Supplement for:
Copy Number Variants and Segmental Duplications Show Different Formation Signatures

Philip M. Kim*, Hugo Y.K. Lam*, Alexander E. Urban, Jan O. Korbel, Xueying Chen,
Michael Snyder and Mark B. Gerstein

Supplement Tables

Table S1

Repeat elements and their association with SDs and CNVs in different datasets. See attached Excel table. Also variation of bin sizes.

Table S2

Association of CNVs with repeat elements according to CNV size. As can be seen, different sizes of CNV associate with different significance with repeat elements. While the >1Mb column may be misleading (since very few events were recorded here), we see as a general trend that larger CNVs tend to associate better with repeats, most notably with L1 Lines, Alu elements and SDs.

Association	<10kb	10kb-100kb	100kb-1000kb	>1Mb
Alu	-0.0036	-0.0037	0.0047	0.0015
L1 LINE	-0.0002	0.0131	0.0096	0.002
L2 LINE	0.0073	0.0008	-0.0045	0.0008
Microsatellite	-0.0018	0.0077	0.0213	0.0065
SD	0.0114	0.0511	0.0694	0.0219

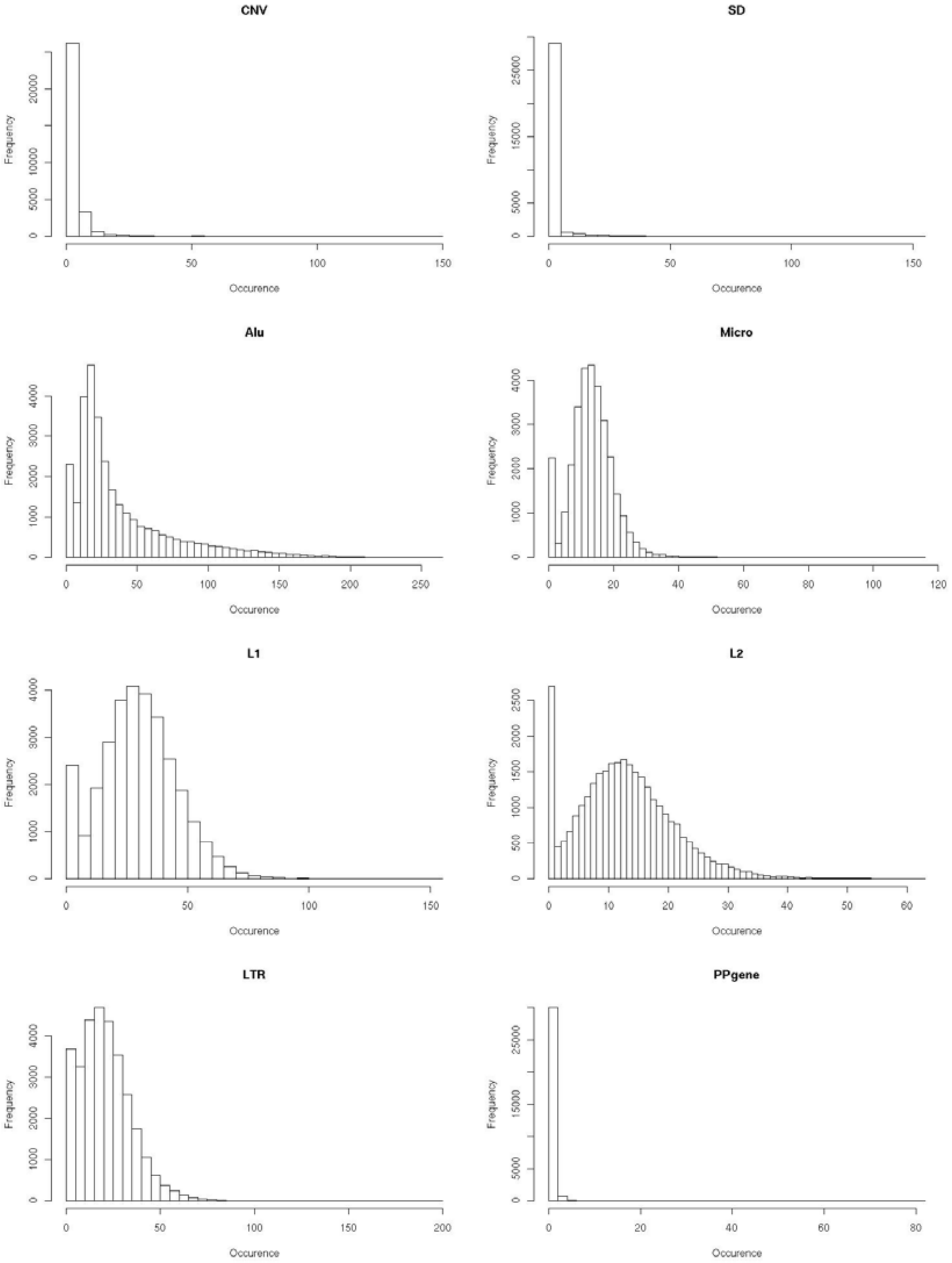
p-value	<10kb	10kb-100kb	100kb-1000kb	>1Mb
Alu	0.0465	0.0376	0.0088	0.4074
L1 LINE	0.9135	0	0	0.2704
L2 LINE	0.0001	0.6441	0.012	0.6491
Microsatellite	0.3097	0	0	0.0003
SD	0	0	0	0

