DNA Breaks and Gaps Target Retroviral Integration

Gayan Senavirathne[1], Anne Gardner[1], James London[1], Ryan K. Messer[1], Yow Yong Tan[1], Kristine E. Yoder[1,*] and Richard Fishel[1,*]

[1]Department of Cancer Biology and Genetics, The James Comprehensive Cancer Center and Ohio State University Wexner Medical Center, Columbus, OH 43210, US

*Correspondence may also be addressed to: R. Fishel (rafishel@gmail.com) or K.E. Yoder (yoder.176@osu.edu)

Integration into a host genome is essential for retrovirus infection and is catalyzed by a nucleoprotein complex (Intasome) containing the virus-encoded integrase (IN) and the reverse transcribed (RT) virus copy DNA (cDNA). Previous studies suggested that integration was limited by intasome-host DNA recognition progressions. Using single molecule Förster resonance energy transfer (smFRET) we show that PFV intasomes pause at nicked and gapped DNA, which targeted site-directed integration without inducing significant intasome conformational alterations. Base excision repair (BER) components that affect retroviral integration *in vivo* produce similar nick/gap intermediates during DNA lesion processing. Intasome pause dynamics was modified by the 5'-nick/gap chemistry, while an 8-oxo-guanine lesion, a mismatch, or a nucleotide insertion that induce backbone flexibility and/or static bends had no effect. These results suggest that dynamic often non-productive intasome-DNA interactions may be modulated to target retroviral integration.

(137 words)

Retroviruses must integrate into a host genome to cause pathogenic diseases such as AIDS or leukemia [1]. Adaptation of these genomic integration properties has been employed in the development of gene delivery vehicles designed to correct a variety of monogenic diseases [2]. Lentiviral vectors largely derived from the HIV-1 retrovirus appear most easily manipulated and can infect non-dividing cells [3]. However, lentivirus often integrate into open actively transcribing host genomic regions [4] which may unintentionally alter cellular attributes or activate oncogenesis [5].

Modulating retrovirus target site choice is one scheme for modifying genome targeting. Retroviral integration is catalyzed by a poorly described pre-integration complex (PICs) that at a minimum consists of the viral integrase (IN) in complex with the two long terminal repeat (LTR) sequences that flank the viral cDNA, termed an intasome [6-8]. IN initially excises two nucleotides (nt) from the LTR [9] producing recessed 3'-OH's that provide the substrate for consecutive IN-

catalyzed S_N2 strand transfer reactions. Strand transfer covalently links the LTR ends across one major groove of a target DNA separated by 4-6 bp, depending on the retrovirus genera [8,10,11]. The newly integrated provirus is then flanked by 4-6 nt gaps containing a 2 nt 5'-flap, which appear to be restored to fully duplex DNA by the host DNA repair machinery [12-14].

Most retroviruses prefer to integrate within cryptic DNA sequence that for HIV-1 (T c **G** * G/t T N A C * C/a g A; lowercase disfavored, * strand transfer site) often positions a guanine (G) adjacent to the 5'-LTR [15,16]. Importantly, mutation of the base excision repair (BER) glycosylases *OGG1* or *MUTYH* reduced HIV-1 integration and altered the preference for G adjacent to the 5'-LTR [12,13,17]. OGG1 principally excises 8-oxo-guanine (8-oxo-G) residues in DNA that are the most frequent oxidative damage in cells [18,19]. MUTYH excises adenine (A) residues that are frequently misincorporated across from an 8-oxo-G lesion during replication. Mutation of the BER polymerase β (*Polβ*) also significantly decreased HIV-1 integration [17]. However, the defect was linked to the 5'-deoxyribose phosphate (5'-dRP) lyase activity and not the polymerase activity of Polβ [17]. The 5'-dRP lyase removes the deoxyribose sugar and associated phosphate after base excision by the glycosylase, leaving a 1 nt gap in the DNA [20-22]. OGG1 also contains an intrinsic 5'-dRP lyase that may be exploited in concert with MUTYH damage processing [17].

Historical studies have suggested that bent or flexible DNA is a favored target for retroviral integration *in vivo* [23,24] and *in vitro* [25-32]. Prototype foamy virus (PFV) is a member of the *spumavirus* genera and has been extensively studied since it shares similar catalytic site geometry, chemistry and therapeutic inhibitor sensitivity with the pathogenic lentivirus HIV-1 [33-35]. Structures of the PFV intasome target capture complex (TCC) and the strand transfer complex (STC) have been solved and represent key steps in retroviral integration [25,33]. While several single molecule studies have detailed the dynamic progressions between PFV intasome and target DNA [26,32], the processes that drive integration site choice remain enigmatic.

RESULTS

PFV intasomes transiently pause at nicks and gaps in the DNA. Single molecule Förster resonance energy transfer (smFRET) [36] was developed to probe the real-time interactions between a purified PFV intasome and DNA containing a variety of DNA lesions and BER intermediates that introduce localized bends or flexibility into DNA [37-40] (**Fig. 1a**; **Extended Data Table 1**). PFV intasomes were assembled with pre-processed viral 5'-LTR oligonucleotides (vDNA) containing a Cy3 FRET-donor flourophore on the complementary strand 11 bp 5' of the recessed 3'-OH. A surface immobilized 60 bp target DNA (tDNA) contained a Cy5 FRET-acceptor flourophore located on the undamaged strand 11 bp 5' of a lesion as well as an Alexa488 marker flourophore on the damaged strand 3 bp from the 3' end (**Fig. 1a**). We first confirmed that concerted integration of the Cy3-PFV intasomes into a supercoiled target DNA was similar to unlabeled PFV intasomes, suggesting that the Cy3-labeled vDNA does not affect integration activity (**Extended Data Fig. 1**) [26].

Introduction of Cy3-PFV into an smFRET flow-cell containing surface immobilized duplex DNA and illuminating with a 532 nm laser resulted in Cy5 (660 nm) FRET emission events consistent with dynamic TCC interactions at 100 msec time resolution (**Fig. 1b and 1c**, **left**). Background correction and Hidden Markov Modeling (HMM) generated a transition density plot [41] that was used to determine FRET efficiency (E) and reversibility (**Fig. 1c**, **middle**; **Extended Data Fig. 2 and 3**). The total number of FRET events (n) from numerous DNA molecules (N) were aligned to the initial PFV intasome infusion time to produce a post-synchronized histogram (N = 552, n = 46; **Fig. 1c**, **right**) [42]. A narrow range of Cy5 flourescence (E ~ 0.06) was detected in both the transition density plot and post-synchronized histogram that after inspection of the real-time movies were found to represent excursions of free Cy3-PFV intasomes or aggregates across the evanescent field containing the substrate DNA (**Extended Data Movie 1**). These events were easily recognized since they saturated the Cy3 channel and bled into the Cy5 channel generating a pseudo-FRET signal (**Fig. 1c**, **right**). Rare higher FRET events (E > 0.1)

were detected that appeared to correspond to PFV intasome binding and dissociation near the Cy5 flourophore on the DNA (**Fig. 1c**, **right**). A similar pattern was observed when the DNA contained an 8-oxo-G/C nucleotide lesion, a G/T mismatch or a +T nucleotide insertion (**Extended Data 4a-c**). The lack of FRET events above the pseudo-FRET background suggests very little if any specific binding of PFV intasomes to these DNA substrates.

In contrast, a DNA substrate containing a 1 nt gap with a 3'-hydoxyl (OH) and 5'-phosphate (1 nt GAP 5'-P) that is similar to a 5'-dRP lyase intermediate consistently displayed bursts of Cy3 transitions accompanied by stronger Cy5 intensity jumps consistent with FRET (**Fig. 1d, left**). Single molecule movies confirmed that these FRET events represented PFV intasome-lesion interactions arising from colocalized diffraction-limited Cy3-Cy5 spots (**Extended Data Movie 2**). The FRET bursts from individual traces (**Fig. 1d**, **center**) converged into a distinct ensemble FRET peak ($E_{1nt\ Gap\ (5'\text{-}P)}$ ~ 0.16 ± 0.05; N = 540, n = 3267; **Fig. 1d**, **right**; **Extended Data Table 4**) that was clearly above the pseudo-FRET background. We conclude that these higher FRET events are stably but transient TCCs where the PFV intasome has bound to the DNA lesion (**Fig. 1b**; **Fig. 1d**). DNA substrates containing a strand scission (Nick 5'-P; $E_{Nick\ (5'\text{-}P)}$ ~ 0.15 ± 0.08; N = 549, n = 1084), a 2 nt gap (2 nt GAP 5'-P; $E_{2nt\ Gap(5'\text{-}P)}$ ~ 0.18 ± 0.10; N = 474, n = 1175) as well as a nick ($E_{Nick\ (5'\text{-}OH)}$ ~ 0.17 ± 0.07; N = 558, n = 507), 1 nt gap ($E_{1nt\ Gap\ (5'\text{-}OH)}$ ~ 0.19 ± 0.05; N = 496, n = 2784) and 2 nt gap ($E_{2nt\ Gap\ (5'\text{-}OH)}$ ~ 0.23 ± 0.11; N = 540, n = 799) without the corresponding 5'-P (5'-OH) also yielded FRET events consistent with TCC's (**Extended Data Fig. 4d-h**; **Extended Data Table 2**). The number of transitions per molecule also increased with DNA substrates that showed clear FRET events (**Extended Data Fig. 5**). Taken together, these observations strongly suggest that the FRET transitions indicated Cy3-PFV binding to specific DNA lesions.

To quantify the relative efficiency between lesions, we normalized the frequency of TCC FRET binding events to the pseudo-FRET excursion events that are relatively constant at a fixed Cy3-PFV intasome concentration between experiments, DNA substrates, and time resolution

(**Fig. 1e-g**; **Extended Table 3**). As expected, there were few FRET events outside the normalized pseudo-FRET background (E ~ 0.06) in the presence of duplex DNA, or DNA containing an 8-oxo-G/C lesion, a G/T mismatch or a +T nucleotide insertion (**Fig. 1e**; **Extended Data Fig. 6a-d**). Since the G/T mismatch and +T insertion introduce significant local flexibility into the DNA and the +T insertion produces a relatively stable DNA bend, these observations suggest that the preference for flexible or bent DNA by retrovirus intasomes [23,43,44] may be more subtle than previously appreciated.

In contrast, a nick, 1 nt gap or 2 nt gap, with or without the 5'-P, displayed FRET peaks that were clearly separated from the pseudo-FRET background (**Fig. 1f,g**; **Extended Data Fig. 6e-j**). The relative frequency of FRET TCC events suggested that Cy3-PFV preferred a 1 nt Gap >> nick $\geq$ 2 nt Gap. A broadening of the FRET TCC peak was observed with a 2 nt GAP (**Fig. 1f,g**). Because the 2 nt Gap extends further toward the Cy5 FRET acceptor, these results appear to suggest an increased number of different asymmetric binding events (degrees of freedom) occurring 5' of the gap. We note that the post-synchronized histograms and the relative binding efficiency pattern of FRET TCC events was similar at 1 sec resolution (**Extended Data Fig. 7**; **Extended Table 4**).

Using HMM we determine the binding (τ_{on}) and dissociation (τ_{off}) kinetics (**Extended Data Fig. 8a**). Even at the 100 msec frame rate binding to duplex DNA and DNA containing an 8-oxo-G/C lesion, a G/T mismatch or a +T nucleotide insertion was extremely rare, often << 1 event per molecule over the 3 min observation window, which resulted in statistically insignificant τ_{on} and τ_{off} values (**Extended Data Fig. 5a-d**; **Extended Data Fig. 8b**). However, the 1 nt GAP (5'-P) and 1 nt Gap (5'-OH) experienced an average of ~6 binding events per molecule with a distribution of 1-14 binding events for the vast majority of molecules (**Extended Data Fig. 5f,i**). Less, but statistically significant binding events were recorded for DNAs containing a nick and 2 nt gap regardless of the 5'-end configuration (**Extended Data Fig. 6e,g,i,j**). Fitting these events to a

single exponential decay resulted in a τ_{on} that varied from 230-483 msec and an τ_{off} that varied from 6.8-21.7 sec (**Fig. 1h,i**; **Extended Data Fig. 8c-h**). We note that the τ_{on} events for the 2 nt gap substrates approach the frame rate which appear to reduce the number of recorded events and in the case of the 2 nt GAP (3'-OH) artificially increased the τ_{off} (**Fig. 1h,i**; **Extended Data Fig. 8e,h**). Blocking the open end of the 1nt GAP (5'-OH) DNA substrate to trap intasomes diffusing along the DNA [45] did not significantly alter the distribution of events per molecule or the post-synchromized FRET histogram slightly (**Extended Data Fig. 9a,b**. However, the τ_{on} was slightly decreased while the τ_{off} increased (**Fig. 1h.i**; **Extended Data Fig. 9c-e**), which significantly increase in the number of normalized binding events (**Fig. 1j**). These observations suggest that at least a fraction of the TCC Cy3-PFV intasomes dissociate by sliding off the ends of the lesion containing oligonucleotide, while most directly dissociate for the DNA.

Integration of PFV intasomes occurs at the pause site. The formation of a strand transfer complex (STC) resulting in integration results in prolonged FRET events (**Fig. 2a,b**). These long-term FRET events often ended with photobleaching of one or both Cy3 flourophores on the Cy3-PFV intasome or photobleaching of the Cy5 flourophore on the DNA substrate (**Fig. 2c**). Post-synchronized histograms displayed well defined STC FRET events for many of the DNA substrates in which both the frequency and FRET efficiency could be determined (**Fig. 2d,e**; **Extended Data Table 2**). Apart from the Nick (5'-P) DNA substrate, the FRET efficiency of TCC events (E_{TCC}) were nearly identical to the STC events (E_{STC}; **Extended Data Table 2**). Where it differed, for example the Nick (5'-P) DNA substrate, the number of events were extremely low, often below 5, making interpretation of the histograms impractical (**Extended Data Fig. 10)**.

To confirm the nature of the STC events we assembled PFV intasomes with viral donor oligonucleotides containing a termina di-deoxy-adenosine (ddA) that is incapable of integration (**Fig. 2f,g**). As expected, we observed significant numbers of TCC events with the 1 nt Gap (5'-

OH) that displayed a similar FRET efficiency to *wild type* PFV intasome interactions (**Fig. 2f**; **Extended Data Table 2**). However, we observed only two potential STC events (**Fig. 2g**). These results strongly support the conclusion that the prolonged FRET events observed with normal viral DNA donor oligonucleotides represent covalent STC formation.

To visualize the dynamic interactions of PFV INT with different DNA lesions, we performed a series of smFRET experiments by introducing a Cy3 labeled INT (NTS-P-2Cy3-PFVwt INT) on to surface-immobilized 60 bp DNA containing an A488, a Cy5 and different lesions. A488 and Cy5 were placed respectively 18 bp and 11 bp away from the damages to minimize any interference with the INT activities (**Figs. 1a-f,2a-g, Supplementary Tables 1,2**). First, we examined if the oxidative damage 8-OxoG can promote TC (**Fig. 1b**). The INT-tDNA binding should generate a Cy3 signal. TCC formation centering the damage should result in a robust Cy5 intensity by ~50% FRET from Cy3 (**Supplementary Fig. 2a**). However, the SM traces at 100 ms resolution only contained occasional sharp Cy3 spikes with corresponding weak Cy5 intensities leading to 0.06 FRET (**Fig. 1b, panels 1,2**). SM movies revealed these Cy3 excursions were not binding events, but rather aggregates of PFV INT[29] floating in the bulk solution saturating the Cy3 channel (**Supplementary Movie 1**). Saturating Cy3 intensities bled into the Cy5 channel generating apparent Cy5 signals with a pseudo-FRET value ($E_{aggn.2Cy3\text{-}INT}$) at 0.06.

A transition density plot (TDP) generated by the Hidden Markov Model (HMM) analysis[41] of FRET traces indicated two narrows off-diagonal peaks (0→0.06, 0.06→0) representing the passage of aggregates through DNA pixels (**Fig.1b, panel 3, Supplementary Fig. 4a**). Similarly, a post-synchronized histogram (PSH) produced by aligning ensemble of trajectories to the initial INT infusion[42] showed a narrow FRET density at $E_{aggn.2Cy3\text{-}INT}$ = 0.06 (**Fig.1b, panel 4**). A time averaged smFRET histogram also showed a distribution centered around the same FRET value (**Fig. 1f, panel 1**). An undamaged fully complementary (FC) DNA reproduced indistinguishable behavior (**Supplementary Figs. 4b,5a**). Thus, any interaction of PFV INT with the 8-OxoG or undamaged DNA should be <100 ms temporal resolution of the experiment.

PFV INT transiently capture SSDBs. Next, we wondered if DNA intermediates associated with the SPBER pathway and similar damages enable specific TC. Single-nt gaps are obligatory intermediates of the enzymatic removal of 8-OxoG[20-22]. These may induce bent DNA conformations similar to the tDNA in the TCC crystal structure (**Supplementary Figs. 1a,2a**). A single-nt gap substrate with 5'-Phosphate (P) at the lesion (5'-P 1nt Gap DNA) indeed displayed markedly different INT TC dynamics compared to the 8-Oxo G and FC DNAs (**Fig. 1c**). In addition to INT aggregations, intensity trajectories showed bursts of Cy3 transitions accompanied by stronger Cy5 intensity jumps (**Fig. 1c, panel 1)**. The resulting FRET bursts in individual traces converged into a distinct ensemble TC state at $E_{TCC.5'\text{-}P\ 1nt}$ = 0.15 (**Figs. 1c,1f, panel 1, Supplementary Fig. 4c**). SM movies confirmed $E_{TCC.5'\text{-}P\ 1nt}$ represented a true INT-DNA bound state originating from co-localized diffraction-limited Cy3-Cy5 spots (**Supplementary Movie 2**). However, the $E_{TCC.5'\text{-}P\ 1nt}$ = 0.15 was noticeably lower than the predicted value of ~0.5 from the crystal structure (**Supplementary Fig. 2a**)

While INT binding to 8-OxoG and FC DNAs is undetectable, the 5'-P 1nt Gap substrate displayed ~6 binding transitions per trace on average in a distribution spanning up to ~13-15 counts (**Supplementary Figs. 6a-c,7a-d**). The formation of these specific TCCs were highly dynamic and transient. From the HMM analysis, we determined the TCC lifetime (τ_{on}) to be 436 ± 100 ms and the average time between bindings (τ_{off}) to be 14.6 ± 0.8 s at 5 nM INT (**Fig. 1f, panels 2,3, Supplementary Fig. 7a,d**).

DNA nicks are the last intermediates of SPBER reaction[20-22]. Nicks may also introduce flexibility to dsDNA, albeit with lower efficiency[39,40], promoting TC by PFV INTs. In our smFRET TC assay, a nicked substrate (5'-P Nick DNA) showed localized PFV INT binding similar 5'-P 1nt Gap DNA (**Fig. 1d,1f, panel 1, Supplementary Fig. 4d**). However, the smFRET and binding count histograms revealed ~3-fold reduction in binding compared to the gap substrate (**Fig. 1f**

panel, 1, Supplementary Fig. 6d), indicating a possible correlation between tDNA flexibility and TCC formation. We then tested this notion by using a target containing a two-nt gap (5'-P 2nt Gap DNA) (**Fig. 1e**). Based on the flexibility, this DNA should produce more TCCs compared to previous two. Contrary to this expectation we observed a ~3-fold lower binding similar to the nicked DNA (**Fig. 1f, panel 1, Supplementary Fig. 6e**). Moreover, the FRET distribution showed slightly shifted center at $E_{TCC.5'\text{-}P\ 2nt}$ = 0.18 (**Fig 1f, panel 1, Supplementary Fig. 4e**). Although the number of binding events was lower, the TCC lifetime (τ_{on}) and the binding rate ($1/\tau_{off}$) did not change appreciably for nicked and two-nt gap substrates (**Fig. 1f, panel 2,3, Supplementary Fig. 7e,f**).

tDNA bending and the TC. To further evaluate the relationship between tDNA bending and TC, we employed two additional substrates containing a G•T mismatch (GT MM) or a T-Bulge (**Figs. 2a,b**). The pre-bent DNA conformations induced by these lesions assist the recognition by MutS homologs[46-48]. Mu transposase, a DDE family recombinase similar to PFV INT, was also shown to efficiently bind G•G substrates because of DNA bending[28,46]. However, at 100 ms resolution, PFV INT showed the same undetectable binding characteristics as 8-OxoG and FC DNA (**Figs. 2a,b,f, panel 1, Supplementary Figs. 4f,g**).

