Chapter 2

DNA insertions and deletions in the human lineage

In this chapter, we investigate short (1-100 bp) DNA insertions and deletions that occurred in the human lineage since speciation from its common ancestor with chimpanzee. We find that the majority of insertions are tandem duplications of directly adjacent sequence segments with conserved polarity. Indels in microsatellites comprise only a small fraction. The underlying molecular processes of indel generation do not necessarily rely on the presence of preexisting duplicates, as would be expected for unequal crossing over, as well as replication slippage. Instead, our findings point towards a mechanism that preferentially occurs in the male germline and is not recombination-mediated. In protein-coding regions of the human genome we find that indels are subject to distinct levels of selective pressure with regard to their structural impact on the amino acid sequence, as well as to general properties of the genes they are located in. These observations confirm that many commonly accepted characteristics of selective constraints for substitutions are also valid for amino acid indels. Surprisingly, non-frameshifting tandem duplications and deletions in coding regions still occur at approximately 50% of their genomic background rates. As is already well established in the context of gene and segmental duplications, our results indicate that duplications are also likely to constitute an important process for rapid generation of new genetic material and function on smaller scales.

2.1 Identification of indels in the human lineage

Indel identification from multiple alignments Our comparative genomics analysis to identify indels in the human branch since its split from the common ancestor with chimp utilizes the recently available University of California Santa Cruz (UCSC) whole-genome multiple alignments of 16 vertebrates including human. From these multiple alignments we extracted the human (hg18, Mar 2006), chimp (panTro1, Nov 2003), and rhesus (rheMac2, Jan 2006) tracks. For our purposes, the considered dataset represent the most suitable set available at the moment; major advantages are good sequence quality, high coverage, and small divergence among all three species (pairwise similarities with respect to single nucleotide mutations are

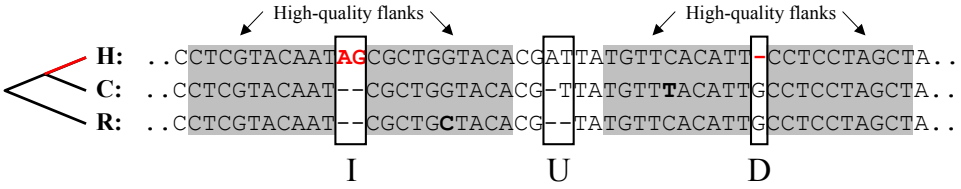


Figure 2.1: Exemplary multiple alignment of orthologous sequence segments in human (H), chimp (C) and rhesus (R). The gap containing regions I and D can unambiguously be explained by a single insertion (I) or deletion (D) event in the human lineage since its speciation from the common ancestor with chimp. In contrast, region U has non-overlapping gaps in chimp and rhesus and therefore requires at least two indel events. These scenarios are always ambiguous. For example, U can be explained by an insertions in human and a deletion in rhesus, but also by a deletion in chimp and a deletion in rhesus.

98.8% between human and chimp, 93.5% between human and rhesus, and 93.3% between chimp and rhesus). The resulting three species alignments cover 85% of the human genome and feature gap lengths of up to 100 bp in each species. If not a result of erroneous alignment, these gaps correspond to insertion or deletion events along branches of the phylogenetic tree ((human,chimp),rhesus). Using rhesus as outgroup, we can explicitly partition indels into insertions and deletions in the human branch by means of maximum parsimony [145, 31]. Since we aim at the identification of reliable characteristics of insertions and deletions, we performed rigorous filtering keeping only unambiguous events located in high quality alignments.

We define a situation as an insertion in the human lineage since speciation from the common ancestor with chimp if the alignment has a segment of gaps in the chimp and rhesus sequences, whereas no gaps are present in the corresponding segment of the human sequence. Additionally, we require the gap segments in the chimp and rhesus sequences to start and end at the same position (case I in Fig. 2.1). This is necessary because alignment regions with not exactly overlapping gap segments in chimp and rhesus cannot be explained by only one insertion event. They require at least two indel events and it is not possible to assign the events to particular branches of the phylogenetic tree in an unambiguous manner (see e. g. case U in Fig. 2.1). Accordingly, we define an event as a deletion in the human lineage if the multiple alignment has a segment of gaps in the human sequence where no gaps are present in the chimp and rhesus sequences (case D in Fig. 2.1). If a deletion occurred in human, the corresponding chimp sequence is taken as an approximation of the ancestral sequence.

The ranking of all different gap motifs with more than 10^5 hits in the three species alignments is shown in Table 2.1. Motifs 3-6 represent elementary events that can unambiguously be explained by a single insertion or deletion event [145, 31]. In particular, motifs 3 and 4 are insertions and deletions in chimp, 5 and 6 are deletions and insertions in human. Motifs 1 and 2 can also be explained by a single indel event, but due to the unknown status of the root we cannot distinguish whether the event occurred in the rhesus lineage, or on the branch from the root to the common

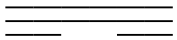
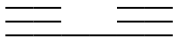
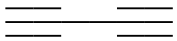
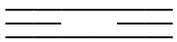
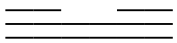
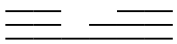
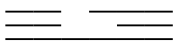
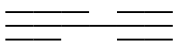
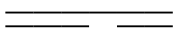
	motif	hits	kbp
1		6124068	27948
2		5457464	32481
3		1171399	4915
4		1129247	3935
5		749968	2416
6		450737	1773
7		116221	1099
8		114681	1122
9		113403	1183
10		105025	1120
other		555805	8664

Table 2.1: Ranking of gap motifs. The three rows of each gap motif correspond to the three species tracks and indicate presence (bars) and absence (empty spaces) of sequence segments of variable length in the three species multiple alignments. Species order is human (top row), chimp (middle row), and rhesus (bottom row). The overall number of basepairs for each motif was calculated by summing the distances between the 5' and 3' ungapped flanks for all gap-containing regions in our multiple alignments, which feature the given gap motif.

ancestor of human and chimp. Overall, motifs 1-6 comprise 93.75% of the number of all events and account for 84.75% of basepairs comprised in gap motifs. For our analysis, we focused on unambiguous insertions and deletions in the human branch after its split from the common ancestor with chimp (motifs 6 and 5, respectively, in Table 2.1).

To further increase the quality of our set we performed a second filtering step excluding those indels from our analysis, which have more than one mismatch or gap in the three species alignments of their 10 bp upstream or downstream flanks (see also Fig. 2.1). This second step additionally filtered out approximately 50% of events along the human branch for the sake of the resulting set now being highly unlikely to result from alignment errors.

Insertion and deletion statistics Our resulting dataset contains 225,744 insertions and 429,048 deletions of high quality in the human branch. A table containing chromosomal position, length, and inserted/deleted sequence of all identified insertions and deletions is provided online [107]. Insertions overall comprise 0.76 Mbp of sequence, deletions 1.36 Mbp. Their length distributions are shown in Fig. 2.2. In the 2384 Mbp of human genomic sequence covered by our multiple alignments this accounts on average for one inserted base per 3.1 kbp, and one deleted base per 1.8 kbp. These rates should be regarded as conservative lower bounds of actual insertion and deletion rates in the human genome. Rather than to derive true indel rates, our study is designed to investigate detailed characteristics and possible origin of inserted and deleted segments. For that purpose, we require a reliable set of high quality indels and applied a strictly conservative filtering procedure. Our numbers are therefore much smaller compared to previous studies. For instance, in 2 Mbp of pairwise human-chimp alignments Britten et al. have measured a cumulative total of 20 kbp located in gaps of length 1-100 bp [24]. Assuming all of these gaps to reflect indel events (which is likely to overestimation the number of actual events due to

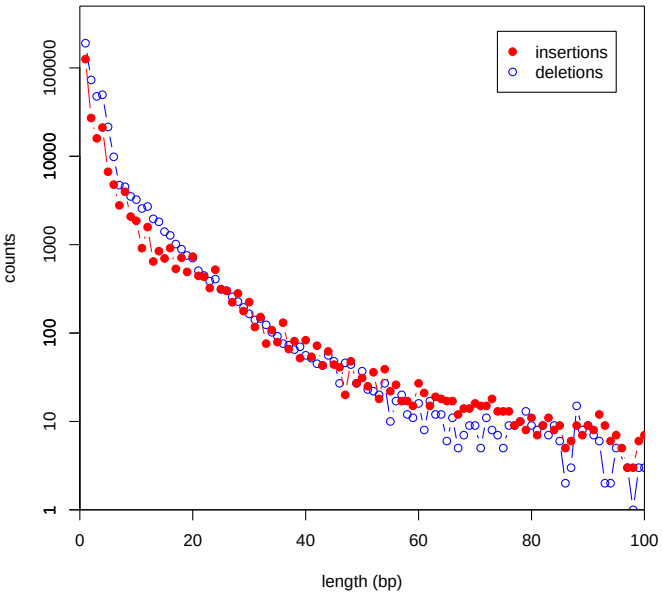


Figure 2.2: Length distribution of the identified 225,744 insertions and 429,048 deletions in our dataset. Short indels comprise by far the largest number of all indels. Single nucleotide insertions (deletions) already account for 56% (44%) of all events on the investigated scales. Note that these numbers are only lower bounds on true indel numbers due to our conservative filtering for high-quality events.

the known low quality of the chimp sequence) and assuming equal indel rates in the human and chimp lineage, this would indicate approximately 5 times higher rates compared to our lower bounds.

