OPINION

Safe harbours for the integration of new DNA in the human genome

Michel Sadelain, Eirini P. Papapetrou and Frederic D. Bushman

Abstract | Interactions between newly integrated DNA and the host genome limit the reliability and safety of transgene integration for therapeutic cell engineering and other applications. Although targeted gene delivery has made considerable progress, the question of where to insert foreign sequences in the human genome to maximize safety and efficacy has received little attention. In this Opinion article, we discuss 'genomic safe harbours' — chromosomal locations where therapeutic transgenes can integrate and function in a predictable manner without perturbing endogenous gene activity and promoting cancer.

The modification of the human genome by the stable insertion of functional transgenes and other genetic elements is of great value in biomedical research and medicine. Several diseases have now been successfully treated with gene therapy, including some haematological and metabolic disorders[1–4]. Genetically modified human cells are also valuable for the study of gene function, and for tracking and lineage analyses using reporter systems. All these applications depend on the reliable function of the introduced genes in their new environments. However, randomly inserted genes are subject to position effects and silencing, making their expression unreliable and unpredictable[5–9]. Centromeres and sub-telomeric regions are particularly prone to transgene silencing[10,11]. Reciprocally, newly integrated genes may affect the surrounding endogenous genes and chromatin, potentially altering cell behaviour or favouring cellular transformation[12]. Despite the successes of therapeutic gene transfer, there have been several cases of malignant transformation associated with insertional activation of oncogenes following stem cell gene therapy[4], emphasizing the importance of where newly integrated DNA locates.

The most common approach to stable transgene delivery in human cells makes use of retroviral vectors, consisting of either γ-retroviral vectors that are derived from murine leukaemia viruses (MLVs) or lentiviral vectors that are derived from HIV-1. These afford semi-random insertion spanning the entire genome with a high predilection for integration in the vicinity of transcribed genes, which accounts for about two-thirds of all integrations[13–15]. The malfunction of retroviral-encoded transgenes, especially their silencing[5–9], and their most dramatic effects in the context of gene therapy — clonal expansion, myelodysplasia and leukaemia[16–20] — have been extensively studied (BOX 1). It is well established that integration near or within cancer-related genes poses the greatest threat in the context of therapeutic cell engineering[4,21,22], making the avoidance of proximity to such genes a priority.

Considerable efforts are underway to prevent the two major shortcomings of semi-random DNA integration — variable transgene expression and insertional oncogenesis. These efforts rest on the design of vectors with a reduced susceptibility to position effects and silencing[23], the tethering of retroviral pre-integration complexes to selected DNA-binding proteins to restrict vector integration[24] and the design of vectors with tissue-specific expression patterns or other features that are designed to limit unwanted interactions with the host genome, including the removal of retroviral enhancers (self-inactivating (SIN) vectors)[23,25,26] and the inclusion of insulator elements[27]. A review of all these strategies[4,28] is beyond the scope of this article, but each strategy ultimately aims to minimize unwanted interactions between the inserted genetic material and the chromosomal region harbouring it.

An alternative approach is to target the genetic material to a predetermined genomic site. In recent years, several technologies have been developed and applied to human cells for the targeted delivery of foreign DNA. Targeted gene delivery exploits DNA repair mechanisms that occur in response to DNA double-strand breaks (DSBs) to introduce new genetic material[29,30]. The frequency of targeting can be increased by the introduction of DNA DSBs at the target site using specific rare-cutting endonucleases[31]. Several methodologies to induce site-specific DSBs in the targeted site are now available, including zinc-finger nucleases[32], meganucleases[33] and transcription activator-like effector (TALE) nucleases[34].

Although these tools represent an important advance, the question of where to introduce transgenes or non-coding RNAs to maximize safety and efficacy has not been comprehensively addressed. In the case of gene repair, the natural destination for the new DNA sequence is the mutant gene locus. However, the question of where to introduce new genes with reporter, suicide or selectable functions has not received much consideration. Thus, one may target integration to genes that are thought to be dispensable or, alternatively, to extragenic regions. Some permissive sites may be appropriate for some tissues or lineages but not for others. The effect of the transgene on adjacent genes should be either null or fully assessed and understood — but how to do so?

In this Opinion article, we discuss approaches for identifying and validating genomic safe harbours (GSHs). GSHs are intragenic or extragenic regions of the human genome that are able to accommodate the predictable expression of newly integrated DNA without adverse effects on the host cell or organism. A useful safe harbour must permit sufficient transgene expression to yield desired levels of the

Box 1 | **Human cancer following therapeutic gene transfer**

The most feared risk that is posed by therapeutic transgene integration is that of malignant transformation[4]. Whereas semi-random gene integration obligatorily results in insertional mutagenesis, some integrations (depending on the nature of the vector, the site of integration and the affected cell type) may cause or facilitate cancer. Alongside successful proof-of-principle studies in the realm of haematopoietic stem cell gene therapy has come the realization that the integration of γ-retroviral vectors that are driven by their long-terminal repeat (LTR) may facilitate the emergence of clonal expansion, myelodysplasia or overt leukaemia[16–20]. The development of leukaemia has been dramatically illustrated by gene therapy for X-linked severe combined immunodeficiency (SCID) and Wiskott–Aldrich syndrome (WAS), with five of 20 and one of ten patients, respectively, so far developing T cell leukaemias. A striking feature of these clonal transformations, which are all linked to the integration of an LTR-driven γ-retroviral vector in the vicinity of an oncogene, is the recurrent involvement in all but one case of the gene encoding LIM domain only protein 2 (LMO2)[16,19]. Clonal expansion has also been associated with vector-mediated transactivations of other proto-oncogenes such as *MECOM* (also known as *MDS1* and *EVI1* complex locus) and *PRDM1* (REFS 16,17). Vector integration sites are typically located just upstream of transcription start sites or in introns near the transcript 5′ end, most commonly within 50 kb of the transcriptional start site. Few, if any, long-distance interactions spanning more than 300 kb have been documented. Less commonly, proto-oncogenes can also be activated by the formation of new proteins, sometimes involving the fusion of viral and cellular sequences, or by truncating messages to remove 3′ negative regulatory sequences[20].

vector-encoded protein or non-coding RNA. A GSH also must not predispose cells to malignant transformation nor alter cellular functions. What distinguishes a GSH from a fortuitous good integration event is the predictability of outcome, which is based on prior knowledge.

Proximity to cancer-related genes

Extensive studies of the insertional activation of cancer genes provide a detailed picture of unsafe harbours for integration. Historically, one of the main techniques for identifying genes that promote transformation has been surveying retroviral integration sites in tumours in model organisms. Cases in which integration sites are recovered repeatedly near the same gene in cancer cells but not in normal cells provide evidence for the involvement of these genes in transformation. Supporting this inference, many of the proto-oncogenes that have been identified in such screens have been independently validated by other methods, such as by forcing transformation by deliberate overexpression of these proto-oncogenes[35–37].

