Supplementary Information

1. Cloning vector pGIS4a for GIS-PET library

2. Cloning vector pGIS3h for ChIP-PET library

3. Mapping of GIS-PET identified transcripts

4. Distribution of homopolymer errors in GIS-PETs

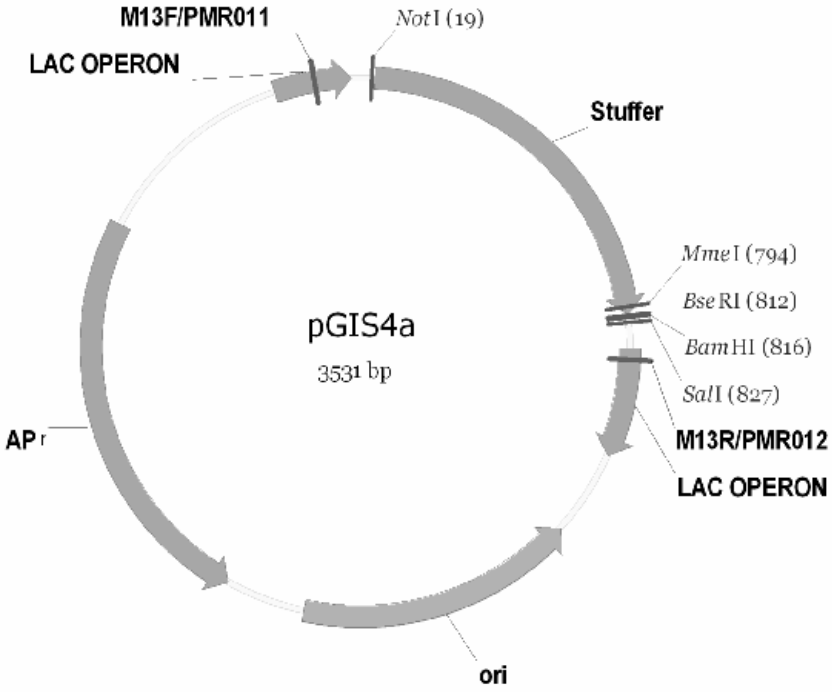
5. Transcripts identified by MS-PET analysis of transcripts