Quality assessment We further tested whether our set of identified insertions and deletions might be likely to have originated from sequencing or assembly errors by analyzing sequence quality values. For each contig in the human assembly (Ensembl version 38, Apr 2006) "Base Quality tracks", if available, were retrieved from GenBank using the asn2fsa tool of the NCBI toolbox. This way we could obtain quality values for 420 Mbp, i.e. more than 10% of the human genomic sequence. About 325 Mbp (77%) of these bases are of high quality with quality values of 90 or more. 95 Mbp (23%) are of low quality. For 96 kbp of inserted sequence segments base quality information is available. In this set, 74 kbp (77%) have high quality values of 90 or more. Similarly, we tested the two bases 5' and 3' to deleted sequence segments. In total there are 106 kbp of flanking bases with base quality information, 83 kbp (78%) of which have quality values of 90 and more. Sequence quality in inserted sequence segments and around deleted segments hence reflects that of the genomic background, disproving that indels are preferentially identified in low quality sequence regions.

In addition, assuming that a considerable amount of indels in our set (which overall comprises more than 2 Mbp, i.e. 0.06% of the entire genome) reflects sequencing or assembly errors would also imply much lower sequence accuracy than the claimed 99.99% of the human genome sequence [141]. Sequencing or assembly errors in chimp and rhesus are much less likely to give rise to wrongly identified insertions or deletions, as this would require equal errors in both species' sequences. We conclude that sequencing errors are unlikely to play a major role in our analysis.

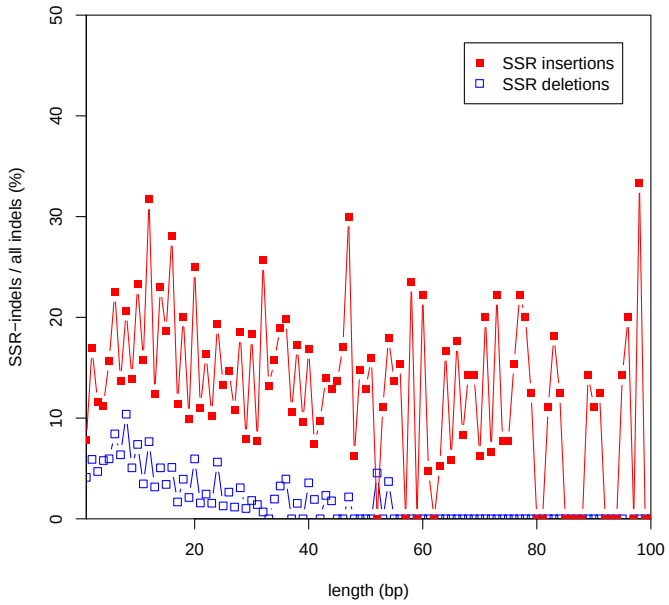


Figure 2.3: Measured relative ratios of SSR insertions and deletions among all insertions and deletions in our set per indel length l. Overall, SSR insertions were found to account for 15% (112 kbp) of the number of all insertions and 5% (70 kbp) of all deletions. The relative ratio of SSR insertions to non-SSR insertions does not significantly decrease for longer insertions, whereas SSR deletions are restricted to short segments.

Indels in Microsatellites Elongation and contraction of microsatellites – tracts of short simple sequence repeats – pose an established mechanism for the generation of short indels [158]. Microsatellites comprise about 3% of the human genome and show a high degree of copy number variation between species and polymorphism within the human population [68].

We annotated indels as simple sequence repeats (SSRs) if at least half of their sequence was identified as SSR by the DUST module of BLAST [6]. According to this classification, we find SSR insertions to account for 15% (112 kbp) of the number of all insertions and 5% (70 kbp) of all deletions in our dataset. The measured relative ratios of SSR insertions and deletions among all insertions and deletions in our dataset per indel length l are shown in Fig. 2.3. Although SSR indels in microsatellites occur at a higher rate compared to non-SSR indels in the genomic background, they make up only a small fraction of all indels in our set. The prevalence of SSR insertions over deletions strongly supports the hypothesis of an overall microsatellite expansion in the human lineage [7, 172].

2.2 Tandem duplications and molecular mechanisms

In the following, we focus on the characteristics and possible origin of non SSR-related indels, which constitute 85% of all insertions and 95% of all deletions. Two mechanisms are commonly regarded as the primary processes capable of inserting and deleting short DNA segments, replication slippage (RS) and unequal crossing

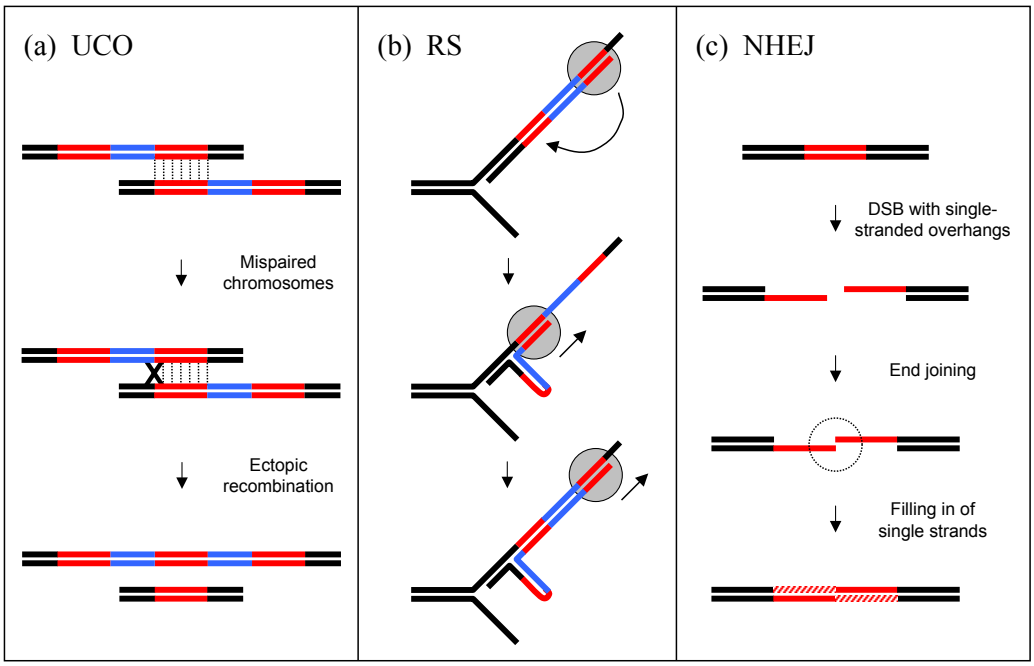


Figure 2.4: Molecular mechanisms of indel generation: (a) UCO can occur during recombination of two homologous chromosomes that contain preexisting duplicates (red sequence segments). If the first copy of the duplicate in one chromosome mispairs with the second copy in the other chromosome, a succeeding crossing over event can result in the generation of a chromosome containing an additional copy of the preexisting duplicate or a deletion of one of its copies [89]. (b) Chromosomal regions that contain preexisting duplicates are also prone to RS. During replication, the DNA polymerase can dissociate right after having passed the second copy of the duplicate and then slip backwards re-associating erroneously at the position of the first copy thereby forming a bulge in the newly synthesized strand [161]. After the next round of replication the new strand will contain an additional copy of the preexisting duplicate. In an analogous manner, forward slippage of the polymerase can cause a deletion. (c) NHEJ (described later in this section) can occur after a double-strand break (DSB) with single-stranded overhangs. If during repair both ends are ligated at the tips of the overhangs the succeeding filling in of the remaining single-stranded intervals will generate tandem duplication insertions. Ligation following the excision of nucleotides at the overhangs can result in deletions [137, 132].

over (UCO). Both processes are also assumed responsible for SSR length variation. The generation of indels by RS and UCO is illustrated in Fig. 2.4 (a) and (b).

