



Universiteit  
Leiden  
The Netherlands

## **Tandem substitutions in somatic hypermutation**

Sepulveda Yanez, J.H.; Saravia, D.A.; Pilzecker, B.; Schouwenburg, P.A. van; Burg, M. van den; Veelken, H.; ... ; Koning, M.T.

### **Citation**

Sepulveda Yanez, J. H., Saravia, D. A., Pilzecker, B., Schouwenburg, P. A. van, Burg, M. van den, Veelken, H., ... Koning, M. T. (2022). Tandem substitutions in somatic hypermutation. *Frontiers In Immunology*, 12. doi:10.3389/fimmu.2021.807015

Version: Publisher's Version

License: [Creative Commons CC BY 4.0 license](https://creativecommons.org/licenses/by/4.0/)

Downloaded from: <https://hdl.handle.net/1887/3566358>

**Note:** To cite this publication please use the final published version (if applicable).



# Tandem Substitutions in Somatic Hypermutation

Julieta H. Sepúlveda-Yáñez<sup>1,2</sup>, Diego Alvarez Saravia<sup>2</sup>, Bas Pilzecker<sup>3,4</sup>, Pauline A. van Schouwenburg<sup>5</sup>, Mirjam van den Burg<sup>5</sup>, Hendrik Veelken<sup>1</sup>, Marcelo A. Navarrete<sup>2</sup>, Heinz Jacobs<sup>4</sup> and Marvyn T. Koning<sup>1\*†</sup>

<sup>1</sup> Department of Hematology, Leiden University Medical Center, Leiden, Netherlands, <sup>2</sup> School of Medicine, University of Magallanes, Punta Arenas, Chile, <sup>3</sup> Department of Tumor Immunology, Radboud Institute for Molecular Life Sciences, Nijmegen, Netherlands, <sup>4</sup> Division of Tumor Biology and Immunology, Netherlands Cancer Institute, Amsterdam, Netherlands, <sup>5</sup> Department of Pediatrics, Leiden University Medical Center, Leiden, Netherlands

## OPEN ACCESS

### Edited by:

Sylvie Hermouet,  
U1232 Centre de Recherche en  
Cancérologie et Immunologie Nantes  
Angers (CRCINA)(INSERM), France

### Reviewed by:

Yu Zhang,  
Western Michigan University,  
United States  
Richard Chahwan,  
University of Zurich, Switzerland

### \*Correspondence:

Marvyn T. Koning  
m.t.koning@lumc.nl

### †Present address:

Marvyn T. Koning,  
Department of Hematology, Leiden  
University Medical Center, Leiden,  
Netherlands

### Specialty section:

This article was submitted to  
B Cell Biology,  
a section of the journal  
Frontiers in Immunology

**Received:** 01 November 2021

**Accepted:** 16 December 2021

**Published:** 07 January 2022

### Citation:

Sepúlveda-Yáñez JH,  
Alvarez Saravia D, Pilzecker B,  
van Schouwenburg PA,  
van den Burg M, Veelken H,  
Navarrete MA, Jacobs H  
and Koning MT (2022)  
Tandem Substitutions in  
Somatic Hypermutation.  
Front. Immunol. 12:807015.  
doi: 10.3389/fimmu.2021.807015

Upon antigen recognition, activation-induced cytosine deaminase initiates affinity maturation of the B-cell receptor by somatic hypermutation (SHM) through error-prone DNA repair pathways. SHM typically creates single nucleotide substitutions, but tandem substitutions may also occur. We investigated incidence and sequence context of tandem substitutions by massive parallel sequencing of V(D)J repertoires in healthy human donors. Mutation patterns were congruent with SHM-derived single nucleotide mutations, delineating initiation of the tandem substitution by AID. Tandem substitutions comprised 5,7% of AID-induced mutations. The majority of tandem substitutions represents single nucleotide juxtalocations of directly adjacent sequences. These observations were confirmed in an independent cohort of healthy donors. We propose a model where tandem substitutions are predominantly generated by translesion synthesis across an apyrimidinic site that is typically created by UNG. During replication, apyrimidinic sites transiently adopt an extruded configuration, causing skipping of the extruded base. Consequent strand decontraction leads to the juxtalocation, after which exonucleases repair the apyrimidinic site and any directly adjacent mismatched base pairs. The mismatch repair pathway appears to account for the remainder of tandem substitutions. Tandem substitutions may enhance affinity maturation and expedite the adaptive immune response by overcoming amino acid codon degeneracies or mutating two adjacent amino acid residues simultaneously.

**Keywords:** apyrimidinic site, tandem dinucleotide substitutions (TDNS), strand slippage, tandem substitution, translesion synthesis (TLS), uracil-N-glycosylase (UNG)

## INTRODUCTION

To effectively counter the virtually limitless possibilities of pathogen-derived immune challenges, B lymphocytes - representing a critical arm of the adaptive immune system - can generate a virtually limitless repertoire of structural B-cell antigen receptor (BCR) variants through somatic hypermutation (SHM) (1–7). Upon encountering an antigen, antigen-activated B cells initiate SHM, as well as class-switch recombination, through activity of activation-induced cytosine deaminase (AID) (8, 9).

AID deaminates cytosine (C) to uracil (U) preferentially in nucleotide motif WRCY (where W denotes A or T; R denotes A or G; and Y denotes C or T) on both DNA strands (10–14). This deamination locally instigates various substitutions through various mutagenic processing pathways (13–17).

When the U remains unmodified it will instruct a template T to all polymerases, resulting in C to T and G to A transitions. However, a U in the DNA is usually efficiently detected by uracil DNA glycosylase (UNG), which cleaves the base from the sugar-phosphate backbone, thereby generating a non-instructive apyrimidinic (AP) site. The AP site normally initiates faithful base excision repair (BER) involving an AP endonuclease (APE) and POLB (18, 19). During SHM however, the AP site can serve as a non-instructive template for the translesion synthesis (TLS) polymerase REV1, a dCMP transferase that can only insert a C opposite the newly generated AP site, thereby enabling C to G and G to C transversions (20–22). Alternatively, a single strand break at the AP site generated by APE allows POLH in complex with monoubiquitinated homotrimeric DNA clamp and replication processivity factor PCNA (PCNA-Ub) to access the site and generate about 8% of all A/T mutations by error-prone long-patch BER (6, 23–26).

Otherwise, the uracil is recognized as a U/G mismatch by the MSH2-MSH6 mismatch recognition complex and initiates the formation of a single-stranded gap flanking the U/G mismatch (17, 27). Upon binding of the MSH2-MSH6 complex, exonuclease 1 excises the uracil-containing strand to initiate non-canonical mismatch repair (ncMMR), where POLH in complex with PCNA-Ub generates the majority of A/T mutations (24, 26, 28–31).

Taken together, low fidelity translesion synthesis (TLS) DNA polymerases enable the generation of a large part of the spectrum of nucleotide substitutions. In these short- and long-patch ‘repair’ pathways, error-prone TLS DNA polymerases such as REV1, POLH, POLZ and perhaps POLI introduce mutations both at the position of the deaminated cytosine as well as in its near vicinity (28, 32–34).

In principle, all these mechanisms lead to single nucleotide substitutions (SNS). However, the occurrence of AID-instigated contiguous or ‘tandem’ substitutions, especially tandem dinucleotide substitutions (TDNS), has been described for several species. The term tandem substitution refers to contiguous substitutions resulting from a single DNA damage event, which could in theory lead to multiple discrete single nucleotide substitutions during its repair. The contribution of tandem substitutions to all AID-induced substitutions differs per species, ranging from 1,6% in mice (35) to nearly 60% in sharks (36, 37). Data on the frequency of this phenomenon from human *ex vivo* experiments in the immunoglobulin loci are currently lacking, but genome-wide the frequency of TDNS is estimated to range from 0,1% to 1% (38–41).

The biological consequences of tandem substitutions remain unknown. One advantage for TDNS is that they overcome the redundancy in the amino acid code in positions in the BCR where non-synonymous mutations are beneficial. This would in theory allow for faster repertoire diversification and hence more effective affinity maturation and humoral immunity (35, 42).

The molecular mechanism underlying the generation of tandem substitutions is debated. Evidently, a proportion of the observed contiguous substitutions are not, in fact, single event tandem substitutions, but multiple independent SNS occurring in contiguous nucleotides (43, 44). This may hold especially true in and around canonical AID hot spot motifs, where consequently SHM is most efficient (37). It appears however, that the incidence of tandem substitutions is far higher than expected from clustered SNS alone, indicating the existence of a mechanism creating contiguous substitutions in a single event (35).

Despite its potentially substantial role in the adaptive immune response, the mechanisms responsible for tandem substitutions and their relative contribution to the process of SHM has not been investigated in humans since the introduction of massive parallel sequencing. Here, we investigated the incidence of tandem substitutions in human peripheral blood B cells (PBMC) in healthy donors and patients with DNA repair deficiency and analysed their substitution motifs. We aimed to confirm association of previously implicated DNA repair mechanisms involved in the creation of tandem substitutions, as well as identify other potential mechanisms. Additionally, we investigated whether tandem substitutions indeed overcome amino acid redundancy by analysing the topographical distribution of TDNS in the V allele.

## MATERIALS AND METHODS

### Sample Collection and Preparation

Peripheral blood samples were obtained with written informed consent from twelve healthy stem cell donors in compliance with the biobanking regulations of Leiden University Medical Center. Mononuclear cells were isolated by Ficoll separation and cryopreserved in aliquots. B cells were purified from aliquots of thawed cells by removal of non-B cells with magnetic beads (B cell isolation kit II; Miltenyi Biotec, Leiden, The Netherlands), routinely yielding a purity of >99% CD19<sup>+</sup> B cells as assessed by flow cytometry.

### V(D)J Library Generation

Aliquots of  $2 \times 10^6$  B cells were processed according to the ARTISAN PCR protocol for unbiased amplification of BCR repertoires, which has very low amplification and sequencing error rates of  $0.126 \times 10^{-3}$  per bp (45). Full-length IgM and IgG VDJ were amplified from all twelve donors, while IgA and IgE VDJ. VJ-kappa and VJ-lambda were amplified from six donors each (46). Libraries were barcoded, pooled and amplified as single molecules in rolling circles on a total of fourteen SMRT cells on the RSII system (Pacific Biosciences, Menlo Park, CA, USA). Output sequence files were filtered with SMRT portal software for a minimum of eight sequencing passes. All sequences were annotated by IMGT HighV-QUEST (47).

Additional datasets of healthy donors, as well as UNG-deficient and MSH2- and MSH6-deficient patients were obtained from publicly available sources (48, 49)

(**Supplementary Table S1**). Its amplification and sequencing methodology was previously validated and shown to be reliable for the identification VDJ of naïve and memory B cells alike (50).

