, high-penetrance CSGs for breast cancer and CRC have been made but no gene with a similar profile to *BRCA1*, *BRCA2* or the MMR genes has been identified. If additional CSGs exist, as is the case in CRC with *POLE*[22] and *NTHL1*[23] variants, each will account for only a small proportion of the familial risk (*i.e*. <1%)[19]. These data, coupled with the previously described high estimates of cancer heritability from twin studies, suggest that much of the missing heritability will be polygenic. Here, the co-inheritance of genetic variants, each of which has a modest individual effect, can cause a wide range of risk in the population.

Paradoxically therefore, while the reasoning behind Anderson's statement was incorrect, it is now recognized that much of the genetic architecture of cancer susceptibility is in fact explained by polygenic inheritance. Thus, a high proportion of cancers may arise in a genetically susceptible minority of individuals in a population - a consequence of the combined effects of common low-penetrance alleles and rare disease-causing variants that confer moderate cancer risks. In appreciation of this, the past decade has seen a shift in gene discovery efforts from models of predisposition based on high-penetrance single-gene mutations (*i.e*. causative of cancer syndromes) to multi-genic models. This Review focuses on the major findings from association studies, in particular genome-wide association studies (GWAS), both in terms of understanding the allelic architecture of cancer susceptibility and its functional basis as well as ongoing challenges and future perspectives.

[H1] The advent of GWAS

Association studies have detected two main classes of cancer susceptibility variants with different levels of risk and prevalence in the population. Firstly, rare moderate-penetrance variants (risk allele frequency **[G]** <2%; odds ratios (ORs) **[G]** >2.0) have been identified through the direct

interrogation of candidate genes. For example, other genes encoding proteins involved in the DNA-damage response pathway, in addition to *BRCA1* and *BRCA2*, such as *ATM*[24,25], *CHEK2*[26] and *PALB2*[27,28], have been associated with an increased breast cancer risk. Based on their risk allele frequencies of 0.1% to 0.5% and the modest two-fold increase in risk associated with each, variation in these genes contribute little to the familial risk of breast cancer. Secondly, common low-penetrance alleles (risk allele frequency >5%; OR <1.5) have been identified by GWAS. It is likely however that the spectrum of penetrance and frequency of risk alleles for many cancers occurs on a continuum. This dichotomy probably reflects the subgroups of risk alleles that are most readily detected, rather than the underlying biological or evolutionary constraints.

GWAS were made possible by improved insight into common genetic variation coupled with technological developments in high-throughput genotyping. Through an agnostic genome-based approach, GWAS compares the frequency of common DNA variations in a large series of unrelated cancer patients and matched healthy individuals (referred to as 'controls' from hereon), to identify genetic variants associated with cancer risk (**Fig. 2**). GWAS of most of the common cancers have now been performed and genomic variants associated with their risks identified, providing direct evidence of polygenic susceptibility.

[H1] Study design for GWAS

From 2006 onwards, knowledge of single nucleotide polymorphisms (SNPs, the most common genetic variant) gained from the Human Genome Project[29,30] and the International Hap Map Project[31], together with technical advances in high-throughput genotyping technology made large-scale GWAS a viable option.

The underlying basis of GWAS is that adjacent stretches of DNA tend to be non-independently co-inherited. This non-random association of alleles (linkage disequilibrium (LD) **[G]**) allows certain SNPs to act as proxies, or tagSNPs, for adjacent SNPs[32]. Therefore, the number of SNPs that need to be genotyped to capture most common variants (*i.e.* minor allele frequency >5%) is reduced to around 300,000[33]. By determining which SNP alleles occur more or less frequently in patients compared with healthy individuals, genomic regions associated with risk can be identified (**Fig. 2**).

There is a general need for patients and controls to be appropriately matched. This matching is to ensure adequate statistical power, and to minimise biases or confounding factors leading to false-

positive associations. For example, methods have been developed that can correct for potential population differences between patients and controls[34]. To mitigate the issue of multiple comparisons and reduce false-positive associations, stringent statistical thresholds are necessary. The Bonferroni correction is commonly applied, whereby a *P*-value of 5.0×10^{-8} corresponds to genome-wide significance at the 5% threshold (*i.e*. 0.05/1,000,000 SNPs) [**BOX 1**]. The strength of associations, however, have to be interpreted with caution owing to the "winner's curse". Thereby, an overestimation of effect size **[G]** is likely to occur if, for example, initial discovery studies have low sample size and statistical power[35,36]. More reliable effect sizes can be estimated through validation in independent cohorts.

Historically, to offset the high cost of commercial SNP arrays but retain statistical power GWAS were generally designed based on a staged strategy. That is, promising associations from the initial genome-wide analysis were followed up by targeted genotyping of independent case-control series. The significantly reduced cost of arrays and the formation of international consortia have led to many analyses being solely based on the meta-analysis of genome-wide SNP data. While intrinsically attractive, the combination of data from different arrays can raise issues relating to varying quality of genotyping between array technologies as well as the density of SNP genotypes. Fortunately, the imputation of untyped genotypes using sequenced reference panels of individuals available through initiatives such as the 1000 Genomes Project[37], UK10K consortium[38] and haplotype reference consortium[39], has facilitated the harmonisation of data generated by different array formats. This has allowed SNP alleles with frequencies as low as 0.1% to be accurately imputed[39] extending the utility of GWAS to decipher the allelic structure of cancer susceptibility. Of note, the role of structural variations, such as copy number variations (CNVs), is largely unappreciated, because existing arrays are not ideally formatted to capture them.

[H1] Cancer risk loci identified

Over the past decade, multiple GWAS have been reported for each of the major cancers in European populations, including breast[40-42], prostate[43-45], lung[46-49], colorectal[50-56], pancreatic[57-60], gastric[61,62], renal[63-65] and bladder cancer[66,67]. For many of these, East Asian and African-American population specific risk loci have also been identified, reflecting differences in LD structure between ethnicities[68]. GWAS have also been reported for malignant melanoma[69-71], ovarian cancer[72-75], basal cell carcinoma[76-79], glioma[80-82], meningioma[83], testicular germ cell tumour (TGCT)[84-86], thyroid cancer[87] and several of the haematological malignancies including the major B-

cell tumours - acute lymphocytic leukaemia (ALL)[88-91], chronic lymphocytic leukaemia (CLL)[92-94], multiple myeloma (MM)[95-98], Hodgkin lymphoma (HL)[99-101] follicular lymphoma[102] and diffuse large B-cell lymphoma[103]. Additionally, common risk alleles have been identified through GWAS for several paediatric solid cancers including Wilms tumour[104] and neuroblastoma[105]. Each of these studies reported well-validated disease loci. Currently, more than 430 cancer associations at 262 distinct genomic regions have been identified by GWAS (**Supplementary Fig. 1, Supplementary Table 1**).

Breast and prostate cancer GWAS have so far yielded the greatest number of risk loci[41,45]. This high output is likely because of greater statistical power owing to the large sample size of the respective GWAS, each of which involved the genotyping of over 120,000 individuals[41,45]. For other cancers, differences in their heritability are likely to have influenced the performance of GWAS in identifying risk loci. For example, in CLL, which is strongly heritable and has an eight-fold familial RR[106], GWAS have led to the identification of 43 risk loci, despite it being based on the analysis of only 6,200 patients and 17,598 controls[92]. In contrast, a GWAS analysis of 29,266 patients and 56,450 controls has led to the discovery of only 18 risk loci for all lung cancer subtypes[49], reflecting the importance of non-genetic risk factors in the aetiology of this cancer.

[H1] Pleiotropy at cancer risk loci

Most SNP associations identified to date have been cancer-specific, which is consistent with the epidemiological observations of most familial cancer risks[1]. However, approximately one third of SNPs map to genomic loci associated with multiple cancers. Classically pleiotropic loci would be those where the exact same association signal (and therefore presumed molecular mechanism) encompasses multiple cancers. A broader and perhaps more pragmatic definition encompasses cancer-specific "hotspots" with a presumed shared (but less direct) molecular mechanism, allowing *e.g.* for cancer- or tissue-specific enhancer effects. This definition of pleiotropy **[G]** enables the grouping of cancers or loci that can be instructive in our understanding of cancer by highlighting shared mechanisms or hallmarks, for example, telomere-related loci[107] at 3q26.2 (*TERC*), 5p15.33 (*TERT*), 10q24.33 (*OBFC1*) and 20q13.33 (*RTEL1*) are associated with risks of multiple cancers[107]. In particular, the SNP rs2736100 at 5p15.33 (*TERT*) is associated with risk of many cancer types, including glioma[81] as well as bladder[108] and lung[109] cancer. Similarly, the locus at 9p21.3 (*CDKN2A-CDKN2B*) has been found to influence glioma[81], melanoma[69], ALL[110] and lung

cancer[111] risk, as well as naevi density[112]. For some loci the immediate cancer-specific mechanism of predisposition may not be shared, though ultimately they might converge on the same oncogenic mechanism. The SNP rs6983267 at chromosome 8q24.21 was found through scans of both prostate cancer[113] and CRC[114]. However, this locus has also been shown to harbour risk SNPs for other cancers. These SNPs localise within distinct LD blocks and likely reflect tissue-specific effects on cancer risk[64,66,75,81,93,99], through regulation of *MYC* (**Fig. 3**).

Additional insights can be gained from the cancer types themselves implicated at "pleiotropic" loci. For example, 16q24.3 harbours multiple associations for skin cancers, including melanoma[69], non-melanoma skin cancer[115] and cutaneous squamous cell carcinoma[116], which is likely indicative of a common, perhaps tissue-specific mechanism of action. Furthermore, additional multiple-cancer regions are consistent with known familial co-clusters *e.g*. 19p13.11 and breast[117] and ovarian[74] cancer. For other regions containing multiple cancer associations, the shared genomic location is likely due to chance and the molecular basis of associations completely independent. In these cases there is no additional insight to be gained from collectively considering the multiple cancer and risk associations. Exploring the nature of pleiotropic loci will likely be the focus of future work and lead to increased insight into cancer susceptibility and aetiology.

[H1] Insights into cancer biology

One of the anticipated deliverables from GWAS was that the identification of variants of genes in specific pathways would provide new insights into cancer biology. Few of the genes implicated by GWAS had previously been evaluated in targeted association studies, emphasizing that the candidate gene approach was hampered by a limited knowledge of tumour biology. Moreover, insights into new pathways of tumorigenesis for different cancer types have emerged; for example, the role of B-cell developmental and immune response genes (*e.g*. *IKZF1, CEBPE, IRF4*, *IRF8, GATA3* and *ARID5B*) as key determinants of the risk of B-cell tumours[88-92,95,97-99,118-120]. Similarly, GWAS implicated genes involved in developmental transcriptional regulation, microtubule and chromosomal assembly, and components of the KIT-MAPK signalling pathway in TGCT oncogenesis[121-124].

Given the considerable difficulties in unambiguously identifying causative exposures for many cancers, genetic associations have the potential to endorse current aetiological hypotheses, or

suggest new ones that merit testing through gene- or environment-specific hypotheses. Examples of loci that demonstrate an effect on cancer risk mediated by lifestyle or environmental exposure include a SNP at 15q25 (*CHRNA3-CHRNA5)* locus that is indirectly associated with lung cancer risk through nicotine addiction[46,125]. The genotype at this locus influences the ability to quit smoking[126] and smokers carrying two copies of the *CHRNA3-CHRNA5* risk allele smoke on average two more cigarettes per day (CPD) than those homozygous for non-risk alleles[48].

Other examples of genotype indirectly influencing cancer risk are provided by a SNP at 8p22 (*NAT2)*modifying the effect of smoking on bladder cancer[127] and the skin pigmentation loci that are associated with skin cancer[69,76]. Such data implies that heritable factors may well have a greater impact on cancer incidence than previously thought.

