Published in final edited form as:

DNA Repair (Amst). 2009 January 1; 8(1): 137–141. doi:10.1016/j.dnarep.2008.09.006.

SHMTool: a webserver for comparative analysis of somatic hypermutation datasets

Thomas MacCarthy¹, Sergio Roa², Matthew D. Scharff², and Aviv Bergman^{1,*}

¹Dept of Systems and Computational Biology, Albert Einstein College of Medicine, 1300 Morris Park Ave, Bronx, NY 10461, USA

²Dept of Cell Biology, Albert Einstein College of Medicine, 1300 Morris Park Ave, Bronx, NY 10461, USA

Abstract

The somatic hypermutation (SHM) of Immunoglobulin variable (V) regions is a key process in the generation of antibody diversity. The growing number of datasets of point mutations that occur during SHM in mice and humans often include comparisons between wild-type and individuals or strains genetically defective in the repair mechanisms that contribute to SHM. However, it has been difficult to compare the results of different studies because the analyses have not been standardized for criteria such as correction for base composition and the inclusion of unique mutations. If many mutations are involved, the analysis can also be time consuming. To overcome these problems and facilitate a standardized analysis and display of similar data, we present a webserver (SHMTool) for comparing SHM datasets, available at <http://scb.aecom.yu.edu/shmtool>.

Keywords

somatic hypermutation; mutation analysis; standardization

Somatic hypermutation (SHM) is a key process in the generation of antibody diversity that normally operates in antibody-forming B cells by introducing point mutations into the variable regions of immunoglobulin (Ig) heavy and light chain genes (reviewed in [1]). SHM is initiated when the highly mutagenic enzyme activation-induced deaminase (AID) [2] generates C→U mutations by deaminating cytosines preferentially at WRC hot spots motifs (where W=A/T, R=G/A and C is the mutated base) [3,4]. The uracils may be carried forward as mutations (C→T transitions) if bypassed during DNA replication. Alternatively, mutations may be introduced via UNG-mediated base excision repair or MSH2/MSH6-mediated mismatch repair pathways (recruited to repair the U:G mismatch), leading to C transversions (C→G, C→A) and mutations at neighboring A and T bases [5]. Previous research has made considerable progress in the characterization of the quantitative and qualitative features of SHM (reviewed in [5]), typically via analysis of mutated Ig locus sequences. At the same time, technological advances have facilitated the production of increasing amounts of sequence data. The criteria used to analyze and interpret this data have been heterogeneous [6] leading to situations in which reinterpretation of previously published data conflicts with the original results [7,8], as

*To whom correspondence should be addressed: tel +1(718)678-1063, fax +1(718)678-1018, email aviv@aecom.yu.edu.

Conflict of Interest statement The authors declare that there are no conflicts of interest

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

discussed in the example below. Analyses of different genomic regions with distinct base pair composition or datasets obtained from different sources (e.g. spleen or Peyer's patches) may be subject to similar problems. For example, a recent controversy over the role of DNA polymerase θ in SHM has led to a call for a standardized analysis method [9].

In general terms, most experiments investigating SHM compare sequence sets (case vs. control) with respect to their mutation profile. For example, Rada et al. [7] analyzed double knockout mice deficient for UNG (base excision repair pathway) and MSH2 (mismatch repair pathway), comparing these to wild-type controls. They found mutations at A:T sites were almost entirely ablated, leaving mostly C→T (and complementary G→A) mutations. Presumably these mutations were primarily caused by replication bypass, thus reflecting the original pattern of AID activity without the complicating subsequent base excision and mismatch repair. The authors also concluded there was no strand bias (mutability differences between transcribed and non-transcribed strands), given the similarity in the number of mutations at C-sites compared to G-sites (of a total of 520 mutations, 238 accumulated at C-sites vs. 270 at G-sites) [5,7]. A subsequent study [8] repeated the analysis while correcting for base composition (there are 94 C-sites vs. 160 G-sites in the unmutated germline sequence) and found a statistically significant strand bias.

