OXFORD JOURNAL
OF STUDENT
SCHOLARSHIP

Genomic Safe Harbours and Human Gene Therapy: A Review of Candidate Sites, Evaluation Criteria, and Functional Applications

Nandini Pangal
nandini.pangal@gmail.com

ABSTRACT

Genetic engineering, particularly insertion therapies like CRISPR-Cas9, depends on the stability and predictability of transgene integration, modification, or deletion into human DNA. Unintended outcomes can occur due to random or inadequate targeting, ranging from increased oncogene expression to insertional mutagenesis. Therefore, Genomic Safe Harbours are defined as sites capable of supporting transgene integration, modification, or deletion (i) without disrupting essential cellular functions, and (ii) allowing predictable, stable expression across different contexts. In spite of thorough research across chromosomes, no universally applicable, clinically validated Genomic Safe Harbour site has yet been established. While some widely used sites, such as AAVS1, hRosa26, and CCR5, are often considered Genomic Safe Harbour sites, they do not meet the previously mentioned criteria and show variation in expression and integration outcomes. Furthermore, current identification criteria lack consistency across studies, revealing a fundamental gap: the absence of a standardized framework for identifying and comparing such sites. This review synthesizes current research on Genomic Safe Harbour sites, including their genomic landscape and distribution, covering the number of known sites and their conservation, evaluation criteria such as distinguishing between putative and proven sites, functional applications, prediction tools, and future directions for the standardized and quantitative identification and prioritization of these sites.

Keywords:
Genomic safe harbours, Gene therapy, Targeted genome integration, CRISPR-Cas systems

1. INTRODUCTION AND BACKGROUND

Modern biology has been transformed by the ability to modify the genome, especially with the discovery of the Cas9 enzyme, which has enabled the modification, deletion, or insertion of genes that can eliminate diseases and alleviate suffering. (Jinek et al.; Anzalone et al.) However, for the clinical application of genetic engineering, it is necessary that the target sites selected for the deployment of gRNA, or RNA-guided endonuclease technology, must have functional predictability and must have minimal

variability in the consequences of integrating, modifying, or deleting a variety of different transgenes. (Papapetrou and Schambach) Insertion at inappropriate sites can have a myriad of consequences, such as oncogene activation leading to cancer, unintended immune responses and reactions that can result in organ failure, or off-target effects resulting in insertion at the incorrect site. (Hacein-Bey-Abina et al.; Bushman) Therefore, it is imperative to have reliable genomic loci where inserted transgenes can be stably expressed over the long-term without affecting endogenous cellular processes. (Papapetrou and Schambach) Moreover, integration near transcription start sites (TSSs) can result in dysregulation, which can lead to changes in promoter efficacy or isoform diversity, and integration near oncogenes is a safety risk. (Sadelain et al.) Therefore, the natural course of action would be to insert into relatively inactive regions with an absence of genetic elements, but this would then result in poor expression of the transgene. (Papapetrou and Schambach) Therefore, Genomic Safe Harbour sites must also strike a balance between safety and expression. Therefore, Genomic Safe Harbour (GSH) sites must balance transcriptional accessibility with genomic safety. An ideal locus should permit strong transgene expression while remaining sufficiently distant from oncogenes, essential genes, transcription start sites (TSSs), and other regulatory elements associated with insertional mutagenesis. (Sadelain et al.) The effectiveness of any integration site is further modulated by cell-type-specific chromatin architecture, including topological domain organisation, histone modifications, and enhancer–promoter interactions. These factors make genomic safety an inherently context-dependent property (Dixon et al.; Hyun et al.; Terranova et al.) Historically, three loci have been constantly used as Genomic Safe Harbour candidate sites: AAVS1, CCR5, and hRosa26, due to high accessibility and ease of CRISPR-Cas9 targeting. However, these genes are located within transcription units or in gene-rich regions, exhibiting variety in their expression. (Klatt et al.; Bhagwan et al.) This variability and volatility raise concerns for clinical applications. Consequently, the identification of a true, ubiquitous genetic safe harbour site is an unresolved problem. While studies have shown the use of safe harbour sites in non-human species such as pigs and rice, this review will be limited to the use of such sites in human gene therapies. (Chen et al.; Cantos et al.) A further complication is the absence of standardised evaluation criteria, and published studies use diverse methods of identification, ranging from investigating datasets containing endogenous parvoviral elements (EPVs) to computational approaches that use binary filtering mechanisms. (Autio et al.; Quezada-Ramírez et al.; Leitão et al.) Recent advances have shifted from a primarily biological approach to one that integrates computational analyses, but these approaches still face the aforementioned issue with differing criteria. (Autio et al.; Leitão et al.) Reliable and context-specific Genomic Safe Harbour sites are even more important in light of recent clinical advances. Transfer efficiency, editing efficiency, and long-term expression of current gene editing therapies and pipelines vary significantly depending on the cell type and delivery method. (Dunbar et al.) For instance, editing efficiencies of >80% have been achieved with ex vivo delivery of CRISPR in CD34+ hematopoietic stem and progenitor cells (HSPCs) in therapies like Casgevy, compared to *in vivo* delivery methods with more tissue-specific and lower editing efficiencies. (Frangoul et al.) Likewise, adeno-associated viral (AAV) vectors and lipid nanoparticles have different tissue tropisms and different degrees of persistence of the transgene, making it clear that the site of genomic integration is important and must be predictable and context-dependent. (Naldini; Dunbar et al.) This review synthesises current research and identification strategies to provide a

comprehensive overview of Genomic Safe Harbour sites.