The chemical groups at the damage and the TC. The dynamic TC may be a result of shallow (~6-7 k_BT) potential energy wells[49] presented by the SSDBs for PFV INT binding (See the DISCUSSION and ONLINE METHODS). These may act as kinetic traps that stall INTs. Next, we assessed if the chemical nature at the SSDBs can affect the energy barrier by using a set of lesions bounded by 3'-,5'-OH instead of canonical 3'-OH,5'-P (**Figs. 2c-g**). The presence of 3'-,5'-OH on a DNA gap was shown to significantly affect the gap filling activity of Polβ. The 5'-OH 1nt Gap DNA recapitulated the TC characteristics of the 5'-P counterpart with only minor changes

(**Figs. 2c,f,g, Supplementary Figs. 4h,6f,7g**). Compared to the 5'-OH nick, 5'-P nick showed similar TCC lifetime (τ_{on}) and a binding rate ($1/\tau_{off}$) albeit with ~2-fold lower biding frequency (**Figs. 2d,f,g, Supplementary Figs. 4i,6g,7h**). The 5'-OH 2nt Gap DNA maintained a shifted FRET distribution similar to the 5'-P counterpart but had ~1.5-fold lower binding frequency (**Figs. 2e,2f panle1, Supplementary Figs. 4j,6h)**. The kinetic analysis for this DNA revealed a ~2-fold lower τ_{on} and a slight decrease in the binding rate ($1/\tau_{off}$) compared to other DNAs (**Figs. 2f, panel 2,2g, Supplementary Fig. 7i)**. Interestingly, the FRET distributions for the all three 5'-OH substrates displayed minute but noticeable shift towards higher FRET (ΔE_{TCC}~0.02-0.05) compared to the 5'-P substrates (**Figs. 1c-f and 2c-f**).

A separate PFV INT containing Cy3 at a different location (TS-D-2Cy3-PFVwt INT) (**Supplementary Figs. 2b,8a**) displayed similar activities, albeit with a lower E_{TCC}~0.07-0.08 indicating virtually no effect by Cy3 labeling (**Supplementary Figs. 8a-e**).

Damage site dependent PFV INT catalyzed ST. In addition to the transient TCCs, a subset of FRET traces also contained stable transitions that lasted for several minutes (**Supplementary Fig. 9**). The lifetimes of these extended events were limited by fluorophore photobleaching or the end of observation. To examine if they represented the real-time formation of STCs, we repeated the smFRET experiments at 1 s resolution. This allowed longer (10 min) observations without fast photobleaching (**Fig. 3, Supplementary Figs. 8, 10-15**).

Incubation of NTS-P-2Cy3-PFVwt INT with 8-OxoG, FC, GT MM, and T-Bulge DNAs predominantly resulted in INT aggregations ($E_{aggn.2Cy3\text{-}INT}$ = 0.06) similar to the previous experiments (**Supplementary Figs. 10-12, Supplementary Movie 3**). However, FC and GT MM DNAs respectively showed 2 and 3 longer SM traces that correspond to ~0.5% (N_{Total} = 586) and ~0.6% (N_{Total} = 617) of all the analyzed traces (**Supplementary Fig. 13, Supplementary Table 3**). The FRET values for these traces diverged from the E_{TCC}. Conversely, the SSDBs with either

5'-P or 5'-OH reproduced the same TC pattern observed with high-resolution experiments (**Supplementary Figs. 10-12, Supplementary Movie 4**). Moreover, these DNA substrates produced prolonged FRET events with an E_{STC}~0.16-0.24 comparable to E_{TCC} (**Figs. 3a-e, Supplementary Fig. 14, Supplementary Table 3**). The TS-D-2Cy3-PFVwt INT produced similar data, confirming our observations are independent of the INT labeling scheme (**Supplementary Figs. 8f-i,9g-l,15, Supplementary Table 3**).

Interestingly, the fraction of prolonged FRET traces depended on the tDNA (**Figs. 3d-f, Supplementary Figs.14e,15e,f Supplementary Table 3**). The SSDBs were efficient at producing these events compared to FC DNA or other tested lesions (**Fig. 3f, Supplementary Table 3**). The frequency of stable FRET formation followed an order; single-nt gap > two-nt gap > nick indicating a lack of clear correlation with the tDNA flexibility. Surprisingly, the 5'-OH targets produced more ST than their 5'-P counterparts (**Figs. 3d-f, Supplementary Figs. 15e,f, Supplementary Table 3**). The tDNAs with enough molecules to construct FRET distributions exhibit a pattern; $E_{STC.5'\text{-}P\ 1nt}$ = 0.16 < $E_{STC.5'\text{-}OH\ 1nt}$ = 0.18 < $E_{STC.5'\text{-}OH\ 2nt}$ = 0.24 (**Figs. 3d,e, Supplementary Figs. 14e,f**). These values were closer to their respective E_{TCC} values that followed the same pattern; $E_{TCC.5'\text{-}P\ 1nt}$ = 0.15 < $E_{TCC.5'\text{-}OH\ 1nt}$ = 0.17 < $E_{TCC.5'\text{-}OH\ 2nt}$ = 0.23 (**Figs. 1,2**).

Occasional Cy3 photobleaching revealed E_{STC} is a combination of two distinct FRET states originating from the two Cy3s in the INT (**Figs. 1a,3b,c, Supplementary Figs. 2a,b,9g-j,14d,15b-d**)[42]. Minor differences seen between these two states are consistent with slightly different Cy3-Cy5 distance ($R_{Cy3\text{-}Cy5}$) estimations (**Supplementary Figs. 2a,b**). Under photoprotective SM observations (ONLINE METHODS), the Cy3 photobleaching was inefficient and entirely random[50]. Therefore, most of the STC traces displayed only the combined E_{STC} (**Fig. 3a, Supplementary Figs. 9a-e,k,14a-c,15a**). Consequently, for the overall analyses, we combined all the subpopulations of traces (**Figs. 3d,e**, **Supplementary Figs. 9f,l,14e,f,15e,f**). Likewise, the E_{TCC} must be a twin state, but the shorter TCC lifetimes (τ_{on}) prevented any deconvolution by photobleaching.

The 5'-OH 1nt Gap DNA showed the highest percentage (~20-22%) of prolonged FRET traces (**Figs. 3e,f, Supplementary Fig. 15f, Supplementary Table 3**). To confirm if they represented true STCs, we tested a PFV INT that contained 3'-dideoxy A (3'-ddA, NTS-P-2Cy3-PFV 3'-ddA INT). This INT retains the active site Mg^{2+}-coordination and DNA binding activities of the wt, although 3'-ddAs completely abolish the ST (**Supplementary Figs. 7j,16,17**)[25,28]. In the smFRET assays it efficiently produced short-lived TCCs on 5'-OH 1nt Gap DNA at 100 ms or 1 s resolution similar to the wt (**Supplementary Figs. 7j,16a-d**). However, the formation of long traces was rare (2 in 564, 0.4%), **Supplementary Table 3**), confirming nearly all the stable FRET observed with wt represented catalytic events. STC traces were then used to calculate the overall ST times ($t_{ST,overall}$), measured from INT injection to the appearance of stable Cy3 /FRET signals (**Supplementary Table 4**). The data showed $\bar{t}_{ST,overall}$ is ~3-5 min, corresponding to an overall strand transfer rate ($k_{ST,overall}$) ~0.2-0.3 min^{-1} (**Supplementary Table 4**). In contrast to the product formation efficiency, the $k_{ST,overall}$ was largely independent of the DNA target (**Supplementary Tables 3,4**).

Single ST event in the vicinity of DNA damages . The distinct E_{TCC} and E_{STC} states demonstrated site-specific binding and ST reactions (**Figs.1-3, Supplementary Figs. 4,8-11,14-16**). However, these FRET values only report the positioning of vDNAs relative to the tDNA in three dimensions (3D) (**Fig. 1a, Supplementary Figs. 2a,b,8a**). Therefore, to map the exact integration sites, we performed biochemical assays using the DNAs from smFRET reactions and unlabeled or TS-P-2Cy3-PFVwt INTs (**Figs. 3f,g, Supplementary Figs. 18-20**)[51]. The A488 and Cy5 on tDNAs were used to analyze ST positions with single-nt resolution by denaturing urea-PAGE [51,52].

The 8-Oxo G, FC, GT MM, and T-Bulge DNAs were all inert towards PFV integration (**Figs. 3f,g, Supplementary Figs. 18-20**). However, SSDBs readily produced specific Cy5 product bands, corresponding to integration in the vicinity of the damage sites. The presence of these

localized bands is consistent with the well-defined E_{TCC} and E_{STC}. The fractional intensities of these products were higher compared %STC found in SM assays as expected from a higher reaction temperature (37 ^{0}C vs. room temperature, **ONLINE METHODS**). Still, the two assays followed a similar trend where 5'-OH SSDBs produced higher activities than 5'-P substrates (**Fig. 3f)**. Interestingly A488 labeled strands were intact, indicating no integration into them (**Fig. 3g, Supplementary Figs. 18-20**). Reactions with TS-P-2Cy3-PFVwt INT confirmed no integration into the unlabeled strands either (**Supplementary Fig. 20**). Altogether these data showed single ST events into the Cy5-strand mainly 4 bp upstream of the damage along with other bands that vary by only a few nts (**Fig. 3g, Supplementary Figs. 18-20**). Surprisingly the two STs in PFV catalyzed concerted integrations take place 4 bp apart[25,27,51]. Therefore, it is conceivable that one of the vDNA ends associate with the 5'-moiety of the SSDBs allowing the other one to undergo ST. This behavior leads to asymmetric TCC and STC formation (**Fig. 3g, Supplementary Figs. 18-20**). A plasmid-based biochemical test[26,29] showed wt INTs are fully capable catalyzing concerted integration, with virtually no half-site products that result from a single ST (**Supplementary Figs. 17**).

Asymmetric kinetic trapping leading to asymmetric ST. The sufficient distances of fluorophores from the lesions should not cause asymmetric binding and integration of INT (**Fig. 1a, Supplementary Figs. 2a,b,8a, Supplementary Table 1,2**). To further confirm this is the case, we tested another 5'-OH 1nt Gap DNA that contained a Cy5 11 bp downstream of the gap (**Supplementary Figs. 2c,21a, Supplementary Table 1,2**). This Rev-Cy5 5'-OH 1nt Gap DNA reproduced the biochemical and SM activities of the 5'-OH 1nt Gap DNA (**Supplementary Figs. 20,21, Supplementary Tables 3,4**). With the NTS-P-2Cy3-PFVwt INT, this target produced $E_{TCC.Rev\text{-}Cy5\ 5'\text{-}OH\ 1nt}$ = 0.23, and $E_{STC.Rev\text{-}Cy5\ 5'\text{-}OH\ 1nt}$ = 0.29 consistent with shorter $R_{Cy3\text{-}Cy5}$ (**Supplementary Figs. 2c,21b-g**). However, the placement of Cy5 adjacent to G, A bases in this

DNA construct caused undesirable photophysics[53] that lead to ~6% (N_{Total} = 527) of Cy3 only STCs (**Supplementary Tables 1-4, Supplementary Figs. 21d**).

The higher FRET associated with this tDNA allowed us to clearly discriminate $E_{TCC.Rev\text{-}Cy5\ 5'\text{-}OH\ 1nt}$ from $E_{STC.Rev\text{-}Cy5\ 5'\text{-}OH\ 1nt}$ (**Supplementary Fig. 21e**)[36]. Similarly, the individual states that constituted $E_{STC.Rev\text{-}Cy5\ 5'\text{-}OH\ 1nt}$ were more distinguishable (**Supplementary Figs. 21f,g**). However, the observed FRET values were still lower than the predicted values from the crystal structures (**Supplementary Figs. 2,21**).

Interestingly the lifetimes of STCs were only limited by the fluorophore photobleaching or the observation time window (**Figs. 3a-e Supplementary Figs. 9,13-15,21**). Therefore, STCs are able to firmly hold the entire structure even after creating dsDNA breaks (DSDBs) from SSDBs (**Fig 3, Supplementary Figs. 14,15, 21**).

The progression of transient TC into stable ST. The TCC and STC crystal structures of PFV INT overlay nearly perfectly, indicating no detectable structural changes during integration (**Supplementary Fig. 1b**). Consistent with this observation, our smFRET data revealed only minor or no differences between E_{TCC} and E_{STC} values for several different Cy3-Cy5 geometries (**Supplementary Figs. 2a-c,9f,l,14f,21e**). Moreover, STCs were formed in abrupt single steps with no discernable intermediates even at 100 ms resolution (**Supplementary Fig. 9**). STC traces seemed to be static except for the photophysical fluctuations (**Figs. 3a-c, Supplementary Figs. 9,13-15,21**). However, E_{STC} only indicate the positioning of vDNA relative to the tDNA. The presence of two Cy3 FRET donors and low FRET values may obscure a sensitive measurement of smaller distance changes[36]. Therefore, we generated another PFV INT with a Cy3-Cy5 FRET pair attached to the vDNAs (NTS-Cy3-Cy5-PFV INT) to measure internal structural changes with high sensitivity (**Supplementary Fig. 2d**).

We introduced this INT onto 5'-OH 1nt Gap DNA at 100 ms after photobleaching the Cy5 on the DNA (**Fig. 4a**). These steps were done in real-time consecutively to distinguish INT-tDNA

interactions from the surface bindings (**ONLINE METHODS, Supplementary Movie 5**). The data showed short FRET bursts as before, indicating the dynamics of transient TC (**compare Figs. 1c,2c and 4b**). The smFRET traces, TDP and PSH analyses showed these events converged to an $E_{TCC.5'\text{-}OH\ 1nt.FRET\ INT}$ = 0.62 state (**Fig. 4b**). Interestingly the overall FRET distribution appeared to be broader compared to previous experiments (**compare Figs. 1c,2c with 4b**).

Similar to 2Cy3-INT experiments, a fraction of DNAs contained long-lived traces marking the ST at $E_{STC.5'\text{-}OH\ 1nt.FRET\ INT}$ = 0.66 (**Supplementary Figs. 22a-e**). Therefore, we repeated the experiments at 1 s resolution for 10 min to capture the full scale of STs (**ONLINE METHODS, Fig. 4c**). A Majority (80%, N_{Total} = 609) of the traces included only TC events (**Supplementary Fig. 23a**). Control experiments with FC DNA showed the TC transitions were contaminated with FRET-INT aggregations ($E_{aggn.FRET\ INT}$ = 0.58) (**Supplementary Figs. 23b,c Supplementary Movie 6**). However, 10% of the 5'-OH 1nt Gap DNA molecules produced exceptionally stable FRET similar to the previous experiments indicating ST (**Fig. 4C, Supplementary Movie 7**). The INT aggregates also produced a small shoulder peak at ~0.4 FRET in the average analyses of these STC traces (**Fig. 4C, Supplementary Figs. 22,23**).

We noted that the TC efficiency was only half of the ~20% value obtained in previous experiments (**Supplementary Table 3**). However, this %STC value matched well with the expected 50% formation efficiency of FRET-INT in the assembly reaction (**ONLINE METHODS**). A 6% of Cy3 only STCs also followed the predicted proportion of Cy3 only INT in the assembly mixture (**Supplementary Fig. 24, ONLINE METHODS**). The missing 5% of STCs should correspond to Cy5 only INTs that appear dark under green laser excitation (**Fig. 4a, ONLINE METHODS**).

The $E_{TCC.5'\text{-}OH\ 1nt.FRET\ INT}$ distribution was slightly off compared to the $E_{STC.5'\text{-}OH\ 1nt.FRET\ INT}$ (**Supplementary Figs. 22f,23d**). This discrepancy may indicate two slightly different conformations for TCC and STC at physiological conditions. However, the difference is minor and the transition from TCC to STC may be fast (<100 ms) to materialize in individual FRET

traces. Therefore, STs appeared in abrupt single steps (**Fig. 4c, panel 1,2 Supplementary Fig. 22a-e**). In summary, FRET-INT recapitulated the other smFRET data and further confirmed the lack of large changes in the PFV INT core during an integration reaction.

DISCUSSION

Retroviruses prefer bent and flexible DNA for integration[Pruss, 1994 #12816;Bor, 1995 #12812;Bor, 1995 #12812;Lesbats, 2016 #5;Michieletto, 2019 #24;Maertens, 2010 #11431;Maertens, 2010 #11431;Jones, 2016 #15561;Maskell, 2015 #11379;Fuller, 2017 #26;Jones, 2019 #52;Wilson, 2019 #59;Yin, 2016 #16;Vanderlinden, 2019 #28;Vanderlinden, 2019 #28]. Using high-resolution experiments and well-characterized PFV INT, we tested if the integration can be enhanced by flexible or pre-bent tDNAs including SPBER associated intermediates.

8-OxoG, the principal precursor for SPBER, showed indistinguishable INT activities from undamaged DNA. The 8-OxoG:C base pairing does not provide flexibility to a 60 bp tDNA. The energy penalty for deforming a dsDNA shorter than the ~150 bp persistent length requires strong protein binding. The binding energy may then compensate for the energy required for DNA deformation[28,49]. On a 24 kbp surface tethered DNA, at physiological conditions, PFV INT diffuse fast with a 1D diffusion coefficient (D_{1D}) of ~$7x10^5$ bp^2s^{-1} before dissociating in ~2 s[26]. This DNA sliding indicates a weaker binding energy of ~1 k_BT. The value of D_{1D} also predicts that INT should traverse the entire 60 bp DNA in ~2.5 ms (**ONLINE METHODS**). Transient INT binding to 8-OxoG or undamaged DNA would be therefore, undetectable and too short for catalysis[26]. Surprisingly the TCC and STC crystal structures contain even shorter (18 bp) tDNA (**Supplementary Figs. 1,2**)[25]. The crystallization conditions may have provided enough stability for INT binding and time to undergo ST[54,55].

We presented evidence for DNA lesions such as SSDBs promoting TC and ST. It is conceivable that the local DNA flexibility introduced by SSDBs helps tDNAs to adopt

conformations present in PFV INT crystal structures at least transiently[37-40]. This model predicts several outcomes; a symmetric TC centering the damages leading to symmetric concerted integrations (**Fig. 5a**), DNA flexibility dependent TC and ST reactions, the promotion of TC and ST by GT MM, T-Bulge DNAs due to pre-bent conformations[28].

The smFRET data from two DNA substrates containing Cy5 at either side of the damages showed asymmetric FRET values indicating asymmetric TCC and STC formation (**Figs. 2c,3e Supplementary Fig. 21**). The integration site analysis further established lopsided half-site integrations (**Fig. 3g, Supplementary Figs. 18-20**). The data for GT MM and T-Bulge DNAs were indistinguishable from that of FC DNA (**Figs. 2a,b,f,3f,g, Supplementary Figs. 4,5, 11a,b,18,19**). All these observations contradict a symmetric model based on tDNA bending. Instead, the SSDBs may act as potential energy wells that hold one of the vDNA ends to stall INTs (**Fig. 5**). PFV INT may engage with these discontinuities while tracking the helicity of the tDNA in a rotational coupled diffusion[26]. The local DNA deformations at the damage site could enhance the recognition by INT, similar to the mechanism proposed for DNA repair complex Rad4/XPC[56]. The degree of DNA deformations depends on the nature of the lesions and may explain the differences observed in SM TC data. The side chain of PFV INT β 1/ β2 loop amino acid Arg 329 was shown to interact with the tDNA major grove. This resulted in a subtle sequence preference for integration[25]. It will be interesting to test if this residue may also differentially interact with various DNA damages to produce different dynamics.