6. Reduction of noise from MS-PET analyzed ChIP-PET clusters

7. List of references used in Supplementary Information

1. Cloning vector pGIS4a for GIS-PET library

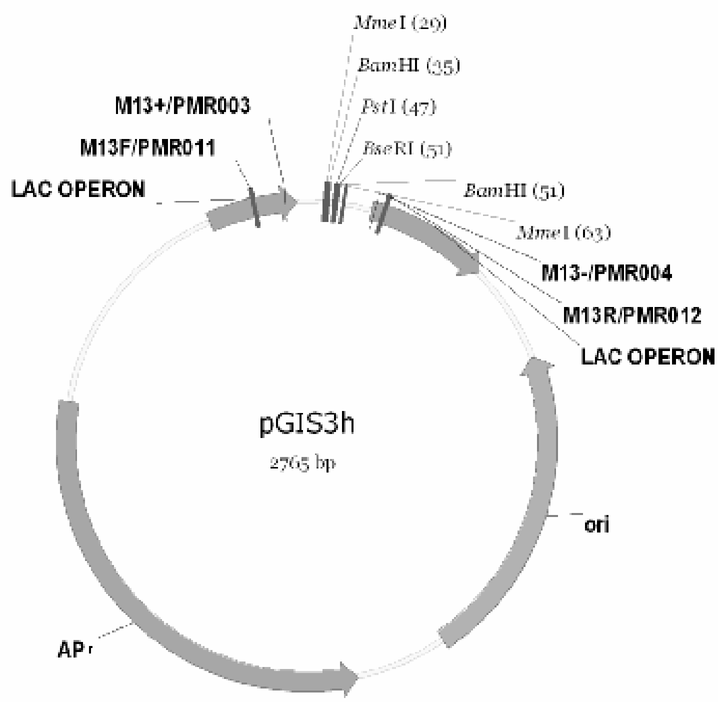
Supplementary Figure 1. Cloning vector pGIS4a and relevant vector sequence.
The pGIS4a vector is designed for flcDNA cloning. Sequential BseRI and BamHI digestion releases an asymmetric PET that can be subsequently dimerized into diPETs for MS-PET sequencing analysis.




```
                      NotI
                     ~~~~~~~~
   1  GGGCGAATTC GATATCGCGG CCGCGCCTGG ATAAAGTCAG CAGCTTCCAC GCCAGCTTCA CACAAAAAGT GACTGACGGT AGCGGCGCGG CGGTGCAGGA
      CCCGCTTAAG CTATAGCGCC GGCGCGGACC TATTTCAGTC GTCGAAGGTG CGGTCGAAGT GTGTTTTTCA CTGACTGCCA TCGCCGCGCC GCCACGTCCT
 101  AGGTCAGGGC GATCTGTGGG TGAAACGTCC AAACTTATTC AACTGGCATA TGACACAACC TGATGAAAGC ATTCTGGTTT CTGACGGTAA AACACTGTGG
      TCCAGTCCCG CTAGACACCC ACTTTGCAGG TTTGAATAAG TTGACCGTAT ACTGTGTTGG ACTACTTTCG TAAGACCAAA GACTGCCATT TTGTGACACC
 201  TTCTATAACC CGTTCGTTGA GCAAGCTACG GCAACCTGGC TGAAAGATGC CACCGGTAAT ACGCCGTTTA TGCTGATTGC CCGCAACCAG TCCAGCGACT
      AAGATATTGG GCAAGCAACT CGTTCGATGC CGTTGGACCG ACTTTCTACG GTGGCCATTA TGCGGCAAAT ACGACTAACG GGCGTTGGTC AGGTCGCTGA
 301  GGCAGCAGTA CAATATCAAA CAGAATGGCG ATGACTTTGT CCTGACGCCG AAAGCCAGCA ATGGCAATCT GAAGCAGTTC ACCATTAACG TGGGACGTGA
      CCGTCGTCAT GTTATAGTTT GTCTTACCGC TACTGAAACA GGACTGCGGC TTTCGGTCGT TACCGTTAGA CTTCGTCAAG TGGTAATTGC ACCCTGCACT
 401  TGGCACAATC CATCAGTTTA GCGCGGTGGA GCAGGACGAT CAGCGCAGCA GTTATCAACT GAAATCCCAG CAAAATGGGG CTGTGGATGC AGCGAAATTT
      ACCGTGTTAG GTAGTCAAAT CGCGCCACCT CGTCCTGCTA GTCGCGTCGT CAATAGTTGA CTTTAGGGTC GTTTTACCCC GACACCTACG TCGCTTTAAA
 501  ACCTTCACCC CGCCGCAAGG CGTCACGGTA GATGATCAAC GTAAGTAGAG GCACCTGAGT GAGCAATCTG TCGCTCGATT TTTCGGATAA TACTTTTCAA
      TGGAAGTGGG GCGGCGTTCC GCAGTGCCAT CTACTAGTTG CATTCATCTC CGTGGACTCA CTCGTTAGAC AGCGAGCTAA AAAGCCTATT ATGAAAAGTT
 601  CCTCTGGCCG CGCGTATGCG GCCAGAAAAT TTAGCACAGT ATATCGGCCA GCAACATTTG CTGGCTGCGG GGAAGCCGTT GCCGCGCGCT ATCGAAGCCG
      GGAGACCGGC GCGCATACGC CGGTCTTTTA AATCGTGTCA TATAGCCGGT CGTTGTAAAC GACCGACGCC CCTTCGGCAA CGGCGCGCGA TAGCTTCGGC
 701  GGCATTTACA TTCTATGATC CTCTGGGGGC CGCCGGGTAC CGGCAAAACA ACTCTCGCTG AAGTGATTGC CCGCTATGCG AACGCTGATG TGGAACGTAT
      CCGTAAATGT AAGATACTAG GAGACCCCCG GCGGCCCATG GCCGTTTTGT TGAGAGCGAC TTCACTAACG GGCGATACGC TTGCGACTAC ACCTTGCATA
                    MmeI   BseRI
                   ~~~~~~  ~~~~~~~
                      BamHI        SalI
                      ~~~~~~~     ~~~~~~~
 801  TTCTGCCGTA AGTCGGATCC TCCTCGTCGA CCTGCAGGCA TGCAAGCTTG AGTATTCTAT AGTGTCACCT AAATAGCTTG GCGTAATCAT GGTCATAGCT
      AAGACGGCAT TCAGCCTAGG AGGAGCAGCT GGACGTCCGT ACGTTCGAAC TCATAAGATA TCACAGTGGA TTTATCGAAC CGCATTAGTA CCAGTATCGA
```

2. Cloning vector pGIS3h for ChIP-PET library

Supplementary Figure 2. Cloning vector pGIS 3h and relevant vector sequence. The pGIS3h vector is used for ChIP-PET library construction to identify transcription factor binding sites. Digestion with BseRI followed by alkaline phosphatase treatment, and BamHI digestion, releases an asymmetric PET that can subsequently be dimerized via the BamHI cohesive site for diPET construction and MS-PET analysis.




```
                                            MmeI      MmeI
                                          ~~~~~~~   ~~~~~~~
                                     BamHI     PstI    BamHI     BseRI
                                     ~~~~~~    ~~~~~~  ~~~~~~~   ~~~~~~~
1   GGGCGAATTC GATATCGCGG CCGCGAGGAT TATGGATCCG ACTGCAGTCG GATCCATACT CCTCATTGCA GGCATGCAAG CTTGAGTATT CTATAGTGTC
    CCCGCTTAAG CTATAGCGCC GGCGCTCCTA ATACCTAGGC TGACGTCAGC CTAGGTATGA GGAGTAACGT CCGTACGTTC GAACTCATAA GATATCACAG
```

3. Mapping of GIS-PET identified transcripts

Supplementary Table 1. Transcript mapping results of MS-PET analyzed MCF7 GIS-PET library
The table provides an overview of the PET mapping statistics for the library without homopolymer error analysis.
"N.A.", not available.

