Universidade de Lisboa

Faculdade de Ciências

Departamento de Biologia Vegetal

Screening for RAG activity in haematopoietic tumours using a novel reporter strategy

Inês Gomes de Noronha Trancoso

Mestrado em Biologia Molecular Humana

2009

Universidade de Lisboa

Faculdade de Ciências

Departamento de Biologia Vegetal

Screening for RAG activity in haematopoietic tumours using a novel reporter strategy

Inês Gomes de Noronha Trancoso

Dissertação de Mestrado orientada por:

Doutora Leonor Morais Sarmento – Instituto Gulbenkian de Ciência, Oeiras

Professora Doutora Maria Margarida Telhada – Faculdade de Ciências da Universidade de Lisboa, Lisboa

Mestrado em Biologia Molecular Humana

2009

Table of contents

Acknowledgements

À Leonor Sarmento, por me ter proposto este projecto, pelo empenho na orientação do mesmo, por me ter iniciado no verdadeiro trabalho de investigação, pelos numerosos ensinamentos de "truques" laboratoriais e também pela partilha dos sucessos e insucessos, pelos dias longos e ainda pelas brincadeiras.

À Professora Margarida Telhada, por ter aceite a co-orientação desta dissertação, pela receptividade e entusiasmo que sempre demonstrou pelo meu trabalho. Também pelo enorme trabalho que lhe dei para poder ultrapassar todas as burocracias.

To Jocelyne Demengeot, for opening the doors of her group, making me feel at home. Thank you for the constant interest for my work, guidance and availability.

Aos Doutores João Barata e Andres Yunes por toda a informação que me permitiu estudar a mutação encontrada num paciente leucémico, bem como pelo fornecimento das linhas celulares.

Ao Vasco Barreto por ter contribuído com sugestões valiosas no decurso deste trabalho.

Ao Jorge Carneiro pelo aconselhamento estatístico, pelas discussões e ensinamentos.

Ao Rui Gardner e à Telma Lopes pelo apoio técnico em citometria de fluxo, pela partilha do entusiasmo nos bons resultados e pela incansável disponibilidade, até aos fins-de-semana e feriados.

A todos os colegas com quem partilhei o laboratório durante este ano, pela troca de experiências, pela boa disposição e pelo excelente ambiente de trabalho. Em particular aqueles que tiveram uma partilha mais directa do trabalho de todos os dias: Ana Catarina Martins, Andreia Lino, Marie Louise Bergman e Ricardo Paiva.

Aos colegas de Faculdade que seguiram a vida de investigação no IGC, ou fora, juntamente comigo: Alexandre Leitão, Ana Ferreira, Isadora Monteiro, João Osório, Marta Marialva, Pedro Lima, Pedro Patraquim e Sara Esteves pela partilha de experiências, pelas discussões e pelo companheirismo.

Ao Vitor por ter estado sempre a meu lado e por ter trocado jantares por companhia durante o meu trabalho fora de horas.

À minha família pelo apoio transmitido ao longo de todo o meu trabalho, em especial aos meus pais.

Ao IGC, pela hospitalidade e oportunidade de desenvolver a minha tese num Instituto de excelência, onde a qualidade científica se faz partilhando e discutindo.

List of Abbreviations

AML Acute Myeloid Leukaemia

AR Antigen Receptor

APL Acute Promyelocytic Leukaemia

ALL Acute Lymphoblastic Leukaemia

BCR B Cell Receptor

Bp base pair

cDNA complementary Deoxyribonucleic Acid

CML Chronic Myelogenous Leukaemia

CMV Citomegalovirus

cRSS cryptic Recombination Signal Sequence

DN Double-negative

DNA Deoxyribonucleic Acid

DSB Double-Strand Break

ER Efficiency of recombination

GFP Green Fluorescent Protein

HEK Human Embyonic Kidney

HRP Horse Radish Pedroxidase

IVRA *In vitro* Recombination Assay

mRFP monomeric Red Fluorescent Protein

MSCV Murine Stem Cell Virus

NHEJ Non-homologous end joining

pt post-transfection

RAG Recombination-Activating Genes

RNA Ribonucleic Acid

RSS Recombination Signal Sequence

TCR T Cell Receptor

TdT Deoxynucleotidyl Transferase

µL microliter

List of figures

Appendix

List of tables in Appendix

Abstract (Portuguese)

A recombinação V(D)J, o rearranjo somático de segmentos génicos *variable* (V), *diversity* (D) e *joining* (J), é o fenómeno responsável pela imensa diversidade de reportório de receptores de antigénio (ARs) que medeiam o reconhecimento molecular pelo sistema imunitário dos vertebrados. Os *Recombination-Activating Genes* (RAG) 1 e 2 formam a endonuclease RAG que reconhece sequências sinalizadoras de reconhecimento (RSSs) adjacentes aos segmentos génicos do AR, gerando quebras na dupla cadeia de ADN. A junção e reparação destas quebras é realizada pela actividade da maquinaria de reparação não homóloga (*non-homologous end joining* ou NHEJ) em coordenação com a RAG, ultimamente gerando ARs funcionais.

A recombinação V(D)J é regulada a vários níveis. Pensava-se que a expressão de RAG estaria restrita a determinados estadios do desenvolvimento linfocitário (células pró/pré B e T). A nível molecular, a actividade RAG dá-se pelo reconhecimento direccionado de RSSs. Estes apresentam uma estrutura *consensus*, sendo constituída por um heptâmero, um espaçador de 12 ou 23 nucleótidos e um nonâmero, apresentando poucas sequências *consensus*. Estas sequências são portanto consideravelmente degeneradas, favorecendo uma panóplia de interacções com a RAG e consequentemente de diversidade de ARs. Contudo, existem sequências semelhantes a RSSs, os RSSs crípticos (cRSSs), que se presume existirem ao longo do genoma, devido à aleatoriedade da composição nucleotídica do mesmo ou simplesmente pela acção de mecanismos evolutivos. Alguns destes cRSSs são realmente substratos funcionais da RAG. Devido a tal, a RAG também tem sido denominada de "transposase promíscua". Actualmente, existem três modelos de actividade ilegítima de RAG, o modelo de *substrate selection error*, envolvendo rearranjos entre um RSS e um cRSS ou dois cRSSs; o modelo de *end-donation*, em que os rearranjos são feitos entre um RSS e uma quebra de ADN em cadeia dupla independente de RAG; o modelo de *transposition* que ocorre no âmbito de uma reacção legítima em que, por extensão do tempo de vida do fragmento excisado, a RAG realiza uma reacção de transposição com reinserção do mesmo fragmento. Estes modelos têm como objectivo explicar como a actividade ilegítima de RAG pode eventualmente originar assinaturas nucleotídicas, e mutações como translocações, que estão na base da instabilidade genómica encontrada em tumores linfóides.

O padrão de expressão de RAG durante a hematopoiese permanece controverso. Existem evidências de expressão de RAG em estadios precoces do desenvolvimento hematopoiético, anteriormente ao estadio de comprometimento da linhagem linfóide. Adicionalmente, a presença de expressão de RAG, bem como de assinaturas nucleotídicas reminiscentes de rearranjos, foi descrita em doenças malignas hematopoiéticas quer de

fenótipo linfóide como mielóide, nomeadamente em leucemias promielocíticas agudas (APL) e leucemias mielóides agudas (AML). Contudo, o papel da RAG na origem, manutenção e progressão de instabilidade genómica na transformação hematopoiética permanece obscuro.

Vários repórteres moleculares de actividade RAG têm vindo a ser desenvolvidos, com vista a compreender os fenómenos bioquímicos e fisiológicos de actividade RAG. As técnicas de quantificação de rearranjos moleculares mediados por RAG têm normalmente requerido grandes quantidades de material biológico, bem como procedimentos intrincados e morosos. Apesar de algumas limitações terem sido ultrapassadas, e de alguns repórteres possuírem um esqueleto retroviral (em vez de extracromossomal), vantajoso por mimetizar uma condição mais fisiológica, a detecção de actividade RAG requer ainda passos secundários. Até à data, nenhum repórter reuniu um substrato retroviral com um método simples e directo de leitura de actividade RAG.

O nosso laboratório gerou recentemente o GFPi, uma nova ferramenta molecular episssomal/retroviral baseada em fluorescência, que permite uma fácil avaliação da actividade RAG em vários tipos celulares por permitir o cálculo de um valor de eficiência de recombinação (ER) de RAG através da análise de células por citometria de fluxo.

Como primeiro objectivo deste trabalho, procurámos estabelecer este novo repórter optimizando a detecção de actividade RAG no ensaio de recombinação *in vitro*, através de um aumento de expressão dos genes RAG1 e RAG2, localizados nos plasmídios utilizados neste ensaio. Concluímos que o referido aumento de expressão permitiu alargar a janela de detecção de valores de eficiência de recombinação *in vitro*.

Com vista avaliar a sensibilidade do GFPi na detecção de actividade RAG em presença de sequências de reconhecimento degeneradas, reproduzimos o *setup* experimental abordado por Kelsoe e colaboradores com o repórter clássico pJH290, desta vez com o GFPi. Para tal, variámos a sequência espaçadora de 23 nucleótidos do RSS contido no GFPi e submetemos cada repórter a ensaios de recombinação *in vitro*, de maneira a observar a repercussão em valores relativos de actividade RAG. Observámos que o GFPi é realmente sensível a variações nucleotídicas nesta estrutura nucleotídica em particular, apresentando uma tendência de valores de eficiência de recombinação semelhante à literatura, apesar de não ter reproduzido fielmente a mesma.

Acoplando as vantagens do GFPi à optimização obtida na detecção de valores de eficiência de recombinação *in vitro*, nomeadamente por amplificação ERs até então residuais, decidimos testar a aplicabilidade do GFPi na medição de ERs de uma colecção de RSSs crípticos (cRSSs) conhecidos por estarem envolvidos no desenvolvimento leucémico. Não só validámos esta ferramenta como apta a ser aplicada nestas condições como analisámos a funcionalidade destes cRSSs, já descritos na literatura. Ao contrário do que é

mencionado nesta, o GFPi permitiu classificar o cRSS do gene *stem cell leukaemia* (SCL) como possuindo uma actividade RAG detectável, apesar de residual.

Este novo ensaio permitiu-nos ainda quantificar a eficiência de recombinação RAG de um cRSS previamente não detectado, encontrado no primeiro exão do gene *phosphatase and tensin homologue deleted on chromosome 10* (PTEN), conhecido por estar envolvido na origem de leucemias. Provámos que o cRSS do gene PTEN é tão funcional como outros RSSs dos loci V(D)J, e que esta actividade RAG se correlaciona com uma mutação recentemente encontrada no mesmo local, num paciente oncológico com Leucemia Aguda Linfóide de células T.

Também usufruímos da forma retroviral do GFPi, com vista a quantificar actividade RAG endógena em linhas celulares hematopoiéticas tumorais humanas. Para tal, desenvolvemos a produção de pseudoretrovirus de GFPi, procedendo à infecção das mesmas linhas celulares, quer de origem linfóide como mielóide, de maneira a obter a forma integrada do GFPi em número baixo de integrantes por célula. Apesar de termos encontrado diferenças em actividade RAG específicas de linhagem (nomeadamente na linha de células B Reh que apresentou actividade RAG intensa), não encontrámos nenhum sinal de actividade RAG endógena promíscua num curto espaço de tempo, nas linhas celulares mielóides.

Não obstante, descobrimos que o repórter GFPi não só consegue providenciar uma medição quantitativa de forma transiente *in vitro* (na sua forma episssomal), como também o faz na sua forma integrada *ex vivo* (na sua forma retroviral), como substrato estável e mais próximo de uma condição fisiológica.

Assim, concluímos com este trabalho que o repórter GFPi reúne todos os requisitos para um repórter de actividade RAG eficiente, acoplado a um método fácil e de leitura directa. Como resultado, possuímos agora uma ferramenta robusta adequada a várias possíveis aplicações. Prevemos que o GFPi possa possibilitar a descoberta da relação entre a actividade promíscua da RAG e o desenvolvimento leucémico, que foi preliminarmente abordada neste trabalho, bem como aplicações bioquímicas (na compreensão da fisiologia da recombinação V(D)J) e biomédicas (numa perspectiva clínica, como ferramenta de diagnóstico).

Palavras-chave: RAG, recombinação V(D)J, repórter molecular, GFPi, cRSS, desenvolvimento leucémico, instabilidade genómica.

Abstract (English)

Recombination-Activating Genes (RAG) 1 and 2, form the site specific recombinase that mediates V(D)J recombination at the antigen receptor loci, a process responsible for lymphocyte diversity. RAG can also interact with degenerated recognition signal sequences (cRSSs) distributed throughout the genome and potentially induce genomic instability. To this date, no available reporter of RAG activity has gathered the versatile features of a molecular tool with a simple method of readout. Our laboratory has recently generated GFPi, a novel episomal/retroviral fluorescence-based molecular tool that allows for fast assessment of RAG activity in various cell types.

We have found that the GFPi reporter can not only provide *in vitro* quantitative measurement transiently, being sensitive to RSS sequence degeneration, but also *ex vivo* in its integrated form as a stable substrate.

We have also optimized the *in vitro* recombination assay (IVRA) for the detection of a broader window of RAG activity values. This novel assay allowed us to quantify the efficiency of RAG recombination of a selected set of cRSSs, namely a previously undetected cRSS found in the first exon of the *phosphatase and tensin homologue deleted on chromosome 10* (PTEN) gene, which is known to be involved in leukemogenesis. PTEN cRSS was proved to be as functional as other V(D)J loci RSSs, and this activity correlates with a novel mutation which was recently found in a leukaemic patient in this same site.

We also made use of the retroviral-based form of GFPi in order to quantify endogenous RAG activity in human haematopoietic tumour cell lines of lymphoid and myeloid origin. Although we have found lineage-specific differences on RAG activity (namely in Reh B-cell line which presented intense RAG activity), we found no signs of promiscuous endogenous RAG activity in myeloid cell lines in a short timespan.

Thus, we believe the GFPi reporter gathers all the requisites for an efficient RAG activity reporter, coupled with a fast and direct method of readout. Moreover, we now possess a robust tool that is suitable for a number of applications, namely unravelling the relation between RAG promiscuous activity and leukaemogenesis, which was preliminary addressed in this work.

Key-words: RAG, V(D)J recombination, molecular reporter, GFPi, cRSS, leukaemogenesis, genomic instability.

1. INTRODUCTION

1.1. Evolution of the Immune System Specificity

The generation of immune diversity is nowadays seen as a long story from the evolutionary point of view. Ever since pathogens exist, the host's immune system has endured the difficult role of detecting foreign and harmful entities that are extremely diverse, ensuring their ablation. Currently two models contain evolutionary explanations regarding the host's solutions for this struggle: a discrimination mechanism, between self and non-self entities[1] and a danger discrimination mechanism[2], with damage signals eliciting immune responses. In both cases, the mechanism allowed for an amelioration of the specificity for pathogen recognition. This immune specificity was achieved by taking advantage of DNA recombination, gene conversion[3] and hypermutation of recognition receptors[4], generating a diverse defence line. However, the innate immune system cannot cope with a great level of diversity since the generation of its germline-encoded receptors is limited by genome size[5].

The adaptive immune system, which is to date only described in vertebrates, owes its success to the appearance of new cell types, new morphological innovations[6], new mechanisms of diversity generation and increased specificity for pathogen recognition[5]. Moreover, these specialized cell types enhanced memory strategies, enabling a more efficient response to the pathogen's subsequent attack. The crosstalk within and between the innate and the adaptive immune systems was also extremely important in this process[7].

Recombination-Activating genes 1 and 2 (RAG1 and RAG2) had a pivotal role on generating the diversity of the adaptive immune system [8-11]. They codify for the endonuclease or recombinase complex (herein referred as RAG unless individual proteins are mentioned) which mediates the recombination of DNA segments, giving rise to unique antigen receptors (ARs) present in B and T lymphocytes, a phenomenon named as V(D)J recombination[11]. Due to its "cutting and stitching" properties, RAG allows for a major manipulation of the germline information contained in the DNA of lymphocytes[11, 12]. The RAG genes are present in all jawed vertebrates and RAG's target sequences are highly conserved as well (**Fig.I**, **App.**[5]). Despite the existing discussion about the time of appearance of these genes, it is thought that RAG evolved from a single transposition event which occurred approximately 500 million years ago[13]. Nevertheless, a RAG-like sequence was also found in Echinoderms[14] which may provide alternative explanations to RAG's evolution.

1.2. V(D)J Recombination

V(D)J recombination, the somatic rearrangement of variable (V), diversity (D) and joining (J) segments, is the phenomenon responsible for the immense diversity of AR repertoires that mediate molecular recognition by the vertebrate immune system[12]. The RAG

endonuclease specifically recognizes recombination signal sequences (RSSs) adjacent to AR gene segments and generates DNA double strand breaks (DSBs)[11, 12]. Joining and repair of the nicked strands are accomplished by the activity of the non-homologous end-joining (NHEJ) machinery, in coordination with RAG, to give rise to functional ARs of B and T cells (BCR and TCRs, respectively)[15].

In mammals, the final protein structure originating the AR is a heterodimer: the BCRs are composed by an immunoglobulin heavy chain (IgH) which can be coupled either with an Igκ or Igλ light chain, whereas the TCRs are either αβ or γδ protein dimmers. Thus, there are seven AR loci present in the mammalian genome.

V(D)J recombination defines several stages of lymphocyte development. Concerning the development of B cells, taking place in the bone marrow, these cells recombine the IgH locus (from Pre-Pro to Pro-B stage), rearranging one of the possible light chain loci, Igκ or Igλ, (from Pro to Pre-B stage), in order to produce the BCR. If rearranging the Igκ, the cell represses the Igλ locus rearrangement: a mechanism known as isotypic exclusion. The cells are further subjected to negative and positive selection pressures, avoiding auto-reactivity and increasing specificity to a given antigen.

