Statistical Methods for Identifying Sequence Motifs Affecting Point Mutations

Yicheng Zhu[1], Teresa Neeman[2], Von Bing Yap[3], and Gavin Huttley[1]

[1]**Research School of Biology, The Australian National University, Canberra ACT 2601, Australia**
[2]**Statistical Consulting Unit, The Australian National University, Canberra ACT 2601, Australia**
[3]**Department of Statistics and Applied Probability, National University of Singapore, 117546, Singapore**

ABSTRACT

Mutation processes differ between types of point mutation, genomic locations, cells, and biological species. For some point mutations, specific neighbouring bases are known to be mechanistically influential. Beyond these cases, numerous questions remain unresolved including: what are the sequence motifs that affect point mutations? how large are the motifs? are they strand symmetric? and, do they vary between samples? We present new log-linear models that allow explicit examination of these questions along with sequence logo style visualisation to enable identifying specific motifs. We demonstrate the performance of these methods by analysing mutation processes in human germline and malignant melanoma. We recapitulate the known CpG effect and identify novel motifs, including a highly significant motif associated with A→G mutations. We show that major effects of neighbours on germline mutation lie within ±2 of the mutating base. Models are also presented for contrasting the entire mutation spectra (the distribution of the different point mutations). We show the spectra vary significantly between autosomes and X-chromosome, with a difference in T→C transition dominating. Analyses of malignant melanoma confirmed reported characteristic features of this cancer including statistically significant strand asymmetry and markedly different neighbouring influences. The methods we present are made freely available as a Python library `https://bitbucket.org/pycogent3/mutationmotif`.

Keywords: log-linear model, context dependent mutation, germline mutation, somatic mutation

INTRODUCTION

Understanding the contributions of mutation processes to genetic diversity has broad relevance to topics ranging from estimating genetic divergence (Huttley, 2004; Schluter, 2009; Harris, 2015) to the aetiology of disease (Peltomaki and Vasen, 1997; Ying and Huttley, 2011; Nik-Zainal et al., 2012; Alexandrov et al., 2013a). While mutations occur on many scales, from single nucleotide point mutations to substantial genomic rearrangements, we restrict our attention here to point mutation processes. A multitude of mechanisms have been characterised that cause DNA lesions (Cooke et al., 2003; Helleday et al., 2014). Similarly, an array of processes repairing DNA lesions have also been described (Helleday et al., 2014). From examination of sequence composition alone it is apparent that mechanisms of mutagenesis (lesion formation and subsequent failure of DNA repair) differ between genomic locations (Francioli et al., 2015), between cell types (Nishino et al., 1996) and between species (Karlin et al., 1998). In evaluating natural systems, where only the starting and ending sequence states may be known, establishing the mechanistic origins remains a challenge. In mammals, an informative exception is the case of C→T point mutations. In this instance, a 3' G strongly implies a mechanism of 5-methyl-cytosine (5mC) deamination. This is due to the binding affinity of DNA methylases for the CpG sequence motif (Vinson and Chatterjee, 2012) and the greatly elevated mutation rate of 5mC (Coulondre et al., 1978). As the CpG example illustrates, predicting the contribution of a specific mechanism requires knowledge of a characteristic mutation sequence signature. Motivated by this, we focus here on development of a statistical method and associated visualisation approach for revealing signature sequence motifs associated with point mutations. We refer to these as mutation motifs.

Considerable evidence indicates that the influence of neighbouring bases on point mutations is a general phenomenon. Early studies on inherited, and thus germline, mutations in humans supported the hypermutability of the CpG dinucleotide as the dominant origin of C→T mutations (Cooper, 1995). Subsequent work further suggested that the remaining 11 point mutations are also affected by neighbouring bases (Krawczak et al., 1998). From analyses of mutations in human disease genes, Krawczak et al. (1998) inferred the influence of neighbours are confined to the positions immediately flanking the mutated location. The work on human polymorphisms demonstrated these results applied more generally across the genome (Zhao and Boerwinkle, 2002). Recently, using trinucleotides where the mutated base is central, distinctive mutation signatures that discriminate human cancer types have been identified (Alexandrov et al., 2013a). These results demonstrate the influence of neighbouring bases generalises to somatic mutations. Early influential work on plant cpDNA completes the demonstration of the generality of neighbouring influences across the tree-of-life (Morton et al., 1997). While Krawczak et al. (1998) and Zhao and Boerwinkle (2002) identified the influence of neighbours is proportional to distance, the work of Alexandrov et al. (2013a) was focused on the immediate flanking bases.

The influence of neighbouring bases on mutagenesis can have multiple causes. The chemical properties of DNA alone can confer a neighbour influence on mutation susceptibility. Adjacent pyrimidines are vulnerable to a dimerisation in the presence of UV light (Brown, 2002, p 426) with TpT being most susceptible. As the influence of DNA methylase preference for CpG dinucleotides demonstrates, DNA binding properties of macromolecules are a further likely source of neighbouring base influences. With numerous DNA-protein binding interactions central to DNA repair processes, any affinity to specific sequence motifs of these molecules may result in those motifs being under-represented in mutated sequences.

Analysis techniques for estimation of neighbouring base influences on mutation draw on different approaches. Krawczak *et al* (Krawczak et al., 1998) quantified neighbouring base influence by contrasting observed base frequencies against an equiprobable frequency distribution via a Euclidean distance. Zhao and Boerwinkle (Zhao and Boerwinkle, 2002) used just the base frequencies per position except beyond $\pm$10bp where averages across position ranges were used. In both these approaches, the background sequence distribution is assumed to be random occurrence of bases. These approaches therefore potentially obscure the real signal by confounding it with the non-random occurrence of bases characteristic of DNA sequences.

The distinctive mutagenic biology of cancer has motivated development of methods to identify specific mutation signatures across all point mutations. The related methods of (Alexandrov et al., 2013b) and (Shiraishi et al., 2015) tackle the problem of resolving the signatures of different mutational processes. As these signatures can contain instances of the different point mutation directions, they are a composite of distinct underlying mutational processes operating across multiple types of point mutations. The different mutation signatures may, therefore, contain component(s) that are identical and are not well suited to examining the influence of neighbouring bases on single point mutation directions.

More recently, the influence of neighbouring bases has been examined by using a probability of polymorphism that was conditioned on the sequence context (Aggarwala and Voight, 2016). A 7-mer context was identified as accounting for a median of 81% of the variability in the probability of polymorphism across point mutations. This result indicated inclusion of higher order (three-way and greater) interactions accounted for as much as 50% of the model predictive power. However, k-mers exhibit a non-random distribution within the human genome (Karlin, 1998; Chor et al., 2009). Moreover, variation in sequence composition is correlated with variation in substitution rate (Hodgkinson and Eyre-Walker, 2011). These suggest that by averaging across all occurrences of the sequence context, the results of (Aggarwala and Voight, 2016) could reflect the relationship between genomic location and the probability of polymorphism rather than the mechanistic influence of neighbours on mutation.

Detection of functional sequence motifs is a related problem to which information theoretic techniques have been extensively applied. Mutual information (MI) per position in a sequence alignment is computed by subtracting the position's Shannon entropy from entropy of the uniform distribution (Shannon, 1948). Coupling of this metric with the sequence logo visualisation approach has led to its widespread application for discovery of functional motifs (Schneider and Stephens, 1990). The display used the MI statistic to define a stack of colour coded letters, representing the sequence states, with each letter's height scaled proportional to its contribution to the total MI (Schneider and Stephens, 1990). For this application it is conventional to assume an independent equifrequent reference distribution. As removing the constraint

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2236v3 | CC BY 4.0 Open Access | rec: 16 Nov 2016, publ:

of equal frequencies can lead to negative values of MI, which are not readily interpretable, MI is not appropriate for examination of most DNA sequences as the equifrequent property typically does not hold.

Many of the developed techniques are confounded by common properties of genome DNA sequences. The ordering of nucleotides in DNA sequences is not random (Karlin, 1998). For the genomes of many organisms, such as vertebrates, there is also considerable within genome variation in k-mer frequencies (Karlin, 1998; Chor et al., 2009). For instance, trinucleotide frequencies within a protein coding exon are not well explained by the product of their monomer frequencies. Moreover, trinucleotide frequencies can differ between protein coding and non-protein coding sequences due to the differing influence of natural selection. Thus, neighbour affect analyses on exons may exhibit greater error rates unless such confounding is accounted for. Most available methods also do not distinguish contributions from independent positions compared with joint contributions from multiple positions. For instance, are mutations affected by the sequence of bases present at two positions (Zhang and Mathews, 1995)? Log-linear models allow flexible parameterisations for hierarchical hypothesis testing of categorical data and have been previously applied to examination of neighbouring influences (Huttley et al., 2000). Their generality allows for controlling of potential confounding differences, such as differences in sample size and nucleotide composition. The support for comparing hypotheses in a hierarchical manner enables explicit examination of hypotheses such as strand symmetry and absence of higher-order effects, which have been assumed by some approaches (Aggarwala and Voight, 2016). Thus, they provide an objective basis for identifying parametrically succinct models.

