APPLICATIONS OF NEXT-GENERATION SEQUENCING

Role of non-coding sequence variants in cancer

Ekta Khurana[1–4], Yao Fu[5], Dimple Chakravarty[2,6], Francesca Demichelis[2,3,7], Mark A. Rubin[1,2,6] and Mark Gerstein[8–10]

Abstract | Patients with cancer carry somatic sequence variants in their tumour in addition to the germline variants in their inherited genome. Although variants in protein-coding regions have received the most attention, numerous studies have noted the importance of non-coding variants in cancer. Moreover, the overwhelming majority of variants, both somatic and germline, occur in non-coding portions of the genome. We review the current understanding of non-coding variants in cancer, including the great diversity of the mutation types — from single nucleotide variants to large genomic rearrangements — and the wide range of mechanisms by which they affect gene expression to promote tumorigenesis, such as disrupting transcription factor-binding sites or functions of non-coding RNAs. We highlight specific case studies of somatic and germline variants, and discuss how non-coding variants can be interpreted on a large-scale through computational and experimental methods.

Exome sequencing
Sequencing the protein-coding portion of the genome using target-enrichment and high-throughput sequencing technology.

Driver mutations
Sequence variants that confer growth advantage to tumour cells.

Passenger mutations
Sequence variants that do not contribute to cancer growth.

Germline variants
Heritable variants that are transmitted to offspring. These variants are constitutional (that is, present in all cells of the body).

Correspondence to E.K., M.A.R. and M.G.
ekk2003@med.cornell.edu; rubinma@med.cornell.edu; mark.gerstein@yale.edu

See Author addresses box for address list.

doi:10.1038/nrg.2015.17
Published online 19 Jan 2016

Sequencing of thousands of tumour samples has revealed the landscape of somatic mutations in protein-coding genes[1]. Most previous studies of cancer genomes have used exome sequencing rather than whole-genome sequencing (WGS) owing to lower costs and a focus on the regions that are considered to be most functionally relevant. However, the decreasing costs of sequencing have enabled WGS of thousands of tumours by individual research groups and efforts such as TCGA (The Cancer Genome Atlas) and ICGC (International Cancer Genome Consortium). One of the most important benefits of WGS is the identification of variants in non-coding regions of the genome. Indeed, most of the variants obtained from WGS of tumour genomes lie there (FIG. 1). There is an increased realization of the importance of non-coding variants in cancer, and an ongoing collaboration between TCGA and ICGC, called the Pan-Cancer Analysis of Whole Genomes (PCAWG) aims to analyse non-coding variants in ~2,500 tumour and matched normal whole genomes. One of the biggest challenges of analysing non-coding or coding variants is to identify driver mutations and to distinguish them from passenger mutations.

The link between inherited germline variants and complex disorders, including cancer, has been probed previously by numerous genome-wide association studies (GWASs) using DNA from non-disease cells (usually blood). These studies implicated various protein-coding genes in tumorigenesis (such as DNA repair and cell-cycle control genes)[2]. Importantly, these studies also revealed that many loci that are associated with cancer susceptibility lie in non-coding regions of the genome[3,4].

In this Review, we discuss our current understanding of the role of non-coding sequence variants in cancer development and growth. We first describe distinctions in the nature of somatic versus germline sequence variants and then provide brief overviews of the various non-coding annotations. We then discuss diverse molecular mechanisms by which somatic and germline variants are known to lead to tumorigenesis, including their functional interplay. Finally, to interpret non-coding variants linked to cancer, we describe how bioinformatics and experimental approaches can be used to prioritize them and validate their functional relevance. Throughout our Review, we focus on the effects of DNA sequence variants in non-coding regions. However, we note that besides sequence alterations, other changes can occur in non-coding regions of cancer genomes, such as epigenetic changes at regulatory elements and transcriptional dysregulation of non-coding RNAs (ncRNAs); for further information on these phenomena, the reader is referred to REFS 5–9.

Genomic sequence variants

Most of the genome is non-coding and most DNA sequence variants occur in non-coding regions. Hence, the general properties of sequence variants are applicable to non-coding variants. They range from single nucleotide variants (SNVs) to small insertions and deletions of less than 50 base pairs in length (indels), to

© 2016 Macmillan Publishers Limited. All rights reserved

Author addresses

[1]Meyer Cancer Center, Weill Cornell Medical College, New York, New York 10065, USA.
[2]Institute for Precision Medicine, Weill Cornell Medical College, New York, New York 10065, USA.
[3]Institute for Computational Biomedicine, Weill Cornell Medical College, New York, New York 10021, USA.
[4]Department of Physiology and Biophysics, Weill Cornell Medical College, New York, New York 10065, USA.
[5]Bina Technologies, Roche Sequencing, Redwood City, California 94065, USA.
[6]Department of Pathology and Laboratory Medicine, Weill Cornell Medical College, New York, New York 10065, USA.
[7]Centre for Integrative Biology, University of Trento, 38123 Trento, Italy.
[8]Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut 06520, USA.
[9]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520, USA.
[10]Department of Computer Science, Yale University, New Haven, Connecticut 06520, USA.

larger structural variants. Structural variants can be copy number variants (CNVs; such as deletions and duplications) or copy-number neutral (such as inversions and translocations). An average human genome contains roughly 4 million germline sequence variants relative to the reference genome[10], whereas a tumour genome typically contains thousands of variants relative to the same individual's germline DNA[11] (FIG. 1). Although somatic variants are known to be present in healthy tissues[12,13], they are relatively rare compared with the number of somatic variants in tumours. Thus, most studies of somatic variants have focused on tumours, and in this article we refer to somatic variants as the ones specific to tumour cells. Somatic mutation frequency varies considerably across different cancer types[11,14], as do the relative proportions of non-coding and coding variants (FIG. 1).

A discussion of germline variants is important, as cancer is known to have a familial component, and several non-coding variants are known to play a part in cancer development. The two-hit hypothesis is a widely known mechanism by which germline variants can promote oncogenesis (discussed below). Rare, non-coding germline variants with high penetrance may be directly responsible for tumorigenesis (for example, as observed in familial cancer cases[15]), whereas variants with low penetrance may modulate the effects of somatic variants[2]. The number of germline variants relative to the reference human genome per individual differs by ethnicity, and individuals from different populations show varied profiles of rare and common variants[10]. With the exception of paediatric cancers, most cancer cases occur at an older age. Thus, the germline variants associated with increased cancer susceptibility for non-paediatric cancers do not typically have a fitness effect at reproductive age, which is perhaps the reason for the continued prevalence of such variants in the population. In addition, germline variants often show linkage disequilibrium (LD) — that is, association of alleles at multiple loci. LD between variants, especially the common ones with high allele frequencies, presents a unique set of challenges in disentangling the causal disease variants from the ones that they are linked with. Owing to their low allele frequencies, rare variants do not exhibit strong LD with common or other rare variants[16].

Germline and somatic variants exhibit many distinct features. First, although both of them comprise SNVs, indels and structural variants, a much higher fraction of somatic variants consists of structural variants, including large genomic rearrangements. For example, fusion events between distant genes have been observed in many cancer types but are rare in germline sequences. Similarly, complex genomic rearrangements, including chromoplexy[17] and chromothripsis[18], are known to occur in cancer cells. Chromosomal aneuploidy, whereby an entire chromosome may be lost or gained, is also often observed in cancer[19]. Second, unlike somatic variants, germline variants occur in all tissues of the body. Thus, they must be compatible with organismal viability, which is probably why they are generally not as disruptive as the major chromosomal rearrangements or as aneuploidy observed in tumour cells. However, the functional effect of germline variants might not be manifested in all tissues, for example if they occur in regions of closed chromatin or if they disrupt a binding site of a transcription factor (TF) that is not expressed in the tissue. Third, somatic sequence variants may not be shared by all cells in the tumour tissue. Such tumour heterogeneity makes interpretation of somatic variants more complex. Fourth, various phenomena, such as kataegis[20] and genome-wide mutational signatures[11], are characteristic only of somatic variants. In particular, more than 20 mutational signatures have been identified in 30 different cancer types. Some signatures (such as the one associated with the APOBEC family of cytidine deaminases) are common across many different cancer types, whereas other signatures (such as the one observed in malignant melanoma and linked with damage caused by ultraviolet light) are specific to particular tumour classes[11]. Finally, unlike germline variants, somatic variants are not inherited. Thus, they are not subject to the recombinatorial effects of meiosis and hence do not show LD (discussed above).

Non-coding element annotations

To understand the effect of sequence variants in non-coding regions, we first need to examine the role of various non-coding functional elements. Below, we discuss these elements and the approaches used to annotate them in the genome.

Non-coding elements can have diverse roles in the regulation of protein-coding genes. Broadly speaking, they consist of *cis*-regulatory regions[21] and ncRNAs. These elements are generally identified by functional genomics approaches or sequence conservation and often display cell- and tissue-type specificity (FIG. 2).

Cis-regulatory regions include promoters and distal elements (enhancers, silencers and insulators), which regulate gene expression following binding by TFs. TFs bind to specific DNA sequences (motifs) within their larger regions of occupancy (peaks), which can be identified using chromatin immunoprecipitation followed by sequencing (ChIP-seq) assays. They bind DNA in regions of open (non-nucleosomal) chromatin, which can be identified using DNase I hypersensitivity assays, and DNase I footprinting can also help to identify

Genome-wide association studies
(GWASs). Studies that interrogate multiple common genetic variants along the genome in large cohorts of individuals to evaluate whether any variant is associated with a specific trait.

Single nucleotide variants
DNA sequence changes at single nucleotides.

Somatic variants
Variants that are not inherited from a parent and are not transmitted to offspring.

Penetrance
The proportion of individuals carrying an allele (or a genotype) that also express the trait (phenotype) associated with it.

Chromoplexy
(From the Greek *pleko*, meaning to weave, or to braid). A class of complex somatic DNA rearrangements whereby abundant DNA deletions and intra- and inter-chromosomal translocations that have originated in an interdependent way occur within a single cell cycle.

Chromothripsis
(From the Greek *thripsis*, meaning shattering into pieces). A clustered chromosomal rearrangement in confined genomic regions that results from a single catastrophic event, usually limited to one chromosome.

