WindowMasker: window-based masker for sequenced genomes

by Aleksandr Morgulis, E. Michael Gertz, Alejandro A. Schäffer and Richa Agarwala

National Center for Biotechnology Information, National Institutes of Health, Department of Health and Human Services, Building 38A, Room 1003N, 8600 Rockville Pike, Bethesda, MD 20894, USA

Journal Club presentation by Maido Remm
03.11.2006

Why was this work necessary?

- The problem: 25%-75% of eukaryotic genomes contain highly repeated sequence motifs
- The specific problem: Homology search (for example using **BLAST** program) returns too many false positive results.

Repeat types in human genome:

1. Tandem repeats
2. Low complexity repeats
3. Interspersed repeats
4. Duplications

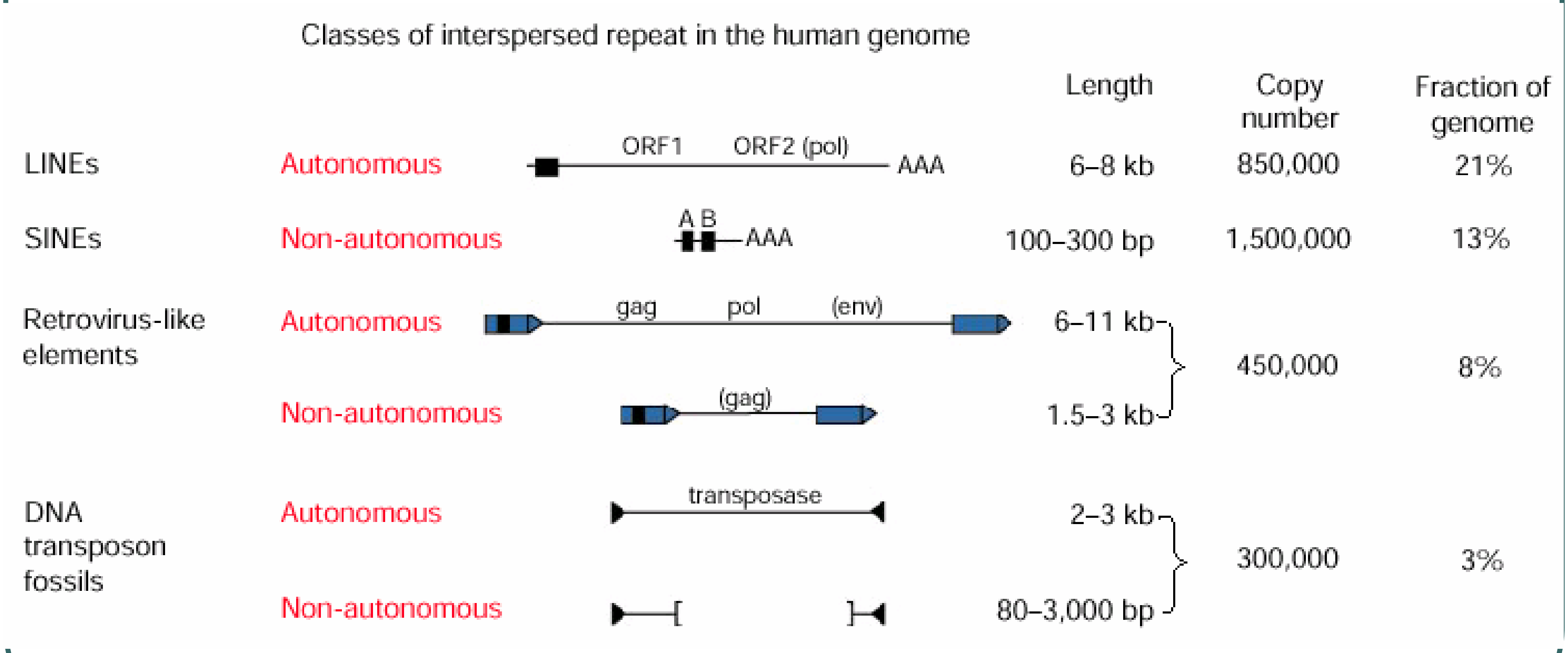
1. Tandem repeats

```
AATTTTTGTATTTTTTTTTAGAGACGGGGTTTCACCATGTTGGTCAGGCTGACTATGGAGT
TATTTTAAGGTTAATATATATAAAGGGTATGATAGAACACTTGTCATAGTTTAGAACGAA
CTAACgatagatagatagatagatagatagatagatagatagatagatagatagacagat
tgataGTTTTTTTTTATCTCACTAAATAGTCTATAGTAAACATTTAATTACCAATATTTG
GTGCAATTCTGTCAATGAGGATAAATGTGGAATCGTTATAATTCTTAAGAATATATATTC
CCTCTGAGTTTTTGATACCTCAGATTTTAAGGCC
```

2. Low-complexity regions

```
CAGGTTTAAGCTATCTTCCTACCTCAGATTCCCAAGTAGCTGGGACTATAGGCGCATACC
GCCACACACTGGCttttttttttttttttttttttttttttGTAGAGACGGGTCTCATT
GTGTTGCCCAGGATGGTCTTGAATTCCTGGGCTCAAGCAATCCTCCCACTTTGGCCTCCC
AAAGTGTTGAGACTGCAAGCATGAGCCACTGTGCTGGCCCAGAGTGACTCATAAAAAATG
GCCTTACttccctctctctcctctctccccacccacccccacctctctctcGCGCTCTGA
GGCCTCCAAAATCCTGGAGAAAACCTGCCCCTGACAAACTTCCCTCTCTGCCTTTCTGAA
CCTCGCATCTTCTCTTCTCTCAATTCTGAATGGCAAAAGCCCAAAGAACCAGCCCAAAAG
AAGAGAGCCCTGCTCAGACGGGGCCACACCCCTGCAATGGGAGGGGAAGAGTGTGGGCGA
TTCTTTCTACATAAGTTCTCTATGCAGaaaaaaaaGCATTTGAATCCAATATTCAACAGT
AAGCAAGCACTATGCTAATTAAATGACACGTCATATTGTTTACACTGATTTTGGCTGTCT
TATACTGAAAGTTATAAAATATATCAGTTAATGGGAAGAAAAGGATTATAATTGGTAAGC
TGAACATAGATTGACAACAAAGGTTATTCAGTCAAAGTagaaagaaagagagagggagag
agaAAAGTGAGCAAGCAGATGCTGTTAGAAAATGTTCCAGCAACCCCACACTTGAAGAAT
```

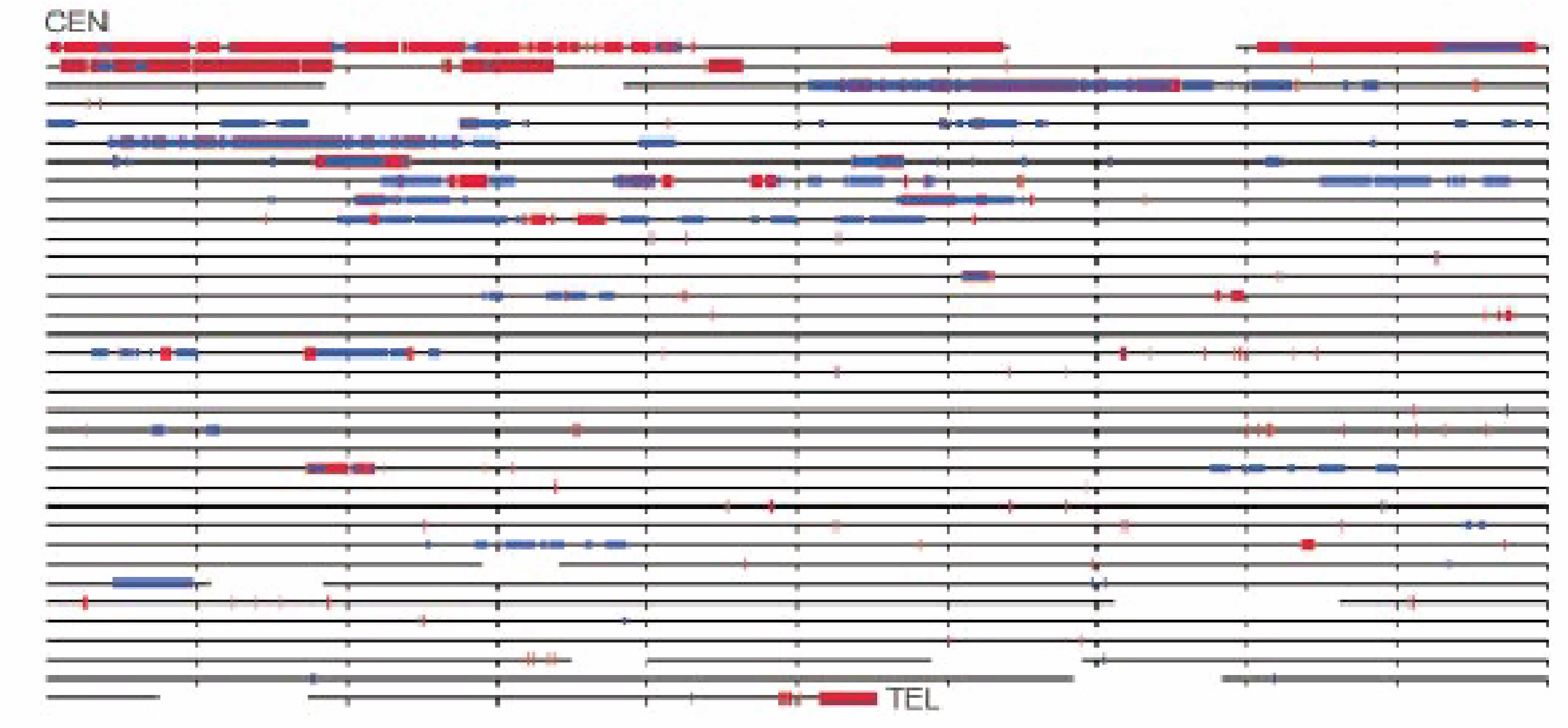
3. Interspersed repeats



Classes of interspersed repeat in the human genome

			Length	Copy number	Fraction of genome
LINEs	Autonomous	ORF1 ORF2 (pol) AAA	6–8 kb	850,000	21%
SINEs	Non-autonomous	A B AAA	100–300 bp	1,500,000	13%
Retrovirus-like elements	Autonomous	gag pol (env)	6–11 kb	450,000	8%
	Non-autonomous	(gag)	1.5–3 kb		
DNA transposon fossils	Autonomous	transposase	2–3 kb	300,000	3%
	Non-autonomous		80–3,000 bp		

4. Duplications



Duplication landscape of human chromosome 22q. Intrachromosomal (blue) and interchromosomal (red) duplications > 1kb long and with >90% nucleotide identity are shown. The duplications were found using PARASIGHT computer program (Bailey and Eichler, unpublished).

Current situation:

- Human and mouse genomes can be masked by using **REPEATMASKER** software

- It compares genome against sequence library of known repeats and masks all regions in the genome that are similar to repeats (eg. >80% identical to any known repeat)

Problems with current situation:

- **REPEATMASKER** repeat libraries are available only for limited number of genomes. Newly sequenced genomes cannot be masked by **REPEATMASKER**
- **REPEATMASKER** uses Smith-Waterman algorithm for finding gapped alignments between genome and known repeats which takes ca 1000 hours per genome

WindowMasker

Todays topics:

- Working principle of WindowMasker
- Test 1: What regions are masked?
- Test 2: What homologous matches are different in BLAST search (comparing WM and RM)?
- Test 3: What homologous matches are different in BLAST search (comparing WM and UM)?
- Test 4: How fast is WindowMasker?

WindowMasker: working principle

WindowMasker finds and masks all repeats from given genome in two passes. First pass finds repeats, second pass masks repeats.

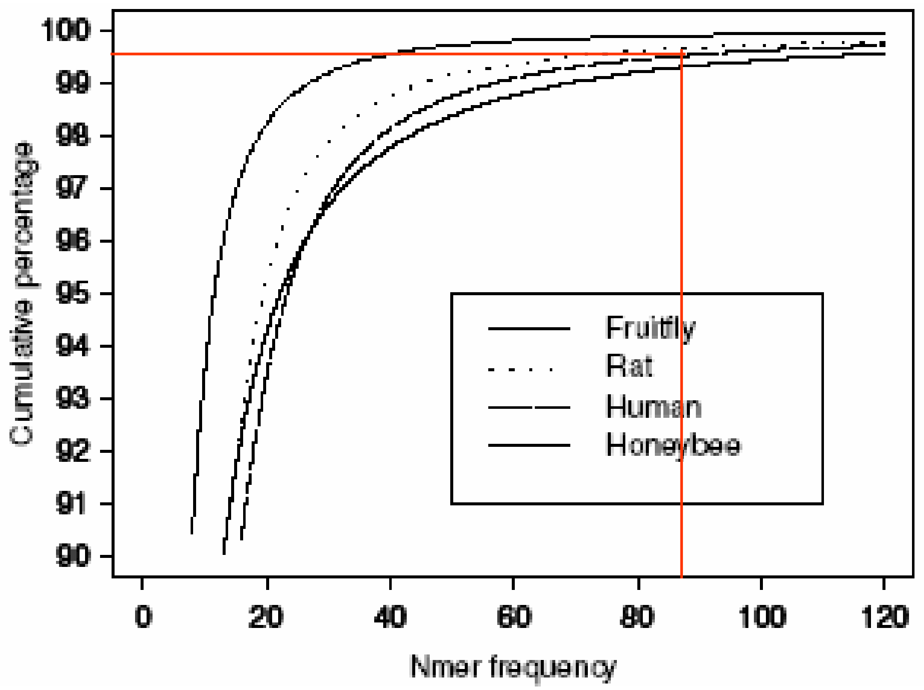
Repeat detection is based on number of occurrences of N-mer windows. N is fixed for any given genome.

WM was optimized for human genome and then tested on mouse, rat, 2 fruitflies, honeybee genomes.

WindowMasker: repeat detection

1) Determine N based on genome length L
   N is smallest integer which satisfies: $L/(4^N) < 5$
2) Scan the genome to determine frequency of all N-mer oligonucleotides S
3) Sort N-mers with freq(S)>0 by their frequency *freq(S)*
4) Find values of the following percentiles:
   T_{low} (90.0%), T_{high} (99.8%), $T_{threshold}$ (99.5%), T_{extend} (99.0%)
5) Recalculate all frequencies using T_{low} and T_{high}:
   $freq(S) = T_{high}$ *if* $(freq(S) > T_{high})$
   $freq(S) = T_{low}/2$ *if* $(freq(S) < T_{low})$

WindowMasker: repeat detection



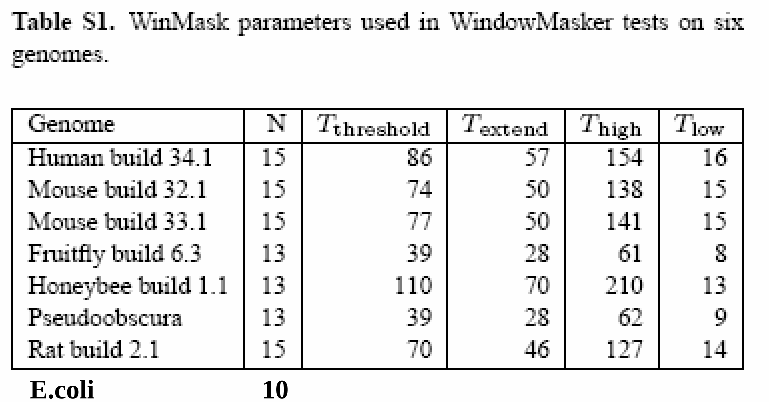


Table S1. WinMask parameters used in WindowMasker tests on six genomes.

Genome	N	$T_{\text{threshold}}$	T_{extend}	T_{high}	T_{low}
Human build 34.1	15	86	57	154	16
Mouse build 32.1	15	74	50	138	15
Mouse build 33.1	15	77	50	141	15
Fruitfly build 6.3	13	39	28	61	8
Honeybee build 1.1	13	110	70	210	13
Pseudoobscura	13	39	28	62	9
Rat build 2.1	15	70	46	127	14
E.coli	**10**				

$T_{\text{threshold}}$ is 99.5% cumulative **percentage of different N-mers** in genome

WindowMasker: repeat masking

1. Scan all windows W with length N+4 in the genomic sequence. This can also be subsequence from given genome.

2. Assign score to each window
   score(W) = int(average score of composite N-mers)

3. Mask all nucleotides in window if
   $score(W) >= T_{threshold}$

4. Mask any interval between two consecutive windows if **every base is in window** that has score
   $score(W) >= T_{extend}$ (99.0% percentile of N-mers)

WindowMasker: Test 1

Which regions are masked differently between RepeatMasker and WindowMasker?

At their default settings RepeatMasker masks 48% and WindowMasker 37% of the human genome. The overlapping part was ca 30% of the genome.

Two sets of sequences were used for comparing their masking:
R1. The 50 longest contigious regions that were masked by WM but not with RM **(RM-/WM+ regions)**

R2. The 50 longest contigious regions that were masked by RM but not with WM **(RM+/WM- regions)**

WindowMasker: Test 1

Why are some regions masked differently between RepeatMasker and WindowMasker?

Match type	Total	Number of BLASTN matches in human genome		
		1–10	11–50	>50
(RM-/WM+): Large TR pattern	44	0	1	**43**
(RM-/WM+): No TR pattern	6	4	0	**2**
(RM+/WM-): No TR pattern	50	**50**	0	0

WindowMasker: Test 2

Which matches are different between BLAST homology searches from genomes masked by RepeatMasker or WindowMasker?

Two sets of sequences were used for comparing their masking:

R3. The matches from MegaBLAST search that were retrieved from RM-masked genome, but not from WM-masked genome **(RM matches)**

R4. The matches from MegaBLAST search that were retrieved from WM-masked genome, but not from RM-masked genome **(WM matches)**

MegaBLAST search was run with 300 sample query sequences (0.5, 10 and 100 kb sizes). Match is >92 bp long and >95% identical to query.

WindowMasker: Test 2

Why are some BLAST matches different if genome is masked with RepeatMasker or WindowMasker?

Match type	Total	Number of BLASTN matches in human genome		
		1–10	11–50	>50
(RM matches): Large TR pattern	7	0	0	**7**
(RM matches): No TR pattern	75	6	7	**62**
(WM matches): Large TR pattern	2	**0**	2	0
(WM matches): No TR pattern	63	**49**	14	0

WindowMasker: Test 3

Which matches are different between BLAST homology searches from genomes masked by WindowMasker (WM) and unmasked genomes (UM)?

One set of sequences were used for illustrating the effect of masking:

R5. The matches from MegaBLAST search that were retrieved from RM-masked genome, but not from WM-masked genome **(UM matches)**

MegaBLAST search was run with 300 sample query sequences (0.5, 10 and 100 kb sizes). Match is >92 bp long and >95% identical to query.

WindowMasker: Test 3

Why are some BLAST matches different if genome is masked with RepeatMasker or WindowMasker?

Match type	Total	Number of BLASTN matches in human genome		
		1–10	**11–50**	**>50**
Honeybee genome				
(UM matches): Small TR pattern	49			
(UM matches): Large TR pattern	111	2	2	107
(UM matches): No TR pattern	134	95	39	0
Rat genome				
(UM matches): Small TR pattern	251			
(UM matches): Large TR pattern	33	0	0	33
(UM matches): No TR pattern	4189	2334	442	1413

WindowMasker: Test 4

How long does it take to mask human genome?

RepeatMasker: **1045 hours** CPU time
WindowMasker: **11 hours** CPU time

1.2 hours for first pass
(counting N-mer frequencies)

9.8 hours for second pass
(masking the genome sequence)

Conclusions:

1. WindowMasker is a fast and efficient program for masking repeats in novel genomes?

2. It does not provide well–understandable interpretation of masked regions (because different thresholds are used in masking)

3. It provides reference for citation of DUST program