Supplementary Material

for J. A. Greenbaum, B. Pang and T. D. Tullius, Construction of a genome-scale structural map at single-nucleotide resolution

DNA templates and primers

The single-stranded random 40mer (R40) and pentamer DNA templates were purchased from GeneLink Inc. (Westchester, NY) and were purified by PAGE. All primers, including Cy5-labeled primers, were purchased from Integrated DNA Technology (Coralville, IA) and were purified by HPLC.

Primers (all shown 5' to 3') that were used for complementary strand synthesis and amplification prior to cloning were:

$R_{forward}$: TGTAACTGAAAATGGCTGAAGGT

$R_{reverse}$: TGTTGGATAAATGTCTTGTGGTG

To clone members of the R40 and pentamer libraries, PCR products were ligated and transformed into one-shot *E. coli* cells (Invitrogen) followed by plating and overnight incubation at 37 °C. After colonies were picked from the Petri dish, plasmid DNA was extracted using a Qiagen Maxiprep kit.

The $R_{forward}$ and $R_{reverse}$ primers were used for amplifying an individual member of the R40 library prior to treatment with hydroxyl radical.

The following primers were used for amplification of individual pentamer library members prior to treatment with hydroxyl radical:

$P_{forward}$: GAATTAACCCTCACTAAAGGGACT

$P_{reverse}$: CACTATAGGGCGAATTGAATTTAG

Algorithm for generating a sequence containing all possible pentanucleotides

In order to design a sequence that contained all 1024 possible pentanucleotide sequences exactly one time (see Figure S1), the following algorithm was implemented:

1. Start with a random tetranucleotide sequence.

2. Add a nucleotide to the 3' end.

3. If the resulting pentanucleotide has already been incorporated into the forward or reverse strand, remove the added nucleotide and iterate to the next one (e.g., if NNNNA has been incorporated, then try NNNNC, and so on).

4. If NNNNA, NNNNC, NNNNG, and NNNNT have already been incorporated, go to step 1.

Repeat steps 2 and 3 until all possible pentamers have been incorporated.

The pentamer templates were shorter than the R40 template by approximately 40 nucleotides, so we designed new PCR primers ($P_{forward}$ and $P_{reverse}$, see above) that were complementary to sequences in the vector rather than the insert sequence. This resulted in a longer PCR product, 220 bp, and made the pentamer sequence fragments the proper size to be accurately resolved and quantified.

PCR amplification for hydroxyl radical cleavage

Amplification of DNA for use in the hydroxyl radical cleavage reaction required two sets of primers. Each primer set contained one Cy5-labeled primer and one unlabeled primer. Each library member was amplified with both primer sets, in separate reactions, in order to obtain the cleavage pattern of both strands. The PCR reaction yielded duplex DNA molecules 158 bp in length for the N40 library, and 220 bp in length for the pentamer library. PCR products were loaded onto a 3% NuSieve agarose gel (8 cm x 7 cm) and run for 1.5 hr at 100 V in 1X TAE buffer. The gel box was covered with aluminum foil during the run, to minimize degradation of the Cy5 dye. Following electrophoresis the gel was scanned using a Storm 860 scanner (Molecular Dynamics) on the "Red fluorescence" setting. Using the scan as a template, the PCR product was excised from the gel and purified using the Quantum Prep Freeze 'N Squeeze DNA Gel Extraction Spin Column (Bio-Rad). Following gel purification, 600 μl of 100% ethanol and 30 μl of 3 M sodium acetate were added to the sample. The sample was stored at -80 °C for at least one hour. DNA was pelleted by centrifugation at 14,000 rpm for 30 min at 4 °C in an Eppendorf 5417C refrigerated microcentrifuge. The supernatant was removed and the pellet was washed with 400 μl of 70% ethanol. The sample was centrifuged at 14,000 rpm for 5 min and the supernatant was removed. The pellet was dried in a DNA120 Speed Vac Concentrator (Savant) for 15 min. The dry pellet was dissolved in 70 μl water. 1 μl was spotted onto Whatman 3MM filter paper and the spot was scanned with the Storm 860 in order to verify the presence of labeled DNA.