Concerning T cell development, αβ-T cells and γδ-T cells arise in the thymus from a common progenitor cell originated in the bone marrow[16]. T cells rearrange the TCRβ, γ, and δ loci first at the double-negative (DN) stages (from DN2 to DN3 stage) but there is a strong bias for β-chain production. This chain couples to the invariant pre-TCRα chain, forming the pre-TCR. The pre-TCR checkpoint occurs from DN3 to DN4 stage enabling survival, proliferation and differentiation of these cells. Finally, the TCRα locus is rearranged (from DN4 to double-positive, DP, cell stage) and the cell produces the αβ-TCR. It is the TCR that allows for the positive-selection stage, where the cell commits either to the CD4 or CD8 single-positive (SP) lineage, before being subjected to negative selection pressures (which eliminate auto-reactive T cells) and exiting the thymus to the peripheral organs. γδ-T cell fate is still poorly understood.

The AR loci vary in number and order of V, D and J segments (also in between organisms[17]) and only the IgH, TCRβ and TCRγ loci bear D segments, which are located between the Vs and Js. All segments are flanked by RSSs, which are composed by a palindromic seven-nucleotide consensus sequence (heptamer), followed by a 12 or 23-nucleotide spacer and a nine nucleotide sequence (nonamer) exhibiting few consensus positions[18, 19] (**Fig.1 A**). Even though this signature is essential for recombination, these three regions can vary in nucleotide composition, allowing for a degenerated permission and favouring a wide range of interactions with RAG, consequently affecting the efficiency of the rearrangement in course. During recombination, RAG obeys a 12/23 rule, meaning that, with few exceptions, it only assembles a 12-RSS with a 23-RSS, allowing for a less promiscuous

and more targeted activity[20]. The RSSs' orientation is also important, since it determines if the sequence between the rearranged segments will be inverted or excised[21]. The usual organization of the RSS in the AR loci is such that the rearranged segments (coding sequences) remain in the chromosome and the intermediate sequence is excised, along with the RSSs (**Fig.1 A**).

V(D)J recombination is triggered during lymphocyte development by specific environmental conditions. The model nowadays accepted states that RAG targets and binds the RSSs[20] (likely binding the 12-RSS before the 23-RSS[22]), nicking one of the DNA strands precisely at the joint between the coding sequence and the RSS heptamer, exposing a 3' hydroxyl group[19] (**Fig.1 Ba**). The nicking is followed by the pairing of the recruited RSSs, with the help of RAG, forming the synapsis[19] (**Fig.1 Bb**). The DSB occurs due to a nucleophilic attack of the nicked strand, which invades the 5'-phosphorylated signal-end of the other (in a transesterification reaction, similarly to a transposition mechanism) forming a hairpin structure[19] (**Fig.1 Bc**), resulting in the excision of the fragment in between the coding sequences (RSSs included). The hairpins formed at both coding sequences, together with RAG and the blunt signal ends, form the post-cleavage complex[23]. The ubiquitous Non-homologous end joining (NHEJ) repair machinery is then recruited, resolves this structure and forms the coding joint by nicking the hairpins in a non-targeted manner (a process undertaken by the protein Artemis[24]) and connecting them to each other[19] (**Fig.1 Bd**), allowing for the loss or addition of nucleotides (in the last case provided by Deoxynucleotidyl Transferase, TdT) in the joining[25]. The signal joint is composed by the excised fragment connected through the RSSs, forming an excision circle which will be lost during cell division. Alternatively, when the signal end is joined to the coding end of the other signal, rather than forming a signal joint, forms a hybrid joint, resulting in an unproductive recombination. The same may happen when the same pairs of coding and signal joints are

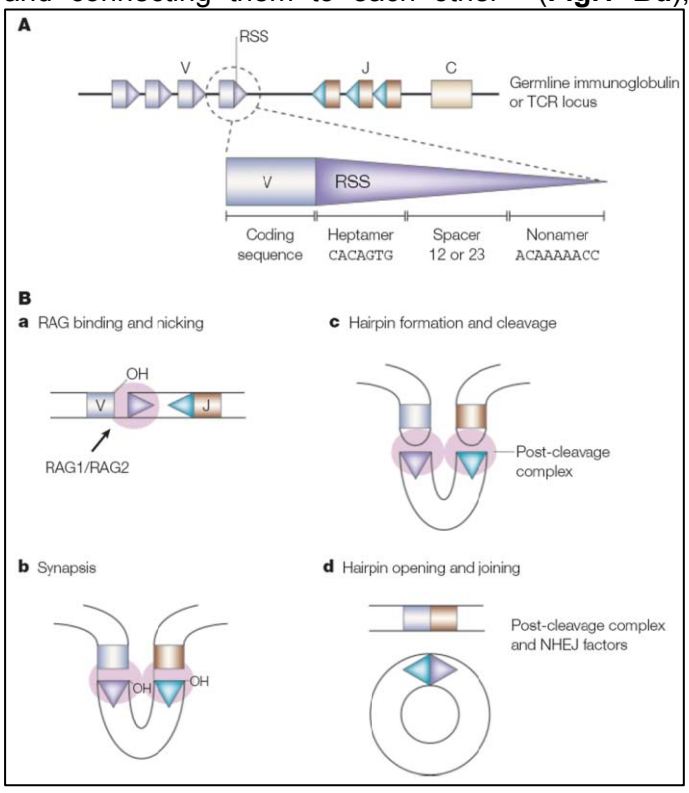


Figure 1. Schematic representation of V(D)J recombination. A. AR organization **B.** Snapshots of V-J reaction **Ba.** RAG binding and nicking **Bb.** Synapsis **Bc.** Hairpin formation and cleavage **Bd.** Haripin opening and joining.Triangles represent RSSs and rectangles represent AR gene segments. *In* Roth et al., 2003.

rejoined, forming an open-and-shut joint[26].

Due to the combinatorial properties of the V(D)J segment assembly (which can generate 10^7 possible ARs[27]), the randomness of Artemis and TdT activity, pairing of the AR heterodimers and to somatic hypermutation (in B cells) the diversity provided for AR generation is approximately 10^{10} possible ARs in B cells and 10^{17} in T cells[28].

1.2.1. Regulation of V(D)J Recombination

V(D)J recombination is regulated at several levels. Lineage and developmental-stage restrictions of V(D)J recombination are controlled by RAG expression (which occurs in two waves of expression during lymphocyte development, mentioned in **1.2.**) and by RAG's differential accessibility to the loci[29]. Despite the little information on this mechanism, it is known that locus accessibility is controlled by large-scale chromosome dynamics[30], specific histone methylation profiles[31], secondary chromatin structures, chromatin remodelers[32], cis-acting elements[33-35], transcriptional factors[36] and RSS nucleotide composition itself[37]. All of these factors dictate the chromosome region and the AR locus to be exposed to RAG, the segments to be recruited for rearrangement[38], the order of AR and segment rearrangement[39], and some of them even regulate localized RAG deposition[36]. Concerning the cell cycle regulation, V(D)J recombination is restricted to the G0/G1 stage, which is when RAG is expressed. The extracellular environment also conditions V(D)J recombination since it influences RAG and repair machinery components' accumulation and degradation in the cell (eg. **Fig.II, App.**,[30, 40]). Besides serving as an intermediate for V(D)J recombination, RAG also plays a regulatory role in several stages of V(D)J recombination, namely in the hairpin opening stage, in stabilization of the post-cleavage complex and in the recruitment of the NHEJ machinery.

Allelic exclusion is also a regulatory mechanism inherent to V(D)J recombination, namely to all B cell and some T cell AR loci[15]. It guarantees the expression of a single BCR in the cell, by silencing the allele which is not being rearranged at the AR locus, through a feedback inhibition mechanism. If the rearrangement is unproductive, then this restriction is withdrawn and the other allele is rearranged. Recently, new data provided evidence that RAG is implicated in the regulation of allelic exclusion[41].

Additionally, V(D)J recombination does not require but is enhanced *in vitro* by the presence of the high mobility group proteins 1 and 2 (HMGB1 and 2)[42], which suggests that these proteins may regulate V(D)J recombination *in vivo*.

Due to this complex orchestrated sequence of events, a model of V(D)J recombination was recently published, hypothesizing that all of these regulatory features (chromatin remodelling complexes, RAG, transcription factors and repair machinery) were localized in a single sub-nuclear compartment known as a "V(D)J factory", having RAG as a

nucleating agent[30]. This nuclear configuration would facilitate the alterations of the chromosome architecture, allelic pairing, V(D)J segment sub-nuclear repositioning, 12/23 long-range RSS assembly, all due to a more efficient coupling of events and a more regulated environment.

1.3. RAG

The organization of the RAG locus is not common, since both genes are located close to each other (separated by approximately 8kb), contain single large coding exons, and are convergently transcribed. This organization, as well as RAG's ability to potentially induce a transposition-like reaction during V(D)J recombination, was the starting point for RAG's evolutionary hypothesis as having been originated from a transposition event[43] (mentioned in **1.1.**).

The murine RAG proteins have 1040 (RAG1) and 527 (RAG2) residues each. Core RAG1 bears the binding domains required for activity: a heptamer and a nonamer binding domain (recently confirmed by crystal structure to bind the nonamer[44]) which bind the RSS, and a RAG2 binding site. Interestingly, the Zn ions found to bind core RAG-1 are determinant for RAG-cleavage activity[45]. Although it is not known if RAG2 has additional RSS binding sites (since some RAG-2 mutants impair DNA binding[46]), it is accepted that, as RAG1, RAG2 holds domains that induce DNA cleavage and hairpin formation[46, 47].

Paradoxically, RAG1 was demonstrated to be more promiscuous than previously thought, since it may bind non-RSS sequences. This effect is masked by the RAG2 protein, which was recently proved to avoid RAG1 mistargeting of RSS binding[48]. RAG2 was also described as playing a role at enhancing RSS recognition, increasing RAG1 affinity for DNA by 20-fold[49]. The non-core RAG domain functions lie mainly on the C-terminal portion of RAG2[50]. This region suppresses RAG-induced transposition[51] (mentioned in **1.3.2.**), restricts V(D)J recombination to the G0/G1 cell cycle stage and contains a PHD motif which recognizes H3 methylated histone (found in the V(D)J segments in enriched amounts at specific residues)[31]. Overall, RAG1 is currently seen has having the primary specific binding and cleavage activities, with RAG2 and HMGB1/2 functioning as its cofactors.

1.3.1. Regulation of RAG expression

As previously referred in **1.2.1.**, regulation of RAG expression directly influences V(D)J recombination. The controlling mechanisms underlying RAG expression fluctuations during lymphocyte development are still not well understood. The regulation of the RAG locus by cis-acting elements and transcription factors in a lineage-specific manner[52-54] does not fully explain this phenomenon. However, greater attention has been given to other regulatory mechanisms such as cell cycle signalling, which regulates RAG degradation at the

protein level[55] and AR signalling[53, 56] which, in turn, regulates RAG transcription levels. Additionally, it was thought that the RAG locus was controlled in a similar way to the AR loci, i.e. that it was meant to be closed unless an instructive environmental signal induced chromatin opening (such as cis-acting elements or transcription factors). Nevertheless, it was not until recently that a new finding provided the hint that the locus is exposed: NWC, a third evolutionarily conserved gene located within the RAG locus, was found to be ubiquitously expressed and controlled by the intragenic RAG2 promoter. Moreover, the exception was observed in immature and B/T mature lymphocytes in which its expression was driven by the RAG1 promoter, transcribed as a RAG1-NWC hybrid mRNA, therefore possibly representing a natural antisense transcript for RAG2 and consequently regulating RAG expression in a lineage-restricted and time-dependent manner[57]

It is also thought that receptor editing requires regulation of the RAG locus in immature B lymphocytes; although it is an unresolved issue, some claim mature B lymphocytes re-acquire RAG expression and undergo receptor editing too.

RAG expression pattern during haematopoiesis remains controversial. There is evidence showing that RAG is expressed at earlier stages of haematopoietic development, prior to lymphoid commitment[58]. Furthermore, occasional V(D)J rearrangements have been observed in dendritic cells[59, 60], suggesting that RAG activity it is not restricted to the lymphoid haematopoietic compartment.

1.3.2. RAG and genomic instability

The genomic instability elicited by RAG during V(D)J recombination can be seen as the counterpart of the immune system's physiology. However, that counterpart is astonishingly probable: taking place on a daily basis, RAG induces loads of targeted DSBs waiting to be repaired in an adequate manner within the lymphocyte population. Due to this demanding task of accurately remodelling the genome, V(D)J recombination has been seen as "a disaster waiting to happen"[61]. Oddly, RAG-mediated genomic instability events seem to rarely trigger oncogenesis in a lifetime. It is thought that the cell's protective regulatory properties and their influence on cell proliferation may be the reason [19].

Sequences similar to RSSs, cryptic RSSs (cRSSs), are predicted to exist throughout the genome[62], originated by chance or due to evolutionary mechanisms[63]. They usually bear conserved heptamer and nonamer motifs, which flank a 12- or 23-spacer-like sequence. Although some nucleotide positions are more required than others, all of these regions may vary in nucleotide composition. In the human genome, the predicted number of cRSSs is controversial: Lewis et al. prediction reaches 10 million[62] whereas the *in silico* prediction from Cowel et al. states it is ten-fold less[63]. Some of these cRSSs are functional substrates for RAG. Therefore, RAG has also been named "promiscuous transposase"[19]. Currently, three

models of RAG's illegitimate activity have been proposed[19]: the substrate selection error model (**Fig.2 A**), involving rearrangements between a RSS and a cRSS or two cRSSs; the end-donation model (**Fig.2 B**), in which the rearrangements involve a RSS and a DSB with a RAG-independent origin; the transposition model[10] (**Fig.2 C**), which occurs in the frame of a legitimate reaction, in which by extending the half-life of the excised fragments, RAG drives a transposition reaction and fragment reinsertion. In all three types of events, the resulting configuration of the DNA sequence serves as a rearrangement signature of the event that took place. These signatures turned out to be extremely useful when studying RAG activity outside the lymphoid compartment and/or its improper activity in pathological conditions (as being associated to oncogenesis). All of these three phenomena may trigger oncogenesis in several ways (for example, bringing an oncogene in proximity to a strong promoter or enhancer, truncating a tumour suppressor protein or shifting the DNA sequence of a tumour suppressor gene out of frame).

Haematopoietic malignancies display a very high incidence of chromosomal aberrations and the mechanisms underlying this genomic instability are still under intense scrutiny. Regarding lymphoid leukaemias, it has been shown that, in some T-cell Acute Lymphoblastic Leukaemias (T-ALL), RAG expression persists and is found in T cell maturation stages normally devoided of RAG activity[64].

Some cRSS were found to be recruited for translocation events, leading to leukaemogenesis. Following the substrate selection error model, cRSSs near oncogenes/tumour suppressors have been described to be recruited for translocations involving one of the AR loci: the t(11;14)(q13;q32) translocation of the cyclin D1 gene and IgH leading to mantle cell lymphoma[65]; the t(9;14)(p13;q32) translocation of Pax5 and the IgH locus leading to lymphoplasmacytic lymphomas[66]; the t(11;14)(p13;q11) translocation of Lmo2 and TCRδ locus[67], the t(1;14)(p34;q11) translocation of TAL-1 (or SCL) and TCRδ locus[68] and the t(7;9)(q34;q32) translocation of TAL-2 and TCRβ locus[68], all leading to T-ALL. cRSSs can also be recruited along with other cRSSs: that is the case of the 1p32 SCL/SIL mutation, which is the most frequent genetic event leading

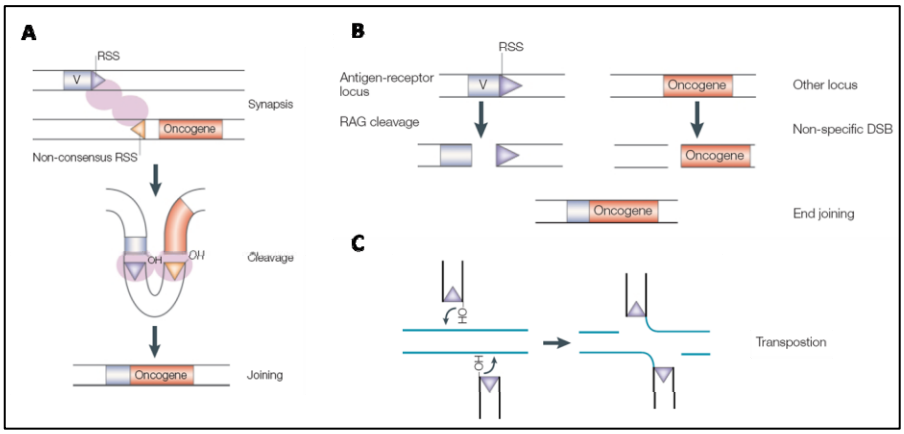


Figure 2. Models of RAG illegitimate activity. A. Substrate selection error model; **B.** End-donation model; **C.** Transposition model. Triangles represent RSSs, blue rectangles represent AR gene segments and red rectangles represent an oncogene. *In* Roth et al. 2003

to T-ALL, in which a 100kb is deleted[69].

RAG expression and rearrangement signatures have been described in haematopoietic malignancies with a myeloid phenotype, such as Acute Promyelocytic Leukaemia (APL)[70] and Acute Myeloid Leukaemia (AML)[71, 72]. Little is known about RAG's influence in solid tumourigenesis. However, an isoform of the EGF receptor (EGFRvIII) is generated in the brain through internal deletion mutations on specific exons. Since the gene is flanked by cRSSs it was suggested to be rearranged by a RAG-mediated event, leading to gliomas[73].