Recently, researchers have suggested that "replicative" errors contribute substantially to cancer aetiology alongside environmental and heritable factors, inferred from observations of a correlation between total stem cell divisions and cancer incidence in various cancer types[132, 133]. However, such assertions warrant further scrutiny, as the relative contributions of factors havebeen calculated based on the assumption of independent effects, and have solely been based on high-penetrance inherited mutations, which contribute little to the population attributable risk **[G]** of a given cancer. All in all, it is likely that all of the posited components interact, with common GWAS susceptibility alleles also playing a role.

[H1] Genetic risk in non-Europeans

GWAS have been conducted in a number of non-European populations, either for cancers common in all populations (such as prostate, breast and colorectal cancer) or for those common in specific populations (*e.g.* hepatocellular carcinoma in East Asians). Approximately 56% of GWAS risk loci show association with only European populations, and 29% of GWAS risk loci show association with multiple populations, predominantly European, East Asian and African-American (**Supplementary Table 1**). Undoubtedly, cancer GWAS have been dominated by studies of European populations, so the proportion of GWAS risk loci associated with non-European populations is likely to increase as more studies in different populations are undertaken. Intriguingly, within certain risk loci exhibiting association with multiple populations, there may be population-specific association signals. For example for prostate cancer there appear to be several susceptibility regions at 8q24.21 with differing specificities for African American, Japanese

American, Native Hawaiian, Latino and European American populations[128]. This might reflect population-specific disease mechanisms.

Additionally, founder mutations arising in small populations can inform on cancer genetic risk. Examples outside of GWAS include bi-allelic *NTHL1* mutations as a cause of recessive CRC discovered through sequencing Dutch families[23], as well as *APC* p.Ile1307Lys (rs1801155) as a basis of low-penetrance susceptibility to CRC in individuals of Ashkenazi Jewish ancestry[129].

[H1] Common variation and heritable risk

The loci identified through GWAS tend to exhibit dosage effects, with homozygous carriers of the risk allele having an excess risk approximately twice that of heterozygous carriers of the risk allele. This might partly reflect the fact that a log-additive model **[G]** has been used for the primary discovery, as even large GWAS will be underpowered to demonstrate significant deviation from this model[130].

Nearly all the cancer susceptibility loci identified to date are associated with modest increases in risk, with ORs generally less than 1.5. Exceptions to this are the SNPs at 9p21 (*JAK2*) for myeloproliferative neoplasms[131], 12q21.32 (*KITLG*) for TGCT[121,122] and 8q24.21 (*CCDC26*) for IDH-mutated glioma[132], each of which are associated with a three-fold increased risk of the respective cancer. These cancers are notable in having large familial risks but showing little evidence for Mendelian predisposition[133-135].

These GWAS data provide general insights into the allelic architecture of cancer susceptibility. Even though the cancer risks associated with these SNPs are modest, the variants are common and therefore, each of them contributes to the risk of the respective cancer type in a large proportion of the population. The number of common variants each of which could explain more than 1% of inherited risk is very low. However, as the SNPs identified by GWAS have to pass a very stringent significance threshold, there are likely multiple SNPs with weak effect sizes that do not meet these criteria but still contribute to the heritable risk of a given cancer. Quantifying the heritability explained by both known and potential susceptibility SNPs is therefore important to verify the aetiological basis of cancer and understand its genetic architecture. Calculating the proportion of phenotypic variance explained by a large number of SNPs for complex human

diseases is a significant challenge. Methods such as Genome-wide Complex Trait Analysis[136] (GCTA) **[G]** , which estimate the polygenic variance (*i.e.* heritability) ascribable to all GWAS SNPs simultaneously, have shown that common variation is likely to explain a high proportion of heritable risk of many cancers, with estimates of 10% for oestrogen receptor negative breast cancer[137], 38% for prostate cancer[137] and 17% for CRC[138]. More recent methods have attempted to improve GCTA by accounting for minor allele frequency, LD and genotype uncertainty. Such methods appear to produce higher estimates of heritability ascribed to common genetic variation[139], hence the contribution of polygenic inheritance to the heritable risk of cancer may currently be underestimated.

Given the sample size of the GWAS that have been conducted, it is unlikely that there are many (or any) common disease loci with minor allele frequencies (MAFs) >20% in European populations that have stronger effects than those already identified for the major cancer types. Many of the loci identified have ORs of 1.1 or less (**Supplementary Table 1**), and the statistical power of most studies will be too low to detect effects of this magnitude for uncommon alleles (*i.e.* MAF < 10%)**.** As a consequence of the low statistical power as well as submaximal tagging the identification of risk variants conferring ORs of 1.05-1.1 will be problematic for all except the largest of studies[140].

[H1] Deciphering risk loci

The underlying premise of GWAS is that an association reveals the effect of a highly correlated functional variant that is in LD with the tag SNP. Therefore, the genotyped SNP is not generally a strong candidate for causality, and elucidation of the causal variant poses a considerable challenge. Specifically, it is difficult to establish which of a set of closely linked variants that are in LD with each other is the most functionally relevant. While a minority of GWAS tag SNPs are directly functional, for example the 8q24.21 SNP rs6983267 for CRC[141], most are likely in LD with the causal SNP. A key step in deciphering risk loci therefore is fine-mapping **[G]**, which is aided by imputation of untyped genotypes[142,143]. Moreover, fine-mapping can also resolve association signals, for example the 8q24.21 association for glioma where the imputed SNP rs55705857 has been shown to be sufficient to explain two tag SNP signals previously thought to be independent[132].

Many functional classes of genetic variation have been implicated as the basis of GWAS risk loci (**Fig. 4**). To date relatively few risk loci have been comprehensively studied. However, insights into the genetic and biological basis of cancer susceptibility mediated through common variation are emerging.

A small number of the identified cancer GWAS loci directly impact on the amino acid sequence of the expressed protein. The mechanistic interpretation of such variants is presumed to be relatively simple, due to the implied direct relationship between genotype and function. Examples include *BRCA2* p.Lys3326Ter (rs11571833) and *CHEK2* p.Ile157Thr (rs17879961) for lung[47] and breast cancer[42]. Similarly a direct relationship can be inferred for those affecting RNA processing such as the SNP in the 3'UTR (poly-A tail) of *TP53* (rs78378222) associated with prostate cancer and glioma risk[144,145], and those affecting splice sites such as the inhibitory splice isoform rs10069690 variant at 5p15.33 (*TERT*), resulting in decreased telomerase activity[146]. However, it is possible that coding variants could have more subtle effects that do not necessarily involve disrupting protein function[147], but instead involve tagging functional non-coding variants.

Most risk loci map to non-coding regions of the genome (*e.g.* gene introns or promoters and intergenic regions), which is perhaps unsurprising given that these regions comprise approximately 99% of the genome and the common, low–penetrance nature of these risk polymorphisms is more compatible with subtle, regulatory effects. Indeed GWAS risk loci have been demonstrated to map to genomic regions of cell-type specific active chromatin and show an over-representation of expression quantitative trait loci[148], methylation quantitative trait loci[149] and transcription factor (TF) binding[150]. Chromatin conformation studies have helped link regulatory regions, which SNPs identified by GWAS localise to, with their respective target genes[151-153]. Specifically, they have demonstrated that cis-regulatory effects mediated by chromatin looping interactions between enhancers and promoter regions within topologically associated domains (TADs)[154] are likely to be the functional basis of many GWAS signals.

There have been significant efforts to understand the regulatory mechanisms perturbed at cancer risk loci. Such studies have been aided by statistical methodologies such as Summary-data-based Mendelian Randomization[149] and initiatives such as the ENCODE[155], Roadmap Epigenomics[156] and BLUEPRINT epigenome[157] consortiums which have generated publicly available maps of regulatory regions. Furthermore, network-based approaches have yielded insights into higher-order

structures governing disease susceptibility. For example, binding of specific TFs can be enriched at risk loci. Such TFs are frequently mutated in tumours and have relevant biological activity[158].

The 8q24.21 region is one of the most intriguing and important loci to emerge from GWAS and is a good example of such regulatory mechanisms. The genomic interval at 128-130 megabases harbours multiple independent loci with distinct tumour specificities for CRC, glioma, CLL, MM, HL, and prostate, breast, and bladder cancers within the same TAD (**Fig. 4**). However, the region to which these cancer associations map is devoid of protein-coding transcripts. The 8q24.21 SNP rs6983267, which is associated with CRC and prostate cancer, resides in an evolutionarily conserved region. The two allelic variants of rs6983267 show differential binding of the transcription factor TCF7L2 to an enhancer element that physically interacts with the *MYC* promoter, which is 300 kilobases telomeric to rs6983267[141,159]. The *MYC* oncogene is commonly amplified or overexpressed in many cancers. Recent Hi-C analysis **[G]** of this region has demonstrated a more complicated regulatory mechanism, implicating various lincRNAs that mediate effects at risk loci for example *CCAT1, PCAT1* and *CCDC26* for CRC, prostate cancer and glioma respectively[152,153]. While studies to fully elucidate the regulatory mechanisms underpinning the 8q24.21 locus and risks of various cancers are in their relative infancy, such endeavours will likely involve exploration of tissue-specific effects in appropriate model systems and CRISPR/Cas9-mediated disruption of candidate regulatory elements[160].

[H1] Subtype-specific associations

Many cancers have distinct molecular profiles due to different aetiological pathways. The relationship between SNP genotype and tumour phenotype is becoming apparent for many cancer subtypes. In lung cancer, the 5p15.33 (*TERT-CLPTM1L*) and 3q28 (*TP63*) SNPs significantly influence lung cancer histology, and are principally associated with adenocarcinoma, while 13q12 (*BRCA2*) and human leucocyte antigen (HLA) associations are specific for squamous lung cancer[111]. Similarly, many glioma risk loci are subtype-specific, such as associations at 5p15.33, 20q13.33 and 7p11.2 for glioblastoma (GBM) and at 11q23.3 and 8q24.21 for non-GBM glioma[82]. Two of the most striking genotype-phenotype relationships identified to date are the 10q21.2 (*ARID5B*) ALL association, which seems to be highly selective for the subset of B-cell precursor ALL with hyperdiploidy (HD)[88] and the *CCND1* c.870G>A SNP which is specific for myeloma that has the (11;14)(q13;q32) translocation[97]. Presumably, such subtype-specific associations reflect particular

mutational signalling contexts; thereby potentially providing insight into tumour development. Susceptibility alleles increasing cancer risk might confer a selective advantage and therefore be preferentially enriched in the given cancer relative to the non-risk allele. Evidence for such a phenomenon has been demonstrated recently for the SNP rs7090445 at 10q21.2. Here, the risk allele is preferentially retained in HD-ALL blasts, consistent with inherited genetic variation contributing to arrest of normal lymphocyte development, and this facilitates leukaemic clonal expansion[118]. Similarly, the risk allele of the missense variant *CDKN2A* p.Ala148Thr (rs3731249), which increases ALL risk, has been shown to be preferentially selected during clonal evolution[161].

The recently proposed omnigenic model of complex disease susceptibility proposes that any gene with regulatory variants in disease relevant tissues will have an effect on disease risk. In this model, genes are defined as "core" if they have a specific role in disease aetiology and "peripheral" if their role is indirect. Given that there are more peripheral genes than core genes, and the range of effect sizes observed, a large fraction of the total genetic contribution to disease is thought to arise from peripheral genes that do not play direct roles in disease. Therefore, under this model, peripheral genes affect the regulation and function of core genes through networks, in a relatively subtle manner. This model is based on our admittedly limited understanding of cancer and network biology and remains to be proven experimentally[162].

[H1] Clinical relevance

As well as offering the prospect of risk stratification, cancer genetics provides for a better understanding of the developmental basis of cancer at a fundamental level. Such information can have direct clinical application in a number of contexts (**Fig. 5**).