In this study, we develop log-linear approaches for examination of mutation processes. Our work is distinguished from previous methods by conditioning on the mutation event, rather than the sequence context, and employs a control distribution that is matched for genomic location. We present hierarchical hypothesis tests for evaluating whether: (i) neighbouring bases associate with mutation direction, (ii) neighbouring base associations are equal between samples, and (iii) the spectrum of mutations (the relative abundance of the 12 point mutations) are equal between samples. A sequence logo inspired visualisation approach is also presented. We demonstrate application of the models by applying them to data previously reported to exhibit distinctive mutation processes; namely, germline mutations in different sequence classes (e.g. transcribed, untranscribed) and chromosome classes (e.g. autosome and sex-chromosome), and somatic mutations in cancer. Mutation events in both human germline and somatic tissues were inferred from single nucleotide genetic variants available in Ensembl. In addition to replicating the well known CpG effect, our results indicate that neighbourhood size can be quite large and, as we demonstrate for the A→G transition mutation, the influence of neighbours does not decay monotonically with distance. We further show that both independent and dependent position influences contribute to mutational process. Through formal testing of equivalence between samples, we demonstrate significant differences between sequence classes, chromosome classes and between melanoma and germline mutations. Software implementing all these methods, released under an open source license, is made available `https://bitbucket.org/pycogent3/mutationmotif`.

MATERIALS AND METHODS

Data sampling

We infer mutation events in humans from published genetic variant records. Germline mutations were inferred from single nucleotide polymorphic (SNP) sites. Somatic mutations were inferred from genetic variants identified in cancers. In both cases, the mutation direction, location and associated flanking sequence were sampled from Ensembl (Flicek et al., 2013) release 79 using PyCogent's Ensembl querying capabilities (Knight et al., 2007). The Ensembl variation database records whether a variant is classified as somatic. We sampled germline SNPs using that flag and required the Ensembl record indicate the SNP was validated, had an inferred ancestral allele and that its flanking sequence matched the reference genome. For each such filtered SNP, we recorded the alleles, ancestral allele, strand, sequence class (exonic, intronic or intergenic), genomic coordinates and 300bp of flanking sequence either side of the SNP location.

Sampling somatic genetic variants involved both the COSMIC (Forbes et al., 2015) and Ensembl databases. Complete mutant export data was obtained from COSMIC, which included variant identifiers and the primary pathology from which a variant had been reported. Flanking sequence was derived by obtaining the Ensembl records for the variant identifiers, ensuring the record was flagged as somatic and

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2236v3 | CC BY 4.0 Open Access | rec: 16 Nov 2016, publ:

then following the same procedure as for the germline variants. We restricted our attention to variants identified from malignant melanoma.

Determining base counts

For each mutation direction (e.g. C→T) we obtained base counts from paired mutated and reference base locations. Neighbour positions were indexed relative to the position of the chosen location. For a mutated base, the chosen location was the annotated site of the variant (Figure 1). With knowledge of the mutation direction, a location with the same starting base as that affected by the mutation was randomly sampled within 300bp of the annotated variant. (e.g. a random choice of a position with a C in the case of a C→T mutation), but excluded the variant location. This is the paired reference base. In each case, a 5bp long sequence centred on the chosen location was extracted and the bases observed per relative position were recorded. We refer to these as neighbourhoods. As the total number of possible neighbourhoods was 256, a single file was written with counts for each of the possible neighbourhoods for both the mutated and reference locations. This approach to identifying the reference distribution also confers a substantial computational advantage, both in terms of memory required and compute time.

C→T	T	G	A	G	C	C	G	G	G	C	A
	-5	-4	-3	-2	-1	0	1	2	3	4	5
Reference C	C	T	G	G	G	C	A	T	G	A	G
	-1	0	1	2	3	4	5	-5	-4	-3	-2

Figure 1. Sampling mutated and reference base neighbourhoods. The neighbourhood of a position at which a C→T mutation occurred is compared with the neighbourhood of a reference occurrence of C randomly selected from within ±300bp of the C→T mutation. (The example sequence is greatly shortened to simplify the figure.) The location of the C→T variant is the central position for the mutated base and is assigned the index 0. The C at position 4 was randomly chosen as the reference location and the sequence is shifted so it is centred on this position (see *Determining base counts* for fuller explanation).

Log-linear modelling of neighbour effects

We first demonstrate the general approach of applying log-linear models for understanding neighbour influences on mutation by focusing on the influence of a single neighbouring position. Expected counts are modelled using the Poisson distribution for all the log-linear models described in this work. We then consider the extension of comparing neighbour contributions between samples. Both of these analyses are concerned with the independent contribution of bases at a position to mutation status.

For a single position, we evaluate whether *base* and mutation *status* occur independently using a straightforward log-linear model. Under the most saturated log-linear model, the log of the expected frequency f_{is} for *base* i and mutation *status* s can be expressed as

$$\ln f_{is} = \lambda + \lambda_i^{base} + \lambda_s^{status} + \lambda_{is}^{base:status} \quad (1)$$

where λ represents the intercept (i.e., common to all counts), λ_i^{base}, the contribution to the frequency of being *base* i, λ_s^{status} the contribution to the frequency of being mutation *status* s, and the interaction between *base* and *status* $\lambda_{is}^{base:status}$. The latter expresses the degree of non-independence between *base* and mutation *status*. The number of levels for each factor are: *base*, 4 levels (A, C, G, T); and mutation *status*, 2 levels (mutated, M and reference, R). Because the total counts for M and R are identical by design, $\lambda^{status} = 0$ for all s. The fit of a log-linear model is measured as the deviance (D). We specify the null hypothesis that bases occur independent of mutation status by setting $\lambda^{base:status} = 0$ for all i, s. The alternate is the fully saturated model. The difference in D between the null and alternate, nested models, is taken as χ^2 with degrees of freedom equal to the difference in the number of free parameters. In this instance, the degrees of freedom is 3.

When comparing groups, e.g. autosome versus X-chromosome, we add another factor (λ^{group}) to the log-linear model (2). The fully parameterised version of this log-linear model requires addition of three interaction parameters: 2 two-way interactions and the three-way interaction parameter $\lambda^{base:status:group}$.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2236v3 | CC BY 4.0 Open Access | rec: 16 Nov 2016, publ:

This parameter represents the influence of group on the *base* : *status* interaction. We therefore evaluate the null hypothesis of no difference between samples by setting all $\lambda^{base:status:group} = 0$ and compare this against the fully saturated model. If the group factor has only 2 levels, then the degrees of freedom for the resulting D is 3.

$$\begin{aligned}\ln f_{isg} = \lambda &+ \lambda_i^{base} + \lambda_s^{status} + \lambda_g^{group} \\ &+ \lambda_{is}^{base:status} + \lambda_{ig}^{base:group} + \lambda_{sg}^{status:group} \\ &+ \lambda_{isg}^{base:status:group}\end{aligned} \tag{2}$$

We now extend this approach to consider the simultaneous influence on mutation status of bases at multiple positions. To illustrate, consider the two neighbours following the base C in Figure 1. There are sixteen possible dinucleotides at the $1,2$ positions. The goal of this model is to establish whether the dinucleotides at these two positions associate with mutation status of C after taking account of the independent contributions of these positions. In order to achieve this, our two-position interaction model extends the independent contribution model (1), adding factors for the additional position and then interaction terms between the parameters. The fully saturated two-position interaction model is

$$\begin{aligned}\ln f_{ijs} = \lambda &+ \lambda_i^{base_1} + \lambda_j^{base_2} + \lambda_s^{status} + \lambda_{is}^{base_1:status} \\ &+ \lambda_{js}^{base_2:status} + \lambda_{ij}^{base_1:base_2} + \lambda_{ijs}^{base_1:base_2:status}\end{aligned} \tag{3}$$

where λ^{base_1} and λ^{base_2} represent the base contributions at positions one and two. In addition to including factors for the independent contributions of the two positions on mutation status, the $\lambda^{base_1:base_2}$ accounts for non-independent occurrence of bases at the positions, a key property of DNA sequences. The null hypothesis of no interaction between dinucleotides and mutation status is specified by setting all $\lambda^{base_1:base_2:status} = 0$ and comparing this against the fully saturated model. The resulting D has 9 degrees of freedom. For a given mutation direction, we perform this analysis for all possible combinations of pairs of sites.

These approaches are further extended to consider interactions amongst three positions, amongst four positions and for comparison of these effects amongst groups.

Log-linear model of mutation spectra

For analysis of mutation spectra, we evaluate the null hypothesis that the distribution of mutations is the same between groups. The opportunity for a specific mutation direction is affected by the total occurrence of the starting base. This quantity can be difficult to ascertain, such as in cancers where there may be major genomic rearrangements (e.g. deletions) relative to a reference group. To avoid this uncertainty, we restrict the analysis to point mutations from a specific base, comparing the relative counts of each of the 3 possible mutations between groups. This is a test of independence between ending base and group.

