**Universidad Autónoma de Madrid**

Programa de Doctorado en Biociencias Moleculares


# Molecular mechanisms of Activation Induced Deaminase specificity

PhD thesis

Ángel F. Álvarez Prado

**Madrid, 2018**

**Departamento de Bioquímica**

Facultad de Medicina

**Universidad Autónoma de Madrid**

# Molecular mechanisms of Activation Induced Deaminase specificity

Memoria presentada por el Licenciado en Biotecnología

Ángel F. Álvarez Prado

para optar al título de Doctor por la Universidad Autónoma de Madrid.

Directora de tesis:

Dra. Almudena Rodríguez Ramiro

Esta tesis doctoral ha sido realizada en el Laboratorio de Biología de Linfocitos B del Centro Nacional de Investigaciones Cardiovasculares Carlos III (CNIC).

Madrid, 2018

Memoria presentada por **Ángel F. Álvarez Prado**, Licenciado en Biotecnología, para optar al grado de Doctor por la Universidad Autónoma de Madrid.

Esta tesis doctoral ha sido realizada en el Laboratorio de Biología de Linfocitos B del Centro Nacional de Investigaciones Cardiovasculares Carlos III (CNIC), bajo la dirección de la **Doctora Almudena Rodríguez Ramiro**, y para que así conste y a los efectos oportunos, firma el siguiente certificado;

En Madrid, a 28 de Mayo de 2018.

Almudena Rodríguez Ramiro

# Acknowledgements

I feel privileged to have worked under the direction of Dr. Almudena R. Ramiro, a brilliant scientist and excellent mentor. My first words of gratitude are devoted to her trust, careful supervision, inspiring insights and patience. I would also like to thank our collaborators, Dr. J.R. Regueiro and Dr. Juan Méndez for the stimulating opportunity to work in other disciplines and CNIC core facilites for their instrumental technical support.

My special thanks go to the first generation of "hypermutants" for sharing their passion for science, teaching me so much in such a short period of time and being the living proof that hard work and fun are not only compatible but highly advisable. Thank you for making me feel at home from the first day.

A word of gratitude is also due to my former and current labmates for their relentless support, scientific and non-scientific discussions, their efforts to survive my intense data seminars and for operating as a safety valve when stress reached dangerous levels.

I am truly indebted to my friends. They encouraged me to keep on going when things went tough and backed me in the hardest moments. Some of them have gone through this exciting but yet exhausting stage and perfectly understand how important is to count on someone else. For those of you who have not, lucky you! Living outside the greedy tentacles of science will provide enough spare time to pursue crazy ideas and gather unforgettable moments. Please, do share some of them with us scientists, we poor things.

There are a number of others who stepped in at some crucial point to push me forward, sometimes inadvertently or just by pure chance. I am somehow indebted to them as well. Whether you are reading this or not, thank you.

Finally, I owe my deepest gratitude to my parents, my brother and my partner. Not only have they been the stand point from which to move the world, but they have been my world at the same time. This thesis work is dedicated to you.

Ángel F. Álvarez.

Madrid, May 2018.

## Resumen

Los linfocitos B ejercen una función fundamental en la inmunidad humoral mediante la secreción de anticuerpos. El evento más característico de la biología de los linfocitos B maduros es la diversificación secundaria de sus genes de inmunoglobulinas durante la reacción de centro germinal (CG) para generar un repertorio prácticamente ilimitado de anticuerpos con distintas especificidades. La desaminasa inducida por activación (AID, de sus siglas en inglés) inicia la diversificación secundaria de anticuerpos en linfocitos B de CG mediante la desaminación de citosinas en los genes de inmunoglobulinas. Sorprendentemente, AID también puede ejercer su actividad en otras regiones del genoma, dando lugar a mutaciones o translocaciones cromosómicas con potencial oncogénico. Por tanto, es fundamental comprender los mecanismos responsables de la especificidad de diana de esta enzima. Sin embargo, su estudio se ha visto limitado por la extrema dificultad para detectar mutaciones inducidas por AID, ya que ocurren en muy baja frecuencia.

En este trabajo hemos desarrollado una novedosa aproximación basada en captura y enriquecimiento para la identificación de dianas mutacionales de AID en linfocitos B de CG. Secuenciamos 1588 regiones genómicas con una elevada profundidad de lectura e identificamos 275 genes mutados por AID, incluyendo 30 de las 35 dianas de AID previamente descritas. Además, hemos identificado un nuevo "punto caliente" (*hotspot*) para la actividad de AID. Basándonos en las características moleculares observadas en los genes mutados por AID hemos desarrollamos un modelo de aprendizaje automático (*machine learning*) que permite predecir nuevas dianas mutacionales de AID y lo hemos validado experimentalmente. También encontramos que las vías de reparación por excisión de bases y de desapareamiento de bases se respaldan mutuamente para reparar de forma fiel la mayor parte de las lesiones inducidas por AID. Por último, nuestros datos establecen un nuevo vínculo entre las mutaciones inducidas por AID y el desarrollo de linfomas.

## Abstract

B lymphocytes are key effectors of the humoral immune response through the secretion of antibodies. The most distinctive event in mature B lymphocytes biology is the secondary diversification of their immunoglobulin genes during the germinal center (GC) reaction, which is fundamental to generate a repertoire of antibodies with virtually unlimited specificities. Activation Induced Deaminase (AID) initiates secondary antibody diversification in GC B cells through the deamination of cytosines on immunoglobulin genes. Remarkably, AID can also target other regions in the genome, triggering mutations or chromosome translocations, with major implications for oncogenic transformation. However, understanding the specificity of AID has proved extremely challenging, mostly because of the difficulty to detect AID-induced mutations, which occur at very low frequencies.

In this work we have developed a novel capture-based approach to explore AID mutagenesis in a representation of the B cell genome. We have sequenced at very high depth 1588 genomic regions from GC B cells and identified 275 genes targeted by AID, including 30 of the previously known 35 AID targets. We have also identified the most highly mutated hotspot for AID activity described to date. Further, integrative analysis of the molecular features of mutated genes coupled to machine learning has produced a powerful predictive tool for AID targets, which has been experimentally validated. We have also found that Base Excision Repair and Mismatch Repair pathways back-up each other to faithfully repair most of AID-induced lesions. Finally, our data establishes a novel link between AID mutagenic activity and lymphomagenesis.

# Index

I. Introduction 26

1. The immune system: innate and adaptive immunity 29

2. B cell differentiation and antibody diversification 30
2.1. Early B cell differentiation and primary antibody diversification 32

2.2. The Germinal Center reaction and secondary antibody diversification 33

3. Activation induced deaminase 36

3.1. The Neuberger model for AID deamination 37
3.1.1. Uracil-N-Glycosylase in Somatic Hypermutation 39
3.1.2. Mismatch Repair in Somatic Hypermutation 40

3.2. AID and class switch recombination (CSR) 41

3.3. AID and gene conversion 42

3.4. Target specificity of AID 43
3.4.1. AID targeting to immunoglobulin genes 43
3.4.2. AID targeting to non-immunoglobulin loci 45

3.5. Regulation of AID activity 46
3.5.1. Transcriptional regulation of AID 46
3.5.2. Post-transcriptional regulation of AID 47
3.5.3. Post-translational regulation of AID 47
3.5.4. Cell-cycle linked regulation 48

3.6. AID biology beyond B lymphocytes: expression and activity in non-lymphoid tissues 49

4. AID off-targeting and lymphomagenesis 50

II. Objectives 53

III. Methods 57
Mice 59
Isolation of mouse B cells from secondary lymphoid organs 59
DNA capture library 59
DNA capture and sequencing 60
Target enrichment assessment by qRT-PCR 61
Sanger sequencing 61

PCR-Seq to validate machine learning approach **62**
Gene expression profiling by RNA-seq **64**

Computational analysis **64**
a. Pipeline to identify and annotate AID-induced mutations **64**
b. Sequence context of mutated cytosines **66**
c. Gene expression profiling by RNA-Seq **66**
d. Transcription rate analysis (GRO-Seq) **66**
e. RNAP II and SPT5 recruitment **67**
f. Superenhancers analysis **67**
g. MED12 binding and epigenetic marks analysis **67**
h. Convergent transcription analysis (GRO-Seq) **68**
i. Machine learning to predict AID targets **68**
j. Annotation of AID targets **70**

Raw data availability **70**
Code availability **70**
Statistical analysis **70**

IV. Results **72**

1. Development of a custom protocol to detect AID-induced mutations **74**

1.1. Design of a custom RNA bait capture library **75**

1.2. Validation of the target enrichment protocol **76**

1.3. Development of a bioinformatics pipeline to analyze AID mutational activity **78**

2. Identification and characterization of AID targets in Germinal Center B lymphocytes **81**

2.1. Capture-based deep sequencing allows high throughput identification of AID targets **81**

2.2. Analysis of the local specificity of AID **88**

2.3.- Molecular characterization of AID targets **91**
2.3.1. AID targets are highly transcribed **91**
2.3.2. AID targets recruit high levels of RNAP II and SPT5 **92**
2.3.3. AID targets are enriched in marks associated to active enhancers and transcription elongation **94**
2.3.4.- AID targets are regulated by superenhancers and frequently undergo convergent transcription **96**

3. Prediction of AID targets in Germinal Center B lymphocytes **98**

4. Analysis of the role of the Base Excision Repair (BER) and Mismatch Repair (MMR) pathways in the resolution of AID-induced deaminations **101**

5. Analysis of the contribution of AID off-targeting to the development of Germinal Center derived malignancies **104**

V. Discussion **109**

1. Development of a capture-based NGS approach to detect AID-induced mutations **111**

2. Discovery of a large catalogue of AID mutational targets **112**

3. AGCTNT as a novel hotspot for AID activity **113**

4. Molecular characterization of AID targets **114**

5. Machine learning approach to predict AID off-targeting **117**

6. Role of BER and MMR in the resolution of AID-induced deaminations **119**

7. AID off-targeting and lymphomagenesis **120**

8. Concluding remarks and future prospects **121**

VI. Conclusions **123**

VI. Conclusiones **127**

Bibliography **131**

Annex **135**

Publications **167**

# List of figures


| | | |
|---|---|---|
| Figure 1. | Immunoglobulin structure | 31 |
| Figure 2. | B cell development, activation and terminal differentiation | 35 |
| Figure 3. | Neuberger's deamination model | 38 |
| Figure 4. | AID activity during the Germinal Center Reaction | 51 |
| Figure 5. | Groups of genes included in the custom capture library | 75 |
| Figure 6. | Target enrichment protocol allows efficient enrichment of selected genes | 77 |
| Figure 7. | Schematic representation of the custom bioinformatics pipeline developed for the analysis of AID mutations | 79 |
| Figure 8. | Schematic representation of the experimental approach | 82 |
| Figure 9. | High quality sequencing of captured genes | 83 |
| Figure 10. | 291 reproducible targets were detected by high-throughput analysis of AID-induced mutations | 85 |
| Figure 11. | Overlap analysis of the 291 targets discovered in this study and previously published data on AID-induced mutations, translocations and double strand breaks | 87 |
| Figure 12. | Mutation analysis at WRCY/RGYW hotspots in *Ung*$^{-/-}$*Msh2*$^{-/-}$ GC B cells | 89 |
| Figure 13. | AGCTNT is a novel AID mutational hotspot | 90 |
| Figure 14. | AID targets are highly transcribed | 93 |
| Figure 15. | AID targets show high density binding of RNAP II and SPT5 | 94 |
| Figure 16. | AID targets are enriched in marks associated to transcription and transcription elongation | 95 |
| Figure 17. | AID targets are regulated by superenhancers and frequently undergo convergent transcription | 96 |
| Figure 18. | Molecular features of AID targets | 97 |

| | | |
|---|---|---|
| Figure 19. | Molecular features of AID predict mutability | 99 |
| Figure 20. | Experimental validation of the machine-learning model by PCR-Seq | 100 |
| Figure 21. | BER and MMR back up each other to error-free repair AID-induced lesions | 101 |
| Figure 22. | BER and MMR pathways faithfully repair most AID-induced deaminations | 103 |
| Figure 23. | AID targets are recurrently mutated in human lymphomas | 104 |
| Figure 24. | Genes frequently mutated in human DLBCL are targeted by AID | 105 |
| Figure 25. | Mutation profiles of representative DLBCL genes | 107 |

# List of tables

## Materials and methods


| | | |
|---|---|---|
| Table 1. | Summary of mice used in this study | 60 |
| Table 2. | Primers used for qRT-PCR | 61 |
| Table 3. | Primers used for Sanger sequencing | 62 |
| Table 4. | Primers used for PCR-Seq | 63 |
| Table 5. | Parameters, input and output files of our custom Perl software | 65 |
| Table 6. | Software versions used for computational analysis | 69 |


## Results


| | | |
|---|---|---|
| Table 7. | Cycle threshold values in post-enrichment and input fractions | 76 |
| Table 8. | Summary of depth and sequencing parameters of the capture libraries analyzed | 82 |
| Table 9. | Mutation analysis of representative AID targets in *Ung*$^{-/-}$*Msh2*$^{-/-}$ and *Aicda*$^{-/-}$ mice by Sanger sequencing | 84 |
| Table 10. | List of genes selected for machine learning validation | 100 |
| Table 11. | List of the 18 AID targets mutated in repair-proficient germinal center B cells | 103 |
| Table 12. | Mutations found in *Ung*$^{-/-}$*Msh2*$^{-/-}$ mice that have been identified in cohorts of human lymphoma patients | 106 |

# Abbreviations

# Abbreviations

| Abbreviation | Full name |
|---|---|
| 3'RR | 3' Regulatory Region |
| AID | Activation Induced Deaminase |
| APE1 | Apurinic Endonuclease 1 |
| APE2 | Apurinic Endonuclease 2 |
| BCL6 | B-cell lymphoma 6 |
| BCR | B Cell Receptor |
| BER | Base Excision Repair |
| CDR | Complementary Determining Regions |
| ChIP-Seq | Chromatin Immunoprecipitation Sequencing |
| CHK2 | Checkpoint Kinase 2 |
| ConvT | Convergent Transcription |
| CSR | Class Switch Recombination |
| Ct | Cycle threshold |
| DE | Differentially Expressed |
| DIVAC | Diversification Activator |
| DLBCL | Diffuse Large B Cell Lymphoma |
| DSB | Double-Strand Break |
| dsDNA | Double-Stranded DNA |
| DZ | Dark Zone |
| EEF1A | Elongation factor 1-alpha 1 |
| FWR | Framework Regions |
| G4 | G-quadruplex |
| GALT | Gut-Associated Lymphoid Tissue |
| GC | Germinal Center |
| gDNA | Genomic DNA |
| GRO-Seq | Global Run-On Sequencing |
| HIGM2 | Hyper IgM type II |
| HR | Homologous Recombination |
| HSP90 | Heat Shock Protein 90 |
| Ig | Immunoglobulin |
| IGC | Immunoglobulin Gene Conversion |
| IgH | Immunoglobulin Heavy chain / locus |
| IgL | Immunoglobulin Light chain / locus |
| IL4 | Interleukin 4 |
| INDEL | Insertion or Deletion |
| LPS | Lipopolysaccharide |
| LZ | Light Zone |
| MBD4 | Methyl-CpG Binding Domain protein 4 |
| MMR | Mismatch Repair |
| MSH2 | MutS Protein Homolog 2 |
| MSH6 | MutS Protein Homolog 6 |
| NES | Nuclear Export Signal |
| NGS | Next Generation Sequencing |
| NHEJ | Non-Homologous End Joining |

| Abbreviation | Full name |
|---|---|
| NLS | Nuclear Localization Signal |
| PCNA | Proliferating Cell Nuclear Antigen |
| PCR | Polymerase Chain Reaction |
| PCR-Seq | Sequencing of Polymerase Chain Reaction products |
| qRT-PCR | Quantitative Real Time Polymerase Chain Reaction |
| RAG1 | Recombination Gene Enzyme 1 |
| RAG2 | Recombination Gene Enzyme 2 |
| RNAP II | RNA Polymerase II |
| RNA-Seq | RNA Sequencing |
| RSS | Recombination signal sequence |
| SSB | Single-strand break |
| SE | Superenhancer |
| SHM | Somatic Hypermutation |
| SMUG1 | Single-Strand-Selective Monofunctional Uracil-DNA Glycosylase 1 |
| SNP | Single Nucleotide Polymorphism |
| ssDNA | Single-Stranded DNA |
| TC | Translocation |
| TCR | T Cell Receptor |
| TDG | Thymine-DNA glycosylase |
| TH | T helper cell |
| TLS | Translesion Synthesis |
| TNFα | Tumor Necrosis Factor α |
| TP53BP1 | TP53 Binding Protein 1 |
| TSS | Trascriptional Start Site |
| UNG | Uracyl-N-Glycosylase |
| UTR | Untranslated Region |
| XRCC4 | X-ray repair cross complementing 4 |

# I. INTRODUCTION

## 1. The immune system: innate and adaptive immunity

Immunity emerged during evolution to protect organisms against deleterious agents. From the simplest bacterial immune systems, such as CRISPR-Cas or restriction modification systems, to the complex immune system of vertebrates, all immune responses are triggered by the recognition of an exogenous factor as foreign. This recognition activates effector functions in the cell that culminate in the removal of the factor that prompted the response.

Higher vertebrates present two types of defense mechanisms: innate and adaptive immunity. Innate immunity is rapid, non-specific and acts as a first barrier against infection. The innate response is elicited by the recognition of molecular patterns that are commonly present in pathogens and allows their efficient elimination. In addition, innate immunity is fundamental to activate adaptive immunity. The adaptive response is defined by two features, specificity and memory: specificity to distinguish between different, even closely related, pathogens and molecules and respond to them; and immune memory, which results in a faster and more efficient removal of the pathogen upon reinfection. The main players of the adaptive immune response are T and B lymphocytes. T lymphocytes participate in the removal of intra-cellular pathogens and contribute to the activation of other immune cell subsets, while B lymphocytes are specialized to develop humoral responses against extra-cellular pathogens. The activity of both B and T lymphocytes relies on the expression of the antigen receptors, TCR (T Cell Receptor) and BCR (B Cell Receptor), which specifically recognize antigens (proteins or polysaccharides, generally) present in pathogens. Each T or B lymphocyte expresses a unique receptor which is specific to a single antigen, such that the collection of T or B cells in an organism provide an inmensely diverse TCR and BCR repertoire to proficiently combat virtually any infection.

## 2. B cell differentiation and antibody diversification

The BCR is a multiprotein complex composed of two modules: a membrane bound immunoglobulin (Ig) molecule, which is responsible for antigen binding, and an Igα/β heterodimer (CD79a/CD79b), endowed with signalling function. Igs, also called antibodies in their secreted form, are composed of four polypeptides: two identical heavy chains (IgH) and two identical kappa (κ) or lambda (λ) light chains (IgL) (Figure 1 A,B). Both heavy and light chains are formed by an N-terminal variable region and a C-terminal constant region. The variable region is responsible for antigen recognition, and can be further subdivided into hypervariable regions (or complementarity determining regions, CDRs) and framework regions (FWRs). There are 3 CDRs in the light and heavy chains that determine the specificity of the antibody, and 4 interspersed FWRs that serve as a scaffold to favor the contact between the CDRs and the antigen. On the other hand, the constant region defines the isotype of the antibody and its effector funcion, i.e. the mechanism by which the antigen will be removed.

The BCR drives the differentiation, maintenance and activation of B lymphocytes and is therefore a key molecule for humoral immunity. During early development in the bone marrow, the assembly of a pre-BCR is an absolute requirement for B cells to survive (Kitamura et al., 1991); mature B cell persistence in the periphery is dependent on BCR signalling (Lam et al., 1997); and upon antigen encounter in the secondary lymphoid organs, the BCR is central for B cell activation, clonal selection of B cells and their terminal differentiation to plasmatic and memory cells (reviewed in Victora and Nussenzweig, 2012). Thus, B cells are driven by BCR signals to make vital cell-fate decisions at several stages of their development.

Antibodies can specifically bind and eliminate a practically unlimited number of foreign antigens. It is estimated that mammals can generate in the order of $10^{11}$ different antibodies (Chapter 1, Tasuku Honjo, Michael Reth, Andreas Radbruch, Frederick Alt, 2015). This hugely diverse repertoire originates from active somatic gene editing taking place at the Ig loci. Three different genes encode, respectively, the IgH chains and the Igκ and Igλ light chains. Ig genes have an analogous organization in all mammals, although there are differences in their chromosomal locations and the number and sequence of different gene segments in each locus may vary. The IgH locus is composed of multiple gene segments termed $V_H$ (variable), $J_H$ (joint), $D_H$ (diversity) and $C_H$ (constant). In mice, there are around 100 functional $V_H$, 10-15 $D_H$ and 4 $J_H$ segments encoding the IgH variable region; and 8 $C_H$ segments

encoding the constant region of the different IgH isotypes (Cμ, Cδ, Cγ1, Cγ2a, Cγ2b, Cγ3, Cε and Cα) (Retter, J Imm, 2007) (Figure 1C). On the other hand, the human IgH locus contains 123 Vh, 26 Dh and 6 Jh segments; and there are 11 Ch segments: 9 coding (Cμ, Cδ, Cγ3, Cγ1, Cα1, Cγ2, Cγ4, Cε1 and Cα2) and 2 pseudogenes (Matsuda et al., 1998). The IgL κ and IgL λ loci are composed by $V_L$, $J_L$ and $C_L$ segments and also exhibit some differences in number beween human and mice. As we will explain below, diversification of antibody genes generally occurs in two different stages: primary diversification in the bone marrow and secondary diversification in germinal centers.

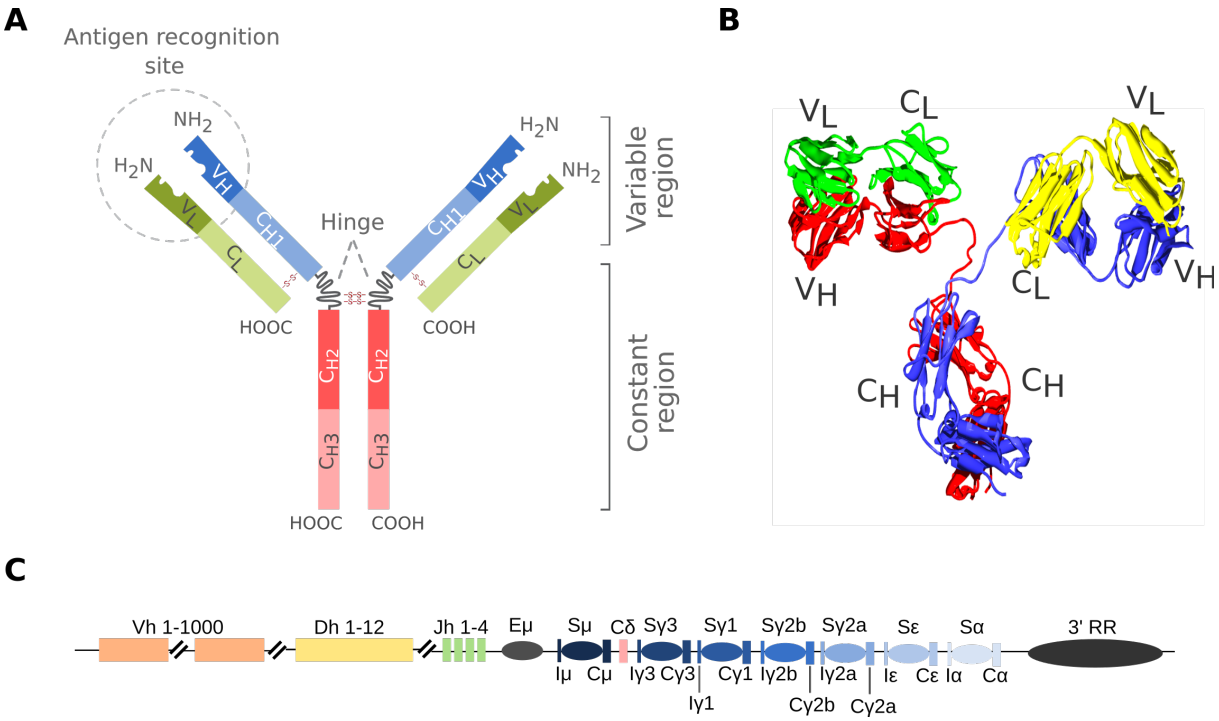



Figure 1 | **Immunoglobulin structure. (A)** Schematic and 3D **(B)** representation of the immunoglobulin molecule. The antibody molecule is a dimer of heterodimers (one heavy and one light chain) connected by disulfide bonds. The antigen binding site is formed by the heavy and light chain variable regions ($V_H$, $V_L$); the constant regions form the effector arm. **(C)** Schematic representation of the germline murine immunoglobulin heavy chain (IgH) locus (not to scale). Variable (V), Diversity (D) and Joint (J) segments undergo V(D)J recombination to produce a unique rearrangement per cell. The rearranged VDJ region will be subject to somatic hypermutation (SHM) during the germinal center reaction. There are eight constant (C) genes specifying different antibody isotypes. Except for IgD -which can be expressed by alternative splicing-, each constant region is preceded by a switch (S) and intronic (I) region. Upon activation, B cells can switch from IgM/D to any of the other six isotypes.

## 2.1. Early B cell differentiation and primary antibody diversification

B cells generate in the bone marrow from hematopoietic stem cells (HSCs) through a tightly regulated process (Figure 2) that concludes with the expression of a functional BCR. The driving force of B cell development is the rearrangement of IgH and IgL loci, which constitutes the primary diversification of antibodies (see below). The rearrangement of the IgH locus by V(D)J recombination is initiated at the proB stage of B cell differentiation in the bone marrow. Successful IgH rearrangement enables the expression of the IgH chain, which pairs with the invariant Surrogate Light Chains VpreB and λ5 and the signal-transducing subunits Igα/β to form the pre-BCR in preB cells (Karasuyama et al., 1990; Tsubata and Reth, 1990). Signals from the pre-BCR trigger an intense proliferative stage of preB cells which gives raise to the largest expansion of differentiating cells. This is followed by the rearrangement of the IgL locus by V(D)J recombination. B cells that have successfully rearranged both the IgH and IgL loci express a functional BCR and are called immature B cells. At this stage, there is a tolerance checkpoint by which immature B cells expressing an autoreactive BCR are subject to deletion (apoptosis), anergy (unresponsiveness to antigen), or receptor editing (further recombination of the IgL locus to replace the BCR by a non self-reacting version) (reviewed in Pelanda and Torres, 2012).

Primary antibody diversification by V(D)J recombination is a site-specific reaction that takes place in an antigen-independent fashion in the bone marrow. During V(D)J recombination, a random combination of $V_H$, $D_H$ and $J_H$ segments is assembled in each individual differentiating B cell, followed by an analogous reaction at the Igκ or Igλ locus. V(D)J recombination is initiated by the Recombination Activating Gene enzymes (RAG1/2), which form a tetrameric complex recognizing the recombination signal sequences (RSS) flanking each V, D and J segment. RAG1 and RAG2 create a loop between two RSS, bring them together and generate a DNA double strand break (DSB) at each RSS. These DSBs are resolved by Non-Homologolous End-Joining (NHEJ) resulting in the assembly of V, D and J segments into the variable domain of IgH (reviewed in Arya and Bassing, 2017; Bassing et al., 2002). The random choice of a single V, D and J segment to be recombined is responsible for the large array of different receptors produced by V(D)J recombination. In addition, the imprecise joining of V, D and J segments implies the excision, duplication and insertion of nucleotides and further increases the variability of the rearranged segments (Max et al., 1979; Sakano et al., 1979; Seidman et al., 1979). Light chains also undergo an analogous process of recombination of their V and J gene segments once the IgH locus has been successfully rearranged. Due to the random nature of V(D)J recombination,

only a fraction of the rearrangements is functional, i.e. contain a productive VDJ (IgH locus) or VJ (IgL loci) exon that can be assembled into a surface-expressed BCR. In the case of the IgH locus, if one allele is productively rearranged, the recombination of the other allele is prevented. In the IgL loci, the productive rearrangement of Igκ precludes Igλ rearrangement. Thus, Igλ exclusively undergoes V(D)J recombination if Igκ is non-productively rearranged or if a self-reactive Igκ light chain is subject to receptor editing. This process is called allelic exclusion and ensures that B lymphocytes are monospecific and express only one functional Ig per cell (reviewed in Vettermann and Schlissel, 2010).

V(D)J recombination, together with the imprecise joining by NHEJ, generates a primary repertoire of about $10^5$-$10^6$ different antibody specifities (Bassing et al., 2002; Di Noia and Neuberger, 2007).

### 2.2. The Germinal Center reaction and secondary antibody diversification

Immature B cells that express a functional IgM/D BCR on their membrane exit the bone marrow and migrate to the periphery, where they can recirculate and populate secondary lymphoid organs. There, they can encounter and respond to foreign antigens. Upon antigen binding by the BCR, a mature B lymphocyte can engage in the Germinal Center (GC) reaction, an essential event for the generation of high affinity antibodies with diverse effector functions. This secondary antibody diversification takes place through Somatic Hypermutation (SHM) and Class Switch Recombination (CSR) (Figure 2). SHM introduces point mutations in the V(D)J rearranged variable region of the Ig loci (reviewed in Di Noia and Neuberger, 2007; Methot and Di Noia, 2017), while CSR is a region specific recombination reaction that replaces the primary IgM isotype with IgG, IgA or IgE isotypes. GCs originate in secondary lymphoid organs which are structured in follicles. Follicles are mostly composed by naïve B cells surrounded by T cells. The GC reaction begins with the acquisition of the antigen by naïve B cells, which migrate to the T cell rich zone of the follicle and are fully activated by co-stimulatory signals from specificity-matching $CD4^+$ helper T cells ($T_H$). This interaction is mainly mediated by CD40 ligand, expressed by the $T_H$ cell, which engages its receptor CD40 in the antigen-stimulated B cell, and triggers an intense proliferation and clonal expansion of the B lymphocyte that initially recognized the antigen. The mature GC can be divided into two compartments: the dark zone (DZ) and the light zone (LZ). The DZ consists of a cluster of highly proliferative B cells called centroblasts that are actively introducing mutations at their Ig variable regions by SHM. These mutations can modify the affinity of the antibody. In the LZ, centrocytes are selected by the affinity of their BCRs for the cognate antigen in

the context of T follicular helper and follicular dendritic cell help. As a result, B cells whose BCRs have gained affinity for the antigen receive anti-apoptotic signals, while B cells bearing a BCR with decreased affinity undergo apopotosis (Figure 2). These proliferation/SHM/selection events can be repeated in iterative cycles and thus result in the increase of the antibody affinity for the antigen that initiated the response, a process called affinity maturation. In addition, a subset of centrocytes can undergo CSR in the LZ. In CSR a recombination occurs between the highly repetitive switch regions of the IgM/D constant chain (Cμ/Cδ) and a downstream switch region (Cα, Cε or Cγ) culminating with the replacement of an IgM/D isotype by either an IgA, IgE or IgG isotype. Thus, CSR modulates the effector function of the antibody allowing a single variable region (i.e. a single specificity) to have several effector capabilites (reviewed in Stavnezer and Schrader, 2014; Stavnezer et al., 2008; Xu et al., 2012). As a result of the GC reaction, B cell clones with high affinity BCRs terminally differentiate into plasmatic or memory B cells (reviewed in Mesin et al., 2016) (Figure 2).

Besides SHM and CSR, birds and some mammals (rabits, cows, sheep and pigs) can diversify their immunoglobulin loci by immunoglobulin gene conversion (IGC), a process of homologous recombination that replaces pieces of the VDJ region with portions of 5'-encoded pseudo-genes (reviewed in Tang and Martin, 2007). Although mechanistically different, SHM, CSR and IGC are triggered by a single enzyme, the activation induced deaminase (AID).

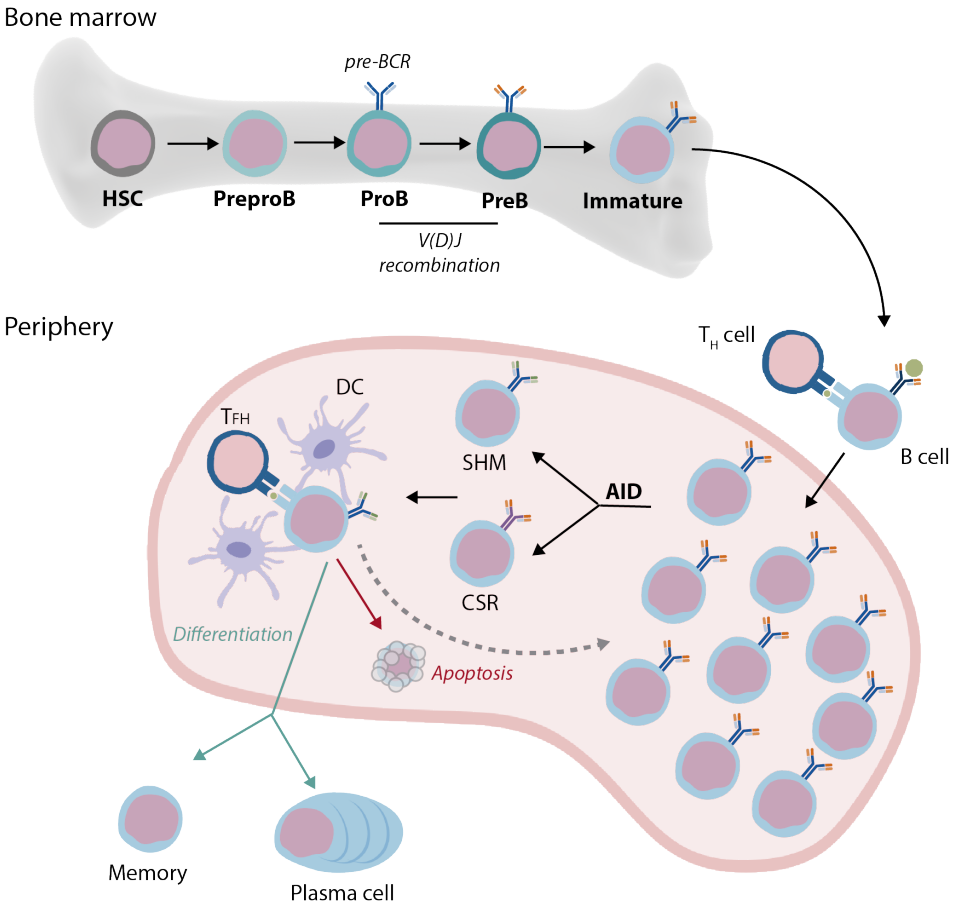



Figure 2 | **B cell development, activation and terminal differentiation.** B cells originate from bone marrow Hematopoietic Stem Cells (HSC). The most distinctive event of B cell differentiation in the bone marrow is V(D)J recombination. This process takes place during the proB and preB stages through recombination of the variable regions of the immunoglobulin (Ig) genes. As a result, a primary repertoire of highly diverse BCRs is generated, with single B cell clones carrying specific rearrangements. Immature B cells expressing a functional rearranged BCR migrate to the secondary lymph organs (periphery) where they can enter the Germinal Center (GC) reaction upon antigen encounter. Interaction with T helper cells ($T_H$) further stimulates the B cell, leading to clonal expansion and initiating the GC reaction. B cell activation increases AID expression which triggers the remodeling of Ig genes by Somatic Hypermutation (SHM) and Class Switch Recombination (CSR). B cell clones that increase the affinity of their BCRs for the cognate antigen will receive stimulatory signals from T follicular helper ($T_{FH}$) and dendritic cells (DC). Depending on the strength of BCR signalling these clones will either undergo a new round of SHM (low strength; dashed grey line) or differentiate into memory (moderate strength) or plasma (high strength) cells. On the other hand, B cell clones that decrease the affinity of the BCR for the antigen will enter apoptosis (red line).

## 3. Activation induced deaminase

AID was first identified in a substractive library screening as an upregulated mRNA in CH12F3 cells stimulated to switch *in vitro* (Muramatsu et al., 1999). Later on, genetic experiments showed that AID is absolutely required for SHM and CSR in mice and humans (Muramatsu et al., 2000; Revy et al., 2000) and for gene conversion in chicken (Arakawa et al., 2002; Harris et al., 2002). Indeed, AID deficiency in humans leads to an immunodeficiency called Hyper IgM type II (HIGM2) syndrome (Revy et al., 2000). The genetic analysis of these patients, together with new cases later identified (Caratão et al., 2013; Durandy, 2009; Imai et al., 2005; Meyers et al., 2011; Quartier et al., 2004) has allowed a better understanding of the structure-function organization of AID protein. AID is a small, 198aa protein with a molecular mass of 24KDa. It is very well conserved between mice and humans, with an aminoacid sequence homology of ~92%. AID comprises a nuclear localization signal (NLS) in the N-terminal region (residues 1-30) and a nuclear export signal (NES) in the C-terminal region (residues 183-198) that control AID shuttling to and from the nucleus (Ito et al., 2004; explained in more detail in chapter 3.4); a catalytic CMP/dCMP-type deaminase domain placed in the central region of the protein (residues 23-129); and an APOBEC-like domain (residues 108-181). Up to date, 40 different AID mutations have been identified in HIGM2 patients (Caratão et al., 2013). AID mutants in the N-terminus have been shown capable of CSR but not SHM (Shinkura et al., 2004), whereas AID mutants in the C-terminus can initiate SHM but not CSR (Barreto et al., 2003; Ta et al., 2003). Pure AID protein has not been crystalized to date, but two different labs have resolved the crystal structure of AID soluble variants, which provided some hints to AID biology: Goodman's lab identified a "specificity loop" accounting for AID preference to deaminate WRC (W = A/T; R = A/G) motifs (Pham et al., 2016, 2017); Wu's lab reported a "bifurcated substrate-binding surface" which explains the preference of AID to act on G4 structures by simultaneous capture of two adjacent single-stranded DNA strands (Qiao et al., 2017).

Due to its homology to APOBEC1, an RNA editing enzyme, AID was first proposed to act on RNA. However, this hypothesis has been abandoned in light of genetic and biochemical evidence proving that AID deaminates cytidine residues on DNA (Chaudhuri et al., 2003; Maul et al., 2011; Petersen-Mahrt et al., 2002; Pham et al., 2003; Ramiro et al., 2003); and targets ssDNA but not dsDNA, RNA or DNA-RNA hybrids (Bransteitter et al., 2003; Chaudhuri et al., 2003; Dickerson et al., 2003; Pham et al., 2003).

In this chapter we will introduce the molecular mechanisms underlying these two key reactions for secondary antibody diversification.

### 3.1. The Neuberger model for AID deamination

In 2002, Neuberger's lab proposed a unifying model to explain how DNA deamination by AID could initiate both SHM and CSR (Petersen-Mahrt et al., 2002) building on a previous speculative model for SHM by Scharff and colleagues (Poltoratsky et al., 2000) and on the idea that all diversification programs might be triggered by a common type of DNA lesion (Ehrenstein and Neuberger, 1999; Maizels, 1995; Sale et al., 2001; Weill and Reynaud, 1996). Cytosine (C) deamination by AID would turn it into a uracil (U) and generate a U:G mismatch in the DNA. The U:G mismatch could be directly replicated so that an adenine residue (A) would be introduced opposed to the uracil, leading to a C→T transition mutation in one of the daughter cells (or a G→A transition if the deamination ocurred in the non-transcribed DNA strand) (Phase Ia of SHM; Petersen-Mahrt et al., 2002; Rada et al., 1998; Wiesendanger et al., 2000; Figure 3). Alternatively, the U:G mismatch could be recognized by the DNA repair machinery, namely the Base Excision Repair (BER) or Mismatch Repair (MMR) pathways. Uracil-N-glycosylase (UNG) could detect the U:G mismatch and remove the uracil, leaving an abasic site that would be replicated through by translesion synthesis (TLS) polymerases enabling the generation of both C→T transitions and C→G or C→A transversions (Phase Ib of SHM; Petersen-Mahrt et al., 2002; Rada et al., 1998; Wiesendanger et al., 2000; Figure 3). Alternatively, the recognition of the U:G mismatch by the MMR pathway via MutS Protein Homolog 2/6 (MSH2/6) dimer would account for mutations at A/T nucleotides (Phase II of SHM; Rada et al., 1998; Wiesendanger et al., 2000; Figure 3). These molecular pathways could explain SHM triggered by a C→U deamination. In addition, abasic sites could be substrates for endonucleases and generate single strand breaks close to the mismatch. Two proximal abasic sites in opposite strands could generate a double strand break (DSB) and thus initiate CSR (Figure 3; further explained in chapter 3.2).

This deamination model has been extensively validated through genetic evidence, which will be explained below.

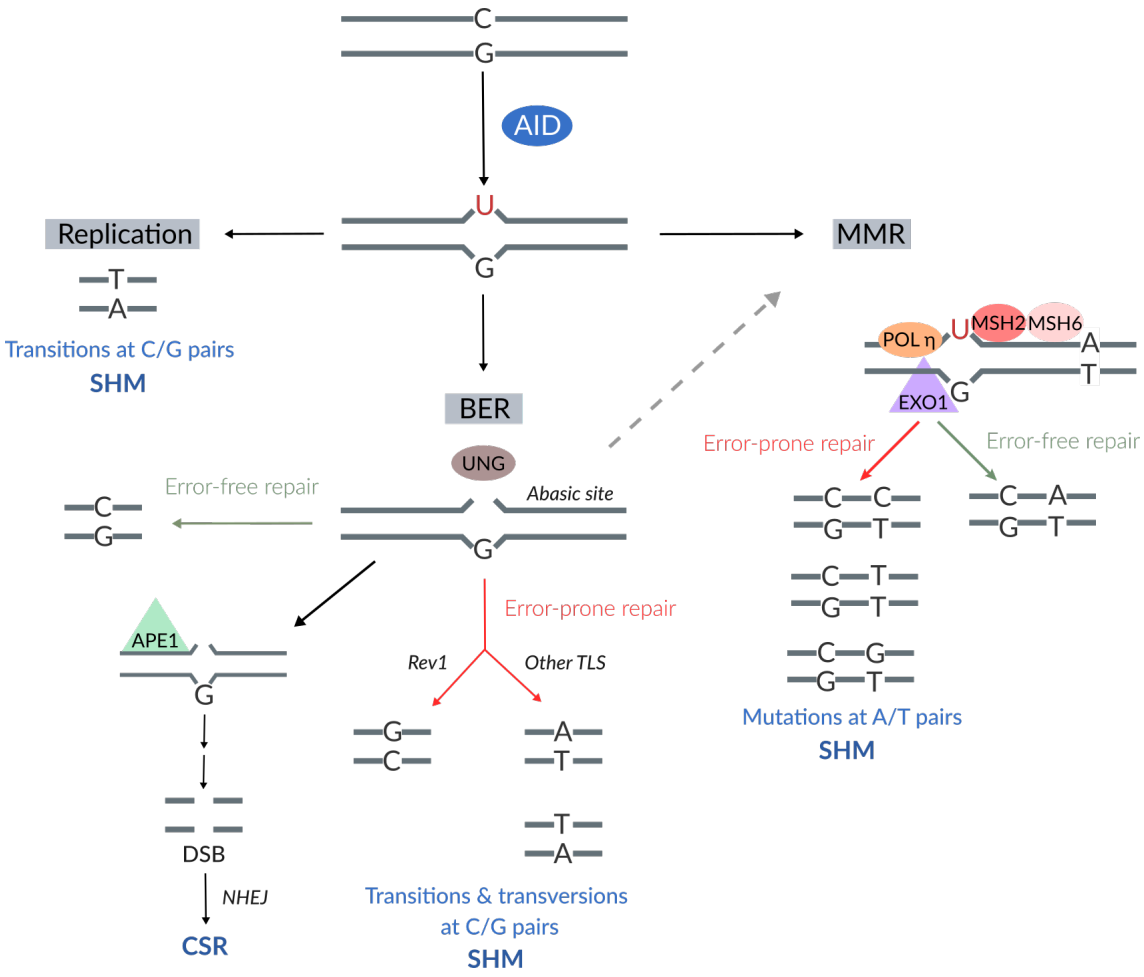



Figure 3 | **Neuberger's deamination model.** Schematic representation of the DNA deamination model. Cytosine deamination by AID generates a U:G mismatch that can be either directly replicated over or further processed by the Base Excision Repair (BER) and/or Mismatch Repair (MMR) pathways to give rise to point mutations (SHM) or double strand breaks -DSB- (CSR). Direct replication of the unrepaired uracil will lead to a C→T transition mutation in one of the daughter cells. Furthermore, the U:G mismatch can be recognized by the MSH2/6 heterodimer which recruits exonuclease I (EXO I) and excises the mismatch and a stretch of the surrounding DNA. The gap will be filled by Pol η, giving rise to mutations at A/T pairs adjacent to the deaminated cytosine. The U:G mismatch can also be recognized by UNG, which generates an abasic site by uracil excision. On one hand, this abasic site can be processed by APE1/2 and lead to DNA double strand breaks that can be sensed by the Non-Homologous End Joining pathway (NHEJ) and trigger CSR. On the other hand, translesion polymerases (TLS), such as *Rev1*, can replicate the abasic site producing transition and transversion mutations at C:G pairs. *See text for further details.*

### 3.1.1. Uracil-N-Glycosylase in Somatic Hypermutation

Evidence for the role of UNG in antibody diversification came from three genetic studies from the Neuberger lab performed in bacteria (Petersen-Mahrt et al., 2002), chicken DT40 cells (Noia and Neuberger, 2002) and mice (Rada et al., 2002). First, processing of AID-induced uracils by UNG was shown in AID-expressing *E. coli* (Petersen-Mahrt et al., 2002). Later, chemical inhibition of UNG in DT40 B cells revealed a shift in the mutational pattern of the Ig variable region to a severe increase in transitions and an equivalent reduction in transversions at C:G pairs (Noia and Neuberger, 2002). Furthermore, a mouse model deficient for UNG showed a similar mutation pattern at C:G, while mutations at A:T pairs remained practically unaltered in the absence of UNG (Rada et al., 2002). Remarkably, this phenotype has been also observed in humans bearing inactivating mutations in UNG (Imai et al., 2003). These studies showed that: 1) UNG is critical for the generation of C/G transversions; 2) in its absence the U:G mismatches can be directly replicated to give raise to C/G transitions or processed by the MMR, main responsible for the introduction of A/T mutations (see below). Multiple evidence has further reinforced the role of UNG in antibody diversification (Krijger et al., 2009; Maul et al., 2011; Ranjit et al., 2011; Saribasak et al., 2006; Sharbeen et al., 2012; Zahn et al., 2013; reviewed in Methot and Di Noia, 2017).

In general, misincorporated nucleotides are recognized by the BER pathway and polymerase β mediates high-fidelity canonical repair of the abasic sites generated by UNG. However, in the context of AID-induced deaminations at Ig loci, these mismatches can be either faithfully repaired by Pol β (Wu and Stavnezer, 2007) or replicated in an error-prone fashion, leading to the mutation pattern explained above. It has been proposed that this mutagenic resolution could be carried out by TLS polymerases. TLS polymerases are specialized in the insertion of bases opposite DNA lesions that stall the replication fork. Due to their large, relaxed catalytic site they are more prone to error than conventional replicative polymerases. There is solid evidence of a role for REV1 in the generation of transversion mutations dowstream UNG (Jansen et al., 2006): REV1 is very efficient introducing deoxycytidine residues opposed to an abasic site (Nelson et al., 1996) and its deficiency in mice leads to a complete loss of C→G transversions in the coding (non-transcribed) strand and a reduction in the non-coding strand (Jansen et al., 2006). This suggests that other TLS polymerases could be involved in the generation of C→G transversions (Reviewed in Weill and Reynaud, 2008). Additionally, the implication of other bypass polymerases, such as polymerase ζ (Reviewed in Seki et al., 2005; Weill and Reynaud, 2008),

polymerase θ (Di Noia and Neuberger, 2007), polymerase ι (Faili et al., 2002; Maul et al., 2016) or polymerase μ (Domínguez et al., 2000; Ruiz et al., 2001, 2004), cannot be ruled out. While the mechanism responsible for the switch from high-fidelty to error-prone polymerases remains unkown, recent studies point to monoubiquitination of PCNA as a mediator (Arakawa, PloS Biol, 2006).

Besides UNG, there are three more DNA glycosylases able to escind uracils in vertebrates: SMUG1, TDG and MBD4. However, these seem to play little if any natural role in antibody diversification (Di Noia et al., 2006; Rada et al., 2004; Visnes et al., 2009).

### 3.1.2. Mismatch Repair in Somatic Hypermutation

In addition to mutations at C:G pairs, which are explained by either direct replication of the U:G mismatch or error-prone BER, SHM also generates mutations at A/T nucleotides. This outcome mostly originates from the action of the MMR pathway, as demonstrated by the severe reduction in A/T mutations found in mice deficient for *Msh2* or *Msh6* (Phung et al., 1998; Rada et al., 1998; Wiesendanger et al., 2000). The residual A/T mutations in this models suggest that UNG could be also contributing to this kind of mutations (Delbos et al., 2007; Rada et al., 2004). Indeed, mice deficient for UNG and MSH2 present an Ig mutation pattern completely devoid of C/G transversions and A/T mutations and exclusively composed of C/G transitions (Rada et al., 2004). This indicates, first, that in the absence of UNG and MSH2 U:G mismatches are directly replicated; and second, that uracil excision by UNG provides a backup for the second phase of SHM (Rada et al., 2004).

The MSH2/6 complex can recognize and excise the U:G mismatch, together with a stretch of the surrounding DNA, by exonuclease I (Bardwell et al., 2004). Monoubiquitinated PCNA then recruits TLS polymerases to fill the gap, which leads to transitions and transversions at A/T nucleotides. Evidence points to polymerase η (pol η) as the main contributor to this error-prone synthesis (Delbos et al., 2005, 2007; Wilson et al., 2005): pol η functionally interacts with the MSH2/MSH6 complex (Wilson et al., 2005); its deficiency produces a C/G biased mutation pattern (85% C/G, 15% A/T mutations) (Delbos et al., 2005); and the combined absence of MSH2 and pol η in mice completely abrogates mutations in A:T pairs (Delbos et al., 2007). Only in the complete absence of pol η, polymerase κ can exert a backup function and contribute to A/T mutations (Faili et al., 2009).

A long standing question in the field is understanding why BER and MMR, which are usually involved in the faithful repair of DNA lesions, contribute to introduce mutations in the context of AID-induced U:G mismatches (Krokan et al., 2014). In this regard, previous work from our lab showed that the choice between an error-free or error-prone outcome by UNG is influenced by the local sequence context in the IgH locus (Pérez-Durán et al., 2012). In addition, Liu and colleagues proposed that the choice between error-prone and error-free repair by BER and MMR is locus specific (Liu et al., 2008). In this work, we will explore the biology of BER and MMR pathways to try to add to the understanding of this relevant question.

### 3.2. AID and class switch recombination (CSR)

CSR is a recombination reaction that takes place between switch regions, highly repetitive sequences located upstream of all the IgH constant genes, with the exception of Cδ. B cells emerging from the bone marrow express IgM or IgD by alternative splicing. CSR replaces the Cμ/Cδ IgH constant region by a downstream (Cα, Cε or Cγ) constant region and involves the generation of DNA double strand breaks (DSBs), which act as a substrate for recombination. CSR is triggered by the deamination of cytosine residues in the IgH switch regions by AID, which generates a U:G mismatch. Uracil removal by UNG is a crucial event for CSR, since mice deficient for UNG (Rada et al., 2002) and humans bearing inactivating mutations in UNG (Imai et al., 2003) have drastically impaired CSR. A single strand DNA break in the abasic site is then introduced by APE1/2 endonucleases (Guikema et al., 2007). If two ssDNA breaks occur close enough on opposite strands of the DNA, they can give raise to a DSB. The formation of DSBs in Sμ (donor) and a downstream switch region -Sα, Sε or Sγ- (acceptor) enables recombination by Non-Homologous End Joining (NHEJ) and results in the replacement of IgH Cμ by a different IgH C segment, thus completing CSR. This process requires the coordinated action of a number of different proteins: MRE11-RAD50-NBS1 (MRN complex) senses the DSB; ATM then binds MRN via NBS1 and phosphorylates NBS1, TP53BP1, P53, CHK2, and H2AX, which causes the accumulation of other repair proteins and results in repair by NHEJ. In addition, Ku70-Ku80 can also bind to DNA and recruit enzymes that effect the recombination, such as XRCC4-ligase IV complex, which can fuse two DSBs to complete the reaction (reviewed in Stavnezer and Schrader, 2014; Stavnezer et al., 2008; Xu et al., 2012).

### 3.3. AID and gene conversion

In contrast to mice and humans, chickens have a single copy of the V and J segments at IgH and IgL loci, and nearly identical D segments in IgH, and they diversify their antibodies primarily by gene conversion. Gene conversion is a diversification mechanism that generates templated changes in the sequence of the IgV region making use of the 25 pseudo variable gene segments (ΨV) upstream the $V_L$ or $V_H$ regions as templates to replace homologous sequences of functionally rearranged VJ and VDJ segments. Gene conversion is initiated by deamination of cytosines into uracils by AID (Arakawa et al., 2002; Harris et al., 2002). Uracils are then excised by UNG (Di Noia JM and Neuberger MS, 2004; Noia and Neuberger, 2002) and the resulting abasic site can give rise to SSBs or DSBs that are resolved by the Homologous Recombination (HR) machinery. Thus, either abasic sites or SSB/DSBs can trigger IGC as long as there are 3' free ends to initiate homology search and prime DNA synthesis. In the presence of donor ΨV sequences, the 3' free end is used for homology search and invasion of the ΨV region by the V region forming a loop. Extension through the loop copies the template ΨV sequence into the V region and the structure is then resolved by the HR machinery, generating diversity in the IgV segments (reviewed in Tang and Martin, 2007). Therefore, the upstream pseudogene only serves as a template and remains unaltered, allowing further rounds of gene conversion. Chickens use gene conversion at different stages of B-cell development. After V(D)J recombination in the yolk sac, B cell precursors colonize the Bursa of Fabricius, and diversify their V regions through several rounds of gene conversion to generate a pool of naïve B cells with a diverse range of receptor specificities. In the secondary lymphoid organs, chicken GC B cells diversify their antibody repertoire both by gene conversion and SHM (Arakawa et al., 1998; reviewed in Tang and Martin, 2007).

## 3.4. Target specificity of AID

### 3.4.1. AID targeting to immunoglobulin genes

One of the most relevant, yet unresolved, questions about AID biology is to understand what are the molecular players that drive it to the Ig genes and what defines the boundaries of mutagenesis within the Ig loci to the variable and switch regions.

Long before the discovery of AID, mutation analysis of collections of human V(D)J rearranged sequences revealed that SHM preferentially focuses in small, degenerate motifs called *hotspots* (Dörner et al., 1998; Rogozin and Diaz, 2004; Rogozin and Kolchanov, 1992). These hotspots were initially defined as RGYW and their reverse complement WRCY (Dörner et al., 1998; Rogozin and Kolchanov, 1992) and further refined to WRCH/DGYW (Rogozin and Diaz, 2004), where W = A/T; R = A/G; Y = C/T, H = A/C/T; D = A/G/T. Biochemical evidence proved that AID preferentially mutates cytosines lying within WRC motifs (Bransteitter et al., 2003; Pham et al., 2003); and *in vivo* analysis of SHM has also demonstrated a preference for AID to mutate WRCY motifs (Pérez-Durán et al., 2012; Yeap et al., 2015; Zarrin et al., 2004). However, it seems obvious that the low complexity of these sequence motifs makes them highly unlikely to be the sole contributors to AID target specificity.

Transcription is an absolute requirement for AID activity on Ig genes. The first hint suggesting an important role for transcription in SHM came from the observation that SHM of an Igκ transgene was dependent on the presence of the 3' kappa transcriptional enhancer (Betz et al., 1994). Later on, the Storb lab showed in a transgenic mouse model that a kappa promoter placed upstream of an Ig Cκ region was enough to trigger SHM in this normally unmutated sequence (Peters and Storb, 1996). Further work revealed a direct correlation between transcription levels and hypermutation (Fukita et al., 1998). Later on, biochemical studies in synthetic DNA substrates and *E. coli* systems proved that AID targets ssDNA that is exposed during transcription (Chaudhuri et al., 2003; Pham et al., 2003; Ramiro et al., 2003). The link between expression and AID activity has since been reinforced by multiple evidences (reviewed in Storb, 2014). However, transcription alone is not enough to explain AID targeting, since genes known to be highly transcribed in B cells do not accumulate mutations (reviewed in Kenter et al., 2016).

A role for secondary structures in AID targeting has also been proposed. Transcription through the low complexity, GC rich IgH switch regions can generate R-loops and G-quadruplexes. R-loops are structures that form during transcription when the nascent RNA hybridizes to the template DNA, thus producing an RNA-DNA hybrid that displaces the non-template ssDNA. R-loops can also arise during replication, where RNA-DNA hybrids prime DNA synthesis (reviewed in Santos-Pereira and Aguilera, 2015). G-quadruplexes (G4) are four-stranded structures formed by the stacking of G-quartets, which arise from intra-strand base pairing of guanines by hydrogen bonding into planar structures. Structural analysis of a soluble form of AID showed that it preferentially binds "structured substrates" such as G-quadruplex and branched DNA (Qiao et al., 2017), and a recent work reported that spliced intronic IgH transcripts that form G4 RNA structures can bind AID and contribute to its targeting to switch regions for mutation (Zheng et al., 2015). Regarding R-loops, the classic view posits that they favor SHM by generating abundant ssDNA substrates for AID activity. However, this hypothesis has not been supported by *in vivo* evidence and the contribution of R-loops to SHM remains speculative (Parsa et al., 2012; Romanello et al., 2016; Ronai et al., 2007; reviewed in Pavri et al., 2017). Conversely, there is solid evidence that R-loops occur in switch regions (Yu et al., 2003) and are required for efficient CSR (Shinkura, Nat Imm, 2003), with R-loop frequency directly correlating with CSR efficiency (Zhang et al., 2014).

Regulatory sequences at the Ig locus also play a role in AID targeting beyond their effect on transcription. A first example was provided by the finding that an intronic 3' enhancer was required for SHM of an Igκ transgene (Betz et al., 1994). More recently, an enhancer called DIVAC (from Diversification Activator) (Blagodatski et al., 2009) has been identified as crucial for SHM in chicken DT40 cells (Blagodatski et al., 2009) and found conserved in mammalian Ig loci (Buerstedde et al., 2014). In addition, the IgH 3' regulatory region (3'RR), long known to be required for CSR (reviewed in Pinaud et al., 2011), has also been shown necessary for SHM (Rouaud et al., 2013). The mechanisms involved in AID targeting by these regulatory elements remain unknown. Given the fact that the DIVAC contains many transcription factor binding sites and the 3'RR spans ~30Kb, it seems very likely that they act as a docking platform to recruit other cofactors that in turn contribute to the recruitment of AID, such that AID would target the Ig loci as part of a multiprotein complex. Back in 1996, a SHM model was proposed by the Storb lab in which a "mutator factor" (AID) travelled together with the transcription machinery (Peters and Storb, 1996; Storb et al., 1998). On these lines, several studies indicate that AID specifically interacts with components of the transcription machinery, such as RNA

polymerase II (RNAP II), SPT5 or the RNA exosome, among others (Basu et al., 2011; Nambu et al., 2003; Nowak et al., 2011; Pavri et al., 2010; Xu et al., 2010). Genome-wide profiling of RNAP II and SPT5 in B cells led to the hypothesis that AID targets DNA at places where RNAP II is stalled, with SPT5 as the main responsible for recruiting AID to halted RNAP II (Pavri et al., 2010). This idea was reinforced by the finding that topoisomerase I inhibition increases SHM by preventing DNA unwinding ahead of the transcription fork and contributing to RNAP II pausing (Maul et al., 2015). Furthermore, the RNA exosome also interacts with RNAP II via SPT5/6 and has been implicated in the targeting of AID to both strands of transcribed dsDNA (Basu et al., 2011). Finally, AID activity at the three Ig loci has been reported to occur within superenhancer (SE) domains interconnected by long-range interactions (Qian et al., 2014). Together, these findings suggest that the transcription machinery, RNAP II stalling, the recruitment of the RNA exosome and transcription regulatory elements favor AID targeting to the Ig loci.

A recent work approached the notion that the Ig genes provide a privileged context for SHM. Yeap and colleagues used a mouse model in which a "passenger" sequence was introduced in place of the endogenous IgV exon on one IgH allele, while the other allele was normal to ensure B cell maturation and GC formation (Yeap et al., 2015). They found that both alleles were equally mutated by AID, indicating that gene location rather than Ig primary sequences plays a major role in SHM.

### 3.4.2. AID targeting to non-immunoglobulin loci

Despite primarily targeting Ig loci, AID also targets other regions in the genome, although at much lower frequencies. The first evidence for AID off-targeting was provided by the finding that *Bcl6* was mutated in human memory, but not naive, B cells (Shen et al., 1998) and in human lymphoma samples (Pasqualucci et al., 1998), with *Bcl6* consistently displaying the mutation hallmarks of Ig SHM. Following this discovery, additional genes such as *Fas* (Müschen et al., 2000), *Cd79a/b* (Gordon et al., 2003), *Myc*, *Pim1*, *Pax5, RhoH* (Pasqualucci et al., 2001) and many others (Liu et al., 2008) were identified to be mutated by AID. *In vitro* transgene mutation assays further support these observations: AID has been shown to mutate transcriptionally active transgenes in lymphoma cell lines (Bachl et al., 2001) and in AID-overexpressing non-B cell lines (Yoshikawa et al., 2002). In addition, AID was proved essential for the generation of c-myc/IgH translocations (Ramiro et al., 2004) by inducing DNA DSBs both at IgH and c-myc loci (Robbiani et al., 2008). These and other studies (Kovalchuk et al.,

2007; Pasqualucci et al., 2008) provided the first link between AID off-target activity and chromosome translocations (TCs). Further work reported that AID can generate DNA lesions (DSBs) in other non-Ig genes, which can lead to their translocation (Robbiani et al., 2009). Finally, high-throughput profiling of AID binding showed widespread interaction of AID with many genes across the genome (Yamane et al., 2011) and genome-wide maps of AID-induced TCs and DSBs suggested that AID can be mistargeted to hundreds of off-targets (Chiarle et al., 2011; Klein et al., 2011; Staszewski et al., 2011).

The mechanisms driving AID off-targeting remain object of intense study. Recent publications indicate that many TCs triggered by AID occur at regions where sense and antisense transcription converge (Meng et al., 2014; Qian et al., 2014). This convergent transcription (ConvT) arises at superenhancers, where antisense transcription originates within sense transcribed genes (Meng et al., 2014; Qian et al., 2014). In the same line, the RNA exosome has also been implicated in targeting AID to divergently transcribed loci (Pefanis et al., 2014). A hypothetical model based on these findings posits that: 1) ConvT induces RNAP II stalling, which helps recruit AID via SPT5; 2) RNA exosome detects and degrades antisense transcripts while contributing to AID recruitment; 3) Both RNAP II stalling and RNA exosome action facilitate AID access to ssDNA. Furthermore, ChIP-Seq studies suggest that AID off-targeting may also have an epigenetic component, with marks of active enhancers and transcription elongation providing nucleosome accessibility to AID (Wang et al., 2014). Together, these findings begin to define a set of transcription-related features relevant for AID off-targeting, but further work will be necessary to unveil the molecular mechanisms responsible for AID aberrant activity (Figure 4).

## 3.5. Regulation of AID activity

### 3.5.1. Transcriptional regulation of AID

*Aicda* expression is mostly restricted to activated B cells and it is triggered by cytokines and cell to cell interactions in the context of an antigen induced activation. *Aicda* is highly expressed in GC B cells, but strongly repressed in memory and plasmatic B cells (Crouch et al., 2007; Shaffer et al., 2002). *Aicda* can be regulated by both activating and repressing transcription factors, such as c-MYB, E2F, ID2 -negative regulators- or PAX5 -a positive regulator- (Gonda et al., 2003; Tran et al., 2010). Indeed, there are 4 well conserved regulatory regions in *Aicda* gene containing binding sites for up to 19 transcription factors

(Stavnezer, 2011). Thus, AID expression is tightly linked to the transcriptional program of the activated B cell and the balance between repressing and activating signals limits AID expression to this specific stage. AID haploinsufficiency (Sernández et al., 2008; Takizawa et al., 2008) further supports that AID levels are physiologically limited, maybe to minimize its deleterious function.

### 3.5.2. Post-transcriptional regulation of AID

*Aicda* expression levels can also be regulated by microRNAs, non-coding RNA molecules of small size (20-23nt) that bind complementary mRNAs and either promote their degradation or block their translation. There is strong evidence for AID regulation by two microRNAs, miR-155 and miR-181b, which bind conserved sites in the 3'-UTR of *Aicda* and repress AID expression. miR-155 parallels AID expression in splenic B cells activated *in vitro* (Teng et al., 2008). Mutation of miR-155 binding site at the 3'-UTR of *Aicda* leads to a 2-3x increase of AID mRNA and protein and increases CSR (Dorsett et al., 2008; Teng et al., 2008) and IgH/c-myc translocations (Dorsett et al., 2008). On the other hand, miR-181b is expressed in resting B cells and progressively downregulated upon B cell activation (Yébenes et al., 2008). miR-181b overexpression downregulates *Aicda* mRNA levels and impairs CSR (Yébenes et al., 2008). Together, miR-155 and miR-181b fine-tune AID expression, with experimental evidence suggesting a non-overlapping function by which miR-181b would prevent premature AID expression in resting B cells and miR-155 would act as safety control to limit AID levels in activated B cells.

### 3.5.3. Post-translational regulation of AID

A further layer of regulation occurs at the post-translational level, where subcelullar localization and stability of AID protein, together with phosphorylation, balance AID quantity and activity. AID is predominantly cytoplasmic in resting B lymphocytes. This is mostly due to the cooperative action of active cytoplasmic retention (Methot et al., 2015; Patenaude et al., 2009) and nuclear export mechanisms (Brar et al., 2004; Ito et al., 2004; McBride et al., 2004). In the cytoplasm, AID interacts with HSP90, which prevents its proteasomal degradation (Orthwein et al., 2010); and EEF1A, which indepently contributes to sequestering AID in the cytoplasm (Methot et al., 2015). Upon B cell activation, AID shuttles to the nucleus by an active transport mechanism (Ito et al., 2004; Patenaude et

al., 2009) where it exerts its deaminating function. Of note, AID is quickly degraded in the nucleus by both ubiquitin dependent and independent pathways (Aoufouchi et al., 2008; Uchimura et al., 2011). Thus, AID is much more stable in the cytoplasm than in the nucleus, which safeguards the B cell genome from its deletereous activity. In addition, AID activity can also be regulated by phosphorylation. There is experimental evidence for phosphorylation of AID in up to five conserved residues (Ser3, Thr27, Ser38, Thr140 and Tyr184), although only three have been proven to affect AID activity when phosphorylated: Ser38 and Thr140, which increase AID activity *in vivo*; and Thr27, which inhibits SHM and CSR *in vitro* (Basu et al., 2005; McBride et al., 2006, 2008; reviewed in Orthwein and Di Noia, 2012).

### 3.5.4. Cell-cycle linked regulation

Recent evidence points to a cell-cycle linked regulation of AID activity that coordinates protein shuttling to the nucleus with a stage of DNA damage tolerance that favours AID mutagenesis. In that sense, it has been shown that nuclear localization of AID in G1 phase is well tolerated, while it compromises cell viability during S-G2M (Le and Maizels, 2015). In addition, AID degradation is slower in G1 than in S or G2-M phase in this system (Le and Maizels, 2015). Further work in primary B cells has demonstrated that IgH loci deamination is restricted to early G1 phase and suggested that nuclear disassembly/reassembly together with post-mitotic transcription resumption renders DNA vulnerable to damage (Wang et al., 2017).

Together, the mechanisms presented above define a tight spatiotemporal regulation of AID activity that balances the introduction of mutations in Ig loci with the maintenance of genome integrity.

### 3.6. AID biology beyond B lymphocytes: expression and activity in non-lymphoid tissues

Although initially believed to be exclusively expressed in germinal center B cells, numerous studies have reported AID expression outside the B cell compartment. AID expression has been detected in a variety of pluripotent tissues, such as oocytes, spermatocytes, primordial germ cells or embryonic stem cells (Bhutani et al., 2010; Morgan et al., 2004; Popp et al., 2010; Schreck S et al., 2006); and has been proposed to play a role in epigenetic programming during early development through the deamination of 5-methylcytidine into thymine (Bhutani et al., 2010; Morgan et al., 2004; Popp et al., 2010). However, the involvement of AID in active demethylation remains controversial (reviewed in Ramiro and Barreto, 2015). In addition, AID is expressed in these tissues at levels orders of magnitude lower than those found in GC B cells, which poses some concern on the physiological relevance of these observations. Other studies have linked inflammation to AID expression through the NF-κB pathway, which normally contributes to the induction of AID in B lymphocytes (Dedeoglu et al., 2004; Tran et al., 2010). For instance, *in vitro* activation of this pathway by TNFα triggers AID expression in different non-B cell types (Endo et al., 2007, 2008; Matsumoto et al., 2007). In addition, infection by *H. pylori* produces aberrant AID expression in human gastric epithelial cells (Matsumoto et al., 2007); liver tissue from hepatitis patients expresses AID (Kou Tadayuki et al., 2006); and mouse models of inflammatory bowel disease present AID expression in colon epithelium (Takai et al., 2012). This ectopic AID expression has been related to gastric carcinogenesis (Matsumoto et al., 2007), hepatocarcinoma (Endo et al., 2007; Kou Tadayuki et al., 2006) and colitis-associated colorectal cancer (Takai et al., 2012) where somatic mutations (presumably introduced by AID) have been identified in *Tp53, c-myc* and *Pim1*. Additionally, several epithelial breast cancer cell lines have been shown to express AID (Babbage et al., 2006). These findings suggest that AID deregulation could contribute to the development of a wide variety of non B-cell neoplasias, particularly in the epithelial context. Previous work from our lab tested this hypothesis by generating conditional mouse models of AID overexpression in pancreas and colon epithelium and showed that AID expression alone, even at levels similar to those found in GC B cells, is not sufficient to promote carcinogenesis in these tissues (Pérez-García et al., 2015).

## 4. AID off-targeting and lymphomagenesis

As we have introduced before, AID targeting is not restricted to the Ig loci. SHM has been found in a number of non-Ig genes both in humans and mice (Liu et al., 2008; Müschen et al., 2000; Pasqualucci et al., 2001; Shen et al., 1998), and AID can generate mutations and DSBs that lead to chromosomal tranlocations between the IgH locus and a proto-oncogen, a hallmark of many mature human B cell lymphomas (Kovalchuk et al., 2007; Pasqualucci et al., 2008; Ramiro et al., 2004, 2006, Robbiani et al., 2008, 2009). AID depletion delays the onset of lymphomagenesis in different *in vivo* models, such as IL6 (Ramiro et al., 2004) and pristane (Kovalchuk et al., 2007) induced plasmacytomas or *Bcl6* overexpression driven lymphomas (Pasqualucci et al., 2008). This establishes a direct link between AID off-targeting and B cell malignant transformation. Moreover, it has been described that the generation of translocations also depends on UNG (Ramiro et al., 2006), which suggests a common mechanism for SHM, CSR and translocations. Finally, a mouse model overexpressing AID in B cells showed extensive genomic damage and widespread DSBs, indicating that AID may be sufficient to produce the DNA lesions underlying lymphomagenesis (Robbiani et al., 2009).

Therefore, there is strong evidence for AID off-target activity and the development of lymphoma through the generation of chromosome translocations. In this work, we aim at improving the understanding of the molecular mechanisms that define AID specificity. This will likely provide insights on how genomic integrity is maintained in hypermutating B lymphocytes and why SHM is mistargeted to genes relevant for carcinogenesis.

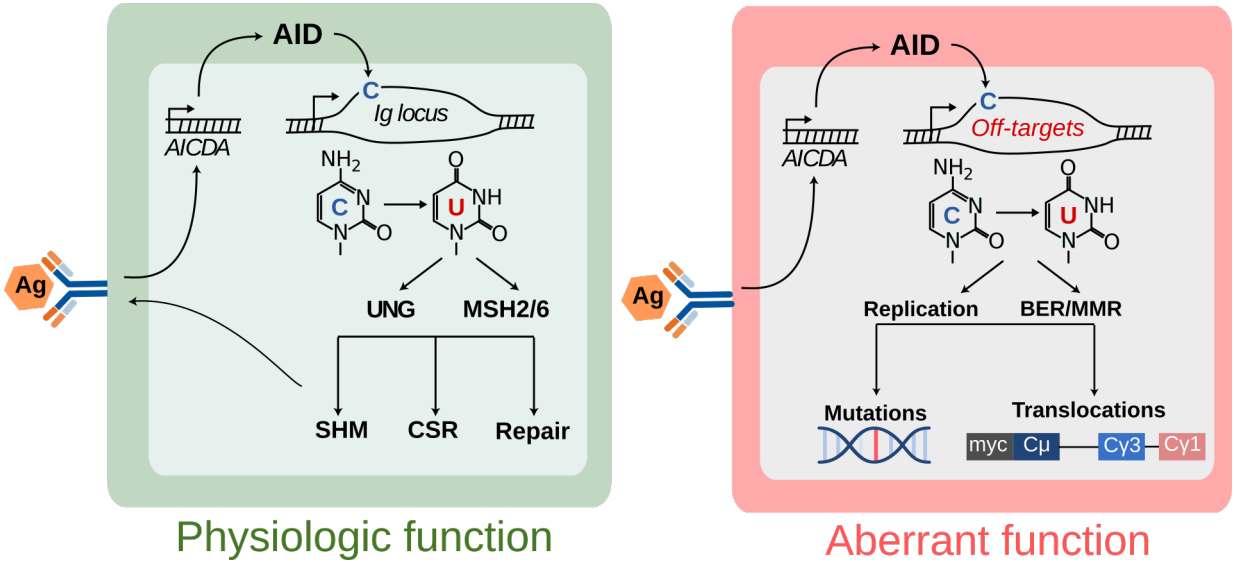



Figure 4 | **AID activity during the Germinal Center Reaction.** Schematic representation of AID activity outcomes in the Germinal Center. Upon activation, B cells express high levels of AID, which is translocated to the nucleus. Once there, AID deaminates cytosine residues in the variable or switch regions of the Ig loci to trigger SHM and CSR, respectively (left panel). However, AID activity is not restricted to Ig genes and can also act on other genes (off-targets), leading to point mutations or translocations with oncogenic potential (right panel). *Adapted from Methot&Di Noia, 2017.*

# II. OBJECTIVES

## Objectives

Activation Induced Deaminase (AID) plays a critical role in the immune response by diversifying the antibody repertoire through the deamination of cytosines in the immunoglobulin loci. However, it can also introduce DNA lesions in other regions of the genome, leading to mutagenic events and translocations with oncogenic potential. However, the mechanisms driving AID target specificity remain poorly understood. In this thesis work, we approached the following objectives:

1. To develop a high throughput strategy, based on target enrichment and next generation sequencing, for the identification of AID targets.

2. To identify and characterize AID targets in Germinal Center B lymphocytes.

3. To analyze the contribution of the Base Excision Repair and Mismatch Repair pathways to the resolution of AID-induced deaminations.

4. To analyze the contribution of AID mutational activity to Germinal Center derived malignancies.

# III. METHODS

## Mice

*Ung* and *Msh2* mutant mice used in this study were generated by crossing *Ung*$^{-/-}$ mice (Nilsen et al., 2000) and *Msh2*$^{-/-}$ mice (Reitmair et al., 1995). *Aicda*$^{-/-}$ mice have been previously described (Muramatsu et al., 2000). Mice were housed in specific pathogen-free conditions. Male and female mice between 20-28 weeks were used for the experiments, unless specified otherwise (gene expression profiling by RNA-Seq). Number of animals per group to detect biologically significant effect sizes was calculated using appropriate statistical sample size formula. All experiments were done in concordance to EU Directive 2010/63EU and Recommendation 2007/526/EC regarding the protection of animals used for experimental and other scientific purposes, enforced in Spanish law under RD 53/2013.

## Isolation of mouse B cells from secondary lymphoid organs

Mice were euthanized by $CO_2$ exposure. Peyer's patches were isolated from the ileum of necropsied mice and stored in ice in "complete" RPMI medium (RPMI-1640 -Sigma Aldrich- supplemented with 10% (v/v) Fetal Bovine Serum (FBS), HEPES (20mM), penicilin (50 U/ml) and streptomicin (50µg/ml) until processed. Organ disgregation was performed in 70µm pore nylon cell strainers (BD Falcon) in complete RPMI medium and cell suspensions were centrifugued 10' at 400 x G and 4ºC. Cell pellets thereof obtained were resususpended in PBS 1X 1% (v/v) FBS.

## DNA capture library

A set of 1588 genomic regions was selected as a representation of the genome. Details on gene selection can be found in "Results" section. Briefly, RNA probes were designed in SureDesign/eArray platform (Agilent) to capture the first 500bp from the TSS of a collection of 1375 different genes. To optimize capture yield, probes covered 50 extra nucleotides at the 5' and 3' boundaries of our regions of interest. Furthermore, they were designed to yield a 5x tiling frequency. This means that each nucleotide to be captured is covered by at least 5 different probes and implies a high density coverage by 20 tightly tiled 120nt-long baits for each of the 1588 regions to be captured. Finally, a custom SureSelect$^{XT}$ capture library was synthesized by the manufacturer.