[H1] Drug discovery and repositioning

Cancer genome sequencing studies provide evidence that regulatory regions and target genes implicated by GWAS are frequently the subject of somatic mutation, reflecting "driver activity"[163-166]. Such studies can aid in deciphering risk loci and offer the prospect of maximising drug discovery efforts. Indeed, there are many successfully approved drugs for which GWAS has provided direct supporting genetic evidence[167]. This evidence has highlighted targets for drug development and identified targets for potential drug repositioning[168,169]. Although not in the

context of cancer, proof-of principle for this has been provided by the use of Ustekinumab, a monoclonal antibody that neutralizes the shared p40 subunit of IL-12 and IL-23[170]. GWAS identified the IL-23 signalling pathway as a risk factor for the development of psoriasis[171] and the *IL23R* p.Arg381Gln (rs11209026) polymorphism was shown to afford protection from multiple inflammatory diseases[172]. Approved and promising therapies in cancer for which GWAS associations exist include BCL2 inhibition in CLL[92,173] and FGFR inhibition in breast cancer[40,174]. However, further work is required to identify target genes and aberrant biological pathways from GWAS associations and to define the germline-somatic continuum to maximise the potential of GWAS in drug discovery[175].

[H1] Stratified screening

The possibility of identifying those at increased risk on the basis of their genotype is of more immediate clinical relevance, since it will help to tailor prevention or screening strategies. The low level of risk associated with most cancer GWAS risk variants has been considered a barrier to the clinical application of these markers in cancer prevention. However, small effect sizes associated with individual SNPs do not necessarily preclude clinical utility. As demonstrated for CRC as well as breast and prostate cancer, the combined effect of multiple risk SNPs has the potential to achieve a degree of risk discrimination that is useful for population-based prevention and screening programmes. For example, a polygenic risk score (PRS) based on the 37 known risk variants for CRC indicates that individuals with the top 10% highest scores will have a 1.8-fold increased risk of CRC and those within the top 1% will have a 2.9-fold increased risk of CRC when compared with the population median (**Fig. 6**)[138,176]. Making use of a PRS has the potential to optimise the efficiency of population-based screening programmes for the early detection of CRC, prostate cancer and breast cancer[138,177]. Furthermore, the observed level of risk discrimination from PRS may be informative in formulating and delivering chemoprevention strategies. Use of PRS has also recently been shown to provide informative cancer risk stratification in the context of Mendelian cancer susceptibility, notably for *BRCA1* and *BRCA2* mutation carriers[178,179].

[H1] Informing prevention

Mendelian Randomisation (MR) analysis enables identification of non-genetic risk factors and possible chemoprevention agents by use of GWAS data [**BOX 2**]. For example, by using genetic

markers as proxies (*i.e.* genetic instruments) for hyperlipidaemia, a causal relationship between hypercholesterolemia and CRC has been demonstrated[180]. Furthermore, a genetic risk score comprising SNPs which lower 3-hydroxy-3-methylglutaryl-CoA reductase (HMGCR) expression, and therefore mimicking the effects of statin therapy to reduce cholesterol levels, was associated with reduced CRC risk[180]. Such data provides support for additional clinical benefit from statins aside from their primary use in the context of coronary heart disease. A significant challenge when conducting MR analysis is to ensure validity of genetic instruments by excluding pleiotropy [**BOX 2**]. Whilst methods have been developed to quantify pleiotropy in such studies, these have received scrutiny in the proposed omnigenic model that proposes the concept of "network pleiotropy", *i.e.* a single variant may affect multiple traits because those traits are mediated through the same regulatory networks in the same cell types rather than because the traits are causally related[162]. Such "network pleiotropy" may not be readily detected in traditional MR-based analyses.

[H1] Informing treatment

As a potential prognostic factor, the concept of germline variation imparting inter-individual variability in tumour development and progression is receiving increasing attention with examples in many cancer types[181-185]. Genetic variation has been linked to treatment response in CLL, and lung and breast cancer, whereby chemotherapies that are CYP3A substrates such as cyclophosphamide, taxanes and mitoxantrone, may be suboptimal for *CYP3A7*1C* carriers[186]. Furthermore, GWAS has been successful in identifying individuals at risk of treatment related toxicity such as anthracycline-induced cardiotoxicity[187] and radiotherapy induced tissue damage[188]. Hence GWAS offers an opportunity to realize the vision of personalized medicine by identifying common genetic variation affecting drug efficacy and drug-induced toxicity. This can improve therapeutic decision making, enabling the possibility of patient-tailored drug selection.

[H1] Conclusions and future challenges

GWAS have demonstrated that much of the heritable risk for most common cancers is polygenic. Hence the architecture of inherited genetic susceptibility to cancer is defined by a montage of predisposition alleles with different levels of risk and prevalence in the population. With the notable exception of breast and prostate cancer, the currently identified loci explain only a small

proportion of the familial risk of many cancers. Many GWAS have long tails of low OR associations, suggesting that larger studies should identify many more new susceptibility loci. Although rare recurrent disease-causing variants may not make a substantial contribution to the heritable risk of cancer, this class of risk variants have probably been under-discovered. Hence subjecting GWAS datasets to imputation using recently developed reference panels to recover sub-polymorphic risk alleles (*i.e*. risk allele frequency <1%) is likely to be a profitable avenue of research.

The loci identified through GWAS have greatly expanded the existing repertoire of genes that influence cancer risk. Determining the functional consequences of GWAS data is however likely to continue to be challenging, but is required to fully exploit GWAS in order to gain a greater understanding of cancer biology and suggest potential targets for therapeutic and preventive strategies. Advances in model systems and strategies such as saturating mutagenesis of risk loci using CRISPR/Cas9 is likely to facilitate such analyses[189,190].

Acknowledgements

We are grateful to Cancer Research UK for Support. AS is in receipt of a Clinical Training Fellowship from Cancer Research UK.

Competing interests statement

The authors declare no competing financial interests.

Author contributions

A.S., B.K. and R.S.H. researched data for the article, made substantial contributions to discussions of the content, wrote the article and reviewed and/or edited the manuscript before submission.

BOXES

Box 1: Genome-wide significance threshold

While historically other thresholds have been proposed[40,191], the commonly accepted threshold for genome-wide significance is $P<5x10^{-8}$, *i.e.* a Bonferroni correction at the 5% significance level for 1,000,000 independent tests. The first published mention of this was through simulation studies by Risch and Merikangas in 1996[192], assuming 100,000 genes with five bi-allelic SNPs per gene, testing for each allele independently. This has proven remarkably close to empirical estimates such as 150 per 500 Kb from the International HapMap Consortium in 2005, which leads to a two-sided significance threshold of $5.5x10^{-8}$ when extending to the whole genome[33].

More recently this threshold has remained in place for sequencing studies or those making use of whole-genome imputation of >10 million common variants, under the assumption that LD between SNPs approximates to ~1 million independent tests.

Box 2: The principles of Mendelian randomization

In observational studies, establishing a causal relationship between two associated variables may not always be possible. Furthermore, unmeasured factors (confounders) may influence both variables and thus explain the observed association. Mendelian randomization (MR) is a technique aimed at unbiased assessment of causal effects and estimation of their magnitude.

MR uses genetic markers known to be associated with a potential risk factor in the assessment of its effect on another trait or disease[193]. These markers, termed instrumental variables (IVs), rely on a number of assumptions, namely that the IVs are solely associated with the trait or disease (*i.e.* absence of pleiotropy), and that the IVs are independent of confounders. This methodology can allow for causality to be assessed without the influence of confounding factors.

With the development of large genomic datasets and establishment of robust IVs in the form of genetic risk variants, MR offers the ability to identify non-genetic risk factors[194], chemopreventative agents[180] and perform safety analysis of therapies[195].

FIGURE LEGENDS

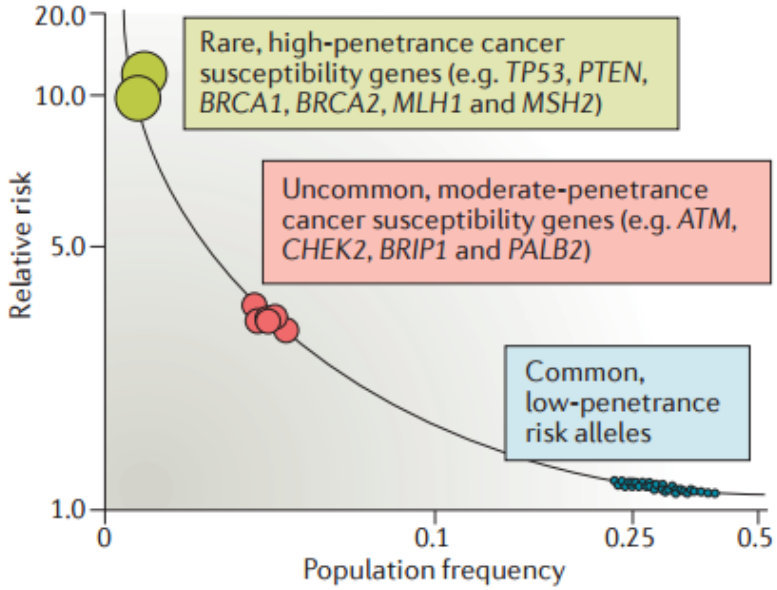

Figure 1: Genetic architecture of cancer risk. This graph depicts the low relative risks (RRs) associated with common, low-penetrance genetic variants, such as single-nucleotide polymorphisms identified in genome-wide association studies; moderate RRs associated with uncommon, moderate-penetrance genetic variants such as *ATM* and *CHEK2*; and a higher RR associated with rare, high-penetrance genetic variants, such as pathogenic mutations in *BRCA1* and *BRCA2* associated with hereditary breast and ovarian cancer.

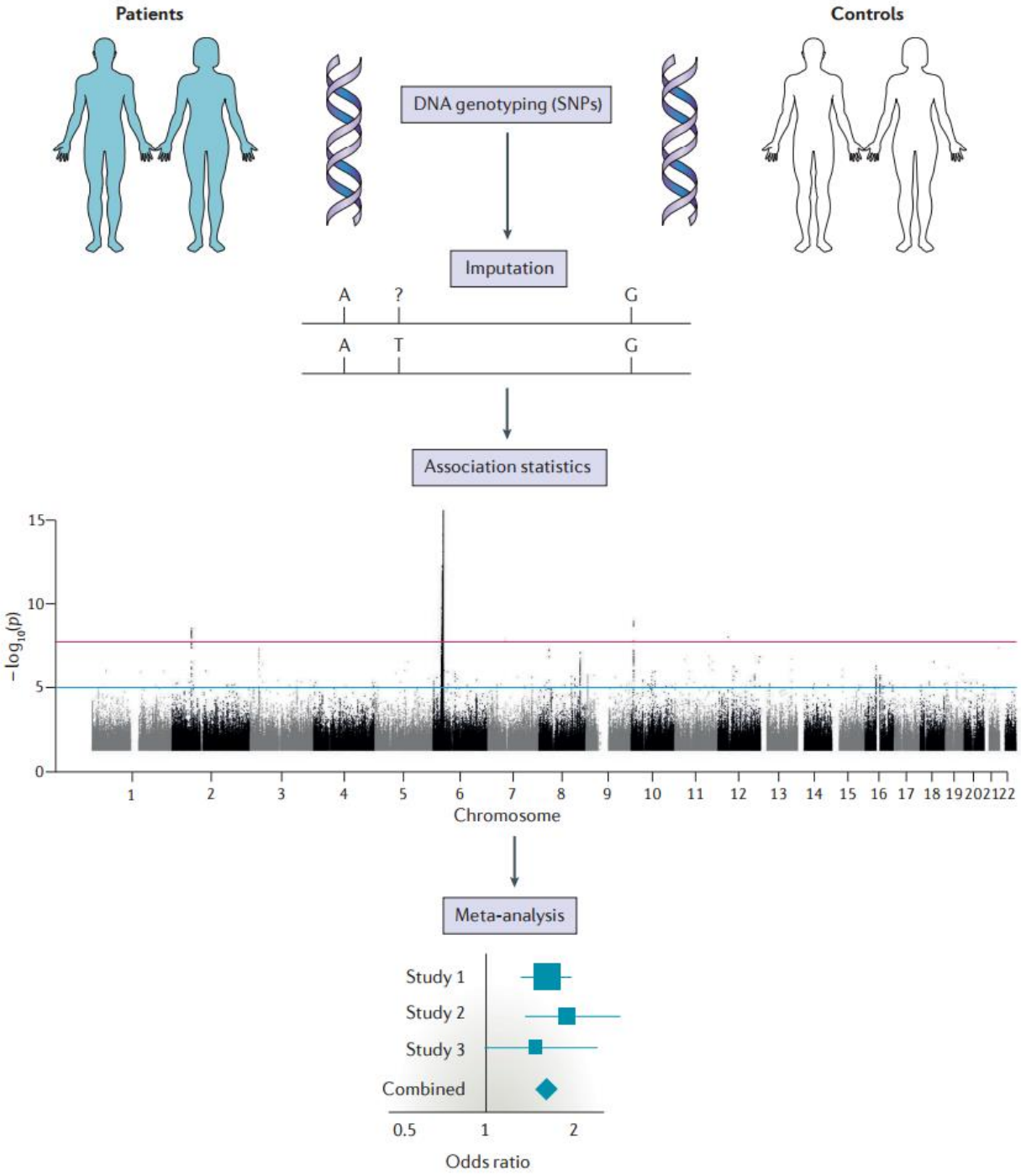

Figure 2: GWAS design. DNA from patients and controls are genotyped using commercially available genome platforms that assess for common genetic variations in the form of single nucleotide polymorphisms (SNPs) across the entire human genome. Data are reviewed to ensure appropriate genotyping quality. Genome imputation allows for recovery of untyped SNPs. Association test statistics are generated to identify genetic risk loci. Where more than one dataset is available, a meta-analysis is conducted to increase study power. Replication in appropriate study populations may also be performed to validate associations.

