MTOpen

Retroviral Integrations in Gene Therapy Trials

Luca Biasco[1], Cristina Baricordi[1] and Alessandro Aiuti[1,2]

[1]San Raffaele Telethon Institute for Gene Therapy (HSR-TIGET), Milan, Italy; [2]University of Rome Tor Vergata, Rome, Italy

γ-Retroviral and lentiviral vectors allow the permanent integration of a therapeutic transgene in target cells and have provided in the last decade a delivery platform for several successful gene therapy (GT) clinical approaches. However, the occurrence of adverse events due to insertional mutagenesis in GT treated patients poses a strong challenge to the scientific community to identify the mechanisms at the basis of vector-driven genotoxicity. Along the last decade, the study of retroviral integration sites became a fundamental tool to monitor vector–host interaction in patients overtime. This review is aimed at critically revising the data derived from insertional profiling, with a particular focus on the evidences collected from GT clinical trials. We discuss the controversies and open issues associated to the interpretation of integration site analysis during patient's follow up, with an update on the latest results derived from the use of high-throughput technologies. Finally, we provide a perspective on the future technical development and on the application of these studies to address broader biological questions, from basic virology to human hematopoiesis.

Received 28 September 2011; accepted 6 December 2011; published online 17 January 2012. doi:10.1038/mt.2011.289

INTRODUCTION

Retroviral vectors (RVs) have been widely used to deliver therapeutic genes in the context of gene therapy (GT) clinical applications for monogenic disorders, cancer, and infectious diseases providing stable and efficient expression of the transgene to patients.[1–5] Although clinical trials for primary immunodeficiencies have clearly demonstrated the therapeutic benefit of retroviral-based approaches[4–11] the field of GT was significantly impacted by the sudden occurrence of severe adverse events linked to insertional mutagenesis due to aberrant vector-on-host interactions.[7,12–15] Thus, insertional profiling, aimed at identifying vector integration sites and studying their potential impact in preclinical and clinical samples, has become an important tool to evaluate the global safety profile of clinical trials.[12,13,16–22] Several groups including ours are currently working to improve insertion retrieval techniques and analysis in order to collect relevant information from integration site distribution. The results of these studies are exploited for the design of vectors and the set up of gene transfer protocol with the aim to couple efficient and regulated transgene expression with a safe insertion profile.[1] In parallel with these efforts now is the appropriate time to retrospectively analyze in details the data collected in the past years from vector integration studies with the goal to provide a proper rendering of the biological impact of retroviral insertions. This review aims to re-evaluate the information available from clinical insertional profiling of RVs with a particular eye on the controversial findings and open issues impacting the common interpretation of data derived from integration sites analyses.

GENOTOXICITY IN GT CLINICAL TRIALS

The potential genotoxicity of RV came to the attention of the GT community after the occurrence of severe adverse events in two clinical trials for X-linked severe combined immunodeficiency (SCID-X1). In these patients autologous bone marrow hematopoietic stem/progenitor cell were transduced *ex vivo* with a Moloney leukemia virus (MLV)-derived vector carrying the common γ-subunit of interleukin-2 receptor-γ (IL2RG) under the long terminal repeat (LTR) promoter. The cells were reinfused back without any preparative conditioning.[7,8] Overall, 17 out of the 20 SCID-X1 patients enrolled in both clinical trials benefited from GT, with sustained transgene expression and immunological reconstitution.[7,23,24] However, the success of SCID-X1 GT was mitigated by the occurrence of serious adverse events.[13] Four patients in the French trial and one in the English trial developed clonal T-cell proliferation that became evident 2–6 years after treatment. Characterization of the leukemic clones revealed the presence of integration sites in proximity of SPAG6, CCND2, and LMO2 genes, and uncovered a significant LMO2 overexpression in the transformed cells.[12] The aberrant T-cell proliferation in these patients was associated with the transactivation of this proto-oncogene by vector enhancer sequences present on the LTR.[16]

LMO2 is normally expressed in hematopoietic stem cells (HSCs) and very early T-cell precursors while it is usually down-regulated upon differentiation and its locus is involved in chromosomal translocation in cases of acute T-cell leukemia (T-ALL).[25] Although vector integrations close to LMO2 gene were found to reside within FRA11E, a common fragile site[26] and chromosomal aberrations were detected in the expanded clones, the phenotype of such clones could not be associated with T-ALL.[13] In addition, no overexpression of IL2RG or constitutive activation of downstream signaling molecules were observed in leukemic T cells from patients. Gain-of-function mutations of the transgene were also excluded by sequencing of the integrated provirus.[12]

Besides these findings, the causative role of vector-mediated transactivation on proto-oncogenes is still controversial. Indeed, evidence supports the hypothesis that RV-driven IL2RG expression

Correspondence: *Alessandro Aiuti, Clinical Research Coordinator, HSR-TIGET, Via Olgettina 58, 20132 Milano, Italy. E-mail: aiuti.alessandro@hsr.it*

may synergistically cooperate with oncogenic transformation, possibly influencing T-cell differentiation.[27,28] In addition, it has been suggested that, at least in murine models of SCID-X1, the non-regulated expression of this transgene could by itself drive to leukemogenesis without any additional integration-related influence,[29] even if, in other studies, vector-mediated IL2RG expression was not able to affect normal T-cell development.[30,31] Other oncogenic factors may reside on the disease-related clonal kinetics where abnormal population of lymphoid progenitor cells arrested in their differentiation path may accumulate additional mutations in the bone marrow during lymphoid reconstitution. A recent report raised an alternative possibility that lymphomagenesis in SCID-X1 could have been independent from both insertional mutagenesis and IL2RG overexpression suggesting the existence of other ill-defined risk factors for oncogenesis, including replicative stress.[32] The recent description of a leukemia case from the Wiskott–Aldrich syndrome (WAS) GT trial associated to a LMO2 insertional activation raised again the interest in the specific interaction of RVs and this particular locus.[33,34]

Another insertional mutagenesis effect of RV leading to aberrant clonal proliferation in patients arose in the context of a GT trial for chronic granulomatosis disease (CGD). CGD is a rare inherited immunodeficiency caused by a functional defect in the microbial killing activity of phagocytes. The clinical protocol used in this trial was based on *ex vivo* retroviral gene transfer of gp91phox complementary DNA into mobilized peripheral blood $CD34^+$ cells through a γ-RV.[35] Autologous cell transfer of the gene-corrected $CD34^+$ cells was able to restore enzymatic activity in transduced phagocytes as testified by the clinical resolution of bacterial and fungal infections in three treated patients.[14,36] However, the high level of correction achieved, resulted from an unexpected *in vivo* expansion of gene-corrected myeloid cells showing clusters of vector integrations in the MDS-EVI1, PRDM16, and SETBP1 loci.[14] The relative contribution of each single clone carrying these integrants changed overtime, nonetheless maintaining a sort of stability in the most dominant clones.

