OPEN ACCESS Freely available online

PLoS GENETICS

Repetitive Elements May Comprise Over Two-Thirds of the Human Genome

A. P. Jason de Koning[1], Wanjun Gu[1¤], Todd A. Castoe[1], Mark A. Batzer[2], David D. Pollock[1*]

1 Department of Biochemistry and Molecular Genetics, School of Medicine, University of Colorado, Aurora, Colorado, United States of America, **2** Department of Biological Sciences, Louisiana State University, Baton Rouge, Louisiana, United States of America

Received August 5, 2011; **Accepted** October 4, 2011; **Published** December 1, 2011

Age Brauer

Bioinformatics Journal Club
27th of Feb, 2012

Repeats in eukaryotic genomes

- **Tandem repeats**
  - Microsatellites 1-6 nt
  - Minisatellites 10-60 nt
- **Transposable elements** (TE)
  - **Class-I: Retrotransposons**
    - SINEs: Alu, MIR
    - LINEs: L1, L2, CR1
    - LTRs
  - **Class-II: DNA transposons**

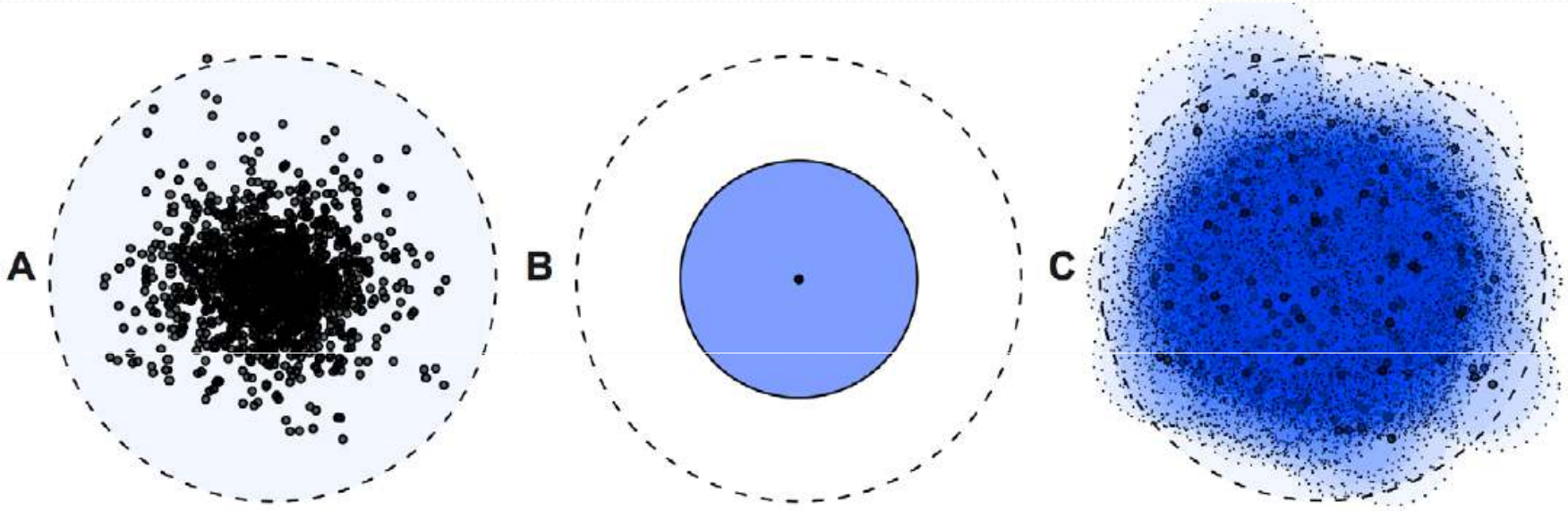


Figure 1. Principles of repeat identification using *P*-clouds. A) True data distribution representing divergence within a TE family from a master element sequence (center). B) Consensus sequence based search throws away information by collapsing observed data to a single sequence. C) *P-clouds* clusters related high-abundance oligos, thus providing better coverage of sequence space.

P-clouds

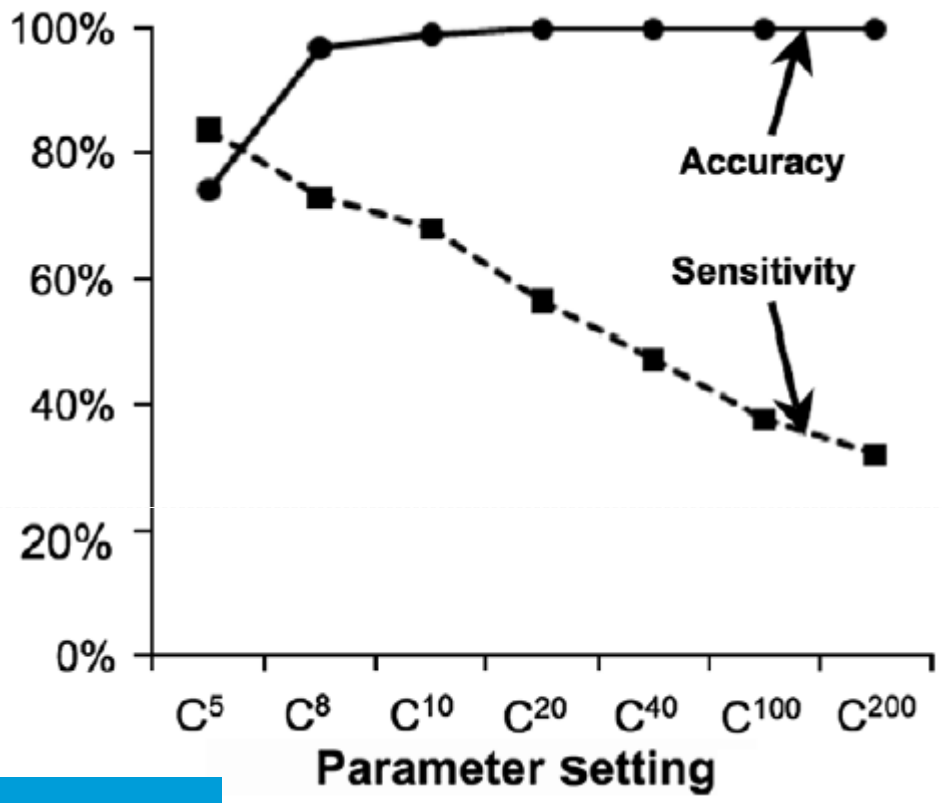
- Find the number of occurences of each specific oligo in a genome.
  - Oligo length $W = log_4(n)+1$
- The highest frequency oligo initiates a cloud.
- Similar high-frequency oligos are added to the cloud.
  - '*similar*': differences of up to 3 nt from core oligo (adjustable)
  - '*high-frequency*': adjustable

Repeat region annotation

- Oligos that were members of *P-clouds* were mapped back to the original genome sequence.
- Segments of the genome with high *P-cloud* oligo density were demarcated as “repeated regions”.
- 80% of every 10 consecutive oligos must be composed of *P-cloud* oligos (adjustable).

False positive assessment

- A simulated random non-repetitive genome sequence constrained to have the same dinucleotide frequencies in 1 Mbp windows as the original human genome.



C^5	(2, 5, 10, 100, 1000)
C^8	(2, 8, 16, 160, 1600)
C^{10}	(2, 10, 20, 200, 2000)
C^{20}	(2, 20, 40, 400, 4000)
C^{40}	(4, 40, 80, 800, 8000)
C^{100}	(10, 100, 200, 2000, 20 000)
C^{200}	(20, 200, 400, 4000, 40 000)

Gu *et al* (2008) Identification of repeat structure in large genomes using repeat probability clouds. *Analytical Biochemistry*, 380, 77-83

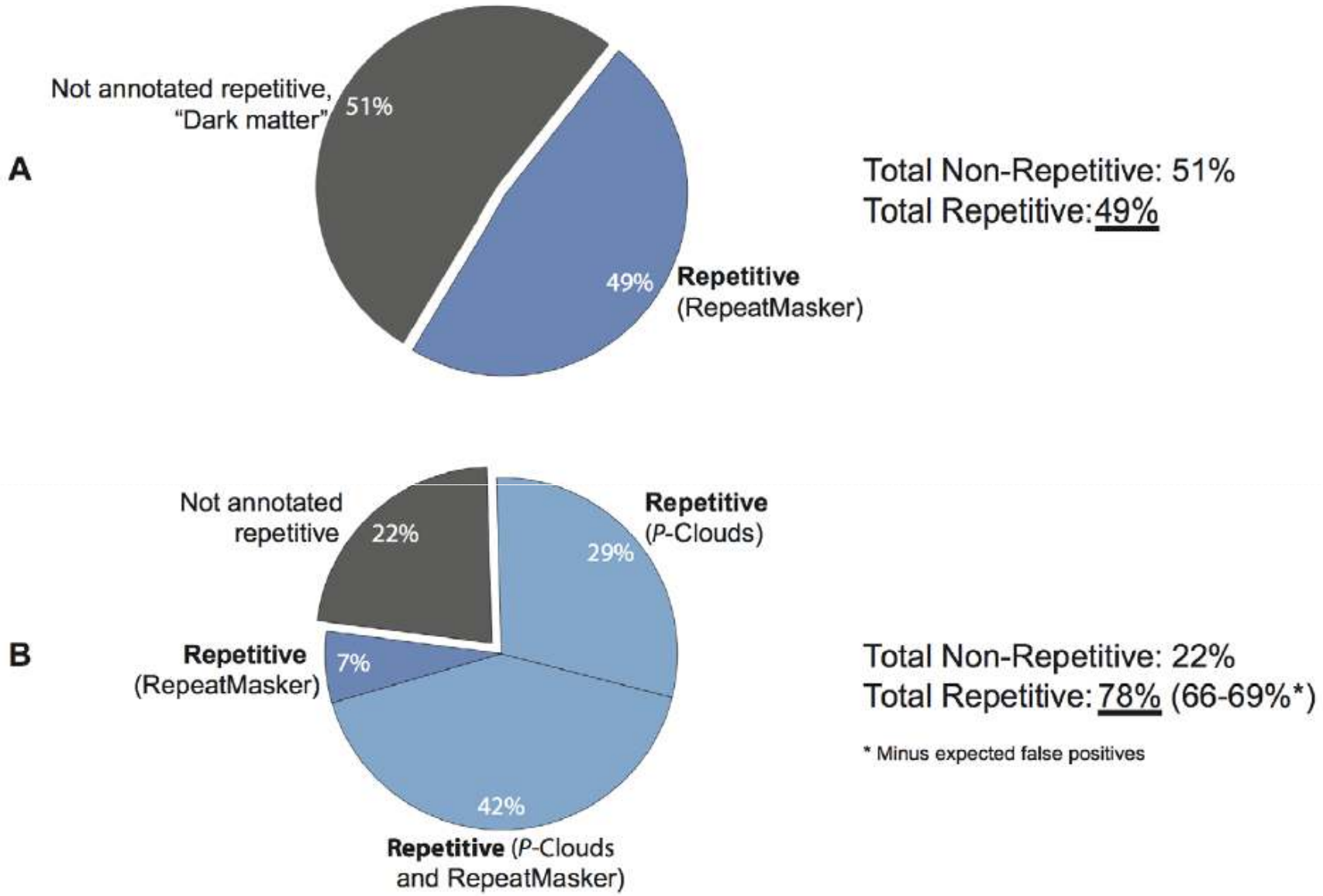


Figure 2. *P*-clouds and RepeatMasker annotation of the repeat structure of the human genome. Results are displayed as a percentage of the ungapped genome assembly length. A) Consensus results prior to this study indicate that <50% of the genome is repetitive (*RepeatMasker*). B) Analysis using *P-clouds* suggests more than two-thirds of the genome is repetitive or repeat-derived.

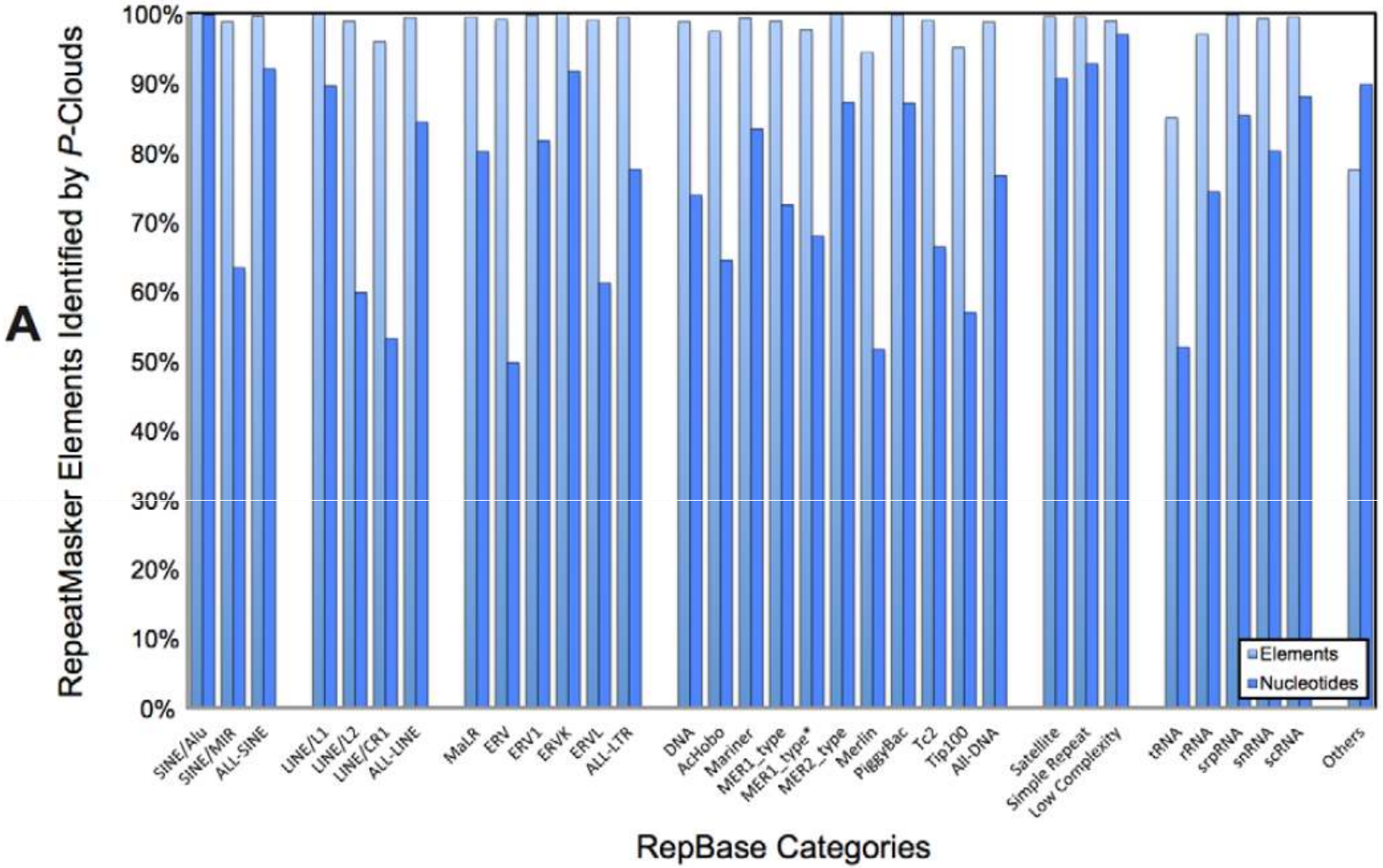


Figure 3. Percentage of previously-identified transposable elements annotated by *P-clouds*. A) The percentage of nucleotides and repeats for each family or repeat classification group. B) The number of nucleotides annotated or missed.

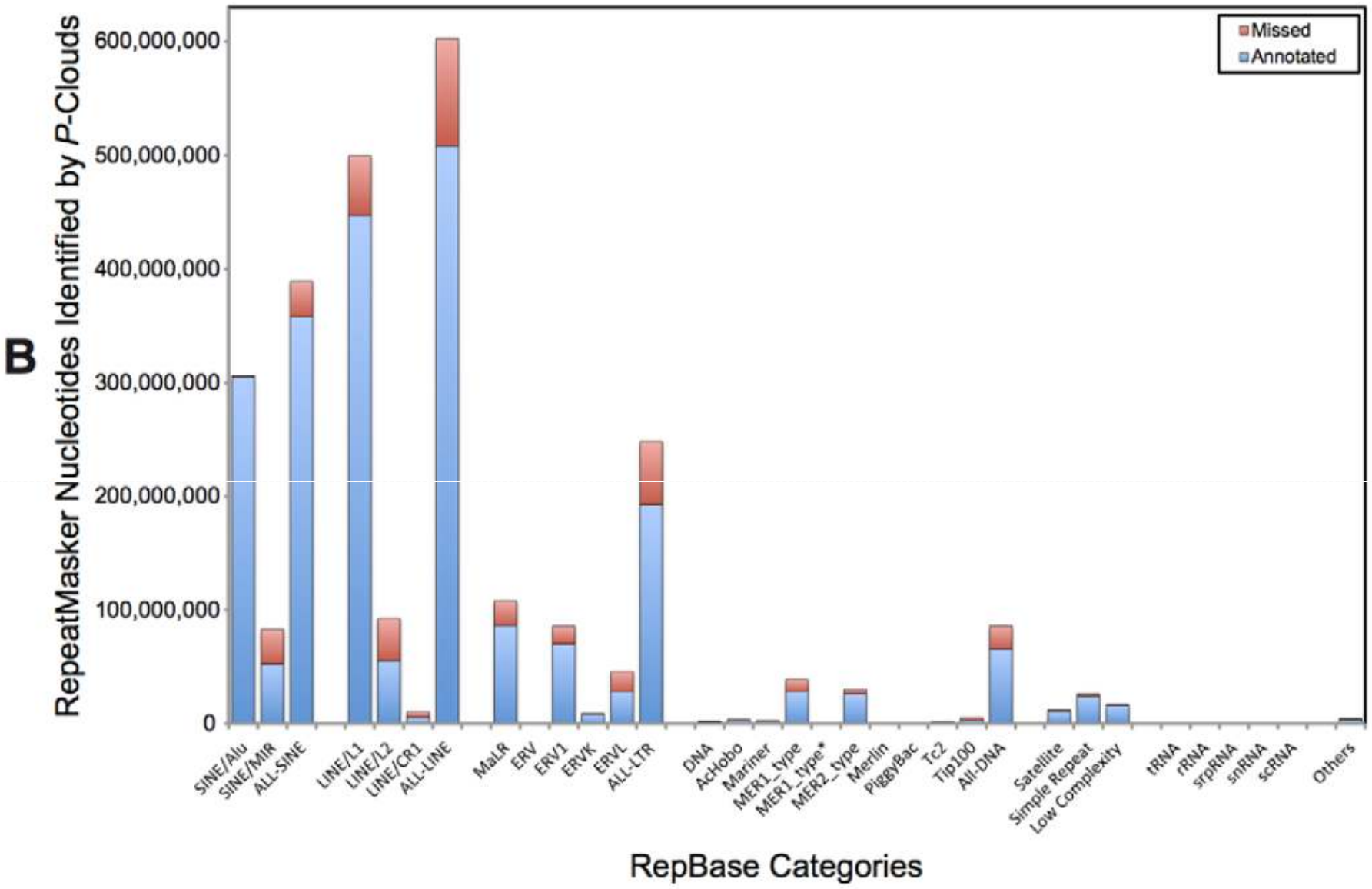


Figure 3. Percentage of previously-identified transposable elements annotated by *P-clouds*. A) The percentage of nucleotides and repeats for each family or repeat classification group. B) The number of nucleotides annotated or missed.

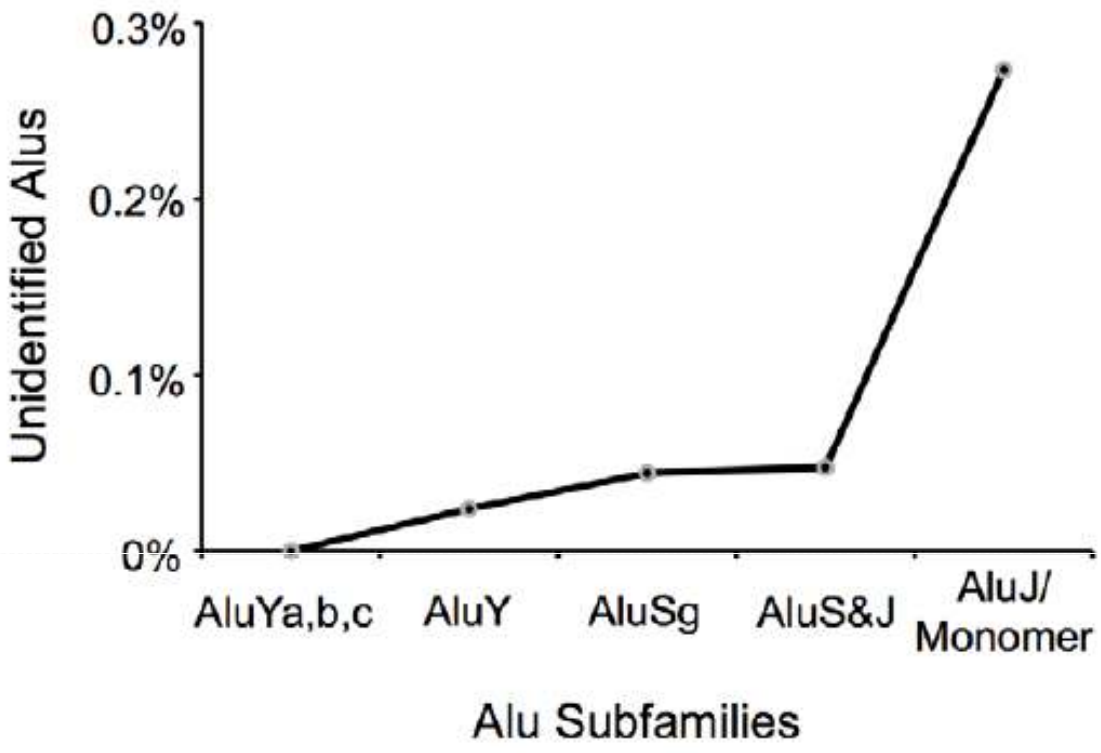


Figure 4. Percentage of Alu elements in different Alu subfamilies not annotated by *P-clouds* analysis. Displayed are elements for which no portion was annotated. The relative age of *Alu* subfamilies increases from left to right.

Table 1. Overlap between genome features and repetitive regions.

Genome Feature	Fraction of Genome	Fraction of RepeatMasker annotations	Fraction of *P-clouds* annotations	Fraction of Novel *P-clouds* annotations
Known Genes (transcribed unit)	37.48%	32.47%	36.02%	41.42%
Segmental Duplications	5.22%	5.33%	5.75%	6.02%
Duplicated Regions (WSSD)	3.53%	3.16%	3.87%	4.63%
Known Genes (exons)	1.12%	0.05%	0.56%	1.29%
Simple Repeats	1.91%	3.00%	2.36%	1.06%
CpG Islands	0.74%	0.07%	0.26%	0.56%
Pseudogenes	0.19%	0.07%	0.16%	0.28%
Total Size:	2.85 Gbp	1.39 Gbp	2.02 Gbp	0.84 Gbp

Total repetitive sequence detected by either RepeatMasker or *P-clouds* was 2.23 Gbp (out of a total 2.85 Gbp sequence in the ungapped assembly).
doi:10.1371/journal.pgen.1002384.t001

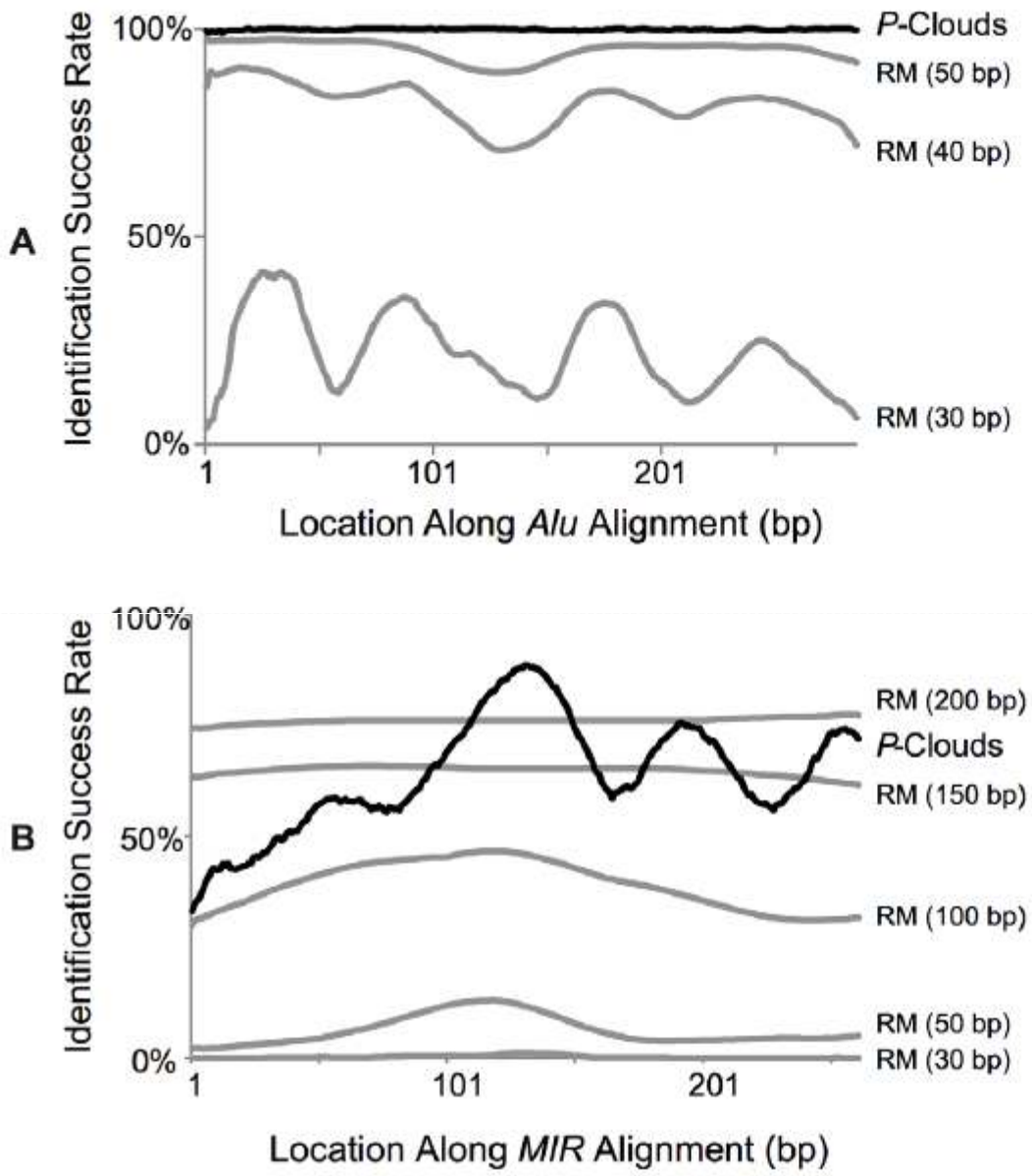


Figure 5. Percent detection success for fragments of known full-length *SINE* elements. A) *Alu* regions. B) *MIR* regions. Identification success is displayed as a running average of 10 bp starting positions.

Element specific *P-clouds* (*ESPs*)

Human genome regions that are not masked by RepeatMasker:

- 749,395 putative Alu regions
  - 20,919,291 bp (FP=22.17%)
- 7,518,362 putative MIR regions
  - 227,472,307 bp (FP=65.42%)

Relationship to other *de novo* estimates

Human chromosome 22

- RepeatMasker(RM) with RepBase 47.9%
- RepeatScout+RM 36.9%
- RepeatScout+RM with RepBase 52.5%
- RepSeek 56.2%

- P-clouds combined with RM+RepBase: 70.6%

Conclusions

- *De novo* search methods can be used to detect repeat-derived sequences that are too diverged or degraded to be easily detected by alignment to known transposable element consensus sequences.
- *P-clouds* predicts >840 Mbp of additional repetitive sequences in human genome.
- >66%-69% of the human genome sequence is repetitive or repeat-derived.
- ESPs identified ~100 Mb of previously unannotated human Alu and MIR elements.