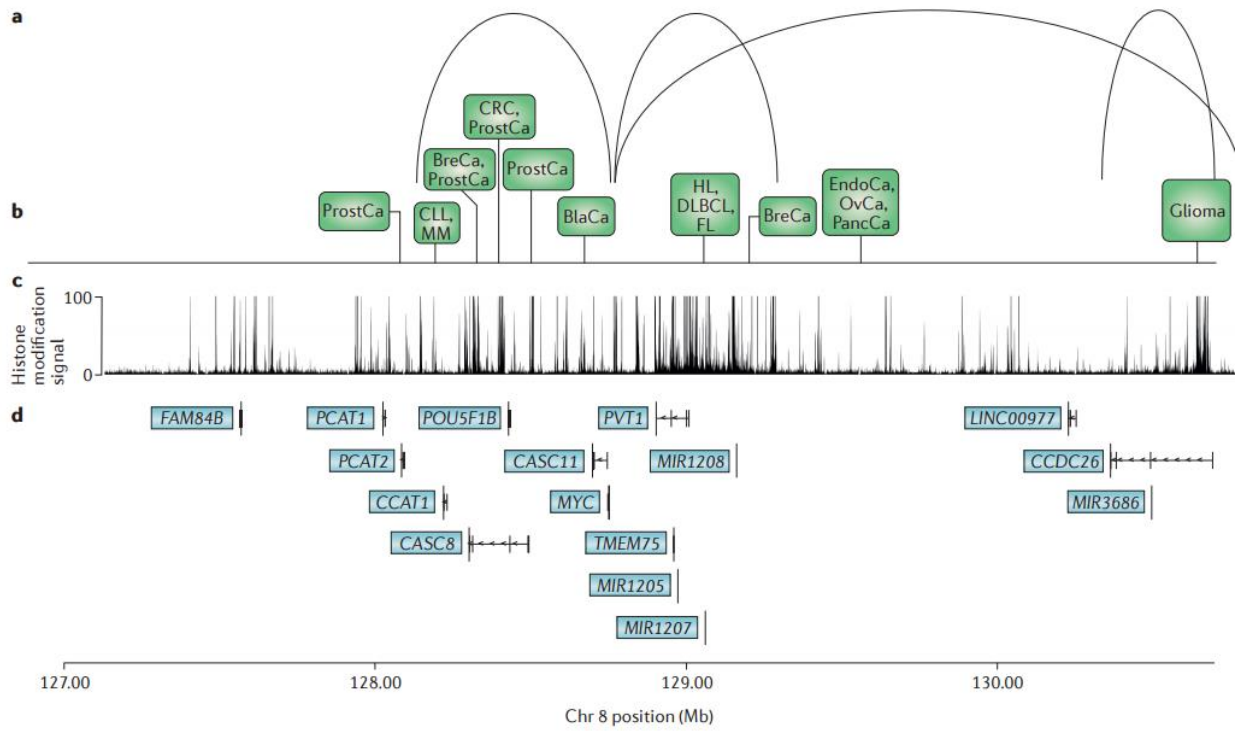

Figure 3: Regulatory interactions at the pleiotropic 8q24.21 risk locus. Plotted region is chr8:127,000,000-130,700,000 (build 37). (a) Looping interactions overlapping cancer association signals; (b) Relative location of cancer GWAS signals (**Supplementary Table 1**); (c) Epigenetic marks – peaks represent histone modifications and indicate DNA with the potential for influencing gene expression of neighbouring genes and through looping interactions, distant genes; (d) Refseq gene annotation (build 37). Abbreviations: PrC, prostate cancer; MM, multiple myeloma; CLL, chronic lymphocytic leukaemia; BrC, breast cancer; CRC, colorectal cancer; BlC, bladder cancer; FL, follicular lymphoma; HL, Hodgkin lymphoma; DLBCL, diffuse large B-cell lymphoma; EC, endometrial carcinoma; OC, ovarian cancer; PaC, pancreatic cancer.

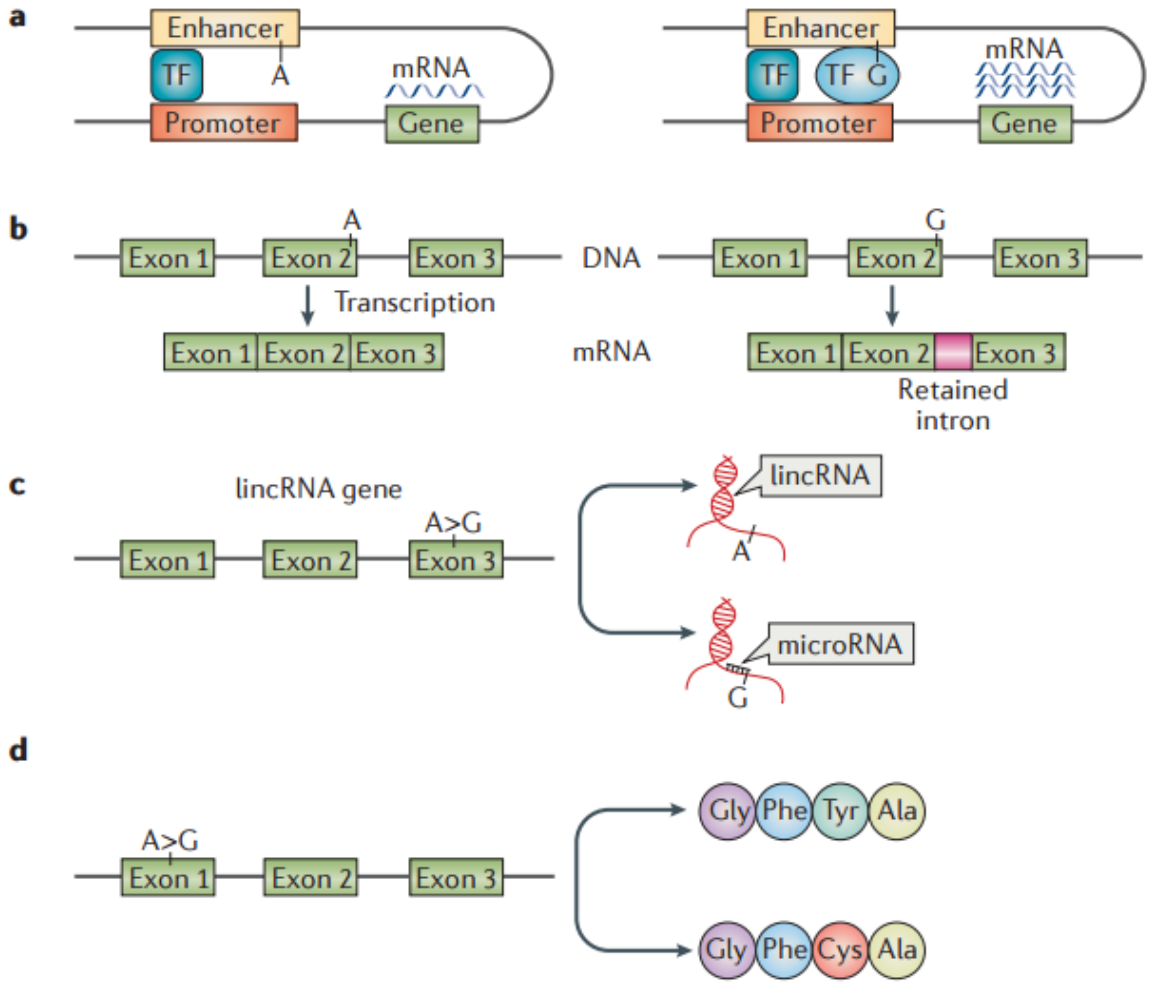

Figure 4: Potential molecular mechanisms of GWAS risk SNPs. (a) The A>G polymorphism is affecting gene transcription through altering transcription factor (TF) binding through looping promoter-enhancer-complex interaction; (b) Affecting mRNA processing (*e.g*. splicing, poly-adenylation). The A>G polymorphism depicted occurs at an intron splice site and results in intron retention; (c) The A>G polymorphism leads to generation of a novel microRNA binding site on a lincRNA; (d) The A>G polymorphism affects the protein sequence by causing amino acid substitution of tyrosine to cytosine)

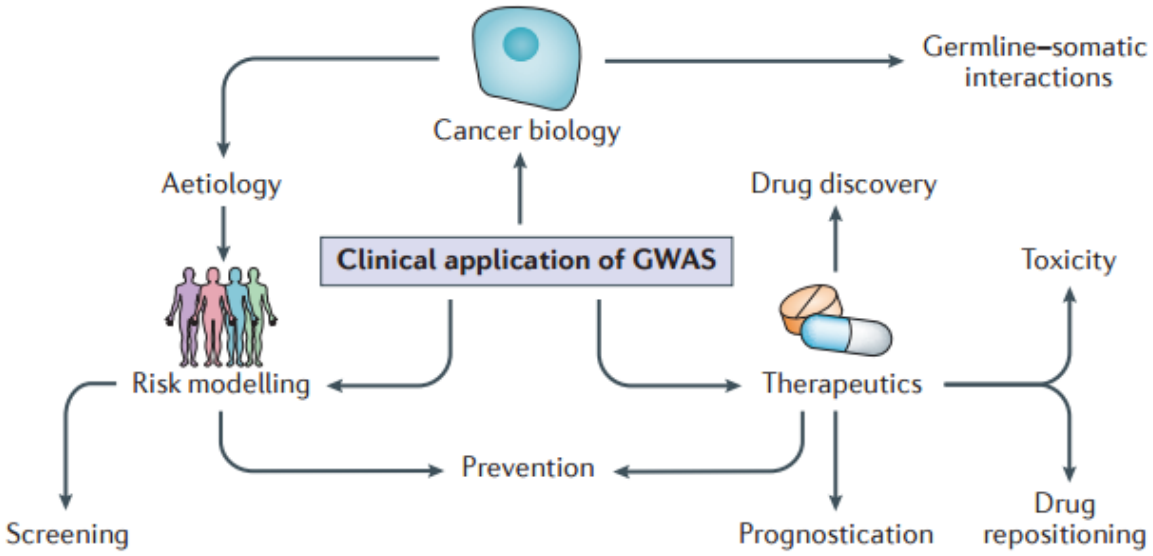

Figure 5: The clinical application of GWAS. Overview of different applications of cancer GWAS, as described in the review. As well as enhancing our knowledge of cancer biology, GWAS can inform on aetiological risk factors for cancer. Through risk modelling, data from GWAS can assist in identifying individuals at increased risk of developing cancer and therefore help prevent cancer and improve early detection through screening. Genes and pathways identified through GWAS may inform drug discovery and repositioning as well as guide clinicians and patients on cancer prognosis and treatment related complications.

