Harms, Klaus; Lunnan, Asbjørn; Hülter, Nils; Mourier, Tobias; Vinner, Lasse; Andam, Cheryl P.; Marttinen, Pekka; Fridholm, Helena; Hansen, Anders Johannes; Hanage, William P.; Nielsen, Kaare Magne; Willerslev, Eske; Johnsen, Pal Jarle

Substitutions of short heterologous DNA segments of intragenomic or extragenomic origins produce clustered genomic polymorphisms

Published in:
Proceedings of the National Academy of Sciences of the United States of America

DOI:
10.1073/pnas.1615819114

Published: 27/12/2016

Document Version
Publisher's PDF, also known as Version of record

Substitutions of short heterologous DNA segments of intragenomic or extragenomic origins produce clustered genomic polymorphisms

Klaus Harms[a,b,1], Asbjørn Lunnan[a], Nils Hülter[c], Tobias Mourier[b], Lasse Vinner[b], Cheryl P. Andam[d], Pekka Marttinen[e], Helena Fridholm[b,f], Anders Johannes Hansen[b], William P. Hanage[d], Kaare Magne Nielsen[g,h], Eske Willerslev[b,1], and Pål Jarle Johnsen[a,1]

[a]Department of Pharmacy, Faculty of Health Sciences, The Arctic University of Norway, 9037 Tromsø, Norway; [b]Centre for GeoGenetics, Natural History Museum of Denmark, University of Copenhagen, 1350 Copenhagen K, Denmark; [c]Genomic Microbiology, Institute of Microbiology, Christian-Albrechts-Universität zu Kiel, 24118 Kiel, Germany; [d]Department of Epidemiology, Harvard T. H. Chan School of Public Health, Boston, MA 02115; [e]Helsinki Institute for Information Technology, Department of Computer Science, Aalto University, FIN-00076 Aalto, Finland; [f]Department of Microbiological Diagnostics and Virology, Statens Serum Institut, 2300 Copenhagen S, Denmark; [g]Department of Life Sciences and Health, Oslo and Akershus University College of Applied Sciences, 0130 Oslo, Norway; and [h]GenØk-Center for Biosafety, 9294 Tromsø, Norway

Edited by John R. Roth, University of California, Davis, CA, and approved November 22, 2016 (received for review September 23, 2016)

In a screen for unexplained mutation events we identified a previously unrecognized mechanism generating clustered DNA polymorphisms such as microindels and cumulative SNPs. The mechanism, short-patch double illegitimate recombination (SPDIR), facilitates short single-stranded DNA molecules to invade and replace genomic DNA through two joint illegitimate recombination events. SPDIR is controlled by key components of the cellular genome maintenance machinery in the gram-negative bacterium *Acinetobacter baylyi*. The source DNA is primarily intragenomic but can also be acquired through horizontal gene transfer. The DNA replacements are nonreciprocal and locus independent. Bioinformatic approaches reveal occurrence of SPDIR events in the gram-positive human pathogen *Streptococcus pneumoniae* and in the human genome.

illegitimate recombination | mutation | microindels

Short patches of clustered nucleotide variations are routinely observed in whole genome comparisons (1, 2). These sequence variations are substrates for natural selection, which shapes prokaryotic (3, 4) and eukaryotic (5, 6) genomes. Clustered nucleotide variations also play a role in oncogenesis where they add to the overall genomic instability (7, 8). Despite their significant biological role, the molecular mechanisms underlying formation of clustered nucleotide variations are not fully understood.

Known mechanisms responsible for clustered nucleotide variations include error-prone DNA polymerases (9) and conversions at imperfect palindromes through template-switching (10) (templated mutagenesis), which can generate tracts of single nucleotide changes, respectively. Down-regulation or loss of genes involved in mismatch repair can also lead to increased genome-wide point mutation frequencies that can result in random single-nucleotide variation (SNV) clusters. Moreover, cumulative SNVs have been described when genes for DNA-modifying enzymes were up-regulated (11). All these mechanisms typically result in tracts of single-nucleotide polymorphisms (SNPs).

More complex clustered genomic polymorphisms may also develop through point mutations accumulating in a small DNA tract over a short time or through independent insertion and deletion events (12). A number of RecA-independent mechanisms have been described and investigated in detail that lead to microdeletions without insertions, or to microinsertions without deletions, in both prokaryotic and eukaryotic organisms. Among these mechanisms are replication slippage (13) or copy number variations in microsatellite DNA (14), illegitimate recombination at microhomologies (15, 16), imprecise nonhomologous end joining (NHEJ) (17), DNA gyrase-mediated strand switching (18), and transposon scars. Two or more temporally independent deletion/insertion events at the same locus can result in clustered polymorphisms, although in retrospective studies, such sequential events are nearly impossible to verify.

The most diverse clusters of nucleotide variations are formed by microhomology-mediated end-joining (MMEJ). MMEJ has been observed in eukaryotes only and can repair DNA double-strand (ds) breaks in an error-prone way. During repair, MMEJ often generates short, direct, or inverted repeats (19) and occasionally incorporates ectopic DNA at the recombinant joints (20). MMEJ results in highly variable clustered polymorphisms at the recombinant joint and is now recognized as a driving force in rapidly evolving oncogenic cells (21). DNA polymerase theta (POLQ) has recently been identified as the key enzyme in MMEJ-directed error-prone repair, but many mechanistic details of its function remain elusive (22). To date, no POLQ-like genes have been identified in prokaryotes.

Due to the immense evolutionary and biomedical implications of how and why genetic diversity is generated in prokaryotic and eukaryotic organisms, the underlying mechanisms are intensively investigated. To study and quantify the formation of clustered polymorphisms, we developed a detection assay in the bacterium

Significance

Clustered genomic polymorphisms in DNA, such as microindels and stretches of nucleotide changes, play an important role in genome evolution. Here, we report a mutation mechanism responsible for such genomic polymorphisms where short, single-stranded DNA molecules invade double-stranded DNA and replace short genomic segments. We show, in a bacterial model organism, that the genomic replacements occur with very low levels of sequence identity (microhomologies). The invading DNA can be of intagenomic or foreign origin. Genotoxic stress, horizontally taken-up DNA, or lack of genome maintenance functions increase the mutation frequency up to 7,000-fold. Bioinformatic approaches suggest that this class of mutations is widespread in prokaryotes and eukaryotes and may have a role in tumorigenesis.