The single-nt gaps are best at stalling PFV INTs regardless of their 5'-moiety (**Figs. 1c,2c, Supplementary Figs. 8,10b,11c**). However, much faster $k_{off} = 1/\tau_{on} \sim$ 2-5 s^{-1} compared to slower $k_{ST,overall}$ ~0.2-0.3 min^{-1} cause most TCCs to depart tDNA before catalyzing ST[57]. Moreover, compared to 5'-OH, 5'-P was inhibitory to ST (**Figs. 3d-g, Supplementary Figs. 15e,f,18,19, Supplementary Table 3**). Since TC kinetics for two chemical groups are the same, it is not clear how the identity of the 5'-end is relayed to the vDNA located upstream of the damage to undergo

differential ST (**Fig. 5a**). There may be undetected rearrangements in the TCC core that may involve tDNA deformations and residues such Arg 329. Mutants of Arg329 were shown to alter the subtle sequence preference for PFV INT[25], and it will be interesting use those mutants to test for a potential signaling mechanism. Alternative to these short range movements, the unstructured outer N and C terminal domains of the PFV INT[29] may undergo large changes when converting from TCC to STC. Experiments with PFV INT containing fluorescent outer domains and outer domain truncation may help finding a satisfactory answer. Besides weaker TCC formation, 5'-OH Nick and 2nt gap DNAs produced higher catalytic efficiency, indicating their ability to adopt conformations needed for efficient ST (**Figs. 2,3f,g, Supplementary Figs. 6,11,12,14, Supplementary Table 3**). Even after half-site integrations and DSDBs in tDNAs, PFV INT still produced exceptionally stable STCs without detectable structural changes (**Fig. 3-5, Supplementary Figs. 9,13-15,21**). An active removal mechanism may necessary for the removal of INs from STCs *In vivo*[32] (**Fig. 5**).

Reactive oxygen species (ROS) cause frequent formation of 8-OxoG in cells. The removal of 8-OxoG is a 'housekeeping' function of healthy cells and may provide an opportunity to enhance HIV or PFV integration. However, our data show that repair of 8-OxoG only leads to half-site PFV integrations at most ~10-30% of the time per damage per 60 bp (**Fig. 3f,g, Supplementary Figs. 18-20**). This seems to contrast previous *in vivo* HIV infection data[12,17] and may reflect the difference between HIV and PFV. However, the retroviral integration is an isoenergetic reaction even on SSDBs[32]. These damages only provide a kinetic favorability but not the free energy required to drive the reaction forward (**Fig. 5b**). Wrapping the DNA into nucleosomes *in vivo* may provide the necessary energy to drive the integration to completion while the damages kinetically localizing the INTs[23,27,30,49,51]. In that case the integration profiles will likely follow the symmetric model and its predicted outcomes described above (**Fig. 5a**). Alternatively, half-site PFV integrations may get resolved or convert to aberrant proviruses by the cellular machinery. Defective U5 ends, HIV drug-resistant IN mutants and suboptimal doses of IN inhibitor Raltegravir

have been shown to promote aberrant HIV proviruses. Further *In vivo* and *in invitro* experiments with different retroviruses are needed to expand the knowledge on retroviral integration in the context of DNA damage repair.

METHODS

Methods are included in a separate document for online publication.

Acknowledgments

The authors would like to thank Rob Levendosky, and Gregory Bowman from John Hopkins University for providing protocols for the sequencing ladder. This work was supported by the National Institutes of Health (AI126742to RF & KEY, GM121284 to KEY and GM080176 to R.F.).

Author contributions

G.S., R.F. and K.E.Y. conceived, designed, and analyzed the smFRET experiments. A.G. performed and analyzed the integration site mapping experiments. J.L. wrote the MATLAB code for smFRET data analysis. R.K.M. and Y.Y.T. produced PFV reagents. All the authors were involved in writing the paper.

Competing financial interests

The authors declare no competing financial interests.

Additional information

Any supplementary information and source data are available in the online version of the paper. Correspondence and requests for materials should be addressed to K.E.Y. or R.F.

FIGURE LEGENDS

Figure 1. Visualizing the PFV intasome target capture dynamics in real-time. (**a**) The smFRET experimental setup for visualizing the target capture by a PFV intasome labeled with Cy3 at non-transfer strands (Cy3-PFV). The intasome was introduced in real-time onto various DNAs containing A488 and Cy5. (**b**) The PFV TCC crystal structure showing the positions of Cy3 and Cy5 attachments to the vDNAs and the target DNA respectively. The estimated inter-dye distances, the corresponding FRET efficiencies, and the expected FRET value (E_{TCC}) are shown. (**c,d**) A representative intensity trajectory (panel 1), the corresponding FRET trajectory with the HMM fit (panel 2), and the PSH (panel 3) displaying intasome interactions with GC (**c**), and 1nt Gap (5'-P) DNA (**d**). *N* is the number of DNA molecules analyzed for each substrate. *n* is the total number of transitions that crossed >0.15 FRET threshold. (**e**) Normalized smFRET histograms and their Gaussian fits showing the distributions of E_{pseudo} for GC, 8-OxoG, G/T, +T DNAs. (**f,g**) Normalized smFRET histograms and their Gaussian fits showing the distributions of E_{pseudo} and E_{TCC} for (5'-P) (**f**) or (5'-OH) (**g**) gap substrates. The GC data was included for comparison. (**h,i**) The TCC lifetime (τ_{on}) (**h**) and the lifetime of the dissociated state **(τ_{off})** (**i**) for different DNA substrates. Error bars represent the errors from the fittings of dwell time distributions shown in **Extended Data Fig. 8.** (**j**) Normalized smFRET histograms and their Gaussian fits showing the distribution of E_{pseudo} and E_{TCC} for the 1nt Gap (5'-OH) DNA with or without blocking the free DNA ends.

Figure 2. Visualizing the PFV intasome catalyzed strand transfer in real-time. (**a**) The PFV STC crystal structure showing the positions of Cy3 and Cy5 attachments to the vDNAs and the target DNA respectively. The estimated inter-dye distances, the corresponding FRET efficiencies, and the expected FRET value (E_{STC}) are shown. (**b,c**) Representative intensity trajectories and corresponding FRET trajectories with the HMM fits showing the strand transfer by the Cy3-PFV.

Green arrows mark the photobleaching of Cy3. (**d,e**) The PSHs and the smFRET histograms corresponding to a collective of STC events for 1nt Gap (5'-P) (**d**) or (5'-OH) (**e**) DNA. The Gaussian fits to the histograms are shown as red lines. These data were collected at 100 ms (**b**) or 1 s (**c-e**) resolution using the smFRET setup in **Fig. 1a**. *n* and the percentages respectively indicate the number and the efficiency of strand transfer events. (**f,g**) The PSHs and the smFRET histograms for Cy3-PFV-ddA interacting with 1nt Gap (5'-OH) DNA representing TCCs at 100 ms resolution (**f**) and long events at 1 s resolution (**g**). These data were obtained using a smFRET setup similar to **Fig. 1a**. *N* is the total number of DNA molecules analyzed for each experiment. *n* and the percentages respectively indicate the number and the frequency of the long events. (**h**) The PFV STC crystal structure showing the positions of Cy3 and Cy5 attachments to the vDNAs and R-Cy5 1nt Gap (5'-OH) DNA respectively. The estimated inter-dye distances, the corresponding FRET efficiencies, and the expected FRET value (E_{TCC}) are shown. (**i**) A representative intensity trajectory and the corresponding FRET trajectory with the HMM fit showing Cy3-PFV binding to R-Cy5 1nt Gap (5'-OH) DNA. (**j**) The PSH and the smFRET histogram produced by averaging *N* = 423 smFRET traces. (**k**) A representative intensity trajectory and the corresponding FRET trajectory with the HMM fit showing Cy3-PFV strand transfer into R-Cy5 1nt Gap (5'-OH) DNA. (**l**) The PSH and the smFRET histogram corresponding to *n* = 71 such events. The percentage is the frequency of strand transfer. The data in (**i-l**) were collected at 1 s resolution using a smFRET setup similar to the **Fig. 1a.**

Figure 3. Analysis of PFV strand transfer activity. (**a**) Single-molecule and gel quantification of PFV integration activity on different DNA substrates expressed as %STC. The smFRET data were obtained using Cy3-PFV, and the gel data were obtained using PFV (**Extended Data Fig. 1**). The error bars in the smFRET data reflect $\frac{\sqrt{n}}{N} \times 100\%$, where *n* is the number of strand transfer events and *N* is the total number of DNA molecules analyzed. Error bars in the gel data are the

standard deviations from triplicates. (**b**) Representative denaturing PAGE gels (top), and integration sites mapping (bottom) showing the results from bulk integration experiments. The lengths of ssDNAs as determined using a sanger sequencing ladder are shown. The blue and red color gel bands indicate DNA fragments containing A488 and Cy5 respectively. The arrows in the mapping show the positions of strand transfer, while the numbers indicate the sizes of the products. (**c**) Gel analysis (top) and sites mapping (bottom) of Cy3-PFV(TS) (**Extended Data Table 1 and Extended Data Fig.1**) integration into different 1nt Gap (5'-OH) substrates. The lengths of ssDNAs as determined using a sanger sequencing ladder are shown. F-Cy5 is equivalent to 1nt Gap (5'-OH). The Unlbld DNA is label-free. The cyan, green and red color gel bands corresponding to DNA fragments containing A488, Cy3 and Cy5 respectively. The yellow or brown color bands contain both Cy3 and Cy5. The green color arrows show the strand transfer events that produce Cy3 only fragments, whereas yellow and brown color arrows indicate the events producing both Cy3 and Cy5 labeled fragments. The numbers accompanied by the arrows are the sizes of the resulting DNA. The circled numbers imply possible concerted integrations.

Figure 4. Probing the structural dynamics of PFV intasome during target capture and strand transfer. (**a**) The smFRET experimental setup for visualizing the structural dynamics of PFV intasome during TCC and STC formation. A PFV intasome labeled with Cy3 and Cy5 at the non-transfer strands (Cy3/Cy5 PFV) was introduced after photobleaching Cy5 on the substrate. (**b**) The PFV TCC (left) and STC (right) crystal structures showing the positions of Cy3 and Cy5 attachments to the vDNAs. The estimated inter-dye distances, the corresponding FRET efficiencies, and the expected FRET values (E_{TCC}, E_{STC}) are shown. (**c**) A representative intensity trajectory and the resulting FRET trajectory with the HMM fit showing Cy3/Cy5-PFV binding to 1nt Gap (5'-OH) DNA (left). The PSH and the smFRET histogram generated by averaging N = 526 of such FRET traces (right). (**d**) A representative intensity trajectory and the resulting FRET trajectory with the HMM fit showing Cy3/Cy5-PFV strand transfer into 1nt Gap (5'-OH) DNA (left).

The PSH and the smFRET histogram generated by averaging n = 61 of such FRET events (right). The percentage is the efficiency of strand transfer. The Gaussian fits to the histograms are shown as red lines. The TCC and STC data were collected at 100 ms and 1 s resolutions respectively.

References

1. Varmus, H. Retroviruses. *Science* **240**, 1427-35 (1988).
2. Yoder, K.E., Rabe, A.J., Fishel, R. & Larue, R.C. Strategies for Targeting Retroviral Integration for Safer Gene Therapy: Advances and Challenges. *Front Mol Biosci* **8**, 662331 (2021).
3. Poeschla, E., Corbeau, P. & Wong-Staal, F. Development of HIV vectors for anti-HIV gene therapy. *Proc Natl Acad Sci U S A* **93**, 11395-9 (1996).
4. Schroder, A.R. et al. HIV-1 integration in the human genome favors active genes and local hotspots. *Cell* **110**, 521-9 (2002).
5. Kohn, D.B., Sadelain, M. & Glorioso, J.C. Occurrence of leukaemia following gene therapy of X-linked SCID. *Nat Rev Cancer* **3**, 477-88 (2003).
6. Bowerman, B., Brown, P.O., Bishop, J.M. & Varmus, H.E. A nucleoprotein complex mediates the integration of retroviral DNA. *Genes Dev* **3**, 469-78 (1989).
7. Wei, S.Q., Mizuuchi, K. & Craigie, R. A large nucleoprotein assembly at the ends of the viral DNA mediates retroviral DNA integration. *EMBO J* **16**, 7511-20 (1997).
8. Lesbats, P., Engelman, A.N. & Cherepanov, P. Retroviral DNA Integration. *Chem Rev* **116**, 12730-12757 (2016).
9. Delelis, O., Carayon, K., Saib, A., Deprez, E. & Mouscadet, J.F. Integrase and integration: biochemical activities of HIV-1 integrase. *Retrovirology* **5**, 114 (2008).
10. Brown, P.O., Bowerman, B., Varmus, H.E. & Bishop, J.M. Retroviral integration: structure of the initial covalent product and its precursor, and a role for the viral IN protein. *Proc Natl Acad Sci U S A* **86**, 2525-9 (1989).
11. Roth, M.J., Schwartzberg, P.L. & Goff, S.P. Structure of the termini of DNA intermediates in the integration of retroviral DNA: dependence on IN function and terminal DNA sequence. *Cell* **58**, 47-54 (1989).
12. Espeseth, A.S. et al. siRNA screening of a targeted library of DNA repair factors in HIV infection reveals a role for base excision repair in HIV integration. *PLoS One* **6**, e17612 (2011).
13. Yoder, K.E. et al. The base excision repair pathway is required for efficient lentivirus integration. *PLoS One* **6**, e17862 (2011).
14. Yoder, K.E. & Bushman, F.D. Repair of gaps in retroviral DNA integration intermediates. *J Virol* **74**, 11191-200 (2000).
15. Holman, A.G. & Coffin, J.M. Symmetrical base preferences surrounding HIV-1, avian sarcoma/leukosis virus, and murine leukemia virus integration sites. *Proc Natl Acad Sci U S A* **102**, 6103-7 (2005).
16. Wu, X., Li, Y., Crise, B., Burgess, S.M. & Munroe, D.J. Weak palindromic consensus sequences are a common feature found at the integration target sites of many retroviruses. *J Virol* **79**, 5211-4 (2005).
17. Bennett, G.R. et al. Repair of oxidative DNA base damage in the host genome influences the HIV integration site sequence preference. *PLoS One* **9**, e103164 (2014).
18. Barnes, D.E. & Lindahl, T. Repair and genetic consequences of endogenous DNA base damage in mammalian cells. *Annu Rev Genet* **38**, 445-76 (2004).
19. Ciccia, A. & Elledge, S.J. The DNA damage response: making it safe to play with knives. *Mol Cell* **40**, 179-204 (2011).
20. Hegde, M.L., Hazra, T.K. & Mitra, S. Early steps in the DNA base excision/single-strand interruption repair pathway in mammalian cells. *Cell Res* **18**, 27-47 (2008).
21. Robertson, A.B., Klungland, A., Rognes, T. & Leiros, I. DNA repair in mammalian cells: Base excision repair: the long and short of it. *Cell Mol Life Sci* **66**, 981-93 (2009).
22. Friedberg, E.C. A history of the DNA repair and mutagenesis field: The discovery of base excision repair. *DNA Repair (Amst)* **37**, A35-9 (2016).

23. Pruss, D., Reeves, R., Bushman, F.D. & Wolffe, A.P. The influence of DNA and nucleosome structure on integration events directed by HIV integrase. *J Biol Chem* **269**, 25031-41 (1994).
24. Bor, Y.C., Bushman, F.D. & Orgel, L.E. In vitro integration of human immunodeficiency virus type 1 cDNA into targets containing protein-induced bends. *Proc Natl Acad Sci U S A* **92**, 10334-8 (1995).
25. Maertens, G.N., Hare, S. & Cherepanov, P. The mechanism of retroviral integration from X-ray structures of its key intermediates. *Nature* **468**, 326-9 (2010).
26. Jones, N.D. et al. Retroviral intasomes search for a target DNA by 1D diffusion which rarely results in integration. *Nat Commun* **7**, 11409 (2016).
27. Maskell, D.P. et al. Structural basis for retroviral integration into nucleosomes. *Nature* **523**, 366-9 (2015).
28. Fuller, J.R. & Rice, P.A. Target DNA bending by the Mu transpososome promotes careful transposition and prevents its reversal. *Elife* **6**(2017).
29. Jones, N.D. et al. Prototype foamy virus intasome aggregation is mediated by outer protein domains and prevented by protocatechuic acid. *Sci Rep* **9**, 132 (2019).
30. Wilson, M.D. et al. Retroviral integration into nucleosomes through DNA looping and sliding along the histone octamer. *Nat Commun* **10**, 4189 (2019).
31. Yin, Z. et al. Crystal structure of the Rous sarcoma virus intasome. *Nature* **530**, 362-6 (2016).
32. Vanderlinden, W. et al. The free energy landscape of retroviral integration. *Nat Commun* **10**, 4738 (2019).
33. Hare, S., Gupta, S.S., Valkov, E., Engelman, A. & Cherepanov, P. Retroviral intasome assembly and inhibition of DNA strand transfer. *Nature* **464**, 232-6 (2010).
34. Cook, N.J. et al. Structural basis of second-generation HIV integrase inhibitor action and viral resistance. *Science* **367**, 806-810 (2020).
35. Passos, D.O. et al. Structural basis for strand-transfer inhibitor binding to HIV intasomes. *Science* **367**, 810-814 (2020).
36. Roy, R., Hohng, S. & Ha, T. A practical guide to single-molecule FRET. *Nat Methods* **5**, 507-16 (2008).
37. Mills, J.B., Cooper, J.P. & Hagerman, P.J. Electrophoretic evidence that single-stranded regions of one or more nucleotides dramatically increase the flexibility of DNA. *Biochemistry* **33**, 1797-803 (1994).
38. Guo, H. & Tullius, T.D. Gapped DNA is anisotropically bent. *Proc Natl Acad Sci U S A* **100**, 3743-7 (2003).
39. Roll, C., Ketterle, C., Faibis, V., Fazakerley, G.V. & Boulard, Y. Conformations of nicked and gapped DNA structures by NMR and molecular dynamic simulations in water. *Biochemistry* **37**, 4059-70 (1998).
40. Zhang, Y. & Crothers, D.M. High-throughput approach for detection of DNA bending and flexibility based on cyclization. *Proc Natl Acad Sci U S A* **100**, 3161-6 (2003).
41. Bronson, J.E., Fei, J., Hofman, J.M., Gonzalez, R.L., Jr. & Wiggins, C.H. Learning rates and states from biophysical time series: a Bayesian approach to model selection and single-molecule FRET data. *Biophys J* **97**, 3196-205 (2009).
42. Senavirathne, G., Mahto, S.K., Hanne, J., O'Brian, D. & Fishel, R. Dynamic unwrapping of nucleosomes by HsRAD51 that includes sliding and rotational motion of histone octamers. *Nucleic Acids Res* **45**, 685-698 (2017).
43. Muller, H.P. & Varmus, H.E. DNA bending creates favored sites for retroviral integration: an explanation for preferred insertion sites in nucleosomes. *EMBO J* **13**, 4704-14 (1994).
44. Pruss, D., Bushman, F.D. & Wolffe, A.P. Human immunodeficiency virus integrase directs integration to sites of severe DNA distortion within the nucleosome core. *Proc Natl Acad Sci U S A* **91**, 5913-7 (1994).