The efficacy of GT was lost upon a progressive decrease of transgene expression on dominant clones due to methylation of the viral promoter.[15] Concomitantly, the three treated patients developed a myelodysplasia with monosomy 7 and one of them died of overwhelming sepsis 27 months after GT.[37] It has been recently hypothesized that the overexpression of EVI1 gene by insertional activation could have led to disruption of normal centrosome duplication resulting in genomic instability, monosomy 7, and clonal progression towards aberrant expansion and myelodysplasia.[15] In this case, the strong enhancer activity of the spleen focus-forming virus (SFFV) LTR may have further favored the transactivation of EVI1 gene to oncogenic expression levels.

Indeed, another GT trial for CGD, based on the use of a γ-retroviral construct not carrying the SFFV promoter element, did not show any sign of clonal outgrowth with evidence of normal hematopoiesis maintained overtime in all the patients treated.[38] Thus, some of the different outcomes among the GT trials may be related to the strength of LTR enhancer sequences included in the vector constructs. It is also possible that the unregulated expression of gp91phox transgene particularly in the hematopoietic stem/progenitor cell compartment could have contributed to DNA damage by reactive oxygen species production,[36] making desirable further improvement in vector design. On this line, animal models may be useful for studying normal hematopoietic dynamics independently from transgene and disease background. Indeed, Calmels *et al.* were able to detect a strong common insertion site (CIS) in MDS-EVI1 locus in myeloid lineages not leading to clonal imbalances years after infusion in rhesus macaque, through the use of an MLV vector carrying a marker "neutral" transgene in a disease-independent setting.[39]

To overcome some of the potential issues associated with integrating vectors, self-inactivating retroviral and lentiviral constructs have been developed which should carry a safer profile as compared to γ-RVs.[40,41] These constructs are specifically designed to reduce the probability of vector-mediated transactivation of neighboring genes by the elimination of the enhancer sequences from viral LTRs. The tendency of lentiviral vectors to integrate inside transcriptional units but not in promoter regions, differently from γ-RVs,[42,43] could play a role in reducing their cellular transformation potential. On this line, a direct comparison between SIN-MLV and MLV with full-length LTR should allow dissecting the influences of vector distribution and of the activity of vector enhancer sequences on genotoxicity potential of RV constructs.

Clinical applications of lentiviral vector constructs was first achieved for the treatment of X-linked adrenoleukodystrophy (ALD)[11] and β-thalassemia,[41] and more recently for metachromatic leukodystrophy and WAS.[1] The first two studies showed clear therapeutic benefit in the few patients treated, leading to the arrest of disease progression in ALD patients and transfusion independence in a thalassemia patient. However, in the latter trial the appearance of a self-limiting dominant clone in the myeloid compartment of a β(E)/β(0)-thalassemia individual was of some concerns. In this clone, a transcriptional activation of HMGA2 occurred due to vector-mediated generation of a truncated transcript.[39] Indeed, a recent study in mice carrying a truncated form of HMGA2 showed how an overexpression of the transcript is associated with the development of proliferative hematopoiesis and clonal expansion.[44] Nevertheless, in the absence of clinical evidences supporting the existence of a preleukemic state or a significant hematopoietic imbalance, this finding could still be interpreted as a stochastic event or as the result of a "benign" clonal expansion. A long-term follow up of the expanded clone in the thalassemia patient will be important to exclude the generation of potential clonal aberrancies in the future.

On the other hand, the GT approaches for adenosine deaminase-severe combined immunodeficieny (ADA-SCID) developed in the last years constitute to date an example of disease correction by γ-RV in absence of long-term vector-driven genotoxicity.[10,45] Over the last 10 years, more than 30 ADA-SCID patients have been treated with GT in different centers displaying in most cases significant benefits from the treatment both at metabolic and immunological levels and, importantly, not showing any case of leukemia.[46] Data from these trials could provide information on a more neutral interaction between RVs and host genome, although the selective advantages of gene-corrected lymphocytes in ADA-SCID should be taken into account as a bias to the general distribution of RV insertions overtime.

The high incidence of insertional mutagenesis events in SCID-X1, CGD and recently WAS trials has raised the attention of the GT community to the importance of performing integration site analysis and has called for immediate investigation of the insertional profile of RVs in clinical applications. The presence of clusters of insertions *in vivo*, the detection of integration sites in proximity of proto-oncogenes, the contribution of each integrant to the pool of transduced cells are now among the main issues to be addressed for the safety assessment and follow up of GT clinical trials. A careful review of data collected from integration site analysis in GT trials should allow delving deeper into each of these topics with the goal to lead to a more precise interpretation of the results of GT clinical approaches.

CIS

Early studies exploited oncogenicity mediated by retroviruses to tag specific insertion sites correlated with experimentally induced tumors in animal models.[47–50] Upon aberrant selection of transduced cells, retroviral integrations were recurrently located and clustered in specific regions called CIS. These loci were defined on the basis of specific frequency of insertions into given genomic windows (e.g., more than two independent insertions in 50 kb, more than three insertions in 100 kb, more than four in 200 kb etc.).

According to these observations, studies were designed to analyze the incidence of CIS detection *in vivo* considering this as a fundamental parameter for measuring the safety of GT protocols. In this regard, assessing the genotoxic potential of retroviral versus lentiviral vectors Montini *et al.* were able to show that acceleration of tumor onset in tumor-prone mice was associated with the detection of CIS in transformed cells.[40] Indeed, the early integration site profiling from GT clinical trials had already showed clusters of insertions *in vivo* from GT treated patients unveiling in some cases a strong correlation with the occurrence of severe adverse events.[17,21] As above mentioned, a number of insertions into the MDS-EVI1 or PRDM16 locus were found in strongly dominant myeloid clones from the two chronic CGD GT treated patients.[14] The LMO2 and CCND2 loci were instead related with the development of leukemia cases in a total of five patients from two GT trials for SCID-X1, being CIS in both the trials.[16,18] Similar findings were recently shown in the context of the clinical trial for WAS[6] in which it was reported a high incidence of vector integrations in these loci. This data supports a possible link between the presence of CIS and an increased risk for patients of developing aberrant clonal selection *in vivo*.

However, when an attempt was made in the French SCID-X1 trial to stratify the patients in individuals who developed leukemia versus those that did not, according to their integration profile,[17] the authors commented that "it was not possible to distinguish retroviral insertion sites in patients with lymphoproliferation from those without and CIS of third order or higher were spread over these two groups of patients". Moreover, CIS were even conserved among diverse GT clinical trials with different outcomes. The very same regions involved by genotoxic integrations in the SCID-X1 trials were also found at the same frequency in the ADA-SCID GT without leading to any abnormal expansion in patients.[19] In the same study, CIS were also detected with comparable percentages in transduced hematopoietic stem/progenitor cell before infusion in patients while another study confirmed the presence of insertional hotspots in transduced human hematopoietic progenitors *in vitro* just after transduction.[51] Our group recently unveiled the epigenetic features at the basis of these *in vitro* insertional preferences showing that they are cell-specific and physiologically maintained in patients even many years after GT in absence of adverse events.[20]

Following this line, a possible explanation for the detection of CIS could depend on the intrinsic biases of the RV vectors that integrate in favored genomic loci at the time of entry into the target cell. Therefore, in the context of GT trials a proper comparison between *in vitro* and *in vivo* data from GT patients is fundamental to draw conclusions on integration site selection. The occurrence of a clonal skewing *in vitro*, due to vector enhancer\promoter activity before reinfusion, could give rise to potential biases, but the short culture time (around 3–4 days in most of the protocols) makes this event unlikely. Other potential caveats of integration site analysis are related to the heterogeneous composition of transduced $CD34^+$ cells of which long-term HSC represents only a small fraction. This cell heterogeneity may impact both the vector distribution *in vitro* and the detection of CIS *in vivo* due to the vector effects on specific progenitors, whose proportion may vary among the trials.

