GENOME RESEARCH

Analysis of copy number variants and segmental duplications in the human genome: Evidence for a change in the process of formation in recent evolutionary history

Philip M. Kim, Hugo Y.K. Lam, Alexander E. Urban, *et al.*

Genome Res. published online Oct 8, 2008;
Access the most recent version at doi:10.1101/gr.081422.108

Analysis of Copy Number Variants and Segmental Duplications in the Human Genome: Evidence for a Change in the Process of Formation in Recent Evolutionary History

Philip M. Kim[1*] Hugo Y. K. Lam[2*] Alexander E. Urban[3], Jan Korbel[1], Xueying Chen[1], Michael Snyder[3] and Mark B. Gerstein[1,2,4¶]

[1]Department of Molecular Biophysics and Biochemistry
[2]Program in Computational Biology and Bioinformatics
[3]Department of Molecular, Cellular and Developmental Biology
[4]Department of Computer Science
Yale University
New Haven, CT 06520

¶ To whom correspondence should be addressed.
Tel: +1 203 432 6105; Fax: +1 360 838 7861;
Email: mark.gerstein@yale.edu

*These authors contributed equally to this work

Abstract:

In addition to variation in terms of single nucleotide polymorphisms, whole genomic regions differ in copy number among individuals. These differences are referred to as Copy Number Variants (CNVs) which recent mapping studies have shown to be prevalent in mammalian genomes. CNVs that reach fixation in the population are give rise to Segmental Duplications (SDs). SDs, in turn, are operationally defined as long (>1kb) stretches of duplicated DNA with high sequence identity. Here, we investigate formation signatures for both phenomena. NAHR employs existing repeats to generate new duplications. Therefore, we examine in detail co-occurrence patterns of different genomic repeat features with both CNVs and SDs. First, we analyzed the localization of SDs with other SDs (i.e. their co-localization) and find that SDs are significantly co-localized with each other, resulting in a highly skewed "power-law" distribution. This observation suggests a preferential attachment mechanism, i.e. existing SDs are likely to be involved in creating new ones nearby. Furthermore, we observe a significant association of CNVs with SDs, but show that a SD-mediated mechanism could only account for a fraction (maximally 28%) of CNVs. As another major contributor to SD formation, Alu elements a type of repeat had previously been identified by virtue of their strong association with SDs. While we also observe this association, we find that it sharply decreases for younger SDs. Continuing this trend, we find only weak associations of CNVs with Alu elements. In the same vein, we report an association of SDs with processed pseudogenes, which is decreasing for younger SDs and absent for CNVs. Finally, we find a number of other repeat elements, namely LINEs and microsatellites, to be significantly more associated with CNVs than SDs, which may explain their formation. Overall, we find that a shift in predominant formation mechanism occurred in the recent evolutionary history. About 40 Mya ago, during a burst in retrotransposition activity (the "Alu burst"), non-allelic homologous recombination (NAHR), mediated by Alus, was the main driver of such genome rearrangement; however, its relative importance has decreased markedly since then, with proportionally more events now being associated with other repeats and with Non-homologous end-joining.

In contrast to the precisely known SD boundaries, most current data on CNVs is of somewhat low resolution, which makes exact conclusions about their surrounding sequences difficult. Therefore, in addition to the coarse-grained analysis above, we performed targeted sequencing of 67 CNV breakpoints and complemented this with previously sequenced ones. We then analyzed the sequence signatures of this combined set of over 600 breakpoints to verify the conclusions that were drawn from the coarse grained analysis. Our findings support the above findings; only few breakpoints show associations with Alu elements, more show formation signatures of NAHR mediated by SDs or LINES.

Keywords:
Segmental Duplication, Copy Number Variant, Non-allelic homologous recombination, Alu element, microsatellite, preferential attachment

Introduction

With the rapid advances in high-throughput technology, the study of human genome variation is emerging as a major research area. A large fraction of variation in terms of SNPs ("point variation") has been mapped and genotyped (The International HapMap Consortium 2005). However, it has recently been recognized that a major fraction of mammalian genetic variation is manifested in an entirely different phenomenon known as copy number variation. In contrast to SNPs, these variations correspond to relatively large (over 1kb according to a widely accepted operational definition) regions in the genome that are either deleted or amplified on certain chromosomes ("block variation") (Freeman et al. 2006; Iafrate et al. 2004; Korbel et al. 2007; Redon et al. 2006; Sebat et al. 2004; Tuzun et al. 2005). They are known as Copy Number Variants (CNVs) and are estimated to cover about 12% of the human genome, thereby accounting for a major portion of human genetic variation (Levy et al. 2007; Redon et al. 2006). Some CNVs reach fixation in the population and (if they correspond to duplications) are then visible in the genome as Segmental Duplications (SD) (Bailey and Eichler 2006). A sizeable fraction (estimated to be 5.2%) of the human genome is covered in these SDs (Bailey and Eichler 2006; Bailey et al. 2002). These are defined as duplicated genomic regions of >1kb with 90% or greater sequence identity among the duplicates. They are especially widespread in the primate lineage (Cheng et al. 2005). SDs enclosing entire genes contribute to the expansion of protein families(Korbel et al. 2008). Some of these duplicated genes may fall out of use, thereby giving rise to pseudogenes. Some duplications that are annotated as SDs may not be fixed in the population, but rather correspond to common CNVs, in particular common ones that are present in the human reference genome. Current efforts to sequence individual human genomes, such as the 1000 genomes project (1000genomes.org), will bring greater certainty about which SDs are fixed and which are polymorphic, and hence are more correctly viewed as CNVs.

Hitherto, not much was known about mechanisms of CNV formation, but it has been suggested that non-allelic homologous recombination (NAHR) during meiosis can lead to the formation of larger deletions and duplications (or to structural variants such as inversions). In general, recombination mechanisms such as NAHR are mediated by pre-existing repeats. Alu elements have been previously implicated in formation of SDs (Bailey et al. 2003; Zhou and Mishra 2005), which is consistent with NAHR-based formation. Likewise, SDs have been suggested as mediating CNV formation (Cooper et al. 2007; Freeman et al. 2006; Sharp et al. 2006). However, not all duplications are thought to arise due to NAHR-based mechanisms: In subtelomeres, a separate mechanism, non-homologous end joining (NHEJ), has been suggested for SD formation (Conrad and Hurles 2007; Linardopoulou et al. 2005). Furthermore, recent studies have uncovered a mechanism that combines both homologous and non-homologous recombination (Bauters et al. 2008; Richardson et al. 1998). Finally, a novel mechanism that involves fork stalling and template switching during replication has been proposed (Lee et al. 2007).

In this work, we examine formation signatures of both SDs and CNVs in an integrated fashion. Specifically, we first survey genomic features in the human and their occurrence. Among the features that we examined are SD and CNV boundaries as well as common repeat elements, such as Alu and LINE retrotransposons and microsatellites. To assess co-localization of the different features, we follow a two-pronged approach: First,

we bin all the features in small sequence bins of 100kb and examine the associations by computing Spearman (rank) correlation coefficients between two features (e.g., Alu elements and CNV breakpoints) as sketched out in Figure 1. This coarse-grained approach is necessary to avoid problems with the comparatively low resolution of current large-scale CNV data (at best 50kb (Coe et al. 2007)). We use the Spearman correlation as a more robust measure to detect non-linear relationships. A high (statistically significant) correlation implies strong co-localization. We interpret statistical enrichment of co-localized elements as an indicator that these elements might be involved in the formation of SDs or CNVs, respectively. Second, to provide further evidence that the co-localization trends found above are due to actual differences formation mechanism, we examined actual breakpoints. Thus far, not many sequences of CNV breakpoints are available. Hence, we performed targeted sequencing, and we analyzed them in combination with a large number of previously sequenced breakpoints. To calculate enrichment of specific features around the breakpoints we compare the number of intersecting features to randomized global and local regions of the genome. Our results show different signatures of formation for SDs and CNVs. While for SDs (and especially older ones), we find a striking enrichment of (among other repeats) Alu elements in the breakpoint regions, suggesting Alu mediated formation, we find little evidence for such a mechanism in CNVs. We present evidence for several alternative features that may contribute to the formation of both SDs and CNVs.

Results

Segmental Duplications follow a power law pattern in the human genome, suggesting a preferential attachment mechanism

SDs are believed to be the result of CNVs reaching fixation. Also, it has been suggested that CNV formation is partly mediated by SDs (Freeman et al. 2006; Sharp et al. 2006; Sharp et al. 2005). Taken together, this would imply that SD formation would preferentially occur in regions with many previously existing SDs. That is, an SD rich region would generate more CNVs than other regions, some of which, in turn, become fixed as SDs. This phenomenon represents one form of a preferential attachment mechanism ("the rich get richer"). This mechanism has been well studied in the physics literature and it is known that it generally lead to a power-law distribution in terms of the regions (Albert and Barabasi 2002). Note, however, while a preferential attachment mechanism does generally lead to a power-law distribution, the inverse is not necessarily the case. A power-law or scale-free distribution corresponds to a distribution with a very long tail (Barabasi and Albert 1999). For our case, this would mean that there should be an extreme imbalance in the distribution of SDs, i.e. a few regions in the genome would be very rich in SDs, while most would contain no or very few SDs. Intuitively, the phenomenon of preferential attachment led to an enrichment of SDs in regions already rich in SDs and hence a highly skewed distribution. Hence, if SD-mediated NAHR is a major factor contributing to new SDs, we would expect the density of Segmental Duplications to be distributed according to a power-law throughout the human genome. Indeed, when analyzing different regions in the human genome for ends of Segmental Duplications harbored, we observe a distinct power-law (See Fig. 2). This power-law behavior is consistent with the existence of rearrangement "hot-spots"(Jiang et al. 2007). This result, taken together with the aforementioned theoretical notions supports the

hypothesis that SD formation is indeed mediated by pre-existing SDs. The power-law distribution is independent of SD size, age, or the binning procedure (See Methods).