Author contributions: K.H., K.M.N., E.W., and P.J.J. designed research; A.J.H. supervised pipeline building; K.H., A.L., N.H., T.M., L.V., C.P.A., P.M., and H.F. performed research; K.H., T.M., C.P.A., P.M., and P.J.J. analyzed data; and K.H., W.P.H., K.M.N., E.W., and P.J.J. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence may be addressed. Email: klaus.harms@spdir.net, ewillerslev@snm.ku.dk, or paal.johnsen@uit.no.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1615819114/-/DCSupplemental.

Acinetobacter baylyi. We demonstrate how regions of clustered, highly variable DNA sequence variations (ranging from 3 to 77 bp) can be formed by two coupled, microhomology-dependent illegitimate recombination (IR) events with free DNA single strands of intragenomic or external origin.

Results

Joined Double Illegitimate Recombinations Generate Clustered Polymorphisms. To quantify and characterize clustered small indels and polymorphisms, we developed an in vivo detection construct (*hisC*::′ND5i′) (23) in the soil bacterium *Acinetobacter baylyi* ADP1. The construct is permissive for small IR events but largely refractory to single-nucleotide mutations. In this construct, two neighboring stop codons in a functionless 228-bp insert prevent expression of a histidine prototrophy marker gene (histidinol-phosphate aminotransferase; Fig. 1*A*). We found that spontaneous histidine-prototrophic (His$^+$) mutants arose at low frequencies. Subsequent DNA sequencing analyses of individual His$^+$ isolates revealed that the ′ND5i′ segment was frequently substituted with different heterologous segments of intragenomic origins. The substituting DNA segments were of similar or shorter length, eliminating or bypassing the stop codons (Fig. 1 *B*–*E* and Dataset S1), and their neighboring upstream and downstream nucleotide stretches were identical with DNA segments in otherwise fully heterologous DNA regions elsewhere in the genome (Fig. S1). Sequence analyses of these donor DNA fragments and the parental DNA sequences strongly suggested that integration occurred through hybridization at microhomologies (short identical DNA stretches) or at extended microhomologies (clusters of microhomologies interrupted by mismatches and gaps in heterologous DNA; Fig. 1 *B*–*E* and *Supporting Information*) followed by illegitimate recombinations. The recombinations occurred either at a single, contiguous microhomology (class 1 events; Fig. 1 *B* and *C*) or at two separate microhomologies on the same molecule (class 2 events; Fig. 1 *D* and *E*). The recombinations were nonreciprocal (*Supporting Information*) and independent of genomic locus and detection construct (*Supporting Information*). Together, these short-patch double illegitimate recombination (SPDIR) events led to highly variable polymorphisms at a single genetic locus and introduced multiple clustered nucleotide exchanges, DNA sequence replacements of variable length, or deletions accompanied by nucleotide changes at the deletion site (Dataset S1), resulting in highly diverse codon changes (Fig. S2). In all characterized SPDIR events, the source DNA of the acquired nucleotide polymorphisms was identified both for intragenomic and extragenomic (see below) origins (Dataset S1). Net nucleotide gains (maximum six base pairs) were observed in only a few cases. Although the SPDIR mechanism depends on microhomologies, the randomness of the genetic changes observed suggests a broad mutagenic potential.

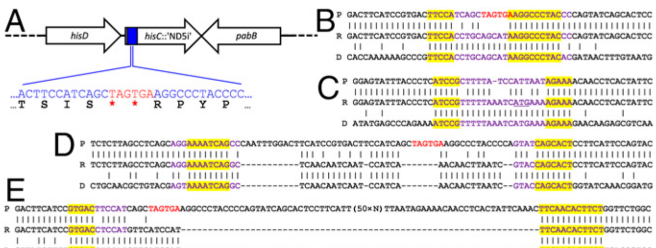

Fig. 1. (*A*) Schematic illustration of the *hisC*::′ND5i′ detection construct for SPDIR. The genomic location and the sequence detail of the two stop codons are indicated (modified after ref. 23). The ′ND5i′ insert is shown in blue, and the translated codons are shown in black, with the two consecutive stop codons indicated in red. (*B*–*E*) Examples of clustered polymorphisms generated by SPDIR, shown as triple DNA alignments of the parental (His$^-$; P), His$^+$ recombinant (R), and donor (D) strands used for the double IR. Stop codons are indicated in red, and recombination sites are highlighted in yellow. Microhomologies (as approximated by ΔG^0_{min}) are in purple typeface. (*B* and *C*) Class 1 SPDIR events formed by two illegitimate joints at a single, contiguous extended microhomology. (*D* and *E*) Class 2 SPDIR events with illegitimate joints at separate (simple or extended) microhomologies, leading to complex replacements or deletions. The donor DNA originated from intragenomic loci [(*B*) Recurrent SPDIR mutation A26 (Dataset S1), putative ACIAD1938 gene; (*C*) SPDIR O106, putative ACIAD1581 gene; and (*D*) SPDIR R159, putative ACIAD2154 gene] except *E*. In *E*, the donor DNA was derived from *Bacillus subtilis* DNA (*ipk* gene) and acquired by *A. baylyi* through natural transformation. The complete set of experimentally found SPDIR sequences is listed in Dataset S1.

Low Frequency of SPDIR Mutations in Wild-Type Cells. We quantified occurrence of SPDIR experimentally in wild-type (WT) *A. baylyi* cells and found that His$^+$ revertants were scarce (1.1×10^{-11}; about 14-fold rarer than single point mutations; Table 1). The fraction of SPDIR mutation events among the His$^+$ reversions was ∼5%, corresponding to a calculated SPDIR frequency of 5.6×10^{-13} (Table 1). This number is likely an underestimation due to limitations in the detection construct because SPDIR-generated substitutions that introduce stop codons or frameshifts or lead to improper protein folding remain undetected.