Still, some of the CIS are specifically detected only after *in vivo* selection, while are not detectable *in vitro* at the time of transduction. Indeed, an alternative mechanism leading to CIS identification *in vivo* (not related to vector genotoxicity) could be linked to the transgene activity. One could speculate that expression variegation, due to positional effects, could confer different survival potentials to different clones. In this respect transcriptionally active regions could influence transgene expression and provide selective advantages to certain gene-corrected cells. As a consequence, a higher proportion of clones carrying integrations in these loci could be found *in vivo* overtime with the appearance of "benign" CIS on the integration profiles of GT patients. We recently suggested that these nongenotoxic CIS could be detected in the ADA-SCID GT where gene-corrected cells expressing higher levels of the ADA transgene, possibly due to the high transcriptional activity of the regions hosting the integration sites, have a selective advantage *in vivo* through better detoxification from ADA substrates.[20]

In the attempt to dissect the different origins of CIS, a recent work exploiting datasets of integration sites from experimental models and an ALD lentiviral GT trial[11] suggested that the distribution of integration sites along the CIS could be predictive of their genotoxic potential.[52] The identification of "sharp peaks" of insertions targeting a single gene has been suggested to constitute the "worst" CIS configuration. Another study recently showing an integrated analysis of >7,000 insertion sites previously retrieved in the context of multiple clinical approaches, revealed the presence of shared CIS among the trials and pointed at a restricted number of specific loci as preferential targets for retroviral integrations *in vitro* and *in vivo*.[53] These types of bioinformatic approaches with the meta-analysis of big insertion site databases available in the literature will be of help in the future to unveil the real nature of CIS detection in GT applications.

PROTO-ONCOGENIC HITS

For safety assessment of GT, the ontology and functional characteristic of genes close to the insertion sites represent an important parameter. The *in vivo* detection of vector integrations in proximity of genes involved in growth control or associated with transformation events is generally considered a hint of a potentially aberrant clonal selection. Indeed, early retroviral tagging experiments in mice showed how integrations isolated from tumor cells were often located in proximity of growth-promoting genes,[48] thus, providing a number of candidate proto-oncogenes now listed in the retroviral tagged cancer gene database (RTCGD, http://rtcgd.ncifcrf.gov/).

More recently, different groups have also observed the association of insertions in proximity of particular gene functional categories (such as cell-cycle control, apoptosis signaling or transcriptional regulation) with clonal dominance, in the context of *in vitro* and *in vivo* genotoxicity assays.[40,54,55] A list of genes derived from *in vitro* clonal dominance assays (insertional dominance database) was built as an additional reference for vector biosafety studies in human GT.[55] On the basis of this information, it was generally assumed that a high incidence of vector integrations in these regions can be associated to an increased risk of developing aberrant clonal selection *in vivo*.

However, one should be reminded that most of the so-called "growth-promoting" genes or proto-oncogenes could be also more neutrally defined as "stemness" genes since they are in general highly expressed in stem cells bearing self-renewing capacity. This was indeed pointed out in the work by Kustikova *et al.* where integrations in dominant clones were reported to significantly mark "stemness" pathways.[55] This feature is of particular relevance for MLV vectors which, in absence of any strong *in vivo* clonal selection, display a tendency to land into certain stem cell associated loci *in vitro* due to the transcriptional activity and epigenetic status of hematopoietic progenitors at the time of transduction.[20,51]

These insertional preferences could explain why different GT trials, irrespectively to their outcomes, displayed the same vector bias for genomic regions like LMO2 and EVI1, which are highly active in hematopoietic progenitors.[56,57] Indeed, by tracking LMO2 integrations overtime in different patients from our ADA-SCID GT trial, we showed that the relative clonal contribution of these integrants was maintained below 1% of all transduced T cells over a long period of time after GT.[19] More recently, we found that this locus is a target for retroviral integrations only in HSC, where several histone modifications markers reflecting an open chromatin configuration are suggestive of a high accessibility of this region to integration events.[20]

In any case, analyzing the frequency of potentially dangerous insertions into the genome represents only an indirect measure of the genotoxic potential of integrating vectors. A more precise indication of insertional genotoxicity would come from a detailed study of the consequences of a specific integration on cell behavior in clinical samples. However, studying potential vector-mediated transactivation of proto-oncogenes as a first hit mechanism for cellular transformation in patients, before the development of aberrant expansion, has been a difficult task. Indeed, by the analysis of single T-cell clones isolated *ex vivo* from patients, two independent studies were not able to show major vector-mediated perturbations of cellular genes.[58,59] In addition, even when these events were detected they did not have any influence on cellular behavior or growth rate, possibly because these subtle changes in the transcriptome could be below the hypothetical threshold required for the induction of transformation events. The detection of additional mutations after, and not before, the proto-oncogenic vector perturbation could also help explain these contrasting findings, but so far, no study was able to show the consequential appearance of these events upon retroviral insertions *in vivo* before aberrant expansions. Furthermore, it was not formally shown that a cell bearing a particular insertion site could become more susceptible to additional independent mutations and acquire a higher spontaneous mutation rate. Indeed, the long-term observations in ADA-SCID GT patients have revealed that the integration of RV into regions associated with leukemic events is not *per se* sufficient to give rise to clonal aberrant expansion in patients.[19]

In summary, the detection of insertions into proto-oncogenes, namely "stemness" genes, *in vivo* in HSC GT could also be considered as a reflection of what is a general physiological "insertional footprint" of gene-corrected hematopoietic stem cell at the time of transduction. Following this line, the biological interpretation of insertion site distribution from patients' samples would strongly benefit from a thorough profiling of integrations coupled with an in-depth analysis of target cell transcriptome and epigenome before infusion.[60]

CLONAL QUANTIFICATION OF GENE-CORRECTED CELLS

To date, ligation-mediated PCR (LM-PCR) or linear amplification-mediated PCR (LAM-PCR) are the most exploited methods to retrieve integration sites from transduced cells. Both the technologies are based on the digestion of genomic DNA with restriction enzymes, the ligation of a linker cassette and the exponential amplification of vector-genomic junctions through primers annealing on the final LTR portion and the linker cassette itself.[61] The final PCR products are then sequenced in order to collect and map the regions flanking the vector LTR, retrospectively identifying the integration sites on the genome of reference. The early protocols based on shotgun cloning into competent bacteria and Sanger sequencing have been replaced with more efficient and cost-effective methods such as the barcode tagging of LAM-PCR products from different cell sources followed by pyrosequencing of the pooled samples.[22]

The idea that data generated by LAM-PCR in combination with high-throughput sequencing are highly representative of the clonal contribution of each integrant in a given sample has been the origin of some potential misinterpretation of data derived from the clinics. The appearance of a gel run of LAM-PCR products from an *in vitro* transduced bulk population is generally a smear of undistinguishable bands corresponding to several hundred thousands diverse integration sites. Differently, samples purified from patients, years after infusion of gene-corrected cells, show more discrete bands with different intensities corresponding to a more limited number of transduced clones engrafted *in vivo*. The clonal contribution of gene-corrected cells in the context of GT clinical follow up is generally gathered from the appearance of oligoclonal or polyclonal repertoires of LAM-PCR products.[16–19,21]

It is then commonly believed that the different intensities of the gel bands (and as a consequence the different relative sequence counts obtained form high-throughput sequencing) are somehow linked to the clonal contributions of each integration site in the original cell pool.