Electrophoretic separation of hydroxyl radical-treated DNA

A 10% photo-initiating acrylamide solution (Zaxis or Amresco) was injected into a Visible Genetics Long-Read gel plate and polymerized. The DNA sample (2 μl) was loaded on the gel, leaving an empty lane between samples. The gel was run for 270 min at 2000 V and 60 °C. The laser was set to 50% power and a sampling rate of 1 sec.

Quantitation of cleavage data

The electrophoresis apparatus used to separate product of hydroxyl radical cleavage measures the fluorescence intensity of the eluate as it flows past a detector and stores this information in a text file. Each lane of the gel is represented as a series of fluorescence data points that form peaks, with each peak corresponding to one cleavage product. The area of each peak represents the amount of cleavage product of a particular length. Because reaction of the hydroxyl radical with a nucleotide in a DNA strand leads to destruction of that nucleotide, the area of a peak in the cleavage pattern in fact represents the extent of reaction of the hydroxyl radical with the nucleotide immediately to its 3' side.

To approximate the instrumental background of the sequencer, we subtract from each data point the average fluorescence intensity of the eluate before the first DNA fragment reaches the detector. We quantify peak areas by using the computer program PeakFit to simultaneously fit 86 peaks (the 40 central nt, 36 nt of common sequence, and 5 nt on either side) of a background-subtracted dataset. Fitting the peaks in the raw data effectively reduces the dataset from a fluorescence trace with over 15,000 data points to an array of 86 peak areas. Fitting also deconvolutes closely spaced peaks in the electrophoretogram, allowing for accurate measurement of peak areas (Shadle et al. 1997). We tested more than 30 functions for their ability to accurately model the experimental data, by fitting a lane that contained only one peak. We began with the Lorentzian function, since this lineshape had been found to be suitable for fitting cleavage data for a radiolabeled sample electrophoresed on a conventional denaturing gel (Shadle et al. 1997). However, we found that the Lorentzian function does a poor job in approximating the shape of the fluorescence curve that is generated by the apparatus we used for these experiments. We eventually found that the Exponentially-Modified Gaussian + Half-Modified Gaussian (EMG + GMG) function included in the PeakFit program proved to be the most accurate for fitting these fluorescence curves, with an R^2 value of 0.999.

Normalization of cleavage data

We use the cleavage patterns of the common flanking sequences in each template (see Fig. 1) to normalize the datasets, so that cleavage patterns for different members of the library can be quantitatively compared. To accomplish this, the areas of the six peaks representing the common sequence AATTCG were summed separately for the flanking common palindromic sequences to the 5' and 3' side of the test sequence. We linearly interpolate these summed areas, and then divide the area of each peak in between by the interpolated area at that position (equation 1).

$$N_i = \frac{A_i}{C_i} \qquad \text{(eq 1)}$$

Here, N_i is the normalized area of the peak at position i, A_i is the experimentally measured area of the peak at position i, and C_i is the interpolated area of the common sequence sum at position i. The normalized peak areas are further adjusted to represent a Z-score relative to the distribution of cleavage intensities in the common sequences. Figure 2 shows that this method yields normalized peak areas with an average standard deviation of 0.19.

Database creation and population

Samples were imported into the ORChID database and assigned a sample ID (sid). The sid consists of 5 digits, with the first digit, either 1 or 2, indicating either the forward strand or reverse complement, respectively. The 2nd digit is either 0 for an R40 template, or a 5 for a pentamer template. The third digit, 0 through 9, specifies the number of times the same sequence appears in the database. The fourth and fifth digits, both 0 through 9, indicate the sample number used in sequencing and cleavage experiments.

Table S1. Template sequences. Listed are the spacer sequences and flanking regions for each DNA template, as diagrammed in Figure 1. To construct the final template, the test sequence was placed between the 5' and 3' spacer/flanking regions. Indicated by bold typeface are the common palindromic dodecanucleotide sequences in the flanking regions (see Fig. 1) that are used for normalization. The length of 118 in parentheses for the Pentamer template represents the single pentamer sequence that is shorter than the others.