Such studies have been extensively carried out in mice, and data collected in the Retroviral Tagged Cancer Gene Database (RTCGD)[36,37] (see Further information). Analysis shows that integration sites in tumours commonly lie near the starting point of transcription, either upstream or just within the transcription unit, often in a 5′ intron. Proviruses at these locations usually increase the rate of transcription either via promoter or via enhancer insertion[38]. For cancers that are associated with human gene therapy, insertional activation by promoter or enhancer insertion seems to be most common[16,17,19,39]. Integration sites are commonly in DNA just upstream of transcription start sites or in introns near transcript 5′ ends. Thus, the findings in humans so far parallel results from model organisms (BOX 1).

Cancer gene lists

How well can we identify regions of the genome that should be avoided to prevent transformation during human gene therapy? Various experimental approaches have been used to implicate specific genes in cancer, and these have been collected into a large, and growing, number of databases[35,36,40–44] (TABLE 1). A few cancer-related genes of special interest have been associated with integration sites in transformed cells that arise during human gene therapy (such as LIM domain only protein 2 (*LMO2*), cyclin D2 (*CCND2*), the polycomb ring finger oncogene *BMI1* and *MECOM* (also known as *MDS1* and *EVI1* complex locus)[16,17,19,39]. Extensive clinical experience with human cancers has of course implicated a large collection of mutations in human genes, as has the experimental induction of cancer in animal models[35–37]. And more complete data are certainly on the way — the Cancer Genome Atlas Project, the UK Sanger Cancer Genome Project and others are sequencing large numbers of human genomes from cancer cells and matched normal cells, so that new data on cancer-associated mutations are accumulating at a remarkable rate (see, for example, REF. 44).

Which cancer-related genes are the most important ones that should be avoided for safe gene correction? In most cases, it will not be easy to know which genes are most dangerous for the combination of disease treated, tissue involved and vector used. For this reason, we have assembled a comprehensive list for first-order screening, which is comprised of the combination of all available functionally defined cancer genes (TABLE 1). To assemble this AllOnco list, human genes and the human homologues of cancer genes from other organisms are identified and added to the collection. The final list contains 2,070 genes, or ~10% of all human genes.

This approach to cancer gene identification is problematic for several reasons. Many genes will be important in cancer in specific tissues and stages of development but benign in others. Cancer genes also certainly do not behave identically in all organisms. Sometimes the human homologues of genes identified in model organisms are not obvious, or there are one-to-many or many-to-one gene mappings between organisms. However, given the fact that cancers that arise during human gene therapy may not be identical to known human cancers, it may be useful to initially screen against as broad a list as possible. Thus, in devising safe harbour criteria, the most conservative approach is to use the most encompassing cancer gene list.

Locations for genomic safe harbours

Should GSHs lie in selected genes or in extragenic regions? Their location in selected genes assumes that certain non-essential genes can be disrupted without pathological consequences. Housekeeping genes may be attractive as potential universal GSHs because of their ubiquitous expression, but this very property argues against dispensability. Non-essential genes with a fairly broad tissue distribution may be more likely candidates. Some recent studies suggest that intragenic sites that lie within gene-rich regions can accommodate the integration of certain expression cassettes without detectable consequences, at least in some cell types[45], although the process of documenting safety is still incomplete. The alternative is to locate GSHs at extragenic sites, where expression may be more problematic, but these are areas that evolutionary considerations suggest may be more benign.

The genomic locations of endogenous retroviruses provide some clues for identifying GSHs. About 8% of the human genome

is comprised of fragments of retroviruses that integrated into the germ line in the mammalian lineage during evolution[46]. We can deduce that these integration events were not harmful by the fact that the great majority are fixed in the human germ line. The endogenous retroviruses are not randomly distributed, but are enriched outside transcription units[47], suggesting that evolutionary selection eliminated integration events within transcription units. The minority of integrated proviruses within transcription units are usually in reverse orientation relative to host gene transcription; this orientation is expected to be the least disruptive because polyadenylation and splicing signals in the provirus are in an antisense orientation and are thus inactive. It might have been that endogenous retroviruses in fact favoured *de novo* integration outside genes and in an antisense orientation within genes, but experimental analysis shows that the initial integration for two endogenous retroviruses is in fact favoured within transcription units, highlighting the importance of selective pressure to yield the observed genomic distribution[48,49]. Similar data are seen for all studied vertebrates. Thus, endogenous retrovirus biology suggests that integration outside transcription units may be most benign.

Candidate genomic safe harbours

Although no GSH has yet been fully validated, specific loci or general criteria for prospectively identifying GSHs have been proposed. Only three sites in the human genome have been used for targeted transgene addition to date: the adeno-associated virus site 1 (AAVS1), the chemokine (CC motif) receptor 5 (*CCR5*) gene locus and the human orthologue of the mouse ROSA26 locus. As discussed below, the information that is currently available regarding the safety features of these loci is too limited to qualify any of them as a GSH, and, from the available data, none seems ideal. We further discuss below our proposed criteria for extragenic GSHs, based on bioinformatic analyses of retroviral integration databases.

AAVS1. The AAVS1 site in chromosome 19 (position 19q13.42) was identified as a repeatedly recovered site of integration of wild-type AAV in the genome of cultured human cell lines that have been infected with AAV *in vitro*[50]. Because a large proportion of the human population has encountered AAV, as evidenced by detectable antibodies against some AAV serotypes, but without any discernable pathology, it was inferred that integration in AAVS1 may be innocuous[51]. However, little is known about natural AAV infection[51]. In the absence of a helper virus, AAV can establish latency. Its genome has been detected in multiple tissues (such as the muscle, spleen, liver, bone marrow, genital tract, heart, brain and kidney) in humans as well as in non-human primates. The few studies that have examined the status of the AAV genome in human and non-human primate tissues showed that it is mostly present in non-integrated (episomal) forms[52,53]. Integrated forms are more frequently found in random genomic sites than in AAVS1 in both *in vitro* and *in vivo* infected human tissues[52,54,55]. The overall frequency of integration of wild-type AAV in AAVS1 has been estimated as less than 0.5% of infectious viral genomes[56]. In fact, AAVS1 integrations after *in vivo* infection with wild-type AAV have been detected collectively in tissues from only two human subjects and three rhesus macaques[57,58]. There is, therefore, little evidence that AAV integration into AAVS1 is a biologically important part of the *in vivo* AAV replication cycle.

Efforts towards directing integration to AAVS1 include targeting by use of the viral-encoded Rep protein, as well as by homologous recombination-based strategies in human T cells, neural stem cells, embryonic stem cells (ESCs) and induced pluripotent stem cells (iPSCs)[59–64]. In the ESCs and iPSCs, the robust expression of reporter transgenes inserted into AAVS1 was observed over several days, as well as following differentiation in multiple tissues deriving from all three embryonic germ layers[60–63], similar to observations in mouse ESCs after gene integration into the orthologous site[65,66]. The widespread expression across cell types may be due to a DNAse I hypersensitive site and an insulator element contained in the AAVS1 locus that may maintain an open chromatin conformation[45,67].