1.3.3. Quantification of RAG activity

Several molecular reporters have been developed in order to understand the biochemical[21] and physiological[74] features of RAG activity (**Fig.3**). Extrachromosomal classical reporters, such as pJH290 served as molecular tools for RAG activity standardization measurements (**Fig.3 A**, [21, 75]). The assessment of RAG efficiency of recombination was very laborious and time-consuming since it required cell transfection, cell lysis, fragment digestion and purification for further transformation into bacteria. The final readout of the number of recombined molecules was provided either through the ratio of different antibiotic-resistant colonies or through PCR profiles. Moreover, the *in vitro* assays involved high amounts of biological material and intricate multi-step procedures[21]. This caveat has been later tackled through the use of integrated recombination substrates[76] (built onto a retroviral backbone, **Fig.3 B, C**). Retroviral strategies are critical as they enable the targeting of poorly transfectable cells and the delivery of a stable integrated, more physiologic

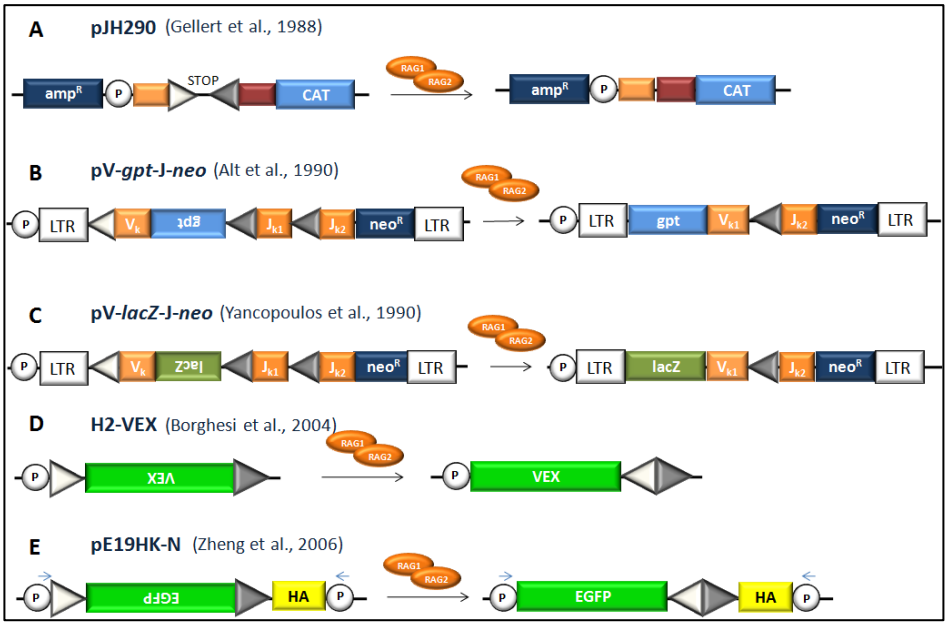


Figure 3. Representation of a number of existing RAG activity reporters. (not to scale) amp^R – ampicillin resistance gene; P-promoter; STOP – STOP codon; CAT- chloramphenicol acetyl transferase; LTR – long terminal repeat; Vk, Jk – coding segments of the Igk locus; neo^R – neomycin resistance gene; gpt – guanine-xantine phosphoribosyl transferase gene; lacZ – β-galactosidase gene; VEX – GFP-variant gene; EGFP – GFP-variant gene; open and closed triangles represent 12- and 23-RSSs, respectively.

reporter substrate. Although the *in vitro* assay's readout system has been improved over the years, through the use of reporters bearing genes that codify for fluorescent proteins detectable by flow cytometry (**Fig.3 D, E**,[77]), the detection still requires secondary steps[76, 77], such as enzymatic reactions or antibody stainings.

Through the use of these valuable molecular tools, quantification of RAG activity has been the vehicle for studying the influence of RSS sequence composition on RAG interaction not only with the V(D)J loci, dictating the diversity of the AR repertoire[78], but also with the whole genome, where cRSSs exist and appear to participate in tumourigenesis[19]. The RSS degeneracy allows for a not so nucleotide-restricted action of RAG: even some V(D)J loci RSSs exhibit variations in the most conserved nucleotide positions (for example in the nonamer or heptamer). Moreover, the synergistic effects of different nucleotide positions spread throughout the RSS are not completely understood, which may also be an influencing factor on RAG's activity.

The first approach to RSS analysis was made by using a linear nucleotide comparison, i.e., the variation of single nucleotides at each position to find their individual importance. Simultaneously, the concept of consensus RSS (herein referred as "Cons") appeared as the sequence that better depicts the average physiological efficiency of recombination of the RSSs found in the AR loci[59]. Recently, Kelsoe's laboratory classified RSSs according to an algorithm that accounts for the RSS' nucleotide relative positional information and estimates its probability of occurrence in the population of mouse RSSs. By examining the nucleotide relationships within a given DNA sequence, the algorithm produces a RSS information content (the RIC score[79]) which is the sequence's theoretical recombination potential. The criterion of functionality for 12-RSS was established to be ≥-40, whereas for 23-RSS it was ≥-60. This *in silico* method defined a different consensus RSS (herein referred as "ConK") as the most conserved 23-spacer (which does not exist in nature), as well as an ideal 23-spacer ("spaMI") defined as having the most frequent nucleotide correlations and differing at five nucleotide positions from ConK (one single and two doublet mutations, named as "(MI)G4", "(MI)G14/15" and "(MI)C19/20", respectively)[78]. spaMI was later found to exist in V(D)J loci mouse RSSs (near gene and pseudogene V_H segments at the IgH locus). spaMI was not associated with consensus heptamer and nonamer but with a low-efficiency non-consensus nonamer, composed of a 5'CAG motif ("CAGnon"). The authors progressed to dissect the contribution of different 23-RSS spacers to the recombination reaction[78], by determining the *in vitro* efficiency of recombination associated with this set of spacers and respective point-mutated sequences in the context of a consensus heptamer and nonamer (**Table I, App.**), using a slightly modified version of the classical pJH290 excision reporter system (**Fig.3 A**, [75]), the p290T, and a modified *in vitro*

recombination assay (IVRA), by electroporation of murine pre-B cells which bear endogenous RAG activity.

1.3.3.1. The GFPi reporter

Our laboratory has generated a RAG activity reporter based on the murine stem cell-retroviral backbone, the MSCV vector[80]. The MSCV-mRFP version (L. M. Sarmento, unpublished) was generated by replacement of the Green Fluorescence Protein (*GFP*) in MSCV-GFP[80] by the monomeric Red Fluorescence Protein (*mRFP1*) gene[81]. The MSCV contains two long terminal repeats – LTRs – placed at both ends; the 5'LTR contains the viral promoter and is followed by a restriction multiple cloning site (MCS). Next, an internal ribosomal entry site (IRES) followed by the *mRFP* gene enables the direct detection of transfected or retroviral-infected cells, as mRFP positive, allowing for the expression of the bicistronic transcript encompassing the coding sequence cloned in the MCS, along with the mRFP marker (**Fig.4 A**).

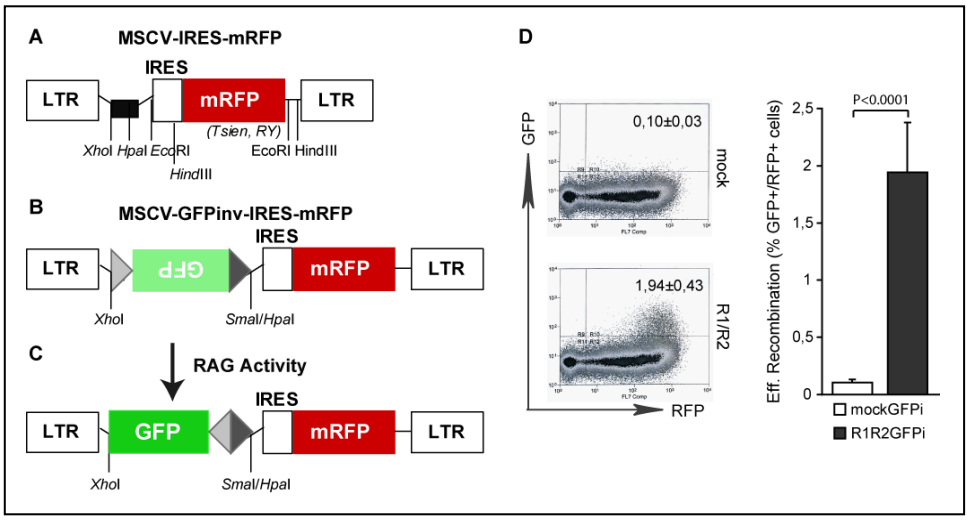


Figure 4. The GFPi reporter. A. Linear representation of the MSCV-IRES-mRFP retroviral region (not to scale); **B.** MSCV-GFPi-IRES-mRFP structure (not to scale); **C.** The resulting GFPi structure in the presence of RAG activity **D.** GFPi *in vitro* recombination assay was performed in the presence (R1/R2) or absence (mock) of RAG activity: Left, GFP/RFP flow cytometry dot plot analysis of 293T cells (representative experiment); right, bar graph with the average ER (n=8, t-test P value < 0.0001).

The RAG reporter structure was cloned into the MCS of the MSCV-mRFP (**Fig.4 B**). It comprises an inverted *GFP* sequence, flanked by two RSSs, a 12- and a 23-spacer RSS, positioned, respectively, at the 5' and 3' ends of the inverted *GFP*. The RSSs are both oriented 5' to 3' (heptamer to nonamer orientation) and their nucleotide composition was defined according to the consensus sequences (Cons) used in the classical pJH290 reporter[21]. Therefore, this particular plasmid was named "GFPi-Cons". The structural relation between the RSSs and their orientation determines that their recognition will lead to an inversion instead of an excision reaction[18]. The delivery of this reporter into cells will first render them mRFP positive. The presence of RAG activity will result in the inversion of the

GFP coding sequence and the acquisition of the GFP fluorescence (**Fig.4 C**). This reporter construct can be used as an extrachromosomal substrate, in an *in vitro* recombination assay (IVRA), by direct transfection, or used to produce retroviral particles to infect primary cells and measure RAG activity *in vivo*.

Preliminary experiments have confirmed the functionality of the GFPi reporter. The GFPi reporter was used to measure RAG activity in an IVRA adapted from Hesse et al[21]. As shown in **Fig.4 D**, in the absence of RAG, expression of GFP was not detected in the mRFP positive population, whereas in the presence of RAG1 and RAG2, an average of 1,94% (and a standard deviation of 0,43) mRFP positive cells were GFP positive. The efficiency of RAG activity was defined as "efficiency of recombination" or "ER", corresponding to the frequency of GFP positive cells detected in the mRFP positive population.

2. AIMS OF THE PROJECT

In spite of the breakthrough of applying molecular reporters to study V(D)J recombination and RAG-mediated genomic instability, so far, no reporter has gathered the retroviral reporter structure with a direct and simple method of readout. In this work, we aimed at establishing a novel retroviral-fluorescent reporter of RAG activity by optimizing the detection of RAG activity (**4.1.**) and proving the reporter's sensitivity to RSS sequence degeneration (**4.2.**). With this new tool, we aimed at assessing the functionality of a set of cRSSs known to be involved in leukaemogenesis (**4.3., 4.4.**) and characterize, qualitatively and quantitatively, the presence of endogenous RAG activity in human leukaemic tumour cell lines (**4.5.**).

3. MATERIALS AND METHODS

3.1. Cell culture

HEK (human embryonic kidney fibroblasts) 293T cells (**4.1.**, **4.2., 4.3.** and **4.4.**) and 3T3 fibroblasts (**4.5.**) were cultured in DMEM[82] media (GIBCO) supplemented with Fetal Calf Serum 10%, Sodium Pyruvate 1%, L-Glutamine 1%, PenStrep 1% (Invitrogen), at a temperature of 37ºC and 5% CO_2 conditions. The cells were harvested, washed and re-seeded at low density every two to three days.

HL-60 ,SUPT1, Nalm6, Jurkat, Reh and K562 tumour cell lines (**4.5.**) were cultured in RPMI Glutamax media (GIBCO) supplemented with Fetal Calf Serum 10%, Sodium Pyruvate 1%, PenStrep 1% , at a temperature of 37ºC and 5% CO_2 conditions. The cells were maintained at a concentration of 0,25 x 10^6 cells/mL. As so, cells were washed and re-seeded every two to three days and purged from dead cells by Ficoll (Pharmacia) gradient when necessary.

3.2. *In Vitro* Recombination Assay (IVRA)

The IVRA protocol was adapted from Hesse et al[21]. Briefly, in **4.1.**, **4.2.**, **4.3.** and **4.4.**, 0,5 x 10^6 HEK 293T cells were seeded in a 6-well plate well in 2,5 mL of media. After 24 hours, cells were co-transfected by replacing 600 μL of media by a transfection mix containing 10 μL of Lipofectamine2000 (Invitrogen), 600μL of Optimem and a total of 10μg of pDNA: with 10μg of mock plasmid DNA or with 5μg of GFPi (or a GFPi-derived plasmid) and 5 μg of mock plasmid (negative control) or 2,5μg of H2K-RAG1 plus 2,5μg of H2K-RAG2 (H2k assay) or 1,6μg of pCMV-RAG1 plus 1,4μg of pCMV-RAG2 plus 2μg of mock plasmid (equimolar CMV assay). After 16 hours of incubation (over-night) in transfection mix-containing media, cultures were washed with fresh media. The cells were further cultured for 48 hours, harvested, and analysed by flow cytometry for GFP and mRFP fluorescence detection at the single cell level. Variations to the protocol are indicated in figure legends or results.

3.3. Statistical analyses

In **4.1.**, **4.2. 4.3.** and **4.4.**, to every type of IVRA, an average and a standard deviation of ER were determined for each independent experiment, by pooling the replicates. When pooling experiments, the background average value was subtracted [corresponding to the (-) GFPi negative control] and the standard deviation was found by the formula $SD_{final} = \sqrt{(standard\ deviation\ x)^2 + (standard\ deviation\ y)^2 + \cdots)}$. When comparing groups of data, a non-parametric t-test was applied using the GraphPrism software. Whenever necessary, a Welch correction added.

3.4. Flow cytometry analyses

In **4.1.**, **4.2.**, **4.3.**, **4.4.** and **4.5.**, flow cytometry data were acquired in MoFlo (RFP was measured on channel 7 (FL7) with a Krypton Crystal Laser CL-2000 Dioded 561nm Yellow Laser and a filter pass D630/75; GFP was measured on channel 1 (FL1) with an Argon Sapphire-200mV 488nm Blue Laser, with D530/40 filter pass). Data was further analysed using the Flowjo software. In **4.5.**, flow cytometry data were also aquired in FACS Calibur in order to calculate viral titers, respectively (GFP was measured on channel 1 (FL1) with a 15mW SPECTRA-PHYSICS Aircooled 488nm laser). Data was further analysed using the CellQuest software.

3.5. Immunoblot analysis

In **4.1.**, HEK 293T cells were transfected with equimolar amounts of H2K-RAG1 or CMV-RAG1 or H2K-RAG2 or CMV-RAG2, cultured for 36 hours, after which the cells were lysed and protein extracts run in a SDS-PAGE. Separated proteins were transferred onto a nitrocellulose membrane, and probed with antibodies anti-RAG1 (rabbit polyclonal diluted

1:100 or 1:500, K-20 SantaCruz) or anti-RAG2 (rabbit polyclonal diluted 1:50 or 1:500, M-300 SantaCruz) antibodies followed by the secondary anti-rabbit-HRP antibody (Thermo Scientific) probing. For the internal controls, the primary antibodies used were anti-α-tubulin (mouse monoclonal, diluted 1:500 or 1:1000, SIGMA) and anti-β-actin (goat polyclonal, diluted 1:1000 or 1:2000, C-19 SantaCruz) and the secondary were anti-mouse-HRP (Thermo Scientific) and anti-goat-HRP (Pierce), respectively. Immunoblot washes and incubations were performed in Tris-Buffer-Saline supplemented with 0.1% Tween-20. Detection was performed using the Thermo Scientific Pierce Fast Western Blot Kit and ECL Substrate chemioluminescent kit (Thermo Fisher). Prior to re-probing, membranes were subjected to a stripping solution of 0,0625M Tris-HCl pH6.8, 2% SDS, 1/100 (v/v) β-mercaptoethanol and incubated at 56ºC for one hour.

3.6. Molecular cloning

3.6.1. CMV-RAG Cloning

In **4.1.** CMV-RAG1 was obtained by excising a 4.7Kb fragment from a Bluescript pKS-RAG1 plasmid (L. Sarmento), comprising the RAG2 5' mini-exon and the RAG1 coding sequence, through the use of *Xho*I/*Not*I restriction enzymes (NEB). This fragment was further ligated to a previously processed 3.6Kb pCMVβ vector (Invitrogen) cut by *Not*I and *Xho*I (NEB). Likewise, CMV-RAG2 was obtained by excising a 2.4Kb fragment containing RAG2 5' mini-exon and RAG2 coding sequence with *Sal*I/*Ase*I (NEB) restriction enzymes. As the *Ase*I site disrupted the TAA stop codon contained in the 3' region of RAG2 coding sequence, the *Ase*I end was filled-in with Klenow (NEB) prior to *Sal*I digestion. Likewise, the 3.6Kb recipient vector pCMVβ was also cut by *Not*I followed by fill-in with Klenow and further digestion with *Xho*I resulting in a *Xho*I/*Not*I(Klenow)-ended fragment allowing for TAG stop codon reconstitution upon ligation with the RAG2 fragment.

CMV-RAG1 was obtained by excising a 4.7Kb fragment from an intermediate Bluescript pKS plasmid, used in the above H2k-RAG1 cloning, comprising RAG1 5' mini-exon and RAG1 coding sequence, through the use of *Xho*I/*Not*I restriction enzymes (NEB). This fragment was further ligated to a previously processed 3.6Kb pCMVβ vector (Invitrogen) cut by *Ecor*V (NEB) followed by *Not*I and *Xho*I, generating CMV-RAG1. Likewise, CMV-RAG2 was obtained by excising a 2.4Kb fragment containing RAG2 5' mini-exon and RAG2 coding sequence with *Sal*I/*Ase*I (NEB) restriction enzymes. As the *Ase*I site disrupted the TAA stop codon contained in the 3' region of RAG2 coding sequence, this fragment was further processed by Klenow (NEB), as well as the 3.6Kb recipient vector pCMVβ, which was also cut by *Not*I, processed by Klenow and further cut by *Xho*I resulting in a *Xho*I/*Not*I(Klenow)-ended fragment allowing for TAG stop codon reconstitution and generating CMV-RAG2.