	PETs	% of total PETs	% of mapped PETs	PET clusters
Total PETs not mapped*	156,286	49.78 %	N.A.	N.A.
Total PETs mapped	157,697	50.22%	100.00%	22,992
mapped once (1)	136,612	43.51%	86.63%	20,864
mapped (2)	8,793	2.80%	5.58%	2,218
mapped (3)	2,929	0.93%	1.86%	907
mapped (4)	1,883	0.60%	1.19%	532
mapped (5)	627	0.20%	0.40%	320
mapped (6)	377	0.12%	0.24%	235
mapped (7)	271	0.09%	0.17%	214
mapped (8)	268	0.09%	0.17%	108
mapped (9)	112	0.04%	0.07%	120
mapped (10)	163	0.05%	10.00%	83
mapped (>10)	5,662	1.80%	3.59%	N.A.
Total PETs	**313,983**	**100.00%**	**N.A.**	**N.A.**

* Before homopolymer error analysis

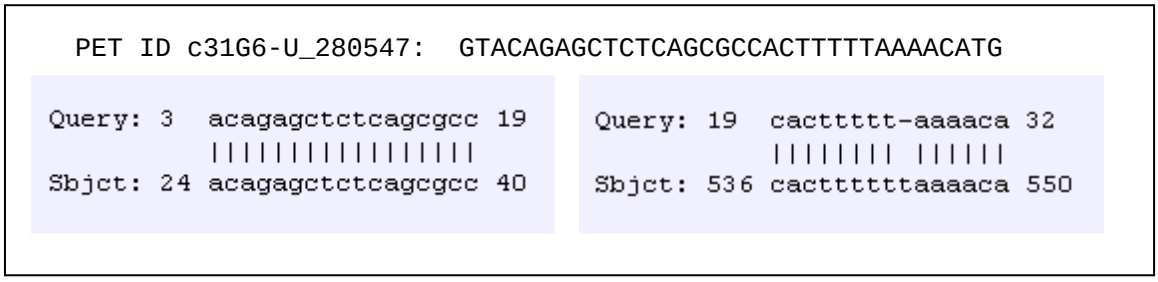
Supplementary Table 2. Transcript mapping results of Sanger capillary sequenced MCF7 GIS-PET library

	PETs	% of total PETs	% of mapped PETs	PET clusters
Total PETs not mapped	33,097	24.38%	N.A.	N.A.
Total PETs mapped	102,660	**75.62%**	100.00%	12,996
mapped once (1)	92,928	68.45%	90.52%	11,513
mapped (2)	4,691	3.46%	4.57%	1,641
mapped (3)	1,311	0.97%	1.28%	616
mapped (4)	754	0.56%	0.73%	363
mapped (5)	332	0.24%	0.32%	271
mapped (6)	190	0.14%	0.19%	172
mapped (7)	119	0.09%	0.12%	144
mapped (8)	61	0.04%	0.06%	96
mapped (9)	29	0.02%	0.03%	81
mapped (10)	32	0.02%	0.03%	52
mapped (>10)	0	0.00%	0.00%	0
Total PETs	**135,757**	**100.00%**	**N.A.**	**N.A.**

4. Distribution of homopolymer errors

The analysis of the distribution of putative homopolymer errors differs from the procedure used to recover originally unmapped PETs described in the main text. In the recovery procedure, the pool of 156,286 originally unmapped PETs was re-mapped by allowing a single 1-base deletion within homopolymer regions, retrieving PETs that mapped to single genomic loci under these conditions, and subsequently re-mapping the remaining PETs by allowing a single 1-base insertion. This sequential recovery procedure ensured maximal recovery of PETs (56,914 PETs recovered), and avoided the inadvertent duplicated recovery of PETs that might contain both overcall and undercall errors.

However, for the error distribution analysis procedure, we were interested solely in examining the spread of either overcalls or undercalls across regions of various homopolymer lengths. To ensure as full a coverage as possible of all apparent errors, we decided to retrieve overcall and undercall errors separately instead of sequentially. Accordingly, from the 156,286 unmapped PETs, by allowing a 1-base deletion and re-mapping, we recovered a set of 35,523 PETs (22.73%) that were putative overcall errors. Similarly, by subjecting the entire pool of 156,286 unmapped PETs to re-mapping after allowing a 1-base insertion, we recovered a separate set of 27,047 PETs (17.31%) containing putative undercall errors. Thus, overcall errors appeared to predominate. An example of a homopolymer error is shown in Supplementary Figure 3.



Supplementary Figure 3. An example of an apparent undercall error within a homopolymer region. In this example, PET ID c31G6-U_280547 (the "Query") should have mapped to its target (the "Subject"; transcript BC001410, chr1:148818057-148822521) except for an extraneous 1-base deletion in the 3' signature of the PET sequence (within a 6-base T-homopolymer site). It was subsequently recovered by allowing a 1-base insertion.