Glossary

Euchromatic portion
A region of chromatin that has lighter packing than heterochromatin and that is generally considered to be richer in actively transcribed genes.

Gene trapping screen
A high-throughput approach used to report and/or inactivate the expression of multiple individual genes across the genome by introducing a reporter gene lacking a promoter (through plasmid or retroviral gene transfer). Selection for expression of the gene requires transcription from a cellular promoter.

Insulator elements
Regulatory DNA elements that create boundaries in chromatin, delineating the ranges over which neighbouring regulatory elements function. They can have enhancer-blocking activity, which prevents communication between discrete sequence elements (typically enhancers and promoters) when insulators are positioned between them, and/or barrier activity, which prevents the spread of heterochromatin.

Intergenic transcription
Transcription of chromosomal DNA sequences between known genes.

Locus control regions
A class of *cis*-acting DNA regulatory elements that confer high level, tissue-specific, site-of-integration-independent, copy number-dependent expression on linked transgenes located at ectopic chromatin sites.

Matrix attachment regions
AT-rich sequences of DNA that bind to a proteinaceous nuclear scaffold called the nuclear matrix.

Meganucleases
Sequence-specific endonucleases with long recognition sites (> 12 bp). They are naturally occurring enzymes that are harnessed as tools for targeted genome engineering by the modification of their recognition sequence.

Proviruses
The duplex DNA form of the retroviral genome linked to a cellular chromosome. The provirus is produced by reverse transcription of the RNA genome and subsequent integration into the chromosomal DNA of the host cell.

Retroviral pre-integration complexes
Complexes of viral and cellular proteins with retroviral DNA made by reverse transcription, which together are capable of integrating the viral DNA into a target DNA.

Sub-telomeric regions
Regions adjacent to the telomeres or tips of chromosomes that are often heterochromatic.

Transcription activator-like effector (TALE) nucleases
Artificial endonucleases generated by fusing a TALE DNA-binding domain to the catalytic domain of an endonuclease that introduces double-strand breaks. Similar to zinc-finger nucleases and meganucleases, TALE nucleases can also be engineered to target user-specified DNA sequences within complex genomes.

Zinc-finger nucleases
A class of synthetic proteins that are generated by fusing a zinc-finger DNA-binding domain to the cleavage domain of the FokI restriction endonuclease. The DNA-binding domain can be engineered to induce double-strand breaks in desired DNA sequences, thus facilitating site-specific homologous recombination by the endogenous DNA repair machinery and targeted editing of a genomic locus (insertion, deletion and single-base substitution).

Table 1 | **Databases of genes implicated in cancer**

Gene set*	Number of genes	Species	Description	Refs
Atlas	999	Human	This gene set is from the Atlas of genetics and cytogenetics in oncology and haematology. It lists both hybrid genes found in at least one cancer case and gene amplifications or homozygous deletions found in a significant subset of cases in a given cancer type	41
Miscellaneous	187	Multiple	This gene set is from *Retroviruses* (Cold Spring Harbor Laboratory Press), an early version of the CIS database, a list from T. Hunter, The Salk Institute, La Jolla, California, USA, and miscellaneous additions from the scientific literature	35
CAN genes	192	Human	This gene set includes 192 common genes that were mutated at significant frequency in all tumours of human breast and colorectal cancers	42
CIS (RTCGD)	593	Mouse	This gene set is from the Mouse Variation Resource and lists retroviral insertional mutagenesis in mouse haematopoietic tumours	36
Human lymphoma	38	Human	This gene set is a list of lymphoid-specific oncogenes that was compiled by M. Cavazzana-Calvo and colleagues, Hôpital Necker, Paris, France	
Sanger	452	Human	This gene set is from the Cancer Gene Census, a compilation from the scientific literature of "mutated genes that are causally implicated in oncogenesis." (REF. 43)	43
Waldman	455	Human	This gene set is from the Waldman gene database and lists cancer genes sorted by chromosomal locus and includes links to OMIM	
AllOnco	2,070	Mouse and human	This database is a master set of the seven sets described above in which all genes are converted to their human homologues	

*Gene lists and links to original sources are available at The Bushman lab cancer gene list website (see Further information). CAN, cancer; CIS, common insertion site; OMIM, Online Mendelian Inheritance in Man; RTCGD, Retroviral Tagged Cancer Gene Database.

Integration in the AAVS1 locus disrupts the gene phosphatase 1 regulatory subunit 12C (*PPP1R12C*; also known as *MBS85*), which encodes a protein with a function that is not clearly delineated. The organismal consequences of disrupting one or both alleles of *PPP1R12C* are currently unknown. No gross abnormalities or differentiation deficits were observed in human and mouse pluripotent stem cells harbouring transgenes targeted in AAVS1 (REFS 61,63,65). However, most of these studies used Rep-mediated targeting, which seems to preserve the functionality of the targeted allele and maintains the expression of *PPP1R12C* at levels that are comparable to those in non-targeted cells (possibly owing to the duplication of endogenous sequences during Rep-mediated integration)[65]. This is not the case with homologous recombination-mediated targeting, which disrupts one allele, or often both alleles, when an endonuclease is used[45,60]. A recent study showed that cryptic splice acceptor sites that may be present in some expression cassettes (within the commonly used promoters phosphoglycerate kinase (*PGK*) and eukaryotic translation elongation factor 1α (*EF1A*)) can interfere with the transcription of *PPP1R12C*, resulting in a 50% to 100% reduction in its expression levels[45]. This type of interference can be dealt with by modifying the exogenous DNA by the removal of splice and polyadenylation sites[45]. Importantly, the AAVS1 locus is extremely gene-rich (FIG. 1a), and some integrated promoters can indeed transactivate neighbouring genes, the consequence of which in different tissues is presently unknown. One study suggests that integration at AAVS1 is benign in cultured human T cells[45], but the safety of integration at this site in other cell types, as well as in laboratory animals or human subjects, remains undefined.

CCR5. *CCR5*, which is located on chromosome 3 (position 3p21.31), encodes the major co-receptor for HIV-1. The discovery that homozygosity for a naturally occurring null mutation (CCR5Δ32) confers resistance to HIV-1 infection and is not associated with any major pathology, triggered intense interest in disrupting the *CCR5* gene for HIV/AIDS therapy and prompted the development of zinc-finger nucleases that target its third exon[68,69].

Endogenous CCR5 expression is high in haematopoietic cells of T lymphocyte, monocyte and macrophage lineages but not in B lymphocytes or dendritic cells[70]. CCR5 is also expressed in neurons, microglia, endothelium and smooth muscle, albeit variably[70]. The normal functions of CCR5 remain to be fully deciphered, and potential sequelae of its disruption have not been fully evaluated. *CCR5*$^{-/-}$ mice display no gross abnormalities, but have impaired leukocyte migration and increased susceptibility to certain infections, although they are more resistant to others. Genetic association studies in humans have established that homozygous knockouts for *CCR5* have increased susceptibility to West Nile virus infection[71]. Reporter genes such as that encoding green fluorescent protein (GFP) have been knocked into this locus in cord blood CD34$^+$ haematopoietic progenitors, T cells and human ESC lines[45,72]. GFP expression was significantly lower than that afforded by the AAVS1 site in human T cells[45]. As is the case for AAVS1, the genomic locus where *CCR5* resides contains several genes, including cancer-related genes that can be dysregulated by integrated transgenes[45] (FIG. 1b).