For a specific base, the log of the expected frequency is defined as

$$\ln f_{dg} = \lambda + \lambda_d^{direction} + \lambda_g^{group} + \lambda_{dg}^{direction:group} \tag{4}$$

where the factor $\lambda^{direction}$ represents the counts of the 3 different point mutation directions, λ^{group} the counts in the different groups, and $\lambda^{direction:group}$ the interaction between these factors. We specify the null hypothesis of equivalent proportions between the groups by setting $\lambda^{direction:group} = 0$. For two groups, comparing against the fully saturated model, the D has 2 degrees of freedom.

Visualisation

Sequence logo's display motifs using the mutual information as the letter stack height, and the fraction contributed to the mutual information (MI) by individual bases is derived from their individual terms in the MI calculation. We adopt a similar approach here. Instead of using MI, we use relative entropy (RE). The log likelihood ratio D is converted to RE by dividing by twice the sample size. RE from a log-linear analysis specifies the letter stack height. We use the terms in the RE equation to determine the proportion

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2236v3 | CC BY 4.0 Open Access | rec: 16 Nov 2016, publ:

of the stack height attributable to a specific base. We differ from the conventional sequence logo approach by distinguishing between bases that are under or over represented in the mutated class, relative to the unmutated class. Under-represented bases are indicated by a 180° rotation.

Interpretation of the logo is straightforward. A higher RE value indicates that a position(s) has a greater influence on mutation. Support for concluding a stack height reflects a meaningful influence on mutation derives from the p-value, from the log-linear model, that the data arose under the null hypothesis. The magnitudes and orientations of letters further conveys meaning in that ordinary letter orientation is indicative of over representation in the mutated group while inverted orientation indicates under representation. We note here that we make a choice to use residuals from the mutated class for display. Using residuals from the unmutated class would generate an image with the opposite letter orientations.

For multi-position models (e.g. equation 3), the stack height is equal between the indicated positions. For the two-position model, the characters for the nucleotide pair at the two positions share the same proportion and orientation. For the more complicated analyses involving contrasting neighbour effects between groups, the reference category is the one provided first to the software.

Differences in mutation spectra are visualised using a grid with rows corresponding to the starting base and columns the base resulting from the mutation. Each row corresponds to a single log-linear test for equivalent distribution of the possible point mutations from the base indicated by the row label (see *Log-linear model of mutation spectra*). The RE for each row is computed from the deviance of the corresponding spectra test. Letter heights for each base are scaled proportional to the corresponding term in the RE equation. The sum of letter heights in a row is the total RE for that test. Bases over-represented in the reference group are oriented in the conventional manner while under-represented bases are rotated 180°. In the spectra analysis, the largest base in the grid is the dominant mutation product difference between the groups.

Availability of data and materials

`MutationMotif` is a Python 3.5 compatible library for performing the statistical analyses outlined in this work that is freely available under an open source license. The project homepage is at `https://bitbucket.org/pycogent3/mutationmotif` and the version employed for the reported work is available in Zenodo (DOI 10.5281/zenodo.166388). It draws on R (Ihaka and Gentleman, 1996) for log-linear modelling, via the glm function, using the rpy2 Python binding to R. Sequence logo's are drawn using custom Python code included in `MutationMotif`. Other dependencies include PyCogent (Knight et al., 2007), `pandas`, `numpy`, `matplotlib` and `scitrack`.

The scripts performing the data sampling and applying the analyses reported in this work are freely available under the GPL at `https://bitbucket.org/gavin.huttley/analysemutations` and the version employed for the reported work is available in Zenodo (DOI 10.5281/zenodo.166387). `AnalyseMutations` includes the counts data required by `MutationMotif` and the complete set of results contained in this work. These counts data were produced from data sampled from the Ensembl and COSMIC databases, as described in *Data sampling*. Because the data files from which the counts files were produced are so large, they are available separately in Zenodo (DOIs 10.5281/zenodo.53158 `https://zenodo.org/record/53158` and 10.5281/zenodo.53164 `https://zenodo.org/record/53164`) under the Creative Commons Attribution-Share Alike license. Data files are typically gzip compressed standard formats; tab delimited text files, fasta formatted sequence files, serialised data is stored as json or pickle (Python's native serialised format).

File S1 contains tables and figures from additional analyses.

RESULTS

Overview of notation and neighbour effect log-linear models

The notation X→Y refers to a point mutation from starting base X to ending base Y, X→Y* refers to a point mutation and its strand symmetric counterpart, e.g. C→T* is C→T or G→A. The sampled region around a mutated base is called a neighbourhood with neighbours being the individual positions within the neighbourhood. A mutation motif is a specific neighbourhood that is enriched in mutated sequences compared to the reference distribution.

The log-linear model of neighbour influence evaluates the null hypothesis that a neighbouring base(s) flanking a specific point mutation is the same as that flanking a random occurrence of the starting base.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2236v3 | CC BY 4.0 Open Access | rec: 16 Nov 2016, publ:

For instance, does the distribution of bases at sites flanking C→T mutations differ from that flanking all C's? As the frequency of bases varies between genomic locations (Bernardi, 2000; Karlin, 1998; Chor et al., 2009), matching of the mutated and reference locations reduces possible confounding. We achieve this matching by deriving a reference location proximal to each mutated location. The sampling process is shown in Figure 1. We sampled 300bp of flanking genomic sequence each side of a variant and within this segment chose, at random, another occurrence of the starting base affected by the mutation event. Unless stated otherwise, we limited our analysis of neighbouring influence to $\pm$2bp either side of the mutated position, resulting in 256 possible neighbourhoods. For any given mutation direction, counts of these different neighbourhoods are obtained from both the sample centred on the mutated base and the sample centred on a random occurrence of the starting base. These counts are used to construct the contingency tables for the log-linear analysis. This approach achieves the objectives of controlling for compositional variation across the genome and controlling for the non-random occurrence of bases. See *Determining base counts* for more detail on this procedure.

The log-linear models used to examine the effect of neighbours on point mutation include parameters that represent an interaction between neighbouring base(s) and mutation status (see *Log-linear modelling of neighbour effects*). The contribution of this parameter to model fit is measured as a Deviance which, along with the residual degrees-of-freedom, is used to calculate the corresponding p-value for the null hypothesis. We convert the Deviance to relative entropy (hereafter, RE) as this measures the information content of the data under the model in a manner that is robust to sample size, allowing comparisons among analyses.

As we are concerned with whether flanking positions individually or jointly affect mutation process we describe the influence of neighbouring bases as independent or dependent/joint effects respectively. The influence of a base at a single neighbouring position on a point mutation will be referred to as an "independent" effect. The case when bases at two or more neighbouring positions influence a point mutation will be referred to as a "dependent" interactive effect or the joint influence of multiple bases. The number of positions involved in a dependent effect is referenced as the "order" of the interaction. An independent effect, the influence of a single position on mutation, is a first order effect while the joint influence of two positions on mutation is a second order effect. Flanking locations are indexed relative to the mutated position. The immediate flanking 5' base is at position -1 while the immediate flanking 3' base is at position $+1$ (see Figure 1). A series of positions are indicated by the relative indices in parentheses e.g. $(-2,-1)$ are two positions 5' to the mutated base. We note here that in the case of a dependent effect the actual positions are not necessarily physically adjacent, e.g. $(-2,2)$.

Log-linear models recapitulate the CpG effect and reveal higher order effects

In the analyses we report below, we focus principally on analyses of intergenic autosomal data. We also sampled variants from introns and exons. We relegate all results from analysis of other genomic regions to supplementary material as the results are substantively the same as those from the intergenic sequence class.

We benchmarked our method by examining the influence of neighbouring bases on C→T point mutations in the autosomal intergenic sample. (As none of the strand symmetry tests were significant for the intergenic autosomal mutations, we limit our discussion to the "plus" strand directions only.) We expected the influence of methylation induced deamination at CpG to reveal a strong G effect at the +1 position (Cooper and Youssoufian, 1988). This prediction was confirmed in the results of the hypothesis test (Table S1) and visually in the mutation motif logo (Figure 2B). The analysis established that while all positions made highly significant independent contributions to mutation (all p-values were estimated as ≈ 0, Table S1) the magnitude of their influence was small compared to that at the +1 position and only one of these was evident in the mutation logo, that of A at the -1 position (Figure 2B). (Results from the equivalent analysis of autosomal exon data are shown in Figure S1.)

Specific combinations of bases at multiple positions also significantly affected C→T mutations. All higher order interactions were statistically significant (all p-values $< 10^{-22}$, Table S1). A feature of the second and third order joint effects was that bases physically adjacent to each other or to the mutated position had the strongest association: $(-2,-1)$, $(-1,+1)$, $(+1,+2)$ second order interactions (Table S1 and Figure 2C), and the $(-2,-1,+1)$ third order interaction (Figure 2D).