Clonal dominance occurring *in vitro* in preclinical assays was associated to a reduction in the polyclonality of LAM-PCR products and an increased intensity of single bands overtime.[54] The sequence counts of vector-genome junctions from 454 sequencing correlated well with the dynamics of the leukemic clones bearing LMO2 integrations in the high-throughput analysis of samples from one SCID-X1 GT trial in patients before and after chemotherapic treatment.[16] More recently the relative abundance of a clone carrying an insertion in the HMGA2 locus was also measured by means of sequence counts in the β-thalassaemia patient treated with lentiviral GT.[41] Thus, is the low diversity and relatively high contribution of single LAM-PCR products in a cell pool predictive of a potentially skewed profile with dominant clones emerging from a mixed transduced population?

To fairly answer this question it is important to take into account that the LAM-PCR technique itself has specific technical biases linked to the use of restriction enzymes.[62] For example, the same sample could show few or many bands according to the cutting frequency of the enzymes used in the LAM-PCR protocol. This technical constraint affects not only the number of unique bands but also their relative intensity and ultimately the number of reads retrieved from high-throughput sequencing for each particular vector-genomic junction. As a consequence, a given integration site could be more easily retrieved due to its proximity to a specific restriction site and along the LAM-PCR protocol more favorably amplified and sequenced irrespectively of the contribution of the relative clone in the original cell pool. Indeed, the group of Christopher Baum reported recently that, in the absence of self-evident clonal expansions, there is often a strong discrepancy between insertion site frequency measured by 454 sequencing and the results of specific quantitative PCR designed on the same vector-genome junctions.[63] This aspect should be carefully taken into account since many groups are now exploiting more than one restriction enzyme for LAM-PCR protocols that are then differently biased.

Overcoming some of these issues, a significant improvement of LAM-PCR technology has been recently proposed by Paruzinsky *et al.*[64] The authors showed the exploitation of a restriction site-independent method of insertion site retrieval, which proved to be efficacious but yet limited by the amount of DNA required for the analysis. Another unbiased method for the recovery of integration sites is based on a phage Mu-mediated introduction of an adaptor sequence allowing the amplification of vector-genome junctions without the need for restriction enzymes.[65] This protocol has been exploited with success to estimate the relative abundance of gene-modified cells in clinical trials samples. Additional technologies providing a theoretically unbiased high-throughput access to the genome have been recently developed based on sonication of genomic DNA.[66] Once these and other new nonrestrictive platforms will be consolidated in terms of sensitivity they should provide the techniques of choice for future vector integration studies.

Importantly, sequence counts of vector-genome junctions may be informative of the relative frequency of a clone bearing a given insertion only within the population of gene-corrected cells but not on the whole set of blood cells in a given lineage. Thus, before drawing conclusions on a clonal amplification of potential clinical relevance, a relatively high sequence count (e.g., 30% of the total sequencing reads) should be always normalized to the frequency of transduced cells in the analyzed sample (which in some cases is less than 5% of the total cell population in the patient).

INTEGRATION SITE ANALYSIS AND NEXT GENERATION SEQUENCING

Despite the caveats and the open questions, over the past years the analysis of integration profiles in retroviral-based GT has allowed investigators to address several issues related to insertional genotoxicity and safety of GT approaches.[6,16–21,40,55] The next generation sequencing technologies have dramatically boosted these studies by exponentially increasing the amount of insertion sites available for in-depth analysis.

The other side of the coin is the potential "toxicity of information" deriving from the overflow of data originated from deep sequencing. There is already a strong requirement for the development of new tools both for the automatic processing of raw sequences and for the algorithms downstream of the analysis pipeline. A proper handling of the sequence datasets is now required to effectively manage thousands of sequencing reads at one time. A major challenge to this approach will be the storage and use of large amount of sequence data, which will require an integrated platform of database structures flexible to the future development of sequencing technologies. The next years will also see an increasing demand for more refined tools to discriminate between *bona fide* vector-genomic junctions and by-products derived from LAM-PCR/deep sequencing systematic biases. Indeed, although extremely time and cost-effective high-throughput technologies still carry an average error rate significantly higher than the typical observed from high-quality Sanger sequencing.[67] Downstream analysis will also need to be adapted to the size of the deep sequencing data to avoid misinterpretations.

A prototypical example involves the identification of CIS. When statistical algorithms designed for low-throughput generated databases are exploited for the detection of cluster of insertions in larger insertional data, the incidence of CIS could be over-estimated, and may lead to not fully statistically supported biological conclusions. Concerns about the statistical definition of CIS were already raised in the early insertion site analysis[68] and in a work based on a meta-analysis of all the CIS identified from retroviral tagging experiments and listed in the retroviral tagged cancer gene database (RTCGD).[69] By the development of a kernel convolution framework, de Ridder and co-authors were able to detect CIS in a noisy environment while controlling the probability of detecting false clusters of insertions. Strikingly, they found that 53% of the previously defined CIS did not reach the significance threshold in this setting. To date, the genomic windows initially established for the characterization and identification of CIS are in many case too wide to properly define an insertional hotspot and they should be adapted in terms of distances between two or more insertion sites to the currently available databases. On this

line, new statistical tools for CIS definition are under development for the analysis of large insertional datasets.[20,52,70,71]

THE FUTURE OF INTEGRATION PROFILING

Although in the next years insertion site analysis will still generate information relevant to the GT field, the impact of these data will progressively broaden to address more general biological question.

An obvious extension of these studies is based on the use of integration site analysis to get new insights on the biology of wild-type retroviruses with a particular focus on HIV. Indeed, increasing amount of data from insertional profiling have been produced along the last years enlightening molecular aspect linked to tethering, nuclear transportation, and latency of HIV.[72–75] To date, one of the best-supported models of HIV shuttle to nucleus, involves interactions between HIV integrase (IN) and LEDGF/p75 protein.[76,77] A study based on insertional profiling of wild-type HIV on cells lacking LEDGF/p75 showed reduced frequency of integration in transcription units[72] proposing a tethering function of LEDGF/p75 through the binding of IN. However, since the reduction in targeting of genes appeared overall modest, the IN–LEDGF/p75 interaction did not seem to fully explain HIV integration preferences, indicating that other molecules may be involved. A more recent work showed that depletion of transportin-3 and RanBP2 altered integration targeting for HIV suggesting a role of these molecules in the nuclear transportation of HIV.[75] Another integration site analysis was also able to address issues regarding the latency of HIV. Working on the hypothesis that integration in resting cells may contribute to formation of the latent reservoir, Brady *et al.* showed that HIV insertions were more frequently found in relatively less gene-dense regions in resting cells than in activated cells, suggesting that, when landing in such gene deserts HIV, may be more prone to forming latent proviruses.[74] High-throughput genetic and epigenetic mapping combined with refined insertion site analysis will be in the next years crucial to extrapolate more factors involved in HIV insertional mechanism potentially uncovering molecular elements with a pivotal role in HIV life cycle and reactivation.