45. Jeong, C. et al. MutS switches between two fundamentally distinct clamps during mismatch repair. *Nat Struct Mol Biol* **18**, 379-85 (2011).
46. Rossetti, G. et al. The structural impact of DNA mismatches. *Nucleic Acids Res* **43**, 4309-21 (2015).
47. Obmolova, G., Ban, C., Hsieh, P. & Yang, W. Crystal structures of mismatch repair protein MutS and its complex with a substrate DNA. *Nature* **407**, 703-10 (2000).
48. Hanne, J. et al. MutS homolog sliding clamps shield the DNA from binding proteins. *J Biol Chem* **293**, 14285-14294 (2018).
49. Lequieu, J., Cordoba, A., Schwartz, D.C. & de Pablo, J.J. Tension-Dependent Free Energies of Nucleosome Unwrapping. *ACS Cent Sci* **2**, 660-666 (2016).
50. Senavirathne, G., Lopez, M.A., Jr., Messer, R., Fishel, R. & Yoder, K.E. Expression and purification of nuclease-free protocatechuate 3,4-dioxygenase for prolonged single-molecule fluorescence imaging. *Anal Biochem* **556**, 78-84 (2018).
51. Mackler, R.M. et al. Nucleosome DNA unwrapping does not affect prototype foamy virus integration efficiency or site selection. *PLoS One* **14**, e0212764 (2019).
52. Sabantsev, A., Levendosky, R.F., Zhuang, X., Bowman, G.D. & Deindl, S. Direct observation of coordinated DNA movements on the nucleosome during chromatin remodelling. *Nat Commun* **10**, 1720 (2019).
53. Levitus, M. & Ranjit, S. Cyanine dyes in biophysical research: the photophysics of polymethine fluorescent dyes in biomolecular environments. *Q Rev Biophys* **44**, 123-51 (2011).
54. Janin, J. Specific versus non-specific contacts in protein crystals. *Nat Struct Biol* **4**, 973-4 (1997).
55. Acharya, K.R. & Lloyd, M.D. The advantages and limitations of protein crystal structures. *Trends Pharmacol Sci* **26**, 10-4 (2005).
56. Chakraborty, S. et al. Enhanced spontaneous DNA twisting/bending fluctuations unveiled by fluorescence lifetime distributions promote mismatch recognition by the Rad4 nucleotide excision repair complex. *Nucleic Acids Res* **46**, 1240-1255 (2018).
57. Steinfeld, J.I., Francisco, J.S. & Hase, W.L. *Chemical kinetics and dynamics*, x, 518 p. (Prentice Hall, Upper Saddle River, N.J., 1999).

Fig. 1

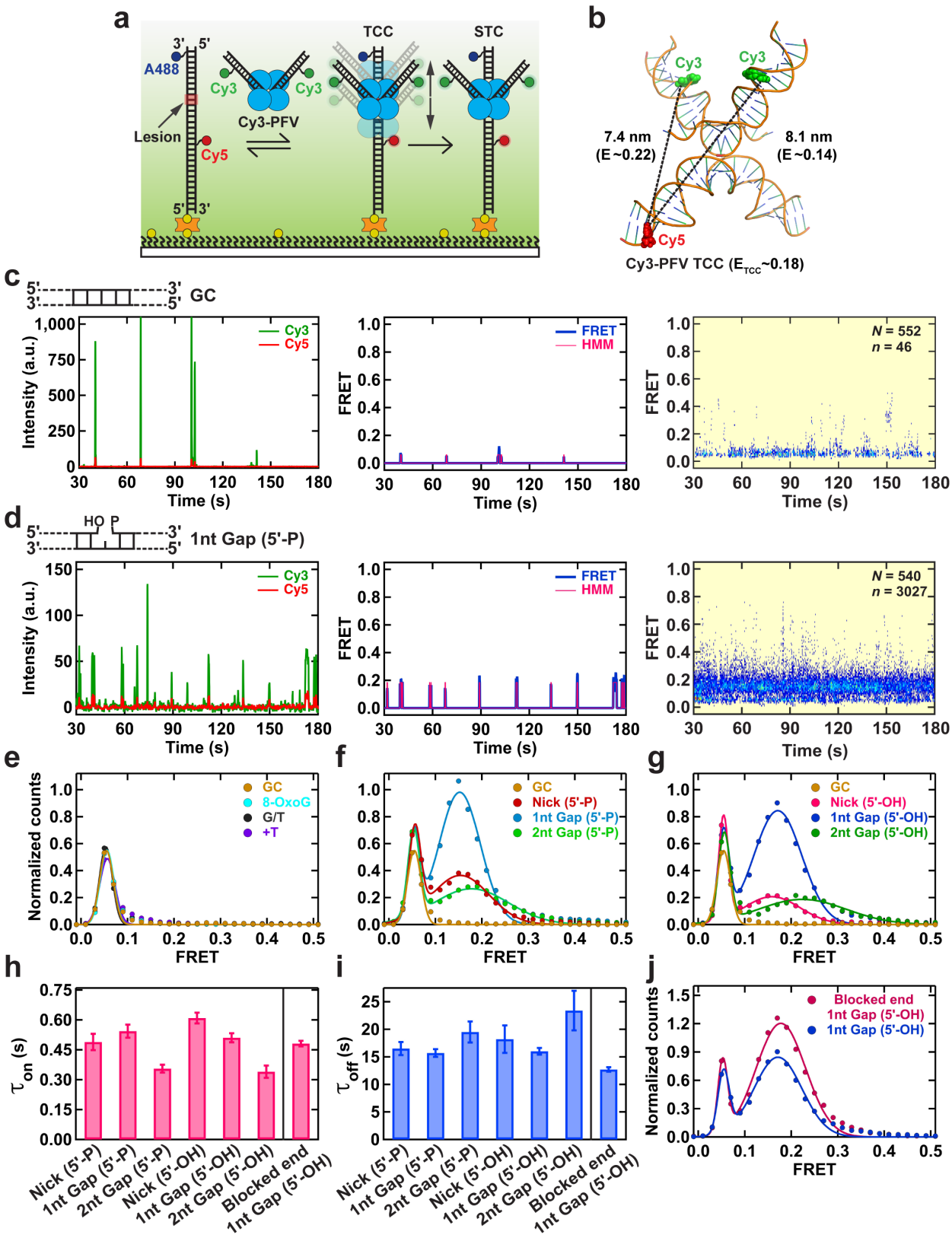

a
3' 5'
A488
Lesion
Cy5
5' 3'
Cy3
Cy3
Cy3-PFV
TCC
STC
b
Cy3
Cy3
7.4 nm
(E ~0.22)
8.1 nm
(E ~0.14)
Cy5
Cy3-PFV TCC (E_TCC~0.18)
c
5' 3'
3' 5'
GC
Intensity (a.u.)
Cy3
Cy5
Time (s)
FRET
HMM
N = 552
n = 46
d
HO P
1nt Gap (5'-P)
N = 540
n = 3027
e
Normalized counts
GC
8-OxoG
G/T
+T
FRET
f
GC
Nick (5'-P)
1nt Gap (5'-P)
2nt Gap (5'-P)
g
GC
Nick (5'-OH)
1nt Gap (5'-OH)
2nt Gap (5'-OH)
h
τon (s)
Nick (5'-P)
1nt Gap (5'-P)
2nt Gap (5'-P)
Nick (5'-OH)
1nt Gap (5'-OH)
2nt Gap (5'-OH)
Blocked end
1nt Gap (5'-OH)
i
τoff (s)
j
Blocked end
1nt Gap (5'-OH)
1nt Gap (5'-OH)

Fig. 2

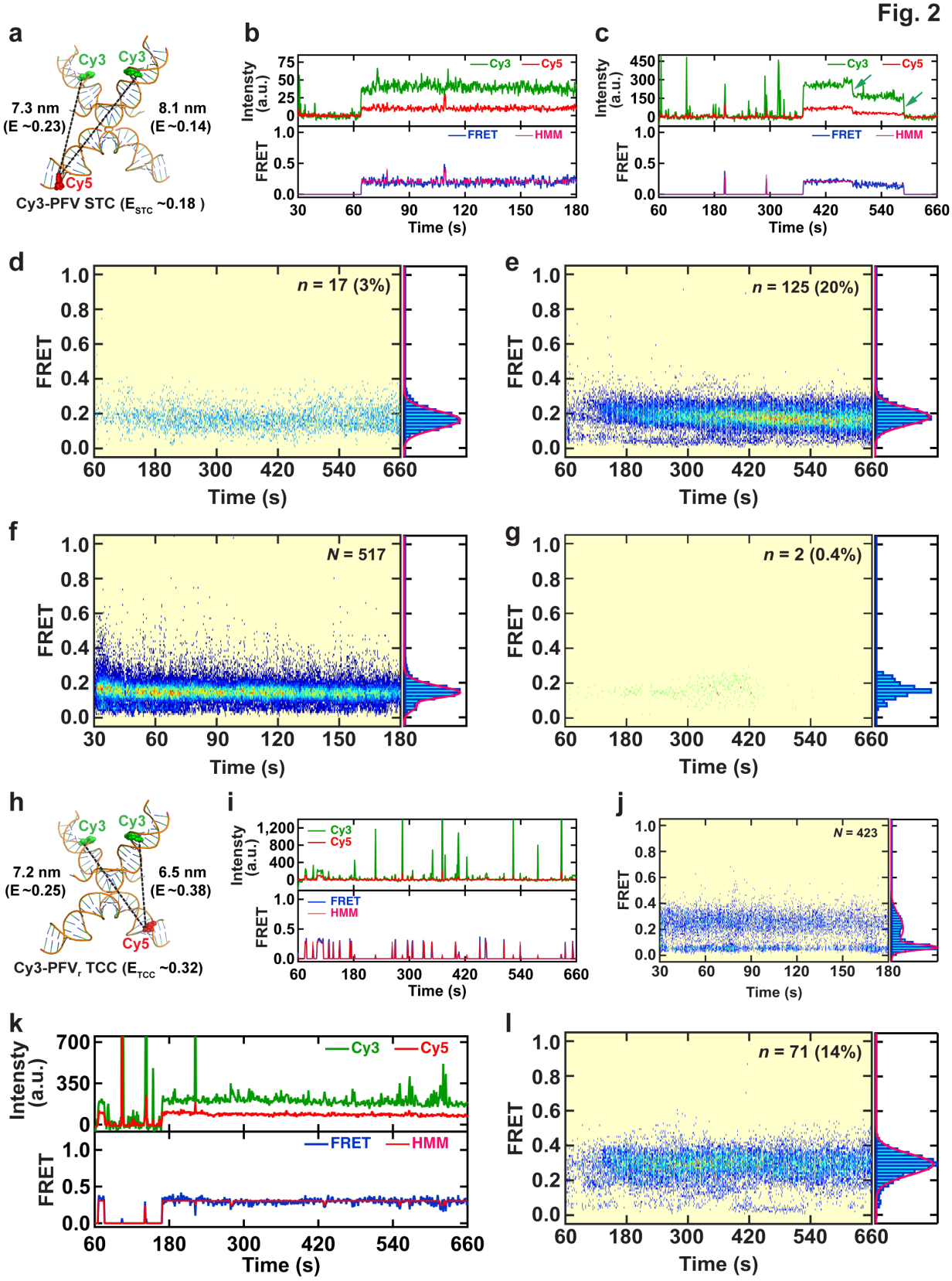

a
Cy3
Cy3
7.3 nm
(E ~0.23)
8.1 nm
(E ~0.14)
Cy5
Cy3-PFV STC (E_STC ~0.18)
b
Intensity (a.u.)
Cy3
Cy5
FRET
HMM
Time (s)
c
Intensity (a.u.)
Cy3
Cy5
FRET
HMM
Time (s)
d
n = 17 (3%)
FRET
Time (s)
e
n = 125 (20%)
FRET
Time (s)
f
N = 517
FRET
Time (s)
g
n = 2 (0.4%)
FRET
Time (s)
h
Cy3
Cy3
7.2 nm
(E ~0.25)
6.5 nm
(E ~0.38)
Cy5
Cy3-PFV_r TCC (E_TCC ~0.32)
i
Intensity (a.u.)
Cy3
Cy5
FRET
HMM
Time (s)
j
N = 423
FRET
Time (s)
k
Intensity (a.u.)
Cy3
Cy5
FRET
HMM
Time (s)
l
n = 71 (14%)
FRET
Time (s)

Fig. 3

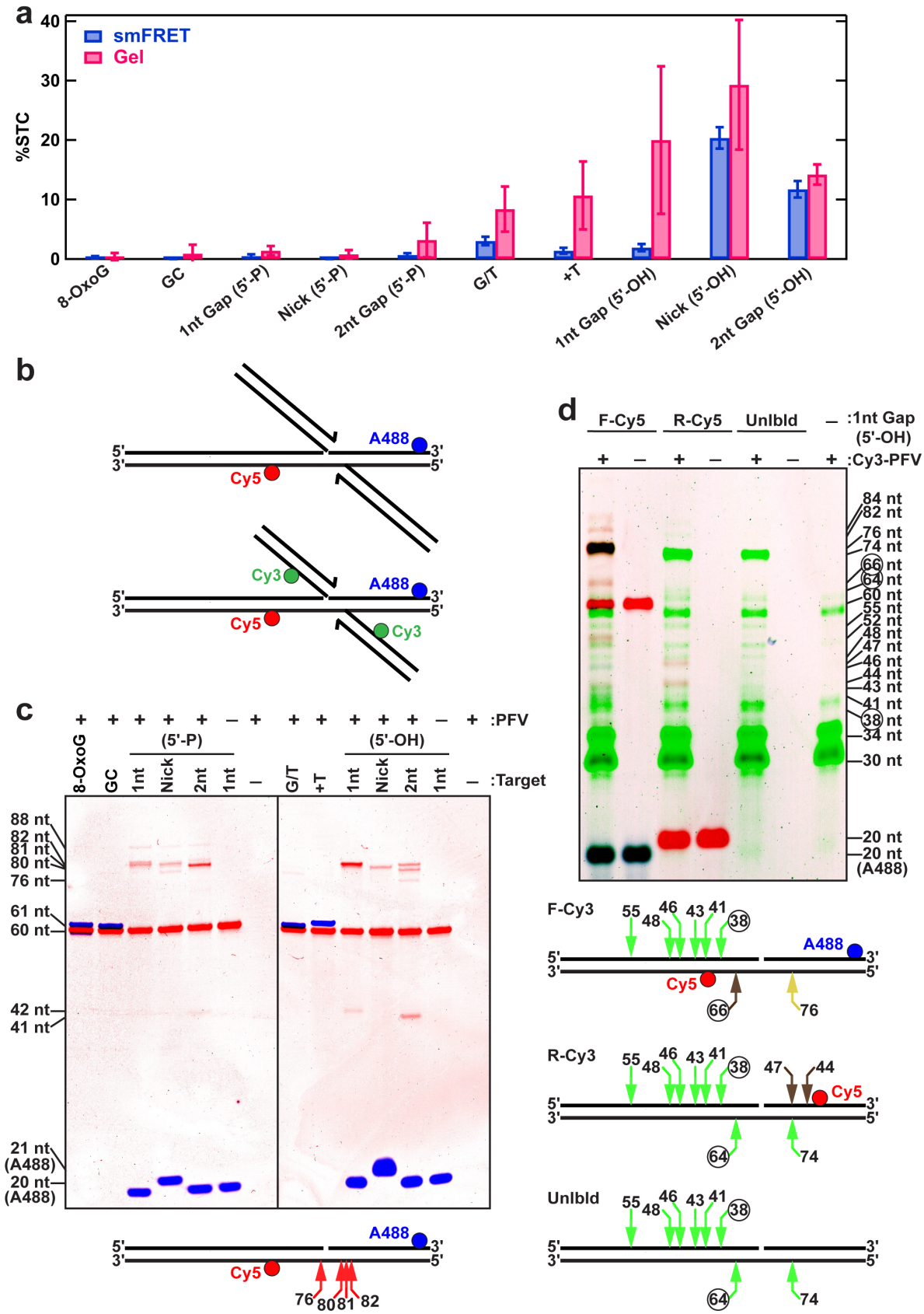
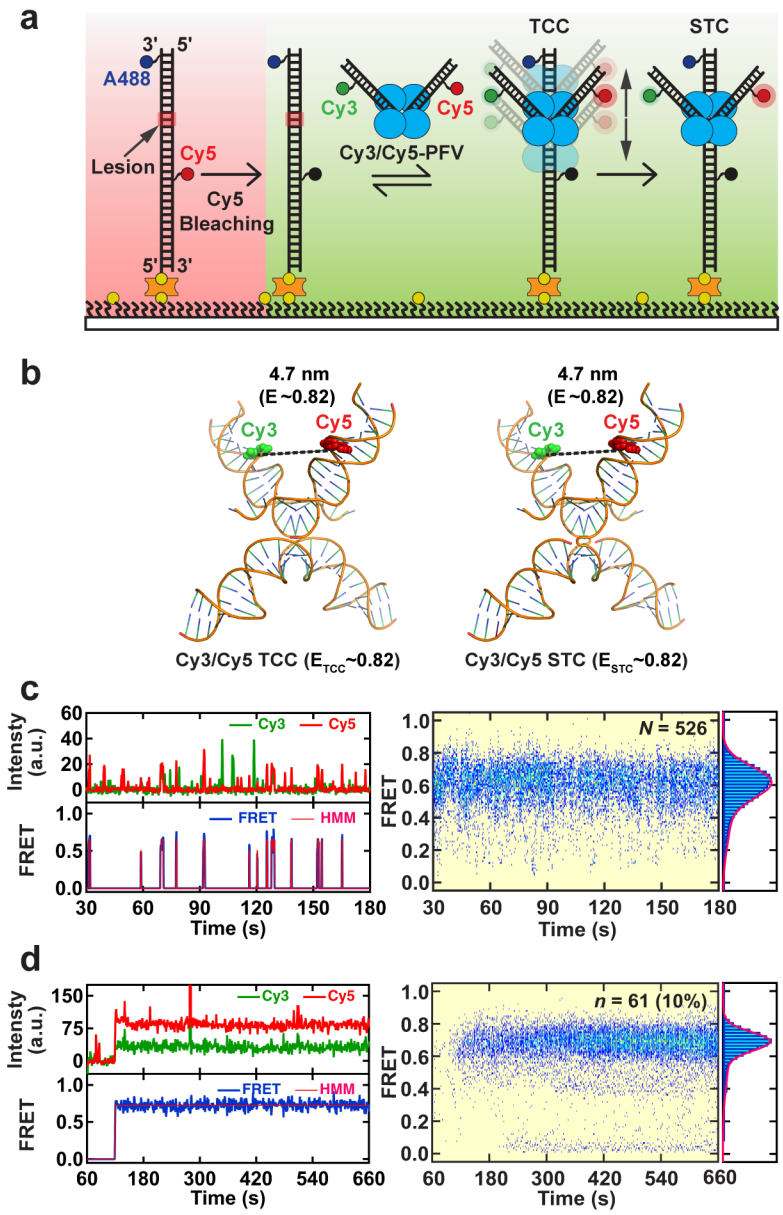

a
smFRET
Gel
%STC
0
10
20
30
40
8-OxoG
GC
1nt Gap (5'-P)
Nick (5'-P)
2nt Gap (5'-P)
G/T
+T
1nt Gap (5'-OH)
Nick (5'-OH)
2nt Gap (5'-OH)
b
A488
Cy5
Cy3
5'
3'
c
:PFV
:Target
(5'-P)
(5'-OH)
88 nt
82 nt
81 nt
80 nt
76 nt
61 nt
60 nt
42 nt
41 nt
21 nt (A488)
20 nt (A488)
76
80
81
82
d
F-Cy5
R-Cy5
Unlbld
:1nt Gap (5'-OH)
:Cy3-PFV
84 nt
82 nt
76 nt
74 nt
66 nt
64 nt
60 nt
55 nt
52 nt
48 nt
47 nt
46 nt
44 nt
43 nt
41 nt
38 nt
34 nt
30 nt
20 nt
20 nt (A488)
F-Cy3
R-Cy3
Unlbld
55
48
46
43
41
38
47
44
66
76
64
74

Fig. 4


a
3'
5'
A488
Lesion
Cy5
Cy5
Bleaching
5'
3'
Cy3
Cy5
Cy3/Cy5-PFV
TCC
STC
b
4.7 nm
(E ~0.82)
Cy3
Cy5
Cy3/Cy5 TCC (E_TCC~0.82)
4.7 nm
(E ~0.82)
Cy3
Cy5
Cy3/Cy5 STC (E_STC~0.82)
c
Intensty (a.u.)
Cy3
Cy5
FRET
FRET
HMM
Time (s)
N = 526
FRET
Time (s)
d
Intensty (a.u.)
Cy3
Cy5
FRET
FRET
HMM
Time (s)
n = 61 (10%)
FRET
Time (s)

ONLINE METHODS

Preparation of tDNAs

ssDNA DNA oligonucleotides used in this study are listed in **Extended Data Table 1**. The 8-Oxo G ssDNA was purchased from Midland Certified Reagent Company, Inc. All the other ssDNAs were purchased from Integrated DNA Technologies. ssDNAs were labeled with NSH-ester of A488 (GE Healthcare) or sulfo-Cy5 (Lumiprobe) at C6 amino modifications following the standard labeling protocol[1]. The labeled and unlabeled DNAs were separated on a C18 column (Agilent Technologies) using reverse-phase HPLC. Selected HPLC fractions were concentrated using 0.5 mL, 3 kDa molecular weight cutoff Amicon Ultra centrifugal filters after evaporating the organic solvent in a SppedVac vacuum concentrator (Thermo Fisher Scientific). The DNAs were buffer exchanged into 20 mM Tris-HCl pH 8.0, 1 mM EDTA, and stored in -20 ^{0}C until further purification.