Segmental Duplications co-occur best with other Segmental Duplications of similar age

Furthermore, an SD mediated NAHR mechanism would imply that recent SDs should co-occur with older Segmental Duplications. If we bin SDs according to sequence similarity between the duplicates (viewing sequence similarity between the duplicates as approximate age since they diverge after duplication) we should see a significant co-occurrence between SDs most similar in sequence identity. Indeed, we observe a significant correlation between SDs in different age bins (sequence identity) (See Fig. 3). Strikingly, we observe that the best co-occurrence for the SDs of any given age-bin is with the SDs in the "neighboring" bin (i.e., the bin slightly older), consistent with a SD-mediated NAHR. Note that this result would also be consistent with different regions being susceptible to chromosomal rearrangements at different times. However, without a preferential attachment mechanism, we are very unlikely to observe a power-law distribution as in Fig 2. Finally, we observe that this correlation is best for old SDs and gets successively worse as we move towards more and more recent SDs. This may be indicative of a trend of changing SD formation behavior, as we will discuss below.

Alu mediated NAHR is an additional mechanism to preferential attachment

As another mechanism for SD formation, NAHR mediated by Alu retrotransposons has been proposed (Bailey et al. 2003). Note that Alu repeats are the most common repeat element in the human genome with about a million copies. We set out to examine this mechanism and find that indeed, SDs show highly significant co-localization with Alu elements (See Fig. 4B and Table S1), consistent with earlier reports (Bailey and Eichler 2006; Zhou and Mishra 2005). This trend is decreasing rapidly for younger SDs (See Fig 4B), the oldest (most divergent) SDs associate most strongly with Alus. In line with this result, we find that most SDs have low (~90%) sequence identity, similar to Alu elements (See Fig 5). The abundance of both retrotransposed elements and SDs then decreases with rising sequence identity, in sync. SDs appear to co-localize with LINE/L1 repeats, but this association is much weaker and might be reflective of co-localization of Alus and L1 repeats (Kazazian 2004). LINE repeats are very prevalent as well with about 900,000 copies. Therefore, as previously pointed out, Alu elements appear to be mediating SD formation. We also find evidence that Alu mediated mechanisms and preferential attachment mechanisms may be complementary. That is, SDs that co-localize strongly with Alus show weaker correlation with pre-existing SDs (See Fig 4A) than those that appear in Alu-poor regions. This result holds true for SDs of any sequence identity bin. This result suggests that a certain group of SDs is likely to have been formed by an Alu-mediated mechanism and another disjoint group is a more likely candidate for a mechanism involving pre-existing SDs.

Processed pseudogenes show significant association with SDs and a small, but significant number of SDs are flanked by matching pseudogenes

Processed pseudogenes were formed in a way similar to Alu retrotransposons, i.e., they parasitize the same LINE retrotransposition machinery and are also thought to have been

mostly formed during the Alu burst ~40 Mya ago (Zhang et al. 2002). The obvious difference is that there is a much greater variety of pseudogenes than Alu elements. Therefore it is less likely for any given processed pseudogene to find a nearby matching partner, to recombine with, which is a prerequisite for genome rearrangement via homologous recombination. Despite this, we find a strong enrichment of processed pseudogenes with SDs (See Fig 6). To evaluate whether these pseudogenes actually contributed to the formation of SDs, we performed a detailed breakpoint analysis of SDs. For a number of cases (144), we find matching processed pseudogenes at the matching SD junction regions of duplicated regions. In additional 78 cases, we find processed pseudogenes at both SD junctions that have different parent genes, but are highly similar (>95% sequence identity) over stretches of at least 200bp. Note that many pseudogenes have different parents but still show high sequence identity. While these numbers are highly significant (p-values<<0.001), they are relatively small compared to the total number of processed pseudogenes in the human genome (9747, www.pseudogene.org). One reason for this may be that the recombination process requires the pairing of two separate and matching pseudogenes. Since there are far fewer matching pseudogenes than Alu elements, they led to the formation of much fewer SDs. These results suggest that pseudogenes did contribute to SD formation, albeit only in a small number of cases.

Copy Number Variants co-occur with Segmental Duplications

It has been noted previously that Copy Number Variants co-occur with Segmental Duplications and SD mediated NAHR has been suggested as a possible mechanism of CNV formation (Freeman et al. 2006; Goidts et al. 2006; Perry et al. 2006; Sharp et al. 2006). In light with this, CNVs have been viewed as the drifting, polymorphic form of SDs, i.e. SDs correspond to CNVs that have been fixed. This view implies that CNVs should follow a similar pattern of distribution as very young SDs (i.e., SDs of very high sequence similarity), since they would have been created by similar mechanisms. When analyzing SD and CNV distributions in the genome, we indeed find that there is a significant overlap (See Fig. 7A). However, the correlation between SD and CNV occurrence may be smaller than expected. We find that maximally 28% of CNVs were formed by an SD-mediated mechanism, i.e. lie in a region with a nearby SD. This is an upper bound estimate, since proximity does not imply causality. From another perspective, one may (perhaps naively) expect that the similarity in distribution of CNVs and SDs of >99% sequence identity should be comparable to the similarity between the distributions of SDs of >99% sequence identity and SDs of 98-99% identity. However, we find that the correlation for CNVs and young SDs (rank correlation of 0.11) is lower than the one for “very old” (90-92% sequence identity) and “very young” (>99% sequence identity) SDs (rank correlation of 0.24). In other words, about 80% of “very young” SDs could be the result of NAHR mediated by older SDs. Conversely, the same can be said of only 28% of CNV. This may be consistent with the fact that CNVs are polymorphic whereas SDs are fixed.

Copy Number Variants do not show any significant association with Alu elements, but associate with other repeats

If CNVs and SDs are formed by similar processes, one might assume that CNVs would also show association with Alu elements. However, we find that CNVs show no

significant association with Alu elements (See Fig. 7B). Previous studies found weak associations of CNVs with Alu elements(Cooper et al. 2007), but they are much weaker than the ones found for SDs (of any sequence identity bracket). Indeed, when controlling for SD content, the association becomes even weaker (See Supplement). These associations may be due to the low resolution of the data that underlies most of these studies (250K for data from competitive genome hybridization (CGH) on BAC microarrays). Indeed, for all studies that used technology of higher resolution(Khaja et al. 2006; Korbel et al. 2007; Tuzun et al. 2005) no association with Alu elements was found (See Table S1).

This result implies that an Alu mediated mechanism is an unlikely candidate for CNV formation. It is consistent with reports that Alu mediated NAHR was most common during or shortly after the burst of Alu activity ~40Mya ago and has since declined (Jurka 2004). Hence, the formation of CNVs and some SDs is probably mediated by different phenomena. One might argue that some of this difference is due to the different methods of experimental determination – SDs are read directly from the genome and CNVs used in this study are determined using microarrays. Therefore, we computed associations between Alus and CNVs that were determined using very different methodologies, including different kinds of microarrays and paired-end sequencing (See Table S1). Therefore, we conclude that Alu elements, while active in genome rearrangements in the past, do not currently play a major role in the formation of CNVs. It should be pointed out that this result does not contradict the notion of CNVs as drifting SDs – it simply suggests that the mechanism of CNV/SD formation may have undergone significant change in the past 40 million years.

The absence of association with Alu elements and the weakness of co-localization with SDs leads to the question of which genomic features are relevant for CNV formation. It has been suggested that microsatellite repeats have a role in mediation of chromosome rearrangements (Ugarkovic and Plohl 2002). An association of SD junctions with microsatellites has previously been pointed out (Bailey et al. 2003). Hence, we examined whether they would associate with known CNVs. We indeed find that microsatellite repeats show a highly significant co-localization with CNVs (See Fig 7B and C and Table 1), even after correcting for SD abundance.

Analysis of sequenced breakpoints

A difference between SDs and most of the current CNV data is that SD breakpoints are known exactly, whereas for CNVs only their approximate location is known (based on CGH experiments). As mentioned above, most of the current data has a resolution of at best 50kb (Coe et al. 2007). To make authoritative statements about formation signatures one has to analyze the exact sequences surrounding the breakpoints. Hence, we performed targeted sequencing of a number of representative CNV breakpoints, and identified a total of 132 breakpoints (See Table 2). We combined this with previously sequenced breakpoints (Korbel et al. 2007) to analyze a total set of 534 breakpoints, representative of all CNV events. To verify the trends we identified using the large-scale data, we analyzed the enrichment of different repeat elements in the immediate vicinity of the breakpoints and the existence of matching repeats flanking both sides of the breakpoints. To control for local sequence biases, we calculated the enrichment both with respect to the entire genome (global enrichment) and a 5kb region around the breakpoint

(local enrichment) (See Table 1). We find only an extremely weak association with Alu elements, confirming the above trend. In total we find 29% of the breakpoints to be associated with LINE repeats and another 2% to be associated with SDs. 9% were flanked by other repeat elements (e.g., LTR and others). The remainder (60%) of breakpoints did not show any homology signature. We should note here that the PEM (using short sequence reads) approach is likely to bias somewhat against repeat-rich regions, and hence the fraction of NAHR-mediated CNVs may be higher in reality. This may also explain the discrepancy between the above found fraction of SD mediated CNVs (maximally 28%) with the one found here (about 2%). However, many exhibit signatures that may be indicative of non-homologous end-joining (NHEJ). Specifically, 40% of the breakpoints show the so-called microhomologies that can be a sign of NHEJ(Lieber et al. 2003). Another 14% exhibit micro-insertions which have also been implicated in NHEJ. We hence estimate that the latter CNVs may have been formed by double strand breakage and NHEJ. Aside from these sequence signatures, there is also biophysical evidence: breakpoints are enriched in regions that are known to be genomically unstable: We find that breakpoint regions tend to lie in GC poor regions (See Table 1), which are known to be thermodynamically less stable. Moreover, NHEJ breakpoints tend to lie in significantly less stable regions than NAHR breakpoints (p-value<0.01). Moreover, we find that a few NHEJ breakpoints lie in the unstable subtelomeric regions, while no NAHR breakpoints do. We hence hypothesize that random breakage followed by NHEJ is one major mechanism for CNV formation.