2. KNOWN SITES

2.1 AAVS1

The adeno-associated virus site 1 (AAVS1) is the naturally occurring and preferred integration site of the non-pathogenic AAV virus on chromosome 19, situated within the first intron of the protein phosphatase 1 regulatory subunit 12C (PPP1R12C) gene. (Papapetrou and Schambach) Historically, it has been one of the most widely used sites in gene editing applications; however, studies have demonstrated that transgene silencing, especially in the case of weaker cell type-specific promoters, can occur at the AAVS1 locus of human pluripotent stem cells (PSCs) and can impede transgene expression during differentiation. (Klatt et al.; Bhagwan et al.) Transgene silencing is often associated with epigenetic modifications such as DNA methylation or heterochromatin formation during lineage commitment. (Klatt et al.) Furthermore, despite its accessibility, AAVS1 resides within a transcription unit, and integration here can result in position effects dependent on the density and strength of promoters. (Papapetrou and Schambach; Sadelain et al.)

2.2 hROSA26

The human ortholog of the mouse Rosa26 locus is a locus extensively validated in the murine setting for the insertion of ubiquitously expressed transgenes. (Irion et al.) It is located on chromosome 3 within an intron of the THUMPD3 gene. (Irion et al.) Unlike the locus in mice, the human ortholog is located near critical genes, increasing the chance of their disruption, raising questions regarding its clinical application. (Irion et al.; Sadelain et al.) Although the murine Rosa26 locus satisfies many GSH criteria, the human ortholog does not fully replicate this functionality. (Irion et al.) This exposes a key limitation in that cross-species mapping is not directly transferable due to inherent differences in chromosomal organisation and architecture. (Papapetrou and Schambach)

2.3 CCR5

The chemokine (C-C motif) receptor 5 (CCR5) gene is a chemokine receptor gene known as an HIV-1 coreceptor. (Papapetrou and Schambach) It is situated on the short (p) arm of human chromosome 3 at position 21 (3p21). (Papapetrou and Schambach) However, it is not a fully functional Genomic Safe Harbour site because its disruption affects immune cell trafficking and can impair host defense, as seen with increased susceptibility to infections like the West Nile virus. (Glass et al., 2005; Glass et al., 2006) Rather than acting as a safe harbour site, CCR5 is more commonly used for functional knockouts due to high tolerance. (Papapetrou and Schambach; Sadelain et al.) Functional knockout tolerance refers to the ability of a gene to undergo loss-of-function without causing major phenotypic consequences. Thus, while CCR5 disruption is used in clinical contexts, it is often not the optimal site for transgene integration. (Papapetrou and Schambach)

2.4 H11 (Hipp11) Locus

The H11 (Hipp11) locus is an intergenic locus that was initially discovered in the mouse genome on chromosome 11 between the Eif4enif1 and Drg1 genes. (Papapetrou and Schambach) Its capacity to allow stable and ubiquitous expression of transgenes without interfering with endogenous gene activity has made it a popular Genomic Safe Harbour in murine models. (Papapetrou and Schambach) The locus is found in a transcriptionally permissive site, with little interference from surrounding regulatory factors, and is therefore appropriate for the expression of the locus in various tissues. (Papapetrou and Schambach)

H11 orthologous regions have been studied in humans, but are less well-characterized than classical sites like AAVS1. (Papapetrou and Schambach) Early research indicates that these areas could have comparable intergenic spacing and chromatin accessibility patterns, yet their safety and functional stability are not adequately tested. (Papapetrou and Schambach) Notably, although the murine H11 locus exhibits numerous features of a perfect Genomic Safe Harbour, cross-species variations in genome structure, regulatory framework, and chromatin dynamics restrict direct translation. (Papapetrou and Schambach)

Therefore, although H11 is an encouraging model to use to identify intergenic safe harbour sites, additional research is needed to establish whether similar loci in the human genome can satisfy strict clinical standards. (Papapetrou and Schambach)

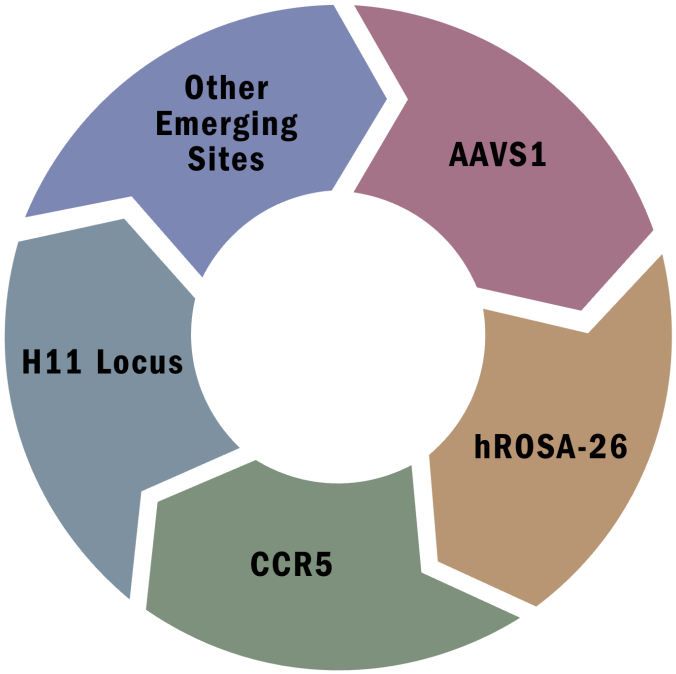

Figure1. Graphical representation of known sites (original figure created by the author)

