Letter

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,[1,8] Hugo Y.K. Lam,[2,8] Alexander E. Urban,[3] Jan O. Korbel,[1,7] Jason Affourtit,[4] Fabian Grubert,[5] Xueying Chen,[1] Sherman Weissman,[5] Michael Snyder,[3] and Mark B. Gerstein[1,2,6,9]

[1]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520, USA; [2]Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut 06520, USA; [3]Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, Connecticut 06520, USA; [4]454 Life Sciences, Branford, Connecticut 06405, USA; [5]Department of Genetics, Yale University, New Haven, Connecticut 06520, USA; [6]Department of Computer Science, Yale University, New Haven, Connecticut 06520, USA; [7]European Molecular Biology Laboratory, 69117 Heidelberg, Germany

Segmental duplications (SDs) are operationally defined as >1 kb stretches of duplicated DNA with high sequence identity. They arise from copy number variants (CNVs) fixed in the population. To investigate the formation of SDs and CNVs, we examine their large-scale patterns of co-occurrence with different repeats. *Alu* elements, a major class of genomic repeats, had previously been identified as prime drivers of SD formation. We also observe this association; however, we find that it sharply decreases for younger SDs. Continuing this trend, we find only weak associations of CNVs with *Alu*s. Similarly, we find an association of SDs with processed pseudogenes, which is decreasing for younger SDs and absent entirely for CNVs. Next, we find that SDs are significantly co-localized with each other, resulting in a highly skewed "power-law" distribution and chromosomal hotspots. We also observe a significant association of CNVs with SDs, but find that an SD-mediated mechanism only accounts for some CNVs (<28%). Overall, our results imply that a shift in predominant formation mechanism occurred in recent history: ~40 million years ago, during the "*Alu* burst" in retrotransposition activity, non-allelic homologous recombination, first mediated by *Alu*s and then the by newly formed CNVs themselves, was the main driver of genome rearrangements; however, its relative importance has decreased markedly since then, with proportionally more events now stemming from other repeats and from non-homologous end-joining. In addition to a coarse-grained analysis, we performed targeted sequencing of 67 CNVs and then analyzed a combined set of 270 CNVs (540 breakpoints) to verify our conclusions.

[Supplemental material is available online at www.genome.org.]

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 single nucleotide polymorphisms (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 (>1 kb according to a widely accepted operational definition) regions in the genome that are either deleted or amplified on certain chromosomes ("block variation") (Iafrate et al. 2004; Sebat et al. 2004; Tuzun et al. 2005; Freeman et al. 2006; Redon et al. 2006; Korbel et al. 2007). They are known as "copy number variants" (CNVs) and are estimated to cover ~12% of the human genome, thereby accounting for a major portion of human genetic variation (Redon et al. 2006; Levy et al. 2007). Some CNVs reach fixation in the population and (if they correspond to duplications) are then visible in the genome as Segmental Duplications (SDs) (Bailey and Eichler 2006). A sizeable fraction (estimated to be 5.2%) of the human genome is covered in these SDs (Bailey et al. 2002; Bailey and Eichler 2006). These are defined as duplicated genomic regions of >1 kb 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

[8]These authors contributed equally to this work.
[9]Corresponding author.
E-mail mark.gerstein@yale.edu; fax (360) 838-7861.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.081422.108. Freely available online through the *Genome Research* Open Access option.

(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 the 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 (Freeman et al. 2006; Sharp et al. 2006; Cooper et al. 2007). However, not all duplications are thought to arise because of NAHR-based mechanisms: In subtelomeres, a separate mechanism, non-homologous end-joining (NHEJ), has been suggested for SD formation (Linardopoulou et al. 2005; Conrad and Hurles 2007). Furthermore, recent studies have uncovered a mechanism that combines both homologous and non-homologous recombination (Richardson et al. 1998; Bauters et al. 2008). Finally, a novel mechanism that involves fork stalling and template switching during replication has been proposed (Lee et al. 2007).