Discussion

We have presented results that suggest changes in the formation of large genome rearrangements over the past 40 Mya. Our results suggest that shortly after the burst in Alu activity, Alu- or pseudogene-mediated mechanisms were predominant in the formation of SDs. The formed SDs then presented highly homologous regions themselves and were active shortly after formation in generating new SDs. However, it is striking to see that the association of SDs with Alu elements is decreasing with decreasing age of the SD (increasing sequence similarity between the duplicates) (Fig 4B). Likewise, the co-localization of SDs with their younger counterparts is decreasing. These trends are indicative of a lesser contribution of homology mediated mechanisms for SD formation. At almost the same rate, preference of SDs for subtelomeric regions in the genome is increasing (Fig 4B). Genesis of SDs in subtelomeric regions is largely due to a mechanism based on non-homologous end joining (NHEJ) mediated by micro-homologies (<25bp homology), rather than a NAHR mechanisms mediated by larger matching repeats (Linardopoulou et al. 2005). Note that an alternative hypothesis for the enrichment of SD breakpoints in Alu rich regions is the clustering of Alu elements(Jurka et al. 2004).

The lack of association of CNVs with Alu elements is quite surprising, as concurrent Alu-Alu recombination has been reported in the literature (Deininger and Batzer 1999; Nystrom-Lahti et al. 1995). However, our results indicate that while Alu-Alu recombination used to be a major force in shaping genome rearrangements, in the very recent genome evolution it did not leave a significant signature. Furthermore, our sequenced breakpoints confirm the absence of Alu elements near the breakpoints. Do note however, that there may be some bias of the sequencing method against Alu repeats.

Moreover, it is in line with the emerging trend of decreasing Alu association of SDs. It is likely the result of the decrease in Alu activity since the Alu burst, which led to continuing Alu divergence and hence, diminishing probability of Alu mediated NAHR. This finding is further bolstered by the fact that most SDs have a similar sequence divergence (age) as most Alus, i.e. they were likely created around the Alu burst. While association does not imply causality, the lack of association (such as here, with Alu elements and CNVs) certainly implies lack of causality. In other words, it would be hard to argue that Alu elements are the predominant mediator of CNV formation solely based on the observation of co-localization. Thus, our observations provide strong evidence against the involvement of Alu elements in CNV formation.

On the other hand it has previously been suggested that CNVs associate with SD elements, and we find this trend persisting. However, SDs mediated CNV formation can only account for a minority of the CNVs found (less than 10% based on our sequenced breakpoints). Therefore, other mechanisms have to be at work as well. We suggest the following two possibilities for alternative mechanisms: First, we find associations of CNVs with other repeats, namely microsatellites and LINE repeats. Large-scale associations only gives weak evidence for this connection, but the presence of matching repeats in the immediate vicinity of the sequenced breakpoints makes a stronger case for microsatellites and LINE involvement in CNV formation. Moreover, since microsatellites have been implicated in genome rearrangements, an involvement in CNV formation would certainly be sensible (Ugarkovic and Plohl 2002). Second, our findings are also suggestive of an increased role of NHEJ based mechanisms for the generation of CNVs which accounts for many of the breakpoints that were not associated with any known repeat. Indeed, we find an association of CNVs towards subtelomeric regions (p-value<0.001), where double strand breakage and NHEJ is known to be prevalent. Moreover, in the sequenced breakpoint data we find some indication that NHEJ is an alternative mechanism for CNV formation, such as the microhomologies present in many breakpoint sequences.

In summary, we find evidence for formation of duplications via NAHR that was mediated by repeat elements. While the co-localization does not imply causality, this mechanism has been proposed before and is supported by several pieces of data for SDs. It also explains nicely the decrease of co-localization of SDs with Alus and with each other. This leads to a coherent picture: about 40 Mya ago, there was a peak in Alu activity, known as the Alu burst (see Figure 8). The burst created a high number of repeat elements that served as templates for NAHR. Hence, ectopic recombination took place at a high rate and set off extensive genome rearrangement, thereby creating many SDs. The SDs themselves then could also serve as NAHR templates, “feeding the fire” of recombination. This also nicely explains the existence of the rearrangement hot-spots in the current human genome. Therefore, the majority of SDs that we find have low sequence identiy (~90%), similar to Alu elements stemming from the burst, suggesting that they were formed during a similar time. Moreover, the number of SDs decreases with rising sequence identity, in sync with the decrease of Alu repeats (correlation r=0.92, p<0.001, see Figure 5). This is consistent with our hypothesis, that the decline in retrotransposition activity then led to an overall decline in genome rearrangements. Moreover, the relative importance of other repeat elements, such as LINE elements or microsatellites in terms of mediating NAHR increased; while they were created in the

genome at a basal level, the strong effect of the Alu burst had previously masked their influence. This is why we find a stronger signature of enrichment of these elements with CNV breakpoint regions. Finally, other mechanisms, namely NHEJ, play a much bigger role in reshaping the genome today, again consistent with the fact, that a majority of current CNV breakpoints exhibit signatures suggesting a formation through NHEJ.

Aside from the factors discussed above, selection could have influenced the sequence signatures found around SDs or CNVs. Many SDs may have undergone some kind of selection during their way to fixation. By contrast, most CNVs are likely to be neutral, even though, analogous to SNPs, some may have been selected for or against(Cooper et al. 2007; Hurles et al. 2008; Korbel et al. 2007). Hence, one may assume that the differences between CNVs and SDs pointed out above could be due to selection. The most striking difference is certainly the difference in association with Alu elements; if selection were responsible for this difference, two scenarios are possible: First, Alu elements in the vicinity of SDs could lead to preferential fixation of these SDs. It is hard to imagine how Alu elements in the genomic neighborhood should influence the fixation of SDs, therefore we deem this scenario very unlikely. Second, Alu elements in the vicinity of CNV were removed by negative selection. This possibility is equally unlikely, and we believe that the far more parsimonious explanation is that Alu elements had a predominant role in past SD, but not in present CNV formation.

Conclusions

We present evidence for different formation mechanisms of structural variants in the human genome. Our main result suggests that currently occurring Copy Number Variants appear to follow a pattern somewhat similar to young Segmental Duplications and decidedly different from older Segmental Duplications. We show a shift from a prevalence of Alu-mediated generation of old SDs towards other mechanisms for more recent SDs. The weakness of association of CNVs with Alu elements can be viewed as the natural extension of this trend, as CNVs (that correspond to amplifications) are “very young” SDs. This trend is consistent with current models that propose a decrease of Alu activity after the “Alu-burst” about 40Mya ago. Finally, we present results suggesting that while some CNVs are formed through NAHR, a large fraction of them are formed through non-homologous end-joining. These trends are present in the large amounts of low-resolution data as well as found confirmed in the substantial number of sequenced breakpoints.

Methods

Sequence data preparation

We used the segmental duplications database from the University of Washington (http://humaparalogy.gs.washington.edu/dups) based on the build 36 genome (Bailey et al. 2002). We binned all existing SDs into sequence identity categories and different size categories (See Supplement). To enable comparison with low-resolution copy number variation data, we finally binned all segmental duplications according to genomic coordinate. We varied the binsize from 10kb to 1Mb. Because the copy number variant mapping resolution is at most 50kb for the techniques employed in the used datasets(Coe et al. 2007), we report the results for calculations with a binsize of 100kb. Calculations using other binsizes are reported in Table S1. For copy number variants we used three

separate datasets, based on three different assay methodologies. The three-way comparison should avoid biases that may have been introduced by a single method. First, we used the recent set from the Human Copy Variation Consortium, which was based on microarray methods(Redon et al. 2006). Secondly, the structural variation data based on Fosmid-paired-end sequencing was used(Tuzun et al. 2005). Finally, a comparison of two different genome assemblies has revealed putative copy number variations(Khaja et al. 2006). The results from the latter two CNV datasets are reported in Table S1.

Breakpoint sequencing

A total of 67 individual breakpoints identified by the Paired end matching (PEM) were sequenced using the following approach. PCR fragments were extracted either by gel-purification or gel-extraction with Millipore Ultrafree®-DA centrifugal filter devices (Millipore Corp., Bedford, MA) or by bead-purification from the reaction mixture with Agencourt® AMPure® (Agencourt Biocience Corporation, Beverly, MA). Amplified fragment pools (50 – 150 fragments each) were randomly sheared by nebulization, converted to blunt-ends, and adaptors ligated with the GS DNA Library Preparation kit according to the manufacturer's protocols (454 Life Sciences, Branford, CT; Roche Diagnostics, Indianapolis, IN). The resulting single stranded DNA shotgun libraries were then sequenced with 454 Sequencing. Both the resulting reads (median length=250bp) and contigs generated by 454's de novo assembler Newbler (see software user manual, 454 Life Sciences and Roche Diagnostics) were scanned for the respective SV-breakpoints with BLAST (S4) alignment against the human reference genome; we required best-hits to the genome for both portions of a read/contig matching on either side of a candidate breakpoint junction.

Repeat Analysis

Different kinds of repeats were identified using the genome annotation on the UCSC genome browser, based on the output of Repeatmasker. As above, distributions of Alu elements, LINE elements, and microsatellites were binned according to their genomic coordinates. Recombination hotspot data was taken from the HapMap recombination data(The International HapMap Consortium 2005). Data for the processed pseudogenes was obtained from Pseudogene.org(Karro et al. 2007).