Human ROSA26. The mouse *Rosa26* strain was derived from a retroviral gene trapping screen[73]. The trapped locus was named after the mouse strain and has become a standard locus for transgenesis in mouse ESCs. Irion *et al.*[74] identified the human ROSA26 locus by means of homology in chromosome 3 (position 3p25.3). A red fluorescent protein (RFP) reporter gene without a promoter targeted to this locus was expressed in cells of all three germ layers. Endogenous transcripts are detected in multiple adult human tissues at variable levels, but their role is presently unknown. No further studies have assessed the utility or safety of the human locus. As is the case with the two sites discussed above, the human ROSA26 locus is also located near genes that can potentially be dysregulated by transgene targeting into this locus (FIG. 1c). Thus, the efficacy and safety profile of this site is also uncertain.

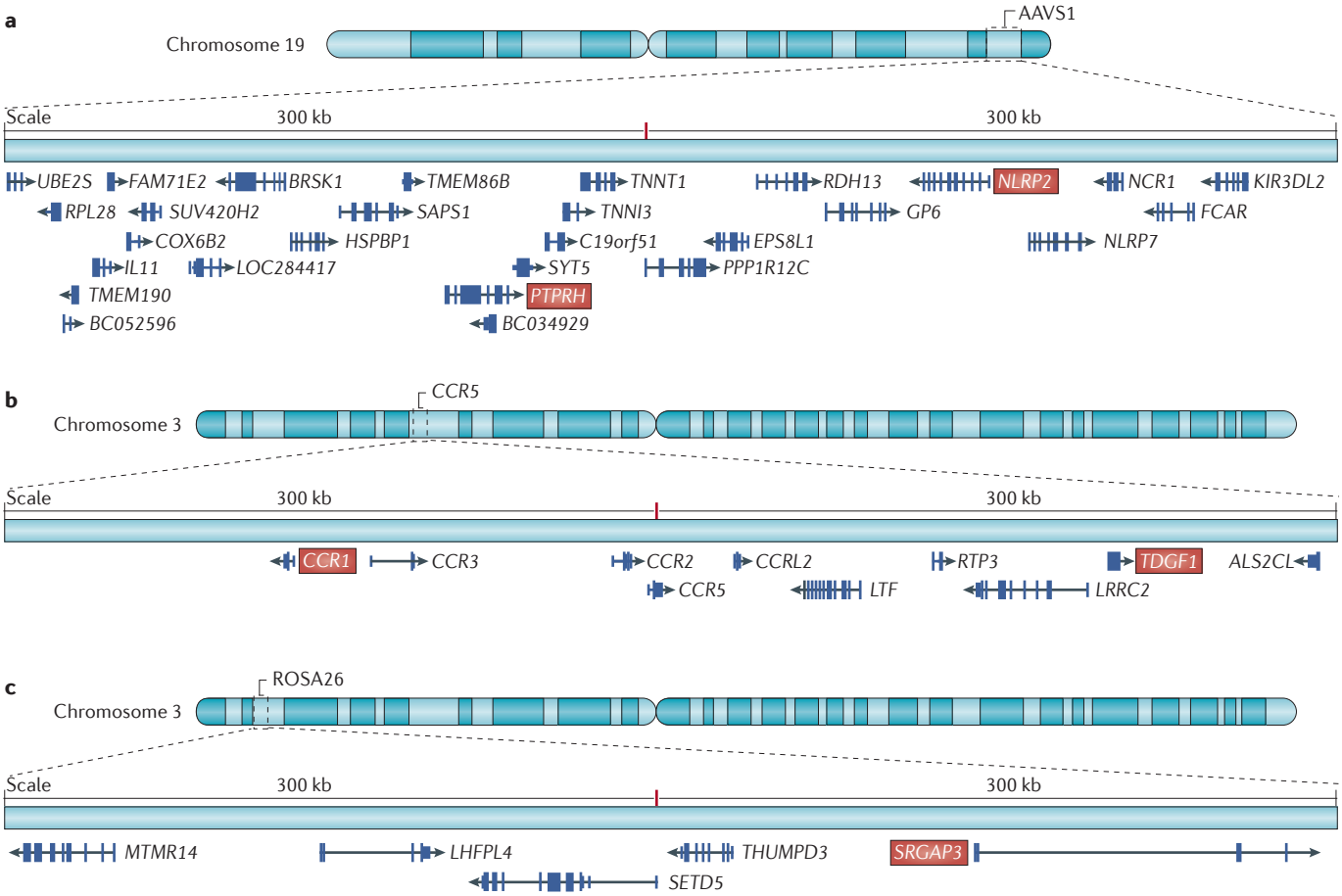

Figure 1 | **Intragenic candidate GSHs.** Chromosome ideograms (depicted in upper panels) and graphics (depicted in lower panels) depicting 300 kb of human genome on both sides of each site (shown by a vertical red line) are shown for adeno-associated virus site 1 (AAVS1) (part **a**), the chemokine (CC motif) receptor 5 (*CCR5*) gene locus (part **b**) and the human orthologue of the mouse ROSA26 locus (part **c**). Chromosome ideograms and graphics were generated with the University of California, Santa Cruz (UCSC) Genome Graphs tool. All UCSC genes present in the genomic region spanning 600 kb illustrated in the graphic are shown. Genes that are implicated in cancer are shown in red boxes. The arrows indicate the direction of transcription. GSH, genomic safe harbour.

Criteria for defining extragenic GSHs. We previously proposed five criteria (FIG. 2) to facilitate the identification of prospective GSHs[75]. These criteria aim to exclude the disruption of endogenous coding genes and ultra-conserved regions, and to minimize the possibility of long-range interactions between vector-encoded transcriptional activators and the promoters of adjacent genes, particularly cancer-related and microRNA genes[76–78]. Distances from these elements were chosen based on comparison with data from the RTCGD insertional activation database and may evolve as further information accumulates. No site that meets all of the above criteria has yet been validated, but we previously identified some candidates (FIG. 2).

There are challenges to meeting these criteria, which ultimately stem from the still incomplete functional annotation of the human genome. One issue is the incompleteness of cancer gene lists, as discussed above, so that our ability to identify dangerous sites is still evolving. Another issue is the distinction between bona fide transcription units and intergenic transcription. The roughly 3.4 billion bases of the euchromatic portion of the human genome are believed to encode ~20,000 proteins, with exons comprising about 1.5% and the transcription units about 33% of the genome. However, these numbers could still change drastically. Non-protein-coding genes are hard to detect and may be quite abundant, and several reports suggest that low-level transcription may commonly take place in what were thought to be intergenic regions[79]. It could still be discovered that almost all the euchromatic human DNA is transcribed at some level, with the recognized transcription units simply transcribed more frequently. Thus, although some safe harbour criteria aim to avoid well-documented transcription units, it remains possible that extragenic safe harbours will still be transcribed at a low level.