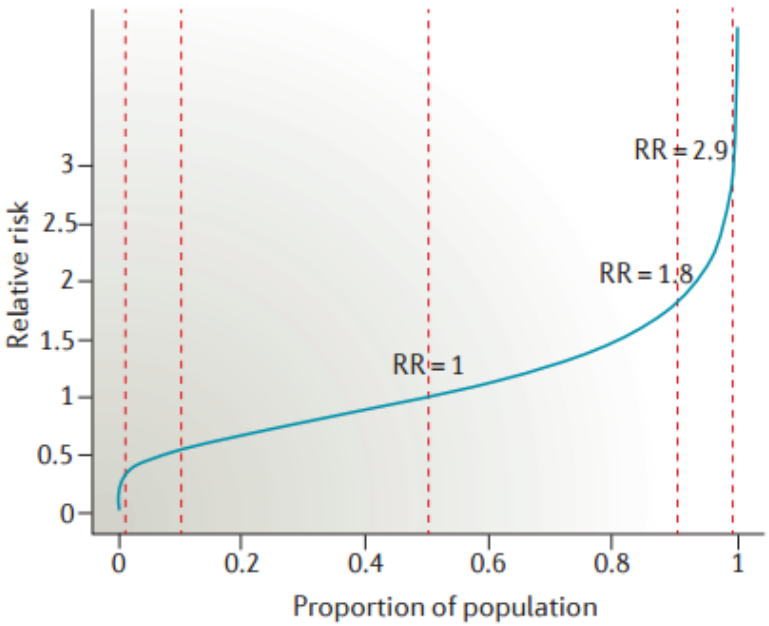

Figure 6: Population distribution of polygenic risk score for CRC ordered by RR. Based on the known 37 risk SNPs. Vertical red lines (left to right) correspond to 1%, 10%, 50%, 90%, and 99% centile, respectively. Individuals within the top 10% of genetic risk have a 1.8-fold increased risk of CRC and those within the top 1% (*i.e.* 27–54 risk alleles) have a 2.9-fold increased risk of CRC when compared with the population median.

GLOSSARY

Relative Risk (RR) – ratio of disease occurrence in one group versus another (*e.g*. cancer risk in patient relatives compared with the general population). The RR estimate associated with common risk alleles identified through GWAS is usually a per-allele RR (co-dominant log-additive genetic model).

Mendelian predisposition – This occurs when germline mutation in a single gene (*e.g.* cancer susceptibility gene) is sufficient to cause cancer in a majority of patients (*e.g.* female carriers of *BRCA1* mutations have ~80% lifetime risk of developing breast cancer). These mutations can be dominant or recessive, caused by mono-allelic and bi-allelic mutations respectively.

Cancer Susceptibility Gene (CSG) – genes in which inherited mutations (commonly high-penetrance) predispose to cancer.

Penetrance – proportion of individuals carrying a particular allele (e.g. in a cancer susceptibility gene) that go on to develop cancer. High-penetrance mutations confer a high risk of causing cancer, whereas low-penetrance polymorphisms confer a small risk.

Linkage Disequilibrium (LD) – non-random association of alleles at different sites in a given population. Alleles in high LD are those where their shared frequency combinations are greater than would be expected if they were inherited independently. LD can be affected by factors such as natural selection and genetic drift, as well as rates of mutation and recombination.

Heritability – estimate of the proportion of variation in a trait in a given population that is due to genetic variation. In particular, narrow-sense heritability (h^2) is the proportion of variance in a trait due to additive genetic factors, whereas broad-sense heritability (H^2) is the proportion of variance in a trait due to all genetic factors (e.g. including dominance, gene-gene interactions).

Genome-wide complex trait analysis (GCTA) - computational method by which the narrow-sense heritability of a trait can be estimated through case-control GWAS genotypes and estimates of trait incidence.

Fine-mapping – process of refining GWAS association signals and prioritising likely causative variants *e.g.* through *in silico* annotations of putative functional effect.

Hi-C analysis – a form of chromosome conformation capture, in which cross-linked DNA fragments are sequenced in order to infer the three-dimensional structure of the genome and identify potential regulatory interactions.

Effect size – quantitative measurement statistic of the strength of an association between two variables e.g. SNP genotype and cancer risk.

Risk allele frequency – frequency of risk allele (B) in a given population at a bi-allelic site with non-risk allele (A), derived from genotype counts through formula (2xBB+AB)/(2x(AA+AB+BB))

Odds ratio – odds that an outcome will occur given a particular exposure compared to the absence of that exposure *e.g*. comparing variant site allele frequency in cancer patients and controls.

Population attributable risk – number of cases of disease among exposed individuals that can be attributed to that exposure (*e.g.* carriers of a particular risk SNP).

Pleiotropy – occurs when a risk locus is associated with multiple phenotypic traits. In some cases the same variant is presumed to influence multiple traits, while in other cases different traits map to distinct locations within the risk locus.

SELECTED REFERENCES

[2] Lichtenstein, P. 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* **343**, 78-85 (2000). **Landmark paper estimating heritability of common cancers from an analysis of 44,788 twins.**

[31] International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789-96 (2003). **Insights from the HapMap genotyping project, demonstrating that patterns of LD between common variants can allow design of arrays of 200,000 – 1,000,000 "tag SNPs" to capture a large proportion of common SNPs (~10 million).**

[37] The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-73 (2010). **Initial findings from the 1000 genomes project, which characterised genetic variation in different populations after sequencing 1,092 individuals.**

[39] McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics* **48**, 1279-1283, doi:10.1038/ng.3643 (2016). **Demonstrating use of population reference haplotypes in imputation of GWAS arrays can allow genotype estimation at allele frequencies as low as 0.1%.**

[125] Thorgeirsson, T.E. et al. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* **452**, 638-42 (2008). **GWAS of smoking quantity, describing gene-lifestyle interaction between variation at 15q24 and nicotine dependence, leading to indirect association with lung cancer risk.**

[40] Easton, D.F. et al. Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087-93 (2007). **First breast cancer GWAS, describing the discovery of five risk loci.**

[153] Jager, R. et al. Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nat Commun* **6**, 6178 (2015). **Use of targeted capture approach to greatly enrich for Hi-C contacts within CRC risk regions, aiding functional interrogation of DNA regulatory interactions at the known risk loci.**

[141] Pomerantz, M.M. et al. The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nat Genet* **41**, 882-4 (2009). **One of first attempts to functionally characterise a cancer risk locus, demonstrating allele-specific differential binding of TCF7L2 to rs6983267, which encompasses an enhancer element that interacts with MYC.**

[160] Sur, I. K. *et al.* Mice lacking a Myc enhancer that includes human SNP rs6983267 are resistant to intestinal tumors. *Science* **338**, 1360-1363, doi:10.1126/science.1228606 (2012). **Genetic engineering in the mouse demonstrated an *in vivo* effect of a risk SNP in a regulatory element on tumour development.**

[109] Wang, Y. *et al.* Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nature genetics* **40**, 1407-1409, doi:10.1038/ng.273 (2008). **First lung cancer GWAS to identify common genetic factors influencing lung cancer risk in smokers.**

[114] Tomlinson, I. *et al.* A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nature genetics* **39**, 984-988, doi:10.1038/ng2085 (2007). **Paper reporting on the first GWAS of colorectal cancer, with discovery of rs6983267 at 8q24.21.**

REFERENCES

1 Houlston, R. & Peto, J. Genetics and the common cancers. *In: Genetic predisposition to cancer. eds Eeles, RA, Easton DF, Ponder BAJ, Eng, C.* **2nd Edition** 235-248 (2004).

2 Lichtenstein, P. *et al.* Environmental and heritable factors in the causation of cancer--analyses of cohorts of twins from Sweden, Denmark, and Finland. *The New England journal of medicine* **343**, 78-85, doi:10.1056/NEJM200007133430201 (2000).

3 Wiemels, J. L. *et al.* Prenatal origin of acute lymphoblastic leukaemia in children. *Lancet* **354**, 1499-1503 (1999).

4 Peto, J. & Houlston, R. S. Genetics and the common cancers. *European journal of cancer* **37 Suppl 8**, S88-96 (2001).

5 Anderson, D. E. Genetic study of breast cancer: identification of a high risk group. *Cancer* **34**, 1090-1097 (1974).

6 Miki, Y. *et al.* A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. *Science* **266**, 66-71, doi:10.1126/science.7545954 (1994).

7 Wooster, R. *et al.* Localization of a breast cancer susceptibility gene, BRCA2, to chromosome 13q12-13. *Science* **265**, 2088-2090 (1994).

8 Hall, J. M. *et al.* Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* **250**, 1684-1689 (1990).

9 Peltomaki, P. *et al.* Genetic mapping of a locus predisposing to human colorectal cancer. *Science* **260**, 810-812 (1993).

10 Lindblom, A., Tannergard, P., Werelius, B. & Nordenskjold, M. Genetic mapping of a second locus predisposing to hereditary non-polyposis colon cancer. *Nature genetics* **5**, 279-282, doi:10.1038/ng1193-279 (1993).

11 Kinzler, K. W. *et al.* Identification of FAP locus genes from chromosome 5q21. *Science* **253**, 661-665 (1991).

12 Fishel, R. *et al.* The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell* **75**, 1027-1038 (1993).

13 Leach, F. S. *et al.* Mutations of a mutS homolog in hereditary nonpolyposis colorectal cancer. *Cell* **75**, 1215-1225, doi:http://dx.doi.org/10.1016/0092-8674(93)90330-S (1993).

14 Cannon-Albright, L. A. *et al.* Assignment of a locus for familial melanoma, MLM, to chromosome 9p13-p22. *Science* **258**, 1148-1152 (1992).

15 Hussussian, C. J. *et al.* Germline p16 mutations in familial melanoma. *Nature genetics* **8**, 15-21, doi:10.1038/ng0994-15 (1994).

16 Ballinger, M. L. *et al.* Monogenic and polygenic determinants of sarcoma risk: an international genetic study. *The Lancet. Oncology* **17**, 1261-1271, doi:10.1016/S1470-2045(16)30147-4 (2016).

17 Aaltonen, L., Johns, L., Jarvinen, H., Mecklin, J. P. & Houlston, R. Explaining the familial colorectal cancer risk associated with mismatch repair (MMR)-deficient and MMR-stable tumors. *Clinical cancer research : an official journal of the American Association for Cancer Research* **13**, 356-361, doi:10.1158/1078-0432.CCR-06-1256 (2007).

18 Peto, J. *et al.* Prevalence of BRCA1 and BRCA2 gene mutations in patients with early-onset breast cancer. *Journal of the National Cancer Institute* **91**, 943-949 (1999).

19 Chubb, D. *et al.* Rare disruptive mutations and their contribution to the heritable risk of colorectal cancer. *Nat Commun* **7**, 11883, doi:10.1038/ncomms11883 (2016).

20 Anglian Breast Cancer Study Group. Prevalence and penetrance of BRCA1 and BRCA2 mutations in a population-based series of breast cancer cases. *Br J Cancer* **83**, 1301-1308, doi:10.1054/bjoc.2000.1407 (2000).

21 Lubbe, S. J., Webb, E. L., Chandler, I. P. & Houlston, R. S. Implications of familial colorectal cancer risk profiles and microsatellite instability status. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* **27**, 2238-2244, doi:10.1200/JCO.2008.20.3364 (2009).

22 Palles, C. *et al.* Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nature genetics* **45**, 136-144, doi:10.1038/ng.2503 (2013).

23 Weren, R. D. *et al.* A germline homozygous mutation in the base-excision repair gene NTHL1 causes adenomatous polyposis and colorectal cancer. *Nature genetics* **47**, 668-671, doi:10.1038/ng.3287 (2015).

24 Swift, M., Reitnauer, P. J., Morrell, D. & Chase, C. L. Breast and other cancers in families with ataxia-telangiectasia. *The New England journal of medicine* **316**, 1289-1294, doi:10.1056/NEJM198705213162101 (1987).