## DNA capture and sequencing

Germinal center (*Cd19*$^{+}$*Fas*$^{+}$*GL7*$^{+}$) B cells were isolated from Peyer's patches of *Ung*$^{+/-}$*Msh2*$^{+/-}$, *Ung*$^{-/-}$*Msh2*$^{+/-}$, *Ung*$^{+/-}$*Msh2*$^{-/-}$, *Ung*$^{-/-}$*Msh2*$^{-/-}$ mouse littermates and *Aicda*$^{-/-}$ mice (Table 1) by sorting in a BD Biosciences FACSAria cell sorter after staining with anti-mouse antibodies to *Cd19*, *Fas* and *GL7* (BD Biosciences). Staining was perfomed in PBS 1X 1% (v/v) FCS; cells were washed in PBS 1X 1% (v/v) FCS (10' centrifugation at 400 x G and 4ºC), resuspended in "sorting buffer" (PBS 1X 2% (v/v) FCS 15mM HEPES) and filtered through a pre-separation 70µm pore filter (BD Biosciences). FACS sorted cells were lysed in lysis buffer (50mM Tris pH=8, 200mM NaCl, 10mM EDTA pH=8, 1% SDS) and proteinase K (1/50 from a 20mg/ml stock). Genomic DNA (gDNA) was purified by a two-step extraction by phenol:chloroform:isoamyl alcohol (25:24:1) and chloroform followed by precipitation in absolute ethanol plus Pellet Paint coprecipitant (Merck Millipore) and washing with 75% ethanol. Quantification was done in an Invitrogen Qubit Fluorometer. DNA capture, library preparation and DNA sequencing was performed by the Genomics Unit at CNIC following manufacturer's instructions. Briefly, gDNA (1,1µg per sample) was fragmented in a Covaris sonicator to ~150-200 nucleotide long (average size) fragments and purified using Agencourt AMPure XP beads. Quality was assessed with the 2100 Bioanalyzer (Agilent). Then, fragment ends were repaired, adapters were ligated, and the resulting library was amplified and hybridized with our custom SureSelect$^{XT}$ (Agilent) library of RNA probes. DNA-RNA hybrids were then captured by magnetic bead selection. After indexing, libraries were single-end sequenced in an Illumina HiSeq 2500 platform following manufacturer's instructions.

| Genotype | # mice exp1 | # mice exp2 |
|---|---|---|
| *Ung*$^{-/-}$*Msh2*$^{-/-}$ | 37 | 8 |
| *Ung*$^{-/-}$*Msh2*$^{+/-}$ | 46 | 8 |
| *Ung*$^{+/-}$*Msh2*$^{-/-}$ | 46 | 2 |
| *Ung*$^{+/-}$*Msh2*$^{+/-}$ | 10 | 11 |
| *Aicda*$^{-/-}$ | 31 | - |

Table 1 | **Summary of mice used in this study.**

## Target enrichment assessment by qRT-PCR

*Noxa1*, *Ostn* and *Pcna* amplifications were quantified with SYBR green assay (Applied Biosystems) in an AbiPrism AB7900 Standard real-time PCR system. *Gapdh* amplifications were used as normalization controls. Primers used for the amplifications are indicated in Table 2. SDS software (Applied Biosystems) was used for the analysis of the data.

| Oligonucleotide | | Sequence (5'-3') |
|---|---|---|
| *Gapdh* | Forward | TGA AGC AGG CAT CTG AGG G |
| | Reverse | CGA AGG TGG AAA GTG GGA G |
| *Ostn* | Forward | CAT AGT GTT GCT GTG GTT |
| | Reverse | CAT TAT ATT GGT CTG CTG TT |
| *Noxa1* | Forward | CGC GGG ACA GCA ATG AGA AG |
| | Reverse | CCA TCT ACT CAG TTT CAA GGA |
| *Pcna* | Forward | CTC CAG CAC CTT CTT CAG |
| | Reverse | TCT CAT CTA GTC GCC ACA |

Table 2 | **Primers used for qRT-PCR.**

## Sanger sequencing

Regions to be sequenced were amplified from 160-200ng genomic DNA in 4 independent reactions (40-50ng DNA each) to minimize possible PCR biases. Primers used are indicated in Table 3. Amplification reactions were carried in a final volume of 25µl using 2.5U of Pfu Ultra HF DNA polymerase (Agilent) and the following PCR setup: 95º for 2 min; 25 (*Cd19*, *Cdk4*) or 26 cycles (*miR142, Hist1h1b*) of denaturation at 94º for 30 s, annealing at 57º (*miR142, Hist1h1b*) or 58º (*Cd19, Cdk4*) for 30 s, extension at 72º for 1 min; final stage of 72º for 10 min.

3' A-tailing was performed to make PCR products suitable for TA cloning by adding 0,5µl Taq polymerase (New England Biolabs) to each reaction (10' extension at 72ºC) inmediately after PCR amplification. PCR products were purified from a 1% agarose gel (Illustra Gel Band Purification Kit, GE Healthcare) and cloned into pGEM-T Easy vector (Promega) following manufacturer instructions. Competent DH5α E. Coli bacteria were heat-shock transformed with the constructs and grown

overnight in LB-Amp IPTG/X-Gal (40µL X-gal; 20µl IPTG per plate) plates. Individual, white (X-Gal negative) colonies (192-288 per gene) were picked into 96 well plates. Plasmidic DNA was then isolated (Plasmid MiniPrep Kit, Millipore) and sequenced by Sanger sequencing using SP6 universal primer. Sequence analysis was performed using SeqMan software (Lasergene).

| Oligonucleotide | | Sequence (5'-3') |
|---|---|---|
| *Hist1h1b* | Forward | ATG CCT TAG ACT TCA CCG CC |
| | Reverse | TTG TAA CCT TGA GTC GCC GC |
| *miR142* | Forward | CGG TCC CTG GGA AGT TAC AC |
| | Reverse | AAC GAG AGG CAA ACA GTC TTC A |
| *Cd19* | Forward | GCC CCT CTT CCC TCC TCA TA |
| | Reverse | CCT GCA CCC ACT CAT CTG AA |
| Cdk4 | Forward | TCT GGC AGC TGG TCA CAT GG |
| | Reverse | GAT CAC CAG CTA GTC GTC CC |

Table 3 | **Primers used for Sanger sequencing.**

## PCR-Seq to validate machine learning approach

40-50ng of genomic DNA were amplified using the primers included in Table 4. Amplification reactions were carried in a final volume of 25µl using 2.5U of Pfu Ultra HF DNA polymerase (Agilent) (95º for 2 min; 26 cycles of 94º for 30 s, 55º for 30 s and 72º for 1 min; final stage of 72º for 10 min). PCR products were purified (Illustra Gel Band Purification Kit, GE Healthcare) following manufacturer's protocol, fragmented using a Covaris sonicator and checked for integrity and size distribution in a 2100 Bioanalyzer. Libraries were then prepared by the CNIC Genomics Unit (NEBNext Ultra DNA Library Prep; New England Biolabs) following manufacturer's instructions. Sequencing was performed in an Illumina HiSeq 2500 platform. Mutation analysis was performed as previously described (Pérez-Durán et al., 2012).

| Oligonucleotide | | Sequence (5'-3') |
|---|---|---|
| *Apobec3* | Forward | GTC TTC CAT AGC CTG CTC ACA |
| | Reverse | TAG CTG ACT GGT GTG GTT CC |
| *Aurkaip1* | Forward | ACT TGT CAC TTC CGC AGT CC |
| | Reverse | CCA TCC CCA AGT CAG GTG TG |
| *Ccdc17* | Forward | TCT TTT CTG TCC AGT CCG CC |
| | Reverse | ACA AAT GGG CAG AGT CAG GG |
| *Cd52* | Forward | TAC TGC CGC ACA CAT GAC TC |
| | Reverse | TGA GGT GGG AAG CCA AAC AT |
| *Cd68* | Forward | AGG GGC TGG TAG GTT GAT TG |
| | Reverse | GGA GTC AGG ACT GGA TTT GAC |
| *Cd69* | Forward | TCT AAA GGT TTT GAG ACC CCC |
| | Reverse | TGA AGC CTC ATC AAC GCA CT |
| *Clec2d* | Forward | GGC TCC TGA CCT TGA AAT GC |
| | Reverse | AGG CAA CTT CTG CCA CTA TGC |
| *Coro1a* | Forward | AGG GCT CTG GGG TTC TAC TT |
| | Reverse | GGA AAT GAC CAC GGG GGT TT |
| *Hist1h1c* | Forward | CTC TAT CGG CGT ACT GCC AC |
| | Reverse | ATC GAG TCC CTT GCA ACC TT |
| *Il4i* | Forward | ATT CCC GAG GGA GGT GAG TG |
| | Reverse | GGT AGC TTC TCT CCG TCA CAC |
| *Maz* | Forward | GTC AAC AAA GAA CCC CTC CCT |
| | Reverse | CAC CTG TCC CCT GAG TTG TG |
| *Trex1* | Forward | GCC TAA CAG GTT TGA TTG TCC T |
| | Reverse | TAG GCT GAG CAC TCC CAG TC |

Table 4 | **Primers used for PCR-Seq.**

## Gene expression profiling by RNA-seq

Germinal center ($CD19^+FAS^+GL7^+$) and resting ($CD19^+FAS^+GL7^-$) B cells were sorted from Peyer's patches of littermate 12 weeks old WT C57BL/6 mice. Three biological replicates were analyzed, each composed of a pool of 5 female mice. RNA was purified from pellets of 2-2.5x$10^4$ cells and DNAse I treatment applied to avoid DNA contamination (Qiagen RNAeasy MiniKit). RNA quality was assessed with the 2100 Bioanalyzer showing high RNA purity and integrity. Sequencing libraries were prepared by CNIC Genomic Unit following manufacturer's protocol (NEB NEXT Ultra RNAseq Library Prep Kit, New England Biolabs) from 100ng RNA per replicate and sequenced in an Illumina HiSeq 2500 platform.

## Computational analysis

a. Pipeline to identify and annotate AID-induced mutations

Raw reads were demultiplexed by *CASAVA* (Illumina) to generate a fastq file that was aligned to mouse genome (NCBIm37 v61 Feb 2011) with *Novoalign* (Novocraft) (command line options: `-o SAM -F ILM1.8 -H -r None -q 2`). Samfiles were proccessed with *samtools* (Li et al., 2009) to generate a sorted bamfile (`samtools view` and `samtools sort` commands) that was piped to a custom Perl script for the analysis of AID mutations. The script depends on the ENSEMBL Perl API: core database, functional genomics, comparative genomics and variation data APIs. Briefly, the software analyzes the regions of interest in the bamfile, annotates hotspots, localizes and suppresses annotated SNP positions (Sanger Mouse Genomes Project SNP and Indel Release v2) and reports relevant information about AID activity. Details on parameters, filters and input/output files can be found in Table 5.

AID targets were identified as those genes accumulating significantly more C→T transition mutations in *Ung*$^{-/-}$*Msh2*$^{-/-}$ than in *Aicda*$^{-/-}$ mice (FDR ≤0.05, One-tail Fisher test and Benjamini-Hochberg correction).

Mutation frequencies were calculated as follows:

$$\text{Total mut.freq.} = \frac{\text{Total number of mutations}}{\text{Total sequenced length}}$$

$$\text{Mut.freq.}_{C/G} = \frac{(\text{Mutated cytosines} + \text{Mutated guanines})}{(\text{Seq length cytosines} + \text{Seq length guanines})}$$

$$\text{Mut.freq.}_{\text{WRC(Y)/(R)GYW}} = \frac{(\text{Mutated cytosines}_{\text{WRC(Y)}} + \text{Mutated guanines}_{\text{(R)GYW}})}{(\text{Seq length cytosines}_{\text{WRC(Y)}} + \text{Seq length guanines}_{\text{(R)GYW}})}$$

(Only cytosines in WRC(Y) and guanines in (R)GYW were considered to calculate mutation frequency at hotspots).

| Parameter | Defines | Details |
|---|---|---|
| --reffile | Input file (*required*) | Indexed reference genome in *fasta* format. |
| --bamfile | Input file (*required*) | Sorted *bamfile* (alignment of reads to the reference genome). |
| --posfile | Input file (*required*) | *Bedfile* containing the genomic coordinates of the genomic regions to be analysed: chr, start, end, name, strand. |
| --hsfile | Input file (*required*) | Text file containing Perl regular expressions* matching the motif of each of the hotspots to be analysed. |
| --snpfile | Input file (*required*) | Text file containing the name of the mouse strains that should be considered for SNP removal. |
| --snpfile | Input file (*required*) | *Bcf* file with information relative to the SNP calling of mouse strains as retrieved from the MGP. |
| --qbcut | Filter (*optional*) | Defines the base calling quality threshold for a nucleotide read to be considered for the analysis. Default is Q20 PHRED score |
| --qmcut | Filter (*optional*) | Defines the mapping quality threshold for a read to be considered for the analysis. Default is inactive. |
| --outfile | Output file (*required*) | Textfile containing the final report. |
| --ohs | Output file (*optional*) | Defines whether the SNPs found in the regions analysed should be reported to an output file. |
| --osnps | Output file (*optional*) | Defines whether identified hotspots should be reported to an output file. |
| --time | Technical report | Time and memory usage. |
| --help | Help | Display information about software usage. |

Table 5 | **Parameters, input and output files of our custom Perl software.**

b. Sequence context of mutated cytosines

The sequence context of mutated cytosines (C→T transition frequency ≥ $4x10^{-3}$) was analyzed in a window of 10 nucleotides using an *in-house* built Python script. LOGO representation was done using *WebLogo3* (http://weblogo.threeplusone.com/create.cgi) and percentage of each nucleotide in each position surrounding the mutated cytosine was calculated by a custom Perl script. Enrichment for adenosine, guanine, cytosine or thymine was tested against the sequence context of all cytosines present in the 1588 regions analyzed in this study (one-tailed t-test + Bonferroni correction).

c. Gene expression profiling by RNA-Seq

After demultiplexing by *CASAVA*, read quality was assessed by *FastQC* and sequencing adaptors were removed from sequence reads by *cutadapt* (Martin, 2011). The resulting reads were aligned to and quantified on the mouse transcriptome (NCBIm38 v75 Feb 2014) using *RSEM* (Li and Dewey, 2011) with the following parameters: `-p 3 --time --output-genome-bam --sampling-for-bam --bowtie-e 60 --bowtie-m 30 --bowtie-chunkmbs 512 --fragment-length-mean 180 --fragment-length-sd 50`.

Data were then processed with a differential expression analysis pipeline that used Bioconductor package EdgeR (Robinson et al., 2010) for normalization and differential expression testing. Transcript quantification was performed at the gene level, and only those genes bearing ≥ 3 counts per million were considered for differential expression analysis. Genes were considered differentially expressed at a q-value ≤ 0.05 (Benjamini-Hoschberg FDR).

d. Transcription rate analysis (GRO-Seq)

Reads were mapped to the mouse genome (mm9/NCBI37) using *bowtie2* (Langmead and Salzberg, 2012) and uniquely mapped, non-redundant reads were kept. Reads mapping in +/-1Kb from TSS were quantified and summarized at the gene level using *HTSeq*.

e. RNAP II and SPT5 recruitment

Quantification of RNAP II and SPT5 recruitment was extracted from (Pavri et al., 2010) (Table S3A in their manuscript).

f. Superenhancers analysis

Data was extracted from the catalog of superenhancers that overlap with gene bodies identified in germinal center B cells as published in (Meng et al., 2014) (TableS3 in their manuscript).

g. MED12 binding and epigenetic marks analysis

Sequencing data (fastq files) from MED12 and H3K4me1, H3K36me3, H3K79me2 epigenetic marks ChIP-Seq experiments was aligned to the mouse genome (NCBIm37 v61 Feb 2011) using *bowtie* (Langmead et al., 2009) (command line options: `--best -m1 -n2 -p2`). Alignment files were processed by s*amtools* to generate a sorted bamfile. Peak calling was done using *MACS2* (Zhang et al., 2008) following the optimal parameters for a histone modification status profiling as reported by the creators of the tool (callpeak function with command line options: `-q 0.01 -g 1.87e9 –nomodel –nolambda`) (Feng et al., 2011). Mapping of annotated peaks to genes was done using *GREAT* (Mc Lean, 2010) with the following parameters: **Species assembly:** Mouse NCBI build 37 (UCSC mm9, Jul/2007); **Background regions:** Whole genome; **Association rule:** Basal plus extension (Proximal: 5Kb upstream, 1Kb downstream; plus Distal: up to 1000Kb); Curated regulatory domains were not included for the analysis.

To remove background noise from data representation, percentage of positive targets and non-targets for a given mark was referred to the percentage of positive regions in the whole genome before plotting. For instance, 75,6% AID targets and 50% non-targets showed MED12 binding with 36,4% of the genome bearing a positive signal. Thus, percentage of AID targets binding MED12 was represented as 39,2% and percentage of non-targets binding MED12 as 13,6%. Nonetheless, statistics were calculated using original numbers.

h. Convergent transcription analysis (GRO-Seq)

Convergent transcription data analysis was performed as described in Meng et al., 2014. In brief, reads were mapped to the mouse genome (mm9/NCBI37) using *bowtie2* and uniquely mapped, non-redundant reads were kept. *HOMER* (Heinz et al., 2010) was used with default parameters to identify transcribed regions from both strands and *bedtools* (intersect function) (Quinlan, 2002) to find and annotate "Convergent transcription (ConvT) regions" (regions where a greater of 100pb sense and antisense transcription overlap occurs).

i. Machine learning to predict AID targets

The conditional inference tree for the classification was built using *ctree* function from *party* R package (Hothorn et al., 2006; Libbrecht and Noble, 2015) with default parameters. Genes with a background mutation frequency above $5x10^{-4}$ were excluded to avoid artifacts. The resulting 1339 genes were divided into two groups: AID target (272; 20% of total) and non-target (1067; 80% of total) genes and the following variables were fed into the model: expression, transcription rate, RNAP II and SPT5 recruitment (quantitative, continuous); MED12 recruitment, H3K4me1 recruitment, H3K36me3 recruitment, H3K79me2 recruitment, regulation by superenhancers and occurrence of convergent transcription (qualitative, discrete). All variables were assigned equal weights to fit the model.

| Software | Version |
|---|---|
| Bedtools | 2.24 |
| BioPerl | 1.6.924 |
| Bowtie | 1.1.1 |
| Bowtie2 | 2.2.4 |
| CASAVA | 1.8 |
| CIRCOS | 0.69-5 |
| Cutadapt | 1.9 |
| Debian GNU/linux (x86-64) | 7 (Wheezy) and 8 (Jessie) |
| ENSEMBL Perl API | ensembl65 |
| FastQC | 0.10.1 |
| GNU bash | 4.2 and 4.3.30 |
| GREAT | 3.0.0 |
| HOMER | 4.6 |
| HTSeq (python-htseq) | 0.5.4 |
| Novoalign | 2.08.01 |
| MACS2 | 2.1.0.20140616 |
| Perl | 5.20.2 |
| Python | 2.7.9 |
| R | 3.1.1 |
| R package: *edgeR* | 3.16.5 |
| R package: *gplots* | 3.0.1 |
| R package: *limma* | 3.30.13 |
| R package: *party* | 1.2-3 |
| R package: *pheatmap* | 1.0.8 |
| RSEM | 1.2.25 |
| sratoolkit | 2.8.1.3 |
| Weblogo | 3 |

Table 6 | **Software versions used for computational analysis.**

j. Annotation of AID targets

Annotation of AID targets was performed based on public data on sequencing of human diffuse large B cell lymphoma, Burkitt lymphoma and follicular lymphoma tumors (Lohr et al., 2012; Love et al., 2012; Miranda et al., 2014; Morin et al., 2013; Okosun et al., 2014; Zhang et al., 2013).

## Raw data availability

Sequencing data generated for this thesis are available through the Gene Expression Omnibus: targeted DNA deep sequencing (GSE102944); RNA-Seq (GSE98086).

The rest of the datasets analyzed are publicly available through the Gene Expression Omnibus and/or Sequence Read Archive: GRO-Seq (GSE62296): Germinal Center B cells (SRR1611832, SRR1611833, SRR1611834), Naïve B cells (SRR1611829, SRR1611830, SRR1611831); ChIP-Seq of Pol II and SPT5 (GSE24178); ChIP-Seq data of epigenetic marks: MED12 (SRX347810), H3K4me1 (SRX347815), H3K36me3 (SRX185869), H3K79me2 (SRX185843).

## Code availability

Source code developed for the analysis described in this thesis will be provided upon reasonable request.

## Statistical analysis

Statistical analyses were performed with stats R package v3.1.1. Error bars in figures represent standard error of the mean (SEM). Student's t-test was applied to continuous data and Fisher test was used to assess differences between categorical variables. P-values were corrected for multiple hypothesis testing by Benjamini-Hochberg or Bonferroni method where appropriate. Differences were considered statistically significant at $p \leq 0.05$ or $q \leq 0.05$.

# IV. RESULTS

## 1. Development of a custom protocol to detect AID-induced mutations

While AID is crucial for the humoral immune response, its activity poses a risk to genome integrity through the bystander deamination of cytosine residues in non-Ig genes. Thus, it is fundamental to understand the mechanisms that drive AID target specificity. However, this issue remains unresolved, mostly due to the technical challenge to reliably detect AID-mediated mutagenesis. There are two main reasons that have classically hampered the discovery of AID targets: first, AID-induced mutations occur at very low frequencies (ranging from $10^{-2}$-$10^{-3}$ -in the case of the Ig locus- to $10^{-4}$ or lower -in off-targets-); second, they are non-clonal (different cells carry different mutations) which obscures their faithful identification. Therefore, the detection of mutations derived from AID activity requires high fidelity and high depth (i.e. reading each nucleotide a high number of times) sequencing of candidate genes. Until now, only a relatively small number of genes has been directly assayed for AID mutations (Liu et al., 2008; Pasqualucci et al., 1998, 2001; Shen et al., 1998). These studies were performed by PCR amplification of individual genes, cloning and Sanger sequencing of single colonies, a time-consuming approach that is not suitable to evaluate large collections of genes. On the other hand, genome-wide studies of AID specificity have made use of high-throughput analysis of AID binding, which does not warrant AID activity, or AID-induced DSBs or TCs, which involve complex processing of the initial deamination by AID (Chiarle et al., 2011; Klein et al., 2011; Meng et al., 2014; Qian et al., 2014; Staszewski et al., 2011; Yamane et al., 2011). In this work we propose to directly measure AID mutagenic activity in a wide representation of the B cell genome by next generation sequencing (NGS). Our lab previously showed that NGS of PCR products (PCR-Seq) can increase sequencing depth by a thousand fold when compared to Sanger sequencing and permits the detection of AID-induced mutations in the IgH locus (Pérez-Durán et al., 2012). Here, we coupled the power of NGS to a target enrichment protocol that provided the high depth and high sensitivity necessary to evaluate AID mutagenic activity in a very large collection of genes.

### 1.1. Design of a custom RNA bait capture library

To explore the scope of AID-induced mutations at a high-throughput scale, we made use of a target enrichment protocol and designed a library of biotinylated RNA probes to capture a collection of 1588 gene fragments (corresponding to 1375 different genes) as a representation of the B cell genome (Annex I; Figure 5). Gene selection included three different groups. The first group accounts for 85% of all genes and was randomly selected from within genes annotated to be protein coding in AmiGO database. Bioinformatics analysis were performed to ensure even representation of chromosomal location and unbiased biological function in this set of genes. The second group accounts for 13% of the capture library and includes all genes that had already been tested for AID activity in the literature (Gordon et al., 2003; Müschen et al., 2000; Pasqualucci et al., 2001; Pavri et al., 2010; Robbiani et al., 2009; Shen et al., 1998) and IgH probes ($J_H4$, Sμ and Eμ regions) as positive controls. The third group contains a small set of genes related to cancer development, and comprises 2% of all genes in the library. Since AID activity in Ig genes has been reported to concentrate in the vicinity of the transcriptional start site (TSS) (Besmer et al., 2006; Peters and Storb, 1996; Storb, 2014), probes were designed to capture the first 500bp downstream of the TSS of each of the 1375 genes. As various genes contained more than one predicted TSS, the library includes a total of 1588 different genomic regions (Figure 6A,B). This library design enables the capture of about 0.8Mb of genomic sequence and coupled to high depth sequencing will theoretically allow reading each nucleotide at a depth ≥ 2000x.

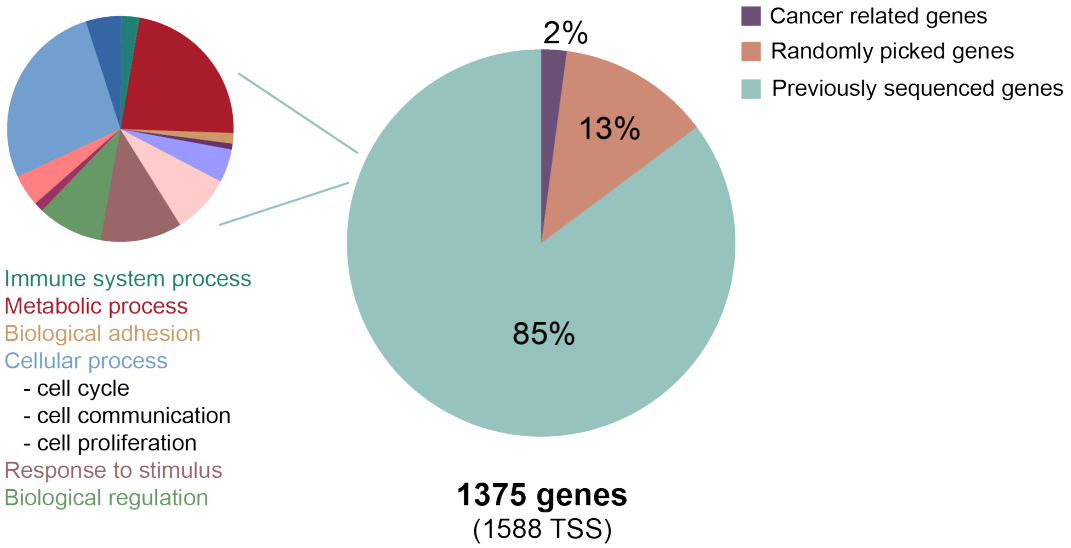



Figure 5 | **Groups of genes included in the custom capture library.**

### 1.2. Validation of the target enrichment protocol

To test the efficiency of the target enrichment protocol we performed real-time qPCR to measure the relative DNA abundance of genes included (*Noxa* and *Pcna*) and not included (*Ostn*) in our capture library before and after enrichment. Equivalent amounts of DNA were used for each amplification and cycle threshold values ($C_t$) were measured in all the samples. $C_t$ is defined as the number of amplification cycles that results in a fluorescent signal above the detection threshold and is inversely proportional to the amount of DNA in the sample. Therefore, we would expect those genes included in our capture library to have lower Ct values after enrichment than before, and vice versa for those genes not included. Indeed, we found that in the case of *Noxa* and *Pcna*, enriched fractions consistently amplified 11 cycles earlier ($\Delta C_t = Ct_{input} - Ct_{enriched} = 11$) than input fractions (Table 7). This means that the capture protocol yielded, approximately, a 2000 fold enrichment ($2^{\Delta CT} = 2048$) (Figure 6C). Conversely, *Ostn* amplified 11 cycles later in the enriched fraction than in the input fraction ($\Delta C_t = Ct_{input} - Ct_{enriched} = -11$) (Table 7) revealing a depletion of approximately 2000 fold ($2^{\Delta CT} = 1/2048$) (Figure 6C). Notably, a 2000 fold enrichment of the target regions implies an equivalent increase in the sensitivity of the system to detect AID-induced mutations. Thus, we can conclude that our target enrichment approach allows efficient enrichment of targets and provides improved sensitivity to analyze AID mutagenic activity.

| | | Ct value | | |
|---|---|---|---|---|
| | | ***Noxa*** | ***Pcna*** | ***Ostn*** |
| **Experiment 1** | Enriched | 18 | 16 | 32 |
| | Input | 29 | 27 | 24 |
| **Experiment 2** | Enriched | 18 | 16 | 36 |
| | Input | 29 | 27 | 24 |

Table 7 | **Cycle threshold values in post-enrichment and input fractions.** Each Ct value represents the mean of two independent technical replicates (standard deviation equals zero in all cases).

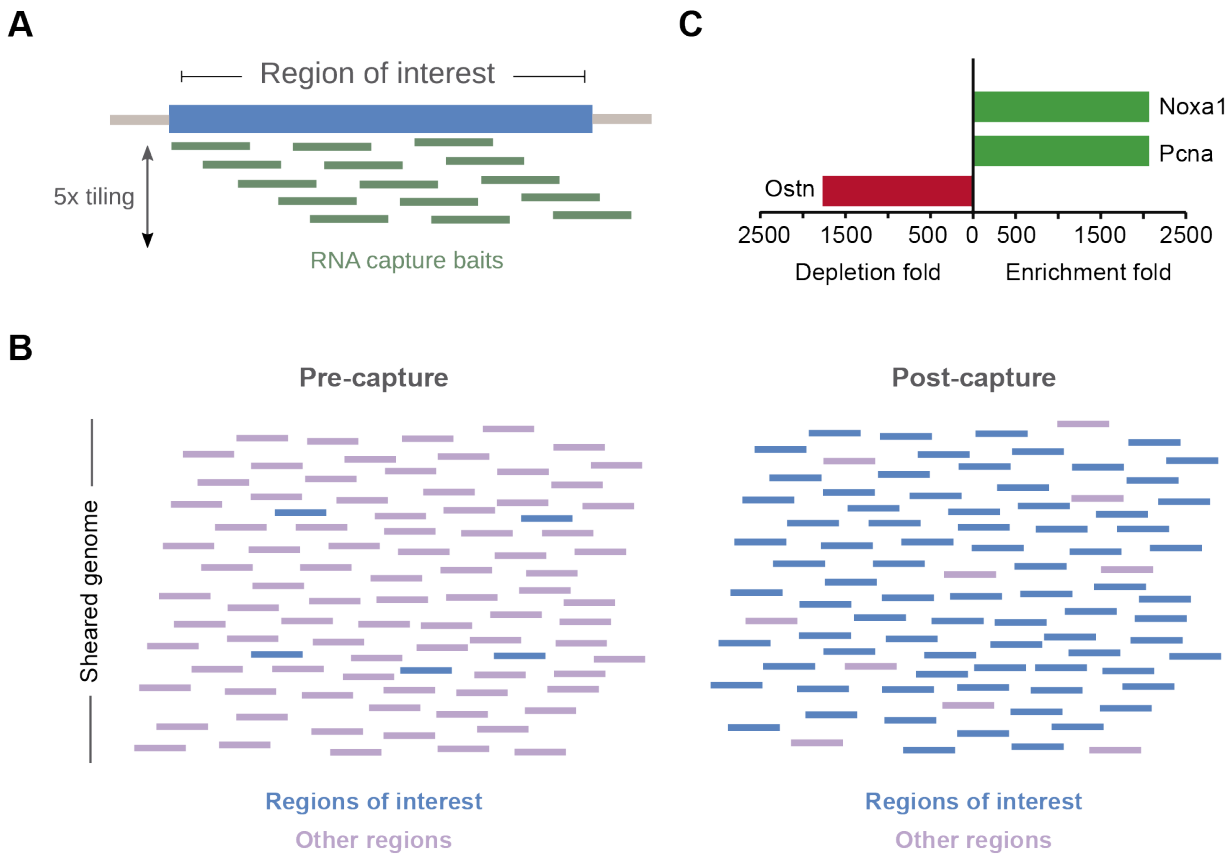



Figure 6 | **Target enrichment protocol allows efficient enrichment of selected genes.** (**A** and **B**) Schematic representation of the DNA capture approach. (**A**) Biotinylated RNA baits are designed to capture 500bp from the Transcription Start Site (TSS) of a collection of 1375 genes. DNA-RNA hybrids are purified by a magnetic field. After denaturalization of the duplex, DNA is amplified by PCR and deep-sequenced (**B**) Before capture, regions of interest constitute a small proportion of the sample, (left); after capture, they are highly enriched and represent a major fraction of the sample (right) (**C**) Genomic DNA corresponding to genes included (green bars) and not included (red bar) in the SureSelect library was quantified by qRT-PCR before and after DNA capture enrichment. Graph represents fold depletion or enrichment calculated as $2^{(\text{CT input} - \text{CT enriched fraction})}$. Mean of two independent experiments is represented.

### 1.3. Development of a bioinformatics pipeline to analyze AID mutational activity

To perform the mutation analysis of the sequences obtained with our target enrichment and high depth sequencing protocol, we developed a custom bioinformatics pipeline that gathers, summarizes and reports information for AID activity in a per TSS manner (Figure 7). Briefly, the pipeline includes two major steps: 1) alignment of sequences to the reference genome; 2) processing of the alignments to reliably detect AID induced mutations and summarize and report data comprehensively. Our pipeline uses Novoalign for the alignment of reads for a number of reasons: it takes into account base calling quality for the alignment; it has built-in adapter and base quality trimming; it permits the alignment of reads with INDELS; it can align mismatches that cover up to 50% of the read length; and it scores among the lowest in terms of incorrectly mapped reads and among the highest in terms of proportion of reads aligned (Hatem et al., 2013; Li and Homer, 2010). Details on the fine-tuning of the aligner can be found in materials and methods section. Here, we will focus on the second step. A custom analytic software was developed with the aim of getting a modular and versatile tool for the detection and annotation of mutations from NGS genome-wide data. The software is written in Perl and heavily relies on ENSEMBL Perl API[1]. This enables code reuse and quick adaptability to potential new requirements. The pipeline workflow operates as follows (Figure 7): first, sequences stored in a fastq file are aligned to the reference genome by Novoalign; second, the samfile containing the alignments is processed by samtools (Li et al., 2009) to generate a sorted bamfile; third, the sorted bamfile is piped to our Perl software for the analysis and classification of mutations. In brief, the software connects to ENSEMBL database and downloads the sequence of the regions of interest, finds hotspots in those regions and annotates them. Then, it does a pileup[2] of the alignments and processes them base by base. At this point, genomic positions where SNPs have been identified by the Sanger Mouse Genomes Project (Keane et al., 2011) in the selected mouse strains are removed from the analysis. Nucleotide reads that do not pass the quality thresholds are filtered out as well. When the analysis of a genomic region is finished, the software creates a summary containing both technical parameters related to sequencing and information about mutations and prints it to an output file. It also annotates all hotspots found and all genomic positions removed from the analysis due to known SNPs.

1 Application Programming Interface, a standard set of functions, protocols and tools used to create software.
2 Summary of the alignment at the nucleotide level, including: genomic coordinates of the reference nucleotide, number of reads covering the site, the specific nucleotides read and their base calling qualities.

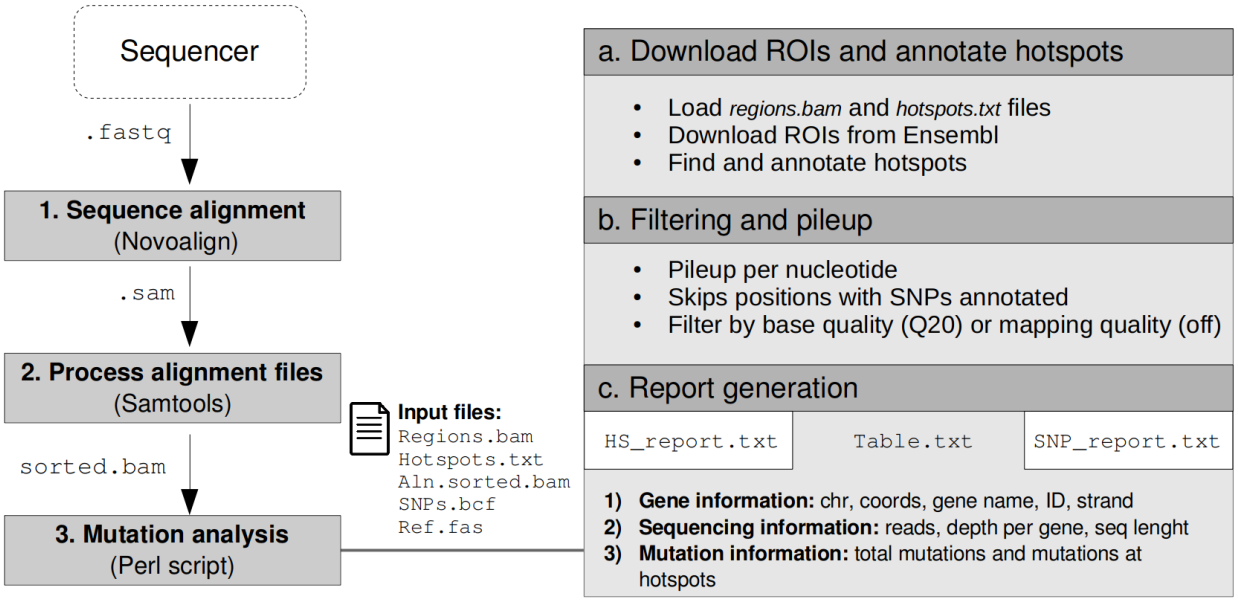



Figure 7 | **Schematic representation of the custom bioinformatics pipeline developed for the analysis of AID mutations.** After demultiplexing, reads are aligned to the mouse genome; alignments are processed and fed to a custom Perl script that removes SNPs, finds and annotates hotspots, filters sequencing reads by base and mapping quality and generates a comprehensive report that includes relevant sequencing technical parameters and information about mutations.

The final report generated by the software is a plain text tabbed file containing one row per TSS and 84 columns with the following information: 1) **Gene information.** Genomic coordinates, strand, ENSEMBL ID and gene name; 2) **Sequencing technical information.** Number of reads, average depth[3], total sequenced length[4], percentage of nucleotides covered by more than 300 reads, percentage of nucleotides with more than 90% of reads passing the established base calling quality filter; 3) **Mutation information.** a) Total mutations: total number of mutations, number of each of the $4^3$ possible mutations, number of mutations at C:G pairs, number of transitions and transversions at C:G pairs, number of mutations at A:T pairs and frequencies for all of the previously mentioned parameters. b) Mutations at hotspots: total number of mutations, total number of transitions and transversions and frequencies for WR<u>C</u>(Y) and (R)<u>G</u>YW hotspots.

3 Number of times a given genomic region has been read on average.
4 Total sequence volume covering a particular genome region.

Additionally, this pipeline can generate two additional files: 1) **Hotspots file:** reports information about the hotspots identified by the software at the gene level. First line contains ENSEMBL identifier, gene name, coordinates and number of hotspots in tabbed columns. Second line includes the DNA sequence of the region analyzed. Subsequent lines (one per identified hotspot) contain coordinates of the genomic position where the hotspot starts, position of the hotspot relative to the region being analyzed and type(s) of hotspot(s) found in that position. 2) **SNPs file:** *vcf* file that contains detailed information about the SNPs identified by the MGP in our regions of interest for the mouse strains selected.

In conclusion, we built a custom bioinformatic pipeline to mine high-throughput NGS data and analyze AID mutational activity.

## 2. Identification and characterization of AID targets in Germinal Center B lymphocytes

### 2.1. Capture-based deep sequencing allows high throughput identification of AID targets

To investigate AID mutational activity in the B cell genome, we made use of a mouse model deficient for UNG and MSH2, key components of the BER and MMR pathways. In their absence, AID induced U:G mismatches remain unprocessed and are replicated over, thus leaving a mutational signature of C→T and G→A transitions that reveals the footprint of AID deamination events on DNA (Methot and Di Noia, 2017; Rada et al., 2004). We used GC B cells from Peyer's patches, secondary lymphoid organs that form part of the gut-associated lymphoid tissue. Peyer's patches act as an immune sensor of the small intestine, with lymphoid cells undergoing chronic stimulation very likely due to antigens present in the gut environment (González-Fernández and Milstein, 1993). GCs develop spontaneously in mouse Peyer's patches in the absence of immunization, with high rates of SHM accumulating in Ig genes (González-Fernández and Milstein, 1993). GC ($CD19^+FAS^+GL7^+$) B cells were purified from Peyer's patches of *Ung*$^{-/-}$*Msh2*$^{-/-}$ and from *Aicda*$^{-/-}$ mice as negative controls. Genomic DNA was isolated, subjected to target enrichment and deep-sequenced (Figure 8). This approach allowed an extremely high sequencing quality[5] and depth, with each nucleotide being read ~2300 times on average (Figures 8, 9; Table 8). In two independent experiments we found a set of 291 genomic regions (corresponding to 275 different genes) that were consistently mutated in *Ung*$^{-/-}$*Msh2*$^{-/-}$ GC B cells when compared to *Aicda*$^{-/-}$ GC B cells (FDR ≤ 0.05; Figure 10; Annex II). Moreover, we found a strong correlation between the mutation frequencies of the 1588 regions measured in the two biological replicates ($R^2$ = 0.86, Figure 10C left panel). This correlation is even stronger for AID targets ($R^2$ = 0.99, Figure 10C right panel). Thus, these results indicate that we have reproducibly found a set of 275 targets that are mutated by AID.

5 Sequencing quality refers to the probability of a base call being wrong, the higher quality, the lower probability of error.

| | *Ung*$^{+/-}$ *Msh2*$^{+/-}$ | | *Ung*$^{-/-}$ *Msh2*$^{+/-}$ | | *Ung*$^{+/-}$ *Msh2*$^{-/-}$ | | *Ung*$^{-/-}$ *Msh2*$^{-/-}$ | | *Aicda*$^{-/-}$ |
|---|---|---|---|---|---|---|---|---|---|
| | Exp1 | Exp2 | Exp1 | Exp2 | Exp1 | Exp2 | Exp1 | Exp2 | Exp1 |
| **Average depth (reads/nt)** [a] | 2581 | 2264 | 2100 | 2163 | 2303 | 2358 | 2175 | 2412 | 2303 |
| **Total length (Gb)** [b] | 2 | 1.75 | 1.62 | 0.94 | 1.78 | 1.82 | 1.68 | 1.86 | 1.78 |
| **Total length per region (Mb)** [c] | 1.25 | 1.1 | 1.02 | 1.05 | 1.12 | 1.15 | 1.06 | 1.17 | 1.12 |
| **NT>300 (%)** [d] | 96.4 | 96.4 | 95.7 | 95.9 | 96.2 | 96.2 | 96.1 | 96 | 96.2 |
| **NT>90Q20 (%)** [e] | 97.9 | 97.9 | 97.9 | 97.8 | 97.9 | 97.9 | 97.9 | 97.8 | 97.9 |

Table 8 | **Summary of depth and sequencing parameters of the capture libraries analyzed.** (**a**) Number of times each nucleotide of each of the 1588 captured regions was read on average. (**b**) Total number of bases read for the 1588 genomic regions analyzed. (**c**) Average number of bases read per captured region. (**d**) Average proportion of nucleotides read more than 300 times within each captured region. (**e**) Average proportion of nucleotides covered by more than 90% of reads passing Q20 quality threshold.

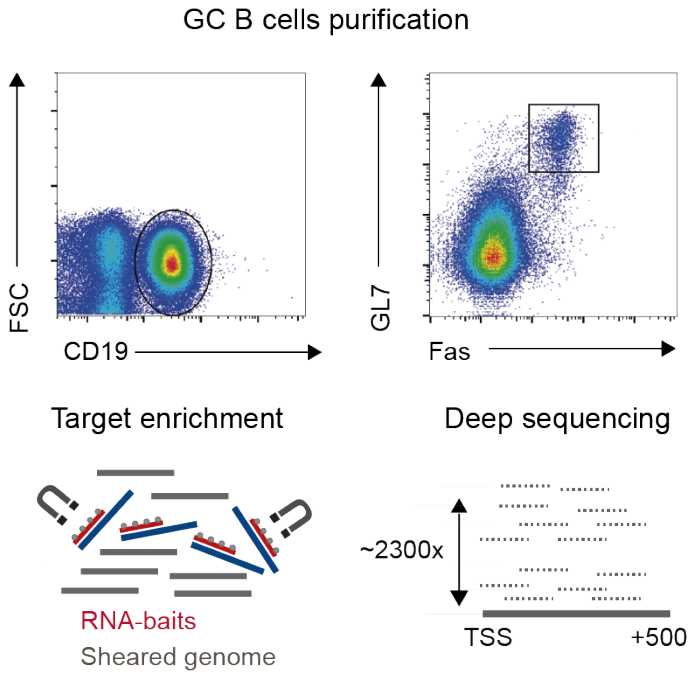



Figure 8 | **Schematic representation of the experimental approach.** GC ($CD19^{+}FAS^{+}GL7^{+}$) B cells from Peyer's patches were isolated by cell sorting and genomic DNA was extracted, sheared and captured with a custom library of RNA probes. Enriched DNA was subject to NGS to achieve a mean of 2300 reads per nucleotide.

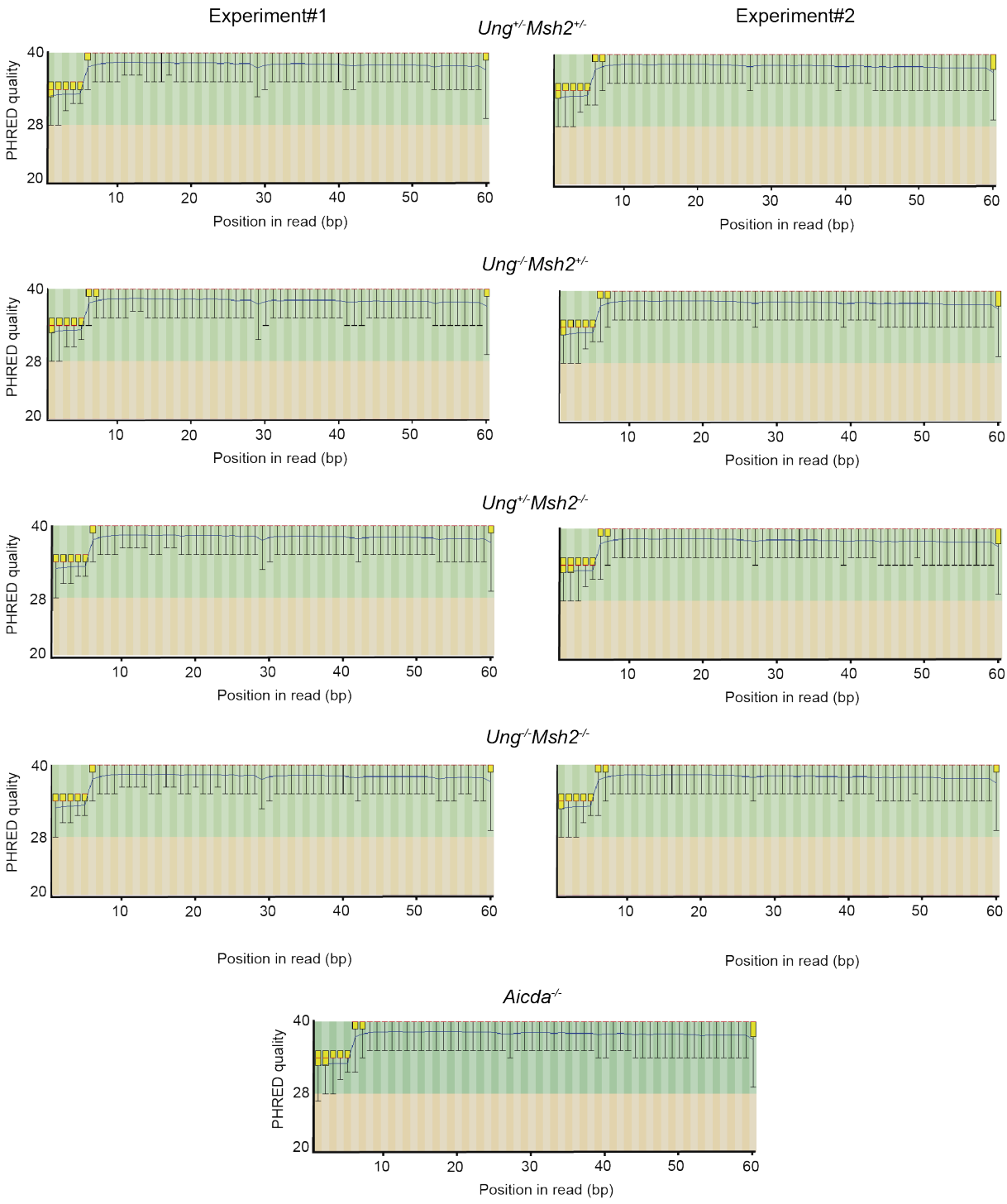



Figure 9 | **High quality sequencing of captured genes.** FastqQC plots showing average base calling quality along the 60nt reads in the libraries sequenced. A PHRED quality score between 28 and 40 corresponds to a base calling accuracy between 99,8% and 99,99%.

We found that the average C/G transition frequency of the 291 targets is close to $6.5x10^{-4}$, more than 4-fold above the background frequency of $1.5x10^{-4}$ detected in control cells (Figure 10D). Interestingly, mutations concentrate at the SHM hotspots WRC(Y)/(R)GYW (W = A/T; R = A/G; Y = C/T) (Rogozin and Kolchanov, 1992; Dörner et al., 1998; Rogozin and Diaz, 2004), where mutation frequencies are 8 to 12-fold above background (Figure 10D), reinforcing the idea that the mutations we detected are the result of AID mediated deamination.

To further validate our results, we performed Sanger sequencing of some of the genes we found mutated: *Hist1h1b*, *miR142*, *Cd19* and *Cdk4*. Primers were designed to amplify roughly the same region that we previously analyzed by NGS. We sequenced 50-100Kb (equivalent to 100-200 colonies; Table 9) per gene and found a significantly higher mutation frequency in *Ung*$^{-/-}$*Msh2*$^{-/-}$ than in control GC B cells in all the genes analyzed (Figure 10E).

| Gene | Genotype | Mutations[(a)] | Length (bp)[(b)] | Frequency[(c)] | P-value [(d)] |
|---|---|---|---|---|---|
| *Hist1h1b* | *Ung*$^{-/-}$*Msh2*$^{-/-}$ | 27 | 51000 | $5.3 \times 10^{-4}$ | $8.7 \times 10^{-8}$ |
| | *Aicda*$^{-/-}$ | 1 | 52000 | $1.9 \times 10^{-5}$ | |
| *miR142* | *Ung*$^{-/-}$*Msh2*$^{-/-}$ | 32 | 53000 | $6.0 \times 10^{-4}$ | $4.3 \times 10^{-8}$ |
| | *Aicda*$^{-/-}$ | 2 | 52250 | $3.8 \times 10^{-5}$ | |
| *Cd19* | *Ung*$^{-/-}$*Msh2*$^{-/-}$ | 14 | 107620 | $1.3 \times 10^{-4}$ | $1.5 \times 10^{-2}$ |
| | *Aicda*$^{-/-}$ | 1 | 56950 | $1.8 \times 10^{-5}$ | |
| *Cdk4* | *Ung*$^{-/-}$*Msh2*$^{-/-}$ | 15 | 110999 | $1.4 \times 10^{-4}$ | $4.3 \times 10^{-2}$ |
| | *Aicda*$^{-/-}$ | 2 | 54010 | $3.7 \times 10^{-5}$ | |

Table 9 | **Mutation analysis of representative AID targets in *Ung*$^{-/-}$*Msh2*$^{-/-}$ and *Aicda*$^{-/-}$ mice by Sanger sequencing**. (**a**) Total number of mutations detected. (**b**) Total number of bases sequenced. (**c**) Mutation frequency calculated as (Number of mutations / Number of bases sequenced). (**d**) Statistical comparison of the mutation frequency found in *Ung*$^{-/-}$*Msh2*$^{-/-}$ with that of *Aicda*$^{-/-}$ control mice (one-tail Fisher test).

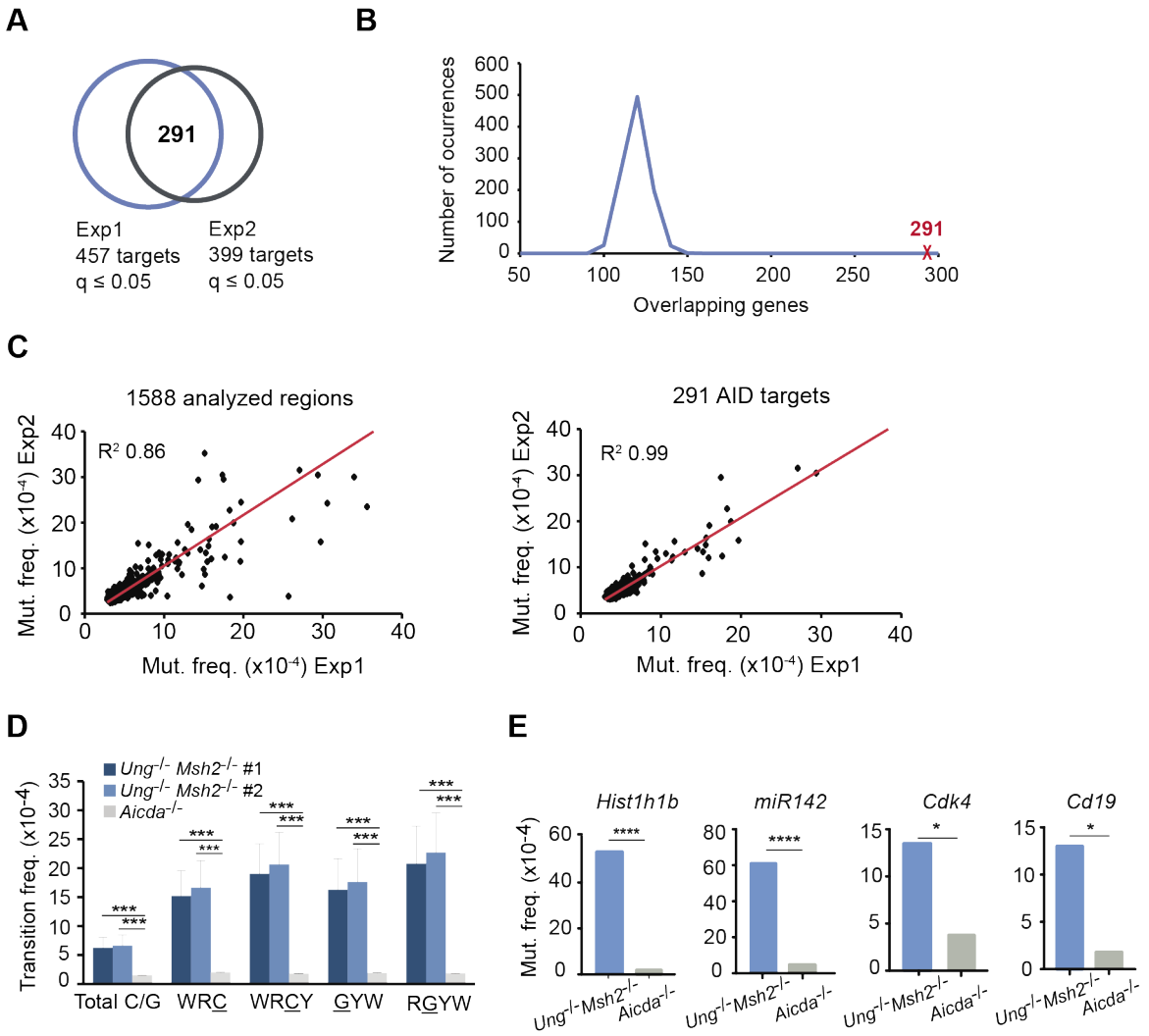



Figure 10 | **291 reproducible targets were detected by high-throughput analysis of AID-induced mutations.** **(A)** Two independent experiments were performed (Annex II) with 457 mutated targets found in Exp1 and 399 in Exp2. An overlap of 291 AID targets was found between Exp1 and Exp2. **(B)** *In silico* simulation to quantify the reliability of the 291 regions reproducibly found mutated in Exp1 and Exp2. Graph represents the experimental distribution of random overlaps after 1000 iterations. For each iteration, random groups of 457 and 399 genes were selected from the genes included in the SureSelect capture library, overlapped and the number of coincident genes reported. The probability to find an overlap of 291 genes by chance is below 1 out of each $10^{16}$ times tested. Two-tailed Fisher test; P ~ $10^{-16}$. **(C-D)** Comparison of mutation frequencies found in Exp1 and Exp2. **(C)** Mutation frequencies of the 1588 TSS proximal regions analyzed and the 291 targets found in two independent experiments. **(D)** Mean transition frequency in total C/G nucleotides and in C/G within WRCY/RGYW hotspots (W=A/T; R=A/G; Y=C/T) of the 291 AID targets (two-tailed Student's t-test; two independent experiments). **(E)** Validation of representative AID targets by Sanger sequencing (one-tail Student's t-test; Table 9). *, P ≤ 0.05; ***, P ≤ $10^{-3}$; ****, P ≤ $10^{-4}$. Error bars depict SEM.

Importantly, the collection of targets we identified includes 30 of the 35 previously known AID targets (Figure 11A), such as *Bcl6*, *Pim1*, *RhoH*, *Pax5* or *Cd83* (Shen et al., 1998; Pasqualucci et al., 1998; Müschen et al., JEM, 2000; Pasqualucci et al., 2001; Gordon et al., PNAS, 2003; Liu et al., 2008). We also performed overlap analysis of our identified AID mutational targets with published data on genome-wide analysis of DNA breaks (DSBs) and chromosome translocations induced by AID (Chiarle et al., 2011; Klein et al., 2011; Staszewski et al., 2011; Qian et al., 2014; Meng et al., 2014) (Figure 11B-G). Only a fraction of the targets in each of these studies was assessed in our capture-based SHM assay (1375 genes); but all 275 AID targets reported here were assayed in DSB and TC genome-wide studies (Figure 11B). Only 22 of the translocation sites described by Meng et al were included in our 1375-gene library. However, from those we found 19 (86%) genes mutated (Figure 11E, G). Likewise, 28 of the translocation sites described by Klein et al were included in our library, 21 of which (75%) were mutated (Figure 11F, G). In contrast, the fraction of mutated genes found in our study that undergo either DSBs or chromosome translocations is much smaller. For instance, only 19 of the 275 AID targets identified by us (7%) undergo chromosome translocations according to Meng et al. (Figure 11E, G). This likely reflects that while most of the DSBs / TCs detected in these studies come from AID-induced mutations in off-targets, not all AID mutations give raise to DSB / TC events.

We conclude that our capture-based deep sequencing approach has allowed the discovery of an unprecedented, massive collection of AID targets.

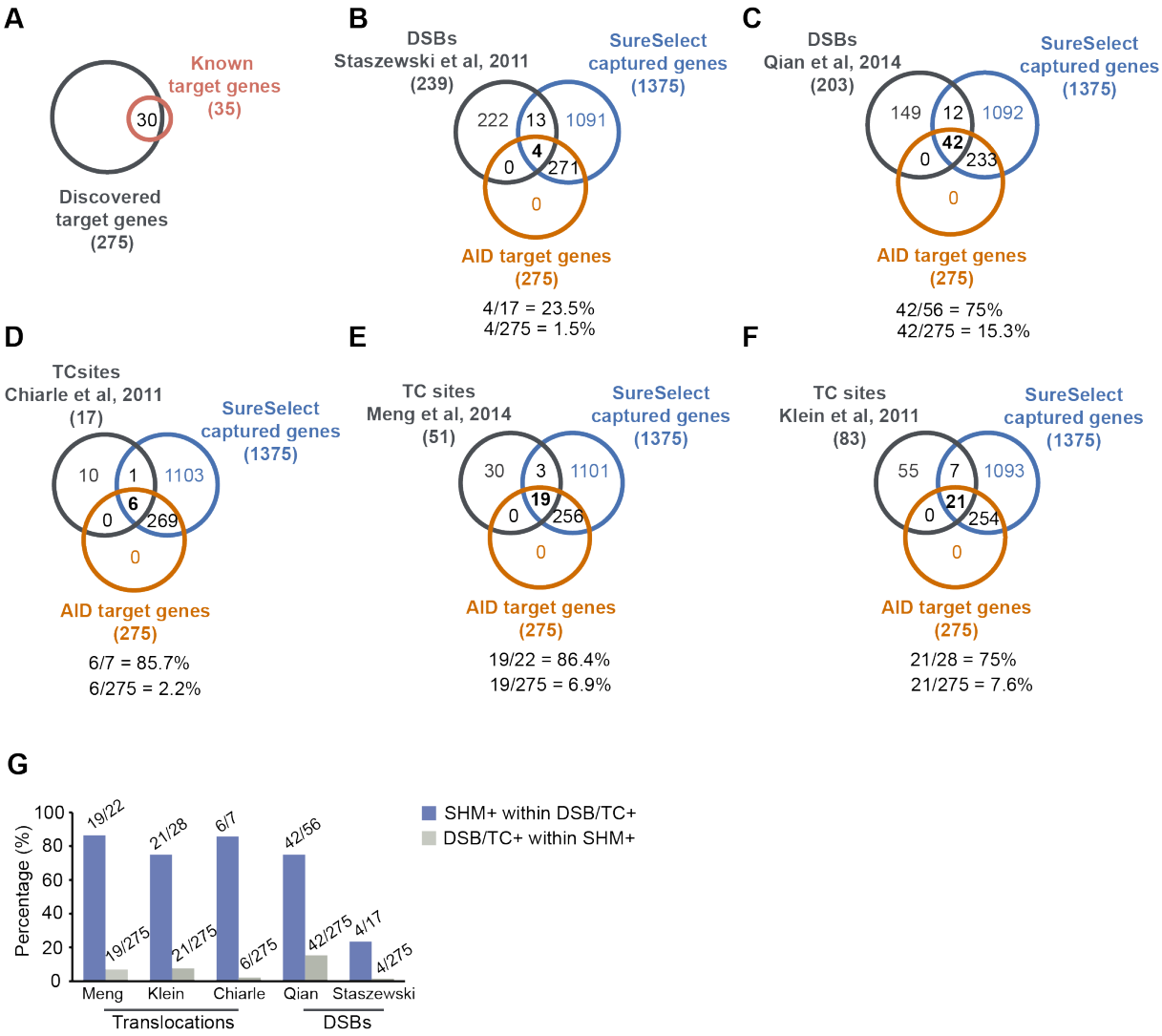



Figure 11 | **Overlap analysis of the 291 targets discovered in this study and previously published data on AID-induced mutations, translocations and double strand breaks. (A)** Overlap between the targets discovered in this study and previously reported AID targets. (**B-F**) Venn diagrams showing overlaps between the genes included in our capture library (SureSelect captured genes, n=1375), AID targets identified in this study (n=275) and genes undergoing DSB or TCs according to the indicated studies. **(G)** Percentage of genes undergoing DSB/TCs according to the indicated studies within the AID targets described in this study ($SHM^+$; n=275) and percentage of $SHM^+$ genes within $DSB/TC^+$ genes (see Materials and methods for further details).

## 2.2. Analysis of the local specificity of AID

To gain insights into the local sequence preference of AID, we first analyzed the mean mutation frequency at individual WRCY/RGYW hotspots across all 291 AID targets and found a wide range of mutability, with AACT and AGCT as the top mutated hotspots in both strands of DNA (Figure 12). As these results were obtained in the combined absence of UNG and MSH2, we think this may reflect an intrinsic preference for AID to deaminate cytosine residues lying within these motifs. Interestingly, this result is consistent with previously published studies. For instance, using heterologous and *ex vivo* systems to test SHM, our lab previously found AGCT as the most mutated hotspot in a transgene and in the IgH Sμ region (Pérez-Durán et al., 2012). In addition, Yeap and colleagues identified AGCT as the most robust SHM hotspot in the absence of antigen selection in an *in vivo* model (Yeap et al., 2015). This indicates that AGCT is a strong hotspot for AID local specificity. Next, we performed an unbiased analysis of the sequence context of mutated cytosines in the 291 AID targets. We found that A, G and T nucleotides were the preferred nucleotides at -2, -1 and +1 positions, respectively, but we further uncovered a significant preference for T at +3 (Figure 13A). Indeed, cytosines lying at the AGCTNT motif were significantly more mutated than those in AGCTNV (where V is A, C or G) or than other WRCY/RGYW hotspots (Figure 13B, C). To analyze how frequently this novel hotspot is mutated during SHM, we calculated the percentage of cytosines lying within the AGCTNT context bearing 1 or more mutations. Throughout the AID target sequences covered by our capture library (roughly 145Kb in 291 different regions) AGCTNT hotspot is found 134 times and we found mutations in 104 of them (78%). For the sake of comparison, we performed the same analysis for AGCTNV (254/605, 42% of mutated cytosines) and for two non-hotspot 4-nucleotide motifs, CTCA (146/673, 22% of mutated cytosines) and GGCA (229/1480, 16% of mutated cytosines), and found that AGCTNT is mutated more frequently than AGCTNV or any of the non-hotspot sequences (Figure 13D). We did not find a difference in mutation frequency or in the proportion of sequences mutated between sense and antisense strands (AGCTNT vs ANAGCT, Figure 13C), suggesting that the preference for AID to mutate cytosines within the AGCTNT motif is not dependent on the DNA strand.

Together, these results show that AGCTNT is a novel and the most highly mutated AID hotspot identified so far.

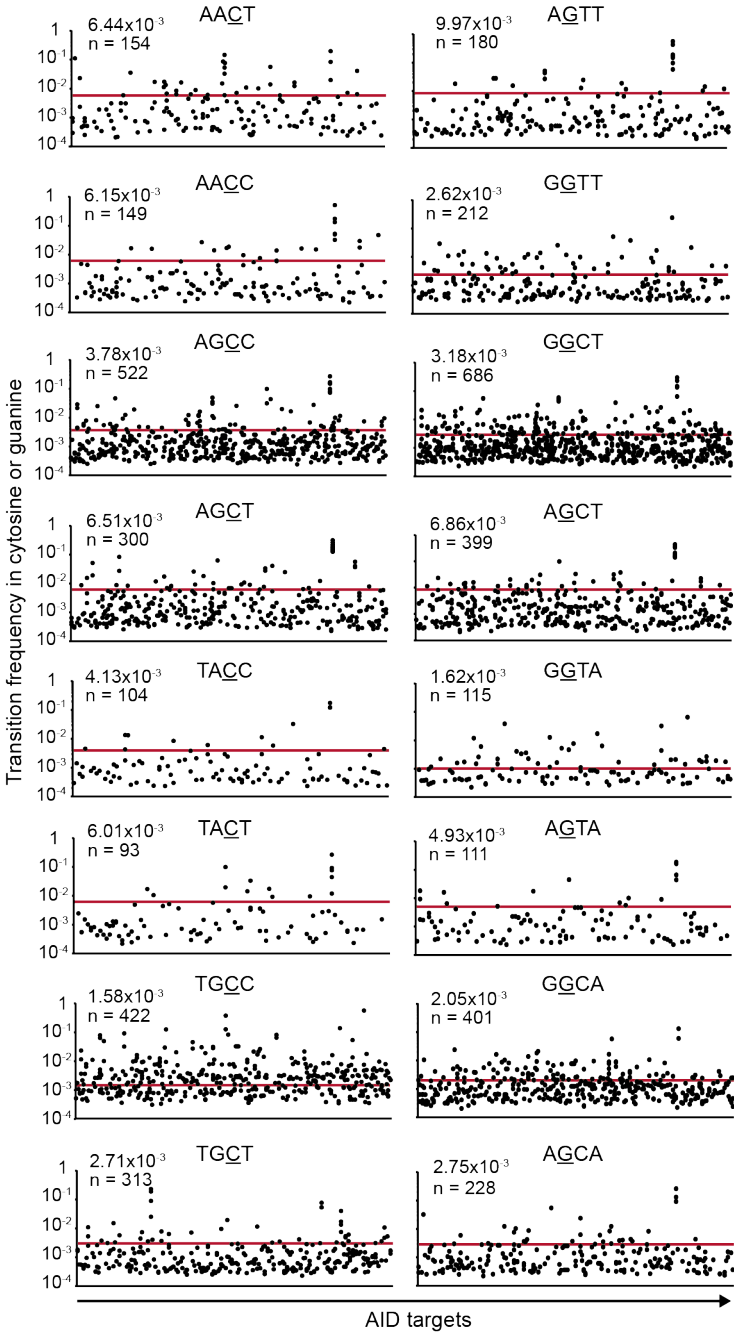



Figure 12 | **Mutation analysis at WRCY/RGYW hotspots in *Ung*$^{-/-}$*Msh2*$^{-/-}$ GC B cells**. Plots show mutated individual hotspots (WRCY, left; RGYW, right). Each dot represents an individual cytosine (C) or guanine (G) within a WRCY/RGYW motif found mutated at least once. Each position in the X axis corresponds to a different gene, and the Y axis shows mutation frequency at the underlined C or G of each individual hotspot within a gene. Mean mutation frequency is indicated and depicted with a red line. Number of mutated hotspots is indicated.

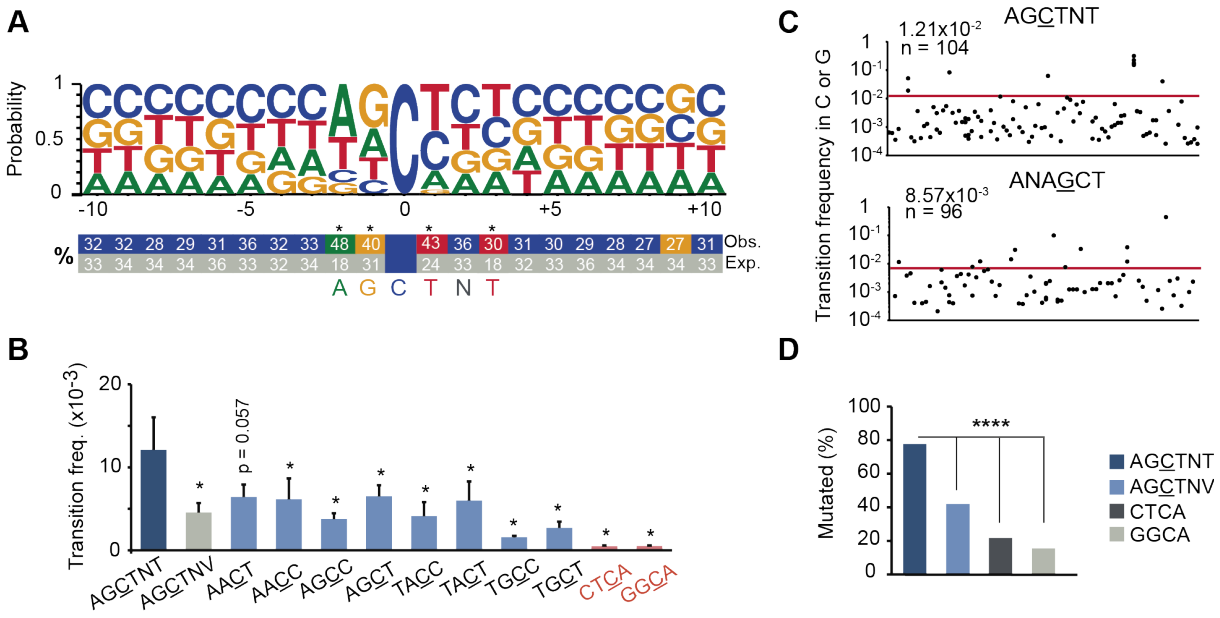



Figure 13 | **AGCTNT is a novel AID mutational hotspot. (A)** Logo representation of the sequence context of mutated cytosines (mutation frequency ≥ $4x10^{-3}$). Statistically significant enrichment of nucleotides surrounding the mutated C is indicated (One-tail Fisher test and Bonferroni correction; *, FDR ≤ $10^{-3}$; see Materials and methods for details), and numbers indicate percentages. **(B)** Mean mutation frequency of cytosines within the indicated motifs (dark blue bar, newly identified hotspot; gray bar, control motif for newly identified hotspot; light blue bars, WRCY hotspots; red bars, random four-nucleotide motifs; two-tailed Mann-Whitney test; *, P ≤ 0.05; error bars depict SEM). **(C)** Mutation analysis of AGCTNT and ANAGCT motifs. Within each plot, each dot represents an individual AGCTNT/ANAGCT motif found mutated at least once. Each position in the X axis corresponds to a different gene, and the Y axis shows mutation frequency of each individual hotspot within a gene. Mean mutation frequency is indicated and depicted with a red line. Number of mutated hotspots is indicated. **(D)** Percentage of mutated cytosines within AGCTNT and AGCTNV hotspot motifs and CTCA and GGCA non-hotspot motifs (Fisher test; ****, P ≤ $10^{-13}$).

## 2.3.- Molecular characterization of AID targets

The uniquely large set of AID-mutated genes identified in this study allows, for the first time, a high throughput examination of the molecular features that associate to SHM and the evaluation of their potential role defining AID target specificity. Since AID activity in the Ig loci has been classically linked to transcription, we performed a comprehensive analysis of transcription-related features of AID targets, including steady-state transcription levels, transcription rate, epigenetic marks and regulatory sequences.

### 2.3.1. AID targets are highly transcribed

Given the link between AID activity and transcription (Betz et al., 1994; Chaudhuri et al., 2003; Fukita et al., 1998; Peters and Storb, 1996; Pham et al., 2003; Ramiro et al., 2003; reviewed in Storb, 2014) we evaluated the transcriptional state of the 1375 genes included in our capture library. Whole transcriptome sequencing (RNA-Seq) was performed in sorted GC ($CD19^+FAS^+GL7^+$) and naïve ($CD19^+FAS^-GL7^-$) B cells from Peyer's patches of WT mice littermates. We carried out differential expression (DE) analysis in GC versus naïve B cells and identified 8868 genes that are DE in the two populations (FDR ≤ 0.05) (Figure 14A). This is consistent with previously published RNA-Seq data on lymph node GC vs naïve B cells (Kuchen et al., 2010) where a similar number of DE genes was reported and a big proportion of them (~90%) match the DE genes identified by our study (data not shown). In our experiment, approximately half of the DE genes were upregulated (4412 genes; 49.7%) and the other half were downregulated (4470 genes; 50.3%) in GC versus naïve B cells (Figure 14A,B). From the 4412 genes significantly upregulated in GC B cells, 407 were included in our capture library and thus assayed for mutations. Notably, approximately 33% of them (133/407) are mutated by AID, suggesting that GC specific genes are frequently affected by AID off-targeting. However, this group of genes only accounts for 48% of the 275 AID targets we identified. This means that AID off-site activity is not restricted to genes specifically activated during the GC reaction, but also affects other transcriptionally active genes. Indeed, the remaining 52% AID targets are well expressed both in naïve and activated states. In agreement with these results, we found that AID targets are significantly more expressed than non-targets and that this difference is even larger for highly mutated targets (Figure 14C).

Next, we analyzed publicly available data on Global Run-On Sequencing (GRO-Seq) assays from mouse GC B cells to estimate the transcription rate of AID targets and non-targets (see Materials and Methods for details). In contrast to RNA-Seq, which quantifies RNA steady state levels, GRO-Seq quantitatively measures nascent RNA transcripts at the genome-wide scale by a pulse and chase approach. With this technique, modified NTPs are incorporated during transcriptional elongation, which are then captured allowing specific sequencing of newly synthesized mRNA. Thus, GRO-Seq provides a snapshot of the total number of engaged transcriptional complexes in the cell, and allows the quantification of the rate at which a particular gene is transcribed in a fashion that is independent of the stability of the resulting mRNA. In line with the results obtained from RNA-Seq data, we found that AID targets are transcribed at significantly higher rates than non-targets, and that the difference is larger when we compare highly mutated to non-mutated genes (Figure 14D).

In conclusion, these results indicate that AID targets are highly transcribed genes.

### 2.3.2. AID targets recruit high levels of RNAP II and SPT5

A few years ago, the Nussenzweig lab reported that AID interacts with SPT5 (Pavri et al., 2010) which facilitates the association between AID and RNAP II. They proposed that AID takes advantage of RNAP II stalling to gain access to ssDNA and that the concomitant decrease in the elongation rate provides increased time of residence for AID and favors its mutagenic activity. To further explore the link between AID induced mutation and RNAP II / SPT5 recruitment, we analyzed publicly available ChIP-Seq data from splenic LPS+IL4 activated B cells. Since this *ex vivo* setting differs from the *in vivo* context in which we identified AID targets, we first compared the transcription levels of the genes included in our capture library in LPS+IL4 activated B cells and in GC B cells. We found a good correlation between the two expression profiles, so we reasoned that RNAP II and SPT5 binding data from LPS+IL4 could be well extrapolated to the GC B cells used in our study (Figure 15A). Thus, we integrated LPS+IL4 ChIP-Seq data with our collection of captured genes and found a significantly higher binding density of RNAP II and SPT5 to AID targets than to non-targets. Furthermore, we detected the highest binding of RNAP II and SPT5 in highly mutated genes (Figure 15B, C). This agrees with our data on expression and transcription rate of AID target genes (Figure 14C, D). Hence, AID targets highly transcribed genes with high density binding of RNAP II and SPT5.

