Table S1 Primers used in primer extension

Gene	ORF	Sequence(5'→3')	5' Start position (related to ATG)
BGL2	*YGR282C*	AGTAGCTGCAGTAGCGAGTGTAG	+36
EGD2	*YHR193C*	TGGCGTTTTCTGGGATAGCAGAC	+25
ERG20	*YJL167W*	CTTTTTCTGAAGCCATTTCTACG	+16
COF1	*YLL050C*	GTCATTGAAAGCGGTAAGGGATTC	+54
YPT7	*YML001W*	GAAGGGGTTTCTATATGGATAAG	-21
HOR7	*YMR251W-A*	GGAAACAACAACTTGAGATAACTTC	+27
TPI1	*YDR050C*	CCACCGACAAAGAAAGTTCTAGCCAT	+26
EGD1	*YPL037C*	GCAGACAACTTTTGTAGCTTAGC	+47
TEF2	*YBR118W*	GACTTCTCTTTACCCATGTTTAGTTA	+17
TDH3	*YGR192C*	CTCTAACCATTTTGTTTGTTTATGTGTG	+10
GCN4	*YEL009C*	TTCGCTAGTGAAACTGATGGGC	-485
SBA1	*YKL117W*	CCTTTGAGCCCATGCAACTTGAGG	+45

Table S2 Non-coding RNA gene verification

RNA	Tag Sequence	Tag Occurrence	Anotated TSS	TSS in this study
RUF3	ACCCAAAAACATCAAGA	1	0	0
snR31	AAGCAAAATTACACC	1	0	1
snR33	AATGCCCTCTTTGTAC	1	0	6
snR37	AACATCAGTAGTGGTT	1	0	0
snR4	ATCAACTATTACAGTC	1	0	-3
snR43	CACTCCTGTTTCTGCCT	2	0	0
snR51	TTATTATGATGATTTTT	1	0	-1
snR61	CTACAATGATGATAAAA	1	0	0
SRG1	AATGCCTTTGTTTGGC	6	0	1

The 5' SAGE tags were mapped to 9 non-coding RNA genes by comparing the tag 5' position to the annotated start position, which is denoted as 0, of the corresponding RNA gene. Most TSS are very close to the annotated RNA 5' end.

Table S3 uORF-containing *S. cerevisiae* genes.

ORF	Gene	TSS[a]	uORF#	uORF sequences
YAR073W	*IMD1*	-106	1	MCN*[b]
YBR073W	*RDH54*	-399	5	MDVA* MDVLPTSTSGKGIFTRSR* MHQENKK* MTVIFD* MWLEIWMSCLLLLLGKAFLLDRVNICIKKIKNKTRRAKKKKENLL*
YBR158W	*AMN1*	-263	1	MC*
YCL007C		-346	4	MSFKTLPRDMGSNYGEDQALKWPVYTF* MPLYFYSYLFLCILILLIFLILRKRHLA* MEKIKL* MARLHILIQP*
YDL125C	*HNT1*	-174	4	MCVRLESGEHYEVKGQSARLPDF* M* MK* MAGPFSSEEVTALLKHCCTIV*
YDR005C	*MAF1*	-122	1	MRKI*
YDR524C	*AGE1*	-403	6	MEY* MSF* M* MFIFFFVVFSAN* MLVFFPIAVVIIVLDEERRKVKI* MFHQ*
YER053C	*PIC2*	-98	1	MTHC*
YER130C		-95	1	MIFCFEFVQ*
YER159C	*BUR6*	-192	2	MH* MNLRKKGQCTEK*
YGL059W		-23	1	MDK*
YHR071W	*PCL5*	-362	2	MHSSVF* MRAIEDTGNNRIY*
YIL094C	*LYS12*	-65	1	MRLLY*
YKL186C	*MTR2*	-258	5	MLL* MIDHDSLYVLQRLPLLTKKS* MTPCMYCRDCHY* MLRLY* MRKRIEQVTNLFL*
YKL194C	*MST1*	-49	1	MQ*
YML106W	*URA5*	-308	1	MKDTQKN*
YMR009W		-447	4	MIHMCASRIVCLDSEMPSK* MHIEKHQSIKEQRGWHLQIKKNNNHQTRH* MQK* MKVISASFLCIDD*
YMR040W		-157	3	M* MYSWPSLRKASRLEVNKYAECSLEHELCE* MNYVNRKILSKISKSS*
YNL182C	*IPI3*	-60	1	MR*
YNR034W-A		-199	1	M*
YOR031W	*CRS5*	-24	1	MSNSTQ*
YOR051C		-459	3	MRL* MFTIHF* MKAGVTSPFLFPFHEVVAEKQFSAYIVLRPKEVKKKVREKELFFSSFEMLHILKNFNIN*
YOR222W	*ODC2*[c]	-459	8	MRHG* MLLFYARQIPLELFFFQFFRGELFLPVYLERSCC* M* MVDQLVLLR* MK* MPGKFLWNFFFFSFFVVSFFCLSISKDLVVDDSVQSRSLSTRHAFASNSVIKRSLRFFSLKLFCTFLLG* MLTCVYCFRRSAFIKCKDNETWLTN* MTQCSRDHFRLDMHLQVILL*
YPR165W	*RHO1*	-289	1	MF*

a, The TSS site was determined by most 5' multiple occurrence unitag of corresponding gene.

b, uORFs are translated in 3 forward frames. "*" indicates translational stop codon.

c, *ODC2* was predicted as containing separate RNA coding transcripts in its 5' UTR (see Results section)

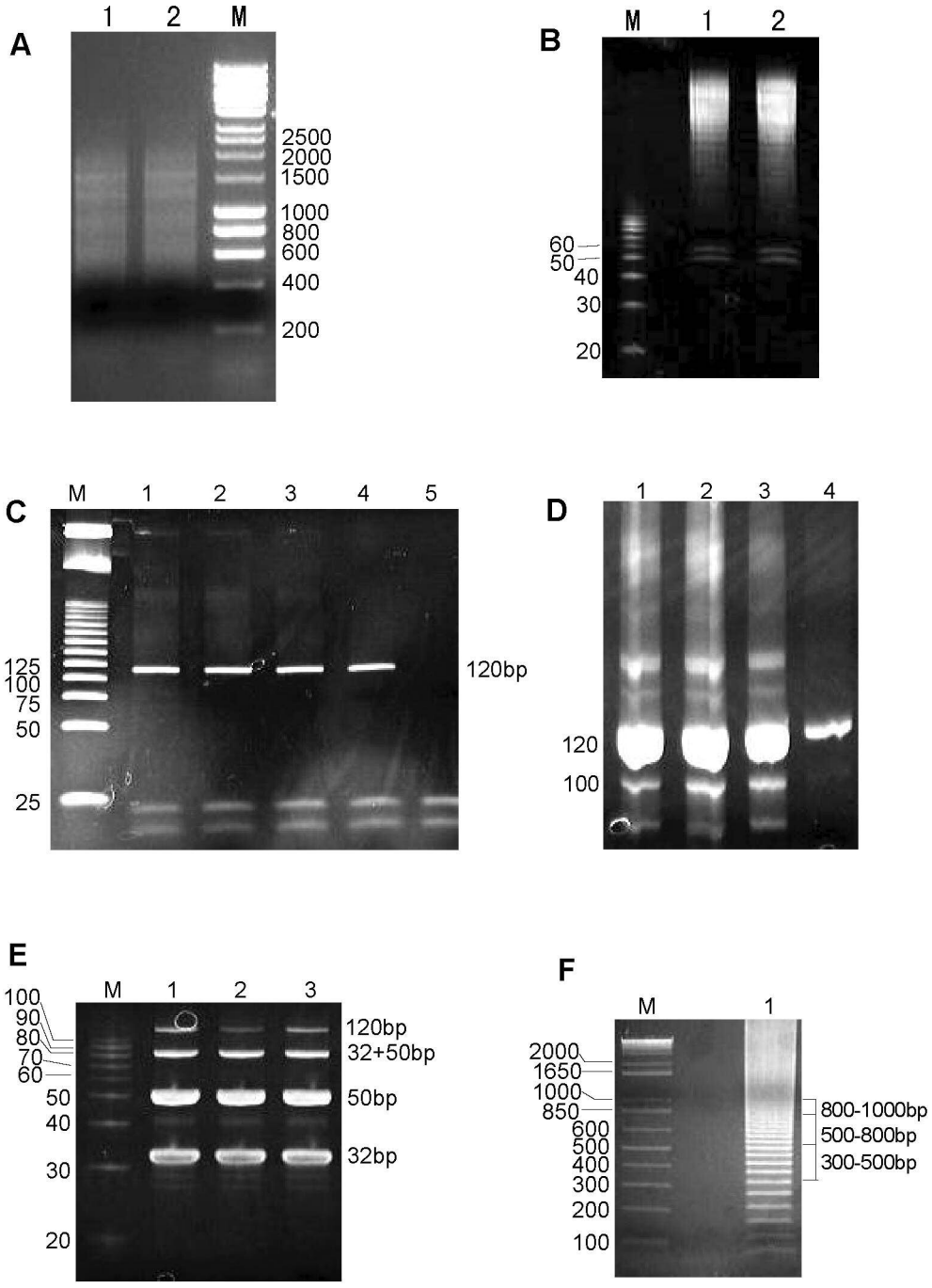
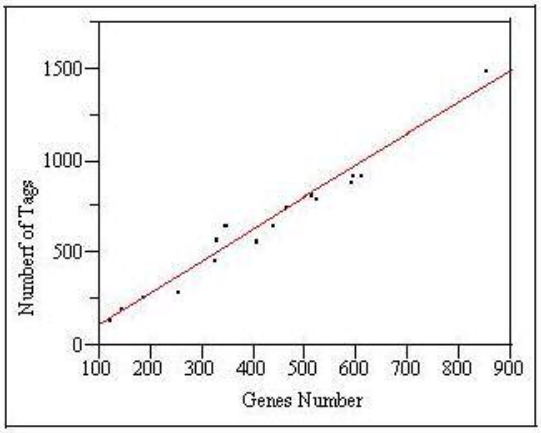

A
1
2
M
2500
2000
1500
1000
800
600
400
200
B
M
1
2
60
50
40
30
20
C
M
1
2
3
4
5
125
100
75
50
25
120bp
D
1
2
3
4
120
100
E
M
1
2
3
100
90
80
70
60
50
40
30
20
120bp
32+50bp
50bp
32bp
F
M
1
2000
1650
1000
850
600
500
400
300
200
100
800-1000bp
500-800bp
300-500bp

Figure S1

Chr	Genes	Length(bp)	Unitags
1	121	230210	150
2	460	813141	717
3	185	316616	265
4	850	1531914	1381
5	328	576869	552
6	143	270148	190
7	593	1090945	880
8	325	562641	429
9	250	439885	280
10	404	745666	531
11	346	666452	614
12	588	1078174	814
13	513	924429	753
14	439	784330	594
15	607	1091287	856
16	520	948060	732



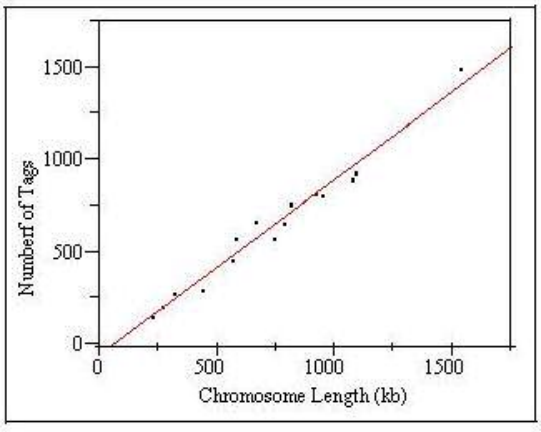


Figure S2

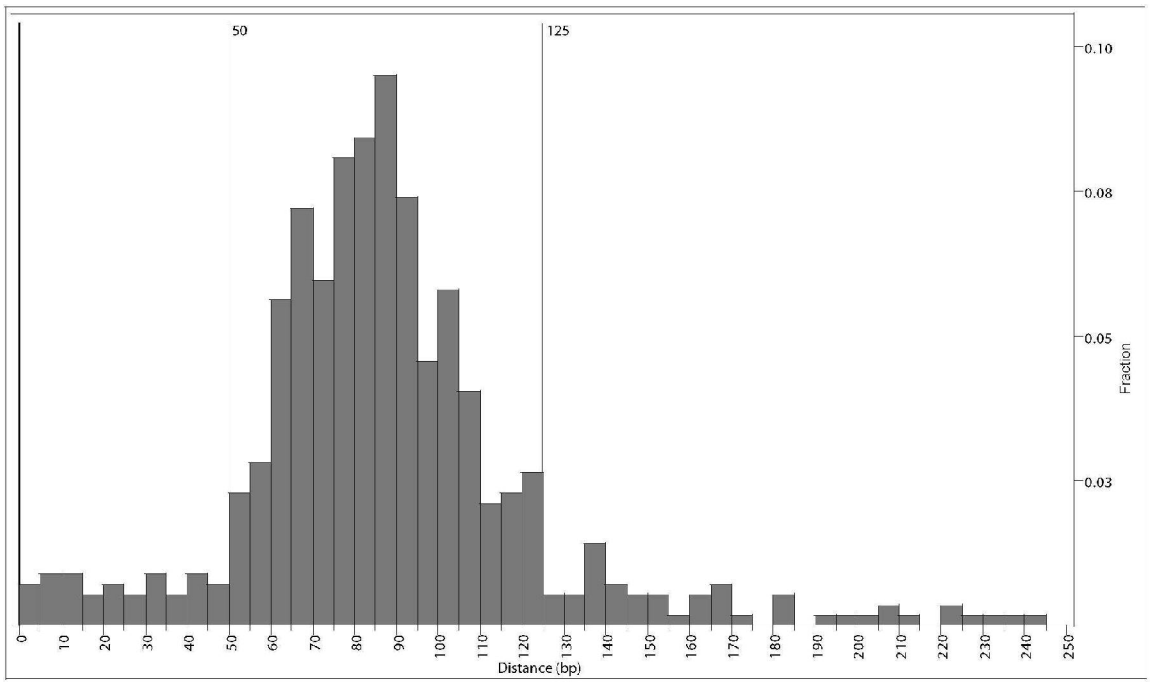

50
125
0.10
0.08
0.05
0.03
Fraction
0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
210
220
230
240
250
Distance (bp)

Figure S3

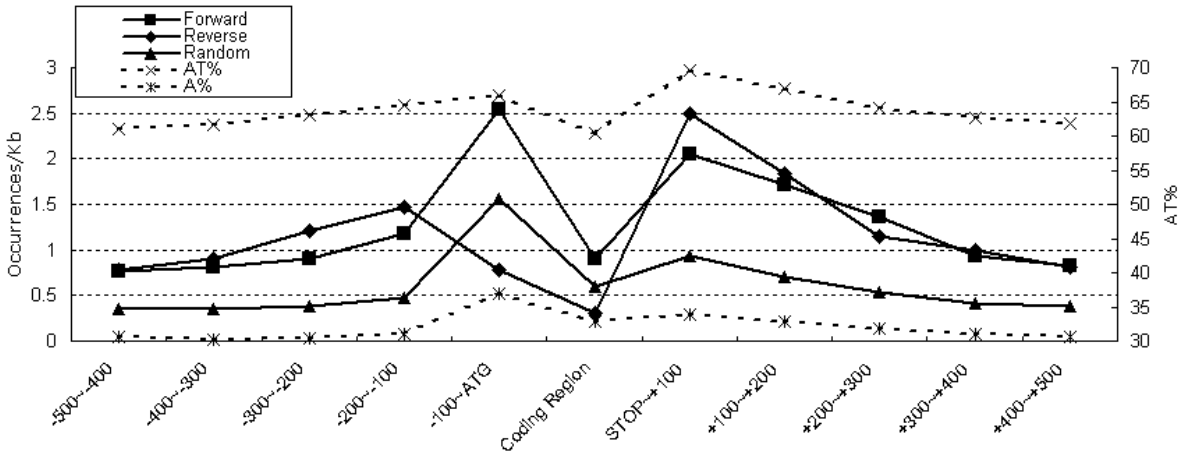

Forward
Reverse
Random
AT%
A%
Occurrences/Kb
AT%
3
2.5
2
1.5
1
0.5
0
70
65
60
55
50
45
40
35
30
-500~-400
-400~-300
-300~-200
-200~-100
-100~ATG
Coding Region
STOP~+100
+100~+200
+200~+300
+300~+400
+400~+500

Figure S4

Figure S1. Gel pictures of 5' SAGE library construction. *(A)* Double strand cDNA after primer extension on 2% Agarose-EtBr gel. Lane 1: Anchoring primer A + CDS primers; Lane 2: Anchoring primer B + CDS primers. *(B) Mme*I digestion product on 15% TBE polyacrylamide gel. Lane 1: Anchoring primer A + CDS primers; Lane 2: Anchoring primer B + CDS primers. *(C)* Ditag amplifying optimization on 15% TBE polyacrylamide gel. Lane 1, template 1/10 dilution; Lane 2, 1/20; Lane 3, 1/40; Lane 4, 1/80; Lane 5, no ligation control *(D)* Ditag scale up amplification on 12% TBE polyacrylamide gel, using 1/80 dilution template. Lane1-3, 30x50µl reaction per lane; Lane 4: 50µl reaction. *(E)* 120-bp ditag BamHI digestion on 12% polyacrylamide gel. Lane 1-3: digestion product. *(F)* Concatenation product on 2% Agarose-EtBr gel. Lane 1: concatenated ditags after ligation. Three different size ranges of fragments were cloned recovered and cloned. M: DNA size markers.

Figure S2. Correlation between gene number/chromosome length and number of tags. The tags with single hit are evenly distributed throughout the whole *S. cerevisiae* genome. Each dot in the plot represents data from a single chromosome. The highly correlation value indicates the 5' SAGE library result is not biased.

Figure S3. **TATA-TSS distances.** Distribution of distances between TATA element and TSS were calculated from 224 gene reported having single TATA element (1) and having TSS start sites (this study). A total of 568 pairs of TATA-TSS were calculated. Among them, 449 (79%) pairs fall into the range 50 to 125bp, which is consistent with previously reported result (2).

Figure S4. The frequency of consensus pattern around the *S. cerevisiae* genes. The occurrences of the TSS consensus sequence 500bp upstream, coding region, and downstream of 5888 genes in *S. cerevisiae*, were calculated in a 100bp window except the coding region, which is calculated as a single unit (Forward). The occurrences on the minus strand (Reverse) strand and random shuffled sequences (Random) are shown. The A% and AT% of corresponding region is also shown. The pattern of consensus sequences is treated as $A(A_{rich>=3})_5N(C/T)A(A/T)NN(A_{rich>=4})_6$ in program search.

Supplemental References

1. Basehoar, A.D., Zanton, S.J. and Pugh, B.F. (2004) Identification and distinct regulation of yeast TATA box-containing genes. *Cell*, **116,** 699-709.
2. Hampsey, M. (1998) Molecular genetics of the RNA polymerase II general transcriptional machinery. *Microbiol Mol Biol Rev*, **62,** 465-503.