## Sequence Selection

Sequences with identical V, D (if applicable), and J genes, identical CDR3 length and  $\geq 95\%$  pairwise identity in the nucleotide CDR3 sequence were analysed as a single sequence. Clonal expansions were reduced to the least mutated sequence to minimise the presence of amplification or sequencing errors. For the Leiden cohort healthy donors, sequences with no mutations, as well as sequences with  $>5\%$  mutations in their V region were excluded from further analysis to minimize chance occurrences of consecutive mutated nucleotides whilst maintaining a reasonable number of mutations per sequence. Such selection of oligomutated sequences could not be performed for the publicly available datasets, as these datasets contained solely (often highly mutated) IgG and IgA rearrangements which, upon equally stringent selection, would have led to greatly reduced library sizes, precluding any meaningful analysis (46). From the publicly available data, only sequences containing one of the 50 most commonly rearranged IGHV alleles were considered, offering the option to completely perform *in silico* correction of these sequences.

For fair comparison, sequences with  $>5\%$  mutations in the Leiden cohort were grouped in bins of 5-10% mutations, 10-20% mutations and  $>20\%$  mutations and analysed identically to the less mutated sequences (**Supplemental Data**). Because the main findings in these data sets in essence corresponded to the findings in the less mutated sequences, yet were more susceptible to methodological imperfections such as less accurate 'false tandem' correction (due to lower ratios of observed:simulated TDNS, see below), these sequences were not incorporated in the main analysis.

## Identification of Mutations

We identified all SNS, TDNS, and longer contiguous substitutions in the V region of the 50 most commonly used IGHV alleles, 25 most commonly used IGKV alleles, and 25 most commonly used IGLV alleles, IGLV9-49\*01 or IGLV10-54\*01 (the latter two to have all IGLV families represented). We identified the germline nucleotide(s), mutated nucleotide(s), context (1 nucleotide on either side of the substitution), and their position within the V allele using Geneious software v10.1.3 (Biomatters Ltd., Auckland, New Zealand). Insertions and deletions were disregarded in this analysis.

The similarity analysis between TDNS and the Doublet Base Substitution (DBS) Signatures deposited in the COSMIC catalog (v3.1 - June 2020) (44) was measured by cosine distance using R (51).

## Predicted Frequency of Contiguous Substitutions From Multiple Mutation Events

The contribution of contiguous substitutions resulting from two or more independently occurring adjacent SNS, was modelled *in silico*. To have adequate numbers of control sequences, only the 20 most abundant V alleles were used for modelling.

Considering the observed mutation frequency of SNS per position and the distribution of particular substitutions per position, germline sequences were mutated *in silico* to a total number of mutations matching the mutation load of sequences observed in the *in vivo* data. These calculations were performed 100.000 times each for all rearrangements. Output sequences were analysed for single and contiguous substitutions, SNS and TDNS motifs and nonsense mutations. Any mutation clusters thus observed were considered to represent 'false' tandem substitutions and were subtracted from the actually observed frequency of tandem substitutions to obtain the true frequency of tandem substitutions in the dataset. Calculations for the complete dataset were extrapolated from this majority subset.

For the detailed modelling strategy, see: <https://github.com/CATGUMAG/tandem-substitution-simulation>.

The CMMRD patient repertoires mostly lacked IgM and both the patient and Rotterdam healthy donor repertoires were sequenced using a different approach and sequencing platform than the Leiden healthy donor cohort, which resulted in higher mutation loads and more false TDNS. Therefore, we did not pool these repertoires in our *in silico* mutation algorithm, but processed them separately and then independently predicted SNS mutation clusters as described for the Leiden cohort.

The paucity of unique sequences in the UNG-deficient patient's library precluded detailed *in silico* correction as performed for the other datasets. To obtain the best approximation of corrected tandem substitution incidence in this dataset, *in silico* predictions were performed on pooled sequences of the IGHV3 and IGHV4 family.

## Mutational Resistance Scoring

A score was assigned to each position in IGHV, IGKV or IGLV to represent the chance that any random single nucleotide substitution will cause a synonymous mutation. For each codon in the human amino acid code, all 9 possible substitutions were classified as synonymous, non-synonymous or nonsense mutations. Since nonsense mutations would normally cause the BCR to be deleted from the repertoire and therefore not be represented in this dataset, these substitutions were disregarded and the chance of a random substitution leading to an amino acid change was calculated from the remaining options. Consecutively, for each position in the V region, proportional germline codon usage was calculated and multiplied by the chance that a codon would incur a synonymous mutation upon any random substitution. The results of all codons per position were added together to obtain the repertoire-broad chance of synonymous mutations per position (**Supplementary Table S2**).

This theoretically expected proportion of synonymous mutations was compared with the observed proportion of synonymous single nucleotide substitutions. When the observed frequency of synonymous mutations is higher than expected, this might indicate that BCR with non-synonymous mutations in this position have undergone negative selection and are no longer found in the repertoire. Conversely, lower observed synonymous mutation rates than expected may indicate a selective advantage for BCR altering their sequence in this

position. However, it should be noted that this method assumes random substitutions across the V region whereas in reality mutations preferentially cluster around AID motifs and some substitutions occur preferentially over others (52). This method is therefore more suited for analysis of V region-wide topographical distribution of mutations rather than scrutiny of a single position.

After calculating the permissiveness of every position to incur non-synonymous mutations, the number of TDNS per position was counted and compared to the pattern of SNS to assess whether tandem substitutions are distributed randomly or preferentially cluster in certain areas of the V region.

We applied a Quasi-Poisson regression to establish if the number of tandem substitutions for each position in the V allele is correlated with the mutational resistance score. The analysis was performed using the R Stats Package (51).

## RESULTS

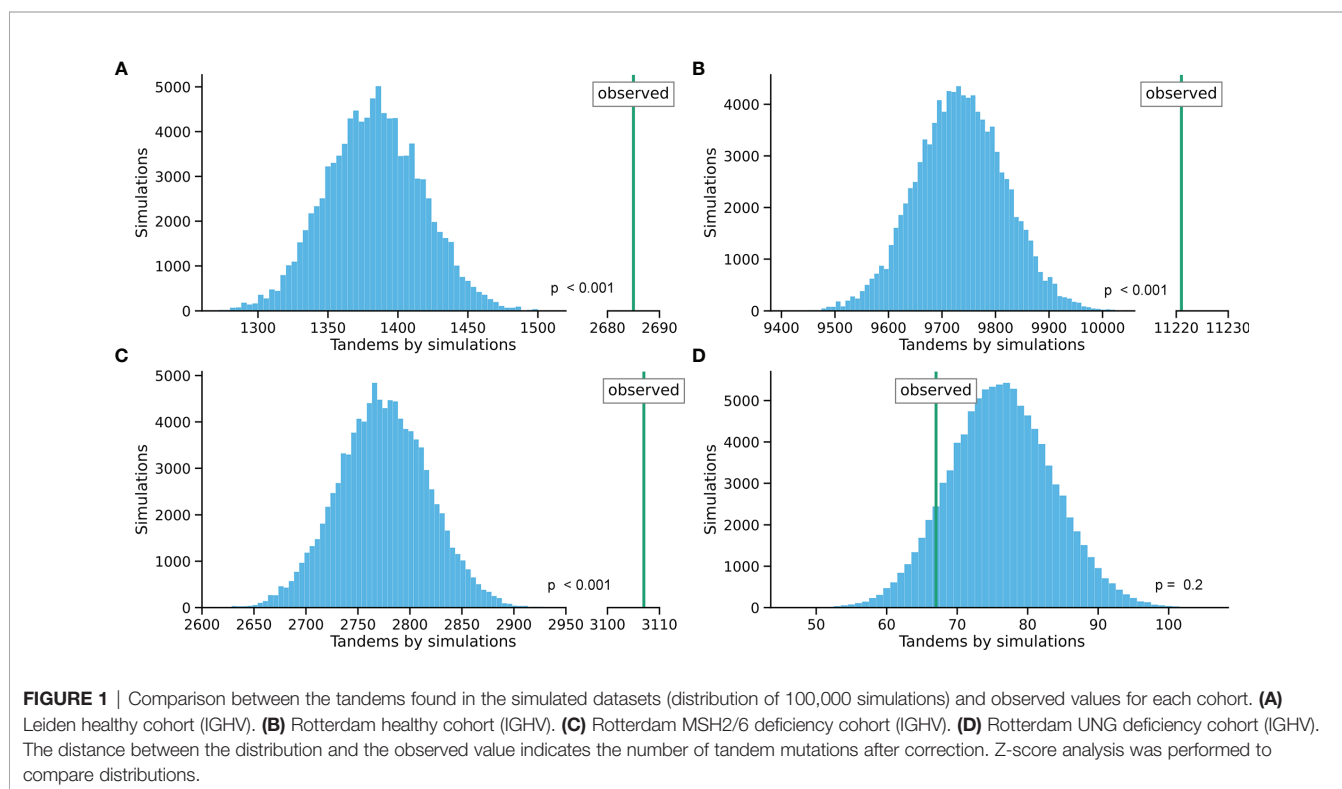
### 6% of Substitutions Are Tandems

Twelve healthy donor peripheral blood B-cell receptor repertoires were sequenced using full-length, unbiased, massive parallel V(D)J sequencing. Selection of unique, clonally unrelated, antigen-experienced sequences carrying up to 5% mutations yielded 13,532 VDJ, 7,952 VJ-kappa and 7,598 VJ-lambda. Comparison to the closest germline allele allowed for identification of a total of 122,878 single nucleotide substitutions (SNS), 10,735 tandem dinucleotide substitutions (TDNS) and 2,615 longer contiguous substitutions. The longest contiguous

substitutions were 8 nucleotides in length (**Supplementary Table S3**).

Since mutation clusters of independently generated, but adjacent SNS are indistinguishable from single event tandem substitutions, we calculated the number of expected clustered SNS for each individual repertoire through *in silico* modeling (IGHV **Figure 1**; IGKV and IGLV and pooled IGV **Supplementary Figure S1**; **Supplemental Data 1.1-1.23**). Using the 20 most abundantly observed IGV alleles, synthetic immunoglobulin repertoires of matching size and V allele usage distribution were mutated 100,000 times *in silico*. These *in silico* repertoires were designed to exactly match the mutation rate for each nucleotide position in the sequenced repertoire. Since all mutations within this model were introduced in single steps, observed mutation clusters in the modelled data were annotated as 'false' tandem substitutions.

Predicted 'false' tandem substitutions were subtracted from the total number of observed mutation clusters to obtain the frequency of single event tandem substitutions. Of all TDNS in the *in silico* simulated subset, 46,2% were predicted to represent two adjacent SNS generated in independent events. Extrapolating to the whole dataset (and counting the 'false' tandem substitutions as multiple SNS each), we observed 133,577 SNS and 5,775 TDNS. Therefore, the incidence of true, single event TDNS in human BCR *in vivo* was found to be 4,10%. Similar calculations for trinucleotide, tetranucleotide and pentanucleotide contiguous substitutions respectively showed incidence rates of 1,18%, 0,31% and 0,12%, containing 12,5%, 3,0% and 1,0% false positives, respectively. Cumulatively, 5,7% of all substitutions in human BCR were single event tandem



substitutions, making their incidence rate much higher than observed in mice and in genomic human sequences outside the immunoglobulin loci (**Supplementary Table S4**).