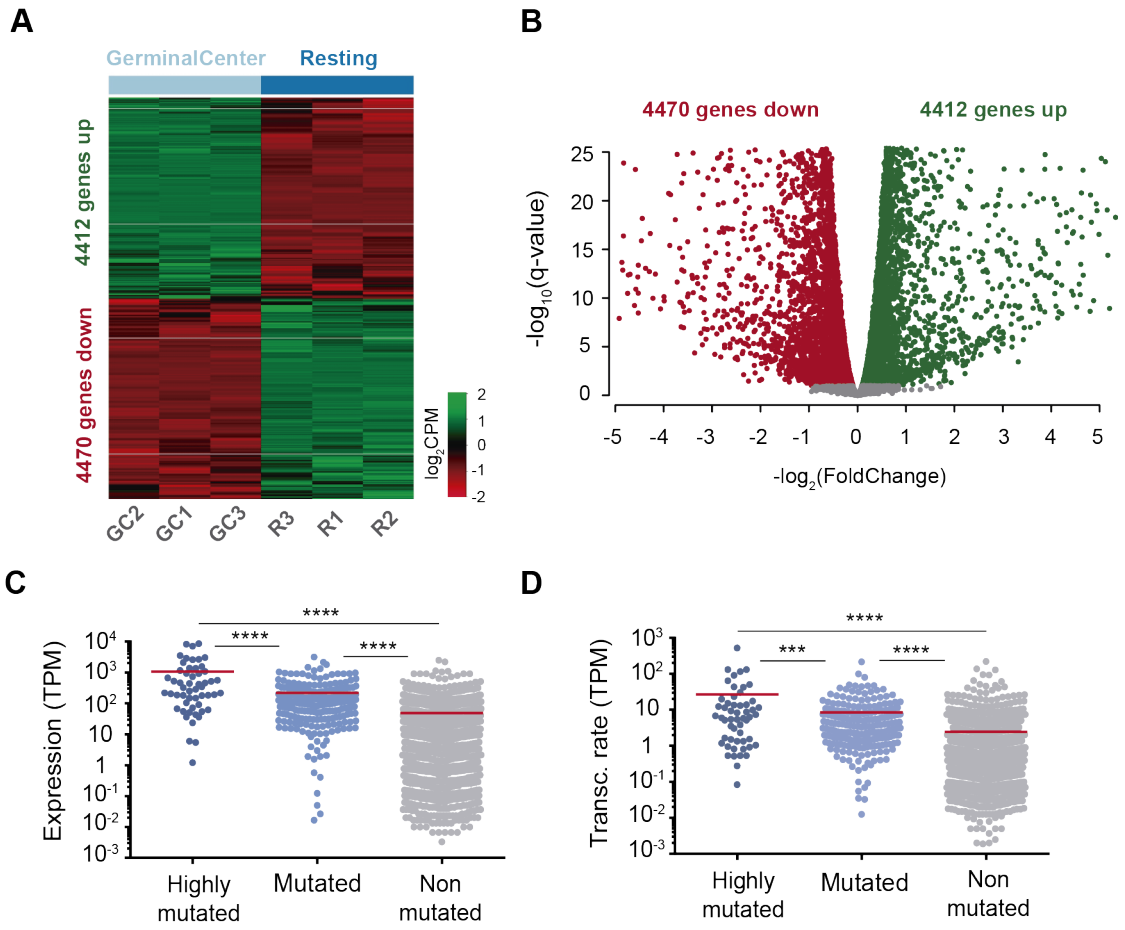



Figure 14 | **AID targets are highly transcribed. (A)** Heatmap representation of the differential expression analysis of Peyer's patch GC and naïve B cells by RNA-Seq. Three biological replicates were analyzed per condition (n=5 each); genes were considered as differentially expressed at FDR ≤ 0.05 (see Materials and methods for details) **(B)** Volcano plot representation of the differential expression analysis depicted in panel A. Green dots, genes significantly upregulated in GC vs naïve B cells (n=4412); red dots, genes significantly downregulated in GC vs naïve B cells (n=4470); Grey dots, unchanged genes (n=2638). **(C)** Expression level of highly mutated (top 20% mutated genes, C→T transition frequency > $3x10^{-4}$), mutated (rest of mutated) and non-mutated genes in Peyer's patch GC B cells as measured by RNA-Seq. **(D)** Transcription rate of highly mutated, mutated and non-mutated genes in GC B cells from lymph nodes as measured by GRO-Seq. TPM, Transcripts Per Million.

### 2.3.3. AID targets are enriched in marks associated to active enhancers and transcription elongation

Mediator is a large protein complex that is generally required for transcription by RNAP II in eukaryotes. The Mediator complex interacts with RNAP II and enhances its recruitment stabilizing transcription complexes at gene promoters (Soutourina, 2018; Taatjes, 2010). The main function of Mediator is exerted at enhancers, where it transduces signals from transcription activators to control transcription initiation. Mediator forms a complex with cohesin to promote DNA loops that connect enhancers and gene promoters (Kagey et al., 2010). Thus, we measured MED12 (a subunit of the mediator complex) binding as a marker of active enhancers in B cells. ChIP-Seq data from LPS+IL4 activated B cells revealed that MED12 recruitment is ~3 times more frequent in AID targets than in non-targets (Figure 16). This suggests that Mediator binding to gene enhancers can favor AID activity, very likely by aiding to the recruitment of RNAP II to gene promoters and thus facilitating transcriptional activation.

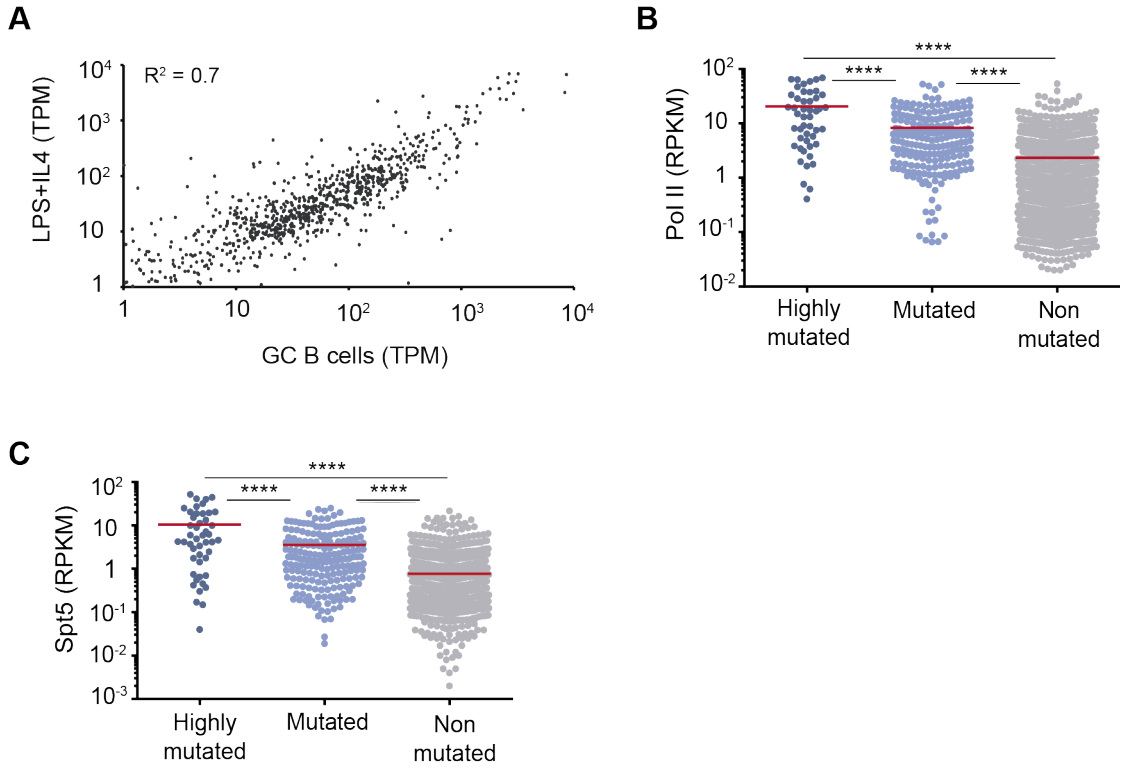



Figure 15 | **AID targets show high density binding of RNAP II and SPT5. (A)** Correlation between the expression profiles of the 1375 genes included in SureSelect capture library in Peyer's patch GC B cells (X axis, this study) and LPS+IL4 activated B cells (Y axis, Pérez-García et al., 2017) as measured by RNA-Seq. TPM, Transcripts Per Million. **(B)** Recruitment of RNAP II and **(C)** SPT5 to AID targets and non-targets measured in in vitro activated splenic B cells by ChIP-Seq. RPKM, Reads Per Kilobase per Million reads mapped.

Epigenetics also plays a pivotal role in the regulation of gene expression. Epigenetic modifications are reversible changes of chromatin structure that do not affect the primary sequence of DNA, but regulate transcriptional activation or repression. They can directly affect DNA (DNA methylation) or DNA-associated proteins (covalent modification of histone proteins). Here, we evaluated the presence of epigenetic marks in AID targets and non-targets. We analyzed publicly available datasets on ChIP-Seq experiments measuring marks related to active transcription and transcription elongation in B cells (Kieffer-Kwon et al., 2013; ENCODE/LICR project). Specifically, we looked at the following modifications at histone H3: monomethylation of lysine four (H3K4me1, mark of active promoters), trimethylation of lysine thirty-six (H3K36me3, mark of elongation) and dimethylation at lysine seventy-nine (H3K79me2, mark of elongation). Interestingly, we found a significant increase in the proportion of AID targets that bear epigenetic marks of active elongation (H3K36me3 and H3K79me2) when compared to non-targets (Figure 16). These results are in accordance with previously published data reporting the occurrence of specific epigenetic marks in genes affected by AID-mediated translocations, where authors propose that epigenetic features mediate AID recruitment to off-target sequences (Wang et al., 2014)

In conclusion, AID targets are enriched in marks of active enhancers and transcriptional elongation.

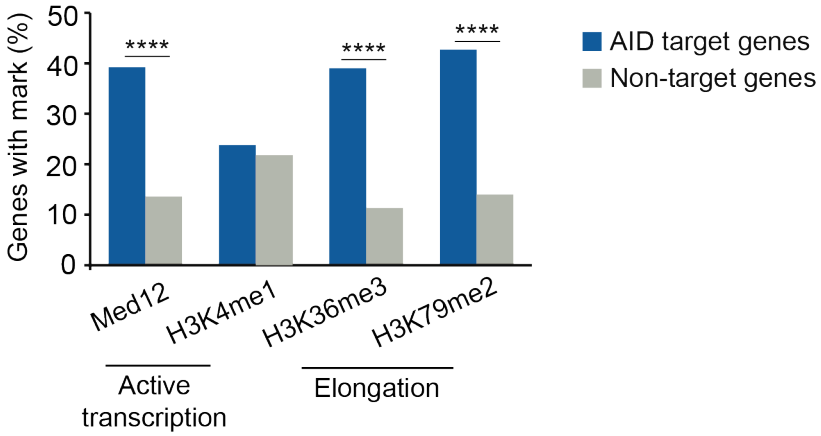



Figure 16 | **AID targets are enriched in marks associated to transcription and transcription elongation.** Transcription and transcription elongation marks in AID targets and non-targets by ChIP-Seq analysis of in vitro activated splenic B cells (MED12, H3K4me1, H3K36me3, H3K79me2).

2.3.4.- AID targets are regulated by superenhancers and frequently undergo convergent transcription

Superenhancers (SE) are clusters of enhancers that recruit lineage specific transcription factors and can establish long range interactions, thus regulating both proximal and distal loci (Hnisz et al., 2013). SE have been recently linked to AID off-targeting, because they initiate antisense transcription within sense transcribed genes, which provokes the collapse of the transcription machinery due to convergent transcription (ConvT). This may contribute to AID targeting in three different ways: first, ConvT results in RNAP II stalling, which can lead to AID recruitment (Pavri et al., 2010); second, ConvT generates ssDNA substrates which are amenable to AID action (Meng et al., 2014; Qian et al., 2014); third, ConvT contributes to the recruitment of the RNA exosome which helps AID access ssDNA (Basu et al., 2011; Pefanis et al., 2014). In light of these findings, we wondered whether the AID targets identified in this work are regulated by SE or undergo ConvT. With that aim, we integrated the data on SE and ConvT published in Meng et al., 2014 and Qian et al., 2014 with our mutational data. Interestingly, we found that primary AID targeting, as measured by AID mutations in the absence of repair, focuses preferentially in the vicinity of superenhancers (Figure 17A) and in regions subject to convergent transcription (Figure 17B) (Meng et al., 2014; Qian et al., 2014).

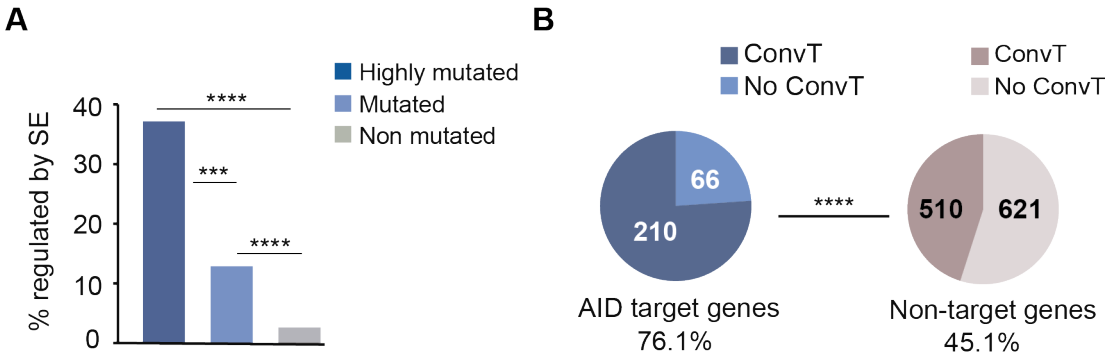



Figure 17 | **AID targets are regulated by superenhancers and frequently undergo convergent transcription. (A)** Proportion of highly mutated, mutated and non-mutated genes regulated by superenhancers in GC B cells (see Materials and Methods for details). **(B)** GRO-Seq analysis of convergent transcription (ConvT) in AID targets and non-targets from GC splenic B cells obtained from SRBC-immunized mice.

Together, the results presented in this chapter show that several transcription-associated events are linked to AID activity, and thus provide proof for this previously reported idea (Meng et al., 2017; Nambu et al., 2003; Pavri et al., 2010; Qian et al., 2014; Storb, 2014; Wang et al., 2014) in a broad collection of targets (Figure 18). However, our data also indicate that AID targeting cannot be defined by any of these features alone.

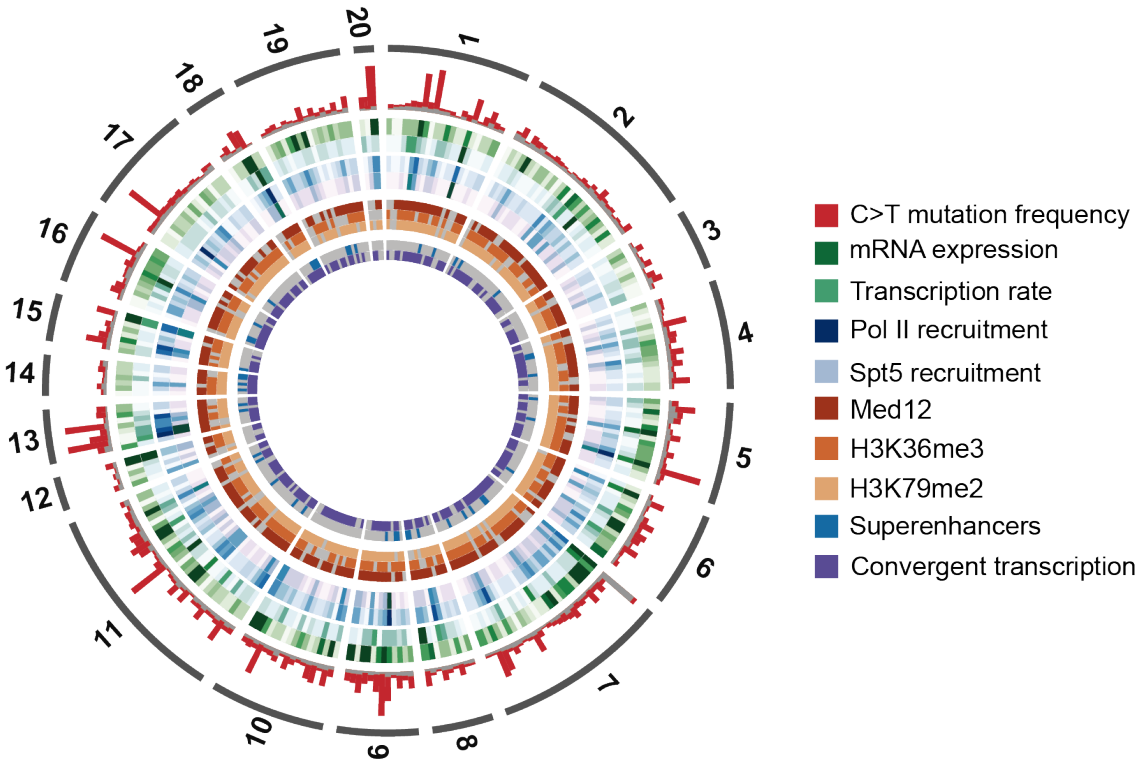



Figure 18 | **Molecular features of AID targets.** Circos plot representation of the AID targets identified in this study and their associated molecular features. The outer ring shows chromosome location and is followed by C→T transition frequency in *Ung*$^{-/-}$*Msh2*$^{-/-}$ (red) and *Aicda*$^{-/-}$ (grey) GC B cells; mRNA expression (dark green); Transcription rate (light green); RNAP II (dark blue) and SPT5 (light blue) binding; MED12 (dark red), H3K36me3 (dark orange), H3K79me2 (light orange) marks; regulation by superenhancers (blue) and ocurrence of convergent transcription (purple).

## 3. Prediction of AID targets in Germinal Center B lymphocytes

Given our finding that there are several features that associate to SHM (Figure 18), we reasoned that, conceivably, a combination of them could be used to predict AID targeting. To approach this hypothesis, we made use of a supervised machine learning algorithm known as recursive partitioning (Tom Mitchell, 1997; Trevor Hastie, Robert Tibshirani, Jerome Friedman, 2009; see Materials and Methods for details). In brief, the algorithm uses a training dataset[6] to generate a classification tree that splits data into different nodes based on a collection of variables. This partitioning allows the identification of the variables that better define each of the initial groups to be classified. There are some advantages to using recursive partitioning trees: they are non-parametric and allow the incorporation of categorical data; they are robust with regards to outliers in the training data and they are very easy to interpret visually (Chapter 3, Tom Mitchell, 1997; Chapters 9-10, Trevor Hastie, Robert Tibshirani, Jerome Friedman, 2009). However, there are also some drawbacks that need to be taken into account: they cannot make predictions beyond the upper and lower limits established by the training data and they tend to overfit (Chapter 3, Tom Mitchell, 1997; Chapters 9-10, Trevor Hastie, Robert Tibshirani, Jerome Friedman, 2009; Libbrecht and Noble, 2015). A good prediction model learns patterns from the training data and is able to generalize them on new data. Overfitting occurs when the model learns patterns that are too specific of the training data and cannot be extrapolated to make reliable predictions.

The recursive partitioning algorithm was trained with a dataset composed of two different groups, AID target (272) and non-target genes (1067). For each of the genes we included the 10 variables measured in chapter 2.3: gene expression, transcription rate, binding of transcription cofactors (RNAP II, SPT5, MED12) and epigenetic marks (H3K36me3, H3K79me2, H3K4me1), regulation by SE and occurrence of ConvT. To control for overfitting we used tree pruning (Chapter 3, Tom Mitchell, 1997; Chapters 9-10, Trevor Hastie, Robert Tibshirani, Jerome Friedman, 2009). This technique removes the nodes of the tree that are less relevant for the classification and reduces the complexity of the final classifiers. If the model is initially overfitted, tree pruning will produce a different classification tree with improved prediction accuracy. Notably, identical trees were produced before and after pruning, which indicates that our prediction model was not initially overfitted. The outcome of the machine learning approach

6 The training dataset is the information used by the algorithm to fit the parameters of the classification model, in other words, the set of examples from which the algorithm "learns".

revealed that the combination of high-density RNAP II and SPT5 binding predicts AID targeting with a 77% probability (P < 0.001; Figure 19). This combination of features is found in 2.3% of the genes (~430 genes) in the whole genome (Figure 19B; AnnexIII). Other combinations bear some predictive power as well but are much less efficient classifying AID targets. Conversely, low RNAP II binding combined with low gene expression predicted the absence of mutation in 95% of the genes (Figure 19).

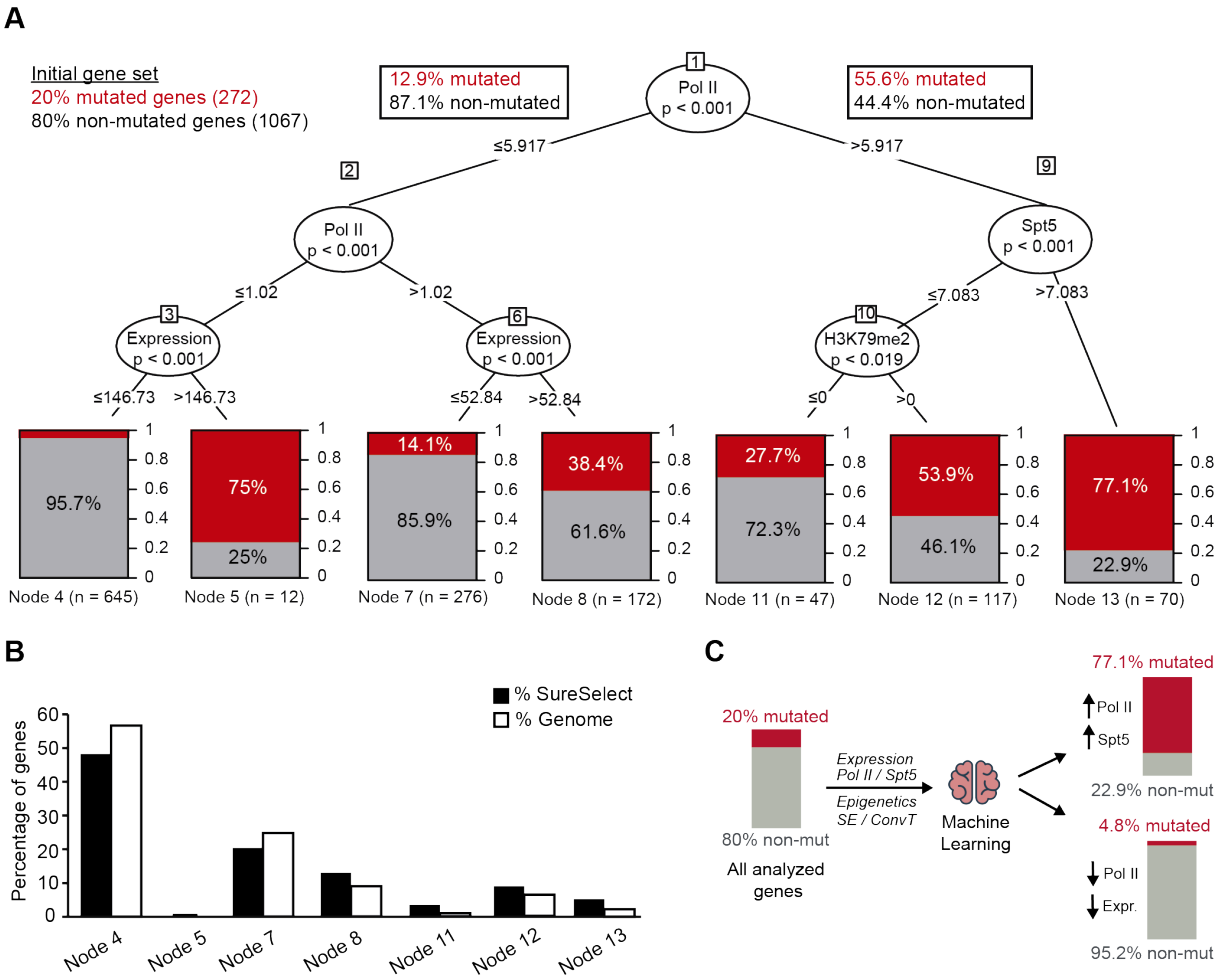



Figure 19 | **Molecular features of AID predict mutability. (A)** Recursive partitioning tree model classifies AID targets based on different molecular features: mRNA expression, RNAP II and SPT5 recruitment, and presence of H3K79me2 epigenetic mark (see Materials and Methods for details). Each node splits the genes into two significantly different groups based on a particular feature. Numbers within the branches indicate the thresholds used to make the groups; P-values of each decision are included below the parameter measured in each node. **(B)** Bar graph depicting the proportion of SureSelect genes (1339 genes, closed bars) or of total genes in the mouse genome (17858 genes; open bars) that meet the thresholds established in each node. **(C)** Simplified schematic representation of the machine-learning approach used for AID target prediction.

To experimentally validate the accuracy of the prediction model, we randomly picked a collection of genes (not included in our capture library) with high-density RNAP II and SPT5 binding (Figure 20A) and analyzed their mutation profiles by PCR-Seq. We found that 11/12 of the analyzed genes were significantly mutated, with mutations focusing at AID mutational hotspots (Figure 20B; Table 10). Indeed, two of those genes (*Hist1h1c* and *Clec2d*) were mutated at the range of the top 20% mutated genes at frequencies similar to those found in *Pax5* or *RhoH* (Table 10; Annex II). Thus, we have built a powerful predictive tool for AID activity that identified high RNAP II and SPT5 binding as the best predictors of AID specificity.

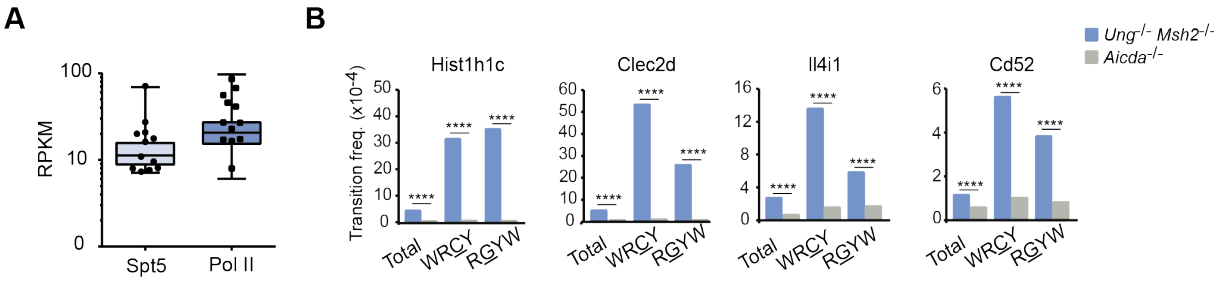



Figure 20 | **Experimental validation of the machine-learning model by PCR-Seq. (A)** Box plot represents genome-wide data of RNAP II and SPT5 recruitment in in vitro activated B cells. Black dots and squares depict the 12 genes selected for the validation of the model prediction. RPKM, Reads Per Kilobase per Million reads mapped. **(B)** Validation of representative genes predicted to be mutated by the model by PCR-Seq (Table 10). Two-tailed Student's t-test, ****P ≤ $10^{-4}$.

| | | | TS freq. at C/G (x$10^{-5}$)[a] | | WRCY TS freq. (x$10^{-5}$)[b] | | RGYW TS freq. (x$10^{-5}$)[c] | | |
|---|---|---|---|---|---|---|---|---|---|
| **Gene** | **POLII (RPKM)** | **SPT5 (RPKM)** | ***Ung$^{-/-}$Msh2$^{-/-}$*** | ***Aicda$^{-/-}$*** | ***Ung$^{-/-}$Msh2$^{-/-}$*** | ***Aicda$^{-/-}$*** | ***Ung$^{-/-}$Msh2$^{-/-}$*** | ***Aicda$^{-/-}$*** | ***FDR*** |
| *Apobec3* | 26.17 | 10.67 | 36.54 | 27.69 | 210.08 | 130.38 | 392.08 | 271.37 | 0.56 |
| ***Aurkaip1*** | 16.74 | 7.88 | 13.88 | 8.77 | 75.31 | 11.13 | 21.58 | 18.90 | **$4.01x10^{-151}$** |
| ***Ccdc17*** | 16.10 | 7.46 | 6.28 | 5.73 | 13.48 | 11.91 | 14.91 | 12.15 | **$3.26x10^{-9}$** |
| ***Cd52*** | 26.43 | 9.32 | 11.54 | 5.81 | 56.15 | 10.19 | 38.29 | 8.21 | **0** |
| ***Cd68*** | 40.33 | 20.30 | 7.43 | 6.98 | 16.24 | 10.99 | 13.08 | 9.17 | **$2.58x10^{-7}$** |
| ***Cd69*** | 7.76 | 17.26 | 4.64 | 3.95 | 11.56 | 10.35 | 13.17 | 10.39 | **$7.98x10^{-13}$** |
| ***Clec2d*** | 22.32 | 19.53 | 51.11 | 4.39 | 532.38 | 9.70 | 257.45 | 7.19 | **0** |
| ***Coro1a*** | 44.71 | 15.77 | 7.81 | 6.47 | 14.89 | 11.90 | 31.95 | 10.13 | **$3.28x10^{-30}$** |
| ***Hist1h1c*** | 84.00 | 69.40 | 60.99 | 8.38 | 412.43 | 11.42 | 460.45 | 10.10 | **0** |
| ***Il4i1*** | 54.63 | 7.14 | 27.22 | 6.73 | 135.76 | 15.75 | 58.41 | 16.92 | **$1.38x10^{-8}$** |
| ***Maz*** | 16.97 | 7.76 | 5.99 | 5.44 | 14.47 | 8.20 | 11.60 | 9.14 | **$4.19x10^{-4}$** |
| ***Trex1*** | 65.81 | 26.68 | 7.07 | 6.26 | 15.74 | 11.94 | 15.19 | 11.84 | **$1.58x10^{-32}$** |

Table 10 | **List of genes selected for machine learning validation.** (**a**) Transition (TS) mutation frequency at C/G nucleotides calculated as (Total TS mutations in C + Total TS mutations in G) / (Total sequenced C + Total sequenced G) (**b**) Transition mutation frequency at cytosines contained in WRCY hotspots calculated as (Total TS mutations in WRCY) / (Total sequenced C in WRCY) (**c**) (c) Transition mutation frequency at guanines containted in RGYW hotspots calculated as (Total TS mutations in RGYW) / (Total sequenced RGYW).

## 4. Analysis of the role of the Base Excision Repair (BER) and Mismatch Repair (MMR) pathways in the resolution of AID-induced deaminations

BER and MMR act downstream of AID-induced U:G mismatches so that UNG is critical for the generation of transversions at C:G pairs while MSH2 facilitates the introduction of mutations at A:T pairs (Frey et al., 1998; Methot and Di Noia, 2017; Phung et al., 1998; Rada et al., 1998, 2002, 2004). UNG and MSH2 can also promote conventional, faithful repair of AID-induced U:G mismatches (Pérez-Durán et al., 2012 Roa et al., 2010; Liu et al., 2008). Indeed, our lab previously reported that the local sequence context can confer specificity to the resolution of AID-induced lesions by UNG, directing it towards an error-free or error-prone repair outcome. More specifically, AA<u>C</u>T, TA<u>C</u>T, A<u>G</u>TA, A<u>G</u>TT, G<u>G</u>TA and G<u>G</u>TT motifs were shown to drive faithful repair by UNG while AG<u>C</u>T, AGCT and AG<u>C</u>A promoted error-prone repair by UNG in IgH Sμ region (Pérez-Durán et al., 2012).

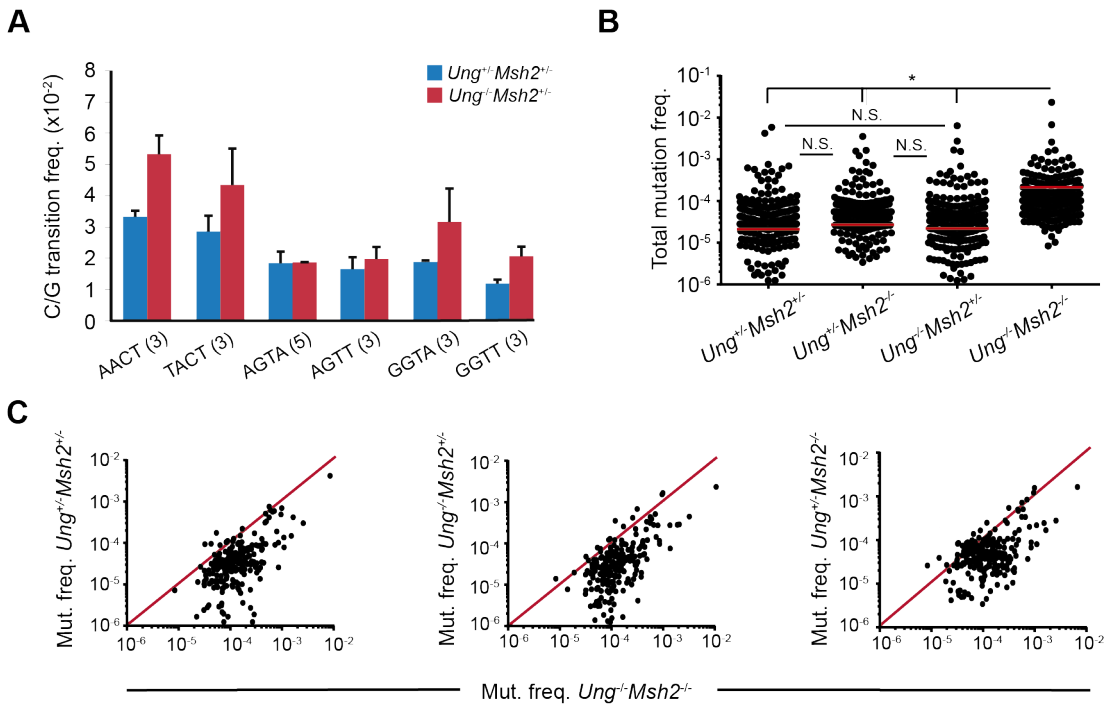



Figure 21 | **BER and MMR back up each other to error-free repair AID-induced lesions. (A)** Mutation frequency of AA<u>C</u>T, TA<u>C</u>T, A<u>G</u>TA, A<u>G</u>TT, G<u>G</u>TA and G<u>G</u>TT hotspots in IgH Sμ region. Number of each hotspot is indicated between parenthesis. **(B and C)** Total mutation frequency of AID targets in *Ung*$^{+/-}$*Msh2*$^{+/-}$ , *Ung*$^{-/-}$*Msh2*$^{+/-}$ ,*Ung*$^{+/-}$*Msh2*$^{-/-}$ GC B cells compared with that of *Ung*$^{-/-}$*Msh2*$^{-/-}$ GC B cells. Mutation frequency found in *Aicda*$^{-/-}$ GC B cells was substracted before plotting. Two-tailed Student's t-test, *, P ≤ 0.05. Error bars depict SEM.

To explore the contribution of BER and MMR to AID mutagenic activity from a more general perspective, we purified GC ($CD19^+FAS^+GL7^+$) B cells from Peyer's patches of single-deficient $Ung^{+/-}Msh2^{-/-}$ and $Ung^{-/-}Msh2^{+/-}$ mice and from control $Ung^{+/-}Msh2^{+/-}$ mice; isolated genomic DNA and performed target enrichment followed by deep-sequencing. Next, we carried out mutation analysis of the 1588 TSS-proximal regions included in our capture library. Unfortunately, our genome-wide approach provided us with limited resolution to evaluate the impact of UNG loss at the level of individual hotspot sequences in AID off-targets. It is interesting though, that when we focused on IgH Sμ region, which accumulates mutations at one order of magnitude higher than off-targets, we did find a trend that exactly reproduces the results published in Pérez-Durán et al., 2012 (Figure 21A). We compared the mutation frequency of the 291 AID target regions identified in this study in $Ung^{+/-}Msh2^{+/-}$, $Ung^{+/-}Msh2^{-/-}$, $Ung^{-/-}Msh2^{+/-}$ and $Ung^{-/-}Msh2^{-/-}$ cells (Figure 21B,C; Annex II). We found similar average mutation frequencies in B cells deficient for UNG alone, MSH2 alone or proficient for both, while AID targets harboured significantly more mutations in the combined absence of UNG and MSH2 (Figure 21B,C). Indeed, only a small proportion (~6%) of the genes mutated in $Ung^{-/-}Msh2^{-/-}$ cells harbour a detectable mutation load in single knockout and double heterozygous cells (Figure 22A, Table 11; Annex II). Moreover, we find that classical AID off-targets, such as *Bcl6* or *Pim1*, while mutated in all 4 genotypes, harbour a significantly bigger load of mutations in $Ung^{-/-}Msh2^{-/-}$ cells than in $Ung^{+/-}Msh2^{-/-}$, $Ung^{-/-}Msh2^{+/-}$ or $Ung^{+/-}Msh2^{+/-}$ cells (Figure 22B). Together, these data indicate that BER and MMR backup each other to faithfully repair most AID-induced lesions in GC B cells.

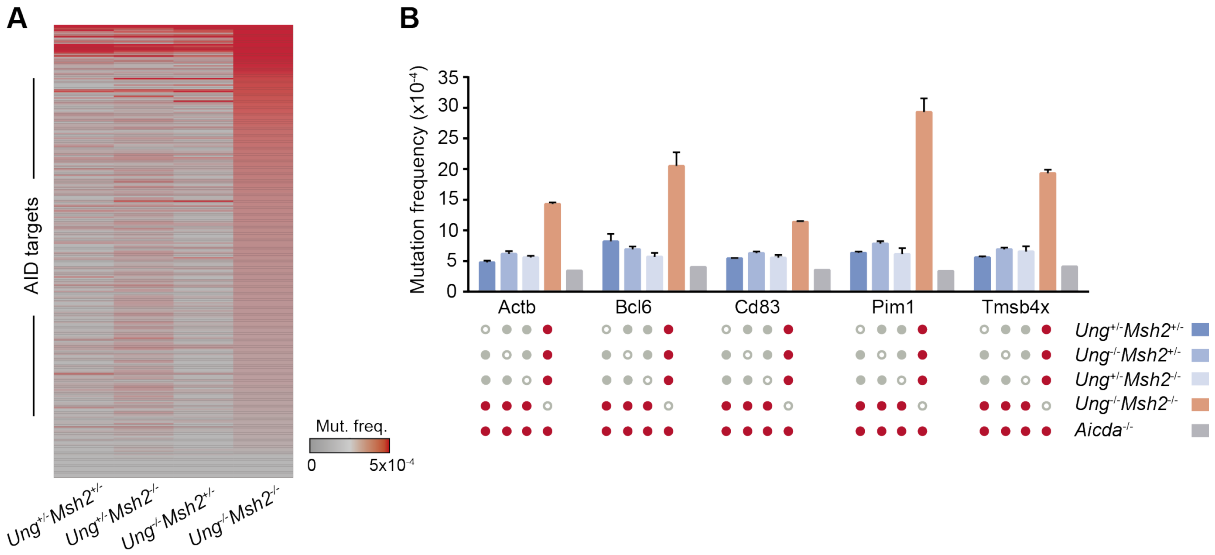



Figure 22 | **BER and MMR pathways faithfully repair most AID-induced deaminations. (A)** Heatmap representation of AID targets in *Ung*$^{+/-}$*Msh2*$^{+/-}$, *Ung*$^{-/-}$*Msh2*$^{+/-}$, *Ung*$^{+/-}$*Msh2*$^{-/-}$ and *Ung*$^{-/-}$*Msh2*$^{-/-}$ GC B cells. Mutation frequency found in *Aicda*$^{-/-}$ GC B cells was substracted before plotting. **(B)** Mutation frequency of representative genes in *Ung*$^{+/-}$*Msh2*$^{+/-}$, *Ung*$^{-/-}$*Msh2*$^{+/-}$, *Ung*$^{+/-}$*Msh2*$^{-/-}$, *Ung*$^{-/-}$*Msh2*$^{-/-}$ and *Aicda*$^{-/-}$ GC B cells. Red dots indicate statistically different mutation frequencies between the indicated genotypes.

| Gene name | | |
|---|---|---|
| *Acot 7* | *Dusp6* | *Ildr2* |
| *Actb* | *Dyrk3* | *Pim1* |
| *Bcl6* | *Galnt1* | *Prdx1* |
| *Cd19* | *IgH Eμ* | *Rps6* |
| *Cd83* | *IgH Jh4* | *Srsf10* |
| *Cdk4* | *IgH Sμ* | *Tmsb4x* |

Table 11 | **List of the 18 AID targets mutated in repair-proficient germinal center B cells.**

## 5. Analysis of the contribution of AID off-targeting to the development of Germinal Center derived malignancies

We have shown here that AID off-targeting occurs widely in the B cell genome during the GC reaction, and while MMR together with BER takes care of faithfully repairing many of these lesions, they leave behind a proportion of them, resulting in the accumulation of mutations in non-Ig genes. This poses an evident risk to genome integrity and can have an impact in the initiation and/or progression of oncogenic transformation. Genetic mouse models have shown that AID is responsible for chromosome translocations relevant for the etiology of lymphoma (Kovalchuk et al., 2007; Pasqualucci et al., 2008; Ramiro et al., 2004, 2006, Robbiani et al., 2008, 2009). However, the implication of AID-induced mutations in the malignant transformation of B cells remains object of study. We thus assessed the contribution of AID off-target mutations to B-cell derived malignancies by making use of available sequencing data on human lymphoma tumors. We found that AID targets are significantly enriched in genes mutated in human B cell lymphoma (see materials and methods section for details) (Figure 23).

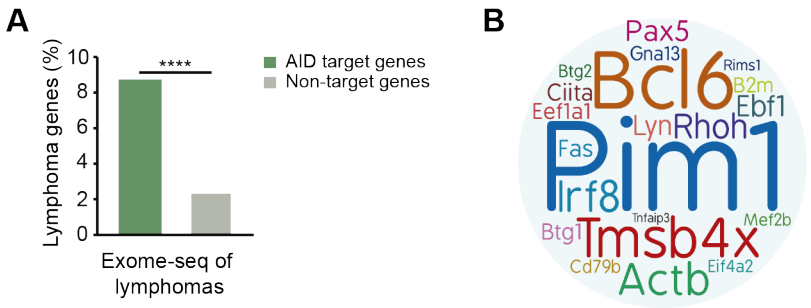



Figure 23 | **AID targets are recurrently mutated in human lymphomas. (A)** AID targets are enriched in genes frequently mutated in human lymphomas ("lymphoma genes"). Percentage of lymphoma genes within AID target and non-target genes. Annotation was done from public data on human lymphoma sequencing (see Materials and methods for details; Two-tailed Fisher test, ****, $P \leq 10^{-4}$). **(B)** Wordcloud representation of the lymphoma genes found to be mutated by AID in this study. Word size is proportional to mutation frequency.

Indeed, 21/275 (7.6%) of our set of AID target genes are mutated in diffuse large B cell lymphoma (DLBCL) (Figure 24; Annex IV), a highly prevalent, aggressive form of the disease (Shaffer and Staudt, 2012). In agreement with previous reports, these 21 genes include *Bcl6, RhoH, Pim1, Ebf1, Eif4a2* and *Pax5* (Liu et al., 2008; Pasqualucci et al., 2001; Shen et al., 1998). Additionally, we identified 9 novel genes mutated in human DLBCL that accumulate AID-induced mutations (Figure 24; Annex IV):

*Btg2, Ciita, Eef1a1, Gna13, Irf8, Lyn, Mef2b, Rims1* and *Tnfaip3.* Alterations in most of these genes are relevant for lymphomagenesis: *Gna13* codes for a small GTPase whose loss in GC B cells impairs apoptosis and promotes lymphoma (Healy et al., 2016); *Btg2* is an antiproliferative gene and well known tumor suppressor (Mao et al., 2014) involved in B-cell differentiation (Tijchon et al., 2016); *Ciita* is a frequent gene fusion partner in many lymphoid cancers and its alteration reduces patient survival (Steidl et al., 2011), most likely by suppressing antigen presentation and allowing immune evasion (Green et al., 2015); *Irf8* frequently translocates to IgH and drives DLBCL (Bouamar et al., 2013); *Lyn* is a proto-oncogene (Ingley, 2012; Tauzin et al., 2008) coding for a kinase that acts as a negative regulator of the BCR signalling (Chan et al., 1997; Hibbs et al., 1995); *Tnfaip3* deregulation constituvely activates NF-κB signalling and has been reported as a tumor supressor in Hodgkin (Schmitz et al., 2009) and Non-Hodgkin lymphoma (Compagno et al., 2009); and *Mef2b* encodes an acetyltransferase whose overexpression increases cell migration and favours epithelial-mesenchymal transition (Pon et al., 2015). On the other hand, the contribution to lymphoma of *Eef1a1,* a translation elongation factor, and *Rims1*, a member of the *Ras* superfamily of genes that regulates exocytosis, has not been thoroughly demonstrated. However, given the fact that they play key functions in the cell and appear frequently mutated in human lymphoma, their involvement in disease does not seem a far-fetched possibility.

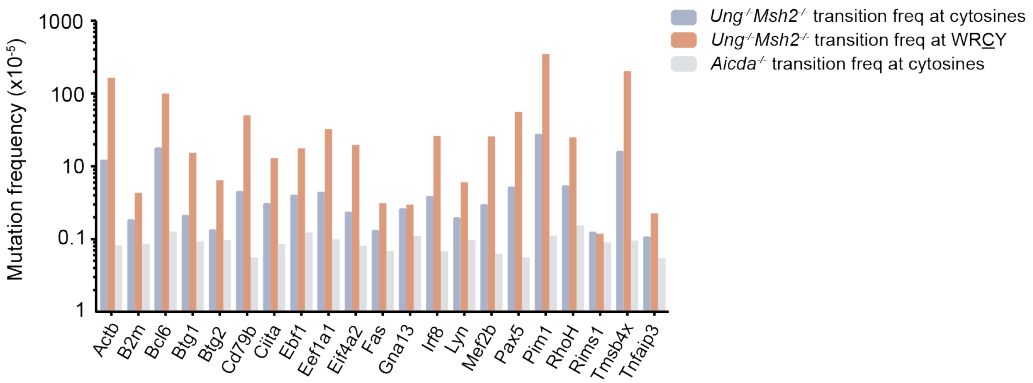



Figure 24 | **Genes frequently mutated in human DLBCL are targeted by AID.** Mutation frequency in total C/G nucleotides and C/G nucleotides within WRCY/RGYW hotspots of the 21 genes frequently mutated in human DLBCL and targeted by AID in *Ung*$^{-/-}$*Msh2*$^{-/-}$ and *Aicda*$^{-/-}$ mouse GC B cells (mean of two independent experiments, see Material and methods for details).

In addition to the functional relevance of the above-mentioned AID target genes, it is remarkable that we found many instances where the exact same mutations described in human lymphoma tumours were also found in our study in non-transformed mouse GC B cells (Figure 25; Table 12). For instance, we found a C→G transversion in *Gna13* that changes a tyrosine in residue 89 to a stop codon or a G→A transition affecting a tryptophan in residue 85 of *Tnfaip3* that truncates the protein as well (see Table 12 for a complete list of mutations). Interestingly, many of these mutations frequently lie within hotspot sequences (Figure 25, highlighted by asterisks), suggesting that they may have originated from AID activity in human tumors, and affect amino acid residues that are evolutionarily well conserved in mammals, which highlights their importance for protein function.

| Gene | Position | nt change | aa change | Mut. freq. (x10-4)[(a)] | Pathology | Reference |
|---|---|---|---|---|---|---|
| *Btg2* | 1:135975497 | C → A | A45T \| A45E | 4 | DLBCL | [1] |
| *Btg2* | 1:135975530 | G→T | R34I | 3.6 | NHL | [8] |
| *Btg2* | 1:135975535 | G→T | E32D | 14 | NHL | [8] |
| *Btg2* | 1:135975590 | C→T | A14V | 4 | NHL | [8] |
| *Gna13* | 11:109224379 | C→T | S31F | 35 | DLBCL | [1] |
| *Gna13* | 11:109224450 | T→C | L38P | 6.1 | DLBCL | [8] |
| *Gna13* | 11:109224554 | C→G | Y89STOP | 7.1 | NHL | [8] |
| *Tnfaip3* | 10:18731328 | G→A | W85STOP | 6 | Hodgkin | [2] |
| *Tnfaip3* | 10:18731338 | A→T | K82STOP | 6 | FL | [3] |
| *Tnfaip3* | 10:18731530 | G→C | A18P | 31 | DLBCL | [4] |
| *Tnfaip3* | 10:18731575 | G→C | E3STOP | 5.6 | DLBCL | [4] |
| *Pim1* | 17:29628092 | C→T , C→G | L2F \| L2V | 136 | FL | [1] [3] [5] |
| *Pim1* | 17:29628109 | C→G | N7K | 555 | NHL | [6] |
| *Pim1* | 17:29628113 | C→T | L9F | 9.3 | DLBCL | [7] |
| *Pim1* | 17:29628117 | C→G | A10T | 38.8 | DLBCL | [7] |
| *Pim1* | 17:29628146 | C→G | L20V | 147 | NHL | [6] |
| *Pim1* | 17:29628156 | C→G , C→T | T23R \| T23I | 35.1 | FL | [3] |
| *Pim1* | 17:29628160 | G→C | K24N | 313 | FL | [3] |
| *Pim1* | 17:29628162 | C→G | L25V | 11.6 | NHL | [1] |
| *Pim1* | 17:29628270 | G→A , G→C | G28D \| G28A | 166 | DLBCL. FL | [3] [6] [7] |
| *Pim1* | 17:29628275 | G→A | E30K | 11.9 | NHL | [6] |
| *Pim1* | 17:29628284 | C→T | P33S | 970 | FL | [3] |
| *Pim1* | 17:29628298 | G→C | Q37H | 63.6 | NHL | [1] |
| *Pim1* | 17:29628301 | C→A | Y38STOP | 58.5 | DLBCL | [3] |
| *Pim1* | 17:29628302 | C→T | Q39STOP | 64.1 | DLBCL. FL | [3] [7] |

Table 12 | **Mutations found in *Ung*$^{-/-}$*Msh2*$^{-/-}$ mice that have been identified in cohorts of human lymphoma patients.** (**a**) Total net mutation frequency calculated as: (Number of mutations / Number of bases sequenced)$_{\text{UngMsh2 dKO}}$ − (Number of mutations / Number of bases sequenced)$_{\text{AID KO}}$

[1] Bruno et al., 2014; [2] Compagno et al., 2009; [3] Fabbri et al., 2013; [4] Morin et al., 2011; [5] Pasqualucci et al., 2014; [6] Schmitz et al., 2009; [7] Zhang et al., 2013;[8] International Cancer Genome Consortium (DFKZ. Germany).

Together these results suggest that off-target AID mutagenic activity can contribute to GC-associated lymphomagenesis.

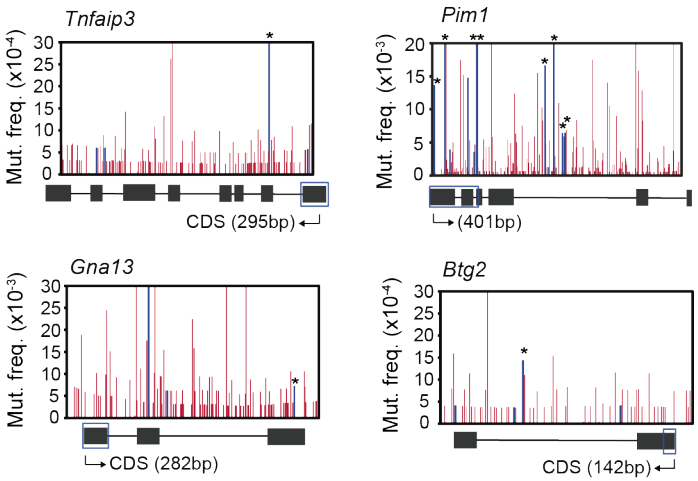



Figure 25 | **Mutation profiles of representative DLBCL genes.** Per nucleotide analysis of the mutation frequency of representative DLBCL genes in *Ung*$^{-/-}$*Msh2*$^{-/-}$ mouse GC B cells. Blue bars indicate mutations identical to those found in human lymphoma tumor samples (Table 12); asterisks indicate mutations occurring in a WRCY hotspot. The diagrams below the graphs represent the complete gene (not to scale), and blue boxes indicate the region depicted above. Mutation frequency found in each nucleotide in *Aicda*$^{-/-}$ GC B cells was substracted before plotting.

# V. DISCUSSION

## 1. Development of a capture-based NGS approach to detect AID-induced mutations

In this thesis work, we studied the specificity of AID targeting at a high-throughput scope. With that aim, we developed a novel strategy based on Next Generation Sequencing (NGS) to analyze AID-induced mutations in the genome of germinal center B lymphocytes. It is important to note that AID-induced mutations are particularly challenging to detect, since they occur at a very low frequency ($10^{-2}$-$10^{-3}$ in Ig genes; at least one order of magnitude lower in off-targets) and do not show clonal recurrence, with different cells carrying different mutations. Thus, it is critical that the sequencing approach to detect AID-induced mutations: 1) allows high depth, i.e. each nucleotide must be read a large number of times; 2) has a low sequencing error rate, i.e. below the mutation frequency introduced by AID. To accomplish these goals, we designed a protocol that couples target enrichment to NGS. Target enrichment effectively reduces the total length of DNA that is subject to sequencing; thus, for the same total volume of sequences obtained, target enrichment allows a great increase in the number of times each nucleotide is read. In our case, we captured 500bp from the TSS of a selection of ~1600 genomic locations, covering a total of 0.8Mb, approximately 1/3000th of the mouse genome. Accordingly, this target enrichment approach yielded roughly a 2000 fold enrichment of the target DNA regions and a consistent sequencing depth of ~2300x, i.e. each nucleotide was read approximately 2300 times on average. Importantly, this increases by more that 20 fold the sequencing depth achieved by previous studies based on conventional Sanger sequencing (Liu et al., 2008). In contrast, sequencing the whole mouse exome or genome at 2300x depth would have meant producing a sequence volume of ~1.5Tb and ~6Tb respectively, a largely unaffordable setup. Therefore, with our experiment we have covered a reasonably broad length sequence with affordable means.

Regarding sequencing background, our protocol involves several PCR steps during target enrichment and library preparation that can introduce unspecific mutational noise. We found however, that the major source of error stems from the inherent chemistry of the sequencer, which cannot be modified. To control for sequencing background noise and false positive signals, i) we have used AID deficient mice to get an estimation of the sequencing error; and ii) we have developed a custom bioinformatics pipeline that allowed us to discard genomic variation derived from mixed genetic backgrounds. In addition, computational noise (errors coming from base calling and/or sequence alignments) was also strictly controlled by filtering data based on base calling and mapping qualities. Although the

background noise of the NGS setup presented here is much higher than that of Sanger sequencing, the huge increase in read depth and the genome-wide scope provided by our capture-based approach largely compensates for this loss of sensitivity. Indeed, this approach has allowed, for the first time, the evaluation of AID off-target mutagenesis in more than 1500 different genomic locations. As a reference, this is ~15 times larger than the largest study published to date. Remarkably, our capture-based approach can be easily extended to analyze broader collections of genes and our bioinformatics pipeline would allow the analysis of AID-induced mutations up to a whole-genome level.

## 2. Discovery of a large catalogue of AID mutational targets

Up to date, only a limited number of genes had been directly interrogated for AID-mediated mutagenesis (Shen et al., 1998; Pasqualucci et al., 1998; Müschen et al., 2000; Pasqualucci et al., 2001; Gordon et al., 2003; Liu et al., 2008; Robbiani et al., 2009; Pavri et al., 2010). Instead, genome-wide AID specificity has been inferred from high throughput analysis of AID binding (Yamane et al., 2011) or AID-induced DSBs or TCs (Chiarle et al., 2011; Klein et al., 2011; Staszewski et al., 2011, Meng et al., 2014; Qian et al., 2014). The Casellas lab performed a ChIP-Seq experiment of AID in *in vitro* LPS+IL4 activated B cells and reported AID binding to ~5900 genes (~12000 genomic locations) (Yamane et al., 2011), although serious concerns have been raised on different aspects of this study (Hogenbirk et al., 2012). Regardless of these issues, at best ChIP-Seq studies would provide information of AID binding, but not necessarily activity. On the other hand, the generation of DSB or TCs involves a complex processing of the initial lesion induced by AID that occurs only in a fraction of the deamination events. Therefore, measuring DSBs or TCs as a surrogate of AID activity will oversee all the lesions that are fixed as mutations, either by replication or after BER/MMR activities.

In contrast, the use of a mouse model deficient for BER and MMR provides a clean, unbiased view of the primary deamination events produced by AID. Thus, we identified the largest collection of AID mutational targets -275 genes- to date, expanding in more than 10 times the number of previously known targets. Several evidences support the reliability of these data. First, the mutation frequencies measured in AID targets are well above background signal. On average, AID targets showed a mutation frequency 4x higher than background in the *Ung*$^{-/-}$*Msh2*$^{-/-}$ setting. Second, we found hotspot focusing, a hallmark of AID activity. In hotspots, mutation frequencies peaked to up to an 8-12 fold increase over

background. Third, AID targets are highly reproducible, as supported by the good correlation found between two biological replicates ($R^2$ = 0.99). Fourth, we validated some of the newly discovered targets by conventional Sanger sequencing. Finally, we confirmed the vast majority of previously identified AID targets (Shen et al., 1998; Pasqualucci et al., 1998; Müschen et al., 2000; Pasqualucci et al., 2001; Gordon et al., 2003; Liu et al., 2008; Robbiani et al., 2009; Pavri et al., 2010). In addition, integration of our mutational targets with DSB/TC data (Chiarle et al., 2011; Klein et al., 2011; Staszewski et al., 2011; Qian et al., 2014; Meng et al., 2014) revealed that while not all deaminations produced by AID give raise to DSB/TCs, most of the events identified in these studies originate from AID activity.

## 3. AGCTNT as a novel hotspot for AID activity

We took advantage of the large collection of AID targets identified here to systematically evaluate the intrinsic local sequence preference of AID. Analysis of classic AID hotspots revealed that they are highly sensitive to genomic location, as shown by the wide range of mutation frequencies observed for the same hotspot within the same gene and between different genes (Figure 12). Consistent with previous findings (Pérez-Durán et al., 2012; Yeap et al., 2015), we identified AGCT as the most mutated hotspot within the classical WRCY/RGYW motifs. However, we decided to explore the possibility that a wider sequence context could also influence the efficiency of AID targeting and analyzed a window of ±10 nucleotides from mutated cytosines. Thus, it is important to note here that we focused on the mutated cytosine for our hotspot search, while the classic WRCY/RGYW hostpots were defined considering the whole motif as a possible target. We found that the +4 nucleotide position from the mutated cytosine fine-tunes AID local specificity. Specifically, we found that AGCTNT is a novel hotspot for AID activity and that cytosines lying within this motif are significantly more mutated than those lying within AGCTNV. Furthermore, 78% of the AGCTNT motifs identified in AID targets are mutated, as compared to 42% of AGCTNV motifs. Interestingly, AGCTNT motif is enriched in IgH Sμ region, where we identified a total of 6 AGCTNT motifs, three times more than expected by chance. Remarkably, AGCTNT was mutated in 100% of the instances mentioned above. Last, this motif has also been identified in murine IgH Sγ2a, Sγ2a and Sγ3 consensus tandem repeats. Together, these findings suggest that AGCTNT could be relevant for CSR. Regarding its potential role in AID off-targeting, even if the occurrence of this novel hotspot is more restricted than that of classic

WRCY/RGYW hotspots (1/1024 bp versus 1/256 bp), it seems unlikely that it suffices to define AID off-target specificity. In addition, AGCTNT appears at the random expected frequency in off-targets (not shown), which would suggest that it does not contribute to the recruitment of AID to these loci. Thus, we propose that, similar to the classic WRCY/RGYW hotspots, AGCTNT defines a local preference for AID deaminating activity once it has been targeted to a particular locus. In this respect, the accumulation and frequent mutation of AGCTNT hotspots in Ig switch regions suggests that AID mediated deamination in these motifs could be a major event for the initiation of CSR. Equivalently, AGCTNT may contribute to AID-induced deamination in off-targets. We therefore speculate that although the primary targeting of AID is not defined by the sequence context alone, once AID gains access to ssDNA, specific sequence contexts can confer local specificity for SHM. However, the mechanisms by which AID is recruited to Ig and non-Ig loci remain to be defined, as we will discuss in the next chapter.

## 4. Molecular characterization of AID targets

In this study we have performed a comprehensive characterization of the molecular features that associate to off-target SHM. Given that active transcription is a well-known requirement for AID activity, we first analyzed the transcriptional profiles of GC and resting B cells. As expected, we found diverging transcriptional programs in both subsets. It is interesting that approximately one third of the genes specifically upregulated during the GC reaction are mutated by AID. Nevertheless, they only account for about half of total AID targets, the other half being genes highly expressed both in the naïve and activated states. This indicates that AID off-target activity is not exclusively directed to GC specific genes, but to a wider set of highly expressed loci. Of interest, we observed some instances of highly expressed genes that were not mutated by AID. These results support the idea that active transcription is an absolute requirement for AID activity but argue that AID specificity is not driven by transcription levels. On the other hand, we also observed a few AID targets which were apparently expressed at very low levels or not expressed at all. Since this is a very small proportion of all detected AID targets, we think it may reflect a technical issue with the RNA-Seq data or annotation; this should be experimentally tested by an alternative measurement of mRNA levels. We also analyzed transcription rate, i.e. the number of engaged transcriptional complexes per gene, and found that AID

targets are transcribed at higher rates than non-targets, reinforcing the notion that AID preferentially targets highly transcribed loci.

A number of genetic screenings have shown AID binding to different components of the transcription machinery, which would facilitate the association of AID with transcribed switch regions and modulate CSR. Some examples would be the 14-3-3 adaptor proteins (Xu et al., 2010), PTBP2 protein (Nowak et al., 2011), the RNA exosome (Basu et al., 2011), RNA polymerase II (Nambu et al., 2003) and the elongation factor SPT5 (Pavri et al., 2010). Based on the position of hypermutation relative to transcription start sites, Storb proposed that RNAP II pausing could be linked to mutation (Peters&Storb, 1996). Further, a genome-wide study of RNAP II and SPT5 in activated B cells showed a strong correlation between RNAP II and SPT5 genome occupancy (Pavri et al., 2010). In this study, authors proposed that SPT5 facilitates the interaction between AID and RNAP II and pointed to polymerase stalling as a mechanism by which AID would get access to ssDNA, with decreased elongation rates providing more time for AID to deaminate the target sequence (Pavri et al., 2010). To explore the contribution of RNAP II and SPT5 binding to AID off-targeting, we analyzed their recruitment to the AID targets identified in our study. In line with our findings on transcription, we observed that AID targets bind high levels of RNAP II and SPT5, suggesting that their binding to off-targets is relevant for AID recruitment to these loci. It is interesting that, on average, genes included in the "highly mutated" group accumulated significantly higher levels of RNAP II and SPT5 than those in the "mutated" or "non-mutated" groups, which suggests a direct correlation between the binding of these transcription cofactors and the accumulation of AID-induced mutations. However, at the gene level this correlation was poor, so we must be cautious in the interpretation of these results.

Epigenetic modifications have been suggested to contribute to AID targeting to Ig (reviewed in Sheppard et al., 2018) and non-Ig loci (Wang et al., 2014). Wang et al compared AID-induced TCs in B cells and mouse embrionary fibroblasts upon AID overexpression and identified some genes similarly transcribed in both experimental systems that were only mutated in B cells. Moreover, they found a common set of distinctive epigenetic features associated to these genes in B cells but not in MEFs. These data led them hypothesize that epigenetic modifications are mediators of AID recruitment to off-target sites in a fashion that is not dependent on the transcription of the target gene. Here we have analyzed the presence of epigenetic marks associated to active transcription and transcription elongation in AID targets. Consistent with the findings by Wang and colleagues, we observed that AID targets are enriched in H3K36me3 and H3K79me2 marks. However, given the correlative nature of this

observation, we can only suggest that these epigenetic marks define a chromatin accessibility status that likely favors AID activity.

Recent studies have linked AID off-targeting to super-enhancer (SE) regulated loci (Qian et al., 2014; Meng et al., 2014). SE are regulatory clusters where chromatin accessibility and transcriptional activity are much higher than at other transcriptional active sites (Whyte et al., 2013). Further, they can establish long-range interactions and play an important role in cell specific processes (Hnisz et al., 2013, 2015; Whyte et al., 2013). SE have been related to AID targeting at two different levels. On one hand, they have been proposed to establish regulatory clusters that drive "networks of cooperating elements" that generate "the proper conditions for AID promiscuous activity" (Qian et al., 2014). On the other hand, SE initiate antisense transcription within sense transcribed genes (convergent transcription). This phenomenon contributes to RNAP II stalling and exosome recruitment, which contribute to AID recruitment and activity (Meng et al., 2014). In agreement with these hypothesis, we found that AID targets are frequently regulated by SE, and that primary AID targeting often occurs in regions undergoing convergent transcription. We also observed that highly mutated genes were more frequently regulated by SE than mutated and non-mutated genes. It is important to underline that although ConvT arises from SE, not all genes regulated by SE necessarily undergo ConvT. However, we found that all AID off-targets regulated by SE also underwent ConvT, which suggests that, together, SE regulation and ConvT may favor a microenvironment suitable for AID activity. Following Meng and Qian hypotheses we would speculate that: i) SE would provide improved chromatin accessibility to AID by increasing transcription and potentially coordinate regulated networks of B cell specific transcription factors that could contribute to AID off-targeting; ii) ConvT would lead to Pol II stalling and exosome recruitment, which in turn favor AID activity at off-targets.

AID activity occurs at much higher levels on Ig genes than on off-targets. Although this preferential targeting has been known for years, the mechanisms underlying are not completely understood. Several features have been proposed to make the Ig loci privileged for SHM, including specific enhancers (Buerstedde et al., 2014), regulatory regions (Rouaud et al., 2013), epigenetic marks (reviewed in Sheppard et al., 2018) or transcriptional-related mechanisms, such as RNAP II stalling (reviewed in Storb, 2014), exosome recruitment (Basu et al., 2011) or SE regulation (Qian et al., 2014; Meng et al., 2014). Remarkably, some of these features are also present in AID off-targets. This raises the question of whether AID off-targeting may be driven by the same Ig-like features and only quantitative differences account for the widely differing mutation frequencies in Ig versus non-Ig targets, or

alternatively, if off-targets are made accessible to AID activity by distinct molecular features. In this respect, ConvT is the only phenomenon linked to AID targeting that has been observed in off-targets but not in Ig genes, although this may be due to technical reasons (Meng et al., 2014). The finding that IgH Sμ is >7 times more mutated than the most highly mutated off-targets in *Ung*$^{-/-}$*Msh2*$^{-/-}$ GC B cells suggests that AID primary targeting to Ig genes is much more frequent than to off-targets and that BER and MMR mediated processing of AID-induced lesions does not account for the different mutation frequencies found at on and off-targets. Indeed, we observed that AID-mediated deaminations were repaired to a similar extent in IgH Sμ and in off-targets (~70-80% of lesions faithfully repaired). Thus, we believe that a combination of known and unknown Ig specific features may explain AID preference for Ig loci. Similarly, we think that AID off-targeting is not random, but also defined by a specific set of features. In this regard, we found that several mechanisms linked to transcription contribute to AID off-targeting, but none of them alone suffices to define AID specificity. We envision that AID off-target specificity is driven by a complex combination of factors, including DNA repair pathways, specific sequence contexts, chromatin accessibility, transcriptional cofactors and other transcriptional and architectural regulatory mechanisms. In this regard, we think a thorough analysis of the dynamics of genome compartmentalization and loci interactions during the GC reaction would help complete the picture and probably establish a definitive link between transcription and the definition of AID target specificity.

## 5. Machine learning approach to predict AID off-targeting

Here, we have integrated our mutation data with the collection of molecular features described above to feed a machine learning algorithm. According to the classification tree generated by our model, the combined binding of SPT5 and RNAP II at high density is the best predictor for AID mutability, although other transcription-associated traits, such as high expression levels or presence of H3K79me2 mark, bear some predictive power as well. Further, we have performed experimental validation of the model by randomly picking 12 genes predicted as targets and sequencing them. It is important to note that none of those genes was initially included in our capture library. Notably, 11/12 (91%) were actually mutated, suggesting that the accuracy of the model is even higher than expected. Indeed, two of those genes were mutated at the range of the top 20% mutated AID targets. Theoretically, our machine learning approach could be improved by using other classic machine learning algorithms,

such as random forests or support vector machines, which are less prone to overfitting, or deep learning, which is more accurate than classic machine learning methods. However, we do not think the use of these models would have a major impact on our predictions, i) because the recursive partitioning model we used was not overfitted; ii) because deep learning requires extremely large training datasets to be properly trained and outperform classic machine learning approaches. Instead, we believe that incorporating new parameters to the prediction, such as 3D genome organization information or data on binding of other transcriptional cofactors, could indeed be very useful to fine-tune the predictions and contribute to better delimit the extent of AID off-target activity in the B cell genome.

A first attempt to predict AID off-targeting was published a few years ago but the prediction model lacked experimental validation (Duke et al., 2013). In addition, the moderate number of AID targets known at that time limited the prediction power of the approach. A total of 83 genes, which had been previously assayed for AID mutations in *Ung*$^{-/-}$*Msh2*$^{-/-}$ B cells (Liu et al., 2008), was used to train the model. This collection of genes was divided in three different groups: A (15 highly significant mutated genes), B (21 mutated genes) and C (47 non-mutated genes). According to their classification tree, a total of 6073 genes were predicted to be AID targets, of which 1896 genes were predicted to be highly mutated (group A). To get an estimation of the prediction efficiency of the model, we compared their predicted set of AID targets with collections of genes that bear AID-induced DSB/TC (Staszewski et al., 2011; Chiarle et al., 2011; Qian et al., 2014; Meng et al., 2014). We only found a minor fraction of the genes that undergo DSB/TCs within their most reliable set of predicted targets (group A): 10% (Staszewski et al., 2011) and 19% (Qian et al., 2014) of reported DSBs; and 22% (Klein et al., 2011); 18% (Chiarle et al., 2011) and 16% (Meng et al., 2014) of reported TCs. This suggests that the efficiency of the prediction model was far from optimal. Similarly, only a small portion of the 275 AID targets that we experimentally identified in GC B cells were predicted by their model: 14% (39/275) if we compare to group A; 16% (44/275) if we compare to their complete set of predicted genes (groups A+B).

Thus, to the best of our knowledge, the machine learning approach presented in this thesis constitutes the first instance of a tool that successfully predicts the potential of a gene to be targeted by AID. There are roughly 430 genes predicted to be AID off-targets by our machine learning model, including many cancer drivers and genes recurrently mutated in human lymphoma. Therefore, we think this collection of potential off-targets constitutes a good starting point to guide new SHM studies both from the perspective of AID biology and in the context of carcinogenesis.

## 6. Role of BER and MMR in the resolution of AID-induced deaminations

With regards to the fate of AID-induced lesions, BER and MMR have been long known to broaden the diversity of SHM with an apparent perverted recruitment of error prone polymerases, and to do so in a cooperative manner (Di Noia and Neuberger, 2007; Methot and Di Noia, 2017; Rada et al., 2004). The mechanisms responsible for the error-free versus error-prone activity of UNG and MSH2 are far from understood, and both local sequence and gene-specific contexts may play a role in defining the fate of the U:G resolution (Liu et al., 2008; Pérez-Durán et al., 2012). Pérez-Durán et al. showed that the sequence context infuences the outcome of AID activity on the IgH locus, with particular hotspots favouring error-free or error-prone repair of U:G mismatches. On the other hand, Liu et al. examined a collection of 118 genes that are highly expressed in GC B cells by Sanger sequencing and identified 23 highly mutated genes (~20%) in a repair proficient background (*Aicda*$^{+/+}$ vs *Aicda*$^{-/-}$ GC B cells). Further, they sequenced 83 genes from B cells deficient for UNG and MSH2, and found 36 (43%) significantly mutated. Analysis of individual off-targets in WT, *Ung*$^{-/-}$, *Msh2*$^{-/-}$ and *Ung*$^{-/-}$*Msh2*$^{-/-}$ GC B cells revealed that different genes behaved differently: although statistical significance was precluded by the limited number of mutations detected, authors observed differential trends when comparing mutation frequencies in the four mentioned genotypes. These data led them to the hypothesis that the B cell genome is protected at two different levels: 1) by the selective targeting of AID (since they found ~40% genes targeted by AID); 2) by a balance between error-free and error-prone repair, which would be defined by gene-specific features.

Interestingly, our data demonstrates that the fate of the majority of off-target lesions induced by AID is to undergo faithful repair by BER and MMR and that both pathways can backup each other in this task. It is worth mentioning here that all the 10 AID off-targets analyzed in Liu et al. in the repair-deficient setting showed a consistent tendency to accumulate more mutations in the absence of BER and MMR than in any of the other genotypes, which is concordant with our results. Further, they found a higher proportion of mutated genes in the repair-deficient (43%) than in the repair proficient background (20%).

Given the major repair function of BER and MMR, our experimental approach did not provide enough resolution to evaluate individual hotspots in AID off-targets in *Ung*$^{-/-}$*Msh2*$^{+/-}$, *Ung*$^{+/-}$*Msh2*$^{-/-}$ or *Ung*$^{+/-}$*Msh2*$^{+/-}$ GC B cells. However, we did find a trend that reproduces the results published by Pérez-Durán et al. when we analyzed AA<u>C</u>T, TA<u>C</u>T, A<u>G</u>TT, A<u>G</u>TA, G<u>G</u>TA and G<u>G</u>TT hotspots in IgH Sμ (which accumulates at least one order of magnitude more muations than off-targets) of *Ung*$^{+/-}$*Msh2*$^{+/-}$ and *Ung*$^{-/-}$*Msh2*$^{+/-}$ B cells.

We also observed that the error-free repair of AID off-targets by BER/MMR is not absolute. Indeed, a minor fraction of the mutations escaped repair and we identified a collection of genes (18; ~6.5% of total AID targets) consistently mutated regardless of the genotype analyzed. This finding can be interpreted from different perspectives. On one hand, as suggested by Liu et al., 2008, it is possible that different genes possess different specific qualities that tip the scales towards a higher or lower level of repair. Alternatively, since these 18 genes correspond to the most mutated AID off-targets, it seems reasonable to speculate that the repair pathways might be overwhelmed by an excessive mutation load. While further investigation will be required to ascertain this issue, our data suggests that faithful repair of off-target mutations is the prevalent outcome of the BER and MMR pathways, with different genes being repaired with different efficiencies.

## 7. AID off-targeting and lymphomagenesis

As we have discussed above, AID mistargeting to non-Ig loci is a frequent event during the GC reaction and affects a large number of genes. Despite most AID-induced lesions being faithfully repaired by the BER and MMR pathways, there is a fraction that remains unrepaired and could contribute to genome instability and carcinogenesis. We found that the AID off-targets identified in this study are significantly enriched in genes that are recurrently mutated in human lymphoma tumors, which may suggest a contribution of AID-induced mutations to lymphomagenesis. More specifically, our collection of targets included 21 genes that are usually mutated in human DLBCL. Of these, 9 were revealed as AID targets for the first time in this study. Although many of these 21 genes are well-known tumor suppressors (*Gna13, Btg2, Tnfaip3, etc*) or proto-oncogenes (*Bcl6, Lyn, Pim1, etc*), it remains possible that some of the mutations observed in human lymphomas, although of AID origin, are mere

passengers. Regardless of oncogenic relevance, it is remarkable that even though our study was performed in non-transformed cells we could detect AID mutations in the exact same residues that have been recurrently found mutated in human lymphomas. Thus, our results yield a novel perspective on the contribution of AID activity to B cell transformation through the introduction of mutations. Regarding the carcinogenic process, we would speculate that a minor fraction of unrepaired mutations in pro-lymphomagenic genes could be enough to provide cell growth advantage and account for the predominance of AID-mediated mutations found in lymphomas. However, there is extensive evidence of the relevance of repair deficiencies in the development of cancer, with *Msh2* probably being one of the most notable examples (reviewed in Baretti and Le, 2018). In addition, polymorphisms in *Ung* have been associated with cancer, although they are much less frequent than mutations in *Msh2* (reviewed in Wallace et al., 2012). Therefore, we cannot completely rule out the possibility that the repair pathways are somehow crippled in the lymphomagenesis context, thus allowing AID-induced lesions to escape repair.

## 8. Concluding remarks and future prospects

We envision that the advent of new sequencer chemistries and the evolving reduction in the costs of sequencing will soon allow the whole-genome evaluation of AID off-site mutational activity. In the meanwhile, we believe that the catalog of AID targets provided by our study constitutes a very valuable resource for the field. Together with the experimental and *in silico* strategies developed here, the discovery of this large collection of AID targets will help tackle relevant research questions, such as the evaluation of novel molecular mechanisms involved in AID targeting, the prediction of new targets or the assessment of cancer-associated mutations. Furthermore, our capture-based approach can be easily translated to the human setting. In this regard, it would be of immediate interest to perform an exhaustive profiling of AID off-target activity in human memory B cells. Optimally, the analysis of paired tumor-healthy samples from the same donor would help clarify not only what is the extent of AID mistargeting in a repair proficient background, but also what is the precise contribution of AID mutations to lymphomagenesis or other malignancies. We think approaches similar to ours would be very interesting to broaden our knowledge on the biology of AID in humans and its contribution to disease.