3.6.2. GFPi-Cons Cloning

In **4.1.**, The GFPi-Cons reporter was generated by PCR with the use of specific primers, a plasmid containing the GFP gene as a template (MSCV-GFP) and *Pfu* polymerase (Promega). The core of this strategy relied on the PCR primers' design. As an example, an expected sequence of the GFPi-Cons PCR product is depicted in **Fig.III** and the respective primers used in **Table I,** both in **Appendix I (App.)**.

Left to right (5' to 3'), this sequence displays a *Xho*I restriction site, a 5' to 3' consensus 12-RSS, the reverse complementary sequence of a Kozak-*GFP* fusion, a 5' to 3' consensus 23-RSS and a *Sma*I site. The regions used for primers' design are underlined and at least 20 nucleotides for *GFP* hybridization were included.

The fragment was further submitted to a *Taq* polymerase (Invitrogen) poly(A)-tail extension in order to allow for their ligation to the intermediate vector pGEM-Teasy (Promega). This vector was further amplified in *Escherichia coli*, and the selected clones were used to extract the final *Xho*I/*Sma*I fragments to be ligated to the MSCV-mRFP vector previously digested with *Xho*I and *Hpa*I, generating GFPi-Cons (Sarmento, unpublished). The final vector was sequenced for fragment confirmation at the IGC Sequencing Facility.

3.6.3. GFPi-23-RSS reporters Cloning

In **4.2.**, GFPi-ConK, GFPi-spaMI, GFPi-(MI)G4, GFPi-(MI)G14/15, GFPi-(MI)C19/20 and GFPi-spaMI-CAGnon fragments were cloned using the exact same methodology as in **3.6.2.**, with the exception of the reverse primer oligonucleotide used for each 23-RSS reporter cloning. The 12-RSS forward primer was used in all PCR reactions and its sequence is an exact match of that underlined in **Fig. III (App.)**. **Table I (App.)** lists the 23-RSS sequences contained on each GFPi-reporter and the respectively designed reverse primers.

3.6.4. GFPi-cRSS reporters Cloning

In **4.3.** and **4.4.,** GFPi-V_H, GFPi-Jβ, GFPi-Lmo2, GFPi-SCL and GFPi-PTEN reporter versions were originated using the exact same methodology as in **3.6.2.**, with the exception of the reverse primer oligonucleotide used for each 12-RSS reporter cloning. **Table II (App.)** lists the 12-RSS and cRSS sequences used in this work, as well as their respective primer sequences.

3.7. Viral production and titers

In **4.5.**, the viral production was adapted by Naldini et al. Briefly, 2 x 10^6 HEK 293T cells were seeded on a pre-coated Poly-L-Lysine (SIGMA) 20cm^2 6cm-diameter dish in 5 mL of media and incubated for 24 hours. For viral production, cells were co-transfected for 16hours in 4 mL of media using the calcium phosphate method with 5μg of pKAT retroviral packaging vector, 5μg of pCMV-VSV-G envelope glycoprotein encoding plasmid and 10μg of

MSCV-GFPinv in a 1 mL transfection mix containing 0.165M $CaCl_2$ in Hydrogen-Phosphate-Buffer-Saline (HBS with 0.75mM Na_2HPO4, 140mM NaCl, 5mM KCl and 6mM Dextrose). Transfections were concluded by replacing the transfection media by 3mL of fresh media supplemented with HEPES 20mM for viral harvesting. Virus-containing media were collected, filtered with 0,45μ filters at 24, 48 and 72h post-transfection conclusion.

Viral titers were estimated by using virus volumes (**V.Vol**) of 100μL or 200μL of 24, 48 or 72h stocks to infect $1x10^5$ 3T3 fibroblasts (pre-seeded 24 hours prior at $0,5x10^5$/condition) in the presence of 8μg/mL of Polybrene (Hexadimethrine bromide, SIGMA) by spin-infection at 2500 rpm, 30 min., RT. Frequencies of infection (**FI**) were determined by measuring the cells positive for the IRES-coupled fluorescence by flow cytometry 4 to 5 days after infection. The viral titer (**VT**) was calculated with FIs equal or lower than 0,5 since the n. of infectious particles (**IP**) is directly proportional to the number of cells (**NC**) when Multiplicity Of Infection (MOI) is equal or lower than 1 and considering that a retroviral infection carried out with a MOI of 1, i.e n. infectious particles (**IP**) = n. cells (**NC**), gives rise to a FI of 0,5. Therefore **VT** (IP /mL) = [FI x NC x (1/V.Vol)] /0,5 . The viral titers obtained were $1-2x10^6$ IP /mL.

3.8. Cell line infections

In **4.5.**, 1 x 10^5 cells of HL-60, K-562, SUPT1, Jurkat, NALM-6 and Reh tumour cell lines were infected on a previously retronectin-coated (5μg/cm^2, TaKara) 2cm^2 well-plate. Viral suspensions of MSCV-mRFP or MSCV-GFPi were added at a MOI 5:1 and the cells were centrifuged for 2,5h at 2500rmp RT. Polybrene (Hexadimethrine bromide, SIGMA) was added in a concentration of 4μg/mL in order to facilitate viral adhesion. 24 hours later, the cells were submitted to a second round of infection in the exact same conditions.

4. RESULTS

4.1. Optimization of the GFPi IVRA

The efficiency of RAG activity detected in the preliminary GFPi *in vitro* experiment (ER = 1,94% ± 0,43) is lower than that described when using other reporters bearing the same consensus RSSs[63]. This could be either due to the presence of inhibitory nucleotide sequences flanking the reporter cassette (in the MSCV-mRFP vector) and/or to suboptimal levels of RAG expression when using the Major-Histocompatibility-Class I (MHC-I) promoter, named H2K, present in the RAG expression plasmids used (H2k-RAG). To address this issue, we subcloned RAG1 and RAG2 coding sequences under the control of the CMV promoter to test the ER of GFPi-Cons by IVRA in a context of RAG overexpression. Firstly, H2k and CMV-driven RAG protein levels were compared by immunoblot analysis of 293T cells transfected with equimolar amounts of H2k-RAG1 or CMV-RAG1 (**Fig.IV A** and **B,**

App.), or H2k-RAG2 or CMV-RAG2 (**Fig.IV C** and **D, App.**). Since we have found a remarkable difference between H2k-RAG and CMV-RAG protein levels, we optimized the detection in order to better determine such different levels. CMV-RAG1 and RAG2 were successfully detected when titrating cell extracts to half, 1/5 and 1/10 of their original concentrations and using standard antibody concentrations (**Fig.IV A** and **C, App.**). Conversely, H2k-RAG1 and RAG2 were detected by doubling the amount of loaded cell extracts and using a higher antibody concentration, allowing for a comparison between H2K and CMV-driven expression (**Fig.IV B** and **D, App.**). Overall, this analysis indicates that CMV-driven expression of RAG gives rise to at least 10 to 20-fold more protein than the H2k promoter.

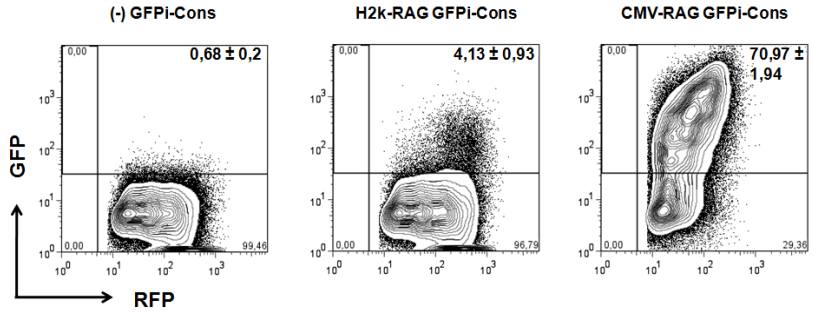


Figure 5. Representative flow cytometry analysis of control (-) GFPi-Cons (left), H2k-RAG GFPi IVRA (middle) and CMV-RAG GFPi IVRA (right). Events were gated on RFP^+ (x axis) and GFP^+ (y axis); the frequency of double-positive events (ER) average and standard deviation is highlighted in bold; n=3 replicates of one independent experiment.

Subsequently, CMV-RAG1 and CMV-RAG2 were used in an IVRA, along with the GFPi-Cons reporter, in equimolar amounts to H2k-RAG (**Fig.5**). CMV-RAG exhibited a 20-fold higher ER than H2k-RAG (60,85 ± 1,75 vs. 3,01% ± 1,01; n=9). This efficiency greatly exceeds the average of 4-6% described in the literature when using pJH290[63, 78], which measures the number of recombined substrate molecules. Detailed analysis of recombination in the RFP^+ population, bearing the maximum number of reporters per cell and which display maximum efficiency of recombination (MIF_{RFP}=150), shows that CMV-RAG expression, when compared to H2k-RAG, not only generates 10-fold higher frequency of GFP^+ cells (95,92 ± 0,39 vs 8,49 ± 2,22) but also 10-fold higher number of recombined substrates in average per cell, represented by the value of the mean intensity of GFP fluorescence (MIF_{GFP}, 1145,12 ± 120,52 vs 88,75 ± 9,89). This data shows that the new GFPi reporter is a competent substrate for RAG activity and that the designed IVRA, either using H2k-RAG or CMV-RAG, provides a broader window for quantitation of RAG activity.

We next sought to understand whether various baseline levels of recombination could be obtained when using the GFPi-Cons reporter along with the CMV-RAG expression system. We aimed at defining the assay that better allowed observing increasing or decreasing differences between different GFPi constructs. To that aim, we titrated down the amounts of CMV-RAG plasmids used along with GFPi-Cons in the cell transfection and proceed to determine RAG activity. Progressive 2-fold decreases in CMV-RAG plasmid

amounts, from 1,4-1,6 μg to 0,175-0,2 μg, did not correlate with significant differences in ER or with differences proportional to the plasmid reduction (**Fig.6, Fig.7**).

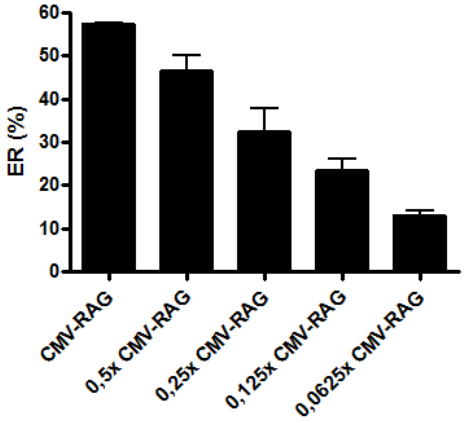


Figure 6. Efficiency of recombination values for GFPi-Cons upon CMV-RAG plasmid titration in IVRAs. CMV-RAG1 and CMV-RAG2 expression plasmids were titrated down by half from 1,4μg and 1,6 μg (CMV-RAG) to 0,175 μg and 0,2 μg (0,0625x CMV-RAG), respectively; one to two independent experiments, each with n=3 replicates; ER = Efficiency of Recombination.

This observation suggested that CMV-RAG plasmid amounts express levels of protein which approach functional saturation. This conclusion is also supported by the ER values obtained for high amounts of substrate which reach 91-97% for a RFP Mean Intensity of Fluorescence (MIF_{RFP}) around 150 and 400 (data not shown).

Therefore, we propose that the CMV-RAG GFPi assay cannot be used to compare RAG activity at least of consensus-like RSSs since non-saturating conditions can only be reached with very small amounts of CMV-RAG DNA which may be limiting in the transfection reaction. Overall, we have successfully optimized the GFPi assay creating a broader window of RAG activity measurement with the use of H2k-RAG and CMV-RAG, which may be useful to study cryptic RSSs.

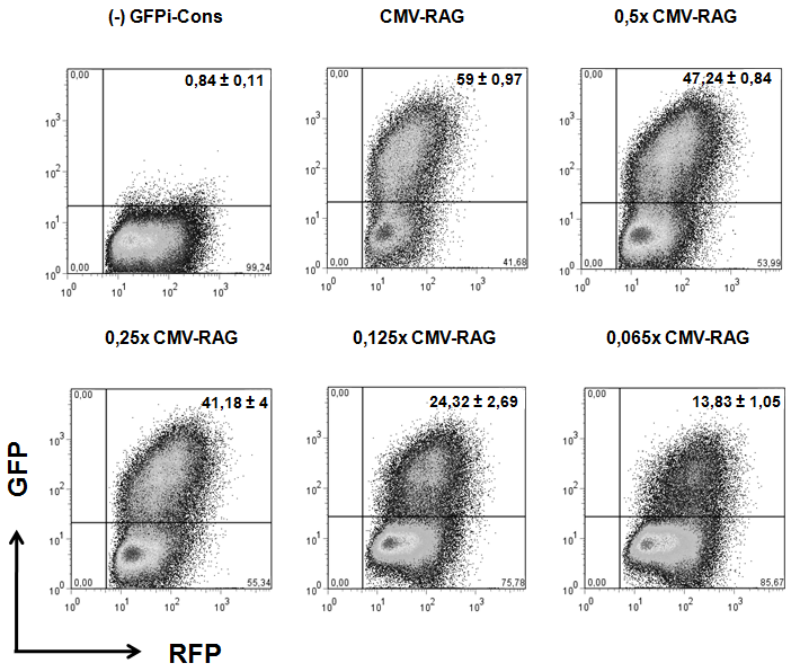


Figure 7. Representative flow cytometry analysis of successive CMV-RAG GFPi IVRAs with decreasing titrated levels of CMV-RAG plasmid amounts. GFPi-Cons IVRA in absence of CMV-RAG was used as a negative control (upper left); CMV-RAG1 and CMV-RAG2 expression plasmids were titrated down by half from 1,4μg and 1,6 μg (CMV-RAG) to 0,175 μg and 0,2 μg (0,0625x CMV-RAG), respectively; events were gated on RFP^+ (x axis) and GFP^+ (y axis); the frequency of double-positive events (ER) average and standard deviation is highlighted in bold; n=3 replicates of one independent experiment.

4.2. GFPi sensitivity to 23-RSS sequence degeneration

We aimed at determining whether the GFPi reporter was able to provide a fine quantitation of RAG activity *in vitro*, paralleling the pJH290 system. As so, Kelsoe and

colleagues' RSS analysis was here reproduced by using the GFPi reporter. Six variations to the GFPi-Cons reporter were generated (GFPi-ConK, GFPi-spaMI, GFPi-(MI)G4, GFPi-(MI)G14/15, GFPi-(MI)C19/20 and GFPi-spaMI-CAGnon), bearing the same 12-RSS but differing in the 23-RSS (as described in [78]), and were further used in IVRAs for ER determination.

Despite the high values of ER obtained with CMV-RAG in this work, we decided to use H2k-RAG to test the 23-RSSs due to the fact that some RSS sequences which were predicted to be highly efficient[78] (eg. spaMI) would not be properly assessed in the saturating conditions observed previously with GFPi-Cons.

As a first approach, the 23-RSSs were analysed with the RIC algorithm developed by Kelsoe et al.[79]; all tested 23-RSS sequences were predicted to be functional, since their score was ≥-60 (**Table III, App.**).

When comparing the literature's 23-RSS analysis in the p290T setup (**Fig.8 A**,[78]) to the H2k-RAG GFPi setup (**Fig.8 B, Fig.9**), we find a narrower window of ER values in the case of H2k-RAG GFPi, which ranged from an ER of 0,32% to 5,89%, contrasting to p290T which ranged from 0% to 33,3%. Nevertheless, we also observed a remarkable difference between standard deviations of both assays: H2k-RAG GFPi exhibited ratios average/standard deviation ranging from 15% to 37% while p290T ranged from 18% to 100%. Since the data points of the literature's experiment were not available, it was not

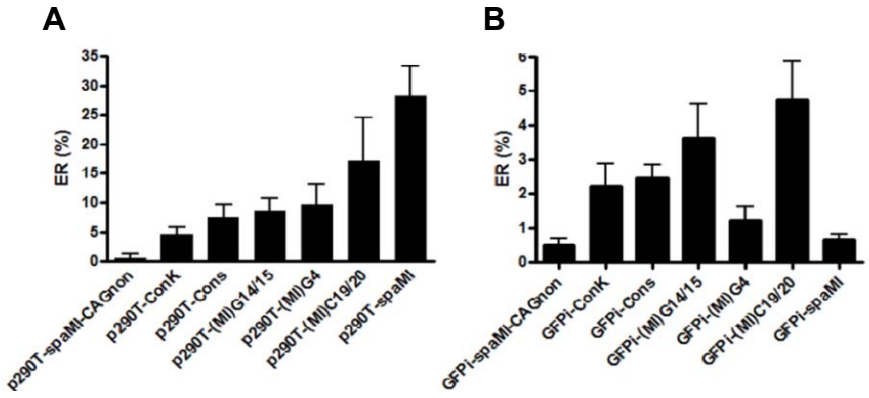


Figure 8. p290T and H2k-RAG GFPi 23-RSS reporters' ERs. A. Graphical representation of Kelsoe and colleagues p290T-23-RSS reporters IVRA results[78] **B.** GFPi-23-RSS reporters IVRAs using H2k-RAG plasmids; two to three pooled independent experiments each with n=3 replicates; ER = Efficiency of Recombination.

possible to compare the p290T and GFPi IVRA RAG activity measurements by using powerful statistical testing. Thus, we took a simple though suitable statistical approach: the fold difference coefficient calculation between ER values of 23-RSS pT290 and GFPi constructs. Fold-difference values were calculated for every set of two constructs, by dividing the respective ER values, either in the p290T (**Table IV, App.**) or the H2k-RAG GFPi assay (**Table V, App.**). These were then compared between the two assays, by calculating the respective coefficient value for every pair of constructs (**Table VI, App.**). If both assays were

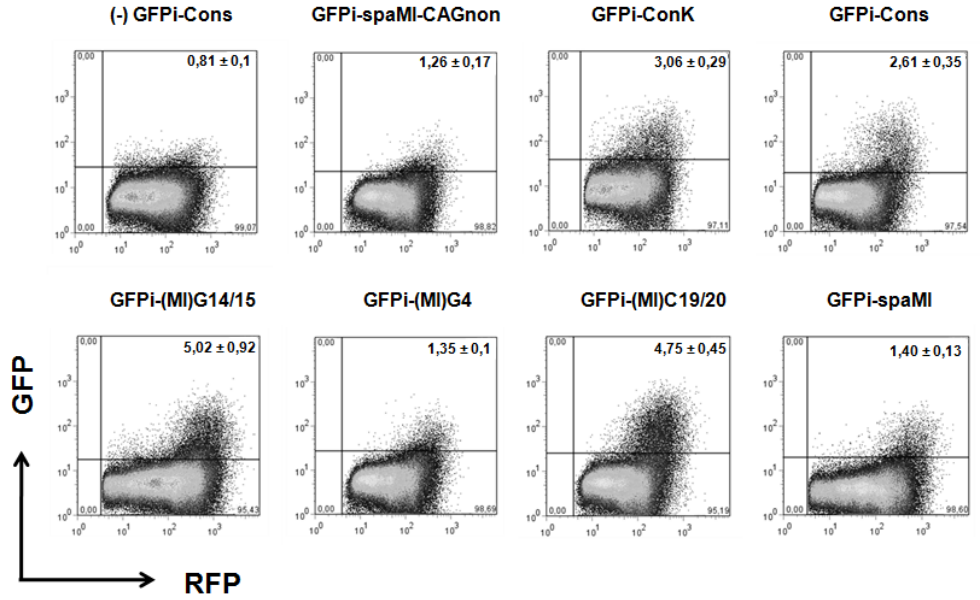


Figure 9. Representative flow cytometry analysis plots of negative control GFPi-Cons IVRA (upper left) and H2k-RAG 23-RSS GFPi IVRAs. GFPi-Cons IVRA in absence of CMV-RAG was used as a negative control (upper left); Events were gated on RFP^+ (x axis) and GFP^+ (y axis); the frequency of double-positive events (ER) average and standard deviation is highlighted in bold; n=3 replicates of one independent experiment.

similar, **Table VI. (App.)** should present the same value in every cell. The coefficient values obtained were observed to vary in a narrow interval, from a coefficient of 0,8 to 3,1. The bigger differences were found in (MI)G4 and spaMI coefficients. Discarding these, the analysis is consistent to a general similarity between the p290T and the GFPi assays.

