SUPPORTING ONLINE MATERIAL

Indel Detection

Small (<=30-bp) indels occurring in the human lineage since its divergence from chimpanzee were identified from the human-chimpanzee-macaque (hg18-panTro2-rheMac2) 3-way Multi-Z alignments (Blanchette et al. 2004), following the methods described in (Kvikstad et al. 2007). Briefly, as part of the macaque consortium, we derived a substitution rate matrix (Chiaromonte et al. 2002) and determined gap penalties appropriate for the human-chimpanzee divergence. Resulting alignments were analyzed for accuracy using an alignment diagnostic termed gap attraction (Lunter et al. 2006; Lunter et al. 2008) that was found to be minimal (Kvikstad et al. 2007).

Furthermore, we developed a computational pipeline employing rigorous filtering criteria to remove potential false positives that could be attributed to the alignment of draft quality sequences to the finished human genome; filtering was applied to gaps occurring in overlapping alignment blocks that could be due to duplicated regions, to gaps of unequal lengths among species that could be due to sequence errors and/or multiple events, and to gaps flanked (+/- 3 nucleotides) by low quality (Phred score <=20) nucleotides in either draft genome (Kvikstad et al. 2007). Additionally, indels were excluded if they occurred in microsatellite, simple repeat or low complexity regions (Smit et al. 1996-2004) for the sake of sequence, assembly, and alignment accuracy. Thus, our filtered data set likely represents a conservative estimate of the actual number of indel mutations that have accumulated in the human genome since divergence from chimpanzee.

Non-coding, Non-repetitive (NCNR) Genome

We focused our analyses on the NCNR portion of the genome for several reasons. First, indel rates and patterns may differ considerably between coding and non-coding DNA due to the influence of natural selection (e.g., Lunter et al. 2006). Second, the detection of small sequence motifs in repetitive DNA may be biased due to base composition and chromosomal preferences of various transposable element families (Lander et al. 2001). Yet, previously we analyzed

genome-wide heterogeneity in indel rates using ancestral repeats (ARs) as a model for neutral DNA (Kvikstad et al. 2007). Applying similar methodology, we observed that variation in insertion and deletion rates at the 1-Mb scale is similar between the AR and NCNR portions of the genome (data not shown). Thus, our choice to focus here on NCNR as the neutral portion of the genome is unlikely to introduce any significant biases in evaluating the forces that shape indel rates and patterns.

Finally, we took advantage of available indel polymorphism data in order to conduct a direct comparison of observed vs. expected rates of insertions (deletions separately) in the two presumably neutral data sets. We compared the rates of chromosome 1 polymorphic (from Mills et al. 2006) vs. fixed insertions and (separately) deletions occurring in NCNR sequences (insertions: 3.3×10^{-5} polymorphic, 8.4×10^{-5} fixed; deletions: 3.4×10^{-5} polymorphic, 1.8×10^{-4} fixed) to those in AR sequences (insertions: 1.8×10^{-5} polymorphic, 1.6×10^{-4} fixed; deletions: 1.7×10^{-5} polymorphic, 2.6×10^{-4} fixed) using a modified Hudson-Kreitman-Aguade test (Hudson et al. 1987). The test results were not significant ($p>0.95$ for both insertions and deletions), suggesting that indels identified in NCNR regions are unlikely to be strongly affected by different forces than those acting on AR regions, with the latter regions widely accepted as a model of neutral evolution (Hardison et al. 2003; Lunter et al. 2006).

Wavelet Transformation Methodology

A wavelet transform is a type of decomposition that allows for the localization of a signal in time (or an otherwise defined natural order) and variation frequency or scale (Lio 2003; Percival and Walden 2006). The input signal (X) is *dilated* over scales (j=1…J), and *translated* over times by inner product with a so-called wavelet filter (Φ). In the case of a discrete wavelet transform (DWT), the resulting wavelet coefficients (W_j) describe the signal in terms of changes in the averages of its values over various scales, while scaling coefficients (V_j) are associated with the averages themselves (Percival and Walden 2006). By accounting for multiple scales simultaneously, the coefficients

produced by a wavelet transform represent both global trends and local fluctuations in the original signal, which is decomposed as:

$$X = \Phi^T \mathbf{W} = \sum_{j}^{J} \Phi_j{}^T \mathbf{W}_j + V_J^T \mathbf{V}_J$$