# VI. CONCLUSIONS

# Conclusions

1. We have developed a capture-based NGS approach that allows the detection of AID-induced mutations at a high-throughput level.

2. We have identified the largest collection of AID off-targets to date, composed by 275 different genes.

3. We have identified AGCTNT as a new hotspot for AID activity.

4. AID targets are highly transcribed loci that bind high levels of the transcription cofactors RNAP II and SPT5 and bear epigenetic marks associated to active transcription and transcription elongation.

5. AID targets are frequently regulated by superenhancers and undergo convergent transcription.

6. We have developed and validated a machine learning algorithm that predicts AID off-targeting.

7. BER and MMR pathways backup each other to faithfully repair AID-induced deaminations.

8. AID-induced mutations are recurrently found in human lymphomas.

# VI. CONCLUSIONES

## Conclusiones

1. Hemos desarrollado un protocolo de captura seguido de secuenciación masiva que permite la detección de mutaciones inducidas por AID a gran escala.

2. Hemos identificado la colección de dianas de AID más amplia hasta la fecha, compuesta por 275 genes distintos.

3. AGCTNT es un nuevo "punto caliente" (*hotspot*) para la actividad de AID.

4. Las dianas de AID son *loci* altamente transcritos que unen niveles elevados de los cofactores transcripcionales ARN polimerasa II y SPT5 y presentan marcas epigenéticas asociadas a transcripción activa y elongación transcripcional.

5. Las dianas de AID están reguladas por súper activadores transcripcionales y sufren transcripción covergente frecuentemente.

6. Hemos desarrollado y validado experimentalmente un modelo de aprendizaje automático (*machine learning*) capaz de predecir nuevas dianas de AID.

7. Las vías de reparación por escisión de bases (*base excision repair*) y de desapareamiento de bases (*mismatch repair*) se respaldan mutuamente para reparar de forma fiel la mayor parte de las lesiones inducidas por AID.

8. Las mutaciones inducidas por AID se encuentran de forma recurrente en linfomas humanos.

# Bibliography

Aoufouchi, S., Faili, A., Zober, C., D'Orlando, O., Weller, S., Weill, J.-C., and Reynaud, C.-A. (2008). Proteasomal degradation restricts the nuclear lifespan of AID. Journal of Experimental Medicine *205*, 1357–1368.

Arakawa, H., Kuma, K., Yasuda, M., Furusawa, S., Ekino, S., and Yamagishi, H. (1998). Oligoclonal Development of B Cells Bearing Discrete Ig Chains in Chicken Single Germinal Centers. The Journal of Immunology *160*, 4232–4241.

Arakawa, H., Hauschild, J., and Buerstedde, J.-M. (2002). Requirement of the Activation-Induced Deaminase (AID) Gene for Immunoglobulin Gene Conversion. Science *295*, 1301–1306.

Arya, R., and Bassing, C.H. (2017). V(D)J Recombination Exploits DNA Damage Responses to Promote Immunity. Trends in Genetics *33*, 479–489.

Babbage, G., Ottensmeier, C.H., Blaydes, J., Stevenson, F.K., and Sahota, S.S. (2006). Immunoglobulin Heavy Chain Locus Events and Expression of Activation-Induced Cytidine Deaminase in Epithelial Breast Cancer Cell Lines. Cancer Res *66*, 3996–4000.

Bachl, J., Carlson, C., Gray-Schopfer, V., Dessing, M., and Olsson, C. (2001). Increased Transcription Levels Induce Higher Mutation Rates in a Hypermutating Cell Line. The Journal of Immunology *166*, 5051–5057.

Bardwell, P.D., Woo, C.J., Wei, K., Li, Z., Martin, A., Sack, S.Z., Parris, T., Edelmann, W., and Scharff, M.D. (2004). Altered somatic hypermutation and reduced class-switch recombination in exonuclease 1–mutant mice. Nature Immunology *5*, 224.

Baretti, M., and Le, D.T. (2018). DNA mismatch repair in cancer. Pharmacology & Therapeutics.

Barreto, V., Reina-San-Martin, B., Ramiro, A.R., McBride, K.M., and Nussenzweig, M.C. (2003). C-Terminal Deletion of AID Uncouples Class Switch Recombination from Somatic Hypermutation and Gene Conversion. Molecular Cell *12*, 501–508.

Bassing, C.H., Swat, W., and Alt, F.W. (2002). The Mechanism and Regulation of Chromosomal V(D)J Recombination. Cell *109*, S45–S55.

Basu, U., Chaudhuri, J., Alpert, C., Dutt, S., Ranganath, S., Li, G., Schrum, J.P., Manis, J.P., and Alt, F.W. (2005). The AID antibody diversification enzyme is regulated by protein kinase A phosphorylation. Nature *438*, 508–511.

Basu, U., Meng, F.-L., Keim, C., Grinstein, V., Pefanis, E., Eccleston, J., Zhang, T., Myers, D., Wasserman, C.R., Wesemann, D.R., et al. (2011). The RNA Exosome Targets the AID Cytidine Deaminase to Both Strands of Transcribed Duplex DNA Substrates. Cell *144*, 353–363.

Betz, A.G., Milstein, C., González-Fernández, A., Pannell, R., Larson, T., and Neuberger, M.S. (1994). Elements regulating somatic hypermutation of an immunoglobulin κ gene: Critical role for the intron enhancer/matrix attachment region. Cell *77*, 239–248.

Bhutani, N., Brady, J.J., Damian, M., Sacco, A., Corbel, S.Y., and Blau, H.M. (2010). Reprogramming towards pluripotency requires AID-dependent DNA demethylation. Nature *463*, 1042–1047.

Blagodatski, A., Batrak, V., Schmidl, S., Schoetz, U., Caldwell, R.B., Arakawa, H., and Buerstedde, J.-M. (2009). A cis-Acting Diversification Activator Both Necessary and Sufficient for AID-Mediated Hypermutation. PLOS Genetics *5*, e1000332.

Bouamar, H., Abbas, S., Lin, A.-P., Wang, L., Jiang, D., Holder, K.N., Kinney, M.C., Hunicke-Smith, S., and Aguiar, R.C.T. (2013). A capture-sequencing strategy identifies IRF8, EBF1, and APRIL as novel IGH fusion partners in B-cell lymphoma. Blood *122*, 726–733.

Bransteitter, R., Pham, P., Scharff, M.D., and Goodman, M.F. (2003). Activation-induced cytidine deaminase deaminates deoxycytidine on single-stranded DNA but requires the action of RNase. PNAS *100*, 4102–4107.

Brar, S.S., Watson, M., and Diaz, M. (2004). Activation-induced Cytosine Deaminase (AID) Is Actively Exported out of the Nucleus but Retained by the Induction of DNA Breaks. J. Biol. Chem. *279*, 26395–26401.

Buerstedde, J.-M., Alinikula, J., Arakawa, H., McDonald, J.J., and Schatz, D.G. (2014b). Targeting Of Somatic Hypermutation By immunoglobulin Enhancer And Enhancer-Like Sequences. PLoS Biol *12*, e1001831.

Caratão, N., Cortesão, C.S., Reis, P.H., Freitas, R.F., Jacob, C.M.A., Pastorino, A.C., Carneiro-Sampaio, M., and Barreto, V.M. (2013). A novel activation-induced cytidine deaminase (AID) mutation in Brazilian patients with hyper-IgM type 2 syndrome. Clinical Immunology *148*, 279–286.

Chan, V.W.F., Meng, F., Soriano, P., DeFranco, A.L., and Lowell, C.A. (1997). Characterization of the B Lymphocyte Populations in Lyn-Deficient Mice and the Role of Lyn in Signal Initiation and Down-Regulation. Immunity *7*, 69–81.

Chaudhuri, J., Tian, M., Khuong, C., Chua, K., Pinaud, E., and Alt, F.W. (2003). Transcription-targeted DNA deamination by the AID antibody diversification enzyme. Nature *422*, 726–730.

Chiarle, R., Zhang, Y., Frock, R.L., Lewis, S.M., Molinie, B., Ho, Y.-J., Myers, D.R., Choi, V.W., Compagno, M., Malkin, D.J., et al. (2011). Genome-Wide Translocation Sequencing Reveals Mechanisms of Chromosome Breaks and Rearrangements in B Cells. Cell *147*, 107–119.

Compagno, M., Lim, W.K., Grunn, A., Nandula, S.V., Brahmachary, M., Shen, Q., Bertoni, F., Ponzoni, M., Scandurra, M., Califano, A., et al. (2009). Mutations of multiple genes cause deregulation of NF-κB in diffuse large B-cell lymphoma. Nature *459*, 717–721.

Crouch, E.E., Li, Z., Takizawa, M., Fichtner-Feigl, S., Gourzi, P., Montaño, C., Feigenbaum, L., Wilson, P., Janz, S., Papavasiliou, F.N., et al. (2007). Regulation of AID expression in the immune response. J Exp Med *204*, 1145–1156.

Dedeoglu, F., Horwitz, B., Chaudhuri, J., Alt, F.W., and Geha, R.S. (2004). Induction of activation-induced cytidine deaminase gene expression by IL-4 and CD40 ligation is dependent on STAT6 and NFkappaB. Int. Immunol. *16*, 395–404.

Delbos, F., Smet, A.D., Faili, A., Aoufouchi, S., Weill, J.-C., and Reynaud, C.-A. (2005). Contribution of DNA polymerase η to immunoglobulin gene hypermutation in the mouse. Journal of Experimental Medicine *201*, 1191–1196.

Delbos, F., Aoufouchi, S., Faili, A., Weill, J.-C., and Reynaud, C.-A. (2007). DNA polymerase η is the sole contributor of A/T modifications during immunoglobulin gene hypermutation in the mouse. Journal of Experimental Medicine *204*, 17–23.

Di Noia, J.M., and Neuberger, M.S. (2007). Molecular Mechanisms of Antibody Somatic Hypermutation. Annual Review of Biochemistry *76*, 1–22.

Di Noia, J.M., Rada, C., and Neuberger, M.S. (2006). SMUG1 is able to excise uracil from immunoglobulin genes: insight into mutation versus repair. The EMBO Journal *25*, 585–595.

Di Noia JM, and Neuberger MS (2004). Immunoglobulin gene conversion in chicken DT40 cells largely proceeds through an abasic site intermediate generated by excision of the uracil produced by AID-mediated deoxycytidine deamination. European Journal of Immunology *34*, 504–508.

Dickerson, S.K., Market, E., Besmer, E., and Papavasiliou, F.N. (2003). AID Mediates Hypermutation by Deaminating Single Stranded DNA. Journal of Experimental Medicine *197*, 1291–1296.

Domínguez, O., Ruiz, J.F., Laín de Lera, T., García-Díaz, M., González, M.A., Kirchhoff, T., Martínez-A, C., Bernad, A., and Blanco, L. (2000). DNA polymerase mu (Pol μ), homologous to TdT, could act as a DNA mutator in eukaryotic cells. EMBO J *19*, 1731–1742.

Dörner, T., Foster, S.J., Farner, N.L., and Lipsky, P.E. (1998). Somatic hypermutation of human immunoglobulin heavy chain genes: targeting of RGYW motifs on both DNA strands. Eur. J. Immunol. *28*, 3384–3396.

Dorsett, Y., McBride, K.M., Jankovic, M., Gazumyan, A., Thai, T.-H., Robbiani, D.F., Di Virgilio, M., San-Martin, B.R., Heidkamp, G., Schwickert, T.A., et al. (2008). MicroRNA-155 Suppresses Activation-Induced Cytidine Deaminase-Mediated Myc-Igh Translocation. Immunity *28*, 630–638.

Duke, J.L., Liu, M., Yaari, G., Khalil, A.M., Tomayko, M.M., Shlomchik, M.J., Schatz, D.G., and Kleinstein, S.H. (2013). Multiple Transcription Factor Binding Sites Predict AID Targeting in Non-Ig Genes. J Immunol *190*, 3878–3888.

Durandy, A. (2009). Immunoglobulin class switch recombination: study through human natural mutants. Philosophical Transactions of the Royal Society of London B: Biological Sciences *364*, 577–582.

Ehrenstein, M.R., and Neuberger, M.S. (1999). Deficiency in Msh2 affects the efficiency and local sequence specificity of immunoglobulin class-switch recombination: parallels with somatic hypermutation. The EMBO Journal *18*, 3484–3490.

Endo, Y., Marusawa, H., Kinoshita, K., Morisawa, T., Sakurai, T., Okazaki, I.-M., Watashi, K., Shimotohno, K., Honjo, T., and Chiba, T. (2007). Expression of activation-induced cytidine deaminase in human hepatocytes via NF-*κ*B signaling. Oncogene *26*, 5587.

Endo, Y., Marusawa, H., Kou, T., Nakase, H., Fujii, S., Fujimori, T., Kinoshita, K., Honjo, T., and Chiba, T. (2008). Activation-Induced Cytidine Deaminase Links Between Inflammation and the Development of Colitis-Associated Colorectal Cancers. Gastroenterology *135*, 889–898.e3.

Faili, A., Aoufouchi, S., Flatter, E., Guéranger, Q., Reynaud, C.-A., and Weill, J.-C. (2002). Induction of somatic hypermutation in immunoglobulin genes is dependent on DNA polymerase iota. Nature *419*, 944.

Faili, A., Stary, A., Delbos, F., Weller, S., Aoufouchi, S., Sarasin, A., Weill, J.-C., and Reynaud, C.-A. (2009). A Backup Role of DNA Polymerase κ in Ig Gene Hypermutation Only Takes Place in the Complete Absence of DNA Polymerase η. The Journal of Immunology *182*, 6353–6359.

Feng, J., Liu, T., and Zhang, Y. (2011). Using MACS to Identify Peaks from ChIP-Seq Data. In Current Protocols in Bioinformatics, (John Wiley & Sons, Inc.),

Frey, S., Bertocci, B., Delbos, F., Quint, L., Weill, J.-C., and Reynaud, C.-A. (1998). Mismatch Repair Deficiency Interferes with the Accumulation of Mutations in Chronically Stimulated B Cells and Not with the Hypermutation Process. Immunity *9*, 127–134.

Fukita, Y., Jacobs, H., and Rajewsky, K. (1998). Somatic Hypermutation in the Heavy Chain Locus Correlates with Transcription. Immunity *9*, 105–114.

Gonda, H., Sugai, M., Nambu, Y., Katakai, T., Agata, Y., Mori, K.J., Yokota, Y., and Shimizu, A. (2003). The Balance Between Pax5 and Id2 Activities Is the Key to AID Gene Expression. Journal of Experimental Medicine *198*, 1427–1437.

González-Fernández, A., and Milstein, C. (1993). Analysis of somatic hypermutation in mouse Peyer's patches using immunoglobulin kappa light-chain transgenes. PNAS *90*, 9862–9866.

Gordon, M.S., Kanegai, C.M., Doerr, J.R., and Wall, R. (2003). Somatic hypermutation of the B cell receptor genes B29 (Igβ, CD79b) and mb1 (Igα, CD79a). PNAS *100*, 4126–4131.

Green, M.R., Kihira, S., Liu, C.L., Nair, R.V., Salari, R., Gentles, A.J., Irish, J., Stehr, H., Vicente-Dueñas, C., Romero-Camarero, I., et al. (2015). Mutations in early follicular lymphoma progenitors are associated with suppressed antigen presentation. Proc. Natl. Acad. Sci. U.S.A. *112*, E1116-1125.

Guikema, J.E.J., Linehan, E.K., Tsuchimoto, D., Nakabeppu, Y., Strauss, P.R., Stavnezer, J., and Schrader, C.E. (2007). APE1- and APE2-dependent DNA breaks in immunoglobulin class switch recombination. J. Exp. Med. *204*, 3017–3026.

Harris, R.S., Sale, J.E., Petersen-Mahrt, S.K., and Neuberger, M.S. (2002). AID Is Essential for Immunoglobulin V Gene Conversion in a Cultured B Cell Line. Current Biology *12*, 435–438.

Hatem, A., Bozdağ, D., Toland, A.E., and Çatalyürek, Ü.V. (2013). Benchmarking short sequence mapping tools. BMC Bioinformatics *14*, 184.

Healy, J.A., Nugent, A., Rempel, R.E., Moffitt, A.B., Davis, N.S., Jiang, X., Shingleton, J.R., Zhang, J., Love, C., Datta, J., et al. (2016). GNA13 loss in germinal center B cells leads to impaired apoptosis and GC B cell persistence and promotes lymphoma in vivo. Blood blood-2015-07-659938.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. Mol. Cell *38*, 576–589.

Hibbs, M.L., Tarlinton, D.M., Armes, J., Grail, D., Hodgson, G., Maglitto, R., Stacker, S.A., and Dunn, A.R. (1995). Multiple defects in the immune system of Lyn-deficient mice, culminating in autoimmune disease. Cell *83*, 301–311.

Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-André, V., Sigova, A.A., Hoke, H.A., and Young, R.A. (2013). Super-Enhancers in the Control of Cell Identity and Disease. Cell *155*, 934–947.

Hnisz, D., Schuijers, J., Lin, C.Y., Weintraub, A.S., Abraham, B.J., Lee, T.I., Bradner, J.E., and Young, R.A. (2015). Convergence of Developmental and Oncogenic Signaling Pathways at Transcriptional Super-Enhancers. Molecular Cell *58*, 362–370.

Hogenbirk, M.A., Velds, A., Kerkhoven, R.M., and Jacobs, H. (2012). Reassessing genomic targeting of AID. Nat Immunol *13*, 797–798.

Hothorn, T., Hornik, K., and Zeileis, A. (2006). Unbiased Recursive Partitioning: A Conditional Inference Framework. J. Comput. Graph. Stat. *15*, 651–674.

Imai, K., Slupphaug, G., Lee, W.-I., Revy, P., Nonoyama, S., Catalan, N., Yel, L., Forveille, M., Kavli, B., Krokan, H.E., et al. (2003). Human uracil–DNA glycosylase deficiency associated with profoundly impaired immunoglobulin class-switch recombination. Nat Immunol *4*, 1023–1028.

Imai, K., Zhu, Y., Revy, P., Morio, T., Mizutani, S., Fischer, A., Nonoyama, S., and Durandy, A. (2005). Analysis of class switch recombination and somatic hypermutation in patients affected with autosomal dominant hyper-IgM syndrome type 2. Clinical Immunology *115*, 277–285.

Ingley, E. (2012). Functions of the Lyn tyrosine kinase in health and disease. Cell Commun. Signal *10*, 21.

Ito, S., Nagaoka, H., Shinkura, R., Begum, N., Muramatsu, M., Nakata, M., and Honjo, T. (2004). Activation-induced cytidine deaminase shuttles between nucleus and cytoplasm like apolipoprotein B mRNA editing catalytic polypeptide 1. PNAS *101*, 1975–1980.

Jansen, J.G., Langerak, P., Tsaalbi-Shtylik, A., van den Berk, P., Jacobs, H., and de Wind, N. (2006). Strand-biased defect in C/G transversions in hypermutating immunoglobulin genes in Rev1-deficient mice. J. Exp. Med. *203*, 319–323.

Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orlando, D.A., van Berkum, N.L., Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S., et al. (2010). Mediator and Cohesin Connect Gene Expression and Chromatin Architecture. Nature *467*, 430–435.

Karasuyama, H., Kudo, A., and Melchers, F. (1990). The proteins encoded by the VpreB and lambda 5 pre-B cell-specific genes can associate with each other and with mu heavy chain. Journal of Experimental Medicine *172*, 969–972.

Keane, T.M., Goodstadt, L., Danecek, P., White, M.A., Wong, K., Yalcin, B., Heger, A., Agam, A., Slater, G., Goodson, M., et al. (2011). Mouse genomic variation and its effect on phenotypes and gene regulation. Nature *477*, 289–294.

Kenter, A.L., Kumar, S., Wuerffel, R., and Grigera, F. (2016). AID hits the jackpot when missing the target. Current Opinion in Immunology *39*, 96–102.

Kieffer-Kwon, K.-R., Tang, Z., Mathe, E., Qian, J., Sung, M.-H., Li, G., Resch, W., Baek, S., Pruett, N., Grøntved, L., et al. (2013). Interactome Maps of Mouse Gene Regulatory Domains Reveal Basic Principles of Transcriptional Regulation. Cell *155*, 1507–1520.

Kitamura, D., Roes, J., Kühn, R., and Rajewsky, K. (1991). A B cell-deficient mouse by targeted disruption of the membrane exon of the immunoglobulin μ chain gene. Nature *350*, 423–426.

Klein, I.A., Resch, W., Jankovic, M., Oliveira, T., Yamane, A., Nakahashi, H., Di Virgilio, M., Bothmer, A., Nussenzweig, A., Robbiani, D.F., et al. (2011). Translocation-Capture Sequencing Reveals the Extent and Nature of Chromosomal Rearrangements in B Lymphocytes. Cell *147*, 95–106.

Kou Tadayuki, Marusawa Hiroyuki, Kinoshita Kazuo, Endo Yoko, Okazaki Il-mi, Ueda Yoshihide, Kodama Yuzo, Haga Hironori, Ikai Iwao, and Chiba Tsutomu (2006). Expression of activation-induced cytidine deaminase in human hepatocytes during hepatocarcinogenesis. International Journal of Cancer *120*, 469–476.

Kovalchuk, A.L., duBois, W., Mushinski, E., McNeil, N.E., Hirt, C., Qi, C.-F., Li, Z., Janz, S., Honjo, T., Muramatsu, M., et al. (2007). AID-deficient Bcl-xL transgenic mice develop delayed atypical plasma cell tumors with unusual Ig/Myc chromosomal rearrangements. Journal of Experimental Medicine *204*, 2989–3001.

Krijger, P.H.L., Langerak, P., Berk, P.C.M. van den, and Jacobs, H. (2009). Dependence of nucleotide substitutions on Ung2, Msh2, and PCNA-Ub during somatic hypermutation. J Exp Med *206*, 2603–2611.

Krokan, H.E., Sætrom, P., Aas, P.A., Pettersen, H.S., Kavli, B., and Slupphaug, G. (2014). Error-free versus mutagenic processing of genomic uracil—Relevance to cancer. DNA Repair *19*, 38–47.

Kuchen, S., Resch, W., Yamane, A., Kuo, N., Li, Z., Chakraborty, T., Wei, L., Laurence, A., Yasuda, T., Peng, S., et al. (2010). Regulation of MicroRNA Expression and Abundance during Lymphopoiesis. Immunity *32*, 828–839.

Lam, K.-P., Kühn, R., and Rajewsky, K. (1997). In Vivo Ablation of Surface Immunoglobulin on Mature B Cells by Inducible Gene Targeting Results in Rapid Cell Death. Cell *90*, 1073–1083.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. *10*, R25.

Le, Q., and Maizels, N. (2015). Cell Cycle Regulates Nuclear Stability of AID and Determines the Cellular Response to AID. PLoS Genet *11*.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.

Li, H., and Homer, N. (2010). A survey of sequence alignment algorithms for next-generation sequencing. Brief. Bioinform. *11*, 473–483.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Subgroup, 1000 Genome Project Data Processing (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Libbrecht, M.W., and Noble, W.S. (2015). Machine learning applications in genetics and genomics. Nature Reviews Genetics *16*, 321–332.

Liu, M., Duke, J.L., Richter, D.J., Vinuesa, C.G., Goodnow, C.C., Kleinstein, S.H., and Schatz, D.G. (2008). Two levels of protection for the B cell genome during somatic hypermutation. Nature *451*, 841–845.

Lohr, J.G., Stojanov, P., Lawrence, M.S., Auclair, D., Chapuy, B., Sougnez, C., Cruz-Gordillo, P., Knoechel, B., Asmann, Y.W., Slager, S.L., et al. (2012). Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. PNAS *109*, 3879–3884.

Love, C., Sun, Z., Jima, D., Li, G., Zhang, J., Miles, R., Richards, K.L., Dunphy, C.H., Choi, W.W.L., Srivastava, G., et al. (2012). The genetic landscape of mutations in Burkitt lymphoma. Nat Genet *44*, 1321–1325.

Maizels, N. (1995). Somatic hypermutation: How many mechanisms diversify V region sequences? Cell *83*, 9–12.

Mao, B., Zhang, Z., and Wang, G. (2014). BTG2: A rising star of tumor suppressors (Review). International Journal of Oncology.

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal *17*, 10–12.

Matsuda, F., Ishii, K., Bourvagnet, P., Kuma, K., Hayashida, H., Miyata, T., and Honjo, T. (1998). The Complete Nucleotide Sequence of the Human Immunoglobulin Heavy Chain Variable Region Locus. Journal of Experimental Medicine *188*, 2151–2162.

Matsumoto, Y., Marusawa, H., Kinoshita, K., Endo, Y., Kou, T., Morisawa, T., Azuma, T., Okazaki, I.-M., Honjo, T., and Chiba, T. (2007). Helicobacter pylori infection triggers aberrant expression of activation-induced cytidine deaminase in gastric epithelium. Nature Medicine *13*, 470.

Maul, R.W., Saribasak, H., Martomo, S.A., McClure, R.L., Yang, W., Vaisman, A., Gramlich, H.S., Schatz, D.G., Woodgate, R., Iii, D.M.W., et al. (2011). Uracil residues dependent on the deaminase AID in immunoglobulin gene variable and switch regions. Nature Immunology *12*, 70–76.

Maul, R.W., Saribasak, H., Cao, Z., and Gearhart, P.J. (2015). Topoisomerase I deficiency causes RNA polymerase II accumulation and increases AID abundance in immunoglobulin variable genes. DNA Repair *30*, 46–52.

Maul, R.W., MacCarthy, T., Frank, E.G., Donigan, K.A., McLenigan, M.P., Yang, W., Saribasak, H., Huston, D.E., Lange, S.S., Woodgate, R., et al. (2016). DNA polymerase ι functions in the generation of tandem mutations during somatic hypermutation of antibody genes. J Exp Med jem.20151227.

Max, E.E., Seidman, J.G., and Leder, P. (1979). Sequences of five potential recombination sites encoded close to an immunoglobulin kappa constant region gene. Proc Natl Acad Sci U S A *76*, 3450–3454.

McBride, K.M., Barreto, V., Ramiro, A.R., Stavropoulos, P., and Nussenzweig, M.C. (2004). Somatic Hypermutation Is Limited by CRM1-dependent Nuclear Export of Activation-induced Deaminase. Journal of Experimental Medicine *199*, 1235–1244.

McBride, K.M., Gazumyan, A., Woo, E.M., Barreto, V.M., Robbiani, D.F., Chait, B.T., and Nussenzweig, M.C. (2006). Regulation of hypermutation by activation-induced cytidine deaminase phosphorylation. PNAS *103*, 8798–8803.

McBride, K.M., Gazumyan, A., Woo, E.M., Schwickert, T.A., Chait, B.T., and Nussenzweig, M.C. (2008). Regulation of class switch recombination and somatic mutation by AID phosphorylation. Journal of Experimental Medicine *205*, 2585–2594.

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol. *28*, 495–501.

Meng, F.-L., Du, Z., Federation, A., Hu, J., Wang, Q., Kieffer-Kwon, K.-R., Meyers, R.M., Amor, C., Wasserman, C.R., Neuberg, D., et al. (2014). Convergent Transcription at Intragenic Super-Enhancers Targets AID-Initiated Genomic Instability. Cell *159*, 1538–1548.

Mesin, L., Ersching, J., and Victora, G.D. (2016). Germinal Center B Cell Dynamics. Immunity *45*, 471–482.

Methot, S.P., and Di Noia, J.M. (2017). Molecular Mechanisms of Somatic Hypermutation and Class Switch Recombination. Advances in Immunology *133*, 37–87.

Methot, S.P., Litzler, L.C., Trajtenberg, F., Zahn, A., Robert, F., Pelletier, J., Buschiazzo, A., Magor, B.G., and Noia, J.M.D. (2015). Consecutive interactions with HSP90 and eEF1A underlie a functional maturation and storage pathway of AID in the cytoplasm. Journal of Experimental Medicine *212*, 581–596.

Meyers, G., Ng, Y.-S., Bannock, J.M., Lavoie, A., Walter, J.E., Notarangelo, L.D., Kilic, S.S., Aksu, G., Debré, M., Rieux-Laucat, F., et al. (2011). Activation-induced cytidine deaminase (AID) is required for B-cell tolerance in humans. PNAS *108*, 11554–11559.

Miranda, N.F.C.C. de, Georgiou, K., Chen, L., Wu, C., Gao, Z., Zaravinos, A., Lisboa, S., Enblad, G., Teixeira, M.R., Zeng, Y., et al. (2014). Exome sequencing reveals novel mutation targets in diffuse large B-cell lymphomas derived from Chinese patients. Blood *124*, 2544–2553.

Morgan, H.D., Dean, W., Coker, H.A., Reik, W., and Petersen-Mahrt, S.K. (2004). Activation-induced Cytidine Deaminase Deaminates 5-Methylcytosine in DNA and Is Expressed in Pluripotent Tissues IMPLICATIONS FOR EPIGENETIC REPROGRAMMING. J. Biol. Chem. *279*, 52353–52360.

Morin, R.D., Mungall, K., Pleasance, E., Mungall, A.J., Goya, R., Huff, R.D., Scott, D.W., Ding, J., Roth, A., Chiu, R., et al. (2013). Mutational and structural analysis of diffuse large B-cell lymphoma using whole-genome sequencing. Blood *122*, 1256–1265.

Muramatsu, M., Sankaranand, V.S., Anant, S., Sugai, M., Kinoshita, K., Davidson, N.O., and Honjo, T. (1999). Specific Expression of Activation-induced Cytidine Deaminase (AID), a Novel Member of the RNA-editing Deaminase Family in Germinal Center B Cells. J. Biol. Chem. *274*, 18470–18476.

Muramatsu, M., Kinoshita, K., Fagarasan, S., Yamada, S., Shinkai, Y., and Honjo, T. (2000). Class Switch Recombination and Hypermutation Require Activation-Induced Cytidine Deaminase (AID), a Potential RNA Editing Enzyme. Cell *102*, 553–563.

Müschen, M., Re, D., Jungnickel, B., Diehl, V., Rajewsky, K., and Küppers, R. (2000). Somatic Mutation of the Cd95 Gene in Human B Cells as a Side-Effect of the Germinal Center Reaction. J Exp Med *192*, 1833–1840.

Nambu, Y., Sugai, M., Gonda, H., Lee, C.-G., Katakai, T., Agata, Y., Yokota, Y., and Shimizu, A. (2003). Transcription-Coupled Events Associating with Immunoglobulin Switch Region Chromatin. Science *302*, 2137–2140.

Nelson, J.R., Lawrence, C.W., and Hinkle, D.C. (1996). Deoxycytidyl transferase activity of yeast REV1 protein. Nature *382*, 729–731.

Nilsen, H., Rosewell, I., Robins, P., Skjelbred, C.F., Andersen, S., Slupphaug, G., Daly, G., Krokan, H.E., Lindahl, T., and Barnes, D.E. (2000). Uracil-DNA Glycosylase (UNG)-Deficient Mice Reveal a Primary Role of the Enzyme during DNA Replication. Molecular Cell *5*, 1059–1065.

Noia, J.D., and Neuberger, M.S. (2002). Altering the pathway of immunoglobulin hypermutation by inhibiting uracil-DNA glycosylase. Nature *419*, 43.

Nowak, U., Matthews, A., Zheng, S., and Chaudhuri, J. (2011). The splicing regulator PTBP2 is an AID interacting protein and promotes binding of AID to switch region DNA. Nat Immunol *12*, 160–166.

Okosun, J., Bödör, C., Wang, J., Araf, S., Yang, C.-Y., Pan, C., Boller, S., Cittaro, D., Bozek, M., Iqbal, S., et al. (2014). Integrated genomic analysis identifies recurrent mutations and evolution patterns driving the initiation and progression of follicular lymphoma. Nat Genet *46*, 176–181.

Orthwein, A., and Di Noia, J.M. (2012). Activation induced deaminase: How much and where? Seminars in Immunology *24*, 246–254.

Orthwein, A., Patenaude, A.-M., Affar, E.B., Lamarre, A., Young, J.C., and Noia, J.M.D. (2010). Regulation of activation-induced deaminase stability and antibody gene diversification by Hsp90. Journal of Experimental Medicine *207*, 2751–2765.

Parsa, J.-Y., Ramachandran, S., Zaheen, A., Nepal, R.M., Kapelnikov, A., Belcheva, A., Berru, M., Ronai, D., and Martin, A. (2012). Negative Supercoiling Creates Single-Stranded Patches of DNA That Are Substrates for AID–Mediated Mutagenesis. PLOS Genetics *8*, e1002518.

Pasqualucci, L., Migliazza, A., Fracchiolla, N., William, C., Neri, A., Baldini, L., Chaganti, R.S., Klein, U., Küppers, R., Rajewsky, K., et al. (1998). BCL-6 mutations in normal germinal center B cells: evidence of somatic hypermutation acting outside Ig loci. Proc. Natl. Acad. Sci. U.S.A. *95*, 11816–11821.

Pasqualucci, L., Neumeister, P., Goossens, T., Nanjangud, G., Chaganti, R.S.K., Küppers, R., and Dalla-Favera, R. (2001). Hypermutation of multiple proto-oncogenes in B-cell diffuse large-cell lymphomas. Nature *412*, 341–346.

Pasqualucci, L., Bhagat, G., Jankovic, M., Compagno, M., Smith, P., Muramatsu, M., Honjo, T., Morse, H.C., Nussenzweig, M.C., and Dalla-Favera, R. (2008). AID is required for germinal center–derived lymphomagenesis. Nat Genet *40*, 108–112.

Pasqualucci, L., Dominguez-Sola, D., Chiarenza, A., Fabbri, G., Grunn, A., Trifonov, V., Kasper, L.H., Lerach, S., Tang, H., Ma, J., et al. (2011). Inactivating mutations of acetyltransferase genes in B-cell lymphoma. Nature *471*, 189–195.

Pasqualucci, L., Khiabanian, H., Fangazio, M., Vasishtha, M., Messina, M., Holmes, A.B., Ouillette, P., Trifonov, V., Rossi, D., Tabbò, F., et al. (2014). Genetics of follicular lymphoma transformation. Cell Rep *6*, 130–140.

Patenaude, A.-M., Orthwein, A., Hu, Y., Campo, V.A., Kavli, B., Buschiazzo, A., and Noia, J.M.D. (2009). Active nuclear import and cytoplasmic retention of activation-induced deaminase. Nature Structural & Molecular Biology *16*, 517–527.

Pavri, R. (2017). R Loops in the Regulation of Antibody Gene Diversification. Genes *8*, 154.

Pavri, R., Gazumyan, A., Jankovic, M., Di Virgilio, M., Klein, I., Ansarah-Sobrinho, C., Resch, W., Yamane, A., San-Martin, B.R., Barreto, V., et al. (2010). Activation-Induced Cytidine Deaminase Targets DNA at Sites of RNA Polymerase II Stalling by Interaction with Spt5. Cell *143*, 122–133.

Pefanis, E., Wang, J., Rothschild, G., Lim, J., Chao, J., Rabadan, R., Economides, A.N., and Basu, U. (2014). Noncoding RNA transcription targets AID to divergently transcribed loci in B cells. Nature *advance online publication*.

Pelanda, R., and Torres, R.M. (2012). Central B-Cell Tolerance: Where Selection Begins. Cold Spring Harb Perspect Biol *4*, a007146.

Pérez-Durán, P., Belver, L., Yébenes, V.G. de, Delgado, P., Pisano, D.G., and Ramiro, A.R. (2012). UNG shapes the specificity of AID-induced somatic hypermutation. J Exp Med *209*, 1379–1389.

Pérez-García, A., Pérez-Durán, P., Wossning, T., Sernandez, I.V., Mur, S.M., Cañamero, M., Real, F.X., and Ramiro, A.R. (2015). AID-expressing epithelium is protected from oncogenic transformation by an NKG2D surveillance pathway. EMBO Molecular Medicine *7*, 1327–1336.

Pérez-García, A., Marina-Zárate, E., Álvarez-Prado, Á., Ligos, J.M., Galjart, N., and Ramiro, A.R. (2017). CTCF orchestrates the germinal centre transcriptional program and prevents premature plasma cell differentiation., CTCF orchestrates the germinal centre transcriptional program and prevents premature plasma cell differentiation. Nat Commun *8*, *8*, 16067–16067.

Peters, A., and Storb, U. (1996). Somatic Hypermutation of Immunoglobulin Genes Is Linked to Transcription Initiation. Immunity *4*, 57–65.

Petersen-Mahrt, S.K., Harris, R.S., and Neuberger, M.S. (2002). AID mutates E. coli suggesting a DNA deamination mechanism for antibody diversification. Nature *418*, 99–104.

Pham, P., Bransteitter, R., Petruska, J., and Goodman, M.F. (2003). Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. Nature *424*, 103–107.

Pham, P., Afif, S.A., Shimoda, M., Maeda, K., Sakaguchi, N., Pedersen, L.C., and Goodman, M.F. (2016). Structural analysis of the activation-induced deoxycytidine deaminase required in immunoglobulin diversification. DNA Repair (Amst) *43*, 48–56.

Pham, P., Afif, S.A., Shimoda, M., Maeda, K., Sakaguchi, N., Pedersen, L.C., and Goodman, M.F. (2017). Activation-induced deoxycytidine deaminase: Structural basis for favoring WRC hot motif specificities unique among APOBEC family members. DNA Repair *54*, 8–12.

Phung, Q.H., Winter, D.B., Cranston, A., Tarone, R.E., Bohr, V.A., Fishel, R., and Gearhart, P.J. (1998). Increased Hypermutation at G and C Nucleotides in Immunoglobulin Variable Genes from Mice Deficient in the MSH2 Mismatch Repair Protein. J Exp Med *187*, 1745–1751.

Pinaud, E., Marquet, M., Fiancette, R., Péron, S., Vincent-Fabert, C., Denizot, Y., and Cogné, M. (2011). Chapter 2 - The IgH Locus 3′ Regulatory Region: Pulling the Strings from Behind. In Advances in Immunology, F.W. Alt, K.F. Austen, T. Honj, F. Melchers, J.W. Uhr, and E.R. Unanue, eds. (Academic Press), pp. 27–70.

Poltoratsky, V., Goodman, M.F., and Scharff, M.D. (2000). Error-Prone Candidates Vie for Somatic Mutation. Journal of Experimental Medicine *192*, F27–F30.

Pon, J.R., Wong, J., Saberi, S., Alder, O., Moksa, M., Grace Cheng, S.-W., Morin, G.B., Hoodless, P.A., Hirst, M., and Marra, M.A. (2015). MEF2B mutations in non-Hodgkin lymphoma dysregulate cell migration by decreasing MEF2B target gene activation. Nature Communications *6*, 7953.

Popp, C., Dean, W., Feng, S., Cokus, S.J., Andrews, S., Pellegrini, M., Jacobsen, S.E., and Reik, W. (2010). Genome-wide erasure of DNA methylation in mouse primordial germ cells is affected by AID deficiency. Nature *463*, 1101–1105.

Qian, J., Wang, Q., Dose, M., Pruett, N., Kieffer-Kwon, K.-R., Resch, W., Liang, G., Tang, Z., Mathé, E., Benner, C., et al. (2014). B Cell Super-Enhancers and Regulatory Clusters Recruit AID Tumorigenic Activity. Cell.

Qiao, Q., Wang, L., Meng, F.-L., Hwang, J.K., Alt, F.W., and Wu, H. (2017). AID Recognizes Structured DNA for Class Switch Recombination. Molecular Cell *67*, 361–373.e4.

Quartier, P., Bustamante, J., Sanal, O., Plebani, A., Debré, M., Deville, A., Litzman, J., Levy, J., Fermand, J.-P., Lane, P., et al. (2004). Clinical, immunologic and genetic analysis of 29 patients with autosomal recessive hyper-IgM syndrome due to Activation-Induced Cytidine Deaminase deficiency. Clinical Immunology *110*, 22–29.

Quinlan, A.R. (2002). BEDTools: The Swiss-Army Tool for Genome Feature Analysis. In Current Protocols in Bioinformatics, (John Wiley & Sons, Inc.)

Rada, C., Ehrenstein, M.R., Neuberger, M.S., and Milstein, C. (1998). Hot Spot Focusing of Somatic Hypermutation in MSH2-Deficient Mice Suggests Two Stages of Mutational Targeting. Immunity *9*, 135–141.

Rada, C., Williams, G.T., Nilsen, H., Barnes, D.E., Lindahl, T., and Neuberger, M.S. (2002). Immunoglobulin Isotype Switching Is Inhibited and Somatic Hypermutation Perturbed in UNG-Deficient Mice. Current Biology *12*, 1748–1755.

Rada, C., Di Noia, J.M., and Neuberger, M.S. (2004). Mismatch Recognition and Uracil Excision Provide Complementary Paths to Both Ig Switching and the A/T-Focused Phase of Somatic Mutation. Molecular Cell *16*, 163–171.

Ramiro, A.R., and Barreto, V.M. (2015). Activation-induced cytidine deaminase and active cytidine demethylation. Trends in Biochemical Sciences *40*, 172–181.

Ramiro, A.R., Stavropoulos, P., Jankovic, M., and Nussenzweig, M.C. (2003). Transcription enhances AID-mediated cytidine deamination by exposing single-stranded DNA on the nontemplate strand. Nat Immunol *4*, 452–456.

Ramiro, A.R., Jankovic, M., Eisenreich, T., Difilippantonio, S., Chen-Kiang, S., Muramatsu, M., Honjo, T., Nussenzweig, A., and Nussenzweig, M.C. (2004). AID Is Required for c-myc/IgH Chromosome Translocations In Vivo. Cell *118*, 431–438.

Ramiro, A.R., Jankovic, M., Callen, E., Difilippantonio, S., Chen, H.-T., McBride, K.M., Eisenreich, T.R., Chen, J., Dickins, R.A., Lowe, S.W., et al. (2006). Role of genomic instability and p53 in AID-induced c-myc–Igh translocations. Nature *440*, 105–109.

Ranjit, S., Khair, L., Linehan, E.K., Ucher, A.J., Chakrabarti, M., Schrader, C.E., and Stavnezer, J. (2011). AID Binds Cooperatively with UNG and Msh2-Msh6 to Ig Switch Regions Dependent upon the AID C Terminus. The Journal of Immunology *187*, 2464–2475.

Reitmair, A.H., Schmits, R., Ewel, A., Bapat, B., Redston, M., Mitri, A., Waterhouse, P., Mittrücker, H.-W., Wakeham, A., Liu, B., et al. (1995). MSH2 deficient mice are viable and susceptible to lymphoid tumours. Nat Genet *11*, 64–70.

Retter, I., Chevillard, C., Scharfe, M., Conrad, A., Hafner, M., Im, T.-H., Ludewig, M., Nordsiek, G., Severitt, S., Thies, S., et al. (2007). Sequence and characterization of the Ig heavy chain constant and partial variable region of the mouse strain 129S1. J. Immunol. *179*, 2419–2427.

Revy, P., Muto, T., Levy, Y., Geissmann, F., Plebani, A., Sanal, O., Catalan, N., Forveille, M., Dufourcq-Lagelouse, R., Gennery, A., et al. (2000). Activation-Induced Cytidine Deaminase (AID) Deficiency Causes the Autosomal Recessive Form of the Hyper-IgM Syndrome (HIGM2). Cell *102*, 565–575.

Roa, S., Li, Z., Peled, J.U., Zhao, C., Edelmann, W., and Scharff, M.D. (2010). MSH2/MSH6 Complex Promotes Error-Free Repair of AID-Induced dU:G Mispairs as well as Error-Prone Hypermutation of A:T Sites. PLOS ONE *5*, e11182.

Robbiani, D.F., Bothmer, A., Callen, E., Reina-San-Martin, B., Dorsett, Y., Difilippantonio, S., Bolland, D.J., Chen, H.T., Corcoran, A.E., Nussenzweig, A., et al. (2008). AID Is Required for the Chromosomal Breaks in c-myc that Lead to c-myc/IgH Translocations. Cell *135*, 1028–1038.

Robbiani, D.F., Bunting, S., Feldhahn, N., Bothmer, A., Camps, J., Deroubaix, S., McBride, K.M., Klein, I.A., Stone, G., Eisenreich, T.R., et al. (2009). AID Produces DNA Double-Strand Breaks in Non-Ig Genes and Mature B Cell Lymphomas with Reciprocal Chromosome Translocations. Molecular Cell *36*, 631–641.

Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinforma. Oxf. Engl. *26*, 139–140.

Rogozin, I.B., and Diaz, M. (2004). Cutting Edge: DGYW/WRCH Is a Better Predictor of Mutability at G:C Bases in Ig Hypermutation Than the Widely Accepted RGYW/WRCY Motif and Probably Reflects a Two-Step Activation-Induced Cytidine Deaminase-Triggered Process. J Immunol *172*, 3382–3384.

Rogozin, I.B., and Kolchanov, N.A. (1992). Somatic hypermutagenesis in immunoglobulin genes. II. Influence of neighbouring base sequences on mutagenesis. Biochimica et Biophysica Acta *1171*, 11–18.

Romanello, M., Schiavone, D., Frey, A., and Sale, J.E. (2016). Histone H3.3 promotes IgV gene diversification by enhancing formation of AID-accessible single-stranded DNA. EMBO J *35*, 1452–1464.

Ronai, D., Iglesias-Ussel, M.D., Fan, M., Li, Z., Martin, A., and Scharff, M.D. (2007). Detection of chromatin-associated single-stranded DNA in regions targeted for somatic hypermutation. Journal of Experimental Medicine *204*, 181–190.

Rouaud, P., Vincent-Fabert, C., Saintamand, A., Fiancette, R., Marquet, M., Robert, I., Reina-San-Martin, B., Pinaud, E., Cogné, M., and Denizot, Y. (2013). The IgH 3′ regulatory region controls somatic hypermutation in germinal center B cells. Journal of Experimental Medicine *210*, 1501–1507.

Rouault, J.-P., Falette, N., Guéhenneux, F., Guillot, C., Rimokh, R., Wang, Q., Berthet, C., Moyret-Lalle, C., Savatier, P., Pain, B., et al. (1996). Identification of BTG2, an antiproliferative p53–dependent component of the DNA damage cellular response pathway. Nat Genet *14*, 482–486.

Ruiz, J.F., Domínguez, O., Lera, T.L. de, García–Díaz, M., Bernad, A., and Blanco, L. (2001). DNA polymerase mu, a candidate hypermutase? Philosophical Transactions of the Royal Society of London B: Biological Sciences *356*, 99–109.

Ruiz, J.F., Lucas, D., García-Palomero, E., Saez, A.I., González, M.A., Piris, M.A., Bernad, A., and Blanco, L. (2004). Overexpression of human DNA polymerase μ (Pol μ) in a Burkitt's lymphoma cell line affects the somatic hypermutation rate. Nucleic Acids Res *32*, 5861–5873.

Sakano, H., Hüppi, K., Heinrich, G., and Tonegawa, S. (1979). Sequences at the somatic recombination sites of immunoglobulin light-chain genes. Nature *280*, 288–294.

Sale, J.E., Calandrini, D.M., Takata, M., Takeda, S., and Neuberger, M.S. (2001). Ablation of XRCC2/3 transforms immunoglobulin V gene conversion into somatic hypermutation. Nature *412*, 921.

Santos-Pereira, J.M., and Aguilera, A. (2015). R loops: new modulators of genome dynamics and function. Nature Reviews Genetics *16*, 583.

Saribasak, H., Saribasak, N.N., Ipek, F.M., Ellwart, J.W., Arakawa, H., and Buerstedde, J.-M. (2006). Uracil DNA Glycosylase Disruption Blocks Ig Gene Conversion and Induces Transition Mutations. The Journal of Immunology *176*, 365–371.

Schmitz, R., Hansmann, M.-L., Bohle, V., Martin-Subero, J.I., Hartmann, S., Mechtersheimer, G., Klapper, W., Vater, I., Giefing, M., Gesk, S., et al. (2009). TNFAIP3 (A20) is a tumor suppressor gene in Hodgkin lymphoma and primary mediastinal B cell lymphoma. J. Exp. Med. *206*, 981–989.

Schreck S, Buettner M, Kremmer E, Bogdan M, Herbst H, and Niedobitek G (2006). Activation-induced cytidine deaminase (AID) is expressed in normal spermatogenesis but only infrequently in testicular germ cell tumours. The Journal of Pathology *210*, 26–31.

Seidman, J.G., Max, E.E., and Leder, P. (1979). A kappa-immunoglobulin gene is formed by site-specific recombination without further somatic mutation. Nature *280*, 370–375.

Seki, M., Gearhart, P.J., and Wood, R.D. (2005). DNA polymerases and somatic hypermutation of immunoglobulin genes. EMBO Rep *6*, 1143–1148.

Sernández, I.V., de Yébenes, V.G., Dorsett, Y., and Ramiro, A.R. (2008). Haploinsufficiency of Activation-Induced Deaminase for Antibody Diversification and Chromosome Translocations both In Vitro and In Vivo. PLoS ONE *3*, e3927.

Shaffer, A.L., and Staudt, L.M. (2012). Pathogenesis of Human B Cell Lymphomas. Annual Review of Immunology *30*, 565–610.

Shaffer, A.L., Lin, K.-I., Kuo, T.C., Yu, X., Hurt, E.M., Rosenwald, A., Giltnane, J.M., Yang, L., Zhao, H., Calame, K., et al. (2002). Blimp-1 Orchestrates Plasma Cell Differentiation by Extinguishing the Mature B Cell Gene Expression Program. Immunity *17*, 51–62.

Sharbeen, G., Yee, C.W.Y., Smith, A.L., and Jolly, C.J. (2012). Ectopic restriction of DNA repair reveals that UNG2 excises AID-induced uracils predominantly or exclusively during G1 phase. Journal of Experimental Medicine *209*, 965–974.

Shen, H.M., Peters, A., Baron, B., Zhu, X., and Storb, U. (1998). Mutation of BCL-6 Gene in Normal B Cells by the Process of Somatic Hypermutation of Ig Genes. Science *280*, 1750–1752.

Shen, H.M., Tanaka, A., Bozek, G., Nicolae, D., and Storb, U. (2006). Somatic Hypermutation and Class Switch Recombination in Msh6−/−Ung−/− Double-Knockout Mice. J Immunol *177*, 5386–5392.

Sheppard, E.C., Morrish, R.B., Dillon, M.J., Leyland, R., and Chahwan, R. (2018). Epigenomic Modifications Mediating Antibody Maturation. Front. Immunol. *9*.

Shinkura, R., Ito, S., Begum, N.A., Nagaoka, H., Muramatsu, M., Kinoshita, K., Sakakibara, Y., Hijikata, H., and Honjo, T. (2004). Separate domains of AID are required for somatic hypermutation and class-switch recombination. Nature Immunology *5*, 707–712.

Soutourina, J. (2018). Transcription regulation by the Mediator complex. Nature Reviews Molecular Cell Biology *19*, 262.

Staszewski, O., Baker, R.E., Ucher, A.J., Martier, R., Stavnezer, J., and Guikema, J.E.J. (2011). Activation-Induced Cytidine Deaminase Induces Reproducible DNA Breaks at Many Non-Ig Loci in Activated B Cells. Molecular Cell *41*, 232–242.

Stavnezer, J. (2011). Complex regulation and function of activation-induced cytidine deaminase. Trends in Immunology *32*, 194–201.

Stavnezer, J., and Schrader, C.E. (2014). IgH Chain Class Switch Recombination: Mechanism and Regulation. The Journal of Immunology *193*, 5370–5378.

Stavnezer, J., Guikema, J.E.J., and Schrader, C.E. (2008). Mechanism and Regulation of Class Switch Recombination. Annual Review of Immunology *26*, 261–292.

Steidl, C., Shah, S.P., Woolcock, B.W., Rui, L., Kawahara, M., Farinha, P., Johnson, N.A., Zhao, Y., Telenius, A., Neriah, S.B., et al. (2011). MHC class II transactivator CIITA is a recurrent gene fusion partner in lymphoid cancers. Nature *471*, 377–381.

Storb, U. (2014). Chapter Seven - Why Does Somatic Hypermutation by AID Require Transcription of Its Target Genes? In Advances in Immunology, F.W. Alt, ed. (Academic Press), pp. 253–277.

Storb, U., Klotz, E.L., Hackett, J., Kage, K., Bozek, G., and Martin, T.E. (1998). A hypermutable insert in an immunoglobulin transgene contains hotspots of somatic mutation and sequences predicting highly stable structures in the RNA transcript. J. Exp. Med. *188*, 689–698.

Ta, V.-T., Nagaoka, H., Catalan, N., Durandy, A., Fischer, A., Imai, K., Nonoyama, S., Tashiro, J., Ikegawa, M., Ito, S., et al. (2003). AID mutant analyses indicate requirement for class-switch-specific cofactors. Nature Immunology *4*, 843–848.

Taatjes, D.J. (2010). The human Mediator complex: a versatile, genome-wide regulator of transcription. Trends in Biochemical Sciences *35*, 315–322.

Takai, A., Marusawa, H., Minaki, Y., Watanabe, T., Nakase, H., Kinoshita, K., Tsujimoto, G., and Chiba, T. (2012). Targeting activation-induced cytidine deaminase prevents colon cancer development despite persistent colonic inflammation. Oncogene *31*, 1733.

Takizawa, M., Tolarová, H., Li, Z., Dubois, W., Lim, S., Callen, E., Franco, S., Mosaico, M., Feigenbaum, L., Alt, F.W., et al. (2008). AID expression levels determine the extent of cMyc oncogenic translocations and the incidence of B cell tumor development. Journal of Experimental Medicine *205*, 1949–1957.

Tang, E.S., and Martin, A. (2007). Immunoglobulin gene conversion: Synthesizing antibody diversification and DNA repair. DNA Repair *6*, 1557–1571.

Tasuku Honjo, Michael Reth, Andreas Radbruch, Frederick Alt (2015). Molecular Biology of B Cells - 2nd Edition (Elsevier Academic Prints).

Tauzin, S., Ding, H., Khatib, K., Ahmad, I., Burdevet, D., van Echten-Deckert, G., Lindquist, J.A., Schraven, B., Din, N.-U.-, Borisch, B., et al. (2008). Oncogenic association of the Cbp/PAG adaptor protein with the Lyn tyrosine kinase in human B-NHL rafts. Blood *111*, 2310–2320.

Teng, G., Hakimpour, P., Landgraf, P., Rice, A., Tuschl, T., Casellas, R., and Papavasiliou, F.N. (2008). MicroRNA-155 Is a Negative Regulator of Activation-Induced Cytidine Deaminase. Immunity *28*, 621–629.

Tijchon, E., Emst, L. van, Yuniati, L., Schenau, D. van I., Havinga, J., Rouault, J.-P., Hoogerbrugge, P.M., Leeuwen, F.N. van, and Scheijen, B. (2016). Tumor suppressors BTG1 and BTG2 regulate early mouse B-cell development. Haematologica *101*, e272–e276.

Tom Mitchell. Machine Learning (McGraw Hill, 1997).

Tran, T.H., Nakata, M., Suzuki, K., Begum, N.A., Shinkura, R., Fagarasan, S., Honjo, T., and Nagaoka, H. (2010). B cell-specific and stimulation-responsive enhancers derepress Aicda by overcoming the effects of silencers. Nat Immunol *11*, 148–154.

Trevor Hastie, Robert Tibshirani, Jerome Friedman. The Elements of Statistical Learning (Springer, 2009).

Tsubata, T., and Reth, M. (1990). The products of pre-B cell-specific genes (lambda 5 and VpreB) and the immunoglobulin mu chain form a complex that is transported onto the cell surface. Journal of Experimental Medicine *172*, 973–976.

Uchimura, Y., Barton, L.F., Rada, C., and Neuberger, M.S. (2011). REG-γ associates with and modulates the abundance of nuclear activation-induced deaminase. Journal of Experimental Medicine *208*, 2385–2391.

Vettermann, C., and Schlissel, M.S. (2010). Allelic exclusion of immunoglobulin genes: models and mechanisms. Immunol Rev *237*, 22–42.

Victora, G.D., and Nussenzweig, M.C. (2012). Germinal Centers. Annual Review of Immunology *30*, 429–457.

Wallace, S.S., Murphy, D.L., and Sweasy, J.B. (2012). Base Excision Repair and Cancer. Cancer Lett *327*, 73–89.

Wang, Q., Kieffer-Kwon, K.-R., Oliveira, T.Y., Mayer, C.T., Yao, K., Pai, J., Cao, Z., Dose, M., Casellas, R., Jankovic, M., et al. (2016). The cell cycle restricts activation-induced cytidine deaminase activity to early G1. J. Exp. Med. jem.20161649.

Wang, Q., Oliveira, T., Jankovic, M., Silva, I.T., Hakim, O., Yao, K., Gazumyan, A., Mayer, C.T., Pavri, R., Casellas, R., et al. (2014). Epigenetic targeting of activation-induced cytidine deaminase. PNAS *111*, 18667–18672.

Weill, J.-C., and Reynaud, C.-A. (1996). Rearrangement/hypermutation/gene conversion: when, where and why? Immunology Today *17*, 92–97.

Weill, J.-C., and Reynaud, C.-A. (2008). DNA polymerases in adaptive immunity. Nature Reviews Immunology *8*, 302–312.

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R.A. (2013). Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes. Cell *153*, 307–319.

Wiesendanger, M., Kneitz, B., Edelmann, W., and Scharff, M.D. (2000). Somatic Hypermutation in Muts Homologue (Msh)3-, Msh6-, and Msh3/Msh6-Deficient Mice Reveals a Role for the Msh2–Msh6 Heterodimer in Modulating the Base Substitution Pattern. Journal of Experimental Medicine *191*, 579–584.

Wilson, T.M., Vaisman, A., Martomo, S.A., Sullivan, P., Lan, L., Hanaoka, F., Yasui, A., Woodgate, R., and Gearhart, P.J. (2005). MSH2–MSH6 stimulates DNA polymerase η, suggesting a role for A:T mutations in antibody genes. J Exp Med *201*, 637–645.

Wu, X., and Stavnezer, J. (2007). DNA polymerase β is able to repair breaks in switch regions and plays an inhibitory role during immunoglobulin class switch recombination. Journal of Experimental Medicine *204*, 1677–1689.

Xu, Z., Fulop, Z., Wu, G., Pone, E.J., Zhang, J., Mai, T., Thomas, L.M., Al-Qahtani, A., White, C.A., Park, S.-R., et al. (2010). 14-3-3 adaptor proteins recruit AID to 5′-AGCT-3′-rich switch regions for class switch recombination. Nat Struct Mol Biol *17*, 1124–1135.

Xu, Z., Zan, H., Pone, E.J., Mai, T., and Casali, P. (2012). Immunoglobulin class-switch DNA recombination: induction, targeting and beyond. Nature Reviews Immunology *12*, 517.

Xue, K., Rada, C., and Neuberger, M.S. (2006). The in vivo pattern of AID targeting to immunoglobulin switch regions deduced from mutation spectra in msh2−/− ung−/− mice. J Exp Med *203*, 2085–2094.

Yamane, A., Resch, W., Kuo, N., Kuchen, S., Li, Z., Sun, H., Robbiani, D.F., McBride, K., Nussenzweig, M.C., and Casellas, R. (2011). Deep-sequencing identification of the genomic targets of the cytidine deaminase AID and its cofactor RPA in B lymphocytes. Nat Immunol *12*, 62–69.

Yeap, L.-S., Hwang, J.K., Du, Z., Meyers, R.M., Meng, F.-L., Jakubauskaitė, A., Liu, M., Mani, V., Neuberg, D., Kepler, T.B., et al. (2015). Sequence-Intrinsic Mechanisms that Target AID Mutational Outcomes on Antibody Genes. Cell *163*, 1124–1137.

Yébenes, V.G. de, Belver, L., Pisano, D.G., González, S., Villasante, A., Croce, C., He, L., and Ramiro, A.R. (2008). miR-181b negatively regulates activation-induced cytidine deaminase in B cells. Journal of Experimental Medicine *205*, 2199–2206.

Yoshikawa, K., Okazaki, I., Eto, T., Kinoshita, K., Muramatsu, M., Nagaoka, H., and Honjo, T. (2002). AID Enzyme-Induced Hypermutation in an Actively Transcribed Gene in Fibroblasts. Science *296*, 2033–2036.

Yu, K., Chedin, F., Hsieh, C.-L., Wilson, T.E., and Lieber, M.R. (2003). R-loops at immunoglobulin class switch regions in the chromosomes of stimulated B cells. Nature Immunology *4*, 442.

Zahn, A., Daugan, M., Safavi, S., Godin, D., Cheong, C., Lamarre, A., and Noia, J.M.D. (2013). Separation of Function between Isotype Switching and Affinity Maturation In Vivo during Acute Immune Responses and Circulating Autoantibodies in UNG-Deficient Mice. The Journal of Immunology *190*, 5949–5960.

Zarrin, A.A., Alt, F.W., Chaudhuri, J., Stokes, N., Kaushal, D., Pasquier, L.D., and Tian, M. (2004). An evolutionarily conserved target motif for immunoglobulin class-switch recombination. Nature Immunology *5*, 1275.

Zhang, J., Grubor, V., Love, C.L., Banerjee, A., Richards, K.L., Mieczkowski, P.A., Dunphy, C., Choi, W., Au, W.Y., Srivastava, G., et al. (2013). Genetic heterogeneity of diffuse large B-cell lymphoma. PNAS *110*, 1398–1403.

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nusbaum, C., Myers, R.M., Brown, M., Li, W., et al. (2008). Model-based Analysis of ChIP-Seq (MACS). Genome Biol. *9*, R137.

Zhang, Z.Z., Pannunzio, N.R., Han, L., Hsieh, C.-L., Yu, K., and Lieber, M.R. (2014). The Strength of an Ig Switch Region Is Determined by Its Ability to Drive R Loop Formation and Its Number of WGCW Sites. Cell Reports *8*, 557–569.

Zheng, S., Vuong, B.Q., Vaidyanathan, B., Lin, J.-Y., Huang, F.-T., and Chaudhuri, J. (2015). Non-coding RNA Generated following Lariat Debranching Mediates Targeting of AID to DNA. Cell *161*, 762–773.

# ANNEX

**Annex I.** Genes included in SureSelect capture library

| Gene | Origin | Status | Gene | Origin | Status |
|---|---|---|---|---|---|
| *0610007P08Rik* | A | Non-mutated | *Ebf1* | A | Mutated |
| *Ada* | A | Mutated | *Eif4a2* | A | Mutated |
| *Agxt2l2* | A | Non-mutated | *Erh* | A | Mutated |
| *Aicda* | A | Mutated | *Ets1* | A | Mutated |
| *Alk* | A | Non-mutated | *Ezh2* | A | Non-mutated |
| *Anxa2* | A | Non-mutated | *Fas* | A | Mutated |
| *Atm* | A | Non-mutated | *Fcgr2b* | A | Non-mutated |
| *B2m* | A | Mutated | *Fchsd2* | A | Mutated |
| *Bcar3* | A | Non-mutated | *Fen1* | A | Mutated |
| *Bcl11a* | A | Mutated | *Fgfr3* | A | Non-mutated |
| *Bcl2* | A | Non-mutated | *Fh1* | A | Non-mutated |
| *Bcl2l11* | A | Non-mutated | *Fli1* | A | Mutated |
| *Bcl6* | A | Mutated | *Fnbp1* | A | Mutated |
| *Birc5* | A | Non-mutated | *Fosb* | A | Non-mutated |
| *Blimp1* | A | Non-mutated | *Gadd45b* | A | Mutated |
| *Blk* | A | Mutated | *Gas5* | A | Mutated |
| *Blnk* | A | Non-mutated | *Grap* | A | Mutated |
| *Brg1* | A | Non-mutated | *H2afx* | A | Mutated |
| *Btg1* | A | Mutated | *H3f3b* | A | Mutated |
| *Btg2* | A | Mutated | *Hdac1* | A | Mutated |
| *Canx* | A | Non-mutated | *Hdgf* | A | Mutated |
| *Card11* | A | Non-mutated | *Hivep3* | A | Non-mutated |
| *Casp3* | A | Non-mutated | *Hnrnpa2b1* | A | Mutated |
| *Cbfb* | A | Non-mutated | *Hnrnpf* | A | Mutated |
| *Ccnb2* | A | Non-mutated | *Id3* | A | Mutated |
| *Ccnd1* | A | Non-mutated | *Igf1* | A | Non-mutated |
| *Ccnd2* | A | Non-mutated | *Igj* | A | Mutated |
| *Ccnd3* | A | Non-mutated | *Igsf8* | A | Non-mutated |
| *Cd19* | A | Mutated | *Ikzf1* | A | Mutated |
| *Cd22* | A | Mutated | *Il21r* | A | Mutated |
| *Cd37* | A | Mutated | *Il4ra* | A | Mutated |
| *Cd40* | A | Non-mutated | *Irf4* | A | Mutated |
| *Cd40* | A | Non-mutated | *Jak2* | A | Non-mutated |
| *Cd53* | A | Mutated | *Kpna2* | A | Mutated |
| *Cd79a* | A | Mutated | *Kras* | A | Non-mutated |
| *Cd79b* | A | Mutated | *Ksr1* | A | Non-mutated |
| *Cd82* | A | Non-mutated | *Lbr* | A | Non-mutated |
| *Cd83* | A | Mutated | *Lcp1* | A | Non-mutated |
| *Cdc20* | A | Non-mutated | *Lmo2* | A | Non-mutated |
| *Cdc37* | A | Non-mutated | *Lmo4* | A | Non-mutated |
| *Cenpa* | A | Non-mutated | *Lrmp* | A | Mutated |
| *Chd2* | A | Mutated | *Ltb* | A | Mutated |
| *Ciita* | A | Mutated | *Lyl1* | A | Non-mutated |
| *Cks2* | A | Non-mutated | *Lyn* | A | Mutated |
| *Crip1* | A | Non-mutated | *Mafb* | A | Non-mutated |
| *Ctsh* | A | Non-mutated | *Man1a* | A | Mutated |
| *Cyth1* | A | Mutated | *Mbd4* | A | Non-mutated |
| *Dad1* | A | Non-mutated | *Mef2b* | A | Mutated |
| *Dock10* | A | Non-mutated | *Mef2c* | A | Non-mutated |
| *Ebf1* | A | Mutated | *Mid1* | A | Non-mutated |

| Gene | Origin | Status |
|---|---|---|
| *Mir142* | A | Mutated |
| *Mll1* | A | Non-mutated |
| *Ms4a1* | A | Mutated |
| *Msh6* | A | Mutated |
| *Mybl2* | A | Non-mutated |
| *Myc* | A | Mutated |
| *Mycn* | A | Non-mutated |
| *Ncl* | A | Mutated |
| *Nedd8* | A | Non-mutated |
| *Npm1* | A | Non-mutated |
| *Nr5a1* | A | Non-mutated |
| *Nras* | A | Non-mutated |
| *Odc1* | A | Non-mutated |
| *Pax5* | A | Mutated |
| *Pcna* | A | Mutated |
| *Pebp4* | A | Non-mutated |
| *Pgk1* | A | Non-mutated |
| *Phip* | A | Mutated |
| *Pik3ap1* | A | Mutated |
| *Pim1* | A | Mutated |
| *Pold4* | A | Mutated |
| *Polm* | A | Non-mutated |
| *Pou2af1* | A | Mutated |
| *Ppp1r16b* | A | Mutated |
| *Prdx1* | A | Mutated |
| *Prkcd* | A | Non-mutated |
| *Ptma* | A | Mutated |
| *Ptprc* | A | Mutated |
| *Rac2* | A | Non-mutated |
| *Rad51* | A | Mutated |
| *Raf1* | A | Non-mutated |
| *Ran* | A | Non-mutated |
| *Ranbp1* | A | Mutated |
| *Rapgef1* | A | Non-mutated |
| *Rasa3* | A | Non-mutated |
| *Rassf2* | A | Non-mutated |
| *Rbm39* | A | Mutated |
| *Rel* | A | Mutated |
| *Rela* | A | Non-mutated |
| *Rev3l* | A | Mutated |
| *Rfc2* | A | Mutated |
| *Rhoh* | A | Mutated |
| *Rpia* | A | Mutated |
| *Rpl3* | A | Mutated |
| *Rps12* | A | Mutated |
| *Sema4d* | A | Non-mutated |
| *Set* | A | Non-mutated |
| *Sfi1* | A | Non-mutated |
| *Sh3bp5* | A | Mutated |
| *Slbp* | A | Mutated |
| *Snx29* | A | Non-mutated |
| *Snx5* | A | Mutated |
| *Sorcs2* | A | Non-mutated |
| *Sp110* | A | Non-mutated |
| *Spib* | A | Mutated |
| *Srsf10* | A | Mutated |
| *St6gal1* | A | Mutated |
| *Stmn1* | A | Non-mutated |
| *Syk* | A | Mutated |
| *Tapbp* | A | Mutated |
| *Tcea1* | A | Mutated |
| *Tcl1* | A | Non-mutated |
| *Tex14* | A | Mutated |
| *Tff1* | A | Non-mutated |
| *Tkt* | A | Non-mutated |
| *Top1* | A | Mutated |
| *Tpt1* | A | Non-mutated |
| *Trp53* | A | Non-mutated |
| *Txn1* | A | Non-mutated |
| *Tyms* | A | Non-mutated |
| *Ubac2* | A | Mutated |
| *Ung* | A | Non-mutated |
| *Usf1* | A | Non-mutated |
| *Yes1* | A | Non-mutated |
| *Akap13* | B | Non-mutated |
| *Apc* | B | Non-mutated |
| *Axl* | B | Non-mutated |
| *Csf1r* | B | Non-mutated |
| *Ctnnb1* | B | Non-mutated |
| *Ddost* | B | Non-mutated |
| *Fes* | B | Non-mutated |
| *Fgfr2* | B | Non-mutated |
| *Gip* | B | Non-mutated |
| *Gli2* | B | Non-mutated |
| *Gnas* | B | Non-mutated |
| *Irs2* | B | Non-mutated |
| *Lmo1* | B | Mutated |
| *Mas1* | B | Non-mutated |
| *Mcf2* | B | Non-mutated |
| *Mos* | B | Non-mutated |
| *Mre11a* | B | Non-mutated |
| *Mycl1* | B | Non-mutated |
| *Nfkb2* | B | Mutated |
| *Ntrk1* | B | Non-mutated |
| *Prkcz* | B | Non-mutated |
| *Runx1* | B | Non-mutated |
| *Sis* | B | Non-mutated |
| *Tal1* | B | Non-mutated |
| *Tal2* | B | Non-mutated |
| *Tiam1* | B | Non-mutated |

| Gene | Origin | Status |
|---|---|---|
| *Tlx1* | B | Non-mutated |
| *Tnfaip3* | B | Mutated |
| *Tsc2* | B | Non-mutated |
| *1810030J14Rik* | C | Non-mutated |
| *1810065E05Rik* | C | Mutated |
| *2010107E04Rik* | C | Non-mutated |
| *2200002D01Rik* | C | Non-mutated |
| *2210415F13Rik* | C | Non-mutated |
| *Aanat* | C | Non-mutated |
| *Aatf* | C | Non-mutated |
| *Aatk* | C | Mutated |
| *Abca1* | C | Non-mutated |
| *Abi1* | C | Non-mutated |
| *Abi2* | C | Non-mutated |
| *Abl1* | C | Non-mutated |
| *Abl1* | C | Non-mutated |
| *Abl2* | C | Mutated |
| *Ablim1* | C | Mutated |
| *Ablim2* | C | Non-mutated |
| *Abr* | C | Non-mutated |
| *Abt1* | C | Non-mutated |
| *Acaca* | C | Non-mutated |
| *Acacb* | C | Non-mutated |
| *Acer1* | C | Non-mutated |
| *Acnat1* | C | Non-mutated |
| *Acot1* | C | Non-mutated |
| *Acot3* | C | Non-mutated |
| *Acot7* | C | Mutated |
| *Acr* | C | Non-mutated |
| *Acsl6* | C | Non-mutated |
| *Acsm1* | C | Non-mutated |
| *Acss1* | C | Non-mutated |
| *Actb* | C | Mutated |
| *Acvr1* | C | Non-mutated |
| *Adamts20* | C | Non-mutated |
| *Adamts9* | C | Non-mutated |
| *Adamtsl4* | C | Non-mutated |
| *Adar* | C | Mutated |
| *Adc* | C | Non-mutated |
| *Adcy9* | C | Non-mutated |
| *Adcyap1* | C | Non-mutated |
| *Adh1* | C | Non-mutated |
| *Adipoq* | C | Non-mutated |
| *Adora2b* | C | Non-mutated |
| *Adss* | C | Non-mutated |
| *Aebp1* | C | Non-mutated |
| *Aes* | C | Non-mutated |
| *Aff2* | C | Non-mutated |
| *Agk* | C | Mutated |
| *Agr2* | C | Non-mutated |

| Gene | Origin | Status |
|---|---|---|
| *Agtpbp1* | C | Non-mutated |
| *Aifm1* | C | Non-mutated |
| *Aire* | C | Non-mutated |
| *Akap2* | C | Non-mutated |
| *Akap8* | C | Mutated |
| *Akt1* | C | Non-mutated |
| *Akt2* | C | Non-mutated |
| *Alcam* | C | Non-mutated |
| *Alpk3* | C | Non-mutated |
| *Als2cl* | C | Non-mutated |
| *Amd1* | C | Non-mutated |
| *Ank3* | C | Non-mutated |
| *Anxa5* | C | Non-mutated |
| *Anxa6* | C | Non-mutated |
| *Anxa7* | C | Non-mutated |
| *Aoah* | C | Non-mutated |
| *Apex1* | C | Mutated |
| *Apip* | C | Non-mutated |
| *Apoa1* | C | Non-mutated |
| *Apoa4* | C | Non-mutated |
| *Apobec1* | C | Mutated |
| *Apobec2* | C | Non-mutated |
| *Apoc3* | C | Non-mutated |
| *Apoe* | C | Mutated |
| *Aptx* | C | Non-mutated |
| *Aqp2* | C | Non-mutated |
| *Aqp8* | C | Non-mutated |
| *Arhgap31* | C | Non-mutated |
| *Arl4a* | C | Non-mutated |
| *Arpp21* | C | Non-mutated |
| *Arrb1* | C | Non-mutated |
| *Art2a-ps* | C | Non-mutated |
| *Ascl3* | C | Non-mutated |
| *Asf1b* | C | Non-mutated |
| *Aspn* | C | Non-mutated |
| *Atf5* | C | Mutated |
| *Atg13* | C | Non-mutated |
| *Atp12a* | C | Non-mutated |
| *Atp4a* | C | Non-mutated |
| *Atp5b* | C | Mutated |
| *Atp5e* | C | Mutated |
| *Atp5j2* | C | Non-mutated |
| *Atp5o* | C | Mutated |
| *Atpif1* | C | Non-mutated |
| *Atr* | C | Non-mutated |
| *Atrnl1* | C | Non-mutated |
| *Atrx* | C | Non-mutated |
| *Atxn1* | C | Non-mutated |
| *Atxn2* | C | Non-mutated |
| *Aurka* | C | Non-mutated |