The non-SPDIR His$^+$ mutations were in most cases (>90% in WT) conferred by in frame deletions in ′ND5i′ [i.e., single illegitimate recombination (IR) events], both with and without microhomologies, and occasionally by different classes of mutations (*Supporting Information*). The fact that SPDIR occurred in the WT close to the detection limit in our specific experimental setup can explain lack of prior experimental discovery.

Single-Strand–Specific DNA Exonucleases Control SPDIR in Wild-Type Cells. Microhomology-mediated IR events have been observed in prokaryotes and eukaryotes (15, 16) and are initiated by annealing of DNA single-strand ends. We hypothesized that SPDIR was initiated by hybridization of genomic dsDNA at exposed single-stranded (ss) gaps, loops, or replication forks, with ssDNA segments. In prokaryotes, free cytoplasmic DNA single strands are attacked by ss-specific DNA exonucleases (24) (ssExo), and in *A. baylyi*, these ssExo have been revealed as RecJ and ExoX (23). We therefore quantified SPDIR in ssExo-deficient mutants and found that the SPDIR frequency was elevated approximately sevenfold in Δ*recJ* and fourfold in Δ*exoX* mutants (Table 1). The frequency was increased 28-fold in a Δ*recJ* Δ*exoX* double mutant, which lacked all ssExo activity. In the Δ*recJ* Δ*exoX* strain, SPDIR events produced about 34% of all His$^+$ mutation events, whereas in WT and in the single mutants the proportion of SPDIR events was at least sixfold lower than in the Δ*recJ* Δ*exoX* mutant (Table 1). These results confirmed that SPDIR is suppressed by ssExo in WT cells and indicate that SPDIR events depend on the presence of ssDNA in the cytoplasm.

SPDIR Is Inhibited by RecA Protein. Cytoplasmic ssDNA is a cellular genome damage signal and can be bound by RecA protein to initiate recombinational repair and to trigger the SOS response (25). We deleted the *recA* gene of *A. baylyi*, and in the Δ*recA* mutant we observed an about sixfold SPDIR frequency increase. Remarkably, in a Δ*recA* Δ*recJ* Δ*exoX* triple mutant, the SPDIR frequency was >7,700-fold higher than that of the WT, and SPDIR was the most common His$^+$ mutation (80%; Table 1). The strong synergy effect suggests that SPDIR is controlled by factors beyond elimination of free cytoplasmic DNA. It is conceivable that binding of RecA protein to ssDNA efficiently prevents hybridization of ssDNA molecules, and molecules that escape RecA-binding frequently anneal at microhomologies. In WT cells, these microhomology-annealed molecules are attacked

GENETICS

Table 1. His$^+$ and SPDIR frequencies in *A. baylyi* strains without and with genotoxic stress or addition of DNA

A. baylyi ADP1 *hisC*::′ND5i′ relevant genotype	Amendment			Median His$^+$ frequency	SPDIR fraction‡	Calculated SPDIR frequency		*n*
	CIP* (MIC)	UV, $mJ_{260\ nm}$	DNA†			Absolute	Relative	
Wild type	—	—	—	1.1×10^{-11}	5% (2/40)	5.6×10^{-13}	=1	10
Δ*exoX*	—	—	—	3.1×10^{-11}	8% (2/25)	2.4×10^{-12}	4.3	17
Δ*recJ*	—	—	—	1.1×10^{-10}	4% (1/25)	4.3×10^{-12}	7.7	9
Δ*recJ* Δ*exoX*	—	—	—	4.6×10^{-11}	34% (19/56)	1.6×10^{-11}	28	11
Δ*recA*	—	—	—	4.4×10^{-11}	8% (2/25)	3.5×10^{-12}	6.2	15
Δ*recA* Δ*recJ* Δ*exoX*	—	—	—	5.4×10^{-9}	80% (32/40)	4.3×10^{-9}	7,722	14
Wild type	0.1	—	—	1.5×10^{-9}	2% (1/50)	3.0×10^{-11}	53	11
	0.25	—	—	7.1×10^{-9}	5% (2/40)	3.6×10^{-10}	631	10
Wild type	—	3.6	—	2.8×10^{-10}	4% (2/46)	1.2×10^{-11}	21	13
	—	10.8	—	8.2×10^{-9}	0% (0/67)	$<1.2\times10^{-10}$	<216	12
Δ*recJ* Δ*exoX*	—	3.6	—	1.3×10^{-9}	25% (2/8)	3.3×10^{-10}	594	5
	—	10.8	—	8.5×10^{-9}	25% (3/12)	2.1×10^{-9}	3,782	5
Wild type	—	—	BS	6.5×10^{-11}	4% (1/25)§	2.6×10^{-12}	4.6	10
Δ*recJ* Δ*exoX*	—	—	AB	1.4×10^{-9}	47% (8/17)	6.6×10^{-10}	1,173	10
	—	—	SS	7.4×10^{-10}	33% (7/21)§	2.5×10^{-10}	439	5
	—	—	BS	5.5×10^{-10}	51% (24/47)¶	2.8×10^{-10}	500	9
Δ*recJ* Δ*exoX* Δ*comA*	—	—	—	6.5×10^{-11}	35% (8/23)	2.3×10^{-11}	40	10
hisC$^+$ *trpE27*	—	—	—	1.5×10^{-10}#	n.a.	n.a.	n.a.	11

n.a., not applicable.

*CIP, ciprofloxacin supplemented at concentrations relative to the minimal inhibitory concentration (MIC) for *A. baylyi* wild type (62.5 $ng\cdot mL^{-1}$; modified Etest).

†Supplemented with 300 $ng\cdot mL^{-1}$ genomic DNA from the following sources: BS, *Bacillus subtilis* 168; AB, *A. baylyi hisC*::′ND5i′; SS, salmon sperm DNA.