2.5 Other Emerging Sites

Recent research has also found other candidate Genomic Safe Harbour sites that seek to overcome the shortcomings of classical loci by focusing on intergenic positioning and less regulatory interference. (Pellenz et al.) The CLYBL locus on chromosome 13 represents an early-stage candidate that has demonstrated encouraging expression stability *in vitro* but requires extensive long-term and *in vivo* evaluation before clinical consideration. (Pellenz et al.) In contrast to AAVS1, CLYBL is not located in a transcription unit, which minimizes the chances of insertional mutagenesis and transcriptional interference. (Pellenz et al.) Functional analyses have shown that it is expressed in a consistent manner in a variety of cell types, but long-term stability and safety in developmental stages need to be confirmed. (Pellenz et al.)

The TIGRE locus (Transgene Integration by Genomic Regulatory Elements), first described in mouse models, is a strategy, not a fixed site. (Papapetrou and Schambach) It takes advantage of genomic regions that have good regulatory conditions to allow high and stable expression of transgenes. (Papapetrou and Schambach) TIGRE-based systems often incorporate enhancer elements and insulator sequences to minimize positional effects and silencing. (Papapetrou and Schambach) Although they are very effective in controlled experimental systems, their application to human genomic systems is still hindered by the variability in regulatory landscapes. (Papapetrou and Schambach)

Together, these emerging loci reflect a shift toward intergenic, transcriptionally permissive regions and the incorporation of regulatory engineering approaches. (Pellenz et al.) But none of them has so far been validated to the extent that they can be used universally in clinical practice. (Papapetrou and Schambach; Sadelain et al.)

3. INNOVATIVE APPROACHES FOR IDENTIFICATION OF PUTATIVE SITES

3.1 COMPUTATIONAL APPROACH

A computational analysis identified 25 unique putative universal Genomic Safe Harbour sites in the human genome. It filtered for safety criteria such as distance from genes, oncogenes, miRNAs, lncRNAs, location outside transcription units, ultra-conserved regions, and DNaseI hypersensitivity clusters, as well as accessibility features like residing in active chromosomal compartments and association with ubiquitously expressed, low-variance genes. (Autio et al.) Of these 25 sites, 3 were experimentally validated *in vitro* using human embryonic stem cells (hESCs) and their differentiated progeny: (i) Pansio-1 (chromosome 1), (ii) Olônne-18 (chromosome 18), and (iii) Keppel-19 (chromosome 19). (Autio et al.) These three sites showed stable transgene expression, minimal disruption of the native transcriptome, and no interference with directed differentiation into various cell lineages (neuronal, liver, cardiac, and pancreatic β-cells). (Autio et al.)

3.2 ENDOGENOUS PARVOVIRAL ELEMENTS

A novel approach identified 102 potential human Genomic Safe Harbour sites by investigating a dataset of endogenous parvoviral elements (EPVs) integrated into host species germlines over millions of years. (Quezada-Ramírez et al.) From these, 17 loci (8 intergenic, 9 intronic) were selected for characterization in primary human CD34+ hematopoietic stem and progenitor cells (HSPCs). (Quezada-Ramírez et al.) This approach identified at least 9 promising Genomic Safe Harbour sites suitable for gene therapy of blood disorders, with sustained transgene expression in both erythroid and myeloid phenotypes. (Quezada-Ramírez et al.) These include Dep. 2, Dep. 3, Dep. 55, Prot. 218, Prot. 2, Prot. 181 (widespread expression), and Dep. 13, Dep. 1, Ap. 102, Prot. 176 (myeloid lineage-biased expression). (Quezada-Ramírez et al.) Further evaluation *in vivo* is necessary for these sites. (Quezada-Ramírez et al.) Additional candidate loci may emerge from the remaining mapped loci and may be suitable for other stem and progenitor cells. (Quezada-Ramírez et al.)

4. CONSERVATION OF SITES

The role of conservation in the identification of Genomic Safe Harbour sites remains controversial. One perspective argues that ultra-conserved regions, that is, segments of DNA that show complete identity between various species, should be avoided. (Sadelain et al.) The fact that these sites have had no mutations across lineages implies that they are essential and must not be tampered with. (Sadelain et al.) On the other hand, the second school of thought, as seen in the method of identification using endogenous parvovirus elements, focuses on these ultra-conserved regions since they have historically exhibited stability and an aversion to mutation. (Quezada-Ramírez et al.) This highlights a fundamental limitation in the usage of conservation as a criterion. (Papapetrou and Schambach) Most analyses subscribe to the first school of thought and avoid ultraconserved regions, but they also demonstrate robustness as previously discussed. (Autio et al.) However, this stability does not always equate to functional predictability. (Papapetrou and Schambach) Oftentimes, these regions can overlap with other elements, such as non-coding RNA or oncogenes. (Hauptman and Glavač; Bracken et al.) Increased expression or modification of these elements could result in unforeseen consequences. (Hauptman and Glavač; Bracken et al.) Therefore, it is ill-advised to consider the conservation of the site as a standalone criterion and instead consider it within the context of the larger genome. (Papapetrou and Schambach; Sadelain et al.) Integrating a variety of factors, including but not limited to tissue type, chromatin accessibility, histone markers, genetic density, transcriptional start site distance, distance from exons, and distance from oncogenes, would produce a more nuanced assessment of the site's ability to integrate transgenes predictably and ensure functional stability. (Autio et al.; Hyun et al.; Terranova et al.) Further, conservation as a criterion should be paired with cell type, since the regulatory landscape itself is prone to changes across tissues. (Papapetrou and Schambach) This ambiguity in the implications of conservation further strengthens the argument for a context-dependent and situation-specific method of identification for putative sites. (Papapetrou and Schambach; Sadelain et al.) Ultimately, conservation should act as an informing criterion for the selection of sites in conjunction with other parameters, but should not act as