Another source of heterogeneity in analysis relates to how mutations are counted. Multiple sequences derived from a single independent source (e.g. a clonal lineage identifiable via a unique CDR3 region) will frequently contain the same mutation in more than one sequence (e.g. a G→T at codon 33 in region V186.2, producing an aminoacid replacement that is strongly selected during NP response [10]). It is usually impossible to determine whether the mutation occurred just once (with the different sequences representing different sub-lineages) or several times independently. Accordingly, in some studies, such mutations are reported once (as *unique* or “nonclonal” mutations), or as many times as observed (*non-unique*, or “total” mutations), or both ways.

Here we present SHMTool, a webserver developed to offer automated analysis of mutated SHM sequences. The process is outlined in Figure 1. SHMTool receives FASTA sequence files in two categories (CONTROL and CASE, e.g. wild-type and genetically modified) to be compared. One or more files can be uploaded for each category. Within the files, each of which may contain many sequences, identical mutations (i.e. same mutation, same site) will be considered *unique* and counted only once (analysis of *non-unique* counts, where every sequence is considered independent, is available separately). Separate files should be submitted for sequences originating from independent sources (e.g. different mice, different B cell clones from one mouse defined by CDR3 sequence or clones of tissue culture cells). A single consensus (germline) sequence must also be designated. The user may also specify a subregion (potentially non-contiguous subset of sites) S of the consensus sequence to be analyzed separately. The complementary subregion S' (all sites not in S) is also analyzed. The subregion feature will typically be used, for example, to analyze complementary determining regions that form the antigen binding sites separately from the framework regions that position the CDRs in the variable region, or to exclude known polymorphic sites that are not mutations and would lead to an overestimate of the mutation frequency. Statistics comparing S and S' are also generated.

Unique mutations are classified (by base pair from→to, e.g. C→T) and aggregated separately for CONTROL and CASE categories. For each from→to mutation class, a 2×2 contingency table F is generated. The first column of F simply consists of the number of CONTROL and CASE mutations respectively (e.g., if there are x wild-type and y transgenic mutations, then $F_{1,1}=x$ and $F_{2,1}=y$). For the second column, the “number of trials” (the test assumes a binomial distribution), L is assigned for both CONTROL (L_{CT}) and CASE (L_{CA}). We consider it correct

to set L to the theoretical maximum number of mutations, i.e. the number of (G, C, A or T) sites multiplied by the number of groups (files) in the category. Continuing with the example, if the consensus sequence contains N C-sites, and there are 3 wild-type and 4 transgenic groups respectively (uploaded as separate files of sequences), then $L_{CT}=3N$ and $L_{CA}=4N$, and $F_{1,2}=(3N)-x$ and $F_{2,2}=(4N)-y$. With the contingency table assigned, a χ^2 test [as implemented by the R function `prop.test` [11]] is applied.

For *non-unique* mutation counts, all Q sequences in each category are considered independently and $L=NQ$. The actual frequency of independent mutational events is expected to lie somewhere between the *unique* and *non-unique* mutation frequencies [7]. Because the sequencing reaction often leads to sequences of differing lengths, it may be tempting in these circumstances to count mutations up to the end of each sequence, especially if mutations are rare. However, to ensure mutation frequencies are calculated correctly, we require each mutated sequence to span the exact length of the consensus. Otherwise, we are likely to underestimate the *unique* mutation frequency since not all sites would be represented by Q sequences. This does mean that any sequences shorter than the consensus cannot be used. Note that preprocessing of sequences (e.g., alignment, vector removal) is not performed by SHMTool and may be required prior to upload.

SHM analyses also typically present mutation counts as percentages of the total number of mutations. Clearly though, this can be misleading (e.g. an observed percentage increase in C→T mutations may be caused by an absolute increase in C→T mutation frequency, or a decrease in frequency of other mutations). We therefore consider mutation frequency (number of mutations / L) to be a more useful summary statistic. However, for convenience SHMTool presents both percentages and mutation frequencies.

To maintain base composition correction when reporting consolidated mutation counts (e.g. mutations from G and C), the number of mutations and L , are both aggregated. For *unique* mutations from a single base (e.g., all mutations from C: C→G, C→A, or C→T), L is multiplied by 3, since 3 possible substitutions now occur at each site (however, for *non-unique* mutations, L remains unchanged because the 3 substitutions are mutually exclusive).