‡Identical genotypes were regarded as siblings originating from a single mutation event.

§The SPDIR events formed with endogenous AB DNA.

¶Eight SPDIR events were formed with BS, and 15 events were formed with AB DNA. One donor DNA segment was present in both donor genomes.

#Point mutation frequency, given as median Trp$^+$ frequency.

by ssExo and prevented from genomic integration, as observed in *Escherichia coli* (24) and *A. baylyi* (23). Alternatively, faithful recombinational DNA damage repair mediated by RecA together with ssExo prevents production of ssDNA remnants (26) (e.g., displaced strand fragments or flaps) that could act as donor molecules for SPDIR. These explanations are not mutually exclusive.

Exposure to Genotoxic Stress Increases SPDIR Frequencies. IR frequencies are increased with accumulating genomic DNA damages, and the increase has been attributed to microhomology-mediated DNA end-joining events leading to deletions and other genomic rearrangements (27). We determined whether introduction of DNA strand breaks affected SPDIR frequency in *A. baylyi.* For this purpose, we treated growing cultures with subinhibitory concentrations of ciprofloxacin (a fluoroquinolone antibiotic interfering with DNA gyrase activity) (28), or with variable doses of UV (UV) light. Both agents result in replication blocks and lead to genome fragmentation (29, 30). We found that the His$^+$ frequencies were increased up to at least 600-fold with increasing doses of ciprofloxacin or UV until viability was affected, and SPDIR events were detected at low proportions (2–5%) except after UV irradiation with 10.8 mJ (Table 1).

When we repeated the UV experiments with the Δ*recJ* Δ*exoX* mutant, SPDIR events accounted for ∼25% of His$^+$ events with both UV doses tested (Table 1). This ratio was lower than in untreated cells (34%), indicating that SPDIR is increased by two to three orders of magnitude with increasing DNA damage levels, which is in agreement with previous reports on IR (27). However, the increase of SPDIR events is lower than that of IR-mediated mutations such as deletions.

Natural Transformation Increases Frequency and Variability of SPDIR Events. To explore the effect of exogenous DNA on SPDIR formation, we exploited the constitutive competence for natural transformation of WT *A. baylyi* cells (23). DNA molecules are taken up by the cells into the cytoplasm as single strands (31). We found that exposure to foreign DNA isolated from *Bacillus subtilis* resulted in a fourfold to fivefold elevated SPDIR frequency (Table 1). We repeated the experiments with the Δ*recJ* Δ*exoX* mutant, using *B. subtilis* DNA, isogenic *A. baylyi* His$^-$ DNA, and DNA isolated from salmon sperm as donor DNA substrates. In the Δ*recJ* Δ*exoX* strain, addition of the DNA substrates led to SPDIR frequencies about 15- to 40-fold higher than without added DNA (Table 1). Notably, when exposed to foreign DNA, about two thirds of the SPDIR mutations were formed with cognate DNA, and approximately one third were formed with taken-up DNA. This result is consistent with findings of previous reports showing that recombination attempts during natural transformation frequently result in DNA strand breaks and thus can damage genomic DNA (32). The DNA damages then lead to increased SPDIR frequencies, as observed in the experiments with ciprofloxacin and UV light. The RecA-independent recombination at the MH was strand orientation-specific (*Supporting Information*). In a transformation-deficient Δ*comA* Δ*recJ* Δ*exoX* triple mutant [lacking the ComA DNA uptake pore (23)], the SPDIR frequency was not different from that of the Δ*recJ* Δ*exoX* mutant (Table 1).

These results confirm that SPDIR is primarily an intragenomic process and also demonstrate that natural transformation can be mutagenic through the SPDIR pathway. Consequently, clustered polymorphisms in the genome of some bacterial species can be the result of foreign DNA acquisition. However, in retrospective genome analyses it may often not be possible to identify the origin of donor DNA molecules due to the short length of the SPDIR-generated polymorphisms.

The Two IR Events of SPDIR Are Temporally Linked. Three lines of evidence strongly suggest that SPDIR mutations form within a single generation before selection. First, we frequently found intragenomic donor DNA segments in SPDIR isolates with reverse complement orientation relative to the *hisC*::′ND5i′ allele (Table S1). In these cases, temporally independent IR events would result in lariat chromosome intermediates that cannot be replicated by the cell. Second, SPDIR events with foreign DNA (that is taken up by the cell as ssDNA fragments) require genomic integration through two IR events in a single generation to prevent potentially lethal dsDNA breaks. Third, many His^+ colonies from the same primary cultures frequently carried unique, identical SPDIR mutations, both with intragenomic or exogenous donor DNA molecules. Such jackpot events strongly suggest that SPDIR mutations preexisted in the bulk culture (see *Supporting Information* and Table S1 for details). In our frequency calculations, identical mutations from the same assay were treated as single mutation events.

A Model for SPDIR Caused by Cytoplasmic ssDNA Molecules. We show that SPDIR events depend on the presence of ssDNA and are suppressed by key components of the genome maintenance machinery. A genomic integration model is depicted in Fig. 2. In that model, microhomologies are used by the cell to join unrelated DNA molecule ends, as has been demonstrated and quantified in previous studies for single (15) or multiple (24, 33) IR events. The model further builds on a proposed mechanism for strand orientation-specific, RecA-independent integration of short DNA molecules (23), in which we showed that fully homologous oligodeoxynucleotides (≥20 bp) could be chromosomally integrated in a single event during replication, acting as primers for Okazaki fragments (23, 24) (*Supporting Information*). In the present study, we demonstrate that microhomologies are sufficient for chromosomal integration at low but detectable frequencies during lagging strand DNA synthesis (Fig. 2, Fig. S3, and *Supporting Information*).