Wavelet coefficients are scale-specific and orthogonal across scales, thus enabling the decomposition of signal features (i.e. functions of the signal) across scales. For example, a scale-by-scale analysis of the variance of the wavelet coefficients decomposes the variance of the input signal into the contributions attributable to each scale (Percival and Walden 2006); because the coefficients are uncorrelated, signal variability is resolved into component changes at each scale, without propagation from smaller to larger scales. The same can be done for second moments or for cross-moments when considering more than one input signal. Thus, wavelet analysis provides a useful framework for the investigation of fluctuations in signals and patterns in data that might otherwise be overlooked by *a priori* selection of scale, a fact that is crucial to the analyses presented in this article.

Wavelet techniques have been employed in several areas of biological research including ecological time series (e.g., Dale and Mah 1998; Keitt and Urban 2005; 2006), protein structure prediction (Hirakawa et al. 1999; Lio 2003), and amino acid substitution rate modeling (Morozov et al. 2000). Applications of wavelets to DNA sequence data have remained rather sparse (reviewed in Lio 2003). Early studies utilized wavelet transformations to analyze small data sets of protein coding genes and discern the underlying long-range correlations in DNA base composition (Arneodo et al. 1995; 1998). Bacterial genomes composed of single chromosomes were examined for presence of novel pathogenic islands via patterns in GC content (Lio and Vannucci 2000). More recently, signatures of nucleosome positioning were revealed via comparative wavelet analyses of eukaryotic DNA (Audit et al. 2001; 2002; Thurman et al. 2007; Yuan and Liu 2008). Finally, wavelets were used to investigate associations among multiple signals - nucleotide diversity, recombination, and other sequence features on human chromosome 20 (Spencer et al. 2006). Yet, wavelet analysis of the

human genome has remained elusive, due in part to considerable differences among chromosomes in many sequence characteristics, e.g., gene content (Lander et al. 2001) and base composition (Schmidt and Frishman 2008).

Because spatial patterns in motif occurrences could reflect underlying variation in base composition and/or substitution rates across the genome (see above; Arneodo et al. 1995), and because the accuracy of functions computed on wavelet coefficients decreases at large scales (larger wavelet scales have fewer coefficients; Percival and Walden 2006), we therefore implemented *a random permutation scheme* to assess significance accounting for these compositional and accuracy effects (Dale and Mah 1998; Keitt and Fischer 2006). For each wavelet analysis, significance was assessed computationally by permutation of the ordered time series in the frequency profiles prior to wavelet transformation and multi-scale analysis, allowing us to derive empirical *p*-values for each statistic of interest (second moments or cross-moments). Corresponding empirical *p*-values were computed for each motif, each event type (insertion, deletion), each flank (5', 3') and each scale, and adjusted for multiple testing as to control the false discovery rate (FDR; Benjamini and Hochberg 1995) at 5% (significance was reported in all cases with an adjusted *p*-value < 0.05).

Analysis of 1-bp and Polymorphic Indels

We further investigated any potential context biases due to heterogeneity in indel sizes or evolutionary times of occurrence that could affect our analysis of sequence motifs involved in indel formation. We used 1-bp events to study potential biases due to size differences, since single nucleotide insertions and deletions constitute ~50% of small indels (Kvikstad et al. 2007). Additionally, we used polymorphic indels segregating in the human population (Mills et al. 2006; see above) to study potential biases due to varying divergence times, since such indels are "young" events and less likely to have undergone millions of years of selection and/or drift. Here we provide a preliminary comparison of motif behaviors flanking each of these indel types, restricting attention to chromosome

1 and to one of the main analyses, namely the detection of significant spatial patterns (enrichment profiles).