Indels resulting from RS and UCO feature common intrinsic characteristics of the inserted or deleted sequence segments and their immediate vicinity. As shown in Fig. 2.4 (a) and (b), both processes require the original presence of two close copies of a DNA segment, UCO for the ectopic recombination between the two copies, RS for the slipped strand misalignment [88]. Signatures of UCO and RS on sequence level are both of the form ABA→ABABA (insertions) and ABA→A (deletions), where A's denote copies of a DNA segment, which might be separated by a spacer

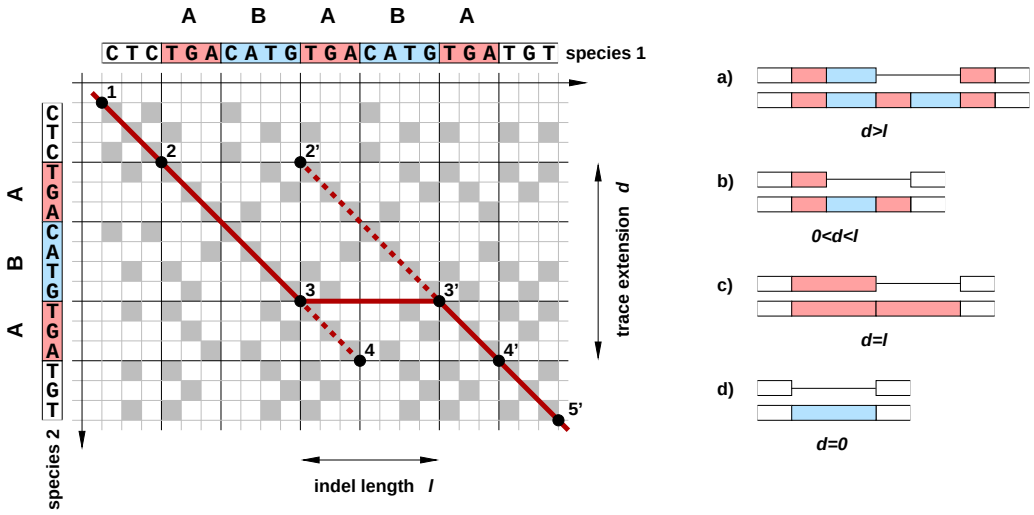


Figure 2.5: Identifying indel duplication signatures by the trace extension method. In the left part of the figure the dot-matrix of an exemplary indel event is shown. Different relations between the indel length l and its trace extension d corresponds to different classes of events. The four major classes are sketched in the right part of the plot.

segment B. Consequently, insertions resulting from either of the two processes are tandem duplications of juxtapositional sequence and deletions result in loss of the spacer B and one copy of the preexisting duplicate A.

If UCO and RS are indeed the predominant mechanisms of indel-generation on a genome-wide level, one should expect to see clear signatures of the above described sequence characteristics among most indels in our set. Investigating this prediction constitutes the aim of the following analysis in which we introduce the measurement of the, so-called, trace extension of an indel as a powerful method to identify tandem duplication insertions and determine the length of preexisting duplicates.

The trace extensions method In our analysis, we want to judge whether an insertion of a sequence segment is in fact a duplication of an adjacent sequence segment, rather than just a random piece of DNA. If so, we would further like to know whether duplicates were already present at the insertion site before the duplication event occurred. Likewise, we want to detect deletion events which resulted from removing one of the two copies of a preexisting duplicate. Measuring the trace extension of an indel allows us to address these questions in a quantitative way.

The trace extension of an indel is a quantity derived from the alignment dot-matrix in the vicinity of the indel event. In our case, dot-matrices are constructed from the homologous sequence segments of human and chimp, which we extracted from the three species multiple alignments. A pairwise alignment of two sequences corresponds to a specific path in the dot-matrix of the two sequences. An exemplary dot-matrix is shown in the left part of Fig. 2.5. The given alignment between the two sequences (solid line connecting points 1,2,3,3',4', and 5') describes a situation where either a sequence segment was inserted in species 1 or deleted in species 2. If by comparison with the out-group species (rhesus) the indel is identified as an insertion in human,

species 1 is assigned to human and species 2 to chimp, vice versa in case of a deletion. The inserted or deleted sequence is identified as the corresponding segment in species 1 between points 3 and 3'. However, for the given scenario the two paths (1,2,2',3',4',5') and (1,2,3,4,4',5') describe alternative alignments which do not involve a different number of gaps and mismatches and will therefore get assigned equal alignment scores. Hence, it is not possible to unambiguously identify the precise sequence of the inserted or deleted segment. This type of ambiguity is generic among indel events involving tandem duplicates. Which specific alignment is actually chosen by the alignment algorithm depends on the particular implementation. The length l of an indel can be defined as the horizontal distance between the starting and end points 3 and 3' of the indel in the original pairwise alignment. It is independent of the particular alignment chosen.

To define the trace extension of an indel we locate two specific points (4 and 2') in the dot-matrix according to the following procedure: From the starting point 3 of the indel in the original pairwise alignment we measure how far we can extend an alternative "forward" alignment path (dashed line) that has to obey positive score along the entire path and is not allowed to rejoin the original alignment path (solid line). We use a gapped Smith-Waterman alignment algorithm with a scoring function that assigns a score of +1 to matching nucleotides, whereas mismatches and gaps are penalized equally with a score of -9. This results in positive alignment scores between the two aligned sequences if their similarity according to edit distance (number of gaps+mismatches) is larger than 90%. As mentioned in the text, in one particular analysis we only require 80% sequence similarity which is achieved by a higher match-score of +2. Point 4 is then defined as the point of maximal score of such an alternative forward alignment. Likewise, 2' is derived by an analogous "backward" alignment starting at 3'. We now define the trace extension d as the vertical distance between points 4 and 2'. Alike the indel length l, the trace extension d is also independent of the particular alignment path chosen if there are ambiguous paths.

Trace extensions and indel duplication characteristics Calculating the trace extension d of an insertion or deletion event allows to identify its duplication characteristics because different relations between d and l corresponds to different classes of events. The four major classes are sketched in the right part of Fig. 2.5: a) The case $d > l$ indicates tandem duplication insertions of the form ABA→ABABA, or deletion events ABABA→ABA, and corresponds to the sketched scenario in the left dot-matrix of Fig. 2.5. The length of the preexisting duplicate A is $d - l$. b) Partially duplicated insertions A→ABA, or deletion events of the form ABA→A, have $0 < d < l$. In this case, the length of A is specified by d. c) Events with $d = l$ are tandem duplication insertions A→AA lacking preexisting duplicates, or complete deletions of one copy of a preexisting tandem duplicate AA→A. d) Non-duplication insertions, or deletions which did not result from taking out one copy of a preexisting duplicate, have $d = 0$. Notice that insertions with $d \geq l$ are always tandem duplications, irrespective of the length of a possibly present preexisting duplicate. This corresponds to cases a) and c). In contrast, b) is not considered as a tandem duplication since the fragment B was not present in the ancestral sequence.

Duplication characteristics of human indels To test whether generation of non-SSR indels is compatible with UCO or RS we computed the trace extension d for all indels in our dataset. As an example, the distribution of d for $l = 8$ bp long indels is shown in Fig. 2.6 (a). Additional plots for different indel lengths are presented in Fig. 2.7. The trace extension analysis revealed that indeed 84% of all non-SSR insertions are tandem duplications, indicated by $d \geq l$. As shown in Fig. 2.6 (b), the proportion of duplications is generally higher for short insertions compared to longer insertions. For instance, 91% of all single nucleotide insertions are tandem duplications in contrast to only 42% of all 30 bp long insertions. We further checked whether the remaining 16% of insertions, which are not identified as tandem duplications ($d < l$), are inverted copies, complementary duplications, or complementary inversions of adjacent sequence and found none of these classes to yield significant contributions. However, as shown in Fig. 2.6 (b), the fraction of insertions with $d < l$ could be substantially reduced by relaxing the required similarity between duplicates from 90% to 80%. We hence suggest that many insertions with $d < l$ might also have originated from tandem duplication events, but divergence between the two duplicates is too high, or multiple indel events might have occurred at the same locus. At first glance, sequence divergence of more than 10% during the investigated rather short evolutionary timescale seems exceptionally large. Yet this could be due to a possibly much higher rate of insertions and deletions in generically unstable regions of the genome. In fact, the generation of tandem duplication insertions will make these regions even more prone for additional insertions and deletions to occur because the newly generated tandem copies can promote further UCO and RS events.