Labeled ssDNAs prepared in this manner still significantly contaminated with truncations from the synthesis. Therefore, each ssDNA was gel purified to ~100% purity using 12% Acrylamide:Bis 19:1/7M Urea PAGE. The gel extracted DNAs were concentrated as above and stored at -20 ^{0}C in 20 mM Tris-HCl pH 8.0, 1 mM EDTA.

The substrate DNAs in **Extended Data Table 1** were obtained by annealing equimolar mixtures of corresponding ssDNAs in 20 mM Tris-HCl pH 8.0, 1 mM EDTA, 100 mM NaCl. Annealing reactions were done in a thermal cycler (Thermo Fisher Scientific) by heating the samples to 95 ^{0}C and slowly cooling down to 15 ^{0}C. Fully annealed dsDNAs were enriched by ion-exchanging HPLC on a Gen-Pak Fax column (Waters). The purity of HPLC fractions was tested by 5% Acrylamide:Bis 59:1 native PAGE. Fractions containing ~100% pure dsDNA were pooled together and concentrated as above. The concentrations of dsDNAs were determined by UV-Vis absorbance of the DNA at 280 nm, AF488 at 490 nm, and Cy5 at 650 nm using a nanodrop device (Thermo Fisher Scientific). These DNAs were stored at -20 ^{0}C in 20 mM Tris-HCl pH 8.0, 1 mM EDTA, 100 mM NaCl.

PFV intasome assembly

PFV intasomes were assembled as described previously using recombinant integrase and dsDNA mimicking the U5 vDNA ends (**Extended Data Table 1**)[2-4]. vDNAs were prepared as described above using appropriate ssDNA oligos (**Extended Data Table 1**). The intasome assemblies were performed by salt dialysis followed by chromatographic purification on a size exclusion column (SEC). The activities of SEC fractions were tested using a standard plasmid-based integration assays before flash freezing and storing at -80 ^{0}C for further use. Catalytically deficient Cy3-PFV-ddA construct was assembled in the same way using integrase and a vDNA containing 3'-ddA (**Extended Data Table 1**)[5,6]. The intasome containing the Cy3 and Cy5 FRET pair (Cy3/Cy5-PFV) was prepared by mixing an equimolar Cy3 and Cy5 vDNAs during the assembly reaction. The expected outcomes for different species are: 50% Cy3 and Cy5 containing intasomes, 25% two Cy3 intasomes, and 25% two Cy5 intasomes.

***$R_{Cy3-Cy5}$* and *$E_{TCC\ or\ STC}$* estimations from the crystal structures**

The PFV intasome TCC and STC crystal structures (PDB code: 3OS1, 3OS0) were downloaded from the PDB data base. To measure the distances ($R_{Cy3-Cy5}$) between the fluorophores in vDNAs and the target DNAs, we extended the target DNA chains in both crystal structures due to absence of some base pairs in the structures compared to the target DNA used in the smFRET experiments. This was done by loading the structures into the discover studio and generating a B-helix of double strands DNA structures (sequence: CCCGAG), which then ligated to the ends of the target DNA. The resultant structures were then opened in PyMol 2.1 (Schrödinger, Inc) to estimate inter dye distances ($R_{Cy3-Cy5}$) for various complexes. Although the dyes were attached through C6 amino linkers, the $R_{Cy3-Cy5}$ were measured between DNA bases using the distance measurement module in PyMol. These $R_{Cy3-Cy5}$ were converted to corresponding FRET efficiencies using the Förster equation, assuming the orientation factor (κ^2) is 2/3 (complete free rotations of the dyes) and the Förster radius for Cy3-Cy5 ($R_{0,Cy3-Cy5}$) is 6 nm[7].

$$E = \frac{1}{1+\left(\frac{R_{Cy3-Cy5}}{R_{0,Cy3-Cy5}}\right)^6} \quad (1)$$

For the construct that showed two different FRET state because of the presence of two FRET pairs the expected FRET (E_{TCC} or E_{STC}) was determined as arithmetic mean of the two states.

smFRET imaging

smFRET imaging was done on a home built inverted fluorescence microscope (Olympus), as described previously[1]. Prism-based total internal reflection (TIRF) of a green (532 nm) or red (635 nm) laser was used to excite fluorophores attached to a surface of flow cell. The fluorescence from individual fluorophores was collected through a 60X water immersion objective (Olympus) and directed onto an emCCD chip (Princeton Instruments) after magnifying another 1.6X and separating Cy3, Cy5 emissions using a Dual View device (Photometrics).

The quartz surface of the flow cells was passivated with a 1:20 ratio of biotin-PEG and methoxy-PEG (5000 MW, Layson Bio, Inc). Biotin-PEG was used to immobilize target DNAs by biotin-neutravidin-biotin interactions at ~0.2 molecules/μm^2 surface density. The methoxy-PEG brush minimizes the surface interactions of biomolecules[1,7]. The imaging buffer (Buffer-I) for all the experiments consisted of 30 mM Bis-tris propane, pH 7.5, 110 mM NaCl, 2 mM $MgSO_4$, 4 μM $ZnCl_2$, 0.1 mM DTT, 0.2 mg/mL BSA, 0.02% IGPEPAL. Buffer-I also included a cocktail of saturated (~2 mM) Trolox and an oxygen scavenging system (OSS) to minimize photo-blinking and photobleaching of fluorophores respectively[1]. The OSS consisted of 25 mM protocatechuic acid (PCA) and 20 nM protocatechuate dioxygenase (PCD)[1,8]. All the experiments were performed at 24 ± 1 ^{0}C.

smFRET target capture assays

The imaging for target capture was done at 100 ms time resolution to capture transient events. Single-molecule movies were initiated by exciting Cy5-DNA in Buffer-I with a 635 nm red laser at ~2 mW. After 20 s, the excitation was switched to a 532 nm green laser maintained at ~6 mW. 10 s after the green laser exposure, 5 nM Cy3-PFV in Buffer-I was infused in real-time into the flow cells. Data recording was continued under continuous green laser excitation for 2.5 min from the injection (**Extended Data Fig. 2a**).

Cy3/Cy5-PFV experiments were performed the same way with the following modifications. The initial red laser exposure was used to bleach Cy5 in the field of view (FOV) entirely within 20 s. The fast photobleaching was achieved by eliminating the OSS in Buffer-I. The subsequent intasome injection and imaging were done exactly as above with OSS in Buffer-I (**Extended Data Fig. 2b**).

smFRET strand transfer assays

Strand transfer assays were recorded with 1 s time resolution and at lower laser powers to improve fluorophore lifetimes, allowing longer observations. Movies were initiated by Cy5 excitation with the red laser at ~1 mW. After 30 s, the excitation was switched to the green laser maintained at ~4 mW. 30 s after the green laser exposure, 5 nM Cy3-PFV in Buffer-I was infused in real-time into the flow cell. The imaging was continued under continuous green laser excitation for 10 min from the injection (**Extended Data Fig. 2c**).

Cy3/Cy5-PFV experiments were performed the same way with the following modifications. The initial red laser exposure was used to bleach Cy5 in the FOV entirely within 30 s. The fast photobleaching was achieved by eliminating the OSS in Buffer-I. The subsequent intasome injection and imaging were done exactly as above with OSS in Buffer-I (**Extended Data Fig. 2d**).

The Rev-Cy5 1nt Gap (5'-OH) experiments suffer from fast photobleaching and photophysical fluctuations of Cy5 due to its closeness to purines (A,G)[9] in the DNA sequence

(**Extended Data Table 1**). Therefore, we performed only 1 s resolution experiments at ~2 mW reduced power for 532 nm laser in regular experiments and at ~4 mW for fast photobleaching.

Initial processing of SM movies

Single-molecule movies were acquired using the Micro-Manager imaging software as described previously (**Supplementary Movies 1-7**)[1]. We used a custom-written MATLAB (MathWorks) program to extract intensity and FRET data from these movies as follows.

Prior to the analysis the program requires mapping of the Cy3 and Cy5 channels. This was accomplished by using the emissions of 0.2 μm crimson carboxylate modified microspheres (Thermo Fisher Scientific). Then using the program, the movies were manually inspected to identify single-molecules as diffraction-limited spots. The initial Cy5 excitation provided excellent signal to noise (S/N) for identifying individual DNA molecules. Each molecule is then marked with a circle of adjustable radius. However, the crowding of DNAs in the FOV may produce overlapping circles. To avoid interference, we chose smaller three-pixel radius that encircled an entire molecule without including neighbors. In addition, we set a five pixels distance cutoff from a molecule to another or to the edges of the FOV. The eccentricity (a measure of circularity) cutoff for the spots was chosen as 0.2. All these stringent selection criteria allowed us to identify ~500-600 well separated DNAs per movie.

Generation of SM traces and HMM analysis

After the selection algorithm, pixel intensities within a given circle were integrated to generate raw Cy3 and Cy5 intensities. This procedure was continued for all the movie frames to collect emissions as a function of time and to build intensity traces without background corrections **(Extended Data Fig. 2a-d, panel 1).** The graphical user interface (GUI) of our MATLAB program allowed direct comparison of these traces with their corresponding overlapped Cy3, Cy5 spots in the single-molecule movies.

The infusion of Cy3-PFV led to marked increase in both Cy3 and Cy5 backgrounds. These intensity jumps were used as references for trace truncations and background corrections **(Extended Data Fig. 2a-d, panel 2).** The 100 ms traces were smoothed with a three-point averaging and the 1 s traces were not smoothed. Truncated traces were subjected to vbFRET HMM algorithm built into our program[10] to first adjust the backgrounds and then to identify states in FRET traces. FRET was calculated from corrected intensities (I) as $I_{Cy5}/(I_{Cy3}+I_{Cy5})$. Four states were used as the initial guess for HMM. When an intasome is not bound to a target DNA, both I_{Cy3} and I_{Cy5} approached zero in the background corrected intensity traces. This led to erratic fluctuations in FRET for intasome unbound regions **(Extended Data Fig. 2a-d, panel 3).** As a solution, the FRET and the corresponding HMM state was assigned zero when either I_{Cy3} or I_{Cy5} approached zero. These 'cleaned up' FRET traces and their HMM fittings were manually inspected for correct fittings **(Extended Data Fig. 2a-d, panel 4).** Moreover, traces containing Cy3 photobleaching indicated negligible spectral bleed-through to the Cy5 channel. Therefore, Cy3 bleed-though correction was omitted from I_{Cy5}. Also, intasome aggregates with saturating Cy3 only contributed to a minor Cy5 signal and 0.06 FRET.

Further selection and categorization of smFRET traces

Intasome injections sometimes created photophysical fluctuations at the beginning of some of the FRET traces. When necessary these aberrant data were excluded by truncating the traces (up to 100 frames). In other cases, the surface binding of an intasome within a target DNA selection circle resulted in steady Cy3 signals without DNA binding. However, visual inspection of movies allowed us to identify these pseudo-events and elimination by truncation. Only molecules that contained at least 100 frames worth data were included in the final analyses.

Traces containing FRET or colocalized Cy3-Cy5 for prolonged time windows (up to minutes) were categorized as STCs. Traces that lack these long events were categorized as TCCs. The data for these two categories were analyzed and presented separately. $\%STC_{FRET}$

and their error estimates ($\Delta\%STC_{FRET}$) were calculated from the number of DNAs that showed strand transfer events (*n*) and total number of DNA molecules (*N*), using equation (2) and (3) respectively[11].

$$\%STC_{FRET} = {}^{n}/_{N} \times 100\% \quad (2)$$

$$\Delta\%STC_{FRET} = {}^{\sqrt{n}}/_{N} \times 100\% \quad (3)$$

Transition Density Plots (TDPs) for target capture smFRET traces

Using our MATLAB program TDPs were generated by compiling idealized FRET traces resulting from the HMM fittings[10]. The peaks in a TDP represent transitions from a given initial FRET state to a given final FRET state. Also, the peak heights represent total number of transitions between corresponding states[10]. The use of 4 states as an initial guess for HMM, resulted in occasional overfitting of the data and low populated off diagonal peaks.

Post-Sync Histogram analysis

Using our MATLAB program FRET traces were synchronized to the injection of intasomes to create PSH plots. The assigned 0 FRET corresponding to the intasome unbound state was eliminated for clarity. For 100 ms resolution experiments, 0.01 FRET bins and 300 ms time bins were used to construct PSHs. For 1 s experiments, 0.01 FRET bins and 0.91 s time bins were used to construct PSHs.

smFRET histograms

smFRET histograms were built with 0.2 FRET bins by molecule and time-averaging the >0 FRET states in the traces. The heights of the bins (counts) depend on the number of target DNA molecules included in the analysis. For TCC histograms, this dependency was eliminated by recalculating the counts per DNA. To directly compare between Cy3-PFV experiments the

histograms were re-normalized to the number of events in the pseudo-FRET ~0.06 peak. This was done by integrating the area under the pseudo-FRET peak to calculate the raw counts and then dividing the whole histogram by that number.

All the smFRET histograms were prepared and fitted using Igor Pro 8 (WaveMetrics). The following Gaussian equation was used for fittings.

$$y = \sum_{i=1}^{n} 1 \Big/ \sqrt{2\pi\sigma_i^2}\, e^{\frac{(x-x_{0,i})^2}{2\sigma_i^2}} \tag{4}$$

In most cases a single Gaussian (n =1) fit the data well. Occasionally n = 2 or n = 3 were needed to get the best fits.

Dwell time and transition count histograms

The dwell times of the bound (t_{on}), and unbound (t_{off}) states were extracted from the TCC traces using the HMM fittings, as shown in **Extended Data Fig. 8a**. For Cy3-PFV intasome experiments, using a FRET threshold of 0.1, the FRET > 0.1 was defined as bound, and 0 FRET was defined as unbound (**Extended Data Fig. 8a**). The dwell times from many molecules were binned according to the web-based bin optimization algorithm (https://www.neuralengine.org//res/histogram.html), and histograms were generated in Igor Pro 8 (WaveMetrics). Following single exponential functions were used to calculate the average bound (τ_{on}) and unbound (τ_{off}) dwell times.

$$Counts_{bound} = A_1 e^{-t_{on}/\tau_{on}} \tag{5}$$

$$Counts_{unbound} = A_2 e^{-t_{off}/\tau_{off}} \tag{6}$$

The transition counts for Cy3-PFV TCC traces were calculated using FRET thresholds; 0.1. Upward FRET jumps that crossed the threshold was defined as transitions. Transition counts from individual traces were binned with 1-transition bins to build histograms in Igor pro 8 (WaveMetrics).

Overall ST times

Strand transfer times (t_{ST}) is defined as the time from the intasome injection to the first frame of that a stable FRET or Cy3 signal appear (**Extended Data Table 6)**. A MATLAB script was used to extract t_{ST} from individual traces. The mean ($\bar{t}_{ST}$) and standard deviation (σ_{ST}.) for a given target DNA was calculated as,

$$\bar{t}_{ST} = \frac{\sum_{i=1}^{N} t_{ST}}{n} \quad (7)$$

$$\sigma_{ST} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(t_{ST} - \bar{t}_{ST})^2} \quad (8)$$

Where *n* is the number DNA molecules that showed strand transfer events.

Ensemble biochemical experiments

Plasmid-based integration assay

This assay was performed using the standard protocol described previously[2,3]. Briefly, 25 nM of PFV INTs were incubated with 50 ng of supercoiled (SC) plasmid DNA (pGEM®-T Easy) in 10 mM Bis-tris propane, pH7.5, 110 mM NaCl, 5 mM $MgSO_4$, 4 μM $ZnCl_2$ 10 mM DTT in 15 μl reactions for 5 min at 37 ^{0}C. The reactions were terminated by adding 0.1% SDS, 2.5 mM EDTA, 1 mg/ml proteinase K and incubated at 55 ^{0}C for an hour. The products were mixed with 5% glycerol before resolving on a 1% agarose gel in 1X TAE at 105 V for an hour. Gels were stained with 0.1 μg/mL ethidium bromide and scanned on a Sapphire Biomolecular Imager. The quantification of supercoiled (SC) and linear DNA was done using the AzureSpot software (Azure Biosystems) as described previously[2,3] and presented in **Extended Data Fig. 1**.

Integration site mapping experiments

Integration site mapping experiments were also performed as described previously[4] using target DNAs in **Extended Data Table 2**. Briefly, 10 nM unlabeled PFV intasomes were incubated with

5 nM targets in 30 mM Bis-tris propane, pH7.5, 110 mM NaCl, 2 mM $MgSO_4$, 4 μM $ZnCl_2$, 10 mM DTT in 15 μl reactions for 5 min at 37 °C. Reactions were terminated by adding 0.1% SDS, 2.5 mM EDTA, 1 mg/ml proteinase K and incubating at 37°C for 20 min. Deproteinated samples were denatured by heating to 95 °C with 50% formamide for 10 min, then iced. Sequencing ladders were generated using Thermo Sequenase Dye Primer Manual Cycle Sequencing Kit 792601 according to the manufacturer's directions[12] with pDrive-601NPS[1] as the template and Cy5, Alexa488 or Cy3 labeled oligos complementary to the 5' end of 601NPS as primers. 1 pmol primer and I pmol template per reaction were used. Annealing temperature was 55 °C with 45 cycles. Sequencing reactions were diluted with an equal volume of formamide, heated to 75 °C for 10 min and stored at -20 °C. Products were resolved on 0.8 mm 8% Acrylamide:Bis 19:1/7M Urea PAGE gels in 1XTBE at 40 W for various times. Gels were scanned on a Sapphire Biomolecular Imager and quantified using AzureSpot software (Azure Biosystems). Alignment of the products with the sequencing ladder was used to determine the integration sites. The integration efficiency was calculated as the fractional intensity of a band relative to the lane. Reactions without INTs were used as controls for target input.

The site mapping experiments comparing different 1nt Gap (5'-OH) DNAs were performed as described above with the following modifications in the reaction conditions. An intasome containing shorter vDNAs and Cy3 at the transfer strand (Cy3-PFV(TS), **Extended Data Table 1**) at 100 nM was reacted with 25 nM DNA targets to improve the efficiency of the reaction. In addition, the reaction buffer was supplemented with 5 nM PCA to improve the lifetime of the intasome[3]. The PAGE separation of resultant products, the site size determination and imaging were done same as above.