Bioinformatic Analyses Reveal Putative SPDIR Events in *Streptococcus pneumoniae*. We hypothesized that SPDIR is a general genetic mechanism forming microindels and clustered polymorphisms with intragenomic DNA. To test this hypothesis, we searched for variations consistent with SPDIR in the gram-positive human pathogen *Streptococcus pneumoniae* and in human genomic DNA samples using bioinformatic approaches. We performed initial DNA sequence analyses on 203 pairwise genome alignments from the well-characterized *S. pneumoniae* PMEN1 lineage (34) collected

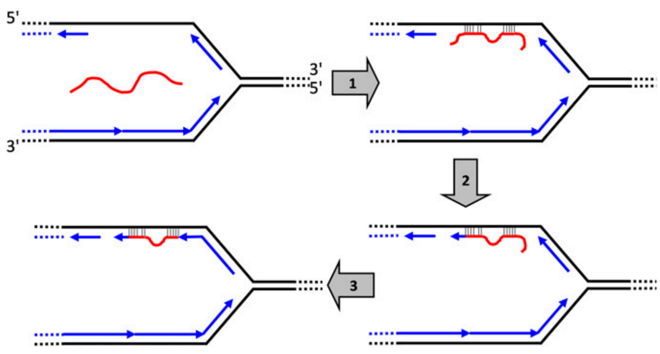

Fig. 2. Model for SPDIR mechanism illustrated with a DNA replication fork (black indicates parental DNA strands; blue arrows indicate newly synthesized DNA strands). The proposed mechanism expands on a synthesis of several microhomology-dependent IR models (15, 23, 24, 33). In step 1, an ssDNA molecule (red) anneals at one or more microhomologous regions with exposed ssDNA segments at the discontinuously synthesized arm. In step 2, the potential 3′-extension is processed, and the hybridized molecule is extended by a DNA polymerase. In step 3, the potential 5′-overhang is removed, and the processed end is covalently joined with the newly synthesized 3′-end of the next Okazaki fragment.

over 30 y. We called clustered polymorphisms as a set of ≥3 cumulative single-nucleotide polymorphisms (SNPs) with no more than eight base pairs (bp) between each SNP (*Supporting Information*). We subsequently identified genomic DNA segments that could have served as potential donor molecules for SPDIR events.

For each microhomology, we calculated the minimal free energy of hybridization (35) (ΔG^0_{min}) as a proxy for the annealing stability properties of a microhomology. Conservatively, we only considered DNA segments that displayed a lower ΔG^0_{min} than the weakest microhomology found in the experimental studies with *A. baylyi* (*Supporting Information* and Dataset S1). Using these criteria, we obtained a set of eight putative SPDIR events that are in accordance with the thermodynamical requirements identified experimentally (Dataset S2).

Although identification of false-positive donor molecules cannot be excluded using this retrospective approach, the likelihood of random occurrence of identical DNA segments of typically 13 or more bp occurring in intragenomic DNA is low (*Supporting Information*). False-positives due to accumulated point mutations or alternative microindel-generating processes cannot be completely ruled out. On the basis of estimates of yearly point mutation rates in the PMEN1 lineage (35) (1.57×10^{-6}) of ∼3.3 single-nucleotide changes per genome per year, the probability of multiple adjacent SNPs mimicking SPDIR events while the remainder of the genome remains unchanged is extremely low.

Bioinformatic Analyses Reveal Putative SPDIR Events in Human Genomes. For humans, we isolated DNA from blood samples and colon cancer tissues from three individuals (36) and sequenced the DNA on an Illumina HiSeq 2000. We called clustered polymorphisms with donor molecules for SPDIR largely as described above (see *Supporting Information* for details). Altogether, we identified 94 putative SPDIR events (Table S2 and Dataset S3). Detailed analyses showed that more than half of these events were short clustered nucleotide variations present in various human sequence databases including alternative genome assemblies, suggesting that SPDIR contributes to the generation of human heterozygous alleles and that SPDIR is a mutation mechanism operative in humans.

The remaining insertion–deletion sequences were not found in available databases and were considered novel (Table S2). Seven novel putative SPDIR events were uniquely identified in DNA from blood, whereas a total of 33 putative novel events were identified in DNA from colon cancer tissue only (Table 2). Remarkably, 16 novel SPDIR events from cancer tissue formed at predicted hairpins and led to microinversions (Fig. 3) that in two cases were imprecise (Fig. 3*B*). These microinversions predictively formed through donor ssDNA molecules that originated from the same locus but reannealed with the DNA single strand as reverse complement. Donor DNA molecules from loci very close to the SPDIR site were also observed in the experimental studies (*A. baylyi* SPDIR isolates A4 and K49; Dataset S1), but the *hisC*::′ND5i′ detection construct did not contain stem–loop secondary structures. Close proximity between donor locus and recombinant microindel locus may increase the likelihood for a SPDIR event.

Three novel SPDIR events, including a single microinversion, were identified both from cancer tissue and from blood (Table 2), suggesting somatic mutations early in embryogenesis, spread of genetic material within the body, or previously unknown heterozygous alleles. The observed predominance of SPDIR in colon cancer tissue possibly reflects the reduced activity of genome maintenance functions generally observed in cancer cells (37, 38). This observation is consistent with the increased SPDIR frequencies of genome maintenance mutants in our experimental bacterial system (Table 1).

GENETICS

Table 2. Combined numbers of novel SPDIR events from three human individuals

Putative novel SPDIR events	Cancer	Cancer and blood	Blood
Total	33	3	7
Associated with genes	20	1	4
ORFs	15	0	3
Potential control regions	5	1	1
Tumorigenesis	2	0	0
Growth and proliferation, differentiation, apoptosis, DNA binding, and transcription	8	1	2
Other functions	10	0	2
Not associated with genes	13	2	3
Microinversions	16	1	0

The potential SPDIR numbers for each human individual are listed in Table S2.

Discussion

In this study we identified a previously unrecognized mechanism, SPDIR, which generates clustered DNA polymorphisms. We show that SPDIR facilitates the formation of SNP clusters, microindels, and mosaic genes (experimentally observed substitutional insertion of up to 26 codons; Dataset S1). SPDIR occurs by ssDNA segments of intragenomic or extragenomic origins that invade and replace genomic DNA through two IR events.