| Gene | Origin | Status |
|---|---|---|
| *Aurkb* | C | Non-mutated |
| *Avil* | C | Non-mutated |
| *Avp* | C | Non-mutated |
| *Azi2* | C | Non-mutated |
| *B4galt1* | C | Non-mutated |
| *Bad* | C | Mutated |
| *Bak1* | C | Non-mutated |
| *Barx2* | C | Non-mutated |
| *Bax* | C | Non-mutated |
| *Baz2a* | C | Non-mutated |
| *Bbs4* | C | Non-mutated |
| *Bcam* | C | Non-mutated |
| *Bcap31* | C | Non-mutated |
| *Bcat1* | C | Non-mutated |
| *Bcl11b* | C | Non-mutated |
| *Bcl3* | C | Non-mutated |
| *Bcl6b* | C | Non-mutated |
| *Bco2* | C | Non-mutated |
| *Bcr* | C | Non-mutated |
| *Best2* | C | Non-mutated |
| *Bgn* | C | Non-mutated |
| *Bid* | C | Mutated |
| *Bmf* | C | Non-mutated |
| *Bmi1* | C | Non-mutated |
| *Bmp2k* | C | Mutated |
| *Bnip3* | C | Non-mutated |
| *Bop1* | C | Non-mutated |
| *Brca1* | C | Mutated |
| *Brca2* | C | Non-mutated |
| *Brf1* | C | Non-mutated |
| *Brms1* | C | Non-mutated |
| *Brsk1* | C | Non-mutated |
| *Btg4* | C | Non-mutated |
| *Btrc* | C | Non-mutated |
| *Bub1* | C | Non-mutated |
| *Bub1b* | C | Non-mutated |
| *Bub3* | C | Non-mutated |
| *C1d* | C | Non-mutated |
| *C2cd3* | C | Mutated |
| *C3* | C | Non-mutated |
| *Cabp4* | C | Non-mutated |
| *Cacna1a* | C | Non-mutated |
| *Cacna1e* | C | Non-mutated |
| *Cacna1s* | C | Non-mutated |
| *Cacnb2* | C | Non-mutated |
| *Cacnb3* | C | Non-mutated |
| *Cacng4* | C | Mutated |
| *Cadm1* | C | Non-mutated |
| *Cadps2* | C | Non-mutated |
| *Calca* | C | Non-mutated |
| *Calm1* | C | Mutated |
| *Camk1d* | C | Non-mutated |
| *Capn2* | C | Non-mutated |
| *Capn3* | C | Non-mutated |
| *Car1* | C | Non-mutated |
| *Car2* | C | Non-mutated |
| *Car4* | C | Non-mutated |
| *Carf* | C | Non-mutated |
| *Casc3* | C | Non-mutated |
| *Casp1* | C | Non-mutated |
| *Casr* | C | Non-mutated |
| *Cat* | C | Mutated |
| *Cav2* | C | Non-mutated |
| *Cbll1* | C | Non-mutated |
| *Cbr2* | C | Non-mutated |
| *Cbx1* | C | Non-mutated |
| *Cby1* | C | Non-mutated |
| *Ccdc56* | C | Non-mutated |
| *Cckar* | C | Non-mutated |
| *Ccl1* | C | Non-mutated |
| *Ccl28* | C | Non-mutated |
| *Ccnb1ip1* | C | Non-mutated |
| *Ccne1* | C | Non-mutated |
| *Ccne2* | C | Mutated |
| *Ccr1* | C | Non-mutated |
| *Ccr7* | C | Non-mutated |
| *Ccrn4l* | C | Non-mutated |
| *Cd1d1* | C | Non-mutated |
| *Cd1d2* | C | Non-mutated |
| *Cd209b* | C | Non-mutated |
| *Cd24a* | C | Mutated |
| *Cd27* | C | Non-mutated |
| *Cd28* | C | Non-mutated |
| *Cd300lf* | C | Non-mutated |
| *Cd36* | C | Non-mutated |
| *Cd3e* | C | Non-mutated |
| *Cd40lg* | C | Non-mutated |
| *Cd48* | C | Mutated |
| *Cd5* | C | Non-mutated |
| *Cd74* | C | Mutated |
| *Cd9* | C | Non-mutated |
| *Cdc42bpa* | C | Non-mutated |
| *Cdc42ep1* | C | Non-mutated |
| *Cdh1* | C | Non-mutated |
| *Cdh4* | C | Non-mutated |
| *Cdk1* | C | Non-mutated |
| *Cdk11b* | C | Mutated |
| *Cdk4* | C | Mutated |
| *Cdk5* | C | Non-mutated |
| *Cdkn1a* | C | Non-mutated |

| Gene | Origin | Status |
|---|---|---|
| *Cdkn1b* | C | Non-mutated |
| *Cdkn2a* | C | Non-mutated |
| *Cdkn2b* | C | Non-mutated |
| *Cdkn2c* | C | Non-mutated |
| *Cdt1* | C | Non-mutated |
| *Cebpa* | C | Non-mutated |
| *Cebpb* | C | Non-mutated |
| *Cebpd* | C | Non-mutated |
| *Cebpe* | C | Non-mutated |
| *Cebpg* | C | Non-mutated |
| *Cela1* | C | Non-mutated |
| *Cela2a* | C | Non-mutated |
| *Celf1* | C | Non-mutated |
| *Celf2* | C | Non-mutated |
| *Celf3* | C | Non-mutated |
| *Celf4* | C | Non-mutated |
| *Cenpf* | C | Non-mutated |
| *Cenpk* | C | Non-mutated |
| *Cetn1* | C | Non-mutated |
| *Chd1* | C | Non-mutated |
| *Chd7* | C | Non-mutated |
| *Chek1* | C | Mutated |
| *Chek2* | C | Non-mutated |
| *Chrna1* | C | Non-mutated |
| *Chrna6* | C | Non-mutated |
| *Chrnb2* | C | Non-mutated |
| *Cirbp* | C | Non-mutated |
| *Cish* | C | Non-mutated |
| *Ckmt1* | C | Non-mutated |
| *Clca1* | C | Non-mutated |
| *Clcn1* | C | Non-mutated |
| *Clcn2* | C | Non-mutated |
| *Clcnka* | C | Non-mutated |
| *Cldn1* | C | Non-mutated |
| *Clk1* | C | Mutated |
| *Cln6* | C | Non-mutated |
| *Cnbp* | C | Mutated |
| *Cnga2* | C | Non-mutated |
| *Col1a1* | C | Non-mutated |
| *Col5a3* | C | Non-mutated |
| *Cox10* | C | Non-mutated |
| *Cox4i1* | C | Mutated |
| *Cox6a1* | C | Mutated |
| *Cox6b1* | C | Non-mutated |
| *Cox6c* | C | Non-mutated |
| *Cox8a* | C | Mutated |
| *Cpa1* | C | Non-mutated |
| *Cpeb1* | C | Non-mutated |
| *Cradd* | C | Mutated |
| *Crcp* | C | Non-mutated |

| Gene | Origin | Status |
|---|---|---|
| *Creb1* | C | Non-mutated |
| *Crebbp* | C | Non-mutated |
| *Crem* | C | Non-mutated |
| *Crip3* | C | Non-mutated |
| *Crlf1* | C | Non-mutated |
| *Cry2* | C | Non-mutated |
| *Csf1* | C | Non-mutated |
| *Csf2* | C | Non-mutated |
| *Csk* | C | Mutated |
| *Csnk1d* | C | Mutated |
| *Csrnp2* | C | Non-mutated |
| *Ctbp2* | C | Non-mutated |
| *Ctcf* | C | Non-mutated |
| *Ctcfl* | C | Non-mutated |
| *Ctgf* | C | Non-mutated |
| *Ctnna3* | C | Non-mutated |
| *Ctnnd1* | C | Non-mutated |
| *Ctsd* | C | Non-mutated |
| *Ctse* | C | Non-mutated |
| *Ctsg* | C | Non-mutated |
| *Ctss* | C | Non-mutated |
| *Cxadr* | C | Non-mutated |
| *Cybrd1* | C | Non-mutated |
| *Cycs* | C | Non-mutated |
| *Cyfip1* | C | Non-mutated |
| *Cyp11b1* | C | Non-mutated |
| *Cyp11b2* | C | Non-mutated |
| *Cyp1a1* | C | Non-mutated |
| *Cyp2c55* | C | Non-mutated |
| *Cyp39a1* | C | Non-mutated |
| *Cyth3* | C | Non-mutated |
| *Dand5* | C | Non-mutated |
| *Dao* | C | Non-mutated |
| *Dap3* | C | Non-mutated |
| *Daxx* | C | Mutated |
| *Dbh* | C | Non-mutated |
| *Dbi* | C | Non-mutated |
| *Dcn* | C | Non-mutated |
| *Dctn3* | C | Non-mutated |
| *Ddb2* | C | Mutated |
| *Ddit3* | C | Non-mutated |
| *Ddr1* | C | Non-mutated |
| *Ddx20* | C | Mutated |
| *Ddx5* | C | Mutated |
| *Deaf1* | C | Non-mutated |
| *Dedd* | C | Non-mutated |
| *Defb8* | C | Non-mutated |
| *Dffa* | C | Non-mutated |
| *Dgat2* | C | Non-mutated |
| *Dgki* | C | Non-mutated |

| Gene | Origin | Status |
|---|---|---|
| *Dgkz* | C | Non-mutated |
| *Dguok* | C | Non-mutated |
| *Dhrs4* | C | Non-mutated |
| *Dicer1* | C | Non-mutated |
| *Dlc1* | C | Non-mutated |
| *Dld* | C | Non-mutated |
| *Dmap1* | C | Non-mutated |
| *Dmc1* | C | Non-mutated |
| *Dnaja1* | C | Non-mutated |
| *Dnajb11* | C | Non-mutated |
| *Dnd1* | C | Non-mutated |
| *Dnmt1* | C | Mutated |
| *Dnmt3a* | C | Non-mutated |
| *Dnmt3l* | C | Non-mutated |
| *Dntt* | C | Non-mutated |
| *Doc2a* | C | Non-mutated |
| *Dock7* | C | Non-mutated |
| *Dok1* | C | Mutated |
| *Dpf1* | C | Non-mutated |
| *Drd5* | C | Non-mutated |
| *Dsg4* | C | Mutated |
| *Dsp* | C | Non-mutated |
| *Dtx2* | C | Non-mutated |
| *Dub1* | C | Non-mutated |
| *Dub1a* | C | Non-mutated |
| *Duox2* | C | Non-mutated |
| *Dusp22* | C | Non-mutated |
| *Dusp6* | C | Mutated |
| *Dyrk3* | C | Mutated |
| *E2f1* | C | Mutated |
| *E2f2* | C | Mutated |
| *E2f4* | C | Non-mutated |
| *E2f7* | C | Non-mutated |
| *E4f1* | C | Mutated |
| *Ear2* | C | Non-mutated |
| *Ebf3* | C | Non-mutated |
| *Ebf4* | C | Non-mutated |
| *Ect2* | C | Non-mutated |
| *Eda* | C | Non-mutated |
| *Edil3* | C | Non-mutated |
| *Eef1a1* | C | Mutated |
| *Egf* | C | Non-mutated |
| *Egfl7* | C | Non-mutated |
| *Egfr* | C | Non-mutated |
| *Egr2* | C | Non-mutated |
| *Ehhadh* | C | Non-mutated |
| *Ehmt2* | C | Non-mutated |
| *Eif2a* | C | Mutated |
| *Eif2ak4* | C | Non-mutated |
| *Eif3a* | C | Mutated |
| *Eif3b* | C | Non-mutated |
| *Eif3d* | C | Mutated |
| *Eif5a* | C | Mutated |
| *Elane* | C | Non-mutated |
| *Elf1* | C | Non-mutated |
| *Elf4* | C | Non-mutated |
| *Elf5* | C | Non-mutated |
| *Elk1* | C | Non-mutated |
| *Ell* | C | Mutated |
| *Eln* | C | Non-mutated |
| *Elp2* | C | Non-mutated |
| *Enox1* | C | Non-mutated |
| *Enox2* | C | Non-mutated |
| *Epb4.1l5* | C | Non-mutated |
| *Epcam* | C | Non-mutated |
| *Epgn* | C | Non-mutated |
| *Ephb2* | C | Non-mutated |
| *Ephx1* | C | Non-mutated |
| *Erap1* | C | Non-mutated |
| *Erbb2* | C | Non-mutated |
| *Erbb3* | C | Non-mutated |
| *Ercc1* | C | Non-mutated |
| *Ercc2* | C | Non-mutated |
| *Ercc4* | C | Non-mutated |
| *Ercc5* | C | Non-mutated |
| *Ereg* | C | Non-mutated |
| *Eri1* | C | Mutated |
| *Esam* | C | Non-mutated |
| *Esrra* | C | Non-mutated |
| *Esrrb* | C | Non-mutated |
| *Esrrg* | C | Non-mutated |
| *Etnk2* | C | Non-mutated |
| *Etv4* | C | Non-mutated |
| *Etv6* | C | Non-mutated |
| *Exo1* | C | Non-mutated |
| *Ezr* | C | Non-mutated |
| *Fabp2* | C | Non-mutated |
| *Fabp5* | C | Non-mutated |
| *Fabp7* | C | Non-mutated |
| *Faf1* | C | Non-mutated |
| *Faim* | C | Non-mutated |
| *Fancc* | C | Non-mutated |
| *Fancd2* | C | Non-mutated |
| *Fancg* | C | Non-mutated |
| *Far1* | C | Non-mutated |
| *Fasl* | C | Non-mutated |
| *Fau* | C | Non-mutated |
| *Fbl* | C | Non-mutated |
| *Fbxl12* | C | Non-mutated |
| *Fbxo11* | C | Non-mutated |

| Gene | Origin | Status |
|---|---|---|
| *Fbxo5* | C | Non-mutated |
| *Fbxo8* | C | Non-mutated |
| *Fcgr1* | C | Non-mutated |
| *Fcgr3* | C | Non-mutated |
| *Ferd3l* | C | Non-mutated |
| *Ffar1* | C | Non-mutated |
| *Fgd4* | C | Non-mutated |
| *Fgf1* | C | Non-mutated |
| *Fgf2* | C | Non-mutated |
| *Fgf3* | C | Non-mutated |
| *Fgf4* | C | Non-mutated |
| *Fgf6* | C | Non-mutated |
| *Fgf7* | C | Non-mutated |
| *Fgfr1* | C | Non-mutated |
| *Fgfr1* | C | Non-mutated |
| *Figf* | C | Non-mutated |
| *Flt4* | C | Non-mutated |
| *Fntb* | C | Non-mutated |
| *Fos* | C | Non-mutated |
| *Fosl1* | C | Non-mutated |
| *Fosl2* | C | Non-mutated |
| *Foxa1* | C | Non-mutated |
| *Foxc1* | C | Non-mutated |
| *Foxc2* | C | Non-mutated |
| *Foxd1* | C | Non-mutated |
| *Foxe3* | C | Non-mutated |
| *Foxf2* | C | Non-mutated |
| *Foxg1* | C | Non-mutated |
| *Foxj1* | C | Non-mutated |
| *Foxl2* | C | Non-mutated |
| *Foxo3* | C | Non-mutated |
| *Foxo4* | C | Mutated |
| *Foxp1* | C | Non-mutated |
| *Foxp3* | C | Non-mutated |
| *Fpr1* | C | Non-mutated |
| *Frat1* | C | Non-mutated |
| *Frem1* | C | Non-mutated |
| *Fscn2* | C | Non-mutated |
| *Fth1* | C | Mutated |
| *Ftl1* | C | Mutated |
| *Fyb* | C | Non-mutated |
| *Fyn* | C | Non-mutated |
| *Gab3* | C | Non-mutated |
| *Gabrg1* | C | Non-mutated |
| *Gadd45a* | C | Non-mutated |
| *Gadd45g* | C | Mutated |
| *Gale* | C | Non-mutated |
| *Galnt1* | C | Mutated |
| *Gamt* | C | Non-mutated |
| *Gap43* | C | Non-mutated |

| Gene | Origin | Status |
|---|---|---|
| *Gas1* | C | Non-mutated |
| *Gas7* | C | Non-mutated |
| *Gata2* | C | Non-mutated |
| *Gata3* | C | Non-mutated |
| *Gata4* | C | Non-mutated |
| *Gata6* | C | Non-mutated |
| *Gck* | C | Non-mutated |
| *Gdap1* | C | Non-mutated |
| *Gdf5* | C | Non-mutated |
| *Gdf7* | C | Non-mutated |
| *Gdi2* | C | Mutated |
| *Ghr* | C | Non-mutated |
| *Ghrhr* | C | Non-mutated |
| *Gigyf1* | C | Non-mutated |
| *Gins1* | C | Non-mutated |
| *Gja1* | C | Non-mutated |
| *Gja10* | C | Non-mutated |
| *Gjc2* | C | Non-mutated |
| *Gkap1* | C | Non-mutated |
| *Glis1* | C | Non-mutated |
| *Gls* | C | Non-mutated |
| *Gm1821* | C | Non-mutated |
| *Gm5409* | C | Non-mutated |
| *Gna12* | C | Non-mutated |
| *Gna13* | C | Mutated |
| *Gnao1* | C | Non-mutated |
| *Gng2* | C | Non-mutated |
| *Gnpat* | C | Non-mutated |
| *Got1* | C | Non-mutated |
| *Gpat2* | C | Non-mutated |
| *Gpd1* | C | Non-mutated |
| *Gphn* | C | Non-mutated |
| *Gpnmb* | C | Non-mutated |
| *Gpr12* | C | Non-mutated |
| *Gpx1* | C | Non-mutated |
| *Gpx2* | C | Non-mutated |
| *Grik2* | C | Non-mutated |
| *Grip1* | C | Non-mutated |
| *Grlf1* | C | Non-mutated |
| *Grn* | C | Non-mutated |
| *Grsf1* | C | Non-mutated |
| *Gsdma3* | C | Non-mutated |
| *Gsdmd* | C | Non-mutated |
| *Gsk3b* | C | Non-mutated |
| *Gss* | C | Non-mutated |
| *Gsta3* | C | Mutated |
| *Gstk1* | C | Non-mutated |
| *Gstp1* | C | Non-mutated |
| *Guca2a* | C | Non-mutated |
| *Gucy1a2* | C | Non-mutated |

| Gene | Origin | Status |
|---|---|---|
| *Gucy1a3* | C | Non-mutated |
| *Gulo* | C | Non-mutated |
| *Gyk* | C | Non-mutated |
| *Gykl1* | C | Non-mutated |
| *Gypa* | C | Non-mutated |
| *Gzmb* | C | Non-mutated |
| *H2-Aa* | C | Non-mutated |
| *H2-Bl* | C | Non-mutated |
| *H47* | C | Mutated |
| *H60c* | C | Non-mutated |
| *Hba-a1* | C | Non-mutated |
| *Hck* | C | Non-mutated |
| *Hdac2* | C | Non-mutated |
| *Hdac3* | C | Non-mutated |
| *Hdac6* | C | Non-mutated |
| *Hdac7* | C | Non-mutated |
| *Hdac9* | C | Mutated |
| *Hdc* | C | Non-mutated |
| *Hectd1* | C | Non-mutated |
| *Hells* | C | Mutated |
| *Hif3a* | C | Non-mutated |
| *Hint1* | C | Non-mutated |
| *Hist1h1a* | C | Mutated |
| *Hist1h1b* | C | Mutated |
| *Hist1h1t* | C | Non-mutated |
| *Hlf* | C | Non-mutated |
| *Hltf* | C | Non-mutated |
| *Hopx* | C | Non-mutated |
| *Hoxa3* | C | Non-mutated |
| *Hoxa3* | C | Non-mutated |
| *Hoxb1* | C | Non-mutated |
| *Hoxb5* | C | Non-mutated |
| *Hoxd12* | C | Non-mutated |
| *Hoxd13* | C | Non-mutated |
| *Hpx* | C | Non-mutated |
| *Hrg* | C | Non-mutated |
| *Hrh1* | C | Non-mutated |
| *Hsd17b12* | C | Non-mutated |
| *Hsf4* | C | Mutated |
| *Hspa1a* | C | Non-mutated |
| *Hspa1b* | C | Non-mutated |
| *Hspb2* | C | Non-mutated |
| *Htr3b* | C | Non-mutated |
| *Htr5a* | C | Non-mutated |
| *Hvcn1* | C | Mutated |
| *Hyal5* | C | Non-mutated |
| *Icam1* | C | Non-mutated |
| *Icmt* | C | Non-mutated |
| *Icosl* | C | Non-mutated |
| *Id1* | C | Non-mutated |
| *Id2* | C | Non-mutated |
| *Id4* | C | Non-mutated |
| *Ifitm1* | C | Non-mutated |
| *Ifnar1* | C | Non-mutated |
| *Ifnar2* | C | Non-mutated |
| *Igf2* | C | Non-mutated |
| *Igfbp3* | C | Non-mutated |
| *Ighg1* | C | Non-mutated |
| *Ighg2b* | C | Non-mutated |
| *Ighg2c* | C | Non-mutated |
| *Ighm* | C | Non-mutated |
| *Igsf5* | C | Non-mutated |
| *Igsf9* | C | Non-mutated |
| *Ikbkb* | C | Non-mutated |
| *Ikbkg* | C | Non-mutated |
| *Ikzf2* | C | Non-mutated |
| *Il17c* | C | Non-mutated |
| *Il19* | C | Non-mutated |
| *Il21* | C | Non-mutated |
| *Il25* | C | Non-mutated |
| *Il2ra* | C | Non-mutated |
| *Il3* | C | Non-mutated |
| *Il4* | C | Non-mutated |
| *Il6* | C | Non-mutated |
| *Il7* | C | Non-mutated |
| *Il7r* | C | Non-mutated |
| *Ildr2* | C | Mutated |
| *Impact* | C | Non-mutated |
| *Ing1* | C | Non-mutated |
| *Insig1* | C | Non-mutated |
| *Ipmk* | C | Mutated |
| *Ipo11* | C | Non-mutated |
| *Ipo11* | C | Non-mutated |
| *Ipo7* | C | Mutated |
| *Irf6* | C | Non-mutated |
| *Irg1* | C | Non-mutated |
| *Irs1* | C | Non-mutated |
| *Islr2* | C | Non-mutated |
| *Itch* | C | Non-mutated |
| *Itga1* | C | Non-mutated |
| *Itga2b* | C | Non-mutated |
| *Itga4* | C | Mutated |
| *Itga5* | C | Non-mutated |
| *Itgal* | C | Mutated |
| *Itgam* | C | Non-mutated |
| *Itgb1bp3* | C | Non-mutated |
| *Itgb2* | C | Mutated |
| *Itpka* | C | Non-mutated |
| *Jag2* | C | Non-mutated |
| *Jak3* | C | Non-mutated |

| Gene | Origin | Status |
|---|---|---|
| *Jmjd6* | C | Non-mutated |
| *Jun* | C | Non-mutated |
| *Junb* | C | Mutated |
| *Jund* | C | Non-mutated |
| *Kcna5* | C | Non-mutated |
| *Kcnj8* | C | Non-mutated |
| *Kif17* | C | Non-mutated |
| *Kif1b* | C | Non-mutated |
| *Kifap3* | C | Non-mutated |
| *Kiss1* | C | Non-mutated |
| *Kiss1r* | C | Non-mutated |
| *Kit* | C | Non-mutated |
| *Kitl* | C | Non-mutated |
| *Klb* | C | Non-mutated |
| *Klk1* | C | Non-mutated |
| *Kpnb1* | C | Mutated |
| *Krt20* | C | Non-mutated |
| *Krt27* | C | Non-mutated |
| *L1cam* | C | Non-mutated |
| *Lag3* | C | Non-mutated |
| *Lamtor3* | C | Non-mutated |
| *Large* | C | Non-mutated |
| *Lass2* | C | Non-mutated |
| *Lats2* | C | Non-mutated |
| *Lbp* | C | Non-mutated |
| *Lbx1* | C | Non-mutated |
| *Lcat* | C | Non-mutated |
| *Lck* | C | Non-mutated |
| *Lef1* | C | Non-mutated |
| *Lepr* | C | Non-mutated |
| *Lepre1* | C | Non-mutated |
| *Lgals1* | C | Non-mutated |
| *Lgals4* | C | Non-mutated |
| *Lifr* | C | Non-mutated |
| *Lig4* | C | Non-mutated |
| *Lims1* | C | Non-mutated |
| *Lin7a* | C | Non-mutated |
| *Lipa* | C | Non-mutated |
| *Lipg* | C | Non-mutated |
| *Lmbr1* | C | Non-mutated |
| *Lmf1* | C | Non-mutated |
| *Loxhd1* | C | Non-mutated |
| *Lpar1* | C | Non-mutated |
| *Lpin2* | C | Non-mutated |
| *Lsp1* | C | Mutated |
| *Lst1* | C | Non-mutated |
| *Lta* | C | Non-mutated |
| *Lxn* | C | Non-mutated |
| *Ly6e* | C | Mutated |
| *Lyz1* | C | Non-mutated |

| Gene | Origin | Status |
|---|---|---|
| *Madd* | C | Non-mutated |
| *Maf* | C | Non-mutated |
| *Maoa* | C | Non-mutated |
| *Map2k1* | C | Non-mutated |
| *Mapk1* | C | Non-mutated |
| *Mapk3* | C | Non-mutated |
| *Mapk7* | C | Non-mutated |
| *Mapk8ip3* | C | Non-mutated |
| *Mapkbp1* | C | Non-mutated |
| *Mast2* | C | Non-mutated |
| *Mb* | C | Non-mutated |
| *Mbd2* | C | Non-mutated |
| *Mbd3* | C | Non-mutated |
| *Mbl1* | C | Non-mutated |
| *Mbl2* | C | Non-mutated |
| *Mcm2* | C | Mutated |
| *Mcm6* | C | Mutated |
| *Mcm7* | C | Mutated |
| *Mcpt4* | C | Non-mutated |
| *Mdfi* | C | Mutated |
| *Mdh2* | C | Mutated |
| *Mdk* | C | Non-mutated |
| *Mdm2* | C | Non-mutated |
| *Mecp2* | C | Non-mutated |
| *Mef2a* | C | Non-mutated |
| *Mest* | C | Non-mutated |
| *Met* | C | Non-mutated |
| *Mgst3* | C | Non-mutated |
| *Mia1* | C | Non-mutated |
| *Mknk1* | C | Non-mutated |
| *Mlh1* | C | Non-mutated |
| *Mlh3* | C | Mutated |
| *Mll3* | C | Non-mutated |
| *Mlst8* | C | Mutated |
| *Mlx* | C | Non-mutated |
| *Mmp2* | C | Non-mutated |
| *Mnat1* | C | Non-mutated |
| *Mogat1* | C | Non-mutated |
| *Mogat2* | C | Non-mutated |
| *Mov10l1* | C | Non-mutated |
| *Mpdz* | C | Non-mutated |
| *Mrpl51* | C | Mutated |
| *Ms4a8a* | C | Non-mutated |
| *Msh2* | C | Mutated |
| *Msh3* | C | Non-mutated |
| *Msh4* | C | Non-mutated |
| *Msh5* | C | Mutated |
| *Mstn* | C | Non-mutated |
| *Msx1* | C | Non-mutated |
| *Msx3* | C | Non-mutated |

| Gene | Origin | Status |
|---|---|---|
| *Mt1* | C | Non-mutated |
| *Mtap1a* | C | Non-mutated |
| *Mtbp* | C | Non-mutated |
| *Mthfd1* | C | Non-mutated |
| *Mtor* | C | Mutated |
| *Mtss1* | C | Mutated |
| *Myb* | C | Non-mutated |
| *Mybbp1a* | C | Mutated |
| *Mycbp2* | C | Mutated |
| *Myh11* | C | Non-mutated |
| *Myl12b* | C | Non-mutated |
| *Myo1f* | C | Non-mutated |
| *Myo6* | C | Non-mutated |
| *Nampt* | C | Non-mutated |
| *Nanog* | C | Non-mutated |
| *Nanos2* | C | Non-mutated |
| *Nat3* | C | Non-mutated |
| *Ncf1* | C | Non-mutated |
| *Nckap1* | C | Non-mutated |
| *Ncoa1* | C | Non-mutated |
| *Ncor1* | C | Non-mutated |
| *Ndrg1* | C | Non-mutated |
| *Ndufa2* | C | Non-mutated |
| *Ndufa3* | C | Non-mutated |
| *Ndufa7* | C | Non-mutated |
| *Ndufb9* | C | Non-mutated |
| *Nek1* | C | Non-mutated |
| *Neurl1a* | C | Non-mutated |
| *Nf1* | C | Non-mutated |
| *Nf2* | C | Non-mutated |
| *Nfil3* | C | Non-mutated |
| *Nfkb1* | C | Non-mutated |
| *Nfkbid* | C | Non-mutated |
| *Nfx1* | C | Non-mutated |
| *Nfyb* | C | Non-mutated |
| *Nkx2-2* | C | Non-mutated |
| *Nkx3-1* | C | Non-mutated |
| *Nlgn3* | C | Non-mutated |
| *Nlk* | C | Non-mutated |
| *Nlrc4* | C | Non-mutated |
| *Nmnat1* | C | Non-mutated |
| *Nod1* | C | Non-mutated |
| *Nod2* | C | Non-mutated |
| *Nodal* | C | Non-mutated |
| *Nog* | C | Non-mutated |
| *Nos3* | C | Non-mutated |
| *Notch1* | C | Non-mutated |
| *Notch2* | C | Non-mutated |
| *Nox1* | C | Non-mutated |
| *Noxa1* | C | Non-mutated |
| *Noxo1* | C | Non-mutated |
| *Npc1l1* | C | Non-mutated |
| *Npy* | C | Non-mutated |
| *Nr3c2* | C | Non-mutated |
| *Nr4a2* | C | Non-mutated |
| *Ntan1* | C | Mutated |
| *Nup160* | C | Non-mutated |
| *Nup62* | C | Non-mutated |
| *Nupr1* | C | Non-mutated |
| *Oit1* | C | Non-mutated |
| *Oprd1* | C | Non-mutated |
| *Pafah1b1* | C | Non-mutated |
| *Pak1* | C | Non-mutated |
| *Parg* | C | Non-mutated |
| *Parp1* | C | Mutated |
| *Parp2* | C | Non-mutated |
| *Parp4* | C | Non-mutated |
| *Pax2* | C | Non-mutated |
| *Pax6* | C | Non-mutated |
| *Paxip1* | C | Non-mutated |
| *Pcsk1* | C | Non-mutated |
| *Pdcd1lg2* | C | Non-mutated |
| *Pde6g* | C | Non-mutated |
| *Pdgfa* | C | Non-mutated |
| *Pdgfb* | C | Non-mutated |
| *Pdgfc* | C | Non-mutated |
| *Pdgfra* | C | Non-mutated |
| *Pea15b* | C | Non-mutated |
| *Perp* | C | Non-mutated |
| *Pex13* | C | Mutated |
| *Pgm1* | C | Non-mutated |
| *Phf21a* | C | Non-mutated |
| *Phgr1* | C | Non-mutated |
| *Pias1* | C | Mutated |
| *Pick1* | C | Mutated |
| *Pik3c2b* | C | Non-mutated |
| *Pik3cd* | C | Non-mutated |
| *Pim2* | C | Non-mutated |
| *Pin1* | C | Non-mutated |
| *Pip5k1b* | C | Non-mutated |
| *Pitx3* | C | Non-mutated |
| *Pkd1* | C | Non-mutated |
| *Pkn1* | C | Non-mutated |
| *Pla2g15* | C | Non-mutated |
| *Plaa* | C | Non-mutated |
| *Plac8* | C | Mutated |
| *Plat* | C | Non-mutated |
| *Plcb2* | C | Mutated |
| *Plch2* | C | Non-mutated |
| *Plcl1* | C | Non-mutated |

| Gene | Origin | Status |
|---|---|---|
| *Plk2* | C | Non-mutated |
| *Plscr3* | C | Non-mutated |
| *Pms2* | C | Mutated |
| *Pnliprp2* | C | Non-mutated |
| *Pnp* | C | Non-mutated |
| *Pnpla6* | C | Non-mutated |
| *Pola1* | C | Mutated |
| *Pola2* | C | Non-mutated |
| *Polb* | C | Non-mutated |
| *Pold1* | C | Mutated |
| *Polg* | C | Non-mutated |
| *Polh* | C | Non-mutated |
| *Polk* | C | Non-mutated |
| *Polr1b* | C | Non-mutated |
| *Pomc* | C | Non-mutated |
| *Postn* | C | Non-mutated |
| *Pot1a* | C | Non-mutated |
| *Pot1b* | C | Non-mutated |
| *Pou2f1* | C | Mutated |
| *Pou2f2* | C | Mutated |
| *Pou2f3* | C | Non-mutated |
| *Pou3f2* | C | Non-mutated |
| *Pou3f3* | C | Non-mutated |
| *Pou4f1* | C | Non-mutated |
| *Pou4f2* | C | Non-mutated |
| *Pou4f3* | C | Non-mutated |
| *Ppard* | C | Mutated |
| *Ppargc1a* | C | Non-mutated |
| *Ppia* | C | Mutated |
| *Ppm1j* | C | Non-mutated |
| *Ppm1l* | C | Non-mutated |
| *Ppp1r15a* | C | Non-mutated |
| *Ppp1r15b* | C | Mutated |
| *Ppp3ca* | C | Non-mutated |
| *Prdm1* | C | Non-mutated |
| *Prdx6* | C | Non-mutated |
| *Prkaa1* | C | Non-mutated |
| *Prkcc* | C | Non-mutated |
| *Prkci* | C | Non-mutated |
| *Prkdc* | C | Non-mutated |
| *Prkra* | C | Non-mutated |
| *Prl4a1* | C | Non-mutated |
| *Prlr* | C | Non-mutated |
| *Prss1* | C | Non-mutated |
| *Prss3* | C | Non-mutated |
| *Psap* | C | Non-mutated |
| *Psmb4* | C | Non-mutated |
| *Psmc3* | C | Mutated |
| *Psme1* | C | Non-mutated |
| *Psme2* | C | Non-mutated |

| Gene | Origin | Status |
|---|---|---|
| *Ptbp2* | C | Mutated |
| *Pten* | C | Mutated |
| *Ptf1a* | C | Non-mutated |
| *Ptk2* | C | Non-mutated |
| *Ptk2b* | C | Non-mutated |
| *Ptk7* | C | Non-mutated |
| *Ptp4a1* | C | Non-mutated |
| *Ptpn1* | C | Non-mutated |
| *Ptpra* | C | Non-mutated |
| *Ptprm* | C | Non-mutated |
| *Ptprt* | C | Non-mutated |
| *Ptrf* | C | Non-mutated |
| *Pttg1* | C | Non-mutated |
| *Pvrl1* | C | Non-mutated |
| *Pycard* | C | Non-mutated |
| *Qtrtd1* | C | Non-mutated |
| *Rab27a* | C | Non-mutated |
| *Rab3a* | C | Non-mutated |
| *Rab3c* | C | Non-mutated |
| *Rab4a* | C | Non-mutated |
| *Rab5b* | C | Non-mutated |
| *Rab5c* | C | Non-mutated |
| *Rac1* | C | Mutated |
| *Rac3* | C | Non-mutated |
| *Rad17* | C | Non-mutated |
| *Rad23a* | C | Mutated |
| *Rad23b* | C | Non-mutated |
| *Rad51c* | C | Non-mutated |
| *Rad52* | C | Non-mutated |
| *Rad54b* | C | Non-mutated |
| *Rad54b* | C | Non-mutated |
| *Rad54l* | C | Non-mutated |
| *Rad9* | C | Mutated |
| *Rag1* | C | Non-mutated |
| *Rala* | C | Non-mutated |
| *Ramp1* | C | Non-mutated |
| *Rasgrf1* | C | Non-mutated |
| *Rb1* | C | Non-mutated |
| *Rb1cc1* | C | Non-mutated |
| *Rbbp7* | C | Non-mutated |
| *Rbl1* | C | Non-mutated |
| *Rbm15* | C | Mutated |
| *Rbm19* | C | Mutated |
| *Rbpjl* | C | Non-mutated |
| *Rdx* | C | Non-mutated |
| *Rec8* | C | Non-mutated |
| *Recql4* | C | Mutated |
| *Reg4* | C | Non-mutated |
| *Rell2* | C | Non-mutated |
| *Ret* | C | Non-mutated |

| Gene | Origin | Status |
|---|---|---|
| *Rev1* | C | Non-mutated |
| *Rfx3* | C | Non-mutated |
| *Rgma* | C | Non-mutated |
| *Rgmb* | C | Non-mutated |
| *Rgs13* | C | Mutated |
| *Rgs9bp* | C | Non-mutated |
| *Rhoa* | C | Non-mutated |
| *Rhob* | C | Non-mutated |
| *Rictor* | C | Non-mutated |
| *Rims1* | C | Mutated |
| *Rims2* | C | Non-mutated |
| *Rnaseh1* | C | Non-mutated |
| *Rnf11* | C | Non-mutated |
| *Rnf130* | C | Non-mutated |
| *Rnf2* | C | Mutated |
| *Ror2* | C | Non-mutated |
| *Ros1* | C | Non-mutated |
| *Rpa1* | C | Mutated |
| *Rpl29* | C | Mutated |
| *Rpl32* | C | Mutated |
| *Rpl35* | C | Mutated |
| *Rpl4* | C | Mutated |
| *Rpl41* | C | Mutated |
| *Rplp0* | C | Mutated |
| *Rps14* | C | Mutated |
| *Rps5* | C | Mutated |
| *Rps6* | C | Mutated |
| *Rps9* | C | Mutated |
| *Rras* | C | Non-mutated |
| *Rrn3* | C | Non-mutated |
| *Rsc1a1* | C | Non-mutated |
| *Rsu1* | C | Non-mutated |
| *Rtel1* | C | Non-mutated |
| *Runx1t1* | C | Mutated |
| *Ryr2* | C | Non-mutated |
| *S100a16* | C | Non-mutated |
| *S100a6* | C | Non-mutated |
| *Saa1* | C | Non-mutated |
| *Safb* | C | Mutated |
| *Satb2* | C | Non-mutated |
| *Scai* | C | Non-mutated |
| *Scarf2* | C | Non-mutated |
| *Scn1a* | C | Non-mutated |
| *Scnm1* | C | Non-mutated |
| *Sdc1* | C | Non-mutated |
| *Selenbp1* | C | Non-mutated |
| *Selp* | C | Non-mutated |
| *Sema3a* | C | Non-mutated |
| *Serpinb9* | C | Non-mutated |
| *Serpine1* | C | Non-mutated |
| *Sprr2a2* | C | Non-mutated |
| *Sprr2b* | C | Non-mutated |
| *Sra1* | C | Mutated |
| *Src* | C | Non-mutated |
| *Srcin1* | C | Non-mutated |
| *Srf* | C | Non-mutated |
| *Srpk2* | C | Non-mutated |
| *Srrd* | C | Non-mutated |
| *Srsf1* | C | Non-mutated |
| *Ss18l1* | C | Non-mutated |
| *Ssbp1* | C | Non-mutated |
| *Sst* | C | Non-mutated |
| *Sstr2* | C | Non-mutated |
| *Stac* | C | Non-mutated |
| *Stap1* | C | Non-mutated |
| *Star* | C | Non-mutated |
| *Stat3* | C | Non-mutated |
| *Stat4* | C | Non-mutated |
| *Stat5a* | C | Non-mutated |
| *Stat5b* | C | Non-mutated |
| *Stau1* | C | Mutated |
| *Steap4* | C | Non-mutated |
| *Strap* | C | Non-mutated |
| *Stub1* | C | Non-mutated |
| *Stx3* | C | Mutated |
| *Sumo3* | C | Non-mutated |
| *Sv2a* | C | Non-mutated |
| *Sv2b* | C | Mutated |
| *Swap70* | C | Mutated |
| *Syn1* | C | Non-mutated |
| *Sync* | C | Non-mutated |
| *Syngr1* | C | Non-mutated |
| *Syngr3* | C | Non-mutated |
| *Synpo* | C | Non-mutated |
| *Syp* | C | Non-mutated |
| *Sypl2* | C | Non-mutated |
| *Sytl1* | C | Non-mutated |
| *Taf10* | C | Non-mutated |
| *Taf4a* | C | Mutated |
| *Taf7l* | C | Non-mutated |
| *Taf8* | C | Non-mutated |
| *Taf9* | C | Mutated |
| *Taf9* | C | Mutated |
| *Tbp* | C | Non-mutated |
| *Tbpl1* | C | Non-mutated |
| *Tbpl2* | C | Non-mutated |
| *Tcf21* | C | Non-mutated |
| *Tcf3* | C | Mutated |
| *Tcf4* | C | Mutated |
| *Tcf7l2* | C | Non-mutated |

| Gene | Origin | Status |
|---|---|---|
| *Tdrd1* | C | Non-mutated |
| *Tdrd6* | C | Non-mutated |
| *Tead2* | C | Non-mutated |
| *Tek* | C | Non-mutated |
| *Tekt2* | C | Non-mutated |
| *Tenc1* | C | Non-mutated |
| *Tep1* | C | Non-mutated |
| *Tet2* | C | Mutated |
| *Tfap2a* | C | Non-mutated |
| *Tfap2e* | C | Non-mutated |
| *Tfap4* | C | Mutated |
| *Tfcp2l1* | C | Non-mutated |
| *Tfdp1* | C | Mutated |
| *Tfdp2* | C | Non-mutated |
| *Tfe3* | C | Non-mutated |
| *Tfeb* | C | Non-mutated |
| *Tfec* | C | Non-mutated |
| *Tff3* | C | Non-mutated |
| *Tgfa* | C | Non-mutated |
| *Tgfb1* | C | Non-mutated |
| *Tgfb2* | C | Non-mutated |
| *Tgfb3* | C | Non-mutated |
| *Tgfbi* | C | Non-mutated |
| *Tgfbr1* | C | Non-mutated |
| *Tgif2* | C | Non-mutated |
| *Th* | C | Non-mutated |
| *Thra* | C | Non-mutated |
| *Tial1* | C | Non-mutated |
| *Tinf2* | C | Non-mutated |
| *Tk2* | C | Non-mutated |
| *Tlr1* | C | Non-mutated |
| *Tlr2* | C | Non-mutated |
| *Tmlhe* | C | Non-mutated |
| *Tmod1* | C | Non-mutated |
| *Tmsb4x* | C | Mutated |
| *Tnf* | C | Mutated |
| *Tnfaip1* | C | Non-mutated |
| *Tnfrsf11b* | C | Non-mutated |
| *Tnfrsf12a* | C | Non-mutated |
| *Tnfrsf4* | C | Non-mutated |
| *Tnfsf11* | C | Non-mutated |
| *Tnks2* | C | Non-mutated |
| *Tnn* | C | Non-mutated |
| *Tnp2* | C | Non-mutated |
| *Top2a* | C | Mutated |
| *Top2b* | C | Non-mutated |
| *Top3b* | C | Non-mutated |
| *Topors* | C | Mutated |
| *Tpi1* | C | Non-mutated |
| *Tpm1* | C | Non-mutated |
| *Traf3ip2* | C | Non-mutated |
| *Traf4* | C | Non-mutated |
| *Traf6* | C | Non-mutated |
| *Trhr2* | C | Non-mutated |
| *Trim63* | C | Non-mutated |
| *Trp53bp1* | C | Non-mutated |
| *Trp53bp2* | C | Non-mutated |
| *Trpt1* | C | Non-mutated |
| *Tsg101* | C | Non-mutated |
| *Tspan1* | C | Non-mutated |
| *Ttf1* | C | Mutated |
| *Ttll8* | C | Non-mutated |
| *Tyr* | C | Non-mutated |
| *U2af2* | C | Mutated |
| *Uba3* | C | Mutated |
| *Ubb* | C | Mutated |
| *Ube2b* | C | Mutated |
| *Ube2n* | C | Mutated |
| *Ube3a* | C | Non-mutated |
| *Ube4b* | C | Mutated |
| *Ubox5* | C | Mutated |
| *Ubtf* | C | Mutated |
| *Uchl1* | C | Non-mutated |
| *Ucp3* | C | Non-mutated |
| *Ufm1* | C | Non-mutated |
| *Ufm1* | C | Non-mutated |
| *Uhmk1* | C | Mutated |
| *Ulbp1* | C | Non-mutated |
| *Uncx* | C | Non-mutated |
| *Uqcrh* | C | Non-mutated |
| *Uqcrq* | C | Non-mutated |
| *Usp22* | C | Non-mutated |
| *Usp8* | C | Non-mutated |
| *Vav1* | C | Mutated |
| *Vav2* | C | Mutated |
| *Vav3* | C | Non-mutated |
| *Vcan* | C | Non-mutated |
| *Vcl* | C | Mutated |
| *Vdac1* | C | Mutated |
| *Vdr* | C | Non-mutated |
| *Vezt* | C | Non-mutated |
| *Vhl* | C | Non-mutated |
| *Vim* | C | Non-mutated |
| *Vnn3* | C | Non-mutated |
| *Vrk1* | C | Non-mutated |
| *Vsx1* | C | Non-mutated |
| *Vtn* | C | Non-mutated |
| *Vwa1* | C | Non-mutated |
| *Vwc2* | C | Non-mutated |
| *Wisp2* | C | Non-mutated |

| Gene | Origin | Status |
|---|---|---|
| *Wnt7b* | C | Non-mutated |
| *Wt1* | C | Non-mutated |
| *Wwp1* | C | Non-mutated |
| *Xab2* | C | Non-mutated |
| *Xbp1* | C | Mutated |
| *Xiap* | C | Non-mutated |
| *Xpa* | C | Non-mutated |
| *Xpc* | C | Non-mutated |
| *Xrcc6* | C | Non-mutated |
| *Zbtb7a* | C | Non-mutated |
| *Zc3h15* | C | Mutated |
| *Zdhhc1* | C | Non-mutated |
| *Zdhhc8* | C | Non-mutated |
| *Zeb1* | C | Non-mutated |
| *Zfand6* | C | Non-mutated |
| *Zfp467* | C | Non-mutated |
| *Zfp488* | C | Non-mutated |
| *Zg16* | C | Non-mutated |
| *Zglp1* | C | Non-mutated |
| *Zhx2* | C | Non-mutated |
| *Zmiz1* | C | Non-mutated |
| *Zmpste24* | C | Non-mutated |
| *Zp3* | C | Non-mutated |
| *Zscan21* | C | Non-mutated |

**Annex II.** List of the 291 AID targets discovered in *Ung*$^{-/-}$*Msh2*$^{-/-}$ GC B cells.

| | | | Transitions at C/G pairs | | | | | | FDR | | C/G Transition Frequency | | | Total mutation frequency | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Ung$^{-/-}$Msh2$^{-/-}$Exp1 | | Ung$^{-/-}$Msh2$^{-/-}$Exp2 | | *Aicda*$^{-/-}$ | | | | *Ung*$^{-/-}$*Msh2*$^{-/-}$ | | *Aicda*$^{-/-}$ | *Ung*$^{+/-}$*Msh2*$^{+/-}$ | | *Ung*$^{+/-}$*Msh2*$^{-/-}$ | | *Ung*$^{-/-}$*Msh2*$^{+/-}$ | | *Ung*$^{-/-}$*Msh2*$^{-/-}$ | | *Aicda*$^{-/-}$ |
| Gene | Ensembl ID | Coordinates | C/G transitions | C/G sequenced | C/G transitions | C/G sequenced | C/G transitions | C/G sequenced | Exp1 | Exp2 | Exp1 | Exp2 | Exp1 | Exp1 | Exp2 | Exp1 | Exp2 | Exp1 | Exp2 | Exp1 | Exp2 | Exp1 |
| *1810065E05Rik* | ENSMUSG00000013653 | 11:58234613-58235112 | 108 | 578612 | 123 | 654456 | 66 | 616727 | 1,28E-03 | 7,46E-04 | 1,87E-04 | 1,88E-04 | 1,07E-04 | 4,66E-04 | 4,07E-04 | 4,29E-04 | 4,82E-04 | 4,14E-04 | 4,18E-04 | 4,35E-04 | 4,10E-04 | 4,83E-04 |
| *Aatk* | ENSMUSG00000025375 | 11:119907530-119908029 | 115 | 627801 | 132 | 689096 | 70 | 665980 | 8,87E-04 | 1,69E-04 | 1,83E-04 | 1,92E-04 | 1,05E-04 | 3,63E-04 | 3,35E-04 | 3,49E-04 | 3,80E-04 | 3,64E-04 | 3,30E-04 | 3,70E-04 | 3,55E-04 | 3,12E-04 |
| *Abl2* | ENSMUSG00000026596 | 1:158488918-158489417 | 24 | 127053 | 29 | 134774 | 4 | 120744 | 9,98E-04 | 1,60E-04 | 1,89E-04 | 2,15E-04 | 3,31E-05 | 5,54E-04 | 4,58E-04 | 5,18E-04 | 5,99E-04 | 5,21E-04 | 5,12E-04 | 4,88E-04 | 4,77E-04 | 3,93E-04 |
| *Ablim1-1* | ENSMUSG00000025085 | 19:57193000-57193499 | 158 | 871990 | 154 | 977772 | 103 | 917256 | 5,96E-04 | 2,13E-02 | 1,81E-04 | 1,58E-04 | 1,12E-04 | 3,53E-04 | 3,00E-04 | 3,03E-04 | 3,41E-04 | 3,29E-04 | 2,82E-04 | 3,62E-04 | 3,15E-04 | 3,15E-04 |
| *Acot7-2* | ENSMUSG00000028937 | 4:151560243-151560742 | 414 | 600746 | 779 | 686292 | 86 | 655520 | 1,81E-57 | 5,60E-132 | 6,89E-04 | 1,14E-03 | 1,31E-04 | 5,83E-04 | 6,28E-04 | 5,25E-04 | 1,06E-03 | 5,30E-04 | 5,01E-04 | 6,55E-04 | 8,77E-04 | 2,96E-04 |
| *Actb* | ENSMUSG00000029580 | 5:143667904-143668403 | 1283 | 664606 | 1337 | 727186 | 88 | 706775 | 6,36E-287 | 8,41E-278 | 1,93E-03 | 1,84E-03 | 1,25E-04 | 5,08E-04 | 4,43E-04 | 5,85E-04 | 5,30E-04 | 5,68E-04 | 6,63E-04 | 1,46E-03 | 1,41E-03 | 3,40E-04 |
| *Ada* | ENSMUSG00000017697 | 2:163575414-163575913 | 142 | 839004 | 144 | 900476 | 100 | 867497 | 9,12E-03 | 3,03E-02 | 1,69E-04 | 1,60E-04 | 1,15E-04 | 3,71E-04 | 2,90E-04 | 3,53E-04 | 3,47E-04 | 3,77E-04 | 3,40E-04 | 3,69E-04 | 3,46E-04 | 3,39E-04 |
| *Adar-2* | ENSMUSG00000027951 | 3:89534640-89535139 | 158 | 844603 | 190 | 951051 | 122 | 906859 | 1,58E-02 | 2,15E-03 | 1,87E-04 | 2,00E-04 | 1,35E-04 | 3,31E-04 | 3,21E-04 | 3,48E-04 | 3,38E-04 | 3,16E-04 | 2,92E-04 | 3,64E-04 | 3,52E-04 | 3,15E-04 |
| *Agk* | ENSMUSG00000029916 | 6:40275477-40275976 | 128 | 661157 | 128 | 721156 | 87 | 712317 | 2,98E-03 | 2,00E-02 | 1,94E-04 | 1,77E-04 | 1,22E-04 | 3,73E-04 | 3,34E-04 | 3,28E-04 | 3,47E-04 | 3,45E-04 | 2,94E-04 | 3,88E-04 | 3,19E-04 | 3,27E-04 |
| *Aicda* | ENSMUSG00000040627 | 6:122503827-122504326 | 419 | 622206 | 336 | 688299 | 63 | 656677 | 2,98E-68 | 3,96E-42 | 6,73E-04 | 4,88E-04 | 9,59E-05 | 4,26E-04 | 3,31E-04 | 3,35E-04 | 3,40E-04 | 3,67E-04 | 4,17E-04 | 6,19E-04 | 4,89E-04 | 2,90E-04 |
| *Akap8* | ENSMUSG00000024045 | 17:32457599-32458098 | 271 | 921841 | 216 | 991927 | 136 | 937960 | 4,50E-11 | 7,25E-04 | 2,94E-04 | 2,18E-04 | 1,45E-04 | 3,74E-04 | 3,35E-04 | 3,62E-04 | 3,82E-04 | 3,41E-04 | 3,18E-04 | 4,56E-04 | 3,78E-04 | 3,18E-04 |
| *Apex1* | ENSMUSG00000035960 | 14:51544696-51545195 | 112 | 534368 | 133 | 634471 | 70 | 479669 | 3,76E-02 | 3,41E-02 | 2,10E-04 | 2,10E-04 | 1,46E-04 | 3,60E-04 | 2,93E-04 | 3,80E-04 | 3,55E-04 | 3,39E-04 | 3,23E-04 | 4,14E-04 | 3,77E-04 | 4,19E-04 |
| *Apobec1* | ENSMUSG00000040613 | 6:122551963-122552462 | 59 | 216939 | 78 | 243624 | 28 | 242037 | 6,11E-04 | 7,05E-06 | 2,72E-04 | 3,20E-04 | 1,16E-04 | 4,01E-04 | 2,99E-04 | 3,67E-04 | 3,49E-04 | 3,30E-04 | 3,08E-04 | 3,97E-04 | 4,20E-04 | 3,28E-04 |
| *Apoe* | ENSMUSG00000002985 | 7:20284016-20284515 | 217 | 808907 | 187 | 904147 | 105 | 862508 | 6,65E-11 | 5,24E-05 | 2,68E-04 | 2,07E-04 | 1,22E-04 | 3,63E-04 | 3,01E-04 | 3,53E-04 | 3,70E-04 | 3,31E-04 | 3,16E-04 | 3,83E-04 | 4,08E-04 | 3,06E-04 |
| *Atf5-2* | ENSMUSG00000038539 | 7:52071529-52072028 | 221 | 661275 | 272 | 692644 | 122 | 656875 | 7,03E-07 | 1,14E-11 | 3,34E-04 | 3,93E-04 | 1,86E-04 | 4,29E-04 | 3,24E-04 | 3,65E-04 | 3,99E-04 | 3,79E-04 | 3,80E-04 | 5,05E-04 | 4,96E-04 | 3,40E-04 |
| *Atp5b-1* | ENSMUSG00000025393 | 10:127520363-127520862 | 245 | 702451 | 235 | 774079 | 87 | 734085 | 4,76E-19 | 3,08E-14 | 3,49E-04 | 3,04E-04 | 1,19E-04 | 3,62E-04 | 3,04E-04 | 3,39E-04 | 3,78E-04 | 3,47E-04 | 3,21E-04 | 4,56E-04 | 4,38E-04 | 3,10E-04 |
| *Atp5e* | ENSMUSG00000016252 | 2:174289103-174289602 | 150 | 756165 | 169 | 826748 | 109 | 799734 | 8,19E-03 | 3,15E-03 | 1,98E-04 | 2,04E-04 | 1,36E-04 | 3,83E-04 | 3,17E-04 | 3,65E-04 | 3,94E-04 | 3,49E-04 | 3,78E-04 | 3,93E-04 | 3,69E-04 | 3,19E-04 |
| *Atp5o* | ENSMUSG00000022956 | 16:91931376-91931875 | 126 | 627614 | 160 | 689116 | 89 | 675021 | 7,13E-03 | 6,90E-05 | 2,01E-04 | 2,32E-04 | 1,32E-04 | 9,69E-04 | 8,26E-04 | 7,58E-04 | 5,83E-04 | 7,95E-04 | 8,06E-04 | 8,04E-04 | 8,53E-04 | 1,47E-03 |
| *B2m* | ENSMUSG00000060802 | 2:121973423-121973922 | 181 | 865665 | 399 | 979732 | 128 | 900105 | 2,76E-03 | 1,73E-27 | 2,09E-04 | 4,07E-04 | 1,42E-04 | 4,73E-04 | 3,27E-04 | 3,45E-04 | 3,69E-04 | 3,78E-04 | 3,22E-04 | 3,93E-04 | 4,93E-04 | 3,39E-04 |
| *Bad* | ENSMUSG00000024959 | 19:7016345-7016844 | 156 | 951417 | 208 | 1053172 | 119 | 969598 | 3,72E-02 | 1,27E-04 | 1,64E-04 | 1,97E-04 | 1,23E-04 | 3,70E-04 | 3,09E-04 | 3,66E-04 | 3,49E-04 | 3,81E-04 | 2,93E-04 | 3,73E-04 | 3,48E-04 | 3,38E-04 |
| *Bcl11a-1* | ENSMUSG00000000861 | 11:23978056-23978555 | 93 | 308111 | 117 | 348752 | 52 | 329931 | 6,13E-04 | 1,68E-05 | 3,02E-04 | 3,35E-04 | 1,58E-04 | 4,14E-04 | 4,00E-04 | 4,10E-04 | 5,15E-04 | 4,29E-04 | 4,09E-04 | 4,51E-04 | 4,90E-04 | 4,01E-04 |
| *Bcl6* | ENSMUSG00000022508 | 16:23988199-23988698 | 1404 | 417740 | 1939 | 440223 | 120 | 436513 | 8,99E-289 | 0,00E+00 | 3,36E-03 | 4,40E-03 | 2,75E-04 | 9,42E-04 | 6,94E-04 | 6,33E-04 | 5,07E-04 | 6,47E-04 | 7,36E-04 | 1,83E-03 | 2,27E-03 | 4,01E-04 |
| *Bid* | ENSMUSG00000004446 | 6:120866339-120866838 | 183 | 754213 | 284 | 811723 | 141 | 768463 | 2,85E-02 | 1,09E-09 | 2,43E-04 | 3,50E-04 | 1,83E-04 | 3,28E-04 | 3,61E-04 | 3,90E-04 | 3,70E-04 | 3,72E-04 | 3,41E-04 | 3,97E-04 | 4,89E-04 | 3,53E-04 |
| *Blk* | ENSMUSG00000014453 | 14:64035525-64036024 | 102 | 527591 | 106 | 596350 | 69 | 556744 | 1,18E-02 | 4,56E-02 | 1,93E-04 | 1,78E-04 | 1,24E-04 | 3,39E-04 | 3,31E-04 | 3,35E-04 | 4,03E-04 | 3,48E-04 | 3,07E-04 | 3,74E-04 | 3,53E-04 | 3,24E-04 |
| *Bmp2k* | ENSMUSG00000034663 | 5:97426708-97427207 | 47 | 139451 | 62 | 141754 | 23 | 136675 | 1,58E-02 | 2,14E-04 | 3,37E-04 | 4,37E-04 | 1,68E-04 | 3,94E-04 | 3,85E-04 | 4,03E-04 | 3,45E-04 | 3,89E-04 | 3,84E-04 | 5,48E-04 | 5,49E-04 | 3,79E-04 |
| *Brca1* | ENSMUSG00000017146 | 11:101412770-101413269 | 152 | 767386 | 161 | 864750 | 96 | 817166 | 2,50E-04 | 1,23E-03 | 1,98E-04 | 1,86E-04 | 1,17E-04 | 3,54E-04 | 3,15E-04 | 3,65E-04 | 3,45E-04 | 3,51E-04 | 3,24E-04 | 3,62E-04 | 3,82E-04 | 2,96E-04 |
| *Btg1* | ENSMUSG00000036478 | 10:96079635-96080134 | 110 | 427551 | 180 | 467161 | 60 | 430699 | 4,76E-04 | 3,74E-12 | 2,57E-04 | 3,85E-04 | 1,39E-04 | 3,74E-04 | 3,10E-04 | 3,58E-04 | 3,48E-04 | 3,51E-04 | 3,23E-04 | 4,07E-04 | 4,73E-04 | 3,44E-04 |
| *Btg2* | ENSMUSG00000020423 | 1:135975233-135975732 | 145 | 666628 | 159 | 728119 | 104 | 675155 | 1,79E-02 | 1,59E-02 | 2,18E-04 | 2,18E-04 | 1,54E-04 | 3,42E-04 | 3,00E-04 | 3,71E-04 | 3,95E-04 | 3,52E-04 | 3,04E-04 | 3,95E-04 | 3,76E-04 | 3,18E-04 |
| *C2cd3* | ENSMUSG00000047248 | 7:107520743-107521242 | 289 | 760425 | 198 | 846933 | 120 | 795981 | 6,74E-18 | 5,33E-04 | 3,80E-04 | 2,34E-04 | 1,51E-04 | 3,87E-04 | 3,27E-04 | 3,61E-04 | 3,65E-04 | 3,57E-04 | 3,49E-04 | 4,85E-04 | 4,04E-04 | 3,21E-04 |
| *Cacng4* | ENSMUSG00000020723 | 11:107655279-107655778 | 122 | 555422 | 110 | 598492 | 69 | 564512 | 3,74E-04 | 2,16E-02 | 2,20E-04 | 1,84E-04 | 1,22E-04 | 3,87E-04 | 3,45E-04 | 3,67E-04 | 4,11E-04 | 3,76E-04 | 3,23E-04 | 3,80E-04 | 3,46E-04 | 3,14E-04 |
| *Calm1* | ENSMUSG00000001175 | 12:101437751-101438250 | 218 | 582011 | 171 | 632117 | 88 | 605252 | 4,53E-14 | 7,73E-06 | 3,75E-04 | 2,71E-04 | 1,45E-04 | 4,47E-04 | 3,88E-04 | 3,67E-04 | 3,75E-04 | 3,70E-04 | 3,46E-04 | 4,98E-04 | 4,26E-04 | 3,16E-04 |
| *Cat* | ENSMUSG00000027187 | 2:103324811-103325310 | 149 | 750450 | 172 | 833393 | 108 | 790008 | 8,85E-03 | 2,59E-03 | 1,99E-04 | 2,06E-04 | 1,37E-04 | 3,50E-04 | 3,05E-04 | 3,76E-04 | 3,82E-04 | 3,43E-04 | 3,17E-04 | 3,58E-04 | 3,60E-04 | 3,15E-04 |
| *Ccne2-2* | ENSMUSG00000028212 | 4:11118856-11119355 | 138 | 512119 | 213 | 558252 | 98 | 548934 | 5,54E-03 | 1,10E-09 | 2,69E-04 | 3,82E-04 | 1,79E-04 | 6,62E-04 | 6,46E-04 | 7,17E-04 | 7,24E-04 | 6,39E-04 | 6,59E-04 | 6,70E-04 | 7,33E-04 | 6,45E-04 |
| *Cd19* | ENSMUSG00000030724 | 7:133557885-133558384 | 585 | 590847 | 1082 | 636713 | 80 | 628528 | 6,06E-101 | 1,87E-220 | 9,90E-04 | 1,70E-03 | 1,27E-04 | 8,39E-04 | 9,48E-04 | 7,26E-04 | 5,96E-04 | 6,82E-04 | 5,25E-04 | 7,99E-04 | 1,17E-03 | 3,23E-04 |
| *Cd22-1* | ENSMUSG00000030577 | 7:31663548-31664047 | 210 | 508233 | 170 | 583742 | 101 | 552839 | 3,34E-11 | 7,25E-04 | 4,13E-04 | 2,91E-04 | 1,83E-04 | 4,20E-04 | 4,07E-04 | 4,21E-04 | 4,68E-04 | 4,05E-04 | 3,96E-04 | 5,10E-04 | 4,80E-04 | 3,92E-04 |
| *Cd22-2* | ENSMUSG00000030577 | 7:31664862-31665361 | 121 | 594181 | 117 | 648737 | 67 | 628974 | 7,86E-05 | 1,91E-03 | 2,04E-04 | 1,80E-04 | 1,07E-04 | 3,35E-04 | 2,87E-04 | 3,29E-04 | 3,62E-04 | 3,40E-04 | 2,78E-04 | 3,62E-04 | 3,51E-04 | 2,84E-04 |
| *Cd24a* | ENSMUSG00000047139 | 10:43298975-43299474 | 430 | 668120 | 455 | 706462 | 96 | 690421 | 1,75E-52 | 8,83E-54 | 6,44E-04 | 6,44E-04 | 1,39E-04 | 4,21E-04 | 3,76E-04 | 4,08E-04 | 4,01E-04 | 4,40E-04 | 4,13E-04 | 7,11E-04 | 6,59E-04 | 3,97E-04 |
| *Cd37* | ENSMUSG00000030798 | 7:52493710-52494209 | 228 | 622269 | 239 | 693742 | 76 | 655705 | 1,63E-19 | 8,06E-18 | 3,66E-04 | 3,45E-04 | 1,16E-04 | 3,31E-04 | 3,34E-04 | 3,37E-04 | 3,65E-04 | 3,27E-04 | 3,03E-04 | 4,90E-04 | 4,75E-04 | 3,44E-04 |
| *Cd48* | ENSMUSG00000015355 | 1:173612186-173612685 | 153 | 469597 | 217 | 521041 | 56 | 477045 | 4,50E-11 | 1,59E-19 | 3,26E-04 | 4,16E-04 | 1,17E-04 | 3,74E-04 | 3,15E-04 | 3,47E-04 | 3,28E-04 | 3,46E-04 | 3,21E-04 | 4,30E-04 | 4,73E-04 | 3,02E-04 |
| *Cd53* | ENSMUSG00000040747 | 3:106592568-106593067 | 149 | 758706 | 177 | 884426 | 96 | 842871 | 1,33E-04 | 3,20E-05 | 1,96E-04 | 2,00E-04 | 1,14E-04 | 3,77E-04 | 3,30E-04 | 3,48E-04 | 3,79E-04 | 3,51E-04 | 3,13E-04 | 3,63E-04 | 3,64E-04 | 3,06E-04 |
| *Cd74* | ENSMUSG00000024610 | 18:60963503-60964002 | 257 | 339907 | 354 | 398425 | 15 | 258186 | 2,31E-43 | 1,04E-55 | 7,56E-04 | 8,88E-04 | 5,81E-05 | 5,66E-04 | 7,69E-04 | 9,54E-04 | 5,41E-04 | 1,13E-03 | 5,97E-04 | 1,60E-03 | 1,21E-03 | 4,08E-03 |
| *Cd79a* | ENSMUSG00000003379 | 7:25682530-25683029 | 216 | 364021 | 269 | 401722 | 38 | 382713 | 3,39E-32 | 2,16E-40 | 5,93E-04 | 6,70E-04 | 9,93E-05 | 3,61E-04 | 3,13E-04 | 3,19E-04 | 3,61E-04 | 3,42E-04 | 3,59E-04 | 5,86E-04 | 5,84E-04 | 3,02E-04 |
| *Cd79b* | ENSMUSG00000040592 | 11:106175344-106175843 | 391 | 507479 | 480 | 564481 | 53 | 538105 | 1,34E-67 | 2,94E-81 | 7,70E-04 | 8,50E-04 | 9,85E-05 | 5,53E-04 | 4,89E-04 | 5,34E-04 | 6,53E-04 | 5,02E-04 | 5,03E-04 | 8,56E-04 | 8,71E-04 | 4,49E-04 |
| *Cd83* | ENSMUSG00000015396 | 13:43880476-43880975 | 1044 | 798791 | 1150 | 867530 | 134 | 848537 | 1,85E-185 | 3,80E-195 | 1,31E-03 | 1,33E-03 | 1,58E-04 | 5,49E-04 | 5,38E-04 | 6,00E-04 | 5,07E-04 | 6,56E-04 | 6,00E-04 | 1,13E-03 | 1,15E-03 | 3,54E-04 |
| *Cdk11b* | ENSMUSG00000029062 | 4:154998978-154999477 | 170 | 846241 | 187 | 894894 | 116 | 865838 | 2,53E-03 | 6,09E-04 | 2,01E-04 | 2,09E-04 | 1,34E-04 | 3,33E-04 | 3,17E-04 | 3,56E-04 | 3,81E-04 | 3,40E-04 | 3,02E-04 | 4,01E-04 | 3,48E-04 | 3,09E-04 |
| *Cdk4* | ENSMUSG00000006728 | 10:126500659-126501158 | 186 | 616638 | 223 | 661600 | 114 | 709287 | 5,15E-07 | 3,59E-10 | 3,02E-04 | 3,37E-04 | 1,61E-04 | 7,33E-04 | 9,87E-04 | 5,26E-04 | 1,11E-03 | 8,69E-04 | 5,08E-04 | 6,40E-04 | 1,03E-03 | 3,51E-04 |
| *Chd2* | ENSMUSG00000078671 | 7:80686133-80686632 | 198 | 729975 | 171 | 809722 | 81 | 760913 | 1,18E-12 | 1,17E-06 | 2,71E-04 | 2,11E-04 | 1,06E-04 | 3,09E-04 | 2,96E-04 | 3,41E-04 | 3,55E-04 | 3,59E-04 | 3,21E-04 | 4,11E-04 | 3,64E-04 | 3,18E-04 |
| *Chek1* | ENSMUSG00000032113 | 9:36533744-36534243 | 129 | 587896 | 219 | 655426 | 86 | 638564 | 1,55E-03 | 7,32E-13 | 2,19E-04 | 3,34E-04 | 1,35E-04 | 3,57E-04 | 3,62E-04 | 3,57E-04 | 4,12E-04 | 3,99E-04 | 3,31E-04 | 4,08E-04 | 4,49E-04 | 3,06E-04 |
| *Ciita* | ENSMUSG00000075404 | 16:10489874-10490373 | 286 | 598029 | 368 | 667910 | 89 | 635646 | 9,12E-27 | 3,45E-37 | 4,78E-04 | 5,51E-04 | 1,40E-04 | 5,61E-04 | 3,77E-04 | 4,20E-04 | 4,87E-04 | 4,43E-04 | 3,90E-04 | 5,84E-04 | 6,54E-04 | 4,04E-04 |
| *Ciita-2* | ENSMUSG00000022504 | 16:10488271-10488770 | 413 | 938622 | 350 | 1053849 | 107 | 995011 | 5,28E-46 | 4,19E-27 | 4,40E-04 | 3,32E-04 | 1,08E-04 | 3,27E-04 | 3,11E-04 | 3,35E-04 | 3,81E-04 | 3,11E-04 | 3,29E-04 | 4,91E-04 | 4,06E-04 | 3,09E-04 |
| *Clk1* | ENSMUSG00000026034 | 1:58480433-58480932 | 220 | 688612 | 263 | 739819 | 105 | 703889 | 3,08E-10 | 3,90E-14 | 3,19E-04 | 3,55E-04 | 1,49E-04 | 3,53E-04 | 3,33E-04 | 3,62E-04 | 3,91E-04 | 3,80E-04 | 3,22E-04 | 4,84E-04 | 4,66E-04 | 3,35E-04 |
| *Cnbp* | ENSMUSG00000030057 | 6:87800601-87801100 | 153 | 516603 | 148 | 540939 | 82 | 518739 | 1,61E-05 | 2,37E-04 | 2,96E-04 | 2,74E-04 | 1,58E-04 | 3,72E-04 | 3,02E-04 | 3,49E-04 | 3,79E-04 | 3,51E-04 | 3,17E-04 | 4,39E-04 | 4,32E-04 | 3,11E-04 |
| *Cox4i1* | ENSMUSG00000031818 | 8:123192190-123192689 | 133 | 610439 | 149 | 665493 | 94 | 631579 | 1,24E-02 | 5,91E-03 | 2,18E-04 | 2,24E-04 | 1,49E-04 | 4,00E-04 | 3,38E-04 | 3,71E-04 | 4,22E-04 | 3,90E-04 | 3,21E-04 | 4,18E-04 | 4,02E-04 | 3,37E-04 |
| *Cox6a1* | ENSMUSG00000041697 | 5:115798465-115798964 | 139 | 640715 | 136 | 668150 | 94 | 646443 | 7,82E-03 | 3,07E-02 | 2,17E-04 | 2,04E-04 | 1,45E-04 | 3,38E-04 | 3,54E-04 | 3,45E-04 | 3,81E-04 | 3,89E-04 | 3,61E-04 | 3,97E-04 | 4,04E-04 | 3,56E-04 |
| *Cox8a* | ENSMUSG00000035885 | 19:7291607-7292106 | 229 | 841672 | 242 | 893226 | 139 | 828713 | 2,81E-05 | 2,60E-05 | 2,72E-04 | 2,71E-04 | 1,68E-04 | 7,43E-04 | 7,21E-04 | 1,01E-03 | 8,63E-04 | 9,99E-04 | 1,09E-03 | 7,90E-04 | 8,26E-04 | 6,90E-04 |
| *Cradd* | ENSMUSG00000045867 | 10:94786232-94786731 | 207 | 456848 | 181 | 526336 | 130 | 495059 | 4,78E-06 | 4,35E-02 | 4,53E-04 | 3,44E-04 | 2,63E-04 | 5,40E-04 | 5,28E-04 | 5,75E-04 | 6,10E-04 | 5,35E-04 | 5,52E-04 | 5,86E-04 | 5,22E-04 | 5,03E-04 |
| *Csk* | ENSMUSG00000032312 | 9:57492488-57492987 | 298 | 489386 | 358 | 544273 | 65 | 530077 | 1,20E-39 | 3,33E-47 | 6,09E-04 | 6,58E-04 | 1,23E-04 | 4,92E-04 | 4,10E-04 | 4,73E-04 | 4,68E-04 | 4,64E-04 | 4,37E-04 | 7,64E-04 | 7,21E-04 | 3,64E-04 |
| *Csnk1d* | ENSMUSG00000025162 | 11:120852148-120852647 | 152 | 555365 | 142 | 585236 | 103 | 591508 | 1,38E-03 | 2,72E-02 | 2,74E-04 | 2,43E-04 | 1,74E-04 | 4,01E-04 | 3,57E-04 | 3,86E-04 | 4,09E-04 | 3,56E-04 | 3,26E-04 | 4,17E-04 | 3,86E-04 | 3,70E-04 |
| *Cyth1* | ENSMUSG00000017132 | 11:118109407-118109906 | 214 | 747423 | 223 | 763562 | 101 | 755887 | 6,20E-10 | 1,38E-10 | 2,86E-04 | 2,92E-04 | 1,34E-04 | 4,02E-04 | 2,96E-04 | 3,71E-04 | 3,95E-04 | 3,59E-04 | 3,21E-04 | 4,84E-04 | 4,67E-04 | 3,68E-04 |
| *Daxx* | ENSMUSG00000002307 | 17:34046443-34046942 | 137 | 518202 | 123 | 569451 | 75 | 537203 | 3,28E-05 | 8,57E-03 | 2,64E-04 | 2,16E-04 | 1,40E-04 | 4,12E-04 | 3,17E-04 | 4,13E-04 | 4,33E-04 | 3,27E-04 | 3,58E-04 | 4,19E-04 | 3,54E-04 | 3,39E-04 |
| *Ddb2* | ENSMUSG00000002109 | 2:91076724-91077223 | 166 | 679079 | 173 | 748034 | 82 | 688244 | 3,05E-07 | 2,39E-06 | 2,44E-04 | 2,31E-04 | 1,19E-04 | 3,49E-04 | 3,22E-04 | 3,88E-04 | 3,79E-04 | 3,48E-04 | 3,07E-04 | 4,33E-04 | 3,72E-04 | 3,08E-04 |
| *Ddx20* | ENSMUSG00000027905 | 3:105489990-105490489 | 182 | 523935 | 174 | 558189 | 80 | 531241 | 8,32E-10 | 1,96E-07 | 3,47E-04 | 3,12E-04 | 1,51E-04 | 4,12E-04 | 3,45E-04 | 4,27E-04 | 3,64E-04 | 3,98E-04 | 3,13E-04 | 5,09E-04 | 4,58E-04 | 3,29E-04 |
| *Ddx5* | ENSMUSG00000020719 | 11:106649309-106649808 | 272 | 590855 | 242 | 644855 | 110 | 596950 | 2,55E-16 | 1,41E-09 | 4,60E-04 | 3,75E-04 | 1,84E-04 | 5,38E-04 | 5,52E-04 | 5,23E-04 | 6,26E-04 | 5,45E-04 | 5,20E-04 | 6,97E-04 | 6,23E-04 | 5,63E-04 |
| *Dnmt1* | ENSMUSG00000004099 | 9:20756818-20757317 | 322 | 641931 | 188 | 700547 | 124 | 689646 | 1,82E-23 | 1,81E-03 | 5,02E-04 | 2,68E-04 | 1,80E-04 | 4,36E-04 | 3,78E-04 | 4,23E-04 | 4,25E-04 | 3,77E-04 | 3,55E-04 | 6,13E-04 | 4,42E-04 | 3,96E-04 |
| *Dok1* | ENSMUSG00000068335 | 6:82982966-82983465 | 138 | 790501 | 161 | 859276 | 105 | 807835 | 4,56E-02 | 1,05E-02 | 1,75E-04 | 1,87E-04 | 1,30E-04 | 3,50E-04 | 3,03E-04 | 3,53E-04 | 3,65E-04 | 3,44E-04 | 2,83E-04 | 3,51E-04 | 3,79E-04 | 3,10E-04 |