Validation of genomic safe harbours

A potential GSH requires functional validation, and such validations will help to refine the criteria that are shown in FIG. 2. The first validation step is to measure the effect of the integrated transgene on neighbouring gene expression in cultured cells. Although the site of integration of new DNA can be mapped to the nucleotide by a number of techniques[80–82], the study of gene expression is dependent on the availability of relevant tissues. This analysis is most effective in clonable cell types, such as pluripotent stem cells or T lymphocytes. We have previously quantified dysregulation of gene expression by transgenes inserted in human iPSCs and their erythroid progeny[75] (FIG. 2). These analyses revealed the

a

Criteria for GSH	Site 1	Site 19	Site X
Distance of > 50 kb from 5′ end of any gene	✓		✓
Distance of > 300 kb from cancer-related genes	✓		
Distance of > 300 kb from any microRNA	✓	✓	✓
Outside a gene transcription unit	✓	✓	✓
Outside of ultra-conserved regions	✓	✓	✓

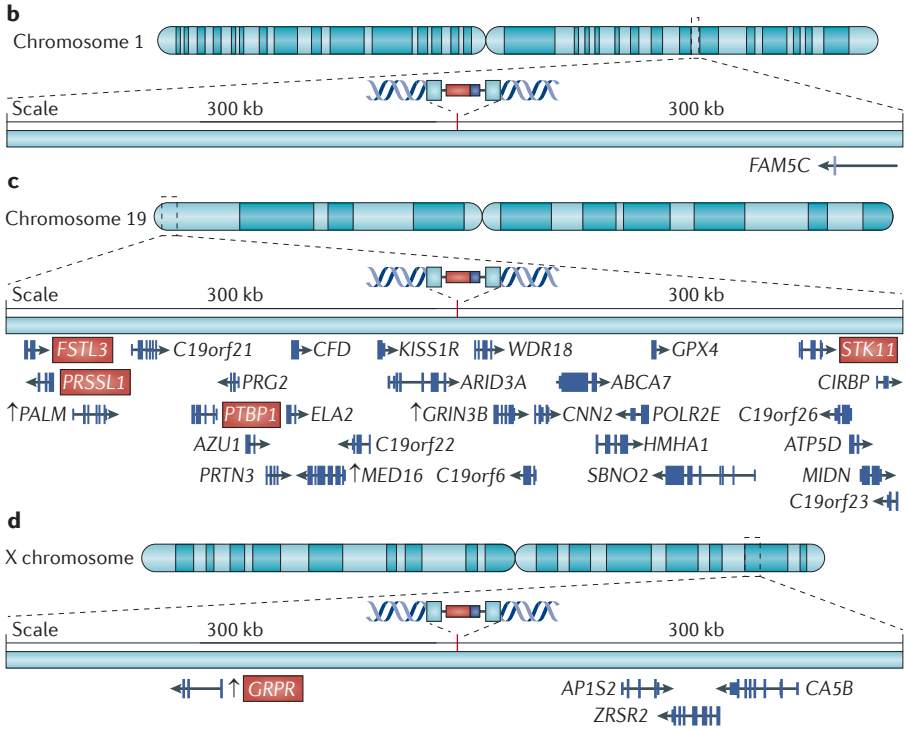

Figure 2 | **Prospective extragenic GSHs in human iPS cell clones harbouring single-copy globin transgenes.** Criteria for selecting extragenic genomic safe harbours (GSHs) based on location relative to genes and other genetic elements are shown (part **a**)[75]. Transcription units are defined by genome-wide sequencing of cDNAs, then mapping the exons back onto the genome scaffold to determine the parts of the genome that are transcribed, based on the Refseq database of human genes. Ultra conserved regions are non-coding intragenic or intergenic regions that are completely conserved in the human, mouse and rat genomes. A site (Site 1 in part **a**) that meets all five GSH criteria is shown (part **b**). Examples of sites (Site 19 and Site X in part **a**) that support globin transgene expression but that have been excluded on the basis of one or more of the GSH criteria shown in part **a** are shown (parts **c** and **d**). Chromosome ideograms (upper panels) and graphics (lower panels) depicting 300 kb of human genome on both sides of the globin vector integration were generated with the University of California, Santa Cruz (UCSC) Genome Graphs tool. The region shown in part **b** corresponds to nucleotides 187,783,272–188,383,272 (with integration at nucleotide 188,083,272 of chromosome 1); the region shown in part **c** corresponds to nucleotides 626,157–1,226,157 (with integration at nucleotide 926,157 of chromosome 19); and the region shown in part **d** corresponds to nucleotides 16,209,610–15,609,610 (with integration at nucleotide 15,909,610 of the X chromosome) of the respective chromosomes using the hg18 human genome assembly. A vertical red line depicts the position of the vector insertion. All UCSC genes present in the illustrated genomic region are shown. Genes that are implicated in cancer are shown in red boxes. The horizontal arrows indicate the direction of transcription. Vertical arrows mark genes the expression of which was found to be upregulated by the integrated vector either in the undifferentiated state or in the erythroid progeny of the induced pluripotent stem cell clone[75].

dysregulation of gene expression up to a distance of 275 kb from the vector insertion[75]. Analysing the transcriptome in mixed cell populations harbouring a variety of integration sites is challenging, but could become feasible if primary cells are transduced using a targeted approach[72,83,84]. This kind of study would still pose logistical challenges in tissue types that are difficult to access; for example, neuronal subsets or endocrine cells. The derivation of related cell types from pluripotent stem cells could make such studies feasible, provided that the gene regulation in cells that are differentiated *in vitro* mirrors that of their *in vivo* counterparts. In addition to gene expression data, epigenetic features of the targeted locus before and after transgene knock-in may be assessed, although the interpretation of the effect of different epigenetic marks on transcription may not be straightforward.

The qualification of a site as a GSH implies that it can accommodate different newly integrated transcription units. This evaluation would require the targeting or the exchange of a range of different genetic elements, including promoters, enhancers and chromatin determinants (locus control regions, matrix attachment regions and insulator elements)[6,7,23,27]. The validation of GSHs is likely to build on the contributions of many laboratories, as exemplified by studies of the murine ROSA26 locus.

In vitro studies should be extended to *in vivo* studies to assess possible tissue malfunction or transformation. This is especially important for the validation of sites that are modified by targeted nucleases, which, unlike retroviral integrants or unassisted homologous recombination, are likely to target both alleles[45,60]. Such an assessment is not equally feasible for all tissues. For example, the derivation of engraftable haematopoietic stem cells (HSCs) from human ESCs or iPSCs is not yet attainable[85], thus precluding serial transplantation studies in immunodeficient mice. *In vivo* evaluation of safe harbours could alternatively be attempted in model organisms; for example, long-term studies in transgenic mice bearing transgenes that are integrated in syntenic regions. However, differences in genome structure, the extent of synteny or biological differences between murine and human oncogenic pathways may confound such studies. Finally, some insights may be obtained through searches in the databases of common retroviral integration sites derived from patients treated with retroviral vectors (see, for example, REF. 86). This approach, however, is limited by the lack of information on transgene expression levels at these sites, and the difficulty of distinguishing between vector-driven clonal expansion and vector-independent clonal accumulation following a fortuitous initial integration.