Separate counts and tests are performed for subregions, as described above, and for motif sites representing both hot-and cold-spots (WRC, GYW, WA, TW, SYC, GRS, ADK, MHT, AA, TT, RGYW, WRCY, DGYW, WRCH, where W=A/T, R=A/G, Y=C/T, S=G/C, D=G/A/T, K=G/T, M=C/A, H=C/A/T) [3-5,12,13].

The web tool produces further analyses and displays typically presented in SHM studies including counts of mutations per sequence, and spatial plots of mutation counts per site (both aggregated as well as filtered by base: G, C, A or T). To enable users to test the webserver, an example *in vivo* dataset from a previous study [14] is available for download from the website. This particular dataset contains 3 wild-type and 4 *Msh6*^{-/-} FASTA files containing mutations in a Jh2-Jh4 intron region. Using this dataset, the user will be able to confirm key results previously reported for MSH6-deficient mice [14,15], namely decreased overall mutation frequency, decreased mutations at A and T sites and an increase in transition mutations at G and C sites (note there are minor quantitative differences with the reported results since the more stringent preprocessing required for SHMTool was not used for the original study). The comparison is summarized in Table 1, which presents the analysis of datasets from reference (14) using SHMTool in the two left hand columns, the analysis that was originally published (14) in the second column from the right and the results from another report on the same MSH6-deficient mice (15) in the right hand column. The two right hand columns illustrate how different groups analyze their data using different criteria such as unique vs non-unique mutations. A comparison of the left hand with the right hand columns highlights some of the

more detailed data and statistics available through SHMTool, such as changes in the mutation rates at AID and Pol η hotspots [16,17], which were previously reported either qualitatively or not at all.

SHMTool extracts from each set of sequences the many different sorts of analysis that have been reported by different investigators in the past and displays it in a variety of ways so that investigators can select how best to present their results. It also provides the appropriate statistical analysis enabling the further investigation of complex issues such as mutation saturation and strand bias. For example, counts of mutations per sequence are presented as histograms, which can be used to evaluate differences in degree of mutation saturation. For strand bias, for example, if there is no significant difference (between CASE and CONTROL) in mutation rates at WRC hotspots, yet there is at GYW hotspots, then this suggests a strand-specific difference.

Mutation profile analysis, similar to that offered by SHMTool, is used widely throughout immunology and cancer research. For example, in investigations of DNA mismatch repair [18], and APOBEC3-induced DNA lesions [19]. In its current form, SHMTool should be useful to investigators in these fields. However, to further broaden the scope of the tool, in future versions we will add features such as area-specific motifs and analysis of insertions and deletions. The webserver software is implemented using Perl, C++ and R (source code available upon request from the authors).

Acknowledgements

TM and AB are supported in part by The Seaver Foundation Center for Bioinformatics at the Albert Einstein College of Medicine, and NIH grants 1-R01-AG028872, and 1-P01-AG027734. SR is supported by Postdoctoral Fellowship EX-2006-0732 from the Spanish Ministry of Education and Science. MDS is supported by RO1CA72649 and RO1CA102705 and by the Harry Eagle Chair provided by the National Women's Division of the Albert Einstein College of Medicine.

References

1. Peled JU, Kuang FL, Iglesias-Ussel MD, Roa S, Kalis SL, Goodman MF, Scharff MD. The Biochemistry of Somatic Hypermutation. *Annu Rev Immunol* 2008;26:481–511. [PubMed: 18304001]
2. Muramatsu M, Sankaranand VS, Anant S, Sugai M, Kinoshita K, Davidson NO, Honjo T. 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;274:18470–18476. [PubMed: 10373455]
3. Rogozin IB, Kolchanov NA. Somatic hypermutagenesis in immunoglobulin genes. II. Influence of neighbouring base sequences on mutagenesis. *Biochim Biophys Acta* 1992;1171:11–18. [PubMed: 1420357]
4. Pham P, Bransteitter R, Petruska J, Goodman MF. Processive AID-catalysed cytosine deamination on single-stranded DNA simulates somatic hypermutation. *Nature* 2003;424:103–107. [PubMed: 12819663]
5. Di Noia JM, Neuberger MS. Molecular mechanisms of antibody somatic hypermutation. *Annu Rev Biochem* 2007;76:1–22. [PubMed: 17328676]
6. Longerich S, Basu U, Alt F, Storb U. AID in somatic hypermutation and class switch recombination. *Curr Opin Immunol* 2006;18:164–174. [PubMed: 16464563]
7. 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;16:163–171. [PubMed: 15494304]
8. Xiao Z, Ray M, Jiang C, Clark AB, Rogozin IB, Diaz M. Known components of the immunoglobulin A:T mutational machinery are intact in Burkitt lymphoma cell lines with G:C bias. *Mol Immunol* 2007;44:2659–2666. [PubMed: 17240451]