© 2016 Macmillan Publishers Limited. All rights reserved

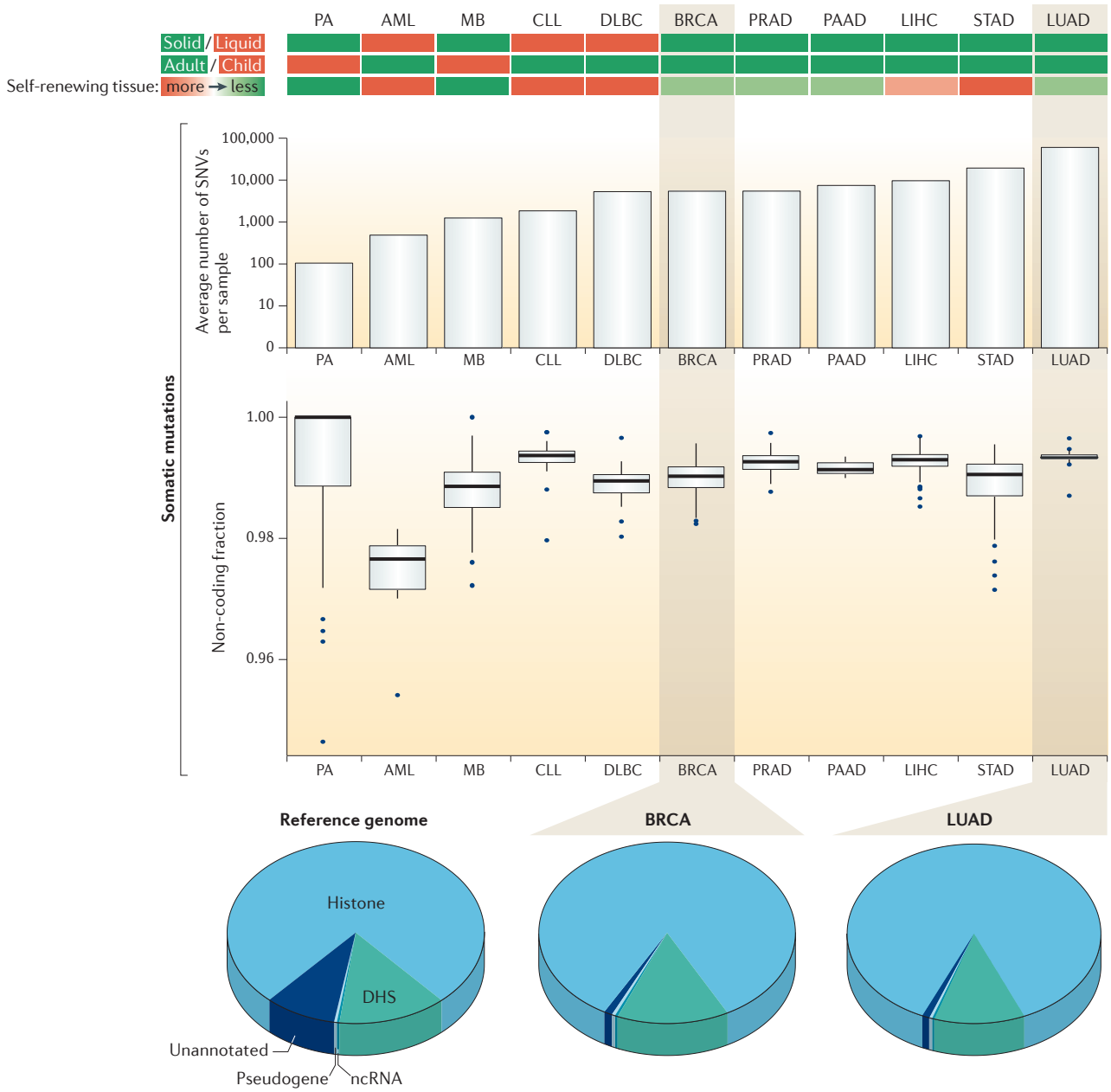


Figure 1 | **Somatic mutations in various cancer types.** Somatic tumour mutations were obtained from whole genome studies in REFS 11,17,137, and cancer types are ordered from left to right based on average number of single nucleotide variants (SNVs) per sample. Bar plots denote the average number of SNVs. Box plots show the fraction of non-coding SNVs (based on GENCODE 19). Spearman correlation between total number of mutations and non-coding fraction = 0.32, $P = 2.20 \times 10^{-15}$. Note that this correlation is when we exclude pilocytic astrocytoma (PA), which shows great variability in the number of mutations and has been hypothesized to be a single-pathway disease. This positive correlation could be due to the higher number of passenger mutations in tumours with high total numbers of mutations, and most non-coding mutations corresponding to passenger events. Reference genome pie-chart shows the coverage of different non-coding categories in the reference human genome. BRCA (breast cancer) and LUAD (lung adenocarcinoma) pie-charts show mean SNVs per sample in each category for these two cancer types. As one region or variant can overlap multiple categories, the following hierarchy is used to categorize non-coding variants: non-coding RNA (ncRNA), pseudogene > DNase1 hypersensitive site (DHS) > histone > unannotated. ncRNAs and pseudogenes are from GENCODE 19, DHS from 125 cell lines from REF. 138 and histone modifications from the Encyclopedia of DNA Elements Data Coordinating Center (ENCODE DCC; version March 2012). The fraction of non-coding SNVs that is not annotated is lower in BRCA and LUAD genomes relative to the fraction of non-coding genome that is not annotated in the reference genome. This could be a real effect reflecting enrichment of SNVs in functional annotations or bias due to difficulties in identifying SNVs in complex repetitive regions that are also hard to annotate. AML, acute myeloid leukaemia; CLL, chronic lymphocytic leukaemia; DLBC, diffuse large B cell lymphoma; LIHC, liver hepatocellular carcinoma; MB, medulloblastoma; PAAD, pancreatic adenocarcinoma; PRAD, prostate adenocarcinoma; STAD, stomach adenocarcinoma.

© 2016 Macmillan Publishers Limited. All rights reserved

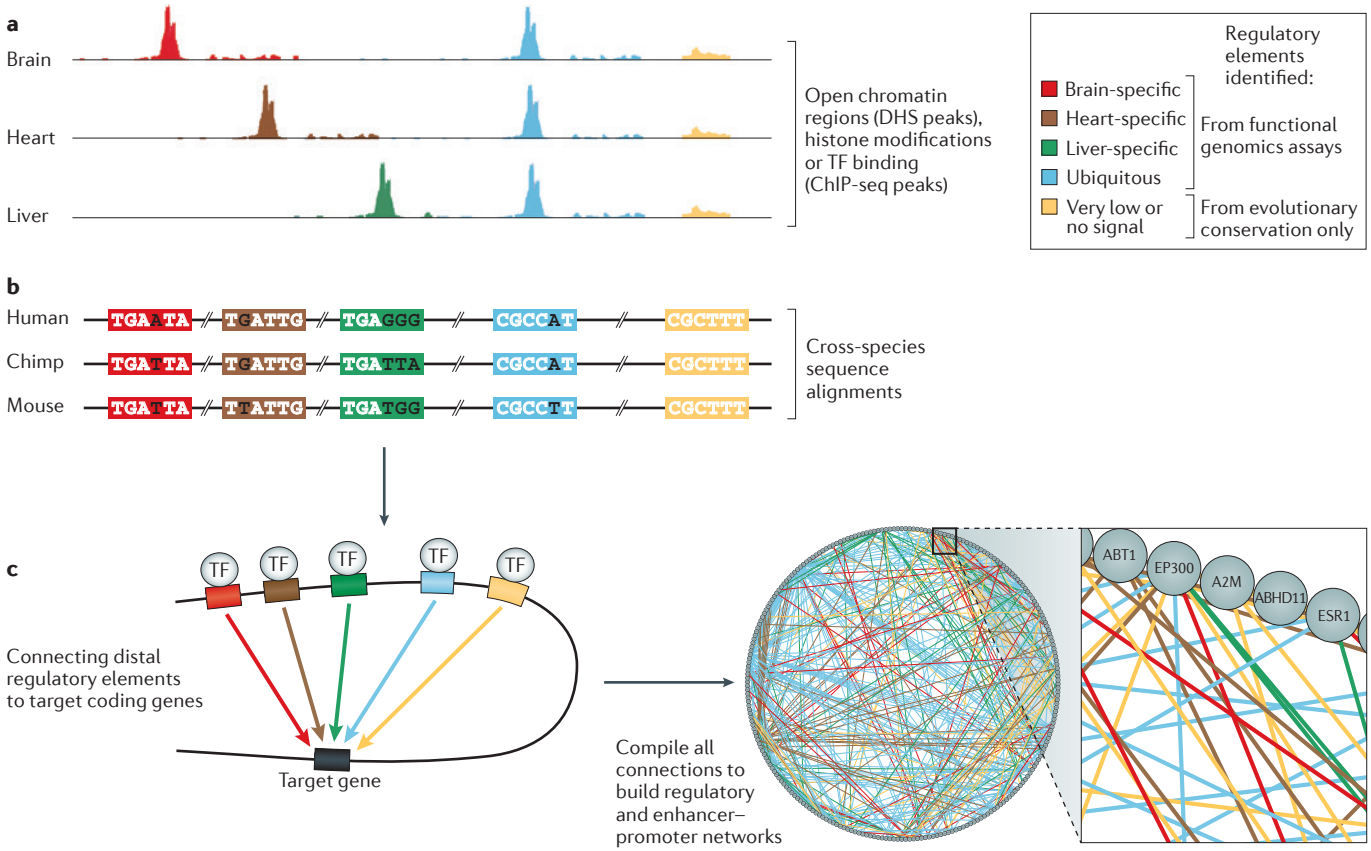


Figure 2 | **Identification of *cis*-regulatory elements using functional genomics assays and evolutionary conservation.** **a** | Regulatory elements can show differential activity across tissues and can be identified by various functional genomics assays, including DNase I hypersensitive sites (DHSs, for regions of open chromatin) or ChIP-seq peaks (for histone modifications and transcription factor (TF) binding). **b** | Some elements (yellow) may not show activity in limited functional genomics experiments and are identified by evolutionary conservation only. Bold white text indicates conservation. **c** | Distal regulatory elements can be connected to target coding genes using various approaches (including chromosome conformation capture (3C) assays, correlation of enhancer activity and gene expression, and expression quantitative trait loci (eQTLs)). All the connections can then be compiled into regulatory (TF to target gene) and enhancer–promoter networks. A regulatory network is shown with TFs and target genes as nodes and interactions between them as edges. The interactions can be tissue specific (red, brown and green edges) or ubiquitous (blue and yellow edges).

Kataegis
(From the Greek *kataigis*, meaning thunder).
A phenomenon that is characterized by large clusters of mutations (hypermutation) in the genome of cancer cells. An APOBEC family enzyme might be responsible for the kataegis process.

Cis-regulatory regions
Regions that regulate the expression of genes on the same DNA molecule. These include promoters, enhancers, silencers, insulators and untranslated regions.

Enhancers
Distal *cis*-regulatory regions bound by transcription factors that activate genes by helping the recruitment of RNA polymerase to the promoters.

high-resolution TF-binding sites within the larger DNase I hypersensitive sites (DHSs)[22,23]. Furthermore, DNA methylation and other histone modifications can modulate TF accessibility to DNA. Indeed, several histone marks are associated with specific putative functions: for example, trimethylated Lys4 of histone 3 (H3K4me3) with active promoters, acetylation of Lys27 of histone 3 (H3K27ac) with active promoters and enhancers, and H3K27me3 with repressed regions[24]. Such sites of histone modifications can also be identified using ChIP-seq assays. Although most sequence-specific TFs and some chromatin marks lead to highly localized ChIP-seq signals (hundreds of nucleotides), other marks (such as H3K9me3 and H3K36me3) are associated with large genomic domains that can cover up to a few megabases. Thus, overall, epigenetic changes can alter TF accessibility in different cellular states and may be thought of as changing the activity of regulatory elements, resulting in cell-type specificity of their associated genes. This is similar to the way one thinks of differential activity of a universal gene set in different cell types.

Distal regulatory elements may regulate gene expression by interacting with promoters in the three-dimensional (3D) structure of the genome. Linking the distal elements to their target protein-coding genes in the 3D chromatin structure is crucial to understand the effects of sequence variants in them. Multiple approaches have been used to link *cis*-regulatory regions to their target genes. For example, chromosome conformation capture (3C) technology has demonstrated that regulatory sequences can control transcription by looping to and physically contacting target coding genes that are located tens or hundreds of kilobases away[25,26]. The 3C technology probes one-versus-one contacts in the 3D space of the genome. Further variations of the 3C technology have since been developed that probe one-versus-all (4C), many-versus-many (5C) and all-versus-all (HiC) contacts[26]. Other approaches that have been used to link distal regulatory elements to their target genes include correlation of histone marks at enhancer regions and target gene expression across multiple cell lines[27] and links between expression quantitative trait loci (eQTLs)

© 2016 Macmillan Publishers Limited. All rights reserved

and associated genes (discussed below)[28]. The resulting linkages can then be studied as a comprehensive network[29] (FIG. 2).

Several large-scale efforts such as ENCODE[24] and the National Institutes of Health (NIH) Roadmap Epigenomics Consortium[30,31] were launched to create a comprehensive map of regulatory regions. Besides TF-binding sites, many other modes of *cis*-regulation exist in the genome. The 3′ untranslated regions (3′ UTRs) of mRNAs contain binding sites for microRNAs (miRNAs; discussed below) and also have a role in mRNA stability and translation[32]. 5′ UTRs contain regulatory elements for both transcription and translation stages, such as the 5′-cap structure, translation initiation motifs and internal ribosome entry sites[33]. High-throughput RNA sequencing (RNA-seq) yields functional insights into the genome. Correlation of gene expression with the occurrence of sequence variants helps to identify eQTLs in non-coding regions, which in turn point to the putative functional role of the region[34]. Gene expression studies across various tissues can reveal regulatory regions that are associated with tissue-specific expression[35]. The Genotype–Tissue Expression (GTEx) project has provided an atlas of gene expression across multiple tissues and many individuals, which can be used to identify potential regulatory regions[28,35].

RNA-seq also reveals non-coding transcripts, which can be further confirmed to not have protein-coding ability by the absence of open reading frames or proteomic analysis. Certain histone modifications associated with active transcription can also indicate ncRNA activity, such as H3K4me3 for promoters of ncRNA-encoding loci and H3K36me3 for actively transcribed ncRNAs. ncRNAs can be divided into several categories, such as tRNAs, rRNAs, small nucleolar RNAs (snoRNAs), small nuclear (snRNAs), miRNAs and long ncRNAs (lncRNAs; which are >200 nucleotides)[36]. All these RNAs act through different mechanisms to modulate gene expression, and many are known to have an important role in cancer biology, in particular miRNAs and lncRNAs[7]. miRNAs inhibit target gene expression by binding to the 3′ UTRs of their mRNAs and causing mRNA degradation or repression of translation. The precise mechanisms of action of many lncRNAs remain unclear. That said, a number of lncRNAs have been shown to act as molecular scaffolds that bind proteins, DNA or other RNA molecules, and are able to modulate gene expression[37–41].