Perspectives

A GSH is a chromosomal site where transgenes can be stably and reliably expressed in all tissues of interest without adversely affecting endogenous gene structure or expression. A strictly defined GSH has not yet been identified. The availability of such sites would be extremely useful to express reporter genes, suicide genes, selectable genes or therapeutic genes. Three intragenic sites have been proposed as GSHs (AAVS1, *CCR5* and ROSA26). All three, however, are in fairly gene-rich regions and are near genes that have been implicated in cancer. Genes that are adjacent to AAVS1 may be spared by some promoters[45], but safety validation in multiple tissues remains to be carried out. Also, the dispensability of the disrupted gene, especially after biallelic distruption, as is often the case with endonuclease-mediated targeting, remains to be functionally investigated. The identification of more sites would be highly valuable, especially at extragenic loci, for which we have proposed criteria based on studies with recombinant retroviruses (FIG. 2). We suggest some methodological principles for selecting and validating GSHs, including bioinformatics, expression arrays to query nearby genes, *in vitro*-directed differentiation or *in vivo* reconstitution assays in xenogeneic transplant models, transgenesis in syntenic regions and analyses of patient databases from individuals harbouring infectious or recombinant retroviruses. This is admittedly not a simple task, but it is one that promises important rewards for human cell engineering if true GSHs are to be validated. The development of technologies and tools for targeted integration[32–34], which has preceded the identification of appropriate target sites, will greatly benefit from studies based on screening of clonal cell lines with semi-random integrations[75]. The discovery and validation of GSHs in the human genome will ultimately benefit human cell engineering and especially stem cell therapies.

Michel Sadelain and Eirini P. Papapetrou are at the Center for Cell Engineering, Molecular Pharmacology and Chemistry Program, Memorial Sloan-Kettering Cancer Center, New York, New York 10065, USA.

Frederic D. Bushman is at the Department of Microbiology, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA.

