Homologous recombination repair creates mutations in the non-coding genome that alter Topoisomerase-1 cleavage sites & orchestrates irinotecan resistance

Santosh Kumar[1†], Valid Gahramanov[1†], Julia Yaglom[1], Shivani Patel[1†], Lukasz Kaczmarczyk[1], Ivan Alexandrov[2], Gabi Gerlitz[1], Mali Salmon-Divon[1], Michael Y. Sherman[1*]

Affiliations:

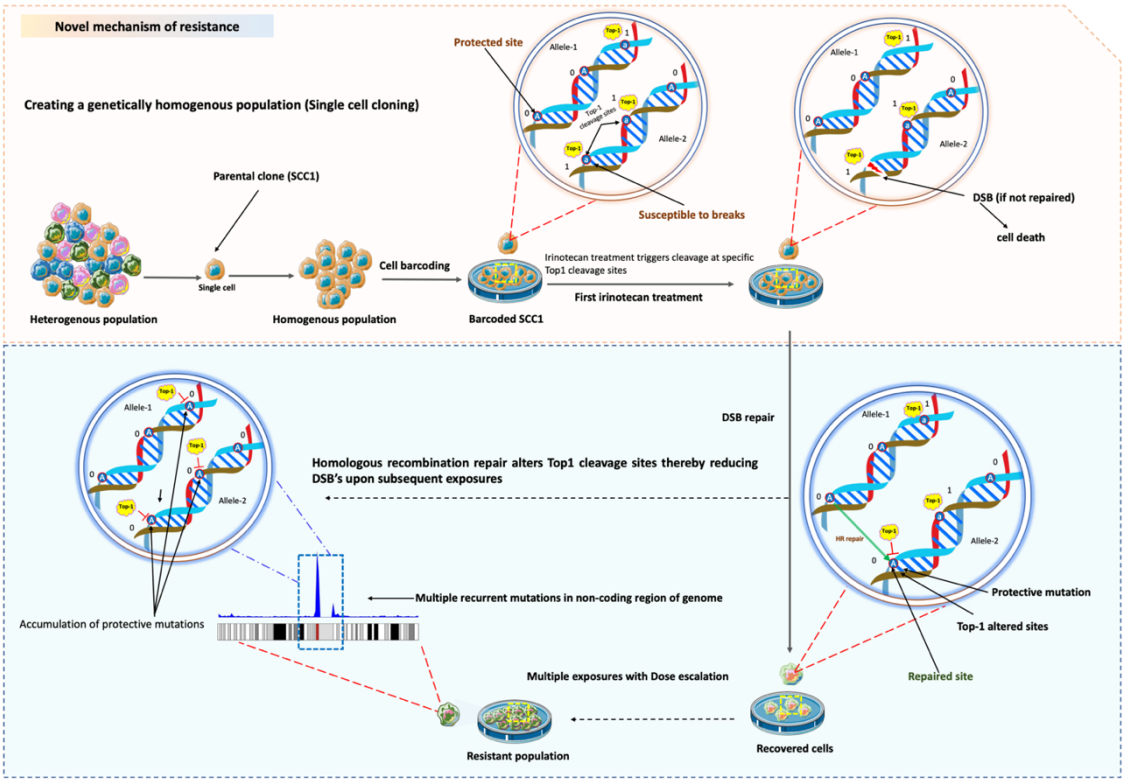
1. Department of Molecular Biology, Ariel University, Israel-40700.
2. Research Center of Biotechnology of the Russian Academy of Sciences, Moscow, Russia-119071.

[†]Equal contribution
* Corresponding author

Email: sherma1@ariel.ac.il

Keywords: Drug-adaptation, senescence-like state, topoisomerase 1 inhibitor, resistance, mutation.

Graphical abstract



Abstract

Development of resistance to chemotherapy is a leading cause of treatment failure. Drug-resistance mechanisms typically involve mutations in specific proteins or changes in their levels of

expression. It is commonly understood that such mutations happen randomly prior to treatment and are positively selected during the treatment. We realized that selection of drug-resistant cancer cell mutants in culture could be achieved by multiple drug exposures of cloned genetically identical cells, and thus is unlikely to result from selection of pre-existent mutations. Accordingly, adaptation must involve generation of mutations de-novo in the process of drug treatment. Here we explored both the mechanism of resistance and the origin of resistant mutations in response to a widely used Top1 inhibitor irinotecan, which triggers DNA breaks, resulting in cytotoxicity. Resistance mechanism was based on gradual accumulation of recurrent mutations in non-coding regions of DNA at sequence-specific Top1 cleavage sites. Surprisingly, cancer cells had much higher number of such sites than reference genome, which may define their increased sensitivity to irinotecan. Homologous recombination repair of DNA double strand breaks at these cleavage sites following initial drug exposures gradually reverted cleavage-sensitive cancer-related sequences back to cleavage-resistant "normal" genome sequences. These genetic changes reduced generation of double strand DNA breaks upon subsequent exposures, thus gradually increasing the drug resistance. Together, large target size for resistance mutations and their generation specifically at Top1 cleavage spots lead to gradual and rapid accumulation of such mutations, synergistically accelerating development of resistance.

Introduction

Resistance to chemotherapy is a leading cause of treatment failure. In line with prior knowledge of development of antibiotic-resistant bacteria (1–6), the major concept in the cancer field is that resistance mutations happen randomly prior to a treatment and are positively selected during the treatment. This notion was directly supported by experiments with cell barcoding that studied selection of resistant clones followed by the barcode analysis (7–11). In these experiments a fraction of clones selected in parallel independent treatments had same barcodes, strongly

suggesting that these clones originated from the same parental cells which carried resistant mutations prior to the beginning of drug treatment (12–14). However, other selected resistant clones that carried different barcodes were not further investigated in this study and may result from mutations that were generated in response drug treatment. Overall, in selection of drug-resistant mutations in culture, cells are adapted to low drug concentrations, and then via multiple passages with dose escalation, resistant mutants are selected (6,15–22). The process usually takes several months and provides resistance to higher than initial drug concentrations, but not to full drug resistance. Importantly, selection of drug-resistant mutants in dose escalation experiments could be achieved from cloned genetically identical cell population, suggesting that adaptation must involve generation of mutations de-novo in the process of drug treatment.

Exploration of forces driving drug-resistance under the dose-escalation, may help to uncover novel mechanisms of development of resistance in clinical setting (16,17,23–32). Indeed, multiple administration of drug doses in clinic appears to mimic the in vitro scheme of drug resistance selection. Understanding the mechanisms of the development of drug resistance can also provide novel insights to guide the design of drug combinations and treatment strategies.

Here, we investigated how these adaptive mutations may emerge in colon cancer cells with the example of resistance to a widely used anti-cancer drug irinotecan. Irinotecan is a pro-drug, which is converted into the active drug SN-38, which binds to Top1 (33–36). The binding, in turn, allows Top1 to make DNA breaks, but prevents re-ligation (34). Top1-mediated single strand breaks may facilitate double strand DNA breaks that, if unrepaired, cause cancer cell death (33). Top1 works both during transcription and replication (34). A number of mechanisms of resistance to Irinotecan have been described, including mutations in Top1, upregulation of multidrug-resistance pumps and associated enzyme system (36–40). It is difficult, however, to understand why a dose escalation scheme could be important for development of drug resistance based on these mechanisms. Here,

we evaluated the efficiency of the development of irinotecan resistance and uncovered a novel resistance mechanism based on the active generation of a large number of mutations in Top1-dependent DNA breaking sites that reduce the chances of double strand breaks upon consequent irinotecan exposures in the process of dose escalation.

Results

Experimental design with multiple irinotecan treatments

Since irinotecan is widely used against colon cancer, we investigated its effects on colon cancer cell line HCT116. In order to achieve genetic uniformity of the population, we cloned the cells and isolated seven independent clones. Genetic uniformity within each clone indicates that any drug-resistance mutation selected in our experiments occurs either in the process the colony growth from a single cell prior to the drug treatment or is actively generated in the process of drug treatment. Sensitivity to irinotecan differed dramatically between the clones, with minimal toxic concentration between 1nM and 40nM. We chose two clones, SSC1 and SSC7, with high sensitivity for further experiments.

To understand development of drug resistance, SSC7 cells were exposed to 4nM irinotecan, which led to the death of a significant fraction of the population, and the cell cycle arrest of the rest of the population. Cells remained in the arrested state without divisions for 14 days, and then resumed growth. Second treatment with 4nM irinotecan led to a shorter period of the cell cycle arrest. Such a treatment cycle was conducted five times in total. At the fifth cycle, practically no cell death or growth inhibition was seen, indicating that cells became fully adapted to this concentration of irinotecan (Fig. 1A, B).

Further, we tested if a similar pattern occurs when cells develop adaptation to high doses (40nM) of irinotecan. The fraction of cells that survived 40nM stayed in a senescent state for more than three months, after which cells resumed propagation and filled the plate. Upon subsequent exposure of the recovered population to 40nM irinotecan the period of the growth arrest was only about one month. The process was repeated three more times, and each time the fraction of dying cells was lower and the time period of growth arrest was shorter compared to the previous round of selection, so that following the fourth exposure, cells spent approximately one week in the arrested state (Fig 1A, B). Altogether, these findings indicate similarities of adaptation to low and high doses of irinotecan.

Most of the survived cells recover from the senescence like growth arrest

A large fraction of cells in the population could adapt to the treatment and resume growth after the cell cycle arrest. Alternatively, a small fraction of cells that could be originally resistant to the drug continued to propagate, which became detectable only when they began outgrowing the rest of the arrested population. To distinguish between these possibilities, SSC7 cells were infected with the cell cycle reporter virus (41) and exposed to 4nM irinotecan. Survived treated cells stopped dividing, acquired senescence-like phenotype, and according to the reporter, underwent G1 growth arrest (Fig.1C, D). Cells remained in G1 for four days, after which a majority of cells exited G1 and entered the cell cycle (Fig.1C, D). Entering the cycle following the senescence-like arrest was surprisingly slow, and only by day 8 almost 100% of cells reached G2. There were no localized shifts of cells to G2, indicating the lack of clonal expansion. This observation indicates that cells underwent true adaptation to irinotecan, rather than reflects the expansion of a small fraction of initially resistant cells.

A large fraction of cells gets adapted to irinotecan

To quantitatively assess the process of adaptation, 10 million SSC1 cells were individually barcoded using Cellecta 50M barcodes lentiviral library (42). When cells filled the plate, half of them were collected for the barcode analysis, and half were used for further dose-escalation experiments with sequential passaging over 2, 4, 6, 8, and 15nM of irinotecan. With each passage, we observed a reduction of the population of dying cells, and a reduction of the period of growth arrest (Fig. 1E). At different stages of the experiment, we collected cells for the analysis of barcodes, as described in Materials and Methods.

Comparison of barcodes that were detected in the control population and population treated with 2nM showed that about $4^{*}10^{-3}$ of the original clones survived the selection (Fig. 1E). Analysis of barcodes in the population of cells after the 15nM irinotecan selection demonstrated that $2^{*}10^{-3}$ of original clones survived the entire series of selections, which is only two times lower than the number of clones survived the first round of selection at 2nM. These findings indicate that (a) the probability of the survival of clones is high compared to the usual probability of spontaneous or even mutagen-induced mutations (according to classical studies usually not higher than 10^{-4}(3,4,15,17–21,43,44)), and (b) almost 50% of clones that survive 2nM selection survived the entire series of selections, suggesting that if cells are able to survive initial treatment they can also survive the dose escalation treatment.

To test if the dose escalation process is critical for the development of the resistant forms, barcoded SSC1 cells were exposed directly to 15nM irinotecan. Analysis of barcodes indicated that only $2^{*}10^{-5}$ of clones survived (Fig. 1E), which is 100X lower than the survival rate of 15nM irinotecan in the dose escalation experiment. These data indicate that the dose escalation protocol is important for effective development of resistant variants.

Changes in transcriptome may not be involved in the resistance development

To explore the mechanisms of the adaptation, the survived population of multiple rounds of treatment with 40nM irinotecan, was cloned again. Barcodes from several clones were isolated and sequenced. Three clones with different barcodes (MSC1, MSC2 and MSC3) were chosen for further analysis. Since the barcoding of cells was done prior to the entire series of irinotecan treatments, the fact that these clones carry different barcodes indicated that they did not split from the same clone somewhere in the middle of the irinotecan treatment. In other words, they represent the progeny of cells that underwent the entire series of the treatments independently of each other. We would like to reiterate that the original barcoded population was genetically homogenous because of the initial cloning.

To uncover potential mechanisms related to changes in gene expression, parental SSC1 clone and individual mutant isolates MSC1, MSC2 and MSC3 were exposed to 10nM irinotecan and their transcriptome was analyzed by RNAseq. A number of differentially expressed genes was observed in the survived clones. Importantly, we did not observe changes in either *MDR1* or other *ABC* transporters involved in drug efflux, or Top1, or DNA repair genes (Table S1), suggesting that the mechanism of resistance in these clones may not be related to expected changes in the transcriptome.