For each event type, indels were restricted to regions in our defined NCNR portion of chromosome 1 (Table S5). Chromosome 1 was chosen because it represents ~10% of the human genome (Lander et al. 2001). To be consistent with our criteria, polymorphic indels identified in (Mills et al. 2006) were further filtered to exclude those due to "repeat expansions" as identified by the authors. For comparison, we also created a subset of our original indels restricted to chromosome 1. Total frequency profiles (see Methods) were constructed for each motif in each subgenome for each data set: chromosome 1 1-bp indels, chromosome 1 polymorphic indels, chromosome 1 original indels. Control profiles were built by sampling the NCNR control subgenome in equal size to each insertion/deletion data set.

Results for enrichment profiles are summarized in Tables S6, S7, S8. Notably, significant motifs for each data set largely represent a subset of the motifs found significant in the main results. For example, motifs with significant enrichment profiles surrounding indels restricted to chromosome 1 are mostly a subset of the genome-wide results (4/5 for deletions and 7/9 for insertions; Table S6). We detected three motifs with significant enrichment profiles on chromosome 1 (but not genome-wide); notably, these were significant genome-wide before FDR correction was applied, but failed the 5% threshold after correction. Analysis of the 1-bp events reveals that all motifs identified flanking deletions were significant in the main findings, and the majority for insertions as well (3/5; Table S7). Finally, the results for motifs' behavior flanking polymorphic indels are again largely consistent with our main findings (3/3 deletions; 5/9 insertions, with all 4 motifs significant before FDR correction; S8).

Thus, while results for deletions are consistent with our main findings, motif behaviors flanking insertions show slightly more heterogeneity depending on size (1 bp) or evolutionary time (polymorphic), yet these subtle differences consist of motifs that were detected genome-wide, although failing to pass an FDR cut-off of 5% -- we therefore do not consider them as novel findings. Hence,

we conclude that variation due to indel sizes and "ages" is not sufficient to alter the main findings presented here.

SUPPLEMENTAL FIGURE LEGENDS

Figure S1. Total frequency profiles for an example motif, topoisomerase cleavage site 4, in 5' (blue) and 3' regions (black) flanking deletions. Red bands correspond to the 95th percentile distribution of total frequency profiles obtained from permutation of the 5' and 3' position labels under the null hypothesis of no positional difference (see Methods). Flanks not significantly different from permutation testing are indicated in red. As expected, the null distribution shows remarkable symmetry irrespective of position relative to deletion mutation. Note, however, the significant positional difference for approximately 36% of the data points (Table S1) in the real total frequency profiles (blue points 5' and black points 3', respectively). P-values are provided for the flanks closest to the breakpoint that show extreme behavior for this motif and roughly 25% of the motifs analyzed (Table S1). Green bands correspond to the 95% distribution of the total frequency profiles in the control subgenome.

Figure S2. Mutli-scale analysis of enrichment profiles: Indel vs. control. Here an example motif's total frequency profile in an indel-related subgenome is compared with that in the corresponding control subgenome (black and green lines in Fig. 2A, respectively). The difference, i.e. the enrichment profile, is wavelet-transformed (left panel), and its size is measured by wavelet-based second moments computed at multiple scales (right panel; black line). Significance is assessed by randomly permuting the original frequency profiles, and recomputing enrichment profile, wavelet transform and second moments following each permutation – the red bands in the lower right panel capture 95% of the resulting "null" second moments (shown prior to FDR correction). Due to the decreasing number of available wavelet coefficients, the power of this analysis decreases as the scale increases. Yet, the topoisomerase cleavage site

4 motif still presents a significant enrichment profile at large scales (observed second moments outside the red bands).

Figure S3. Insertions vs. deletions: mutli-scale analysis of similarity between profiles. This is investigated comparing the example motif's total frequency profile in a deletion-related subgenome with that in the corresponding insertion-related subgenome (e.g. the black lines, Fig. 2B). Each of the two profiles is wavelet transformed (left and middle panels), and their similarity is measured through wavelet-based Kendall's tau correlations computed at multiple scales (right panel; black line). Significance is again assessed by randomly permuting the original frequency profiles, and recomputing wavelet transforms and correlations following each permutation – the red bands in the lower right panel capture 95% of the resulting "null" Kendall's taus, with the expected increasing width as scale increases. The topoisomerase cleavage site 4 motif presents significantly dissimilar spatial patterns 5' of deletions and insertions at the 80-bp scale (observed Kendall's tau is indeed outside the red bands; shown prior to FDR correction).