Computation of associations

Coarse-grained co-localization was assessed by computing the spearman rank correlations between the binned distributions of each feature (SD occurrence, CNV occurrence or repeat occurrence) per bin. This measure is an accurate and robust measure of association and is independent of any assumptions of the distribution of the respective features. We used a binsize of 100kb for the analysis, but changes in the binning procedure did not have an effect on our results (See Supplement). This coarse grained approach can identify larger scale trends. It is especially suitable for the analysis of CNV associations because of the current low resolution mapping of their breakpoints. However, it may not be able to pinpoint exact breakpoint characteristics.

For sequenced breakpoints we calculated enrichments both in a global and local context. In a global context, we compared the average number of a random nucleotide in the genome intersecting with a given genomic element with the average number a breakpoint

did. Since this may be biased by local genomic context, we also calculated the average number a random nucleotide would intersect a given genomic element in a 50kb window around the breakpoint.

Detailed SD breakpoint analysis for processed pseudogenes

For a detailed analysis of processed pseudogene enrichment at SD breakpoints we analyzed all SD junctions for overlap with pseudogenes. Because of potential sequencing and alignment errors, we defined the SD junction as +/- 5 basepairs around the annotated breakpoint. We then looked for SDs where pseudogenes overlapped either the SD start or end junction in both duplicated segments. For each of these, we then compared the parent genes of the two pseudogenes that overlapped the SD junctions. For pseudogenes with different parent genes, we compared their sequence similarity using FASTA.

To assess the significance of the overlap between the processed pseudogenes and SD junctions, we first picked genomic regions of the same size and number as SDs at random and compared the overlap with processed pseudogenes. No matching junctions that had matching pseudogenes were found. As a second procedure that captures potential sequence biases, we randomized the SD junctions in a 5kb window around the actually junction and calculated the overlap with matching pseudogenes.

CNV breakpoint analysis

To complement the coarse-grained approach, we analyzed a set of 536 sequenced breakpoints, a combination of the breakpoints from Korbel et al. and the newly sequenced breakpoints above. We analyzed the occurrence of breakpoints in known repeat sequences from Repeatmasker. Furthermore, we analyzed each breakpoint for the occurrence of microhomologies and microinsertions. All calculations were carried out using custom code in Matlab, R and Perl.

All data and supplementary material is available on our website:
http://www.gersteinlab.org/proj/sdcnvcorr

Acknowledgments

We thank George Perry for careful reading of the manuscript and many insightful comments. We also thank Tara Gianoulis, Prianka Patel and Deyou Zheng for comments on the manuscript, technical assistance and helpful suggestions. We acknowledge support from the NIH and from the AL Williams Professorship funds.

References

Albert, R. and A.L. Barabasi. 2002. Statistical Mechanics of Complex Networks. *Review of Modern Physics* **74:** 47-97.

Bailey, J.A. and E.E. Eichler. 2006. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat Rev Genet* **7:** 552-564.

Bailey, J.A., Z. Gu, R.A. Clark, K. Reinert, R.V. Samonte, S. Schwartz, M.D. Adams, E.W. Myers, P.W. Li, and E.E. Eichler. 2002. Recent segmental duplications in the human genome. *Science* **297:** 1003-1007.

Bailey, J.A., G. Liu, and E.E. Eichler. 2003. An Alu transposition model for the origin and expansion of human segmental duplications. *Am J Hum Genet* **73:** 823-834.

Barabasi, A.L. and R. Albert. 1999. Emergence of Scaling in Random Networks. *Science* **286:** 509-512.

Bauters, M., H. Van Esch, M.J. Friez, O. Boespflug-Tanguy, M. Zenker, A.M. Vianna-Morgante, C. Rosenberg, J. Ignatius, M. Raynaud, K. Hollanders, K. Govaerts, K. Vandenreijt, F. Niel, P. Blanc, R.E. Stevenson, J.P. Fryns, P. Marynen, C.E. Schwartz, and G. Froyen. 2008. Nonrecurrent MECP2 duplications mediated by genomic architecture-driven DNA breaks and break-induced replication repair. *Genome Res* **18:** 847-858.

Cheng, Z., M. Ventura, X. She, P. Khaitovich, T. Graves, K. Osoegawa, D. Church, P. DeJong, R.K. Wilson, S. Paabo, M. Rocchi, and E.E. Eichler. 2005. A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* **437:** 88-93.

Coe, B.P., B. Ylstra, B. Carvalho, G.A. Meijer, C. Macaulay, and W.L. Lam. 2007. Resolving the resolution of array CGH. *Genomics* **89:** 647-653.

Conrad, D.F. and M.E. Hurles. 2007. The population genetics of structural variation. *Nat Genet* **39:** S30-36.

Cooper, G.M., D.A. Nickerson, and E.E. Eichler. 2007. Mutational and selective effects on copy-number variants in the human genome. *Nat Genet* **39:** S22-29.

Deininger, P.L. and M.A. Batzer. 1999. Alu repeats and human disease. *Mol Genet Metab* **67:** 183-193.

Freeman, J.L., G.H. Perry, L. Feuk, R. Redon, S.A. McCarroll, D.M. Altshuler, H. Aburatani, K.W. Jones, C. Tyler-Smith, M.E. Hurles, N.P. Carter, S.W. Scherer, and C. Lee. 2006. Copy number variation: new insights in genome diversity. *Genome Res* **16:** 949-961.

Goidts, V., D.N. Cooper, L. Armengol, W. Schempp, J. Conroy, X. Estivill, N. Nowak, H. Hameister, and H. Kehrer-Sawatzki. 2006. Complex patterns of copy number variation at sites of segmental duplications: an important category of structural variation in the human genome. *Hum Genet* **120:** 270-284.

Hurles, M.E., E.T. Dermitzakis, and C. Tyler-Smith. 2008. The functional impact of structural variation in humans. *Trends Genet* **24:** 238-245.

Iafrate, A.J., L. Feuk, M.N. Rivera, M.L. Listewnik, P.K. Donahoe, Y. Qi, S.W. Scherer, and C. Lee. 2004. Detection of large-scale variation in the human genome. *Nat Genet* **36:** 949-951.

Jiang, Z., H. Tang, M. Ventura, M.F. Cardone, T. Marques-Bonet, X. She, P.A. Pevzner, and E.E. Eichler. 2007. Ancestral reconstruction of segmental duplications reveals punctuated cores of human genome evolution. *Nat Genet* **39:** 1361-1368.

Jurka, J. 2004. Evolutionary impact of human Alu repetitive elements. *Curr Opin Genet Dev* **14:** 603-608.

Jurka, J., O. Kohany, A. Pavlicek, V.V. Kapitonov, and M.V. Jurka. 2004. Duplication, coclustering, and selection of human Alu retrotransposons. *Proc Natl Acad Sci U S A* **101:** 1268-1272.

Karro, J.E., Y. Yan, D. Zheng, Z. Zhang, N. Carriero, P. Cayting, P. Harrrison, and M. Gerstein. 2007. Pseudogene.org: a comprehensive database and comparison platform for pseudogene annotation. *Nucleic Acids Res* **35:** D55-60.

Kazazian, H.H., Jr. 2004. Mobile elements: drivers of genome evolution. *Science* **303:** 1626-1632.

Khaja, R., J. Zhang, J.R. MacDonald, Y. He, A.M. Joseph-George, J. Wei, M.A. Rafiq, C. Qian, M. Shago, L. Pantano, H. Aburatani, K. Jones, R. Redon, M. Hurles, L. Armengol, X. Estivill, R.J. Mural, C. Lee, S.W. Scherer, and L. Feuk. 2006. Genome assembly comparison identifies structural variants in the human genome. *Nat Genet* **38:** 1413-1418.

Korbel, J.O., P.M. Kim, X. Chen, A.E. Urban, M. Snyder, and M.B. Gerstein. 2008. The current excitement about copy-number variation: how does it relate to gene duplication and protein families? *Curr Opin Struct Biol* **in press**.

Korbel, J.O., A.E. Urban, J.P. Affourtit, B. Godwin, F. Grubert, J.F. Simons, P.M. Kim, D. Palejev, N.J. Carriero, L. Du, B.E. Taillon, Z. Chen, A. Tanzer, A.C. Saunders, J. Chi, F. Yang, N.P. Carter, M.E. Hurles, S.M. Weissman, T.T. Harkins, M.B. Gerstein, M. Egholm, and M. Snyder. 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318:** 420-426.

Lee, J.A., C.M. Carvalho, and J.R. Lupski. 2007. A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell* **131:** 1235-1247.

Levy, S., G. Sutton, P.C. Ng, L. Feuk, A.L. Halpern, B.P. Walenz, N. Axelrod, J. Huang, E.F. Kirkness, G. Denisov, Y. Lin, J.R. MacDonald, A.W. Pang, M. Shago, T.B. Stockwell, A. Tsiamouri, V. Bafna, V. Bansal, S.A. Kravitz, D.A. Busam, K.Y. Beeson, T.C. McIntosh, K.A. Remington, J.F. Abril, J. Gill, J. Borman, Y.H. Rogers, M.E. Frazier, S.W. Scherer, R.L. Strausberg, and J.C. Venter. 2007. The diploid genome sequence of an individual human. *PLoS Biol* **5:** e254.

Lieber, M.R., Y. Ma, U. Pannicke, and K. Schwarz. 2003. Mechanism and regulation of human non-homologous DNA end-joining. *Nat Rev Mol Cell Biol* **4:** 712-720.

Linardopoulou, E.V., E.M. Williams, Y. Fan, C. Friedman, J.M. Young, and B.J. Trask. 2005. Human subtelomeres are hot spots of interchromosomal recombination and segmental duplication. *Nature* **437:** 94-100.