A striking feature of the measured trace extension distributions is the distinct peak at $d = l$ for non-SSR insertions. This peak indicates tandem duplications of the form A→AA and is common among insertions in all investigated length classes (Fig. 2.7). Insertions with $d = l$ are unlikely to have originated by UCO or RS, since no preexisting duplicates were present prior to the insertion event. The steep right flank of the peak at $d = l$ for non-SSR insertions and the rapid decay of the distribution for $d > 0$ among non-SSR deletions indicates a general lack of preexisting duplicates longer than a few bp (the length of a preexisting duplicate is $d - l$ for insertions and d for deletions). Only 25% of all non-SSR insertions and 17% of deletions show signatures of preexisting duplicates longer than 4 bp. However, the conclusions which can be drawn from the length of a preexisting duplicate about a likely molecular mechanism of indel generation depend on the length of the indel. Most 1 bp long non-SSR indels, for example, originate from preexisting duplicates of 1-4 bp. This indicates the presence of a mononucleotide stretch prior to the indel event, short enough not to be annotated as SSR. A one nucleotide slippage of DNA polymerase within this stretch poses a likely scenario for the origin of most such indels. On the other hand, for indels considerably longer than 4 bp the presence of 1-4 bp long preexisting duplicates implies that both copies of the preexisting duplicates are separated by a distance much larger compared to their lengths (see Fig. 2.5). It is rather unlikely that such short and far-spaced duplicates can trigger ectopic recombination between the two copies in case of UCO, or lead to a far backward or forward slippage of DNA polymerase

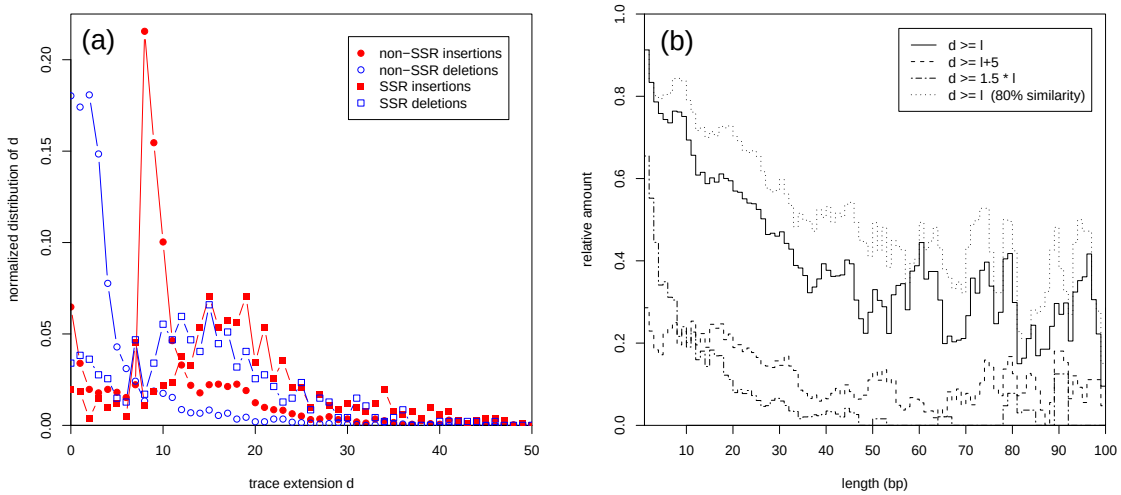


Figure 2.6: (a) Normalized distributions of the trace extension d for $l = 8$ bp long indels. The distinct peak at $d = l$ for non-SSR insertions represents tandem duplications of the form A→AA. In contrast to non-SSR indels, SSR insertions and deletions have broad distributions, and d is significantly larger than l for most indels, as expected for UCO or RS. (b) Duplication signatures of non-SSR insertions. The solid line is the ratio of tandem duplications ($d \geq l$) among all non-SSR insertions per insertion length l, which overall comprise 84% of all non-SSR insertions. The dashed line is the proportion of non-SSR insertions with $d \geq l + 5$, i.e. tandem duplications of the form ABA→ABABA featuring preexisting duplicates A longer than 4 bp. This class comprises 25% of all non-SSR insertions. In comparison to insertions with a fixed preexisting duplicate length, the relative amount of insertions with preexisting duplicates at least half as long as the indel decreases rapidly with increasing indel length (dot-dashed line). The dotted line demonstrates how the proportion of tandem duplications among all non-SSR insertions can be increased by reducing the required sequence similarity for our trace extension analysis from 90% to 80%. All curves have been smoothed using running averages over 3 bp.

during replication. Thus, in order to determine whether an indel is likely to have originated by UCO or RS, it is more adequate to investigate the ratio $(d - l)/l$ of preexisting duplicate length to insertion length (for deletions, the ratio is d/l). For a reasonable ratio of 1.5, i. e. the length of the preexisting duplicate is at least as long as the spacer between the two copies, it is shown in Fig. 2.6 (b) that the ratio of non-SSR insertions with $(d-l)/l \geq 1.5$ decreases rapidly from approximately 65% of all 1 bp non-SSR insertions to less than 10% of all 20 bp long non-SSR insertions. Similar behavior is observed for deletions (data not shown). We conclude from this data that the majority of short indels are compatible with UCO or RS, while many longer indels ($l > 5$) are unlikely to have originated by these processes.

Indel generation by nonhomologous end joining Instead of UCO or replication slippage we propose that a considerable fraction of longer indels might be generated by a different mechanism, based on the imperfect repair of DNA double-strand breaks by nonhomologous end joining (NHEJ) [97, 160]. NHEJ is the most

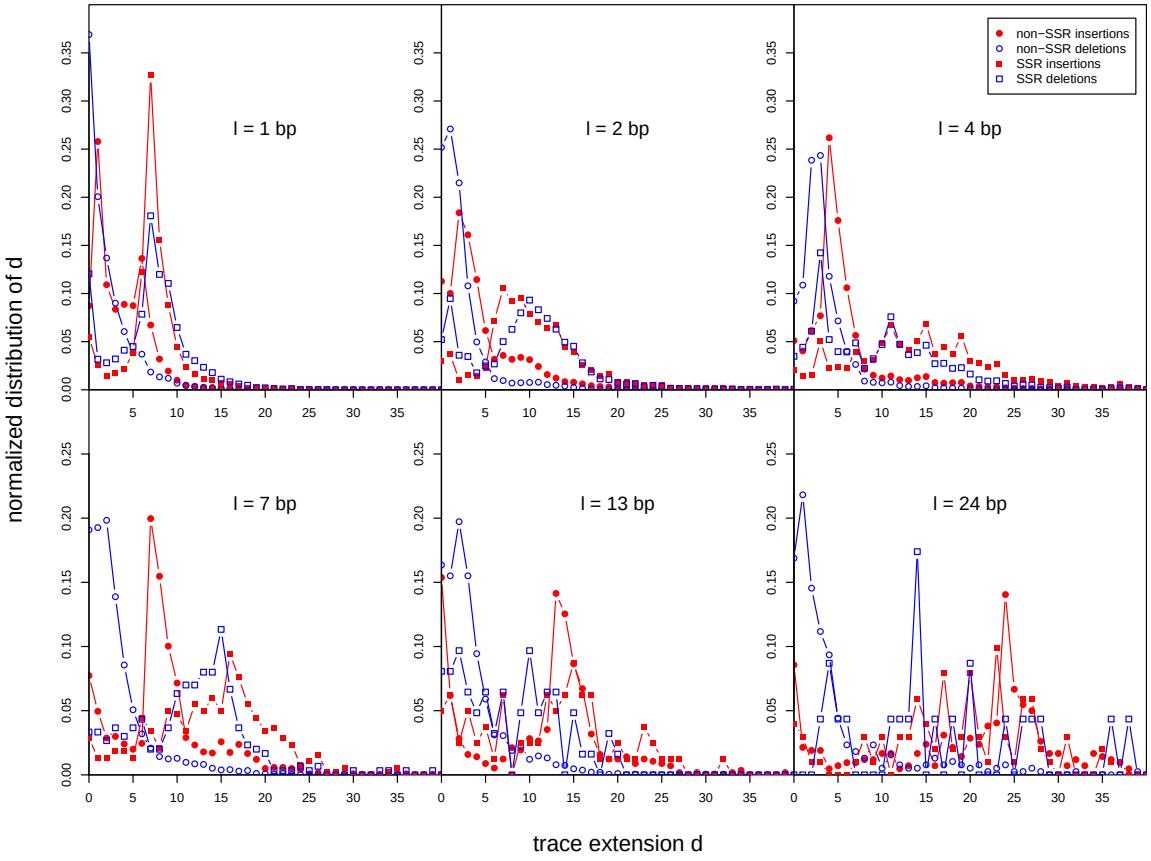


Figure 2.7: Normalized distributions of trace extensions d for different indel lengths.

common double-strand break repair pathway in many organisms and is evolutionary conserved throughout all kingdoms of life. After a DNA break with single-stranded overhangs both ends rejoin by basepairing between opposite single strands. This process is known to often result in gain or loss of DNA, especially if overhangs are damaged [137, 132]. The generation of indels by NHEJ is schematically depicted in Fig 2.4 (c). If basepairing erroneously occurs between microhomologies at the tips of the overhangs, the succeeding filling in of the remaining single-stranded intervals generates tandem duplication insertions. Ligation following the excision of nucleotides at the overhangs can result in deletions. NHEJ requires only short microhomologies of 1-4 bp and can even ligate overhangs without homologies at all [137]. Since double-strand breaks are especially deleterious, it is not surprising that the repair mechanism accepts changes in the nucleotide sequence for the sake of preserved chromosomal integrity. In accordance with our findings long preexisting duplicates are not crucial for indels to be generated by this mechanism.