the primary mechanism behind this identification. (Papapetrou and Schambach)

5. CRITERIA FOR IDENTIFICATION

Sites are required to be outside gene transcription units and distant from coding and non-coding genes. (Sadelain et al.) Proposals suggest being at least 50 kb away from the transcriptional start site (TSS) to avoid common mechanisms of insertional oncogenesis, where vector integration often occurs upstream of dysregulated genes. (Sadelain et al.; Hacein-Bey-Abina et al.) Genomic Safe Harbours are required to be at least 300 kb away from genes known to play a role in cancer in humans or model organisms, as this minimises the risk of activating cancer-related pathways and increasing the expression of oncogenes. (Sadelain et al.; Autio et al.) Similarly, a distance of at least 300 kb from miRNA genes is recommended, since this prevents disruption of regulatory non-coding elements that control gene expression networks. (Sadelain et al.; Bracken et al.) Moreover, the sites should be more than 100 kb away from any lncRNA. (Sadelain et al.) Disruption or dysregulation of ncRNAs, including lncRNAs, can lead to diseases such as cancer. (Hauptman and Glavač; Ishibashi et al.) To avoid disrupting essential genetic elements, Genomic Safe Harbours should also be located outside ultra-conserved regions. (Sadelain et al.) Inactivation of genes indispensable for cell viability (approximately 2,000 genes in humans) should be avoided, although essentiality can be context-dependent. (Wang et al.) Additionally, sites should reside in active chromosomal compartments that remain consistently open across multiple cell and tissue types, ensuring accessibility of the genomic region for transcriptional machinery. (Autio et al.) Regions generally should avoid DNase I hypersensitive clusters, as these are enriched in transcription factor binding sites and regulatory elements that may interfere with transgene expression. (Autio et al.; Sadelain et al.) Favorable epigenetic marks for open chromatin (e.g., H3K4me3, DNase I hypersensitive sites) are associated with improved integration efficiency and expression. (Hyun et al.; Terranova et al.) The location relative to topologically associated domains (TADs) is also important, and it is preferable to avoid TADs enriched in cancer-related genes or to target TAD borders. (Dixon et al.) Candidate Genomic Safe Harbours in one study were located more than 80 kb away from TAD borders. (Autio et al.) Markers of active chromatin (e.g., H3K4me1, H3K27ac) are desirable, as they correlate with high transcriptional activity, whereas integration into repressive chromatin (e.g., H3K27me3) may lead to silencing. (Hyun et al.; Terranova et al.) A convergent orientation of flanking genes is recommended to reduce interference with promoter regions, as regulatory elements are typically upstream of the gene. (Sadelain et al.)

5.1 FUNCTIONAL VALIDATION REQUIREMENTS

Integration should result in little or no change in mRNA expression of nearby genes, with most differentially expressed genes (DEGs) overlapping with wild-type controls. (Autio et al.) The targeting system (e.g., CRISPR/Cas9) should not create unintended edits elsewhere in the genome. (Anzalone et al.) The transgene should remain integrated over multiple generations without reducing cellular fitness. (Papapetrou and Schambach; Sadelain et al.) Expression should be stable across numerous passages in pluripotent stem cells and remain consistent across differentiated cell types (e.g., neuronal, liver, cardiac, pancreatic β-cells) under both constitutive and inducible promoters. (Autio et al.) Maintenance of

pluripotency and differentiation potential should be preserved, such that integration does not disrupt the cell's ability to differentiate into multiple lineages. (Autio et al.) Genomic Safe Harbours ideally should accommodate multiplex gene insertion and complex genetic circuits without affecting neighboring genes. (Pellenz et al.)

5.2 CONTEXT DEPENDENCY

While the criteria outlined in Section 5.1 provide a structural foundation for GSH identification, they cannot be defined as absolute constants. (Papapetrou and Schambach) Their effectiveness depends on variables including cell type and developmental stage, since every tissue has a unique function and therefore a unique genetic architecture. (Papapetrou and Schambach; Sadelain et al.) Chromatin accessibility, enhancer–promoter interactions, and transcription factor availability all shift based on context, meaning a locus that supports stable transgene expression in one cell type may be silenced or dysregulated in another. (Hyun et al.; Dixon et al.; Papapetrou and Schambach; Sadelain et al.)

Computational identification frameworks face the same limitation. By applying uniform filtering criteria to static, population-level genomic data, methods such as Autio et al.'s pipeline and SHIP inherently abstract away the cell-type-specific chromatin dynamics that determine real-world GSH behaviour. It is worth noting, however, that variability within a defined cell type is considerably more predictable than variability across the genome at large, making context-specific identification a more reliable strategy than genome-wide searches optimised for universality.