In this study, 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 survey are SD and CNV boundaries as well as common repeat elements, such as *Alu* and LINE retrotransposons and microsatellites. To assess colocalization of the different features, we follow a two-pronged approach: First, we bin all the features into small sequence bins of 100 kb 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

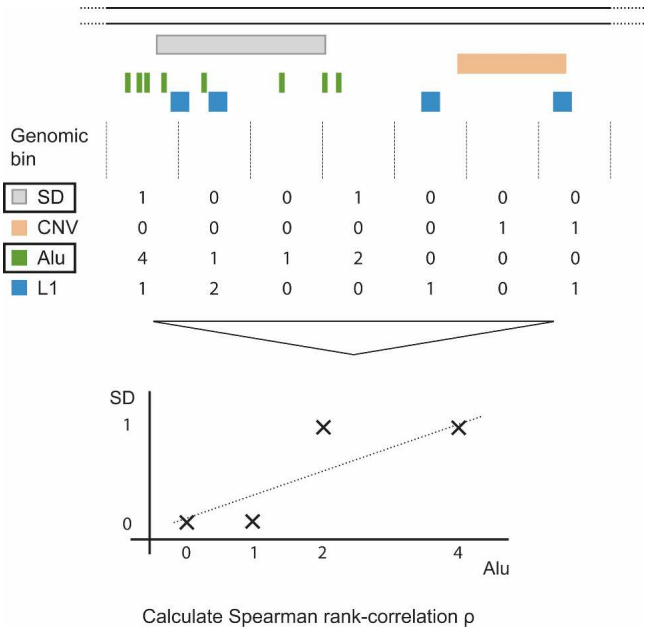

Figure 1. Schematic representation of the overall analysis methodology. For the coarse-grained analysis, genomic features are surveyed. First, the number of features in each genomic bin is counted. Then the overall pairwise correlation is measured (using Spearman rank correlation or Wilcoxon rank-sum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

is necessary to avoid problems with the comparatively low resolution of current large-scale CNV data (at best 50 kb) (Coe et al. 2007). We use the Spearman correlation as a more robust measure to detect nonlinear relationships. A high (statistically significant) correlation implies strong colocalization. We interpret statistical enrichment of colocalized 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 colocalization trends found above are due to actual differences in formation mechanisms, we examined actual breakpoints. Thus far, not many sequences of CNV breakpoints are available. Hence, we performed targeted sequencing of breakpoints, and we analyzed them in combination with a large number of previously sequenced ones. 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 (especially older ones), we find a striking enrichment of *Alu* elements and other repeats in the breakpoint regions, suggesting *Alu*-mediated formation, we find little evidence for such a mechanism in CNVs. Here, 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. 2005, 2006). 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 leads to a power-law distribution in terms of the regions (Albert and Barabasi 2002). Note, however, that 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, that is, 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 resulting in a highly skewed distribution. Hence, if SD-mediated NAHR is a major factor contributing to new SDs, we would expect the density of SDs to be distributed according to a power law throughout the human genome. Indeed, when analyzing different regions in the human genome for ends of SDs 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 mediated by pre-existing SDs. The power-law distribution is independent of SD size, age, or the binning procedure (see Methods).