Indel rate variations between autosomes and sex-chromosomes Further insight into the partial contributions of UCO, RS, and NHEJ to indel generation can be obtained by a separate measurement of indel rates in autosomes (chromosomes 1-22) and the two sex chromosomes X and Y. For example, differences in indel rates between autosomes and sex chromosomes can be used to investigate the importance

of replication in indel generation because the X chromosome spends less time in the male germline and thus undergoes fewer replications than autosomes [102, 96, 49]. Indel rates in the Y chromosome can also help us to elucidate the contribution of recombination because 95% of the chromosome do not recombine in the human lineage [89, 146, 59].

An accurate estimation of insertion and deletion rates in the human Y chromosome on the sole basis of presently available human-chimp-rhesus alignments is limited by the fact that only a female rhesus has been sequenced [136]. At first glance, it is therefore surprising that still 11.4 Mbp of the overall 27.1 Mbp euchromatic sequence of the human Y chromosome are covered in the UCSC three-species alignments, but this is likely to reflect the intricate evolutionary history of the Y chromosome in the human lineage [87, 71, 59]. Alignment blocks where human and chimp sequence segments are located on the Y chromosome whereas the homologous rhesus sequence resides on a different chromosome, account for 4.1 Mbp of the total 11.4 Mbp. We can use these alignment for our analysis of indel rates in the human Y chromosome because the chromosomal disparity presumably results from interchromosomal rearrangement events that either occurred in the rhesus lineage, or on the branch from the last common ancestor with rhesus to the last common ancestor of human and chimp. A second suitable class are, so-called, X-transposed sequences. These sequence regions of the human Y chromosome are considered to have originated from a massive X-to-Y transposition in the human lineage that occurred about 4 million years ago, shortly after the divergence of the human and chimpanzee lineage [127, 119, 142, 146]. Consequently, X-transposed regions are characterized by alignment blocks where human Y-chromosomal sequence is aligned to X-chromosomal sequence in chimp and rhesus. This class comprises another 2.4 Mbp in our multiple alignments. From the combined set of both classes we further removed alignment blocks that are located in the two recombining pseudo-autosomal regions PAR1 and PAR2, as we want to focus explicitly on those regions of the Y chromosome which are not subject to frequent recombination [41, 143, 56]. This leaves us with 6.2 Mbp of human Y-chromosomal sequence in the UCSC data set that we can utilize for the estimation of indel rates in our analysis.

In Table 2.2 insertion and deletion rates are listed for all human chromosomes. We find that rates of insertions and deletions in the X chromosome are significantly lower compared to autosomes ($I_X/I_A = 0.85$, $D_X/D_A = 0.83$). In the Y chromosome, insertion rates do not significantly deviate from the autosomal average, whereas deletions occur at significantly higher rates ($I_Y/I_A = 1.00$, $D_Y/D_A = 1.23$). As shown in Fig. 2.8, different indel length-classes show qualitatively similar behavior.

Our findings point towards several characteristics of the underlying molecular processes of indel generation: First, insertions and deletions are preferentially generated in the male germline indicated by the suppression of indel rates in the X chromosome compared to autosomes. This observation is also compatible with an important role of replication errors because the X chromosome undergoes fewer numbers of replications compared to autosomes. Second, indel generation does not seem to necessarily

chr	I (bp/Mbp)	D (bp/Mbp)	chr	I (bp/Mbp)	D (bp/Mbp)
1	317	550	13	321	586
2	320	579	14	327	559
3	315	567	15	339	599
4	309	603	16	339	522
5	311	585	17	352	550
6	315	590	18	324	572
7	323	586	19	310	541
8	312	579	20	327	539
9	333	554	21	352	667
10	325	545	22	352	586
11	322	565	X	273	475
12	323	561	Y	327	705

Table 2.2: Indel rates per chromosome. Insertion rates (I) and deletion rates (D) were calculated by dividing the overall number of inserted and deleted basepairs per chromosome by the length of its sequence covered in the analyzed multiple alignments. Average insertion and deletion rates in autosomes are $I_{\mathrm{A}} = 326 \pm 13.5$ bp/Mbp and $D_{\mathrm{A}} = 572 \pm 29.9$ bp/Mbp. Insertion and deletion rates in the X chromosome are significantly lower than the autosomal average, deletion rates in the Y chromosome are significantly higher (all differences are larger than three standard deviations of the interautosomal rate variations).

rely on recombination events. If so, we would expect to observe lower indel rates in the Y chromosome, which is not supported by our data. Yet, our findings do not exclude a potential involvement of recombination for a subset of indels. An increase of replication-mediated indels resulting from the higher number of replications in the male germline could compensate for the suppression of recombination-mediated indels in the Y chromosome. Involvement of recombination in at least some fraction of indel events has indeed been supported by the finding that local recombination rates are positively correlated with indel rates on length-scales of 1 Mbp [86].

Nevertheless, both results clearly do not speak in favor of UCO as the predominant mechanism of indel generation. RS and NHEJ, in contrast, do not require recombination events and the suggested preferential occurrence of indels in the male germline is well consistent with the higher number of germ-cell division in males making them more prone to double-strand breaks as well as replication errors.

It has been postulated recently that the contributions of different molecular processes might in fact be unequal for insertions and deletions due to non-trivial mechanistic differences between the two types of mutations [86]. Distinct motifs have for example been found for insertion and deletion hotspots [84, 17]. In particular, it has been claimed that replication-related factors are mostly important for deletions, whereas recombination-related effects are more pronounced for insertions [86]. Both claims are compatible with our measured disparity between insertion and deletion rates in the Y chromosome. Deletion rates are significantly higher in the Y chromosome compared to autosomes, insertion rates are not. No corresponding differences could

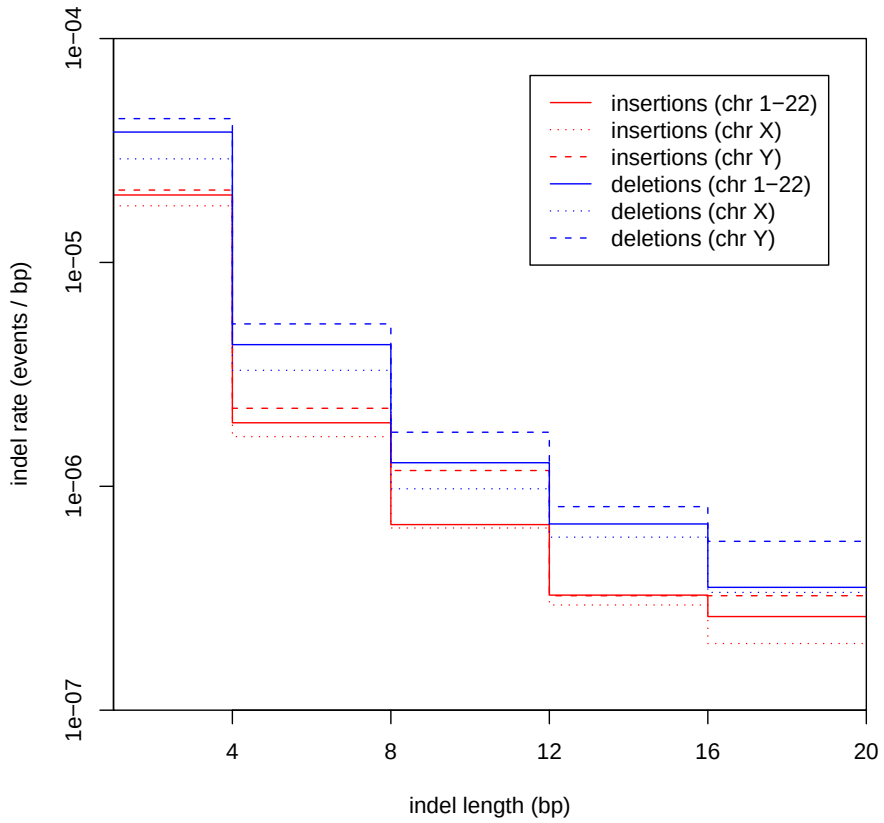


Figure 2.8: Indel rates in autosomes and sex-chromosomes. Rates were calculated by dividing the number of all identified insertions and deletions in the length classes 1-4 bp, 5-8 bp, 9-12 bp, 13-16 bp, and 17-20 bp for autosomes and the two sex-chromosomes by the respective length of chromosomal sequence covered in our analyzed multiple alignments ($L_A = 2274.7$ Mbp, $L_X = 98.3$ Mbp, $L_Y = 6.2$ Mbp). Deletion rates in the Y chromosome are higher than autosomal deletion rates in all investigated length classes, but no clear bias can be observed for Y-chromosomal insertions. In contrast, deletion and insertion rates in the X chromosome are generally lower compared to autosomal rates.

be measured in the X chromosome, but there the magnitude of the effect is also expected to be smaller since the X chromosome still spends one third of its time in the male germline.