Transcribed pseudogenes are a particular type of ncRNA that bear a clear resemblance to functioning protein-coding genes. Pseudogenes are copies of coding genes that have lost their ability to code for proteins owing to disabling mutations, such as premature stop codons and frameshift insertions or deletions. They can be divided into duplicated and processed, based on their formation from duplication or retrotransposition of the parent gene, respectively[42,43]. Processed pseudogenes typically lack the promoter sequence and intronic structure and contain a 3′-poly(A) tail. Although pseudogenes cannot code for proteins, they can be transcribed and can regulate the expression of their parent genes, for example by generating endo-siRNAs (endogenous small interfering RNAs) and participating in the RNA interference pathway[44,45] or by acting as molecular sponges, competing with parent gene mRNAs for miRNA binding[7].

Evolutionary conservation of genomic sequence across multiple species is also used to annotate non-coding regions[46,47]. Comparative analysis of human with mouse, rat and dog genomes showed that at least ~5% of the genome is conserved[48–50]. Because only ~1.5% of the genome codes for proteins, the remaining ~3.5% conserved regions likely contain regulatory elements and ncRNAs. Furthermore, 481 segments that are at least 200 base pairs long are 100% conserved between mice, rats and humans. These regions, termed ultra-conserved elements, cover ~107 kilobases of the genome and also exhibit high conservation among vertebrates[51]. Of these 481 ultra-conserved elements, 370 do not overlap protein-coding exons. Analysis of the sequence variants in these non-coding, ultra-conserved elements is important because they have been shown to have a role in cancer biology. Some of them are transcribed and act as ncRNAs that exhibit aberrant expression in tumorigenesis and indeed can be used to differentiate cancer types[52,53].

Besides selection constraint across multiple species, non-coding elements also exhibit conservation among humans. Negative selection within the human population can be estimated using various metrics, such as enrichment of rare alleles and reduced density of single nucleotide polymorphisms (SNPs)[24,54,55]. These metrics can be especially important to identify elements that show human-specific conservation and function[56]. By estimating conservation in hundreds of functional non-coding categories, sensitive and ultra-sensitive elements were identified[54]. These elements show strong depletion of common polymorphisms and enrichment of known disease-causing mutations. It has also been shown that negative selection among humans can be used to identify candidate cancer driver mutations[54].

The functional activity of conserved non-coding elements can be tested using various assays. For example, hundreds of evolutionarily conserved regions (including ultra-conserved elements) have been tested for their *in vivo* activity as enhancers and the results are available from the VISTA database[57].

We summarize the various sources of non-coding element annotations in TABLE 1. The web links for file downloads are provided at the end of the article.

Roles for somatic variants in cancer

In this section, we discuss some known cases of somatic variants and their likely role in tumorigenesis. Most of the examples noted below were identified in focused studies of cancer genes and their regulatory regions, and only a few were identified through WGS of tumour genomes. Indeed, few studies have tried to explore the role of non-coding somatic variants in cancer development in a systematic manner through large-scale analysis of tumour whole genomes[54,58–63]. In FIG. 3, we show study design strategies to probe non-coding mutations in tumour genomes. We note that variant calling in non-coding regions from next-generation sequencing of cancer whole-genomes can be especially challenging.

Silencers
Distal *cis*-regulatory regions bound by transcription factors that repress gene expression by preventing RNA polymerase from binding to the gene promoter.

Insulators
Regions that block the interaction between enhancers and promoters.

DNase I footprinting
A method to detect the exact binding sites of DNA-binding proteins based on the fact that a protein bound to DNA protects it from cleavage by DNase I.

Chromosome conformation capture
(3C). A biochemical method whereby the three-dimensional organization of chromatin in living cells is fixed and analysed.

Expression quantitative trait loci
(eQTLs). Loci in which DNA sequence variants are related with expression levels of mRNAs.

Endo-siRNAs
Endogenously produced small interfering RNAs that regulate gene expression by binding and cleaving mRNA targets or mediating heterochromatin formation.

Negative selection
Selective pressure that results in the removal of deleterious alleles.

Single nucleotide polymorphisms
(SNPs). Single nucleotide variants that show variability in the human population. As used in the context of this Review, they may be common (with high allele frequency) or rare (with low allele frequency).

© 2016 Macmillan Publishers Limited. All rights reserved

Table 1 | **Non-coding annotations**

Annotation	Resource
Transcription start sites	GENCODE FANTOM
Transcription factor-binding sites and motifs	ENCODE Roadmap epigenomics JASPAR Transfac CIS-BP
DHS (regions of open chromatin)	ENCODE Roadmap epigenomics
Histone marks	ENCODE Roadmap epigenomics
Integrated chromatin states (including enhancers)	ENCODE Roadmap epigenomics (derived from methods such as ChromHMM and Segway) FANTOM
Enhancer–promoter linkages	ENCODE Roadmap epigenomics FunSeq2
Transcription factor–target gene linkages	ENCODE (derived from ChIP-seq) ENCODE (derived from DHS) Roadmap epigenomics
Topologically associated domains from HiC	ENCODE
Various types of ncRNAs	GENCODE miRBase snoRNABase GtRNAdb MiTranscriptome

DHS, DNase I hypersensitivity site; ncRNA, non-coding RNA.

This is partly because roughly half of the human genome consists of repetitive DNA, which presents technical challenges in alignment and variant calling using short reads, especially for the identification of structural variants[64,65]. Because cancer genomes contain a higher fraction of structural variants than germline genomes, variant detection becomes even more challenging. In addition, the depth of coverage for cancer WGS needs to be more than typically used for germline WGS to account for decreased purity and increased ploidy and heterogeneity of tumour samples[66].

As more whole genomes are analysed, we are likely to see new types of mutational effects; for example, most known point mutations related to oncogenesis lead to gain of TF motifs, and we expect to see examples of mutations leading to loss of motifs. Below, we discuss some case studies in which non-coding variants disrupt regulatory elements or create new ones (FIG. 4).

Gain of TF-binding sites. *TERT* (telomerase reverse transcriptase) encodes the catalytic subunit of the enzyme telomerase. Telomerase lengthens telomeres, allowing cells to escape apoptosis and become cancerous. *TERT* expression is generally repressed in normal somatic cells, but can be overexpressed in cancer[67]. In the past few years, numerous studies have reported recurrent mutations in the promoter of *TERT* in many different cancer types[15,67–69]. These mutations create binding motifs for the ETS family of TFs, including TCFs (ternary complex factors), leading to their binding to the *TERT* promoter and subsequent upregulation of gene expression (FIG. 4B). Tumours in tissues with relatively low rates of self-renewal (including melanomas, urothelial carcinomas and medulloblastomas) tend to exhibit higher frequencies of *TERT* promoter mutations[69]. The high occurrence of these mutations points to their role as driver as opposed to passenger mutations.

Gain of TF-binding sites has also been observed for enhancers, which constitute important distal *cis*-regulatory elements and play a major part in gene transcription. In particular, super-enhancers are regions that recruit many TFs and drive the expression of genes that define cell identity[70]. Recently, it was reported that somatic mutations create MYB-binding motifs in T cell acute lymphoblastic leukaemia (T-ALL), forming a super-enhancer upstream of *TAL1* (T cell acute lymphocytic leukaemia 1), which results in its overexpression[71]. *TAL1* is an oncogene that codes for a basic helix–loop–helix TF, which has an important role in erythroid cell differentiation and is implicated in haematopoietic malignancies.

Fusion events due to genomic rearrangements. Genomic rearrangements can lead to fusion of active regulatory elements with oncogenes. For example, the 5′ UTR of *TMPRSS2* is frequently fused with ETS family genes (for example, *ERG* and *ETV1*) in prostate cancer[72]. This leads to *ERG* overexpression, which disrupts androgen receptor (AR) signalling by inhibiting the expression of AR and its target genes and inducing repressive epigenetic programmes[73]. AR has an important role in lineage-specific differentiation of the prostate, and its misregulation is linked with cellular dedifferentiation and malignant transformation. Furthermore, genomic rearrangements are also significantly associated with AR-binding sites in a subset of prostate cancers, indicating that AR binding may drive the formation of structural rearrangements[74,75].

In another example, it was reported that somatic structural variants juxtapose coding sequences of *GFI1* or *GFI1B* (growth factor independent 1 family oncogenes) proximal to active enhancers (an event known as 'enhancer hijacking') in medulloblastoma[76] (FIG. 4C). Activated *GFI1* and *GFI1B*, combined with *MYC* activation, can then promote medulloblastoma pathogenesis. Similarly, in T-ALL the *TAL1*-coding sequence is fused with the promoter of the ubiquitously expressed *SIL* (SCL-interrupting locus) gene, leading to overexpression of *TAL1* (REF. 77), a rearrangement found in 25% of cases of human T-ALL. Thus, *TAL1* may be overexpressed in T-ALL either because it is fused to the promoter of another gene or because of the gain of TF-binding sites caused by point mutations (discussed above). These examples of fusion events add to the list of known examples in lymphoid malignancies, whereby genomic rearrangements bring oncogenes (including *MYC* and *BCL2* (B cell lymphoma 2)) adjacent to active promoters or enhancers[78].

Oncogene
A gene that is often upregulated in cancer and can lead to or promote cancer growth.

© 2016 Macmillan Publishers Limited. All rights reserved

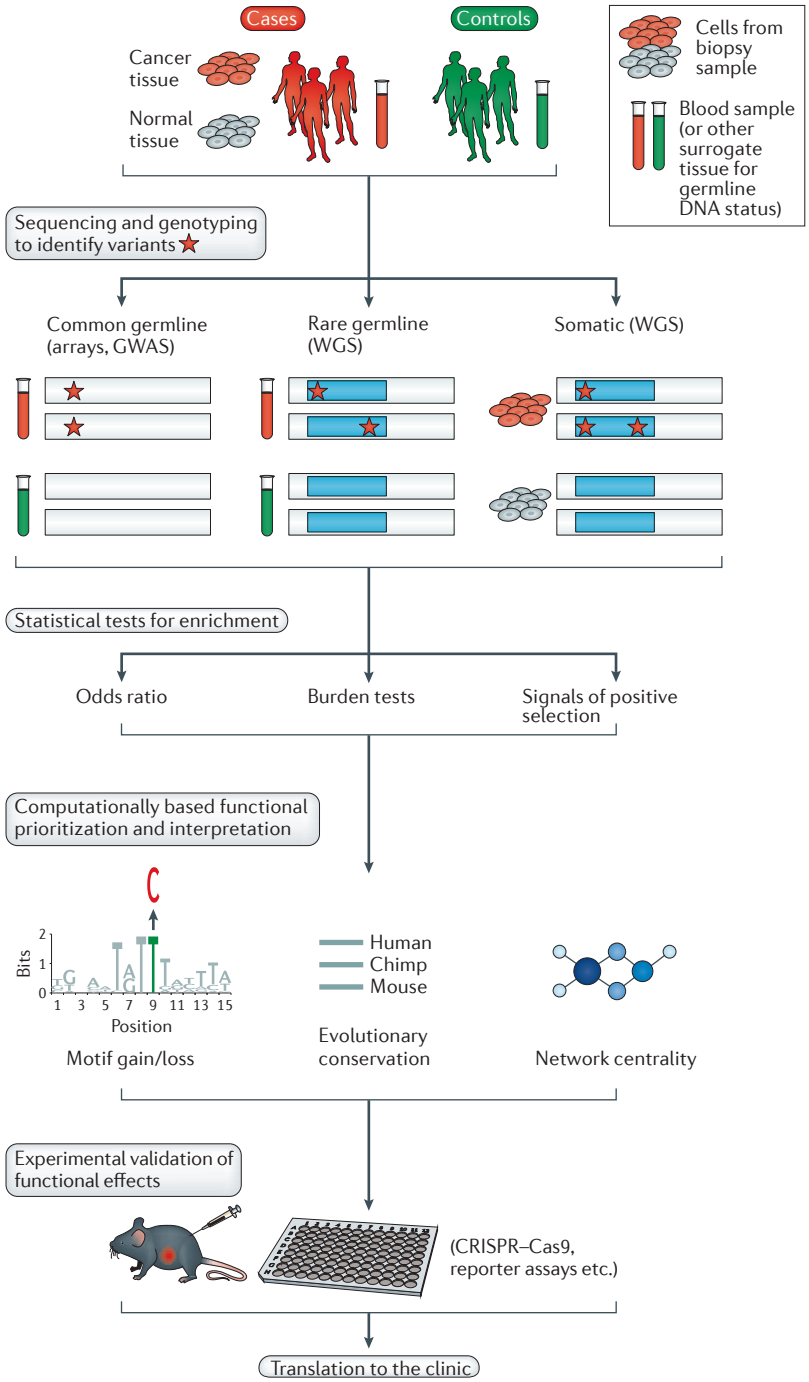


Figure 3 | **Study designs commonly used to identify germline and somatic non-coding sequence variants linked with tumorigenesis.** Common germline variants have mostly been probed by genome-wide association studies (GWASs) using blood samples, in which the genotypes of cases are compared with those of controls. Effect sizes can be estimated using statistical tests, such as measuring the odds ratio[139]. Rare germline variants need whole-genome sequencing (WGS) data, and variants are often collapsed in the same element (blue blocks) to gain statistical power (burden tests)[88]. We note that WGS enables the identification of both common and rare variants linked with tumorigenesis. Somatic sequence variants also need WGS data from tumour versus matched normal tissue (blood is used when normal tissue is not available), and driver mutations are often identified by aggregating variants across elements[58]. Computational methods (such as those listed in TABLE 2) can be used to prioritize the variants and interpret their functional effects. Experimental approaches (such as those shown in FIG. 5) can be used for functional validation. Generally, the variants that pass all these steps are then translated to the clinic for diagnostic and therapeutic purposes.