25 Renwick, A. *et al.* ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. *Nature genetics* **38**, 873-875, doi:10.1038/ng1837 (2006).

26 Meijers-Heijboer, H. *et al.* Low-penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in noncarriers of BRCA1 or BRCA2 mutations. *Nature genetics* **31**, 55-59, doi:10.1038/ng879 (2002).

27 Rahman, N. *et al.* PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nature genetics* **39**, 165-167, doi:10.1038/ng1959 (2007).

28 Erkko, H. *et al.* A recurrent mutation in PALB2 in Finnish cancer families. *Nature* **446**, 316-319, doi:10.1038/nature05609 (2007).

29 Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304-1351, doi:10.1126/science.1058040 (2001).

30 Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921, doi:10.1038/35057062 (2001).

31 International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789-796, doi:10.1038/nature02168 (2003).

32 Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J. & Lander, E. S. High-resolution haplotype structure in the human genome. *Nature genetics* **29**, 229-232, doi:10.1038/ng1001-229 (2001).

33 International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299-1320, doi:10.1038/nature04226 (2005).

34 Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics* **38**, 904-909, doi:10.1038/ng1847 (2006).

35 Ioannidis, J. P., Ntzani, E. E., Trikalinos, T. A. & Contopoulos-Ioannidis, D. G. Replication validity of genetic association studies. *Nature genetics* **29**, 306-309, doi:10.1038/ng749 (2001).

36 Lohmueller, K. E., Pearce, C. L., Pike, M., Lander, E. S. & Hirschhorn, J. N. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature genetics* **33**, 177-182, doi:10.1038/ng1071 (2003).

37 The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061-1073, doi:10.1038/nature09534 (2010).

38 Huang, J. *et al.* Improved imputation of low-frequency and rare variants using the UK10K haplotype reference panel. *Nat Commun* **6**, 8111, doi:10.1038/ncomms9111 (2015).

39 McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nature genetics* **48**, 1279-1283, doi:10.1038/ng.3643 (2016).

40 Easton, D. F. *et al.* Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087-1093, doi:10.1038/nature05887 (2007).

41 Michailidou, K. *et al.* Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nat Genet* **47**, 373-380, doi:10.1038/ng.3242 (2015).

42 Michailidou, K. *et al.* Large-scale genotyping identifies 41 new loci associated with breast cancer risk. *Nature genetics* **45**, 353-361, 361e351-352, doi:10.1038/ng.2563 (2013).

43 Amundadottir, L. T. *et al.* A common variant associated with prostate cancer in European and African populations. *Nature genetics* **38**, 652-658, doi:10.1038/ng1808 (2006).

44 Eeles, R. A. *et al.* Multiple newly identified loci associated with prostate cancer susceptibility. *Nature genetics* **40**, 316-321, doi:10.1038/ng.90 (2008).

45 Al Olama, A. A. *et al.* A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. *Nat Genet* **46**, 1103-1109, doi:10.1038/ng.3094 (2014).

46 Amos, C. I. *et al.* Genome-wide association scan of tag SNPs identifies a susceptibility locus for lung cancer at 15q25.1. *Nature genetics* **40**, 616-622, doi:10.1038/ng.109 (2008).

47 Wang, Y. *et al.* Rare variants of large effect in BRCA2 and CHEK2 affect risk of lung cancer. *Nature genetics* **46**, 736-741, doi:10.1038/ng.3002 (2014).

48 Wang, Y., Broderick, P., Matakidou, A., Eisen, T. & Houlston, R. S. Chromosome 15q25 (CHRNA3-CHRNA5) variation impacts indirectly on lung cancer risk. *PLoS One* **6**, e19085, doi:10.1371/journal.pone.0019085 (2011).

49 McKay, J. D. *et al.* Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nature genetics* **49**, 1126-1132, doi:10.1038/ng.3892 (2017).

50 Dunlop, M. G. *et al.* Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk. *Nature genetics* **44**, 770-776, doi:10.1038/ng.2293 (2012).

51 Houlston, R. S. *et al.* Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. *Nature genetics* **42**, 973-977, doi:10.1038/ng.670 (2010).

52 Jaeger, E. *et al.* Common genetic variants at the CRAC1 (HMPS) locus on chromosome 15q13.3 influence colorectal cancer risk. *Nature genetics* **40**, 26-28, doi:10.1038/ng.2007.41 (2008).

53 Study, C. *et al.* Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nature genetics* **40**, 1426-1435, doi:10.1038/ng.262 (2008).

54 Tomlinson, I. P. *et al.* A genome-wide association study identifies colorectal cancer susceptibility loci on chromosomes 10p14 and 8q23.3. *Nature genetics* **40**, 623-630, doi:10.1038/ng.111 (2008).

55 Schumacher, F. R. *et al.* Genome-wide association study of colorectal cancer identifies six new susceptibility loci. *Nature communications* **6**, 7138, doi:10.1038/ncomms8138 (2015).

56 Orlando, G. *et al.* Variation at 2q35 (PNKD and TMBIM1) influences colorectal cancer risk and identifies a pleiotropic effect with inflammatory bowel disease. *Human molecular genetics* **25**, 2349-2359, doi:10.1093/hmg/ddw087 (2016).

57 Amundadottir, L. *et al.* Genome-wide association study identifies variants in the ABO locus associated with susceptibility to pancreatic cancer. *Nature genetics* **41**, 986-990, doi:10.1038/ng.429 (2009).

58 Petersen, G. M. *et al.* A genome-wide association study identifies pancreatic cancer susceptibility loci on chromosomes 13q22.1, 1q32.1 and 5p15.33. *Nature genetics* **42**, 224-228, doi:10.1038/ng.522 (2010).

59 Childs, E. J. *et al.* Common variation at 2p13.3, 3q29, 7p13 and 17q25.1 associated with susceptibility to pancreatic cancer. *Nature genetics* **47**, 911-916, doi:10.1038/ng.3341 (2015).

60 Wolpin, B. M. *et al.* Genome-wide association study identifies multiple susceptibility loci for pancreatic cancer. *Nature genetics* **46**, 994-1000, doi:10.1038/ng.3052 (2014).

61 Abnet, C. C. *et al.* A shared susceptibility locus in PLCE1 at 10q23 for gastric adenocarcinoma and esophageal squamous cell carcinoma. *Nature genetics* **42**, 764-767, doi:10.1038/ng.649 (2010).

62 Helgason, H. *et al.* Loss-of-function variants in ATM confer risk of gastric cancer. *Nature genetics* **47**, 906-910, doi:10.1038/ng.3342 (2015).

63 Purdue, M. P. *et al.* Genome-wide association study of renal cell carcinoma identifies two susceptibility loci on 2p21 and 11q13.3. *Nature genetics* **43**, 60-65, doi:10.1038/ng.723 (2011).

64 Gudmundsson, J. *et al.* A common variant at 8q24.21 is associated with renal cell cancer. *Nature communications* **4**, 2776, doi:10.1038/ncomms3776 (2013).

65 Scelo, G. *et al.* Genome-wide association study identifies multiple risk loci for renal cell carcinoma. *Nature communications* **8**, 15724, doi:10.1038/ncomms15724 (2017).

66 Kiemeney, L. A. *et al.* Sequence variant on 8q24 confers susceptibility to urinary bladder cancer. *Nature genetics* **40**, 1307-1312, doi:10.1038/ng.229 (2008).

67 Rothman, N. *et al.* A multi-stage genome-wide association study of bladder cancer identifies multiple susceptibility loci. *Nature genetics* **42**, 978-984, doi:10.1038/ng.687 (2010).

68 Reich, D. E. *et al.* Linkage disequilibrium in the human genome. *Nature* **411**, 199-204, doi:10.1038/35075590 (2001).

69 Bishop, D. T. *et al.* Genome-wide association study identifies three loci associated with melanoma risk. *Nature genetics* **41**, 920-925, doi:10.1038/ng.411 (2009).

70 Law, M. H. *et al.* Genome-wide meta-analysis identifies five new susceptibility loci for cutaneous malignant melanoma. *Nature genetics* **47**, 987-995, doi:10.1038/ng.3373 (2015).

71 Barrett, J. H. *et al.* Genome-wide association study identifies three new melanoma susceptibility loci. *Nature genetics* **43**, 1108-1113, doi:10.1038/ng.959 (2011).

72 Song, H. *et al.* A genome-wide association study identifies a new ovarian cancer susceptibility locus on 9p22.2. *Nature genetics* **41**, 996-1000, doi:10.1038/ng.424 (2009).

73 Kuchenbaecker, K. B. *et al.* Identification of six new susceptibility loci for invasive epithelial ovarian cancer. *Nature genetics* **47**, 164-171, doi:10.1038/ng.3185 (2015).

74 Pharoah, P. D. *et al.* GWAS meta-analysis and replication identifies three new susceptibility loci for ovarian cancer. *Nature genetics* **45**, 362-370, 370e361-362, doi:10.1038/ng.2564 (2013).

75 Goode, E. L. *et al.* A genome-wide association study identifies susceptibility loci for ovarian cancer at 2q31 and 8q24. *Nature genetics* **42**, 874-879, doi:10.1038/ng.668 (2010).

76 Gudbjartsson, D. F. *et al.* ASIP and TYR pigmentation variants associate with cutaneous melanoma and basal cell carcinoma. *Nature genetics* **40**, 886-891, doi:10.1038/ng.161 (2008).

77 Chahal, H. S. *et al.* Genome-wide association study identifies 14 novel risk alleles associated with basal cell carcinoma. *Nature communications* **7**, 12510, doi:10.1038/ncomms12510 (2016).

78 Stacey, S. N. *et al.* New common variants affecting susceptibility to basal cell carcinoma. *Nature genetics* **41**, 909-914, doi:http://www.nature.com/ng/journal/v41/n8/suppinfo/ng.412_S1.html (2009).

79 Stacey, S. N. *et al.* New basal cell carcinoma susceptibility loci. *Nature communications* **6**, 6825, doi:10.1038/ncomms7825 (2015).

80 Kinnersley, B. *et al.* Genome-wide association study identifies multiple susceptibility loci for glioma. *Nat Commun* **6**, 8559, doi:10.1038/ncomms9559 (2015).

81 Shete, S. *et al.* Genome-wide association study identifies five susceptibility loci for glioma. *Nature genetics* **41**, 899-904, doi:10.1038/ng.407 (2009).

82 Melin, B. S. *et al.* Genome-wide association study of glioma subtypes identifies specific differences in genetic susceptibility to glioblastoma and non-glioblastoma tumors. *Nature genetics*, doi:10.1038/ng.3823 (2017).

83 Dobbins, S. E. *et al.* Common variation at 10p12.31 near MLLT10 influences meningioma risk. *Nature genetics* **43**, 825-827, doi:10.1038/ng.879 (2011).

84 Litchfield, K. *et al.* Identification of 19 new risk loci and potential regulatory mechanisms influencing susceptibility to testicular germ cell tumor. *Nature genetics* **49**, 1133-1140, doi:10.1038/ng.3896 http://www.nature.com/ng/journal/v49/n7/abs/ng.3896.html#supplementary-information (2017).

85 Litchfield, K. *et al.* Identification of four new susceptibility loci for testicular germ cell tumour. *Nature communications* **6**, 8690, doi:10.1038/ncomms9690 (2015).

86 Wang, Z. *et al.* Meta-analysis of five genome-wide association studies identifies multiple new loci associated with testicular germ cell tumor. *Nature genetics* **49**, 1141-1147, doi:10.1038/ng.3879 (2017).

87 Gudmundsson, J. *et al.* A genome-wide association study yields five novel thyroid cancer risk loci. *Nature communications* **8**, 14517, doi:10.1038/ncomms14517 (2017).

88 Papaemmanuil, E. *et al.* Loci on 7p12.2, 10q21.2 and 14q11.2 are associated with risk of childhood acute lymphoblastic leukemia. *Nature genetics* **41**, 1006-1010, doi:10.1038/ng.430 (2009).