Our genetic studies in *A. baylyi* with specific deletion mutants, together with the genotoxic stress and transformation experiments, clearly show that cytoplasmic ssDNA segments are responsible for SPDIR (Fig. 2). In wild-type cells, cytoplasmic ssDNA is a genomic damage signal, and the formation of ssDNA is tightly controlled (25). SPDIR can be classified both as a recombination and as a replication-associated mutation mechanism for clustered polymorphisms, with rare ssDNA segments acting as mutagens. Although oligonucleotides are known to recombine intracellularly or in the course of horizontal gene transfer (23, 24), and synthetic oligonucleotides are now widely used in targeted mutagenesis approaches, these events are based on DNA homology. SPDIR depends exclusively on microhomologies in otherwise heterologous DNA that can be as short as 12 bp and interrupted by mismatches and gaps.

SPDIR occurs rarely in *A. baylyi* wild-type cells. However, DNA damages increase the SPDIR frequency by orders of magnitude. Consequently, the cells turn into transient phenotypic mutators for microindels under genotoxic stress. The transient mutator phenotype does not require mutations in DNA repair genes, as frequently observed in mismatch repair-deficient mutators of prokaryotes and eukaryotes (39). It is conceivable that increased SPDIR frequencies can provide cells with a competitive advantage in fluctuating environments, as reported for genotypic mutators (40, 41). SPDIR can generate near-random genetic variations and alter entire protein domains in a single generation. It is thus tempting to speculate that SPDIR may be an important mechanism in protein evolution (42) following gene amplification and duplication events (43) (*Supporting Information*).

Our in silico identification of potential SPDIR events in both the gram-positive pathogen *S. pneumoniae* and in the human genome strongly suggests that SPDIR is a general mutation mechanism with relevance beyond our model organism *A. baylyi*. The identified microindel variants, together with the presence of intragenomic donor molecules, are consistent with the experimentally obtained SPDIR events and thus biologically plausible. Typical SPDIR-generated sequence changes are inaccessible by known point mutation or recombination processes, such as replication slippage, microhomology-dependent IR, NHEJ, DNA gyrase-mediated strand switching, or transpositions. However, sequence variations caused by SPDIR are comparable with those produced by MMEJ, a highly mutagenic DNA repair mechanism in eukaryotes (20). MMEJ is tightly down-regulated in healthy cells but often operative in tumor tissue. DNA double-strand breaks are repaired by MMEJ in an error-prone way, frequently leading to incorporation of ectopic DNA segments at the joints (20).

In our human tumor samples, we determined that 16 uniquely identified clustered polymorphisms were microinversions at predicted hairpins (Fig. 3 and Dataset S3). Microinversions at hairpins have been reported (44–46), but the mechanistic details of their formation remain elusive (45) and are considered unrelated to templated mutagenesis at imperfect hairpins (46). The formation of microinversions is also not consistent with our current understanding of the MMEJ or of other mutation mechanisms, and microinversions at hairpins have not been reported in MMEJ surveys (19, 20). However, microinversions can be explained most parsimoniously by SPDIR where the inverted repeats of the hairpins act as microhomologies and are used for the illegitimate joints (Fig. 3*A*), consistent with the model shown in Fig. 2. Our results indicate that SPDIR-caused mutations occur in colon cancer at elevated frequencies but not in the whole blood control. In many cancers, including those with up-regulated MMEJ, genome maintenance functions such as Rad51 (eukaryotic RecA homolog) and ssExo are down-regulated (47, 48). It is conceivable that SPDIR occurs at elevated frequencies in such tumor cells, as experimentally observed in the *A. baylyi* Δ*recA* Δ*recJ* Δ*exoX* triple mutant (Table 1). The role of SPDIR in cancer progression requires further exploration.

Materials and Methods

The *A. baylyi* mutant strains were constructed as described (23, 32, 49) with standard procedures (*Supporting Information*) and are listed in Table S3. The mutation experiments were conducted in liquid cultures that were inoculated with a single colony of a His$^-$ strain and aerated for 15 h at 30 °C in LB broth. The cells were washed, plated on M9 minimal medium with 10 mM succinate (M9S; His$^+$ mutant titer) and in appropriate dilution on LB (total cell titer), and incubated at 30°. When applicable, ciprofloxacin or DNA was added before inoculation. When UV was used as DNA-damaging agent, the cells were grown for 11 h, washed in PBS, irradiated with a germicidal lamp, and then grown in LB for another 4 h. On the M9S selective plates, His$^-$ cells grow less than one generation.

His$^+$ colonies on M9S were picked after 40 h (*recA*$^+$ strains) or 64 h (Δ*recA* strains) and restreaked on M9S, and the recombinant *hisC* segment was amplified by PCR and Sanger-sequenced (*Supporting Information*). To identify ectopic inserts, the sequencing results were aligned with the *A. baylyi* genome and, when donor DNA for natural transformation was used, with donor DNA sequences, using BLAST (50).

The bioinformatic approaches are described in detail in *Supporting Information*. The R scripts are available from the authors upon request.

Two ethical boards reviewed the protocol for investigation of the human samples included in this study: the Regional Committee on Health Research Ethics (Case H-2–2012-FSP2) and the National Committee on Health

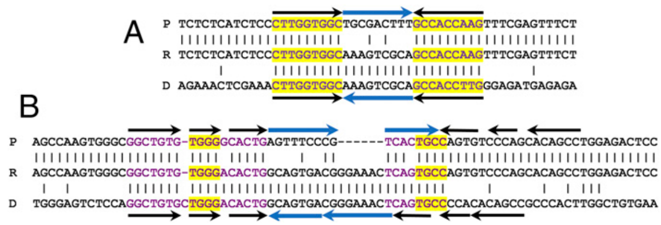

Fig. 3. Examples of microinversions at predicted hairpins identified in cancer tissue from human individuals. Black arrows indicate the inverted repeats (IR), and blue arrows indicate the loop orientation. Other color codings are the same as in Fig. 1 *B–E*. All potential SPDIR events found in the human genomes are listed in Dataset S3. (*A*) The class 2 microinversion Z441, located in the proto-oncogene *SASH1* of colon cancer tissue from individual 1. (*B*) Example of a microinversion that was fully annealed at the left IR but misannealed at the right IR (using an alternative microhomology for the right illegitimate joint), resulting in a net gain of six bp (Z2579; individual 3, colon cancer, intergenic region).