Correspondence to M.S.
e-mail: m-sadelain@ski.mskcc.org

doi:10.1038/nrc3179
Published online 1 December 2011

1. Cartier, N. *et al.* Hematopoietic stem cell gene therapy with a lentiviral vector in X-linked adrenoleukodystrophy. *Science* **326**, 818–823 (2009).
2. Aiuti, A. *et al.* Multilineage hematopoietic reconstitution without clonal selection in ADA-SCID patients treated with stem cell gene therapy. *J. Clin. Invest.* **117**, 2233–2240 (2007).
3. Gaspar, H. B. *et al.* Hematopoietic stem cell gene therapy for adenosine deaminase-deficient severe combined immunodeficiency leads to long-term immunological recovery and metabolic correction. *Sci. Transl. Med.* **3**, 97ra80 (2011).
4. Riviere, I., Dunbar, C. & Sadelain, M. Hematopoietic stem cell engineering at a crossroads. *Blood* 17 Nov 2011 (doi:10.1182/blood-2011-09-349993).
5. Martin, D. I. & Whitelaw, E. The vagaries of variegating transgenes. *Bioessays* **18**, 919–923 (1996).
6. Kioussis, D. & Festenstein, R. Locus control regions: overcoming heterochromatin-induced gene inactivation in mammals. *Curr. Opin. Genet. Dev.* **7**, 614–619 (1997).
7. Rivella, S. & Sadelain, M. Genetic treatment of severe hemoglobinopathies: the combat against transgene variegation and transgene silencing. *Semin. Hematol.* **35**, 112–125 (1998).
8. Bestor, T. H. Gene silencing as a threat to the success of gene therapy. *J. Clin. Invest.* **105**, 409–411 (2000).
9. Ellis, J. Silencing and variegation of gammaretrovirus and lentivirus vectors. *Hum. Gene Ther.* **16**, 1241–1246 (2005).
10. Karpen, G. H. Position-effect variegation and the new biology of heterochromatin. *Curr. Opin. Genet. Dev.* **4**, 281–291 (1994).
11. Weiler, K. S. & Wakimoto, B. T. Heterochromatin and gene expression in Drosophila. *Annu. Rev. Genet.* **29**, 577–605 (1995).
12. Nusse, R., van Ooyen, A., Cox, D., Fung, Y. K. & Varmus, H. Mode of proviral activation of a putative mammary oncogene (int-1) on mouse chromosome 15. *Nature* **307**, 131–136 (1984).
13. Schroder, A. R. *et al.* HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* **110**, 521–529 (2002).
14. Wu, X., Li, Y., Crise, B. & Burgess, S. M. Transcription start regions in the human genome are favored targets for MLV integration. *Science* **300**, 1749–1751 (2003).
15. Mitchell, R. S. *et al.* Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biol.* **2**, E234 (2004).
16. HaceinBeyAbina, S. *et al.* LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. *Science* **302**, 415–419 (2003).
17. Ott, M. G. *et al.* Correction of X-linked chronic granulomatous disease by gene therapy, augmented by insertional activation of MDS1-EVI1, PRDM16 or SETBP1. *Nature Med.* **12**, 401–409 (2006).
18. Stein, S. *et al.* Genomic instability and myelodysplasia with monosomy 7 consequent to EVI1 activation after gene therapy for chronic granulomatous disease. *Nature Med.* **16**, 198–204 (2010).
19. Howe, S. J. *et al.* Insertional mutagenesis combined with acquired somatic mutations causes leukemogenesis following gene therapy of SCID-X1 patients. *J. Clin. Invest.* **118**, 3143–3150 (2008).
20. Cavazzana-Calvo, M. *et al.* Transfusion independence and HMGA2 activation after gene therapy of human beta-thalassaemia. *Nature* **467**, 318–322 (2010).
21. Kustikova, O. S. *et al.* Dose finding with retroviral vectors: correlation of retroviral vector copy numbers in single cells with gene transfer efficiency in a cell population. *Blood* **102**, 3934–3937 (2003).
22. Modlich, U. *et al.* Leukemias following retroviral transfer of multidrug resistance 1 (MDR1) are driven by combinatorial insertional mutagenesis. *Blood* **105**, 4235–4246 (2005).
23. Chang, A. H. & Sadelain, M. The genetic engineering of hematopoietic stem cells: the rise of lentiviral vectors, the conundrum of the ltr, and the promise of lineage-restricted vectors. *Mol. Ther.* **15**, 445–456 (2007).
24. Gijsbers, R. *et al.* LEDGF hybrids efficiently retarget lentiviral integration into heterochromatin. *Mol. Ther.* **18**, 552–560 (2010).
25. Modlich, U. *et al.* Cell-culture assays reveal the importance of retroviral vector design for insertional genotoxicity. *Blood* **108**, 2545–2553 (2006).
26. Montini, E. *et al.* The genotoxic potential of retroviral vectors is strongly modulated by vector design and integration site selection in a mouse model of HSC gene therapy. *J. Clin. Invest.* **119**, 964–975 (2009).
27. Emery, D. W. The use of chromatin insulators to improve the expression and safety of integrating gene transfer vectors. *Hum. Gene Ther.* **22**, 761–774 (2011).
28. Persons, D. A. & Baum, C. Solving the problem of gamma-retroviral vectors containing long terminal repeats. *Mol. Ther.* **19**, 229–231 (2011).
29. Capecchi, M. R. Gene targeting in mice: functional analysis of the mammalian genome for the twenty-first century. *Nature Rev. Genet.* **6**, 507–512 (2005).
30. Moynahan, M. E. & Jasin, M. Mitotic homologous recombination maintains genomic stability and suppresses tumorigenesis. *Nature Rev. Mol. Cell Biol.* **11**, 196–207 (2010).
31. Jasin, M. Genetic manipulation of genomes with rare-cutting endonucleases. *Trends Genet.* **12**, 224–228 (1996).
32. Porteus, M. H. & Carroll, D. Gene targeting using zinc finger nucleases. *Nature Biotechnol.* **23**, 967–973 (2005).
33. Paques, F. & Duchateau, P. Meganucleases and DNA double-strand break-induced recombination: perspectives for gene therapy. *Curr. Gene Ther.* **7**, 49–66 (2007).
34. Boch, J. TALEs of genome targeting. *Nature Biotechnol.* **29**, 135–136 (2011).
35. Rosenberg, N., Jolicouer, P., Coffin, J. M., Hughes, S. H. & Varmus, H. E. in *Retroviruses* 475–586 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, USA, 1997).
36. Akagi, K., Suzuki, T., Stephens, R. M., Jenkins, N. A. & Copeland, N. G. RTCGD: retroviral tagged cancer gene database. *Nucleic Acids Res.* **32**, D523–D527 (2004).
37. Kim, R. *et al.* Genome-based identification of cancer genes by proviral tagging in mouse retrovirus-induced T-cell lymphomas. *J. Virol.* **77**, 2056–2062 (2003).
38. Kohn, D. B., Sadelain, M. & Glorioso, J. C. Occurrence of leukaemia following gene therapy of X-linked SCID. *Nature Rev. Cancer* **3**, 477–488 (2003).
39. HaceinBeyAbina, S. *et al.* Efficacy of gene therapy for X-linked severe combined immunodeficiency. *N. Engl. J. Med.* **363**, 355–364 (2010).
40. Higgins, M. E., Claremont, M., Major, J. E., Sander, C. & Lash, A. E. CancerGenes: a gene selection resource for cancer genome projects. *Nucleic Acids Res.* **35**, D721–D726 (2007).
41. Huret, J. L., Minor, S. L., Dorkeld, F., Dessen, P. & Bernheim, A. Atlas of genetics and cytogenetics in oncology and haematology, an interactive database. *Nucleic Acids Res.* **28**, 349–351 (2000).
42. Sjoblom, T. *et al.* The consensus coding sequences of human breast and colorectal cancers. *Science* **314**, 268–274 (2006).
43. Futreal, P. A. *et al.* A census of human cancer genes. *Nature Rev. Cancer* **4**, 177–183 (2004).
44. The Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615 (2011).
45. Lombardo, A. *et al.* Site-specific integration and tailoring of cassette design for sustainable gene transfer. *Nature Methods* **8**, 861–869 (2011).
46. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
47. Smit, A. F. Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.* **9**, 657–663 (1999).
48. Brady, T. *et al.* Integration target site selection by a resurrected human endogenous retrovirus. *Genes Dev.* **23**, 633–642 (2009).
49. Barr, S. D., Leipzig, J., Shinn, P., Ecker, J. R. & Bushman, F. D. Integration targeting by avian sarcoma-leukosis virus and human immunodeficiency virus in the chicken genome. *J. Virol.* **79**, 12035–12044 (2005).
50. Kotin, R. M., Linden, R. M. & Berns, K. I. Characterization of a preferred site on human chromosome 19q for integration of adeno-associated virus DNA by non-homologous recombination. *EMBO J.* **11**, 5071–5078 (1992).
51. Henckaerts, E. & Linden, R. M. Adeno-associated virus: a key to the human genome? *Future Virol.* **5**, 555–574 (2010).
52. Schnepp, B. C., Jensen, R. L., Chen, C. L., Johnson, P. R. & Clark, K. R. Characterization of adeno-associated virus genomes isolated from human tissues. *J. Virol.* **79**, 14793–14803 (2005).