89 Vijayakrishnan, J. *et al.* A genome-wide association study identifies risk loci for childhood acute lymphoblastic leukemia at 10q26.13 and 12q23.1. *Leukemia* **31**, 573-579, doi:10.1038/leu.2016.271 (2017).

90 Vijayakrishnan, J. *et al.* The 9p21.3 risk of childhood acute lymphoblastic leukaemia is explained by a rare high-impact variant in CDKN2A. *Sci Rep* **5**, 15065, doi:10.1038/srep15065 (2015).

91 Migliorini, G. *et al.* Variation at 10p12.2 and 10p14 influences risk of childhood B-cell acute lymphoblastic leukemia and phenotype. *Blood* **122**, 3298-3307, doi:10.1182/blood-2013-03-491316 (2013).

92 Law, P. J. *et al.* Genome-wide association analysis implicates dysregulation of immunity genes in chronic lymphocytic leukaemia. *Nature communications* **8**, 14175, doi:10.1038/ncomms14175 (2017).

93 Crowther-Swanepoel, D. *et al.* Common variants at 2q37.3, 8q24.21, 15q21.3 and 16q24.1 influence chronic lymphocytic leukemia risk. *Nat Genet* **42**, 132-136, doi:10.1038/ng.510 (2010).

94 Di Bernardo, M. C. *et al.* A genome-wide association study identifies six susceptibility loci for chronic lymphocytic leukemia. *Nature genetics* **40**, 1204-1210, doi:10.1038/ng.219 (2008).

95 Broderick, P. *et al.* Common variation at 3p22.1 and 7p15.3 influences multiple myeloma risk. *Nature genetics* **44**, 58-61, doi:10.1038/ng.993 (2011).

96 Chubb, D. *et al.* Common variation at 3q26.2, 6p21.33, 17p11.2 and 22q13.1 influences multiple myeloma risk. *Nature genetics* **45**, 1221-1225, doi:10.1038/ng.2733 (2013).

97 Weinhold, N. *et al.* The CCND1 c.870G>A polymorphism is a risk factor for t(11;14)(q13;q32) multiple myeloma. *Nature genetics* **45**, 522-525, doi:10.1038/ng.2583 (2013).

98 Mitchell, J. S. *et al.* Genome-wide association study identifies multiple susceptibility loci for multiple myeloma. *Nature communications* **7**, 12050, doi:10.1038/ncomms12050 (2016).

99 Enciso-Mora, V. *et al.* A genome-wide association study of Hodgkin's lymphoma identifies new susceptibility loci at 2p16.1 (REL), 8q24.21 and 10p14 (GATA3). *Nature genetics* **42**, 1126-1130, doi:10.1038/ng.696 (2010).

100 Frampton, M. *et al.* Variation at 3p24.1 and 6q23.3 influences the risk of Hodgkin's lymphoma. *Nature communications* **4**, 2549, doi:10.1038/ncomms3549 (2013).

101 Cozen, W. *et al.* A meta-analysis of Hodgkin lymphoma reveals 19p13.3 TCF3 as a novel susceptibility locus. *Nature communications* **5**, 3856, doi:10.1038/ncomms4856 (2014).

102 Skibola, C. F. *et al.* Genome-wide association study identifies five susceptibility loci for follicular lymphoma outside the HLA region. *Am J Hum Genet* **95**, 462-471, doi:10.1016/j.ajhg.2014.09.004 (2014).

103 Cerhan, J. R. *et al.* Genome-wide association study identifies multiple susceptibility loci for diffuse large B cell lymphoma. *Nature genetics* **46**, 1233-1238, doi:10.1038/ng.3105 (2014).

104 Turnbull, C. *et al.* A genome-wide association study identifies susceptibility loci for Wilms tumor. *Nature genetics* **44**, 681-684, doi:10.1038/ng.2251 (2012).

105 Diskin, S. J. *et al.* Copy number variation at 1q21.1 associated with neuroblastoma. *Nature* **459**, 987-991, doi:10.1038/nature08035 (2009).

106 Goldin, L. R., Pfeiffer, R. M., Li, X. & Hemminki, K. Familial risk of lymphoproliferative tumors in families of patients with chronic lymphocytic leukemia: results from the Swedish Family-Cancer Database. *Blood* **104**, 1850-1854, doi:10.1182/blood-2004-01-0341 (2004).

107 Codd, V. *et al.* Identification of seven loci affecting mean telomere length and their association with disease. *Nature genetics* **45**, 422-427, doi:10.1038/ng.2528 (2013).

108 Rafnar, T. *et al.* Sequence variants at the TERT-CLPTM1L locus associate with many cancer types. *Nature genetics* **41**, 221-227, doi:10.1038/ng.296 (2009).

109 Wang, Y. *et al.* Common 5p15.33 and 6p21.33 variants influence lung cancer risk. *Nature genetics* **40**, 1407-1409, doi:10.1038/ng.273 (2008).

110 Sherborne, A. L. *et al.* Variation in CDKN2A at 9p21.3 influences childhood acute lymphoblastic leukemia risk. *Nature genetics* **42**, 492-494, doi:10.1038/ng.585 (2010).

111 Timofeeva, M. N. *et al.* Influence of common genetic variation on lung cancer risk: meta-analysis of 14 900 cases and 29 485 controls. *Human molecular genetics* **21**, 4980-4995, doi:10.1093/hmg/dds334 (2012).

112 Falchi, M., Spector, T. D., Perks, U., Kato, B. S. & Bataille, V. Genome-wide search for nevus density shows linkage to two melanoma loci on chromosome 9 and identifies a new QTL on 5q31 in an adult twin cohort. *Human molecular genetics* **15**, 2975-2979, doi:10.1093/hmg/ddl227 (2006).

113 Yeager, M. *et al.* Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nature genetics* **39**, 645-649, doi:http://www.nature.com/ng/journal/v39/n5/suppinfo/ng2022_S1.html (2007).

114 Tomlinson, I. *et al.* A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nature genetics* **39**, 984-988, doi:10.1038/ng2085 (2007).

115 Nan, H. *et al.* Genome-wide association study identifies novel alleles associated with risk of cutaneous basal cell carcinoma and squamous cell carcinoma. *Human molecular genetics* **20**, 3718-3724, doi:10.1093/hmg/ddr287 (2011).

116 Chahal, H. S. *et al.* Genome-wide association study identifies novel susceptibility loci for cutaneous squamous cell carcinoma. *Nat Commun* **7**, 12048, doi:10.1038/ncomms12048 (2016).

117 Antoniou, A. C. *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. *Nature genetics* **42**, 885-892, doi:10.1038/ng.669 (2010).

118 Studd, J. B. *et al.* Genetic and regulatory mechanism of susceptibility to high-hyperdiploid acute lymphoblastic leukaemia at 10p21.2. *Nature communications* **8**, 14616, doi:10.1038/ncomms14616 (2017).

119 Law, P. J. *et al.* Genome-wide association analysis of chronic lymphocytic leukaemia, Hodgkin lymphoma and multiple myeloma identifies pleiotropic risk loci. *Sci Rep* **7**, 41071, doi:10.1038/srep41071 (2017).

120 Speedy, H. E. *et al.* A genome-wide association study identifies multiple susceptibility loci for chronic lymphocytic leukemia. *Nature genetics* **46**, 56-60, doi:10.1038/ng.2843 (2014).

121 Rapley, E. A. *et al.* A genome-wide association study of testicular germ cell tumor. *Nature genetics* **41**, 807-810, doi:10.1038/ng.394 (2009).

122 Kanetsky, P. A. *et al.* Common variation in KITLG and at 5q31.3 predisposes to testicular germ cell cancer. *Nature genetics* **41**, 811-815, doi:10.1038/ng.393 (2009).

123 Turnbull, C. *et al.* Variants near DMRT1, TERT and ATF7IP are associated with testicular germ cell cancer. *Nature genetics* **42**, 604-607, doi:10.1038/ng.607 (2010).

124 Ruark, E. *et al.* Identification of nine new susceptibility loci for testicular cancer, including variants near DAZL and PRDM14. *Nature genetics* **45**, 686-689, doi:10.1038/ng.2635 (2013).

125 Thorgeirsson, T. E. *et al.* A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* **452**, 638-642, doi:10.1038/nature06846 (2008).

126 Freathy, R. M. *et al.* A common genetic variant in the 15q24 nicotinic acetylcholine receptor gene cluster (CHRNA5-CHRNA3-CHRNB4) is associated with a reduced ability of women to quit smoking in pregnancy. *Human molecular genetics* **18**, 2922-2927, doi:10.1093/hmg/ddp216 (2009).

127 Garcia-Closas, M. *et al.* Common genetic polymorphisms modify the effect of smoking on absolute risk of bladder cancer. *Cancer research* **73**, 2211-2220, doi:10.1158/0008-5472.CAN-12-2388 (2013).

128 Haiman, C. A. *et al.* Multiple regions within 8q24 independently affect risk for prostate cancer. *Nature genetics* **39**, 638-644, doi:10.1038/ng2015 (2007).

129 Woodage, T. *et al.* The APC I1307K allele and cancer risk in a community-based study of Ashkenazi Jews. *Nature genetics* **20**, 62-65 (1998).

130 Bush, W. S. & Moore, J. H. Chapter 11: Genome-wide association studies. *PLoS Comput Biol* **8**, e1002822, doi:10.1371/journal.pcbi.1002822 (2012).

131 Kilpivaara, O. *et al.* A germline JAK2 SNP is associated with predisposition to the development of JAK2(V617F)-positive myeloproliferative neoplasms. *Nature genetics* **41**, 455-459, doi:10.1038/ng.342 (2009).

132 Enciso-Mora, V. *et al.* Deciphering the 8q24.21 association for glioma. *Human molecular genetics* **22**, 2293-2302, doi:10.1093/hmg/ddt063 (2013).

133 Malmer, B. *et al.* GLIOGENE an International Consortium to Understand Familial Glioma. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* **16**, 1730-1734, doi:10.1158/1055-9965.EPI-07-0081 (2007).

134 Hemminki, K., Sundquist, J. & Bermejo, J. L. Associated cancers in parents and offspring of polycythaemia vera and myelofibrosis patients. *British journal of haematology* **147**, 526-530, doi:10.1111/j.1365-2141.2009.07874.x (2009).

135 Litchfield, K. *et al.* Rare disruptive mutations in ciliary function genes contribute to testicular cancer susceptibility. *Nat Commun* **7**, 13840, doi:10.1038/ncomms13840 (2016).

136 Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature genetics* **42**, 565-569, doi:10.1038/ng.608 (2010).

137 Sampson, J. N. *et al.* Analysis of Heritability and Shared Heritability Based on Genome-Wide Association Studies for 13 Cancer Types. *JNCI: Journal of the National Cancer Institute* **107**, djv279-djv279, doi:10.1093/jnci/djv279 (2015).

138 Frampton, M. J. *et al.* Implications of polygenic risk for personalised colorectal cancer screening. *Annals of oncology : official journal of the European Society for Medical Oncology* **27**, 429-434, doi:10.1093/annonc/mdv540 (2016).

139 Speed, D. *et al.* Reevaluation of SNP heritability in complex human traits. *Nature genetics* **49**, 986-992, doi:10.1038/ng.3865 (2017).

140 Sham, P. C. & Purcell, S. M. Statistical power and significance testing in large-scale genetic studies. *Nat Rev Genet* **15**, 335-346, doi:10.1038/nrg3706 (2014).

141 Pomerantz, M. M. *et al.* The 8q24 cancer risk variant rs6983267 shows long-range interaction with MYC in colorectal cancer. *Nature genetics* **41**, 882-884, doi:10.1038/ng.403 (2009).

142 Howie, B. N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics* **5**, e1000529, doi:10.1371/journal.pgen.1000529 (2009).

143 Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3* **1**, 457-470, doi:10.1534/g3.111.001198 (2011).

144 Stacey, S. N. *et al.* A germline variant in the TP53 polyadenylation signal confers cancer susceptibility. *Nature genetics* **43**, 1098-1103, doi:10.1038/ng.926 (2011).