Nystrom-Lahti, M., P. Kristo, N.C. Nicolaides, S.Y. Chang, L.A. Aaltonen, A.L. Moisio, H.J. Jarvinen, J.P. Mecklin, K.W. Kinzler, B. Vogelstein, and et al. 1995. Founding mutations and Alu-mediated recombination in hereditary colon cancer. *Nat Med* **1:** 1203-1206.

Perry, G.H., J. Tchinda, S.D. McGrath, J. Zhang, S.R. Picker, A.M. Caceres, A.J. Iafrate, C. Tyler-Smith, S.W. Scherer, E.E. Eichler, A.C. Stone, and C. Lee. 2006. Hotspots for copy number variation in chimpanzees and humans. *Proc Natl Acad Sci U S A* **103:** 8006-8011.

Redon, R., S. Ishikawa, K.R. Fitch, L. Feuk, G.H. Perry, T.D. Andrews, H. Fiegler, M.H. Shapero, A.R. Carson, W. Chen, E.K. Cho, S. Dallaire, J.L. Freeman, J.R. Gonzalez, M. Gratacos, J. Huang, D. Kalaitzopoulos, D. Komura, J.R. MacDonald, C.R. Marshall, R. Mei, L. Montgomery, K. Nishimura, K. Okamura, F. Shen, M.J. Somerville, J. Tchinda, A. Valsesia, C. Woodwark, F. Yang, J. Zhang, T. Zerjal, L. Armengol, D.F. Conrad, X. Estivill, C. Tyler-Smith, N.P. Carter, H. Aburatani, C. Lee, K.W. Jones, S.W. Scherer, and M.E. Hurles. 2006. Global variation in copy number in the human genome. *Nature* **444:** 444-454.

Richardson, C., M.E. Moynahan, and M. Jasin. 1998. Double-strand break repair by interchromosomal recombination: suppression of chromosomal translocations. *Genes Dev* **12:** 3831-3842.

Sebat, J., B. Lakshmi, J. Troge, J. Alexander, J. Young, P. Lundin, S. Maner, H. Massa, M. Walker, M. Chi, N. Navin, R. Lucito, J. Healy, J. Hicks, K. Ye, A. Reiner, T.C. Gilliam, B. Trask, N. Patterson, A. Zetterberg, and M. Wigler. 2004. Large-scale copy number polymorphism in the human genome. *Science* **305:** 525-528.

Sharp, A.J., Z. Cheng, and E.E. Eichler. 2006. Structural variation of the human genome. *Annu Rev Genomics Hum Genet* **7:** 407-442.

Sharp, A.J., D.P. Locke, S.D. McGrath, Z. Cheng, J.A. Bailey, R.U. Vallente, L.M. Pertz, R.A. Clark, S. Schwartz, R. Segraves, V.V. Oseroff, D.G. Albertson, D. Pinkel, and E.E. Eichler. 2005. Segmental duplications and copy-number variation in the human genome. *Am J Hum Genet* **77:** 78-88.

The International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **437:** 1299-1320.

Tuzun, E., A.J. Sharp, J.A. Bailey, R. Kaul, V.A. Morrison, L.M. Pertz, E. Haugen, H. Hayden, D. Albertson, D. Pinkel, M.V. Olson, and E.E. Eichler. 2005. Fine-scale structural variation of the human genome. *Nat Genet* **37:** 727-732.

Ugarkovic, D. and M. Plohl. 2002. Variation in satellite DNA profiles--causes and effects. *Embo J* **21:** 5955-5959.

Zhang, Z., P. Harrison, and M. Gerstein. 2002. Identification and analysis of over 2000 ribosomal protein pseudogenes in the human genome. *Genome Res* **12:** 1466-1482.

Zhou, Y. and B. Mishra. 2005. Quantifying the mechanisms for segmental duplications in mammalian genomes by statistical analysis and modeling. *Proc Natl Acad Sci U S A* **102:** 4051-4056.

Tables

Table 1: Association of SV breakpoints with several classes of repetitive elements. The relative enrichment (global) gives the enrichment relative to the global genomic background. The local relative enrichment gives the enrichment relative to a 50kb window around the breakpoint.

Repeat Type	Frequency	Global enrichment	p-value	Local enrichment	p-value
Alu	0.09	0.94	3.24E-01	1.13	1.74E-01
SD	0.41	2.57	2.14E-07	1.17	2.64E-01
L1	0.24	1.48	1.03E-07	1.12	7.16E-02
L2	0.01	0.47	1.72E-02	0.52	2.31E-02
Microsatellite	0.03	3.91	6.74E-11	3.11	2.99E-07
LTR	0.09	1.14	1.71E-01	0.89	1.97E-01
PPgene	0.01	2.08	9.55E-02	1.66	1.98E-01
GC	0.39	0.96	7.24E-03	0.97	3.00E-02

Table 2: Newly sequenced CNV breakpoints. Most sequenced breakpoints show small homologies indicative of NHEJ. Furthermore, some breakpoints have microinsertions, which also indicate an NHEJ mechanism. Finally, some breakpoints show larger homologies, which suggest NAHR.

Chromosome	Start	End	Mechanism	Repeat
1	147600602	147986401	NAHR 272bp homology	SD
1	154793347	154795560	NAHR 19bp homology	None
1	157227979	157232826	NHEJ 4bp microhomology	None
1	208144678	208152601	NHEJ 6bp microinsertion	None
1	246118115	246124262	NAHR 14bp homology	None
2	126159721	126168302	NHEJ 4bp microhomology	None
2	146579091	146593333	NHEJ 2bp microhomology	None
2	54418997	54420978	NHEJ 3bp microinsertion	None
2	90959251	90972058	NAHR 205bp homology	Satellite
3	10201175	10203945	NHEJ 4bp microhomology	None
3	121644332	121647642	NHEJ 10bp microinsertion	None
3	188063727	188068042	NHEJ 45bp microinsertion	None
3	47465673	47468445	NHEJ 2bp microhomology	None

3	62639438	62670706	NHEJ 3bp microhomology	None
4	106926782	106936575	NAHR (repeat)	LINE/L1
4	108347263	108351179	NHEJ 11bp microinsertion	None
4	142450233	142452513	NHEJ 5bp microhomology	None
4	165024355	165039560		None
4	42457435	42464300	NAHR (repeat)	LINE/L1
4	58180961	58185488	NAHR (repeat)	LINE/L1
4	79488158	79494220	NAHR 14bp homology	None
5	10579961	10585291	NAHR 105bp homology	SINE/Alu
5	177754281	177756656	NHEJ 8bp microhomology	None
5	49471345	49476325	NAHR 303bp homology	Satellite/centr
5	57715747	57721855	NHEJ 4bp microhomology	None
5	71386	76029	NHEJ 3bp microhomology	SD
6	165644659	165652123	NHEJ 3bp microhomology	None
6	34045807	34050676	NHEJ 8bp microinsertion	None
7	113203412	113209444	NAHR 15bp homology	None
8	2116965	2122377	NHEJ 1bp microhomology	None
8	25122602	25126570	NHEJ 7bp microhomology	None
8	584397	589415	NHEJ 3bp microinsertion	None
8	73950329	73956378	NAHR 10bp homology	None
9	112516996	112519927	NHEJ 4bp microhomology	None
9	70927942	70933175	NHEJ 2bp microhomology	None
9	73446481	73449953	NHEJ 3bp microhomology	None
9	84854269	84860328	NAHR 15bp homology	None
10	114102173	114106649	NHEJ 2bp microhomology	None
10	128578838	128582206	NHEJ 10bp microinsertion	None
10	4427701	4431391	NHEJ 1bp microhomology	None
10	5627110	5677111	NHEJ 6bp microhomology	None
10	84117799	84120345	NHEJ 5bp microhomology	None
12	11075858	11142017	NAHR 170bp	SD

			homology	
12	128624266	128628228		None
12	15909933	15912931	NHEJ 1bp microinsertion	None
12	38587965	38602082	NHEJ 13bp microinsertion	None
12	55618220	55663208	NAHR (repeat)	SD
12	94757723	94760459	NAHR 11bp homology	None
13	33033730	33042822		None
13	56650541	56686865	NHEJ 3bp microhomology	None
13	71705623	71710360	NHEJ 5bp microinsertion	None
14	105282154	105397044	NHEJ 3bp microhomology	None
14	34184839	34192011	NHEJ 2bp microhomology	None
14	73076457	73108631	NAHR 256bp homology	LINE/L1
14	81568863	81573084	NHEJ 10bp microinsertion	None
15	22009161	22111478	NAHR (repeat)	LTR/ERVL
15	68808907	68814563	NAHR 14bp homology	LINE/L1
16	29167046	86811700	NAHR 264bp homology	SD
16	76929139	76942400		None
18	14542177	14558726	NHEJ 8bp microhomology	SD
18	45948971	45952385	NHEJ 4bp microinsertion	None
20	28122727	28149711	NAHR (repeat)	SD
20	42760727	42762938	NHEJ 1bp microhomology	None
20	7044793	7050847	NAHR 12bp homology	None
21	19758801	19765198		None
22	27963089	27965391	NHEJ 3bp microhomology	None

Figure Captions

Figure 1: Schematic representation of the overall analysis methodology. For the coarse grained analysis, genomic features are surveyed. First, the number of features in each genomic bin is counted. Then the overall pairwise correlation is measured (using Spearman rank correlation or Wilcoxom ranksum tests).