DSBs significantly contribute to irinotecan-induced cell death

To further explore potential mechanisms of the resistance, we performed a pooled shRNA screen to identify genes important for survival of irinotecan treatment. SSC1 cells were infected with focused lentiviral shRNA library Decipher Module-1(45) that targets signaling pathways. This library covers about 20% of human genes. Cells were treated with 10nM irinotecan for 24 hours and cells that survived the treatment after 5 days were collected and the barcodes isolated, sequenced and analyzed. As control, we used the same population of infected cells but without

irinotecan treatment. Among genes, depletion of which showed sensitizing effects, was a group of genes that plays a role in DNA double strand breaks repair, predominantly representing the homologous recombination (HR) and nonhomologous end joining (NHEJ) DNA repair pathways, including *POLE, POLE3, POLE4, KAT5, RAD51C, RAD54L, RAD1, RAD9A, H2AFX, LIG4* and *PARP2* (Table S2). We also observed a number of genes involved in translesion DNA synthesis (Table S2). These data reinforce the understanding that (a) generation of double strand DNA breaks is critical for cell death caused by irinotecan and (b) HR and to some extent NHEJ repair pathways play an important role in the irinotecan survival in naïve cells (though these pathways are not upregulated in the resistant clones (Table S1)).

Development of irinotecan resistance associates with emergence of multiple non-random mutations

To study mutations that emerged in the survived clones, we performed whole genome sequencing. Genomes of the survived clones MSC1, MSC2 and MSC3 were compared with the genome of original population SSC1, and each of them with the Reference Genome (GRCh38/hg38) in the UCSC (46) database. We observed that the parental SSC1 genome had hundreds of thousands of SNPs and InDels compared to the Reference Genome.

These mutations may reflect the fact that HCT116 cells were isolated from a different individual than a group of individuals whose sequence compose the Reference Genome. Alternatively, these mutations could arise in the process of cancer development and/or further culturing of HCT116 cells in the laboratory conditions. Overall analysis of mutation in SSC1 indicated that 93% of them do not correspond to known SNPs in the human population, suggesting that the vast majority of the mutations simply reflect either the cancer nature of these cells or genetic instability upon culturing. Accordingly, they will be further called “cancer alleles”. When genomes of the irinotecan-resistant

isolates MSC1, MSC2 and MSC3 were compared with the genome of SSC1, we identified hundreds of thousands of InDels and SNPs that arose in the process of adaptation to irinotecan.

Strikingly a very large fraction of these mutations was common between the independently isolated clones (Fig. 2A, B). If compared by pairs, i.e. MSC1/MSC2, MSC1/MSC3 and MSC2/MSC3, in each pair between 17% and 45% of mutations were common, and between 7% and 15% were common between all three independent isolates (46,099 mutations, of which 28.4% were InDels and 71.6% were point mutations (Table S4)). Even considering that irinotecan may have a high mutagenic activity and triggers protective mutations with the rate as high as 10^{-4} per nucleotide, the probability of overlap of a mutation in three independent clones will be 10^{-12} (i.e. way less than one triple mutation per $3*10^{9}$ bp genome), which is many orders of magnitude lower than seen in the experiment. Thus, the overlapping (common between three clones) mutations clearly point to a non-random mutation mechanism. Generation of these mutations seem to be guided by a mechanism, understanding of which may clarify the adaptation pathway.

Analysis of the ENCODE (promotors and enhancer datasets (47)), GEO (GSE57628, mapping of Top1 sites in human HCT116 cells)(48) and UCSC(46) database datasets uncovered that the vast majority of the mutations are present in heterochromatin (high content of histone H3K9me3) in the noncoding regions (Fig. 2C). There was a small fraction of mutations present in promotors/enhancers (<1%) and the coding regions (2%), most of them in the exons (Fig. 2B, C). Importantly, the genes that have mutations in their coding and regulatory regions did not belong to known pathways associated with either cell survival or DNA repair, and therefore are unlikely to be involved in the adaptation process (Table S3, also see mutation landscape section in supplement for description on an exception case). To avoid dilution of the focus of this paper, we present more detailed analysis of mutations in the Supplement information (see supplement section Mutation Landscape, mutation mapping to regions in genome and its correlations to distinct signatures such

as nature of repetitive elements (Fig. S1), identification of pericentromeric and heterochromatin region based on methylation signatures of H3k9me3 and H3k27me3 (Fig. S2), Genome wide density plot for triple mutations (Fig. S3)). Altogether, these data suggest that adaptation to irinotecan was not associated with either mutations in functional genes or changes of expression of these genes.

Mutations result from repair of Top1-generated DSBs

To understand how hundreds of thousands of mutations in repeats and untranslated regions (see Mutation landscape section in the Supplement) could be involved in adaptation to irinotecan, we proposed that they could result from DNA breaks generated by Top1. Indeed, the mutation sites strongly co-localized with sites of DNA cleavage by Top1 (48) (19.8% of cases). By applying hypergeometric distribution algorithm (see Materials and Methods) we established that such co-localization is highly statistically significant ($p<<e^{-200}$). This percent of overlap is probably a strong underestimation since experimental conditions in two studies were different, i.e. long-term generation of mutations vs short-term Top1 inhibitor treatment. Furthermore, when overlaid distribution of triple mutations along the chromosomes over distribution of Top1-cleavage sites, we observed a strong overlap of peaks (Fig. 2D). Therefore, mutations take place at Top1-cleavage sites.

There are several lines of evidence that most of the mutations were generated either by HR or NHEJ repair of DSBs. 16% of the mutations in the isolates were clustered (7,510 out of 46,099 triple mutations sample), where 2-7 mutations were present within regions of up to 100 bp (we chose this length of DNA for definition of clustered mutations, see Materials and Methods, (Fig. 3A & Tables S5, S6). Since random positioning predicts that mutations should be on average separated by about 10,000 bp (about 300,000 mutations per clone distributed over 3 billion base pairs of the total genome), such clustering demonstrated their non-random appearance and suggested a mechanism

of their generation. In all these cases, these were loss of heterozygosity (LOH) mutations. These LOH mutations did not result from deletions of one of the alleles (since the number of reads corresponding to these regions were similar to the average number of reads along the genome), but by copying of an allele from one chromosome to another, including copying of the entire mutation cluster (Fig. 3B). Such copying could result only from the HR repair of DSBs. LOH occurred in 78% of common mutations (36,228 mutations out of 46,099), and similar fraction of LOH was found with the overall set of mutations in resistant clones (Table S7), suggesting that the majority of mutations were generated by the HR repair system.

In the resistant clones 28% of de-novo mutations that were not LOH resulted from NHEJ (9,871 de novo common mutations, and similar fraction of NHEJ was found in the overall set of mutations). They resulted from NHEJ because in case of insertions, these mutations have a very specific signature of duplication of a neighboring region. This duplication results from pairing of broken ends at the terminal nucleotides and filling the gaps on both strands via translesion DNA synthesis (Fig. 3B). This structure allows precise identification of the site of the DNA break, i.e. at the site of pairing between the duplication regions (Fig. 3B) (see also Mutation landscape section and Table S8). Overall, de novo appearing InDels can be used as hallmarks of DNA breaks that were repaired via NHEJ, while LOH mutations can be used as hallmarks of DNA breaks repaired via HR.

Very importantly, a high fraction of the overlapping mutations between independent resistant clones suggests that the irinotecan-inhibited Top1 generates DNA breaks at specific sites in the chromatin (possibly specific Top1 binding or activation sites), which further leads to generation of mutations upon HR or NHEJ repair of DSBs.

Mechanism of adaptation to irinotecan

The following considerations brought a framework for understanding the mechanism of adaptation. At each genome location, the parental SSC1 population could have alleles either with no mutations compared to the Reference Genome (0/0), with mutation in one allele (0/1) (heterozygosity), or both alleles (1/1) (Fig. 4A, B). Sites where two alleles in SSC1 had different mutations compared to the Reference Genome (1/2) were extremely rare. Accordingly, mutations that appear in MSC clones compared to parental SSC1 could be mutations de-novo generated by NHEJ (e.g. 0/0→0/1) or loss of heterozygosity generated by HR (0/1→0/0; or 0/1→1/1) (Fig. 4C, D).

In sites with LOH repaired by HR, a heterozygous allele could revert to either the Reference Genome allele (0/1 to 0/0) or to the "cancer" allele seen in the parental clone SSC1 (0/1 to 1/1). A surprising key observation that led to understanding the mechanism of adaptation was that the frequency of 0/1 to 0/0 shifts was 5.44 times higher than 0/1 to 1/1 shifts (Fig. 5A). 0/1 to 0/0 shift means that the allele from the Reference Genome was copied to the DSB at the "cancer" allele (an allele in SSC1 parental cells that differs from the Reference Genome), which ultimately means that the probability of double strand DNA breaks in "cancer" alleles is 5.44 time higher compared to the Reference Genome allele. The probability of breaks in "cancer" allele was even higher in the pericentromeric regions, where the ratio of breaks in Reference Genome allele to "cancer" allele was 1/20 (Fig. 5B). In the chromosome arms this ratio was 1/3.4. Therefore, surprisingly, alleles that acquired mutations in the course of cancer development were significantly more prone to Top1-induced double strand breaks than normal human genome alleles (Fig 5B).

This unexpected finding provides the mechanism of gradual adaptation to irinotecan. Initial exposures to irinotecan generate reversion of a number of "cancer" alleles to the Reference Genome alleles, which are more resistant to Top1-induced DSBs, which in turn creates a protective mechanism against following exposures. In other words, with each exposure there will be fewer and fewer potential Top1-cleavage sites, which will lead to stronger and stronger adaptation. This

novel mechanism of adaptation to irinotecan does not involve expression of any protective proteins or mutations in them, but rather involves a high number of changes in the DNA structure that make it less prone to Top1-induced breaks.

Adaptation associates with reduced ability of irinotecan to trigger DNA breaks

This mechanism predicts that in the process of adaptation, following multiple exposures to the same concentration of irinotecan, cells should experience fewer DSBs, while the rate of repair of DSBs remains the same. To test this prediction, we took cells that have not been drug exposed, and that underwent five cycles of exposure to 4nM of irinotecan. Both populations were subjected to 4nM of irinotecan for 24 hours, and the number of γ-H2AX foci was assessed by immunofluorescence. While irinotecan exposure of naïve cells caused a dramatic increase in the number of foci, exposure of cells that were preadapted by five cycles of irinotecan treatment barely caused foci formation (Fig. 6A, B). On the other hand, the rate of DSB repair (recovery of γH2AX foci) was not accelerated (Fig. 6C). Importantly, lower foci formation correlated with the lack of cell death and growth arrest. Similarly, there was a lower overall number of DNA breaks in adapted cells, as judged by the comet assay (Fig. 7A-E). Therefore, indeed reversion of "cancer" alleles to the Reference Genome alleles appears to be associated with fewer DSBs.

Discussion

Here, we addressed why the approach towards the selection of drug-resistant mutants in cancer cells requires multiple exposures to drugs and dose escalation. Such a selection scheme suggests that (a) either cells are somehow adapted to the low concentrations of drugs (e.g., via epigenetics mechanisms), and this adaptation guides further selection of the resistant mutant forms; or alternatively (b), development of drug resistance involves acquiring of a large number of mutations, each of which provides a fraction of the resistance, but gradually they accumulate and get selected

in the process of drug escalation. At least with an inhibitor of Top1 irinotecan, we show that the second possibility is correct. It appeared that a very high number of mutations is generated in the process of selection of irinotecan-resistant mutants in the dose escalation experiment. Surprisingly, the absolute majority of them were in the non-coding and silenced regions of the genome, suggesting that these mutations do not affect expression or function of specific genes involved in the irinotecan resistance. Accordingly, this mechanism of adaptation is fundamentally different from previously known mechanisms related to changes in drug target, drug metabolism or transport.