Tandem duplications and short paired duplicates in the genome Our findings provide direct evidence that the generation of tandem duplications is the predominant process of DNA insertion on small length scales. This observation can also explain the ubiquity of short paired duplicates in mammalian genomes. For example, copies of 25-100 bp long segments (which do not include known repetitive elements) have been found to be highly overrepresented in vertebrate genomes, yet the two copies are often separated by spacers ranging from a few bp up to several kbp [155, 154]. It has been proposed by Achaz et al. that spaced duplets arise by direct

tandem duplications and separation evolves by subsequent insertion or rearrangement events [1]. However, the nature of the separation mechanism is still controversial [155]. Our analysis shows that spaced duplicates are indeed likely to have originated from juxtapositional copies.

2.3 Indels in protein-coding regions

The analysis presented up to this point focused on general characteristics and molecular origin of short DNA insertions and deletions in the human genome. We thereby aimed at deriving statements on typical properties of short indel events on a genome-wide level without differentiating between distinct genomic substructures, e.g. genes, non-coding regions, repetitive elements, etc. As it is commonly believed that a large proportion of the human genome is evolving under approximate selective neutrality, the results we obtained on indels that have already been fixed in the population (assuming that most events in our set are not polymorphic), are likely to also reflect general properties of the underlying mutational processes responsible for the generation of new indel containing alleles in individuals. For instance, in neutrally evolving regions of the genome the rate of fixation of mutants in the population resembles the rate at which they are generated in individuals.

Selection will break this symmetry. Reduction of mutational rates in specific genomic regions compared to presumably neutrally evolving ones can therefore be indicative for selective constraints associated with the mutational processes in the investigated regions. In the following analysis, we will use this approach to investigate selective forces on insertions and deletions of short DNA segments in particularly important areas of the human genome, the regions that code for proteins.

Identification of coding indels In contrast to the previous analysis in Sections 2.1 and 2.2, where indels were identified from UCSC multiple alignments, the subsequent analysis of indels in protein coding regions utilizes more recent human-chimp-rhesus multiple alignments obtained from the Ensembl database (version 41, October 2006) [67]. They are based on the releases homo sapiens core 41_36c, pan troglodytes core 41_21 and macaca mulatta core 41_10a, and were generated by MLAGAN [25]. Identification of insertions and deletions was conducted in an analogous manner as specified in Section 2.1 for the UCSC alignments. A detailed description of the identification process is presented in [46]. In the resulting set, 724 indels were detected to be located within protein-coding sequence segments according to the Ensembl (version 41) annotation of the human genome [67].

Selective pressure on coding indels The identified 724 indels in protein-coding regions account for only 0.14% of all indels identified from the Ensemble multiple alignments on a genome-wide scale. Comparison of this fraction with the density of protein coding segments, which is about 1.2% for the human genome [69], indicates that indels in coding regions are highly suppressed relative to those in the genomic

background. This can be expected since coding indels will always change the amino acid sequence of the translated protein (in contrast to nucleotide substitutions, which can be synonymous). The effects of indels on the protein sequence can range from insertions or deletions of amino acids if indel lengths are multiples of 3 bp (non-frameshifting indels), up to complete non-functionalization of the protein in case of frameshifting indels. Mutants carrying frameshifting indels are consequently more likely to be removed from the population by purifying selection than those with frameshifting indels [153].

To quantify the amount of purifying selection associated with coding indels in more detail, we calculated the ratio of indel rates in coding regions and those measured in the non-coding background for each indel length l. The resulting ratios are shown in Fig. 2.9. As expected, frameshifting indels are highly suppressed. Non-frameshifting indels, on the other hand, have $I_c/I_{nc} \sim D_c/D_{nc} \sim 0.5$.

This ratio is surprisingly high compared to the ratio of the non-synonymous single nucleotide mutation rate in coding regions and the mutation rate in non-coding regions, which is only $K_A/K_I \sim 0.23$ between human and chimp [35]. Assuming the majority of indels in non-coding regions to be selectively neutral, the observed ratio implies that every second non-frameshifting indel in a coding region is not sufficiently deleterious to be removed by natural selection, in contrast to only one out of four non-synonymous substitutions [138]. Hence, in most of the cases amino acid insertions or deletions seem to have a considerably smaller impact on protein structure and function than substituting one amino acid by another.

Frameshifting indels Despite the approximately 10 times higher suppression of frameshifting indels compared to non-frameshifting indels, we still find 324 events in our set to be frameshifting. This number is unexpectedly high concerning the presumably profound impact of frameshifts on the translated protein sequence. One possible scenario could be that there is only a small number of wrongly predicted Ensemble genes that give rise to many frameshifting indels. Yet this is not supported by our data; there is no gene containing more than two indels of our set, and only 9 (13) genes have two non-frameshifting (frameshifting) indels. Another likely origin of frameshifting indels could be falsely annotated coding regions. To further investigate this possibility, we checked the fraction of indels that are located in experimentally validated RefSeq peptides [121]. While 259 of the 400 non-frameshifting indels (65%) occurred in exons of Ensembl transcripts that could be mapped to RefSeq peptides with at least 99% target and query identity, it was possible for only 108 of the 324 frameshifting indels (33%). This disproportionality is indeed indicative for a significant fraction of frameshifting indels being located in erroneously predicted Ensembl exons, but still a substantial number of events cannot be explained this way.

So far, we are not able to rate what fraction of the remaining frameshifting events is biologically meaningful, and what are the contributions of alignment, sequencing, and other sources of error. In principle, it is also possible that a frameshift caused by one indel can be compensated by a second frameshifting indel. If both events occur within a close distance, changes in the amino acid sequence can be minimized.