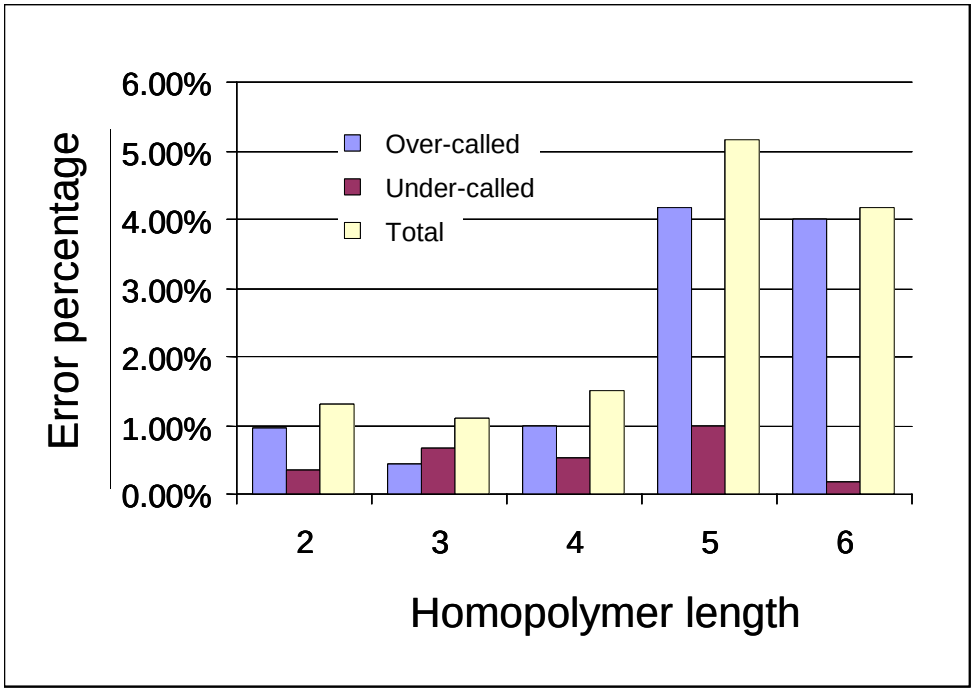
Next, we wanted to determine the distribution of homopolymer sequencing errors. For accuracy, we focused only on a subset of high-copy recovered PETs (arbitrarily, >=10 PETs per matching transcript) from each of the overcall- or undercall-error ditag sets (above), that could be mapped onto known genes (data within any of the sub-bases RefSeq, KnownGene, GenBank mRNA, or MGC).

Accordingly, we identified 4,185 recovered PETs that contained putative overcall errors, that corresponded to 39 transcripts (these transcripts were defined by an existing

set of 11,077 PETs within the single-locus-match (PET1) category). Conversely, we were able to identify 2,032 recovered PETs (putative undercall errors) corresponding to 46 transcripts defined by an existing 12,812 PETs within the PET1 category. By aligning the newly-recovered PETs with the corresponding pre-existing PETs (that were readily mapped to the same transcript during first-pass mapping), the percentage of error in each category could then be calculated by the formula:

> Percentage error per homopolymer category = **x / (n * y)**, where **x** = the number of cases (PETs) recovered, and **n** = total number of sites of homopolymer length **y** (in both the recovered ditags plus the matching pre-existing PETs), and the multiplication by y is necessary to take into consideration the total number of nucleotides that were sequenced.

Our data (Supplementary Figure 4) shows that the occurrence of errors within the homopolymer regions appears to increase with homopolymer length, with a peak at homopolymer length = 5 bases. Furthermore, insertion errors (overcalls) are more prevalent. This is consistent with our scavenging results above, where we found that the parameter "allow-1-deletion" was more important than "allow-1-insertion" with regard to recovering accurate ditags (in other words, there were quantitatively more ditags that were rendered unmappable due to the extraneous insertion of bases in homopolymer regions).



Supplementary Figure 4. Error distribution across different homopolymer lengths
The "Over-called" homopolymer errors contain an extra base in the homopolymer stretch compared to the reference genome sequence. The "Under-called" errors missed a base in the homopolymer stretch when compared to the reference genome sequence. "Total" errors are the sum of the two.

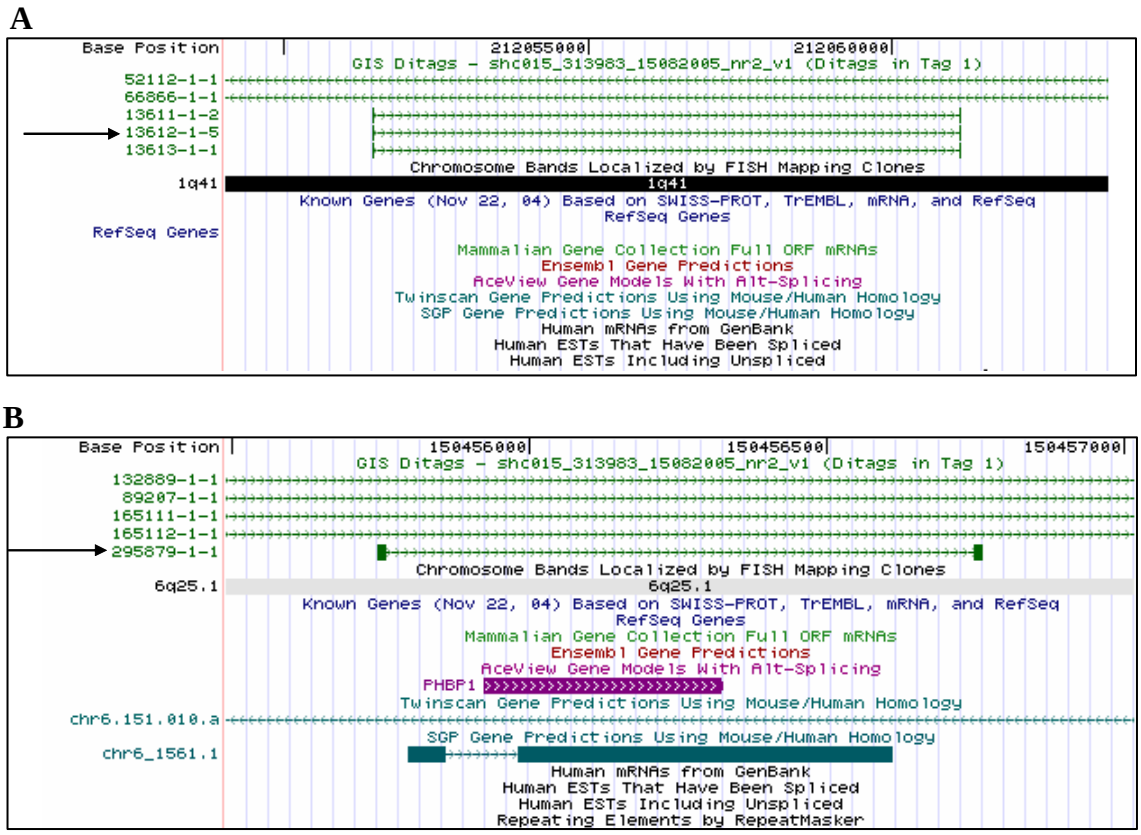
Errors in homopolymer regions are a known artifact of 454-sequencing, and the overall trend in error distribution, *viz*. that errors increase with increasing homopolymer length, is similar between our data and earlier published results (1). Although the published results indicated that undercall errors were predominant, in direct contrast with our own analysis, we have since established that this is a random rather than systematic phenomenon, and the relative proportion of insertion and deletion errors appear to be vary from one library to another (Du Lei, pers. comm.).