## The Distribution of TDNS Matches SNS

The distribution of all SNS and TDNS occurring in the most abundantly rearranged V alleles was mapped geographically up to IMGT amino acid position C104. As expected, mutation frequencies in complementarity determining regions (CDR) exceeded those of framework regions (FR). Specifically, structurally essential FR residues such as C23, W41 and C104 were strikingly less mutated than surrounding residues. Overall, the distribution of TDNS resembled those of SNS, indicating that mechanisms governing the generation and selection of TDNS are closely related to those of SNS (**Figure 2, Supplementary Figure S2A and Supplementary Figure S3A**). Also, the SNS follow an equal distribution pattern of mutations (Ts/Tv: 1,05) corresponding to previously described datasets (53) (**Supplementary Table S5A**). Despite their overall similarities, the relative abundance of TDNS differed from SNS in a number of positions. TDNS were overrepresented in a number of FR residues in the IGLV dataset (**Figure 2, Supplementary Figure S2B and Supplementary Figure S3B**). More commonly, however, FR residues contained fewer TDNS than expected. This relative scarcity of TDNS was most profound around structurally essential FR residues, where the observed SNS generally encoded for synonymous mutations (**Figure 2, Supplementary Figure S2C and Supplementary Figure S3C**). We hypothesized that TDNS in FR would be selected against, given how their higher potential for non-synonymous/replacement mutations (see below) would easily lead to deleterious effects on BCR integrity.

To test this hypothesis, we predicted the proportion of synonymous mutations at each position (Materials and Methods) and compared this to the observed proportion of synonymous mutations at that position. Subtracting the former from the latter resulted in a mutational resistance score, where progressively higher scores represented positions with more synonymous mutations than expected. Putatively, positions with high mutational resistance scores result from negative selection of BCR with non-synonymous mutations in these positions.

Indeed, we observed relatively few TDNS in positions with higher mutational resistance scores and we confirmed a negative selection of tandem substitutions in these structurally important positions using a Quasi-Poisson regression (**Figure 2, Supplementary Figures S2D, S3D, Supplementary Table S6, p-value<0,05**).

## Tandem Mutations Overcome Codon Redundancies

As more nucleotides are mutated at once, the chance of non-synonymous, i.e. amino acid replacement mutations, increases. To determine how often this occurred, TDNS were assigned to three groups according to their position in the codon: position 1 for substitutions of the 5' base in the coding strand and the

middle base, position 2 for the middle and the coding strand 3' base, and position 3 for the 3' base of one codon and the 5' base of the subsequent codon. Substitutions in position 2 may only cause non-synonymous or nonsense mutations, whilst substitutions in position 1 and 3 may rarely cause synonymous mutations. Overall, 98,6% of TDNS encoded for at least one amino acid replacement, while only 71,6% of SNS encoded a non-synonymous mutation. Apparently, TDNS have the potential to expedite amino acid changes and thereby are likely to enhance adaptive immune responses.

Additionally, TDNS in position 3 may potentially mutate two adjacent residues simultaneously, which was observed in 18,5% of such position 3-situated TDNS. It appears that this potential for accelerated affinity maturation comes at a price, because position 3-situated TDNS were less abundant than the other positions, supposedly due to negative selection of structurally unsound BCR or nonsense mutations. An exception was found in the IGLV library, where some of the previously described hotspots were situated in position 3 and compensated for the lower abundance overall (**Figure 3**). Whether such relative scarcity of position 3 TDNS indeed resulted from increased negative selection pressure, could not be tested in this reverse immunology investigation.

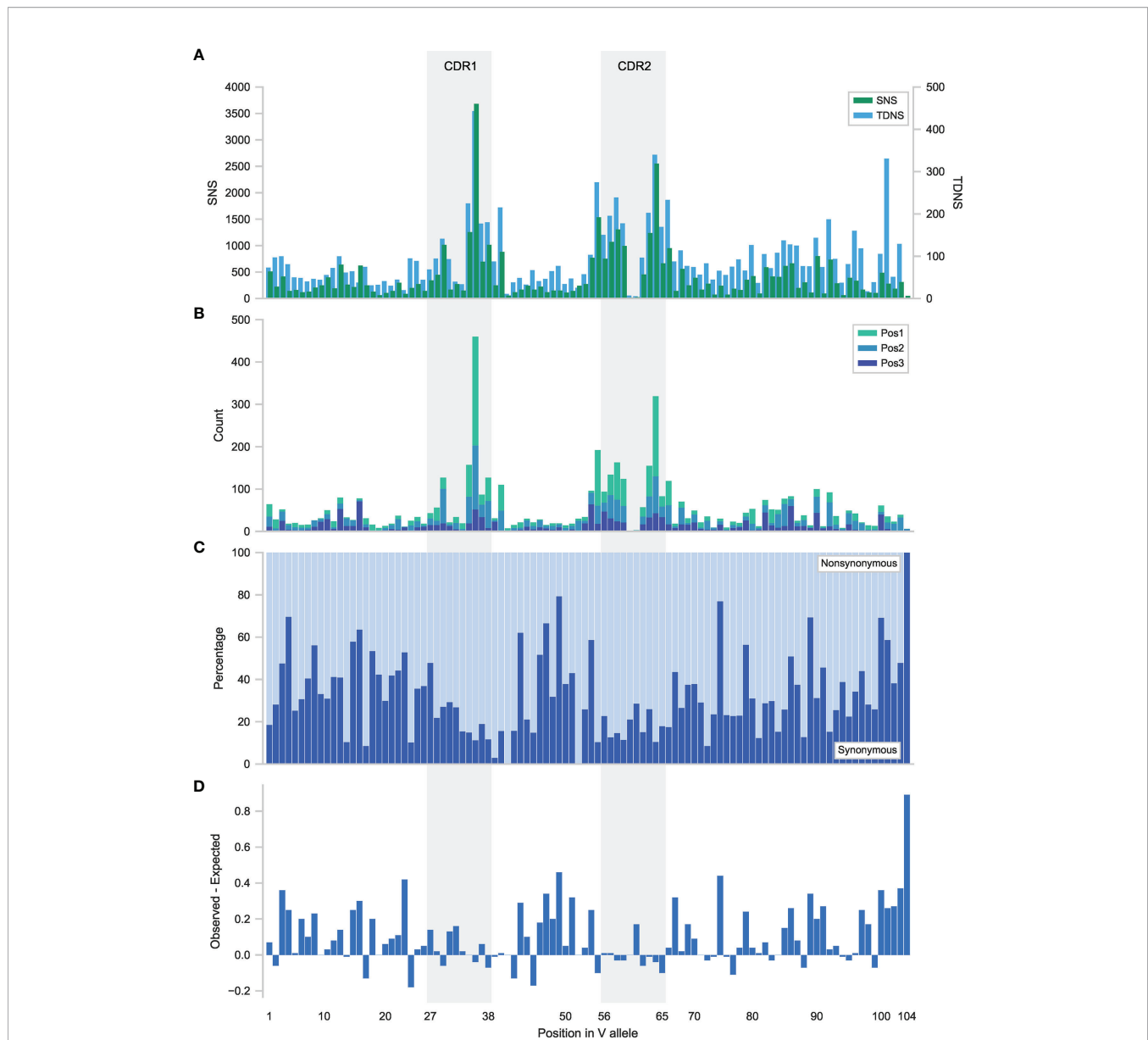
## Most Tandem Mutations Are Juxtalocations

After subtraction of *in silico* predicted 'false tandem' TDNS, substitution tables were generated for IGHV, IGKV and IGLV collections and for the complete dataset (IGV **Figure 4; IGHV, IGKV and IGLV Supplementary Figure S4; Supplementary Tables S7-S9 and Supplemental Data 1.1-1.23**). Heavy and light chain substitution tables were comparable, except for CT to TA and GA to AG TDNS, which were more prevalent in IGLV, and IGKV+IGLV, respectively. We postulate that these observations stem from differences in the reference sequences, in which these bases may be more prone to mutation, or such mutations are more permissible and lead to less negative selection.

TDNS did not occur in equal frequencies, nor did they follow substitution patterns expected from SNS tables. Instead, half (49,9%) of all TDNS occurred in dipurine or dipyrimidine motifs and virtually all mutations swapped the positions of either one of the reference bases (i.e. AG to CA; 51,5% versus 44,4% expected; Student's t-test:  $p < 0,0001$ ), or both bases at once, leading to inversions (i.e. AG to GA; 37,3% versus 11,1% expected; Student's t-test:  $p < 0,0001$ ). Only 10,4% (expected: 44,4%; Student's t-test:  $p < 0,0001$ ) of TDNS did not follow this pattern (**Figure 4 and Supplementary Table S10A**).

Additionally, we compared the TDNS signature with the recently described doublet base substitutions (DBS) signatures from COSMIC (v3.1 - May 2020). We found that the tandem substitution signature in immunoglobulins does not correspond to a previously described pattern (cosine similarity  $< 0,5$ , **Supplementary Table S11**).

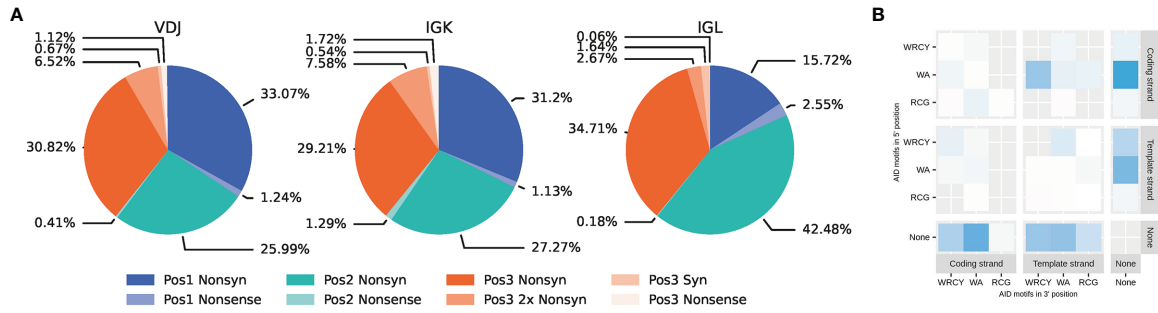
Whenever a TDNS reference sequence was frequently found in the TDNS table, their reverse complement would also be



highly prevalent. However, within such pairs, the motifs that contained a cytosine on the transcribed/non-coding strand, which during replication serves as a template for lagging strand synthesis, were more frequently mutated than their non-transcribed partner, following the pattern previously described for AID-induced mutations (53) (**Supplementary Figure S5**). Indeed, a strand-bias analysis revealed that tandem substitutions preferentially occurred on the template strand, and that canonical (WRCY) and non-canonical (WA; POLH-

associated) (43, 54) AID motifs were most commonly found around the 5' base of the template strand sequence (**Figure 3**). These observations implicate that tandem substitutions preferentially arise at AID-induced DNA lesions.