These considerations collectively imply that the validation of a putative GSH must be performed in the specific cell type and developmental context in which the therapeutic transgene will be expressed. Context-specificity should therefore be treated not as a complicating factor but as a design criterion for future identification frameworks. (Papapetrou and Schambach; Autio et al.)

6. COMPUTATIONAL AND MODELLING TOOLS

Beyond primarily biological methods of identification of safe harbour sites, there have been a myriad of methods using computational and modelling methods. (Autio et al.; Quezada-Ramírez et al.; Leitão et al.) First, Autio et al.'s Workflow, a Python pipeline filtering genomic data, uses binary exclusion and inclusion criteria to determine sites that adhere to stringent identification boundaries using ENCODE, ENSEMBL, COSMIC, UCbase 2.0, and GTEx databases. (Autio et al.) They excluded regions within 50 kb of any TSS, 300 kb of cancer genes or miRNAs, and 100 kb of lncRNAs. (Autio et al.) Further, they excluded ultraconserved regions and DNase I hypersensitivity clusters, identifying which regions were in active chromosomal compartments across cell types. (Autio et al.) Then, they used a BLAT search to remove all the non-unique sequences to filter down the search results. (Autio et al.) Secondly, SHIP (Safe Harbour Identification Program) is a GUI-based Python program used to find intergenic GSH candidates in annotated eukaryotic genomes. (Leitão et al.) The users of this platform are able to define intergenic length and gene orientation while the code outputs regulatory profiles. (Leitão et al.) It excludes genes,

ncRNAs, pseudogenes, centromeres/telomeres, mobile elements, replication origins, recombination regions, ESTs, TF binding sites, tRNAs, and lincRNAs. (Leitão et al.) Thirdly, the previously discussed method of EPV modelling by Quezada-Ramírez et al maps endogenous parvovirus elements (EPVs) across species and identifies orthologous loci in humans via genome alignments and synteny, following which it confirms the locations of these sites with BLAT or BLASTN. (Quezada-Ramírez et al.) Computational tools have been crucial in the identification of putative sites, some of which have been proven to be applicable in clinical contexts. (Autio et al.; Quezada-Ramírez et al.; Leitão et al.) The implications of this context-specificity for computational identification methods are discussed in Section 5.2. To enhance these methods of identification, it would be advisable to insert more context-specific data, such as the type of transgene to be inserted and the intention of this insertion. (Papapetrou and Schambach; Autio et al.) A direct comparison of their relative strengths and applicability is provided in Section 6.1.

6.1 COMPARATIVE ANALYSIS OF IDENTIFICATION FRAMEWORKS

Framework	**Approach**	**Key Criteria**	**Strengths**	**Limitations**	**Translatability**
Autio et al.	Binary exclusion Python pipeline using ENCODE / ENSEMBL / COSMIC / GTEx	TSS, oncogene, miRNA, lncRNA distances; ultra-conserved regions; DNase I clusters	Stringent; reproducible; experimentally validated in hESCs	Human-specific; static data; no dynamic chromatin; ignores cell-type variation	3 sites validated *in vitro* across multiple lineages
SHIP	GUI Python tool for intergenic GSH candidates in any annotated eukaryote	Intergenic length, gene orientation, ncRNAs, mobile elements, TF binding sites	Species-agnostic; user-configurable; outputs regulatory profiles	No in vitro validation pipeline; output quality depends on annotation completeness	No human clinical validation yet
EPV mapping	Evolutionary: maps ancient parvovirus integration sites conserved across species	Orthologous loci; synteny; BLAT/BLASTN confirmation; erythroid/myeloid expression	Biologically motivated; identifies sites tolerated over millions of years	Relies on evolutionary conservation as a proxy; limited to parvovirus-integration loci	9 sites validated in primary human CD34+ HSPCs

Autio et al. and SHIP both use annotation-driven binary filtering, but differ in scope and flexibility. Autio et al. prioritises stringency with fixed distance thresholds calibrated to oncogenic risk, while SHIP is user-configurable and organism-agnostic. Neither approach intrinsically validates expression; they identify candidates only. EPV mapping rests on a fundamentally different rationale: it treats the

evolutionary stability of ancient viral integration sites as a biological proxy for genomic tolerance. This is more empirically grounded than annotation filtering but is constrained to loci where parvoviral integration happens to occur, which may not represent the optimal expression landscape for therapeutic transgenes. All three methods rely on static genomic datasets and are therefore unable to capture dynamic, cell-type-specific chromatin remodelling. None integrates single-cell epigenomic data (e.g., scATAC-seq) or three-dimensional genome architecture (Hi-C).

7. FUNCTIONAL APPLICATIONS FOR HUMAN GENOME SITES

7.1 MODELING OF B-THALASSEMIA AND THERAPEUTIC GENE EXPRESSION (*IN VITRO* STUDIES; iPSC MODEL)

In a disease model of β-thalassemia, a therapeutic β-globin transgene was successfully inserted into genomic sites meeting the Genomic Safe Harbour criteria in patient-derived induced pluripotent stem cells (iPSCs). (Papapetrou et al.) This achieved therapeutic expression levels without dysregulating endogenous genes. (Papapetrou et al.) This approach is considered a potentially clinically translatable method for autologous cell and gene therapy. (Papapetrou et al.)