| | | | Transitions at C/G pairs | | | | | | | | C/G Transition Frequency | | | | | | Total mutation frequency | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Ung-/-Msh2-/-Exp1 | | Ung-/-Msh2-/-Exp2 | | Aicda-/- | | FDR | | Ung-/-Msh2-/- | | Aicda-/- | | Ung+/-Msh2+/- | | Ung+/-Msh2-/- | | Ung-/-Msh2+/- | | Ung-/-Msh2-/- | | Aicda-/- | |
| Gene | Ensembl ID | Coordinates | C/G transitions | C/G sequenced | C/G transitions | C/G sequenced | C/G transitions | C/G sequenced | Exp1 | Exp2 | Exp1 | Exp2 | Exp1 | | Exp1 | Exp2 | Exp1 | Exp2 | Exp1 | Exp2 | Exp1 | Exp2 | Exp1 | |
| *Dsg4* | ENSMUSG00000001804 | 18:20594676-20595175 | 234 | 643280 | 260 | 741818 | 194 | 713232 | 7,90E-03 | 2,03E-02 | 3,64E-04 | 3,50E-04 | 2,72E-04 | | 4,35E-04 | 4,16E-04 | 4,54E-04 | 5,35E-04 | 4,38E-04 | 3,65E-04 | 4,57E-04 | 4,27E-04 | 4,09E-04 | |
| *Dusp6* | ENSMUSG00000019960 | 10:98725865-98726364 | 679 | 620088 | 822 | 691274 | 82 | 654408 | 3,56E-123 | 5,45E-144 | 1,10E-03 | 1,19E-03 | 1,25E-04 | | 1,04E-03 | 1,11E-03 | 7,99E-04 | 1,27E-03 | 9,95E-04 | 5,16E-04 | 8,49E-04 | 9,21E-04 | 3,19E-04 | |
| *Dyrk3* | ENSMUSG00000016526 | 1:133034312-133034811 | 729 | 464930 | 337 | 500073 | 70 | 436631 | 1,55E-128 | 2,52E-34 | 1,57E-03 | 6,74E-04 | 1,60E-04 | | 1,01E-03 | 1,14E-03 | 1,32E-03 | 1,32E-03 | 1,37E-03 | 6,26E-04 | 1,52E-03 | 8,63E-04 | 4,82E-04 | |
| *E2f1* | ENSMUSG00000027490 | 2:154395089-154395588 | 130 | 629443 | 151 | 667660 | 93 | 659209 | 1,33E-02 | 1,21E-03 | 2,07E-04 | 2,26E-04 | 1,41E-04 | | 3,41E-04 | 3,02E-04 | 3,43E-04 | 3,87E-04 | 3,55E-04 | 2,94E-04 | 3,87E-04 | 3,58E-04 | 3,25E-04 | |
| *E2f2* | ENSMUSG00000018983 | 4:135728309-135728808 | 185 | 454184 | 169 | 506415 | 72 | 490306 | 1,14E-13 | 1,05E-08 | 4,07E-04 | 3,34E-04 | 1,47E-04 | | 4,08E-04 | 3,43E-04 | 3,80E-04 | 4,35E-04 | 3,48E-04 | 3,95E-04 | 5,51E-04 | 4,74E-04 | 3,01E-04 | |
| *E4f1* | ENSMUSG00000024137 | 17:24591757-24592256 | 205 | 755753 | 155 | 816018 | 102 | 735467 | 9,44E-08 | 3,36E-02 | 2,71E-04 | 1,90E-04 | 1,39E-04 | | 3,73E-04 | 3,29E-04 | 3,54E-04 | 3,50E-04 | 3,14E-04 | 3,51E-04 | 4,36E-04 | 3,80E-04 | 3,48E-04 | |
| *Ebf1* | ENSMUSG00000057098 | 11:44431636-44432135 | 233 | 449050 | 345 | 478696 | 87 | 464618 | 1,20E-16 | 1,43E-34 | 5,19E-04 | 7,21E-04 | 1,87E-04 | | 3,65E-04 | 3,38E-04 | 3,79E-04 | 3,94E-04 | 4,11E-04 | 3,58E-04 | 6,21E-04 | 6,99E-04 | 3,53E-04 | |
| *Eef1a1* | ENSMUSG00000037742 | 9:78329032-78329531 | 459 | 715067 | 512 | 769746 | 107 | 737892 | 6,90E-54 | 2,84E-59 | 6,42E-04 | 6,65E-04 | 1,45E-04 | | 2,50E-03 | 2,36E-03 | 1,43E-03 | 2,43E-03 | 1,82E-03 | 2,38E-03 | 1,60E-03 | 1,91E-03 | 2,95E-03 | |
| *Eif2a* | ENSMUSG00000027810 | 3:58329743-58330242 | 146 | 830559 | 175 | 915815 | 116 | 890209 | 3,56E-02 | 4,44E-03 | 1,76E-04 | 1,91E-04 | 1,30E-04 | | 3,27E-04 | 3,29E-04 | 3,61E-04 | 3,62E-04 | 3,40E-04 | 3,01E-04 | 3,66E-04 | 3,59E-04 | 3,35E-04 | |
| *Eif3a* | ENSMUSG00000024991 | 19:60866097-60866596 | 455 | 996063 | 306 | 1057507 | 166 | 1022373 | 1,68E-32 | 7,47E-09 | 4,57E-04 | 2,89E-04 | 1,62E-04 | | 3,54E-04 | 3,52E-04 | 3,52E-04 | 3,79E-04 | 3,63E-04 | 3,34E-04 | 5,78E-04 | 4,27E-04 | 3,28E-04 | |
| *Eif3d* | ENSMUSG00000016554 | 15:77800755-77801254 | 243 | 796707 | 261 | 873457 | 126 | 820784 | 1,17E-09 | 2,25E-09 | 3,05E-04 | 2,99E-04 | 1,54E-04 | | 3,71E-04 | 3,24E-04 | 3,65E-04 | 3,93E-04 | 3,31E-04 | 3,61E-04 | 4,46E-04 | 4,17E-04 | 3,52E-04 | |
| *Eif4a2* | ENSMUSG00000022884 | 16:23107552-23108051 | 404 | 957514 | 414 | 1038510 | 135 | 980345 | 2,33E-32 | 2,17E-29 | 4,22E-04 | 3,99E-04 | 1,38E-04 | | 3,12E-04 | 3,45E-04 | 3,50E-04 | 3,52E-04 | 3,65E-04 | 3,45E-04 | 5,09E-04 | 4,83E-04 | 3,11E-04 | |
| *Eif5a-1* | ENSMUSG00000078812 | 11:69733633-69734132 | 374 | 742478 | 368 | 770568 | 114 | 757801 | 2,89E-33 | 3,74E-30 | 5,04E-04 | 4,78E-04 | 1,50E-04 | | 4,94E-04 | 3,01E-04 | 3,63E-04 | 3,77E-04 | 3,59E-04 | 3,59E-04 | 5,53E-04 | 5,45E-04 | 3,21E-04 | |
| *Eif5a-3* | ENSMUSG00000078812 | 11:69734088-69734587 | 236 | 627743 | 173 | 664189 | 121 | 668428 | 1,74E-10 | 6,64E-03 | 3,76E-04 | 2,60E-04 | 1,81E-04 | | 6,87E-04 | 5,96E-04 | 6,78E-04 | 6,69E-04 | 6,66E-04 | 6,61E-04 | 7,91E-04 | 6,64E-04 | 5,69E-04 | |
| *Eif5a-4* | ENSMUSG00000078812 | 11:69734278-69734777 | 315 | 699639 | 210 | 758573 | 107 | 729400 | 1,18E-25 | 3,01E-07 | 4,50E-04 | 2,77E-04 | 1,47E-04 | | 5,43E-04 | 4,19E-04 | 4,65E-04 | 4,70E-04 | 4,79E-04 | 4,33E-04 | 6,90E-04 | 5,16E-04 | 4,17E-04 | |
| *Eif5a-5* | ENSMUSG00000078812 | 11:69734389-69734888 | 289 | 721139 | 206 | 778918 | 109 | 754042 | 1,27E-20 | 1,24E-06 | 4,01E-04 | 2,64E-04 | 1,45E-04 | | 5,04E-04 | 3,85E-04 | 4,28E-04 | 4,34E-04 | 4,41E-04 | 3,95E-04 | 6,24E-04 | 4,59E-04 | 3,88E-04 | |
| *Ell* | ENSMUSG00000070002 | 8:73063574-73064073 | 106 | 492639 | 112 | 555718 | 66 | 522688 | 2,23E-03 | 7,73E-03 | 2,15E-04 | 2,02E-04 | 1,26E-04 | | 3,74E-04 | 3,05E-04 | 3,16E-04 | 4,18E-04 | 3,34E-04 | 3,05E-04 | 3,77E-04 | 4,13E-04 | 3,28E-04 | |
| *Emu* | NoCode002 | 12:114665496-114665807 | 576 | 394268 | 609 | 459225 | 52 | 439275 | 1,03E-123 | 1,21E-114 | 1,46E-03 | 1,33E-03 | 1,18E-04 | | 7,80E-04 | 6,68E-04 | 5,67E-04 | 7,36E-04 | 6,56E-04 | 5,88E-04 | 8,96E-04 | 8,51E-04 | 3,10E-04 | |
| *Erh* | ENSMUSG00000021131 | 12:81744349-81744848 | 143 | 659794 | 192 | 706210 | 104 | 691754 | 1,22E-02 | 4,31E-06 | 2,17E-04 | 2,72E-04 | 1,50E-04 | | 4,13E-04 | 2,96E-04 | 3,93E-04 | 4,14E-04 | 3,81E-04 | 3,31E-04 | 4,29E-04 | 4,27E-04 | 3,21E-04 | |
| *Eri1* | ENSMUSG00000031527 | 8:36558088-36558587 | 99 | 526061 | 163 | 553046 | 62 | 536641 | 7,53E-03 | 3,74E-10 | 1,88E-04 | 2,95E-04 | 1,16E-04 | | 3,46E-04 | 3,56E-04 | 3,69E-04 | 4,02E-04 | 3,61E-04 | 3,04E-04 | 3,71E-04 | 4,36E-04 | 3,20E-04 | |
| *Ets1* | ENSMUSG00000032035 | 9:32503627-32504126 | 287 | 524073 | 267 | 586650 | 90 | 544529 | 1,25E-25 | 1,02E-17 | 5,48E-04 | 4,55E-04 | 1,65E-04 | | 4,24E-04 | 3,63E-04 | 3,90E-04 | 3,71E-04 | 4,55E-04 | 3,63E-04 | 6,23E-04 | 5,30E-04 | 3,62E-04 | |
| *Fas* | ENSMUSG00000024778 | 19:34365149-34365648 | 126 | 518153 | 119 | 551763 | 53 | 459575 | 1,37E-05 | 4,49E-04 | 2,43E-04 | 2,16E-04 | 1,15E-04 | | 4,24E-04 | 3,86E-04 | 4,54E-04 | 3,94E-04 | 4,07E-04 | 3,80E-04 | 5,23E-04 | 4,38E-04 | 3,23E-04 | |
| *Fchsd2* | ENSMUSG00000030691 | 7:108257289-108257788 | 476 | 745466 | 659 | 816912 | 120 | 776490 | 6,81E-53 | 1,45E-83 | 6,39E-04 | 8,07E-04 | 1,55E-04 | | 4,66E-04 | 4,97E-04 | 4,03E-04 | 4,21E-04 | 4,56E-04 | 4,87E-04 | 7,20E-04 | 8,07E-04 | 3,57E-04 | |
| *Fen1* | ENSMUSG00000024742 | 19:10277934-10278433 | 172 | 862644 | 190 | 976838 | 117 | 912862 | 8,38E-04 | 1,33E-03 | 1,99E-04 | 1,95E-04 | 1,28E-04 | | 3,14E-04 | 3,00E-04 | 3,28E-04 | 3,91E-04 | 3,50E-04 | 3,01E-04 | 3,68E-04 | 3,37E-04 | 3,33E-04 | |
| *Fli1* | ENSMUSG00000016087 | 9:32348454-32348953 | 237 | 862713 | 258 | 904025 | 135 | 892881 | 1,19E-07 | 7,45E-09 | 2,75E-04 | 2,85E-04 | 1,51E-04 | | 6,02E-04 | 3,11E-04 | 3,72E-04 | 3,68E-04 | 3,76E-04 | 3,11E-04 | 4,22E-04 | 4,47E-04 | 3,49E-04 | |
| *Fnbp1-2* | ENSMUSG00000075415 | 2:30997029-30997528 | 175 | 492567 | 169 | 514751 | 78 | 510802 | 1,02E-09 | 5,75E-08 | 3,55E-04 | 3,28E-04 | 1,53E-04 | | 4,02E-04 | 3,41E-04 | 4,04E-04 | 4,26E-04 | 3,91E-04 | 3,54E-04 | 5,10E-04 | 4,37E-04 | 3,63E-04 | |
| *Foxo4* | ENSMUSG00000042903 | X:98449867-98450366 | 83 | 564888 | 93 | 537789 | 45 | 556546 | 3,78E-03 | 8,87E-05 | 1,47E-04 | 1,73E-04 | 8,09E-05 | | 3,51E-04 | 3,17E-04 | 3,33E-04 | 3,68E-04 | 3,69E-04 | 3,16E-04 | 3,41E-04 | 3,39E-04 | 2,79E-04 | |
| *Fth1* | ENSMUSG00000024661 | 19:10057193-10057692 | 157 | 739271 | 193 | 809262 | 107 | 746410 | 5,13E-03 | 8,72E-05 | 2,12E-04 | 2,38E-04 | 1,43E-04 | | 3,62E-04 | 2,98E-04 | 3,46E-04 | 3,89E-04 | 3,57E-04 | 2,84E-04 | 4,19E-04 | 3,71E-04 | 3,37E-04 | |
| *Ftl1* | ENSMUSG00000050708 | 7:52714757-52715256 | 92 | 304615 | 78 | 335544 | 46 | 315947 | 1,84E-04 | 3,06E-02 | 3,02E-04 | 2,32E-04 | 1,46E-04 | | 4,41E-04 | 4,24E-04 | 4,21E-04 | 4,59E-04 | 4,59E-04 | 4,51E-04 | 5,70E-04 | 4,99E-04 | 3,58E-04 | |
| *Gadd45b* | ENSMUSG00000015312 | 10:80392836-80393335 | 464 | 868383 | 364 | 916664 | 112 | 892678 | 1,40E-52 | 1,49E-29 | 5,34E-04 | 3,97E-04 | 1,25E-04 | | 3,72E-04 | 3,36E-04 | 3,79E-04 | 3,70E-04 | 4,01E-04 | 3,29E-04 | 5,98E-04 | 4,77E-04 | 2,83E-04 | |
| *Gadd45g* | ENSMUSG00000021453 | 13:51942044-51942543 | 114 | 590842 | 285 | 657055 | 84 | 619943 | 3,19E-02 | 3,94E-23 | 1,93E-04 | 4,34E-04 | 1,35E-04 | | 4,52E-04 | 3,41E-04 | 3,48E-04 | 3,58E-04 | 3,42E-04 | 3,13E-04 | 3,51E-04 | 5,00E-04 | 3,16E-04 | |
| *Galnt1-1* | ENSMUSG00000000420 | 18:24363845-24364344 | 104 | 316197 | 118 | 337701 | 40 | 323734 | 2,29E-07 | 1,22E-08 | 3,29E-04 | 3,49E-04 | 1,24E-04 | | 6,08E-04 | 4,27E-04 | 4,32E-04 | 5,13E-04 | 5,17E-04 | 5,93E-04 | 4,64E-04 | 4,72E-04 | 2,97E-04 | |
| *Galnt1-2* | ENSMUSG00000000420 | 18:24364495-24364994 | 79 | 359102 | 145 | 398682 | 46 | 379759 | 3,97E-03 | 3,21E-11 | 2,20E-04 | 3,64E-04 | 1,21E-04 | | 3,44E-04 | 3,31E-04 | 3,70E-04 | 3,72E-04 | 3,28E-04 | 3,76E-04 | 3,25E-04 | 4,68E-04 | 2,90E-04 | |
| *Gas5* | ENSMUSG00000053332 | 1:162965297-162965796 | 577 | 615634 | 794 | 668642 | 107 | 628751 | 2,26E-79 | 8,85E-120 | 9,37E-04 | 1,19E-03 | 1,70E-04 | | 2,69E-03 | 2,50E-03 | 8,57E-04 | 2,41E-03 | 1,83E-03 | 2,55E-03 | 1,75E-03 | 2,95E-03 | 4,36E-03 | |
| *Gdi2* | ENSMUSG00000021218 | 13:3537321-3537820 | 253 | 596124 | 279 | 635569 | 150 | 606710 | 6,46E-07 | 5,23E-08 | 4,24E-04 | 4,39E-04 | 2,47E-04 | | 4,75E-04 | 4,22E-04 | 4,90E-04 | 5,03E-04 | 5,34E-04 | 5,15E-04 | 6,55E-04 | 6,23E-04 | 5,17E-04 | |
| *Gna13* | ENSMUSG00000020611 | 11:109224108-109224607 | 208 | 642595 | 296 | 684780 | 101 | 642910 | 6,78E-09 | 1,13E-19 | 3,24E-04 | 4,32E-04 | 1,57E-04 | | 3,66E-04 | 3,45E-04 | 3,84E-04 | 4,33E-04 | 4,01E-04 | 3,43E-04 | 4,99E-04 | 6,36E-04 | 3,59E-04 | |
| *Grap* | ENSMUSG00000004837 | 11:61466823-61467322 | 240 | 391575 | 193 | 439740 | 54 | 437463 | 1,27E-32 | 2,40E-18 | 6,13E-04 | 4,39E-04 | 1,23E-04 | | 4,10E-04 | 3,06E-04 | 3,43E-04 | 3,46E-04 | 3,24E-04 | 3,40E-04 | 5,80E-04 | 4,69E-04 | 3,35E-04 | |
| *Gsta3* | ENSMUSG00000025934 | 1:21230670-21231169 | 128 | 406344 | 179 | 480189 | 92 | 456338 | 3,59E-03 | 5,86E-06 | 3,15E-04 | 3,73E-04 | 2,02E-04 | | 6,00E-04 | 4,46E-04 | 4,92E-04 | 5,72E-04 | 4,53E-04 | 4,70E-04 | 5,04E-04 | 5,01E-04 | 4,19E-04 | |
| *H2afx* | ENSMUSG00000049932 | 9:44142798-44143297 | 523 | 663628 | 531 | 716953 | 118 | 697873 | 9,33E-65 | 1,07E-59 | 7,88E-04 | 7,41E-04 | 1,69E-04 | | 3,86E-04 | 3,84E-04 | 3,60E-04 | 3,78E-04 | 4,71E-04 | 4,38E-04 | 7,85E-04 | 7,24E-04 | 3,51E-04 | |
| *H3f3b* | ENSMUSG00000016559 | 11:115885319-115885818 | 209 | 768881 | 252 | 840753 | 116 | 787845 | 4,42E-07 | 5,42E-10 | 2,72E-04 | 3,00E-04 | 1,47E-04 | | 3,67E-04 | 3,18E-04 | 4,11E-04 | 4,66E-04 | 3,44E-04 | 3,30E-04 | 4,33E-04 | 4,65E-04 | 3,68E-04 | |
| *H47* | ENSMUSG00000075701 | 7:73224535-73225034 | 219 | 865446 | 279 | 939021 | 118 | 906324 | 2,04E-08 | 5,67E-14 | 2,53E-04 | 2,97E-04 | 1,30E-04 | | 3,45E-04 | 3,19E-04 | 3,35E-04 | 3,59E-04 | 3,23E-04 | 2,93E-04 | 4,00E-04 | 4,22E-04 | 2,92E-04 | |
| *Hdac1* | ENSMUSG00000028800 | 4:129219391-129219890 | 159 | 754440 | 180 | 805875 | 111 | 797257 | 2,66E-03 | 3,31E-04 | 2,11E-04 | 2,23E-04 | 1,39E-04 | | 3,69E-04 | 3,20E-04 | 3,72E-04 | 4,24E-04 | 3,98E-04 | 3,72E-04 | 4,09E-04 | 4,09E-04 | 3,26E-04 | |
| *Hdac9* | ENSMUSG00000004698 | 12:35213255-35213754 | 118 | 540256 | 139 | 587555 | 69 | 507953 | 5,03E-03 | 5,32E-04 | 2,18E-04 | 2,37E-04 | 1,36E-04 | | 5,75E-04 | 5,70E-04 | 5,26E-04 | 5,96E-04 | 5,46E-04 | 5,07E-04 | 4,92E-04 | 5,29E-04 | 4,78E-04 | |
| *Hdgf* | ENSMUSG00000004897 | 3:87710243-87710742 | 222 | 680627 | 193 | 751993 | 106 | 691046 | 3,79E-10 | 7,45E-05 | 3,26E-04 | 2,57E-04 | 1,53E-04 | | 3,73E-04 | 3,38E-04 | 3,73E-04 | 3,95E-04 | 3,53E-04 | 3,45E-04 | 4,45E-04 | 3,91E-04 | 3,47E-04 | |
| *Hells* | ENSMUSG00000025001 | 19:39005480-39005979 | 168 | 557458 | 221 | 605650 | 104 | 603496 | 3,14E-05 | 6,55E-10 | 3,01E-04 | 3,65E-04 | 1,72E-04 | | 3,55E-04 | 3,62E-04 | 3,68E-04 | 4,17E-04 | 3,54E-04 | 3,72E-04 | 4,21E-04 | 4,71E-04 | 3,56E-04 | |
| *Hist1h1a* | ENSMUSG00000049539 | 13:23855537-23856036 | 323 | 452994 | 440 | 491598 | 66 | 480903 | 7,78E-44 | 1,11E-65 | 7,13E-04 | 8,95E-04 | 1,37E-04 | | 4,63E-04 | 3,37E-04 | 3,33E-04 | 3,59E-04 | 3,56E-04 | 3,59E-04 | 6,84E-04 | 7,91E-04 | 3,02E-04 | |
| *Hist1h1b* | ENSMUSG00000058773 | 13:21871924-21872423 | 576 | 387884 | 735 | 449949 | 63 | 426862 | 6,12E-114 | 6,37E-137 | 1,48E-03 | 1,63E-03 | 1,48E-04 | | 3,49E-04 | 3,69E-04 | 4,25E-04 | 4,17E-04 | 4,13E-04 | 4,47E-04 | 1,15E-03 | 1,22E-03 | 3,33E-04 | |
| *Hnrnpa2b1-1* | ENSMUSG00000004980 | 6:51419394-51419893 | 461 | 842176 | 382 | 944784 | 177 | 988573 | 1,24E-39 | 2,19E-19 | 5,47E-04 | 4,04E-04 | 1,79E-04 | | 3,83E-04 | 3,59E-04 | 3,67E-04 | 3,91E-04 | 3,98E-04 | 3,62E-04 | 5,99E-04 | 4,92E-04 | 3,39E-04 | |
| *Hnrnpf-2* | ENSMUSG00000042079 | 6:117856800-117857299 | 251 | 674985 | 271 | 724519 | 83 | 720551 | 3,83E-22 | 4,24E-23 | 3,72E-04 | 3,74E-04 | 1,15E-04 | | 3,96E-04 | 3,30E-04 | 3,61E-04 | 4,53E-04 | 3,80E-04 | 3,49E-04 | 5,29E-04 | 4,93E-04 | 3,20E-04 | |
| *Hnrnpf-3* | ENSMUSG00000042079 | 6:117857822-117858321 | 334 | 935765 | 479 | 1007257 | 159 | 964305 | 1,51E-15 | 5,09E-34 | 3,57E-04 | 4,76E-04 | 1,65E-04 | | 3,71E-04 | 3,47E-04 | 3,98E-04 | 3,77E-04 | 4,07E-04 | 3,52E-04 | 4,83E-04 | 5,30E-04 | 3,33E-04 | |
| *Hnrnpf-4* | ENSMUSG00000042079 | 6:117867312-117867811 | 165 | 477813 | 249 | 521172 | 71 | 508346 | 2,73E-10 | 2,65E-22 | 3,45E-04 | 4,78E-04 | 1,40E-04 | | 4,41E-04 | 2,83E-04 | 3,01E-04 | 3,54E-04 | 3,36E-04 | 3,27E-04 | 4,44E-04 | 5,20E-04 | 3,10E-04 | |
| *Hsf4* | ENSMUSG00000033249 | 8:107793774-107794273 | 144 | 826674 | 146 | 932286 | 98 | 886188 | 1,77E-03 | 2,11E-02 | 1,74E-04 | 1,57E-04 | 1,11E-04 | | 3,01E-04 | 3,32E-04 | 3,53E-04 | 3,32E-04 | 3,09E-04 | 3,10E-04 | 3,72E-04 | 3,41E-04 | 3,08E-04 | |
| *Hvcn1-1* | ENSMUSG00000064267 | 5:122659746-122660245 | 142 | 717436 | 164 | 793379 | 104 | 752454 | 1,40E-02 | 4,23E-03 | 1,98E-04 | 2,07E-04 | 1,38E-04 | | 3,61E-04 | 3,18E-04 | 3,32E-04 | 3,89E-04 | 3,62E-04 | 3,00E-04 | 3,72E-04 | 3,97E-04 | 3,27E-04 | |
| *Id3* | ENSMUSG00000007872 | 4:135699737-135700236 | 271 | 878340 | 280 | 962896 | 139 | 943036 | 3,58E-12 | 1,57E-10 | 3,09E-04 | 2,91E-04 | 1,47E-04 | | 3,50E-04 | 3,28E-04 | 4,47E-04 | 4,37E-04 | 3,68E-04 | 3,29E-04 | 4,54E-04 | 4,20E-04 | 3,00E-04 | |
| *Igj* | ENSMUSG00000067149 | 5:88956423-88956922 | 239 | 344459 | 344 | 408061 | 162 | 357676 | 1,18E-04 | 1,78E-10 | 6,94E-04 | 8,43E-04 | 4,53E-04 | | 3,78E-04 | 5,49E-04 | 5,93E-04 | 5,79E-04 | 6,26E-04 | 6,64E-04 | 7,31E-04 | 8,27E-04 | 4,79E-04 | |
| *Ikzf1* | ENSMUSG00000018654 | 11:11586216-11586715 | 561 | 644630 | 607 | 719160 | 89 | 695595 | 5,50E-91 | 1,99E-90 | 8,70E-04 | 8,44E-04 | 1,28E-04 | | 3,63E-04 | 3,36E-04 | 3,55E-04 | 3,33E-04 | 3,43E-04 | 3,43E-04 | 7,22E-04 | 6,89E-04 | 3,24E-04 | |
| *Il21r* | ENSMUSG00000030745 | 7:132746943-132747442 | 501 | 654441 | 543 | 723059 | 104 | 691656 | 1,32E-66 | 5,66E-67 | 7,66E-04 | 7,51E-04 | 1,50E-04 | | 5,37E-04 | 3,29E-04 | 4,17E-04 | 3,51E-04 | 3,99E-04 | 4,45E-04 | 7,12E-04 | 6,72E-04 | 3,48E-04 | |
| *Il4ra* | ENSMUSG00000030748 | 7:132695796-132696295 | 173 | 681497 | 386 | 764757 | 114 | 716952 | 4,21E-04 | 1,02E-30 | 2,54E-04 | 5,05E-04 | 1,59E-04 | | 3,97E-04 | 3,00E-04 | 3,10E-04 | 3,47E-04 | 3,46E-04 | 3,14E-04 | 4,32E-04 | 5,56E-04 | 3,28E-04 | |
| *Ildr2* | ENSMUSG00000040612 | 1:168184270-168184769 | 95 | 282593 | 107 | 303801 | 24 | 310856 | 7,82E-12 | 3,42E-13 | 3,36E-04 | 3,52E-04 | 7,72E-05 | | 1,15E-03 | 8,72E-04 | 2,53E-03 | 1,25E-03 | 1,45E-03 | 2,51E-03 | 1,30E-03 | 1,33E-03 | 3,19E-04 | |
| *Ipmk* | ENSMUSG00000060733 | 10:70810541-70811040 | 110 | 541434 | 129 | 590055 | 74 | 604134 | 2,58E-03 | 2,56E-04 | 2,03E-04 | 2,19E-04 | 1,22E-04 | | 3,55E-04 | 3,48E-04 | 3,97E-04 | 4,10E-04 | 3,36E-04 | 3,49E-04 | 3,89E-04 | 3,76E-04 | 3,12E-04 | |
| *Ipo7* | ENSMUSG00000066232 | 7:117161939-117162438 | 165 | 775158 | 182 | 839214 | 120 | 792026 | 1,22E-02 | 6,89E-03 | 2,13E-04 | 2,17E-04 | 1,52E-04 | | 3,72E-04 | 3,25E-04 | 3,22E-04 | 3,59E-04 | 3,56E-04 | 3,57E-04 | 3,78E-04 | 3,60E-04 | 3,35E-04 | |
| *Irf4* | ENSMUSG00000021356 | 13:30841127-30841626 | 157 | 458687 | 143 | 527486 | 68 | 507831 | 1,50E-10 | 5,28E-06 | 3,42E-04 | 2,71E-04 | 1,34E-04 | | 3,69E-04 | 3,20E-04 | 3,57E-04 | 3,91E-04 | 4,17E-04 | 3,69E-04 | 4,69E-04 | 3,94E-04 | 2,73E-04 | |
| *Irf8* | ENSMUSG00000041515 | 8:123260276-123260775 | 328 | 550422 | 456 | 601283 | 66 | 564248 | 5,67E-43 | 4,99E-66 | 5,96E-04 | 7,58E-04 | 1,17E-04 | | 3,76E-04 | 3,38E-04 | 3,76E-04 | 3,81E-04 | 3,40E-04 | 3,44E-04 | 6,53E-04 | 6,73E-04 | 3,38E-04 | |
| *Itga4* | ENSMUSG00000027009 | 2:79095583-79096082 | 164 | 740245 | 237 | 815845 | 101 | 788194 | 6,23E-05 | 7,88E-12 | 2,22E-04 | 2,90E-04 | 1,28E-04 | | 3,85E-04 | 3,23E-04 | 3,66E-04 | 3,58E-04 | 5,35E-04 | 3,61E-04 | 4,20E-04 | 4,04E-04 | 3,19E-04 | |
| *Itgal* | ENSMUSG00000030830 | 7:134439862-134440361 | 104 | 563705 | 138 | 620692 | 72 | 604187 | 1,19E-02 | 6,28E-05 | 1,84E-04 | 2,22E-04 | 1,19E-04 | | 3,70E-04 | 2,94E-04 | 3,41E-04 | 3,68E-04 | 3,14E-04 | 3,24E-04 | 3,63E-04 | 3,43E-04 | 3,22E-04 | |
| *Itgb2* | ENSMUSG00000000290 | 10:76993093-76993592 | 128 | 536106 | 127 | 595348 | 69 | 580511 | 1,10E-05 | 3,19E-04 | 2,39E-04 | 2,13E-04 | 1,19E-04 | | 3,41E-04 | 3,00E-04 | 3,49E-04 | 4,11E-04 | 3,41E-04 | 3,28E-04 | 3,80E-04 | 3,95E-04 | 3,18E-04 | |
| *Jh4* | NoCode003 | 12:114666194-114666736 | 11619 | 822724 | 14950 | 920989 | 97 | 948760 | 0,00E+00 | 0,00E+00 | 1,41E-02 | 1,62E-02 | 1,02E-04 | | 5,34E-03 | 3,65E-03 | 2,09E-03 | 1,78E-03 | 2,77E-03 | 3,25E-03 | 6,47E-03 | 7,59E-03 | 2,91E-04 | |
| *Junb* | ENSMUSG00000052837 | 8:87502118-87502617 | 311 | 737087 | 329 | 796520 | 112 | 752693 | 1,45E-22 | 3,61E-22 | 4,22E-04 | 4,13E-04 | 1,49E-04 | | 3,94E-04 | 3,21E-04 | 3,60E-04 | 3,96E-04 | 3,71E-04 | 4,63E-04 | 5,53E-04 | 4,76E-04 | 3,62E-04 | |
| *Kpna2* | ENSMUSG00000018362 | 11:106860340-106860839 | 244 | 519174 | 316 | 561836 | 96 | 546553 | 1,05E-16 | 4,31E-26 | 4,70E-04 | 5,62E-04 | 1,76E-04 | | 7,87E-04 | 7,45E-04 | 7,40E-04 | 4,73E-04 | 6,43E-04 | 4,72E-04 | 8,67E-04 | 1,00E-03 | 4,37E-04 | |

| | | | Transitions at C/G pairs | | | | | | FDR | | C/G Transition Frequency | | | Total mutation frequency | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Ung$^{-/-}$Msh2$^{-/-}$Exp1 | | Ung$^{-/-}$Msh2$^{-/-}$Exp2 | | *Aicda*$^{-/-}$ | | | | Ung$^{-/-}$Msh2$^{-/-}$ | | *Aicda*$^{-/-}$ | Ung$^{+/-}$Msh2$^{+/-}$ | | Ung$^{+/-}$Msh2$^{-/-}$ | | Ung$^{-/-}$Msh2$^{+/-}$ | | Ung$^{-/-}$Msh2$^{-/-}$ | | *Aicda*$^{-/-}$ |
| **Gene** | **Ensembl ID** | **Coordinates** | **C/G transitions** | **C/G sequenced** | **C/G transitions** | **C/G sequenced** | **C/G transitions** | **C/G sequenced** | **Exp1** | **Exp2** | **Exp1** | **Exp2** | **Exp1** | **Exp1** | **Exp2** | **Exp1** | **Exp2** | **Exp1** | **Exp2** | **Exp1** | **Exp2** | **Exp1** |
| *Kpnb1* | ENSMUSG00000001440 | 11:97048707-97049206 | 55 | 181493 | 86 | 198709 | 22 | 178540 | 9,57E-04 | 6,24E-08 | 3,03E-04 | 4,33E-04 | 1,23E-04 | 4,96E-04 | 4,11E-04 | 4,86E-04 | 4,05E-04 | 3,97E-04 | 5,21E-04 | 5,11E-04 | 6,78E-04 | 4,26E-04 |
| *Lmo1* | ENSMUSG00000036111 | 7:116313323-116313822 | 136 | 720717 | 160 | 822820 | 100 | 802471 | 5,05E-03 | 1,62E-03 | 1,89E-04 | 1,94E-04 | 1,25E-04 | 3,36E-04 | 3,23E-04 | 3,56E-04 | 3,81E-04 | 3,72E-04 | 3,22E-04 | 3,96E-04 | 3,65E-04 | 3,39E-04 |
| *Lrmp* | ENSMUSG00000030263 | 6:145070259-145070758 | 199 | 696203 | 261 | 775233 | 94 | 684396 | 1,22E-08 | 4,23E-14 | 2,86E-04 | 3,37E-04 | 1,37E-04 | 3,46E-04 | 4,97E-04 | 6,54E-04 | 3,42E-04 | 4,14E-04 | 3,58E-04 | 7,02E-04 | 4,52E-04 | 2,45E-03 |
| *Lsp1-1* | ENSMUSG00000018819 | 7:149646775-149647274 | 125 | 628466 | 134 | 716411 | 85 | 672175 | 4,01E-03 | 1,37E-02 | 1,99E-04 | 1,87E-04 | 1,26E-04 | 1,29E-03 | 7,00E-04 | 1,43E-03 | 9,88E-04 | 1,70E-03 | 1,61E-03 | 1,57E-03 | 1,64E-03 | 2,29E-03 |
| *Ltb* | ENSMUSG00000024399 | 17:35331452-35331951 | 219 | 634910 | 501 | 711874 | 85 | 662443 | 4,38E-15 | 5,72E-65 | 3,45E-04 | 7,04E-04 | 1,28E-04 | 4,23E-04 | 3,40E-04 | 3,61E-04 | 3,85E-04 | 3,78E-04 | 3,39E-04 | 4,42E-04 | 6,49E-04 | 3,48E-04 |
| *Ly6e-1* | ENSMUSG00000022587 | 15:74785481-74785980 | 284 | 644773 | 661 | 730789 | 100 | 690693 | 2,83E-23 | 6,66E-94 | 4,40E-04 | 9,05E-04 | 1,45E-04 | 3,89E-04 | 3,50E-04 | 3,69E-04 | 3,77E-04 | 4,08E-04 | 3,26E-04 | 5,88E-04 | 8,14E-04 | 4,00E-04 |
| *Ly6e-2* | ENSMUSG00000022587 | 15:74785501-74786000 | 287 | 648185 | 658 | 734202 | 103 | 692343 | 7,94E-23 | 2,55E-91 | 4,43E-04 | 8,96E-04 | 1,49E-04 | 3,92E-04 | 3,52E-04 | 3,80E-04 | 3,78E-04 | 4,13E-04 | 3,28E-04 | 5,85E-04 | 8,11E-04 | 3,98E-04 |
| *Ly6e-3* | ENSMUSG00000022587 | 15:74785536-74786035 | 271 | 643424 | 480 | 728497 | 105 | 684529 | 2,70E-19 | 5,58E-52 | 4,21E-04 | 6,59E-04 | 1,53E-04 | 3,67E-04 | 3,34E-04 | 3,66E-04 | 3,76E-04 | 3,98E-04 | 3,18E-04 | 5,37E-04 | 6,53E-04 | 3,48E-04 |
| *Lyn* | ENSMUSG00000042228 | 4:3605268-3605767 | 205 | 837139 | 291 | 913231 | 117 | 850871 | 2,51E-06 | 1,19E-14 | 2,45E-04 | 3,19E-04 | 1,38E-04 | 3,51E-04 | 3,23E-04 | 3,69E-04 | 3,75E-04 | 3,63E-04 | 3,16E-04 | 4,11E-04 | 4,60E-04 | 3,28E-04 |
| *Man1a* | ENSMUSG00000003746 | 10:53795103-53795602 | 371 | 734648 | 437 | 819639 | 112 | 784848 | 4,89E-36 | 4,63E-42 | 5,05E-04 | 5,33E-04 | 1,43E-04 | 3,83E-04 | 3,65E-04 | 3,74E-04 | 3,99E-04 | 3,77E-04 | 4,38E-04 | 5,94E-04 | 5,90E-04 | 3,44E-04 |
| *Mcm2* | ENSMUSG00000002870 | 6:88848275-88848774 | 324 | 563045 | 476 | 597167 | 88 | 574445 | 1,19E-32 | 8,59E-61 | 5,75E-04 | 7,97E-04 | 1,53E-04 | 4,26E-04 | 2,87E-04 | 3,43E-04 | 3,67E-04 | 3,65E-04 | 3,54E-04 | 5,97E-04 | 7,02E-04 | 3,24E-04 |
| *Mcm6* | ENSMUSG00000026355 | 1:130255734-130256233 | 251 | 751486 | 415 | 818149 | 119 | 796013 | 8,99E-13 | 1,97E-36 | 3,34E-04 | 5,07E-04 | 1,49E-04 | 1,24E-03 | 1,05E-03 | 5,92E-04 | 1,16E-03 | 9,89E-04 | 1,09E-03 | 9,37E-04 | 1,34E-03 | 1,95E-03 |
| *Mcm7* | ENSMUSG00000029730 | 5:138612591-138613090 | 182 | 665798 | 211 | 713331 | 107 | 689303 | 1,36E-05 | 1,86E-07 | 2,73E-04 | 2,96E-04 | 1,55E-04 | 4,60E-04 | 4,19E-04 | 4,30E-04 | 5,19E-04 | 4,34E-04 | 4,35E-04 | 4,62E-04 | 4,96E-04 | 4,27E-04 |
| *Mdfi-2* | ENSMUSG00000032717 | 17:47971141-47971640 | 165 | 892383 | 184 | 944059 | 120 | 930634 | 7,70E-03 | 1,50E-03 | 1,85E-04 | 1,95E-04 | 1,29E-04 | 3,37E-04 | 3,33E-04 | 4,09E-04 | 3,74E-04 | 3,56E-04 | 3,15E-04 | 3,62E-04 | 3,51E-04 | 3,17E-04 |
| *Mdh2* | ENSMUSG00000019179 | 5:136254519-136255018 | 158 | 750847 | 155 | 841360 | 105 | 783547 | 1,19E-03 | 3,00E-02 | 2,10E-04 | 1,84E-04 | 1,34E-04 | 3,53E-04 | 3,17E-04 | 3,36E-04 | 3,79E-04 | 3,59E-04 | 3,27E-04 | 3,84E-04 | 3,56E-04 | 3,32E-04 |
| *Mef2b* | ENSMUSG00000079033 | 8:72676677-72677176 | 105 | 215355 | 155 | 247435 | 26 | 224352 | 2,89E-12 | 7,37E-20 | 4,88E-04 | 6,26E-04 | 1,16E-04 | 4,34E-04 | 2,84E-04 | 3,12E-04 | 3,51E-04 | 3,38E-04 | 3,09E-04 | 5,15E-04 | 6,69E-04 | 3,06E-04 |
| *Mir142* | ENSMUSG00000065420 | 11:87570366-87570865 | 828 | 600466 | 1212 | 671624 | 93 | 658987 | 1,55E-161 | 1,29E-242 | 1,38E-03 | 1,80E-03 | 1,41E-04 | 4,84E-04 | 3,81E-04 | 4,24E-04 | 3,72E-04 | 4,76E-04 | 4,15E-04 | 1,06E-03 | 1,30E-03 | 3,31E-04 |
| *Mlh3* | ENSMUSG00000021245 | 12:86611015-86611514 | 161 | 923662 | 188 | 1019032 | 124 | 990252 | 1,45E-02 | 2,65E-03 | 1,74E-04 | 1,84E-04 | 1,25E-04 | 3,39E-04 | 3,33E-04 | 3,31E-04 | 3,99E-04 | 3,48E-04 | 3,20E-04 | 3,76E-04 | 3,46E-04 | 3,13E-04 |
| *Mlst8* | ENSMUSG00000024142 | 17:24615524-24616023 | 230 | 843924 | 290 | 909513 | 118 | 835142 | 2,07E-08 | 8,11E-14 | 2,73E-04 | 3,19E-04 | 1,41E-04 | 3,62E-04 | 3,06E-04 | 3,65E-04 | 3,96E-04 | 3,97E-04 | 3,14E-04 | 4,23E-04 | 4,40E-04 | 3,39E-04 |
| *Mrpl51* | ENSMUSG00000030335 | 6:125142218-125142717 | 132 | 659742 | 142 | 715217 | 102 | 702555 | 3,37E-02 | 3,79E-02 | 2,00E-04 | 1,99E-04 | 1,45E-04 | 3,12E-04 | 2,96E-04 | 3,56E-04 | 3,46E-04 | 3,31E-04 | 2,91E-04 | 3,85E-04 | 3,69E-04 | 3,44E-04 |
| *Ms4a1* | ENSMUSG00000024673 | 19:11340142-11340641 | 599 | 908541 | 833 | 1020544 | 123 | 976298 | 5,63E-82 | 8,72E-122 | 6,59E-04 | 8,16E-04 | 1,26E-04 | 3,51E-04 | 3,16E-04 | 3,29E-04 | 3,47E-04 | 3,36E-04 | 3,23E-04 | 5,66E-04 | 6,59E-04 | 3,07E-04 |
| *Msh2* | ENSMUSG00000024151 | 17:88071897-88072396 | 137 | 510486 | 118 | 558651 | 81 | 563898 | 3,29E-05 | 2,11E-02 | 2,68E-04 | 2,11E-04 | 1,44E-04 | 3,66E-04 | 3,50E-04 | 3,82E-04 | 3,95E-04 | 3,63E-04 | 2,97E-04 | 4,06E-04 | 3,39E-04 | 3,12E-04 |
| *Msh5-1* | ENSMUSG00000007035 | 17:35183052-35183551 | 182 | 629303 | 156 | 681009 | 85 | 635597 | 1,07E-08 | 2,28E-04 | 2,89E-04 | 2,29E-04 | 1,34E-04 | 3,58E-04 | 3,16E-04 | 3,53E-04 | 4,01E-04 | 3,47E-04 | 2,79E-04 | 4,28E-04 | 3,59E-04 | 3,15E-04 |
| *Msh5-2* | ENSMUSG00000007035 | 17:35183169-35183668 | 126 | 521555 | 118 | 567992 | 71 | 523734 | 3,71E-04 | 1,30E-02 | 2,42E-04 | 2,08E-04 | 1,36E-04 | 3,57E-04 | 3,51E-04 | 3,77E-04 | 4,04E-04 | 3,68E-04 | 2,95E-04 | 4,16E-04 | 3,63E-04 | 3,27E-04 |
| *Msh6* | ENSMUSG00000005370 | 17:88374390-88374889 | 190 | 764672 | 184 | 817788 | 128 | 750411 | 3,14E-03 | 3,79E-02 | 2,48E-04 | 2,25E-04 | 1,71E-04 | 3,97E-04 | 3,29E-04 | 3,47E-04 | 3,89E-04 | 3,83E-04 | 3,27E-04 | 4,25E-04 | 3,80E-04 | 3,51E-04 |
| *Mtor* | ENSMUSG00000028991 | 4:147822691-147823190 | 97 | 499716 | 134 | 557502 | 63 | 529821 | 7,27E-03 | 1,27E-05 | 1,94E-04 | 2,40E-04 | 1,19E-04 | 3,71E-04 | 3,10E-04 | 3,16E-04 | 3,55E-04 | 3,73E-04 | 3,42E-04 | 4,02E-04 | 4,06E-04 | 3,15E-04 |
| *Mtss1* | ENSMUSG00000022353 | 15:58913082-58913581 | 216 | 738806 | 316 | 819654 | 130 | 795017 | 6,87E-07 | 9,62E-17 | 2,92E-04 | 3,86E-04 | 1,64E-04 | 3,48E-04 | 3,23E-04 | 3,74E-04 | 3,73E-04 | 3,78E-04 | 3,76E-04 | 4,20E-04 | 4,84E-04 | 3,61E-04 |
| *Mybbp1a* | ENSMUSG00000040463 | 11:72254880-72255379 | 141 | 555854 | 154 | 614681 | 95 | 597996 | 1,47E-03 | 1,61E-03 | 2,54E-04 | 2,51E-04 | 1,59E-04 | 3,85E-04 | 3,09E-04 | 3,80E-04 | 3,82E-04 | 3,45E-04 | 3,55E-04 | 3,88E-04 | 4,54E-04 | 3,46E-04 |
| *Myc* | ENSMUSG00000022346 | 15:61816896-61817395 | 362 | 480006 | 459 | 522364 | 80 | 490714 | 6,61E-44 | 2,95E-59 | 7,54E-04 | 8,79E-04 | 1,63E-04 | 4,00E-04 | 3,39E-04 | 3,89E-04 | 4,39E-04 | 4,61E-04 | 3,76E-04 | 7,59E-04 | 7,97E-04 | 3,36E-04 |
| *Mycbp2* | ENSMUSG00000033004 | 14:103745518-103746017 | 42 | 159683 | 46 | 164458 | 14 | 161191 | 7,02E-04 | 2,20E-04 | 2,63E-04 | 2,80E-04 | 8,69E-05 | 5,32E-04 | 3,28E-04 | 3,19E-04 | 3,87E-04 | 3,56E-04 | 3,13E-04 | 4,12E-04 | 4,58E-04 | 2,79E-04 |
| *Ncl* | ENSMUSG00000026234 | 1:88255531-88256030 | 287 | 983500 | 284 | 1034488 | 141 | 973071 | 2,46E-11 | 1,61E-09 | 2,92E-04 | 2,75E-04 | 1,45E-04 | 3,51E-04 | 3,14E-04 | 3,47E-04 | 3,59E-04 | 3,77E-04 | 3,60E-04 | 4,58E-04 | 4,28E-04 | 3,41E-04 |
| *Nfkb2-1* | ENSMUSG00000025225 | 19:46379227-46379726 | 148 | 747171 | 164 | 821171 | 116 | 777691 | 4,49E-02 | 3,79E-02 | 1,98E-04 | 2,00E-04 | 1,49E-04 | 3,64E-04 | 3,25E-04 | 3,19E-04 | 3,87E-04 | 3,80E-04 | 3,14E-04 | 3,78E-04 | 3,62E-04 | 3,21E-04 |
| *Nfkb2-2* | ENSMUSG00000025225 | 19:46380185-46380684 | 120 | 560027 | 121 | 620957 | 66 | 547790 | 5,96E-04 | 5,04E-03 | 2,14E-04 | 1,95E-04 | 1,20E-04 | 3,76E-04 | 3,42E-04 | 3,66E-04 | 4,00E-04 | 3,77E-04 | 3,53E-04 | 4,05E-04 | 3,85E-04 | 3,00E-04 |
| *Nfkb2-3* | ENSMUSG00000025225 | 19:46380420-46380919 | 273 | 621600 | 151 | 696165 | 73 | 653745 | 2,18E-29 | 1,04E-05 | 4,39E-04 | 2,17E-04 | 1,12E-04 | 3,31E-04 | 3,23E-04 | 3,56E-04 | 3,60E-04 | 3,68E-04 | 3,41E-04 | 5,29E-04 | 4,03E-04 | 3,20E-04 |
| *Ntan1* | ENSMUSG00000022681 | 16:13819370-13819869 | 134 | 679722 | 129 | 758918 | 89 | 723768 | 1,87E-03 | 4,36E-02 | 1,97E-04 | 1,70E-04 | 1,23E-04 | 3,82E-04 | 2,85E-04 | 3,47E-04 | 3,90E-04 | 3,62E-04 | 3,49E-04 | 3,93E-04 | 3,28E-04 | 3,15E-04 |
| *Parp1* | ENSMUSG00000026496 | 1:182499106-182499605 | 151 | 707486 | 150 | 771792 | 91 | 742589 | 1,17E-04 | 1,72E-03 | 2,13E-04 | 1,94E-04 | 1,23E-04 | 3,34E-04 | 3,10E-04 | 3,87E-04 | 3,94E-04 | 3,30E-04 | 3,38E-04 | 4,01E-04 | 3,66E-04 | 3,10E-04 |
| *Pax5* | ENSMUSG00000014030 | 4:44722813-44723312 | 554 | 503476 | 682 | 551020 | 64 | 515254 | 3,64E-99 | 5,88E-121 | 1,10E-03 | 1,24E-03 | 1,24E-04 | 5,42E-04 | 4,87E-04 | 5,00E-04 | 5,41E-04 | 4,88E-04 | 4,35E-04 | 8,81E-04 | 9,36E-04 | 4,20E-04 |
| *Pcna* | ENSMUSG00000027342 | 2:132078417-132078916 | 146 | 617312 | 162 | 677457 | 89 | 656611 | 1,41E-04 | 6,24E-05 | 2,37E-04 | 2,39E-04 | 1,36E-04 | 3,42E-04 | 2,88E-04 | 3,17E-04 | 3,66E-04 | 3,36E-04 | 3,25E-04 | 4,16E-04 | 3,58E-04 | 3,13E-04 |
| *Pex13* | ENSMUSG00000020283 | 11:23565436-23565935 | 187 | 684392 | 158 | 749783 | 106 | 697002 | 5,64E-06 | 2,48E-02 | 2,73E-04 | 2,11E-04 | 1,52E-04 | 3,85E-04 | 3,91E-04 | 3,79E-04 | 3,87E-04 | 3,62E-04 | 3,18E-04 | 4,51E-04 | 4,10E-04 | 3,33E-04 |
| *Phip* | ENSMUSG00000032253 | 9:82868597-82869096 | 176 | 494894 | 129 | 511028 | 63 | 495442 | 1,16E-12 | 2,67E-05 | 3,56E-04 | 2,52E-04 | 1,27E-04 | 3,60E-04 | 3,33E-04 | 3,34E-04 | 3,72E-04 | 3,70E-04 | 3,35E-04 | 4,89E-04 | 4,06E-04 | 3,50E-04 |
| *Pias1* | ENSMUSG00000032405 | 9:62828187-62828686 | 155 | 556985 | 282 | 612404 | 76 | 599841 | 5,35E-08 | 6,11E-27 | 2,78E-04 | 4,60E-04 | 1,27E-04 | 8,75E-04 | 8,13E-04 | 7,72E-04 | 7,10E-04 | 7,04E-04 | 7,91E-04 | 7,13E-04 | 7,31E-04 | 6,68E-04 |
| *Pick1-2* | ENSMUSG00000068206 | 15:79059812-79060311 | 130 | 873283 | 167 | 967303 | 100 | 912093 | 4,41E-02 | 1,15E-03 | 1,49E-04 | 1,73E-04 | 1,10E-04 | 3,46E-04 | 3,03E-04 | 3,18E-04 | 3,79E-04 | 3,48E-04 | 3,13E-04 | 3,46E-04 | 3,42E-04 | 3,13E-04 |
| *Pik3ap1* | ENSMUSG00000025017 | 19:41459061-41459560 | 192 | 797382 | 167 | 882959 | 96 | 797313 | 9,03E-08 | 1,41E-03 | 2,41E-04 | 1,89E-04 | 1,20E-04 | 2,84E-03 | 3,30E-03 | 2,18E-03 | 3,31E-03 | 1,63E-03 | 2,93E-03 | 2,94E-03 | 3,05E-03 | 5,88E-03 |
| *Pim1* | ENSMUSG00000024014 | 17:29627990-29628489 | 2145 | 614306 | 2751 | 657842 | 98 | 642524 | 0,00E+00 | 0,00E+00 | 3,49E-03 | 4,18E-03 | 1,53E-04 | 6,56E-04 | 6,14E-04 | 7,08E-04 | 5,19E-04 | 7,42E-04 | 8,24E-04 | 2,71E-03 | 3,15E-03 | 3,36E-04 |
| *Plac8* | ENSMUSG00000029322 | 5:101000725-101001224 | 81 | 340658 | 136 | 383491 | 50 | 361198 | 7,70E-03 | 1,63E-08 | 2,38E-04 | 3,55E-04 | 1,38E-04 | 3,65E-04 | 3,10E-04 | 3,43E-04 | 4,04E-04 | 3,41E-04 | 3,13E-04 | 3,88E-04 | 4,36E-04 | 3,42E-04 |
| *Plcb2* | ENSMUSG00000040061 | 2:118553675-118554174 | 149 | 910849 | 170 | 1020135 | 103 | 919968 | 8,67E-03 | 4,49E-03 | 1,64E-04 | 1,67E-04 | 1,12E-04 | 2,38E-03 | 1,64E-03 | 2,11E-03 | 3,01E-03 | 3,38E-03 | 1,15E-03 | 1,77E-03 | 1,24E-03 | 5,42E-03 |
| *Pms2* | ENSMUSG00000079109 | 5:144670870-144671369 | 189 | 709817 | 157 | 798404 | 101 | 774062 | 2,36E-08 | 4,20E-03 | 2,66E-04 | 1,97E-04 | 1,30E-04 | 3,47E-04 | 2,81E-04 | 3,82E-04 | 3,28E-04 | 3,32E-04 | 3,15E-04 | 4,00E-04 | 3,37E-04 | 3,02E-04 |
| *Pola1* | ENSMUSG00000006678 | X:90876995-90877494 | 282 | 580875 | 500 | 549551 | 66 | 571667 | 5,31E-31 | 8,85E-86 | 4,85E-04 | 9,10E-04 | 1,15E-04 | 3,54E-04 | 3,29E-04 | 3,44E-04 | 3,95E-04 | 3,17E-04 | 3,67E-04 | 5,34E-04 | 7,57E-04 | 3,49E-04 |
| *Pold1* | ENSMUSG00000038644 | 7:51803686-51804185 | 339 | 702046 | 470 | 793770 | 110 | 757641 | 5,45E-31 | 4,51E-49 | 4,83E-04 | 5,92E-04 | 1,45E-04 | 3,55E-04 | 3,17E-04 | 3,80E-04 | 4,20E-04 | 3,72E-04 | 3,23E-04 | 5,72E-04 | 6,31E-04 | 3,64E-04 |
| *Pold4* | ENSMUSG00000024854 | 19:4231899-4232398 | 169 | 610086 | 201 | 700534 | 75 | 640097 | 1,01E-09 | 3,33E-11 | 2,77E-04 | 2,87E-04 | 1,17E-04 | 4,42E-04 | 4,16E-04 | 4,96E-04 | 5,24E-04 | 4,41E-04 | 4,52E-04 | 5,42E-04 | 5,26E-04 | 4,55E-04 |
| *Pou2af1* | ENSMUSG00000032053 | 9:51021795-51022294 | 1438 | 581980 | 1399 | 657671 | 83 | 650248 | 0,00E+00 | 6,93E-304 | 2,47E-03 | 2,13E-03 | 1,28E-04 | 4,19E-04 | 3,91E-04 | 4,35E-04 | 3,79E-04 | 4,42E-04 | 3,62E-04 | 1,53E-03 | 1,34E-03 | 3,27E-04 |
| *Pou2f1-1* | ENSMUSG00000026565 | 1:167864532-167865031 | 99 | 492890 | 97 | 575689 | 63 | 550754 | 1,63E-03 | 4,00E-02 | 2,01E-04 | 1,68E-04 | 1,14E-04 | 3,13E-04 | 2,87E-04 | 3,10E-04 | 3,44E-04 | 2,99E-04 | 3,26E-04 | 3,67E-04 | 3,08E-04 | 3,27E-04 |
| *Pou2f2* | ENSMUSG00000008496 | 7:25916980-25917479 | 108 | 407122 | 115 | 452927 | 70 | 452772 | 1,50E-03 | 3,47E-03 | 2,65E-04 | 2,54E-04 | 1,55E-04 | 3,67E-04 | 2,80E-04 | 3,21E-04 | 3,69E-04 | 3,44E-04 | 3,23E-04 | 4,09E-04 | 4,10E-04 | 3,84E-04 |
| *Ppard* | ENSMUSG00000002250 | 17:28369699-28370198 | 138 | 562238 | 119 | 599180 | 83 | 585038 | 3,15E-04 | 4,36E-02 | 2,45E-04 | 1,99E-04 | 1,42E-04 | 4,81E-04 | 4,02E-04 | 4,03E-04 | 4,44E-04 | 4,28E-04 | 4,17E-04 | 4,80E-04 | 4,07E-04 | 3,31E-04 |
| *Ppia* | ENSMUSG00000071866 | 11:6315873-6316372 | 103 | 409774 | 109 | 450054 | 56 | 378333 | 4,14E-03 | 7,85E-03 | 2,51E-04 | 2,42E-04 | 1,48E-04 | 3,47E-04 | 3,03E-04 | 3,50E-04 | 4,38E-04 | 3,76E-04 | 3,45E-04 | 4,36E-04 | 4,24E-04 | 3,33E-04 |
| *Ppp1r15b* | ENSMUSG00000046062 | 1:135027743-135028242 | 156 | 783542 | 160 | 841987 | 108 | 807394 | 4,55E-03 | 1,38E-02 | 1,99E-04 | 1,90E-04 | 1,34E-04 | 3,69E-04 | 3,31E-04 | 3,66E-04 | 3,45E-04 | 3,48E-04 | 3,18E-04 | 3,84E-04 | 3,67E-04 | 3,51E-04 |
| *Ppp1r16b-2* | ENSMUSG00000037754 | 2:158492871-158493370 | 156 | 557644 | 239 | 619563 | 109 | 590836 | 2,79E-03 | 3,74E-10 | 2,80E-04 | 3,86E-04 | 1,84E-04 | 4,26E-04 | 3,59E-04 | 3,62E-04 | 3,84E-04 | 4,22E-04 | 3,21E-04 | 4,47E-04 | 5,14E-04 | 3,48E-04 |
| *Prdx1* | ENSMUSG00000028691 | 4:116358204-116358703 | 317 | 838736 | 697 | 935580 | 158 | 875549 | 6,39E-14 | 2,03E-72 | 3,78E-04 | 7,45E-04 | 1,80E-04 | 1,17E-03 | 1,20E-03 | 7,12E-04 | 1,42E-03 | 6,83E-04 | 1,22E-03 | 8,06E-04 | 1,51E-03 | 5,60E-04 |
| *Psmc3* | ENSMUSG00000002102 | 2:90894173-90894672 | 251 | 880879 | 278 | 980450 | 112 | 893512 | 6,93E-13 | 2,88E-13 | 2,85E-04 | 2,84E-04 | 1,25E-04 | 3,37E-04 | 3,15E-04 | 3,89E-04 | 3,86E-04 | 3,29E-04 | 3,35E-04 | 4,49E-04 | 4,01E-04 | 3,31E-04 |
| *Ptbp2* | ENSMUSG00000028134 | 3:119485807-119486306 | 219 | 520503 | 223 | 581299 | 96 | 557036 | 3,58E-13 | 9,30E-11 | 4,21E-04 | 3,84E-04 | 1,72E-04 | 3,71E-04 | 3,07E-04 | 3,71E-04 | 3,65E-04 | 3,20E-04 | 3,18E-04 | 5,67E-04 | 5,09E-04 | 3,52E-04 |
| *Pten* | ENSMUSG00000013663 | 19:32832067-32832566 | 142 | 429732 | 101 | 462809 | 58 | 450616 | 2,01E-09 | 4,28E-03 | 3,30E-04 | 2,18E-04 | 1,29E-04 | 3,98E-04 | 3,22E-04 | 4,28E-04 | 4,00E-04 | 3,90E-04 | 3,09E-04 | 5,40E-04 | 4,03E-04 | 3,27E-04 |
| *Ptma* | ENSMUSG00000026238 | 1:88423311-88423810 | 600 | 584974 | 862 | 621012 | 102 | 593937 | 1,04E-86 | 1,34E-141 | 1,03E-03 | 1,39E-03 | 1,72E-04 | 4,49E-04 | 4,14E-04 | 3,90E-04 | 3,65E-04 | 5,22E-04 | 4,48E-04 | 9,58E-04 | 1,19E-03 | 3,41E-04 |
| *Ptprc* | ENSMUSG00000026395 | 1:140071383-140071882 | 173 | 560187 | 234 | 614866 | 75 | 584306 | 3,43E-10 | 1,83E-17 | 3,09E-04 | 3,81E-04 | 1,28E-04 | 3,42E-04 | 3,07E-04 | 3,49E-04 | 3,43E-04 | 3,09E-04 | 2,93E-04 | 3,86E-04 | 4,16E-04 | 2,99E-04 |
| *Rac1* | ENSMUSG00000001847 | 5:144288362-144288861 | 195 | 836186 | 226 | 911049 | 117 | 868552 | 1,11E-05 | 2,89E-07 | 2,33E-04 | 2,48E-04 | 1,35E-04 | 3,86E-04 | 3,22E-04 | 3,82E-04 | 3,82E-04 | 3,77E-04 | 3,61E-04 | 4,07E-04 | 4,48E-04 | 3,56E-04 |
| *Rad23a* | ENSMUSG00000003813 | 8:87364041-87364540 | 132 | 486609 | 162 | 519096 | 78 | 496719 | 4,86E-04 | 1,99E-06 | 2,71E-04 | 3,12E-04 | 1,57E-04 | 3,40E-04 | 3,15E-04 | 3,86E-04 | 3,77E-04 | 3,53E-04 | 3,39E-04 | 4,17E-04 | 4,42E-04 | 3,45E-04 |
| *Rad51* | ENSMUSG00000027323 | 2:118938553-118939052 | 194 | 901196 | 284 | 991878 | 145 | 903256 | 1,83E-02 | 4,89E-08 | 2,15E-04 | 2,86E-04 | 1,61E-04 | 3,55E-04 | 3,14E-04 | 3,69E-04 | 3,70E-04 | 3,53E-04 | 3,33E-04 | 3,93E-04 | 4,06E-04 | 3,36E-04 |
| *Rad9* | ENSMUSG00000024824 | 19:4201104-4201603 | 220 | 1076714 | 208 | 1165317 | 127 | 1061661 | 6,22E-06 | 1,31E-03 | 2,04E-04 | 1,78E-04 | 1,20E-04 | 3,45E-04 | 2,85E-04 | 3,59E-04 | 3,72E-04 | 3,23E-04 | 3,31E-04 | 3,62E-04 | 3,40E-04 | 3,19E-04 |
| *Ranbp1* | ENSMUSG00000005732 | 16:18248288-18248787 | 170 | 711915 | 229 | 752275 | 129 | 728583 | 2,41E-02 | 3,43E-06 | 2,39E-04 | 3,04E-04 | 1,77E-04 | 3,89E-04 | 3,49E-04 | 3,57E-04 | 3,56E-04 | 3,62E-04 | 3,33E-04 | 3,86E-04 | 4,64E-04 | 3,69E-04 |
| *Rbm15* | ENSMUSG00000048109 | 3:107135708-107136207 | 232 | 629055 | 147 | 708227 | 75 | 683164 | 1,12E-21 | 2,50E-05 | 3,69E-04 | 2,08E-04 | 1,10E-04 | 3,48E-04 | 3,06E-04 | 3,46E-04 | 3,81E-04 | 3,58E-04 | 3,03E-04 | 4,83E-04 | 3,67E-04 | 3,14E-04 |