To identify a common mutational signature, we mapped the nucleotides directly adjacent to tandem substitutions. Scrutiny of these contexts revealed that tandem substitutions commonly already contained the mutated sequence in the overlap between the reference sequence and its context (**Figure 5**). In other words,



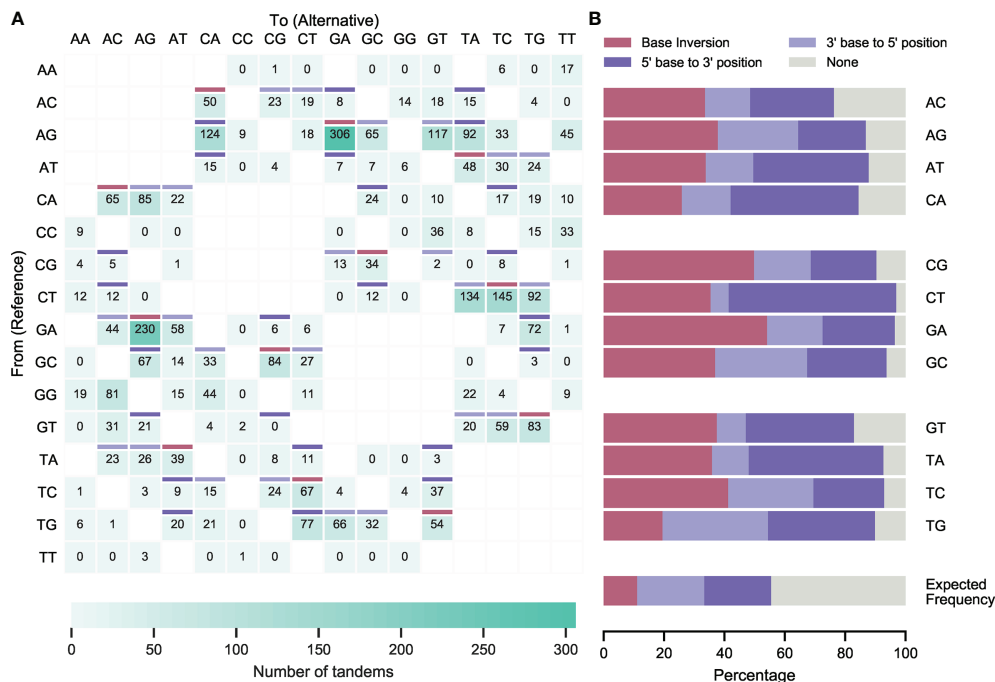
**FIGURE 3 |** Relative contributions of synonymous, non-synonymous and nonsense tandem dinucleotide substitutions for VDJ, VJ-kappa and VJ-lambda. **(A)** Results are split per position in the codon. Position 1 represents 5' and middle base on the coding strand, position 2 represents middle and 3' base and position 3 represents the 3' base of the codon and the 5' base of the downstream codon. Tandem dinucleotide substitutions causing two separate synonymous mutations in position 3 are displayed as a separate category. **(B)** Analysis of AID motifs present in tandem substitution for IGHV sequences. The 3 motifs associated with AID activity (WRCY, WA, and RCG) were identified in position 5' or 3' of the tandem. The motifs were considered in forward (coding) and reverse (template) direction.

tandem substitutions apparently derive from small juxtaposition events, where directly adjacent sequences move a single position upstream or downstream, thus explaining the abovementioned observation of one of the original bases moving to the other position in the motif. This mechanism also explains the high incidence of inverting substitutions, as they may result from both upstream and downstream juxtapositions. Although, not found

in sufficient abundance to allow for statistical testing, it appeared that tandem substitutions longer than TDNS followed similar patterns.

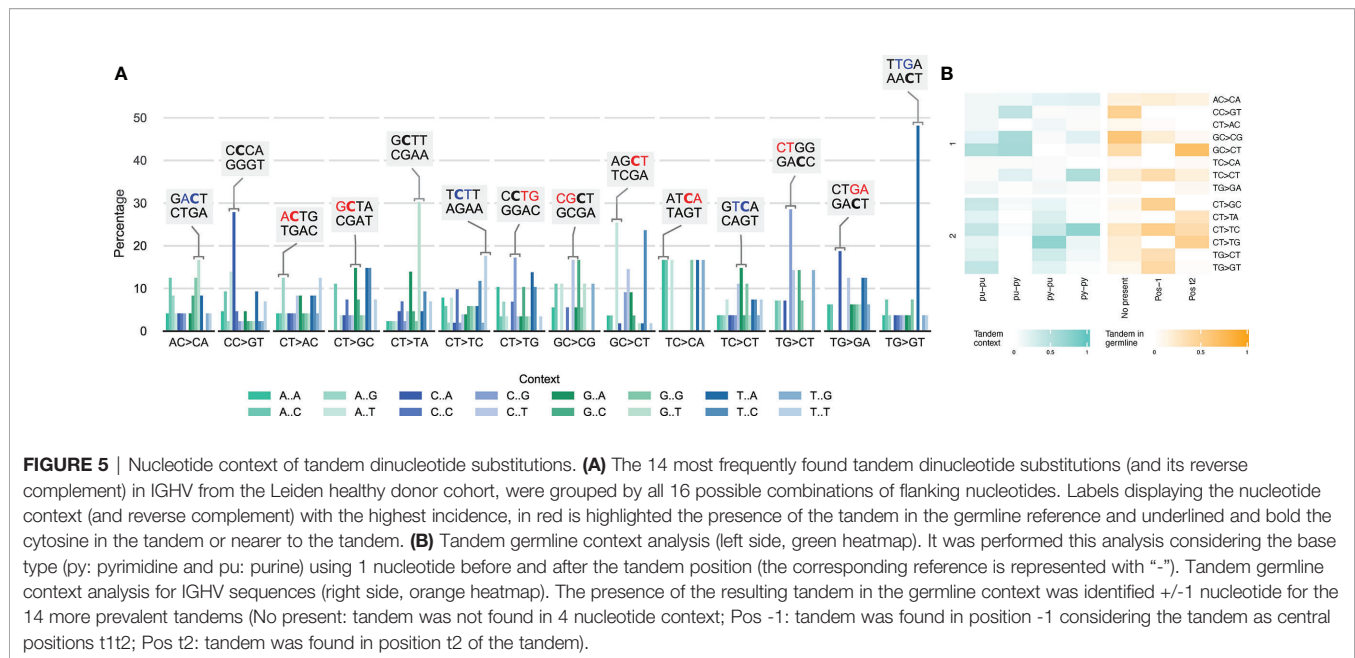
### Polydipyrimidine Stretches Are Favored

Following the observations that TDNS occur preferentially in dipyrimidine motifs, that the cytosine-containing strand is



**FIGURE 4 |** Corrected incidence of tandem dinucleotide substitutions in healthy donors. **(A)** Dinucleotide substitutions from unique IGHV, IGKV and IGLV sequences from the Leiden healthy donor cohort and corrected after *in silico* predictions of dinucleotide substitutions that did not occur in tandem. Burgundy cells represent sequence inversions, light and dark purple cells represent juxtapositions of the 5' and 3' base in the pair (as seen from the non-transcribed strand), respectively. For unshaded cells, juxtaposition could not be assessed due to one or more nucleotides in the reference sequence matching the mutated sequence. **(B)** Relative contribution of sequence inversions and juxtapositions.





dominantly targeted, and that tandem substitutions in majority represent single nucleotide juxtapositions, tandem substitutions should preferentially arise from polydipyrimidine stretches. Indeed, the substitution tables show that most of the dominantly observed TDNS motifs fit within this hypothesis (**Figure 4** and **Supplementary Table S10**).

Notable exceptions to this rule were the commonly observed AG to GA and GC to CG inverting substitutions, which contained the mutated sequence in their germline context in less than half of instances.

## VDJ Tandem Substitutions in UNG and MMR Deficiency

After AID deaminates a C into a U, immunoglobulin gene mutational patterns are governed by UNG-initiated BER and the non-canonical MMR pathway. Based on knockout mouse models, tandem substitutions were described to predominantly occur in the MMR pathway (55). Since experimental models to mimic SHM in humans are not available, the closest approximation to test the relative contribution of BER and MMR is to analyze BCR repertoires from DNA repair deficient patients.

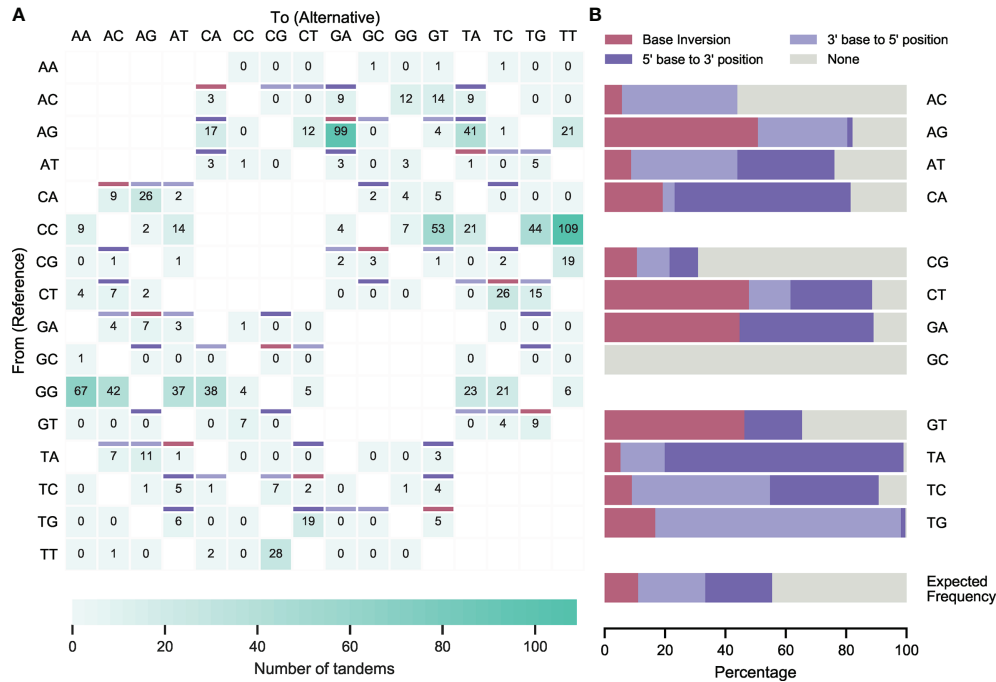
We obtained a massive parallel sequencing library with VDJ from one MSH2-deficient and three MSH6-deficient patients. All patients carried biallelic defects in their respective genes, leading to Constitutional Mismatch Repair Deficiency (CMMRD) syndrome (56, 57). An additional VDJ library was obtained from an UNG-deficient patient. As internal controls, we analyzed an independent cohort of healthy donors generated by the same methodology as these patient datasets (henceforth referred to as: 'Rotterdam healthy donor cohort') (49). Sequences were filtered identically as described for the Leiden healthy donor cohort, except for the selection by mutation load (see *Materials*

and *Methods*), yielding 7.654 unique VDJ sequences from the CMMRD patients and 104 unique VDJ sequences from the UNG-deficient patient.