We also examined the distribution of homopolymers of varying lengths within both the pool 56,914 PETs that were recovered as described in the main text, and the pool of 136,612 PETs that could be mapped to unique chromosomal loci in the first-pass mapping. Supplementary Table 3 shows that the longer homopolymers (more specifically, the total number of bases sequenced in longer homopolymeric stretches) are more well-represented in the pool of recoverable PETs, compared to bases in homopolymer stretches of the same length within the first-pass-mapped PETs. We believe that this reflects the increased incidence of multiplex-sequencing errors with increasing homopolymer length, possibly coupled with decreased tag complexity, both these factors contributing to poorer mapping rates.

Homopolymer distribution in 56,914 recovered PETs	H2	H3	H4	H5	H6	H7+
Total homopolymers in each class (n)	287,966	86,084	32,739	22,251	3,567	0
Total bases sequenced in each class (n * homopolymer category)	575,932	258,252	130,956	111,255	21,402	0
Percentage of 2,054,944 total bases sequenced in all 56,914 PETs	**28.03**	**12.57**	**6.37**	**5.41**	**1.04**	**0.00**
Homopolymer distribution in 136,612 PET1s	H2	H3	H4	H5	H6	H7+
Total homopolymers in each class (n)	676,142	216,325	74,030	32,764	5,252	0
Total bases sequenced in each class (n * homopolymer category)	1,352,284	648,975	296,120	163,820	31,512	0
Percentage of 4,907,904 total bases sequenced in all 136,612 PET1s	**27.55**	**13.22**	**6.03**	**3.33**	**0.64**	**0.00**

Supplementary Table 3. Representation of homopolymers in recovered PET0s *vs* first-pass-mapped PET1s.
The table shows that there is a greater percentage of sequenced bases from long homopolymers in the PETs recovered by 1-base insertion/deletion (see main text), compared with sequenced bases from the same homopolymer category in PET1s. H2, 2-mer homopolymers; H7+, homopolymer stretches of 7 or more bases; PET1, PETs mapped to single (unique) chromosomal loci.

5. Transcripts identified by MS-PET



Supplementary Figure 5. Examples of transcripts identified by MS-PET sequencing
A. A total count of 8 GIS-PETs (3 PET sequences) mapped to a novel transcript within a gene desert region. **B.** One GIS-PET sequence was mapped to and validates a predicted gene.

6. Reduction of noise from MS-PET analyzed ChIP-PET clusters

In the 8,896 PETs (88.64% of the total 10,036 mapped ChIP-PET sequences) that mapped to single chromosomal loci, we found 843 PET sequences that were not identical but which nonetheless mapped to identical chromosomal locations, and therefore required merging to eliminate redundancy. This sequence variability we attributed to the phenomenon of MmeI enzyme slippage which we previously observed (2), resulting in uncertainty at the interface of 5' and 3' signatures within each PET. This merging process further reduced the number of PET sequences to 8,053. The majority (7,529) of these 8,053 PETs were aligned along the genome as singletons (i.e., only 1 PET per mapping locus), and were thus also removed as possible non-specific background noise, as authentic ChIP-enriched targets would be expected to form a cluster of PETs around the binding consensus sequence The remaining 524 unique PETs formed 253 clusters containing 2 to 6 individual PETs per cluster, and therefore could be considered to be potential p53 binding sites (Supplementary Table 4).

Supplementary Table 4. Mapping statistics of MS-PET sequenced p53 ChIP-PET data.
The table shows a detailed breakdown of the process of eliminating background noise.