ncRNAs and their binding sites. Dysregulation of ncRNAs is a cancer signature, and at least in some cases it could be due to the presence of somatic variants in them. *MALAT1* (metastasis-associated lung adenocarcinoma transcript 1; also known as NEAT2) is a lncRNA that regulates the expression of genes that are associated with metastasis. It was first identified in lung cancer but was later observed to be upregulated in many different cancers, including bladder cancer[79,80]. However, the reasons for its upregulation are not clear. In a pan-cancer analysis of ~3,200 tumours from 12 cancer types, *MALAT1* was found to be significantly mutated in bladder cancer[1]. This indicates that mutations in the *MALAT1* sequence might be under positive selection in the tumour. It is known that positive selection in coding genes can be linked with their dysregulation, whereby the mutations lead to loss or gain of function. Similarly mutations in *MALAT1* might be related to its overexpression in bladder cancer. In another example, copy number amplification of a lncRNA, *lncUSMycN*, is thought to contribute to neuroblastoma progression[81,82]. *lncUSMycN* binds to the RNA-binding protein NonO, leading to upregulation of *MYCN* oncogene, which further leads to tumorigenesis.

Besides sequence variants in ncRNAs themselves, mutations in their binding sites are also linked to cancer. For example, miR-31 targets the *AR* mRNA directly, and a mutation disrupting a miR-31-binding site in the *AR* gene can lead to overexpression of AR in prostate cancer[83] (FIG. 4D).

Role of pseudogenes in modulating the expression of a parental gene. Because of their resemblance to their parental protein-coding genes, transcribed pseudogenes are thought to have a natural way of affecting and regulating their parental counterparts[44]. In particular, pseudogene deletion or amplification can affect competition for miRNA binding with the parent gene, which in turn could affect the expression of the parent gene. *PTEN* and *BRAF* are well-studied cancer genes that are often dysregulated in tumours. They both have pseudogenes that can act as miRNA sponges. *PTEN* is a tumour suppressor that negatively regulates the AKT (also known as PKB) signalling pathway[84]. The *PTEN* pseudogene is deleted in cancer, and as a result more miRNAs bind to the 3′ UTR of the parental *PTEN* mRNA, leading to downregulation of its expression[84] (FIG. 4E).

By contrast, overexpression of the *BRAF* pseudogene in mice leads to increased miRNA sequestration, which results in overexpression of the *BRAF* gene[85]. *BRAF* is an oncogene that plays an important part in the mitogen-activated protein kinase (MAPK) signalling pathway, and its overexpression leads to increased MAPK activity. Genomic aberrations of the *BRAF* pseudogene have been observed in many cancer types[85].

Roles for germline variants in cancer

Most of the non-coding germline variants that are associated with cancer susceptibility have been identified by GWASs, although some were discovered through focused studies of specific cancer genes or pathways. GWASs have typically used SNP-sensitive microarrays

© 2016 Macmillan Publishers Limited. All rights reserved

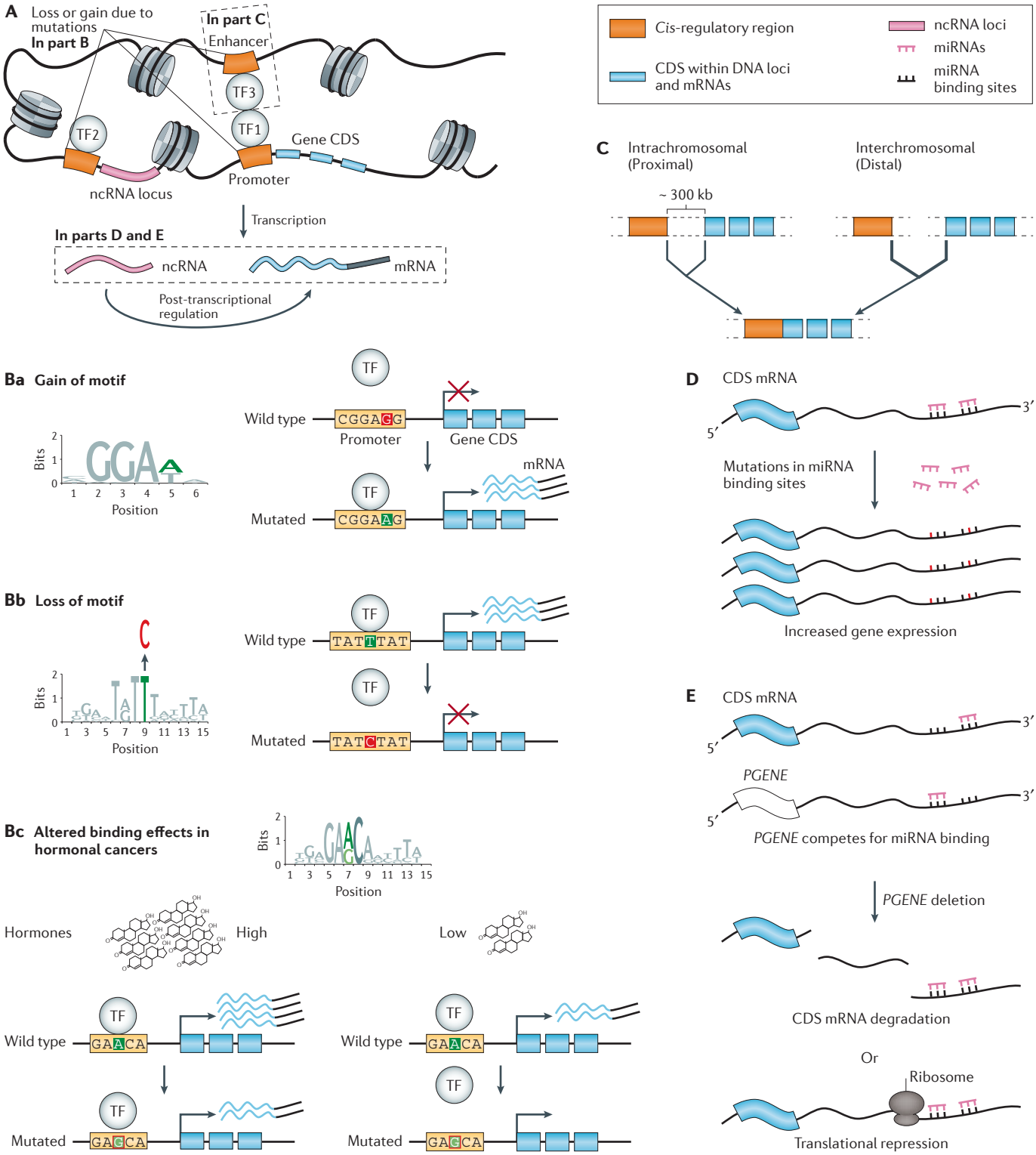


Figure 4 | **Effect of sequence variants in non-coding regions in tumorigenesis. A** | Overview of the non-coding elements that can be affected. Specific cases are shown in parts **B**–**E**. **B** | Mutations can lead to gain (part **Ba**) or loss (part **Bb**) of transcription factor (TF)-binding motifs. Other sequence variants, such as amplification or deletion of the motif, can have similar effects. The effects of a single nucleotide polymorphism that reduces nuclear receptor (NR) binding affinity to DNA are observed at lower NR levels as a result of reduced hormone levels (part **Bc**). **C** | Structural variants (SVs) juxtapose the oncogene (for example, the growth factor independent 1 family oncogenes *GFI1* and *GFI1B*) next to a regulatory element (such as a super enhancer). Deletions, tandem duplications, inversions, translocations or other complex SVs can juxtapose the gene next to an enhancer, leading to its transcription. Either the enhancer or gene can overlap SVs. **D** | Mutations in microRNA (miRNA)-binding sites prevent miRNA binding, leading to increased target gene expression. **E** | PTEN pseudogene (*PGENE*) loss. Pseudogene deletion leads to more miRNAs binding to the parent gene, leading to mRNA silencing through its degradation or translational repression. CDS, coding sequence; ncRNA, non-coding RNA.

© 2016 Macmillan Publishers Limited. All rights reserved

(for example, iCOGS array[86]) that also include variants in non-coding regions. However, a major limitation of the array-based GWAS approach is that only common variants are represented, so associations with disease risk can only be identified for these common variants. The availability of WGS data from matched normal tissue of cancer patients with proper controls from healthy individuals allows the investigation of both rare and common variants linked with cancer risk (FIG. 3). In general, the highly deleterious variants with stronger functional impact are more likely to be rare[54,87]. However, rare variant association studies of individual variants are usually underpowered owing to their low frequencies. Thus, rare variants in the same element are often collapsed into a single variable to gain statistical power (in the framework of burden tests)[88].

Like somatic variants, germline non-coding variants can also affect gene expression in many different ways; for example, point mutations in binding motifs of sequence-specific TFs may disrupt their binding, and large deletions may remove entire TF-binding sites or enhancer elements (FIG. 4). However, we note that GWAS SNPs associated with cancer risk might not be the causal variants and might instead be in LD with them. Functional studies of GWAS SNPs and those in LD with them can help to identify the causal variants and shed light on their mechanism of action. Below, we discuss a few examples of non-coding germline variants that are related to cancer susceptibility and functional effects of which have been studied.

Promoter mutations. Germline mutations in the *TERT* promoter are associated with familial melanoma[15]. Similarly to the effects of somatic mutations, these create binding motifs for the ETS family of TFs, including TCFs (FIG. 4B). The functional effects of these mutations are more likely to be observed in the tissues where these TFs are expressed. Increased expression of the TCF *ELK1* gene is observed in female-specific tissues, such as ovary and placenta. Horn *et al.*[15] reasoned that besides melanoma, this may be related to the increased ovarian cancer risk in women who are carriers of the mutation.

Moreover, a SNP in the *MDM2* promoter is associated with accelerated tumour formation in many different cancer types[89,90]. This SNP likely increases the binding affinity of the Sp1 TF, leading to upregulation of *MDM2*. MDM2 is a negative regulator of the tumour suppressor p53, so *MDM2* overexpression leads to suppression of the p53 pathway, resulting in increased tumour formation.

SNPs in enhancers. Multiple SNPs in a gene desert on chromosome 8q24 upstream of *MYC* are related to increased risk for many cancer types (breast, prostate, ovarian, colon and bladder cancers and chronic lymphocytic leukaemia)[91]. Several observations, such as histone methylation and acetylation marks and 3C assays, suggest that these 8q24 SNPs occur in regions that act as enhancers for *MYC* in a tissue-specific manner. Tissue-specificity of these regulatory regions might be the reason why these SNPs are associated with specific cancer types. In another example, a prostate cancer-associated SNP occurs in a cell-type specific enhancer and leads to increased HOXB13 binding[92]. This, in turn, increases the expression of *RFX6*, which is linked to cell growth in prostate cancer. In addition, a recent study showed that a polymorphism in a super-enhancer leads to differential binding of the GATA TF and influences neuroblastoma susceptibility[93].

Hormone-regulated cancers (such as prostate, breast, ovarian and endometrial cancer) present an interesting case in which the effect of mutations in TF-binding sites might vary with age owing to differential TF activity during a person's lifetime. AR and oestrogen receptor (ER) are nuclear receptor TFs that are activated by the androgen and oestrogen hormones, respectively. The production of these hormones varies substantially during an individual's lifetime, leading to varied activity of AR and ER. Thus, the effect of germline polymorphisms in the promoters or enhancers that are recognized by these TFs can be age-dependent and can vary depending on the abundance of TFs[94]. For example, the effect of a mutation in the DNA motif that decreases the binding affinity of the TF may be felt more strongly when the TF abundance is low (FIG. 4B). This is because the high TF abundance may compensate for the slight loss of binding affinity.