From these results, we conclude that H2k-RAG GFPi assay holds a power of discrimination which lies mainly on lower ER values. GFPi was able to partially reproduce the 23-RSS ranking, correctly discriminating the ER of four constructs out of seven. These data show that the GFPi reporter is at least partially sensitive to RSS sequence variations allowing for the detection of absolute values of efficiency of recombination in a range at least equal to that exhibited by the analysis with p290T, therefore providing a tool to detect differences in RAG activity.

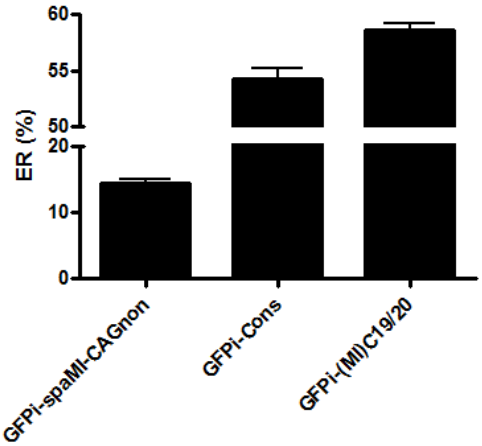


Figure 10. CMV-RAG 23-RSS GFPi reporters' ERs. GFPi-spaMI-CAGnon, GFPi-Cons and GFPi-(MI)C19/20 23-RSS reporters IVRAs using CMV-RAG plasmids; one to two pooled independent experiments each with n=3 replicates; ER = Efficiency of Recombination.

The CMV-RAG GFPi system appeared to be ideal to study RSSs with low efficiency of recombination, particularly cryptic RSSs. Therefore, we decided to better analyze the 23-RSS GFPi constructs which had provided the most extreme ER values in the H2k-RAG GFPi assay (GFPi-(MI)C19/20 and the low-

efficiency GFPi-spaMI-CAGnon reporters) by using the CMV-RAG GFPi system and compare them to the consensus construct (GFPi-Cons) used initially in this IVRA setup.

Herein, the window of ER values was much broader, as expected (**Table VII** in **App., Fig.10, Fig.11**). The value observed for GFPi-Cons remained as the highest value (ER = 54,2 ± 1), since GFPi-(MI)C19/20 did not exhibit higher value in this assay (ER = 58,58 ± 0,63), contrarily to what was observed with H2k-RAG. This consolidates the previous observation that in these conditions the assay is not discriminatory for RSSs with a high efficiency. Interestingly, the ER of GFPi-spaMI-CAGnon (ER = 14,49% ± 0,52) increased 28-fold when compared to the H2k-RAG ER value (ER = 0,51 ± 0,19) and the approximate 4-fold increase in ER from GFPi-spaMI-CAGnon to GFPi-Cons observed in the H2k assay was maintained in the CMV analysis. These data indicate that the CMV-RAG assay is most suitable to study cRSSs with predicted residual RAG activity or previously undetected ER.

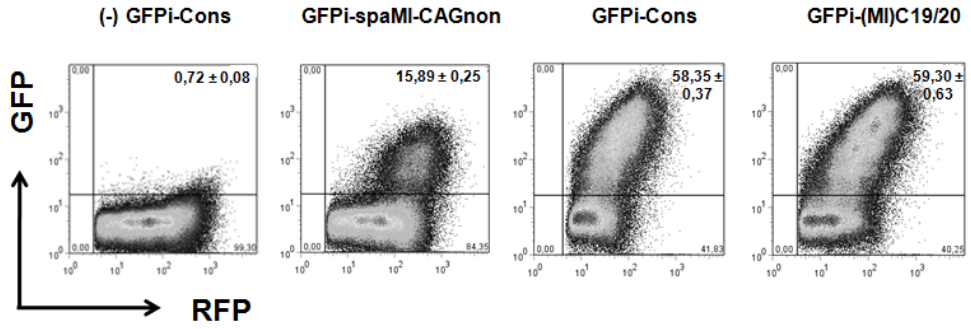


Figure 11. Representative flow cytometry analysis plots of CMV-RAG 23-RSS GFPi IVRAs. GFPi-Cons IVRA in absence of CMV-RAG was used as a negative control (left); events were gated on RFP^+ (x axis) and GFP^+ (y axis); the frequency of double-positive events (ER) average and standard deviation is highlighted in bold; n=3 replicates of one independent experiment.

With these results, we concluded that the GFPi assay bears the ability to measure RAG activity on a broad range of ER levels, since H2k-RAG is suitable for highly efficient RSS ER measurement whereas CMV-RAG is able to detect low efficiencies of recombination, therefore providing a versatile tool for further application.

4.3. Assessment of RAG-mediated cRSS functionality *in vitro*

The question underlying the relationship between RAG and its effect on oncogenesis by cRSS targeting has been addressed for long (referred in **1.3.2.**). So far, the molecular studies undertaken have required laborious and time-demanding work. At this stage, due to the advantages of the GFPi reporter, we decided to validate this tool for RAG activity detection in cRSSs.

For the validation of the CMV-RAG cRSS-GFPi assay, we have selected a set of described 12-cRSSs and two V(D)J loci 12-RSSs: Jβ2-2, herein named as Jβ, and V_H1/87(181), or V_H, selected from different AR loci (TCRβ and IgH, respectively). Both were previously shown to exhibit barely detectable functionality levels[63, 83, 84], thus reflecting low usage of these segments by RAG[83, 85]. 12-cRSSs were selected from Lmo2 and SCL genes

(known to be involved in T-ALL onset) the first having been described as functionally efficient[85] and the second considered undetectable[85] (as shown in **Table VIII, App.**).

As a preliminary approach, we have addressed RSS/cRSS functionalities by running the RIC score algorithm for all the referred sequences and found that all sequences exhibited RIC scores lower than the functionality threshold, with the exception of Jβ2-2 which presented a RIC score of -37,8 (**Table VIII, App.**).

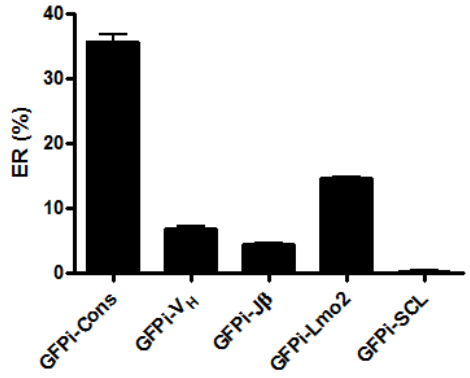


Figure 12. Efficiency of recombination obtained for each GFPi-RSS/cRSS reporter. GFPi-Cons ER value is used as a control; one to two pooled independent experiments each with n=3 replicates; ER = Efficiency of Recombination.

Subsequently, four GFPi-reporters were generated in a similar manner to the GFPi 23-RSS constructs (referred in **4.2.**), all containing the same 23-RSS sequence (Cons) but differing in the 12-RSS: GFPi-V_H, GFPi-Jβ, GFPi-Lmo2 and GFPi-SCL. For each reporter, the ER was determined by IVRA using CMV-RAG expression plasmids in order to successfully detect residual RAG activity (**Table VIII, App.**).

Concerning the comparison between ERs of each GFPi-RSS/cRSS reporter, all reporters presented lower ERs than the reference ER value of GFPi-Cons (**Fig.12, Fig.13**). In these assays, GFPi-Lmo2 was shown to be much more functional (exhibiting an ER of 14,67% ± 0,16) than the two V(D)J loci RSS reporters (ER = 6,77% ± 0,39 for GFPi-V_H and ER = 4,41% ± 0,13 for GFPi-Jβ). GFPi-SCL presented residual RAG activity levels (with a low though detectable ER of 0,28% ± 0,1).

Hence, the CMV-RAG GFPi IVRA assay was successfully established as a cRSS assay, which allowed for the successful detection of low levels of RAG

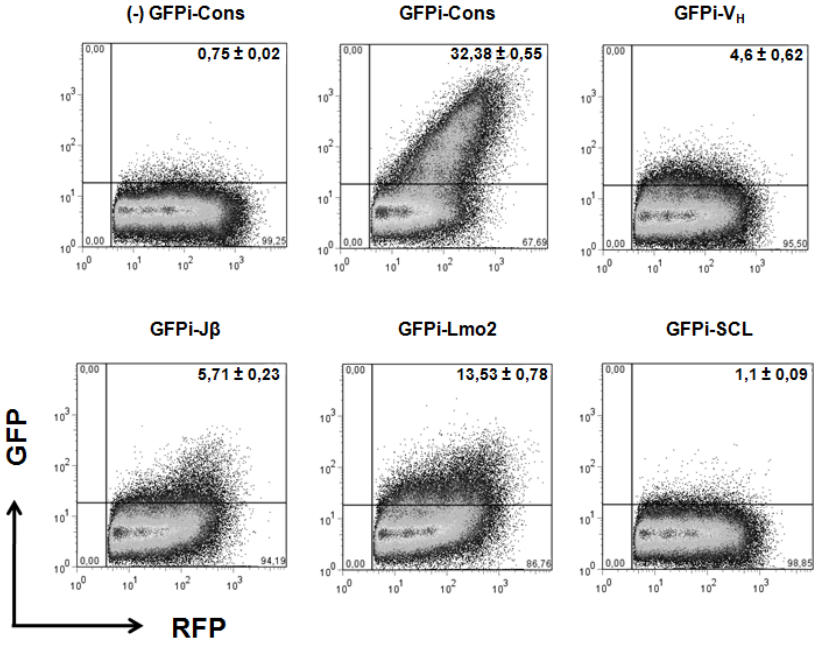


Figure 13. Representative flow cytometry analysis plots of CMV-RAG 12-RSS/cRSS-GFPi IVRAs. GFPi-Cons IVRA in absence of CMV-RAG was used as a negative control (left); events were gated on RFP^+ (x axis) and GFP^+ (y axis); the frequency of double-positive events (ER) average and standard deviation is highlighted in bold; n=3 replicates of one independent experiment.

activity not only in low-efficient RSSs but also in a cRSS with previously undetectable ER levels.

4.4. Assessment of RAG-mediated PTEN cRSS functionality *in vitro*

PTEN (*phosphatase and tensin homologue deleted on chromosome 10*) gene encodes for a phosphatase that is part of the phosphatidylinositol-3 kinase (PI3K) /PTEN/protein kinase B (Akt) pathway which plays an important role on cell metabolism, proliferation, cell cycle progression and survival[86].

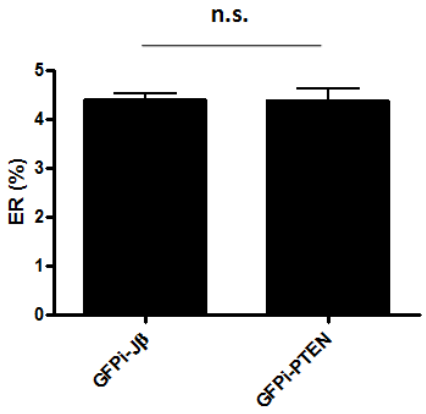
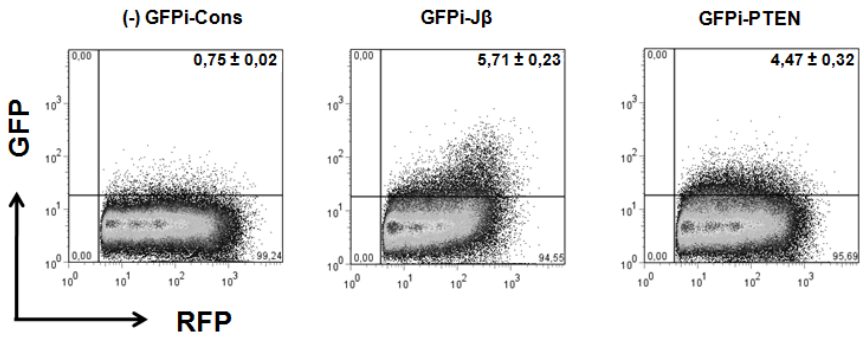


Figure 14. Efficiency of recombination obtained for GFPi-PTEN cRSS reporter. GFPi-Jβ ER value is used as a control; one independent experiment containing n=3 replicates.

Mutations in the PTEN gene have been associated with brain tumours and more recently to T-ALL generation[87]. Recently, this gene was found mutated in a T-ALL leukaemia patient (unpublished data, Barata et al.). Sequencing analysis has revealed a novel mutation located within PTEN's first exon, exhibiting a deletion of a 12-nucleotide fragment (which disrupted the ATG start codon) and an addition of 14 nucleotides in a proximate region. We have used the RIC score algorithm to screen for cRSSs in this particular region of the PTEN gene. A 12-cRSS sequence was predicted to exist adjacently to the referred mutation site, bearing a RIC score of -45,5 (near the -40 threshold of functionality). In an attempt to understand a possible relation between this cRSS and the referred mutation, we assessed its functionality upon RAG activity by generating GFPi-PTEN and testing this reporter in the CMV-RAG cRSS GFPi assay, along with the previously tested low efficiency Jβ cRSS, as a control.



GFPi-PTEN exhibited similar ER values to GFPi-Jβ (confirmed by a t-test comparison, p>0,05), exhibiting an ER of 4,39% ± 0,23 (**Fig.14, Fig.15**).

Figure 15. Representative flow cytometry analysis plots of negative control GFPi-Cons IVRA (left), and CMV-RAG GFPi-Jβ (middle) and GFPi-PTEN (right) cRSS IVRAs. Events were gated on RFP^+ (x axis) and GFP^+ (y axis); the frequency of double-positive events (ER) average and standard deviation is highlighted in bold; n=3 replicates of one independent experiment.

4.5. Endogenous RAG activity detection in human haematopoietic tumour cell lines

Leukaemias, representing a large proportion of haematological malignancies, are described as having disrupted haematopoietic differentiation[88]. Acute leukaemias are characterized by a haematopoietic developmental arrest of immature cells, accompanied by a boost of cell proliferation, whereas chronic leukaemias are originated from more mature cell stages and develop at a much slower rate. Although leukaemias are classified according to their lymphoid or myeloid origin, in some cases, the leukaemic cell type may carry biphenotipic markers[72], which may suggest that both lineages are not so well defined. Considering the promiscuity of RAG expression[58] and activity[19], together with the genetic plasticity of transformed cells in the context of haematopoiesis, we hypothesised a role for RAG in the origin and/or progression of leukaemic genomic instability, either inside or beyond the lymphoid compartment.

As an initial approach, we performed a systematic screen for endogenous RAG activity in lymphoid and myeloid human tumour cell lines, as it had previously been assessed with classical replicative extrachromosomal reporters[89, 90]. However, we aimed at creating a more physiological setup by using the GFPi-Cons retroviral-based reporter as an integrated and low copy number substrate. For that, pseudo-retroviruses of control (MSCV-mRFP) and GFPi-Cons reporters were produced and used to infect a collection of myeloid and lymphoid human tumour cell lines, kindly provided by the Unidade de Biologia do Cancro, Instituto de Medicina Molecular and Instituto Português de Oncologia. HL-60 and K-562 myeloid cell lines were originally collected from a leukaemic patient with APL and from CML (Chronic Myelogenous Leukaemia) in blast crisis (i.e., the terminal phase of CML), respectively. SUP-T1 and Jurkat lymphoid cell lines were originated from T-ALL whereas Reh and NALM-6 were from B-ALL. The assessment of RAG activity was performed in the reporter-infected cells by flow cytometry analysis, in a similar manner to the above tasks of this work.

The optimum conditions for cell analysis were found at two weeks post-infection (**Fig. 16**). The dynamics of RAG activity in a time-dependent manner could not be assessed, since the rate of infected cells was not stable throughout time. Regarding the ER values obtained with the integrative GFPi-Cons reporter (**Table IX, App.**) and the respective flow cytometry profiles (**Fig.16**), all lymphoid cell lines presented RAG activity, whereas HL-60 and K-562 myeloid cell lines seemed devoided of it (the ER of HL-60 was 0,05% and K-562 was 0%). Both B-ALL cell lines displayed a clear double-positive population: NALM-6 had an ER of 1,19% and Reh presented the highest ER of all tested cell lines (ER = 13,61%). Concerning T-ALL cell lines, both presented a very scattered double-positive population, different from the profile found in B-ALL. Jurkat's ER value of 1,11% was quite similar to the one of NALM-6 and SUP-T1 presented very low values of endogenous RAG activity (ER = 0,24%).