Figure S4. Motif X vs. motif Y: mutli-scale analysis of co-location in profiles. Along with topoisomerase cleavage site 4 (X), here we consider DNA Pol pause/frameshift hotpsot 1 (Y). Co-location is investigated comparing the total frequency profiles of X and Y in an indel-related subgenome (e.g. the black lines for X on the left, and Y on the right in Fig. 2C. Each of the two profiles is wavelet transformed (left and middle panels), their multi-scale similarity is measured through wavelet-based Kendall's tau correlations (right panel; black line), and significance is assessed through random permutations of the original frequency profiles, resulting in the 95% "null" red bands in the right panel (again, shown prior to FDR correction). Topoisomerase cleavage site 4 and DNA Pol pause/frameshift hotpsot 1 (Y) present a significant co-location only at very small scales. The anticorrelation at large scales is not significant.

REFERENCES FOR THE SUPPLEMENTAL MATERIAL

Arneodo, A., E. Bacry, P.V. Graves, and J.F. Muzy. 1995. Characterizing long-range correlations in DNA sequences from wavelet analysis. *Physical Review Letters* **74:** 3293-3296.

Arneodo, A., Y. D'aubenton-Carafa, B. Audit, E. Bacry, J.F. Muzy, and C. Thermes. 1998. What can we learn with wavelets about DNA sequences? *Physica A* **249:** 439-448.

Audit, B., C. Thermes, C. Vaillant, Y. D'aubenton-Carafa, J.F. Muzy, and A. Arneodo. 2001. Long-range correlations in genomic DNA: a signature of the nucleosomal structure. *Physical Review Letters* **86:** 2471-2474.

Audit, B., C. Vaillant, A. Arneodo, Y. D'aubenton-Carafa, and C. Thermes. 2002. Long-range corelations between DNA bending sites: relation to the structure and dynamics of nucleosomes. *Journal of Molecular Biology* **316:** 903-918.

Benjamini, Y. and Y. Hochberg. 1995. Comtrolling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **57:** 289-300.

Blanchette, M., W.J. Kent, C. Riemer, L. Elnitski, A.F.A. Smit, K.M. Roskin, R. Baertsch, K. Rosenbloom, H. Clawson, E.D. Green et al. 2004. Aligning Multiple Genomic Sequences With the Threaded Blockset Aligner. *Genome Research* **14:** 708-715.

Chiaromonte, F., V.B. Yap, and W. Miller. 2002. Scoring pairwise genomic sequence alignments. In *Pacific Symp Biocomput*, pp. 115-126.

Dale, M. and M. Mah. 1998. The use of wavelets for spatial pattern analysis in ecology. *Journal of Vegetation Science* **9:** 805-814.

Hardison, R.C., K.M. Roskin, S. Yang, M. Diekhans, W.J. Kent, R. Weber, L. Elnitski, J. Li, M. O'Connor, D. Kolbe et al. 2003. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Research* **13:** 13-26.

Hirakawa, H., S. Muta, and S. Kuhara. 1999. The hydrophobic cores of proteins predicted by wavelet analysis. *Bioinformatics* **15:** 141-148.

Hudson, R., M. Kreitman, and M. Aguade. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics* **116:** 153-159.

Keitt, T.H. and J. Fischer. 2006. Detection of scale-specific community dynamics using wavelets. *Ecology* **87:** 2895-2904.

Keitt, T.H. and D.L. Urban. 2005. Scale-specific inference using wavelets. *Ecology* **86:** 2497-2504.

Kvikstad, E., S. Tyekucheva, F. Chiaromonte, and K. Makova. 2007. A macaque's-eye view of human insertions and deletions: differences in mechanisms. *Public Library of Sciences Computational Biology* **3:** e176.

Lander, E.S. L.M. Linton B. Birren C. Nusbaum M.C. Zody J. Baldwin K. Devon K. Dewar M. Doyle W. FitzHugh et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860-921.