Variants in introns. Variants in introns can affect splice sites and also cause loss of regulatory repressor elements. For instance, a rare mutation in an intron of *BRCA2* causes aberrant splicing and is related to Fanconi anaemia (a rare recessive disorder associated with a high cancer risk)[95]. Also, germline CNVs spanning intronic inhibitor regulatory elements can lead to the overexpression of target transcripts, potentially modulating cell proliferation or cell migration. For example, the loss of an intronic regulatory element in *MGAT4C* (α-1,3-mannosyl-glycoprotein 4-β-N acetylglucosaminyltransferase C) was found to be associated with an increased risk of developing aggressive prostate cancer in a population-based study[96].

SNPs in ncRNAs and their binding sites. In an effort to find germline variants that are associated with ovarian cancer, Chen *et al.* performed targeted sequencing of miRNAs and 3′ UTRs (as they contain miRNA-binding sites) of ~6,000 cancer-associated genes from 31 patients with ovarian cancer[97]. They found enrichment of a variant in the 3′ UTR of *PCM1* (pericentriolar material 1) in cases versus controls. *PCM1* associates with the centrosome complex in a cell-cycle dependent manner and is misregulated in ovarian cancer. Although the mechanism of action of the variant is not clear, it is possible that it alters miRNA binding (because it is located in the 3′ UTR), resulting in differential *PCM1* mRNA expression.

Whereas most cancer-associated polymorphisms are related to increased risk, some of them can also be beneficial and reduce susceptibility. A SNP in miR-27a impairs the processing of pre-mir-27a to its mature version. The reduced miR-27a level results in increased expression of its target gene, *HOXA10* (homeobox A10), and reduced susceptibility to gastric cancer[98]. *HOXA10* codes for a

Burden tests
Statistical methods to test the cumulative effect of multiple variants in a genomic region.

© 2016 Macmillan Publishers Limited. All rights reserved

TF that plays a central part in tumour biology, likely by regulating the expression of key downstream genes, such as *TP53* (REF. 99).

The examples above do not constitute an exhaustive list of all known cases of non-coding germline variants that are associated with altered cancer risk, but are meant to illustrate the diverse ways in which many regulatory polymorphisms exhibit their functional effects. Other methods of identifying variants with potential functional consequences, such as eQTLs and allele-specific expression analyses, have been used to interpret cancer-associated loci identified through GWAS[100–102]. Such studies reveal germline determinants of gene expression in tumours and help to establish a link between non-coding risk loci and their target coding genes.

Interplay between germline and somatic variants

As is apparent from several of the case studies described above, cancer results from a complex interplay of inherited germline and acquired somatic variants. Knudson's two-hit hypothesis is widely known, whereby one allele of a tumour suppressor gene is disrupted by a germline variant and the second through somatic mutation, resulting in oncogenesis[103]. Loss of heterozygosity (LOH) events affecting non-coding elements have also been observed. In these cases, somatic variants disrupt the only functioning copy of the non-coding element, as one copy is already disabled by germline variants. For example, one study reported that many miRNA-encoding loci are located at regions undergoing LOH — overall 80 miRNAs were located at regions undergoing LOH or amplification[104]. The loss of these miRNAs may promote oncogenesis by leading to overexpression of their target oncogenes. In addition, two lncRNAs, *LOC285194* and *BC040587*, were noted to be in a region that frequently undergoes deletions and LOH in osteosarcoma[105]. Deletions of these ncRNAs are also associated with poor patient survival. A separate study later found that indeed *LOC285194* acts as a tumour suppressor[106]. Overall, we expect the availability of thousands of tumour–normal matched whole genomes from the PCAWG project will enable the high-resolution analysis of such LOH events in different cancer types and the simultaneous probing of non-coding elements for the presence of germline and somatic variants.

In a contrasting scenario to LOH events, a common SNP (rs2853669) in the *TERT* promoter weakens the effects of somatic *TERT* promoter mutations. As observed in bladder cancer, patients with somatic lesions in the *TERT* promoter who also carried this germline SNP showed better survival[107]. From a mechanistic viewpoint, the common SNP might weaken the effect of somatic mutations because it disrupts a pre-existing ETS2-binding site. Thus, the multiple germline and somatic variants in the *TERT* promoter particularly demonstrate the complex relationship of regulatory variants with cancer susceptibility, oncogenesis and patient survival.

Computational methods for identifying drivers

Computational prediction of drivers is a challenging task. Below we discuss the various methods that exist to predict drivers in coding and non-coding regions.

Broadly speaking, driver identification uses two lines of evidence: detection of signals of positive selection (that is, the presence of more recurrent mutations than expected by random chance)[14] or prediction of mutations with high functional impact[108]. Analysis of the recurrence of somatic variants from tumour samples in functional elements to identify regions under positive selection is similar to the burden test strategy that is used to associate rare germline variants with complex traits[109]. Such analyses can be done for a specific cancer type or across multiple cancers[110]. In addition, computational tools that try to identify driver genes by detecting positive selection signatures need to account for genomic mutation rate covariates (such as transcriptional activity and DNA replication timing) that lead to mutational heterogeneity across the genome[14,111]. Methods that aim to predict the functional impact of nonsynonymous mutations in coding genes (for example, SIFT[112] and PolyPhen[113]) can be used for both germline and somatic variants. These methods use many features, such as evolutionary conservation, protein structural information and physicochemical properties of amino acid changes[108]. Nevertheless, there is significant room for improvement of these methods[108].

Computational identification of non-coding drivers is in many ways even more challenging than coding drivers owing to their complex and varied modes of action (as discussed in this Review) and our poor understanding of non-coding regions in general. Non-coding mutations are also more abundant than coding ones and thus the key mutations with functional impact have to be distinguished from a larger set of passenger events.

Some methods of identifying non-coding drivers analyse the recurrence of somatic variants from tumour samples in functional elements[58,59,61,111]. We note that such methods that try to identify driver non-coding elements (that is, those undergoing positive selection in tumours) also need to account for genomic mutation rate covariates like the methods for driver analyses of coding genes, as discussed above[14,111,114]. A number of computational tools also exist to annotate and prioritize potentially functional non-coding variants with high impact. A list of these tools with their key features and corresponding references is provided in TABLE 2. Most of these tools can interpret both SNVs and indels, and some tools (for example, ANNOVAR, VEP and GEMINI) also analyse structural variants. Many tools first annotate variants with various functional annotations (such as TF-binding sites and ncRNAs). Some of them try to interpret the effect of *cis*-regulatory mutations at a nucleotide-level resolution by computing whether they create new TF-binding motifs or lead to loss of existing ones[115]. In addition, various biological networks can provide information about the connectivities of the target genes of non-coding variants. In particular, mutations in regulatory regions of highly connected genes in protein–protein interaction and regulatory networks have been suggested to have a higher functional impact than those targeting peripheral genes in the network[54,116]. Also, high conservation among humans and across multiple species tends to be an indicator of function, and hence it is used as a feature by many tools. Some tools are

Positive selection
Directed selection that forces the allele frequency of advantageous mutations to increase.

© 2016 Macmillan Publishers Limited. All rights reserved

Table 2 | **Computational methods to prioritize non-coding variants with functional effects**

Tool	Variant type	Functional annotation	Conservation	LD calculation	Somatic mutation recurrence	Scoring scheme	Refs
SeattleSeq	SNV, indel	Y	Y	N	N	N	138
SNPnexus	SNV, indel	Y	Y	N	N	N	140,141
ANNOVAR	SNV, indel, SV	Y	Y	N	N	N	142
VEP	SNV, indel, SV	Y	N	N	N	N	143
OncoCis	SNV, indel	Y	Y	N	N	N	144
GEMINI	SNV, indel, SV	Y	Y	N	N	N	145
FunciSNP	SNP	Y	N	Y	N	N	146
HaploReg	SNP, indel	Y	Y	Y	N	N	147
GWAS3D	SNP	Y	Y	Y	N	Y	148
is-rSNP	SNV	N	N	N	N	Y	149
RegulomeDB	SNV	Y	N	N	N	Y	150
SInBaD	SNV	N	Y	N	N	Y	151
CADD	SNV, indel	Y	Y	N	N	Y	152
FunSeq	SNV, indel	Y	Y	N	Y	Y	54,115
GWAVA	SNV, indel	Y	Y	N	N	Y	153
FitCons	SNV	Y	Y	N	N	Y	154
DeepSEA	SNV, indel	Y	Y	N	N	Y	155

Indel, insertion and deletion; LD, linkage disequilibrium; SNP, single nucleotide polymorphism; SNV, single nucleotide variant; SV, structural variant.

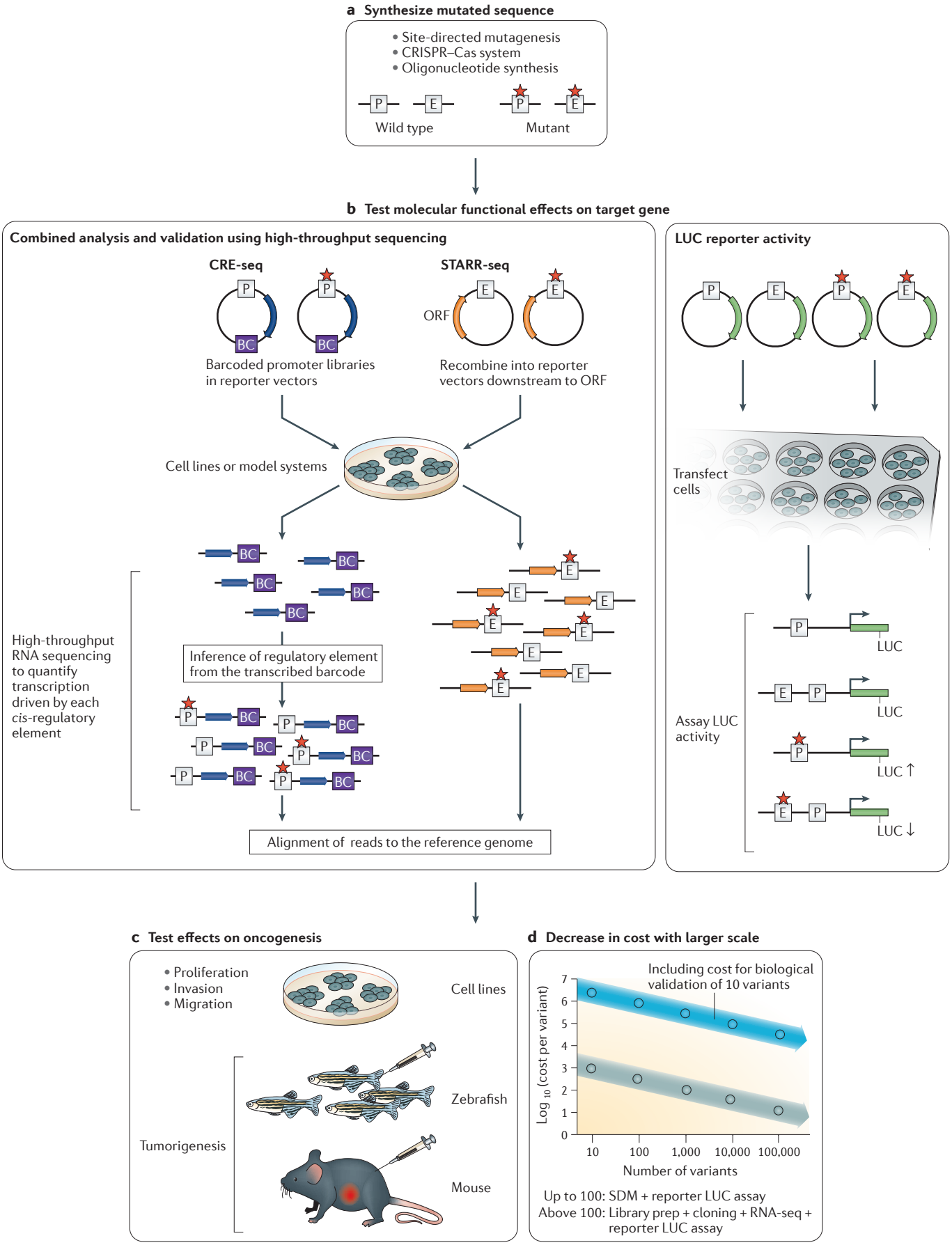
designed specifically for common GWAS variants (for example, FunciSNP, Haploreg and GWAS3D) and try to identify candidate regulatory SNPs that are correlated with GWAS SNPs. This is because the GWAS hit may not be the causal variant but might be in LD with it. Thus, they identify putative causal variants for complex disorders, including cancer susceptibility. Finally, some methods integrate all the features to provide a score for the likely functional impact of each variant (for example, RegulomeDB, CADD, FunSeq and FitCons).