Integration site analysis has also been exploited to uncover the properties of Zinc-finger nucleases (ZFNs) molecules. ZFNs-based technology is able, by introducing a double-strand break at a predetermined locus, to drive efficiently site-specific integration of a transgene of interest in human target cells.[78–80] Until recently, a direct genome-wide measurement of the degree of specificity of ZFNs activity at the desired locus was missing. By combining ZNFs with an integrase-defective lentiviral vector, Gabriel *et al.* were able to tag the sites of double-strand breaks and to retrieve the insertions of integrase-defective lentiviral vector with conventional LAM-PCR/pyrosequencing-based techniques.[81] Thus, insertion profiling has allowed studying with unprecedented resolution the incidence of ZFNs off-target activity moving the technology forward towards its application in translational research. Studies based on integration site analysis will be important to provide information for the utilization of new molecular tools in the clinic, as well as to enlighten new mechanisms at the basis of protein–DNA interactions.

A broader application of integration profiling involves the exploitation of specific integrants as tags to track single cell clones *in vivo*. Upon transduction each target hematopoietic cell becomes marked by a unique vector integration site, which will be inherited by all its progeny. Consequently, if two or more cell clones belonging to different lineages share an identical integration, it is likely they are derived from a common upstream progenitor. Based on this principle, insertion site analysis of different hematopoietic lineages after infusion of transduced cells, could provide a unique tool to study *in vivo* hematopoietic progenitors survival, fate decisions, and dynamics overtime directly in patients. Indeed, while it has long been established that multipotent, self-renewing HSCs sustain the lifelong replenishment of the whole mature blood cell compartment, the complete hierarchy of human hematopoiesis is still far from being elucidated.[82–86] Reaching a consensus on a model of hematopoiesis has been hampered essentially by to the lack of experimental settings allowing high-throughput clonal analysis of hematopoietic progenitors behaviour *in vivo*. Addressing this issue, retroviral insertion databases from SCID-X1 GT patients have been exploited to assess the long-term clonal output of hematopoietic progenitors, up to 10 years after infusion of transduced cells.[87] However, the limited lineage engraftment of transduced cells in patients, allowed only an estimation of the clonal diversity of the reconstituted hematopoietic system based on a capture–recapture method applied to insertions retrieved from T cells. The ALD and WAS GT clinical trials,[6,11] showed instead a multilineage engraftment of gene-corrected cells up to 2 years postinfusion of transduced cells. A high degree of shared identical integrations between myeloid and lymphoid lineages was observed, potentially marking multipotent progenitors but the analysis was performed at early timepoints, lacking a longer patients follow up.

To draw a comprehensive picture of human hematopoietic dynamics through "insertional tagging" both multilineage engraftment and long-term follow up will be needed. Indeed, only under these conditions the detection of shared identical integrations among bone marrow $CD34^+$ progenitors and multiple myeloid and lymphoid lineages persisting overtime would identify, directly in humans, cell clones meeting the definition of long-term reconstituting HSCs. Similarly, the consistent finding of integrants shared between some hematopoietic lineages but not others, would indicate the persistence of marked lineage-restricted progenitors. Monitoring these clones over defined periods of time in humans, as performed in mice,[88] would provide important information concerning the lifespan and possible fluctuations in lineage output of hematopoietic progenitors.

On this line, a promising tool for quantitative and unbiased studies in a mixed population of transduced cells is represented by retroviral oligonucleotide barcoding, exploiting PCR and/or array-based techniques to detect complex libraries of RVs with unique sequence tags,.[65,89,90] Nonetheless, the exploitation of this technique remains to date confined to animal studies, since the utilization of vector libraries is not feasible in human GT applications.

CONCLUSIONS

Overall, in the coming years, studies based on integration site analysis will play a fundamental role in answering questions

going from the basic biology to the translational research and clinical applications. The increasing amount of data will require investigators to deal with the potential overload of information by constantly reviewing and adapting the tools for the analysis of insertion sites at all levels. In addition, more insights from *in vitro* and *in vivo* experiments will be needed to solve the controversies associated with the functional role of vector–host interactions. These steps will be crucial to provide a proper interpretation of biological data while the technology of retroviral gene transfer is moving rapidly towards new and more complex applications.

ACKNOWLEDGMENTS

The authors are grateful for support of our work by Italian Telethon Foundation, CLINIGENE-NoE, CELL-PID and EUROCGD (E-rare). We thank Kevin Goudy for his revision of the manuscript. The authors declared no conflict of interest.