Lio, P. 2003. Wavelets in bioinformatics and computational biology: state of art and perspectives. *Bioinformatics* **19:** 2-9.

Lio, P. and M. Vannucci. 2000. Finding pathogenicity islands and gene transfer events in genome data. *Bioinformatics* **16:** 932-940.

Lunter, G., C.P. Ponting, and J. Hein. 2006. Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Comput Biol* **2:** e5.

Lunter, G., A. Rocco, N. Mimouni, A. Heger, A. Caldeira, and J. Hein. 2008. Uncertainty in homology inferences: Assessing and improving genomic sequence alignment. *Genome Research* **18:** 298-309.

Mills, R.E., C.T. Luttig, C.E. Larkins, A. Beauchamp, C. Tsui, W.S. Pittard, and S.E. Devine. 2006. An initial map of insertion and deletion (INDEL) variation in the human genome. *Genome Research* **16:** 1182-1190.

Morozov, P., T. Sitnikova, G. Churchill, F.J. Ayala, and A. Rzhetsky. 2000. A new method for characterizing replacement rate variation in molecular sequences: application of the fourier and wavelet models to Drosophila and mammalian proteins. *Genetics* **154:** 381-395.

Percival, D.B. and A.T. Walden. 2006. *Wavelet methods for time series analysis*. Cambridge U Press, NY.

Schmidt, T. and D. Frishman. 2008. Assignment of isochores for all completely sequenced vertebrate genomes using a consensus. *Genome Biology* **9:** R104.

Smit, A.F.A., R. Hubley, and P. Green. 1996-2004. RepeatMasker.

Spencer, C., P. Deloukas, S. Hunt, J. Mullikin, S. Myers, B. Silverman, P. Donnelly, D. Bentley, and G. McVean. 2006. The influence of recombination on human genetic diversity. *Public Library of Sciences Genetics* **2:** e148.

Thurman, R.E., N. Day, W.S. Noble, and J.A. Stamatoyannopoulos. 2007. Identification of higher-order functional domains in the human ENCODE regions. *Genome Research* **17:** 917-927.

Yuan, G.-C. and J.S. Liu. 2008. Genomic sequence is highly predictive of local nucleosome depletion. *Public Library of Sciences Computational Biology* **4:** e13.