7.2 CORRECTION OF HEMOPHILIA B IN MICE (*IN VIVO* STUDIES; MURINE MODEL)

A human coagulation factor IX gene was inserted into the liver-specific albumin gene locus using a recombinant adeno-associated virus vector. (Li et al.) This intervention was demonstrated to ameliorate hemophilia B in mice, highlighting the utility of carefully studied loci as Genomic Safe Harbours. (Li et al.)

7.3 DISEASE MODELING AND DRUG TESTING FOR HYPERTROPHIC CARDIOMYOPATHY (HCM) USING AAVS1 (*IN VITRO* STUDIES; hiPSC-CM MODEL)

The AAVS1 locus in human iPSCs was targeted to insert a genetically encoded calcium indicator (R-GECO1.0). (Bhagwan et al.) These hiPSCs were differentiated into cardiomyocytes (hiPSC-CMs) to model HCM-associated mutations (c.MYH7C9123T and c.ACTC1G301A). (Bhagwan et al.) This system enabled live Ca2+ imaging in hiPSC-CMs to investigate abnormal Ca2+ transients caused by HCM mutations. (Bhagwan et al.) Crucially, the aberrant Ca2+ transient phenotype was successfully rescued with drug treatment using a combination of ranolazine and dantrolene. (Bhagwan et al.) This demonstrated the system's utility as an *in vitro* alternative for drug testing for HCM, potentially reducing reliance on animal models. (Bhagwan et al.)

7.4 GENOMIC EDITING AND GENE ACTIVATION IN HUMAN CANCER CELL LINES USING SHS231 (*IN VITRO* STUDIES; CANCER CELL LINES)

The newly identified SHS231 site on human chromosome 4 supported the stable expression and function of various transgene-encoded proteins, including fluorescent markers and Cas9 protein variants. (Pellenz et al.) SHS231-integrated Cas9 proteins were used to introduce a large (17 kb) gRNA-specified deletion in the PAX3/FOXO1 fusion oncogene in human rhabdomyosarcoma cells. (Pellenz et al.) As a Cas9-VPR

fusion protein, it was shown to upregulate the expression of the muscle-specific transcription factor MYF5 in human rhabdomyosarcoma cells. (Pellenz et al.) These demonstrate precise gene editing and regulation capabilities. (Pellenz et al.)

7.5 STABLE, LINEAGE-SPECIFIC TRANSGENE EXPRESSION IN HUMAN HEMATOPOIETIC STEM AND PROGENITOR CELLS (HSPCS) (*IN VITRO* STUDIES; PRIMARY HUMAN CD34+ CELLS)

Seventeen endogenous parvovirus elements (EPV) orthologs were evaluated as potential Genomic Safe Harbours in primary human CD34+ HSPCs. (Quezada-Ramírez et al.) Six of these loci demonstrated sustained transgene (eGFP) expression in both erythroid and myeloid phenotypes, while three loci showed myeloid-specific expression. (Quezada-Ramírez et al.) This led to the identification of at least nine promising Genomic Safe Harbour sites suitable for gene therapy of blood disorders. (Quezada-Ramírez et al.) The edited CD34+ HSPCs maintained their multipotency with either constitutive or lineage-specific transgene expression, such as preferential myeloid expression from Ap. 102 and Dep. 1 loci. (Quezada-Ramírez et al.) Persistent transgene expression was observed in reticulocytes derived from most analyzed loci (except Prot. 218). (Quezada-Ramírez et al.)

7.6 STABLE CONSTITUTIVE AND INDUCIBLE TRANSGENE EXPRESSION ACROSS DIVERSE HUMAN CELL TYPES (*IN VITRO* STUDIES; hESC MODEL)

Computationally defined sites Pansio-1, Olônne-18, and Keppel-19 were validated *in vitro* using human embryonic stem cells (hESCs). (Autio et al.) They supported stable constitutive transgene expression (using a Clover fluorophore) in hESCs and their differentiated progeny, including neuronal, liver (hepatocyte-like), cardiac (cardiomyocyte-like), and pancreatic β-cells. (Autio et al.) These sites also demonstrated the ability to support inducible transgene expression (Tet-inducible GFP transgene) in hESCs. (Autio et al.) The targeted hESCs were able to form teratomas comprising tissues from all three germ lineages (endoderm, ectoderm, and mesoderm), indicating no interference with differentiation. (Autio et al.)

7.7 CHALLENGES AND LIMITATIONS:

Although significant advances have been made in the discovery of candidate Genomic Safe Harbour sites, there are still a number of critical issues. (Papapetrou and Schambach; Naldini) The main weakness is that there are no standardized and universally accepted criteria for defining and validating such sites. (Papapetrou and Schambach; Sadelain et al.) The existing literature uses heterogeneous methods, such as computational filtering and experimental validation, which lead to discrepancies and a lack of comparability between the results. (Autio et al.; Leitão et al.)

The other significant problem is the context-specificity of the genome. The level of chromatin accessibility, epigenetic modifications, and regulatory interactions differ greatly among cell types, developmental stages, and environmental conditions. (Hyun et al.; Dixon et al.) Consequently, a locus that is a safe harbour in one situation can have erratic conduct in a different situation, which negates the idea

of a universally applicable site. (Papapetrou and Schambach)

Also, many candidate sites have been tested only *in vitro* or in small cell populations, and there is a lack of *in vivo* or long-term data to support stability, safety, and the lack of deleterious effects such as insertional oncogenesis. (Hacein-Bey-Abina et al.; Bushman) Subtle transcriptional perturbations, which might not be obvious, can have cumulative or delayed effects. (Hacein-Bey-Abina et al.)