The key to understanding the nature of the resistance was the observation that these mutations result from the repaired DSBs via the HR and NHEJ pathways. Though cuts by Top1 generate single strand breaks (SSB), these breaks can develop into DSBs (34). In HR-dependent DSB repair, we observed loss of heterozygocity associated with copying of intact alleles to the allele with DSB, which allowed precise identification of the allele with DSB. Strikingly, more than 80% of DSBs took place in the alleles with mutations associated with the cancer nature of the parental cells, and accordingly, the HR repair led to restoration of the original "normal" alleles. As an example of such sequences, we found a large fraction of breaks in the polyC sequences of 30-40bp, which were present in the parental cancer cells (See Mutation landscape, Supplement Information). As a result of HR repair, these alleles were changed to alleles with interrupted polyC regions which were present in the reference human genome. Accordingly, at these sites, DSBs appear to require extended polyC, and thus an allele that has interruption of the polyC tract must have a lower probability of breaks. Therefore, reversion of extended polyC to the interrupted tract of polyC protects this site from further breaks, and thus contributes to the overall development of resistance to irinotecan (Fig. S4). This observation ultimately means that upon the first exposure to the drug, a fraction of sites with high probability of Top1-induced breaks will be reverted to sites with low

probability of breaks. Therefore, with each cycle of exposure to irinotecan fewer and fewer sites with high probability of DSBs will remain in the genome, which ultimately must increase the chance of survival. Indeed, we demonstrated that the number of DSB is reduced following cycles of exposure to irinotecan, and this effect was associated with lower probability of breaks rather than with more efficient DNA repair. Therefore, development of resistance to irinotecan involves acquiring a large number of mutations, each of which provides a fraction of the overall resistance, which gradually accumulate and are selected in the process of drug escalation.

This adaptation mechanism that associates with gradual reduction of the number of DNA breaks by Top1, suggests lower efficiency of Top1-dependent relaxation of DNA supercoils in selected clones. Possibly, the number of mutations that provide adaptation to irinotecan may be limited by the necessity to carry out the DNA relaxation activity. Alternatively, in these clones Top2 can take over essential DNA relaxation.

This work also illuminates novel aspects of function of Top1. Though it was reported that Top1 associates with active RNA polymerase to relieve DNA supercoils generated in the process of transcription(48–50), our data suggests that Top1 can also functions in a transcription-independent relief of supercoils since a majority of mutations was seen in heterochromatin (H3K9me3 and H3K27me3 patterns, see Fig. S2). The fact that there was a very large fraction of DSBs common between the three-independent irinotecan-resistant isolates indicates that there are preferable sites of breakage. This idea is reinforced by the finding that the mutation sites highly overlapped with sites of DNA cleavage by Top1 (48). Possibly, these mutation sites are preferable sites of binding of Top1 or some Top1 activating factors to DNA, suggesting that Top1 or its activators have a sequence binding preference. Alternatively, Top1 may bind anywhere on the DNA and when moving together with RNApol and possibly DNApol, stalls at these regions to increase the

probability of cuts. Another attractive possibility is that these repeat regions are preferable sites where SSBs are converted to DSBs.

The novel drug resistance mechanism may have interesting implications for understanding evolutionary processes. Indeed, it is possible that DSBs generated by Top1 take place predominantly at the sites of mutations that deviate from “normal” genomes. Accordingly, these DSBs can be repaired by the HR pathway, which will lead to the restoration of normal genome homozygocity. In other words, a stabilizing evolutionary selection may take place even in the absence of the selection pressure, simply as a result of the Top1 and HR repair function. Accordingly, the overall diversity of SNPs and InDels in the plurality of normal genomes has limitations that are shaped by the function of the Top1 and HR repair systems.

Materials and Methods

Cell culture and reagents. Cell lines were obtained from the ATCC. HCT116 cells (ATCC Cat# CCL-247, RRID:CVCL_0291) were cultured in McCoy's 5A Medium supplemented 10% FBS, 4mM L-glutamine (Cat#03-020-1B, BI-Biologicals), 2mM L-alanyl-L-glutamine (Cat#03-022-1B, BI-Biologicals), 1% penstrep (Cat#03-031-1B, BI-Biologicals), and were grown in a humidified incubator at 37°C and 5% CO2 (51). Cell cultures were checked for mycoplasma contamination routinely at six-week intervals by PCR using EZ-PCR™ Mycoplasma Detection Kit (Biological Industries, Israel) and confirmed to be mycoplasma-free. Irinotecan was purchased from Sigma-Aldrich (MO, USA).

Single cell line cloning. Single cell cloning was performed by limiting dilution method. Further using cloning discs individual clones were isolated (Merck, Ca#Z374431). Individual clones were further grown and stored (10% DMSO in FBS) as stocks in liquid nitrogen.

Virus Preparation for the barcoding library and shRNA screens. Lentivirus for barcoding library and shRNA screens were prepared using the corresponding plasmids along with virus coat packaging plasmids using Lipofectamine3000 infection reagent in 293LTX cells. Briefly, cells were passaged and grown at 80-90% confluency. For transfection, reagents were mixed in opti-MEM (Thermo Scientific, MA, USA) supplemented with 4mM glutamine and co-incubated for overnight. Next day media was changed with OptiMEM supplemented with 5% FBS, 4mM glutamine and kept further for 24 hours. Viruses were harvested using 0.45 µm filter and kept in -80C for further use. Upon infection with corresponding libraries, we chose the MOI of about 20%, so that on average each cell receives only one viral particle. After infection, cells carrying lentiviruses were selected with puromycin and further divided into groups for drug treatment in culture(52).

Cell barcoding. Cloned cell population was barcoded with 50 million (50M) library according to manufacturer protocols developed by Cellecta Inc (Cellecta CloneTracker 50M Lentiviral Barcode Library, RRID:SCR_021827)(42). Briefly, cells were infected with the barcoding lentiviruses, selected with puromycin and further divided into groups for drug treatment. After the treatments, cells were allowed to recover, genomic DNA was purified and barcodes were isolated by nested PCR, all samples were multiplexed for sequencing. Detailed procedure of PCR and primers details for 50M libraries is available in Supplementary datasheet "Data S1" (Tables S9 & S10).

Pooled shRNA Genetic Screen. For the shRNA screen, we used human decipher module 1 library (RRID:Addgene_28289)(45). Cells were infected with this pooled shRNA library with low MOI.

Cells were treated with irinotecan for 24 hours, recovered for 4 days, and collected for DNA isolation and further processing. Barcodes were isolated by nested PCR, sequenced and analyzed, according to the manufacturer's protocol.

Genomic DNA extraction and amplification of library barcodes. Isolation of genomic DNA from cultured cells was performed by Wizard genomic DNA isolation kit (Promega, WI, USA). Amplification of the barcodes was carried out by nested PCR. Detailed procedure of PCR and primer details for shRNA screen is available in Supplementary datasheet "Data S1" (Tables S11& S12). Briefly, 1-st PCR (PCR 1) was performed using Titanium Taq DNA Polymerase (# 639209, Takara Bio, CA, USA). Separation of the PCR products from primers and gel purification was done by QIAquick PCR & Gel Cleanup Kit (Qiagen, Germany). 2-nd PCR (PCR 2) was carried out using nested primers either generic or having unique sample barcodes. PCR 2 was performed using Phusion High-Fidelity PCR Master Mix (Thermo Scientific, MA, USA). Samples were multiplexed by adding an additional sample barcode during the second round of PCR. Samples were normalized individually, then pooled together, and purification of the PCR products was completed using AmpureXP magnetic beads (Beckman Coulter, CA, USA) following manufacturer protocols. Next, we sequenced the barcodes using Ion Torrent.

Analysis of the barcode data. We used a combination of custom-tailored applications to analyze sequencing reads along with the R programming language. Data were first checked for quality of reads through FastQC (v0.11.7, RRID:SCR_014583), further using barcode-splitter (v0.18.6, barcode splitter, RRID:SCR_021825) reads were demultiplexed based on sample barcodes (1 error as mismatch or deletion was allowed for sample barcodes while demultiplexing). Obtained FASTQ files were used to count the library barcodes by using python-based applications that were custom-made for this purpose. Quantification of the unique barcodes that were abundantly enriched or lost after treatment has been done via a python-based script (software version 3.10.0,

RRID:SCR_008394). For data cleaning and visualizations tidyverse-v1.0.0 (RRID:SCR_019186) and ggplot2-v3.3.3 (RRID:SCR_014601) of R packages were utilized.

Transcriptome analysis. RNA was extracted from cells using the RNeasy Mini kit (Cat#74104, Qiagen). Library preparation strategy (BGISEQ-500, RRID:SCR_017979) was adopted and performed by BGI, China. Briefly, mRNA molecules were purified from total RNA using oligo(dT) attached magnetic beads and fragmented into small pieces using fragmentation reagent after reaction a certain period at proper temperature. First-strand cDNA was generated using random hexamer-primed reverse transcription, followed by a second strand cDNA synthesis. The synthesized cDNA was subjected to end-repair and then was 3' adenylated. Adapters were ligated to the ends of these 3' adenylated cDNA fragments. This process amplified the cDNA fragments with adapters from previous step. PCR products were purified with Ampure XP Beads (AGENCOURT), and dissolved in EB solution. Library was validated on the Agilent Technologies 2100 bioanalyzer (2100 Bioanalyzer Instrument, RRID:SCR_018043). The double stranded PCR products were heat denatured and circularized by the splint oligo sequence. The single strand circle DNA (ssCir DNA) were formatted as the final library. The library was amplified with phi29 to make DNA nanoball (DNB) which had more than 300 copies of one molecular. The DNBs were load into the patterned nanoarray and single end 50 (pair end 100) bases reads were generated in the way of sequenced by synthesis. For in house pipeline was developed to analyze the data where reads were first trimmed and clipped for quality control in trim_galore (v0.5.0, RRID:SCR_011847) and checked for each sample using FastQC (v0.11.7, RRID:SCR_014583). Data was aligned by Hisat2 (v2.1.0, RRID:SCR_015530) using hg38, GRch38.97. High-quality reads were then imported into samtools (v1.9 using htslib 1.9, RRID:SCR_002105) for conversion into SAM files and later to BAM file. Gene-count summaries were generated with featureCounts (v1.6.3, RRID:SCR_012919): A numeric matrix of raw read counts was generated, with genes in

rows and samples in columns, and used for differential gene expression analysis with the Bioconductor, RRID:SCR_006442 (edgeR v3.32.1, RRID:SCR_012802 (54), Limma v3.46.0, RRID:SCR_010943) packages to calculate differential expression of genes. For normalization "voom" function was used, followed by eBayes and decideTests functions to compute differential expression of genes.

Human Whole Genome (HWG) sequencing and analysis. Isolated DNA was sent for 30X whole genome sequencing as a service by commercial provider Dante labs, Italy. After obtaining reads, data were first checked for quality of reads through FastQC (v0.11.7, RRID:SCR_014583), further using barcode-splitter (v0.18.6) reads were demultiplexed based on sample barcodes (1 error as mismatch or deletion was allowed for sample barcodes while demultiplexing). Reads were aligned to reference genome hg38 with BWA-MEM (RRID:SCR_010910)(55) default settings. Aligned SAM files were converted to BAM using samtools (ver1.9, RRID:SCR_002105). Obtained BAM files were used to call variants using Genome Analysis Tool Kit (GATK v4.1.8.0, RRID:SCR_001876)(56) from Broad's Institute, we called for combined genomic variants as Variant Call Factor (VCF) file as output. In above analysis we followed strictly the best practices guidelines of GATK(57). All further analysis was done using either suitable R packages (tidyverse-v1.0.0 (RRID:SCR_019186) and ggplot2-v3.3.3(RRID:SCR_014601), karyoploteR v1.18.0(karyoploteR, RRID:SCR_021824)(58) for data, we also used SnpEff (v3.11, RRID:SCR_005191)(59) program for annotation of genes and bedtools v2.27.1(RRID:SCR_006646)(60) to match obtained variants to available bed files from UCSC(UCSC Cancer Genomics Browser, RRID:SCR_011796)(46), GEO (Gene Expression Omnibus (GEO), RRID:SCR_005012) datasets and other relevant databases. To extract data from metadata file and for comparative analysis we used Microsoft excel (RRID:SCR_016137) as tool. Raw data files and associated common mutation analysis in form of VCF file is available at SRA

(NCBI Sequence Read Archive (SRA), RRID:SCR_004891) under accession number (PRJNA738674).

Assessment of cluster mutations in resistant mutants. For estimating the number of mutations present in close proximity or whether it is clustered, we took 100bp small windows and calculated whether the mutations were present in close proximity in these small intervals. We used custom written python codes to first, separate the whole genome into small bins of 100bp respectively. Further we defined common mutation's genomic locations to it and calculated the conditions that is how many of these mutations fall within 100bp window. The output was generated as text file where genomic locations and number of mutations in 100bp window were calculated.

Classification of repeats and computing mutations in low complexity regions of genome. For classifying the nature of repeats where these mutations have been accumulated, we used RepeatMasker v4.1.0 (RepeatMasker, RRID:SCR_012954, database Dfam_3.1 (RRID:SCR_021168) and rmblastn version 2.9.0)(61) that screens DNA sequences for known repeats and low complexity DNA sequences. Reference sequence hg38 was used for this purpose. Common mutations (total-46099) were aligned to reference database according to their occurrences based on genomic locations. Program defines the nature of the sequence where these mutations have been accumulated. The output of the program is a detailed annotation of repeats that are present in the query sequence as well as a modified version of the query sequence in which all the annotated repeats are classified. Sequence comparisons were performed on Unix environment using cross-match, an efficient implementation of the Smith-Waterman-Gotoh algorithm developed by Phil Green, or by WU-Blast developed by Warren Gish(61). Output file is matrix of repeated elements found in query sequence compared to reference libraries reported as percentage in the tabulated form (62).