References

1. Senavirathne, G., Mahto, S.K., Hanne, J., O'Brian, D. & Fishel, R. Dynamic unwrapping of nucleosomes by HsRAD51 that includes sliding and rotational motion of histone octamers. *Nucleic Acids Res* **45**, 685-698 (2017).
2. Jones, N.D. et al. Retroviral intasomes search for a target DNA by 1D diffusion which rarely results in integration. *Nat Commun* **7**, 11409 (2016).
3. Jones, N.D. et al. Prototype foamy virus intasome aggregation is mediated by outer protein domains and prevented by protocatechuic acid. *Sci Rep* **9**, 132 (2019).
4. Mackler, R.M. et al. Nucleosome DNA unwrapping does not affect prototype foamy virus integration efficiency or site selection. *PLoS One* **14**, e0212764 (2019).
5. Maertens, G.N., Hare, S. & Cherepanov, P. The mechanism of retroviral integration from X-ray structures of its key intermediates. *Nature* **468**, 326-9 (2010).
6. Fuller, J.R. & Rice, P.A. Target DNA bending by the Mu transpososome promotes careful transposition and prevents its reversal. *Elife* **6**(2017).
7. Roy, R., Hohng, S. & Ha, T. A practical guide to single-molecule FRET. *Nat Methods* **5**, 507-16 (2008).
8. Senavirathne, G., Lopez, M.A., Jr., Messer, R., Fishel, R. & Yoder, K.E. Expression and purification of nuclease-free protocatechuate 3,4-dioxygenase for prolonged single-molecule fluorescence imaging. *Anal Biochem* **556**, 78-84 (2018).
9. Levitus, M. & Ranjit, S. Cyanine dyes in biophysical research: the photophysics of polymethine fluorescent dyes in biomolecular environments. *Q Rev Biophys* **44**, 123-51 (2011).
10. Bronson, J.E., Fei, J., Hofman, J.M., Gonzalez, R.L., Jr. & Wiggins, C.H. Learning rates and states from biophysical time series: a Bayesian approach to model selection and single-molecule FRET data. *Biophys J* **97**, 3196-205 (2009).
11. Steinfeld, J.I., Francisco, J.S. & Hase, W.L. *Chemical kinetics and dynamics*, x, 518 p. (Prentice Hall, Upper Saddle River, N.J., 1999).
12. Sabantsev, A., Levendosky, R.F., Zhuang, X., Bowman, G.D. & Deindl, S. Direct observation of coordinated DNA movements on the nucleosome during chromatin remodelling. *Nat Commun* **10**, 1720 (2019).

Extended Data

DNA Breaks and Gaps Target Retroviral Integration

Gayan Senavirathne[1], Anne Gardner[1], James London[1], Ryan K. Messer[1], Yow Yong Tan[1], Richard Fishel[1,*] and Kristine E. Yoder[1,*]

[1]Department of Cancer Biology and Genetics, The James Comprehensive Cancer Center and Ohio State University Wexner Medical Center, Columbus, OH 43210, US

*Correspondence may also be addressed to: R. Fishel (rafishel@gmail.com) or K.E. Yoder (yoder.176@osu.edu)

Extended Data Table 1. Target DNAs and vDNAs used in this study.

Oligo names	Sequence	DNA Substrate
KEY22 KEY21	5′-Bio-CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTCTAGCACCGCTTAA-3′ 3′-GACCTCTTAGGGCCACGGCTCCGGCGAGTTAACCAGCATCTGTCGAGATCGTGGCGAATT-5′	GC
KEY25 KEY21	8-OxoG 5′-Bio-CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTCTAGCACCGCTTAA-3′ 3′-GACCTCTTAGGGCCACGGCTCCGGCGAGTTAACCAGCATCTGTCGAGATCGTGGCGAATT-5′	8-OxoG
KEY22 KEY28	5′-Bio-CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTCTAGCACCGCTTAA-3′ 3′-GACCTCTTAGGGCCACGGCTCCGGCGAGTTAACCAGCATTTGTCGAGATCGTGGCGAATT-5′	G/T
KEY29 KEY21	T 5′-Bio-CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTCTAGCACCGCTTAA-3′ 3′-GACCTCTTAGGGCCACGGCTCCGGCGAGTTAACCAGCATCTGTCGAGATCGTGGCGAATT-5′	+T
KEY23 KEY24 (5'-P) KEY 21	HO P 5′-Bio-CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTCTAGCACCGCTTAA-3′ 3′-GACCTCTTAGGGCCACGGCTCCGGCGAGTTAACCAGCATCTGTCGAGATCGTGGCGAATT-5′	Nick (5'-P)
KEY23 KEY26 (5'-P) KEY 21	HO P 5′-Bio-CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTA ACAGCTCTAGCACCGCTTAA-3′ 3′-GACCTCTTAGGGCCACGGCTCCGGCGAGTTAACCAGCATCTGTCGAGATCGTGGCGAATT-5′	1nt Gap (5'-P)
KEY27 KEY26 (5'-P) KEY 21	HO P 5′-Bio-CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGT ACAGCTCTAGCACCGCTTAA-3′ 3′-GACCTCTTAGGGCCACGGCTCCGGCGAGTTAACCAGCATCTGTCGAGATCGTGGCGAATT-5′	2nt Gap (5'-P)
KEY23 KEY24 (5'-OH) KEY 21	HO OH 5′-Bio-CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTAGACAGCTCTAGCACCGCTTAA-3′ 3′-GACCTCTTAGGGCCACGGCTCCGGCGAGTTAACCAGCATCTGTCGAGATCGTGGCGAATT-5′	Nick (5'-OH)
KEY23 KEY26 (5'-OH) KEY 21	HO OH 5′-Bio-CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTA ACAGCTCTAGCACCGCTTAA-3′ 3′-GACCTCTTAGGGCCACGGCTCCGGCGAGTTAACCAGCATCTGTCGAGATCGTGGCGAATT-5′	1nt Gap (5'-OH)
KEY27 KEY26 (5'-OH) KEY 21	HO OH 5′-Bio-CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGT ACAGCTCTAGCACCGCTTAA-3′ 3′-GACCTCTTAGGGCCACGGCTCCGGCGAGTTAACCAGCATCTGTCGAGATCGTGGCGAATT-5′	2nt Gap (5'-OH)
KEY23 Cy5-KEY26 (5'-OH) KEY 21	HO OH 5′-Bio-CTGGAGAATCCCGGTGCCGAGGCCGCTCAATTGGTCGTA ACAGCTCTAGCACCGCTTAA-3′ 3′-GACCTCTTAGGGCCACGGCTCCGGCGAGTTAACCAGCATCTGTCGAGATCGTGGCGAATT-5′	R-Cy5 1nt Gap (5'-OH)
KEY675 Cy3-KEY616	5′-CTGTTCGGGCGCCACTCAATATACAAAATTCCATGACA-3′ 3′-GACAAGCCCGCGGTGAGTTATATGTTTTAAGGTACTGTTA-5′	Cy3-PFV
3'-ddA KEY675 Cy3-KEY616	5′-CTGTTCGGGCGCCACTCAATATACAAAATTCCATGAC[2′3′ddA]-3′ 3′-GACAAGCCCGCGGTGAGTTATATGTTTTAAGGTACTGTTA-5′	Cy3-PFV-ddA
KEY675 KEY616	5′-CTGTTCGGGCGCCACTCAATATACAAAATTCCATGACA-3′ 3′-GACAAGCCCGCGGTGAGTTATATGTTTTAAGGTACTGTTA-5′	PFV
Cy3-KEY675(2) KEY616	5′-GCGCCACTCAATATACAAAATTCCATGACA-3′ 3′-CGCGGTGAGTTATATGTTTTAAGGTACTGTTA-5′	Cy3-PFV (TS)
KEY675 Cy5-KEY616	5′-CTGTTCGGGCGCCACTCAATATACAAAATTCCATGACA-3′ 3′-GACAAGCCCGCGGTGAGTTATATGTTTTAAGGTACTGTTA-5′	Cy5-PFV

The constituent DNA oligonucleotides, the target DNA substrates, and the vDNAs used in this study. The labeling positions of A488, Cy3 and Cy5 are marked with blue, green, and red colors respectively. Bio = 5'-Biotin, HO = Hydroxyl, P = Phosphate.

Extended Data Table 2. E_{TCC} and E_{STC} values for different smFRET experiments.

DNA Substrate	$E_{TCC} \pm \sigma_{TCC}$	$E_{STC} \pm \sigma_{STC}$
GC	–	0.54 ± 0.06
G/T	–	0.32 ± 0.08
Nick (5'-P)	0.15 ± 0.08	0.27 ± 0.04, 0.10± 0.04
1nt Gap (5'-P)	0.16 ± 0.05	0.16 ± 0.05
2nt Gap (5'-P)	0.18 ± 0.10	0.21 ± 0.05
Nick (5'-OH)	0.17 ± 0.07	0.17 ± 0.05
1nt Gap (5'-OH)	0.19 ± 0.05	0.18 ± 0.05
2nt Gap (5'-OH)	0.23 ± 0.11	0.24 ± 0.05
Blocked end 1nt Gap (5'-OH)	0.18 ± 0.05	0.19 ± 0.05
1nt Gap (5'-OH) + Cy3-PFV-ddA	0.15 ± 0.05	-
R-Cy5 1nt Gap (5'-OH)	0.21 ± 0.10	0.29 ± 0.06

These values were calculated as described in **ONLINE METHODS** by fitting TCC and STC FRET distributions for each experiment with a single or a combination of two Gaussians. E and σ indicate the centers and the standard deviations of the peaks respectively.

Extended Data Table 3. Statistics for the number of counts under the pseudo-FRET peaks of the TCC histograms recorded at 100 ms resolution.

DNA Substrate	Number of DNA molecules (N)	Pseudo FRET	Pseudo FRET
		Total number of counts (n)	Number of counts per DNA (n/N)
GC	552	283	0.513
8-OxoG	496	258	0.521
G/T	574	342	0.596
+T	606	442	0.730
Nick (5'-P)	549	314	0.571
1nt Gap (5'-P)	540	381	0.706
2nt Gap (5'-P)	474	322	0.680
Nick (5'-OH)	558	346	0.621
1nt Gap (5'-OH)	496	356	0.718
2nt Gap (5'-OH)	540	194	0.359
Blocked end 1nt Gap (5'-OH)	483	297	0.614

$\bar{x} = 0.603, \sigma = \pm 0.110$

These numbers were calculated as described in **ONLINE METHODS** by fitting FRET distributions for each DNA with a single or a combination of two Gaussians and integrating the area under the curve for the 0.06 FRET peak. $\bar{x}$ and σ show the average and the standard deviation for the counts per DNA calculated from all the DNA substrates.

Extended Data Table 4. Statistics for the number of counts under the pseudo-FRET peaks of the TCC histograms recorded at 1 s resolution.

DNA Substrate	Number of DNA molecules (N)	Pseudo FRET	Pseudo FRET
		Total number of counts (n)	Number of counts per DNA (n/N)
GC	583	606	1.04
8-OxoG	624	2359	3.78
G/T	613	1765	2.88
+T	531	1529	2.88
Nick (5'-P)	583	2268	3.89
1nt Gap (5'-P)	542	2000	3.69
2nt Gap (5'-P)	498	1604	3.22
Nick (5'-OH)	614	1105	1.8
1nt Gap (5'-OH)	489	2308	4.72
2nt Gap (5'-OH)	534	951	1.78

$\bar{x} = 2.97, \sigma = \pm 1.14$

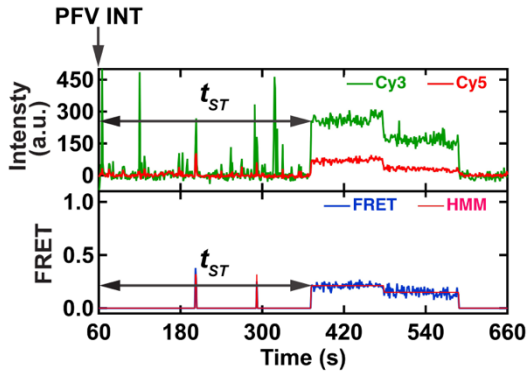
These numbers were calculated as described in **ONLINE METHODS** by fitting FRET distributions for each DNA with a single or a combination of two Gaussians and integrating the area under the curve for the 0.06 FRET peak. $\bar{x}$ and σ show the average and the standard deviation for the counts per DNA calculated from all the DNA substrates.

Extended Data Table 5. Percentages of STC events for different experiments.

Cy3-PFV			PFV
DNA Substrate	**$\%STC_{smFRET}$ ± $\Delta\%STC_{smFRET}$**	**Number of DNA molecules (N)**	**$\%STC_{Gel}$ ± $\Delta\%STC_{Gel}$**
GC	0.3 ± 0.2	585	0.4 ± 0.6
8-OxoG	0 ± 0	624	0.9 ± 1.5
G/T	0.5 ± 0.3	616	1.4 ± 0.8
+T	0 ± 0	531	0.8 ± 0.7
Nick (5'-P)	0.7 ± 0.3	587	3.2 ± 2.9
1nt Gap (5'-P)	3 ± 0.7	559	8.4 ± 3.8
2nt Gap (5'-P)	1.4 ± 0.5	506	10.7 ± 5.7
Nick (5'-OH)	2 ± 0.6	626	20 ± 12.4
1nt Gap (5'-OH)	20 ± 1.8	614	29.3 ± 10.9
2nt Gap (5'-OH)	12 ± 1.4	605	14.2 ± 1.7
Blocked end 1nt Gap (5'-OH)	24 ± 2.0	547	
R-Cy5 1nt Gap (5'-OH)	20 ± 1.8	527	
DNA Substrate	**$\%STC_{smFRET}$ ± $\Delta\%STC_{smFRET}$**	**Number of DNA molecules (N)**	
GC	0 ± 0	547	
1nt Gap (5'-P)	5 ± 0.9	579	
1nt Gap (5'-OH)	22 ± 2.0	569	
Cy3-PFV-ddA			
DNA Substrate	**$\%Long_{smFRET}$ ± $\Delta\%Long_{smFRET}$**	**Number of DNA molecules (N)**	
1nt Gap (5'-OH)	0.4 ± 0.2	564	
Cy3/Cy5-PFV			
DNA Substrate	**$\%STC_{smFRET}$ ± $\Delta\%STC_{smFRET}$**	**Number of DNA molecules (N)**	
GC	0.4 ± 0.3	548	
1nt Gap (5'-OH)	16 ± 1.6	609	

These percentages were calculated as described in **ONLINE METHODS**. For Cy3-PFV-ddA this number was defined with a different name to signify the catalytic deficiency. *N* indicates the total number of DNA molecules analyzed for each experiment. *n* is the number of molecules that showed strand transfer events. The errors ($\Delta\%STC_{smFRET}$) reflect $\frac{\sqrt{n}}{N} \times 100\%$.

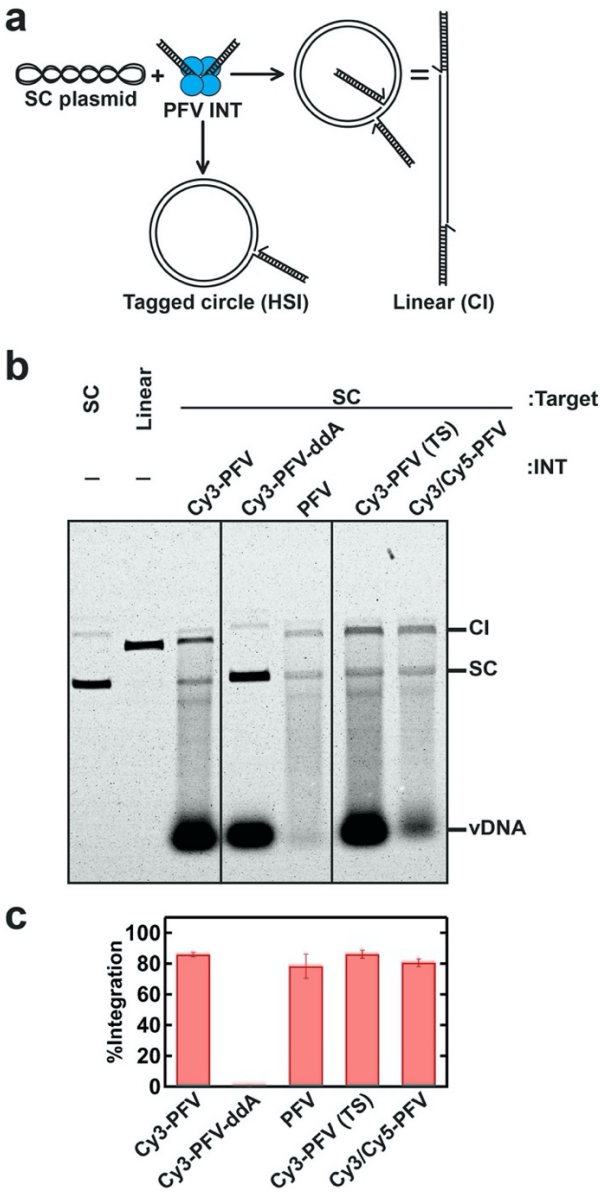
Extended Data Table 6. The strand transfer times for different DNA experiments.



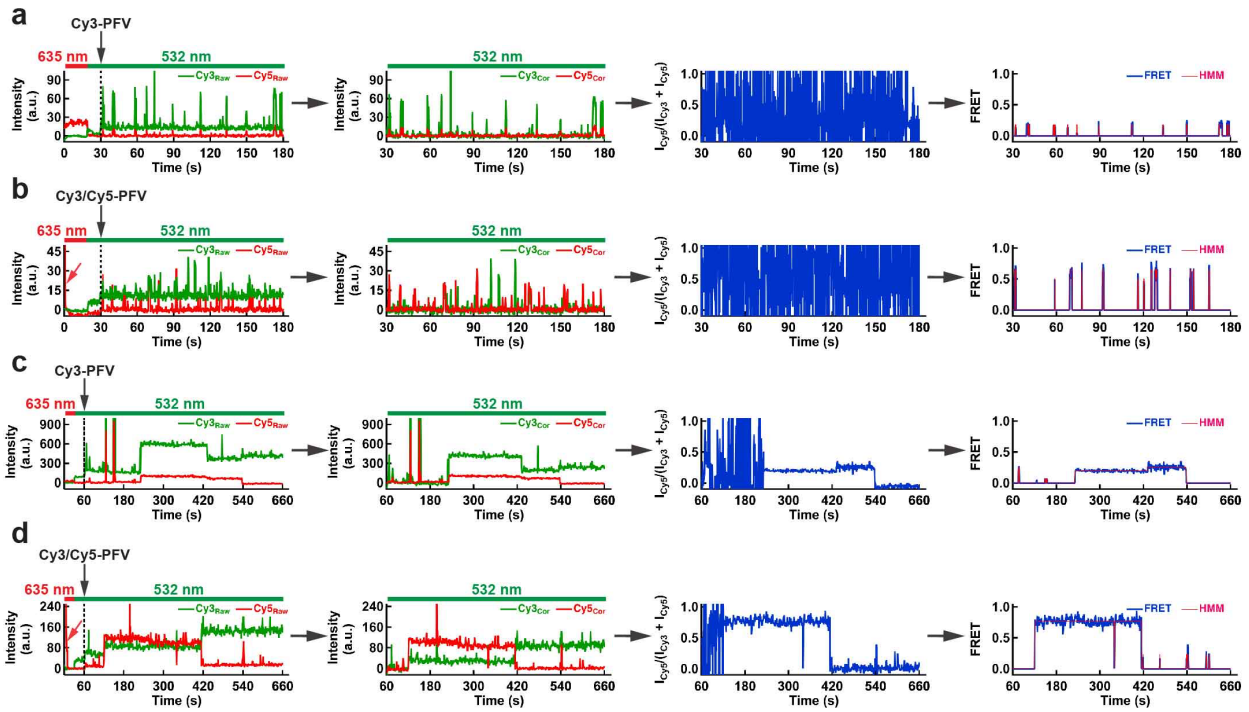
Cy3-PFV		
DNA Substrate	$\bar{t}_{ST} \pm \sigma_{ST}$ **(s)**	**Number of events (n)**
GC	172 ± 87	2
G/T	328 ± 228	3
Nick (5'-P)	191 ± 143	4
1nt Gap (5'-P)	254 ± 202	17
2nt Gap (5'-P)	264 ± 176	7
Nick (5'-OH)	180 ± 147	12
1nt Gap (5'-OH)	249 ± 150	125
2nt Gap (5'-OH)	277 ± 148	71
Blocked end 1nt Gap (5'-OH)	260 ± 145	134
R-Cy5 1nt Gap (5'-OH)	216 ± 152	104
Cy3-PFV (TS1)		
DNA Substrate	$\bar{t}_{ST} \pm \sigma_{ST}$ **(s)**	**Number of events (n)**
1nt Gap (5'-P)	229 ± 166	28
1nt Gap (5'-OH)	255 ± 159	124
Cy3-PFV-ddA		
DNA Substrate	$\bar{t}_{long} \pm \sigma_{long}$ **(s)**	**Number of events (n)**
1nt Gap (5'-OH)	186 ± 103	2
Cy3/Cy5-PFV		
DNA Substrate	$\bar{t}_{ST} \pm \sigma_{ST}$ **(s)**	**Number of events (n)**
GC	140 ± 38	2
1nt Gap (5'-OH)	251 ± 156	95