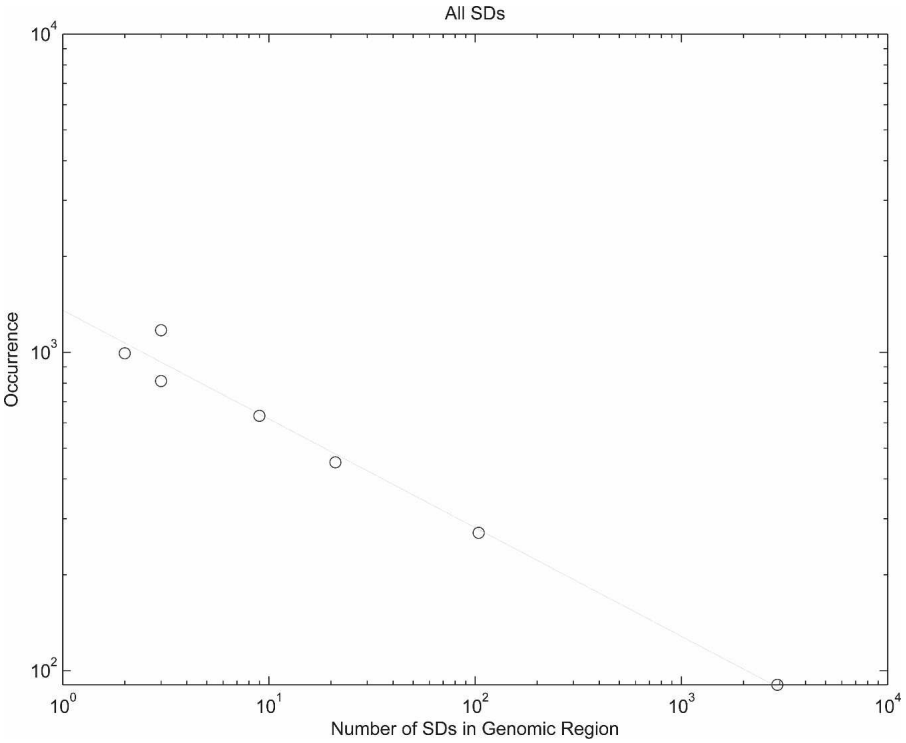

Figure 2. Segmental duplications are distributed according to a power law in the human genome. As can be seen, segmental duplications follow a power-law distribution, that is, while most regions in the genome are relatively poor in SDs, there are a small number of regions with much higher SD occurrence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

colocalization with *Alu* elements (see Fig. 4B and Supplemental Table S1), consistent with earlier reports (Zhou and Mishra 2005; Bailey and Eichler 2006). This trend is decreasing rapidly for younger SDs (see Fig. 4B), while the oldest (most divergent) SDs associate most strongly with *Alu*s. In line with this result, we find that most SDs have a sequence identity similar to *Alu* elements (90%) (see Fig. 5). The abundance of both retrotransposed elements and SDs then decreases with rising sequence identity, in sync. SDs also appear to colocalize with LINE/L1 repeats, but this association is much weaker and might be reflective of colocalization of *Alu*s and L1 repeats (Kazazian Jr. 2004). We also find evidence that *Alu*-mediated mechanisms and preferential attachment mechanisms may be complementary. That is, SDs that colocalize strongly with *Alu*s 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. It 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.

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. Hence, 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 different bins. Indeed, we observe a significant correlation between SDs in different age groups (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 Figure 2. Finally, we observe that this correlation is best for old SDs and gets successively worse as we move toward more and more recent SDs. This may be indicative of a trend of changing SD formation behavior, as we 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 SDs show highly significant

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* retro transposons, that is, they parasitize the same LINE retrotransposition machinery and are also thought to have been mostly formed during the *Alu* burst ~40 million years ago (Mya) (Zhang et al. 2002). The obvious difference is that there are 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 an 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 200 bp. Note that many pseudogenes have different parents but still show high sequence identity. While these numbers are highly significant (P-values $\ll 0.001$), they are relatively small compared to the total number of processed pseudogenes in the human genome (9747; www.pseudogene.org). One reason 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, this may have led to the formation of much fewer SDs.