Assessment of satellites and simple repeats. In-depth analysis was done to identify the nature of satellites and its classification using RepeatMasker v4.1.0 (RepeatMasker, RRID:SCR_012954)(61) advanced options wherein we investigated the type of satellites and simple repeats. Firstly, fasta sequences for the genomic locations for common mutations were fetched using bedtools "getfasta" algorithm (BEDTools, RRID:SCR_006646)(60). It was further analyzed by RepeatMasker v4.1.0(61) to identify the repeated elements. Next, "fasta.out" was used to further investigate and quantify the classification of the satellites type. While "fasta.tbl" was used to compute overall percent representation of the repeated elements. Additionally, we used Tandem repeat finder (RRID:SCR_005659)(63) and UCSC microsatellite track to investigate and compute the microsatellites and simple repeats.

Double strand break quantification. Estimation of double strand break occurring in the system of wild type cell population and resistant lines once adapted to drug was done using (a) γH2AX assay(64) and (b) Comet assay(65).

γ-H2AX assay. Briefly, cells were plated in 24 well plates (Cellvis, Cat#P24-1.5H-N) at 30% confluency. It was treated with 4nM Irinotecan for 24 hours. Drug was removed from the plates and cells were allowed to recover back. This treatment cycle was repeated for 5 times after each recovery. Once cells were resistant to 4nM irinotecan termed as "T5_resistant population", was used for further assay. For quantification of foci, cells were treated parallelly both wild type and T5_resistant cells with respective controls. Cells were washed with 1xPBS and fixed with 0.2% formaldehyde. Permeabilization was done by 0.2% Triton X-100 in PBS for 10 minutes at room temperature. Then cells were blocked with Bovine serum albumin (BSA, 5% w/v in Phosphate buffered saline with 0.05% tween-20: PBST) for 1 hour. Washed 3 times with PBST. Incubation with primary antibody (Phospho-Histone H2A.X (Ser139) Antibody, Cell Signaling Technology Cat# 2577, RRID:AB_2118010) was done overnight at 4^0C. Cells were washed with 1xPBST five

times followed by incubation with secondary antibody (Goat Anti-Rabbit IgG H&L, Alexa Fluor® 488, Abcam Cat# ab150077, RRID:AB_2630356) for 1 hour. Antibody was removed and washing was done for five times to remove non-specific antibody binding. Imaging was performed using Hermes Wiscon Imaging System (IDEA Bio-Medical Ltd., WiScan Hermes High Content Imaging System, RRID:SCR_021786) and image analysis was performed using inbuilt software package system (Athena Wisoft, Ver1.0.10). Statistics calculation was performed with GraphPad Prism version 9.0.0 (GraphPad Prism, RRID:SCR_002798), California USA. The significance of differences was determined using unpaired Welch's correction, two-tailed t-test (*$P < 0.0332$, **$P < 0.0021$, ***$P < 0.0002$, ****$P < 0.0001$).

Single cell gel electrophoresis (SCGE). To estimate the change in break of DNA strands before and after treatment in normal cells and resistant cell population, a well-known method of single cell gel electrophoresis (SCGE) or COMET assay was performed(66). Cells were taken and embedded in 1% agarose on a microscope slide are lysed with detergent and high salt.

Glass slides were precoated with 1% agarose. Cells were treated with irinotecan or left untreated to serve as control. Preparation of sample was done by scrapping cells gently with 0.05% trypsin. Washed 3 times with ice cold 1X-PBS and roughly $0.1*10^6$ cells/ml was taken. 50uL of cell suspension and 50uL of 1% low molten agarose kept at 37^0C was mixed together. 75uL was used to make bubbles on the slide and was left to solidify at 4^0C until it formed clear ring. Lysis was performed by placing the sample slides in lysis buffer [(Nacl (2.5M), EDTA-pH 10 (100mM), Tris-Base pH 10 (10mM), Triton X100 (1% freshly added before use), buffer pH was maintained at 10 before adding Triton-X100] overnight at 4^0C. Unwinding was performed by rinsing the slides with fresh water to remove salts and detergents. Slides were immerged in unwinding buffer [(NaOH(300mM), EDTA (1mM), buffer pH 13)] for 1 hour at 4^0C. After unwinding, samples were run in electrophoresis alkali buffer [(NaOH (12g/L), EDTA (500mM pH 8)], slides were kept in

electrophoresis tank and buffer was poured to cover the slides. Running was performed at 22V(constant), 400mA(constant) for 40 minutes. Neutralization was performed by dipping the slides in tris buffer (0.4M, pH 7.5) for 5 minutes. Slides were then immerged in 70% ethanol for 15 minutes, air dried for 30 minutes, Staining was performed by incubating in DAPI (1ug/ul, 1:5000 dilution). Washed with ice cold water and air dried in dark. Imaging was performed using Olympus IX81 Inverted Fluorescence Automated Live Cell Microscope (18MP CMOS USB camera) with associated, in-built software package (Olympus IX 81 Inverted Fluorescence Automated Live Cell Microscope, RRID:SCR_020341). Quantification was performed using OpenComet v1.3.1(OpenComet, RRID:SCR_021826)(67). Statistics calculation was performed with GraphPad Prism version 9.0.0 (GraphPad Prism, RRID:SCR_002798), California USA. The significance of differences was determined using unpaired Welch's correction, two-tailed t-test (*$P < 0.0332$, **$P < 0.0021$, ***$P < 0.0002$, ****$P < 0.0001$).

Hypergeometric distribution: To predict the significance of overlap between triple mutations and Top1-cleavage sites we used R language tool (see code availability section for applicable codes). The problem of DNA sites overlap is described by a hypergeometric distribution where one list defines the number of cleavage sites and the other list defines the number of mutations. Assume the total genome size is n, the number of points in the first list is a, and the number of points in the second list is b. If the intersection between the two lists is t, the probability density of seeing t can be calculated as: *Probability of occurrence=dhyper (t, a, n – a, b)*. Based on the existing data, $b < a$. So, the largest possible value for t is b. Therefore, the p-value of seeing intersection t is: *sum(dhyper(t:b, a, n - a, b))*. Taking, top1 cleavage sites (a) cover $2.5*10^7$ base pairs in the genome ($2*10^5$ sites X 50bp per peak), mutations (b)=46099, total genome (n)=$3*10^9$ bp, t=9000 overlap out of 46099. Sum of overlap was computed to be absolute zero with null probability.

Author Contributions: Conceptualization: Santosh Kumar, Michael Sherman; Methodology and pipeline design: Santosh Kumar; Investigation: Santosh Kumar, Shivani Patel, Julia Yaglom; Bioinformatics analysis: Santosh Kumar, Valid Gahramanov, Shivani Patel; Suggestions in data analysis: Ivan Alexandrov, Gabi Gerlitz, Mali Salmon-Divon, Lukasz Kaczmarczyk; Supervision: Michael Sherman; Writing-original draft: Michael Sherman, Santosh Kumar; Organizing manuscript and visualization of data: Santosh Kumar; Review and approval: All the contributors.

Competing Interest Statement: The authors declare no competing interests.

Funding:

National Institutes of Health grant RA1800000163

Israel Science Foundation Grant: ISF-1444/18, ISF-2465/18.

Data and materials availability: Data such as raw FASTQ files for human whole genome sequencing has been submitted to publicly accessible database SRA under accession number PRJNA738674. Processed VCF showing identical mutations among all resistant subclones compared to parental line (available in supplement information Table S13) is included in this paper. All list genes from transcriptome analysis and shRNA screens are available in this paper. Raw data and files after post analysis for transcriptome analysis is deposited and is available at GEO data base with accession number GSE189366. Analyzed data is presented in supplement table at relevant section.

Data referred from literature: GSE57628, mapping of Top-1 binding and cleavage sites reported for HCT116 cells, Summary of epigenome ENCSR309SGV (ENCODE database that present the summary of various methylation signatures such as H3K9me3 and H3K27me3 for genome of HCT116).

Code availability: All the codes essentially used in analysis of sequencing data, either human whole genome, transcriptome analysis or shRNA screens has been submitted to public domain and can be accessed through Github (https://github.com/santoshbiowarrior333/Irinotecan_resistance).

References

Alfarouk KO, Stock C-M, Taylor S, Walsh M, Muddathir AK, Verduzco D, Bashir AHH, Mohammed OY, Elhassan GO, Harguindey S, et al. 2015. Resistance to cancer chemotherapy: failure in drug response from ADME to P-gp. *Cancer Cell Int* **15**. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4502609/ (Accessed December 8, 2020).

Baranello L, Wojtowicz D, Cui K, Devaiah BN, Chung H-J, Chan-Salis KY, Guha R, Wilson K, Zhang X, Zhang H, et al. 2016. RNA Polymerase II Regulates Topoisomerase 1 Activity to Favor Efficient Transcription. *Cell* **165**: 357–371. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4826470/ (Accessed September 8, 2021).

Benson G. 1999. Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research* **27**: 573–580. https://doi.org/10.1093/nar/27.2.573 (Accessed August 31, 2021).

Cahill DP, Kinzler KW, Vogelstein B, Lengauer C. 1999. Genetic instability and darwinian selection in tumours. *Trends in Cell Biology* **9**: M57–M60. https://linkinghub.elsevier.com/retrieve/pii/S096289249901661X (Accessed September 4, 2021).

Cairns J. 1975. Mutation selection and the natural history of cancer. *Nature* **255**: 197–200. http://www.nature.com/articles/255197a0 (Accessed September 3, 2021).

Cairns J, Overbaugh J, Miller S. 1988. The origin of mutants. *Nature* **335**: 142–145. https://www.nature.com/articles/335142a0 (Accessed June 19, 2021).

Cellecta. CloneTracker™ 50M Lentiviral Barcode Library. *Cellecta*. https://cellecta.com/products/clonetracker-50m-lentiviral-barcode-library (Accessed June 19, 2021a).

Cellecta. DECIPHER Pooled shRNA Library Viability Screen. *Cellecta*. https://cellecta.com/pages/decipher-pooled-shrna-library-viability-screen (Accessed June 19, 2021b).

Champoux JJ. 2001. DNA Topoisomerases: Structure, Function, and Mechanism. *Annu Rev Biochem* **70**: 369–413. https://www.annualreviews.org/doi/10.1146/annurev.biochem.70.1.369 (Accessed June 19, 2021).

Chen Y, Lun AT, McCarthy DJ, Ritchie ME, Phipson B, Hu Y, Zhou X, Robinson MD, Smyth GK. 2021. *edgeR: Empirical Analysis of Digital Gene Expression Data in R*. Bioconductor version: Release (3.13) https://bioconductor.org/packages/edgeR/ (Accessed August 3, 2021).

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin)* **6**: 80–92.

Demerec M. 1937. Frequency of Spontaneous Mutations in Certain Stocks of Drosophila Melanogaster. *Genetics* **22**: 469–478. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1208982/ (Accessed September 4, 2021).

DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C, Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* **43**: 491–498. http://www.nature.com/articles/ng.806 (Accessed August 7, 2021).

Drake JW. 1999. The Distribution of Rates of Spontaneous Mutation over Viruses, Prokaryotes, and Eukaryotes. *Annals of the New York Academy of Sciences* **870**: 100–107. https://nyaspubs.onlinelibrary.wiley.com/doi/abs/10.1111/j.1749-6632.1999.tb08870.x (Accessed August 7, 2021).

Drake JW, Charlesworth B, Charlesworth D, Crow JF. 1998. Rates of Spontaneous Mutation. *Genetics* **148**: 1667. http://www.genetics.org/content/148/4/1667.abstract.

Duesberg P, Stindl R, Hehlmann R. 2000. Explaining the high mutation rates of cancer cells to drug and multidrug resistance by chromosome reassortments that are catalyzed by aneuploidy. *Proc Natl Acad Sci U S A* **97**: 14295–14300. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC18912/ (Accessed June 19, 2021).

Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, Smit AF. 2019. *RepeatModeler2: automated genomic discovery of transposable element families*. Genomics http://biorxiv.org/lookup/doi/10.1101/856591 (Accessed December 22, 2020).

Gel B, Serra E. 2017. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data ed. J. Hancock. *Bioinformatics* **33**: 3088–3090. https://academic.oup.com/bioinformatics/article/33/19/3088/3857734 (Accessed September 4, 2021).

Geraldine A. Van der Auwera, Brian D. O'Connor. 2020. *Genomics in the Cloud*. First.

Gottesman MM. 2002. Mechanisms of cancer drug resistance. *Annu Rev Med* **53**: 615–627.