The strand transfer time (t_{ST}) is the time it takes to observe a stable FRET (or Cy3) signal from the intasome injection, as shown by the figure in the first raw. The mean values ($\bar{t}$) and standard deviations (σ) were calculated as described in **ONLINE METHODS**. For Cy3-PFV-ddA, this

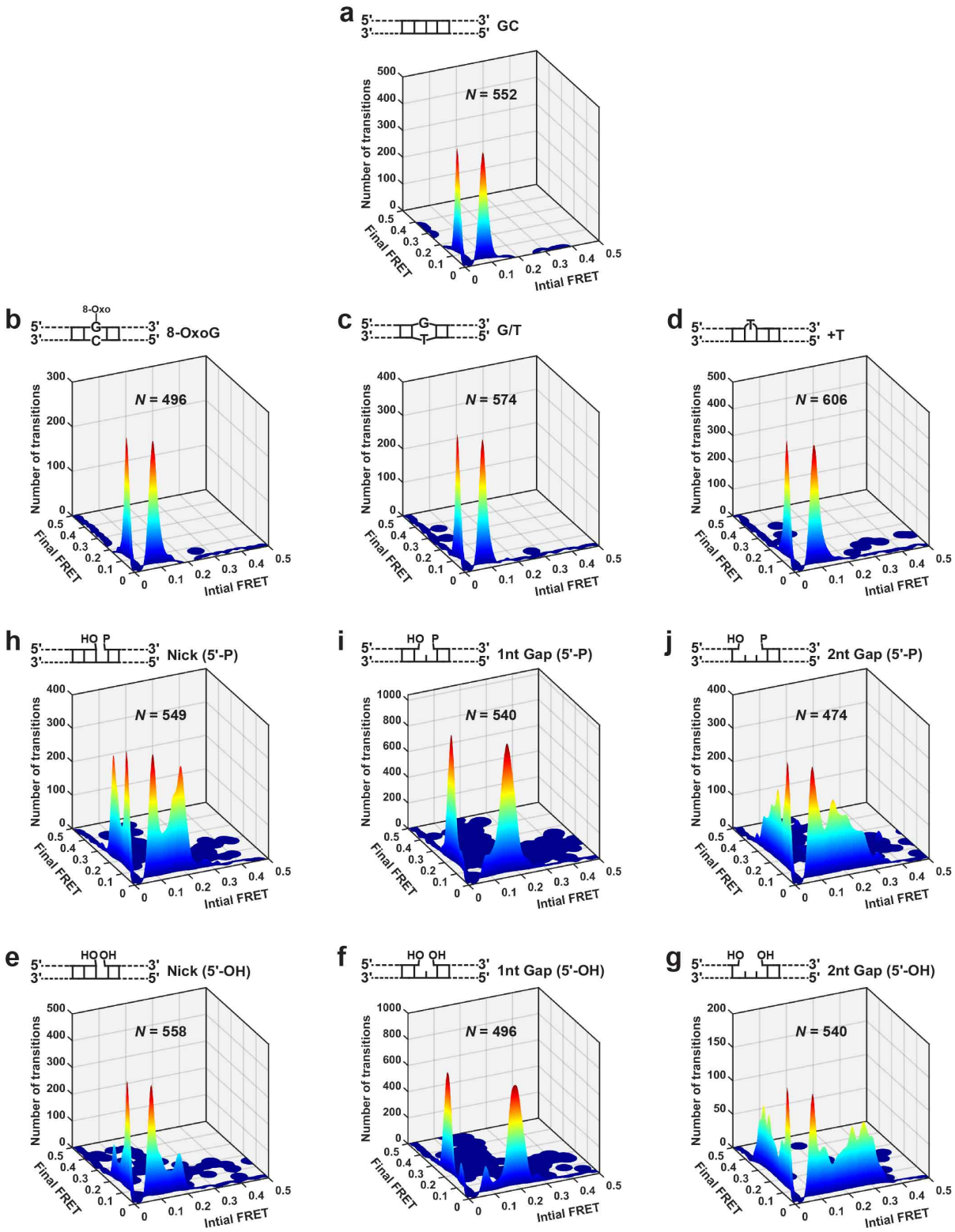
number was defined with a different name to signify the catalytic deficiency. *n* indicates the number of DNAs that showed strand transfer events.



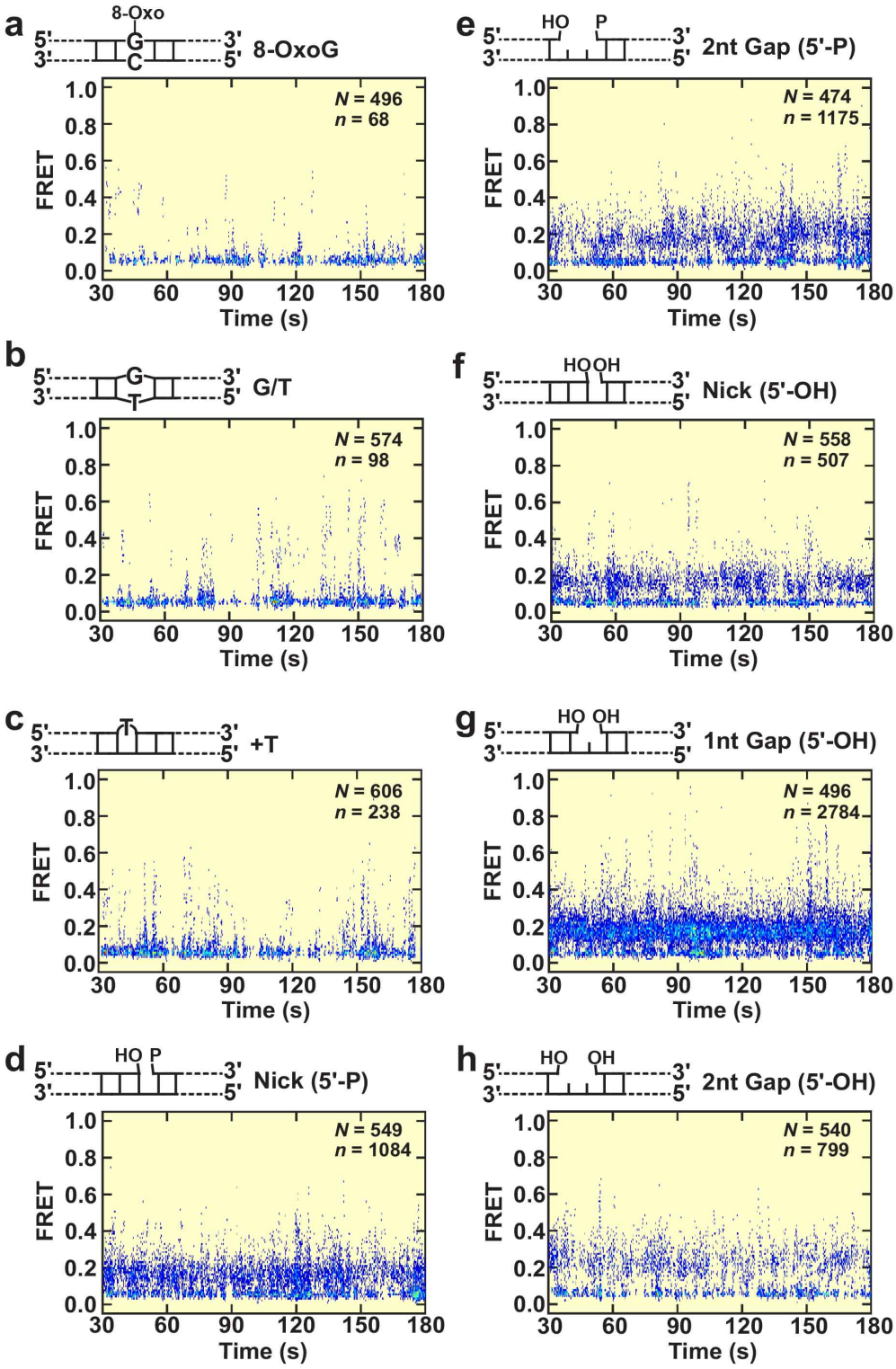
Extended Data Figure 1. Plasmid DNA-based integration activity assay of the intasomes. (**a**) A schematic of the assay. When a PFV intasome consisting of DNA oligonucleotides mimicking vDNA ends catalyzes concerted integration into a supercoiled (SC) DNA, linear DNA is produced. A relaxed tagged circle is created when only one strand transfer occurs in a half-site integration (HSI). (**b**) An ethidium bromide-stained agarose gel showing the integration activity of all the intasomes used in this study. Competing aggregation and auto integration reactions continuously deplete the active pool of intasomes available for the reaction and cause the incomplete conversion of the SC DNA. (**c**) Quantification of the gel data using the ethidium bromide fluorescence. Error bars indicate standard deviations from two independent experiments. HSI products that should migrate slower than the linear DNA were undetectable with ethidium bromide. The catalytically deficient Cy3-PFV-ddA did not produce any detectable activity.



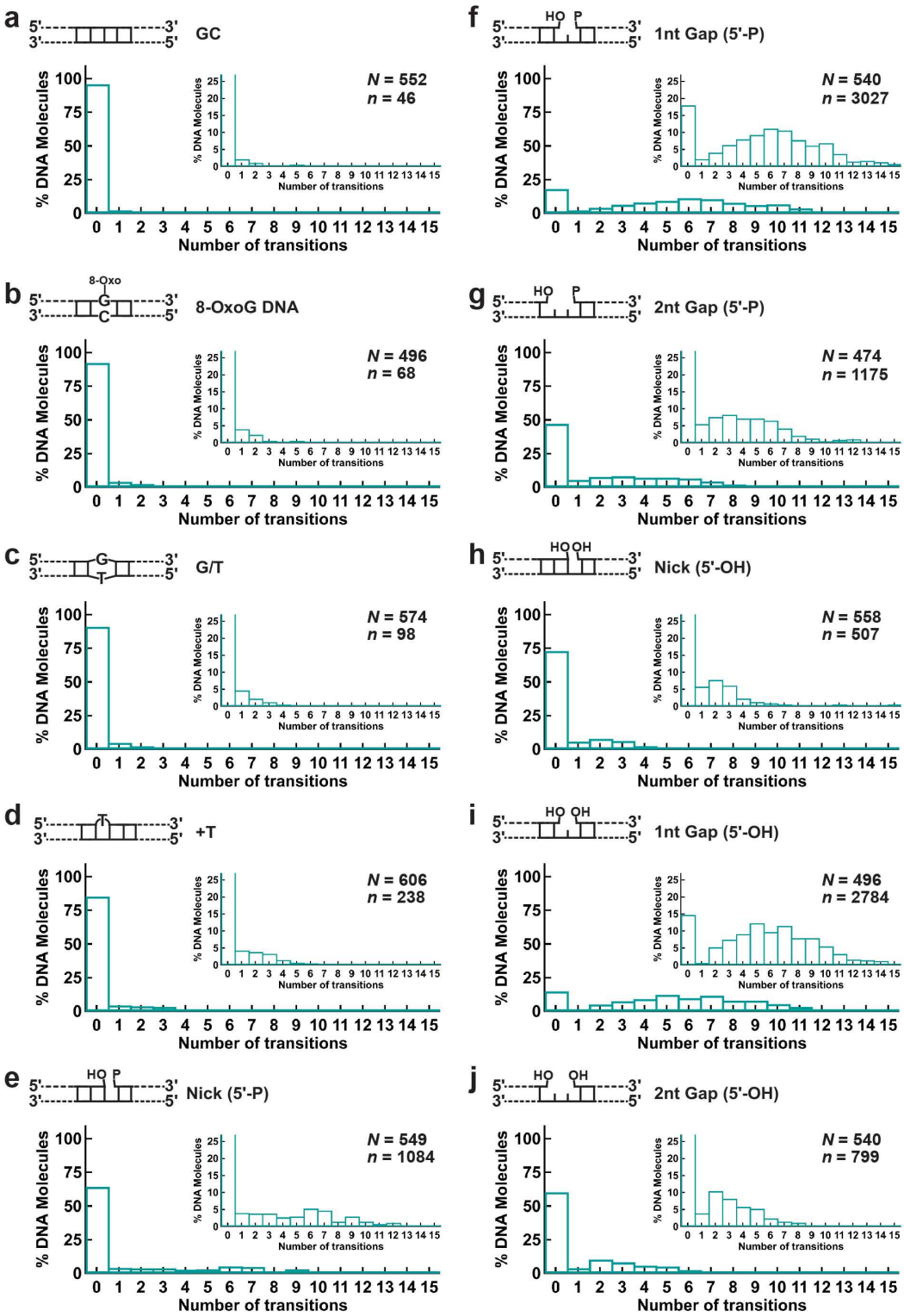
Extended Data Figure 2. A summary of the algorithm used to truncate, background correct, and perform HMM analysis of the smFRET trajectories. (**a**) A representative TCC trajectory for Cy3-PFV collected at 100 ms resolution. The raw intensity trajectory shows the initial red laser excitation, switching to the green laser, and real-time intasome injection (panel 1). The beginning part of the raw intensity trajectory was truncated at the intasome injection, and then Cy3, Cy5 backgrounds were corrected (panel 2). The ratiometric FRET for the truncated trajectory shows erratic fluctuations when both Cy3 and Cy5 signals approach zero (panel 3). Therefore, zero values were assigned for calculated FRET and the HMM fitting when Cy3 or Cy5 intensity reaches the background noise level (panel 4). (**b**) A representative TCC trajectory for Cy3/Cy5-PFV collected at 100 ms resolution. The experiment and the analysis were done similarly to (**a**), except the initial red laser exposure was used to photobleach Cy5 on the substrate (red arrow in panel 1). (**c**) A representative STC trajectory for Cy3-PFV collected at 1 s resolution. The experiment and the analysis were done similarly to (**a**). (**d**) A representative STC trajectory for Cy3/Cy5-PFV collected at 1 s resolution. The experiment and the analysis were done similarly to (**a**), except the initial red laser exposure was used to photobleach Cy5 on the substrate (red arrow in panel 1).



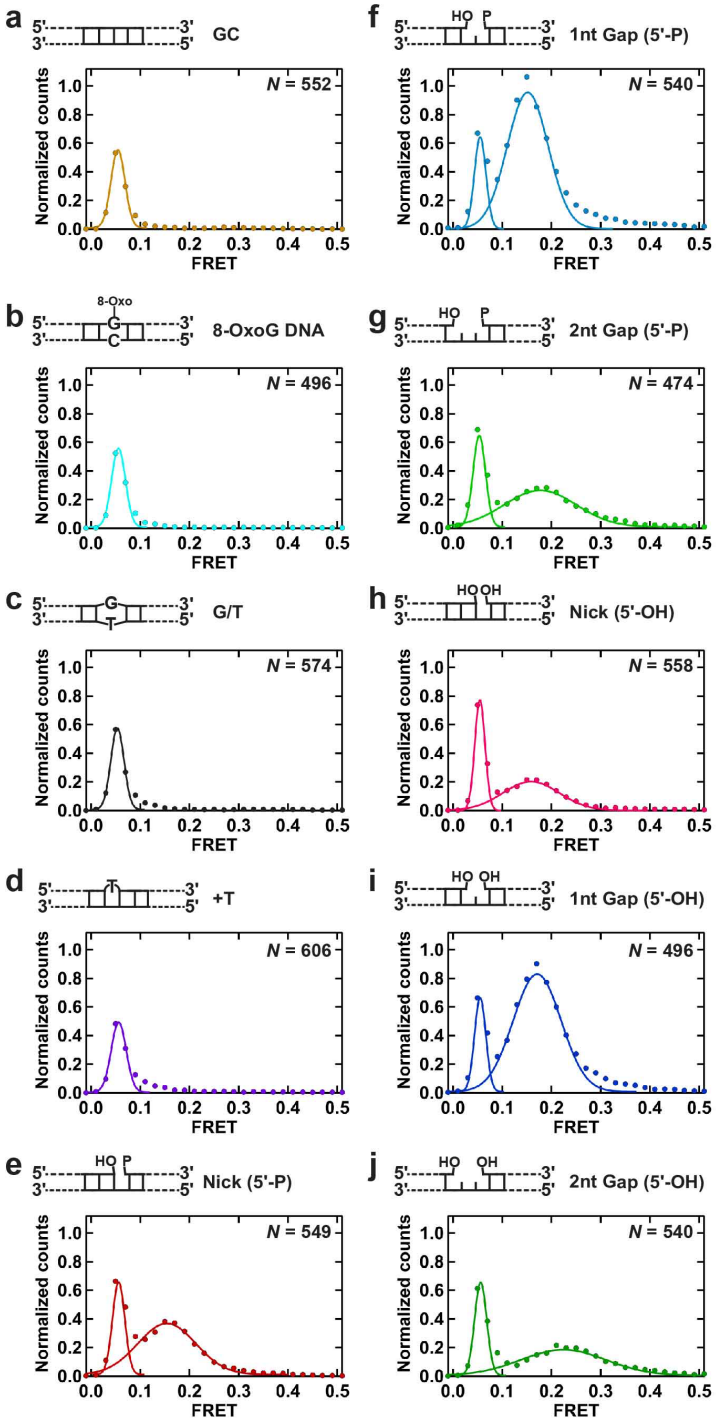
Extended Data Figure 3. Transition Density Plots (TDPs) for Cy3-PFV TCC formation with different DNA targets. (**a-g**) These TDPs show the number of transitions between a given initial and a final FRET state. The identities of DNAs and the number of DNA molecules (*N*) used to build each TDP are shown.