Figure 2: Segmental duplications are distributed according to a power-law in the human genome. As can be seen, segmental duplications follow a power-law distribution, i.e., while most regions in the genome are relatively poor in SDs, there is a small number of regions with much higher SD occurrence ($p(x) \sim x^{-0.31}$). This is indicative of a preferential attachment ("rich get richer") mechanism

Figure 3: Heatmap of associations of SDs in different sequence identity bins. SDs co-occur best with pre-existing SDs of similar age and this trend appears be stronger for older SDs. Associations are given as spearman rank correlations of number of occurrence in genomic bins. All correlations are highly significant (p-value<<0.00001)

Figure 4:A) Alu mediated NAHR and preferential attachment are two complementary mechanisms for SD formation. In Alu rich regions (>10 Alu elements per 10kb), the association of SDs and pre-existing SDs is much lower than in Alu poor regions (No Alu elements per 100kb). Associations are given as spearman rank correlations of number of occurrence in genomic bins. All correlations are highly significant (p-value<<0.00001)
B) Association of Alu elements and SDs is highest for the oldest (~40Mya old) SDs and drops significantly for recent SDs. At the same time, preference for subtelomeric regions and a presumed NHEJ mechanism rises. Associations are given as spearman rank correlations of number of occurrence in genomic bins. All correlations are highly significant (p-value<<0.00001)

Figure 5: Sequence divergence of repeat elements in the human genome. As approximate age, the sequence divergence shows a burst of Alu activity roughly 40 Mya ago, and a marked decrease afterwards. The distribution of (active) LINE elements is somewhat more even. The relative number of SDs decreases in a fashion similar to the Alu elements.

Figure 6:A) Pseudogene association with SDs. Just like Alu elements, pseudogenes co-localize very strongly with old SDs and less so with younger SDs. All correlations are highly significant (p-value<<0.00001)
B) Detailed SD junction analysis. A total of 144 SDs showed matching processed pseudogenes at both junctions, i.e. both peudogenes have the same parent gene and show high homology. When picking random genomic regions of the same size and number as SDs, no matching pseudogenes were ever found to overlap both SD junctions. When using an randomized offset of +/- 5kb to account for potential sequence biases, an average of 40 matching pseudogenes were found, but in 1000 trials, never more than 43.

C) Schematic of matching processed pseudogenes at SD junctions. The processed pseudogenes overlap matching SD junctions at both duplicated segments, making them likely candidates for having mediated NAHR.

Figure 7: A) Association of SDs and CNVs. Shown is the association of SDs (90-99% sequence identity) with "young" SDs (>99% sequence identity, left bar) and CNVs (right bar).CNVs co-localize with SDs, but much weaker than very young SDs. Associations are given as spearman rank correlations of number of occurrence in genomic bins. All correlations are highly significant (p-value<<0.00001)
B) CNV association with different human repeat elements. CNVs associate weakly with L1 elements and microsatellites, but show no association with Alu elements. C) CNV association with human repeat elements after correcting for SD content. There is almost no significant association, the observed depletion in Alu elements may be due to a preference of CNVs for subtelomeric regions. Associations are given as spearman rank correlations of number of occurrence in genomic bins. p-values of the correlations are given in the bubbles.

Figure 8: A schematic of the change of formation mechanism over the last 40 million years in the mammalian lineage.

Figure 1

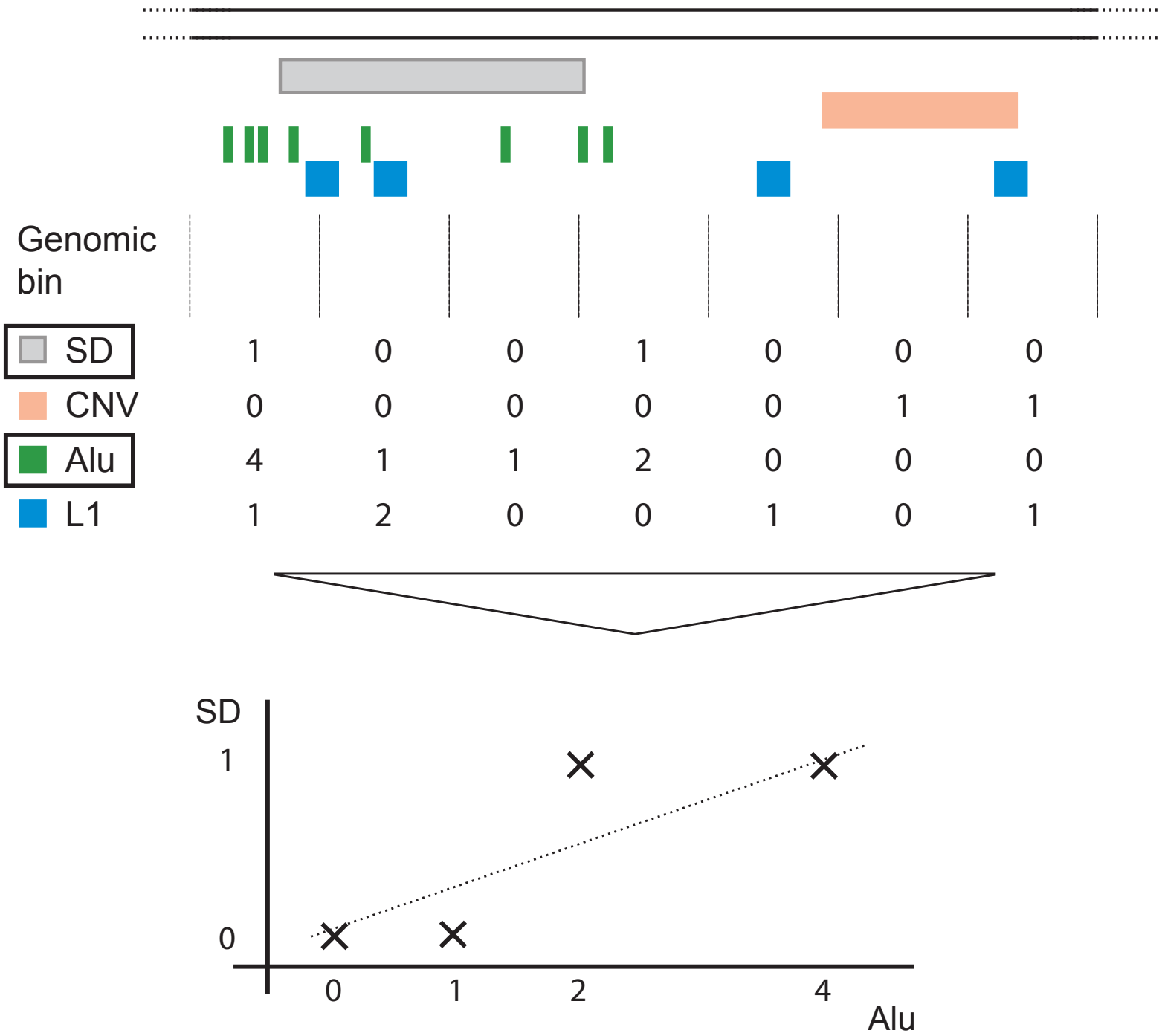

Genomic bin
SD
CNV
Alu
L1
1 0 0 1 0 0 0
0 0 0 0 0 1 1
4 1 1 2 0 0 0
1 2 0 0 1 0 1
SD
1
0
0
1
2
4
Alu
Calculate Spearman rank-correlation ρ