9. Gearhart PJ. Response to “Mutation frequency vs. mutation patterns: A comparison of the results in spleen and Peyer’s patches”. *DNA Repair (Amst)* 2008;7:1411–1412.
10. Rajewsky K, Forster I, Cumano A. Evolutionary and somatic selection of the antibody repertoire in the mouse. *Science* 1987;238:1088–1094. [PubMed: 3317826]
11. Bates, D.; Chambers, J.; Dalgaard, P.; Gentleman, R.; Hornik, K.; Iacus, S.; Ihaka, R.; Leisch, F.; Lumley, T.; Maechhler, M.; Masarotto, G.; Murdoch, D.; Murrell, P.; Plummer, M.; Ripley, B.; Schwate, H.; Temple Lang, D.; Tierney, L. *The R Project for Statistical Computing (Web Site)*. 1996. <http://www.r-project.org>
12. Bhattacharya P, Grigera F, Rogozin IB, McCarty T, Morse HC 3rd, Kenter AL. Identification of murine B cell lines that undergo somatic hypermutation focused to A:T and G:C residues. *Eur J Immunol* 2008;38:227–239. [PubMed: 18081040]
13. 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;172:3382–3384. [PubMed: 15004135]
14. Li Z, Zhao C, Iglesias-Ussel MD, Polonskaya Z, Zhuang M, Yang G, Luo Z, Edelmann W, Scharff MD. The mismatch repair protein Msh6 influences the in vivo AID targeting to the Ig locus. *Immunity* 2006;24:393–403. [PubMed: 16618598]
15. Martomo SA, Yang WW, Gearhart PJ. A role for Msh6 but not Msh3 in somatic hypermutation and class switch recombination. *J Exp Med* 2004;200:61–68. [PubMed: 15238605]
16. Masuda K, Ouchida R, Hikida M, Kurosaki T, Yokoi M, Masutani C, Seki M, Wood RD, Hanaoka F, OW J. DNA polymerases eta and theta function in the same genetic pathway to generate mutations at A/T during somatic hypermutation of Ig genes. *J Biol Chem* 2007;282:17387–17394. [PubMed: 17449470]
17. Mayorov VI, Rogozin IB, Adkison LR, Gearhart PJ. DNA polymerase eta contributes to strand bias of mutations of A versus T in immunoglobulin genes. *J Immunol* 2005;174:7781–7786. [PubMed: 15944281]
18. Scherer SJ, Avdievich E, Edelmann W. Functional consequences of DNA mismatch repair missense mutations in murine models and their impact on cancer predisposition. *Biochem Soc Trans* 2005;33:689–693. [PubMed: 16042575]
19. Vartanian JP, Guetard D, Henry M, Wain-Hobson S. Evidence for editing of human papillomavirus DNA by APOBEC3 in benign and precancerous lesions. *Science* 2008;320:230–233. [PubMed: 18403710]

Figure 1. Outline of SHMTool process

The raw CONTROL and CASE datasets (far left) require user preprocessing as described in text before submission to SHMTTool. The webserver produces mutation counts and comparative statistics for both unique and non-unique mutations, as well as spatial plots. Optionally, a subregion S of the sequence can be analyzed separately and compared to its complementary subregion S'.

Table 1
Comparison of SHMTool results from *Msh6*^{-/-} and control mice with those previously reported.