REFERENCES

1. Naldini, L (2011). Ex vivo gene transfer and correction for cell-based therapies. *Nat Rev Genet* **12**: 301–315.
2. Rossi, JJ, June, CH and Kohn, DB (2007). Genetic therapies against HIV. *Nat Biotechnol* **25**: 1444–1454.
3. Deeks, SG, Wagner, B, Anton, PA, Mitsuyasu, RT, Scadden, DT, Huang, C *et al.* (2002). A phase II randomized study of HIV-specific T-cell gene therapy in subjects with undetectable plasma viremia on combination antiretroviral therapy. *Mol Ther* **5**: 788–797.
4. Mavilio, F, Pellegrini, G, Ferrari, S, Di Nunzio, F, Di Iorio, E, Recchia, A *et al.* (2006). Correction of junctional epidermolysis bullosa by transplantation of genetically modified epidermal stem cells. *Nat Med* **12**: 1397–1402.
5. Ciceri, F, Bonini, C, Stanghellini, MT, Bondanza, A, Traversari, C, Salomoni, M *et al.* (2009). Infusion of suicide-gene-engineered donor lymphocytes after family haploidentical haemopoietic stem-cell transplantation for leukaemia (the TK007 trial): a non-randomised phase I-II study. *Lancet Oncol* **10**: 489–500.
6. Boztug, K, Schmidt, M, Schwarzer, A, Banerjee, PP, Díez, IA, Dewey, RA *et al.* (2010). Stem-cell gene therapy for the Wiskott-Aldrich syndrome. *N Engl J Med* **363**: 1918–1927.
7. Hacein-Bey-Abina, S, Hauer, J, Lim, A, Picard, C, Wang, GP, Berry, CC *et al.* (2010). Efficacy of gene therapy for X-linked severe combined immunodeficiency. *N Engl J Med* **363**: 355–364.
8. Cavazzana-Calvo, M, Hacein-Bey, S, de Saint Basile, G, Gross, F, Yvon, E, Nusbaum, P *et al.* (2000). Gene therapy of human severe combined immunodeficiency (SCID)-X1 disease. *Science* **288**: 669–672.
9. Aiuti, A, Vai, S, Mortellaro, A, Casorati, G, Ficara, F, Andolfi, G *et al.* (2002). Immune reconstitution in ADA-SCID after PBL gene therapy and discontinuation of enzyme replacement. *Nat Med* **8**: 423–425.
10. Aiuti, A, Cattaneo, F, Galimberti, S, Benninghoff, U, Cassani, B, Callegaro, L *et al.* (2009). Gene therapy for immunodeficiency due to adenosine deaminase deficiency. *N Engl J Med* **360**: 447–458.
11. Cartier, N, Hacein-Bey-Abina, S, Bartholomae, CC, Veres, G, Schmidt, M, Kutschera, I *et al.* (2009). Hematopoietic stem cell gene therapy with a lentiviral vector in X-linked adrenoleukodystrophy. *Science* **326**: 818–823.
12. Hacein-Bey-Abina, S, Von Kalle, C, Schmidt, M, McCormack, MP, Wulffraat, N, Leboulch, P *et al.* (2003). LMO2-associated clonal T cell proliferation in two patients after gene therapy for SCID-X1. *Science* **302**: 415–419.
13. Hacein-Bey-Abina, S, von Kalle, C, Schmidt, M, Le Deist, F, Wulffraat, N, McIntyre, E *et al.* (2003). A serious adverse event after successful gene therapy for X-linked severe combined immunodeficiency. *N Engl J Med* **348**: 255–256.
14. Ott, MG, Schmidt, M, Schwarzwaelder, K, Stein, S, Siler, U, Koehl, U *et al.* (2006). Correction of X-linked chronic granulomatous disease by gene therapy, augmented by insertional activation of MDS1-EVI1, PRDM16 or SETBP1. *Nat Med* **12**: 401–409.
15. Stein, S, Ott, MG, Schultze-Strasser, S, Jauch, A, Burwinkel, B, Kinner, A *et al.* (2010). Genomic instability and myelodysplasia with monosomy 7 consequent to EVI1 activation after gene therapy for chronic granulomatous disease. *Nat Med* **16**: 198–204.
16. Hacein-Bey-Abina, S, Garrigue, A, Wang, GP, Soulier, J, Lim, A, Morillon, E *et al.* (2008). Insertional oncogenesis in 4 patients after retrovirus-mediated gene therapy of SCID-X1. *J Clin Invest* **118**: 3132–3142.
17. Deichmann, A, Hacein-Bey-Abina, S, Schmidt, M, Garrigue, A, Brugman, MH, Hu, J *et al.* (2007). Vector integration is nonrandom and clustered and influences the fate of lymphopoiesis in SCID-X1 gene therapy. *J Clin Invest* **117**: 2225–2232.
18. Howe, SJ, Mansour, MR, Schwarzwaelder, K, Bartholomae, C, Hubank, M, Kempski, H *et al.* (2008). Insertional mutagenesis combined with acquired somatic mutations causes leukemogenesis following gene therapy of SCID-X1 patients. *J Clin Invest* **118**: 3143–3150.
19. Aiuti, A, Cassani, B, Andolfi, G, Mirolo, M, Biasco, L, Recchia, A *et al.* (2007). Multilineage hematopoietic reconstitution without clonal selection in ADA-SCID patients treated with stem cell gene therapy. *J Clin Invest* **117**: 2233–2240.
20. Biasco, L, Ambrosi, A, Pellin, D, Bartholomae, C, Brigida, I, Roncarolo, MG *et al.* (2011). Integration profile of retroviral vector in gene therapy treated patients is cell-specific according to gene expression and chromatin conformation of target cell. *EMBO Mol Med* **3**: 89–101.
21. Schwarzwaelder, K, Howe, SJ, Schmidt, M, Brugman, MH, Deichmann, A, Glimm, H *et al.* (2007). Gammaretrovirus-mediated correction of SCID-X1 is associated with skewed vector integration site distribution in vivo. *J Clin Invest* **117**: 2241–2249.
22. Schmidt, M, Schwarzwaelder, K, Bartholomae, CC, Glimm, H and von Kalle, C (2009). Detection of retroviral integration sites by linear amplification-mediated PCR and tracking of individual integration clones in different samples. *Methods Mol Biol* **506**: 363–372.
23. Fischer, A and Cavazzana-Calvo, M (2008). Gene therapy of inherited diseases. *Lancet* **371**: 2044–2047.