The CMMRD patient repertoires mostly lacked IgM and both the patient and Rotterdam healthy donor repertoires were sequenced using a different approach and sequencing platform than the Leiden healthy donor cohort, which resulted in higher mutation loads and relatively more false TDNS. Therefore, we did not pool these repertoires in our *in silico* mutation algorithm, but processed them separately and then independently predicted SNS mutation clusters as described for the Leiden cohort.

After *in silico* correction, tandem dinucleotide substitutions in the Rotterdam healthy donor cohort confirmed the previously observed paradigm of juxtapositions in TCT-motifs (**Supplementary Figure S6**, **Supplementary Table S10B**, **Supplemental Data 2.1-2.42**). In this dataset, inversions and base swaps were less dominant in motifs other than the dipyrimidine ones, putatively as a result of the more challenging *in silico* correction of this more artefact-prone dataset compared to the Leiden healthy donor cohort. Nevertheless, in the corrected tables, TDNS still represented 1,9% of all substitutions, almost twice the frequency of the previously highest estimate.

After removal of *in silico* predicted 'false' TDNS, the dinucleotide substitution table of the CMMRD patients also showed great similarities to the Leiden healthy donor cohort, suggesting that the mismatch repair machinery does not play a major role in the formation of tandem substitutions in humans (**Figure 6**, **Supplementary Table S10C**, **Supplemental Data 2.1-2.32**). The corrected proportion of TDNS in all substitutions was 2,0%, comparable to the 1,9% in the equally extensively mutated sequences from the Rotterdam healthy donor cohort. Notably, the SNV show a moderately increased Ts : Tv ratio of 1,91,



**FIGURE 6 |** Corrected incidence of tandem dinucleotide substitutions in MSH2/6 deficiency. **(A)** Dinucleotide substitutions from unique IGHV sequences obtained from constitutional mismatch repair deficiency (CMMRD) patients and corrected after *in silico* predictions of dinucleotide substitutions that did not occur in tandem. Burgundy cells represent sequence inversions, light and dark purple cells represent juxtalocations of the 5' and 3' base in the pair (as seen from the non-transcribed strand), respectively. For unshaded cells, juxtalocation could not be assessed due to one or more nucleotides in the reference sequence matching the mutated sequence. **(B)** Relative contribution of sequence inversions and juxtalocations.

consistent with DNA repair deficiency during somatic hypermutation (Supplementary Table S5B).

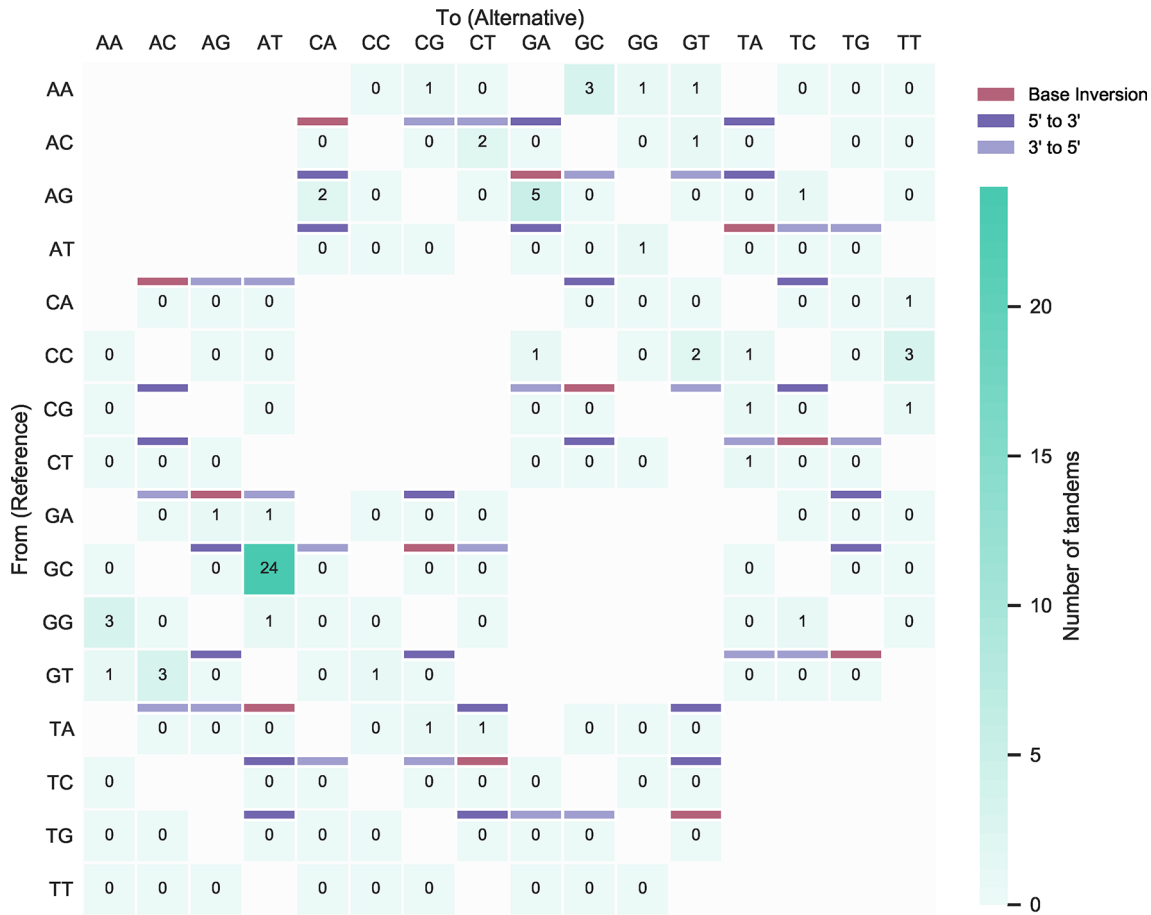
Conversely, the previously identified substitution patterns could not be reproduced in the dinucleotide substitution table of the UNG-deficient patient. In the SNS, the Ts : Tv ratio was skewed towards transitions (3,84), consistent with the ‘replication without repair’ pathway leading to C to T transitions in the absence of BER (Supplementary Table S5C). Rather than base swaps and inversions, contiguous substitutions dominantly also resulted from this ‘replication without repair’ pathway, representing AID-induced C to T transitions in dinucleotide motifs consisting solely of cytosine and guanine bases (Figure 7). This observation remained consistently different from the other datasets when these were randomly sampled down to the smaller library size of this (Supplementary Figure S7). Indeed, 81% of these ubiquitous GC to AT transitions were located in AID hotspots, which are known to favor transversions only with UNG (over)expression (58). The paucity of unique sequences in this library precluded detailed *in silico* correction as performed for the other datasets. To obtain the best approximation of the incidence of corrected tandem substitutions in this dataset, *in silico* predictions were performed on pooled sequences of the IGHV3 and IGHV4 families, respectively. These analyses predicted 76 mutation clusters resulting from adjacent SNS events, of which 67 were actually

observed in the dataset (Figure 1 and Supplemental Data 4.1-4.3). Thus, no tandem substitutions were observed in the absence of the critical, BER-initiating component UNG. Indeed, this is in line with our previous observation that tandem substitutions are more abundant in immunoglobulin loci than the genome at large, as UNG activity is mainly directed to these genetic regions. Additionally, relative absence of faithful DNA repair mechanisms during SHM makes UNG-derived AP sites more vulnerable to tandem substitutions than genome-wide events of spontaneous deamination (59).

While these findings are congruent with the hypothesis that UNG may be involved in the formation of the majority of tandem substitutions in humans, it should be noted that the n=1 sample size precludes any definitive conclusions.

## DISCUSSION

We describe the ubiquitous presence of tandem substitutions in human V(D)J rearrangements. Tandem substitutions contribute to the acquisition of mutations during SHM in a frequency at least two to five times higher than the previously highest estimate. In fact, tandem substitution events are likely to be still underestimated by this study, since any TDNS events resulting in a mutation where one of the bases matches the