53. Gao, G. *et al.* Adeno-associated viruses undergo substantial evolution in primates during natural infections. *Proc. Natl Acad. Sci. USA* **100**, 6081–6086 (2003).
54. Drew, H. R., Lockett, L. J. & Both, G. W. Increased complexity of wild-type adeno-associated virus-chromosomal junctions as determined by analysis of unselected cellular genomes. *J. Gen. Virol.* **88**, 1722–1732 (2007).
55. Huser, D. *et al.* Integration preferences of wildtype AAV-2 for consensus rep-binding sites at numerous loci in the human genome. *PLoS Pathog* **6**, e1000985 (2010).
56. McCarty, D. M., Young, S. M., Jr. & Samulski, R. J. Integration of adeno-associated virus (AAV) and recombinant AAV vectors. *Annu. Rev. Genet.* **38**, 819–845 (2004).
57. Mehrle, S., Rohde, V. & Schlehofer, J. R. Evidence of chromosomal integration of AAV DNA in human testis tissue. *Virus Genes* **28**, 61–69 (2004).
58. Hernandez, Y. J. *et al.* Latent adeno-associated virus infection elicits humoral but not cell-mediated immune responses in a nonhuman primate model. *J. Virol.* **73**, 8549–8558 (1999).
59. DeKelver, R. C. *et al.* Functional genomics, proteomics, and regulatory DNA analysis in isogenic settings using zinc finger nuclease-driven transgenesis into a safe harbor locus in the human genome. *Genome Res.* **20**, 1133–1142 (2010).
60. Zou, J. *et al.* Oxidase-deficient neutrophils from X-linked chronic granulomatous disease iPS cells: functional correction by zinc finger nuclease-mediated safe harbor targeting. *Blood* **117**, 5561–5572 (2011).
61. Ramachandra, C. J. *et al.* Efficient recombinase-mediated cassette exchange at the AAVS1 locus in human embryonic stem cells using baculoviral vectors. *Nucleic Acids Res.* **39**, e107 (2011).
62. Smith, J. R. *et al.* Robust, persistent transgene expression in human embryonic stem cells is achieved with AAVS1-targeted integration. *Stem Cells* **26**, 496–504 (2008).
63. Yang, L. *et al.* Human cardiovascular progenitor cells develop from a KDR+ embryonicstemcell-derived population. *Nature* **453**, 524–528 (2008).
64. Hockemeyer, D. *et al.* Efficient targeting of expressed and silent genes in human ESCs and iPSCs using zinc-finger nucleases. *Nature Biotechnol.* **27**, 851–857 (2009).
65. Henckaerts, E. *et al.* Site-specific integration of adeno-associated virus involves partial duplication of the target locus. *Proc. Natl Acad. Sci. USA* **106**, 7571–7576 (2009).
66. Dutheil, N. *et al.* Characterization of the mouse adeno-associated virus AAVS1 ortholog. *J. Virol.* **78**, 8917–8921 (2004).
67. Ogata, T., Kozuka, T. & Kanda, T. Identification of an insulator in AAVS1, a preferred region for integration of adeno-associated virus DNA. *J. Virol.* **77**, 9000–9007 (2003).
68. Liu, R. *et al.* Homozygous defect in HIV-1 coreceptor accounts for resistance of some multiply-exposed individuals to HIV-1 infection. *Cell* **86**, 367–377 (1996).
69. Perez, E. E. *et al.* Establishment of HIV-1 resistance in CD4+ T cells by genome editing using zinc-finger nucleases. *Nature Biotechnol.* **26**, 808–816 (2008).
70. Rottman, J. B. *et al.* Cellular localization of the chemokine receptor CCR5. Correlation to cellular targets of HIV-1 infection. *Am. J. Pathol.* **151**, 1341–1351 (1997).
71. Lim, J. K., Glass, W. G., McDermott, D. H. & Murphy, P. M. CCR5: no longer a "good for nothing" gene--chemokine control of West Nile virus infection. *Trends Immunol.* **27**, 308–312 (2006).
72. Lombardo, A. *et al.* Gene editing in human stem cells using zinc finger nucleases and integrase-defective lentiviral vector delivery. *Nature Biotechnol.* **25**, 1298–1306 (2007).
73. Zambrowicz, B. P. *et al.* Disruption of overlapping transcripts in the ROSA beta geo 26 gene trap strain leads to widespread expression of beta-galactosidase in mouse embryos and hematopoietic cells. *Proc. Natl Acad. Sci. USA* **94**, 3789–3794 (1997).
74. Irion, S. *et al.* Identification and targeting of the ROSA26 locus in human embryonic stem cells. *Nature Biotechnol.* **25**, 1477–1482 (2007).
75. Papapetrou, E. P. *et al.* Genomic safe harbors permit high beta-globin transgene expression in thalassemia induced pluripotent stem cells. *Nature Biotechnol.* **29**, 73–78 (2011).
76. Li, Q., Peterson, K. R., Fang, X. & Stamatoyannopoulos, G. Locus control regions. *Blood* **100**, 3077–3086 (2002).
77. Bejerano, G. *et al.* Ultraconserved elements in the human genome. *Science* **304**, 1321–1325 (2004).
78. Fraser, P. Transcriptional control thrown for a loop. *Curr. Opin. Genet. Dev.* **16**, 490–495 (2006).
79. Gingeras, T. R. Origin of phenotypes: genes and transcripts. *Genome Res.* **17**, 682–690 (2007).
80. Ciuffi, A. *et al.* Methods for integration site distribution analyses in animal cell genomes. *Methods (San Diego, Calif.)* **47**, 261–268 (2009).
81. Gabriel, R. *et al.* Comprehensive genomic access to vector integration in clinical gene therapy. *Nature Med.* **15**, 1431–1436 (2009).
82. Brady, T. *et al.* A method to sequence and quantify DNA integration for monitoring outcome in gene therapy. *Nucleic Acids Res.* **39**, e72 (2011).
83. Ogino, H., McConnell, W. B. & Grainger, R. M. High-throughput transgenesis in Xenopus using I-SceI meganuclease. *Nature Protoc.* **1**, 1703–1710 (2006).
84. Cermak, T. *et al.* Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic Acids Res.* **39**, e82 (2011).
85. Papapetrou, E. P. & Sadelain, M. Reconstructing blood from induced pluripotent stem cells. *F1000 Med. Rep.* **2**, 44 (2010).
86. Wang, G. P. *et al.* Dynamics of gene-modified progenitor cells analyzed by tracking retroviral integration sites in a human SCID-X1 gene therapy trial. *Blood* **115**, 4356–4366 (2010).

Acknowledgements
The authors thank N. Malani, S. Roth and A. Bailey for help with the Cancer Gene database. This work was supported by US NIH grants CA059350, HL053750, DK087923, AI052845 and AI082020 and by the New York State Stem Cell Science (NYSTEM) grant N08T060.

Competing interests statement
The authors declare no competing financial interests.

FURTHER INFORMATION
Michel Sadelain's homepage: http://www.mskcc.org/research/lab/michel-sadelain
Atlas of Genetics and Cytogenetics in Oncology and Haematology: http://atlasgeneticsoncology.org
Cancer Gene Census: http://www.sanger.ac.uk/genetics/CGP/Census/
Retroviral Tagged Cancer Gene Database (RTCGD): http://variation.osu.edu/rtcgd
The Bushman lab cancer gene list: http://microb230.med.upenn.edu/protocols/cancergenes.html
Waldman gene database: http://cc.ucsf.edu/people/waldman/GENES/completechroms.html

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

OPINION

Programmed cell removal: a new obstacle in the road to developing cancer

Mark P. Chao, Ravindra Majeti and Irving L. Weissman

Abstract | The development of cancer involves mechanisms by which aberrant cells overcome normal regulatory pathways that limit their numbers and their migration. The evasion of programmed cell death is one of several key early events that need to be overcome in the progression from normal cellular homeostasis to malignant transformation. Recently, we provided evidence in mouse and human cancers that successful cancer clones must also overcome programmed cell removal. In this Opinion article, we explore the role of programmed cell removal in both normal and neoplastic cells, and we place this pathway in the context of the initiation of programmed cell death.

The progression from a normal cell to a fully malignant cancer cell involves a number of steps that result in the selection for poorly regulated self-renewal, deregulated proliferation, successful competition for growth-promoting niches, inhibition of differentiation, promotion of invasion, prolonged proliferative lifespan related to telomere maintenance and, importantly, survival. Survival includes avoiding programmed cell death (including, apoptosis, senescence and autophagy), which is activated by several pathways, and many cancers have mutations that enable its evasion[1,2]. Survival also involves an escape from cell death that is imposed by the innate and adaptive immune responses.

Recently, we showed that a novel mechanism — programmed cell removal by macrophages, which are a part of the innate immune system — can also regulate cancer cell survival. Programmed cell removal is a key mechanism that links programmed cell death to the removal of the dying cell. Given that it occurs before the final steps of apoptosis, it could prevent the release of pro-inflammatory signals from dying cells into the surrounding tissue. Although it seems probable that programmed cell death and programmed cell removal are triggered by common signals in a cell, we have shown that neutrophils that are prevented from undergoing programmed cell death by the enforced expression of BCL-2