Research Ethics (Case 1304226). Both review boards approved the human research and waivered the requirement for informed consent, in accordance with national legislation (Sundhedsloven) (36).

ACKNOWLEDGMENTS. We thank Sören Abel, Pia Abel zur Wiesch, Terje Johansen, Trond Lamark, Vidar Sørum, and Wilfried Wackernagel for helpful discussions and Jose Victor Moreno Mayar for assistance with bioinformatic approaches. We thank BGI Europe and the Danish National High Throughput Sequencing Centre for sequencing of the cancer samples and for technical assistance. The cancer work was supported by The Danish National Advanced Technology foundation (The GenomeDenmark platform, Grant 019-2011-2). This work was supported by The Arctic University of Norway and the Danish National Research Foundation (K.H.), the Academy of Finland Grant 251170 (to P.M.), the Finnish Centre of Excellence in Computational Inference Research Grant 259272 (to P.M.), and also by Norwegian Research Council Grant 204263/F20 (to P.J.J.).

1. Cooper GM, et al. (2004) Characterization of evolutionary rates and constraints in three Mammalian genomes. *Genome Res* 14(4):539–548.
2. Mostowy R, et al. (2014) Heterogeneity in the frequency and characteristics of homologous recombination in pneumococcal evolution. *PLoS Genet* 10(5):e1004300.
3. Gibbons HS, et al. (2012) Comparative genomics of 2009 seasonal plague (Yersinia pestis) in New Mexico. *PLoS One* 7(2):e31604.
4. Chewapreecha C, et al. (2014) Comprehensive identification of single nucleotide polymorphisms associated with beta-lactam resistance within pneumococcal mosaic genes. *PLoS Genet* 10(8):e1004547.
5. Mills RE, et al. (2011) Natural genetic variation caused by small insertions and deletions in the human genome. *Genome Res* 21(6):830–839.
6. Huang S, Li J, Xu A, Huang G, You L (2013) Small insertions are more deleterious than small deletions in human genomes. *Hum Mutat* 34(12):1642–1649.
7. Pu Y, et al. (2014) Association of an insertion/deletion polymorphism in IL1A 3′-UTR with risk for cervical carcinoma in Chinese Han Women. *Hum Immunol* 75(8):740–744.
8. Ahmad F, Lad P, Bhatia S, Das BR (2015) Molecular spectrum of c-KIT and PDGFRA gene mutations in gastro intestinal stromal tumor: Determination of frequency, distribution pattern and identification of novel mutations in Indian patients. *Med Oncol* 32(1):424.
9. Kunkel TA (2004) DNA replication fidelity. *J Biol Chem* 279(17):16895–16898.
10. Viswanathan M, Lacirignola JJ, Hurley RL, Lovett ST (2000) A novel mutational hotspot in a natural quasipalindrome in Escherichia coli. *J Mol Biol* 302(3):553–564.
11. Nik-Zainal S, et al.; Breast Cancer Working Group of the International Cancer Genome Consortium (2012) Mutational processes molding the genomes of 21 breast cancers. *Cell* 149(5):979–993.
12. Amos W (2010) Even small SNP clusters are non-randomly distributed: Is this evidence of mutational non-independence? *Proc Biol Sci* 277(1686):1443–1449.
13. Lovett ST, Drapkin PT, Sutera VA, Jr, Gluckman-Peskind TJ (1993) A sister-strand exchange mechanism for recA-independent deletion of repeated DNA sequences in Escherichia coli. *Genetics* 135(3):631–642.
14. Bois P, Jeffreys AJ (1999) Minisatellite instability and germline mutation. *Cell Mol Life Sci* 55(12):1636–1648.
15. Ehrlich SD, et al. (1993) Mechanisms of illegitimate recombination. *Gene* 135(1-2): 161–166.
16. Lee JA, Carvalho CM, Lupski JR (2007) A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell* 131(7):1235–1247.
17. Moore JK, Haber JE (1996) Cell cycle and genetic requirements of two pathways of nonhomologous end-joining repair of double-strand breaks in Saccharomyces cerevisiae. *Mol Cell Biol* 16(5):2164–2173.
18. Naito A, Naito S, Ikeda H (1984) Homology is not required for recombination mediated by DNA gyrase of Escherichia coli. *Mol Gen Genet* 193(2):238–243.
19. Yu AM, McVey M (2010) Synthesis-dependent microhomology-mediated end joining accounts for multiple types of repair junctions. *Nucleic Acids Res* 38(17):5706–5717.
20. Mateos-Gomez PA, et al. (2015) Mammalian polymerase θ promotes alternative NHEJ and suppresses recombination. *Nature* 518(7538):254–257.
21. Lorenz S, et al. (2016) Unscrambling the genomic chaos of osteosarcoma reveals extensive transcript fusion, recurrent rearrangements and frequent novel TP53 aberrations. *Oncotarget* 7(5):5273–5288.
22. Kent T, Chandramouly G, McDevitt SM, Ozdemir AY, Pomerantz RT (2015) Mechanism of microhomology-mediated end-joining promoted by human DNA polymerase θ. *Nat Struct Mol Biol* 22(3):230–237.
23. Overballe-Petersen S, et al. (2013) Bacterial natural transformation by highly fragmented and damaged DNA. *Proc Natl Acad Sci USA* 110(49):19860–19865.
24. Dutra BE, Sutera VA, Jr, Lovett ST (2007) RecA-independent recombination is efficient but limited by exonucleases. *Proc Natl Acad Sci USA* 104(1):216–221.
25. Shinagawa H (1996) SOS response as an adaptive response to DNA damage in prokaryotes. *EXS* 77:221–235.
26. Lyamichev V, Brow MA, Dahlberg JE (1993) Structure-specific endonucleolytic cleavage of nucleic acids by eubacterial DNA polymerases. *Science* 260(5109):778–783.