24. Gaspar, HB, Cooray, S, Gilmour, KC, Parsley, KL, Adams, S, Howe, SJ *et al.* (2011). Long-term persistence of a polyclonal T cell repertoire after gene therapy for X-linked severe combined immunodeficiency. *Sci Transl Med* **3**: 97ra79.
25. Rabbitts, TH (1998). LMO T-cell translocation oncogenes typify genes activated by chromosomal translocations that alter transcription and developmental processes. *Genes Dev* **12**: 2651–2657.
26. Bester, AC, Schwartz, M, Schmidt, M, Garrigue, A, Hacein-Bey-Abina, S, Cavazzana-Calvo, M *et al.* (2006). Fragile sites are preferential targets for integrations of MLV vectors in gene therapy. *Gene Ther* **13**: 1057–1059.
27. Davé, UP, Jenkins, NA and Copeland, NG (2004). Gene therapy insertional mutagenesis insights. *Science* **303**: 333.
28. Davé, UP, Akagi, K, Tripathi, R, Cleveland, SM, Thompson, MA, Yi, M *et al.* (2009). Murine leukemias with retroviral insertions at Lmo2 are predictive of the leukemias induced in SCID-X1 patients following retroviral gene therapy. *PLoS Genet* **5**: e1000491.
29. Woods, NB, Bottero, V, Schmidt, M, von Kalle, C and Verma, IM (2006). Gene therapy: therapeutic gene causing lymphoma. *Nature* **440**: 1123.
30. Pike-Overzet, K, de Ridder, D, Weerkamp, F, Baert, MRM, Verstegen, MM, Brugman, MH *et al.* (2006). Gene therapy: is IL2RG oncogenic in T-cell development? *Nature* **443**: E5.
31. Pike-Overzet, K, van der Burg, M, Wagemaker, G, van Dongen, JJ and Staal, FJ (2007). New insights and unresolved issues regarding insertional mutagenesis in X-linked SCID gene therapy. *Mol Ther* **15**: 1910–1916.
32. Ginn, SL, Liao, SH, Dane, AP, Hu, M, Hyman, J, Finnie, JW *et al.* (2010). Lymphomagenesis in SCID-X1 mice following lentivirus-mediated phenotype correction independent of insertional mutagenesis and gammac overexpression. *Mol Ther* **18**: 965–976.
33. Avedillo Diez, I, Zychlinski, D, Coci, EG, Galla, M, Modlich, U, Dewey, RA *et al.* (2011). Development of novel efficient SIN vectors with improved safety features for Wiskott-Aldrich syndrome stem cell based gene therapy. *Mol Pharm* **8**: 1525–1537.
34. Trobridge, GD (2011). Genotoxicity of retroviral hematopoietic stem cell gene therapy. *Expert Opin Biol Ther* **11**: 581–593.
35. Ott, MG, Merget-Millitzer, H, Ottmann, OG, Martin, H, Brüggenolte, N, Bialek, H *et al.* (2002). Mobilization and transduction of CD34(+) peripheral blood stem cells in patients with X-linked chronic granulomatous disease. *J Hematother Stem Cell Res* **11**: 683–694.
36. Grez, M, Reichenbach, J, Schwäble, J, Seger, R, Dinauer, MC and Thrasher, AJ (2011). Gene therapy of chronic granulomatous disease: the engraftment dilemma. *Mol Ther* **19**: 28–35.
37. Seger, RA (2010). Chronic granulomatous disease: recent advances in pathophysiology and treatment. *Neth J Med* **68**: 334–340.
38. Kang, EM, Choi, U, Theobald, N, Linton, G, Long Priel, DA, Kuhns, D *et al.* (2010). Retrovirus gene therapy for X-linked chronic granulomatous disease can achieve stable long-term correction of oxidase activity in peripheral blood neutrophils. *Blood* **115**: 783–791.
39. Calmels, B, Ferguson, C, Laukkanen, MO, Adler, R, Faulhaber, M, Kim, HJ *et al.* (2005). Recurrent retroviral vector integration at the Mds1/Evi1 locus in nonhuman primate hematopoietic cells. *Blood* **106**: 2530–2533.
40. Montini, E, Cesana, D, Schmidt, M, Sanvito, F, Bartholomae, CC, Ranzani, M *et al.* (2009). 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.
41. Cavazzana-Calvo, M, Payen, E, Negre, O, Wang, G, Hehir, K, Fusil, F *et al.* (2010). Transfusion independence and HMGA2 activation after gene therapy of human ß-thalassaemia. *Nature* **467**: 318–322.
42. Mitchell, RS, Beitzel, BF, Schroder, AR, Shinn, P, Chen, H, Berry, CC *et al.* (2004). Retroviral DNA integration: ASLV, HIV, and MLV show distinct target site preferences. *PLoS Biol* **2**: E234.
43. Wu, X, Li, Y, Crise, B and Burgess, SM (2003). Transcription start regions in the human genome are favored targets for MLV integration. *Science* **300**: 1749–1751.
44. Ikeda, K, Mason, PJ and Bessler, M (2011). 3'UTR-truncated Hmga2 cDNA causes MPN-like hematopoiesis by conferring a clonal growth advantage at the level of HSC in mice. *Blood* **117**: 5860–5869.
45. Gaspar, HB, Cooray, S, Gilmour, KC, Parsley, KL, Zhang, F, Adams, S *et al.* (2011). 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.
46. Ferrua, F, Brigida, I and Aiuti, A (2010). Update on gene therapy for adenosine deaminase-deficient severe combined immunodeficiency. *Curr Opin Allergy Clin Immunol* **10**: 551–556.
47. Shin, MS, Fredrickson, TN, Hartley, JW, Suzuki, T, Akagi, K, Agaki, K *et al.* (2004). High-throughput retroviral tagging for identification of genes involved in initiation and progression of mouse splenic marginal zone lymphomas. *Cancer Res* **64**: 4419–4427.
48. Mikkers, H and Berns, A (2003). Retroviral insertional mutagenesis: tagging cancer pathways. *Adv Cancer Res* **88**: 53–99.
49. Kim, R, Trubetskoy, A, Suzuki, T, Jenkins, NA, Copeland, NG and Lenz, J (2003). Genome-based identification of cancer genes by proviral tagging in mouse retrovirus-induced T-cell lymphomas. *J Virol* **77**: 2056–2062.