Gyori BM, Venkatachalam G, Thiagarajan PS, Hsu D, Clement M-V. 2014. OpenComet: an automated tool for comet assay image analysis. *Redox Biol* **2**: 457–465.

Hall BG. 1988. Adaptive evolution that requires multiple spontaneous mutations. I. Mutations involving an insertion sequence. *Genetics* **120**: 887–897. https://doi.org/10.1093/genetics/120.4.887 (Accessed June 19, 2021).

Hammond WA, Swaika A, Mody K. 2016. Pharmacologic resistance in colorectal cancer: a review. *Ther Adv Med Oncol* **8**: 57–84. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4699262/ (Accessed December 30, 2020).

Hegedüs É, Kókai E, Nánási P, Imre L, Halász L, Jossé R, Antunovics Z, Webb MR, El Hage A, Pommier Y, et al. 2018. Endogenous single-strand DNA breaks at RNA polymerase II promoters in Saccharomyces cerevisiae. *Nucleic Acids Research* **46**: 10649–10668. https://doi.org/10.1093/nar/gky743 (Accessed September 8, 2021).

Hu T, Li Z, Gao C-Y, Cho CH. 2016. Mechanisms of drug resistance in colon cancer and its therapeutic strategies. *WJG* **22**: 6876. http://www.wjgnet.com/1007-9327/full/v22/i30/6876.htm (Accessed December 3, 2020).

Ivashkevich AN, Martin OA, Smith AJ, Redon CE, Bonner WM, Martin RF, Lobachevsky PN. 2011. γH2AX foci as a measure of DNA damage: a computational approach to automatic analysis. *Mutat Res* **711**: 49–60. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3101310/ (Accessed August 7, 2021).

Jackson AL, Loeb LA. 1998. On the origin of multiple mutations in human cancers. *Seminars in Cancer Biology* **8**: 421–429. https://linkinghub.elsevier.com/retrieve/pii/S1044579X98901134 (Accessed August 7, 2021).

Jensen NF, Stenvang J, Beck MK, Hanáková B, Belling KC, Do KN, Viuff B, Nygård SB, Gupta R, Rasmussen MH, et al. 2015. Establishment and characterization of models of chemotherapy resistance in colorectal cancer: Towards a predictive signature of chemoresistance. *Mol Oncol* **9**: 1169–1185. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5528766/ (Accessed November 6, 2021).

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler and D. 2002. The Human Genome Browser at UCSC. *Genome Res* **12**: 996–1006. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC186604/ (Accessed August 10, 2021).

Koh S-B, Mascalchi P, Rodriguez E, Lin Y, Jodrell DI, Richards FM, Lyons SK. 2017. A quantitative FastFUCCI assay defines cell cycle dynamics at a single-cell level. *J Cell Sci* **130**: 512–520.

Kohler SW, Provost GS, Fieck A, Kretz PL, Bullock WO, Sorge JA, Putman DL, Short JM. 1991. Spectra of spontaneous and mutagen-induced mutations in the lacI gene in

transgenic mice. *Proceedings of the National Academy of Sciences* **88**: 7958–7962. http://www.pnas.org/cgi/doi/10.1073/pnas.88.18.7958 (Accessed August 7, 2021).

Kumar S, Subramanian S. 2002. Mutation rates in mammalian genomes. *Proc Natl Acad Sci U S A* **99**: 803–808. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC117386/ (Accessed August 7, 2021).

Lee D-W, Han S-W, Cha Y, Bae JM, Kim H-P, Lyu J, Han H, Kim H, Jang H, Bang D, et al. 2017. Association between mutations of critical pathway genes and survival outcomes according to the tumor location in colorectal cancer. *Cancer* **123**: 3513–3523.

Li H, Durbin R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**: 589–595.

Libbrecht MW, Rodriguez OL, Weng Z, Bilmes JA, Hoffman MM, Noble WS. 2019. A unified encyclopedia of human functional DNA elements through fully automated annotation of 164 human cell types. *Genome Biol* **20**: 180. https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1784-2 (Accessed November 16, 2021).

Liu Y-Q, Li W-Q, Morris-Natschke SL, Qian K, Yang L, Zhu G-X, Wu X-B, Chen A-L, Zhang S-Y, Nan X, et al. 2015. Perspectives on Biologically Active Camptothecin Derivatives: BIOLOGICALLY ACTIVE CAMPTOTHECIN DERIVATIVES. *Med Res Rev* **35**: 753–789. https://onlinelibrary.wiley.com/doi/10.1002/med.21342 (Accessed September 4, 2021).

Loeb LA, Loeb KR, Anderson JP. 2003. Multiple mutations and cancer. *Proceedings of the National Academy of Sciences* **100**: 776–781. http://www.pnas.org/cgi/doi/10.1073/pnas.0334858100 (Accessed September 4, 2021).

Luria SE, Delbrück M. 1943. Mutations of Bacteria from Virus Sensitivity to Virus Resistance. *Genetics* **28**: 491–511. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1209226/ (Accessed June 19, 2021).

Meriin AB, Narayanan A, Meng L, Alexandrov I, Varelas X, Cissé II, Sherman MY. 2018. Hsp70–Bag3 complex is a hub for proteotoxicity-induced signaling that controls protein aggregation. *PNAS* **115**: E7043–E7052. https://www.pnas.org/content/115/30/E7043 (Accessed April 5, 2021).

Neri M, Milazzo D, Ugolini D, Milic M, Campolongo A, Pasqualetti P, Bonassi S. 2015. Worldwide interest in the comet assay: a bibliometric study. *Mutagenesis* **30**: 155–163. https://doi.org/10.1093/mutage/geu061 (Accessed August 7, 2021).

Pang B, Qiao X, Janssen L, Velds A, Groothuis T, Kerkhoven R, Nieuwland M, Ovaa H, Rottenberg S, van Tellingen O, et al. 2013. Drug-induced histone eviction from open

chromatin contributes to the chemotherapeutic effects of doxorubicin. *Nat Commun* **4**: 1908. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3674280/ (Accessed November 9, 2019).

Patnaik S, Anupriya. 2019. Drugs Targeting Epigenetic Modifications and Plausible Therapeutic Strategies Against Colorectal Cancer. *Front Pharmacol* **10**. https://www.frontiersin.org/articles/10.3389/fphar.2019.00588/full (Accessed October 31, 2019).

Phipps AI, Limburg PJ, Baron JA, Burnett-Hartman AN, Weisenberger DJ, Laird PW, Sinicrope FA, Rosty C, Buchanan DD, Potter JD, et al. 2015. Association between molecular subtypes of colorectal cancer and patient survival. *Gastroenterology* **148**: 77-87.e2.

Piwko W, Buser R, Peter M. 2011. Rescuing stalled replication forks: MMS22L-TONSL, a novel complex for DNA replication fork repair in human cells. *Cell Cycle* **10**: 1703–1705.

Pommier Y. 2013. Drugging Topoisomerases: Lessons and Challenges. *ACS Chem Biol* **8**: 82–95. https://pubs.acs.org/doi/10.1021/cb300648v (Accessed September 4, 2021).

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**: 841–842. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2832824/ (Accessed August 7, 2021).

Raskov H, Søby JH, Troelsen J, Bojesen RD, Gögenur I. 2019. Driver Gene Mutations and Epigenetics in Colorectal Cancer. *Ann Surg*.

Russo M, Crisafulli G, Sogari A, Reilly NM, Arena S, Lamba S, Bartolini A, Amodio V, Magrì A, Novara L, et al. 2019. Adaptive mutability of colorectal cancers in response to targeted therapies. *Science* **366**: 1473–1480. https://science.sciencemag.org/content/366/6472/1473 (Accessed June 19, 2021).

Sadanandam A, Lyssiotis CA, Homicsko K, Collisson EA, Gibb WJ, Wullschleger S, Ostos LCG, Lannon WA, Grotzinger C, Del Rio M, et al. 2013. A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nat Med* **19**: 619–625.

Saredi G, Huang H, Hammond CM, Alabert C, Bekker-Jensen S, Forne I, Reverón-Gómez N, Foster BM, Mlejnkova L, Bartke T, et al. 2016. H4K20me0 marks post-replicative chromatin and recruits the TONSL–MMS22L DNA repair complex. *Nature* **534**: 714–718.

Singh NP, McCoy MT, Tice RR, Schneider EL. 1988. A simple technique for quantitation of low levels of DNA damage in individual cells. *Experimental Cell Research* **175**: 184–191. https://linkinghub.elsevier.com/retrieve/pii/0014482788902650 (Accessed September 3, 2021).

Singh S, Szlachta K, Manukyan A, Raimer HM, Dinda M, Bekiranov S, Wang Y-H. 2020. Pausing sites of RNA polymerase II on actively transcribed genes are enriched in DNA double-stranded breaks. *J Biol Chem* **295**: 3990–4000. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7086017/ (Accessed September 8, 2021).

Sinicrope FA, Okamoto K, Kasi PM, Kawakami H. 2016. Molecular Biomarkers in the Personalized Treatment of Colorectal Cancer. *Clin Gastroenterol Hepatol* **14**: 651–658.

Smit, AFA, Hubley, R & Green, P. *RepeatMasker Open-4.0,2013-2015<http://www.repeatmasker.org>*. <http://www.repeatmasker.org>.

Sprouffske K, Aguilar-Rodríguez J, Sniegowski P, Wagner A. 2018. High mutation rates limit evolutionary adaptation in Escherichia coli. *PLOS Genetics* **14**: e1007324. https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1007324 (Accessed October 11, 2021).

Starrett JH, Guernet AA, Cuomo ME, Poels KE, van Alderwerelt van Rosenburgh IK, Nagelberg A, Farnsworth D, Price KS, Khan H, Ashtekar KD, et al. 2020. Drug sensitivity and allele-specificity of first-line osimertinib resistance EGFR mutations. *Cancer Res* **80**: 2017–2030. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7392201/ (Accessed November 21, 2021).

Vanhoefer U, Harstrick A, Achterrath W, Cao S, Seeber S, Rustum YM. 2001. Irinotecan in the Treatment of Colorectal Cancer: Clinical Overview. *JCO* **19**: 1501–1518. http://ascopubs.org/doi/10.1200/JCO.2001.19.5.1501 (Accessed September 4, 2021).

Venook AP, Niedzwiecki D, Lenz H-J, Innocenti F, Fruth B, Meyerhardt JA, Schrag D, Greene C, O'Neil BH, Atkins JN, et al. 2017. Effect of First-Line Chemotherapy Combined With Cetuximab or Bevacizumab on Overall Survival in Patients With KRAS Wild-Type Advanced or Metastatic Colorectal Cancer: A Randomized Clinical Trial. *JAMA* **317**: 2392–2401.

Vitiello PP, Martini G, Mele L, Giunta EF, De Falco V, Ciardiello D, Belli V, Cardone C, Matrone N, Poliero L, et al. 2021. Vulnerability to low-dose combination of irinotecan and niraparib in ATM-mutated colorectal cancer. *J Exp Clin Cancer Res* **40**: 15. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7789007/ (Accessed November 6, 2021).

Webster LT. 1937. INHERITANCE OF RESISTANCE OF MICE TO ENTERIC BACTERIAL AND NEUROTROPIC VIRUS INFECTIONS. *J Exp Med* **65**: 261–286. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2133487/ (Accessed June 19, 2021).

Yaglom JA, Wang Y, Li A, Li Z, Monti S, Alexandrov I, Lu X, Sherman MY. 2018. Cancer cell responses to Hsp70 inhibitor JG-98: Comparison with Hsp90 inhibitors and finding synergistic

drug combinations. *Sci Rep* **8**: 3010. https://www.nature.com/articles/s41598-017-14900-0 (Accessed September 4, 2021).