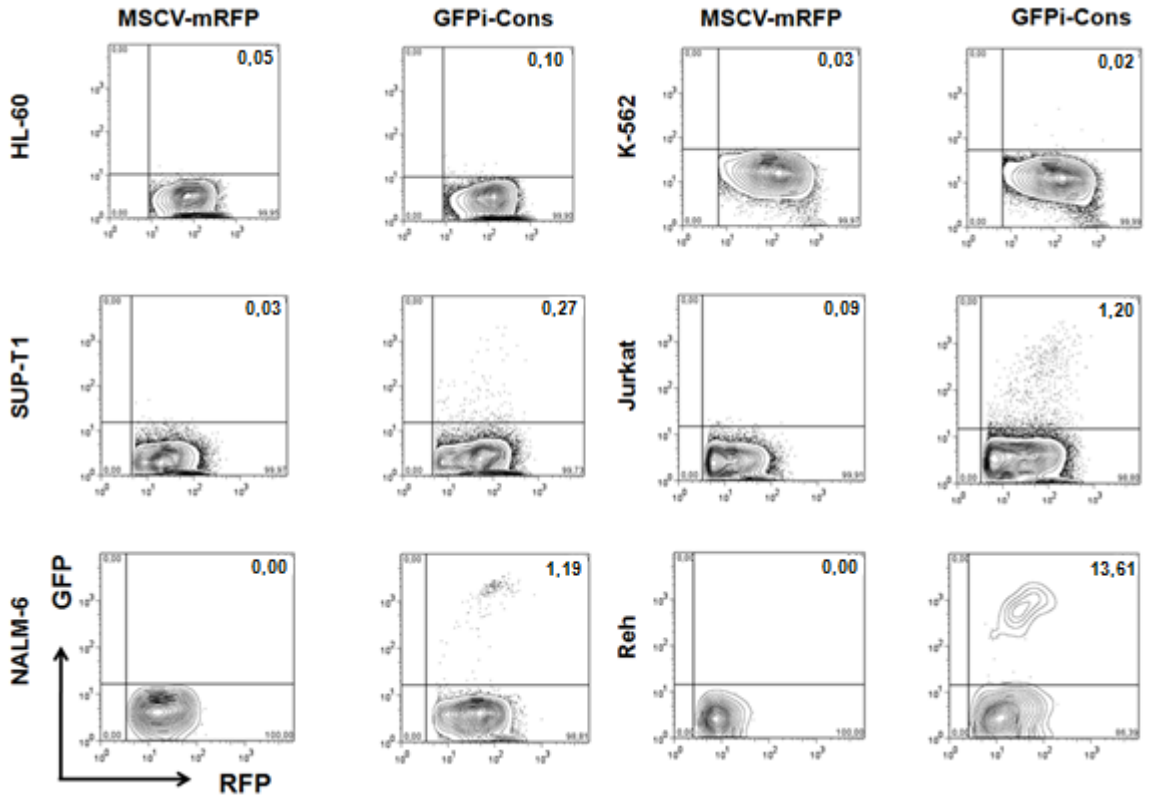


Figure 16. Flow cytometry plots of human tumour cell lines (HL-60, K-562, SUP-T1, Jurkat, NALM-6 and Reh) infected with control (MSCV-mRFP) or GFPi-Cons retroviral-based reporters at three weeks post-infection. All acquisitions are on RFP-positive gated populations, where the GFP-positive events are indicated in frequency on the upper right corner of each plot. This frequency represents the respective ER of a single experiment. The final ER of each cell line is calculated by the following manner: ER (cell line) = ER (GFPi-Cons) – ER (MSCV-mRFP).

This work showed that the GFPi retroviral-based system is efficient in detecting endogenous RAG activity. Moreover, it also demonstrated that GFPi is sensitive to different RAG activity levels (as previously observed *in vitro* in this work).

5. DISCUSSION

5.1. Optimization of the GFPi IVRA

In this work, we were able to establish the GFPi reporter as a highly efficient tool for RAG activity measurement, coupled with a simpler and straightforward method of readout. We were able to optimize this assay, making it highly efficient and obtaining ER values reaching 60%, with a simple IVRA protocol, requiring transfection and dual-fluorescence flow cytometry analysis.

We have found that GFPi is correctly built, successfully presenting a readout of RAG activity. Flanking sequences present on the MSCV vector were found not to be limiting RAG efficiency of recombination since an increase in RAG expression, provided by the CMV-RAG expression plasmids, produced a parallel increment on the ER value. Furthermore, we

observed H2k-driven expression was indeed limiting the ER since accounting for the unexpected low values of ER first obtained with the GFPi dual fluorescence readout.

CMV-RAG levels of expression in the established GFPi-Cons assay produced a substantially higher ER than that measured with pJH290 under endogenous RAG expression[78]. More strikingly, the MSCV-GFPi readout detected a much higher ER than pJH290 system when transfecting similar amounts of CMV-RAG constructs[40]. Furthermore, we concluded that RAG protein saturation levels are achieved for higher numbers of reporter molecules (MIF_{RFP} > 150), as seen when analyzing CMV-RAG GFPi-Cons flow cytometry plots, which present an ER of almost 100% in the highest peak of RFP intensity. These analyses have also shown that saturation levels are achieved in a RFP intensity of 10^2 which indicate an increasing probability of RAG finding reporter molecules and that the assay is limited in RAG activity for the low copy number transfectants. Therefore, we hypothesize that dispersed substrates are more difficult to be targeted by RAG in this case.

It would be interesting to determine the number of reporter plasmids within a cell with a given RFP brightness (namely, by qPCR). With that, it would be possible to determine the number of recombined plasmids within a cell, conferring another type of sensitivity to the ER values (molecule-level instead of cell-level) and better compare the results to those of molecule-determination assays[21]. We found an indication of this value by calculating the mean intensity of GFP fluorescence (MIF_{GFP}), proportional to the average number of recombined reporters per cell, times the percentage of GFP^+ cells, for a population with a given RFP intensity.

A broader window of RAG activity measurement has been achieved by increasing RAG expression levels in the IVRA. Therefore, we conclude that the GFPi assay is suitable for conditions when expecting higher ERs (by using H2k-RAG) and lower ERs (by using CMV-RAG). This optimization allowed for further validation and application of the GFPi reporter in this work.

5.2. GFPi sensitivity to 23-RSS sequence degeneration

Through the use of the GFPi reporter system, we have successfully proven its general sensitivity to differences in 23-RSS spacer sequence composition and consequently to quantification of RAG activity *in vitro*.

We have found that the results obtained follow a general trend to reproduce Kelsoe and colleagues 23-RSS analysis by using the GFPi reporter system, obtaining a gradient of ER values similar to that of the p290T 23-RSS reporters[78], with the exception of GFPi-spaMI and GFPi-(MI)G4. We have confirmed GFPi-spaMI and GFPi-(MI)G4 sequence composition by sequencing analysis. We were particularly surprised by spaMI RIC score and ER results. spaMI, published as the best 23-RSS spacer, should have reflected the best construct

provided by Kelsoe's algorithm (i.e. the highest RIC score). Moreover, since all other sequences are derived from it to the consensus spacer (Cons), its ER was believed to be the highest.

We hypothesize that these differences are due to the type of measurement: GFPi assay differs from Kelsoe and colleagues' in the sense that we have used cell lines devoided of RAG expression for further lipofectamine-mediated transfection with RAG and reporter plasmids (and not a pre-B cell line with endogenous RAG expression, transfected by electroporation with the reporter) and measured RAG activity at the cell level (and not at the molecule level). To support our findings, we used an indication of the number of recombined reporter molecules per cell, MIF_{GFP}, in GFPi-spaMI and GFPi-(MI)C19/20 and compared both MIF_{GFP} values. It is usually considered in flow cytometry analysis that MIF depicts (or is proportional) to the number of existing fluorescent molecules. As fluorescence is a function of the number of gene copies being expressed, we reasoned that the MIF_{GFP} should provide a measure of the number of recombined events (i.e. number of copies in frame) occurring in each analysed cell. Our findings still stand even when performing the MIF_{GFP} analysis (data not shown).

To further confirm this difference we have specifically tested GFPi-spaMI in IVRAs using CMV-RAG expression plasmids to find that GFPi-spaMI still originates a lower ER than that of GFPi-Cons or GFPi-(MI)C19/20 (data not shown).

Regarding the RIC score algorithm, Kelsoe and colleagues state that the RIC score algorithm provides a quantitative value of function. Our analysis of ER and RIC score does not generally corroborate this statement for all constructs tested. Strikingly, the spaMI-CAGnon 23-RSS sequence, containing the CAG-motif nonamer which confers low RAG efficiency[78] to the otherwise spaMI sequence, presents the highest RIC score. This finding demonstrates the algorithm's inability to measure the functional inefficiency of the nonamer coupled with the efficient spaMI heptamer and spacer. The opinions are controversial regarding the informative power of the RIC score. While Lieber et al. support its quantitative sensitivity[83], others have shown, like us, that the algorithm is only powerful qualitatively[91].

In our results, we have observed a narrower window of ER values than in Kelsoe and colleagues work, resulting in a less discriminative assay. However, the replicates increased the reliability of the GFPi assay since its standard deviation values are much narrower than those of p290T, revealing its statistical robustness.

Another factor that might as well have influenced the p290T and GFPi different results is the reporter sequence composition flanking RSS. Roth et al. has made an extensive study on the impact of RSS sequence neighbourhoods on ER. We found that GFPi bears 5'GG/3'CC motifs flanking the 23-RSS sequence and those were predicted to be 5% less functional than the p290T 5'CT/3'AG motifs[92]. Moreover, we speculate that different

neighbourhoods may also influence RSSs in a specific manner, namely by having a synergistic effect in combination with specific spacers, although we have no such evidence.

Finally, we expect these experiments to establish our reporter system as an alternative tool to the classic pJH290 system. It was certainly demonstrated that the GFPi reporter structure is able to provide an excellent quantitative readout. Therefore, an efficient and sensitive RAG activity reporter system was generated, by relying on a single-step analysis, eliminating the expensive, time-consuming and technically challenging procedures inherent to the pJH290 reporter.

5.3. Assessment of RAG-mediated cRSS functionality *in vitro*

We have successfully applied the CMV-RAG GFPi assay as a more suitable approach for assessing cRSS functionality in terms of RAG activity. We have found that all tested 12-cRSSs are targeted by RAG and are engaged in aberrant recombination rearrangements, though with varying frequencies.

The SCL (*stem cell leukaemia*) gene, (also known as TAL-1 - *T-cell acute lymphoblastic leukaemia-1* - or TCL-5) is located on chromosome 1 in the human genome and encodes a transcription factor which plays a role in blood vessel formation, endothelial development and T-cell development during haematopoiesis[93]. Aberrant SCL rearrangements are the most common defects associated with T-ALL, being described in 60% of T-ALL patients[94]. The most frequent, affecting 25% of T-ALL patients, is the SCL-SIL submicroscopic deletion, which is an 82kb interstitial deletion that replaces regulatory SCL 5V sequences by those of an upstream gene, known as SCL-interrupting locus (SIL)[93]. SCL-SIL mutation was described as being induced by RAG illegitimate activity. We have studied the 12-cRSS found to exist within the rearranged region.

Lmo2 (*LIM domain only-2*) gene is located on chromosome 11 and encodes for a transcriptional cofactor that was shown to be involved in vascular endothelial remodelling and hematopoietic development. After differentiation of haematopoietic stem cells, Lmo2 is downregulated in T-lymphocytes[95]. However, RAG illegitimate activity was shown to induce deletion or translocation causing aberrant rearrangements of this gene, ectopic induction of Lmo2 expression and T-ALL onset[85].

When comparing ER values described in the literature to those obtained with GFPi, SCL presented a lower ER than Lmo2 as in Lieber et al. but V_H1/87(181) and Jβ2-2 reversed their ER ranking, contrasting with Kelsoe et al. observations[84, 85]. Nevertheless, it is important to bear in mind that V_H1/87(181) and Jβ2-2 were previously tested through the use of a murine cell line, whereas Lmo2 and SCL in a Reh human pre-B tumour cell line, both already bearing endogenous RAG expression.

Regarding the comparison of RIC score and RSS/cRSS functionality, we found no quantitative relation between the two. All tested sequences were classified as non-functional by the RIC score, with the exception of the control Cons and the Jβ2-2 V(D)J loci RSSs. However, *in vitro*, GFPi-Jβ presented a lower ER than the presumably non-functional V_H. Therefore, in the case of cRSS characterization, the *in vitro* approach assay appears to more accurately predict cRSS functionality than the *in silico* approach.

Surprisingly, two out of the three tested cRSSs (Lmo2 and PTEN) presented similar or higher ER values than the tested V(D)J loci RSSs. We hypothesize that regulatory mechanisms acting upstream of RAG targeting[19], such as chromatin regulation, may be preventing these particular cRSSs from being targeted *in vivo*. Indeed, we observed a unique pattern of recombination frequencies along the RFP gradient (i.e increasing number of reporters) for these cRSSs: for the low copy number of reporter, both Lmo2 and PTEN presented ER values higher than those of Cons; moreover, the proportional increase in ER with the increasing reporter copy number observed in Cons was not observed for these cRSSs, which kept the ER value approximately constant. This suggests that Lmo2 and PTEN cRSSs, without further regulation, have higher probability of recruiting RAG yet low DSB and/or rearrangement capabilities. This further implies that GFPi assay may provide additional, unforeseen, information about the dynamics of RAG/RSS recombination process.

Additionally, GFPi-SCL presented a lower ER than GFPi-Lmo2, in spite of SCL-SIL deletion being more or equally frequent than Lmo2 translocation in T-ALL patients. In this case, we propose a regulatory mechanism having an effect on the incidence of the mutation (in a given cell population), i.e. that the 1p32 SCL/SIL deletional mutation may provide a cellular or populational selective advantage, when compared to the Lmo2 translocation. It would be interesting to explore the contribution of the different mutated regions to cell survival and cell proliferation.

This tool may provide a better understanding of RAG's illegitimate activity, for example by assessing the functionality of other sequences related or still not related to RAG's illegitimate activity (such as the remaining sequences recruited for chromosomal aberrations mentioned in **1.3.2.**). The GFPi reporter system could also be used for testing coupled pairs of RSSs/cRSSs, mimicking a condition similar to a pre-translocation setup, as has been previously done with a few sequence pairs in reporters requiring laborious procedures[68].

5.4. Assessment of RAG-mediated PTEN cRSS functionality *in vitro*

Our results show for the first time that the PTEN tumour suppressor gene is targeted by RAG *in vitro.* We demonstrated that the 12-cRSS adjacent to the start codon of PTEN's

first exon efficiently engages in RAG-mediated recombination *in vitro* and that this phenomenon correlates with a novel mutation found at this site in a T-ALL patient.

This particular cRSS presented a high ER when tested with the GFPi reporter: while GFPi-SCL presented a 52-fold lower ER than Lmo2, GFPi-PTEN was only 3-fold lower. Moreover, PTEN ER value was similar to those of V_H and Jβ V(D)J loci RSSs.

There is the indication of a relation between the mutation found and the occurrence of a RAG-mediated illegitimate event. When performing *in vivo* mutational analysis of the T-ALL patient, this novel deletional mutation, which completely abolished PTEN expression, did not display all the signatures characteristic of a RAG-mediated rearrangement, such as the strand donation signature for hairpin processing. However, we find nucleotide loss and adjacent addition of nucleotides in a potential RAG-induced breakage site and an intact signal end region adjacent to a cRSS that we found that is in fact highly functional. We hypothesize that this region was processed by exonuclease activity followed by nucleotide-addition compatible with TdT activity.

Although we have found that the cRSS present in the first exon of the PTEN gene showed a high efficiency of recombination *in vitro*, this particular exon seems to be rarely targeted for mutation. There is now increasing evidence for exon 7 being the preferencial target for mutation with several mutations from different T-ALL patients already described[96]. It has been hypothesized that the mutational hotspot of exon 7 may relate to RAG activity.

In summary, the cRSS present within the first exon of the PTEN gene was tested for the first time with the use of the GFPi reporter and found to be a substrate as functional as a V(D)J loci RSS for RAG activity. This result thus indicates that PTEN cRSS and V(D)J loci RSS sequences, although similar in functionality, are clearly under different regulatory processes.

5.5. Endogenous RAG activity detection in human haematopoietic tumour cell lines

We have found that GFPi retroviral-based reporter system successfully functioned as an integrative low copy number substrate, being a more physiologic approach of RAG activity measurement in human lymphoid tumour cell lines.

We predicted that GFPi copy number could not exceed four copies per cell [since we established a viral multiplicity of infection (MOI) of four and observed an efficiency of infection of approximately 50%]. When comparing our results to Lieber et al., we find that integrated GFPi displays comparable or higher ER values to those obtained with pJH290-like extrachromosomal substrate[89, 90]. Both integrated GFPi and pJH290-like reporters indicate that Reh bears the highest RAG activity and myeloid cell lines as having the lowest. However, NALM-6 and SUP-T1 presented differing results: integrated GFPi displayed similar ER values for both cell lines whereas Lieber et al. states that NALM-6 presents a 20-fold

higher ER than SUP-T1. Myeloid cell lines were found to be devoided of RAG activity, as expected for a cell line of myeloid origin. Nevertheless, the few observed GFP^+RFP^+ double-positive events in K-562 GFPi-Cons flow cytometry analysis should be taken into account when performing a time-based assessment of RAG activity, since we could be witnessing residual RAG activity that may well become enriched throughout time.

Due to the limited time available for the execution of this task and to its exploratory nature, there were no replicates/independent experiments done in this study. Further work should be done to replicate the infections and perform independent analysis not only to validate the results obtained but also to establish stable infections and analyse alterations in recombination events throughout time.