Table S3
Data summary. Analyzed genomic elements and their occurrence in the human genome build36 are given.

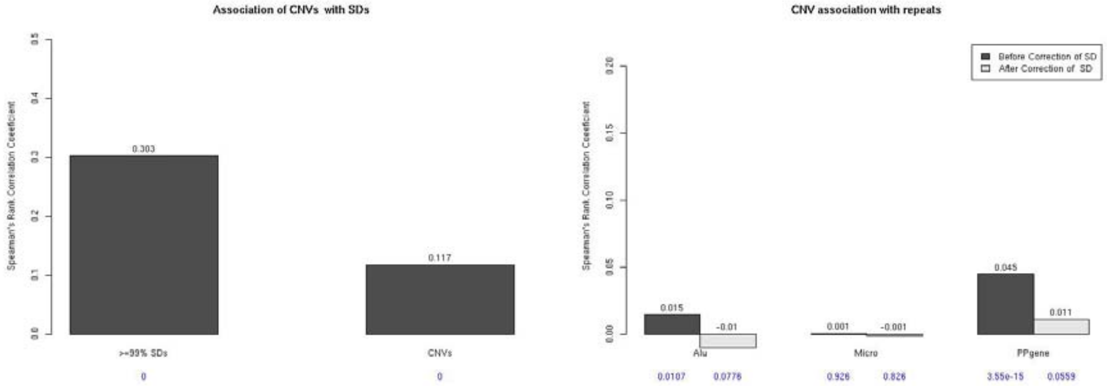
Repeat	**Count**
CNV	11,942
Redon	6,458
Cooper	4,124
Korbel	1,293
Korbel (newly Sequenced)	67
SD	51,838
Alu	1,193,407
L1 LINE	927,393
L2 LINE	409,271
LTR	656,486
Microsatellite	422,698
Processed Pseudogene	10,999

Supplement Figures:

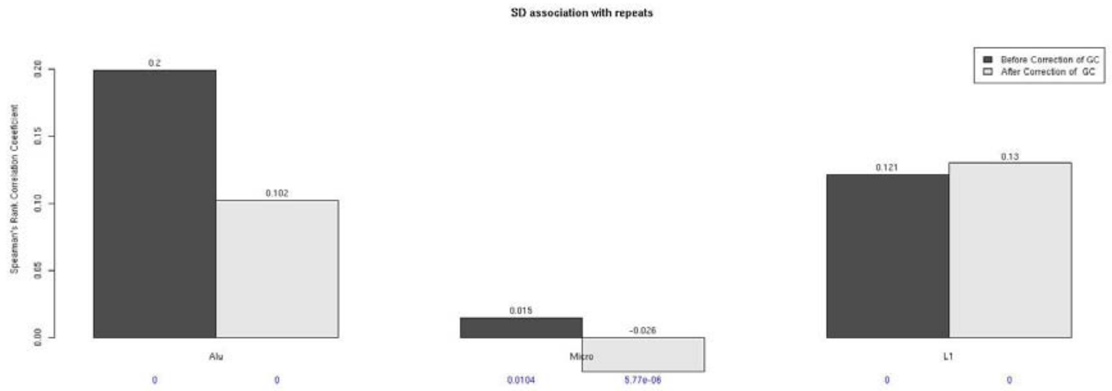
Supplement figure S1: Data summary by bin. The frequency of 100K genomic bins containing different repeat elements are given. E.g., most genomic bins (~4800, equivalent to 4.8 Gb) contain 15 to 20 Alu elements, whereas only a very small fraction contains more than 200 Alu elements.



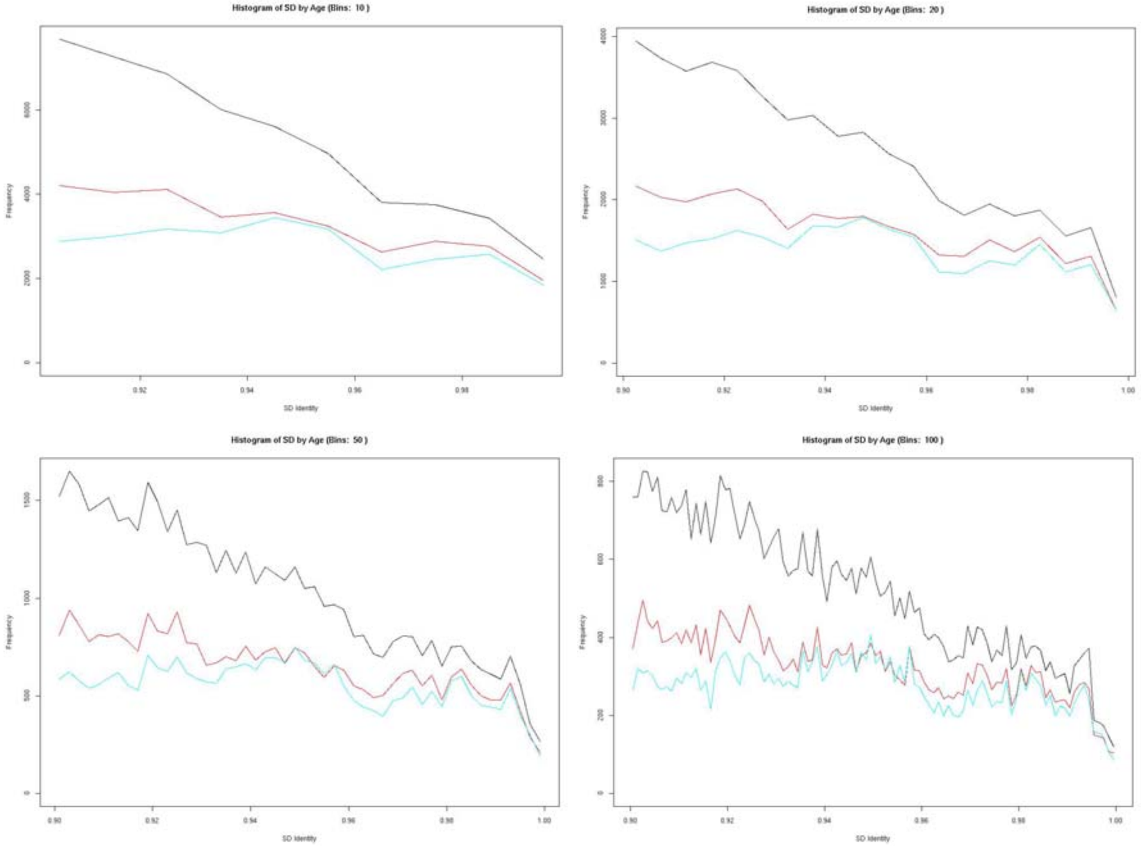
Supplement figure S2: Analysis of compiled CNV dataset (Cooper et al. 2007, Korbel et at. 2007, Redon et al. 2006).



Supplement figure S3: Association of SD and repeats with correction of GC content.



Supplement figure S4: SD binned into sequence identity categories with different number bins. (Black line: SD; Red line: SD associated with ALU; Blue line: SD associated with L1)


Histogram of SD by Age (Bins: 10)
Frequency
SD Identity
Histogram of SD by Age (Bins: 20)
Frequency
SD Identity
Histogram of SD by Age (Bins: 50)
Frequency
SD Identity
Histogram of SD by Age (Bins: 100)
Frequency
SD Identity