Figures and Tables

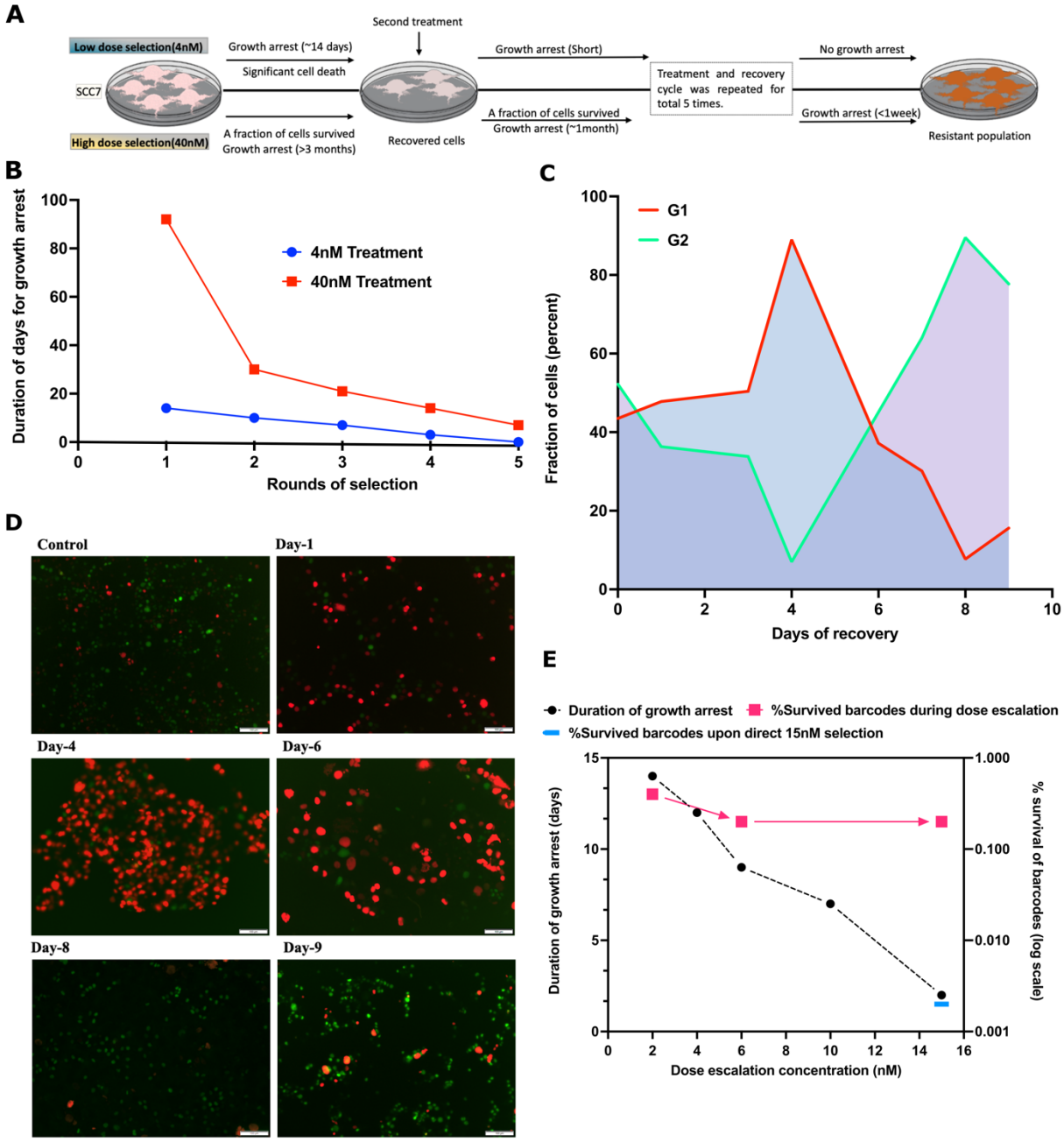


Figure 1. Adaptation to irinotecan associates with reduction in duration of growth arrest: (A) Graphical sketch of the experiment plan, showing low dose selection (upper panel), high dose selection (lower panel). **(B)** clone of HCT116 cells was exposed to the same dose of irinotecan (either 4nM or 40nM for 24 hours) multiple times, and periods of the cell cycle arrest following

exposures were measured. **(C, D)** Simultaneous recovery of irinotecan-treated cells from the cell cycle arrest. Cells were infected with cell cycle reporter plasmid (see Materials and Methods). Cells were treated with 4nM irinotecan, and following the cell cycle arrest (Day 1 and 4) allowed to recover (Day 8). Imaging was performed at the indicated days following irinotecan exposure using TexRed (red) and FITC (green) channels. Green fluorescence represents G2, Red fluorescence represents G1 and overlapping of red-green shows S phase. Quantification of data for fraction of G1, G2 and S phase cells during recovery from 4nM irinotecan treatment was quantified by imageJ. **(E)** Cells were exposed to rounds of treatments with 2, 4, 6, 10 and 15nM irinotecan and the periods of recovery were measured. On the same graph survival of barcodes after these treatments is shown. Cells were barcoded using the 50 million lentiviral barcoding libraries (Cellecta), and cells recovered after the treatments were collected, barcodes isolated, sequenced by IonTorrent and analyzed. Quantification of barcode survival is represented on right Y axis. 0.4% of total barcodes survived 2nM dose compared to untreated control (100%). Following further 6nM and 15nM 0.2% of initial barcodes survived, i.e. approximately 50% of barcodes that survived the 2nM dose. Also shown the fraction of barcodes that survived 15nM dose without preadaptation to lower doses (0.002%).

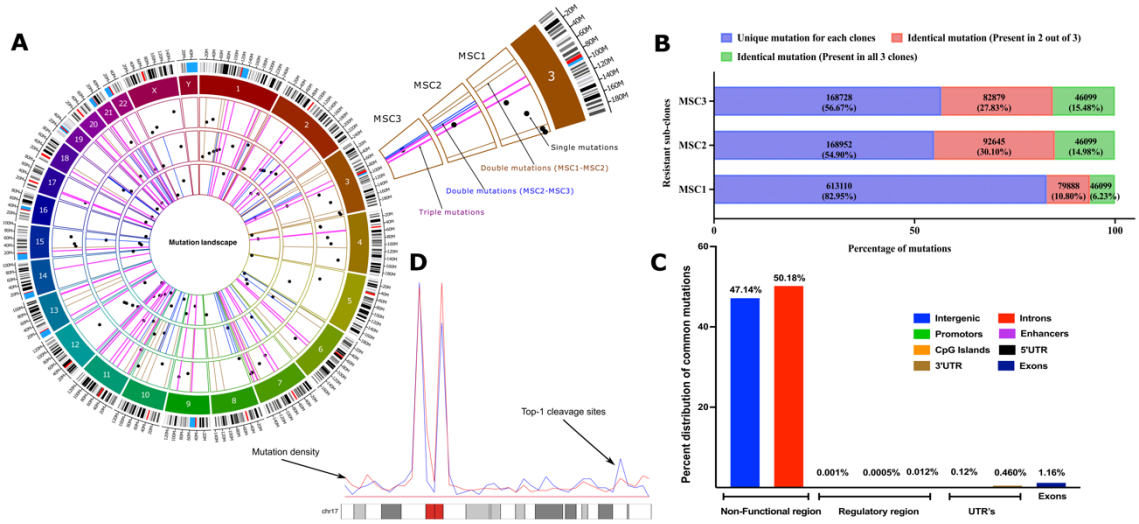


Figure 2. Irinotecan resistant subclones harbors identical mutations in genome at non-coding regions: **(A)** Circos plot to represent the category of mutations based on their occurrence in three independent clones. Three circles inside the chromatogram notations represent three individual selected resistant clones (MSC1, MSC2 and MSC3 from outside to inside). Dots (black) inside the respective circles represent mutations that are unique in the clones and not overlapping to others. Bars (brown) represent the overlapping mutations in clone MSC1 and MSC2. Bars (blue) represent the overlapping mutations in clone MSC2 and MSC3. Bars (pink) represent mutations that are overlapping in all three clones. To plot, only 50 mutations of each category was chosen randomly and are plotted for visualization purpose) to their genomic locations. Plot was prepared using commercial license from OMGenomics (Circa, RRID:SCR_021828) application. **(B)** Graph represent the total quantification of overlapping mutations in the irinotecan-resistant clones that emerge in the adaptation process. Stack graph showing the overlapping mutations among resistant sub-clones. Mutations in the resistant clones that differ them from SSC1 were compared to each other. These mutations represent three categories; (blue) mutations that are unique to each clone and not found in other clones, (red) mutations that are identical in two clones, (green) mutations that are identical in all three clones, **(C)** Categorical distribution of mutations to the non-functional,

regulatory and UTR's region of genome computed from annotation using SnpEff, **(D)** An example representation of chromosome-17 showing overlapped mutations density (blue line) and Top-1 cleavage sites (red line).

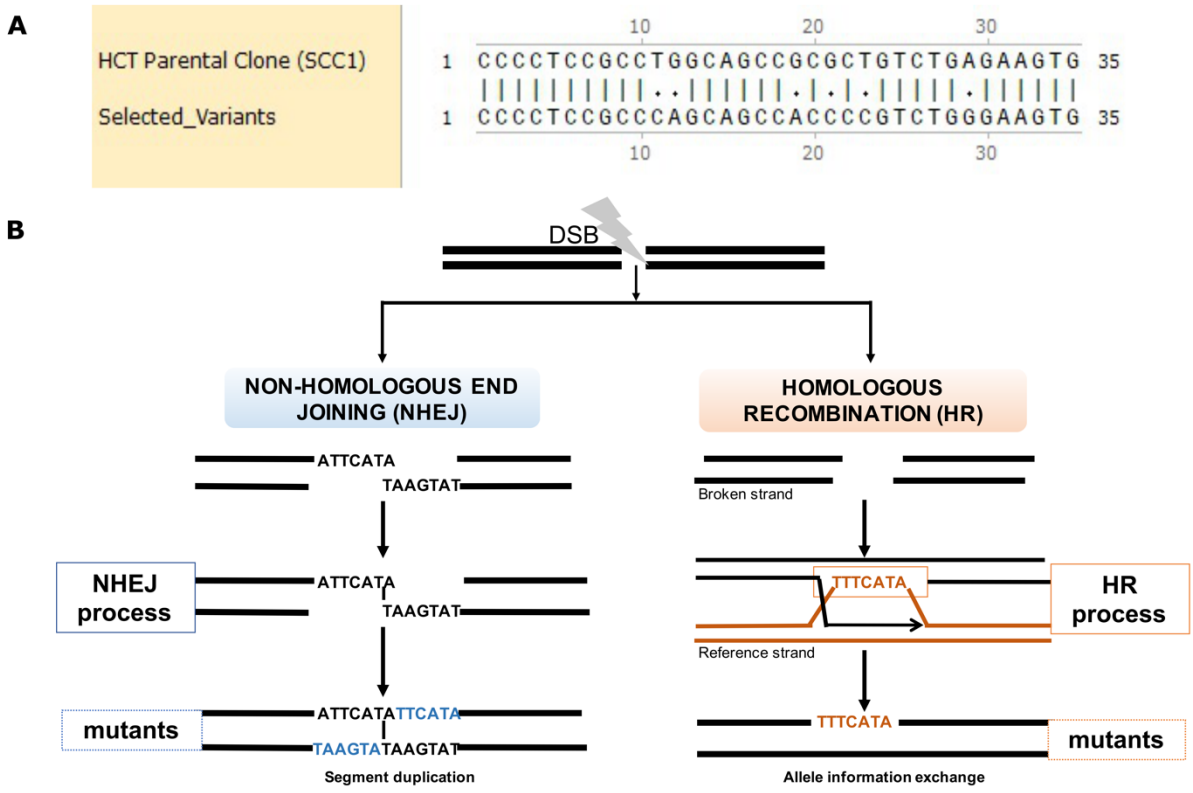


Figure 3. Mutations emerges in clustering patterns resulting from HR and NHEJ repair: (A) Example for clustering pattern of mutations compared to parental SSC1. **(B)** Sketch diagram of the mechanisms of DSB repair that leave specific mutation signatures. Real mutations found in all three resistant clones that result from NHEJ and HR repair are shown as examples. NHEJ repair both DNA ends are either trimmed, which results in InDels, or alternatively are paired via short homology region, which results in duplication of strands on both strands (left panel). In case of HR, an unbroken allele at the homologous chromosome serves as template that is copied to the broken strand.