Despite the highly significant associations between combinations of positions and interactions, the independent position contributions dominated. All effect orders were significantly associated with

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2236v3 | CC BY 4.0 Open Access | rec: 16 Nov 2016, publ:

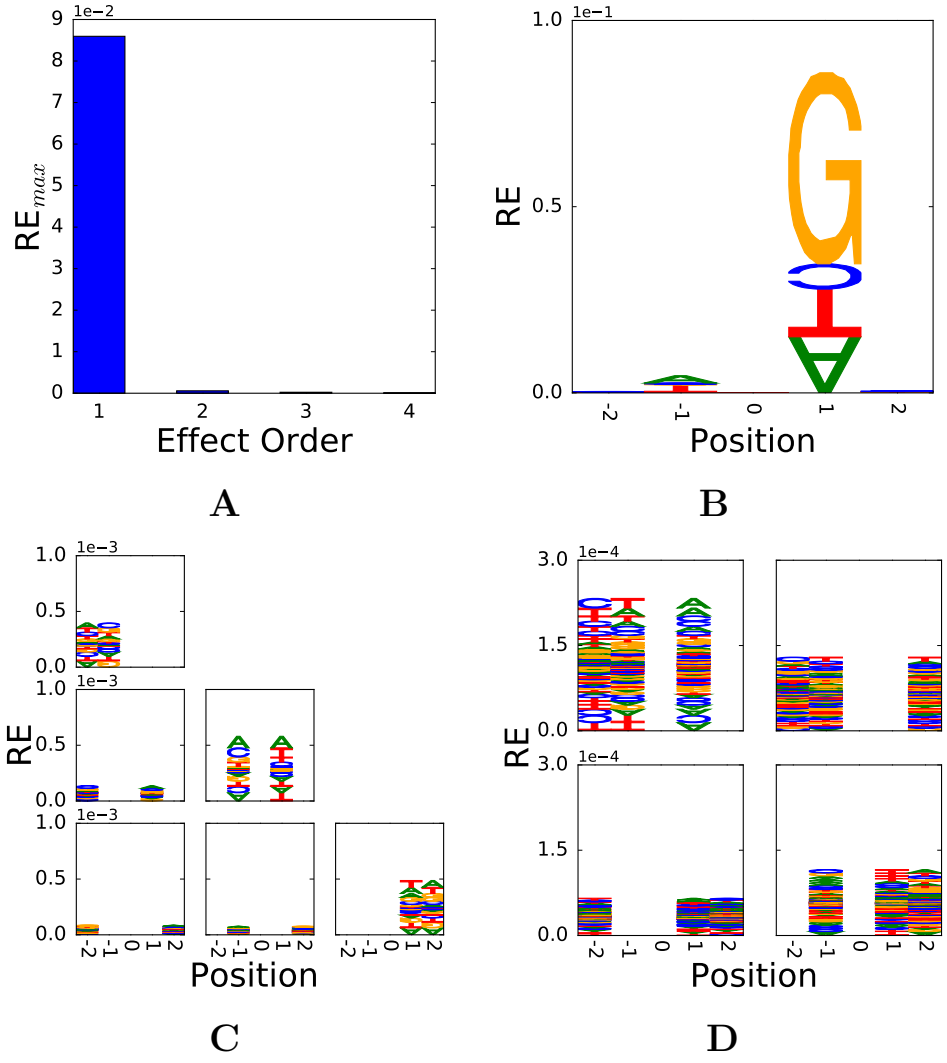

Figure 2. Neighbours influence C→T mutations. **A** First order effects are the dominant neighbour influence, RE$_{max}$ (y-axis) is the maximum RE from the possible evaluations for a motif length (x-axis), **B** Single position effects, **C** Two-way effects, and **D** Three-way effects. For b-d, the y-axis is RE and the x-axis is the position index relative to the mutated base. For details on interpreting the logo see *Visualisation*

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2236v3 | CC BY 4.0 Open Access | rec: 16 Nov 2016, publ:

mutation status even when using the sequential Holm-Šidák correction for 15 tests (Holm, 1979). These results reflect the enormous statistical power resulting from the large sample sizes, e.g. over 1 million C→T intergenic variants. Contrasting the magnitudes of these different effects by displaying the maximum RE value from each effect order (RE_{max}, Figure 2A) provide a useful indicator of their relative influence; $RE_{max}(1)$ is the maximum RE score for first position effects across all positions (e.g. +1 in this case), $RE_{max}(2)$ the maximum RE score from combinations of two positions, and so on for the higher orders. This display established that the 3'-G influence dominates all other neighbouring base effects on C→T mutation. Furthermore, contrasting these values between the point mutations (Table 1) affirms that neighbours have the strongest effect on C→T mutations (Figure S2).

Direction	$RE_{max}(1)$	Pos.(1)	$RE_{max}(2)$	Pos.(2)	$RE_{max}(3)$	Pos.(3)
A→C	0.0039	-1	0.0016	(+1, +2)	0.0012	(-2, -1, +1)
A→G	0.0188	+1	0.0030	(-2, -1)	0.0007	(-2, -1, +1)
A→T	0.0095	+1	0.0051	(-1, +1)	0.0023	(-1, +1, +2)
C→A	0.0091	+1	0.0044	(-1, +1)	0.0015	(-1, +1, +2)
C→G	0.0054	-2	0.0025	(+1, +2)	0.0010	(-1, +1, +2)
C→T	0.0860	+1	0.0006	(-1, +1)	0.0002	(-2, -1, +1)

Table 1. Summary of neighbour associations with plus strand mutations with an autosomal intergenic location. $RE_{max}(\#)$ is the maximum RE for order # and Pos.(#) the corresponding position(s). All point mutations had at least one significant test after correcting for 15 tests (see Table S1) using the Holm-Šidäk procedure.

A→G mutations are also strongly affected by neighbours

The A→G transition mutation exhibited the next strongest influence of neighbouring bases (Table 1). As for C→T, all effect orders were highly significant after correcting for 15 tests (all p-values $< 10^{-47}$, Table S2). All positions showed significant first order influences, but the $-2, -1, +1$ positions were particularly strong (Figure 3B). Two of these, $(-2, -1)$, also exhibited a prominent second order interaction (Figure 3C) while all three contributed the strongest third order interaction (Figure 3D). For A→G mutations, our analysis indicated that while first order effects dominated, higher order effects were important factors affecting this mutation direction (Figure 3A). Again, combinations of bases that were physically adjacent were most influential. (Results from the equivalent analysis of autosomal exon data are shown in Figure S3.)

Transversion mutations are affected by neighbours

All transversion mutations had significant neighbour influences but to a lesser extent than that evident for transition mutations (Table 1, Figure S2). The transversion mutations showed $RE_{max}(1)$ that were 20-fold less than for the C→T mutations. However, higher order effects were typically more pronounced for transversions than transitions. The A→T and C→A transversion mutations showed the greatest influence of neighbours at all levels. The dominant influences were immediately adjacent to the mutating base except for C→G, where position -2 had the strongest effect.

The size of the neighbourhood

Our analyses above indicated first order effects exerted the strongest influence on mutations. Accordingly, we limited our examination of neighbourhood size to first order effects and sampled intergenic autosomal variants with a flank size of $\pm$10bp for an analysis. After correcting for multiple tests, all 20 flanking positions were significant for all point mutations (Table S3). This suggests a neighbourhood size $\geq$10. The tendency for even very distant positions to be highly significant in this analysis likely reflects the enormous sample sizes employed for this analysis and does not necessarily reflect the magnitude of a positions influence. Therefore, for each mutation we estimated the most distant position with a RE that was $\geq$10% of $RE_{max}(1)$. For the transition mutations, the neighbourhood size was restricted to positions within $\pm$2bp (Figure S4) whilst for transversion mutations, the neighbourhood size was within $\pm$4bp (Table S3).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2236v3 | CC BY 4.0 Open Access | rec: 16 Nov 2016, publ:

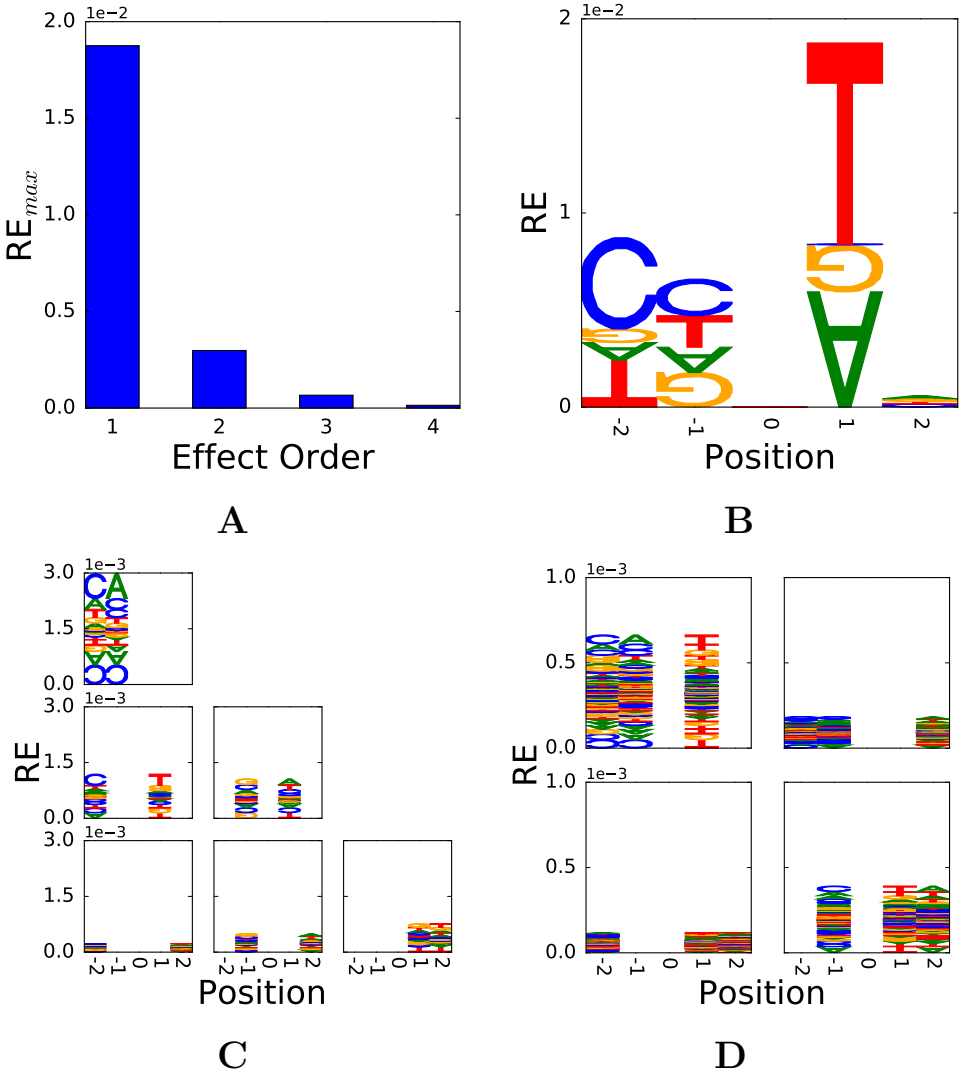

Figure 3. Neighbours influence A→G mutation in autosomal intergenic sequences. **A** First order effects are the dominant neighbour influence, **B** Single position effects, **C** Two-way effects, and **D** Three-way effects. For b-d, the y-axis is RE and the x-axis is the position index relative to the mutated base.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2236v3 | CC BY 4.0 Open Access | rec: 16 Nov 2016, publ:

Some germline point mutations exhibited different neighbouring effects between sequence classes

The operation of transcription coupled DNA repair processes suggested a possible difference in neighbour effect may exist between transcribed and untranscribed sequences. This predicts a difference in mutation profile between intergenic and intronic sequences. Our analysis of neighbour contributions to mutation established that for first order effects, every point mutation was significantly different between the sequence classes (Table S4). For second order effects, only the transition mutations showed significant differences. The biggest difference between the regions was for A→T*. While these effects were highly significant, their $RE_{max}(1)$ were ≈100 fold lower than the overall influence of neighbours on intergenic A→T.

Neighbouring effects differ between chromosome classes

Differences in germline biology between males and females predict distinct mutation profiles between sequences located on the autosomes and X-chromosome (Huttley et al., 2000). Our test of the hypothesis of no difference in flanking base effect between autosome and X-chromosome mutations in intergenic sequences was rejected for first order influences on several of the point mutations, after correcting for 15 tests using the Holm-Šidák procedure (Holm, 1979) (Table S5). Interestingly, A→G* and C→T* showed comparable differences in flanking base effect between the chromosome classes (Deviances ≈26.0 and ≈25.4 respectively). In all cases, the effect exists at the same position as that identified as $RE_{max}(1)$ in the intergenic analysis (Table 1). While the transition mutations were the most statistically significant, their RE lay within the range of the other point mutations (Table S5) indicating their significance reflects greater abundance and thus a greater rate.

Analysis of germline mutation spectra

Our log-linear model for analysis of mutation spectra compares counts of point mutations from the same starting base between groups. By considering only mutations from a single base between different locations, differences in the abundance of the starting base between groups are controlled for. This approach can be applied to groups representing different strands, different genomic regions or different biological materials (e.g. germline and somatic).

Our analysis of germline mutation spectra indicated point mutations were uniformly strand symmetric but different between sequence categories. No sequence category exhibited strand asymmetry in mutation spectra for autosomal data. Significant differences in autosomal mutation spectra were evident between intergenic and intronic regions. The major differences were for transversion mutations, specifically C→A and its strand complement (Table S6).

Significant differences between chromosome classes were evident (Figure 4 and Table S7). For the intergenic sequence class, A→G* transition mutations were in strong excess on autosomes compared with X-chromosome (Figure 4). Comparable results were evident for intronic sequences (Table S8).

Melanoma mutations exhibit strikingly different neighbour effects and spectra

Mutation processes in malignant melanoma are known to be distinctive and to include strand asymmetric mutation processes within genes (Pleasance et al., 2010). Our analysis confirm that the profile of point mutations in the malignant melanoma sample was strikingly different to the germline mutations (Tables S12, S13). The grid of all point mutations (Figure 5) demonstrates that neighbouring influences were most pronounced for C→T point mutations and much stronger influence of neighbouring bases on transversion mutations. The neighbour effects were also significantly strand asymmetric (Table S9), a distinctive characteristic for melanoma. Only substitutions affecting C were significantly different in spectra between strands with the C→T direction being over abundant on the + strand (Figure 6, Table S10).

DISCUSSION

While it has long been appreciated that sequence neighbourhoods affect point mutations, statistical methods for disentangling how neighbours contribute have been limited. Here we addressed this using a novel determination of the reference distribution and log-linear models. This methodological combination is robust to complexity in the genomic background of nucleotide composition. It further enables hierarchical hypothesis testing for establishing the significance and relative importance of neighbour effects. We illustrated utility of the models by applying them to analyses of mutations from samples reported to exhibit

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2236v3 | CC BY 4.0 Open Access | rec: 16 Nov 2016, publ:

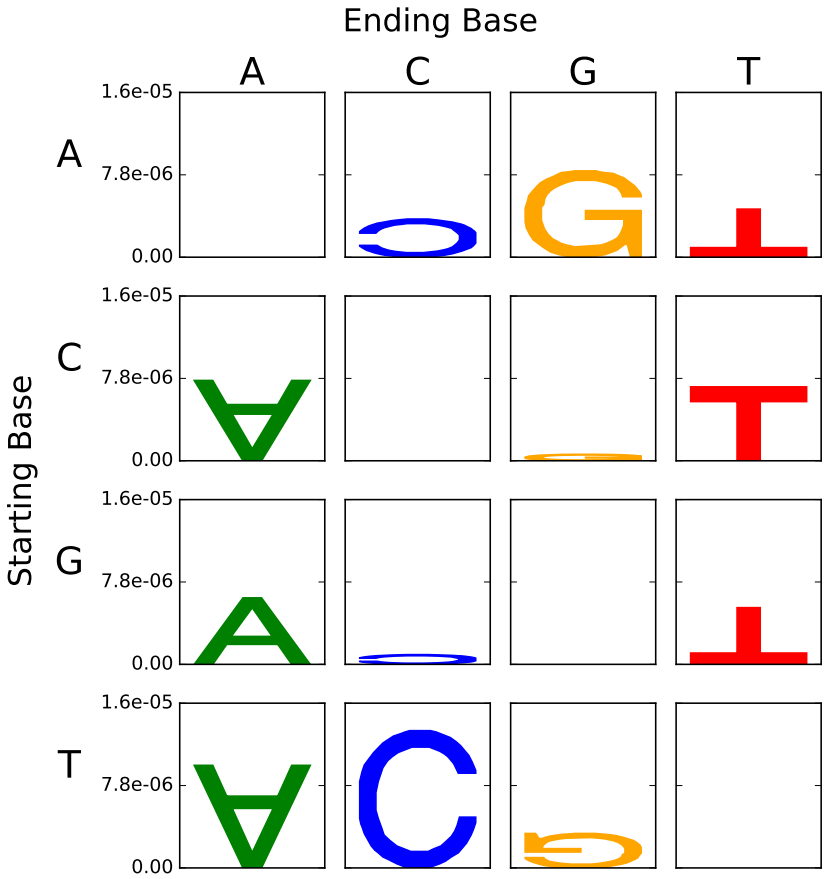

Figure 4. Significant differences in mutation spectra between autosomal and X-chromosomal intergenic sequence regions. Starting base, Ending Base correspond to X, Y respectively in X→Y. The y-axis is RE from the spectra hypothesis test and letter heights are as for the mutation motif logo. Letters in the normal orientation indicate an excess of that mutation direction in autosomal relative to the X-chromosomal mutations. Inverted letters indicate a deficit in autosomal relative to the X-chromosomal mutations.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2236v3 | CC BY 4.0 Open Access | rec: 16 Nov 2016, publ:

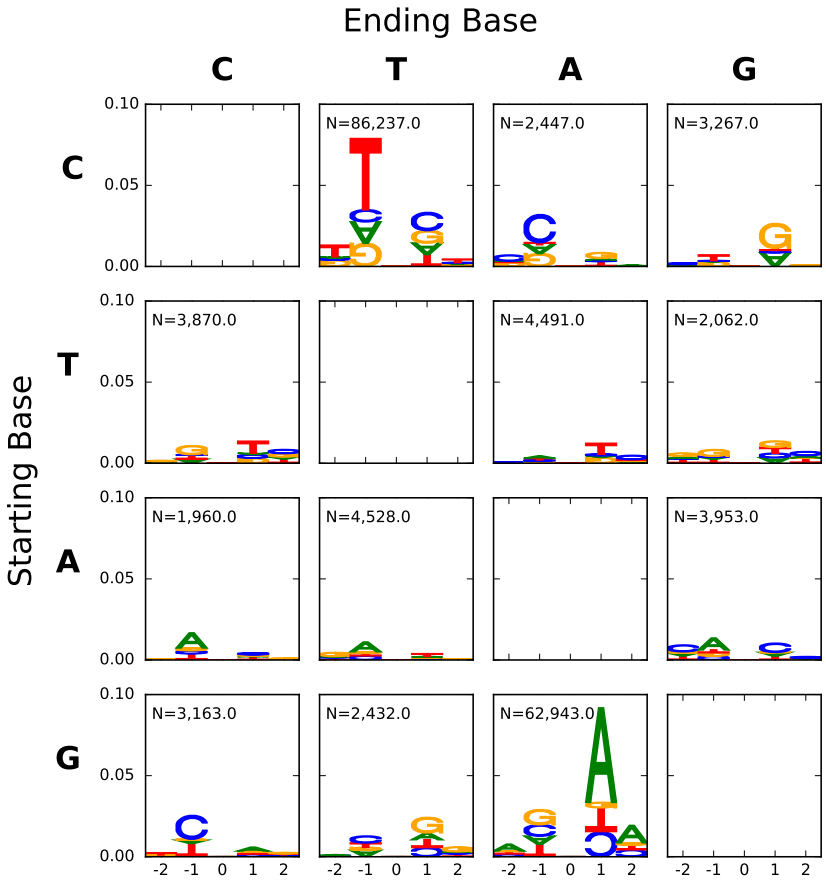

Figure 5. Panel of first order effects from all 12 point mutations from the malignant melanoma sample. Starting base, Ending Base correspond to X, Y respectively in X→Y. The y-axis is RE and the x-axis is the position index relative to the mutated base. N refers to the number of variants from which the logo was derived.

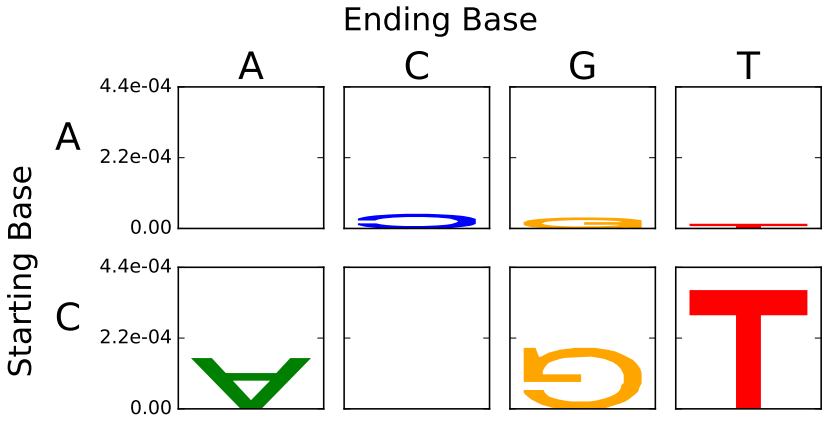

Figure 6. Strand asymmetry in malignant melanoma. Only mutations from C were statistically significant. Starting base, Ending Base correspond to X, Y respectively in X→Y. The y-axis is RE from the spectra hypothesis test and letter heights are as for the mutation motif logo. Letters in the normal orientation indicate an excess of that mutation direction on the + strand. Inverted letters indicate a deficit on the + strand.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2236v3 | CC BY 4.0 Open Access | rec: 16 Nov 2016, publ:

distinctive properties. Our analyses recapitulated well-known effects in terms of neighbour dependence and in terms of differences between genomic regions and somatic and germline, supporting the accuracy of the methods. The results revealed previously unreported neighbour effects that extends beyond immediate flanking positions. Analyses of mutation spectra complemented the neighbour analyses, confirming known features of point mutations in malignant melanoma and identifying novel differences in germline point mutation abundance between sex-chromosomes and autosomes.

The hypermutability of C→T in CpG dinucleotides is the exemplar of context dependent mutation and a gold standard that a method of analysis should correctly recover. We established that the conventional sequence logo analysis approach did not recapitulate the dominant influence of a 3'-G (Figure 7). As this method shares the assumption of equifrequent bases with that of (Krawczak et al., 1998), the failure suggests the Euclidean distance approach will also be flawed. In contrast, as shown in Table S1 and Figure 2, our analysis successfully recapitulated this known effect. The RE_{max} values (Figure 2B) further affirm C→T as most strongly affected by neighbouring bases.

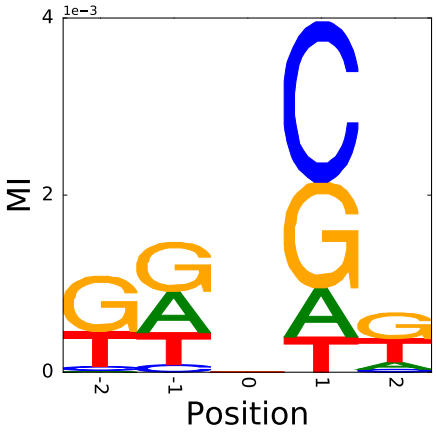

Figure 7. The CpG effect on C→T is not revealed by applying the conventional sequence logo method to autosomal intergenic mutations. MI is mutual information.

In order to sensibly interpret the results of our analyses we de-emphasise the importance of statistical significance and focus instead on effect magnitude. Due to the very large number of inferred mutations, our analyses possess very high power to detect small effects. This is illustrated by the very small p-values associated with, for example, third order effects for the C→T mutation (Table S1). Yet, the magnitude of these effects is relatively small in comparison with the first order effects (Figure 2A). Consequently, and in addition to considering whether effects are statistically significant according to standard criteria, we contrast RE statistics to establish relative importance.

Our analysis identified numerous novel properties of neighbouring sequence influence on point mutation in the germline. First, all mutations were significantly affected by neighbouring bases with transition mutations showing a larger influence of neighbours than transversions. Interestingly, as illustrated by the A→G* mutations, these influences did not decay monotonically with distance from the mutation (Figure 3B). This point mutation further illustrated that multiple neighbouring positions can influence mutation outcome. Comparing RE values to that for C→T indicates that the first order neighbour effects of other point mutations were ∼5−20 fold less, with those values corresponding to A→G and A→C mutations respectively (Table 1). Second, all mutations were significantly affected by higher order effects (interactions between adjacent bases). These were evident in a manner such that bases contiguous with each other and the mutated location showed the largest RE. This may reflect the importance of interactions amongst adjacent bases (base-stacking) in affecting DNA stability (Karlin and Burge, 1995; Yakovchuk et al., 2006). For all point mutations, the RE terms from first order effects were markedly stronger than those for higher order effects. These results were replicated in our analysis of intronic variants (Table S11).

The evidence for neighbouring influence on mutation raised the important question of how far these effects of flanking sequence extend? While there was strong statistical significance of positions as far as 10bp from the mutating base (Table S3), considering the relative magnitude of RE values indicated a

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2236v3 | CC BY 4.0 Open Access | rec: 16 Nov 2016, publ:

very rapid decay away from the mutated position. In particular, that the magnitude of the effect decayed below an order of magnitude within 2 bases for transition mutations. This trend is illustrated by the mutation motif logo displays (Figure S4). While transversion mutations exhibited a slower decay in effect magnitude, and hence a larger neighbourhood, these reflect the smaller $\text{RE}_{max}(1)$ of transversions which constitute a less stringent cut-off.

Our results regarding the importance of higher order interactions indicate that considering 5-mers accounts for the majority of model fit. The deviances from the first order effects of A→G* and C→T* transition mutations accounted for 81% and 98% of the total deviance respectively in the autosomal intergenic sample. Inclusion of second order effects increased both these to $> 96\%$ (Tables S1, S2). Across all point mutations in the autosomal intergenic sample, combining first and second order effects accounted for a median 91% of the total deviance of the 5-mer model. These differences are further illustrated by the motif [C/T]CAAT[C/G/T]N, reported as exhibiting an odds ratio of ~ 6000 for enrichment in mutated sequences (Aggarwala and Voight, 2016). Our results (Table S2 and Figure 3D) identified the CAAT core of this motif as highly significant. However, this is a third-order interaction and the RE for this specific combination of sites is 28 fold less than the strongest first order effect and accounts for only 1.5% the total deviance. We estimated odds ratio for the CAAT mutation motif as ~ 4.0 which is less than the ~ 5.7 odds ratio we estimated for the 7-mer of Aggarwala and Voight (2016). (We note here that our odds ratios are closer to what Aggarwala and Voight (2016) term "fold change".)