Figure 2

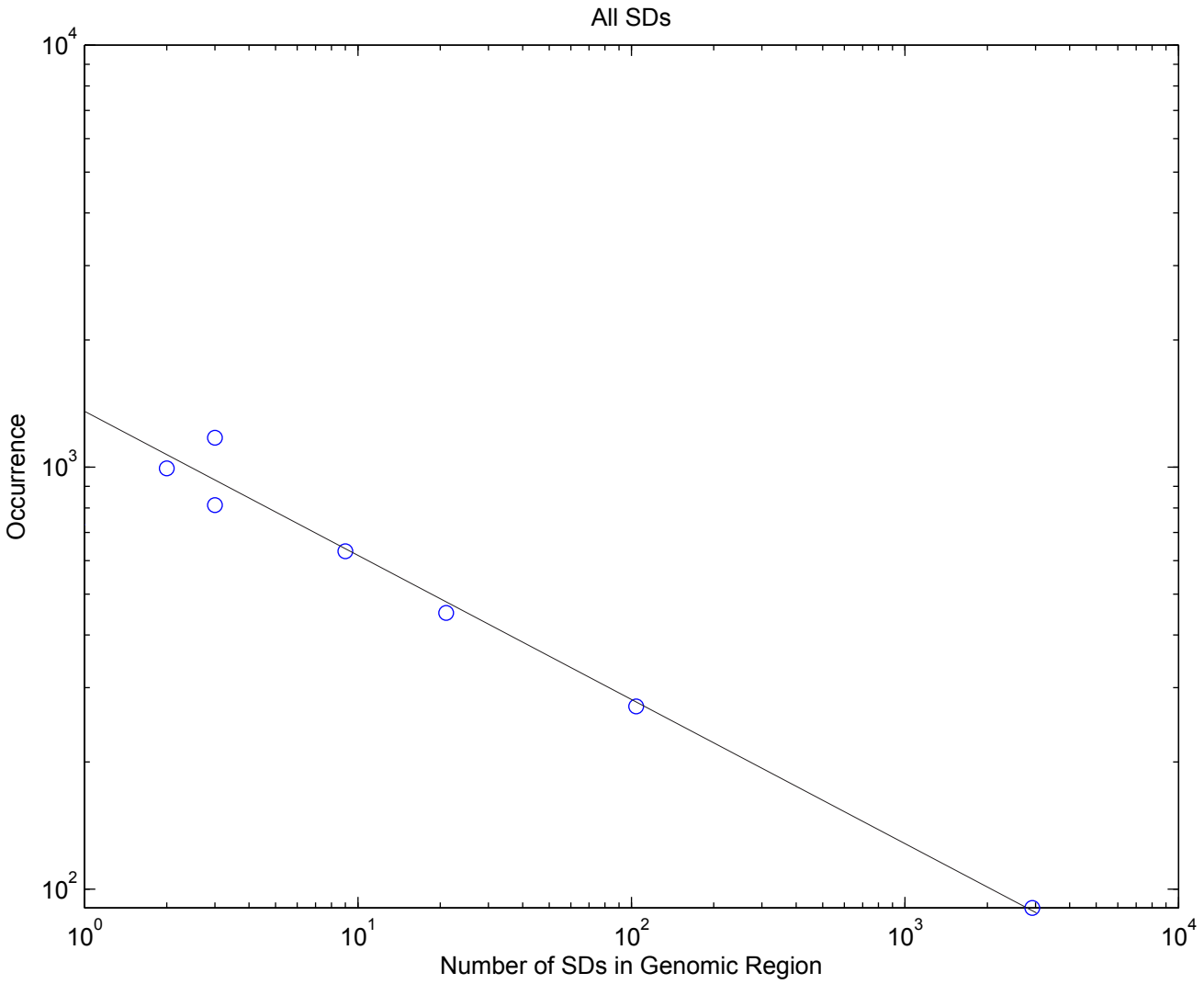


Figure 3

A

SD/SD association with SDs by age

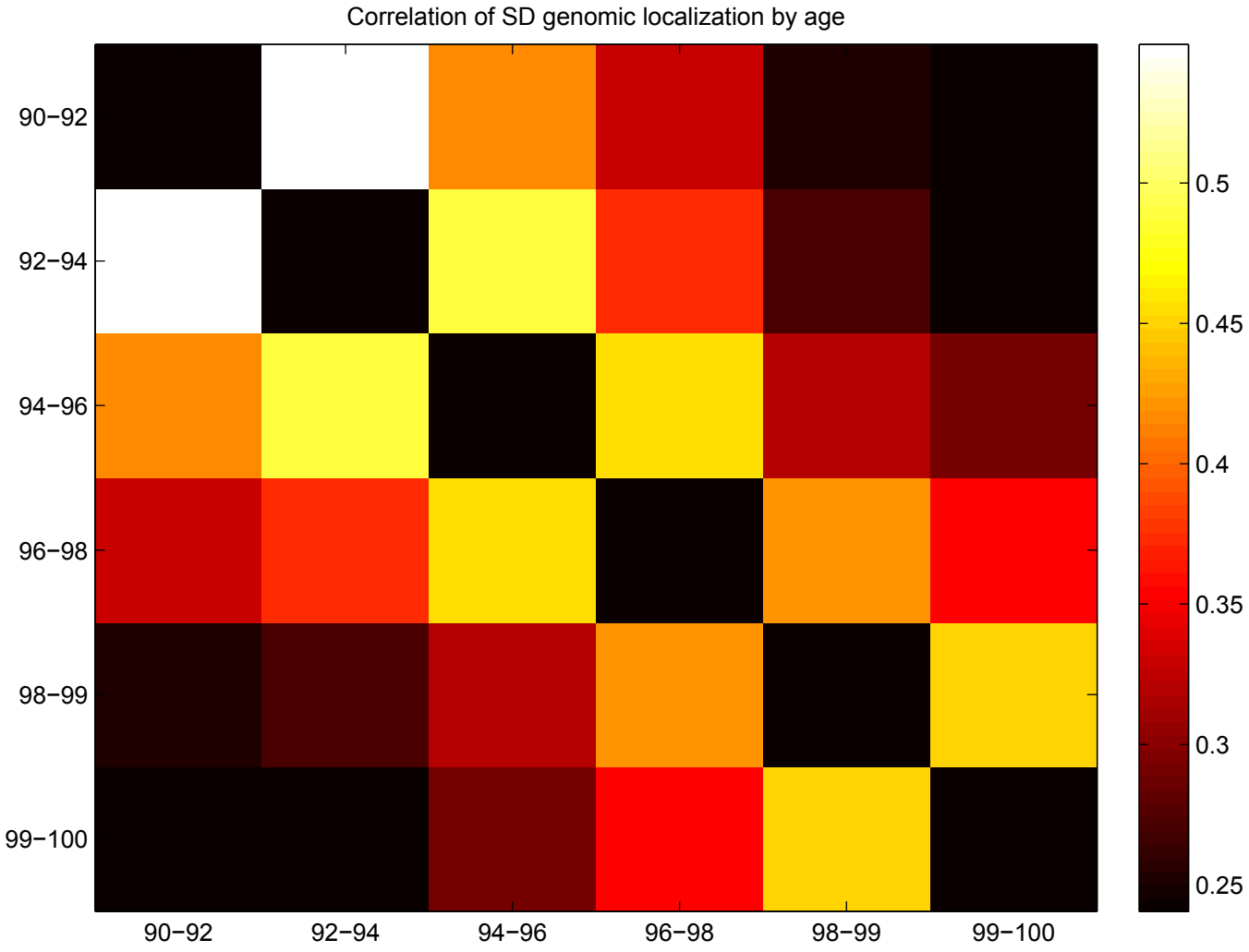


Figure 4

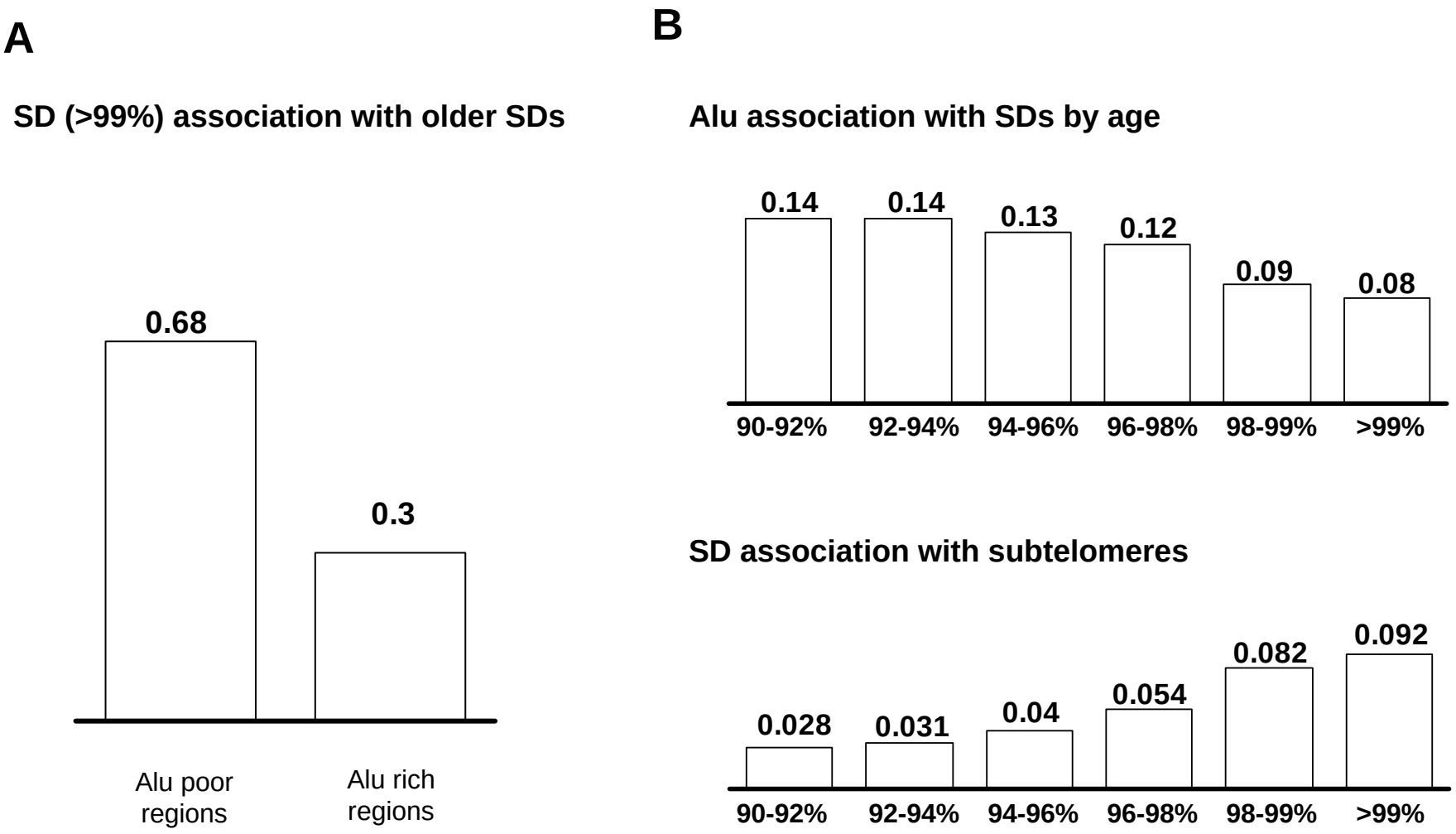


Figure 5

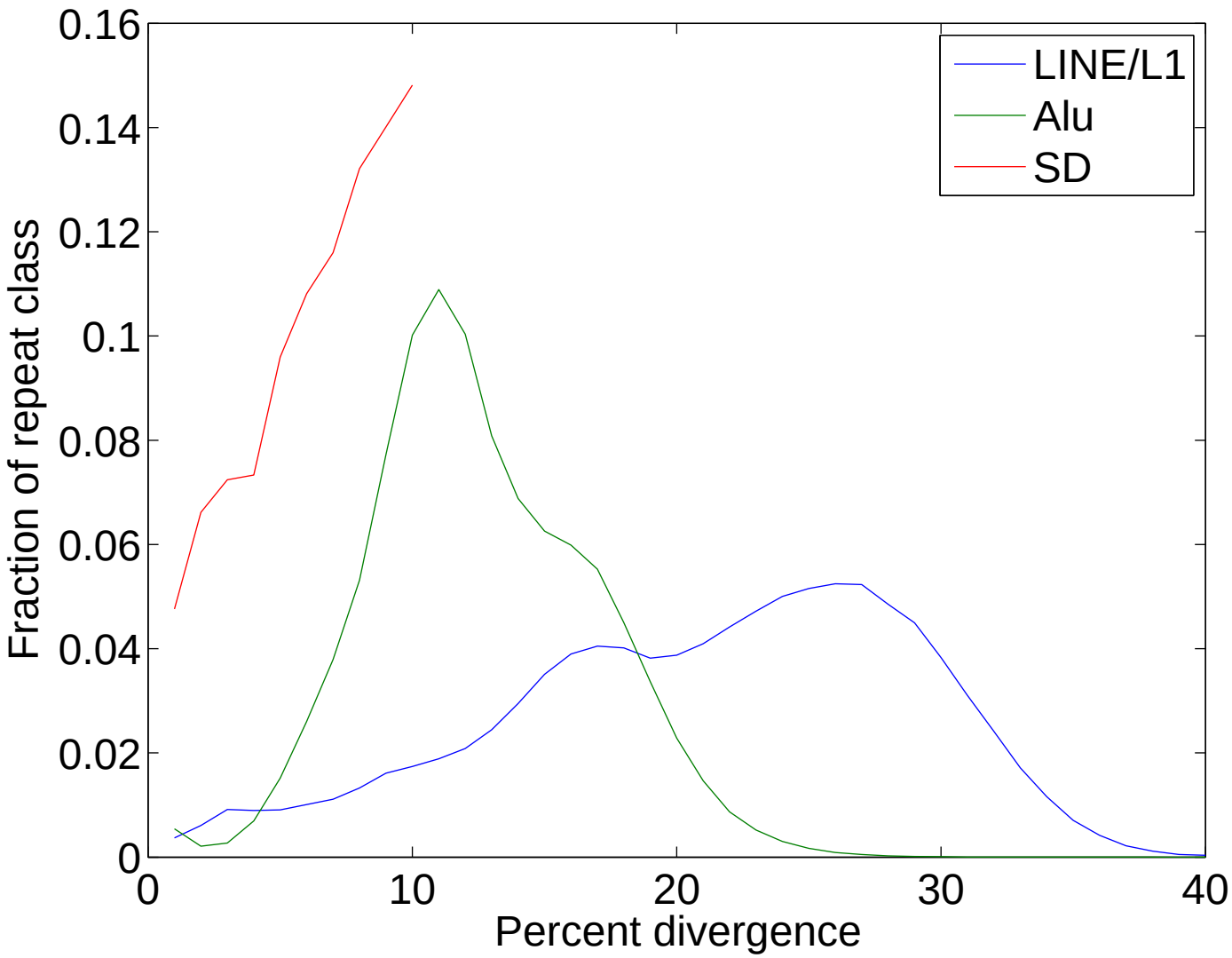
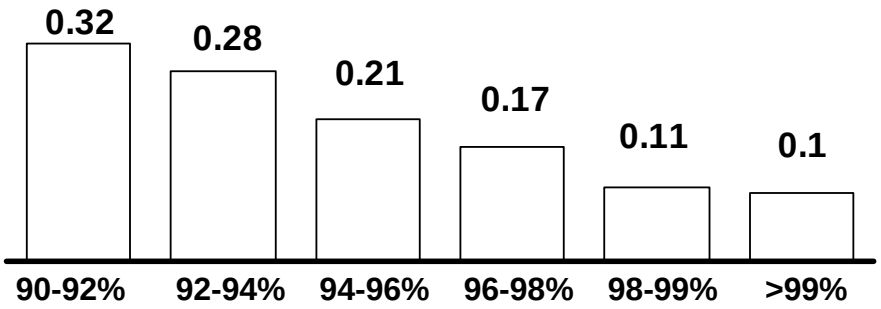


Figure 6

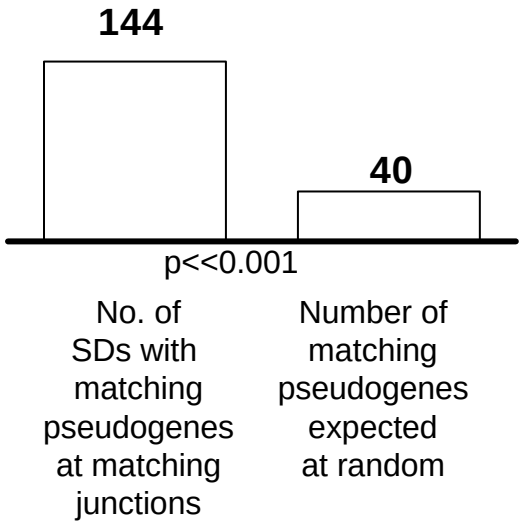
A

Processed pseudogene association with SDs by age



B

Processed pseudogenes at SD junctions