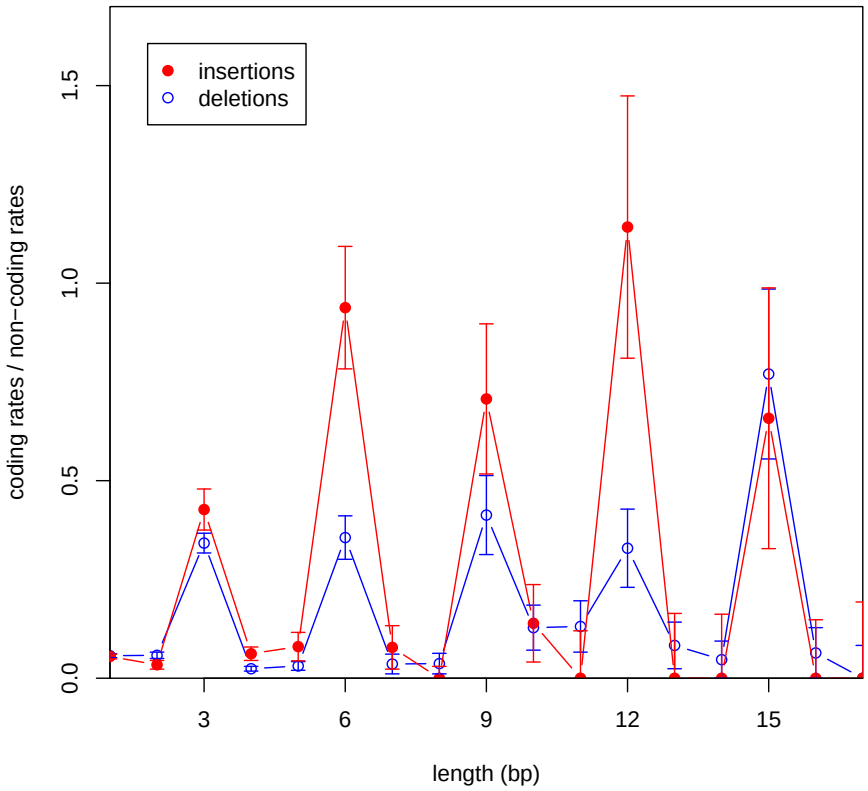


Figure 2.9: Ratios of insertion and deletion rates between coding and non-coding regions. Rates were calculated by dividing the numbers of coding (N_{c}) and non-coding (N_{nc}) autosomal insertions or deletions of length l by the overall length of autosomal coding and non-coding sequence in our multiple alignments ($L_{\mathrm{nc}} = 2695.8$ Mbp, $L_{\mathrm{c}} = 35.0$ Mbp). We further assumed that errors of estimated indel rates are $\sqrt{N_{\mathrm{nc}}}/L_{\mathrm{nc}}$ (non-coding) and $\sqrt{N_{\mathrm{c}}}/L_{\mathrm{c}}$ (coding). If $N = 0$, errors were obtained by setting $N = 1$. Using Gaussian error propagation, errors of the ratios are then determined by $\Delta = (L_{\mathrm{nc}}/L_{\mathrm{c}})(N_{\mathrm{c}}/N_{\mathrm{nc}}^2 + N_{\mathrm{c}}^2/N_{\mathrm{nc}}^3)^{1/2}$.

Inserted and deleted amino acids In contrast to frameshifting indels, which are generally "global" events causing changes on a protein scale, we want to focus on the contribution of indels to protein evolution on a "local" scale in our subsequent analysis. We therefore restricted our set to the 151 insertions and 249 deletions which are non-frameshifting. Their length distribution is shown in Fig. 2.10. It is strongly peaked at 3 bp and rapidly decays for larger indel lengths. A table containing chromosomal position, length, and inserted/deleted sequence of all identified non-frameshifting indels in coding regions is provided online [108].

To investigate whether indels in protein coding regions preferentially induce insertions or deletions of specific amino acids, we counted the distributions of inserted and deleted amino acids in our set. Amino acid sequences of insertions were derived by translating all codons that overlap with the inserted DNA segments. In case

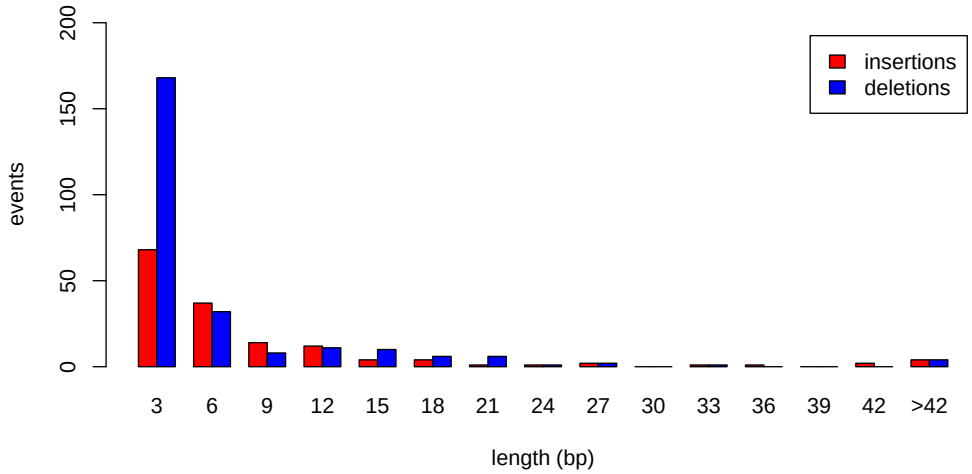


Figure 2.10: Length distributions of non-frameshifting insertions and deletions.

of deletions, the deleted segments were re-inserted in the human sequence and all overlapping codons were translated. Frequencies of the different amino acids were obtained by counting their occurrences in the inserted/deleted amino acid sequences, divided by the overall number of all amino acids in inserted/deleted sequences.

The distributions of inserted and deleted amino acids were compared to the overall abundance of amino acids in proteins of the human genome, which we obtained by measuring the frequencies of amino acids in all protein coding regions of the human genome annotated by Ensembl. As shown in Fig. 2.11 (a), both distributions are significantly different from the background abundance ($p < 10^{-14}$ for insertions, $p < 10^{-10}$ for deletions, Chi Square Test).

We calculated the statistical significance of over- or underrepresentation for all amino acids separately to determine which amino acids contribute most to the observed differences between indel and background distributions. P-values were calculated using $p = \mathrm{erfc}(z/\sqrt{2})$. Throughout this section, z-scores always measure the differences between observed values and background values in standard deviations. All P-values were further Bonferroni corrected for multiple testing (20 tests).

We found that glycine ($p < 0.06$) and alanine ($p < 0.02$) were inserted more often than expected under the assumption that insertion frequencies of different amino acids follow the average distribution of amino acid frequencies in all coding regions of the human genome. Glycine is the smallest among all proteinogenic amino acids, it can therefore be located in parts of the protein that are structurally forbidden to all other amino acids (e. g. tight turns). Alanine is the second smallest amino acid, it is very non-reactive and thus rarely involved directly in protein function [19]. Among deletions, glutamic acid is significantly overrepresented ($p < 0.06$). It is negatively charged and polar, and prefers to be located on the surface of proteins.

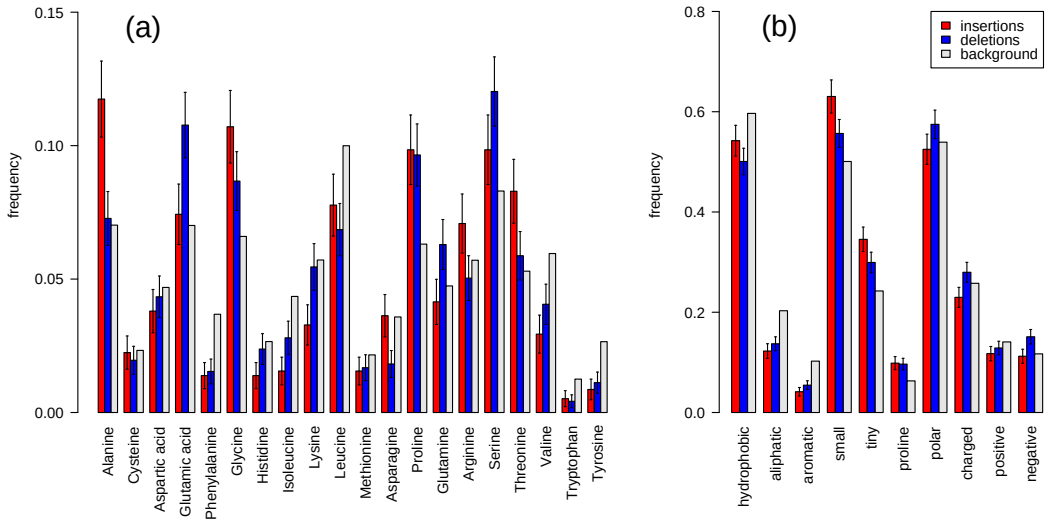


Figure 2.11: (a) Frequency distribution of inserted/deleted amino acids resulting from coding indels in our set compared to the background amino acid frequencies in all human proteins. (b) Frequencies of inserted/deleted amino acids grouped according to 10 different physico-chemical categories. Notice that amino acids can be assigned to more than one category. Error bars in (a) and (b) are standard deviations calculated by $\Delta f_i = \sqrt{N_i}/\sum_j N_j$, where N_i is the total number of inserted/deleted amino acids in category i.

On the other hand, for insertions and deletions phenylalanine and tyrosine (both $p < 0.003$), for insertions isoleucine ($p < 10^{-4}$), lysine ($p < 0.03$) and valine ($p < 0.0006$), and for deletions asparginine, leucine and tryptophan (all $p < 0.03$) are significantly underrepresented among indels. Most of these amino acids prefer to be buried within protein hydrophobic cores (phenylalanine, tyrosine, isoleucine, valine, leucine and tryptophan). Leucine is preferentially located in alpha helices, isoleucine and valine are often found in beta sheets. Asparginine and lysine predominantly reside on the surfaces of proteins [19]. Generally, all significantly underrepresented amino acids are restricted to particular positions in the protein structure. Insertions and deletions of these amino acids are likely to cause major changes in protein structure, stability and function, and are therefore strongly suppressed by purifying selection.