Experimental approaches for functional validation

Most functional validation studies of mutations have focused on the coding portion of the genome. With an expanding appreciation that non-coding mutations have an important role in cancer progression, several recent studies have begun to explore methods to functionally assess non-coding mutations. For example, experimental approaches to understand the effects of *cis*-regulatory mutations in promoters and enhancers on cellular functions are illustrated in FIG. 5. The main strategies first require introducing the sequence variants (FIG. 5a), determining the resulting molecular level effect on transcription using high- and low-throughput functional assays (FIG. 5b) and demonstrating direct biological significance as manifested by alteration in oncogenic properties (for example, increased invasion, proliferation or colony formation; FIG. 5c).

One way to introduce sequence variants involves the use of the CRISPR–Cas9 system to edit the genome[117] (FIG. 5a). Oligonucleotides containing the multiple mutations may also be synthesized directly for high-throughput screening. Then, the direct consequence of non-coding mutations can be evaluated using high-throughput sequencing-based assays[118–120] or low- to medium-throughput luciferase reporter assays[68] (FIG. 5b). Analysis of functional consequences of non-coding mutations in promoters can be achieved in a high-throughput manner using a modification of *cis*-regulatory element analysis by sequencing (CRE-seq)[118]. In this approach, synthetic promoter libraries drive the expression of a common reporter gene and a downstream unique barcode sequence that identifies the upstream promoter. RNA-seq then reveals the effects of promoter variants on the expression levels of their paired barcode sequence. Unlike for promoters, the activity of enhancers is thought to be independent of their location, so enhancer libraries can be incorporated into high-throughput reporter assays using different reporter construct arrangements[121]. In CRE-seq-based approaches[120,122], the enhancer is placed upstream of the reporter gene and the identifying barcode, whereas for self-transcribing active regulatory region sequencing (STARR-seq)[119] the enhancer library is placed downstream of the reporter construct and is itself expressed at the RNA level, and hence can be identified directly by RNA-seq rather than requiring a separate barcode sequence. The cloned libraries can be transfected into eukaryotic cells in a high-throughput, pooled format, and RNA-seq is used to assess the resulting expression level of the reporter (specifically, the expressed unique barcode or downstream enhancer) driven by each variant

© 2016 Macmillan Publishers Limited. All rights reserved



© 2016 Macmillan Publishers Limited. All rights reserved

Figure 5 | **Methods for functional validation of non-coding variants. a** | Mutations in cloned DNA fragments can be generated using site-directed mutagenesis or the CRISPR–Cas system. Synthetic oligonucleotides with a wild-type or mutant sequence can also be chemically synthesized. **b** | Functional output of the non-coding mutations can be determined either using a single or combinatorial approach involving high-throughput sequencing and/or luciferase (LUC) reporter assays. In high-throughput sequencing, effects of mutations in *cis*-regulatory elements (promoters and enhancers) can be studied by an approach called *cis*-regulatory element analysis by sequencing (CRE-seq)[118,122]. For CRE-seq, synthetic regulatory element constructs with wild-type and mutated sequence are cloned into reporter construct, which is tagged at the 3′ end using a specific nucleotide barcode that identifies the upstream promoter or enhancer element. In an alternative method for characterizing enhancer variants, self-transcribing active regulatory region sequencing (STARR-seq)[119], enhancer libraries are flanked by synthetic adaptor DNA sequences and cloned downstream of a transcription reporter construct. For both approaches, RNA transcripts from these libraries are used for cDNA synthesis followed by high-throughput sequencing. The expression driven by each element is measured by the ratio of the fraction of reads in the cDNA pool and the genomic DNA pool for each library construct; the particular element driving the expression of each transcript is identified based on the sequence of the transcribed barcode (for CRE-seq) or the transcribed enhancer (for STARR-seq). This enables accurate quantification of the reporter transcript as a direct measure of the regulatory element activity. For the LUC reporter assays, DNA fragments cloned into the reporter vectors are transfected in cells followed by measuring the reporter activity. **c** | Oncogenic properties, such as cell proliferation, migration and invasion, can be tested *in vitro* using cell lines, and tumorigenesis can also be tested *in vivo* using model organisms. **d** | The cost of functional validation per mutation changes with the techniques used and is the highest when *in vitro* and *in vivo* oncogenic validation studies are included. Approximate cost per variant for functional validation from 10 up to 100 variants is calculated using a combination of site directed mutagenesis (SDM; ~US$100 per variant) and reporter luciferase assays (~$180 per variant). However, for functional validation of over 1,000 variants, cost per variant is optimized with oligonucleotide library synthesis with and without the mutation, cloning (~$1,250 per variant), transfection into cells, RNA extraction and high-throughput sequencing (~$15 per variant) and reporter assays. The light blue line depicts the huge increase in cost when biological validation (*in vitro* and *in vivo* tumorigenic assays) are also carried out.

element. In a separate approach, visible reporter assays using synthetic transcription reporter constructs (such as luciferase reporter assays) that contain the regulatory sequences of the reporter gene enable direct validation of non-coding mutations. This approach is distinct from the sequencing-based approaches and involves preparation of control and variant constructs based on sequence information. Low- to medium-throughput can be achieved by testing the luciferase expression driven of constructs individually in multi-well plates. A major limitation of this well-by-well approach is the variable efficiency of transfecting cells with the reporter vector, which prevents this approach from becoming a high-throughput assay.

Besides the strategies discussed above, which validate the effects of variants in promoters and enhancers on gene expression, other approaches are needed to validate variants in ncRNAs, UTRs and introns. For instance, the effects of mutations in 3′ and 5′ UTRs have been tested by constructing large-scale mutant libraries with thousands of sequence variants and measuring their effects on mRNA and protein expression[32,33]. This approach revealed that most mutations in the 3′ UTRs in model organisms such as yeast have a minor effect on expression. However, some mutations in a TA-rich element linked with 3′ end processing had a strong effect, with up to tenfold change in expression[32]. Among the 5′ UTR variants, the UTR nucleotides at positions −3 to −1, those affecting mRNA secondary structure and out-of-frame upstream AUGs had the strongest effects on protein levels[33]. To test the effects of intronic variants on splicing, minigene assays can be used. In these assays, the variant sequence is cloned into transcription-competent minigene vectors and transfected into mammalian cells. This is followed by examination of the splicing patterns of the transcripts generated from wild-type and variant constructs[123,124].

Minigene assays
Assays using a plasmid with a minimal gene fragment necessary for the gene to be expressed. It can include exons as well as introns, and it serves as a tool for evaluating splicing patterns.

Functional screening approaches help to identify the best candidates but still need tumour type-specific validation. For example, in melanoma samples, mutation in the *NDUFB9* promoter significantly altered promoter activity, as assessed by a luciferase assay in the COLO-829 cell line, but the expression of this promoter did not differ significantly between patients that carry the mutation versus those that do not[125]. Thus, tissue, tumour and genomic context are important factors and require validation.

Functional validation requires demonstrating oncogenic properties that are increased owing to the variant in question (FIG. 5c). To achieve this, wild type (control) and mutants are compared *in vitro* in transfection-based functional assays in cell lines, and *in vivo* using model organisms (for example, zebrafish or mice). Cell line experiments can be used to assess increased cell proliferation, colony formation and the ability of cells to invade through a barrier. *In vivo* models compare tumour growth and ability to achieve metastases between the control and mutant variants. In both *in vitro* and *in vivo* experiments, the selection of model systems is crucial and potentially limiting; ideally, one tries to achieve the context that most closely resembles the situation in which the mutations arose. For some tumour types, like breast, lung and melanoma, there are abundant cell lines, but for others, such as prostate cancer, only few cell lines exist. New approaches in developing patient-derived spheroids or organoids have been proposed to bridge this gap[126–128].

Thus, overall functional validation of non-coding variants is extremely important to understand their biological consequence. High-throughput analysis of variants substantially reduces the cost per variant tested (FIG. 5d). Among the current methods for functional validation of variants, the biological validation for oncogenic properties is the most costly because of the model systems used and the amount of time it takes to achieve a biologically relevant readout. For example, if a mouse model is used to test xenografts, it may take months to determine meaningful growth objectives. When comparing the influence of non-coding variants between humans and mice it will be important to consider species-specific differences and selection of mouse strains with appropriate genetic backgrounds[129]. If a genetically engineered mouse model is required, *de novo* tumour development could take years and hundreds of mice. However, using CRISPR–Cas9 strategies to develop such models may help to accelerate this process in the near future[130,131]. Thus, in general, high-throughput prioritization of putative functional mutations is crucial before the testing of the

© 2016 Macmillan Publishers Limited. All rights reserved

Precision medicine
Medical care tailored to the individual patient, usually using the patient's genomic sequence.

most promising candidates in *in vivo* systems, given the lengthier developmental time and high costs of *in vivo* assays compared with other validation steps.

Conclusions

The current belief is that cancer arises because of the accumulation of multiple driver mutations[132], some of which are non-coding. This may be particularly the case for cancer types in which coding driver mutations have not been identified in major subpopulations of patients, such as non-small cell lung cancer[133]. The current published literature is biased against driver mutations in non-coding regions as these sequences have yet to be explored to the same extent as coding genes owing to the lower costs of exome sequencing and difficulty of interpreting the consequences of mutations in non-coding regions.

Recent studies have shown that small changes in gene expression caused by non-coding mutations can have large phenotypic impact (for example, a SNP in the enhancer of the *KITLG* gene causes 20% change in gene expression and is responsible for blond hair colour[134]). Thus, the combined effect of small changes in expression due to non-coding mutations in cancer might be more important than currently appreciated. Indeed, evidence is emerging that the cumulative effect of co-suppression or co-activation of multiple genes can lead to tumorigenesis. In particular, certain oncogenes and tumour-suppressor genes seem to have a range of oncogenic potential, and cancer results from the combined effect of their CNVs[135,136]. Thus, genomic variants contribute to oncogenesis with continuously varying effects, as opposed to their binary classification into driver mutations and passenger mutations. The effects of somatic variants also depend on the existing genetic background, for example, the presence of risk alleles in inherited germline DNA. Although some somatic variants may have a direct role (such as *TERT* promoter mutations found in many different cancer types[69]), others may modulate important cancer pathways indirectly.

Currently, there is a debate in the community about whether we should analyse whole genomes or exomes. Studies of somatic non-coding mutations are currently reserved for research purposes and have not been incorporated into precision medicine cancer care approaches in the clinic. This is primarily because current therapeutic approaches attempt to target proteins. However, as discussed in this Review, non-coding driver mutations are linked with the expression of the protein-coding genes they regulate. Thus, identification of non-coding driver mutations can enable therapeutic approaches that target the linked protein. Moreover, identification of non-coding germline variants associated with increased cancer susceptibility is also important for risk assessment and potentially for preventive approaches.

The above discussion highlights the importance of accurately determining the links between *cis*-regulatory regions and their target genes to interpret the functional effects of regulatory variants. Although many approaches exist (as discussed in this Review), this remains an active and important area of research, especially the development of high-throughput 3C-derived technologies. We note that even when the links between regulatory regions and target genes are known, it will be important to study the effects of mutations in all elements controlling gene expression in a comprehensive manner. Thus, network approaches will be important to understand the role of non-coding mutations in cancer. We might also be able to identify new pathways or new participants in known pathways that are important in cancer.