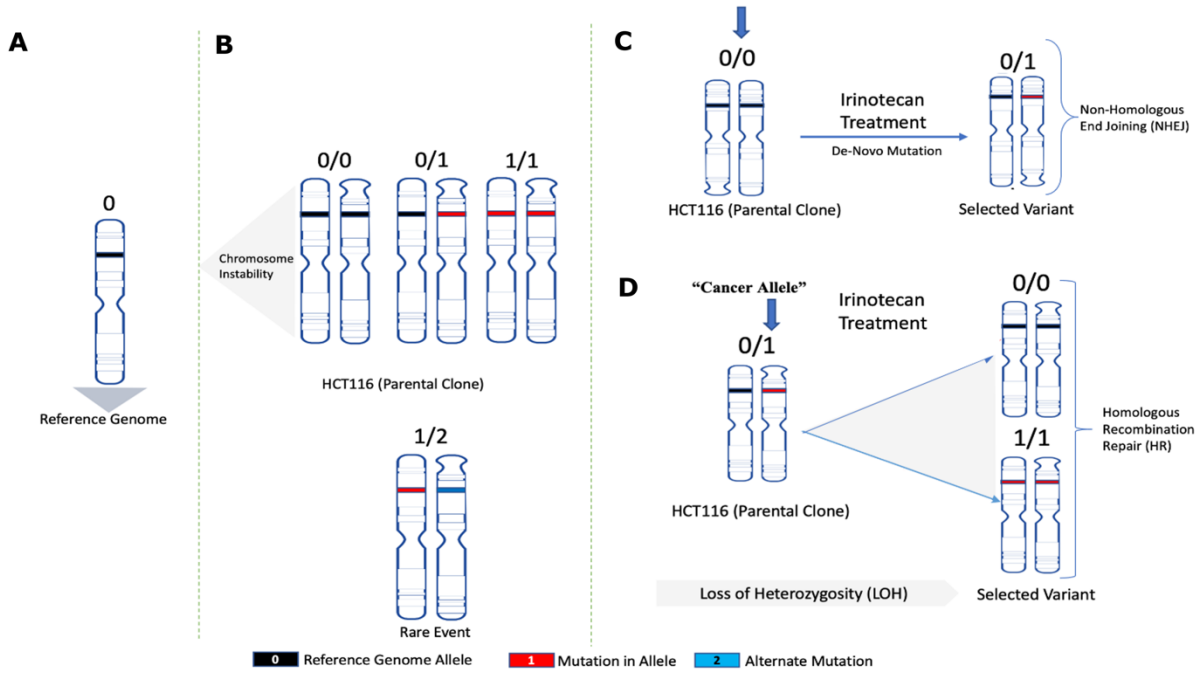


Figure 4. A framework for analyses of the role of mutations in adaptation to irinotecan: Schematic representation for **(A)** Normal reference genome allele, **(B)** Alleles found in parental SSC1 clone, **(C)** De-novo mutations in MSC clones after NHEJ repair, **(D)** Allele change upon loss of heterozygosity after HR repair (Loss of heterozygosity herein is referred to as the change of mutated heterozygous alleles reversed to homozygous reference in selected variants).

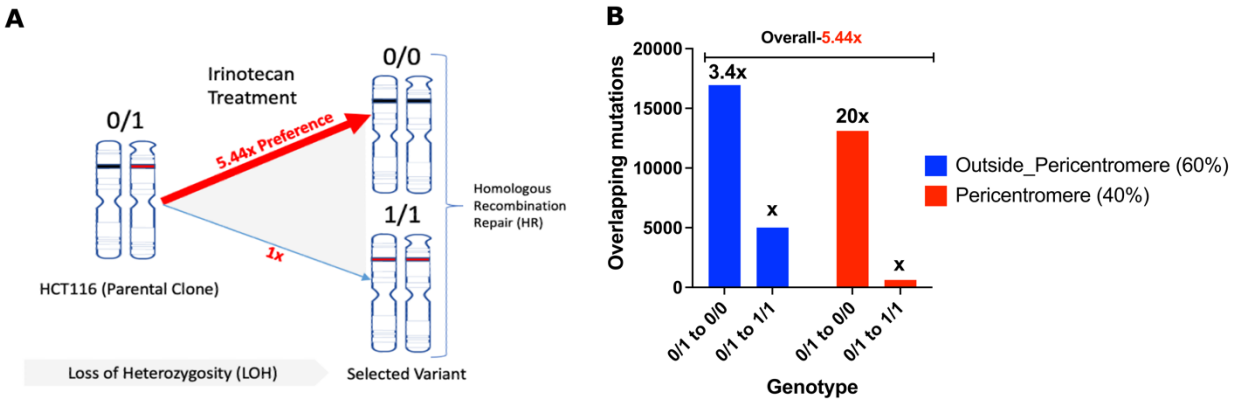


Figure 5. Cancer alleles are more susceptible to breaks compared to reference genome: (A) HR repair generated mutations revert to the reference genome allele 5.44 times more often than to the cancer allele. Therefore, cancer alleles are broken by Top1 5.44 times more often that reference genome alleles. Accordingly reverting to the reference genome alleles protects from DSBs upon consequent exposures to irinotecan. **(B)** Prevalence of breaks in cancer alleles compared to reference genome alleles is lower in chromosome arms than in the pericentromeric region.

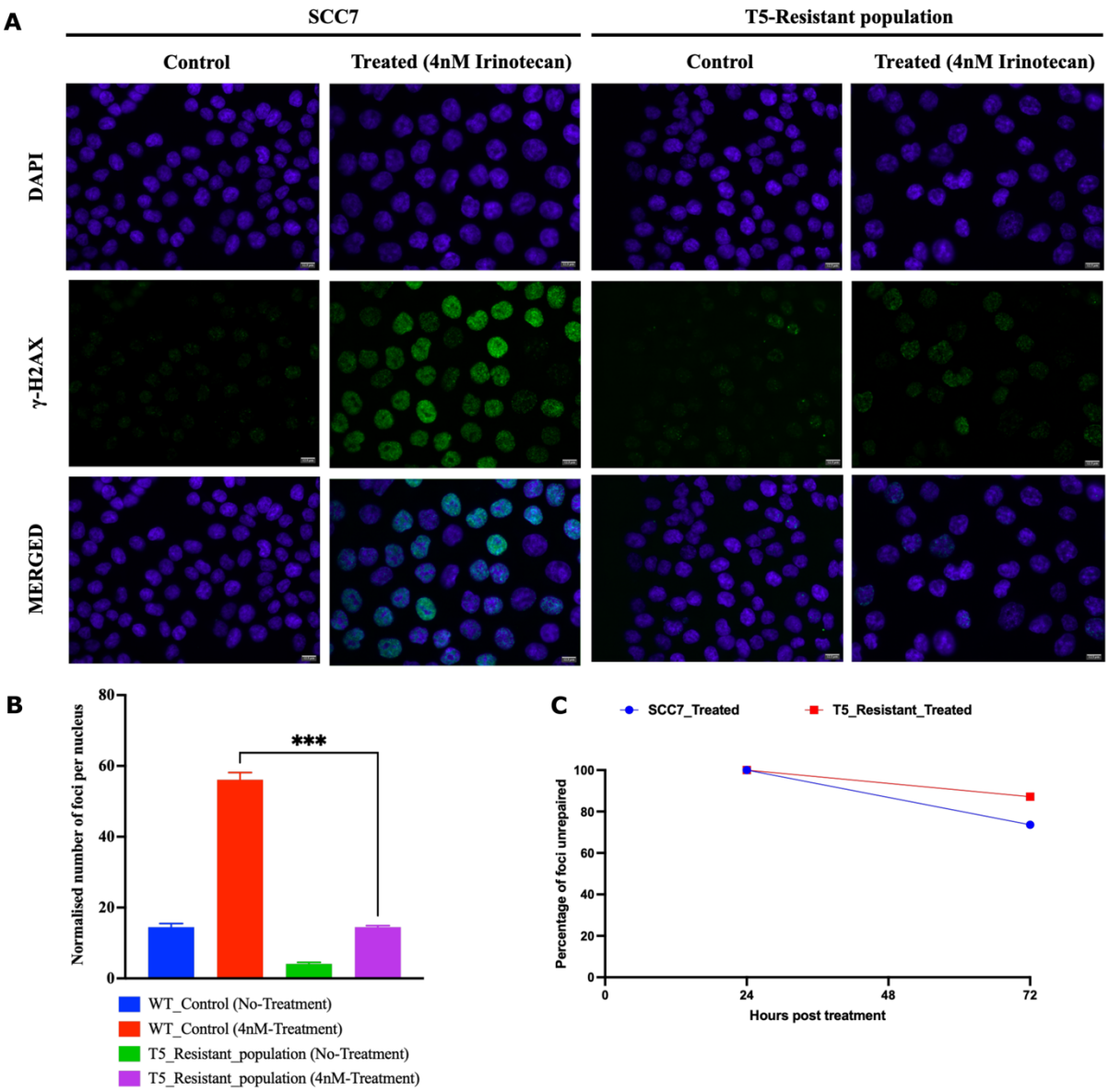


Figure 6. Adapted population experiences lower frequency of DSB following irinotecan exposure. **(A)** γH2AX foci in cells exposed to irinotecan (4nM for 24 hours). SCC7 cells are compared to the same clone but after five cycles of adaptation to 4nM irinotecan. Experiment was conducted in biological triplicates. **(B)** Quantification of data presented in Fig. 6A of the number of foci generated 24-hour post treatment, n=368 images were analyzed with the integrated software (refer material and methods). **(C)** Line plot shows the number of foci remaining after 72-hour of

recovery from irinotecan (SCC7-73%, T5_resistant-87%), indicating that the rate of DSB repair is not faster in adapted cells, n=431 images were analyzed using integrated software (refer material methods). Statistics were calculated using GraphPad Prism version 9.0.0, California USA. The significance of differences was determined using unpaired Welch's correction, two-tailed t-test (*$P < 0.0332$, **$P < 0.0021$, ***$P < 0.0002$, ****$P < 0.0001$) denoted in above in figure (B).

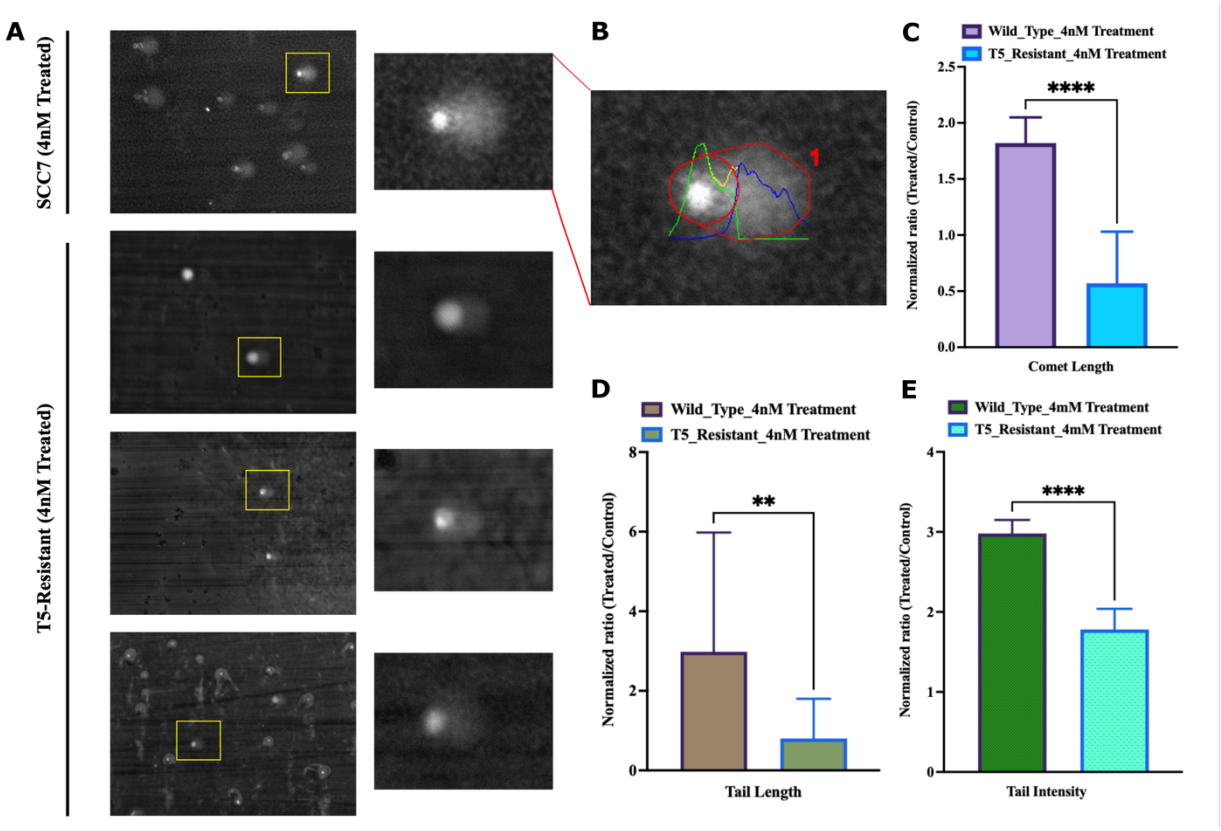


Figure 7. Adapted population experiences lower frequency of overall breaks following irinotecan exposure (A) Representative image of the field of SCC1 cells treated with irinotecan and corresponding comets. Insert shows enlarged image of the comet. Lower three images with inserts show representative images of the adapted population. **(B)** Representative of quantified output image used to compute parameters for comet analysis. **(C)** Comet length and **(D)** Comet Tail length, **(E)** Intensity of DNA in tail was estimated using OpenComet v1.3.1. Bar plot represents normalized ratio of treated and control comets in SCC7 and T5-adapted cells separately. Statistics were calculated as mean of n=21 comets in each using GraphPad Prism version 9.0.0, California USA. The significance of differences was determined using unpaired Welch's correction, two-tailed t-test (*P < 0.0332, **P < 0.0021, ***P < 0.0002, ****P <0.0001) denoted in above figures (C, D, E).