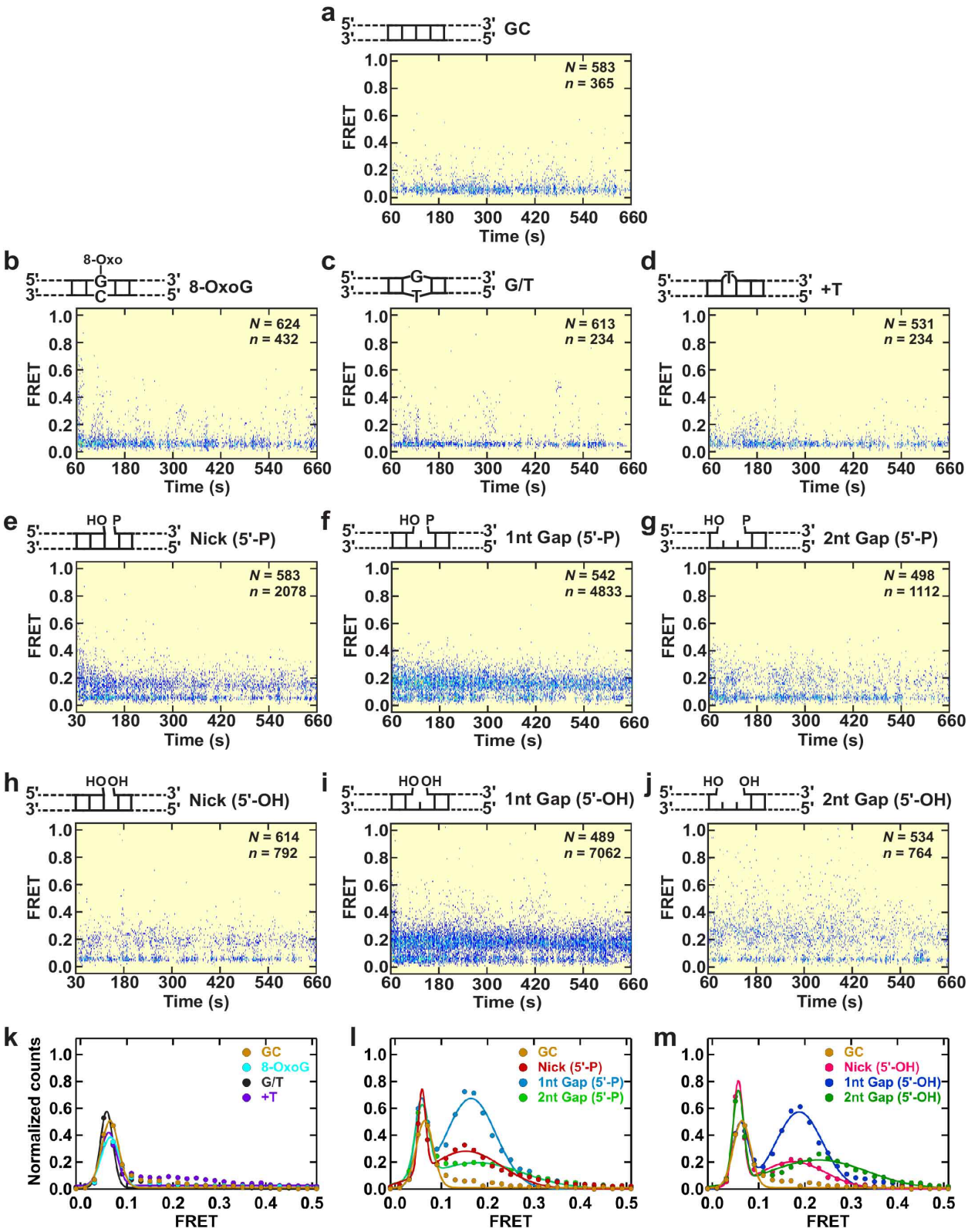
Extended Data Figure 4. Post-Sync Histograms (PSHs) of smFRET trajectories showing Cy3-PFV TCC formation on different DNA targets. (**a-h**) These PSHs were generated by aligning smFRET trajectories from numerous (*N*) DNA molecules. The identities of DNAs and the total number of transitions (*n*) that crossed >0.15 FRET threshold are also shown.



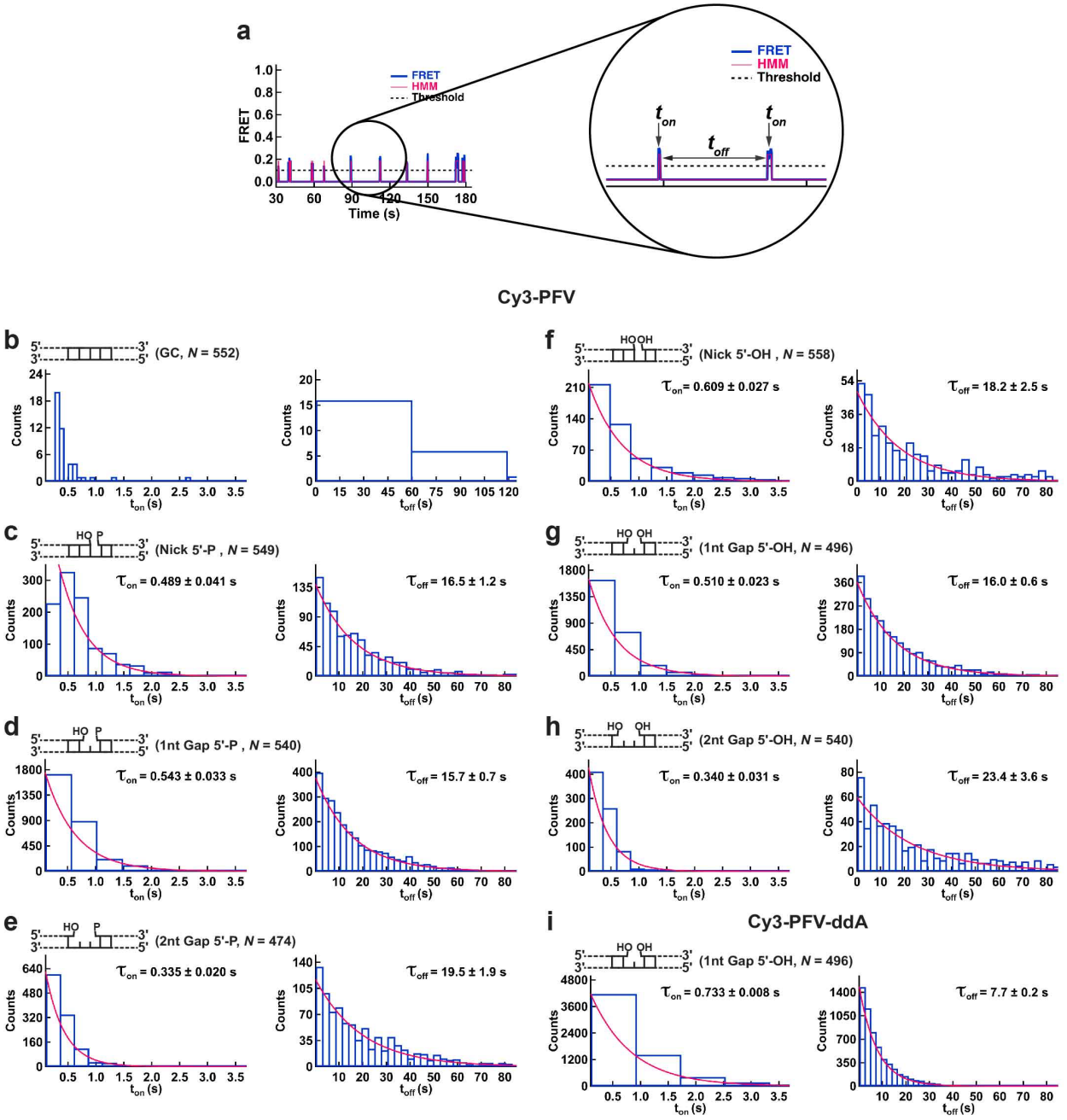
Extended Data Figure 5. Transition count histograms of TCC trajectories collected at 100 ms resolution. (**a-j**) These histograms depict the fraction of DNA molecules that showed a given number of transitions during the observation window of 2.5 min. The identities of DNAs and the total number of DNA molecules (*N*) used in the analysis are shown.



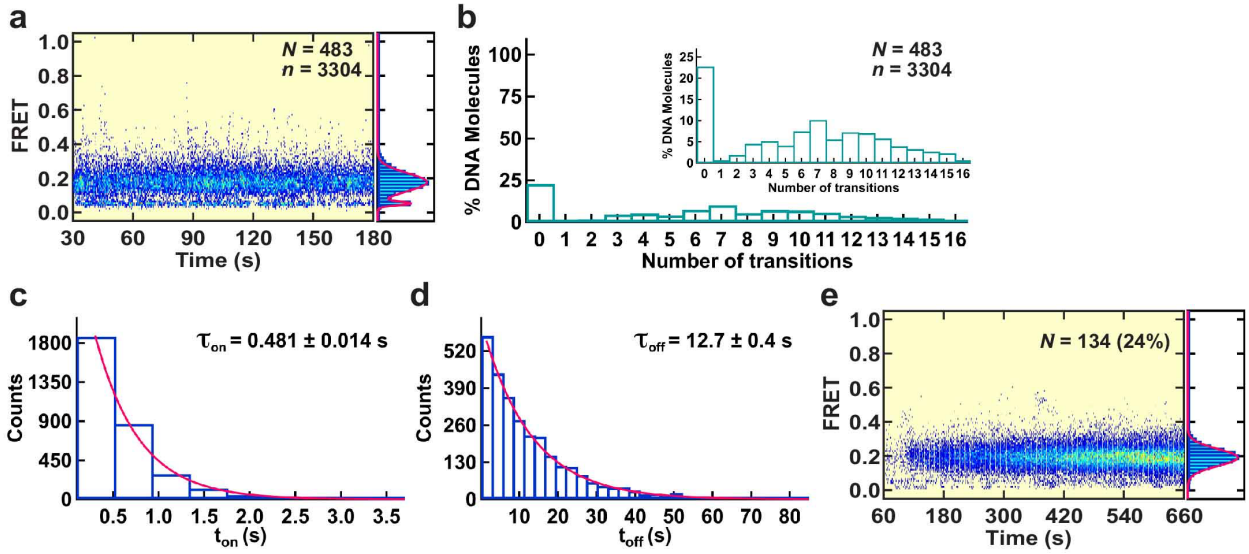
Extended Data Figure 6. Gaussian fittings of the TCC histograms collected at 100 ms resolution. (a-j) The histograms were normalized with respect to the area under the pseudo (0.06) FRET peak as described in ONLINE METHODS. The normalized histograms for GC, 8-OxoG, G/T, +T DNAs fit well with a single Gaussian. A combination of two Gaussians adequately models the histograms for the gap substrate, where the lower FRET peak corresponds to pseudo-FRET, and the higher FRET peak corresponds to intasome stalling on DNA. *N* is the number of DNA molecules included in each histogram.


a
GC
N = 583
n = 365
b
8-OxoG
N = 624
n = 432
c
G/T
N = 613
n = 234
d
+T
N = 531
n = 234
e
Nick (5'-P)
N = 583
n = 2078
f
1nt Gap (5'-P)
N = 542
n = 4833
g
2nt Gap (5'-P)
N = 498
n = 1112
h
Nick (5'-OH)
N = 614
n = 792
i
1nt Gap (5'-OH)
N = 489
n = 7062
j
2nt Gap (5'-OH)
N = 534
n = 764
FRET
Time (s)
k
l
m
Normalized counts
GC
8-OxoG
G/T
+T
Nick (5'-P)
1nt Gap (5'-P)
2nt Gap (5'-P)
Nick (5'-OH)
1nt Gap (5'-OH)
2nt Gap (5'-OH)

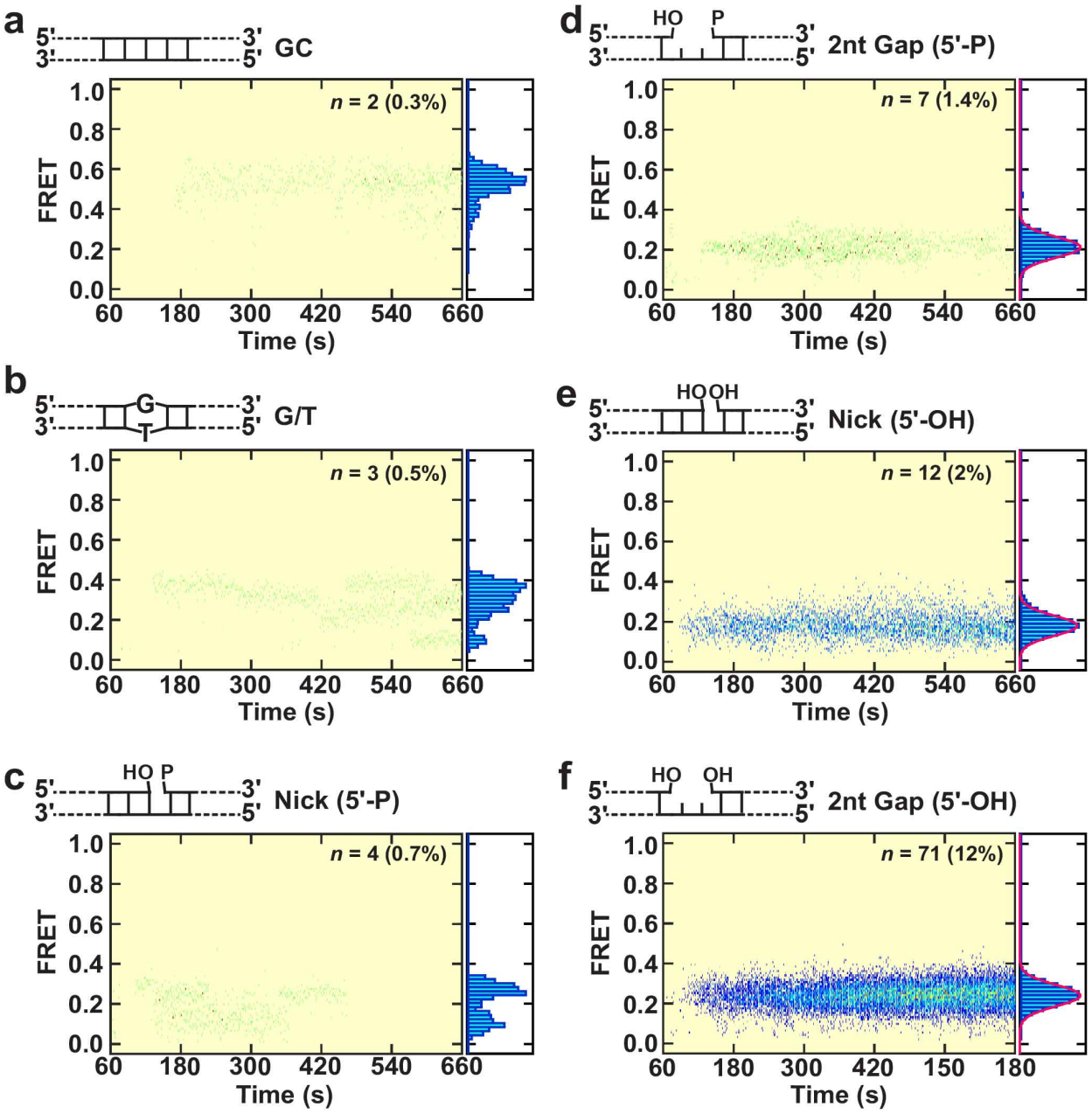
Extended Data Figure 7. Visualizing the PFV intasome target capture dynamics in real-time at 1 s resolution. (**a-j**) **PSHs** displaying intasome interactions with different DNA substrates as. *N* is the number of DNA molecules analyzed for each substrate. *n* is the total number of transitions that crossed >0.15 FRET threshold. (**k**) Normalized smFRET histograms and their Gaussian fits showing the distributions of E_{pseudo} for GC, 8-OxoG, G/T, +T DNAs. (**l,m**) Normalized smFRET histograms and their Gaussian fits showing the distributions of E_{pseudo} and E_{TCC} for (5'-P) (**l**) or (5'-OH) (**m**) gap substrates.



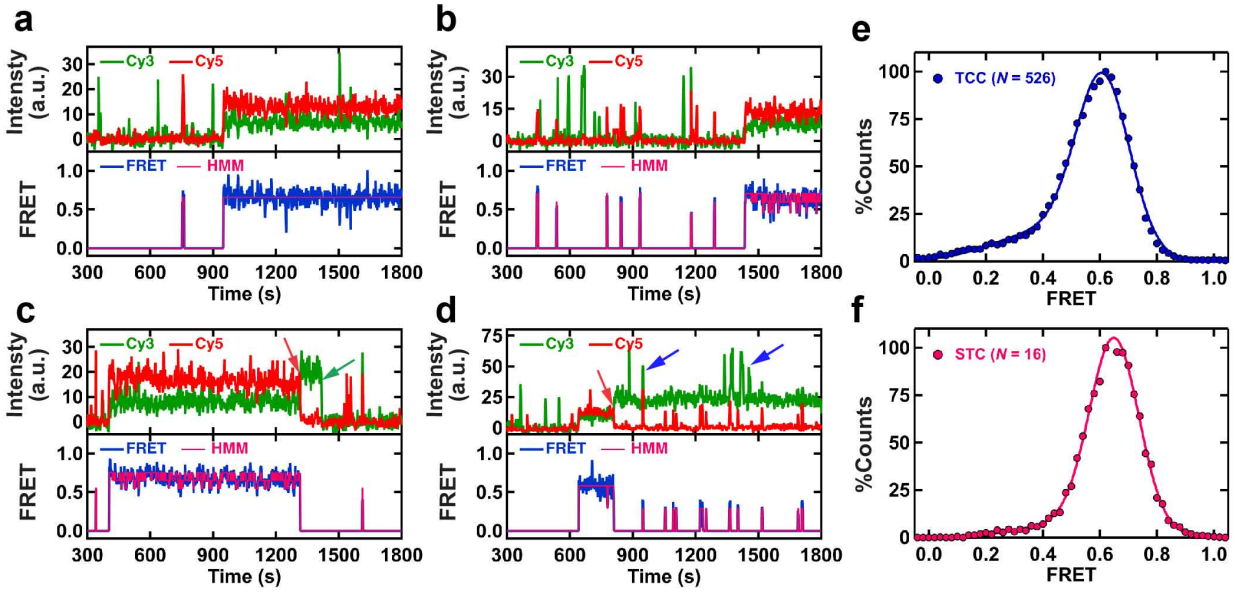
Extended Data Figure 8. Dwell time analysis for PFV intasome target capture at 100 ms resolution. (**a**) An exemplary FRET trajectory showing the transient target capture by Cy3-PFV. The magnified view illustrates the time spent in the bound (t_{on}) and the unbound (t_{off}) states. (**b-h**) Distributions of t_{on} and t_{off} for different DNA substrates. (**i**) Distributions of t_{on} and t_{off} for Cy3-PFV-ddA on 1nt Gap (5'-OH) DNA. The single exponential fits (red lines), the resulting average dwell times, and the errors from the fittings are shown for each distribution. The identities of the DNAs and the number of molecules (*N*) used to build each histogram are also shown.



Extended Data Figure 9. Interactions of Cy3-PFV with the blocked end DNA. (**a**) The PSHs and the smFRET histogram corresponding to a collective of target capture events for blocked end 1nt Gap (5'-OH) DNA. The Gaussian fits to the histograms is shown as a red line. Transition count histogram showing the fraction of DNA molecules that showed a given number of transitions during the observation window of 2.5. (**c,d**) Distributions of t_{on} and t_{off} for the target capture events. The single exponential fits (red lines), the resulting average dwell times, and the errors from the fittings are shown for each distribution. (**e**) The PSH and the smFRET histogram corresponding to *n* = 131 strand transfer events into blocked end 1nt Gap (5'-OH) DNA recorded at 1 s resolution. The percentage is the frequency of strand transfer.



Extended Data Figure 10. Post-Sync Histograms (PSHs) and smFRET histograms of Cy3-PFV strand transfer into different DNA targets. (**a-f**) These PSHs were generated by aligning smFRET trajectories that showed strand transfer events. The identities of DNAs, the total number of strand transfer events (*n*) included in each analysis, and the efficiencies of strand transfer as percentages are shown. The Gaussian fits to the smFRET histograms are shown as red lines.



Extended Data Figure 11. Probing the structural dynamics of PFV intasomes during target capture and strand transfer. (**a-d**) Several different intensity trajectories and the resultant FRET trajectories with the HMM fits showing strand transfer by Cy3/Cy5-PFV into 1nt Gap (5'-OH) DNA. For the noisy and longer STC FRET states, HMM tends to fit fluctuations that are not originating from anti-correlated Cy3, Cy5 intensities. The photobleaching of Cy3 and Cy5 are denoted by green and red color arrows respectively. The blue color arrows indicate random floating aggregates of the intasome that occasionally produce pseudo-FRET transitions. (**e**) smFRET histogram with the Gaussian fit for the target capture by Cy3/Cy5-PFV on 1nt Gap (5'-OH) DNA. (**f**) smFRET histogram with the Gaussian fit for strand transfer by Cy3/Cy5-PFV into 1nt Gap (5'-OH) DNA. These data were collected at 100 ms resolution using a smFRET experimental setup similar to **Fig. 4a**. The number of DNA molecules (*N*) included in each histogram is indicated.