1. Kandoth, C. *et al.* Mutational landscape and significance across 12 major cancer types. *Nature* **502**, 333–339 (2013).
2. Easton, D. F. & Eeles, R. A. Genome-wide association studies in cancer. *Hum. Mol. Genet.* **17**, R109–R115 (2008).
3. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
4. Chen, C. Y., Chang, I. S., Hsiung, C. A. & Wasserman, W. W. On the identification of potential regulatory variants within genome wide association candidate SNP sets. *BMC Med. Genomics* **7**, 34 (2014).
5. Akhtar-Zaidi, B. *et al.* Epigenomic enhancer profiling defines a signature of colon cancer. *Science* **336**, 736–739 (2012).
6. Kron, K. J., Bailey, S. D. & Lupien, M. Enhancer alterations in cancer: a source for a cell identity crisis. *Genome Med.* **6**, 77 (2014).
7. Prensner, J. R. & Chinnaiyan, A. M. The emergence of lncRNAs in cancer biology. *Cancer Discov.* **1**, 391–407 (2011).
8. Iyer, M. K. *et al.* The landscape of long noncoding RNAs in the human transcriptome. *Nat. Genet.* **47**, 199–208 (2015).
9. Stirzaker, C., Taberlay, P. C., Statham, A. L. & Clark, S. J. Mining cancer methylomes: prospects and challenges. *Trends Genet.* **30**, 75–84 (2014).
10. The 1000 Genomes Project Consortium. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491**, 56–65 (2012).
11. Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
12. Abyzov, A. *et al.* Somatic copy number mosaicism in human skin revealed by induced pluripotent stem cells. *Nature* **492**, 438–442 (2012).
13. De, S. Somatic mosaicism in healthy human tissues. *Trends Genet.* **27**, 217–223 (2011).
14. Lawrence, M. S. *et al.* Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**, 214–218 (2013).
Shows how mutational heterogeneity in the genome can lead to false positives during the identification of cancer driver genes.
15. Horn, S. *et al.* *TERT* promoter mutations in familial and sporadic melanoma. *Science* **339**, 959–961 (2013).
One of the first papers showing prevalence of *TERT* promoter mutations in cancer.
16. Daye, Z. J., Li, H. & Wei, Z. A powerful test for multiple rare variants association studies that incorporates sequencing qualities. *Nucleic Acids Res.* **40**, e60 (2012).
17. Baca, S. C. *et al.* Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666–677 (2013).
18. Stephens, P. J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27–40 (2011).
19. Holland, A. J. & Cleveland, D. W. Boveri revisited: chromosomal instability, aneuploidy and tumorigenesis. *Nat. Rev. Mol. Cell Biol.* **10**, 478–487 (2009).
20. Nik-Zainal, S. *et al.* Mutational processes molding the genomes of 21 breast cancers. *Cell* **149**, 979–993 (2012).
21. Wittkopp, P. J. & Kalay, G. *Cis*-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat. Rev. Genet.* **13**, 59–69 (2012).
22. Galas, D. J. & Schmitz, A. DNAse footprinting: a simple method for the detection of protein–DNA binding specificity. *Nucleic Acids Res.* **5**, 3157–3170 (1978).
23. Neph, S. *et al.* An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* **489**, 83–90 (2012).
24. Dunham, I. *et al.* An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
Discussion of functional annotations from the ENCODE project.
25. Hughes, J. R. *et al.* Analysis of hundreds of *cis*-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat. Genet.* **46**, 205–212 (2014).
26. de Laat, W. & Dekker, J. 3C-based technologies to study the shape of the genome. *Methods* **58**, 189–191 (2012).
27. Yip, K. Y. *et al.* Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol.* **13**, R48 (2012).
28. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* **348**, 648–660 (2015).
29. Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE data. *Nature* **489**, 91–100 (2012).
30. Chadwick, L. H. The NIH Roadmap Epigenomics Program data resource. *Epigenomics* **4**, 317–324 (2012).
31. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).

© 2016 Macmillan Publishers Limited. All rights reserved

32. Shalem, O. *et al.* Systematic dissection of the sequence determinants of gene 3′ end mediated expression control. *PLoS Genet.* **11**, e1005147 (2015).
33. Dvir, S. *et al.* Deciphering the rules by which 5′-UTR sequences affect protein expression in yeast. *Proc. Natl Acad. Sci. USA* **110**, E2792–E2801 (2013).
34. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
35. The GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
36. Morris, K. V. & Mattick, J. S. The rise of regulatory RNA. *Nat. Rev. Genet.* **15**, 423–437 (2014).
37. Guttman, M. & Rinn, J. L. Modular regulatory principles of large non-coding RNAs. *Nature* **482**, 339–346 (2012).
38. Zhao, J., Sun, B. K., Erwin, J. A., Song, J. J. & Lee, J. T. Polycomb proteins targeted by a short repeat RNA to the mouse X chromosome. *Science* **322**, 750–756 (2008).
39. Rinn, J. L. *et al.* Functional demarcation of active and silent chromatin domains in human *HOX* loci by noncoding RNAs. *Cell* **129**, 1311–1323 (2007).
40. Penny, G. D., Kay, G. F., Sheardown, S. A., Rastan, S. & Brockdorff, N. Requirement for *Xist* in X chromosome inactivation. *Nature* **379**, 131–137 (1996).
41. Schmitz, K. M., Mayer, C., Postepska, A. & Grummt, I. Interaction of noncoding RNA with the rDNA promoter mediates recruitment of DNMT3b and silencing of rRNA genes. *Genes Dev.* **24**, 2264–2269 (2010).
42. Zhang, Z. *et al.* PseudoPipe: an automated pseudogene identification pipeline. *Bioinformatics* **22**, 1437–1439 (2006).
43. Khurana, E. *et al.* Segmental duplications in the human genome reveal details of pseudogene formation. *Nucleic Acids Res.* **38**, 6997–7007 (2010).
44. Sasidharan, R. & Gerstein, M. Genomics: protein fossils live on as RNA. *Nature* **453**, 729–731 (2008).
45. Tam, O. H. *et al.* Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* **453**, 534–538 (2008).
46. Loots, G. G. *et al.* Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. *Science* **288**, 136–140 (2000).
47. Pennacchio, L. A. & Rubin, E. M. Genomic strategies to identify mammalian regulatory sequences. *Nat. Rev. Genet.* **2**, 100–109 (2001).
48. Waterston, R. H. *et al.* Initial sequencing and comparative analysis of the mouse genome. *Nature* **420**, 520–562 (2002).
49. Gibbs, R. A. *et al.* Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428**, 493–521 (2004).
50. Lindblad-Toh, K. *et al.* Genome sequence, comparative analysis and haplotype structure of the domestic dog. *Nature* **438**, 803–819 (2005).
51. Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304**, 1321–1325 (2004).
52. Peng, J. C., Shen, J. & Ran, Z. H. Transcribed ultraconserved region in human cancers. *RNA Biol.* **10**, 1771–1777 (2013).
53. Calin, G. A. *et al.* Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. *Cancer Cell* **12**, 215–229 (2007).
54. Khurana, E. *et al.* Integrative annotation of variants from 1092 humans: application to cancer genomics. *Science* **342**, 1235587 (2013).
One of the first methods for genome-wide identification of non-coding candidate cancer drivers.
55. Katzman, S. *et al.* Human genome ultraconserved elements are ultraselected. *Science* **317**, 915 (2007).
56. Ward, L. D. & Kellis, M. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. *Science* **337**, 1675–1678 (2012).
57. Visel, A., Minovitsky, S., Dubchak, I. & Pennacchio, L. A. VISTA Enhancer Browser — a database of tissue-specific human enhancers. *Nucleic Acids Res.* **35**, D88–D92 (2007).
58. Weinhold, N., Jacobsen, A., Schultz, N., Sander, C. & Lee, W. Genome-wide analysis of noncoding regulatory mutations in cancer. *Nat. Genet.* **46**, 1160–1165 (2014).
Analysis of hundreds of cancer whole-genomes to identify driver mutations in non-coding regions.
59. Fredriksson, N. J., Ny, L., Nilsson, J. A. & Larsson, E. Systematic analysis of noncoding somatic mutations and gene expression alterations across 14 tumor types. *Nat. Genet.* **46**, 1258–1263 (2014).
60. Smith, K. S. *et al.* Signatures of accelerated somatic evolution in gene promoters in multiple cancer types. *Nucleic Acids Res.* **43**, 5307–5317 (2015).
61. Melton, C., Reuter, J. A., Spacek, D. V. & Snyder, M. Recurrent somatic mutations in regulatory regions of human cancer genomes. *Nat. Genet.* **47**, 710–716 (2015).
62. Katainen, R. *et al.* CTCF/cohesin-binding sites are frequently mutated in cancer. *Nat. Genet.* **47**, 818–821 (2015).
63. Puente, X. S. *et al.* Non-coding recurrent mutations in chronic lymphocytic leukaemia. *Nature* **526**, 519–524 (2015).
64. Treangen, T. J. & Salzberg, S. L. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nat. Rev. Genet.* **13**, 36–46 (2012).
65. Mijuškovic´, M. *et al.* A streamlined method for detecting structural variants in cancer genomes by short read paired-end sequencing. *PLoS ONE* **7**, e48314 (2012).
66. Meyerson, M., Gabriel, S. & Getz, G. Advances in understanding cancer genomes through second-generation sequencing. *Nat. Rev. Genet.* **11**, 685–696 (2010).
67. Heidenreich, B., Rachakonda, P. S., Hemminki, K. & Kumar, R. *TERT* promoter mutations in cancer development. *Curr. Opin. Genet. Dev.* **24**, 30–37 (2014).
68. Huang, F. W. *et al.* Highly recurrent *TERT* promoter mutations in human melanoma. *Science* **339**, 957–959 (2013).
One of the first papers showing prevalence of *TERT* promoter mutations in cancer.
69. Killela, P. J. *et al.* *TERT* promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal. *Proc. Natl Acad. Sci. USA* **110**, 6021–6026 (2013).
70. Hnisz, D. *et al.* Super-enhancers in the control of cell identity and disease. *Cell* **155**, 934–947 (2013).
71. Mansour, M. R. *et al.* An oncogenic super-enhancer formed through somatic mutation of a noncoding intergenic element. *Science* **346**, 644–648 (2014).
72. Tomlins, S. A. *et al.* Recurrent fusion of *TMPRSS2* and ETS transcription factor genes in prostate cancer. *Science* **310**, 1373–1377 (2005).
73. Yu, J. *et al.* An integrated network of androgen receptor, polycomb, and *TMPRSS2–ERG* gene fusions in prostate cancer progression. *Cancer Cell* **17**, 443–454 (2010).
74. Berger, M. F. *et al.* The genomic complexity of primary human prostate cancer. *Nature* **470**, 214–220 (2011).
75. Weischenfeldt, J. *et al.* Integrative genomic analyses reveal an androgen-driven somatic alteration landscape in early-onset prostate cancer. *Cancer Cell* **23**, 159–170 (2013).
76. Northcott, P. A. *et al.* Enhancer hijacking activates GFI1 family oncogenes in medulloblastoma. *Nature* **511**, 428–434 (2014).
77. Breit, T. M. *et al.* Site-specific deletions involving the *tal*-1 and *sil* genes are restricted to cells of the T cell receptor α/β lineage: T cell receptor δ gene deletion mechanism affects multiple genes. *J. Exp. Med.* **177**, 965–977 (1993).
78. Nambiar, M., Kari, V. & Raghavan, S. C. Chromosomal translocations in cancer. *Biochim. Biophys. Acta* **1786**, 139–152 (2008).
79. Gutschner, T. & Diederichs, S. The hallmarks of cancer: a long non-coding RNA point of view. *RNA Biol.* **9**, 703–719 (2012).
80. Han, Y., Liu, Y., Nie, L., Gui, Y. & Cai, Z. Inducing cell proliferation inhibition, apoptosis, and motility reduction by silencing long noncoding ribonucleic acid metastasis-associated lung adenocarcinoma transcript 1 in urothelial carcinoma of the bladder. *Urology* **81**, 209.e1–209.e7 (2013).
81. Liu, P. Y. *et al.* Effects of a novel long noncoding RNA, lncUSMycN, on N-Myc expression and neuroblastoma progression. *J. Natl Cancer Inst.* **106**, dju113 (2014).
82. Buechner, J. & Einvik, C. N-myc and noncoding RNAs in neuroblastoma. *Mol. Cancer Res.* **10**, 1243–1253 (2012).
83. Lin, P. C. *et al.* Epigenetic repression of miR-31 disrupts androgen receptor homeostasis and contributes to prostate cancer progression. *Cancer Res.* **73**, 1232–1244 (2013).
84. Poliseno, L. *et al.* A coding-independent function of gene and pseudogene mRNAs regulates tumour biology. *Nature* **465**, 1033–1038 (2010).
85. Karreth, F. A. *et al.* The BRAF pseudogene functions as a competitive endogenous RNA and induces lymphoma *in vivo*. *Cell* **161**, 319–332 (2015).
86. Bahcall, O. G. iCOGS collection provides a collaborative model. *Nat. Genet.* **45**, 343 (2013).
87. MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* **335**, 823–828 (2012).
88. Wang, Q., Lu, Q. & Zhao, H. A review of study designs and statistical methods for genomic epidemiology studies using next generation sequencing. *Front. Genet.* **6**, 149 (2015).
89. Bond, G. L. & Levine, A. J. A single nucleotide polymorphism in the p53 pathway interacts with gender, environmental stresses and tumor genetics to influence cancer in humans. *Oncogene* **26**, 1317–1323 (2007).
90. Bond, G. L. *et al.* A single nucleotide polymorphism in the *MDM2* promoter attenuates the p53 tumor suppressor pathway and accelerates tumor formation in humans. *Cell* **119**, 591–602 (2004).
91. Grisanzio, C. & Freedman, M. L. Chromosome 8q24-associated cancers and MYC. *Genes Cancer* **1**, 555–559 (2010).
92. Huang, Q. *et al.* A prostate cancer susceptibility allele at 6q22 increases *RFX6* expression by modulating HOXB13 chromatin binding. *Nat. Genet.* **46**, 126–135 (2014).
93. Oldridge, D. A. *et al.* Genetic predisposition to neuroblastoma mediated by a *LMO1* super-enhancer polymorphism. *Nature* **528**_418–421 (2015).
94. Garritano, S. *et al.* *In-silico* identification and functional validation of allele-dependent AR enhancers. *Oncotarget* **6**, 4816–4828 (2015).
95. Bakker, J. L. *et al.* A novel splice site mutation in the noncoding region of *BRCA2*: implications for Fanconi anemia and familial breast cancer diagnostics. *Hum. Mut.* **35**, 442–446 (2014).
96. Demichelis, F. *et al.* Identification of functionally active, low frequency copy number variants at 15q21.3 and 12q21.31 associated with prostate cancer risk. *Proc. Natl Acad. Sci. USA* **109**, 6686–6691 (2012).
97. Chen, X. *et al.* Targeted resequencing of the microRNAome and 3′UTRome reveals functional germline DNA variants with altered prevalence in epithelial ovarian cancer. *Oncogene* **34**, 2125–2137 (2015).
98. Yang, Q. *et al.* Genetic variations in miR-27a gene decrease mature miR-27a level and reduce gastric cancer susceptibility. *Oncogene* **33**, 193–202 (2014).
99. Chu, M. C., Selam, F. B. & Taylor, H. S. HOXA10 regulates p53 expression and matrigel invasion in human breast cancer cells. *Cancer Biol. Ther.* **3**, 568–572 (2004).
100. Li, Q. *et al.* Integrative eQTL-based analyses reveal the biology of breast cancer risk loci. *Cell* **152**, 633–641 (2013).
101. Xu, X. *et al.* Variants at *IRX4* as prostate cancer expression quantitative trait loci. *Eur. J. Hum. Genet.* **22**, 558–563 (2014).
102. Ongen, H. *et al.* Putative *cis*-regulatory drivers in colorectal cancer. *Nature* http://dx.doi.org/10.1038/nature13602 (2014).
103. Knudson, A. G. Mutation and cancer: statistical study of retinoblastoma. *Proc. Natl Acad. Sci. USA* **68**, 820–823 (1971).
104. Calin, G. A. *et al.* Human microRNA genes are frequently located at fragile sites and genomic regions involved in cancers. *Proc. Natl Acad. Sci. USA* **101**, 2999–3004 (2004).
105. Pasic, I. *et al.* Recurrent focal copy-number changes and loss of heterozygosity implicate two noncoding RNAs and one tumor suppressor gene at chromosome 3q13.31 in osteosarcoma. *Cancer Res.* **70**, 160–171 (2010).
106. Liu, Q. *et al.* LncRNA loc285194 is a p53-regulated tumor suppressor. *Nucleic Acids Res.* **41**, 4976–4987 (2013).
107. Rachakonda, P. S. *et al.* *TERT* promoter mutations in bladder cancer affect patient survival and disease recurrence through modification by a common polymorphism. *Proc. Natl Acad. Sci. USA* **110**, 17426–17431 (2013).
108. Gnad, F., Baucom, A., Mukhyala, K., Manning, G. & Zhang, Z. Assessment of computational methods for predicting the effects of missense mutations in human cancers. *BMC Genomics* **14**, S7 (2013).