The profile of somatic mutations is expected to exhibit differences to germline mutations due to requisite defects in DNA repair systems. As reported (Nik-Zainal et al., 2012), such defects are characteristic of cancers. Of the characterised cancers, malignant melanoma exhibit the most distinctive mutation signatures. Included in the distinctiveness of malignant melanoma is a striking strand asymmetry (Pleasance et al., 2010). This putatively derives from UV light induced formation of pyrimidine dimers. In transcribed regions, nucleotide excision repair processes coupled to transcription-coupled repair mechanism, results in efficient repair of transcribed strand lesions. As a consequence, mutations are expected to accumulate on the non-transcribed strand. Evidence supporting this, with more C→T mutations on non-transcribed strand than on the transcribed strand, has been reported (Pleasance et al., 2010).

Our analysis demonstrated that point mutations in melanoma were dependent on neighbours in a manner strikingly different from that of germline processes discussed thus far (Figure 5 and Table 2). While C→T mutations were again the point mutation most affected by neighbouring bases, the motif was markedly different to that from the germline process with a 5'-T showing the greatest influence. This difference indicates that 5mC deamination plays a less prominent role in C→T. Since melanoma arises in part due to defect(s) in DNA repair the distinctive mutation motifs in melanoma indicate either a very effective masking of neighbour effects on lesion formation, or that the DNA repair mechanisms inactivated in melanoma are strongly affected by neighbours. Our melanoma analysis also strongly supported strand asymmetry of mutations, with the effect most pronounced for C→T.

Direction	$\text{RE}_{max}(1)$	**Pos.**(1)	$\text{RE}_{max}(2)$	**Pos.**(2)	$\text{RE}_{max}(3)$	**Pos.**(3)
A→C	0.0167	-1	0.0101	(-1, +1)	0.0078	(-2, +1, +2)
A→G	0.0135	-1	0.0118	(-1, +1)	0.0051	(-1, +1, +2)
A→T	0.0110	-1	0.0039	(-2, +1)	0.0033	(-2, -1, +1)
C→A	0.0319	-1	0.0102	(-1, +1)	-	-
C→G	0.0264	+1	0.0035	(-1, +1)	0.0041	(-2, -1, +1)
C→T	0.0788	-1	0.0130	(-1, +1)	0.0006	(-2, -1, +1)
G→A	0.0918	+1	0.0090	(-1, +1)	0.0009	(-1, +1, +2)
G→C	0.0254	-1	0.0028	(-2, +1)	0.0043	(-1, +1, +2)
G→T	0.0242	+1	0.0078	(+1, +2)	0.0052	(-1, +1, +2)
T→A	0.0123	+1	0.0042	(+1, +2)	0.0044	(-1, +1, +2)
T→C	0.0135	+1	0.0244	(-1, +1)	0.0057	(-1, +1, +2)
T→G	0.0137	+1	0.0118	(-1, +1)	0.0074	(-2, +1, +2)

Table 2. Summary of neighbour associations with mutations in malignant melanoma. $\text{RE}_{max}(\#)$ is the maximum RE for order # and Pos.(#) the corresponding position(s). All point mutations had at least one significant test after correcting for 15 tests (see Table S1) using the Holm-Šidák procedure. Non-significant results are indicated by '-'.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2236v3 | CC BY 4.0 Open Access | rec: 16 Nov 2016, publ:

A major asset to the log-linear modelling framework is the ease of extension to enable comparisons between samples. The utility of this is illustrated above in comparing somatic to germline processes. The appeal of this capability, however, is much broader as it further allows evaluation of the processes that contribute to within genome heterogeneity in sequence composition. We have illustrated this application here by considering genomic regions for which the incidence of mutation processes are known to differ (X-chromosome versus autosomes) or where DNA repair processes are known to differ (transcribed versus untranscribed regions).

The notion that there is a systematic tendency for mutations to originate in males has been known since Haldane (Haldane, 1935, 1946, 1948). The most popular hypothesis to account for male biased evolution is the mutation-through-DNA-replication hypothesis (Li et al., 2002; Webster et al., 2005). Other, non-replication based, differences in mutation between the sexes have also been proposed (Huttley et al., 2000). Included in these is evidence for elevated methylation of DNA in the male germline. This suggests the relative contribution of 5mC derived lesions will be greater on the autosomes compared to X-chromosome as the latter spends less time (on average) in males. Our analyses for differences in neighbour influences did lend support to existence of distinct 5mC affecting mutation processes operating between the X-chromosome and autosomes (Table S5), including a reduced magnitude of the +1 influence on the X-chromosome. However, this was not the strongest difference in neighbour effect between the chromosomal classes; A→G showed the strongest statistical significance while C→G showed the greatest RE. The spectra analyses further emphasised the importance of differences in A→G* point mutations (Figure 4). These results therefore indicate more extensive point mutation differences between these chromosome classes than previously appreciated and suggest a corresponding diversity in mutational processes between male and female germlines.

That differences in operation of DNA repair processes may affect mutation is predicted by the localised influence of transcription coupled DNA repair. This process is known to operate in a manner that is strand asymmetric. Differences in base parity – the frequency of A should equal that of T, G should equal C – support an effect of transcription on point mutation (Touchon et al., 2003). Significant differences in neighbour effects for all point mutations were evident between intergenic and intron regions. However, our analysis of strand symmetry for neighbour effects was not significant for intron sequences for any point mutation. This suggests a distinctive mutation profile arising from transcription, rather than the influence of transcription coupled DNA repair.

We have argued that the matched sampling of the reference distribution in our neighbour analysis is important. Briefly recapitulating that approach, the reference distribution is obtained by randomly selecting a paired reference base within $\pm$300bp of each observed mutation (Figure 1). An alternate to this strategy is to obtain the reference base by randomly selecting from the full genome sequence. For a given point mutation direction, only the reference counts can differ between the $\pm$300bp and genome reference approaches, i.e. the observed counts are identical. Consequently, the statistical inferences will likely differ when the k-mer distribution for a sequence class differs from that of the entire genome. An obvious case where this condition arises are protein coding exons. A neighbour analysis of exon sequences where the reference distribution was obtained from the full genome sequence showed significant differences to the $\pm$300bp one. The relative importance of each flanking position and/or the identity of bases at those positions differed for all of the point mutation directions (for a subset see Figure S5). These results, and its considerable computational advantages, support using the $\pm$300bp reference distribution.

As formulated, the neighbour analysis do not evaluate the relative abundance of mutations between samples. For this purpose, we introduce what we termed the mutation spectrum analysis. As the opportunity for mutation is affected by the frequency of the starting base, and base frequency differs between genomic locations, we perform spectrum analysis for each nucleotide separately. The null hypothesis is a very simple one, that the 3 possible point mutations from a starting base occur in equal frequency between samples. As such, this spectrum approach does not consider neighbouring base contributions at all and is therefore complementary to it.

For each of the above analyses comparing groups we also undertook mutation spectrum analyses. There were no significant strand differences for autosomal data. Comparisons between the X-chromosome and autosomes revealed highly significant differences in composition for all bases (Figure 4). The most pronounced difference was an excess of A→G* transition mutations on autosomes. Similarly, all point mutations showed significantly different mutation spectra between intergenic and intronic regions (Table S6). In this case, however, the dominant differences were an excess of transversions creating A/T base

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2236v3 | CC BY 4.0 Open Access | rec: 16 Nov 2016, publ:

pairs in intergenic regions while introns were characterised by an excess of C/G base pair creating mutations.

CONCLUSION

The methods we present enable characterising mutational processes affecting samples. For the neighbour analyses, the critical properties of the methods we present derive from the specification of the reference distribution and utilisation of the well established log-linear modelling framework. This combination has considerable potential for detailed interrogations of mutation properties and should improve our understanding the mechanism of mutations, both germline and somatic. Our application of the method generated mutation motifs consistent with well known effects. We further revealed a pronounced influence of flanking bases on all point mutation processes. From germline mutations we have identified a striking dependence of the A→G transition on multiple positions. The mechanistic basis of this mutation motif is unknown.

The neighbour and spectral analyses examine complementary aspects of mutational process. The former examines the contribution of neighbouring bases to the mutation outcome from a starting base and the latter considers the breakdown of mutations from a single base. While the p-values from the hypothesis tests are sensitive to sample size, a property that may be proportional to mutation rate, neither approach explicitly considers the rate of mutation.

As with all methods that seek to characterise data arising from unobserved processes, there are challenges of interpretation. In both the neighbour and spectral analysis approaches, the data are a composite of mutation events with potentially diverse etiological histories. As a consequence, differences between samples will potentially reflect multiple mechanistic differences. Regardless of these issues, analyses that use measures of genetic distance, such as phylogenetics, cannot rationally rely on models of sequence divergence that assume mutations affect nucleotides independent of their neighbours. Instead, models that accommodate neighbour effects (e.g. Hwang and Green, 2004) to at least ± 2 positions will need to be developed in order to reasonably capture the neighbour influences described here.

COMPETING INTERESTS

The authors declare that they have no competing interests.

AUTHOR'S CONTRIBUTIONS

YZ, GAH conceived the project and designed the experiments. YZ, GAH, VBY and TN specified the statistical analyses. YZ and GAH performed the research. All authors wrote the manuscript.