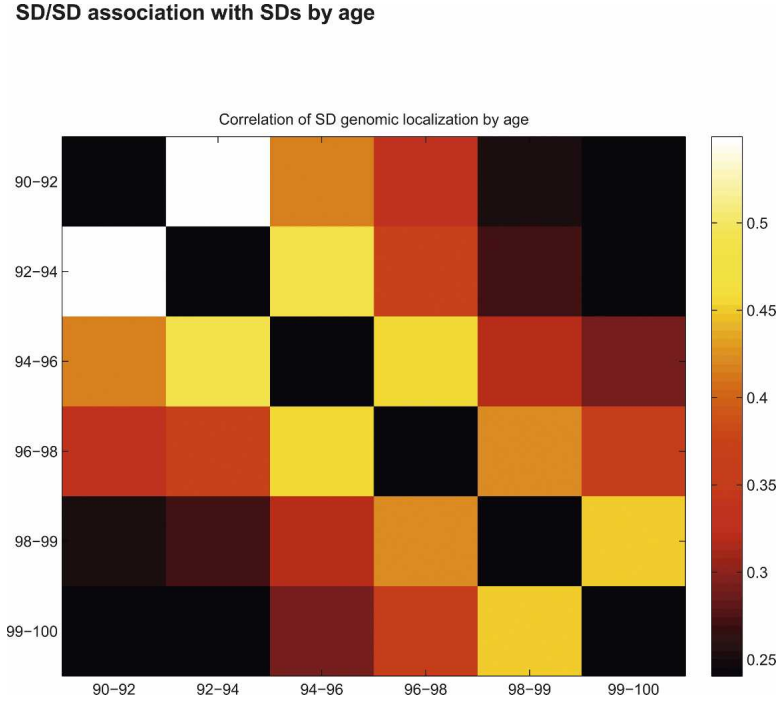

Figure 3. Heatmap of associations of SDs in different sequence identity bins. SDs co-occur best with pre-existing SDs of similar age, and this trend appears to be stronger for older SDs. Associations are given as Spearman rank correlations of the number of occurrences in genomic bins. All correlations are highly significant (*P*-value ≪ 0.00001).

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 CNVs co-occur with SDs, 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 line with this, CNVs have been viewed as the drifting, polymorphic form of SDs, that is, 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 is smaller than may be expected. We find that maximally 28% of CNVs were formed by an SD-mediated mechanism, that is, 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 0.14) 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, ~60% of "very young" SDs could be the result of NAHR mediated by older SDs. Conversely, the same can be said of only 28% of CNVs. 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 Supplemental material).

This result implies that an *Alu*-mediated mechanism is an unlikely candidate for CNV formation. It is consistent with re-

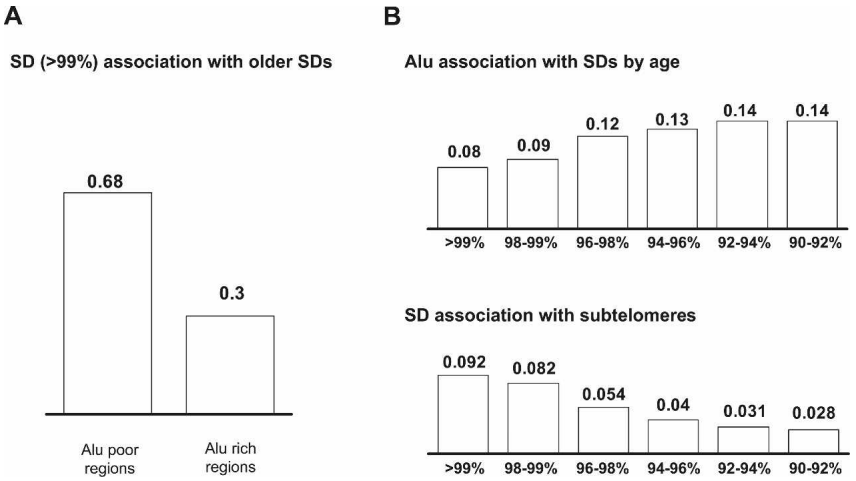

Figure 4. (*A*) *Alu*-mediated NAHR and preferential attachment are two complementary mechanisms for SD formation. In *Alu*-rich regions (>10 *Alu* elements per 10 kb), the association of SDs and pre-existing SDs is much lower than in *Alu*-poor regions (no *Alu* elements per 100 kb). Associations are given as Spearman rank correlations of the number of occurrences in genomic bins. All correlations are highly significant (*P*-value ≪ 0.00001). (*B*) Association of *Alu* elements and SDs is highest for the oldest (~40 Mya) 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 the number of occurrences in genomic bins. All correlations are highly significant (*P*-value ≪ 0.00001).

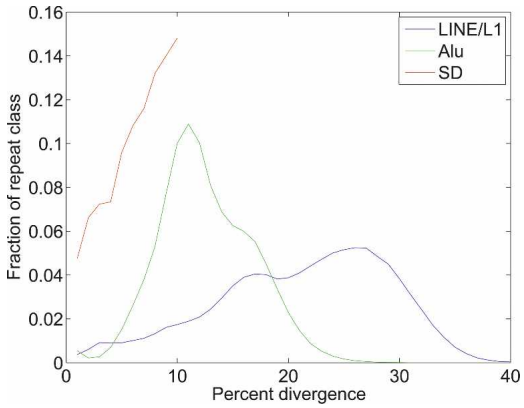

Figure 5 Sequence divergence of repeat elements in the human genome. As approximate age, the sequence divergence shows a burst of *Alu* activity ~40 Mya and a marked decrease afterward. The distribution of (active) LINE elements is somewhat more even. The relative number of SDs decreases in a fashion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