Fig. S1

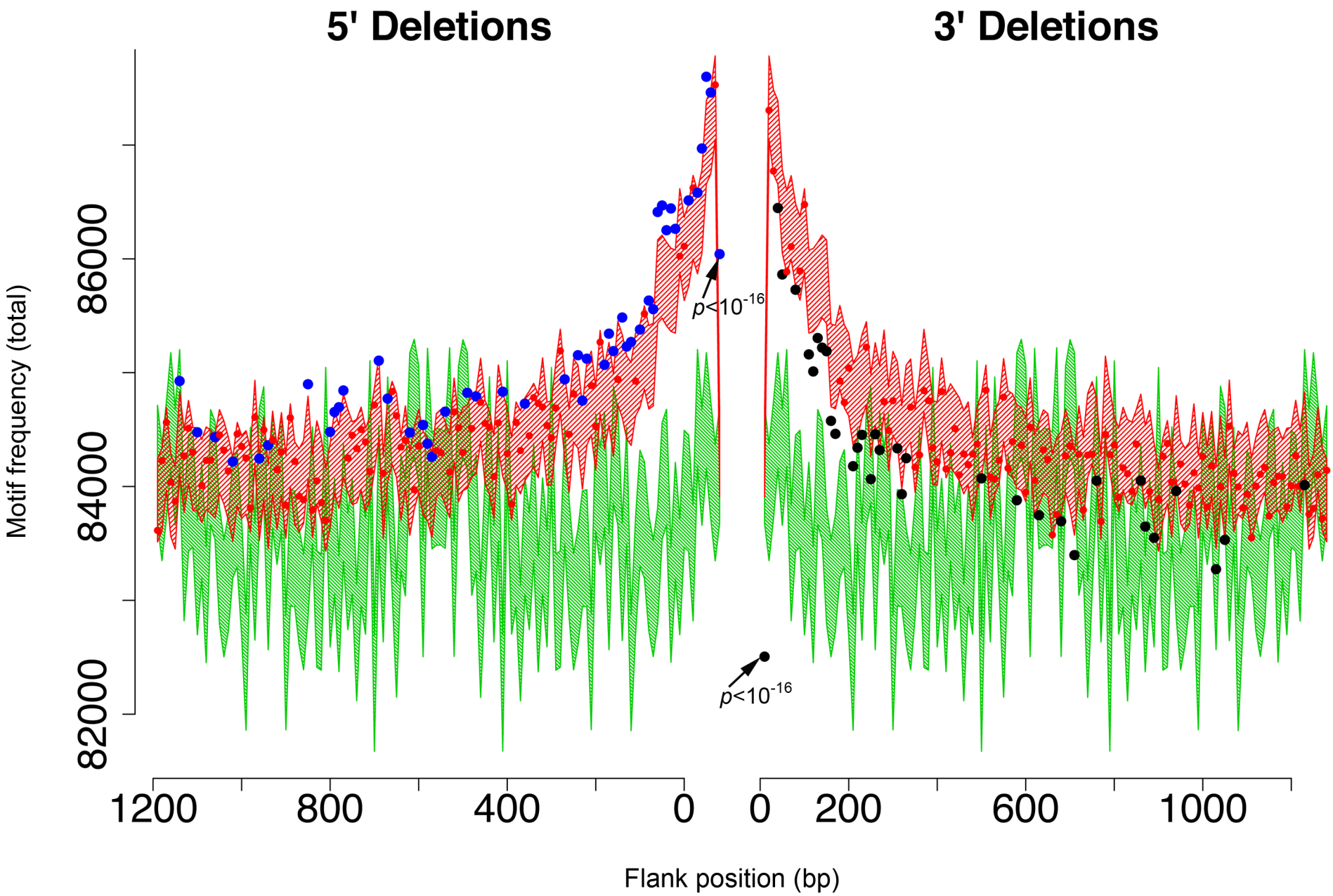

5' Deletions
3' Deletions
Motif frequency (total)
82000
84000
86000
1200
800
400
0
0
200
600
1000
Flank position (bp)
p<10-16
p<10-16