Computational methods are efficient but limited by incomplete or biased genomic data and are not always able to resolve higher-order chromatin interactions and three-dimensional genome architecture. (Autio et al.; Leitão et al.) Moreover, the use of model organisms creates translational constraints because of interspecies genomic variations. (Chen et al.; Cantos et al.)

Collectively, these issues underscore the importance of integrative, context-sensitive models and more rigorous, longitudinal validation approaches. (Papapetrou and Schambach; Naldini)

8. CONCLUSION AND FUTURE PERSPECTIVE

The evidence reviewed here points toward a clear direction for the field: rather than pursuing universally applicable Genomic Safe Harbour sites, the research community should redirect effort toward context-adaptive identification frameworks that account for cell type, developmental stage, and therapeutic application as primary design inputs. Further, as predominantly explored in the case of conservation, no criterion by itself can be used as an absolute parameter and must be analysed through the lens of its location and functionality in the greater genomic architecture. Moreover, to improve accuracy and reduce the considerable time required for solely biological methods, it is advisable to integrate computational and modelling tools, as seen in Autio et al.'s Python workflow, the Safe Harbor Identification Program, and endogenous parvovirus integration site mapping. Further, integrative frameworks combining conservation, epigenetic features, genomic distance metrics, and functional annotations would avoid the reliance on a single filtering criterion and provide a more holistic analysis. While a computational approach has a profusion of benefits, all sites must, naturally, be extensively wet-lab tested before any clinical therapies or applications. In these preliminary tests, the sites must undergo a variety of functional integrations. However, this dual-faceted approach does act as a bottleneck since discrepancies can arise due to either incomplete datasets or biological differences in test subjects. This, in turn, brings out another significant issue. In the past few decades, the quality of chromosomal and DNA-based datasets has improved exponentially; yet, there is still a need for higher-quality and more diverse data that can reduce biases. Overall, rather than prioritising ubiquitously functional Genomic Safe Harbour sites, the research surrounding this pursuit should be applied towards more context-dependent and adaptive frameworks to ensure safer and more predictable transgene integration.

ACKNOWLEDGEMENT:
I would like to thank my parents and school teachers for their constant motivation and support.

CONFLICT OF INTEREST:
There is no potential conflict of interest to disclose.

AUTHOR'S CONTRIBUTION:
Nandini Pangal conceptualised the study, did literature survey, and corrected and edited the manuscript.

FUNDING INFORMATION:
There is no funding information to disclose.

REFERENCES:

1. Jinek, Martin, et al. "A Programmable Dual-RNA-Guided DNA Endonuclease in Adaptive Bacterial Immunity." *Science*, vol. 337, no. 6096, 17 Aug. 2012, pp. 816–21. doi:10.1126/science.1225829.
2. Papapetrou, Eirini P., and Axel Schambach. "Gene Insertion Into Genomic Safe Harbors for Human Gene Therapy." *Molecular Therapy*, vol. 24, no. 4, Apr. 2016, pp. 678–84. doi:10.1038/mt.2016.38.
3. Autio, Matias I., et al. "Computationally Defined and In Vitro Validated Putative Genomic Safe Harbour Loci for Transgene Expression in Human Cells." *eLife*, vol. 13, 2 Jan. 2024, e79592. doi:10.7554/eLife.79592.
4. Quezada-Ramírez, Marco A., et al. "Identification of Genome Safe Harbor Loci for Human Gene Therapy Based on Evolutionary Biology and Comparative Genomics." *Molecular Therapy: Methods & Clinical Development*, vol. 31, 2023, p. 101131. doi:10.1016/j.omtm.2023.101131.
5. Leitão, Marcelo C., et al. "SHIP Identifies Genomic Safe Harbors in Eukaryotic Organisms Using Genomic General Feature Annotation." *Scientific Reports*, vol. 15, no. 1, 28 Feb. 2025, p. 7193. doi:10.1038/s41598-025-91249-9.
6. Pellenz, Sandrine, et al. "New Human Chromosomal Sites with 'Safe Harbor' Potential for Targeted Transgene Insertion." *Human Gene Therapy*, vol. 30, no. 7, July 2019, pp. 814–28. doi:10.1089/hum.2018.169.
7. Klatt, Daniela, et al. "Differential Transgene Silencing of Myeloid-Specific Promoters in the AAVS1 Safe Harbor Locus of Induced Pluripotent Stem Cell-Derived Myeloid Cells." *Human Gene Therapy*, vol. 31, no. 3–4, Feb. 2020, pp. 199–210. doi:10.1089/hum.2019.194.
8. Bhagwan, Jamie R., et al. "Variable Expression and Silencing of CRISPR-Cas9 Targeted Transgenes Identifies the AAVS1 Locus as Not an Entirely Safe Harbour." *F1000Research*, vol. 8, 12 Nov. 2019, p. 1911. doi:10.12688/f1000research.19894.2.
9. Irion, Stefan, et al. "Identification and Targeting of the ROSA26 Locus in Human Embryonic Stem Cells." *Nature Biotechnology*, vol. 25, no. 12, 2007, pp. 1477–82. doi:10.1038/nbt1362.