6. Concluding remarks

We have successfully developed a simple RAG activity measurement tool, the GFPi reporter. This construct allows for transient and stable RAG activity assessment, either by *in vitro* recombination assays or by retroviral integration, gathering an efficient method of readout, provided by dual-fluorescence flow cytometry analysis. The GFPi IVRA was successfully optimized for a broader window of RAG activity measurement and for a functional analysis of cRSSs. We were able to apply this last technique to assess the functionality of a previously undescribed cRSS within the PTEN gene, which correlated with a novel mutation found in a T-ALL patient. This sequence is indeed targeted by RAG, presenting greater efficiency of recombination (ER) than other cRSSs known to be involved in leukaemogenesis and also presenting similar ER values to two tested V(D)J loci RSSs. With the retroviral-based form of GFPi, we were also able to study RAG activity in human leukaemic tumour cell lines with endogenous RAG expression. In this case, we have mimicked a more physiological RAG activity measurement, since we used the integrated form of GFPi in the cell genome in low-copy number. We detected RAG activity in cell lines of lymphoid origin, ranging in ER values in a cell-type specific manner, but not of myeloid origin. Overall, we conclude that the GFPi reporter system satisfies all the conditions for being considered as a suitable tool for RAG activity measurement, coupling a non-laborious assay with a simple readout method. We foresee a number of important applications of this polyvalent tool, ranging from biochemical studies, haematopoiesis and cancer biology research to biomedical applications such as diagnosis.

7. References

1. Janeway, C. A., Jr. Approaching the asymptote? Evolution and revolution in immunology. Cold Spring Harb Symp Quant Biol 54 Pt 1, 1-13 (1989).
2. Matzinger, P. Tolerance, danger, and the extended family. Annu Rev Immunol 12, 991-1045 (1994).
3. Nagawa, F. et al. Antigen-receptor genes of the agnathan lamprey are assembled by a process involving copy choice. Nat Immunol 8, 206-13 (2007).
4. Zhang, S. M., Adema, C. M., Kepler, T. B. & Loker, E. S. Diversification of Ig superfamily genes in an invertebrate. Science 305, 251-4 (2004).
5. Litman, G. W., Cannon, J. P. & Dishaw, L. J. Reconstructing immune phylogeny: new perspectives. Nat Rev Immunol 5, 866-79 (2005).
6. Boehm, T. & Bleul, C. C. The evolutionary history of lymphoid organs. Nat Immunol 8, 131-5 (2007).
7. Horner, C. et al. Role of the innate immune response in sepsis. Anaesthesist 53, 10-28 (2004).
8. Schatz, D. G., Oettinger, M. A. & Baltimore, D. The V(D)J recombination activating gene, RAG-1. Cell 59, 1035-48 (1989).
9. Agrawal, A., Eastman, Q. M. & Schatz, D. G. Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. Nature 394, 744-51 (1998).
10. Hiom, K., Melek, M. & Gellert, M. DNA transposition by the RAG1 and RAG2 proteins: a possible source of oncogenic translocations. Cell 94, 463-70 (1998).
11. Oettinger, M. A., Schatz, D. G., Gorka, C. & Baltimore, D. RAG-1 and RAG-2, adjacent genes that synergistically activate V(D)J recombination. Science 248, 1517-23 (1990).
12. Tonegawa, S. Somatic generation of antibody diversity. Nature 302, 575-81 (1983).
13. Thompson, C. B. New insights into V(D)J recombination and its role in the evolution of the immune system. Immunity 3, 531-9 (1995).
14. Fugmann, S. D., Messier, C., Novack, L. A., Cameron, R. A. & Rast, J. P. An ancient evolutionary origin of the Rag1/2 gene locus. Proc Natl Acad Sci U S A 103, 3728-33 (2006).
15. Jung, D. & Alt, F. W. Unraveling V(D)J recombination; insights into gene regulation. Cell 116, 299-311 (2004).
16. Xiong, N. & Raulet, D. H. Development and selection of gammadelta T cells. Immunol Rev 215, 15-31 (2007).
17. Flajnik, M. F. Comparative analyses of immunoglobulin genes: surprises and portents. Nat Rev Immunol 2, 688-98 (2002).
18. Gellert, M. V(D)J recombination: RAG proteins, repair factors, and regulation. Annu Rev Biochem 71, 101-32 (2002).
19. Roth, D. B. Restraining the V(D)J recombinase. Nat Rev Immunol 3, 656-66 (2003).
20. van Gent, D. C., Ramsden, D. A. & Gellert, M. The RAG1 and RAG2 proteins establish the 12/23 rule in V(D)J recombination. Cell 85, 107-13 (1996).
21. Hesse, J. E., Lieber, M. R., Gellert, M. & Mizuuchi, K. Extrachromosomal DNA substrates in pre-B cells undergo inversion or deletion at immunoglobulin V-(D)-J joining signals. Cell 49, 775-83 (1987).
22. Jones, J. M. & Gellert, M. Ordered assembly of the V(D)J synaptic complex ensures accurate recombination. Embo J 21, 4162-71 (2002).
23. Grundy, G. J. et al. Initial stages of V(D)J recombination: the organization of RAG1/2 and RSS DNA in the postcleavage complex. Mol Cell 35, 217-27 (2009).
24. Ma, Y., Pannicke, U., Schwarz, K. & Lieber, M. R. Hairpin opening and overhang processing by an Artemis/DNA-dependent protein kinase complex in nonhomologous end joining and V(D)J recombination. Cell 108, 781-94 (2002).

25. Lieber, M. R., Hesse, J. E., Mizuuchi, K. & Gellert, M. Lymphoid V(D)J recombination: nucleotide insertion at signal joints as well as coding joints. Proc Natl Acad Sci U S A 85, 8588-92 (1988).
26. Melek, M., Gellert, M. & van Gent, D. C. Rejoining of DNA by the RAG1 and RAG2 proteins. Science 280, 301-3 (1998).
27. Hsu, E., Pulham, N., Rumfelt, L. L. & Flajnik, M. F. The plasticity of immunoglobulin gene systems in evolution. Immunol Rev 210, 8-26 (2006).
28. Goldsby, R., Kindt, T.J., Osborne B.A. & Kuby, J. Immunology (W. H. Freeman and Company, New York, 2003).
29. Stanhope-Baker, P., Hudson, K. M., Shaffer, A. L., Constantinescu, A. & Schlissel, M. S. Cell type-specific chromatin structure determines the targeting of V(D)J recombinase activity in vitro. Cell 85, 887-97 (1996).
30. Matthews, A. G. & Oettinger, M. A. RAG: a recombinase diversified. Nat Immunol 10, 817-21 (2009).
31. Matthews, A. G. et al. RAG2 PHD finger couples histone H3 lysine 4 trimethylation with V(D)J recombination. Nature 450, 1106-10 (2007).
32. Spicuglia, S. et al. Promoter activation by enhancer-dependent and -independent loading of activator and coactivator complexes. Mol Cell 10, 1479-87 (2002).
33. Hawwari, A., Bock, C. & Krangel, M. S. Regulation of T cell receptor alpha gene assembly by a complex hierarchy of germline Jalpha promoters. Nat Immunol 6, 481-9 (2005).
34. Whitehurst, C. E., Chattopadhyay, S. & Chen, J. Control of V(D)J recombinational accessibility of the D beta 1 gene segment at the TCR beta locus by a germline promoter. Immunity 10, 313-22 (1999).
35. Khor, B., Mahowald, G. K., Khor, K. & Sleckman, B. P. Functional overlap in the cis-acting regulation of the V(D)J recombination at the TCRbeta locus. Mol Immunol 46, 321-6 (2009).
36. Wang, X. et al. Regulation of Tcrb recombination ordering by c-Fos-dependent RAG deposition. Nat Immunol 9, 794-801 (2008).
37. Baumann, M., Mamais, A., McBlane, F., Xiao, H. & Boyes, J. Regulation of V(D)J recombination by nucleosome positioning at recombination signal sequences. Embo J 22, 5197-207 (2003).
38. Bates, J. G., Cado, D., Nolla, H. & Schlissel, M. S. Chromosomal position of a VH gene segment determines its activation and inactivation as a substrate for V(D)J recombination. J Exp Med 204, 3247-56 (2007).
39. Sleckman, B. P. et al. Mechanisms that direct ordered assembly of T cell receptor beta locus V, D, and J gene segments. Proc Natl Acad Sci U S A 97, 7975-80 (2000).
40. Lee, J. & Desiderio, S. Cyclin A/CDK2 regulates V(D)J recombination by coordinating RAG-2 accumulation and DNA repair. Immunity 11, 771-81 (1999).
41. Hewitt, S. L. et al. RAG-1 and ATM coordinate monoallelic recombination and nuclear positioning of immunoglobulin loci. Nat Immunol 10, 655-64 (2009).
42. Aidinis, V. et al. The RAG1 homeodomain recruits HMG1 and HMG2 to facilitate recombination signal sequence binding and to enhance the intrinsic DNA-bending activity of RAG1-RAG2. Mol Cell Biol 19, 6532-42 (1999).
43. Schatz, D. G. Antigen receptor genes and the evolution of a recombinase. Semin Immunol 16, 245-56 (2004).
44. Yin, F. F. et al. Structure of the RAG1 nonamer binding domain with DNA reveals a dimer that mediates DNA synapsis. Nat Struct Mol Biol 16, 499-508 (2009).
45. Gwyn, L. M., Peak, M. M., De, P., Rahman, N. S. & Rodgers, K. K. A zinc site in the C-terminal domain of RAG1 is essential for DNA cleavage activity. J Mol Biol 390, 863-78 (2009).
46. Fugmann, S. D. & Schatz, D. G. Identification of basic residues in RAG2 critical for DNA binding by the RAG1-RAG2 complex. Mol Cell 8, 899-910 (2001).

47. Qiu, J. X., Kale, S. B., Yarnell Schultz, H. & Roth, D. B. Separation-of-function mutants reveal critical roles for RAG2 in both the cleavage and joining steps of V(D)J recombination. Mol Cell 7, 77-87 (2001).
48. Zhao, S., Gwyn, L. M., De, P. & Rodgers, K. K. A non-sequence-specific DNA binding mode of RAG1 is inhibited by RAG2. J Mol Biol 387, 744-58 (2009).
49. Godderz, L. J., Rahman, N. S., Risinger, G. M., Arbuckle, J. L. & Rodgers, K. K. Self-association and conformational properties of RAG1: implications for formation of the V(D)J recombinase. Nucleic Acids Res 31, 2014-23 (2003).
50. Akamatsu, Y. et al. Deletion of the RAG2 C terminus leads to impaired lymphoid development in mice. Proc Natl Acad Sci U S A 100, 1209-14 (2003).
51. Curry, J. D. et al. Chromosomal reinsertion of broken RSS ends during T cell development. J Exp Med 204, 2293-303 (2007).
52. Monroe, R. J., Chen, F., Ferrini, R., Davidson, L. & Alt, F. W. RAG2 is regulated differentially in B and T cells by elements 5' of the promoter. Proc Natl Acad Sci U S A 96, 12713-8 (1999).
53. Patra, A. K. et al. PKB rescues calcineurin/NFAT-induced arrest of Rag expression and pre-T cell differentiation. J Immunol 177, 4567-76 (2006).
54. Kuo, T. C. & Schlissel, M. S. Mechanisms controlling expression of the RAG locus during lymphocyte development. Curr Opin Immunol 21, 173-8 (2009).
55. Jiang, H. et al. Ubiquitylation of RAG-2 by Skp2-SCF links destruction of the V(D)J recombinase to the cell cycle. Mol Cell 18, 699-709 (2005).
56. Verkoczy, L. et al. Basal B cell receptor-directed phosphatidylinositol 3-kinase signaling turns off RAGs and promotes B cell-positive selection. J Immunol 178, 6332-41 (2007).
57. Cebrat, M., Miazek, A. & Kisielow, P. Identification of a third evolutionarily conserved gene within the RAG locus and its RAG1-dependent and -independent regulation. Eur J Immunol 35, 2230-8 (2005).
58. Igarashi, H., Gregory, S. C., Yokota, T., Sakaguchi, N. & Kincade, P. W. Transcription from the RAG1 locus marks the earliest lymphocyte progenitors in bone marrow. Immunity 17, 117-30 (2002).
59. Borghesi, L. et al. B lineage-specific regulation of V(D)J recombinase activity is established in common lymphoid progenitors. J Exp Med 199, 491-502 (2004).
60. Pilbeam, K. et al. The ontogeny and fate of NK cells marked by permanent DNA rearrangements. J Immunol 180, 1432-41 (2008).
61. Shaffer, A. L., Rosenwald, A. & Staudt, L. M. Lymphoid malignancies: the dark side of B-cell differentiation. Nat Rev Immunol 2, 920-32 (2002).
62. Lewis, S. M., Agard, E., Suh, S. & Czyzyk, L. Cryptic signals and the fidelity of V(D)J joining. Mol Cell Biol 17, 3125-36 (1997).
63. Cowell, L. G., Davila, M., Yang, K., Kepler, T. B. & Kelsoe, G. Prospective estimation of recombination signal efficiency and identification of functional cryptic signals in the genome by statistical modeling. J Exp Med 197, 207-20 (2003).
64. Bories, J. C., Cayuela, J. M., Loiseau, P. & Sigaux, F. Expression of human recombination activating genes (RAG1 and RAG2) in neoplastic lymphoid cells: correlation with cell differentiation and antigen receptor expression. Blood 78, 2053-61 (1991).
65. Tsujimoto, Y. et al. Clustering of breakpoints on chromosome 11 in human B-cell neoplasms with the t(11;14) chromosome translocation. Nature 315, 340-3 (1985).
66. Iida, S. et al. The t(9;14)(p13;q32) chromosomal translocation associated with lymphoplasmacytoid lymphoma involves the PAX-5 gene. Blood 88, 4110-7 (1996).
67. Cheng, J. T., Yang, C. Y., Hernandez, J., Embrey, J. & Baer, R. The chromosome translocation (11;14)(p13;q11) associated with T cell acute leukemia. Asymmetric diversification of the translocational junctions. J Exp Med 171, 489-501 (1990).
68. Marculescu, R., Le, T., Simon, P., Jaeger, U. & Nadel, B. V(D)J-mediated translocations in lymphoid neoplasms: a functional assessment of genomic instability by cryptic sites. J Exp Med 195, 85-98 (2002).

69. Aplan, P. D. et al. Disruption of the human SCL locus by "illegitimate" V-(D)-J recombinase activity. Science 250, 1426-9 (1990).
70. Chapiro, E. et al. Expression of T-lineage-affiliated transcripts and TCR rearrangements in acute promyelocytic leukemia: implications for the cellular target of t(15;17). Blood 108, 3484-93 (2006).
71. Dupret, C. et al. IgH/TCR rearrangements are common in MLL translocated adult AML and suggest an early T/myeloid or B/myeloid maturation arrest, which correlates with the MLL partner. Leukemia 19, 2337-8 (2005).
72. Foa, R. et al. Rearrangements of immunoglobulin and T cell receptor beta and gamma genes are associated with terminal deoxynucleotidyl transferase expression in acute myeloid leukemia. J Exp Med 165, 879-90 (1987).
73. Fenstermaker, R. A. & Ciesielski, M. J. EGFR Intron Recombination in Human Gliomas: Inappropriate Diversion of V(D)J Recombination? Curr Genomics 8, 163-70 (2007).
74. Borghesi, L. & Gerstein, R. M. Developmental separation of V(D)J recombinase expression and initiation of IgH recombination in B lineage progenitors in vivo. J Exp Med 199, 483-9 (2004).
75. Lieber, M. R. et al. The defect in murine severe combined immune deficiency: joining of signal sequences but not coding segments in V(D)J recombination. Cell 55, 7-16 (1988).
76. Yancopoulos, G. D. et al. A novel fluorescence-based system for assaying and separating live cells according to VDJ recombinase activity. Mol Cell Biol 10, 1697-704 (1990).
77. Zheng, X. & Schwarz, K. Making V(D)J rearrangement visible: quantification of recombination efficiency in real time at the single cell level. J Immunol Methods 315, 133-43 (2006).
78. Cowell, L. G., Davila, M., Ramsden, D. & Kelsoe, G. Computational tools for understanding sequence variability in recombination signals. Immunol Rev 200, 57-69 (2004).
79. Cowell, L. G., Davila, M., Kepler, T. B. & Kelsoe, G. Identification and utilization of arbitrary correlations in models of recombination signal sequences. Genome Biol 3, RESEARCH0072 (2002).
80. Sarmento, L. M. et al. Notch1 modulates timing of G1-S progression by inducing SKP2 transcription and p27 Kip1 degradation. J Exp Med 202, 157-68 (2005).
81. Campbell, R. E. et al. A monomeric red fluorescent protein. Proc Natl Acad Sci U S A 99, 7877-82 (2002).
82. Todaro, G. J. & Green, H. Quantitative studies of the growth of mouse embryo cells in culture and their development into established lines. J Cell Biol 17, 299-313 (1963).
83. Lee, A. I. et al. A functional analysis of the spacer of V(D)J recombination signal sequences. PLoS Biol 1, E1 (2003).
84. Davila, M. et al. Multiple, conserved cryptic recombination signals in VH gene segments: detection of cleavage products only in pro B cells. J Exp Med 204, 3195-208 (2007).
85. Raghavan, S. C., Kirsch, I. R. & Lieber, M. R. Analysis of the V(D)J recombination efficiency at lymphoid chromosomal translocation breakpoints. J Biol Chem 276, 29126-33 (2001).
86. Jiang, B. H. & Liu, L. Z. PI3K/PTEN signaling in tumorigenesis and angiogenesis. Biochim Biophys Acta 1784, 150-8 (2008).
87. Silva, A. et al. PTEN posttranslational inactivation and hyperactivation of the PI3K/Akt pathway sustain primary T cell leukemia viability. J Clin Invest 118, 3762-74 (2008).
88. Tenen, D. G. Disruption of differentiation in human cancer: AML shows the way. Nat Rev Cancer 3, 89-101 (2003).
89. Gauss, G. H., Domain, I., Hsieh, C. L. & Lieber, M. R. V(D)J recombination activity in human hematopoietic cells: correlation with developmental stage and genome stability. Eur J Immunol 28, 351-8 (1998).
90. Lieber, M. R., Hesse, J. E., Mizuuchi, K. & Gellert, M. Developmental stage specificity of the lymphoid V(D)J recombination activity. Genes Dev 1, 751-61 (1987).
91. Zhang, M. & Swanson, P. C. V(D)J recombinase binding and cleavage of cryptic recombination signal sequences identified from lymphoid malignancies. J Biol Chem 283, 6717-27 (2008).