Total no. of PETs	23,283
Unique PET sequences remaining after merging identical PETs (**Noise1**)	22,687
PETs not mapped to hg17	12,651
PETs mapped	10,036
PETs mapped to 1 chromosomal locus	8,896
PETs mapped to 2+ loci	1,140
Of 8,896 single-locus PETs, no. of non-identical PETs mapping to same locus (**Noise2**)	843
no. of singletons (**Noise3**)	7,529
Single-locus PETs remaining after eliminating obvious noise sources (**Noise2** and **Noise3**)	524
no. of PETs overlapping with only <=5 bp end-difference (**Noise4**)*	397
Single-locus PETs remaining after eliminating **Noise4**	127

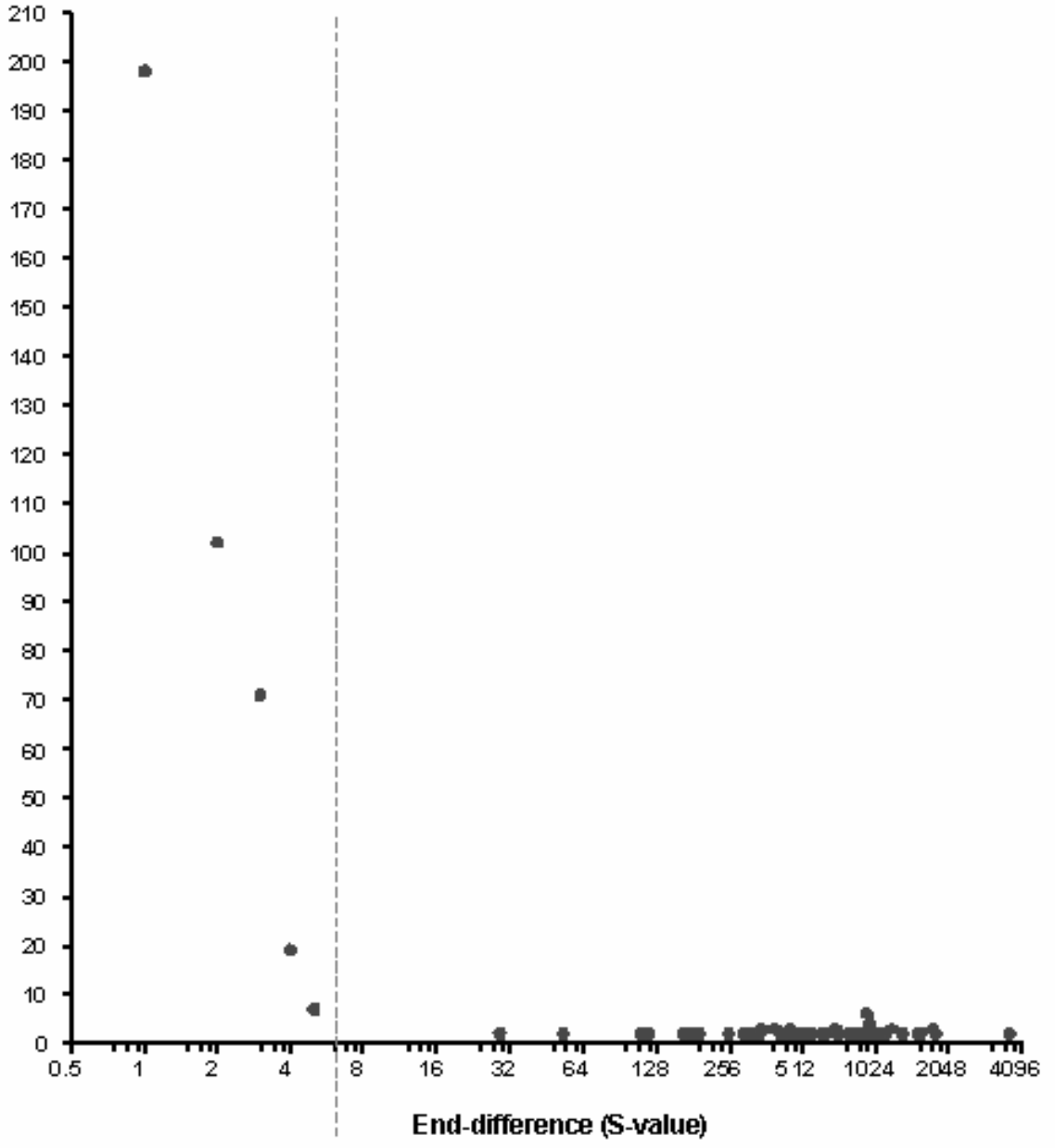
*__Noise4__ was obtained after detailed analysis of the poor initial correlation between this MS-PET analyzed dataset, and a larger Sanger capillary-sequenced ChIP-PET dataset (see text for details).

However, we observed that there was a poor initial correlation (71 of 253; 28.06%) between these 253 clusters and a larger dataset of 1,766 p53 binding sites identified in a previous ChIP-PET experiment (2). Closer examination revealed that a substantial number of PET clusters had their PET members essentially completely overlapping each other, with a difference in mapping of only a few bases at the ends. By determining the sum of end-differences for every PET sequence (S-value; defined as the length of each cluster minus the length of the region overlapped by every PET within that cluster), it was obvious that there was a marked bias in the distribution of PETs, with a far greater number of PETs displaying S-values <= 5. In other words, there appeared to be a transition point at an S-value of 5 (Supplementary Figure 6), separating clusters comprising PETs that were much closer together, from those that were >5 bp apart (both ends considered).