ports that *Alu*-mediated NAHR was most common during or shortly after the burst of *Alu* activity ~40 Mya 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 *Alu*s and CNVs that were determined using very different methodologies, including different kinds of microarrays and paired-end sequencing (see Supplemental 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 colocalization 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 colocalization with CNVs (see Fig. 7B,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 locations are known (based on CGH experiments). As mentioned above, most of the current data has a resolution of at best 50 kb (Coe et al. 2007). To make authoritative statements about formation signatures, one has to analyze the exact sequences surrounding the breakpoints. Therefore, we performed targeted sequencing of a number of representative CNV breakpoints and identified a total of 134 breakpoints (see Table 2). We combined this with previously sequenced breakpoints (Korbel et al. 2007) to analyze a total set of 540 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 50-kb 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. Nine percent 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 paired-end matching (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 (~2%). However, many exhibit signatures that may be indicative of non-homologous end-joining (NHEJ). Specifically, 40% of the breakpoints show the so-called

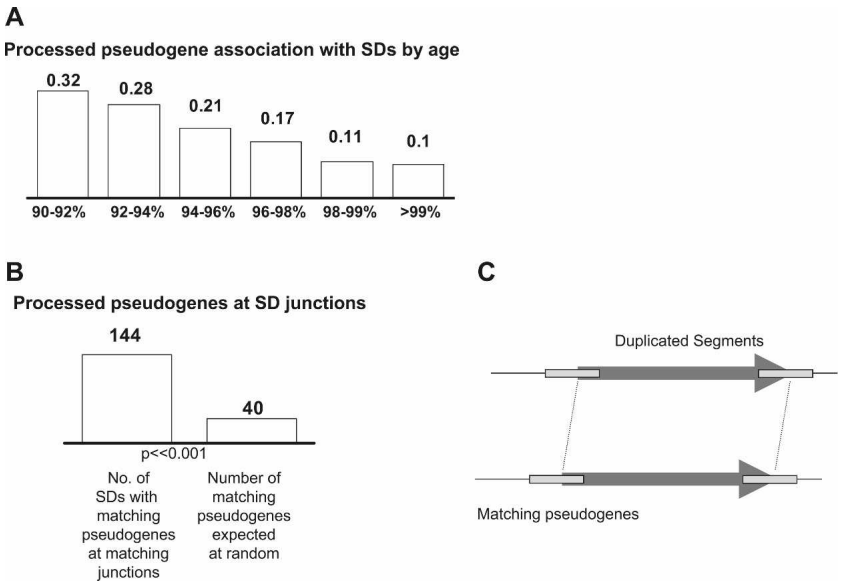

Figure 6. (*A*) Pseudogene association with SDs. Just like *Alu* elements, pseudogenes colocalize very strongly with old SDs and less so with younger SDs. All correlations are highly significant (*P*-value $\ll$ 0.00001). (*B*) Detailed SD junction analysis. A total of 144 SDs showed matching processed pseudogenes at both junctions, that is, both pseudogenes 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 a randomized offset of $\pm$5 kb 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.