92. Wong, S. Y., Lu, C. P. & Roth, D. B. A RAG1 mutation found in Omenn syndrome causes coding flank hypersensitivity: a novel mechanism for antigen receptor repertoire restriction. J Immunol 181, 4124-30 (2008).
93. Cheng, Y., Zhang, Z., Slape, C. & Aplan, P. D. Cre-loxP-mediated recombination between the SIL and SCL genes leads to a block in T-cell development at the CD4- CD8- to CD4+ CD8+ transition. Neoplasia 9, 315-21 (2007).
94. Ferrando, A. A. et al. Biallelic transcriptional activation of oncogenic transcription factors in T-cell acute lymphoblastic leukemia. Blood 103, 1909-11 (2004).
95. Mao, S., Neale, G. A. & Goorha, R. M. T-cell proto-oncogene rhombotin-2 is a complex transcription regulator containing multiple activation and repression domains. J Biol Chem 272, 5594-9 (1997).
96. Gutierrez, A. et al. High frequency of PTEN, PI3K, and AKT abnormalities in T-cell acute lymphoblastic leukemia. Blood 114, 647-50 (2009).

APPENDIX

TABLES

Table I. List of PCR products and respective reverse primers, 23-RSS.

PCR name	Reverse complement primer (23-RSS)	Kelsoe et al. 23-RSSs and GFPi		
GFPi-Cons	5'-CCC GGG GGT TTT TGT ACA GCC AGA CAG TGG AGT ACT ACC ACT GTG CCA CCA TGG TGA GCA AGG GCG AGG AGC-3'	CACAGTG	GT AGT ACT CCA CTG TCT GGC TGT	ACAAAAACC
GFPi-ConK	5'-CCC GGG GGT TTT TGT ACA GGC TCC CGA TGT GGT TCC AAC ACT GTG CCA CCA TGG TGA GCA AGG GCG AGG AGC-3'	CACAGTG	TT GGA ACC ACA TCG GGA GCC TGT	ACAAAAACC
GFPi-spaMI	5'-CCC GGG GGT TTT TGT ACA CAC TCA GGA TGT GGT TGC AAC ACT GTG CCA CCA TGG TGA GCA AGG GCG AGG AGC-3'	CACAGTG	TT GCA ACC ACA TCC TGA GTG TGT	ACAAAAACC
GFPi-(MI)G4	5'-CCC GGG GGT TTT TGT ACA CAC TCA GGA TGT GGT TCC AAC ACT GTG CCA CCA TGG TGA GCA AGG GCG AGG AGC-3'	CACAGTG	TT GGA ACC ACA TCC TGA GTG TGT	ACAAAAACC
GFPi-(MI)G14/15	5'-CCC GGG GGT TTT TGT ACA CAC TCC CGA TGT GGT TGC AAC ACT GTG CCA CCA TGG TGA GCA AGG GCG AGG AGC-3'	CACAGTG	TT GCA ACCACA TCG GGA GTG TGT	ACAAAAACC
GFPi-(MI)C19/20	5'-CCC GGG GGT TTT TGT ACA GGC TCA GGA TGT GGT TGC AAC ACT GTG CCA CCA TGG TGA GCA AGG GCG AGG AGC-3'	CACAGTG	TT GCA ACC ACA TCC TGA GCC TGT	ACAAAAACC
GFPi-spaMI-CAGnon	5'-CCC GGG GGG TTT CTG ACA CAC TCA GGA TGT GGT TGC AAC ACT GTG CCA CCA TGG TGA GCA AGG GCG AGG AGC-3'	CACAGTG	TT GCA ACC ACA TCC TGA GTG TGT	CAGAAACCC

Table II. List of PCR products and respective reverse primers, 12-RSS.

PCR name	Reverse complement primer (12-RSS)	12-RSSs and cRSS		
GFPi-V_H	5'-CTC GAG CAC TAT TAG GAT CAA TCC TTC AAA TCC ATT ACT TGT ACA GCT CGT CCA TGC-3'	CACTATT	AG GAT CAA TCC T	TCAAATCCA
GFPi-Jβ	5'-CTC GAG CAC AGT CGT CGA AAT GCT GGC ACA AAC CTT ACT TGT ACA GCT CGT CCA TGC-3'	CACAGTC	GT CGA AAT GCT G	GCACAAACC
GFPi-Lmo2	5'-CTC GAG AAC ACA CAC AGT ATT GTC TTA CCC AGC AAT AAT TTT ACT TGT ACA GCT CGT CCA TGC-3'	CACAGTA	TT GTC TTA CCC A	GCAATAATT
GFPi-SCL	5'-CTC GAG ACC AAC CAC AGC CTC GCG CAT TTC TGT ATA TTG CTT ACT TGT ACA GCT CGT CCA TGC-3'	CACAGCC	TC GCG CAT TTC T	GTATATTGC
GFPi-PTEN	5'-CTC GAG CAT CAT CAA AGA GAT CGT TAG CAG AAA CAA AAG GTT ACT TGT ACA GCT CGT CCA TGC-3'	CAAAGAG	AT CGT TAG CAG A	AACAAAAGG

Table III. List of 23-RSS reporters and respective RIC score, average efficiency of recombination (ER) and standard deviation (SD) in pT290[78] and H2k-RAG GFPi IVRAs (with two to three independent experiments, each with n=3 replicates).

Reporters	p290T		H2k-RAG GFPi		RIC score
	Average ER (%)	SD	Average ER (%)	SD	
spaMI-CAGnon	0,6	0,6	0,51	0,19	-15,6
ConK	4,6	1,2	2,23	0,66	-21,8
Cons	7,6	2,1	2,46	0,38	-27,3
(MI)G14/15	8,5	2,2	3,64	0,99	-21,7
(MI)G4	9,7	3,5	1,23	0,4	-18
(MI)C19/20	17,2	7,3	4,73	1,16	-17,5
spaMI	28,2	5,1	0,67	0,13	-17,7

Table IV. Fold difference coefficients between ER values of the p290T

Reporters	p290T-spaMI-CAGnon	p290T-ConK	p290T-Cons	p290T-(MI)G14/15	p290T-(MI)G4	p290T-(MI)C19/20	p290T-spaMI
p290T-spaMI-CAGnon		7,7	12,7	14,2	16,2	28,7	47
p290T-ConK			1,7	1,8	2,1	3,7	6,1
p290T-Cons				1,1	1,3	2,3	3,7
p290T-(MI)G14/15					1,1	2	3,3
p290T-(MI)G4						1,8	2,9
p290T-(MI)C19/20							1,6
p290T-spaMI							

Table V. Fold difference coefficients between ER values of the H2k-RAG GFPi

Reporters	GFPi-spaMI-CAGnon	GFPi-ConK	GFPi-Cons	GFPi-(MI)G14/15	GFPi-(MI)G4	GFPi-(MI)C19/20	GFPi-spaMI
GFPi-spaMI-CAGnon		4,4	4,8	7,1	2,4	9,3	1,3
GFPi-ConK			1,1	1,6	0,6	2,1	0,3
GFPi-Cons				1,5	0,5	1,9	0,3
GFPi-(MI)G14/15					0,3	1,3	0,2
GFPi-(MI)G4						3,8	0,5
GFPi-(MI)C19/20							0,1
GFPi-spaMI							

Table VI. Coefficient of fold difference between ER values of p290T and H2k-RAG GFPi .

Reporters	spaMI-CAGnon	ConK	Cons	(MI)G14/15	(MI)G4	(MI)C19/20	spaMI
spaMI-CAGnon		1,8	2,6	2	6,7	3,1	35,8
ConK			1,5	1,1	3,8	1,8	20,4
Cons				0,8	2,6	1,2	13,6
(MI)G14/15					3,4	1,6	18
(MI)G4						0,5	5,3
(MI)C19/20							11,6
spaMI							

Table VII. List of 23-RSS reporters, respective efficiency of recombination (ER) and standard deviation (SD) in CMV-RAG GFPi IVRAs (one to two independent experiments with n=3 replicates each).

Reporters	CMV-RAG GFPi	
	Average ER (%)	SD
spaMI-CAGnon	14,49	0,52
Cons	54,2	1
(MI)C19/20	58,58	0,63
spaMI	46,52	0,58

Table VIII. List of RSS (Cons, V_H and Jβ) and cRSS (Lmo2, SCL) reporters, respective efficiency of recombination (ER) and standard deviation (SD) in literature reporter assays and CMV-RAG GFPi IVRAs (one to two independent experiments with n=3 replicates each).

RSS/cRSS	Literature		GFPi		RIC score
	Average ER (%)	SD	Average ER (%)	SD	
Cons*[1]	7,6	2,1	35,73	1,02	-9,4
VH*[2]	0,01	0,01	6,77	0,39	-44
Jβ*[2]	1,1	0,7	4,41	0,13	-37,8
Lmo2*[3]	0,34	-	14,67	0,16	-42
SCL*[3]	<0,00031	-	0,28	0,1	-53

*[1] Kelsoe et al., 2004 *[2] Kelsoe e tal, 2007 *[3] Lieber e tal, 2001

Table IX. List of human haematopoietic cell lines and respective efficiency of recombination (ER) measured either in the literature (pGG49) or with the retroviral-based GFPi-Cons reporter.

Cell line	ER(%)	
	Replicative pGG49	Integrative GFPi-Cons
HL-60	0,008	0,05
K-562	0,02	0
SUP-T1	0,09	0,24
Jurkat	-	1,11
NALM-6	1,8	0,19
Reh	21,6	13,61

FIGURES

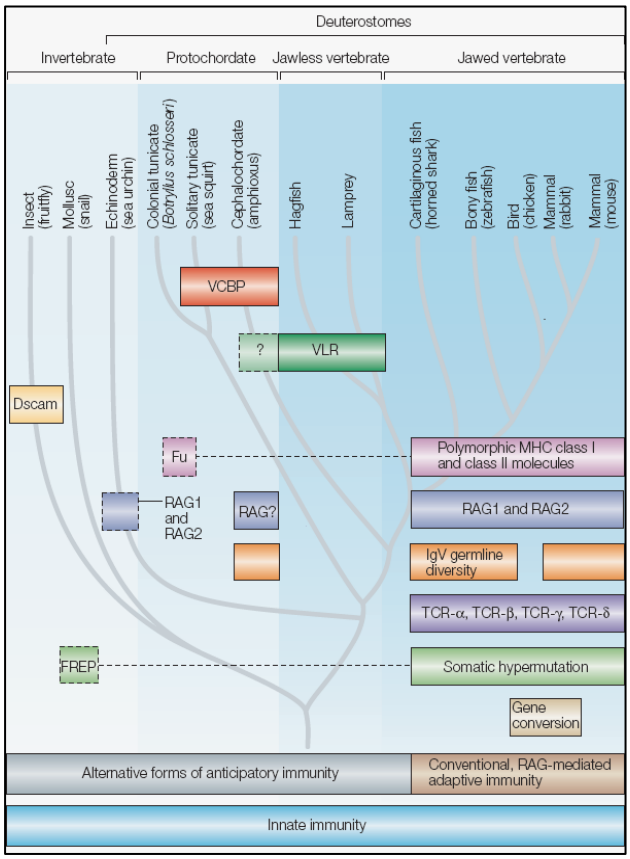


Figure I. Phylogenic representation of immune function characteristics, mechanisms of genetic organization and diversity generation events in selected species. *In* Dishaw et al, 2005.

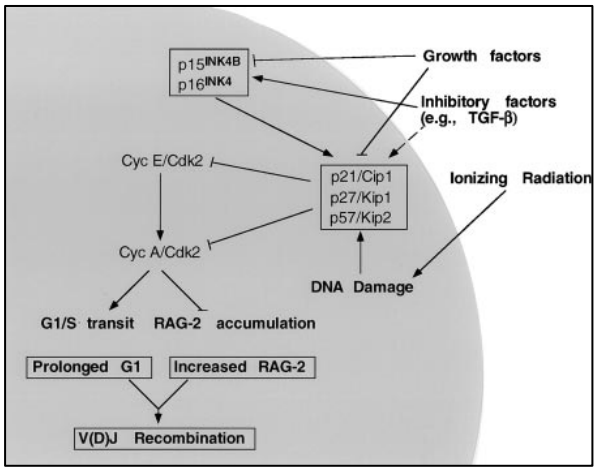


Figure II. Schematic representation exemplifying an environmental regulatory mechanism affecting V(D)J recombination. *In* Desiderio et al., 1999.

CTCGAGCACAGTGCTACAGACTGGAACAAAAACCTTACTTGTACAGCTCGTCCATGCCGAGAGTGATCCCGGCGGCGGTCACGAA
CTCCAGCAGGACCATGTGATCGCGCTTCTCGTTGGGGTCTTTGCTCAGGGCGGACTGGGTGCTCAGGTAGTGGTTGTCGGGCAGC
AGCACGGGGCCGTCGCCGATGGGGGTGTTCTGCTGGTAGTGGTCGGCGAGCTGCACGCTGCCGTCCTCGATGTTGTGGCGGATCT
TGAAGTTCACCTTGATGCCGTTCTTCTGCTTGTCGGCCATGATATAGACGTTGTGGCTGTTGTAGTTGTACTCCAGCTTGTGCCCCAG
GATGTTGCCGTCCTCCTTGAAGTCGATGCCCTTCAGCTCGATGCGGTTCACCAGGGTGTCGCCCTCGAACTTCACCTCGGCGCGGGT
CTTGTAGTTGCCGTCGTCCTTGAAGAAGATGGTGCGCTCCTGGACGTAGCCTTCGGGCATGGCGGACTTGAAGAAGTCGTGCTGCT
TCATGTGGTCGGGGTAGCGGCTGAAGCACTGCACGCCGTAGGTCAGGGTGGTCACGAGGGTGGGCCAGGGCACGGGCAGCTTG
CCGGTGGTGCAGATGAACTTCAGGGTCAGCTTGCCGTAGGTGGCATCGCCCTCGCCCTCGCCGGACACGCTGAACTTGTGGCCGT
TTACGTCGCCGTCCAGCTCGACCAGGATGGGCACCACCCCGGTGAACAGCTCCTCGCCCTTGCTCACCATGGTGGCACAGTGGTAG
TACTCCACTGTCTGGCTGTACAAAAACC**CCCGGG**

Figure III. Expected GFPi-Cons PCR product sequence (5' to 3'). The inverted complementary sequence of the Kozak (in light-green) fused to the *GFP* (green) are flanked at 5' by a 12- and at 3' by a 23-RSS. RSSs are composed by a heptamer (in red), a spacer (in black) and a nonamer (in blue). 5' *Xho*I and 3' *Sma*I restriction sites are indicated in black bold. Primer regions are underlined.

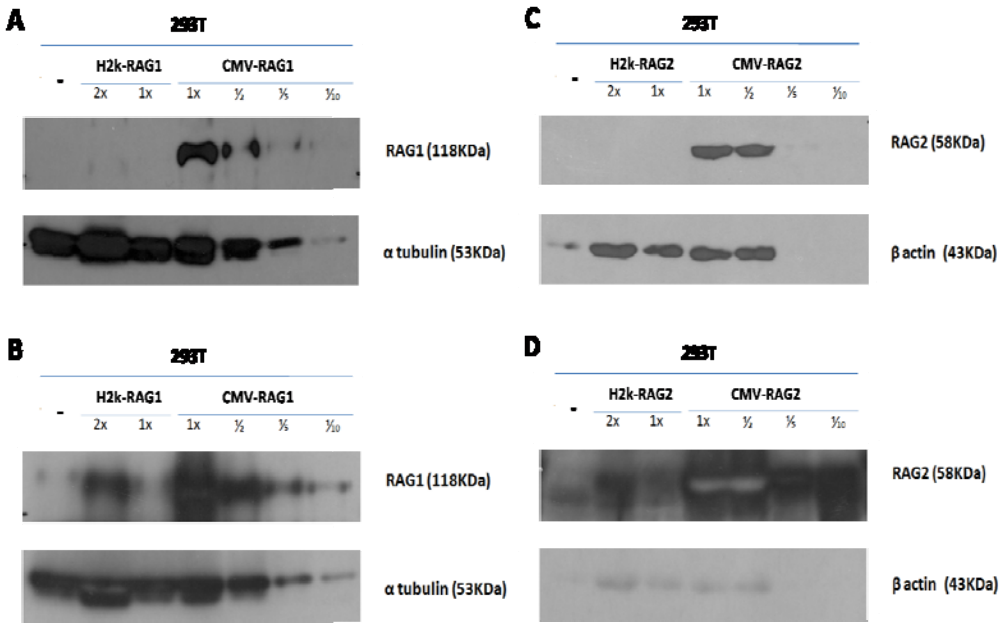


Figure IV. Immunoblot of H2k and CMV-RAG proteins. 293T cells were transfected either with mock (-), H2k-RAG1, CMV-RAG1, H2k-RAG2 or CMV-RAG2 plasmids and the respective protein extracts were further used for Immunoblot analysis. H2k-RAG1, CMV-RAG1 and α-tubulin (internal control) were detected by using **A.** anti-RAG1 antibody in a concentration of 1:500 and anti-α-tubulin in 1:500 and **B.** anti-RAG1 1:100 and anti-α-tubulin 1:1000. H2k-RAG2, CMV-RAG2 and β-actin (internal control) were detected by using **C.** anti-RAG2 antibody in a concentration of 1:500 and anti-β-actin 1:2000 and **D.** anti-RAG2 1:50 and anti-β-actin 1:1000.