The reason for this was revealed by closer visual examination of all 524 PET sequences: 397 PETs (196 clusters) with S-values <= 5 were in fact *identical* apart from minor variations, and had therefore formed artifactual clusters when it should instead have been a singleton. These variations (resulting in a 5 bp error shared between both ends) were due to a combination of homopolymer errors attributable to multiplex sequencing and terminal mismatches (Supplementary Table 5), which are likely an artifact of the end-polishing procedure used in sequencing library construction. Single-base miscalls within homopolymers accounted for 105/196 or 53.57% of the errors.

This noise reduction process enabled the refinement of a final list of 57 clusters comprising 127 PETs of S-values >5, which, as described in the main text, proved to be high-confidence p53 binding sites. In summary, the data showed that MS-PET generated ChIP-PET data could indeed be used to rapidly identify TFBS. Although, compared to Sanger-sequenced ChIP-PETs, additional errors resulting in the formation of artifactual clusters were present, these could be resolved by modifying the clustering algorithms to take into consideration PET sequences that should be merged after allowing for the presence of single-base insertions or deletions within homopolymers.

The 57 putative p53 binding loci identified in this study were compared with p53 binding loci determined by PET clusters in a previous study with a considerably larger dataset generated by Sanger capillary sequencing. The result is presented in Supplementary Table 6.



Supplementary Figure 6. Determining the end-difference cutoff value (that we termed S-value) in all ChIP-PET clusters (253 clusters of 524 PETs). The graph is a plot of the number of PETs at each corresponding S-value (see Supplementary Table 5). The vast majority of PETs (397 PETs in 205 clusters) appear to be concentrated at S-value <=5.

Supplementary Table 5. Noise-reduction analysis on MS-PET analyzed ChIP-PETs.

After eliminating sources of alignment error (both insertion or deletion errors in homopolymer regions, attributable to multiplex sequencing, and errors that were not within homopolymer regions, attributable to molecular cloning procedures), it was discovered that all PETs with S-values =<5 had in fact formed artifactual clusters; conversely, all clusters containing PETs with S-values >5 were verified to be authentic, and were high-confidence p53 targets.

End-difference value (S-value)	No. of PETs	High-confidence clusters (>=2 cluster members)	Homopolymer errors (insertions/deletions)	Non-homopolymer errors	No. of high-confidence clusters after eliminating errors
1	198	99	51	48	0
2	102	50	26	24	0
3	71	35	20	15	0
4	19	9	6	3	0
5	7	12	2	1	0
>5	127	57	0	4	57
	Total PETs = 524	Total clusters = 253	Total errors = 105	Total errors = 95	

"Noise" from false clusters

Supplementary Table 6. The 57 putative p53 binding loci identified by MS-PET sequencing analysis.

After background noise reduction as described in the Supplementary Information, a final 57 PET clusters (containing 127 PETs) identified by MS-PET analysis of ChIP-PET data were matched with high correlation to a large, capillary sequenced dataset (2). "Cluster Size", numbers of individual PETs in each cluster; "p53 binding motif", identified using the p53PET model (2) or with *MatInspector (3); "Nil", no consensus binding site identified. **3'half-site identified.

PET clusters identified in this study by MS-PET sequencing			**PET clusters identified in previous study by Sanger capillary sequencing**		
Cluster ID	Cluster Size	Cluster Location	Matching cluster ID	Cluster Size	p53 binding motif
chr17.65114635	6	chr17:65114635-65115683	chr17.65114449	18	CTGCATGTCAGAACATGCCC
chr1.121096169	4	chr1:121096169-121097192	chr1.121096168	27	Nil
chr8.41795356	3	chr8:41795356-41796055	chr8.41794748	10	TAACTTGCCCAGACATGCCG
chr8.128876297	3	chr8:128876297-128877901	chr8.128875604	10	ATACTGGCAGCGACAAGTTGA**
chr10.129551462	3	chr10:129551462-129552379	chr10.129551119	9	TGACTTGCCCAGACATGTCT
chr19.32428502	3	chr19:32428502-32430399	chr19.32428502	7	Nil
chr19.15863930	3	chr19:15863930-15865190	chr19.15863930	3	CAGCATGCCTTGACATGCCT
chr7.98979755	3	chr7:98979755-98980587	chr 7.98979755	3	TAACATGTAGGGACTTGCCTA*
chr12.104298329	3	chr12:104298329-104299628	chr12.104298329	2	CCACATGGCCCGACCTGACTA*
chr6.36751959	2	chr6:36751959-36752472	chr6.36751902	13	GAACATGTCCCAACATGTTG
chr7.40530025	2	chr7:40530025-40530784	chr7.40529658	10	GGGCATGCCCAGACAAGCCC
chr15.78081823	2	chr15:78081823-78083380	chr15.78081823	9	AGGCGTGTTCGGACATGTCT
chr12.15980549	2	chr12:15980549-15981513	chr12.15980549	9	AGACAGGACAGGACAGGACAG*
chr4.40987514	2	chr4:40987514-40988177	chr4.40986864	8	GGGCATGTTGGGACATGCCT
chr7.150822921	2	chr7:150822921-150823498	chr7.150822098	8	GAGCATGTCTGAACATGTTC
chr6.110309910	2	chr6:110309910-110310423	chr6.110309889	8	AGACTTGCCTGGGCCTGTCC
chr4.78620219	2	chr4:78620219-78621486	chr4.78620134	7	AGGCATGTTTGGACATGTCT
chr8.143893708	2	chr8:143893708-143894556	chr8.143893708	7	ATGCTTGCCCAGGCATGTCC
chr9.84083290	2	chr9:84083290-84083913	chr9.84083088	7	GCACATGCCTGGACATGTTT
chr1.211001176	2	chr1:211001176-211001907	chr1.211000966	7	AAACATGTTGCAACATGTCC
chr19.18335537	2	chr19:18335537-18337042	chr19.18335537	6	CAGCATGCCTTGACATGCCT