50. Lund, AH, Turner, G, Trubetskoy, A, Verhoeven, E, Wientjens, E, Hulsman, D *et al.* (2002). Genome-wide retroviral insertional tagging of genes involved in cancer in Cdkn2a-deficient mice. *Nat Genet* **32**: 160–165.
51. Cattoglio, C, Facchini, G, Sartori, D, Antonelli, A, Miccio, A, Cassani, B *et al.* (2007). Hot spots of retroviral integration in human CD34+ hematopoietic cells. *Blood* **110**: 1770–1778.
52. Biffi, A, Bartolomae, CC, Cesana, D, Cartier, N, Aubourg, P, Ranzani, M *et al.* (2011). Lentiviral vector common integration sites in preclinical models and a clinical trial reflect a benign integration bias and not oncogenic selection. *Blood* **117**: 5332–5339.
53. Deichmann, A, Brugman, MH, Bartholomae, CC, Schwarzwaelder, K, Verstegen, MM, Howe, SJ *et al.* (2011). Insertion sites in engrafted cells cluster within a limited repertoire of genomic areas after gammaretroviral vector gene therapy. *Mol Ther* **19**: 2031–2039.
54. Kustikova, O, Fehse, B, Modlich, U, Yang, M, Düllmann, J, Kamino, K *et al.* (2005). Clonal dominance of hematopoietic stem cells triggered by retroviral gene marking. *Science* **308**: 1171–1174.
55. Kustikova, OS, Geiger, H, Li, Z, Brugman, MH, Chambers, SM, Shaw, CA *et al.* (2007). Retroviral vector insertion sites associated with dominant hematopoietic clones mark "stemness" pathways. *Blood* **109**: 1897–1907.
56. Shimizu, S, Nagasawa, T, Katoh, O, Komatsu, N, Yokota, J and Morishita, K (2002). EVI1 is expressed in megakaryocyte cell lineage and enforced expression of EVI1 in UT-7/GM cells induces megakaryocyte differentiation. *Biochem Biophys Res Commun* **292**: 609–616.
57. Su, AI, Wiltshire, T, Batalov, S, Lapp, H, Ching, KA, Block, D *et al.* (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci USA* **101**: 6062–6067.
58. Cassani, B, Montini, E, Maruggi, G, Ambrosi, A, Mirolo, M, Selleri, S *et al.* (2009). Integration of retroviral vectors induces minor changes in the transcriptional activity of T cells from ADA-SCID patients treated with gene therapy. *Blood* **114**: 3546–3556.
59. Recchia, A, Bonini, C, Magnani, Z, Urbinati, F, Sartori, D, Muraro, S *et al.* (2006). Retroviral vector integration deregulates gene expression but has no consequence on the biology and function of transplanted T cells. *Proc Natl Acad Sci USA* **103**: 1457–1462.
60. Baum, C (2011). Parachuting in the epigenome: the biology of gene vector insertion profiles in the context of clinical trials. *EMBO Mol Med* **3**: 75–77.
61. Schmidt, M, Schwarzwaelder, K, Bartholomae, C, Zaoui, K, Ball, C, Pilz, I *et al.* (2007). High-resolution insertion-site analysis by linear amplification-mediated PCR (LAM-PCR). *Nat Methods* **4**: 1051–1057.
62. Gabriel, R, Eckenberg, R, Paruzynski, A, Bartholomae, CC, Nowrouzi, A, Arens, A *et al.* (2009). Comprehensive genomic access to vector integration in clinical gene therapy. *Nat Med* **15**: 1431–1436.
63. Kustikova, O, Brugman, M, Stahlhut, M, Bartels, S, Schiedlmeier, B, Li, Z *et al.* (2009). Longitudinal analysis of insertional mutants in a murine model of hematopoietic stem cell transplantation (Meeting Abstract). *Hum Gene Ther* **20**: 1507–1508.
64. Paruzynski, A, Arens, A, Gabriel, R, Bartholomae, CC, Scholz, S, Wang, W *et al.* (2010). Genome-wide high-throughput integrome analyses by nrLAM-PCR and next-generation sequencing. *Nat Protoc* **5**: 1379–1395.
65. Brady, T, Roth, SL, Malani, N, Wang, GP, Berry, CC, Leboulch, P *et al.* (2011). A method to sequence and quantify DNA integration for monitoring outcome in gene therapy. *Nucleic Acids Res* **39**: e72.
66. Gillet, NA, Malani, N, Melamed, A, Gormley, N, Carter, R, Bentley, D *et al.* (2011). The host genomic environment of the provirus determines the abundance of HTLV-1-infected T-cell clones. *Blood* **117**: 3113–3122.
67. Kircher, M and Kelso, J (2010). High-throughput DNA sequencing–concepts and limitations. *Bioessays* **32**: 524–536.
68. Wu, X, Luke, BT and Burgess, SM (2006). Redefining the common insertion site. *Virology* **344**: 292–295.
69. de Ridder, J, Uren, A, Kool, J, Reinders, M and Wessels, L (2006). Detecting statistically significant common insertion sites in retroviral insertional mutagenesis screens. *PLoS Comput Biol* **2**: e166.
70. Cattoglio, C, Pellin, D, Rizzi, E, Maruggi, G, Corti, G, Miselli, F *et al.* (2010). High-definition mapping of retroviral integration sites identifies active regulatory elements in human multipotent hematopoietic progenitors. *Blood* **116**: 5507–5517.
71. Presson, AP, Kim, N, Xiaofei, Y, Chen, IS and Kim, S (2011). Methodology and software to detect viral integration site hot-spots. *BMC Bioinformatics* **12**: 367.
72. Ciuffi, A, Llano, M, Poeschla, E, Hoffmann, C, Leipzig, J, Shinn, P *et al.* (2005). A role for LEDGF/p75 in targeting HIV DNA integration. *Nat Med* **11**: 1287–1289.
73. Lewinski, MK, Bisgrove, D, Shinn, P, Chen, H, Hoffmann, C, Hannenhalli, S *et al.* (2005). Genome-wide analysis of chromosomal features repressing human immunodeficiency virus transcription. *J Virol* **79**: 6610–6619.
74. Brady, T, Agosto, LM, Malani, N, Berry, CC, O'Doherty, U and Bushman, F (2009). HIV integration site distributions in resting and activated CD4+ T cells infected in culture. *AIDS* **23**: 1461–1471.
75. Ocwieja, KE, Brady, TL, Ronen, K, Huegel, A, Roth, SL, Schaller, T *et al.* (2011). HIV integration targeting: a pathway involving Transportin-3 and the nuclear pore protein RanBP2. *PLoS Pathog* **7**: e1001313.
76. Cherepanov, P, Maertens, G, Proost, P, Devreese, B, Van Beeumen, J, Engelborghs, Y *et al.* (2003). HIV-1 integrase forms stable tetramers and associates with LEDGF/p75 protein in human cells. *J Biol Chem* **278**: 372–381.
77. Llano, M, Vanegas, M, Fregoso, O, Saenz, D, Chung, S, Peretz, M *et al.* (2004). LEDGF/p75 determines cellular trafficking of diverse lentiviral but not murine oncoretroviral integrase proteins and is a component of functional lentiviral preintegration complexes. *J Virol* **78**: 9524–9537.
78. Urnov, FD, Rebar, EJ, Holmes, MC, Zhang, HS and Gregory, PD (2010). Genome editing with engineered zinc finger nucleases. *Nat Rev Genet* **11**: 636–646.
79. Lombardo, A, Genovese, P, Beausejour, CM, Colleoni, S, Lee, YL, Kim, KA *et al.* (2007). Gene editing in human stem cells using zinc finger nucleases and integrase-defective lentiviral vector delivery. *Nat Biotechnol* **25**: 1298–1306.
80. Urnov, FD, Miller, JC, Lee, YL, Beausejour, CM, Rock, JM, Augustus, S *et al.* (2005). Highly efficient endogenous human gene correction using designed zinc-finger nucleases. *Nature* **435**: 646–651.
81. Gabriel, R, Lombardo, A, Arens, A, Miller, JC, Genovese, P, Kaeppel, C *et al.* (2011). An unbiased genome-wide analysis of zinc-finger nuclease specificity. *Nat Biotechnol* **29**: 816–823.
82. Wada, H, Masuda, K, Satoh, R, Kakugawa, K, Ikawa, T, Katsura, Y *et al.* (2008). Adult T-cell progenitors retain myeloid potential. *Nature* **452**: 768–772.
83. Bell, JJ and Bhandoola, A (2008). The earliest thymic progenitors for T cells possess myeloid lineage potential. *Nature* **452**: 764–767.
84. Doulatov, S, Notta, F, Eppert, K, Nguyen, LT, Ohashi, PS and Dick, JE (2010). Revised map of the human progenitor hierarchy shows the origin of macrophages and dendritic cells in early lymphoid development. *Nat Immunol* **11**: 585–593.
85. Adolfsson, J, Månsson, R, Buza-Vidas, N, Hultquist, A, Liuba, K, Jensen, CT *et al.* (2005). Identification of Flt3+ lympho-myeloid stem cells lacking erythro-megakaryocytic potential a revised road map for adult blood lineage commitment. *Cell* **121**: 295–306.
86. Kawamoto, H, Wada, H and Katsura, Y (2010). A revised scheme for developmental pathways of hematopoietic cells: the myeloid-based model. *Int Immunol* **22**: 65–70.
87. Wang, GP, Berry, CC, Malani, N, Leboulch, P, Fischer, A, Hacein-Bey-Abina, S *et al.* (2010). Dynamics of gene-modified progenitor cells analyzed by tracking retroviral integration sites in a human SCID-X1 gene therapy trial. *Blood* **115**: 4356–4366.
88. Dykstra, B, Kent, D, Bowie, M, McCaffrey, L, Hamilton, M, Lyons, K *et al.* (2007). Long-term propagation of distinct hematopoietic differentiation programs in vivo. *Cell Stem Cell* **1**: 218–229.
89. Gerrits, A, Dykstra, B, Kalmykowa, OJ, Klauke, K, Verovskaya, E, Broekhuis, MJ *et al.* (2010). Cellular barcoding tool for clonal analysis in the hematopoietic system. *Blood* **115**: 2610–2618.
90. Lu, R, Neff, NF, Quake, SR and Weissman, IL (2011). Tracking single hematopoietic stem cells *in vivo* using high-throughput sequencing in conjunction with viral genetic barcoding. *Nat Biotechnol* **29**: 928–933.