145 Enciso-Mora, V. *et al.* Low penetrance susceptibility to glioma is caused by the TP53 variant rs78378222. *Br J Cancer* **108**, 2178-2185, doi:10.1038/bjc.2013.155 (2013).

146 Killedar, A. *et al.* A Common Cancer Risk-Associated Allele in the hTERT Locus Encodes a Dominant Negative Inhibitor of Telomerase. *PLoS genetics* **11**, e1005286, doi:10.1371/journal.pgen.1005286 (2015).

147 Mercer, T. R. *et al.* DNase I-hypersensitive exons colocalize with promoters and distal regulatory elements. *Nature genetics* **45**, 852-859, doi:10.1038/ng.2677 (2013).

148 Grundberg, E. *et al.* Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nature genetics* **44**, 1084-1089, doi:http://www.nature.com/ng/journal/v44/n10/abs/ng.2394.html#supplementary-information (2012).

149 Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nature genetics* **48**, 481-487, doi:10.1038/ng.3538 (2016).

150 Encode Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57-74, doi:10.1038/nature11247 (2012).

151 Mifsud, B. *et al.* Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature genetics* **47**, 598-606, doi:10.1038/ng.3286 (2015).

152 Dryden, N. H. *et al.* Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. *Genome research* **24**, 1854-1868, doi:10.1101/gr.175034.114 (2014).

153 Jager, R. *et al.* Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nature communications* **6**, 6178, doi:10.1038/ncomms7178 (2015).

154 Dixon, J. R. *et al.* Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376-380, doi:10.1038/nature11082 (2012).

155 de Souza, N. The ENCODE project. *Nat Methods* **9**, 1046 (2012).

156 Roadmap Epigenomics, C. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317-330, doi:10.1038/nature14248http://www.nature.com/nature/journal/v518/n7539/abs/nature14248.html#supplementary-information (2015).

157 Stunnenberg, H. G., International Human Epigenome, C. & Hirst, M. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell* **167**, 1145-1149, doi:10.1016/j.cell.2016.11.007 (2016).

158 Castro, M. A. *et al.* Regulators of genetic risk of breast cancer identified by integrative network analysis. *Nature genetics* **48**, 12-21, doi:10.1038/ng.3458 (2016).
159 Tuupanen, S. *et al.* The common colorectal cancer predisposition SNP rs6983267 at chromosome 8q24 confers potential to enhanced Wnt signaling. *Nature genetics* **41**, 885-890, doi:10.1038/ng.406 (2009).
160 Sur, I. K. *et al.* Mice lacking a Myc enhancer that includes human SNP rs6983267 are resistant to intestinal tumors. *Science* **338**, 1360-1363, doi:10.1126/science.1228606 (2012).
161 Walsh, K. M. *et al.* A Heritable Missense Polymorphism in CDKN2A Confers Strong Risk of Childhood Acute Lymphoblastic Leukemia and Is Preferentially Selected during Clonal Evolution. *Cancer research* **75**, 4884-4894, doi:10.1158/0008-5472.CAN-15-1105 (2015).
162 Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* **169**, 1177-1186, doi:10.1016/j.cell.2017.05.038 (2017).
163 Lawrenson, K. *et al.* Cis-eQTL analysis and functional validation of candidate susceptibility genes for high-grade serous ovarian cancer. *Nat Commun* **6**, 8234, doi:10.1038/ncomms9234 (2015).
164 Glodzik, D. *et al.* A somatic-mutational process recurrently duplicates germline susceptibility loci and tissue-specific super-enhancers in breast cancers. *Nature genetics* **49**, 341-348, doi:10.1038/ng.3771 (2017).
165 Ongen, H. *et al.* Putative cis-regulatory drivers in colorectal cancer. *Nature* **512**, 87-90, doi:10.1038/nature13602 (2014).
166 Li, Q. *et al.* Integrative eQTL-Based Analyses Reveal the Biology of Breast Cancer Risk Loci. *Cell* **152**, 633-641 (2013).
167 Nelson, M. R. *et al.* The support of human genetic evidence for approved drug indications. *Nature genetics* **47**, 856-860, doi:10.1038/ng.3314 (2015).
168 Zhang, J. *et al.* Use of Genome-Wide Association Studies for Cancer Research and Drug Repositioning. *PLoS ONE* **10**, e0116477, doi:10.1371/journal.pone.0116477 (2015).
169 Sanseau, P. *et al.* Use of genome-wide association studies for drug repositioning. *Nat Biotech* **30**, 317-320, doi:10.1038/nbt.2151

http://www.nature.com/nbt/journal/v30/n4/abs/nbt.2151.html#supplementary-information (2012).
170 Griffiths , C. E. M. *et al.* Comparison of Ustekinumab and Etanercept for Moderate-to-Severe Psoriasis. *New England Journal of Medicine* **362**, 118-128, doi:10.1056/NEJMoa0810652 (2010).
171 Nair, R. P. *et al.* Genome-wide scan reveals association of psoriasis with IL-23 and NF-kappaB pathways. *Nature genetics* **41**, 199-204, doi:10.1038/ng.311 (2009).
172 Di Meglio, P. *et al.* The IL23R R381Q Gene Variant Protects against Immune-Mediated Diseases by Impairing IL-23-Induced Th17 Effector Response in Humans. *PLOS ONE* **6**, e17160, doi:10.1371/journal.pone.0017160 (2011).
173 Roberts, A. W. *et al.* Targeting BCL2 with Venetoclax in Relapsed Chronic Lymphocytic Leukemia. *New England Journal of Medicine* **374**, 311-322, doi:10.1056/NEJMoa1513257 (2016).
174 Babina, I. S. & Turner, N. C. Advances and challenges in targeting FGFR signalling in cancer. *Nat Rev Cancer*, doi:10.1038/nrc.2017.8 (2017).
175 Pujana, M. A. Integrating germline and somatic data towards a personalized cancer medicine. *Trends in Molecular Medicine* **20**, 413-415, doi:10.1016/j.molmed.2014.05.004 (2014).
176 Pashayan, N. *et al.* Polygenic susceptibility to prostate and breast cancer: implications for personalised screening. *Br J Cancer* **104**, 1656-1663, doi:10.1038/bjc.2011.118 (2011).
177 Seibert, T. M. *et al.* A genetic risk score to guide age-specific, personalized prostate cancer screening. *bioRxiv*, doi:10.1101/089383 (2016).
178 Lecarpentier, J. *et al.* Prediction of Breast and Prostate Cancer Risks in Male BRCA1 and BRCA2 Mutation Carriers Using Polygenic Risk Scores. *Journal of Clinical Oncology* **0**, JCO.2016.2069.4935, doi:10.1200/jco.2016.69.4935 (2017).
179 Kuchenbaecker, K. B. *et al.* Evaluation of Polygenic Risk Scores for Breast and Ovarian Cancer Risk Prediction in BRCA1 and BRCA2 Mutation Carriers. *Journal of the National Cancer Institute* **109**, doi:10.1093/jnci/djw302 (2017).
180 Rodriguez-Broadbent, H. *et al.* Mendelian randomisation implicates hyperlipidaemia as a risk factor for colorectal cancer. *Int J Cancer*, doi:10.1002/ijc.30709 (2017).

181 Hedditch, E. L. *et al.* ABCA Transporter Gene Expression and Poor Outcome in Epithelial Ovarian Cancer. *JNCI: Journal of the National Cancer Institute* **106**, dju149-dju149, doi:10.1093/jnci/dju149 (2014).
182 Perez-Andreu, V. *et al.* Inherited GATA3 variants are associated with Ph-like childhood acute lymphoblastic leukemia and risk of relapse. *Nature genetics* **45**, 1494-1498, doi:10.1038/ng.2803 (2013).
183 Wu, C. *et al.* Genome-wide association study identifies common variants in SLC39A6 associated with length of survival in esophageal squamous-cell carcinoma. *Nature genetics* **45**, 632-638, doi:10.1038/ng.2638

http://www.nature.com/ng/journal/v45/n6/abs/ng.2638.html#supplementary-information (2013).
184 Johnson, D. C. *et al.* Genome-wide association study identifies variation at 6q25.1 associated with survival in multiple myeloma. *Nature communications* **7**, 10290, doi:10.1038/ncomms10290

https://www.nature.com/articles/ncomms10290#supplementary-information (2016).
185 Berndt, S. I. *et al.* Two susceptibility loci identified for prostate cancer aggressiveness. *Nature communications* **6**, 6889, doi:10.1038/ncomms7889 (2015).
186 Johnson, N. *et al.* Cytochrome P450 allele CYP3A7*1C associates with adverse outcomes in chronic lymphocytic leukemia, breast and lung cancer. *Cancer research* **76**, 1485-1493, doi:10.1158/0008-5472.CAN-15-1410 (2016).
187 Aminkeng, F. *et al.* A coding variant in RARG confers susceptibility to anthracycline-induced cardiotoxicity in childhood cancer. *Nature genetics* **47**, 1079-1084, doi:10.1038/ng.3374 (2015).
188 Fachal, L. *et al.* A three-stage genome-wide association study identifies a susceptibility locus for late radiotherapy toxicity at 2q24.1. *Nature genetics* **46**, 891-894, doi:10.1038/ng.3020

http://www.nature.com/ng/journal/v46/n8/abs/ng.3020.html#supplementary-information (2014).
189 Canver, M. C. *et al.* Variant-aware saturating mutagenesis using multiple Cas9 nucleases identifies regulatory elements at trait-associated loci. *Nature genetics*, doi:10.1038/ng.3793 (2017).
190 Canver, M. C. *et al.* BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature* **527**, 192-197, doi:10.1038/nature15521 (2015).
191 Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661-678, doi:10.1038/nature05911 (2007).
192 Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516-1517 (1996).
193 Davey Smith, G. & Ebrahim, S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease?*. *International Journal of Epidemiology* **32**, 1-22, doi:10.1093/ije/dyg070 (2003).
194 Jarvis, D. *et al.* Mendelian randomisation analysis strongly implicates adiposity with risk of developing colorectal cancer. *British journal of cancer* **115**, 266-272, doi:10.1038/bjc.2016.188 (2016).
195 Interleukin-6 Receptor Mendelian Randomisation Analysis (IL6R MR) Consortium *et al.* The interleukin-6 receptor as a target for prevention of coronary heart disease: a mendelian randomisation analysis. *Lancet* **379**, 1214-1224, doi:10.1016/S0140-6736(12)60110-X (2012).

Author Biographies

Amit Sud

Amit Sud is a clinical research fellow in the Molecular and Population Genetics team led by Dr. Richard Houlston at the Institute of Cancer Research, London, United Kingdom. He is interested in the epidemiology, genetic susceptibility and biology of Hodgkin lymphoma and other B-cell malignancies.

Ben Kinnersley

Ben Kinnersley is a postdoctoral researcher working in the Molecular and Population Genetics team led by Dr. Richard Houlston at the Institute of Cancer Research, London, United Kingdom. His primary research interest is genetic susceptibility to glioma, specifically utilising findings from genome-wide association studies to better understand the biology of glioma tumors.

Richard Houlston

Richard Houlston is a professor and Group Leader at The Institute of Cancer Research whose work is focused on understanding inherited susceptibility to cancer. He has discovered the high-risk susceptibility gene for hereditary leiomyomatosis and renal cell cancer. His recent work has centred on the use of genome-wide association studies to identify common risk variants for cancer. He has successfully used this strategy to identify susceptibility genes for colorectal, lung and renal cancer, acute lymphoblastic leukaemia, chronic lymphocytic leukaemia, Hodgkin lymphoma, meningioma and glioma.

Table of contents summary

Genome-wide association studies (GWAS) uncover the impact of genetic variation on the risk of many common cancers. This review discusses current insights and how understanding the biological basis of these associations is required to maximise the clinical benefit of GWAS.

Subject categories

Biological sciences / Genetics / Cancer genomics
[URI /631/208/69]

Biological sciences / Genetics / Cancer genetics
[URI /631/208/68]

Biological sciences / Cancer / Cancer epidemiology
[URI /631/67/2324]

Biological sciences / Cancer / Cancer prevention
[URI /631/67/2195]

Health sciences / Risk factors
[URI /692/499]