chr12.826876	2	chr12:826876-828182	chr12.826876	6	AGGCATGTGCCAACATGCCC
chr5.152171146	2	chr5:152171146-152172316	chr5.152171146	6	TGACTTGCCCAGACATGTCC
chr9.4778404	2	chr9:4778404-4779001	chr9.4778277	6	GAGCATGCCTGTACATGCCT
chr7.152123965	2	chr7:152123965-152125102	chr7.152123892	6	TTACATGCCCGGACATGCCA
chr14.71310532	2	chr14:71310532-71311108	chr14.71310476	6	GGGCTTGTCTAAGACATGCTC
chr2.170903804	2	chr2:170903804-170904541	chr2.170903601	5	GGGCATGCCCCAACATGCCT
chr7.61411725	2	chr7:61411725-61413516	chr7.61411725	5	Nil
chr19.46725987	2	chr19:46725987-46726762	chr19.46725987	5	GAACATGCCTGGGCACATTCA*
chr6.112415080	2	chr6:112415080-112416243	chr6.112414685	5	AGGCATGTCAGGGCCTGTCC
chr8.103317718	2	chr8:103317718-103318672	chr8.103317718	4	AGACATGCCTGGGCATGTCA
chr12.27593036	2	chr12:27593036-27594614	chr12.27593036	4	Nil
chr4.188216741	2	chr4:188216741-188217836	chr4.188216741	4	GGACATGCCCGGGCAAAGGCC*
chr17.46386144	2	chr17:46386144-46387669	chr17.46385667	4	TGACAAGCCCAGACATGCAG
chr2.51394754	2	chr2:51394754-51396351	chr2.51394754	3	GGACATGAATGGACATGTCT
chr17.52149611	2	chr17:52149611-52150336	chr17.52149611	3	GAACATGCCCAGGCAAGCCC
chr7.151206925	2	chr7:151206925-151207691	chr7.151206867	3	GGGCATGTTGGCGCACGTCT
chr11.34663350	2	chr11:34663350-34663804	chr11.34663117	3	TTGCATGGCTGGGCAGGGACT*
chr10.61917221	2	chr10:61917221-61918285	chr10.61917221	3	AGGCATGCTCCACCATGCCT
chr8.57980443	2	chr8:57980443-57981801	chr8.57981054	2	TGACATGTTTGGGCATGTTG
chr10.86274468	2	chr10:86274468-86275311	chr10.86274468	2	GGGCTAGCCTGAGACATGCCC
chr1.227069711	2	chr1:227069711-227070623	chr1.227069711	2	AGACAAGTTGAGACTTGCCC
chr8.95072056	2	chr8:95072056-95072992	chr8.95072056	2	AGACATGCCCAGGCAAACCC
chr9.16627209	2	chr9:16627209-16627911	chr9.16627209	2	GAACATGCAGGGGCAAGCCT
chr12.34742184	2	chr12:34742184-34746104	chr12.34742184	2	TGAGTTGAACACACATGTCAC**
chr1.3650125	2	chr1:3650125-3652312	chr1.3650125	2	GTGCATGTACACGCATGCCTG*
chr6.98126164	2	chr6:98126164-98128173	chr6.98126164	2	AAACATGTCTGTTCATGTTCT*
chr8.87596299	2	chr8:87596299-87597941	Chr8.87596299	2	Nil
chr2.45356351	2	chr2:45356351-45357648	Chr2.45356351	2	Nil
chr8.128235595	2	chr8:128235595-128236621	Chr8.128235595	2	Nil
chr1.26300920	2	chr1:26300920-26302190	Chr1.2630092	2	GGCCATGAAGGGGCTTGGCCT*
chr12.13282893	2	chr12:13282893-13284049	Chr12.13282893	2	Nil

chr18.30704122	2	chr18:30704122-30704951	Chr18.30704122	2	AGAGGGAGATGGGCAGGTCTC**
chr4.139909547	2	chr4:139909547-139910275	Chr4.139909547	2	Nil
chr17.74162019	2	chr17:74162019-74162946	Chr17.74162019	2	Nil
chr5.58290360	2	chr5:58290360-58290516	Chr5.58290360	1	Nil
chr2.66053539	2	chr2:66053539-66053605	chr2.66053539	1	Nil

7. References Used in Supplementary Information

1. Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., Berka, J., Braverman, M.S., Chen, Y.J., Chen, Z. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437, 376-380.

2. Wei, C.L., Wu, Q., Vega, V.B., Chiu, K.P., Ng, P., Zhang, T., Shahab, A., Yong, H.C., Fu, Y., Weng, Z. *et al.* (2006) A global map of p53 transcription-factor binding sites in the human genome. *Cell*, 124, 207-219.

3. Quandt, K., Frech, K., Karas, H., Wingender, E. and Werner, T. (1995) MatInd and MatInspector - New fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res*, 23, 4878-4884.