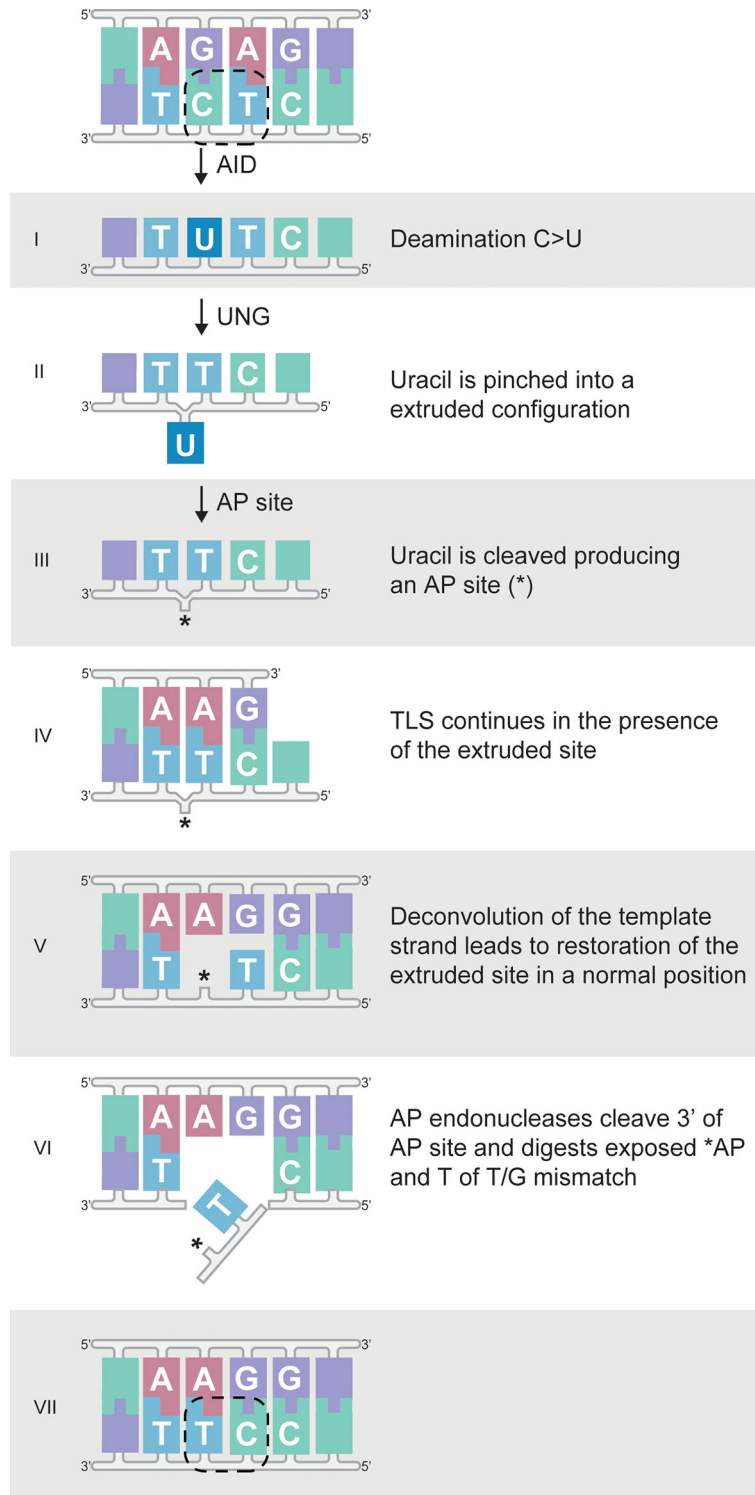


**FIGURE 7** | Incidence of tandem dinucleotide substitutions in UNG deficiency. All dinucleotide substitutions from unique IGHV sequences obtained from an UNG deficient patient. Burgundy cells represent sequence inversions, light and dark purple cells represent juxtalocations of the 5' and 3' base in the pair (as seen from the non-transcribed strand), respectively. For unshaded cells, juxtalocation could not be assessed due to one or more nucleotides in the reference sequence matching the mutated sequence. Results could not be reliably corrected for dinucleotide substitutions that did not occur in tandem for each mutation independently due to the size of the dataset. However, simulation of dinucleotide substitutions resulting from independent single nucleotide substitutions showed an almost identical number of dinucleotide substitutions (n=76) as the number that was observed (n=67), suggesting that this UNG deficient patient has no tandem substitution events.

reference sequence would appear as a regular SNS. Tandem substitutions have the potential to expedite the adaptive immune response by overcoming amino acid code redundancy and by incidentally mutating multiple residues at once. Clustering of such mutations around AID hotspots and their overall distribution indicates that tandem substitutions are an integral part of the SHM spectrum.

Virtually all tandem mutations adhere to a previously unrecognised substitution pattern resulting from single nucleotide juxtalocations. Confirmation of these findings in the independent Rotterdam cohort generated by a different methodical approach, albeit at a lower frequency, ensures that our observations are not a result of methodological artefacts. As an explanation for this phenomenon, we propose the EXPEDITE (EXtruded Pinching Effecting DIrectional Tandem Exchange) model. In this model, an abasic site transiently adopts an extruded configuration leading to misreading of template DNA

during replication. Our data indicate that AID-mediated deamination of a cytosine in polydypyrimidine stretches is the main route to the juxtalocations that underlie tandem substitutions. After recognition of the uracil, UNG cleaves the N-glycosylic bond to create an apyrimidinic (AP) site, a known risk for strand slippage (60). Additionally, UNG creates ~45° kinks at dUTP positions (61), which facilitate the extrusion and subsequent skipping of the abasic site by a DNA damage tolerant DNA polymerase. Considering the polydypyrimidine motifs, such extruded positioning could be facilitated by the relatively small flanking pyrimidine bases, in a similar fashion as was recently described for flanking cytosine bases (62). Indeed, mismatched pairs have an increased propensity for spontaneous base flipping (63), and the presence of UNG further increases the lifetime of open states following spontaneous base-pair dissociation and twists the unstacked nucleotides out of their helical conformation (64–66). These



**FIGURE 8** | Proposed EXPEDITE model for tandem substitutions causing juxtaposition of adjacent residues during DNA replication. (I) Cytosine (C) is deaminated to uracil (U) in a polypyrimidine context (shown here as an example since this context typically contains the most tandem substitutions, refer to ). (II) UNG pinches the U into an extruded configuration and (III) U is cleaved, leaving an AP site (\*). (IV) If during DNA replication the resulting extruded AP site is tolerated, the juxtaposed base acts transiently as a substitute template. (V) After decontraction of the template strand, the AP site is reverted to a normal configuration and replication continues from the ‘mismatch’ position containing the AP site. (VI) Hereafter, the AP site, as well as the mismatch, is removed by canonical base excision repair. (see *Discussion*). (VII) Final repair synthesis and ligation.

mechanisms might be considered intrinsic to the canonical BER pathway, but lead to a previously unrecognized effect.

Subsequent repositioning of the pinched or extruded AP site or base into the strand backbone would move the sequence one base upstream. Replication by a translesion synthesis DNA polymerase capable of extending from the mismatch ensures further extension of the novel DNA strand. Although the events in this model now no longer follow the canonical BER pathway, recruitment of an AP endonuclease (APE), as a downstream component of BER, is conceivable. APE cleaves 3' of the AP site and removes the adjacent mismatched base. Finally, the newly generated gap in the template strand is filled and ligated by faithful DNA repair mechanisms, thereby completing the TDNS process (**Figure 8**). The length of the gap, minus any chance matches at its termini, determines the length of the tandem substitution. It should be stressed that the EXPEDITE model is a best estimate proposal, and the exact proteins and mechanisms involved, especially in the non-canonical mechanism following UNG involvement, are subject to speculation and await verification.

Although AP sites, mismatched bases, and uracil residues have a particular proclivity to adopt extruded configurations (62, 63, 67–69), juxtalocations could theoretically also result from spontaneous base flipping of matched base pairs, especially A-T pairs (66). Indeed, we observe that germline motifs AT and TA, which contain no cytosine bases on either strand and would therefore remain unexplained through the UNG pathway, are the least abundant reference motifs of all tandem substitutions - yet are not absent. Furthermore, they do follow the paradigm of single nucleotide juxtalocated outcomes. Alternatively, these juxtalocations may be part of a longer tandem substitution initiated by a cytosine that is near, but not inside the motif, and other bases in this longer tandem substitution remain identical after juxtalocation, therefore not registering as mutated.

Not all tandem substitutions follow the juxtalocation paradigm. Most importantly, the common (top strand) AG to GA inverting substitution contained the mutated sequence and their direct context in only less than half of cases, suggesting that these may stem from an alternative mechanism. We consider the following: deamination of the C in the bottom strand leads to U, and subsequent replication across U results in a G to A mutation on the top strand (canonical C to T on the bottom strand). Following removal of the U on the bottom strand by either canonical UNG or canonical MMR and patch repair of the bottom strand by POLH, replication across from the new WA (POLH hotspot) introduces a C across from the A. This will generate the observed AG to GA substitution (**Supplementary Figure S8**). Thus it appears that similar to previous observations in mice, MMR is to some extent also involved in tandem substitution generation in humans. However, in contrast to murine studies (42), a majority of tandem substitutions in humans appear to depend on UNG activity in the BER pathway, marking a notable species difference.

A number of specific TDNS, most importantly GC to AA/TT and TC to AA, have previously been attributed to POLZ and/or POLI activity following observations in *Saccharomyces cerevisiae*

and murine models (35, 42, 70, 71). However, these TDNS are among the rarest in each of our datasets and therefore, at least in humans, do not seem to represent significant additional pathways beyond the mechanisms described in this manuscript. Supposedly, these differences derive from a previously remarked species difference concerning tandem substitution formation (35). Indeed, murine translesion polymerases create tandem substitutions through the MMR pathway (42), whilst experiments in human cell lines have implicated POLI in tandem substitution formation, but rather through the involvement of UNG (72, 73). The data in this report propose a new model where the actual tandem substitution inducing lesion, the AP site, does not serve as a non-instructive template but rather causes juxtapositioning by template flipping. Several characteristics identify POLH as the prime candidate translesion synthesis polymerase to create tandem substitution during SHM: The requirement of a large catalytic site, the open active POLH site that can accommodate non-Watson-Crick base pairs, and the high POLH error rate of  $3.5 \times 10^{-2}$  which is associated with its ability to bind dNTP without DNA substrate (74, 75).

All observations in this study derive from a reverse immunology approach. The lack of experimental systems to study SHM in humans currently precludes controlled experimental confirmation of our proposed mechanism. Therefore, analysis of gene-deficient patients as performed in this study serves as a preliminary attempt of experimental confirmation. When testing the EXPEDITE hypothesis against a small cohort of DNA repair deficient patients, we found an apparent absence of tandem substitutions in UNG but not in MSH2/6 deficiency. Although these findings corroborate the described model with a strict dependence on abasic sites, the small number of individuals and sequences and the more error-prone sequencing approach preclude any definitive confirmation. Therefore, the analysis of additional UNG-deficient patients with high-fidelity BCR amplification and sequencing would be highly desirable. Unfortunately, homozygous UNG deficiency is an exceptionally rare disease (76), and the national reference centers for immunodeficiency of The Netherlands (Erasmus Medical Center Rotterdam), France (Hôpital Necker-Enfants Malades, Paris), and Germany (University Medical Center Freiburg) do not have additional UNG-deficient patients in their registries and were also unable to provide us material from the three originally described patients (77). We recommend that whenever additional research material comes available, additional B-cell receptor repertoires are sequenced and results are shared with the scientific community to elucidate the exact role of UNG in the formation of tandem substitutions.

## DATA AVAILABILITY STATEMENT

All the other data supporting the findings of this study are available within the article and in the repository: <https://catg.cl/>

lab/our-articles/jsepulveda-tandem-substitutions/. The code that was written for this project is publicly available at: <https://github.com/catg-umag/tandem-substitution-simulation>.

## ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the local ethics committee (no. B16.039), Leiden University Medical Center. Written informed consent to participate in this study was provided by the participants' legal guardian/next of kin.

## AUTHOR CONTRIBUTIONS

JS-Y and MK designed the study. MK performed experiments. JS-Y and MK performed analyses. DA-S and MN performed bioinformatics experiments and analysis. JS-Y, BP, HJ, and MK proposed the model. HV, MB, MN, and HJ advised on the

project. MK and HJ wrote the manuscript. All authors have read and approved the manuscript.

## FUNDING

MK was supported by a grant from the Professor Steenhuis Fonds. HJ received TOP Grant [91213018] from the Dutch Scientific Organization NWO-ZonMW. JS-Y received a Doctorado Becas Chile [2016-72170683] from Agency for Research and Development (ANID). MN and DA-S work was supported by Fondo Nacional de Desarrollo Científico y Tecnológico (Fondecyt) [1180882], [11140542], and [MAG1895].

## SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: <https://www.frontiersin.org/articles/10.3389/fimmu.2021.807015/full#supplementary-material>

## REFERENCES

- Kinoshita K, Honjo T. Linking Class-Switch Recombination With Somatic Hypermutation. *Nat Rev Mol Cell Biol* (2001) 2:493–503. doi: 10.1038/35080033
- Neuberger MS, Milstein C. Somatic Hypermutation. *Curr Opin Immunol* (1995) 7:248–54. doi: 10.1016/0952-7915(95)80010-7
- Papavasiliou FN, Schatz DG. Somatic Hypermutation of Immunoglobulin Genes: Merging Mechanisms for Genetic Diversity. *Cell* (2002) 109 Suppl: S35–44. doi: 10.1016/S0092-8674(02)00706-7
- Rajewsky K. Clonal Selection and Learning in the Antibody System. *Nature* (1996) 381:751–8. doi: 10.1038/381751a0
- Seki M, Gearhart PJ, Wood RD. DNA Polymerases and Somatic Hypermutation of Immunoglobulin Genes. *EMBO Rep* (2005) 6:1143–8. doi: 10.1038/sj.embor.7400582
- Stavnezer J, Guikema JE, Schrader CE. Mechanism and Regulation of Class Switch Recombination. *Annu Rev Immunol* (2008) 26:261–92. doi: 10.1146/annurev.immunol.26.021607.090248
- Storb U, Peters A, Kim N, Shen HM, Bozek G, Michael N, et al. Molecular Aspects of Somatic Hypermutation of Immunoglobulin Genes. *Cold Spring Harb Symp Quant Biol* (1999) 64:227–34. doi: 10.1101/sqb.1999.64.227
- Muramatsu M, Kinoshita K, Fagarasan S, Yamada S, Shinkai Y, Honjo T. Class Switch Recombination and Hypermutation Require Activation-Induced Cytidine Deaminase (AID), a Potential RNA Editing Enzyme. *Cell* (2000) 102:553–63. doi: 10.1016/S0092-8674(00)00078-7
- Muramatsu M, Sankaranand VS, Anant S, Sugai M, Kinoshita K, Davidson NO, et al. 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* (1999) 274:18470–6. doi: 10.1074/jbc.274.26.18470
- Alvarez-Prado AF, Perez-Duran P, Perez-Garcia A, Benguria A, Torroja C, de Yebenes VG, et al. A Broad Atlas of Somatic Hypermutation Allows Prediction of Activation-Induced Deaminase Targets. *J Exp Med* (2018) 215:761–71. doi: 10.1084/jem.20171738
- Dorner T, Foster SJ, Farner NL, Lipsky PE. Somatic Hypermutation of Human Immunoglobulin Heavy Chain Genes: Targeting of RGYW Motifs on Both DNA Strands. *Eur J Immunol* (1998) 28:3384–96. doi: 10.1002/(SICI)1521-4141(199810)28:10<3384::AID-IMMU3384>3.0.CO;2-T
- Pilzecker B, Jacobs H. Mutating for Good: DNA Damage Responses During Somatic Hypermutation. *Front Immunol* (2019) 10:438. doi: 10.3389/fimmu.2019.00438
- Rogozin IB, Diaz M. 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* (2004) 172:3382–4. doi: 10.4049/jimmunol.172.6.3382
- Rogozin IB, Kolchanov NA. Somatic Hypermutagenesis in Immunoglobulin Genes. II. Influence of Neighbouring Base Sequences on Mutagenesis. *Biochim Biophys Acta* (1992) 1171:11–8. doi: 10.1016/0167-4781(92)90134-L
- Di Noia J, Neuberger MS. Altering the Pathway of Immunoglobulin Hypermutation by Inhibiting Uracil-DNA Glycosylase. *Nature* (2002) 419:43–8. doi: 10.1038/nature00981
- Neuberger MS, Di Noia JM, Beale RC, Williams GT, Yang Z, Rada C. Somatic Hypermutation at A:T Pairs: Polymerase Error Versus dUTP Incorporation. *Nat Rev Immunol* (2005) 5:171–8. doi: 10.1038/nri1553
- Rada C, Di Noia JM, Neuberger MS. Mismatch Recognition and Uracil Excision Provide Complementary Paths to Both Ig Switching and the a/T-Focused Phase of Somatic Mutation. *Mol Cell* (2004) 16:163–71. doi: 10.1016/j.molcel.2004.10.011
- Rada C, Williams GT, Nilsen H, Barnes DE, Lindahl T, Neuberger MS. Immunoglobulin Isotype Switching Is Inhibited and Somatic Hypermutation Perturbed in UNG-Deficient Mice. *Curr Biol* (2002) 12:1748–55. doi: 10.1016/S0960-9822(02)01215-0
- Wiesendanger M, Kneitz B, Edelmann W, Scharff MD. 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. *J Exp Med* (2000) 191:579–84. doi: 10.1084/jem.191.3.579
- Jansen JG, Langerak P, Tsaalbi-Shtylik A, van den Berk P, Jacobs H, de Wind N. Strand-Biased Defect in C/G Transversions in Hypermutating Immunoglobulin Genes in Rev1-Deficient Mice. *J Exp Med* (2006) 203:319–23. doi: 10.1084/jem.20052227
- Krijger PH, Tsaalbi-Shtylik A, Wit N, van den Berk PC, de Wind N, Jacobs H. Rev1 Is Essential in Generating G to C Transversions Downstream of the Ung2 Pathway But Not the Msh2+Ung2 Hybrid Pathway. *Eur J Immunol* (2013) 43:2765–70. doi: 10.1002/eji.201243191
- Tissier A, Kannouche P, Reck MP, Lehmann AR, Fuchs RP, Cordonnier A. Co-Localization in Replication Foci and Interaction of Human Y-Family Members, DNA Polymerase Pol Eta and REV1 Protein. *DNA Repair (Amst)* (2004) 3:1503–14. doi: 10.1016/j.dnarep.2004.06.015
- Delbos F, Aoufouchi S, Faili A, Weill JC, Reynaud CA. DNA Polymerase Eta Is the Sole Contributor of a/T Modifications During Immunoglobulin Gene