	R40 Template
5' spacer/flanking region	TGTAACTGAAAATGGCTGAAGGTACAGACCCTTTAGATCACTAT**CGCGAATTCGCG**ATA
3' spacer/flanking region	TAT**CGCGAATTCGCG**ATACACGAAAACGCAGGGCTGCACCACAAGACATTTATCCAACA
Total Length (including insert)	158
	Pentamer template
5' spacer/flanking region	TGTAACTGAAAATGGCTGAAGGTTAT**CGCGAATTCGCG**ATA
3' spacer/flanking region	TAT**CGCGAATTCGCG**ATACACCACAAGACATTTATCCAACA
Total Length (inclusing insert)	123 (118)

Table S2. *Trimer*s view. All of the trimers in the ORChID database are listed in this view, along with their positions in the sample's sequence and their peak areas.

Column	**Data type**	**Description**
sid	smallint	Sample ID
position	smallint	Position of the trimer sequence within the DNA sequence of the sample that is specified by sid
trimer	character(3)	The trinucleotide sequence
pk1	real	Normalized peak area of the 1st peak in the trimer
pk2	real	Normalized peak area of the 2nd peak in the trimer
pk3	real	Normalized peak area of the 3rd peak in the trimer

Table S3. *Trimer summary* view. All 64 trimers are grouped by sequence. The number of times each trimer occurs in the database is recorded, as well as the mean and standard deviation of the area of each peak of a particular trimer.

Column	Data type	Description
trimer	character(3)	The trimer sequence
count	bigint	Number of times the trimer sequence appears in the ORChID database
pk1 mean	numeric	Mean normalized peak area of the first nucleotide in all instances of the trimer in the database
pk2 mean	numeric	Mean normalized peak area of the second nucleotide in all instances of the trimer in the database
pk3 mean	numeric	Mean normalized peak area of the third nucleotide in all instances of the trimer in the database
pk1 sd	numeric	RMSD of the normalized peak areas of the first nucleotide in all instances of the trimer in the database
pk2 sd	numeric	RMSD of the normalized peak areas of the second nucleotide in all instances of the trimer in the database
pk3 sd	numeric	RMSD of the normalized peak areas of the third nucleotide in all instances of the trimer in the database

Table S4. Correlation of sequence identity and cleavage pattern similarity. Nmers of length 10, 20, and 30 were extracted from the ORChID database and compared with against one another for their degree of similarity in hydroxyl radical cleavage pattern. This metric was compared with the sequence similarity of the Nmers, and a Pearson correlation coefficient was calculated. Here, n represents the number of pairwise correlations calculated. Note the similar correlation at all window sizes.

Window size	n	Correlation
10	3,122,288	0.35
20	1,456,848	0.37
30	418,608	0.36

Table S5. Correlation of cleavage pattern at various levels of sequence identity, for 10mers. The data summarized in Table S4 were binned according to the percent sequence identity (%ID), and the Pearson coefficient for cleavage *vs.* sequence identity. The entry in each cell indicates the number of pairs of 10mer sequences that have a percent identity and Pearson coefficient greater than or equal to the row and column headings, respectively. Note the significant presence of sequence pairs having low sequence identity, but high similarity of cleavage pattern. Conversely, there also exist sequence pairs having high sequence identity but poor cleavage pattern similarity.

	Pearson coefficient									
%ID	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
0	12253	9627	7513	5332	3426	1981	983	320	61	4
10	29721	25172	20207	14681	9812	6248	3274	1327	355	36
20	44165	39637	32942	25882	18538	11714	6561	2632	738	37
30	35382	33934	30996	26192	19611	13175	7818	3365	926	56
40	24670	26009	25192	22960	18751	13801	8597	4325	1270	88
50	9597	11171	11547	11516	10379	8452	6116	3215	1072	117
60	3500	4248	5047	5436	5648	5216	4358	2780	1280	173
70	587	753	962	1192	1480	1673	1704	1353	599	103
80	121	214	292	445	648	842	971	950	661	153

Table S6. Correlation of cleavage pattern at various levels of sequence identity, for 20mers. See the legend to Table S5.