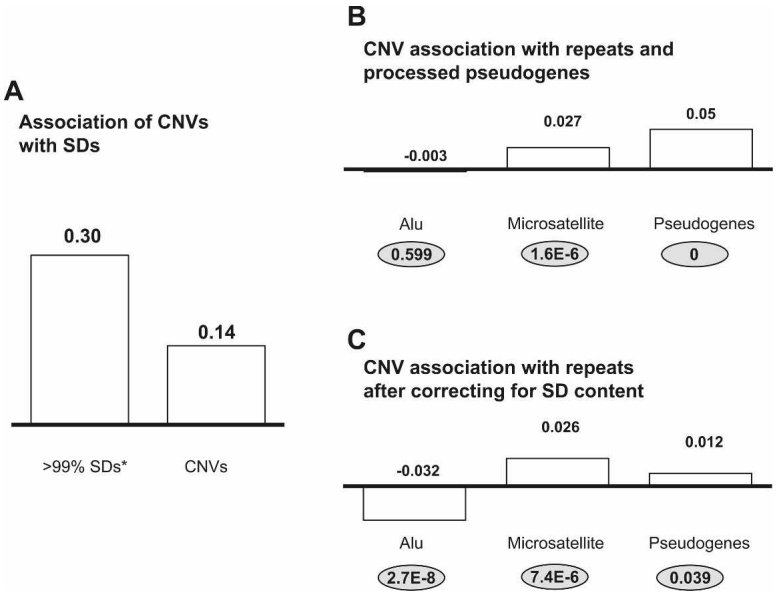

Figure 7. (*A*) Association of SDs and CNVs. Shown is the association of SDs (90%–99% sequence identity) with (*left* bar) "young" SDs (>99% sequence identity) and (*right* bar) CNVs. CNVs colocalize with SDs, but much more weakly than with very young SDs. Associations are given as Spearman rank correlations of the number of occurrences in genomic bins. All correlations are highly significant (*P*-value $\ll$ 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 the number of occurrences in genomic bins. *P*-values of the correlations are given in the bubbles.

microhomologies that can be a sign of NHEJ (Lieber et al. 2003). Another 14% exhibit microinsertions, 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). Also, 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 colocalization 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 NHEJ mediated by microhomologies (<25 bp homology), rather than a NAHR mechanism 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 (Nystrom-Lahti et al. 1995; Deininger and Batzer 1999). 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 that there is no significant enrichment of *Alu* elements near the breakpoints. 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, dimin-

Table 1. Association of SV breakpoints with several classes of repetitive elements

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
Processed pseudogene	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

The relative enrichment (global) gives the enrichment relative to the global genomic background. The local relative enrichment gives the enrichment relative to a 50-kb window around the breakpoint.

Table 2. Newly sequenced CNV breakpoints

Chromosome	Start	End	Repeat
1	147600602	147986401	SD
1	154793347	154795560	None
1	157227979	157232826	None
1	208144678	208152601	None
1	246118115	246124262	None
2	126159721	126168302	None
2	146579091	146593333	None
2	54418997	54420978	None
2	90959251	90972058	Satellite
3	10201175	10203945	None
3	121644332	121647642	None
3	188063727	188068042	None
3	47465673	47468445	None
3	62639438	62670706	None
4	106926782	106936575	LINE/L1
4	108347263	108351179	None
4	142450233	142452513	None
4	165024355	165039560	None
4	42457435	42464300	LINE/L1
4	58180961	58185488	LINE/L1
4	79488158	79494220	None
5	10579961	10585291	SINE/*Alu*
5	177754281	177756656	None
5	49471345	49476325	Satellite/centr
5	57715747	57721855	None
5	71386	76029	SD
6	165644659	165652123	None
6	34045807	34050676	None
7	113203412	113209444	None
8	2116965	2122377	None
8	25122602	25126570	None
8	584397	589415	None
8	73950329	73956378	None
9	112516996	112519927	None
9	70927942	70933175	None
9	73446481	73449953	None
9	84854269	84860328	None
10	114102173	114106649	None
10	128578838	128582206	None
10	4427701	4431391	None
10	5627110	5677111	None
10	84117799	84120345	None
12	11075858	11142017	SD
12	128624266	128628228	None
12	15909933	15912931	None
12	38587965	38602082	None
12	55618220	55663208	SD
12	94757723	94760459	None
13	33033730	33042822	None
13	56650541	56686865	None
13	71705623	71710360	None
14	105282154	105397044	None
14	34184839	34192011	None
14	73076457	73108631	LINE/L1
14	81568863	81573084	None
15	22009161	22111478	LTR/ERVL
15	68808907	68814563	LINE/L1
16	29167046	86811700	SD
16	76929139	76942400	None
18	14542177	14558726	SD
18	45948971	45952385	None
20	28122727	28149711	SD
20	42760727	42762938	None
20	7044793	7050847	None
21	19758801	19765198	None
22	27963089	27965391	None
X	92682955	92688161	None

Most sequenced breakpoints show small homologies indicative of NHEJ. Furthermore, some breakpoints have microinsertions, which also indicate a NHEJ mechanism. Finally, some breakpoints show larger homologies, which suggest NAHR.