© 2016 Macmillan Publishers Limited. All rights reserved

109. Lee, S. *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case–control whole-exome sequencing studies. *Am. J. Hum. Genet.* **91**, 224–237 (2012).
110. Tamborero, D. *et al.* Comprehensive identification of mutational cancer driver genes across 12 tumor types. *Sci. Rep.* **3**, 2650 (2013).
111. Lochovsky, L., Zhang, J., Fu, Y., Khurana, E. & Gerstein, M. LARVA: an integrative framework for large-scale analysis of recurrent variants in noncoding annotations. *Nucleic Acids Res.* **43**, 8123–8134 (2015).
Method that accounts for heterogeneity in mutation rate in non-coding regions to identify regulatory driver mutations.
112. Ng, P. C. & Henikoff, S. SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.* **31**, 3812–3814 (2003).
113. Adzhubei, I. A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
114. Polak, P. *et al.* Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature* **518**, 360–364 (2015).
Shows that somatic mutation density can be predicted based on epigenomic features from the cell of origin.
115. Fu, Y. *et al.* FunSeq2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol.* **15**, 480 (2014).
116. O'Roak, B. J. *et al.* Multiplex targeted sequencing identifies recurrently mutated genes in autism spectrum disorders. *Science* **41**, 177–181 (2012).
117. Konermann, S. *et al.* Genome-scale transcriptional activation by an engineered CRISPR–Cas9 complex. *Nature* **517**, 583–588 (2014).
118. Mogno, I., Kwasnieski, J. C. & Cohen, B. A. Massively parallel synthetic promoter assays reveal the *in vivo* effects of binding site variants. *Genome Res.* **23**, 1908–1915 (2013).
119. Arnold, C. D. *et al.* Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science* **339**, 1074–1077 (2013).
120. Melnikov, A. *et al.* Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol.* **30**, 271–277 (2012).
121. Shlyueva, D., Stampfel, G. & Stark, A. Transcriptional enhancers: from properties to genome-wide predictions. *Nat. Rev. Genet.* **15**, 272–286 (2014).
122. Kwasnieski, J. C., Fiore, C., Chaudhari, H. G. & Cohen, B. A. High-throughput functional testing of ENCODE segmentation predictions. *Genome Res.* **24**, 1595–1602 (2014).
123. Singh, G. & Cooper, T. A. Minigene reporter for identification and analysis of *cis* elements and *trans* factors affecting pre-mRNA splicing. *Biotechniques* **41**, 177–181 (2006).
124. Gaildrat, P. *et al.* Use of splicing reporter minigene assay to evaluate the effect on splicing of unclassified genetic variants. *Methods Mol. Biol.* **653**, 249–257 (2010).
125. Poulos, R. C. *et al.* Systematic screening of promoter regions pinpoints functional *cis*-regulatory mutations in a cutaneous melanoma genome. *Mol. Cancer Res.* **13**, 1218–1226 (2015).
126. van de Wetering, M. *et al.* Prospective derivation of a living organoid biobank of colorectal cancer patients. *Cell* **161**, 933–945 (2015).
127. Boj, S. F. *et al.* Organoid models of human and mouse ductal pancreatic cancer. *Cell* **160**, 324–338 (2015).
128. Gao, D. *et al.* Organoid cultures derived from patients with advanced prostate cancer. *Cell* **159**, 176–187 (2014).
129. Ermann, J. & Glimcher, L. H. After GWAS: mice to the rescue? *Curr. Opin. Immunol.* **24**, 564–570 (2012).
130. Seruggia, D., Fernández, A., Cantero, M., Pelczar, P. & Montoliu, L. Functional validation of mouse tyrosinase non-coding regulatory DNA elements by CRISPR – Cas9-mediated mutagenesis. *Nucleic Acids Res.* **43**, 4855–4867 (2015).
131. Mou, H., Kennedy, Z., Anderson, D. G., Yin, H. & Xue, W. Precision cancer mouse models through genome editing with CRISPR–Cas9. *Genome Med.* **7**, 53 (2015).
132. Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546–1558 (2013).
133. The Cancer Genome Atlas Research Network. Comprehensive molecular profiling of lung adenocarcinoma. *Nature* **511**, 543–550 (2014).
134. Guenther, C. A., Tasic, B., Luo, L., Bedell, M. A. & Kingsley, D. M. A molecular basis for classic blond hair color in Europeans. *Nat. Genet.* **46**, 748–752 (2014).
135. Davoli, T. *et al.* Cumulative haploinsufficiency and triplosensitivity drive aneuploidy patterns and shape the cancer genome. *Cell* **155**, 948–962 (2013).
136. Xue, W. *et al.* A cluster of cooperating tumor-suppressor gene candidates in chromosomal deletions. *Proc. Natl Acad. Sci. USA* **109**, 8212–8217 (2012).
137. Wang, K. *et al.* Whole-genome sequencing and comprehensive molecular profiling identify new driver mutations in gastric cancer. *Nat. Genet.* **46**, 573–582 (2014).
138. Thurman, R. E. *et al.* The accessible chromatin landscape of the human genome. *Nature* **489**, 75–82 (2012).
139. Bush, W. S. & Moore, J. H. Chapter 11: genome-wide association studies. *PLoS Comput. Biol.* **8**, e1002822 (2012).
140. Chelala, C., Khan, A. & Lemoine, N. R. SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. *Bioinformatics* **25**, 655–661 (2009).
141. Dayem Ullah, A. Z., Lemoine, N. R. & Chelala, C. SNPnexus: a web server for functional annotation of novel and publicly known genetic variants (2012 update). *Nucleic Acids Res.* **40**, W65–W70 (2012).
142. Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**, e164 (2010).
143. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069–2070 (2010).
144. Perera, D. *et al.* OncoCis: annotation of *cis*-regulatory mutations in cancer. *Genome Biol.* **15**, 485 (2014).
145. Paila, U., Chapman, B. A., Kirchner, R. & Quinlan, A. R. GEMINI: integrative exploration of genetic variation and genome annotations. *PLoS Comput. Biol.* **9**, e1003153 (2013).
146. Coetzee, S. G., Rhie, S. K., Berman, B. P., Coetzee, G. A. & Noushmehr, H. FunciSNP: an R/bioconductor tool integrating functional non-coding data sets with genetic association studies to identify candidate regulatory SNPs. *Nucleic Acids Res.* **40**, e139 (2012).
147. Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40**, D930–D934 (2012).
148. Li, M. J., Wang, L. Y., Xia, Z., Sham, P. C. & Wang, J. GWAS3D: detecting human regulatory variants by integrative analysis of genome-wide associations, chromosome interactions and histone modifications. *Nucleic Acids Res.* **41**, W150–W158 (2013).
149. Macintyre, G., Bailey, J., Haviv, I. & Kowalczyk, A. is-rSNP: a novel technique for *in silico* regulatory SNP detection. *Bioinformatics* **26**, i524–i530 (2010).
150. Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res.* **22**, 1790–1797 (2012).
151. Lehmann, K. V. & Chen, T. Exploring functional variant discovery in non-coding regions with SInBaD. *Nucleic Acids Res.* **41**, e7 (2013).
152. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
153. Ritchie, G. R., Dunham, I., Zeggini, E. & Flicek, P. Functional annotation of noncoding sequence variants. *Nat. Methods* **11**, 294–296 (2014).
154. Gulko, B., Hubisz, M. J., Gronau, I. & Siepel, A. A method for calculating probabilities of fitness consequences for point mutations across the human genome. *Nat. Genet.* **47**, 276–283 (2015).
155. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).

Acknowledgements
F.D. would like to acknowledge grant IG 13562 from AIRC (Associazione Italiana per la Ricerca sul Cancro).

Competing interests statement
The authors declare no competing interests.

FURTHER INFORMATION
ANNOVAR: http://openbioinformatics.org/annovar/
CADD: http://cadd.gs.washington.edu
CIS-BP: cisbp.ccbr.utoronto.ca
DeepSEA: http://deepsea.princeton.edu
ENCODE (derived from ChIP-seq): encodenets.gersteinlab.org
ENCODE (derived from DHS): regulatorynetworks.org
ENCODE: encodeproject.org
FANTOM: fantom.gsc.riken.jp
FitCons: http://compgen.cshl.edu/fitCons/
FunciSNP: http://bioconductor.org
FunSeq: http://funseq.gersteinlab.org
FunSeq2: funseq2.gersteinlab.org
GEMINI: http://github.com/arq5x/Gemini
GENCODE: gencodegenes.org
GtRNAdb: gtrnadb.ucsc.edu
GWAS3D: http://jjwanglab.org/gwas3d
GWAVA: http://sanger.ac.uk/resources/software/gwava
HaploReg: http://compbio.mit.edu/HaploReg
International Cancer Genome Consortium: http://www.icgc.org
is-Rsnp: http://bioinformatics.research.nicta.com.au/software/is-rsnp/
JASPAR: jasper.genereg.net
miRBase: mirbase.org
MiTranscriptome: mitranscriptome.org
OncoCis: http://powcs.med.unsw.edu.au/OncoCis
RegulomeDB: http://RegulomeDB.org
Roadmap epigenomics: roadmapepigenomics.org
SeattleSeq: http://snp.gs.washington.edu/SeattleSeqAnnotation
SInBaD: http://tingchenlab.cmb.usc.edu/Sinbad
snoRNABase: www-snorna.biotoul.fr
SNPnexus: http://snp-nexus.org
The Cancer Genome Atlas: https://tcga-data.nci.nih.gov
Transfac: biobase-international.com/products
VEP: http://ensembl.org/info/docs/tools/vep

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

© 2016 Macmillan Publishers Limited. All rights reserved