	Pearson coefficient									
%ID	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
0	6219	4438	2889	1520	571	128	2	0	0	0
10	23786	18277	12113	6675	3065	1141	258	35	3	0
20	31766	26376	19075	11935	5838	2194	520	86	2	0
30	24960	23707	19775	13333	7510	2998	849	135	11	0
40	12046	12975	12153	9881	6819	3340	1157	228	11	0
50	3031	4022	4831	4787	3785	2403	1028	241	30	0
60	548	865	1130	1348	1330	1149	623	258	64	0
70	70	101	163	307	389	412	342	113	16	0
80	0	12	40	86	105	194	156	112	54	1

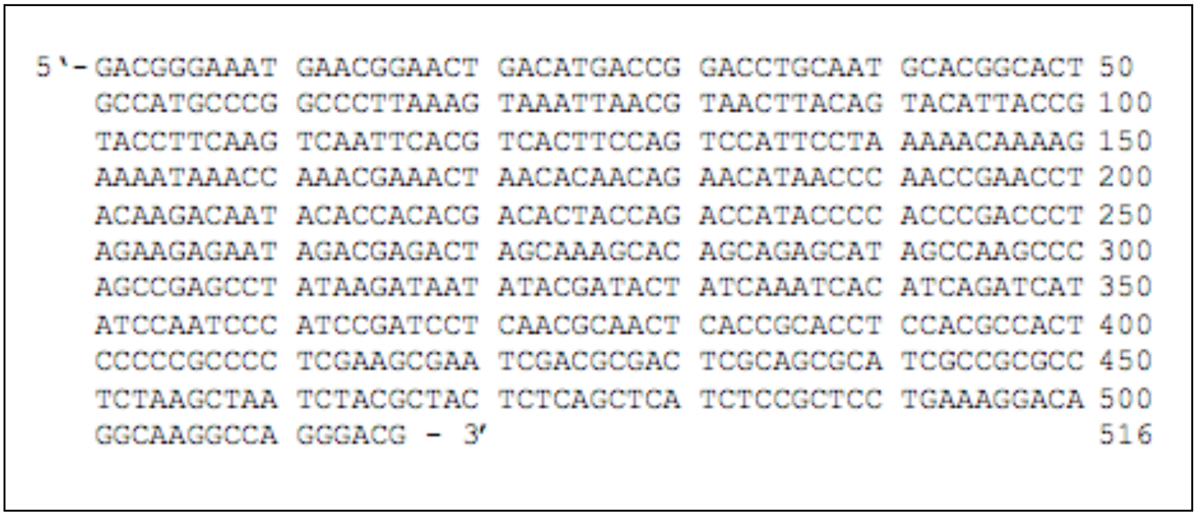
Table S7. Correlation of cleavage pattern at various levels of sequence identity, for 30mers. See the legend to Table S5.

	Pearson coefficient									
%ID	**0**	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**
0	1965	1293	706	303	108	15	5	0	0	0
10	8942	6177	3566	1564	603	109	3	0	0	0
20	11547	9212	6015	3022	1048	256	27	0	0	0
30	8619	8472	6199	3452	1552	411	71	14	0	0
40	3137	3760	3603	2695	1432	477	108	28	1	0
50	488	820	1052	1056	681	372	133	19	0	0
60	98	189	213	219	254	202	120	30	0	0
70	0	17	21	40	95	76	30	9	0	0
80	0	0	6	5	9	11	36	11	10	0

Table S8. Sequences of DNA molecules depicted in Figures showing similarity of hydroxyl radical cleavage patterns of divergent sequences. The Start and End values and sid (sequence ID) are listed for the four sequences in the ORChID database that were used for the cleavage pattern comparisons shown in Fig. 4 and Fig. S3. Positions of sequence identity are indicated by boldface.

Figure	sid	Start	End	Sequence
4	25205	32	41	5′-TTTATTTTCT-3′
4	15105	14	23	5′-AACGAAACTA-3′
S3	25111	14	33	5′-TG**G**CGTGGAGGT**G**CGGTGAG-3′
S3	15003	14	33	5′-CA**G**TACATTACC**G**TACCTTC-3′