Fig. S2

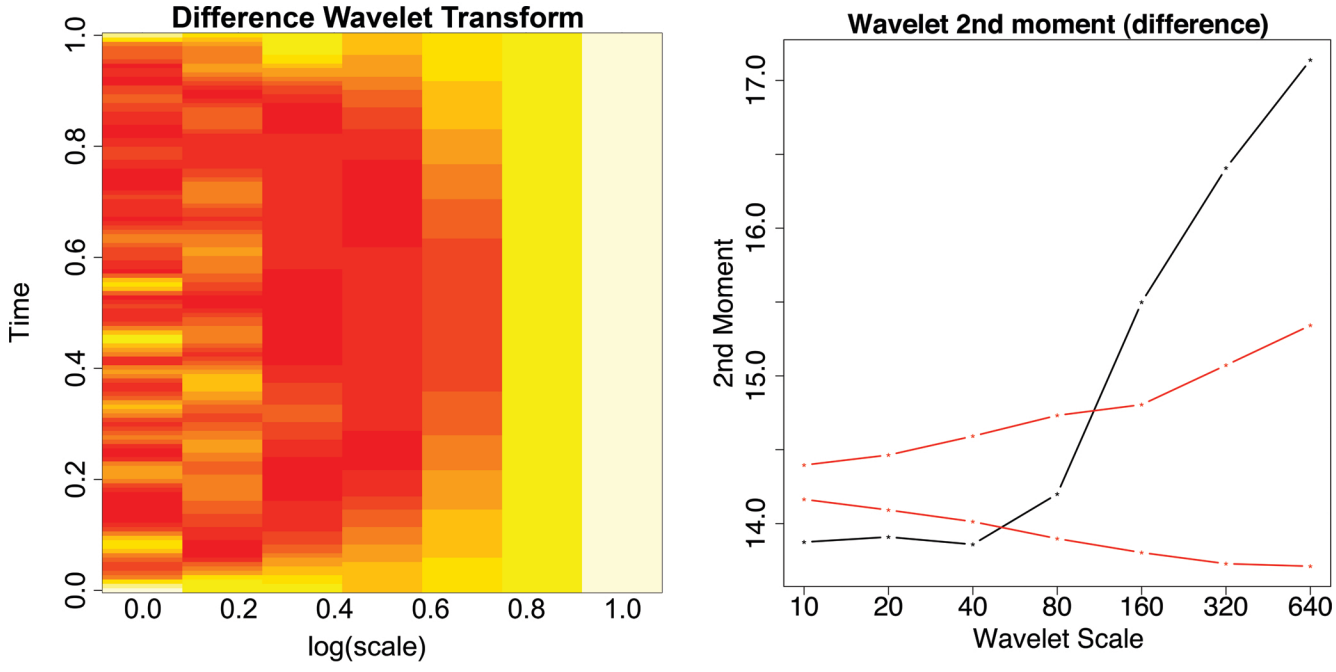

Difference Wavelet Transform
Time
log(scale)
0.0
0.2
0.4
0.6
0.8
1.0
Wavelet 2nd moment (difference)
2nd Moment
Wavelet Scale
14.0
15.0
16.0
17.0
10
20
40
80
160
320
640