General mutation profiles	SHMTool (Jh2-Jh4 region derived from ref. 14)				Ref 14 (Jh2-Jh4 region)		Ref 15 (Jh4 intron)	
	unique mutations		non-unique mutations		unique and non-unique mutations		unique and non-unique mutations	
	Control	Msh6 ^{-/-}	Control	Msh6 ^{-/-}	Control	Msh6 ^{-/-}	Control	Msh6 ^{-/-}
Decreased overall mutation frequency	(based on theoretical maximum)		(based on number of bp sequenced)		Non-unique mutations (based on number of bp sequenced)		Unique mutation (based on number of bp sequenced)	
	17.5 x10 ⁻³	↓ 11.9 x10 ⁻³ ***	9.8 x10 ⁻³	↓ 5x10 ⁻³ ***	10.2 x10 ⁻³	↓ 2.5 x10 ⁻³ ***	6.3 x10 ⁻³	↓ 0.8 x10 ⁻³ ***
	(corrected ¹)		(corrected)					
Decreased A:T mutations	18.2 x10 ⁻³	↓ 1.6 x10 ⁻³ ***	11.9 x10 ⁻³	↓ 0.6 x10 ⁻³ ***	N.R.	N.R.	N.R.	N.R.
	(relative ²)		(relative)		Unique mutations (relative)		Unique mutations (corrected and relative)	
	52.60%	↓ 7% #	61.90%	↓ 6.3% #	54%	↓ 9% ***	54%	↓ 15% ***
	(corrected)		(corrected)					
Increased G:C transitions	27.1 x10 ⁻³	↑ 46.6 x10 ⁻³ *	5 x10 ⁻³	6.3 x10 ⁻³	N.R.	N.R.	N.R.	N.R.
	(relative)		(relative)		Unique mutations (relative)		Reported qualitatively as “most of the G:C mutations were transitions”	
	53.3%	↑ 68.7% #	66.6%	66.6% #	55.5%	↑ 73.2% ***		
Motif mutation profiles								
³ Polη hotspot	W <u>A</u>	4.7 x10 ⁻²	↓ 0.2 x10 ⁻² ***	3 x10 ⁻²	↓ 1.8 x10 ⁻² ***	N.R.	N.R.	
	I <u>W</u>	1.4 x10 ⁻²	↓ 0.4 x10 ⁻² ***	1.2 x10 ⁻²	↓ 0.1 x10 ⁻² ***			
⁴ AID	hotspot W <u>R</u> <u>C</u>	4 x10 ⁻²	7.1 x10 ⁻²	1.9 x10 ⁻²	↑ 3.3 x10 ⁻² ***	N.R. (this work focuses in the comparison of Msh6 ^{-/-} with partially defective mutant Msh6 ^{TD/TD})	In Msh6 ^{-/-} mice, the WGCW subset of the canonical RGYW/WRCY motif was highly targeted, whereas other variants were largely avoided.	
	G <u>Y</u> <u>W</u>	4.5 x10 ⁻²	6.4 x10 ⁻²	2.1 x10 ⁻²	2.9 x10 ⁻²			
	coldspot S <u>Y</u> <u>C</u>	0.6 x10 ⁻²	0.9 x10 ⁻²	0.2 x10 ⁻²	0.3 x10 ⁻²			
	G <u>R</u> <u>S</u>	0.6 x10 ⁻²	1.1 x10 ⁻²	0.08 x10 ⁻²	↑ 0.4 x10 ⁻² ***			
Spatial distribution of mutations	N.A		Similar to mutation positioning reported in ref. 14		Non-unique mutations. Msh6 ^{-/-} mice showed a more focused mutation pattern		Non-unique mutations. Msh6 ^{-/-} mice showed focused peaks of mutation	

* As explained in the text, we consider mutation frequency corrected by base composition to be a more useful summary statistic than mutation frequency relative to total mutations.

Bold and italic indicate statistically significant difference compared to the control group. * denotes 0.01 < *p*-value < 0.05; *** denotes *p*-value < 0.001. Arrows indicate increase or decrease in the frequency of mutation. N.R., not reported. N.A., not applicable.

¹ Base composition correction.

² Relative to the total number of mutations.

³ The decrease in A:T mutations reported above correlates with a highly significant decrease in the mutation of Polη hotspots, especially in WA motifs, which are preferentially targeted by Polη on the non-transcribed strand [16,17]. Underlined sites of the motif are scored. W=A/T, R=A/G, Y=C/T.

⁴ The marginally significant increase in G:C transitions reported above is consistent with a tendency towards an increase in AID hotspot targeting, which results significant only when non-unique (total) mutations are scored.