| | | | Transitions at C/G pairs | | | | | | FDR | | C/G Transition Frequency | | | Total mutation frequency | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Ung$^{-/-}$Msh2$^{-/-}$Exp1 | | Ung$^{-/-}$Msh2$^{-/-}$Exp2 | | *Aicda*$^{-/-}$ | | | | Ung$^{-/-}$Msh2$^{-/-}$ | | *Aicda*$^{-/-}$ | Ung$^{+/-}$Msh2$^{+/-}$ | | Ung$^{+/-}$Msh2$^{-/-}$ | | Ung$^{-/-}$Msh2$^{+/-}$ | | Ung$^{-/-}$Msh2$^{-/-}$ | | *Aicda*$^{-/-}$ |
| **Gene** | **Ensembl ID** | **Coordinates** | **C/G transitions** | **C/G sequenced** | **C/G transitions** | **C/G sequenced** | **C/G transitions** | **C/G sequenced** | **Exp1** | **Exp2** | **Exp1** | **Exp2** | **Exp1** | **Exp1** | **Exp2** | **Exp1** | **Exp2** | **Exp1** | **Exp2** | **Exp1** | **Exp2** | **Exp1** |
| *Rbm19* | ENSMUSG00000029594 | 5:120566522-120567021 | 112 | 536555 | 150 | 604127 | 76 | 558948 | 1,08E-02 | 6,47E-05 | 2,09E-04 | 2,48E-04 | 1,36E-04 | 3,76E-04 | 2,89E-04 | 3,26E-04 | 3,74E-04 | 3,26E-04 | 3,31E-04 | 4,20E-04 | 3,85E-04 | 3,19E-04 |
| *Rbm39* | ENSMUSG00000027620 | 2:156005477-156005976 | 286 | 634965 | 205 | 689250 | 93 | 677920 | 1,18E-25 | 1,30E-09 | 4,50E-04 | 2,97E-04 | 1,37E-04 | 3,91E-04 | 3,39E-04 | 3,48E-04 | 4,14E-04 | 4,15E-04 | 3,27E-04 | 5,57E-04 | 4,29E-04 | 3,32E-04 |
| *Recql4* | ENSMUSG00000033762 | 15:76540407-76540906 | 144 | 581529 | 128 | 646820 | 78 | 628329 | 3,31E-06 | 3,77E-03 | 2,48E-04 | 1,98E-04 | 1,24E-04 | 3,82E-04 | 3,26E-04 | 3,10E-04 | 3,73E-04 | 3,52E-04 | 3,11E-04 | 4,11E-04 | 3,65E-04 | 3,28E-04 |
| *Rel* | ENSMUSG00000020275 | 11:23670471-23670970 | 144 | 603446 | 168 | 672273 | 88 | 626167 | 3,61E-04 | 4,33E-05 | 2,39E-04 | 2,50E-04 | 1,41E-04 | 3,88E-04 | 3,50E-04 | 4,07E-04 | 3,35E-04 | 3,54E-04 | 3,46E-04 | 4,11E-04 | 3,94E-04 | 3,46E-04 |
| *Rev3l* | ENSMUSG00000019841 | 10:39451966-39452465 | 103 | 472492 | 131 | 522108 | 73 | 483878 | 3,58E-02 | 1,61E-03 | 2,18E-04 | 2,51E-04 | 1,51E-04 | 3,80E-04 | 3,34E-04 | 3,48E-04 | 3,60E-04 | 3,26E-04 | 3,68E-04 | 3,98E-04 | 3,91E-04 | 3,02E-04 |
| *Rfc2* | ENSMUSG00000023104 | 5:135058560-135059059 | 159 | 415540 | 143 | 457169 | 72 | 425790 | 2,04E-08 | 6,93E-05 | 3,83E-04 | 3,13E-04 | 1,69E-04 | 3,88E-04 | 3,45E-04 | 3,51E-04 | 3,79E-04 | 3,47E-04 | 3,16E-04 | 5,34E-04 | 4,79E-04 | 3,42E-04 |
| *Rgs13* | ENSMUSG00000051079 | 1:146024003-146024502 | 108 | 571740 | 177 | 666112 | 76 | 606918 | 1,56E-02 | 9,56E-08 | 1,89E-04 | 2,66E-04 | 1,25E-04 | 3,30E-04 | 3,07E-04 | 3,27E-04 | 3,20E-04 | 2,96E-04 | 3,06E-04 | 3,02E-04 | 3,59E-04 | 3,30E-04 |
| *Rhoh* | ENSMUSG00000029204 | 5:66254808-66255307 | 787 | 706562 | 951 | 773153 | 248 | 753845 | 3,88E-72 | 5,44E-92 | 1,11E-03 | 1,23E-03 | 3,29E-04 | 5,33E-04 | 4,08E-04 | 4,24E-04 | 4,27E-04 | 4,26E-04 | 3,78E-04 | 7,88E-04 | 8,29E-04 | 4,03E-04 |
| *Rims1* | ENSMUSG00000041670 | 1:22812064-22812563 | 117 | 589206 | 136 | 692194 | 95 | 677255 | 2,80E-02 | 3,01E-02 | 1,99E-04 | 1,96E-04 | 1,40E-04 | 3,66E-04 | 3,56E-04 | 4,33E-04 | 4,06E-04 | 4,79E-04 | 3,13E-04 | 3,86E-04 | 3,79E-04 | 3,21E-04 |
| *Rnf2* | ENSMUSG00000026484 | 1:153347454-153347953 | 35 | 140958 | 36 | 155865 | 17 | 158125 | 1,23E-02 | 2,53E-02 | 2,48E-04 | 2,31E-04 | 1,08E-04 | 4,72E-04 | 4,61E-04 | 4,81E-04 | 4,92E-04 | 5,05E-04 | 3,86E-04 | 5,40E-04 | 5,04E-04 | 4,35E-04 |
| *Rpa1* | ENSMUSG00000000751 | 11:75161386-75161885 | 178 | 738607 | 153 | 810620 | 106 | 781374 | 1,15E-05 | 2,40E-02 | 2,41E-04 | 1,89E-04 | 1,36E-04 | 3,49E-04 | 3,03E-04 | 3,46E-04 | 3,54E-04 | 3,62E-04 | 3,24E-04 | 3,89E-04 | 3,68E-04 | 3,43E-04 |
| *Rpia* | ENSMUSG00000053604 | 6:70741670-70742169 | 130 | 705834 | 127 | 754821 | 90 | 761126 | 3,91E-03 | 2,75E-02 | 1,84E-04 | 1,68E-04 | 1,18E-04 | 3,62E-04 | 2,82E-04 | 2,94E-04 | 3,98E-04 | 3,80E-04 | 3,36E-04 | 3,76E-04 | 3,34E-04 | 3,19E-04 |
| *Rpl29* | ENSMUSG00000048758 | 9:106331870-106332369 | 228 | 625577 | 147 | 670209 | 85 | 617428 | 7,12E-15 | 2,03E-03 | 3,64E-04 | 2,19E-04 | 1,38E-04 | 3,27E-04 | 3,46E-04 | 3,91E-04 | 4,16E-04 | 3,74E-04 | 3,28E-04 | 4,81E-04 | 4,09E-04 | 3,25E-04 |
| *Rpl3* | ENSMUSG00000060036 | 15:79913337-79913836 | 482 | 627694 | 306 | 689141 | 93 | 621963 | 5,20E-62 | 5,01E-22 | 7,68E-04 | 4,44E-04 | 1,50E-04 | 3,54E-04 | 3,63E-04 | 3,64E-04 | 3,59E-04 | 3,90E-04 | 3,83E-04 | 6,71E-04 | 4,88E-04 | 3,37E-04 |
| *Rpl32* | ENSMUSG00000057841 | 6:115758262-115758761 | 235 | 467993 | 281 | 496836 | 71 | 465442 | 2,15E-20 | 1,12E-26 | 5,02E-04 | 5,66E-04 | 1,53E-04 | 3,70E-04 | 2,95E-04 | 3,41E-04 | 3,58E-04 | 3,91E-04 | 3,53E-04 | 6,01E-04 | 5,94E-04 | 3,34E-04 |
| *Rpl35* | ENSMUSG00000062997 | 2:38860152-38860651 | 171 | 519530 | 213 | 550445 | 82 | 541086 | 2,04E-08 | 3,58E-13 | 3,29E-04 | 3,87E-04 | 1,52E-04 | 3,55E-04 | 3,31E-04 | 3,77E-04 | 3,63E-04 | 3,76E-04 | 3,63E-04 | 4,59E-04 | 4,70E-04 | 3,46E-04 |
| *Rpl4* | ENSMUSG00000032399 | 9:64021194-64021693 | 348 | 727090 | 332 | 819432 | 90 | 792082 | 1,60E-40 | 9,80E-31 | 4,79E-04 | 4,05E-04 | 1,14E-04 | 3,25E-04 | 3,27E-04 | 3,61E-04 | 3,49E-04 | 3,45E-04 | 4,10E-04 | 5,08E-04 | 4,41E-04 | 2,83E-04 |
| *Rpl41* | ENSMUSG00000093674 | 10:127985725-127986224 | 231 | 704789 | 257 | 774035 | 108 | 722845 | 3,19E-11 | 4,38E-12 | 3,28E-04 | 3,32E-04 | 1,49E-04 | 3,81E-04 | 3,07E-04 | 3,53E-04 | 3,86E-04 | 3,56E-04 | 3,13E-04 | 4,63E-04 | 4,38E-04 | 3,49E-04 |
| *Rplp0* | ENSMUSG00000067274 | 5:116009476-116009975 | 311 | 720711 | 388 | 773252 | 126 | 730965 | 1,16E-18 | 8,41E-28 | 4,32E-04 | 5,02E-04 | 1,72E-04 | 3,93E-04 | 3,08E-04 | 3,84E-04 | 3,98E-04 | 3,88E-04 | 2,96E-04 | 5,69E-04 | 6,06E-04 | 3,43E-04 |
| *Rps12* | ENSMUSG00000061983 | 10:23506516-23507015 | 388 | 564928 | 525 | 606055 | 117 | 600164 | 2,34E-37 | 2,24E-60 | 6,87E-04 | 8,66E-04 | 1,95E-04 | 3,87E-04 | 3,28E-04 | 3,68E-04 | 3,40E-04 | 3,69E-04 | 3,50E-04 | 6,84E-04 | 7,59E-04 | 3,44E-04 |
| *Rps14* | ENSMUSG00000024608 | 18:60934250-60934749 | 512 | 783590 | 585 | 875489 | 114 | 766837 | 8,34E-58 | 7,73E-63 | 6,53E-04 | 6,68E-04 | 1,49E-04 | 3,40E-04 | 3,47E-04 | 3,57E-04 | 4,15E-04 | 3,81E-04 | 3,62E-04 | 6,93E-04 | 6,60E-04 | 3,21E-04 |
| *Rps5* | ENSMUSG00000012848 | 7:13507660-13508159 | 394 | 790650 | 426 | 881418 | 111 | 831430 | 5,03E-40 | 2,36E-39 | 4,98E-04 | 4,83E-04 | 1,34E-04 | 3,96E-04 | 2,84E-04 | 3,02E-04 | 3,34E-04 | 3,35E-04 | 3,91E-04 | 5,31E-04 | 5,19E-04 | 2,91E-04 |
| *Rps6* | ENSMUSG00000028495 | 4:86502772-86503271 | 263 | 697680 | 232 | 759107 | 107 | 759467 | 2,80E-18 | 7,03E-11 | 3,77E-04 | 3,06E-04 | 1,41E-04 | 4,87E-04 | 4,64E-04 | 5,76E-04 | 5,19E-04 | 5,19E-04 | 5,36E-04 | 6,14E-04 | 5,78E-04 | 3,03E-04 |
| *Rps9* | ENSMUSG00000006333 | 7:3655643-3656142 | 1393 | 786312 | 1432 | 859642 | 1155 | 822127 | 3,23E-08 | 6,92E-05 | 1,77E-03 | 1,67E-03 | 1,40E-03 | 1,35E-03 | 1,36E-03 | 1,32E-03 | 1,44E-03 | 1,25E-03 | 1,26E-03 | 1,56E-03 | 1,49E-03 | 1,31E-03 |
| *Rrm1* | ENSMUSG00000030978 | 7:109590209-109590708 | 137 | 564554 | 130 | 615072 | 92 | 615904 | 1,15E-03 | 2,83E-02 | 2,43E-04 | 2,11E-04 | 1,49E-04 | 1,94E-03 | 1,60E-03 | 1,85E-03 | 2,46E-03 | 1,81E-03 | 1,78E-03 | 1,97E-03 | 1,58E-03 | 2,45E-03 |
| *Rrm2* | ENSMUSG00000020649 | 12:25393119-25393618 | 182 | 610935 | 201 | 669515 | 104 | 626036 | 7,95E-06 | 3,47E-06 | 2,98E-04 | 3,00E-04 | 1,66E-04 | 3,57E-04 | 3,48E-04 | 4,89E-04 | 3,77E-04 | 3,73E-04 | 3,14E-04 | 4,59E-04 | 5,50E-04 | 3,46E-04 |
| *Runx1t1-2* | ENSMUSG00000006586 | 4:13678444-13678943 | 111 | 552085 | 129 | 630298 | 85 | 604088 | 3,07E-02 | 2,03E-02 | 2,01E-04 | 2,05E-04 | 1,41E-04 | 5,36E-04 | 4,63E-04 | 5,62E-04 | 5,95E-04 | 5,29E-04 | 5,35E-04 | 5,64E-04 | 6,01E-04 | 4,55E-04 |
| *Safb* | ENSMUSG00000071054 | 17:56724405-56724904 | 124 | 537299 | 104 | 587865 | 67 | 555762 | 6,81E-05 | 3,55E-02 | 2,31E-04 | 1,77E-04 | 1,21E-04 | 3,65E-04 | 3,36E-04 | 3,77E-04 | 3,56E-04 | 3,85E-04 | 3,05E-04 | 4,02E-04 | 3,22E-04 | 3,19E-04 |
| *Serinc3* | ENSMUSG00000017707 | 2:163470380-163470879 | 158 | 639032 | 288 | 677569 | 107 | 680762 | 1,08E-03 | 1,96E-19 | 2,47E-04 | 4,25E-04 | 1,57E-04 | 4,04E-04 | 3,04E-04 | 3,73E-04 | 3,70E-04 | 3,70E-04 | 3,44E-04 | 3,93E-04 | 5,72E-04 | 3,67E-04 |
| *Sf3b1* | ENSMUSG00000025982 | 1:55083823-55084322 | 216 | 956671 | 226 | 1029605 | 132 | 934345 | 8,85E-05 | 2,12E-04 | 2,26E-04 | 2,20E-04 | 1,41E-04 | 3,62E-04 | 3,17E-04 | 3,64E-04 | 3,96E-04 | 3,58E-04 | 3,26E-04 | 4,01E-04 | 3,98E-04 | 3,20E-04 |
| *Sh3bp5* | ENSMUSG00000021892 | 14:32248720-32249219 | 151 | 485566 | 172 | 523489 | 96 | 521000 | 2,50E-04 | 2,16E-05 | 3,11E-04 | 3,29E-04 | 1,84E-04 | 4,23E-04 | 2,95E-04 | 3,73E-04 | 3,94E-04 | 4,41E-04 | 3,40E-04 | 4,41E-04 | 4,47E-04 | 3,38E-04 |
| *Sipa1* | ENSMUSG00000056917 | 19:5663208-5663707 | 200 | 664806 | 131 | 711958 | 92 | 687281 | 2,58E-10 | 4,45E-02 | 3,01E-04 | 1,84E-04 | 1,34E-04 | 3,31E-04 | 2,95E-04 | 3,18E-04 | 3,67E-04 | 3,45E-04 | 3,51E-04 | 4,40E-04 | 3,17E-04 | 3,05E-04 |
| *Sirt6-2* | ENSMUSG00000034748 | 10:81089854-81090353 | 143 | 728703 | 186 | 783900 | 89 | 760033 | 5,03E-04 | 1,25E-07 | 1,96E-04 | 2,37E-04 | 1,17E-04 | 3,67E-04 | 3,47E-04 | 3,50E-04 | 4,00E-04 | 3,89E-04 | 3,11E-04 | 3,93E-04 | 4,03E-04 | 3,19E-04 |
| *Slbp* | ENSMUSG00000004642 | 5:33994456-33994955 | 220 | 676978 | 214 | 715314 | 92 | 694595 | 4,84E-13 | 1,04E-10 | 3,25E-04 | 2,99E-04 | 1,32E-04 | 3,66E-04 | 3,11E-04 | 3,68E-04 | 3,84E-04 | 3,45E-04 | 3,49E-04 | 4,40E-04 | 4,24E-04 | 2,95E-04 |
| *Slc30a6* | ENSMUSG00000024069 | 17:74794972-74795471 | 135 | 815974 | 265 | 906712 | 102 | 871918 | 2,02E-02 | 1,74E-15 | 1,65E-04 | 2,92E-04 | 1,17E-04 | 3,41E-04 | 4,48E-04 | 3,37E-04 | 3,68E-04 | 3,25E-04 | 2,86E-04 | 3,40E-04 | 3,97E-04 | 3,05E-04 |
| *Smu* | NoCode001 | 12:114663752-114664475 | 37333 | 734739 | 40702 | 779929 | 136 | 977119 | 0,00E+00 | 0,00E+00 | 5,08E-02 | 5,22E-02 | 1,39E-04 | 6,23E-03 | 6,14E-03 | 4,31E-03 | 3,41E-03 | 5,42E-03 | 8,00E-03 | 2,31E-02 | 2,39E-02 | 3,29E-04 |
| *Snx5* | ENSMUSG00000027423 | 2:144096139-144096638 | 157 | 533761 | 230 | 580477 | 102 | 559053 | 6,44E-04 | 1,33E-10 | 2,94E-04 | 3,96E-04 | 1,82E-04 | 3,51E-04 | 3,38E-04 | 4,44E-04 | 4,46E-04 | 4,18E-04 | 3,54E-04 | 4,81E-04 | 5,00E-04 | 3,70E-04 |
| *Sod1* | ENSMUSG00000022982 | 16:90220987-90221486 | 145 | 690849 | 138 | 772156 | 93 | 724434 | 7,85E-04 | 3,39E-02 | 2,10E-04 | 1,79E-04 | 1,28E-04 | 3,39E-04 | 3,16E-04 | 3,91E-04 | 3,98E-04 | 3,65E-04 | 3,16E-04 | 4,12E-04 | 3,99E-04 | 3,29E-04 |
| *Sp3-2* | ENSMUSG00000027109 | 2:72818004-72818503 | 95 | 300119 | 103 | 323520 | 64 | 300620 | 3,23E-02 | 2,97E-02 | 3,17E-04 | 3,18E-04 | 2,13E-04 | 3,72E-04 | 4,05E-04 | 3,62E-04 | 4,52E-04 | 4,09E-04 | 3,39E-04 | 4,64E-04 | 4,36E-04 | 3,67E-04 |
| *Spib* | ENSMUSG00000008193 | 7:51786942-51787441 | 115 | 381443 | 147 | 413056 | 44 | 404567 | 1,12E-08 | 8,78E-13 | 3,01E-04 | 3,56E-04 | 1,09E-04 | 6,37E-04 | 5,06E-04 | 6,68E-04 | 5,72E-04 | 6,50E-04 | 6,56E-04 | 6,27E-04 | 7,13E-04 | 5,73E-04 |
| *Sra1* | ENSMUSG00000006050 | 18:36829466-36829965 | 158 | 834218 | 197 | 920653 | 107 | 885311 | 1,19E-03 | 7,37E-06 | 1,89E-04 | 2,14E-04 | 1,21E-04 | 3,73E-04 | 2,76E-04 | 3,60E-04 | 3,36E-04 | 3,30E-04 | 2,95E-04 | 3,63E-04 | 3,76E-04 | 3,15E-04 |
| *Srsf10* | ENSMUSG00000028676 | 4:135412007-135412506 | 150 | 328508 | 311 | 354771 | 45 | 346103 | 1,16E-14 | 2,30E-47 | 4,57E-04 | 8,77E-04 | 1,30E-04 | 1,05E-03 | 7,42E-04 | 1,67E-03 | 1,56E-03 | 1,34E-03 | 2,56E-03 | 1,17E-03 | 1,56E-03 | 4,08E-04 |
| *St6gal1-1* | ENSMUSG00000022885 | 16:23224813-23225312 | 191 | 690488 | 261 | 769742 | 97 | 721765 | 2,10E-08 | 2,77E-15 | 2,77E-04 | 3,39E-04 | 1,34E-04 | 3,36E-04 | 3,09E-04 | 3,53E-04 | 3,56E-04 | 3,68E-04 | 2,99E-04 | 4,07E-04 | 4,50E-04 | 3,01E-04 |
| *St6gal1-2* | ENSMUSG00000022885 | 16:23226094-23226593 | 224 | 862113 | 211 | 969163 | 107 | 909042 | 3,36E-11 | 7,61E-07 | 2,60E-04 | 2,18E-04 | 1,18E-04 | 3,53E-04 | 3,48E-04 | 3,25E-04 | 3,42E-04 | 3,30E-04 | 3,67E-04 | 3,83E-04 | 3,51E-04 | 3,29E-04 |
| *Stau1* | ENSMUSG00000039536 | 2:166821279-166821778 | 90 | 456903 | 96 | 484273 | 57 | 450848 | 2,15E-02 | 2,00E-02 | 1,97E-04 | 1,98E-04 | 1,26E-04 | 3,64E-04 | 3,17E-04 | 4,31E-04 | 4,69E-04 | 3,84E-04 | 3,35E-04 | 3,88E-04 | 3,68E-04 | 3,13E-04 |
| *Stx3-2* | ENSMUSG00000041488 | 19:11893394-11893893 | 167 | 695033 | 206 | 798519 | 126 | 688379 | 4,28E-02 | 7,06E-03 | 2,40E-04 | 2,58E-04 | 1,83E-04 | 8,19E-04 | 6,96E-04 | 7,37E-04 | 7,47E-04 | 7,41E-04 | 7,64E-04 | 7,89E-04 | 7,05E-04 | 6,31E-04 |
| *Sv2b-1* | ENSMUSG00000053025 | 7:82452773-82453272 | 126 | 769258 | 141 | 861550 | 100 | 848988 | 3,16E-02 | 3,03E-02 | 1,64E-04 | 1,64E-04 | 1,18E-04 | 3,33E-04 | 3,52E-04 | 3,66E-04 | 3,56E-04 | 3,60E-04 | 3,39E-04 | 3,57E-04 | 3,30E-04 | 3,35E-04 |
| *Swap70* | ENSMUSG00000031015 | 7:117365217-117365716 | 152 | 438736 | 245 | 465057 | 70 | 441820 | 1,55E-07 | 6,00E-21 | 3,46E-04 | 5,27E-04 | 1,58E-04 | 3,57E-04 | 3,35E-04 | 4,11E-04 | 3,42E-04 | 3,74E-04 | 1,08E-03 | 5,28E-04 | 6,09E-04 | 3,09E-04 |
| *Syk* | ENSMUSG00000021457 | 13:52692251-52692750 | 354 | 761725 | 360 | 841588 | 95 | 795086 | 2,34E-37 | 5,56E-33 | 4,65E-04 | 4,28E-04 | 1,19E-04 | 4,70E-04 | 3,38E-04 | 3,15E-04 | 3,41E-04 | 3,54E-04 | 3,54E-04 | 4,90E-04 | 4,42E-04 | 2,99E-04 |
| *Taf4a* | ENSMUSG00000039117 | 2:179710852-179711351 | 30 | 84085 | 36 | 86986 | 11 | 81819 | 1,30E-02 | 2,26E-03 | 3,57E-04 | 4,14E-04 | 1,34E-04 | 3,07E-04 | 3,48E-04 | 3,27E-04 | 3,69E-04 | 4,07E-04 | 4,02E-04 | 5,02E-04 | 5,53E-04 | 3,74E-04 |
| *Taf9-2* | ENSMUSG00000052293 | 13:101421562-101422061 | 311 | 715720 | 182 | 822626 | 86 | 760905 | 7,40E-33 | 8,78E-07 | 4,35E-04 | 2,21E-04 | 1,13E-04 | 3,19E-04 | 3,04E-04 | 3,13E-04 | 3,34E-04 | 3,42E-04 | 3,12E-04 | 4,72E-04 | 3,64E-04 | 3,38E-04 |
| *Tapbp* | ENSMUSG00000024308 | 17:34056423-34056922 | 247 | 753754 | 184 | 844945 | 93 | 799215 | 3,01E-18 | 2,93E-06 | 3,28E-04 | 2,18E-04 | 1,16E-04 | 3,41E-04 | 3,42E-04 | 3,80E-04 | 3,88E-04 | 3,69E-04 | 3,33E-04 | 4,61E-04 | 3,63E-04 | 3,26E-04 |
| *Tcea1-2* | ENSMUSG00000033813 | 1:4848409-4848908 | 191 | 717890 | 164 | 850451 | 89 | 797762 | 2,33E-11 | 1,14E-04 | 2,66E-04 | 1,93E-04 | 1,12E-04 | 3,53E-04 | 3,20E-04 | 3,34E-04 | 3,40E-04 | 3,11E-04 | 3,20E-04 | 3,76E-04 | 3,48E-04 | 3,07E-04 |
| *Tcf3* | ENSMUSG00000020167 | 10:79895899-79896398 | 179 | 678888 | 132 | 695636 | 86 | 667711 | 1,35E-07 | 1,48E-02 | 2,64E-04 | 1,90E-04 | 1,29E-04 | 3,53E-04 | 3,44E-04 | 3,89E-04 | 4,81E-04 | 3,93E-04 | 4,12E-04 | 4,54E-04 | 4,30E-04 | 3,65E-04 |
| *Tcf4-1* | ENSMUSG00000053477 | 18:69504146-69504645 | 122 | 641344 | 155 | 753715 | 98 | 706912 | 4,15E-02 | 6,77E-03 | 1,90E-04 | 2,06E-04 | 1,39E-04 | 3,39E-04 | 3,21E-04 | 3,32E-04 | 3,38E-04 | 3,05E-04 | 3,10E-04 | 3,39E-04 | 3,16E-04 | 3,28E-04 |
| *Telomerase-vert* | ENSMUSG00000064796 | 3:96218257-96218756 | 167 | 847739 | 159 | 913453 | 109 | 860318 | 1,19E-03 | 2,74E-02 | 1,97E-04 | 1,74E-04 | 1,27E-04 | 3,63E-04 | 3,20E-04 | 3,57E-04 | 3,69E-04 | 3,25E-04 | 2,98E-04 | 3,64E-04 | 3,46E-04 | 3,03E-04 |
| *Tet2* | ENSMUSG00000040943 | 3:133206855-133207354 | 151 | 687031 | 193 | 760095 | 106 | 734090 | 2,98E-03 | 1,15E-05 | 2,20E-04 | 2,54E-04 | 1,44E-04 | 4,13E-04 | 3,04E-04 | 4,06E-04 | 3,79E-04 | 3,36E-04 | 3,14E-04 | 3,90E-04 | 3,98E-04 | 2,91E-04 |
| *Tex14-1* | ENSMUSG00000010342 | 11:87218567-87219066 | 70 | 408369 | 81 | 446181 | 47 | 420648 | 4,85E-02 | 2,22E-02 | 1,71E-04 | 1,82E-04 | 1,12E-04 | 3,89E-04 | 4,27E-04 | 3,57E-04 | 3,77E-04 | 3,50E-04 | 2,73E-04 | 3,42E-04 | 3,65E-04 | 3,12E-04 |
| *Tfap4* | ENSMUSG00000005718 | 16:4559221-4559720 | 123 | 652623 | 125 | 696841 | 80 | 646779 | 9,28E-03 | 2,49E-02 | 1,88E-04 | 1,79E-04 | 1,24E-04 | 3,98E-04 | 3,85E-04 | 4,61E-04 | 4,17E-04 | 4,12E-04 | 4,03E-04 | 4,31E-04 | 3,96E-04 | 4,00E-04 |
| *Tfdp1* | ENSMUSG00000038482 | 8:13339674-13340173 | 125 | 409604 | 196 | 477394 | 97 | 463004 | 1,44E-02 | 1,89E-07 | 3,05E-04 | 4,11E-04 | 2,10E-04 | 4,08E-04 | 3,67E-04 | 4,21E-04 | 3,94E-04 | 4,27E-04 | 3,77E-04 | 4,03E-04 | 5,00E-04 | 3,84E-04 |
| *Tmsb4x* | ENSMUSG00000049775 | X:163646651-163647150 | 1326 | 456363 | 1389 | 433376 | 78 | 437003 | 2,98E-280 | 0,00E+00 | 2,91E-03 | 3,21E-03 | 1,78E-04 | 5,43E-04 | 5,74E-04 | 7,41E-04 | 5,67E-04 | 6,65E-04 | 7,20E-04 | 1,87E-03 | 1,99E-03 | 4,09E-04 |
| *Tnf* | ENSMUSG00000024401 | 17:35338442-35338941 | 198 | 828728 | 197 | 940588 | 87 | 884605 | 7,16E-12 | 9,41E-09 | 2,39E-04 | 2,09E-04 | 9,83E-05 | 3,26E-04 | 3,16E-04 | 3,32E-04 | 3,45E-04 | 3,10E-04 | 2,86E-04 | 4,05E-04 | 3,86E-04 | 2,93E-04 |
| *Tnfaip3-1* | ENSMUSG00000019850 | 10:18731270-18731769 | 206 | 815638 | 179 | 893794 | 97 | 856932 | 1,50E-10 | 2,25E-05 | 2,53E-04 | 2,00E-04 | 1,13E-04 | 3,68E-04 | 3,26E-04 | 3,32E-04 | 3,42E-04 | 3,18E-04 | 2,98E-04 | 3,87E-04 | 3,44E-04 | 2,85E-04 |
| *Tnfaip3-2* | ENSMUSG00000019850 | 10:18734717-18735216 | 266 | 789064 | 206 | 866245 | 110 | 830029 | 8,51E-17 | 2,78E-06 | 3,37E-04 | 2,38E-04 | 1,33E-04 | 4,03E-04 | 3,32E-04 | 3,71E-04 | 3,67E-04 | 3,75E-04 | 3,68E-04 | 4,67E-04 | 4,06E-04 | 3,23E-04 |
| *Top1* | ENSMUSG00000070544 | 2:160471633-160472132 | 156 | 671777 | 188 | 732294 | 101 | 707823 | 5,28E-04 | 7,15E-06 | 2,32E-04 | 2,57E-04 | 1,43E-04 | 3,51E-04 | 3,22E-04 | 3,70E-04 | 4,13E-04 | 3,42E-04 | 3,15E-04 | 4,16E-04 | 3,93E-04 | 3,39E-04 |
| *Top2a* | ENSMUSG00000020914 | 11:98885004-98885503 | 185 | 834544 | 219 | 904432 | 119 | 852505 | 3,37E-04 | 5,06E-06 | 2,22E-04 | 2,42E-04 | 1,40E-04 | 3,66E-04 | 3,19E-04 | 3,76E-04 | 3,53E-04 | 3,34E-04 | 3,03E-04 | 3,64E-04 | 3,69E-04 | 3,15E-04 |
| *Topors* | ENSMUSG00000036822 | 4:40216375-40216874 | 144 | 826069 | 197 | 910587 | 111 | 872728 | 2,87E-02 | 2,91E-05 | 1,74E-04 | 2,16E-04 | 1,27E-04 | 3,45E-04 | 3,41E-04 | 3,50E-04 | 3,87E-04 | 3,44E-04 | 2,92E-04 | 3,67E-04 | 3,53E-04 | 3,05E-04 |
| *Tra2b* | ENSMUSG00000022858 | 16:22265503-22266002 | 162 | 636870 | 273 | 687689 | 120 | 651574 | 1,81E-02 | 3,70E-12 | 2,54E-04 | 3,97E-04 | 1,84E-04 | 5,30E-04 | 4,34E-04 | 4,28E-04 | 4,93E-04 | 4,74E-04 | 4,70E-04 | 5,46E-04 | 5,57E-04 | 4,24E-04 |

| | | | Transitions at C/G pairs | | | | | | FDR | | C/G Transition Frequency | | | Total mutation frequency | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | Ung$^{+/-}$Msh2$^{+/-}$Exp1 | | Ung$^{+/-}$Msh2$^{+/-}$Exp2 | | *Aicda*$^{-/-}$ | | | | Ung$^{+/-}$Msh2$^{+/-}$ | | *Aicda*$^{-/-}$ | Ung$^{+/-}$Msh2$^{+/-}$ | | Ung$^{+/-}$Msh2$^{-/-}$ | | Ung$^{+/-}$Msh2$^{+/-}$ | | Ung$^{-/-}$Msh2$^{-/-}$ | | *Aicda*$^{-/-}$ |
| **Gene** | **Ensembl ID** | **Coordinates** | **C/G transitions** | **C/G sequenced** | **C/G transitions** | **C/G sequenced** | **C/G transitions** | **C/G sequenced** | **Exp1** | **Exp2** | **Exp1** | **Exp2** | **Exp1** | **Exp1** | **Exp2** | **Exp1** | **Exp2** | **Exp1** | **Exp2** | **Exp1** | **Exp2** | **Exp1** |
| *Ttf1* | ENSMUSG00000026803 | 2:28915783-28916282 | 120 | 705047 | 220 | 773269 | 91 | 734424 | 4,55E-02 | 3,49E-11 | 1,70E-04 | 2,85E-04 | 1,24E-04 | 3,61E-04 | 3,07E-04 | 3,74E-04 | 3,54E-04 | 3,71E-04 | 3,17E-04 | 3,77E-04 | 4,05E-04 | 3,04E-04 |
| *U2af2* | ENSMUSG00000030435 | 7:5013784-5014283 | 100 | 355362 | 114 | 383400 | 55 | 367985 | 5,74E-04 | 8,83E-05 | 2,81E-04 | 2,97E-04 | 1,49E-04 | 3,46E-04 | 3,08E-04 | 3,61E-04 | 4,46E-04 | 3,86E-04 | 3,18E-04 | 4,31E-04 | 4,24E-04 | 3,32E-04 |
| *Uba3* | ENSMUSG00000030061 | 6:97155137-97155636 | 107 | 387904 | 107 | 436370 | 68 | 421010 | 1,87E-03 | 2,00E-02 | 2,76E-04 | 2,45E-04 | 1,62E-04 | 3,53E-04 | 3,41E-04 | 3,42E-04 | 4,32E-04 | 3,93E-04 | 3,79E-04 | 4,60E-04 | 3,85E-04 | 3,38E-04 |
| *Ubac2* | ENSMUSG00000041765 | 14:122277828-122278327 | 211 | 880732 | 205 | 897741 | 141 | 875519 | 9,40E-04 | 4,49E-03 | 2,40E-04 | 2,28E-04 | 1,61E-04 | 3,75E-04 | 3,26E-04 | 3,29E-04 | 3,84E-04 | 3,97E-04 | 3,67E-04 | 4,05E-04 | 3,83E-04 | 3,51E-04 |
| *Ubb* | ENSMUSG00000019505 | 11:62365006-62365505 | 174 | 664352 | 169 | 729561 | 124 | 708920 | 1,99E-03 | 4,07E-02 | 2,62E-04 | 2,32E-04 | 1,75E-04 | 4,85E-04 | 4,31E-04 | 4,55E-04 | 4,53E-04 | 4,47E-04 | 4,19E-04 | 4,90E-04 | 4,83E-04 | 4,56E-04 |
| *Ube2b* | ENSMUSG00000020390 | 11:51813469-51813968 | 182 | 692159 | 166 | 744358 | 99 | 702589 | 2,19E-06 | 1,04E-03 | 2,63E-04 | 2,23E-04 | 1,41E-04 | 6,75E-04 | 6,15E-04 | 7,05E-04 | 6,78E-04 | 6,26E-04 | 5,91E-04 | 7,21E-04 | 6,21E-04 | 6,21E-04 |
| *Ube2n* | ENSMUSG00000074781 | 10:94977796-94978295 | 122 | 492310 | 123 | 528919 | 60 | 499865 | 1,37E-05 | 8,63E-05 | 2,48E-04 | 2,33E-04 | 1,20E-04 | 4,17E-04 | 3,32E-04 | 3,71E-04 | 3,67E-04 | 3,85E-04 | 3,37E-04 | 4,09E-04 | 3,92E-04 | 3,03E-04 |
| *Ube4b* | ENSMUSG00000028960 | 4:148800241-148800740 | 197 | 841925 | 224 | 910180 | 138 | 878551 | 1,19E-03 | 1,27E-04 | 2,34E-04 | 2,46E-04 | 1,57E-04 | 3,69E-04 | 3,17E-04 | 3,97E-04 | 4,19E-04 | 3,59E-04 | 3,28E-04 | 4,14E-04 | 3,86E-04 | 3,44E-04 |
| *Ubox5* | ENSMUSG00000027300 | 2:130455223-130455722 | 253 | 899101 | 270 | 984227 | 132 | 923318 | 8,57E-10 | 2,37E-09 | 2,81E-04 | 2,74E-04 | 1,43E-04 | 4,19E-04 | 3,07E-04 | 3,47E-04 | 3,61E-04 | 3,41E-04 | 3,28E-04 | 4,67E-04 | 4,19E-04 | 3,32E-04 |
| *Ubtf-1* | ENSMUSG00000020923 | 11:102178102-102178601 | 176 | 430099 | 152 | 463949 | 69 | 426026 | 7,01E-11 | 3,57E-06 | 4,09E-04 | 3,28E-04 | 1,62E-04 | 3,96E-04 | 2,98E-04 | 3,53E-04 | 4,26E-04 | 3,54E-04 | 2,90E-04 | 5,86E-04 | 5,18E-04 | 3,68E-04 |
| *Ubtf-2* | ENSMUSG00000020923 | 11:102179911-102180410 | 156 | 492497 | 156 | 539203 | 78 | 489902 | 2,32E-06 | 5,63E-05 | 3,17E-04 | 2,89E-04 | 1,59E-04 | 4,71E-04 | 3,35E-04 | 3,68E-04 | 4,40E-04 | 3,48E-04 | 3,48E-04 | 4,70E-04 | 4,28E-04 | 3,09E-04 |
| *Uhmk1* | ENSMUSG00000026667 | 1:172145025-172145524 | 215 | 451892 | 252 | 498008 | 81 | 463773 | 5,66E-15 | 5,50E-18 | 4,76E-04 | 5,06E-04 | 1,75E-04 | 3,32E-04 | 5,87E-04 | 1,08E-03 | 7,51E-04 | 8,20E-04 | 1,29E-03 | 6,67E-04 | 7,05E-04 | 3,71E-04 |
| *UPGm13390-2* | ENSMUSG00000087656 | 2:7317326-7317825 | 273 | 554140 | 333 | 651951 | 227 | 620639 | 2,98E-03 | 3,91E-04 | 4,93E-04 | 5,11E-04 | 3,66E-04 | 5,26E-04 | 5,44E-04 | 8,12E-04 | 8,18E-04 | 5,69E-04 | 5,26E-04 | 7,89E-04 | 8,17E-04 | 5,34E-04 |
| *Vav1* | ENSMUSG00000034116 | 17:57418523-57419022 | 121 | 633878 | 149 | 678302 | 80 | 668214 | 3,81E-03 | 4,33E-05 | 1,91E-04 | 2,20E-04 | 1,20E-04 | 3,52E-04 | 3,26E-04 | 3,33E-04 | 3,48E-04 | 3,17E-04 | 3,29E-04 | 3,30E-04 | 3,67E-04 | 3,17E-04 |
| *Vav2* | ENSMUSG00000009621 | 2:27281846-27282345 | 218 | 515435 | 228 | 567378 | 94 | 535758 | 1,16E-12 | 1,79E-11 | 4,23E-04 | 4,02E-04 | 1,75E-04 | 3,72E-04 | 3,51E-04 | 3,72E-04 | 4,29E-04 | 3,45E-04 | 3,70E-04 | 5,40E-04 | 5,26E-04 | 3,58E-04 |
| *Vcl* | ENSMUSG00000021823 | 14:21748655-21749154 | 146 | 747402 | 176 | 825832 | 98 | 776499 | 2,69E-03 | 1,16E-04 | 1,95E-04 | 2,13E-04 | 1,26E-04 | 3,84E-04 | 3,26E-04 | 3,40E-04 | 3,60E-04 | 3,37E-04 | 2,97E-04 | 3,69E-04 | 3,72E-04 | 3,18E-04 |
| *Vdac1* | ENSMUSG00000020402 | 11:52174617-52175116 | 121 | 495642 | 140 | 533042 | 79 | 515183 | 4,04E-03 | 4,59E-04 | 2,44E-04 | 2,63E-04 | 1,53E-04 | 4,25E-04 | 3,62E-04 | 3,62E-04 | 3,79E-04 | 4,35E-04 | 3,62E-04 | 4,30E-04 | 4,03E-04 | 3,65E-04 |
| *Xbp1* | ENSMUSG00000020484 | 11:5420970-5421469 | 217 | 768683 | 148 | 817986 | 104 | 779896 | 7,65E-10 | 4,00E-02 | 2,82E-04 | 1,81E-04 | 1,33E-04 | 3,80E-04 | 3,47E-04 | 3,65E-04 | 3,84E-04 | 3,84E-04 | 3,48E-04 | 4,57E-04 | 3,65E-04 | 3,21E-04 |
| *Zc3h15* | ENSMUSG00000027091 | 2:83484735-83485234 | 118 | 604885 | 135 | 644471 | 78 | 626855 | 6,13E-03 | 8,61E-04 | 1,95E-04 | 2,09E-04 | 1,24E-04 | 3,70E-04 | 3,03E-04 | 3,76E-04 | 3,95E-04 | 3,09E-04 | 3,32E-04 | 3,69E-04 | 3,90E-04 | 3,25E-04 |
| *Zcchc7* | ENSMUSG00000035649 | 4:44769440-44769939 | 240 | 1068500 | 308 | 1144835 | 167 | 1100379 | 4,07E-04 | 9,33E-09 | 2,25E-04 | 2,69E-04 | 1,52E-04 | 3,58E-04 | 3,37E-04 | 3,52E-04 | 3,94E-04 | 3,66E-04 | 3,14E-04 | 4,07E-04 | 4,67E-04 | 3,48E-04 |

**Annex III.** List of genes predicted to be mutated by AID. Genes included in our SureSelect capture library are highlighted in green; genes validated by PCR-Seq are highlighted in orange.

| Gene | SPT5 (RPKM) | RNAP II (RPKM) |
|---|---|---|
| 09-002_145561_chr11 | 19,58 | 30,48 |
| 1110005A03Rik | 16,66 | 14,93 |
| 1110065P20Rik | 13,31 | 24,19 |
| 1500012F01Rik | 35,41 | 41,13 |
| 1700001P01Rik | 7,60 | 9,04 |
| 1700008J07Rik | 10,69 | 18,89 |
| 1700040I03Rik | 7,61 | 14,98 |
| 1810009A15Rik | 13,67 | 21,08 |
| 1810035L17Rik | 7,40 | 8,13 |
| 2010001M09Rik | 17,99 | 37,45 |
| 2310014L17Rik | 12,58 | 16,44 |
| 2310016E02Rik | 7,69 | 16,47 |
| 2310033P09Rik | 9,92 | 26,87 |
| 2310039H08Rik | 19,79 | 28,49 |
| 2700094K13Rik | 10,59 | 16,21 |
| 2810422O20Rik | 12,37 | 6,46 |
| 2810428I15Rik | 8,06 | 12,52 |
| 4930404N11Rik | 9,38 | 15,74 |
| 6230427J02Rik | 7,48 | 20,00 |
| 9130023H24Rik | 7,41 | 12,93 |
| A130010J15Rik | 7,70 | 10,11 |
| A830007P12Rik | 8,66 | 20,30 |
| Abhd11 | 7,58 | 15,92 |
| Actb | 24,97 | 64,52 |
| Actg1 | 8,27 | 23,97 |
| Agxt2l2 | 11,08 | 15,67 |
| AI413582 | 15,87 | 33,75 |
| Aldoa | 8,96 | 16,30 |
| Alkbh1 | 8,10 | 7,68 |
| Apex1 | 12,35 | 21,37 |
| Apobec3 | 10,67 | 26,17 |
| Arf6 | 11,02 | 18,95 |
| Arhgdia | 7,75 | 21,12 |
| Atf4 | 27,82 | 60,46 |
| Aup1 | 9,87 | 21,55 |
| Aurkaip1 | 7,88 | 16,74 |
| AY074887 | 8,26 | 15,96 |
| B2m | 14,74 | 23,02 |
| B3galt4 | 25,22 | 42,94 |
| B3galt6 | 7,92 | 12,87 |
| Banf1 | 11,85 | 23,96 |
| BC031181 | 9,06 | 11,25 |
| BC056474 | 8,47 | 12,46 |
| Bola1 | 10,56 | 20,67 |
| Bola2 | 11,96 | 25,33 |
| Brd2 | 17,98 | 39,92 |
| Bzrap1 | 7,21 | 9,12 |
| Calr | 11,22 | 21,11 |
| Ccdc17 | 7,46 | 16,10 |
| Ccdc84 | 7,78 | 9,08 |

| Gene | SPT5 (RPKM) | RNAP II (RPKM) |
|---|---|---|
| Ccnd2 | 10,25 | 25,29 |
| Ccnl1 | 8,03 | 9,70 |
| Cd37 | 7,16 | 24,05 |
| Cd52 | 9,32 | 26,43 |
| Cd68 | 20,30 | 40,32 |
| Cd69 | 17,26 | 7,76 |
| Cd74 | 31,76 | 64,77 |
| Cd83 | 9,94 | 28,52 |
| Cdt1 | 8,17 | 22,76 |
| Cfl1 | 14,13 | 25,21 |
| Chmp2a | 10,97 | 23,77 |
| Chrac1 | 9,43 | 16,93 |
| Cirbp | 9,29 | 16,72 |
| Clec2d | 19,53 | 22,32 |
| Clic1 | 7,89 | 16,22 |
| Cnbp | 11,13 | 13,08 |
| Coro1a | 15,77 | 44,71 |
| D230037D09Rik | 8,52 | 13,99 |
| D4Wsu53e | 10,78 | 19,06 |
| Dbil5 | 12,67 | 19,83 |
| Dbp | 7,98 | 11,68 |
| Ddit3 | 9,77 | 13,62 |
| Ddx5 | 14,25 | 17,62 |
| Dhps | 7,19 | 16,67 |
| Dusp2 | 13,93 | 26,70 |
| Eef1b2 | 7,92 | 15,72 |
| Eef2 | 9,67 | 29,31 |
| Eif1 | 16,79 | 30,62 |
| Eif4a1 | 15,05 | 29,29 |
| Eif4a2 | 18,77 | 23,05 |
| Eif4g2 | 10,29 | 11,01 |
| Eif5 | 8,32 | 11,45 |
| Eif5a | 14,40 | 24,19 |
| Erp29 | 10,62 | 16,28 |
| Exosc6 | 13,03 | 22,43 |
| Fam36a | 28,47 | 33,32 |
| Fam69b | 9,99 | 18,45 |
| Fau | 10,43 | 25,37 |
| Fbxl22 | 7,77 | 15,22 |
| Fkbp2 | 14,82 | 26,69 |
| Fth1 | 7,26 | 22,31 |
| Ftl1 | 11,04 | 21,11 |
| Fus | 8,46 | 9,50 |
| Gadd45b | 9,28 | 20,21 |
| Gadd45g | 7,87 | 52,74 |
| Gadd45gip1 | 7,29 | 12,29 |
| Gas5 | 44,66 | 37,62 |
| Gimap1 | 15,72 | 25,67 |
| Gimap4 | 11,70 | 14,68 |
| Gimap5 | 13,72 | 20,50 |

| Gene | SPT5 (RPKM) | RNAP II (RPKM) |
|---|---|---|
| Gimap6 | 8,91 | 11,99 |
| Gm11808 | 10,65 | 26,03 |
| Gm15772 | 9,23 | 13,64 |
| Gm5464 | 7,83 | 14,14 |
| Gm5774 | 22,38 | 36,24 |
| Gnb2l1 | 12,21 | 21,68 |
| Gng5 | 7,69 | 6,65 |
| Gnl3 | 11,48 | 14,05 |
| Gps2 | 18,98 | 32,52 |
| Gpx1 | 14,88 | 26,40 |
| Grap | 10,97 | 25,06 |
| Grcc10 | 26,17 | 38,52 |
| Gstt2 | 9,49 | 27,71 |
| H2-Aa | 21,55 | 32,01 |
| H2-Ab1 | 16,59 | 31,48 |
| H2-D1 | 7,73 | 18,42 |
| H2-Eb1 | 14,30 | 27,66 |
| H2-K1 | 8,59 | 18,46 |
| H2-Ke6 | 7,79 | 16,42 |
| H2afj | 12,82 | 16,99 |
| H2afx | 41,29 | 58,91 |
| H2afz | 12,33 | 17,55 |
| H3f3b | 25,04 | 26,99 |
| Hexim1 | 9,26 | 17,51 |
| Higd2a | 8,97 | 14,40 |
| Hirip3 | 8,61 | 15,55 |
| Hist1h1a | 40,08 | 46,78 |
| Hist1h1b | 51,77 | 68,68 |
| Hist1h1c | 69,40 | 84,34 |
| Hist1h1d | 35,85 | 38,24 |
| Hist1h1e | 45,53 | 56,45 |
| Hist1h2ab | 16,71 | 21,05 |
| Hist1h2ac | 18,15 | 17,27 |
| Hist1h2ad | 9,50 | 14,58 |
| Hist1h2ae | 43,69 | 30,53 |
| Hist1h2af | 12,27 | 11,70 |
| Hist1h2ag | 37,44 | 32,66 |
| Hist1h2ah | 28,31 | 30,29 |
| Hist1h2ai | 24,98 | 28,07 |
| Hist1h2ak | 52,73 | 41,50 |
| Hist1h2an | 16,65 | 19,46 |
| Hist1h2bb | 18,95 | 25,00 |
| Hist1h2bf | 17,35 | 14,99 |
| Hist1h2bh | 23,28 | 21,22 |
| Hist1h2bj | 36,67 | 32,45 |
| Hist1h2bk | 25,56 | 24,99 |
| Hist1h2bl | 26,83 | 30,78 |
| Hist1h2bm | 10,86 | 17,60 |
| Hist1h2bn | 49,05 | 42,73 |
| Hist1h2bp | 18,47 | 11,67 |

| Gene | SPT5 (RPKM) | RNAP II (RPKM) |
|---|---|---|
| Hist1h3a | 20,67 | 22,89 |
| Hist1h3g | 44,23 | 40,49 |
| Hist1h3h | 33,45 | 28,03 |
| Hist1h3i | 13,48 | 16,19 |
| Hist1h4a | 47,65 | 55,34 |
| Hist1h4b | 16,41 | 17,00 |
| Hist1h4c | 19,15 | 28,51 |
| Hist1h4d | 22,67 | 32,09 |
| Hist1h4f | 18,47 | 25,92 |
| Hist1h4h | 26,01 | 31,57 |
| Hist1h4i | 47,90 | 35,79 |
| Hist1h4j | 13,59 | 20,81 |
| Hist1h4k | 13,63 | 18,07 |
| Hist2h2ab | 34,81 | 30,95 |
| Hist2h2ac | 36,42 | 42,24 |
| Hist2h2be | 22,86 | 23,98 |
| Hist2h4 | 38,56 | 50,38 |
| Hist3h2a | 22,27 | 21,38 |
| Hist4h4 | 43,56 | 48,86 |
| Hmbs | 7,25 | 7,70 |
| Hmga1 | 10,83 | 22,13 |
| Hmgb1 | 8,12 | 11,93 |
| Hmgb2 | 9,87 | 15,66 |
| Hmgn2 | 7,59 | 12,54 |
| Hnrnpa0 | 13,49 | 26,48 |
| Hnrnpa2b1 | 20,35 | 18,28 |
| Hnrnpa3 | 8,76 | 9,10 |
| Hnrnpab | 13,42 | 22,68 |
| Hnrnpf | 12,08 | 13,18 |
| Hnrnph1 | 7,40 | 10,41 |
| Hnrnpk | 8,86 | 13,38 |
| Hnrnpl | 8,91 | 20,83 |
| Hnrnpu | 13,15 | 14,59 |
| Hnrpdl | 12,18 | 17,43 |
| Hsp90ab1 | 14,13 | 25,11 |
| Hspa5 | 17,02 | 25,55 |
| Hspa8 | 16,08 | 22,58 |
| Hspe1 | 8,11 | 14,18 |
| Htra2 | 10,19 | 20,05 |
| Id3 | 7,95 | 22,18 |
| Idh3b | 17,01 | 22,55 |
| Ier2 | 13,55 | 31,89 |
| Ier5 | 14,92 | 24,93 |
| Igg1 | 15,54 | 67,39 |
| Igm | 37,83 | 97,18 |
| Il4i1 | 7,14 | 54,63 |
| Il4ra | 11,28 | 48,77 |
| Imp3 | 8,15 | 10,43 |
| Ing1 | 8,92 | 20,00 |
| Insm1 | 8,31 | 17,30 |

| Gene | SPT5 (RPKM) | RNAP II (RPKM) |
|---|---|---|
| Jun | 13,53 | 26,07 |
| Junb | 16,99 | 52,00 |
| Jund | 16,29 | 39,50 |
| Lcmt2 | 8,72 | 13,44 |
| Limd2 | 10,27 | 25,86 |
| Lipt2 | 7,48 | 11,87 |
| Lsmd1 | 8,33 | 14,73 |
| Ly6e | 18,75 | 53,15 |
| Lyl1 | 11,19 | 30,77 |
| Mageb3 | 7,14 | 12,74 |
| Manf | 11,69 | 20,05 |
| Mat2a | 15,90 | 15,17 |
| Maz | 7,76 | 16,97 |
| Mbd3 | 8,77 | 18,59 |
| Md1 | 17,37 | 28,93 |
| Mgat2 | 12,06 | 20,89 |
| Mocs3 | 8,81 | 17,50 |
| Mogs | 7,22 | 14,15 |
| Mrp63 | 9,46 | 16,06 |
| Mrpl41 | 10,08 | 21,67 |
| Mrpl49 | 8,19 | 21,71 |
| Mrpl51 | 9,38 | 15,39 |
| Mrpl53 | 9,23 | 17,92 |
| Mrps18b | 10,54 | 16,00 |
| Mrps34 | 9,37 | 22,94 |
| Myc | 9,91 | 33,37 |
| Myl6 | 7,27 | 14,46 |
| Naca | 7,18 | 10,90 |
| Nanos1 | 10,45 | 6,58 |
| Nd | 23,38 | 20,48 |
| Ndufa2 | 8,22 | 13,58 |
| Ndufa4l2 | 9,46 | 12,01 |
| Ndufaf3 | 18,66 | 39,53 |
| Ndufb10 | 14,02 | 22,98 |
| Nfkbia | 12,84 | 33,71 |
| Nme3 | 7,97 | 17,22 |
| Nop10 | 8,35 | 11,30 |
| Nop56 | 18,67 | 27,69 |
| Npm1 | 14,79 | 9,92 |
| Nrp | 15,67 | 19,13 |
| Nt5c | 8,14 | 15,29 |
| Oasl1 | 11,58 | 22,55 |
| Obfc2a | 10,70 | 17,59 |
| Pa2g4 | 7,34 | 10,40 |
| Pcbp1 | 15,33 | 36,40 |
| Pcna | 12,16 | 14,84 |
| Pex12 | 10,22 | 14,09 |
| Pfn1 | 20,74 | 45,00 |
| Pgp | 10,42 | 24,54 |
| Pigw | 9,35 | 13,94 |

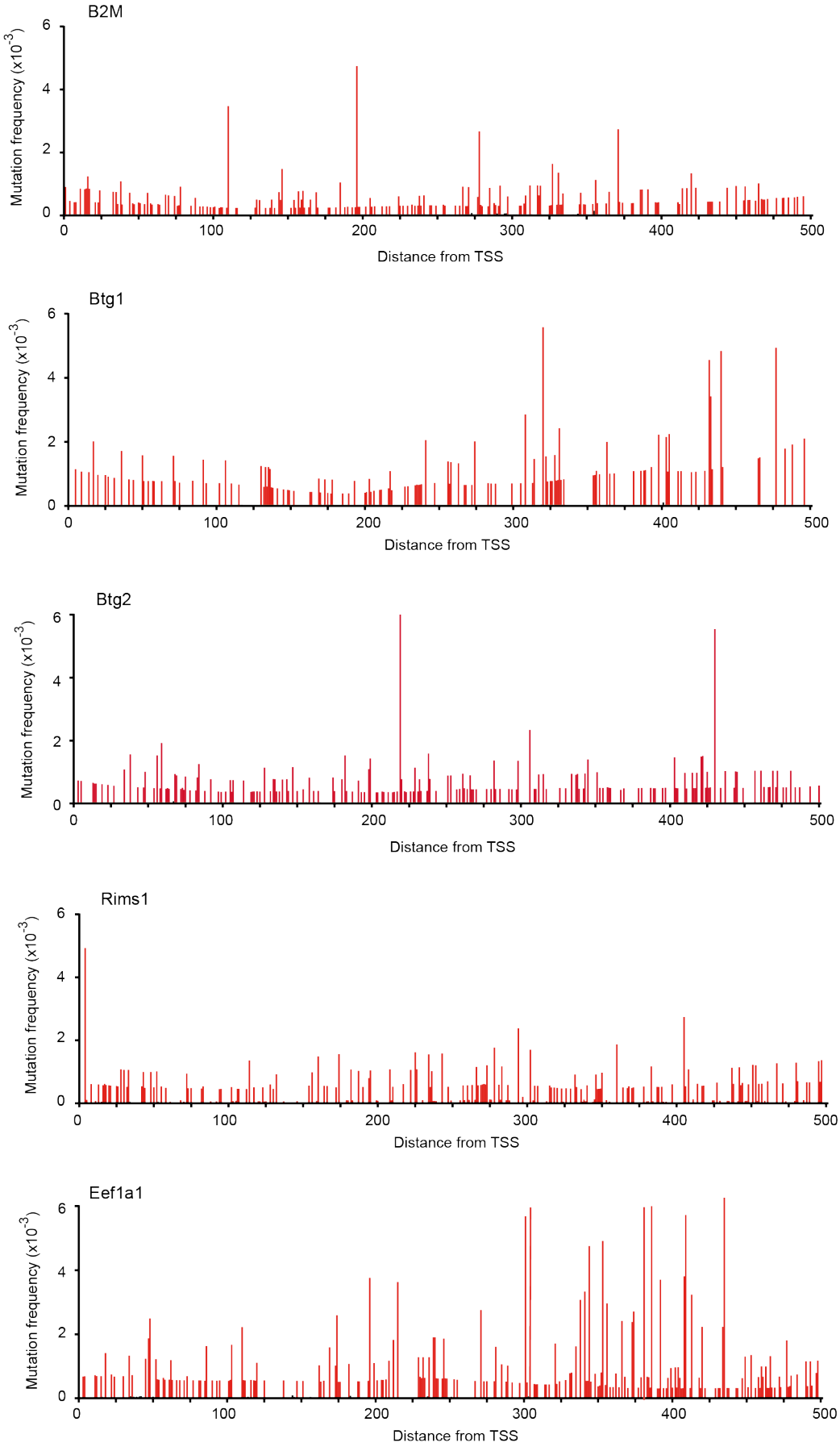
| Gene | SPT5 (RPKM) | RNAP II (RPKM) |
|---|---|---|
| Pim1 | 20,40 | 63,74 |
| Plekhj1 | 15,03 | 27,79 |
| Pnrc2 | 12,39 | 14,93 |
| Polr1c | 7,42 | 12,81 |
| Pop7 | 12,72 | 24,12 |
| Pou2af1 | 7,16 | 19,86 |
| Ppan | 8,37 | 20,65 |
| Ppia | 19,89 | 30,52 |
| Ppp1r10 | 11,30 | 16,08 |
| Ppp1r14b | 8,32 | 17,43 |
| Ptbp1 | 7,65 | 21,51 |
| Ptma | 13,28 | 38,45 |
| Ptprcap | 12,64 | 30,33 |
| Ptrh2 | 9,07 | 8,01 |
| Purb | 8,19 | 16,44 |
| Pycard | 11,13 | 13,89 |
| Rabggtb | 16,28 | 17,44 |
| Rad23a | 12,75 | 16,66 |
| Ran | 8,06 | 15,63 |
| Rbm3 | 13,16 | 17,48 |
| Rbm39 | 10,90 | 7,08 |
| Rdm1 | 12,05 | 14,33 |
| Refbp2 | 42,42 | 38,07 |
| Rfc4 | 7,64 | 8,04 |
| Rnf167 | 18,39 | 36,87 |
| Rnu11 | 24,87 | 41,70 |
| Rnu12 | 27,88 | 43,91 |
| Romo1 | 14,39 | 17,48 |
| Rpl10 | 14,94 | 21,34 |
| Rpl10a | 17,82 | 35,83 |
| Rpl11 | 9,59 | 20,72 |
| Rpl13 | 11,48 | 24,12 |
| Rpl13a | 18,13 | 36,79 |
| Rpl14 | 10,62 | 17,08 |
| Rpl17 | 10,54 | 19,24 |
| Rpl18 | 12,97 | 26,39 |
| Rpl19 | 10,76 | 18,43 |
| Rpl23 | 11,89 | 17,51 |
| Rpl23a | 10,85 | 13,35 |
| Rpl26 | 9,17 | 13,56 |
| Rpl27 | 8,01 | 15,24 |
| Rpl27a | 18,74 | 25,72 |
| Rpl28 | 13,88 | 29,86 |
| Rpl29 | 8,35 | 16,42 |
| Rpl3 | 27,13 | 39,31 |
| Rpl30 | 11,19 | 15,61 |
| Rpl34 | 11,80 | 17,79 |
| Rpl35 | 8,37 | 19,44 |
| Rpl35a | 10,80 | 16,48 |
| Rpl36 | 11,78 | 26,09 |

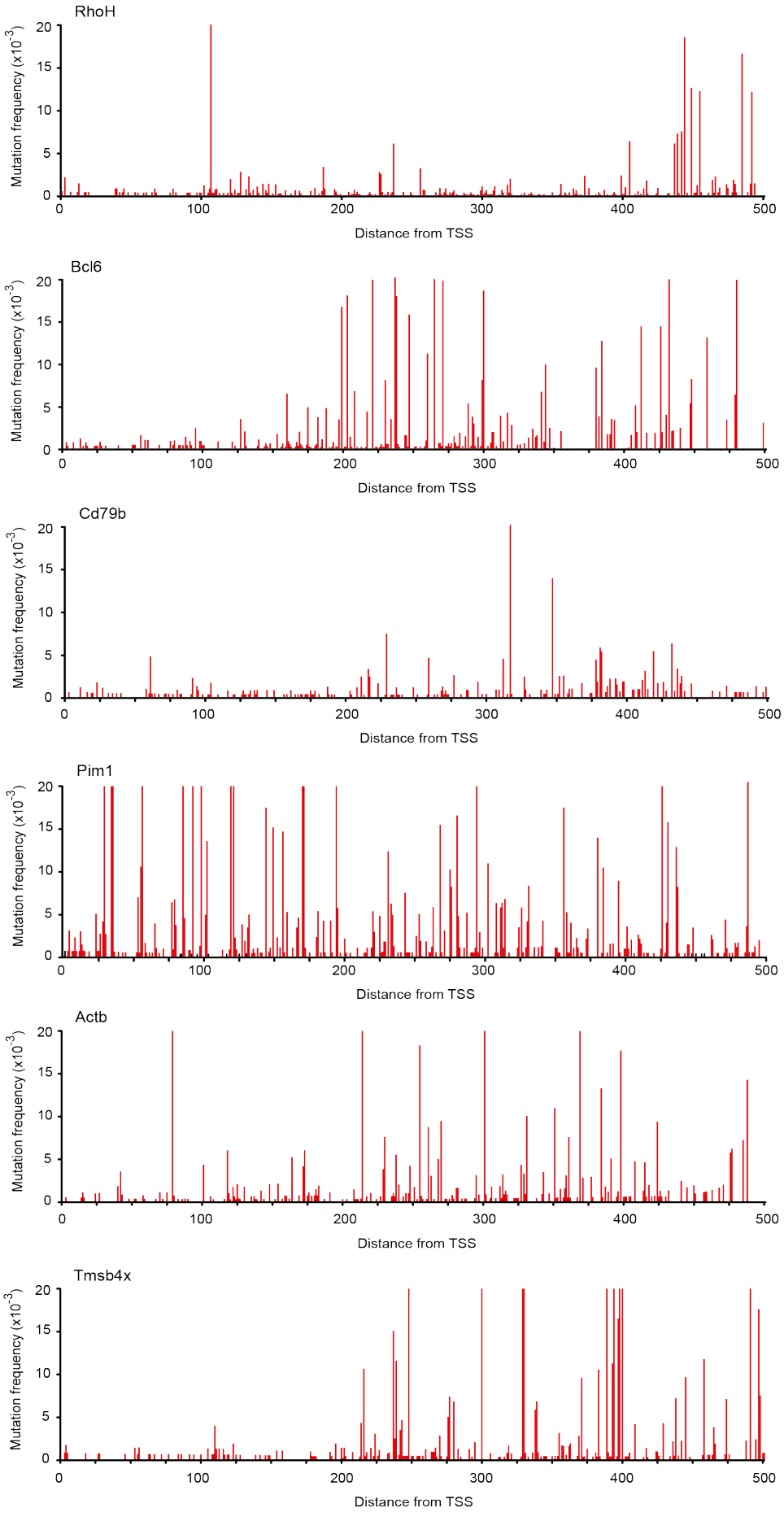
| Gene | SPT5 (RPKM) | RNAP II (RPKM) |
|---|---|---|
| Rpl36a | 8,99 | 14,71 |
| Rpl36al | 14,65 | 29,99 |
| Rpl37 | 8,41 | 22,15 |
| Rpl38 | 13,76 | 22,55 |
| Rpl4 | 12,46 | 15,70 |
| Rpl41 | 21,60 | 28,13 |
| Rpl5 | 9,41 | 12,68 |
| Rpl7a | 11,59 | 27,21 |
| Rpl8 | 11,39 | 24,43 |
| Rpl9 | 10,76 | 20,83 |
| Rplp0 | 11,26 | 18,41 |
| Rplp1 | 14,42 | 25,95 |
| Rplp2 | 11,06 | 17,91 |
| Rps10 | 7,37 | 20,60 |
| Rps11 | 23,77 | 44,19 |
| Rps12 | 19,45 | 29,16 |
| Rps13 | 10,43 | 17,06 |
| Rps14 | 8,93 | 17,55 |
| Rps15 | 10,78 | 26,18 |
| Rps15a | 10,49 | 14,75 |
| Rps17 | 8,66 | 21,00 |
| Rps18 | 13,53 | 25,95 |
| Rps19 | 9,77 | 21,92 |
| Rps2 | 21,38 | 35,04 |
| Rps20 | 16,70 | 27,60 |
| Rps21 | 19,04 | 28,08 |
| Rps24 | 8,04 | 13,73 |
| Rps25 | 9,86 | 17,42 |
| Rps27 | 15,76 | 20,64 |
| Rps27a | 15,30 | 19,80 |
| Rps28 | 14,13 | 22,82 |
| Rps29 | 7,30 | 16,28 |
| Rps3 | 8,56 | 15,35 |
| Rps3a | 8,31 | 14,18 |
| Rps4x | 10,97 | 14,80 |
| Rps5 | 12,97 | 23,55 |
| Rps6 | 11,50 | 20,70 |
| Rps8 | 13,67 | 28,04 |
| Rps9 | 18,72 | 33,62 |
| Rpsa | 12,86 | 23,68 |
| Rraga | 13,17 | 18,87 |
| Rrm2 | 10,26 | 19,22 |
| Sac3d1 | 7,60 | 11,42 |
| Scand1 | 13,63 | 25,40 |
| Sdf2l1 | 11,80 | 23,61 |
| Sdhaf1 | 9,05 | 14,55 |
| Sdr39u1 | 7,66 | 15,05 |
| Serinc3 | 7,81 | 11,91 |
| Serp1 | 11,33 | 14,82 |
| Setd6 | 7,98 | 11,59 |

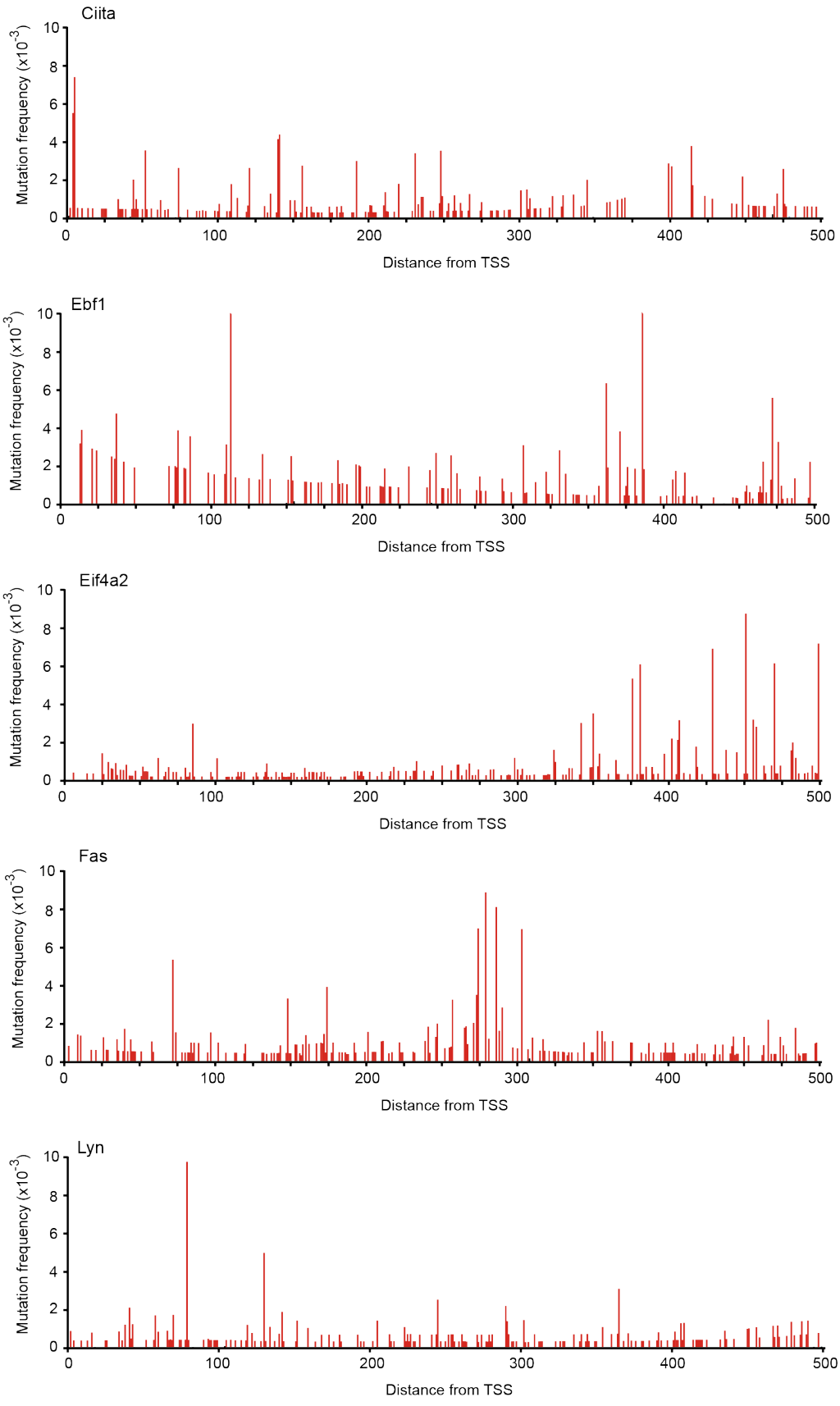
| Gene | SPT5 (RPKM) | RNAP II (RPKM) |
|---|---|---|
| Sfpq | 10,80 | 12,69 |
| Sfrs1 | 14,03 | 13,77 |
| Sfrs13a | 7,18 | 6,89 |
| Sfrs2 | 23,29 | 34,35 |
| Sfrs3 | 11,08 | 12,67 |
| Sfrs5 | 9,37 | 15,78 |
| Sfrs6 | 13,54 | 14,14 |
| Sfrs7 | 12,95 | 12,68 |
| Shisa5 | 14,84 | 41,21 |
| Slc25a11 | 7,90 | 15,71 |
| Slc35b2 | 21,27 | 33,93 |
| Slc39a7 | 9,10 | 18,52 |
| Snapc5 | 14,05 | 15,56 |
| Snora62 | 13,30 | 22,34 |
| Snora64 | 45,15 | 66,25 |
| Snora68 | 8,32 | 25,41 |
| Snora70 | 11,09 | 10,54 |
| Snora74a | 17,06 | 16,19 |
| Snord22 | 32,29 | 22,02 |
| Snord32a | 18,25 | 26,92 |
| Snord33 | 15,72 | 33,09 |
| Snord34 | 14,13 | 39,21 |
| Snord35a | 12,79 | 41,41 |
| Snord35b | 11,37 | 25,04 |
| Snord82 | 13,62 | 13,20 |
| Snord87 | 23,51 | 26,63 |
| Snx5 | 12,60 | 10,70 |
| Socs1 | 14,70 | 54,06 |
| Sra1 | 7,33 | 9,95 |
| Srm | 8,21 | 15,40 |
| Surf1 | 11,54 | 21,03 |
| Syngr2 | 9,11 | 37,94 |
| Syvn1 | 11,70 | 31,39 |
| Taf10 | 8,29 | 15,55 |
| Taf1d | 13,45 | 13,84 |
| Tardbp | 10,69 | 10,85 |
| Tceb2 | 8,30 | 10,34 |
| Tcta | 8,02 | 16,94 |
| Terc | 9,15 | 17,04 |
| Tgif1 | 10,01 | 13,67 |
| Timm13 | 8,50 | 18,26 |
| Timm8b | 11,09 | 18,07 |
| Tk1 | 15,83 | 44,20 |
| Tlcd1 | 22,31 | 32,21 |
| Tmem126a | 9,64 | 6,41 |
| Tmem179b | 14,83 | 29,41 |
| Tmem55b | 7,79 | 16,79 |
| Tmsb10 | 8,58 | 17,01 |
| Tmsb4x | 15,15 | 25,42 |
| Tnf | 9,41 | 26,33 |

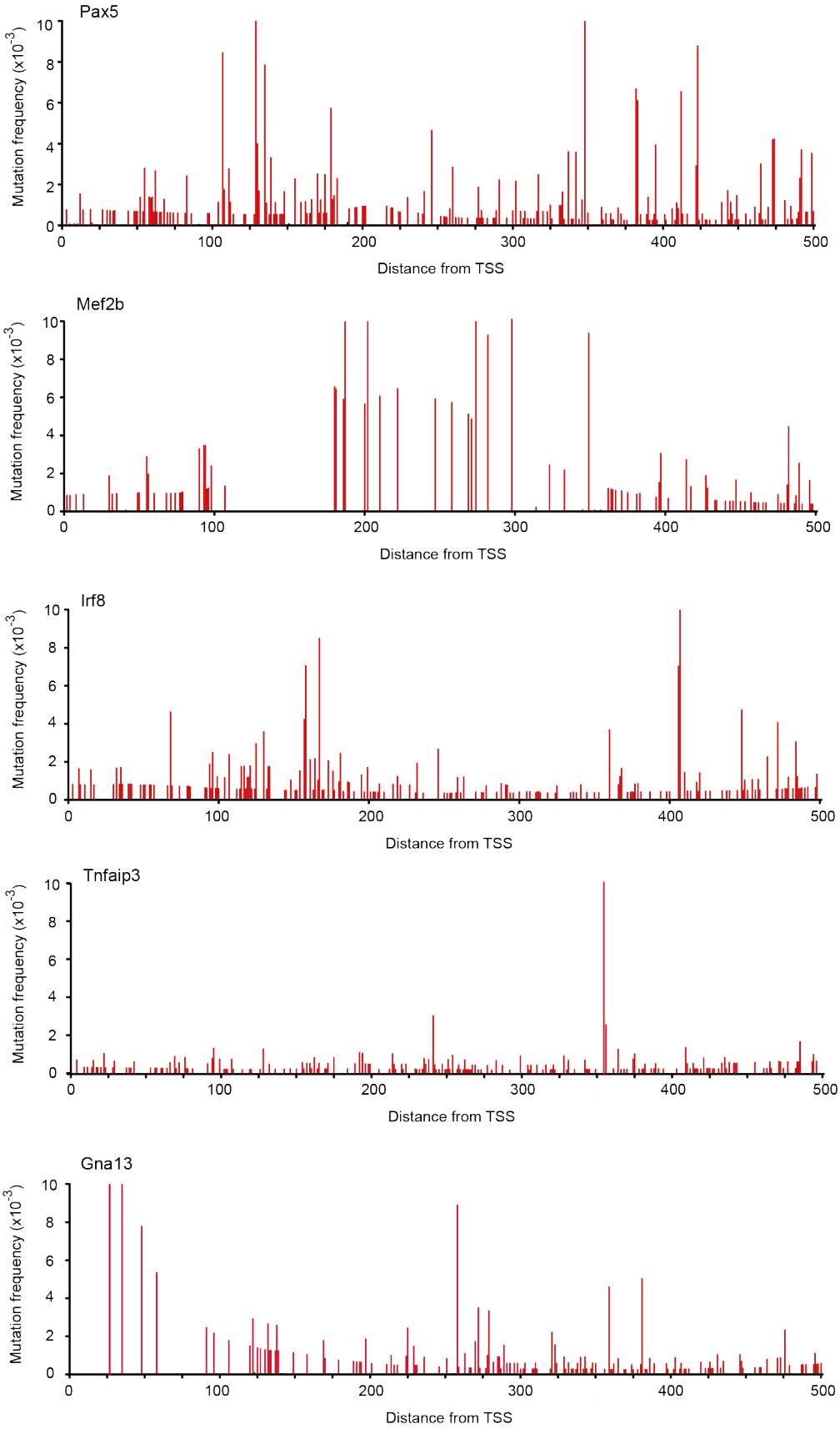
| Gene | SPT5 (RPKM) | RNAP II (RPKM) |
|---|---|---|
| Tnfrsf13b | 10,27 | 27,45 |
| Tnfrsf13c | 7,11 | 25,23 |
| Trex1 | 26,68 | 65,80 |
| Trim41 | 9,78 | 16,12 |
| Trmt112 | 9,23 | 23,70 |
| Trmt2a | 14,10 | 25,34 |
| Tspan31 | 10,23 | 9,77 |
| Tssc4 | 8,35 | 22,25 |
| Tssk6 | 11,67 | 26,23 |
| Ttpal | 8,91 | 11,12 |
| Tuba1b | 13,34 | 20,64 |
| Tubb2c | 10,65 | 23,17 |
| Tubb5 | 10,87 | 23,54 |
| Txnip | 8,37 | 9,94 |
| Uba52 | 10,63 | 25,93 |
| Ubb | 11,41 | 19,46 |
| Ubc | 10,54 | 20,56 |
| Ucp2 | 10,84 | 23,40 |
| Ufsp1 | 11,38 | 25,81 |
| Uqcrq | 7,08 | 11,52 |
| Utp3 | 9,02 | 15,73 |
| Wdr38 | 11,63 | 21,55 |
| Wdr5b | 7,59 | 11,21 |
| Xbp1 | 9,34 | 41,90 |
| Zbtb32 | 12,34 | 18,24 |
| Zc3h10 | 14,90 | 22,80 |
| Zfp207 | 7,59 | 6,05 |
| Zfp36 | 10,13 | 36,28 |
| Zfp36l1 | 8,06 | 21,67 |
| Zfp36l2 | 7,41 | 18,62 |
| Zfp513 | 9,85 | 26,87 |
| Znhit2 | 14,00 | 29,19 |

**Annex IV.** Per nucleotide mutation frequency of the 21 AID targets recurrently found mutated in human DLBCL tumors as measured in *Ung*$^{-/-}$*Msh2*$^{-/-}$ mice. Background, quantified as the mutation frequency found in each nucleotide in *Aicda*$^{-/-}$ mice, was subtracted before plotting.




RhoH
Bcl6
Cd79b
Pim1
Actb
Tmsb4x
Mutation frequency (x10-3)
Distance from TSS
0
5
10
15
20
100
200
300
400
500


Ciita
Mutation frequency (x10-3)
Distance from TSS
Ebf1
Mutation frequency (x10-3)
Distance from TSS
Eif4a2
Mutation frequency (x10-3)
Distance from TSS
Fas
Mutation frequency (x10-3)
Distance from TSS
Lyn
Mutation frequency (x10-3)
Distance from TSS


Pax5
Mutation frequency (x10-3)
Distance from TSS
Mef2b
Mutation frequency (x10-3)
Distance from TSS
Irf8
Mutation frequency (x10-3)
Distance from TSS
Tnfaip3
Mutation frequency (x10-3)
Distance from TSS
Gna13
Mutation frequency (x10-3)
Distance from TSS

# PUBLICATIONS

# PUBLICATIONS

Main publication:

Álvarez-Prado ÁF, Pérez-Durán P, Pérez-García A, Benguria A, Torroja C, de Yébenes VG, Ramiro AR. *A broad atlas of somatic hypermutation allows prediction of activation-induced deaminase targets.* J Exp Med. 2018 Mar 5;215(3):761-771

Collaborations:

Pérez-García A*, Marina-Zárate E*, Álvarez-Prado ÁF, Ligos JM, Galjart N, Ramiro AR. *CTCF orchestrates the germinal centre transcriptional program and prevents premature plasma cell differentiation.* Nat Comm, 2017 Jul 5;8:16067.

Bartolomé-Izquierdo N*, de Yébenes VG*, Álvarez-Prado, AF, Mur SM, Lopez Del Olmo JA, Roa S, Vazquez J, Ramiro AR. *miR-28 regulates the germinal center reaction and blocks tumor growth in preclinical models of non-Hodgkin lymphoma.* Blood, 2017 Apr 27;129(17):2408-2419.

Marín AV, Jiménez-Reinoso A, Briones AC, Muñoz-Ruiz M; Aydogmus C, Pasick LJ, Couso, J, Mazariegos MS, Álvarez-Prado AF, (…) Regueiro JR Garcillán, B. *Primary T-cell immunodeficiency with functional revertant somatic mosaicism in CD247.* J Allergy and Clinical Immunology, 2017 Jan;139(1):347-349.

Published Online: 26 January, 2018 | Supp Info: http://doi.org/10.1084/jem.20171738


JEM



# A broad atlas of somatic hypermutation allows prediction of activation-induced deaminase targets

Ángel F. Álvarez-Prado,[1] Pablo Pérez-Durán,[1] Arantxa Pérez-García,[1] Alberto Benguria,[2] Carlos Torroja,[3] Virginia G. de Yébenes,[1] and Almudena R. Ramiro[1]

[1]B Cell Biology Lab, [2]Genomics Unit, and [3]Bioinformatics Unit, Centro Nacional de Investigaciones Cardiovasculares, Madrid, Spain

**Activation-induced deaminase (AID) initiates antibody diversification in germinal center (GC) B cells through the deamination of cytosines on immunoglobulin genes. AID can also target other regions in the genome, triggering mutations or chromosome translocations, with major implications for oncogenic transformation. However, understanding the specificity of AID has proved extremely challenging. We have sequenced at very high depth >1,500 genomic regions from GC B cells and identified 275 genes targeted by AID, including 30 of the previously known 35 AID targets. We have also identified the most highly mutated hotspot for AID activity described to date. Furthermore, integrative analysis of the molecular features of mutated genes coupled to machine learning has produced a powerful predictive tool for AID targets. We also have found that base excision repair and mismatch repair back up each other to faithfully repair AID-induced lesions. Finally, our data establish a novel link between AID mutagenic activity and lymphomagenesis.**

Journal of Experimental Medicine

## INTRODUCTION

Activation-induced deaminase (AID) is a crucial enzyme for the immune response because it generates high-affinity and switched antibodies in germinal center (GC) B cells by somatic hypermutation (SHM) and class switch recombination (CSR; Muramatsu et al., 2000; Revy et al., 2000). AID initiates SHM and CSR through the deamination of deoxycytidine residues into deoxyuridines on the DNA of Ig genes (Muramatsu et al., 2000; Petersen-Mahrt et al., 2002; Di Noia and Neuberger, 2007; Stavnezer et al., 2008). The resulting U:G mismatch can be alternatively recognized and processed by base excision repair (BER) or mismatch repair (MMR) pathways, leading either to point mutations, in the case of SHM, or to double-strand breaks (DSBs) followed by a recombination reaction, in the case of CSR (Di Noia and Neuberger, 2007; Stavnezer et al., 2008; Reynaud et al., 2009; Methot and Di Noia, 2017). Although AID activity has a strong preference for Ig genes, it can also target other genes, giving rise to point mutations (Shen et al., 1998; Pasqualucci et al., 2001; Liu et al., 2008) or oncogenic chromosome translocations (TCs; Ramiro et al., 2004, 2006; Robbiani et al., 2008). Understanding AID specificity, or targeting, has been hindered by the technical challenge of detecting AID-induced mutations, which occur at very low frequencies. Here, we have used next generation sequencing to directly measure raw AID mutational activity on a broad representation of the genome and thus gather conclusions on AID specificity, DNA repair, and lymphomagenesis.

Correspondence to Almudena R. Ramiro: aramiro@cnic.es

P. Pérez-Durán's present address is Institute for Cancer Genetics, Columbia University, New York, NY.

A. Pérez-García's present address is Beatson Institute for Cancer Research, Glasgow, Scotland, UK.

## RESULTS AND DISCUSSION

### Capture-based deep sequencing allows high-throughput identification of AID targets

To explore the scope of AID-induced mutations at a high-throughput scale, we designed a capture library against 1,588 regions corresponding to 1,379 different genes as a representation of the B cell genome (Table S1; see Design of DNA capture library in the Materials and methods). Genomic DNA from GC B cells was isolated, captured, and deep sequenced (Fig. S1, A and B). We made use of a mouse model deficient for both BER and MMR pathways ($Ung^{-/-}Msh2^{-/-}$ mice). In the absence of BER and MMR, AID-induced U:G mismatches remained unprocessed and were replicated over, thus leaving behind almost solely C→T and G→A transitions, the footprint of AID deamination events on DNA (Rada et al., 2004; Methot and Di Noia, 2017). This approach allowed an extremely efficient enrichment and sequencing depth (Fig. S1, A and B). We found a set of 291 genomic regions (corresponding to 275 different genes) that were reproducibly mutated in $Ung^{-/-}Msh2^{-/-}$ GC B cells when compared with $Aicda^{-/-}$ GC B cells ($q \leq 0.05$; Fig. 1 A; Fig. S1, C–E; and Table S2; representative targets were validated by Sanger sequencing; Fig. 1 B and Table S3). Importantly, the 275-gene target collection included 30 of the 35 previously known AID targets, such as *Bcl6*, *Pim1*, *RhoH*, *Pax5*, and *Cd83* (Fig. 1 C and Table S2; Pasqualucci et al., 2001; Liu et al., 2008; Methot and Di Noia, 2017). Mutations detected in the 291-target regions strongly accumulated in AID mutational hotspots (WRCY/RGYW; underlined letters specify deaminated nucleotides; W = A/T; R = A/G; Y = C/T; Fig. 1 D; Rogozin and Kolchanov, 1992). Finally, we found that our 275-target set in-

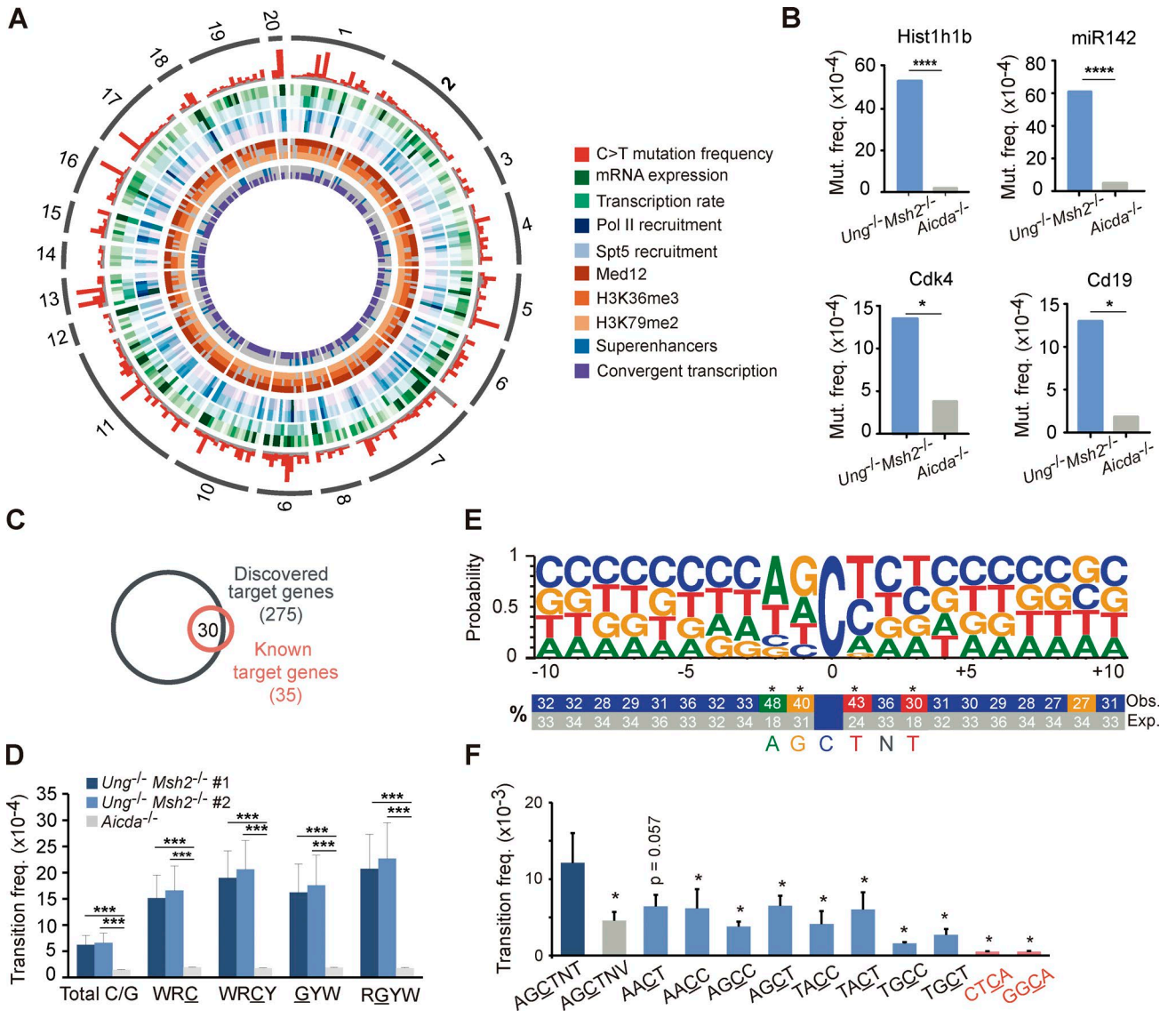



Figure 1. **High-throughput analysis of AID-induced mutations.** DNA from Peyer's patch GC B cells was captured with a probe library for 1,588 genomic regions (Table S1) and deep sequenced. AID targets were identified as those regions accumulating significantly more C→T transition mutations in *Ung*$^{-/-}$*Msh2*$^{-/-}$ than in *Aicda*$^{-/-}$ mice (Table S2; FDR ≤0.05, one-tail Fisher test and Benjamini-Hochberg correction; two independent experiments; see Materials and methods). **(A)** Circos plot representation of the AID targets identified in this study and their associated molecular features. The outer ring shows chromosome location and is followed by C→T transition mutation frequency in *Ung*$^{-/-}$*Msh2*$^{-/-}$ (red) and *Aicda*$^{-/-}$ (gray) mice. **(B)** Validation of representative AID targets by Sanger sequencing (one-tail Fisher test; Table S3). **(C)** Overlap between the targets discovered in this study and previously reported AID targets. **(D)** Mean transition frequency in total C/G nucleotides and in C/G within WRC(Y)/(R)GYW hotspots (W = A/T; R = G/A; Y = C/T) of the 291 AID targets (two-tailed Student's *t* test; two independent experiments). **(E)** Logo representation of the sequence context of mutated cytosines (mutation frequency ≥4 × 10$^{-3}$). Statistically significant enrichment of nucleotides surrounding the mutated C is indicated (*, FDR ≤10$^{-3}$, one-tail Fisher test and Bonferroni correction; see Materials and methods), and numbers indicate percentages. **(F)** Mean mutation frequency of cytosines within the indicated motifs (dark blue bar, newly identified hotspot; gray bar, control motif for newly identified hotspot; light blue bars, WRCY hotspots; red bars, random four-nucleotide motifs; two-tailed Mann-Whitney test). *, P ≤ 0.05; ***, P < 10$^{-3}$; ****, P < 10$^{-4}$. Error bars depict SEM.

cluded a big proportion of genes subject to DSBs or chromosome TCs (Fig. S1 F; Chiarle et al., 2011; Klein et al., 2011; Staszewski et al., 2011; Qian et al., 2014; Dong et al., 2015). Thus, our deep sequencing approach has allowed the discovery of an unprecedented, massive collection of AID targets.