Fig. S3

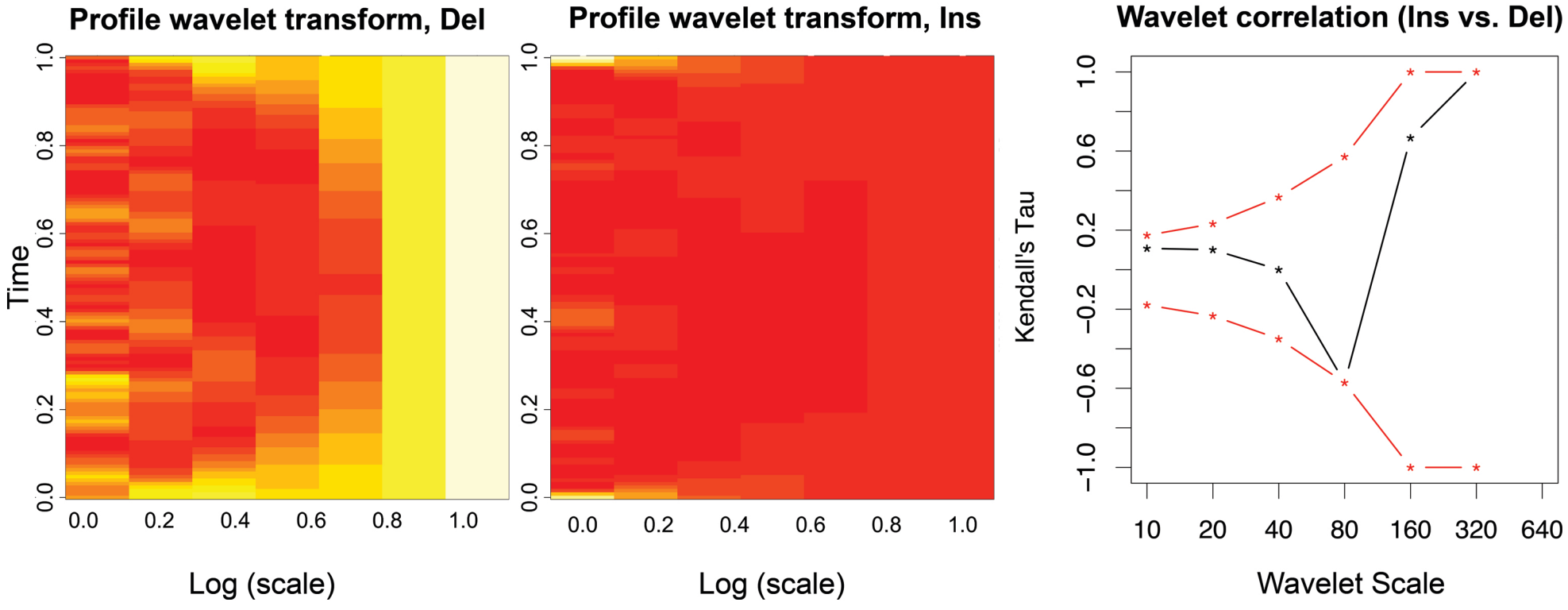
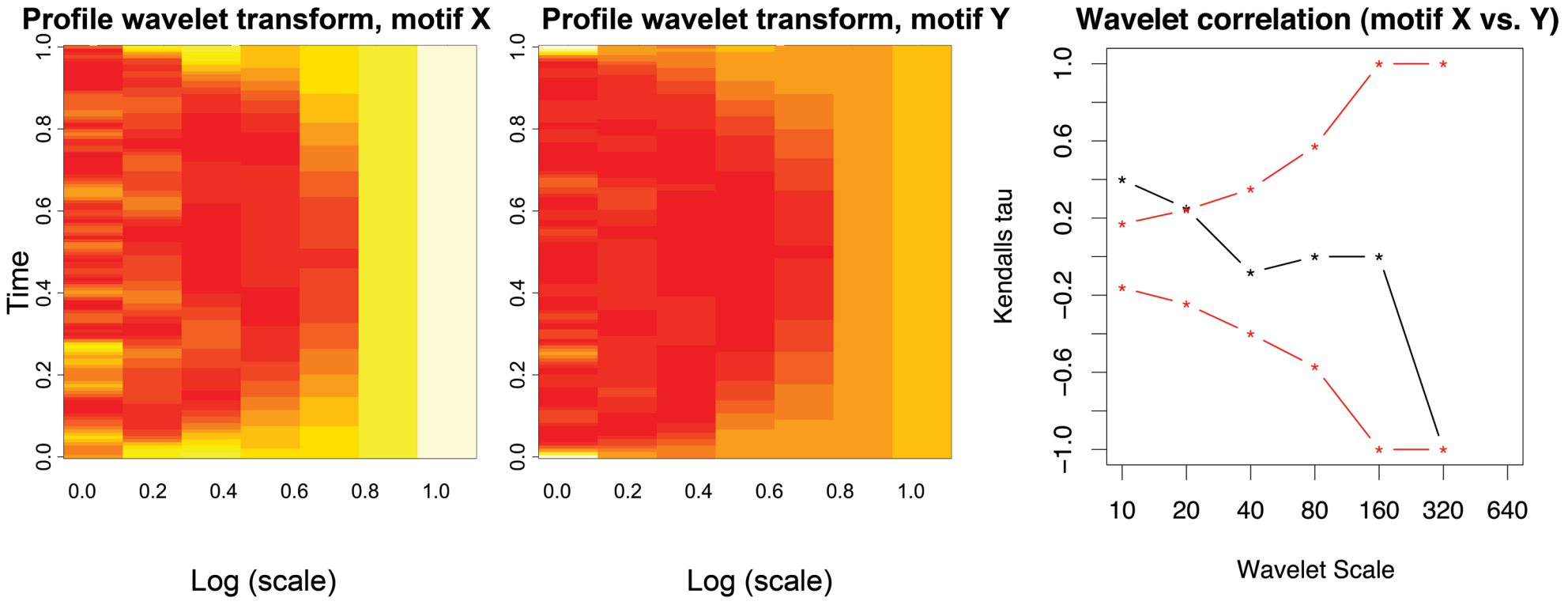

Profile wavelet transform, Del
Profile wavelet transform, Ins
Wavelet correlation (Ins vs. Del)
Time
Log (scale)
Log (scale)
Kendall's Tau
Wavelet Scale
0.0
0.2
0.4
0.6
0.8
1.0
−1.0
−0.6
−0.2
0.2
0.6
1.0
10
20
40
80
160
320
640

Fig. S4


Profile wavelet transform, motif X
Time
0.0
0.2
0.4
0.6
0.8
1.0
Log (scale)
Profile wavelet transform, motif Y
Log (scale)
Wavelet correlation (motif X vs. Y)
Kendalls tau
−1.0
−0.6
−0.2
0.2
0.6
1.0
10
20
40
80
160
320
640
Wavelet Scale