In order to obtain a more general survey of the underlying characteristics that dispose amino acids to be over- or underrepresented in our set, we assigned amino acids to 10 overlapping groups according to their physico-chemical properties: hydrophobic, aliphatic, aromatic, small, tiny, proline, polar, charged, positive and negative [99]. Results are shown in Fig. 2.11 (b). This analysis revealed that indeed small and tiny amino acids are preferentially inserted ($p < 10^{-4}$, all p-values Bonferroni corrected for multiple testing) and tiny amino acids deleted ($p < 0.05$), whereas aliphatic and aromatic amino acids occur less often in inserted ($p < 10^{-6}$) and hydrophobic ($p < 0.002$), aliphatic and aromatic ($p < 10^{-5}$) amino acids in deleted sequence segments compared to their average abundance in protein coding regions.

Insertions and deletions in protein coding regions primarily involve amino acids that have a minor impact on the structure and function of the protein. In contrast, amino acids which are preferentially located in structurally important regions of the protein are highly suppressed. These results agree with the observed dependence of amino acid substitution rates on their local environment within the protein derived from protein alignments [129, 175, 159]. For example, amino acids buried in protein cores have been found to be far more conserved than those at surface positions [126].

Structural preferences of indels To investigate whether the above suggested dependence between indel rate and structural region of a protein can also directly be measured on the structure-level, we retrieved secondary structure information for protein sequences affected by indels from the Protein Data Bank (PDB) [22]. For each indel in our set the sequence of its encompassing protein was blasted against the PDB using blastp from the NCBI QBlast system with default parameters to obtain information on the secondary structure of the protein. In case of a deletion, we blasted the reconstructed ancestral sequence. If more than one hit was reported from the PDB, we chose the first found PDB id which overlaps with the whole indel. The PDB assigns the structural features helix, sheet, turn, or no structure to every amino acid position of the protein. For each of the four structural features we counted the number of indels in our set that reside in a protein region annotated by the structure. If an indel covers more than one structural feature, we weighted each feature by the relative fraction of the length it covers of the indel. For example, a 9 bp long indel where the first 3 bp reside in a protein region annotated as turn, while the last 6 bp are annotated as no structure, adds 1/3 to feature "turn" and 2/3 to feature "no structure". The obtained counts for each structural feature were then divided by the number of all inserted/deleted amino acids with available structural annotation. For the background model, we added for each structural feature the number of amino acids annotated with the feature in all analyzed PDB sequence segments that overlap with the blasted protein sequence, and divided it by the overall length of these segments.

Secondary structure information could be obtained for 343 indels in our set. In Fig. 2.12, we show the distribution of structural features (alpha helix, beta sheet, turn, no structure) among inserted and deleted coding sequence segments in comparison to the background abundance of these features in the analyzed proteins. The analysis corroborates our presumption that coding indels in human preferentially occur in protein regions lacking important secondary structure features, as has already been reported for indels derived from alignments of protein families [129] and coding indels in rodents [153]. In contrast, indels in alpha helices are significantly suppressed ($p < 0.05$, calculated by $p = \mathrm{erfc}(z/\sqrt{2})$ and Bonferroni corrected for 4 tests). This is consistent with the fact that alpha-helices are the most robust secondary structures. Often they form the skeleton of the protein. Amino acid insertions or deletions in protein regions that are supposed to form a alpha helix can have a great impact on the helical structure as they can destroy the internal periodicity of the helix. The observed suppression of indels in these regions is therefore likely to reflect the influence of purifying selection.

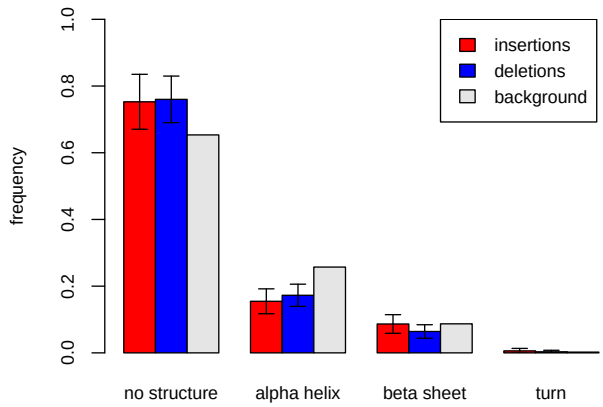


Figure 2.12: Frequency distribution of indel events in the four secondary structure categories helix, sheet, turn and no structure. The background distribution is the relative fraction of amino acids residing in each structure among all analyzed proteins. Error bars were calculated by $\Delta f_i = \sqrt{N_i}/\sum_j N_j$, where N_i is the total number of indels in structure i.

Gene ontology analysis To identify possible correlations between rates of coding indels and categories of proteins that are associated with particular molecular functions, biological processes, or cellular components, a Gene Ontology (GO) [14] analysis was performed among a broad set of 63 GO slim categories [50].

The standard method to investigate whether a certain GO category is over- or underrepresented in a particular subset of genes (e. g. overexpressed genes in a microarray analysis) is to compare the fraction of genes annotated by that GO category in the subset with the fraction of annotated genes in the analyzed background set. However, when analyzing indels such an approach can be misleading if certain GO categories are systematically biased towards shorter or longer genes. The probability of long genes to contain an indel is higher than for short genes. In order to eliminate such possible cross-correlation, we directly measured the rates of coding indels in events per coding sequence length for all genes that could be mapped to our 63 GO slim categories. This way we retrieved category-specific indel rates in events per bp. A list of all 63 estimated rates is presented in [46]. The obtained rates were then compared to the average rate of coding indels in all 16,257 genes of the human genome with available GO annotation. 328 of these genes contain at least one indel of our set. The average rate of coding indels in all annotated genes was calculated to be 1 event per 75 kbp of coding sequence.

To identify GO slim categories with significantly higher or lower indel rates compared to the coding background we assumed that errors of indel rates in a particular GO slim category i are given by $\Delta r_i = \sqrt{N_i}/L_i$, where N_i is the overall number of indels in GO slim category i, and L_i is the total length of all protein coding regions assigned to that category. For GO slim categories with $N_i = 0$ errors were obtained by setting $N_i = 1$. P-values were then calculated using $p = \mathrm{erfc}(z/\sqrt{2})$. The main Gene ontologies cellular component, molecular function, and biological process are independent from each other, but within one group p-values were multiplied by factors 12, 29, and 22, in order, to correct for multiple testing.

We found 6 categories in the ontologies molecular function and biological process which are significantly underrepresented: catalytic activity ($p < 0.04$), ligase activity ($p < 0.0003$), electron transport ($p < 0.003$), amino acid and derivative metabolic process ($p < 10^{-5}$), transport ($p < 0.007$), catabolic process ($p < 0.0002$). All of them are related to biochemical reactions. Suppression of indels in genes associated with these categories may be explained by the fact that biochemical reactions are very specific and are therefore highly conserved throughout evolution.

Chen et al. reported an overrepresentation of indels in genes associated with transcription regulatory activity [31]. We also measure a 1.7-fold higher indel rates in this set of genes. The category transcription regulatory activity characterizes genes that are related to the regulation of other genes. The measured higher indel rates in this class – although not significant after a conservative Bonferroni correction for multiple testing ($p > 0.1$) – conforms well with the hypothesis that many changes between human and chimp took place not only on the amino acid level, but also on the regulatory level [83, 29, 78]. Alongside amino acid substitutions, indels in protein coding regions of regulatory genes could also play an important role among the mutational processes that drive such evolutionary changes. However, the slight overrepresentation may also result from the known enrichment of repetitive sequences in transcription factors [2], which are therefore more prone to frequent indel events.

Evolutionary role of short tandem duplications Gene duplications, large segmental duplications, and entire genome duplications have been widely accepted to promote adaptive evolution and the generation of new genetic functions on large scales [124, 15, 101, 153]. This raises the question to what extend also smaller duplications can contribute to adaptive evolution by generating selectively beneficial variants of proteins or regulatory regions. Several qualitative considerations already point towards a possibly beneficial role of duplications also on intermediate length scales. For example, duplications of small genomic segments have been suggested to accelerate evolution by copy number variations of cis-regulatory motifs [37], or duplication-driven generation of peptide motifs, protein domains, and other functional substructures of genes. It is shown by our analysis that tandem duplication events indeed account for the majority of recently inserted genetic material into the human genome on length scales ranging from short DNA motifs down to single nucleotides. These signatures are also found among coding insertions and deletions; 134 of 151 insertions (89%) in protein-coding regions are tandem duplications. Moreover, we found that non-frameshifting insertions and deletions in protein coding regions are on average less deleterious compared to non-synonymous substitutions. Whether duplications of short DNA motifs are frequently subject to positive selection poses an interesting question for future research that could be addressed by analyzing the degree of polymorphism among small indels.