Figure S1



```
5'-GACGGGAAAT GAACGGAACT GACATGACCG GACCTGCAAT GCACGGCACT 50
   GCCATGCCCG GCCCTTAAAG TAAATTAACG TAACTTACAG TACATTACCG 100
   TACCTTCAAG TCAATTCACG TCACTTCCAG TCCATTCCTA AAAACAAAAG 150
   AAAATAAACC AAACGAAACT AACACAACAG AACATAACCC AACCGAACCT 200
   ACAAGACAAT ACACCACACG ACACTACCAG ACCATACCCC ACCCGACCCT 250
   AGAAGAGAAT AGACGAGACT AGCAAAGCAC AGCAGAGCAT AGCCAAGCCC 300
   AGCCGAGCCT ATAAGATAAT ATACGATACT ATCAAATCAC ATCAGATCAT 350
   ATCCAATCCC ATCCGATCCT CAACGCAACT CACCGCACCT CCACGCCACT 400
   CCCCCGCCCC TCGAAGCGAA TCGACGCGAC TCGCAGCGCA TCGCCGCGCC 450
   TCTAAGCTAA TCTACGCTAC TCTCAGCTCA TCTCCGCTCC TGAAAGGACA 500
   GGCAAGGCCA GGGACG - 3'                                516
```

Figure S1. The minimal length pentamer sequence. This sequence of 516 nucleotides contains all 1024 pentamer sequences when read in the forward and reverse complement directions.

Figure S2

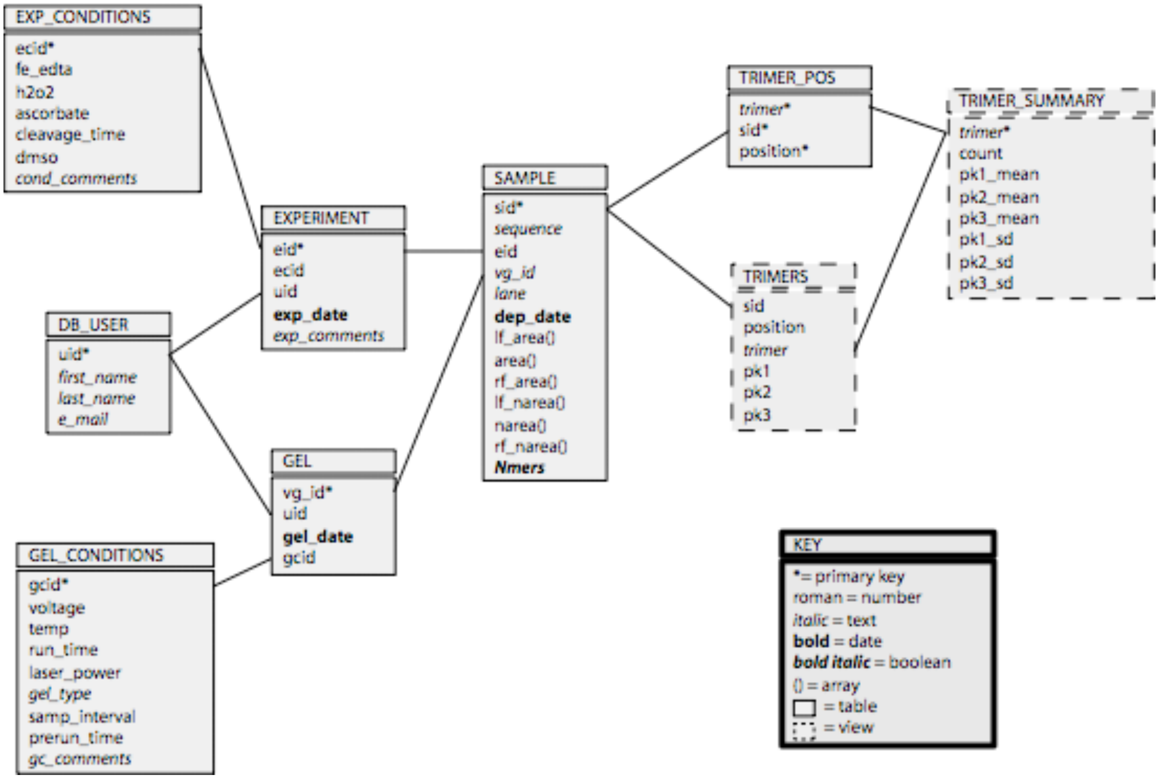


Figure S2. Database schema. Each rectangle in this diagram represents one of the tables in the ORChID database. The name of the table is at the top of the box, and the rows below contain the attributes of the table. The font style of each attribute corresponds to the type of data that it holds, as described in the key. The links between tables represent foreign keys used to cross-reference each table as well as to enforce referential integrity constraints. Note that in addition to the trimers and trimer summary views, there are views corresponding to Nmers of 2 through 7.