Supplementary Information for

Homologous recombination repair creates mutations in the non-coding genome that alter Topoisomerase-1 cleavage sites & orchestrates irinotecan resistance

Santosh Kumar[1†], Valid Gahramanov[1†], Julia Yaglom[1], Shivani Patel[1†], Lukasz Kaczmarczyk[1], Ivan Alexandrov[2], Gabi Gerlitz[1], Mali Salmon-Divon[1], Michael Y. Sherman[1*]

Affiliations:

1. Department of Molecular Biology, Ariel University, Israel-40700.
2. Research Center of Biotechnology of the Russian Academy of Sciences, Moscow, Russia-119071.

[†]Equal contribution

*Professor Michael Y Sherman.

Email: sherma1@ariel.ac.il

This PDF file includes:

Supplementary text

Figures S1 to S4

Other supplementary materials for this manuscript include the following:

Datasets S1 as (.xlsx) containing long tables and meta-datasets

Supplementary Information Text

Mutational landscape in irinotecan resistant clones.

As noted in the main text, mutations in coding and regulatory regions did not affect genes involved in cell sensitivity to irinotecan. The only exception was a mutation in the MMS22L gene that normally works in the base excision repair, but also plays a role in the DSB repair (68,69). However, this mutation was a 1 nucleotide insertion, which most likely inactivated the MMS22L gene. In such a scenario, we expect to see sensitization rather than protection from DSBs. Accordingly, it is unlikely that this mutation associates with adaptation to irinotecan.

To understand the relevance of the mutations to the adaptation process, we manually identified the precise positions of breaks in 1,000 out of 2,829 NHEJ-repaired common mutation sites and analyzed if there are any shared features in sequences of the sites. About 80% of these breaks took place either within or at the edge of 1-4 bp repeat clusters, and the length of these repeats usually was within the range of 5-30bp (see examples of such repeats in Table S8). Though we could not define the breaking points in HR repaired sites precisely, we observed that 75% of these sites are in similar repeats as NHEJ breaking points, and therefore assumed that breaks in these sites occurred in or at the edges of such repeats, as with NHEJ. Overall UCSC analysis showed that the mutations were mostly located in the repetitive elements such as microsatellites (Table S14), simple repeats (Table S15) and satellites (Table S16 for mutations in chromosomal arms and supplement (Table S17) for mutations in pericentromeric genomic locations). Parallel RepeatMasker analysis supports these findings (Fig S1A, Table S18 and S19 for detailed data on repetitive elements and mutation correlation) where specific satellites such as alpha satellites and human satellite-II is shown to have majority of common mutations (Fig S1B). In other words, repetitive elements

accommodate majority of all mutations, including common mutations between the clones, during the adaptation process.

Upon analysis of distribution of mutations along the chromosomes, we found that there was a disproportionally around 40% of the total mutations densely populated regions adjusted to centromere (up to 5Mb depending on the chromosome) at both sides (Fig. S2), which we referred to as "pericentromeric" areas. This region is identified generally by signatures of H3K9me3 and H3K27me3 followed by richness of repetitive elements such as satellites and simple repeats (Fig S2A, B and Fig S3). As a representative example, Figure S2a shows correlation between pericentromeric locations of common mutations on chromosome-1, pericentromeric chromatin modification signatures, and locations of satellites. For comparison, Fig. S2B shows these distributions and parameters in a chromosomal arm of same chromosome-1. Whereas, Fig. S3 visualizes the overall distribution of common mutations in genome and show its specific enriched presence in centromere and pericentromeric regions.

A very common sequence at the breaking points in the chromosome arms was a polyC stretch (25% of all mutation sites in the arms). The length of these polyC sequences associated with breaking points usually were within the range of 20-40bp. To our surprise, in 100% cases of LOH in these regions, original long polyC was substituted with polyC stretch that was interrupted by multiple SNPs (Fig. S4). The fact that these breaks were generated in response to Top1 inhibition by irinotecan suggests that Top1 either binds to such sites specifically, or stalls there in the process of its function, or is activated at these sites.

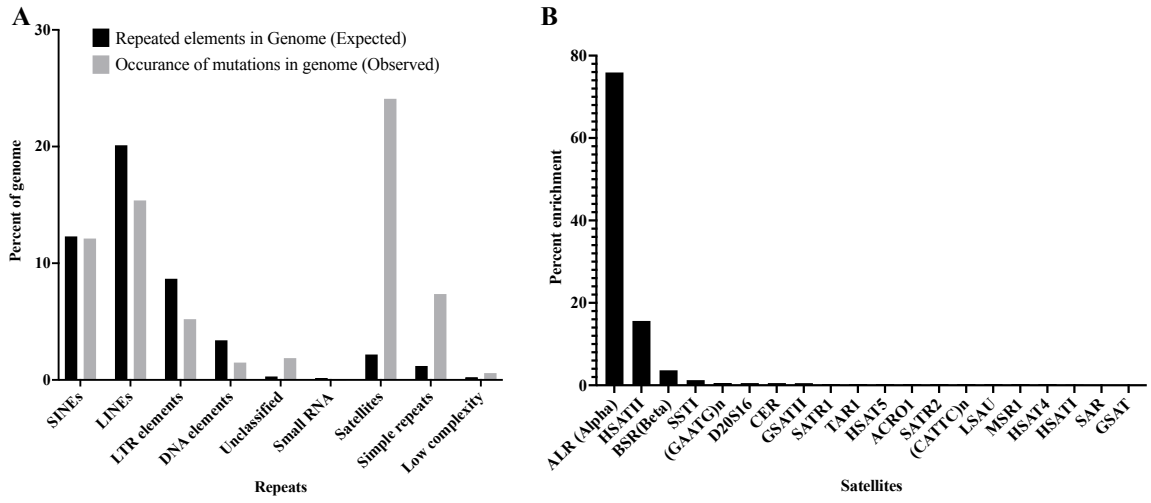


Fig. S1. (A) Distribution of repeated elements in human genome (Expected) and fraction of mutations in MSC clones found in these repeats. common mutation sites were taken randomly (100bp upstream and downstream of SNP's) and run through the RepeatMasker (v4.1.0) using nhmmscan version 3.3.1 (Jul 2020). The presence of mutations in SINE corresponds to the presence of SINE in the genome, indicating lack of the enrichment. Strong enrichment was found in satellites and simple repeats, (B) In-depth analysis for classification of satellites shows vast majority of common mutations are present in ALR and human satellite II followed by BSR type in all three clones.

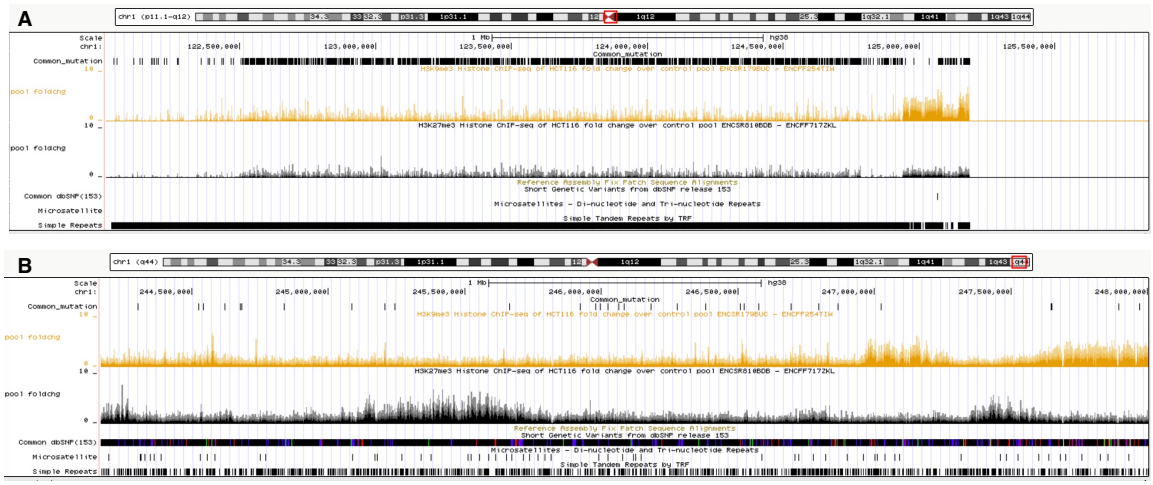


Fig. S2. Representative example for pericentromeric and chromosomal arm: (A) plot shows correlation between common mutations in "pericentromeric region" marked by H3K9me3 and H3K27me3, where mutations are densely accommodated and covered by simple repeats. Mutations in these genomic locations are not common dbSNP's and is accommodated by fewer microsatellites, (B) shows arm portion of chromosome with comparatively much fewer common mutations and scattered simple repeats in sliding window UCSC browser.

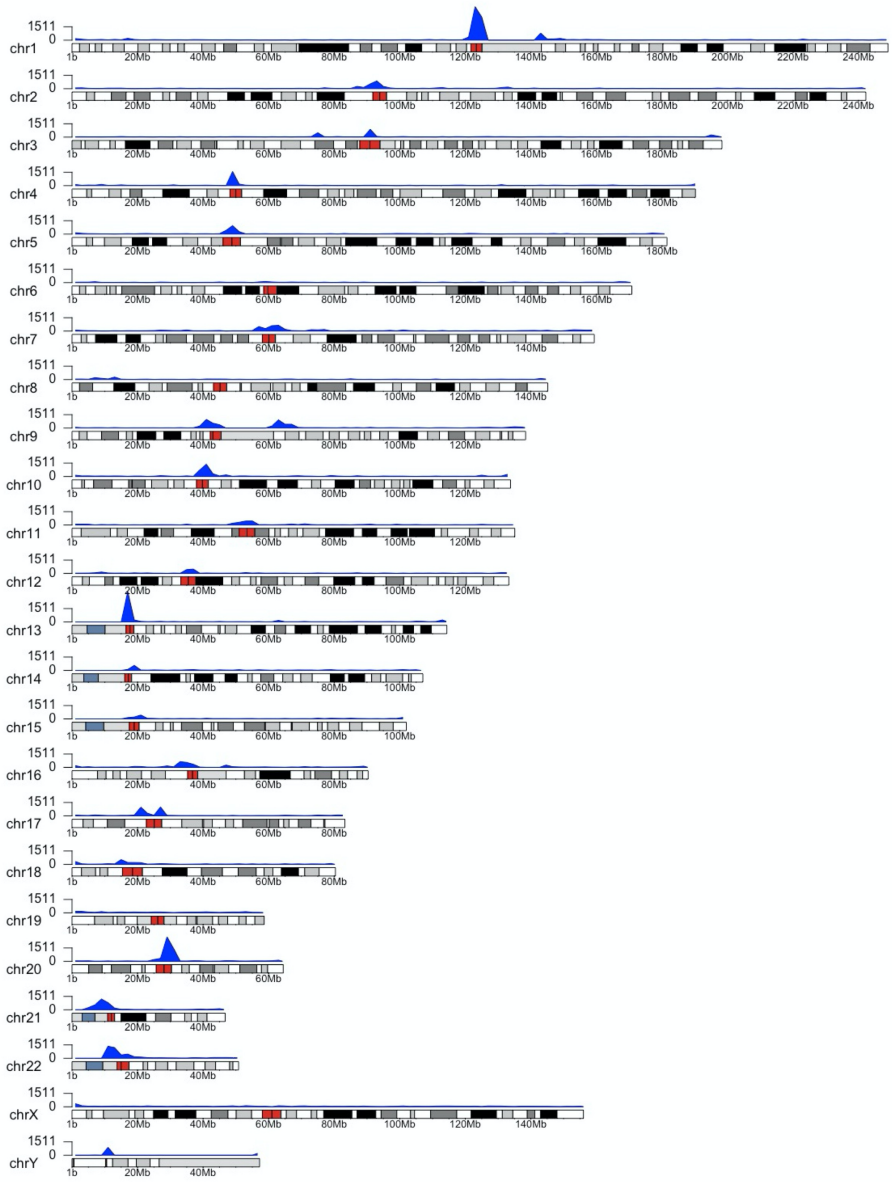


Fig. S3. Disproportionate distribution of common mutations in genome: Karyoplot showing enrichment of mutations (blue peaks) in pericentromeric region (marked with red on cytoband) where genomic positions have been plotted against SNP density in corresponding regions using R package (karyoploteR 1.18.0).

Fig. S4. Representative example mutations in polyC sequences. Comparative alignment of a polyC sequence in SSC1 with reference genome (hg38) and resistant sub-clones. Allele-1 and Allele-2 represent two alleles in homologous chromosomes. Sequence marked by (both red) alleles represents homozygosity (0/0) as shown in reference genome strand and resistant sub-clones (top). Alleles (red and blue) in parental line represent heterozygosity (0/1) with "Cs" base. After the irinotecan adaptation, generated mutations lead to LOH, so that the polyC tract became interrupted as in normal reference genome alleles.