SUPPORTING INFORMATION

S1 File. Supplementary figures and tables.

ACKNOWLEDGEMENTS

We thank Jeremy Widman for allowing us to use his Python implementation of logo drawing code for visualisation. We thank Ben Kaehler and Stephen Haslett for their comments on versions of this work.

REFERENCES

Aggarwala, V. and Voight, B. F. (2016). An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. *Nature Genetics*, 48(4):349–355.

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Aparicio, S. A., Behjati, S., Biankin, A. V., Bignell, G. R., Bolli, N., Borg, A., Børresen-Dale, A.-L., et al. (2013a). Signatures of mutational processes in human cancer. *Nature*.

Alexandrov, L. B., Nik-Zainal, S., Wedge, D. C., Campbell, P. J., and Stratton, M. R. (2013b). Deciphering signatures of mutational processes operative in human cancer. *Cell reports*, 3(1):246–259.

Bernardi, G. (2000). Isochores and the evolutionary genomics of vertebrates. *Gene*, 241(1):3–17.

Brown, T. (2002). *Genomes*. Wiley-Liss.

Chor, B., Horn, D., Goldman, N., Levy, Y., and Massingham, T. (2009). Genomic DNA k-mer spectra: models and modalities. *Genome Biol*, 10(10):R108.

Cooke, M. S., Evans, M. D., Dizdaroglu, M., and Lunec, J. (2003). Oxidative dna damage: mechanisms, mutation, and disease. *FASEB J*, 17(10):1195–214.

Cooper, D. N. (1995). The nature and mechanisms of human gene mutation. *The metabolic and molecular bases of inherited disease*, pages 259–291.

Cooper, D. N. and Youssoufian, H. (1988). The CpG dinucleotide and human genetic disease. *Human genetics*, 78(2):151–155.

Coulondre, C., Miller, J. H., Farabaugh, P. J., and Gilbert, W. (1978). Molecular basis of base substitution hotspots in *Escherichia coli. Nature*, 274(5673):775–780.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., et al. (2013). Ensembl 2014. *Nucleic Acids Res*, page gkt1196.

Forbes, S. A., Beare, D., Gunasekaran, P., Leung, K., Bindal, N., Boutselakis, H., Ding, M., Bamford, S., Cole, C., Ward, S., Kok, C. Y., Jia, M., De, T., Teague, J. W., Stratton, M. R., McDermott, U., and Campbell, P. J. (2015). Cosmic: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res*, 43(D1):D805–D811.

Francioli, L. C., Polak, P. P., Koren, A., Menelaou, A., Chun, S., Renkens, I., van Duijn, C. M., Swertz, M., Wijmenga, C., van Ommen, G., Slagboom, P. E., Boomsma, D. I., Ye, K., Guryev, V., Arndt, P. F., Kloosterman, W. P., de Bakker, P. I. W., and Sunyaev, S. R. (2015). Genome-wide patterns and properties of de novo mutations in humans. *Nat Genet*, 47(7):822–6.

Haldane, J. (1946). The mutation rate of the gene for haemophilia, and its segregation ratios in males and females. *Annals of eugenics*, 13(1):262–271.

Haldane, J. (1948). Croonian lecture: The formal genetics of man. *Proceedings of the Royal Society of London B: Biological Sciences*, 135(879):147–170.

Haldane, J. B. (1935). The rate of spontaneous mutation of a human gene. *Journal of Genetics*, 31(3):317–326.

Harris, K. (2015). Evidence for recent, population-specific evolution of the human mutation rate. *Proceedings of the National Academy of Sciences*, 112(11):3439–3444.

Helleday, T., Eshtad, S., and Nik-Zainal, S. (2014). Mechanisms underlying mutational signatures in human cancers. *Nat Rev Genet*, 15(9):585–598.

Hodgkinson, A. and Eyre-Walker, A. (2011). Variation in the mutation rate across mammalian genomes. *Nat Rev Genet*, 12(11):756–766.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70.

Huttley, G. A. (2004). Modeling the impact of DNA methylation on the evolution of BRCA1 in mammals. *Mol Biol Evol*, 21(9):1760–1768.

Huttley, G. A., Jakobsen, I. B., Wilson, S. R., and Easteal, S. (2000). How important is dna replication for mutagenesis? *Molecular biology and evolution*, 17(6):929–937.

Hwang, D. G. and Green, P. (2004). Bayesian markov chain monte carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proceedings of the National Academy of Sciences of the United States of America*, 101(39):13994–14001.

Ihaka, R. and Gentleman, R. (1996). R: a language for data analysis and graphics. *J. Comput. and Graph. Statistics*, 5:299–314.

Karlin, S. (1998). Global dinucleotide signatures and analysis of genomic heterogeneity. *Curr Opin Microbiol*, 1(5):598–610.

Karlin, S. and Burge, C. (1995). Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet*, 11(7):283–90.

Karlin, S., Campbell, A. M., and Mrázek, J. (1998). Comparative DNA analysis across diverse genomes. *Annual review of genetics*, 32(1):185–225.

Knight, R., Maxwell, P., Birmingham, A., Carnes, J., Caporaso, J. G., Easton, B. C., Eaton, M., Hamady, M., Lindsay, H., Liu, Z., et al. (2007). Pycogent: a toolkit for making sense from sequence. *Genome Biol*, 8(8):R171.

Krawczak, M., Ball, E. V., and Cooper, D. N. (1998). Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. *The American Journal of Human Genetics*, 63(2):474–488.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2236v3 | CC BY 4.0 Open Access | rec: 16 Nov 2016, publ:

Li, W.-H., Yi, S., and Makova, K. (2002). Male-driven evolution. *Current opinion in genetics & development*, 12(6):650–656.

Morton, B. R., Oberholzer, V. M., and Clegg, M. T. (1997). The influence of specific neighboring bases on substitution bias in noncoding regions of the plant chloroplast genome. *J Mol Evol*, 45(3):227–31.

Nik-Zainal, S., Alexandrov, L. B., Wedge, D. C., Van Loo, P., Greenman, C. D., Raine, K., Jones, D., Hinton, J., Marshall, J., Stebbings, L. A., et al. (2012). Mutational processes molding the genomes of 21 breast cancers. *Cell*, 149(5):979–993.

Nishino, H., Buettner, V. L., Haavik, J., Schaid, D. J., and Sommer, S. S. (1996). Spontaneous mutation in Big Blue transgenic mice: analysis of age, gender, and tissue type. *Environ Mol Mutagen*, 28(4):299–312.

Peltomaki, P. and Vasen, H. (1997). Mutations predisposing to hereditary nonpolyposis colorectal cancer: database and results of a collaborative study. the international collaborative group on hereditary nonpolyposis colorectal cancer. *Gastroenterology*, 113(4):1146–1158.

Pleasance, E. D., Cheetham, R. K., Stephens, P. J., McBride, D. J., Humphray, S. J., Greenman, C. D., Varela, I., Lin, M.-L., Ordóñez, G. R., Bignell, G. R., et al. (2010). A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature*, 463(7278):191–196.

Schluter, D. (2009). Evidence for ecological speciation and its alternative. *Science*, 323(5915):737–741.

Schneider, T. D. and Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res*, 18(20):6097–6100.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal*, 27:379–423.

Shiraishi, Y., Tremmel, G., Miyano, S., and Stephens, M. (2015). A Simple Model-Based Approach to Inferring and Visualizing Cancer Mutation Signatures. *PLOS Genetics*, 11(12):e1005657.

Touchon, M., Nicolay, S., Arneodo, A., d'Aubenton Carafa, Y., and Thermes, C. (2003). Transcription-coupled ta and gc strand asymmetries in the human genome. *FEBS Lett*, 555(3):579–82.

Vinson, C. and Chatterjee, R. (2012). Cg methylation. *Epigenomics*, 4(6):655–663.

Webster, M. T., Smith, N. G., Hultin-Rosenberg, L., Arndt, P. F., and Ellegren, H. (2005). Male-driven biased gene conversion governs the evolution of base composition in human alu repeats. *Molecular biology and evolution*, 22(6):1468–1474.

Yakovchuk, P., Protozanova, E., and Frank-Kamenetskii, M. D. (2006). Base-stacking and base-pairing contributions into thermal stability of the dna double helix. *Nucleic Acids Res*, 34(2):564–74.

Ying, H. and Huttley, G. (2011). Exploiting CpG Hypermutability to Identify Phenotypically Significant Variation Within Human Protein-Coding Genes. *Genome Biology and Evolution*, 3:938.

Zhang, X. and Mathews, C. K. (1995). Natural DNA precursor pool asymmetry and base sequence context as determinants of replication fidelity. *J Biol Chem*, 270(15):8401–4.

Zhao, Z. and Boerwinkle, E. (2002). Neighboring-nucleotide effects on single nucleotide polymorphisms: a study of 2.6 million polymorphisms across the human genome. *Genome research*, 12(11):1679–1686.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.2236v3 | CC BY 4.0 Open Access | rec: 16 Nov 2016, publ: