# SUPPLEMENTARY METHODS

## Yeast MA lines

Confirmed point mutations from the yeast MA lines [1] were provided by Dr. Xiaoshu Chen. The list of essential genes was obtained from http://www-sequence.stanford.edu/group/yeast_deletion_project/deletions3.html. The fitness values of single gene deletion strains were from [2].

## Yeast and primate intron sequences

We downloaded *S. cerevisiae* and *S. paradoxus* reference genomes and annotation files from ftp://ftp.sanger.ac.uk/pub/dmc/yeast/latest [3]. From the intron sequences, we eliminated putatively functional regions (30 nucleotides of each end). Introns with remaining lengths < 50 nucleotides were discarded. The intron sequences were aligned using MUSCLE [4] with the default option. In the end, 124 yeast intron sequences from 124 different genes were used in subsequent analyses.

To identify orthologous intron sequences between human and macaque, alignments of the two genomes (hg19/GRCh37) from the 44-way multiz alignments at the UCSC Genome Browser [5] were used. Introns were identified by the genomic coordinates of human ENSEMBL transcripts available at the UCSC Genome Browser [5]. In this study, we used only constitutive introns, which are absent in all Ensembl mRNA transcripts of a given gene. Furthermore, the first and last introns and putative splice site regions (first and last 100 nucleotides of each intron) were eliminated because they may be under functional constraints [6]. In the end, the introns of 3,538 human genes were subject to subsequent analysis.

**Yeast and primate transcriptome data**

Yeast gene expression levels under YPD were determined by Illumina-based RNA-Seq [7]. The human tissue transcriptome data were from two large RNA-Seq datasets. The first dataset [8], compiled from several smaller datasets, included gene expression levels in 12 human tissues (adipose, brain, breast, cerebral cortex, colon, heart, kidney, liver, lung, lymph node, muscle, and testis). The second dataset [9] contained transcriptomic data from six tissues (brain, cerebellum, heart, kidney, liver, and testis) in humans and macaques. In this dataset, some tissues were derived from two sexes, and the mean expression levels from the two sexes were used in such cases.

In some yeast analyses, we used putatively untranscribed regions based on the following criteria: (1) they are located in regions without any annotated feature (e.g., open reading frame, autonomously replicating sequence, centromere, noncoding RNA, and transposable elements); (2) they do not overlap with any RNA-Seq reads [7].

**Confounding factors**

The percentage of GC and CpG dinucleotide frequency are calculated using an in-house Perl script. The nucleosome occupancy data of yeast under YPD [10] and those of human $CD4^+$ T cells [11] were used. Because nucleosome occupancy is largely determined by the DNA sequence, the use of human T cell data should not substantially alter our result [12]. In fact, controlling and not controlling nucleosome occupancy yielded almost identical results in humans. The obtained genomic coordinates (hg18) of human nucleosomes were converted to hg19 using the Lift-Over utility (http://genome.ucsc.edu/cgi-bin/hgLiftOver) in Galaxy (http://g2.bx.psu.edu/). Yeast DNA replication timing was from [13]. Among 58 replication-

timing profiles, mean ratio from four replications (GSM428220, GSM428221, GSM428222, and GSM428223) of the wild-type yeast (BY4741) was used here. The human replication timing profile of human HeLa cells was obtained from http://www.cgm.cnrs-gif.fr/thermes/donnees_sequencage/index.html [14]. Yeast gene expression data during cell cycles were from [15]. In human, for each of these confounding factors, we calculated the difference between the intron and the flanking untranscribed region, as in the case of calculating intron substitution rates.

We followed an earlier study [16] to calculate the transcription-driven mutability index (*TDMI*) of yeast introns. Specifically, the non-transcribed strand of an intron sequence was broken into 30-nucleotide sliding windows with a step size of 1 nucleotide. The folding of each window was estimated using the hybrid-ss-min program in the UNAFold package [17], and both the free energy ($\Delta G$) of its most stable structure and the paired/unpaired state of each site were recorded. The *TDMI* of a site was calculated as the ratio of the sum of $\exp(-\Delta G/RT)$ over all most stable folds in which the site was unpaired and the sum of $\exp(-\Delta G/RT)$ over all most stable folds that include the site, where $T$ is the temperature in degrees Kelvin and $R$ is the ideal gas constant. The *TDMI* of an intron is the mean *TDMI* of all sites in the intron.

**Analysis of mRNA synthesis and degradation rates**

Yeast mRNA synthesis rates were estimated by native elongating transcript sequencing (NET-seq) based on deep sequencing of 3' ends of nascent transcripts associated with RNA polymerase [18]. We only considered sense-strand transcriptions, because antisense transcription rates are much lower [18]. To estimate mRNA degradation rates, we analyzed the mRNA decay profiles with seven time points (0, 5, 10, 20, 30, 40, and 50 min) from microarray

experiments [19]. RNA degradation may be approximated by the first order exponential decay model described by $A_t = A_0\, e^{-kt}$, where $k$ is the degradation rate, $t$ is time, and $A_0$ and $A_t$ are the RNA concentrations at time 0 and $t$, respectively. This formula can be converted to $\ln(A_t) = -kt + \ln(A_0)$. Thus, $-k$ can be estimated from the slope of the linear regression between $t$ and $\ln(A_t)$.

**Multiple regression analysis**

We estimated the relative contributions of multiple predictors to the total variance in yeast mutation rate by calculating the relative contribution of variability explained (*RCVE*) for each predictor using $RCVE = 1 - R^2_{reduced} / R^2_{full}$, where $R^2_{full}$ and $R^2_{reduced}$ are the $R^2$ (square of the correlation coefficient) for the full linear model and the model without the predictor of interest, respectively. To diagnose multicollinearity of each predictor, variance inflation factors (*VIF*s) [20] were calculated. All predictors in the model used had *VIF*s below 2, suggesting that multicollinearity did not adversely affect our model. Multiple linear regression analysis was performed in the R statistical package.

In the human data, we used a multiple linear regression to estimate the impact of expression level and four other factors on mutation rate. A total of five predictors (expression level, GC content, CpG frequency, nucleosome binding, and replication timing) were used to build the regression model. Because the mutation rate used in the model (mean = 0.00105) is the difference between the mutation rate of all introns of a gene and the mutation rate of its flanking untranscribed region, each predictor used is also the difference between the values for the introns minus that for the flanking untranscribed region. The expression level of each flanking untranscribed region is assumed to be 0. We used gene expression levels from human testis in

this analysis, which has a mean value of 0.2297. The estimated coefficient for the expression level predictor is 0.000689 ($P < 0.0071$).

**References**


1. Chen X, Chen Z, Chen H, Su Z, Yang J, et al. (2012) Nucleosomes suppress spontaneous mutations base-specifically in eukaryotes. Science 335: 1235-1238.
2. Deutschbauer AM, Jaramillo DF, Proctor M, Kumm J, Hillenmeyer ME, et al. (2005) Mechanisms of haploinsufficiency revealed by genome-wide profiling in yeast. Genetics 169: 1915-1925.
3. Liti G, Carter DM, Moses AM, Warringer J, Parts L, et al. (2009) Population genomics of domestic and wild yeasts. Nature 458: 337-341.
4. Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792-1797.
5. Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, et al. (2011) The UCSC Genome Browser database: update 2011. Nucleic Acids Res 39: D876-882.
6. Keightley PD, Gaffney DJ (2003) Functional constraints and frequency of deleterious mutations in noncoding DNA of rodents. Proc Natl Acad Sci U S A 100: 13402-13406.
7. Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, et al. (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320: 1344-1349.
8. Xiong Y, Chen X, Chen Z, Wang X, Shi S, et al. (2010) RNA sequencing shows no dosage compensation of the active X-chromosome. Nat Genet 42: 1043-1047.
9. Brawand D, Soumillon M, Necsulea A, Julien P, Csardi G, et al. (2011) The evolution of gene expression levels in mammalian organs. Nature 478: 343-348.
10. Lee W, Tillo D, Bray N, Morse RH, Davis RW, et al. (2007) A high-resolution atlas of nucleosome occupancy in yeast. Nat Genet 39: 1235-1244.
11. Schones DE, Cui K, Cuddapah S, Roh TY, Barski A, et al. (2008) Dynamic regulation of nucleosome positioning in the human genome. Cell 132: 887-898.
12. Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, et al. (2009) The DNA-encoded nucleosome organization of a eukaryotic genome. Nature 458: 362-366.
13. Koren A, Soifer I, Barkai N (2010) MRC1-dependent scaling of the budding yeast DNA replication timing program. Genome Res 20: 781-790.
14. Chen CL, Rappailles A, Duquenne L, Huvet M, Guilbaud G, et al. (2010) Impact of replication timing on non-CpG and CpG substitution rates in mammalian genomes. Genome Res 20: 447-457.
15. Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, et al. (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. Mol Cell 2: 65-73.
16. Hoede C, Denamur E, Tenaillon O (2006) Selection acts on DNA secondary structures to decrease transcriptional mutagenesis. PLoS Genet 2: e176.
17. Markham NR, Zuker M (2005) DINAMelt web server for nucleic acid melting prediction. Nucleic Acids Res 33: W577-581.
18. Churchman LS, Weissman JS (2011) Nascent transcript sequencing visualizes transcription at nucleotide resolution. Nature 469: 368-373.

19. Shalem O, Groisman B, Choder M, Dahan O, Pilpel Y (2011) Transcriptome kinetics is governed by a genome-wide coupling of mRNA production and degradation: a role for RNA Pol II. PLoS Genet 7: e1002273.
20. Kutner MH, Nachtsheim CJ, Neter J, Li W (2005) Applied linear statistical models. McGraw-Hill, New York.

**Legends for supplementary figures**

**Fig. S1.** Correlation between the expression level of a yeast gene and the nucleotide substitution rate in its intron. The nucleotide substitution rate is measured by comparing *S. cerevisiae* and *S. paradoxus* orthologous intron sequences. Each dot represents a gene.

**Fig. S2.** Mutational patterns revealed from the yeast MA lines for the 50% most weakly expressed and 50% most strongly expressed nucleotide positions in the genome. Due to the small sample size, mutation frequency difference between sites with low and high expressions is not significant for any mutational type ($P > 0.2$).

**Fig. S3.** Partial rank correlations between gene expression levels in 12 human tissues and the nucleotide substitution rate between human and macaque, after the simultaneous controls of GC content, CpG frequency, replication timing, and nucleosome binding. The substitution rate is estimated by the rate at intron sites minus that in flanking untranscribed regions. The figure is identical to Fig. 3 except that the flanking regions are 10 kb rather than 5 kb away from the end of 3UTRs. Significant differences from 0 are indicated by * ($P < 0.05$), ** ($P < 0.01$), or *** ($P < 0.001$).

**Fig. S4.** Partial rank correlations between gene expression levels in 12 human tissues and the nucleotide substitution rate between human and macaque, after the simultaneous controls of GC content, CpG frequency, replication timing, and nucleosome binding. The substitution rate is estimated by the rate at synonymous sites (calculated by PAML) minus that in flanking

untranscribed regions. Significant differences from 0 are indicated by * ($P < 0.05$), ** ($P < 0.01$), or *** ($P < 0.001$).

**Fig. S5.** Partial rank correlations between gene expression level and the nucleotide substitution rate, after simultaneous controls of GC content, CpG frequency, replication timing, and nucleosome binding. The substitution rate between human and macaque is estimated by the rate in introns minus that in flanking untranscribed regions. The expression levels are the averages of those of human and macaque. The human liver and human and macaque testis data are from males, whereas those of other tissues are averages of males and females. Significant differences from 0 are indicated by * ($P < 0.05$), ** ($P < 0.01$), or *** ($P < 0.001$).

**Fig. S6.** Lack of significant correlation between the nucleotide substitution rate of an intron and the transcription-driven mutability index (*TDMI*) of the intron. Each dot represents a yeast gene. Nucleotide substitution rates are estimated from orthologous introns of *S. cerevisiae* and *S. paradoxus*.

**Fig. S7.** Yeast indel frequencies in untranscribed regions and introns with different expression levels, determined by comparing *S. cerevisiae* with *S. paradoxus* orthologous sequences. For each indel size, none of the bars for transcribed regions are significantly higher than that for untranscribed regions ($P > 0.1$). Genes with low, intermediate, and high expressions refer to those whose expression levels are in the bottom quartile, middle two quartiles, and top quartile, respectively.

**Fig. S8.** Differences of indel frequencies between human introns and their flanking intergenic regions, determined by human-macaque comparisons. Intermediate and high expression bars are significantly different from each other ($P < 0.05$) for each indel size except 2-bp. Three of the comparisons between low expression and the others are significant ($P < 0.05$): between low and high expressions for the 1-bp indel type; between low and intermediate expressions for the 3-5-bp indel type; between low and high expressions for "all indels". Genes with low, intermediate, and high expressions refer to those whose expression levels are in the bottom quartile, middle two quartiles, and top quartile, respectively.

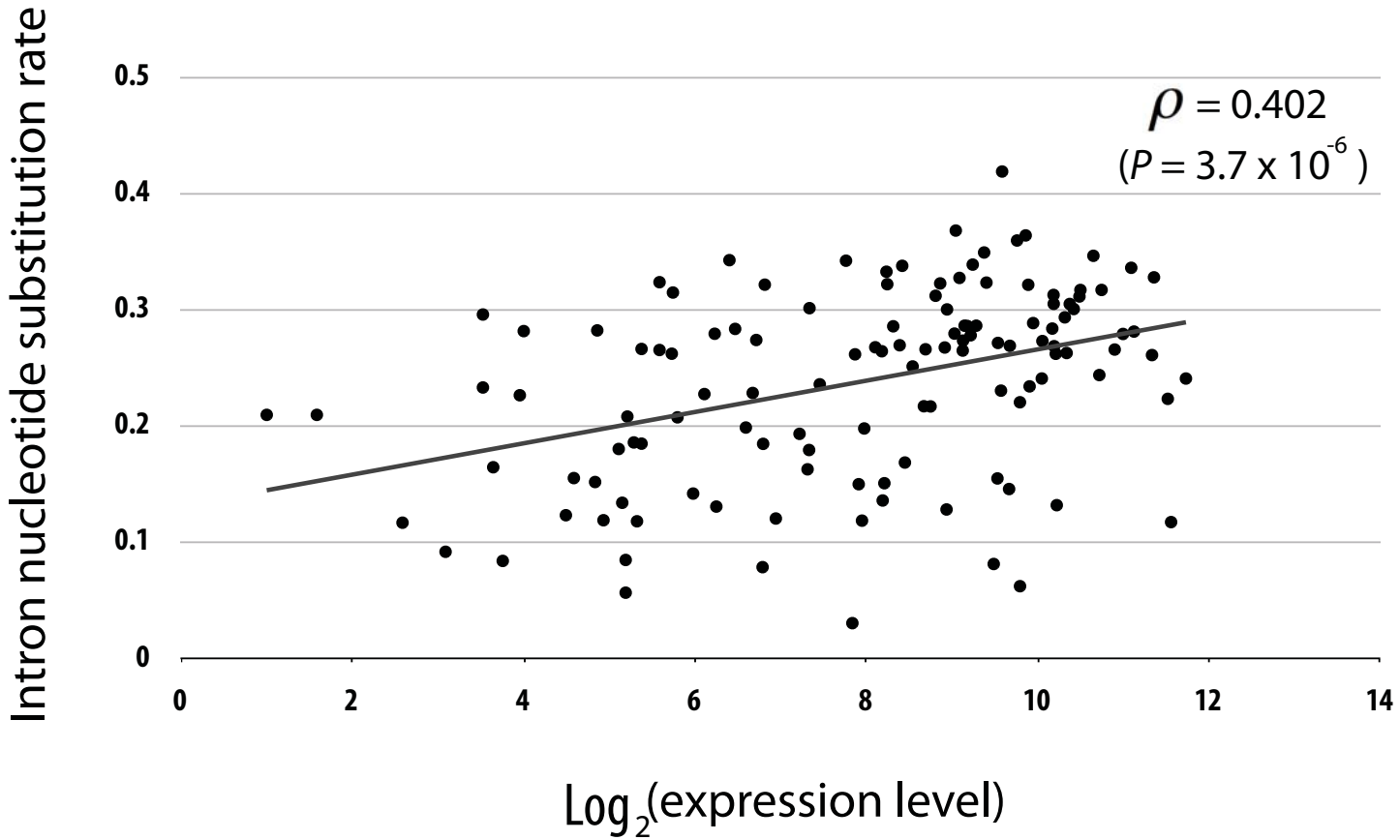
Intron nucleotide substitution rate
0.5
0.4
0.3
0.2
0.1
0
0
2
4
6
8
10
12
14
ρ = 0.402
(P = 3.7 x 10-6 )
Log2(expression level)

Figure S1

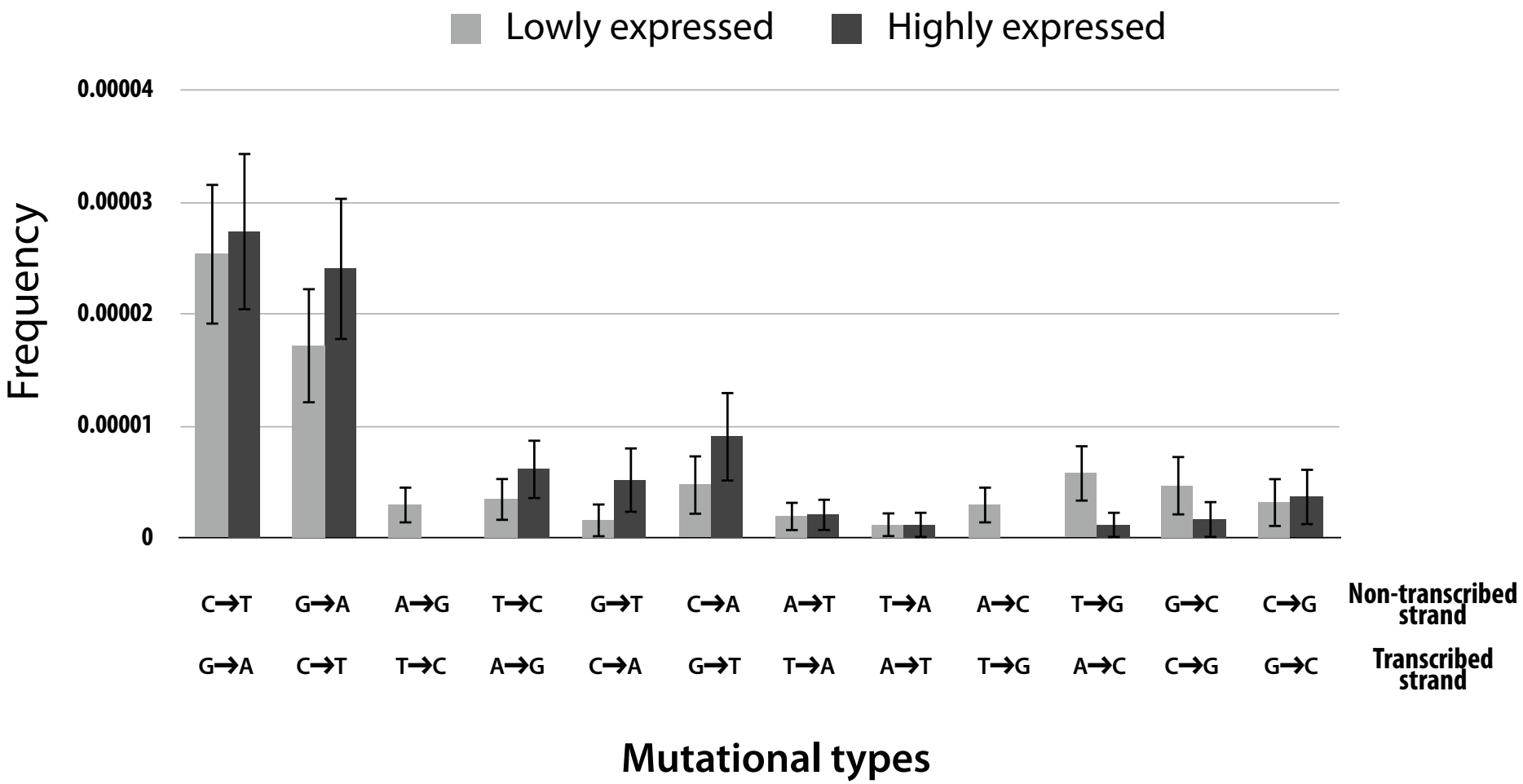
Lowly expressed
Highly expressed
Frequency
0.00004
0.00003
0.00002
0.00001
0
C→T
G→A
A→G
T→C
G→T
C→A
A→T
T→A
A→C
T→G
G→C
C→G
Non-transcribed strand
G→A
C→T
T→C
A→G
C→A
G→T
T→A
A→T
T→G
A→C
C→G
G→C
Transcribed strand
Mutational types

Figure S2

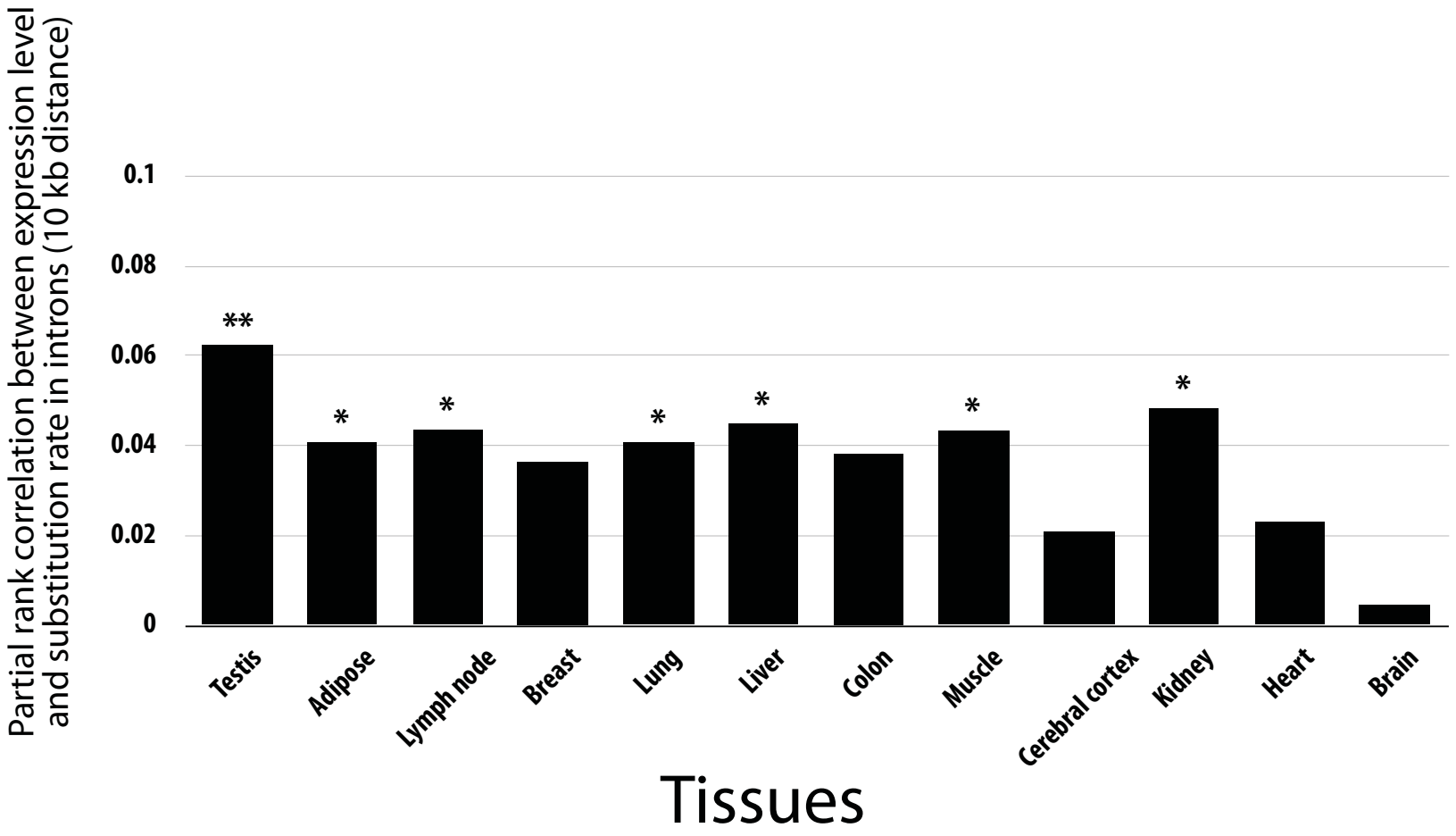
Partial rank correlation between expression level and substitution rate in introns (10 kb distance)
0.1
0.08
0.06
0.04
0.02
0
**
*
*
*
*
*
*
*
Testis
Adipose
Lymph node
Breast
Lung
Liver
Colon
Muscle
Cerebral cortex
Kidney
Heart
Brain
Tissues

Figure S3

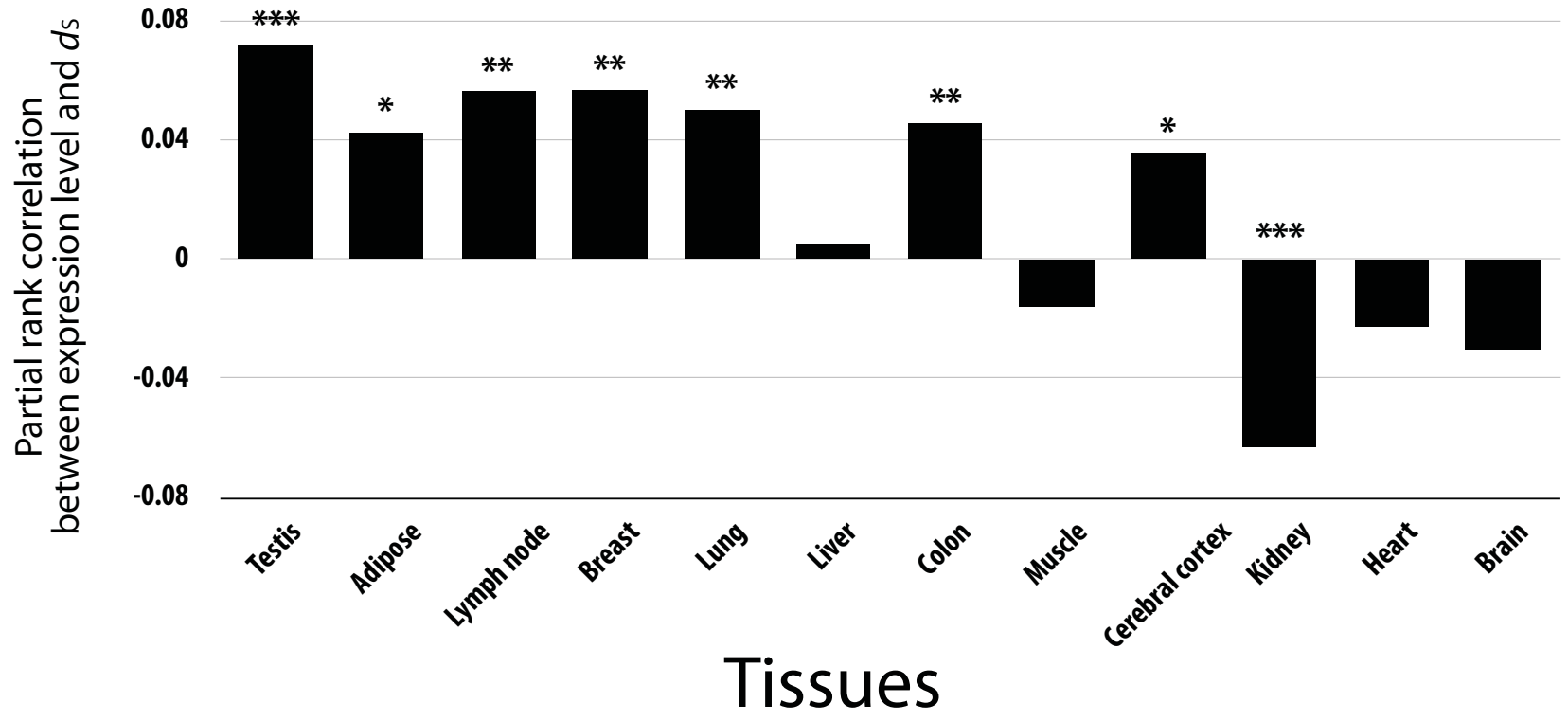
Partial rank correlation between expression level and $d_S$
0.08
0.04
0
-0.04
-0.08
***
*
**
**
**
**
*
***
Testis
Adipose
Lymph node
Breast
Lung
Liver
Colon
Muscle
Cerebral cortex
Kidney
Heart
Brain
Tissues

Figure S4

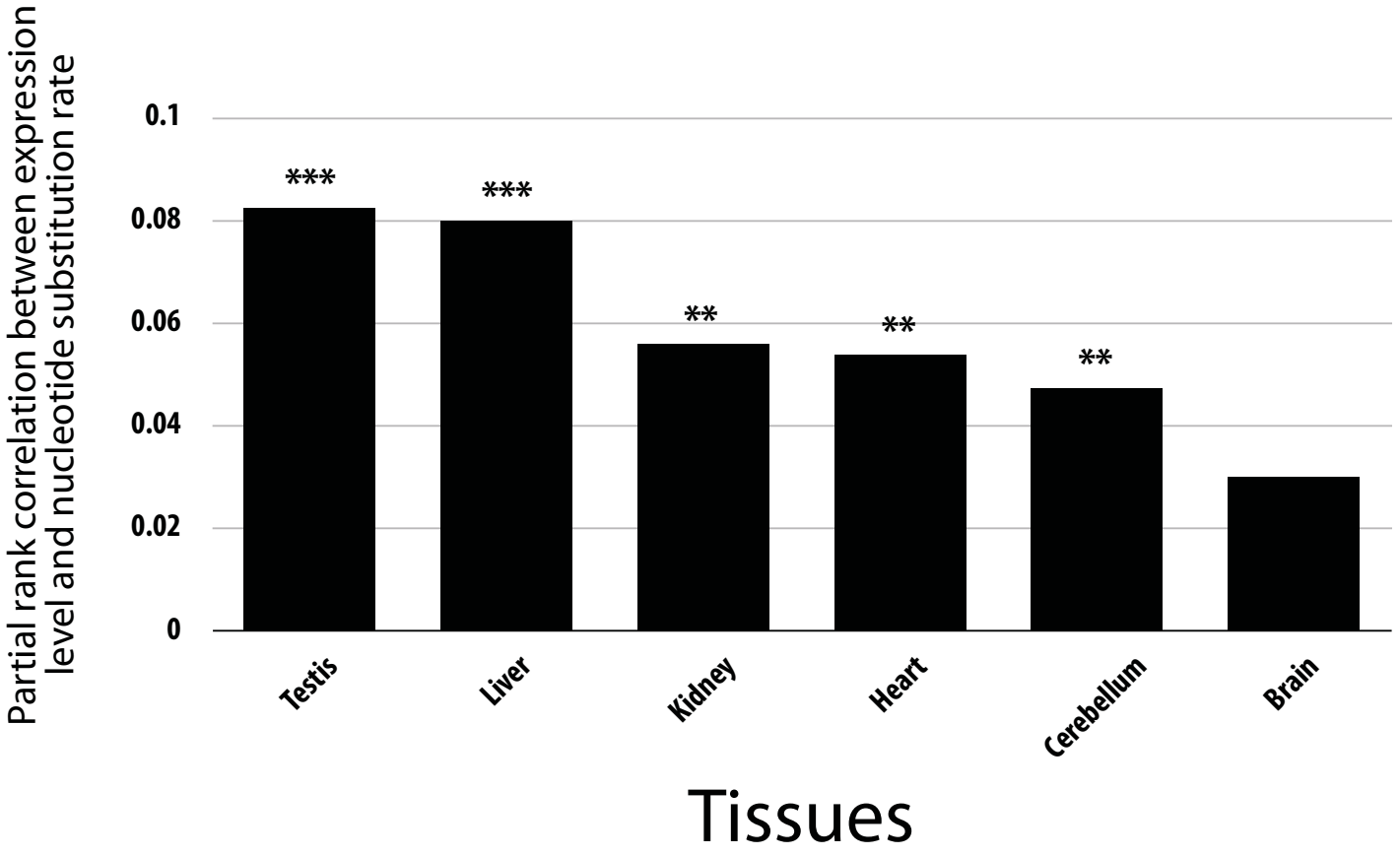
Partial rank correlation between expression level and nucleotide substitution rate
0.1
0.08
0.06
0.04
0.02
0
***
***
**
**
**
Testis
Liver
Kidney
Heart
Cerebellum
Brain
Tissues

Figure S5

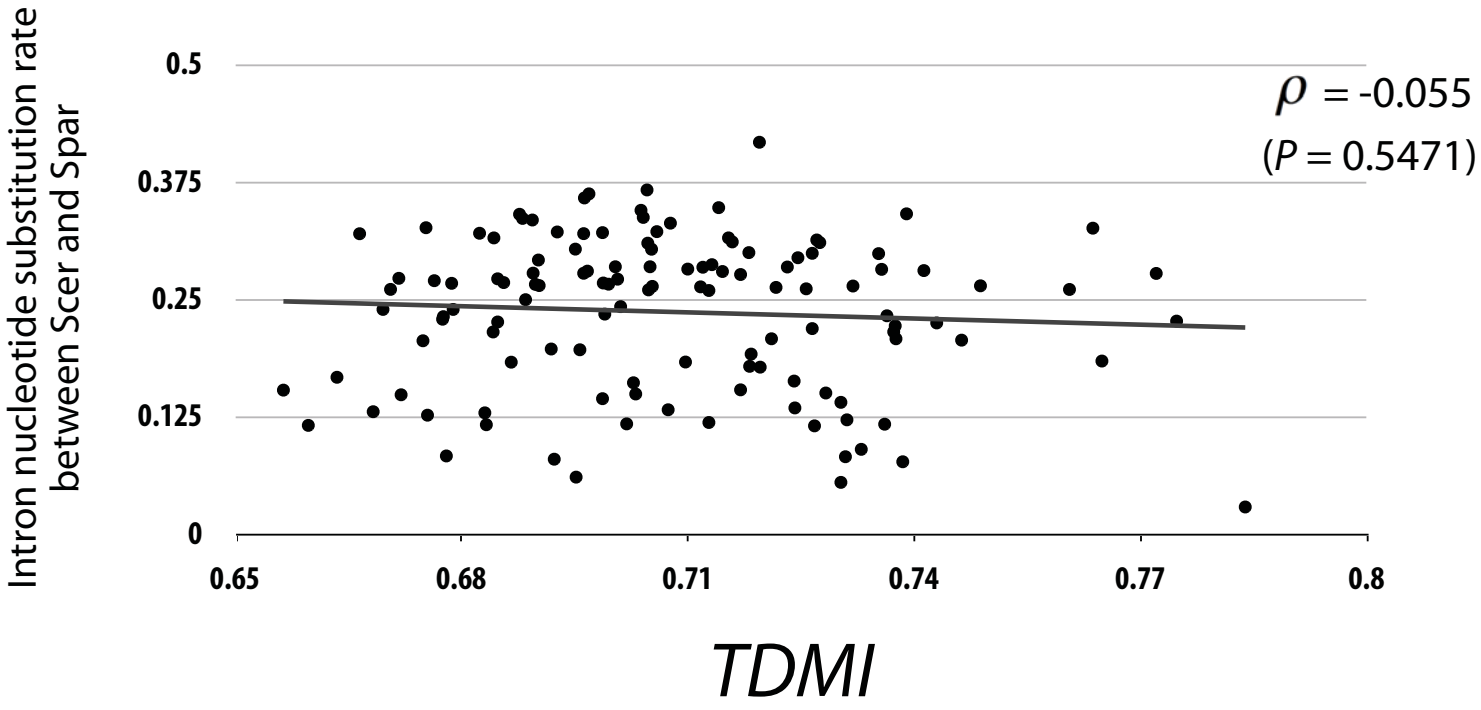
Intron nucleotide substitution rate between Scer and Spar
0.5
0.375
0.25
0.125
0
0.65
0.68
0.71
0.74
0.77
0.8
ρ = -0.055
(P = 0.5471)
TDMI

Figure S6

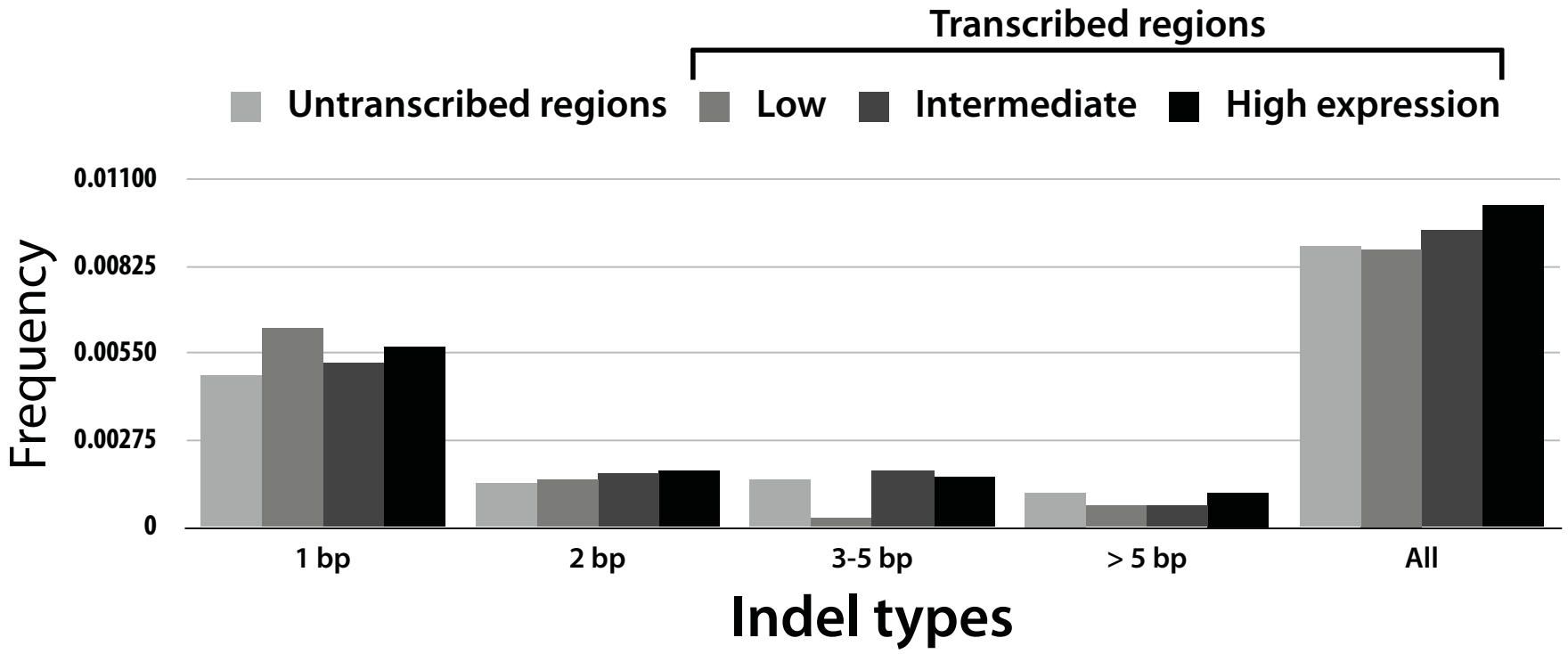
Transcribed regions
Untranscribed regions
Low
Intermediate
High expression
Frequency
0.01100
0.00825
0.00550
0.00275
0
1 bp
2 bp
3-5 bp
> 5 bp
All
Indel types

Figure S7

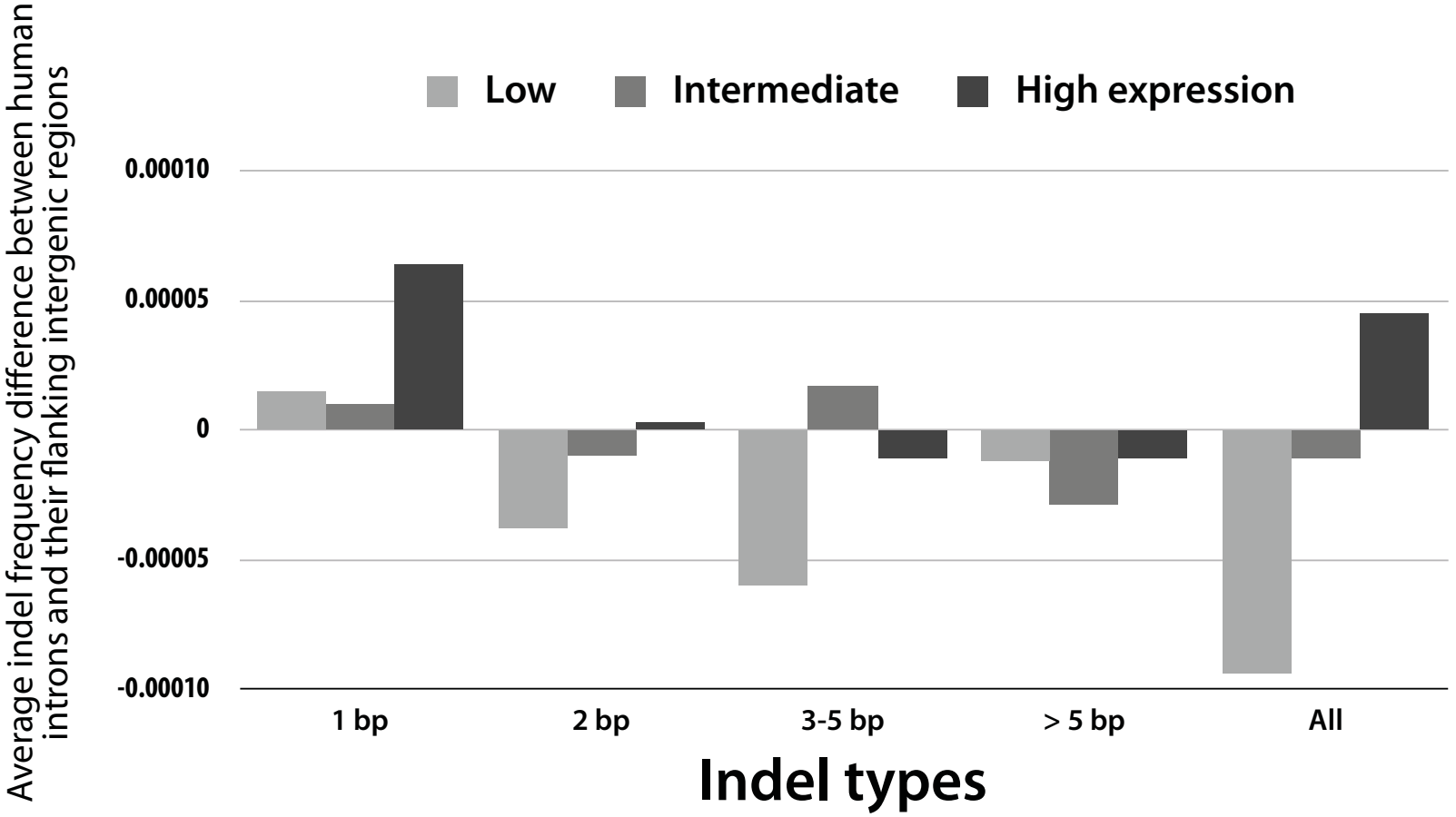
Low
Intermediate
High expression
Average indel frequency difference between human introns and their flanking intergenic regions
0.00010
0.00005
0
-0.00005
-0.00010
1 bp
2 bp
3-5 bp
> 5 bp
All
Indel types

Figure S8