Figure S3

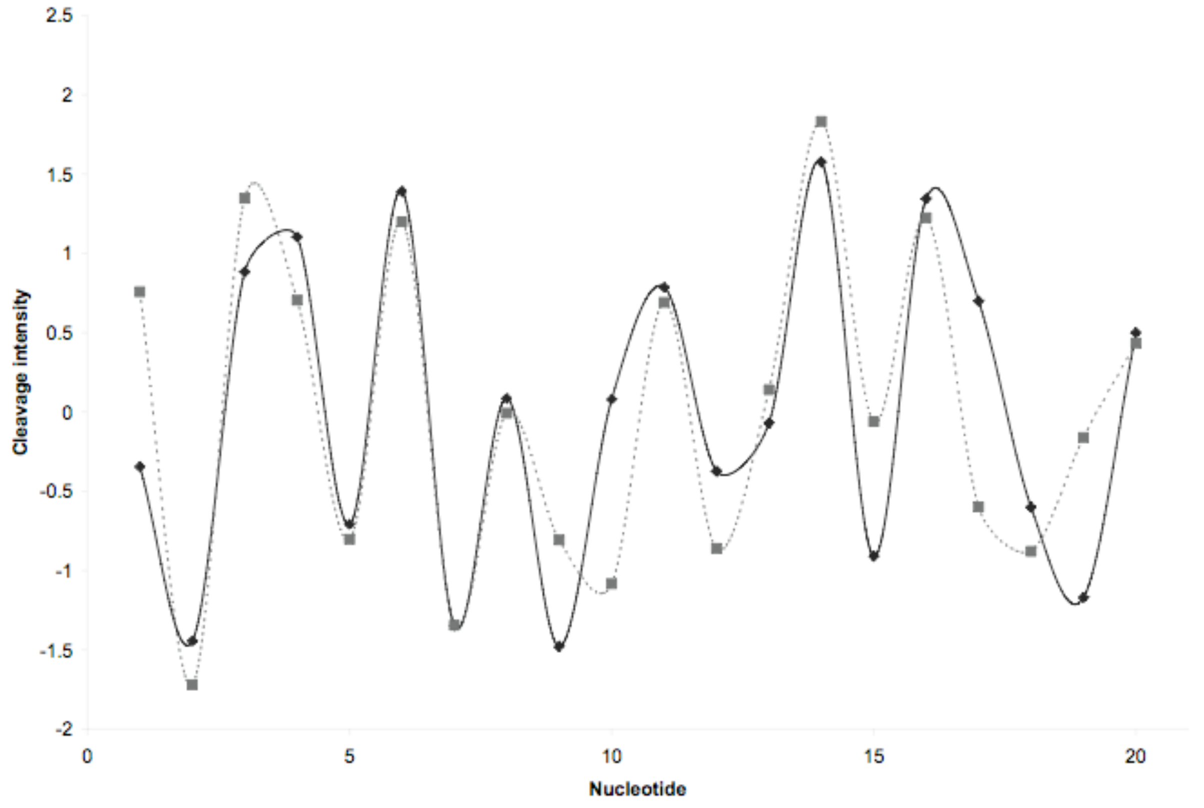


Figure S3: Low sequence identity/high cleavage similarity of 20mer sequences. Plotted are the hydroxyl radical cleavage patterns of two 20mer sequences with 10% sequence identity. Note the significant correlation (R=0.81) of the two patterns throughout most of the plot, even with such low sequence identity. (See Table S8 for sequences.)