C

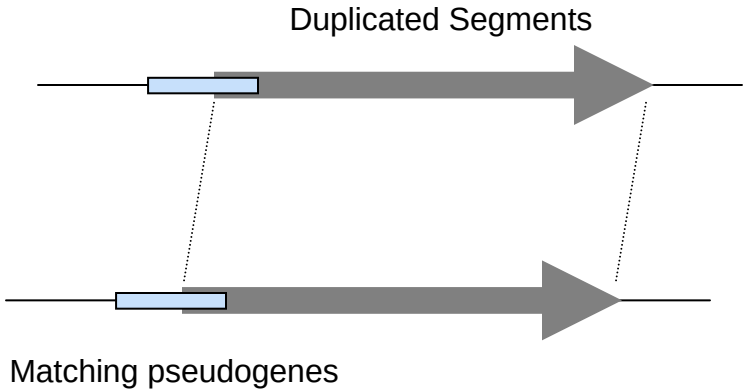


Figure 7

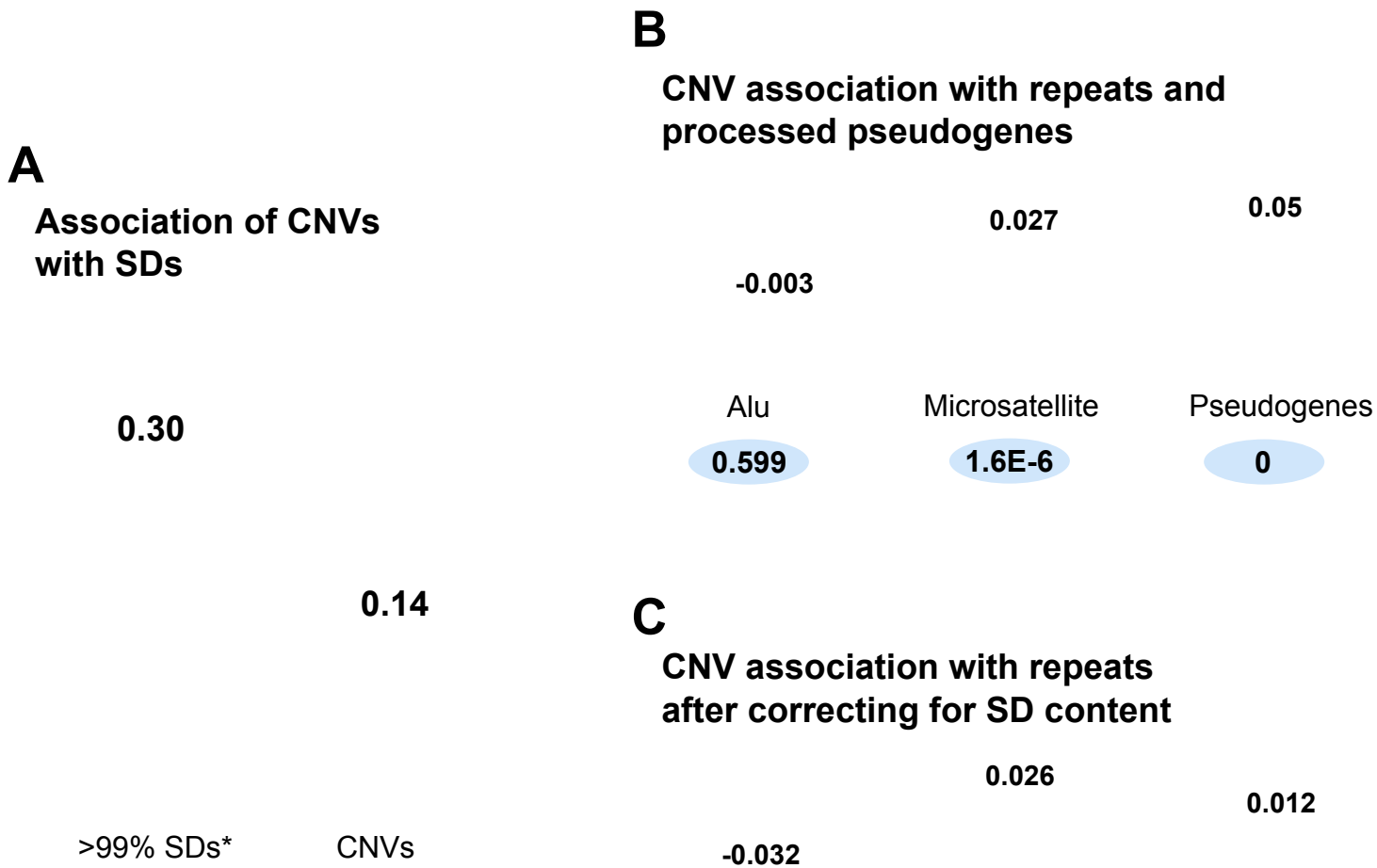

A
Association of CNVs with SDs
0.30
0.14
>99% SDs*
CNVs
B
CNV association with repeats and processed pseudogenes
0.027
0.05
-0.003
Alu
Microsatellite
Pseudogenes
0.599
1.6E-6
0
C
CNV association with repeats after correcting for SD content
0.026
0.012
-0.032
Alu
Microsatellite
Pseudogenes
2.7E-8
7.4E-6
0.039

Figure 8