27. Darmon E, Leach DR (2014) Bacterial genome instability. *Microbiol Mol Biol Rev* 78(1): 1–39.
28. Chen CR, Malik M, Snyder M, Drlica K (1996) DNA gyrase and topoisomerase IV on the bacterial chromosome: Quinolone-induced DNA cleavage. *J Mol Biol* 258(4):627–637.
29. Malik M, Zhao X, Drlica K (2006) Lethal fragmentation of bacterial chromosomes mediated by DNA gyrase and quinolones. *Mol Microbiol* 61(3):810–825.
30. Bonura T, Smith KC (1975) Enzymatic production of deoxyribonucleic acid double-strand breaks after ultraviolet irradiation of Escherichia coli K-12. *J Bacteriol* 121(2): 511–517.
31. Smith HO, Danner DB, Deich RA (1981) Genetic transformation. *Annu Rev Biochem* 50:41–68.
32. Kickstein E, Harms K, Wackernagel W (2007) Deletions of recBCD or recD influence genetic transformation differently and are lethal together with a recJ deletion in Acinetobacter baylyi. *Microbiology* 153(Pt 7):2259–2270.
33. Hülter N, Wackernagel W (2008) Double illegitimate recombination events integrate DNA segments through two different mechanisms during natural transformation of Acinetobacter baylyi. *Mol Microbiol* 67(5):984–995.
34. Croucher NJ, et al. (2011) Rapid pneumococcal evolution in response to clinical interventions. *Science* 331(6016):430–434.
35. Wetmur JG (1996) Nucleic acid hybrids, formation and structure of. *Encyclopedia of Molecular Biology and Molecular Medicine*, ed Meyers RA (VCH Press, New York), Vol 4, pp 235–243.
36. Vinner L, et al. (2015) Investigation of human cancers for retrovirus by low-stringency target enrichment and high-throughput sequencing. *Sci Rep* 5:13201.
37. Hoeijmakers JHJ (2001) Genome maintenance mechanisms for preventing cancer. *Nature* 411(6835):366–374.
38. Hanahan D, Weinberg RA (2011) Hallmarks of cancer: The next generation. *Cell* 144(5):646–674.
39. Modrich P (1991) Mechanisms and biological effects of mismatch repair. *Annu Rev Genet* 25:229–253.
40. Mao EF, Lane L, Lee J, Miller JH (1997) Proliferation of mutators in A cell population. *J Bacteriol* 179(2):417–422.
41. Taddei F, Vulić M, Radman M, Matić I (1997) Genetic variability and adaptation to stress. *EXS* 83:271–290.
42. Alendé N, Nielsen JE, Shields DC, Khaldi N (2011) Evolution of the isoelectric point of mammalian proteins as a consequence of indels and adaptive evolution. *Proteins* 79(5):1635–1648.
43. Sandegren L, Andersson DI (2009) Bacterial gene amplification: Implications for the evolution of antibiotic resistance. *Nat Rev Microbiol* 7(8):578–588.
44. Kelchner SA, Wendel JF (1996) Hairpins create minute inversions in non-coding regions of chloroplast DNA. *Curr Genet* 30(3):259–262.
45. Kim KJ, Lee HL (2005) Widespread occurrence of small inversions in the chloroplast genomes of land plants. *Mol Cells* 19(1):104–113.
46. Schultz GE, Jr, Drake JW (2008) Templated mutagenesis in bacteriophage T4 involving imperfect direct or indirect sequence repeats. *Genetics* 178(2):661–673.
47. Thompson LH, Schild D (2001) Homologous recombinational repair of DNA ensures mammalian chromosome stability. *Mutat Res* 477(1-2):131–153.
48. Chow TY, Choudhury SA (2005) DNA repair protein: Endo-exonuclease as a new frontier in cancer therapy. *Future Oncol* 1(2):265–271.
49. Harms K, Schön V, Kickstein E, Wackernagel W (2007) The RecJ DNase strongly suppresses genomic integration of short but not long foreign DNA fragments by homology-facilitated illegitimate recombination during transformation of Acinetobacter baylyi. *Mol Microbiol* 64(3):691–702.
50. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215(3):403–410.
51. de Vries J, Heine M, Harms K, Wackernagel W (2003) Spread of recombinant DNA by roots and pollen of transgenic potato plants, identified by highly specific biomonitoring using natural transformation of an Acinetobacter sp. *Appl Environ Microbiol* 69(8):4455–4462.
52. Sambrook J, Fritsch EF, Maniatis T (1989) *Molecular Cloning: A Laboratory Manual* (Cold Spring Harbor Lab Press, New York).
53. Seguin-Orlando A, et al. (2013) Ligation bias in illumina next-generation DNA libraries: Implications for sequencing ancient genomes. *PLoS One* 8(10):e78575.
54. Treangen TJ, Ondov BD, Koren S, Phillippy AM (2014) The Harvest suite for rapid core-genome alignment and visualization of thousands of intraspecific microbial genomes. *Genome Biol* 15(11):524.
55. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754–1760.
56. de Berardinis V, et al. (2008) A complete collection of single-gene deletion mutants of Acinetobacter baylyi ADP1. *Mol Syst Biol* 4:174.
57. Barbe V, et al. (2004) Unique features revealed by the genome sequence of Acinetobacter sp. ADP1, a versatile and naturally transformation competent bacterium. *Nucleic Acids Res* 32(19):5766–5779.
58. Sharma B, Hill TM (1995) Insertion of inverted Ter sites into the terminus region of the Escherichia coli chromosome delays completion of DNA replication and disrupts the cell cycle. *Mol Microbiol* 18(1):45–61.
59. Luria SE, Delbrück M (1943) Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* 28(6):491–511.