ishing probability of *Alu*-mediated NAHR. This finding is further bolstered by the fact that most SDs have a similar sequence divergence (age) as most *Alu*s, that is, 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 colocalization. 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 SDs, and we find this trend persisting. However, SDs-mediated CNV formation can only account for a minority of the CNVs found (<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 give 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. 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 toward subtelomeric regions (P-value < 0.001), where double-strand breakage and NHEJ are 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 colocalization 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 colocalization of SDs with *Alu*s and with each other. This leads to a coherent picture: ~40 Mya, there was a peak in *Alu* activity, known as the *Alu* burst (see Fig. 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 identity (~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 Fig. 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 play a much bigger role in

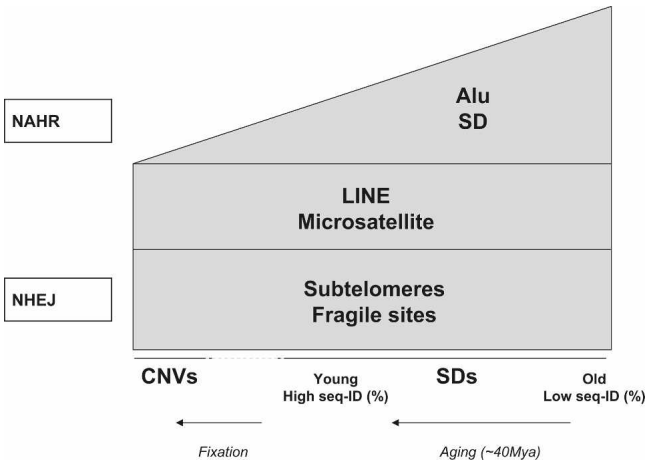

Figure 8. A schematic of the change of formation mechanism over the last 40 million years in the mammalian lineage.

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. In 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; Korbel et al. 2007; Hurles et al. 2008). As a result, 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 toward 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 the current models that propose a decrease of *Alu* activity after the "*Alu* burst" ~40 Mya. Finally, we present results suggesting that while some CNVs are formed through NAHR, a large fraction of them are formed through NHEJ. 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://eichlerlab.gs.washington.edu/database.html) based on the build 36 genome (Bailey et al. 2002). We binned all existing SDs into sequence identity categories and different size categories (see Supplemental material). To enable comparison with low-resolution copy number variation data, we finally binned all segmental duplications according to genomic coordinate. We varied the bin size from 10 kb to 1 Mb. Because the copy number variant mapping resolution is at most 50 kb for the techniques employed in the used data sets (Coe et al. 2007), we report the results for calculations with a bin size of 100 kb. Calculations using other bin sizes are reported in Supplemental Table S1. For copy number variants, we used three separate data sets, 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 data sets are reported in Supplemental Table S1.

Breakpoint sequencing

A total of 67 CNVs 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.) or by bead purification from the reaction mixture with Agencourt AMPure (Agencourt Biocience Corporation). 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; Roche Diagnostics). The resulting single-stranded DNA shotgun libraries were then sequenced with 454 Sequencing. Both the resulting reads (median length = 250 bp) 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 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 hot spot data were taken from the HapMap recombination data (The International HapMap Consortium 2005). Data for the processed pseudogenes were obtained from Pseudogene.org (Karro et al. 2007).

Computation of associations

Coarse-grained colocalization was assessed by computing the Spearman rank correlations between the binned distributions of each feature (SD occurrence, CNV occurrence or repeat occur-

rence) 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 bin size of 100 kb for the analysis, but changes in the binning procedure did not have an effect on our results (see Supplemental material). 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 a local context. In a global context, we compared the average number of a random nucleotide in the genome intersecting with a given genomic element to that of a breakpoint. Since this may be biased by local genomic context, we also calculated the average number of a random nucleotide intersecting with a given genomic element in a 50-kb 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 bp around the annotated breakpoint. We then looked for SDs where pseudogenes overlapped the SD start and end junctions 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 50-kb window around the actual junctions and calculated their overlap with matching pseudogenes.

CNV breakpoint analysis

To complement the coarse-grained approach, we analyzed a set of 540 sequenced breakpoints, a combination of the breakpoints from Korbel et al. (2007) 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 supplemental material are 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, from a Marie Curie Fellowship, and from the A.L. Williams Professorship funds.