- Hypermutation in the Mouse. *J Exp Med* (2007) 204:17–23. doi: 10.1084/jem.20062131
24. Krijger PH, Langerak P, van den Berk PC, Jacobs H. Dependence of Nucleotide Substitutions on Ung2, Msh2, and PCNA-Ub During Somatic Hypermutation. *J Exp Med* (2009) 206:2603–11. doi: 10.1084/jem.20091707
  25. Langerak P, Nygren AO, Krijger PH, van den Berk PC, Jacobs H. A/T Mutagenesis in Hypermutated Immunoglobulin Genes Strongly Depends on PCNAK164 Modification. *J Exp Med* (2007) 204:1989–98. doi: 10.1084/jem.20070902
  26. Zeng X, Winter DB, Kasmer C, Kraemer KH, Lehmann AR, Gearhart PJ. DNA Polymerase Eta Is an A-T Mutator in Somatic Hypermutation of Immunoglobulin Variable Genes. *Nat Immunol* (2001) 2:537–41. doi: 10.1038/88740
  27. Kunkel TA, Erie DA. Eukaryotic Mismatch Repair in Relation to DNA Replication. *Annu Rev Genet* (2015) 49:291–313. doi: 10.1146/annurev-genet-112414-054722
  28. Bardwell PD, Woo CJ, Wei K, Li Z, Martin A, Sack SZ, et al. Altered Somatic Hypermutation and Reduced Class-Switch Recombination in Exonuclease 1-Mutant Mice. *Nat Immunol* (2004) 5:224–9. doi: 10.1038/ni1031
  29. Pena-Diaz J, Bregenhorn S, Ghodgaonkar M, Follonier C, Artola-Boran M, Castor D, et al. Noncanonical Mismatch Repair as a Source of Genomic Instability in Human Cells. *Mol Cell* (2012) 47:669–80. doi: 10.1016/j.molcel.2012.07.006
  30. Rogozin IB, Pavlov YI, Bebenek K, Matsuda T, Kunkel TA. Somatic Mutation Hotspots Correlate With DNA Polymerase Eta Error Spectrum. *Nat Immunol* (2001) 2:530–6. doi: 10.1038/88732
  31. Wilson TM, Vaisman A, Martomo SA, Sullivan P, Lan L, Hanaoka F, et al. MSH2-MSH6 Stimulates DNA Polymerase Eta, Suggesting a Role for A:T Mutations in Antibody Genes. *J Exp Med* (2005) 201:637–45. doi: 10.1084/jem.20042066
  32. Krijger PH, van den Berk PC, Wit N, Langerak P, Jansen JG, Reynaud CA, et al. PCNA Ubiquitination-Independent Activation of Polymerase Eta During Somatic Hypermutation and DNA Damage Tolerance. *DNA Repair (Amst)* (2011) 10:1051–9. doi: 10.1016/j.dnarep.2011.08.005
  33. Martomo SA, Yang WW, Wersto RP, Ohkumo T, Kondo Y, Yokoi M, et al. Different Mutation Signatures in DNA Polymerase Eta- and MSH6-Deficient Mice Suggest Separate Roles in Antibody Diversification. *Proc Natl Acad Sci USA* (2005) 102:8656–61. doi: 10.1073/pnas.0501852102
  34. Sale JE, Lehmann AR, Woodgate R. Y-Family DNA Polymerases and Their Role in Tolerance of Cellular DNA Damage. *Nat Rev Mol Cell Biol* (2012) 13:141–52. doi: 10.1038/nrm3289
  35. Maul RW, MacCarthy T, Frank EG, Donigan KA, McLenigan MP, Yang W, et al. DNA Polymerase Iota Functions in the Generation of Tandem Mutations During Somatic Hypermutation of Antibody Genes. *J Exp Med* (2016) 213:1675–83. doi: 10.1084/jem.20151227
  36. Lee SS, Tranchina D, Ohta Y, Flajnik MF, Hsu E. Hypermutation in Shark Immunoglobulin Light Chain Genes Results in Contiguous Substitutions. *Immunity* (2002) 16:571–82. doi: 10.1016/s1074-7613(02)00300-x
  37. Malecek K, Brandman J, Brodsky JE, Ohta Y, Flajnik MF, Hsu E. Somatic Hypermutation and Junctional Diversification at Ig Heavy Chain Loci in the Nurse Shark. *J Immunol* (2005) 175:8105–15. doi: 10.4049/jimmunol.175.12.8105
  38. Chen JM, Cooper DN, Ferec C. A New and More Accurate Estimate of the Rate of Concurrent Tandem-Base Substitution Mutations in the Human Germline: Approximately 0.4% of the Single-Nucleotide Substitution Mutation Rate. *Hum Mutat* (2014) 35:392–4. doi: 10.1002/humu.22501
  39. Francioli LC, Polak PP, Koren A, Menelaou A, Chun S, Renkens I, et al. Genome-Wide Patterns and Properties of *De Novo* Mutations in Humans. *Nat Genet* (2015) 47:822–6. doi: 10.1038/ng.3292
  40. Harris K, Nielsen R. Error-Prone Polymerase Activity Causes Multinucleotide Mutations in Humans. *Genome Res* (2014) 24:1445–54. doi: 10.1101/gr.170696.113
  41. Schrider DR, Hourmouzdi JN, Hahn MW. Pervasive Multinucleotide Mutational Events in Eukaryotes. *Curr Biol* (2011) 21:1051–4. doi: 10.1016/j.cub.2011.05.013
  42. Saribasak H, Maul RW, Cao Z, Yang WW, Schenten D, Kracker S, et al. DNA Polymerase Zeta Generates Tandem Mutations in Immunoglobulin Variable Regions. *J Exp Med* (2012) 209:1075–81. doi: 10.1084/jem.20112234
  43. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Ng AW, Boot A, et al. The Repertoire of Mutational Signatures in Human Cancer. *Nature* (2020) 578(7793):94–101. doi: 10.1101/322859
  44. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The Repertoire of Mutational Signatures in Human Cancer. *Nature* (2020) 578:94–101. doi: 10.1038/s41586-020-1943-3
  45. Koning MT, Kielbasa SM, Boersma V, Buermans HPJ, van der Zeeuw SAJ, van Bergen CAM, et al. ARTISAN PCR: Rapid Identification of Full-Length Immunoglobulin Rearrangements Without Primer Binding Bias. *Br J Haematol* (2017) 178:983–6. doi: 10.1111/bjh.14180
  46. Koning MT, Trollmann IJM, van Bergen CAM, Alvarez Saravia D, Navarrete MA, Kielbasa SM, et al. Peripheral IgE Repertoires of Healthy Donors Carry Moderate Mutation Loads and Do Not Overlap With Other Isotypes. *Front Immunol* (2019) 10:1543. doi: 10.3389/fimmu.2019.01543
  47. Alamyar E, Duroux P, Lefranc MP, Giudicelli V. IMGT((R)) Tools for the Nucleotide Analysis of Immunoglobulin (IG) and T Cell Receptor (TR) V-(D)-J Repertoires, Polymorphisms, and IG Mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS. *Methods Mol Biol* (2012) 882:569–604. doi: 10.1007/978-1-61779-842-9\_32
  48. IJspeert H, van Schouwenburg PA, Pico-Knijnenburg I, Loeffen J, Brugieres L, Driessen GJ, et al. Repertoire Sequencing of B Cells Elucidates the Role of UNG and Mismatch Repair Proteins in Somatic Hypermutation in Humans. *Front Immunol* (2019) 10:1913. doi: 10.3389/fimmu.2019.01913
  49. IJspeert H, van Schouwenburg PA, van Zessen D, Pico-Knijnenburg I, Driessen GJ, Stubbs AP, et al. Evaluation of the Antigen-Experienced B-Cell Receptor Repertoire in Healthy Children and Adults. *Front Immunol* (2016) 7:410. doi: 10.3389/fimmu.2016.00410
  50. Driessen GJ, IJspeert H, Weemaes CM, Haraldsson A, Trip M, Warris A, et al. Antibody Deficiency in Patients With Ataxia Telangiectasia Is Caused by Disturbed B- and T-Cell Homeostasis and Reduced Immune Repertoire Diversity. *J Allergy Clin Immunol* (2013) 131:1367–75.e9. doi: 10.1016/j.jaci.2013.01.053
  51. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing (2013).
  52. Jacobs H, Fukita Y, van der Horst GT, de Boer J, Weeda G, Essers J, et al. Hypermutation of Immunoglobulin Genes in Memory B Cells of DNA Repair-Deficient Mice. *J Exp Med* (1998) 187:1735–43. doi: 10.1084/jem.187.11.1735
  53. Sohail A, Klapacz J, Samaranyake M, Ullah A, Bhagwat AS. Human Activation-Induced Cytidine Deaminase Causes Transcription-Dependent, Strand-Biased C to U Deaminations. *Nucleic Acids Res* (2003) 31:2990–4. doi: 10.1093/nar/gkg464
  54. Kasar S, Kim J, Improgo R, Tiao G, Polak P, Haradhvala N, et al. Whole-Genome Sequencing Reveals Activation-Induced Cytidine Deaminase Signatures During Indolent Chronic Lymphocytic Leukaemia Evolution. *Nat Commun* (2015) 6:8866. doi: 10.1038/ncomms9866
  55. Maul RW, Gearhart PJ. Refining the Neuberger Model: Uracil Processing by Activated B Cells. *Eur J Immunol* (2014) 44:1913–6. doi: 10.1002/eji.201444813
  56. Tesch VK, IJ H, Raicht A, Rueda D, Dominguez-Pinilla N, Allende LM, et al. No Overt Clinical Immunodeficiency Despite Immune Biological Abnormalities in Patients With Constitutional Mismatch Repair Deficiency. *Front Immunol* (2018) 9:1506. doi: 10.3389/fimmu.2018.01506
  57. Wimmer K, Kratz CP, Vasen HF, Caron O, Colas C, Entz-Werle N, et al. Diagnostic Criteria for Constitutional Mismatch Repair Deficiency Syndrome: Suggestions of the European Consortium 'Care for CMMRD' (C4CMMRD). *J Med Genet* (2014) 51:355–65. doi: 10.1136/jmedgenet-2014-102284
  58. Perez-Duran P, Belver L, de Yébenes VG, Delgado P, Pisano DG, Ramiro AR. UNG Shapes the Specificity of AID-Induced Somatic Hypermutation. *J Exp Med* (2012) 209:1379–89. doi: 10.1084/jem.20112253
  59. Bahjat M, Stratigopoulou M, Pilzecker B, van Dam TP, Mobach S, Bende RJ, et al. DNA Polymerase  $\beta$  Prevents AID-Instigated Mutagenic Non-Canonical Mismatch DNA Repair. *bioRxiv* (2020). doi: 10.1101/2020.01.30.926964
  60. Au RY, Ng KS, Chi LM, Lam SL. Effect of an Abasic Site on Strand Slippage in DNA Primer-Templates. *J Phys Chem B* (2012) 116:14781–7. doi: 10.1021/jp308759k

61. Zharkov DO, Mechetin GV, Nevinsky GA. Uracil-DNA Glycosylase: Structural, Thermodynamic and Kinetic Aspects of Lesion Search and Recognition. *Mutat Res* (2010) 685:11–20. doi: 10.1016/j.mrfmmm.2009.10.017
62. Corbella M, Voityuk AA, Curutchet C. How Abasic Sites Impact Hole Transfer Dynamics in GC-Rich DNA Sequences. *Phys Chem Chem Phys* (2018) 20:23123–31. doi: 10.1039/C8CP03572E
63. Yin Y, Yang L, Zheng G, Gu C, Yi C, He C, et al. Dynamics of Spontaneous Flipping of a Mismatched Base in DNA Duplex. *Proc Natl Acad Sci USA* (2014) 111:8043–8. doi: 10.1073/pnas.1400667111
64. Cao C, Jiang YL, Stivers JT, Song F. Dynamic Opening of DNA During the Enzymatic Search for a Damaged Base. *Nat Struct Mol Biol* (2004) 11:1230–6. doi: 10.1038/nsmb864
65. Carr CE, Khutsishvili I, Gold B, Marky LA. Thermodynamic Stability of DNA Duplexes Comprising the Simplest T → dU Substitutions. *Biochemistry* (2018) 57:5666–71. doi: 10.1021/acs.biochem.8b00676
66. Parker JB, Bianchet MA, Krosky DJ, Friedman JI, Amzel LM, Stivers JT. Enzymatic Capture of an Extrahelical Thymine in the Search for Uracil in DNA. *Nature* (2007) 449:433–7. doi: 10.1038/nature06131
67. Fuxreiter M, Luo N, Jedlovsky P, Simon I, Osman R. Role of Base Flipping in Specific Recognition of Damaged DNA by Repair Enzymes. *J Mol Biol* (2002) 323:823–34. doi: 10.1016/S0022-2836(02)00999-3
68. Imhof P, Zahran M. The Effect of a G:T Mismatch on the Dynamics of DNA. *PLoS One* (2013) 8:e53305. doi: 10.1371/journal.pone.0053305
69. Varnai P, Canalia M, Leroy JL. Opening Mechanism of G:T/U Pairs in DNA and RNA Duplexes: A Combined Study of Imino Proton Exchange and Molecular Dynamics Simulation. *J Am Chem Soc* (2004) 126:14659–67. doi: 10.1021/ja0470721
70. Chen JM, Ferec C, Cooper DN. Patterns and Mutational Signatures of Tandem Base Substitutions Causing Human Inherited Disease. *Hum Mutat* (2013) 34:1119–30. doi: 10.1002/humu.22341
71. Stone JE, Lujan SA, Kunkel TA, Kunkel TA. DNA Polymerase Zeta Generates Clustered Mutations During Bypass of Endogenous DNA Lesions in *Saccharomyces cerevisiae*. *Environ Mol Mutagen* (2012) 53:777–86. doi: 10.1002/em.21728
72. Faili A, Aoufouchi S, Flatter E, Gueranger Q, Reynaud CA, Weill JC. Induction of Somatic Hypermutation in Immunoglobulin Genes Is Dependent on DNA Polymerase Iota. *Nature* (2002) 419:944–7. doi: 10.1038/nature01117
73. Weill JC, Reynaud CA. DNA Polymerases in Adaptive Immunity. *Nat Rev Immunol* (2008) 8:302–12. doi: 10.1038/nri2281
74. Biertumpfel C, Zhao Y, Kondo Y, Ramon-Maiques S, Gregory M, Lee JY, et al. Structure and Mechanism of Human DNA Polymerase Eta. *Nature* (2010) 465:1044–8. doi: 10.1038/nature09196
75. Matsuda T, Bebenek K, Masutani C, Rogozin IB, Hanaoka F, Kunkel TA. Error Rate and Specificity of Human and Murine DNA Polymerase Eta. *J Mol Biol* (2001) 312:335–46. doi: 10.1006/jmbi.2001.4937
76. Davies EG, Thrasher AJ. Update on the Hyper Immunoglobulin M Syndromes. *Br J Haematol* (2010) 149:167–80. doi: 10.1111/j.1365-2141.2010.08077.x
77. Imai K, Slupphaug G, Lee WI, Revy P, Nonoyama S, Catalan N, et al. Human Uracil-DNA Glycosylase Deficiency Associated With Profoundly Impaired Immunoglobulin Class-Switch Recombination. *Nat Immunol* (2003) 4:1023–8. doi: 10.1038/ni974

**Conflict of Interest:** The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

**Publisher's Note:** All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Copyright © 2022 Sepúlveda-Yáñez, Alvarez Saravia, Pilzecker, van Schouwenburg, van den Burg, Veelken, Navarrete, Jacobs and Koning. This is an open-access article distributed under the terms of the Creative Commons Attribution License (CC BY). The use, distribution or reproduction in other forums is permitted, provided the original author(s) and the copyright owner(s) are credited and that the original publication in this journal is cited, in accordance with accepted academic practice. No use, distribution or reproduction is permitted which does not comply with these terms.