Figure S4

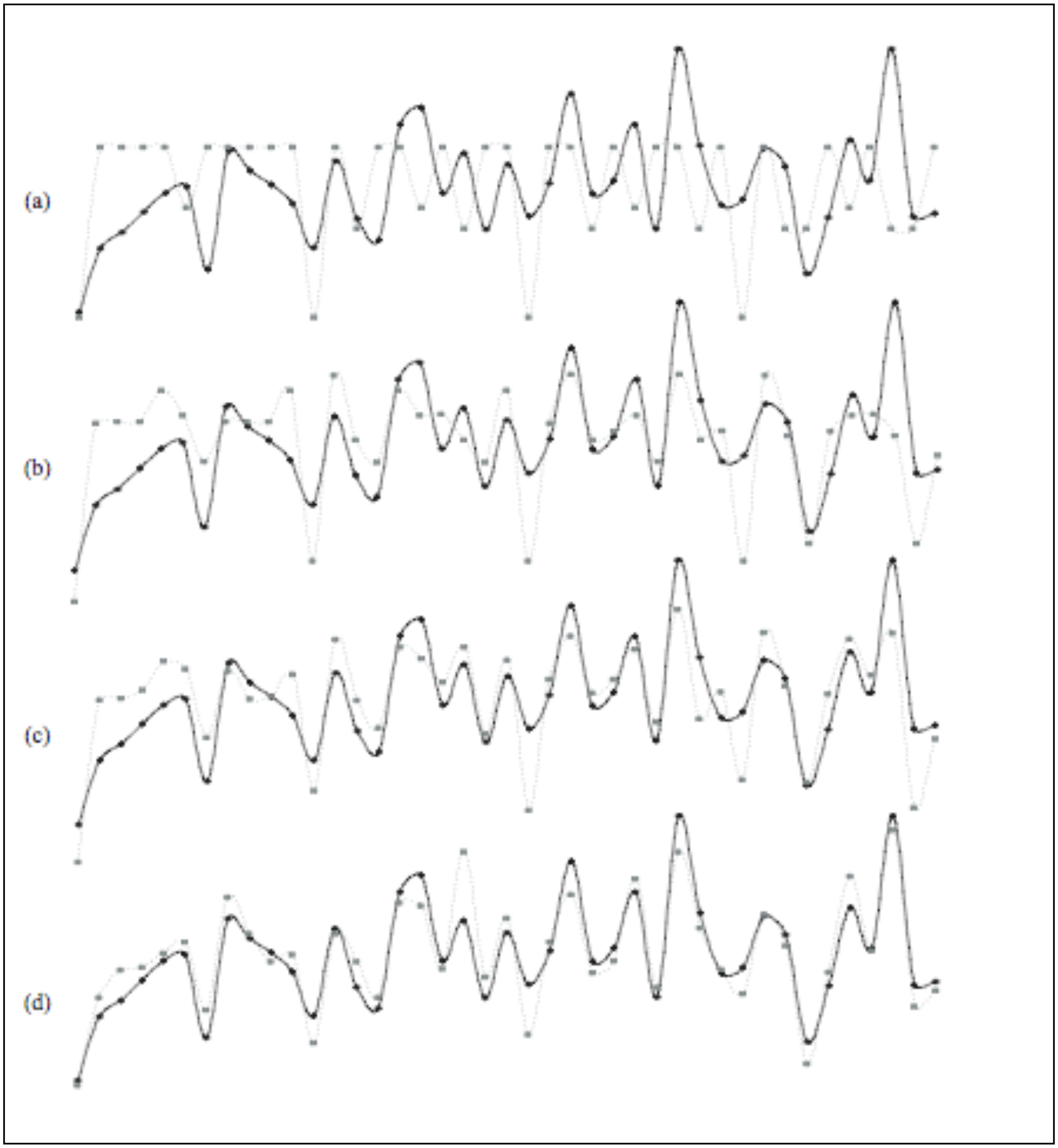


Figure S4: Prediction of hydroxyl radical cleavage intensity using different sliding Nmer window algorithms. The hydroxyl radical cleavage pattern of sample ID 25211 in the ORChID database was predicted using four different sliding Nmer algorithms, to illustrate the improvement in correlation as the model increases in complexity. (a) Monomer window, R=0.18; (b) Dimer window, R=0.61; (c) Trimer window, R=0.77; (d) Tetramer window, R=0.92.

REFERENCES

Shadle, S. E., Allen, D. F., Guo, H., Pogozelski, W. K., Bashkin, J. S., and Tullius, T. D. 1997. Quantitative analysis of electrophoresis data: novel curve fitting methodology and its application to the determination of a protein-DNA binding constant. *Nucleic Acids Res*. **25:** 850-860.