References

Albert, R. and Barabasi, A.L. 2002. Statistical mechanics of complex networks. *Rev. Mod. Phys.* **74:** 47–97.

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

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

Bailey, J.A., Liu, G., and Eichler, E.E. 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 Albert, R. 1999. Emergence of scaling in random networks. *Science* **286:** 509–512.

Bauters, M., Van Esch, H., Friez, M.J., Boespflug-Tanguy, O., Zenker, M., Vianna-Morgante, A.M., Rosenberg, C., Ignatius, J., Raynaud, M., Hollanders, K., et al. 2008. Nonrecurrent MECP2 duplications mediated by genomic architecture-driven DNA breaks and break-induced replication repair. *Genome Res.* **18:** 847–858.

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

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

Conrad, D.F. and Hurles, M.E. 2007. The population genetics of structural variation. *Nat. Genet.* **39:** S30–S36.

Cooper, G.M., Nickerson, D.A., and Eichler, E.E. 2007. Mutational and selective effects on copy-number variants in the human genome. *Nat. Genet.* **39:** S22–S29.

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

Freeman, J.L., Perry, G.H., Feuk, L., Redon, R., McCarroll, S.A., Altshuler, D.M., Aburatani, H., Jones, K.W., Tyler-Smith, C., Hurles, M.E., et al. 2006. Copy number variation: New insights in genome diversity. *Genome Res.* **16:** 949–961.

Goidts, V., Cooper, D.N., Armengol, L., Schempp, W., Conroy, J., Estivill, X., Nowak, N., Hameister, H., and Kehrer-Sawatzki, H. 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., Dermitzakis, E.T., and Tyler-Smith, C. 2008. The functional impact of structural variation in humans. *Trends Genet.* **24:** 238–245.

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

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

Jiang, Z., Tang, H., Ventura, M., Cardone, M.F., Marques-Bonet, T., She, X., Pevzner, P.A., and Eichler, E.E. 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., Kohany, O., Pavlicek, A., Kapitonov, V.V., and Jurka, M.V. 2004. Duplication, coclustering, and selection of human *Alu* retrotransposons. *Proc. Natl. Acad. Sci.* **101:** 1268–1272.

Karro, J.E., Yan, Y., Zheng, D., Zhang, Z., Carriero, N., Cayting, P., Harrrison, P., and Gerstein, M. 2007. Pseudogene.org: A comprehensive database and comparison platform for pseudogene annotation. *Nucleic Acids Res.* **35:** D55–D60.

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

Khaja, R., Zhang, J., MacDonald, J.R., He, Y., Joseph-George, A.M., Wei, J., Rafiq, M.A., Qian, C., Shago, M., Pantano, L., et al. 2006. Genome assembly comparison identifies structural variants in the human genome. *Nat. Genet.* **38:** 1413–1418.

Korbel, J.O., Urban, A.E., Affourtit, J.P., Godwin, B., Grubert, F., Simons, J.F., Kim, P.M., Palejev, D., Carriero, N.J., Du, L., et al. 2007. Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318:** 420–426.

Korbel, J.O., Kim, P.M., Chen, X., Urban, A.E., Weissman, S., Snyder, M., and Gerstein, M.B. 2008. The current excitement about copy-number variation: How it relates to gene duplications and protein families. *Curr. Opin. Struct. Biol.* **18:** 366–374.

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

Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G., et al. 2007. The diploid genome sequence of an individual human. *PLoS Biol.* **5:** e254. doi: 10.1371/journal.pbio.0050254.

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

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

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

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

Redon, R., Ishikawa, S., Fitch, K.R., Feuk, L., Perry, G.H., Andrews, T.D., Fiegler, H., Shapero, M.H., Carson, A.R., Chen, W., et al. 2006. Global variation in copy number in the human genome. *Nature* **444:** 444–454.

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

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., Maner, S., Massa, H., Walker, M., Chi, M., et al. 2004. Large-scale copy number polymorphism in the human genome. *Science* **305:** 525–528.

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

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

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

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

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

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

Received May 27, 2008; accepted in revised form September 30, 2008.

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. 2008 18: 1865-1874 originally published online October 8, 2008
Access the most recent version at doi:10.1101/gr.081422.108