## Identification of AGCTNT as a novel AID hotspot

To gain insights into the local sequence preference of AID, we first analyzed the mean mutation frequency at individual WRCY/RGYW hotspots across all 291 AID targets and found a wide range of mutability, with AACT and AGCT

as the top mutated hotspots in both strands of DNA, which may reflect an intrinsic preference for AID deaminase activity. Next, we performed an unbiased analysis of the sequence context of mutated cytosines. We found that A, G, and T nucleotides were the preferred nucleotides at −2, −1, and +1 positions (Pérez-Durán et al., 2012; Wei et al., 2015; Yeap et al., 2015), respectively, but we further uncovered a significant preference for T at +3 (Fig. 1 E and Fig. S2). Indeed, cytosines lying at the AGCTNT motif were significantly more mutated than those in AGCTNV (where V is A, C, or G) or than other WRCY/RGYW hotspots (Fig. 1 F and Fig. S2, A and B). Thus, our study has revealed AGCTNT as a novel and the most highly mutated AID hotspot identified so far.

### Prediction of AID targets

Using the uniquely large set of AID-mutated genes identified in this study, we performed a comprehensive analysis of molecular features that associate with SHM, including transcription, epigenetic marks, and regulatory sequences (Fig. 1 A; Storb, 2014; Methot and Di Noia, 2017). We first observed that transcription levels and transcription rates are significantly higher in AID targets than in nontargets and that this difference is even higher for highly mutated targets (Fig. 2 A). We also found that RNAPolII and the stalling factor Spt5, previously described to associate with AID (Nambu et al., 2003; Pavri et al., 2010), show higher binding density within AID mutational targets (Fig. 2 B). Likewise, AID targets were enriched in marks of active enhancers and transcriptional elongation, such as Med12, H3K36me3, and H3K79me2 (Fig. 2 C). Finally, we found that primary AID targeting, as measured by AID mutations in the absence of repair, also focuses preferentially in the vicinity of superenhancers (Fig. 2 D) and in regions subject to convergent transcription (Fig. 2 E; Meng et al., 2014; Qian et al., 2014). Together, our mutagenesis study shows that several mechanisms linked to transcription are critical for AID activity, as suggested in previous studies (Nambu et al., 2003; Pavri et al., 2010; Meng et al., 2014; Qian et al., 2014; Wang et al., 2014). Our data also indicate that AID targeting cannot be defined by any of these features alone. To approach whether a combination of these molecular features could be used to predict AID targeting, we developed a prediction model using a machine-learning algorithm, fed with the collection of genes analyzed here together with the set of molecular features described in Fig. 2 (A–E) (Fig. S3, A and B; see Machine learning to predict AID targets in the Materials and methods for details). We found that a combination of high-density RNAPolII and Spt5 binding, found in 2.3% of genes in the whole genome (Fig. S3 B), predicts AID specificity with 77% probability ($P < 0.001$; Fig. 2 F and Fig. S3 A). Conversely, low RNAPolII binding combined with low gene expression predicted the absence of mutations for 95% of genes (Fig. 2 F). To test the accuracy of our prediction model, we analyzed the mutation frequency of a new collection of genes (not included in our capture library) with high-density RNAPolII and Spt5 binding (Fig. S3 C and Table S4). We found that 11/12 of the analyzed genes were significantly mutated (Table S4 and Fig. 2 G). Indeed, two genes (*Hist1h1c* and *Clec2d*) were mutated at the range of the top 20% mutated genes at frequencies similar to those found in *Pax5* or *Rhoh* (Table S2 and Table S4). Thus, we have built a powerful predictive tool for AID activity.

### BER and MMR back up each other to faithfully repair AID-induced lesions

BER and MMR act downstream of AID-induced U:G mismatches so that UNG is critical for the generation of transversions at C:G pairs while MSH2 facilitates the introduction of mutations at A:T pairs (Frey et al., 1998; Phung et al., 1998; Rada et al., 1998, 2002, 2004; Methot and Di Noia, 2017). UNG and MSH2 can also promote conventional, faithful repair of AID-induced U:G mismatches (Liu et al., 2008; Pérez-Durán et al., 2012). To explore the contribution of BER and MMR to AID mutagenic activity, we analyzed GC B cells from single-deficient $Ung^{+/-}Msh2^{-/-}$ and $Ung^{-/-}Msh2^{+/-}$ mice and from control $Ung^{+/-}Msh2^{+/-}$ mice and compared the mutation frequency of the 291 AID target regions identified in this study (Table S2). We found similar mean mutation frequencies in B cells deficient for UNG alone, MSH2 alone, or proficient for both, whereas AID targets harbored significantly more mutations in the combined absence of UNG and MSH2 (Fig. 3, A and B). Indeed, only a small proportion (∼6%) of the genes mutated in $Ung^{-/-}Msh2^{-/-}$ cells was also mutated in single-knockout and double-heterozygous cells (Fig. 3 C and Table S2). Moreover, we found that classical AID off targets, such as *Bcl6* or *Pim1*, although mutated in all genotypes analyzed, harbored a significantly bigger load of mutations in $Ung^{-/-}Msh2^{-/-}$ cells than in $Ung^{+/-}Msh2^{-/-}$, $Ung^{-/-}Msh2^{+/-}$, or $Ung^{+/-}Msh2^{+/-}$ cells (Fig. 3 D). Together, these data indicate that BER and MMR back up each other to faithfully repair most of the AID-induced lesions in GC B cells.

### AID targets are recurrently mutated in human lymphomas

We next assessed the contribution of AID off-target mutations to B cell–derived malignancies by making use of available sequencing data on human lymphomas. We found that AID targets are significantly enriched in genes mutated in human B cell lymphomas (see Annotation of AID targets in the Materials and methods for details; Fig. 4 A). Indeed, 21/275 (7.6%) of our set of AID target genes are mutated in diffuse large B cell lymphomas (DLBCLs; Fig. 4 B), a highly prevalent, aggressive form of lymphoma (Shaffer et al., 2012). Lymphoma genes mutated by AID included *Bcl6*, *RhoH*, *Pim1*, *Ebf1*, *Eif4a2*, and *Pax5*, which is in agreement with previous studies (Shen et al., 1998; Pasqualucci et al., 2001; Liu et al., 2008). In addition, we identified nine novel genes mutated in human DLBCLs that accumulate AID-induced mutations (Fig. 4 B), including *Mef2b*, *Lyn*, *Tnfaip3*, *Gna13*, and *Irf8*. Remarkably, we found many instances where the exact same mutations described in human lymphoma genes

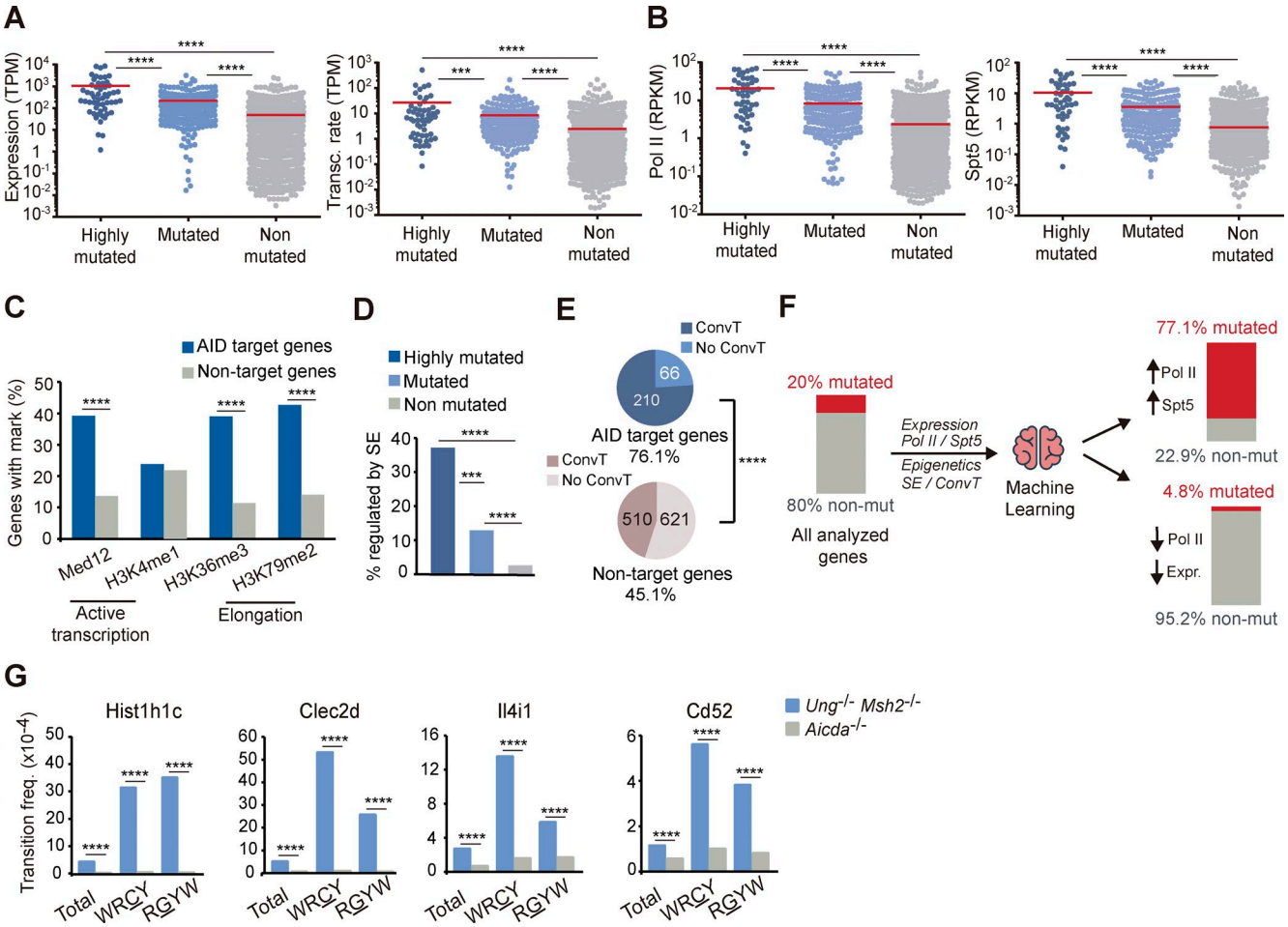



Figure 2. **Molecular features of AID targets predict mutability. (A)** Expression level of highly mutated (top 20% mutated genes, C→T transition frequency >3 × $10^{-4}$), mutated (rest of mutated), and nonmutated genes in Peyer's patch GC B cells as measured by RNA-Seq and transcription rate of AID targets in GC B cells from lymph nodes as measured by GRO-Seq. TPM, transcripts per million. **(B)** Recruitment of RNAPolII and Spt5 to AID targets and nontargets measured in in vitro activated splenic B cells by ChIP-Seq. RPKM, reads per kilobase per million reads mapped. **(C)** Transcription and transcription elongation marks in AID targets and nontargets by ChIP-Seq analysis of in vitro activated splenic B cells (Med12, H3K4me1, H3K36me3, and H3K79me2). **(D)** Proportion of highly mutated, mutated, and nonmutated genes regulated by superenhancers (SE) in GC B cells (see Materials and methods). **(E)** GRO-Seq analysis of convergent transcription (ConvT) in AID targets and nontargets from GC splenic B cells obtained from SRBC-immunized mice. **(F)** Representation of the machine-learning approach used for AID target prediction. **(G)** Validation of representative genes predicted to be mutated by the model by PCR-Seq. Statistical tests: two-tailed Student's *t* test (A, B, and G) and one-tailed Fisher test (C–E). ***, $P < 10^{-3}$; ****, $P < 10^{-4}$.

were also found in the AID targets identified in this study in nontransformed mouse B cells (Fig. 4 C and Table S5). Together, these results suggest that off-target AID mutagenic activity can contribute to GC-associated lymphomagenesis.

Until now, the study of AID specificity has been hindered by the technical challenge of detecting AID-induced mutations; indeed, only a limited number of genes has been directly interrogated for AID-mediated mutagenesis (Pasqualucci et al., 2001; Liu et al., 2008; Methot and Di Noia, 2017). However, genome-wide AID specificity has been inferred from high-throughput analysis of AID binding, which does not warrant AID activity, AID-induced DSBs, or chromosomal TCs, which involve complex processing of the initial lesion induced by AID (Chiarle et al., 2011; Klein et al., 2011; Staszewski et al., 2011; Yamane et al., 2011; Meng et al., 2014; Qian et al., 2014). The strategy developed in this study has provided an unprecedented scope to the analysis of AID targeting: we describe here the broadest collection of AID mutational targets (275 genes) to date, 10-fold larger than the previously known targets. The strength of this analysis is well supported by the confirmation of the vast majority of previously identified AID targets and the validation of targets by conventional Sanger sequencing.

Here, we have integrated our mutation data with a collection of molecular features of GC B cells to feed a machine-learning algorithm. According to the machine-learning tree generated here, the combined binding of Spt5 and RNAPolII at high density is the best predictor for AID mutability, although

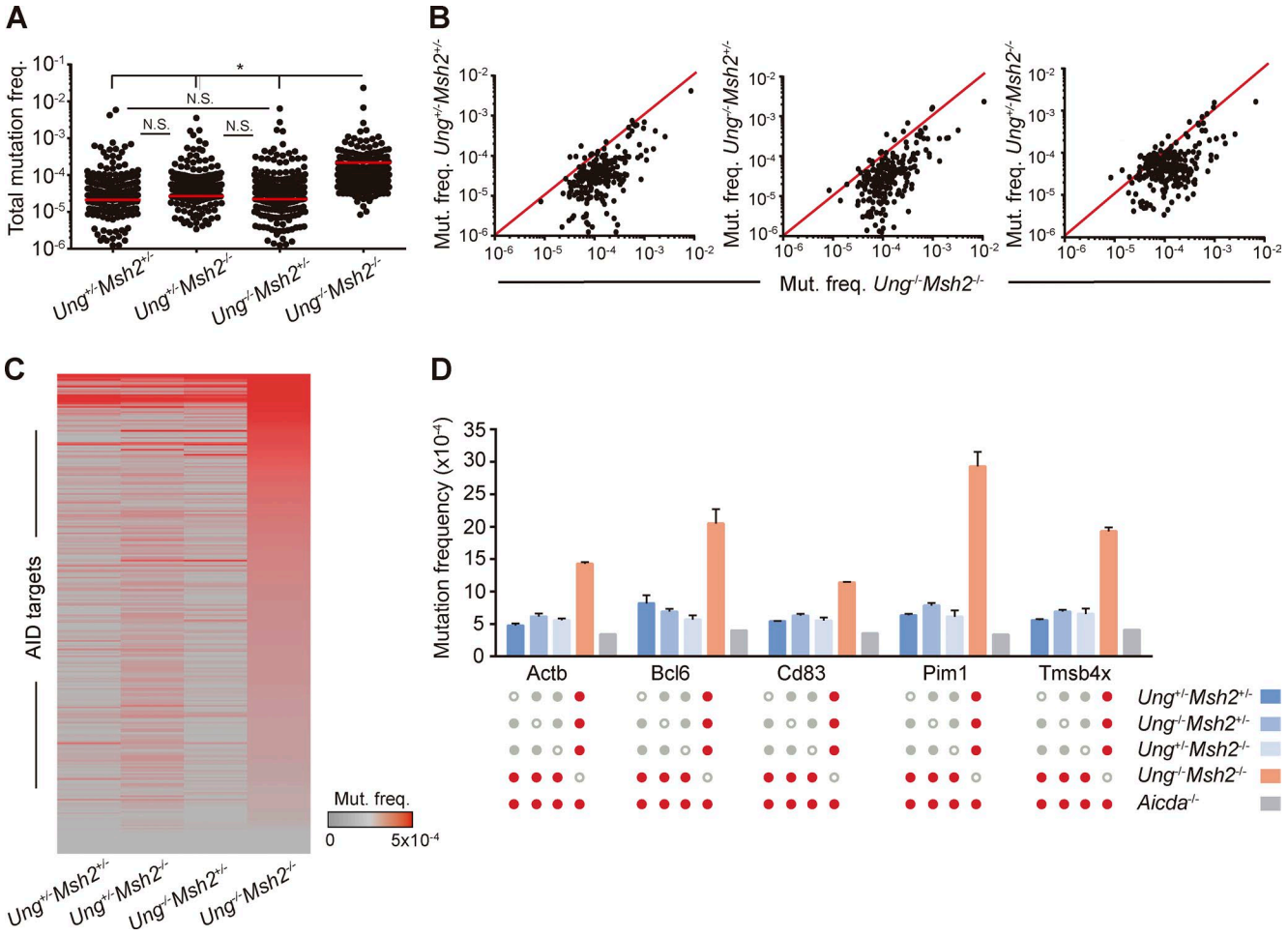



Figure 3. **BER and MMR back up each other to error-free repair AID-induced lesions. (A and B)** Total mutation frequency of AID targets in *Ung*$^{+/-}$*Msh2*$^{+/-}$, *Ung*$^{-/-}$*Msh2*$^{+/-}$, and *Ung*$^{+/-}$*Msh2*$^{-/-}$ GC B cell mice compared with that of *Ung*$^{-/-}$*Msh2*$^{-/-}$ mice (mean of two independent experiments; see Materials and methods; Table S2). **(C)** Heat map representation of AID targets in *Ung*$^{+/-}$*Msh2*$^{+/-}$, *Ung*$^{-/-}$*Msh2*$^{+/-}$, *Ung*$^{+/-}$*Msh2*$^{-/-}$, and *Ung*$^{-/-}$*Msh2*$^{-/-}$ GC B cells. **(D)** Mutation frequency of representative genes in *Ung*$^{+/-}$*Msh2*$^{+/-}$, *Ung*$^{-/-}$*Msh2*$^{+/-}$, *Ung*$^{+/-}$*Msh2*$^{-/-}$, *Ung*$^{-/-}$*Msh2*$^{-/-}$, and *Aicda*$^{-/-}$ GC B cells. Red dots indicate statistically different mutation frequencies between the indicated genotypes. Mutation frequency found in *Aicda*$^{-/-}$ mice was substracted before plotting A–C. (A and D) Two-tailed Student's *t* test; *, $P \leq 0.05$. Error bars depict SEM. N.S., not significant.

additional combinations of transcriptional traits bear some predictive power as well. Furthermore, we have performed independent experimental validation showing that randomly picked Spt5$^{high}$RNAPolII$^{high}$ genes indeed are very frequently mutated by AID. This is, to our knowledge, the first instance of a tool that successfully predicts the potential of a gene to be targeted by AID. Regarding the fate of AID-induced lesions, BER and MMR have long been known to broaden the diversity of SHM with an apparent perverted recruitment of error-prone polymerases and to do so in a cooperative manner (Rada et al., 2004; Di Noia and Neuberger, 2007; Methot and Di Noia, 2017). The mechanisms responsible for the error-free versus error-prone activity of UNG and MSH2 are far from understood, and both gene-specific and local sequence contexts may play a role in defining the fate of the U:G resolution (Liu et al., 2008; Pérez-Durán et al., 2012; Wei et al., 2015). Strikingly, here we show that the fate of the majority of off-target lesions induced by AID is to undergo faithful repair by BER and MMR and that, again, both pathways can back up each other in this task with only a minor fraction of the mutations escaping them. Whether this reflects gene-specific qualities or is the consequence of excessive mutation load will deserve further investigation. We would speculate that a minor fraction of unrepaired mutations in prolymphomagenic genes could provide cell growth advantage and account for the predominance of AID-mediated mutations in lymphomas. Regardless of oncogenic relevance, it is remarkable that even though our mutation analysis was performed in nontransformed cells, we could detect individual AID-induced mutations that are recurrently mutated in lymphoma. Thus, our results yield a novel perspective on the contribution of AID activity to B cell transformation through the introduction of mutations.

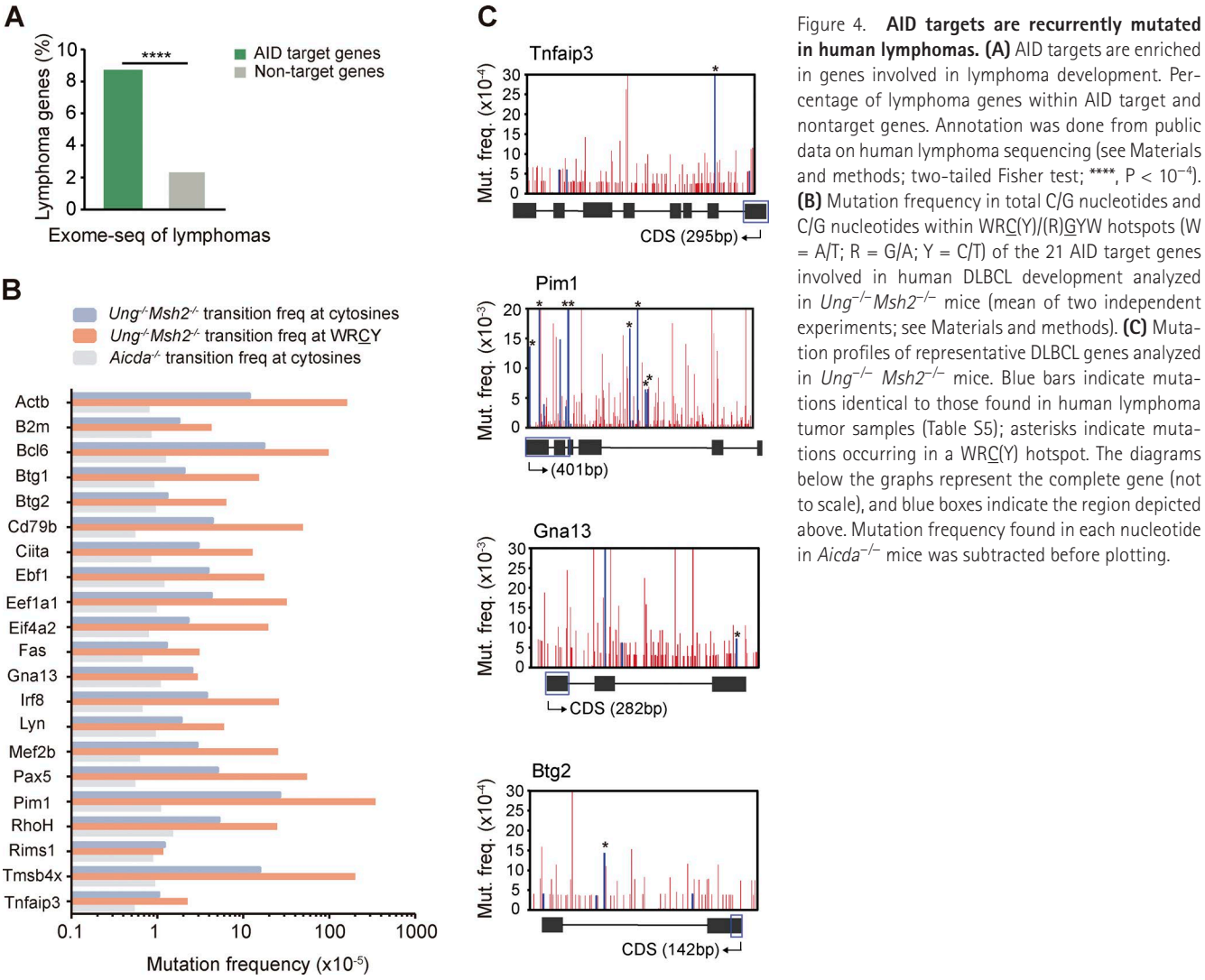



Figure 4. **AID targets are recurrently mutated in human lymphomas. (A)** AID targets are enriched in genes involved in lymphoma development. Percentage of lymphoma genes within AID target and nontarget genes. Annotation was done from public data on human lymphoma sequencing (see Materials and methods; two-tailed Fisher test; ****, $P < 10^{-4}$). **(B)** Mutation frequency in total C/G nucleotides and C/G nucleotides within WRC(Y)/(R)GYW hotspots (W = A/T; R = G/A; Y = C/T) of the 21 AID target genes involved in human DLBCL development analyzed in *Ung*$^{-/-}$*Msh2*$^{-/-}$ mice (mean of two independent experiments; see Materials and methods). **(C)** Mutation profiles of representative DLBCL genes analyzed in *Ung*$^{-/-}$ *Msh2*$^{-/-}$ mice. Blue bars indicate mutations identical to those found in human lymphoma tumor samples (Table S5); asterisks indicate mutations occurring in a WRC(Y) hotspot. The diagrams below the graphs represent the complete gene (not to scale), and blue boxes indicate the region depicted above. Mutation frequency found in each nucleotide in *Aicda*$^{-/-}$ mice was subtracted before plotting.

We expect our mutational study will be valuable for other research questions, including validation of novel molecular mechanisms involved in AID targeting, prediction of novel targets, or assessment of cancer-associated mutations. Furthermore, similar approaches would be of immediate interest to broaden our knowledge on the role of AID or other mutagenic activities not only in B cell lymphomas, but also in malignancies from any origin.

## MATERIALS AND METHODS

### Mice

*Ung* and *Msh2* mutant mice used in this study were generated by crossing *Ung*$^{-/-}$ mice (Nilsen et al., 2000) and *Msh2*$^{-/-}$ mice (Reitmair et al., 1995). *Aicda*$^{-/-}$ mice have been previously described (Muramatsu et al., 2000). Mice were housed in specific pathogen-free conditions. Male and female mice between 20 and 28 wk were used for the experiments. The number of animals per group to detect biologically significant effect sizes was calculated using an appropriate statistical sample size formula. All experiments were done in concordance with EU Directive 2010/63EU and Recommendation 2007/526/EC regarding the protection of animals used for experimental and other scientific purposes, enforced in Spanish law under RD 53/2013.

### Design of DNA capture library

A set of 1,379 mouse genes was selected as a representation of the genome (Table S1). 85% of all genes were randomly picked, ensuring even representation of chromosomal location by bioinformatic analysis and unbiased biological function. ~15% of the library corresponded to previously known AID targets (Müschen et al., 2000; Pasqualucci et al., 2001; Gordon et al., 2003; Liu et al., 2008; Robbiani et al., 2009; Pavri et al., 2010), IgH probes, and other controls. Probes were designed in eArray (Agilent) to capture the first 500 bp downstream of each transcriptional start site (TSS) of each of the 1,379 genes. Because various genes contained more than one predicted TSS, the library includes a total of 1,588 dif-

ferent genomic regions. Library design included 50 extra nucleotides at both ends of each region to optimize the capture yield. A custom target enrichment capture library was then synthesized by the manufacturer (SureSelectXT; Agilent).

### DNA capture and sequencing

GC ($Cd19^{+}Fas^{+}GL7^{+}$) B cells were isolated from Peyer's patches of $Ung^{+/-}Msh2^{+/-}$ ($n_1 = 10$; $n_2 = 11$), $Ung^{-/-}Msh2^{+/-}$ ($n_1 = 46$; $n_2 = 8$), $Ung^{+/-}Msh2^{-/-}$ ($n_1 = 46$; $n_2 = 2$), and $Ung^{-/-}Msh2^{-/-}$ ($n_1 = 37$; $n_2 = 8$) littermates and $Aicda^{-/-}$ ($n = 31$ mice) mice by sorting in a FACSAria cell sorter (BD Biosciences) after staining with anti–mouse antibodies to *Cd19*, *Fas*, and *GL7* (BD Biosciences). Genomic DNA was isolated by standard procedures and quantified in a fluorometer (Qubit; Invitrogen). DNA capture, library preparation, and DNA sequencing were performed by the Genomics Unit at Centro Nacional de Investigaciones Cardiovasculares (CNIC). In brief, DNA was fragmented in a sonicator (Covaris) to ∼200 nucleotide-long (mean size) fragments and purified using AMPure XP beads (Agencourt). Quality was assessed with the 2100 Bioanalyzer (Agilent). Then, fragment ends were repaired, adapters were ligated, and the resulting library was amplified and hybridized with our custom SureSelectXT library of RNA probes. DNA–RNA hybrids were then captured by magnetic bead selection. After indexing, libraries were single-end sequenced in a HiSeq 2500 platform (Illumina).

### Target enrichment assessment by quantitative RT-PCR

*Noxa1*, *Ostn*, and *Pcna* amplifications were quantified with green assay (SYBR; Applied Biosystems) in a real-time PCR system (AB7900 Standard; AbiPrism). *Gapdh* amplifications were used as normalization controls. The following primers were used: *Gapdh* (forward), 5′-TGAAGCAGGCATCTGAGGG-3′; *Gapdh* (reverse), 5′-CGAAGGTGGAAAGTGGGAG-3′; *Ostn* (forward), 5′-CATAGTGTTGCTGTGGTT-3′; *Ostn* (reverse), 5′-CATTATATTGGTCTGCTGTT-3′; *Noxa1* (forward), 5′-CGCGGGACAGCAATGAGAAG-3′; *Noxa1* (reverse), 5′-CCATCTACTCAGTTTCAAGGA-3′; *Pcna* (forward), 5′-CTCCAGCACCTTCTTCAG-3′; and *Pcna* (reverse), 5′-TCTCATCTAGTCGCCACA-3′.

SDS software (Applied Biosystems) was used for analysis of the data.

### Sanger sequencing

Regions to be sequenced were amplified from 160–200-ng genomic DNA in four independent reactions to minimize possible PCR biases. The following primers were used: *Hist1h1b* (forward), 5′-ATGCCTTAGACTTCACCGCC-3′; *Hist1h1b* (reverse), 5′-TTGTAACCTTGAGTCGCCGC-3′; *miR142* (forward), 5′-CGGTCCCTGGGAAGTTACAC-3′; *miR142* (reverse), 5′-AACGAGAGGCAAACAGTCTTCA-3′; *Cd19* (forward), 5′-GCCCCTCTTCCCTCCTCATA-3′; *Cd19* (reverse), 5′-CCTGCACCCACTCATCTGAA-3′; *Cdk4* (forward), 5′-TCTGGCAGCTGGTCACATGG-3′; and *Cdk4* (reverse), 5′-GATCACCAGCTAGTCGTCCC-3′. Amplification reactions were carried in a final volume of 25 µl using 2.5 U Pfu Ultra HF DNA polymerase (Agilent) and the following PCR setup: 95°C for 2 min, 25 (*Cd19* and *Cdk4*) or 26 cycles (*miR142* and *Hist1h1b*) of denaturation at 94°C for 30 s, annealing at 57°C (*miR142* and *Hist1h1b*) or 58°C (*Cd19* and *Cdk4*) for 30 s, extension at 72°C for 1 min, and a final stage of 72°C for 10 min. PCR products were purified from a 1% agarose gel (Illustra Gel Band Purification kit; GE Healthcare) and cloned into pGEMT vector (Promega). Competent DH5α *Escherichia coli* bacteria were transformed with the constructs, and individual colonies (192–288 per gene) were grown in 96-well plates. Plasmidic DNA was then isolated (Plasmid MiniPrep kit; Millipore) and sequenced by Sanger sequencing using SP6 universal primer. Sequence analysis was performed using SeqMan software (Lasergene).

### PCR-Seq to validate the machine-learning approach

40–50 ng of genomic DNA was amplified using the following primers: *Apobec3* (forward), 5′-GTCTTCCATAGCCTGCTCACA-3′; *Apobec3* (reverse), 5′-TAGCTGACTGGTGTGGTTCC-3′; *Aurkaip1* (forward), 5′-ACTTGTCACTTCCGCAGTCC-3′; *Aurkaip1* (reverse), 5′-CCATCCCCAAGTCAGGTGTG-3′; *Ccdc17* (forward), 5′-TCTTTTCTGTCCAGTCCGCC-3′; *Ccdc17* (reverse), 5′-ACAAATGGGCAGAGTCAGGG-3′; *Cd52* (forward), 5′-TACTGCCGCACACATGACTC-3′; *Cd52* (reverse), 5′-TGAGGTGGGAAGCCAAACAT-3′; *Cd68* (forward), 5′-AGGGGCTGGTAGGTTGATTG-3′; *Cd68* (reverse), 5′-GGAGTCAGGACTGGATTTGAC-3′; *Cd69* (forward), 5′-TCTAAAGGTTTTGAGACCCCC-3′; *Cd69* (reverse), 5′-TGAAGCCTCATCAACGCACT-3′; *Clec2d* (forward), 5′-GGCTCCTGACCTTGAAATGC-3′; *Clec2d* (reverse), 5′-AGGCAACTTCTGCCACTATGC-3′; *Coro1a* (forward), 5′-AGGGCTCTGGGGTTCTACTT-3′; *Coro1a* (reverse), 5′-GGAAATGACCACGGGGGTTT-3′; *Hist1h1c* (forward), 5′-CTCTATCGGCGTACTGCCAC-3′; *Hist1h1c* (reverse), 5′-ATCGAGTCCCTTGCAACCTT-3′; *Il4i* (forward), 5′-ATTCCCGAGGGAGGTGAGTG-3′; *Il4i* (reverse), 5′-GGTAGCTTCTCTCCGTCACAC-3′; *Maz* (forward), 5′-GTCAACAAAGAACCCCTCCCT-3′; *Maz* (reverse), 5′-CACCTGTCCCCTGAGTTGTG-3′; *Trex1* (forward), 5′-GCCTAACAGGTTTGATTGTCC T-3′; and *Trex1* (reverse), 5′-TAGGCTGAGCACTCCCAGTC-3′. Amplification reactions were carried in a final volume of 25 µl using 2.5 U Pfu Ultra HF DNA polymerase (Agilent; 95°C for 2 min, 26 cycles of 94°C for 30 s, 55°C for 30 s, 72°C for 1 min, and a final stage of 72°C for 10 min). PCR products were purified and fragmented using a sonicator (Covaris), and libraries were prepared by the CNIC Genomics Unit according to the manufacturer's instructions (NEBNext Ultra DNA Library Prep; New England Biolabs). Sequencing was performed in a HiSeq 2500 platform (Illumina). Analysis was performed as previously described (Pérez-Durán et al., 2012).

### Gene expression profiling by RNA-Seq

GC ($CD19^+Fas^+GL7^+$) and resting ($CD19^+Fas^-GL7^-$) B cells were sorted from Peyer's patches of littermate 12-wk-old WT C57BL/6 mice. Three biological replicates were analyzed, each composed of a pool of five female mice. RNA was purified from pellets of 2–2.5 × $10^4$ cells, and DNaseI treatment was applied to avoid DNA contamination (RNAeasy MiniKit; Qiagen). RNA quality was assessed with the 2100 Bioanalyzer, showing high RNA purity and integrity. Sequencing libraries were prepared by the CNIC Genomic Unit according to the manufacturer's protocol (NEB NEXT Ultra RNaseq Library Prep kit; New England Biolabs) from 100 ng RNA per replicate and sequenced in a HiSeq 2500 platform.

### Computational analysis

**Pipeline to identify and annotate AID-induced mutations.** Raw reads were demultiplexed by Casava v1.8 to generate a fastq file that was aligned to the mouse genome (NCBIm37 v61 Feb 2011) with Novoalign 2.08.01 (command line options: -o SAM -F ILM1.8 -H -r None -q 2). Sam files were processed with Samtools 0.1.19 to generate a sorted bam file that was piped to a custom Perl script for the analysis of AID mutations. In brief, the software analyzes the regions of interest in the bam file, annotates hotspots, localizes and suppresses annotated single nucleotide polymorphism positions (Sanger Mouse Genomes Project SNP and Indel Release v2), and reports relevant information about AID activity. AID targets were identified as those genes accumulating significantly more C→T transition mutations in $Ung^{-/-}Msh2^{-/-}$ than in $Aicda^{-/-}$ mice (false discovery rate [FDR] ≤0.05, one-tail Fisher test and Benjamini-Hochberg correction).

Mutation frequencies were calculated as follows:

$$\text{Total mutation freq} = \frac{\text{Total number of mutations}}{\text{Total sequenced length}},$$

$$\text{Mutation freq}_{C/G} = \frac{(\text{Mutated cytosines} + \text{Mutated guanines})}{(\text{Seq length cytosines} + \text{Seq length guanines})},$$

and

$$\text{Mutation freq}_{WRC(Y)/(R)GYW} = \frac{(\text{Mutated cytosines}_{WRC(Y)} + \text{Mutated guanines}_{(R)GYW})}{(\text{Seq length cytosines}_{WRC(Y)} + \text{Seq length guanines}_{(R)GYW})}.$$

(Only cytosines in WR<u>C</u>(Y) and guanines in (R)<u>G</u>YW were considered to calculate mutation frequency at hotspots.)

**Integration of AID targets with public data on TC and DSB occurrence.** The bar graph included in Fig. S1 F represents overlaps in the 1,375 genes analyzed in this study (divided into mutated and nonmutated genes) and genes where TCs or DSBs occur in B cells: Meng et al. (2014) refer to TC sites identified by HTGTS in αCD40+IL4-activated B cells as published in Table S2 from their study; Klein et al. (2011) refer to TC sites identified by TC-Seq in $IgH^{I\text{-}Sce}$ LPS+IL4-activated B cells as published in Table S4 from their study; Chiarle et al. (2011) refer to TC sites identified by HTGTS in $c\text{-}myc^{25xI\text{-}SceI}$ αCD40+IL4-activated B cells as published in Table S3 (significant hits at P ≤ 0.05) from their study; Qian et al. (2014) refer to DSBs identified by replication protein A (RPA) differential recruitment (RPA–chromatin immunoprecipitation [ChIP]) in Ig*k*AID $53BP1^{-/-}$ in vitro activated B cells as published in Table S1 A from their study; and Staszewski et al. (2011) refer to DSBs identified by Nbs1 binding (ChIP-on-ChIP) in LPS+αIgD-dextran+BLySS-activated B cells as published in Table S1 (P ≤ 0.05) from their study.

**Sequence context of mutated cytosines.** The sequence context of mutated cytosines (C→T transition frequency ≥4 × $10^{-3}$) was analyzed in a window of 10 nucleotides. Logo representation was done using WebLogo3, and the percentage of each nucleotide in each position surrounding the mutated cytosine was calculated by a custom Perl script. Enrichment for adenosine, guanine, cytosine, or thymine was tested against the sequence context of all cytosines present in the 1,588 regions analyzed in this study (one-tailed Student's *t* test + Bonferroni correction).

**Gene expression profiling by RNA-Seq.** After demultiplexing by Casava v1.8, read quality was assessed by FASTQC, and sequencing adapters were removed from sequence reads by cutadapt v1.9. The resulting reads were aligned to and quantified on the mouse transcriptome (NCBIm38 v75, Feb 2014) using RSEM v1.2.25 with the following parameters: -p 3–time–output-genome-bam–sampling-for-bam–bowtie-e 60–bowtie-m 30–bowtie-chunkmbs 512–fragment-length-mean 180–fragment-length-sd 50.

**Transcription rate analysis (GRO-Seq).** Reads were mapped to the mouse genome (mm9/NCBI37) using bowtie2, and uniquely mapped, nonredundant reads were kept. Reads mapping in ±1 kb from TSSs were quantified and summarized at the gene level using HTSeq.

**PolII and Spt5 recruitment.** Quantification of PolII and Spt5 recruitment was extracted from Table S3 A in Pavri et al. (2010).

**Superenhancer analysis.** Data were extracted from the catalog of superenhancers that overlap with gene bodies identified in GC B cells as published in Table S3 in Meng et al. (2014) (GEO accession no. GSE62296).

**Epigenetic mark analysis.** Sequencing data (fastq files) for each epigenetic mark were aligned to the mouse genome (NCBIm37 v61, Feb 2011) using bowtie 1.1.1 (command line options:–best -m1 -n2 -p2). Alignment files were processed by Samtools 0.1.19 to generate a sorted bam file. Peak calling was done using MACS2 (v2.1.0.20140616) according to the optimal parameters for a histone modification status

profiling as reported by the creators of the tool (Feng et al., 2011). Mapping of annotated peaks to genes was done using GREAT (version 3.0.0).

**Convergent transcription analysis (GRO-Seq).** Convergent transcription data analysis was performed as described in Meng et al. (2014). In brief, reads were mapped to the mouse genome (mm9/NCBI37) using bowtie2, and uniquely mapped, nonredundant reads were kept. HOMER (v4.6) was used with default parameters to identify transcribed regions from both strands and bedtools (v2.24) to find and annotate ConvT regions (regions where >100 bp of sense and antisense transcription overlap occurs).

**Machine learning to predict AID targets.** The conditional inference tree for classification was built using the *ctree* function from the party R package with default parameters. Genes with a background mutation frequency $>5 \times 10^{-4}$ were excluded to avoid artifacts. The following variables were fed into the model for each of the 1,339 genes analyzed: expression, transcription rate, PolII recruitment, and Spt5 recruitment (quantitative, continuous); Med12 recruitment, H3K4me1 recruitment, H3K36me3 recruitment, H3K79me2 recruitment, regulation by superenhancers, and occurrence of convergent transcription (qualitative, discrete). All variables were assigned equal weights to fit the model.

**Annotation of AID targets.** Annotation of AID targets was performed based on public data on sequencing of human DLBCLs, Burkitt lymphomas, and follicular lymphoma tumors (Lohr et al., 2012; Love et al., 2012; Morin et al., 2013; Zhang et al., 2013; de Miranda et al., 2014; Okosun et al., 2014).

### Data availability

Sequencing data generated for this study are available through the GEO database: targeted DNA deep sequencing (accession no. GSE102944) and RNA-Seq (accession no. GSE98086).

The rest of the datasets analyzed in the current study are publicly available through the GEO and/or Sequence Read Archive: GRO-Seq (accession no. GSE62296), GC B cells (accession nos. SRR1611832, SRR1611833, and SRR1611834), naive B cells (accession nos. SRR1611829, SRR1611830, and SRR1611831), ChIP-Seq of PolII and Spt5 (accession no. GSE24178), and ChIP-Seq data of epigenetic marks Med12 (accession no. SRX347810), H3K4me1 (accession no. SRX347815), H3K36me3 (accession no. SRX185869), and H3K79me2 (accession no. SRX185843).

### Statistical analysis

Statistical analyses were performed with stats R package v3.1.1. Error bars in figures represent SEM. Student's *t* test was applied to continuous data, and a Fisher test was used to assess differences between categorical variables. P-values were corrected for multiple hypothesis testing by Benjamini-Hochberg or Bonferroni method where appropriate. Differences were considered statistically significant at $P \leq 0.05$ or $q \leq 0.05$.

### Online supplemental material

Fig. S1 shows the experimental workflow used to identify AID targets and technical controls. Fig. S2 shows mutation analysis of WRCY/RGYW hotspots. Fig. S3 shows details on the machine-learning classification tree used for the prediction of AID targets. Table S1 contains a list of the genes included in the capture library. Table S2 A contains a detailed mutation analysis of AID targets in $Ung^{+/-}Msh2^{+/-}$, $Ung^{-/-}Msh2^{+/-}$, $Ung^{+/-}Msh2^{-/-}$, and $Ung^{-/-}Msh2^{-/-}$. Table S2 B contains a list of the 18 AID targets mutated in repair-proficient GC B cells. Table S3 shows mutation analysis of genes validated by Sanger sequencing. Table S4 shows mutation analysis of the genes selected for machine-learning validation. Table S5 contains a list of the mutations found in $Ung^{-/-}Msh2^{-/-}$ GC B cells that have been identified in cohorts of human lymphoma patients.

## ACKNOWLEDGMENTS

We thank all members of the B Cell Biology Laboratory for useful discussions, V. Barreto for critical reading of the manuscript, F. Sánchez-Cabo for advice on statistics analysis, J.M. Ligos for help with flow cytometry, and A. Dopazo for advice on DNA capture and sequencing.

A. Pérez-García was a fellow of the research training program funded by the Ministerio de Educación, Cultura y Deporte (grant FPU-AP2009-1732); A.F. Álvarez-Prado and A.R. Ramiro are supported by Centro Nacional de Investigaciones Cardiovasculares Carlos III (CNIC). This work was funded with the following grants to A.R. Ramiro from Plan Estatal de Investigación Científica y Técnica y de Innovación 2013–2016, Programa Estatal de I+D+i Orientada a los Retos de la Sociedad Retos Investigación: Proyectos I+D+i 2016, Ministerio de Economía, Industria y Competitividad (MEIC; grants SAF2013-42767-R and SAF2016-75511-R). This work is cofunded by Fondo Europeo de Desarrollo Regional and the European Research Council Starting Grant program (grant BCLYM-207844). The CNIC is supported by the MEIC and the Pro CNIC Foundation and is a Severo Ochoa Centre of Excellence (MEIC award SEV-2015-0505).

The authors declare no competing financial interests.

Author contributions: A.F. Álvarez-Prado, P. Pérez-Durán, and A.R. Ramiro designed experiments; A.F. Álvarez-Prado, P. Pérez-Durán, A. Pérez-García, and V.G. de Yébenes performed experiments; A. Benguria performed DNA sequencing; A.F. Álvarez-Prado and C. Torroja developed scripts for data analysis; A.F. Álvarez-Prado and A.R. Ramiro analyzed data and prepared figures; and A.F. Álvarez-Prado and A.R. Ramiro wrote the manuscript.

**Submitted: 20 September 2017**
**Revised: 22 November 2017**
**Accepted: 21 December 2017**

## REFERENCES

Chiarle, R., Y. Zhang, R.L. Frock, S.M. Lewis, B. Molinie, Y.J. Ho, D.R. Myers, V.W. Choi, M. Compagno, D.J. Malkin, et al. 2011. Genome-wide translocation sequencing reveals mechanisms of chromosome breaks and rearrangements in B cells. *Cell.* 147:107–119. https://doi.org/10.1016/j.cell.2011.07.049

de Miranda, N.F., K. Georgiou, L. Chen, C. Wu, Z. Gao, A. Zaravinos, S. Lisboa, G. Enblad, M.R. Teixeira, Y. Zeng, et al. 2014. Exome sequencing reveals novel mutation targets in diffuse large B-cell lymphomas derived from Chinese patients. *Blood.* 124:2544–2553. https://doi.org/10.1182/blood-2013-12-546309

Di Noia, J.M., and M.S. Neuberger. 2007. Molecular mechanisms of antibody somatic hypermutation. *Annu. Rev. Biochem.* 76:1–22. https://doi.org/10.1146/annurev.biochem.76.061705.090740

Dong, J., R.A. Panchakshari, T. Zhang, Y. Zhang, J. Hu, S.A. Volpi, R.M. Meyers, Y.-J. Ho, Z. Du, D.F. Robbiani, et al. 2015. Orientation-specific joining of AID-initiated DNA breaks promotes antibody class switching. *Nature.* 525:134–139. https://doi.org/10.1038/nature14970

Feng, J., T. Liu, and Y. Zhang. 2011. Using MACS to identify peaks from ChIP-Seq data. *Curr. Protoc. Bioinformatics.* 2:14.

Frey, S., B. Bertocci, F. Delbos, L. Quint, J.C. Weill, and C.A. Reynaud. 1998. Mismatch repair deficiency interferes with the accumulation of mutations in chronically stimulated B cells and not with the hypermutation process. *Immunity.* 9:127–134. https://doi.org/10.1016/S1074-7613(00)80594-4

Gordon, M.S., C.M. Kanegai, J.R. Doerr, and R. Wall. 2003. Somatic hypermutation of the B cell receptor genes B29 (Igbeta, CD79b) and mb1 (Igalpha, CD79a). *Proc. Natl. Acad. Sci. USA.* 100:4126–4131. https://doi.org/10.1073/pnas.0735266100

Klein, I.A., W. Resch, M. Jankovic, T. Oliveira, A. Yamane, H. Nakahashi, M. Di Virgilio, A. Bothmer, A. Nussenzweig, D.F. Robbiani, et al. 2011. Translocation-capture sequencing reveals the extent and nature of chromosomal rearrangements in B lymphocytes. *Cell.* 147:95–106. https://doi.org/10.1016/j.cell.2011.07.048

Liu, M., J.L. Duke, D.J. Richter, C.G. Vinuesa, C.C. Goodnow, S.H. Kleinstein, and D.G. Schatz. 2008. Two levels of protection for the B cell genome during somatic hypermutation. *Nature.* 451:841–845. https://doi.org/10.1038/nature06547

Lohr, J.G., P. Stojanov, M.S. Lawrence, D. Auclair, B. Chapuy, C. Sougnez, P. Cruz-Gordillo, B. Knoechel, Y.W. Asmann, S.L. Slager, et al. 2012. Discovery and prioritization of somatic mutations in diffuse large B-cell lymphoma (DLBCL) by whole-exome sequencing. *Proc. Natl. Acad. Sci. USA.* 109:3879–3884. https://doi.org/10.1073/pnas.1121343109

Love, C., Z. Sun, D. Jima, G. Li, J. Zhang, R. Miles, K.L. Richards, C.H. Dunphy, W.W. Choi, G. Srivastava, et al. 2012. The genetic landscape of mutations in Burkitt lymphoma. *Nat. Genet.* 44:1321–1325. https://doi.org/10.1038/ng.2468

Meng, F.L., Z. Du, A. Federation, J. Hu, Q. Wang, K.R. Kieffer-Kwon, R.M. Meyers, C. Amor, C.R. Wasserman, D. Neuberg, et al. 2014. Convergent transcription at intragenic super-enhancers targets AID-initiated genomic instability. *Cell.* 159:1538–1548. https://doi.org/10.1016/j.cell.2014.11.014

Methot, S.P., and J.M. Di Noia. 2017. Molecular Mechanisms of Somatic Hypermutation and Class Switch Recombination. *Adv. Immunol.* 133:37–87. https://doi.org/10.1016/bs.ai.2016.11.002

Morin, R.D., K. Mungall, E. Pleasance, A.J. Mungall, R. Goya, R.D. Huff, D.W. Scott, J. Ding, A. Roth, R. Chiu, et al. 2013. Mutational and structural analysis of diffuse large B-cell lymphoma using whole-genome sequencing. *Blood.* 122:1256–1265. https://doi.org/10.1182/blood-2013-02-483727

Muramatsu, M., K. Kinoshita, S. Fagarasan, S. Yamada, Y. Shinkai, and T. Honjo. 2000. Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID), a potential RNA editing enzyme. *Cell.* 102:553–563. https://doi.org/10.1016/S0092-8674(00)00078-7

Müschen, M., D. Re, B. Jungnickel, V. Diehl, K. Rajewsky, and R. Küppers. 2000. Somatic mutation of the CD95 gene in human B cells as a side-effect of the germinal center reaction. *J. Exp. Med.* 192:1833–1840. https://doi.org/10.1084/jem.192.12.1833

Nambu, Y., M. Sugai, H. Gonda, C.-G. Lee, T. Katakai, Y. Agata, Y. Yokota, and A. Shimizu. 2003. Transcription-coupled events associating with immunoglobulin switch region chromatin. *Science.* 302:2137–2140. https://doi.org/10.1126/science.1092481

Nilsen, H., I. Rosewell, P. Robins, C.F. Skjelbred, S. Andersen, G. Slupphaug, G. Daly, H.E. Krokan, T. Lindahl, and D.E. Barnes. 2000. Uracil-DNA glycosylase (UNG)-deficient mice reveal a primary role of the enzyme during DNA replication. *Mol. Cell.* 5:1059–1065. https://doi.org/10.1016/S1097-2765(00)80271-3

Okosun, J., C. Bödör, J. Wang, S. Araf, C.Y. Yang, C. Pan, S. Boller, D. Cittaro, M. Bozek, S. Iqbal, et al. 2014. Integrated genomic analysis identifies recurrent mutations and evolution patterns driving the initiation and progression of follicular lymphoma. *Nat. Genet.* 46:176–181. https://doi.org/10.1038/ng.2856

Pasqualucci, L., P. Neumeister, T. Goossens, G. Nanjangud, R.S.K. Chaganti, R. Küppers, and R. Dalla-Favera. 2001. Hypermutation of multiple proto-oncogenes in B-cell diffuse large-cell lymphomas. *Nature.* 412:341–346. https://doi.org/10.1038/35085588

Pavri, R., A. Gazumyan, M. Jankovic, M. Di Virgilio, I. Klein, C. Ansarah-Sobrinho, W. Resch, A. Yamane, B. Reina San-Martin, V. Barreto, et al. 2010. Activation-induced cytidine deaminase targets DNA at sites of RNA polymerase II stalling by interaction with Spt5. *Cell.* 143:122–133. https://doi.org/10.1016/j.cell.2010.09.017

Pérez-Durán, P., L. Belver, V.G. de Yébenes, P. Delgado, D.G. Pisano, and A.R. Ramiro. 2012. UNG shapes the specificity of AID-induced somatic hypermutation. *J. Exp. Med.* 209:1379–1389. https://doi.org/10.1084/jem.20112253

Petersen-Mahrt, S.K., R.S. Harris, and M.S. Neuberger. 2002. AID mutates E. coli suggesting a DNA deamination mechanism for antibody diversification. *Nature.* 418:99–104. https://doi.org/10.1038/nature00862

Phung, Q.H., D.B. Winter, A. Cranston, R.E. Tarone, V.A. Bohr, R. Fishel, and P.J. Gearhart. 1998. Increased hypermutation at G and C nucleotides in immunoglobulin variable genes from mice deficient in the MSH2 mismatch repair protein. *J. Exp. Med.* 187:1745–1751. https://doi.org/10.1084/jem.187.11.1745

Qian, J., Q. Wang, M. Dose, N. Pruett, K.R. Kieffer-Kwon, W. Resch, G. Liang, Z. Tang, E. Mathé, C. Benner, et al. 2014. B cell super-enhancers and regulatory clusters recruit AID tumorigenic activity. *Cell.* 159:1524–1537. https://doi.org/10.1016/j.cell.2014.11.013

Rada, C., M.R. Ehrenstein, M.S. Neuberger, and C. Milstein. 1998. Hot spot focusing of somatic hypermutation in MSH2-deficient mice suggests two stages of mutational targeting. *Immunity.* 9:135–141. https://doi.org/10.1016/S1074-7613(00)80595-6

Rada, C., G.T. Williams, H. Nilsen, D.E. Barnes, T. Lindahl, and M.S. Neuberger. 2002. Immunoglobulin isotype switching is inhibited and somatic hypermutation perturbed in UNG-deficient mice. *Curr. Biol.* 12:1748–1755. https://doi.org/10.1016/S0960-9822(02)01215-0

Rada, C., J.M. Di Noia, and M.S. Neuberger. 2004. Mismatch recognition and uracil excision provide complementary paths to both Ig switching and the A/T-focused phase of somatic mutation. *Mol. Cell.* 16:163–171. https://doi.org/10.1016/j.molcel.2004.10.011

Ramiro, A.R., M. Jankovic, T. Eisenreich, S. Difilippantonio, S. Chen-Kiang, M. Muramatsu, T. Honjo, A. Nussenzweig, and M.C. Nussenzweig. 2004. AID is required for c-myc/IgH chromosome translocations in vivo. *Cell.* 118:431–438. https://doi.org/10.1016/j.cell.2004.08.006

Ramiro, A.R., M. Jankovic, E. Callen, S. Difilippantonio, H.-T. Chen, K.M. McBride, T.R. Eisenreich, J. Chen, R.A. Dickins, S.W. Lowe, et al. 2006. Role of genomic instability and p53 in AID-induced c-myc-Igh translocations. *Nature.* 440:105–109. https://doi.org/10.1038/nature04495

Reitmair, A.H., R. Schmits, A. Ewel, B. Bapat, M. Redston, A. Mitri, P. Waterhouse, H.-W. Mittrücker, A. Wakeham, B. Liu, et al. 1995. MSH2 deficient mice are viable and susceptible to lymphoid tumours. *Nat. Genet.* 11:64–70. https://doi.org/10.1038/ng0995-64

Revy, P., T. Muto, Y. Levy, F. Geissmann, A. Plebani, O. Sanal, N. Catalan, M. Forveille, R. Dufourcq-Labelouse, A. Gennery, et al. 2000. Activation-induced cytidine deaminase (AID) deficiency causes the autosomal recessive form of the Hyper-IgM syndrome (HIGM2). *Cell.* 102:565–575. https://doi.org/10.1016/S0092-8674(00)00079-9

Reynaud, C.A., F. Delbos, A. Faili, Q. Guéranger, S. Aoufouchi, and J.C. Weill. 2009. Competitive repair pathways in immunoglobulin gene hypermutation. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 364:613–619. https://doi.org/10.1098/rstb.2008.0206

Robbiani, D.F., A. Bothmer, E. Callen, B. Reina-San-Martin, Y. Dorsett, S. Difilippantonio, D.J. Bolland, H.T. Chen, A.E. Corcoran, A. Nussenzweig, and M.C. Nussenzweig. 2008. AID is required for the chromosomal breaks in c-myc that lead to c-myc/IgH translocations. *Cell.* 135:1028–1038. https://doi.org/10.1016/j.cell.2008.09.062

Robbiani, D.F., S. Bunting, N. Feldhahn, A. Bothmer, J. Camps, S. Deroubaix, K.M. McBride, I.A. Klein, G. Stone, T.R. Eisenreich, et al. 2009. AID produces DNA double-strand breaks in non-Ig genes and mature B cell lymphomas with reciprocal chromosome translocations. *Mol. Cell.* 36:631–641. https://doi.org/10.1016/j.molcel.2009.11.007

Rogozin, I.B., and N.A. Kolchanov. 1992. Somatic hypermutagenesis in immunoglobulin genes: II. Influence of neighbouring base sequences on mutagenesis. *Biochim. Biophys. Acta.* 1171:11–18. https://doi.org/10.1016/0167-4781(92)90134-L

Shaffer, A.L. III, R.M. Young, and L.M. Staudt. 2012. Pathogenesis of human B cell lymphomas. *Annu. Rev. Immunol.* 30:565–610. https://doi.org/10.1146/annurev-immunol-020711-075027

Shen, H.M., A. Peters, B. Baron, X. Zhu, and U. Storb. 1998. Mutation of BCL-6 gene in normal B cells by the process of somatic hypermutation of Ig genes. *Science.* 280:1750–1752. https://doi.org/10.1126/science.280.5370.1750

Staszewski, O., R.E. Baker, A.J. Ucher, R. Martier, J. Stavnezer, and J.E.J. Guikema. 2011. Activation-induced cytidine deaminase induces reproducible DNA breaks at many non-Ig Loci in activated B cells. *Mol. Cell.* 41:232–242. https://doi.org/10.1016/j.molcel.2011.01.007

Stavnezer, J., J.E.J. Guikema, and C.E. Schrader. 2008. Mechanism and regulation of class switch recombination. *Annu. Rev. Immunol.* 26:261–292. https://doi.org/10.1146/annurev.immunol.26.021607.090248

Storb, U. 2014. Why does somatic hypermutation by AID require transcription of its target genes? *Adv. Immunol.* 122:253–277. https://doi.org/10.1016/B978-0-12-800267-4.00007-9

Wang, Q., T. Oliveira, M. Jankovic, I.T. Silva, O. Hakim, K. Yao, A. Gazumyan, C.T. Mayer, R. Pavri, R. Casellas, et al. 2014. Epigenetic targeting of activation-induced cytidine deaminase. *Proc. Natl. Acad. Sci. USA.* 111:18667–18672. https://doi.org/10.1073/pnas.1420575111

Wei, L., R. Chahwan, S. Wang, X. Wang, P.T. Pham, M.F. Goodman, A. Bergman, M.D. Scharff, and T. MacCarthy. 2015. Overlapping hotspots in CDRs are critical sites for V region diversification. *Proc. Natl. Acad. Sci. USA.* 112:E728–E737. https://doi.org/10.1073/pnas.1500788112

Yamane, A., W. Resch, N. Kuo, S. Kuchen, Z. Li, H.W. Sun, D.F. Robbiani, K. McBride, M.C. Nussenzweig, and R. Casellas. 2011. Deep-sequencing identification of the genomic targets of the cytidine deaminase AID and its cofactor RPA in B lymphocytes. *Nat. Immunol.* 12:62–69. https://doi.org/10.1038/ni.1964

Yeap, L.-S., J.K. Hwang, Z. Du, R.M. Meyers, F.-L. Meng, A. Jakubauskaitė, M. Liu, V. Mani, D. Neuberg, T.B. Kepler, et al. 2015. Sequence-Intrinsic Mechanisms that Target AID Mutational Outcomes on Antibody Genes. *Cell.* 163:1124–1137. https://doi.org/10.1016/j.cell.2015.10.042

Zhang, J., V. Grubor, C.L. Love, A. Banerjee, K.L. Richards, P.A. Mieczkowski, C. Dunphy, W. Choi, W.Y. Au, G. Srivastava, et al. 2013. Genetic heterogeneity of diffuse large B-cell lymphoma. *Proc. Natl. Acad. Sci. USA.* 110:1398–1403. https://doi.org/10.1073/pnas.1205299110

SUPPLEMENTAL MATERIAL

Álvarez-Prado et al., https://doi.org/10.1084/jem.20171738

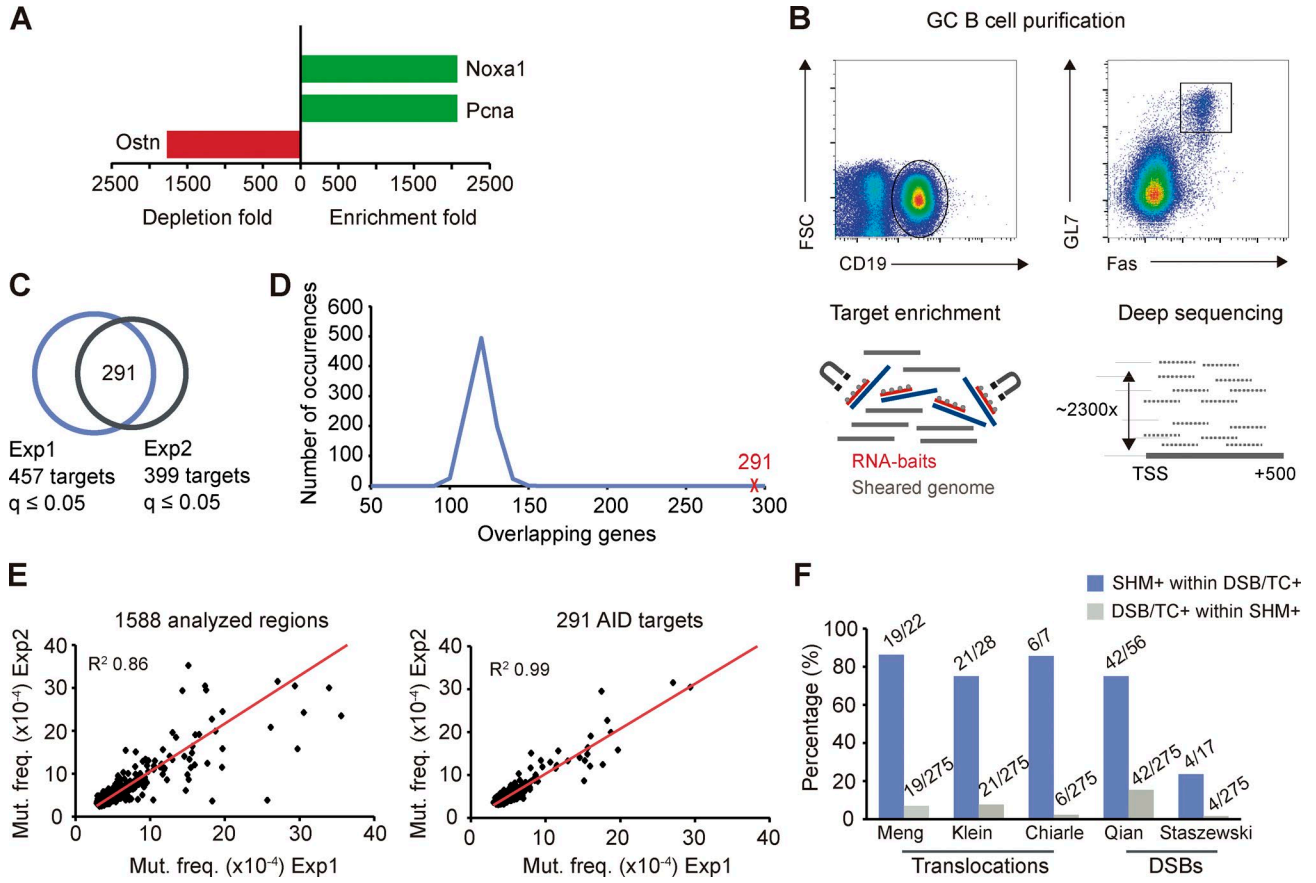



Figure S1. **Identification of AID targets by target enrichment coupled to next generation sequencing. (A)** Target enrichment protocol allows a 2,000-fold enrichment of selected genes. Genomic DNA corresponding to genes included (*Noxa1* and *PCNA*) and not included (*Ostn*) in the SureSelect capture library was quantified by quantitative RT-PCR before and after DNA capture enrichment. Graph represents fold depletion or fold enrichment calculated as $2^{(C_T \text{ Input}-C_T \text{ Enriched fraction})}$. Mean of two independent experiments is represented. **(B)** Schematic representation of the experimental approach used. GC (*CD19*$^+$*Fas*$^+$*GL7*$^+$) B cells from Peyer's patches were isolated by cell sorting, and genomic DNA was extracted, sheared, and captured with a custom library of RNA probes. Enriched DNA was subjected to next generation sequencing to achieve a mean depth of 2,300 reads per nucleotide. **(C)** Two independent experiments were performed (Table S2) with 457 mutated targets found in Exp1 and 399 in Exp2. An overlap of 291 AID targets was found between Exp1 and Exp2. **(D)** Experimental distribution of random overlaps simulated for 1,000 iterations. For each iteration, random groups of 457 and 399 genes were selected from the genes included in the SureSelect capture library, overlapped, and the number of coincident genes reported. The probability to find an overlap of 291 genes by chance is <1 out of each $10^{16}$ times tested. Two-tailed Fisher test; P = $\sim 10^{-16}$. **(E)** Mutation frequencies of the 1,588 TSS proximal regions analyzed and the 291 targets found in two independent experiments. **(F)** Percentage of genes undergoing DSB/TC+ according to the indicated studies within AID mutational targets described in this study (SHM+; 275 genes obtained in two independent experiments) and percentage of SHM+ genes within DSB/TC+ genes (see Materials and methods).

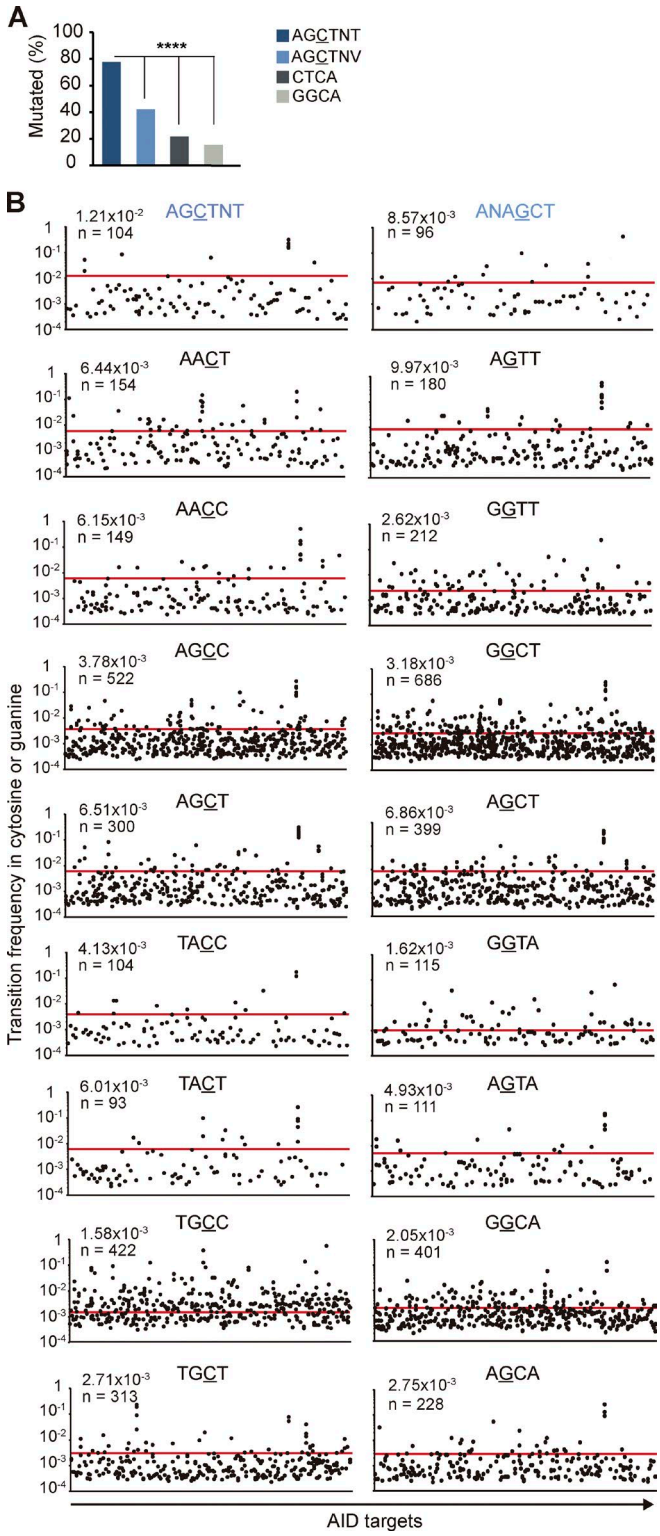



Figure S2. **Mutation analysis at WRCY/RGYW hotspots in *Ung*$^{-/-}$*Msh2*$^{-/-}$ GC B cells. (A)** Percentage of mutated cytosines within AGCTNT and AGCTNV hotspots and CTCA and GGCA non-hotspot motifs (Fisher test; ****, $P < 10^{-13}$). **(B)** Plots show mutated individual hotspots (WRCY, left; RGYW, right). Newly identified AGCTNT/ANAGCT hotspots are shown in the top row. Within each plot, each dot represents an individual WRCY/RGYW motif found mutated at least once. Each position in the x axis corresponds to a different gene, and the y axis shows mutation frequency of each individual hotspot within a gene. Mean mutation frequency is indicated and depicted with a red line. Number of mutated hotspots is indicated.

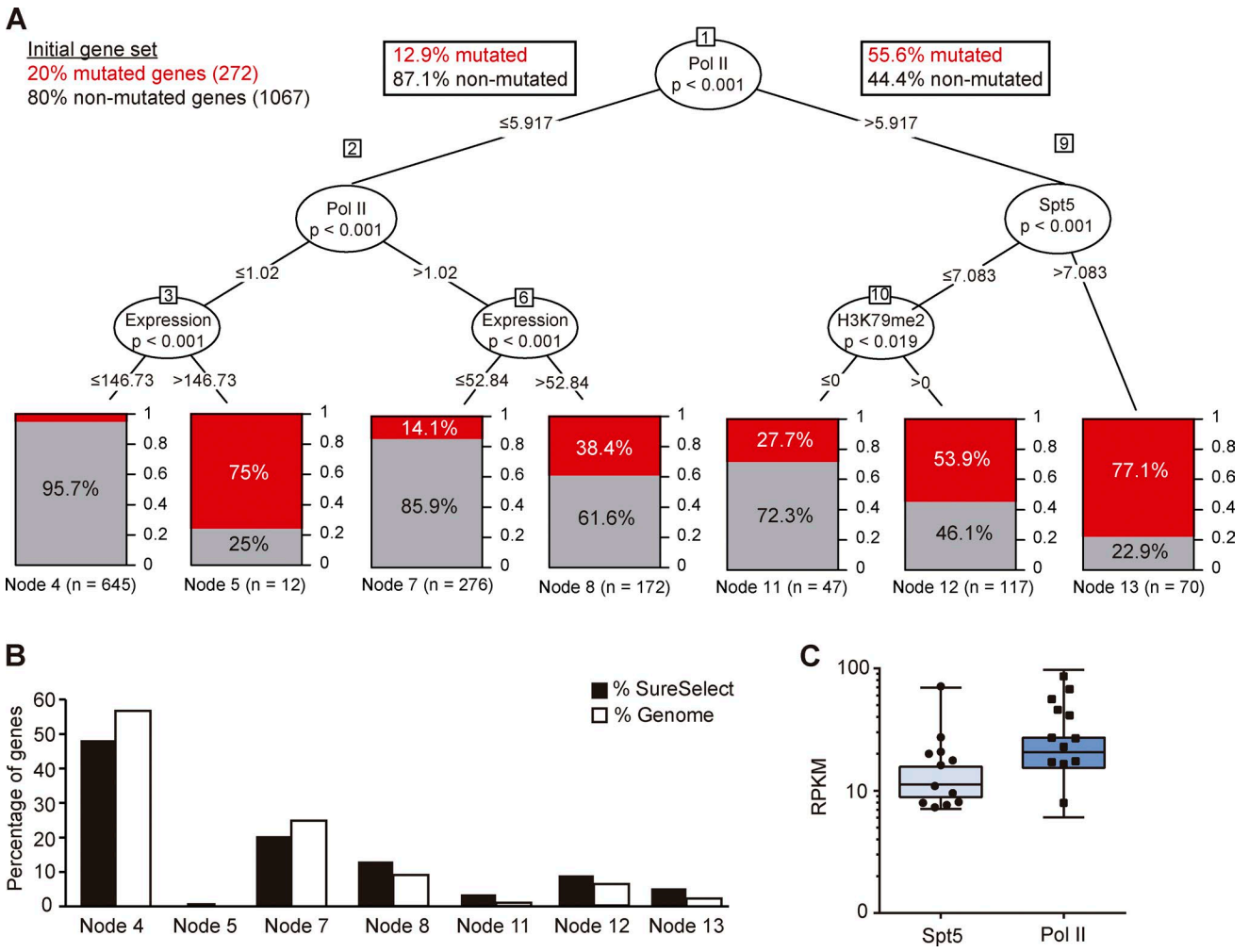



Figure S3. **Machine learning to predict AID targets genome wide. (A)** Recursive partitioning tree model classifies AID targets based on different molecular features: mRNA expression, PolII and Spt5 recruitment, and presence of H3K79me2 epigenetic mark (see Materials and methods). Each node splits the genes into two significantly different groups based on a particular feature. Numbers within the branches indicate the thresholds used to make the groups; p-values of each decision are included below the parameter measured in each node. **(B)** Bar graph depicting the proportion of SureSelect genes (1,339 genes; closed bars) or of total genes in the mouse genome (17,858 genes; open bars) that meet the thresholds established in each node. **(C)** Box plot depicting genome-wide data of PolII and Spt5 recruitment in in vitro activated B cells. Black dots and squares mark the 12 genes selected for the validation of the model prediction. RPKM, reads per kilobase per million reads mapped.

**Tables S1–S5 are provided in separate Excel files.**

**Table S1 contains a list of the genes included in the capture library.**

**Table S2 A contains a detailed mutation analysis of AID targets in *Ung*$^{+/-}$*Msh2*$^{+/-}$, *Ung*$^{-/-}$*Msh2*$^{+/-}$, *Ung*$^{+/-}$*Msh2*$^{-/-}$, and *Ung*$^{-/-}$*Msh2*$^{-/-}$. Table S2 B contains a list of the 18 AID targets mutated in repair-proficient GC B cells.**

**Table S3 shows mutation analysis of genes validated by Sanger sequencing.**

**Table S4 shows mutation analysis of the genes selected for machine-learning validation.**

**Table S5 contains a list of the mutations found in *Ung*$^{-/-}$*Msh2*$^{-/-}$ GC B cells that have been identified in cohorts of human lymphoma patients.**