10. Glass, William G., et al. "Chemokine Receptor CCR5 Promotes Leukocyte Trafficking to the Brain and Survival in West Nile Virus Infection." *Journal of Experimental Medicine*, vol. 202, no. 8, 17 Oct. 2005, pp. 1087–98. doi:10.1084/jem.20042530.
11. Glass, William G., et al. "CCR5 Deficiency Increases Risk of Symptomatic West Nile Virus Infection." *Journal of Experimental Medicine*, vol. 203, no. 1, 2006, pp. 35–40. doi:10.1084/jem.20051970.
12. Papapetrou, Eirini P., et al. "Genomic Safe Harbors Permit High β-Globin Transgene Expression in Thalassemia-Induced Pluripotent Stem Cells." *Nature Biotechnology*, vol. 29, no. 1, Jan. 2011, pp. 73–78. doi:10.1038/nbt.1717.
13. Li, Hao, et al. "CRISPR-Cas9-Mediated In Vivo Gene Integration at the Albumin Locus Recovers Hemostasis in Neonatal and Adult Hemophilia B Mice." *Molecular Therapy: Methods & Clinical Development*, vol. 18, 30 June 2020, pp. 520–31. doi:10.1016/j.omtm.2020.06.025.
14. Frangoul, Haydar, et al. "CRISPR-Cas9 Gene Editing for Sickle Cell Disease and β-Thalassemia." *New England Journal of Medicine*, vol. 384, no. 3, 21 Jan. 2021, pp. 252–60. doi:10.1056/NEJMoa2031054.
15. Chen, Xueying, et al. "Identification of Site in the UTY Gene as Safe Harbor Locus on the Y Chromosome of Pig." *Genes*, vol. 15, no. 8, 1 Aug. 2024, p. 1005. doi:10.3390/genes15081005.
16. Cantos, Cesar, et al. "Identification of 'Safe Harbor' Loci in Indica Rice Genome by Harnessing the Property of Zinc-Finger Nucleases to Induce DNA Damage and Repair." *Frontiers in Plant Science*, vol. 5, 26 June 2014, p. 302. doi:10.3389/fpls.2014.00302.
17. Hacein-Bey-Abina, Salima, et al. "Insertional Oncogenesis in 4 Patients After Retrovirus-Mediated Gene Therapy of SCID-X1." *Journal of Clinical Investigation*, vol. 118, no. 9, 2008, pp. 3132–42. doi:10.1172/JCI35700.
18. Bushman, Frederic D. "Retroviral Insertional Mutagenesis in Humans: Evidence for Four Genetic Mechanisms Promoting Expansion of Cell Clones." *Molecular Therapy*, vol. 28, no. 2, 2020, pp. 352–56. doi:10.1016/j.ymthe.2019.12.009.
19. Hauptman, Nina, and Damjan Glavač. "Long Non-Coding RNA in Cancer." *International Journal of Molecular Sciences*, vol. 14, no. 3, 26 Feb. 2013, pp. 4655–69. doi:10.3390/ijms14034655.
19. Bracken, Cameron P., et al. "A Network-Biology Perspective of MicroRNA Function and Dysfunction in Cancer." *Nature Reviews Genetics*, vol. 17, no. 12, Oct. 2016, pp. 719–32. doi:10.1038/nrg.2016.134.
20. Wang, Tim, et al. "Identification and Characterization of Essential Genes in the Human Genome." *Science*, vol. 350, no. 6264, 27 Nov. 2015, pp. 1096–101. doi:10.1126/science.aac7041.
21. Hyun, Kwang, et al. "Writing, Erasing and Reading Histone Lysine Methylations." *Experimental and Molecular Medicine*, vol. 49, no. 4, 28 Apr. 2017, e324. doi:10.1038/emm.2017.11.
22. Terranova, Rémi, et al. "Histone Modifications for Human Epigenome Analysis." *Journal of Human Genetics*, vol. 58, no. 7, 2013, pp. 437–44. doi:10.1038/jhg.2013.66.
23. Ishibashi, Masaaki, et al. "Long Non-Coding RNAs: Potential Therapeutic Targets." *Journal of Human Genetics*, vol. 57, 2012, pp. 414–18. doi:10.1038/jhg.2012.53.

24. Dixon, Jesse R., et al. "Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions." *Nature*, vol. 485, no. 7398, 17 May 2012, pp. 376–80. doi:10.1038/nature11082.
25. Naldini, Luigi. "Gene Therapy Returns to Centre Stage." *Nature*, vol. 526, no. 7573, 15 Oct. 2015, pp. 351–60. doi:10.1038/nature15731.
26. Dunbar, Cynthia E., et al. "Gene Therapy Comes of Age." *Science*, vol. 359, no. 6372, 12 Jan. 2018, eaan4672. doi:10.1126/science.aan4672.
27. Anzalone, Andrew V., et al. "Genome Editing with CRISPR–Cas Nucleases, Base Editors, Transposases and Prime Editors." *Nature Biotechnology*, vol. 38, no. 7, 2020, pp. 824–44. doi:10.1038/s41587-020-0561-9.
28. Sadelain, Michel, et al. "Safe Harbours for the Integration of New DNA in the Human Genome." *Nature Reviews Cancer*, vol. 12, no. 1, Jan. 2012, pp. 51–58. doi:10.1038/nrc3179.